BW-TREE

Latch-free B+Tree index
→ Threads never need to set latches or block.

Key Idea #1: Deltas
→ No updates in place
→ Reduces cache invalidation.

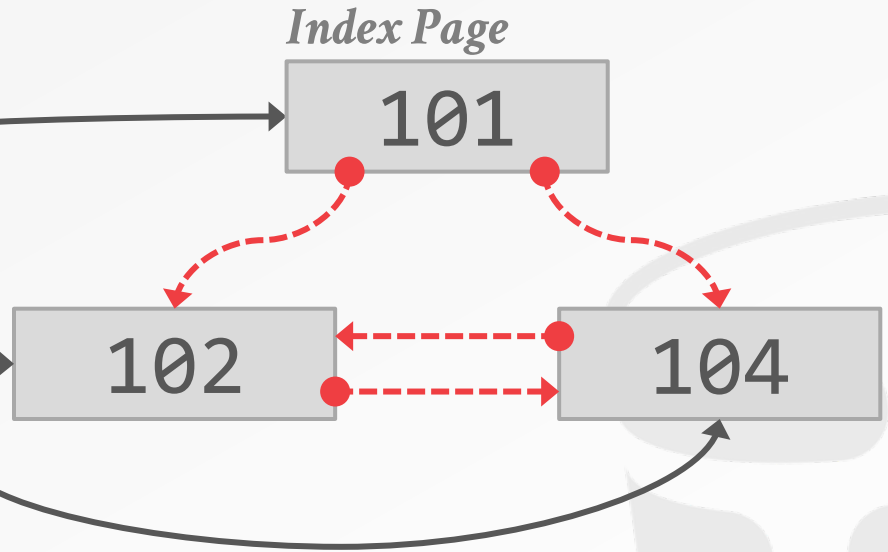
Key Idea #2: Mapping Table
→ Allows for CAS of physical locations of pages.

THE BW-TREE: A B-TREE FOR NEW HARDWARE
ICDE 2013

BW-TREE: MAPPING TABLE

Mapping Table

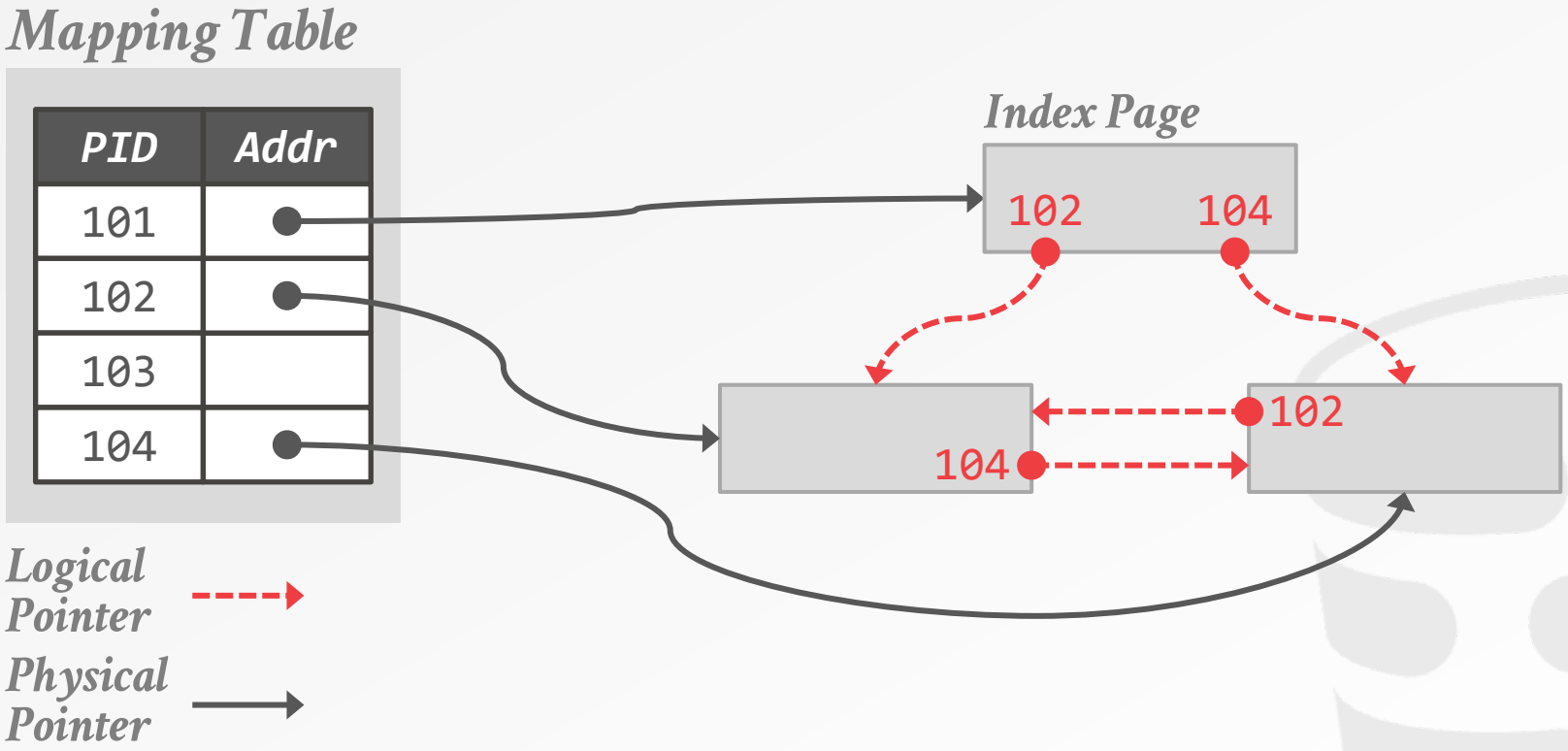
PID	Addr
101	●
102	●
103	
104	●



Logical Pointer ----→

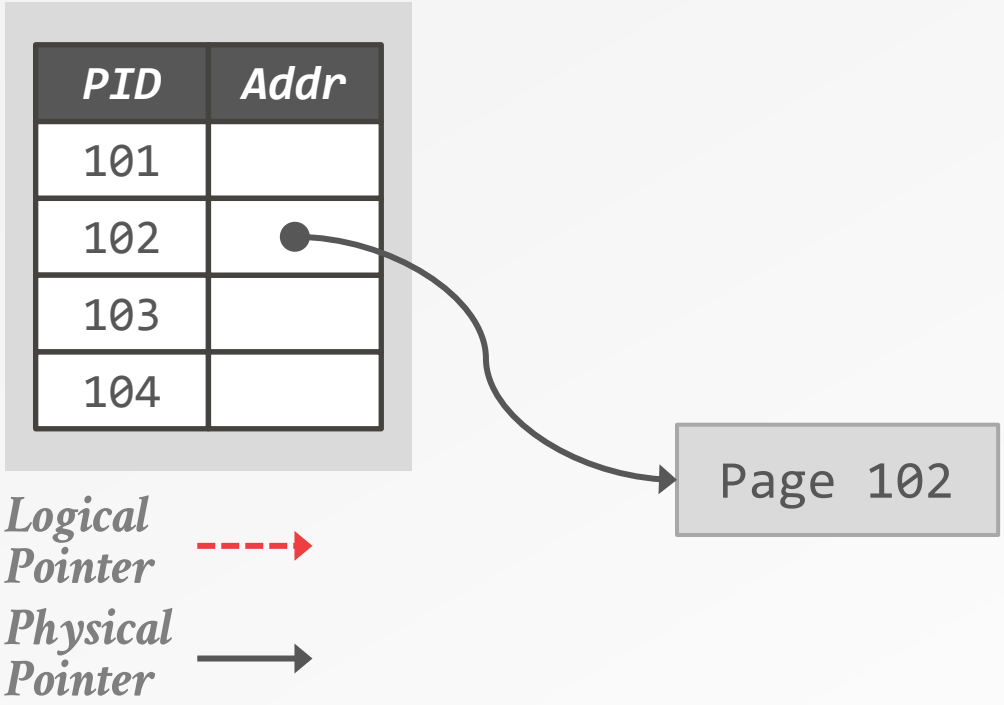
Physical Pointer ——→

BW-TREE: MAPPING TABLE



BW-TREE: DELTA UPDATES

Mapping Table



Each update to a page produces a new delta.

BW-TREE: DELTA UPDATES

Mapping Table

PID	Addr
101	
102	
103	
104	

Logical Pointer ----▸

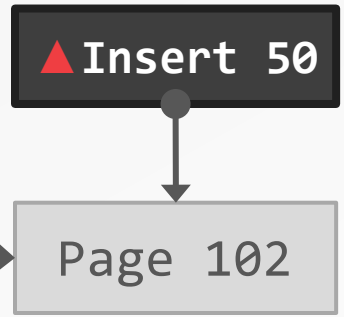
Physical Pointer ——▸

Each update to a page produces a new delta.

BW-TREE: DELTA UPDATES

Mapping Table

PID	Addr
101	
102	
103	
104	



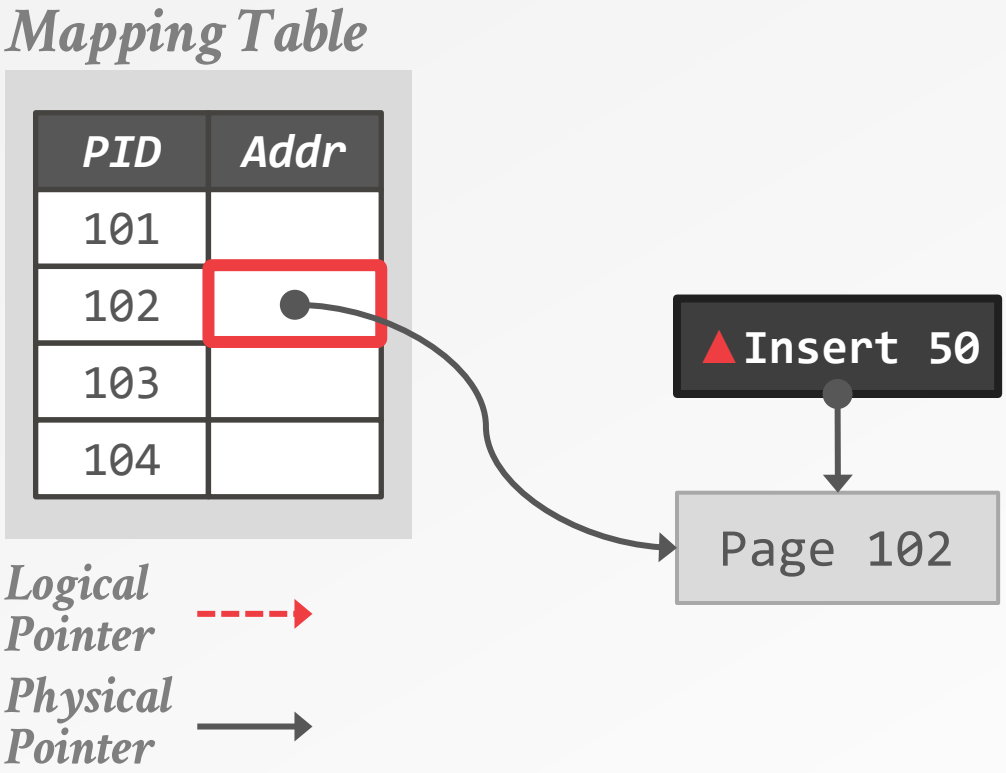
Logical Pointer ----▸

Physical Pointer ——▸

Each update to a page produces a new delta.

Delta physically points to base page.

BW-TREE: DELTA UPDATES



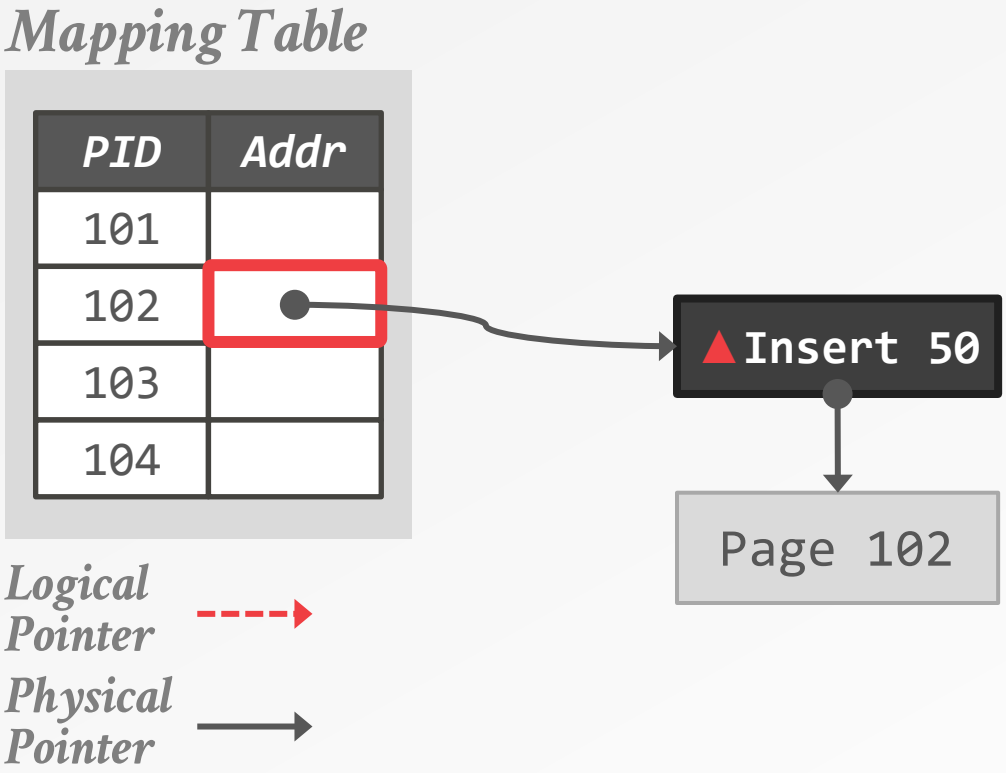
Each update to a page produces a new delta.

Delta physically points to base page.

Install delta address in physical address slot of mapping table using CAS.

