Prefill Workload — Input=1024, Output=1024 (Best Configurations Only)
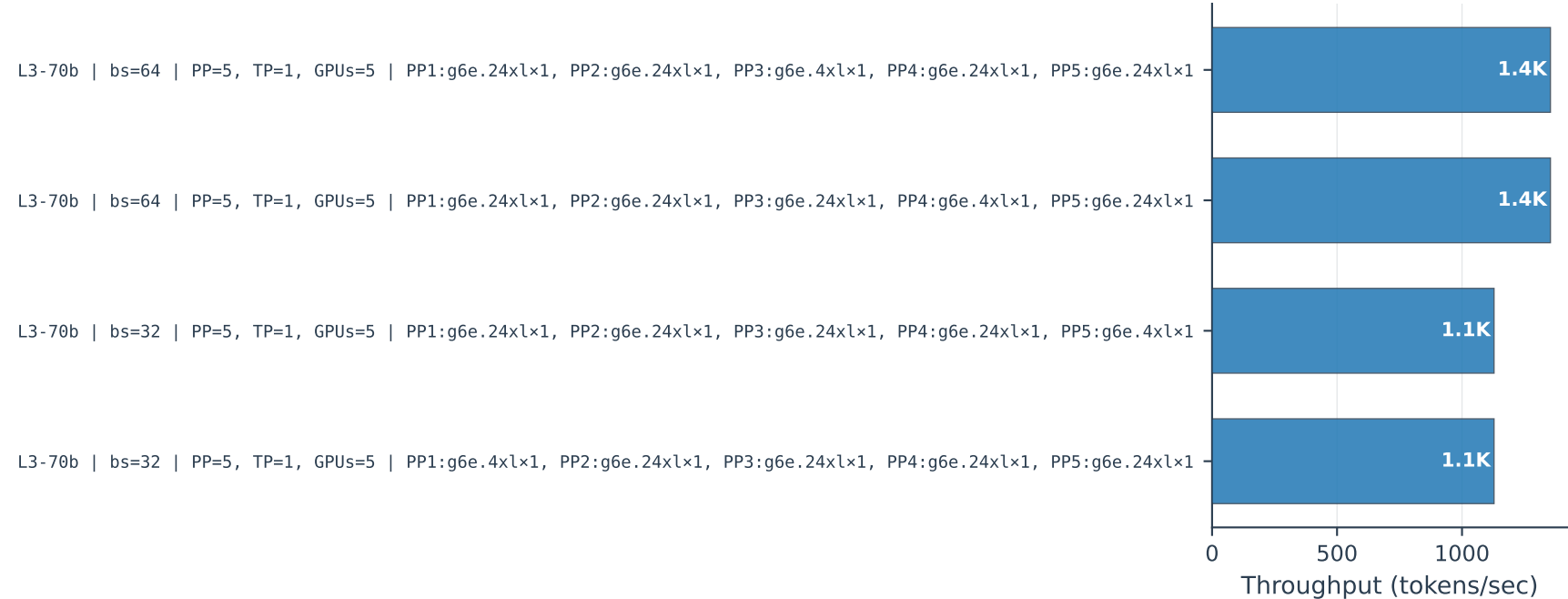
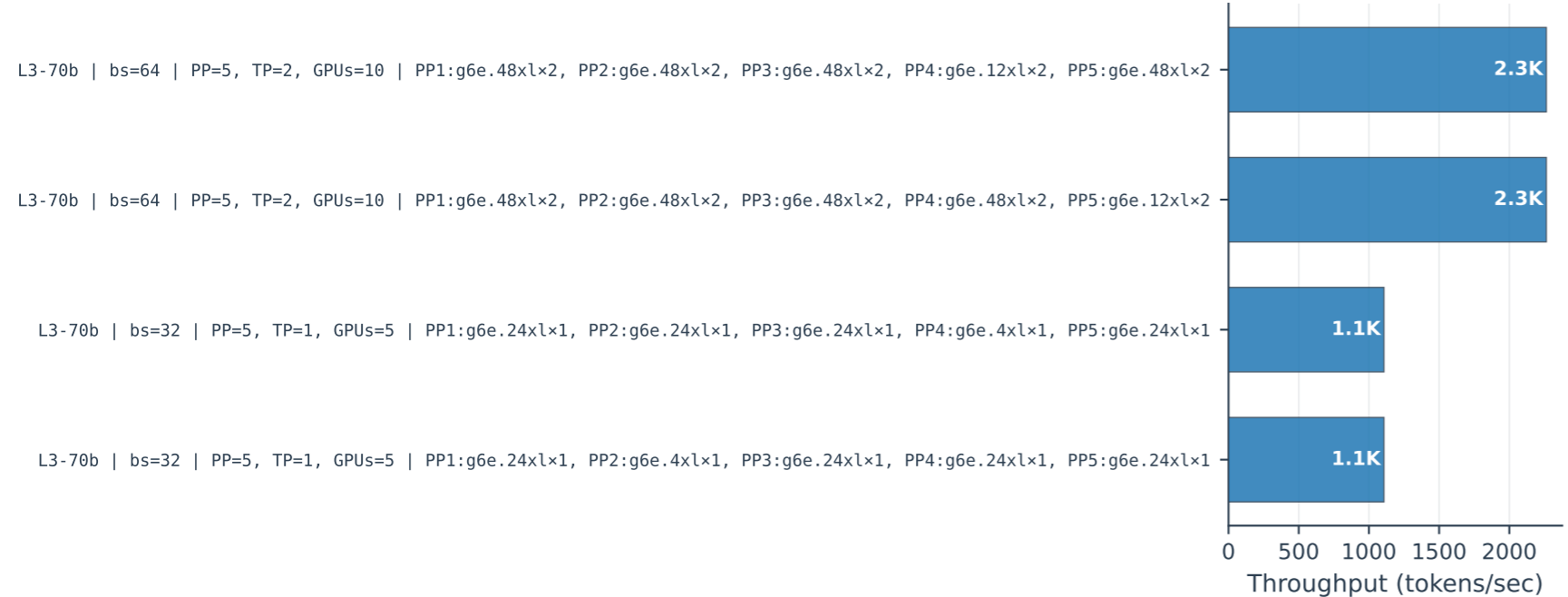## Top 4 by Throughput

