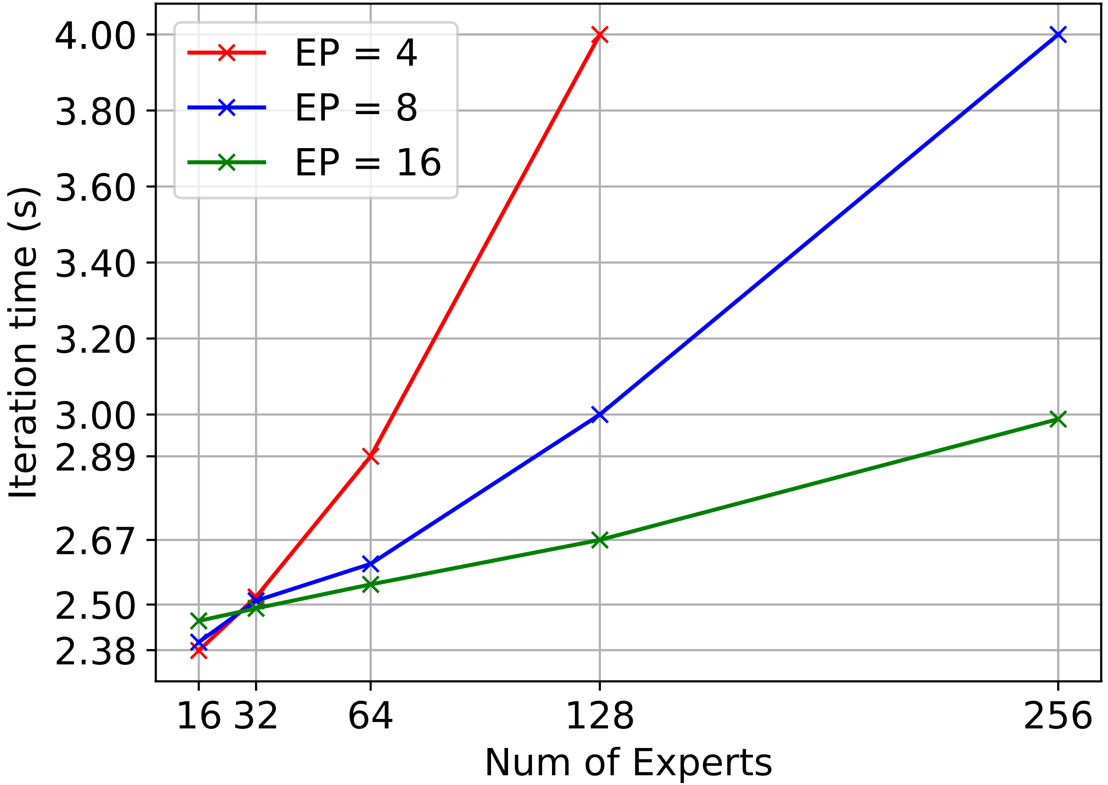
Expert Parallel (EP) vs Iteration Latency (↓ is better)

GPT-3 350M MoE on 4x4 A100 GPUs


EP = 4
EP = 8
EP = 16
Iteration time (s)
Num of Experts