Source: Justin Levandoski

BW-TREE: DELTA UPDATES



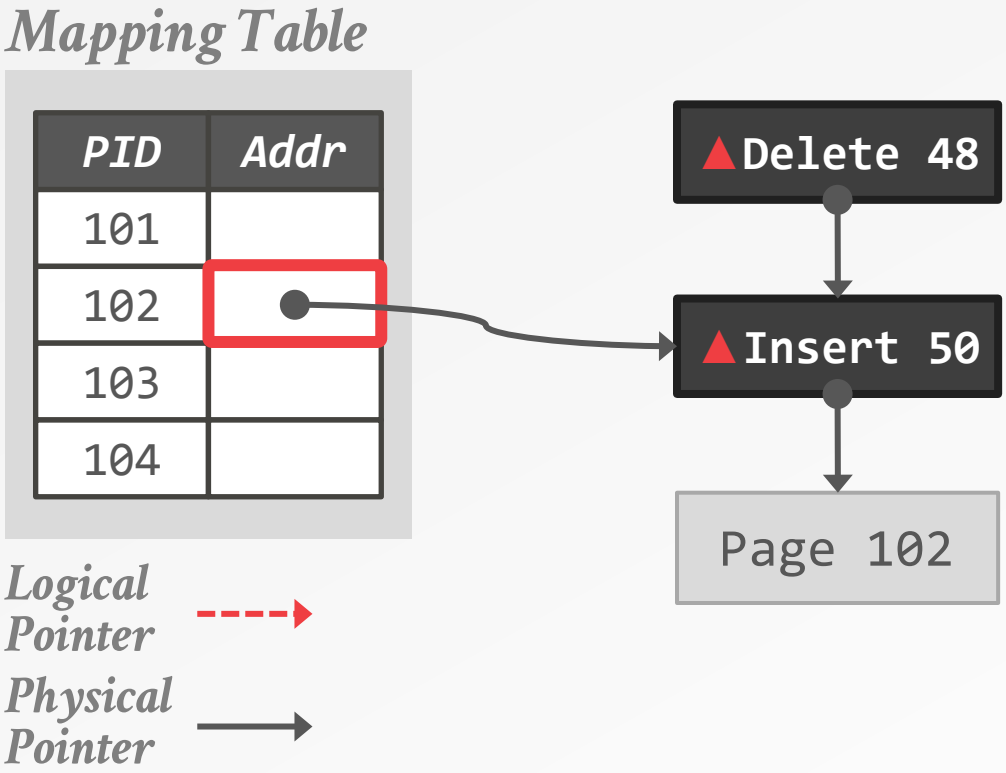
Each update to a page produces a new delta.

Delta physically points to base page.

Install delta address in physical address slot of mapping table using CAS.

Source: Justin Levandoski

BW-TREE: DELTA UPDATES



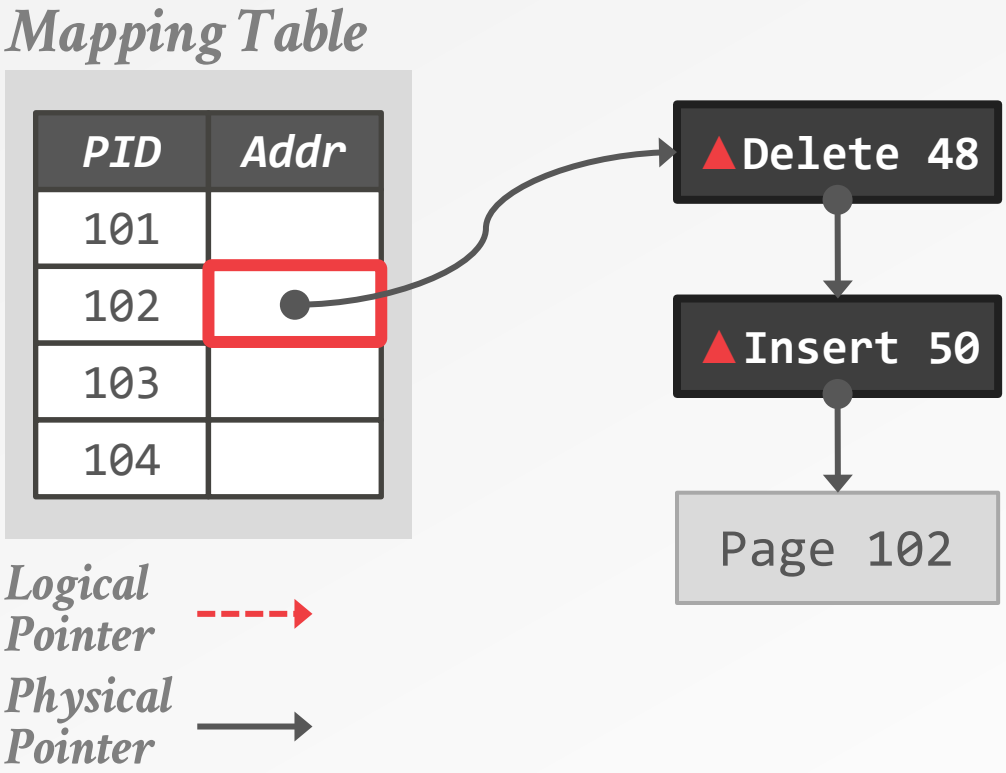
Each update to a page produces a new delta.

Delta physically points to base page.

Install delta address in physical address slot of mapping table using CAS.

Source: Justin Levandoski

BW-TREE: DELTA UPDATES



Each update to a page produces a new delta.

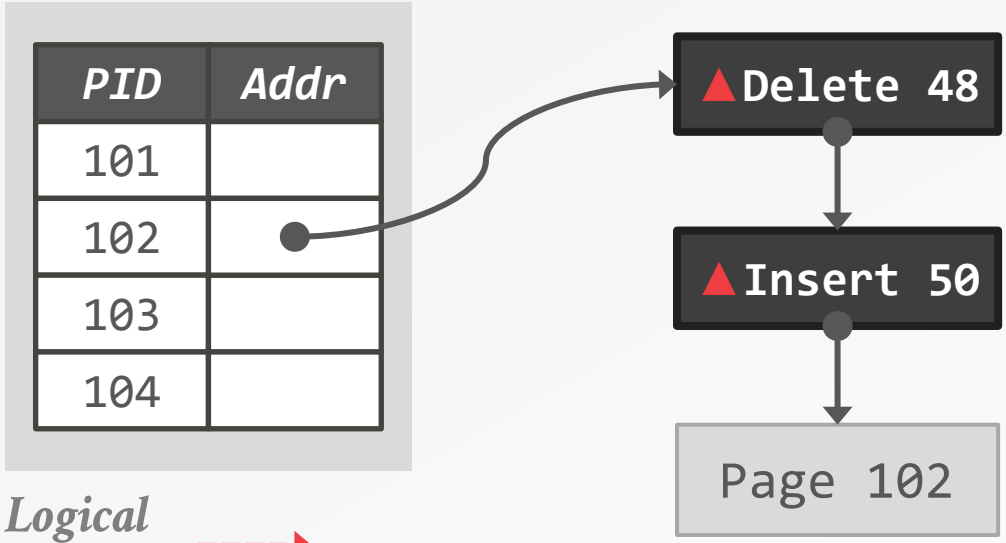
Delta physically points to base page.

Install delta address in physical address slot of mapping table using CAS.

Source: Justin Levandoski

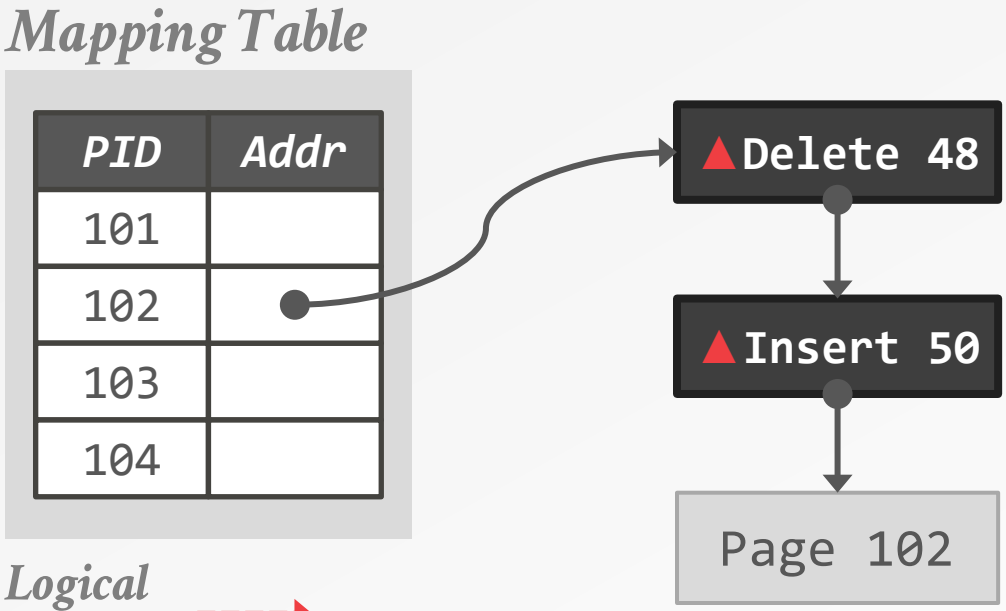
BW-TREE: SEARCH

Mapping Table



Traverse tree like a regular B+tree.

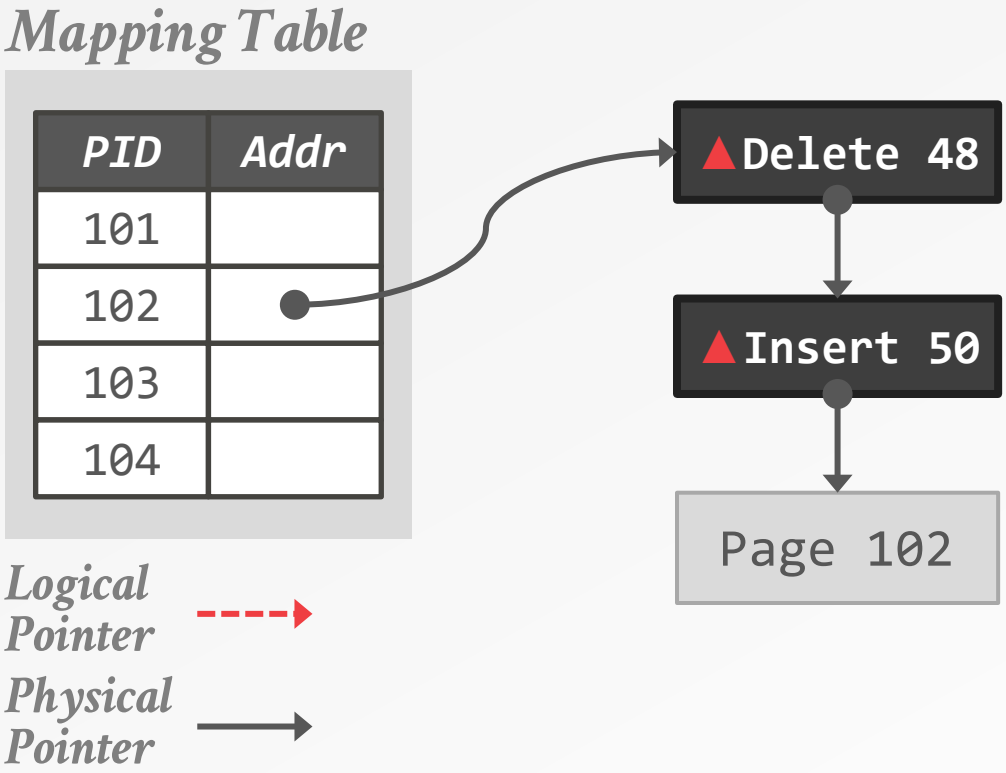
BW-TREE: SEARCH



Traverse tree like a regular B+tree.

If mapping table points to delta chain, stop at first occurrence of search key.

BW-TREE: SEARCH

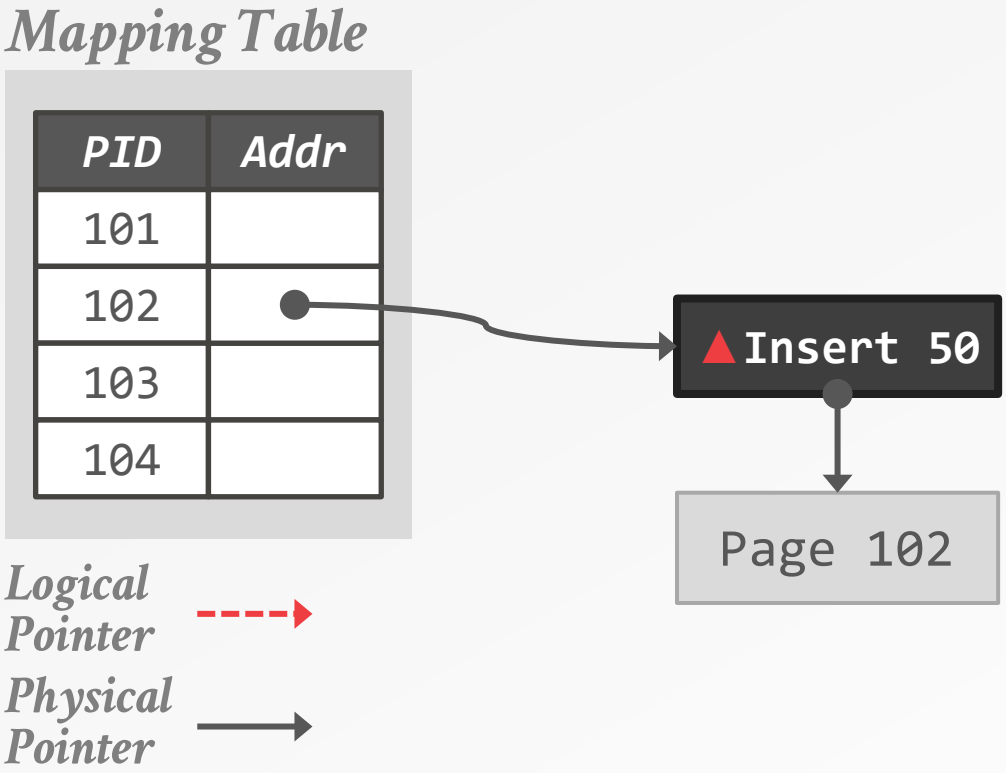


Traverse tree like a regular B+tree.

If mapping table points to delta chain, stop at first occurrence of search key.

Otherwise, perform binary search on base page.

