Performance for Model: 8B, Seq Len: 65536 on A100 176 -180 105.13 105.26 105.57 105.82 106.21 106.34 160 160 104.65 104.90 104.39 105.28 105.71 105.89 - 140 Host Memory (GB) 12 128 144 104.02 103.90 103.69 104.90 105.22 105.48 - 120 CT - TFLOPS/s 103.50 103.50 103.84 104.24 104.52 104.68 100 102.83 103.51 104.04 104.11 102.90 103.24 - 80 96 102.14 102.52 102.94 103.25 102.35 103.34 60 80 78.91 84.41 81.79 87.02 89.47 91.70 30 50 60 40 78 70 Device Memory (GB)