Configuration	Throughput
L3-70b | bs=64 | PP=5, TP=1, GPUs=5 | PP1:g6e.24xl×1, PP2:g6e.24xl×1, PP3:g6e.4xl×1, PP4:g6e.24xl×1, PP5:g6e.24xl×1	1.4K
L3-70b | bs=64 | PP=5, TP=1, GPUs=5 | PP1:g6e.24xl×1, PP2:g6e.24xl×1, PP3:g6e.24xl×1, PP4:g6e.4xl×1, PP5:g6e.24xl×1	1.4K
L3-70b | bs=32 | PP=5, TP=1, GPUs=5 | PP1:g6e.24xl×1, PP2:g6e.24xl×1, PP3:g6e.24xl×1, PP4:g6e.24xl×1, PP5:g6e.4xl×1	1.1K
L3-70b | bs=32 | PP=5, TP=1, GPUs=5 | PP1:g6e.4xl×1, PP2:g6e.24xl×1, PP3:g6e.24xl×1, PP4:g6e.24xl×1, PP5:g6e.24xl×1	1.1K

Throughput (tokens/sec)

Top 4 by Cost Efficiency

Configuration	Cost
L3-70b | bs=64 | PP=5, TP=1, GPUs=5 | PP1:g6e.24xl×1, PP2:g6e.24xl×1, PP3:g6e.4xl×1, PP4:g6e.24xl×1, PP5:g6e.24xl×1	$3.51
L3-70b | bs=64 | PP=5, TP=1, GPUs=5 | PP1:g6e.24xl×1, PP2:g6e.24xl×1, PP3:g6e.24xl×1, PP4:g6e.4xl×1, PP5:g6e.24xl×1	$3.51
L3-70b | bs=32 | PP=5, TP=1, GPUs=5 | PP1:g6e.24xl×1, PP2:g6e.24xl×1, PP3:g6e.24xl×1, PP4:g6e.24xl×1, PP5:g6e.4xl×1	$4.22
L3-70b | bs=32 | PP=5, TP=1, GPUs=5 | PP1:g6e.4xl×1, PP2:g6e.24xl×1, PP3:g6e.24xl×1, PP4:g6e.24xl×1, PP5:g6e.24xl×1	$4.22

Cost ($/million tokens)

