

Transparent Dual Memory Compression Architecture

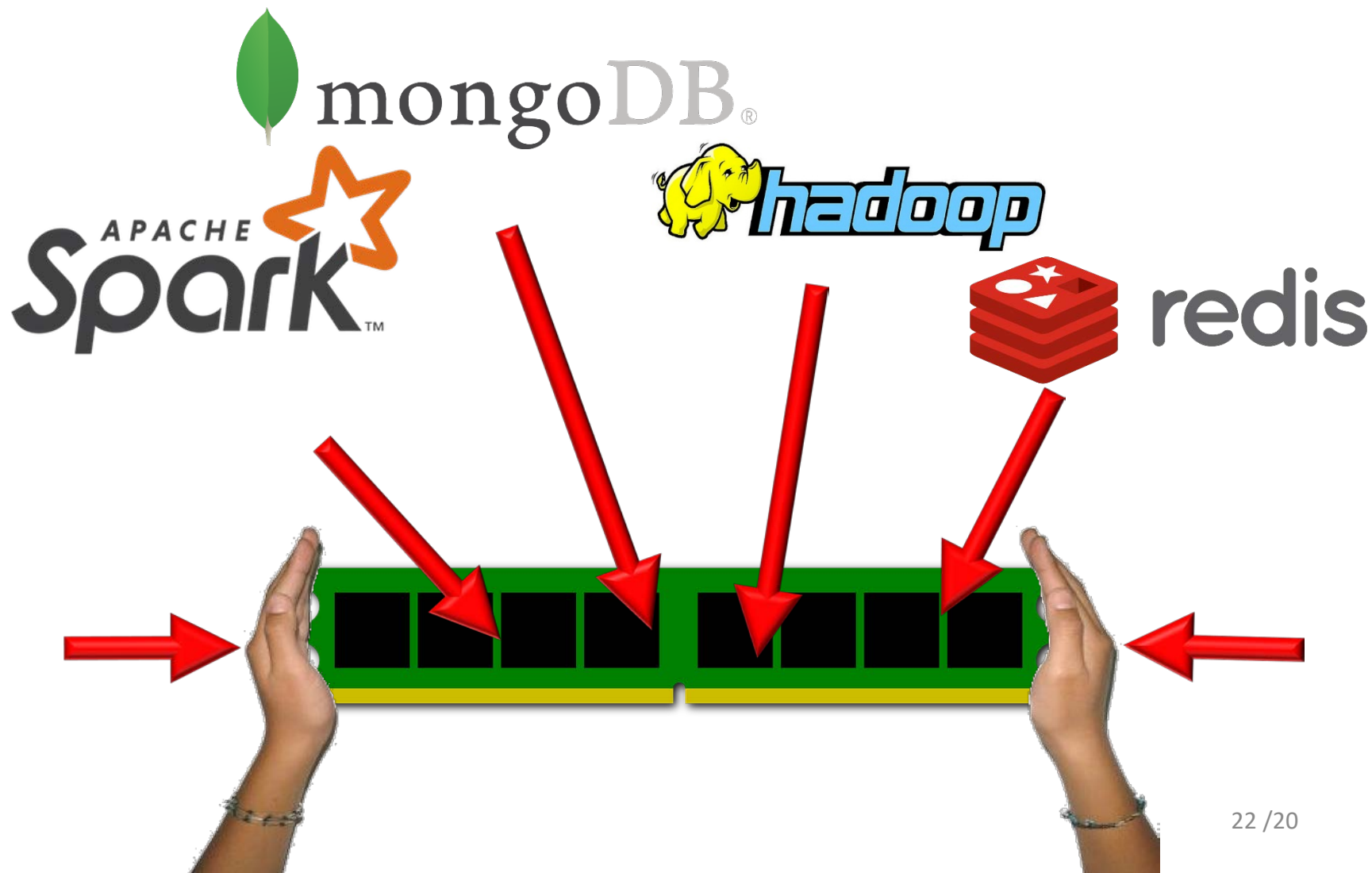
Seikwon Kim^{1,2}, Seonyoung Lee¹, Taehoon Kim¹, Jaehyuk Huh¹

KAIST¹

Samsung Electronics²

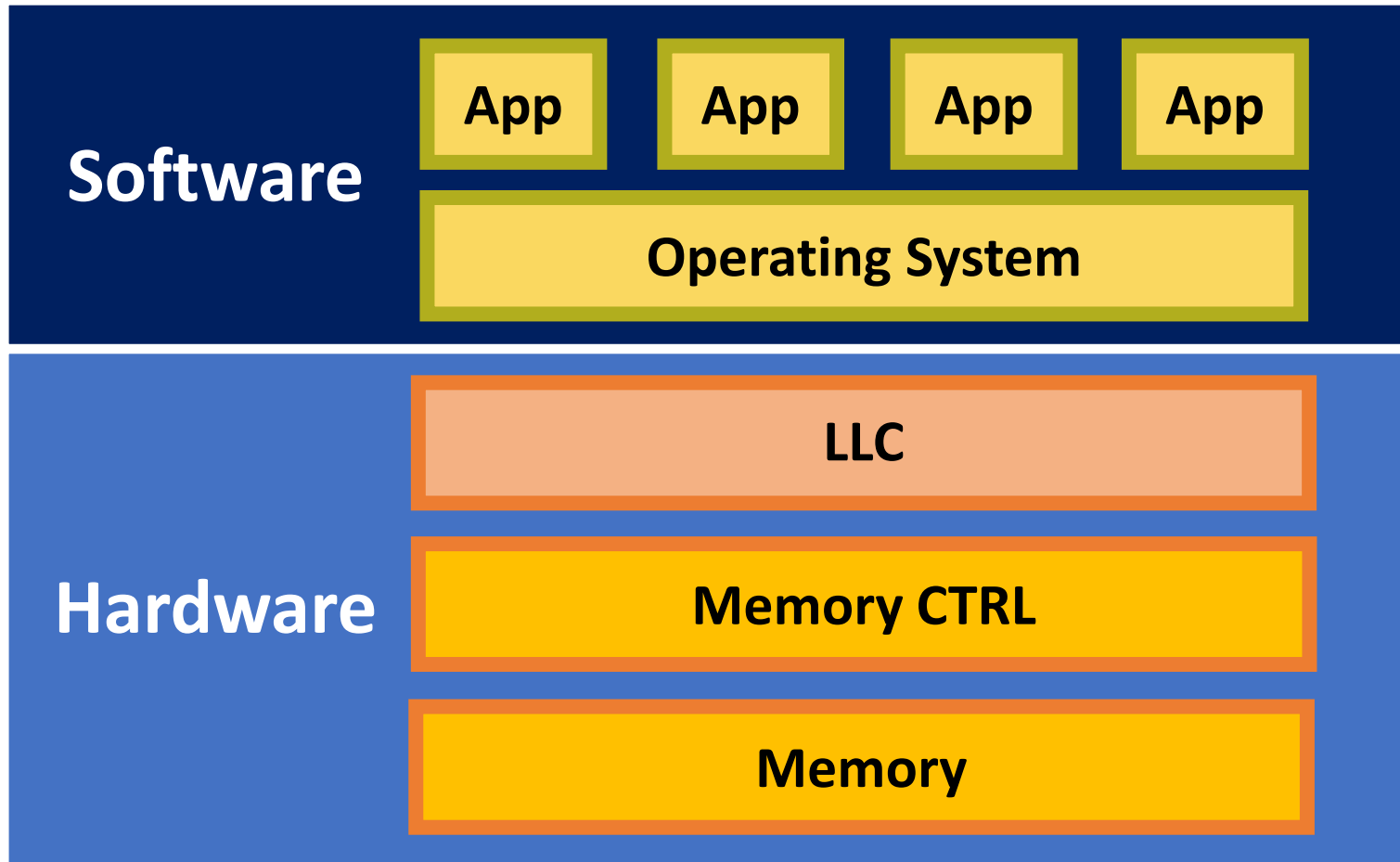
Why Compression?

- Workloads pressuring memory capacity



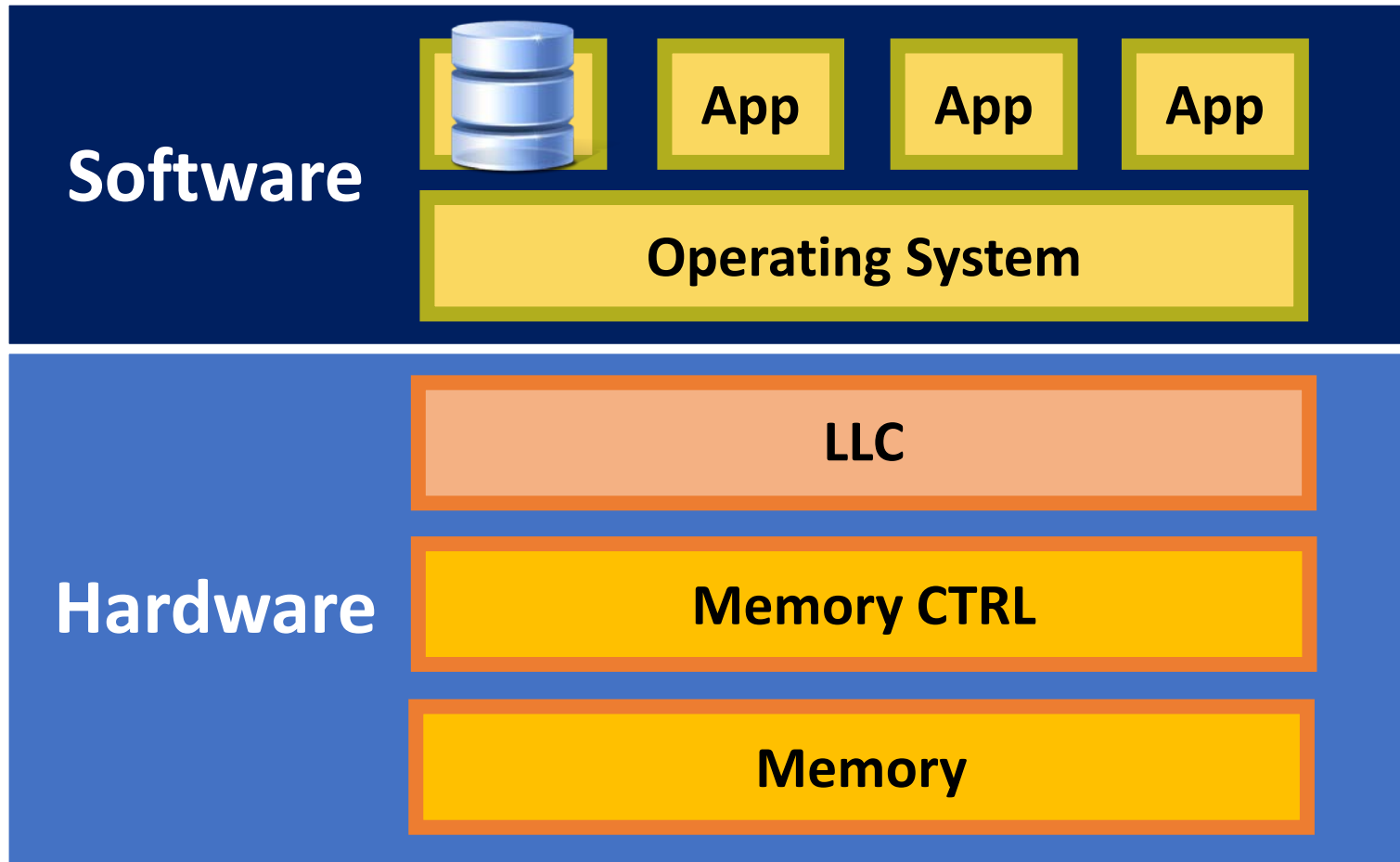
Where to Compress?

- Compression techniques to mitigate pressure



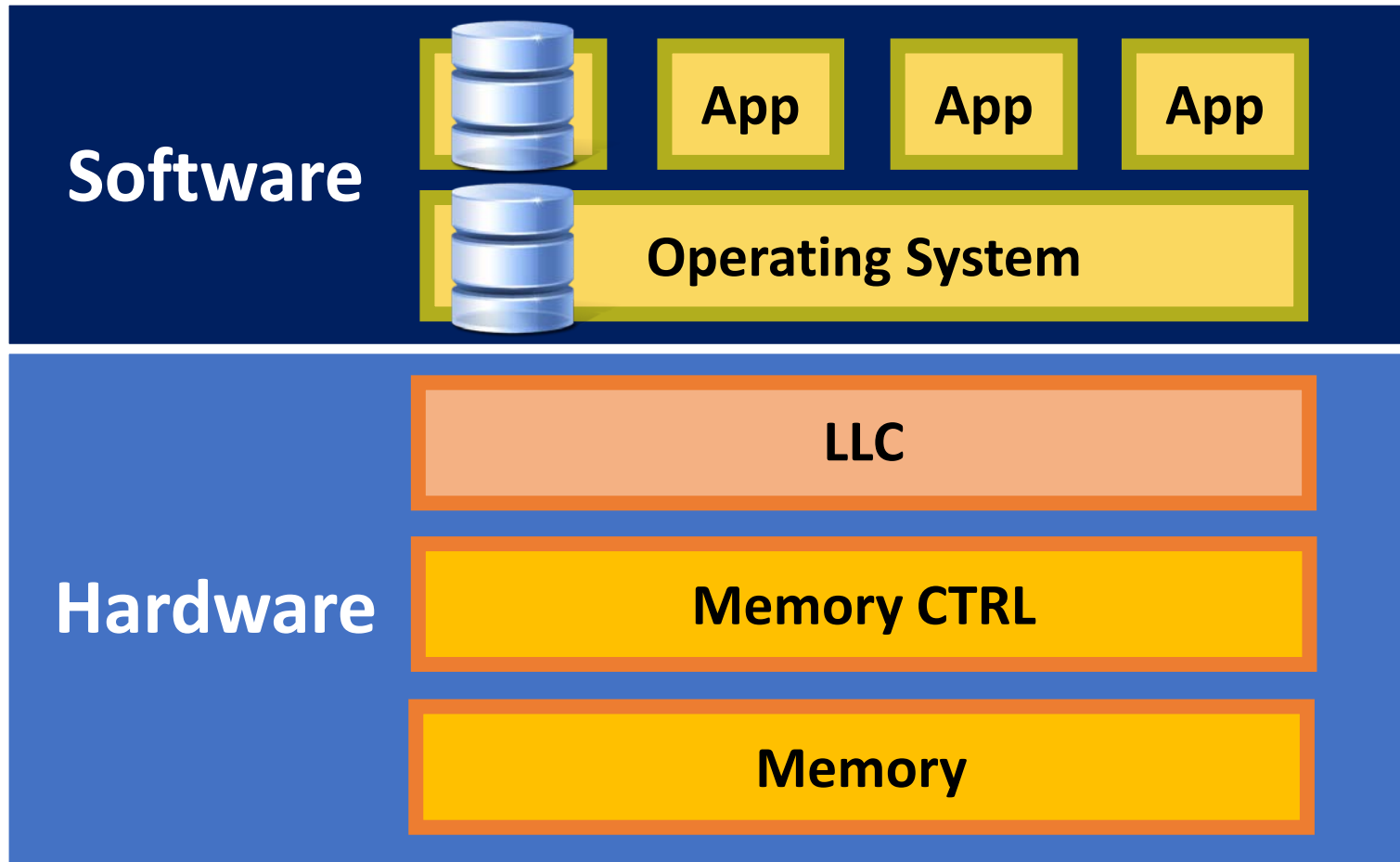
Where to Compress?

- Compression techniques to mitigate pressure



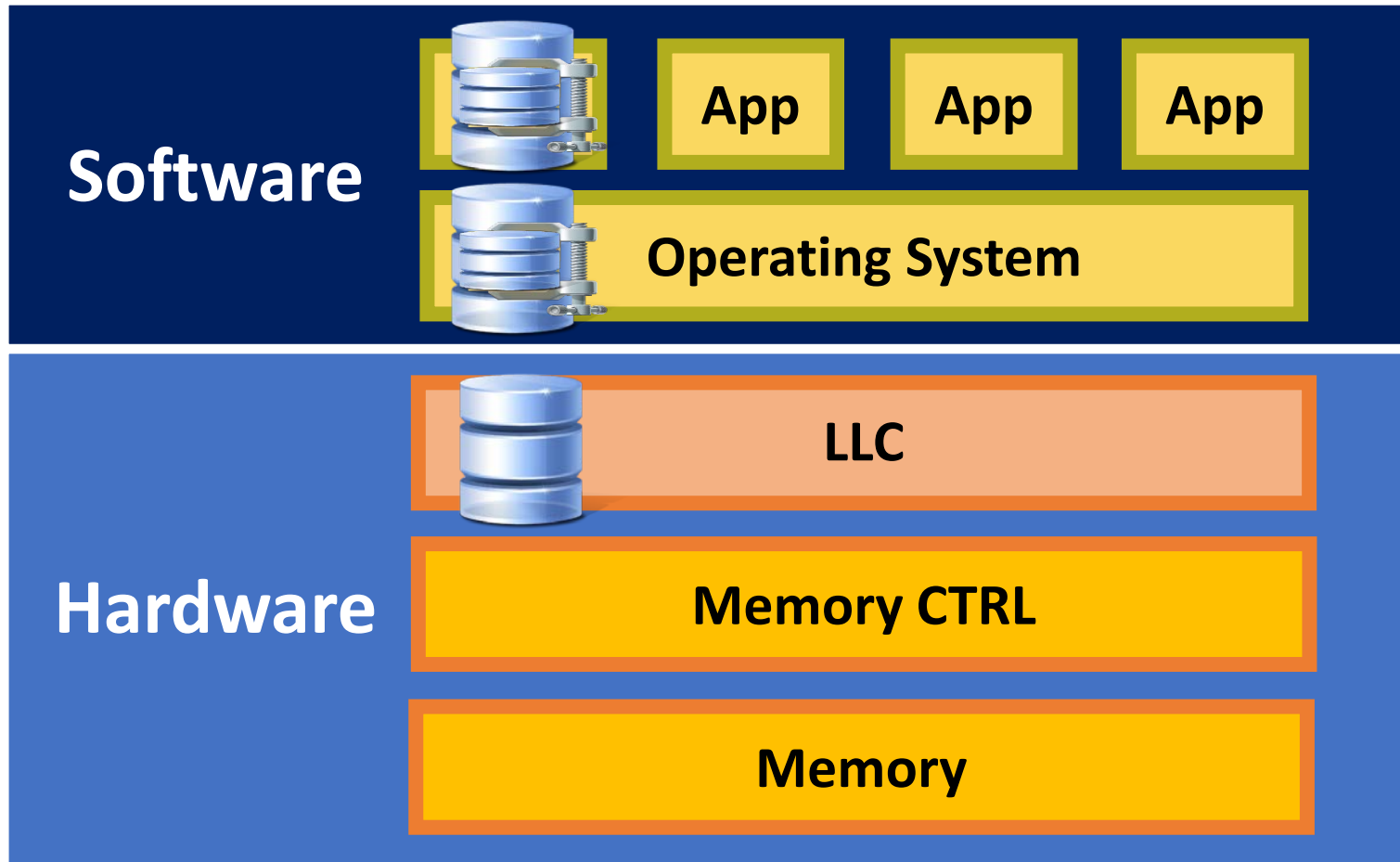
Where to Compress?

- Compression techniques to mitigate pressure



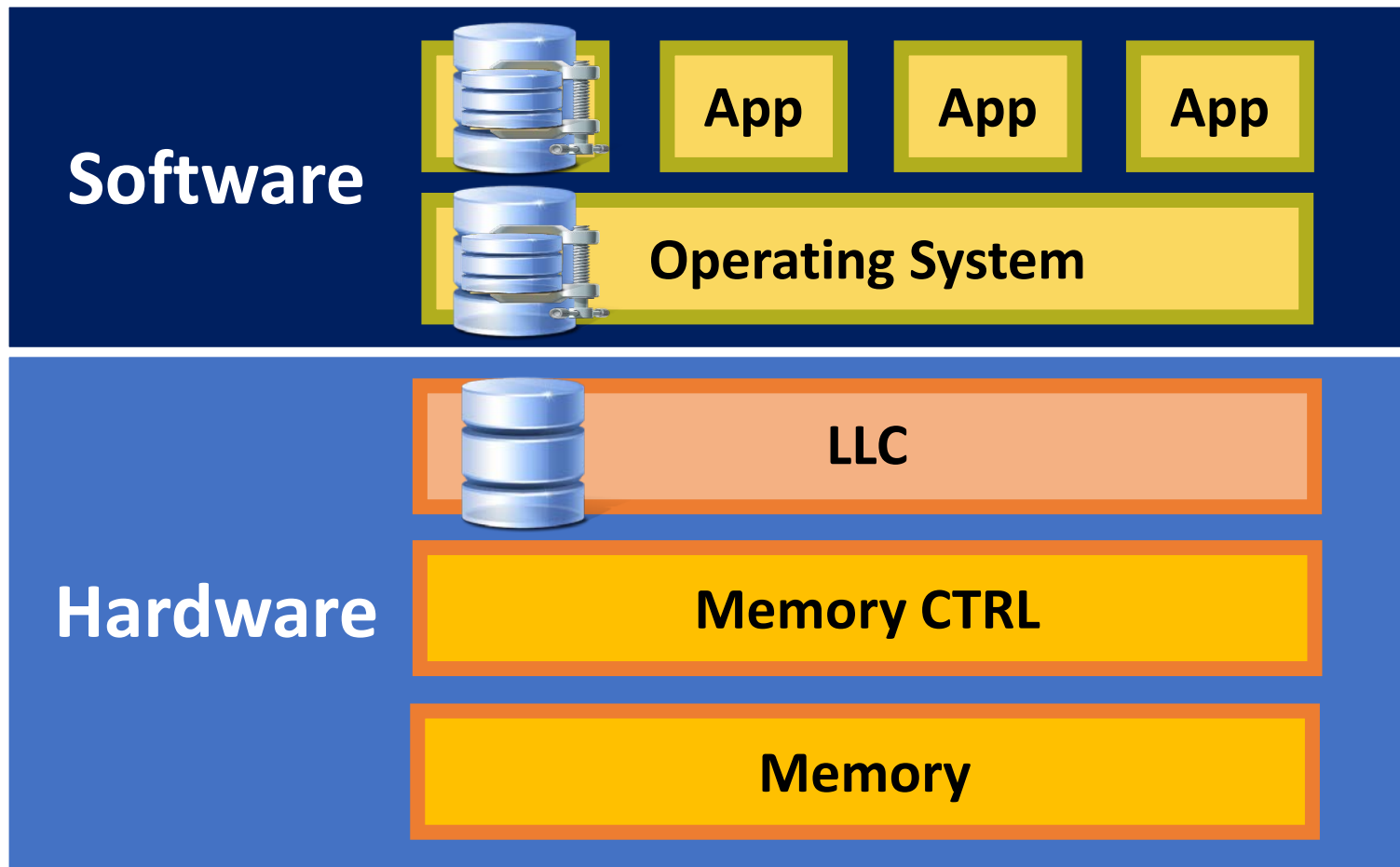
Where to Compress?

- Compression techniques to mitigate pressure



Where to Compress?

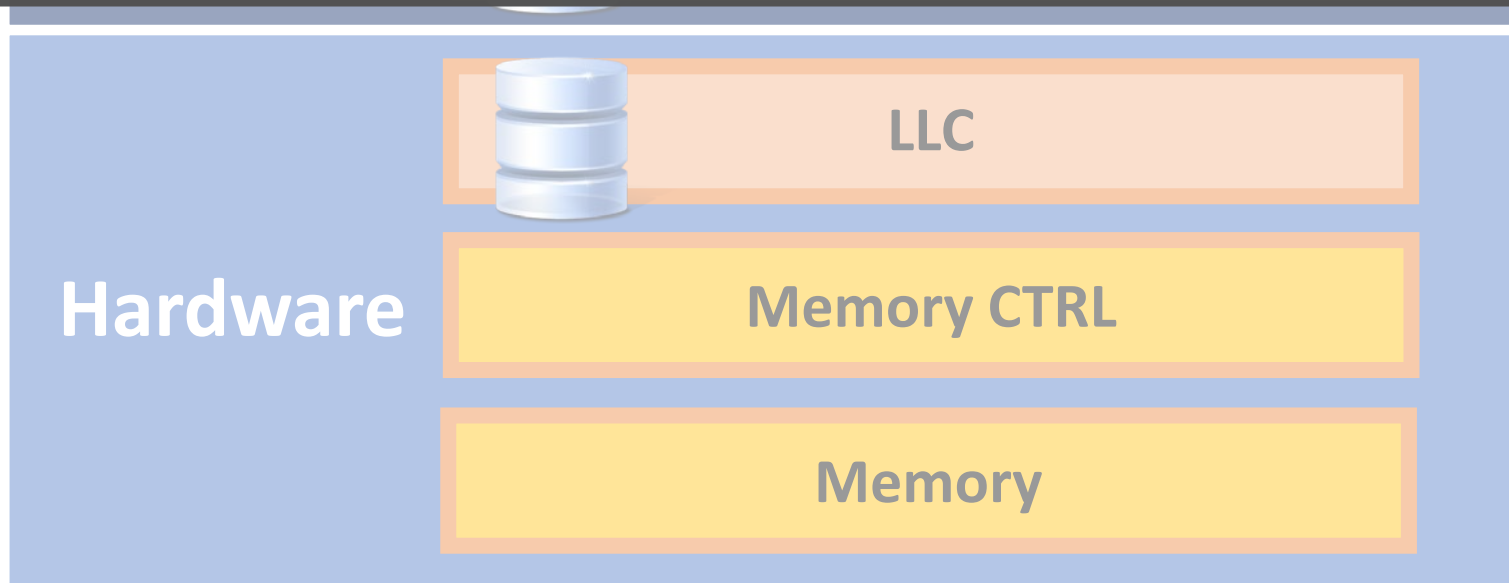
- Compression techniques to mitigate pressure



Where to Compress?

➤ Compression techniques to mitigate pressure

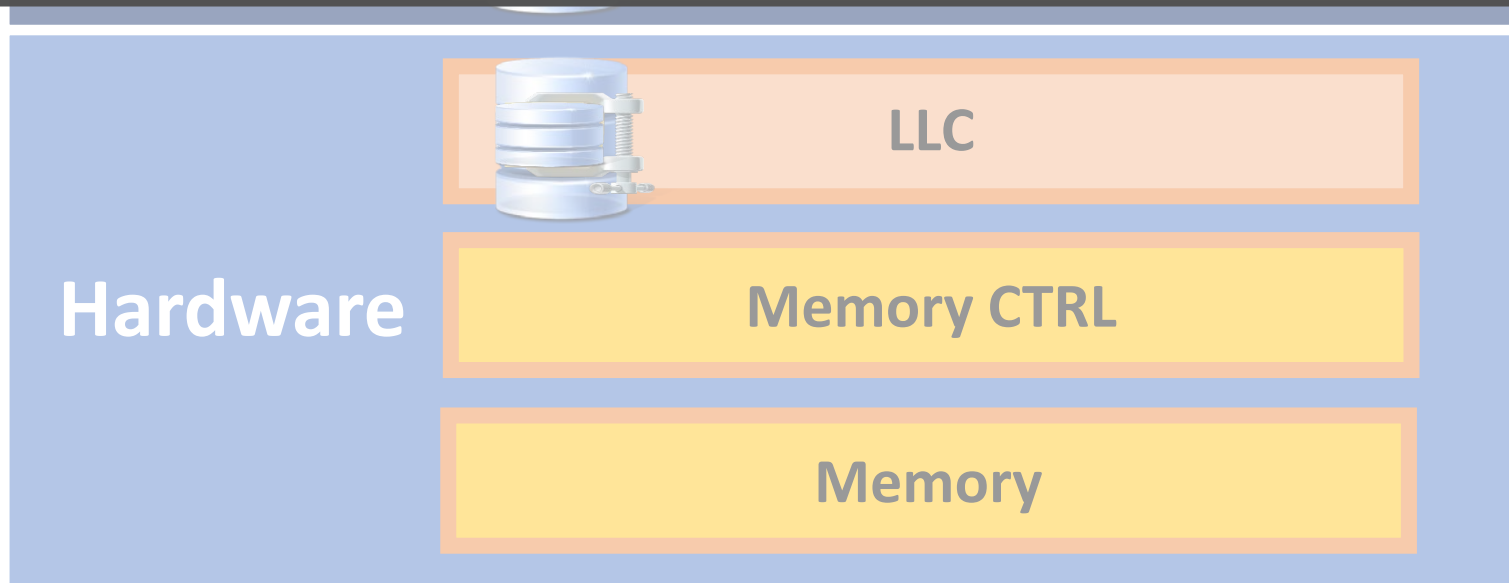
- 👎 Works only for limited workloads
- 👎 Incurs CPU processing overhead
- 👎 Hard to be fine-grained



Where to Compress?

➤ Compression techniques to mitigate pressure

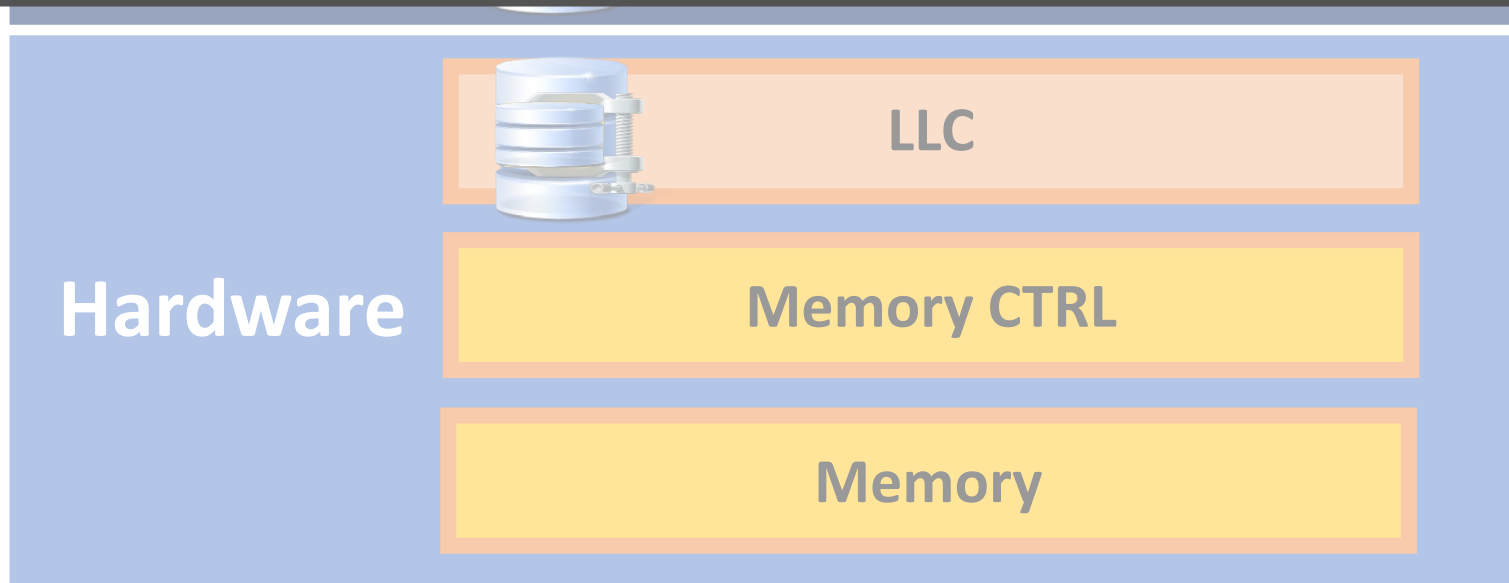
- 👎 Works only for limited workloads
- 👎 Incurs CPU processing overhead
- 👎 Hard to be fine-grained



Where to Compress?

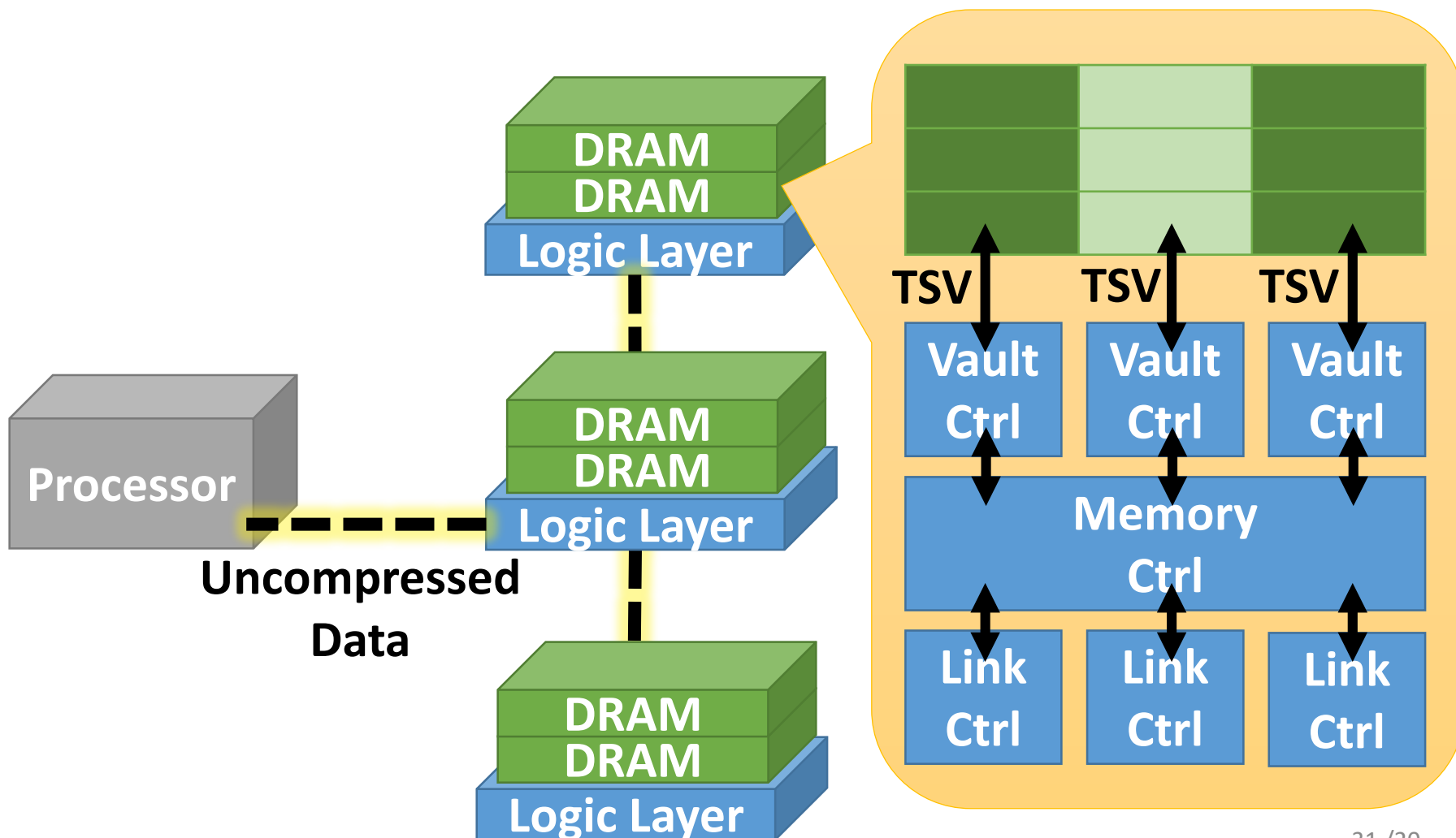
➤ Compression techniques to mitigate pressure

- 👎 Works only for limited workloads
- 👎 Incurs CPU processing overhead
- 👎 Hard to be fine-grained



Where to Compress: HMC

➤ Hybrid Memory Cube(HMC)

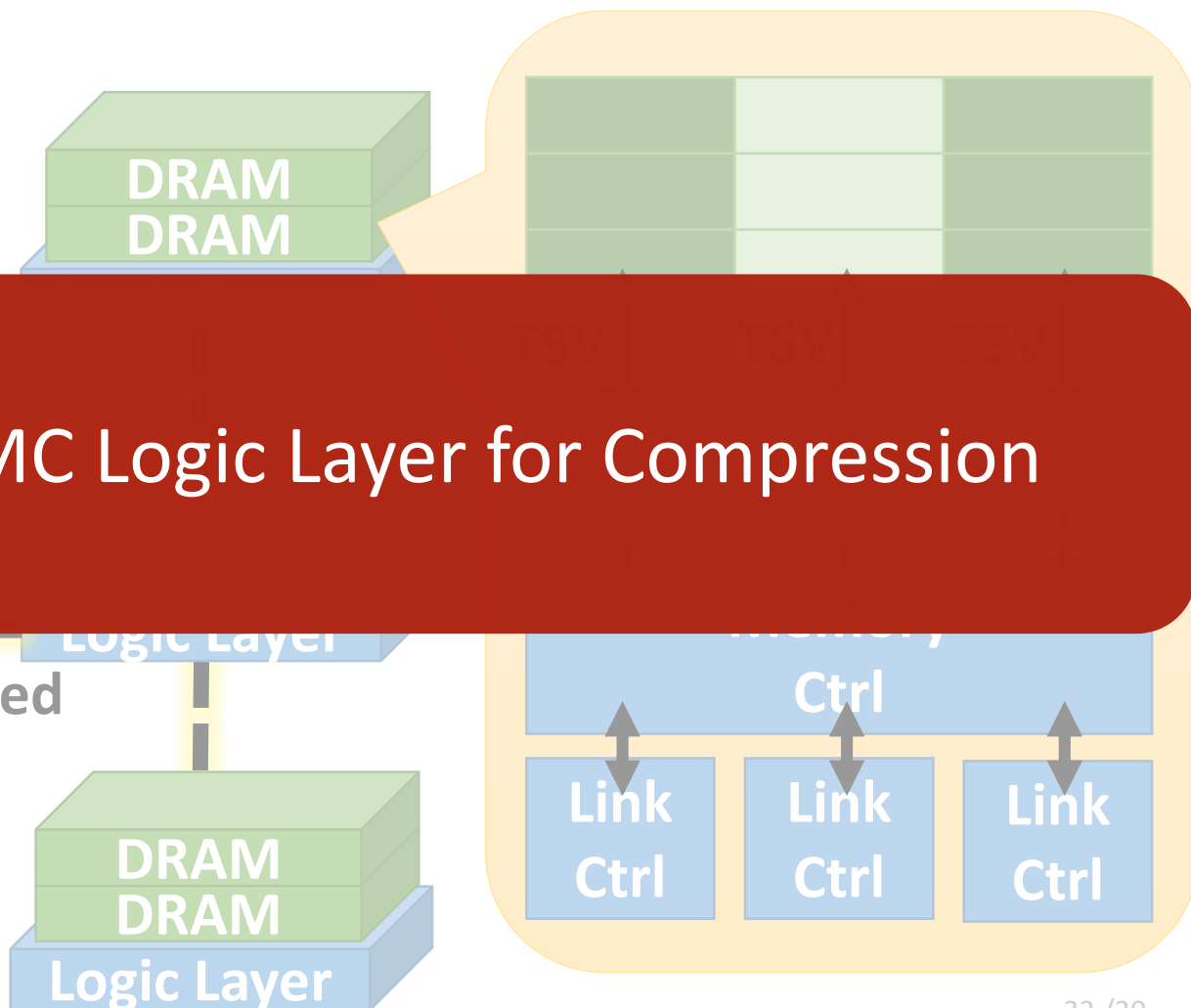


Where to Compress: HMC





➤ Hybrid Memory Cube(HMC)

Utilize HMC Logic Layer for Compression


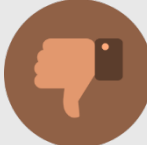


Uncompressed
Data



What Compression Technique?





	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑



What Compression Technique?


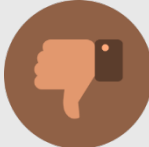


	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)

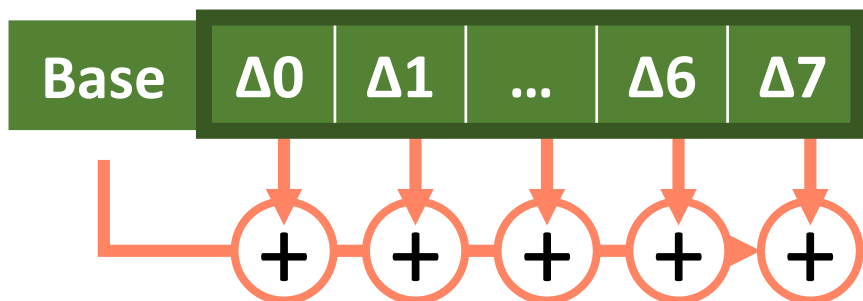


[1] Pekhimenko et al. PACT'12

What Compression Technique?


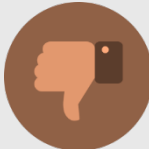


	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)

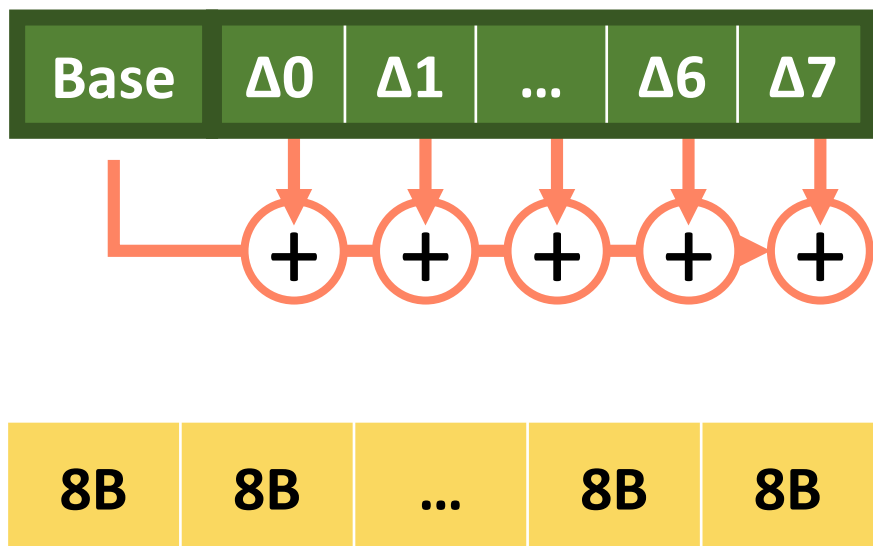


[1] Pekhimenko et al. PACT'12

What Compression Technique?


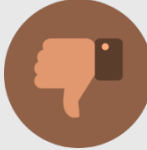


	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)

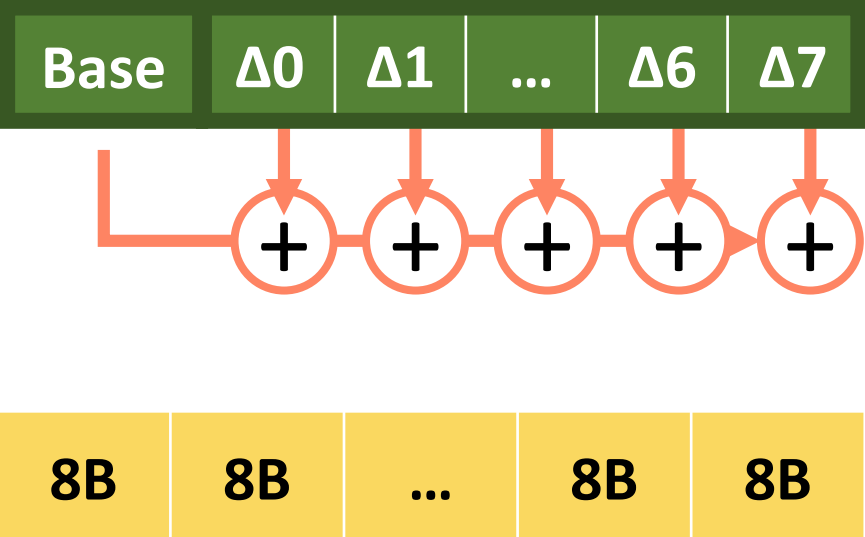


[1] Pekhimenko et al. PACT'12

What Compression Technique?


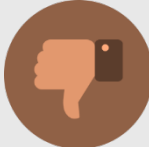


	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)

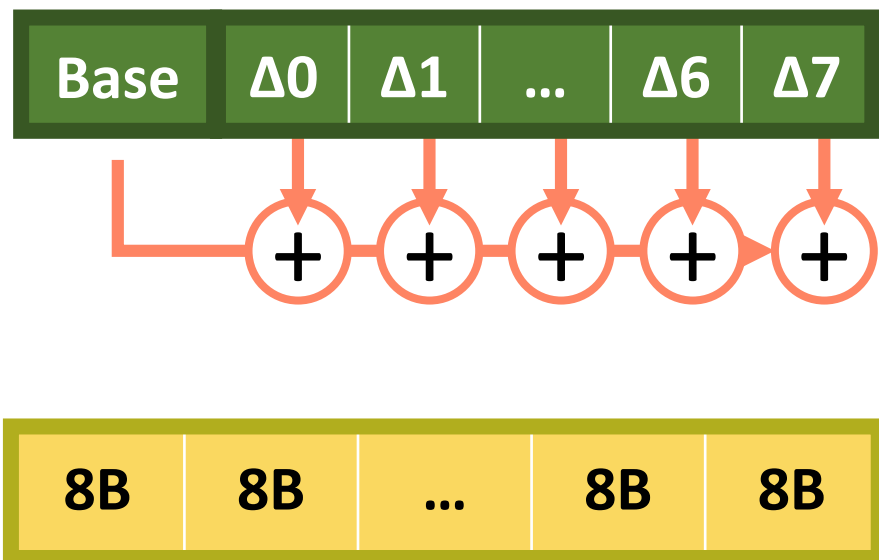


[1] Pekhimenko et al. PACT'12

What Compression Technique?


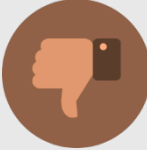


	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)

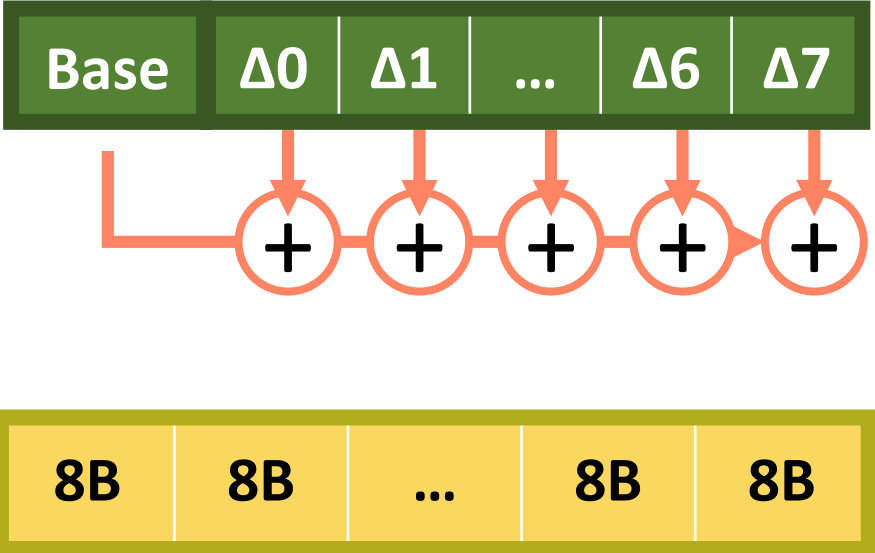


[1] Pekhimenko et al. PACT'12

What Compression Technique?


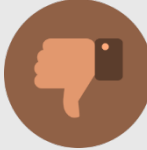


	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)

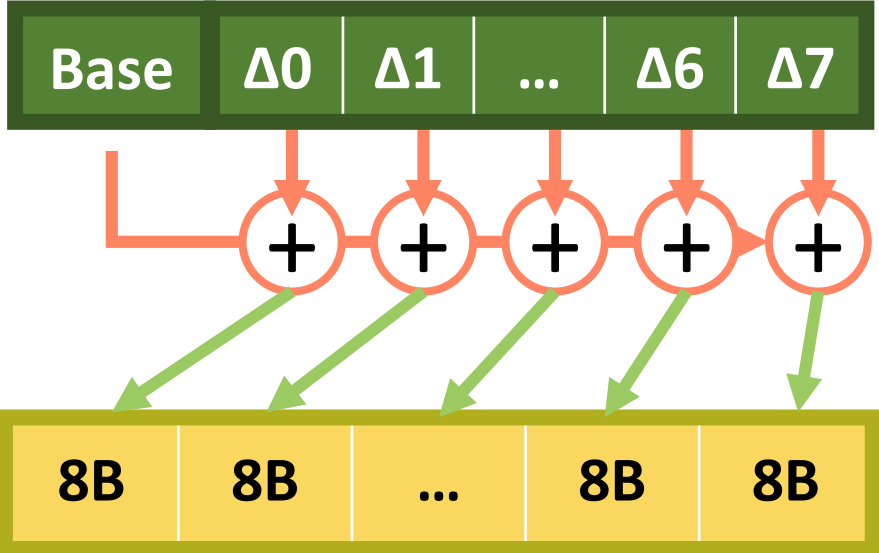


[1] Pekhimenko et al. PACT'12

What Compression Technique?





	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)

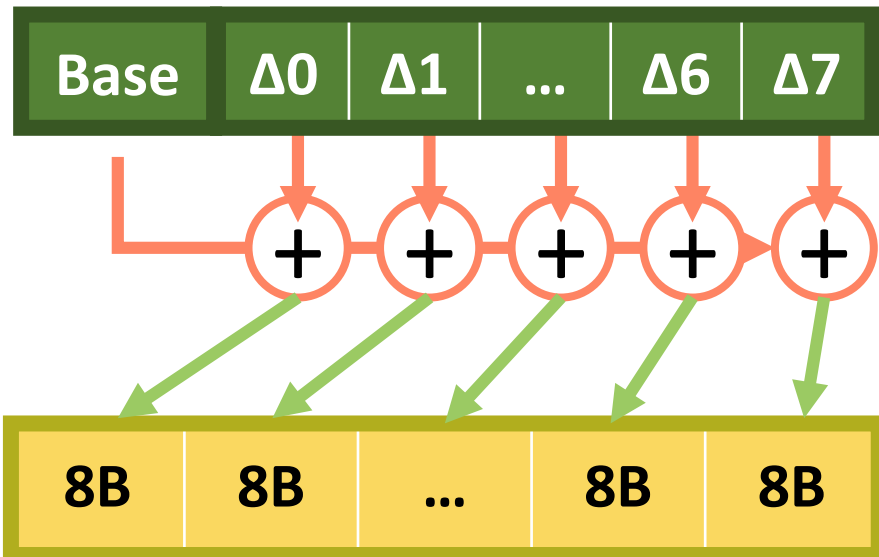


[1] Pekhimenko et al. PACT'12

What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


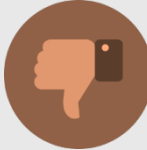




LZ77 (Capacity Optimized)

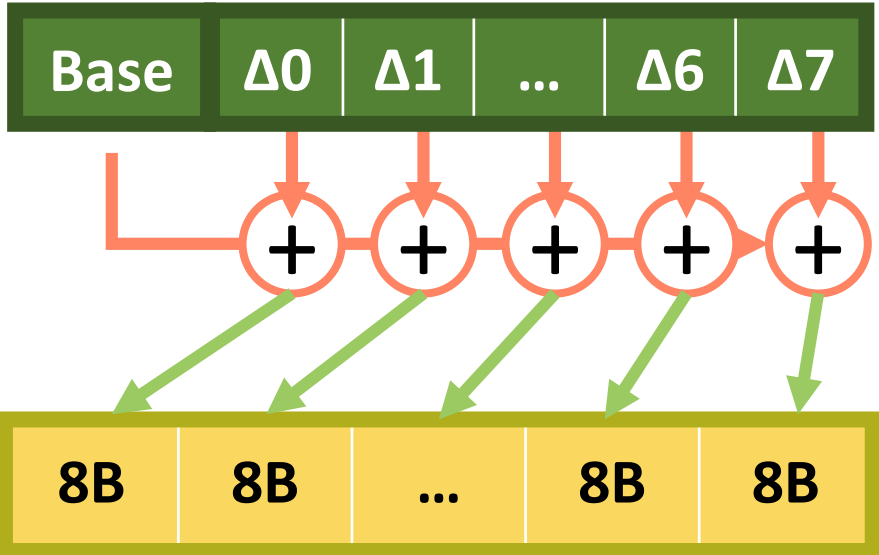


[1] Pekhimenko et al. PACT'12

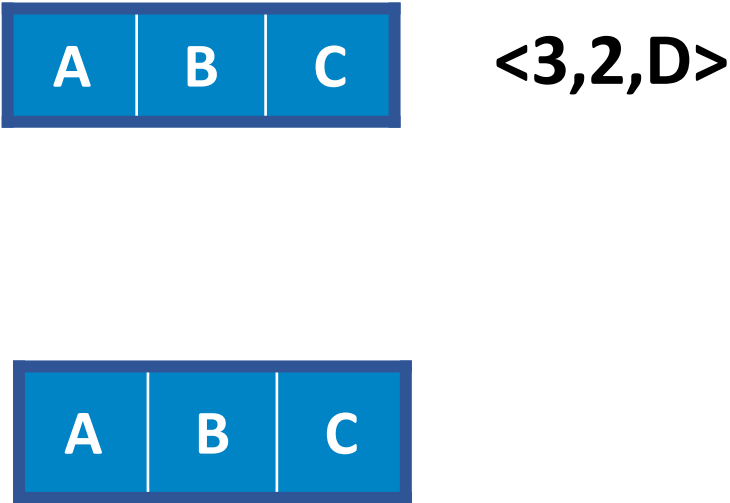
What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


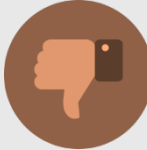




LZ77 (Capacity Optimized)

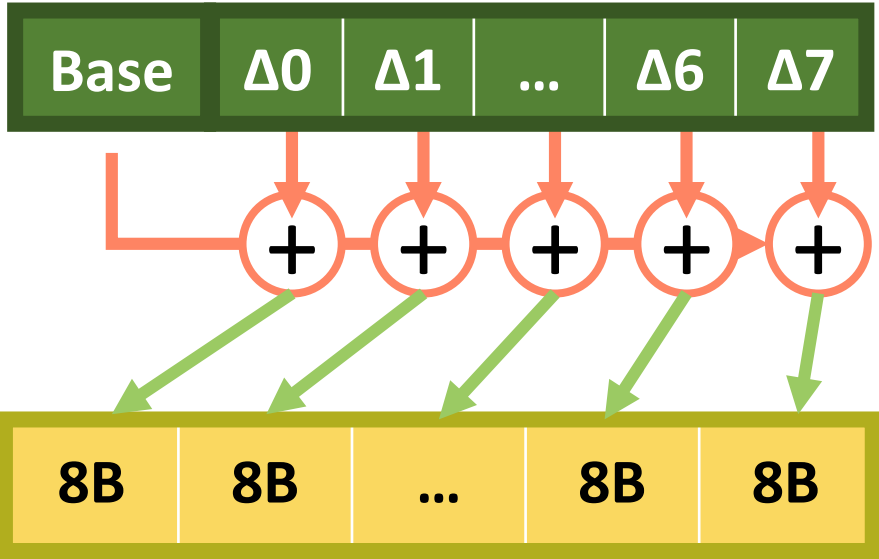


[1] Pekhimenko et al. PACT'12

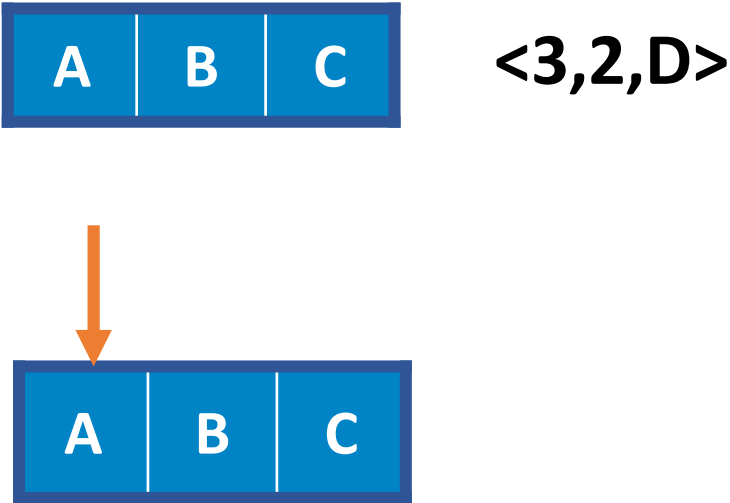
What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


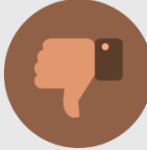




LZ77 (Capacity Optimized)

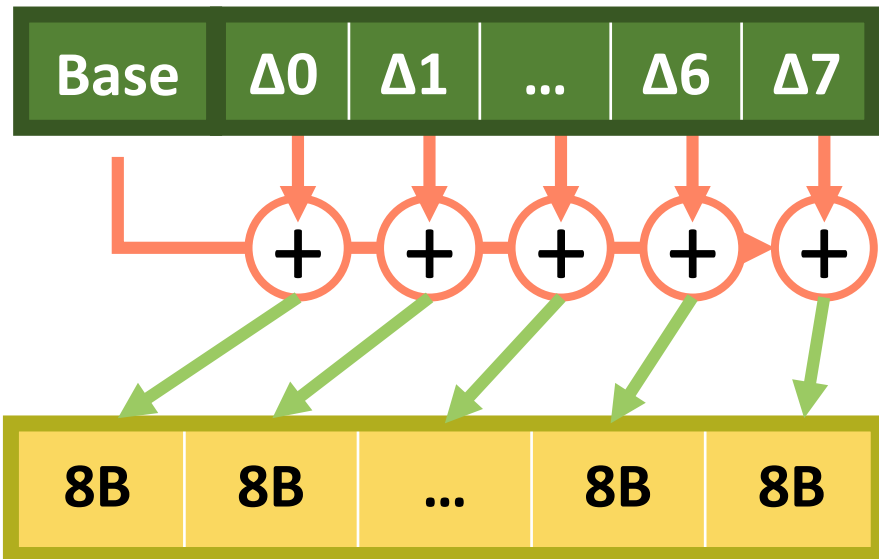


[1] Pekhimenko et al. PACT'12

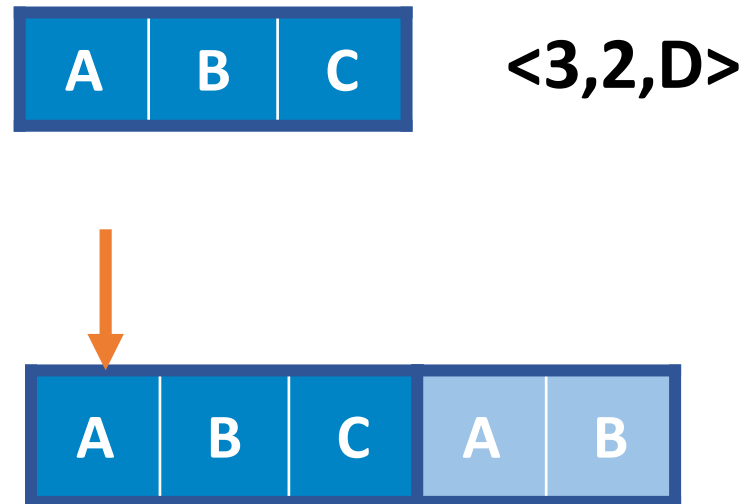
What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


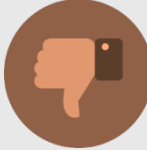




LZ77 (Capacity Optimized)

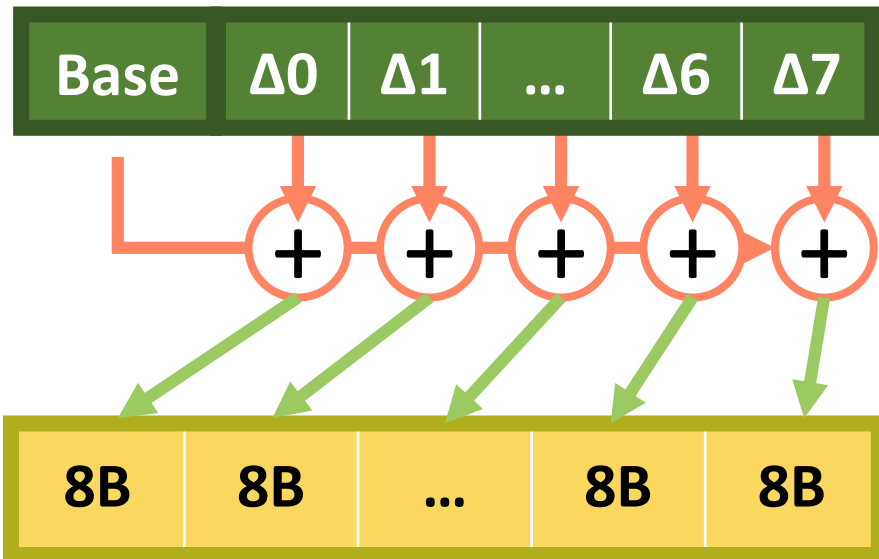


[1] Pekhimenko et al. PACT'12

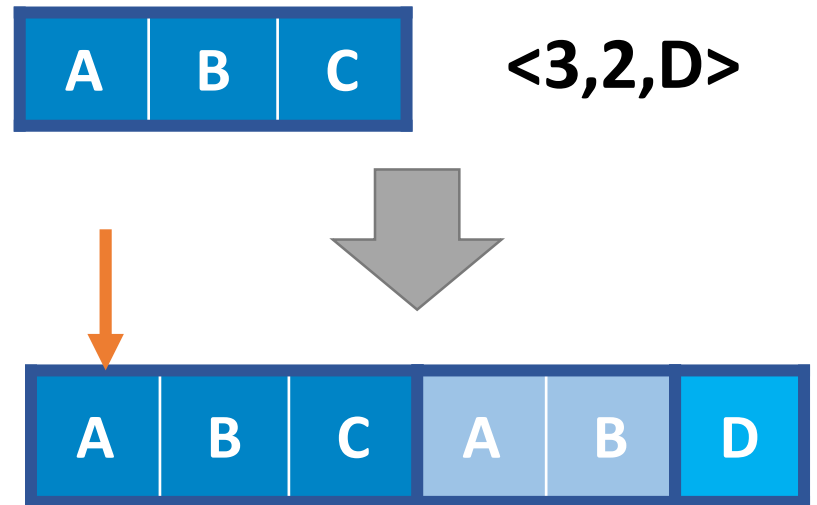
What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


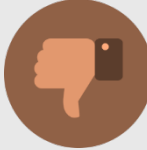




LZ77 (Capacity Optimized)

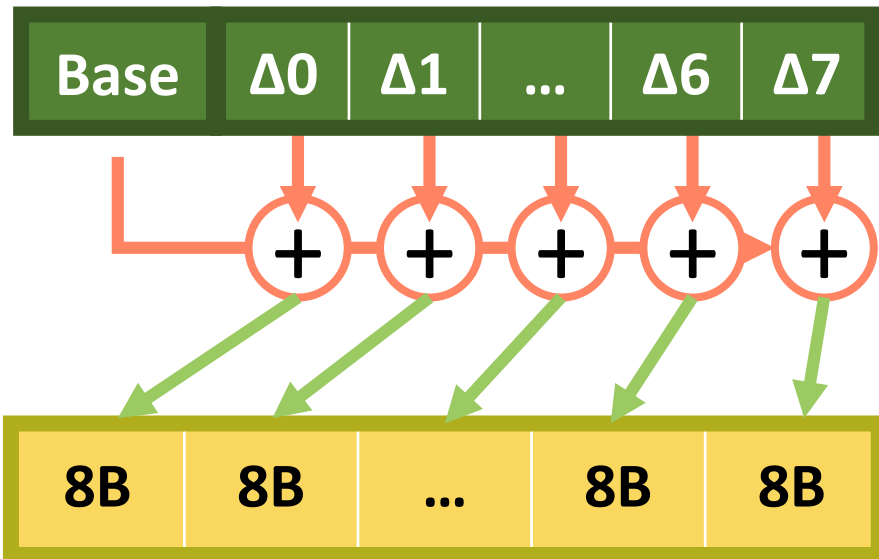


[1] Pekhimenko et al. PACT'12

What Compression Technique?

