Just thought I would save people some time, in case you can’t attend all 600+ Nvidia GTC conference sessions 😊… I put together a “brief” cheat sheet of what was announced/new/highlighted during our GTC conference this week.

Each item has a brief description, so you can scroll though this in a few minutes and be caught up! Feel free to share internally

Also, we launched a cloud service for our H100 technology, as well as AI as a cloud service which the team might find relevant.

If anyone wants to catch the sessions, it’s free: <https://bit.ly/3ZnytL8>

There are specific industry Agendas as well, here are a couple but let me know if you can’t see the specific sector:

**Targeted Agendas by Topic**

* [Healthcare](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.nvidia.com%2Fgtc%2Fsessions%2Fhealthcare%2F&data=05%7C01%7Cjeffchang%40nvidia.com%7C87533e390ba34542b64808db0ba3b303%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638116570230433449%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=S85U7Hdl5MRXa5eCMoYvbuWPWCbk4VZsL4BEUMaq3gc%3D&reserved=0)
* [General Technology](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.nvidia.com%2Fgtc%2Fsessions%2Ftechnology%2F&data=05%7C01%7Cjeffchang%40nvidia.com%7C87533e390ba34542b64808db0ba3b303%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638116570230277238%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=o%2FAqL9MKH1HlNNp%2Bh%2FJoVn%2FRg7FIFxnByi6C109SpZM%3D&reserved=0)
* [Accelerated Networking](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fregister.nvidia.com%2Fevents%2Fwidget%2Fnvidia%2Fgtcspring2023%2F1675398525525001MyZH&data=05%7C01%7Cjeffchang%40nvidia.com%7C87533e390ba34542b64808db0ba3b303%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638116570230277238%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=PbtQg9Me8FSFsuxwx1u6THcr%2FeUsFuXFWwTWmRcX18o%3D&reserved=0)
* [AI-Powered Edge](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fregister.nvidia.com%2Fevents%2Fwidget%2Fnvidia%2Fgtcspring2023%2F1675197522876001JlMs&data=05%7C01%7Cjeffchang%40nvidia.com%7C87533e390ba34542b64808db0ba3b303%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638116570230277238%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=%2F76OsElQ%2FZe14oCzFS14FxBzMfOZtUd3SAlBUWwo9aM%3D&reserved=0)
* [Business Leader](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.nvidia.com%2Fgtc%2Fsessions%2Fexecutives%2F&data=05%7C01%7Cjeffchang%40nvidia.com%7C87533e390ba34542b64808db0ba3b303%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638116570230433449%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=LdY%2FChgfgGxB2wwS%2BEWaGITU8AcxYQyoOMbZpgvTguE%3D&reserved=0)
* [Grace CPU](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fregister.nvidia.com%2Fevents%2Fwidget%2Fnvidia%2Fgtcspring2023%2F1675400584071001xvU5&data=05%7C01%7Cjeffchang%40nvidia.com%7C87533e390ba34542b64808db0ba3b303%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638116570230433449%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=CzIqgNb0mb2X%2BWC1jNEle3IVkumn%2BbWKke1uKrXHzOA%3D&reserved=0)
* [Higher Education and Research (Educators/Researchers)](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fregister.nvidia.com%2Fevents%2Fwidget%2Fnvidia%2Fgtcspring2023%2F1675313370624001vd0G&data=05%7C01%7Cjeffchang%40nvidia.com%7C87533e390ba34542b64808db0ba3b303%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638116570230433449%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=7qmS%2F1y6J4SAOGk1iZfL1umJYX1QWngke5vRUOMaFWc%3D&reserved=0)
* [NVIDIA AI Research Day](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.nvidia.com%2Fgtc%2Fsessions%2Fresearch%2F&data=05%7C01%7Cjeffchang%40nvidia.com%7C87533e390ba34542b64808db0ba3b303%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638116570230589670%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=vbuQ0Lldl0ZonS2LBIrtH5U9xoCMDZx%2Fq0MI73nlSvk%3D&reserved=0)

**ENTERPRISE ANNOUNCEMENTS**

* NVIDIA Inference Platforms
* NVIDIA L4
* NVIDIA H100 NVL
* NVIDIA Grace CPU Superchip and Grace Hopper Superchip
* NVIDIA H100 Production Systems Available
* NVIDIA-Certified Systems
* DGX Cloud
* DGX Platform
* NVIDIA Omniverse Enterprise
* NVIDIA Omniverse Cloud
* NVIDIA Omniverse & Microsoft Azure Collaboration
* NVIDIA OVX 3.0
* NVIDIA Bluefield-3 DPU and DOCA 2.0
* NVIDIA Jetson Orin Nano Developer Kit
* NVIDIA IGX Orin Developer Kits
* NVIDIA AI Enterprise
* NVIDIA NeMo Framework
* NVIDIA NeMo Service
* NVIDIA Picasso Service
* NVIDIA AI Workflows
* NVIDIA Merlin
* NVIDIA cuOpt
* NVIDIA Morpheus
* NVIDIA Riva
* NVIDIA RTX 4000 SFF Ada Generation
* NVIDIA RTX Ada Generation Laptop GPUs
* NVIDIA Project Mellon
* NVIDIA CloudXR 4.0
* DLSS 3 Frame Generation
* RTX Path Tracing
* NVIDIA vGPU
* NVIDIA Quantum Platform
* NVIDIA cuLitho
* NVIDIA Modulus
* NVIDIA RAPIDS RAFT
* NVIDIA TAO Toolkit
* NVIDIA Deepstream SDK
* NVIDIA Metropolis Microservices
* NVIDIA Triton
* NVIDIA JAX
* NVIDIA HPC SDK - Minor Update v23.3
* NVIDIA CUDA Toolkit
* cuNumeric and Legate
* Warp (Omniverse)
* Nsight Developer Tools
* Automotive
* Healthcare - Clara
* NVIDIA Robotics
* AT&T Supercharges Operations with NVIDIA AI

**NVIDIA Inference Platforms**

NVIDIA launched four inference platforms optimized for the diverse set of emerging generative AI applications so developers can quickly build specialized, AI-powered applications. The platforms combine NVIDIA’s full stack of inference software with the latest Ada, Hopper, and Grace Hopper processors.

* **NVIDIA L4 for AI Video** delivers 120X more AI-powered video performance than CPUs and 99% better energy efficiency.
* **NVIDIA L40 for Image Generation** is optimized for graphics and AI-enabled 2D, video, and 3D image generation. The L40 platform is the engine of NVIDIA Omniverse.
* **NVIDIA H100 NVL for Large Language Model Inference** is ideal for deploying massive LLMs like ChatGPT at scale with mainstream servers.
* **NVIDIA Grace Hopper for Recommendation Models** is ideal for graph recommendation models, vector databases, and graph neural networks.

**NVIDIA L4**

The NVIDIA L4 Tensor Core GPU is a universal, energy-efficient GPU powered by the NVIDIA Ada Lovelace architecture. It’s delivered in a low-profile PCIe form factor that provides a cost-effective, energy-efficient solution for high throughput and low latency in every server, from the edge to the data center to the cloud.

