Performance for Model: 1B, Seq Len: 4096 on A100 - 180 2 - 126.36 | 126.46 | 126.24 | 126.06 | 126.35 | 126.29 - 160 - 126.29 | 126.36 | 126.02 | 126.07 | 126.38 | 126.08 | 126.23 | 125.79 | 126.22 140 Host Memory (GB) - 120 FFLOPS/s 126.22 126.15 | 126.24 | 126.26 | 126.19 | 126.18 | 126.20 | 126.19 | 126.09 100 <u>-</u> 126.10 | 126.04 | 125.78 | 125.91 | 125.98 | 126.13 | 125.85 | 125.99 | 125.99 - 80 <u>9</u> - 126.02 | 125.73 | 125.76 | 125.68 | 125.74 | 125.65 | 125.65 | 125.79 | 125.77 60 20 24 28 30 40 50 60 70 78 Device Memory (GB)