Performance for Model: 1B, Seq Len: 16384 on H100 9/ 444.32 444.69 444.20 442.48 442.29 444.56 359.02 442.21 443.03 650 160 358.69 441.52 442.62 443.13 443.06 443.17 443.71 442.11 441.23 144 359.00 442.88 431.55 443.02 444.68 442.98 443.49 444.12 442.19 600 128 440.02 441.08 442.80 442.50 443.26 359.03 443.35 444.06 443.93 - 550 Host Memory (GB) 444.14 442.81 359.08 442.18 442.74 442.74 443.81 443.75 443.89 o O TFLOPS/s 443.65 359.18 441.94 442.63 441.67 441.18 444.25 443.35 442.49 358.94 436.56 441.01 443.19 443.53 443.81 443.71 443.61 443.50 - 450 359.41 442.77 439.30 443.19 442.06 444.63 444.00 443.15 443.61 - 400 358.95 440.83 443.23 443.82 443.04 444.71 443.10 442.71 443.78 358.68 441.20 443.58 444.59 442.30 443.30 443.77 442.46 443.81 - 350 393.05 442.33 444.12 444.77 444.51 444.61 443.68 442.00 442.80 -300 20 24 28 30 40 50 60 70 78 Device Memory (GB)