Performance for Model: 1B, Seq Len: 512 on H100 176 0.40 0.28 0.48 0.48 0.48 0.48 0.48 0.48 0.48 8.0 160 0.28 0.40 0.48 0.48 0.48 0.48 0.48 0.48 0.48 144 0.28 0.40 0.48 0.48 0.48 0.48 0.48 0.48 0.48 0.7 128 0.28 0.40 0.48 0.48 0.48 0.48 0.48 0.48 0.48 Host Memory (GB) 80 96 112 0.28 0.40 0.48 0.48 0.48 0.48 0.48 0.48 0.48 - 0.6 0.28 0.40 0.48 0.48 0.48 0.48 0.48 0.48 0.48 - 0.5 0.28 0.40 0.47 0.48 0.48 0.48 0.48 0.48 0.48 64 0.28 0.40 0.48 0.48 0.48 0.48 0.48 0.48 0.48 -0.448 0.28 0.40 0.48 0.48 0.48 0.48 0.48 0.48 0.48 32 0.28 0.40 0.48 0.48 0.48 0.48 0.48 0.48 0.48 0.3 16 0.31 0.40 0.48 0.47 0.48 0.48 0.48 0.48 0.48 30 40 50 60 70 78 20 24 28 Device Memory (GB)