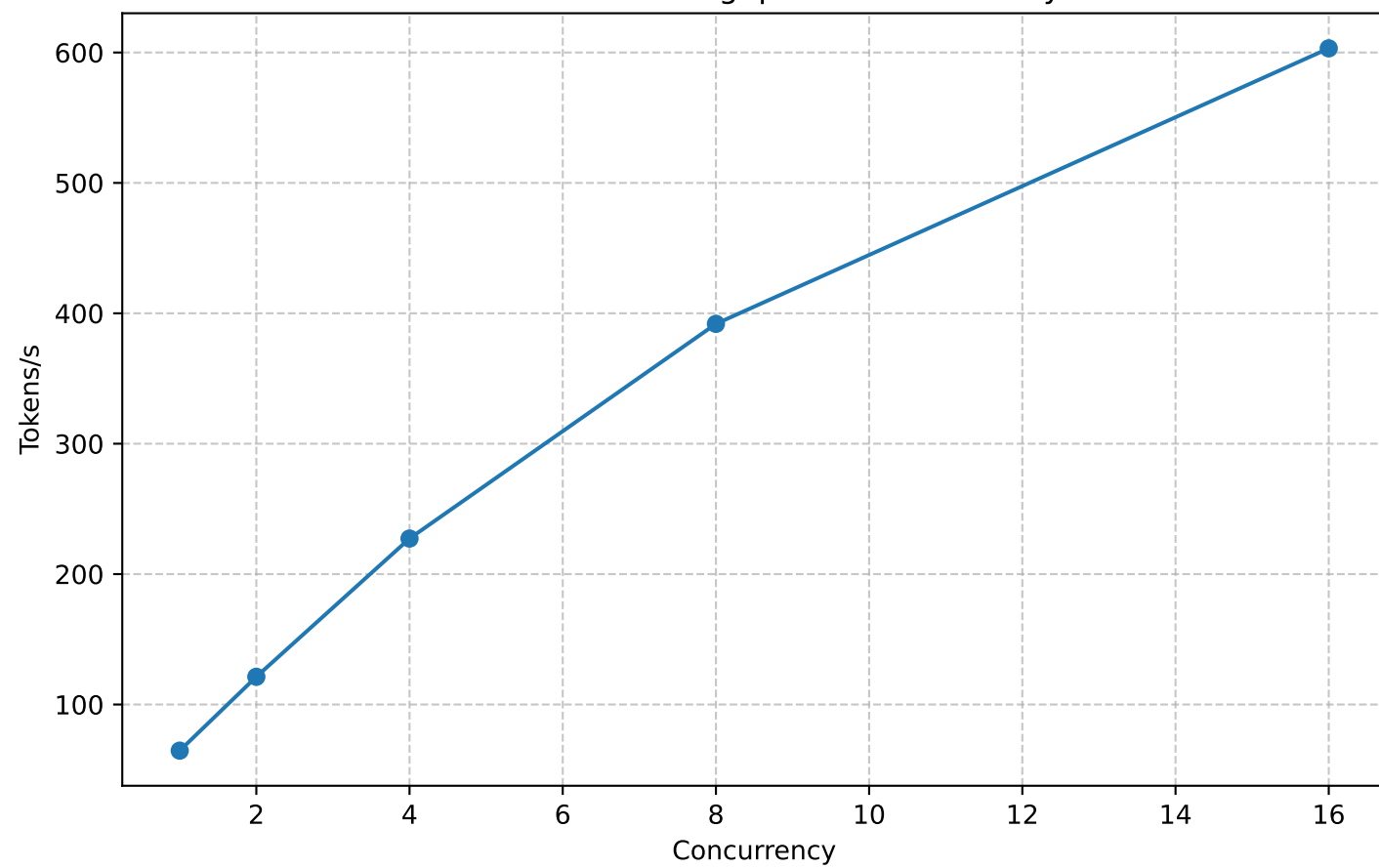
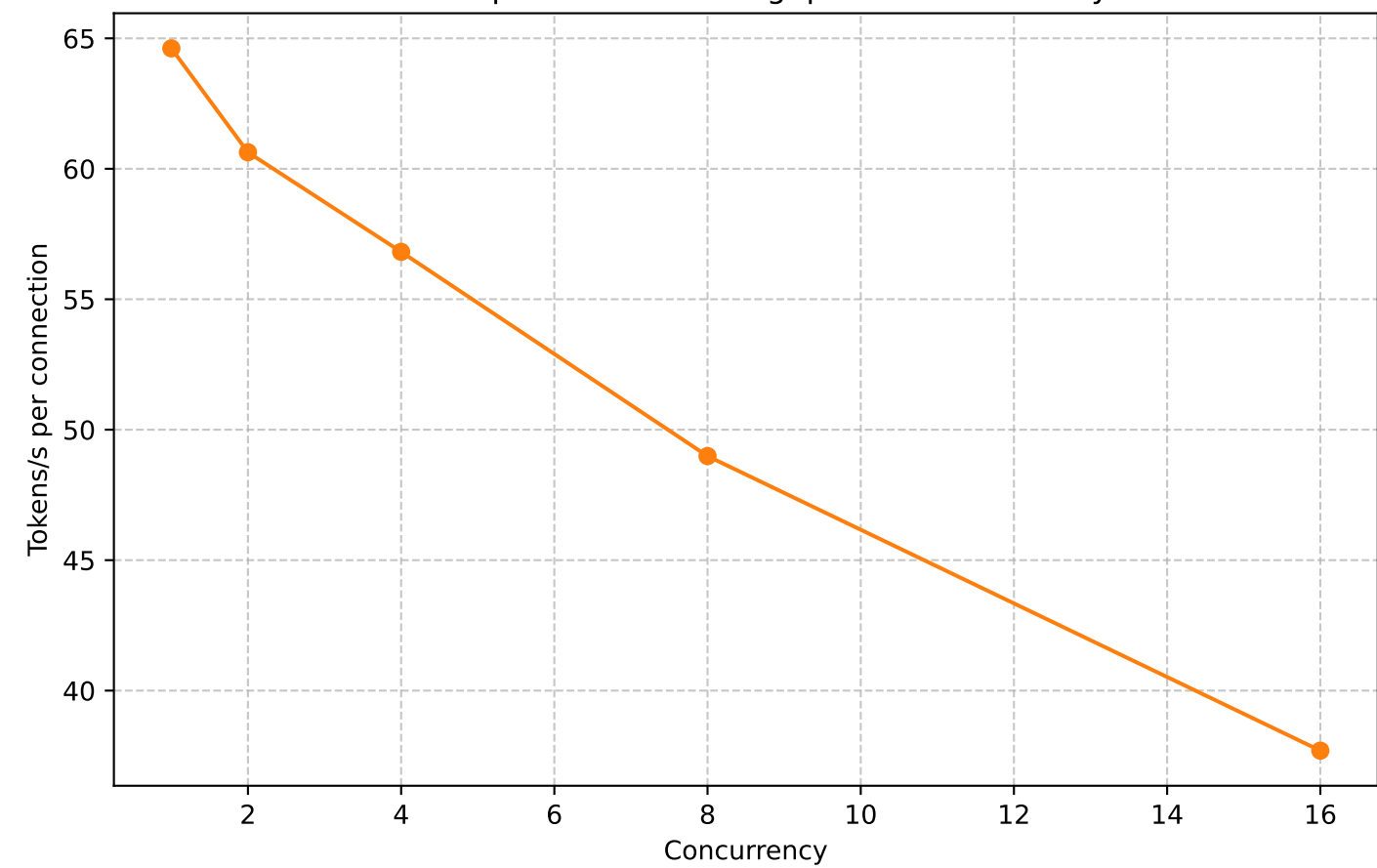


