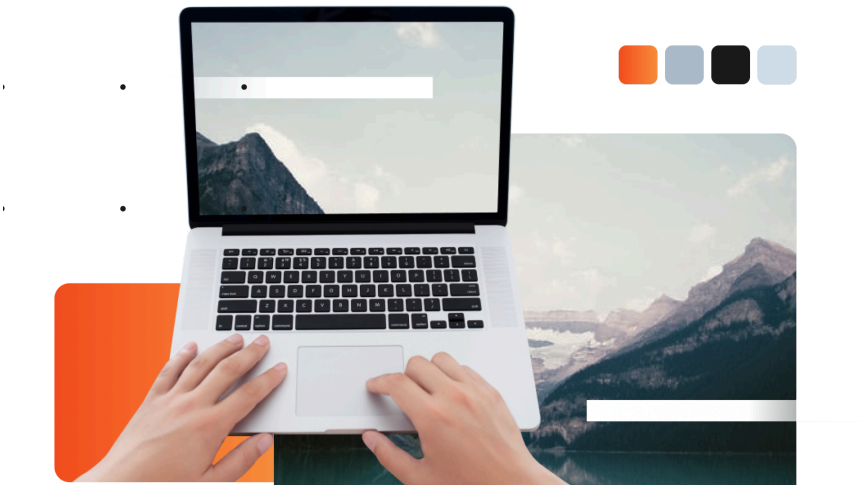
LuxiEdge Performance validation report



To:	Eric Waller, LuxiEdge e@ewaller.com
Prepared by:	TestFort
Date:	December 2025
Contact Person:	Nora Laievska, Director of Partnerships & Growth laievska@qarea.us

1. Executive Summary

TestFort conducted independent performance, stability, and determinism validation of LuxiEdge, a Rust-based deterministic numeric compute engine with NVIDIA CUDA GPU acceleration. Testing was performed on a Lambda Labs H100 SXM instance over multiple sessions, culminating in a 1-hour sustained load test with 200 concurrent virtual users.
Key Findings:

- Raw GPU kernel throughput: 286.94 billion operations/second (aggregate across 7 functions)
- Peak single-function throughput: 331.13 billion operations/second (sqrt)
- API throughput under sustained load: 2.80 million operations/second
- Total operations executed: 444.4 trillion (1-hour test)
- Error rate: 0.00%
- Determinism: Verified via SHA-256 hash consistency across 5 consecutive runs
- CPU and GPU modes produced identical output hashes

2. Test Environment

Parameter	Value
Platform	Lambda Labs
GPU	NVIDIA H100 SXM
CUDA Version	12.8
Operating System	Ubuntu
Test Duration	1 hour (sustained load)
Concurrency	200 virtual users
Load Tool	JMeter, k6
Vector Size	1,000,000 elements per request

3. RAW GPU Kernel Performance

Individual kernel benchmarks were executed with 1,000,000 element vectors:

Function	Throughput (Bops/sec)	Latency (ms)
sin	310.87	32.2
cos	310.78	32.2
exp	206.30	48.5
log	269.38	37.1
sqrt	331.13	30.2
sinh	310.44	32.2
tanh	316.42	31.6
AGGREGATE	286.94	—

Aggregate Performance:

- Total operations: 70.0 billion
- Total time: 0.24 seconds
- Raw throughput: 286.94 billion ops/sec
- Average power: 116.0W
- Efficiency: 2.35 billion ops/joule

4. API Throughput (1-hour Sustained load)

JMeter was used to generate a sustained load against the /eval endpoint for 1 hour with 200 concurrent threads:

Metric	Value
Duration	1 hour

Concurrent Users	200
Total Samples	111,108,790
Total Operations	444.4 trillion
Error Rate	0.00%
Average GPU Power	117.2W

5. Latency Distribution

k6 load test (2-minute sample):

Metric	Value
Average	0.81 ms
Median	0.71 ms
p90	1.23 ms
p95	1.47 ms
Max	9.26 ms

All requests returned HTTP 200. No failed checks observed.

