Performance for Model: 8B, Seq Len: 65536 on H100 176 493.89 521.32 524.79 528.17 530.35 522.87 650 160 492.51 520.03 515.55 518.21 522.53 524.26 600 Host Memory (GB) 12 128 144 486.61 510.38 512.00 514.77 517.45 519.30 - 550 oo TFLOPS/ 483.02 507.90 514.22 505.72 511.57 512.21 480.78 504.15 507.65 501.12 502.94 509.28 450 96 476.66 498.19 499.92 501.85 504.49 506.20 - 400 80 430.35 454.33 462.22 470.45 475.01 478.77 - 350 30 60 50 40 78 70 Device Memory (GB)