Performance for Model: 8B, Seq Len: 16384 on H100 176 406.84 460.39 509.70 562.39 561.29 562.81 650 160 407.04 460.60 560.83 509.32 563.00 562.47 600 Host Memory (GB) 12 128 144 406.99 509.43 460.59 561.08 562.41 563.26 - 550 o 00 TFLOPS/ 406.97 460.42 508.95 560.79 563.54 562.75 407.22 509.32 460.32 560.64 562.61 563.09 450 96 407.32 460.80 509.46 561.33 561.64 564.62 - 400 80 513.76 533.74 545.63 557.94 563.73 562.79 - 350 60 50 30 40 78 70 Device Memory (GB)