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