BW-TREE: CONTENTION UPDATES

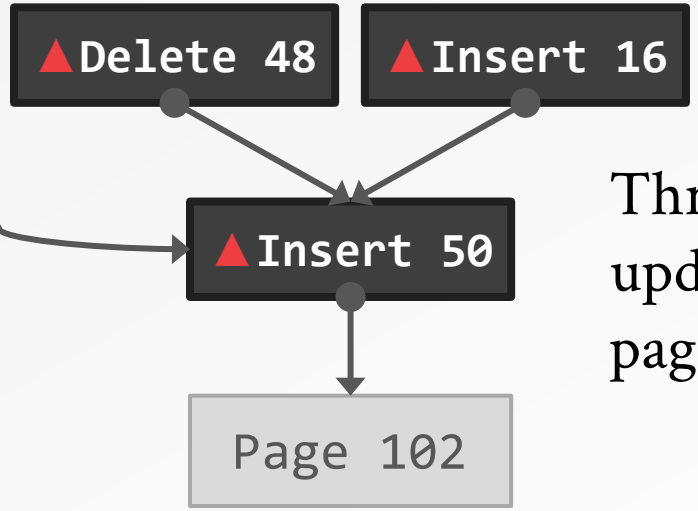


Threads may try to install updates to same state of the page.

BW-TREE: CONTENTION UPDATES

Mapping Table

PID	Addr
101	
102	
103	
104	

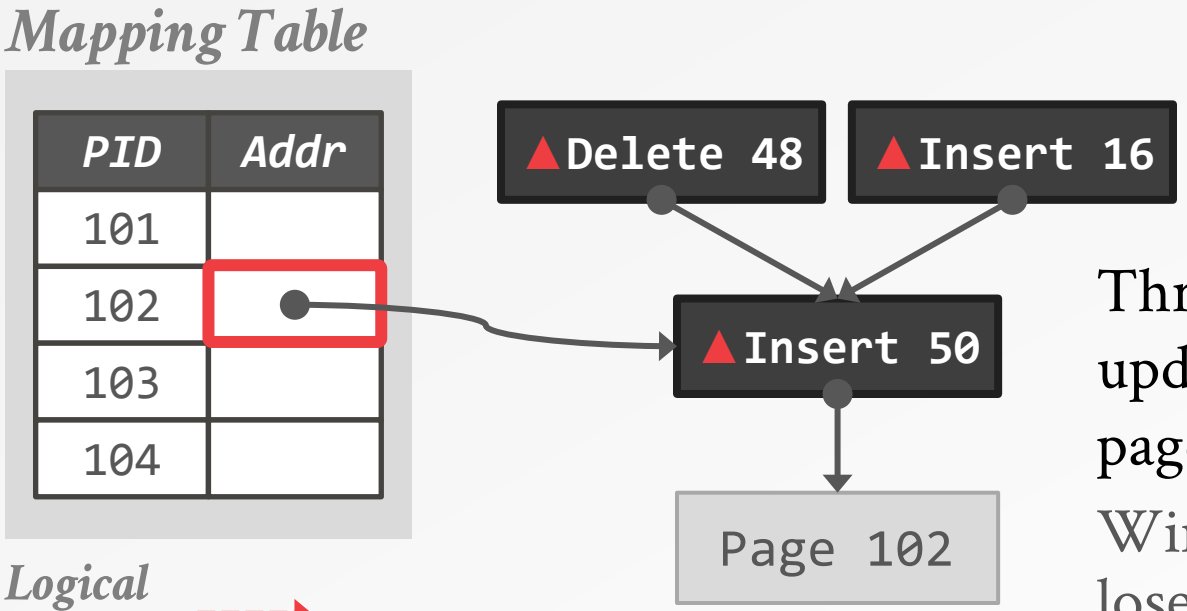


Threads may try to install updates to same state of the page.

Logical Pointer ----▸

Physical Pointer ——▸

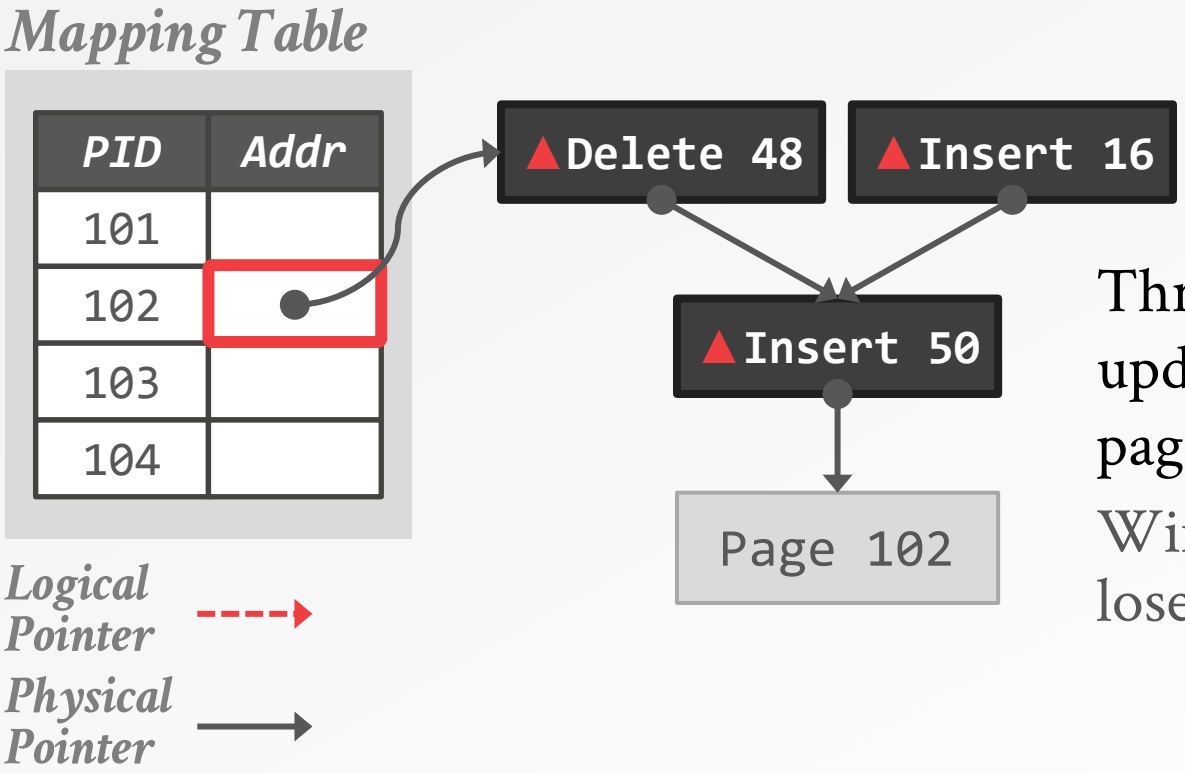
BW-TREE: CONTENTION UPDATES



Threads may try to install updates to same state of the page.

Winner succeeds, any losers must retry or abort

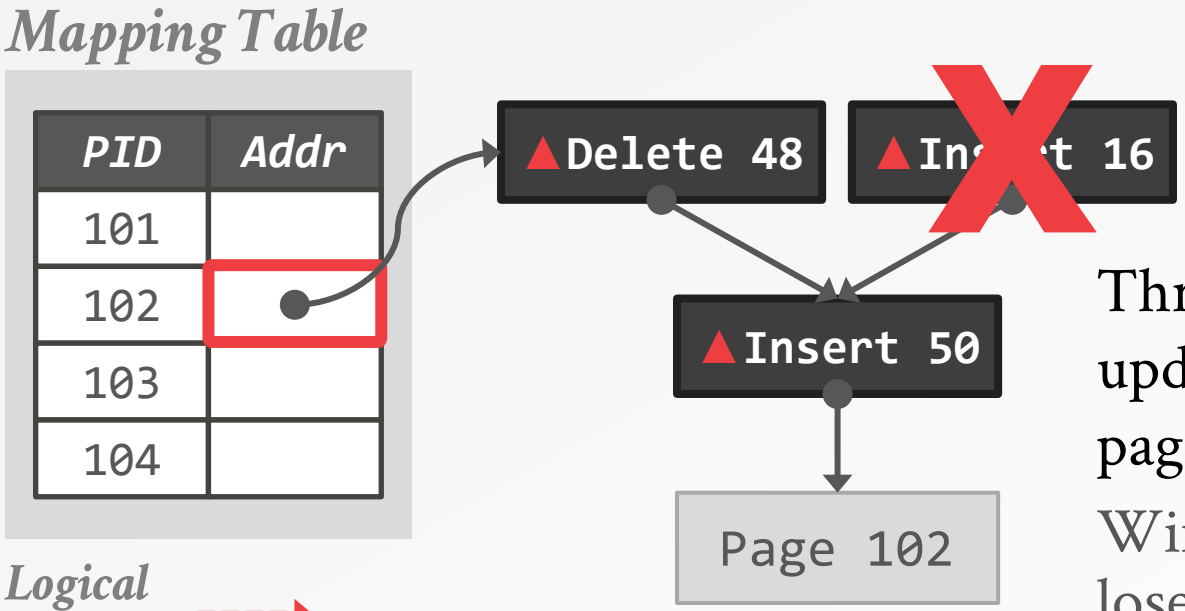
BW-TREE: CONTENTION UPDATES



Threads may try to install updates to same state of the page.

Winner succeeds, any losers must retry or abort

BW-TREE: CONTENTION UPDATES



Threads may try to install updates to same state of the page.

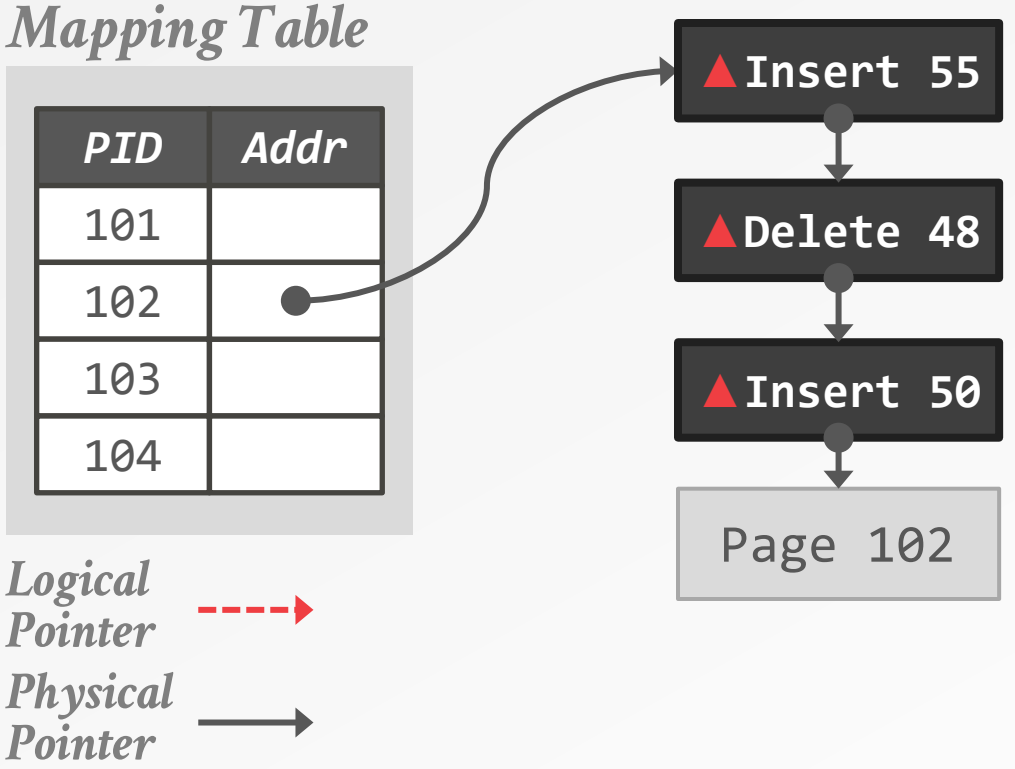
Winner succeeds, any losers must retry or abort

BW-TREE: DELTA TYPES

Record Update Deltas
→ Insert/Delete/Update of record on a page

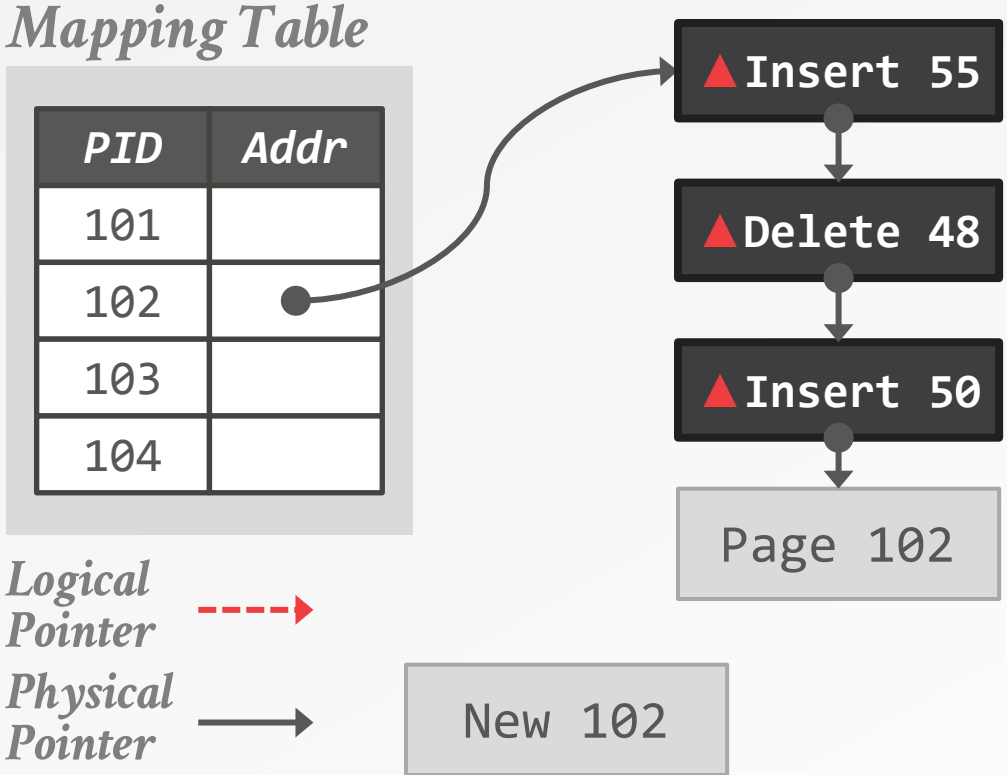
Structure Modification Deltas
→ Split/Merge information

