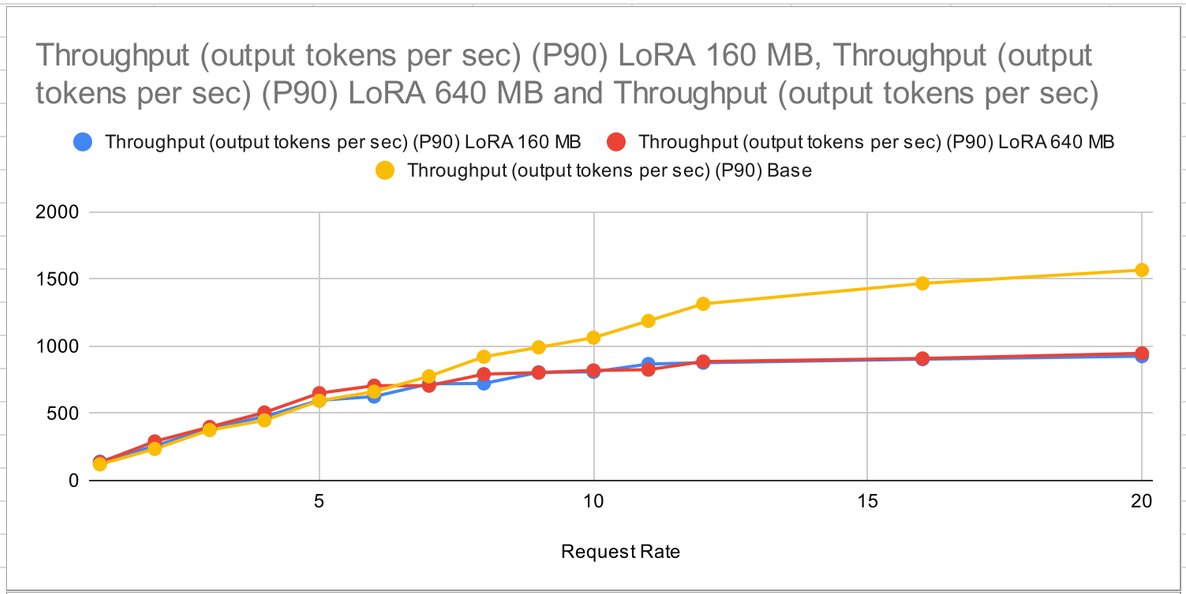
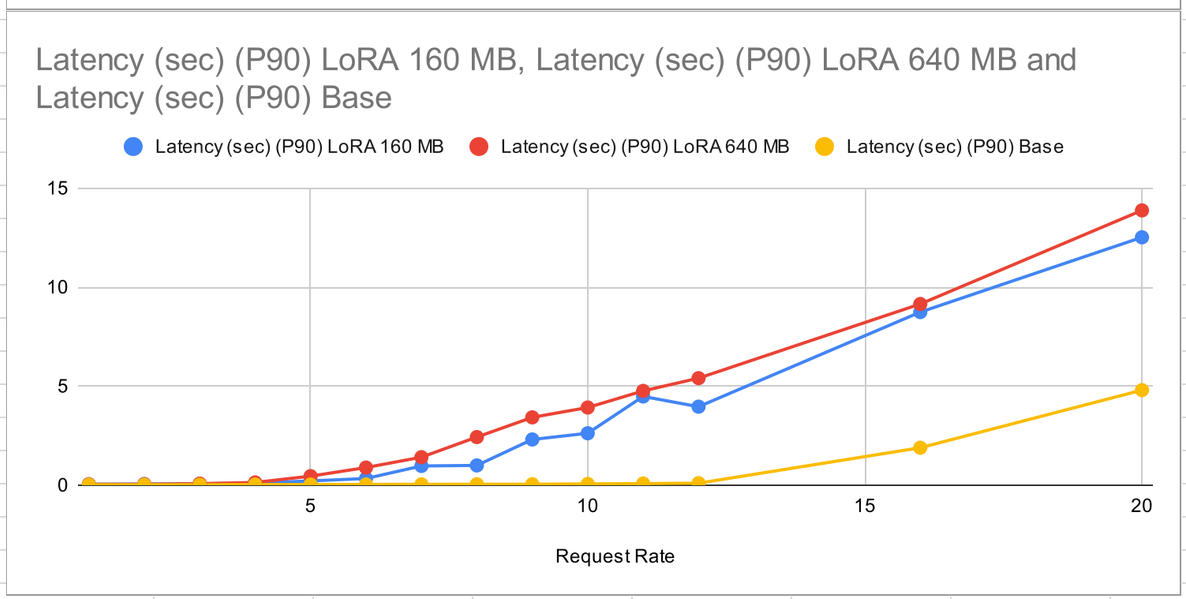

Throughput (output tokens per sec) (P90) LoRA 160 MB, Throughput (output tokens per sec) (P90) LoRA 640 MB and Throughput (output tokens per sec)
Throughput (output tokens per sec) (P90) LoRA 160 MB
Throughput (output tokens per sec) (P90) LoRA 640 MB
Throughput (output tokens per sec) (P90) Base
2000
1500
1000
500
0
5
10
15
20
Request Rate


Latency (sec) (P90) LoRA 160 MB, Latency (sec) (P90) LoRA 640 MB and Latency (sec) (P90) Base
Latency (sec) (P90) LoRA 160 MB
Latency (sec) (P90) LoRA 640 MB
Latency (sec) (P90) Base
15
10
5
0
5
10
15
20
Request Rate