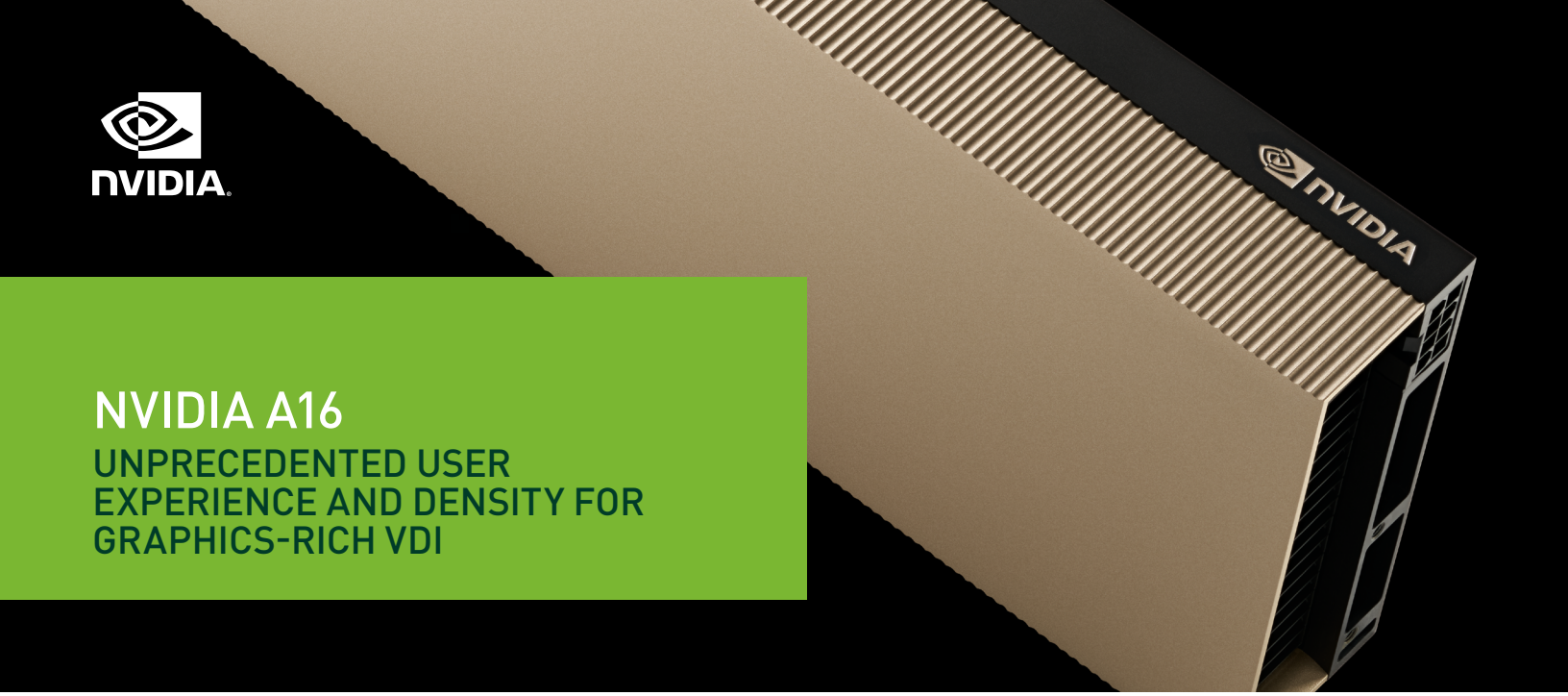

NVIDIA A16

UNPRECEDENTED USER EXPERIENCE AND DENSITY FOR GRAPHICS-RICH VDI

The NVIDIA A16, built on the NVIDIA Ampere architecture, combines with NVIDIA virtual GPU (vGPU) software to raise the bar on user experience for graphics-rich virtual desktop infrastructure (VDI). As more organizations turn to remote work as long term strategy, A16 with the NVIDIA Virtual PC (vPC) or Virtual Applications (vApps) software, enables knowledge workers across every industry to maximize productivity with performance indistinguishable from a native PC. NVIDIA A16 delivers up to 2X the user density, compared with the previous generation M10, reducing the amount of hardware resources needed and lowering the total cost of ownership (TCO).

When combined with NVIDIA RTX Virtual Workstation (vWS) software, the A16 enables affordable entry-level virtual workstations ideal for running workloads such as computer-aided design (CAD). The A16 features a unique quad-GPU board design enabling the provisioning of mixed user profile sizes, so IT can support light virtual PC workloads as well as users with larger memory and graphics requirements. Mixing user types on a board is also supported, enabling the provisioning of virtual PCs, virtual workstations, and even virtualized compute on a single board.

SUPERIOR STREAMING MEDIA PERFORMANCE

The NVIDIA A16 features the highest number of video encoders and decoders with four on-chip hardware encoders (NVENC) and eight decoder (NVDEC) units in a single A16 board. This provides the best encode, decode and transcode performance translating to a maximized number of video streams per A16 board at an attractive price point versus alternative offerings.