BW-TREE: CONSOLIDATION



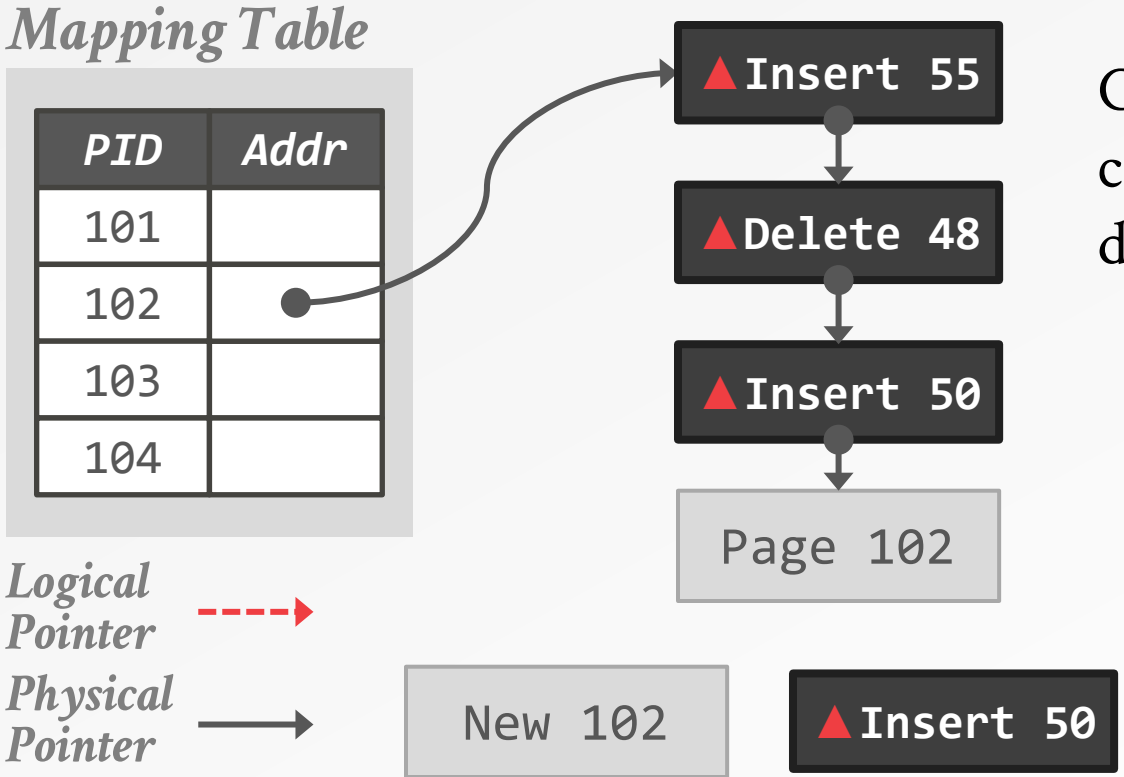
Consolidate updates by creating new page with deltas applied.

BW-TREE: CONSOLIDATION



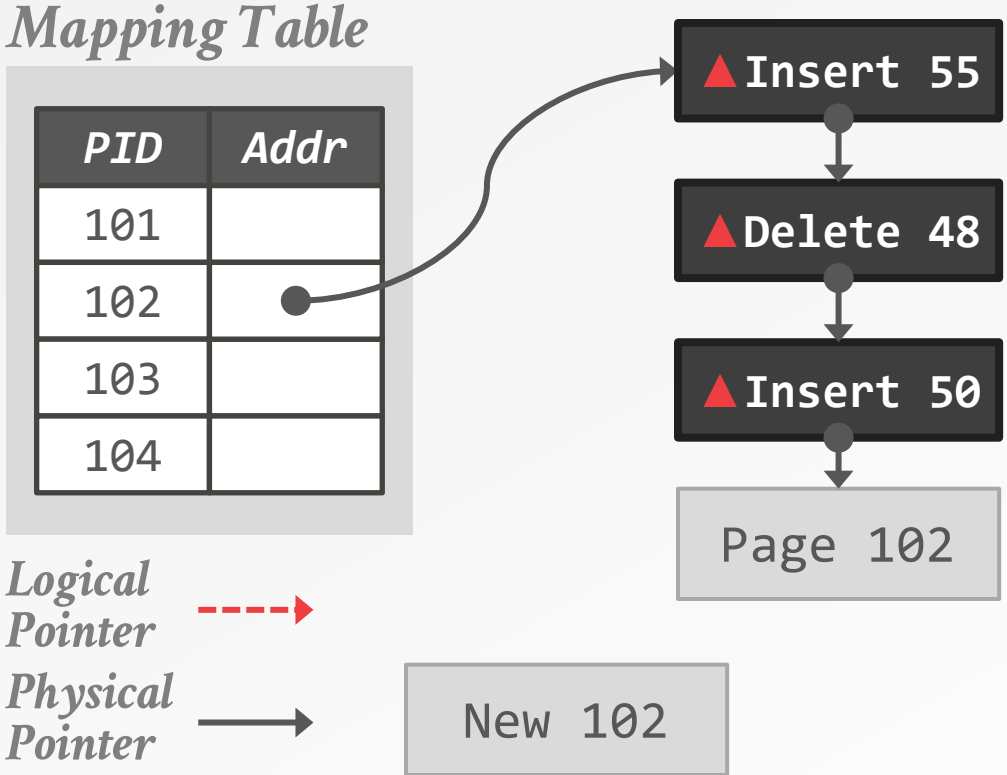
Consolidate updates by creating new page with deltas applied.

BW-TREE: CONSOLIDATION



Consolidate updates by creating new page with deltas applied.

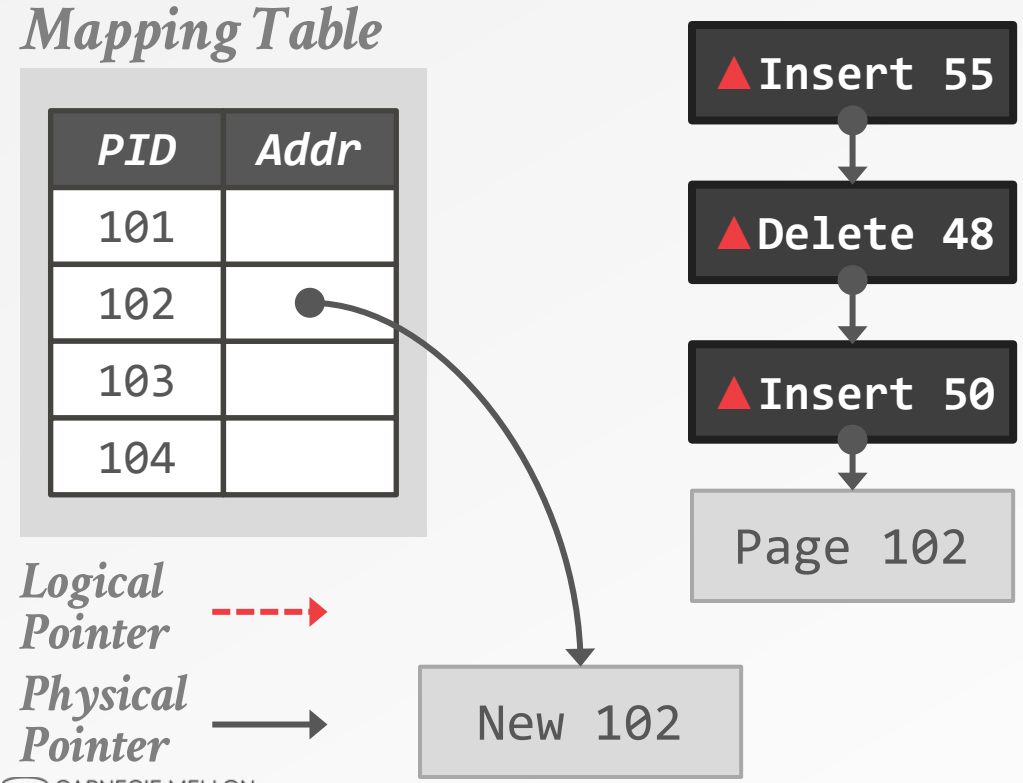
BW-TREE: CONSOLIDATION



Consolidate updates by creating new page with deltas applied.

CAS-ing the mapping table address ensures no deltas are missed.

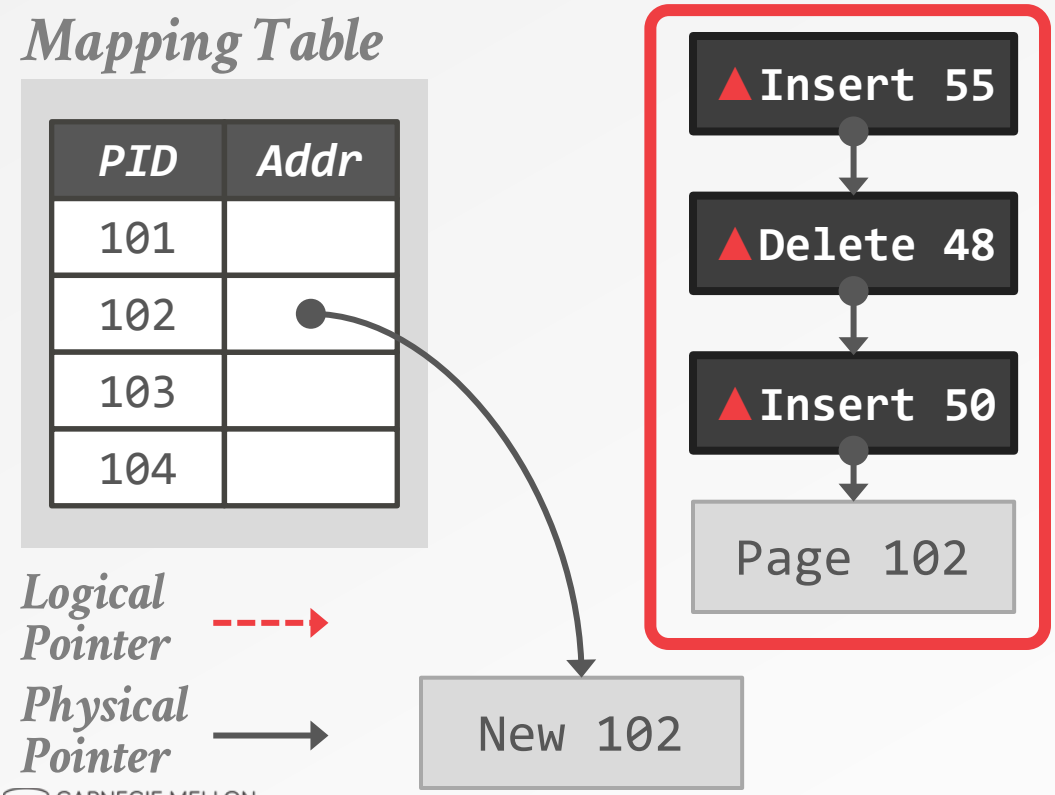
BW-TREE: CONSOLIDATION



Consolidate updates by creating new page with deltas applied.

CAS-ing the mapping table address ensures no deltas are missed.

BW-TREE: CONSOLIDATION



Consolidate updates by creating new page with deltas applied.

CAS-ing the mapping table address ensures no deltas are missed.

Old page + deltas are marked as garbage.

BW-TREE: STRUCTURE MODIFICATIONS

Split Delta Record

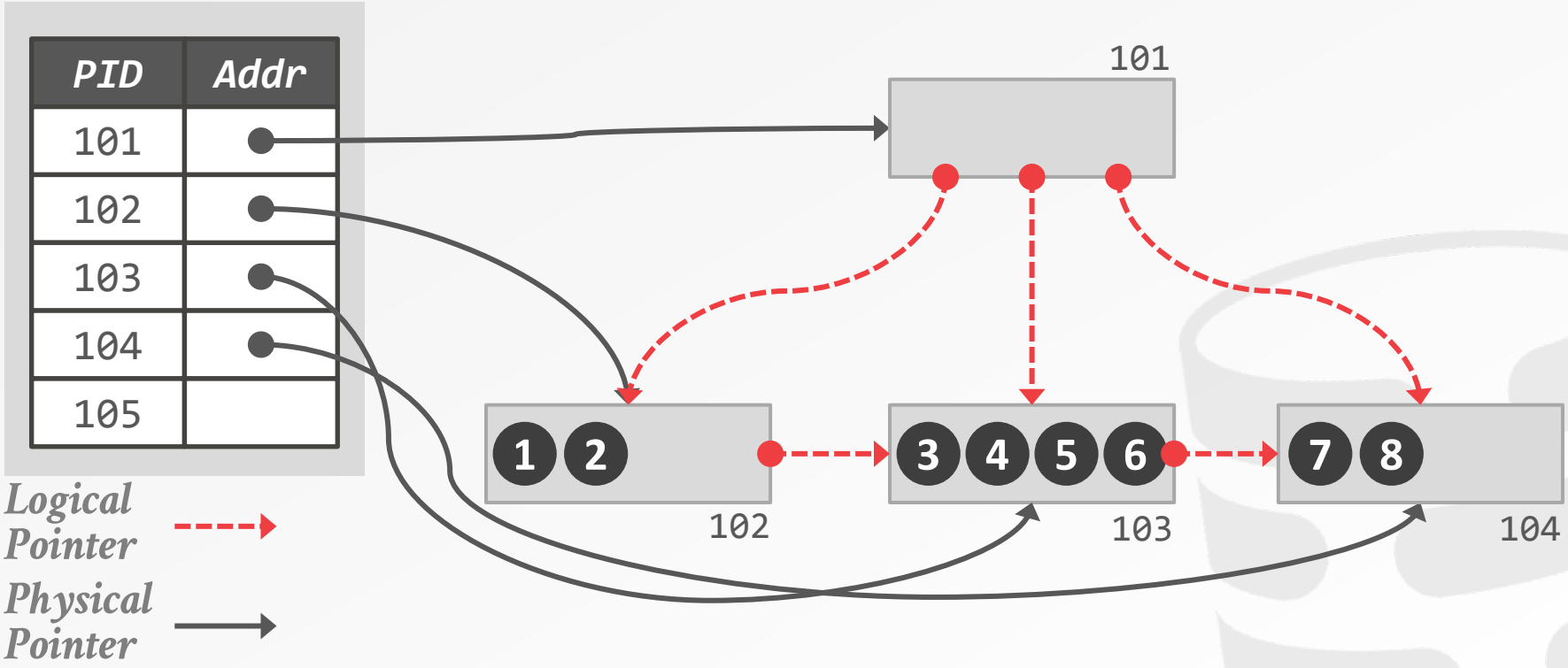
→ Mark that a subset of the base page's key range is now located at another page.
→ Use a logical pointer to the new page.

