# GPU selection guide

## 32 product options for workstation and data center systems

Choose the exact GPU edition before sizing the host. This guide compares 31 PCIe add-in cards and one H200 SXM platform component, then identifies the integration and procurement checks needed for a quote.

| Product group | Options | Use this section for | Page |
|---|---|---|---|
| **Blackwell RTX PRO** | 11 | Server, high-memory workstation and compact desktop choices | 2-3 |
| **Data center accelerators** | 4 | H200 NVL and SXM, plus L40S and L4 | 4 |
| **Ada workstation GPUs** | 8 | Professional graphics and existing Ada deployments | 5 |
| **Ampere workstation GPUs** | 9 | Lifecycle matching and platform-compatible replacements | 6 |
| **Specification and ordering notes** | - | Manufacturer references, exact SKU review and purchase terms | 7 |

### Start with the deployment constraints

- **Workload and memory.** Record the application, model or scene size, precision, context and concurrency. Capacity is stated per GPU; aggregate memory is not automatically a shared pool.
- **Host and cooling.** Match the physical format, electrical link, power connector and airflow to a qualified platform. Passive cards require engineered server airflow.
- **Interconnect and software.** Decide how work spans GPUs. PCIe, NVLink bridges and HGX NVSwitch fabrics are different designs; software and licensing also require review.
- **Procurement baseline.** Specify the manufacturer part number, quantity, condition, required bracket and accepted substitutions before requesting availability.

Product order descends by model tier within each family. It is not a cross-architecture performance ranking. Specifications are reference values; the final manufacturer SKU and qualified system govern the order.

Compare GPUs at nextecus.com/gpus

Find a system at nextecus.com/find-a-system

BLACKWELL 1 OF 2

# High memory RTX PRO options

These six Blackwell options separate server cooling, desktop power envelopes and memory capacities. Equal memory capacity does not imply equal compute performance.

| GPU and edition | Memory per GPU | Memory bandwidth | Board power | Host and cooling |
|---|---|---|---|---|
| **RTX PRO 6000 Blackwell** Server Edition | 96 GB GDDR7 ECC | 1,597 GB/s | 400-600 W | PCIe 5.0 x16 Passive air |
| **RTX PRO 6000 Blackwell** Workstation Edition | 96 GB GDDR7 ECC | 1,792 GB/s | 600 W | PCIe 5.0 x16 Double flow-through |
| **RTX PRO 6000 Blackwell** Max-Q workstation Edition | 96 GB GDDR7 ECC | 1,792 GB/s | 300 W | PCIe 5.0 x16 Active air |
| **RTX PRO 5500 Blackwell** 84 GB workstation Edition | 84 GB GDDR7 ECC | 1,398 GB/s | Up to 600 W | PCIe 5.0 x16 Active air or liquid RXM |
| **RTX PRO 5000 Blackwell** 72 GB workstation | 72 GB GDDR7 ECC | 1,344 GB/s | 300 W | PCIe 5.0 x16 Active |
| **RTX PRO 5000 Blackwell** 48 GB workstation | 48 GB GDDR7 ECC | 1,344 GB/s | 300 W | PCIe 5.0 x16 Active |

Physical fit: passive air Server Edition, Max-Q, RTX PRO 5500 air and RTX PRO 5000 are full-height dual-slot. RTX PRO 6000 Workstation has a taller 5.4-inch dual-slot cooler. The liquid Server Edition is single-slot FHXL; RTX PRO 5500 liquid RXM needs its own qualified integration.

## Select the edition deliberately

- **RTX PRO 6000 Server Edition.** The NexTec eight-GPU reference uses passive air-cooled cards. NVIDIA also offers a liquid-cooled implementation, which needs a different qualified platform.
- **Workstation and Max-Q.** Both are desktop add-in cards. The 600 W Workstation Edition needs greater chassis and cooling clearance; the 300 W Max-Q edition serves a different power envelope.
- **RTX PRO 5500.** 84 GB, 1,398 GB/s and up to 600 W are preliminary manufacturer specifications and may change. Air-cooled and liquid RXM implementations are distinct integration choices.
- **RTX PRO 5000.** 72 GB and 48 GB are separate capacity variants. Put the required capacity and exact SKU on the quote request.

