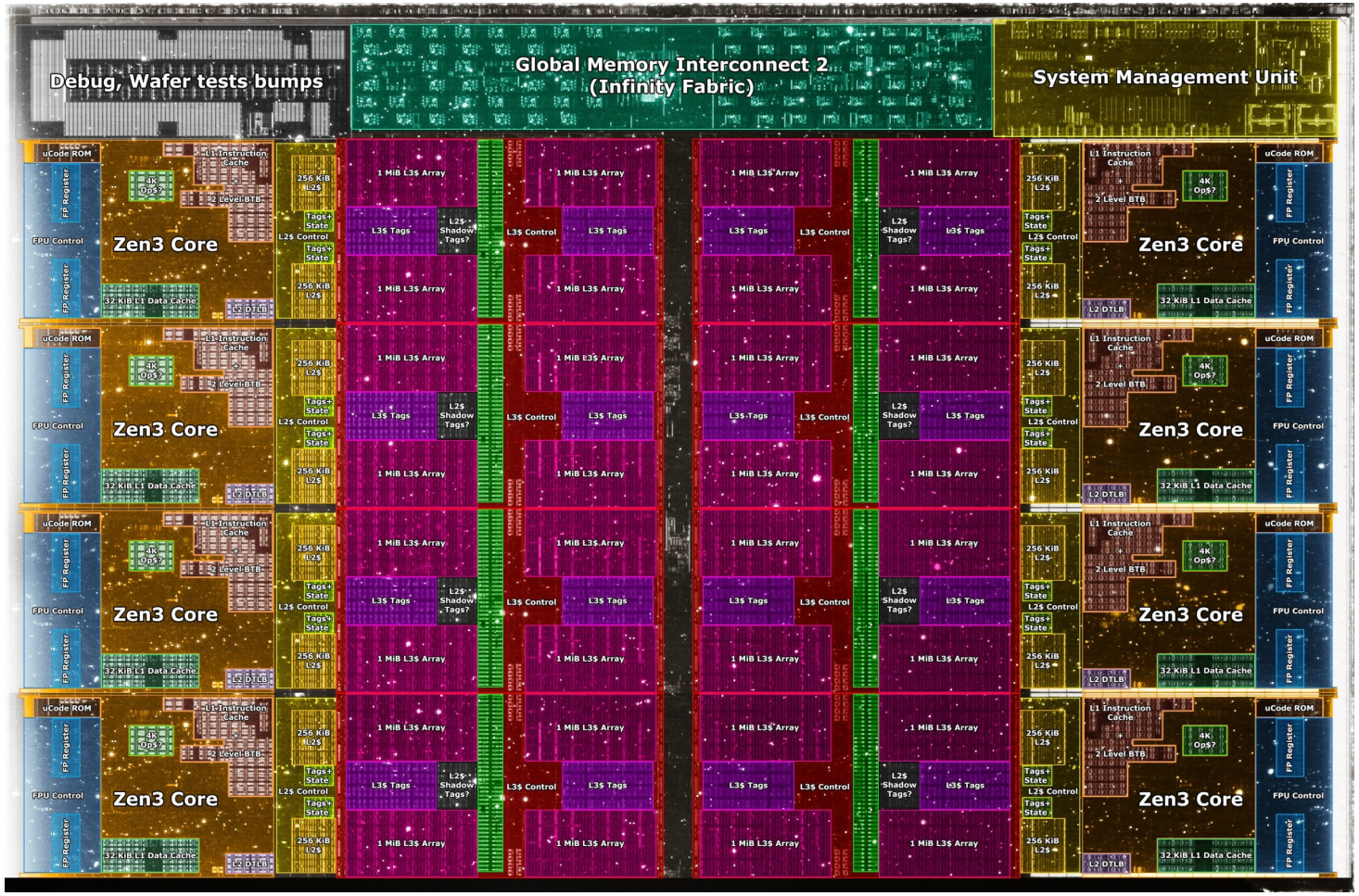
The XOR Cache: A Catalyst for Compression

Zhewen Pan Joshua San Miguel

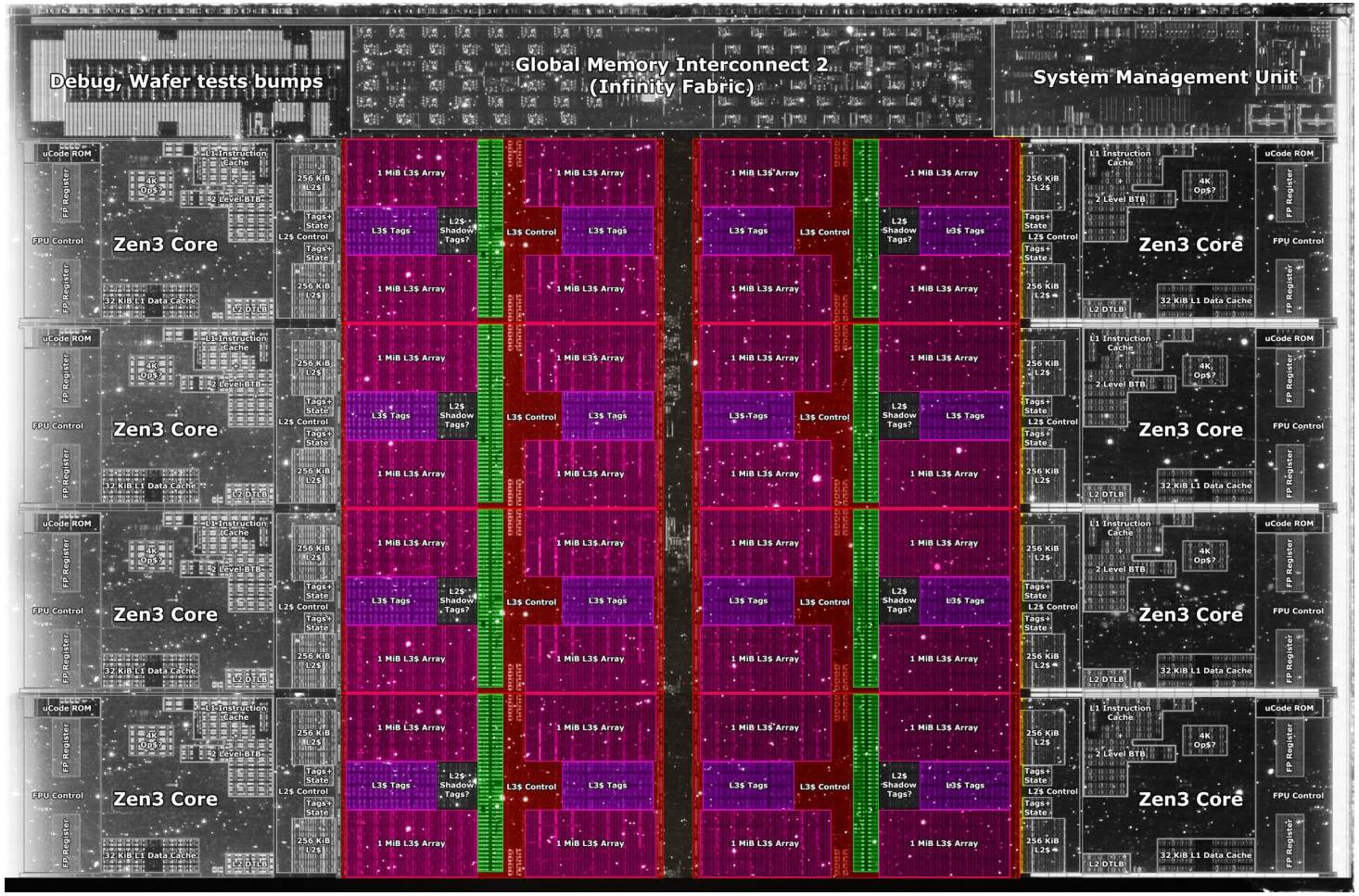
Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Last-level cache (LLC) in today's systems



Mujtaba, H. (2020, November 9). AMD Ryzen 5000 Zen 3 "Vermeer" Undressed, First Ever High-Res Die Shots Close Ups Pictured & Detailed. https://wccftech.com/amd-ryzen-5000-zen-3-vermeer-undressed-high-res-die-shots-close-ups-pictured-detailed/

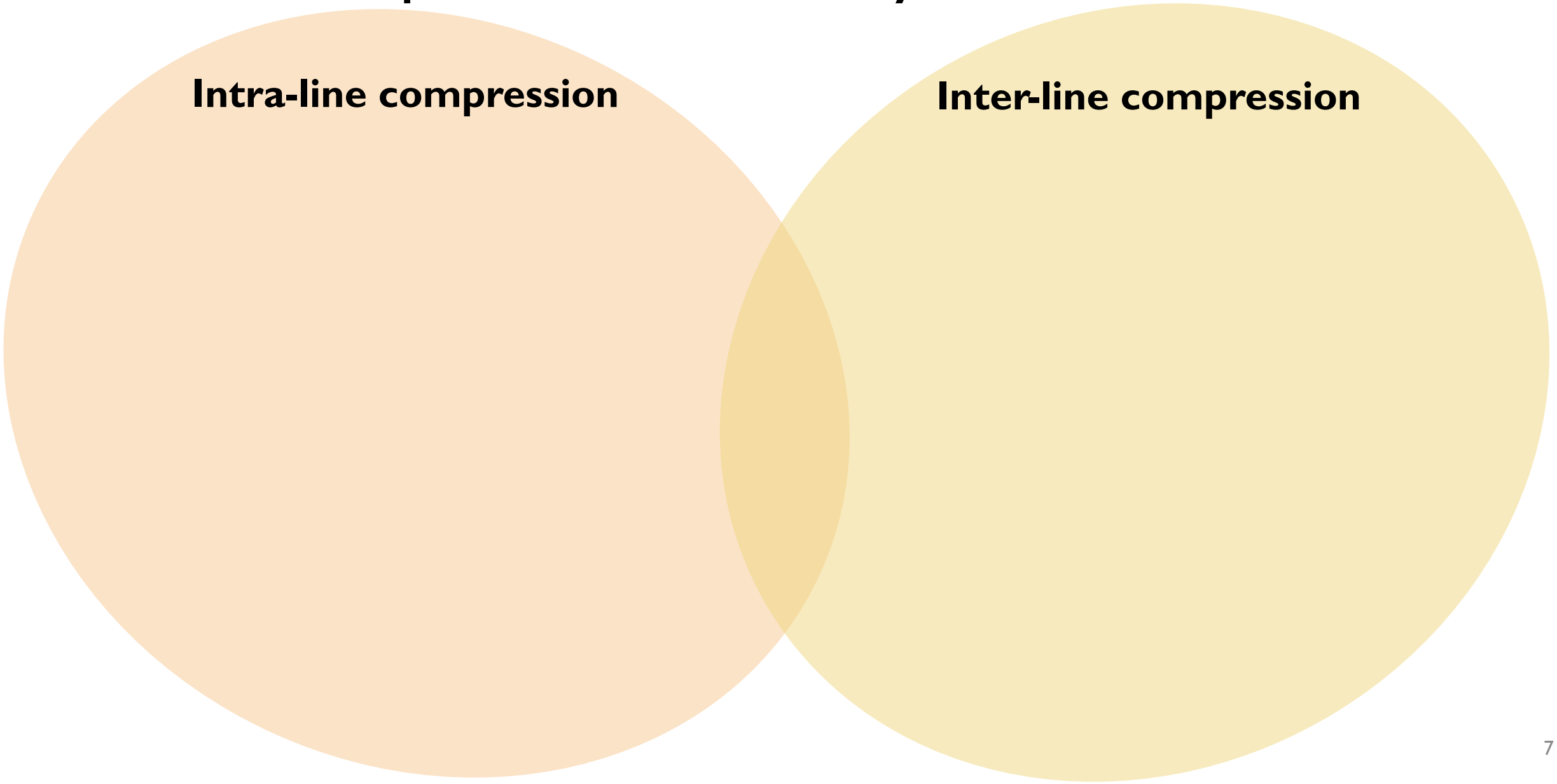
Last-level cache (LLC) in today's systems



32 MB L3 Cache
~40% die area

Mujtaba, H. (2020, November 9). AMD Ryzen 5000 Zen 3 "Vermeer" Undressed, First Ever High-Res Die Shots Close Ups Pictured & Detailed. https://wccftech.com/amd-ryzen-5000-zen-3-vermeer-undressed-high-res-die-shots-close-ups-pictured-detailed/

Cache compression taxonomy



Cache compression taxonomy

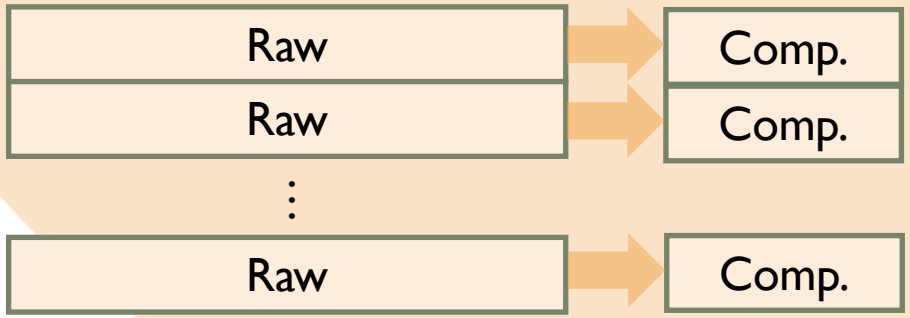
Intra-line compression

BDI [PACT 2012]

Bit-plane [ISCA 2016]

FPC [Tech. Report 2004]

Hycomp [MICRO 2015]



Inter-line compression

Cache compression taxonomy

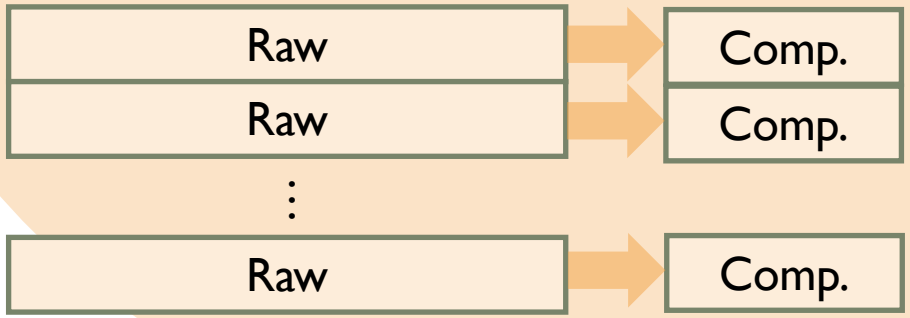
Intra-line compression

BDI [PACT 2012]

Bit-plane [ISCA 2016]

FPC [Tech. Report 2004]

Hycomp [MICRO 2015]



Inter-line compression

Deduplication [ICS 2014]

Bunker [MICRO 2016]

Doppelganger [MICRO 2015]

EPC [HPCA 2022]

Thesaurus [ASPLOS 2020]

Cache compression taxonomy

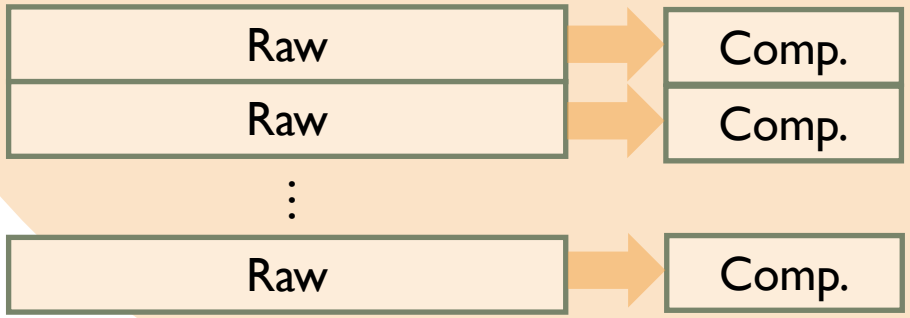
Intra-line compression

BDI [PACT 2012]

Bit-plane [ISCA 2016]

FPC [Tech. Report 2004]

Hycomp [MICRO 2015]



XOR Cache

Inter-line compression

Deduplication [ICS 2014]

Bunker [MICRO 2016]