Separator Delta Record

→ Provide a shortcut in the modified page's parent on what ranges to find the new page.

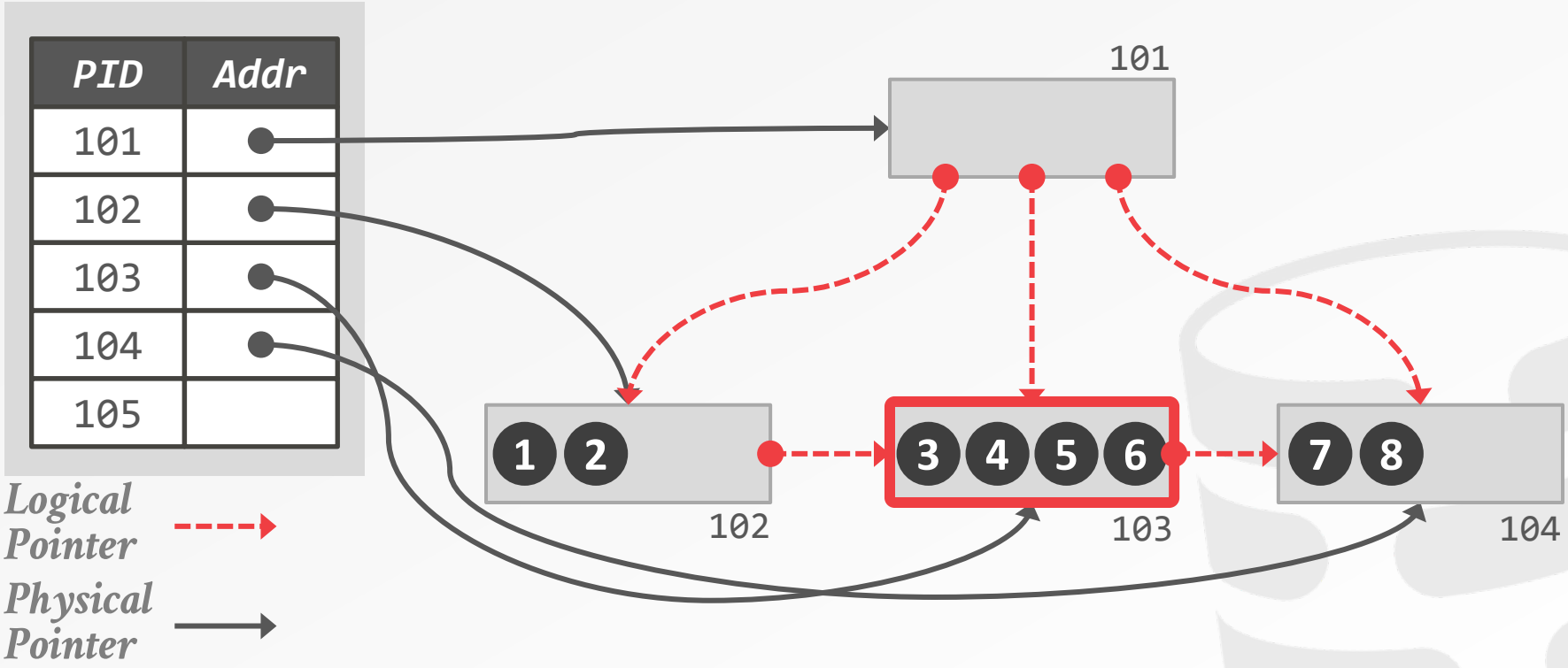
BW-TREE: STRUCTURE MODIFICATIONS

Mapping Table



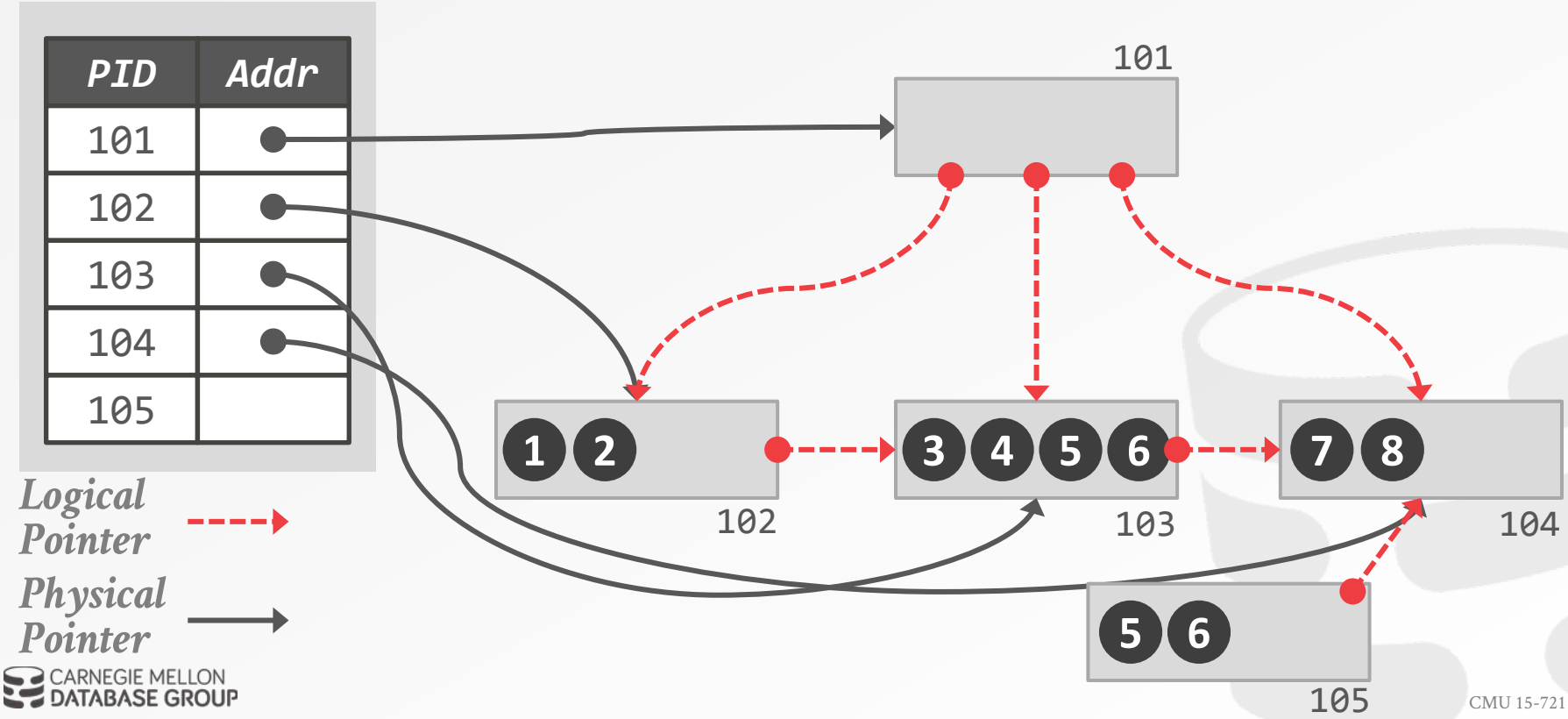
BW-TREE: STRUCTURE MODIFICATIONS

Mapping Table



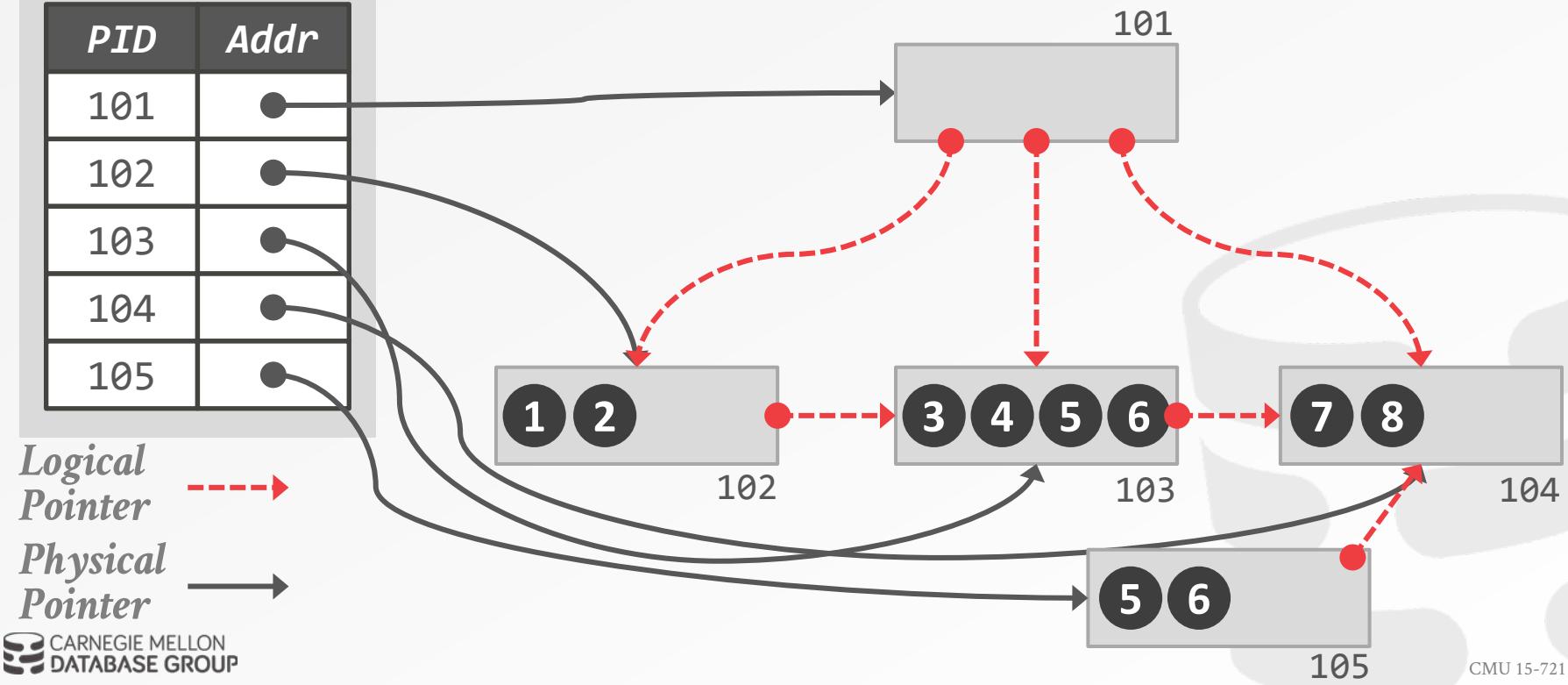
BW-TREE: STRUCTURE MODIFICATIONS

Mapping Table



BW-TREE: STRUCTURE MODIFICATIONS

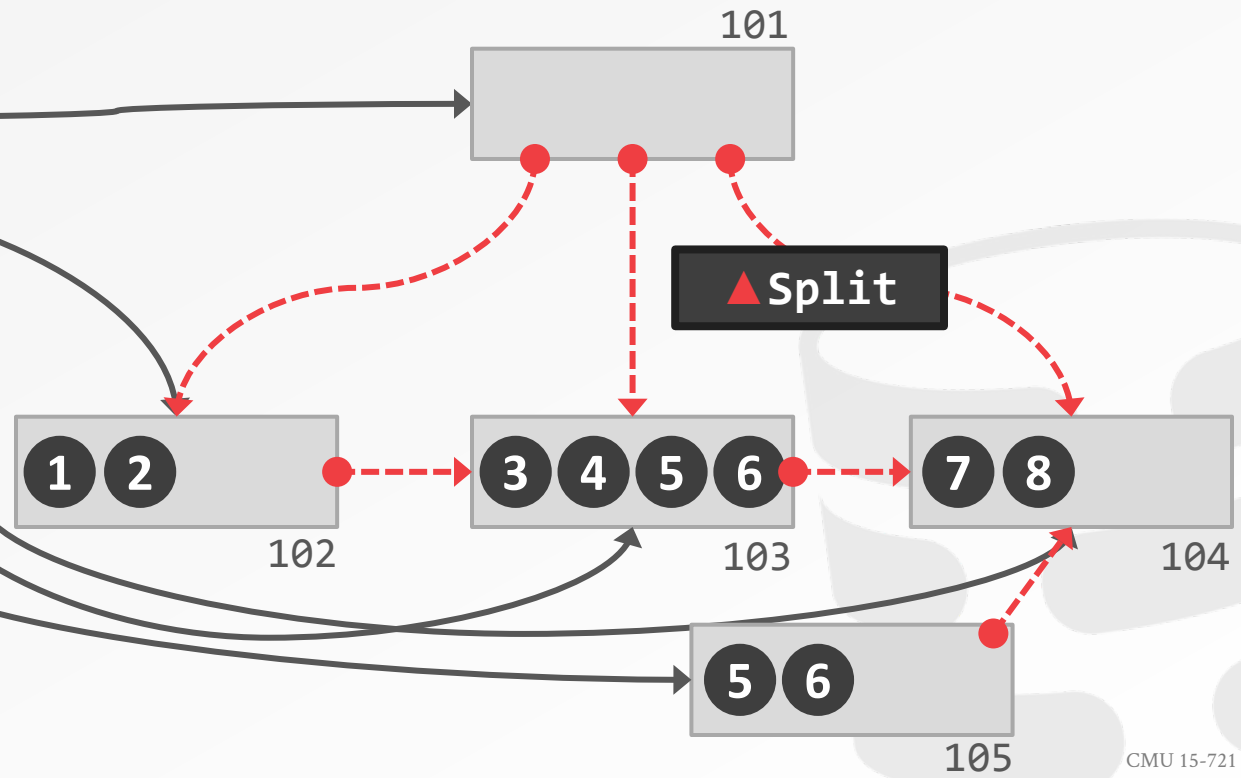
