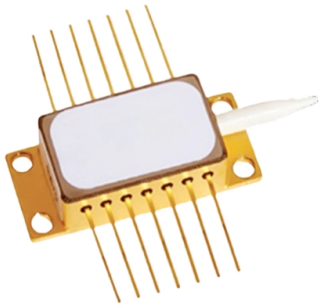
AI server 3090



Overview

May 2026 picks: 2x RTX 3090 (48GB) for the dense-model workhorse; 2x RTX 5060 Ti 16GB (32GB) for budget MoE with --cpu-moe; 2x RTX 2080 Ti 22GB modded for value (Qwen 3. 6 27B at 38 tok/s); 1x 3090 + 1x 4090 for mixed-card pipeline parallelism. cpp and Ollama handle. Want to build a GPU home server for running quantized models?

Here's some tips and tricks for setting up the server. RTX 3090: Two RTX 3090s with NVLink are a common choice for running large AI models. Previously I have built one but only for mining where those GPUs were connected via PCIE x1 risers. 0 x16 so the thing looks slightly different. Building a full DIY rig is a high base cost with inflation with every new recent dual slot capable motherboard checking in above $100. AI from The Basement: My latest side project, a dedicated LLM server powered by 8x RTX 3090 Graphic Cards, boasting a total of 192GB of VRAM. This blogpost was originally posted on my LinkedIn profile in July 2024. Backstory: Sometime in. A 70B model that can't fit on one 24GB card runs at 16-21 tok/s across dual RTX 3090s. You need server-grade platforms.

Article Content

Jan 09, 2026

Building my first AI server with 6 x RTX3090

I am going to build my first AI Server based on 6 x RTX3090 GPUs. Previously I have built one but only for mining where those GPUs were connected

Nov 17, 2025

RTX 3090 · RTX 3080 · RTX 3070 · NVIDIA GPU servers

RTX 3090 3080 3070 dedicated servers, ideal for deep learning, graphics rendering, video transcoding, or crypto mining. 5 minute delivery, 24/7 support, global

Apr 15, 2026

The best Proxmox AI server build for Ollama in 2026

Build a quiet Proxmox AI server for Ollama with clean GPU passthrough, an RTX 3090, 128GB RAM, and a parts list that still makes sense in 2026.

Feb 01, 2026

Quad RTX 3090 AI Rigs: Performance, Motherboards

What multi-GPU AI setups are and how to build quad RTX 3090 rigs with server-grade motherboards, balancing performance and power for large AI models.

Feb 19, 2026

Building a GPU Home Server for AI

RTX 3090: Two RTX 3090s with NVLink are a common choice for running large AI models. NVLink can provide improved communication between

Aug 12, 2025

Best AI, Deep Learning NVIDIA GPU Server in

BIZON ZX9000 – Dual AMD EPYC, 384-core 8 GPU 10 GPU water-cooled NVIDIA RTX H100, H200, A100, A6000, RTX 5090, 4090,

Jun 29, 2025

$1000 Local Ai Home Server on Z440 with 3090 – Digital

Deepseek R1 671b Home Server at $500 Price 24GB VRAM Ai Server Build at Mid-Range $750 Price $1000 Local Ai Server Base Config This server is configured

Jun 02, 2026

Setting Up External Server GPUs for AI/ML/DL

□□ Join me on a deep dive into setting up external GPUs for high-performance AI computations! In this comprehensive guide, I''ll walk you through connecting p...

Oct 18, 2025

Rent 1x3090 at LeaderGPU

Get ultimate computing power with LeaderGPU''s RTX 3090 server. Flexible payment options. Streamline high-performance computing for speed and efficiency.

Aug 02, 2025

$1000 Local Ai Home Server on Z440 with 3090 – Digital

If running one of these fast was a goal, then a 3090 would make decent sense. If however you are happy with tens range models you may not get the same value

May 30, 2026

Cloud servers hosting with RTX 3090

Host a GPU server with RTX 3090: dual Intel Xeon Gold CPUs, up to 3072 GB RAM, SSD storage. Features 328 3rd-gen Tensor Cores and 10,496 CUDA cores for

Feb 08, 2026

Serving AI From The Basement — Part I : A Dedicated

AI from The Basement: My latest side project, a dedicated LLM server powered by 8x RTX 3090 Graphic Cards, boasting a total of 192GB of VRAM. I

Nov 02, 2025

RTX 3090 (24 GB) GPUs für KI-Training und Inferenz

Wir warten unsere Server sorgfältig, wie beispielsweise anhand unserer RTX 3090 24GB-Builds demonstriert wird. Dies ermöglicht Ihnen, sich auf Ihre KI-Projekte – Entwicklung oder Produktion –

Jun 07, 2026

Best Dual-GPU Local AI Setup: RTX 3090, 5060 Ti (2026)

Dual RTX 3090, 2x RTX 5060 Ti, 2x 2080 Ti modded, mixed setups: real configs for Qwen 3.6, MoE, 70B. Tensor vs pipeline parallelism, llama.cpp/vLLM.

Feb 09, 2026

Building a GPU Home Server for AI

Building a GPU Home Server for AI Want to build a GPU home server for running quantized models? Here''s some tips and tricks for setting up the

Dec 01, 2025

Building an AI Homelab with 2X NVIDIA 3090 GPUs

Join me as I walk you through building a custom AI inference rig powered by dual NVIDIA 3090 GPUs. This build is designed for high-performance AI training and inference tasks, featuring an Intel ...

Jul 26, 2025

Dual NVidia RTX 3090 GPU server I have built : LocalLLaMA

Software: I have installed Ozeki AI Server on it for running the AI models. Ozeki AI Server is the best local AI execution framework. It is much faster then other Python based solutions. I had to

Jan 20, 2026

GPU Dedicated Server

The powerful GPU, the RTX PRO 6000 Blackwell, is available for SeiMaxim customers for AI, Metaverse, and HPC workloads. Mission-critical Supermicro,

Nov 10, 2025

Build Your Own AI Server: 2X NVIDIA 3090 for free

□□ Dive into the world of high-performance AI with our detailed server setup tutorial using dual NVIDIA 3090 GPUs! In this video, I''ll guide you through each step of

Sep 09, 2025

Building a High-Performance GPU Server for Large

One of the most critical components of the server is the dual NVIDIA RTX 3090 GPUs connected via an NVLINK bridge. This setup allows for

Aug 25, 2025

Optimizing Qwen3.6-27B Local Inference on RTX 3090 with Native

A deep dive into running the state-of-the-art Qwen3.6-27B model on consumer hardware, achieving 72 tokens per second using native Windows vLLM and implementing hybrid cloud-local

Dec 20, 2025

Rent HPC Servers, Deep Learning Servers, GPU

Rent multi GPU servers and HPC cloud services for deep learning, machine learning & AI. Configurable RTX 4090, RTX A5000/A6000, RTX 6000 Ada GPUs and

Jul 04, 2025

Dual GPU Local AI Server: 3090 vs 4090 | FormulaMod

The 3090 costs roughly twice as much as a 4060 Ti, but it generates images at half the time, so the cost per image is comparable—and you get a card that can handle every model at full

Contact Us

For more information, pricing, or custom solutions, please contact us:

Website: https://www.moletenare-ew.co.za
Email: info@moletenare-ew.co.za
Phone: +86 138 1658 3346
Address: Ningbo, China

This document is for informational purposes only. Specifications subject to change without notice.