The Bunker Cache for Spatio-Value Approximation

Joshua San Miguel

Jorge Albericio

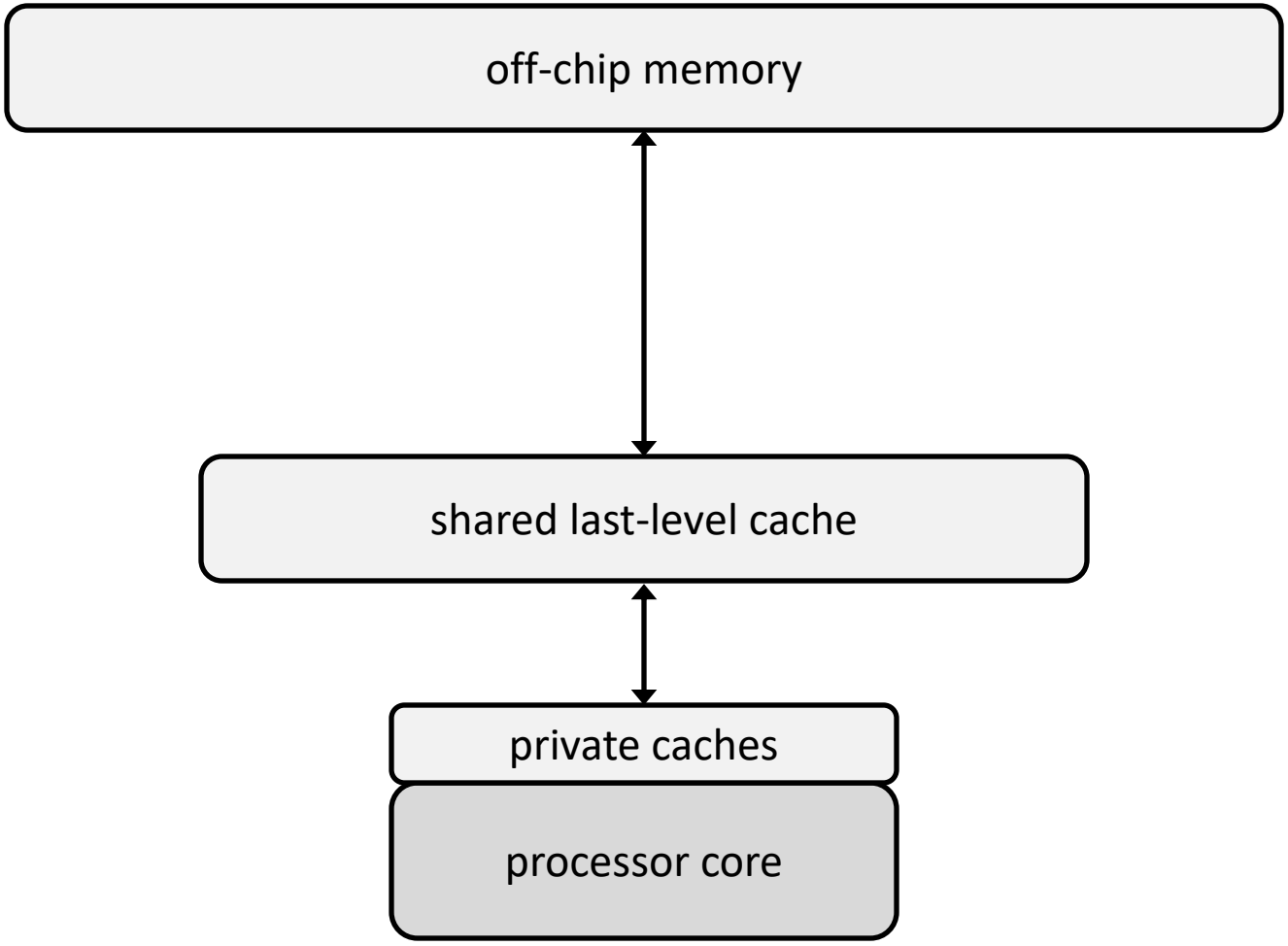
Natalie Enright Jerger

Aamer Jaleel

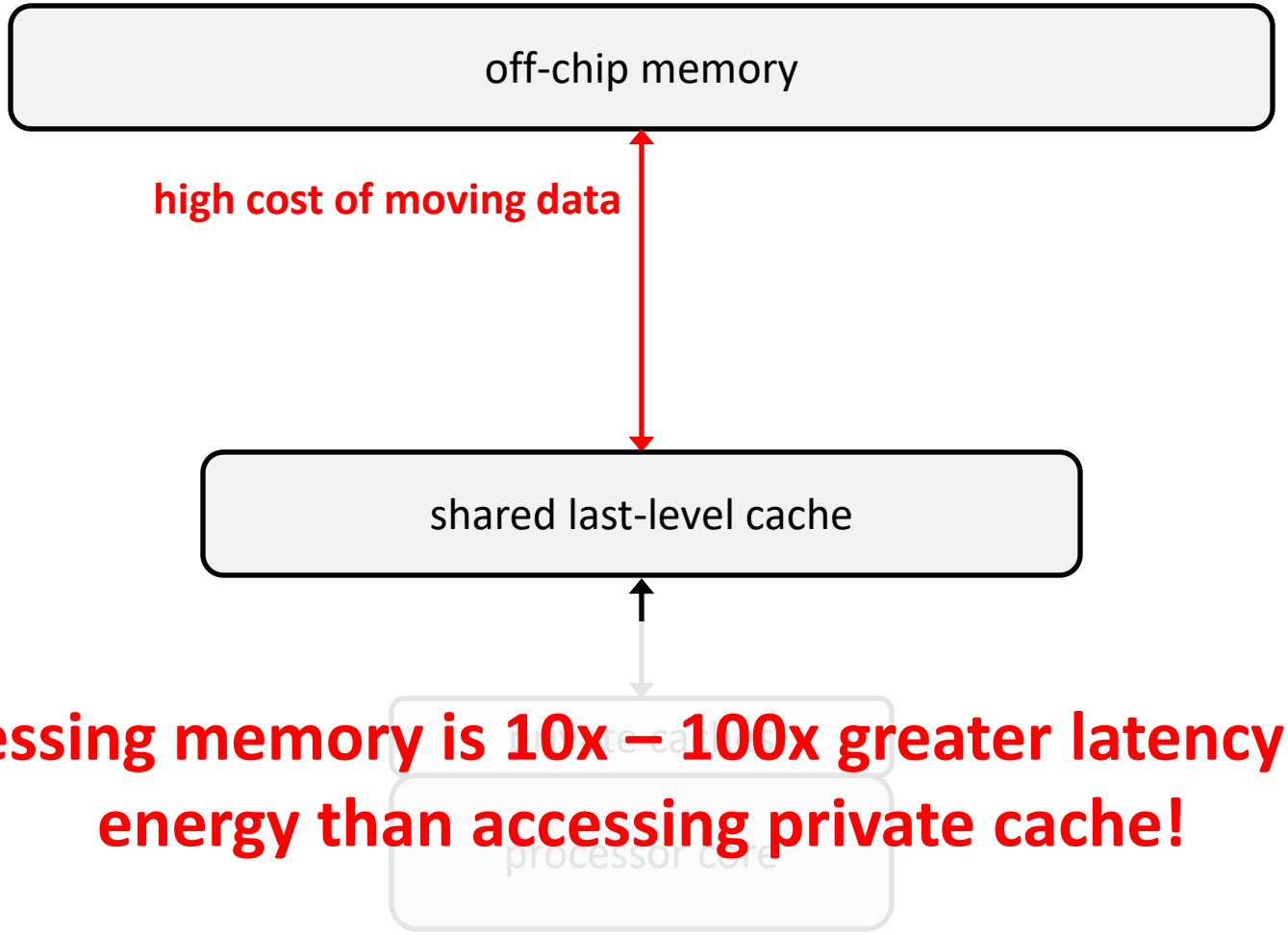
The Edward S. Rogers Sr. Department
of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

NVIDIA

Data Movement and Storage

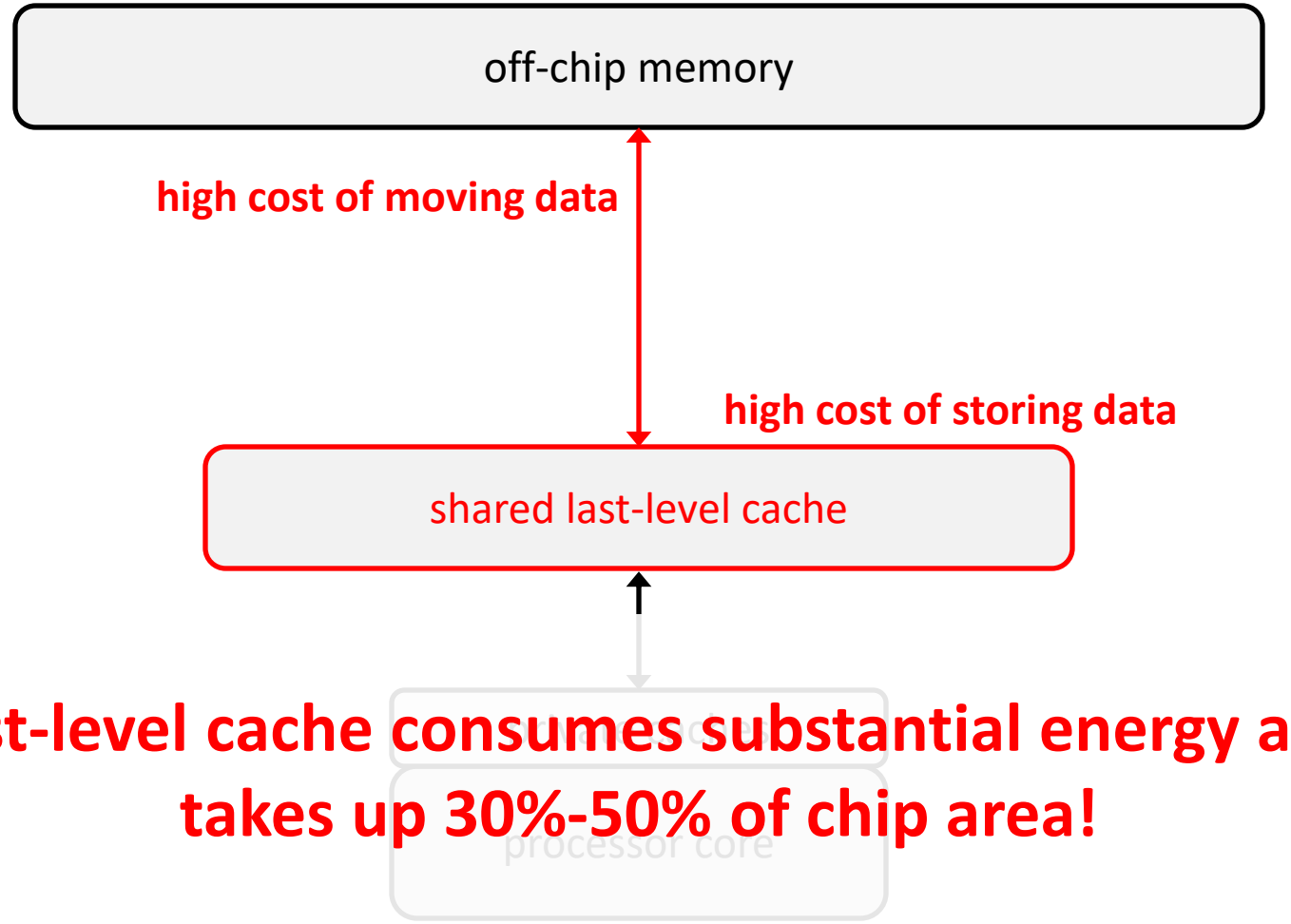


Data Movement and Storage



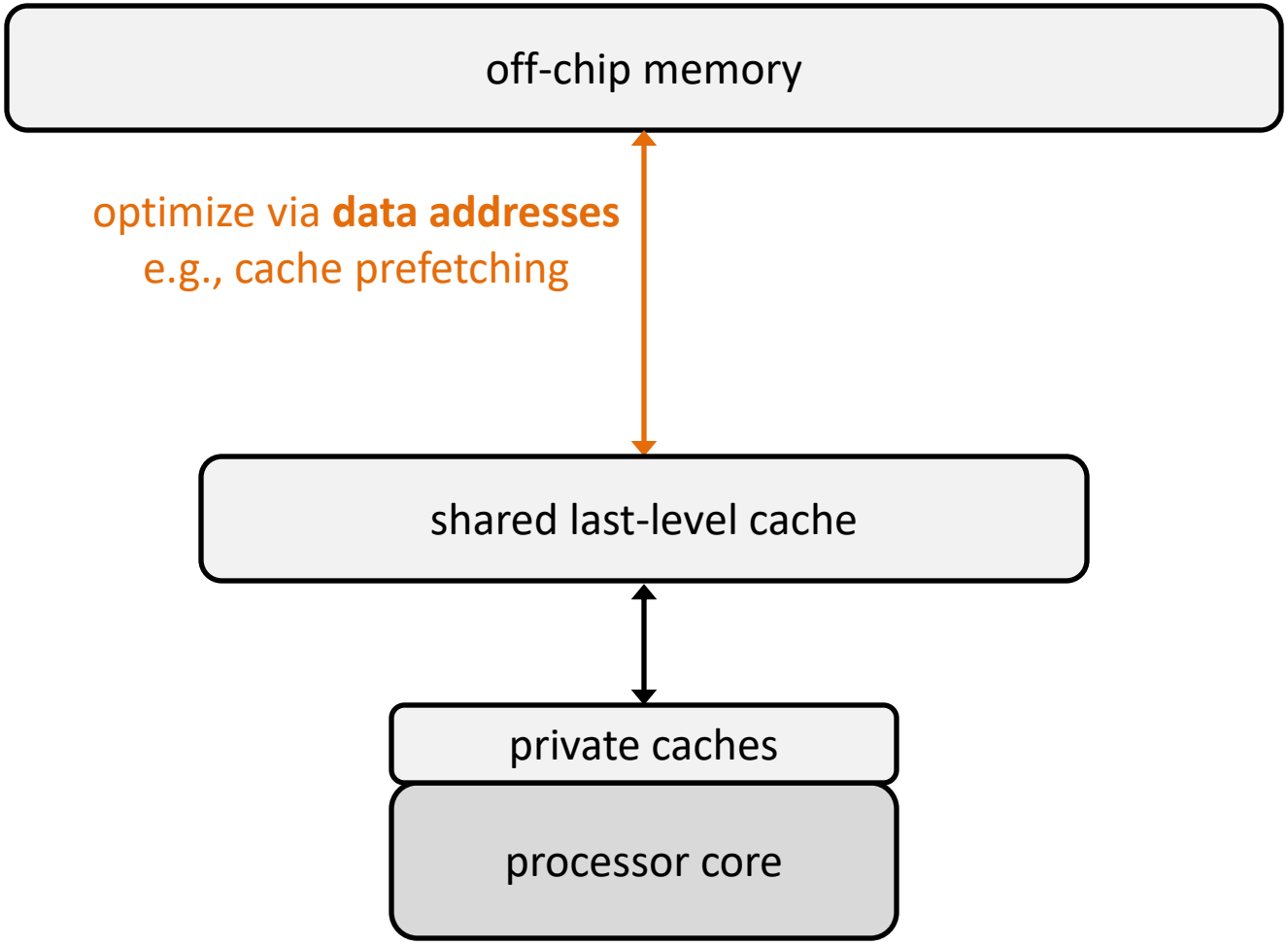
Accessing memory is 10x – 100x greater latency and energy than accessing private cache!

Data Movement and Storage

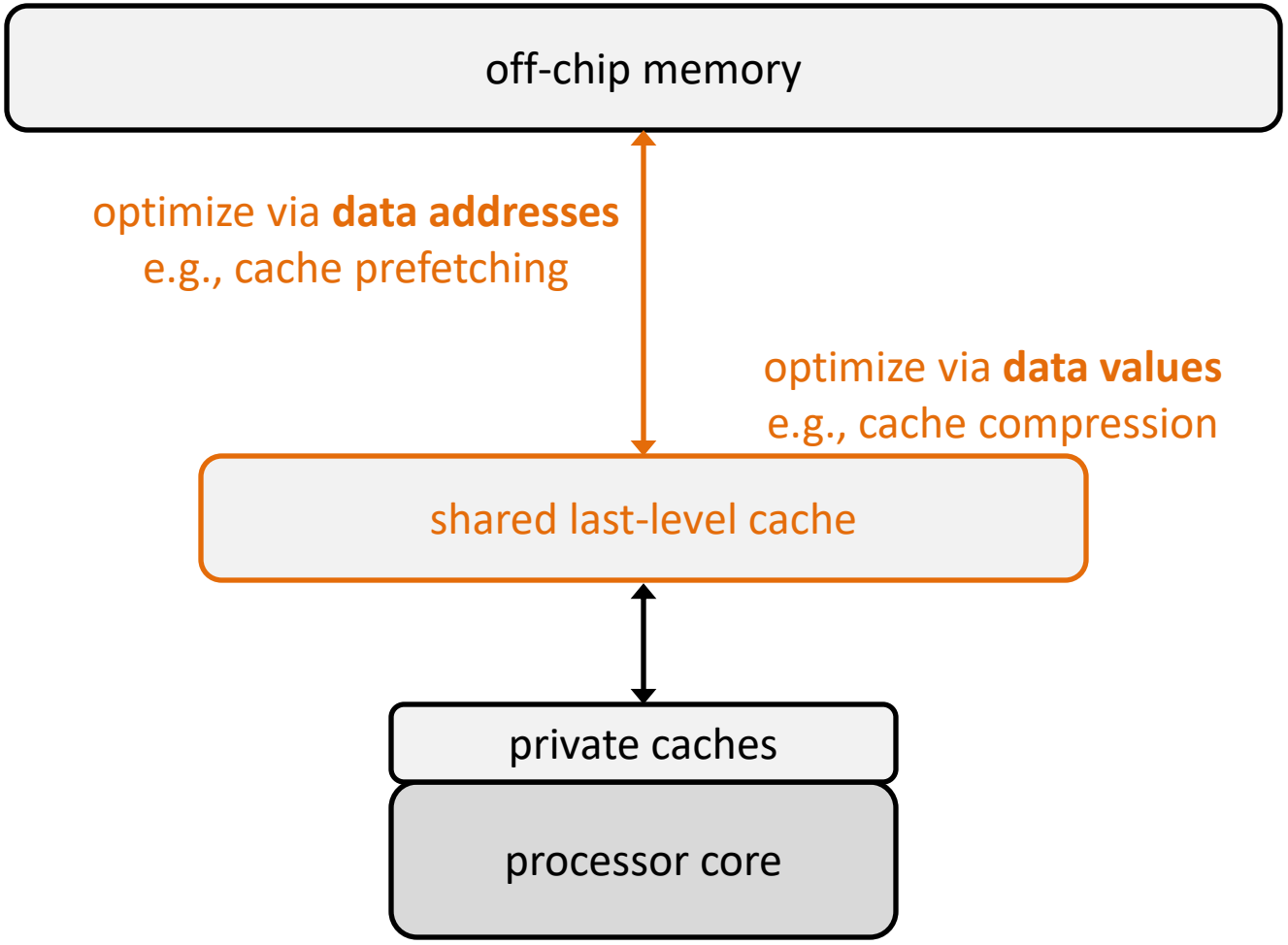


Last-level cache consumes substantial energy and takes up 30%-50% of chip area!

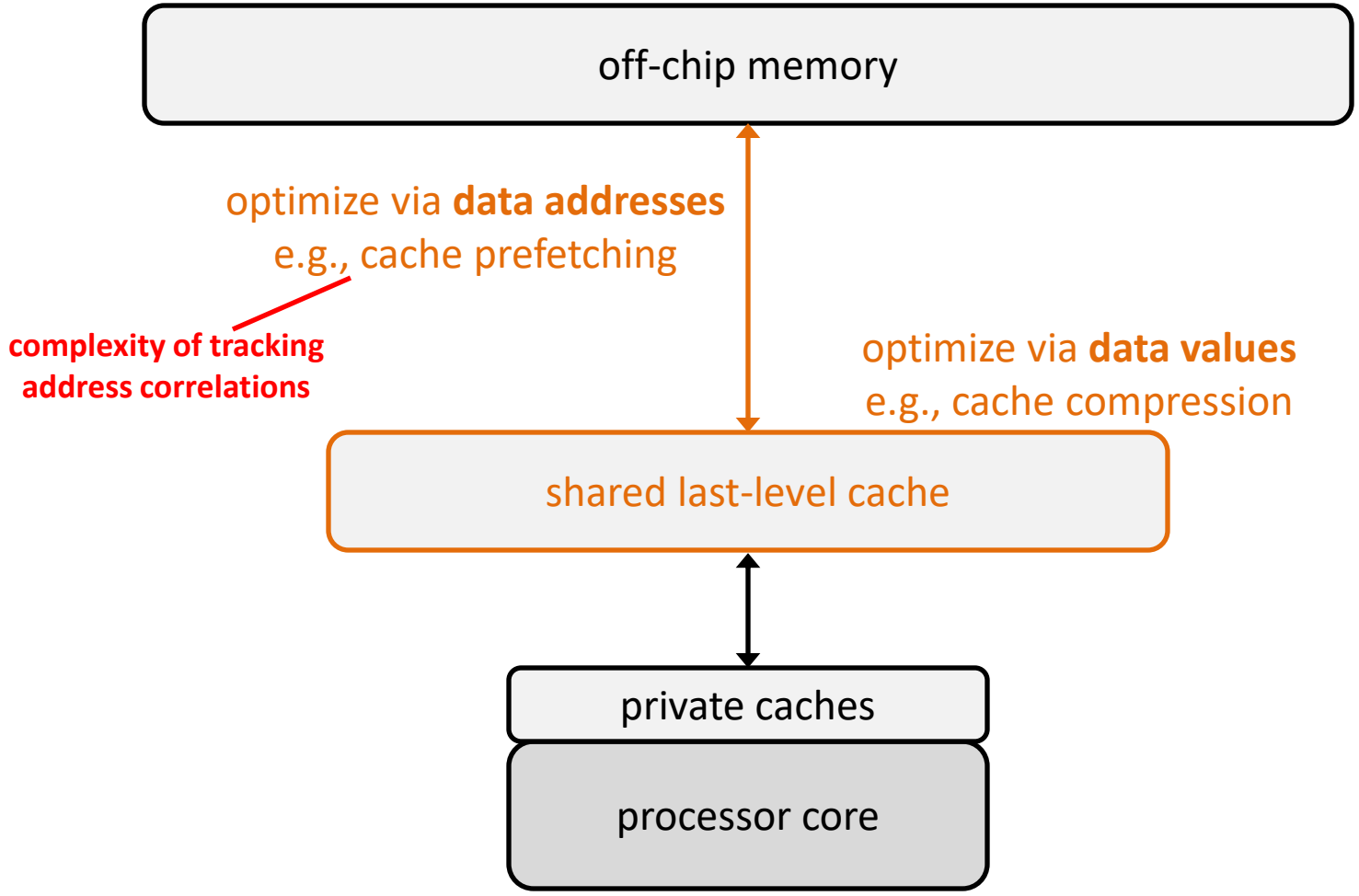
Data Movement and Storage



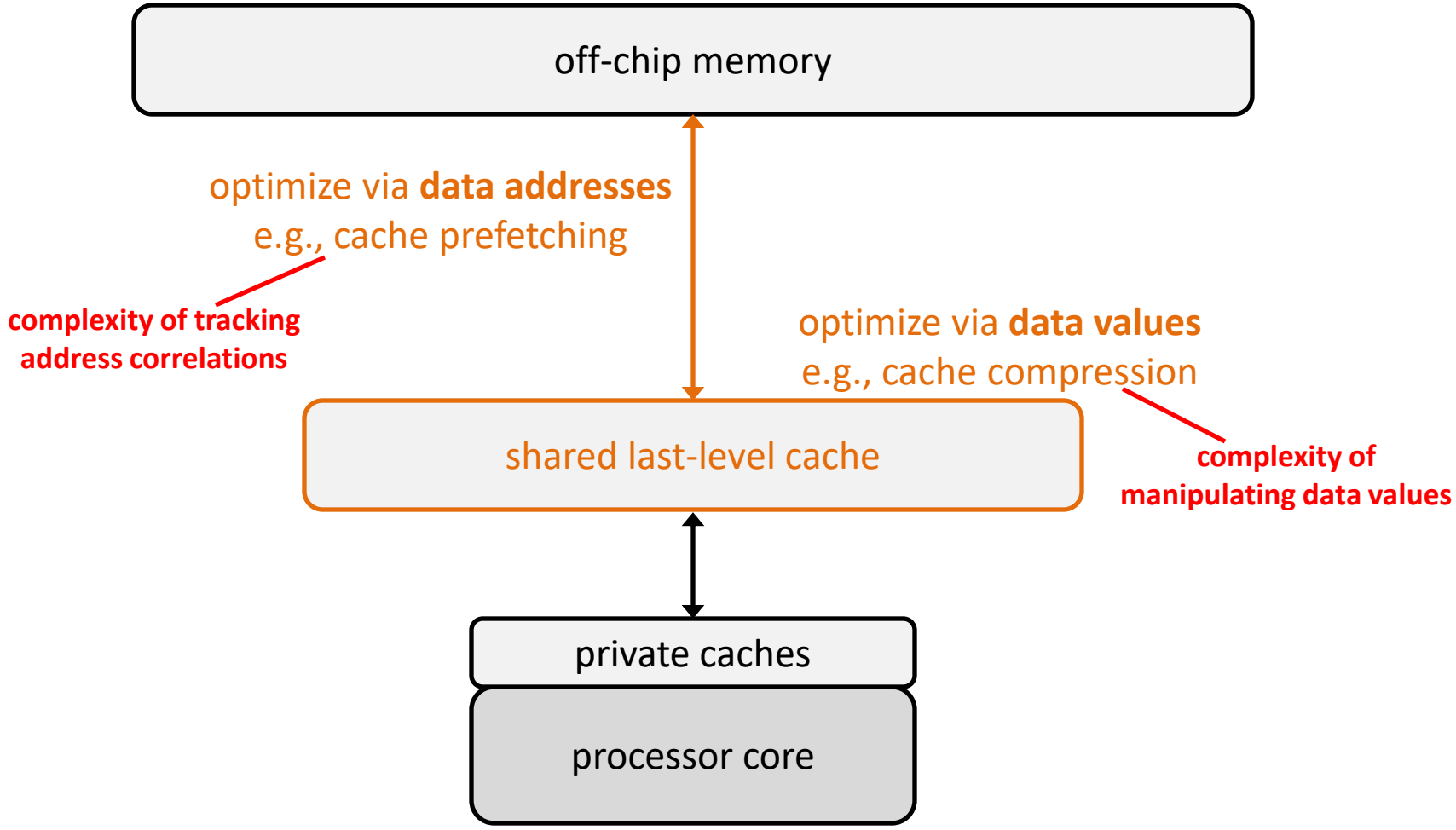
Data Movement and Storage



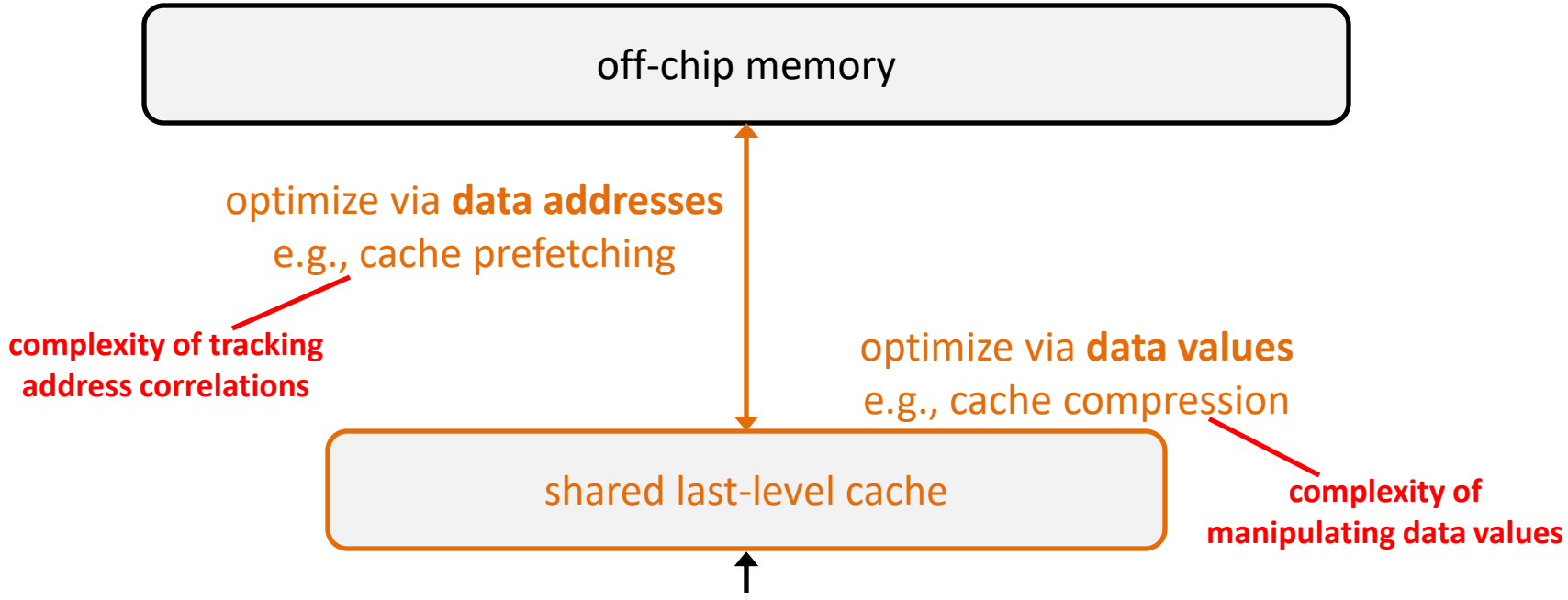
Data Movement and Storage



Data Movement and Storage



Data Movement and Storage



Can we improve data movement and storage simultaneously without the added complexities?

