AXIOMOS

Decide Strategy Audit

FINAL RECOMMENDATION: GPU LOCAL (ON-PREM)

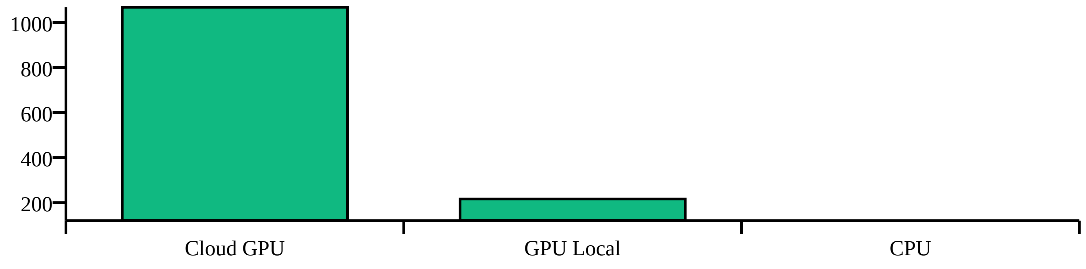
Hardware Target: NVIDIA RTX4090 Cluster

ARCHITECTURE	ANNUAL SAVINGS	BREAK-EVEN	PROJECT ROI
GPU Local (On-Prem)	**$21,406**	**2.5 mo**	**1,327 %**

Financial Arbitrage: This report upgrades your initial technical signal by integrating your specific OpEx constraints.

Cost & Performance Comparison

Option	Monthly Cost	Latency	Verdict
CPU	$120	450 ms	Alternative
GPU Local	$216	60 ms	Recommended
Cloud GPU	$1,068	119 ms	Alternative



36-Month Financial Projection

Timeline	Monthly Savings	Cumulative Cash
Month 1	+$1,784	$-2,716
Month 3	+$1,784	$851
Month 6	+$1,784	$6,203
Month 12	+$1,784	$16,906
Month 18	+$1,784	$27,609
Month 24	+$1,784	$38,312
Month 36	+$1,784	$59,718

Decision Rationale & Strategic Context

EXECUTIVE DECISION REPORT

Workload: INFERENCE | Hardware Target: RTX4090
Usage Profile: CONTINUOUS

RECOMMENDED ARCHITECTURE

■ GPU Local
Confidence Score: HIGH

WHY THIS DECISION IS ROBUST

- Aligned with the requested Latency Target of 200.0ms.
- Optimized for 5,000 daily requests.

FINANCIAL IMPACT & COMPARATIVE ADVANTAGE

Note: Free diagnosis provides a conservative signal estimate. Premium analysis recalculates ROI using your actual monthly spend.

- Initial Investment (CAPEX): $4,500
- Estimated Annual Savings (vs Managed APIs): $21,406
- Break-even point: 2.5 months
- Project ROI (36 months): 1,327.1%

SENSITIVITY ANALYSIS (RISK MITIGATION)

- High Cost Scenario (+25% Exp): $16,054 annual savings
- Optimized Scenario (+15% Eff): $24,617 annual savings

UNIT ECONOMICS (PROPRIETARY VS MANAGED)

- Proprietary Cost per 1M Tokens: $2.815
- Market Benchmark (Public APIs): $0.150
- Efficiency Multiplier: 0.1x better than managed endpoints

MONTHLY TCO COMPARISON

- GPU Local: $216 / month (Recommended)
- GPU Cloud: $1,068 / month
- CPU: $120 / month

STRATEGIC RATIONALE & CONTEXT

- Strategic Context: The ultra-low latency achieved (60ms) is below your 200.0ms target, ensuring a seamless user experience for real-time applications.
- Hardware Strategy: The RTX 4090 workstation-class hardware provides the best performance-per-dollar ratio for mid-scale AI deployments.
- Acceleration Impact: Dedicated hardware provides a 7.5x speedup over CPU, drastically reducing operational bottlenecks and compute waiting times.
- Operational Strategy: Continuous 24/7 workload maximizes hardware utilization, making ownership the most aggressive path to reducing long-term OpEx.
- Cloud Context: AWS provides premium security and compliance, justifying costs for enterprise environments.
- Strategic Moat: Internalizing this workload ensures 100% data sovereignty and alignment with security protocols by avoiding external API dependencies.
- Synthesis: With HIGH confidence, we recommend GPU LOCAL. This decision balances immediate performance targets with the flexibility required as your product evolves.

RISKS, SCALABILITY & COMPLIANCE

- Sovereignty: 100% Data Residency (SOC2, HIPAA & CCPA ready infrastructure)
- Scalability: Supports quantization (4-bit/8-bit) for model upgrades
- Asset Value: High hardware residual value after depreciation

PRIORITIZED NEXT STEPS

- Finalize hardware procurement
- Audit power & cooling requirements
- Setup local inference stack

--- END OF EXECUTIVE REPORT ---