Used Server GPUs for Local AI: Tesla P40, V100, A100, and the eBay Goldmine

February 25, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** The Tesla P40 (24GB, ~$175) is the best budget VRAM play for local AI — $7.30/GB vs $35/GB on an RTX 3090. It runs 30B+ models that gaming GPUs under $200 can't touch. The catch: no display output, no fans, no native FP16, and you'll need a cooling mod. Worth it for tinkerers who want maximum VRAM on minimum budget. Skip it if you want plug-and-play.

📚 **More on this topic:** GPU Buying Guide · Best Used GPUs for Local AI · VRAM Requirements · What Can You Run on 16GB VRAM · Budget AI PC Under $500

Everyone talks about gaming GPUs for local AI. RTX 3060, RTX 3090, maybe an RX 7900 XTX if you're feeling adventurous. But there's a whole parallel market that most hobbyists overlook: used datacenter GPUs on eBay.

Datacenters refresh their hardware every 3-5 years. When they cycle out a rack of Tesla P40s or V100s, those cards hit the secondary market at prices that make the VRAM-per-dollar math look absurd. A Tesla P40 with 24GB of VRAM sells for $150-200 on eBay right now. That's the same VRAM as an RTX 3090 for less than a quarter of the price.

The trade-offs are real: no display output, no fans, questionable driver support. But if you're the type who doesn't mind zip-tying a fan to a heatsink, server GPUs are the most underpriced hardware in the local AI space right now.

Why Server GPUs Are Worth a Look

It all comes down to one number: cost per gigabyte of VRAM.

GPU	VRAM	Used Price	$/GB VRAM
Tesla P40	24GB	~$175	**$7.30**
RTX 3060 12GB	12GB	~$190	$15.80
RTX 3090	24GB	~$800	$33.30
RTX 4090	24GB	~$1,800	$75.00

The P40 costs $7.30 per gigabyte of VRAM. An RTX 3090 costs $33. An RTX 4090 costs $75. If your bottleneck is VRAM — and for local LLMs, it almost always is — server GPUs are hard to beat.

These cards were designed for exactly this workload. NVIDIA built them for inference and training in datacenter racks. They're not gaming cards repurposed for AI. They're AI cards that happen to be available on eBay for a fraction of what datacenters paid.

The Server GPU Lineup: Ranked for Local AI

Here's every server GPU worth considering, with current eBay pricing as of February 2026:

GPU	VRAM	Type	Used Price	$/GB	CUDA Cores	TDP	Architecture
Tesla M40	24GB	GDDR5	~$80-120	$4.20	3,072	250W	Maxwell
Tesla P40	24GB	GDDR5X	~$150-200	$7.30	3,840	250W	Pascal
Tesla P100	16GB	HBM2	~$200-350	$17.20	3,584	250W	Pascal
Tesla V100	32GB	HBM2	~$300-500	$12.50	5,120	300W	Volta
A100 40GB	40GB	HBM2e	~$3,500-5,000	$106	6,912	300W	Ampere
A100 80GB	80GB	HBM2e	~$5,500-7,500	$81	6,912	300W	Ampere

Quick Takes

Tesla M40 (~$80-120): The cheapest 24GB card you can buy. Maxwell architecture is ancient — no INT8 support, slow FP32, and driver support is sketchy. Only worth it if you're stacking multiple cards for raw VRAM capacity and don't care about speed.

Tesla P40 (~$150-200): The sweet spot. 24GB GDDR5X, decent CUDA core count, and the strongest community support of any server GPU for local AI. This is the one most tinkerers buy. More on this below.

Tesla P100 (~$200-350): Interesting but awkward. Only 16GB of VRAM, which isn't much of a step up from a $190 RTX 3060 that also has 12GB plus display output and FP16 support. The HBM2 memory is fast, but the VRAM ceiling limits what models you can load. Hard to recommend unless you find one under $150.

Tesla V100 (~$300-500): This is where things get interesting. 32GB HBM2, native FP16 (Tensor Cores), and Volta architecture that still holds up. It runs Qwen 32B Q4 at about 17 tok/s and can handle 28K context where an RTX 3090 tops out around 10K on the same model. The problem is price variance. Some listings hit $500+, and at that point you're in RTX 3080 Ti territory with way less hassle.

A100 40GB/80GB: Serious hardware for serious budgets. The A100 80GB running at $5,500-7,500 is the only single card that can load 70B models at Q4 without offloading. If you need that, nothing else competes. But most hobbyists aren't spending $5,000+ on a headless GPU card.

The Quirks and Gotchas

This is the section that determines whether server GPUs are for you. Every one of these issues is solvable, but none of them are optional.

No Display Output

Server GPUs have no video ports. Zero. No HDMI, no DisplayPort, nothing. You need a separate GPU or an integrated GPU (iGPU) on your CPU for your monitor.

