Performance for Model: 1B, Seq Len: 512 on H100 9/ 275.33 392.51 470.46 470.56 470.04 471.11 470.80 471.19 160 650 275.53 393.03 470.95 470.42 |471.19|470.04|470.60|470.10|468.87 144 275.43 393.77 470.43 470.08 470.41 471.69 471.55 471.51 468.79 600 128 469.27 470.89 470.46 471.45 472.04 275.30 393.70 471.40 471.19 Host Memory (GB 393.26 469.97 470.51 471.45 |471.06|470.35|470.64|470.75 275.46 - 550 470.06 470.31 469.96 469.19 275.37 392.88 470.21 471.72 471.86 500 臣 275.46 393.32 467.65 470.15 470.85 470.20 471.25 470.41 471.19 275.53 393.14 470.79 471.06 471.47 470.25 471.26 469.87 470.58 - 450 275.41 393.31 470.49 470.83 469.73 472.35 471.75 470.26 471.18 - 400 471.14 | 471.91 | 471.68 | 470.55 | 471.15 393.42 471.76 275.75 470.44 295.29 468.69 471.60 470.80 471.18 471.29 471.14 471.85 393.64 - 350 20 24 30 40 50 60 70 28 78

Device Memory (GB)