Performance for Model: 1B, Seq Len: 1024 on H100 9/ 395.95 469.69 469.45 469.71 278.69 468.54 470.05 468.81 468.97 800 160 278.83 397.42 467.94 469.60 468.94 468.23 469.46 469.43 467.29 144 278.73 395.94 468.14 469.19 468.76 469.15 468.87 469.82 l 467.56 700 128 466.68 468.98 469.25 278.62 397.23 469.64 470.06 469.41 469.40 396.91 468.36 468.95 600 278.66 468.72 468.94 468.28 469.75 468.29 396.51 466.81 467.94 469.86 468.38 278.52 469.17 468.23 468.21 500 278.69 395.59 466.60 469.88 470.65 469.60 470.13 469.70 469.76 469.06 278.79 396.67 467.12 469.33 470.06 | 469.88 | 470.04 | 470.16 - 400 278.69 395.48 468.72 466.87 470.00 468.55 469.32 469.84 469.76 396.66 468.85 469.46 469.43 470.41 469.48 470.00 468.99 278.97 - 300 310.90 396.36 467.34 468.53 467.71 467.55 469.37 468.96 469.96 20 24 30 60 70 28 40 50 78

Device Memory (GB)

Host Memory (GB