NVIDIA® A40 GPU

PRODUCT SPECIFICATIONS

Architecture	NVIDIA Ampere Architecture
Foundry	Samsung
Process Size	8nm
Transistors	28.3 billion
Die Size	628.4 mm^2
CUDA Parallel Processing Cores	10,752
NVIDIA Tensor Cores (3rd Gen)	336
NVIDIA RT Cores (2nd Gen)	84
Peak FP32 TFLOPS (non-Tensor)	37.4
Peak FP16 Tensor TFLOPS with FP16 Accumulate	149.7 | 299.4*
Peak TF32 Tensor TFLOPS	74.8 | 149.6*
Peak BF16 Tensor TFLOPS with FP32 Accumulate	149.7 | 299.4*
Peak INT8 Tensor TOPS	299.3 | 598.6*
Peak INT4 Tensor TOPS	598.7 | 1197.4*
RT Core Performance	73.1 TFLOPS
GPU Memory	48 GB GDDR6 with ECC
Memory Interface	384-bit
Memory Bandwidth	696 GB/s
NVLink[1]	2-way low profile (2-slot)
Interconnect	NVIDIA NVLink 112.5 GB/s (bidirectional) PCIE Gen4 x16 31.5 GB/s (bidirectional)
Max Power Consumption	300W
Graphics Bus	PCI Express 4.0 x16
Display Connectors	DP 1.4 (3) Supports NVIDIA Mosaic and Quadro® Sync[2]
Display Max Resolution / Quantity	Up to four 5K (5120 x 2880) 60 Hz displays or up to two 8K (7680 × 4320) displays.[3]
Form Factor	4.4” H x 10.5” L Dual Slot
Product Weight	990g
Thermal Solution	Passive
vGPU Software Support[3]	NVIDIA RTX® Virtual Workstation, NVIDIA Virtual PC, NVIDIA Virtual Apps, NVIDIA Virtual Compute Server
vGPU Profiles Supported	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 16 GB, 24 GB, 48 GB
Virtual Display Max Resolution / Quantity	Up to four 5K displays or two 8K displays for 8Q+ profiles
Graphics APIs	DirectX 12.07[5], Shader Model 5.17[5], OpenGL 4.68[6], Vulkan 1.18[6]
Compute APIs	CUDA, DirectCompute, OpenCL™, OpenACC®
NVIDIA® 3D Vision® and 3D Vision Pro	Support via 3 pin mini DIN
Frame lock	Compatible (with Quadro Sync II)

NVIDIA A40 TECH SPECS

Power Connector	1x 8-pin CPU
NVENC | NVDEC	1x ENC | 2x DEC (includes AV1 decode)
NEBS Ready	Level 3
Secure and Measured Boot with Hardware Root of Trust	Supported

Peak performance rates are based on GPU Boost Clock.
**Effective TOPS / TFLOPS using the new Sparsity Feature*
1 Connecting two NVIDIA A40 cards with NVLink to scale performance and memory capacity to 96 GB is only possible if your application supports NVLink technology. Please contact your application provider to confirm their support for NVLink.
2 Quadro Sync II card sold separately. Mosaic supported on Windows 10 and Linux.
3 Dual 8K monitors requires monitors support DisplayPort 1.4 DSC for 8K resolution via single cable.
4 A40 is configured for virtualization by default with physical display connectors disabled. The display outputs can be enabled via management software tools.
5 GPU supports DX 12.0 API, Hardware Feature Level 12 + 1.
6 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