Performance for Model: 1B, Seq Len: 8192 on H100 9/ 321.39 431.71 451.14 452.46 453.42 451.27 451.76 452.85 452.07 800 160 321.53 431.64 451.76 451.79 451.06 | 451.65 450.48 450.35 144 450.97 321.44 433.24 449.64 451.39 452.92 451.66 452.11 450,90 700 128 321.53 430.88 448.85 450.32 450.50 452.37 451.69 452.17 451.75 Host Memory (GB) 431.03 452.53 452.69 600 321.70 449.60 452.38 452.31 452.46 452.16 321.34 432.03 450.68 451.15 449.83 450.22 451.90 452.31 451.56 500 321.59 430.31 449.31 451.71 451.88 452.42 451.82 451.86 452.04 321.52 433.47 451.46 451.14 450.62 452.61 451.78 451.90 - 400 321.31 431.84 450.39 452.57 453.03 452.19 452.78 452.80 451.78 321.84 432.90 450.43 453.00 452.31 452.97 451.89 451.37 452.75 - 300 356.77 434.90 450.30 450.37 452.27 452.86 451.99 451.12 451.06 20 24 28 30 40 50 60 70 78 Device Memory (GB)