NVIDIA Turing T4 Cloud GPU Adoption Accelerates

Baidu, Tencent, iFLYTEK, JD.com to Adopt Turing Cloud GPU to Accelerate Hyperscale Datacenters; Inspur, Lenovo, Huawei, Sugon, IPS, H3C Announce T4-Based Servers

GTC China -- Adoption of the NVIDIA® T4 Cloud GPU is accelerating, with more tech giants unveiling products and services based on what is already the fastest-adopted server GPU, NVIDIA announced today.

Following a series of announcements last week, Baidu, Tencent, JD.com and iFLYTEK have begun using T4 to expand and accelerate their hyperscale datacenters. In addition, China's leading computer makers -- including Inspur, Lenovo, Huawei, Sugon, IPS and H3C -- have announced a wide range of new T4 servers.

NVIDIA T4 is being used to accelerate AI inference and training in a broad range of fields, including healthcare, finance and retail, which are key elements in the global high performance computing market for enterprise and hyperscale.

This follows NVIDIA's announcement at the recent SC18 supercomputing show that, just two months after its introduction, T4 is featured in 57 separate server designs from the world's leading computer makers. Additionally, Google Cloud announced T4 availability to its Google Cloud Platform customers.

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

“The continued rapid adoption of T4 makes complete sense, given its unprecedented capabilities," said Ian Buck, vice president of Accelerated Computing at NVIDIA. “Never before have we introduced a GPU that gives public and private clouds the combined performance and energy efficiency they need to more economically run their compute-intensive workloads at scale. And in markets where 'scale' really counts, we expect T4 to be extremely popular."

Based on the new NVIDIA Turing™ architecture, the T4 GPU features multi-precision Turing Tensor Cores and new RT Cores, which, when combined with accelerated containerized software stacks, deliver unprecedented performance at scale.

Among China server companies featuring T4 GPUs are Inspur, Huawei, Lenovo, Sugon, Inspur Power System and H3C. Their new systems include:

- Inspur NF5280M4 /NF5280M5/NF5288M5/NF5468M5
- Huawei G2500/2288 HV5/ 5288V5/G530 V5/G560 V5
- Lenovo ThinkSystem SR630/SR650
- Sugon X580-G30/X745-G30/X780-G30/X780-G35/X785-G30 / X740-H30
- Inspur Power System: FP5295G2
- H3C Uniserver G4900G3

Systems are expected to begin shipping before the end of the year.

Flexible Design, Breakthrough Performance
Designed to meet the unique needs of scale-out public and enterprise cloud environments, NVIDIA 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.

The T4 GPU'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 up to 54 CPU-only servers. For AI training, a server with two T4 GPUs can replace nine dual-socket, CPU-only servers.

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: NVIDIA Turing T4 Cloud GPU adoption accelerating; tech giants unveiling new products and services based on NVIDIA T4 GPUs; the companies beginning to use T4 to expand and accelerate their hyperscale datacenters; T4 servers being announced by leading companies, the companies featuring T4 and what their systems include; NVIDIA T4 GPU being the fastest-ever adopted server GPUs; the number of server designs featuring T4s; T4 being used to accelerate AI inference and training in a broad range of fields including in key elements in the global high performance computing market for enterprise and hyperscale; the availability of T4 to Google Cloud Platform customers; the adoption of T4 making complete sense; the benefits, performance, features and abilities of the NVIDIA T4 GPU, including its unprecedented capabilities, performance, energy efficiency and economy; our expectation for the T4 to be extremely popular; T4 accelerating diverse workloads; T4 featuring multi-precision Turing Tensor Cores and new RT Cores, which when combined with accelerated containerized software stacks, deliver unprecedented at scale; the expected shipping of systems before the end of the year; T4 helping customers efficiently address exploding user and data growth; server designs ranging from a single T4 GPU up to 20 GPUs in a single node; T4 GPUs’ capabilities powering breakthrough AI performance for AI workloads at different levels of precision; and T4 GPUs’ ability to replace CPUs 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