In practice, this means pairing the server GPU with either:

- A cheap GT 710 or GT 1030 ($30-50) for display
- A CPU with integrated graphics (Intel i5/i7 with iGPU, AMD with G-suffix)
- Your existing gaming GPU for display + the server card for inference

This is actually a fine setup for dedicated inference. Run your display off the iGPU, offload all AI work to the P40. llama.cpp and Ollama handle multi-GPU setups well.

Passive cooling: the mandatory mod

Server GPUs ship with bare heatsinks and no fans. They expect server chassis with high-velocity front-to-back airflow. Stick one in a standard desktop case and it will throttle within minutes under load. I've seen reports of 90°C+ before the card shuts itself down.

You have three options:

1. 3D-printed fan shroud ($5-15 + fan): The community favorite. Download a shroud from Printables, print it, and attach a 92mm or 120mm fan. Keeps temps under 75°C at full load. Don't print with PLA. Use PETG or ABS, since PLA softens around 60°C.

2. Zip-tied fans ($10-15): The quick and dirty approach. Zip-tie two 92mm fans directly to the heatsink. Ugly but functional. Nylon zip ties handle the heat fine (they melt at 220°C, and nothing in a PC gets close).

3. Pre-made cooling kits ($20-40): Amazon sells bolt-on blower fan kits designed for Tesla P40/V100/M40 cards. Search "Tesla P40 cooling kit." These use 97x33mm blower fans and bolt directly onto the card.

No matter which option you pick, do the fan mod before you run any sustained workload. These cards will thermal throttle and eventually shut down without active cooling.

The FP16 Problem (P40 and M40)

This is the biggest performance gotcha for the P40 specifically.

The Tesla P40 (compute capability 6.1) does not have native FP16 (half-precision) support. FP16 operations run at 1/64th the speed of full FP32. The P100 and V100 have proper FP16 support; the P40 and M40 do not.

What this means in practice: llama.cpp and most inference engines default to FP16 for GPU computation. On the P40, you need to force FP32 mode or use integer quantization (Q4, Q8). With the right flags, performance is reasonable — but you'll never match an RTX 3060 in raw tokens-per-second at the same model size.

The workaround for llama.cpp: compile with `-DLLAMA_CUDA_FORCE_MMQ=ON` to use integer matrix multiplication instead of FP16 GEMM. This makes the P40 competitive again.

Power Connectors

Most server GPUs use standard 8-pin or 6+2-pin PCIe power connectors — same as gaming GPUs. The P40 takes a single 8-pin. The V100 PCIe takes one 8-pin. No weird adapters needed for these.

The SXM variants (V100 SXM2, A100 SXM4) are a different story. Those use proprietary connectors and need a baseboard. Avoid SXM cards unless you have the matching server chassis. Stick to PCIe versions.

Physical Size

These are full-length, full-height cards. The P40 is 267mm long, which is shorter than most RTX 3090s. But measure your case anyway. The P40 is dual-slot.

Driver and CUDA Support

The Tesla P40 has compute capability 6.1 (Pascal) and supports up to CUDA 12.x with current drivers. Earlier concerns about driver support dropping Pascal have been mostly resolved. NVIDIA still ships datacenter drivers that cover these cards.

The M40 (Maxwell, compute capability 5.2) is the one to watch. Newer CUDA toolkits and frameworks are starting to drop Maxwell support. If you buy an M40 today, expect software compatibility to become a problem within a year or two.

The Tesla P40 Deep Dive

The P40 is the server GPU that the local AI community has adopted as its own. It's cheap, it has headroom, and there's a whole ecosystem of fan mods and config guides built around it.

Why the P40 Works

24GB of VRAM for $175 changes what models you can load. Compare it to what a $190 RTX 3060 12GB can run:

Model	RTX 3060 12GB	Tesla P40 24GB
Llama 3.1 8B Q4	Fits, ~45 tok/s	Fits, ~41 tok/s
Qwen 2.5 14B Q4	Fits (tight), ~20 tok/s	Fits easily, ~16 tok/s
Qwen 30B Q4	Won't fit	Fits, ~15 tok/s
Llama 70B Q3	Won't fit	Partial offload possible
DeepSeek R1 32B Q4	Won't fit	Fits with room to spare

The RTX 3060 is faster per-token on models that fit, but the P40 can load models the 3060 physically cannot. A 30B parameter model at Q4 quantization needs roughly 18-20GB of VRAM. The 3060 tops out at 12GB. The P40 loads it with room left over for context.

Real-World Performance

Benchmarks from LocalScore.ai and community testing:

Model	P40 Performance
LLaMA 7B Q4	~41 tok/s generation, 833 tok/s prompt processing
Qwen 2.5 14B Q4_K_M	~16 tok/s generation, 339 tok/s prompt processing
Qwen3 Coder 30B Q4_K_M	~29 tok/s generation, 288 tok/s prompt processing

Those generation speeds are usable. 16 tok/s on a 14B model is fine for chat. 29 tok/s on a 30B model is faster than most people expect from a $175 card. Prompt processing at 300-800+ tok/s means context loads quickly.