* NVIDIA L4 for AI Video can deliver 120X more AI-powered video performance than CPUs, combined with 99% better energy efficiency. As the most efficient NVIDIA accelerator for mainstream use, servers equipped with L4 deliver 2.7X more generative AI performance over CPU solutions, as well as over 4X more graphics performance than the previous GPU generation.
* Google Cloud is integrating the L4 platform into its Vertex AI model store and it is in private preview starting today. It is the first cloud service provider to offer L4 instances.

**NVIDIA H100 NVL**

As part of the inference platform press release, the new NVIDIA H100 NVL for Large Language Model Inference was introduced.

* NVIDIA H100 NVL for Large Language Model Deployment is ideal for deploying massive LLMs like ChatGPT at scale in mainstream servers.
* 2x PCIe-based GPUs with NVLink Bridge, 188 GB of HBM3 memory, and 7.8 TB/s of memory bandwidth combine to make H100 NVL a supercharged inference platform that can be deployed in mainstream servers at scale.
* H100 NVL and Transformer Engine acceleration deliver up to 12X faster inference performance on GPT-3 compared to the prior generation HGX A100 at data center scale.

**NVIDIA Grace CPU Superchip and Grace Hopper Superchip**

NVIDIA announced the Grace Hopper Inference Platform for recommender systems, graph neural networks, and vector databases. Grace Hopper is part of a family of NVIDIA Inference Platforms, including NVIDIA L4, L40, and H100 all for different size and type of inference problems.

* Grace Hopper has almost 600GB of Fast Access Memory for the GPU which enhances Inference performance on large models. It combines up to 96GB of HBM3 at 4TB/s and up to 480GB of LPDDR5X at 512GB/s, to provide 7X the amount of Fast Access Memory available to traditional GPUs.
* With a 900 GB/s NVLink-C2C connection between CPU and GPU, Grace Hopper can deliver 7X faster data transfers compared to PCIe Gen 5.

NVIDIA also released a blog, [Green Light: NVIDIA Grace CPU Paves Fast Lane to Energy-Efficient Computing for Every Data Center](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DOSD44RUAmUaiZswYKplV3w%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173773147263%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=frRvzQN6aczXMwGN0nqdZR0NrGPCTgKcQLAz05X8zsI%3D&reserved=0), that highlights the performance and energy efficiency of the NVIDIA Grace CPU for Cloud, Enterprise, and HPC applications.

* NVIDIA Grace CPU shows 2X gains over x86 in energy-efficient performance on microservices, analytics, simulations, and more.
* NVIDIA Grace combines 72 Neoverse V2 cores, an ultra fast fabric and the first CPU to use server-class LPDDR5X memory. The memory provides 50% more memory bandwidth at similar cost, but one-eighth the power consumption of typical server memory.

**NVIDIA H100 Production Systems Available**

Production H100 NVIDIA-Certified Systems available from OEM partners and cloud instances announced by Cloud Service Partners.

* H100 systems are available in production across all configurations for on-prem and cloud.
* H100 PCIe systems have been available since last year but are now in the NVIDIA Certified catalog.
* HGX H100 systems are available from major builders in addition to our DGX H100 starting production now.

**DGX Cloud**

NVIDIA DGX Cloud is an AI training service designed for the unique demands of enterprise AI. With integrated software and infrastructure optimized for multi-node training, DGX Cloud provides a full-stack AI developer suite, leadership-class infrastructure, direct access to AI experts, and enterprise support, allowing businesses to get started immediately on AI innovation, with predictable, all-in-one pricing.

* DGX Cloud is a part of the DGX platform product portfolio.
* Enterprises can now access an AI supercomputer in the cloud.
* DGX Cloud is designed and implemented by NVIDIA and DGX Cloud infrastructure is hosted to NVIDIA’s design and specifications through partnerships with Oracle Cloud, Microsoft Azure, and Google Cloud Platform.
* DGX Cloud includes Base Command Platform, NVIDIA AI Enterprise, Enterprise Support, and access to NVIDIA AI Experts.

**NVIDIA DGX Platform**

Built from the ground up for enterprise AI, the NVIDIA DGX platform incorporates the best of NVIDIA software, infrastructure, and expertise in a modern, unified AI development solution spanning clouds and on-premises.

**What’s new:**

* What was previously the DGX “family” or “portfolio”, is now the “DGX platform”. The DGX platform includes: NVIDIA DGX Cloud, NVIDIA Base Command, NVIDIA DGX SuperPOD, NVIDIA DGX BasePOD, NVIDIA DGX H100 systems, NVIDIA DGX A100 systems
* All DGX-related web pages have been updated to support [DGX Platform](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DtHKTIF320E6xLc7ALcFA_g%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173773615926%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=EM%2Br9vrwUhmblzNKCM4jrliru7wL4VebiyukX8twQdM%3D&reserved=0), [DGX Cloud](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DqjWiKaI9Gkepaao-6wr9jQ%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173773615926%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=eBkjtPlCmxkyzgA%2FEH%2B6SEPYw2WiLLTNQlUFUqSjMlc%3D&reserved=0), etc., with a simplified user web journey
* NVIDIA DGX H100 is now in full production

**NVIDIA Omniverse Enterprise**

[NVIDIA Omniverse™ Enterprise](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3Dv3jmSnk2_E67LKrkrpuZWA%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173773772151%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=TVZi1%2Bi9ATZSGx7ZWcTNExsqmKbXz0U9fR12rDcXmx8%3D&reserved=0) is a scalable, end-to-end platform enabling enterprises to develop custom 3D pipelines and simulate large-scale, physically accurate virtual worlds. In the GTC keynote and pre-GTC bootcamp, we outlined our focus on enabling automotive workflows end-to-end.

* New BYOL Enterprise Nucleus Server deployment options for Azure, AWS, and GCP
* Omniverse on Launchpad: Lead with [Bring Your Own Data](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DHugA6S3yB0ejzHD2MqP0Mg%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173773772151%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=331S22zFfT%2BuP9UDEJ63YJHKqA7TuDRIuJQsz5klr%2F4%3D&reserved=0) lab for your customers. Use [SFDC Trial Nomination Form](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DaaBaGJ6zpUiooGpbZR9q-g%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173773928368%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=wtEUUnijRiySM%2FWrxj8IKfV7MAtfKjExZDd4hgSW%2BjA%3D&reserved=0) for qualified contacts.
* New Omniverse for automotive content

**NVIDIA Omniverse Cloud**

NVIDIA Omniverse Cloud is a platform-as-a-service providing developers and enterprises a full-stack cloud environment to design, develop, deploy, and manage industrial metaverse applications.

* Cloud agnostic, deploying on Microsoft Azure first
* Available as enterprise private offer – April 2023 in the Azure Marketplace
* Deployed on NVIDIA OVX 2.0 servers, purpose-built for massive, complex real time 3D workflows

**NVIDIA Omniverse & Microsoft Azure Collaboration**

We announced a collaboration with Microsoft to provide hundreds of millions of Microsoft enterprise users access to powerful industrial metaverse and AI supercomputing resources via the cloud. There are three main initiatives:

* Microsoft Azure selected as first CSP for Omniverse Cloud deployment
* NVIDIA and Microsoft are working to connect Omniverse to the Microsoft 365 suite of applications including Teams, OneDrive, and SharePoint, enabling more seamless enterprise collaboration across digitalization workflows and exposure of Omniverse to Microsoft’s nearly 300M users
* NVIDIA and Microsoft are working to connect Omniverse to Azure Cloud Services – Digital Twins and Internet of Things, enabling enterprises to build and operate more accurate, dynamic, fully functional 3D digital twins that automatically respond to changes in their physical environments.

