Tensor Parallelism

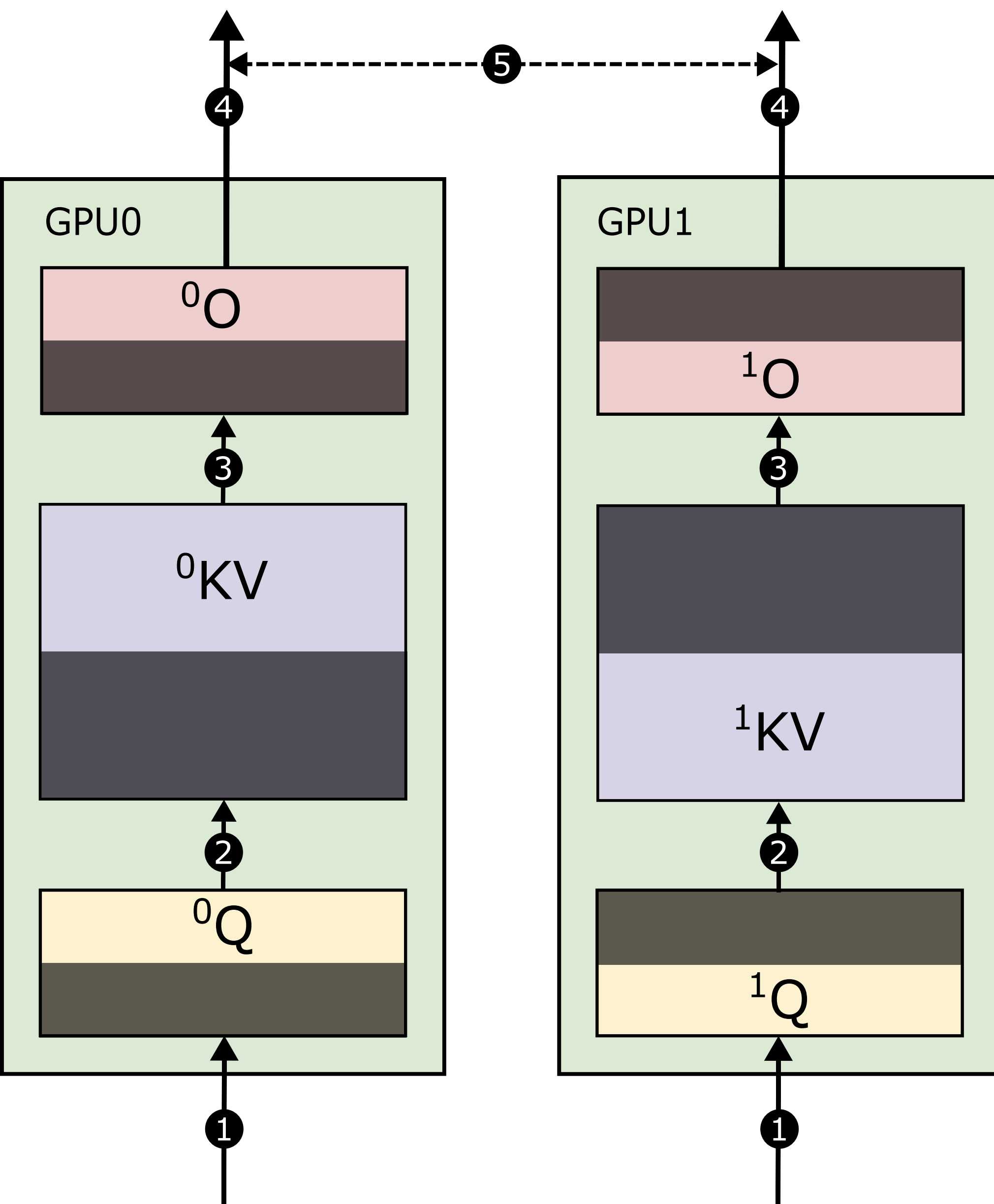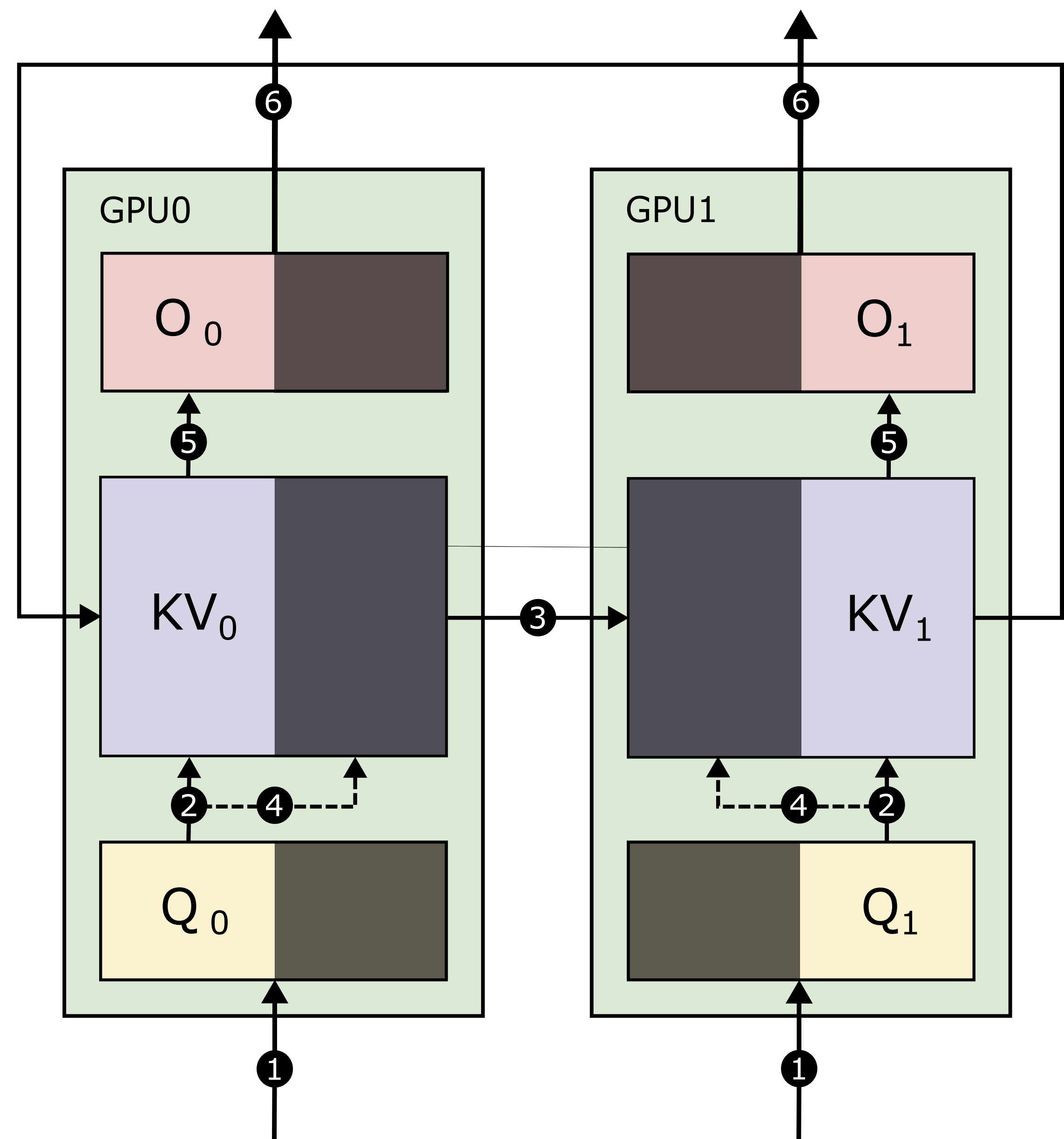**Ring Attention**

Sharded in Sequence Dimension

Sharded in Head Dimension