

T-3 MoE 350M Expert Parallel Communication Group sizes vs Throughput: 4x4 A100 80