Prefill Workload — Input=2048, Output=2048 (Best Configurations Only)
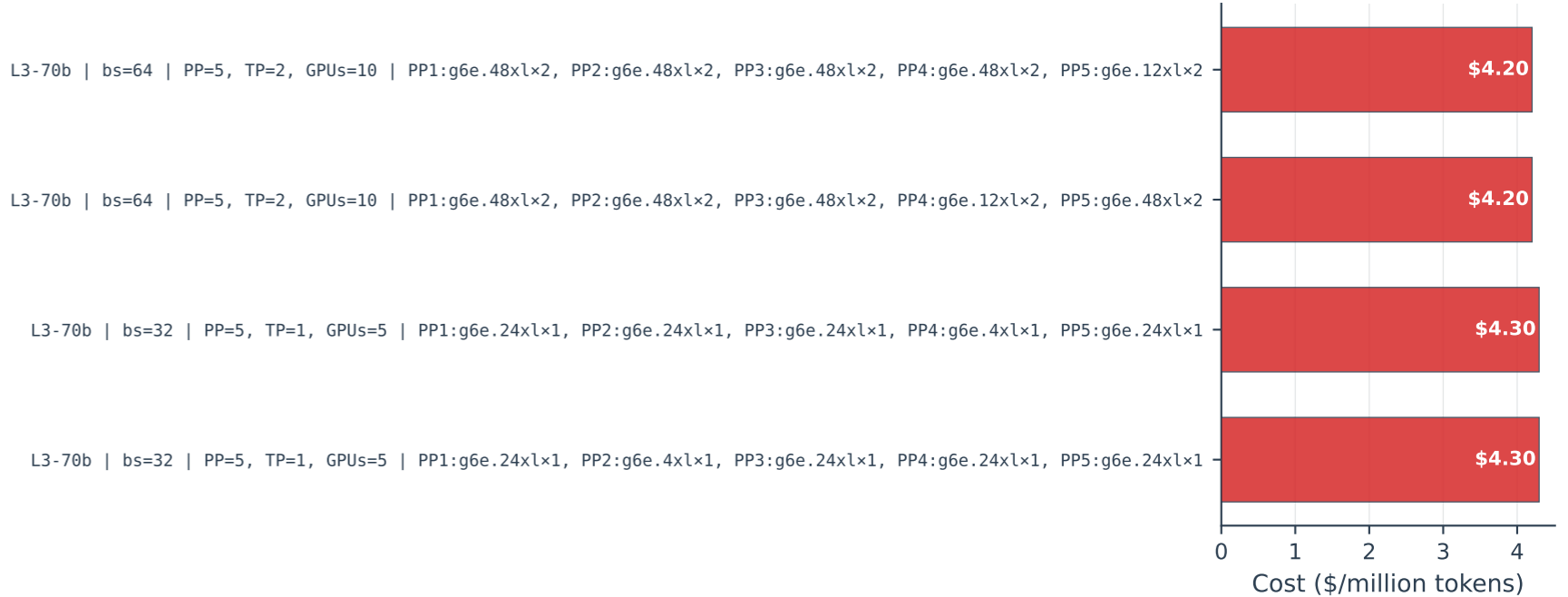
Top 4 by Throughput

Configuration	Throughput
L3-70b | bs=64 | PP=5, TP=2, GPUs=10 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g6e.12xl×2, PP5:g6e.48xl×2	2.3K
L3-70b | bs=64 | PP=5, TP=2, GPUs=10 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g6e.48xl×2, PP5:g6e.12xl×2	2.3K
L3-70b | bs=32 | PP=5, TP=1, GPUs=5 | PP1:g6e.24xl×1, PP2:g6e.24xl×1, PP3:g6e.24xl×1, PP4:g6e.4xl×1, PP5:g6e.24xl×1	1.1K
L3-70b | bs=32 | PP=5, TP=1, GPUs=5 | PP1:g6e.24xl×1, PP2:g6e.4xl×1, PP3:g6e.24xl×1, PP4:g6e.24xl×1, PP5:g6e.24xl×1	1.1K

Top 4 by Cost Efficiency

Configuration	Cost ($/million tokens)
L3-70b | bs=64 | PP=5, TP=2, GPUs=10 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g6e.48xl×2, PP5:g6e.12xl×2	$4.20
L3-70b | bs=64 | PP=5, TP=2, GPUs=10 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g6e.12xl×2, PP5:g6e.48xl×2	$4.20
L3-70b | bs=32 | PP=5, TP=1, GPUs=5 | PP1:g6e.24xl×1, PP2:g6e.24xl×1, PP3:g6e.24xl×1, PP4:g6e.4xl×1, PP5:g6e.24xl×1	$4.30
L3-70b | bs=32 | PP=5, TP=1, GPUs=5 | PP1:g6e.24xl×1, PP2:g6e.4xl×1, PP3:g6e.24xl×1, PP4:g6e.24xl×1, PP5:g6e.24xl×1	$4.30