(where data is located? ⇔ what value is encoded in data?)

Our Work

We explore **Spatio-Value Similarity**:

- there is regularity to where approximately similar values are located in memory

We propose the **Bunker Cache**:

- many-to-one similarity mapping based on memory address
- savings in runtime (1.58x), dynamic energy (1.72x), leakage power (1.65x) at acceptable quality levels

Spatio-Value Similarity

where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

Spatio-Value Similarity

where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

data values can often be approximate and continuous (i.e., smooth)

Spatio-Value Similarity

where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

data values can often be approximate and continuous (i.e., smooth)

data addresses represent a one-dimensional memory space

Spatio-Value Similarity

where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

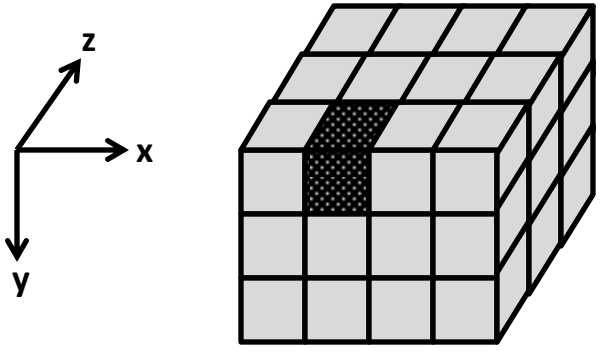
- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.

Spatio-Value Similarity

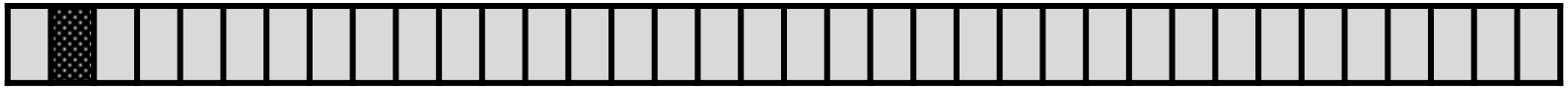
where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.



memory space:

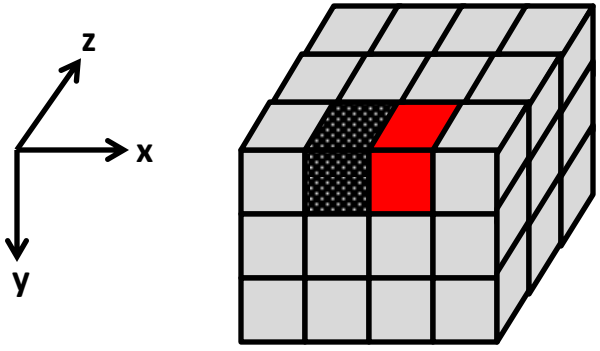


Spatio-Value Similarity

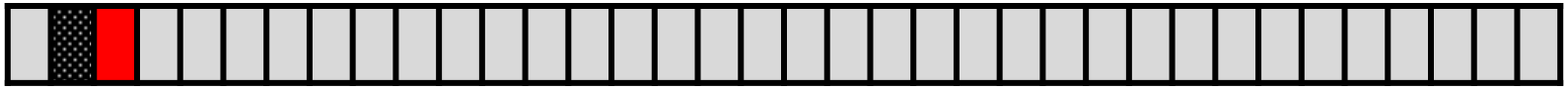
where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.



memory space:

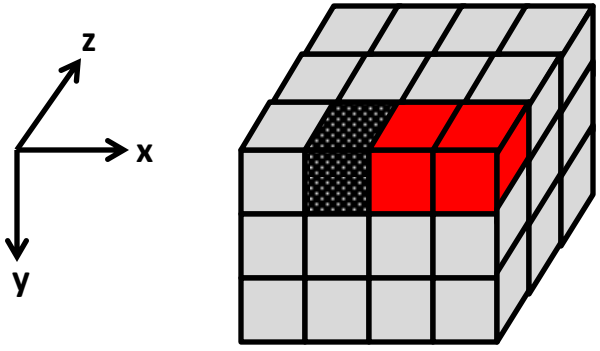


Spatio-Value Similarity

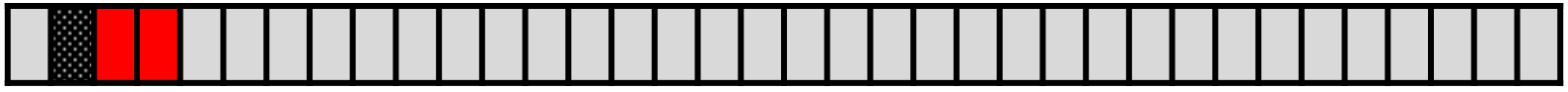
where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.



memory space:

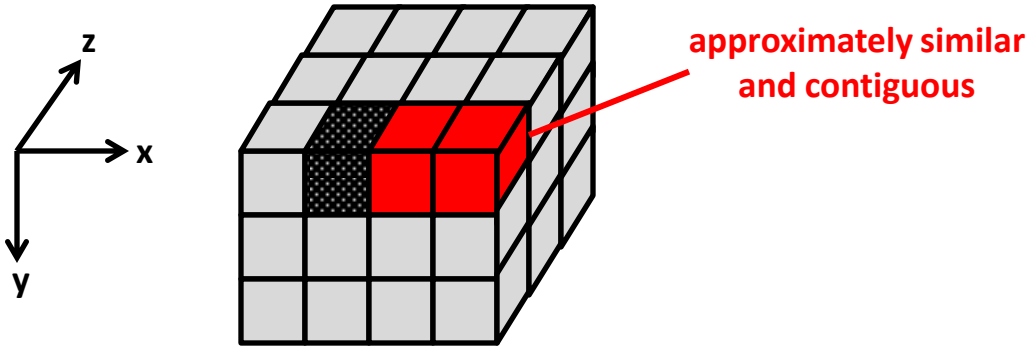


Spatio-Value Similarity

where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.

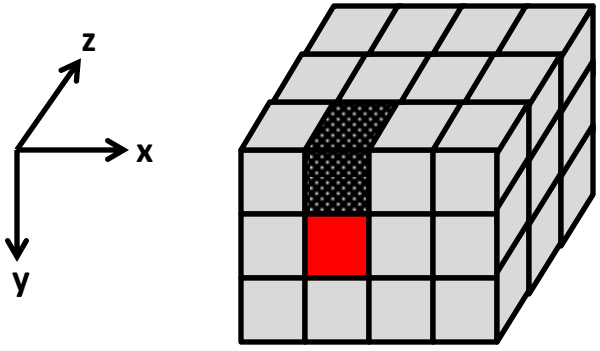


Spatio-Value Similarity

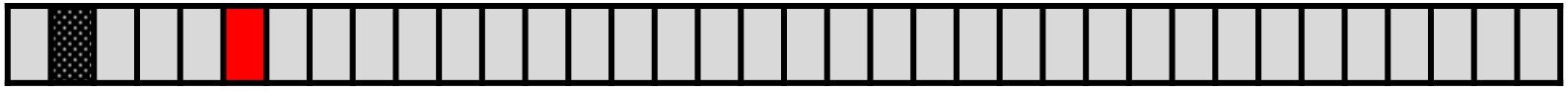
where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.



memory space:

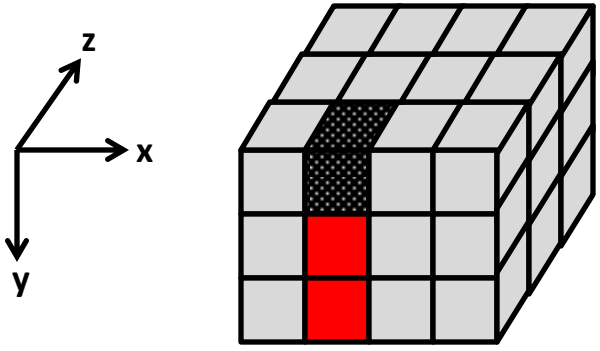


Spatio-Value Similarity

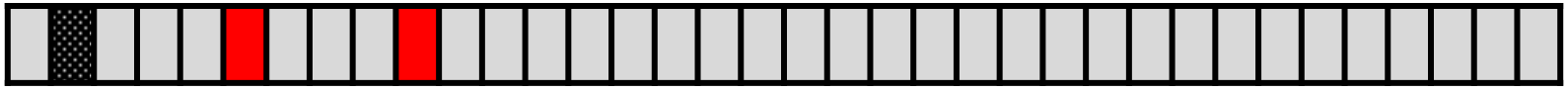
where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.



memory space:

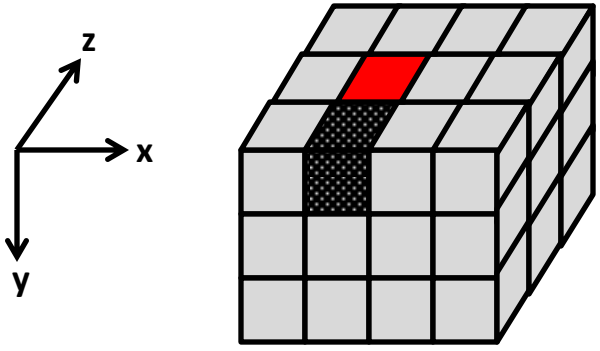


Spatio-Value Similarity

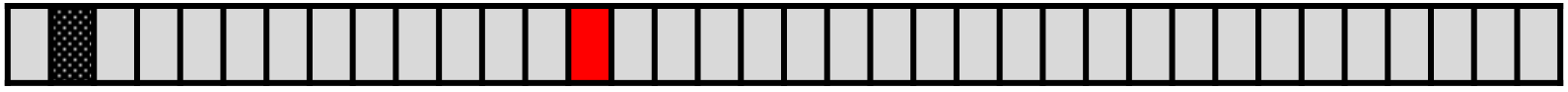
where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.



memory space:

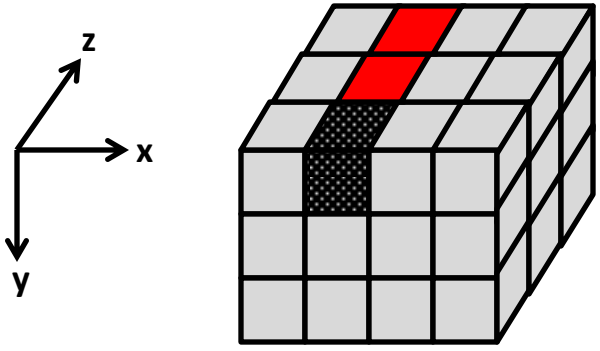


Spatio-Value Similarity

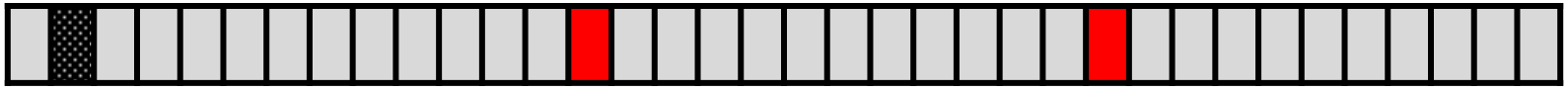
where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.



memory space:

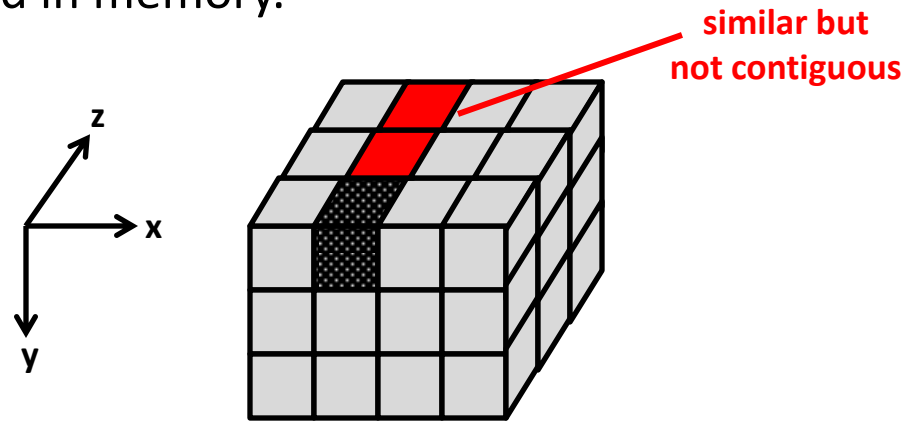


Spatio-Value Similarity

where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.



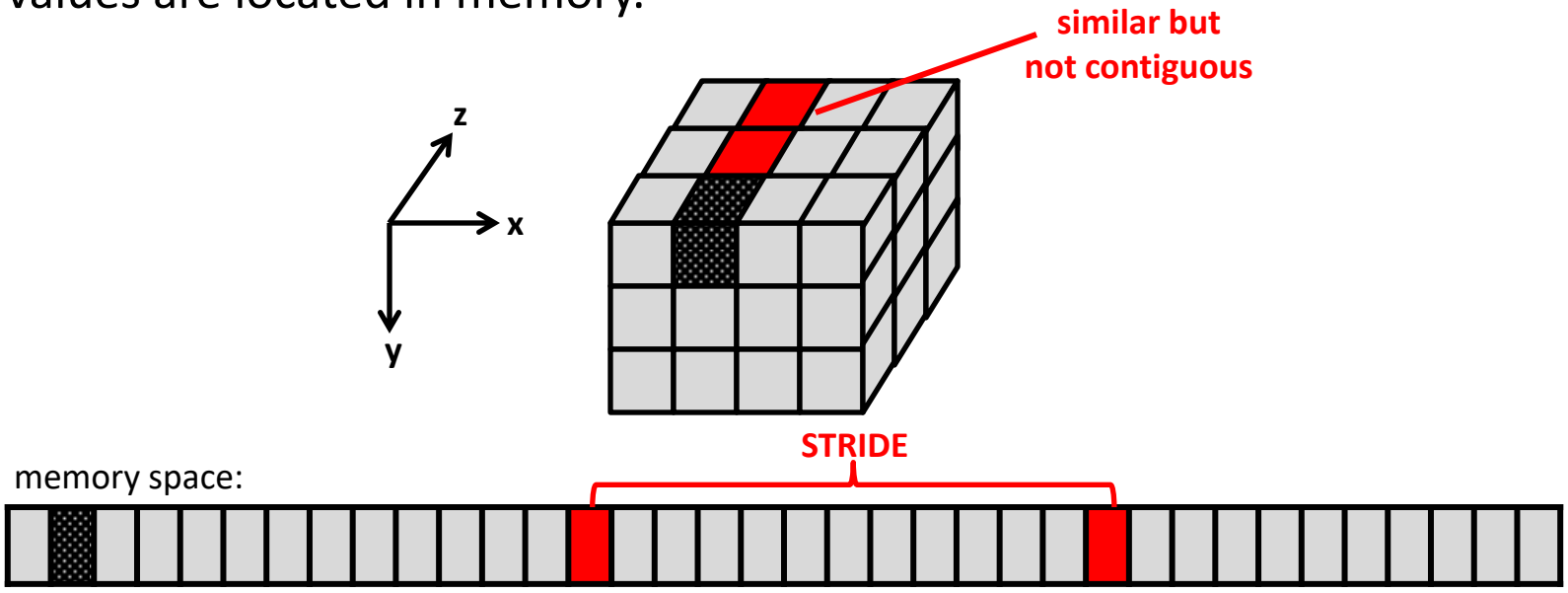
memory space:

Spatio-Value Similarity

where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.



Spatio-Value Similarity

where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.

e.g., image processing

Spatio-Value Similarity

where data is located? ⇔ what value is encoded in data?

The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.

e.g., image processing

similar but not contiguous

STRIDE = image row size

Spatio-Value Similarity

where data is located? ⇔ what value is encoded in data?

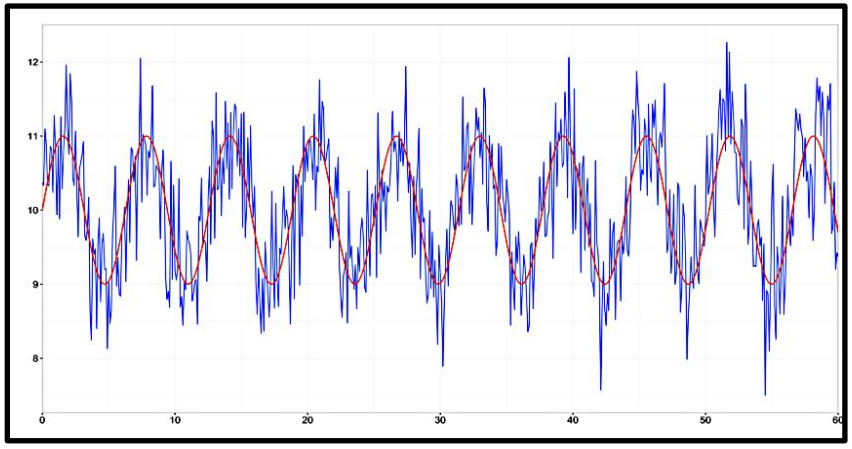
The goal of a processor is to process real-world information, not bits.

- **Spatio-Value Similarity**: there is regularity to where approximately similar values are located in memory.

e.g., image processing

STRIDE = image row size

e.g., signal processing



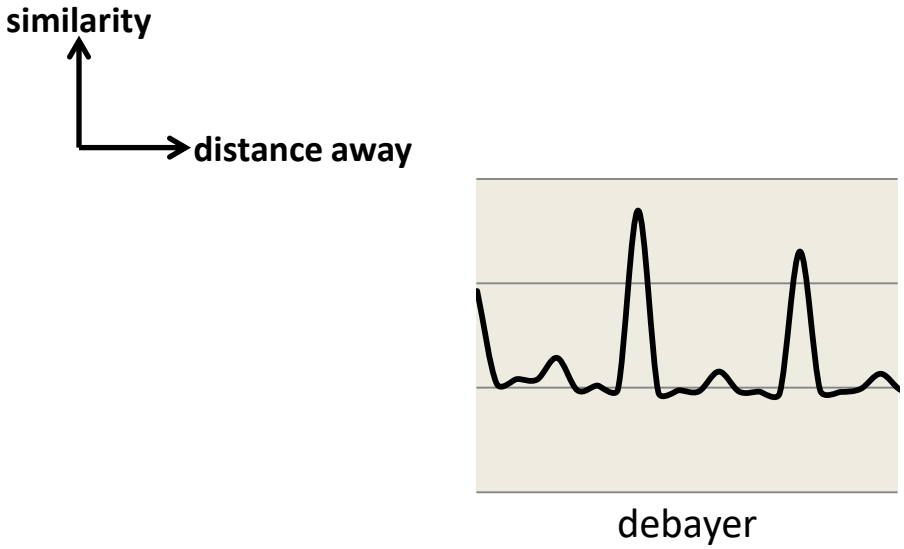
STRIDE = signal period

Spatio-Value Similarity

Given any data block, how similar is it to the block that is distance X away from it?

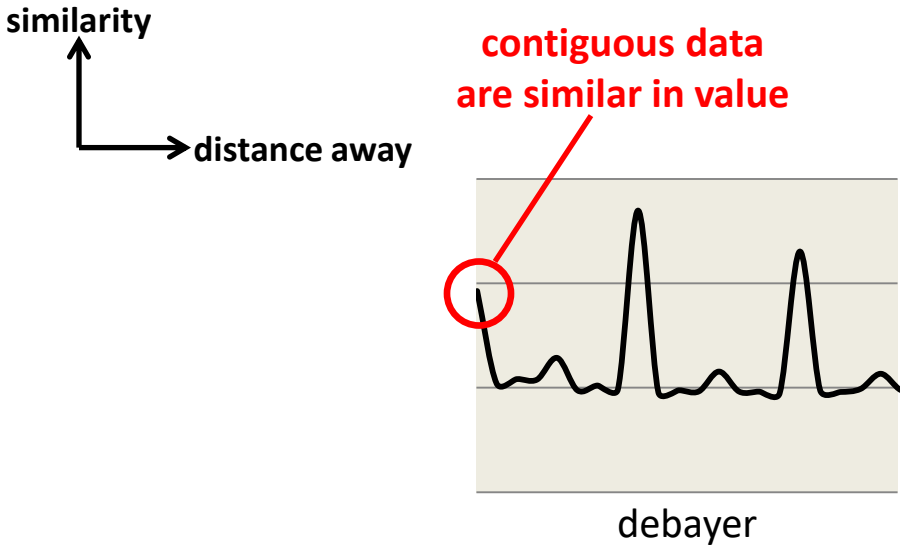
Spatio-Value Similarity

Given any data block, how similar is it to the block that is distance X away from it?



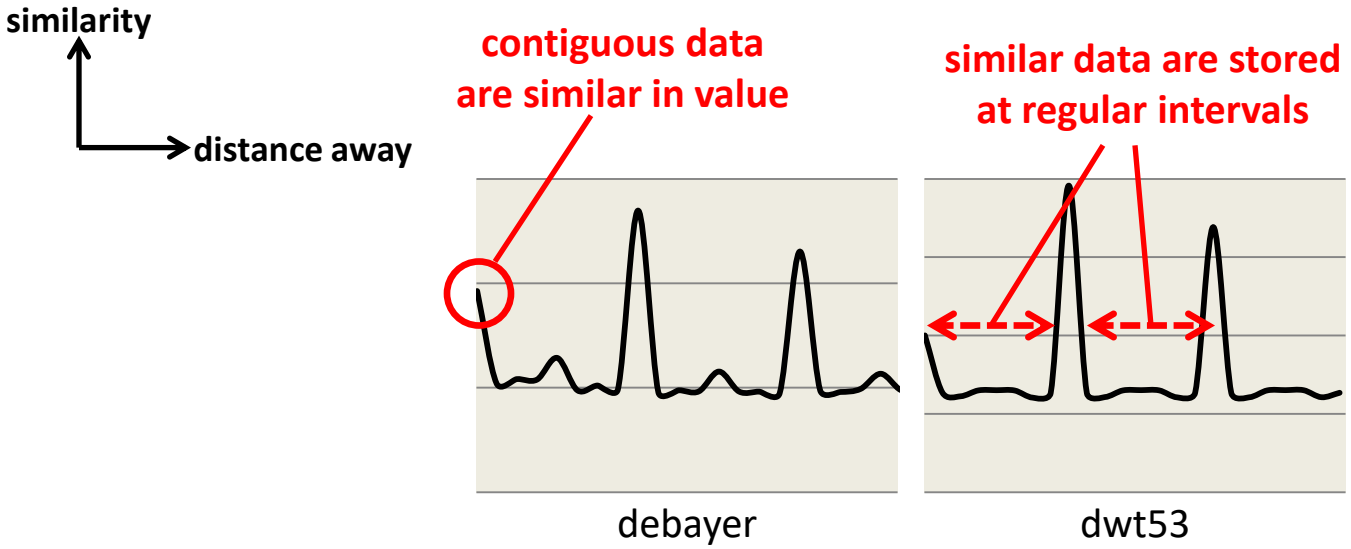
Spatio-Value Similarity

Given any data block, how similar is it to the block that is distance X away from it?



