Performance for Model: 1B, Seq Len: 32768 on H100 56 309.93 418.98 421.39 430.92 430.42 395.10 421.45 650 48 310.06 394.27 421.56 425.83 429.17 430.26 430.20 600 Host Memory (GB) 32 40 310.69 - 550 394.15 413.77 425.83 429.10 422.64 422.48 00 TFLOPS/ 419.62 310.66 395.72 415.00 422.59 422.35 422.20 450 340.79 415.79 420.61 419.82 423.51 430.64 431.48 400 16 368.44 401.19 412.77 420.34 425.93 430.77 432.70 - 350 16 20 28 32 24 64 78 Device Memory (GB)