Performance for Model: 8B, Seq Len: 8192 on H100 176 384.46 417.86 473.04 556.33 565.67 566.24 650 160 384.68 600 Host Memory (GB) 12 128 144 - 550 TFLOPS/s 450 96 -400 80 - 350 60 40 78 30 50 70 Device Memory (GB)