RTX PRO 5500 was listed by NVIDIA as coming soon with preliminary specifications at this review. Availability and the final SKU must be confirmed before ordering. All listed memory capacities include ECC support.

BLACKWELL 2 OF 2

# Mainstream and compact RTX PRO

The Server, Workstation and SFF labels describe materially different power, cooling and physical configurations. Match the card to the host rather than choosing by model number alone.

| GPU and edition | Memory per GPU | Memory bandwidth | Board power | Host and cooling |
|---|---|---|---|---|
| **RTX PRO 4500 Blackwell** Server Edition | 32 GB GDDR7 ECC | 800 GB/s | 165 W | PCIe 5.0 x16 Passive |
| **RTX PRO 4500 Blackwell** Workstation Edition | 32 GB GDDR7 ECC | 896 GB/s | 200 W | PCIe 5.0 x16 Active |
| **RTX PRO 4000 Blackwell** Workstation | 24 GB GDDR7 ECC | 672 GB/s | 145 W | PCIe 5.0 x16 Active |
| **RTX PRO 4000 Blackwell** SFF Edition | 24 GB GDDR7 ECC | 432 GB/s | 70 W | PCIe 5.0 x8 Active |
| **RTX PRO 2000 Blackwell** Compact workstation | 16 GB GDDR7 ECC | 288 GB/s | 70 W | PCIe 5.0 x8 Active |

## Physical fit and electrical lanes

| Model | Reference physical format | Procurement check |
|---|---|---|
| **RTX PRO 4500 Server** | Full-height single-slot | Passive chassis airflow; 165 W server SKU |
| **RTX PRO 4500 Workstation** | Full-height dual-slot | Active workstation cooling; 200 W SKU |
| **RTX PRO 4000** | Full-height single-slot | Different from the 70 W SFF model |
| **RTX PRO 4000 SFF** | Low-profile dual-slot | PCIe 5.0 x8 electrical link; bracket and two-slot clearance |
| **RTX PRO 2000** | Low-profile dual-slot | PCIe 5.0 x8 electrical link; desktop rather than laptop model |

## Define the display and software requirements

Include display, adapter, application and licensing requirements for workstations. For servers, confirm supported firmware, drivers and virtualization entitlement. An x8 specification is the electrical lane count, not a guarantee of fit in a closed-ended x8 slot.

Memory bandwidth is a local GPU-memory figure, not PCIe transfer bandwidth. Board power is not the complete-system electrical load.

DATA CENTER ACCELERATORS

# Choose the platform and fabric

H200 NVL and H200 SXM share HBM3e capacity and local memory bandwidth but are not interchangeable. L40S and L4 are separate Ada-based server accelerator choices.

| GPU and edition | Memory per GPU | Memory bandwidth | Board power | Host and cooling |
|---|---|---|---|---|
| **H200**<br>NVL PCIe | 141 GB<br>HBM3e | 4.8 TB/s | Up to 600 W | PCIe 5.0 x16<br>Passive air |
| **H200**<br>SXM / HGX platform | 141 GB<br>HBM3e | 4.8 TB/s | Up to 700 W | SXM on HGX<br>Platform cooling |
| **L40S**<br>Data-center accelerator | 48 GB<br>GDDR6 ECC | 864 GB/s | 350 W | PCIe 4.0 x16<br>Passive |
| **L4**<br>Low-profile server accelerator | 24 GB<br>GDDR6 ECC | 300 GB/s | 72 W | PCIe 4.0 x16<br>Passive |

Physical fit: H200 NVL and L40S are full-height dual-slot cards. L4 is low-profile single-slot. H200 SXM is an HGX platform component, not a PCIe card.