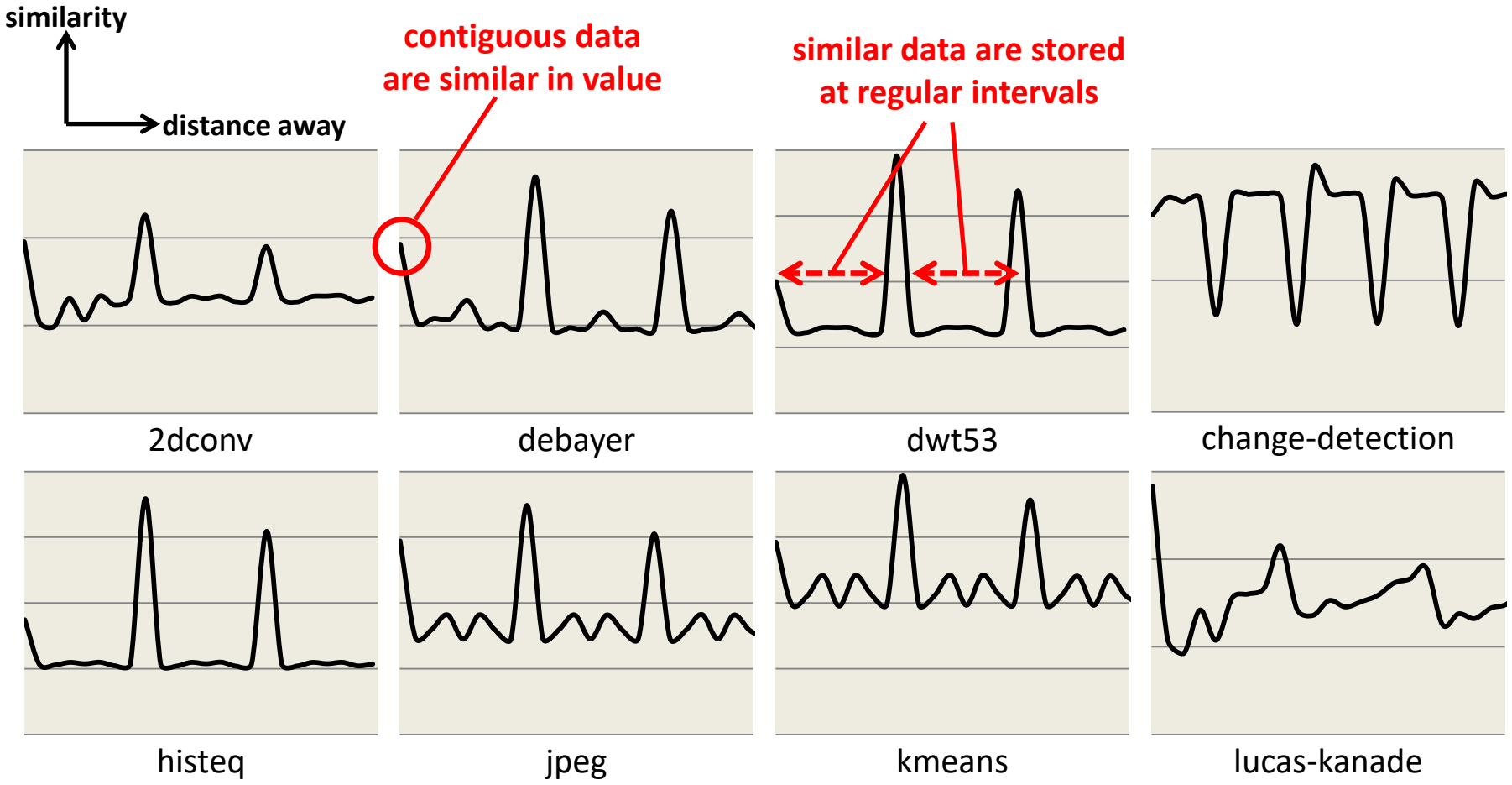
Spatio-Value Similarity

Given any data block, how similar is it to the block that is distance X away from it?

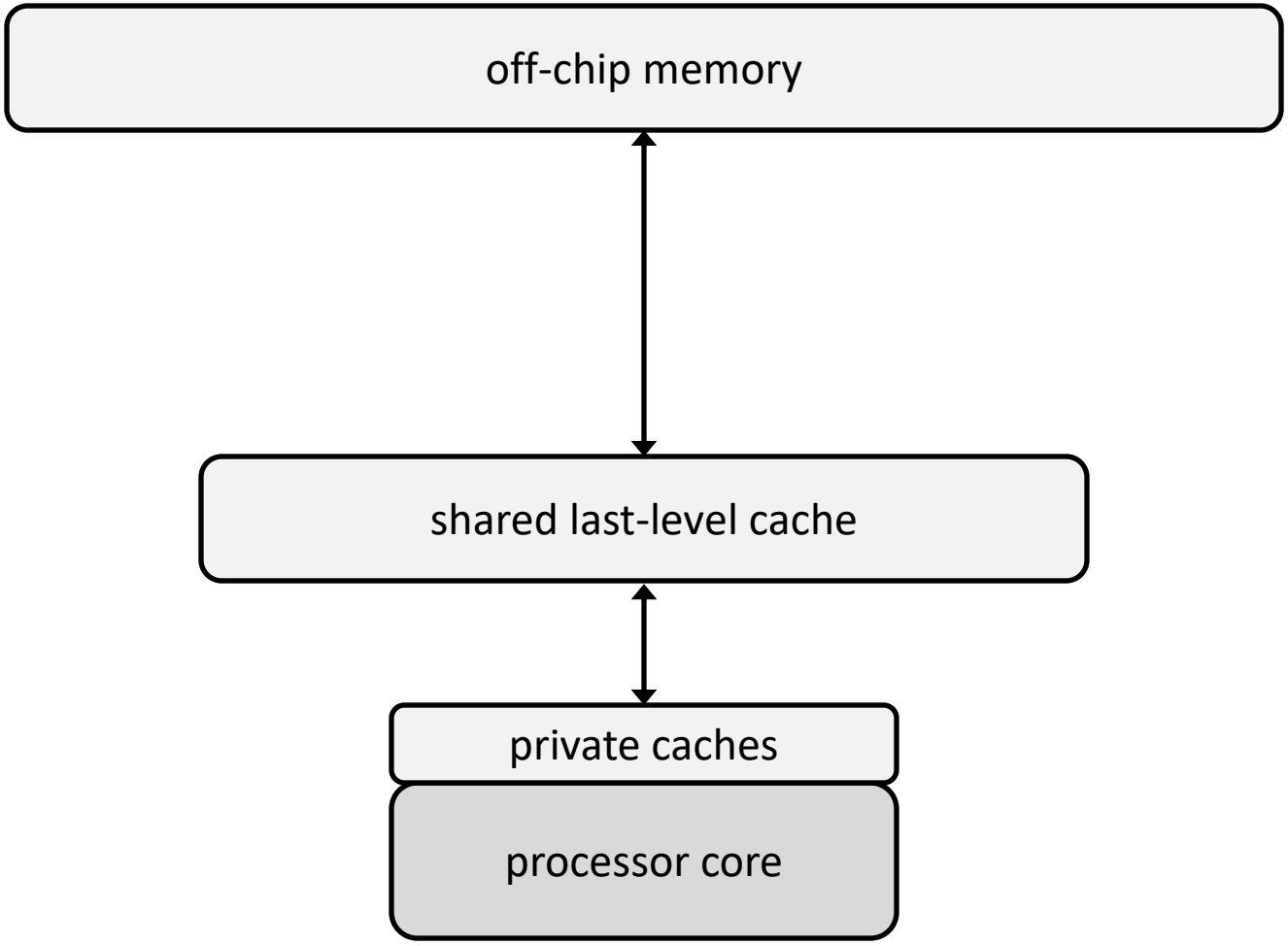


Spatio-Value Similarity

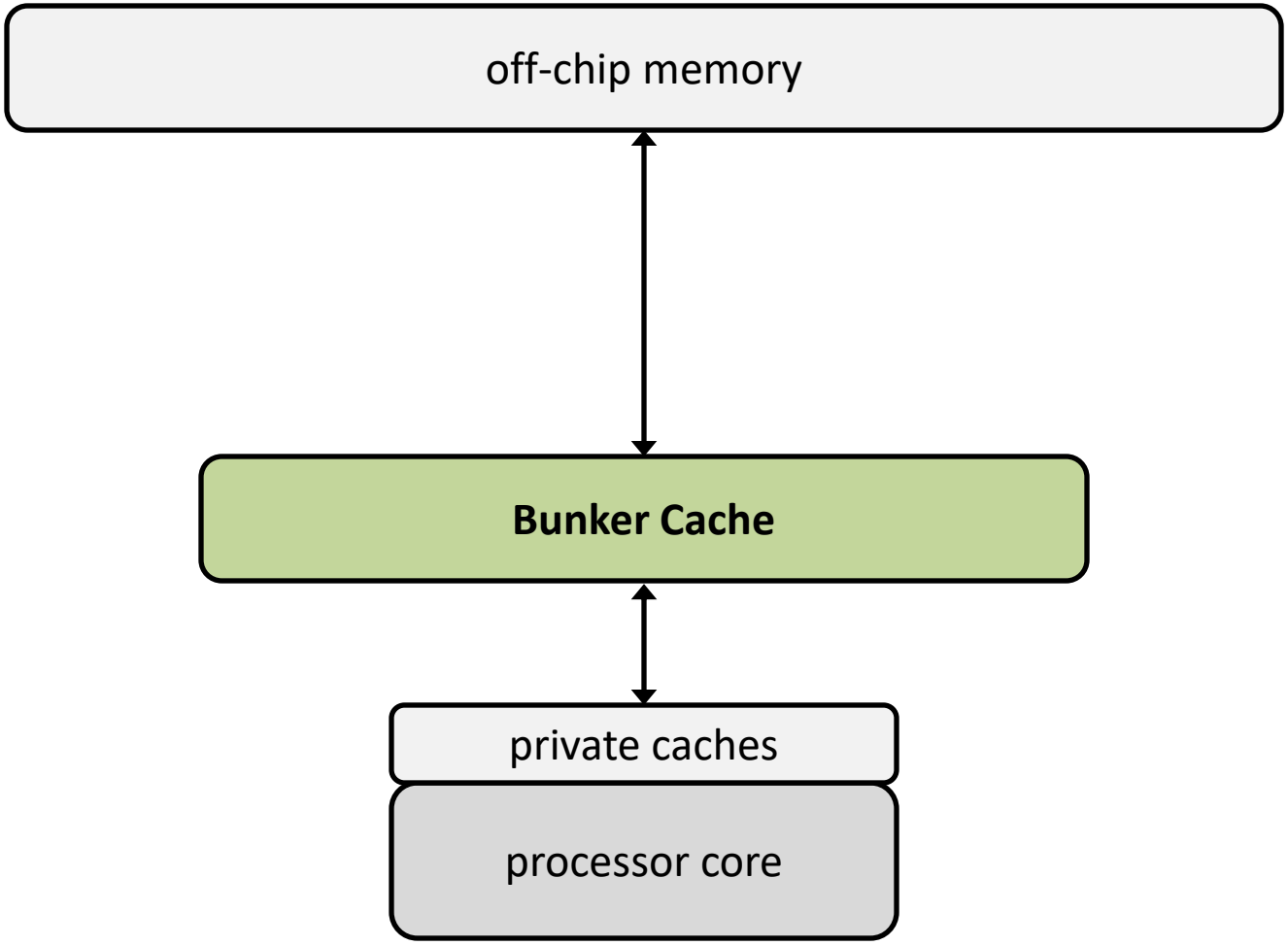
Given any data block, how similar is it to the block that is distance X away from it?



