Performance for Model: 1B, Seq Len: 524288 on H100 3571.893578.623332.49<mark>3569.833567.67</mark>3327.39<mark>3575.87</mark> 3569.04 **Tokens per Second** Host Memory (GB) 80 96 112 - 4500 - 4000 - 3500 Device Memory (GB)