Doppelgänger: A Cache for Approximate Computing

Joshua San Miguel

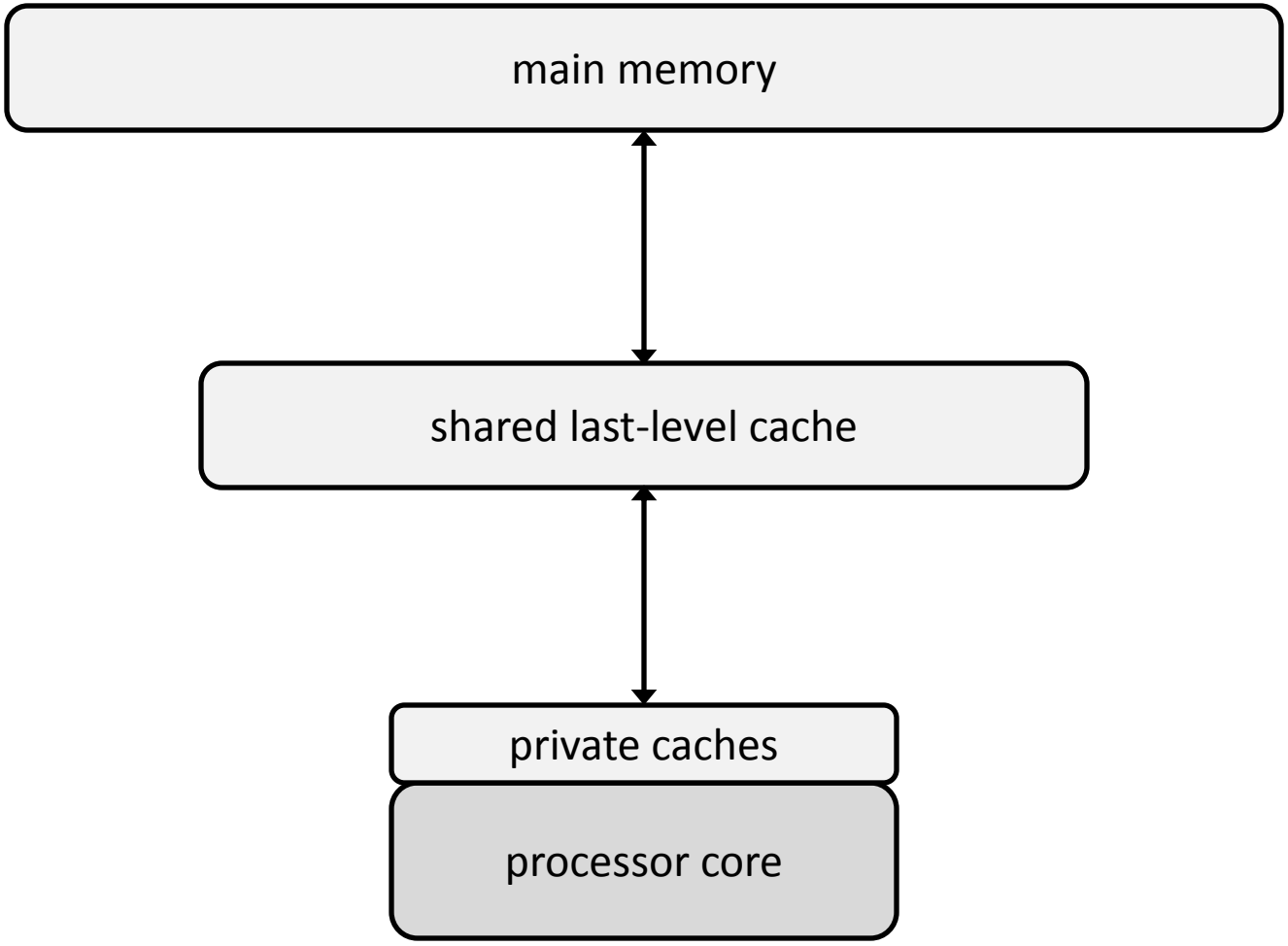
Jorge Albericio

Andreas Moshovos

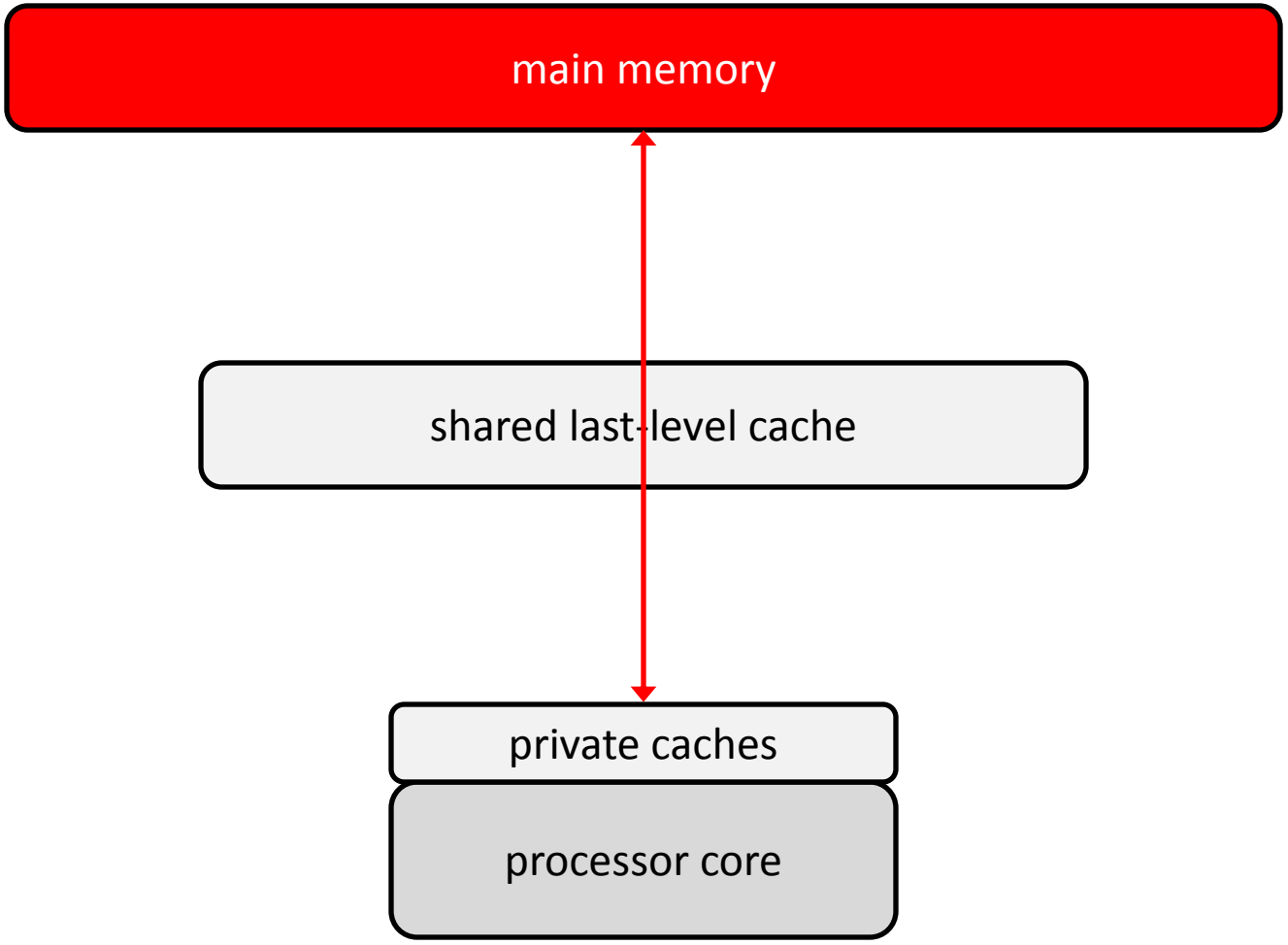
Natalie Enright Jerger

The Edward S. Rogers Sr. Department
of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

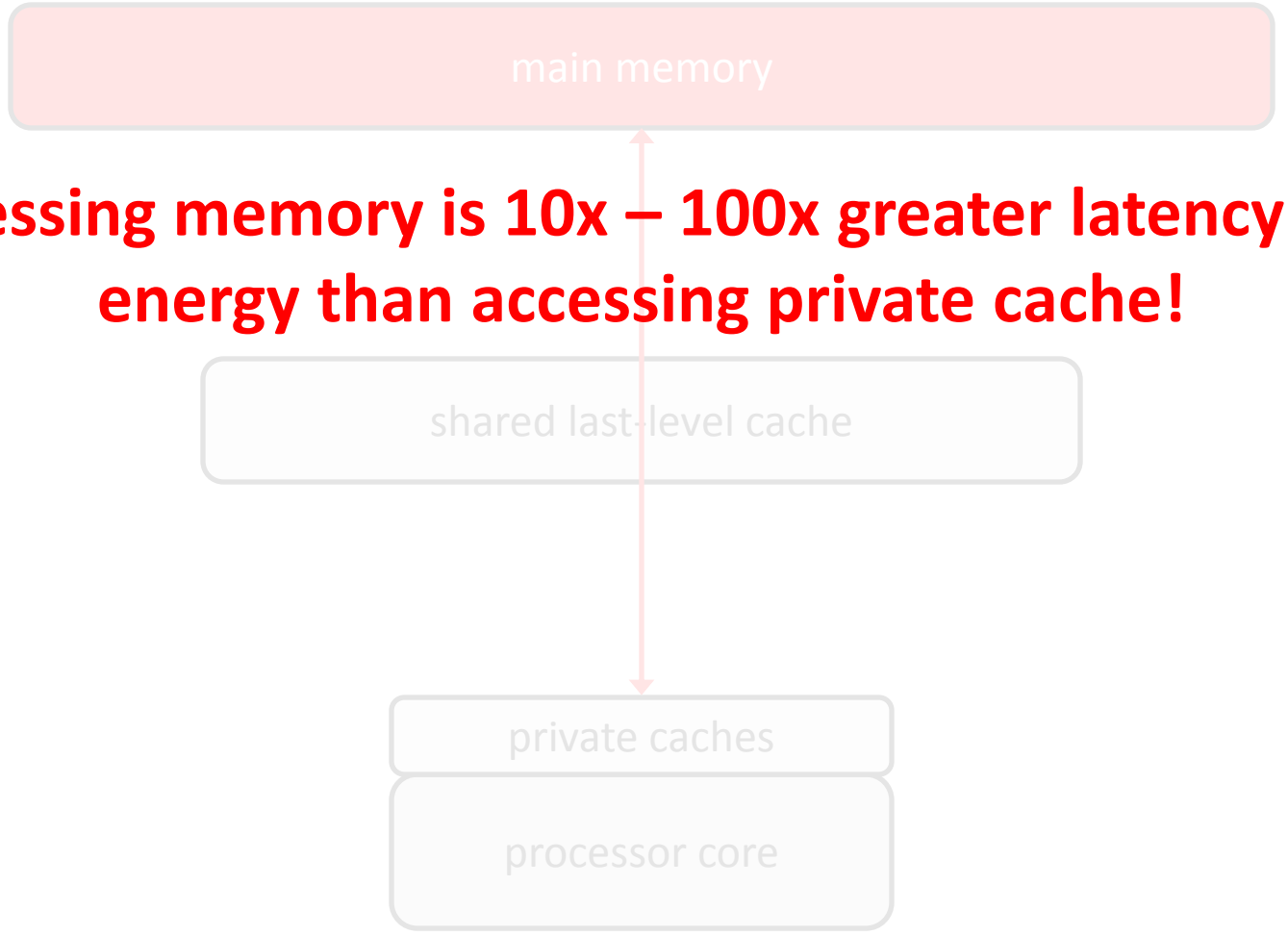
Cache Hierarchy



Cache Hierarchy

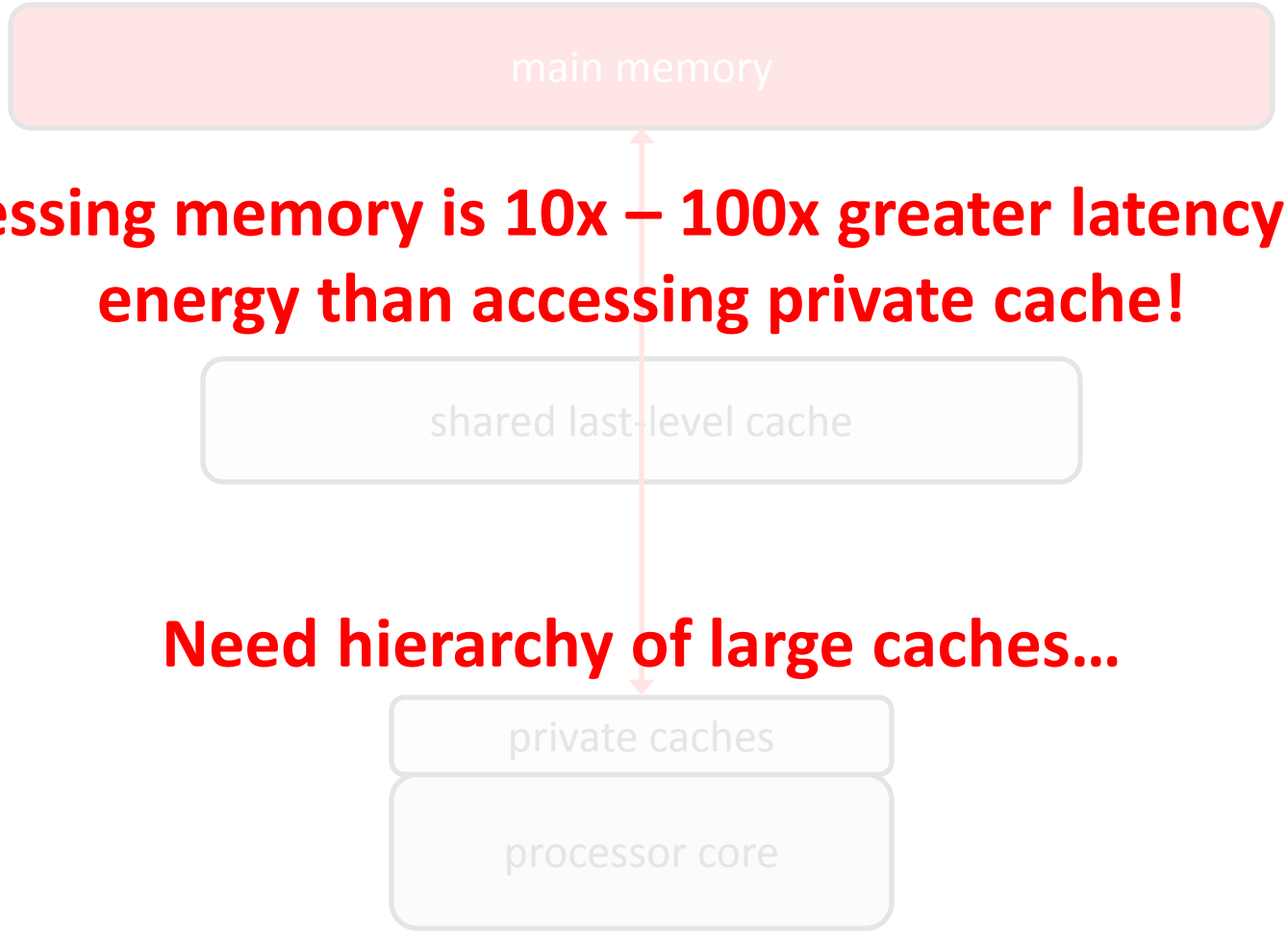


Cache Hierarchy



Accessing memory is 10x – 100x greater latency and energy than accessing private cache!

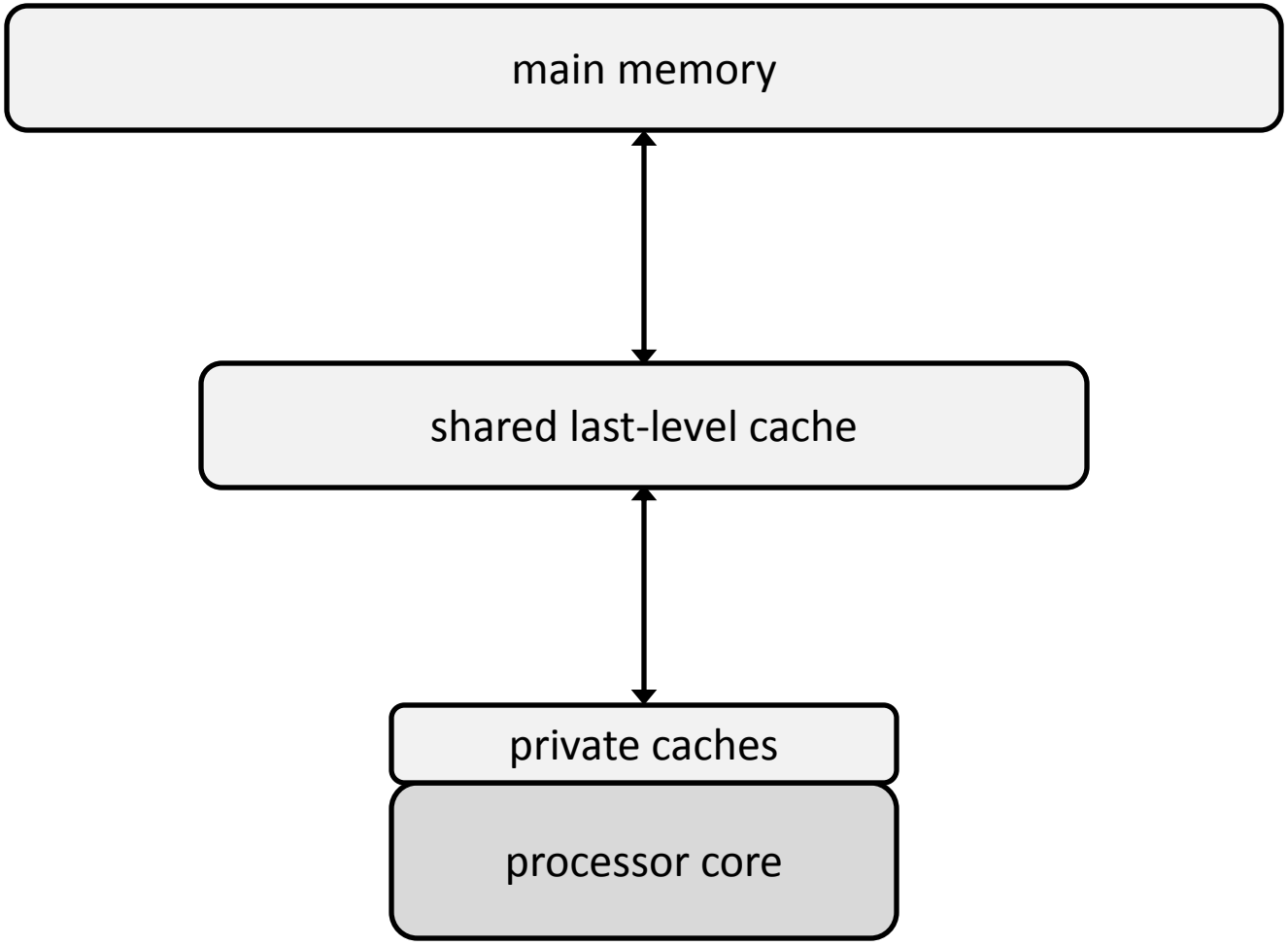
Cache Hierarchy



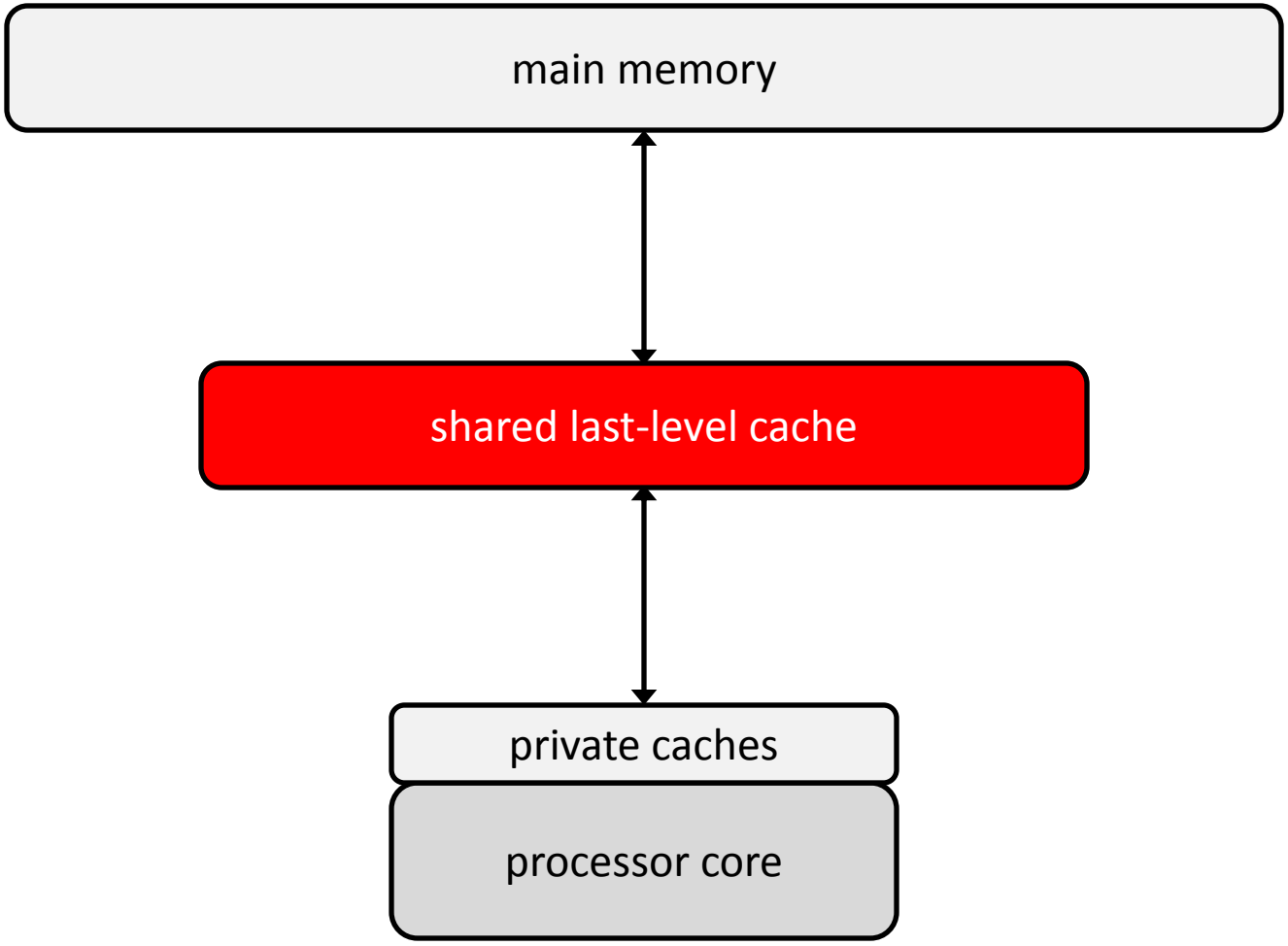
Accessing memory is 10x – 100x greater latency and energy than accessing private cache!

Need hierarchy of large caches…

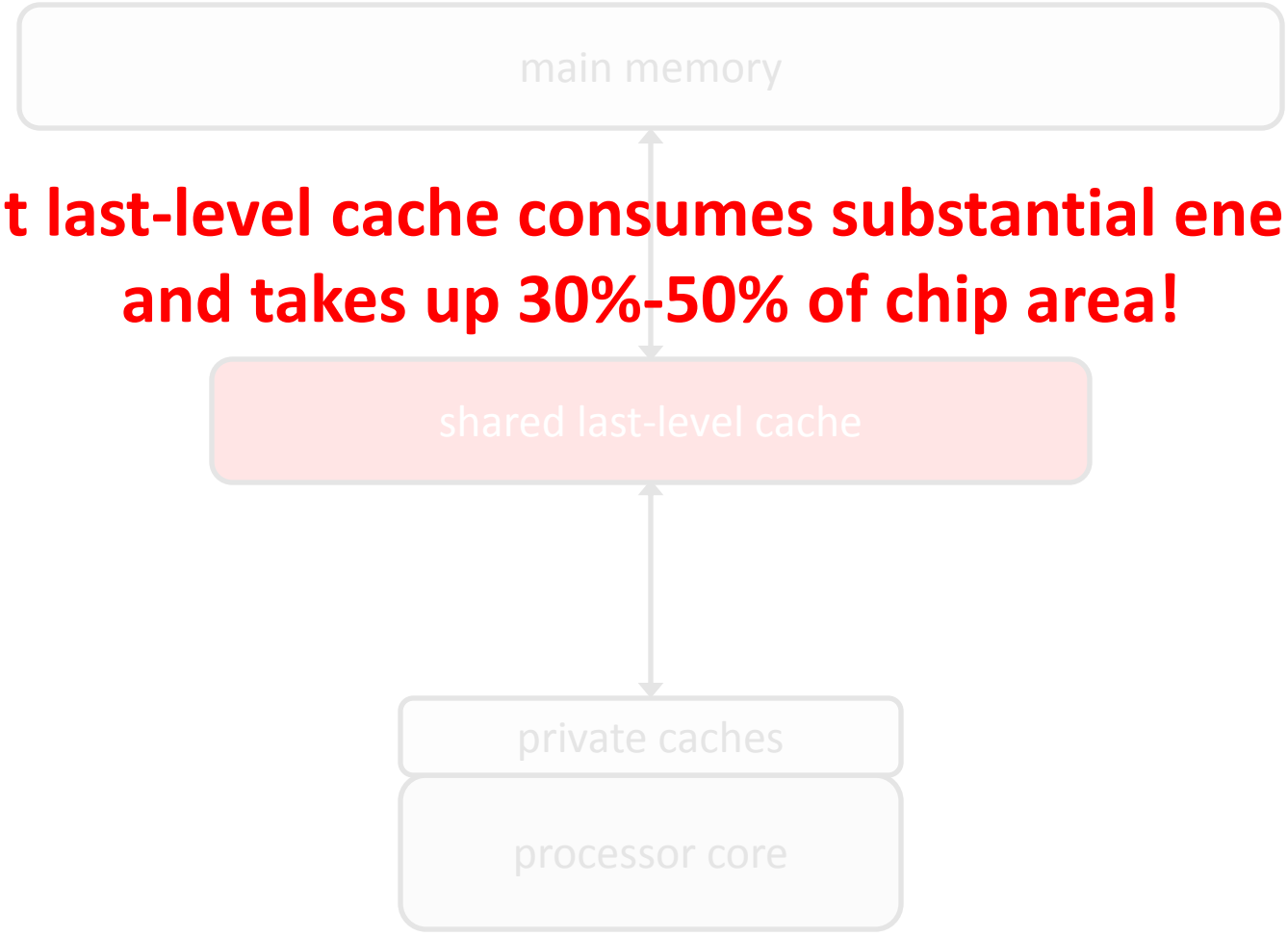
Cache Hierarchy



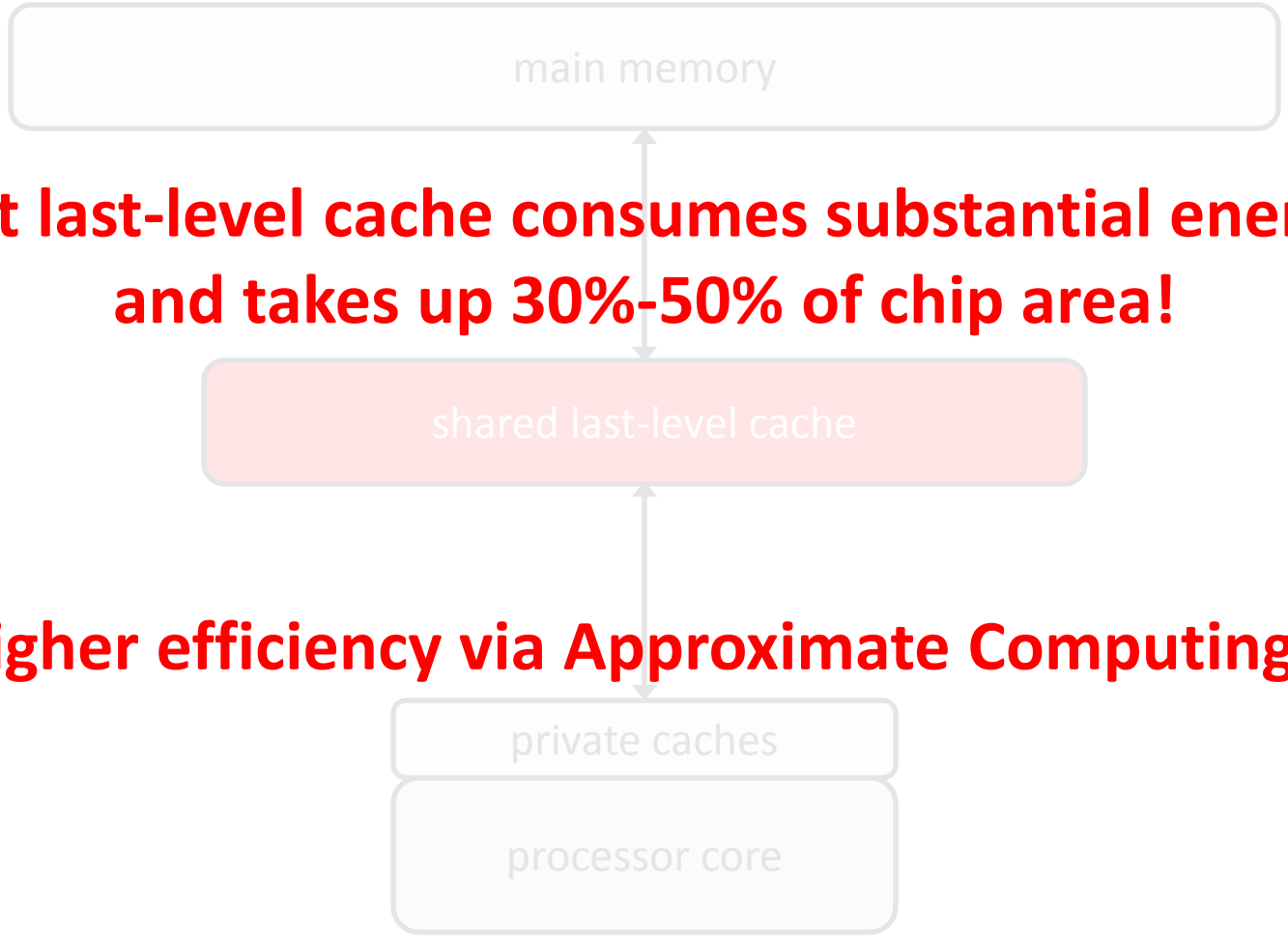
Cache Hierarchy



Cache Hierarchy



Cache Hierarchy



Summary

Doppelgänger Cache:

- Identifies **approximate similarity** in data block values.
  - **77%** cache storage savings of approximable data.

