LangChain

Agent Infrastructure Benchmark

Nirvana ABS vs AWS EBS (gp3 & io2) for AI Agent Workloads

We ran identical LangChain agent workloads across five storage platforms to see which finishes real tasks fastest, and at what cost.

ABS LEAD

14-17%

faster task completion time vs AWS io2

THE WORKLOAD

5M-vector database

Qdrant · Redis · Postgres

Vector search · cache · checkpoints · cold reads

6 storage ops per task

REPRODUCIBILITY

3× cold runs

CONCURRENCY

1,000 agents × **100** tasks

100,000 tasks total

+19%

More throughput

Sustained app IOPS at 100k scale: 172 vs 145.

7.8×

Faster raw disk (fio)

313K IOPS vs io2-64k's 64K provisioned, instance-capped at 40K.

31×

Cheaper than io2

$118/mo vs $3,710/mo (io2-64k).

THE SETUP · SAME WORKLOAD + HARDWARE, FIVE STORAGE TIERS

CONFIG	INSTANCE	VCPU / RAM	STORAGE	PROVISIONED IOPS	TOTAL COST / MO
Nirvana ABS	n1-standard-4	4 / 16 GB · DDR5	ABS 256 GB	20,000 baseline; 600,000 burst, included	$118
gp3-3k	m6i.xlarge	4 / 16 GB · DDR4	gp3 256 GB	3,000	$147
gp3-16k	m6i.xlarge	4 / 16 GB · DDR4	gp3 256 GB	16,000	$212
io2-32k	m6i.xlarge	4 / 16 GB · DDR4	io2 256 GB	32,000	$2,238
io2-64k	m6i.xlarge	4 / 16 GB · DDR4	io2 256 GB	64,000	$3,710

- **Held constant:** instance class, 4 vCPU / 16 GB / 256 GB, same agent code and data on all five.
- **The only variable:** the block storage. ABS includes 20,000 sustained IOPS (600,000 burst), nothing to provision; io2 is billed per provisioned IOPS, driving the $212 → $3,710/mo jump.

NIRVANA LABS

RESULTS // 1000 AGENTS × 100 TASKS × 3 RUNS

TASK COMPLETION AT SCALE · ABS WINS AT EVERY SCALE, LEAD WIDENS WITH LOAD

SCALE	TOTAL TASKS	NIRVANA ABS	IO2-32K	IO2-64K	GP3-16K	GP3-3K	ABS LEAD
100×10	1,000	36s	43s	42s	43s	62s	14-16%
500×20	10,000	351s	420s	424s	430s	433s	16-17%
1000×100	100,000	58 min	69 min	69 min	68 min	71 min	16%

TASK LATENCY DISTRIBUTION AT 100K SCALE · MILLISECONDS

PLATFORM	TASK P50	TASK P95	TASK P99
Nirvana ABS	326	562	725
io2-32k	393	612	779
io2-64k	396	606	771
gp3-16k	390	605	760
gp3-3k	404	643	809

PER-SERVICE BREAKDOWN AT 1000×100 · 100K TASKS · VS IO2-64K

SERVICE	PER TASK	ABS P99	IO2-64K P99	ABS VS IO2
Qdrant · vector search	2 ops	301 ms	140 ms	io2 leads
Redis · cache	2 ops	68.0 ms	70.4 ms	ABS 3% faster
Postgres · checkpoints	2 ops	31.9 ms	37.7 ms	ABS 15% faster
Task p99 (compound)	6 ops	725 ms	771 ms	**ABS 6% faster**
Task completion	100K tasks	58 min	69 min	**ABS 16% faster**

THE VERDICT · ABS VS AWS EBS

FASTER

14-17%

faster than AWS io2 end-to-end on real agent tasks, not synthetic reads.

UNDER LOAD

Holds from **1k to 100k tasks**, widening as you scale. **ABS loves heavy.**

PRICE

$118/mo

31× cheaper than $3,710 (io2-64k).

OPEN SOURCE · RUN IT YOURSELF

github.com/nirvana-labs-examples/langchain-benchmarks

Task completion matters.
Your agents. Faster.