Performance for Model: 1B, Seq Len: 1024 on A100

| Host Memory (GB) \ Device Memory (GB) | 20 | 24 | 28 | 30 | 40 | 50 | 60 | 70 | 78 |
|---|---|---|---|---|---|---|---|---|---|
| 80 | 151.79 | 152.03 | 151.75 | 151.30 | 151.94 | 151.69 | | | |
| 64 | 151.92 | 151.90 | 151.25 | 151.33 | 151.99 | 151.27 | 151.95 | 151.80 | 151.83 |
| 48 | 151.86 | 151.86 | 151.86 | 151.89 | 151.84 | 151.70 | 151.89 | 151.80 | 151.84 |
| 32 | 151.79 | 151.80 | 151.22 | 151.53 | 151.86 | 151.83 | 151.39 | 151.67 | 151.60 |
| 16 | 152.18 | 151.56 | 151.53 | 151.45 | 151.50 | 151.50 | 151.37 | 151.48 | 151.47 |