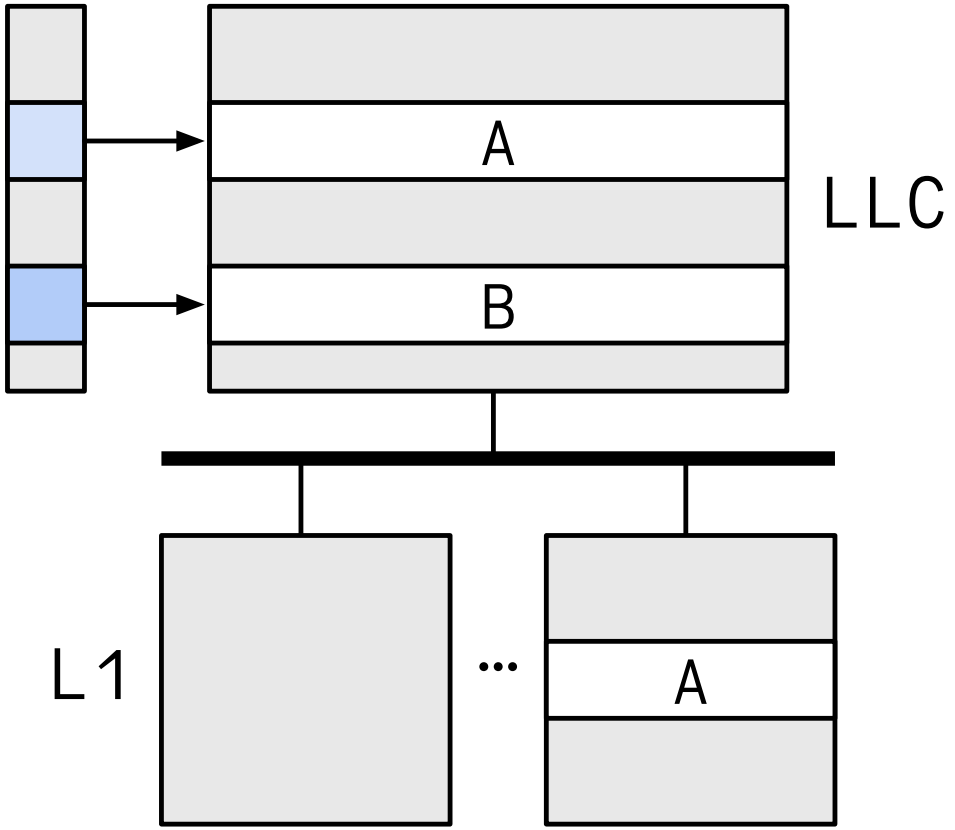
Doppelganger [MICRO 2015]

EPC [HPCA 2022]

Thesaurus [ASPLOS 2020]

