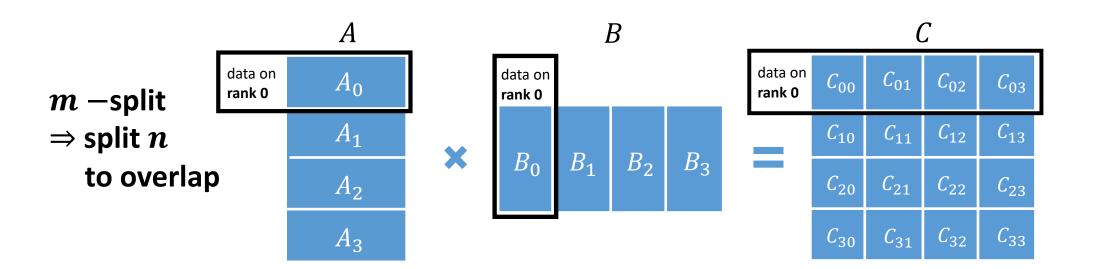
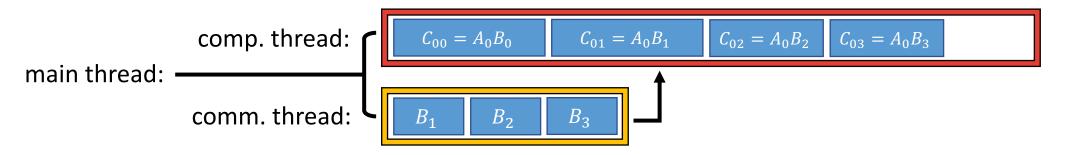


without enforced synchronization

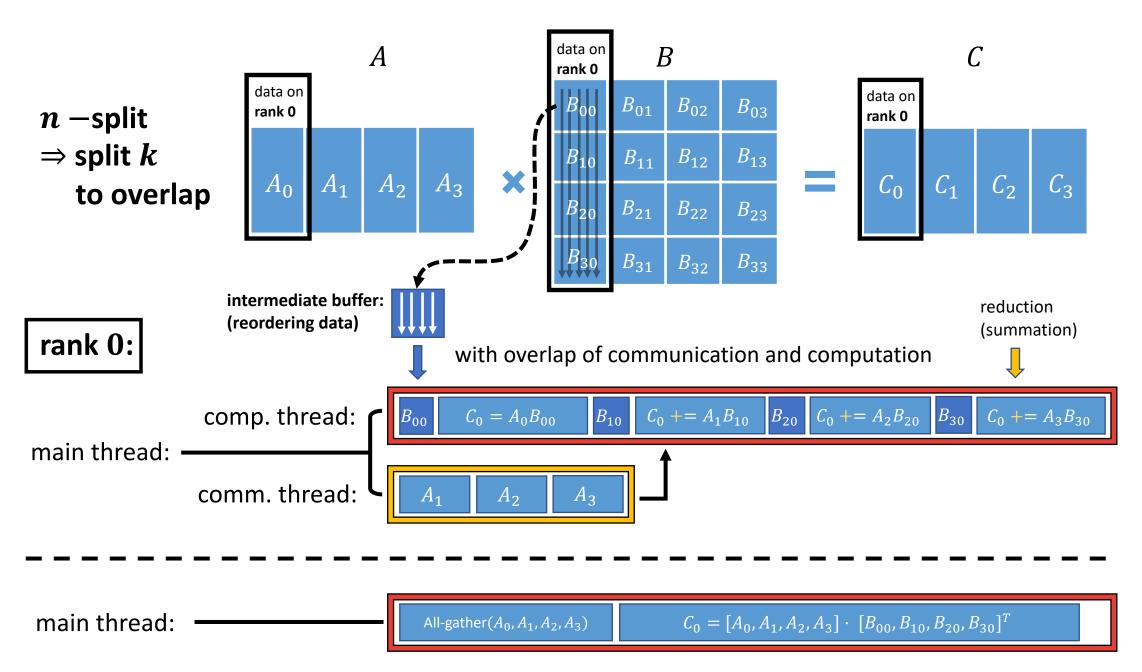




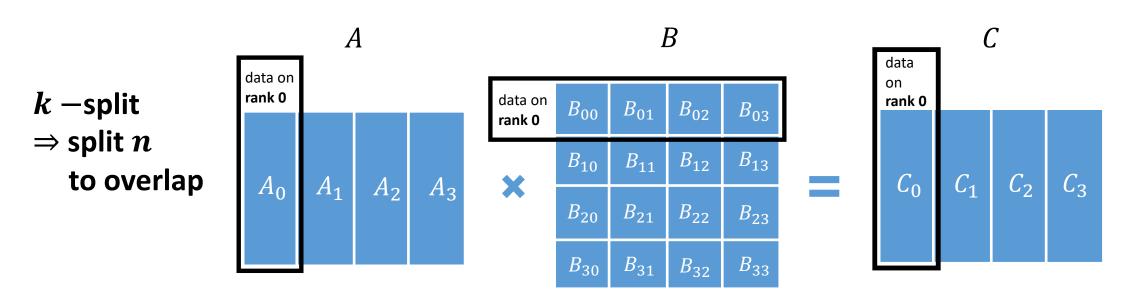
with overlap of communication and computation



without overlap of communication and computation

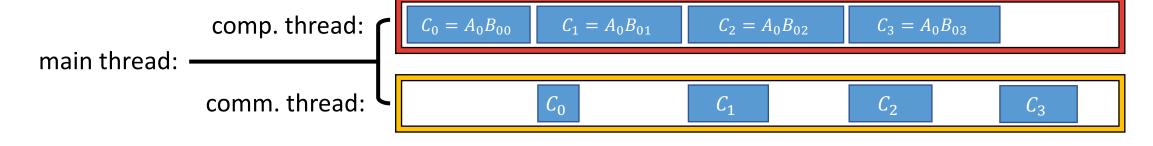


without overlap of communication and computation





with overlap of communication and computation



main thread: $[C_0, C_1, C_2, C_3] = A_0[B_{00}, B_{01}, B_{02}, B_{03}]$ Reduce-scatter (C_0, C_1, C_2, C_3)

without overlap of communication and computation

