Performance for Model: 1B, Seq Len: 4096 on H100 414.51 401.54 465.44 466.34 466.64 467.36 414.36 650 414.58 402.08 414.20 465.40 467.24 466.76 466.78 600 - 550 413.72 402.04 414.08 465.74 467.04 466.64 467.62 00 TFLOPS/ 414.49 402.19 413.87 467.40 466.63 467.31 467.96 450 415.39 401.76 414.54 467.35 468.83 467.25 468.98 400 414.22 401.92 414.08 465.17 468.49 466.68 469.56

32

64

78

28

Device Memory (GB)

24

- 350

56

48

Host Memory (GB) 32 40

16

16

20