Performance for Model: 8B, Seq Len: 1024 on H100 176 345.16 396.94 468.82 546.03 572.06 573.03 650 160 345.38 600 Host Memory (GB) 12 128 144 - 550 TFLOPS/s 450 96 -400 80 - 350 60 40 78 30 50 70 Device Memory (GB)