**NVIDIA OVX 3.0**

NVIDIA OVX™ is a computing system reference architecture optimized to power the creation and operation of Omniverse Enterprise applications at data center scale. NVIDIA OVX systems based on this reference architecture, are purpose-built to deliver the graphics, compute, AI, and networking performance required by the most demanding 3D models and simulations powered by NVIDIA Omniverse Enterprise.

We are announcing the 3rd generation of NVIDIA OVX™, OVX 3.0. The third generation builds upon the capabilities of previous generations by introducing Bluefield-3 as part of the architecture and features a new, balanced, design optimized for performance and scale to accelerate the next-generation of hyperscale Omniverse workloads.

**NVIDIA BlueField-3 DPU and DOCA 2.0**

NVIDIA BlueField-3 data processing units (DPU) are now in full production, and have been adopted by Oracle Cloud Infrastructure (OCI) to achieve higher performance, efficiency, and security performance. By offloading, accelerating, and isolating the data center control-plane, BlueField-3 DPUs create a secure, accelerated, and sustainable infrastructure for running AI and other modern workloads. Accelerated computing with BlueField-3 is saving power and cost, enabling cloud providers to operate massive-scale data centers in the face of power constraints. NVIDIA is integrating BlueField-3 across its data center computing platform including the NVIDIA OVX 3.0 architecture and the NVIDIA HGX H100 AI Supercomputer design. In addition, top OEM partners are building accelerated compute platforms powered by BlueField-3 DPUs.

The NVIDIA DOCA 2.0 software framework now supports BlueField-3 DPUs, empowering thousands of developers to tap into the power of the third-generation DPU platform to rapidly create accelerated applications and services. With extensive libraries, drivers and API's, DOCA is a one stop shop for BlueField DPU developers, and is the key to accelerating infrastructure services in the cloud.

With more than 4,700 Early Access developers already using DOCA to create BlueField applications, NVIDIA is excited to announce DOCA general availability, which opens DOCA access to everyone. In addition, DOCA 2.0 provides a wide range of enhancements, focused on the creation of accelerated networking and security applications.

**NVIDIA Jetson Orin Nano Developer Kit**

Today, NVIDIA announced the new [Jetson Orin Nano Developer Kit](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DQ20mEOO5J0unsNqrW-Gcjg%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173774240812%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=dm3ryEVWU48PZjiidQ4fkPgDPSCKpaeDjKdcv5%2BwaFg%3D&reserved=0) that sets a new standard for creating entry-level AI-powered robots, smart drones, and intelligent vision systems and simplifies getting started with the Jetson Orin Nano series. It delivers 80X the performance compared with the previous generation Jetson Nano, enabling developers to run advanced robotics and vision transformer models. The Jetson Orin Nano Developer Kit has the following key features:

* Jetson Orin Nano 8GB module with heat sink and reference carrier board
* 40 INT8 TOPS powered by Ampere GPU and 6-core Arm® Cortex-A78AE v8.2 64-bit CPU
* 8 GB 128-bit LPDDR5 memory, 68 GB/s
* The carrier board can accommodate all Orin Nano and Orin NX modules due to 100% pin and form factor compatible
* Powerful NVIDIA AI software stack with support for SDKs such as NVIDIA JetPack, NVIDIA RIVA, NVIDIA DeepStream, NVIDIA Isaac ROS, etc.

The full lineup of the [NVIDIA Jetson Orin](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DvkMBw1ErsUKCZxUfk4b84A%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173774240812%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=N6k8SgTYlqQRs%2BinCRnDx6b%2FtgHoBlkUd0ubjnWJriU%3D&reserved=0) family of commercial modules is available now, including the Jetson AGX Orin series, the Jetson Orin NX series, and the Jetson Orin Nano series. The Jetson AGX Orin Developer Kit has also now been upgraded to include 64 GB of memory for the same price of $1,999. JetPack SDK is targeted to release on March 28th.

**NVIDIA IGX Orin Developer Kit**

NVIDIA IGX Orin developer kits are new industrial and medical grade edge AI developer kits that allow developers to build applications that advance human and machine collaboration across industrial and medical environments to improve productivity and safety. IGX Orin developer kits include:

* NVIDIA IGX Orin Module, ConnectX-7 SmartNIC, BMC, Safety MCU
* Chassis and power supply
* NVIDIA SDKs: Metropolis and Holoscan

**NVIDIA AI Enterprise**

With over 50 frameworks, pretrained models and development tools, NVIDIA AI Enterprise, the software layer of the NVIDIA AI platform, is designed to accelerate enterprises to the leading edge of AI, while also simplifying AI to make it accessible to every enterprise. At GTC Spring 2023, NVIDIA announced the latest version of NVIDIA AI Enterprise with new features that reduce the time and cost of enterprise AI, enhancements for hybrid cloud environments, and expanded availability in the cloud.

* **NVIDIA RAPIDS Accelerator for Apache Spark for Accelerated Data Processing:**
  + NVIDIA AI Enterprise 3.1 includes RAPIDS Accelerator for Apache Spark to enable GPU acceleration for Apache Spark 3.x
  + Optimized performance for Spark deployments with full access to enterprise-grade support on certified platforms including Amazon EMR, Google Cloud Dataproc, and Databricks on Azure and AWS
* **New AI Workflows:** two new AI workflows only made available with NVIDIA AI Enterprise subscription
  + Next Item Prediction - building a recommender pipeline to drive customer retention and upsell
  + Route Optimization - optimizing distribution and delivery fleets to reduce cost and time with dynamic constraints
* **NVIDIA AI Enterprise Available on Cloud Marketplaces:**
  + Now available on the Google Cloud Marketplaces, and available soon on the AWS and Microsoft Azure Marketplaces
  + Customers can get started right away by procuring an on-demand instance or a private offer payable with committed cloud spend agreements, both available at the marketplaces.
  + The market place offering includes the full NVIDIA AI Enterprise software stack and enterprise-grade support.
* **CSP Managed Kubernetes Orchestration Services:** NVIDIA AI Enterprise 3.1 supports Kubernetes container hosting environments including Amazon Elastic Kubernetes Service (Amazon EKS) and Google Kubernetes Engine (GKE).
* **Support for Ada Lovelace GPUs and Enhancements for Hybrid Cloud Infrastructure:**
  + Support for NVIDIA Ada Lovelace GPUs - NVIDIA L4, L40, and RTX 6000 Ada Generation - to supercharge diverse generative AI and inference workloads
  + Support for VMware vSphere 8 u1, heterogenous profile types for mixed workloads, and MLOps certifications with ClearML

**NVIDIA NeMo Framework**

NVIDIA announced NVIDIA NeMo, part of the NVIDIA AI platform, an end-to-end, cloud-native enterprise framework to build, customize, and deploy generative AI models with billions of parameters. Highlights include:

