Performance for Model: 8B, Seq Len: 512 on H100 18000 7523.81 12272.46 8645.64 12320.83 10205.87 11744.06 7514.22 16000 Host Memory (GB) 112 128 144 14000 12000 96 - 10000 80 - 8000 40 78 30 70 50 60 Device Memory (GB)