GEARFLOW LAB PRODUCTION SRE RUNBOOK 09 | 36-MONTH CLOUD VPS & GPU VS LOCAL AI HARDWARE TCO

Cloud VPS & GPU vs Local AI Hardware 36-Month TCO & Break-Even Matrix

Open TCO Calculator ->

Engineering Context: Self-hosted AI and DevOps teams often underestimate cloud egress, snapshot storage, and 36-month compounding on `$96-$515/mo` Cloud VPS/GPU instances while overestimating wall-plug electricity on modern UMA workstations (`28W` Apple Silicon, `85W` NVIDIA GB10). Use the deterministic 36-month model below before signing multi-year commitments.

1. Deterministic 36-Month TCO & Break-Even Equations (`30.4375` Gregorian Mean Days/Month)

- **Monthly Wall Energy:** `Monthly_kWh = (Avg_Watts * Hours_Per_Day * 30.4375) / 1000` | `Power_USD = Monthly_kWh * kWh_Rate`
- **Monthly Net Cashflow Delta:** `Net_Monthly_Saving = (Cloud_Base_USD + Cloud_Egress_USD) - Power_USD`
- **Gross Break-Even Month:** `Break_Even_Months = HW_CapEx_USD / Net_Monthly_Saving` (`ceil` = first cashflow-positive billing month)
- **36-Month Net TCO After Resale:** `HW_36M_Net = (HW_CapEx_USD + 36 * Power_USD) - (HW_CapEx_USD * Resale_Ratio_36M)`

2. Empirical 36-Month Cloud vs Local AI Workstation Matrix (`$0.14/kWh` Baseline)

Comparison Pair	Cloud Total/mo	Local CapEx + Power	Break-Even	36M Cloud vs Local Net	36M Net Saved
`DO 16GB` vs `Mac mini M6 32GB` (28W 24h)	`$108/mo ($96+$12)`	`$999 + $2.86/mo`	**`Month 10`** (9.50m)	`$3,888` vs `$702.49` (40% rsl)	**`+$3,185.51`**
`Cloud A10G` vs `RTX 5070 Ti 16GB` (165W 12h)	`$260/mo ($240+$20)`	`$1,850 + $8.44/mo`	**`Month 8`** (7.35m)	`$9,360` vs `$1,321.24` (45% rsl)	**`+$8,038.76`**
`Cloud L40S` vs `NVIDIA GB10 128GB` (85W 24h)	`$515/mo ($480+$35)`	`$3,999 + $8.69/mo`	**`Month 8`** (7.90m)	`$18,540` vs `$2,312.45` (50% rsl)	**`+$16,227.55`**
`Managed 8GB` vs `x86 Mini PC 32GB` (18W 24h)	`$96/mo ($88+$8)`	`$599 + $1.84/mo`	**`Month 7`** (6.36m)	`$3,456` vs `$485.57` (30% rsl)	**`+$2,970.43`**
`DO 32GB` vs `Apple 64GB UMA` (42W 24h)	`$210/mo ($192+$18)`	`$1,999 + $4.29/mo`	**`Month 10`** (9.72m)	`$7,560` vs `$1,254.05` (45% rsl)	**`+$6,305.95`**
`Burst GPU` (<25h/mo) vs `RTX 5070 Ti` (8h)	`$70/mo ($65+$5)`	`$1,850 + $5.62/mo`	`Month 29` (28.74m)	`$2,520` vs `$1,219.99` (45% rsl)	`+$1,300` (Long)

3. Live Host Power & Memory Audit CLI (Linux / macOS)

```
# 1. NVIDIA dGPU (RTX 5070 Ti / 4090) board power draw (W) & VRAM
nvidia-smi --query-gpu=name,power.draw,memory.used,memory.total \
  --format=csv,noheader

# 2. NVIDIA GB10 128GB UMA: read MemAvailable (NVML memory=N/A)
grep -E 'MemTotal|MemAvailable|SwapUsed' /proc/meminfo
nvidia-smi --query-gpu=power.draw,temperature.gpu --format=csv

# 3. macOS Apple Silicon package power & Docker RSS snapshot
sudo powermetrics -i 2000 -n 1 --samplers cpu_power | grep -E "Combined Power"
docker stats --no-stream --format "table {{.Name}}\t{{.MemUsage}}"
```

4. Hybrid Placement Decision Checklist (Cloud vs Local)

- **Keep Public Ingress Lightweight (`$6-$12/mo` Edge VPS):** Terminate webhooks, TLS, or Cloudflare Tunnel on a tiny VPS while routing heavy LLM inference (`Ollama` / `vLLM`) and vector builds to local UMA hardware.
- **Verify VRAM Fit Before Buying:** Confirm quantized weights + GQA `Q8_0` KV cache fit with `>=15%` OS headroom via the **Local LLM VRAM Calculator**.
- **Account for UPS & Off-Host Backups:** Budget `$120` LiFePO4/AVR UPS + automated `pg_dump` + `rclone` encrypted S3/B2 snapshots (`$1-$3/mo`) before retiring managed cloud databases.
- **Stay on Cloud GPU When:** Utilization is `< 15%` (`< $75/mo` burst jobs) or strict customer SOC2 / HIPAA BAA physical datacenter controls apply.

1-Click Prefill TCO Models: DO $96 vs Mac mini M6 32GB | Cloud GPU vs RTX 5070 Ti 16GB | Enterprise GPU vs GB10 128GB | VPS RAM Sizer | All Tools (/tools/)