Performance for Model: 8B, Seq Len: 512 on H100 176 342.31 393.35 464.76 542.18 573.01 574.14 650 160 342.34 600 Host Memory (GB) 12 128 144 - 550 TFLOPS/s 450 96 -400 80 - 350 60 30 40 78 50 70 Device Memory (GB)