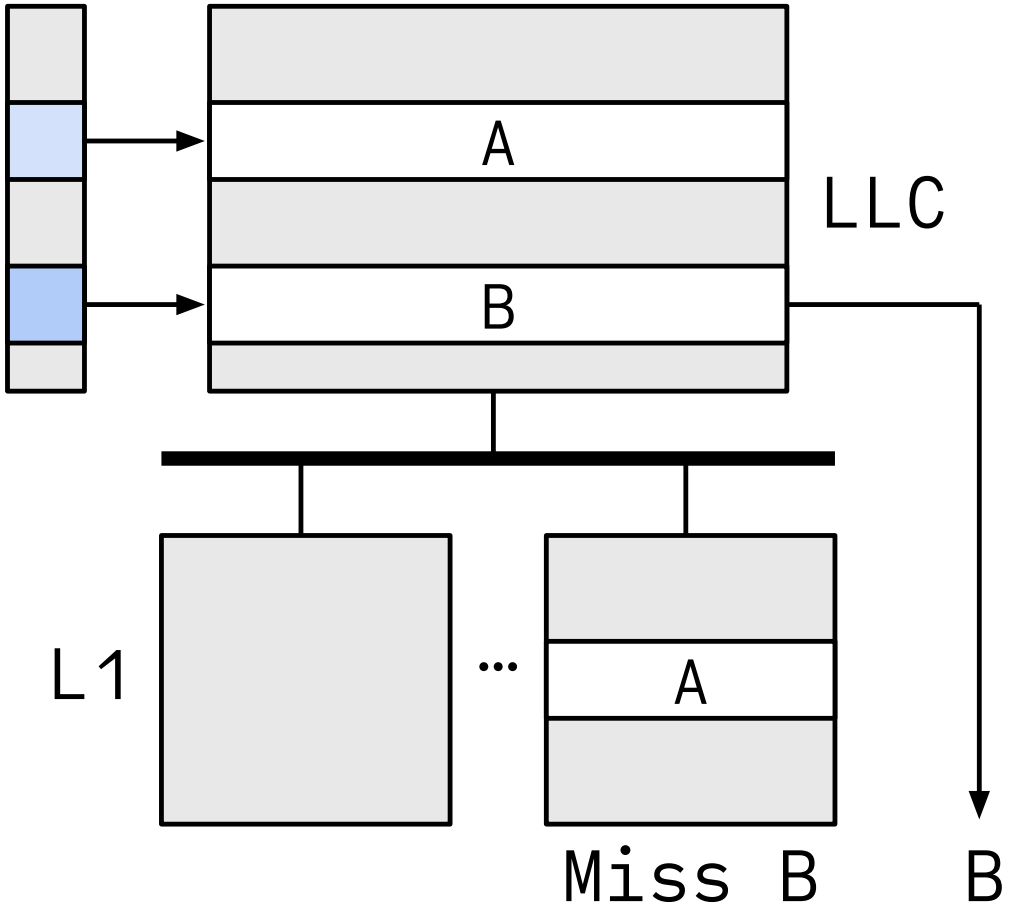
Conventional cache

Insert A

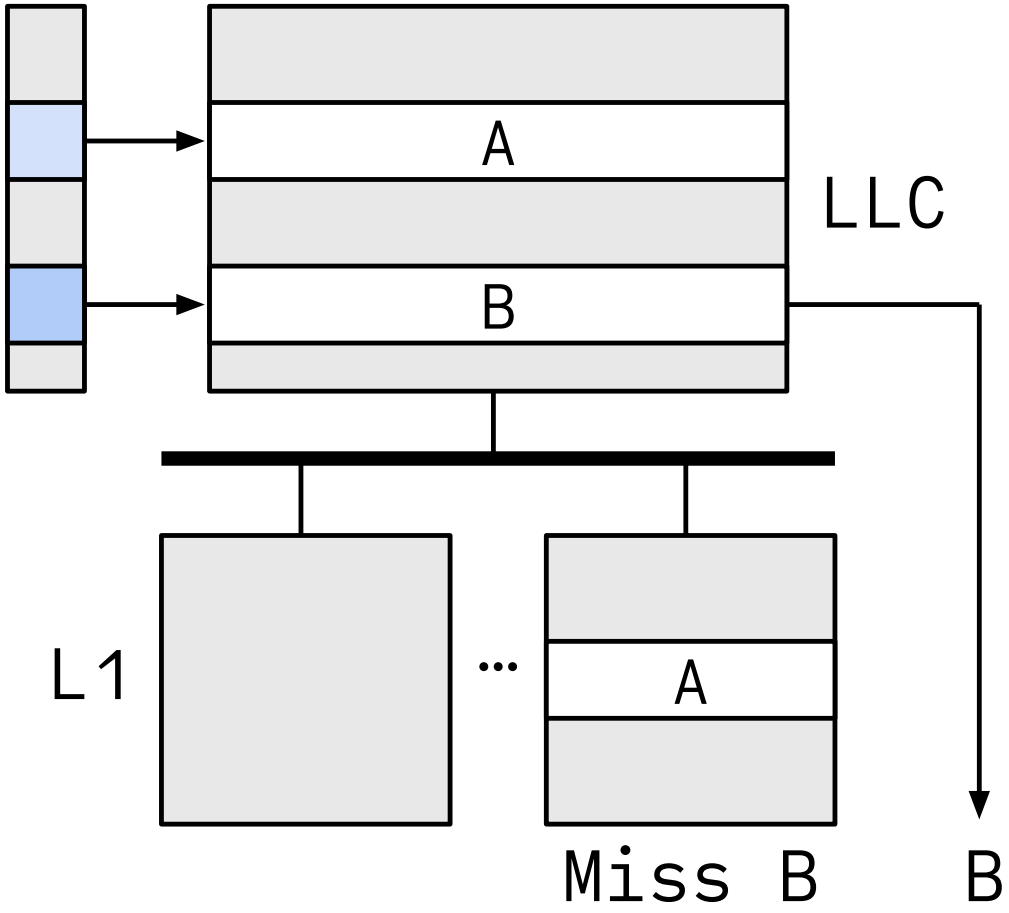


Conventional cache

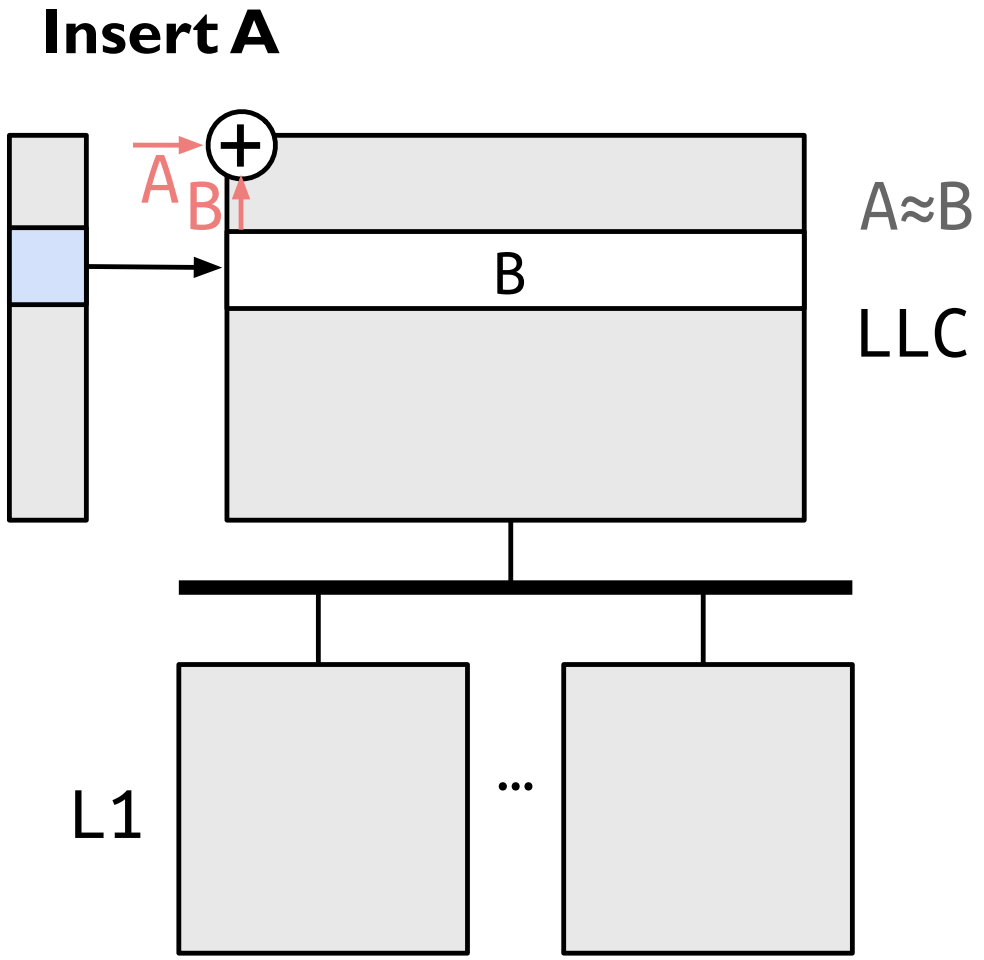
Miss B



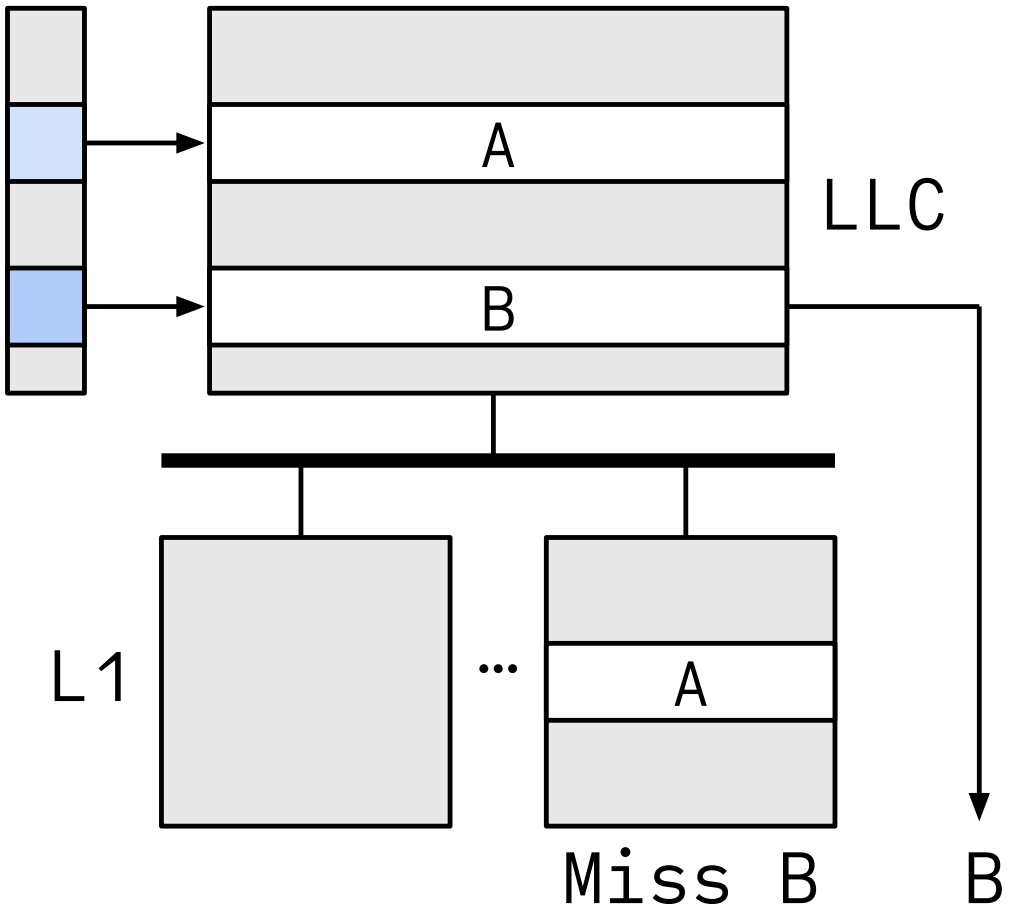
Conventional cache



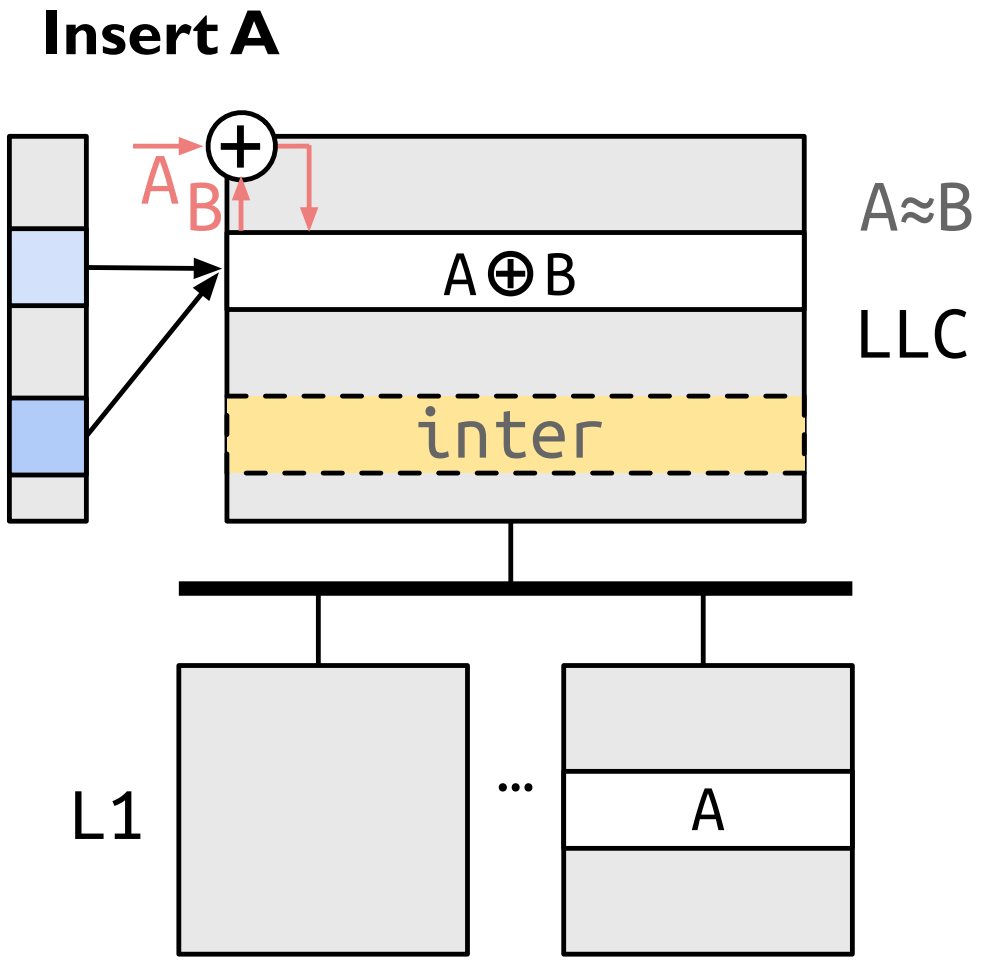
XOR Cache



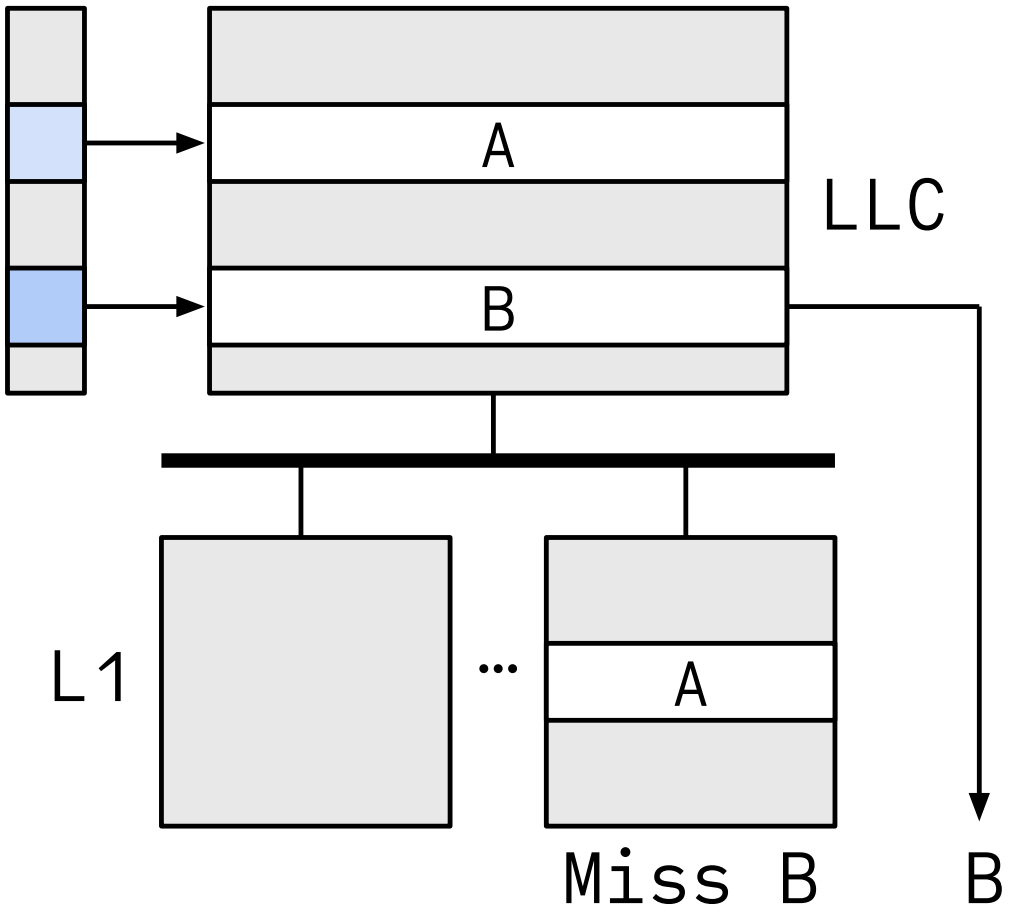
Conventional cache



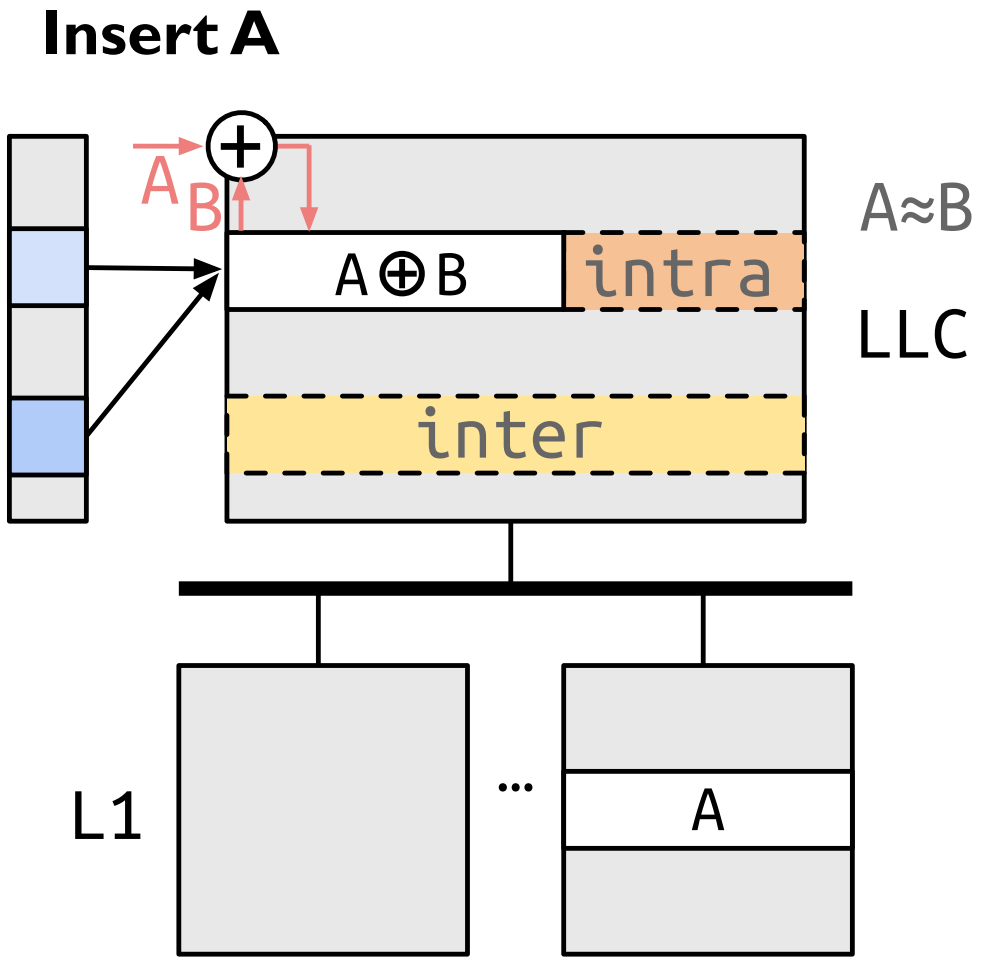
XOR Cache



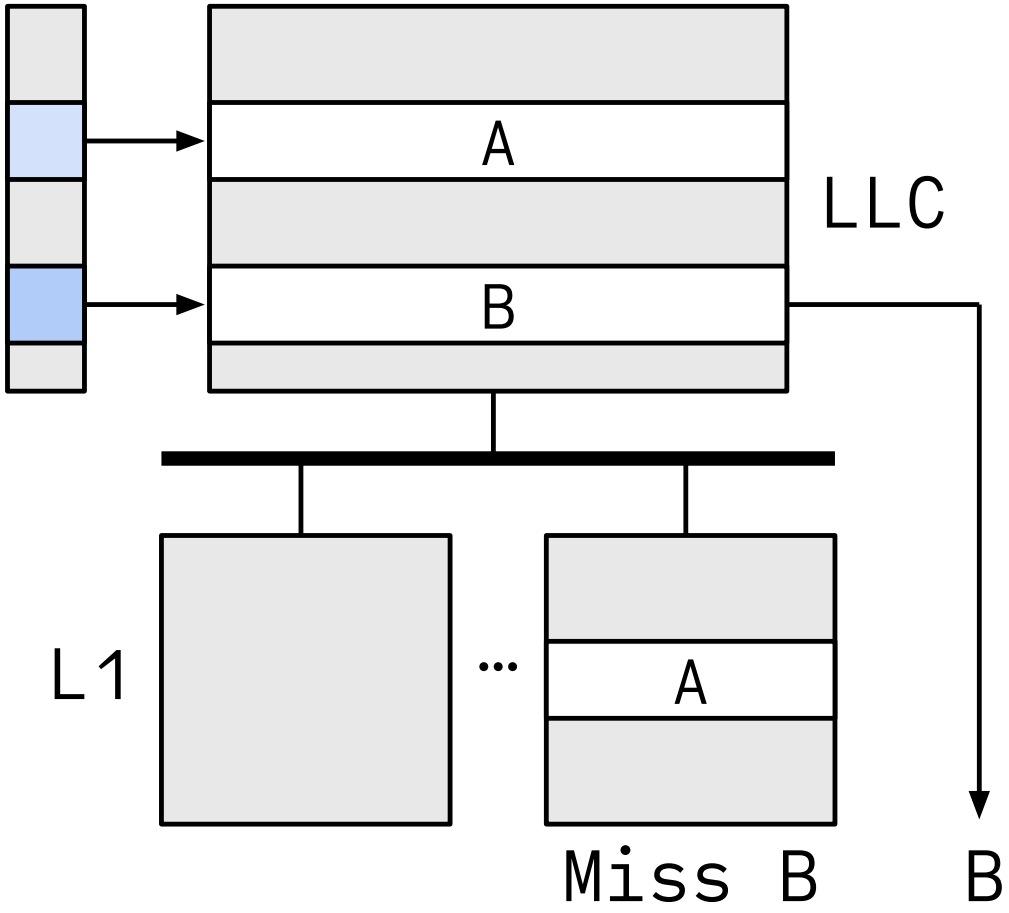
Conventional cache



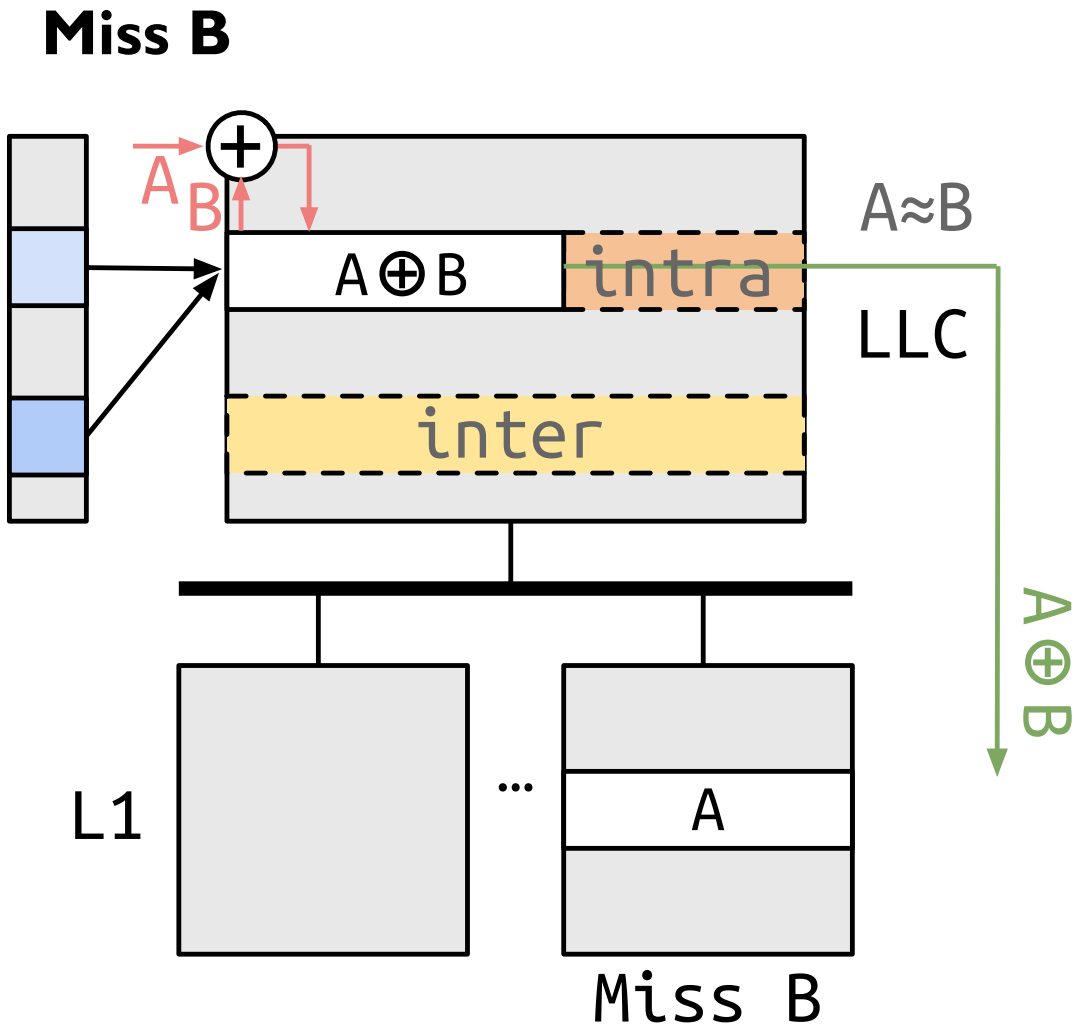
XOR Cache



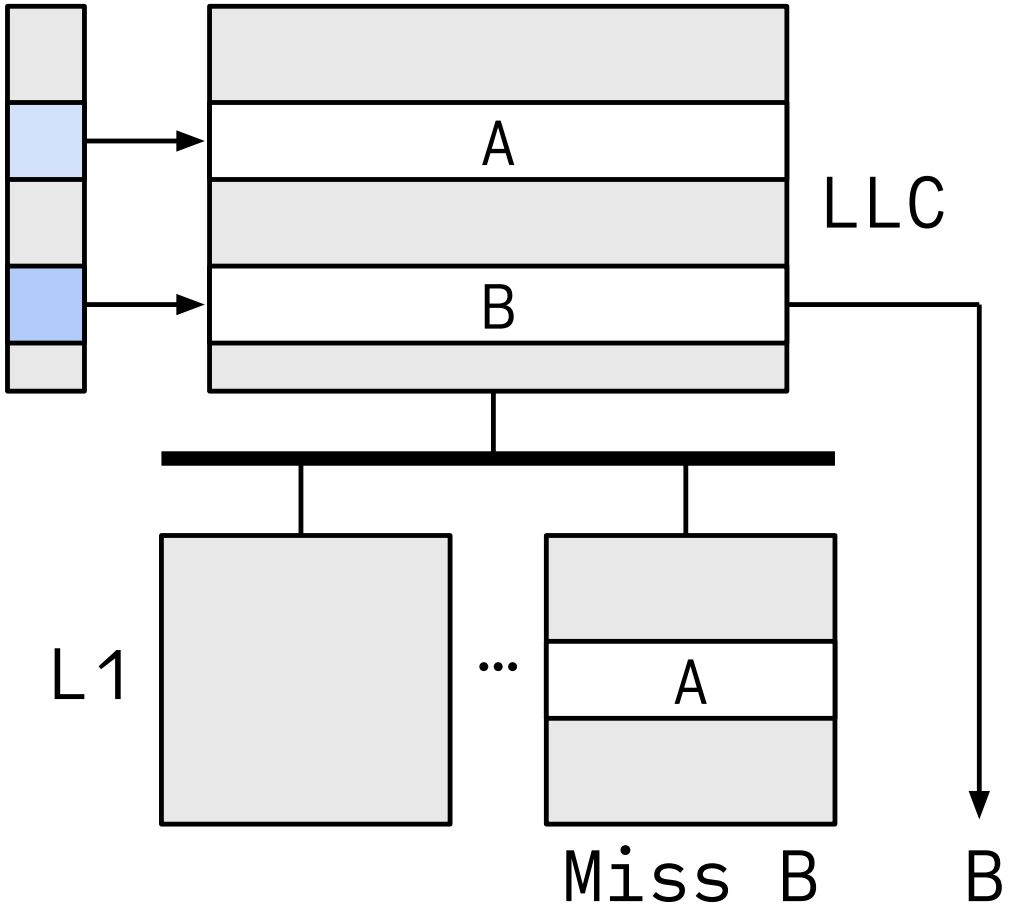
Conventional cache



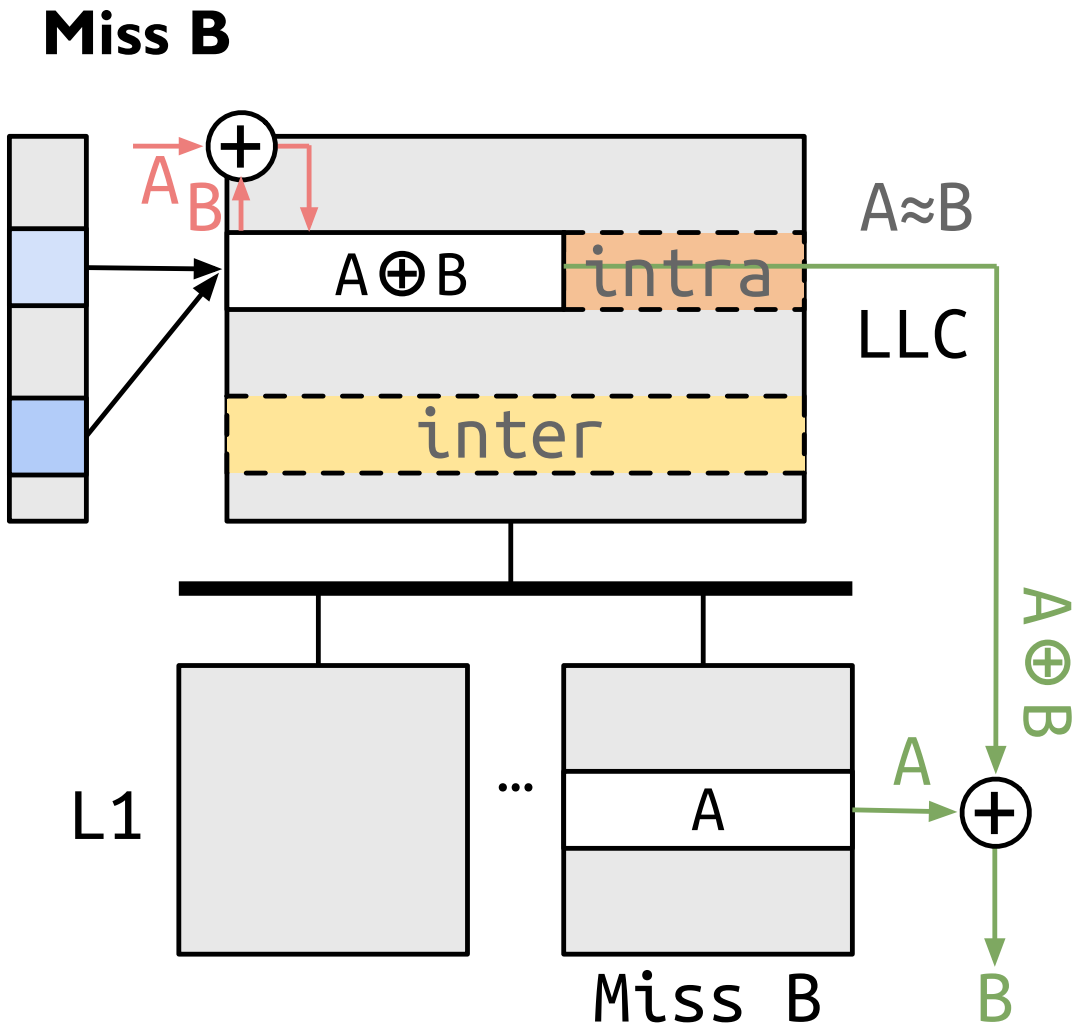
XOR Cache



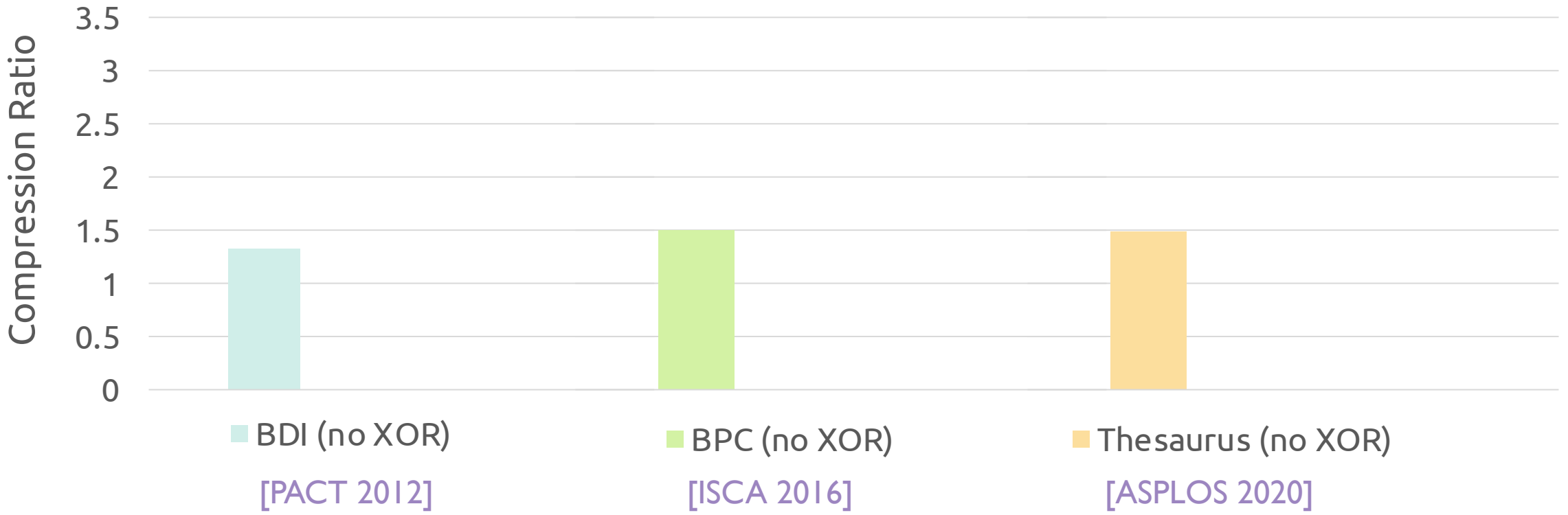
Conventional cache



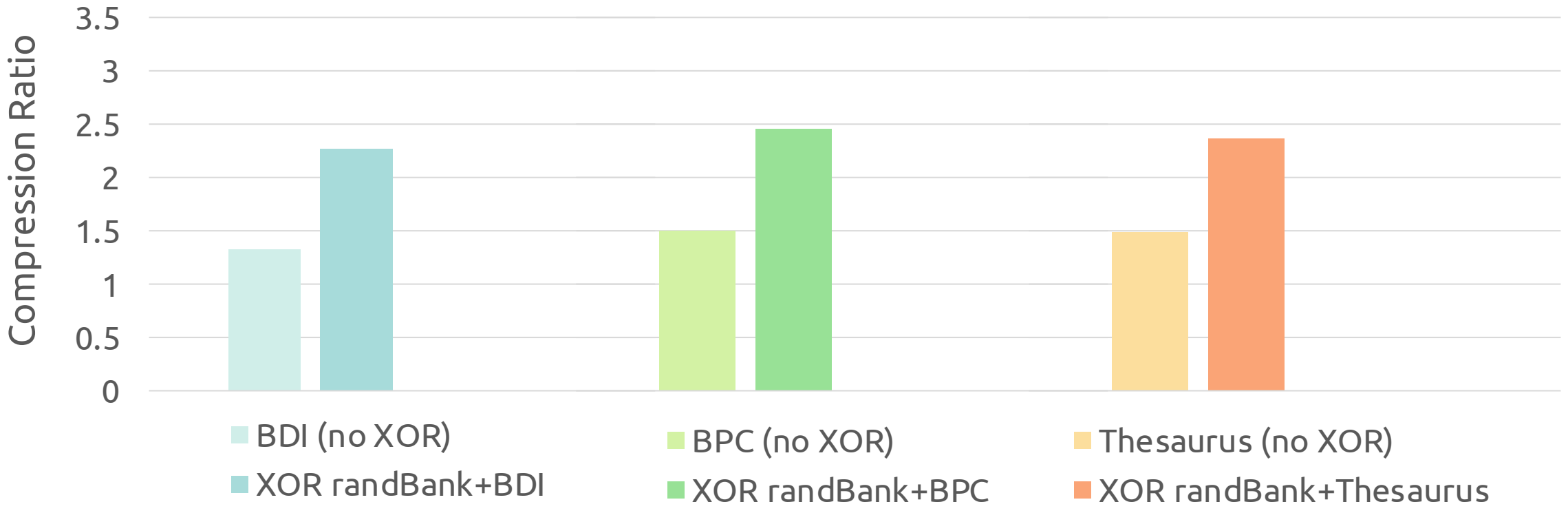
XOR Cache



Compression ratio profiling



