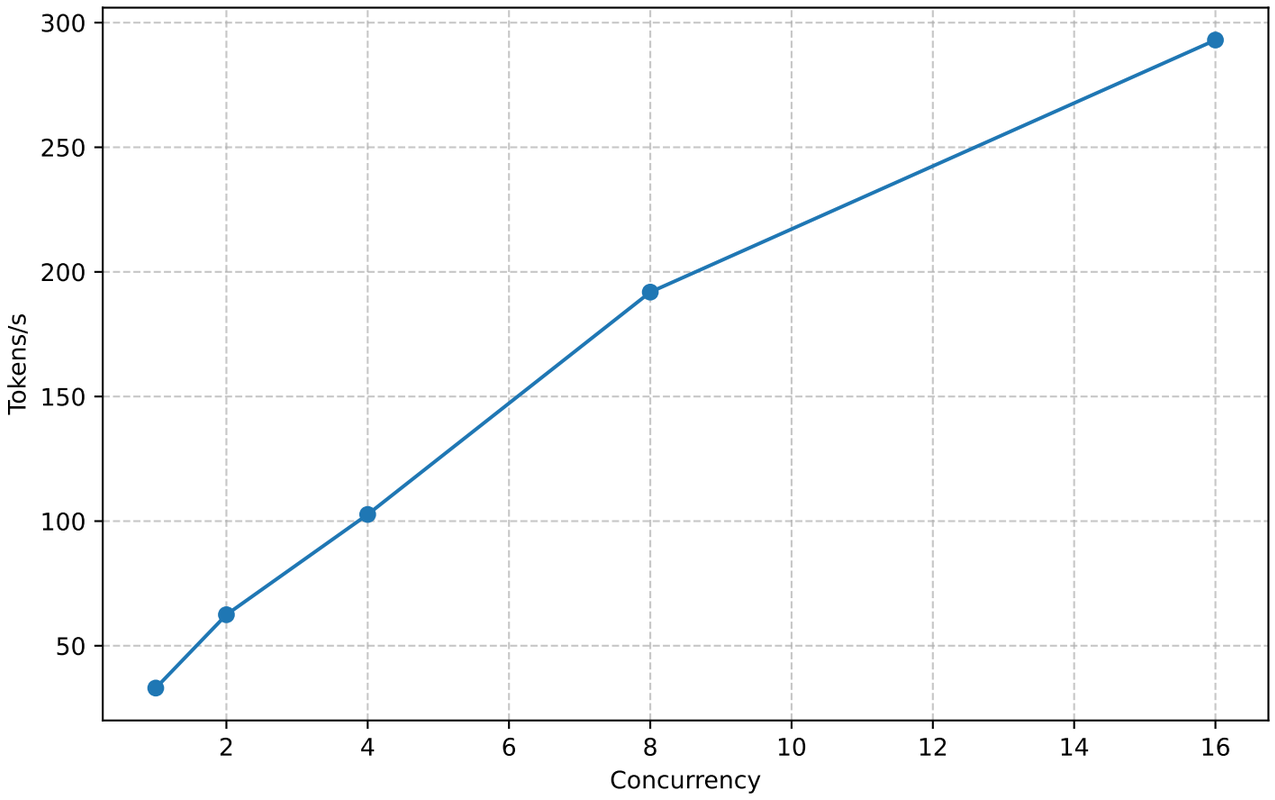
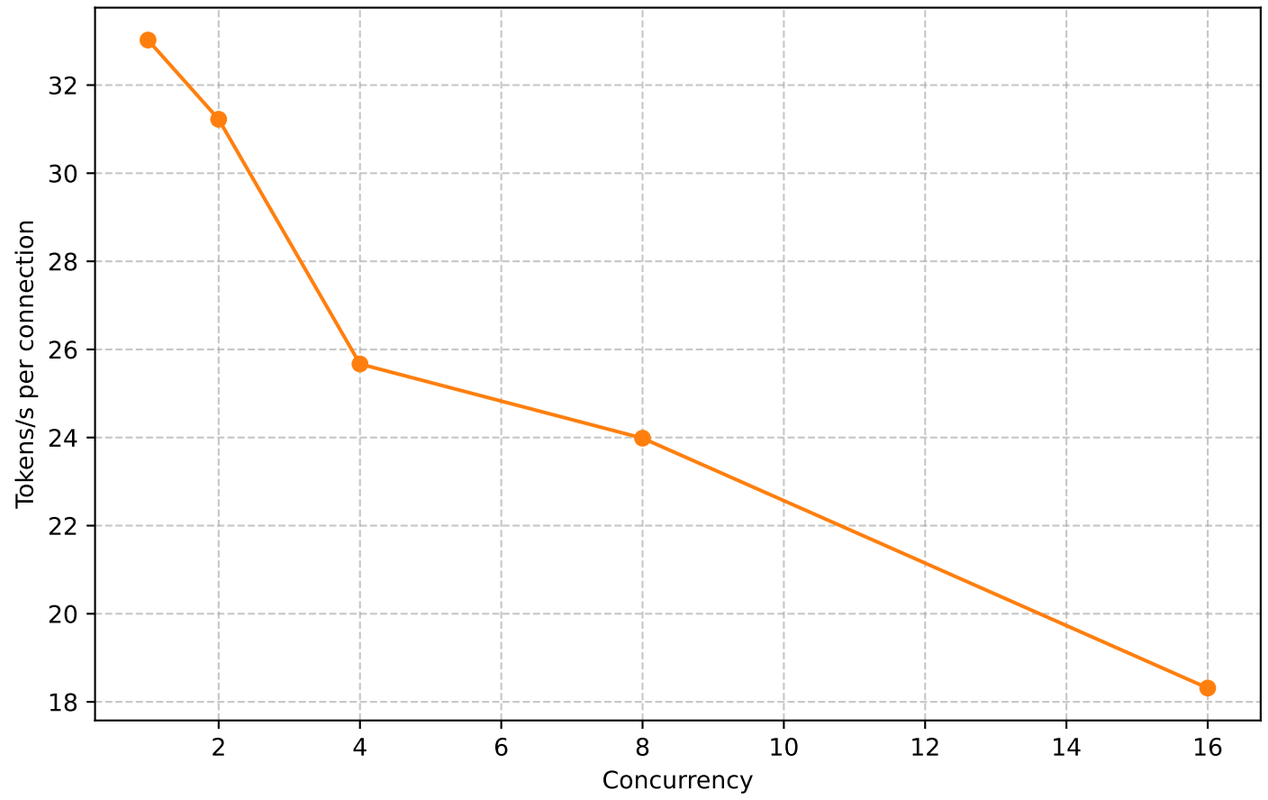
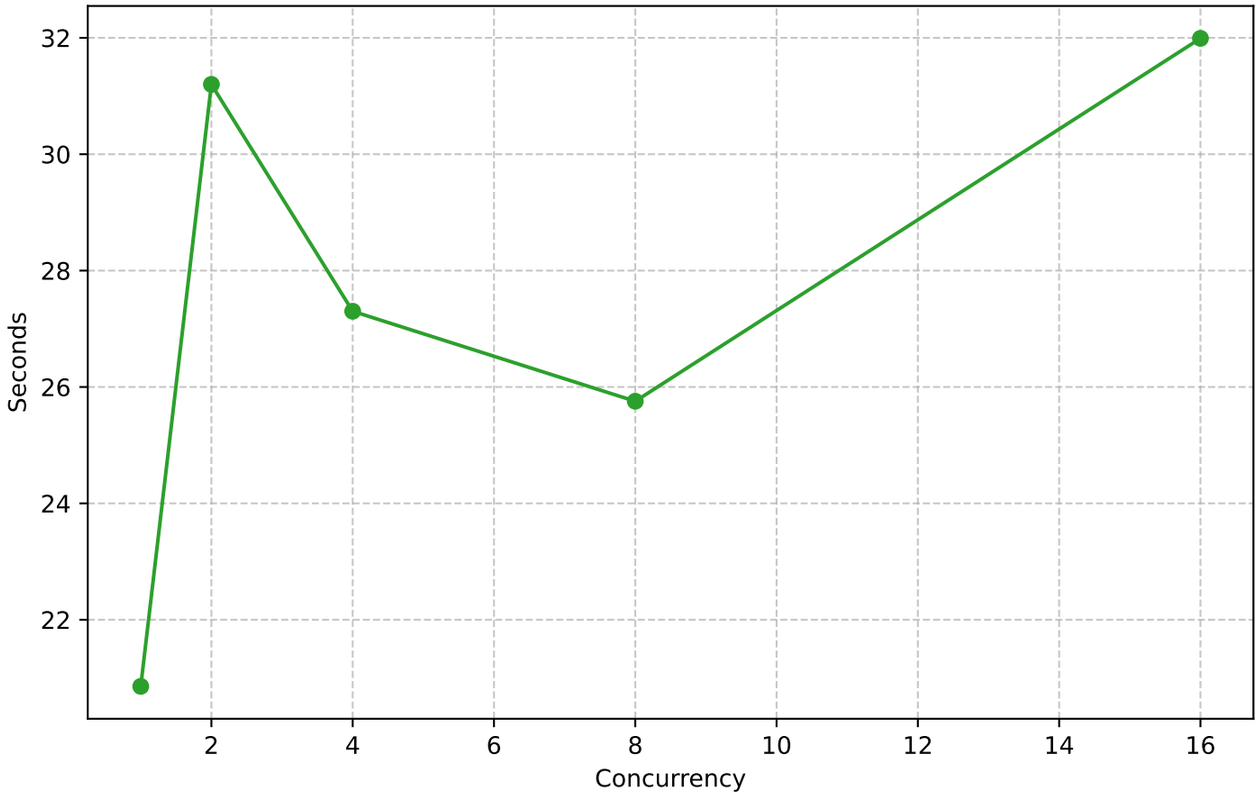
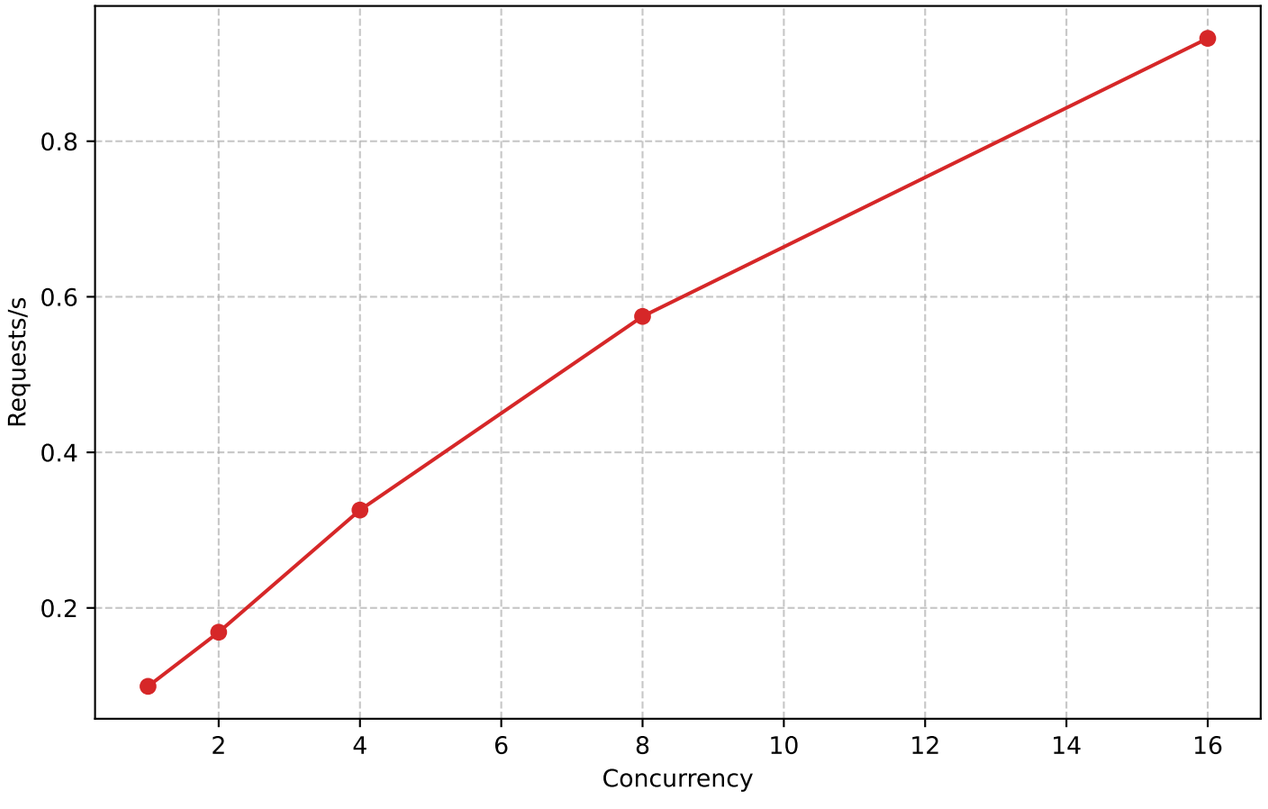

Total Token Throughput vs Concurrency
Tokens/s
Concurrency
300
250
200
150
100
50
2
4
6
8
10
12
14
16


Per-Request Token Throughput vs Concurrency
Tokens/s per connection
Concurrency
32
30
28
26
24
22
20
18
2
4
6
8
10
12
14
16


Average Latency vs Concurrency
Seconds
Concurrency
32
30
28
26
24
22
2
4
6
8
10
12
14
16


Request Throughput vs Concurrency
Requests/s
Concurrency
0.8
0.6
0.4
0.2
2
4
6
8
10
12
14
16