For reference, an RTX 3060 12GB does about 45 tok/s on LLaMA 7B Q4. The P40 is roughly 60-70% of RTX 3060 speed on equivalent models — but it can run models twice the size.

The Ideal P40 Setup

The best use case for a P40 is as a dedicated inference card alongside another GPU:

- **Display GPU:** Your existing gaming card (even a GT 1030 works) or CPU iGPU
- **Inference GPU:** Tesla P40 with fan mod
- **Software:** Ollama or llama.cpp with CUDA offloading to the P40
- **PSU:** 650W minimum (P40 draws 250W at peak)

This setup costs roughly $175 for the P40 + $15 for a cooling solution + $0 for the iGPU you probably already have. Under $200 total for a 24GB inference rig.

Server GPUs vs. consumer GPUs: side by side

	Tesla P40 24GB	RTX 3060 12GB	RTX 3090 24GB
Used Price	~$175	~$190	~$800
VRAM	24GB GDDR5X	12GB GDDR6	24GB GDDR6X
$/GB	$7.30	$15.80	$33.30
Memory Bandwidth	346 GB/s	360 GB/s	936 GB/s
FP16 Support	No (FP32 only)	Yes	Yes
Display Output	None	Yes	Yes
Cooling	DIY required	Included	Included

	Tesla P40 24GB	RTX 3060 12GB	RTX 3090 24GB
7B Q4 Speed	~41 tok/s	~45 tok/s	~90 tok/s
14B Q4 Speed	~16 tok/s	~20 tok/s	~50 tok/s
30B+ Q4	Yes	No (VRAM limit)	Yes
Power Draw	250W	170W	350W
Plug-and-Play	No	Yes	Yes

When each card makes sense

P40 wins when: You need VRAM above all else, you're on a strict budget, you're building a multi-GPU inference rig, or you enjoy the tinkering process. Nobody beats $7.30/GB.

RTX 3060 wins when: You want a single card that does everything — display, gaming, and AI. At $190, it's the best all-around budget card. The 12GB VRAM limits you to 14B models, but for most people starting out, that's enough.

RTX 3090 wins when: You want 24GB VRAM with full FP16 support, proper cooling, and display output. You're paying 4.5x the P40's price for convenience and about 2x the raw speed.

Who Should Buy a Server GPU

Buy a server GPU if:

- You want maximum VRAM for minimum dollars and don't mind DIY
- You're building a dedicated inference machine (headless or with a separate display card)
- You already have a gaming GPU for display and want to add a dedicated AI card
- You're stacking multiple GPUs for a multi-card setup
- You enjoy the tinkering: fan mods, driver troubleshooting, config tweaking

Skip server GPUs if:

- You're building your first AI setup (get an RTX 3060 12GB instead)
- You want one card for gaming + AI + display
- You don't want to deal with cooling mods
- You need native FP16 for training or Stable Diffusion (get a consumer card)
- You want something that works out of the box with zero configuration

Where to Buy

eBay is the primary market. Filter by “Buy It Now” for fixed pricing, or watch auctions for deals. The P40 routinely sells for $150-200 with free shipping. Look for sellers with 98%+ positive feedback and at least 100 sales.

Amazon occasionally has server GPUs, usually at a premium over eBay pricing.

AliExpress has cheap Tesla cards from Chinese sellers. Prices are lower but shipping takes 2-4 weeks and returns are harder. Quality is generally fine — these are genuine datacenter pulls.

Local server surplus dealers — check Craigslist and Facebook Marketplace for IT companies offloading old hardware. You can sometimes score P40s for $100-120 from bulk liquidations.

Buying Tips

- **Test immediately.** Run `nvidia-smi` and a sustained GPU benchmark within 24 hours. eBay buyer protection covers DOA cards, but you need to file within the return window.
- **Check the PCIe connector** for bent pins or corrosion. Datacenter cards get hot-swapped thousands of times.
- **Avoid SXM variants** unless you specifically know what you’re doing. PCIe cards only.
- **Buy the fan mod parts before the card arrives.** You’ll want to run the cooling mod before your first real workload.

The Bottom Line

Server GPUs aren’t for everyone. They’re headless, fanless, and require homework before you can use them. But if you’re chasing VRAM on a budget, nothing in the consumer market comes close.

A Tesla P40 at $175 gives you the same 24GB of VRAM as an $800 RTX 3090. You trade speed, convenience, and FP16 support for a 4.5x price reduction. For a dedicated inference card running 14B-30B models, that trade-off is worth it.

Get the P40 if you want cheap VRAM and enjoy the build. Get the RTX 3060 if you want easy. Get the RTX 3090 if you want both VRAM and speed without compromise.

Source: https://insiderllm.com/guides/used-server-gpus-local-ai/

Free guides for running AI locally