Performance for Model: 8B, Seq Len: 16384 on H100 176 7017.92 7903.25 9521.45 9514.20 8722.88 9515.11 -11000 160 7012.91 7888.42 8715.96 9520.69 9520.40 9535.94 10000 Host Memory (GB) 12 128 144 7014.21 7894.79 9511.62 8700.18 9496.42 9522.41 Tokens/sec 9000 7009.29 9511.69 7893.89 8702.49 9501.83 9519.28 8000 7012.24 7895.33 8700.67 9487.04 9502.69 9511.85 - 7000 7018.34 7914.96 8712.12 9526.40 9528.12 9534.93 - 6000 9537.80 8713.39 9054.05 9191.16 9459.19 9528.07 30 40 50 60 70 78 Device Memory (GB)