Compression ratio profiling



Intra-line synergy example

Two example lines from bodytrack benchmark in PARSEC3.0

Line A	0020 003C 6D7F 0000 7C20 003C 6D7F 0000 7C20 003C …	(16 B)
Line B	0020 004C 6D7F 0000 7C20 004C 6D7F 0000 7C20 004C …	(16 B)

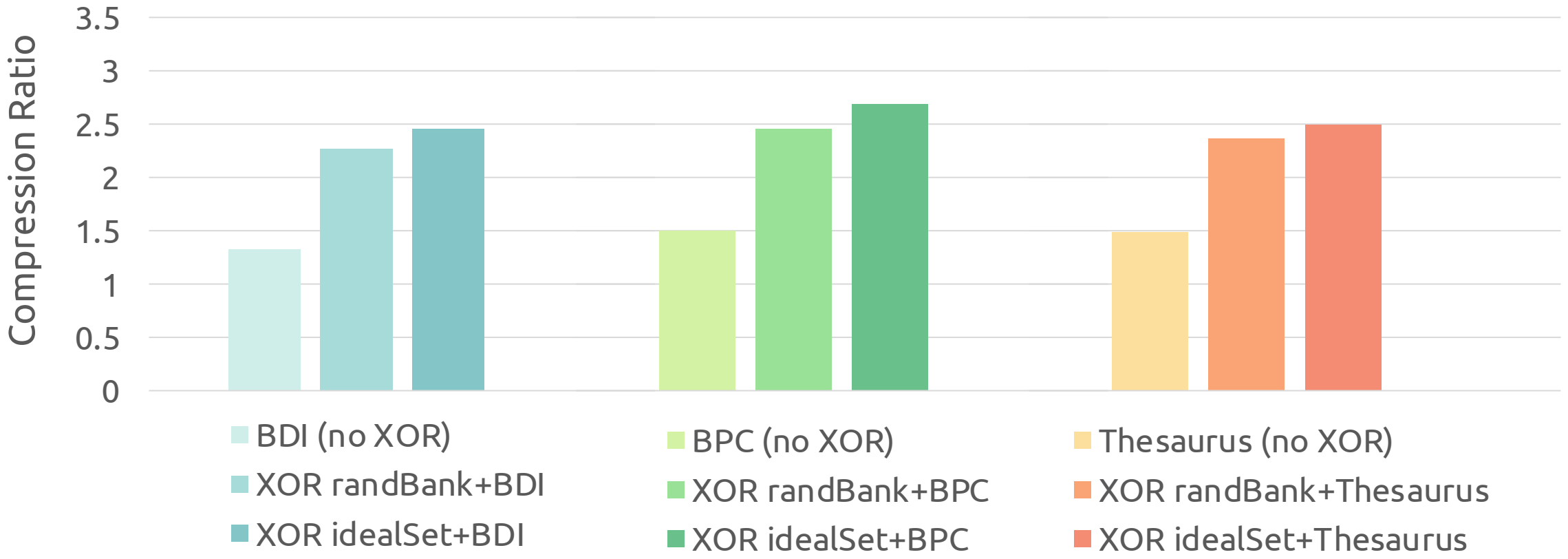
Intra-line synergy example

Two example lines from bodytrack benchmark in PARSEC3.0

Line A	0020 003C 6D7F 0000 7C20 003C 6D7F 0000 7C20 003C …	(16 B)
Line B	0020 004C 6D7F 0000 7C20 004C 6D7F 0000 7C20 004C …	(16 B)
Line A⊕B	0000 0070 0000 0000 0000 0070 0000 0000 0000 0070 …	(8 B)

Greatly reduces entropy when A≈B

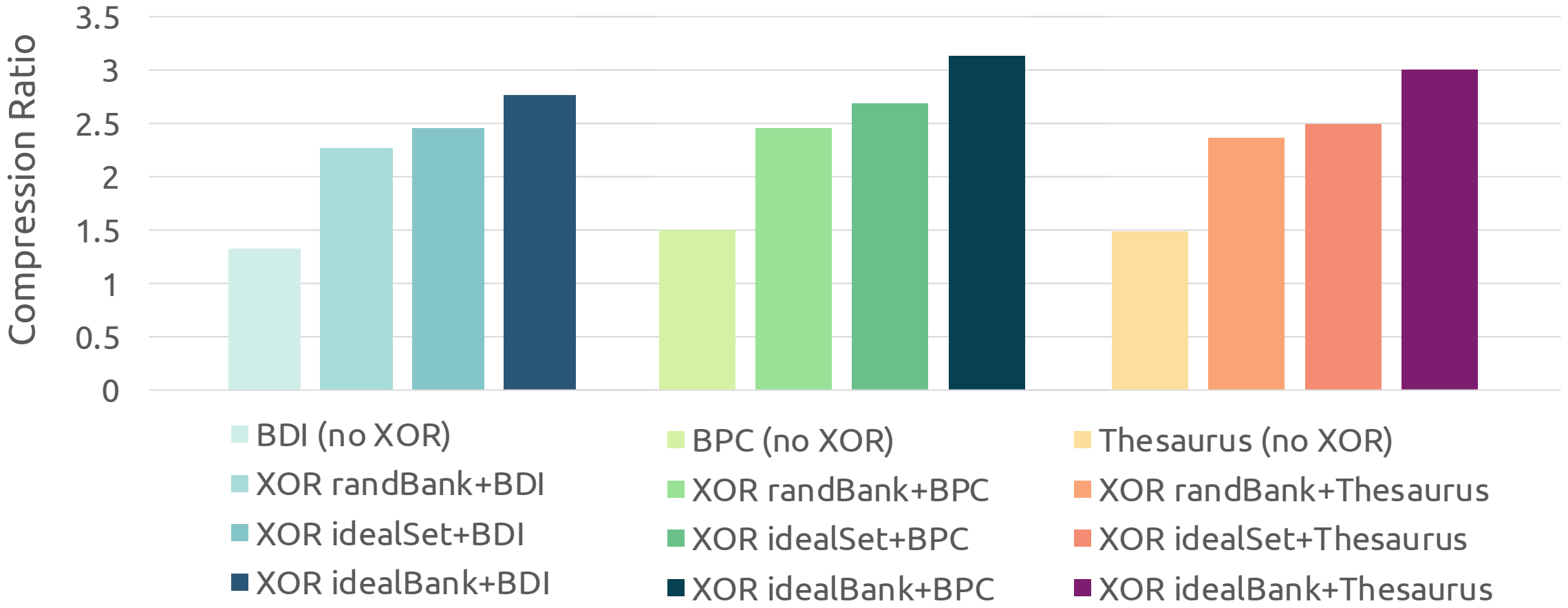
Compression ratio profiling

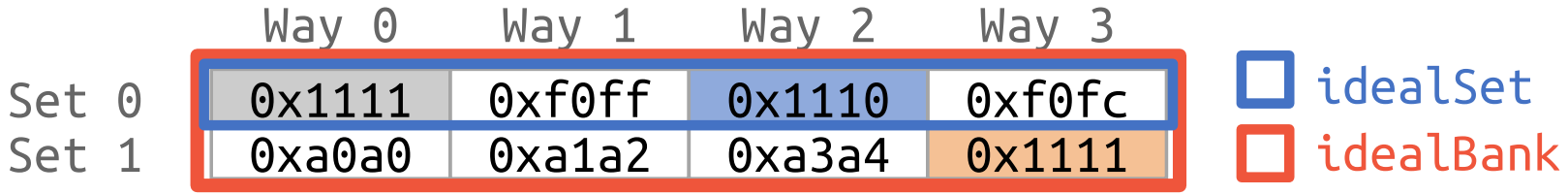


	Way 0	Way 1	Way 2	Way 3
Set 0	0x1111	0xf0ff	0x1110	0xf0fc
Set 1	0xa0a0	0xa1a2	0xa3a4	0x1111