SPECIFICATIONS

PNY Part Number	**NVA16TCGPU-KIT**
GPU Architecture	**NVIDIA Ampere architecture**
CUDA Cores	**5120 | 4x 1280**
Tensor Cores	**160 | 4x 40**
RT Cores	**40 | 4x 10**
GPU memory	**4x 16 GB GDDR6**
Memory bandwidth	**4x 200 GB/s**
Error-correcting code (ECC)	**Yes**
NVIDIA Ampere architecture-based CUDA Cores	**4x 1280**
NVIDIA third-generation Tensor Cores	**4x 40**
NVIDIA second-generation RT Cores	**4x 10**
FP32 | TF32 | TF32[1] (TFLOPS)	**4x 4.5 | 4x 9 | 4x 18**
FP16 | FP16[1] (TFLOPS)	**4x 17.9 | 4x 35.9**
INT8 | INT8[1] (TOPS)	**4x 35.9 | 4x 71.8**
System interface	**PCIe Gen4 (x16)**
Max power consumption	**250W**
Thermal solution	**Passive**
Form factor	**Full height, full length (FHFL) Dual Slot**
Power connector	**8-pin CPU**
Encode/decode engines	**4 NVENC/8 NVDEC (includes AV1 decode)**
Secure and measured boot with hardware root of trust for GPU	**Yes**
vGPU software support	**NVIDIA Virtual PC (vPC), NVIDIA Virtual Applications (vApps), NVIDIA RTX Virtual Workstation (vWS), NVIDIA AI Enterprise, NVIDIA Virtual Compute Server (vCS)**
Graphics APIs	**DirectX 12.07[2], Shader Model 5.17[2], OpenGL 4.68[3], Vulkan 1.18[3]**
Compute APIs	**CUDA, DirectCompute, OpenCL™, OpenACC®**
MIG support	**No**

Modernize Your VDI and Boost Streaming Media Performance

DESIGNED FOR ACCELERATED VDI

Optimized for user density, and combined with NVIDIA vPC software, enables graphics-rich virtual PCs to be accessible from anywhere.

AFFORDABLE VIRTUAL WORKSTATIONS

Large framebuffer per user for entry-level virtual workstations, with NVIDIA RTX vWS software, running workloads such as computer-aided design (CAD).

FLEXIBLY SUPPORT DIVERSE USER TYPES

Unique quad-GPU board design enables the provisioning of mixed user profile sizes and user types, such as virtual PCs and virtual workstations, on a single board.

SUPERIOR USER EXPERIENCE

Provides increased frame rate and lower end-user latency, versus CPU-only VDI, resulting in more responsive applications and a user experience that is indistinguishable from a native PC or workstation.

DOUBLE THE USER DENSITY

Purpose-built for graphics-rich VDI, with support for up to 64 concurrent users per board, in a dual-slot form factor.

HIGH- RESOLUTION DISPLAY

Supports multiple, high-resolution monitors to enable maximum productivity and photorealistic quality in a VDI environment.

MORE THAN 2X THE ENCODER THROUGHPUT

More than double the encoder throughput versus previous generation M10, providing high-performance transcoding and the multiuser performance required for multi-stream video and multimedia.

HIGHEST QUALITY VIDEO

Support for the latest codecs, including H.265 encode/decode, VP9, and AV1 decode for the highest-quality video experiences.

NVIDIA AMPERE ARCHITECTURE

NVIDIA Ampere architecture-based CUDA cores, second-generation RT-Cores, and third- generation Tensor-Cores provide the flexibility to host virtual workstations powered by NVIDIA RTX vWS software, or to leverage unused VDI resources to run compute workloads with NVIDIA AI Enterprise software.

PCI EXPRESS GEN 4

Support for PCI Express Gen 4 data transfer speeds from CPU memory for data-intensive tasks.

Features

- Purpose-built for graphics-rich VDI with NVIDIA vPC
- Provides the lowest cost per virtual workstation user with NVIDIA RTX vWS[4]
- Support for all NVIDIA vGPU software editions: NVIDIA vPC, NVIDIA vApps, NVIDIA RTX vWS, NVIDIA AI Enterprise, and NVIDIA vCS
- PCI Express Gen 4
- Latest CODEC support: H.265 encode/decode, VP9, and AV1 decode

To learn more about the NVIDIA A16, visit **www.pny.com/a16**

[1] Structural sparsity enabled
[2] GPU supports DX 12.0 API, Hardware Feature Level 12 + 1
[3] Product is based on a published Khronos specification and is expected to pass the Khronos conformance testing process when available. Current conformance status can be found at www.krhonos.org/conformance
[4] Comparison of NVIDIA A16 versus T4, RTX 6000, RTX 8000 and A40 virtual workstations.