Performance for Model: 1B, Seq Len: 8192 on H100 176 0.33 0.45 0.47 0.47 0.47 0.47 0.47 0.47 0.47 8.0 160 0.33 0.45 0.47 0.47 0.47 0.47 0.47 0.47 0.47 144 0.33 0.45 0.47 0.47 0.47 0.47 0.47 0.47 0.47 -0.7128 0.33 0.45 0.47 0.47 0.47 0.47 0.47 0.47 0.47 Host Memory (GB) 80 96 112 0.33 0.45 0.47 0.47 0.47 0.47 0.47 0.47 0.47 - 0.6 0.33 0.45 0.47 0.47 0.47 0.47 0.47 0.47 0.47 - 0.5 0.33 0.45 0.47 0.47 0.47 0.47 0.47 0.47 0.47 64 0.33 0.45 0.47 0.47 0.47 0.47 0.47 0.47 0.47 -0.448 0.33 0.45 0.47 0.47 0.47 0.47 0.47 0.47 0.47 32 0.34 0.45 0.47 0.47 0.47 0.47 0.47 0.47 0.47 - 0.3 16 0.38 0.45 0.47 0.47 0.47 0.47 0.47 0.47 0.47 50 24 28 30 40 60 70 78 20 Device Memory (GB)