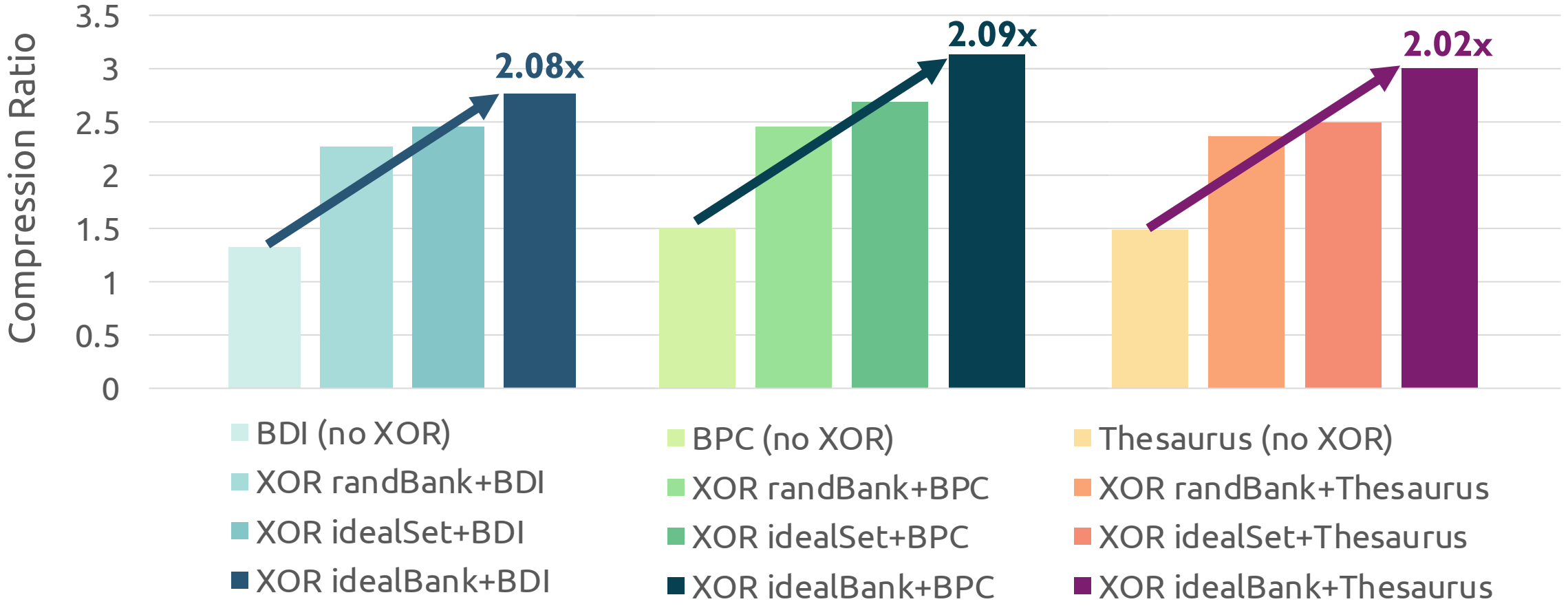
idealSet

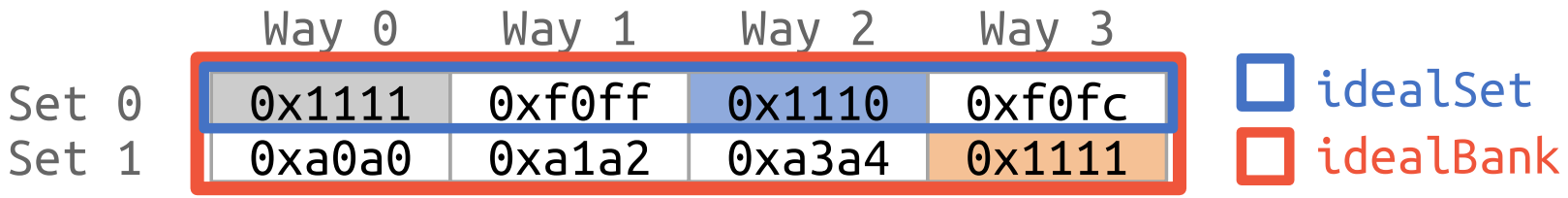
Compression ratio profiling



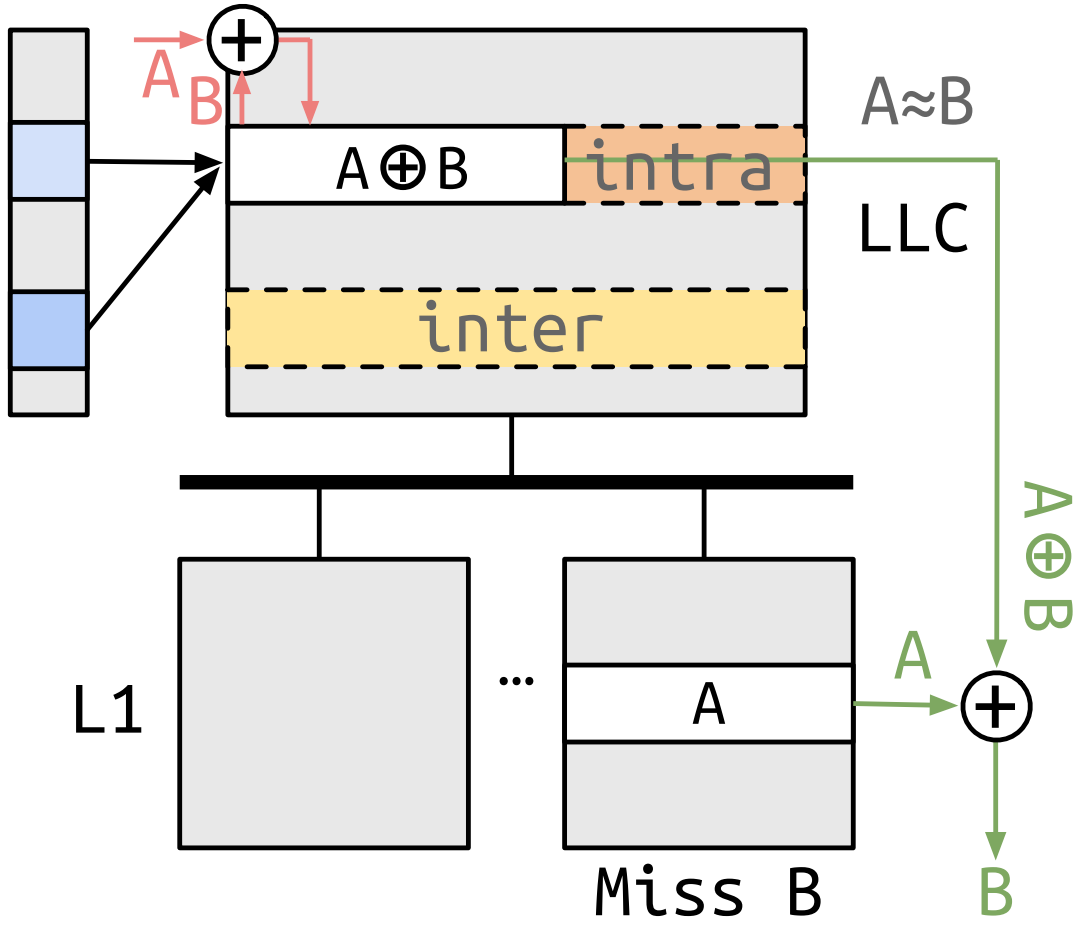


Compression ratio profiling



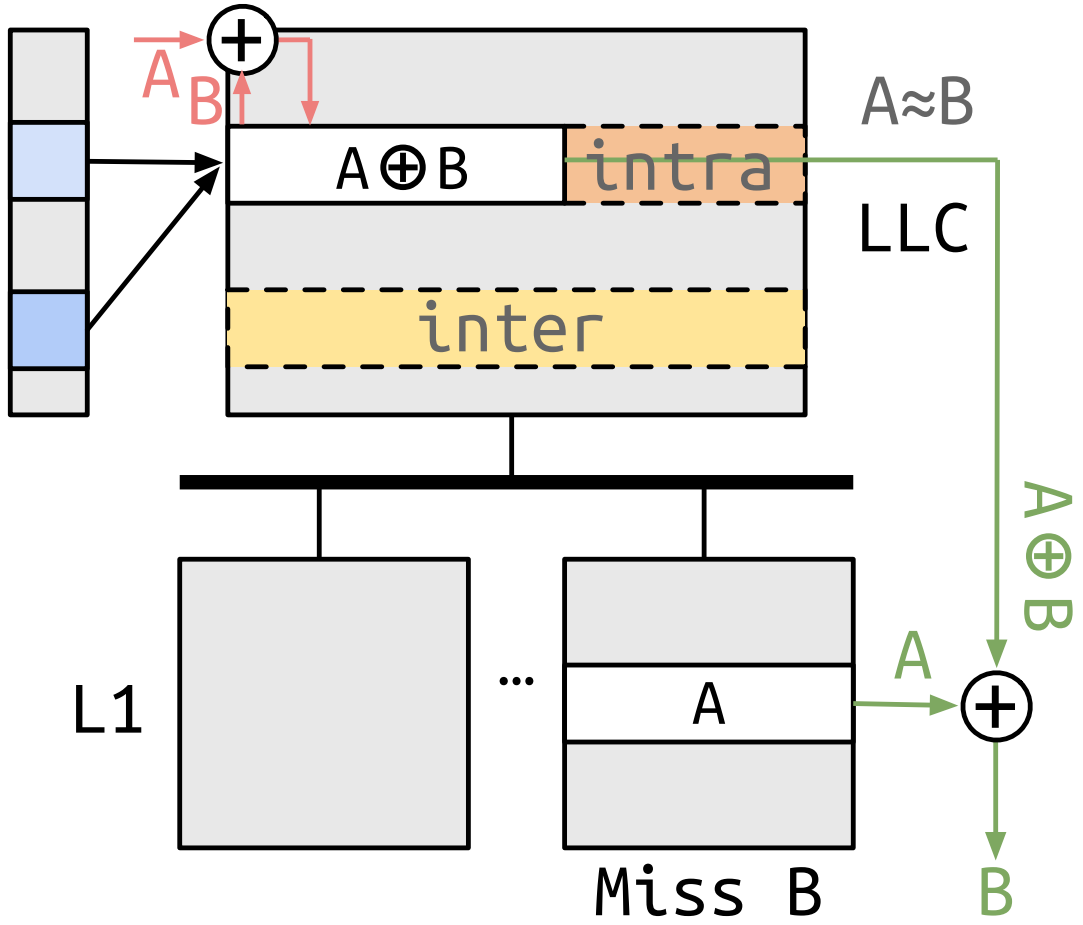


XOR Cache overview



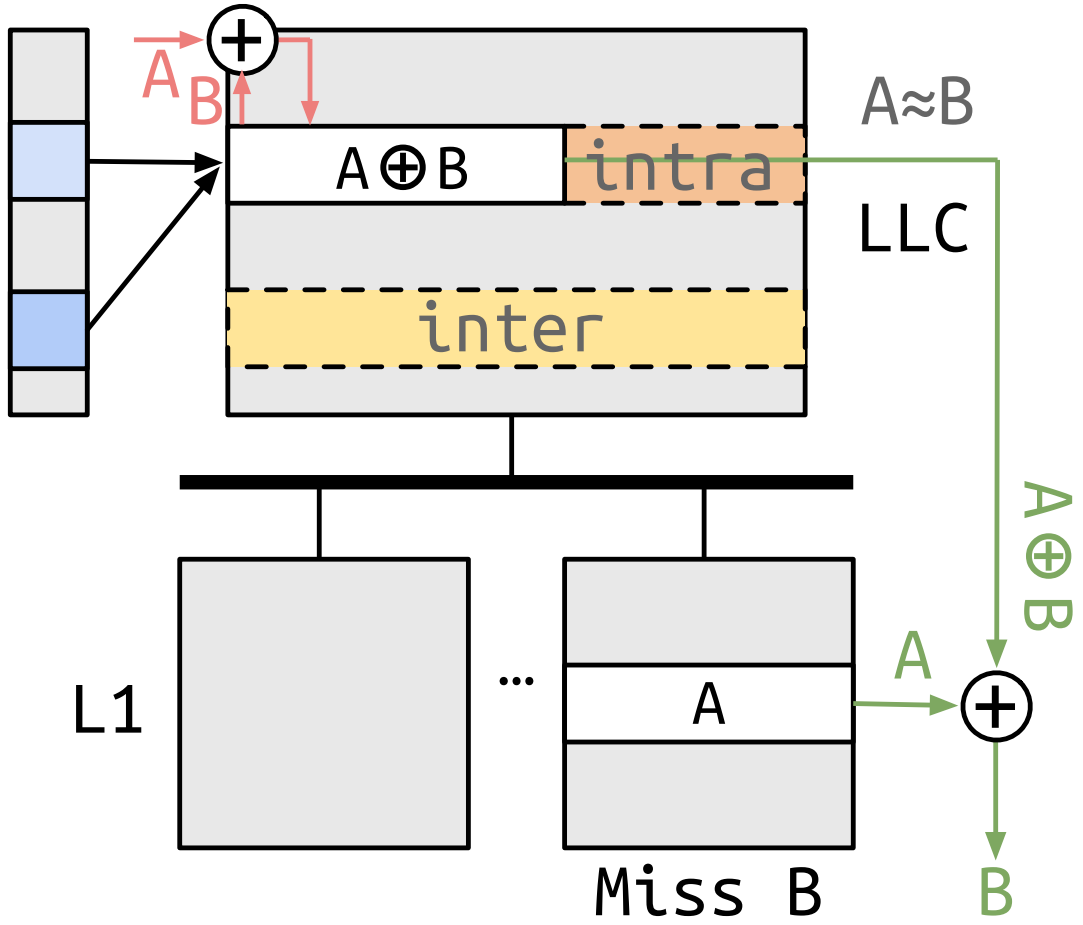
XOR Cache overview

1. Compression



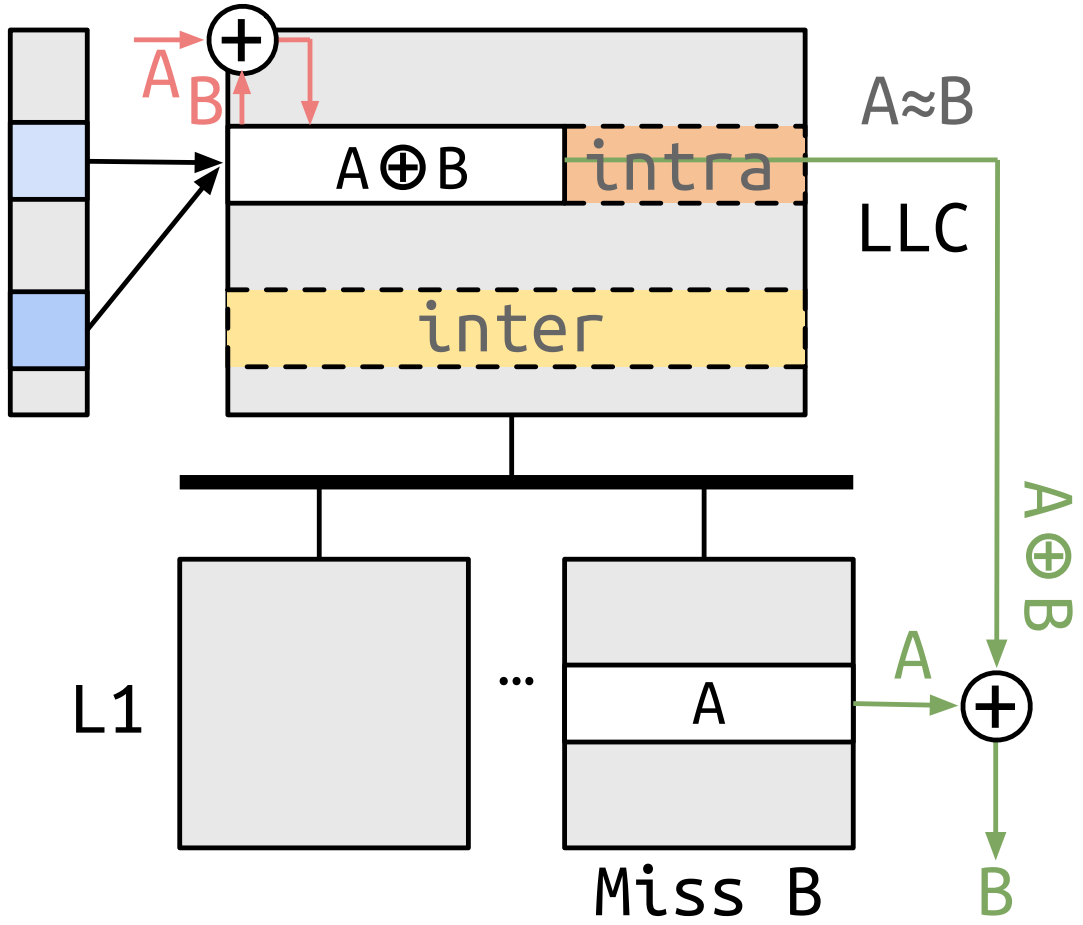
XOR Cache overview

1. **Compression**
2. **Decompression**



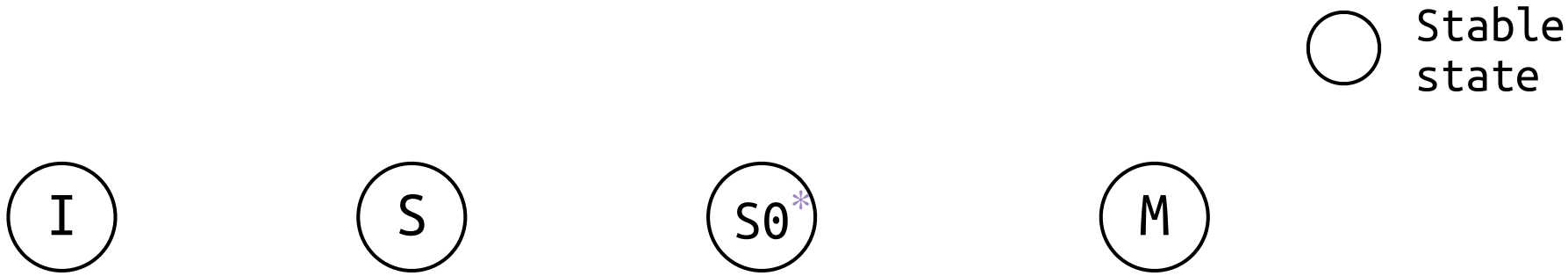
XOR Cache overview

1. **Compression**
2. **Decompression**
3. **UnXORing**



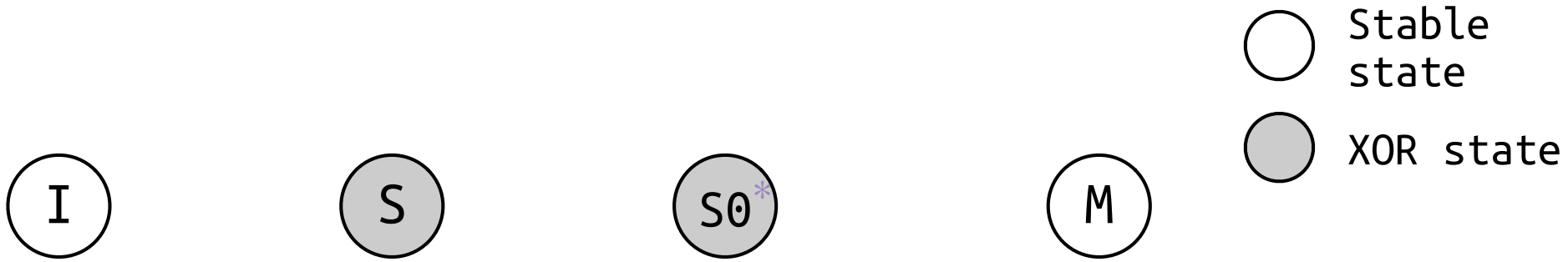
XOR Cache coherence support - MSI

XOR Cache coherence support - MSI



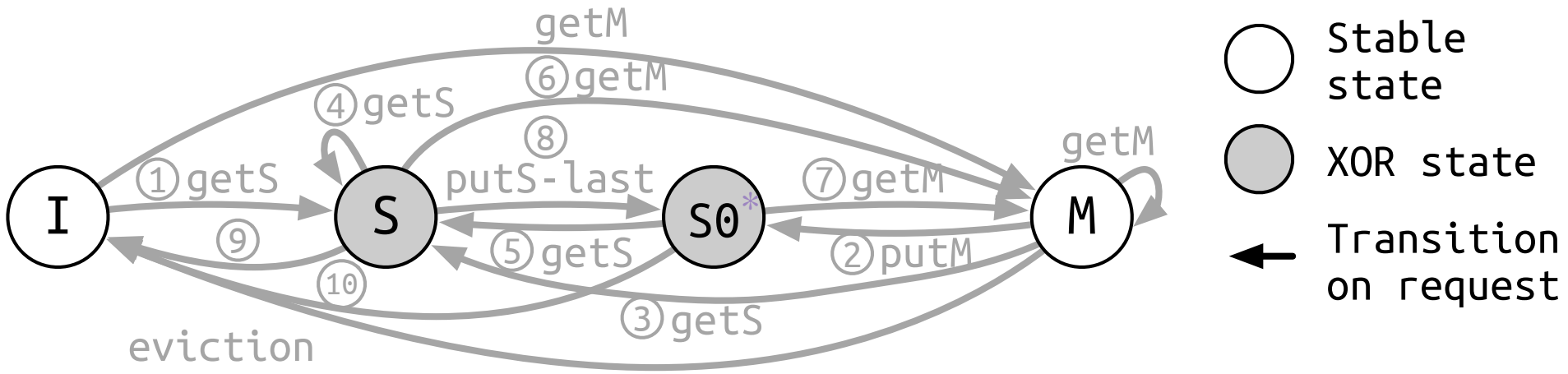
* S0 is a special S state with an empty sharer list.