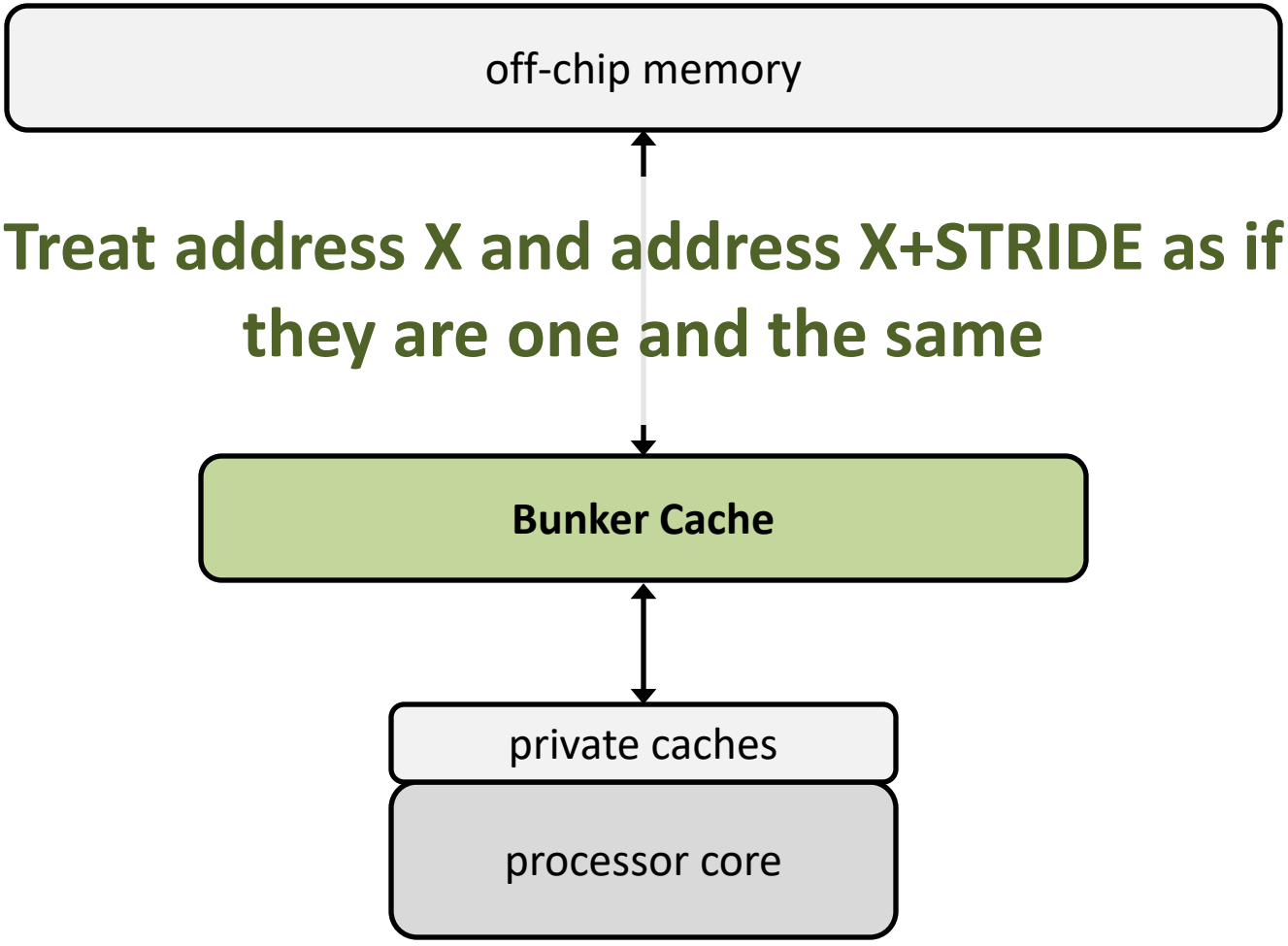
The Bunker Cache



The Bunker Cache

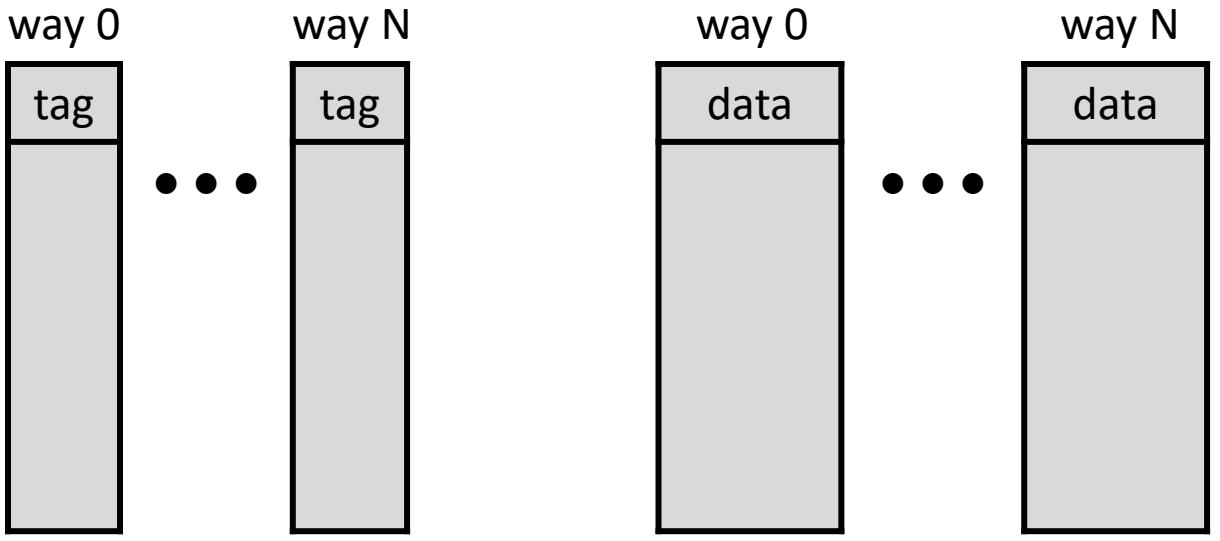


The Bunker Cache



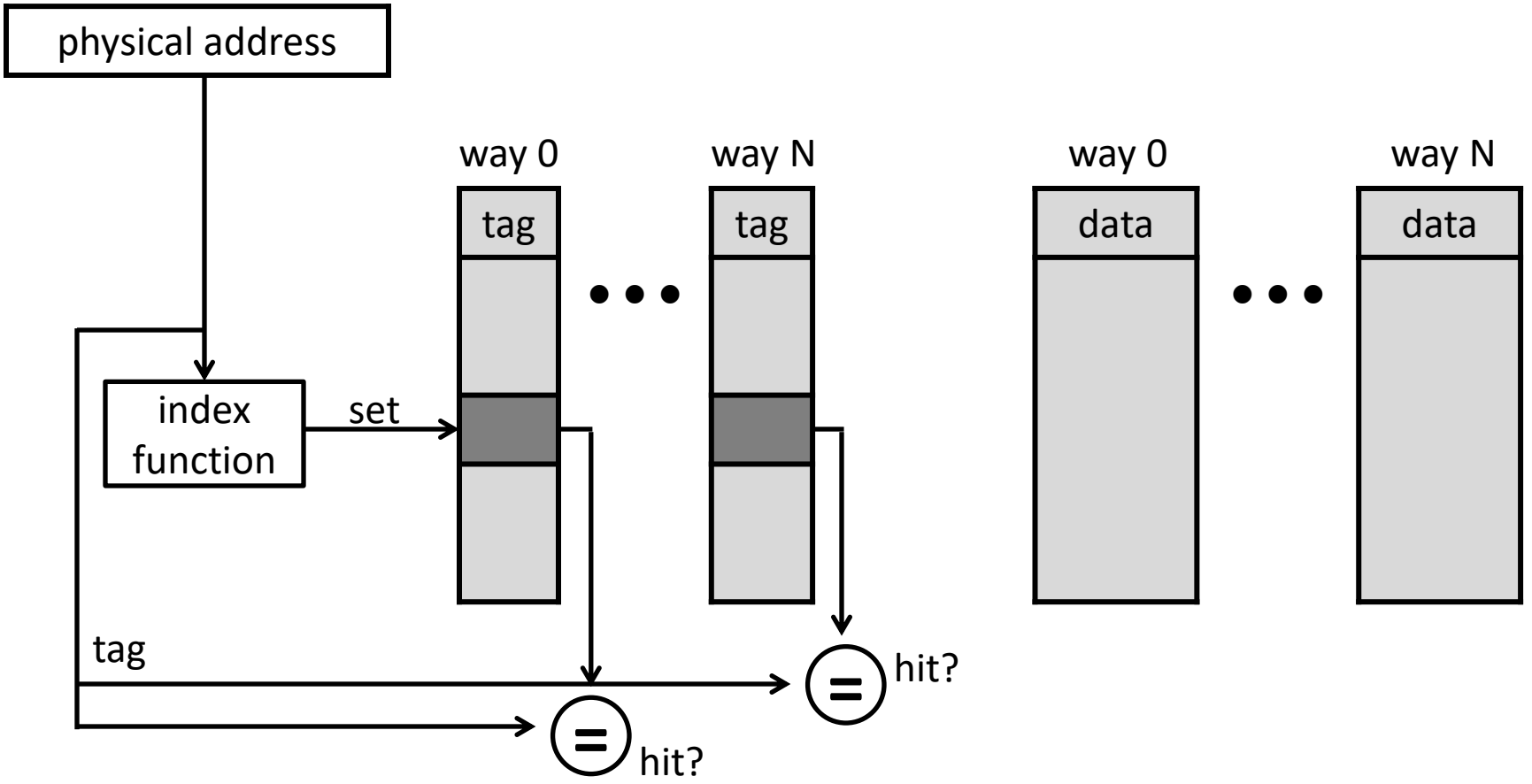
Conventional Cache

physical address

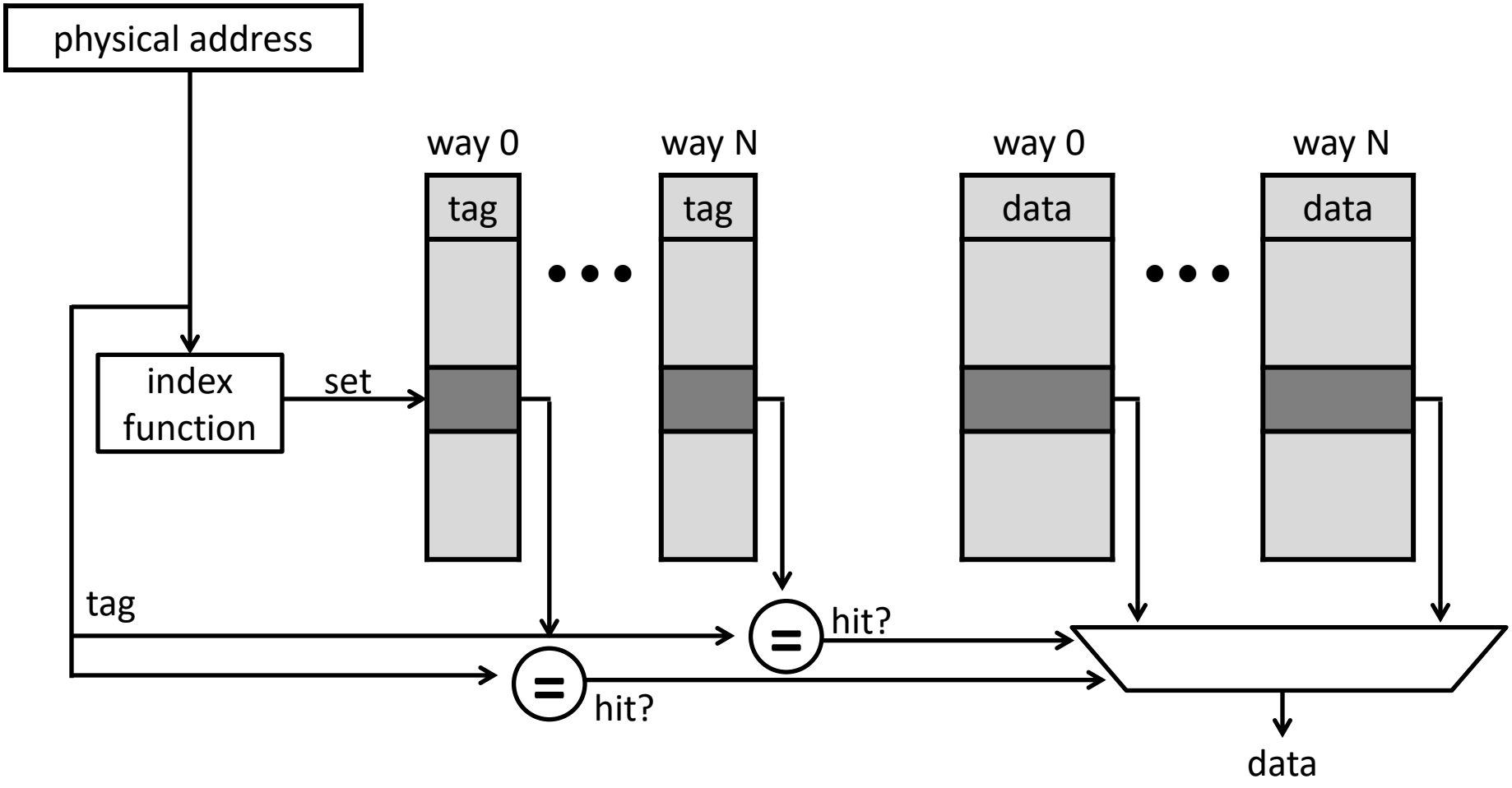


Conventional Cache

Conventional Cache



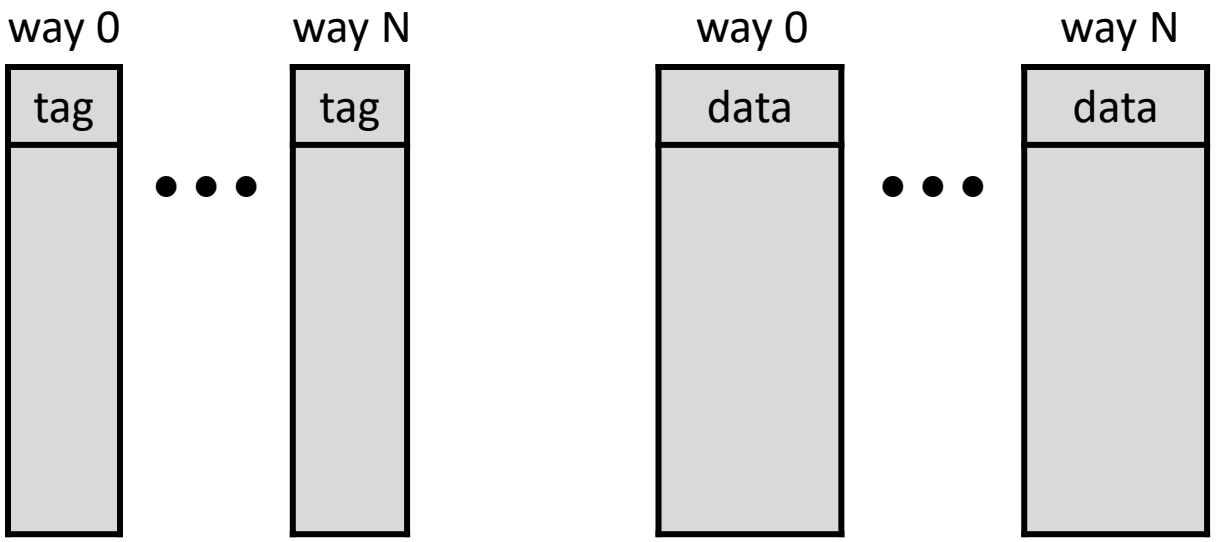
Conventional Cache



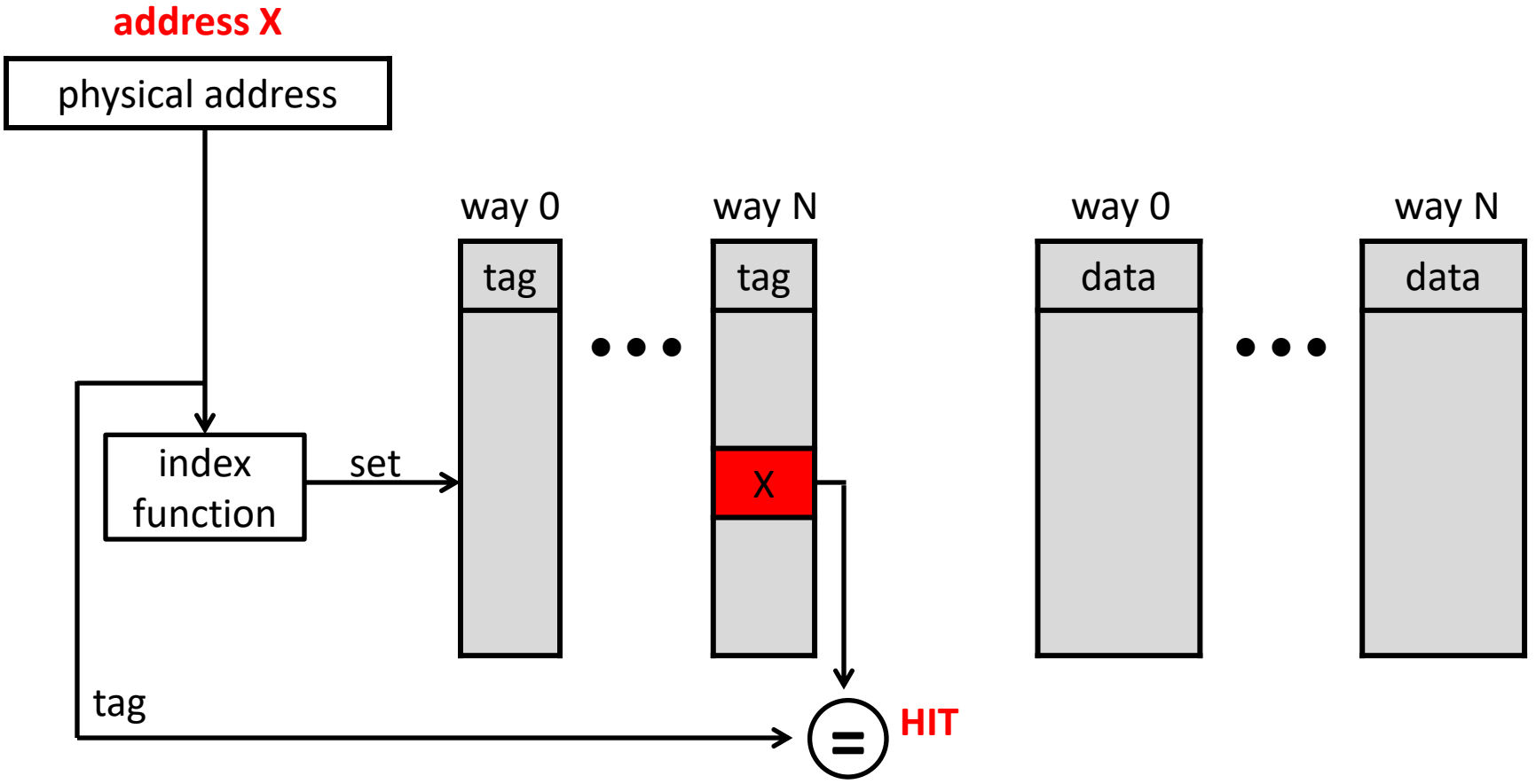
Conventional Cache – Lookup

address X

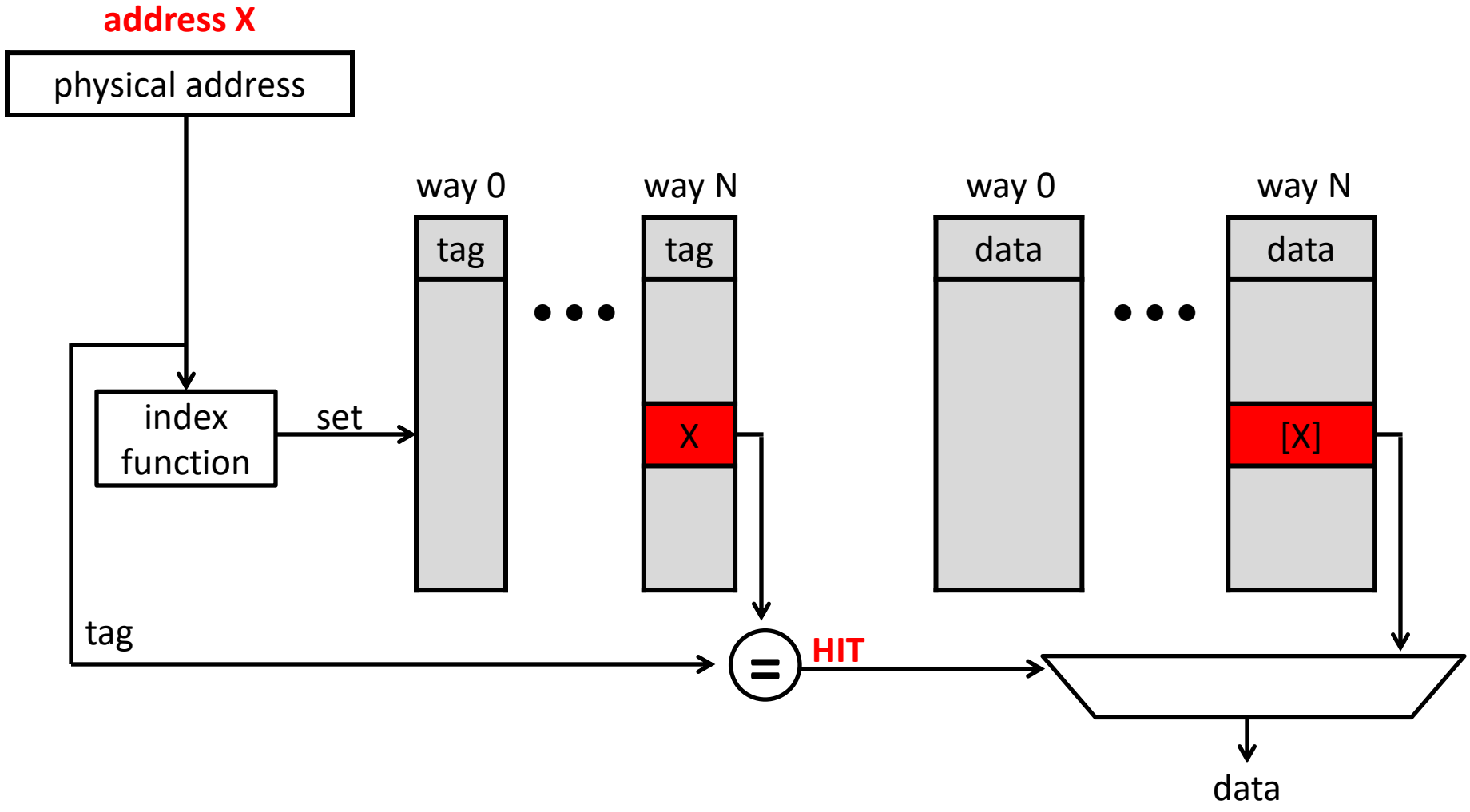
physical address



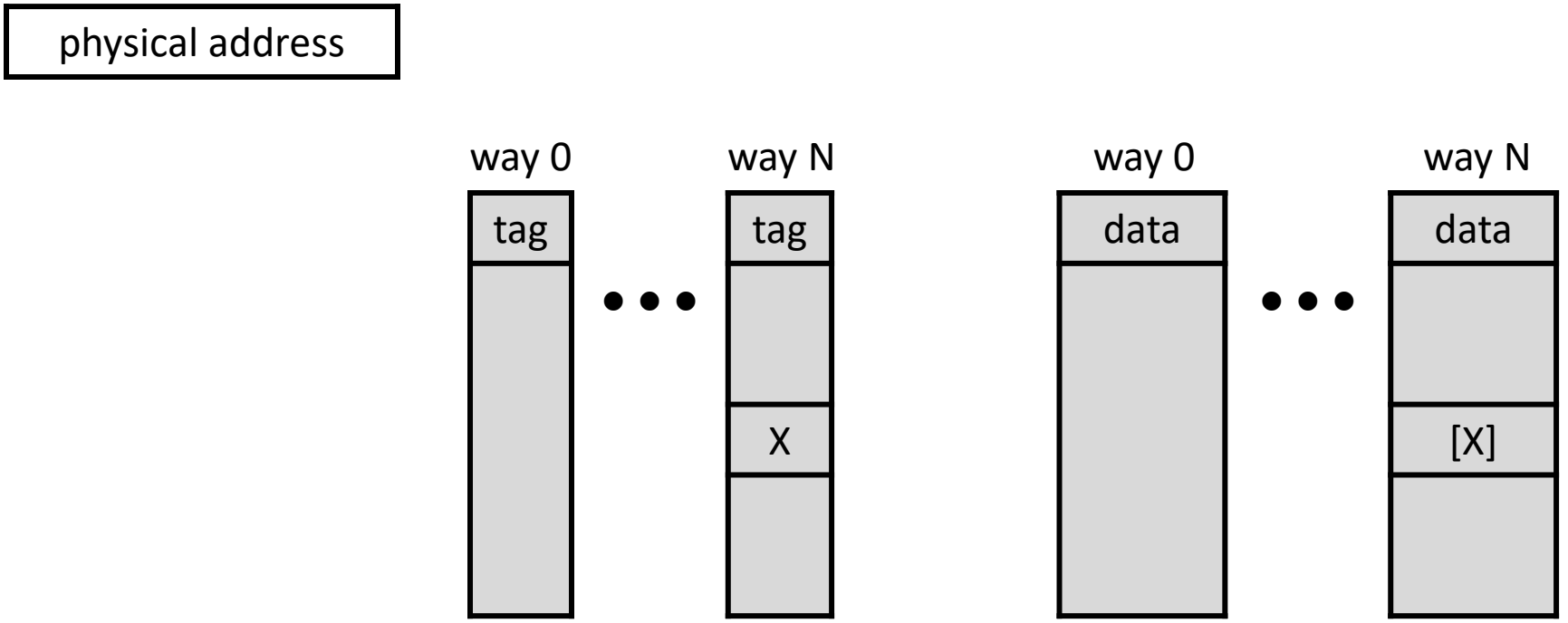
Conventional Cache – Lookup



Conventional Cache – Lookup



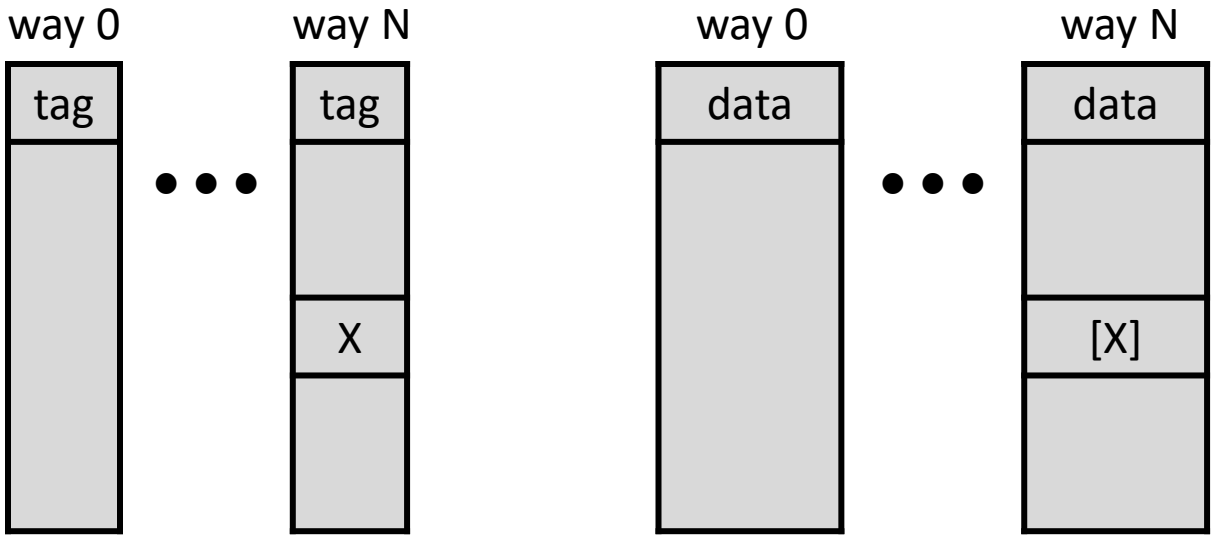
Conventional Cache – Lookup



Conventional Cache – Lookup

address X+STRIDE

physical address

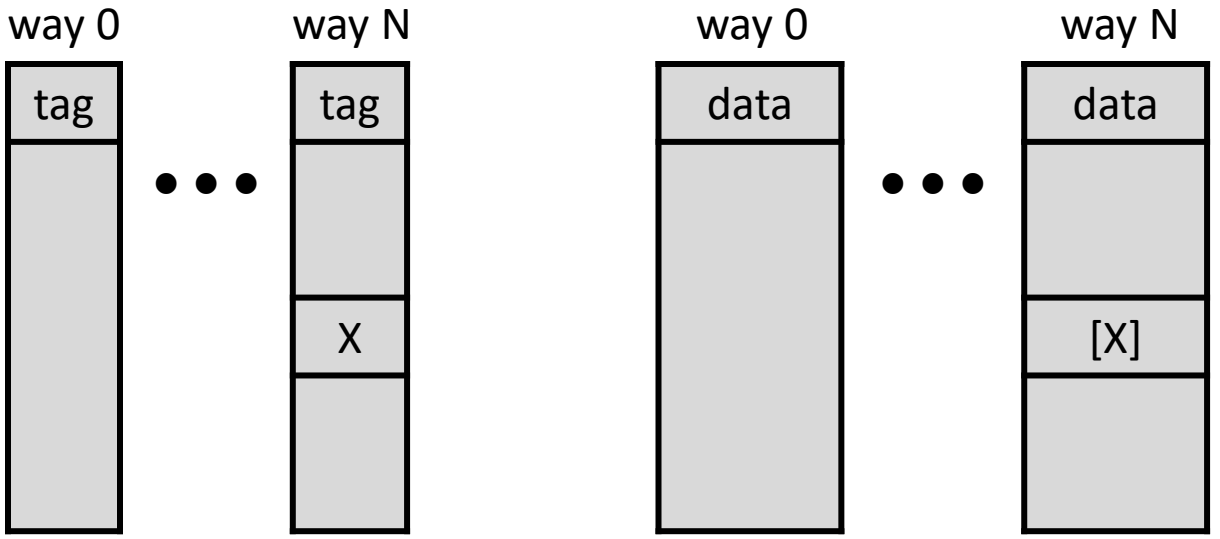


Conventional Cache – Lookup

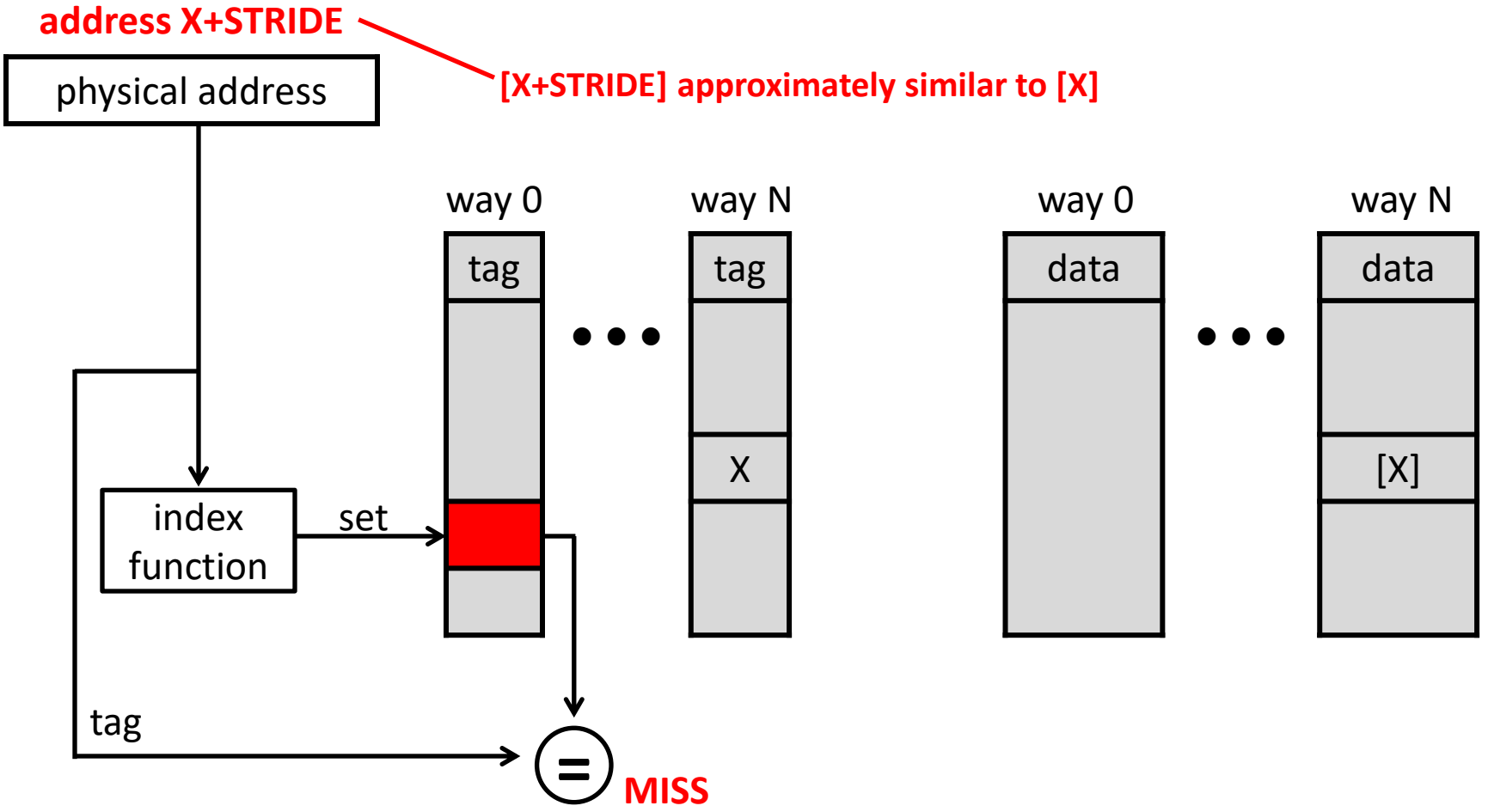
address X+STRIDE

physical address

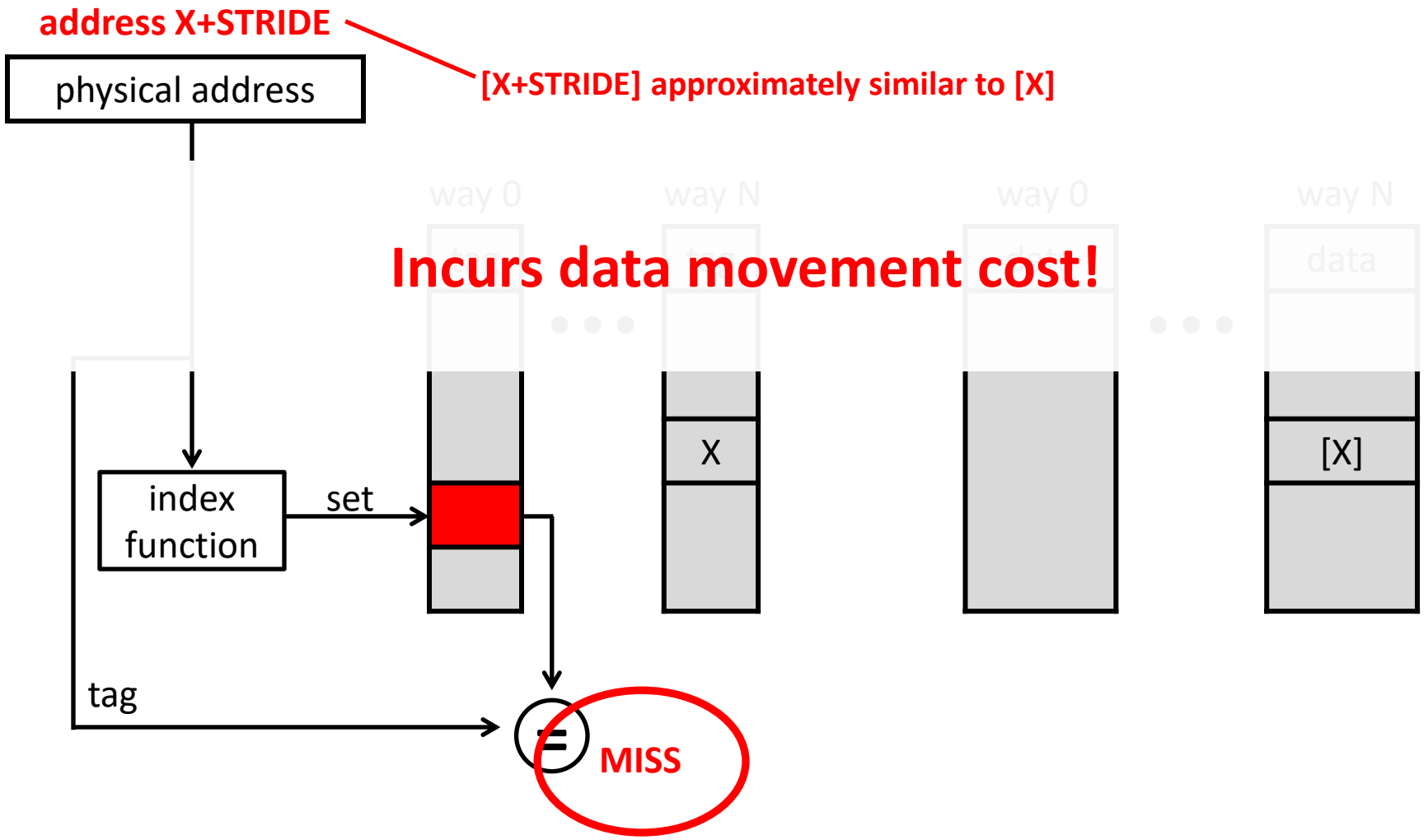
[X+STRIDE] approximately similar to [X]



