Performance for Model: 8B, Seq Len: 32768 on A100 176 -180 124.63 124.56 124.76 124.77 124.95 124.76 160 160 124.79 124.83 124.84 124.18 124.89 125.09 - 140 Host Memory (GB) 12 128 144 124.31 125.09 124.12 124.96 125.23 125.38 - 120 TFLOPS/s 123.36 124.19 124.61 124.95 124.91 124.72 100 121.13 125.08 125.11 121.92 122.83 123.68 - 80 96 118.98 119.76 120.29 121.64 122.62 122.69 60 80 119.62 120.41 117.06 117.75 118.41 120.58 30 50 60 40 78 70 Device Memory (GB)