XOR Cache coherence support - MSI



* S0 is a special S state with an empty sharer list.

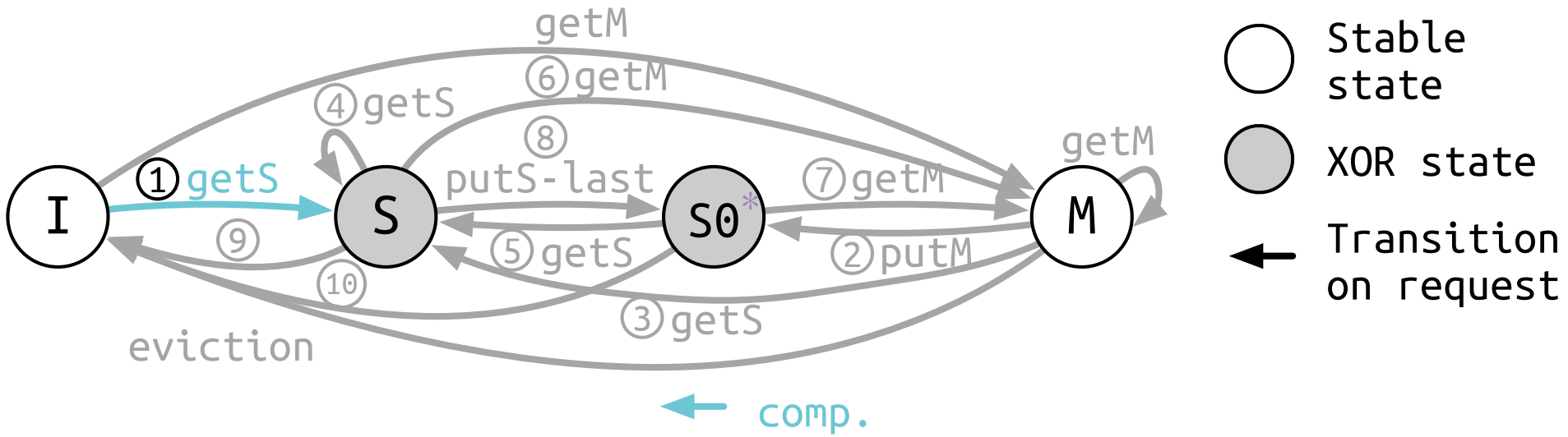
XOR Cache coherence support - MSI



* S0 is a special S state with an empty sharer list.

XOR Cache coherence support - compression

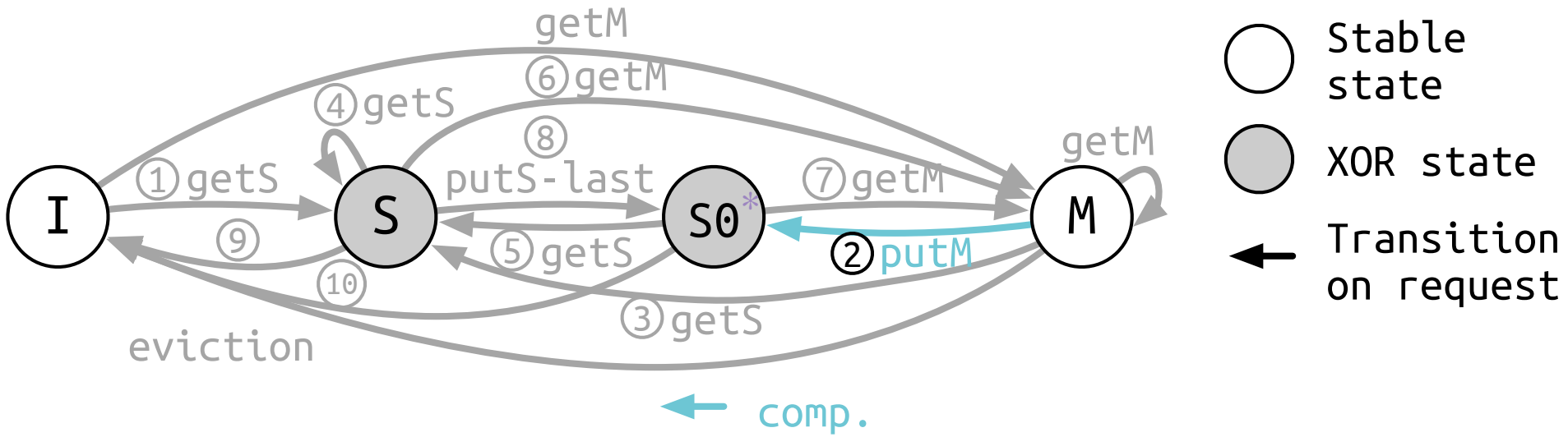
LLC miss



* S0 is a special S state with an empty sharer list.

XOR Cache coherence support - compression

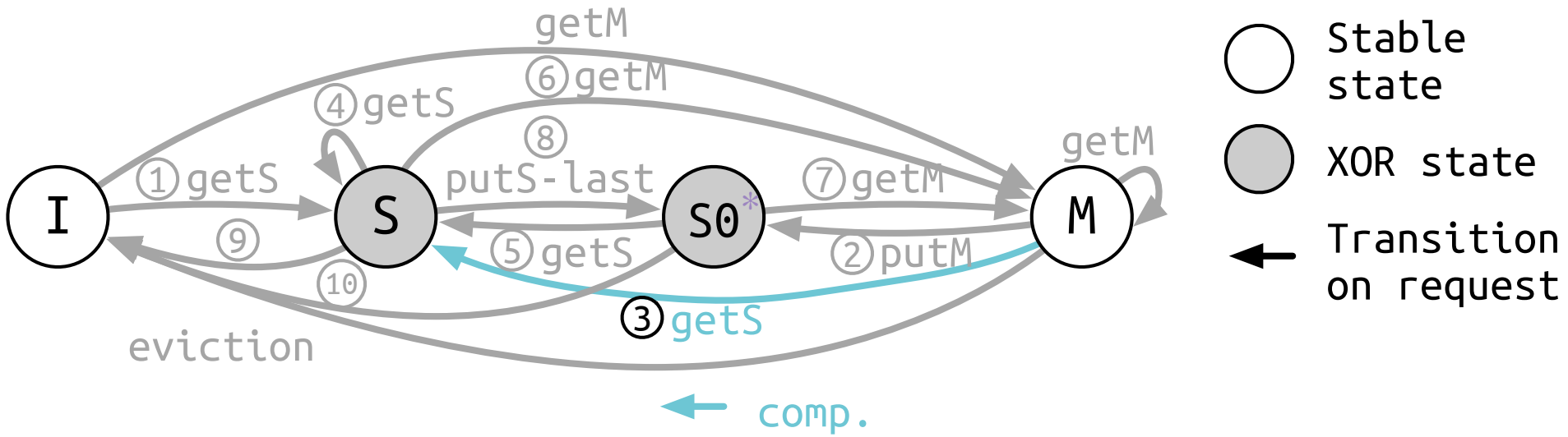
Private cache write-back



* S0 is a special S state with an empty sharer list.

XOR Cache coherence support - compression

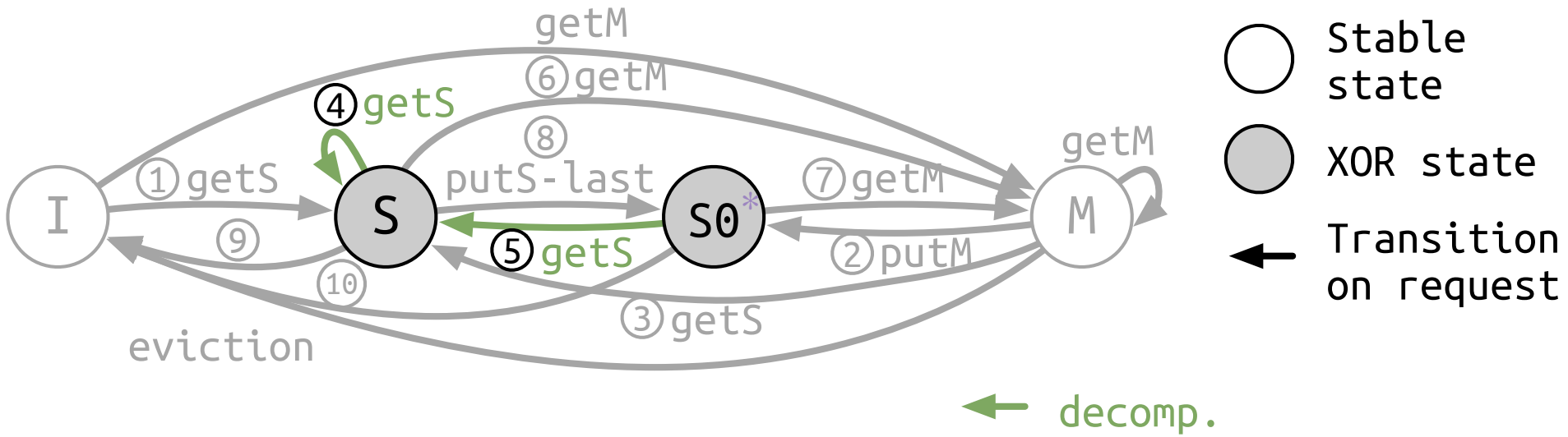
Downgrading



* S0 is a special S state with an empty sharer list.

XOR Cache coherence support - decompression

LLC hit

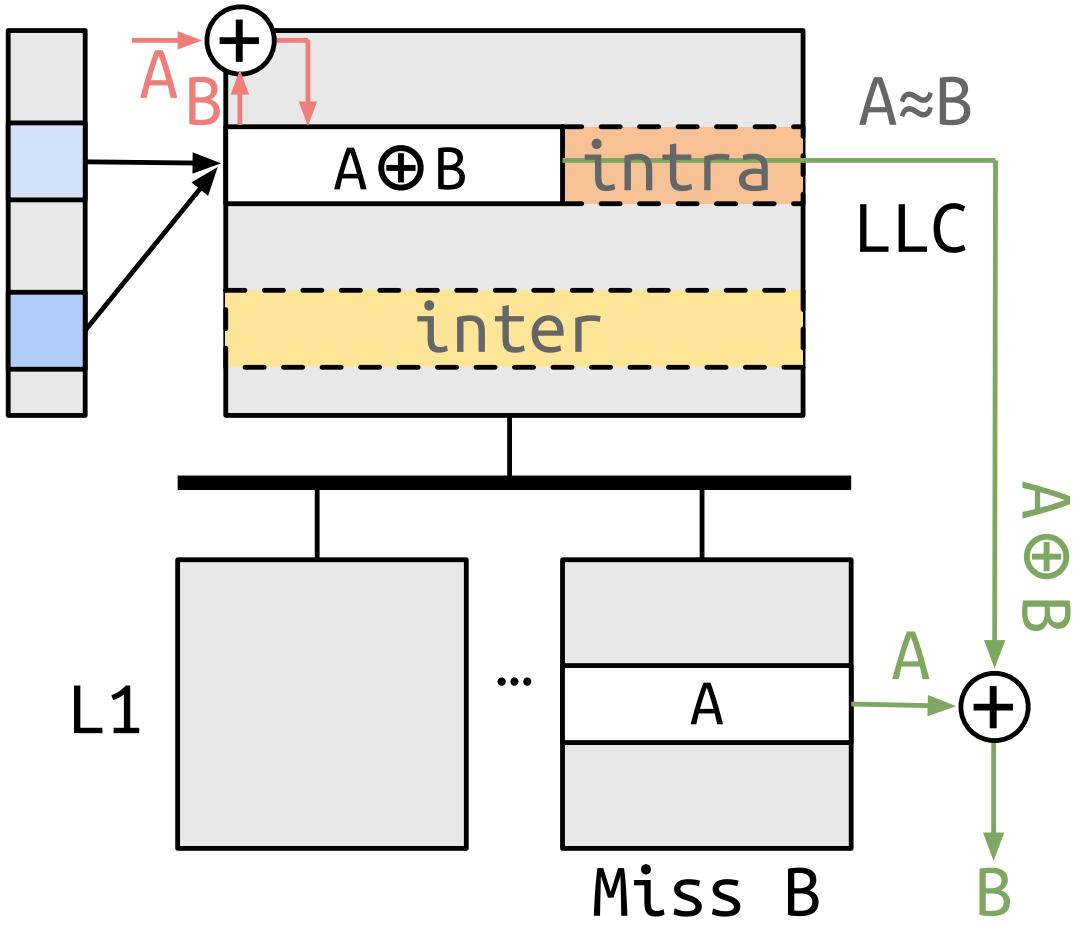


* S0 is a special S state with an empty sharer list.

