Performance for Model: 1B, Seq Len: 16384 on H100 176 0.47 0.47 0.38 0.47 0.47 0.47 0.47 0.47 0.47 8.0 160 0.38 0.47 0.47 0.47 0.47 0.47 0.47 0.47 0.47 144 0.38 0.47 0.46 0.47 0.47 0.47 0.47 0.47 0.47 -0.7128 0.38 0.47 0.47 0.47 0.47 0.47 0.47 0.47 0.47 Host Memory (GB) 80 96 112 0.38 0.47 0.47 0.47 0.47 0.47 0.47 0.47 0.47 - 0.6 0.38 0.47 0.47 0.47 0.47 0.47 0.47 0.47 0.47 - 0.5 0.38 0.46 0.47 0.47 0.47 0.47 0.47 0.47 0.47 64 0.38 0.47 0.47 0.47 0.47 0.47 0.47 0.47 0.47 -0.448 0.38 0.47 0.47 0.47 0.47 0.47 0.47 0.47 0.47 32 0.38 0.47 0.47 0.47 0.47 0.47 0.47 0.47 0.47 - 0.3 16 0.42 0.47 0.47 0.47 0.47 0.47 0.47 0.47 0.47 24 50 70 30 28 40 60 78 20 Device Memory (GB)