Mapping Table



BW-TREE: STRUCTURE MODIFICATIONS

Mapping Table

PID	Addr
101	●
102	●
103	●
104	●
105	●

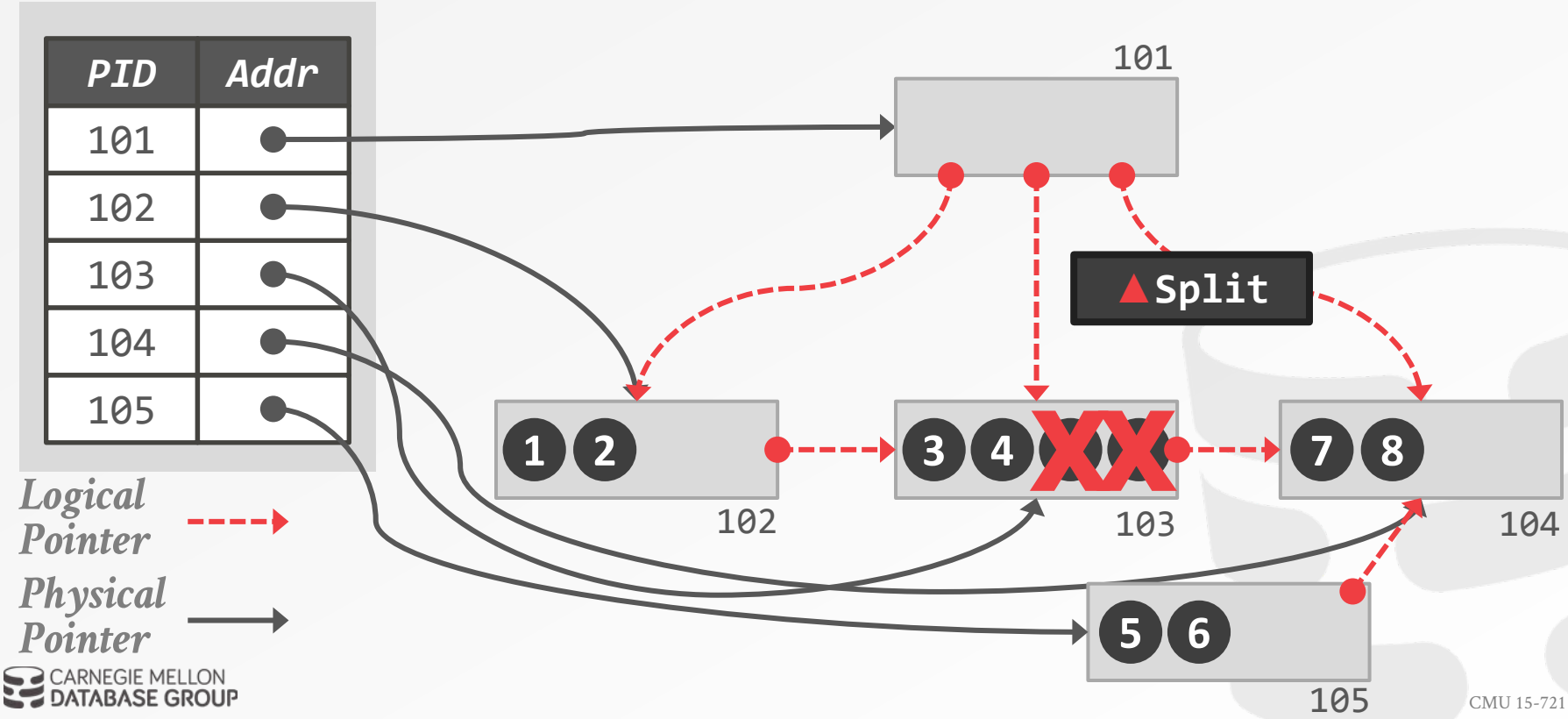


Logical Pointer ----→

Physical Pointer ——→

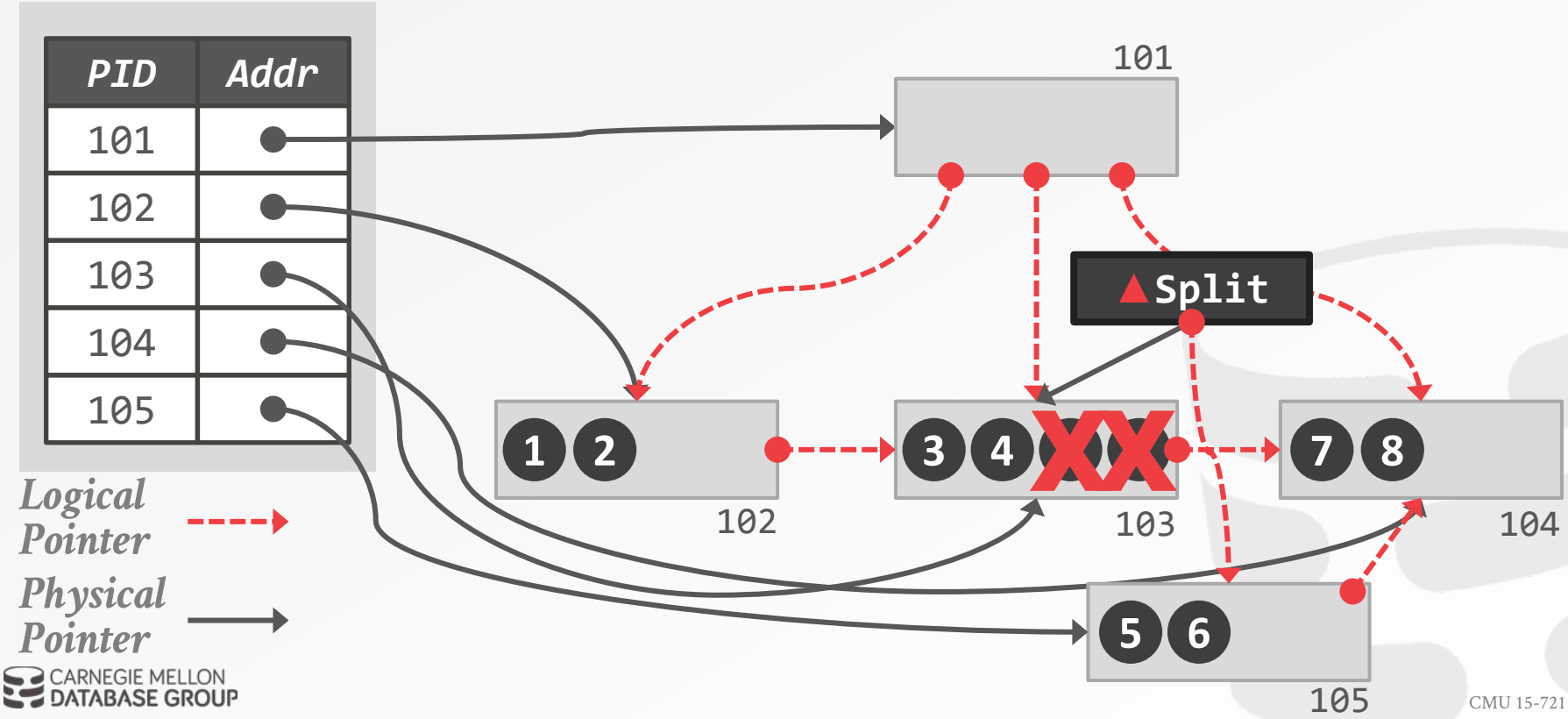
BW-TREE: STRUCTURE MODIFICATIONS

Mapping Table



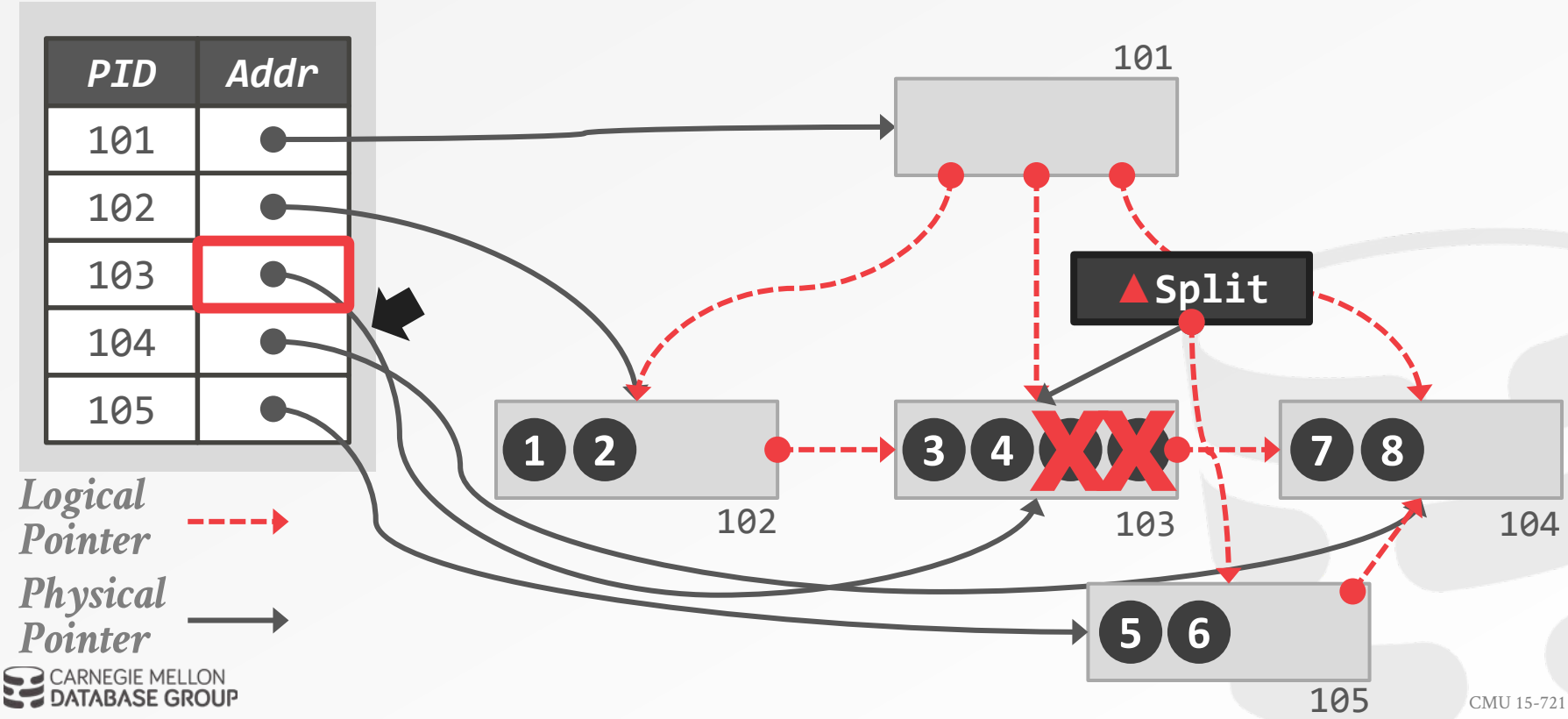
BW-TREE: STRUCTURE MODIFICATIONS

Mapping Table



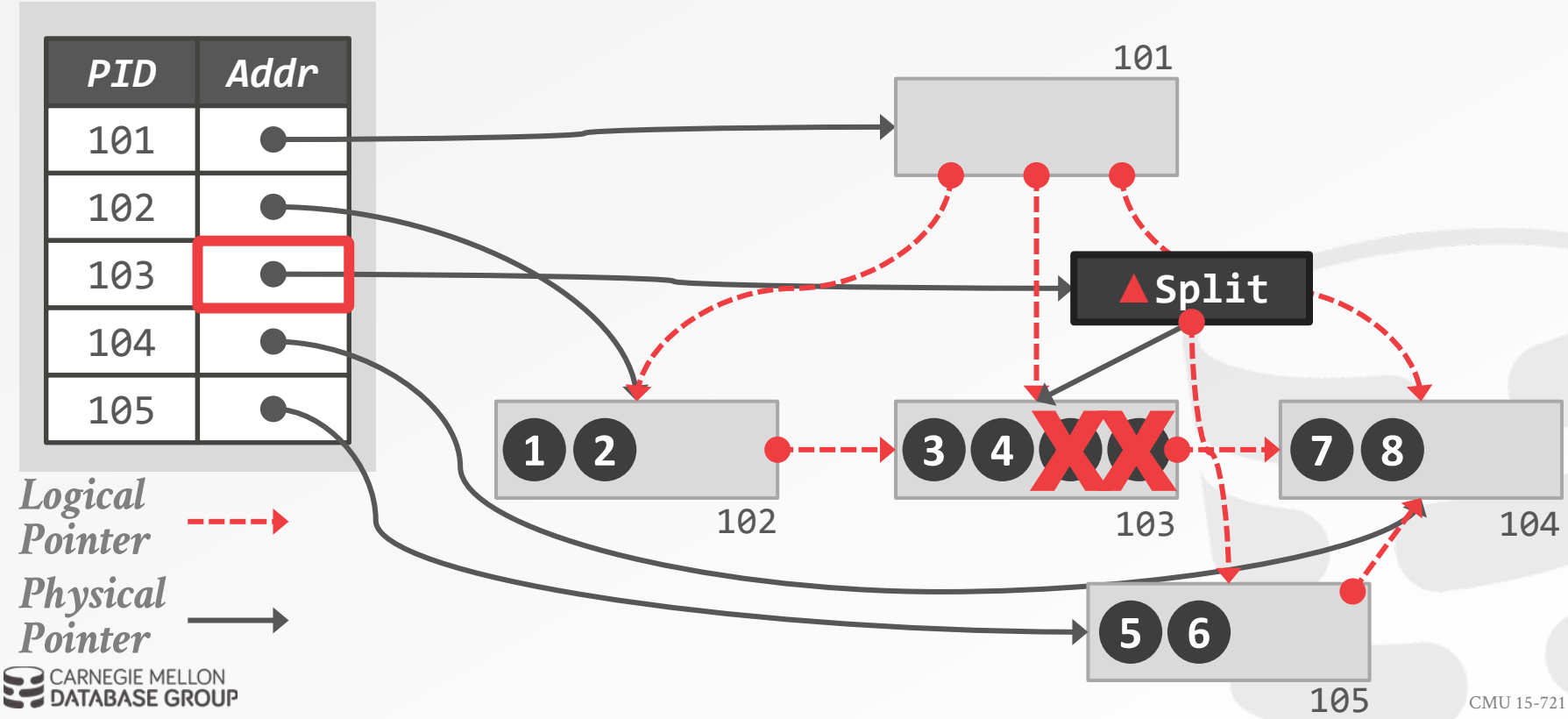
