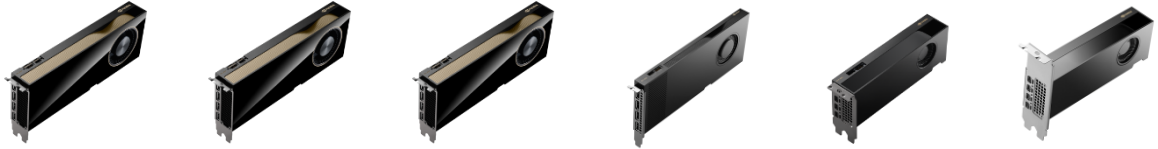


NVIDIA RTX™ GPU Cards Selection Guide



Model Name	NVIDIA RTX 6000 Ada Generation	NVIDIA RTX 5000 Ada Generation	NVIDIA RTX 4500 Ada Generation	NVIDIA RTX 4000 Ada Generation	NVIDIA RTX 4000 SFF Ada Generation	NVIDIA RTX 2000 Ada Generation
GPU Architecture	Ada Lovelace	Ada Lovelace	Ada Lovelace	Ada Lovelace	Ada Lovelace	Ada Lovelace
Memory Capacity	48GB GDDR6 ECC	32GB GDDR6 ECC	24GB GDDR6 ECC	20GB GDDR6 ECC	20GB GDDR6 ECC	16GB GDDR6 ECC
Memory Interface	384-bit	256-bit	192-bit	160-bit	160-bit	128-bit
Memory Bandwidth	960 GB/s	576 GB/s	432 GB/s	360 GB/s	280 GB/s	224 GB/s
CUDA Cores	18,176	12,800	7,680	6,144	6,144	2,816
Tensor Cores	568	400	240	192	192	88
RT Cores	142	100	60	48	48	22
Single-Precision Performance	91.1 TFLOPS [†]	65.3 TFLOPS [†]	39.6 TFLOPS [†]	26.7 TFLOPS [†]	19.2 TFLOPS [†]	12.0 TFLOPS [†]
RT Core Performance	210.6 TFLOPS [†]	151.0 TFLOPS [†]	91.6 TFLOPS [†]	61.8 TFLOPS [†]	44.3 TFLOPS [†]	27.7 TFLOPS [†]
Tensor Performance	1457.0 TFLOPS [‡]	1044.4 TFLOPS [‡]	634.0 TFLOPS [‡]	327.6 TFLOPS [‡]	306.8 TFLOPS [‡]	191.9 TFLOPS [‡]
Graphics Interface	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x8
Max Power Consumption	300W	250W	210W	130W	70W	70W
Thermal Solution	Active	Active	Active	Active	Active	Active
Ambient Operating Temperature	0°C to 45°C	0°C to 45°C	0°C to 45°C	0°C to 45°C	0°C to 45°C	0°C to 50°C
Storage Temperature	-45°C to 75°C	-45°C to 75°C	-45°C to 75°C	-45°C to 75°C	-45°C to 75°C	-45°C to 75°C
Form Factor	4.4" (H) x 10.5" (L) dual slot, full height	4.4" (H) x 10.5" (L) dual slot, full height	4.4" (H) x 10.5" (L) dual slot, full height	4.4" (H) x 9.5" (L) single slot, full height	2.7" (H) x 6.6" (L) dual slot, low profile	2.7" (H) x 6.6" (L) dual slot, low profile
Display Connectors	4x DP 1.4a	4x DP 1.4a	4x DP 1.4a	4x DP 1.4a	4x mDP 1.4a	4x mDP 1.4a
Max Simultaneous Displays	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz
Power Connector	CEM5 16-pin	CEM5 16-pin	CEM5 16-pin	CEM5 16-pin	-	-
Encode/Decode Engines	3x encode, 3x decode (+AV1 encode and decode)	2x encode, 2x decode (+AV1 encode and decode)	2x encode, 2x decode (+AV1 encode and decode)	2x encode, 2x decode (+AV1 encode and decode)	2x encode, 2x decode (+AV1 encode and decode)	1x encode, 1x decode (+AV1 encode and decode)
Graphics APIs	DirectX 12 Shader Model 6.6 OpenGL 4.6 Vulkan 1.3	DirectX 12 Ultimate Shader Model 6.7 OpenGL 4.6 Vulkan 1.3	DirectX 12.07 Shader Model 6.7 OpenGL 4.6 Vulkan 1.3	DirectX 12 Ultimate Shader Model 6.7 OpenGL 4.6 Vulkan 1.3	DirectX 12 Ultimate Shader Model 6.6 OpenGL 4.6 Vulkan 1.3	DirectX 12 Ultimate Shader Model 6.6 OpenGL 4.6 Vulkan 1.3
Compute APIs	CUDA, DirectCompute, OpenCL™	CUDA, DirectCompute, OpenCL™	CUDA, DirectCompute, OpenCL™	CUDA, DirectCompute, OpenCL™	CUDA, DirectCompute, OpenCL™	CUDA, DirectCompute, OpenCL™
NVIDIA NVLink®	No	No	No	No	No	No
Virtualization Ready	Yes	Yes	Yes	Yes	Yes	Yes

Accelerate the Future of Computing
with NVIDIA RTX

NVIDIA RTX Ada Generation
performance for endless possibilities

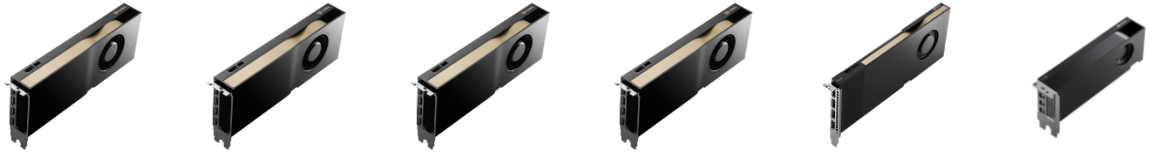


[†]Peak rates based on GPU Boost Clock.

