Performance for Model: 1B, Seq Len: 1024 on H100 176 0.28 0.40 0.47 0.47 0.47 0.48 0.47 0.47 0.47 8.0 160 0.28 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 144 0.28 0.40 0.47 0.47 0.47 0.47 0.47 0.48 0.47 -0.7128 0.28 0.40 0.47 0.47 0.47 0.47 0.48 0.47 0.47 Host Memory (GB) 80 96 112 0.28 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 - 0.6 0.28 0.40 0.47 0.47 0.47 0.47 0.48 0.47 0.47 - 0.5 0.28 0.40 0.47 0.48 0.48 0.47 0.48 0.47 0.47 64 0.28 0.40 0.47 0.47 0.47 0.48 0.48 0.48 0.48 -0.448 0.28 0.40 0.47 0.47 0.47 0.48 0.48 0.47 0.47 32 0.28 0.40 0.47 0.47 0.47 0.48 0.47 0.48 0.47 - 0.3 16 0.31 0.40 0.47 0.47 0.47 0.47 0.48 0.47 0.47 28 30 40 50 60 70 78 20 24 Device Memory (GB)