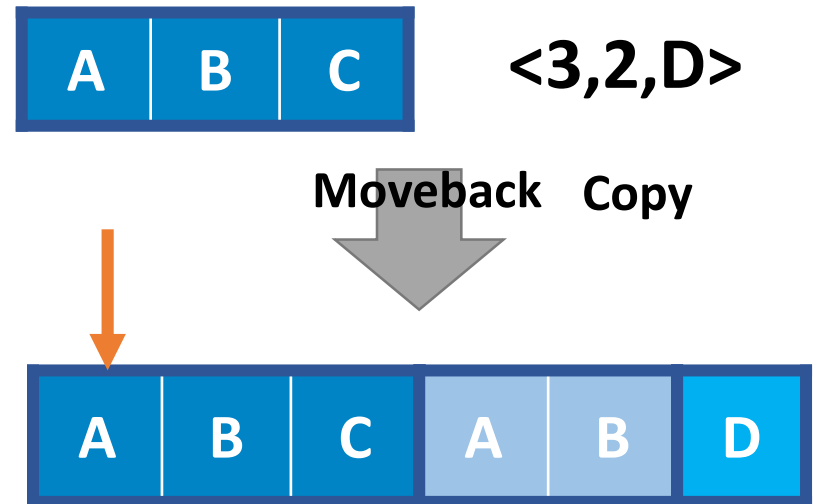
	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


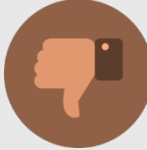




[1] Pekhimenko et al. PACT'12

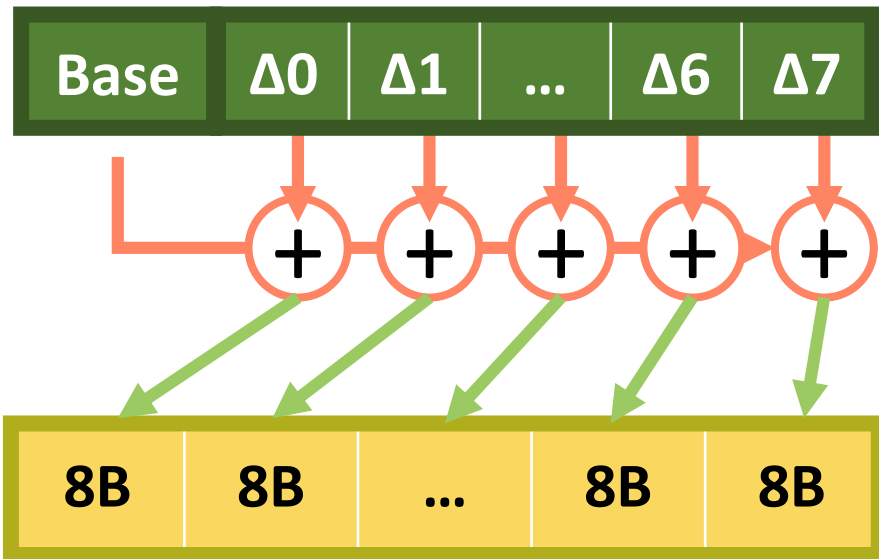
LZ77 (Capacity Optimized)



What Compression Technique?

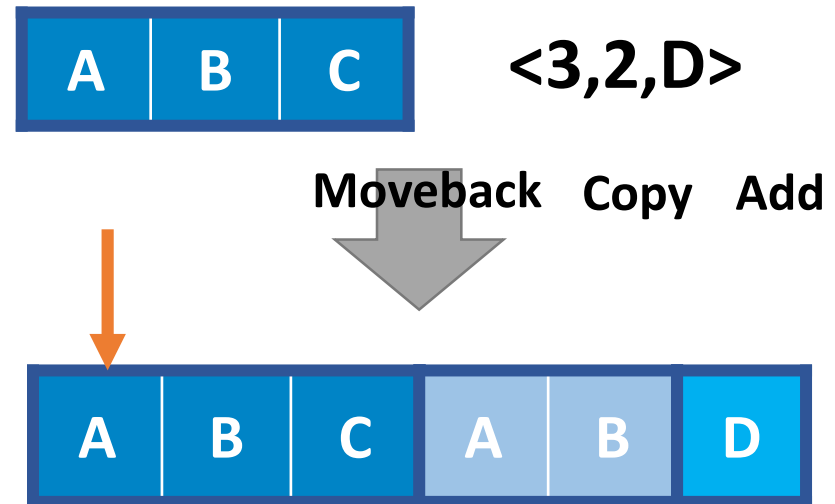
	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


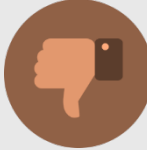




[1] Pekhimenko et al. PACT'12

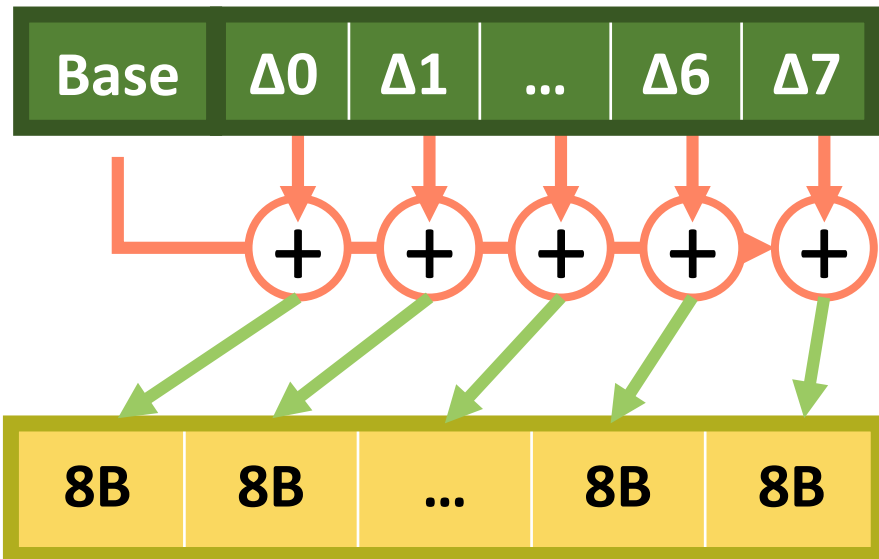
LZ77 (Capacity Optimized)



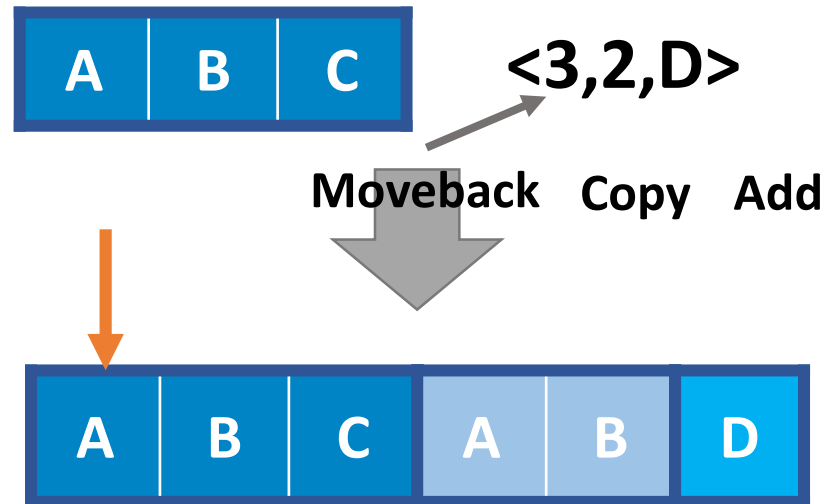
What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


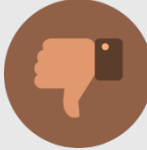




LZ77 (Capacity Optimized)

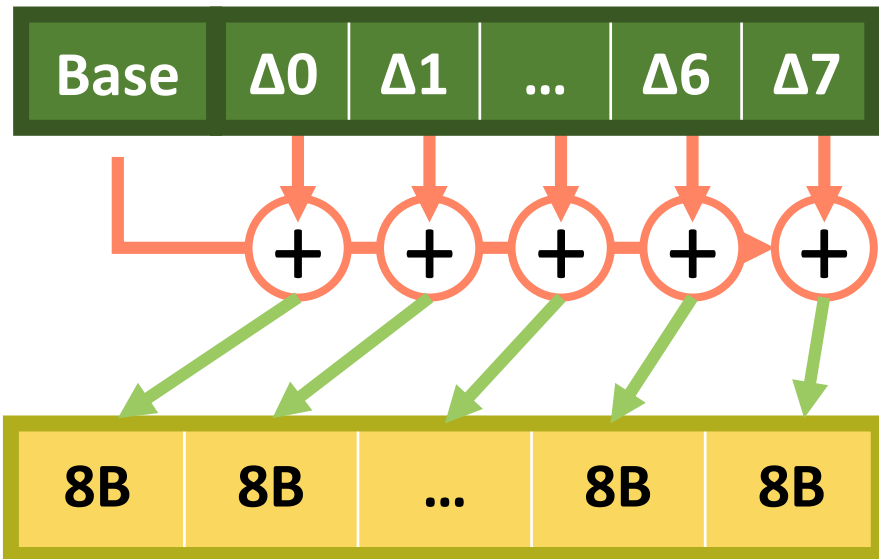


[1] Pekhimenko et al. PACT'12

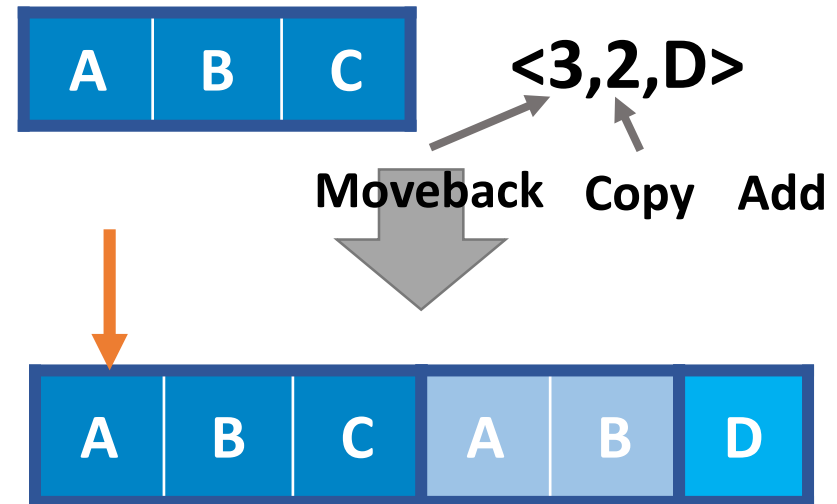
What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


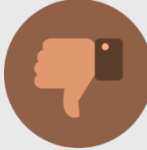




LZ77 (Capacity Optimized)

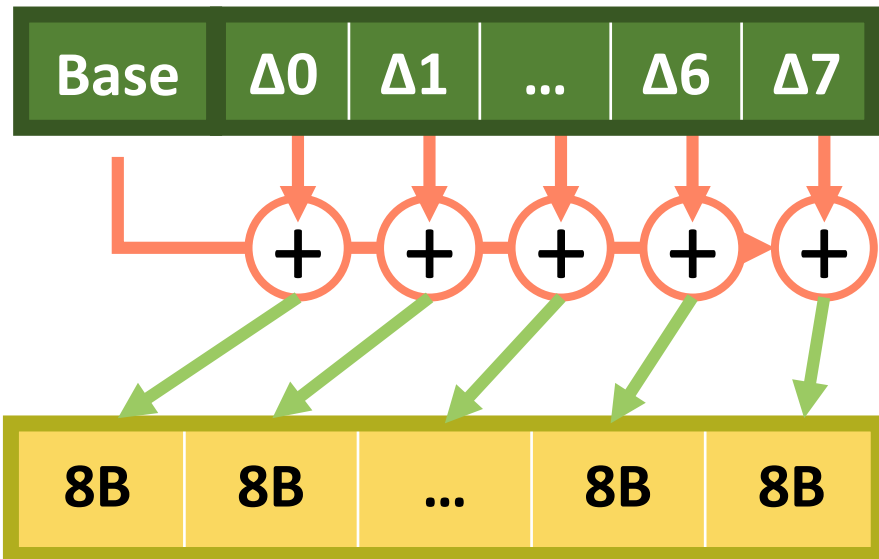


[1] Pekhimenko et al. PACT'12

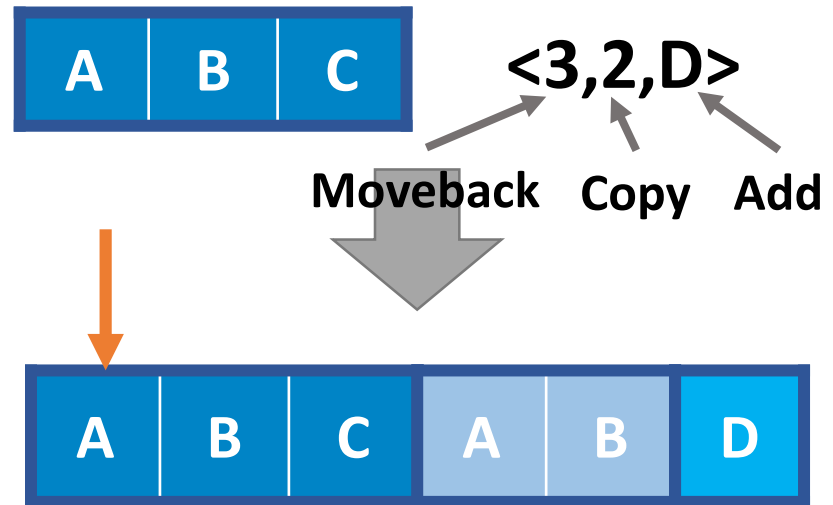
What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


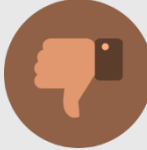




LZ77 (Capacity Optimized)

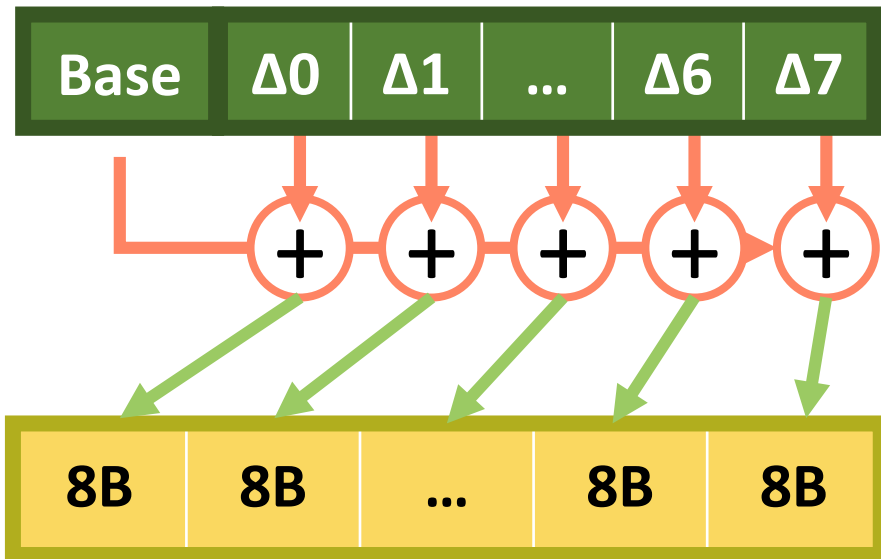


[1] Pekhimenko et al. PACT'12

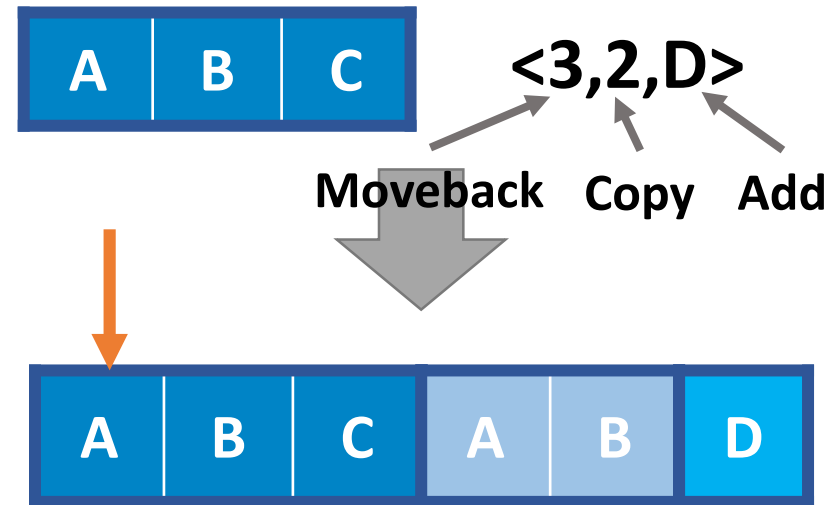
What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


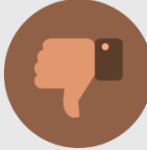




LZ77 (Capacity Optimized)

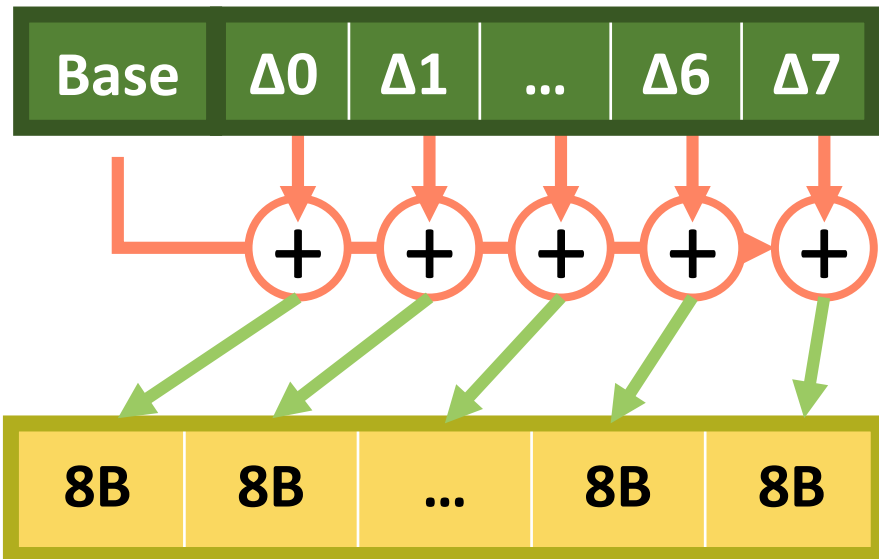


[1] Pekhimenko et al. PACT'12

What Compression Technique?

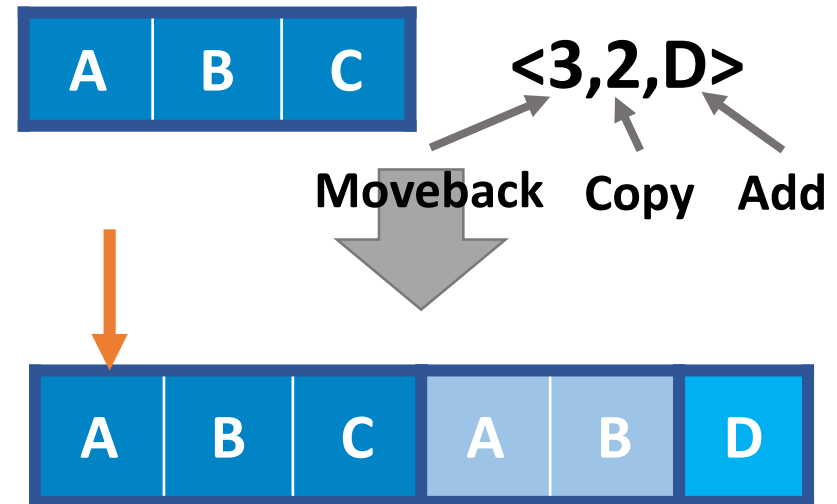
	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


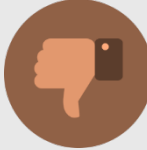




[1] Pekhimenko et al. PACT'12

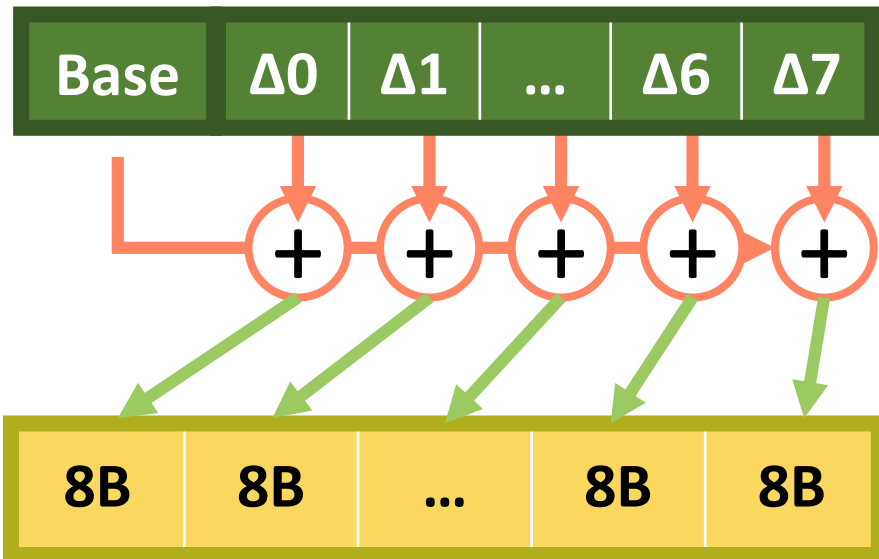
LZ77 (Capacity Optimized)



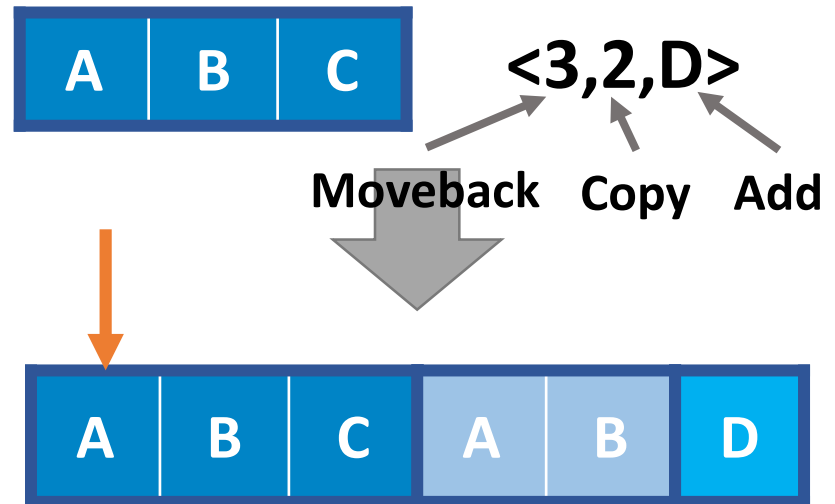
What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)







LZ77 (Capacity Optimized)

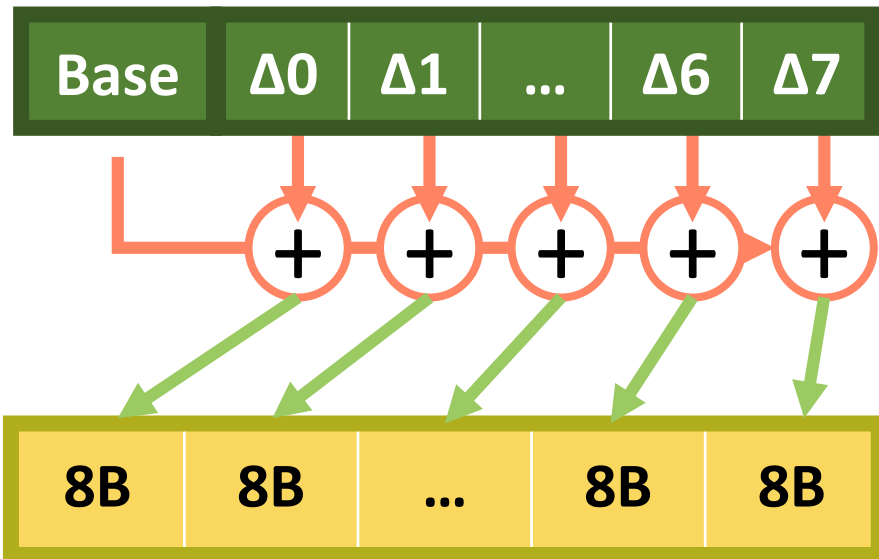


[1] Pekhimenko et al. PACT'12

What Compression Technique?

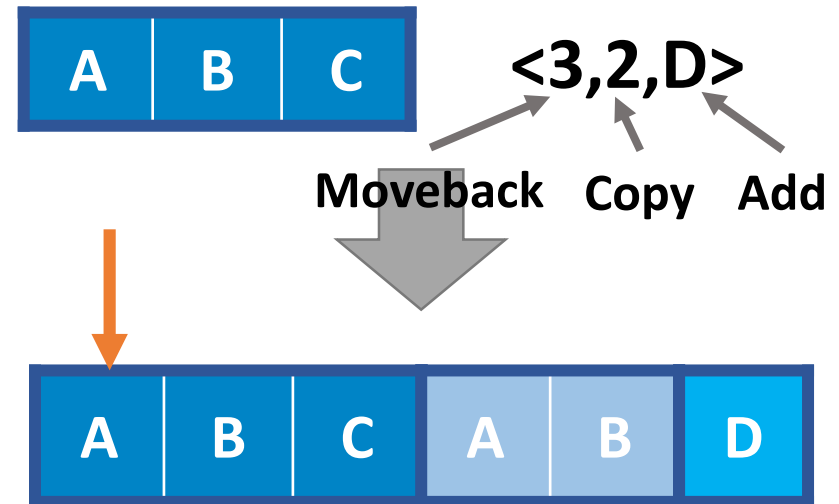
	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


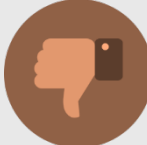




[1] Pekhimenko et al. PACT'12

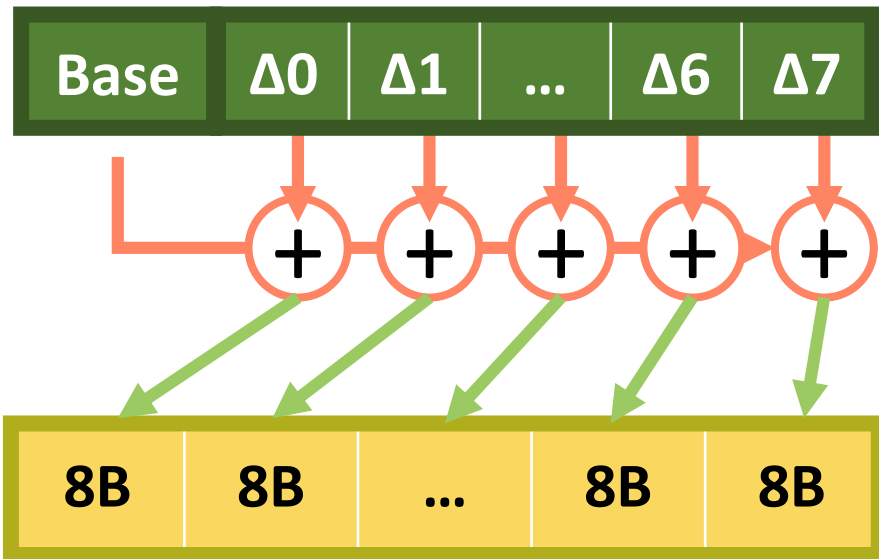
LZ77 (Capacity Optimized)



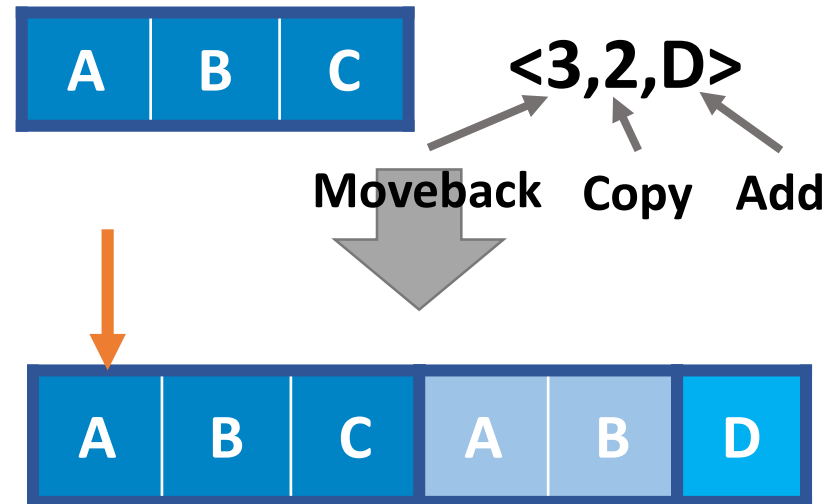
What Compression Technique?

	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)


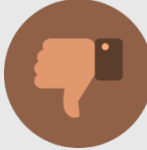




LZ77 (Capacity Optimized)

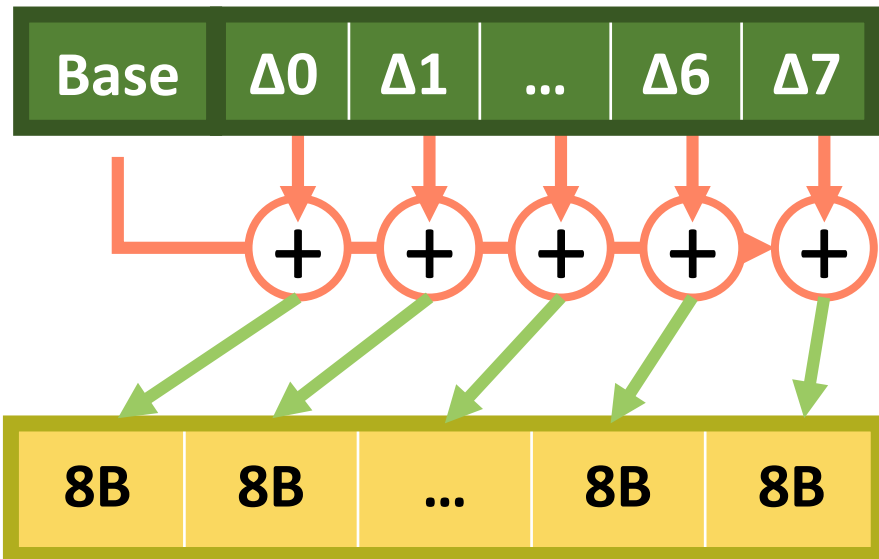


[1] Pekhimenko et al. PACT'12

What Compression Technique?

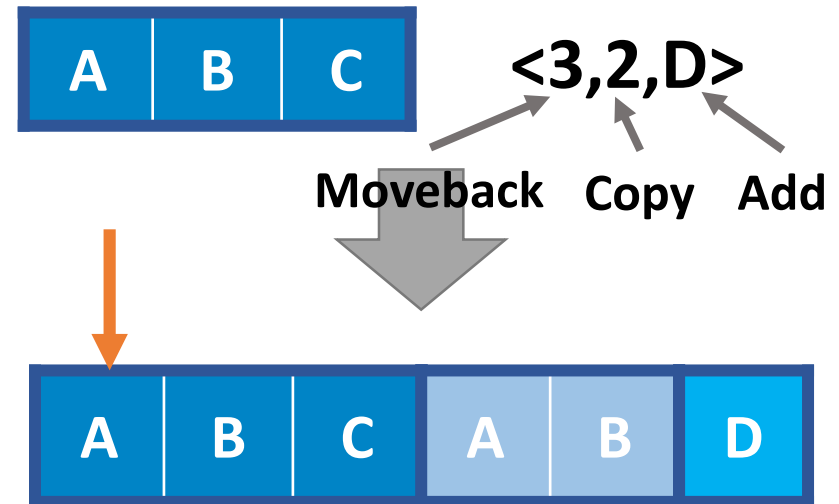
	Latency Opt (BDI ^[1])	Capacity Opt (LZ)
Decompression Latency	 1 cycle	 64 cycle
Compression Ratio	 --	 56% ↑

BDI (Latency Optimized)

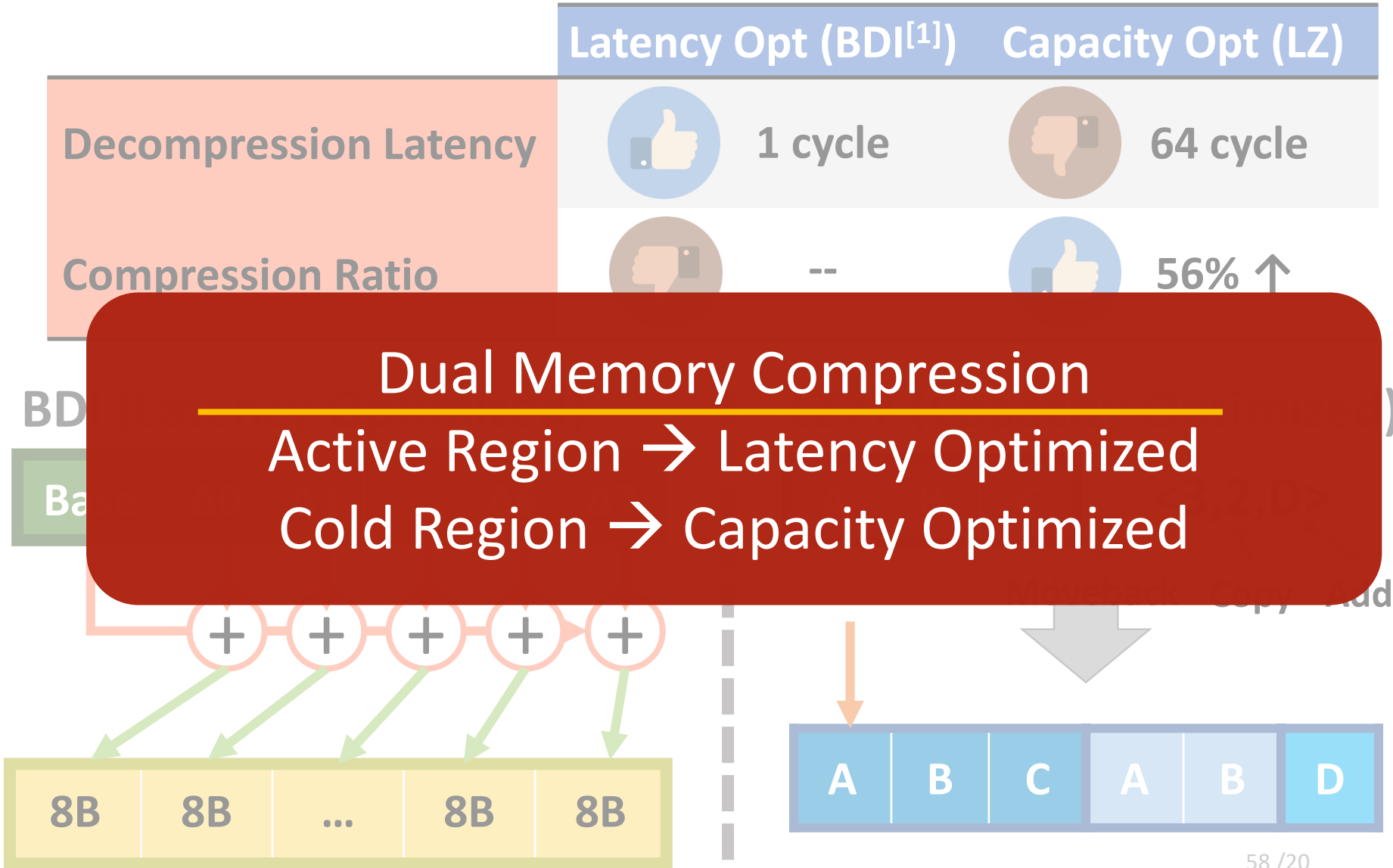


[1] Pekhimenko et al. PACT'12

LZ77 (Capacity Optimized)



What Compression Technique?



[1] Pekhimenko et al. PACT'12

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data

- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



Compressed Page

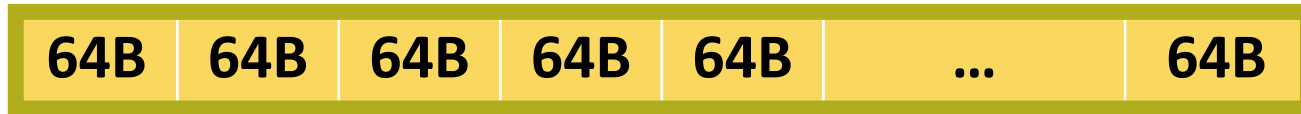
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



Compressed Page

- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



Compressed Page

- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



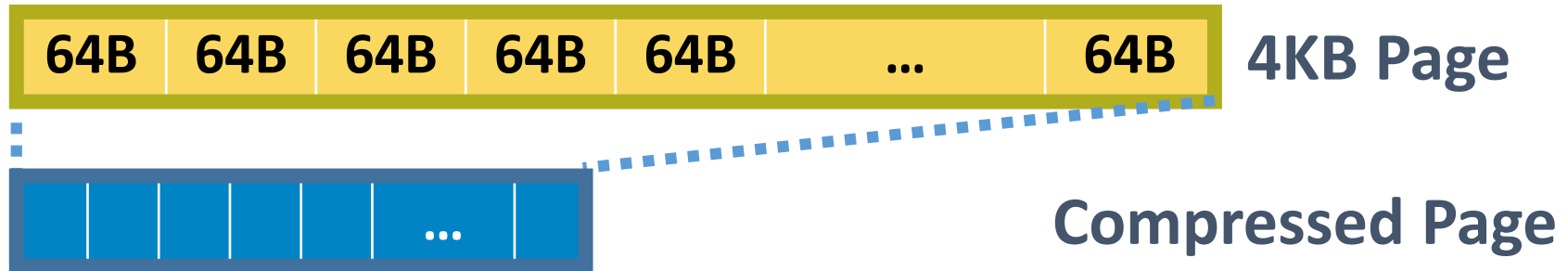
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



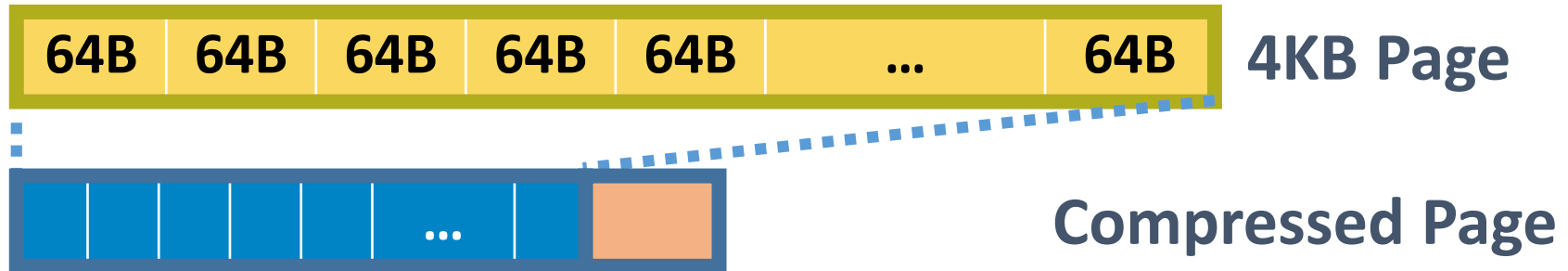
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



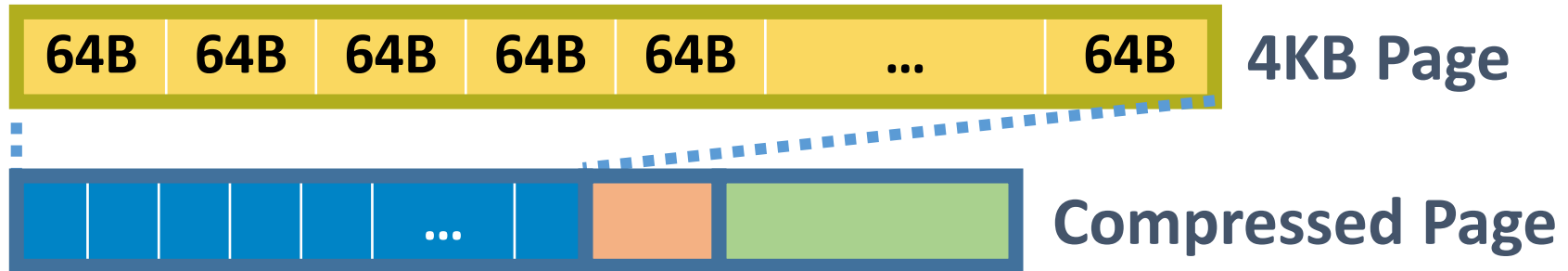
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



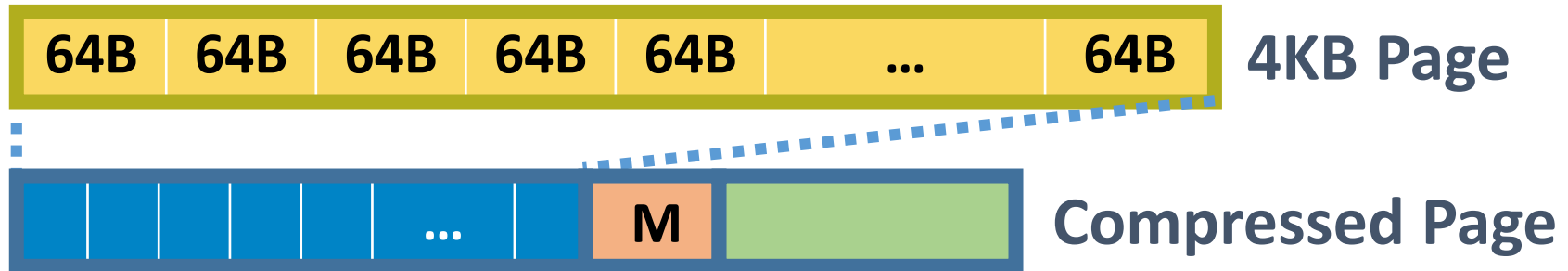
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



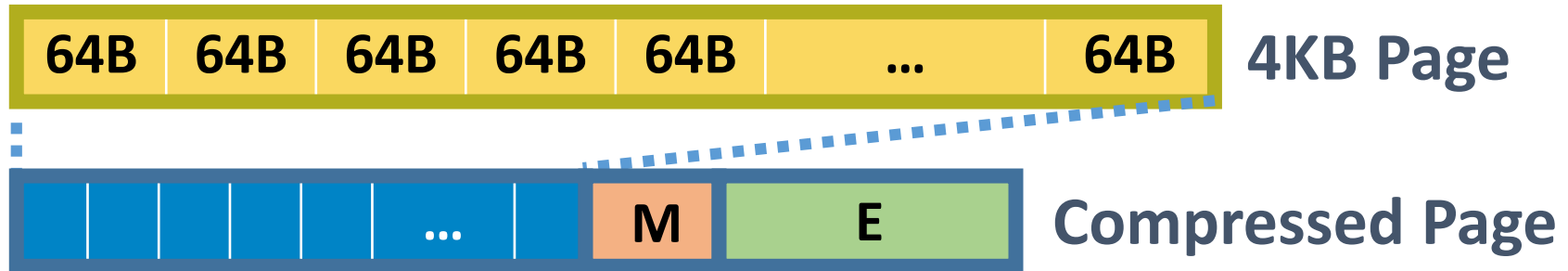
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



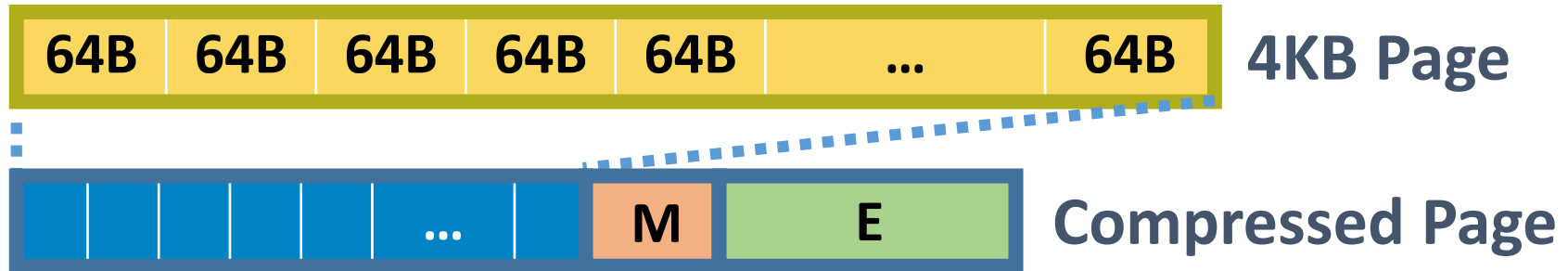
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



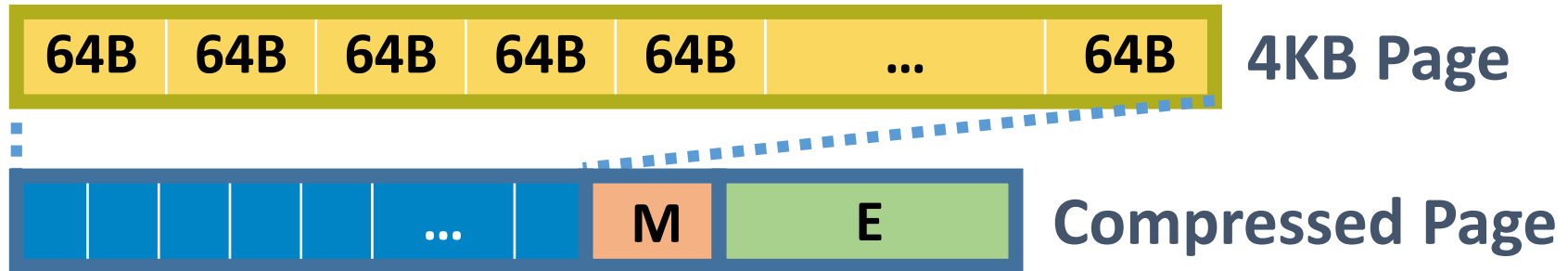
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



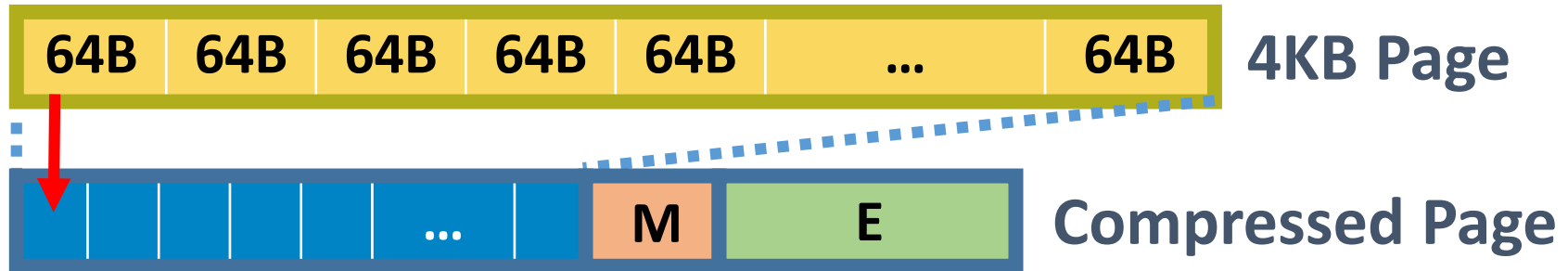
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



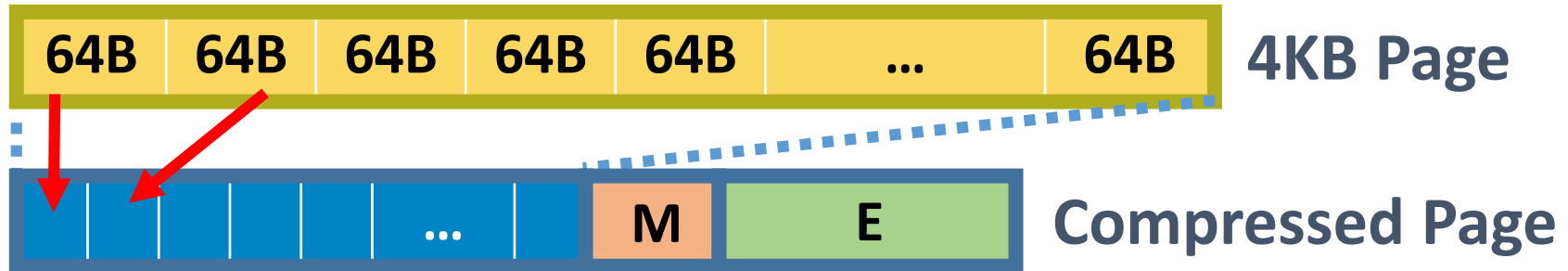
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



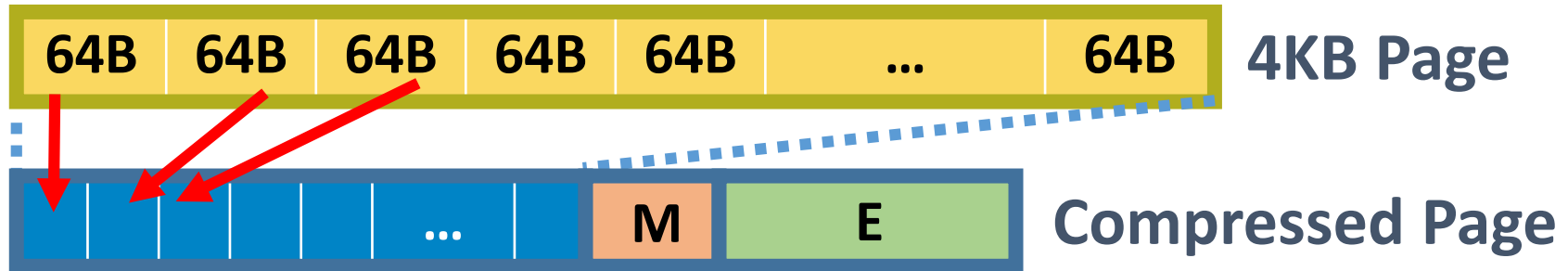
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



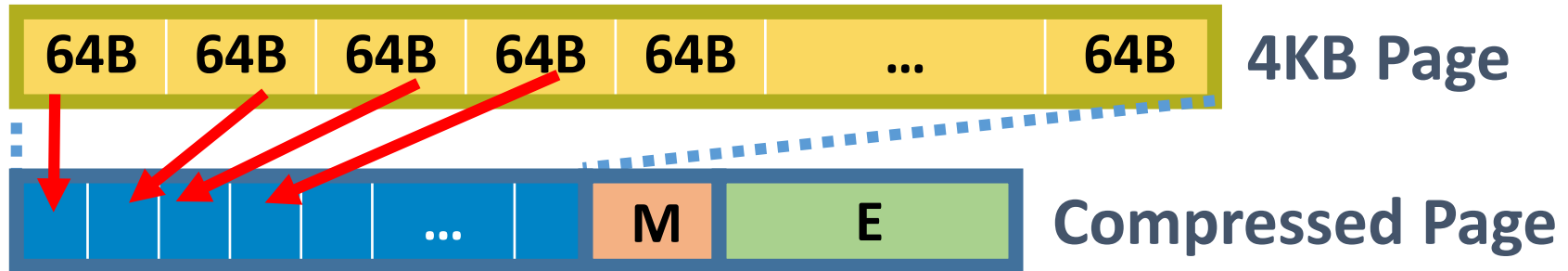
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



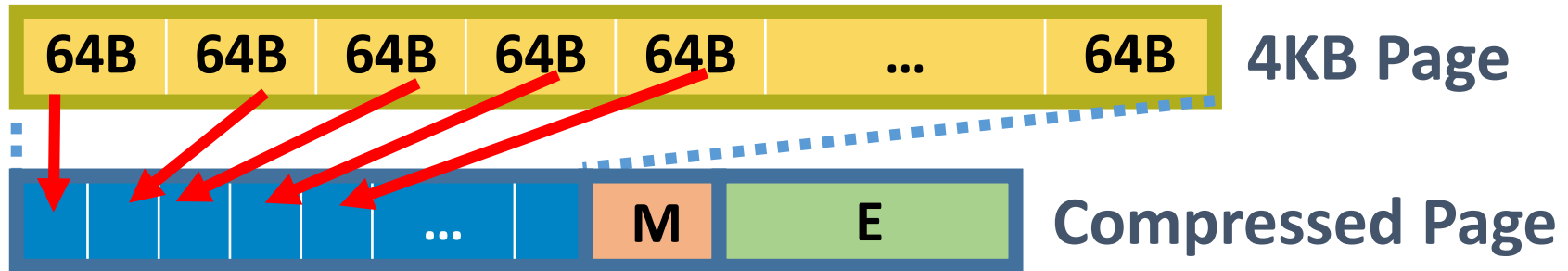
- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP^[2]: OS Page table to locate compressed data



- MXT^[3]: OS transparent approach to locate data

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

➤ LCP[2]: OS Page table to locate compressed data



Simple cacheline offset calculation



Bound to largest compression size

➤ MXT[3]: OS transparent approach to locate data

PTR	PTR	PTR	PTR
PTR	PTR	PTR	PTR

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

➤ LCP[2]: OS Page table to locate compressed data



Simple cacheline offset calculation



Bound to largest compression size

➤ MXT[3]: OS transparent approach to locate data

PTR	PTR	PTR	PTR
PTR	PTR	PTR	PTR

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

➤ LCP[2]: OS Page table to locate compressed data

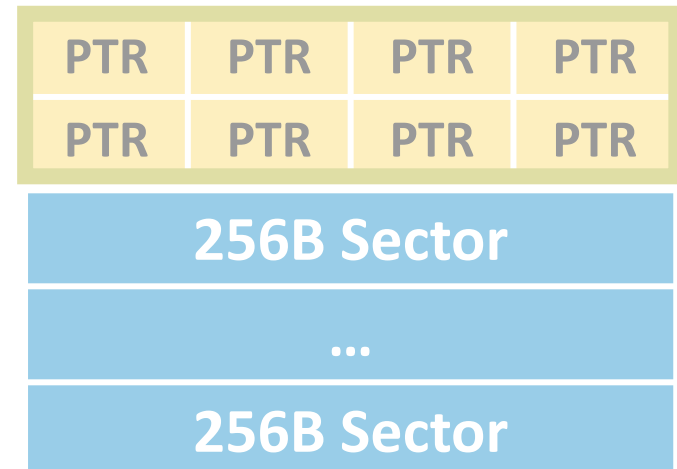


Simple cacheline offset calculation



Bound to largest compression size

➤ MXT[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

➤ LCP[2]: OS Page table to locate compressed data

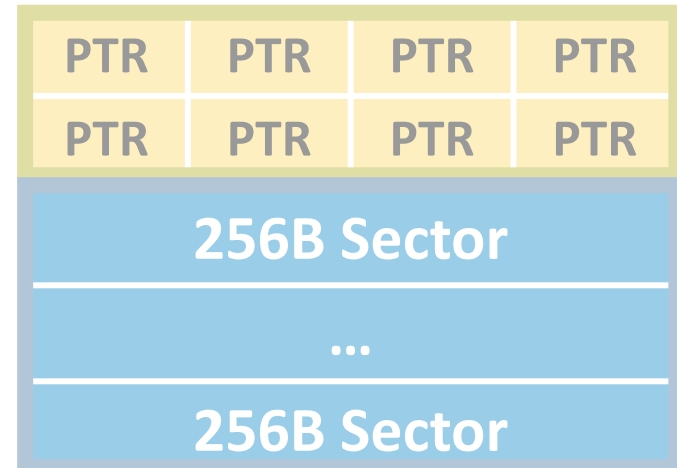


Simple cacheline offset calculation



Bound to largest compression size

➤ MXT[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

➤ LCP[2]: OS Page table to locate compressed data

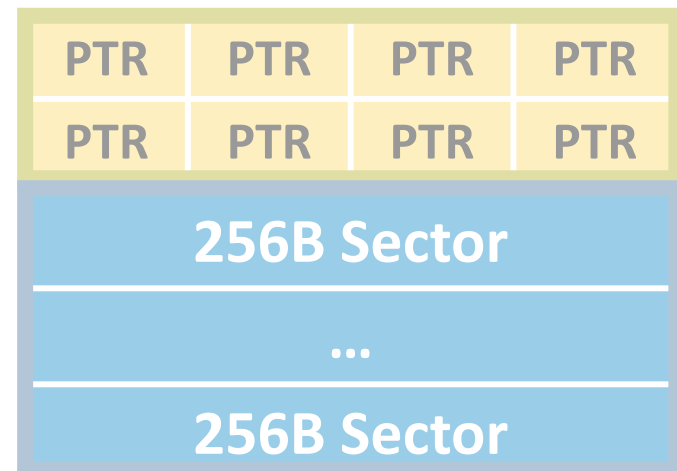


Simple cacheline offset calculation



Bound to largest compression size

➤ MXT[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP[2]: OS Page table to locate compressed data