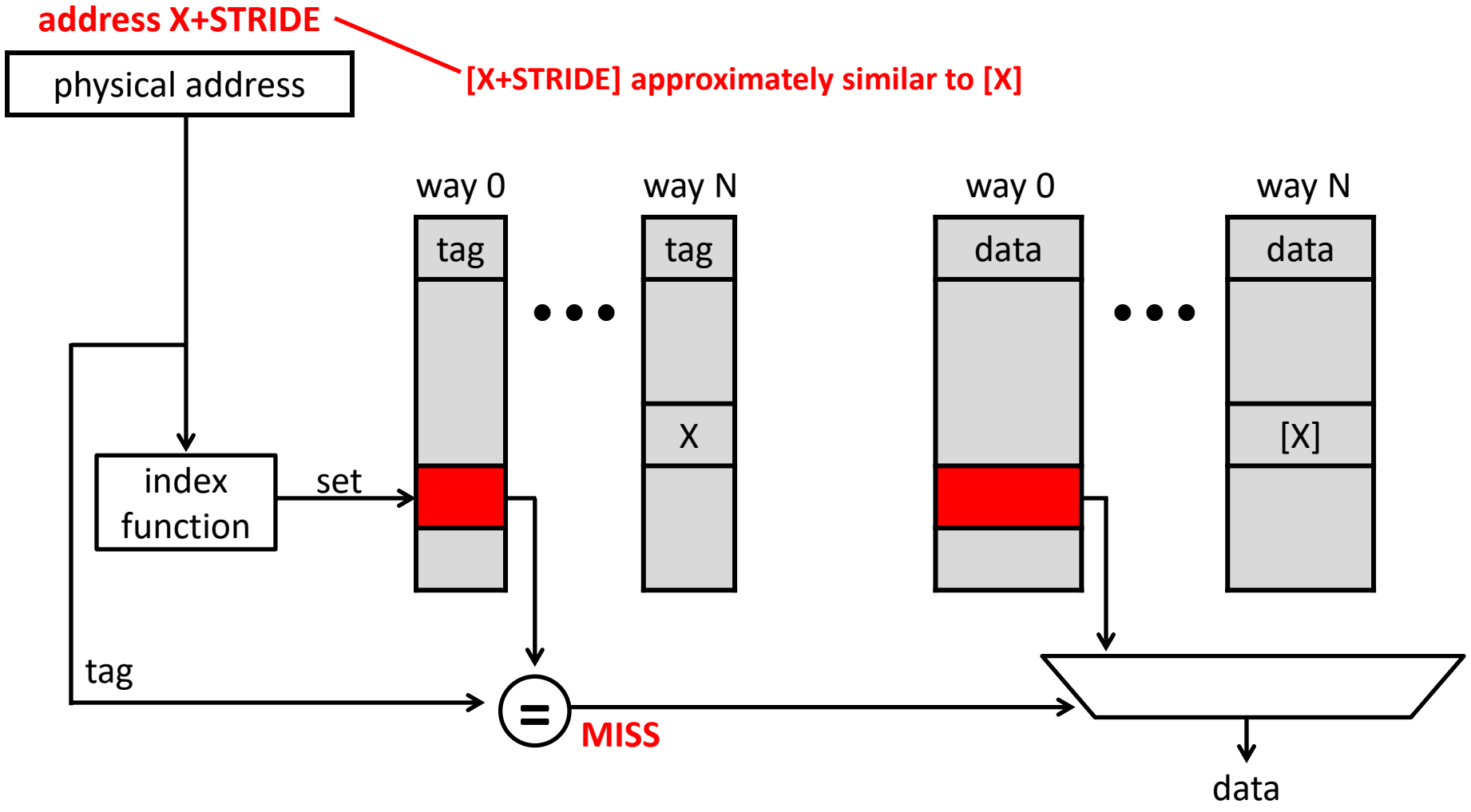
Conventional Cache – Lookup



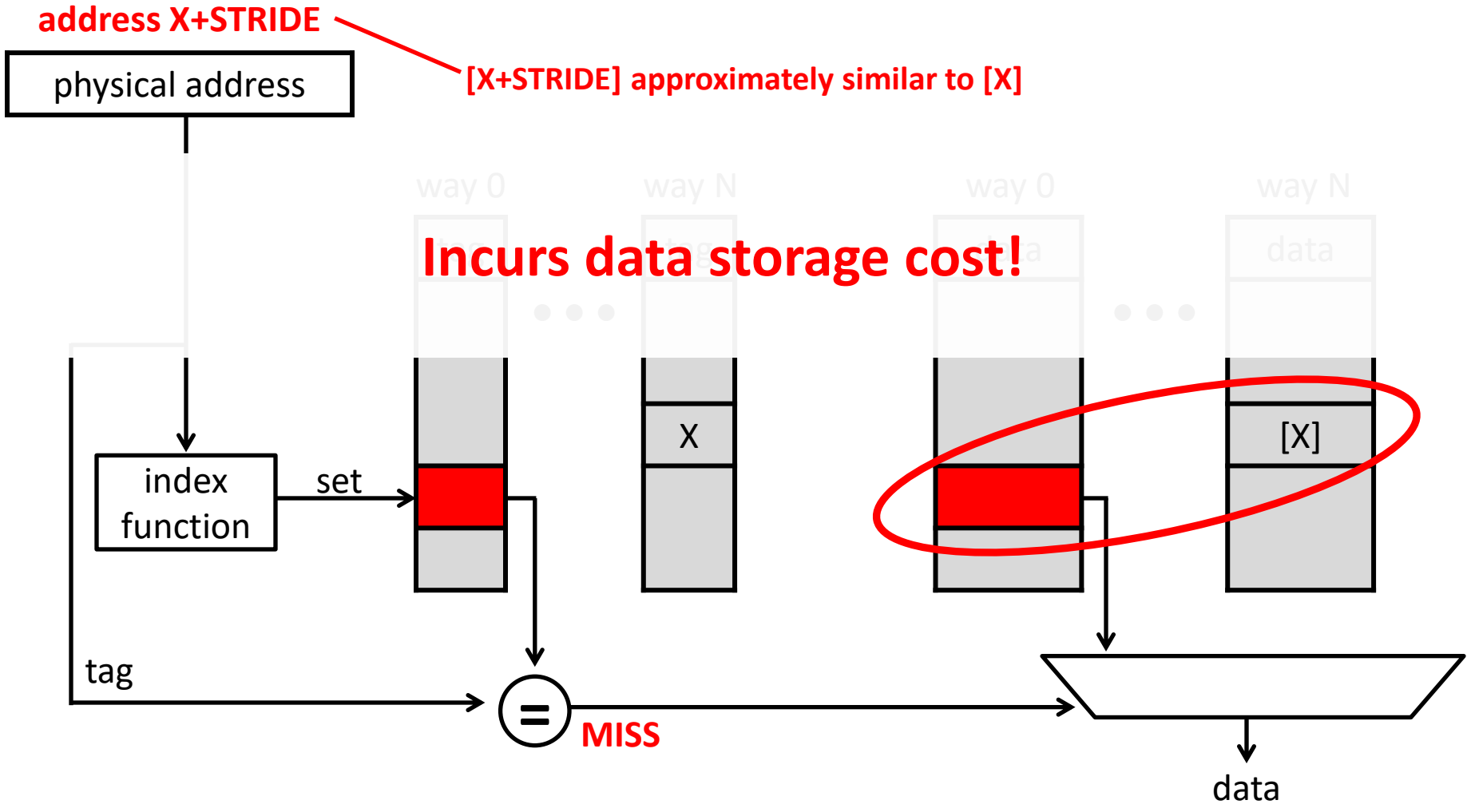
Conventional Cache – Lookup



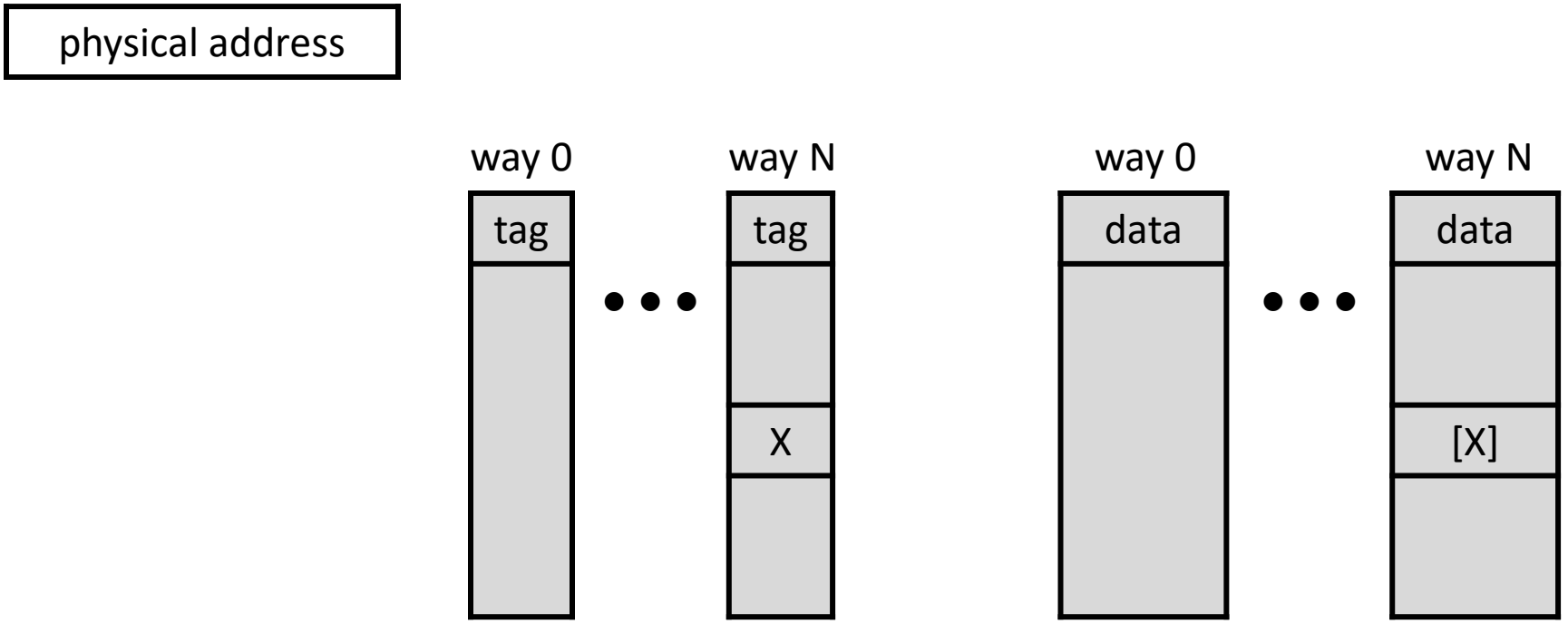
Conventional Cache – Lookup



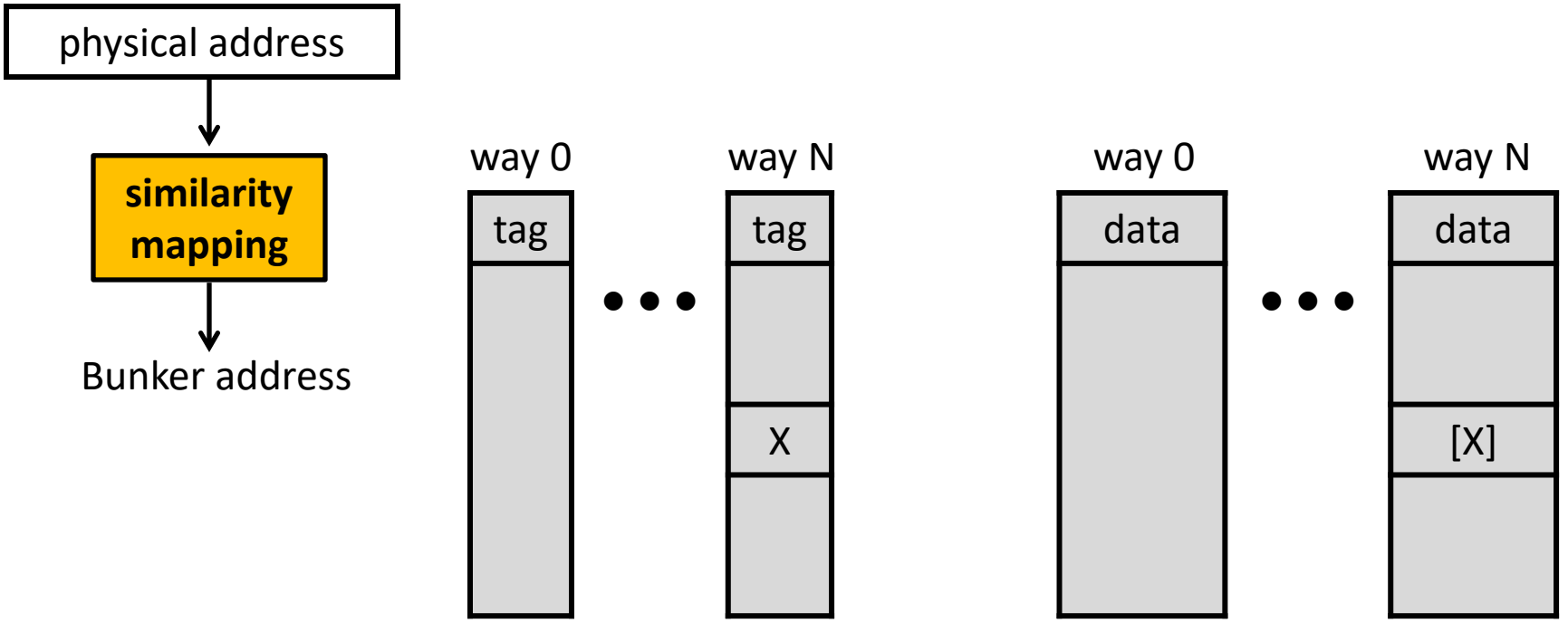
Conventional Cache – Lookup



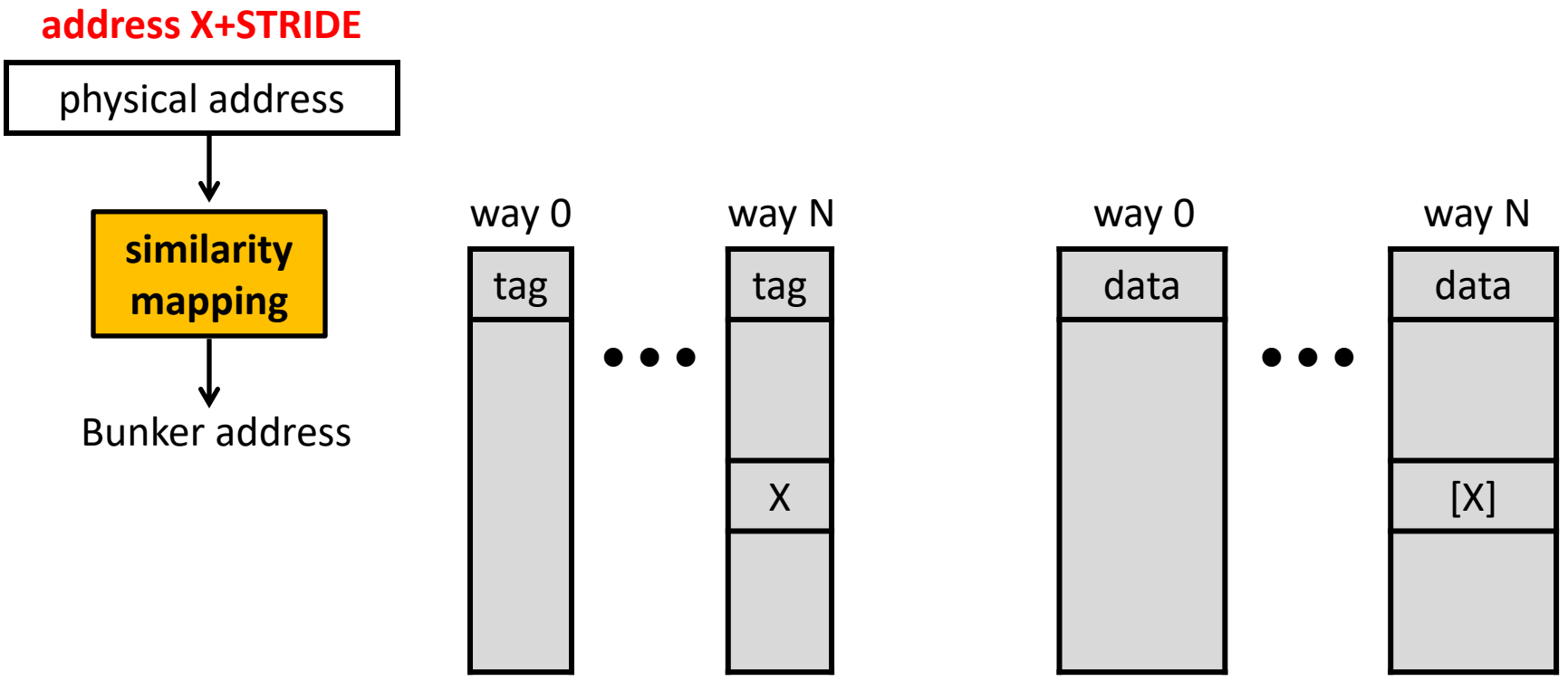
The Bunker Cache



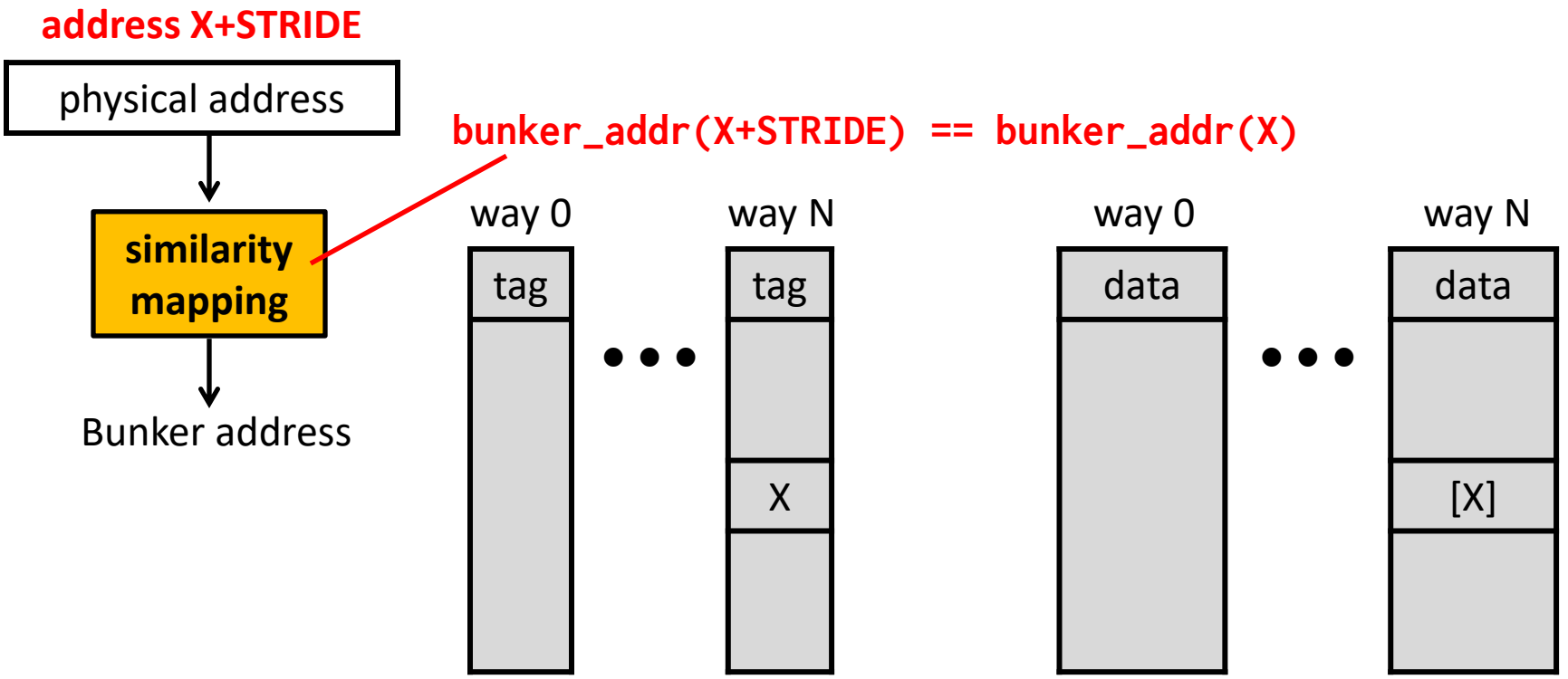
The Bunker Cache



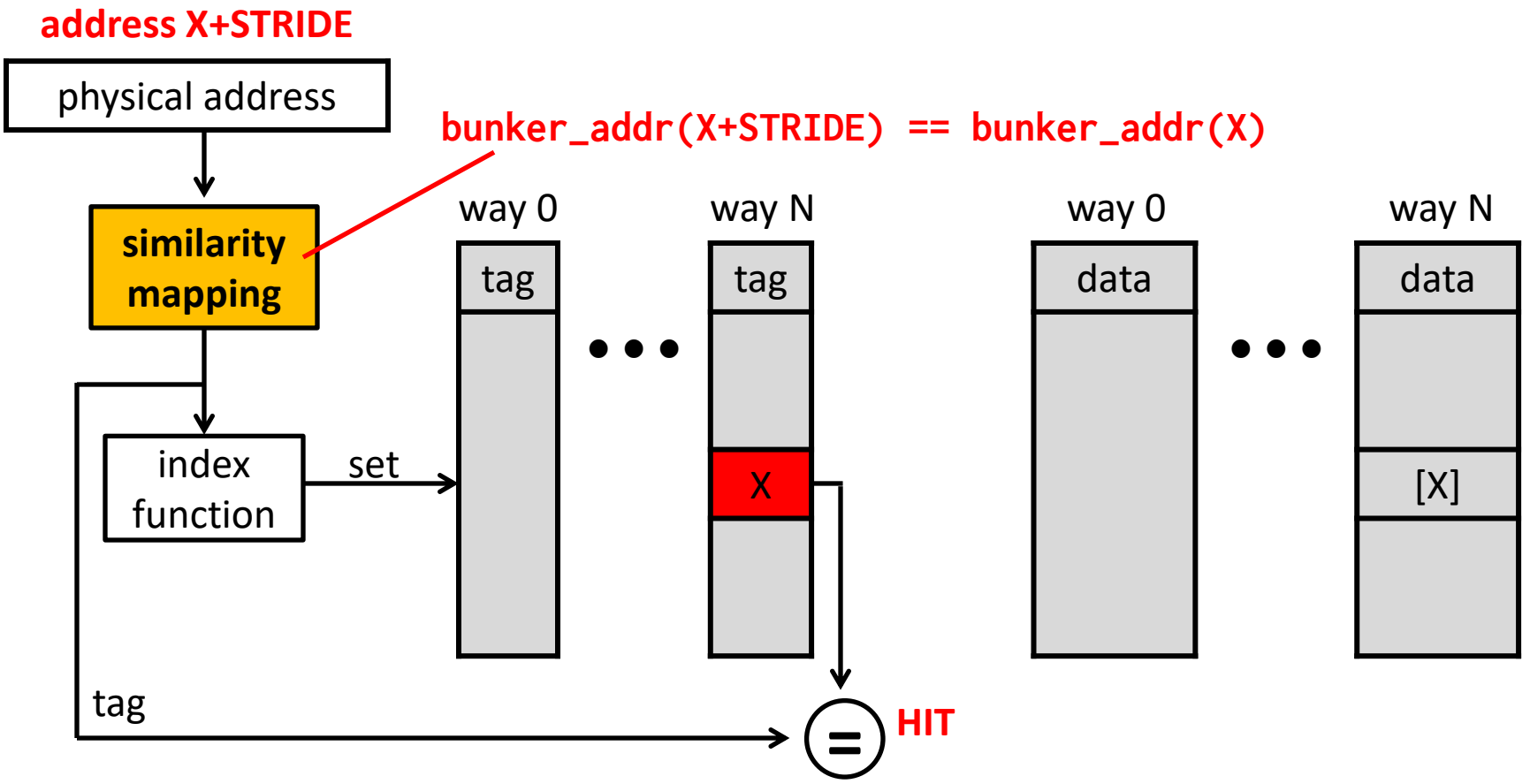
The Bunker Cache – Approximate Lookup



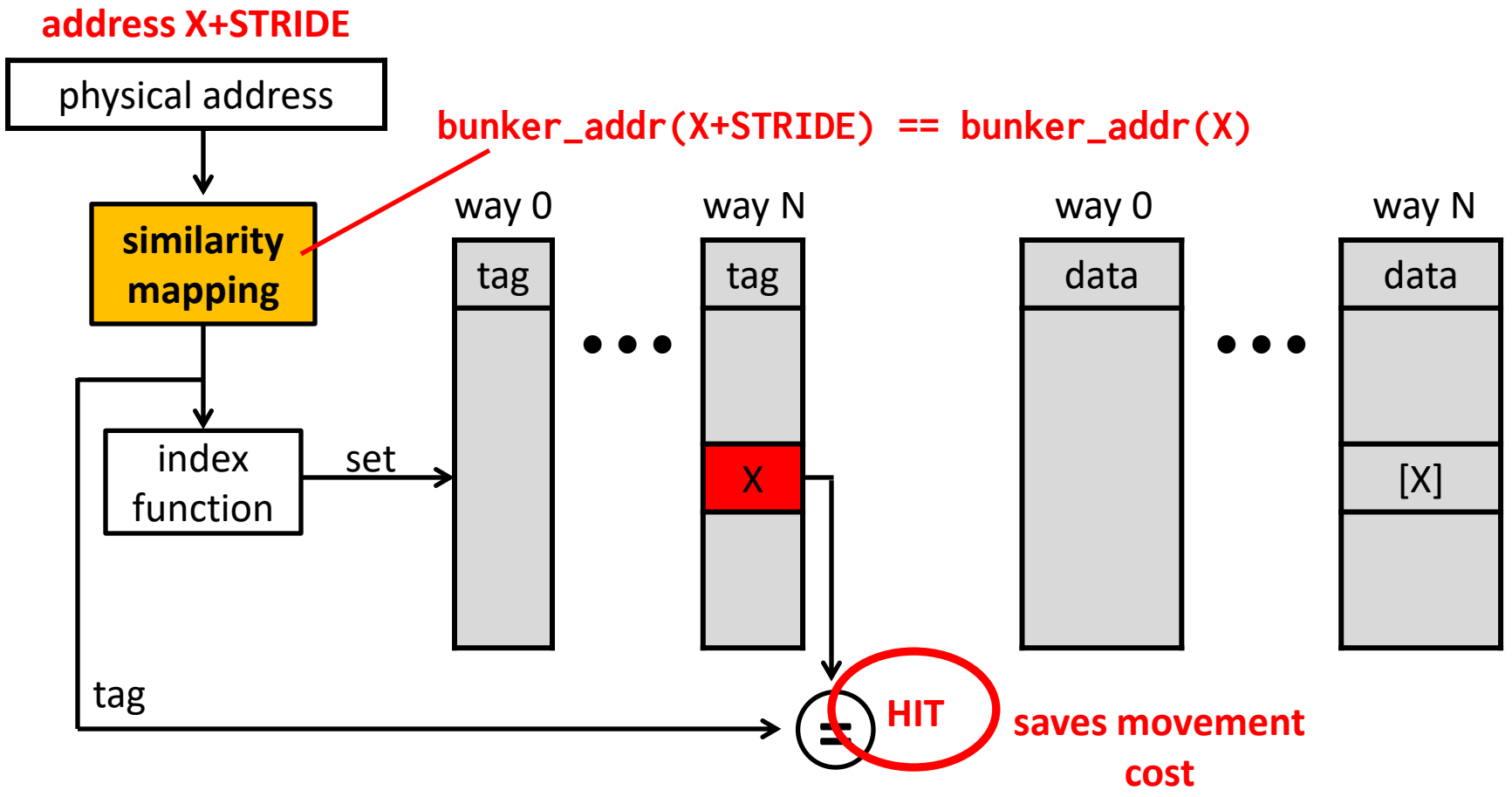
The Bunker Cache – Approximate Lookup



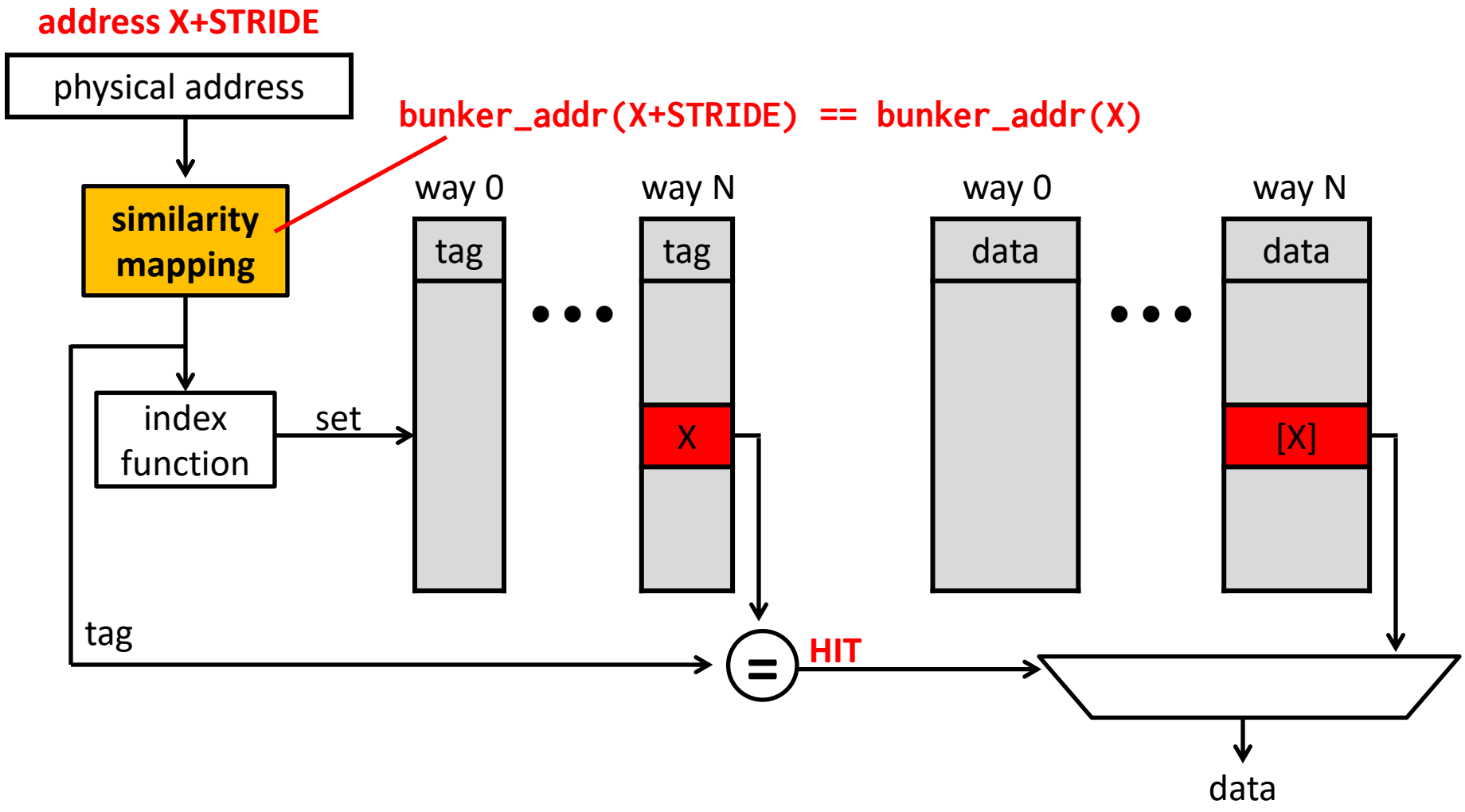
The Bunker Cache – Approximate Lookup



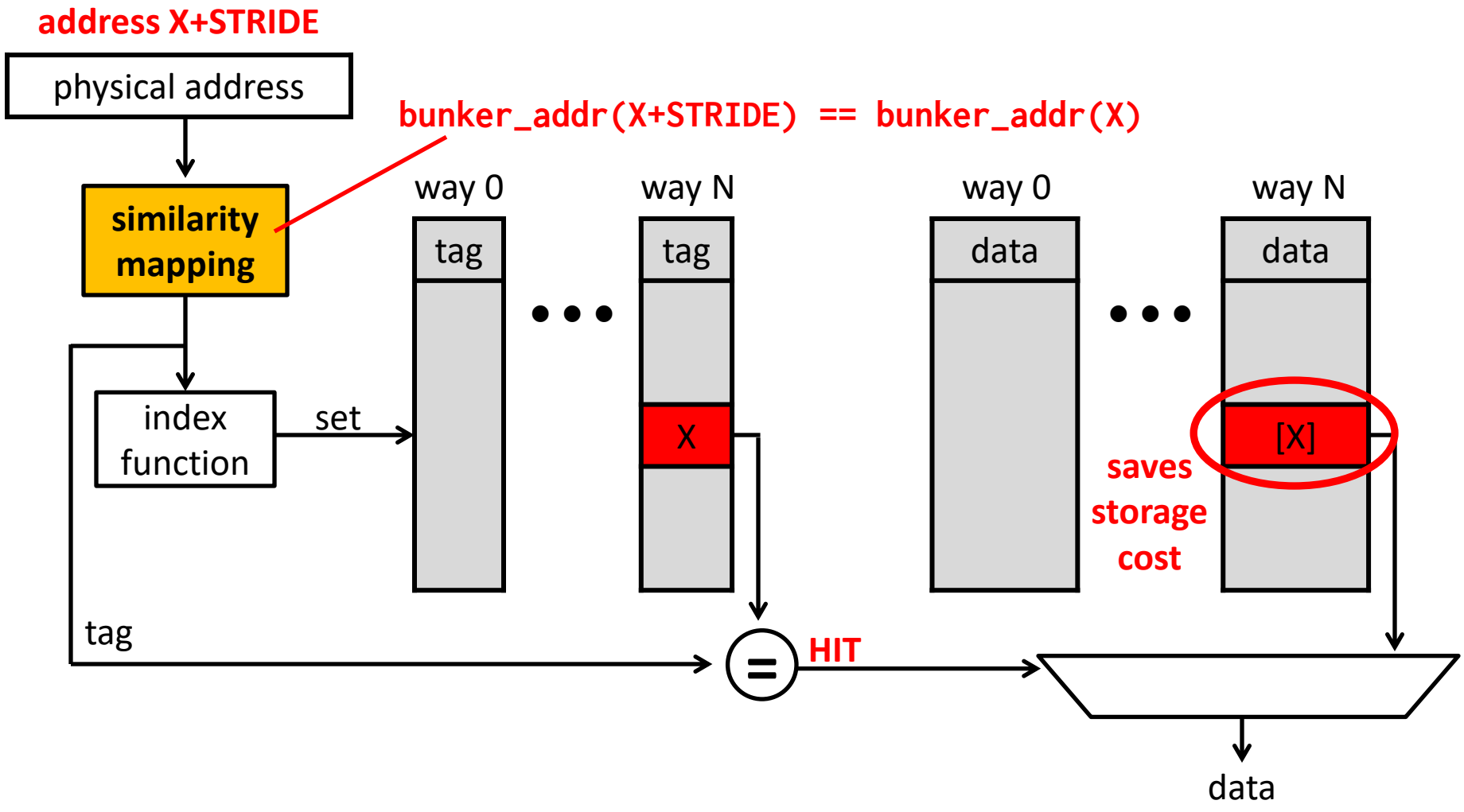
The Bunker Cache – Approximate Lookup



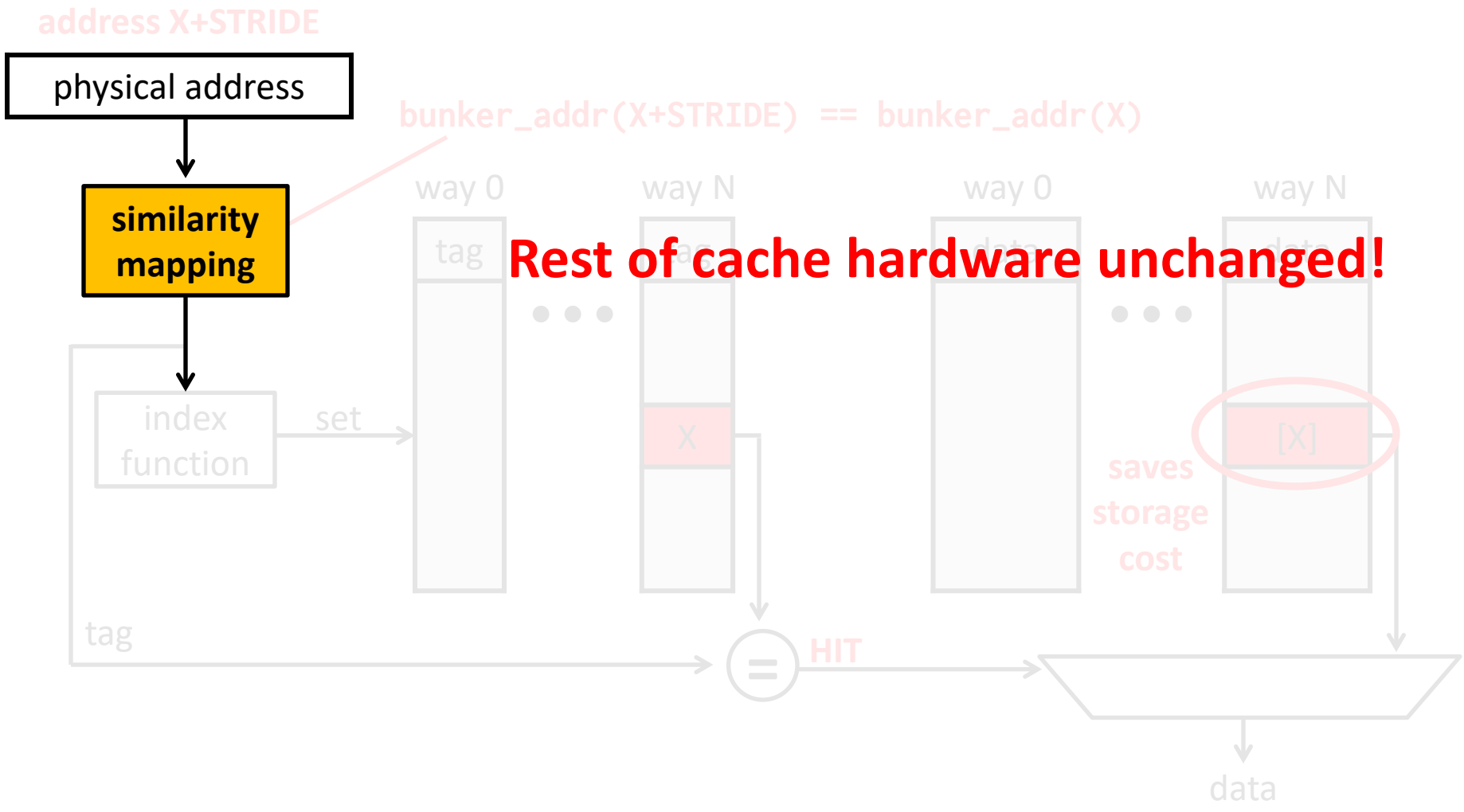
The Bunker Cache – Approximate Lookup



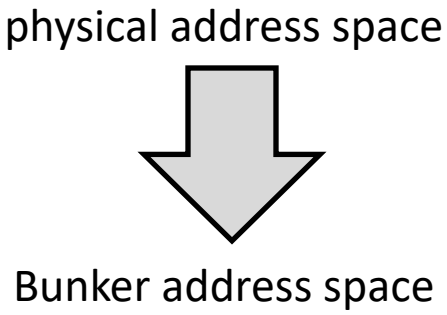
The Bunker Cache – Approximate Lookup



The Bunker Cache – Approximate Lookup



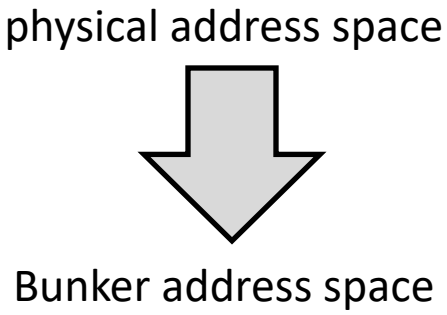
The Bunker Cache – Similarity Mapping



The Bunker Cache – Similarity Mapping

STRIDE: distance between approximately similar blocks

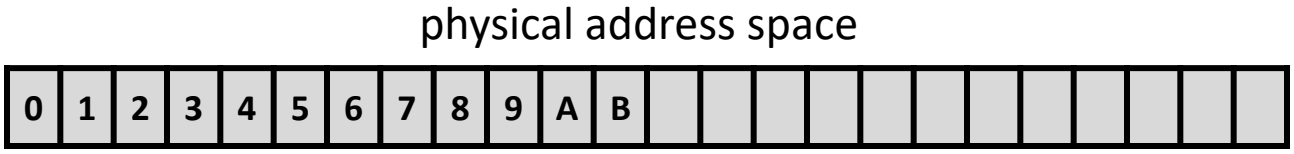
RADIX: degree (i.e., aggressiveness) of approximation



The Bunker Cache – Similarity Mapping

STRIDE: distance between approximately similar blocks

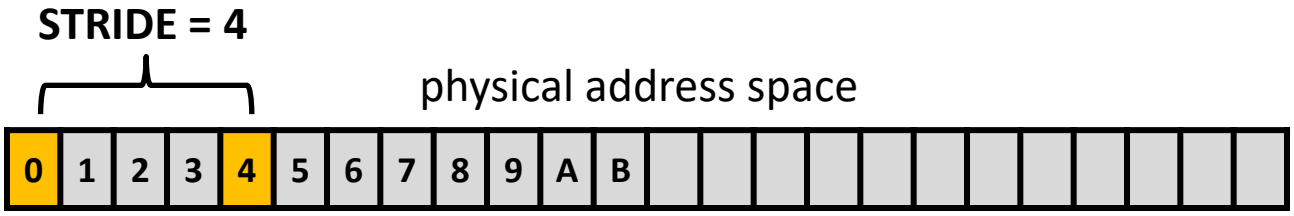
RADIX: degree (i.e., aggressiveness) of approximation



The Bunker Cache – Similarity Mapping

STRIDE: distance between approximately similar blocks

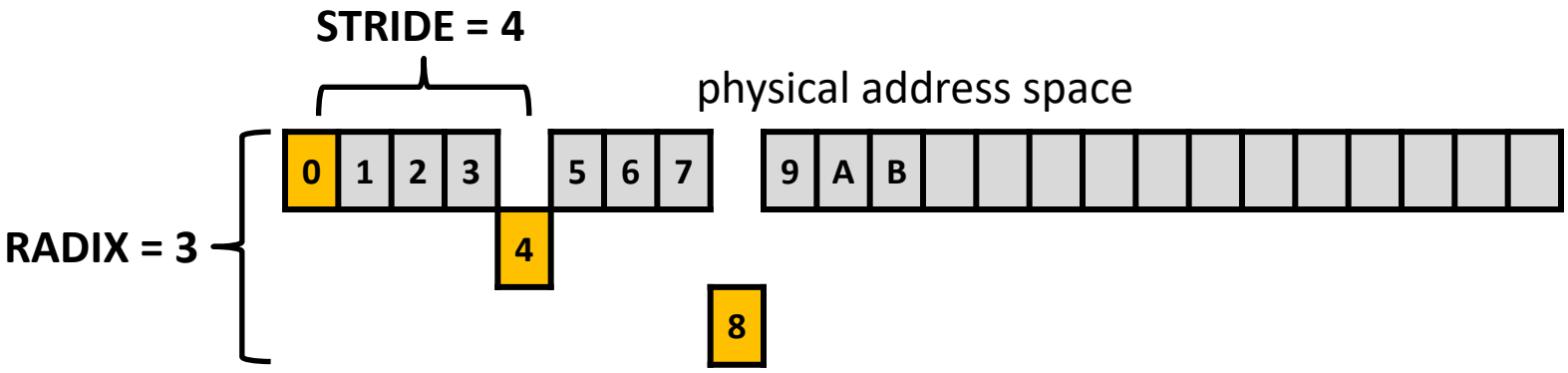
RADIX: degree (i.e., aggressiveness) of approximation