* Expanded support across modalities including speech, language, and images.
  + Model architectures for language: BERT, GPT-3, T5, T5-MoE, Inform.
  + Model architectures for image: Stable Diffusion v1.5, VisionTransformers (ViT), CLIP, Instruct-Pix2Pix
* New LLM customization techniques: Adapters, Reinforcement Learning with Human Feedback (RLHF), and Attention with Linear Biases (AliBi)
* New GPU-accelerated data processing libraries to mine and curate high-quality text data to train large language models

**NVIDIA NeMo Service**

NVIDIA announces NeMo, a cloud service for enterprise hyper personalization and at-scale deployment of intelligent large language models (LLMs). NeMo service provides:

* 3 new [NVIDIA-built foundation models](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DsJ_uOLPdIU6Gxhcp9TiHsA%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775021924%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=bkXlQa5wqVo8pBn7dzUH6W8uHGdbRQRh%2BlObH4CS6tU%3D&reserved=0) to best meet the complexity and latency of their applications
* New community-built T5 model that supports over 101 languages
* New Inform capability that connects proprietary knowledge bases with generative pretrained models to get response on real-time data
* Model customization - define operating domain, add skills, and continuously make models smarter with various customization techniques
* Cloud APIs to customize and deploy models through enterprise applications

**NVIDIA Picasso Service**

NVIDIA is announcing Picasso service, as part of the new NVIDIA AI Foundations, that enables enterprise and ISVs to use NVIDIA developed State of the Art Edify models to generate visual content such as images, videos, and 3D content. NVIDIA is collaborating with:

* Adobe to build the next generation AI capabilities across their product lines. Integrating generative AI into world-class creative and marketing tools.
* Getty to train Edify image and video models on fully licensed image and video data. Getty will pay artists for their contributions.
* Shutterstock to train Edify 3D models on fully licensed data. Shutterstock will pay artists for their contributions.

**NVIDIA AI Workflows**

NVIDIA announced new AI workflows for [next item prediction](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DrQbMfW6ljEaHLZzyIEl85g%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775178148%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=TXR0okipTNadbsd9v2zAEvdYhrIlsMzkxzpNhwZGjS8%3D&reserved=0) and [route optimization](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DyI50Up39KESmU-YKarrvgw%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775178148%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=BPgzM8yJp62%2BC1dqhEGUwTswbPkP4KkkrKe7x6Tzmr4%3D&reserved=0). NVIDIA [AI Workflows](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DBJm3_NnMHEuT_TfLbZZu7Q%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775178148%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=5BHaRmztmVV2mDvJg8sYOHHZe9OEucbylFpCeVojRCc%3D&reserved=0) help customers get a head start with developing and deploying AI solutions. These cloud-native, packaged reference examples illustrate how NVIDIA AI frameworks can be leveraged to build AI solutions to address a range of use cases, including:

* [Intelligent Virtual Assistants](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3D1r8IY9cxI0W_e180XHOKuA%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775178148%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=fxuAfK9m6GjPWDPycTwmalpSE1%2FldUBpF%2Bsv55KHae8%3D&reserved=0)
* [Audio Transcription](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3Dhg-EBI-XPkGKhS1b6Zm05A%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775178148%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=ZMzadUUClIwfMesMoyIPc7dRbtvLF%2B4Nly3qV9LGXDg%3D&reserved=0)
* [Next Item Prediction](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DXQelDtHi3ECvPzXW59XqCQ%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775178148%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=EQ1IhtEq9nzUELMueXmcGkyaOSVYcKgGIeWeUOo6wXY%3D&reserved=0)
* [Digital Fingerprinting](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3D6EAAa1MkQE-FTgi7CGlvUA%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775178148%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=mSRb4tIi3QWXSp2M3o9Wbp3eJX4PPZsNHCj25n2kTdo%3D&reserved=0) to detect cybersecurity threats, and
* [Route Optimization](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DSbBkEkBBh0mKD_2Wj5v2cw%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775178148%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=Ffj0%2BbxVivqzZXXhHrWiBSywqWwEtdYpibfrpmkDd18%3D&reserved=0) for minimizing vehicle routing inefficiencies
* [Retail Loss Prevention](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DWypfQmAFh02XvDr-Eimn-Q%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775178148%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=ksRaO0dSa6XHG9qgSpJjd76wvydQKyy7tmZEjMU056A%3D&reserved=0)
* [Multi-Camera Tracking](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DAvTpvjvgqU6v5rinAO0M4Q%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775178148%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=NnDm15krP%2BjrPQD7B4csmHFgfAgC%2BrUQjkHAGGSfb3g%3D&reserved=0)
* [Retail Store Analytics](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DBCchuvxr-EewVZF4GUgG2g%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775178148%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=QiB%2BEQ5AzVa0lThPFTl%2F4OV0ngYcVda0IDVhuoRrNJs%3D&reserved=0)

**NVIDIA Merlin**

Today, NVIDIA announced updates to [NVIDIA Merlin](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DWF5qSJlmIEuVOSQFNSXBcw%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775646824%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=6REgwzPc3OdN6irjClQbnXRJsM1q19fDn6%2Balro50qc%3D&reserved=0), an end-to-end recommender systems framework designed to accelerate data processing, training, inference, and deployment at any scale. With this latest release, data scientists and machine learning engineers are able to build session-based recommender pipelines and deliver state-of-the-art personalization for new and infrequent users. Merlin's session-based recommenders empower data scientists and machine learning engineers to improve recommendations with little or no user data. Teams are able to enrich user interactions, decrease customer acquisition costs, and increase sales. Highlights include:

* Session-Based Recommenders Lab on Launchpad, personalize the product offer for every individual customer and predict needs and preferences from past interactions
* Next Item Prediction AI Workflow on Launchpad, easily achieve better accuracy for recommendation models
* GTC Session Recommenders Powered by NVIDIA Merlin (Live Now)

**Pricing**: Merlin isfree, Next Item Prediction AI Workflow is available for purchase with NVIDIA AI Enterprise Support

**NVIDIA cuOpt**

Today, NVIDIA announced [cuOpt](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DkoG-YYsZyUKAanMVYaUXXw%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775646824%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=RwUT3aCSyz7dfqAb5%2B4cfoM7bnEU1lMaY71CHsSaHas%3D&reserved=0) breaking world records and how operational research, logistics, supply chain, and industrial teams are able to harness accelerated optimization to save time and reduce infrastructure costs. With billions of potential feasible moves to consider, NVIDIA's flexible solvers allows teams to unlock new use cases, reimagine solutions, find new routes, and discover possibilities previously unknown. Teams are empowered to sign up and learn more about NVIDIA accelerated optimization offering that includes:

* cuOpt, a world record breaking accelerated optimization engine
* a route optimization AI workflow
* an upcoming cloud accelerated optimization service

**NVIDIA Morpheus**

Today, NVIDIA announced updates to Morpheus, a GPU-accelerated cybersecurity framework that uses AI to identify, capture, and act on threats on a scale previously impossible. Highlights include:

* New spear phishing use case that leverages a generative AI training technique which resulted in detection of 90% of targeted spear phishing emails, a 20% improvement compared to a typical phishing detection solution used today.
* Digital fingerprinting workflow enhancements including integrated training and feedback to enable more efficient training of models, and nonlinear pipeline support for easier deployment.

