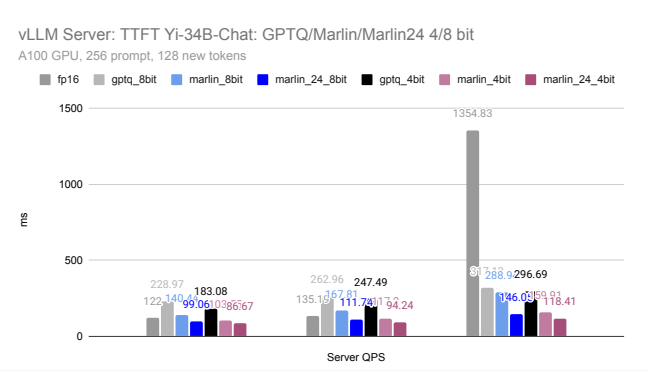
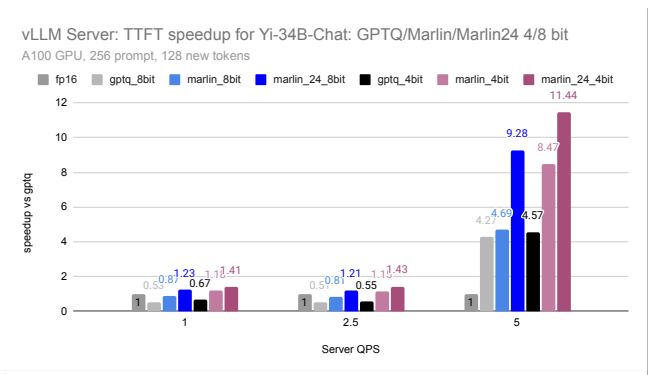
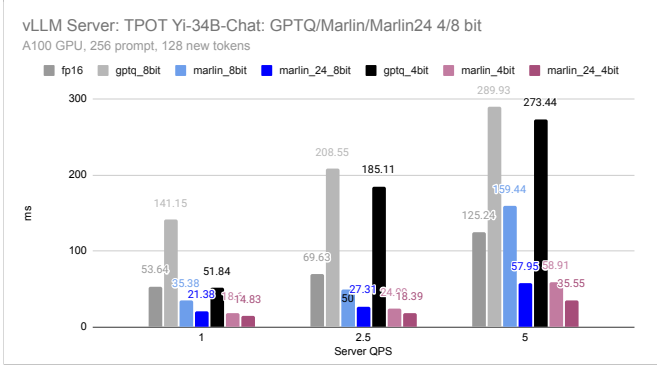
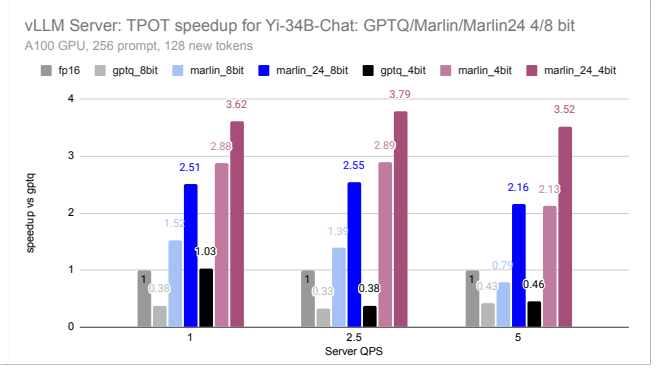

vLLM Server: TTFT Yi-34B-Chat: GPTQ/Marlin/Marlin24 4/8 bit
A100 GPU, 256 prompt, 128 new tokens
fp16
gptq_8bit
marlin_8bit
marlin_24_8bit
gptq_4bit
marlin_4bit
marlin_24_4bit
ms
Server QPS


vLLM Server: TTFT speedup for Yi-34B-Chat: GPTQ/Marlin/Marlin24 4/8 bit
A100 GPU, 256 prompt, 128 new tokens
fp16
gptq_8bit
marlin_8bit
marlin_24_8bit
gptq_4bit
marlin_4bit
marlin_24_4bit
speedup vs gptq
Server QPS


vLLM Server: TPOT Yi-34B-Chat: GPTQ/Marlin/Marlin24 4/8 bit
A100 GPU, 256 prompt, 128 new tokens
fp16
gptq_8bit
marlin_8bit
marlin_24_8bit
gptq_4bit
marlin_4bit
marlin_24_4bit
ms
Server QPS


vLLM Server: TPOT speedup for Yi-34B-Chat: GPTQ/Marlin/Marlin24 4/8 bit
A100 GPU, 256 prompt, 128 new tokens
fp16
gptq_8bit
marlin_8bit
marlin_24_8bit
gptq_4bit
marlin_4bit
marlin_24_4bit
speedup vs gptq
Server QPS