RTX PRO 6000 Server Bill of Materials

ASUS ESC8000A-E13-DR | Eight GPU configuration | Per physical server

Cyfuture AI offers this server configuration for enterprise AI inference, model fine-tuning, graphics and video workloads. This specification details the hardware capacity of one complete server and the configuration selections to be recorded in the customer order.

Capacity at a glance

Resource	Installed configuration	Capacity per server
GPU	8 x 96 GB NVIDIA RTX PRO 6000	768 GB aggregate GPU memory
CPU	2 x AMD EPYC 9554	128 cores / 256 threads
System memory	16 x 32 GB Micron RDIMM	512 GB DDR5 ECC RAM
Local storage	2 x 7.68 TB NVMe SSD	15.36 TB raw capacity

Primary hardware BOM

Component	Qty	Specification
Server platform	1	ASUS ESC8000A-E13-DR, dual-socket rack server with eight GPU positions. Published ESC8000A-E13 enclosure: 4U; final factory SKU governs. [1, 2]
GPU accelerator	8	NVIDIA RTX PRO 6000 Blackwell Server Edition; 96 GB GDDR7 with ECC per GPU; PCIe 5.0 x16; passive server thermal design. [3, 4]
Processor	2	AMD EPYC 9554; 64 cores and 128 threads per CPU; 3.1 GHz base clock, up to 3.75 GHz boost; Socket SP5. [5]
Memory module	16	Micron 32 GB DDR5-4800 ECC RDIMM; 4800 MT/s data rate; 512 GB total. Exact Micron part number to be recorded in the order.
NVMe SSD	2	7.68 TB capacity per drive; 15.36 TB combined raw capacity. SSD manufacturer, model and endurance specification are order selections.
Network adapter	Per order	NVIDIA / Mellanox ConnectX-7 MCX75310AAS. Full suffix, hardware revision and adapter quantity determine the final network specification. [6, 7]

Capacity basis: GPU memory is the sum of eight separate 96 GB devices. Raw SSD capacity is before RAID, filesystem and operating-system allocation. System RAM is separate from GPU memory.

GPU subsystem

Each server contains eight NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs. The following specifications apply to the Server Edition installed in this configuration. [3, 4]

Specification	Per GPU	Eight GPU server
Architecture	NVIDIA Blackwell	8 Blackwell GPUs
GPU memory	96 GB GDDR7 with ECC	768 GB aggregate
CUDA processing cores	24,064	192,512
Tensor Cores	752, fifth generation	6,016
Ray tracing cores	188, fourth generation	1,504
Memory interface	512-bit	Independent on each GPU
Local memory bandwidth	1,597 GB/s	Per-device bandwidth; separate memories
FP32 theoretical peak	120 TFLOPS	960 TFLOPS arithmetic sum
Host interface	PCIe 5.0 x16	8 GPU PCIe interfaces
Maximum GPU board power	Up to 600 W, configurable	Up to 4,800 W for GPUs
Multi Instance GPU	Up to 4 instances at 24 GB	Up to 32 instances in supported configurations
Media acceleration engines	4 NVENC / 4 NVDEC / 4 JPEG	32 of each engine type

Performance figures are manufacturer theoretical peaks. Adding GPU peaks does not establish application throughput or linear scaling. GPU clocks, power limits, precision, software and communication overhead affect measured performance.

Using the available GPU memory

Each GPU has its own 96 GB memory. A workload can use multiple GPUs through an application framework that partitions the model or processing. The 768 GB aggregate is not a single automatically unified memory pool. GPU communication follows the server PCIe and CPU topology; no NVLink or NVSwitch fabric is included in this BOM.

Workload capabilities

The GPU architecture supports AI inference and model adaptation, including low-precision Tensor Core operations such as FP4. It also supports ray tracing, rendering and accelerated video processing. Application sizing should consider model precision, context length, batch size and concurrent users. [3, 4]

MIG is an optional provisioning mode. Instance profiles and availability depend on the supported driver and software stack. The default hardware quantity remains eight physical GPUs; commercial software entitlements are specified separately in the order.

Processors memory and storage

Processor configuration

Specification	Per processor	Per server
Processor model	AMD EPYC 9554	2 processors
Physical cores / threads	64 cores / 128 threads	128 cores / 256 threads
Clock speeds	3.1 GHz base; up to 3.75 GHz boost	Same specification on each CPU
L3 cache	256 MB	512 MB across two CPUs
Socket and generation	SP5; EPYC 9004 series	Dual-socket configuration
Default thermal design power	360 W	720 W CPU TDP total
Memory controller support	12 DDR5 channels; up to 4800 MT/s	24 channels available in total

Processor reference: AMD EPYC 9554. Boost frequencies depend on operating conditions. Thread counts assume simultaneous multithreading is enabled; they do not specify a contracted virtual CPU allocation. [5]

System memory configuration

Parameter	Configured specification
Manufacturer and module type	Micron 32 GB DDR5 ECC registered DIMM
Installed quantity and capacity	16 modules x 32 GB = 512 GB
Data rate	4800 MT/s, commonly marketed as DDR5-4800
Planned socket balance	8 DIMMs / 256 GB per CPU, using the ASUS slot population map
Platform occupancy	16 of 24 DIMM slots populated; 8 slots remain available

Memory expansion option: 24 x 32 GB provides 768 GB and populates all 12 channels per CPU. This is an optional configuration change; the supplied BOM remains 16 x 32 GB. Exact module part numbers and population must satisfy the ASUS qualified memory requirements. [2, 5, 8]

Local NVMe storage

Storage layout	Nominal capacity	Provisioning implication
Two independent drives	2 x 7.68 TB	Separate local volumes; no mirroring
Software mirror if selected	7.68 TB before overhead	Maintains a copy on both drives
Installed raw capacity	15.36 TB	Physical capacity in either layout

Specify the SSD make, model, interface generation, form factor, endurance and power-loss protection in the order. Boot allocation, partitioning and data protection are provisioning selections. A hardware RAID controller, separate boot drives and backup storage are not listed in the primary BOM.

