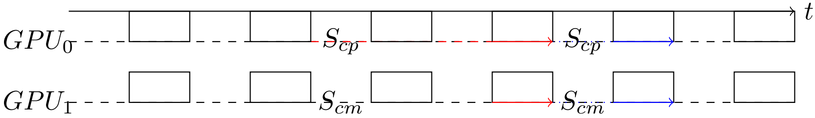

Figure 1: Two types of synchronization: inter-rank synchronization (red dashed line) and intra-rank/inter-stream synchronization (blue dotted line). For simplicity, we assume two GPUs and two streams (S_{cp} and S_{cm}, for compute and communication respectively) per GPU, while CPU op calls are omitted in the plot. GPU kernels are represented by rectangles, and arrows indicate the data dependency between compute and communication kernels.