Performance for Model: 8B, Seq Len: 1024 on H100 176 344.72 396.15 536.92 558.29 559.78 467.41 650 160 344.37 395.59 466.94 535.87 558.68 559.26 600 Host Memory (GB) 12 128 144 344.27 395.62 466.56 536.11 558.01 559.81 - 550 o 00 TFLOPS/ 344.34 395.71 467.07 535.99 557.57 558.72 344.23 395.95 467.08 535.69 557.98 558.54 450 96 344.70 536.65 559.61 396.11 467.37 558.69 - 400 80 468.01 480.32 506.76 542.04 558.50 558.85 - 350 60 50 30 78 40 70 Device Memory (GB)