Performance for Model: 1B, Seq Len: 4096 on H100 176 0.30 0.41 0.46 0.46 0.46 0.46 0.46 0.46 0.46 8.0 160 0.30 0.41 0.46 0.46 0.46 0.46 0.46 0.46 0.46 144 0.30 0.41 0.46 0.46 0.46 0.46 0.46 0.46 0.46 0.7 128 0.30 0.41 0.46 0.46 0.46 0.46 0.46 0.46 0.46 Host Memory (GB) 80 96 112 0.30 0.41 0.46 0.46 0.46 0.46 0.46 0.46 0.46 - 0.6 0.30 0.41 0.46 0.46 0.46 0.46 0.46 0.46 0.46 - 0.5 0.30 0.41 0.46 0.46 0.46 0.46 0.46 0.46 0.46 64 0.30 0.41 0.46 0.46 0.46 0.46 0.46 0.46 0.46 -0.448 0.30 0.41 0.46 0.47 0.46 0.46 0.46 0.46 0.46 32 0.30 0.41 0.46 0.46 0.46 0.46 0.46 0.46 0.47 0.3 16 0.33 0.41 0.46 0.46 0.46 0.46 0.46 0.46 0.46 30 40 50 60 70 20 28 78 24 Device Memory (GB)