Summary

Doppelgänger Cache:

- Identifies **approximate similarity** in data block values.
  - **77%** cache storage savings of approximable data.
- Effectively compresses storage of approximately similar blocks.
  - **3x** better compression ratio than state-of-the-art techniques.

Summary

Doppelgänger Cache:

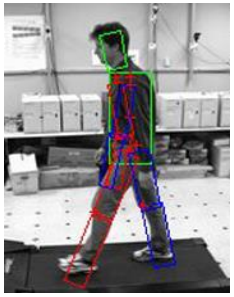
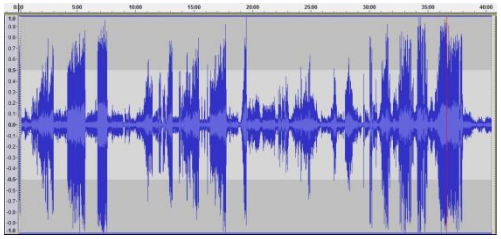
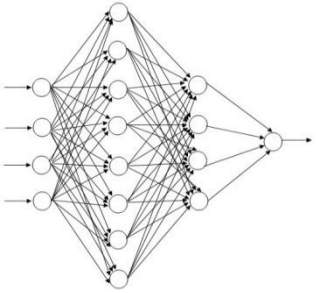
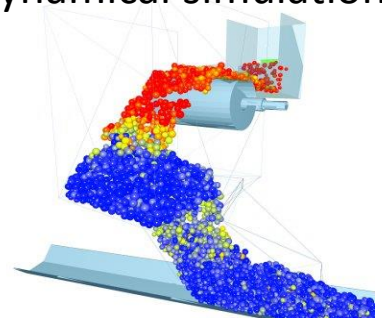
- Identifies **approximate similarity** in data block values.
  - **77%** cache storage savings of approximable data.
- Effectively compresses storage of approximately similar blocks.
  - **3x** better compression ratio than state-of-the-art techniques.
- Significantly reduces area and energy consumption.
  - Reduces total on-chip cache area by **1.36x**.

Outline

- Approximate Computing
  - Approximate Similarity
- Doppelgänger Cache
  - Cache Architecture
  - Similarity Mapping
- Evaluation

Approximate Computing

Not all data/computations need to be precise.

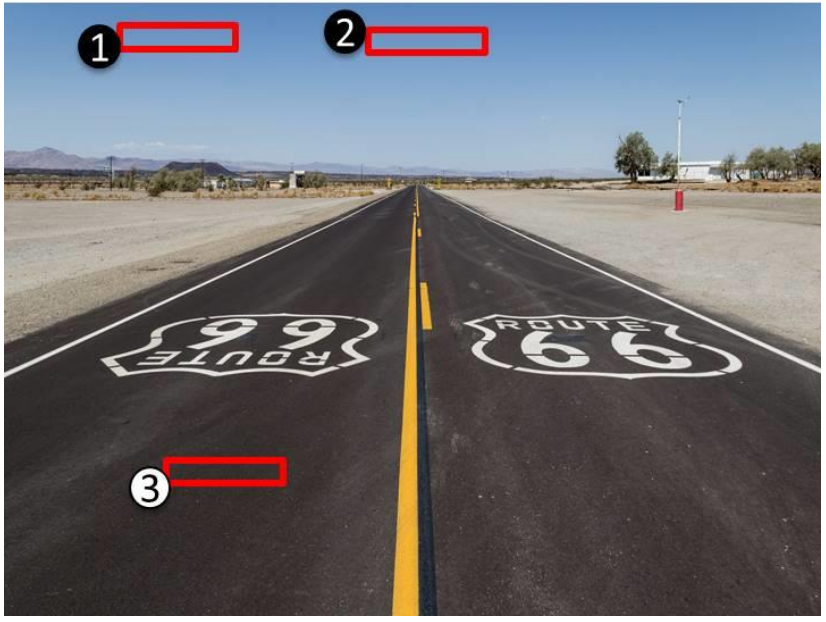
Data mining  http://www.zentut.com/	Computer vision  http://www.cc.gatech.edu/~cnieto6/	Audio and video processing  http://themusicparlour.blogspot.ca/
Gaming  http://www.businessweek.com/	Machine learning  http://www.analyticbridge.com/	Dynamical simulation  http://www.scientific-computing.com/

Approximate Similarity

Two data blocks are ***approximately similar*** (i.e., ***doppelgängers***) if replacing the values of one with the other still results in acceptable application output in the end.

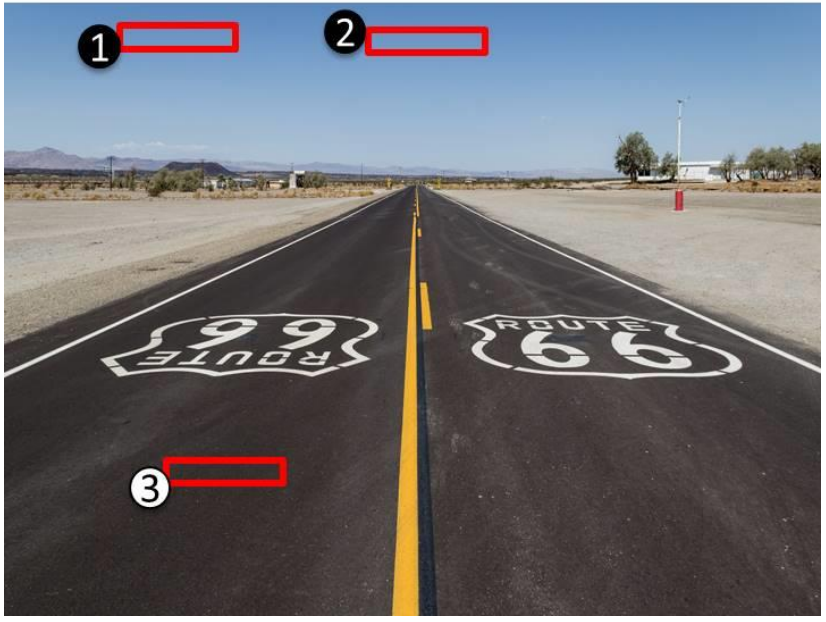
Approximate Similarity

Two data blocks are ***approximately similar*** (i.e., ***doppelgängers***) if replacing the values of one with the other still results in acceptable application output in the end.



Approximate Similarity

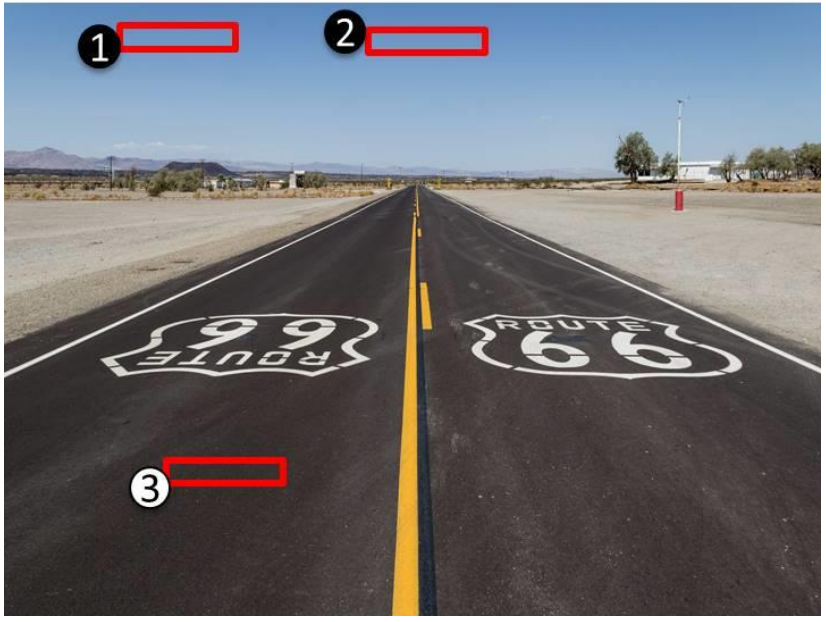
Two data blocks are ***approximately similar*** (i.e., ***doppelgängers***) if replacing the values of one with the other still results in acceptable application output in the end.



1	92	131	183	91	132	186
2	90	131	185	93	133	184
3	35	31	29	43	38	37

Approximate Similarity

Two data blocks are ***approximately similar*** (i.e., ***doppelgängers***) if replacing the values of one with the other still results in acceptable application output in the end.

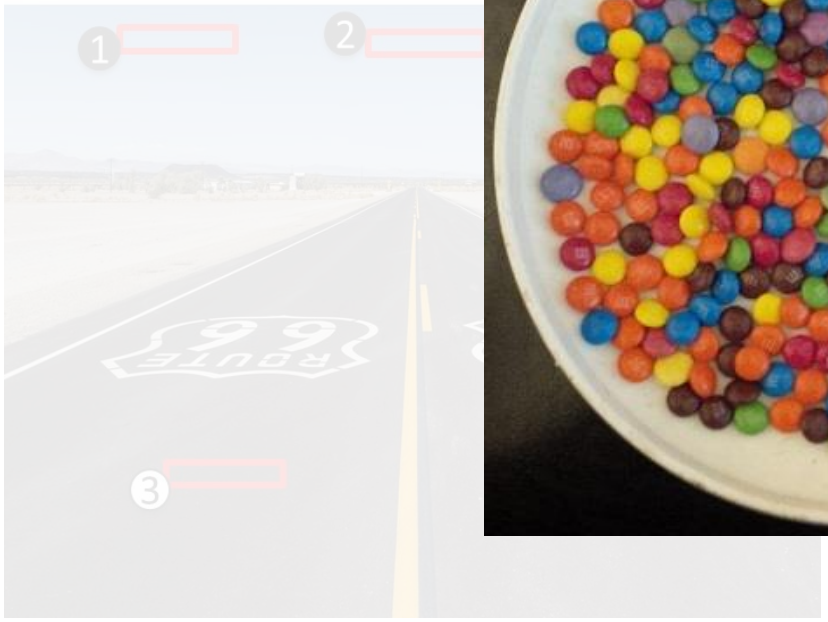


1	92	131	183	91	132	186
2	90	131	185	93	133	184
3	35	31	29	43	38	37

approximately similar

Approximate Similarity

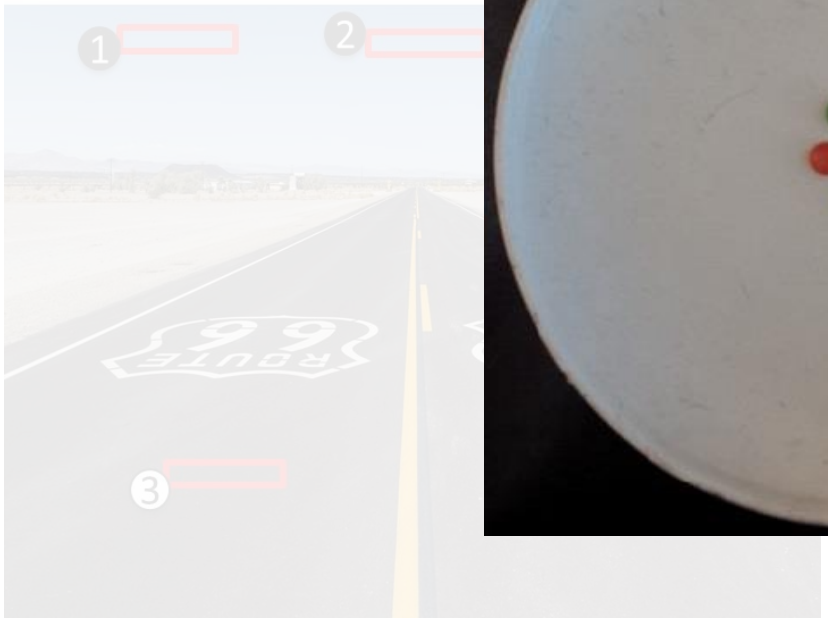
Two data blocks are ***approximately similar*** (i.e., ***doppelgängers***) if replacing the val results in acceptable applica



91	132	186	approximately similar
93	133	184	
43	38	37	

Approximate Similarity

Two data blocks are ***approximately similar*** (i.e., ***doppelgängers***) if replacing the val results in acceptable applica



91 132 186

93 133 184

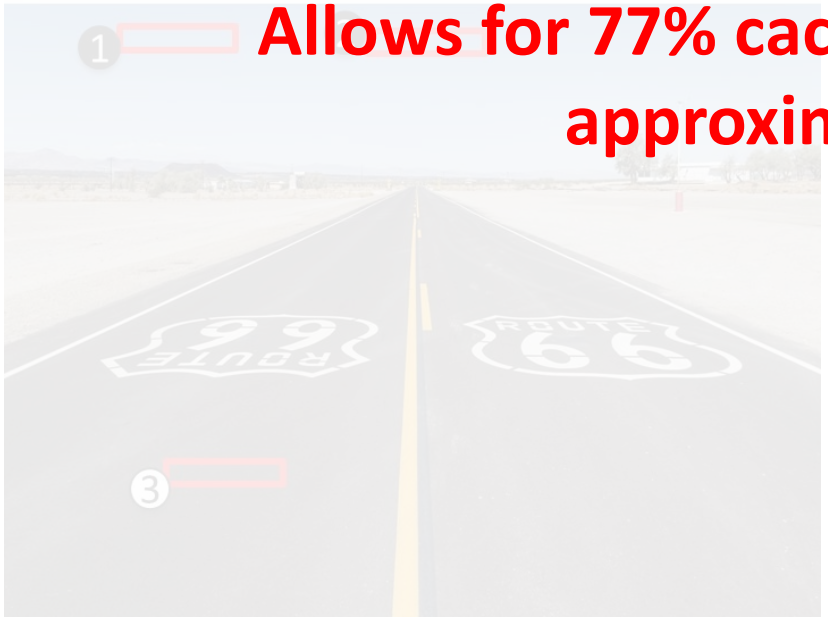
43 38 37

approximately similar

Approximate Similarity

Two data blocks are ***approximately similar*** (i.e., ***doppelgängers***) if replacing the values of one with the other still results in acceptable application output in the end.

Allows for 77% cache storage savings of approximable data!

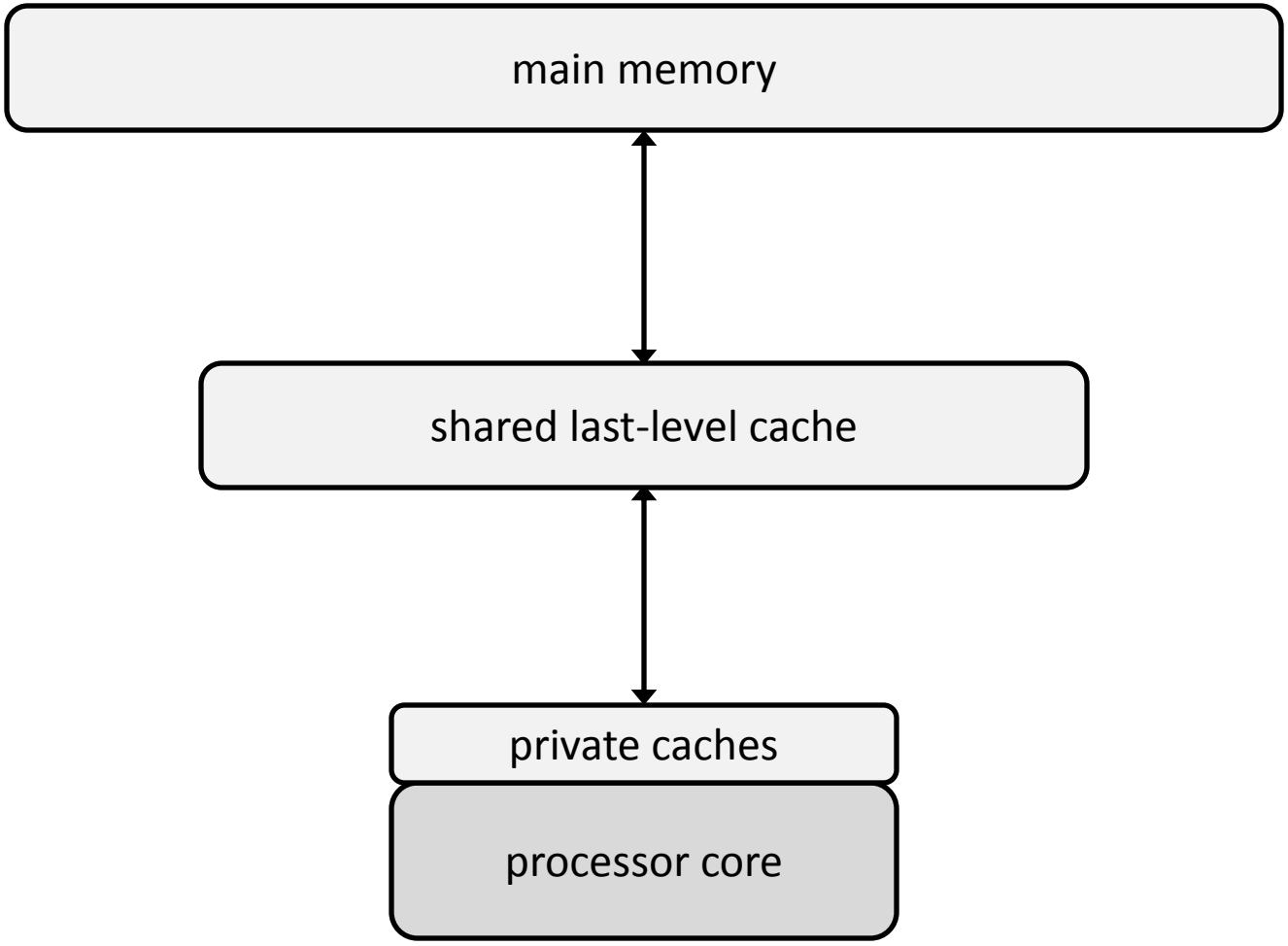


1	92	131	183	91	132	186	approximately similar
2	90	131	185	93	133	184	
3	35	31	29	43	38	37	

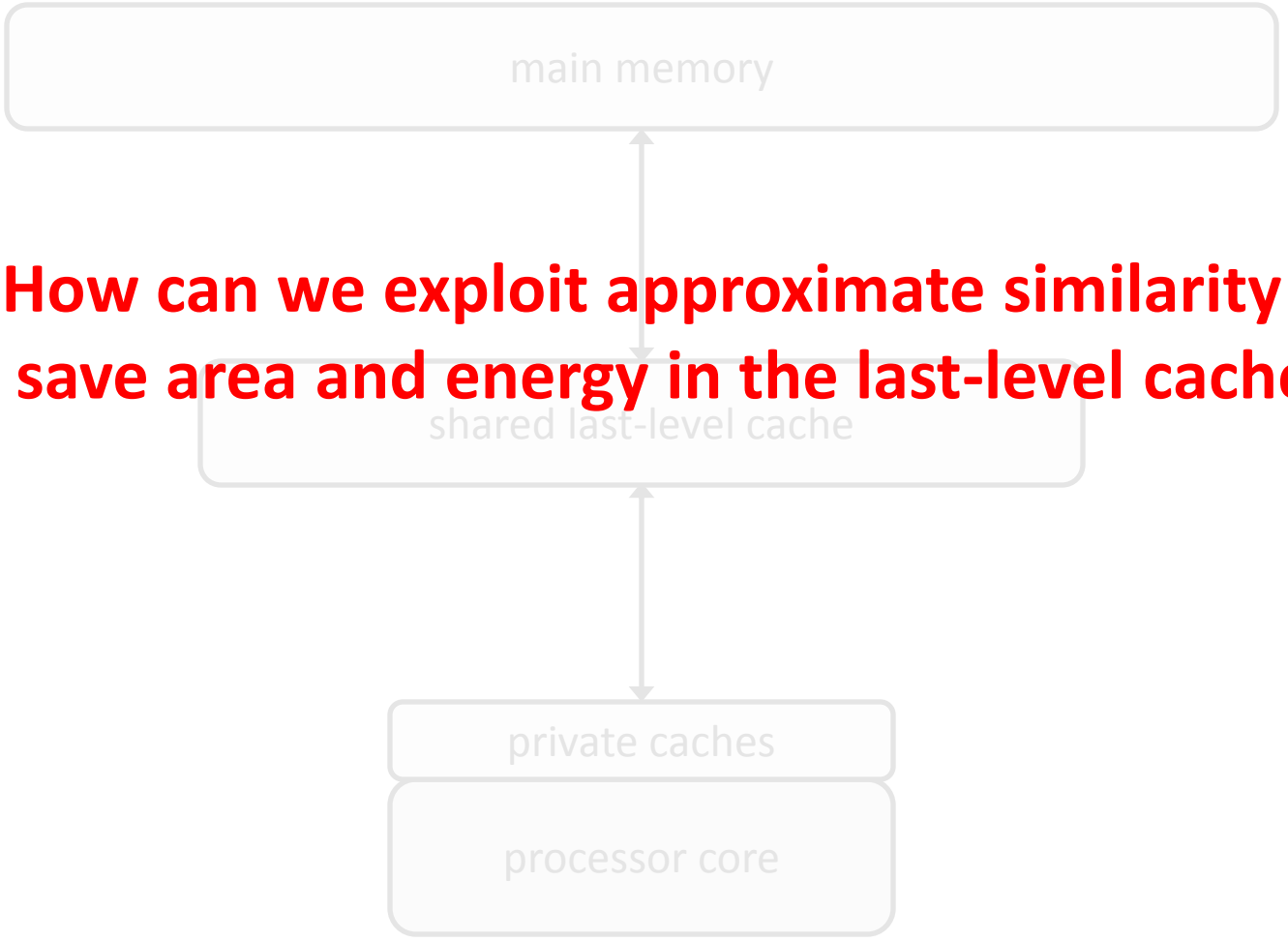
Outline

- Approximate Computing
  - Approximate Similarity
- **Doppelgänger Cache**
  - **Cache Architecture**
  - **Similarity Mapping**
- Evaluation

Doppelgänger Cache

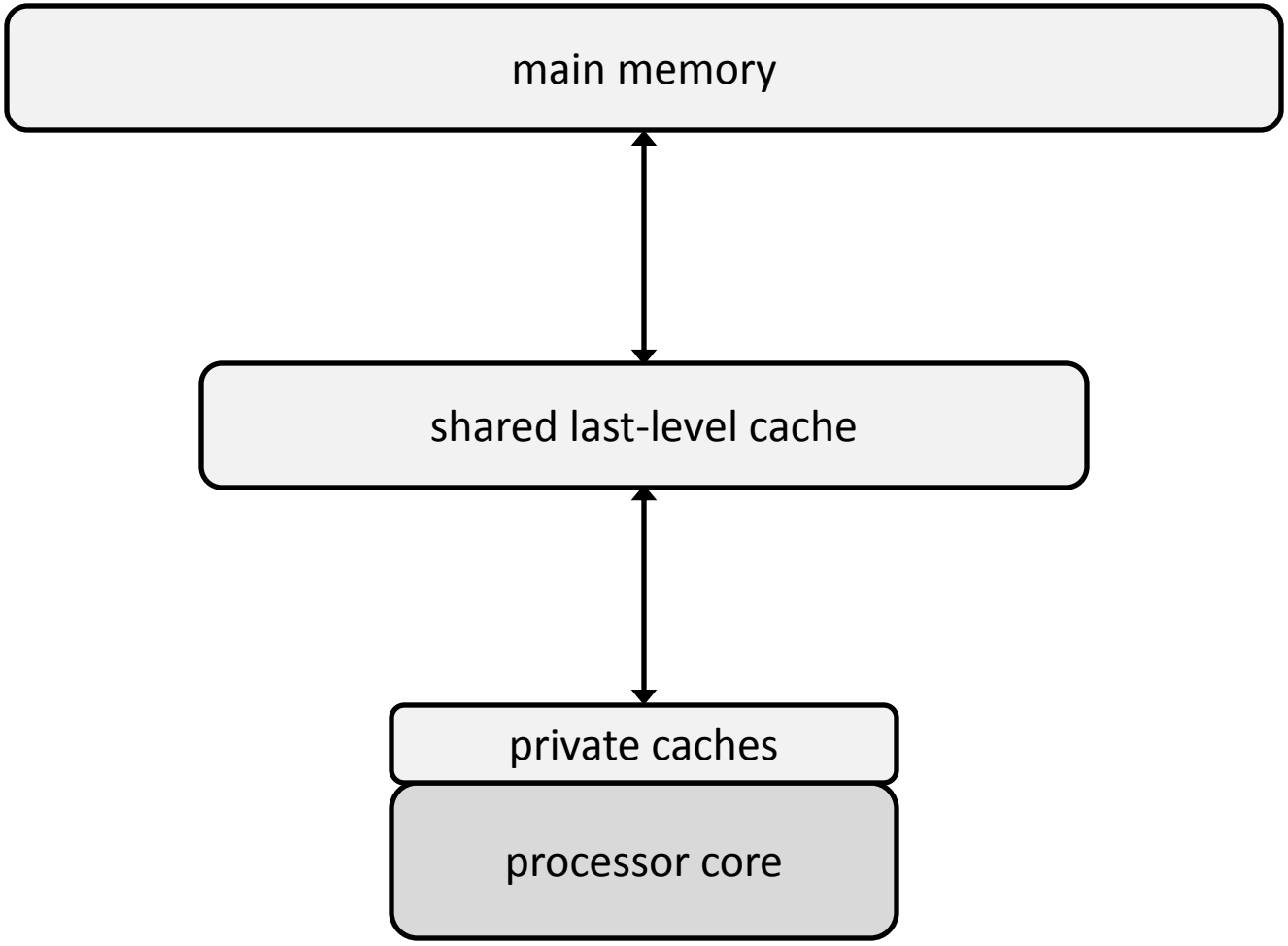


Doppelgänger Cache

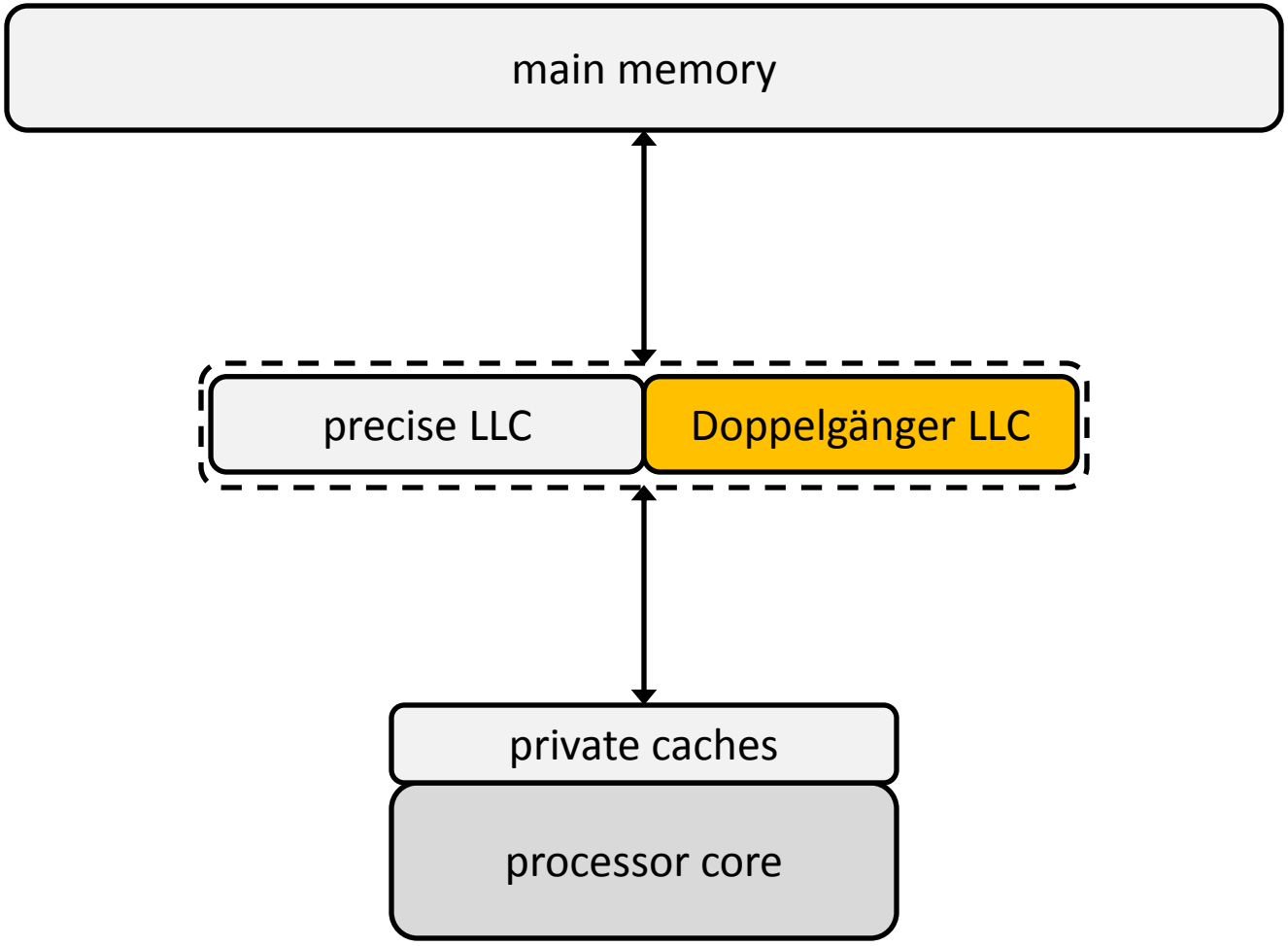


How can we exploit approximate similarity to save area and energy in the last-level cache?

