NVIDIA Announces Record Adoption of New Turing T4 Cloud GPU

Multi-Precision Tensor Core GPU for Diverse Workloads Now Available on Google Cloud; 50+ Server Designs from Major Computer Makers

SC18 -- NVIDIA today announced that the new NVIDIA® T4 GPU has received the fastest adoption of any server GPU.

Two months after its introduction, the T4 is featured in 57 separate server designs from the world's leading computer makers. It is also now available in the cloud, with the first availability of the T4 for Google Cloud Platform customers.

“We have never before seen such rapid adoption of a datacenter processor," said Ian Buck, vice president and general manager of Accelerated Computing at NVIDIA. “Just 60 days after the T4's launch, it's now available in the cloud and is supported by a worldwide network of server makers. The T4 gives today's public and private clouds the performance and efficiency needed for compute-intensive workloads at scale."

The T4 accelerates diverse cloud workloads, including high performance computing, deep learning training and inference, machine learning, data analytics, and graphics. Based on the new NVIDIA Turing™ architecture, it features multi-precision Turing Tensor Cores and new RT Cores, which, when combined with accelerated containerized software stacks, deliver unprecedented performance at scale.

“Real-time visualization and online inference workloads need low latency for their end users. We are delighted to partner with NVIDIA to offer T4 GPU support for Google Cloud customers," said Damion Heredia, senior director of Product Management at Google Cloud. “NVIDIA T4 GPUs for Google Cloud offer a highly scalable, cost-effective, low-latency platform for our ML and visualization customers. Google Cloud's network capabilities together with the T4 offering enable customers to innovate in new ways, speeding up applications while reducing costs."

Interested customers can sign up for Google Cloud's early access program.

Consumer internet companies, including social media and online shopping sites, are among T4's early adopters and largest end-customer base.

Among server companies featuring the T4 are Dell EMC, Hewlett Packard Enterprise, IBM, Lenovo and Supermicro.

Flexible Design, Breakthrough Performance
Designed to meet the unique needs of scale-out public and enterprise cloud environments, T4 maximizes throughput, utilization and user concurrency, helping customers efficiently address exploding user and data growth.

Roughly the size of a candy bar, the low-profile 70-watt T4 GPU has the flexibility to fit into a standard server or any Open Compute Project hyperscale server design. Server designs can range from a single T4 GPU all the way up to 20 GPUs in a single node.

T4's multi-precision capabilities power breakthrough AI performance for a wide range of AI workloads at four different levels of precision, offering 8.1 TFLOPS at FP32, 65 TFLOPS at FP16 as well as 130 TOPS of INT8 and 260 TOPS of INT4. For AI inference workloads, a server with two T4 GPUs can replace 54 CPU-only servers. For AI training, a server with two T4 GPUs can replace nine dual-socket CPU-only servers.

Keep Current on NVIDIA
Subscribe to the NVIDIA blog, follow us on Facebook, Twitter, LinkedIn and Instagram, and view NVIDIA videos on YouTube and images on Flickr.

About NVIDIA
NVIDIA’s (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing — with the GPU acting as the brain of computers, robots and self-driving cars that can perceive and understand the world. More information at http://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the fastest and record adoption of NVIDIA T4 GPUs; the availability of NVIDIA T4 GPUs and its availability on Google Cloud Platform; the rate of adoption of the NVIDIA T4 GPU; support of NVIDIA T4 GPUs by a worldwide network of server makers; the benefits, performance and abilities of the NVIDIA T4 GPUs, including its ability to deliver performance at an unprecedented scale; the benefits of the partnership between NVIDIA and Google Cloud, including its ability to enable customers to innovate in new ways and speeding up applications while reducing costs; NVIDIA T4 GPUs offering a scalable, cost-effective and low-latency platform; consumer internet companies being the early adopters and largest end-customer base for NVIDIA T4 GPUs; the server companies featuring the NVIDIA T4 GPUs; NVIDIA T4 GPUs’ ability to help customers efficiently address exploding user and data growth; NVIDIA T4 GPUs’ flexibility to fit into server designs; and NVIDIA T4 GPUs powering breakthrough AI performance for a wide range of AI workloads are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Media Contacts

Kristin Bryson
+1 203 241 9190
kbryson@nvidia.com