Performance for Model: 8B, Seq Len: 4096 on H100 176 361.65 415.34 468.23 538.01 555.13 556.74 650 160 467.60 361.23 415.13 536.42 556.60 555.83 600 Host Memory (GB) 12 128 144 361.26 415.09 467.92 536.54 555.13 557.13 - 550 TFLOPS/s 361.29 415.26 467.15 536.31 555.03 555.60 - 450 361.24 415.06 467.39 536.16 554.79 555.85 - 400 96 467.81 361.65 415.53 536.15 556.44 556.46 - 350 80 486.29 492.72 510.46 541.82 555.88 557.46 - 300 30 50 60

Device Memory (GB)

78

70

40