Performance for Model: 8B, Seq Len: 16384 on H100 176 406.43 457.70 505.17 551.05 551.41 550.99 650 160 406.14 456.84 504.77 551.37 551.35 552.25 600 Host Memory (GB) 12 128 144 406.21 457.21 503.85 549.96 550.85 551.47 - 550 TFLOPS/s 405.93 550.85 457.16 503.99 550.28 551.29 - 450 406.10 457.24 503.88 549.42 550.33 550.86 - 400 96 406.45 504.54 458.38 551.70 551.80 552.20 - 350 80 504.62 524.35 532.29 547.81 551.80 552.36 - 300 30 50 60 40 78 70 Device Memory (GB)