Performance for Model: 8B, Seq Len: 2048 on H100

Host Memory (GB)	30	40	50	60	70	78
176	7508.54	8632.04	10095.78	11491.91	11955.96	11963.89
160	7499.10	8622.01	10097.31	11483.84	11942.05	11968.39
144	7505.92	8620.93	10081.37	11476.30	11937.04	11961.92
128	7504.23	8626.28	10087.14	11471.32	11938.63	11949.36
112	7504.23	8622.85	10085.38	11476.79	11939.20	11957.24
96	7511.56	8633.49	10099.21	11497.34	11955.31	11960.65
80	10181.10	10361.90	10897.72	11599.12	11952.39	11971.55

Device Memory (GB)