Prefill Workload — Input=4096, Output=4096 (Best Configurations Only)

Top 4 by Throughput

Configuration	Throughput
L3-70b | bs=64 | PP=5, TP=2, GPUs=10 | PP1:g6e.12xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g6e.48xl×2, PP5:g6e.48xl×2	2.2K
L3-70b | bs=64 | PP=5, TP=2, GPUs=10 | PP1:g6e.24xl×2, PP2:g6e.24xl×2, PP3:g6e.24xl×2, PP5:g6e.12xl×2	2.2K
L3-70b | bs=32 | PP=5, TP=2, GPUs=10 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g6e.12xl×2, PP5:g6e.48xl×2	1.8K
L3-70b | bs=32 | PP=3, TP=2, GPUs=6 | PP1:g6e.12xl×2, PP2:g6e.24xl×2, PP3:g6e.24xl×2	1.1K

Throughput (tokens/sec)

Top 4 by Cost Efficiency

Configuration	Cost
L3-70b | bs=64 | PP=5, TP=2, GPUs=10 | PP1:g6e.12xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g6e.48xl×2, PP5:g6e.48xl×2	$4.34
L3-70b | bs=64 | PP=5, TP=2, GPUs=10 | PP1:g6e.24xl×2, PP2:g6e.24xl×2, PP3:g6e.24xl×2, PP4:g6e.24xl×2, PP5:g6e.12xl×2	$4.35
L3-70b | bs=32 | PP=5, TP=2, GPUs=10 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g6e.12xl×2, PP5:g6e.48xl×2	$5.21
L3-70b | bs=32 | PP=3, TP=2, GPUs=6 | PP1:g6e.12xl×2, PP2:g6e.24xl×2, PP3:g6e.24xl×2	$5.21

Cost ($/million tokens)

Prefill Workload — Input=8192, Output=8192 (Best Configurations Only)

Top 3 by Throughput

Configuration	Throughput
L3-70b | bs=64 | PP=5, TP=4, GPUs=20 | PP1:g6e.48xl×4, PP2:g6e.48xl×4, PP3:g6e.48xl×4, PP4:g6e.48xl×4, PP5:g6e.24xl×4	3.3K
L3-70b | bs=64 | PP=5, TP=4, GPUs=20 | PP1:g6e.24xl×4, PP2:g6e.48xl×4, PP3:g6e.48xl×4, PP4:g6e.48xl×4, PP5:g6e.48xl×4	3.3K
L3-70b | bs=32 | PP=5, TP=2, GPUs=10 | PP1:g6e.48xl×2, PP2:g6e.12xl×2, PP3:g6e.48xl×2, PP4:g6e.48xl×2, PP5:g6e.48xl×2	1.7K

Top 3 by Cost Efficiency

Configuration	Cost ($/million tokens)
L3-70b | bs=32 | PP=5, TP=2, GPUs=10 | PP1:g6e.48xl×2, PP2:g6e.12xl×2, PP3:g6e.48xl×2, PP4:g6e.48xl×2, PP5:g6e.48xl×2	$5.55
L3-70b | bs=64 | PP=5, TP=4, GPUs=20 | PP1:g6e.48xl×4, PP2:g6e.48xl×4, PP3:g6e.48xl×4, PP4:g6e.48xl×4, PP5:g6e.24xl×4	$5.73
L3-70b | bs=64 | PP=5, TP=4, GPUs=20 | PP1:g6e.24xl×4, PP2:g6e.48xl×4, PP3:g6e.48xl×4, PP4:g6e.48xl×4, PP5:g6e.48xl×4	$5.73

Throughput (tokens/sec)

Cost ($/million tokens)

Prefill Workload — Input=16384, Output=16384 (Best Configurations Only)

Top 2 by Throughput

Configuration	Throughput
L3-70b | bs=64 | PP=5, TP=4, GPUs=20 | PP1:g6e.48xl×4, PP2:g6e.48xl×4, PP3:g6e.24xl×4, PP4:g6e.48xl×4, PP5:g6e.48xl×4	3.0K
L3-70b | bs=32 | PP=5, TP=4, GPUs=20 | PP1:g6e.48xl×4, PP2:g6e.48xl×4, PP3:g6e.24xl×4, PP4:g6e.48xl×4, PP5:g6e.48xl×4	2.5K

