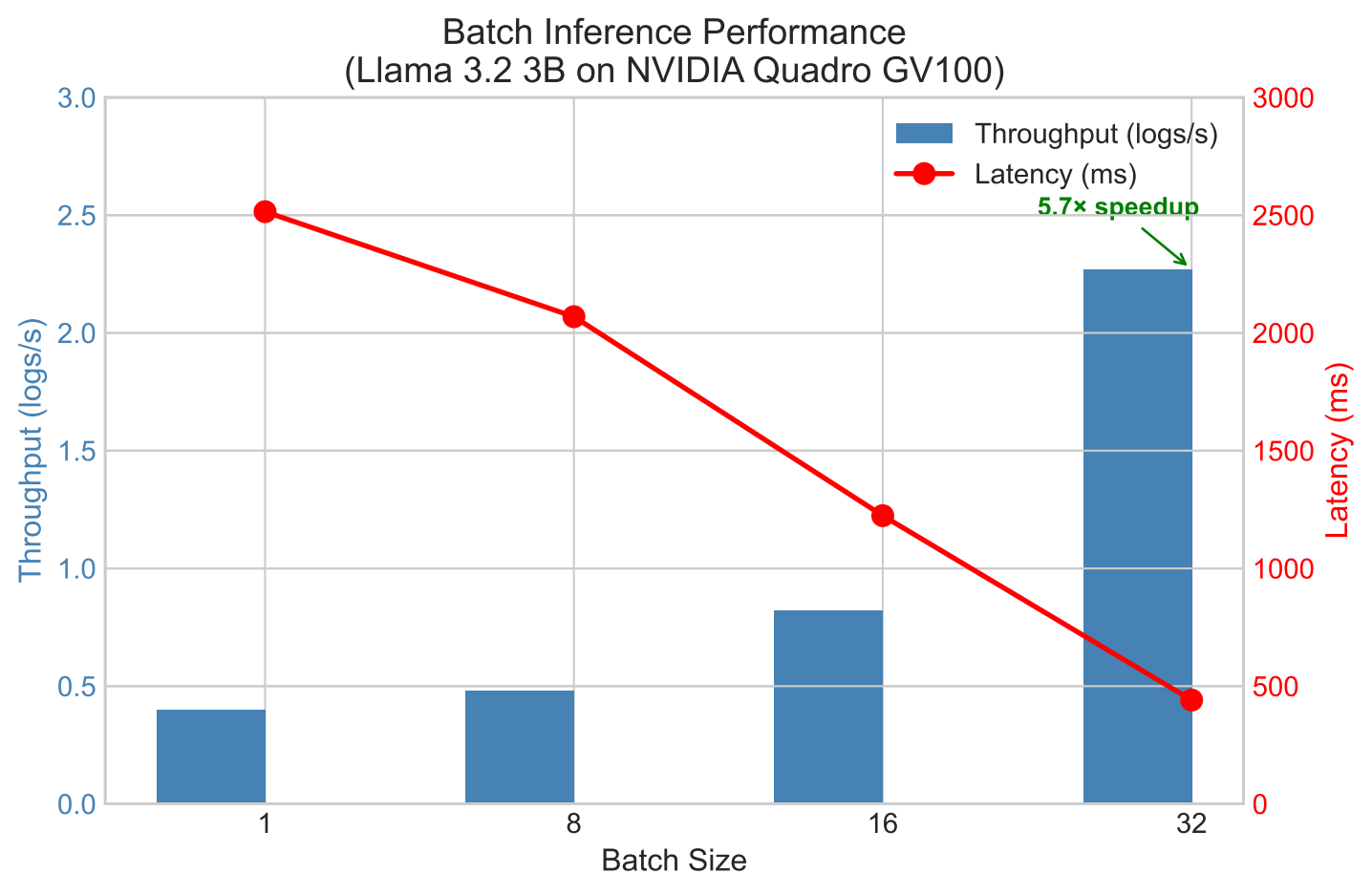
Batch Inference Performance
(Llama 3.2 3B on NVIDIA Quadro GV100)
Throughput (logs/s)
Latency (ms)
5.7× speedup
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3000
2500
2000
1500
1000
500
0
1
8
16
32
Batch Size