The Bunker Cache – Similarity Mapping

STRIDE: distance between approximately similar blocks

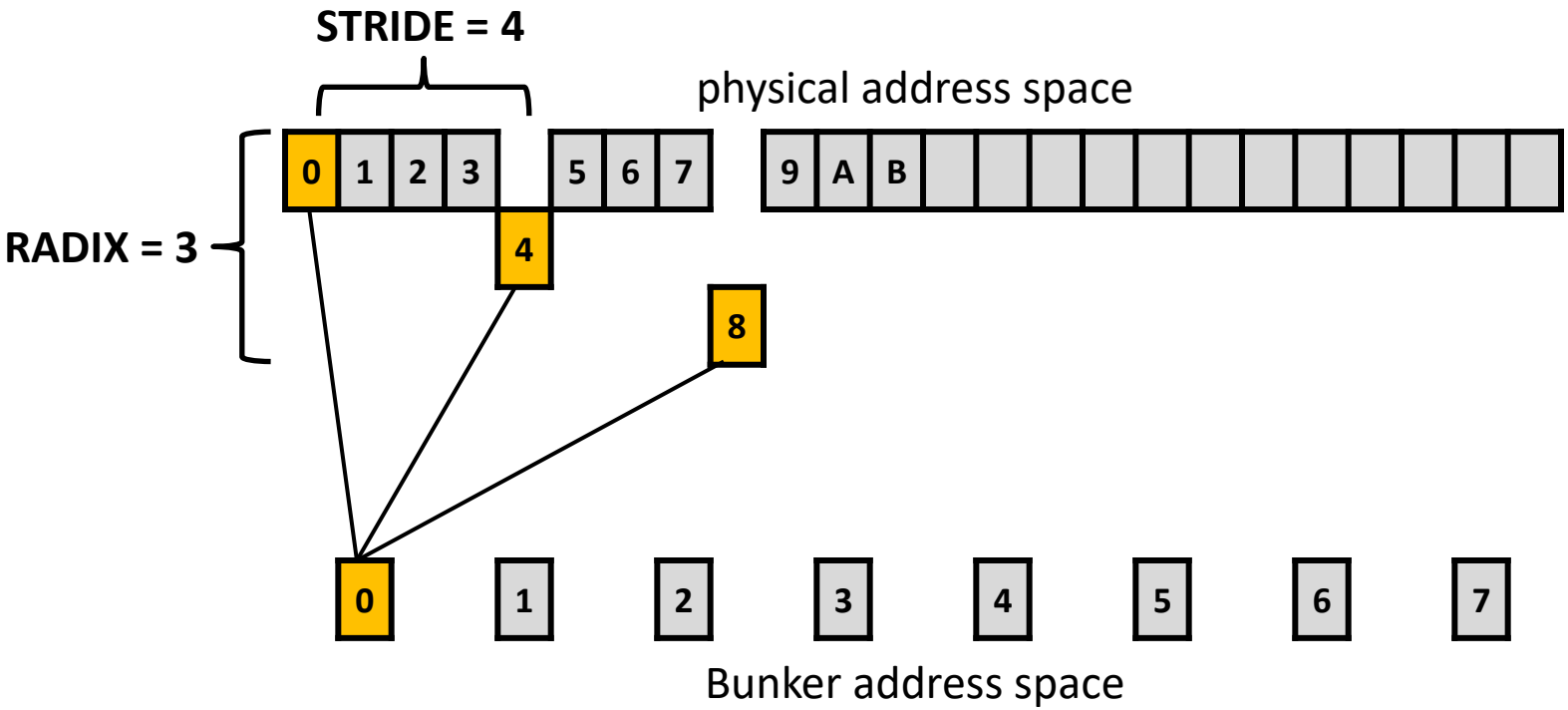
RADIX: degree (i.e., aggressiveness) of approximation



The Bunker Cache – Similarity Mapping

STRIDE: distance between approximately similar blocks

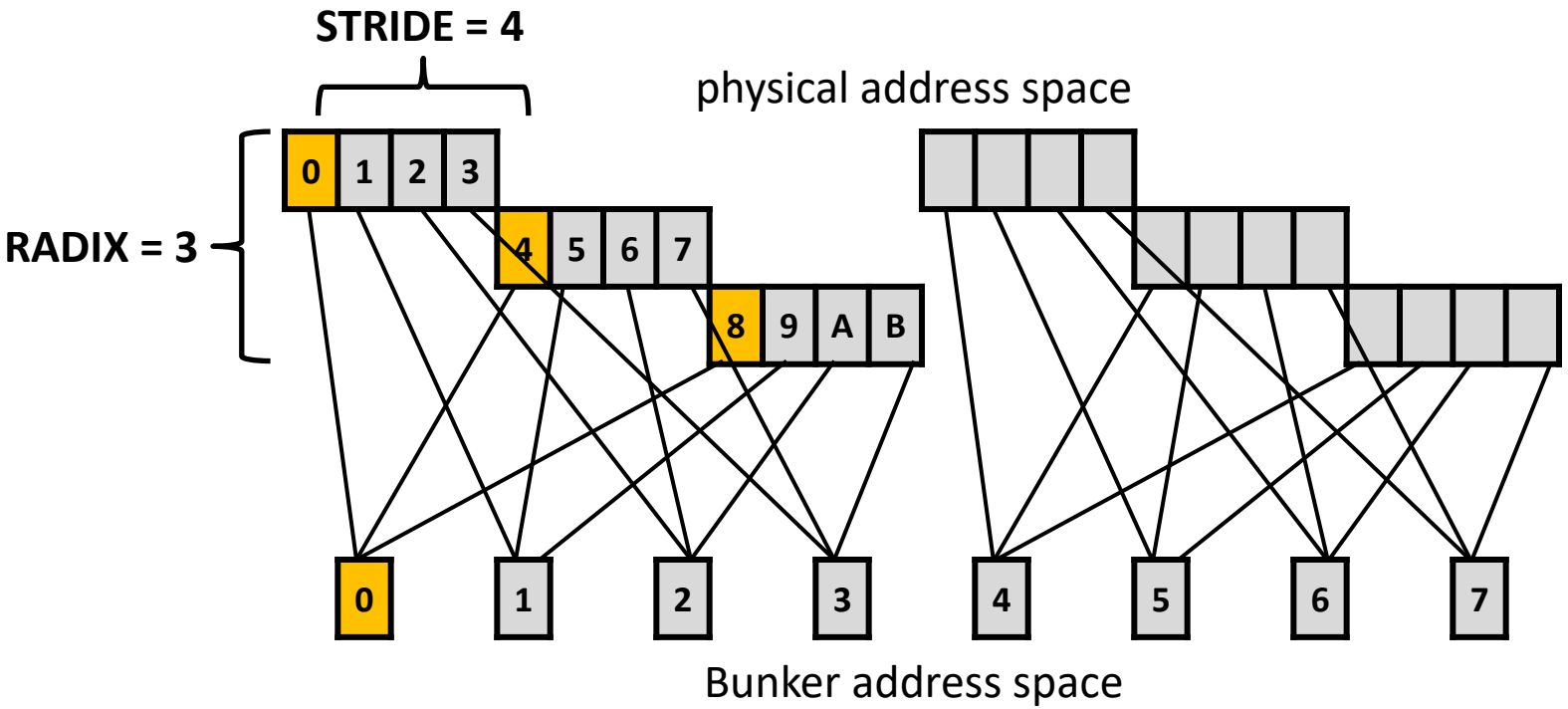
RADIX: degree (i.e., aggressiveness) of approximation



The Bunker Cache – Similarity Mapping

STRIDE: distance between approximately similar blocks

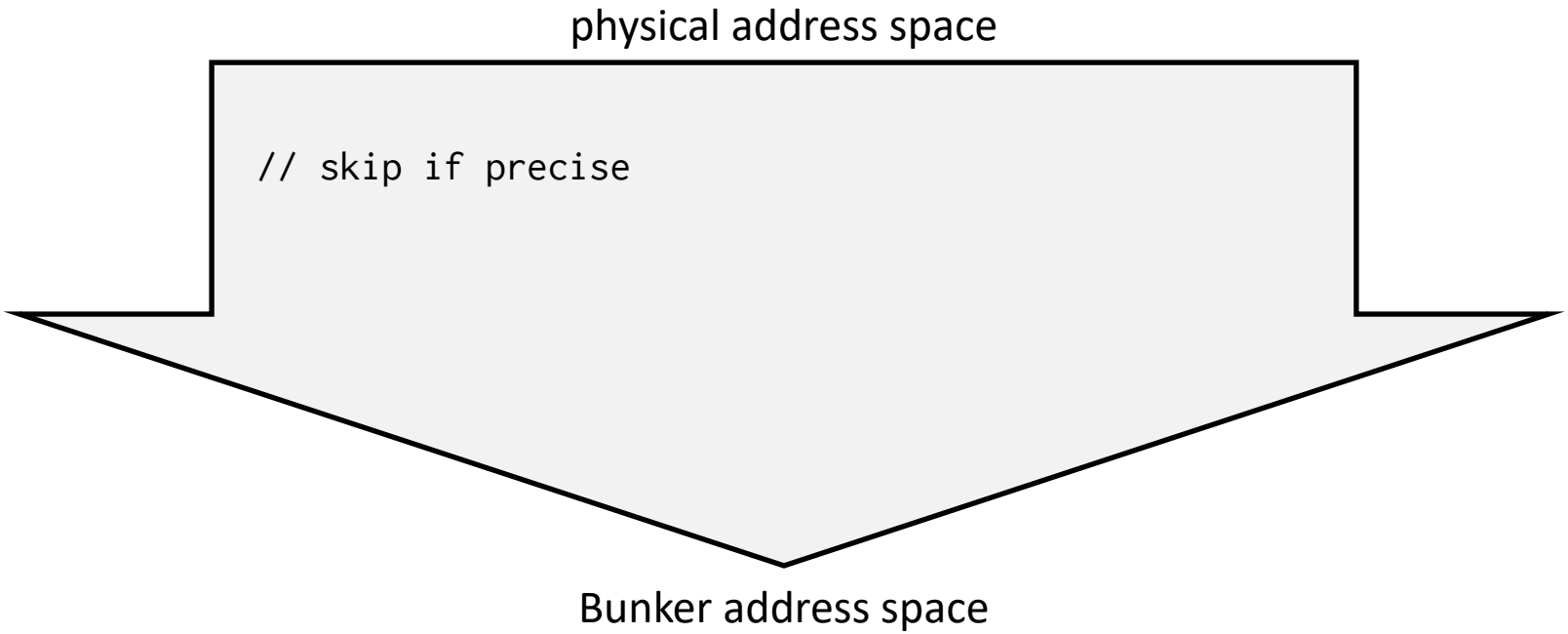
RADIX: degree (i.e., aggressiveness) of approximation



The Bunker Cache – Similarity Mapping

STRIDE: distance between approximately similar blocks

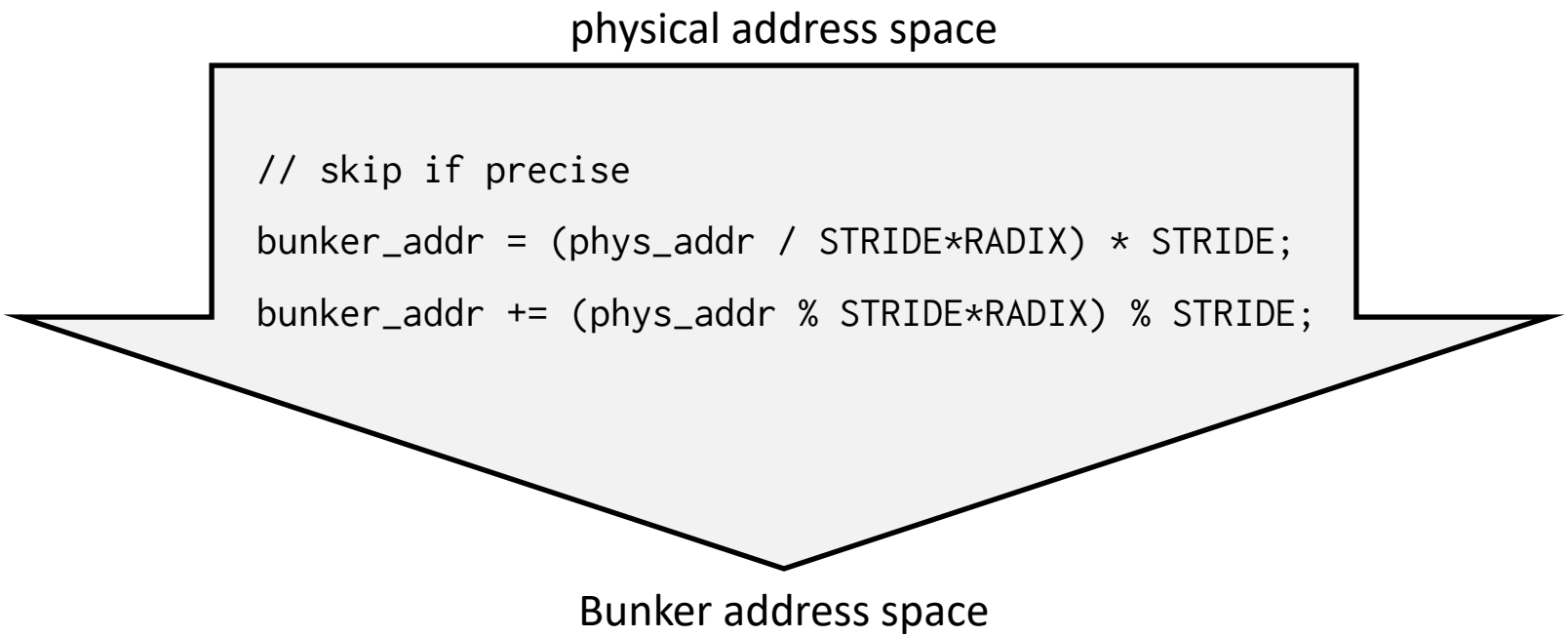
RADIX: degree (i.e., aggressiveness) of approximation



The Bunker Cache – Similarity Mapping

STRIDE: distance between approximately similar blocks

RADIX: degree (i.e., aggressiveness) of approximation



The Bunker Cache – Additional Details

Coherence and dirty state:

- requires a separate directory structure that bypasses similarity mapping

The Bunker Cache – Additional Details

Coherence and dirty state:

- requires a separate directory structure that bypasses similarity mapping

Drowsy blocks:

- many-to-one mapping offers more opportunity for low-leakage storage

The Bunker Cache – Additional Details

Coherence and dirty state:

- requires a separate directory structure that bypasses similarity mapping

Drowsy blocks:

- many-to-one mapping offers more opportunity for low-leakage storage

Dynamic quality control:

