MCP SERVER | NO CODE | CLOUD HOSTED

AI Infrastructure Cost Optimizer MCP.

Model your GPU spend and hosting decisions with precision.

AI Infrastructure Cost Optimizer MCP gives your AI client the math needed to handle heavy computing budgets. It models the financial impact of GPU utilization, compares cloud versus on-premise hosting, and estimates the capital or operating expenses required for your specific AI workloads.

A+ Quality Score 100/100

ai | gpu | cloud | cost-optimization | infrastructure

MAINTAINED BY VINKIUS CLOUD FOR AI AGENTS · SEPTEMBER 2026

vinkius.com/en/ai-agent-connect/ai-infrastructure-cost-optimizer

SCAN TO VIEW

The connectivity layer between AI and the world's software.

Vinkius sits between AI and every application. All communication passes through Vinkius Cloud via the Model Context Protocol (MCP) — with governance, observability, and security at every layer.

Your AI Connections Run Through Vinkius Cloud

The world's largest managed MCP catalog

Vinkius is the connectivity layer where AI connects to the software your business already runs. We handle the hosting, the security, the credentials, the uptime — you get agents that actually do things.

We operate the world's largest managed MCP catalog. Major SaaS platforms, CRMs, databases, and cloud providers — running, monitored, production-ready. This MCP server is hosted and maintained by the Vinkius Cloud for AI Agents.

> *The agent doesn't manage credentials, doesn't manage uptime, doesn't manage security. Vinkius does.*
>
> — Architecture principle

Four Pillars of the Vinkius Runtime

01 — Security by design

Credentials stay encrypted at rest via AES-256. The AI agent never touches raw keys — they're injected into a sandboxed V8 isolate at runtime. Actions are logged, and connections have an emergency kill switch.

02 — Built on MCP Fusion

This MCP server was built with **MCP Fusion**, the open-source framework (Apache 2.0) that powers the entire Vinkius catalog. Schema-as-firewall strips undeclared fields, compiled PII redaction runs at zero overhead, and cryptographic lockfiles produce git-diffable audit trails.

03 — Deterministic observability

Eight immutable metrics per endpoint: request volume, p95 latency, error rate, active connections, cost attribution. A live payload feed logs every tool call with mutation detection.

04 — Autonomous operations

Servers are deployed, monitored, and patched autonomously. New capabilities and security patches ship weekly. Zero-downtime deployments ensure continuous availability across all managed MCP servers.

AES-256	Ed25519	24h TTL	V8 Isolate
Encryption at rest	PKI vault signatures	Ephemeral session keys	Sandboxed execution

One Token. Instant Access.

Every MCP server on Vinkius is accessed through a **Connection Token**. Tokens are generated in the cloud dashboard and produce a unique MCP endpoint URL. Paste this URL into any MCP-compatible client — no SDK required.

A single token can serve **multiple AI clients simultaneously**, or you can issue separate tokens per client for granular access control. Each token tracks its own request count, last activity timestamp, and can be individually enabled or revoked.

MCP ENDPOINT

```
https://edge.vinkius.com/{token}/mcp
```

Claude · Cursor · VS Code · Windsurf · Grok · Gemini

Security Is the Architecture

Security in Vinkius is not a feature — it's the foundation of the runtime. The gateway enforces multiple independent protection layers between AI agents and third-party APIs.

01 — Ed25519 PKI Vault

Every workspace has an Ed25519 Master Key. Session keys are generated ephemerally (24h TTL) and signed by the Master Key. Credentials never leave the vault boundary.

02 — V8 Isolate Sandboxing

Tool code runs inside isolated-vm V8 isolates with 64 MB memory caps and per-request timeouts. No filesystem access, no network access except through the SSRF-guarded fetch bridge.

03 — SSRF Guard

All outbound HTTP requests are DNS-resolved and validated before execution. Private IP ranges (10.x, 172.16-31.x, 192.168.x, AWS metadata 169.254.x) are blocked at the network layer.

04 — DLP & PII Redaction

A ResponseGuard pipeline intercepts every tool response. Configurable redaction patterns strip sensitive fields (emails, SSNs, card numbers) before data reaches the AI agent.

05 — Cryptographic Audit Trail

Every request is signed into a SHA-256 hash chain with Ed25519 signatures. Events form a tamper-proof, SIEM-exportable forensic record.

06 — Honeytoken Trap System

Phantom credentials are injected into isolated environments. If a honeytoken is used outside Vinkius infrastructure, the server is quarantined instantly.

Emergency Kill Switch

EU AI Act Art. 14(1) Compliant

The kill switch is an **emergency halt** mechanism — not a simple toggle. When triggered, it executes three actions atomically:

01 — Server deactivated

The MCP server is immediately taken offline across the entire cluster.

02 — All tokens revoked

Every connection token is invalidated. Total lockout — reconnection blocked until new tokens are issued.

