Performance for Model: 1B, Seq Len: 65536 on H100 9/ 406.33 407.98 410.42 410.74 385.69 407.26 410.40 400.68 800 160 385.53 406.23 406.88 407.92 409.90 410.19 410.55 | 398.58 | 397.50 144 410.91 385.37 406.67 406.91 406.87 398.06 410.34 411.06 409.52 700 128 404.86 405.27 405.36 399.17 410.38 385.62 410.26 411.77 410.84 Host Memory (GB 386.50 405.34 406.91 406.95 409.94 410.27 600 411.44 | 410.38 | 410.67 386.72 406.17 407.35 406.69 407.94 | 407.48 | 411.08 | 410.85 | 411.41 500 386.19 404.30 405.23 407.05 410.26 409.76 410.55 410.84 410.77 386.02 406.85 407.99 407.31 409.64 411.32 |410.18|410.58|411.24 - 400 382.24 403.16 406.05 408.06 410.42 409.72 410.29 411.93 426.80 399.63 401.25 405.59 408.83 409.75 412.27 409.38 - 300 351.19 370.54 377.23 377.58 396.84 |403.14|408.86|411.24|409.46 20 24 30 40 50 60 70 78 28 Device Memory (GB)