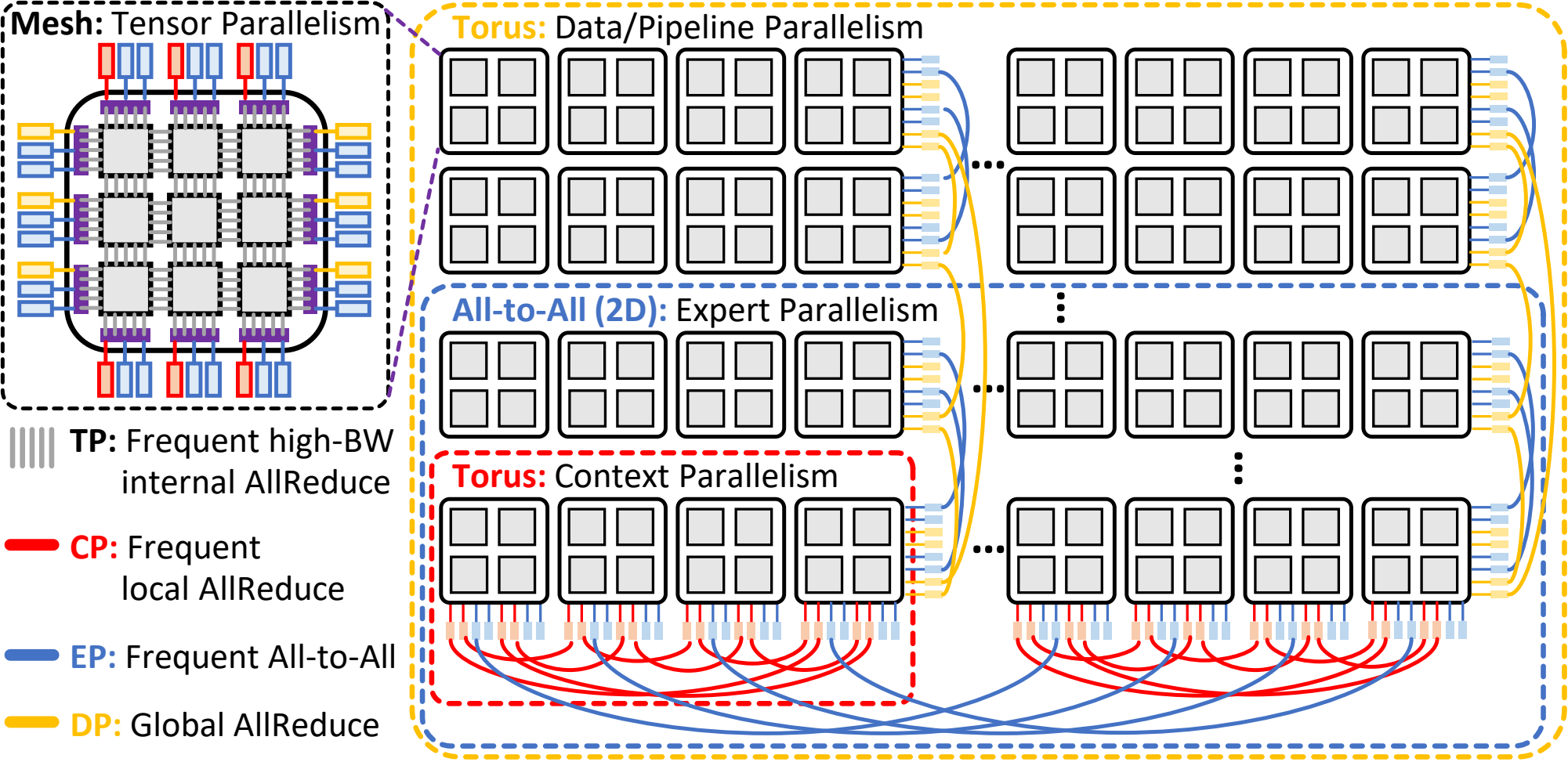
Mesh: Tensor Parallelism
Torus: Data/Pipeline Parallelism
All-to-All (2D): Expert Parallelism
Torus: Context Parallelism
TP: Frequent high-BW internal AllReduce
CP: Frequent local AllReduce
EP: Frequent All-to-All
DP: Global AllReduce