GPT-3 MoE 350M All-to-All Group Size (EP) vs Iteration Time: 4x4 A100 80GB

EP = 1
EP = W

Iteration time (s)

Num of Experts