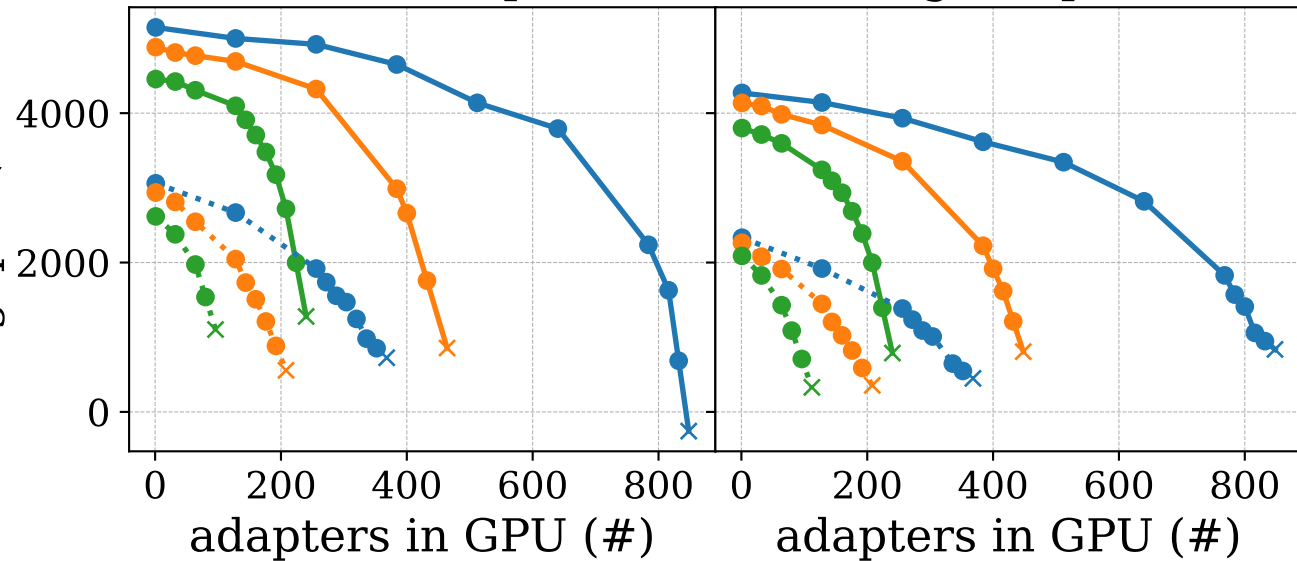


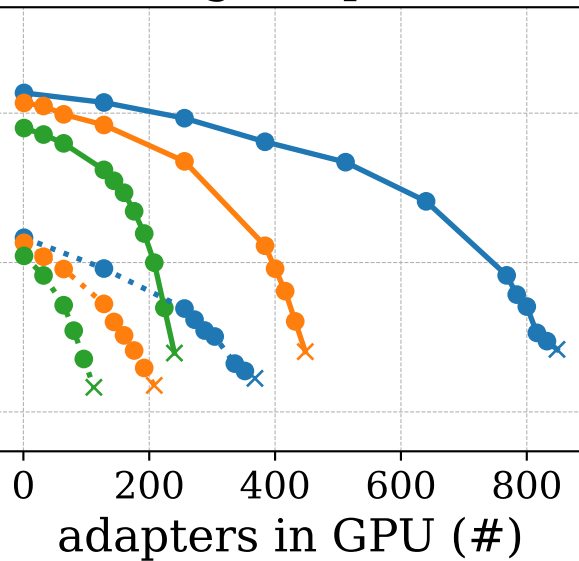


t: throughput (tokens/s)

MediumRequest



LargeRequest



MediumRequest

