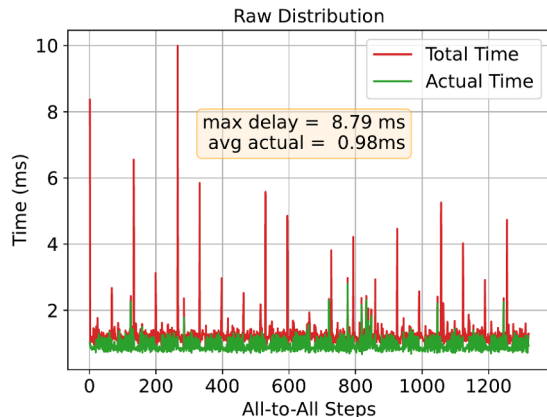
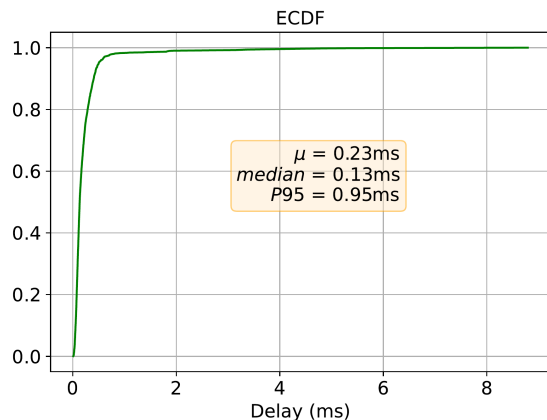


nx4 A100: n Perlmutter GPU nodes, each with 4 A100 80GB
1x8 V100: Azure NDv2 VM

GPT-3 MoE 32x350M All-to-All Straggler Effect: 8x4 A100

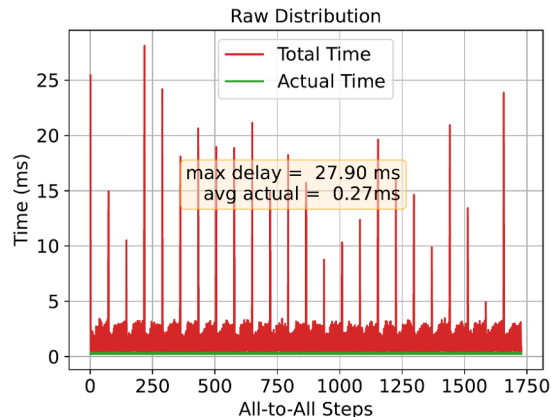


GPT-3 MoE 32x350M All-to-All Straggler Effect: 8x4 A100

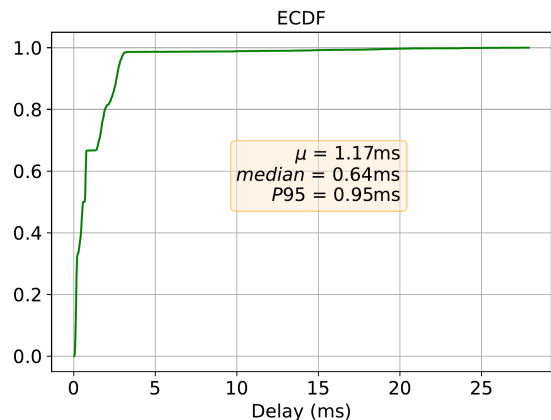


Challenge #1 Straggler Effect

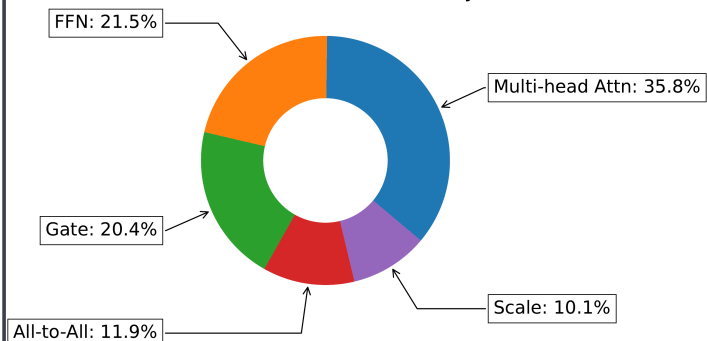
GPT-3 MoE 8x350M All-to-All Straggler Effect: 1x8 V100



GPT-3 MoE 8x350M All-to-All Straggler Effect: 1x8 V100



GPT-3 8x2.7B Forward Pass Kernel Summary: 2x4 A100

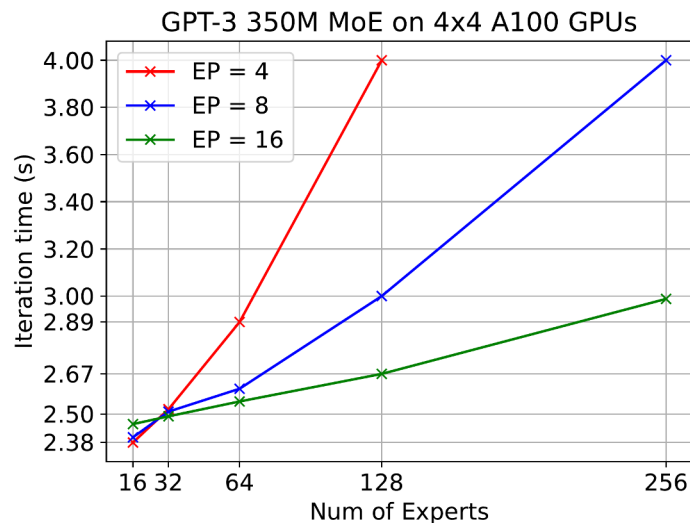


Challenge #2

No overlap of communication and computation.

NB: All-to-All is the *second* most time-consuming kernel, only behind the Attn/FFN GEMM.

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



Challenge #3

Deciding an optimal expert parallel strategy.

NB: EP = 4 yields an OOM error at 256 experts