Performance for Model: 1B, Seq Len: 2048 on H100 9/ 398.09 464.80 464.52 284.96 463.29 465.39 465.81 465.41 466.69 800 160 284.95 397.71 464.63 466.20 465.35 464.60 464.88 465.16 463.53 144 284.95 398.11 465.39 465.36 466.23 465.59 466.73 466.06 463.31 700 128 398.69 465.28 284.94 463.64 464.96 466.62 464.77 465.42 466.51 Host Memory (GB) 80 96 112 464.30 465.96 465.67 464.56 600 284.93 398.72 465.92 464.23 465.58 464.68 466.16 284.78 398.18 463.96 463.34 466.02 466.92 466.49 500 284.95 397.61 462.76 466.48 465.29 466.09 465.67 466.30 465.54 285.03 466.99 396.36 465.62 467.39 466.61 466.81 466.17 464.32 - 400 284.91 397.46 465.56 465.42 467.50 464.92 464.65 464.62 464.33 398.12 465.21 465.46 465.92 465.32 466.03 285.20 466.12 465.79 - 300 317.69 397.60 464.95 466.04 466.88 465.53 466.58 466.12 464.32 20 24 30 40 50 60 70 28 78

Device Memory (GB)