Performance for Model: 8B, Seq Len: 2048 on H100 176 350.69 475.67 402.59 546.90 572.97 572.41 650 160 351.09 -600 Host Memory (GB) 12 128 144 - 550 TFLOPS/s - 450 96 -400 80 - 350 60 40 78 30 50 70 Device Memory (GB)