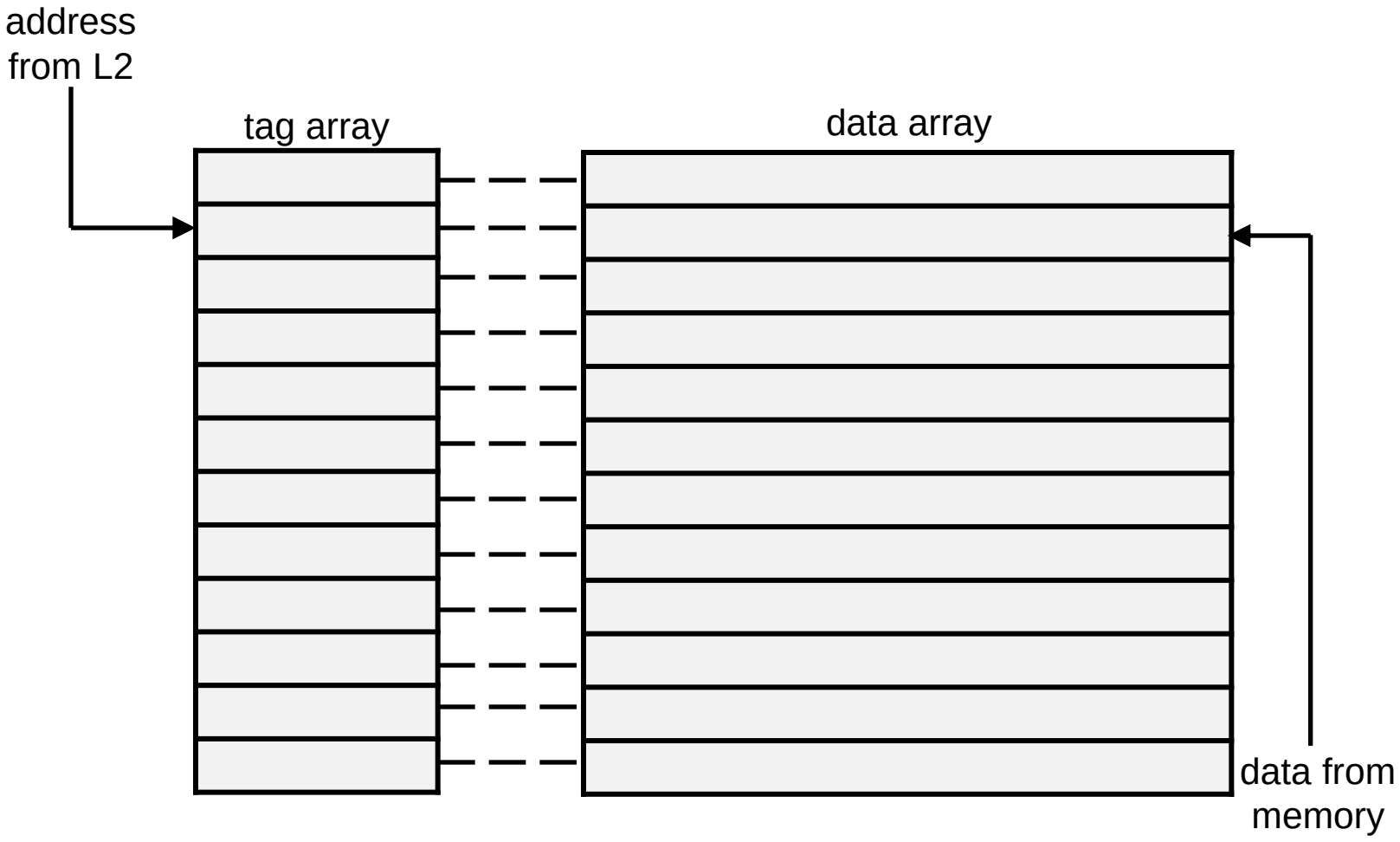
Doppelgänger Cache



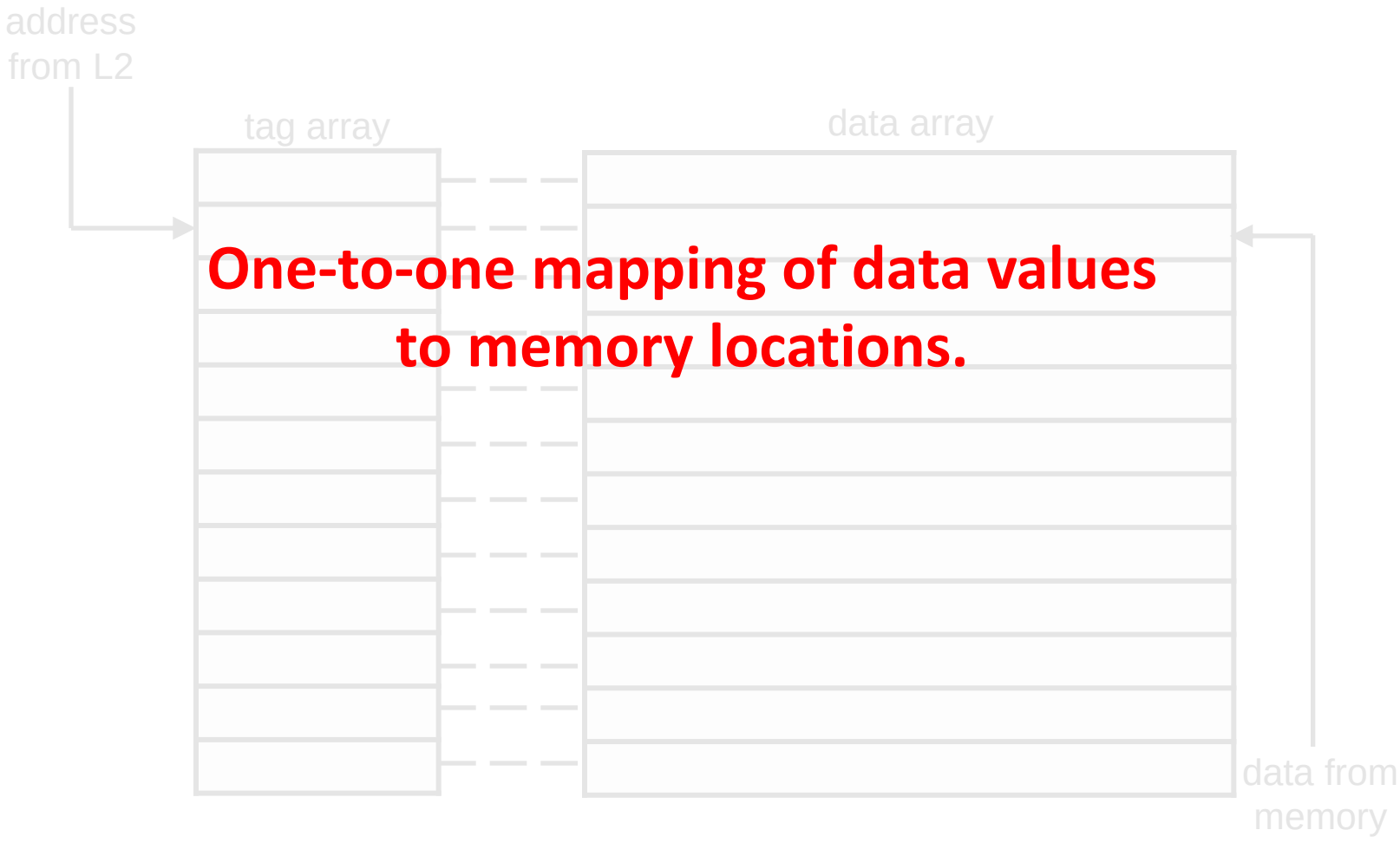
Doppelgänger Cache



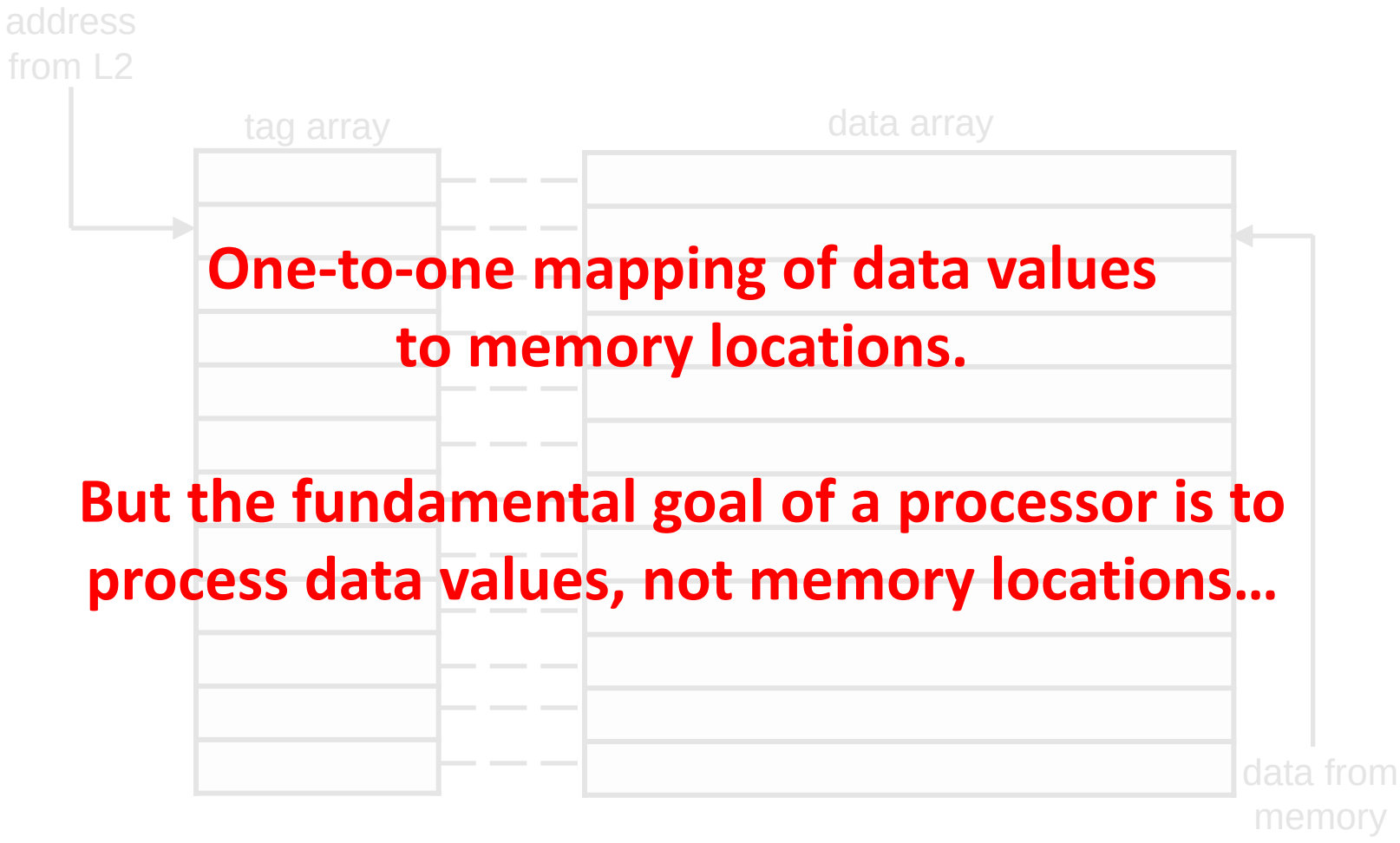
Conventional Cache



Conventional Cache



Conventional Cache



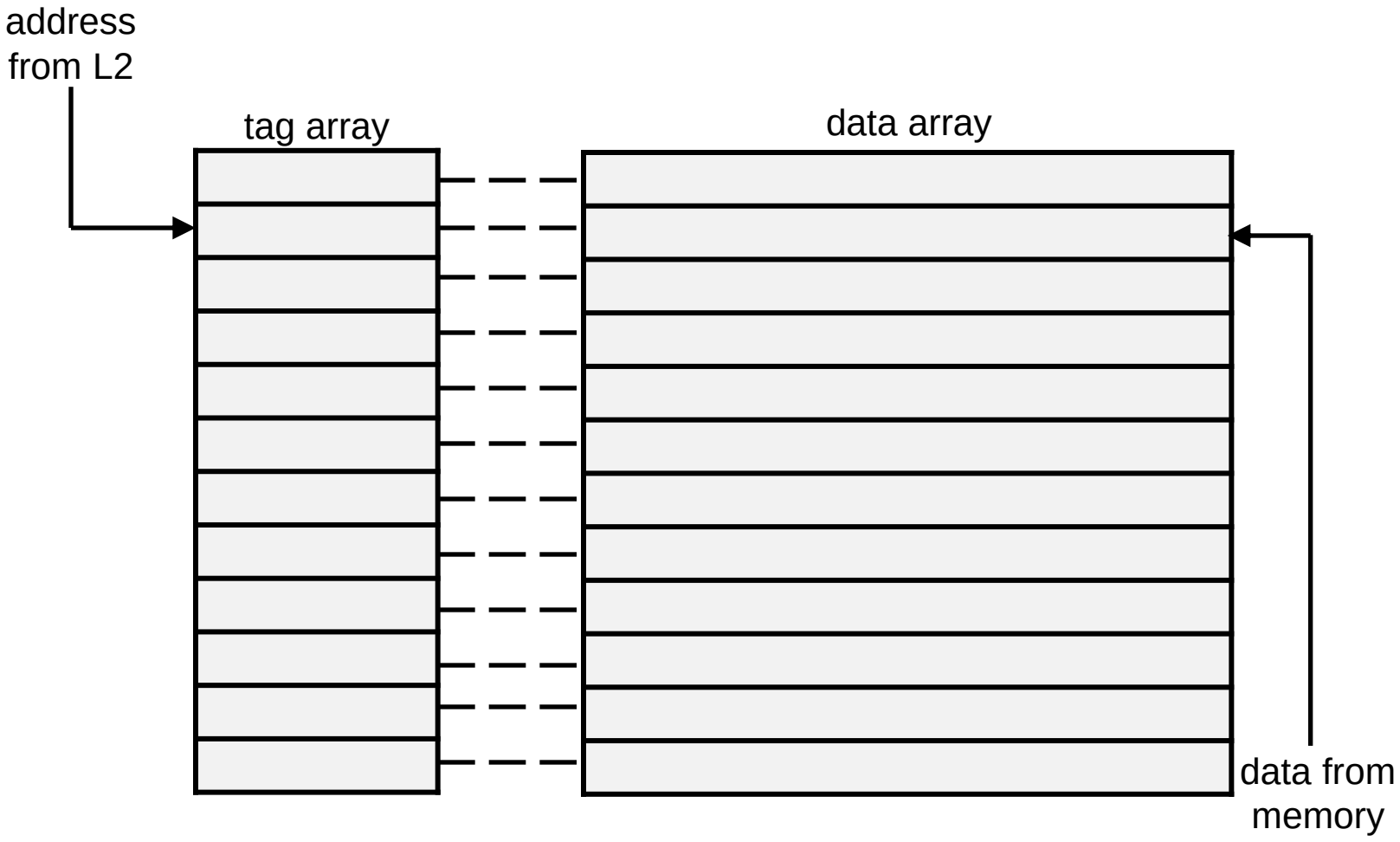
One-to-one mapping of data values to memory locations.

But the fundamental goal of a processor is to process data values, not memory locations…