Platform networking and power

ASUS platform reference

The published ESC8000A-E13 specifications below describe the OEM platform. The exact supplied ESC8000A-E13-DR factory configuration determines installed power modules, accessories and physical details. [1, 2]

Platform feature	Manufacturer reference specification
Rack enclosure	4U; 800 x 439.5 x 175 mm, depth x width x height
GPU expansion	8 PCIe 5.0 x16 positions for dual-slot GPUs
Additional rear PCIe slots	One x16 link and one x8 link; both physical x16 slots
Local drive accommodation	8 front 2.5-inch bays; native configuration supports 2 NVMe drives
Remote management	Dedicated RJ45 management port; ASMB12-iKVM / AST2600 BMC
Power supply reference	4 x 3200 W, 80 PLUS Titanium, arranged as 3+1 redundancy
AC input reference	220-240 V AC, 50/60 Hz; OEM rating 16 A per PSU input
Cooling reference	5 GPU fan modules and 5 system fan modules, hot-swappable

Rack planning: use the final OEM factory specification for space, rail depth, weight, electrical feeds and airflow. The published 4U enclosure dimension must be reconciled with any order description stating 6U.

Network adapter selection

Full ordering part number	Maximum rated link	Connector and host interface
MCX75310AAS-NEAT	400 Gb/s NDR InfiniBand or 400GbE	1 OSFP port; PCIe 5.0 x16
MCX75310AAS-HEAT	200 Gb/s NDR200 InfiniBand or 200GbE	1 OSFP port; PCIe 5.0 x16

NVIDIA specifies InfiniBand and Ethernet support from hardware revision AA for -NEAT and A7 for -HEAT. Confirm the actual card revision, suffix and quantity. The final order selects the applicable adapter and deployed protocol. [6, 7]

Install a 400 Gb/s adapter in the full x16 electrical slot. Link speed and usable throughput depend on the selected protocol, switch, approved optics or cable, firmware and configuration. Network capacity is shared by workloads using that adapter. Switches, optical modules and network cabling require their own order quantities.

Power and thermal planning

At a 600 W setting for each GPU and the 360 W default TDP for each CPU, the GPU and CPU component budget is 5.52 kW. Memory, drives, NICs, fans, motherboard and conversion losses add to this figure. This calculation is not a measured AC input rating. [4, 5]

The final PSU configuration, GPU power limit and combined-load power measurement establish deployment requirements. PSU redundancy does not by itself establish resilience to loss of an entire facility power feed.

Integration and customer handover

Supporting integration items

The following items complete the physical installation. Their inclusion, OEM part numbers and kit contents should be stated in the final server order.

Integration item	Quantity basis	Specification to record
CPU cooling assemblies	2	ASUS assemblies qualified for the EPYC 9554 thermal configuration
GPU installation hardware	1 server set	8 GPU positions, retention hardware and approved GPU power harnesses
Power and cooling assemblies	1 platform set	Exact PSU wattage, count, redundancy arrangement and fan configuration
Storage installation kit	2 drive positions	Compatible carriers, backplane connectivity and internal NVMe cables
Rack installation and AC cords	1 server set	Compatible rail kit; power cords matched to PSU inputs and site outlets
External network link items	Per link design	NIC-side and switch-side interfaces, approved optics or cable, lengths

Order specific selections

Order field	Detail to be agreed
Final hardware identity	ASUS factory SKU and rack height; GPU manufacturer part number; Micron and NVMe part numbers; network adapter suffix, revision and quantity
Software and storage layout	Operating system, boot and data layout, driver / CUDA stack, optional MIG profiles, and any NVIDIA AI Enterprise or vGPU licences
Delivery and support	Hardware sale or hosted service, node quantity, delivery date, warranty / support term and applicable service commitments

For hosted GPU services, this BOM describes the physical server. The customer order defines allocated GPUs, CPU resources, RAM, storage, network access and bandwidth. Customer access to the BMC is subject to the agreed service model.

Handover record

Record component serial numbers, installed capacities, firmware and driver versions, storage layout and negotiated network speed. Validate all eight GPUs under concurrent load and capture memory, storage, thermal and network health results. Any application performance commitment should reference an agreed workload and benchmark method.

Manufacturer references

[1] ASUS ESC8000A-E13 product specification
[2] ASUS ESC8000A-E13 H200 NVL and RTX PRO 6000 datasheet
[3] NVIDIA RTX PRO 6000 Blackwell Server Edition specifications
[4] NVIDIA RTX PRO 6000 Blackwell Server Edition datasheet
[5] AMD EPYC 9554 processor specifications
[6] NVIDIA ConnectX-7 ordering part numbers and revision notes
[7] NVIDIA ConnectX-7 adapter specifications
[8] AMD EPYC 9004 memory population recommendations