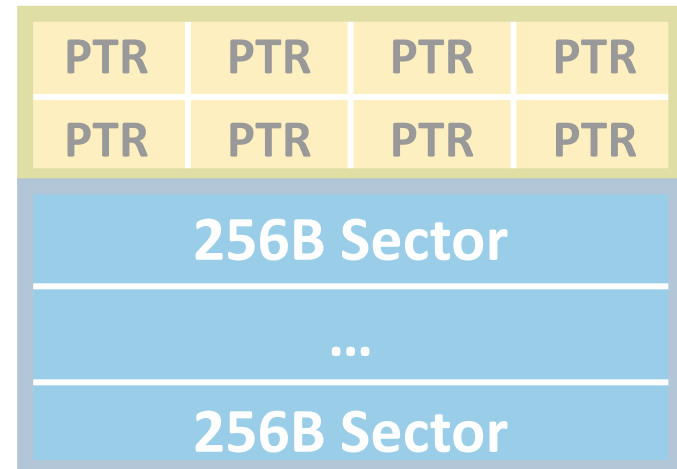
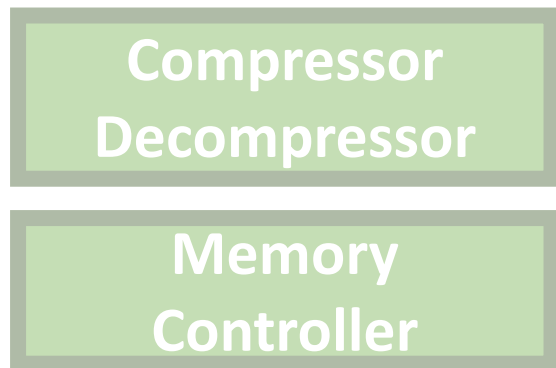


Simple cacheline offset calculation



Bound to largest compression size

- MXT^[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP[2]: OS Page table to locate compressed data

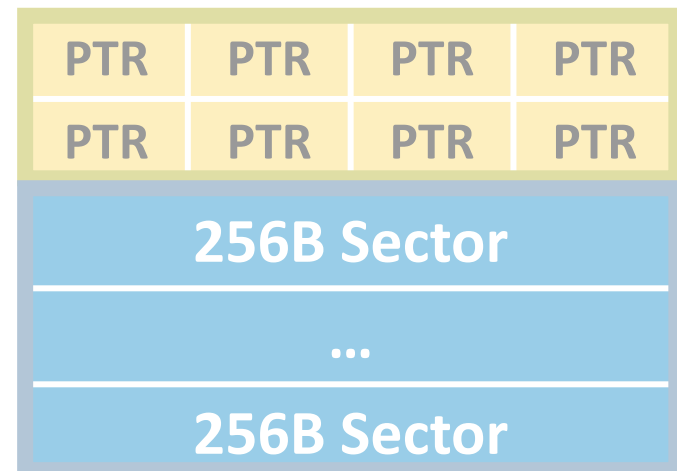
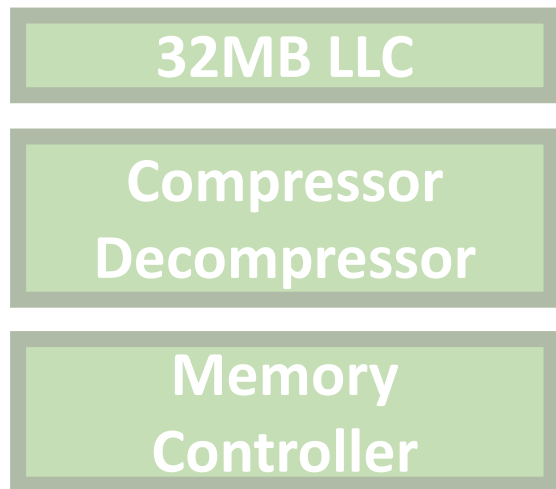


Simple cacheline offset calculation



Bound to largest compression size

- MXT^[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

➤ LCP[2]: OS Page table to locate compressed data

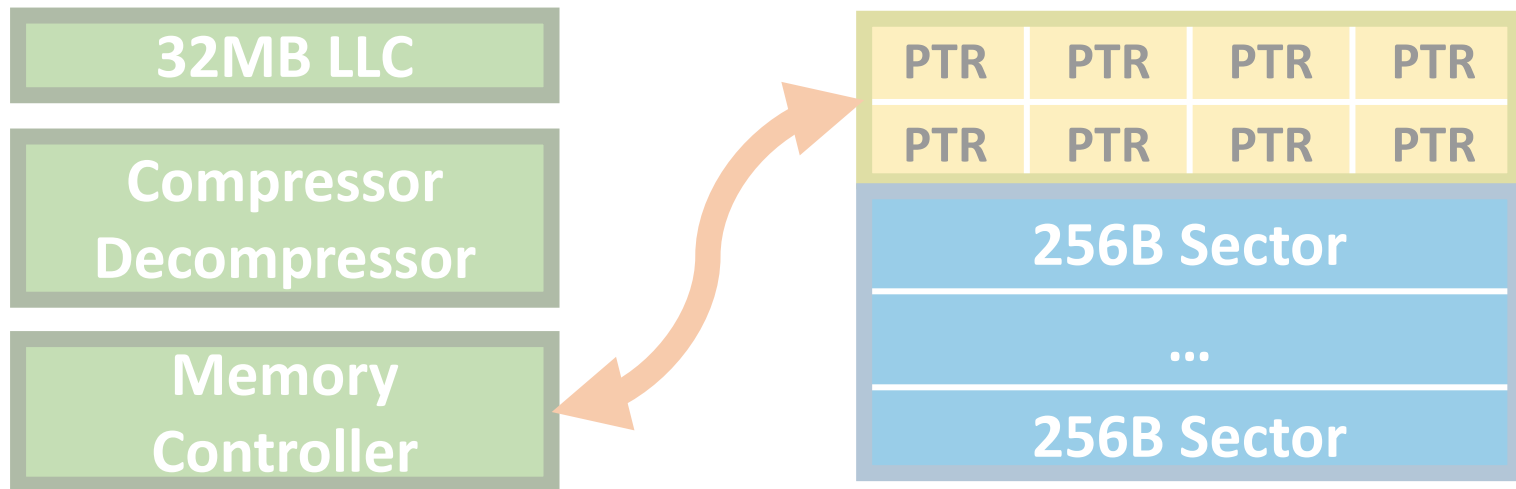


Simple cacheline offset calculation



Bound to largest compression size

➤ MXT[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP[2]: OS Page table to locate compressed data

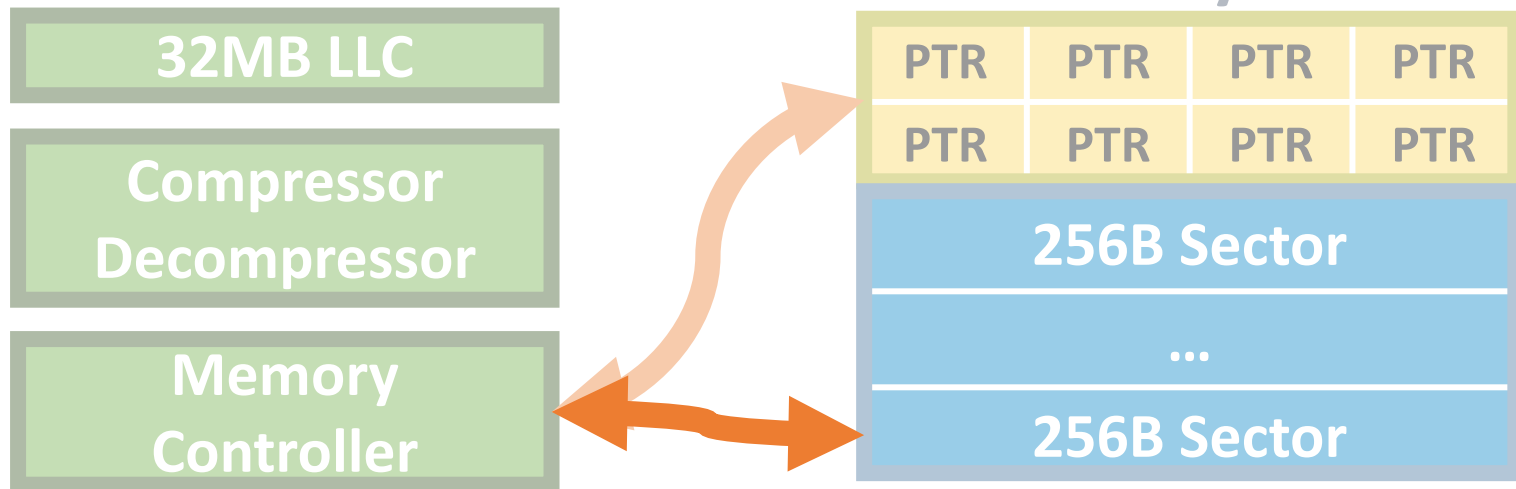


Simple cacheline offset calculation



Bound to largest compression size

- MXT^[3]: OS transparent approach to locate data
Memory-Side



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP[2]: OS Page table to locate compressed data



Simple cacheline offset calculation



Bound to largest compression size

- MXT^[3]: OS transparent approach to locate data

CPU-Side

32MB LLC

Compressor
Decompressor

Memory
Controller

Memory-Side

PTR	PTR	PTR	PTR
PTR	PTR	PTR	PTR

256B Sector

...

256B Sector

[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP[2]: OS Page table to locate compressed data

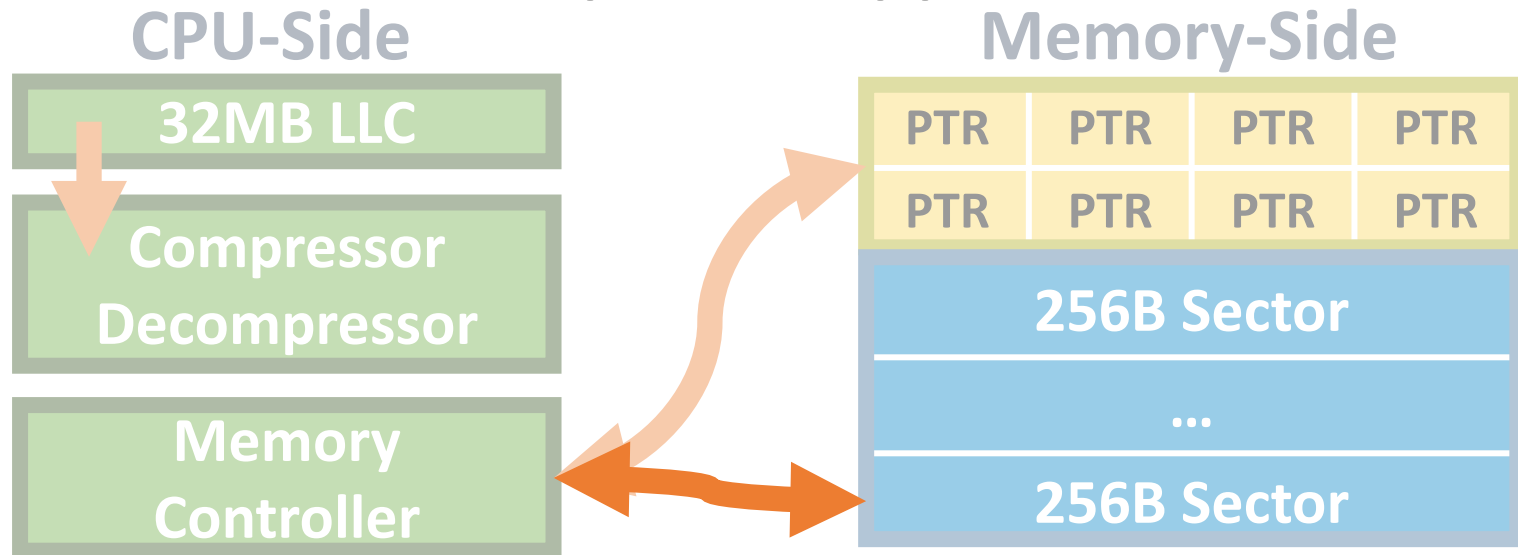


Simple cacheline offset calculation



Bound to largest compression size

- MXT^[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP[2]: OS Page table to locate compressed data

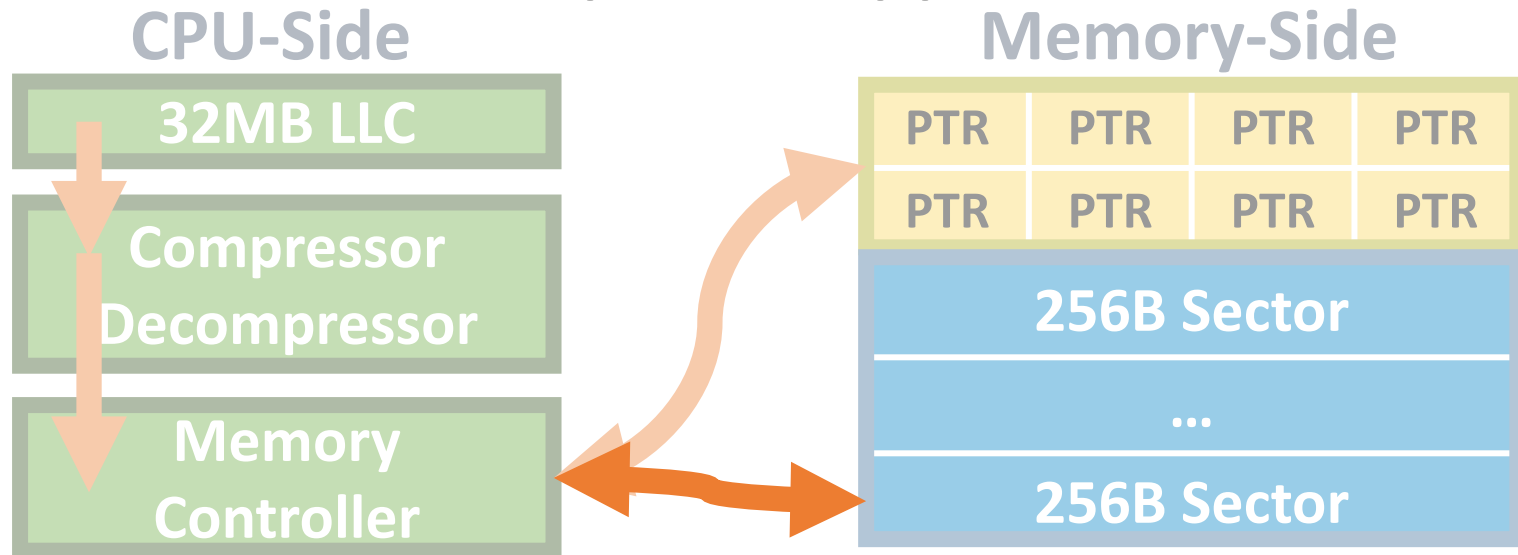


Simple cacheline offset calculation



Bound to largest compression size

- MXT^[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP[2]: OS Page table to locate compressed data

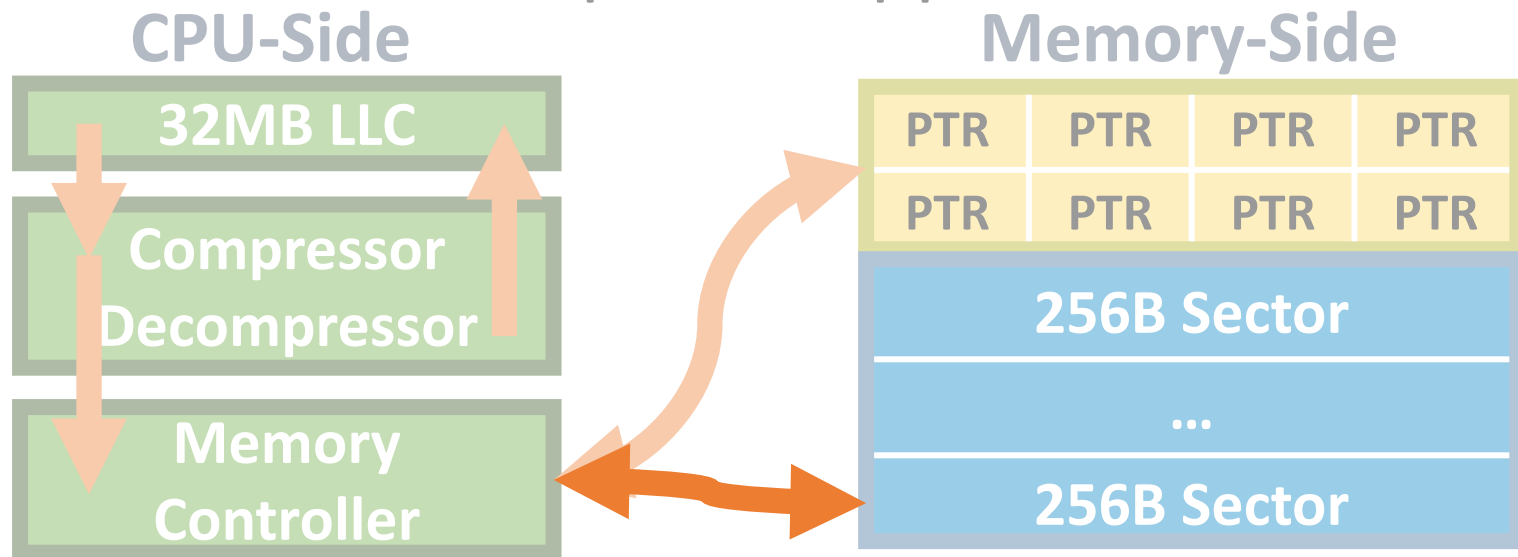


Simple cacheline offset calculation



Bound to largest compression size

- MXT^[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP[2]: OS Page table to locate compressed data

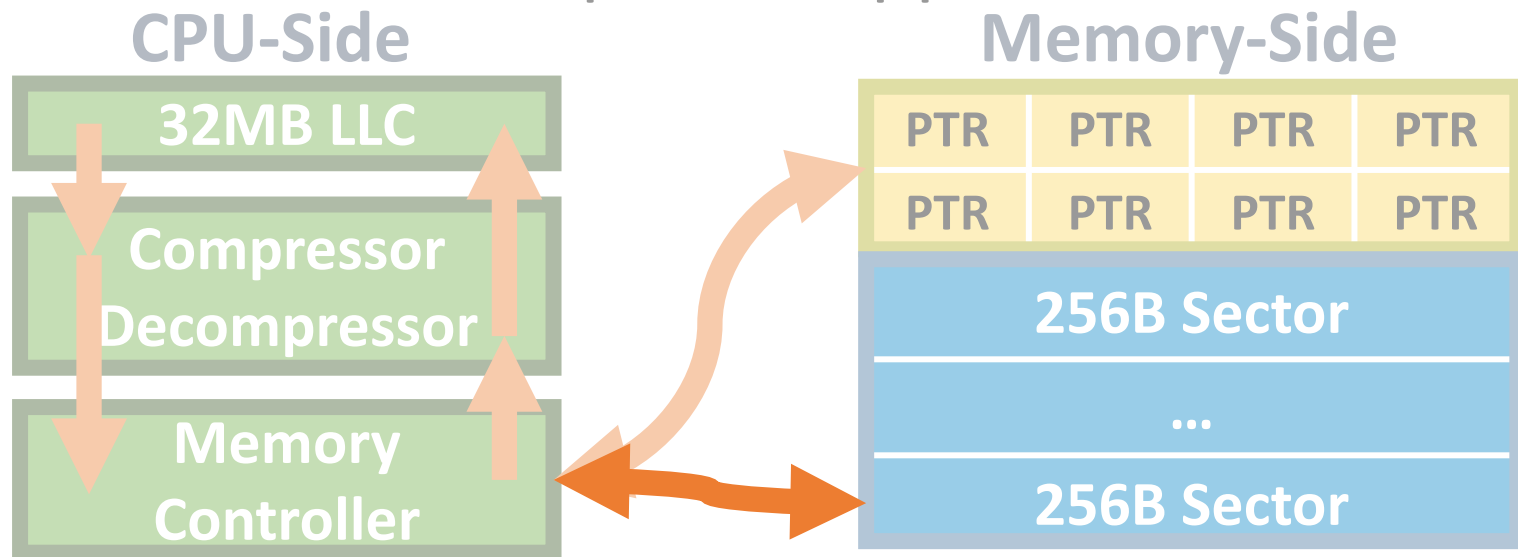


Simple cacheline offset calculation



Bound to largest compression size

- MXT^[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP[2]: OS Page table to locate compressed data

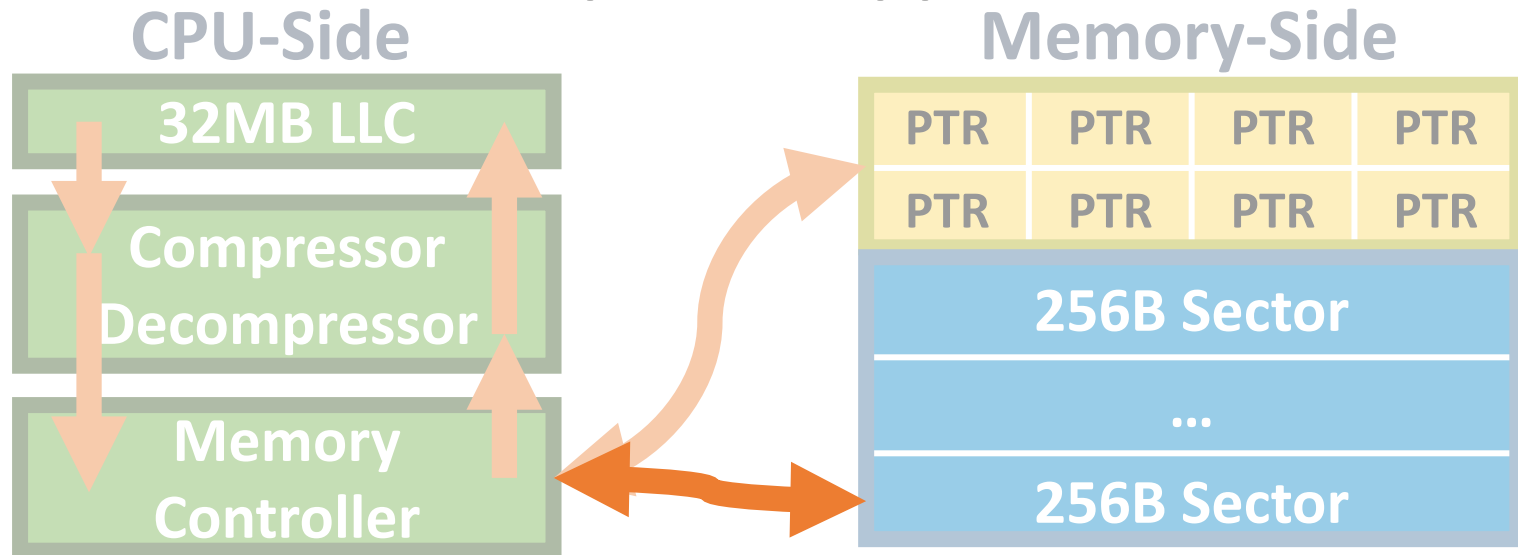


Simple cacheline offset calculation



Bound to largest compression size

- MXT^[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

- LCP[2]: OS Page table to locate compressed data

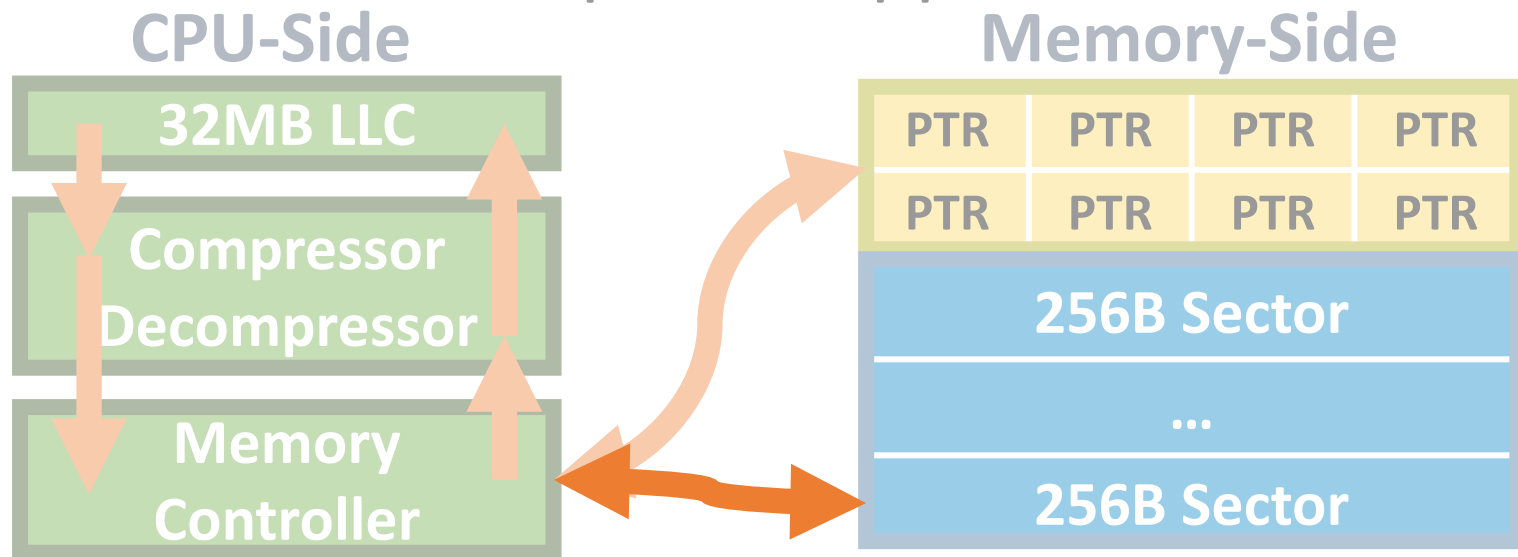


Simple cacheline offset calculation



Bound to largest compression size

- MXT^[3]: OS transparent approach to locate data



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

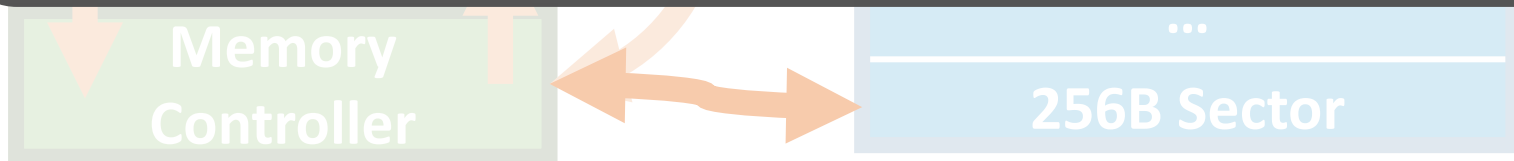
How to Locate Compressed Data?

- LCP[2]: OS Page table to locate compressed data

- 👍 Simple cacheline offset calculation
- 👎 Bound to largest compression size

- MXT[3]: OS transparent approach to locate data

- 👍 OS transparent address translation
- 👎 Requires large (32MB) cache



[2] Pekhimenko et al. MICRO-46

[3] Tremaine et al. IBM Journal of Research and Development, March 2001

How to Locate Compressed Data?

➤ LCP[2]: OS Page table to locate compressed data



Simple cacheline offset calculation



Bound to largest compression size

Dual Memory Compression

Similar to MXT in general

LCP for active region (latency optimized)



OS transparent address translation



Requires large (32MB) cache

Memory
Controller



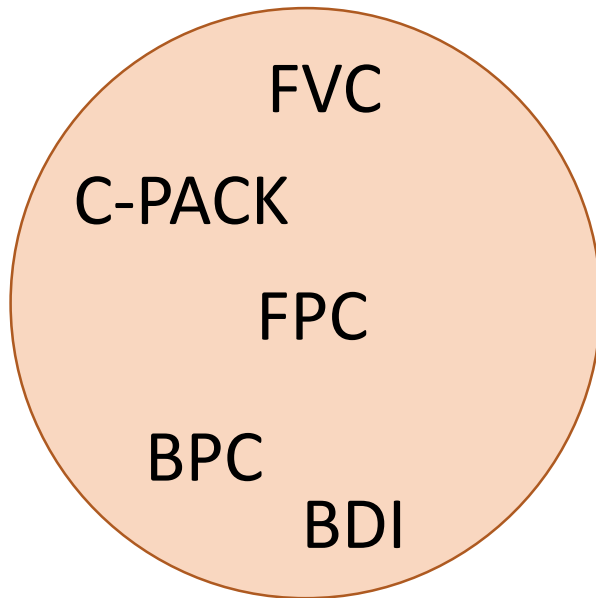
256B Sector

[2] Pekhimenko et al. MICRO-46

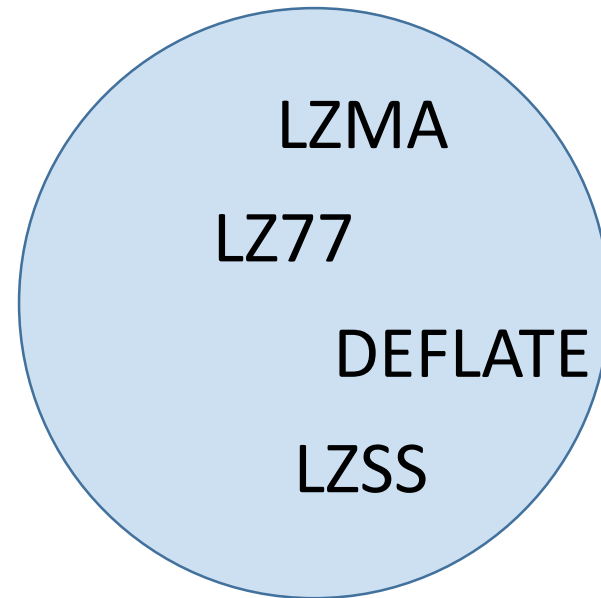
[3] Tremaine et al. IBM Journal of Research and Development, March 2001

Dual Memory Compression (DMC)

- Goal of DMC
 - Short decompression latency
 - High compression ratio
 - OS transparent



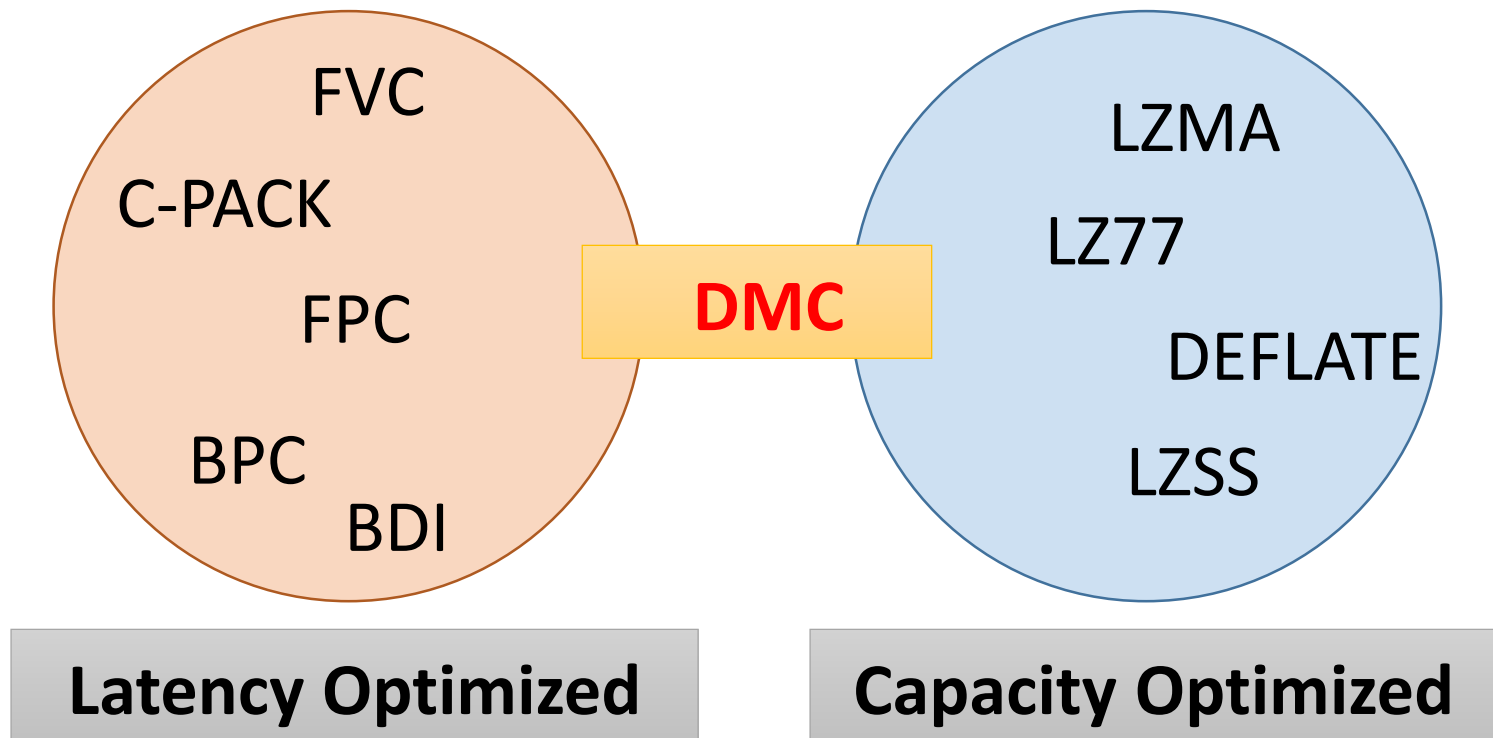
Latency Optimized



Capacity Optimized

Dual Memory Compression (DMC)

- Goal of DMC
 - Short decompression latency
 - High compression ratio
 - OS transparent



