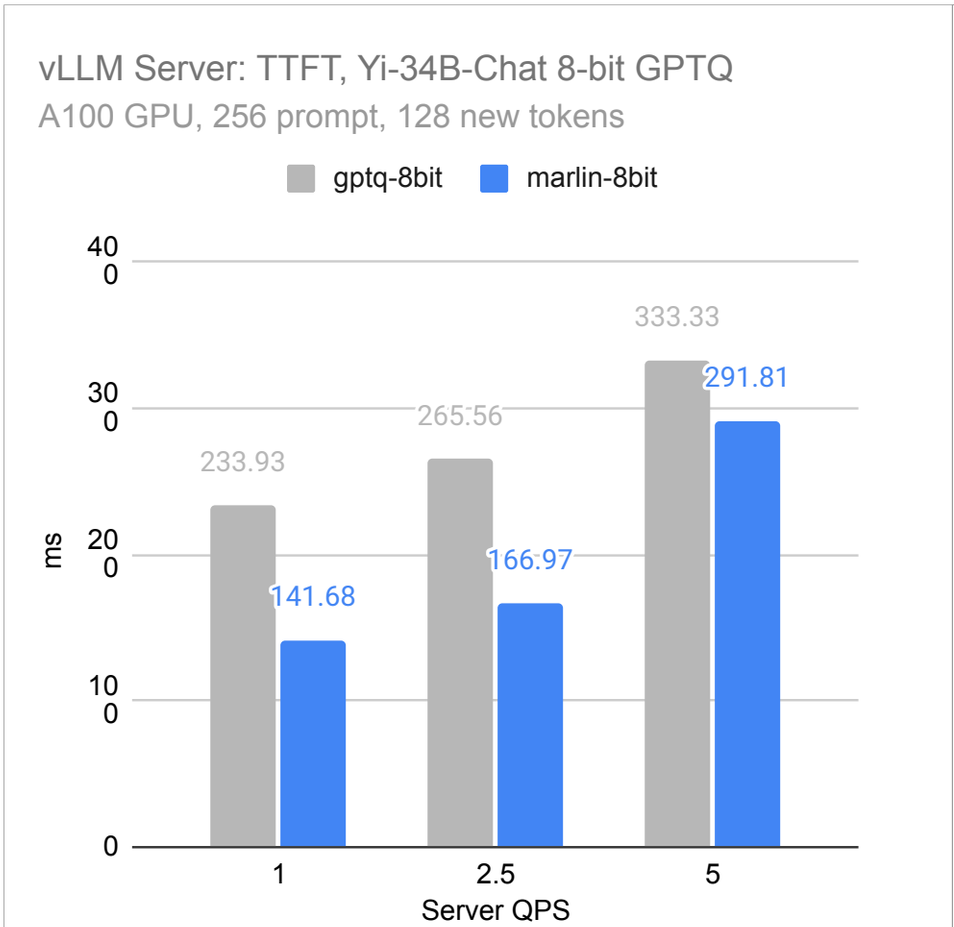
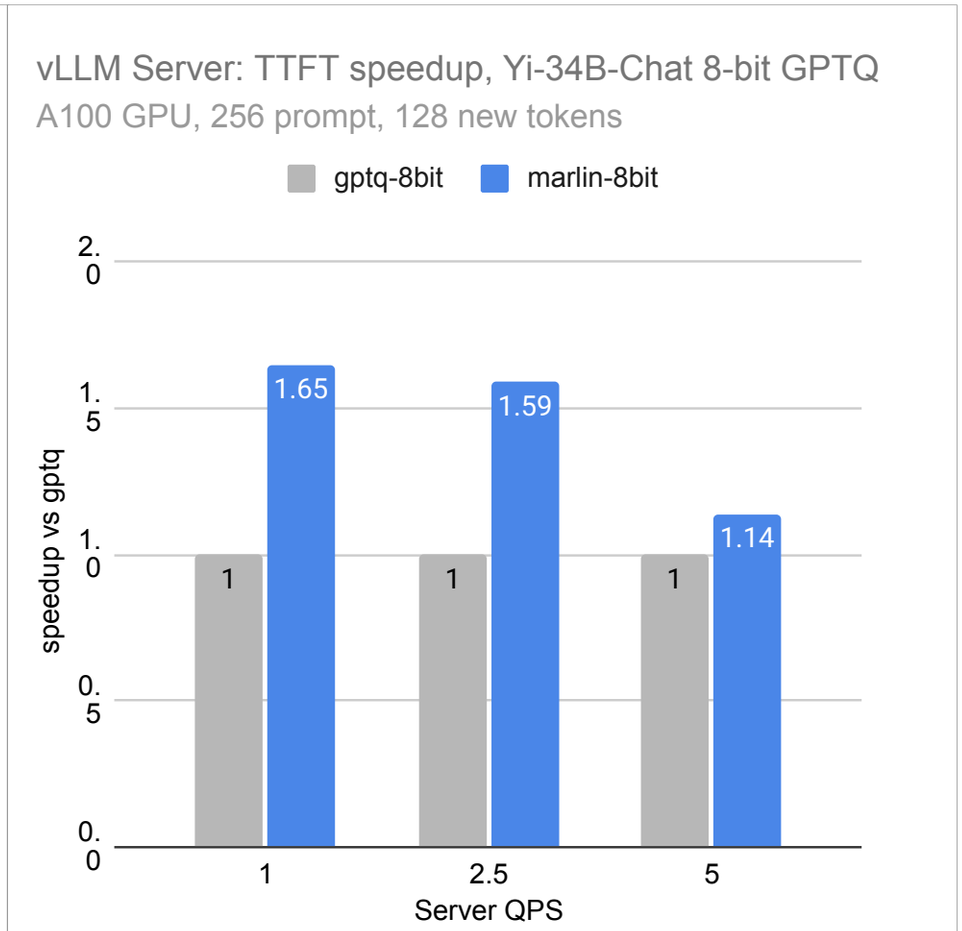
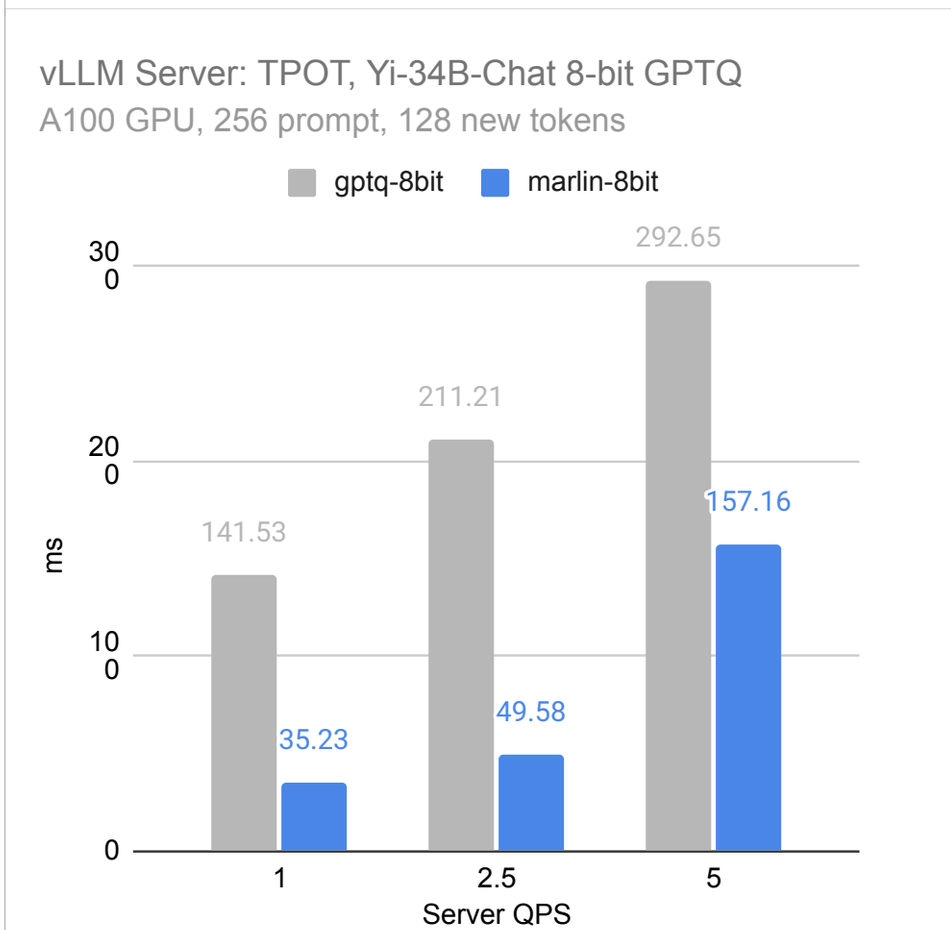
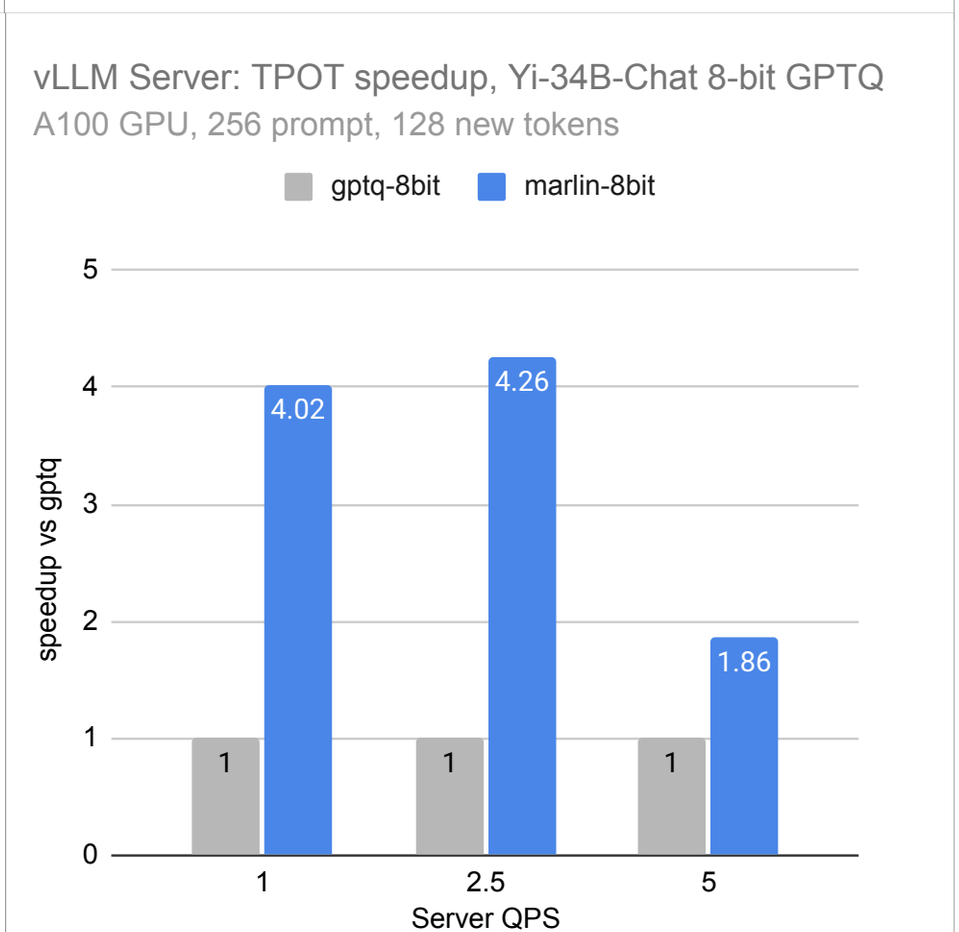

vLLM Server: TTFT, Yi-34B-Chat 8-bit GPTQ
A100 GPU, 256 prompt, 128 new tokens
gptq-8bit
marlin-8bit
ms
40
0
30
0
20
0
10
0
0
233.93
141.68
265.56
166.97
333.33
291.81
1
2.5
5
Server QPS


vLLM Server: TTFT speedup, Yi-34B-Chat 8-bit GPTQ
A100 GPU, 256 prompt, 128 new tokens
gptq-8bit
marlin-8bit
speedup vs gptq
2.
0
1.
5
1.
0
0.
5
0.
0
1
1.65
1
1.59
1
1.14
1
2.5
5
Server QPS


vLLM Server: TPOT, Yi-34B-Chat 8-bit GPTQ
A100 GPU, 256 prompt, 128 new tokens
gptq-8bit
marlin-8bit
ms
30
0
20
0
10
0
0
141.53
35.23
211.21
49.58
292.65
157.16
1
2.5
5
Server QPS


vLLM Server: TPOT speedup, Yi-34B-Chat 8-bit GPTQ
A100 GPU, 256 prompt, 128 new tokens
gptq-8bit
marlin-8bit
speedup vs gptq
5
4
3
2
1
0
1
4.02
1
4.26
1
1.86
1
2.5
5
Server QPS