**NVIDIA Riva**

Today, NVIDIA announced updates to [NVIDIA Riva](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DmUCxqE349UaKEUotkyHQTA%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173775959262%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=wfj11SBpOCyxdn3FJ%2Bn2Yxhq3e9AoH9JXNwbDHkyeOI%3D&reserved=0), a GPU-accelerated speech and translation framework for building and deploying fully customizable real-time multilingual conversational AI applications. Highlights include:

* Translation support for domain-specific use cases across all industries with high quality out-of-the-box multilingual and bilingual models for English to/from Mandarin, Spanish, Russian, German, and French
* Speech recognition for additional 5 languages: Arabic, Italian, Japanese, Korean, Portuguese

(In addition to Mandarin, English (US/UK), French, German, Hindi, Russian, Spanish (Latin America/Spain))

* Audio Transcription and Intelligent Virtual Assistant workflows for accelerated development and deployment at a large scale

**NVIDIA RTX 4000 SFF Ada Generation**

The NVIDIA RTX™ 4000 SFF Ada Generation packs a powerful punch, delivering full-size performance in a compact form factor. Built on the NVIDIA Ada Lovelace architecture, the RTX 4000 SFF combines third-gen RT cores, 4th-gen Tensor cores, and next-gen CUDA cores with 20GB of graphics memory to deliver remarkable acceleration for rendering, AI, graphics, and compute workloads. Workstations powered by the RTX 4000 SFF give professionals the necessary performance, reliability, and versatility to stay ahead of the curve and succeed in today’s competitive landscape without sacrificing performance.

* Third-gen RT cores: Up to 2X the throughput of the previous generation
* Fourth-gen Tensor cores: Up to 2X faster AI performance than the previous generation with support for the FP8 data format
* Latest-gen CUDA cores: Up to 2X the single-precision floating point throughput over the previous generation
* The RTX 4000 SFF features 20GB of GDDR6 memory with ECC (error correcting code), providing more memory bandwidth than the previous generation
* The RTX 4000 SFF features 2X the video encoding performance of the previous generation
* Achieve high-quality, seamless large-scale displays with support for multi-display synchronization

**NVIDIA RTX Ada Generation Laptop GPUs**

NVIDIA is bringing the power of the NVIDIA Ada Lovelace architecture to professional laptops, accelerating advanced visual computing and AI workflows for creators, designers, and engineers on the go.

* New NVIDIA RTX 5000, 4000, 3500, 3000, and 2000 Ada Generation Laptop GPUs feature the latest streaming multiprocessor, RT Core, Tensor Core, and NVENC technology to accelerate next-gen graphics, ray tracing, AI, and video production workflows from anywhere.
* Up to 2X the performance and power efficiency compared to the previous generation.
* New NVIDIA DLSS 3 capability generates new frames with AI to dramatically boost performance.
* Power 3D multi-app collaboration in NVIDIA Omniverse for RTX 3000 and higher GPUs.
* Next-gen GPU efficiency enables RTX 3000-class GPUs in the 14-inch design segment for the first time, bringing new levels of performance to ultra-portable laptops.
* New versions of the NVIDIA RTX A1000 6GB (now with more memory and cores), and RTX A500 (now with display support capability, dependent on OEM implementation) based on the NVIDIA Ampere architecture offer powerful entry solutions to pro graphics in laptops.

**XR: NVIDIA Project Mellon**

Project Mellon is a lightweight Python package capable of harnessing the heavyweight power of speech AI ([NVIDIA Riva](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3Dj9l1Ohy9W0yQPJUhycAtXA%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173776271719%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=0LppTRRhY%2BDeGFIxJViAFFfOkH9RiK8LPwFe7j2hquA%3D&reserved=0)) and large language models ([NVIDIA NeMo service](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DnxQ7kIUnrU66HI5DEGXcZw%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173776271719%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=zlJpIMmBYCNA1uGKEdoS1t%2BLwG%2FBiivas2IiuLwJz7g%3D&reserved=0)) to simplify user interactions in immersive environments. Project Mellon enables developers of almost any application, whether in XR or flat-screen worlds, to easily add natural language understanding to their software as a new kind of human-centered, hands-free user interface. Key release features in Project Mellon 1.0 include:

* Lightweight, easily integrated Python library
* Large Language Model support provides accuracy of natural-language understanding
* Zero Shot Language Models eliminate the need for command-specific training
* Supported natural language commands with conversational and visual context
* Supports asking questions about commands and scene with natural language responses
* Simple Python API for command understanding and execution
* Web-based test application
* Automatic speech recognition (ASR), text-to-speech (TTS), neural machine translation (NMT), and large language models (LLM) can be on prem or hosted remotely, with low latency response times

**XR: NVIDIA CloudXR 4.0**

At GTC we announced the latest release of NVIDIA CloudXR, which enables developers to customize CloudXR for their applications and customers, and scale extended reality (XR) deployment across the cloud, 5G Mobile Compute Edge (MEC) and corporate networks. CloudXR 4.0 introduces a variety of new features and benefits including:

* **The CloudXR Server API**, which gives developers the option to build CloudXR directly into their applications, eliminating the server-side need for an OpenVR or OpenXR runtime. NVIDIA is also working with the team at Collabra to offer the ability to use Monado OpenXR as the API for CloudXR, and CloudXR 4.0 continues to fully support the OpenVR API via the SteamVR runtime.
* **The Unity Plug-in for the CloudXR Client** uses the Unity XR API and other Unity APIs to provide developers a way to create a full featured CloudXR Client built on the Unity engine that can be deployed across many XR client platforms.
* **Low Latency Low Loss Scalable Throughput (L4S) –** CloudXR’s latest version offers a “togglable” implementation of the advanced 5G packet delivery optimization, built to significantly reduce lag in interactive cloud-based video streaming.
* **Generic Controller Support.** With CloudXR 4.0, developers of CloudXR clients can create input mappings for their unique HMD controllers.
* **Call Back-Based Logging.** This allows for greater flexibility in handling log messages where applications can specify the callback function and perform actions with the log messages, such as filtering and writing to a file.
* **Flexible Stream Creation.** Devices are typically associated with a particular type of video stream, for example a mono RGB stream for a tablet or a stereo RGBA stream for a mixed-reality HMD. With CloudXR 4.0, developers can indicate which streams to initiate for a particular device or platform.

**DLSS 3 Frame Generation**

Frame Generation is the new performance multiplier in DLSS 3 that utilizes AI to create entirely new frames. This breakthrough has made real-time path tracing - the next frontier in video game graphics - possible. Since its announcement, over 25 top games and applications use DLSS 3 to deliver realistic graphics with incredible performance, including A Plague Tale Requiem, Portal with RTX, and Cyberpunk 2077. In some cases FPS have almost tripled.