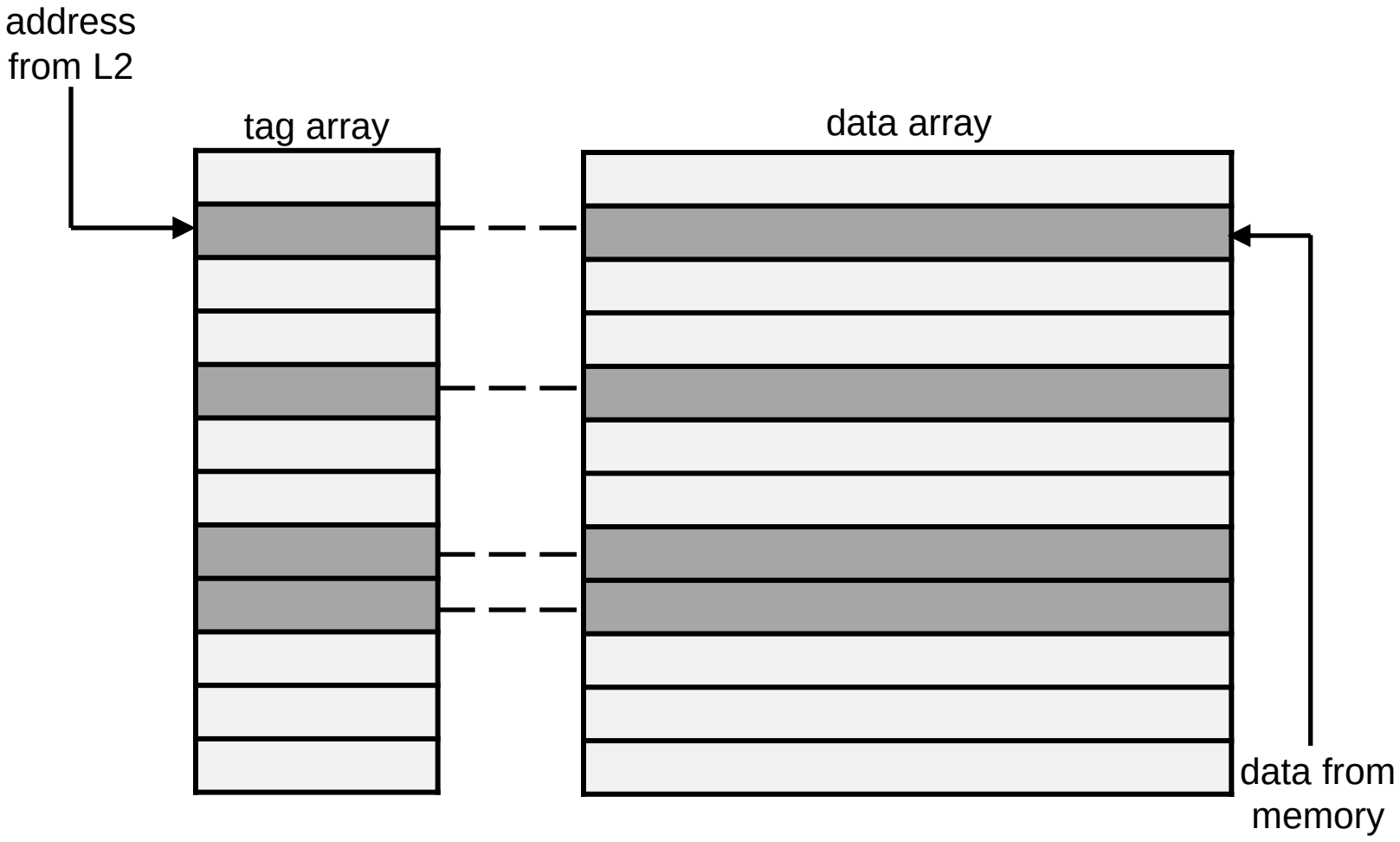
Conventional Cache

Conventional Cache

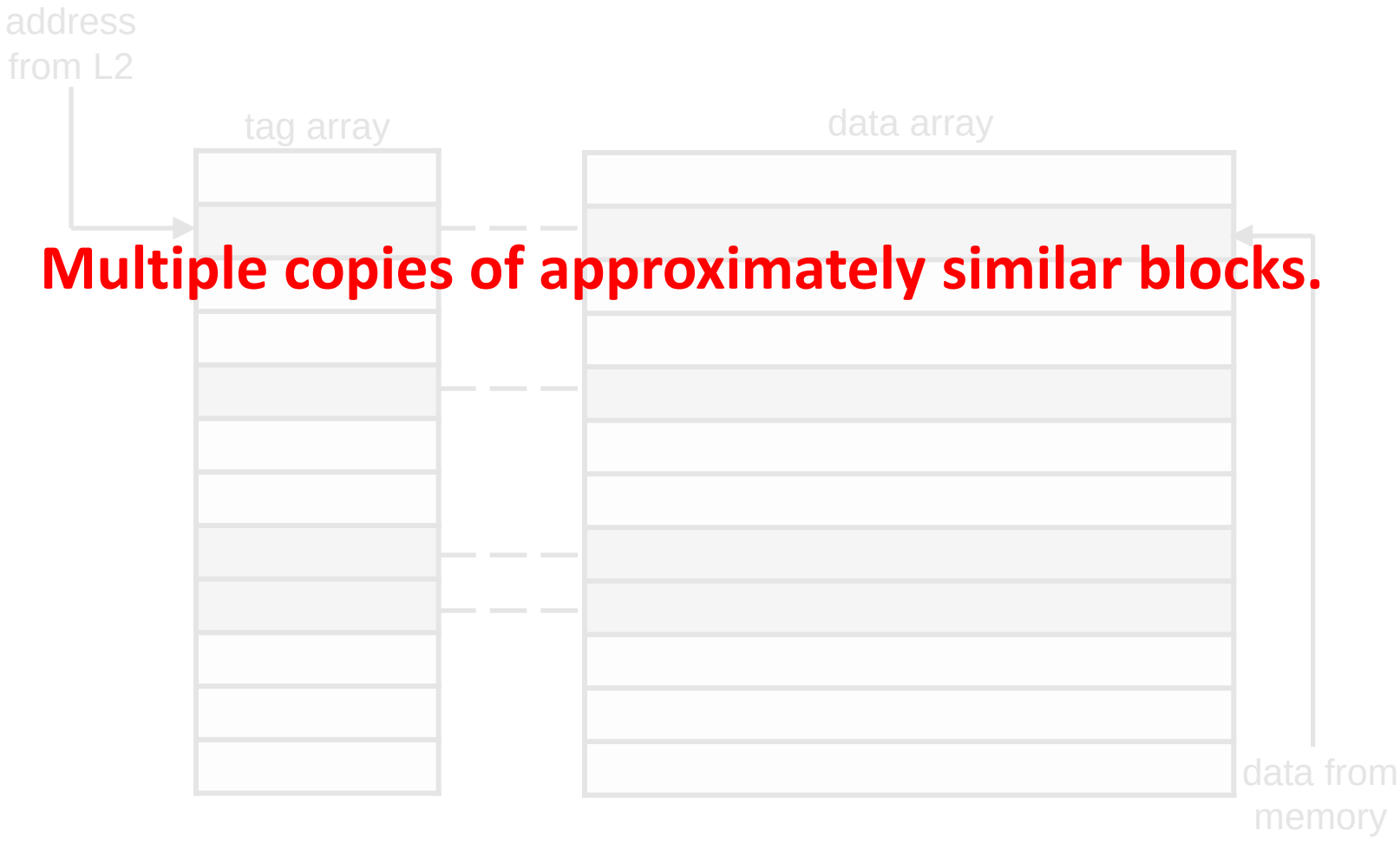
Conventional Cache



Conventional Cache

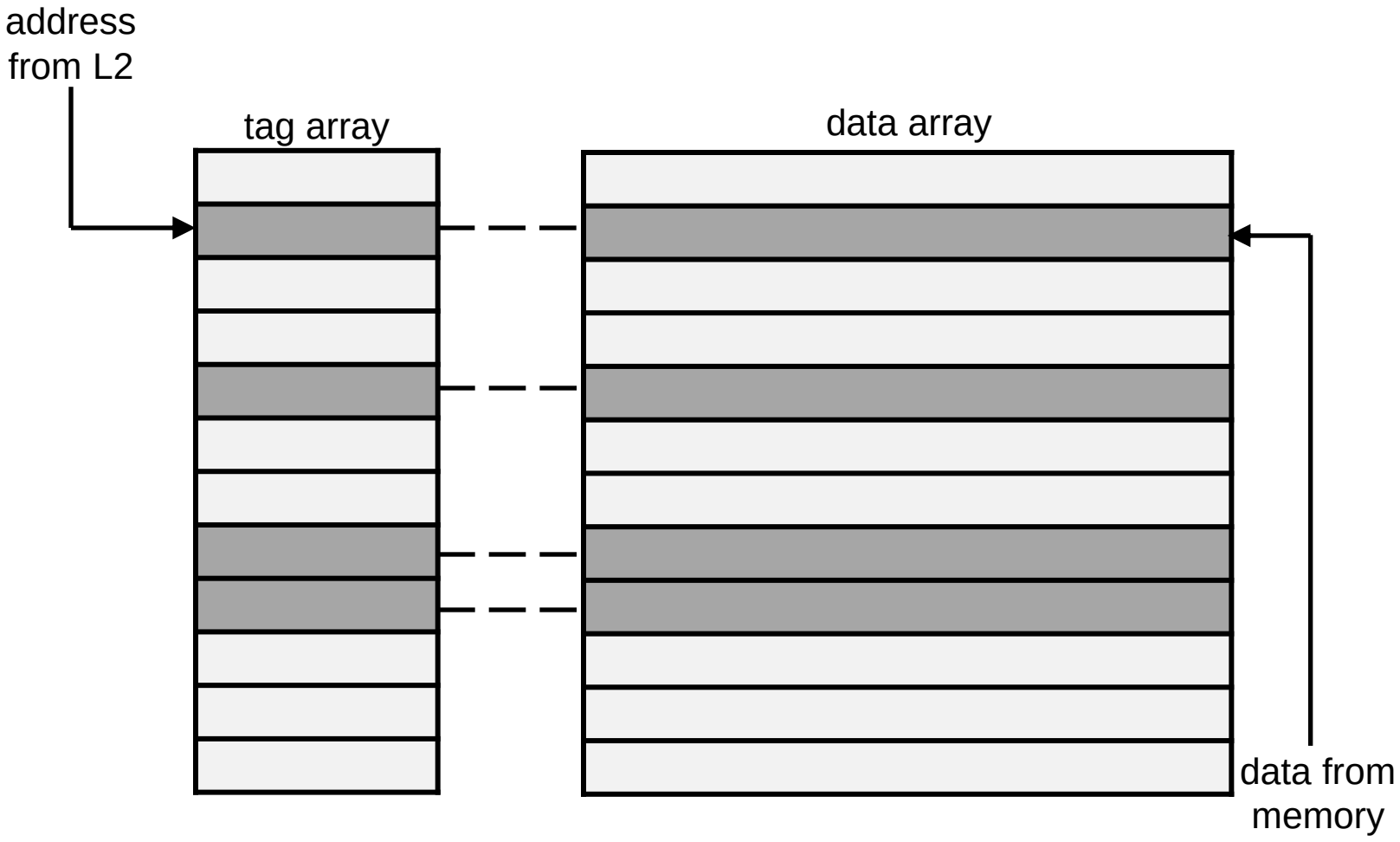


Conventional Cache

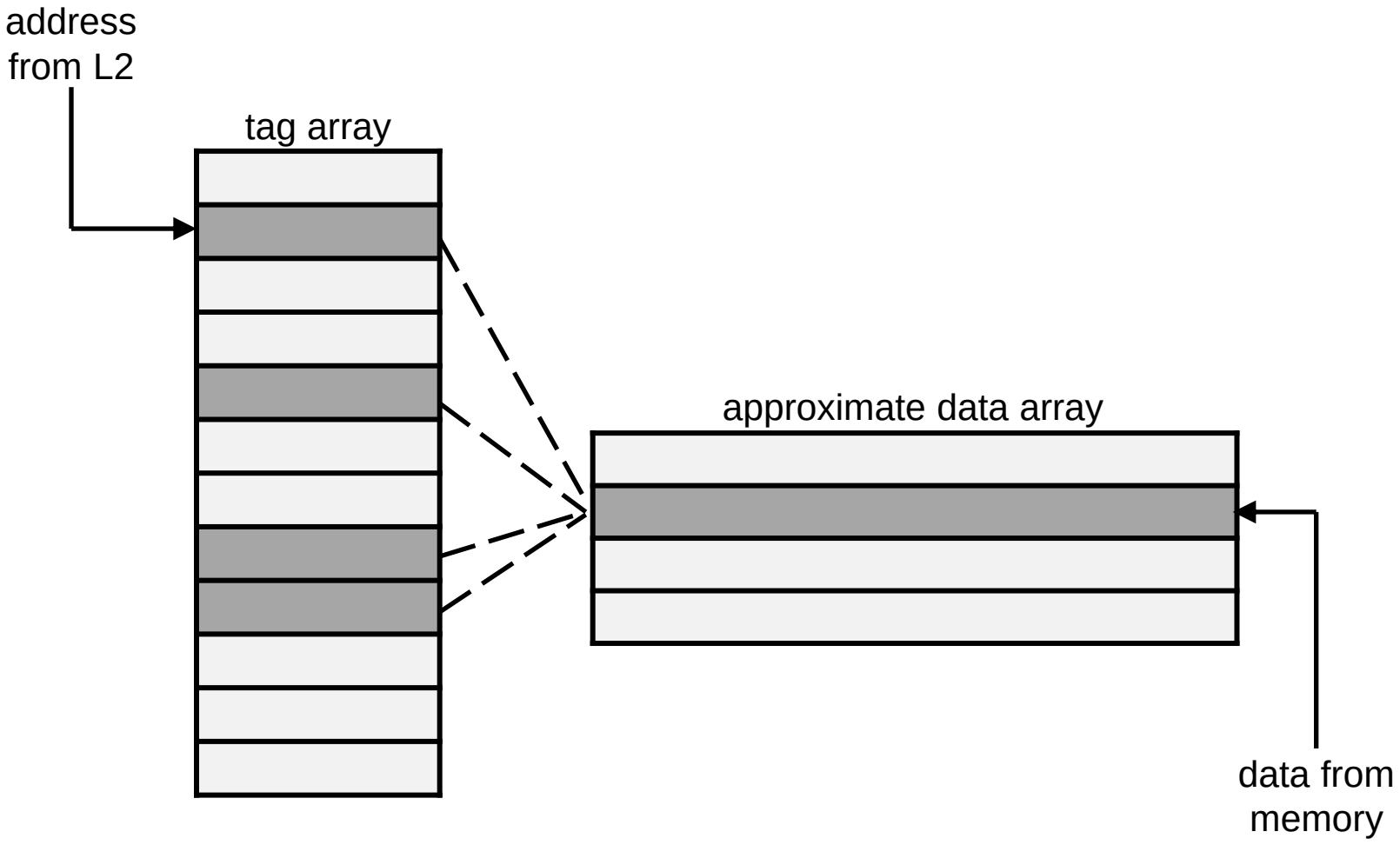


Multiple copies of approximately similar blocks.

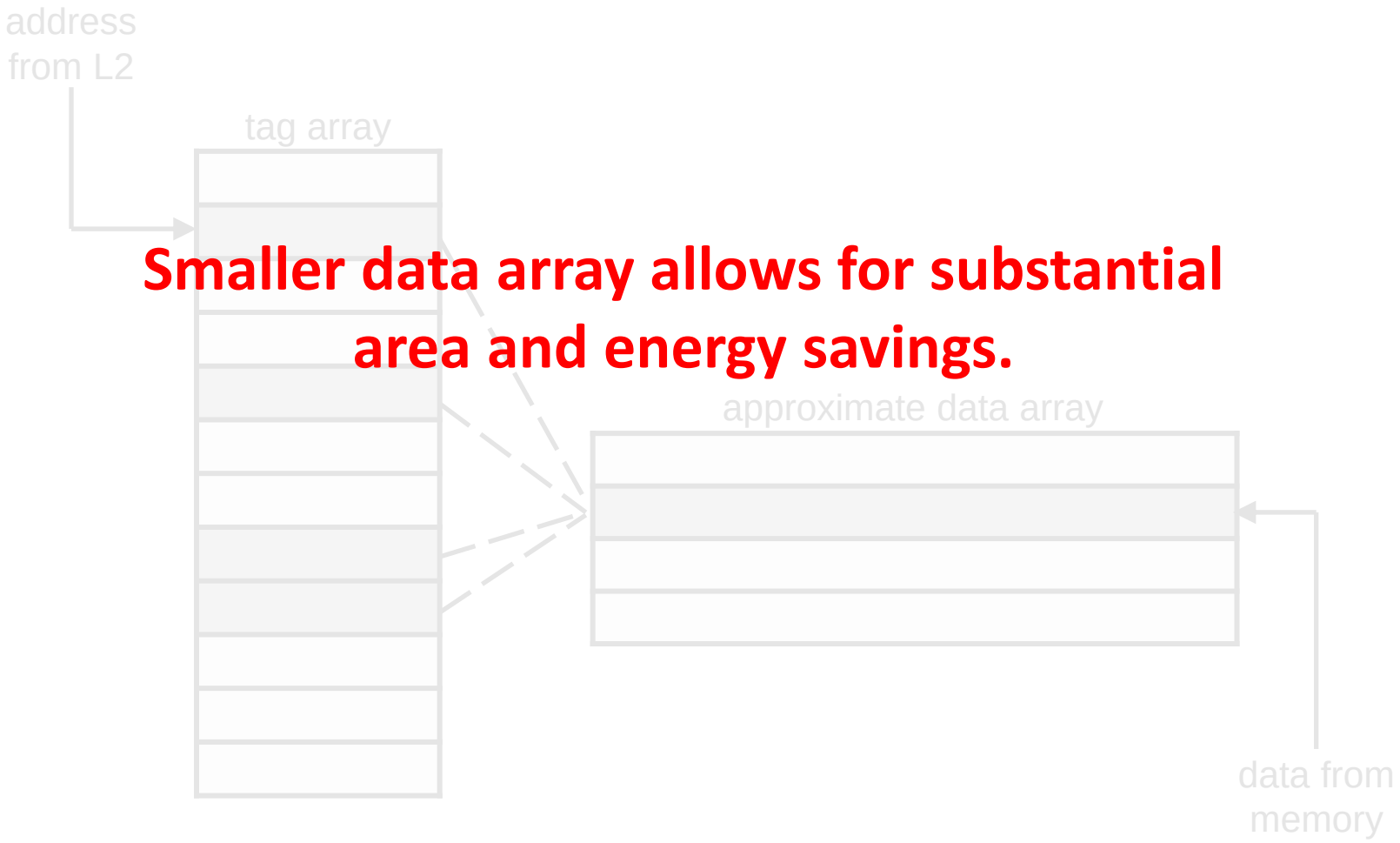
Conventional Cache



Doppelgänger Cache

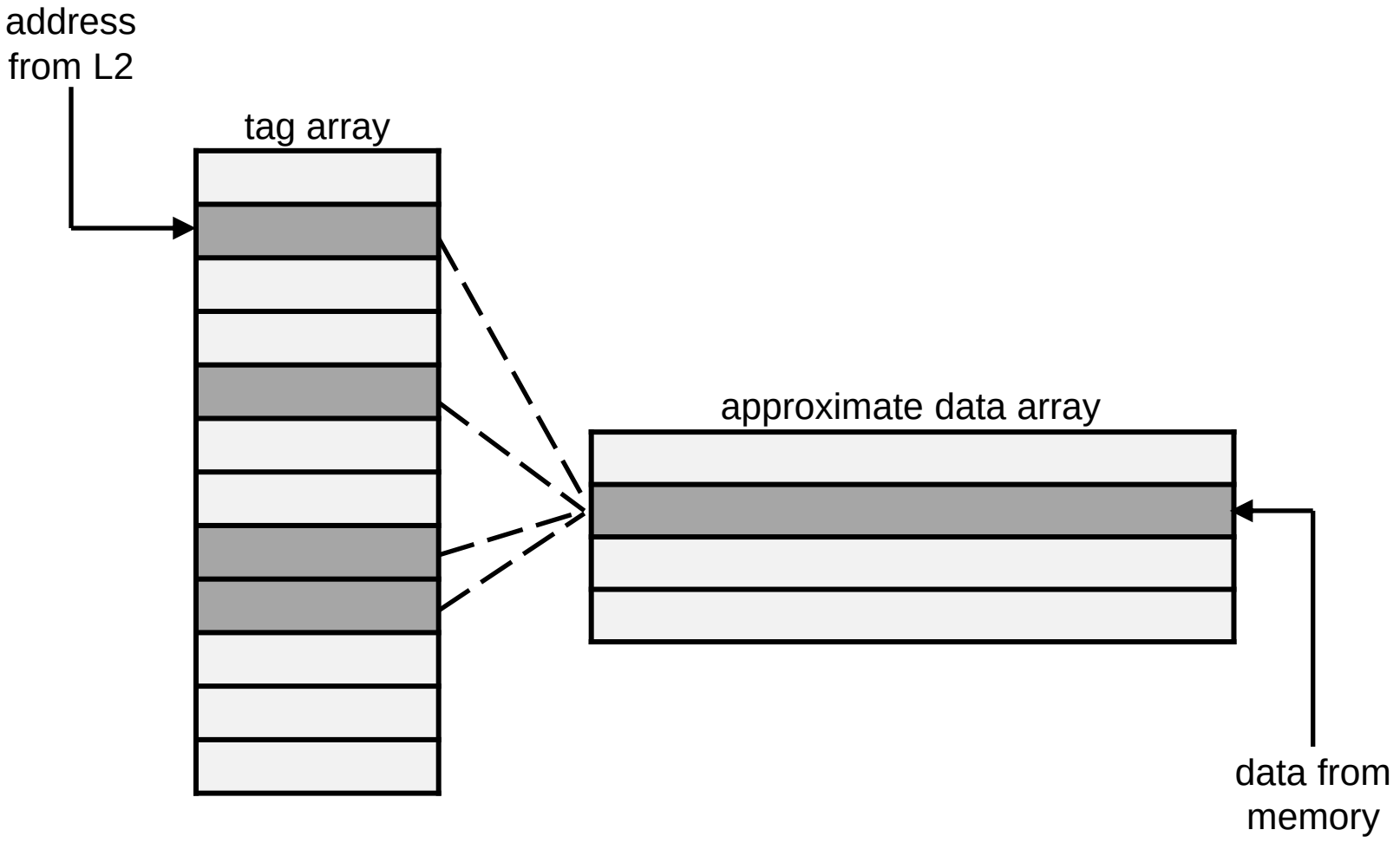


Doppelgänger Cache

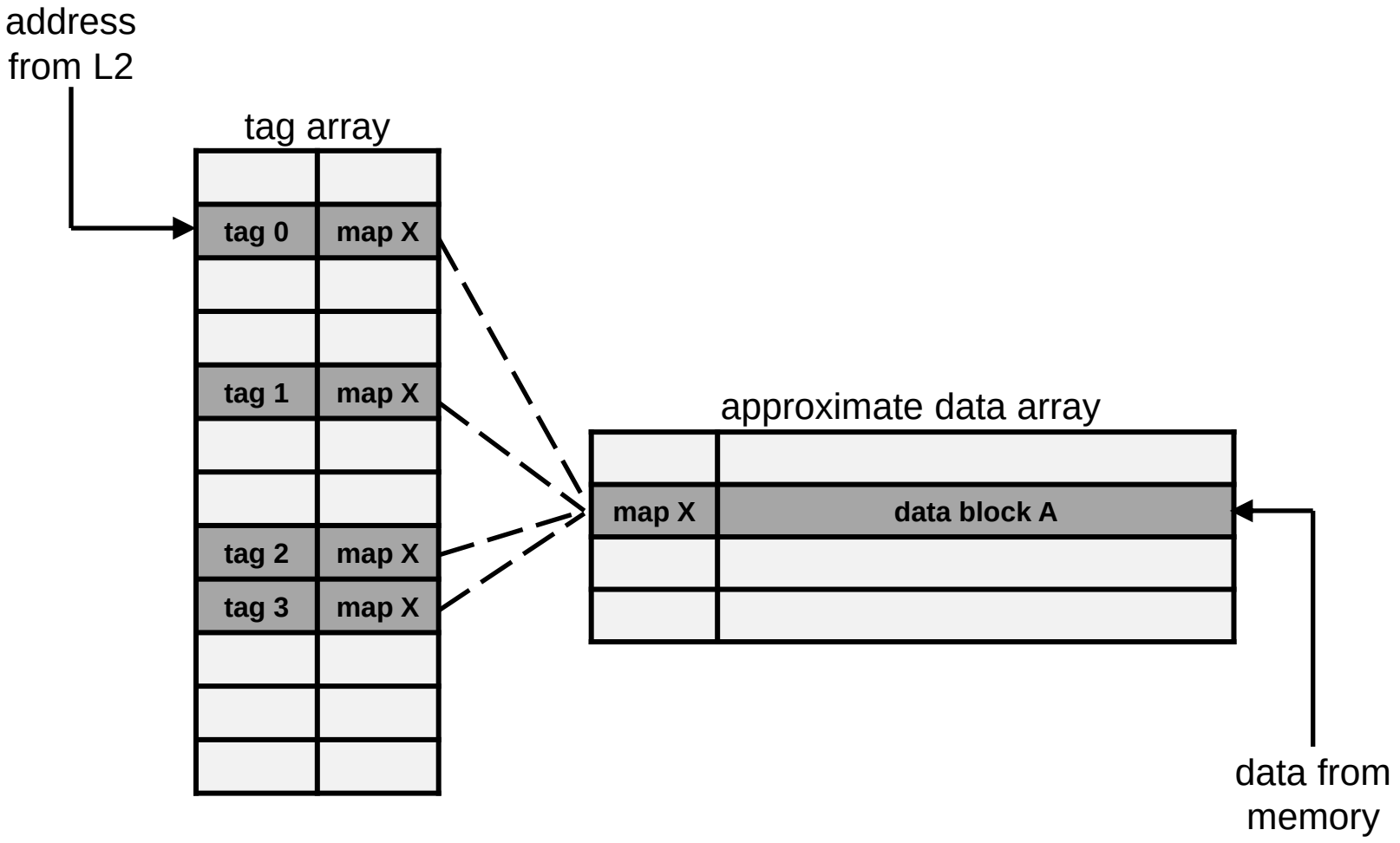


Smaller data array allows for substantial area and energy savings.

