Performance for Model: 1B, Seq Len: 2048 on H100 176 0.29 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 8.0 160 0.29 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 144 0.29 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 -0.7128 0.29 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 Host Memory (GB) 80 96 112 0.29 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 - 0.6 0.29 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 - 0.5 0.29 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 64 0.29 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 -0.448 0.29 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 32 0.29 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 - 0.3 16 0.32 0.40 0.47 0.47 0.47 0.47 0.47 0.47 0.47 28 30 40 50 60 70 78 20 24 Device Memory (GB)