Performance for Model: 8B, Seq Len: 4096 on A100 3569.38 3143.24 3340.93 3595.21 3600.72 3610.88 3609.50 3500 3145.61 3338.61 3579.69 3599.55 3611.04 3615.00 3608.41 3000 Host Memory (GB) 70 75 3145.56 3339.99 3575.09 3600.87 3610.81 3613.65 3614.33 Tokens/sec 2500 3144.75 3608.87 3343.24 3578.42 3601.62 3608.78 3614.64 2000 3162.06 3339.50 3582.23 3599.68 3606.44 3616.44 3614.75 - 1500 3226.61 3292.64 3438.33 3585.40 3604.81 3616.68 3612.19 - 1000 24 30 40 50 60 70 78

Device Memory (GB)