Performance for Model: 1B, Seq Len: 32768 on A100 - 180 67.98 68.34 68.20 67.88 68.15 68.37 160 67.94 68.00 68.24 68.18 68.22 68.39 68.27 68.30 68.23 - 140 L O TFLOPS/s 67.48 67.98 68.15 68.06 68.24 68.15 68.18 68.30 68.12 100 67.86 67.96 67.95 67.94 68.09 68.16 68.10 68.11 68.09 - 80 66.63 67.04 67.39 67.49 67.90 67.56 67.92 67.88 68.04 60

80

64

Host Memory (GB) 48

32

20

24

28

30

40

Device Memory (GB)

50

60

70

78