XOR Cache coherence support - decompression

LLC hit

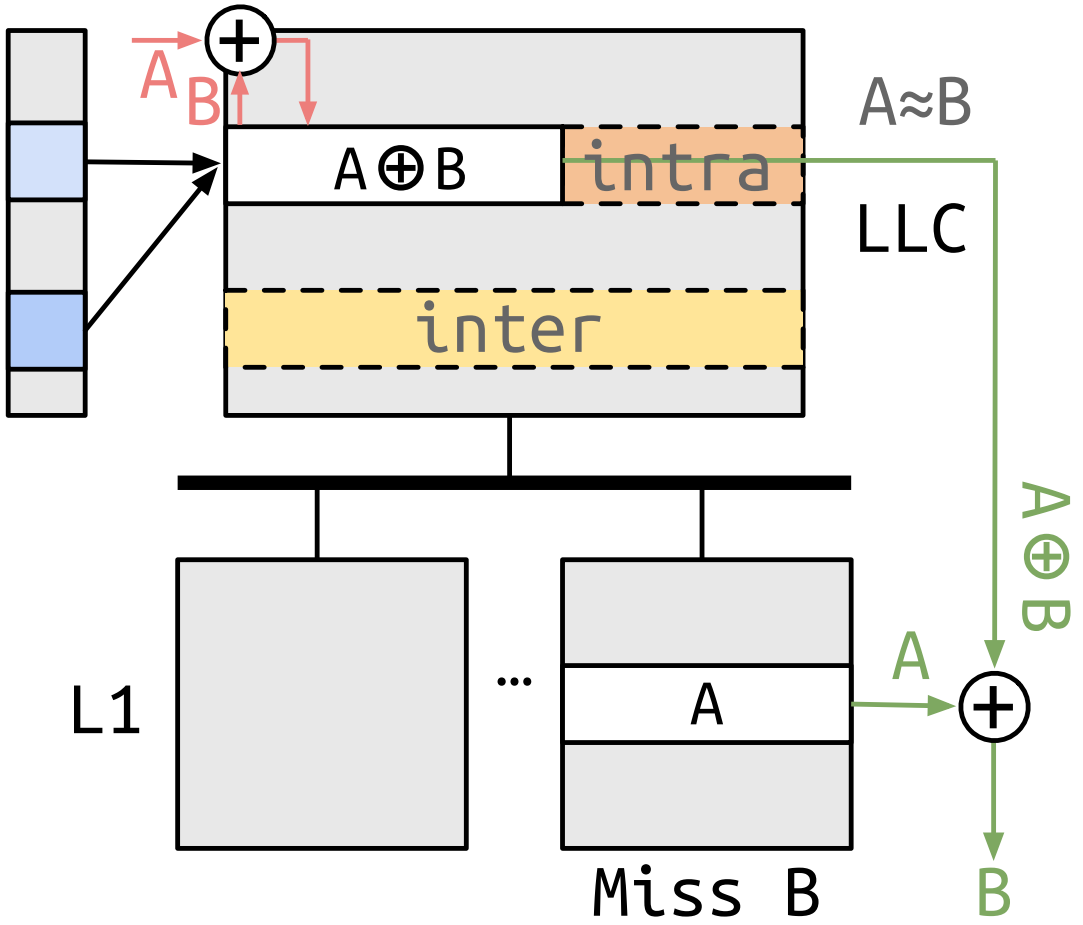
Local recovery



XOR Cache coherence support - decompression

LLC hit

Local recovery

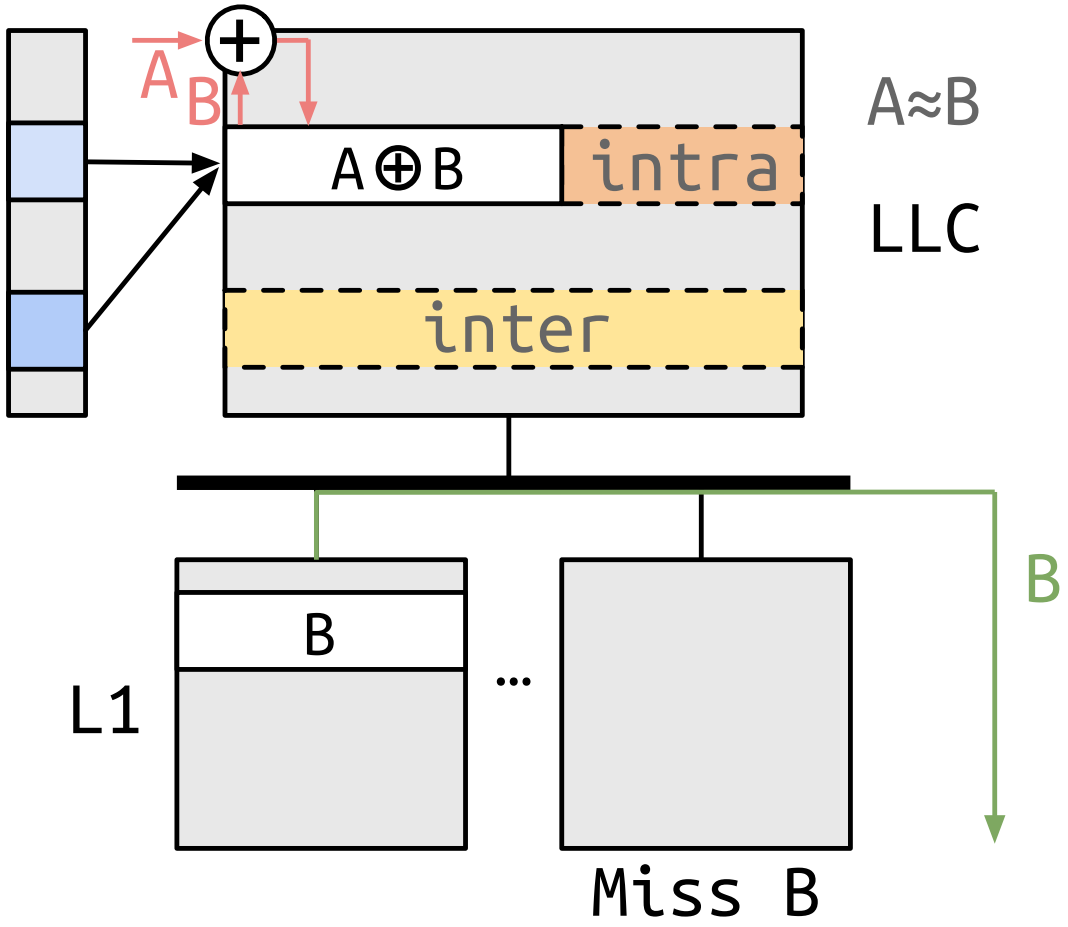


Case	A Dir state (S/S0)	B Dir state (S/S0)	B requestor shares line A
Local recovery	S	S/S0	Yes

XOR Cache coherence support - decompression

LLC hit

Direct forwarding

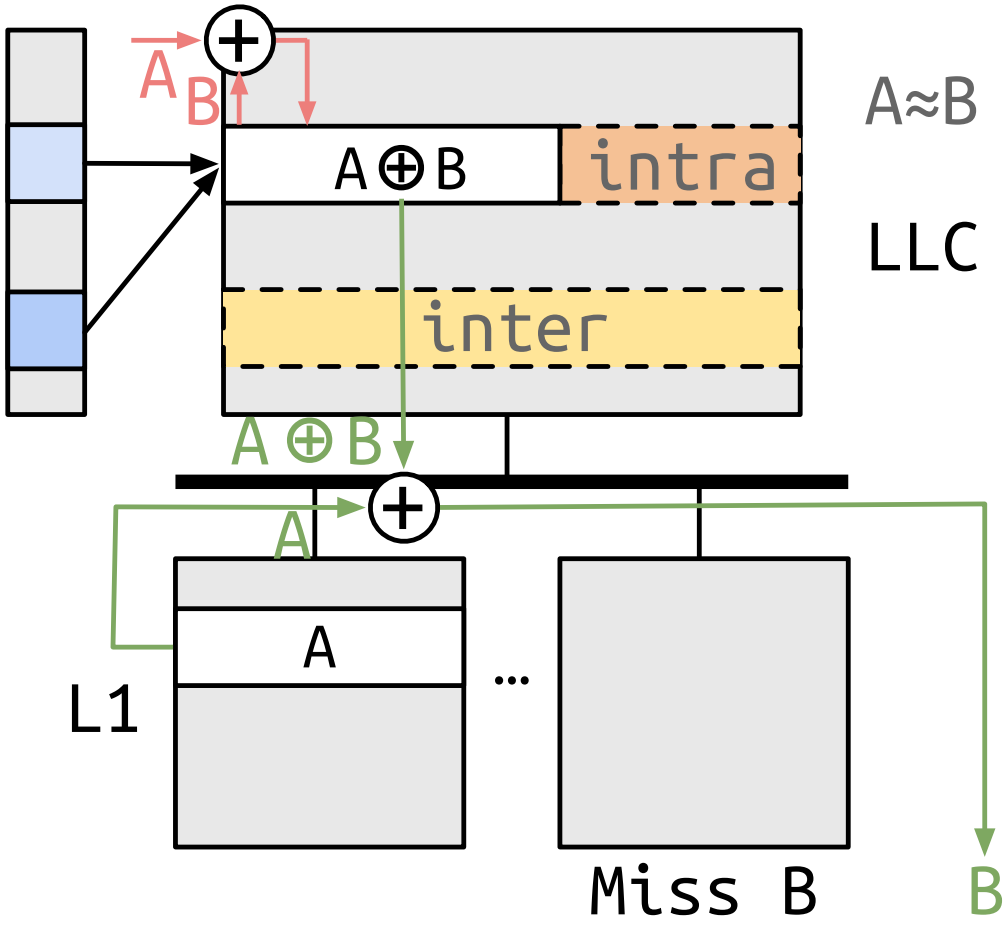


Case	A Dir state (S/S0)	B Dir state (S/S0)	B requestor shares line A
Local recovery	S	S/S0	Yes
Direct forwarding	S/S0	S	No

XOR Cache coherence support - decompression

LLC hit

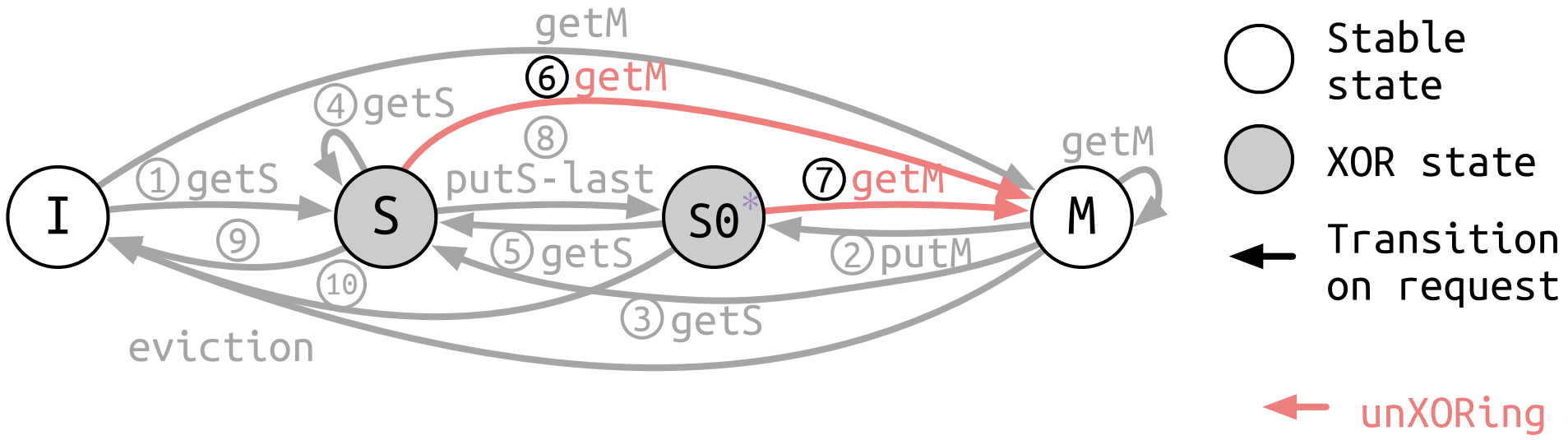
Remote recovery



Case	A Dir state (S/S0)	B Dir state (S/S0)	B requestor shares line A
Local recovery	S	S/S0	Yes
Direct forwarding	S/S0	S	No
Remote recovery	S	S0	No

XOR Cache coherence support - unXORing

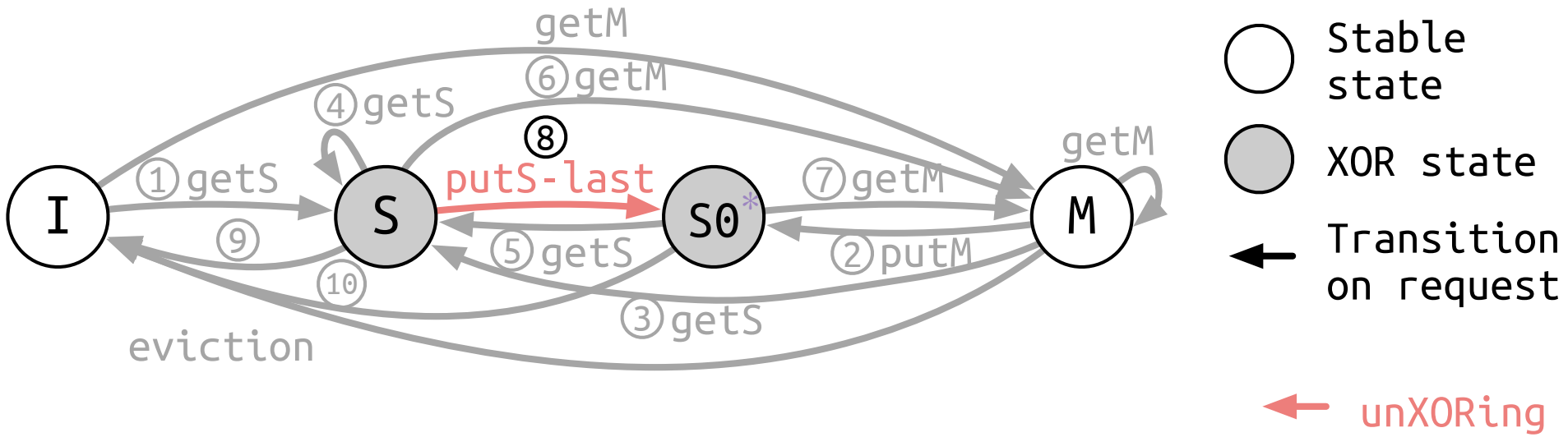
Upgrading



* S0 is a special S state with an empty sharer list.

XOR Cache coherence support - unXORing

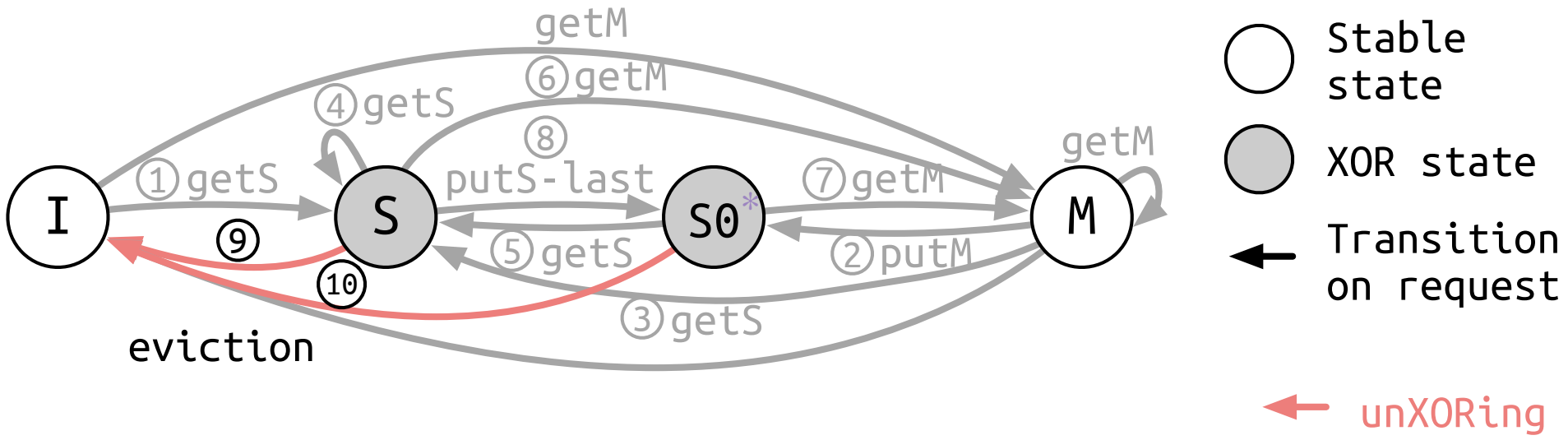
Last sharer eviction



* S0 is a special S state with an empty sharer list.

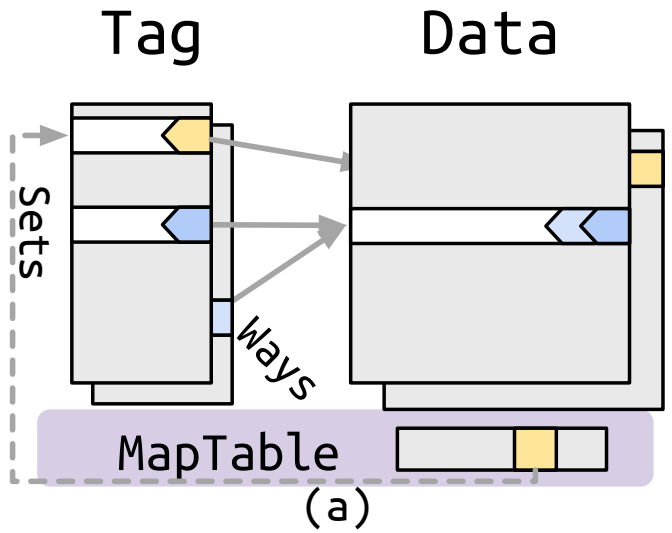
XOR Cache coherence support - unXORing

LLC eviction



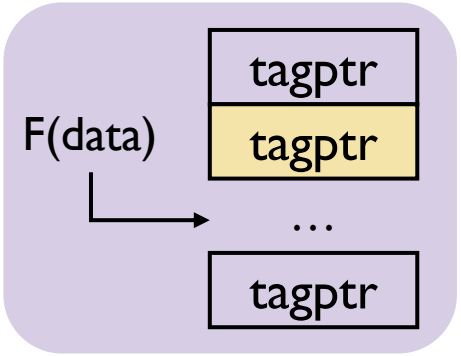
* S0 is a special S state with an empty sharer list.