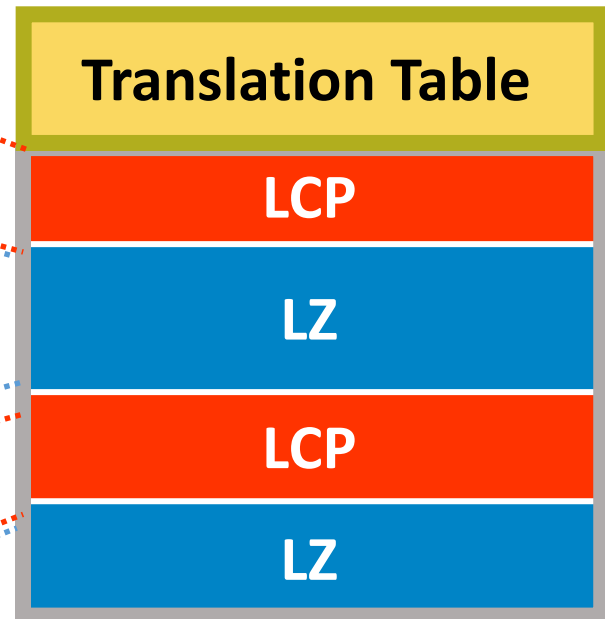
DMC: Basic Idea

- 5.9% 1KB memory block access per 500M instr

Uncompressed Space



Compressed Space



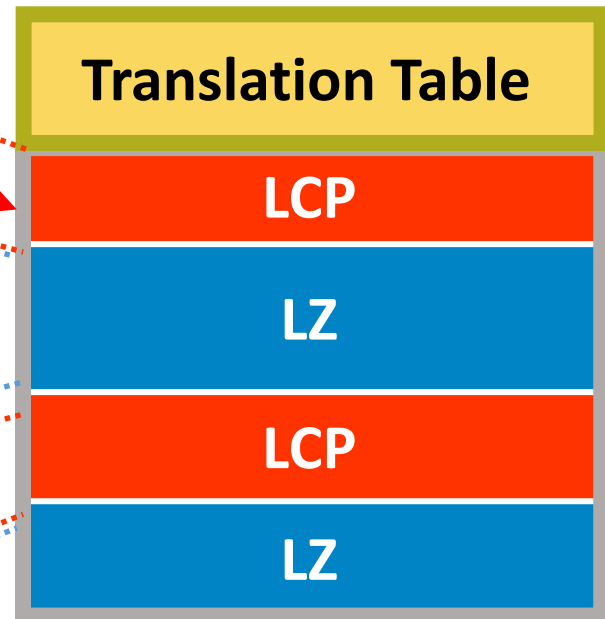
DMC: Basic Idea

- 5.9% 1KB memory block access per 500M instr

Uncompressed Space



Compressed Space



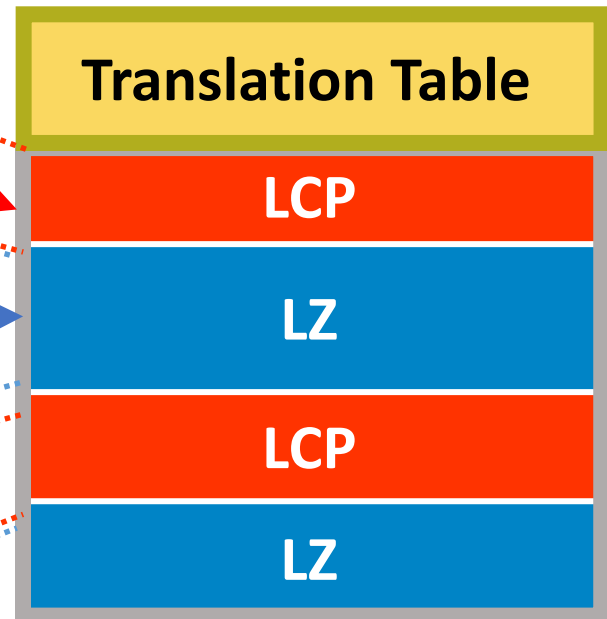
DMC: Basic Idea

- 5.9% 1KB memory block access per 500M instr

Uncompressed Space

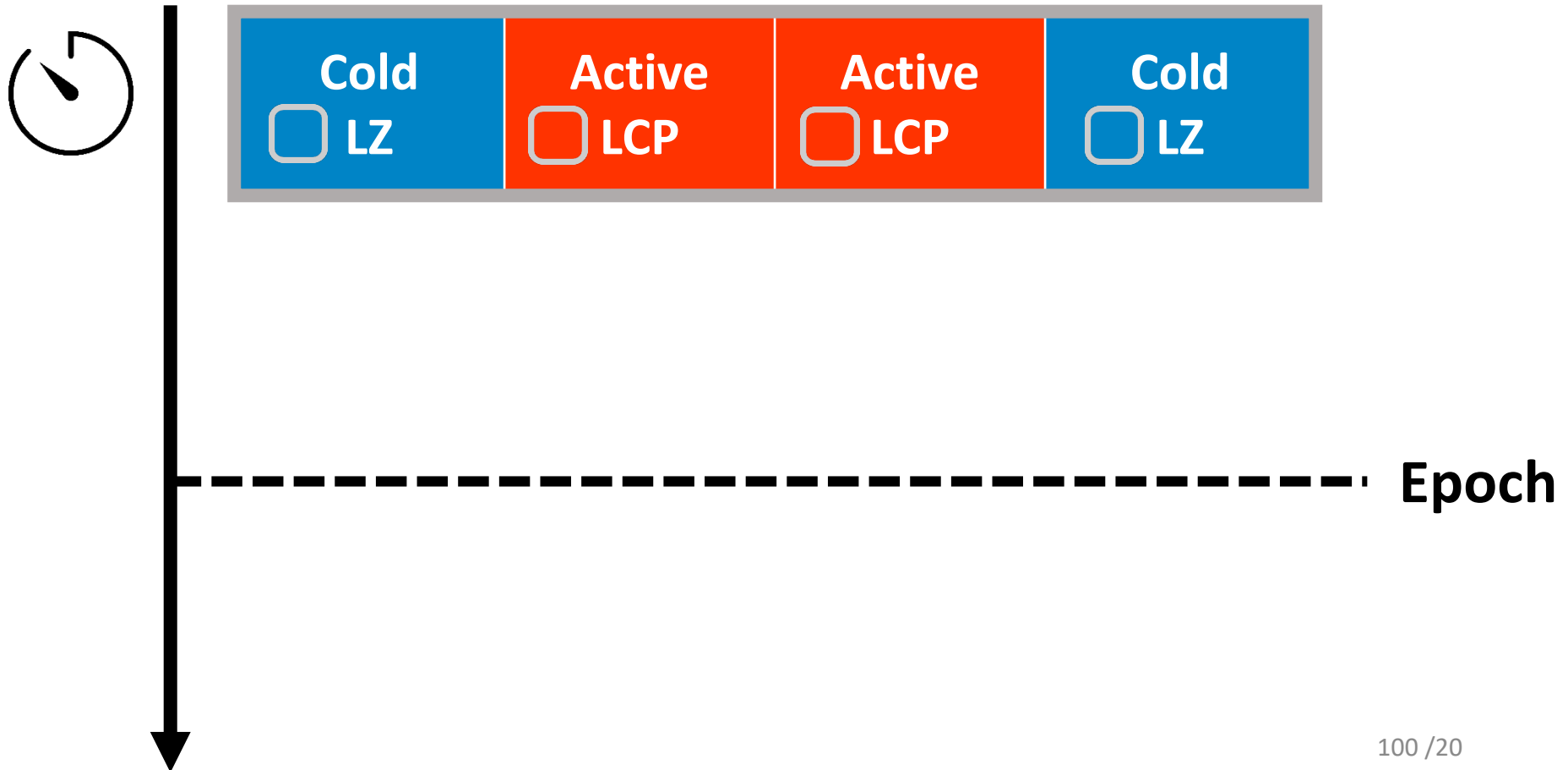


Compressed Space



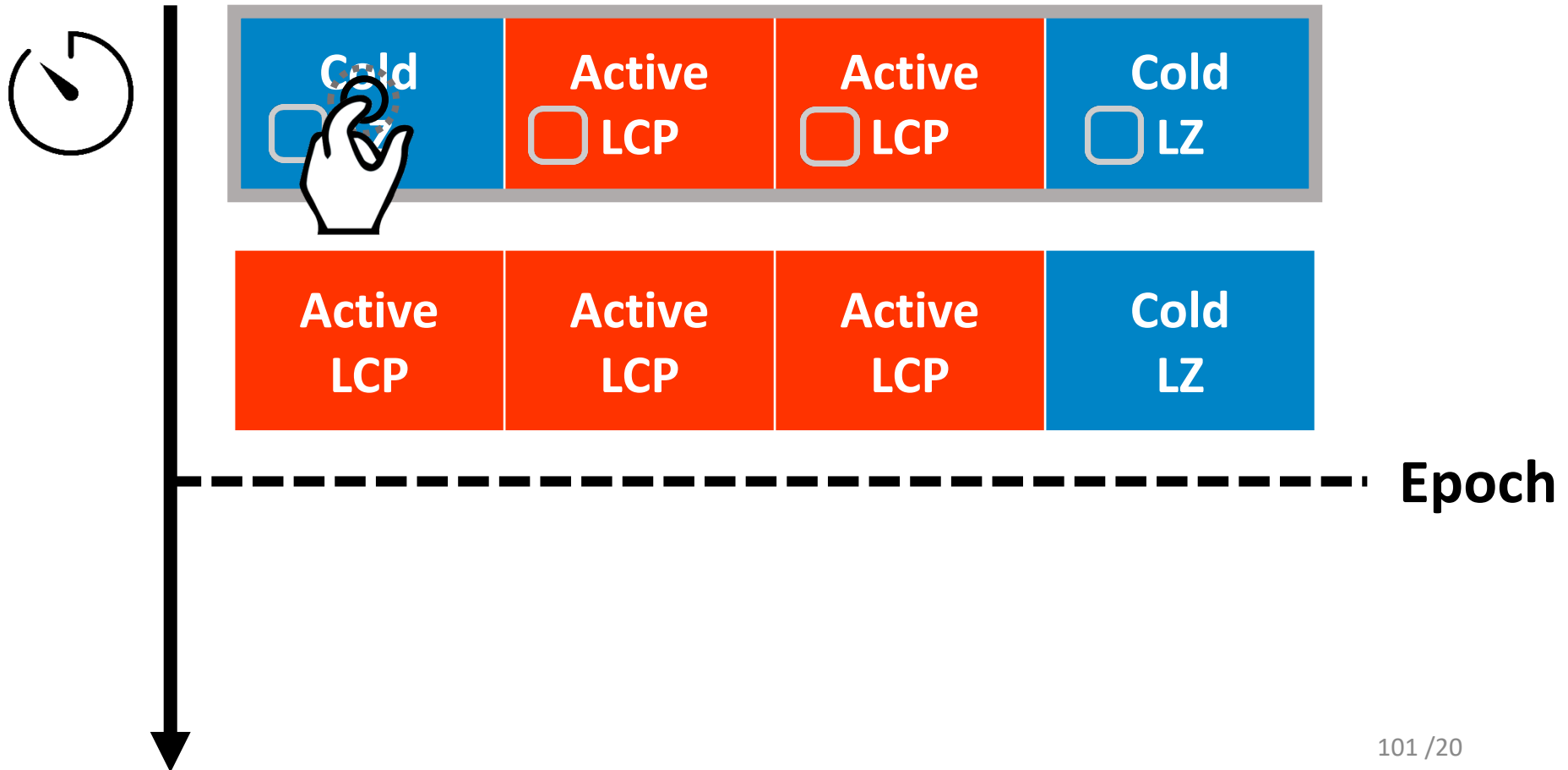
DMC: Transcompression

- On-demand LZ → LCP
- Periodic LCP → LZ



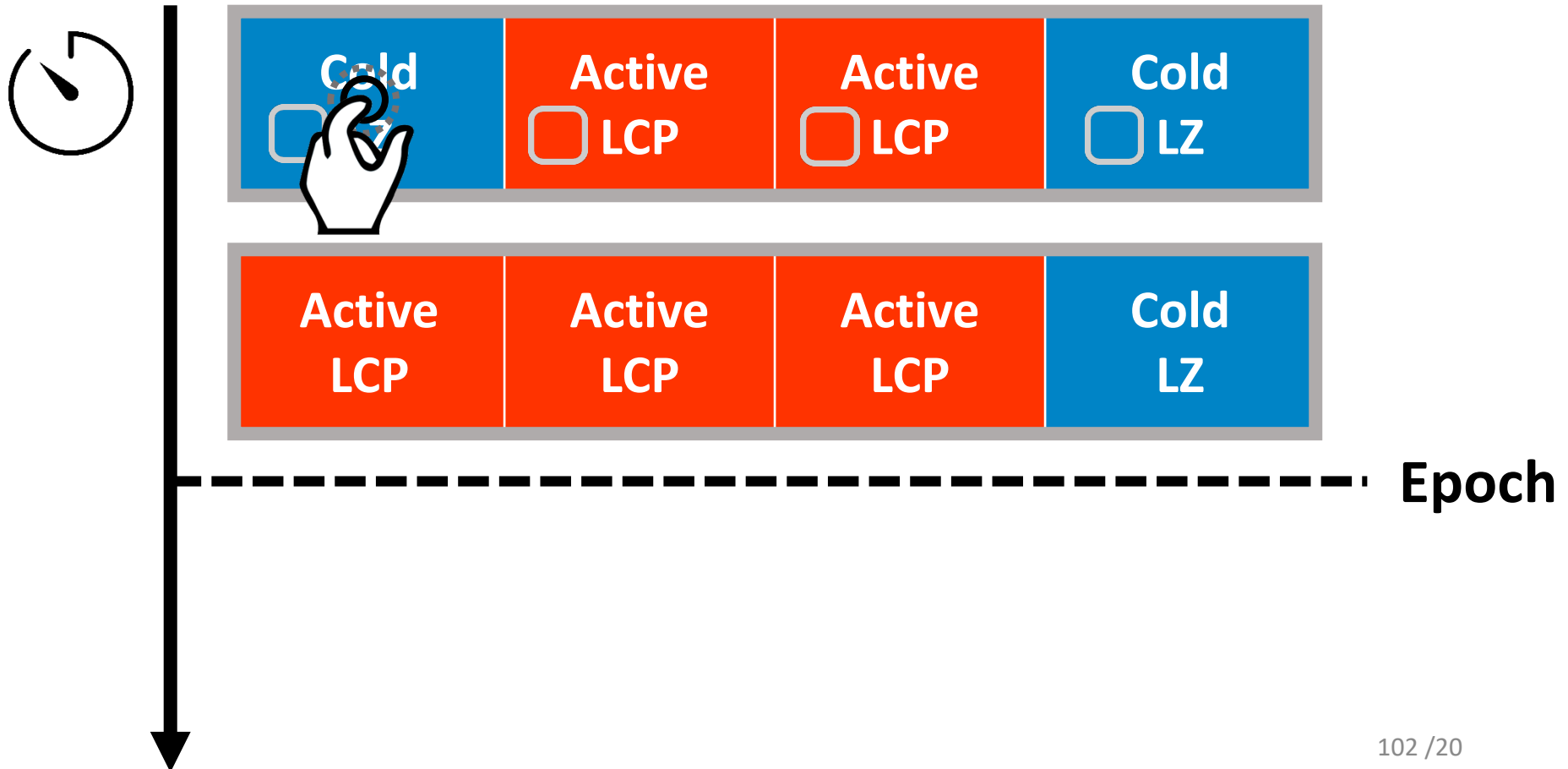
DMC: Transcompression

- On-demand LZ → LCP
- Periodic LCP → LZ



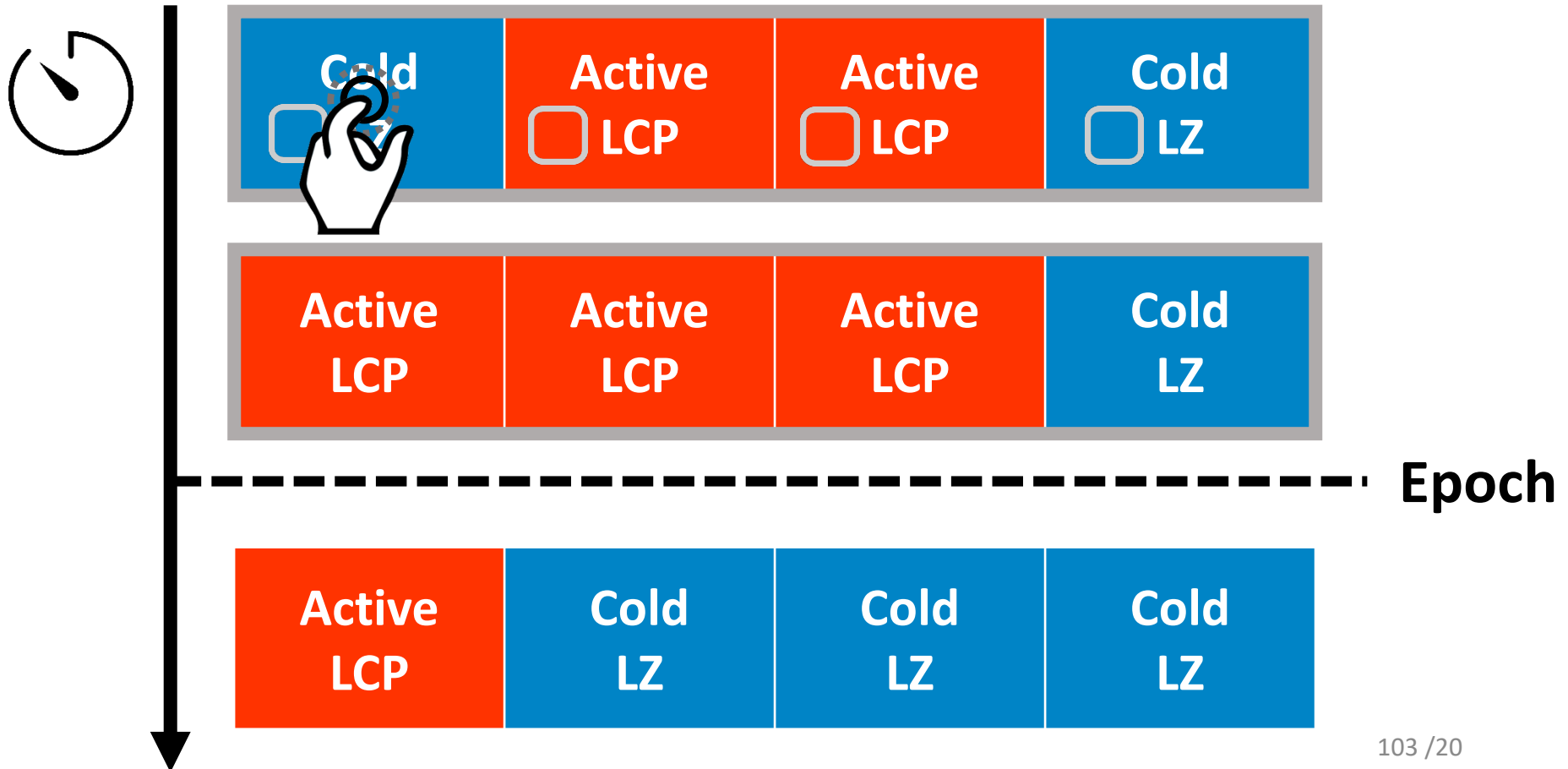
DMC: Transcompression

- On-demand LZ → LCP
- Periodic LCP → LZ



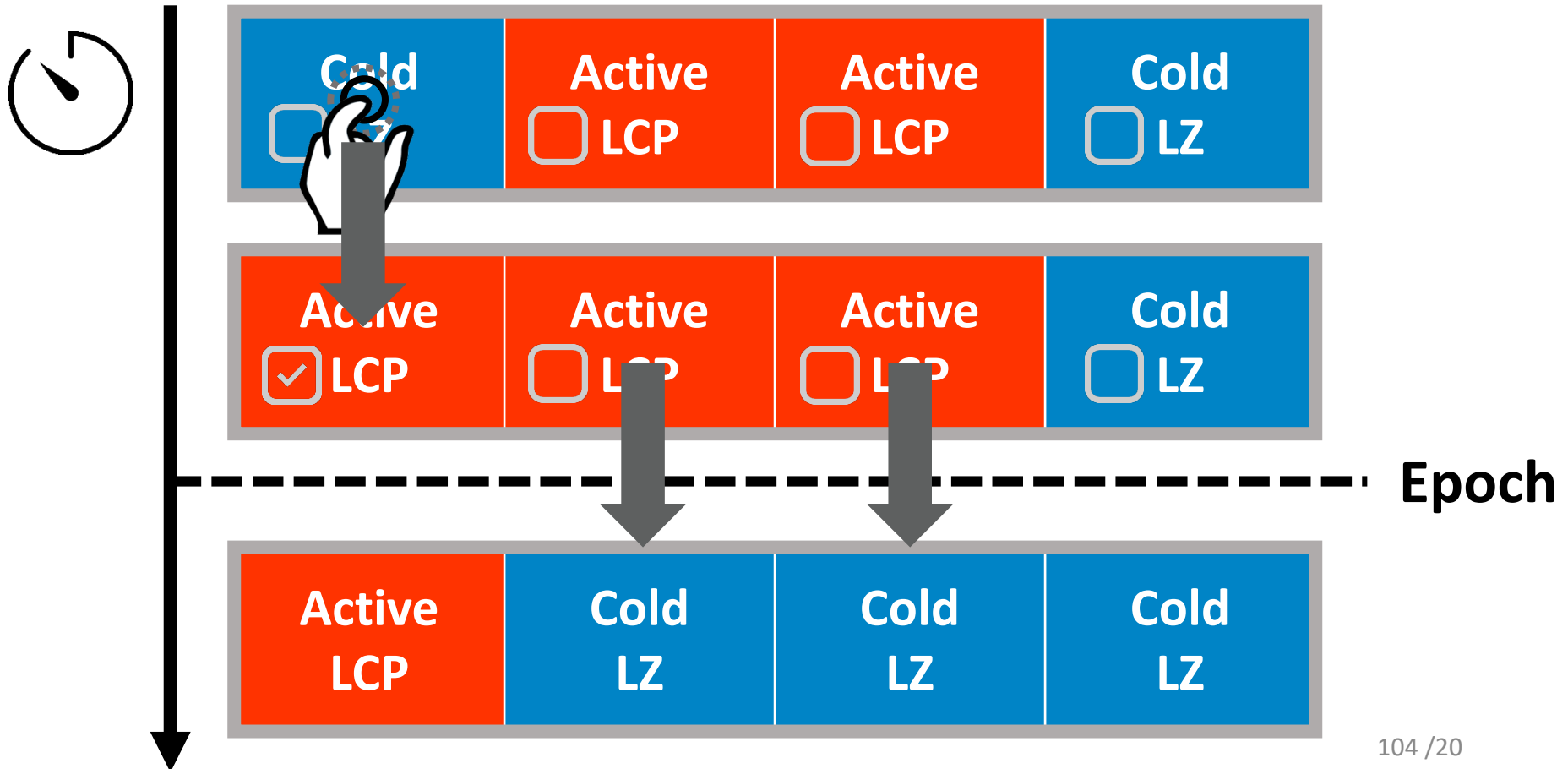
DMC: Transcompression

- On-demand LZ → LCP
- Periodic LCP → LZ



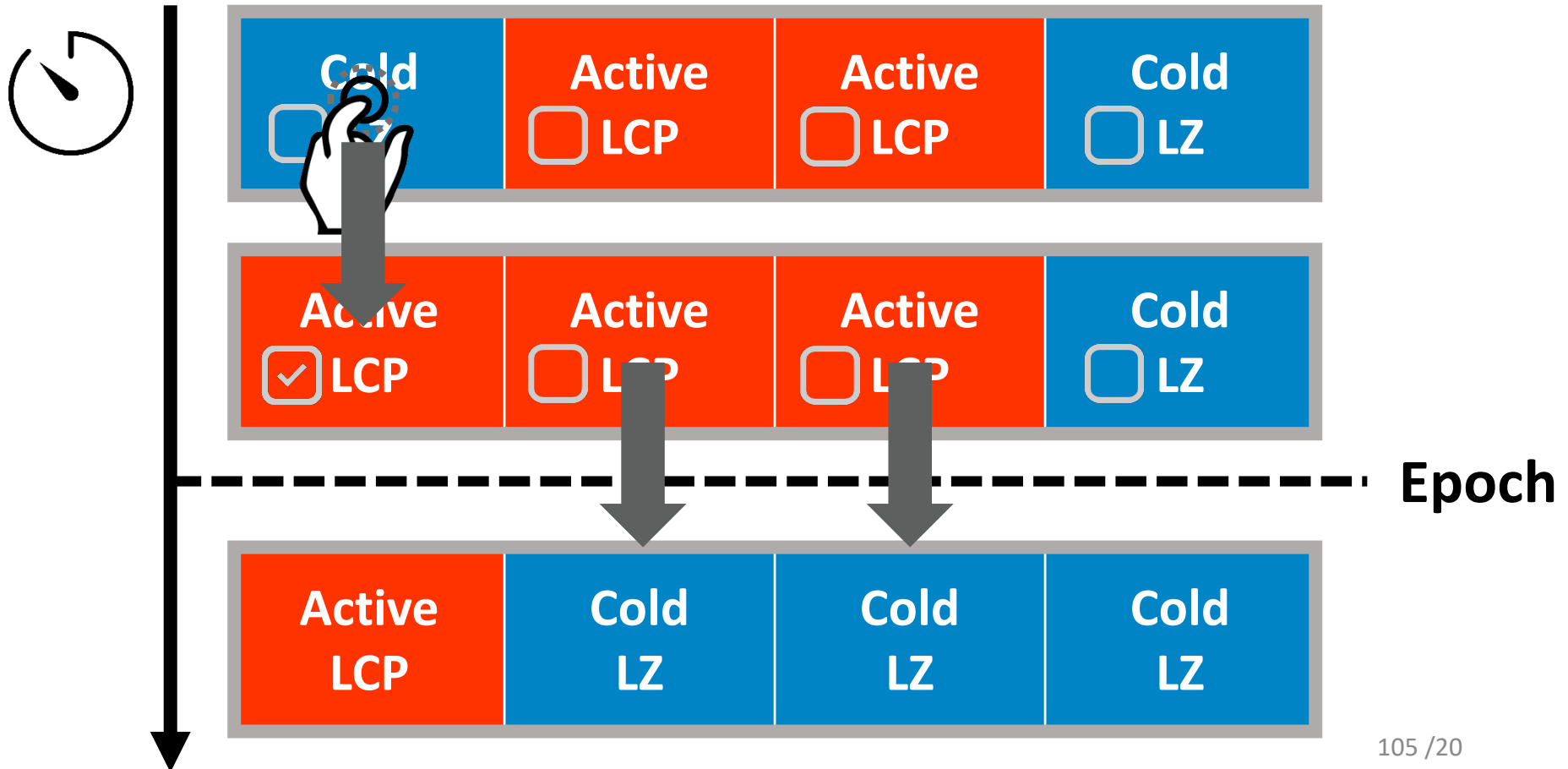
DMC: Transcompression

- On-demand LZ → LCP
- Periodic LCP → LZ



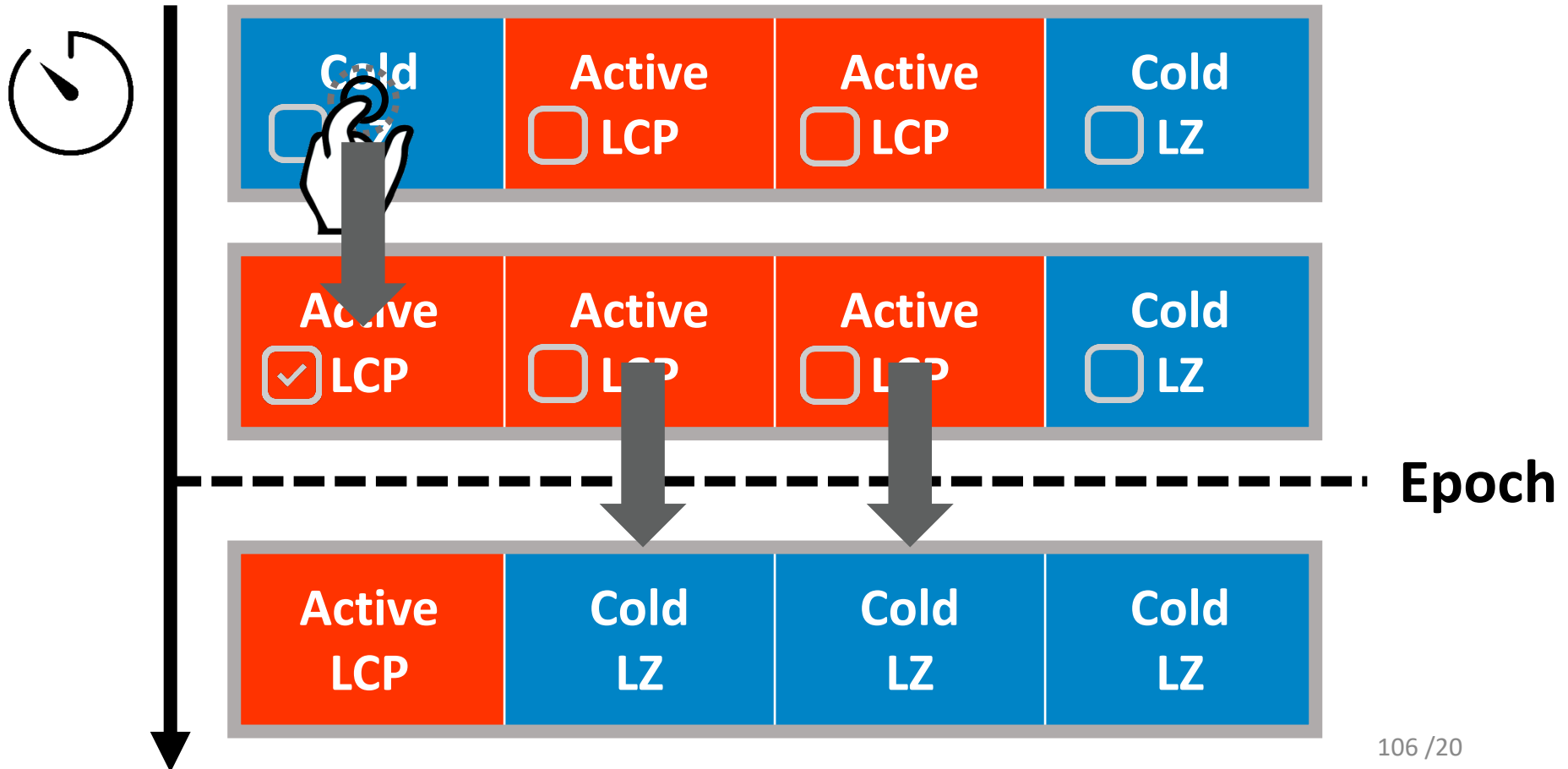
DMC: Transcompression

- On-demand LZ → LCP
- Periodic LCP → LZ



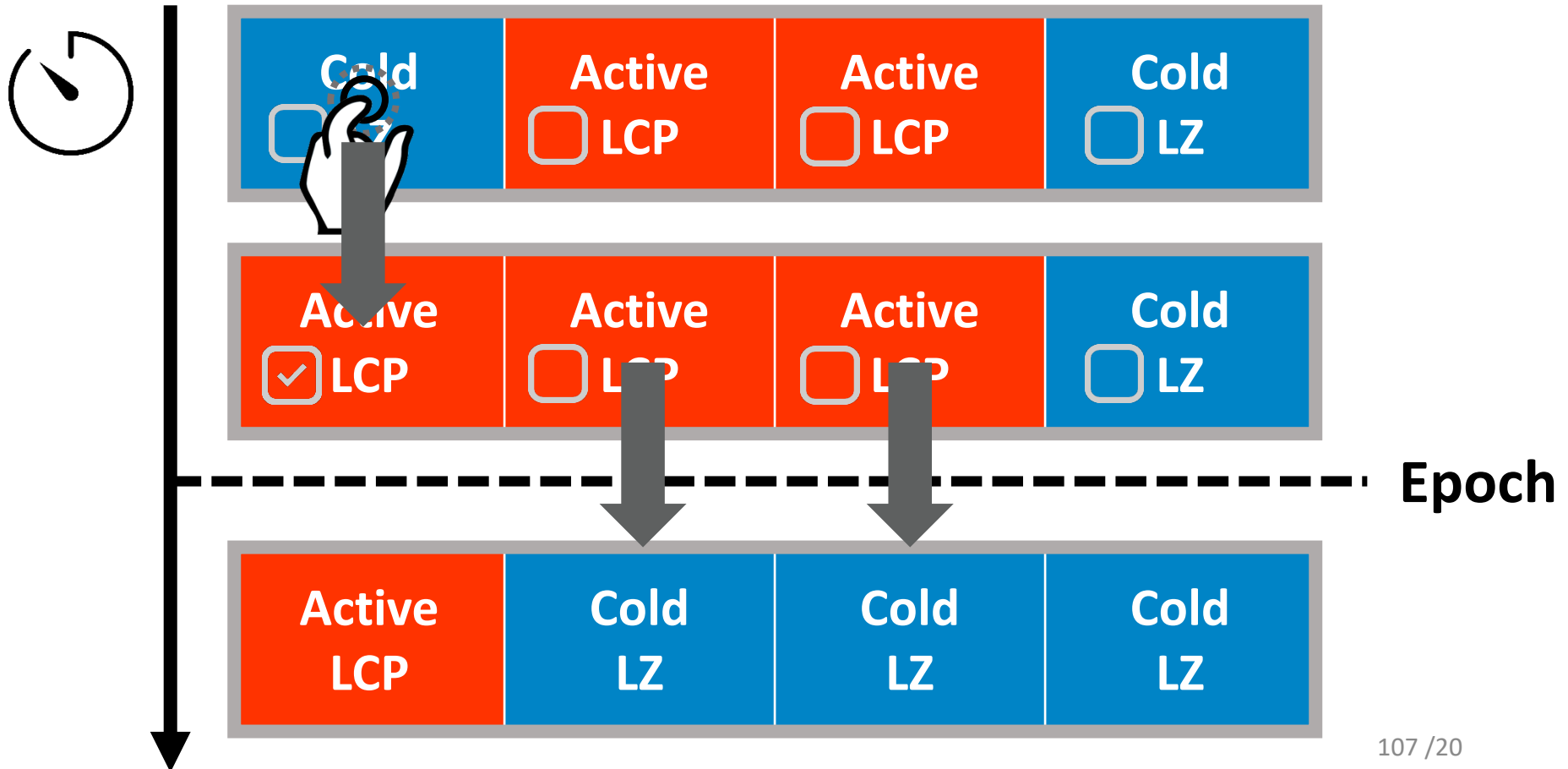
DMC: Transcompression

- On-demand LZ → LCP
- Periodic LCP → LZ



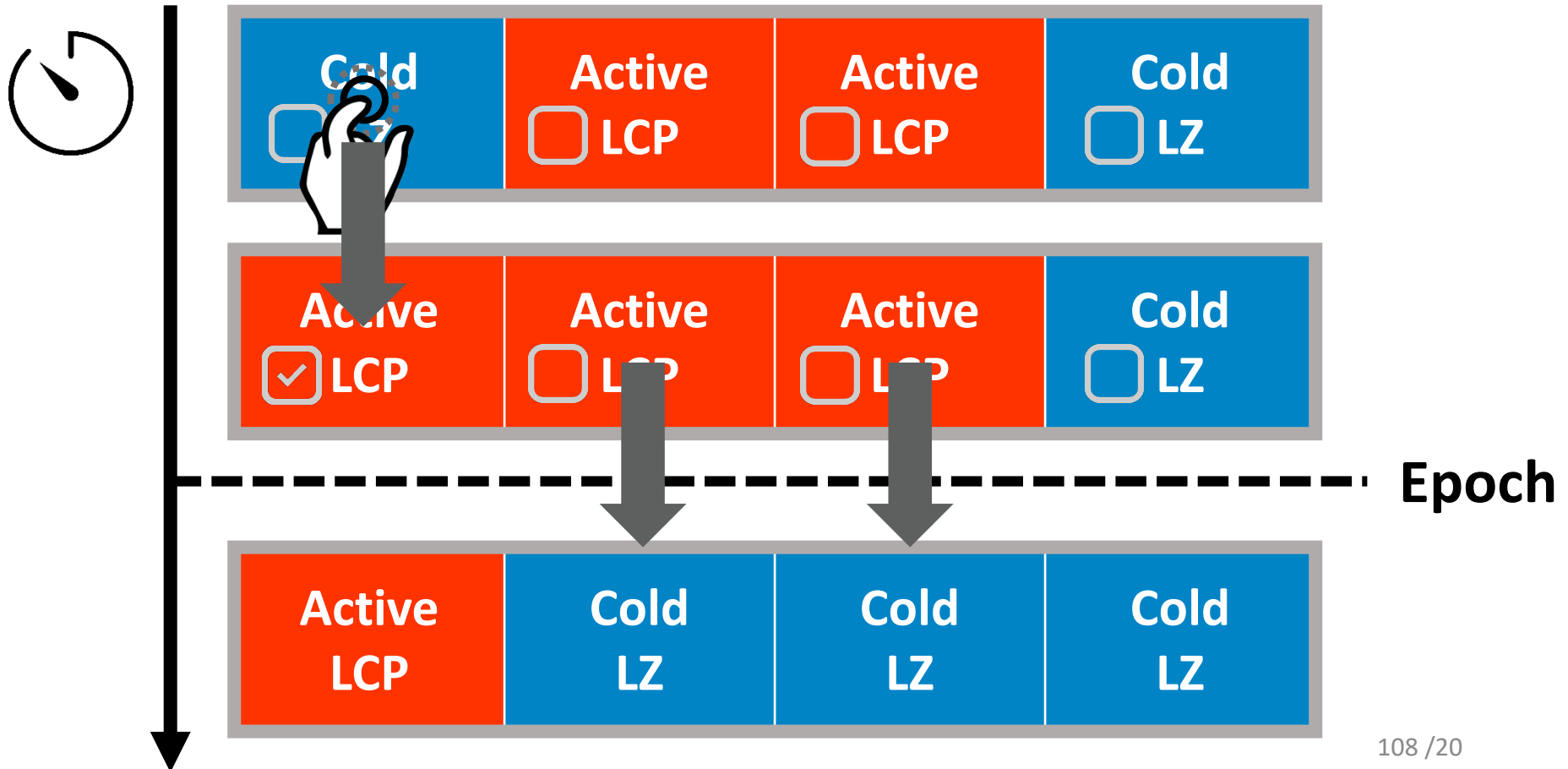
DMC: Transcompression

- On-demand LZ → LCP
- Periodic LCP → LZ



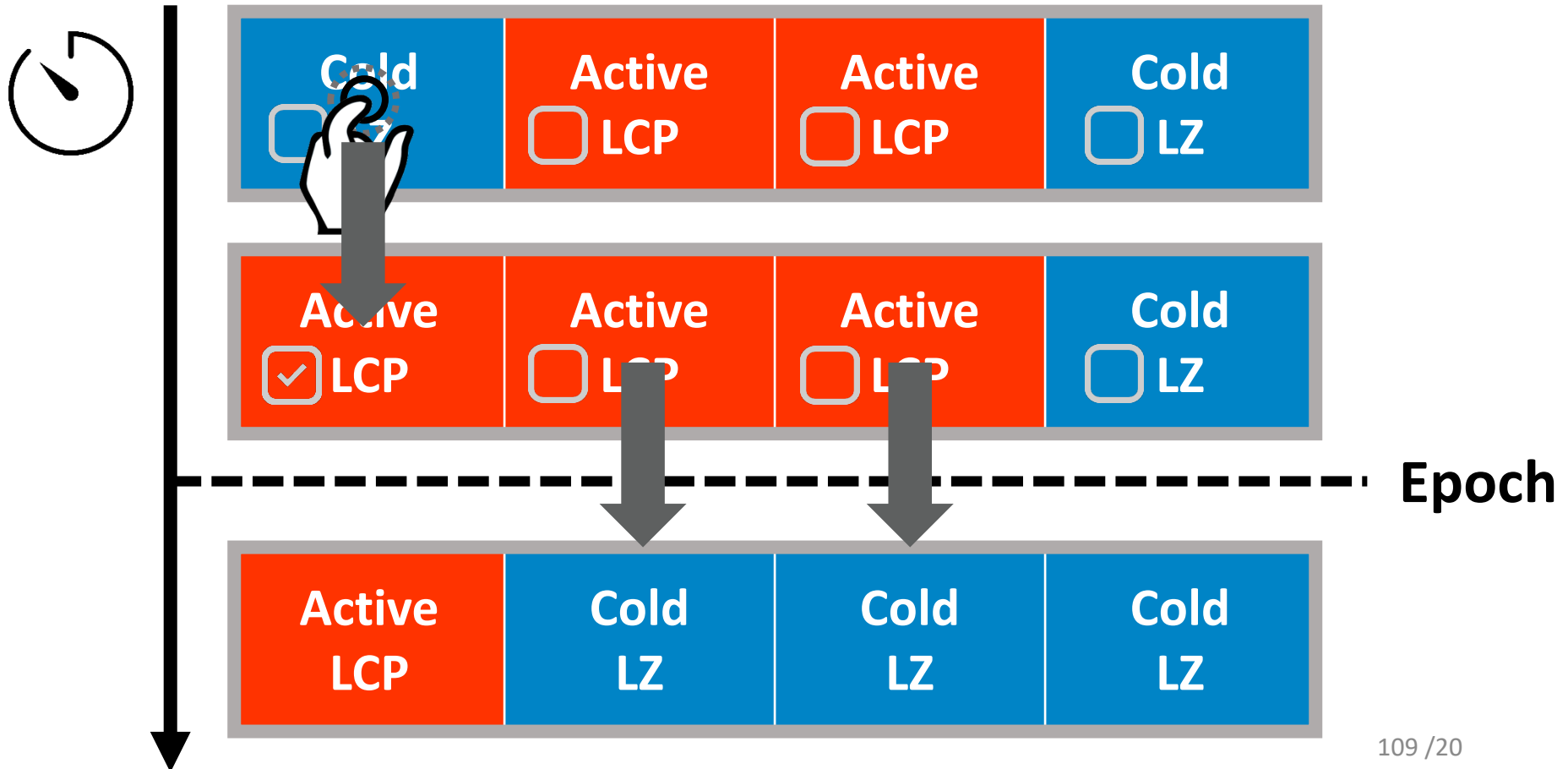
DMC: Transcompression

- On-demand LZ → LCP
- Periodic LCP → LZ



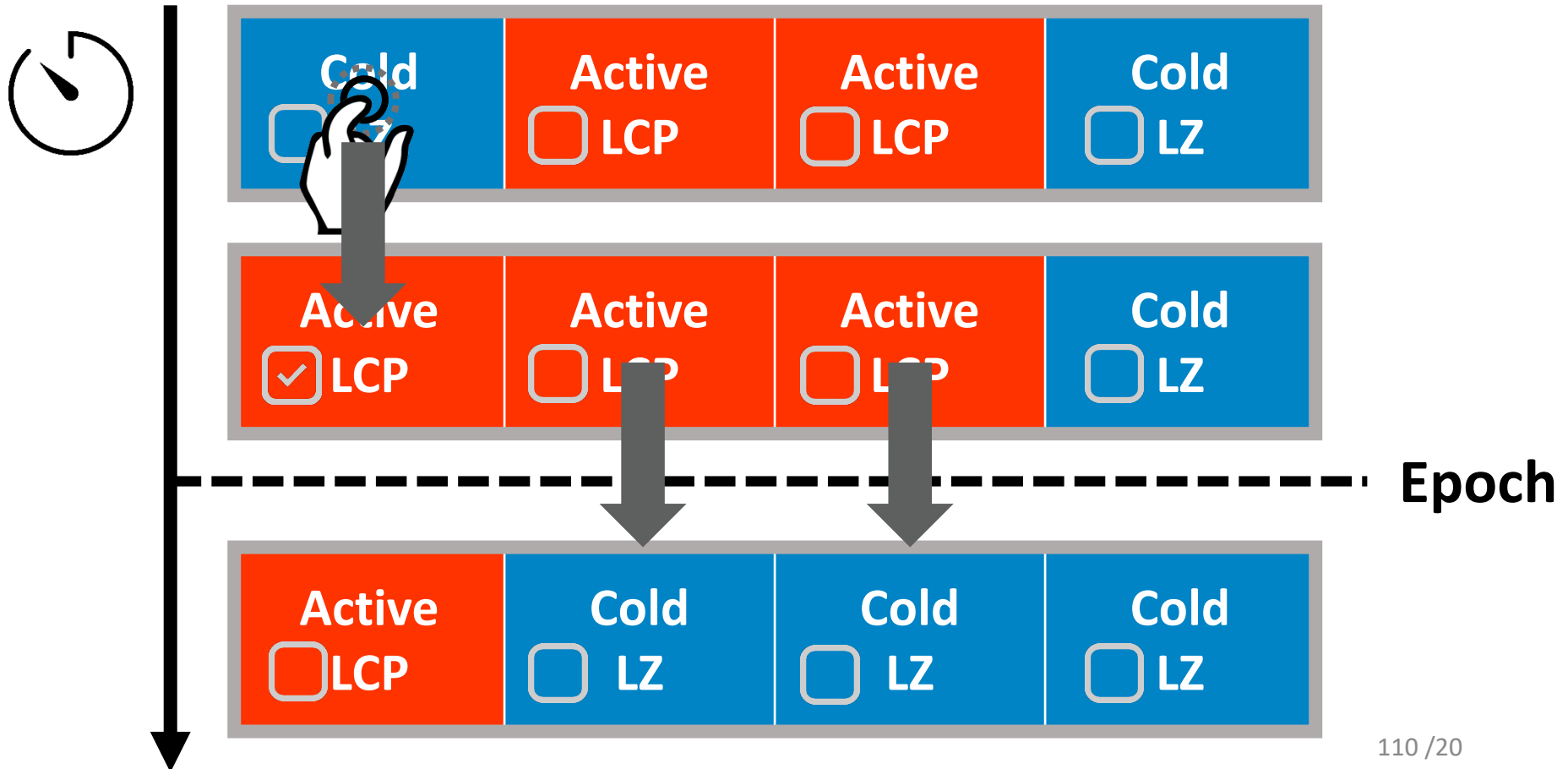
DMC: Transcompression

- On-demand LZ → LCP
- Periodic LCP → LZ



DMC: Transcompression

- On-demand LZ → LCP
- Periodic LCP → LZ



DMC: Accessing Cacheline

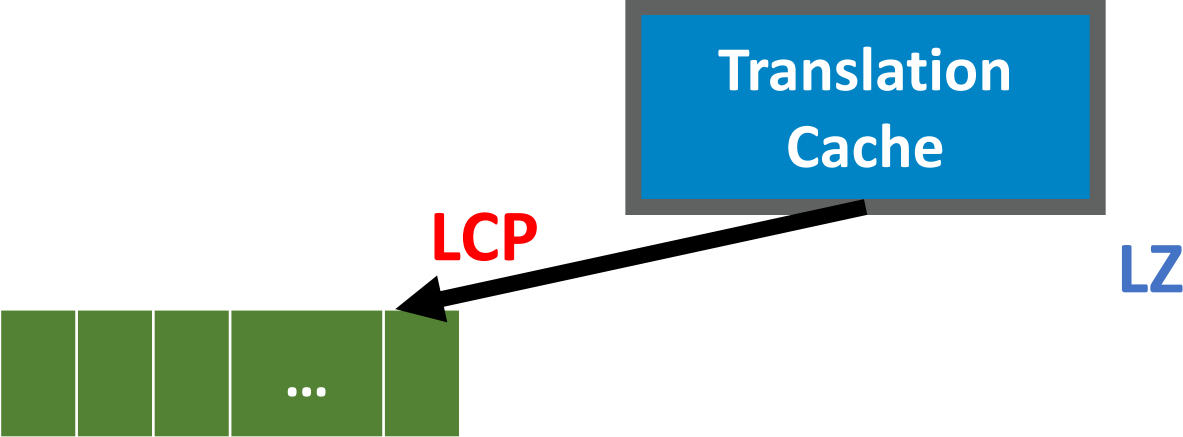


Translation
Cache

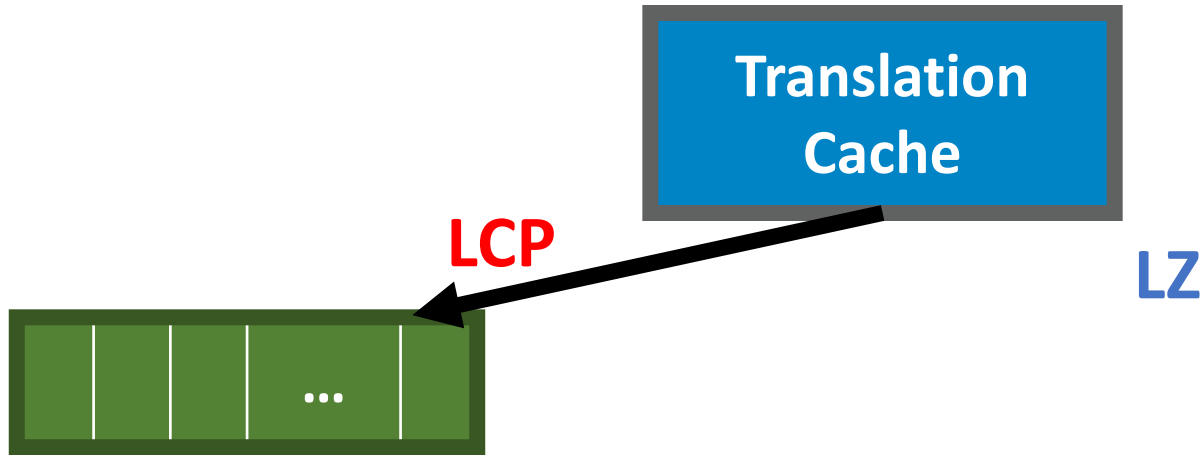
DMC: Accessing Cacheline



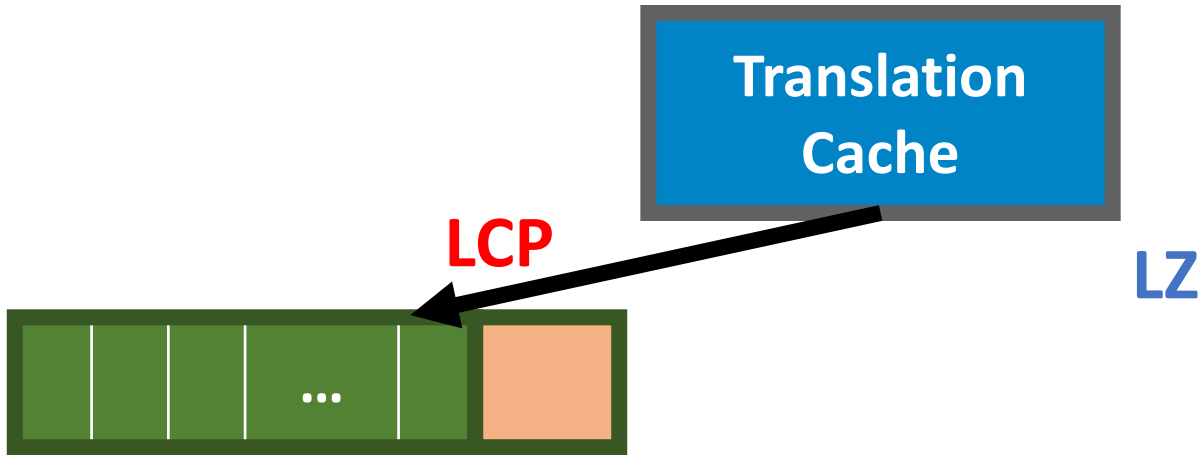
DMC: Accessing Cacheline



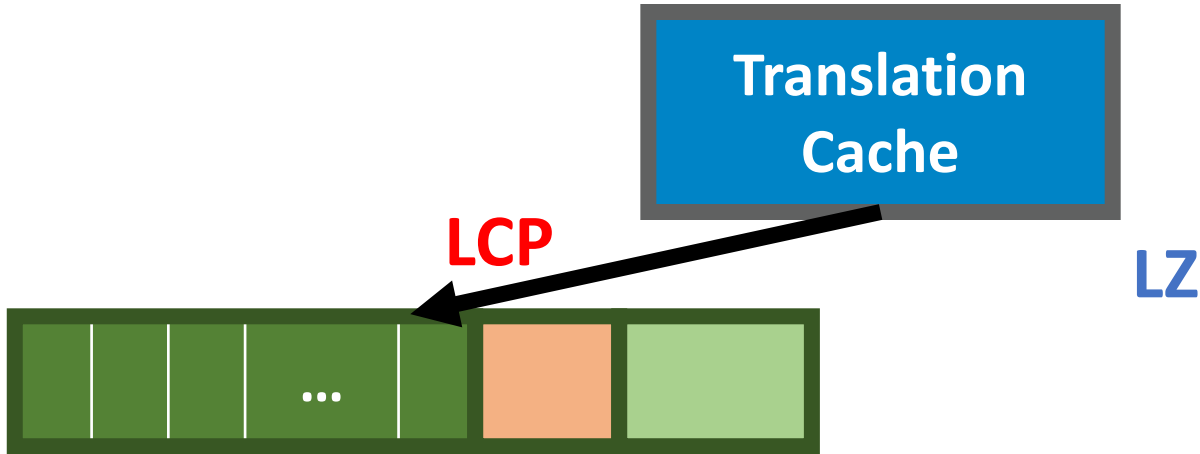