[‡]Effective FP8 teraFLOPS (TFLOPS) using the new sparsity feature.

*All information provided is subject to change without notice. Last updated: March 22, 2024

NVIDIA RTX™ GPU Cards Selection Guide



Model Name	NVIDIA RTX A6000	NVIDIA RTX A5500	NVIDIA RTX A5000	NVIDIA RTX A4500	NVIDIA RTX A4000	NVIDIA RTX A2000 12GB
GPU Architecture	Ampere	Ampere	Ampere	Ampere	Ampere	Ampere
Memory Capacity	48GB GDDR6 ECC	24GB GDDR6 ECC	24GB GDDR6 ECC	20GB GDDR6 ECC	16GB GDDR6 ECC	12GB GDDR6 ECC
Memory Interface	384-bit	384-bit	384-bit	320-bit	256-bit	192-bit
Memory Bandwidth	768 GB/s	768 GB/s	768 GB/s	640 GB/s	448 GB/s	288 GB/s
CUDA Cores	10,752	10,240	8,192	7,168	6,144	3,328
Tensor Cores	336	320	256	224	192	104
RT Cores	84	80	64	56	48	26
Single-Precision Performance	38.7 TFLOPS [†]	34.1 TFLOPS [†]	27.8 TFLOPS [†]	23.7 TFLOPS [†]	19.2 TFLOPS [†]	8.0 TFLOPS [†]
RT Core Performance	75.6 TFLOPS [†]	66.6 TFLOPS [†]	54.2 TFLOPS [†]	46.2 TFLOPS [†]	37.4 TFLOPS [†]	15.6 TFLOPS [†]
Tensor Performance	309.7 TFLOPS [‡]	272.8 TFLOPS [‡]	222.2 TFLOPS [‡]	189.2 TFLOPS [‡]	153.4 TFLOPS [‡]	63.9 TFLOPS [‡]
Graphics Interface	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16
Max Power Consumption	300W	230W	230W	200W	140W	70W
Thermal Solution	Active	Active	Active	Active	Active	Active
Ambient Operating Temperature	0°C to 45°C	0°C to 45°C	0°C to 45°C	0°C to 45°C	0°C to 45°C	0°C to 50°C
Storage Temperature	-45°C to 75°C	-45°C to 75°C	-45°C to 75°C	-45°C to 75°C	-45°C to 75°C	-45°C to 75°C
Form Factor	4.4" (H) x 10.5" (L) dual slot, full height	4.4" (H) x 10.5" (L) dual slot, full height	4.4" (H) x 10.5" (L) dual slot, full height	4.4" (H) x 10.5" (L) dual slot, full height	4.4" (H) x 9.5" (L) single slot, full height	2.7" (H) x 6.6" (L) dual slot, low profile
Display Connectors	4x DP 1.4a	4x DP 1.4a	4x DP 1.4a	4x DP 1.4	4x DP 1.4a	4x mDP 1.4a
Max Simultaneous Displays	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz	4x 4096 x 2160 @ 120 Hz 4x 5120 x 2880 @ 60 Hz 2x 7680 x 4320 @ 60 Hz
Power Connector	CPU 8-pin	PCIe 8-pin	PCIe 8-pin	PCIe 8-pin	PCIe 6-pin	-
Encode/Decode Engines	1x encode, 2x decode (+AV1 decode)	1x encode, 2x decode (+AV1 decode)	1x encode, 2x decode (+AV1 decode)	1x encode, 1x decode (+AV1 decode)	1x encode, 1x decode (+AV1 decode)	1x encode, 1x decode (+AV1 decode)
Graphics APIs	DirectX 12.07 Shader Model 5.17 OpenGL 4.68 Vulkan 1.18	Direct3D 12 Shader Model 6.6 OpenGL 4.6 Vulkan 1.3	DirectX 12 Ultimate Shader Model 6.6 OpenGL 4.6 Vulkan 1.3	DirectX 12 Ultimate Shader Model 6.6 OpenGL 4.6 Vulkan 1.3	DirectX 12 Ultimate Shader Model 6.6 OpenGL 4.6 Vulkan 1.3	DirectX 12 Ultimate Shader Model 6.6 OpenGL 4.6 Vulkan 1.3
Compute APIs	CUDA, DirectCompute, OpenCL™	CUDA, DirectCompute, OpenCL™	CUDA, DirectCompute, OpenCL™	CUDA, DirectCompute, OpenCL™	CUDA, DirectCompute, OpenCL™	CUDA, DirectCompute, OpenCL™
NVIDIA NVLink®	Yes	Yes	Yes	Yes	No	No
Virtualization Ready	Yes	Yes	Yes	Yes	Yes	Yes



[†]Peak rates based on GPU Boost Clock.

[‡]Effective FP8 teraFLOPS (TFLOPS) using the new sparsity feature.

*All information provided is subject to change without notice. Last updated: March 22, 2024