

*Comp.  
Stream*

Gate

**Replication  
Planning**

**Reroute**

*Comm.  
Stream*

Notify  
Dispatch

**Weight  
Distribution**

Token Dispatch

 *Additional overhead upon standard MoE forward*