DMC: Accessing Cacheline



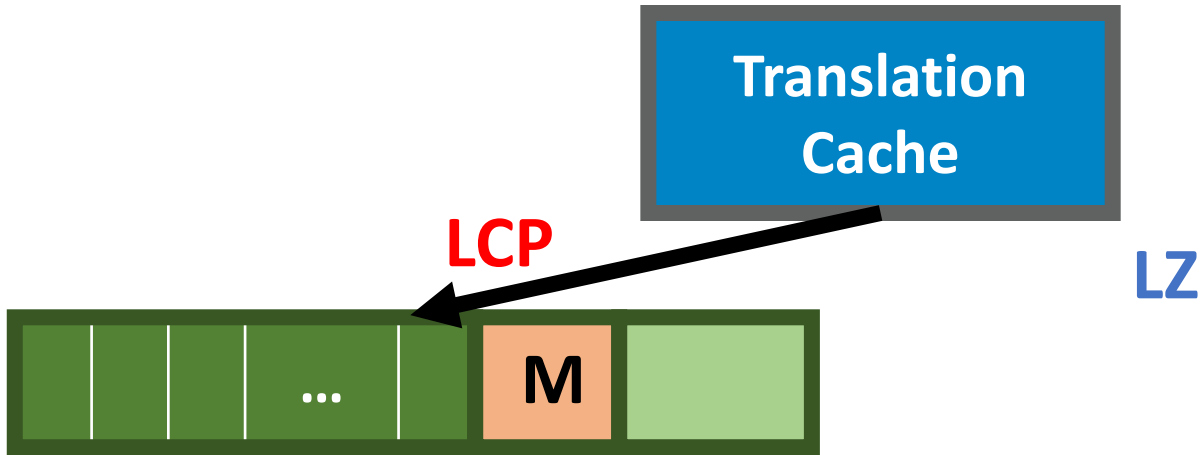
DMC: Accessing Cacheline



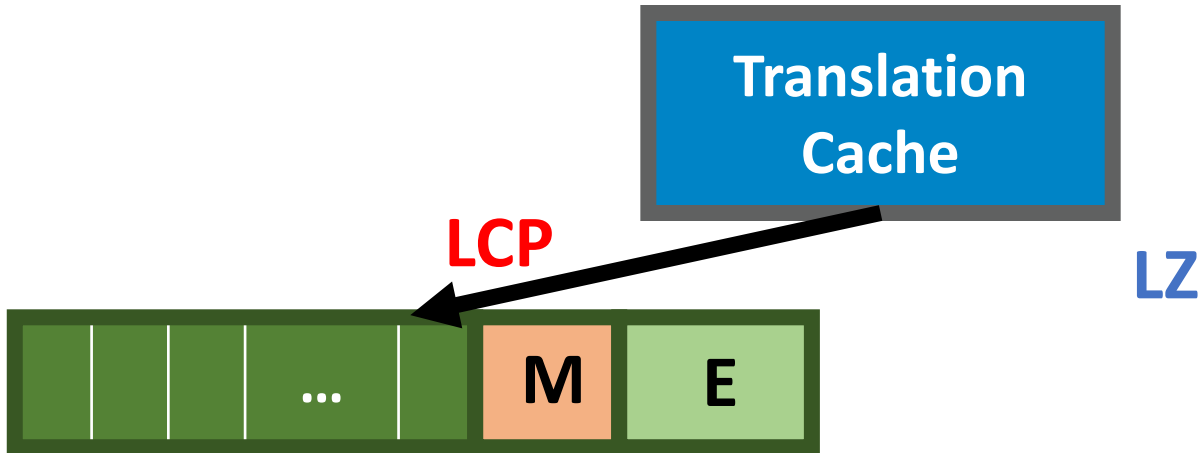
DMC: Accessing Cacheline



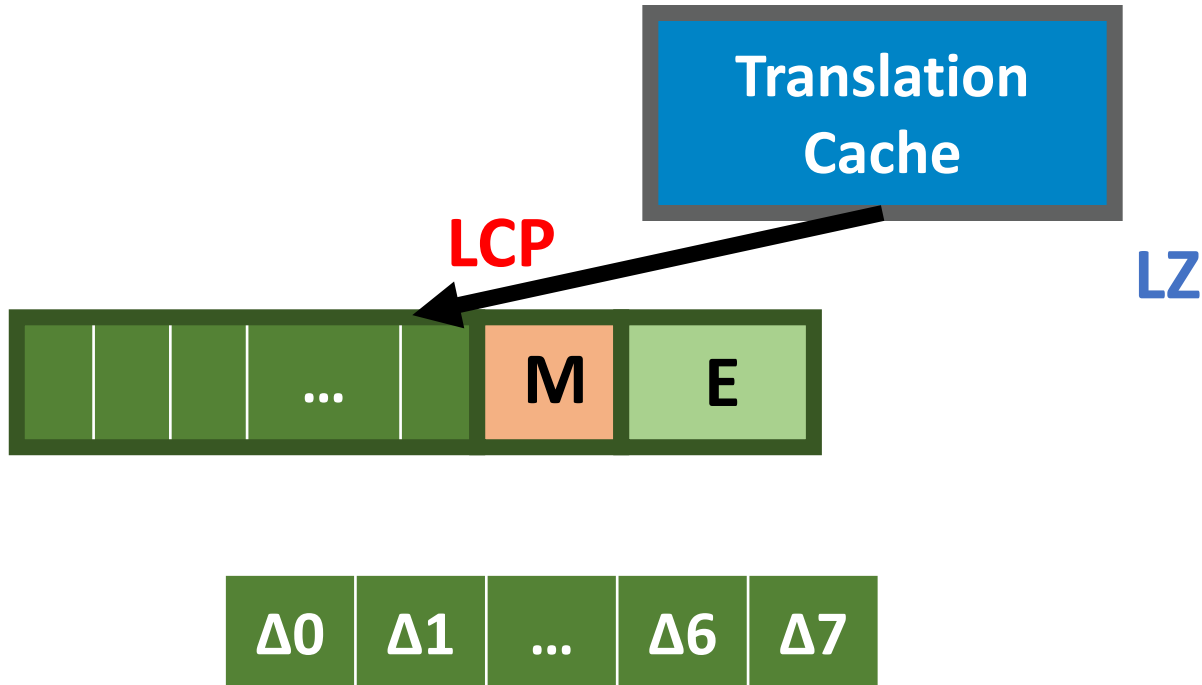
DMC: Accessing Cacheline



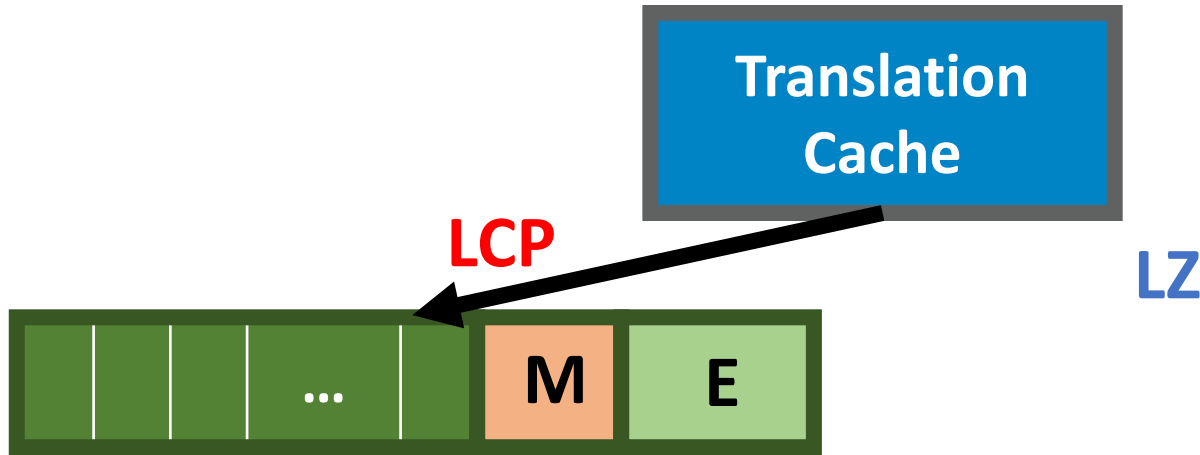
DMC: Accessing Cacheline



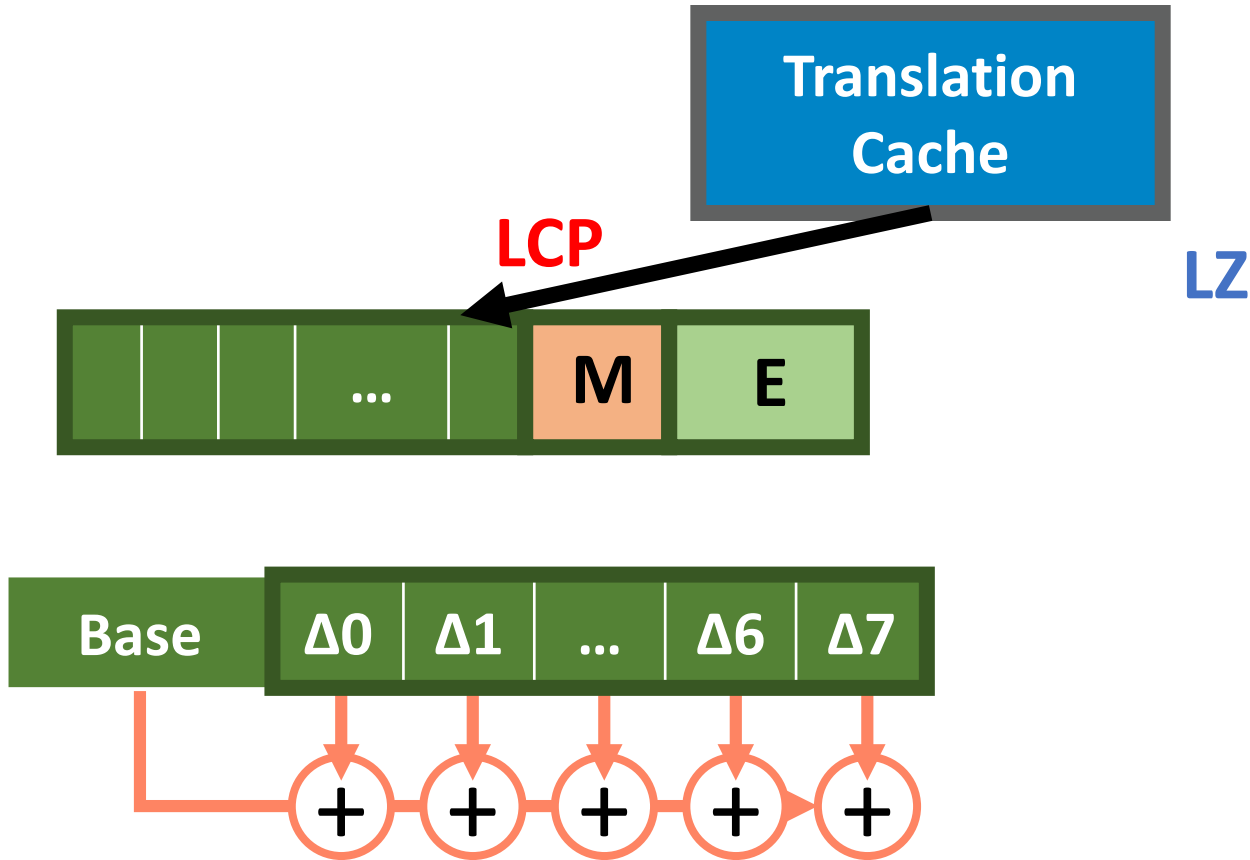
DMC: Accessing Cacheline



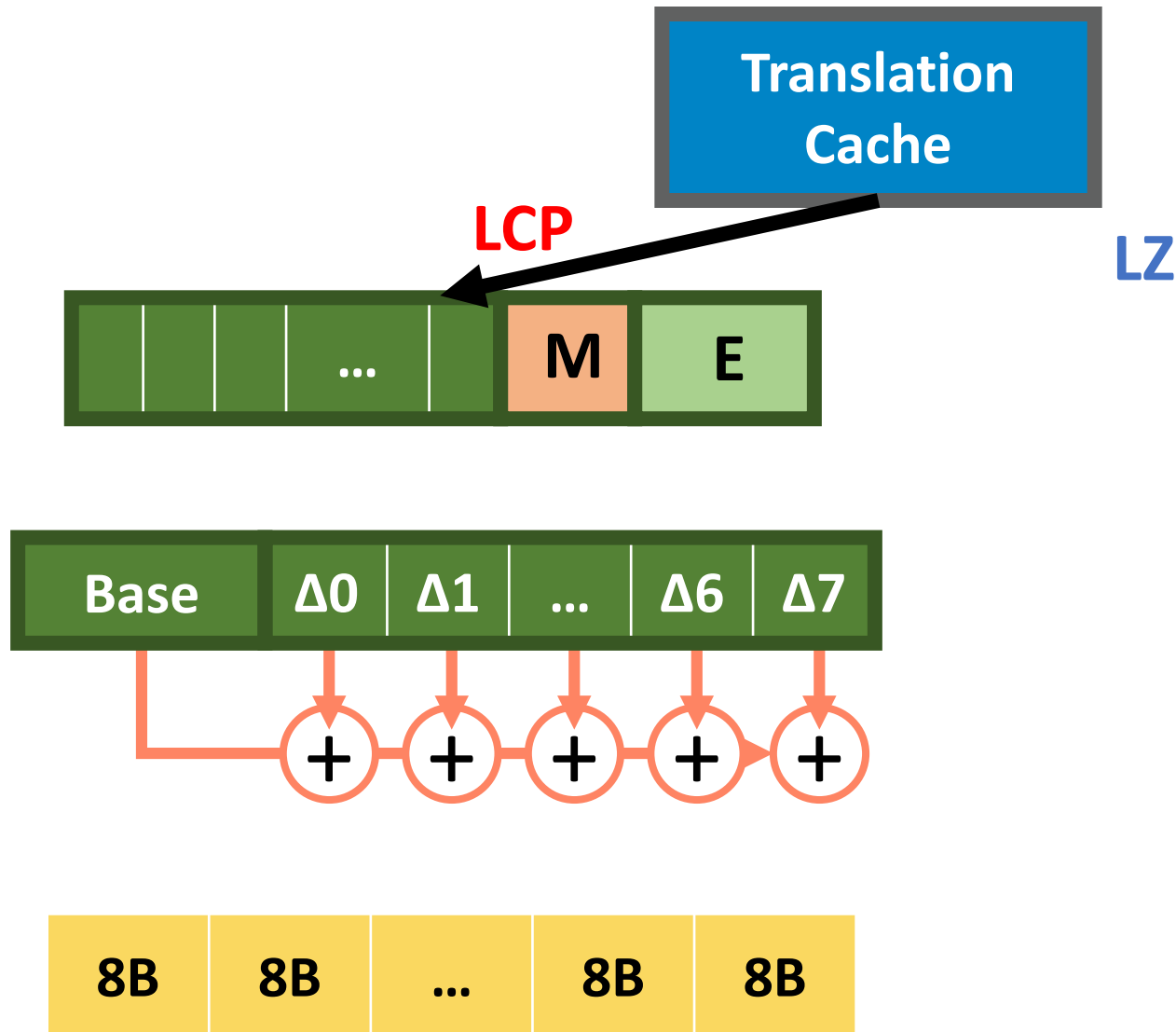
DMC: Accessing Cacheline



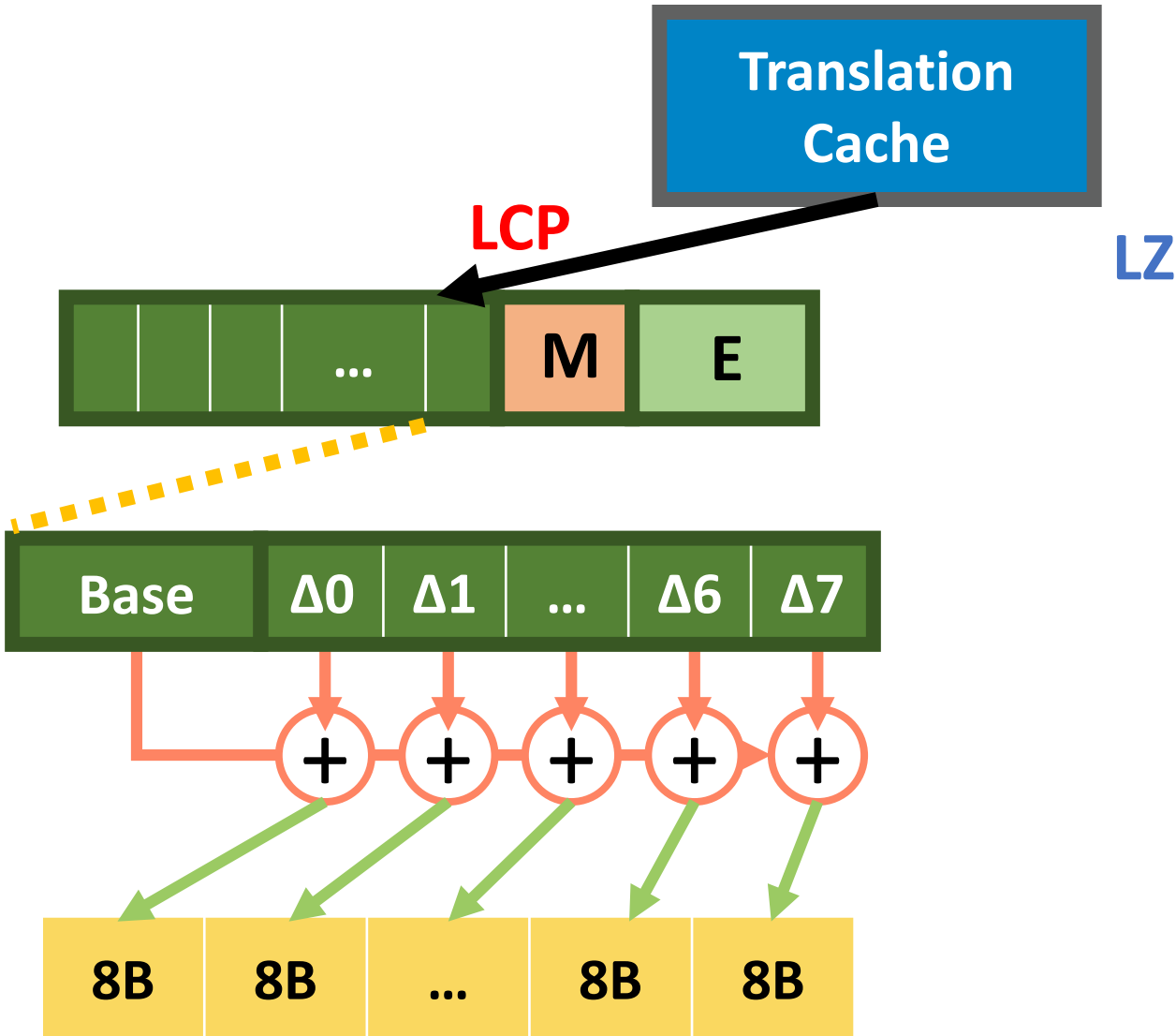
DMC: Accessing Cacheline



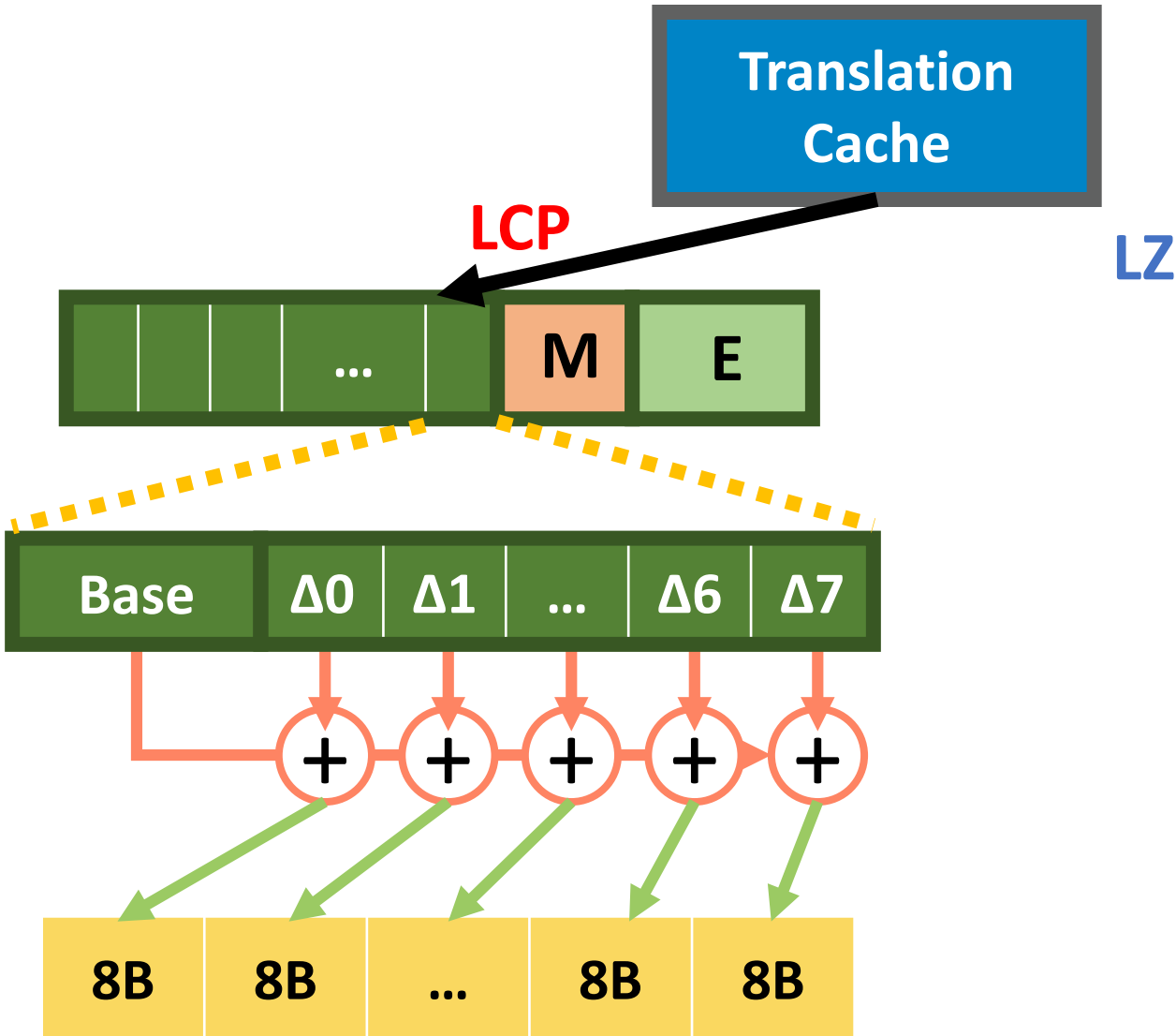
DMC: Accessing Cacheline



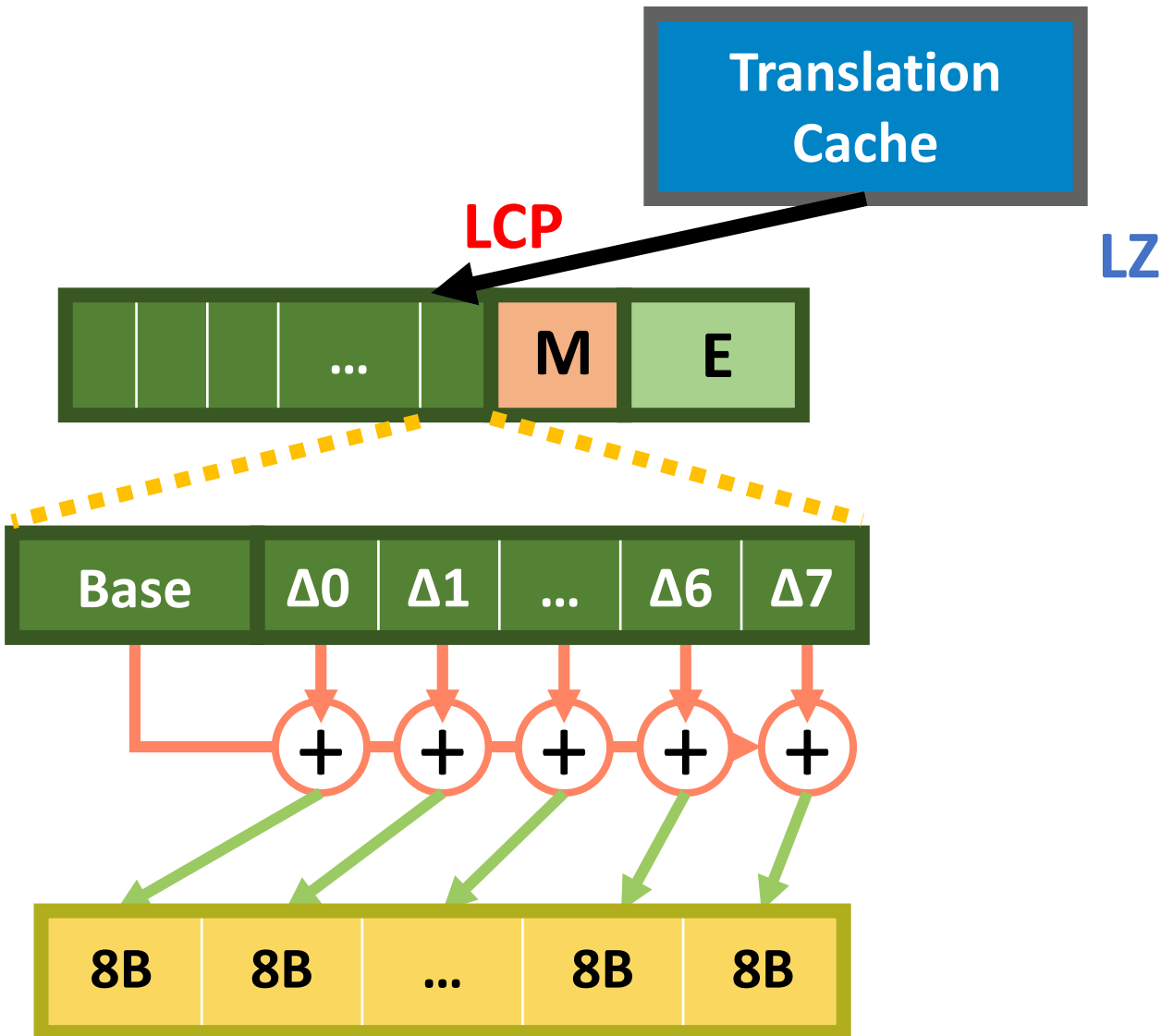
DMC: Accessing Cacheline



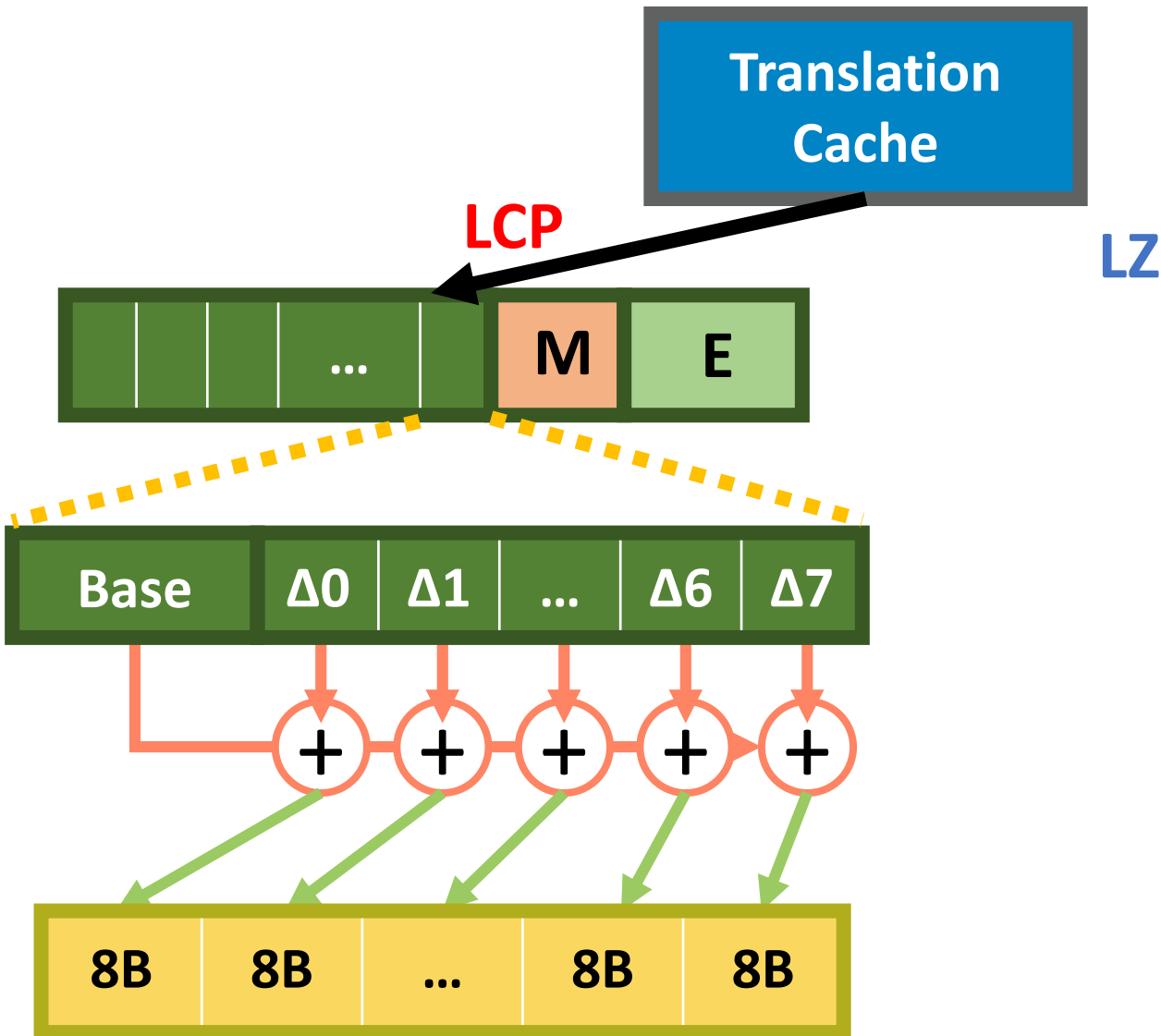
DMC: Accessing Cacheline



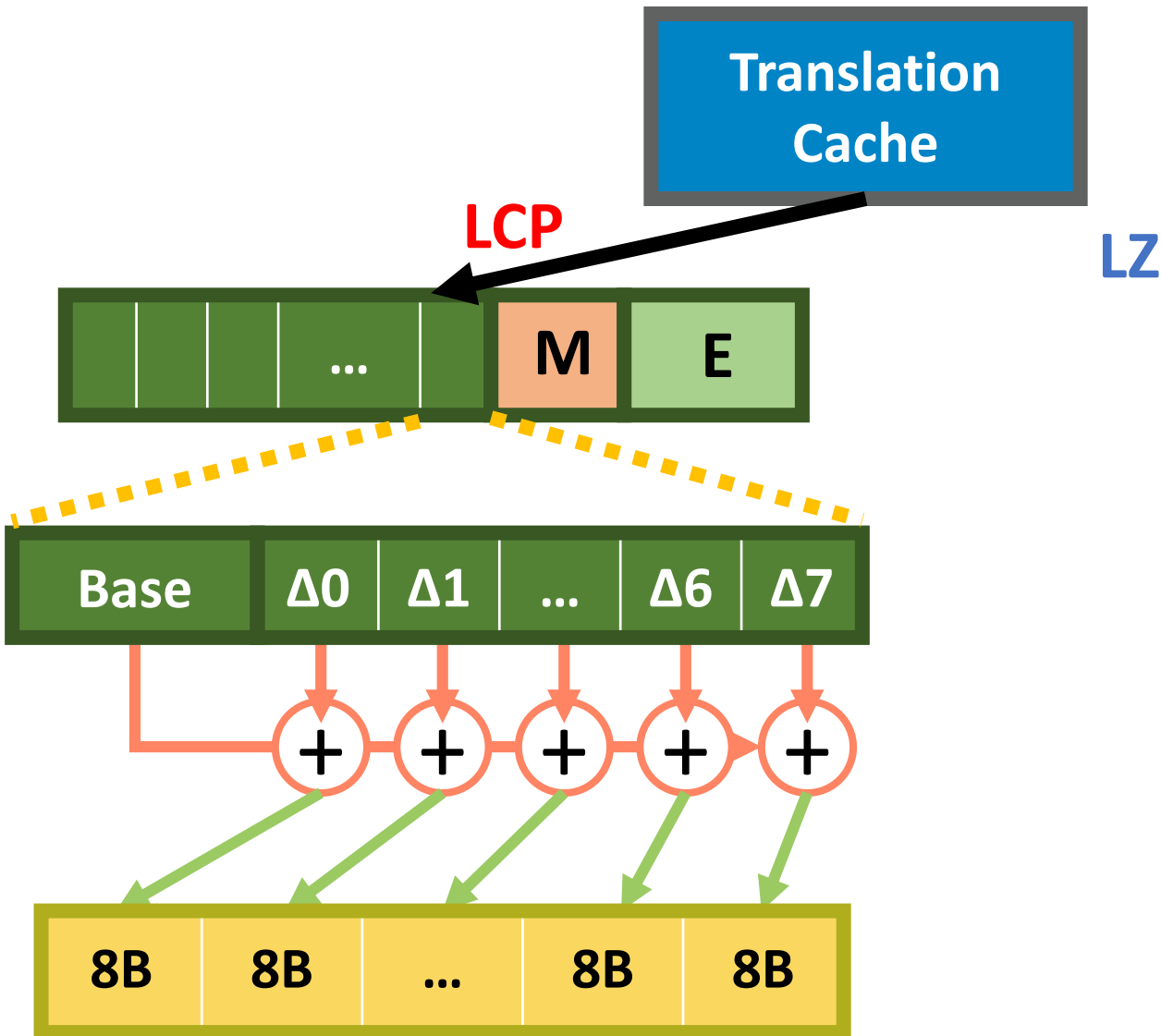
DMC: Accessing Cacheline



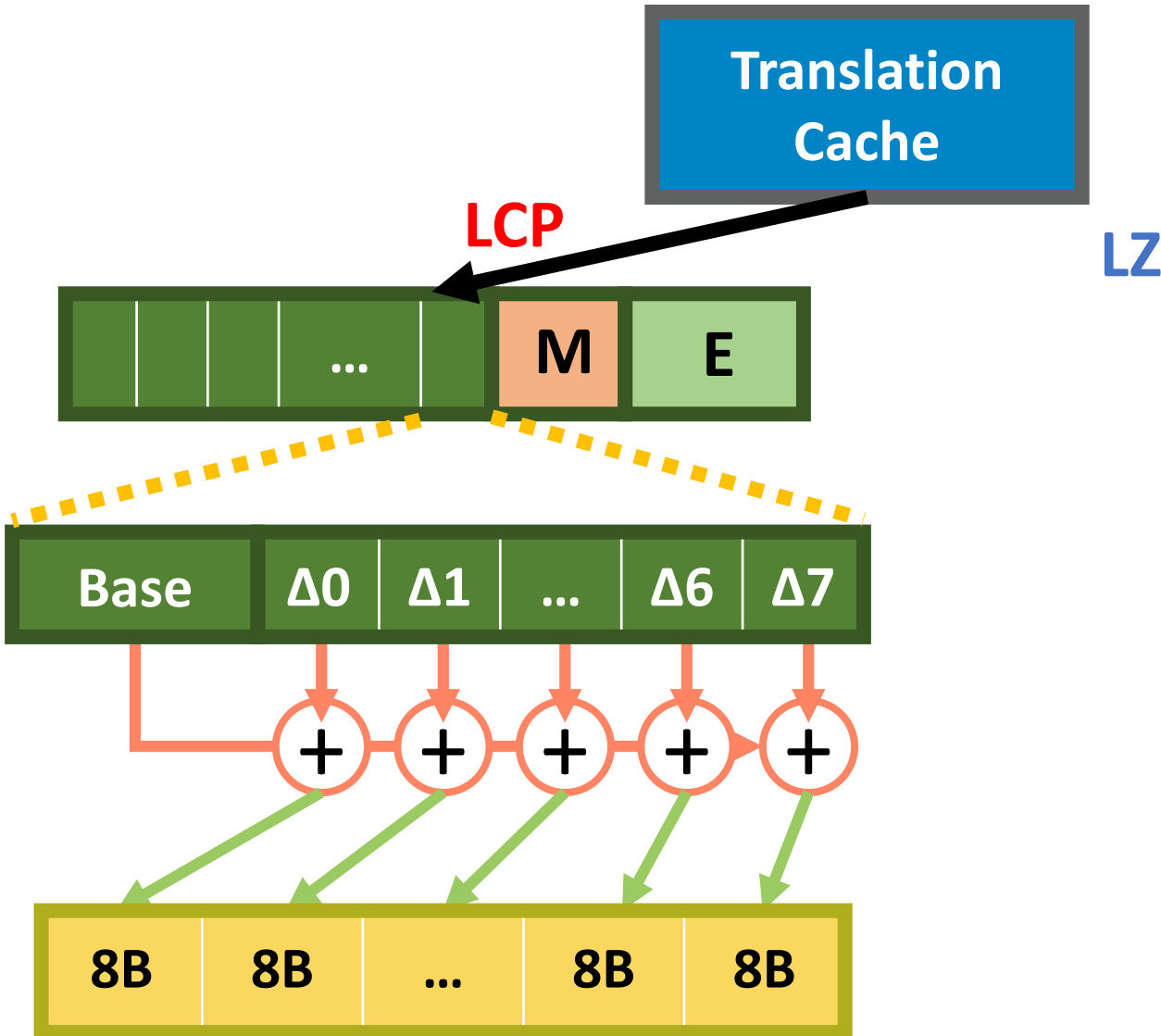
DMC: Accessing Cacheline



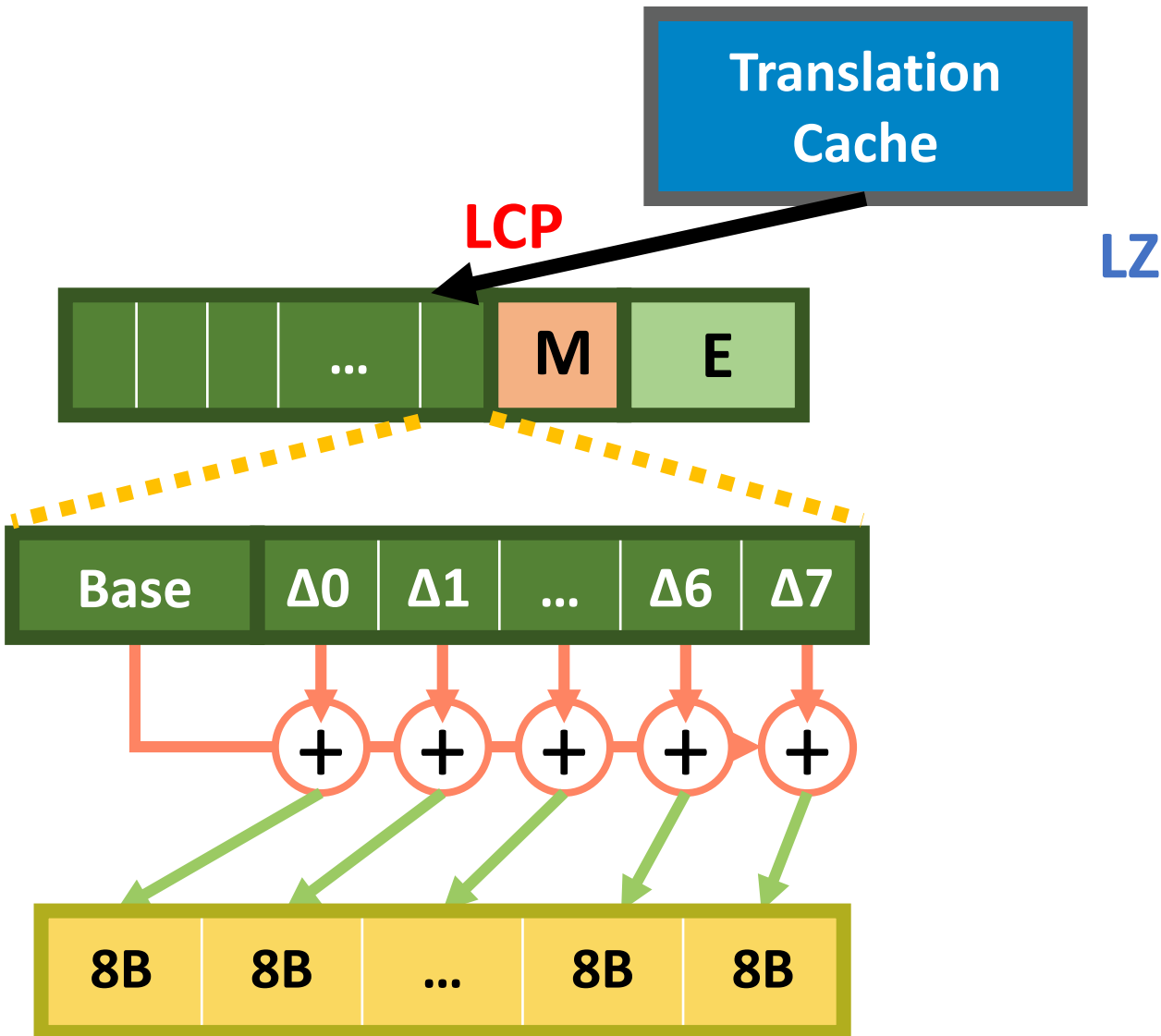
DMC: Accessing Cacheline



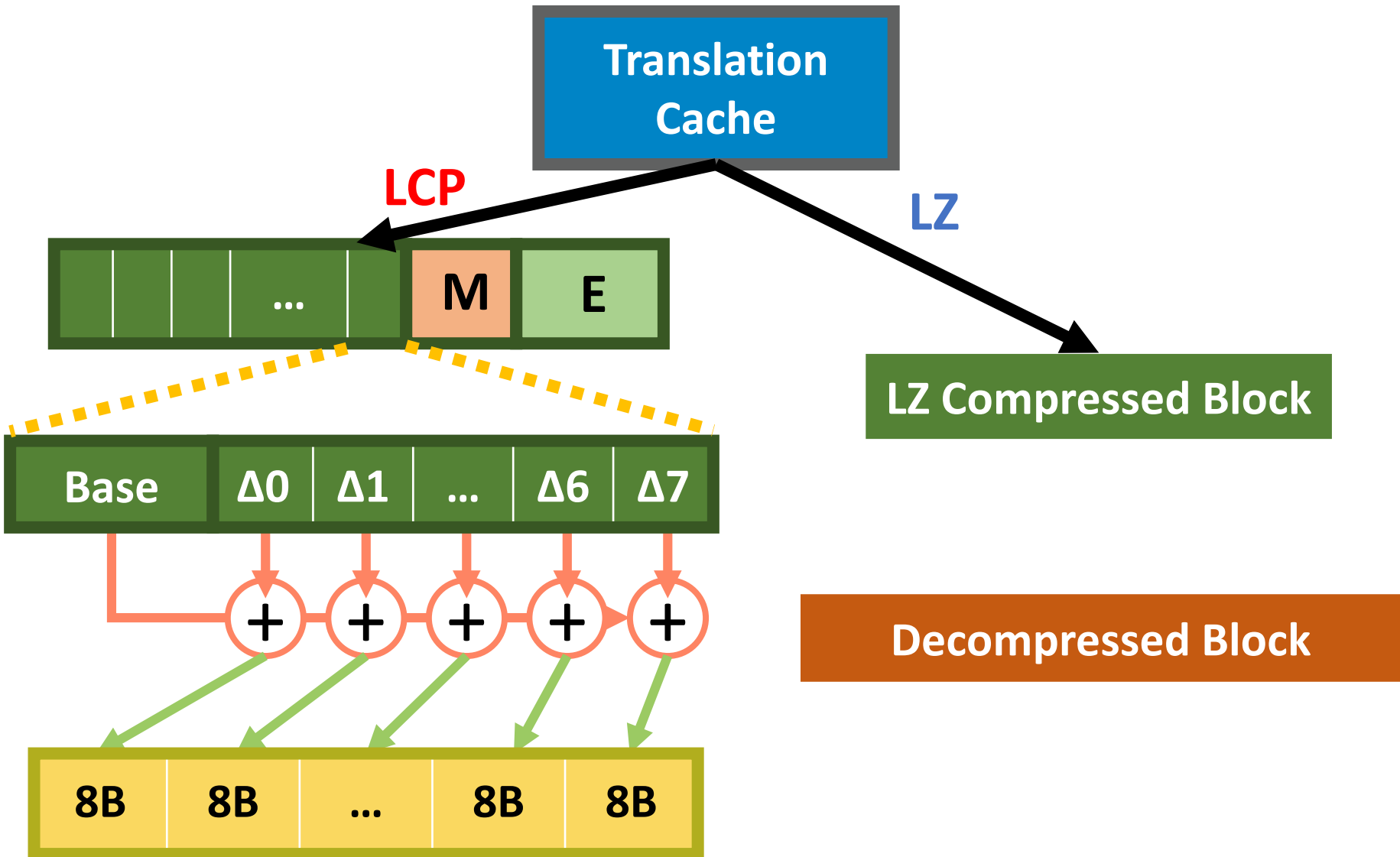
DMC: Accessing Cacheline



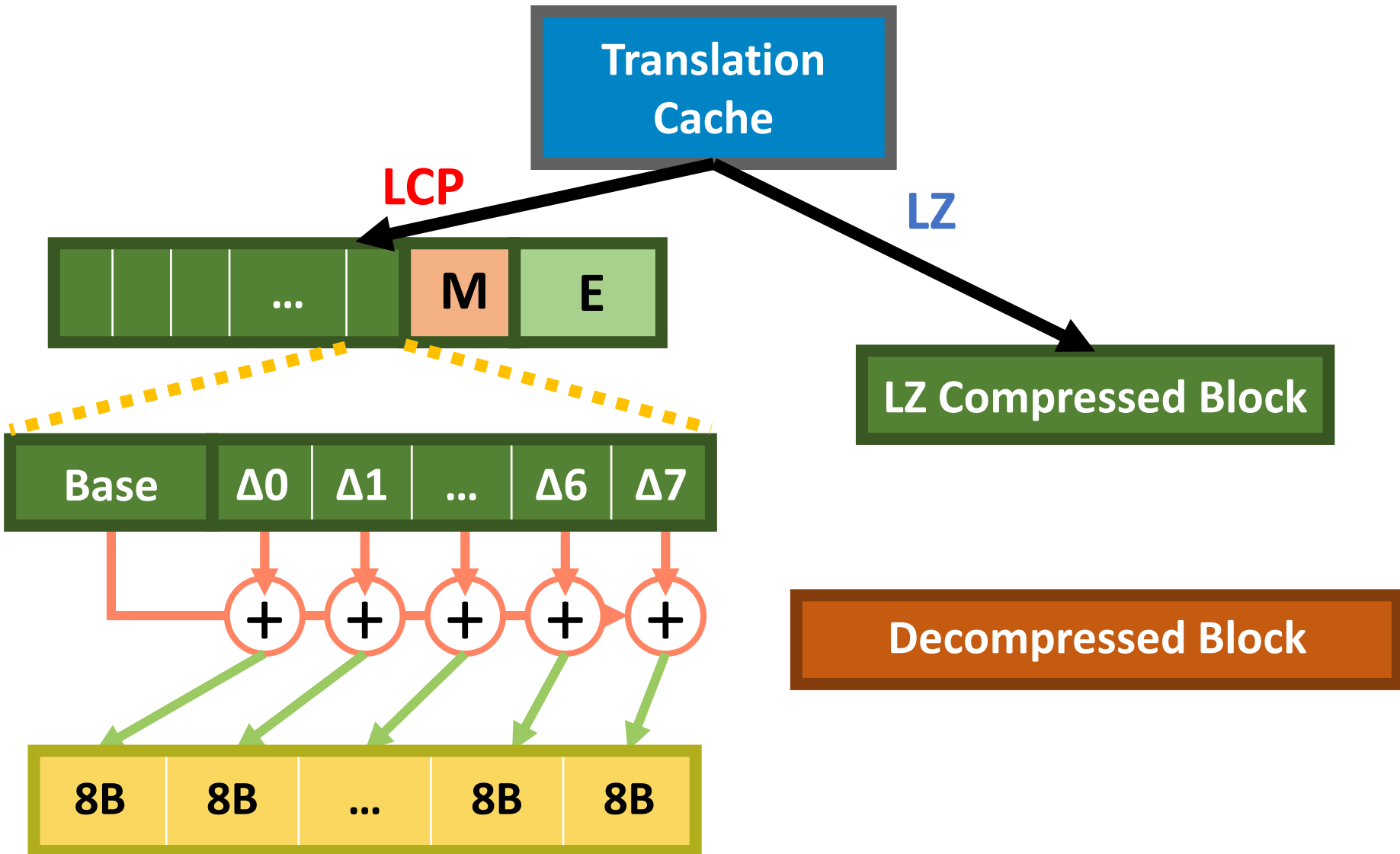
DMC: Accessing Cacheline



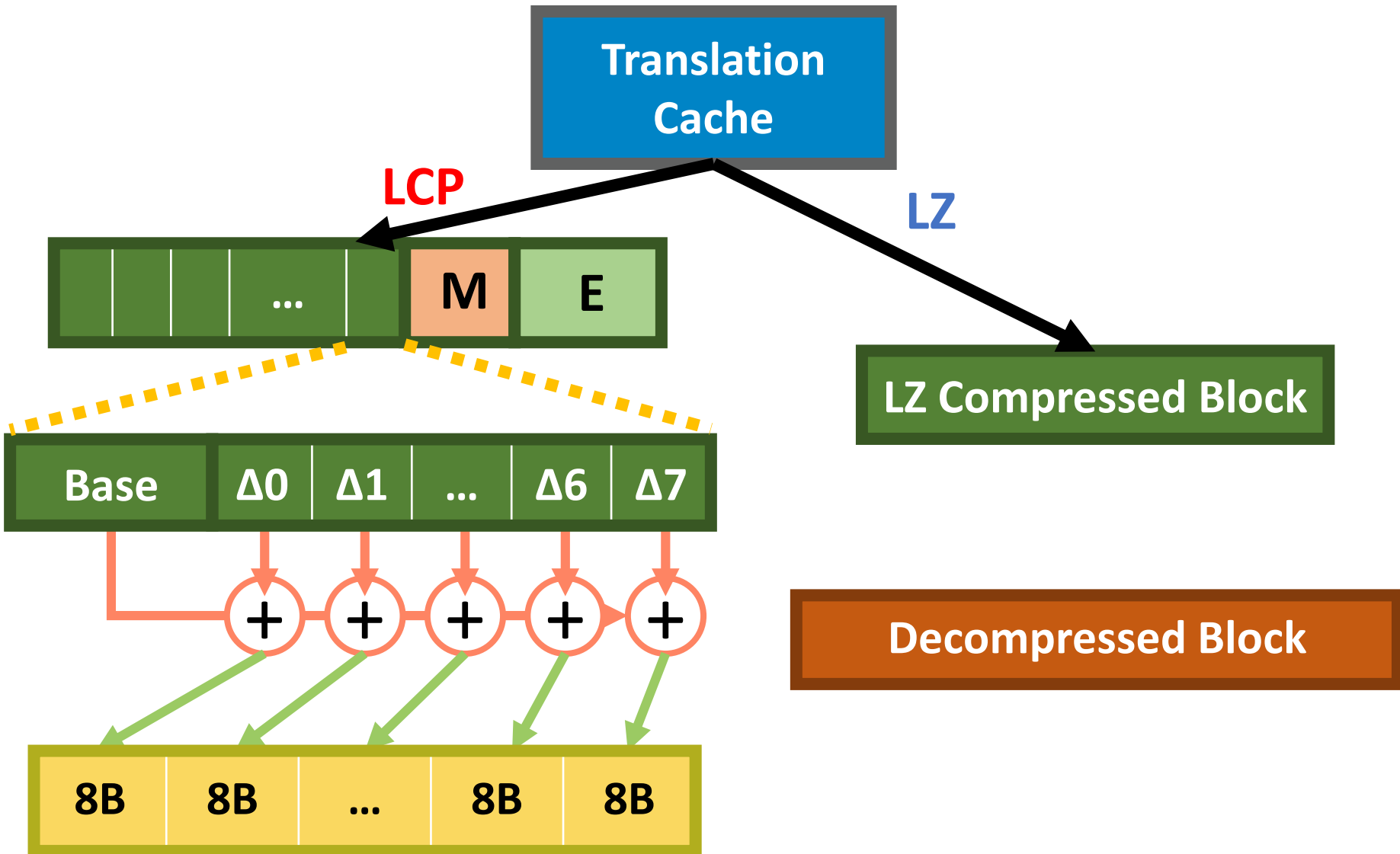
DMC: Accessing Cacheline



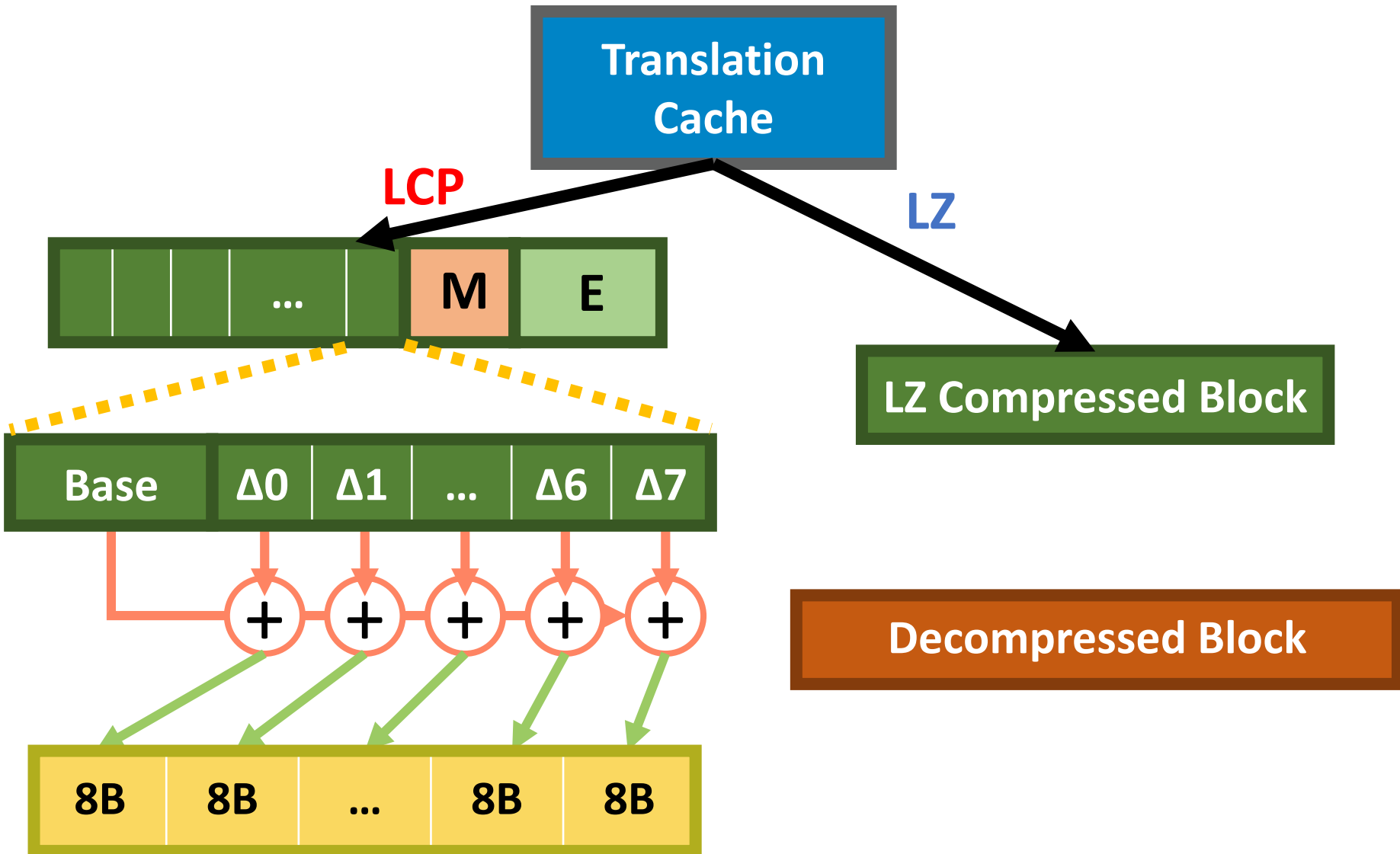
DMC: Accessing Cacheline



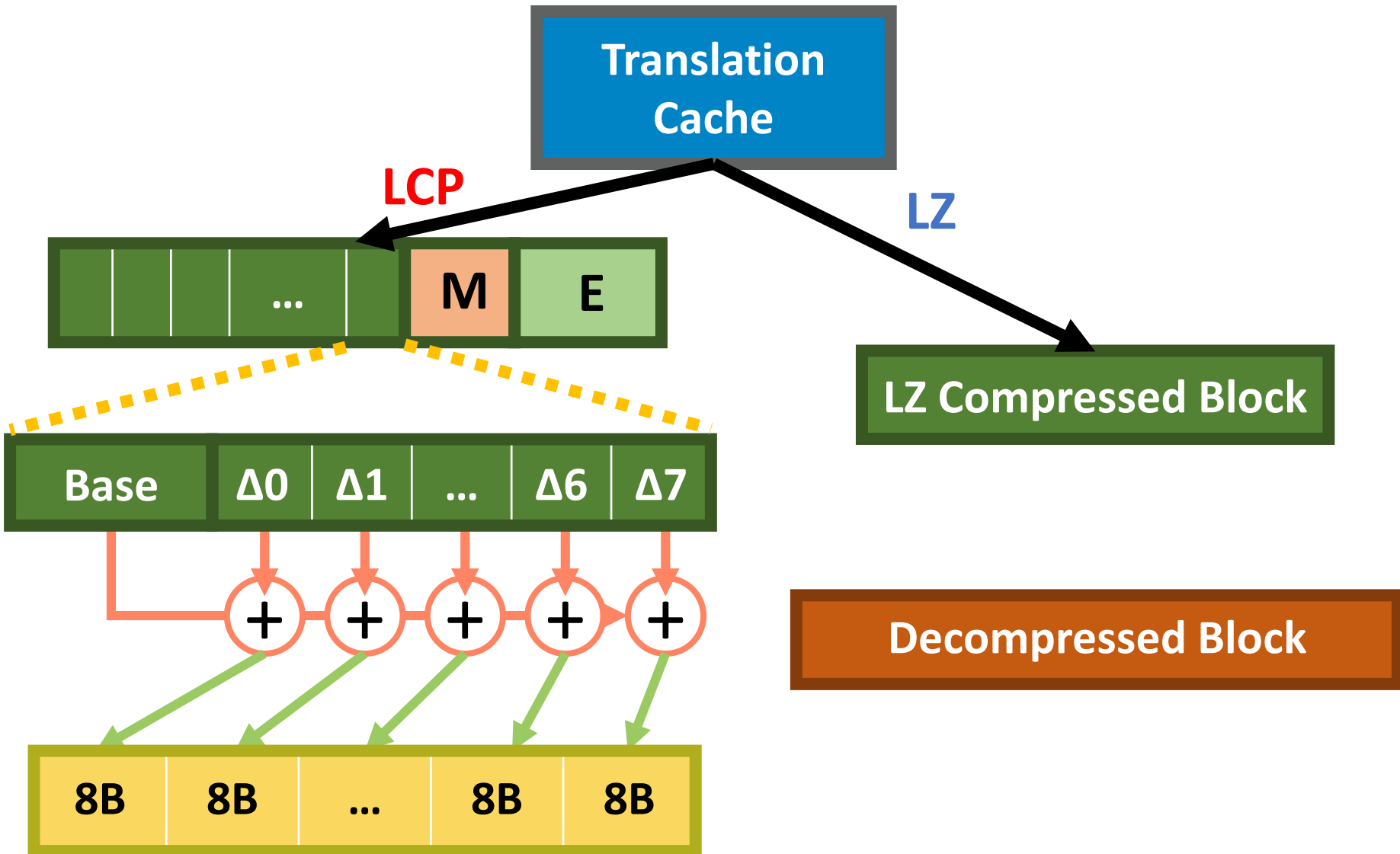
DMC: Accessing Cacheline



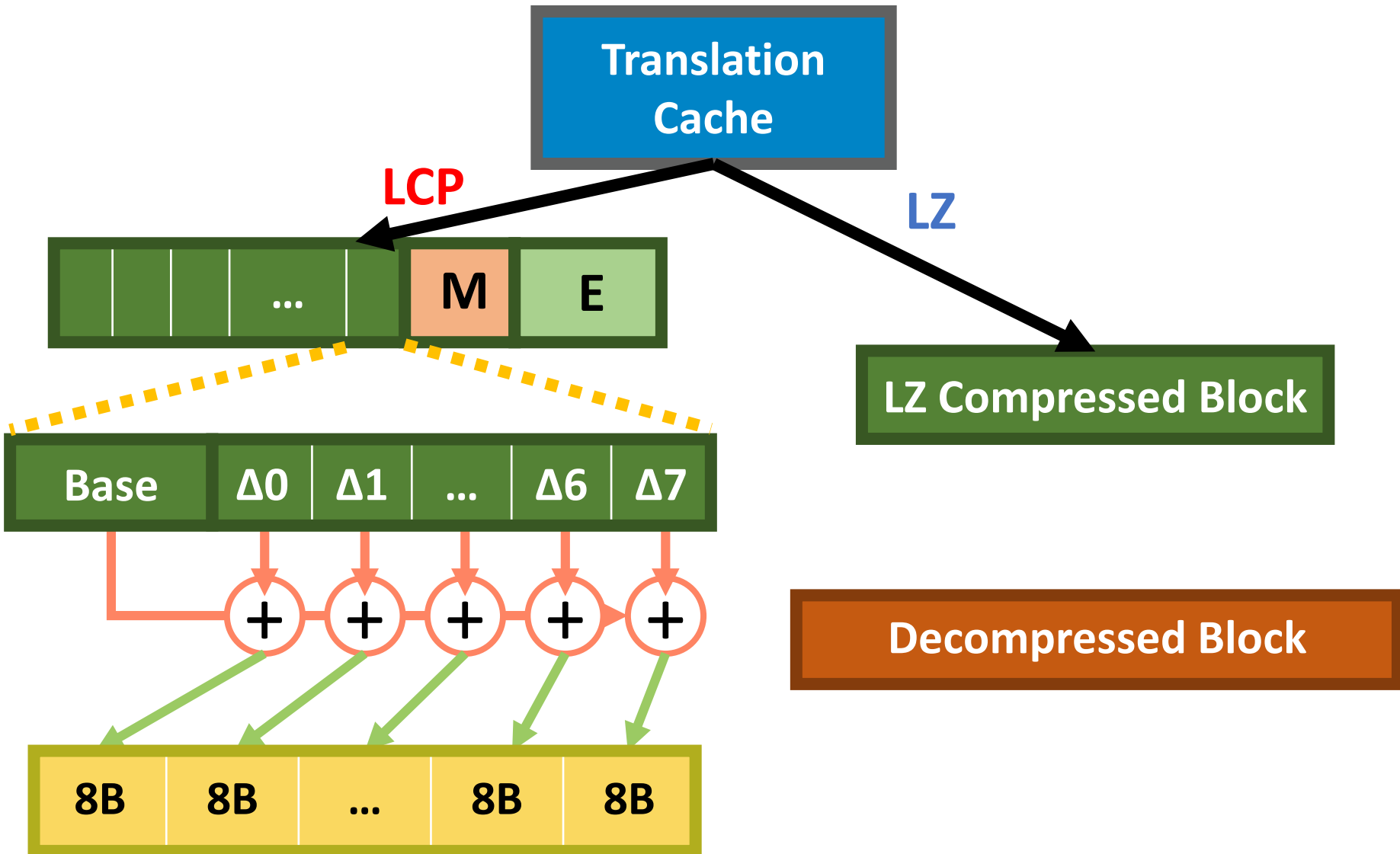
DMC: Accessing Cacheline



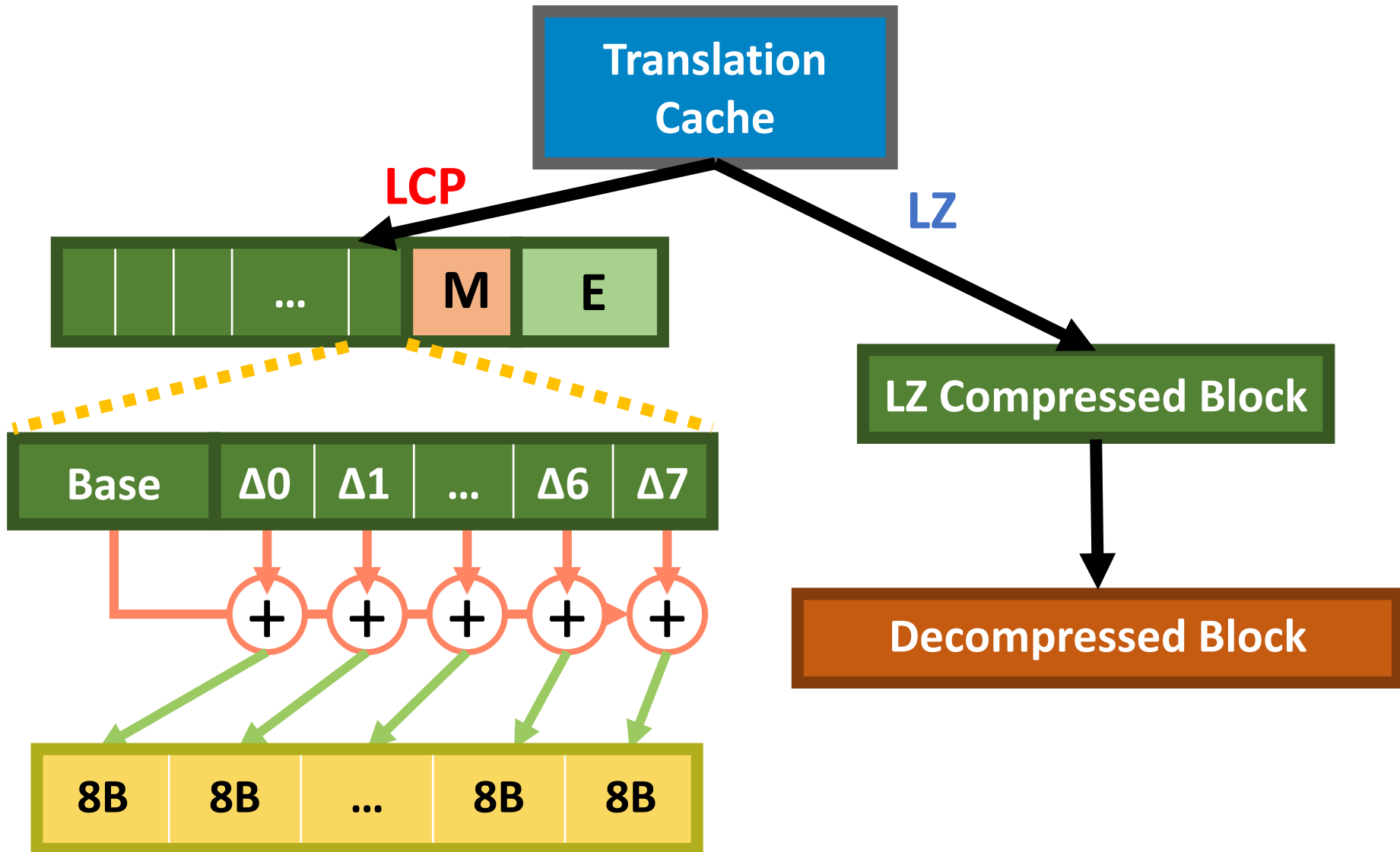
DMC: Accessing Cacheline



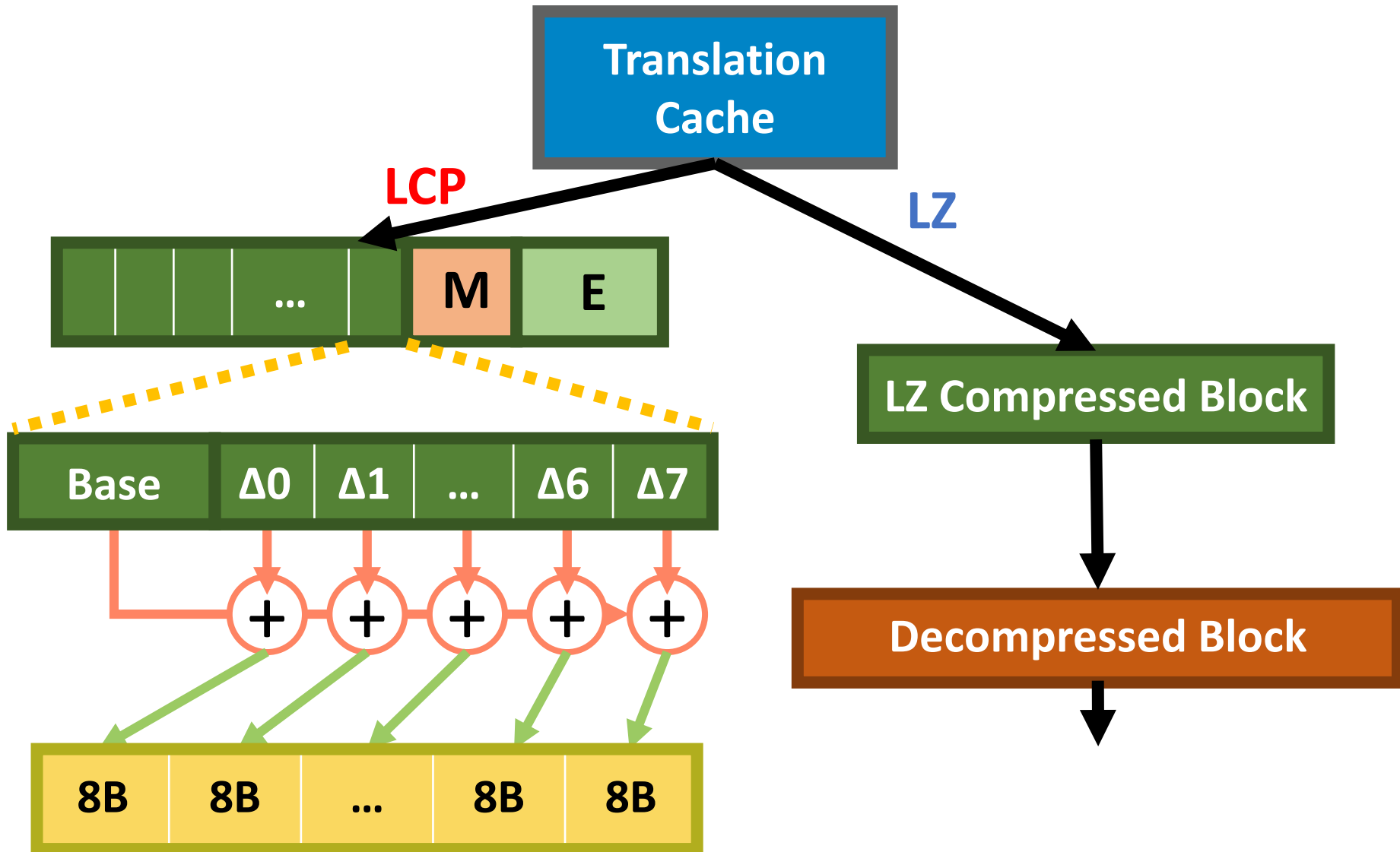
DMC: Accessing Cacheline



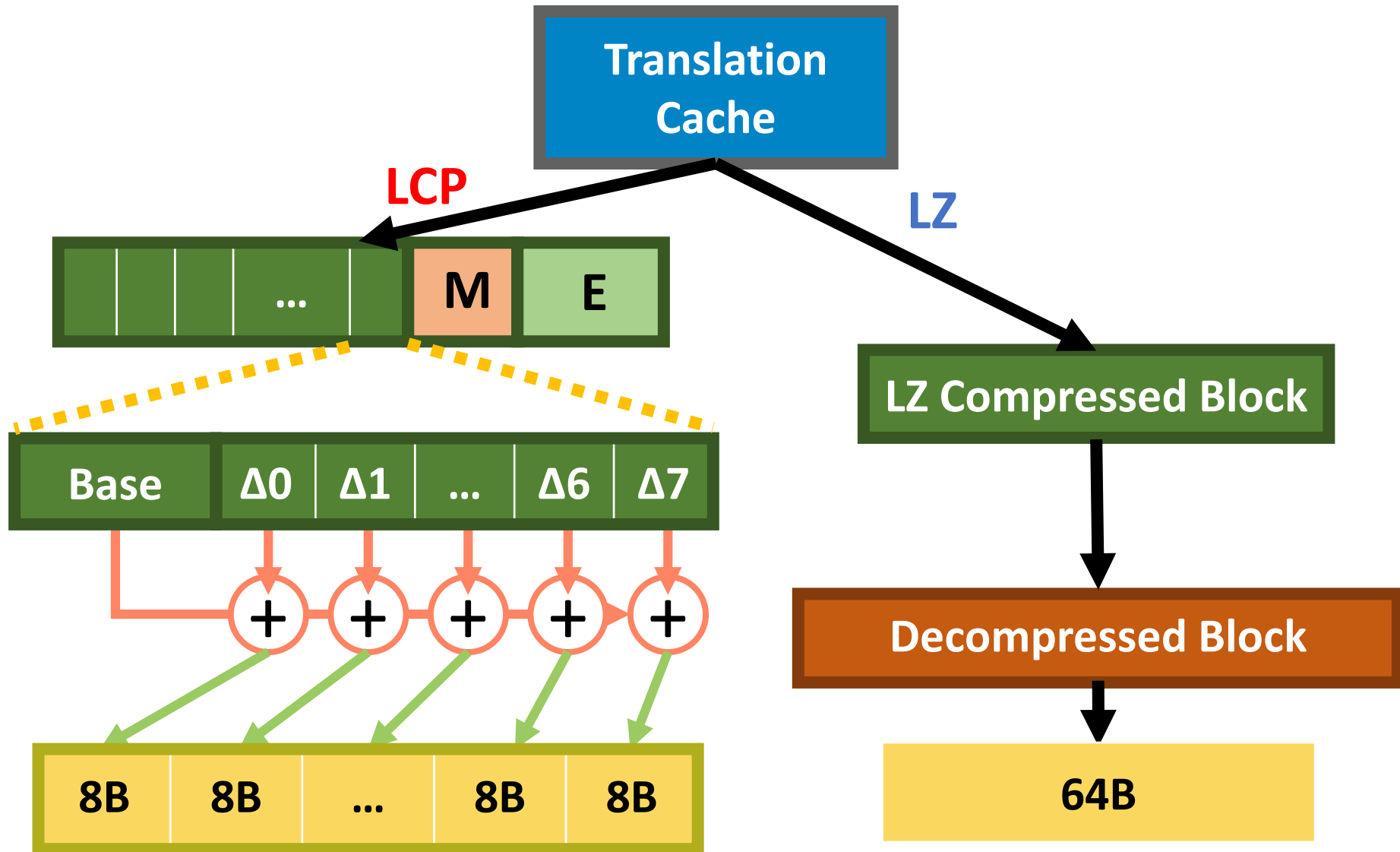
DMC: Accessing Cacheline



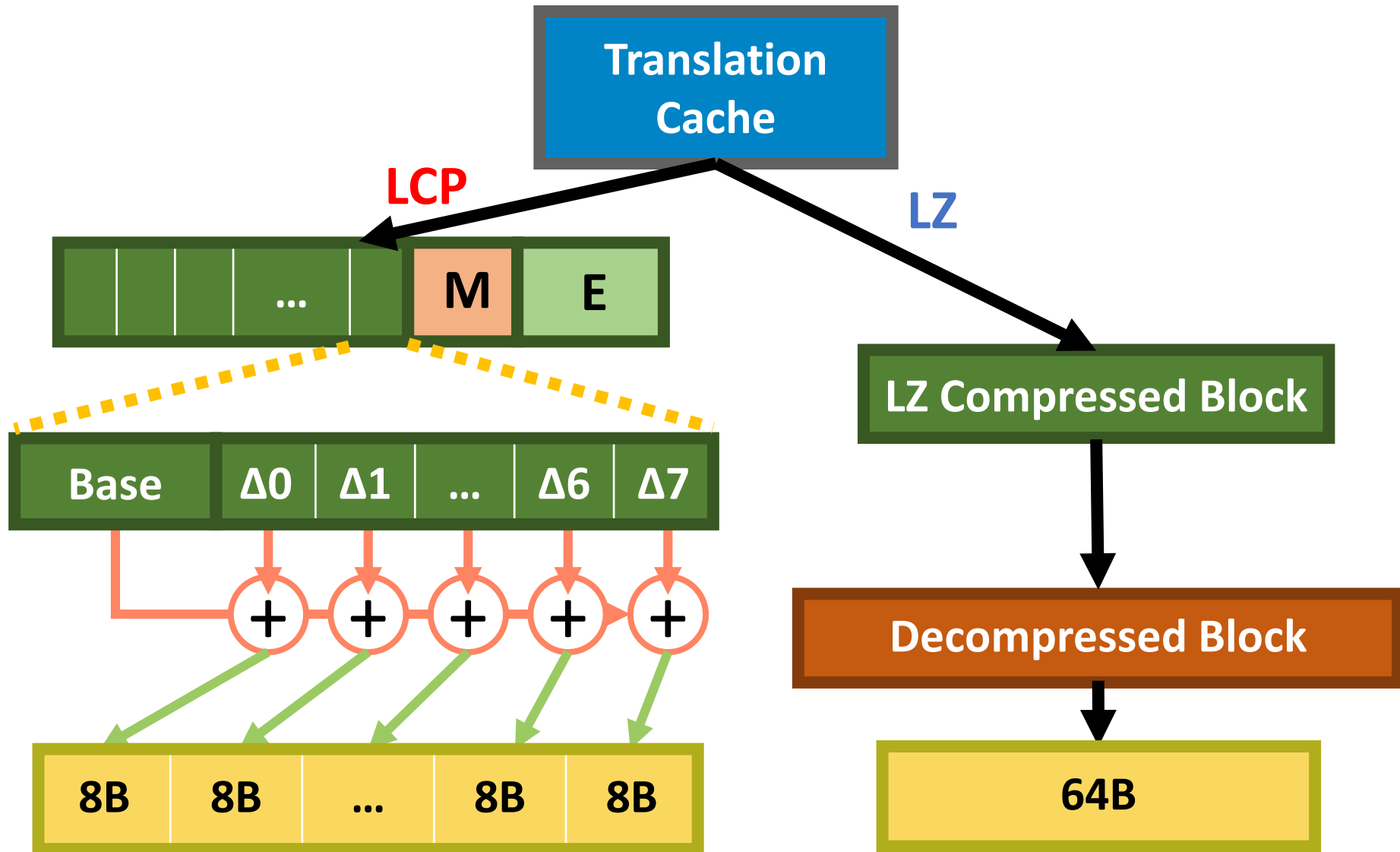
DMC: Accessing Cacheline



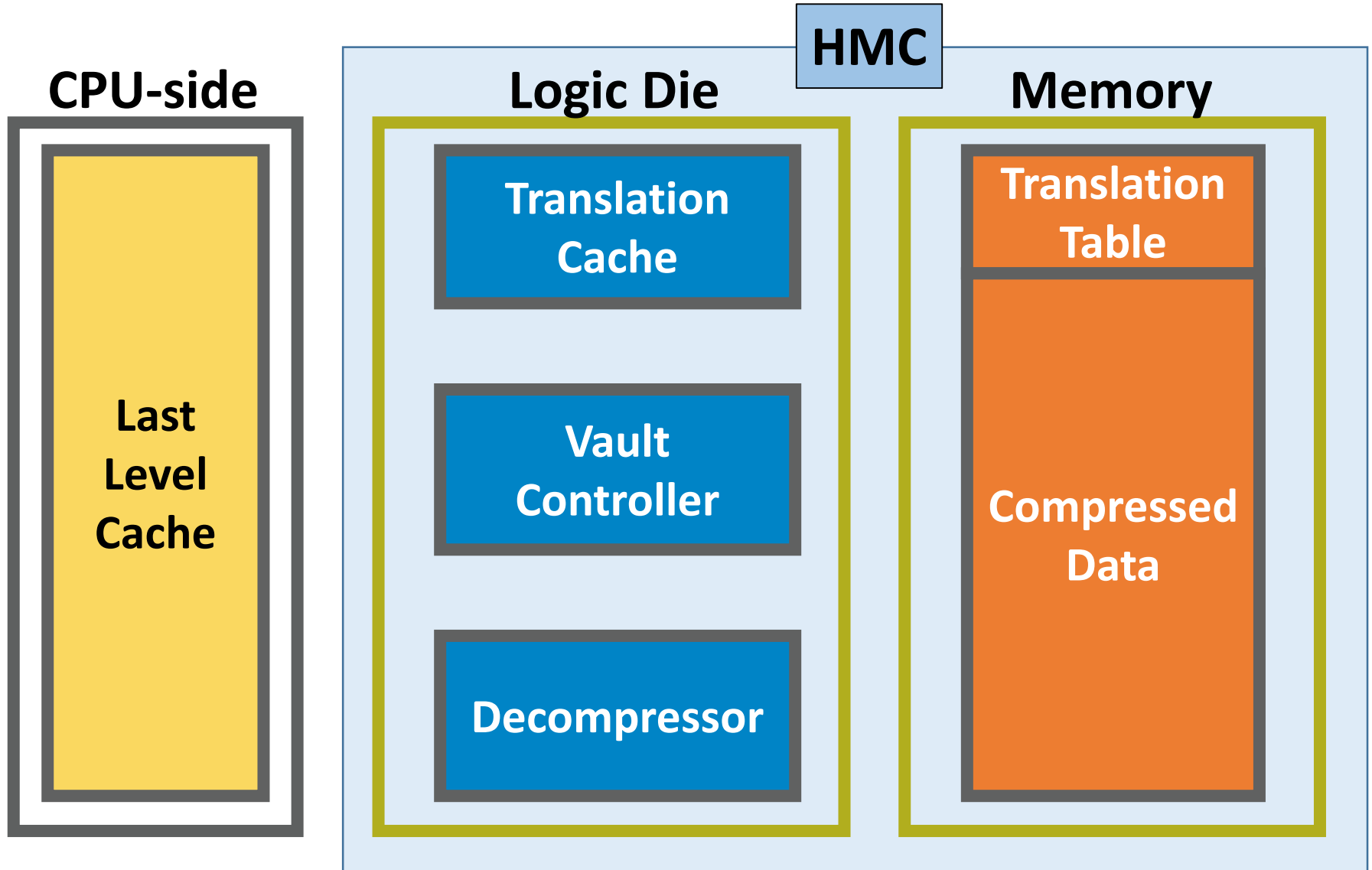
DMC: Accessing Cacheline



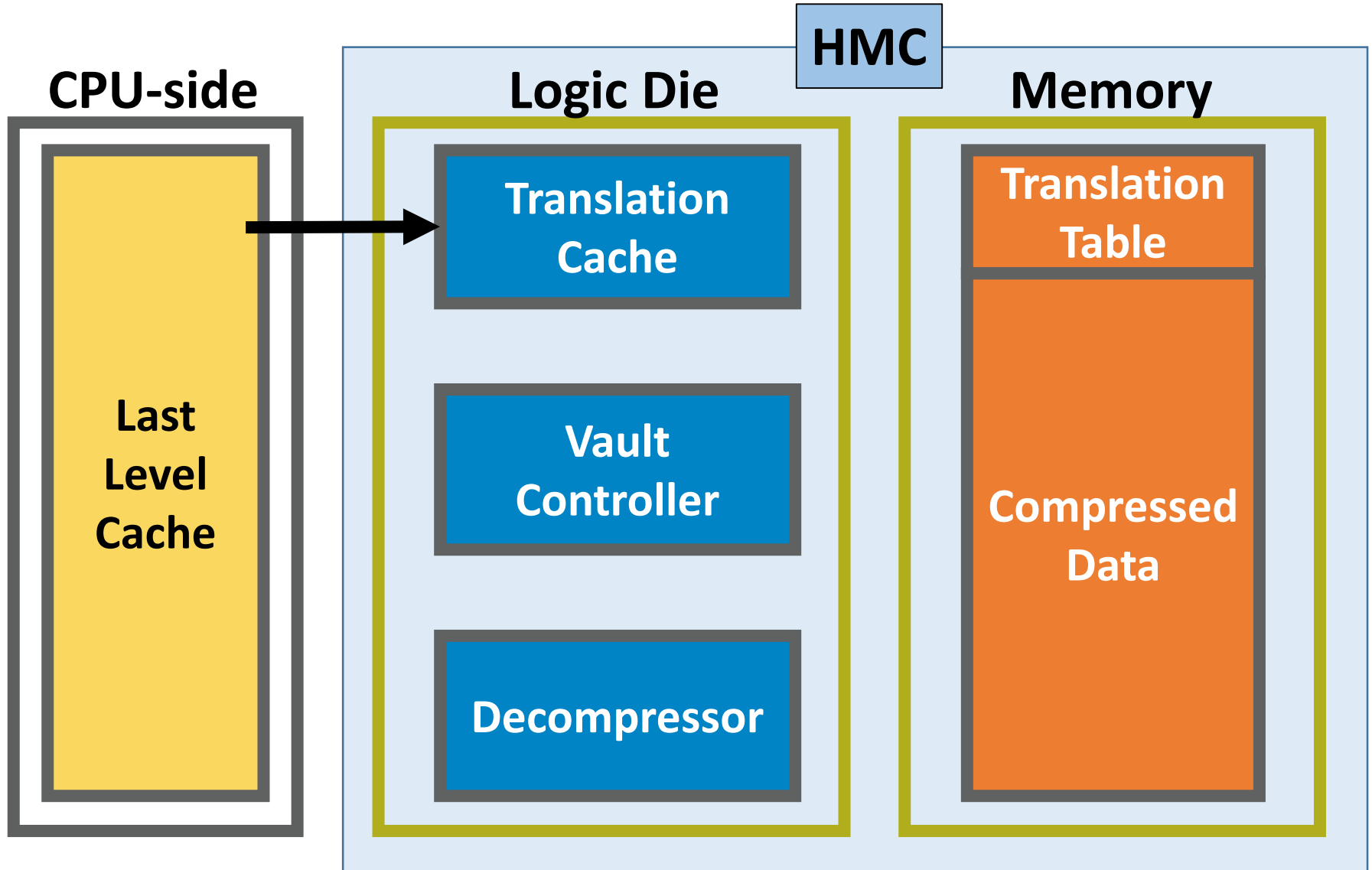
DMC: Accessing Cacheline



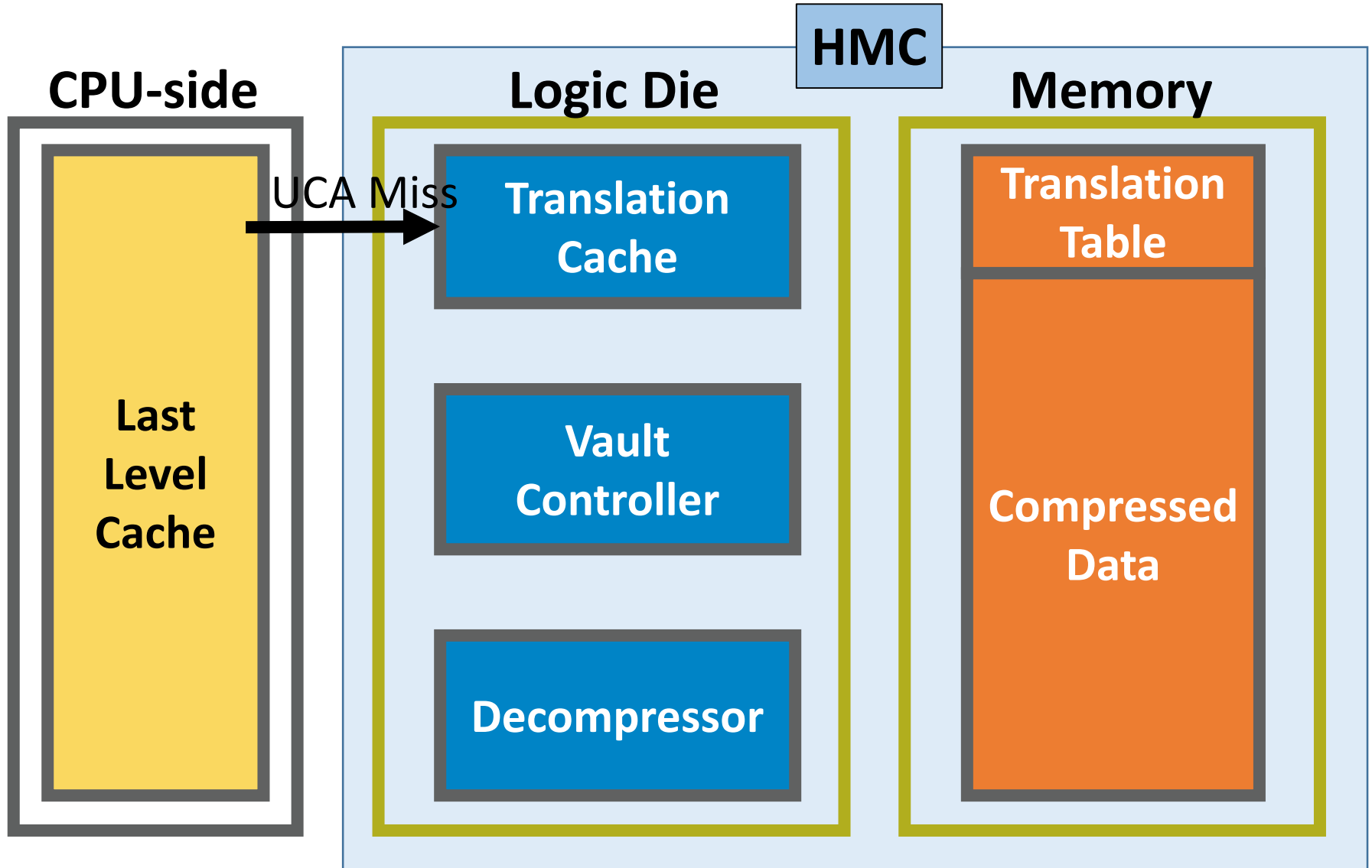
DMC: Components



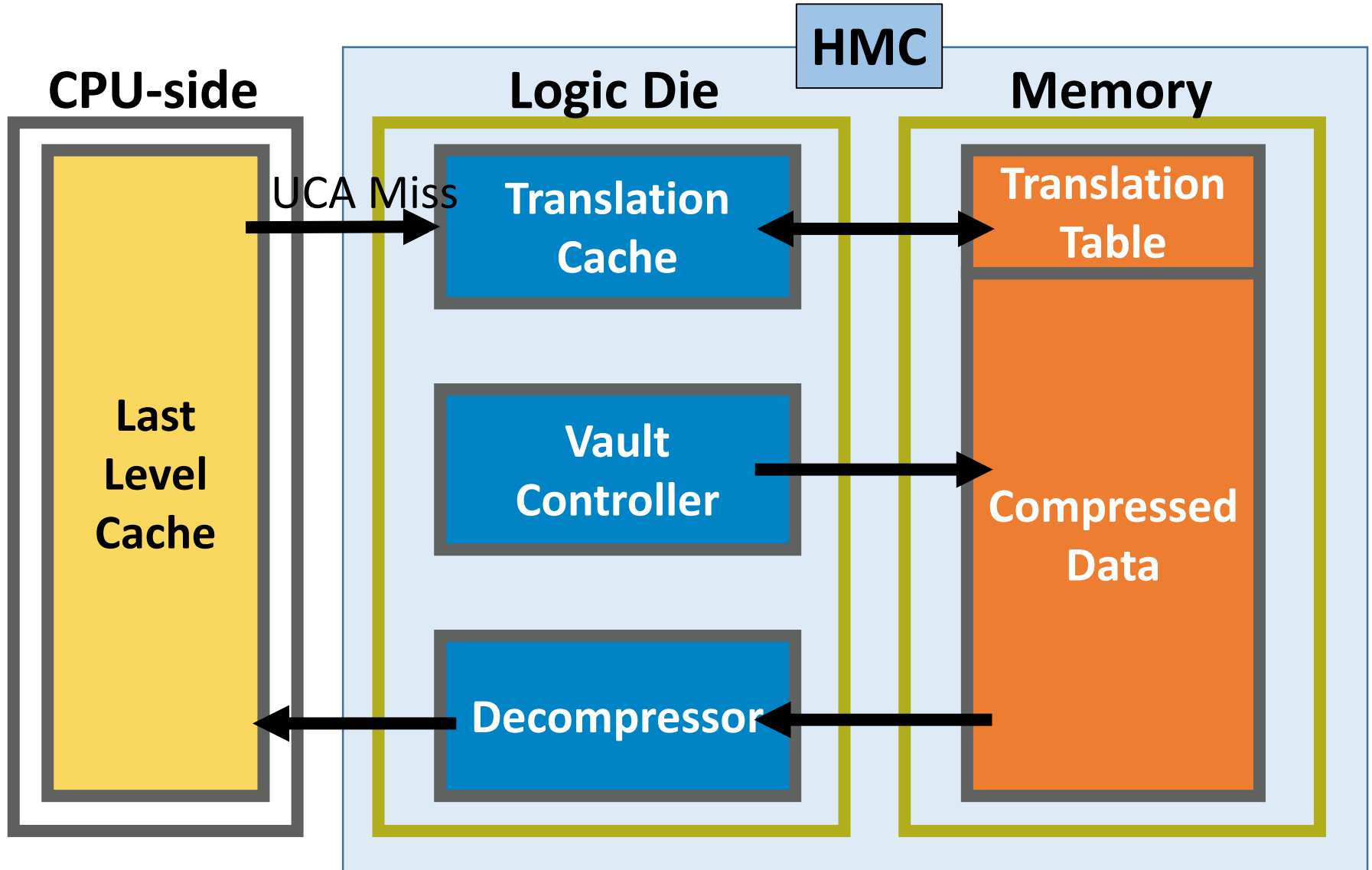
DMC: Components



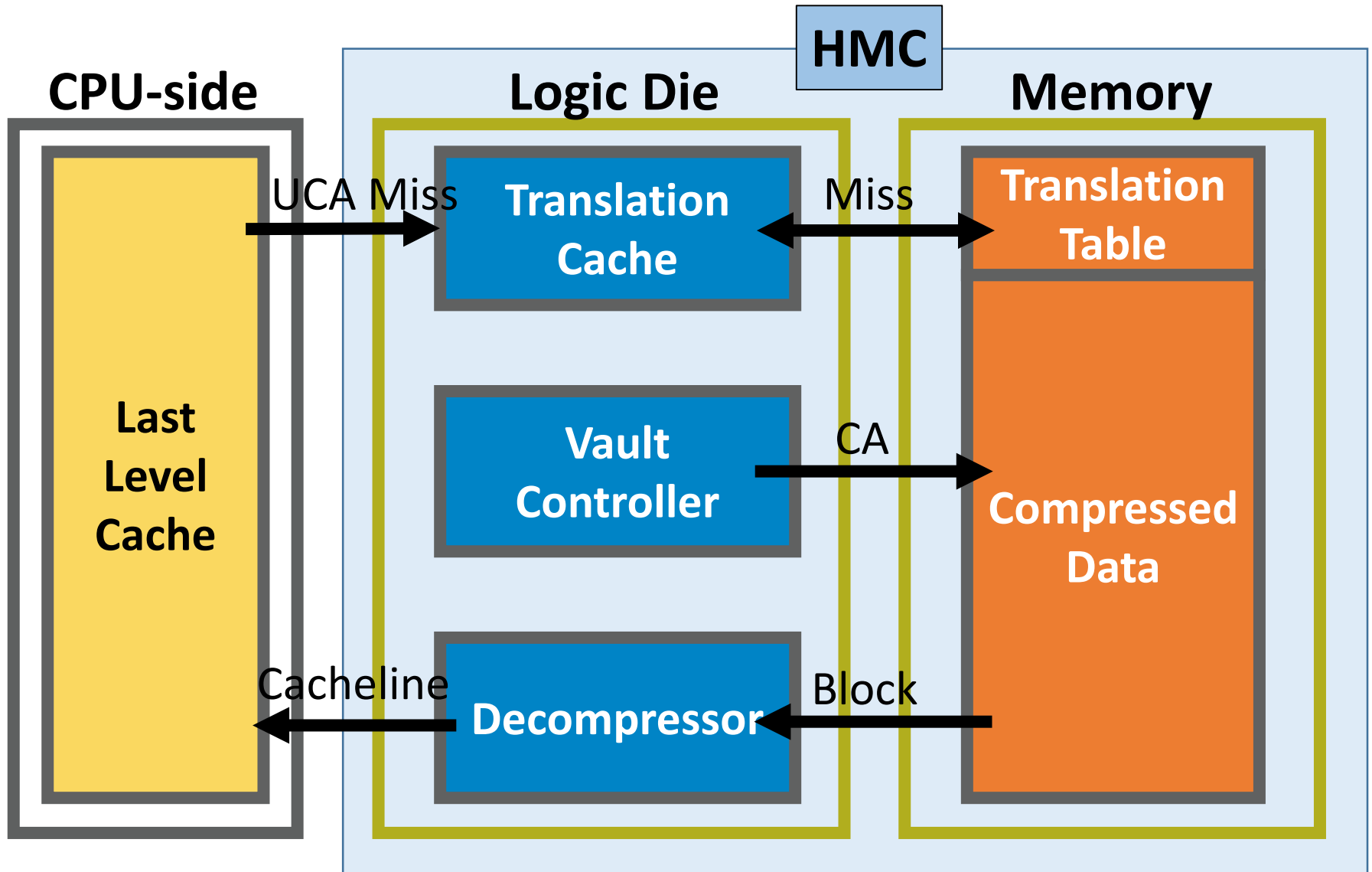
