

FLAGSHIP / 8-GPU BLACKWELL

8× RTX PRO 6000. The complete AI server.

Eight NVIDIA RTX PRO 6000 Blackwell Server Edition accelerators, dual Xeon host compute, ECC memory, enterprise NVMe, and high-speed networking in a configurable 5U platform.



Supermicro SYS-521GE-TNRT manufacturer reference image. The pictured GPU population differs from NexTec's proposed eight-GPU RTX PRO configuration.

GPU configuration	8× RTX PRO 6000 Blackwell Server Edition; passive air; 96 GB GDDR7 ECC each.
Memory & interface	768 GB aggregate across separate GPUs; PCIe 5.0 x16. No HGX NVLink/NVSwitch fabric.
Configuration status	Proposed OEM-based build. Exact BOM, supported power cap, availability, test plan and commercial terms are confirmed in writing.

HOST PLATFORM / BILL OF MATERIALS

Every component, specified.

Reference system	Supermicro SYS-521GE-TNRT · CSE-528GTS-R5K40P chassis · X13DEG-OA motherboard.
Specified GPUs	8× NVIDIA RTX PRO 6000 Blackwell Server Edition, passive air-cooled; 96 GB GDDR7 ECC each.
Host processors	Two LGA-4677 sockets. 4th / 5th Gen Intel Xeon Scalable; up to 64 cores and 350 W per CPU with air cooling.
CPU starting options	2× Xeon Gold 6530: 64 total cores, 270 W TDP each; or 2× Xeon Platinum 8592+: 128 total cores, 350 W TDP each. Final BOM subject to OEM qualification.
ECC system memory	32 DDR5 DIMM slots. OEM maximum: 4 TB at 1 DIMM/channel, up to 5600 MT/s; 8 TB at 2 DIMMs/channel, up to 4400 MT/s. CPU and DIMM selection set actual speed.
Balanced memory population	Proposed 512 GB / 1 TB / 2 TB use 16× 32 / 64 / 128 GB RDIMMs, one per channel across both sockets. Gold 6530 limits 1DPC memory to 4800 MT/s.
Boot devices	2× M.2 NVMe slots. Proposal options include paired enterprise boot drives; mirroring method and OS support are itemized.
Front storage	Default: 8× 2.5-inch NVMe and 8× 2.5-inch SATA hot-swap bays. An optional additional 8-bay NVMe configuration requires the appropriate controller and cabling.
Expansion / topology	13 PCIe 5.0 x16 FHFL slots; switch-based dual-root architecture. Slot link width is distinct from aggregate switch-uplink bandwidth.
Network / management	Onboard 2× 10GbE RJ45 via Intel X710-AT2; dedicated IPMI management. AIOM/OCP 3.0 and PCIe NIC options for the application fabric.

POWER / TOPOLOGY / ACCEPTANCE

Plan the deployment.

High-speed fabric	Quote 100/200/400GbE or InfiniBand with qualified NICs, optics, cables, and switch ports. Define GPU/NIC locality, storage traffic, and GPUDirect requirements.
Cooling	Up to 10 heavy-duty fans with speed control and an air shroud. Passive GPUs rely on chassis airflow. Rack inlet temperature and GPU power settings must follow OEM limits.
Power hardware	4× 2700 W Titanium PSUs; 200-240 VAC input. A 2+2 arrangement provides 5.4 kW nominal surviving output after loss of two modules; the complete DC load must fit that envelope.
GPU power budget	400-600 W configurable per GPU; 3.2-4.8 kW for eight GPUs alone. The suggested 450 W planning profile leaves more headroom; exact supported settings and failover require validation.
Rack envelope	5U · 438 mm wide × 222.5 mm high × 786.1 mm deep. Confirm installed weight, rails, cable clearance, PDU connectors, and front-to-rear airflow.
Integration deliverables	Exact BOM and serial inventory; firmware/driver versions; GPU, ECC, storage and network results; sustained thermal/power logs; workload acceptance results; and the agreed support path.

Workload and acceptance

Scope inference, fine-tuning, rendering or research against a representative application. Validate GPU peer access, CPU/GPU/NIC locality, switch uplinks, storage throughput, temperatures and complete system power. At 600 W per GPU, do not assume 2+2 full-load PSU redundancy.

Technical references

[Supermicro platform specifications ↗](#)[NVIDIA Server Edition specifications ↗](#)[NVIDIA reference architecture ↗](#)[Intel Xeon Gold 6530 ↗](#)[Intel Xeon Platinum 8592+ ↗](#)[Configure this system ↗](#)