Performance for Model: 8B, Seq Len: 32768 on H100 176 487.06 505.18 521.37 542.16 544.31 544.24 650 160 487.12 504.48 520.85 542.15 543.04 543.91 600 Host Memory (GB) 12 128 144 486.79 540.67 504.34 520.70 542.58 545.07 - 550 500.53 522.17 541.69 542.80 511.75 542.47 - 450 517.55 521.66 527.12 533.03 538.51 542.84 - 400 96 505.20 510.51 516.30 522.02 527.18 531.48 - 350 80 493.42 499.15 504.08 508.19 514.40 517.95 - 300 30 50 60 40 78 70 Device Memory (GB)