## H200 integration paths

| Choice | System architecture | Key distinction |
|---|---|---|
| **H200 NVL** | PCIe server with supported 2- or 4-GPU NVLink bridge groups | Up to 600 W per GPU. An eight-card installation does not imply an eight-way NVSwitch fabric. |
| **H200 SXM** | Compatible HGX platform with eight GPUs in the reference system | Up to 700 W per GPU. SXM is a baseboard component, not a PCIe add-in card. |

## L40S and L4 deployment checks

L40S targets mixed AI, graphics and rendering workloads in a qualified passive-cooled server. It does not support NVLink or MIG. L4 provides a 72 W low-profile, single-slot option for inference and video workloads; passive cooling still requires a supported airflow path. L4 ECC is enabled by default and can be disabled in software.

Eight H200 GPUs provide 1,128 GB aggregate device memory. That sum is not automatically one shared memory pool. Select the internal GPU fabric and external network separately.

ADA WORKSTATION GPUS

# Professional Ada options

Use these reference specifications to compare capacities and form factors. Preserve exact generation names when expanding existing fleets or replacing individual cards.

| GPU and edition | Memory per GPU | Memory bandwidth | Board power | Host and cooling |
|---|---|---|---|---|
| **RTX 6000 Ada Generation**<br>Workstation | 48 GB<br>GDDR6 ECC | 960 GB/s | 300 W | PCIe 4.0 x16<br>Active |
| **RTX 5880 Ada Generation**<br>48 GB workstation | 48 GB<br>GDDR6 ECC | 960 GB/s | 285 W | PCIe 4.0 x16<br>Active |
| **RTX 5000 Ada Generation**<br>32 GB workstation | 32 GB<br>GDDR6 ECC | 576 GB/s | 250 W | PCIe 4.0 x16<br>Active |
| **RTX 4500 Ada Generation**<br>24 GB workstation | 24 GB<br>GDDR6 ECC | 432 GB/s | 210 W | PCIe 4.0 x16<br>Active |
| **RTX 4000 Ada Generation**<br>20 GB workstation | 20 GB<br>GDDR6 ECC | 360 GB/s | 130 W | PCIe 4.0 x16<br>Active |
| **RTX 4000 SFF Ada Generation**<br>20 GB SFF workstation | 20 GB<br>GDDR6 ECC | 280 GB/s | 70 W | PCIe 4.0 x16<br>Active |
| **RTX 2000 Ada Generation**<br>16 GB workstation | 16 GB<br>GDDR6 ECC | 224 GB/s | 70 W | PCIe 4.0 x8<br>Active |
| **RTX 2000E Ada Generation**<br>16 GB single-slot | 16 GB<br>GDDR6 ECC | 224 GB/s | 50 W | PCIe 4.0 x8<br>Active fan |

## Physical fit and exact variants

Physical fit: RTX 6000, 5880, 5000 and 4500 Ada are full-height dual-slot. RTX 4000 Ada is full-height single-slot. RTX 4000 SFF and RTX 2000 Ada are low-profile dual-slot. RTX 2000E Ada is low-profile single-slot.

RTX 2000 Ada is a 70 W low-profile dual-slot card. RTX 2000E Ada is a 50 W low-profile single-slot card with active fan cooling. The standard PNY 2000E part number is VCNRTX2000ADAS-LLP; the 40 W Power Limited product is a different variant.

RTX 4000 SFF Ada bandwidth is the NVIDIA reference value of 280 GB/s; vendor-specific listings may differ, so verify the supplied SKU. RTX 5880 regional SKU and sourcing eligibility also require review. RTX 6000 Ada is distinct from Blackwell RTX PRO 6000 and Ampere RTX A6000.