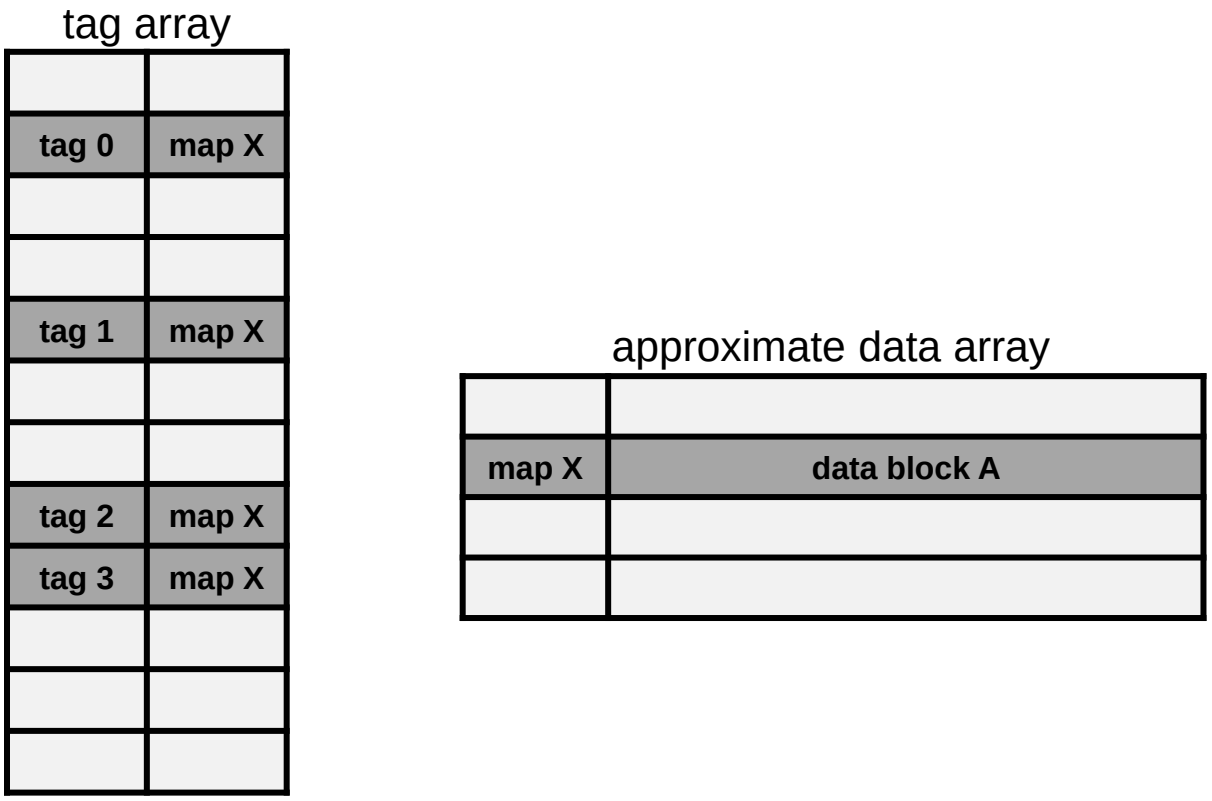
Doppelgänger Cache



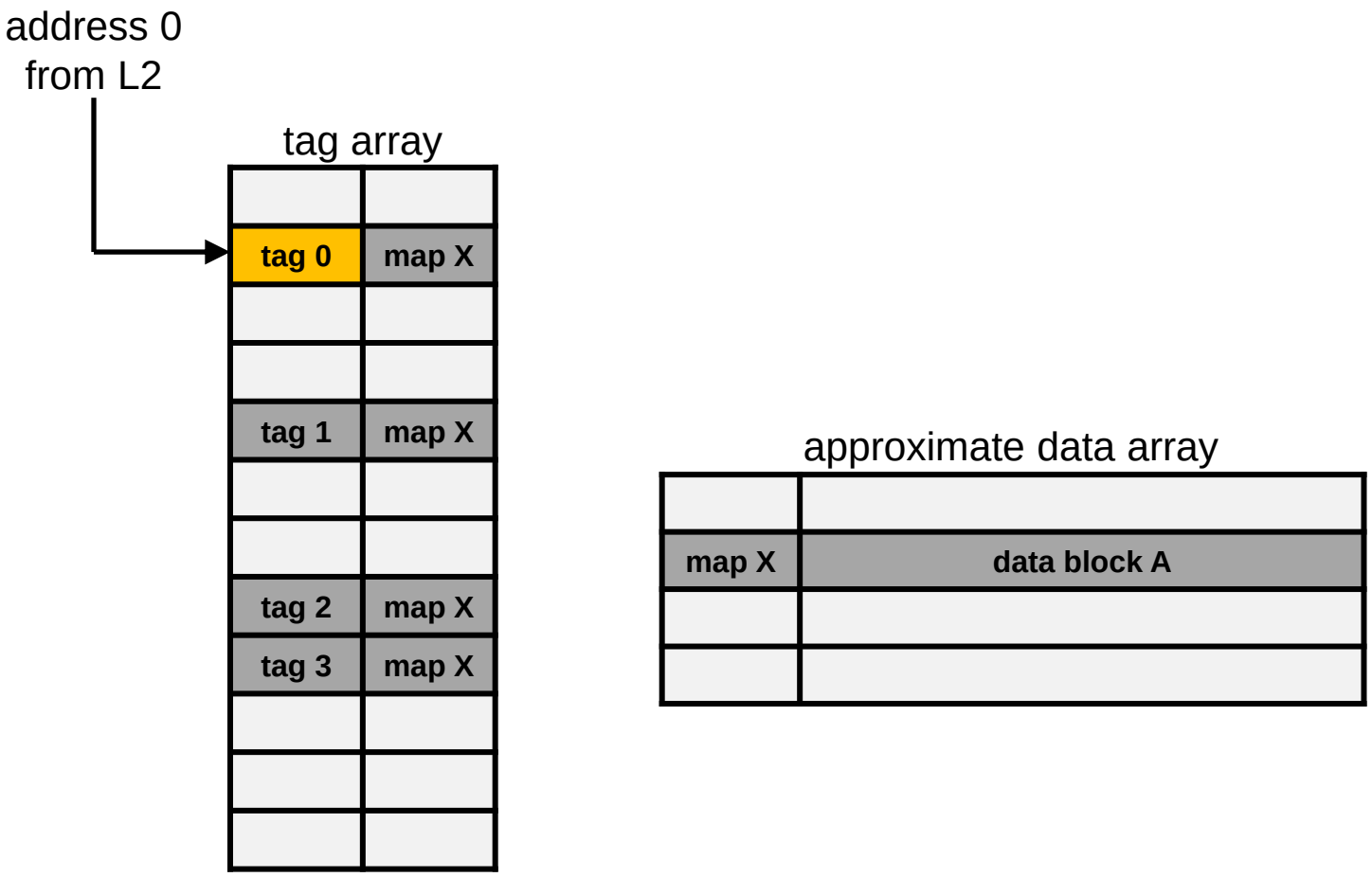
Doppelgänger Cache



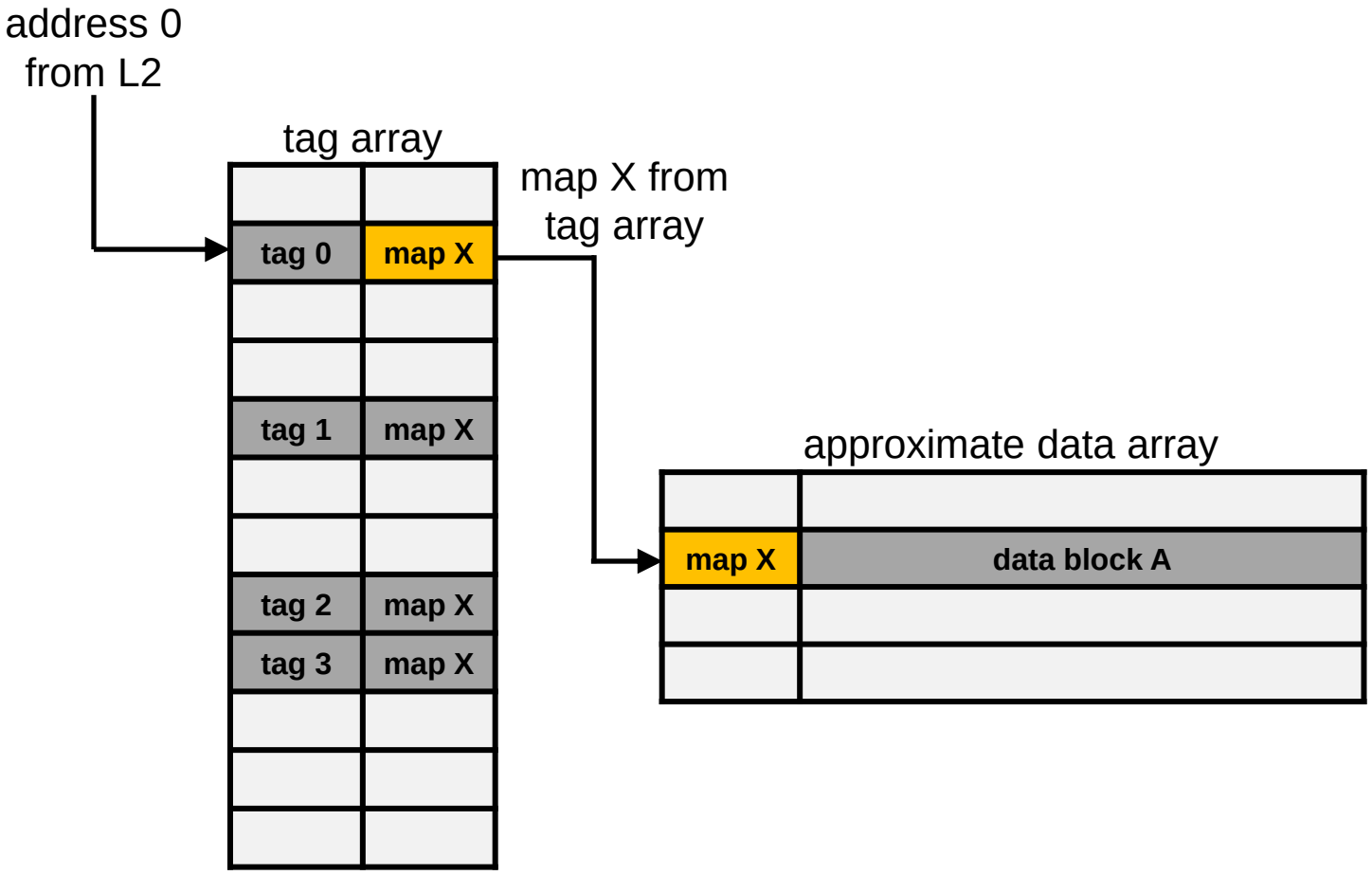
Doppelgänger Cache - Lookups



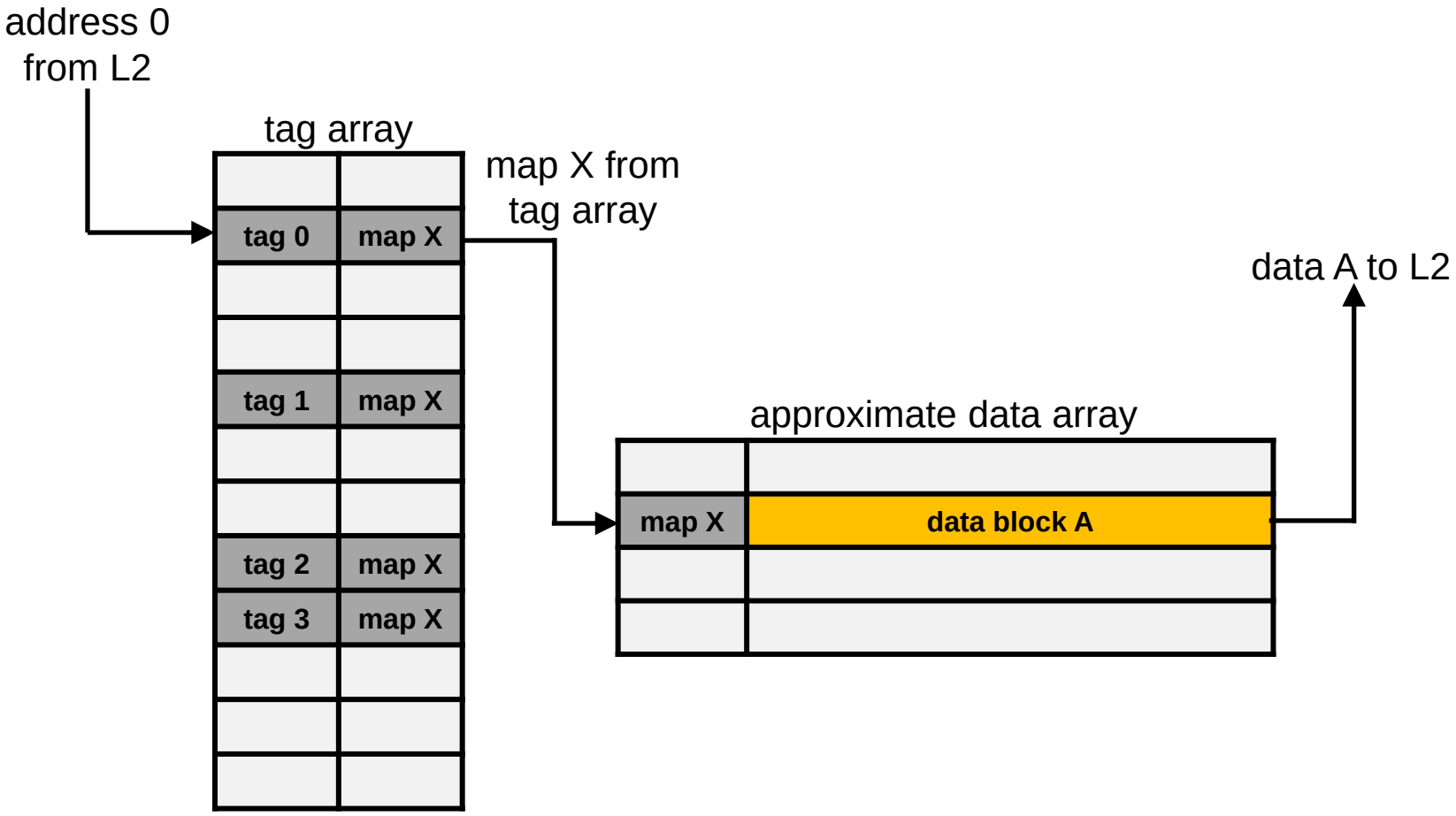
Doppelgänger Cache - Lookups



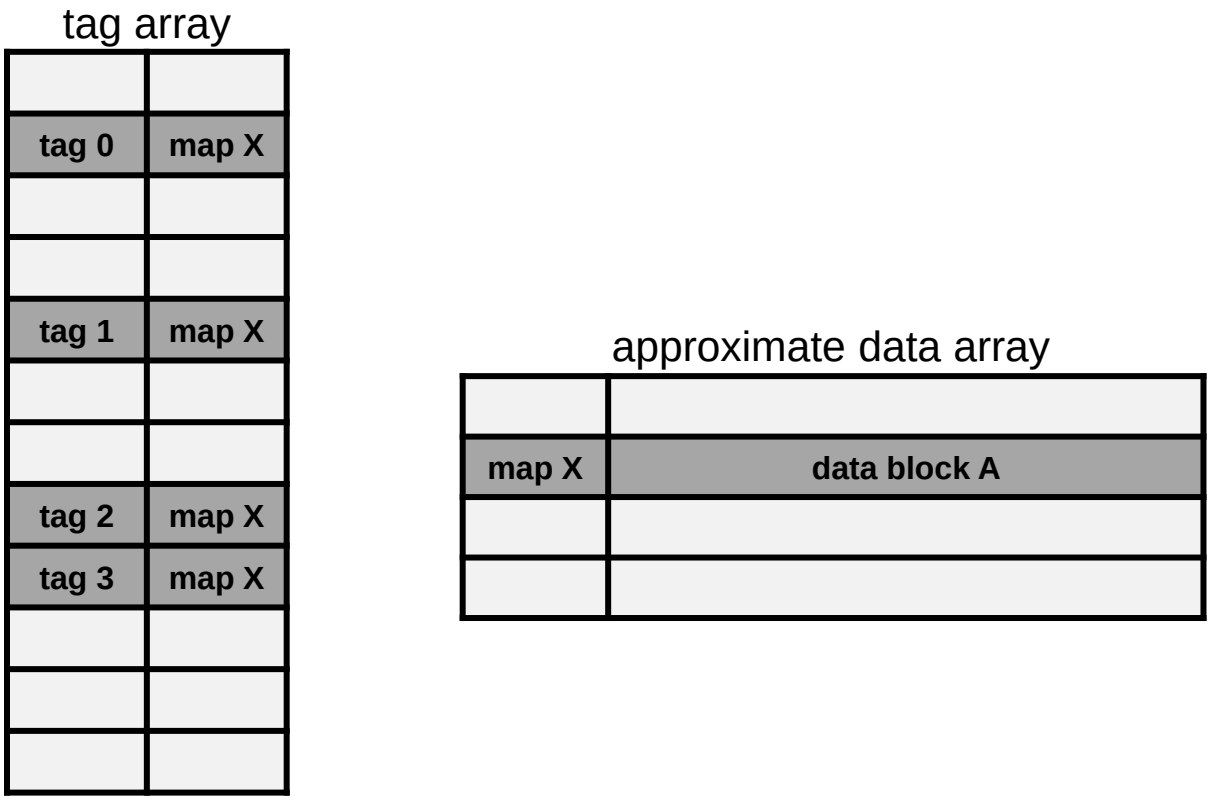
Doppelgänger Cache - Lookups



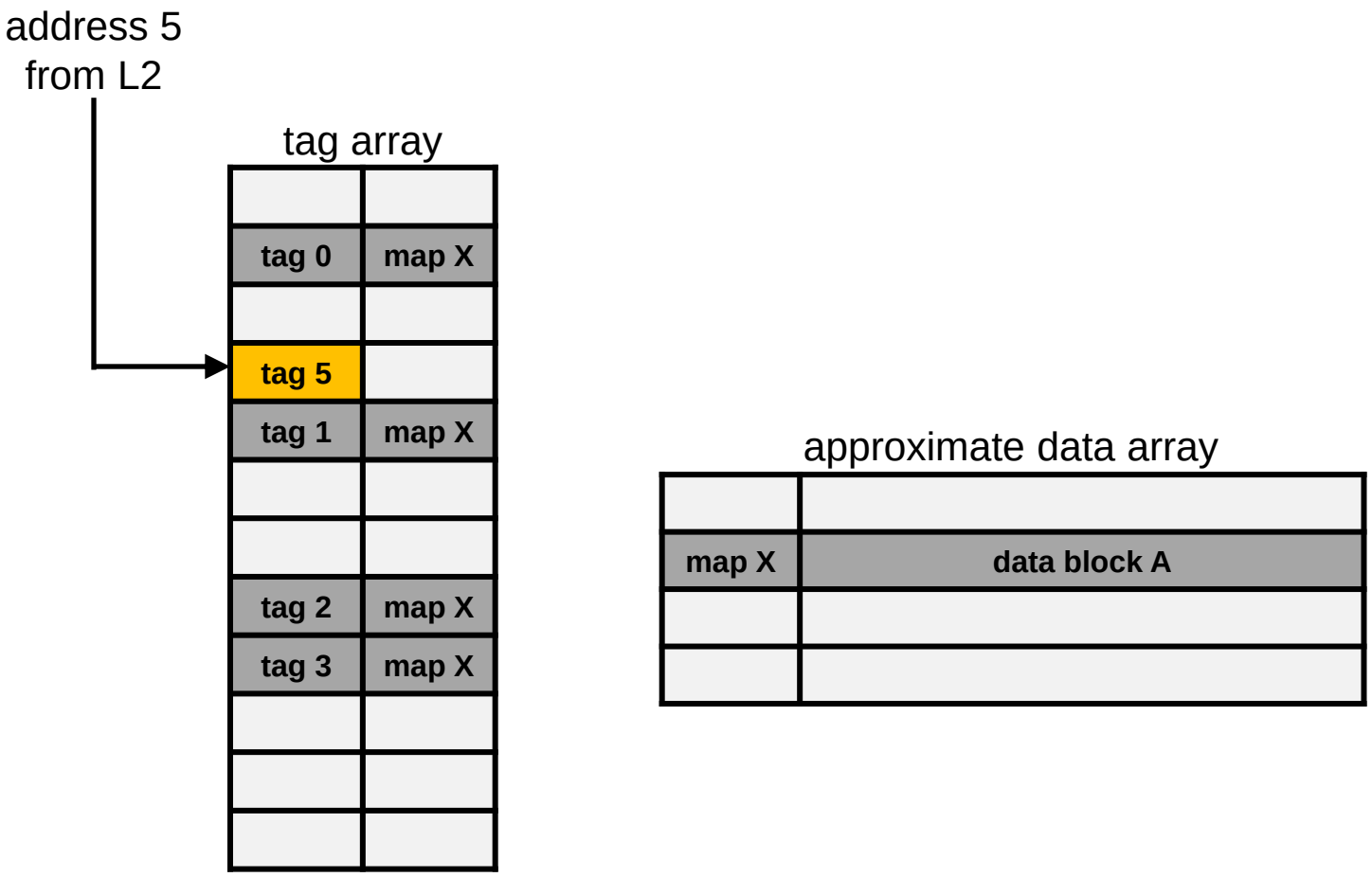
Doppelgänger Cache - Lookups



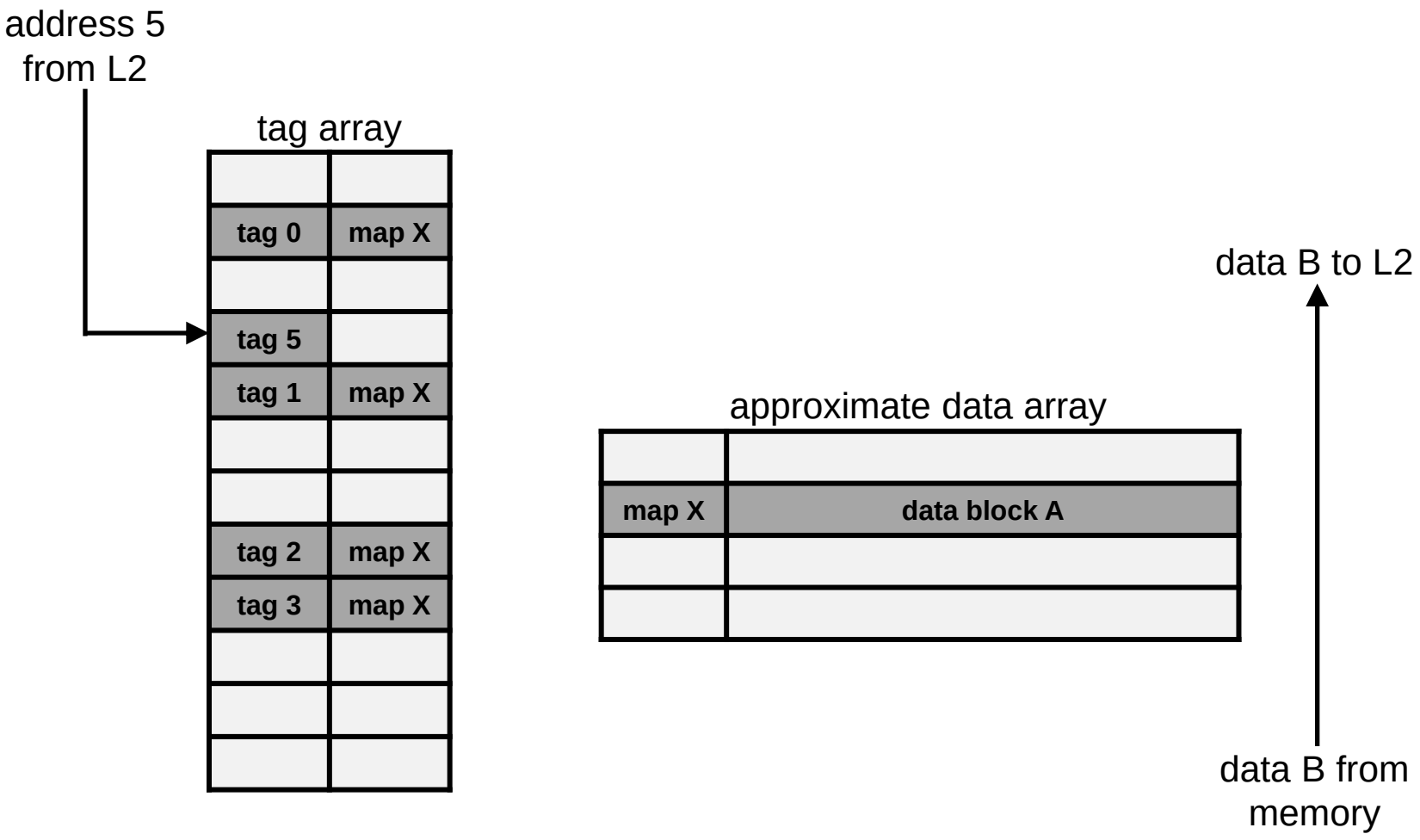
Doppelgänger Cache - Insertions



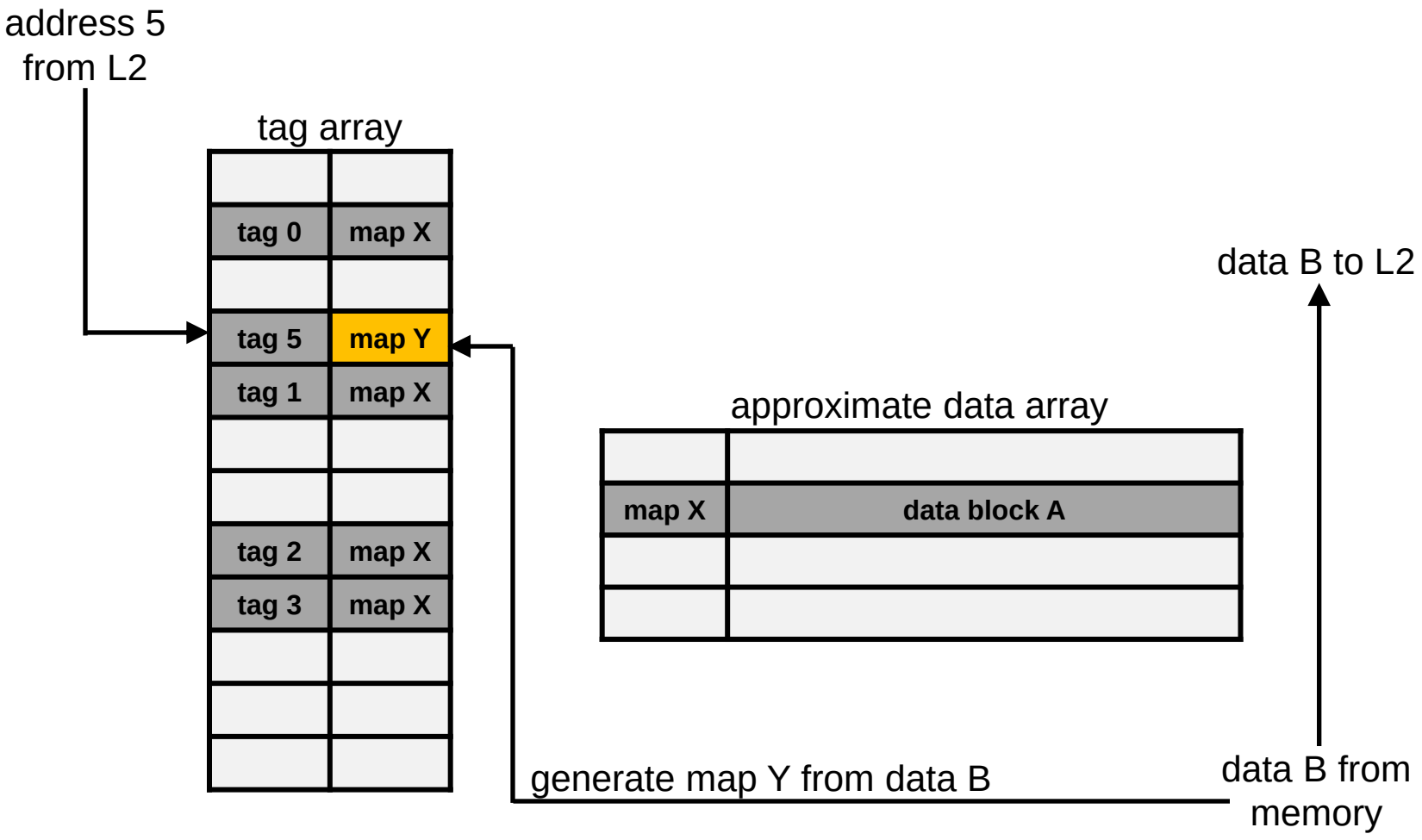
