

Selecting the Right NVIDIA GPU for Virtualization

The GPU that best meets the requirements of your workloads depends on the importance to you of factors such as raw performance, time-to-solution, performance per dollar, performance per watt, form factor, and any power and cooling constraints.

NVIDIA GPUs for Virtualization

Table 4 summarizes the features of the **NVIDIA GPUs for virtualization** workloads based on the NVIDIA Ampere and Ada GPU architectures.

GPUs for graphics workloads based on the NVIDIA Lovelace, and Ampere GPU architectures feature second and third-generation RT Cores. RT Cores are accelerator units that are dedicated to performing ray tracing operations with extraordinary efficiency.

The GPUs in **Table 4** are tested and supported with NVIDIA software for virtualizing GPUs, specifically with NVIDIA virtual GPU software. For the full product support matrices for the NVIDIA software for virtualizing GPUs, refer to the following documentation:

> **Virtual GPU Software Supported Products**

Table 4 NVIDIA GPUs Recommended for Virtualization

Expand

Specification	L40S	L40	L4	A40	A10	A16	A2
GPUs/Board	1	1	1	1	1	4	1
Architecture	Lovelace	Lovelace	Lovelace	Ampere	Ampere	Ampere	Ampere
RTX Technology	✓	✓	✓	✓	✓	✓	✓
Memory Size and	48GB GDDR6	48GB	24GB	48GB	24GB	64GB	16GB



CHAT (beta)

Type		GDDR6	GDDR6	GDDR6	GDDR6	(16GB per GPU) GDDR6	GDDR6
vGPU Profile Sizes (GB)	1, 2, 3, 4, 6, 8, 12, 16, 24, 48	1, 2, 3, 4, 6, 8, 12, 16, 24, 48	1, 2, 3, 4, 6, 8, 12, 24	1, 2, 3, 4, 6, 8, 12, 16, 24, 48	1, 2, 3, 4, 6, 8, 12, 24	1, 2, 4, 8, 16	1, 2, 4, 8, 16
MIG Support	No	No	No	No	No	No	No
NVLink Support	No	No	No	Yes	No	No	No
Form Factor	<ul style="list-style-type: none"> › PCIe 4.0 › Dual Slot FHFL 	<ul style="list-style-type: none"> › PCIe 4.0 › Dual Slot FHFL 	<ul style="list-style-type: none"> › PCIe 4.0 › Single Slot HHHL 	<ul style="list-style-type: none"> › PCIe 4.0 › Dual Slot FHFL 	<ul style="list-style-type: none"> › PCIe 4.0 › Single Slot FHFL 	<ul style="list-style-type: none"> › PCIe 4.0 › Dual Slot FHFL 	<ul style="list-style-type: none"> › PCIe 4.0 › Single Slot HHHL
Power (W)	350	300	72	300	150	250	60
Cooling	Passive	Passive	Passive	Passive	Passive	Passive	Passive
Optimized For ⁴	Performance	Performance	Performance	Performance	Performance	Density	Density
Target Workloads	Deep learning and machine learning training and inference, video	High-end virtual workstations	VDI, mid-level to high-end	High-end virtual workstations	Entry-level to mid-level virtual	Knowledge worker virtual	AI inference, VDI, and virtual

NVIDIA L40S

The NVIDIA® L40S, based on the NVIDIA Ada Lovelace GPU architecture, offers top-tier performance for both **visual computing and AI workloads in data center and edge server deployments**. Featuring 142 third-generation RT Cores and 568 fourth-generation Tensor Cores, it supports hardware ray tracing, revolutionary AI features, advanced shading, and powerful simulation capabilities. The L40S is ideal for deep learning inference and training,



generative AI workloads, batch and real-time rendering, virtual workstations, and cloud gaming. With 48GB of graphics memory, the L40S provides exceptional performance for diverse graphics and compute tasks in modern data centers. When combined with NVIDIA RTX™ Virtual Workstation (vWS) software, it enables powerful virtual workstations with unmatched performance and security, accessible from any device.

NVIDIA L40

The NVIDIA® L40, based on the NVIDIA Ada Lovelace GPU architecture, delivers unprecedented visual computing performance for the data center and provides **revolutionary neural graphics, compute, and AI capabilities to accelerate the most demanding visual computing workloads**. The L40 features 142 third-generation RT Cores that enhance real-time ray tracing capabilities and 568 fourth-generation Tensor Cores with support for the FP8 data format. These new features are combined with the latest generation CUDA Cores and 48GB of graphics memory to accelerate visual computing workloads from high-performance virtual workstation instances to large-scale digital twins in NVIDIA Omniverse. With up to twice the performance of the previous generation at the same power, the NVIDIA L40 is uniquely suited to provide the visual computing power and performance required by the modern data center. When combined with NVIDIA RTX™ Virtual Workstation (vWS) software, the NVIDIA L40 delivers powerful virtual workstations from the data center or cloud to any device. Millions of creative and technical professionals can access the most demanding applications from anywhere with awe-inspiring performance that rivals physical workstations—all while meeting the need for greater security.

