(a) Pre-compute and pre-allocation of large blocks

Output: "[batch, 12, seq_len, 64]"

Beforehand

Runtime
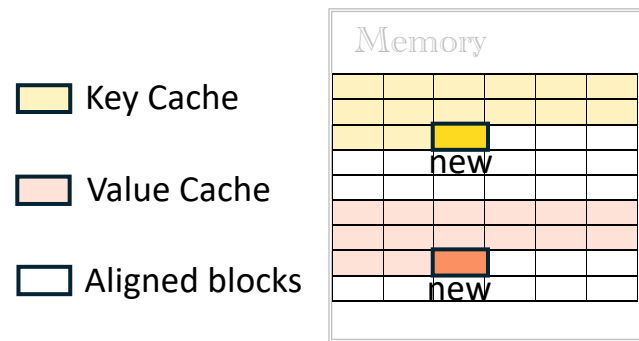
(b) Progressive-Allocation of Aligned Block
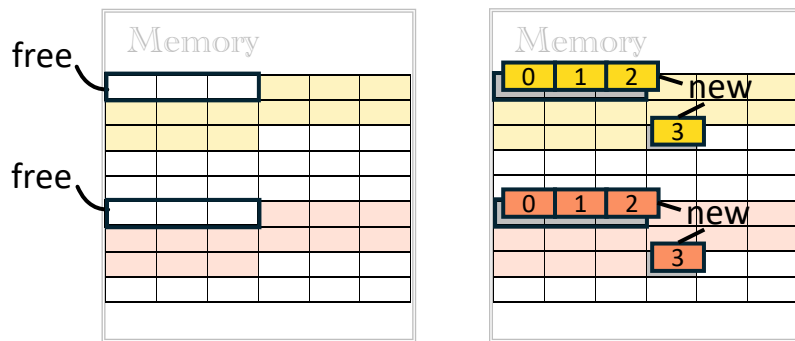
(c) Replacement of Consumed Allocations
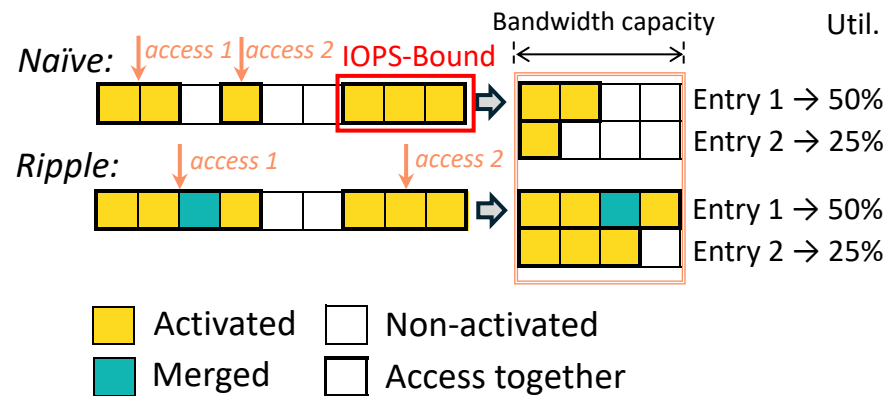
(d) Access Merge of Consecutive Memory