AURORALABS

LOCI AI Optimization Agent for **AI Cloud Infrastructure**

Solution Brief

Executive Summary

Most AI cost overruns come from bad or risky AI Executables hitting your clusters: wasted GPU hours, rollbacks, and customer-visible slowdowns. LOCI adds a pre-run gate that checks the executable (the packaged program you deploy) before it's scheduled. It needs no source code and no hardware time. You get a clear PASS/FAIL against budgets and reliability targets—whether the artifact is built in your CI or delivered by a partner.

Why it matters: You stop waste before it starts, safely increase models per machine, and keep SLOs predictable.

09.25

AURORALABS

The problem LOCI solves

- You only discover performance or cost problems after deployment → wasted GPU hours and rollbacks
- Busy clusters get slowed by noisy builds, hurting other teams and customers
- Partner/marketplace packages arrive without an easy way to trust their cost and reliability

What LOCI is

- A pre-run safety check for AI executables
- Think of it as admission control: if an artifact risks breaching cost or SLO targets, it doesn't ship
- Works on your builds and partner deliverables - **Bring-Your-Executable (BYE)**
- Observability-only: LOCI does not change or "auto-optimize" code; it just makes the go/no-go decision easy

How LOCI works in 3 steps

1. Submit the executable: from CI/CD or upload a partner package (BYE).
2. LOCI checks against your policies: latency budgets, throughput impact, network/fabric safety, and energy/cost budgets.
3. You get PASS/FAIL + a short summary (what to fix or why it's safe). PASS can auto-approve; FAIL can auto-block.

What you get

- Immediate cost protection: fewer wasteful runs on expensive hardware.
- Predictable reliability: risky artifacts never make it to customers.
- More models per machine (safe density): cleaner "neighbors" → higher utilization and revenue capacity.
- Faster approvals: clear, auditable policy gates in CI or via BYE.

AURORALABS

Policies you can set

- Latency: “95% of requests (p95) must be ≤ 120 ms.”
- Throughput: “Don’t admit builds that reduce tokens/sec by more than 10% vs. last good build.”
- Network safety: “Don’t admit builds likely to congest the cluster.”
- Power/Cost: “Energy per token must be within 10% of baseline.”
- Density eligibility: “Only admit builds that are safe to co-locate on the same machine.”

BYE : Partner & Marketplace Intake

- Treat partner packages with the same gates as your own builds - no source code needed
- Optional trust checks: keep an audit trail of PASS/FAIL with rationale
- Privacy note: LOCI reviews executable metadata only - no training data is ingested

Measurable KPIs for Ops

- GPU-hours avoided per week
- Rollbacks avoided
- Models per machine (safe density)
- Revenue per GPU-day or $ per 1K tokens
- SLO tail risk (p95 / p99)

Back-of-envelope ROI Example

(static analysis → power spikes & throttling)

- Catch **(pre-run)**: Inefficient loop → brief **power spikes** → GPU **throttling**
- Cost **avoided: ~12%** lower tokens/sec if not blocked ⇒ **~12% more GPU-hours** + extra **kWh** during spikes
- **Example**: 1,000 GPUs @ $2.50/hr, 20 hr/day → $50,000/day baseline.
- Avoiding 12% throttling ≈ $6,000/day GPU-hour savings; spike (+70 W × 4 hr) ≈ 280 kWh/day → $34/day
- Total ≈ $6,034/day avoided
- **Takeaway: LOCI prevents the waste and SLO hits before deployment**

Calc reference: Daily Savings ≈ (GPU spend/day × throttling_avoided%) + (kWh_avoided/day × price_kWh)

AURORALABS

What changes day-to-day

- Before: Trial-and-error on clusters; find problems late; firefighting and escalations
- After: Auto-gated releases; fewer rollbacks; stable spend; exec-level dashboard of waste avoided & density uplift

What LOCI is Not

- Not a runtime profiler or simulator (no hardware time)
- Not an auto-optimizer (your teams decide how to fix)
- Not limited to your build system—works equally well with partner artifacts

Next Steps

Upload build artifacts from your compilers or SDKs and receive actionable telemetry within minutes.

- **Book a session with our team**
- Contact us at info@auroralabs.com or visit www.auroralabs.com.

www.auroralabs.com

Follow us: