Performance for Model: 8B, Seq Len: 65536 on H100 176 499.56 527.84 532.72 533.49 535.46 529.38 650 160 498.53 522.12 523.79 526.79 528.91 530.24 600 Host Memory (GB) 12 128 144 494.48 519.53 523.86 527.21 521.38 523.76 - 550 oo TFLOPS/ 490.05 514.04 520.95 517.03 518.72 521.80 485.57 510.19 515.90 511.18 512.16 517.30 450 96 482.92 507.85 509.47 505.68 511.39 513.27 - 400 80 435.10 460.85 466.87 477.87 482.05 486.45 - 350 30 60 50 78 40 70 Device Memory (GB)