Top 2 by Cost Efficiency

Configuration	Cost ($/million tokens)
L3-70b | bs=64 | PP=5, TP=4, GPUs=20 | PP1:g6e.48xl×4, PP2:g6e.48xl×4, PP3:g6e.24xl×4, PP4:g6e.48xl×4, PP5:g6e.48xl×4	$6.33
L3-70b | bs=32 | PP=5, TP=4, GPUs=20 | PP1:g6e.48xl×4, PP2:g6e.48xl×4, PP3:g6e.24xl×4, PP4:g6e.48xl×4, PP5:g6e.48xl×4	$7.59

Decode Workload — Input=1024, Output=1024 (Best Configurations Only)

Top 3 by Throughput

L3-70b | bs=64 | PP=5, TP=1-2, GPUs=9 | PP1:g5.12xl×2, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g6e.4xl×1 — 182

L3-70b | bs=64 | PP=5, TP=1-2, GPUs=9 | PP1:g6e.4xl×1, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g5.12xl×2 — 182

L3-70b | bs=32 | PP=5, TP=1-2, GPUs=9 | PP1:g5.12xl×2, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g6e.4xl×1 — 81

Throughput (tokens/sec)

Top 3 by Cost Efficiency

L3-70b | bs=64 | PP=5, TP=1-2, GPUs=9 | PP1:g5.12xl×2, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g6e.4xl×1 — $22.51

L3-70b | bs=64 | PP=5, TP=1-2, GPUs=9 | PP1:g6e.4xl×1, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g5.12xl×2 — $22.51

L3-70b | bs=32 | PP=5, TP=1-2, GPUs=9 | PP1:g5.12xl×2, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g6e.4xl×1 — $50.35

Cost ($/million tokens)

Decode Workload — Input=2048, Output=2048 (Best Configurations Only)

Top 3 by Throughput

Configuration	Throughput (tokens/sec)
L3-70b | bs=64 | PP=5, TP=1-2, GPUs=9 | PP1:g5.12xl×2, PP2:g5.12xl×2, PP3:g6e.4xl×1, PP4:g5.12xl×2, PP5:g5.12xl×2	160
L3-70b | bs=32 | PP=5, TP=1-2, GPUs=9 | PP1:g5.12xl×2, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g6e.4xl×1	76
L3-70b | bs=32 | PP=5, TP=1-2, GPUs=9 | PP1:g6e.4xl×1, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g5.12xl×2	76

Top 3 by Cost Efficiency

Configuration	Cost ($/million tokens)
L3-70b | bs=64 | PP=5, TP=1-2, GPUs=9 | PP1:g5.12xl×2, PP2:g5.12xl×2, PP3:g6e.4xl×1, PP4:g5.12xl×2, PP5:g5.12xl×2	$25.56
L3-70b | bs=32 | PP=5, TP=1-2, GPUs=9 | PP1:g5.12xl×2, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g6e.4xl×1	$54.01
L3-70b | bs=32 | PP=5, TP=1-2, GPUs=9 | PP1:g6e.4xl×1, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g5.12xl×2	$54.01

Decode Workload — Input=4096, Output=4096 (Best Configurations Only)

Top 3 by Throughput

Configuration	Throughput (tokens/sec)
L3-70b | bs=64 | PP=5, TP=2-4, GPUs=12 | PP1:g5.12xl×4, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g6e.48xl×2, PP5:g6e.48xl×2	243
L3-70b | bs=64 | PP=5, TP=2-4, GPUs=12 | PP1:g6e.48xl×2, PP2:g5.12xl×4, PP3:g6e.48xl×2, PP4:g6e.48xl×2, PP5:g6e.48xl×2	243
L3-70b | bs=32 | PP=5, TP=1-2, GPUs=9 | PP1:g5.12xl×2, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g6e.4xl×1	67