03 — WebSocket connections killed

Active connections terminated via Redis pubsub broadcast. Propagates to every runtime node in the cluster.

Full Visibility. Zero Guesswork.

The Vinkius cloud dashboard includes a full MCP Governance suite — real-time analytics and security controls for production AI operations.

Control Plane

KPI dashboard with request volume, latency, success rate, token consumption, and AI-generated operational briefings.

Agent Activity

Which AI clients are connecting, how often, and what they're doing — real-time session tracking.

FinOps

Cost tracking per tool, payload compression savings, budget optimization signals, and consumption trends.

Tool Health

Slowest and most error-prone tools, with actionable root-cause insights and performance baselines.

Firewall & DLP

PII redaction activity, sensitive data protection counters, and security event timeline.

Incident Log

Error trends, failure rates, status-code breakdowns, and forensic audit trail access.

> Get started at **cloud.vinkius.com** — connect your AI agent in under 60 seconds.

AI Infrastructure Cost Optimizer MCP

4 tools available
Cloud-hosted on Vinkius

Managing AI compute costs is a moving target. This MCP gives your agent the ability to run financial models specifically for AI-heavy environments. Instead of guessing how much a cluster of GPUs will cost or whether to move to on-premise hardware, you can use this tool to get concrete numbers. You can model the annual savings from rightsizing instances or switching to spot instances. It also helps you decide between CapEx and OpEx by estimating investment requirements for new hardware. Whether you are deciding between cloud providers or evaluating the long-term cost of running your own hardware, this MCP provides the data to back up your infrastructure decisions.

Core Capabilities

01 — Annual Savings Modeling

Your agent calculates potential yearly reductions in compute spend based on specific optimization tactics.

02 — Hosting Model Comparison

The tool compares the financial viability of cloud versus on-premise setups for your specific workload volume.

03 — Investment Estimation

Your agent determines the required capital or operating expenditures for new infrastructure projects.

04 — Deployment Planning

The tool provides implementation timelines and complexity scores for infrastructure changes.

One Click on Vinkius — From Prompt to Execution

Available at **vinkius.com/en/ai-agent-connect/ai-infrastructure-cost-optimizer** — connect your AI agent in three steps.

01 Connect the MCP to Claude, Cursor, or Windsurf via Vinkius.

02 Provide your workload volume and current utilization data to your agent.

03 Ask your agent to run specific calculations like savings or hosting comparisons.

04 Receive detailed financial estimates and implementation timelines.

Connect the MCP to your client and start running financial models immediately.

Built For

This MCP is built for technical leaders managing the high costs of AI compute. It turns raw workload data into financial models.

AI Infrastructure Engineer

Uses the tool to model the impact of rightsizing and GPU efficiency on existing clusters.

FinOps Analyst

Uses the tool to compare cloud versus on-premise costs to optimize company spend.

CTO

Uses the tool to estimate CapEx and OpEx requirements for upcoming AI projects.

What Changes When You Connect

01 Models annual savings from rightsizing and spot instance usage.

02 Compares cloud and on-premise costs based on workload volume.

03 Estimates CapEx and OpEx requirements for new hardware.

04 Provides implementation timelines and complexity scores for optimizations.

Real-World Applications

Cloud Rightsizing

Your agent calculates how much money you'll save by adjusting your current cloud GPU utilization.

On-Premise Migration

You use the tool to decide if moving workloads from the cloud to local hardware is cheaper at your current volume.

Budget Planning

The tool helps you estimate the total investment needed for a new AI computing environment.

Optimization Scheduling

You determine how long a specific GPU efficiency project will take to complete.

Patterns to Avoid

Guessing GPU savings

AVOID

You ask your agent to guess how much money you'll save by switching to spot instances without any data. This leads to unreliable budget forecasts.

✓ INSTEAD

Use `calculate_savings_potential` to get a concrete estimate of annual reductions based on specific optimization tactics.

Ignoring deployment complexity

AVOID

You decide to migrate all workloads to on-premise hardware immediately without checking how long it takes. This ignores the actual labor and time required.

✓ INSTEAD

Run `project_implementation_timeline` to get a realistic timeframe and a complexity score for your infrastructure changes.

Neglecting CapEx requirements

AVOID

You plan a massive hardware expansion based only on monthly operating costs. You'll run into trouble when you realize you haven't budgeted for the initial purchase.

✓ INSTEAD

Use `estimate_investment_requirements` to determine the necessary CapEx or OpEx for your planned computing environment.

The Right Fit

Technical leaders and FinOps analysts should connect this MCP if they manage high-cost AI compute environments. If you are just running small, low-cost hobbyist models, you don't need this. The deciding factor is whether you need to justify large infrastructure investments or cloud spend to stakeholders.

