Performance for Model: 8B, Seq Len: 32768 on H100 176 492.32 511.52 527.95 549.76 551.18 552.04 650 160 493.06 510.42 527.52 548.89 551.36 552.55 600 Host Memory (GB) 12 128 144 492.85 527.66 549.69 551.03 553.81 511.33 - 550 oo TFLOPS/ 508.79 519.27 549.59 529.99 552.83 552.53 524.04 549.62 533.02 537.95 543.09 552.90 450 96 513.46 520.00 524.63 531.02 536.08 540.40 - 400 80 502.91 508.04 512.50 518.26 523.55 526.19 - 350 60 50 30 40 78 70 Device Memory (GB)