Top 3 by Cost Efficiency

Configuration	Cost ($/million tokens)
L3-70b | bs=64 | PP=5, TP=2-4, GPUs=12 | PP1:g5.12xl×4, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g6e.48xl×2, PP5:g6e.48xl×2	$37.75
L3-70b | bs=64 | PP=5, TP=2-4, GPUs=12 | PP1:g6e.48xl×2, PP2:g5.12xl×4, PP3:g6e.48xl×2, PP4:g6e.48xl×2, PP5:g6e.48xl×2	$37.75
L3-70b | bs=32 | PP=5, TP=1-2, GPUs=9 | PP1:g5.12xl×2, PP2:g5.12xl×2, PP3:g5.12xl×2, PP4:g5.12xl×2, PP5:g6e.4xl×1	$61.33

Decode Workload — Input=8192, Output=8192 (Best Configurations Only)
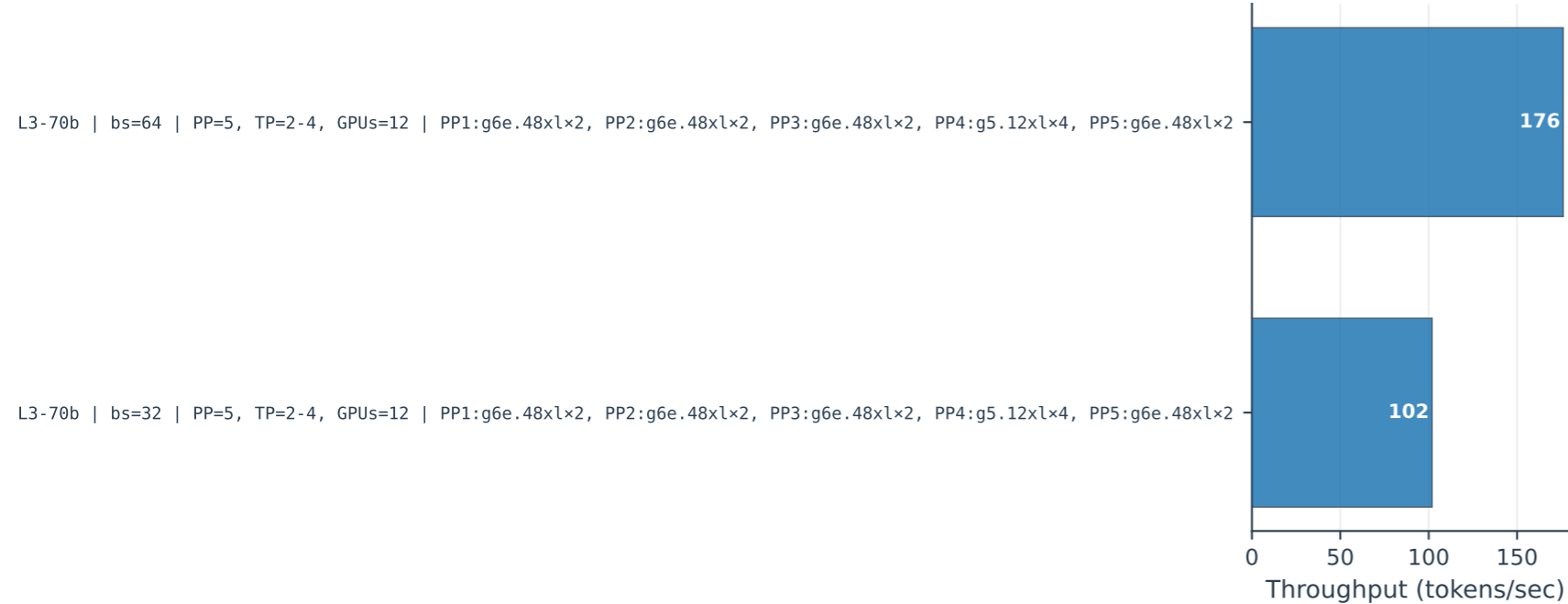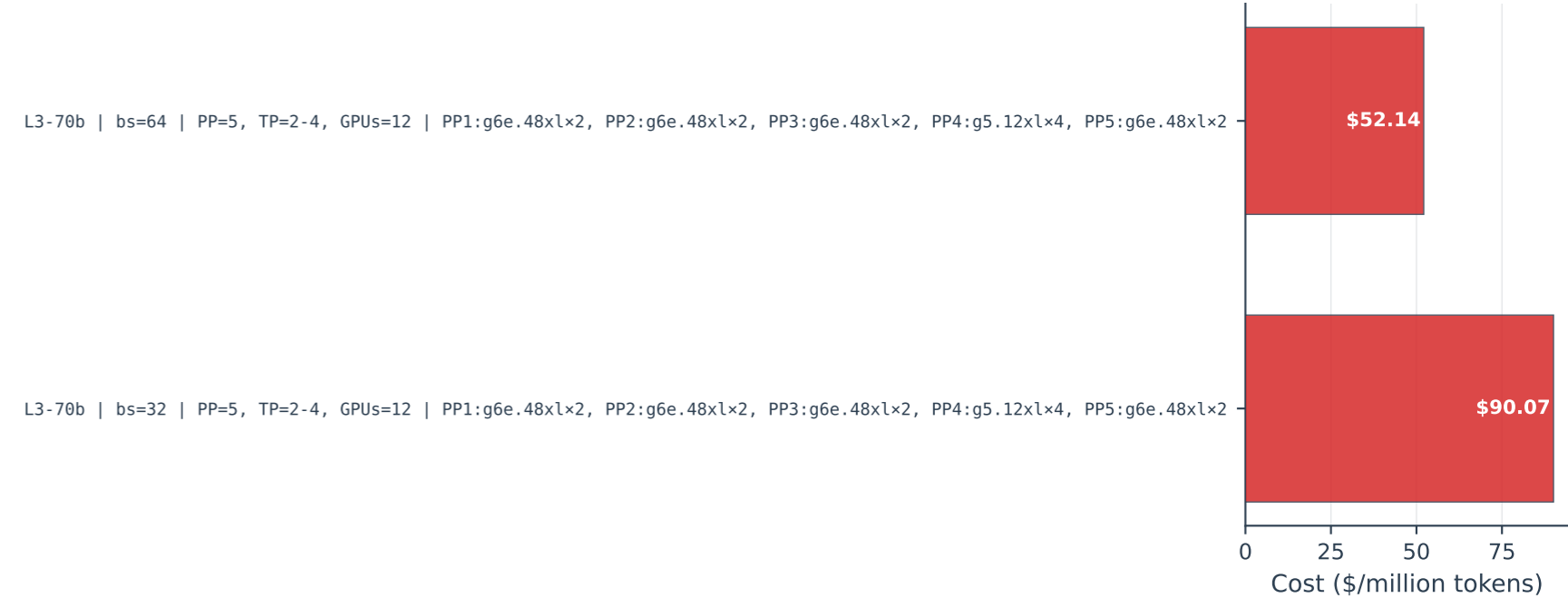
Top 2 by Throughput