XOR Cache architecture - organization



(a)

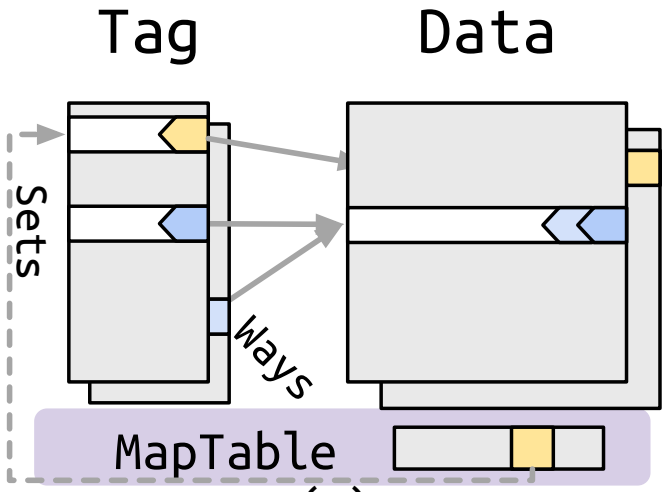
Decoupled tag data organization



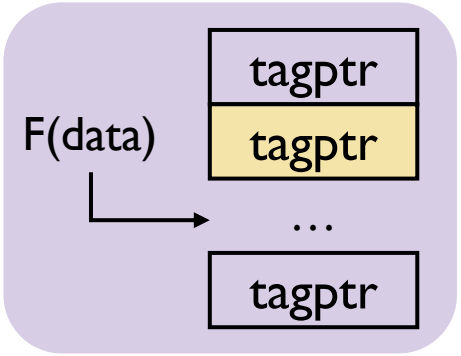
(b)

Map table

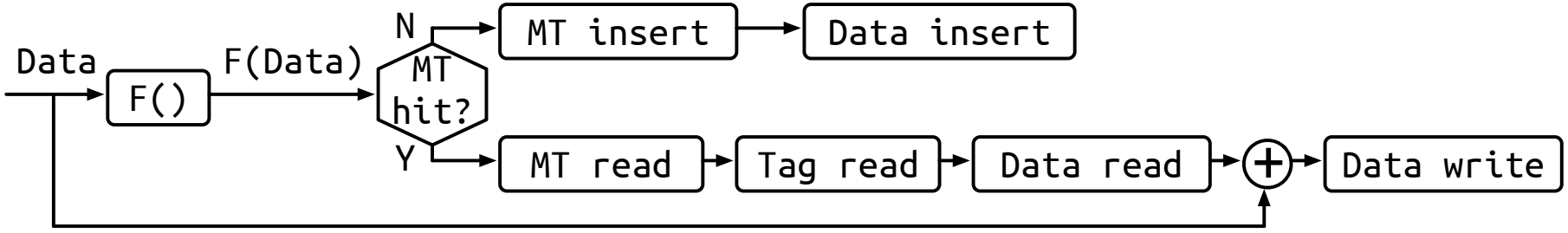
XOR Cache architecture - organization



(a)
Decoupled tag data organization



(b)
Map table



(c)
Insertion flow

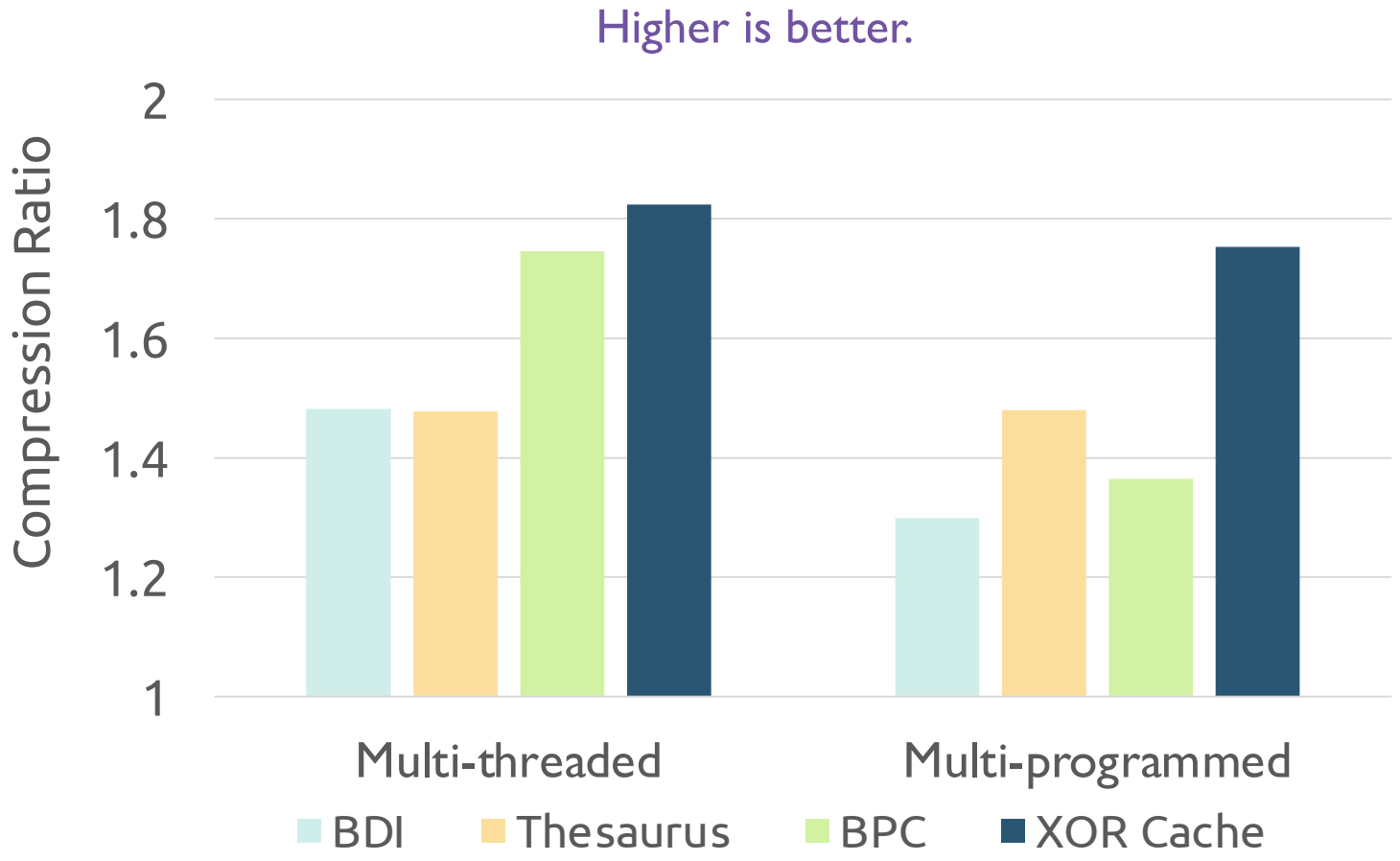
Evaluation - configuration

- Performance: gem5 ruby, full-system
- Benchmark: muti-threaded (perfect, parsec3.0), multi-programmed (spec 2017)
- Baseline configuration:

	Baseline configuration
CPU	4 core, 3GHz x86-64
L1I and L1D	32KiB, 4 way, 4 cycle, 64B line, LRU, Private
L2	256 KiB, 8 way, 9 cycle, 64B line, LRU, Private
L3	1MiB per bank, 16 way, 40 cycle, 64B line, LRU, Shared, 4 banks, mixed-inclusive *
Memory	DualChannelDDR4-2400

* Compressed LLC sized according to profiled compression ratio.

Evaluation - compression ratio improvements

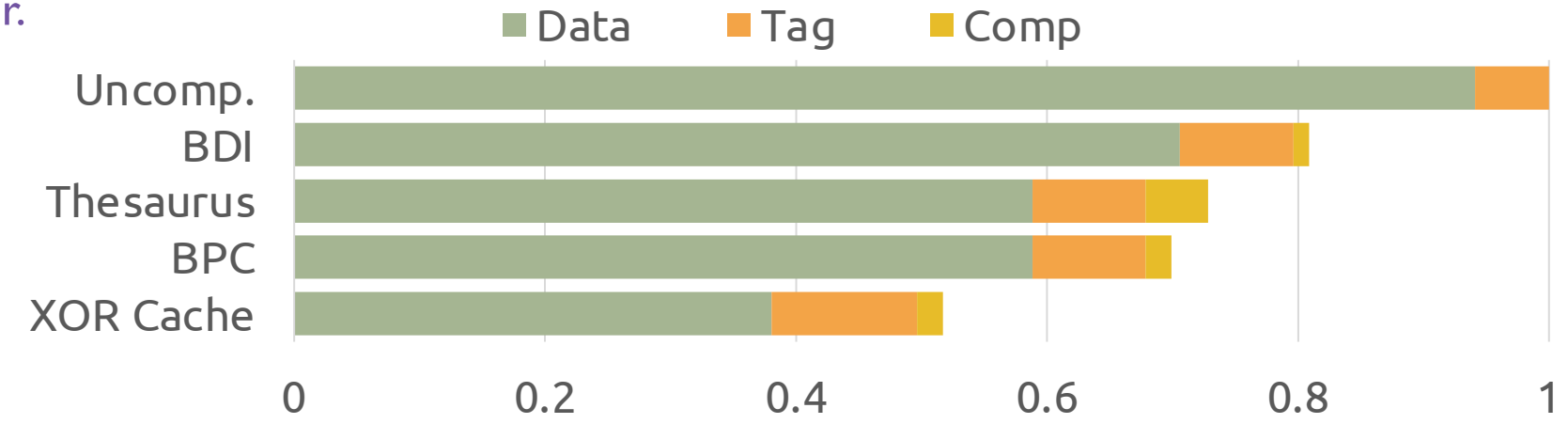


Boosting compression ratio by **23%** and **35%** on multi-threaded and multi-programmed workloads

Evaluation - area and power improvements

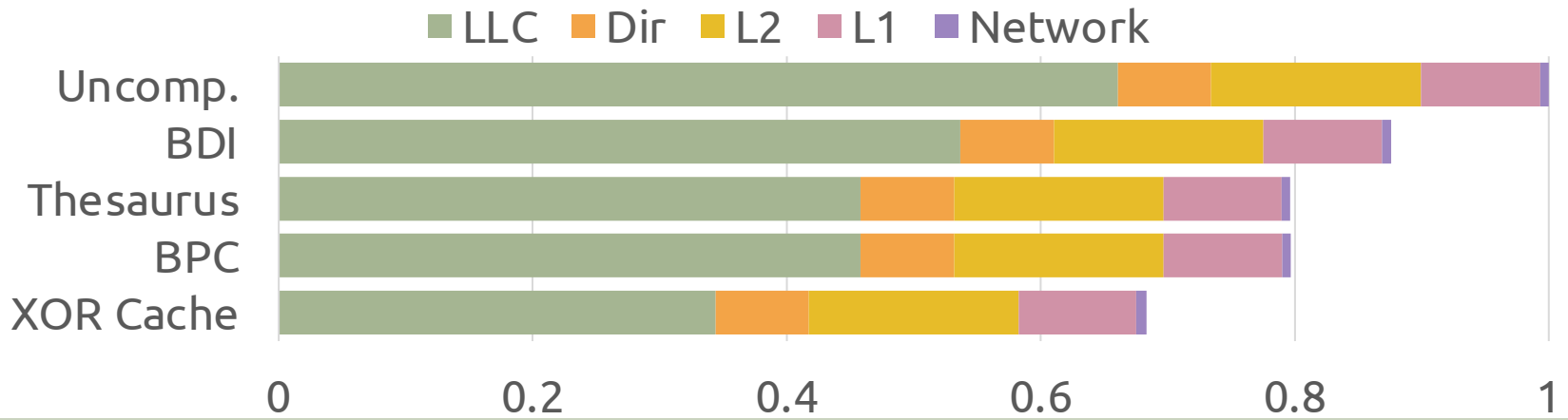
Lower is better.

Area



reducing area by **1.93**x over uncompressed baseline

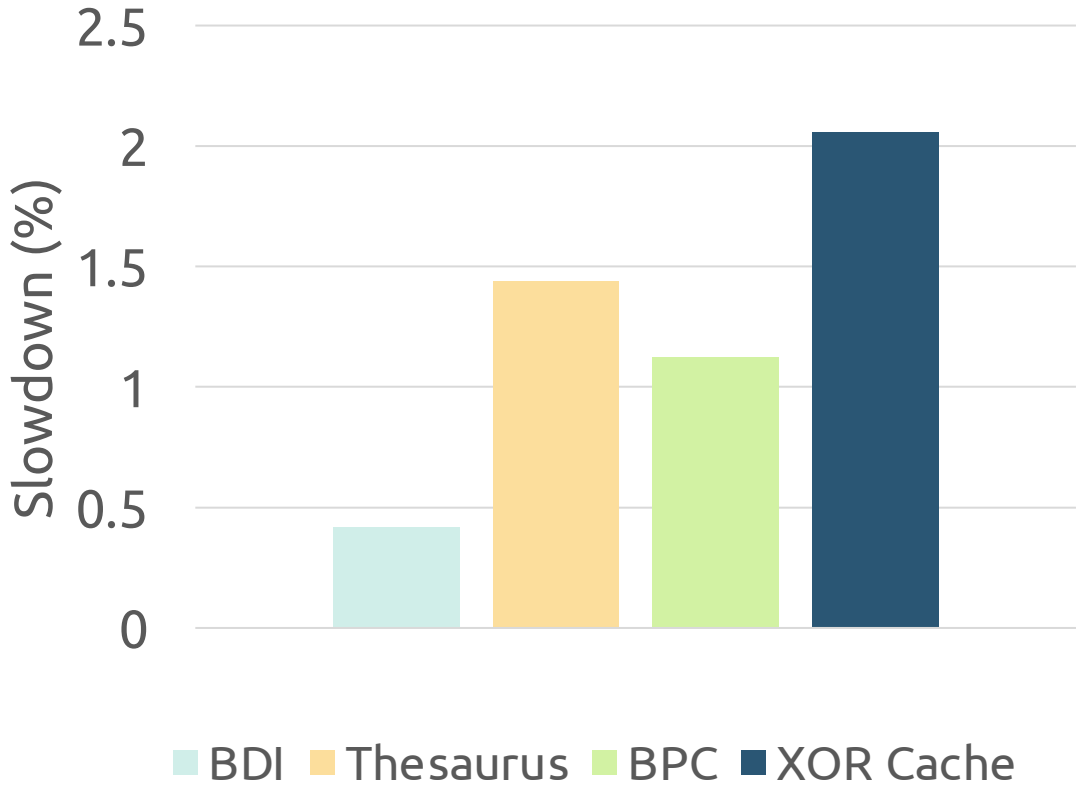
Power

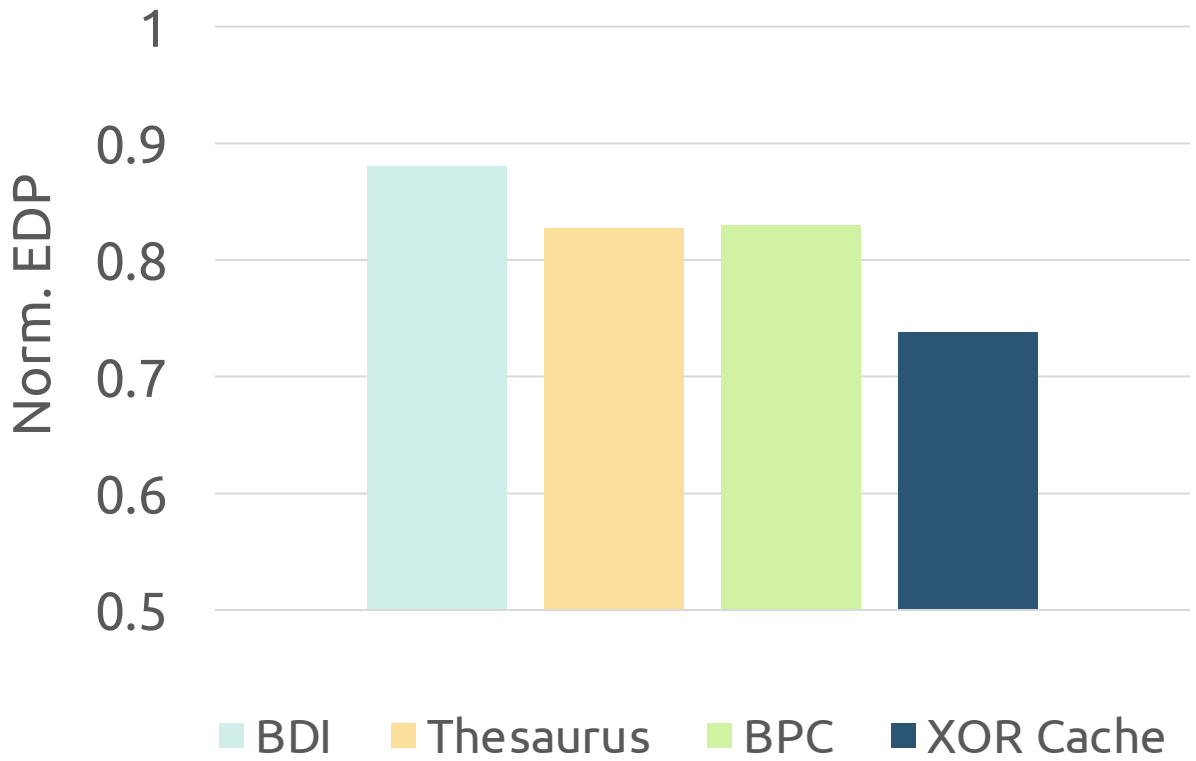


reducing power of LLC by **1.92**x and reduces power of hierarchy by **1.46**x over uncompressed baseline

Evaluation - performance overhead & energy delay product (EDP)

Lower is better.





Incurring a performance overhead of 2.06% while reducing EDP by **26%**

XOR Cache summary

Exploits redundancy due to private caching and inclusion

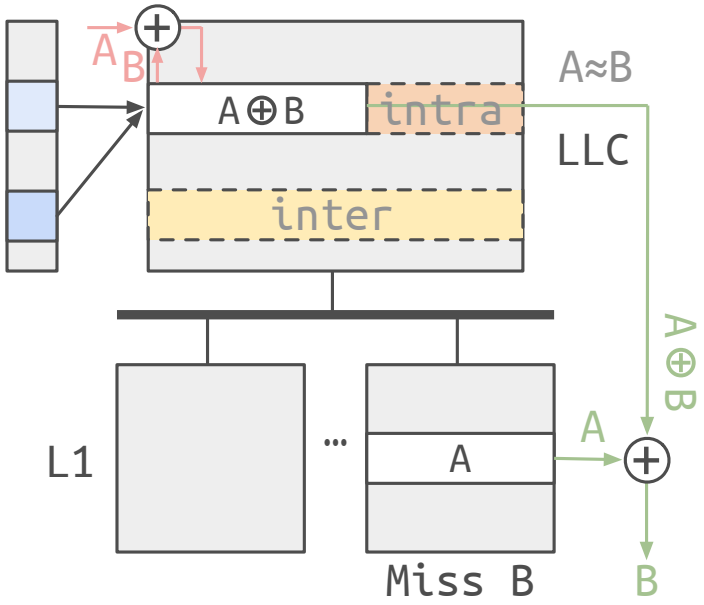
Performs **inter-line** compression and
catalyzes **intra-line** compression opportunities

Reduces LLC area by **1.93**x, power by **1.92**x,
and EDP by **26.3%**

Thank you

The XOR Cache: A Catalyst for Compression

Zhewen Pan Joshua San Miguel



Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN-MADISON