* Frame Generation is available as a plugin for developers through our Streamline 2.0 SDK
* DLSS 3 Frame Generation plugin coming to Unreal Engine in its next release

**Pricing:** Free download under MIT license

**RTX Path Tracing**

Path-Tracing accurately recreates the physics of all light sources in a scene to reproduce what the eye sees in real life. This new SDK gives developers the flexibility and customizability to take advantage of proven NVIDIA technologies to suit the following use cases:

* Building a reference path tracer to ensure your lighting during production is true to life, accelerating the iteration process.
* Building high quality photo modes for RT capable GPUs or real-time ultra quality modes that takes advantage of the Ada architecture

RTX Path Tracing is the culmination of decades of NVIDIA research. This SDK demonstrates best practices for building a path tracer using the latest versions of:

* DLSS 3 for super resolution and frame generation, to multiply performance.
* RTX Direct Illumination for efficient sampling of a high number of shadow casting and dynamic lights.
* NVIDIA Real Time Denoisers for high performance denoising of all light sources.
* Opacity Micro-Map for improving RT performance in scenes with heavy alpha effects
* Shader Execution Reordering for improving shader scheduling, thus increasing performance.

**NVIDIA Virtual GPU**

NVIDIA virtual GPU (vGPU) software enables powerful GPU performance for workloads ranging from graphics-rich virtual workstations to data science and AI, enabling IT to leverage the management and security benefits of virtualization as well as the performance of NVIDIA required for modern workloads. At GTC, we are announcing the following features:

* Support for NVIDIA L4 GPU to accelerate entry to mid-range virtual workstations with NVIDIA RTX Virtual Workstation (vWS) and knowledge worker workloads with NVIDIA Virtual PC (vPC)
* Please see NVIDIA vGPU 15.2 [documentation](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DQOFTYh7RS0KxwUNJ4zYzEQ%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173776427923%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=SqR0Ih9MXxC%2FI97%2BKr10lZZBGPr9SLwa2oo1MOxjnSY%3D&reserved=0) for additional features

**NVIDIA Quantum Platform**

NVIDIA cuQuantum is a SDK for accelerating quantum circuit simulation. cuQuantum enables the quantum computing ecosystem to solve problems at the scale of future quantum advantage, accelerating the development of algorithms and the design/validation of quantum hardware.

NVIDIA CUDA Quantum (formerly known as QODA) is an open QPU-agnostic development platform for hybrid quantum-classical computing. By providing a hybrid quantum/classical programming model that is interoperable with today’s most important scientific computing applications, we are opening up the programming of quantum computers to a massive new class of domain scientists and researchers.

DGX Quantum is a hardware and software system which combines the NVIDIA Grace Hopper Superchip and the CUDA Quantum open-source programming model to create the most potent accelerated computing platform. This computing platform is then integrated with OPX+, the world's most advanced quantum control platform, developed by Quantum Machines, reducing GPU-QPU latency by 1-2 orders of magnitude.

Quantum Computing Platform Highlights:

* Includes: cuQuantum libraries, CUDA Quantum, and DGX Quantum system

cuQuantum Highlights:

* Multi-Node, multi-GPU support in DGX cuQuantum appliance.
* Support for approximate tensor network methods.
* Adoption of cuQuantum continues to gain momentum, including CSPs and industrial quantum groups.

CUDA Quantum Highlights:

* New name, formerly QODA
* As of GTC Spring 2023, CUDA Quantum is in public beta
* Single-source C++ and Python implementations as well as a compiler toolchain for hybrid systems and a standard library of quantum algorithmic primitives.
* QPU agnostic. Partnering with Quantum Hardware companies across a broad range of qubit modalities.
* Delivers nearly 300X speedup over a leading pythonic framework also running on an A100 GPU

DGX Quantum System Highlights:

* Based on Grace Hopper Superchip
* Includes CUDA Quantum
* integrated with OPX+ quantum control platform from Quantum Machines
* The combination reduces GPU-QPU latency by 1-2 orders of magnitude

**cuQuantum and CUDA Quantum Pricing:** Free

**NVIDIA cuLitho**

**Addressing the challenges of computational lithography in semiconductor manufacturing:**

Manufacturing computer chips requires a critical step called computational lithography – a complex computation – involving electromagnetic physics, photochemistry, computational geometry, iterative optimization, and distributed computing.

This computational lithography step is already one of the largest compute workloads in semiconductor production, necessitating massive data centers; the silicon miniaturization evolution process exponentially amplifies the computation requirements over time.

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. This requires advanced computational lithography techniques and increased compute efficiency.

**Announcing NVIDIA cuLitho:**

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 of cuLitho include:

* **Continued Future Silicon Scaling:** Faster OPC, and enabling new lithography innovations
* **Performance:** 40x speed-up of producing semiconductor photomasks
* **Productivity:** 3X~5X more masks per day, masks that took 2 weeks now are produced overnight
* **Savings:** 500 Hopper GPU systems do the work of 40,000 CPU systems
  + Data center efficiencies - 1/9 the power, 1/8 the space
  + Saving 200,000 tons of CO2

**NVIDIA Modulus**

[NVIDIA Modulus](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3D_N9NeMh-4ECNSzW5YpcShg%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173776896607%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=r0dkTu%2FWRYNmhPrQKKOCqlweKe0cYl1kGqPX90I%2BUMs%3D&reserved=0) is a framework that provides a customizable training platform for developing physics machine learning (neural network) models. The Modulus training platform enables combining AI and Physics to create high fidelity, parameterized surrogate models. Such models can be integrated with simulation platforms to accelerate traditional simulations. Such surrogate models can also infer in near real time with high fidelity, serve as robust digital twins and are accelerating the digital transformation in Industrial, Earth science, and Life sciences domains. It’s now available as open source under the Apache 2.0 license. We are releasing Modulus v23.04. Highlights of this release include:

* Re-architected Modulus for improving data driven workflows for AI developers
* New model architectures - graph neural network support, new training recipes e.g. large-scale weather models
* Recipes for deploying Modulus trained models for inference

**NVIDIA RAPIDS RAFT**

NVIDIA announced a new library, RAPIDS RAFT, that enables GPU-accelerated vector search for Large Language Models. By integrating RAFT, vector databases and search engines can now deliver significantly faster performance for tasks such as building indexes, loading data, and executing many different query types. Highlights include:

* RAFT accelerates vector search use cases by offering accelerated Exact and Approximate Nearest Neighbor primitives on GPUs
* RAFT-powered index-building time is up to 95X faster and queries per second are up to 3X faster than CPU implementations

NVIDIA is already working with FAISS, Milvus, and Redis to bring improved vector search performance to their users by building on RAFT.

**Pricing:** Free

**NVIDIA TAO Toolkit**

NVIDIA announced NVIDIA TAO Toolkit 5.0, bringing new features to enhance AI model development. The new features include source-open architecture, transformer-based pretrained models, AI-assisted data annotation, and the capability to deploy models on any platform.