AMPERE WORKSTATION GPUS

# Lifecycle and fleet matching

These Ampere choices support exact-model sourcing discussions. Availability, condition, compatibility, testing and manufacturer warranty eligibility are confirmed in the quotation.

| GPU and edition | Memory per GPU | Memory bandwidth | Board power | Host and cooling |
|---|---|---|---|---|
| **RTX A6000**<br>Ampere workstation | 48 GB<br>GDDR6 ECC | 768 GB/s | 300 W | PCIe 4.0 x16<br>Active |
| **RTX A5500**<br>24 GB workstation | 24 GB<br>GDDR6 ECC | 768 GB/s | 230 W | PCIe 4.0 x16<br>Active |
| **RTX A5000**<br>24 GB workstation | 24 GB<br>GDDR6 ECC | 768 GB/s | 230 W | PCIe 4.0 x16<br>Active |
| **RTX A4500**<br>20 GB workstation | 20 GB<br>GDDR6 ECC | 640 GB/s | 200 W | PCIe 4.0 x16<br>Active |
| **RTX A4000**<br>16 GB workstation | 16 GB<br>GDDR6 ECC | 448 GB/s | 140 W | PCIe 4.0 x16<br>Active |
| **RTX A2000**<br>12 GB workstation | 12 GB<br>GDDR6 ECC | 288 GB/s | 70 W | PCIe 4.0 x16<br>Active |
| **RTX A2000**<br>6 GB workstation | 6 GB<br>GDDR6 ECC | 288 GB/s | 70 W | PCIe 4.0 x16<br>Active |
| **RTX A1000**<br>8 GB workstation | 8 GB<br>GDDR6 | 192 GB/s | 50 W | PCIe 4.0 x8<br>Active |
| **RTX A400**<br>4 GB workstation | 4 GB<br>GDDR6 | 96 GB/s | 50 W | PCIe 4.0 x8<br>Active |

## Match the installed hardware baseline

Physical fit: A6000, A5500, A5000 and A4500 are full-height dual-slot. A4000 is full-height single-slot. Both A2000 capacities are low-profile dual-slot. A1000 and A400 are low-profile single-slot.

Verify the exact replacement SKU, firmware, bracket and host qualification; laptop models differ. RTX A6000 supports a two-GPU NVLink bridge in qualified layouts. Confirm bridge spacing and application support. A2000 12 GB and 6 GB are separate variants.

SPECIFICATION AND ORDERING NOTES

# Prepare a precise quote request

Product names in the comparison tables link to manufacturer references. Use the exact manufacturer part number and final platform qualification to resolve variant-specific differences before ordering.

## Information to include

- **Required hardware.** GPU name, edition, memory capacity, manufacturer SKU, quantity and acceptable condition.
- **Installed environment.** Host model, slot layout, bracket, power supply, connectors, cooling and facility limits.
- **Application and acceptance.** Software stack, representative workload, licensing, test scope and acceptance criteria.
- **Delivery and lifecycle.** Deployment location, required date, approved substitutions and component warranty requirements.

## How to read the specifications

Capacity and bandwidth are per GPU. Board power is not complete-system power. Check physical fit and electrical lanes separately; an x8 link does not guarantee fit in a closed-ended x8 slot. Passive cards need chassis airflow, and GPU count does not define fabric topology.

## Commercial conditions

**Quotation and availability.** Reference specifications and options are not a stock list. Exact parts, condition, compatibility, price, lead time and service scope are confirmed in the written quotation. An inquiry does not place an order or reserve inventory. Sales and delivery are within the United States only.

**Purchase approval.** Buyer vetting and NexTec approval are required before purchase. Manufacturer or distributor approval is required where applicable, including PNY review for applicable orders. Buyer, end-user, intended-use and installation-location details may be requested.

**Manufacturer warranty.** The applicable component manufacturer provides the warranty. Duration, eligibility and claim terms vary by part. Customers contact the manufacturer for claims; NexTec can assist with order and warranty information.

## References and contact

NVIDIA professional desktop GPU portfolio

NVIDIA data center GPU portfolio

Send an RFQ or parts list  nextecus.com/request-quote

General questions  support@nextecus.com

Reference edition: 29 October 2026. Manufacturer trademarks identify third-party products and do not imply endorsement, certification or authorized partner status.