BW-TREE: STRUCTURE MODIFICATIONS

Mapping Table



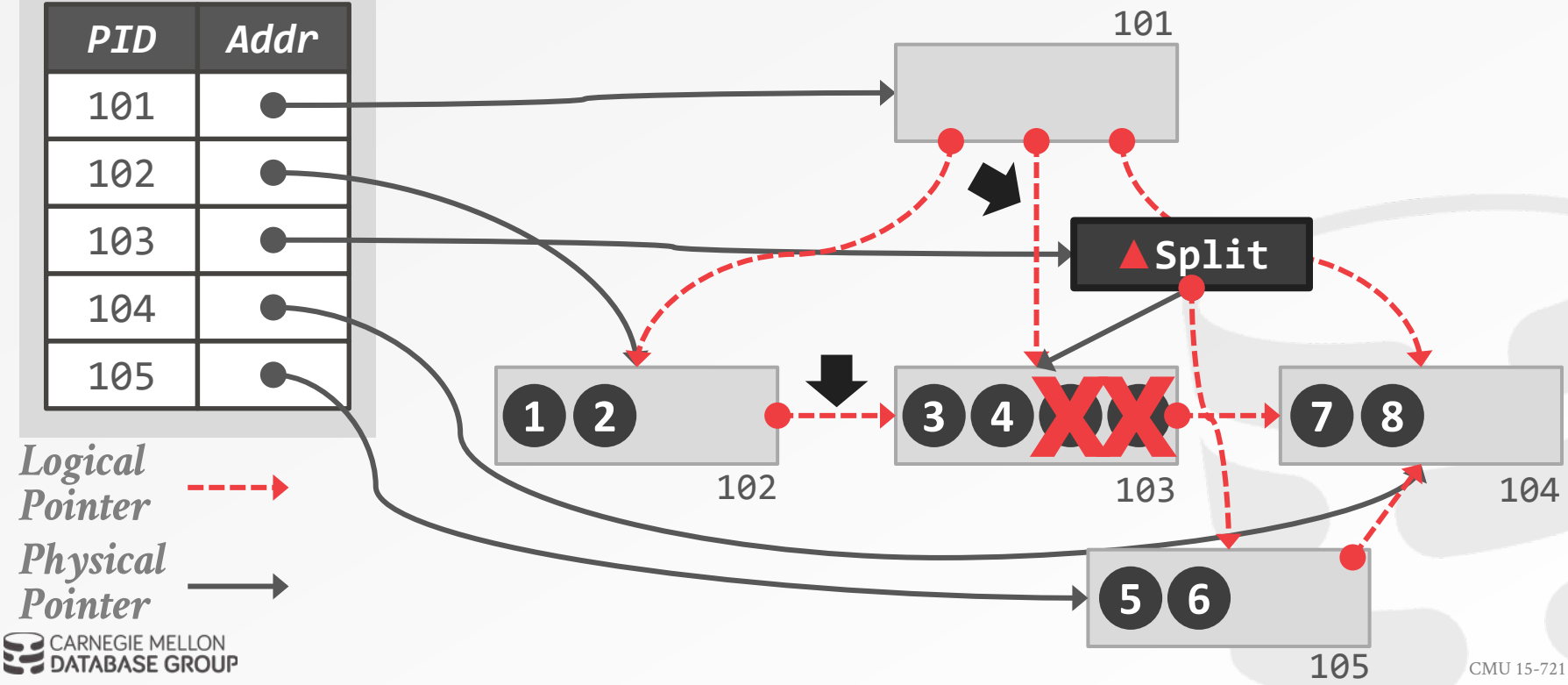
BW-TREE: STRUCTURE MODIFICATIONS

Mapping Table



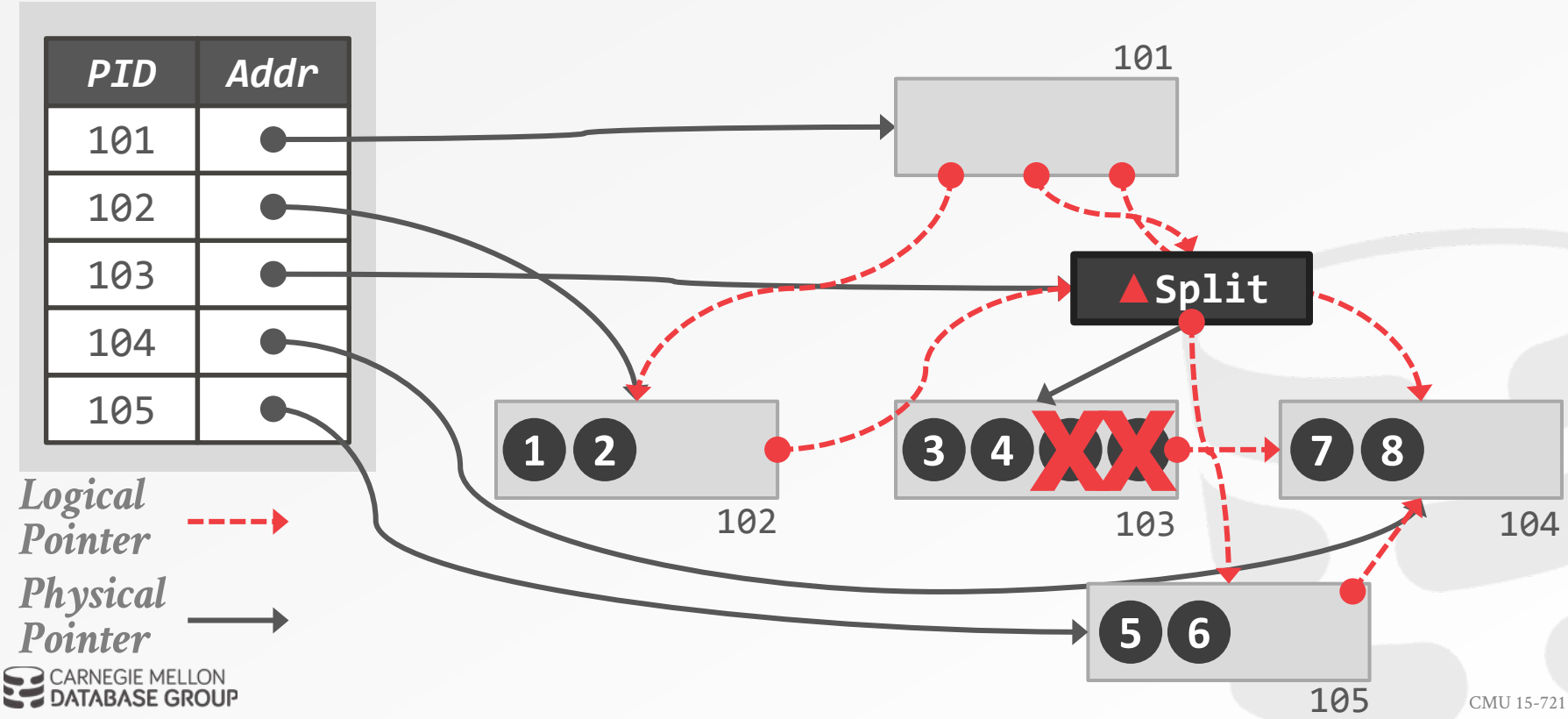
BW-TREE: STRUCTURE MODIFICATIONS

Mapping Table



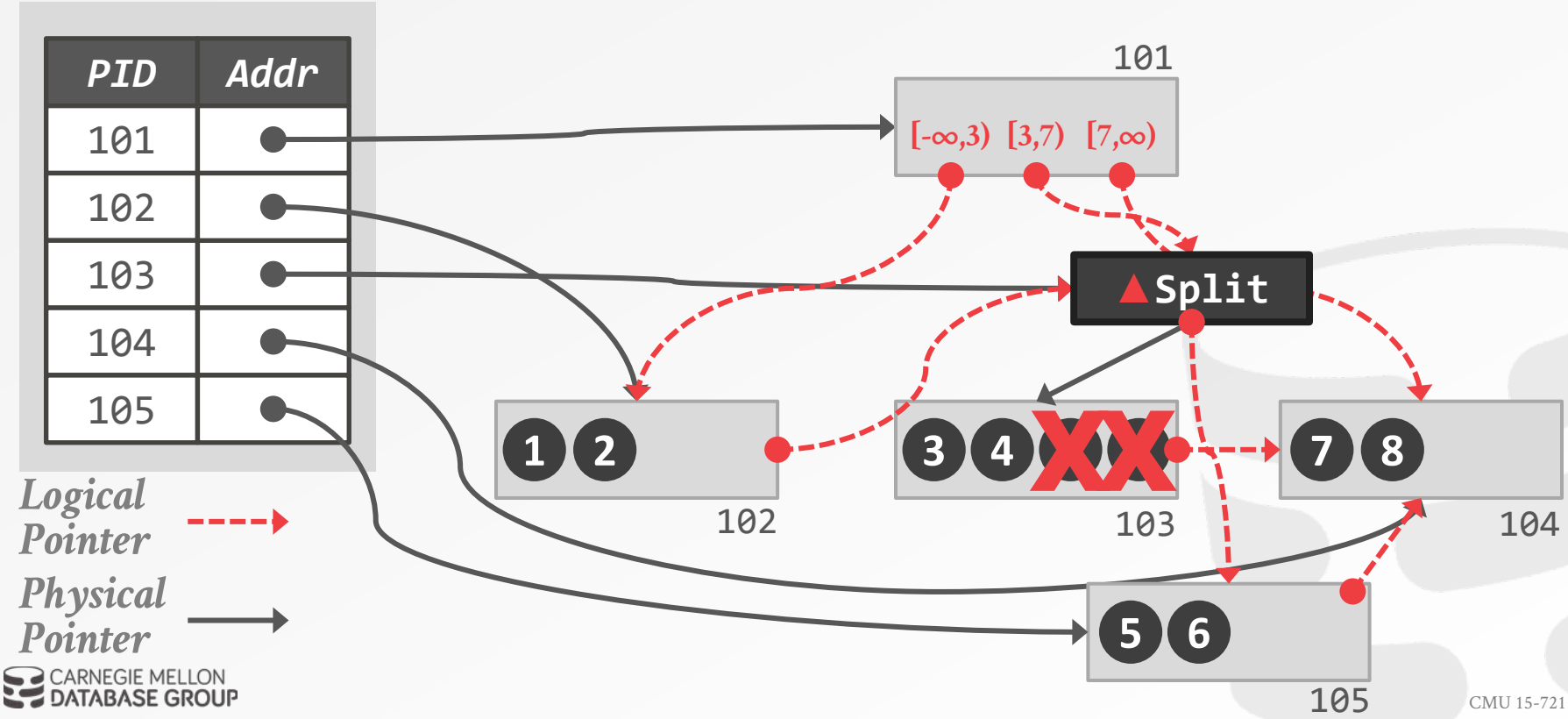
BW-TREE: STRUCTURE MODIFICATIONS

Mapping Table



BW-TREE: STRUCTURE MODIFICATIONS

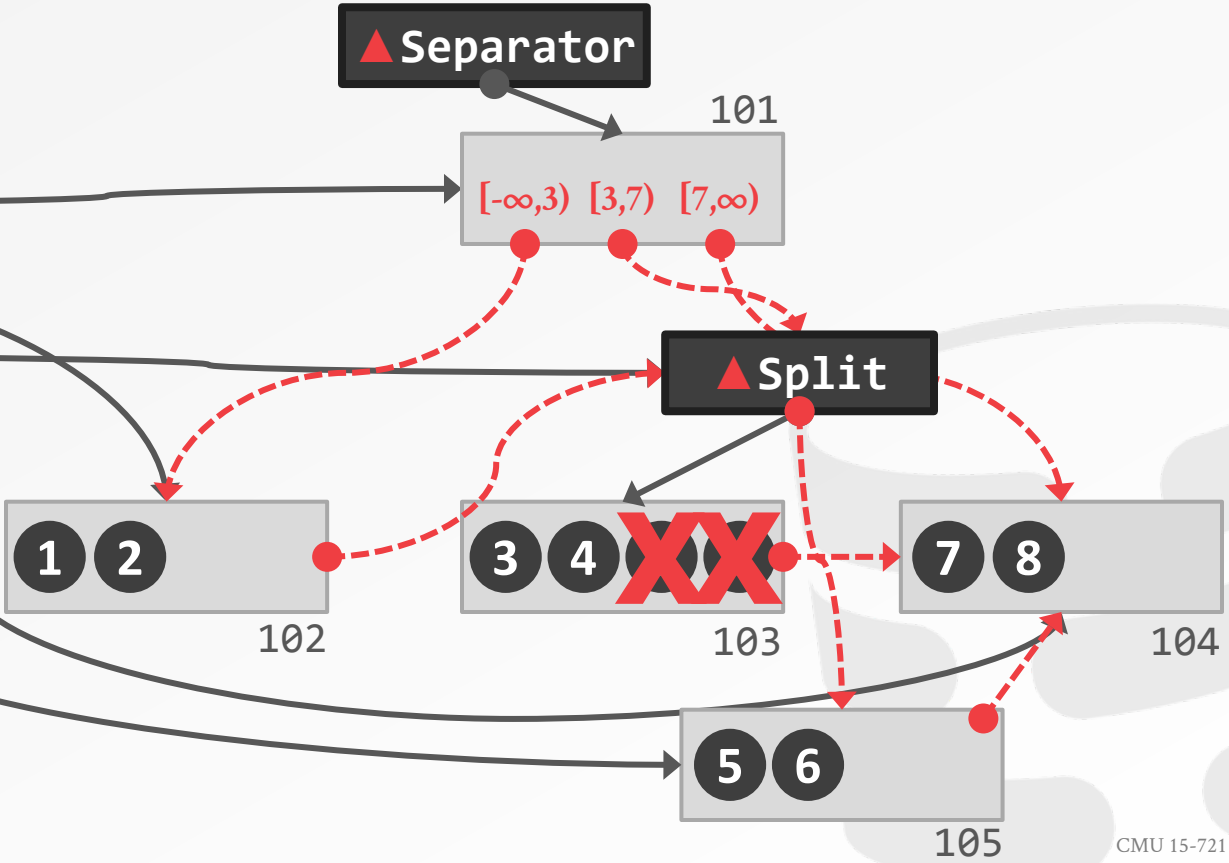
Mapping Table



BW-TREE: STRUCTURE MODIFICATIONS

