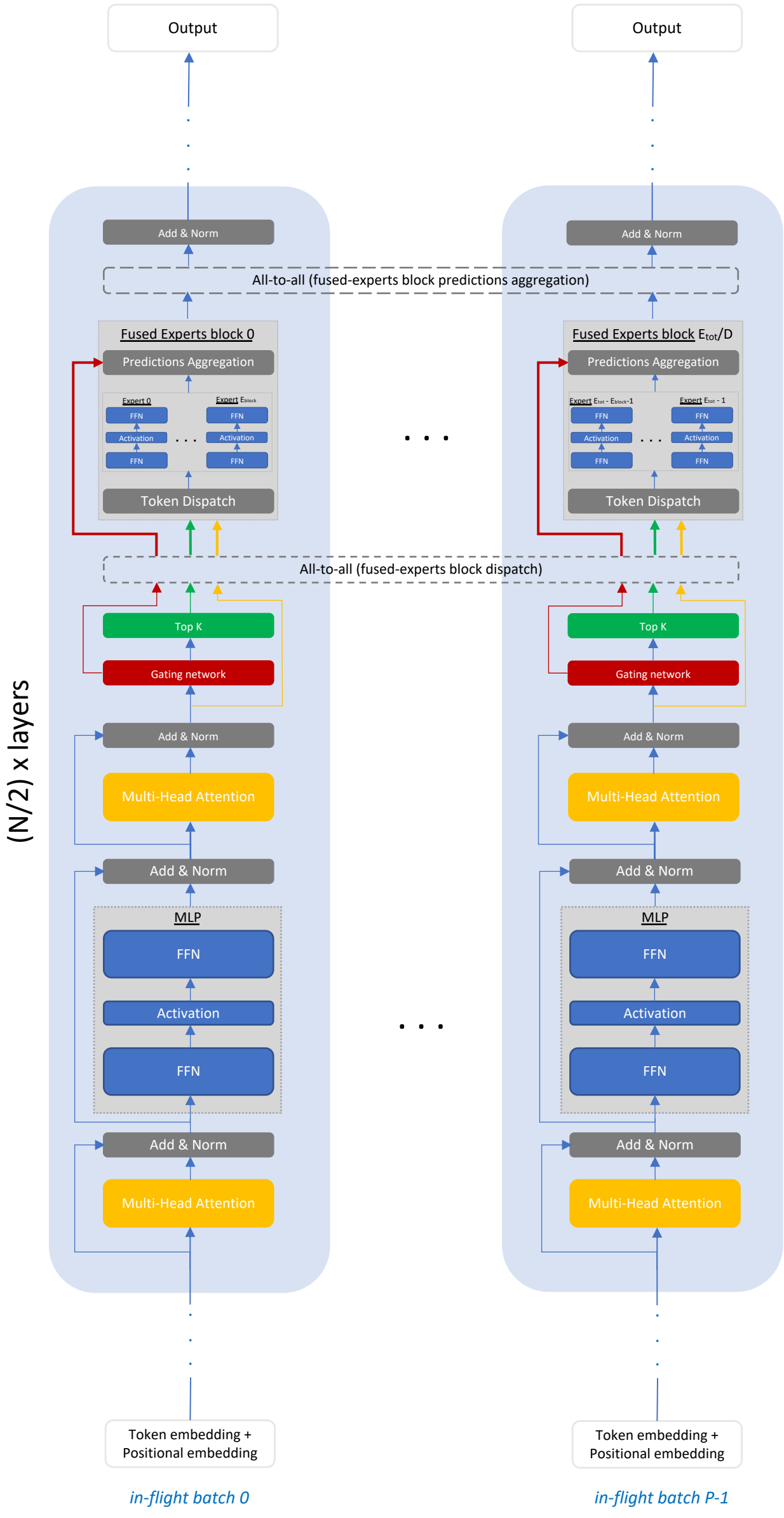
Output
Output
Add & Norm
Add & Norm
All-to-all (fused-experts block predictions aggregation)
Fused Experts block 0
Fused Experts block Etot/D
Predictions Aggregation
Predictions Aggregation
Expert 0
Expert Eblock
Expert Etot - Eblock-1
Expert Etot - 1
FFN
Activation
Token Dispatch
Token Dispatch
All-to-all (fused-experts block dispatch)
Top K
Gating network
Add & Norm
Multi-Head Attention
MLP
(N/2) x layers
(N/2) x layers
Token embedding + Positional embedding
Token embedding + Positional embedding
in-flight batch 0
in-flight batch P-1