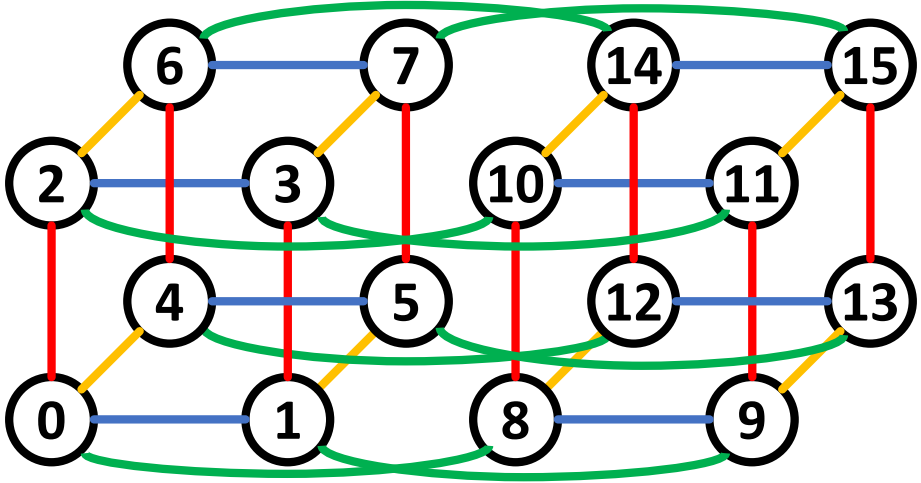


Tensor/Sequence Parallelism

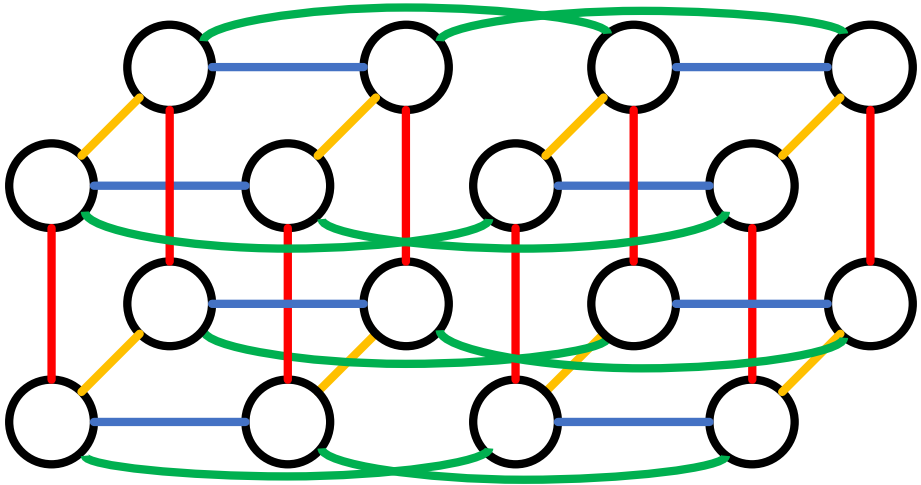
Context Parallelism

Expert Parallelism (FFN layers)

Data Parallelism (FFN layers)

*ADP (Attention) = EDP (FFN) × EP

Pipeline Parallelism



Traffic	Process Group
TP(2):	[0 1]
CP(2):	[0 2]
CPxTP(4):	[0 1 2 3]
EP(2):	**[0 4] (all-to-all)**
DP(4):	[0 4 8 12]
CPxADP(8):	[0 2 4 6 8 10 12 14]
CPxEDP(4):	[0 2 8 10]