NVIDIA L4

The NVIDIA Ada Lovelace L4 Tensor Core GPU delivers universal acceleration and energy efficiency for **video, AI, virtual workstations, and graphics applications** in the enterprise, in the cloud, and at the edge. And with NVIDIA's AI platform and full-stack approach, L4 is optimized for video and inference at scale for a broad range of AI applications to deliver the best in personalized experiences. As the most efficient NVIDIA accelerator for mainstream use, servers equipped with L4 power up to 120X higher AI video performance over CPU solutions and 2.5X more generative AI performance, as well as over 4X more graphics performance than the previous GPU generation. L4's versatility and energy-efficient, single-slot, low-profile form factor makes it ideal for edge, cloud, and enterprise deployments.

NVIDIA A40

Built on the RTX platform, the NVIDIA A40 GPU is uniquely positioned to power high-end virtual workstations running professional visualization applications, accelerating **the most demanding graphics workloads**. The second-generation RT Cores of the NVIDIA A40 enable it to deliver massive speedups for workloads such as photorealistic rendering of movie content, architectural design evaluations, and virtual prototyping of product designs. The NVIDIA A40 features 48 GB of frame buffer, but with the NVIDIA® NVLink® GPU interconnect, it can support up to 96 GB of frame buffer to power virtual workstations that support very large animations, files, or models. Although the NVIDIA A40 has 48 GB of frame buffer, the context switching limit per GPU limits the maximum number of users supported to 32. Refer to [Table 5](#) to see how many VDI users can be supported by each GPU when each user has a vGPU profile with 1 GB of frame buffer.

The NVIDIA A40 is also suitable for running VDI workloads and compute workloads on the same infrastructure. Resource utilization can be increased by using common virtualized GPU accelerated server resources to run virtual desktops and workstations while users are logged on, and compute workloads while users have logged off. Learn more from the NVIDIA whitepaper about [Using NVIDIA Virtual GPUs to Power Mixed Workloads](#).



NVIDIA A16

The NVIDIA A16 is designed to provide the most cost-effective graphics performance for **knowledge worker VDI workloads**. For these workloads, where users are accessing office productivity applications, web browsers, and streaming video, the most important consideration is achieving the best performance per dollar and the highest user density per server. With four GPUs on each board, the NVIDIA A16 is ideal for providing the best performance per dollar and a high number of users per GPU for these workloads.

NVIDIA A10

The NVIDIA A10 is designed to provide cost-effective graphics performance for accelerating and optimizing the performance of **mixed workloads**. When combined with NVIDIA RTX vWS software, it accelerates graphics and video processing with AI on mainstream enterprise servers. Its second-generation RT Cores make the NVIDIA A10 ideal for mainstream professional visualization applications running on high-performance mid-range virtual workstations.

GPU Performance Benchmark Tests

The GPU performance benchmark tests measure GPU performance for virtualized workloads that use NVIDIA GPU virtualization software. To measure the performance of a GPU running a specific virtualized workload, a representative benchmark test for the workload is run on the GPU.

In many cases, cost rather than raw performance is the principal factor in selecting the right virtual GPU solution for a specific workload. For this reason, the GPU performance benchmark tests measure both raw performance **and** performance per dollar.

Unless otherwise stated, the tests are run with vGPU profiles that are allocated all the physical GPU's frame buffer. This vGPU profile size was chosen because the impact of scaling does not vary between different GPUs [5](#).

Table 5 summarizes the results of the benchmark tests to determine which GPUs provide the best raw performance and the best performance per dollar for specific graphic workloads.

Note

When choosing GPUs based on raw performance or performance per dollar, use these results **for general guidance only**. All results are based on the workloads listed in [Table 5](#), which could differ from the applications being used in production.

Table 5 GPU Performance Benchmark Tests and Results

		Best Raw	Most
--	--	----------	------

 **CHAT** (beta)

Workload	Benchmark	Best Raw Performance GPU	Cost-Effective GPU
Knowledge worker VDI	NVIDIA nVector Digital Worker Workload	NVIDIA L4	NVIDIA A16
Professional graphics	SPECviewperf 2020 (3840x2160)	NVIDIA L40S	NVIDIA L4

Knowledge Worker VDI

GPU performance for knowledge worker VDI workloads was measured by using the [NVIDIA nVector Digital Worker Workload](#) benchmark test. NVIDIA nVector Digital Worker Workload is a benchmarking tool that simulates end users' workflows and measures key aspects of the user experience, including end-user latency, framerate, image quality, and resource utilization.

Test Results

The GPUs that provide the best raw performance and cost effectiveness for knowledge worker VDI workloads are listed in [Table 5](#). For knowledge worker VDI workloads, the principal factor in determining cost effectiveness is the combination of performance per dollar and user density.

As more knowledge worker users are added to a server, the server consumes more CPU resources. Adding an NVIDIA GPU for this workload conserves CPU resources by offloading graphics rendering tasks to the GPU. As a result, user experience and performance are improved for end users.

Table 6 Maximum Number of Supported NVIDIA vPC Knowledge Workers (with 1 GB Profile Size)

GPU	Maximum Users per GPU Board	Maximum Boards per Server	Maximum Users per Server
A16	64	3	192
L4	24	6	144
A10	24	6	144

L40S	32	3	96
A40	32	3	96
T4	16	6	96

Table 6 assumes that each user requires a vGPU profile with 1GB of frame buffer. However, to determine the profile sizes that provide the best user experience for the users in your environment, you must conduct a proof of concept (POC).

Please refer to the appropriate sizing guide to build your NVIDIA vGPU environment:

› [NVIDIA vPC Sizing Guide](#)