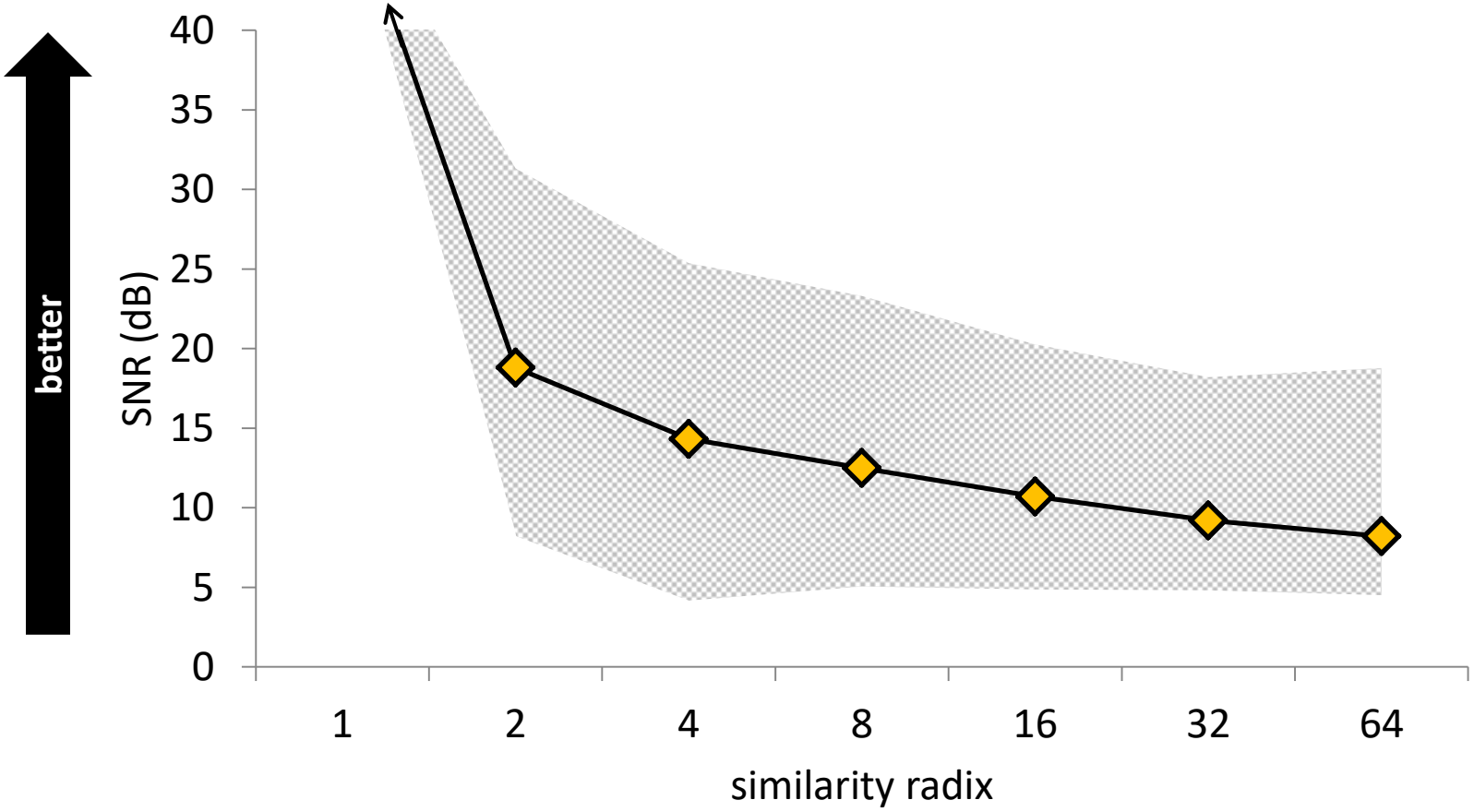
- can tune RADIX and STRIDE on-the-fly via periodic quality checks

More details in paper

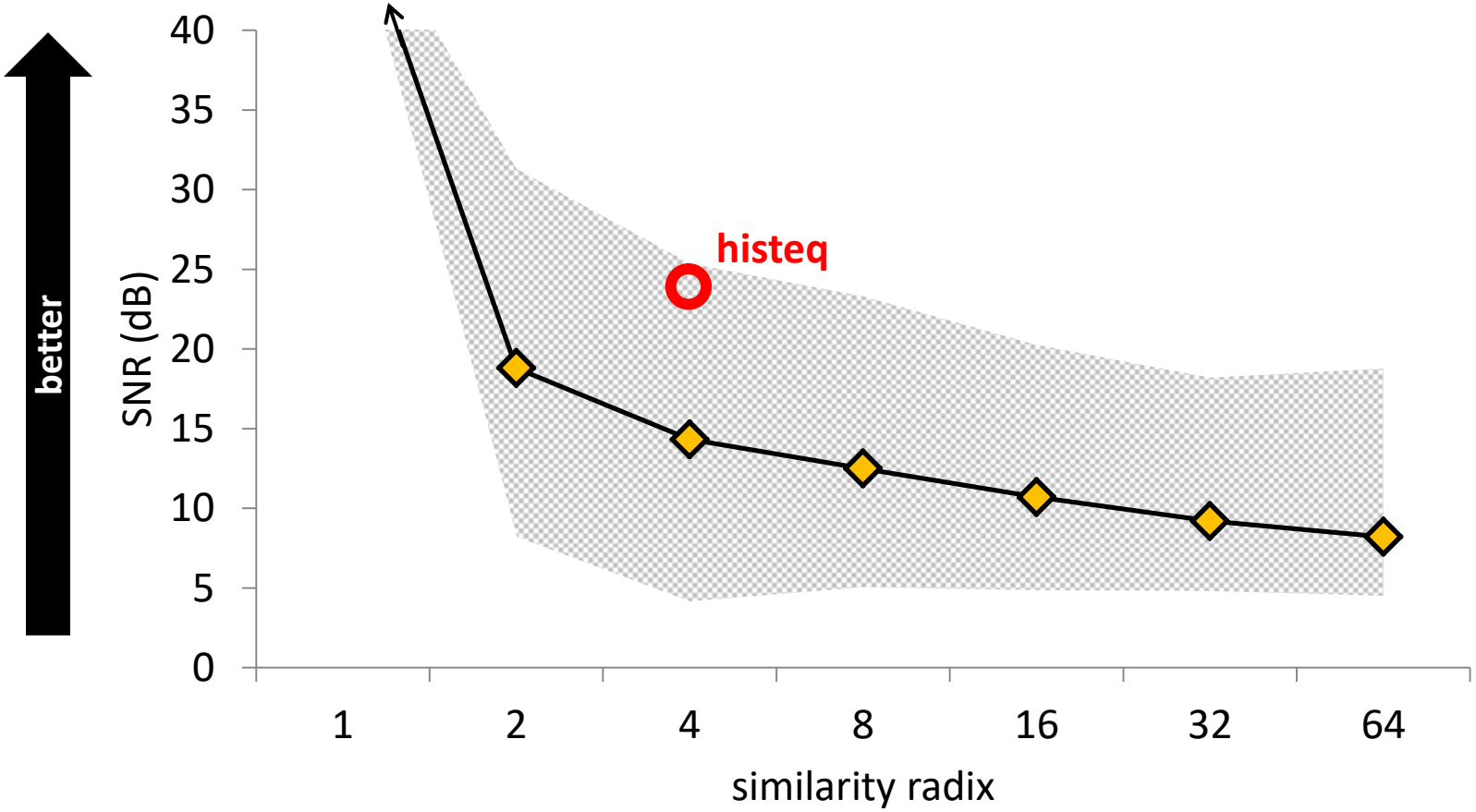
Evaluation

- **Applications**: PERFECT and AxBench
- **Performance**: Full-system cycle-level simulation
- **Energy and Power**: CACTI
- **Quality**: Pin simulation, signal-to-noise-ratio (SNR)
- **Configuration**:
  - 4-core CMP, 16KB private L1, 128KB private L2
  - 2MB shared LLC, 2K-entry directory
  - STRIDE selected based on application's data set dimensions
  - RADIX varied in results

