WhitePaper • VOLTAGEGPU

Bittensor, benchmark compute vs AWS
& inference snapshot

VOLTAGEGPU · 30/09/2025

Executive Summary

I. GPU Compute and AI Inference

A. Problem → Solution → Benefits → Key Figures

Problem

GPU Compute: Access to dedicated GPUs is often expensive and slow; pricing can be unpredictable (egress, idle time, over/under-provisioning).
AI Inference: Putting LLM/vision/audio endpoints into production that handle traffic spikes (autoscale, P50/P95, SLAs) usually demands heavy Ops.

Proposal

GPU Compute: Spin up Dedicated GPU Instances ("GPU Pods") in minutes with pay-per-use pricing, profile choice (GPU/memory/region), and basic monitoring.
AI Inference: Use Managed Model Endpoints: choose a model + resources, get a ready HTTPS endpoint. Powered by a global decentralized network (Bittensor).

Result

GPU Compute: Capacity when you need it, faster time-to-GPU, lower effective cost via competition + pay-per-use, transparent billing, no lock-in.
AI Inference: Ready-to-use endpoints, less Ops, observable latency/throughput, optional SLAs — serve models at scale without managing infra.

B. Key Figures (examples)

Time-to-GPU: from hours → minutes (ex. A100 nodes, Washington US).

Effective cost: ↓ ~70–80% vs AWS on-demand lists (e.g. A100 $6.02/h vs $27–41/h; B200 $41.86/h vs $113.93/h).

Inference unit cost: ↓ up to ~90–99% vs hosted APIs (e.g. DeepSeek-R1 $0.46 in / $1.85 out vs GPT-5 $1.25 in / $10.00 out).

II. Bittensor – How does it work?

Whitepaper Bittensor

A. In one paragraph

Bittensor is a peer-to-peer intelligence market: independent nodes (peers) publish model outputs and evaluate one another. Peers learn weights that score the informational value of their neighbors; scores are recorded on a digital ledger and convert into rewards/stake. An incentive function favors peers trusted by the majority, making collusion (≤ ~50% of network weight) unprofitable. The result is a continuously trained, collectively run market for machine intelligence.

B. Why it matters (engineer view)

- *Measurement by other models (not one static benchmark) rewards useful signal, even for niche/legacy systems.*
- *Incentives + ledger = open pricing and ongoing model improvement without a single central operator.*
- *Standardized "modalities" let the network span text, image, audio—enabling general-purpose inference.*

C. How VoltageGPU uses it (GPU compute & inference)

- *GPU Compute (GPU Pods): we aggregate supply from operators connected to the network → fast, pay-per-use access to dedicated GPUs (profile = GPU/memory/region) with transparent billing.*
- *AI Inference (Managed Model Endpoints): choose a model + resources → get a ready HTTPS endpoint powered by a global decentralized network (Bittensor); run LLM/vision/audio without managing infra.*
- *Market dynamics: competition across providers drives time-to-GPU down and keeps effective costs competitive vs public on-demand lists and hosted APIs.*

III. BENCHMARK

A. Results: VOLTAGEGPU vs AWS GPU COMPUTE

8x NVIDIA A100						
Provider	SKU / GPUs	GPU Mem	Region (exemple)	Billing	$/hour (node)	Source
VOLTAGEGPU	8× A100-SXM4-80GB	8×80 GB	Washington, US	On-demand	$6.02	see links
AWS	p4de.24xlarge (8× A100 80GB)	8×80 GB	us-east-1 / us-west-2	On-demand	~$27.45–$40.97	see links
AWS (alt)	p4de.24xlarge (8× A100 80GB)	8×80 GB	us-east-1 / us-west-2	Capacity Blocks	$14.77	see links

8x NVIDIA H200						
VOLTAGEGPU	8× H200	8×141 GB	Chiyoda, JP	On-demand	$26.60	see links
AWS	p5e.48xlarge (8× H200)	8×141 GB	us-east-2 / us-west-2 / eu-north-1	Capacity Blocks	$34.61–$45.23	see links

8x NVIDIA B200						
VOLTAGEGPU	8× B200	8× 192 GB HBM3e (≈ 8× 179 GiB ≈ 1.44 TB)	Queens, US	On-demand	$41.86	see links
AWS	p6-b200.48xlarge (8× B200)	8× 179 GiB (total 1432 GiB ≈ 1.44 TB)	us-east-1 (N. VA)	On-demand	$113.93	see links

B. Results: VOLTAGEGPU vs OPEN AI - AI INFERENCE

API USAGE							
Provider	Model	Class	$/1M Input	$/1M Output	Notes	Date	Source
VOLTAGEGPU	DeepSeek-R1	Reasoning	$0.46	$1.85	Endpoint managed	Sep 2025	see links
OPENAI	GPT-5	Reasoning	$1.25	$10.00	Flagship	Sep 2025	see links
VOLTAGEGPU	Deepseek R1 0528 Qwen3 8B	General	$0.02	$0.10	Endpoint managed	Sep 2025	see links
OPENAI	GPT-4.1	General	$2.00	$8.00	Large	Sep 2025	see links
OPENAI	GPT-4o mini	General	$0.60	$2.40	Mini / fast	Sep 2025	see links
VOLTAGEGPU	zai-org/GLM-4.5-Air	Compact	FREE	FREE	Endpoint managed	Sep 2025	see links
OPENAI	GPT-4.1 mini	Compact	$0.80	$3.20	Mini	Sep 2025	see links

