Performance for Model: 8B, Seq Len: 8192 on H100 176 540.10 383.68 415.71 470.85 553.82 554.71 800 160 383.63 415.64 470.92 539.41 553.55 553.57 700 Host Memory (GB) 12 128 144 383.55 415.34 470.24 538.59 552.69 553.80 600 383.63 415.72 470.37 539.35 552.65 553.15 500 383.64 415.40 470.23 539.03 552.74 552.82 400 96 383.92 471.01 540.29 554.43 415.97 553.89 - 300 80 494.86 509.84 525.62 542.81 553.59 553.89 60 50 30 78 40 70 Device Memory (GB)