DMC: Components



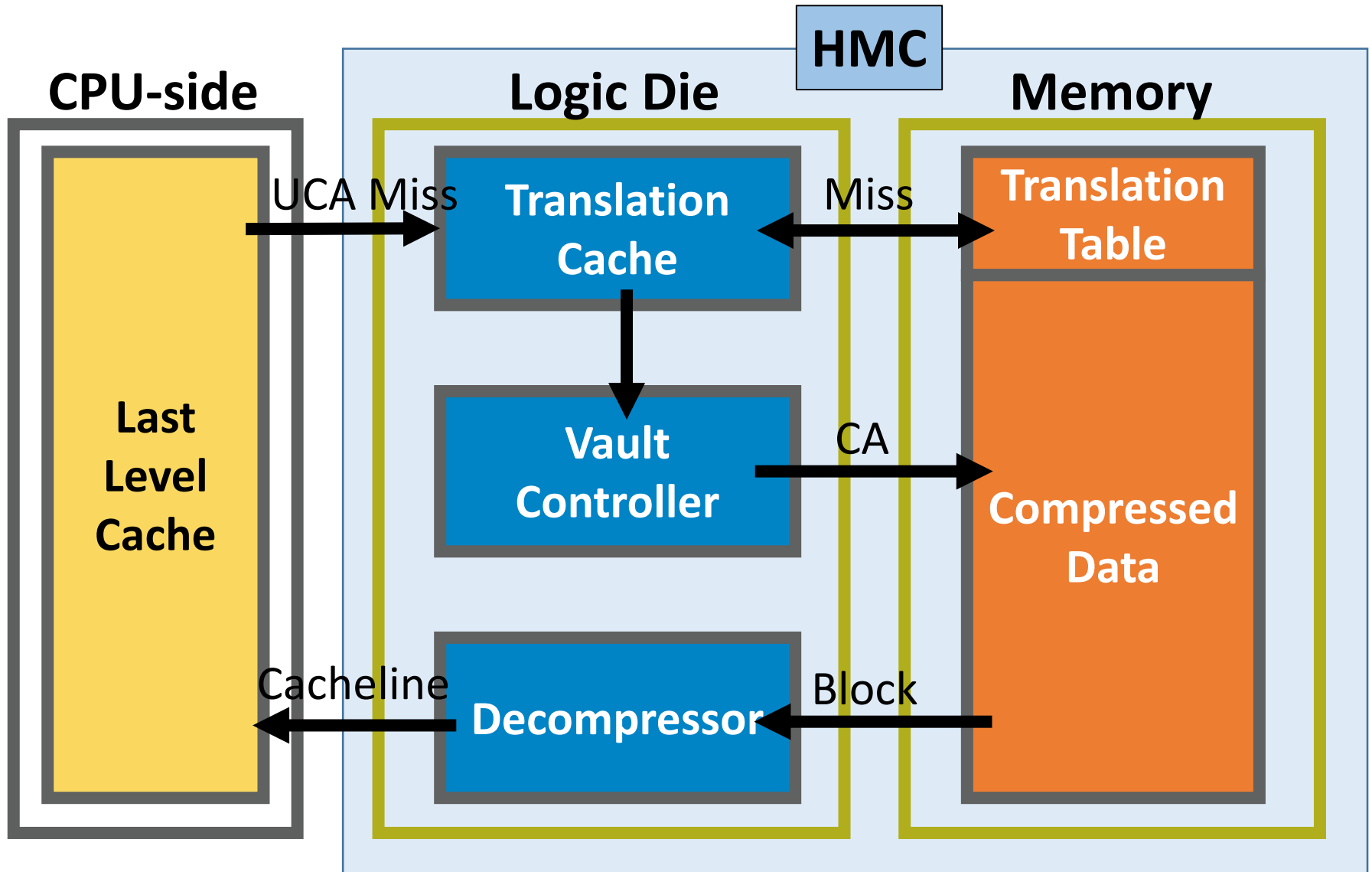
DMC: Components



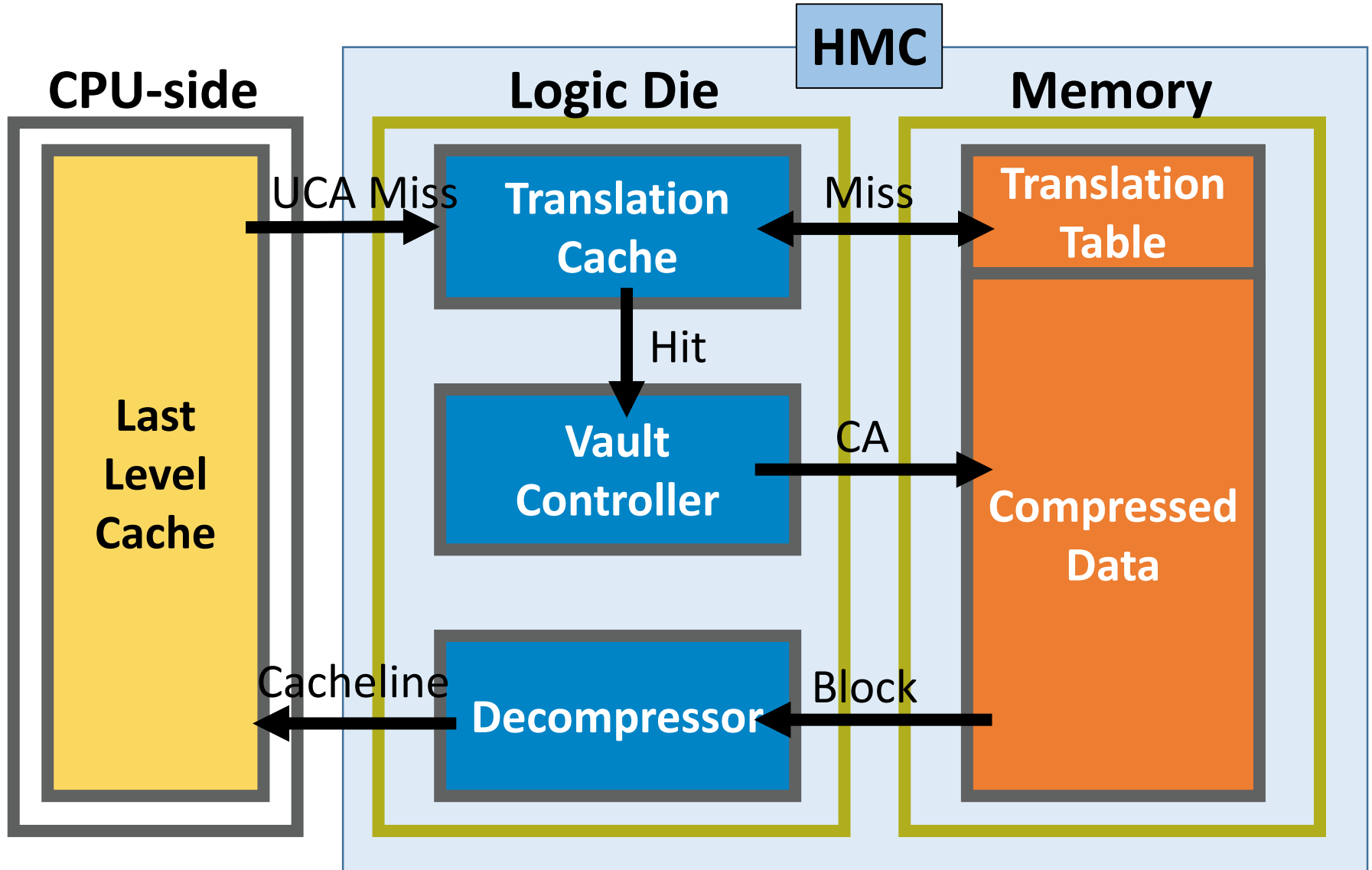
DMC: Components



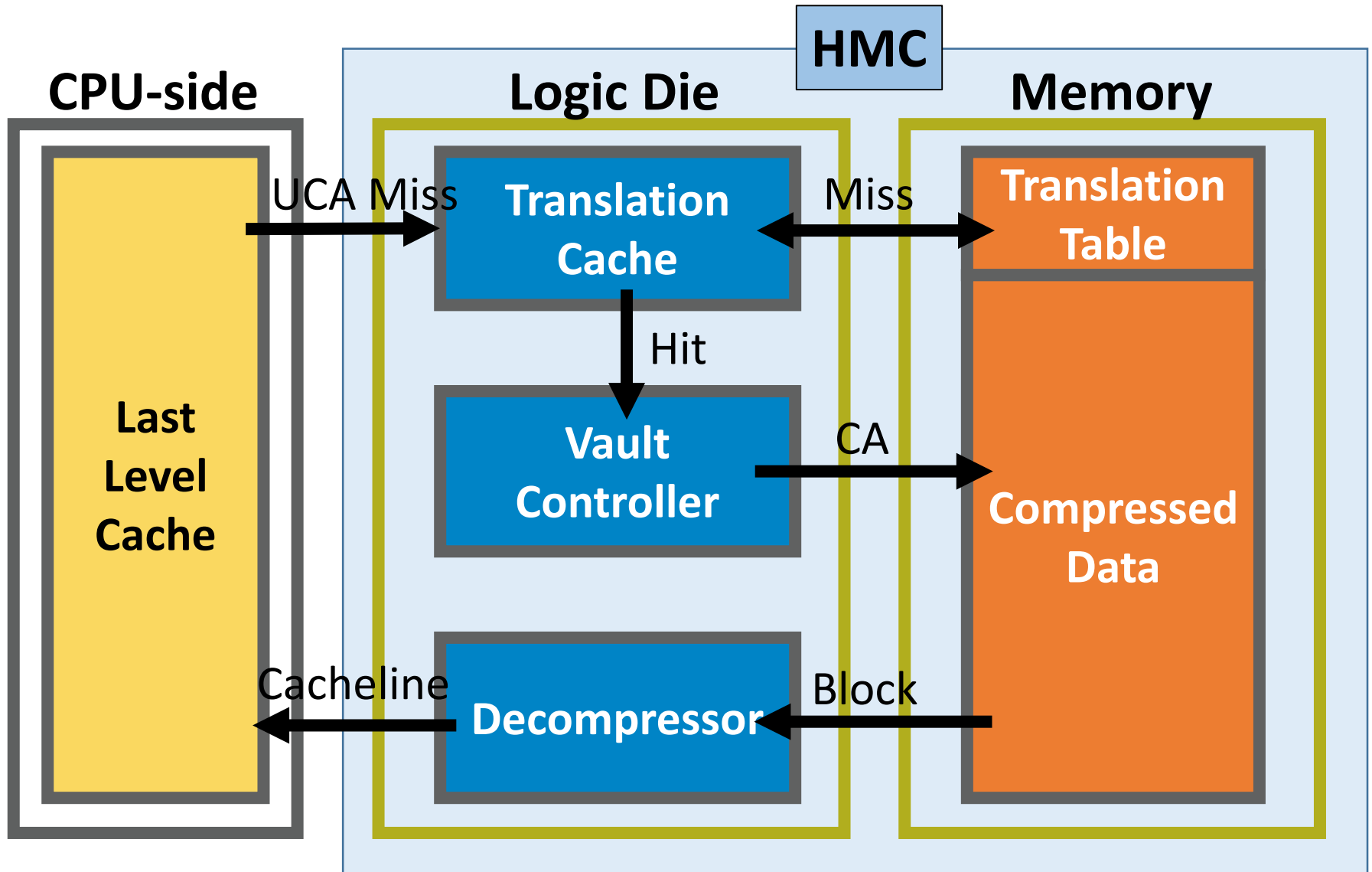
DMC: Components



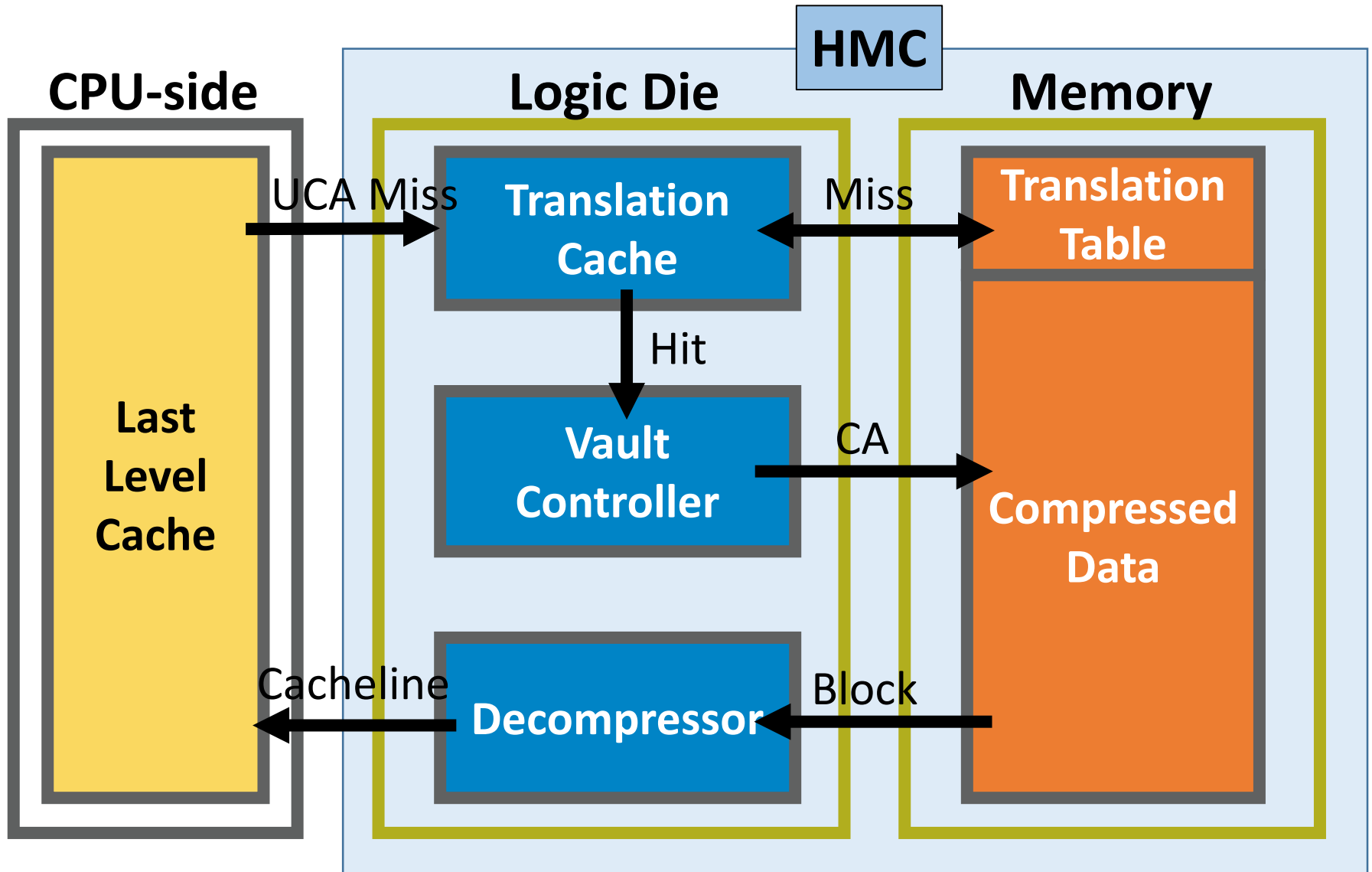
DMC: Components



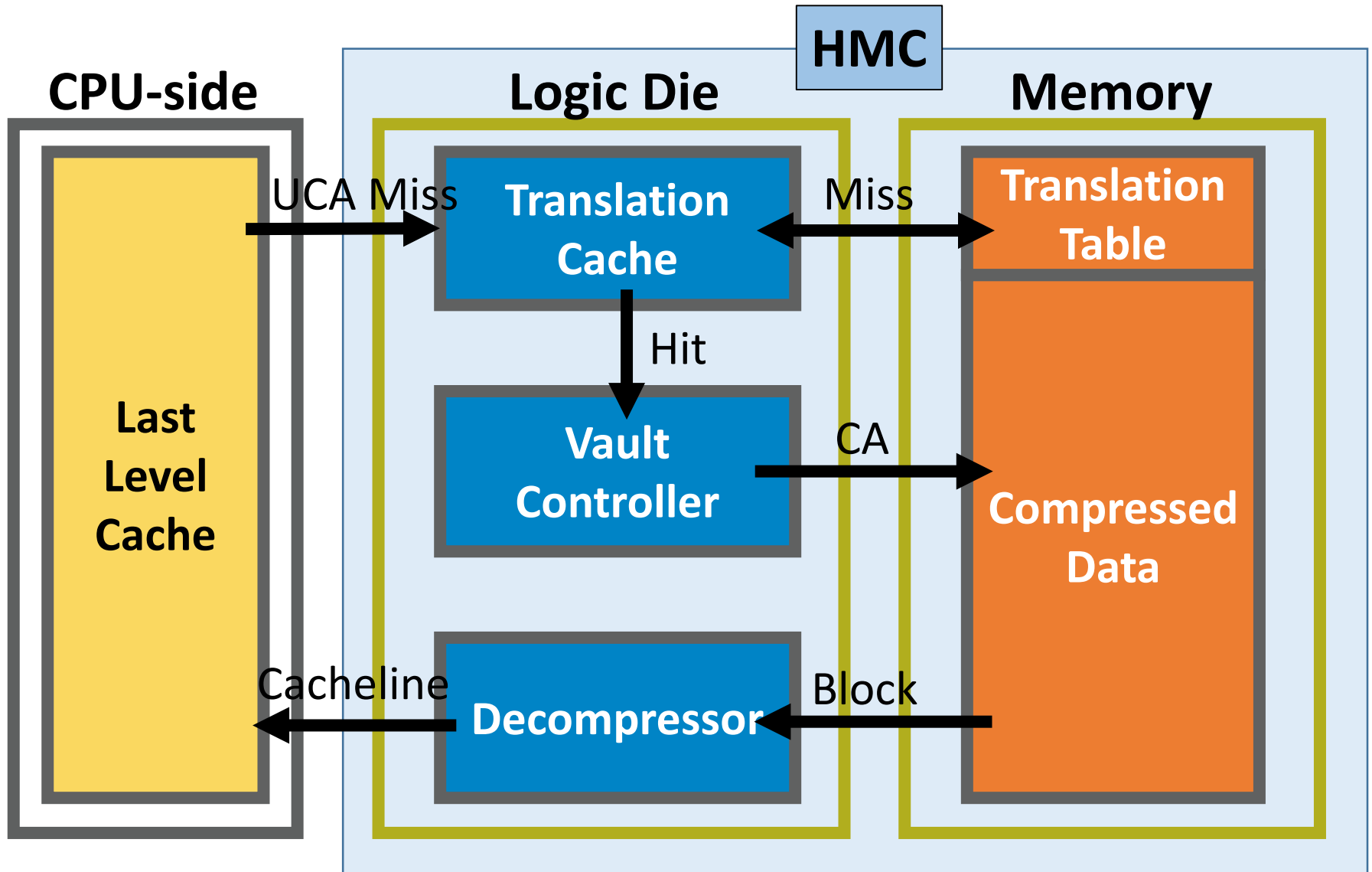
DMC: Components



DMC: Components

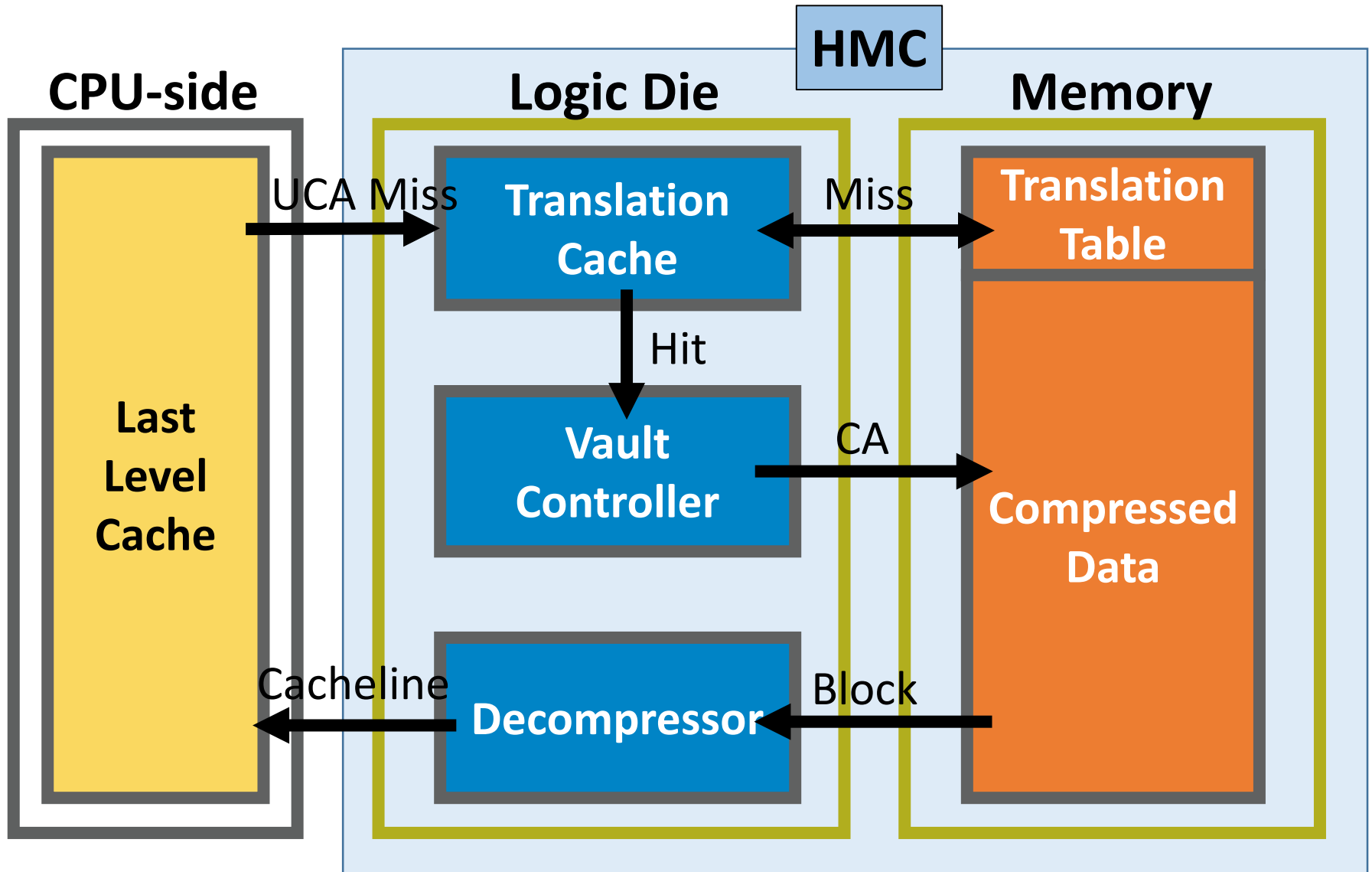


DMC: Components



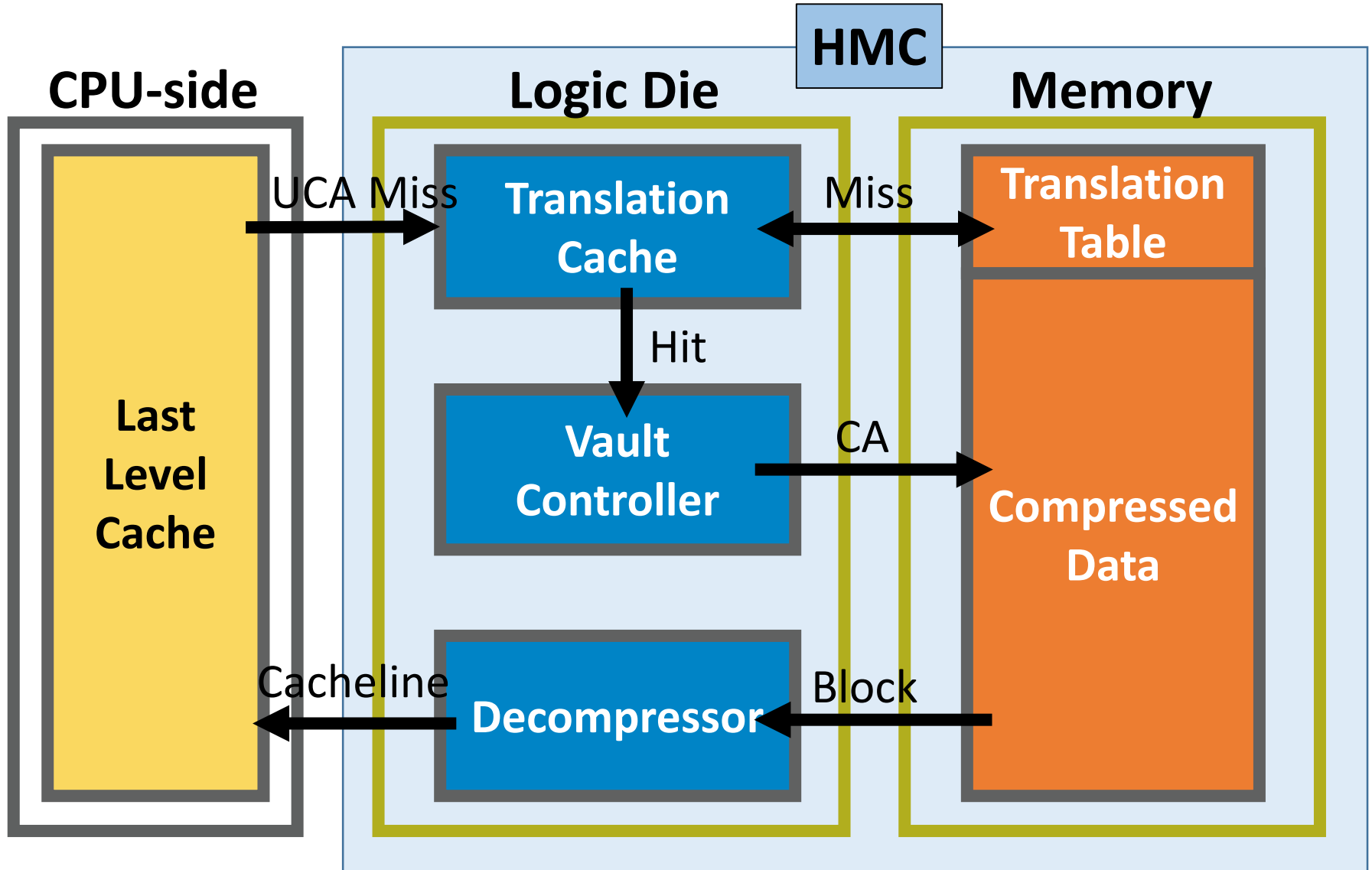
UCA: Uncompressed Address / CA: Compressed Address

DMC: Components



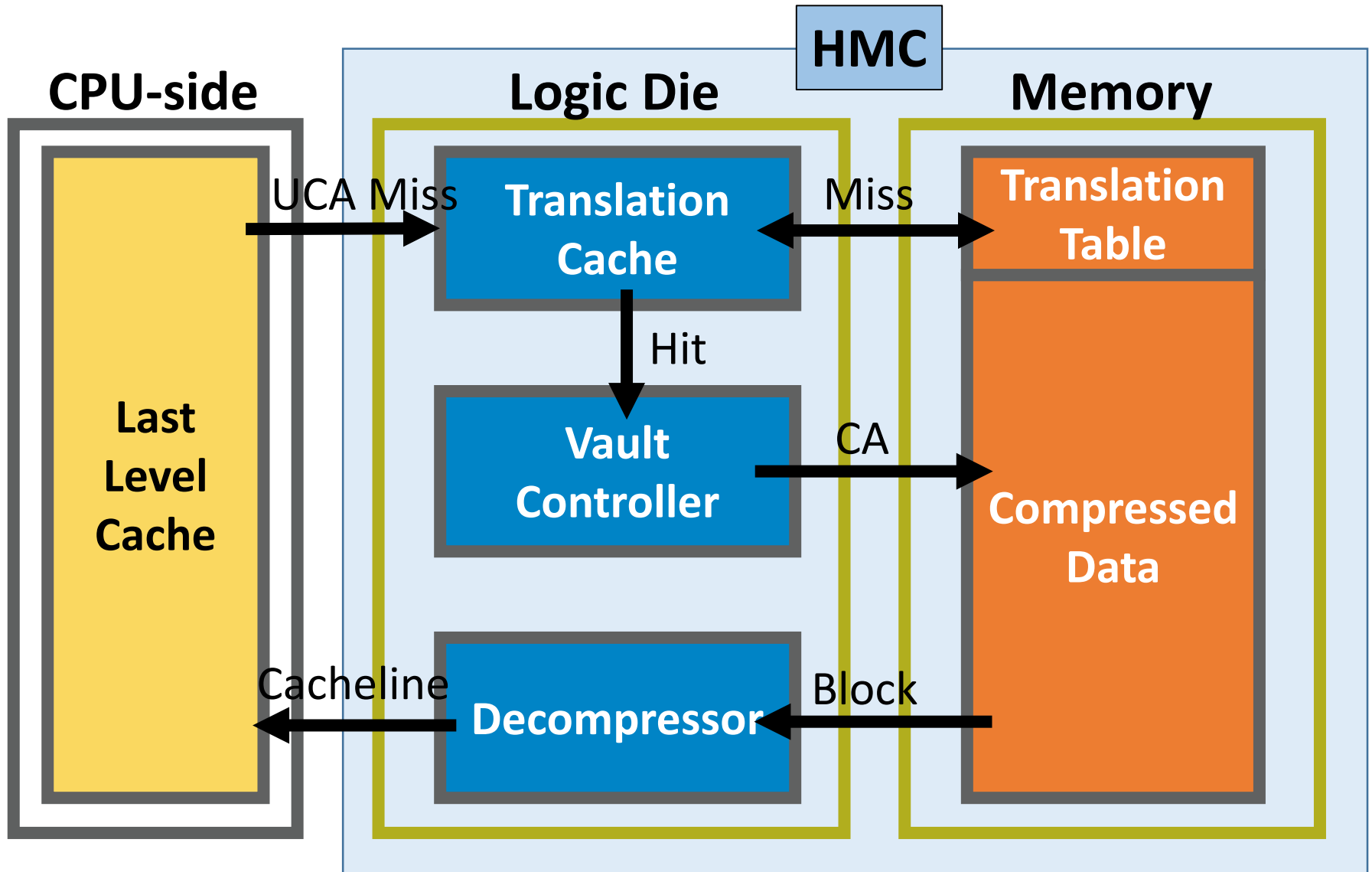
UCA: Uncompressed Address / CA: Compressed Address

DMC: Components



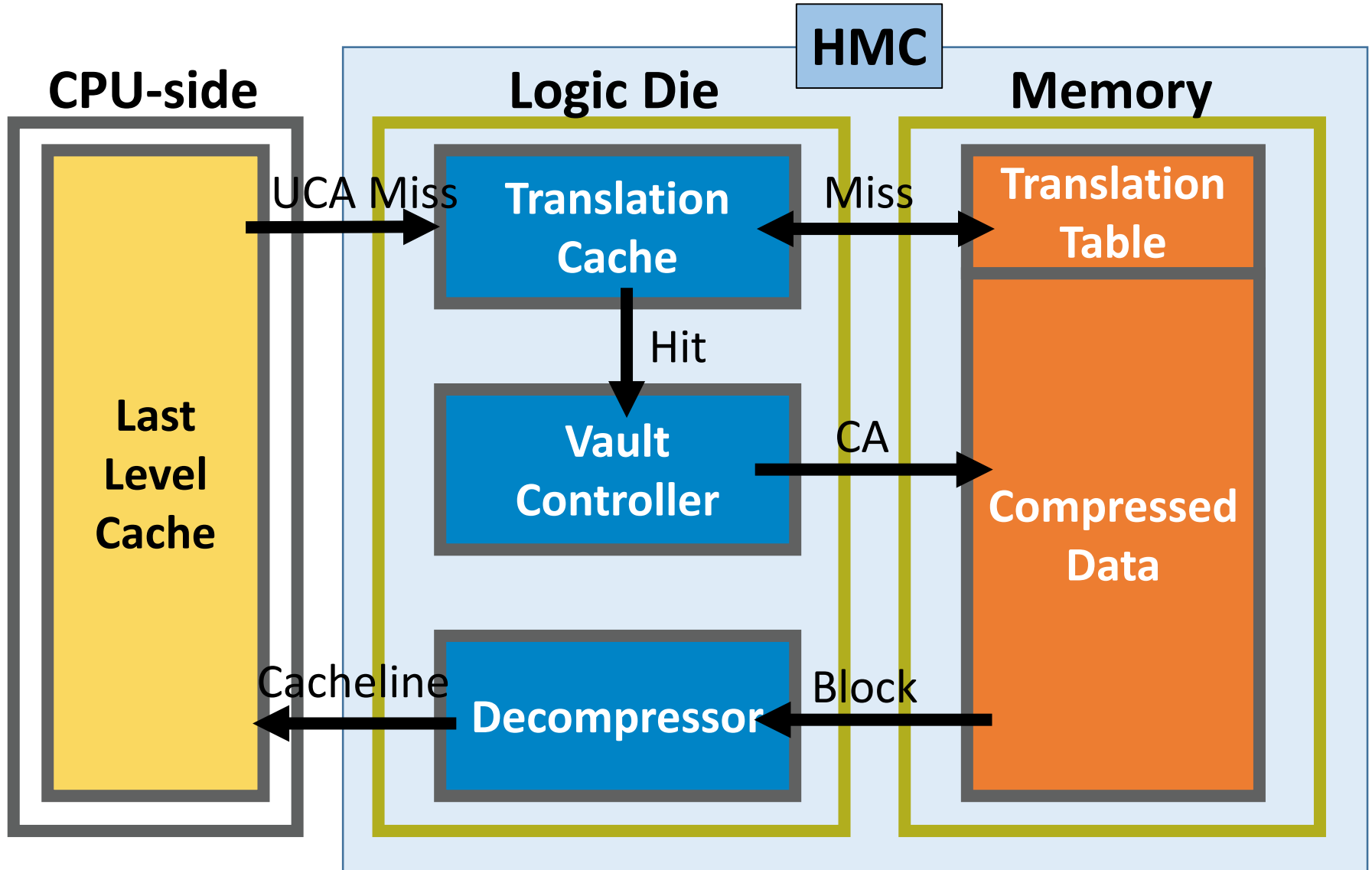
UCA: Uncompressed Address / CA: Compressed Address

DMC: Components



UCA: Uncompressed Address / CA: Compressed Address

DMC: Components



UCA: Uncompressed Address / CA: Compressed Address

Blocksize Considerations

- Block: Unit of compression
- Large block size problem
 - LCP compression ratio drops
 - LZ decompression latency increases
- Small block size problem
 - Translation cache overhead increases
 - Cost of finding cold regions increases

Blocksize Considerations

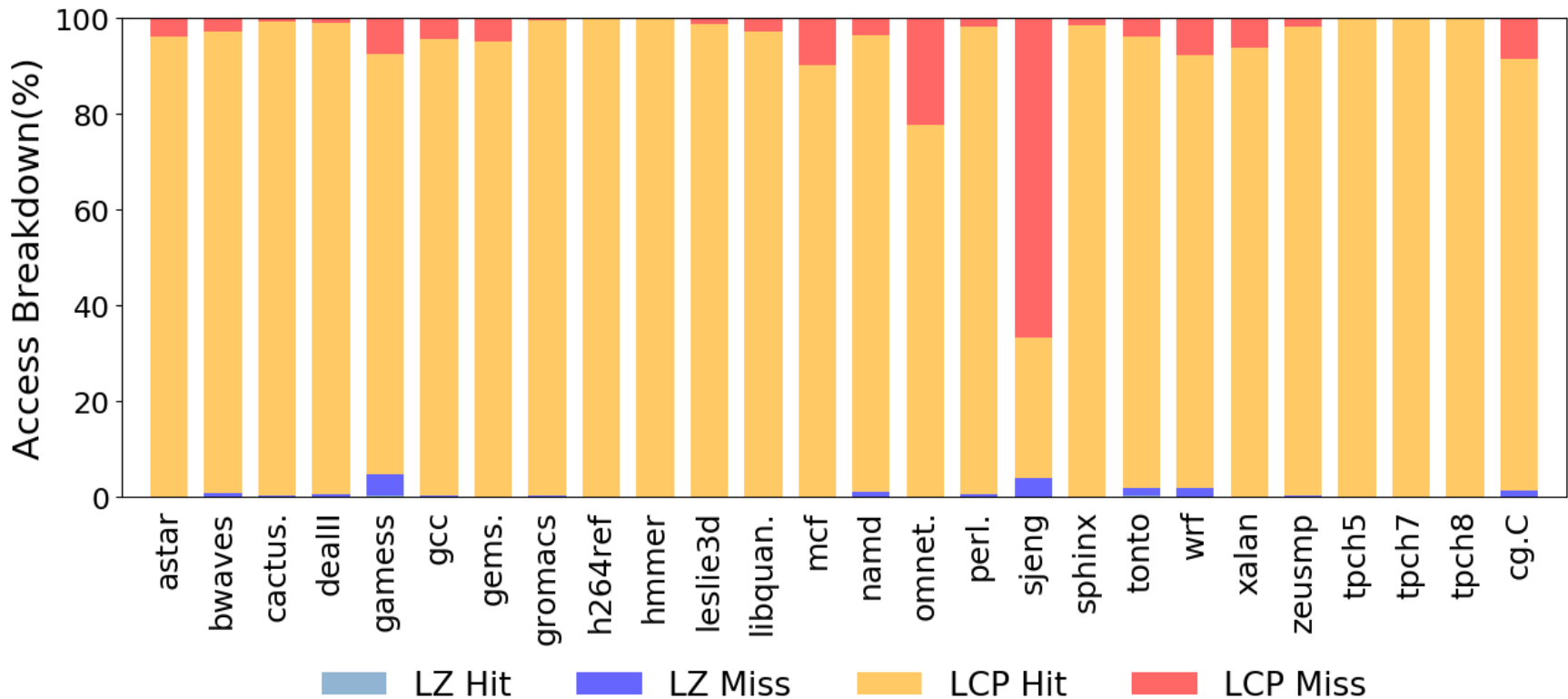
- Block: Unit of compression
- Large block size problem
 - LCP compression ratio drops

Two Different Block Sizes:
One for *LCP*, One for *LZ*

- Cost of finding cold regions increases

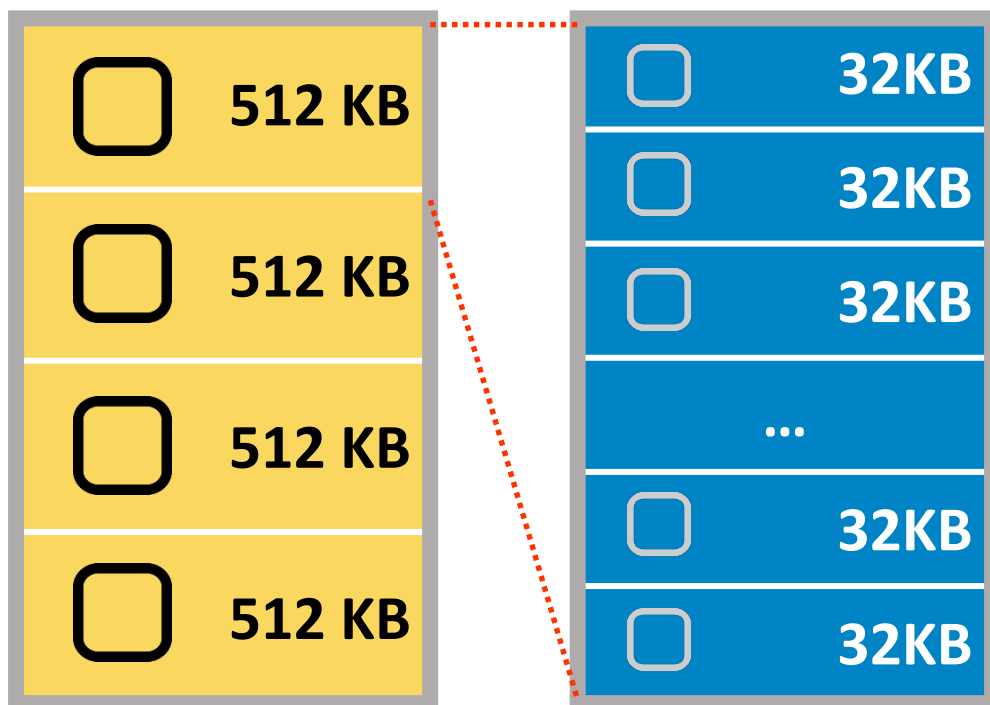
Translation Cache Decomposition

- Mostly LCP hits and misses
- Larger block on LCP
- Smaller block on LZ



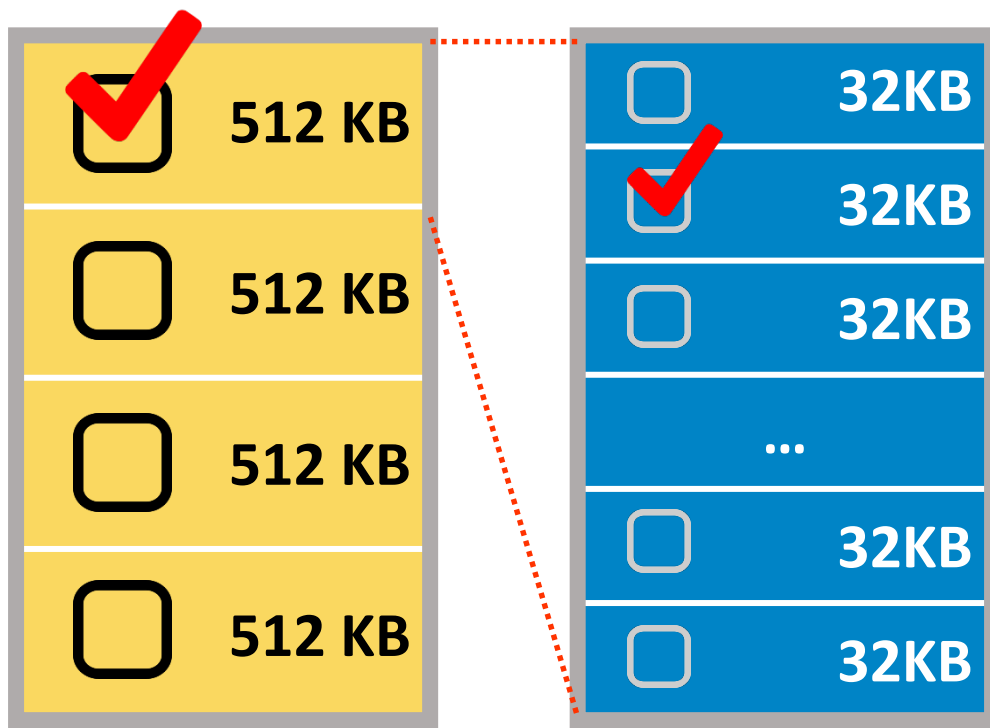
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



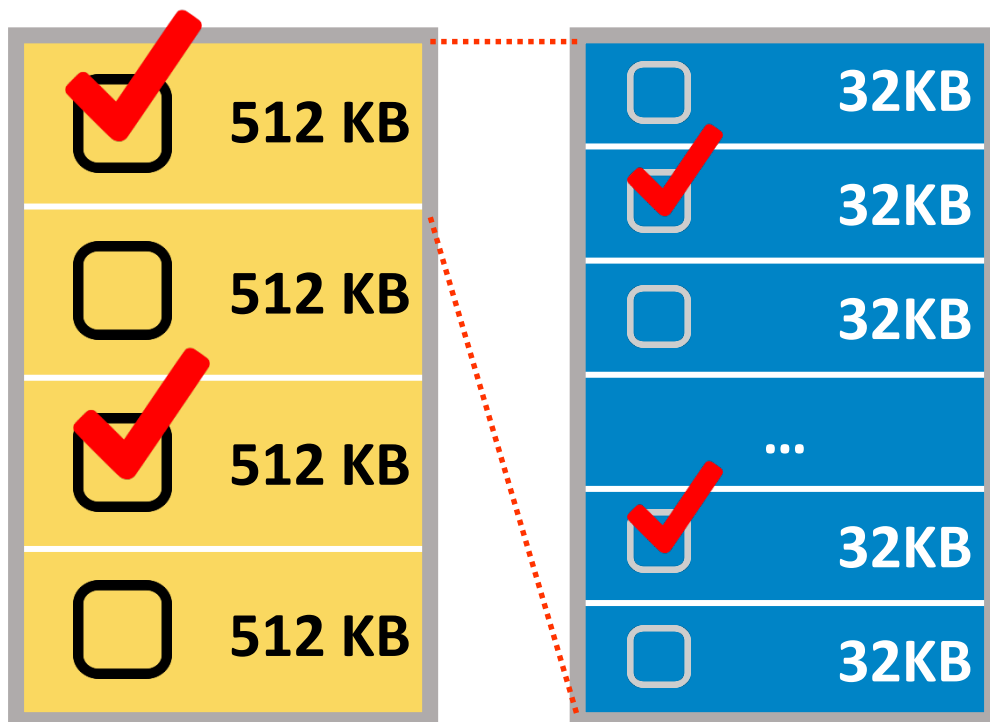
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



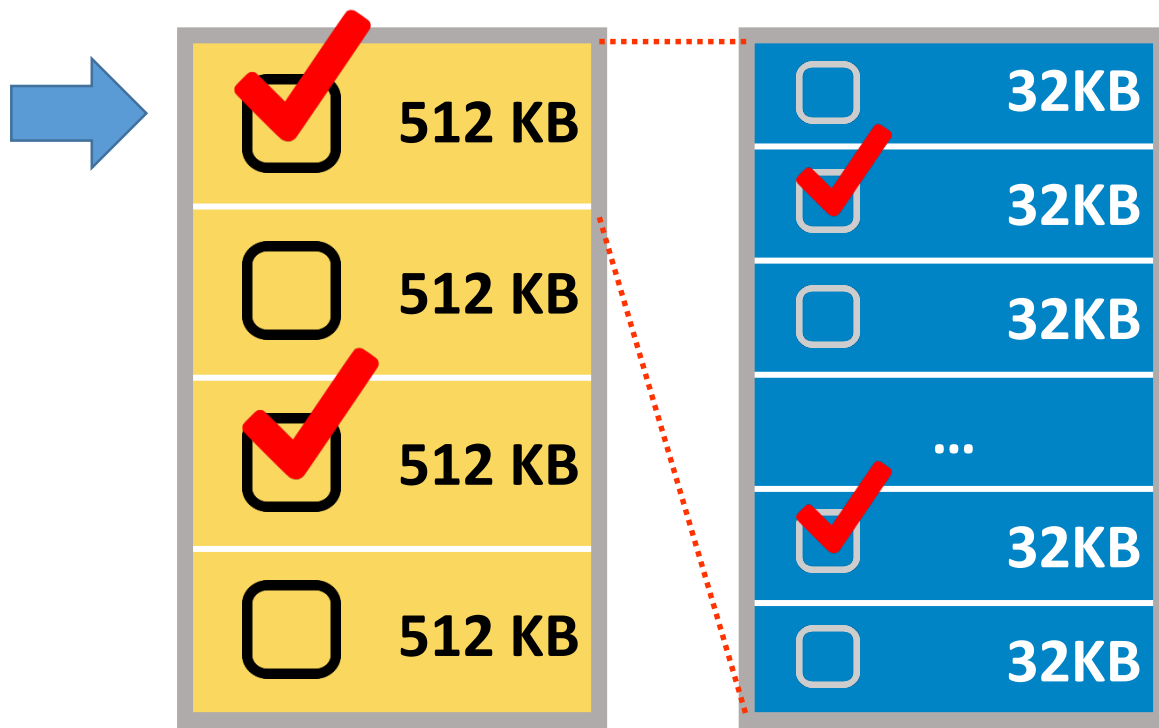
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



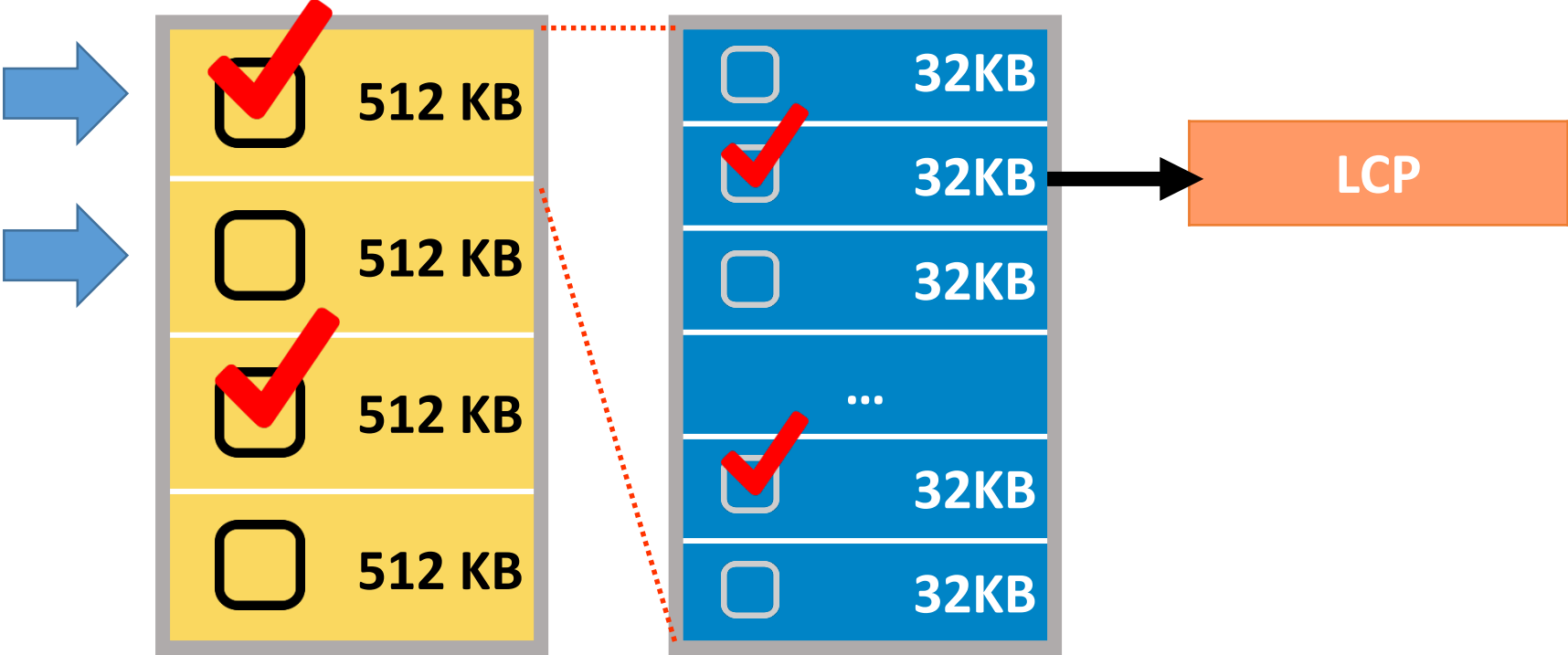
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



