Performance for Model: 1B, Seq Len: 32768 on H100 176 0.40 0.42 0.43 0.43 0.42 0.43 0.43 0.43 0.43 8.0 160 0.40 0.42 0.42 0.42 0.43 0.42 0.42 0.42 0.42 144 0.40 0.42 0.43 0.43 0.42 0.43 0.43 0.43 0.43 -0.7128 0.40 0.41 0.42 0.42 0.43 0.42 0.42 0.42 0.42 Host Memory (GB) 80 96 112 0.40 0.42 0.43 0.43 0.42 0.43 0.43 0.43 0.43 - 0.6 0.40 0.42 0.42 0.42 0.43 0.42 0.42 0.42 0.42 - 0.5 0.40 0.42 0.43 0.42 0.42 0.43 0.43 0.43 0.43 64 0.40 0.42 0.42 0.42 0.43 0.42 0.42 0.42 0.42 -0.448 0.40 0.42 0.43 0.43 0.43 0.43 0.43 0.43 0.42 32 0.40 0.42 0.42 0.42 0.42 0.43 0.42 0.42 0.42 0.3 16 0.40 0.41 0.43 0.42 0.42 0.42 0.43 0.43 0.43 30 40 50 60 70 28 78 20 24 Device Memory (GB)