* ONNX model export that enables TAO mod
* els to be deployed on any devices, such as GPUs, CPUs, MCUs, and more
* AI-assisted annotation tool for faster and less expensive auto-generated segmentation masks
* New SOTA vision transformers for image classification, object detection and segmentation tasks
* Increased AI transparency and explainability by offering TAO available as source open

**Pricing:** Free, Enterprise version - [NVIDIA AI Enterprise](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DY0zFMDlQx0SBvzbDzQ9pCQ%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777052811%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=nff%2BzRLcctdW33w%2BE%2F1vngXJMPPfQ1ePdFTX7K7VD9k%3D&reserved=0)

**NVIDIA DeepStream SDK**

NVIDIA announced the next version of DeepStream SDK, which introduces a new graph execution runtime (GXF) that allows developers to expand beyond the open-source GStreamer multimedia framework. DeepStream’s addition of GXF is a game-changer for users seeking to build applications that require tight execution control, advanced scheduling and critical thread management. This feature unlocks a host of new applications, including those in industrial quality control, robotics and autonomous machines. Updates include:

* New runtime with advanced scheduling options
* New Accelerated Extensions
* Updated Accelerated Plugins

**NVIDIA Metropolis Microservices**

NVIDIA announced an early access program for NVIDIA Metropolis Microservices, which offers cloud-native, customizable, and enterprise-class building blocks that developers can easily adapt to accelerate vision AI applications and services. The early access version includes multiple microservices and three ready-to-use AI workflows - retail loss prevention, multi-camera tracking, and occupancy analytics (including retail store analytics). Highlights include:

* Advanced tracking and analytics capabilities across multiple cameras that help gain insights and reveal patterns from objects’ movements and behaviors.
* Configurability and customizability at all levels, from pretrained models, reference app structure, and APIs, to microservices code itself.
* Cloud-native architecture with well-defined APIs and industry standards.

**NVIDIA Triton**

NVIDIA announced key updates to [Triton Inference Server](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3D1jIQ6fpsC0CI09OHhnrliA%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777209043%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=xDwritUSEvN8vSwjRSGkM1J6DW%2F3uIQ6JChNhdvqS5s%3D&reserved=0), open-source inference-serving software, which brings fast and scalable AI to every application in production. Over 66 features were added in the last year. Notable software updates include:

* [PyTriton](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DoimU1Oe24kGyMPRyHFVG9w%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777365283%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=nRNDTsMZDsbPWH4QwO6Lp2aS9RWWc50CyiMScWz6Doc%3D&reserved=0), a simple interface to run Triton in native Python code. It enables rapid prototyping and easy migration of Python based models. ​
* Support for model ensembles and multiple concurrent model analysis in [Model Analyzer](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DpnDmR6RIlkC7fgiSQd8w2Q%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777365283%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=C37KdRNwgxT5o4H59V5GQQALjHUdC4styhLlRQX3Clc%3D&reserved=0)
* Paddle Paddle support and integration with Paddle Paddle FastDeploy
* [FasterTransformer backend](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3D0VqEPB1P7EOH05yA9xnIyw%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777365283%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=ndUABbWq4GX%2FOHmmkQhFyG5Qfa7NX%2FMFR7PTZoy80gE%3D&reserved=0) with support for BERT, Hugging Face BLOOM & FP8 in GPT
* Triton management service [[Early Access](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DfVi2G-ykCkap8k3JyLIDTw%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777365283%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=hQkDzfamsWKjb%2BXUTOR4foSEsa0mdkxlkD8fXmiZ9JU%3D&reserved=0)]: Automated and resource efficient orchestration of models for inference at scale

**NVIDIA JAX**

JAX is a framework for DL and scientific computing research and offers seamless scale out which is a critical factor in training large Transformer-based (e.g., language) models. Key features include:

* Numpy-like APIs with auto-differentiation for better compatibility
* Batching, device partitioning, and scaling for more efficient GPU usage
* End-to-end compiler optimization (e.g., XLA) with kernel fusion into single CUDA kernel – which improves performance through reduced global memory access

**NVIDIA HPC SDK - Minor Update v23.3**

The [NVIDIA HPC SDK](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DBPenSncseE6iQkkWgkCg4A%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777521510%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=M6j1YUK88%2FfkyyF4Q1b0aTuTtq2s84fLonMFWhSuiD0%3D&reserved=0) 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.

The HPC SDK offers a variety of programming models from performance-optimized drop-in libraries, to standard languages, to directives-based methods, and on to the specialization provided by CUDA. Many of the latest enhancements have been in standard language support for ISO C++, ISO Fortran, and Python. The NVIDIA HPC compilers leverage recent advances in the public specifications for these languages delivering a productive programming environment that is both portable and performant for scaling on GPU accelerated platforms.

This 23.3 update is a minor release including some new features, updated platform support, and bug fixes to the HPC SDK. Platform updates include support enhancements for ARM SLES, LLVM 16, and CUDA Fortran on NVIDIA Hopper GPUs. Updates are made to the Math Libs throughout the year as well.

**CUDA Toolkit**

CTK 12.1 is the latest runtime, driver, and toolkit minor release available today.

* Improved performance for Hopper and Ada Lovelace architectures
* Enhanced libraries (e.g., cuBLAS and cuFFT) supporting Hopper and Ada Lovelace
* Updated Compiler JIT-LTO kernel optimization

**cuNumeric and Legate**

cuNumeric v23.03 and Legate Core v23.03 are libraries designed to accelerate scientific equations in Python. NVIDIA cuNumeric is a Legate library that aims to provide a drop-in replacement library for NumPy, bringing distributed and accelerated computing on the NVIDIA platform to the Python community. Legate democratizes computing by making it possible for all programmers to leverage the power of large clusters of CPUs and GPUs by running the same code that runs on a desktop or a laptop at scale.

A Beta release is now available for download on GitHub, along with [Conda packages](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3Dstge6lduv0GOfjLN2gxBjA%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777677712%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=UPiLdKnpnyXh0pdMinAjWk0mLzUCeQwglFZFgmaV4P8%3D&reserved=0), today.

* Transparently accelerates and scales existing NumPy workflows
* Scales to up to thousands of GPUs optimally
* Requires zero code changes to ensure developer productivity

**Warp (Omniverse)**

NVIDIA Warp is a Python framework that gives coders an easy way to write GPU-accelerated, kernel-based programs in NVIDIA Omniverse™ and OmniGraph. With Warp, Omniverse developers are creating GPU-accelerated, 3D simulation workflows and fantastic virtual worlds.

Warp 0.7.2 is the latest release available in GitHub and PyPi with the following additions:

* A key change has been made to the Warp type system which now supports defining arbitrary length/precision vector and matrix types, and allows us to support a superset of USD types in Warp kernels.
* Enhanced PyTorch compatibility
* Added JAX utilities for zero-copy interop
* Added CUTLASS integration for dense GEMMs

**Nsight Developer Tools**

Nsight Systems 2023.2 releases multi-node data analysis in preview, completing the suite of Nsight Systems multi-node profiling tools. HPC developers can optimize the performance of their systems to achieve faster, more efficient computation.