Doppelgänger Cache - Insertions



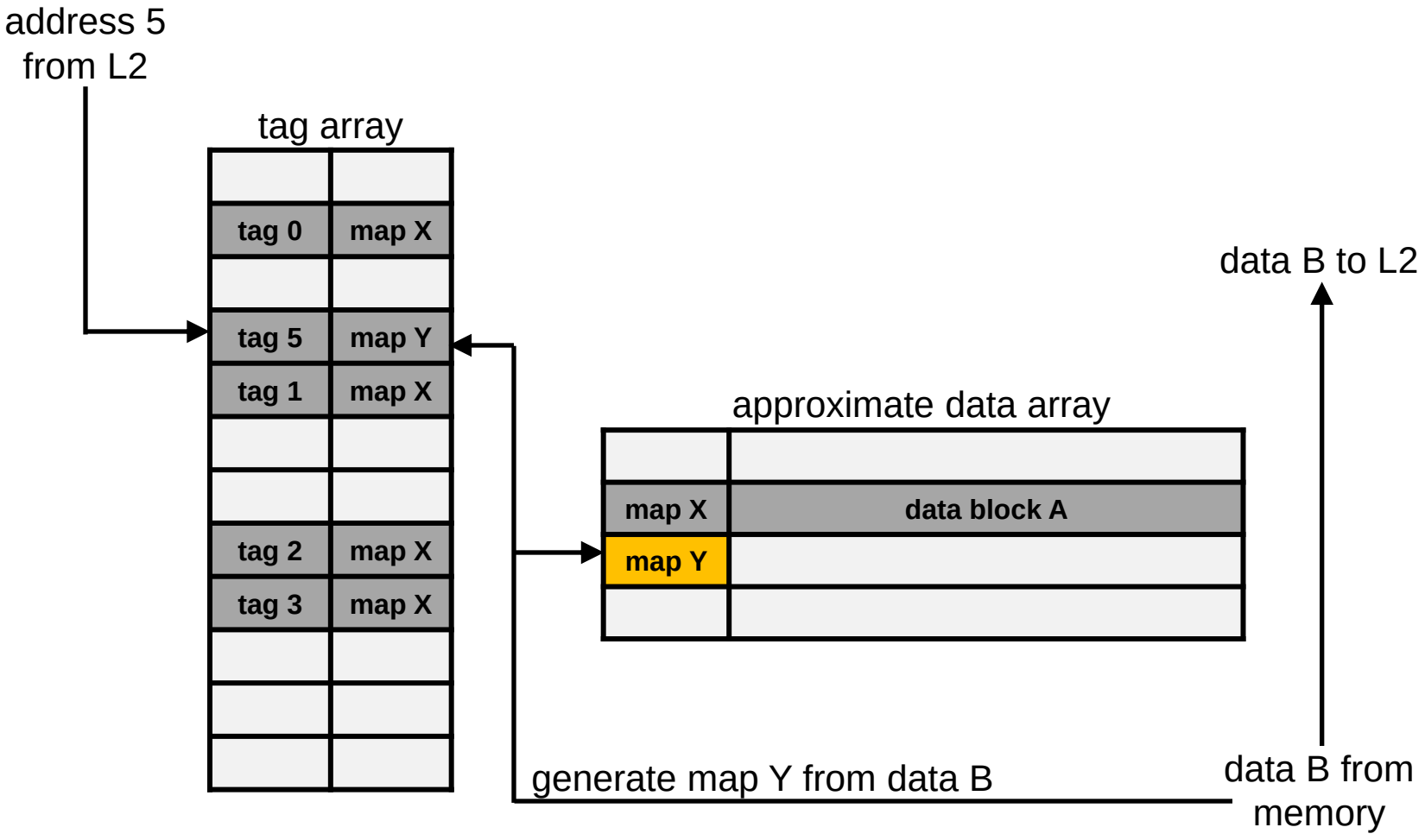
Doppelgänger Cache - Insertions



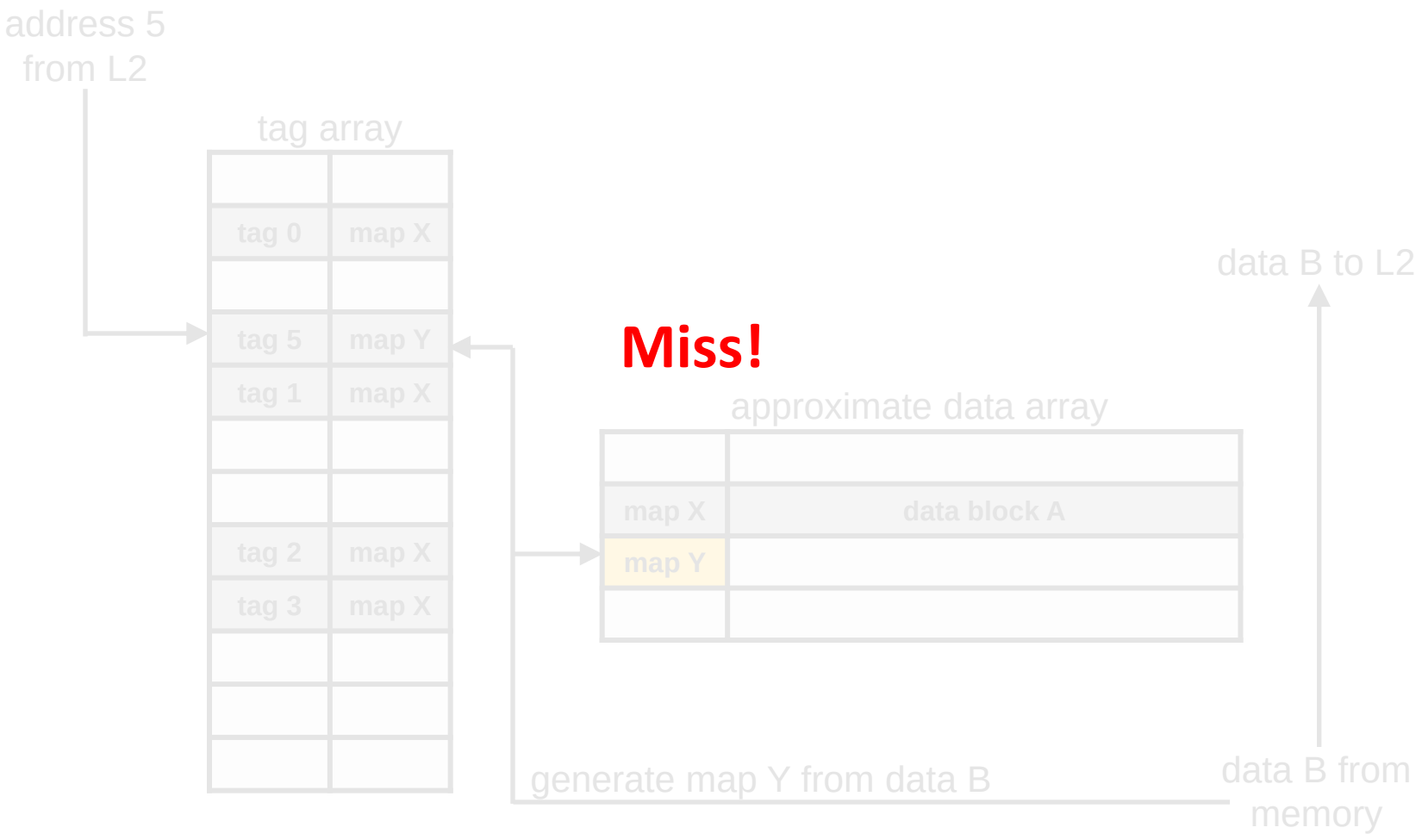
Doppelgänger Cache - Insertions



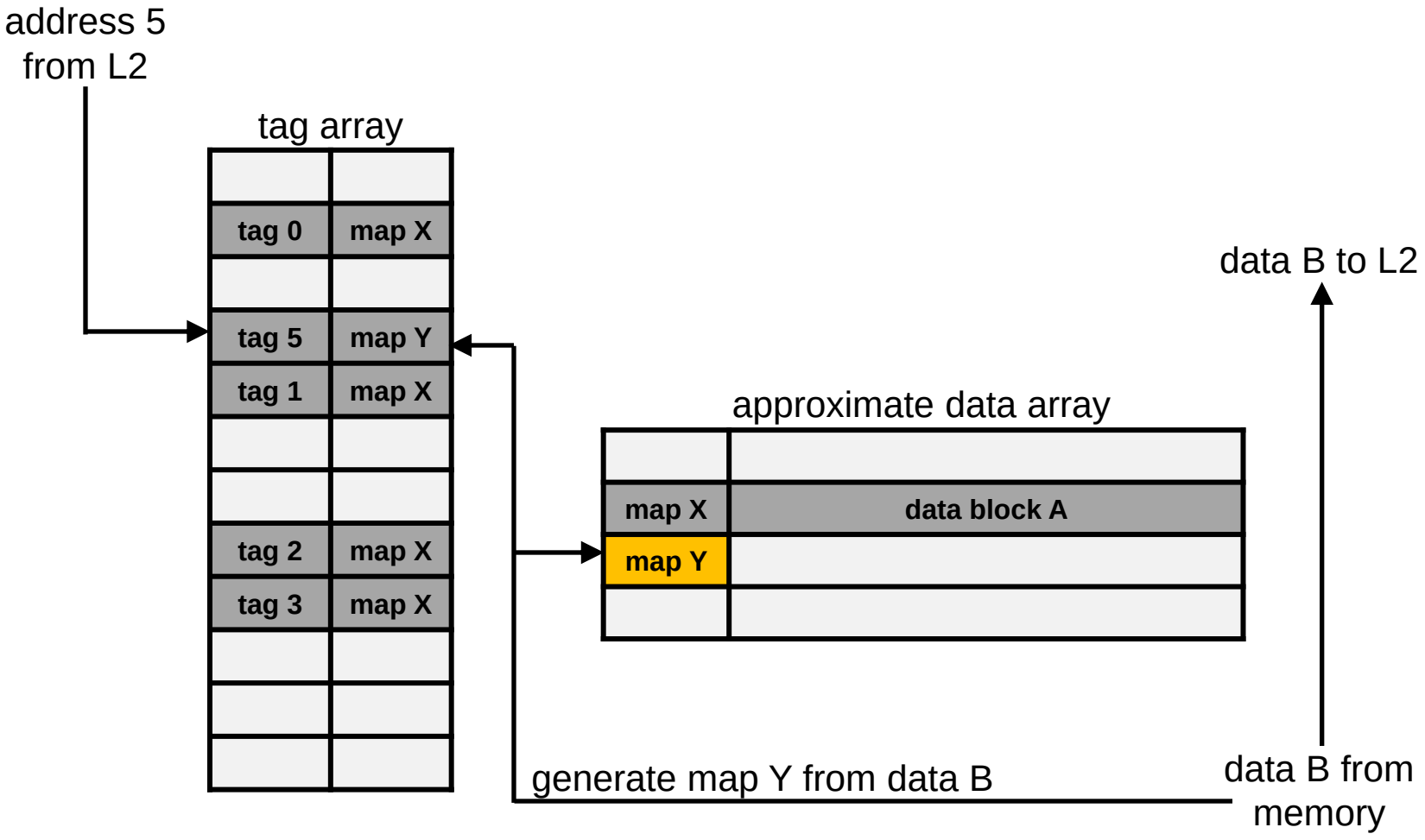
Doppelgänger Cache - Insertions



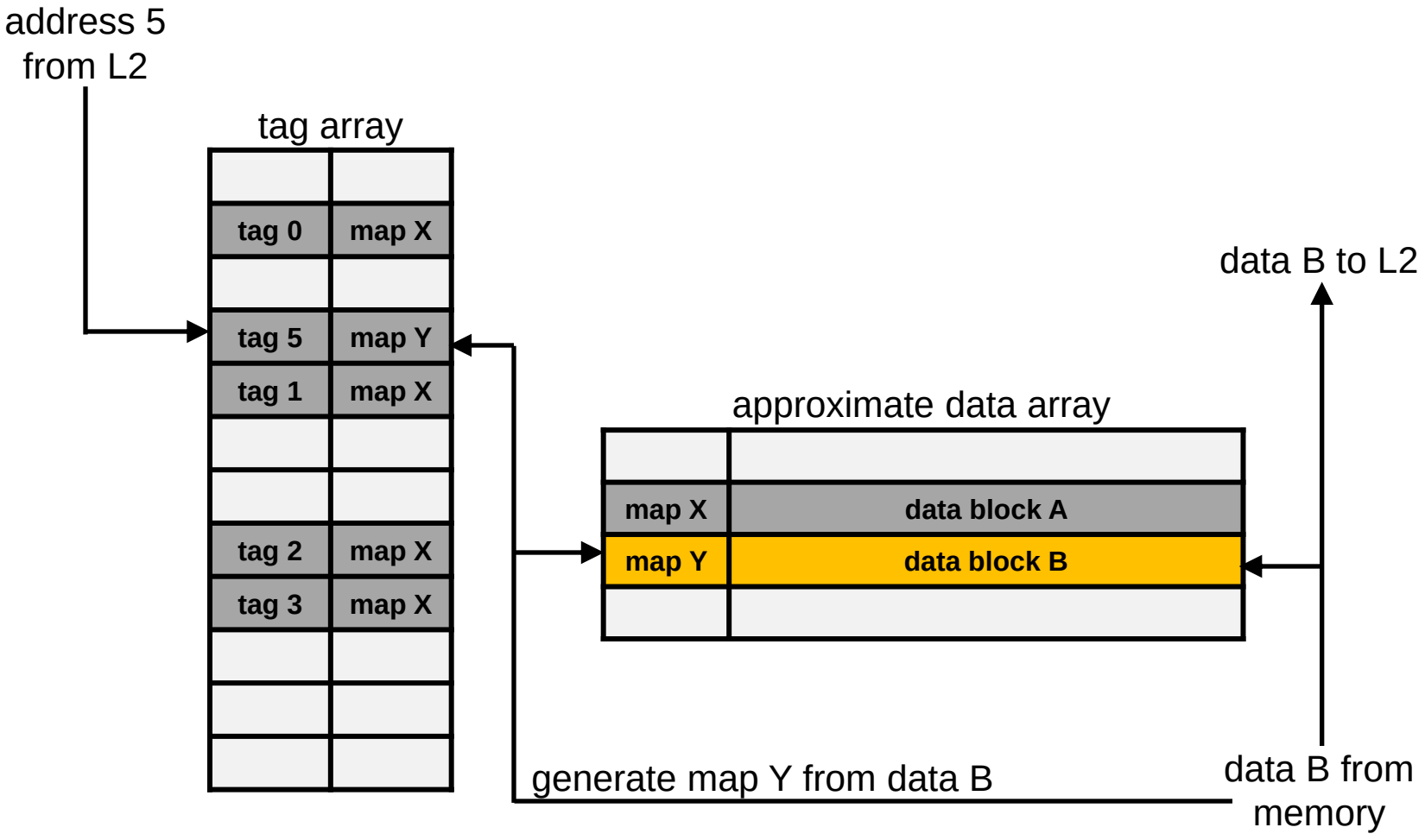
Doppelgänger Cache - Insertions



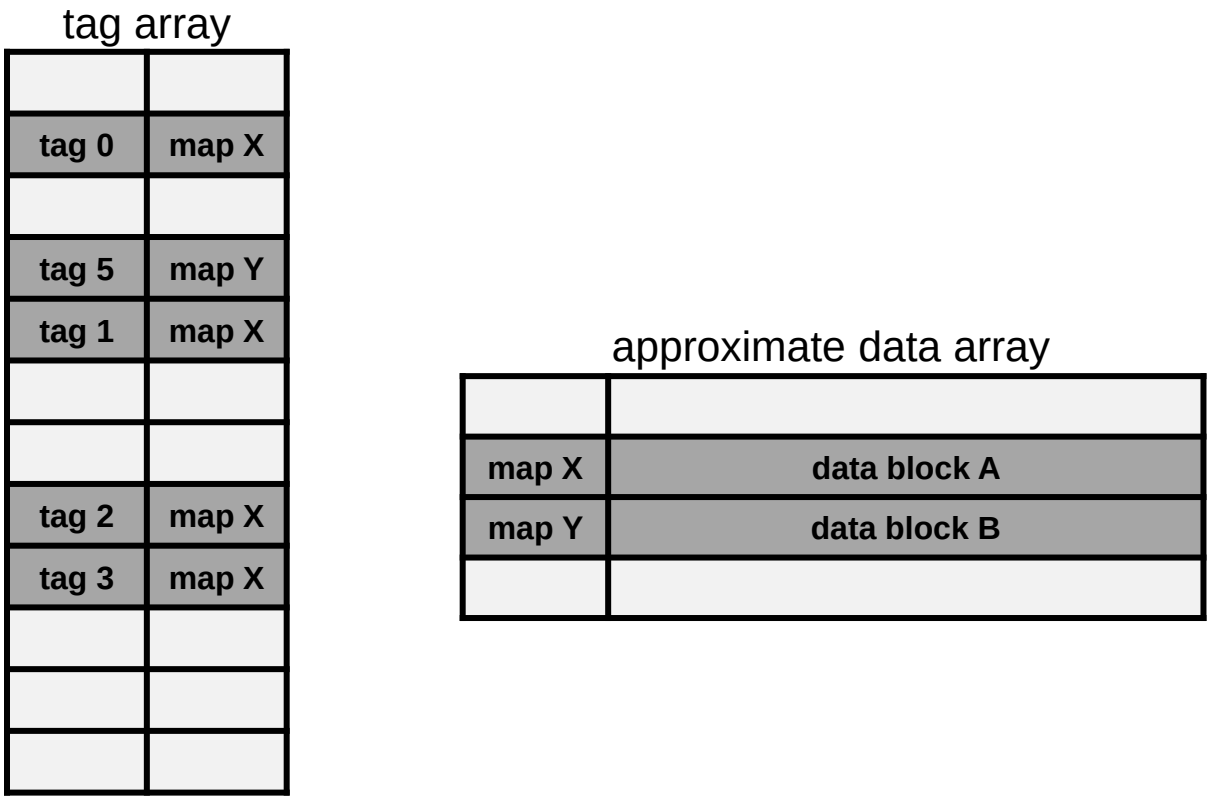
Doppelgänger Cache - Insertions (Miss)



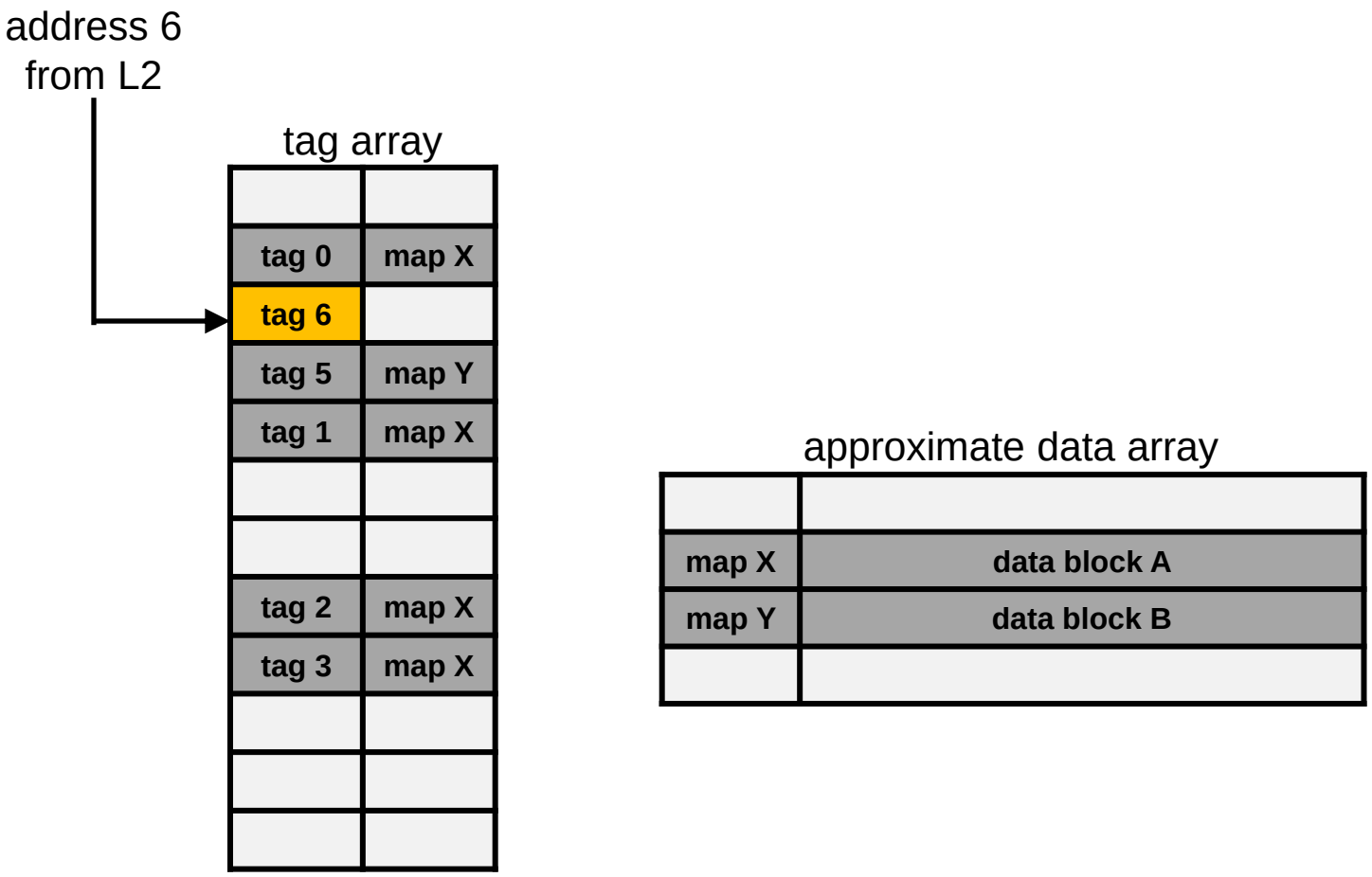
Doppelgänger Cache - Insertions (Miss)