C.API Compatibility

VoltageGPU's inference API is **OpenAI-compatible**: you can point any OpenAI client/SDK to our base URL and keep the same request/response schemas. That means familiar endpoints, headers, and JSON (e.g., messages, choices[0].message.content, usage). Start with **chat** for multi-turn assistants or **completions** for simple prompts. Endpoints:

- POST https://api.voltagegpu.com/v1/chat/completions — generate chat completions (e.g., deepseek-ai/DeepSeek-R1).

- POST https://api.voltagegpu.com/v1/completions — generate text completions with the same model.
  Authentication: Authorization: Bearer YOUR_API_KEY. Optional: set OpenAI-Organization if your tooling expects it.

Run with our API (code examples, reference, schemas)

IV. Value Proposition & Use Cases

A. Cost Formula

Cost = (InputTokens / 1,000,000) × ($/1M Input) + (OutputTokens / 1,000,000) × ($/1M Output)

ROI (%) = (Cost_baseline − Cost_VoltageGPU) / Cost_baseline × 100

a) Use Case 1 — Customer Support Assistant (Reasoning)

Models compared: VoltageGPU DeepSeek-R1-0528 **vs** OpenAI GPT-5

Pricing used:

- VoltageGPU R1-0528 → **$0.46/M in**, **$1.85/M out**
- OpenAI GPT-5 → **$1.25/M in**, **$10.00/M out**

Assumptions (monthly): 2,000,000 chat turns · avg **200 in / 130 out** tokens

- Input = 2,000,000 × 200 = **400M**
- Output = 2,000,000 × 130 = **260M**

Costs (by formula):

- **VoltageGPU:** (400×$0.46) + (260×$1.85) = **$184 + $481 = $665**
- **OpenAI:** (400×$1.25) + (260×$10.00) = **$500 + $2,600 = $3,100**

Savings / ROI (with R1):

- Savings = $3,100 − $665 = **$2,435**
- ROI = 2,435/ 3,100 = **78.6%**

b) Use Case 2 — Bulk Document Summaries / KB Refresh (General)

Models compared: VoltageGPU deepseek-ai/DeepSeek-R1-0528-Qwen3-8B **vs** OpenAI GPT-4.1

Pricing used:

- VoltageGPU Qwen3-8B → **$0.02/M in**, **$0.10/M out**
- OpenAI GPT-4.1 → **$2.00/M in**, **$8.00/M out**

Assumptions (monthly):

Daily 100M tokens in / 30M out ⇒ **3,000M in**, **900M out** per month

Costs:

- **VoltageGPU:** (3,000×$0.02) + (900×$0.10) = **$60 + $90 = $150**
- **OpenAI:** (3,000×$2.00) + (900×$8.00) = **$6,000 + $7,200 = $13,200**

Savings / ROI:

- Savings = $13,200 − $150 = **$13,050**
- ROI = 13,050 / 13,200 = **98.9%**

c) Use Case 3 — Product Content Generator @ Scale (Compact / Budget)

Models compared: VoltageGPU zai-org/GLM-4.5-Air (Free) **vs** OpenAI GPT-4.1 mini
Pricing used:

- VoltageGPU GLM-4.5-Air → **FREE** (in/out)
- OpenAI 4.1 mini → **$0.80/M in**, **$3.20/M out**

Assumptions (monthly): 500,000 requests/day, avg **120 in / 120 out** tokens, 30 days

- Input = 500k × 120 × 30 = **1,800M**
- Output = **1,800M**

Costs:

- **VoltageGPU: $0**
- **OpenAI:** (1,800×$0.80) + (1,800×$3.20) = **$1,440 + $5,760 = $7,200**

Savings / ROI:

- Savings = **$7,200**
- ROI = 7,200 / 7,200 = **100%**
  (Note: subject to Free model availability/rate limits.)

V. Conclusion

Across compute and inference, VoltageGPU delivers **faster time-to-GPU**, **transparent pricing**, and **OpenAI-compatible endpoints** that let you ship without refactoring. For reasoning workloads we benchmark favorably vs flagship APIs; for general and compact classes, we often deliver **2–10× lower $/token** while keeping latency and observability practical for production.

3 Key Takeaways

OpenAI-compatible by design: keep your code and SDKs — just swap the baseURL and your API key to hit VoltageGPU.

Cost efficiency at scale: lower effective $ per hour for multi-GPU nodes and competitive $ per 1M tokens for inference.

Less Ops, more product: launch GPU Pods in minutes and use managed model endpoints with metrics, logs, and optional SLAs.

VI. CTA

CTA + Offres

Get 5% discount: SHA-256-C7E8976BBAF2
Start now: VOLTAGEGPU.COM
Contact: contact@voltagegpu.com
Discord (support & updates): VOLTAGEGPU
X / Twitter: @Voltage_GPU