



(a) GPT-1T MegatronLM traffic matrix GPU 1 to 48 (one pipeline stage)

(b) GPT-1T MegatronLM traffic matrix GPU 1 to 192 (four pipeline stages)