Mapping Table

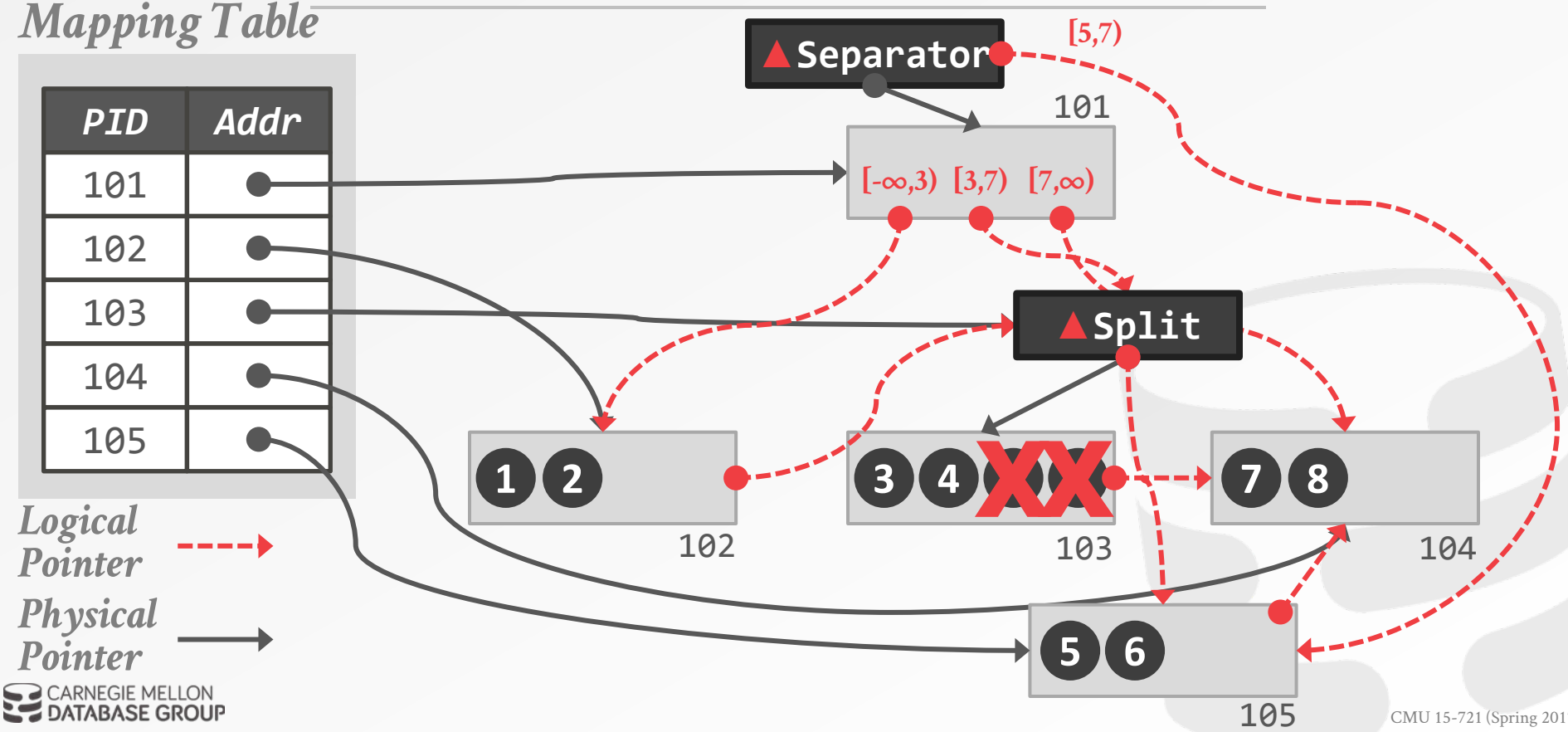
PID	Addr
101	
102	
103	
104	
105	



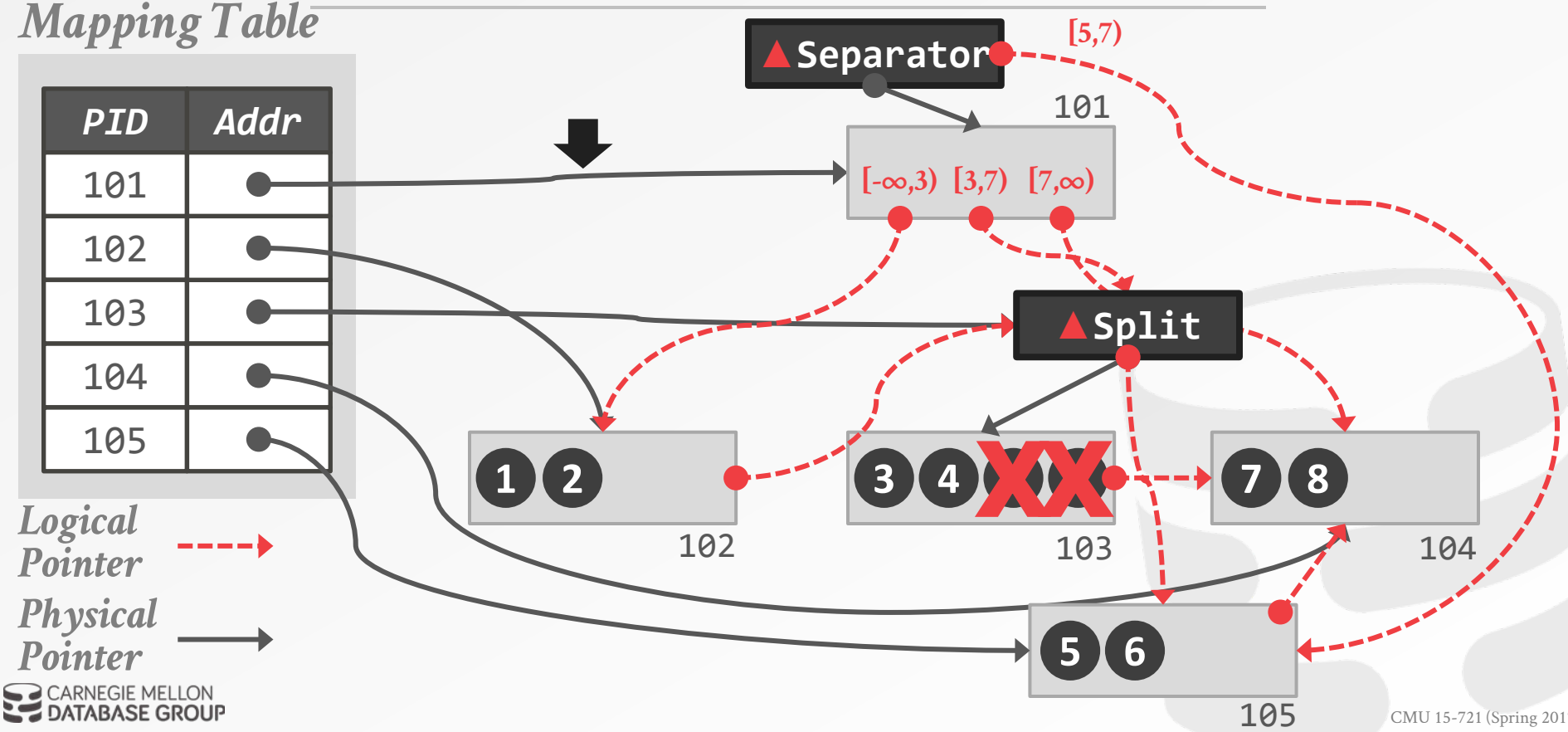
Logical Pointer

Physical Pointer

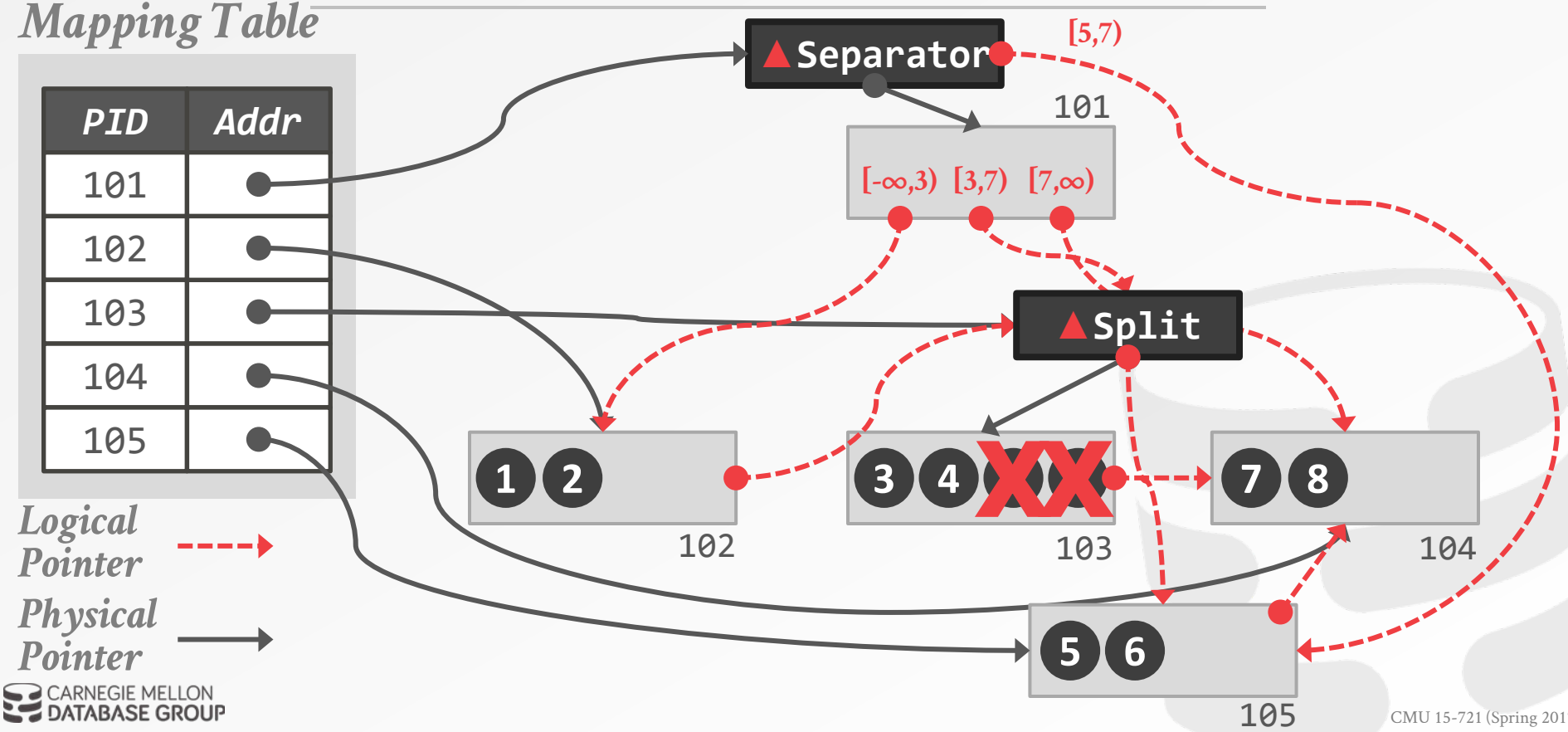
BW-TREE: STRUCTURE MODIFICATIONS



BW-TREE: STRUCTURE MODIFICATIONS

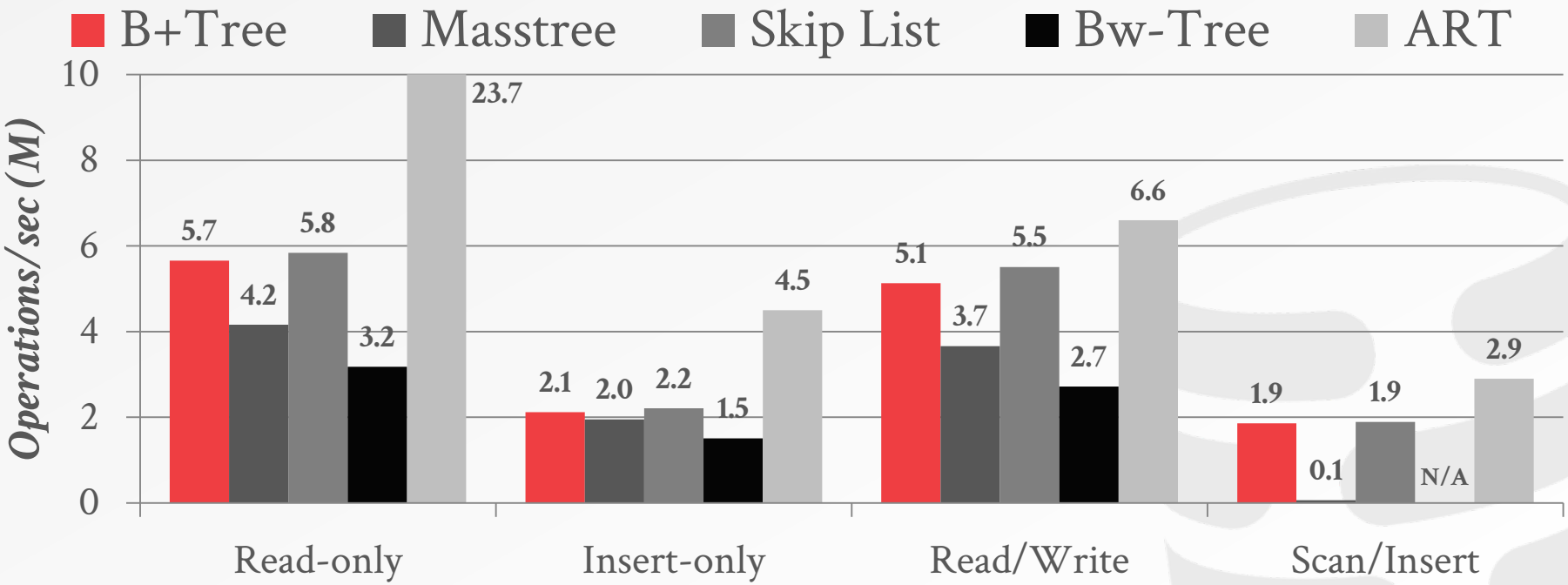


BW-TREE: STRUCTURE MODIFICATIONS



SINGLE-THREADED PERFORMANCE

Data Set: 30m Random 64-bit Integers



Source: Huanchen Zhang