6. Determinism Validation

To verify deterministic execution, the same workload was executed 5 consecutive times in both GPU and CPU modes. SHA-256 hashes were computed on the output vectors.

GPU Mode (5 runs):

Run 1	98bd97026a738671ec7c3d302efa6aa8ff078a5fb9183f7fdf51a1c4ff938321
Run 2	98bd97026a738671ec7c3d302efa6aa8ff078a5fb9183f7fdf51a1c4ff938321
Run 3	98bd97026a738671ec7c3d302efa6aa8ff078a5fb9183f7fdf51a1c4ff938321
Run 4	98bd97026a738671ec7c3d302efa6aa8ff078a5fb9183f7fdf51a1c4ff938321
Run 5	98bd97026a738671ec7c3d302efa6aa8ff078a5fb9183f7fdf51a1c4ff938321

CPU Mode (5 runs):

Run 1	98bd97026a738671ec7c3d302efa6aa8ff078a5fb9183f7fdf51a1c4ff938321
Run 2	98bd97026a738671ec7c3d302efa6aa8ff078a5fb9183f7fdf51a1c4ff938321
Run 3	98bd97026a738671ec7c3d302efa6aa8ff078a5fb9183f7fdf51a1c4ff938321
Run 4	98bd97026a738671ec7c3d302efa6aa8ff078a5fb9183f7fdf51a1c4ff938321
Run 5	98bd97026a738671ec7c3d302efa6aa8ff078a5fb9183f7fdf51a1c4ff938321

Result: All 10 hashes are identical. Deterministic execution confirmed for both GPU and CPU modes. GPU and CPU outputs are bit-exact for this workload.

7. Stability Assessment

Metric	Value
Test Duration	1 hour
Concurrent Users	200
Total Requests	111,108,790

Failed Requests	0
Error Rate	0.00%
System Crashes	0
Memory Leaks Observed	None

The system maintained stable performance throughout the test with no degradation, errors, or anomalies.

8. Conclusion

1. LuxiEdge achieved 286.94 billion operations/second aggregate throughput across 7 transcendental and algebraic functions on NVIDIA H100 SXM.
2. Peak throughput of 331.13 billion ops/sec was observed for sqrt operations.
3. The system processed 444.4 trillion operations over 1 hour with zero errors.
4. Deterministic execution was verified via SHA-256 hash matching across 10 runs (5 GPU, 5 CPU).
5. GPU and CPU modes produced identical output hashes, indicating bit-exact cross-mode parity.
6. Average GPU power consumption was 117.2W under sustained load.
7. p95 API latency was 1.47ms under 200 concurrent users.

Appendix

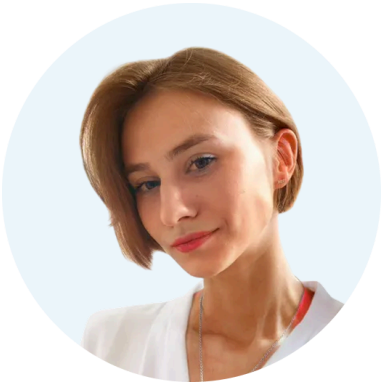
A. Raw Logs:
https://drive.google.com/file/d/1XH4RYoCzMsjHxeJ8-p5oOo8uMb3iO_Hb/view?usp=sharing
B. Test Commands:

- GPU mode: RUST_LOG=info ./target/release/l4_benchmark --listen 0.0.0.0:9090 --mode gpu --vector-size 1000000 --sha256-proof true
- CPU mode: RUST_LOG=info ./target/release/l4_benchmark --listen 0.0.0.0:9090 --mode cpu --vector-size 1000000 --sha256-proof true

C. Load Test Configuration:

- JMeter: 200 threads, 1 hour duration
- k6: 200 VUs, 2 minute sample for latency distribution

Contacts



Nora Laievska

Director of Partnerships & Growth

laievska@qarea.us

Tel: +1 310 388 93 34 USA, UK

testfort.com