



RENT **AI** SERVER

Услугата е базирана на NVIDIA 6000 или NVIDIA 4500.

Предоставя възможност до 24 часа за стартиране на разработка и внедряване на изкуствен интелект.

Услугата дава възможност да започнете без начална инвестиция от десетки или стотици хиляди долара и да надграждате плавно.

РАЗГЛЕДАЙ ПЛАНОВЕТЕ



NVIDIA RTX 6000 PRO BLACKWELL

Enterprise GPU Power for Real Results



Hardware Overview

GPU model	NVIDIA RTX 6000 PRO Blackwell
Architecture	Blackwell
VRAM	96 GB GDDR7 ECC
CUDA cores	24,064
Power (TDP)	600 W (configurable down to 300 W)
Form factor	2-slot blower, fits in 4U server chassis
Driver	NVIDIA RTX Enterprise • Studio (certified)
Host CPU	AMD EPYC / Threadripper PRO
Host RAM	Up to 256 GB ECC DDR5
Storage	Up to 4 TB NVMe
Network	1 Gbps up to 40 Gbps

Available Configurations

Configuration	vCPU	RAM	NVMe	Price / month	Price / day
Basic	16	64 GB	1 TB	€930	€39
Standard	32	128 GB	2 TB	€1,320	€55
Performance	64	256 GB	4 TB	€2,280	€95



HAEMI GPU SERVER

NVIDIA RTX PRO 4500 BLACKWELL

Professional AI Performance at an Excellent Value



Hardware Overview

GPU model	NVIDIA RTX PRO 4500 Blackwell
Architecture	Blackwell
VRAM	24 GB GDDR7 ECC
CUDA cores	10,752
Power (TDP)	200 W (configurable down to 120 W)
Form factor	Dual-slot, professional workstation GPU
Driver	NVIDIA RTX Enterprise • Studio (certified)
Host CPU	AMD EPYC / Threadripper PRO
Host RAM	Up to 256 GB ECC DDR5
Storage	Up to 2 TB NVMe
Network	1 Gbps up to 40 Gbps

Available Configurations

Configuration	vCPU	RAM	NVMe	Price / month	Price / day
Basic	16	64 GB	0.5 TB	€490	€20
Standard	32	128 GB	1 TB	€830	€35
Performance	64	256 GB	2 TB	€1,510	€63



HAEMI GPU SERVER

OPERATING SYSTEMS

OPERATING SYSTEMS



Ubuntu 26.04 LTS

Kernel 6.8 · the default for every template



Ubuntu 24.04 LTS

Kernel 5.15 HWE · for stacks pinned to CUDA 11.8 / 12.1



Debian 12

Bookworm · minimal, driver from the vendor repository



Rocky Linux 9

RHEL-compatible · Slurm, enterprise and HPC stacks



Ubuntu



Ubuntu



Debian

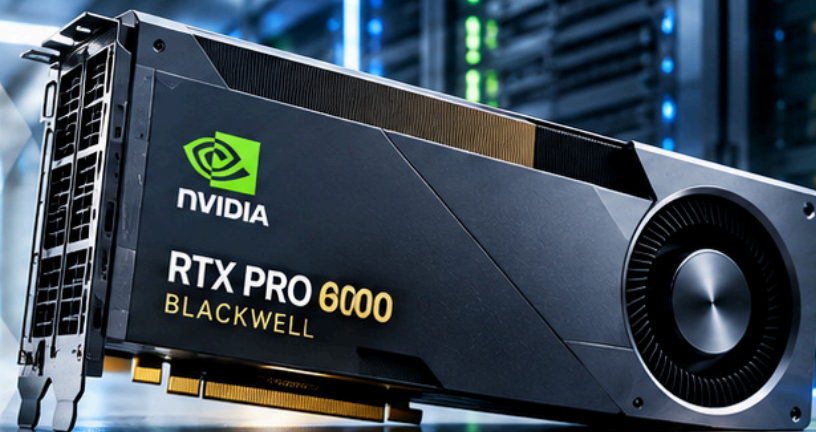


Rocky Linux

REAL TEST RESULTS

Popular AI models on 1× RTX PRO 6000 BLACKWELL

Real test results from our infrastructure



Model	Type	Single request (tokens/s)	16 parallel requests (total tokens/s)
 Gemma 4 26B-A4B	MoE	204	1,619
 Qwen3.8-27B + MTP	Dense	100.7	1,190
 Qwen3.8-27B	Dense	50.3	693
 Qwen3.8-27B (NVFP4)	Dense	58.4	793
 Qwen3.8-27B (NVFP4 + MTP)	Dense	105.8	1,292
 Gemma 4 31B	Dense	43.2	661

Test environment

GPU	1× NVIDIA RTX PRO 6000 Blackwell (96 GB)
OS	Ubuntu 24.04.5
NVIDIA driver	615.71.09
Software	vLLM 0.29.0, Docker 29.8.1



Testing methodology

Parallel requests	2, 4, 8, 16 (3 runs each, 256 output tokens)
Context window	Up to 262,144 tokens
KV cache	FP8 (90% of VRAM)
Disk	2.7 GB/s sequential read, 59 GB model load in 30 s
Network (idle)	Sofia: 13.5 ms, Nuremberg: 36 ms

USE CASES

GPU Use Cases



DATA ANALYTICS

Data Analytics

GPU-accelerated ETL, RAPIDS workflows, and big data processing faster than CPU-only clusters.



AI INFERENCE

Next-gen AI inference

Deploy large language models (LLMs) with low latency and high throughput. Power chatbots, content generation, and knowledge retrieval.



MEDIA & STREAMING

Media & streaming pipelines

Accelerate real-time encoding, transcoding, and streaming with GPU-optimized tools like FFmpeg and Apache Kafka.



RESEARCH

Scientific Research

Climate simulations, genomics, material science, and advanced computational research.



COMPUTER VISION

Computer Vision

Real-time object detection, image processing, and visual AI applications powered by optimized inference.



HPC & SIMULATION

HPC & Simulation

High-performance computing for financial modeling, weather forecasting, molecular dynamics, and engineering simulations.



1. Нещо по въпроса за сигурността на данните



2. Защо е NIS 2 compliant



Свържи се с нас



Поръчай

КОНТАКТИ



 www.persy.com

 070042030

 info@persy.com



 www.varnadatacenter.com

 +359 888 42 05 88

 office@varnadatacenter.com