NVIDIA Brings New Generative AI Capabilities, Groundbreaking Performance to 100 Million Windows RTX PCs and Workstations

NVIDIA RTX GPUs featuring Tensor Cores are accelerating development and deployment of generative AI models, while the upcoming Max-Q low-power AI inferencing is set to improve efficiency.

Author: Jason Paul

Generative AI is rapidly ushering in a new era of computing for productivity, content creation, gaming and more. Generative AI models and applications — like NVIDIA NeMo and DLSS 3 Frame Generation, Meta LLaMa, ChatGPT, Adobe Firefly and Stable Diffusion — use neural networks to identify patterns and structures within existing data to generate new and original content.

When optimized for GeForce RTX and NVIDIA RTX GPUs, which offer up to 1,400 Tensor TFLOPS for AI inferencing, generative AI models can run up to 5x faster than on competing devices. This is thanks to Tensor Cores — dedicated hardware in RTX GPUs built to accelerate AI calculations — and regular software improvements. Enhancements introduced last week at the Microsoft Build conference doubled performance for generative AI models, such as Stable Diffusion, that take advantage of new DirectML optimizations.

As more AI inferencing happens on local devices, PCs will need powerful yet efficient hardware to support these complex tasks. To meet this need, RTX GPUs will add Max-Q low-power inferencing for AI workloads. The GPU will operate at a fraction of the power for lighter inferencing tasks, while scaling up to unmatched levels of performance for heavy generative AI workloads.

To create new AI applications, developers can now access a complete RTX-accelerated AI development stack running on Windows 11, making it easier to develop, train and deploy advanced AI models. This starts with development and fine-tuning of models with optimized deep learning frameworks available via Windows Subsystem for Linux.

Developers can then move seamlessly to the cloud to train on the same NVIDIA AI stack, which is available from every major cloud service provider. Next, developers can optimize the trained models for fast inferencing with tools like the new Microsoft Olive . And finally, they can deploy their AI-enabled applications and features to an install base of over 100 million RTX PCs and workstations that have been optimized for AI.

“AI will be the single largest driver of innovation for Windows customers in the coming years,” said Pavan Davuluri, corporate vice president of Windows silicon and system integration at Microsoft. “By working in concert with NVIDIA on hardware and software optimizations, we’re equipping developers with a transformative, high-performance, easy-to-deploy experience.”

To date, over 400 RTX AI-accelerated apps and games have been released, with more on the way.

During his keynote address kicking off COMPUTEX 2023, NVIDIA founder and CEO Jensen Huang introduced a new generative AI to support game development, NVIDIA Avatar Cloud Engine (ACE) for Games .

This custom AI model foundry service transforms games by bringing intelligence to non-playable characters through AI-powered natural language interactions. Developers of middleware, tools and games can use ACE for Games to build and deploy customized speech, conversation and animation AI models in their software and games.

From servers to the cloud to devices, generative AI running on RTX GPUs is everywhere. NVIDIA's accelerated AI computing is a low-latency, full-stack endeavor. We've been optimizing every part of our hardware and software architecture for many years for AI, including fourth-generation Tensor Cores — dedicated AI hardware on RTX GPUs.

Regular driver optimizations ensure peak performance. The most recent NVIDIA driver, combined with Olive-optimized models and updates to DirectML, delivers significant speedups for developers on Windows 11. For example, Stable Diffusion performance is improved by 2x compared to the previous interference times for developers taking advantage of DirectML optimized paths.

And with the latest generation of RTX laptops and mobile workstations built on the NVIDIA Ada Lovelace architecture, users can take generative AI anywhere. Our next-gen mobile platform brings new levels of performance and portability — in form factors as small as 14 inches and as lightweight as about three pounds. Makers like Dell, HP, Lenovo and ASUS are pushing the generative AI era forward, backed by RTX GPUs and Tensor Cores.

"As AI continues to get deployed across industries at an expected annual growth rate of over 37% now through 2030, businesses and consumers will increasingly need the right technology to develop and implement AI, including generative AI. Lenovo is uniquely positioned to empower generative AI spanning from devices to servers to the cloud, having developed products and solutions for AI workloads for years. Our NVIDIA RTX GPU-powered PCs, such as select Lenovo ThinkPad, ThinkStation, ThinkBook, Yoga, Legion and LOQ devices, are enabling the transformative wave of generative AI for better everyday user experiences in saving time, creating content, getting work done, gaming and more." — Daryl Cromer, vice president and chief technology officer of PCs and Smart Devices at Lenovo

"Generative AI is transformative and a catalyst for future innovation across industries. Together, HP and NVIDIA equip developers with incredible performance, mobility and the reliability needed to run accelerated AI models today, while powering a new era of generative AI." — Jim Nottingham, senior vice president and general manager of Z by HP

"Our recent work with NVIDIA on Project Helix centers on making it easier for enterprises to build and deploy trustworthy generative AI on premises. Another step in this historic moment is bringing generative AI to PCs. Think of app developers looking to perfect neural network algorithms while keeping training data and IP under local control. This is what our powerful and scalable Precision workstations with NVIDIA RTX GPUs are designed to do. And as the global leader in workstations, Dell is uniquely positioned to help users securely accelerate AI applications from the edge to the datacenter." — Ed Ward, president of the client product group at Dell Technologies

"The generative AI era is upon us, requiring immense processing and fully optimized hardware and software. With the NVIDIA AI platform, including NVIDIA Omniverse, which is now preinstalled on many of our products, we are excited to see the AI revolution continue to take shape on ASUS and ROG laptops." — Galip Fu, director of global consumer marketing at ASUS

Soon, laptops and mobile workstations with RTX GPUs will get the best of both worlds. AI inference-only workloads will be optimized for Tensor Core performance while keeping power consumption of the GPU as low as possible, extending battery life and maintaining a cool, quiet system. The GPU can then dynamically scale up for maximum AI performance when the workload demands it.

Developers can also learn how to optimize their applications end-to-end to take full advantage of GPU-acceleration via the NVIDIA AI for accelerating applications developer site .

Original URL: https://blogs.nvidia.com/blog/2023/05/28/computex-generative-ai-rtx/