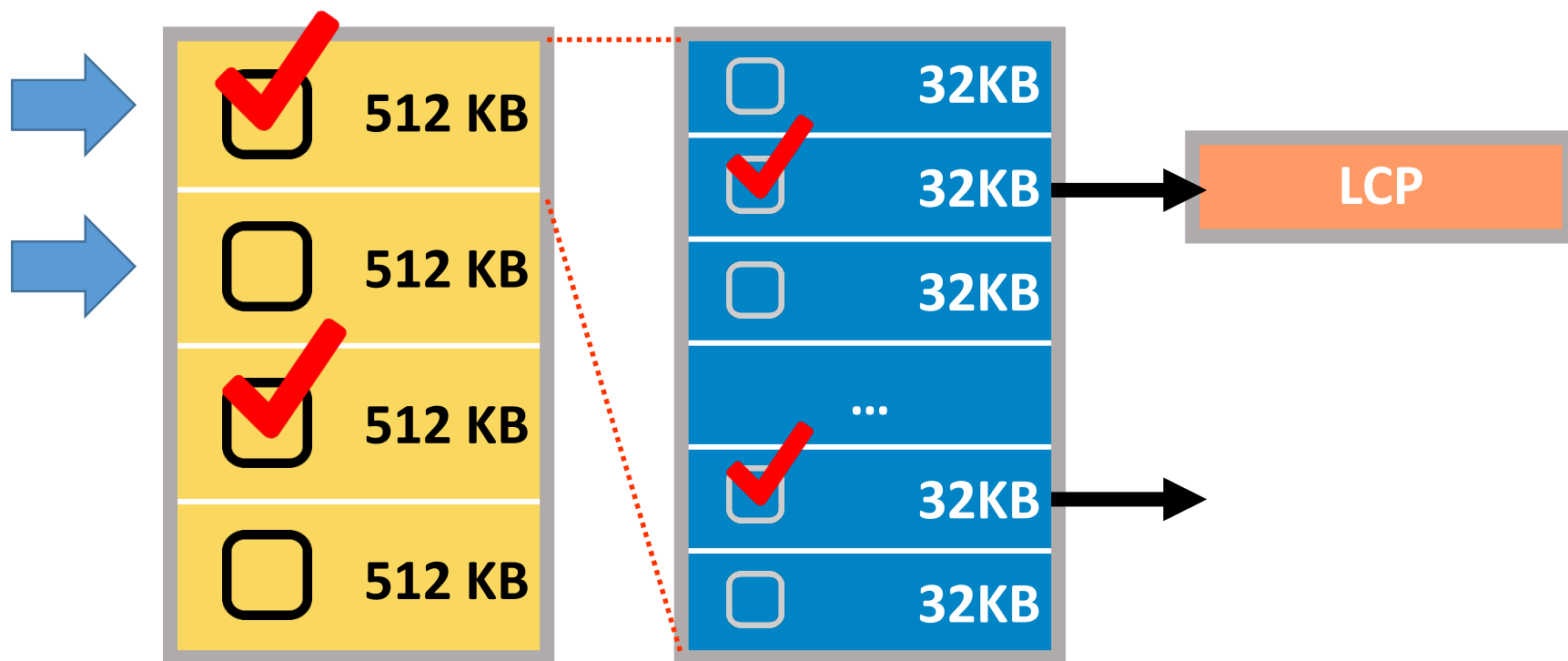
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



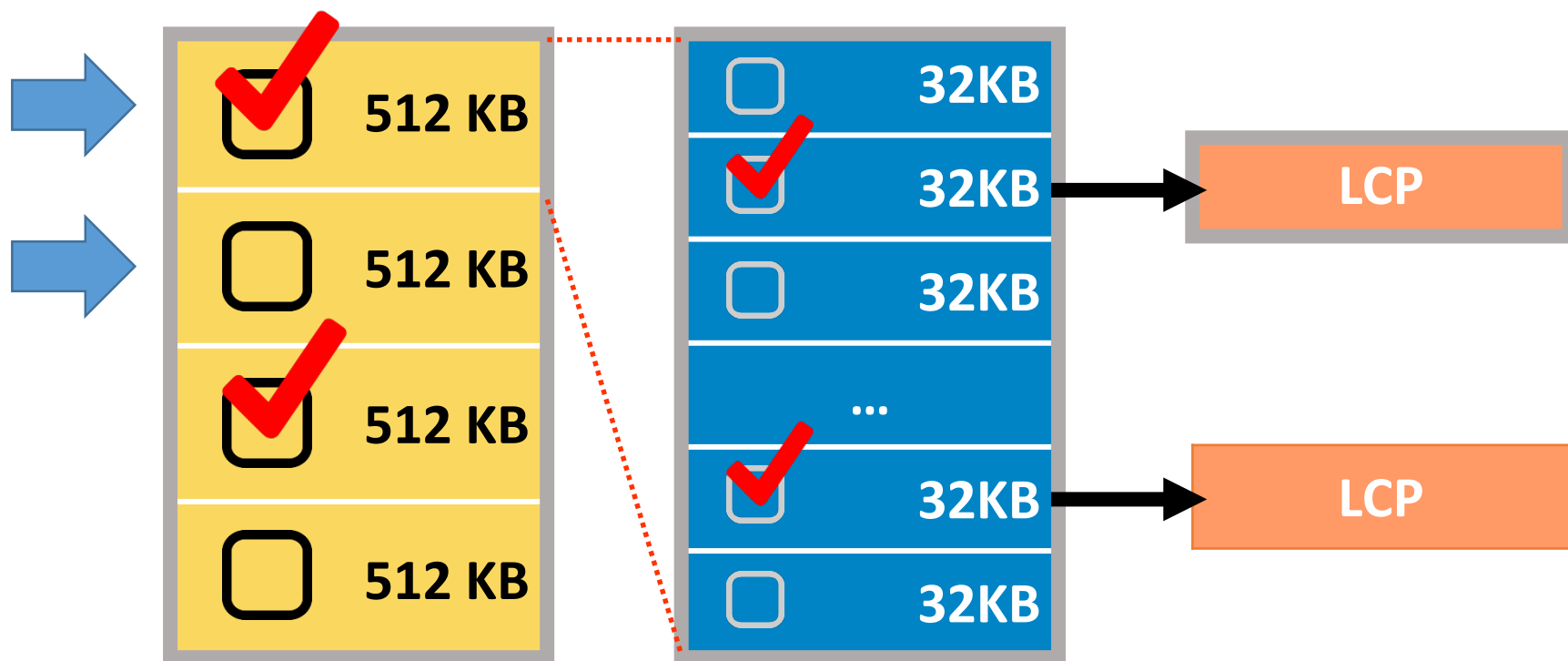
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



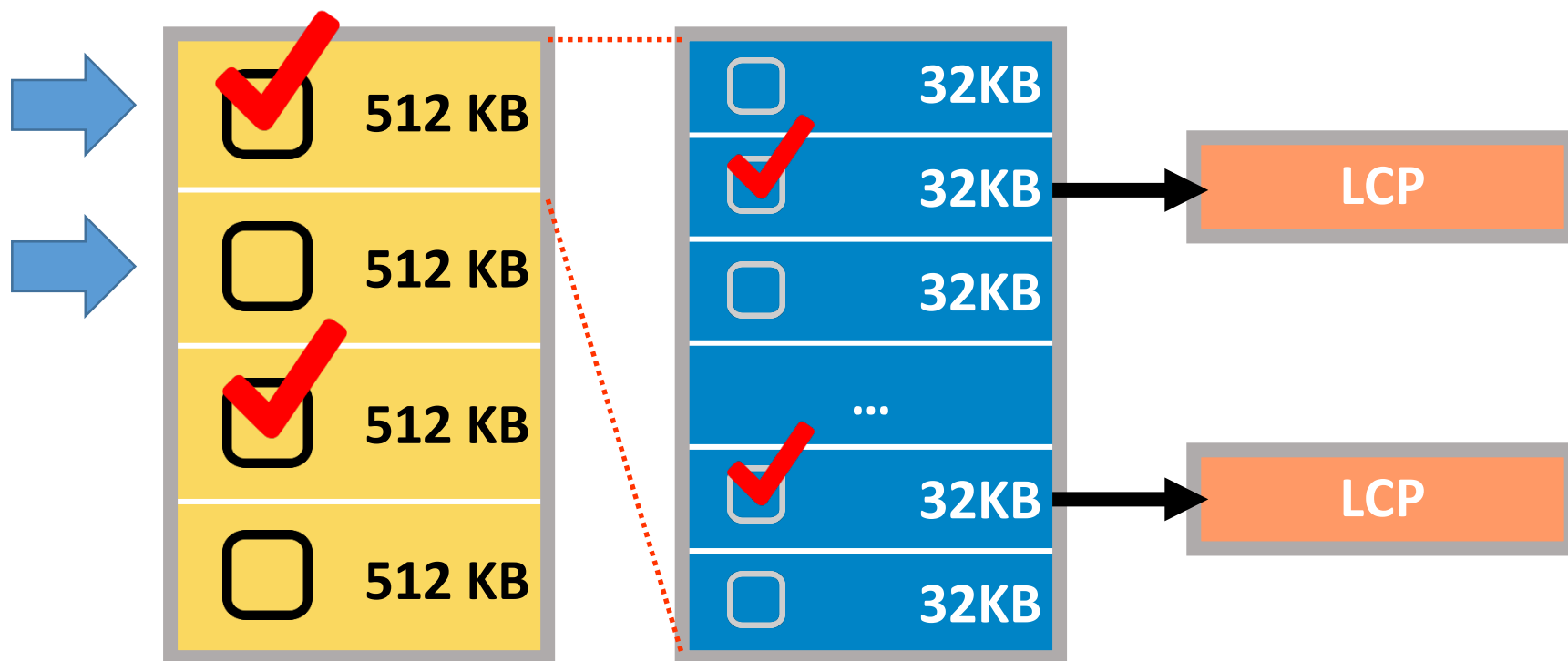
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



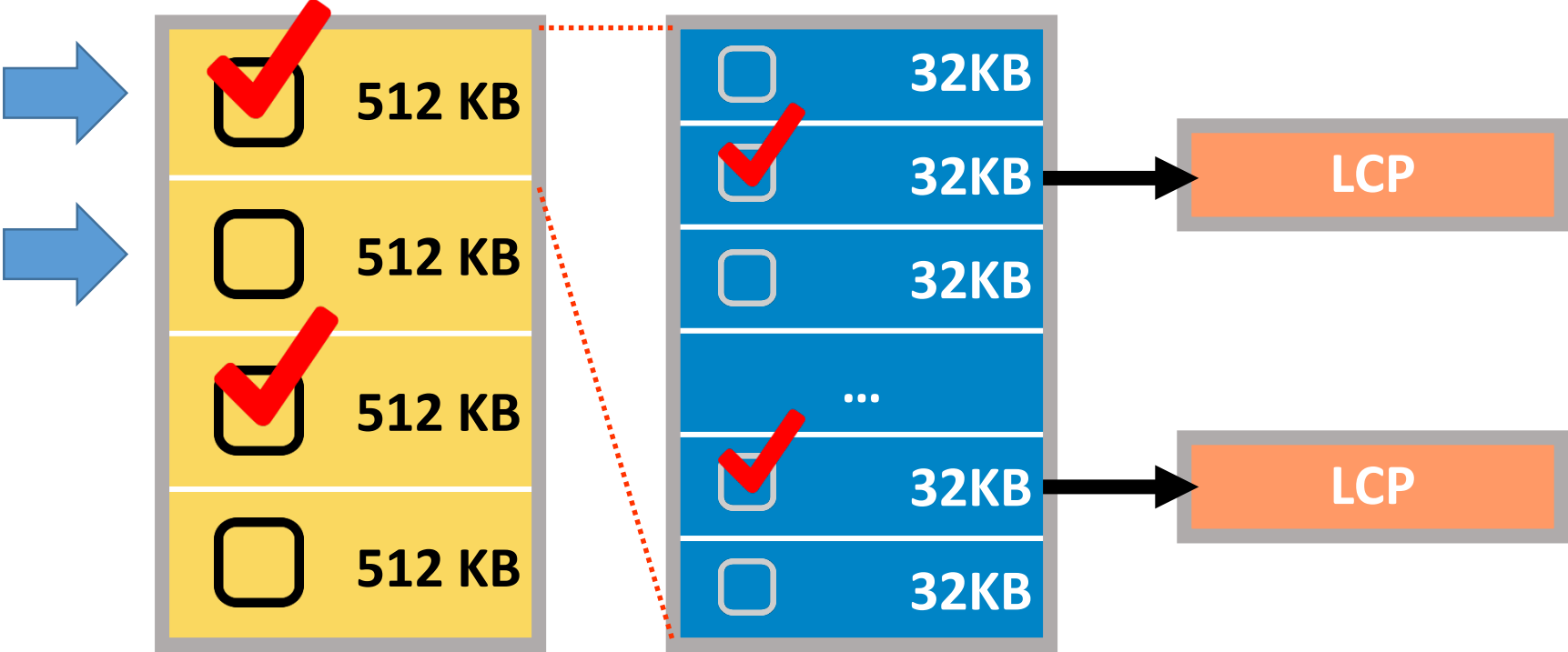
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



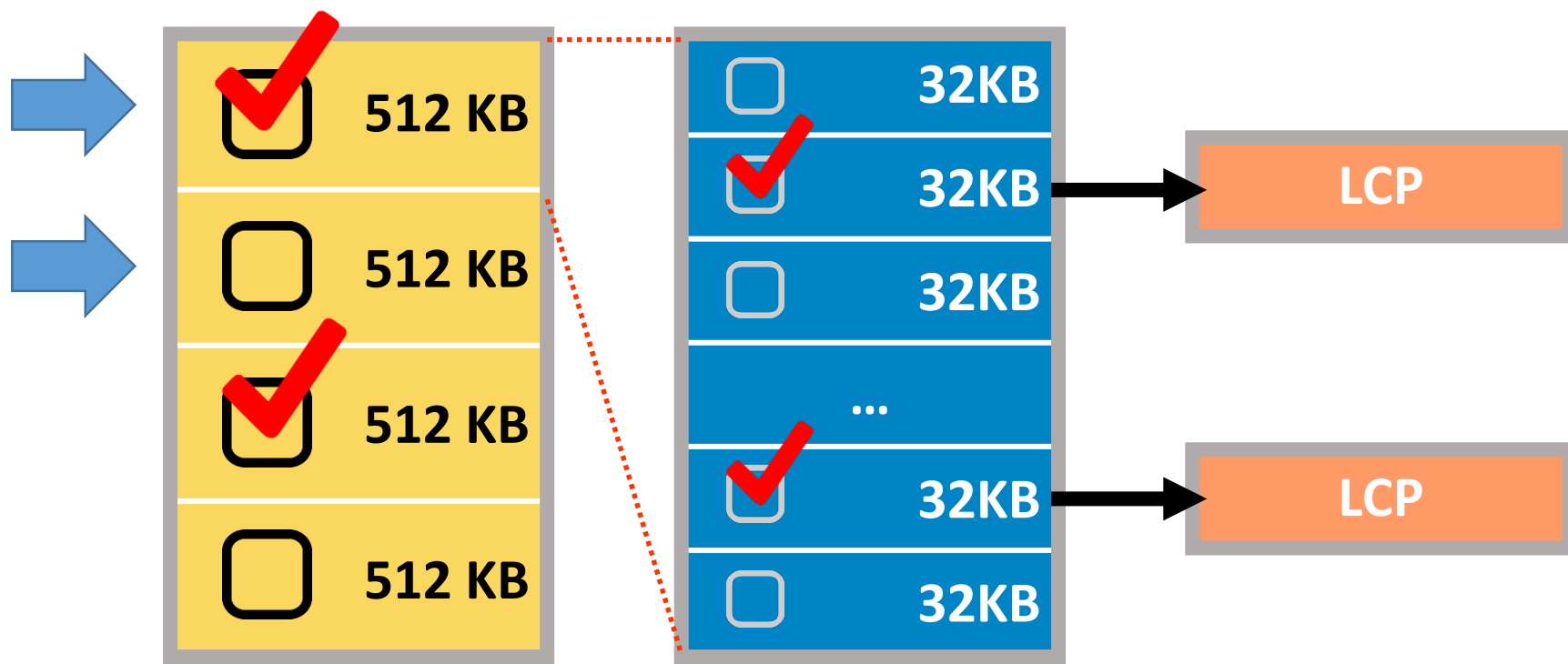
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



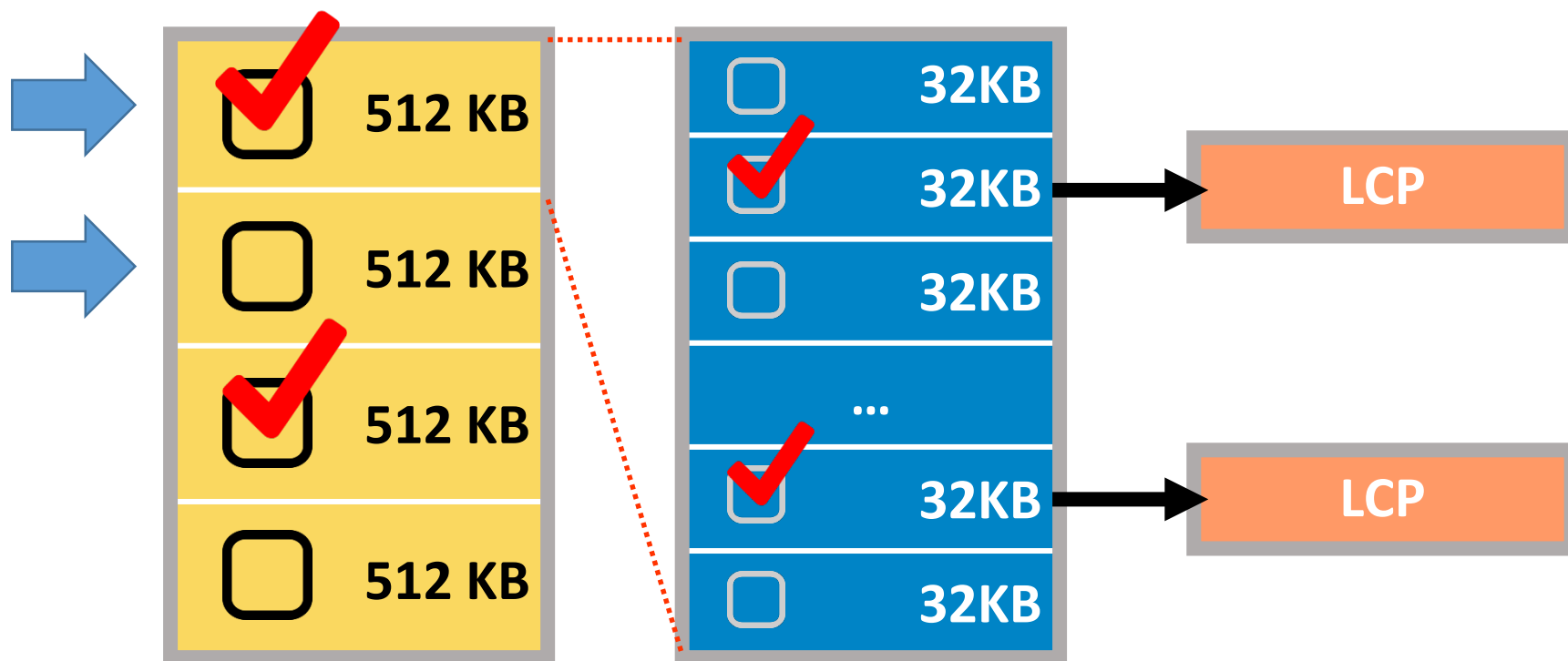
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



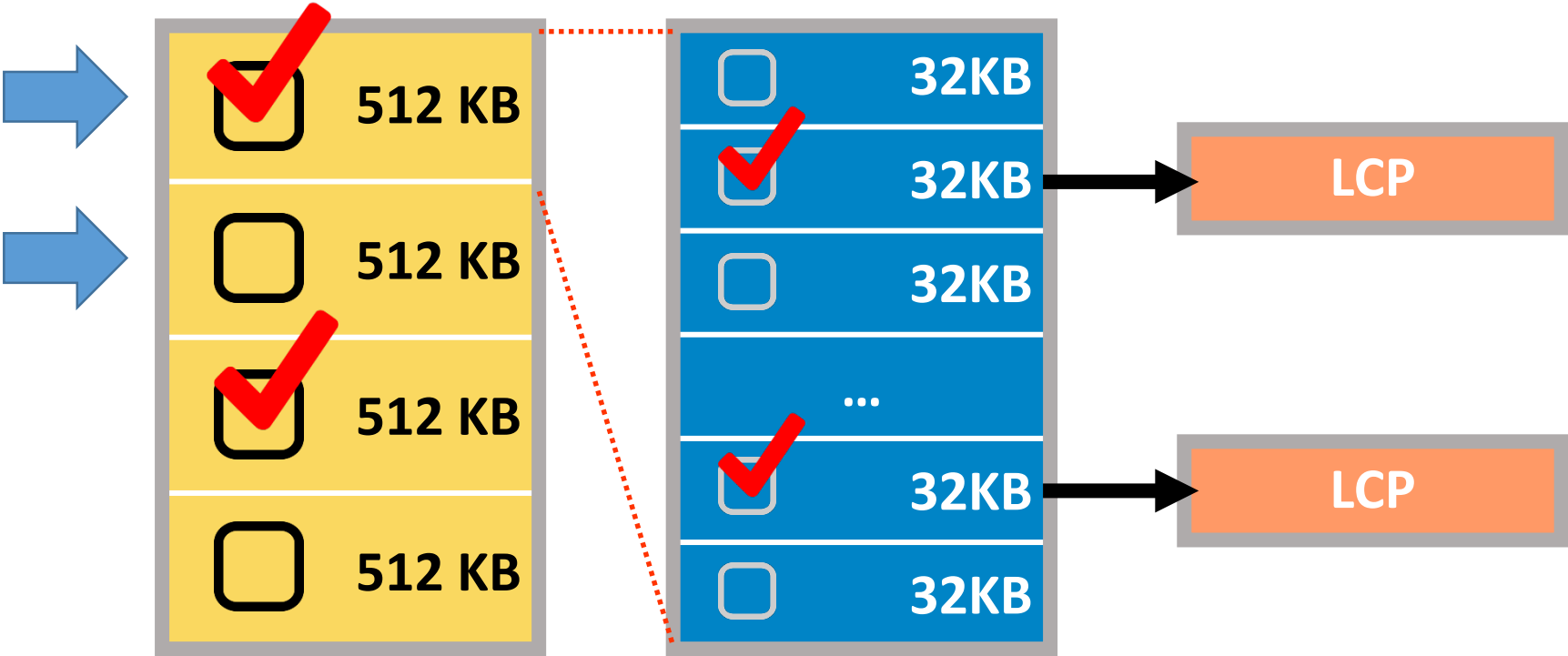
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



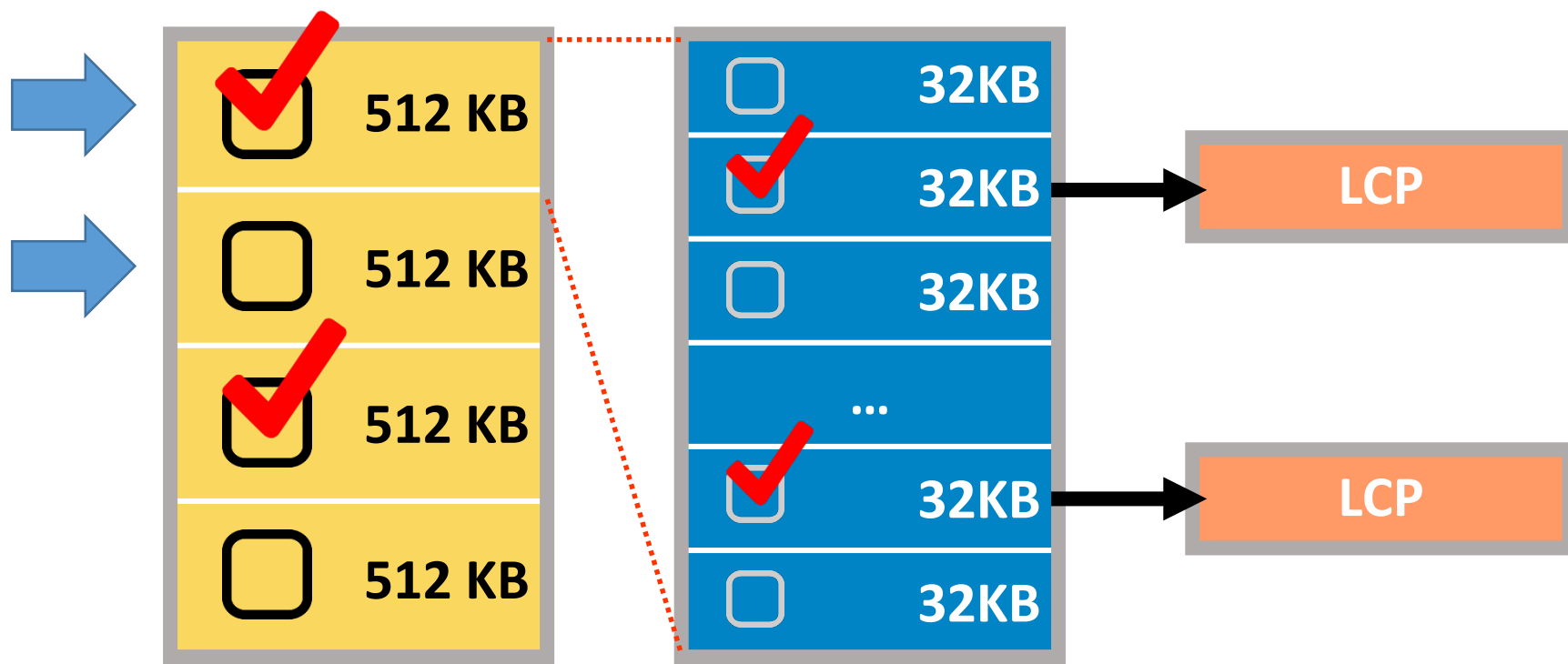
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



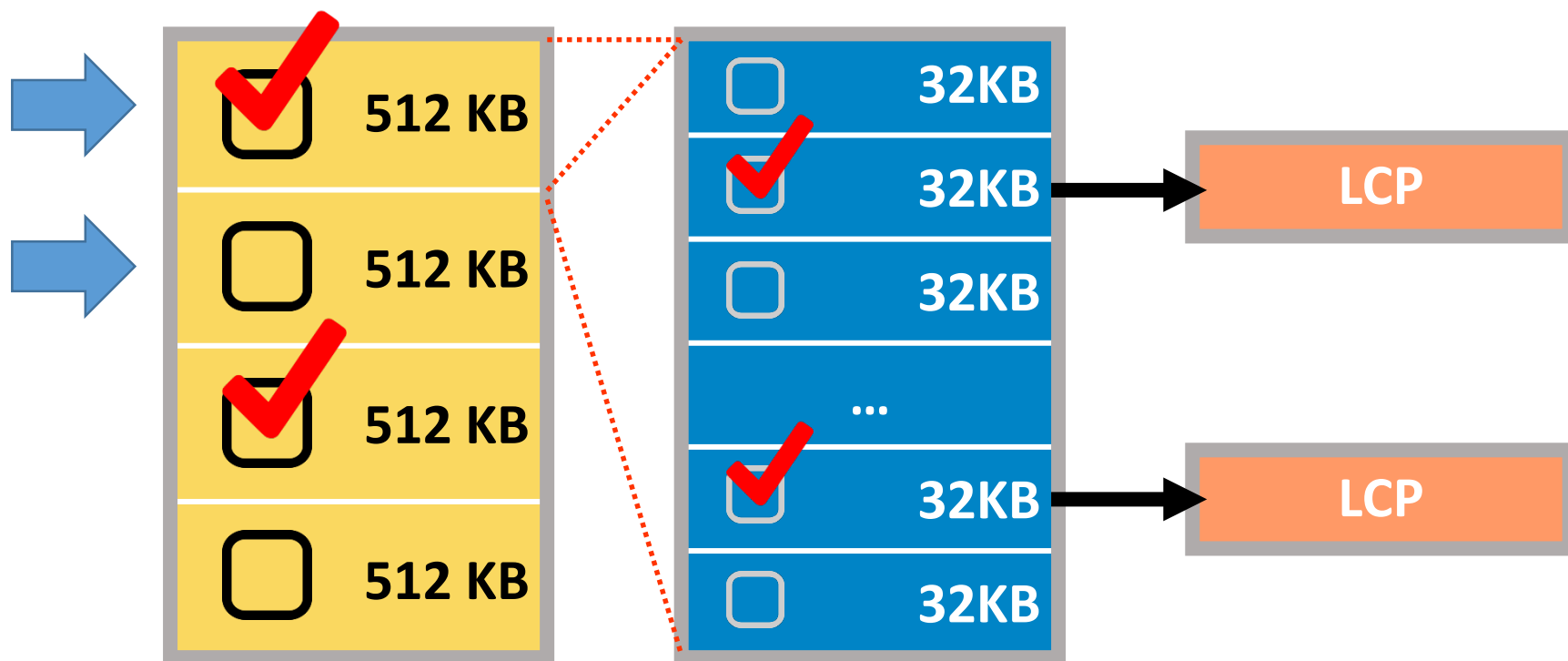
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



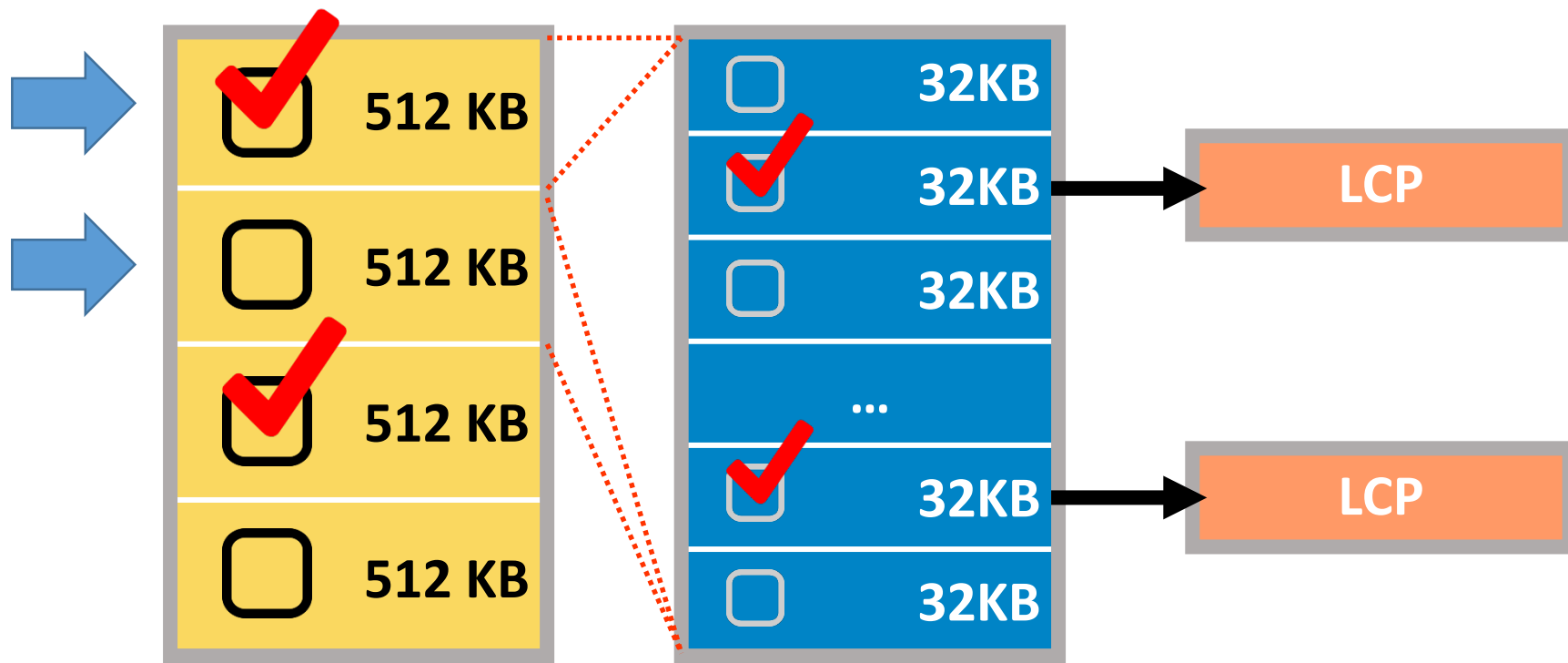
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



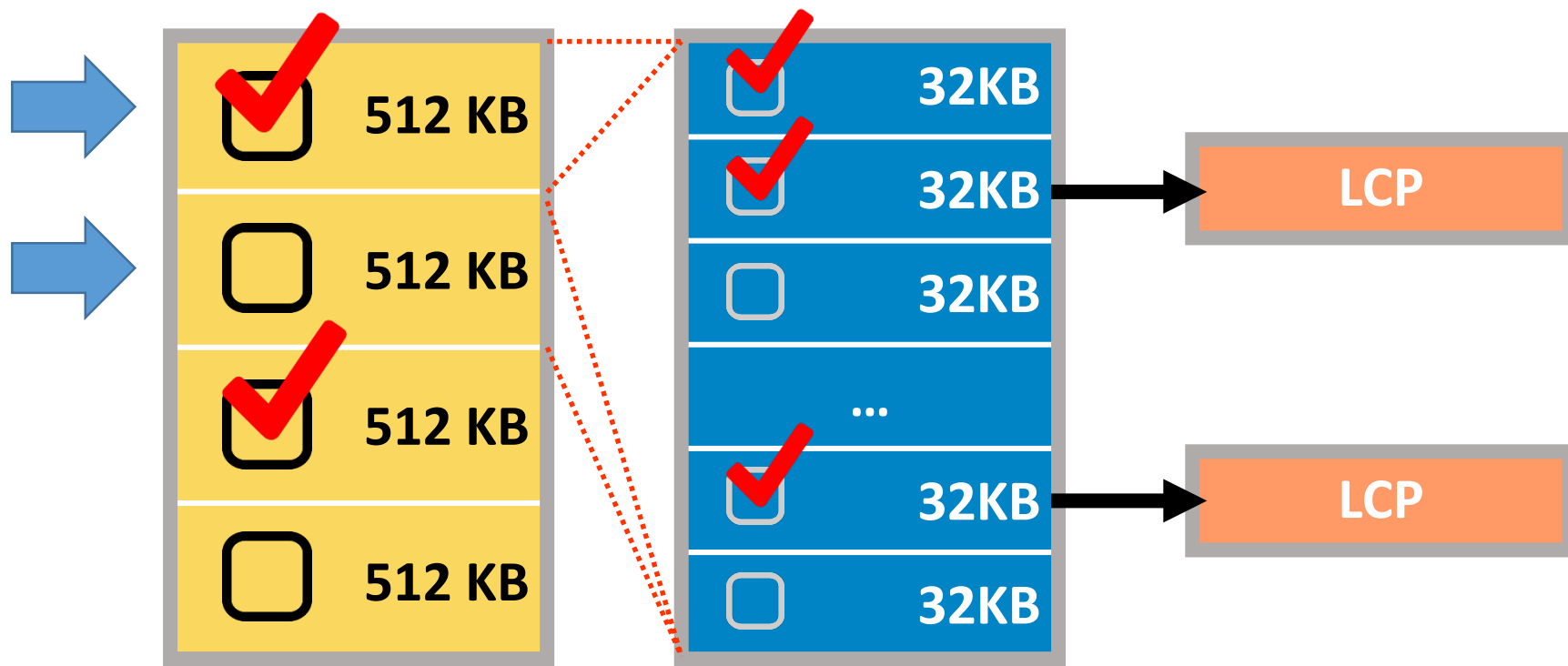
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



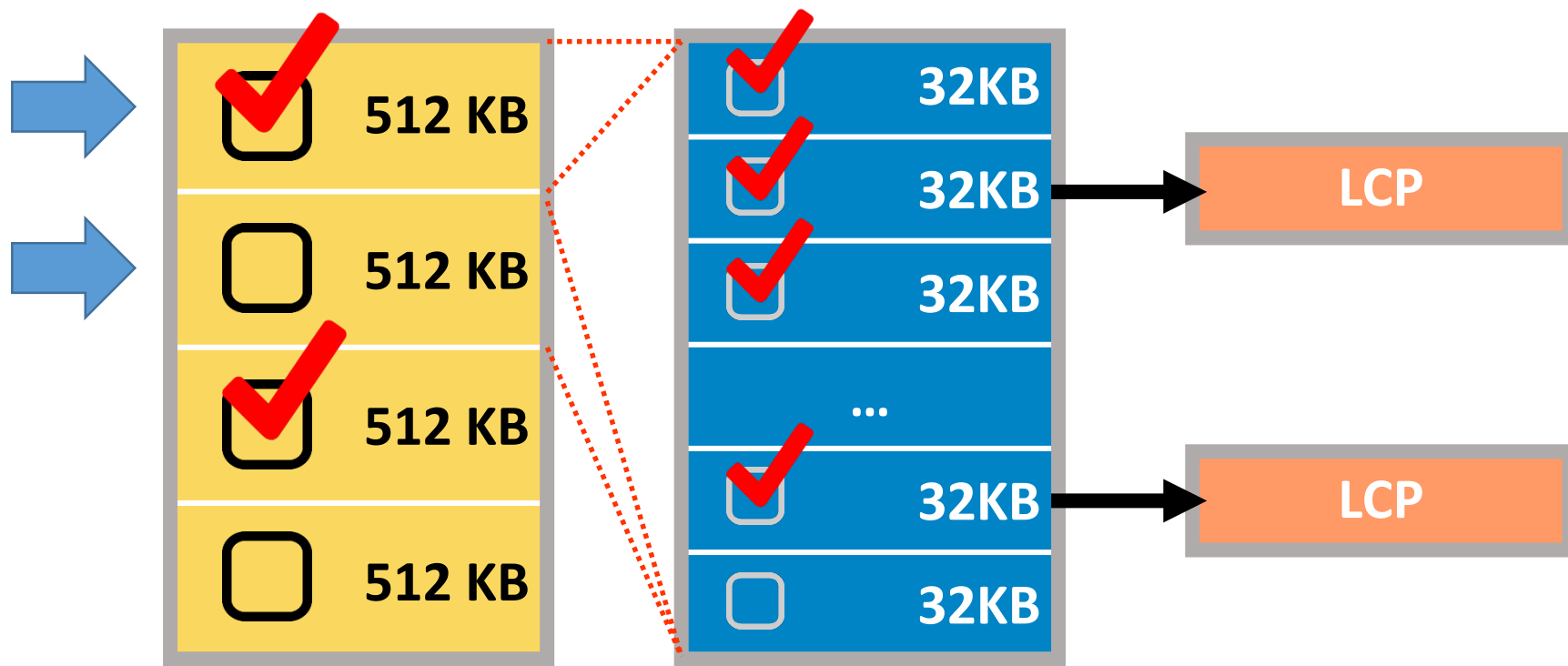
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



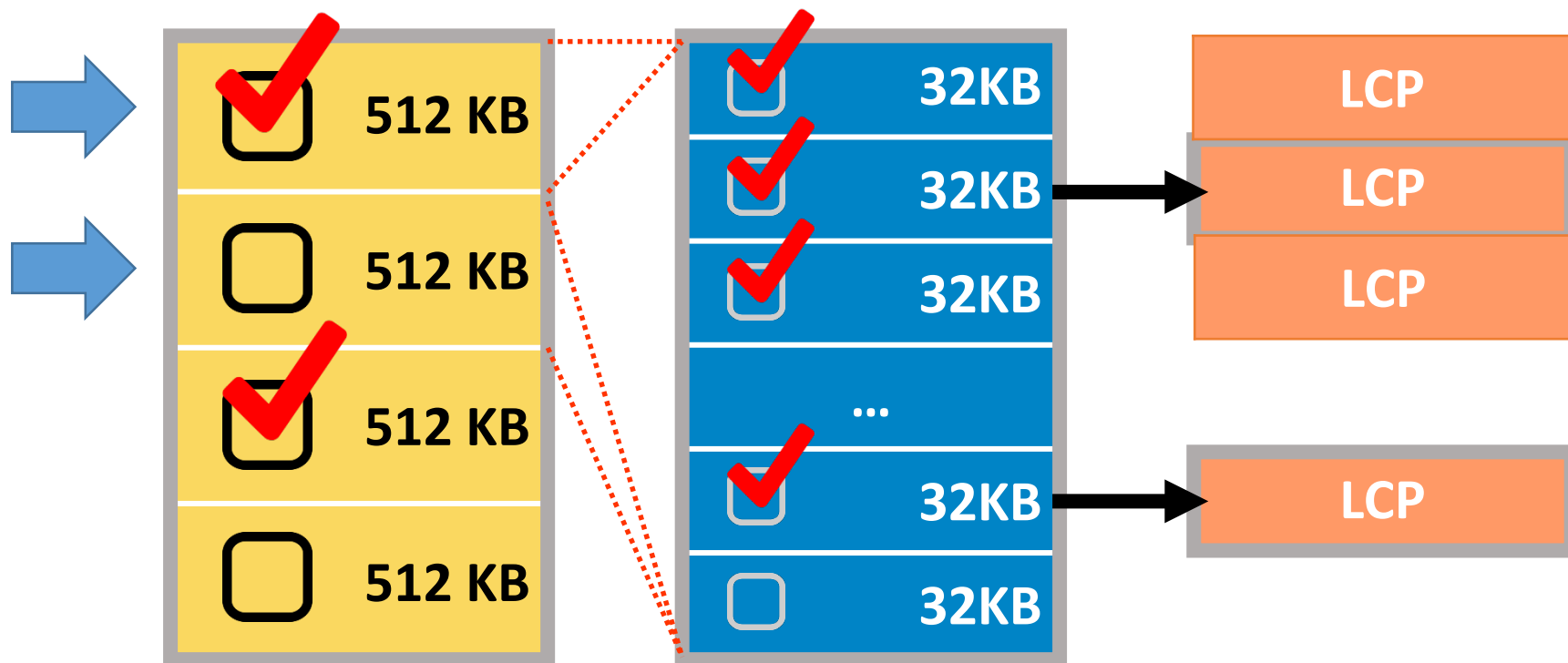
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



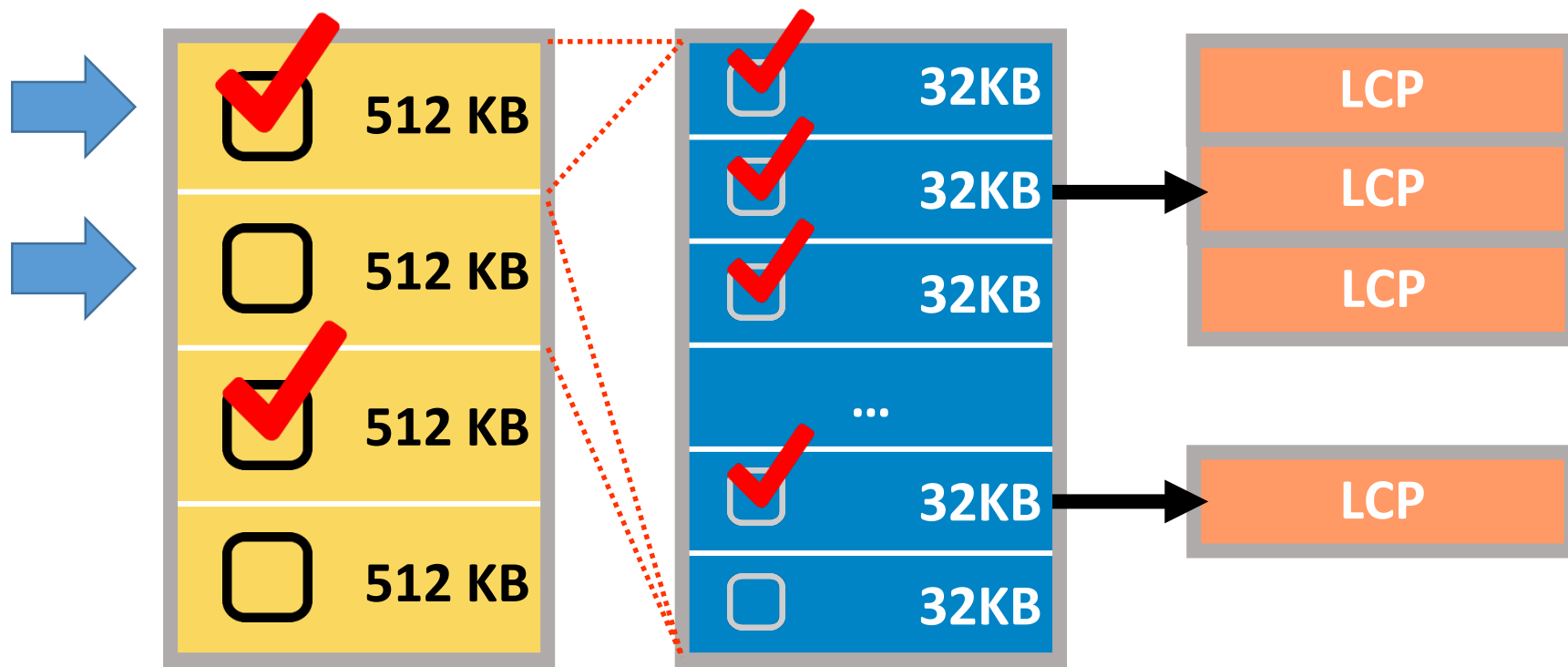
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



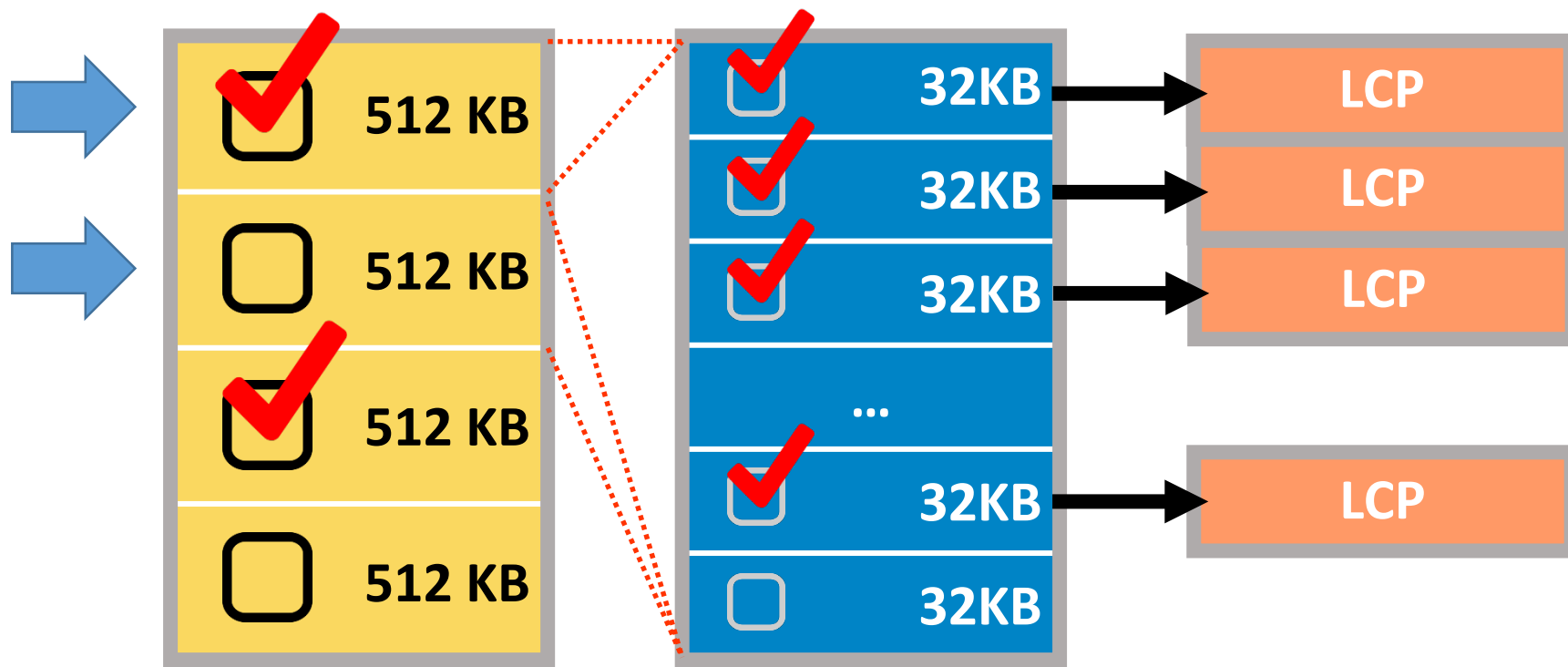
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



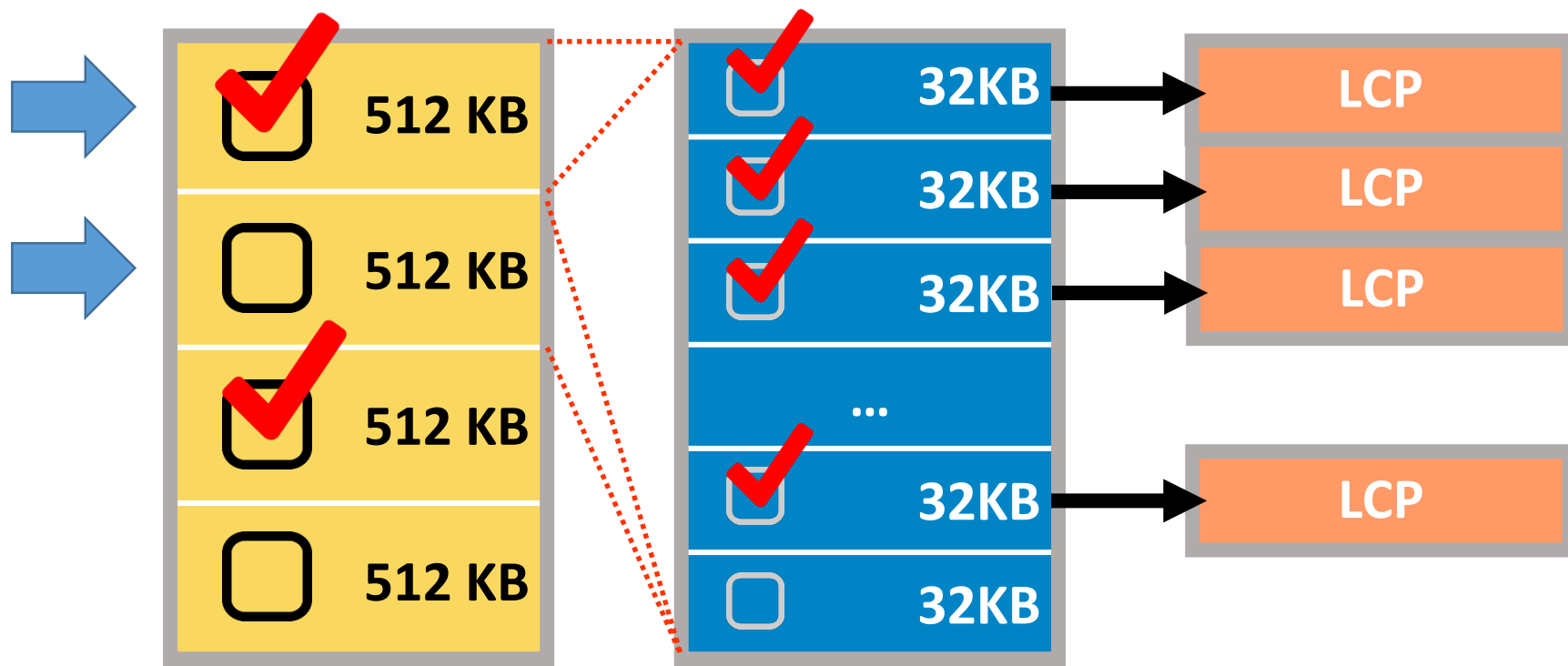
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



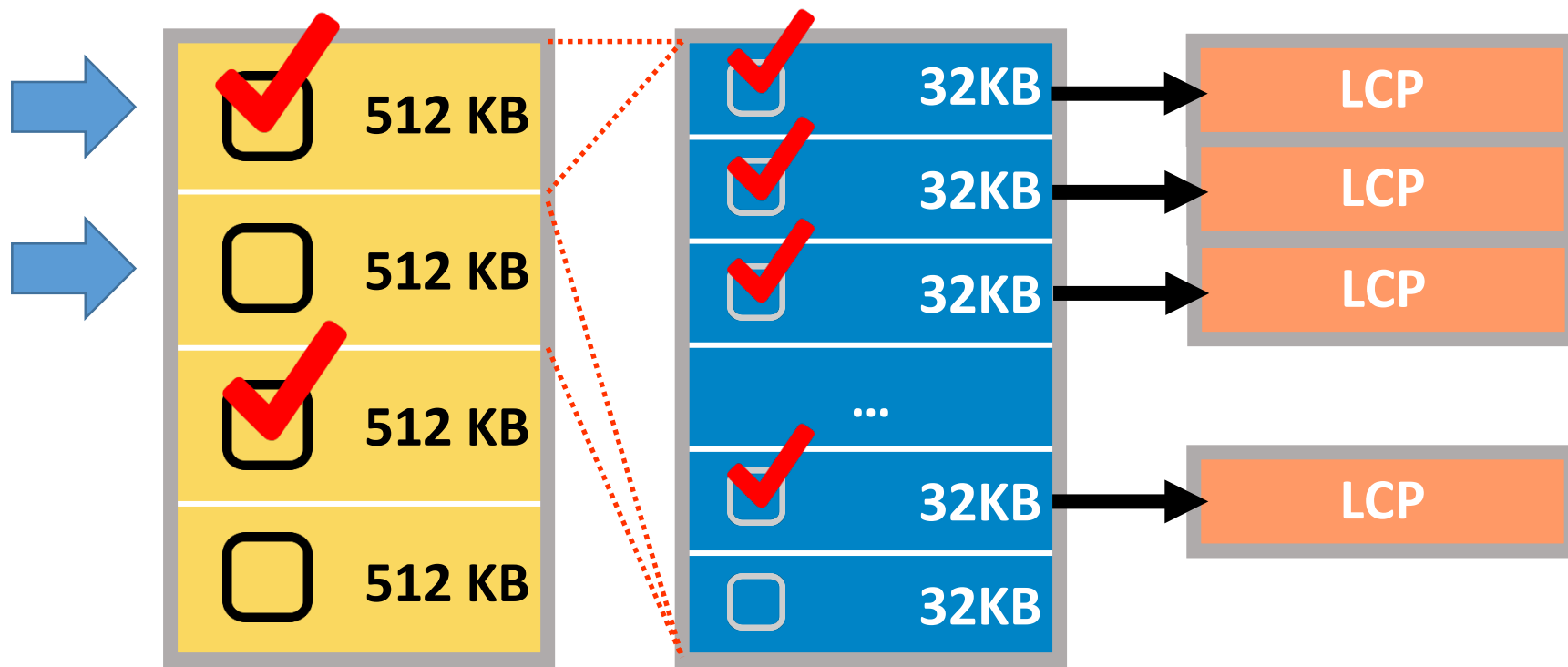
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



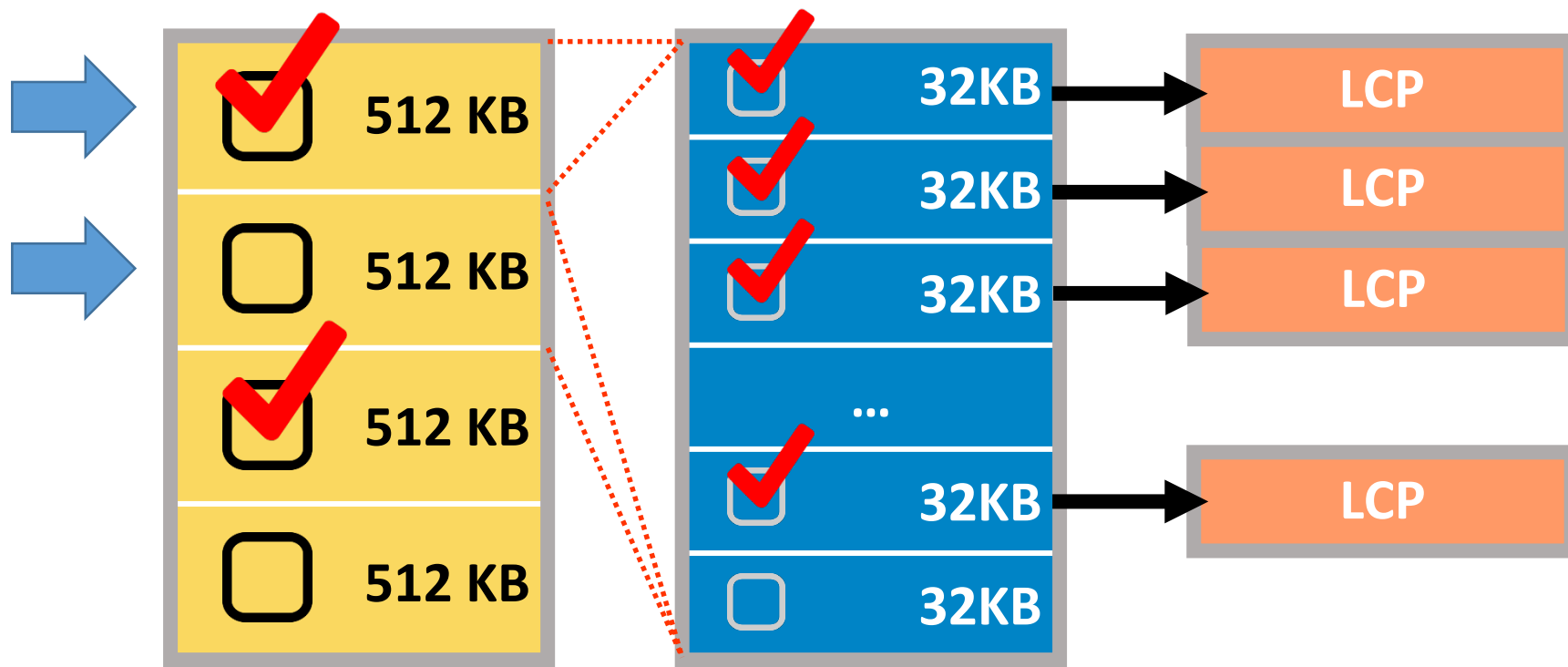
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



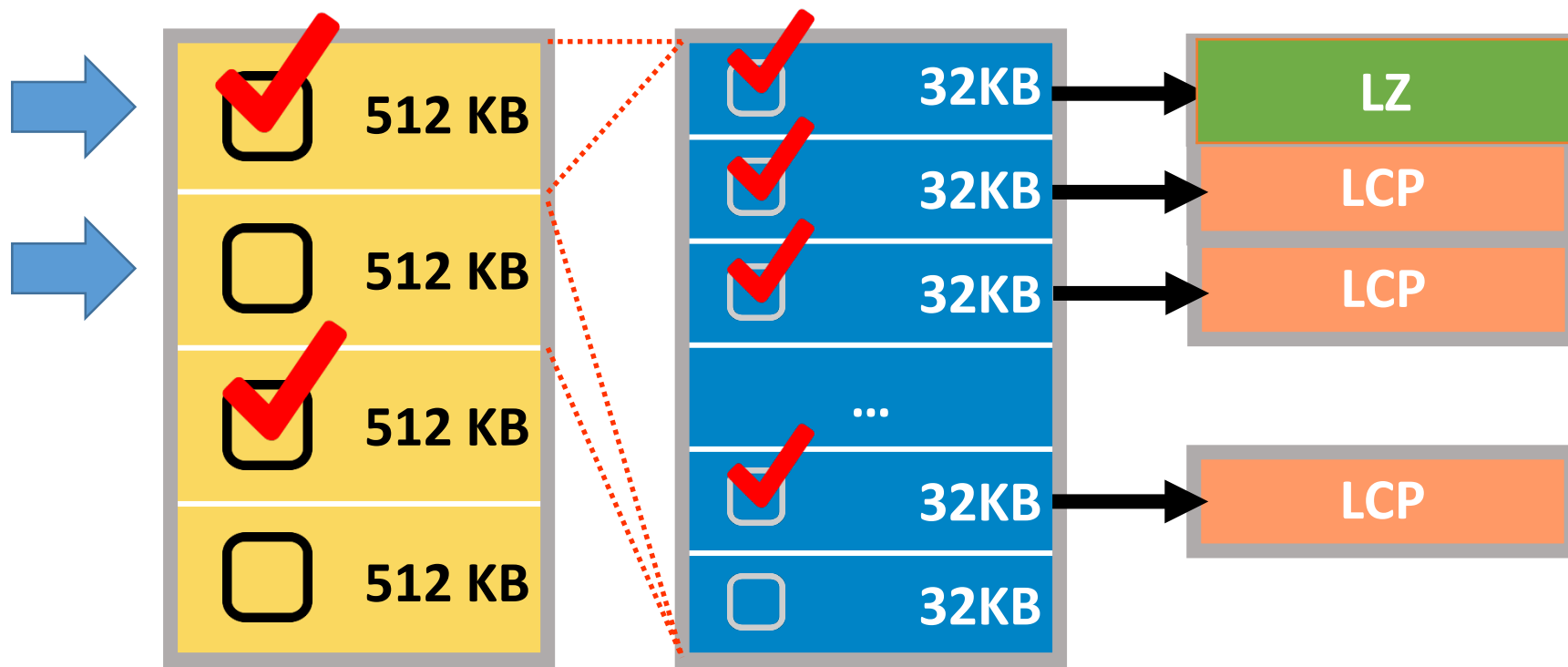
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