4 Tools for AI Infrastructure Financial Modeling

These tools let your agent run specific financial simulations for GPU clusters and cloud workloads. You can model savings, compare hosting environments, and estimate hardware investment needs.

#	TOOL	DESCRIPTION
01	`project_implementation_timeline`	This tool calculates the expected timeframe for deploying specific infrastructure optimizations. It provides a timeline and a complexity score for your planning.
02	`calculate_savings_potential`	This tool estimates how much money you can save annually through optimization strategies like rightsizing or using spot instances.
03	`compare_cloud_vs_onprem`	This tool compares the annual costs of hosting your workloads in the cloud against running them on your own on-premise hardware.
04	`estimate_investment_requirements`	This tool determines the necessary CapEx or OpEx needed to support your planned AI computing infrastructure.

See It in Action

Real prompts you can use once this MCP is connected to your AI agent through Vinkius Cloud.

U How much can I save if I spend 500,000 Euros on cloud AI infrastructure with 40% utilization using rightsizing?

By applying rightsizing to your cloud infrastructure, you can achieve an estimated annual savings of 150,000 Euros.

U Compare cloud vs on-prem for a workload volume of 1000 units with 70% utilization.

For a workload of 1000 units, on-premise is the preferred option with a projected annual cost of 450,000 Euros compared to 600,000 Euros in the cloud.

U How long will it take to implement gpu_efficiency optimizations?

Implementing gpu_efficiency optimizations is expected to take 8 months with a complexity score of 7.

Frequently Asked Questions

01 **What can this MCP calculate?**

It calculates annual savings from optimization, compares cloud versus on-premise costs, and estimates investment requirements.

02 **Which AI clients can use this MCP?**

You can use this MCP with any compatible client like Claude, Cursor, or Windsurf.

03 **Does this tool help with GPU planning?**

Yes, it models the cost impact of GPU utilization and provides implementation timelines for efficiency optimizations.

04 **Can I compare cloud and on-premise costs?**

Yes, the compare_cloud_vs_onprem tool lets you decide the best hosting model based on your workload volume.

05 **How does it handle CapEx and OpEx?**

The estimate_investment_requirements tool determines the specific capital or operating expenditure needs for your infrastructure.

06 **How do I calculate my potential savings?**

You can use the `calculate_savings_potential` tool by providing your current annual spend, environment type, and current utilization rate.

07 **What kind of investment is required for optimizations?**

The `estimate_investment_requirements` tool will tell you if the required investment is CapEx or OpEx based on your chosen optimization levers.

Go Live in 60 Seconds

Get your connection token from **cloud.vinkius.com**, then paste the endpoint URL into any MCP-compatible client.

YOUR MCP ENDPOINT

```
https://edge.vinkius.com/[TOKEN]/mcp
```

CLIENT	WHERE TO CONFIGURE
Claude AI	Profile → Customize → Connectors → "+" → Add custom connector → Paste endpoint
Cursor	Settings → Features → MCP Servers → "+ Add New MCP Server" → Type: SSE → Paste endpoint
VS Code	Ctrl/Cmd+Shift+P → "MCP: Add Server" → add `"ai-infrastructure-cost-optimizer": { "url": "…" }`
Windsurf	MCP Settings → `mcp_settings.json` → Add endpoint URL
ChatGPT	Settings → Tools & plugins → Add MCP server → Paste endpoint
Gemini	Extensions → Add MCP Server → Paste endpoint URL

ASK AN AI ABOUT THIS

Let your preferred AI explain this MCP server

- Ask ChatGPT
- Ask Claude
- Ask Perplexity
- Ask Gemini
- Ask Grok

READY TO CONNECT

AI Infrastructure Cost Optimizer is live on Vinkius Cloud.

Get your connection token, paste it into your AI agent, and start building. No SDK. No deployment. Just results.

Start at cloud.vinkius.com →

vinkius.com · support@vinkius.com

INDEPENDENT PLATFORM DISCLAIMER

Vinkius is an independent platform and is not affiliated with, endorsed by, sponsored by, verified by, or otherwise authorized by AI Infrastructure Cost Optimizer. All third-party trademarks, logos, and brand names are the property of their respective owners. Their use in this document is strictly for informational purposes to identify service compatibility and interoperability.

DOCUMENT INFORMATION

Generated	September 2026
MCP Server	AI Infrastructure Cost Optimizer MCP
Server ID	`01a09da4-698f-72a1-80de-755941ef4ccc`
Platform	Vinkius Cloud for AI Agents
Endpoint	`https://edge.vinkius.com/{token}/mcp`

LICENSE & USAGE

This document is generated automatically by the Vinkius PDF Engine. Content reflects the MCP server configuration at the time of generation and may change as updates are deployed. For the most current information, visit vinkius.com/en/ai-agent-connect/ai-infrastructure-cost-optimizer.