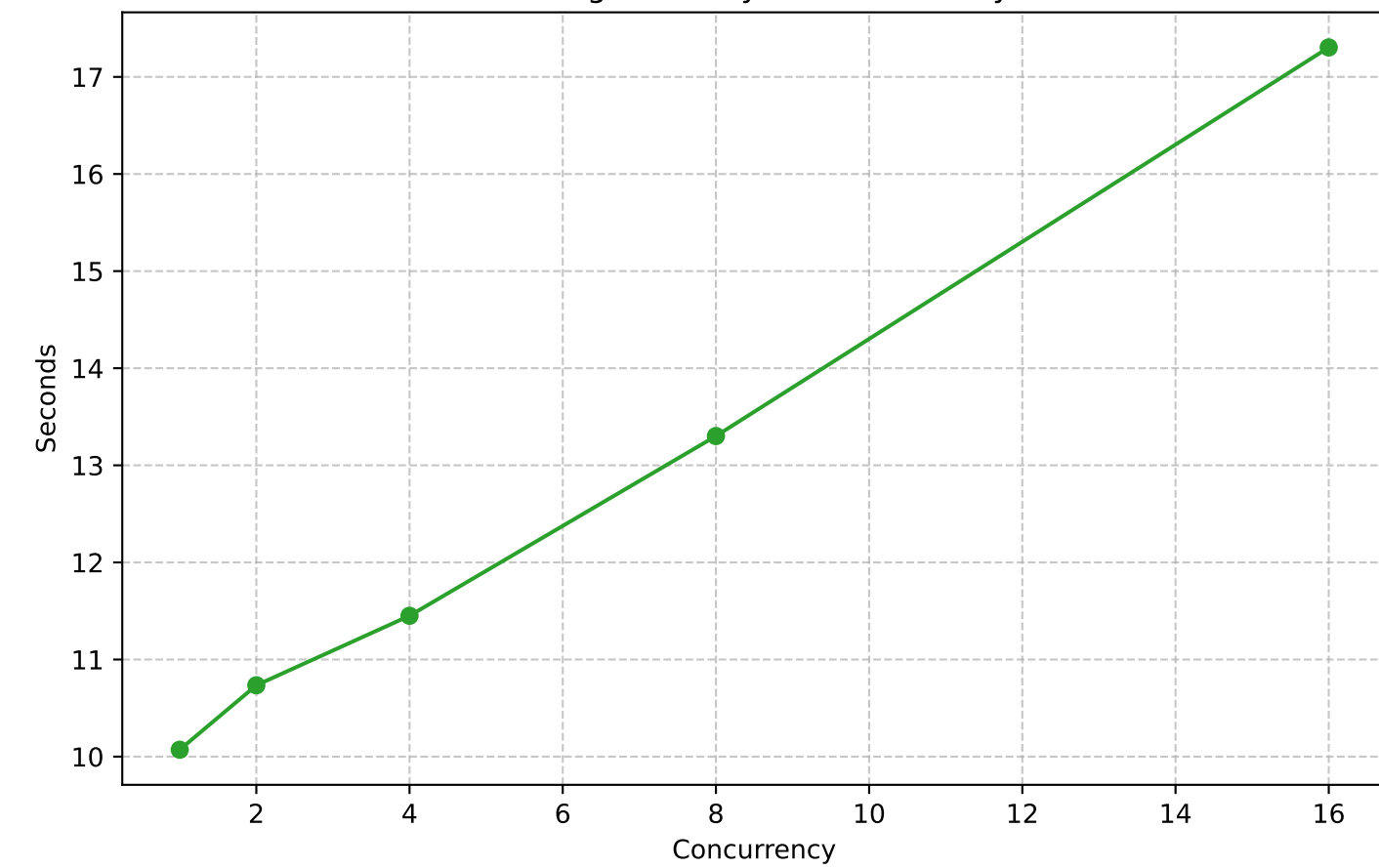
Total Token Throughput vs Concurrency



Per-Request Token Throughput vs Concurrency



Average Latency vs Concurrency



Request Throughput vs Concurrency