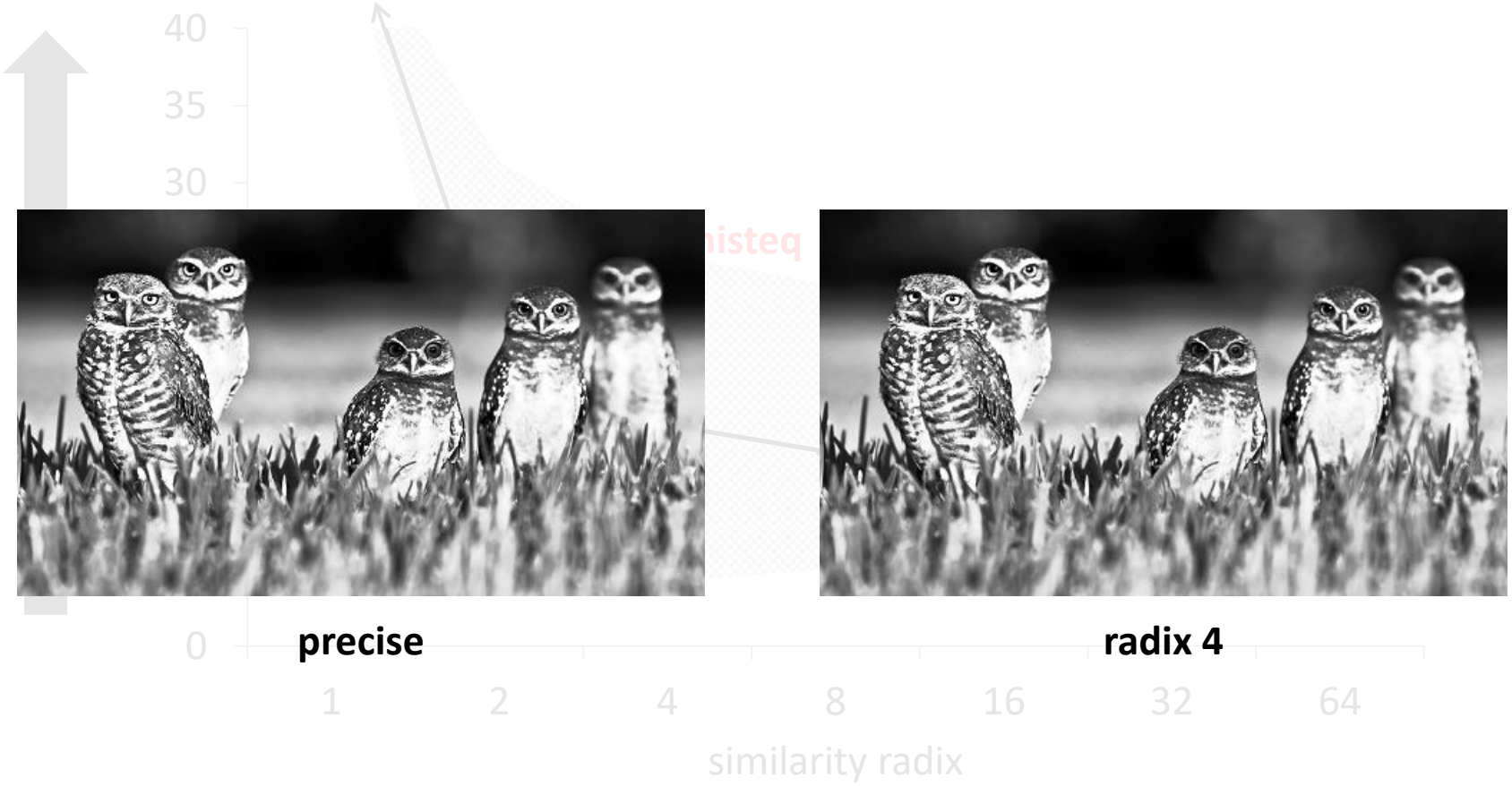
Evaluation – Application Output Quality



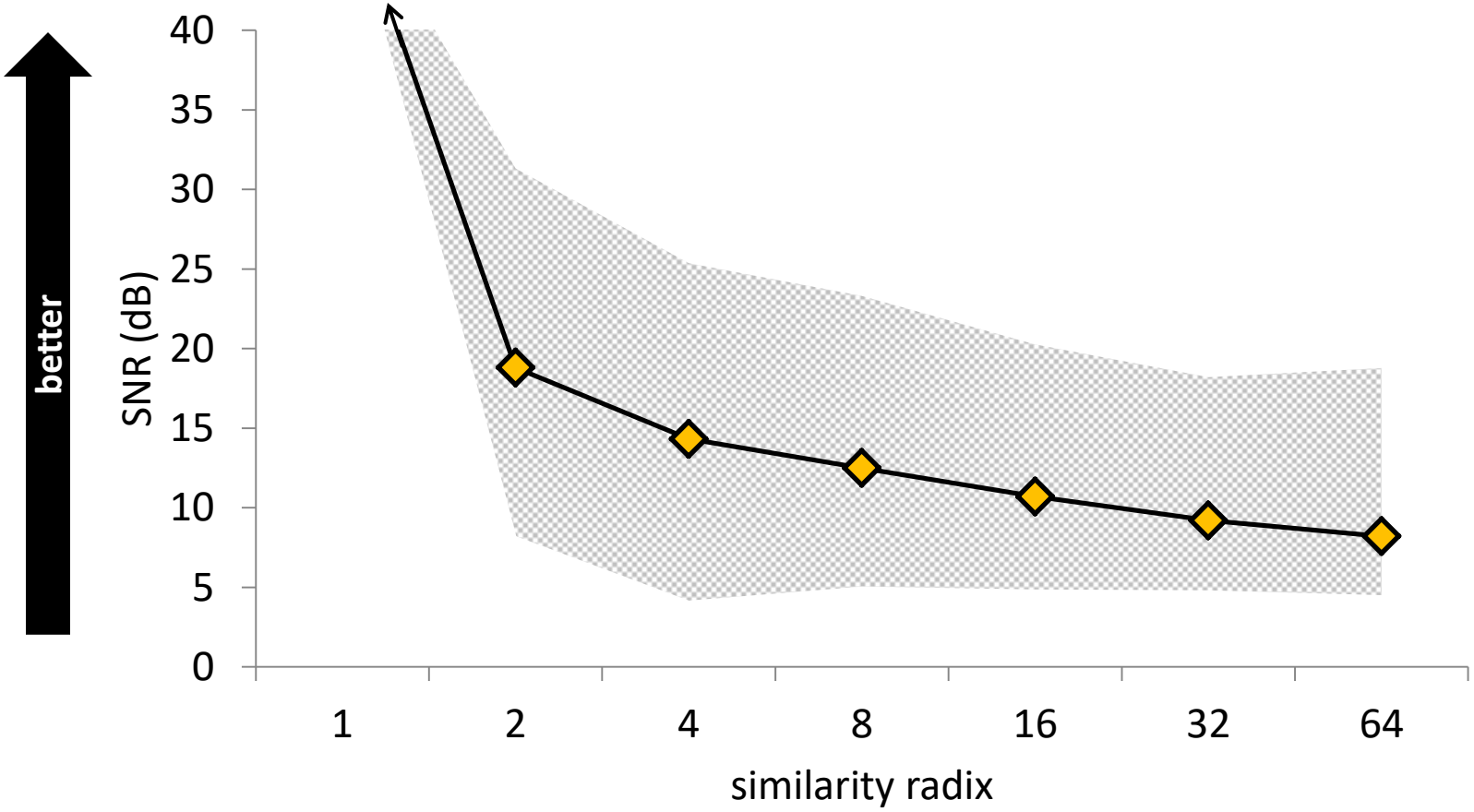
Evaluation – Application Output Quality



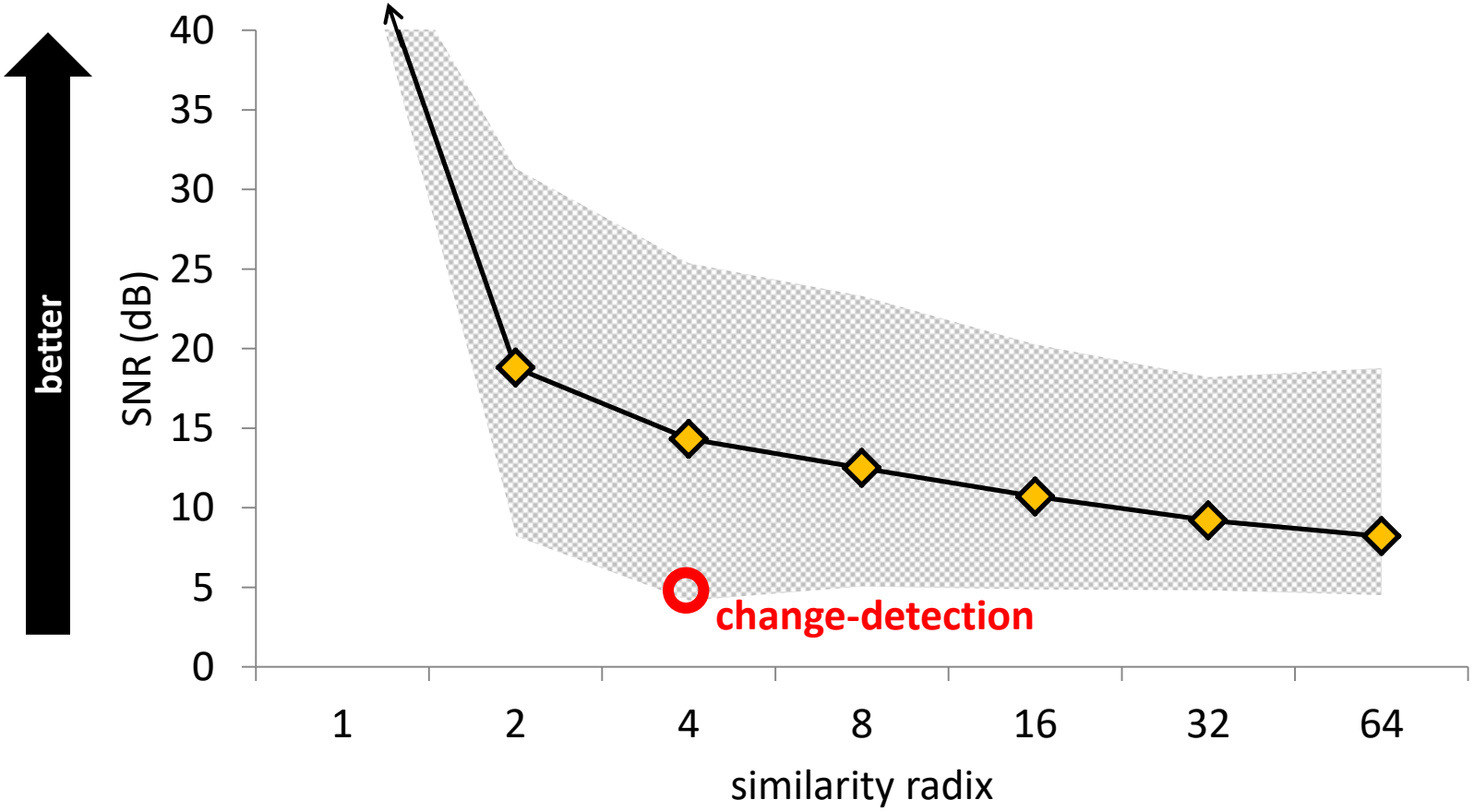
Evaluation – Application Output Quality



Evaluation – Application Output Quality

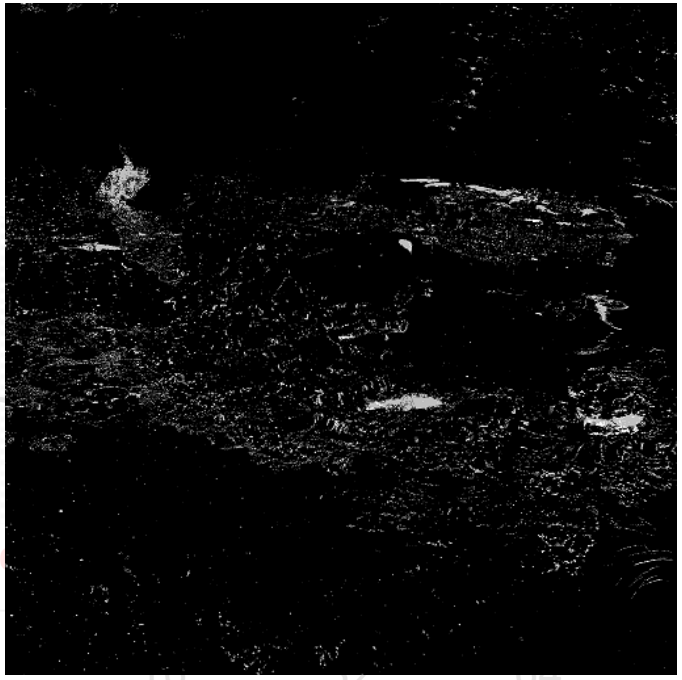


Evaluation – Application Output Quality



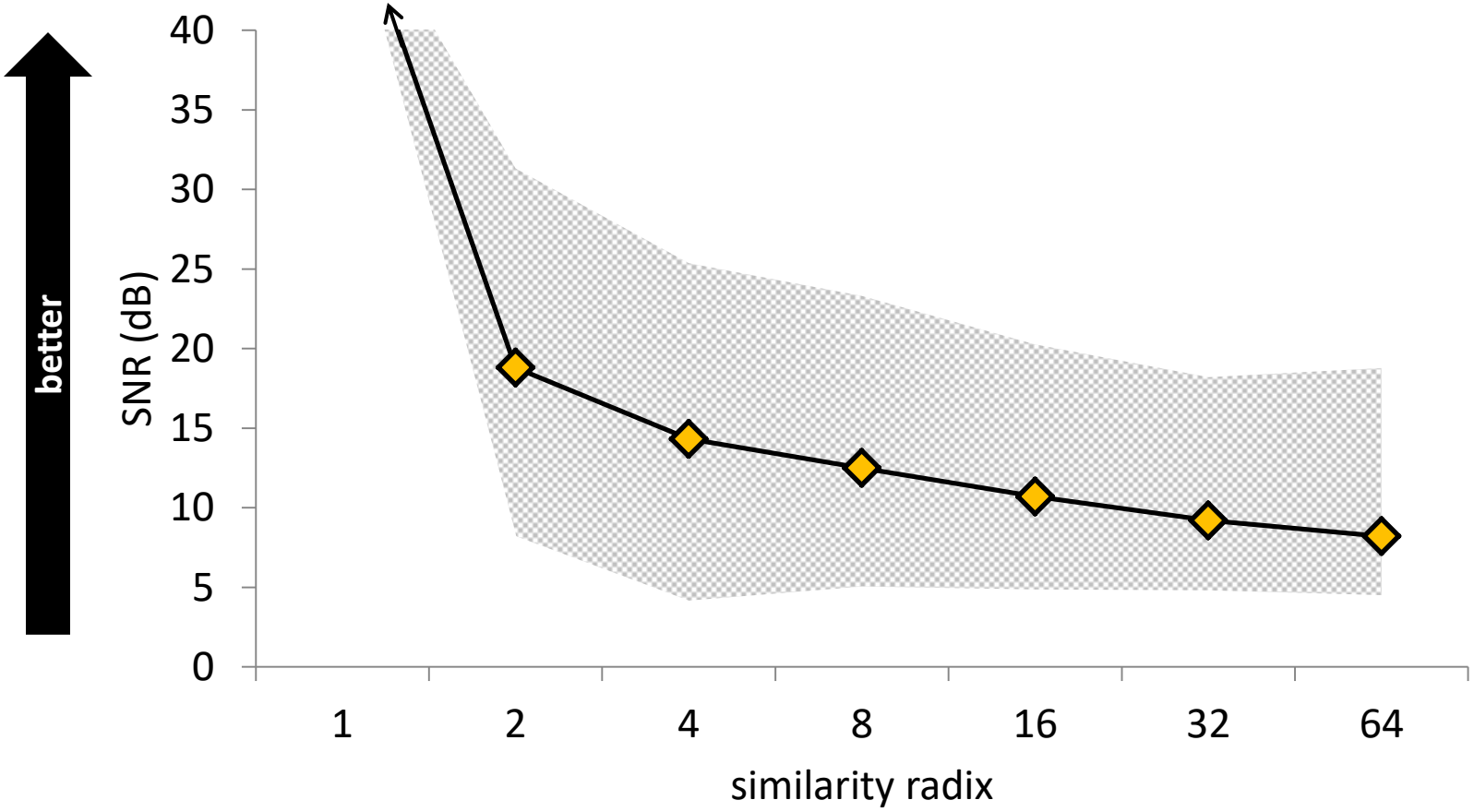
Evaluation – Application Output Quality

precise

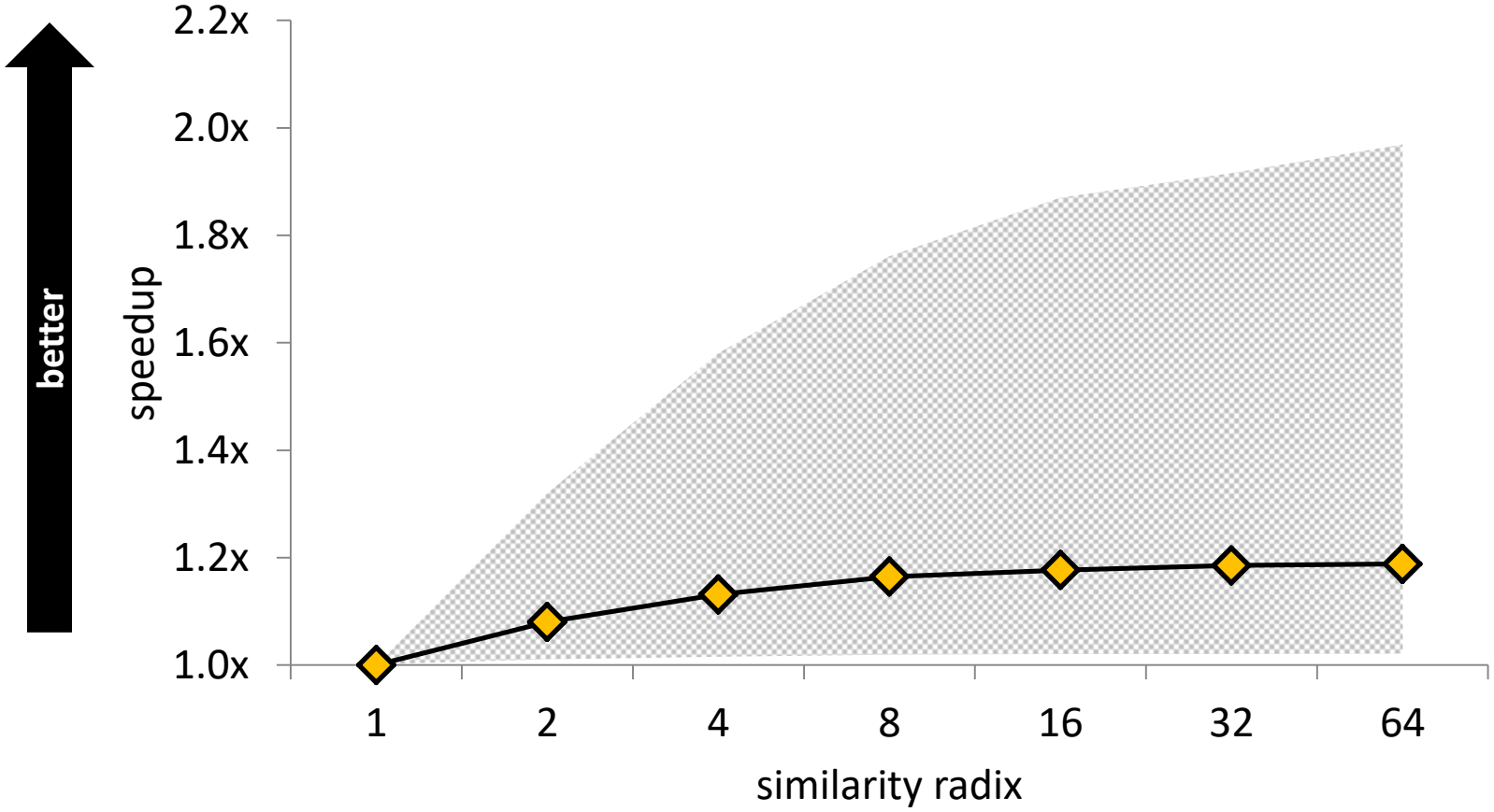


radix 4

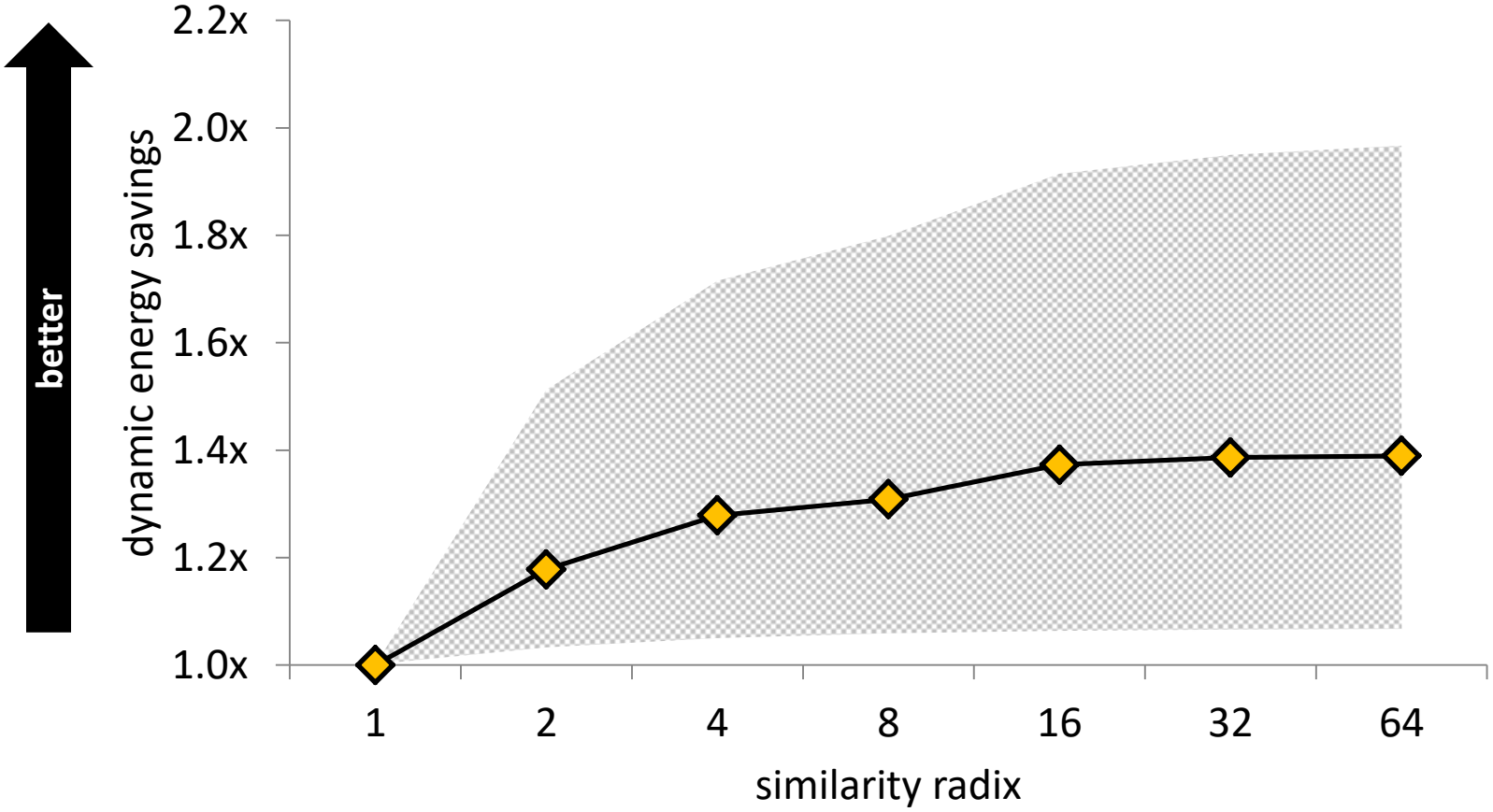
Evaluation – Application Output Quality



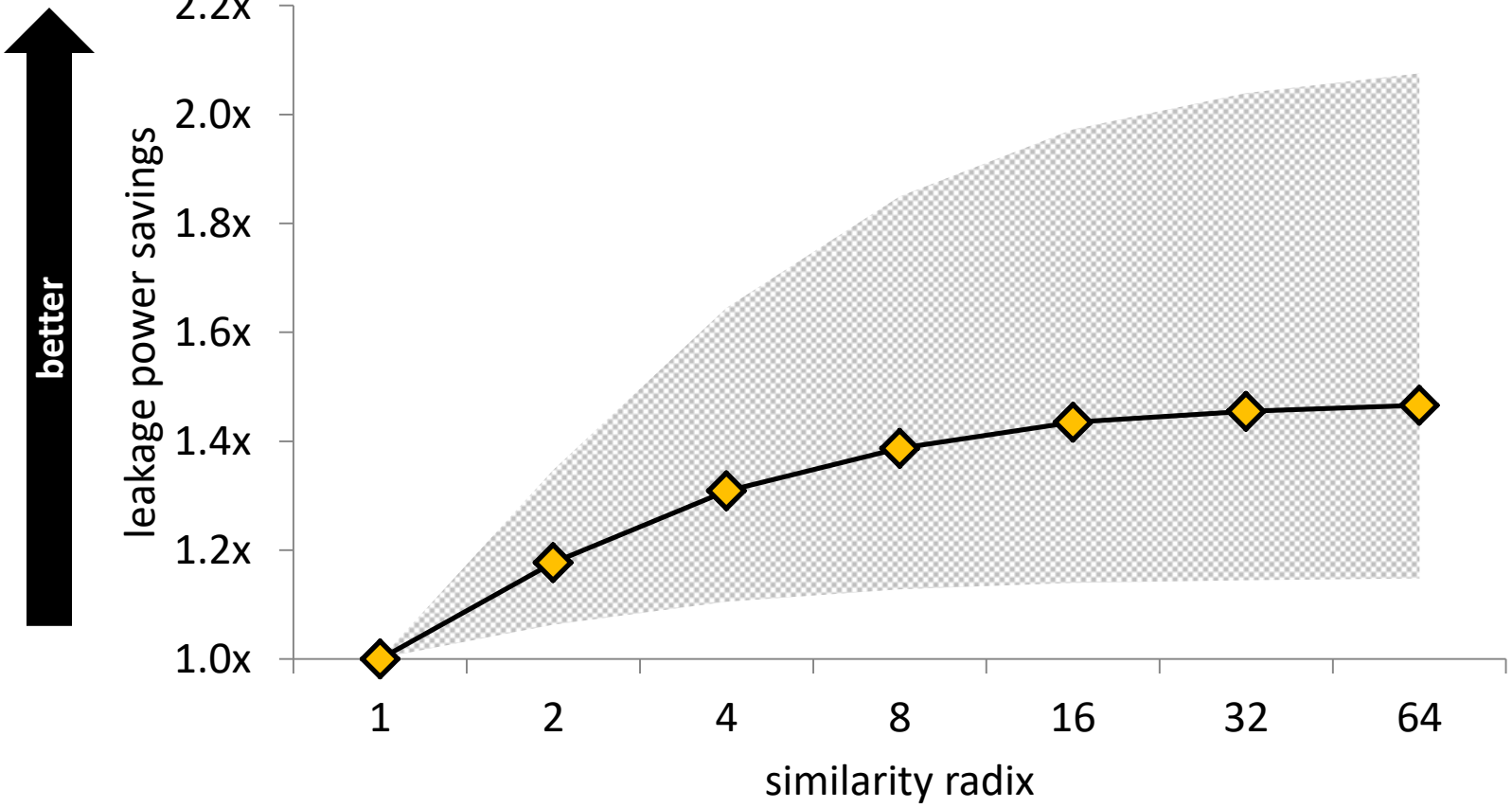
Evaluation – Application Speedup



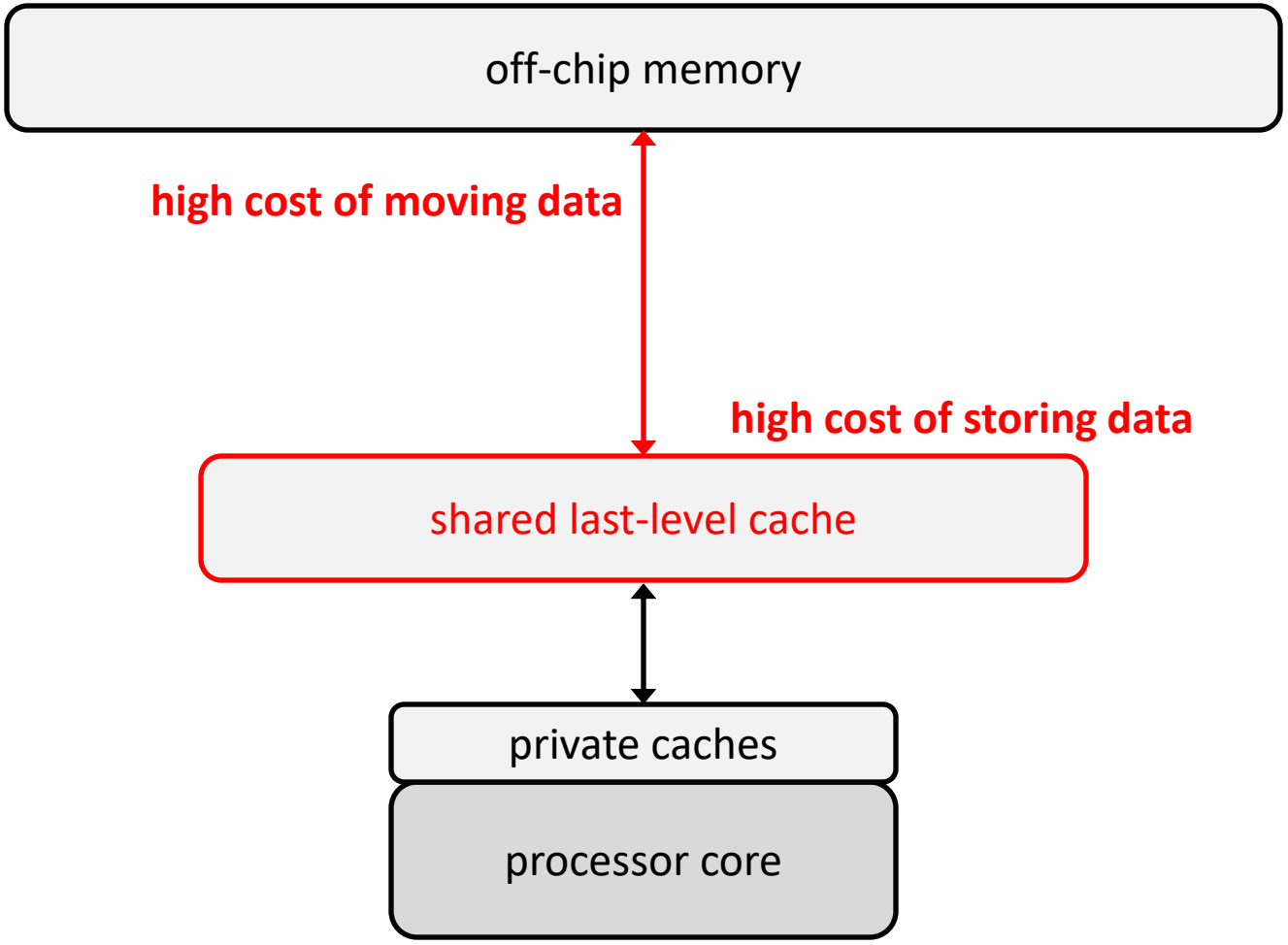
Evaluation – Dynamic Energy Savings



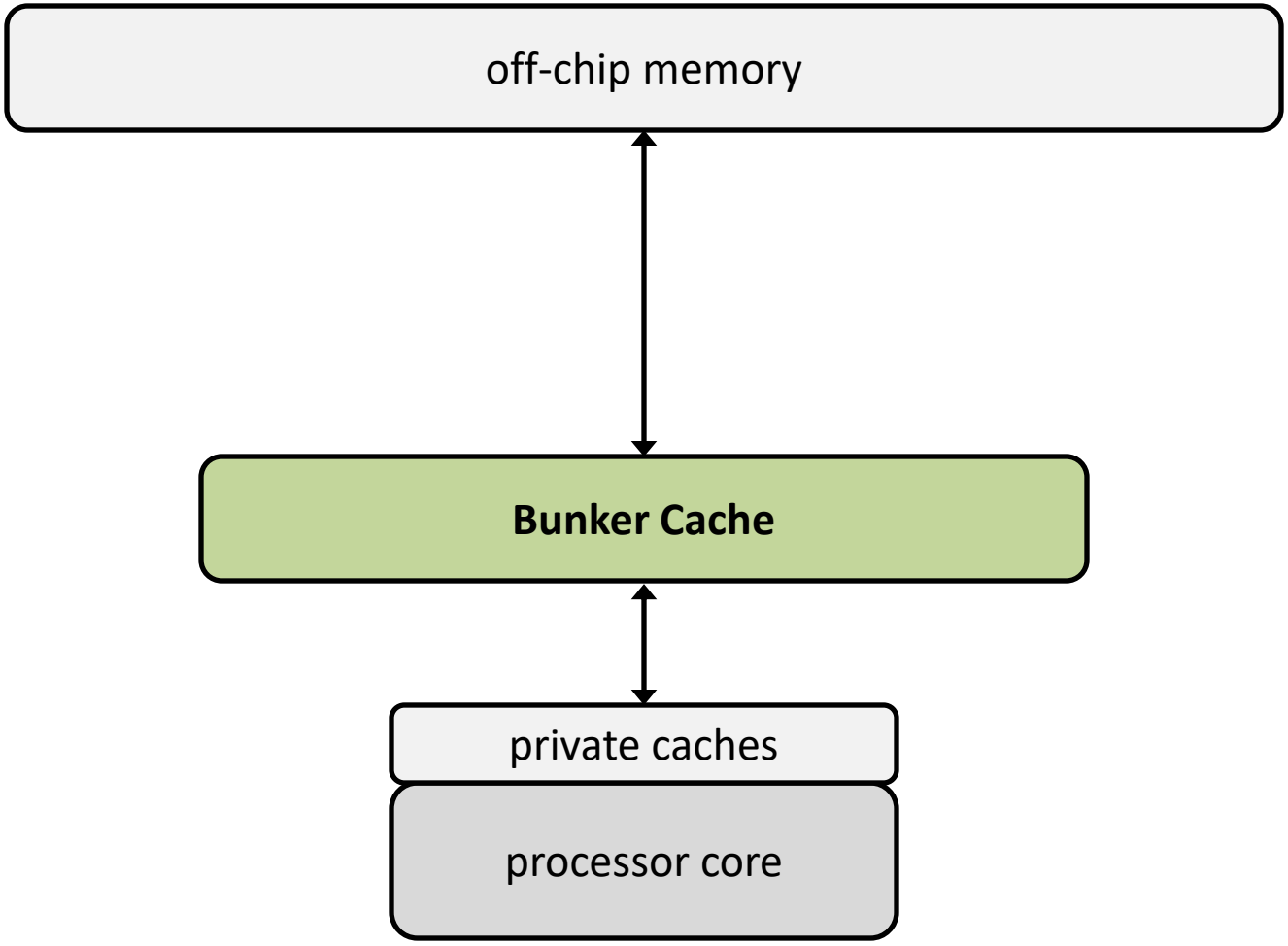
Evaluation – Leakage (Drowsy) Power Savings



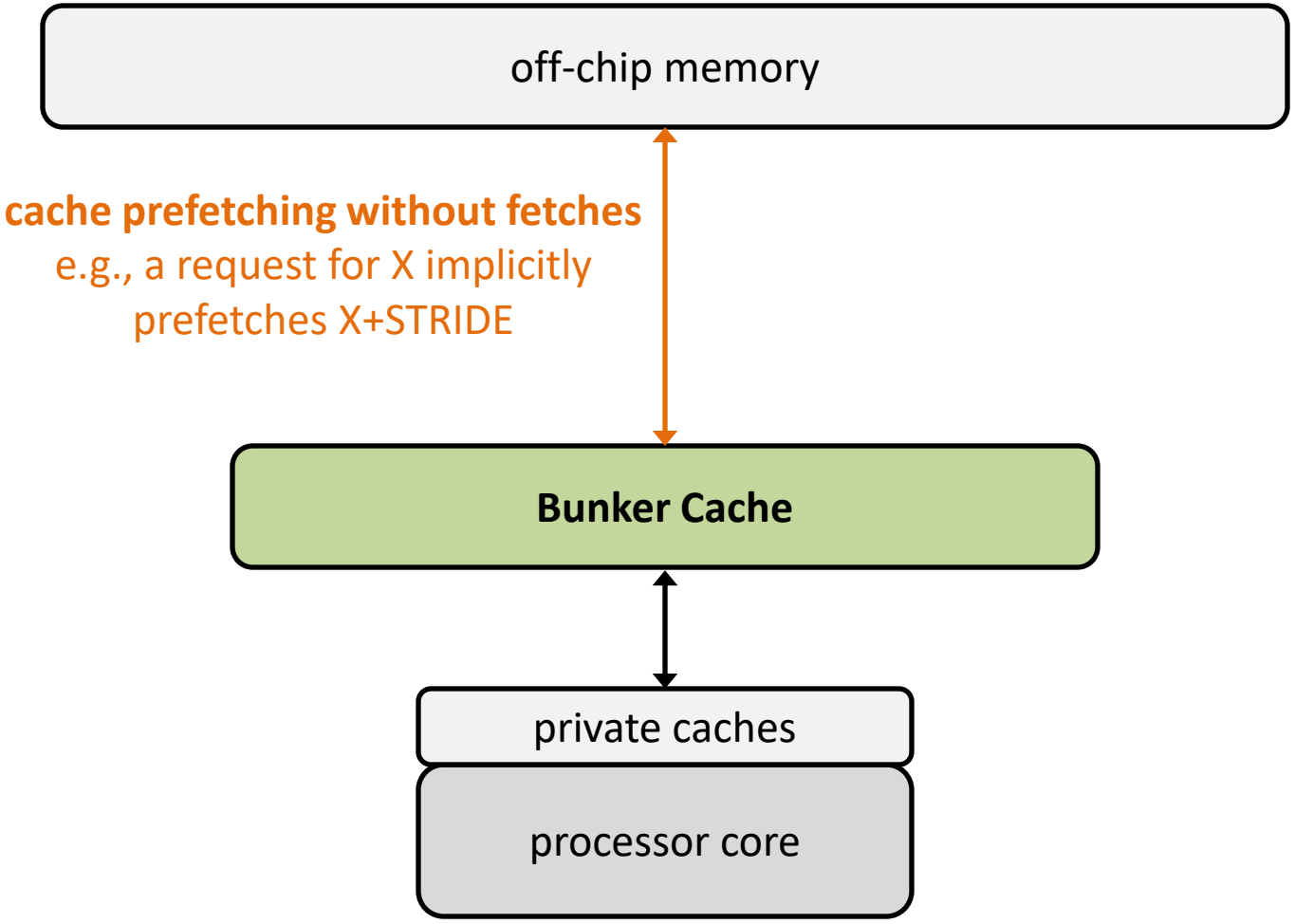
Conclusion


off-chip memory
high cost of moving data
high cost of storing data
shared last-level cache
private caches
processor core

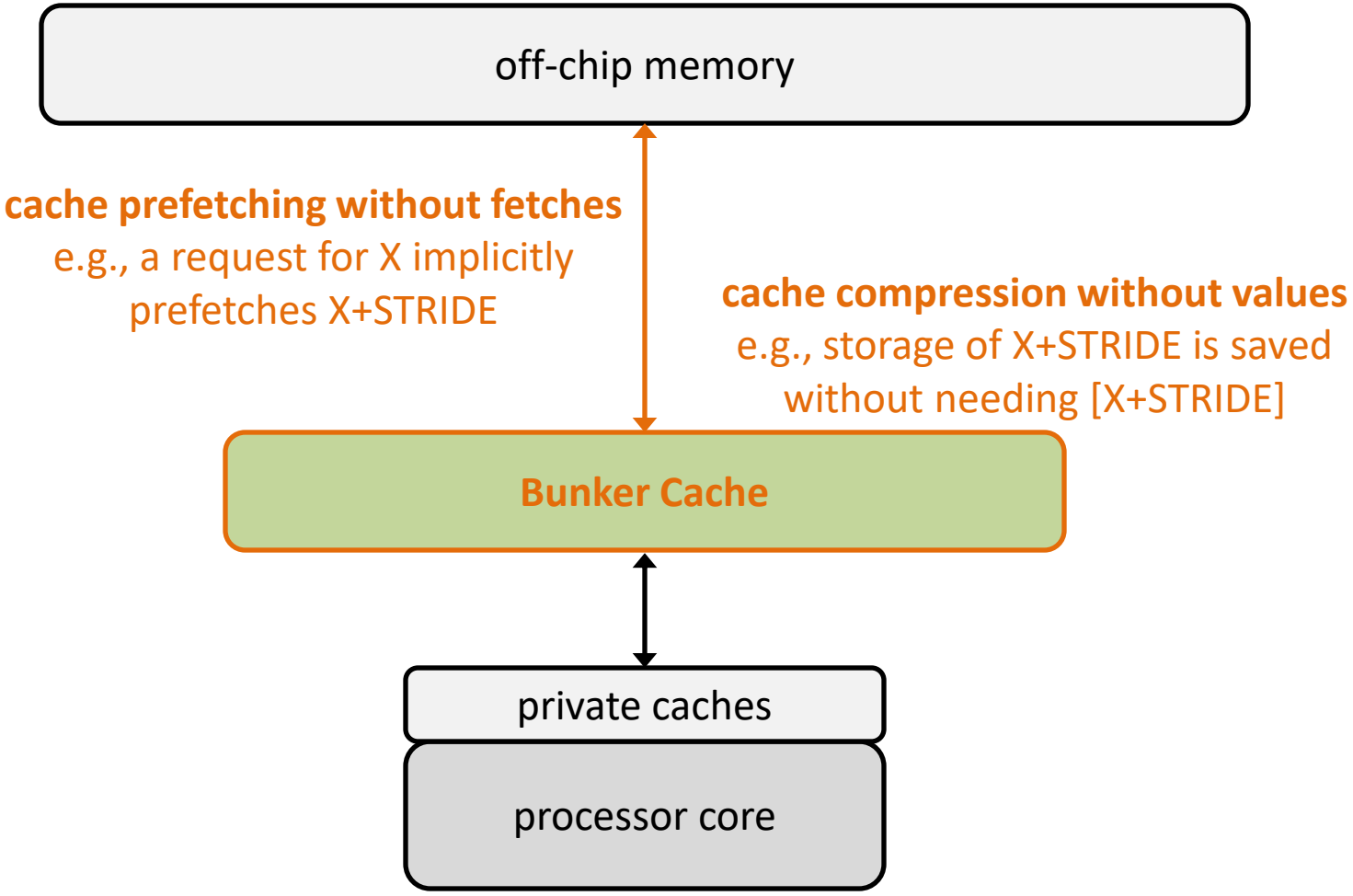
Conclusion


off-chip memory
Bunker Cache
private caches
processor core

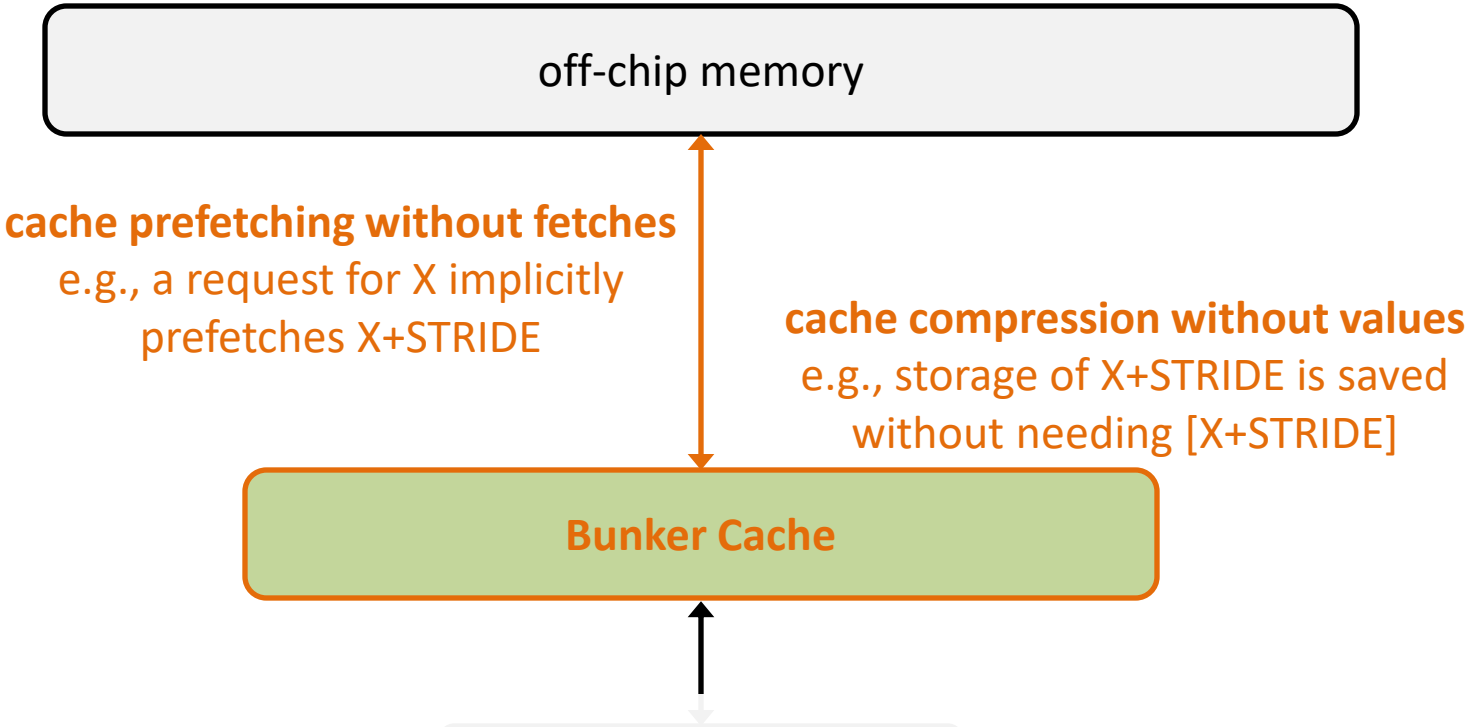
Conclusion


off-chip memory
cache prefetching without fetches
e.g., a request for X implicitly
prefetches X+STRIDE
Bunker Cache
private caches
processor core

Conclusion


off-chip memory
cache prefetching without fetches
e.g., a request for X implicitly prefetches X+STRIDE
cache compression without values
e.g., storage of X+STRIDE is saved without needing [X+STRIDE]
Bunker Cache
private caches
processor core

Conclusion



savings in runtime (1.58x), dynamic energy (1.72x), leakage power (1.65x) at radix 4

Thank you

The Bunker Cache for Spatio-Value Approximation

Joshua San Miguel
Jorge Albericio
Natalie Enright Jerger
Aamer Jaleel

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

NVIDIA