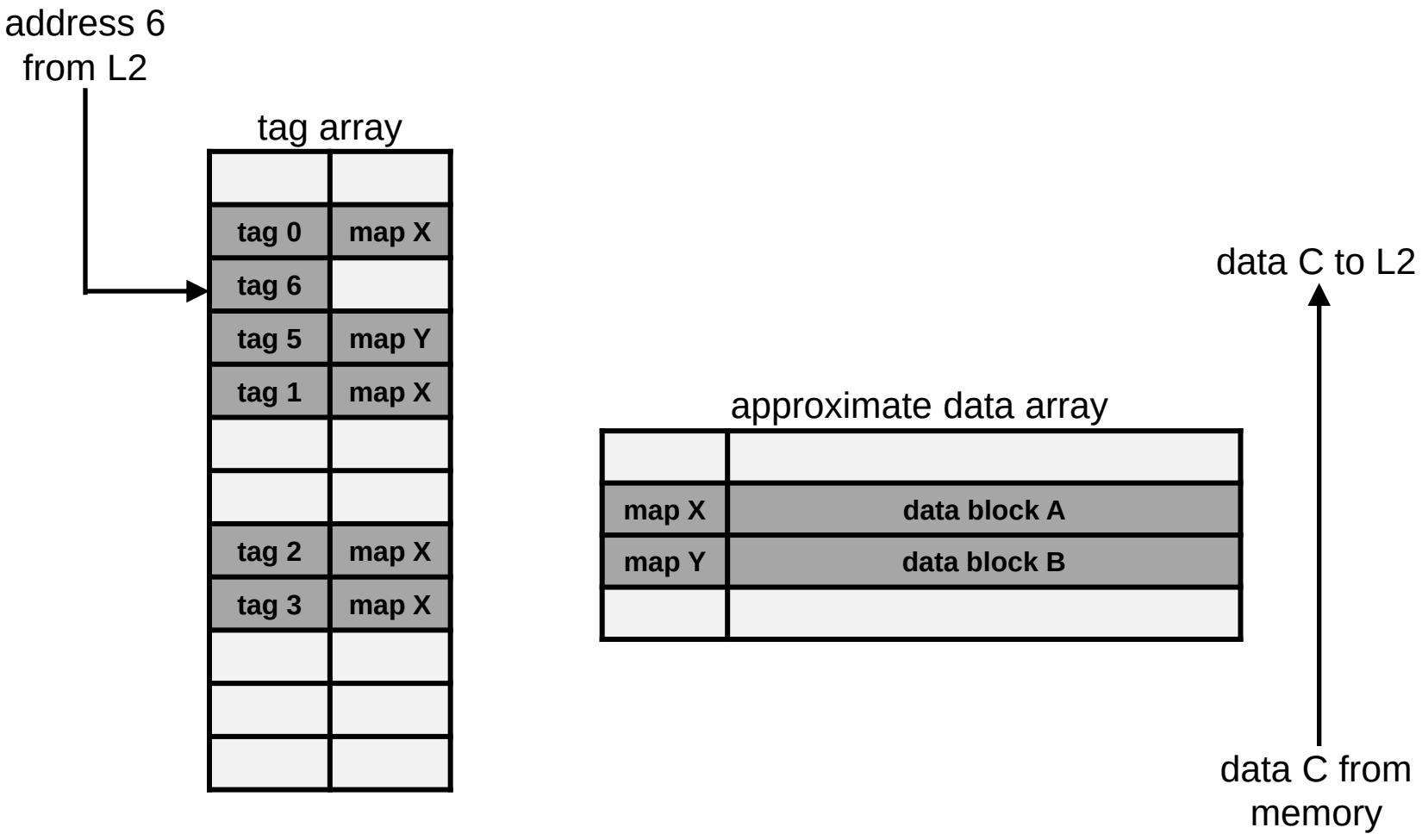
Doppelgänger Cache - Insertions



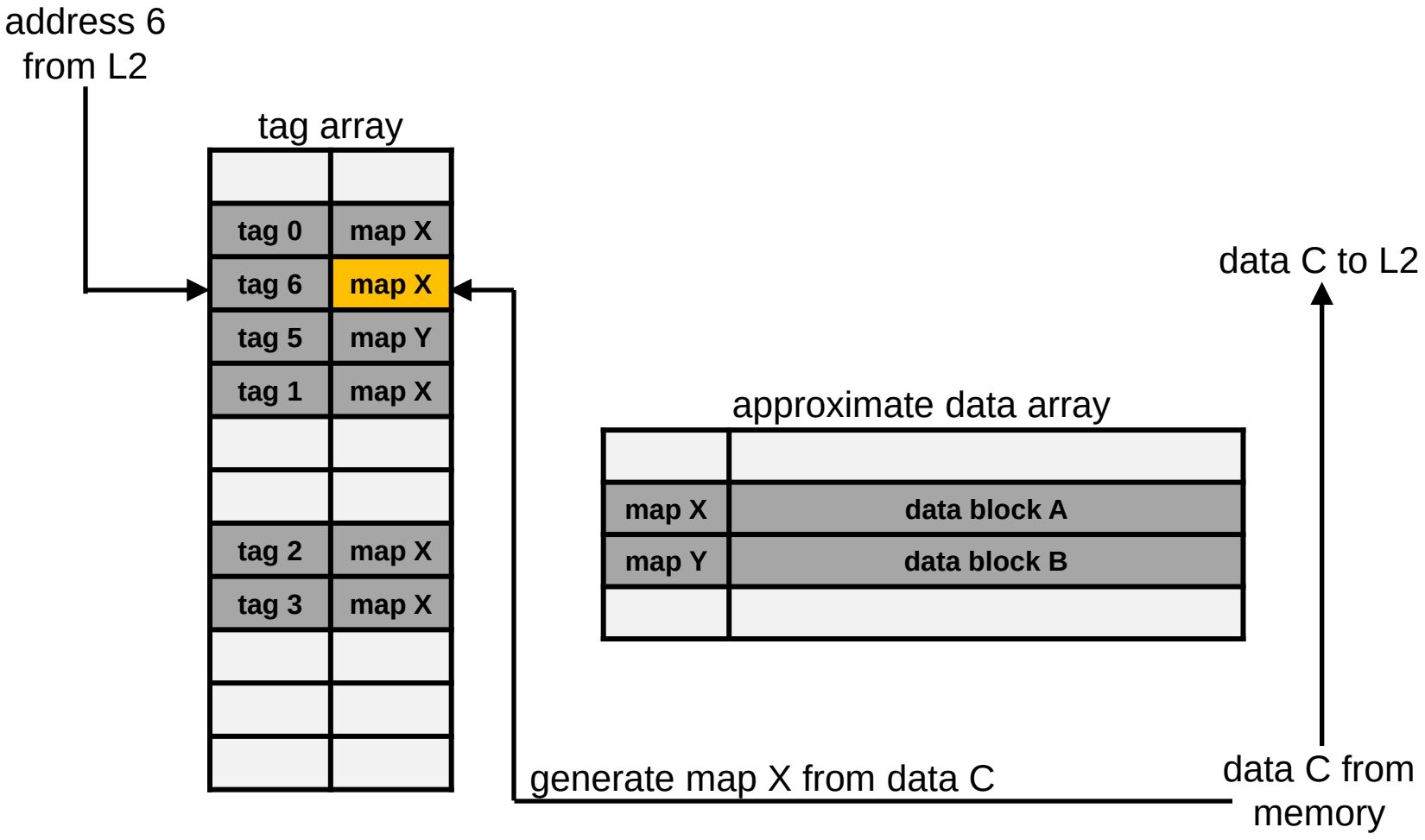
Doppelgänger Cache - Insertions



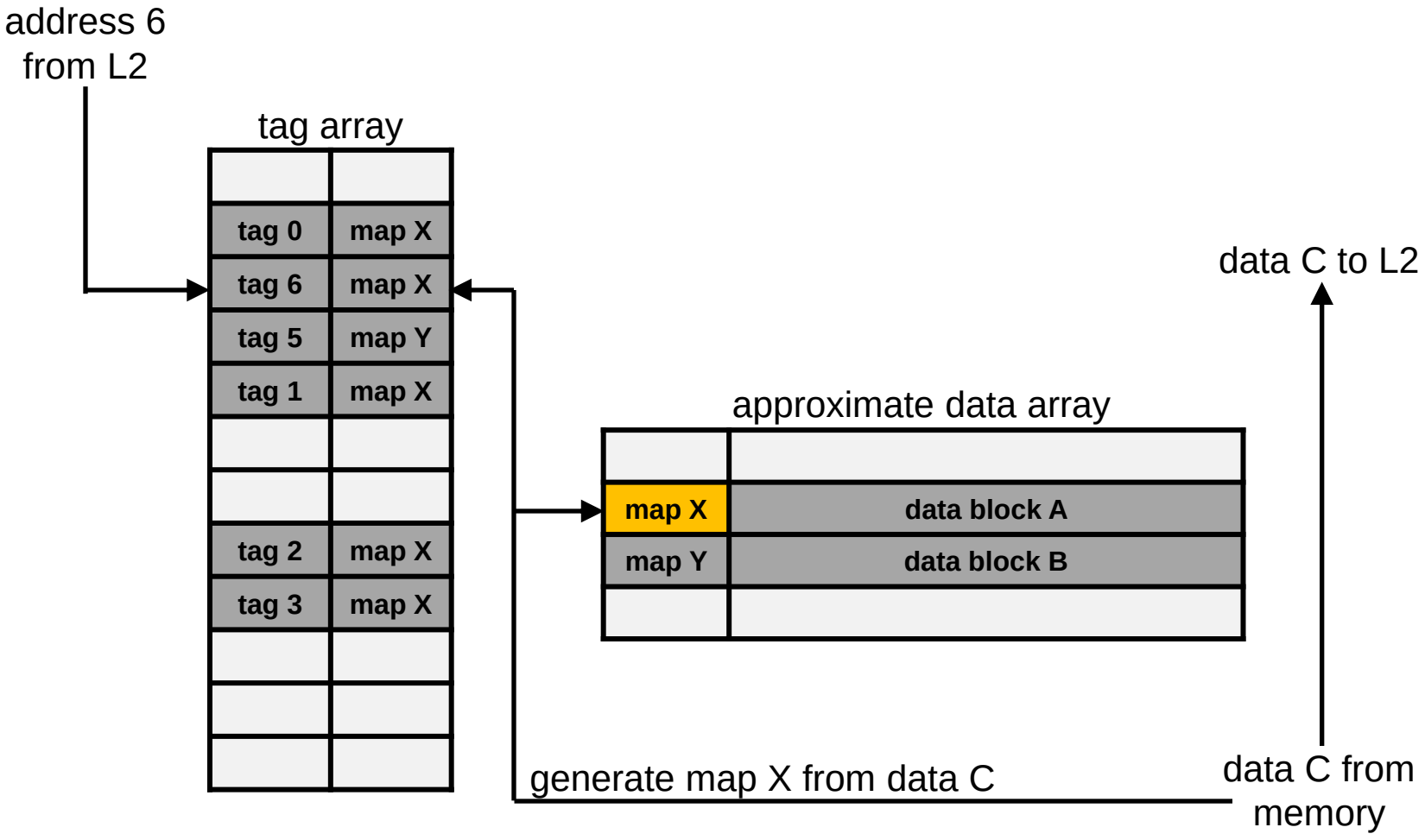
Doppelgänger Cache - Insertions



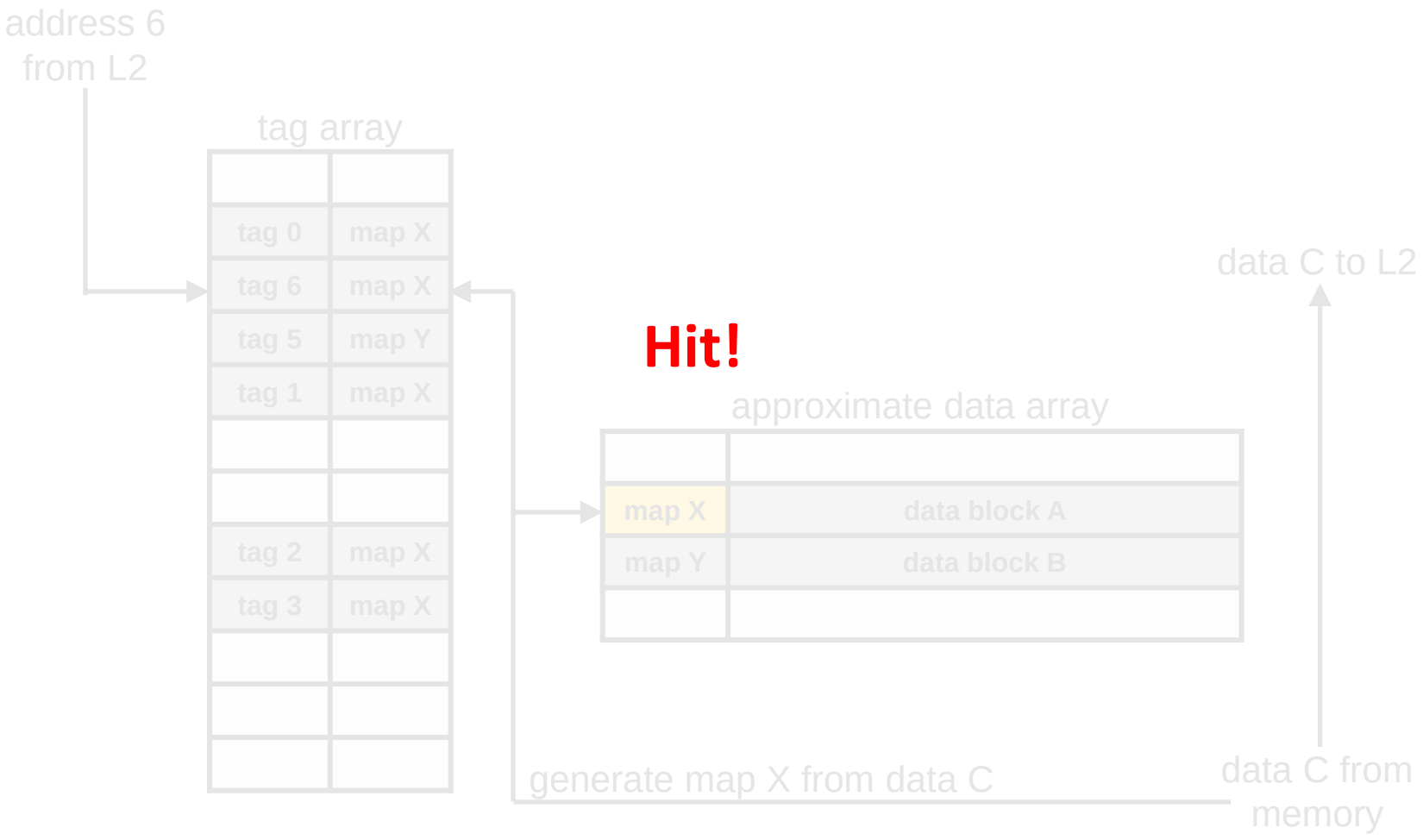
Doppelgänger Cache - Insertions



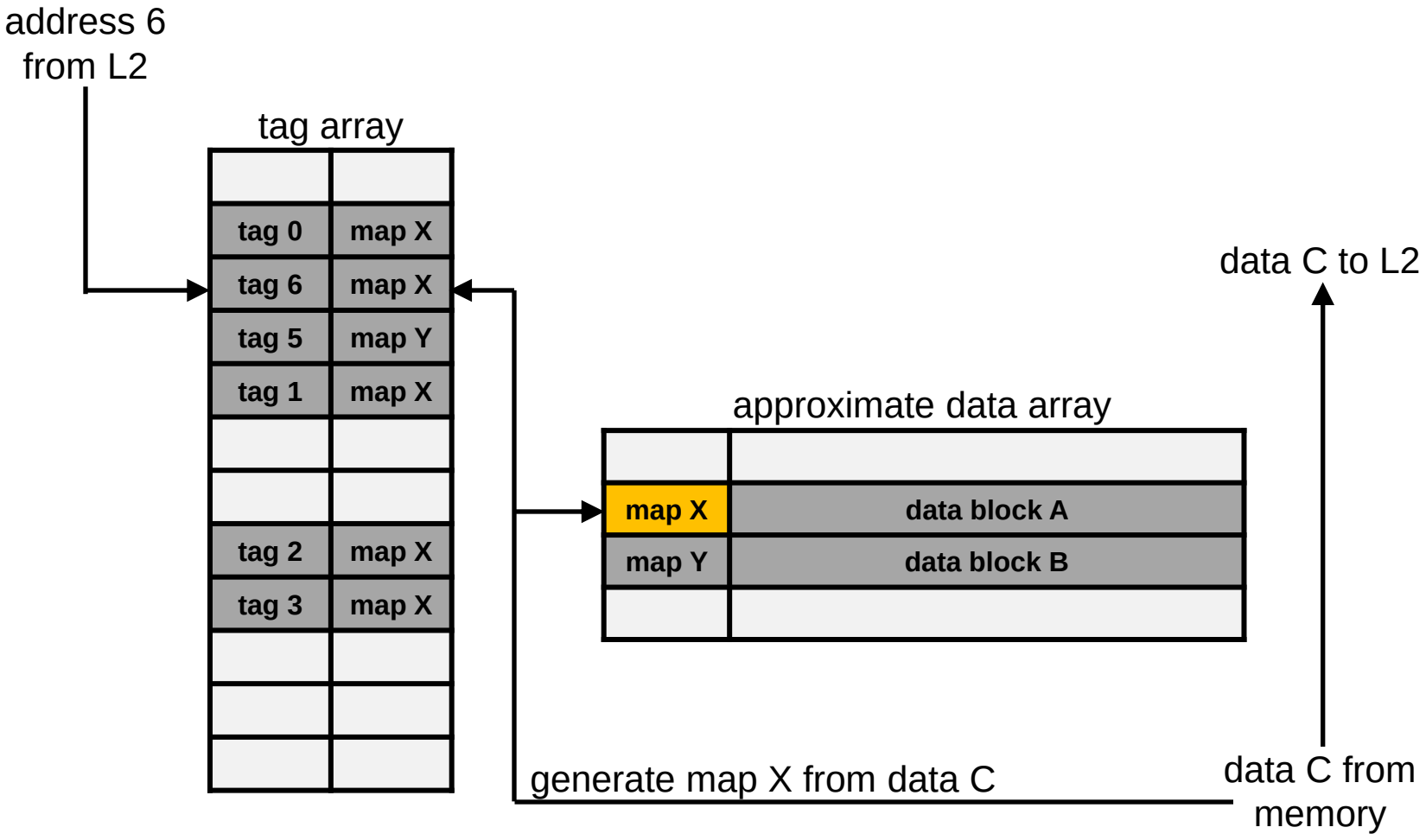
Doppelgänger Cache - Insertions



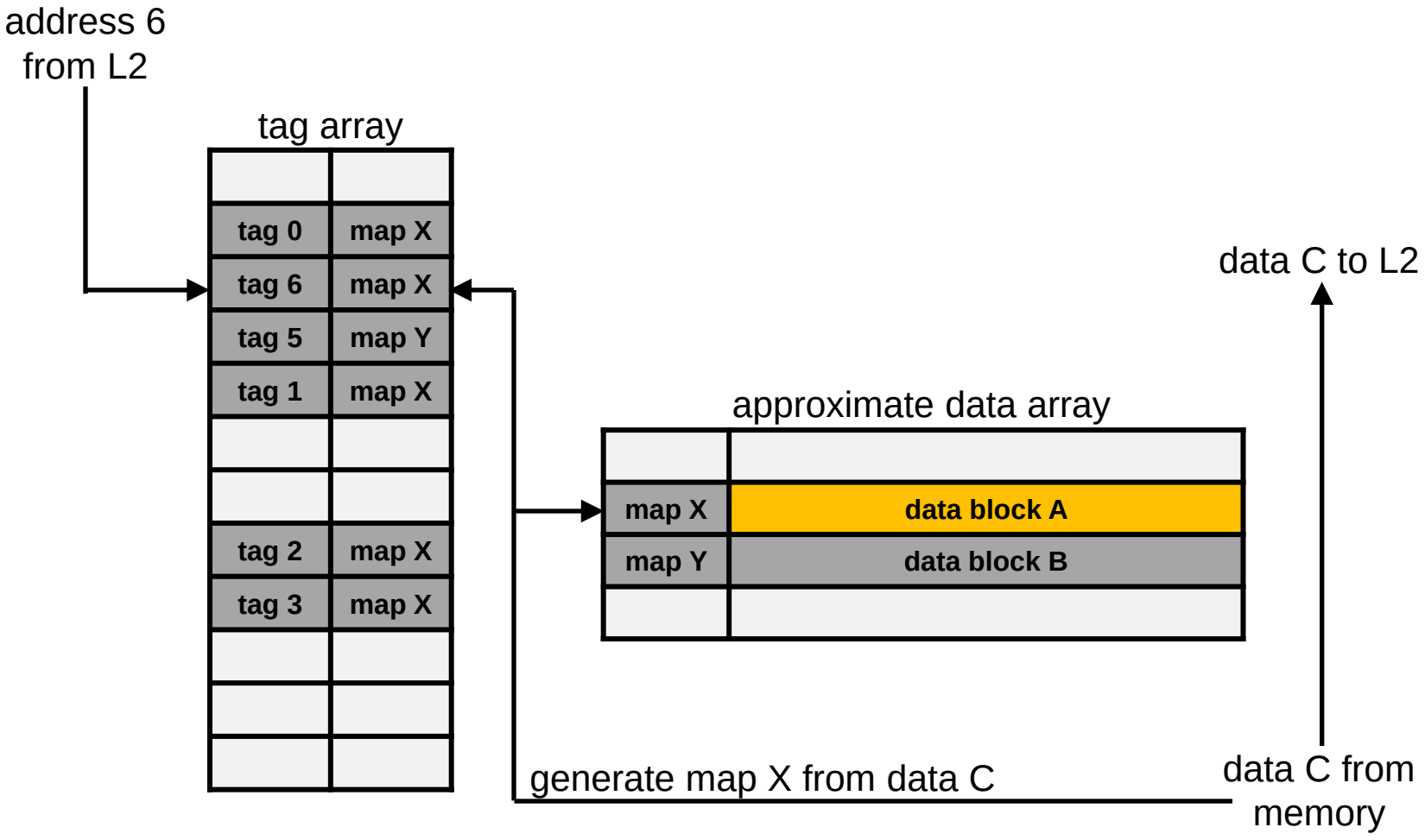
Doppelgänger Cache - Insertions



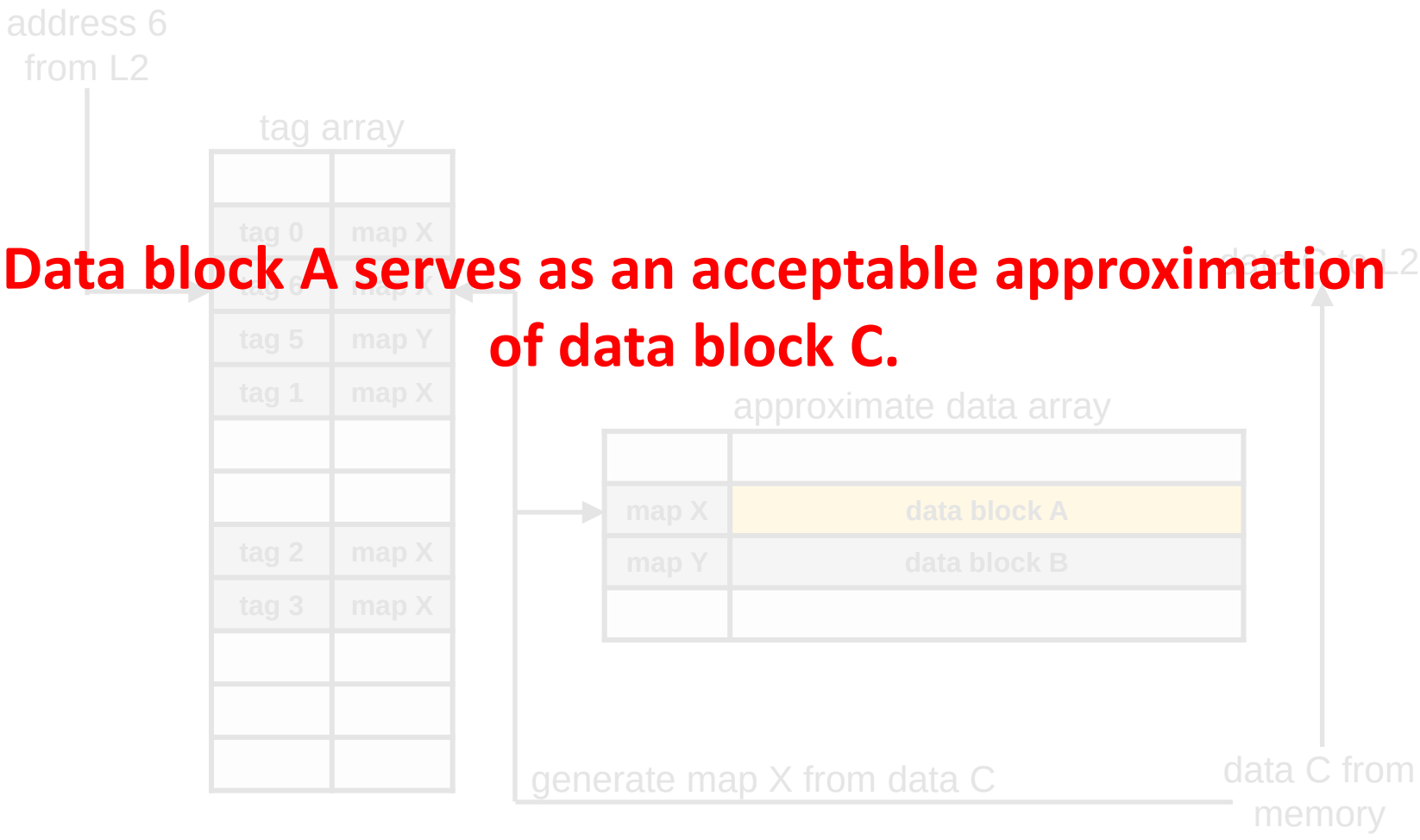
Doppelgänger Cache - Insertions (Hit)



Doppelgänger Cache - Insertions (Hit)

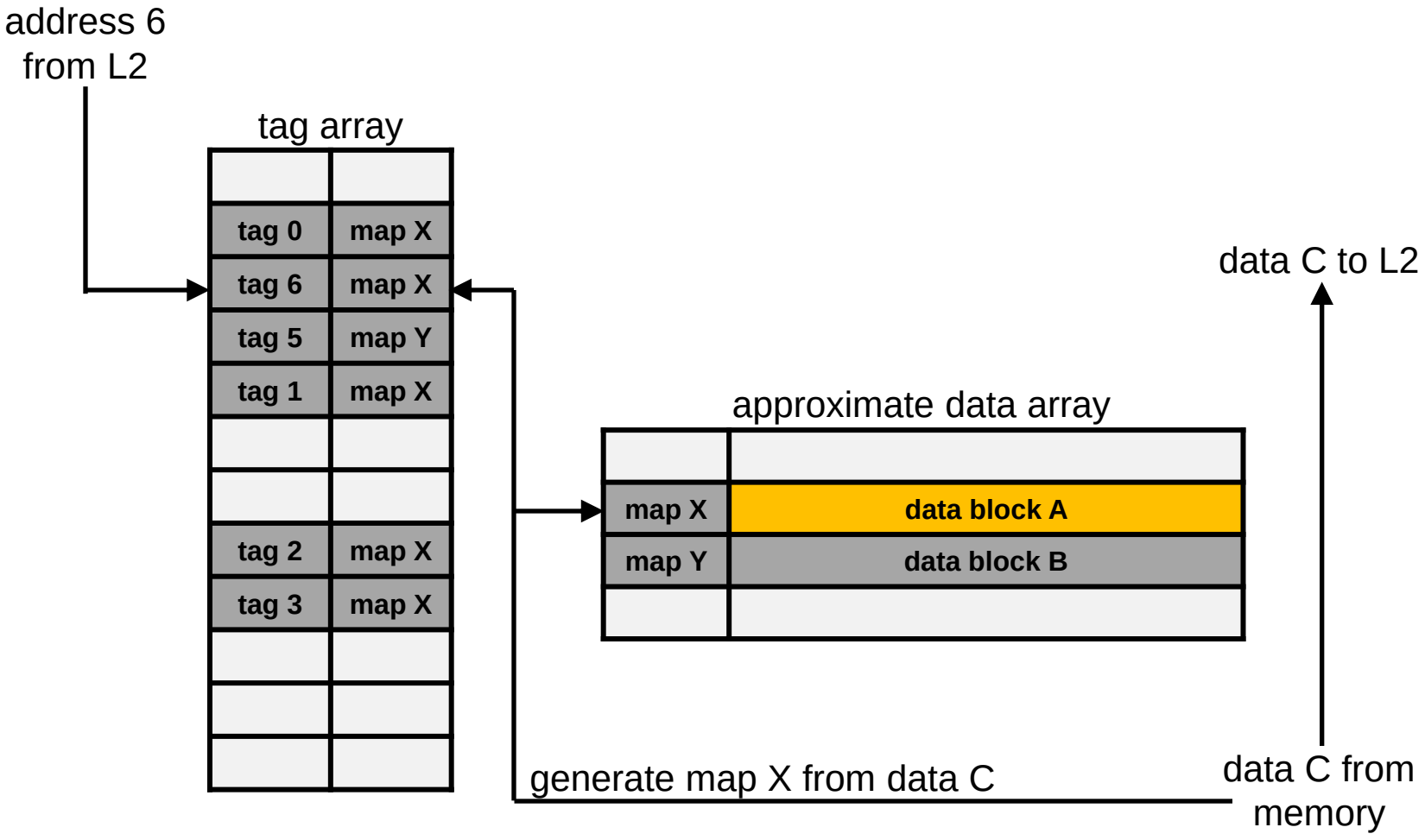


Doppelgänger Cache - Insertions (Hit)



Data block A serves as an acceptable approximation of data block C.

Doppelgänger Cache - Insertions (Hit)

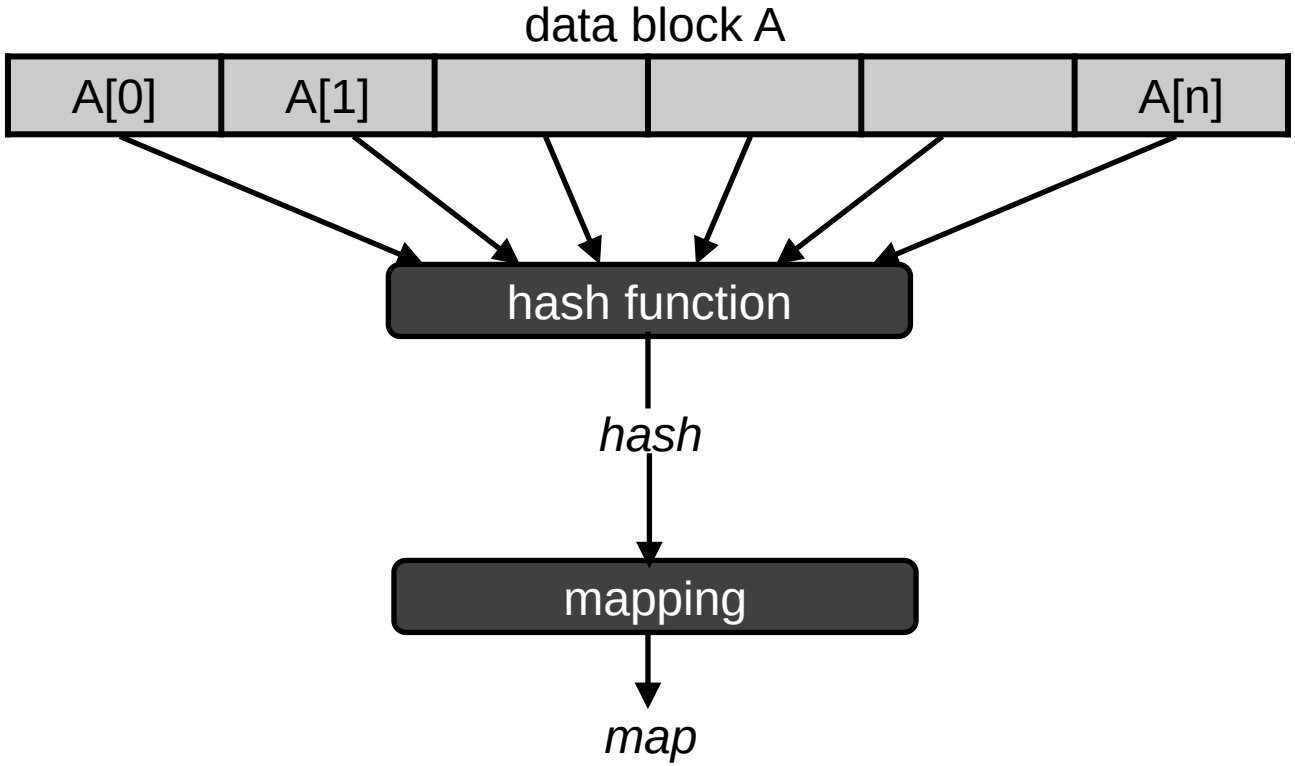


Doppelgänger Cache - Similarity Mapping

The **map** value represents the signature (or **likeness**) of a block. Blocks that generate the same map value are approximately similar.

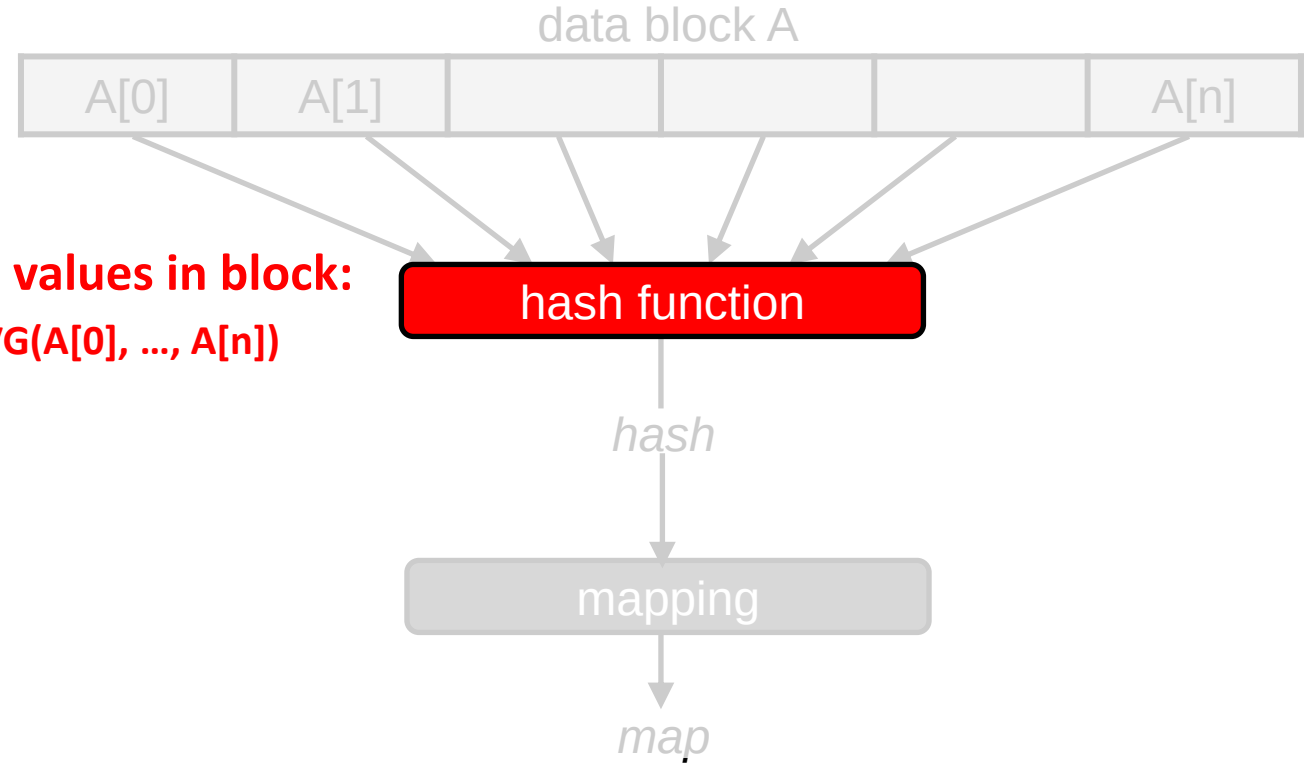
Doppelgänger Cache - Similarity Mapping

The **map** value represents the signature (or **likeness**) of a block. Blocks that generate the same map value are approximately similar.



Doppelgänger Cache - Similarity Mapping

The **map** value represents the signature (or **likeness**) of a block. Blocks that generate the same map value are approximately similar.

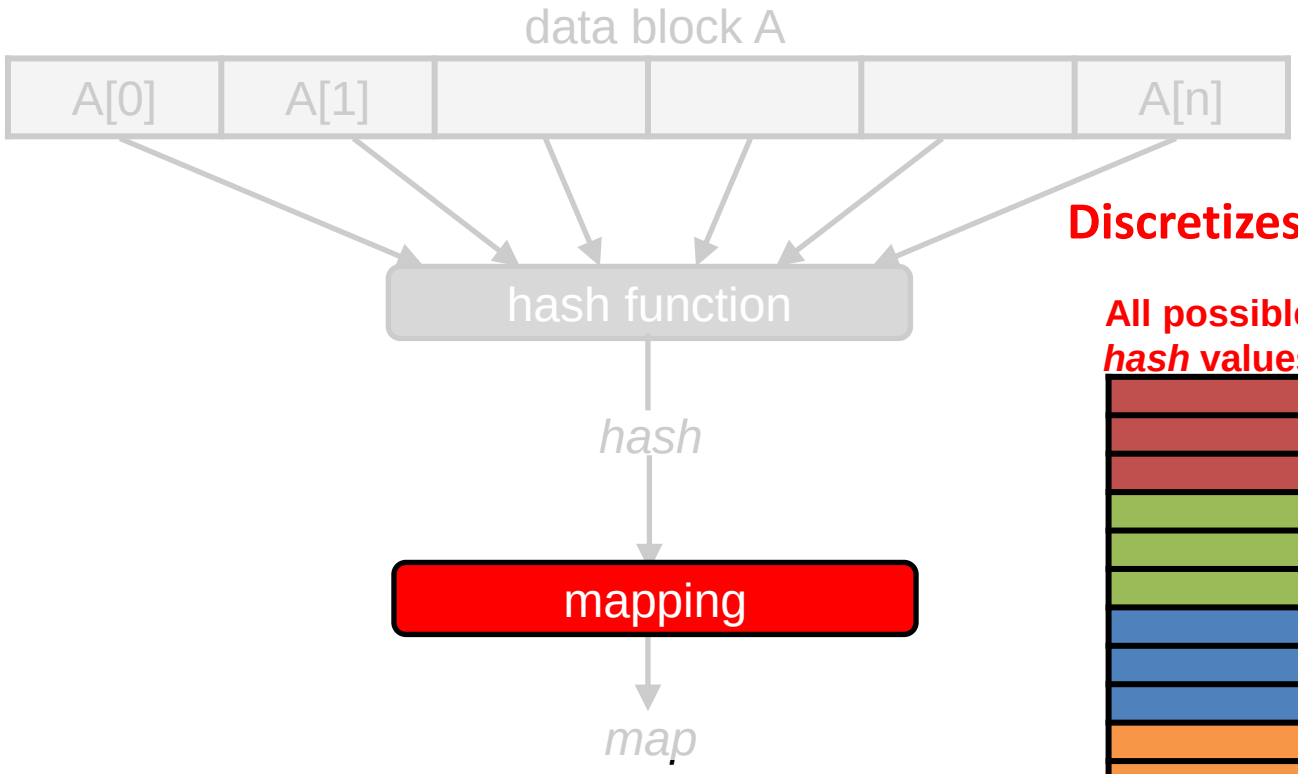


Aggregates values in block:

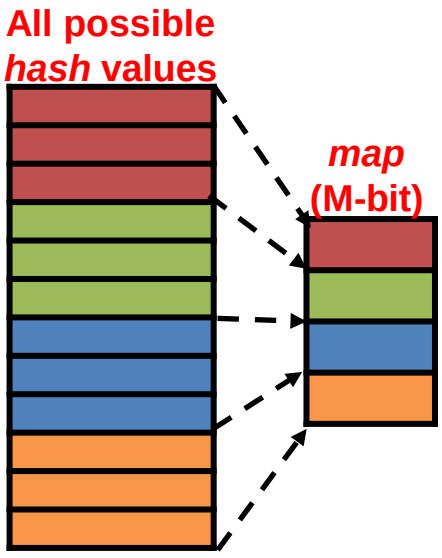
***hash* = AVG(A[0], …, A[n])**

Doppelgänger Cache - Similarity Mapping

The **map** value represents the signature (or **likeness**) of a block. Blocks that generate the same map value are approximately similar.

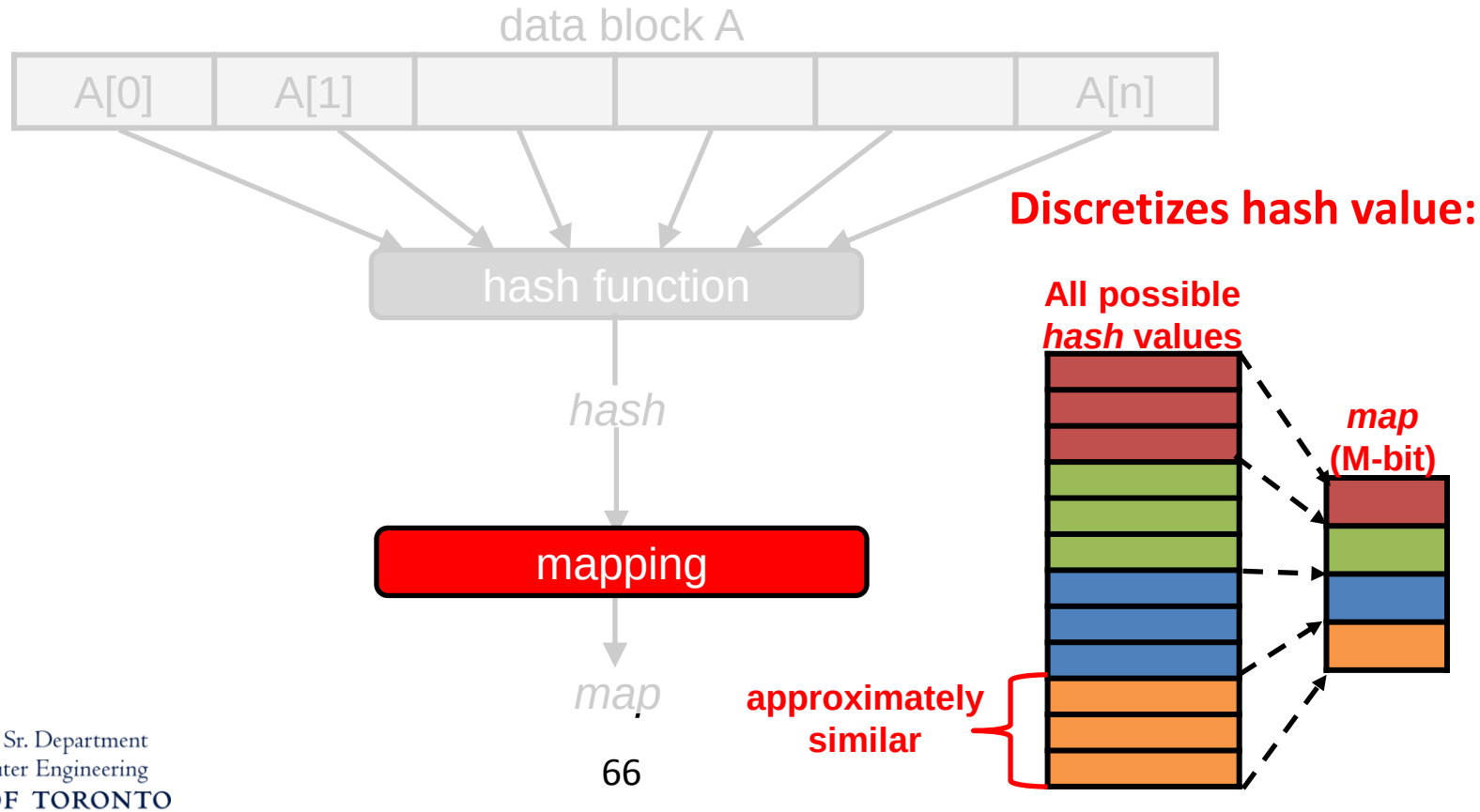


Discretizes hash value:

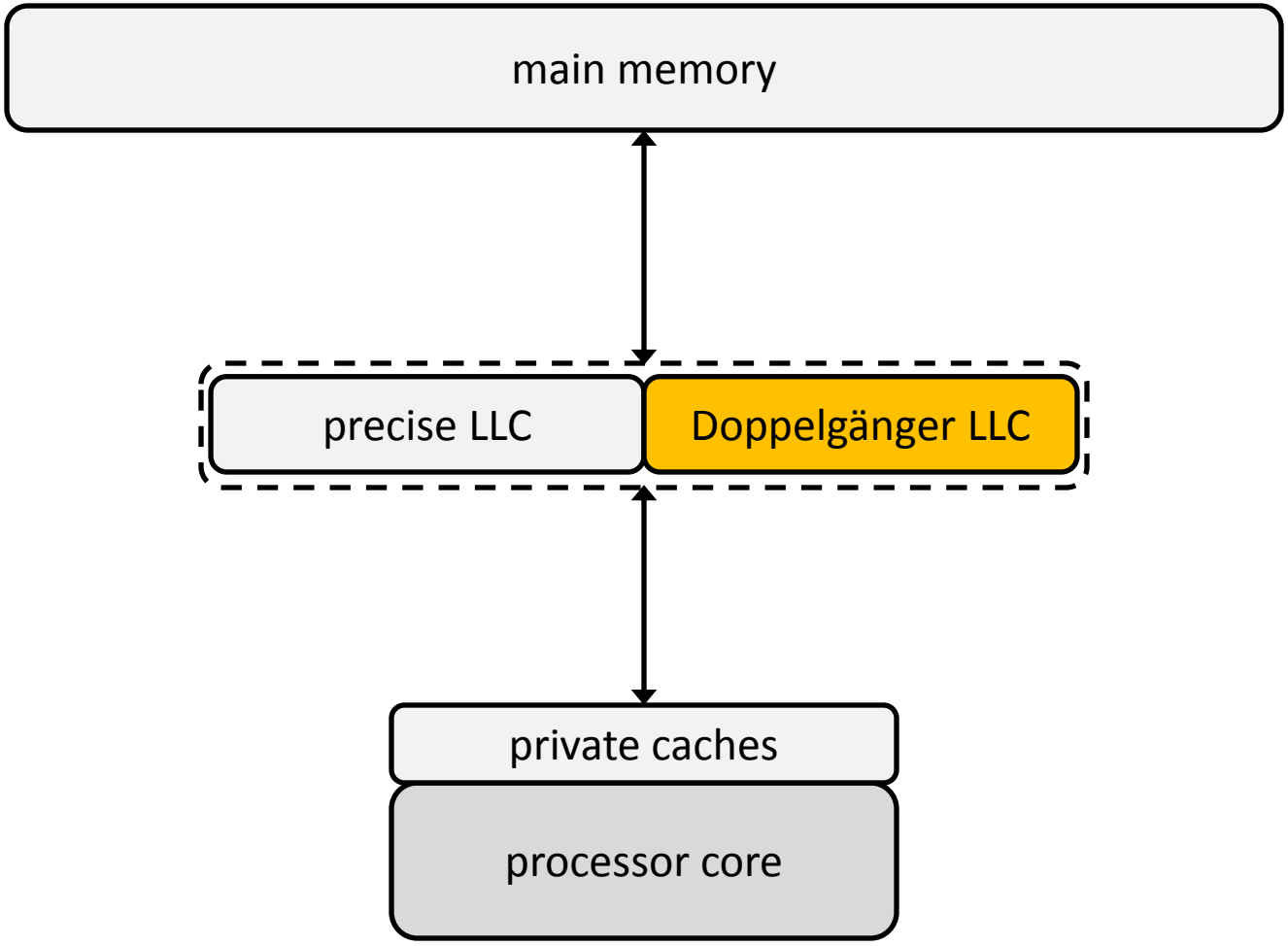


Doppelgänger Cache - Similarity Mapping

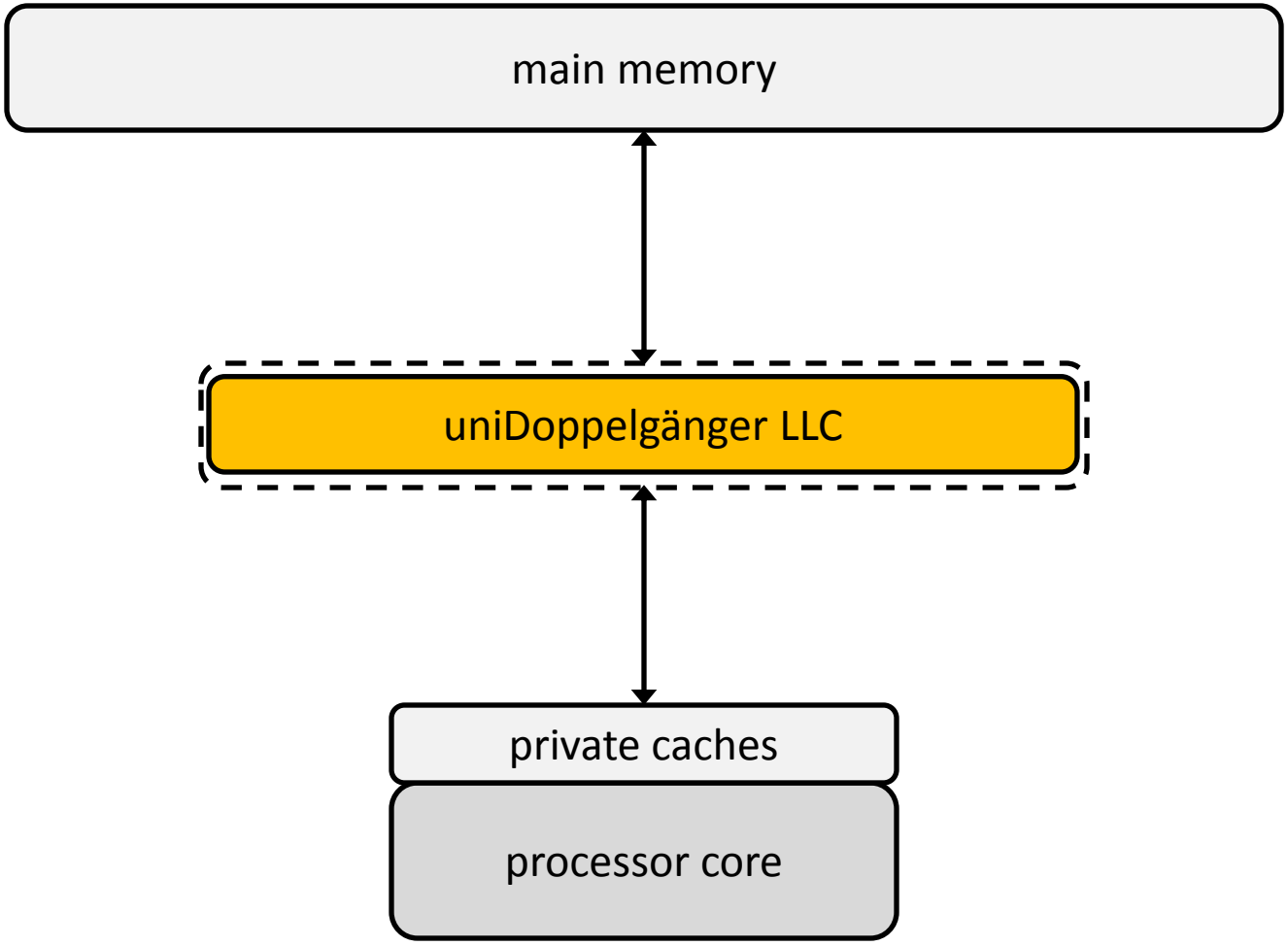
The **map** value represents the signature (or **likeness**) of a block. Blocks that generate the same map value are approximately similar.



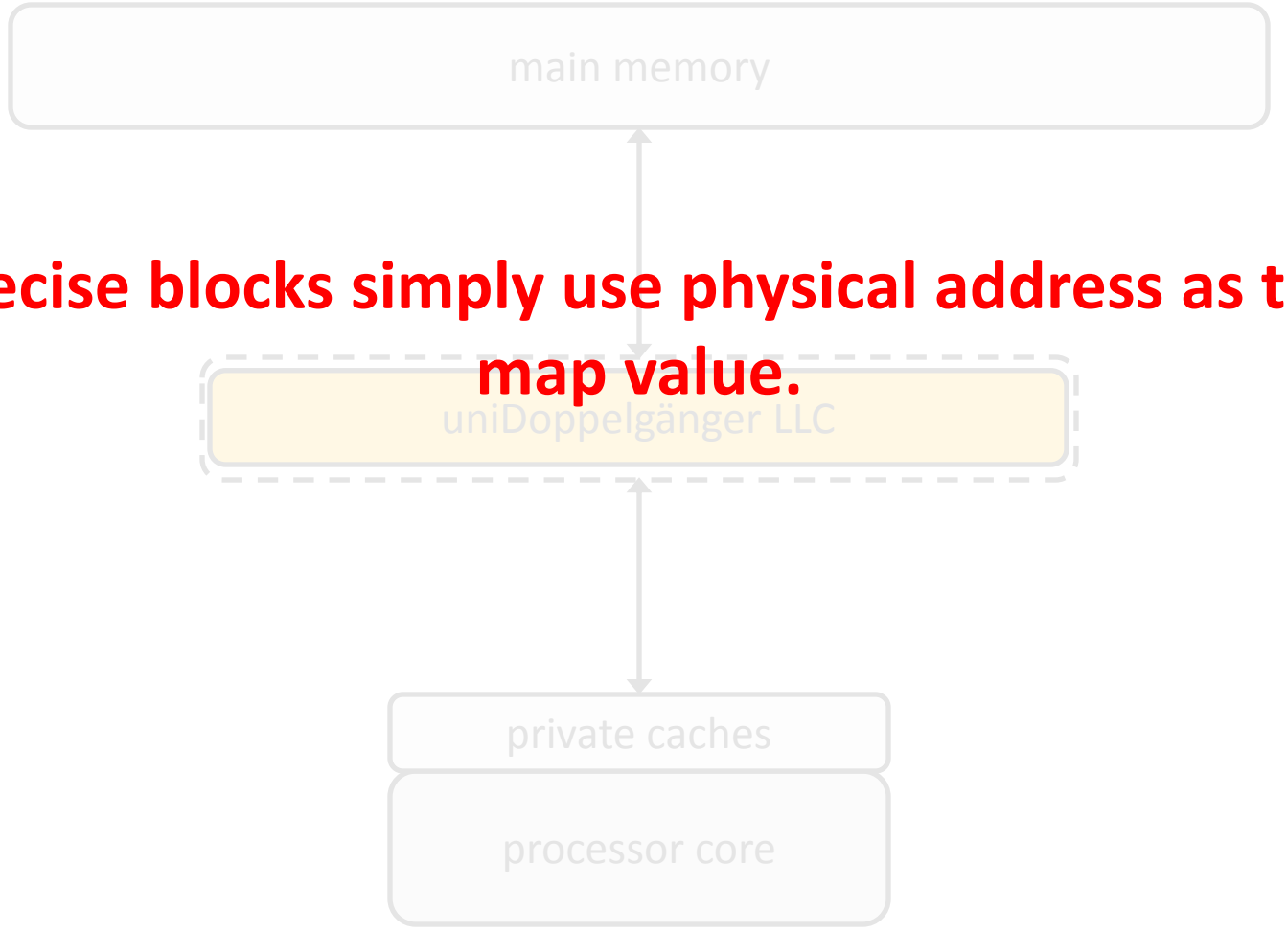
Doppelgänger Cache



uniDoppelgänger Cache



uniDoppelgänger Cache



Precise blocks simply use physical address as the map value.

Doppelgänger Cache

More details in paper:

- Cache writes, replacements and coherence.
- Details on hash functions and mapping.
- Sensitivity to size of map space and data array.
- Evaluation of uniDoppelgänger.

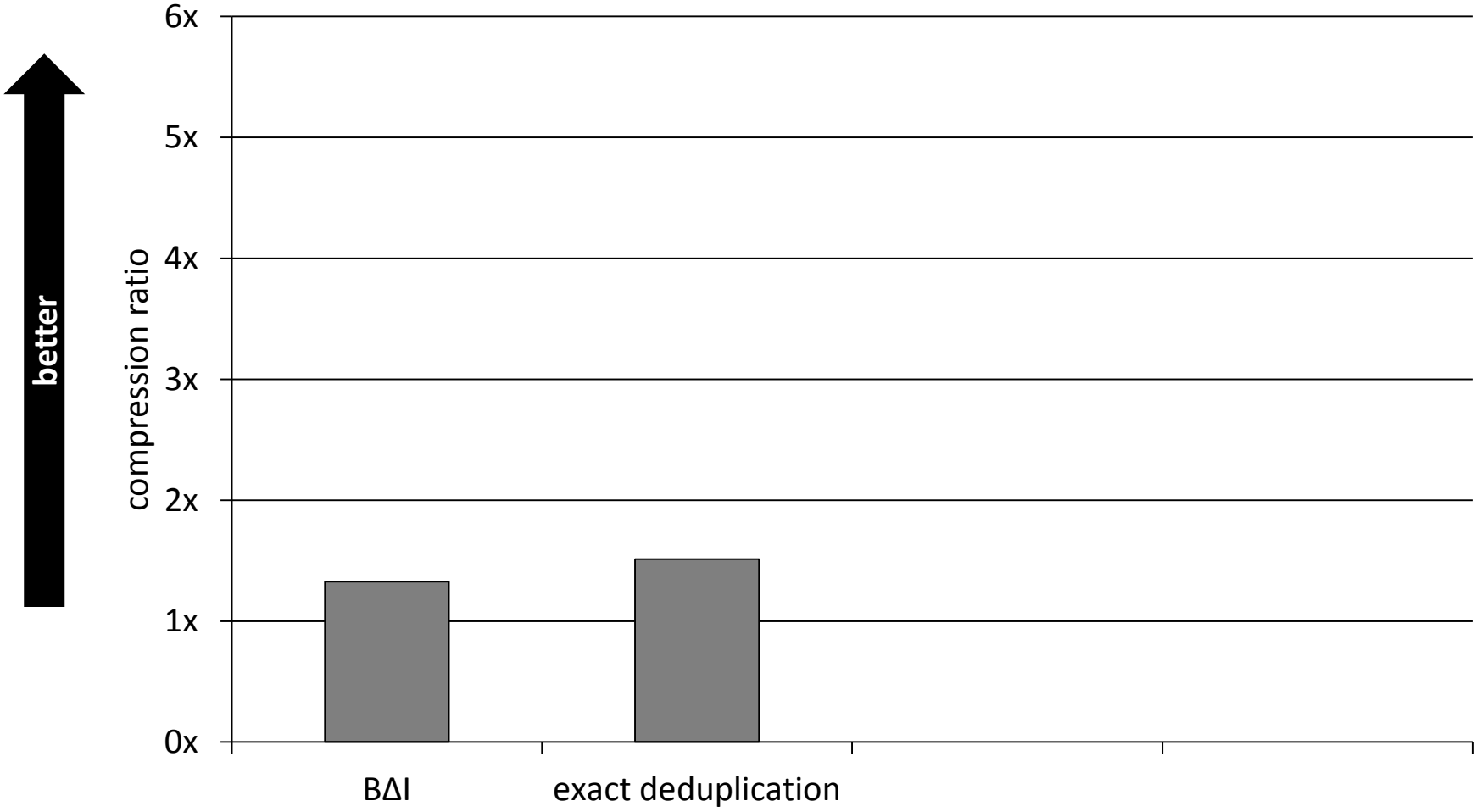
Outline

- Approximate Computing
  - Approximate Similarity
- Doppelgänger Cache
  - Cache Architecture
  - Similarity Mapping
- **Evaluation**

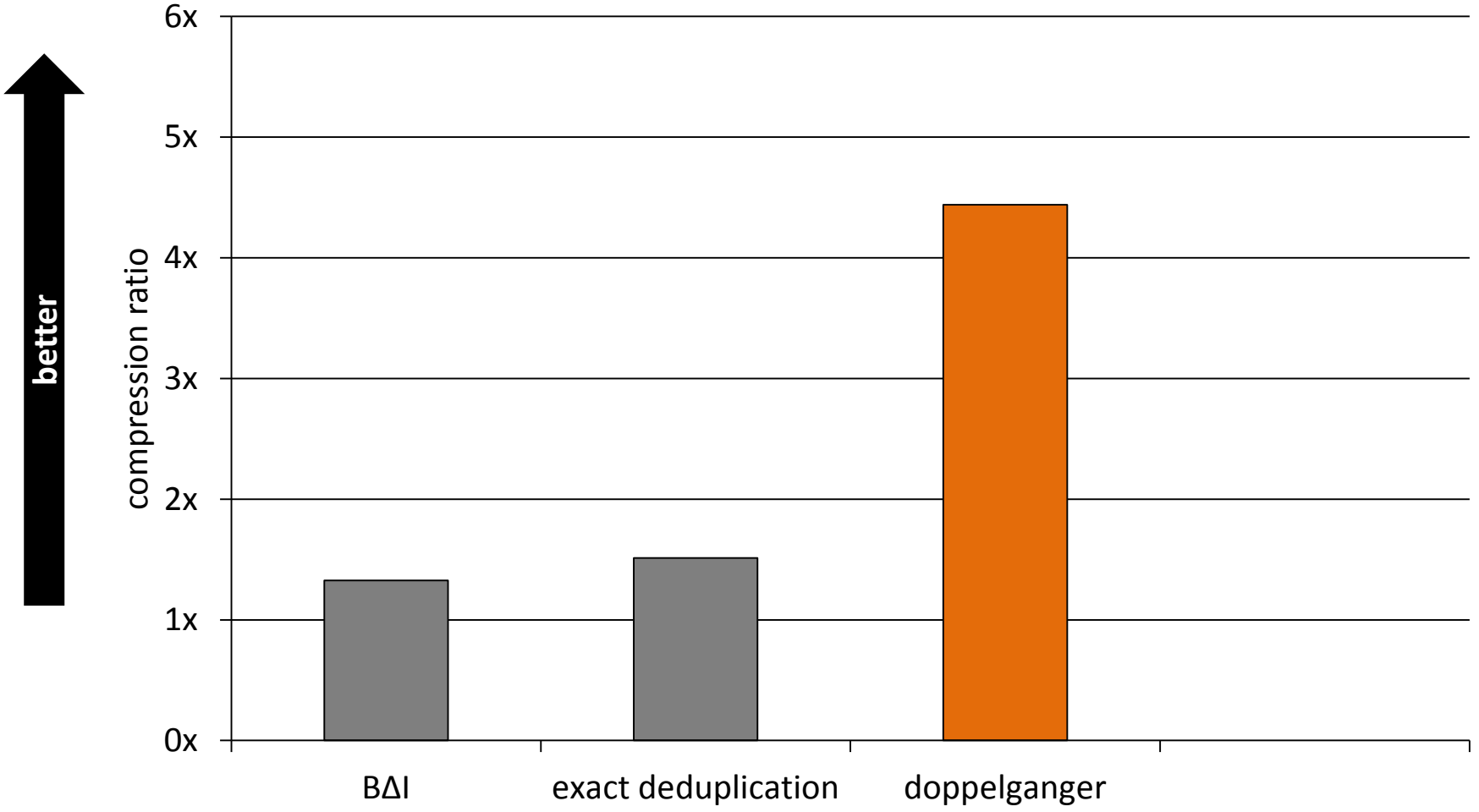
Evaluation

- **Applications**: PARSEC and AxBench
- **Performance**: Full-system cycle-level simulation
- **Error**: Pin simulation
- **Area and Energy**: CACTI
- **Configuration**:
  - 4 cores, private L1 and L2
  - 2MB shared LLC (1MB precise, 1MB Doppelgänger)
  - Doppelgänger: 14-bit similarity map, 1/4 data array

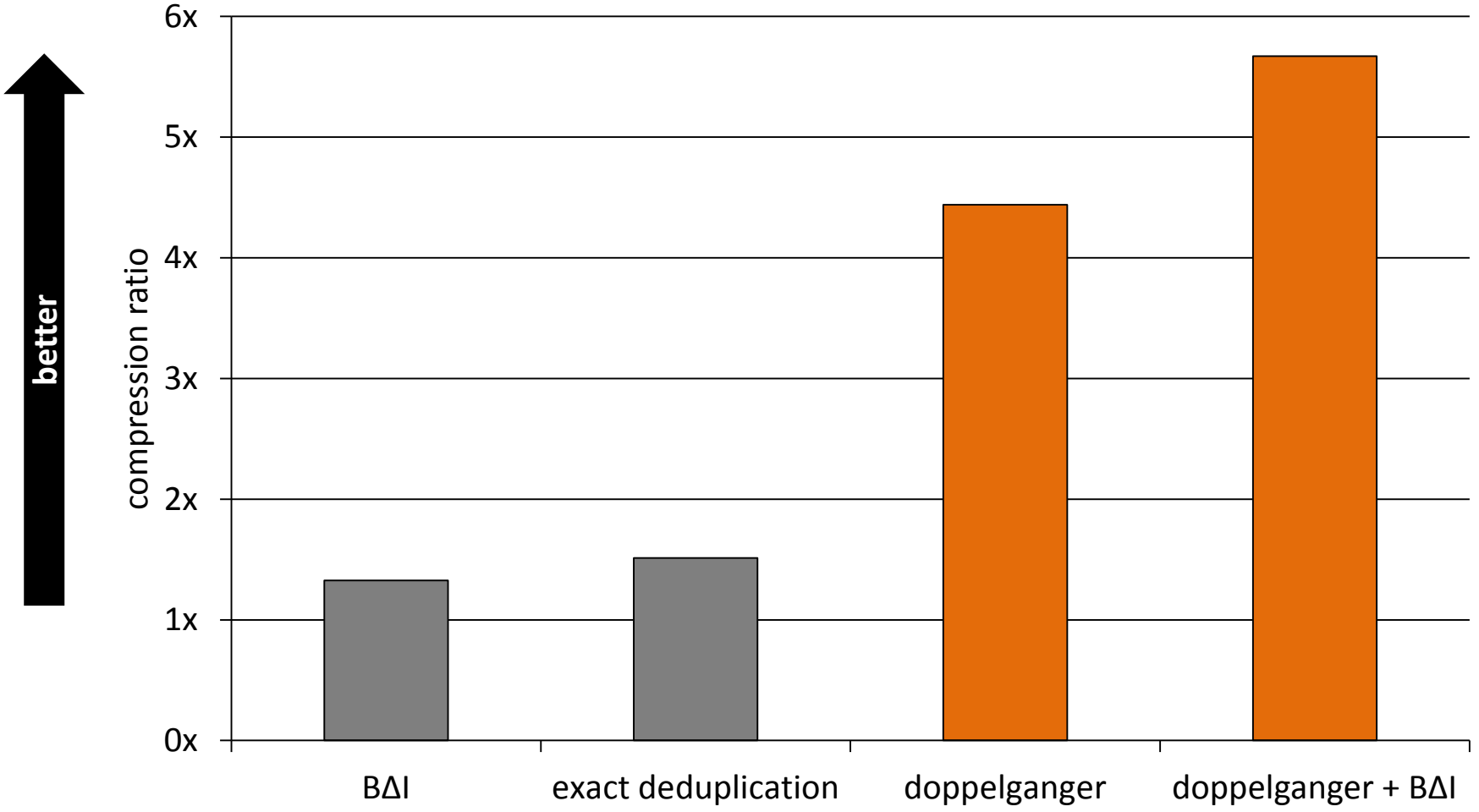
Evaluation - Compression Ratio



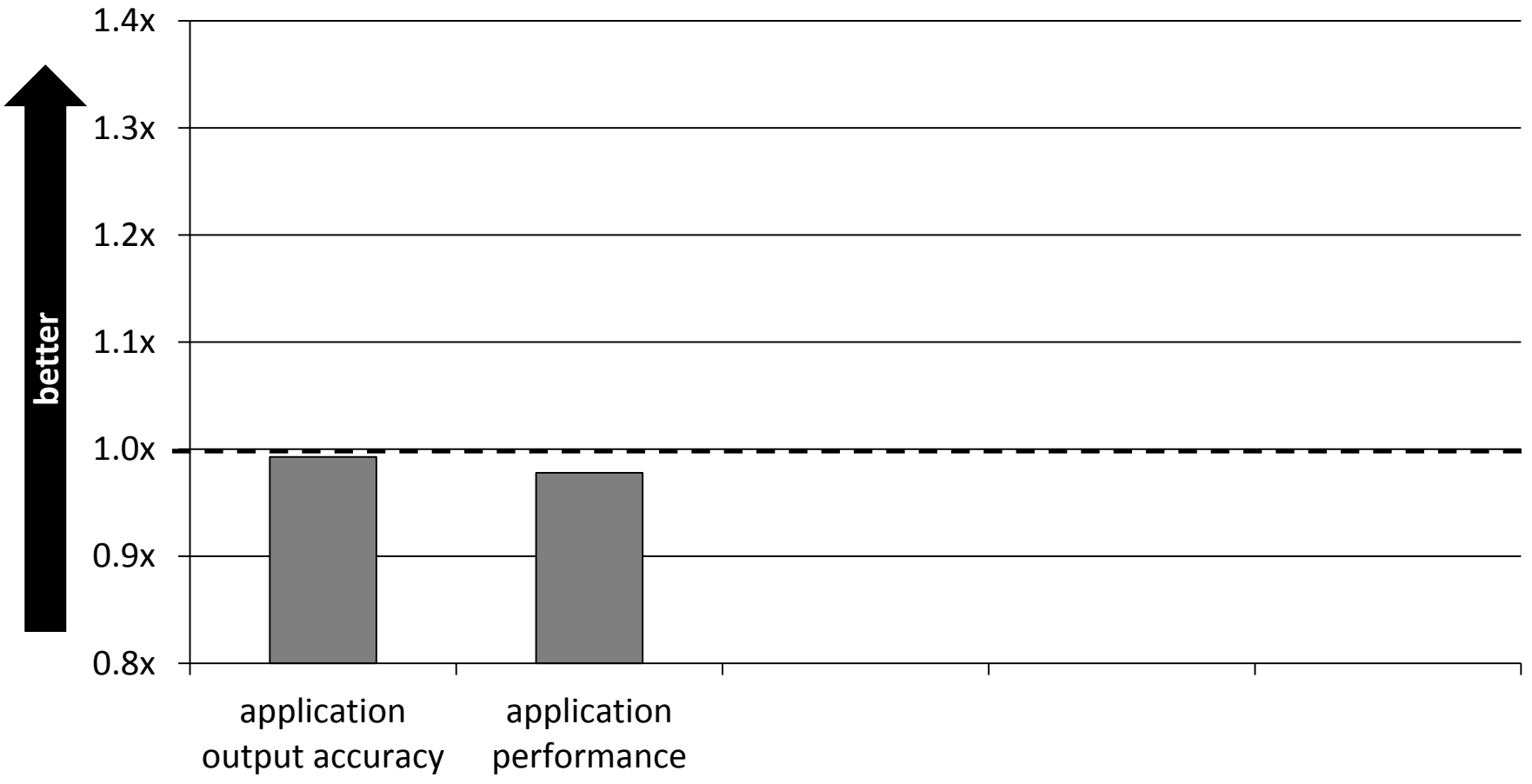
Evaluation - Compression Ratio



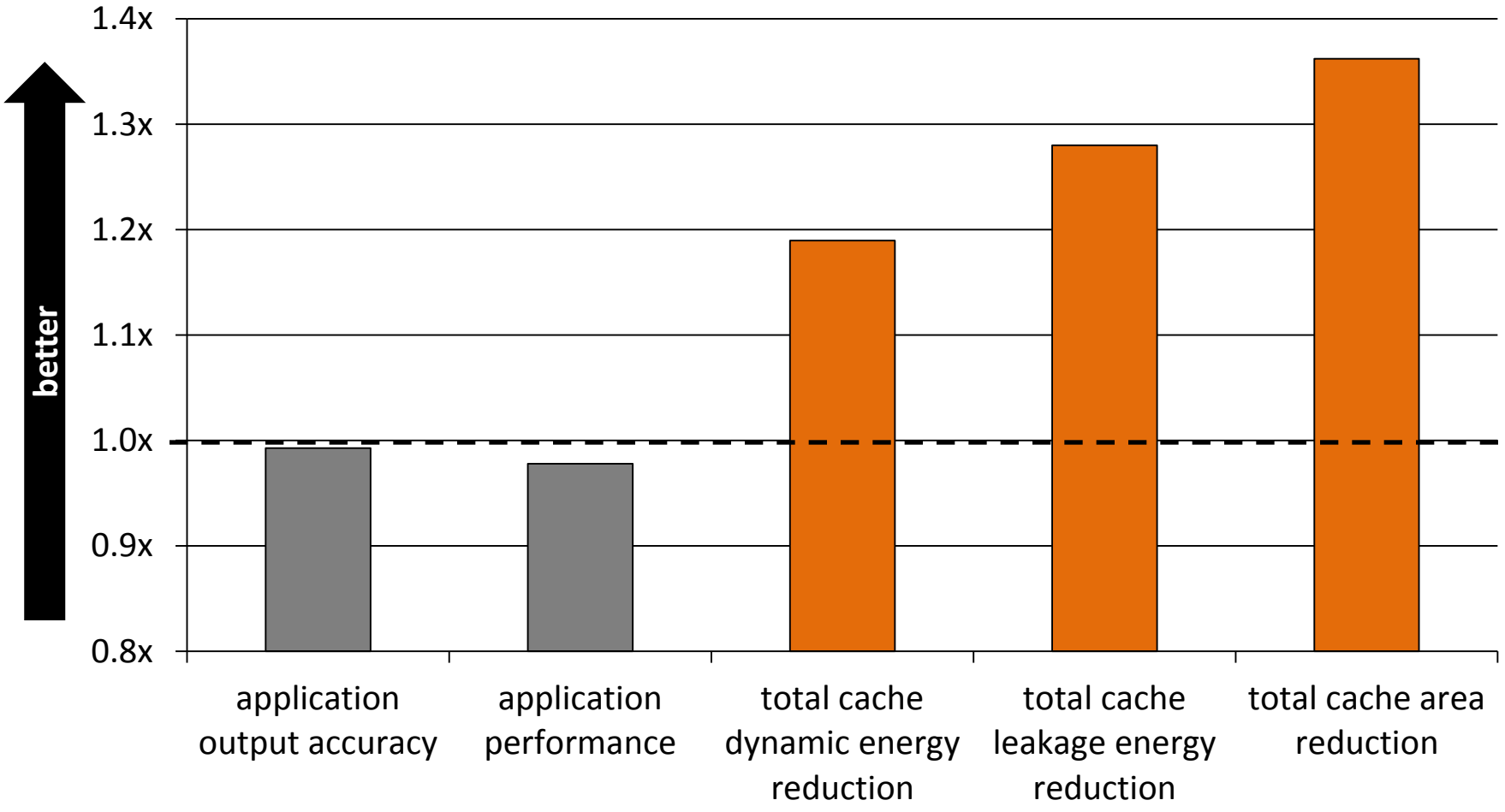
Evaluation - Compression Ratio



Evaluation


1.4x
1.3x
1.2x
1.1x
1.0x
0.9x
0.8x
better
application output accuracy
application performance

Evaluation


1.4x
1.3x
1.2x
1.1x
1.0x
0.9x
0.8x
better
application output accuracy
application performance
total cache dynamic energy reduction
total cache leakage energy reduction
total cache area reduction

Conclusion

Doppelgänger Cache:

- Identifies **approximate similarity** in data block values.
  - **77%** cache storage savings of approximable data.
- Effectively compresses storage of approximately similar blocks.
  - **3x** better compression ratio than state-of-the-art techniques.
- Significantly reduces area and energy consumption.
  - Reduces total on-chip cache area by **1.36x**.

Thank you

Doppelgänger: A Cache for Approximate Computing

Joshua San Miguel
Jorge Albericio
Andreas Moshovos
Natalie Enright Jerger

The Edward S. Rogers Sr. Department
of Electrical & Computer Engineering
UNIVERSITY OF TORONTO