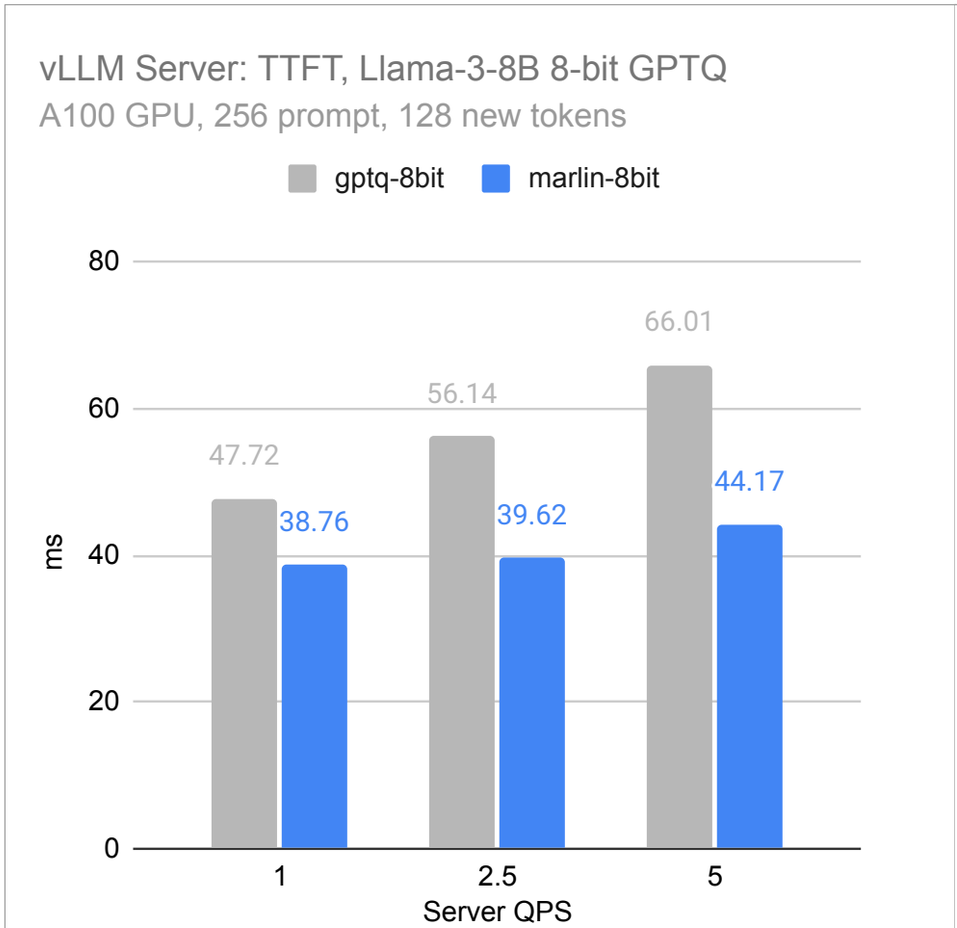
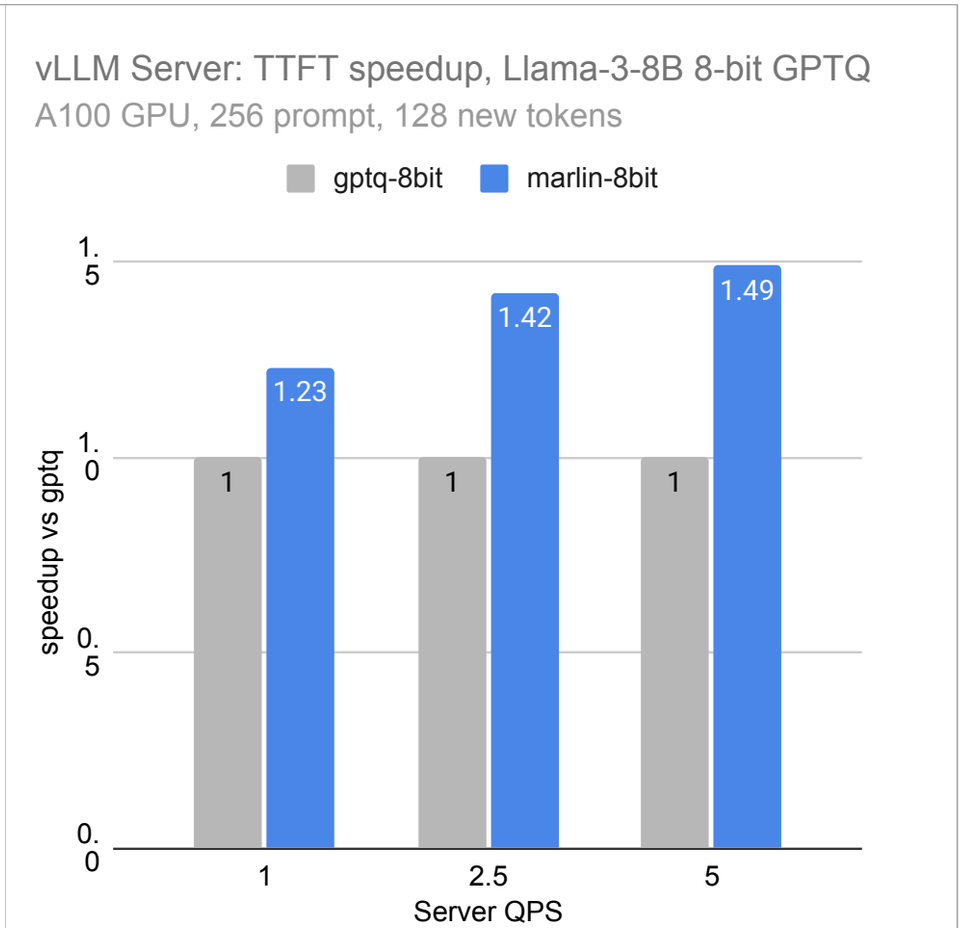
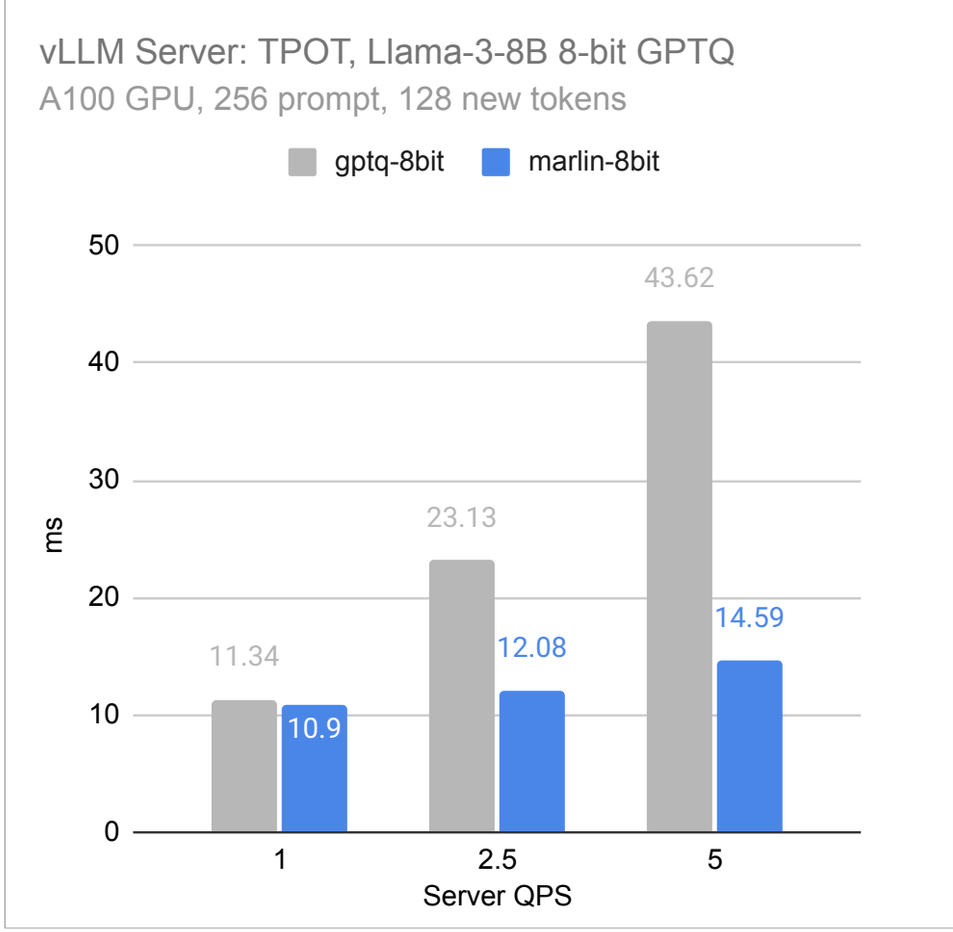
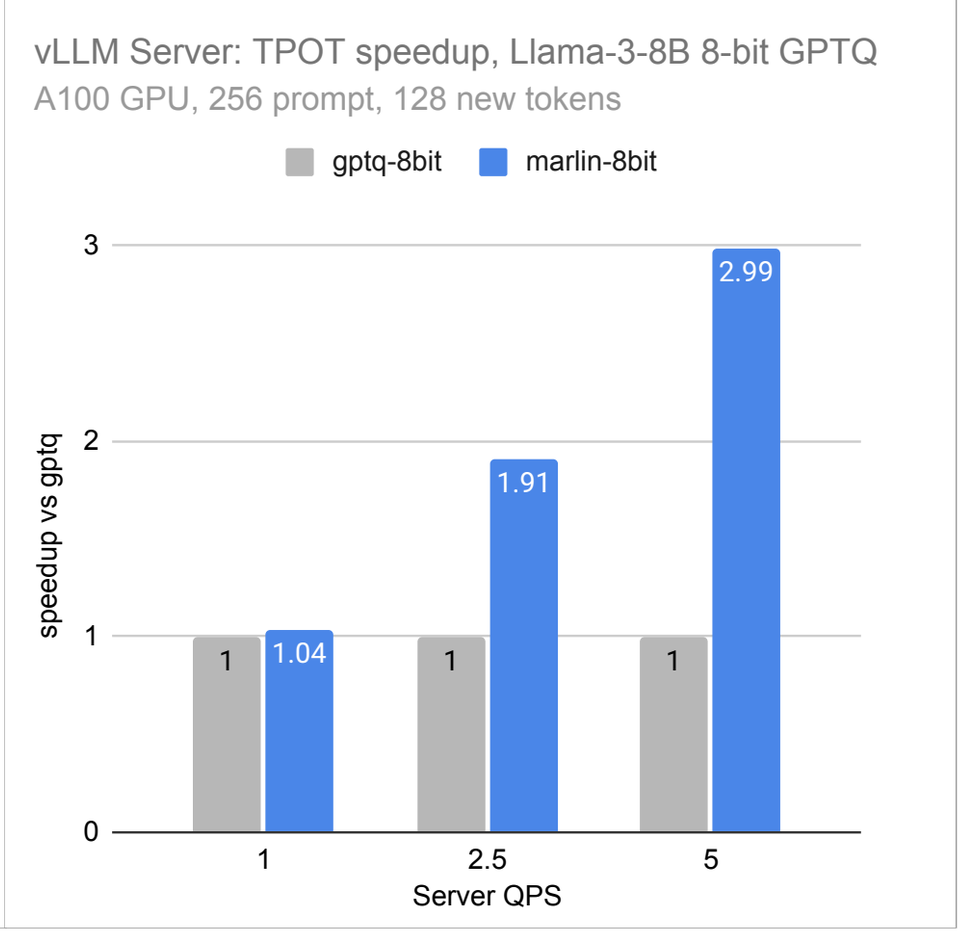

vLLM Server: TTFT, Llama-3-8B 8-bit GPTQ
A100 GPU, 256 prompt, 128 new tokens
gptq-8bit
marlin-8bit
ms
80
60
40
20
0
47.72
38.76
56.14
39.62
66.01
44.17
1
2.5
5
Server QPS


vLLM Server: TTFT speedup, Llama-3-8B 8-bit GPTQ
A100 GPU, 256 prompt, 128 new tokens
gptq-8bit
marlin-8bit
speedup vs gptq
1.5
1.0
0.5
0.0
1
1.23
1
1.42
1
1.49
1
2.5
5
Server QPS


vLLM Server: TPOT, Llama-3-8B 8-bit GPTQ
A100 GPU, 256 prompt, 128 new tokens
gptq-8bit
marlin-8bit
ms
50
40
30
20
10
0
11.34
10.9
23.13
12.08
43.62
14.59
1
2.5
5
Server QPS


vLLM Server: TPOT speedup, Llama-3-8B 8-bit GPTQ
A100 GPU, 256 prompt, 128 new tokens
gptq-8bit
marlin-8bit
speedup vs gptq
3
2
1
0
1
1.04
1
1.91
1
2.99
1
2.5
5
Server QPS