Top 2 by Cost Efficiency

L3-70b | bs=64 | PP=5, TP=2-4, GPUs=12 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g5.12xl×4, PP5:g6e.48xl×2 — **176**

L3-70b | bs=32 | PP=5, TP=2-4, GPUs=12 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g5.12xl×4, PP5:g6e.48xl×2 — **102**

L3-70b | bs=64 | PP=5, TP=2-4, GPUs=12 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g5.12xl×4, PP5:g6e.48xl×2 — **$52.14**

L3-70b | bs=32 | PP=5, TP=2-4, GPUs=12 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g5.12xl×4, PP5:g6e.48xl×2 — **$90.07**

Throughput (tokens/sec)

Cost ($/million tokens)

Decode Workload — Input=16384, Output=16384 (Best Configurations Only)

Top 3 by Throughput

Configuration	Throughput (tokens/sec)
L3-70b | bs=64 | PP=2, TP=8, GPUs=16 | PP1:p4de.24xl×8, PP2:g5.48xl×8	288
L3-70b | bs=64 | PP=2, TP=8, GPUs=16 | PP1:g5.48xl×8, PP2:p4de.24xl×8	288
L3-70b | bs=32 | PP=5, TP=2-4, GPUs=12 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g5.12xl×4, PP5:g6e.48xl×2	74

Top 3 by Cost Efficiency

Configuration	Cost ($/million tokens)
L3-70b | bs=64 | PP=2, TP=8, GPUs=16 | PP1:p4de.24xl×8, PP2:g5.48xl×8	$55.22
L3-70b | bs=64 | PP=2, TP=8, GPUs=16 | PP1:g5.48xl×8, PP2:p4de.24xl×8	$55.22
L3-70b | bs=32 | PP=5, TP=2-4, GPUs=12 | PP1:g6e.48xl×2, PP2:g6e.48xl×2, PP3:g6e.48xl×2, PP4:g5.12xl×4, PP5:g6e.48xl×2	$124.6