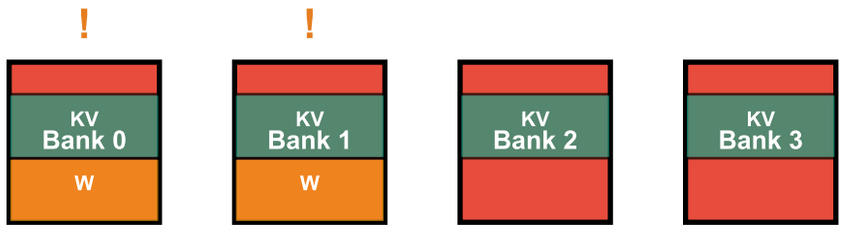
Four KV Cache Placement Policies

1. Naive (Baseline)

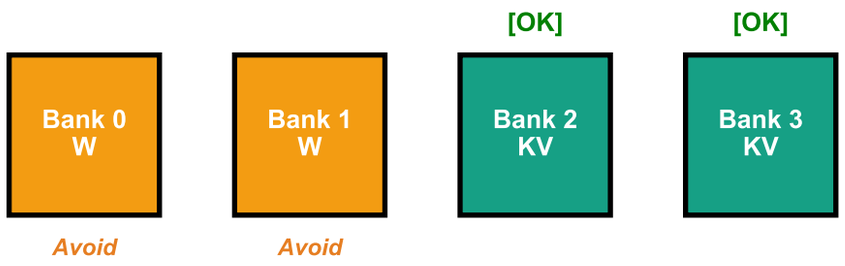


Round-robin allocation
Ignores weights

2. Bank Partitioning

Reserve banks for KV
Zero conflicts

3. Contention-Aware [BEST]



Smart placement
Avoids weight banks

4. Smart Locality

Scores banks
Considers locality