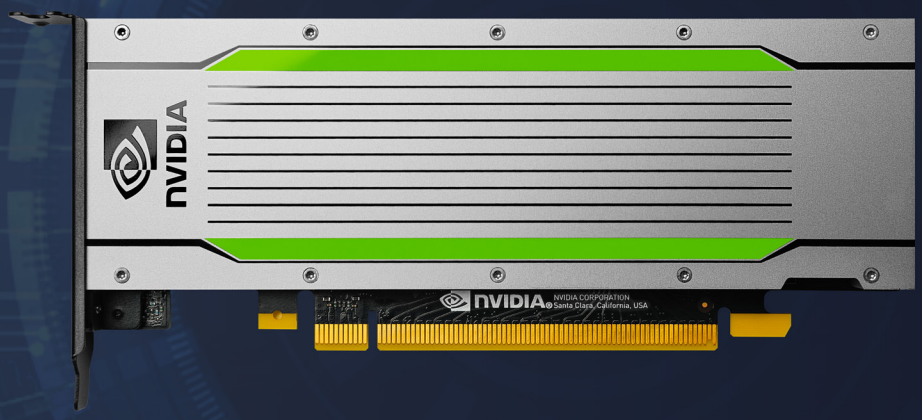
NVIDIA T4G TENSOR CORE GPU

ACCELERATED GRAPHICS AND AI FOR THE ARM-BASED AWS CLOUD



Accelerated Graphics and AI from the AWS Arm-based Cloud

The NVIDIA T4G Tensor Core GPU is the first GPU-accelerated Arm-based instance in the AWS cloud. The Amazon EC2 G5g instances feature Graviton2 processors, based on the 64-bit Arm Neoverse cores, and NVIDIA T4G, bringing together breakthrough graphics performance powered by NVIDIA RTX™ technology, and price performance benefits of AWS Graviton2 processors.

NVIDIA T4G supports next-generation NVIDIA RTX technology for the most advanced visualization experience. When combined with Arm-based Graviton2 processors, developers can create games natively and benefit from the accelerated graphics and encoding capabilities of the NVIDIA T4G, then stream games to mobile devices eliminating the need for emulation software and cross-compilation.

The AWS G5g instance also brings the NVIDIA Arm HPC SDK to cloud computing. With support for NVIDIA T4G and the Arm-based Graviton2 CPU, the NVIDIA Arm HPC SDK provides the tools required to build NVIDIA GPU-accelerated HPC applications in the cloud.

NVIDIA T4G builds on the rich ecosystem of AI frameworks from the NVIDIA NGC™ catalog, CUDA-X™ libraries, over 2.3 million developers, and over 1,800 GPU-optimized applications to help enterprises solve the most critical challenges in their business.

Specifications

FP32	**8.1 TF**
FP16 Tensor Core	**16.3 TF**
INT8 Tensor Core	**32.5 TOPS**
INT4 Tensor Core	**65 TOPS**
RT Cores	**40**
Encode / Decode	**1 encoder 2 decoder**
GPU Memory	**16 GB GDDR6**
GPU Memory Bandwidth	**320 GB/s**
Interconnect	**x16 PCIe Gen3**
Form Factor	**1-slot Low Profile PCIe**
Max TDP Power	**70W**

Highly Efficient Data Center Acceleration

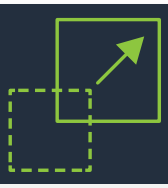
NVIDIA RT Cores

Powerful RT Cores combined with NVIDIA RTX technology, enable real-time ray-traced rendering, delivering photorealistic objects and environments with physically accurate shadows, reflections, and refractions.

NVIDIA Tensor Cores

Turing Tensor Core technology with multi-precision computing for AI powers breakthrough performance from FP32 to FP16 to INT8, as well as INT4 precisions.

16 GB GDDR6



Ultra-fast GDDR6 memory, delivering 320 GB/s of bandwidth for rendering, data science, engineering simulation, and other GPU-memory intensive workloads.

Power-efficient, compact design

Small form-factor, 70-watt design optimized for scale-out environments, providing an incredible 50X higher energy efficiency compared to CPUs. NVIDIA's inference platform has increased efficiency by over 10X and remains the most energy-efficient solution for distributed AI training and inference.

Price performance advantage

NVIDIA T4G availability in Amazon EC2 G5g instances brings world-class graphics and AI performance to the cloud with the price performance benefits of Graviton2 processors.

To learn more about the NVIDIA T4G Tensor Core GPU, visit **https://aws.amazon.com/ec2/instance-types/g5g/**