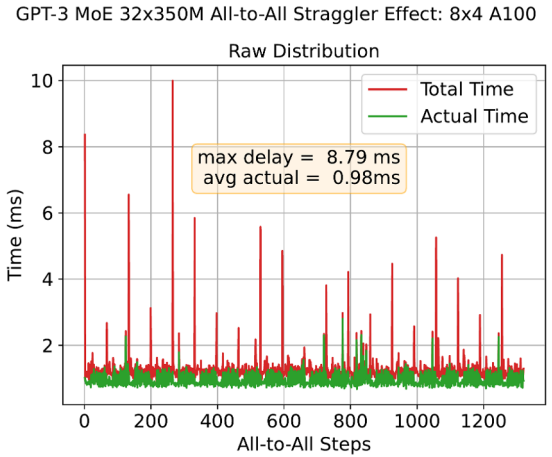
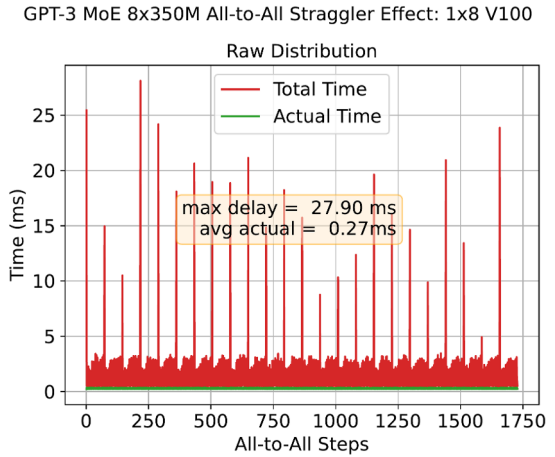
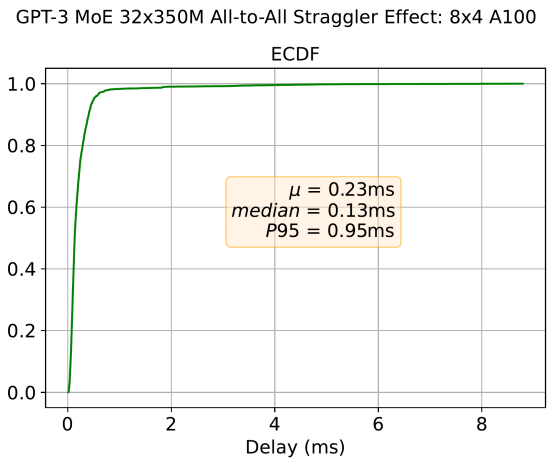
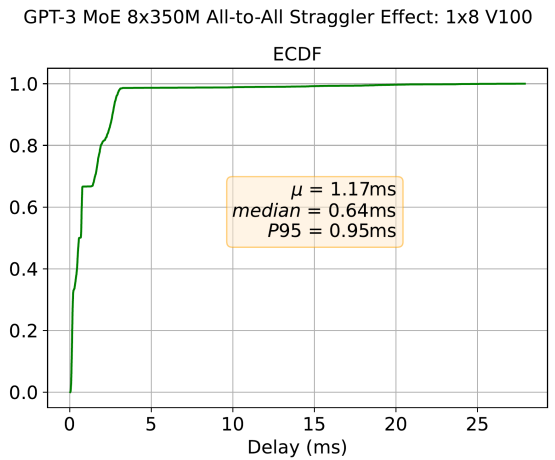
nx4 A100: *n* Perlmutter GPU nodes, each with 4 A100 80GB
1x8 V100: Azure NDv2 VM

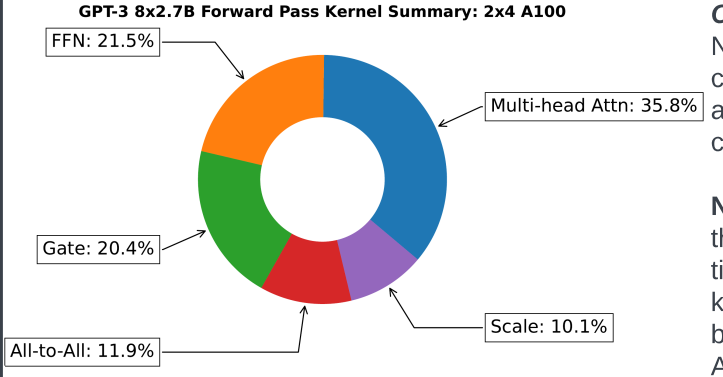








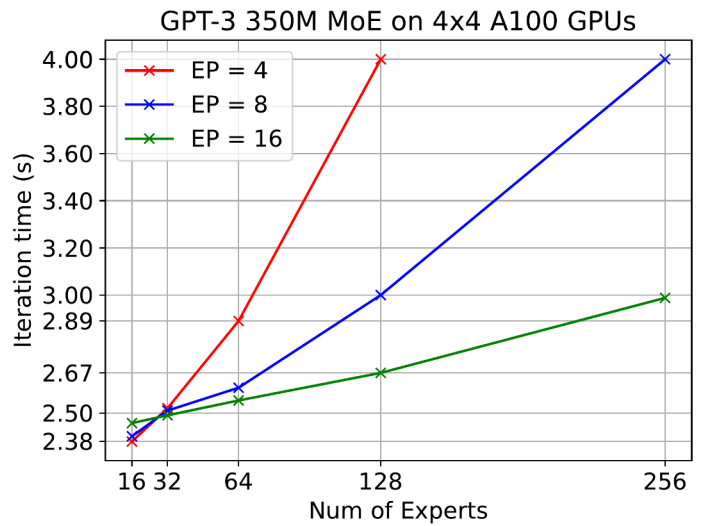
***Challenge* #1** Straggler Effect



***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