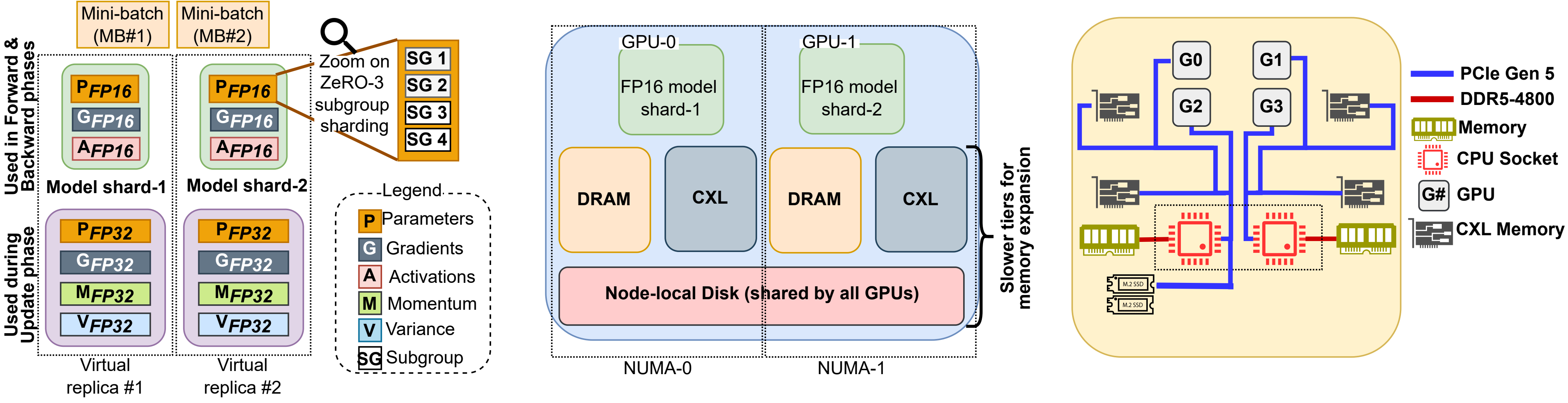

(a) ZeRO-3's hybrid data and model parallelism with subgroup sharding

(b) ZeRO-3 offloading to slower tiers and NUMA mapping

(c) Architecture of our compute node with 4xH100 GPUs