Key/value vectors

Shared Memory

HBM

Indices of vectors to be loaded — HBM

0 3 5 6 8 9 11 14 ●●●

Tile of key/value vectors

Key/value tokens to be loaded

☐ : Vector that should not be loaded

❶ Cooperatively load indices to shared memory

0 3 5 6 8 9 11 14 — **Shared Memory**

❷ Distribute indices to warps

❸: Distribute indices to all threads

↳ : Thread

Warp Group

Warp1 | 0 3
Warp2 | 5 6
Warp3 | 8 9
Warp4 | 11 14
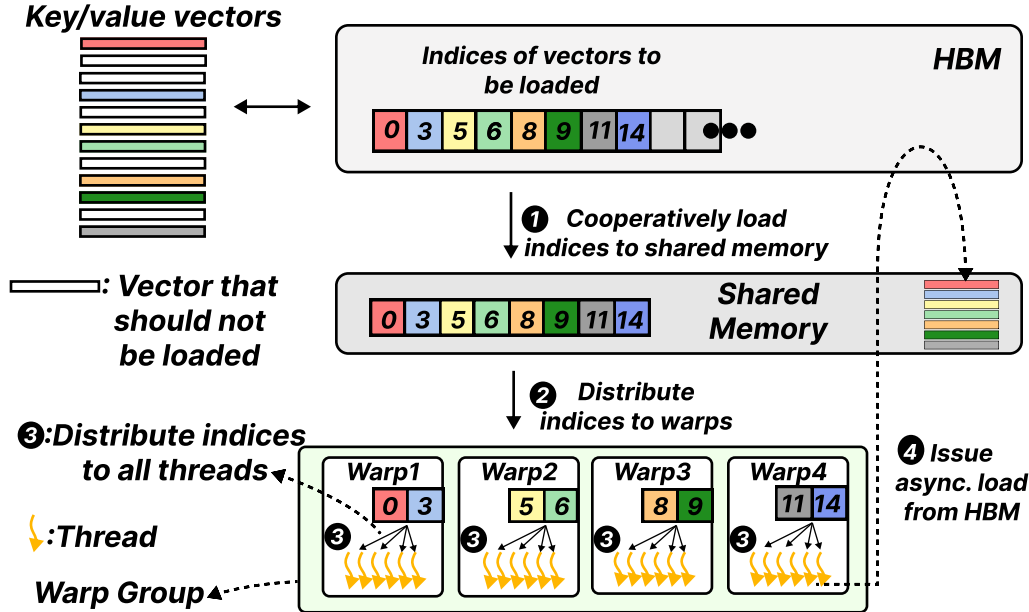
❹ Issue async. load from HBM

Functional depiction of our gather + load

Detailed diagram showing the implementation of gather + load operation