* Data Analysis:
  + Nsight Systems collects performance data in multi-GPU multi-node environments up to cluster scale.
  + Parses profile data and automatically detects how throttles impact overall performance using analysis scripts called “recipes”.
  + Trace performance issues through nodes back to their source and resolve them.
  + Jupyter notebook integration for data visualization and collaboration.
* Python Profiling:
  + Optimizing Python can be challenging due to its interpreted nature and high-level abstractions, but it's the language of choice for HPC developers.
  + Nsight Systems offers automatic Python call stack sampling to support and expedite Python profiling.
* Network Profiling:
  + Inter-node network communication is a crucial facet of datacenter performance.
  + Nsight Systems samples metrics from [NVIDIA Quantum-2 InfiniBand](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DHAG3nOXtX0KYdrhSeai2rg%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777833921%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=jFaYwbVjrWMrbFIKiECG2FQMlaHv%2FtiqrqSfM26RfjU%3D&reserved=0) switches as well as NVIDIA ConnectX smart network interface cards ([smart NICs](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DYcm0ucCZhUWmBqEfNuCMAA%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777833921%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=JY4btOcVuTjqlFbvPG6t0rRmPYXqHgRwmMPZZSBzAgI%3D&reserved=0)).

Also, Nsight Graphics releases version 2023.1, featuring support for profiling opacity micromaps (OMMs) in Vulkan applications, an in-application HUD for GPU Trace, and improvements to Vulkan ray tracing profiling. Nsight Systems also releases support for OMMs.

**Pricing:** Free

**Availability:** Available now

**AUTOMOTIVE**

**NVIDIA Omniverse Cloud**

NVIDIA Expands Omniverse Cloud to Power Industrial Digitalization

* New Platform-as-a-Service Coming Soon to Microsoft Azure. Follows On-Prem Success With BMW Group, Geely Lotus, Mercedes Benz, Toyota, and Volvo Cars
* The new subscription offering for Omniverse Cloud makes it easy for automotive teams — from design and engineering to smart factory to marketing — to digitize their workflows.
* Omniverse Cloud provides automotive customers a full-stack cloud environment to design, develop, deploy and manage industrial metaverse applications.

**NVIDIA Omniverse**

BMW Group Starts Global Rollout of NVIDIA Omniverse: Automaker expands its use of the [NVIDIA Omniverse](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3DbgYhtCdAOUmT2-1x7Ydacg%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777990139%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=4i3bdsqbu%2F%2FagDZxgoh4NiQ%2BjwTqp0Ke8vIDYiXX2g8%3D&reserved=0) platform for building and operating industrial [metaverse](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnvidia.pmail4.com%2FPoliteMail%2Fdefault.aspx%3Fpage%3D0uESlzNWrUmlJ3UVavIt2Q%26ref_id%3DPQvKTA10eEq2I1PbBrLPug&data=05%7C01%7Cmdavidson%40nvidia.com%7Cd52648971bab4ad7c70808db2a33b183%7C43083d15727340c1b7db39efd9ccc17a%7C0%7C0%7C638150173777990139%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=Wv31ydJ3f%2Bw9DnLz2fbwy746HSYQIiaYBm2nugXiP18%3D&reserved=0) applications across its production network around the world, including the planned electric vehicle plant in Debrecen, Hungary, that will only start operations in 2025.

Campaign Marketing

**HEALTHCARE - CLARA**

**BioNeMo**

NVIDIA BioNeMo Service is a cloud service for generative AI in early drug discovery focusing on proteins and small molecules. Nine SOTA AI biomolecular models are now accessible in 1 place for 3D protein structure prediction, proteins and small molecule generation, property predictions, and molecular docking (prediction of the binding structure of a small molecule to a protein.) BioNeMo can be accessed through a web interface or API endpoints and outputs can be visualized and downloaded. AI models in BioNeMo can be fine-tuned and trained on customer’s proprietary data on DGX Cloud. Scientists can now:

* Generate large libraries of proteins
* Build property predictors using embeddings to refine protein libraries
* Generate small-molecules with specific properties
* Rapidly and accurately predict and visualize the 3D structure for billions of proteins
* Run large campaigns of ligand-to-small-molecule pose estimations
* Download proteins, molecules, and predicted 3D structures

**Availability:** Early Access

**Parabricks**

Parabricks 4.1 will be announced at GTC and will be fully released in Q2 2023.

* Parabricks is the only AI and GPU accelerated genomic software suite that includes gold standard applications like GATK and BWA-MEM, plus AI model DeepVariant for variant calling.
* Genomic Sequencer Agnostic - analyzes short read (e.g. Illumina, Singular, Ultima) and now long read genomic data from PacBio.
* Support for genomic analysis of PacBio long read data with accelerated Minimap2 for alignment and accelerated DeepVariant for variant calling.
* Accelerated short read germline analysis at ~16 minutes (was 22 min before) on A100 meaning that 31,000+ genomes/year can be analyzed.
* Deep Variant re-training framework so that anyone can train DeepVariant on their own data to increase accuracy of variant calling.
* Compatibility with the new NVIDIA H100 GPU which includes a powerful DPX instruction for boosting dynamic programming algorithms like Smith Waterman for local sequence alignment.

**Pricing:** Free for researchers and students; NVAIE License for enterprise support.

**MONAI**

MONAI is an enterprise grade, domain-specific, open source framework that accelerates research breakthroughs and drives AI applications into clinical impact.

* MONAI achieves milestone of 1 Million downloads demonstrating widespread adoption
* Generative AI models in MONAI Model Zoo to create large, synthetic medical imaging datasets

**Holoscan**

NVIDIA is the real-time AI platform for streaming sensor data, from surgery to satellites

* NVIDIA is announcing a collaboration with Medtronic to build an AI platform for medical devices
* Medtronic will integrate NVIDIA Holoscan and IGX into their GI Genius intelligent endoscopy module
* ORSI Academy brings Holoscan into the operating room for its first real-world, robotic-assisted surgery

**NVIDIA ROBOTICS**

**Isaac ROS**

We announced that the new Isaac ROS release, Developer Preview 3, will be available soon. Key features include:

* New LIDAR-based Grid Localizer package ​- Automatically find NAV2 Robot POSE in < 0.5 second ​
* New People Detection support in nvblox package​- GPU-accelerated 3D reconstruction for Collision Avoidance​
* Updated VSLAM and Depth Perception GEM (ESS)​
* Source release of NITROS, NVIDIA’s ROS 2 ​hardware acceleration implementation ​
* New Isaac ROS benchmark suite built on ros2\_benchmark ​
* Open-source release for multiple hardware accelerated packages​
* Support for Jetson Orin NX and Orin Nano modules

**Pricing:** N/A

**Isaac Sim**

We announced that Isaac Sim will be available to enterprises through the NVIDIA Omniverse Cloud, a platform-as-a-service providing developers and enterprises a full-stack cloud environment to design, develop, deploy, and manage applications.

* Platform-as-a-service for compute intensive workloads ​like synthetic data generation and CI/CD​
* New APIs reduce need for IT investment in cloud infrastructure

**AT&T Supercharges Operations with NVIDIA AI**

AT&T Corp. and NVIDIA today announced a collaboration in which AT&T will continue to transform its operations and enhance sustainability by using NVIDIA-powered AI for processing data, optimizing service-fleet routing and building digital avatars for employee support and training.

Solutions include:

* NVIDIA RAPIDS
* NVIDIA cuOpt
* NVIDIA Riva
* NVIDIA Omniverse ACE
* NVIDIA Tokkio
* NVIDIA NeMo