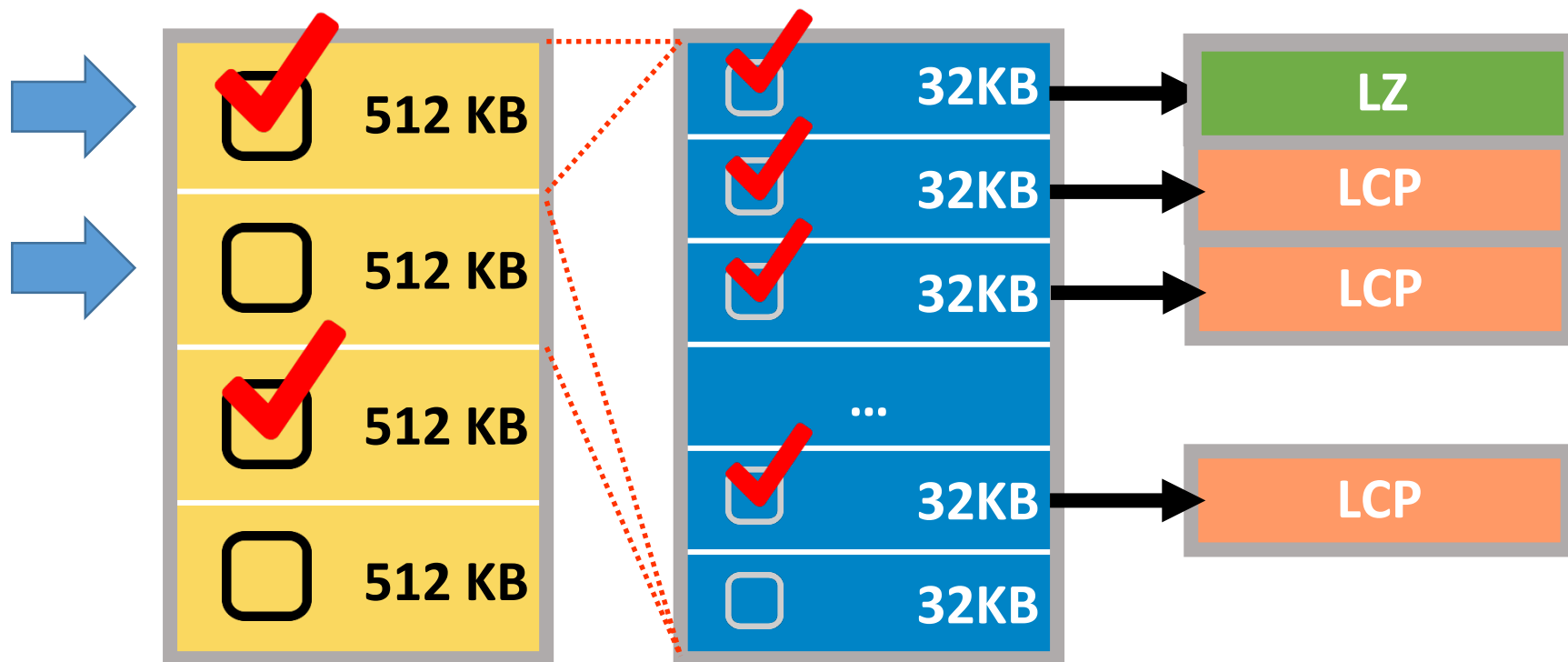
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



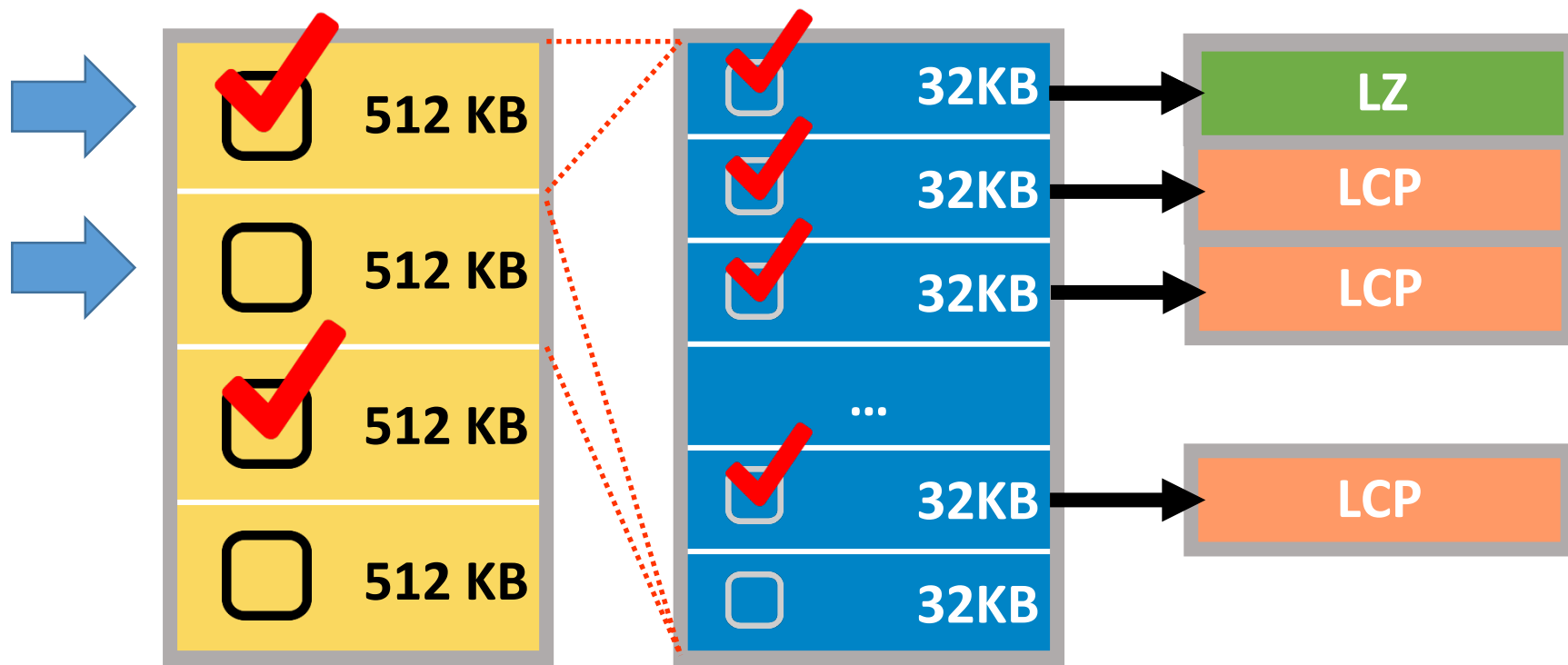
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



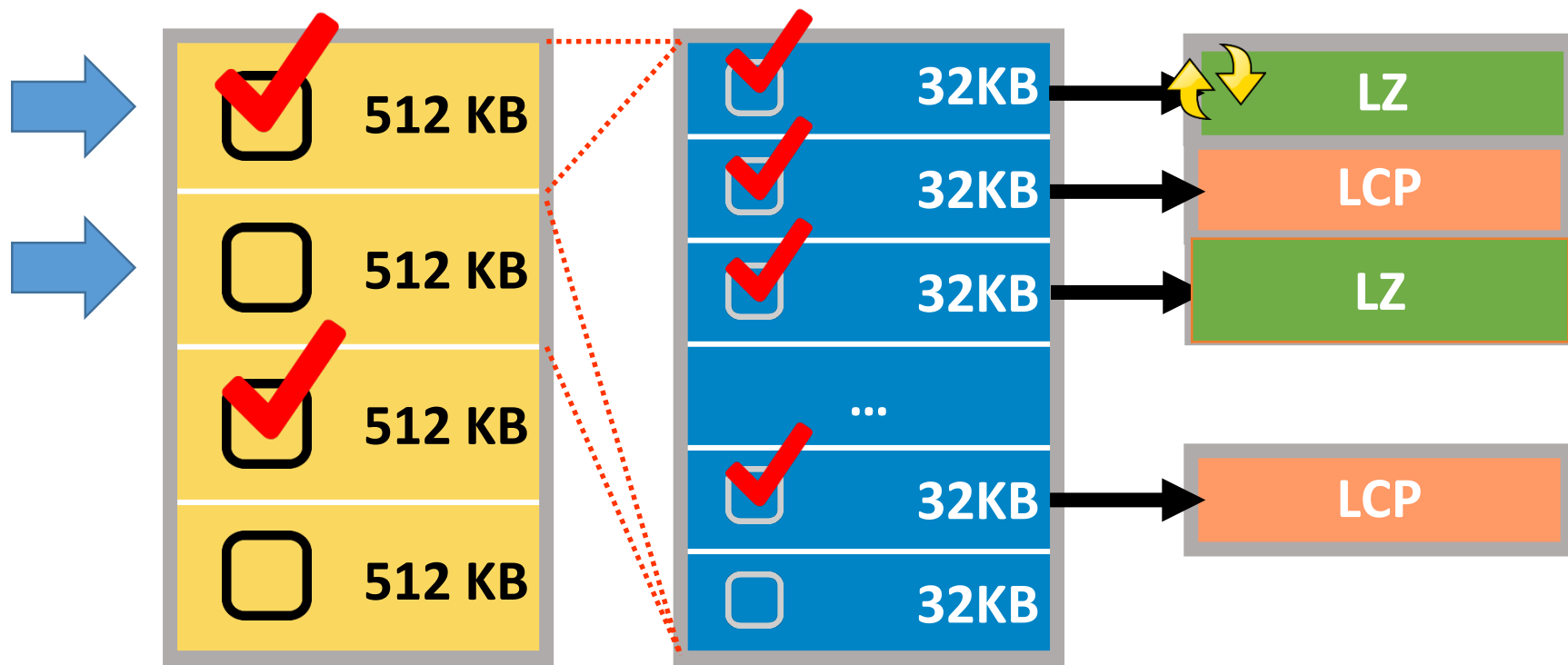
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



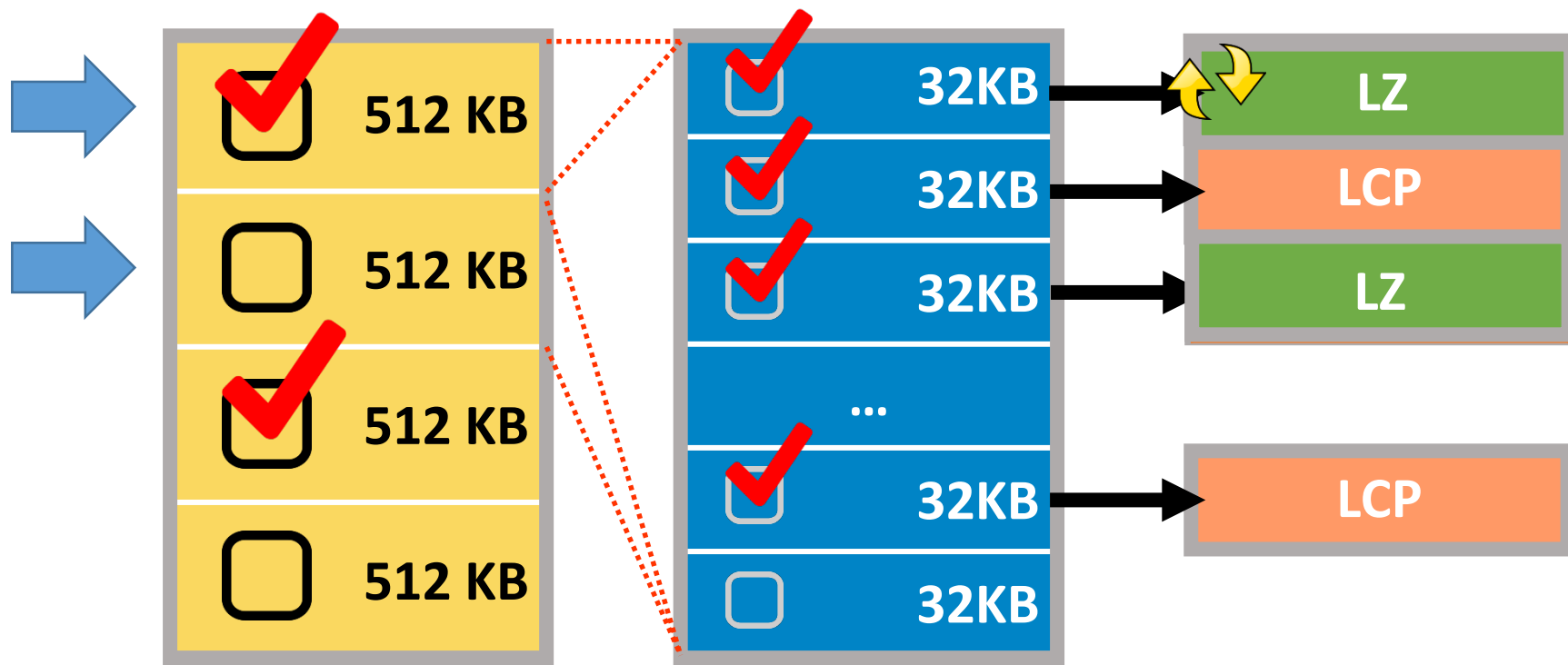
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



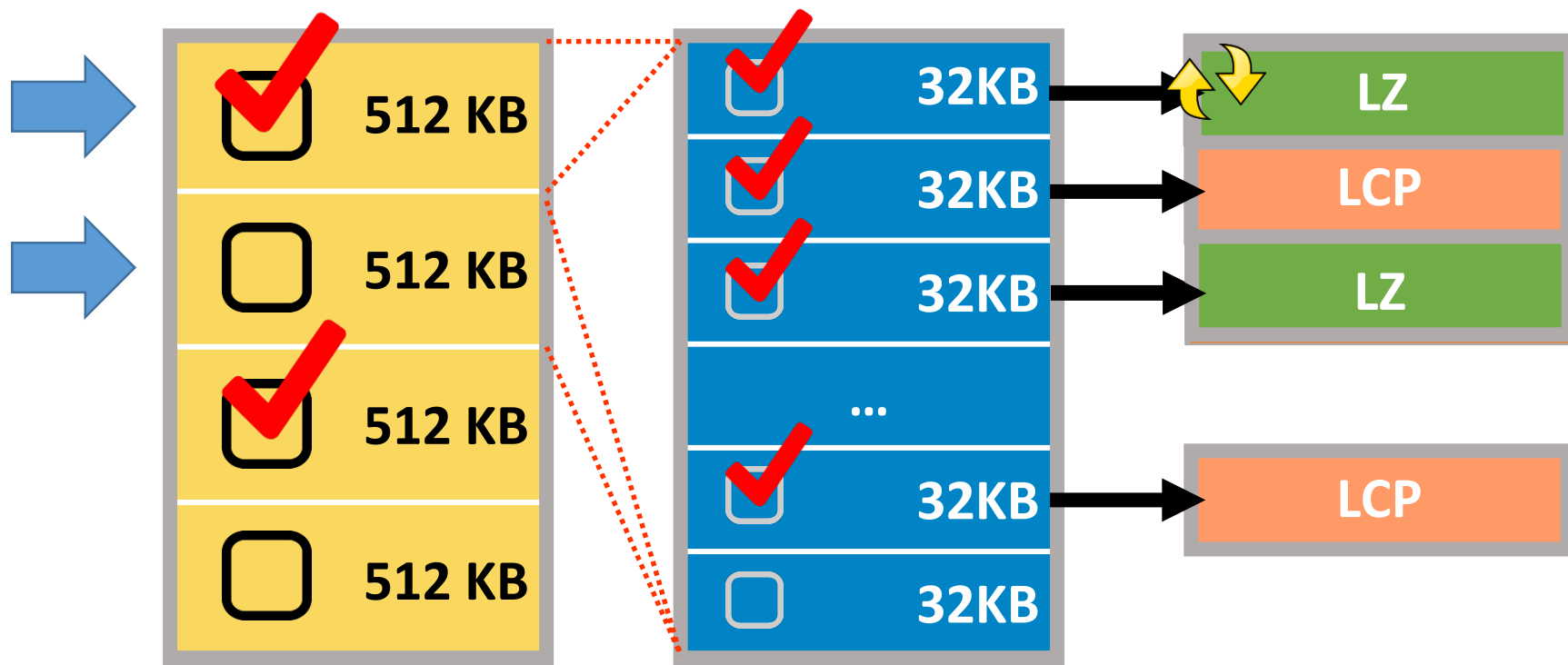
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



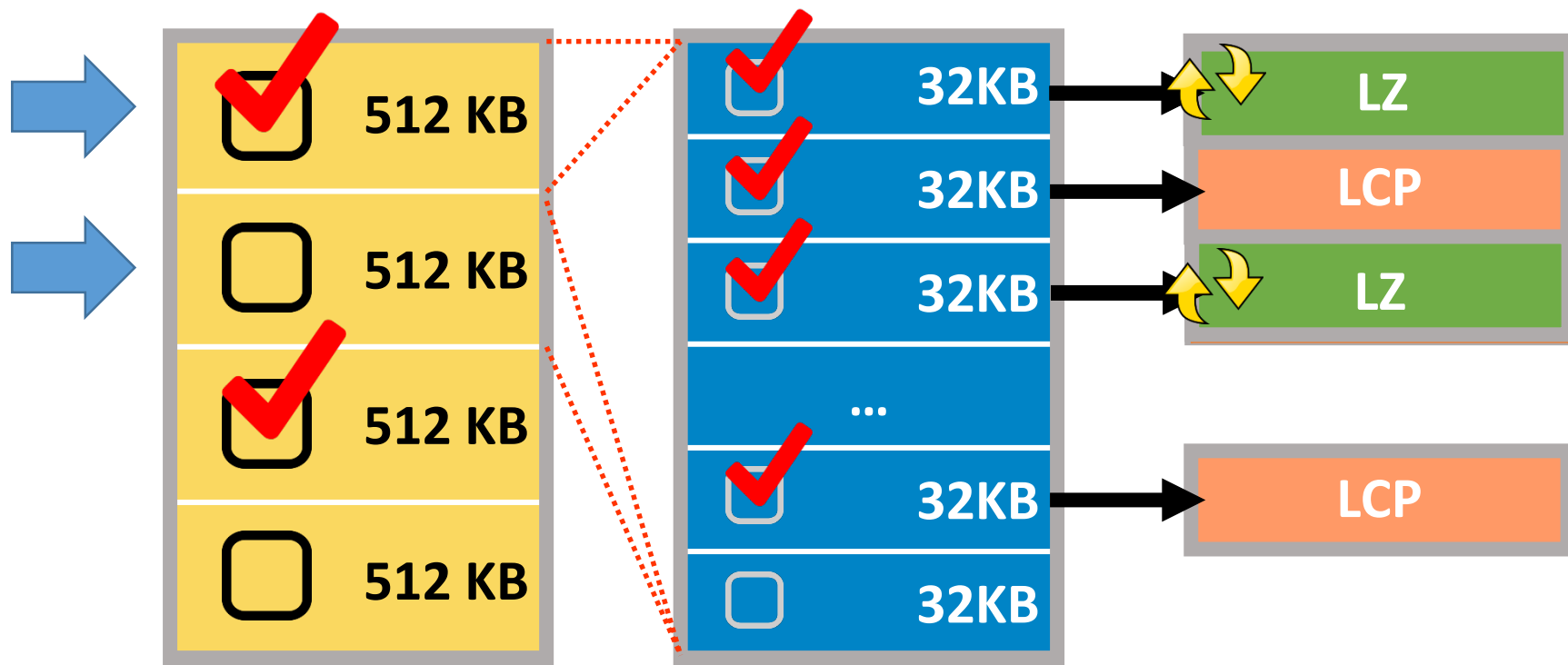
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



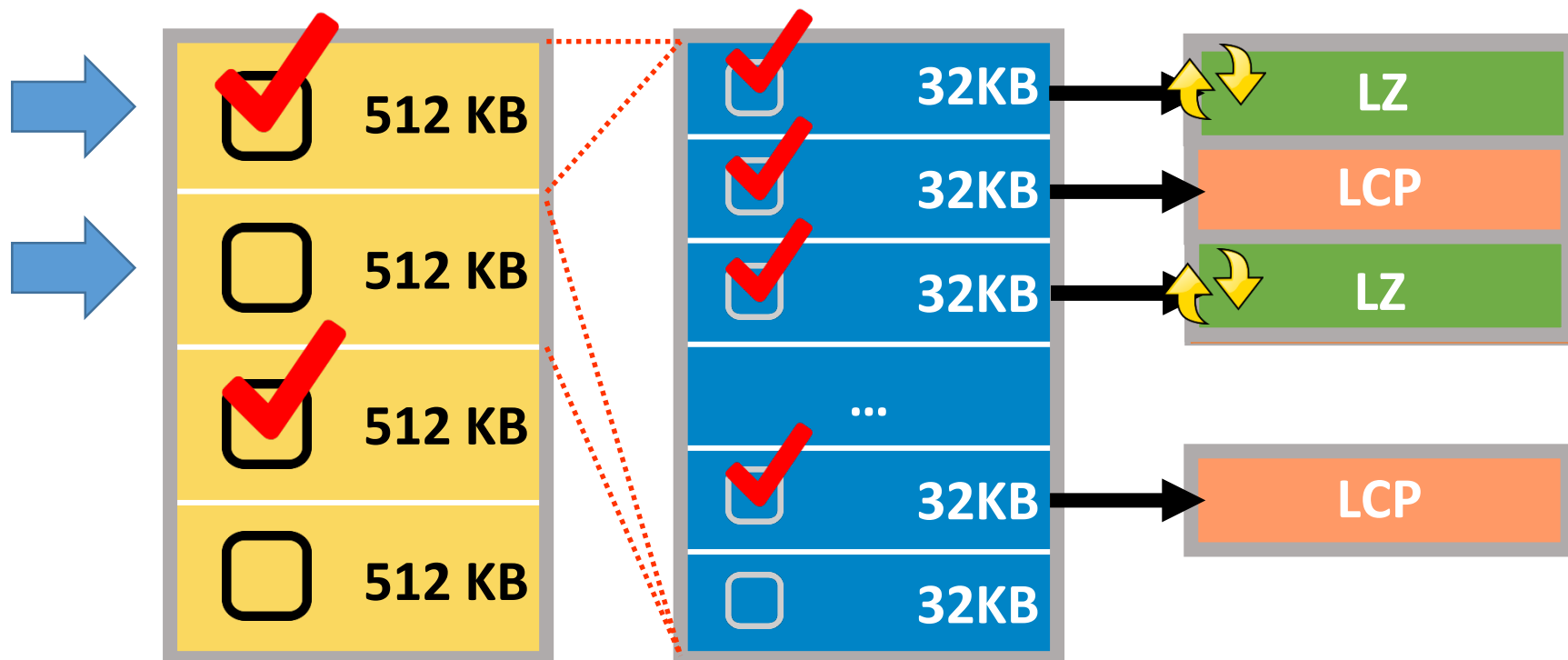
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



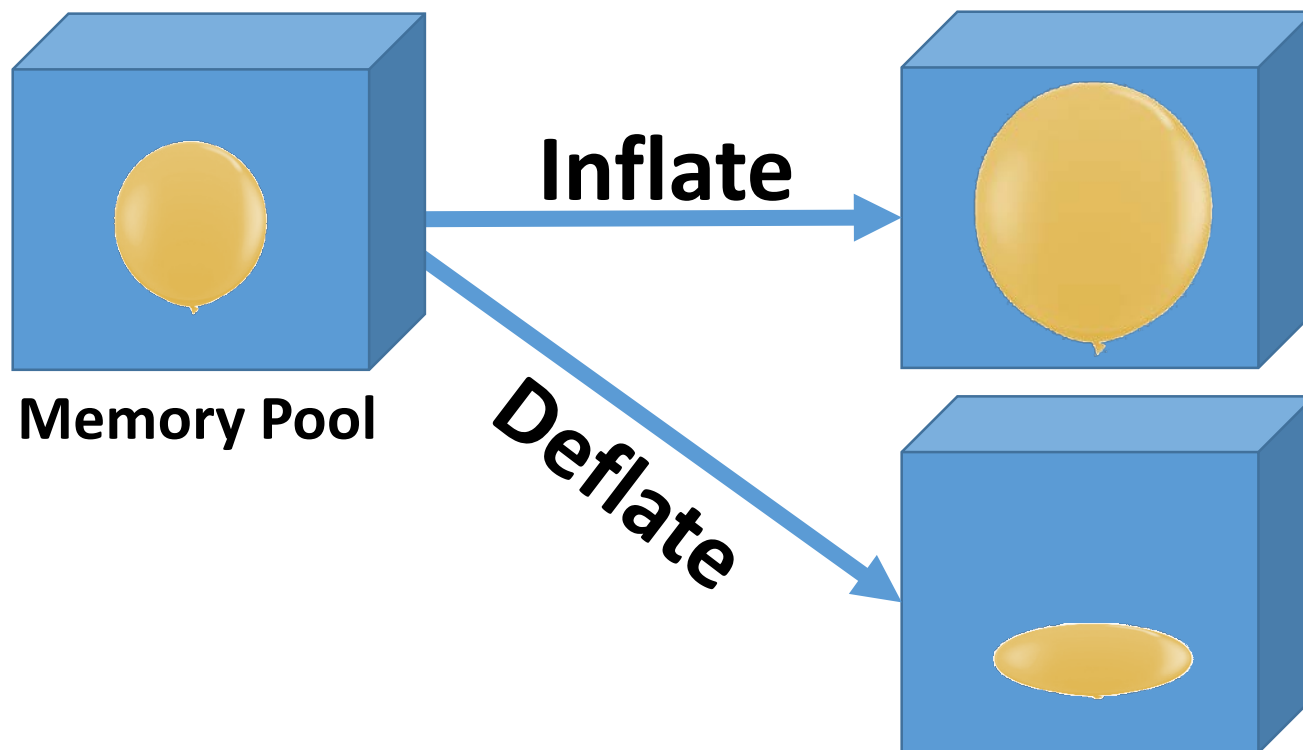
Cold Region Detection

- Periodic LCP to LZ transcompression
- Two-level region-based cold region search



OS Modifications

- Memory balloon driver
 - Inflate: OS is forced to page out
 - Deflate: OS can page in



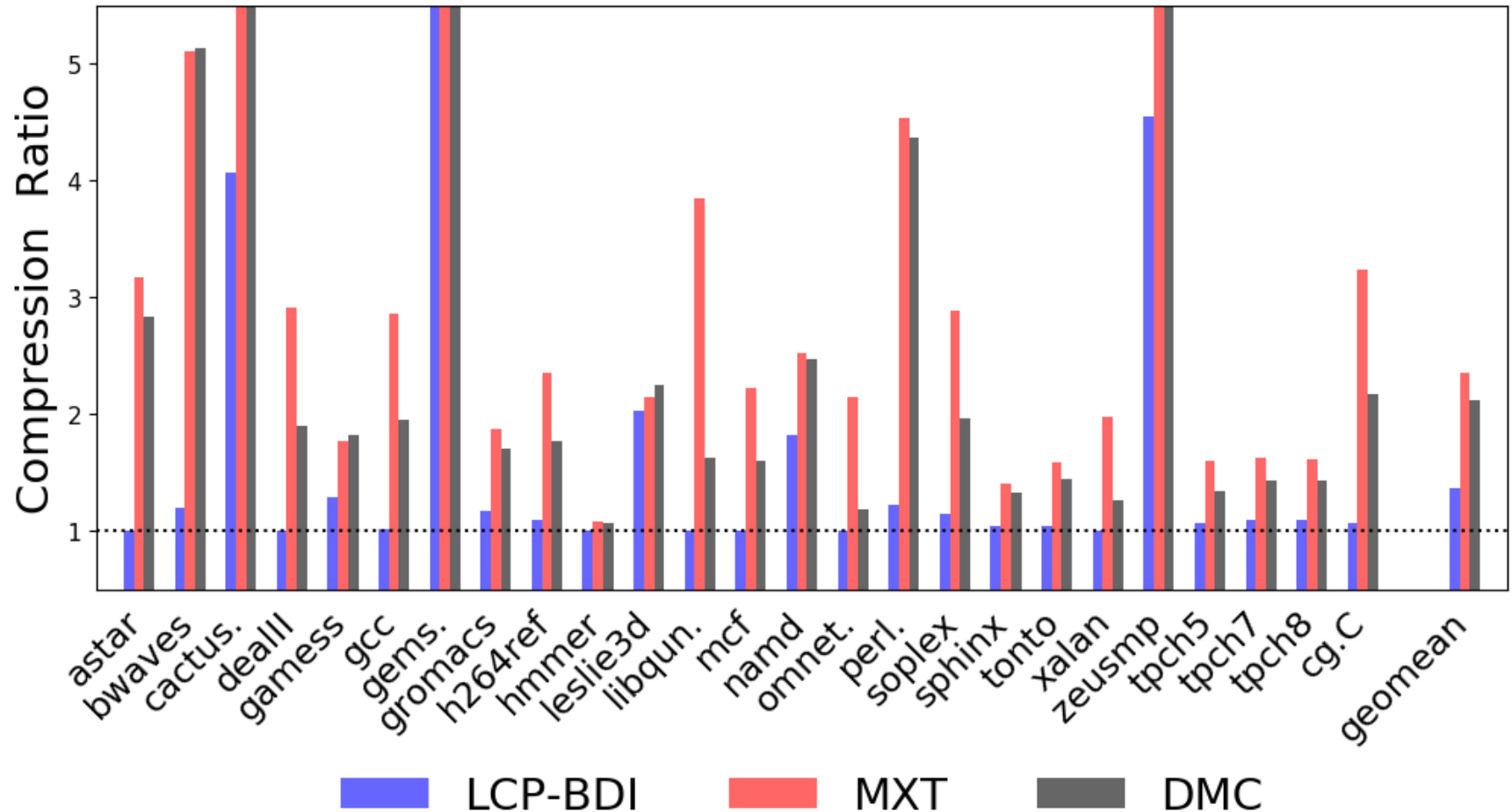
Simulation Configuration

- McSimA+ for core architecture
- GEMS for cache hierarchy

CPU Processor	OoO x86 ISA, 4GHz, 1-4 cores
CPU L1 Cache	32KB I cache, 32KB D cache 64B cacheline, 8 ways
CPU L2 Cache	2MB, 64B cacheline, 32 ways
Epoch Length	50 Million cycles
Transcompression Limit	2400
LCP-BDI Config	1 cycle decomp, Translation cache(X)
MXT Config	64 cycle decomp, Translation cache(0)

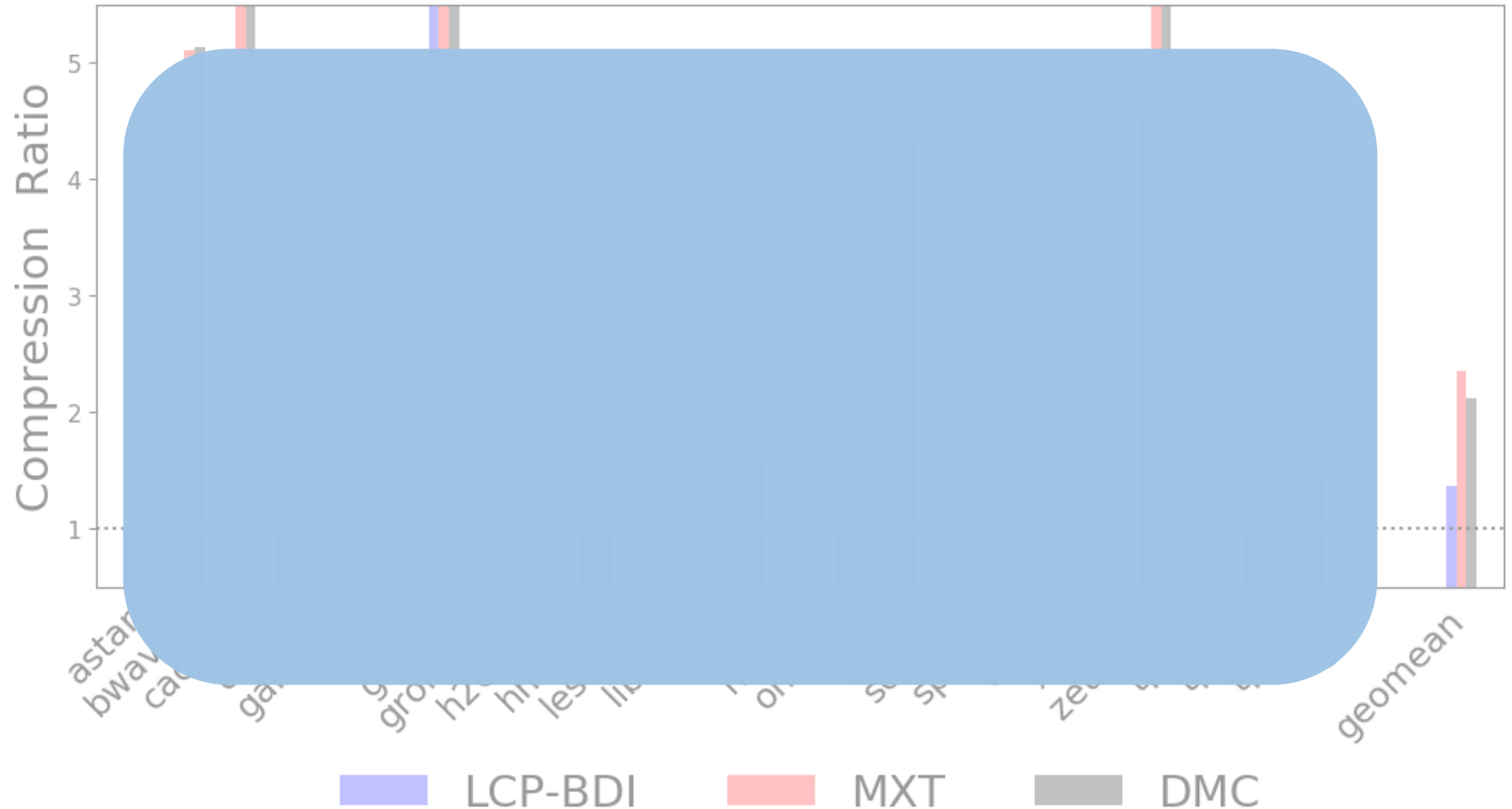
Evaluation: Compression Ratio

➤ Single-core evaluation



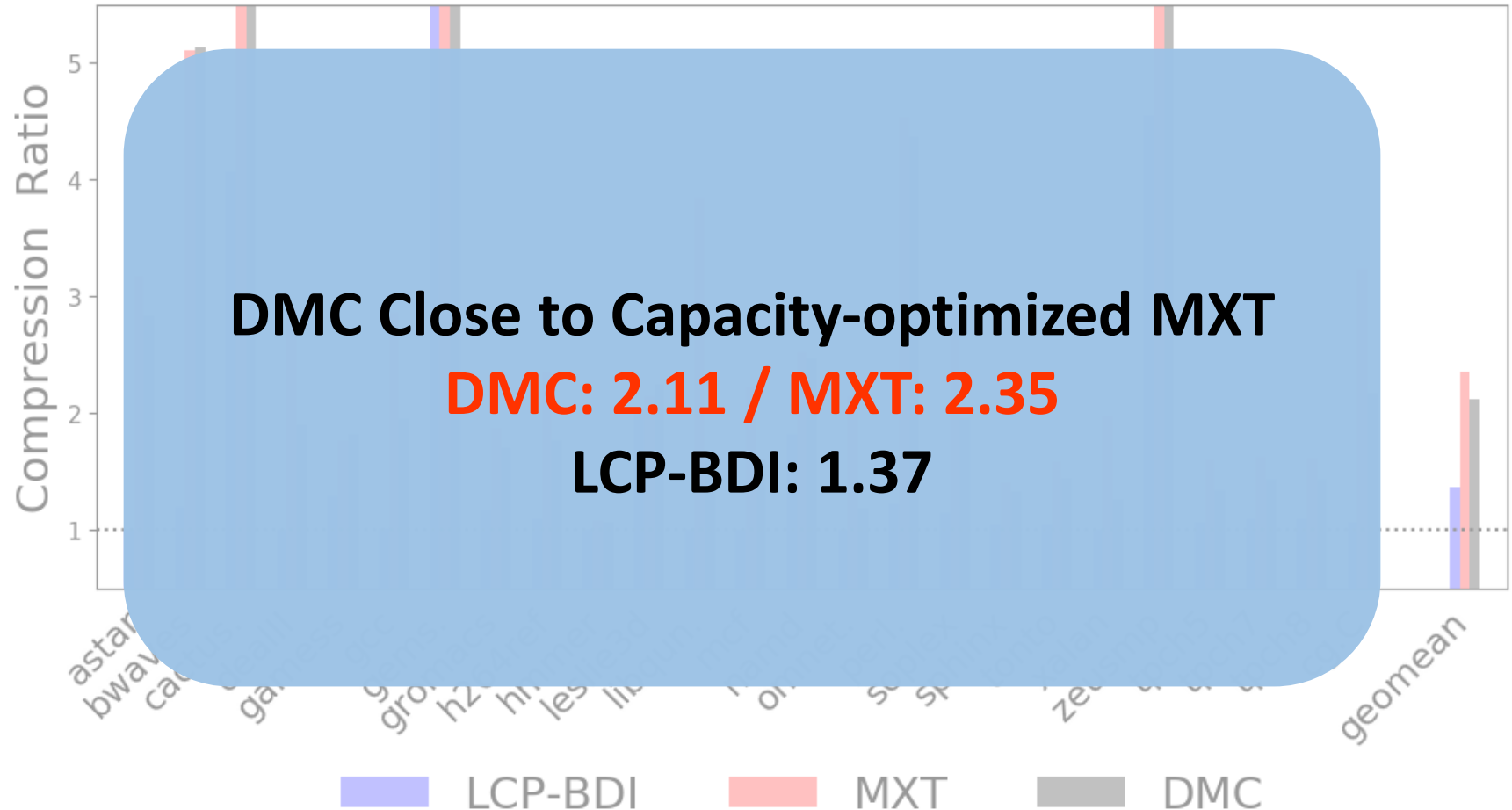
Evaluation: Compression Ratio

➤ Single-core evaluation



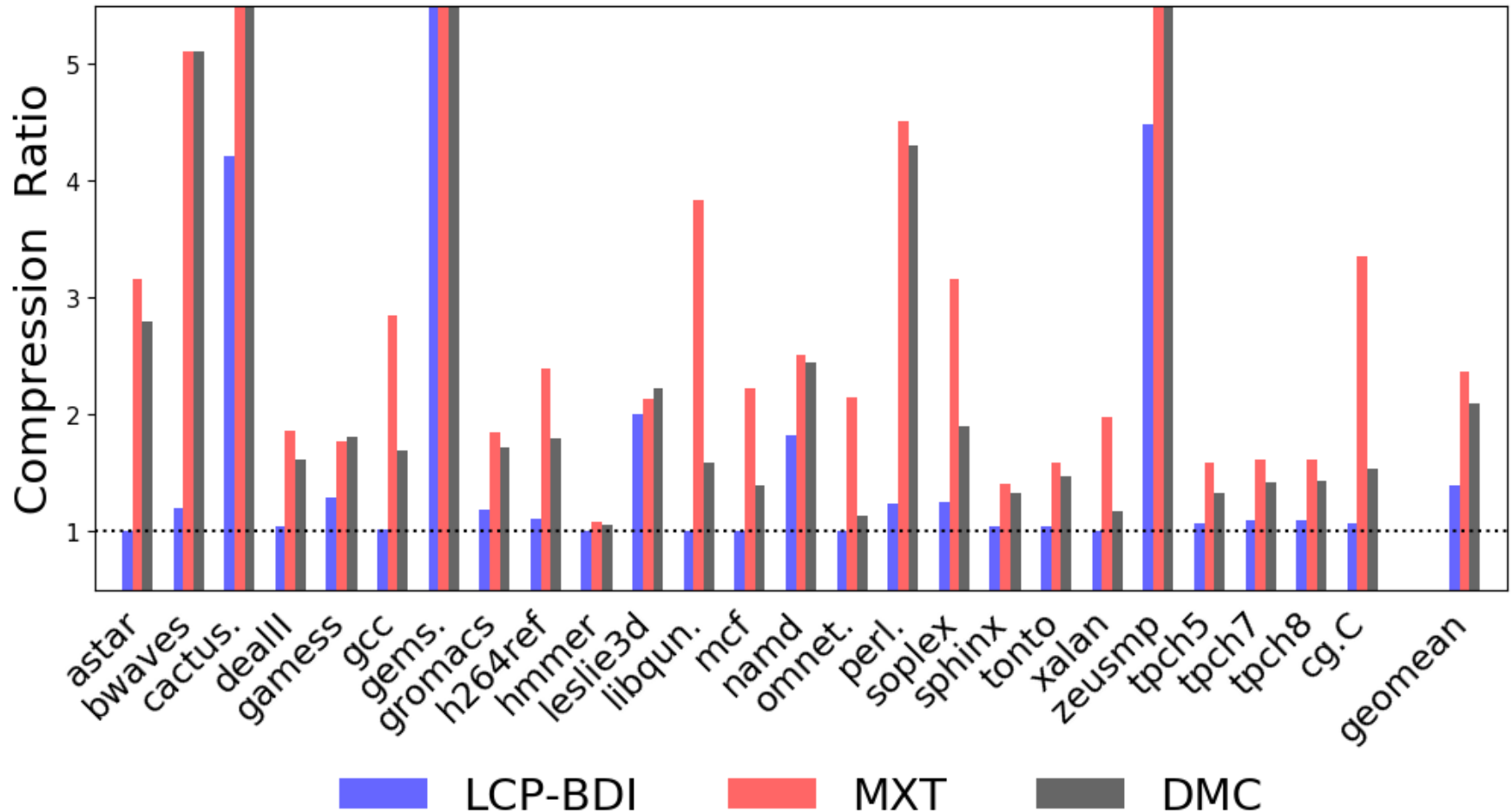
Evaluation: Compression Ratio

➤ Single-core evaluation



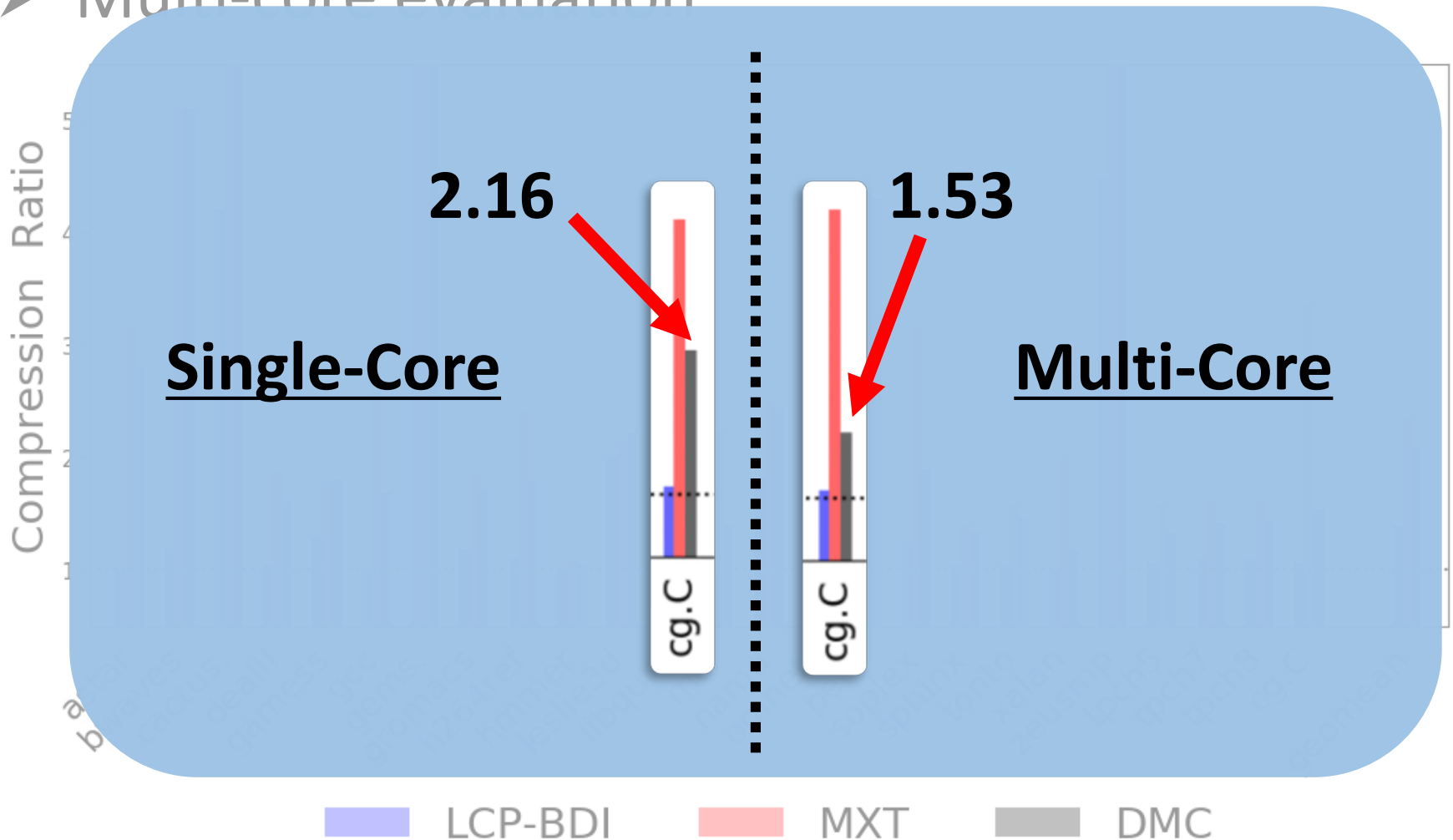
Evaluation: Compression Ratio

➤ Multi-core evaluation



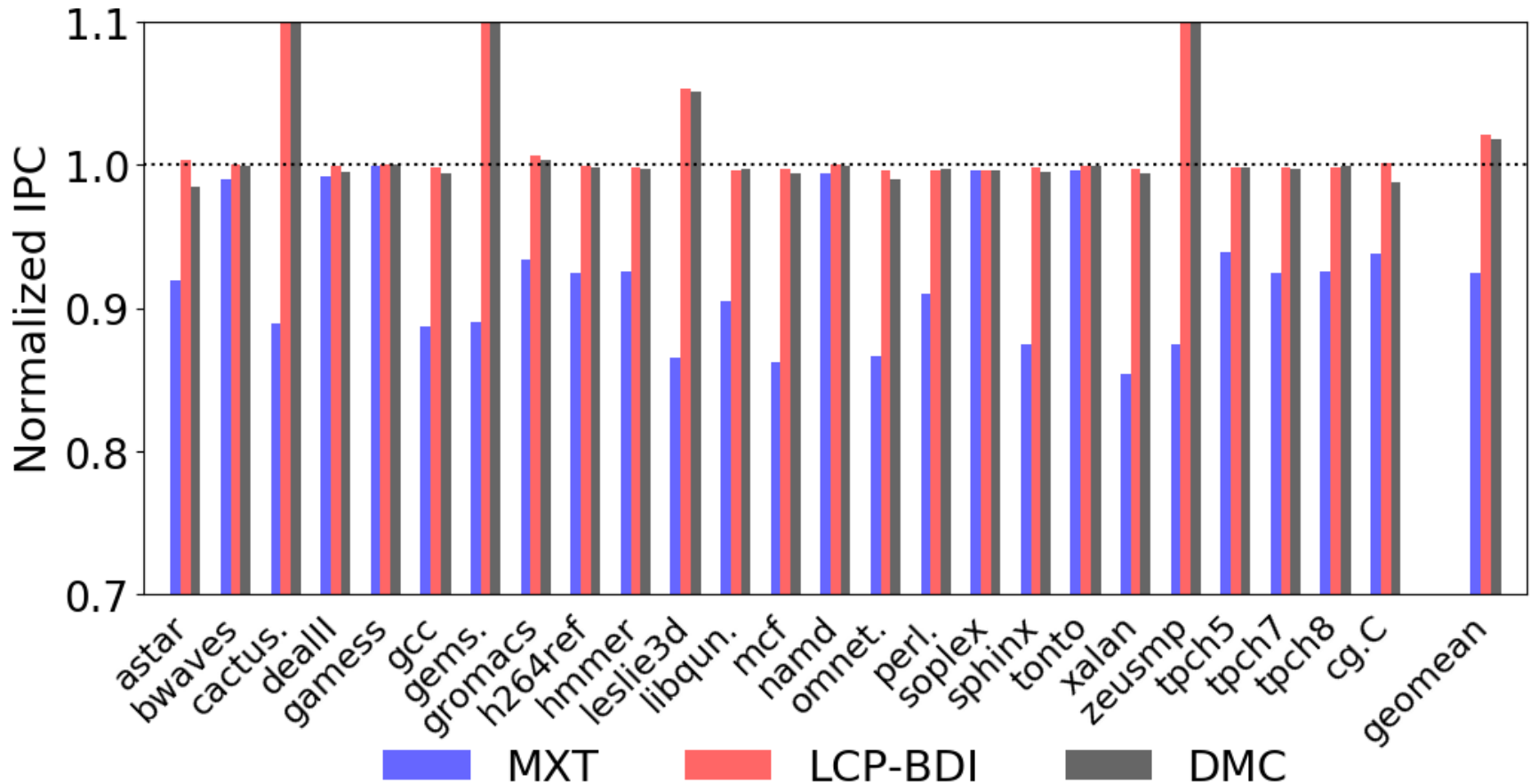
Evaluation: Compression Ratio

➤ Multi-core evaluation



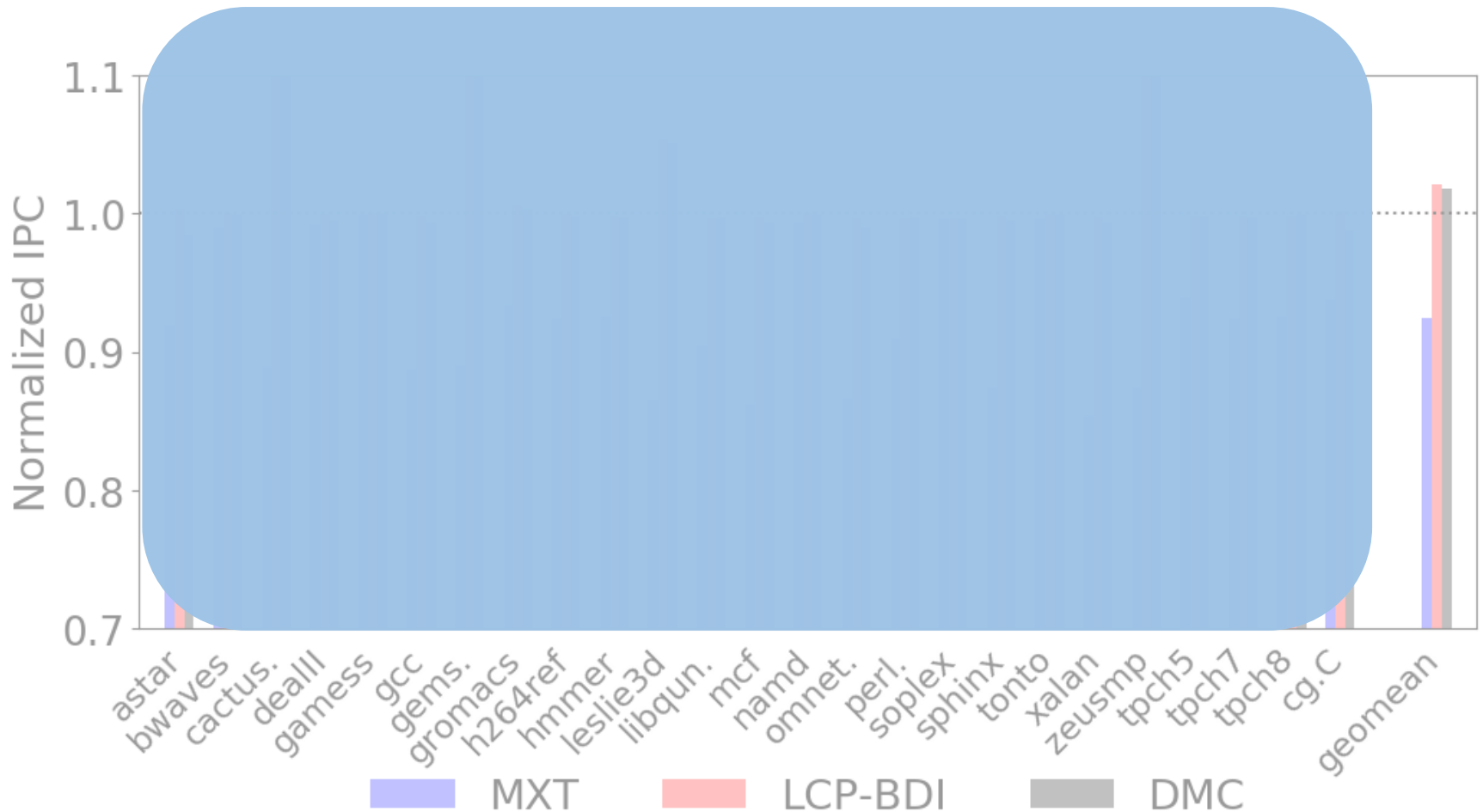
Evaluation: Performance

➤ Multi-core evaluation



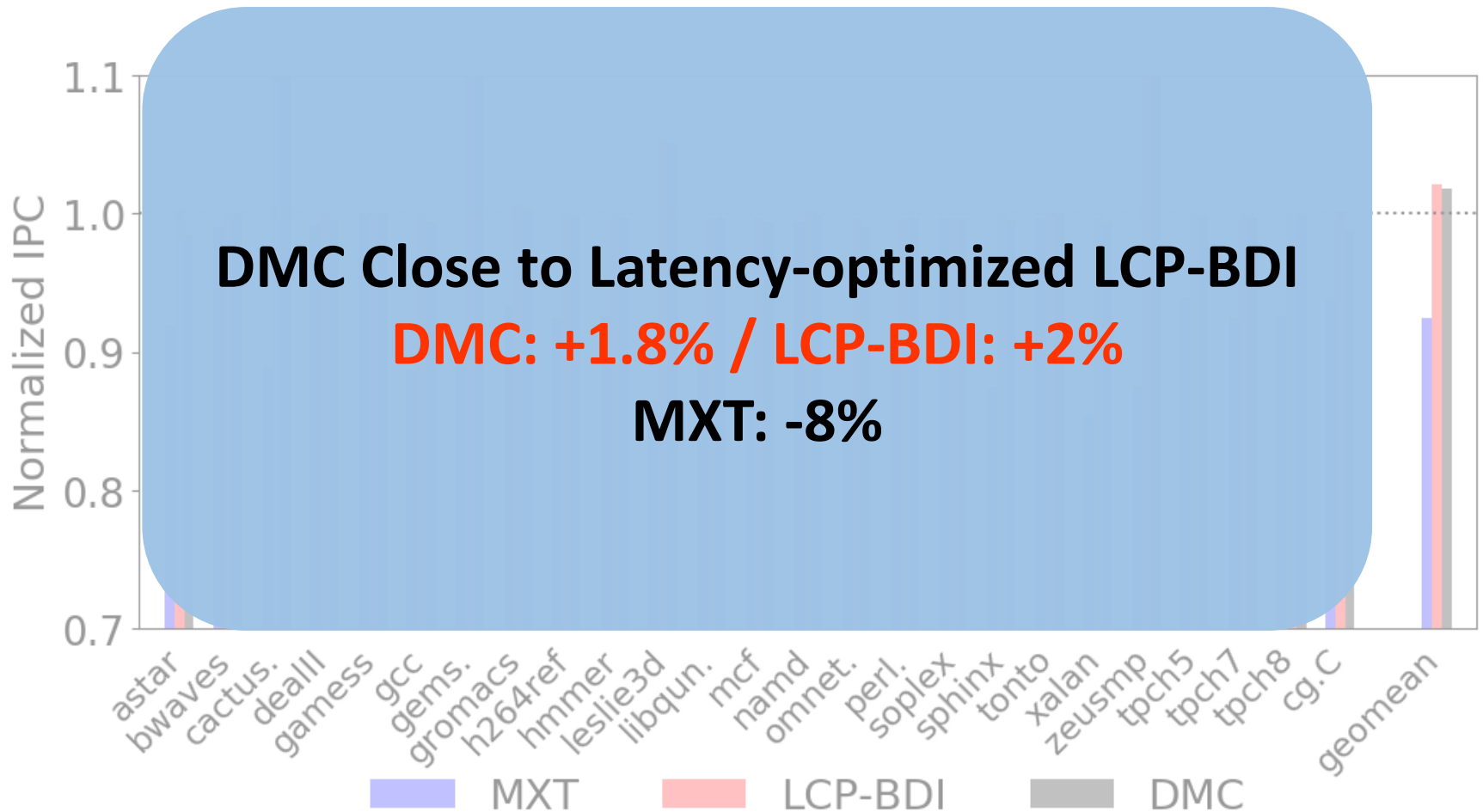
Evaluation: Performance

➤ Multi-core evaluation



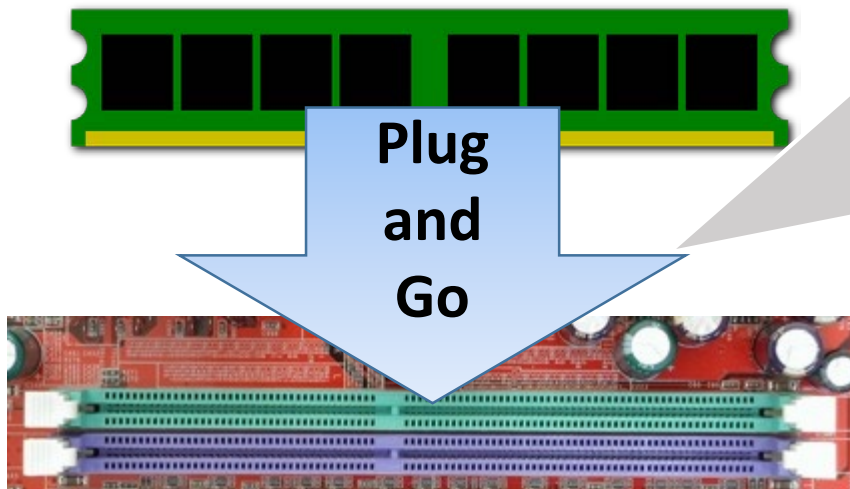
Evaluation: Performance

➤ Multi-core evaluation



Summary and Future Work

- ✓ DMC performs:
 - ✓ Similar to compression ratio of the MXT
 - ✓ Similar to IPC of the LCP
- ✓ Future work: Smart self-managed memory



DMC: First step toward *Self-managed memory*

1. Power management
2. Deduplication
3. Data migration

...