A Skewed Multi-bank Cache for Vector Processors

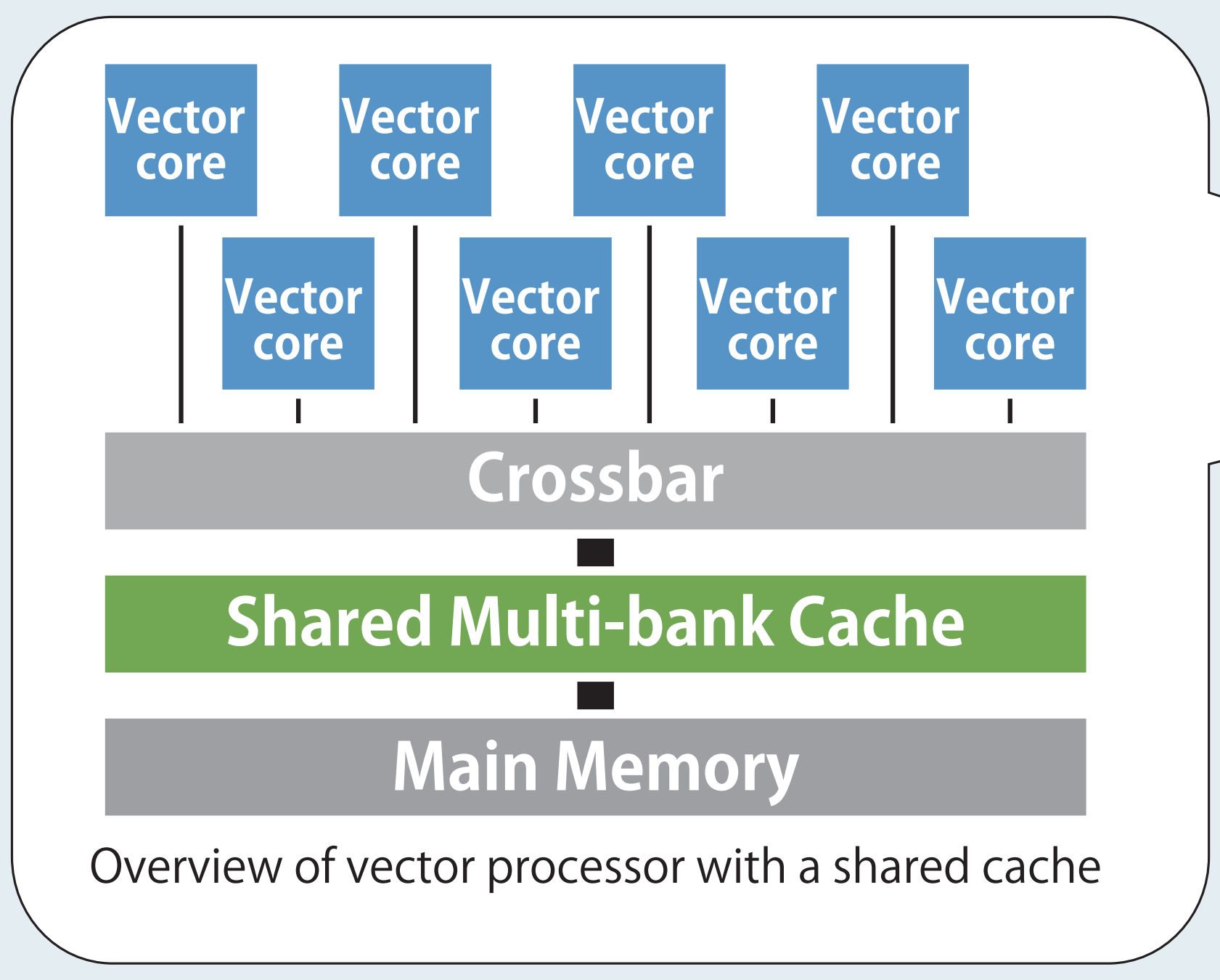
Hikaru Takayashiki, Masayuki Sato, Kazuhiko Komatsu, Hiroaki Kobayashi
Tohoku University

Introduction —Modern Vector Processor—

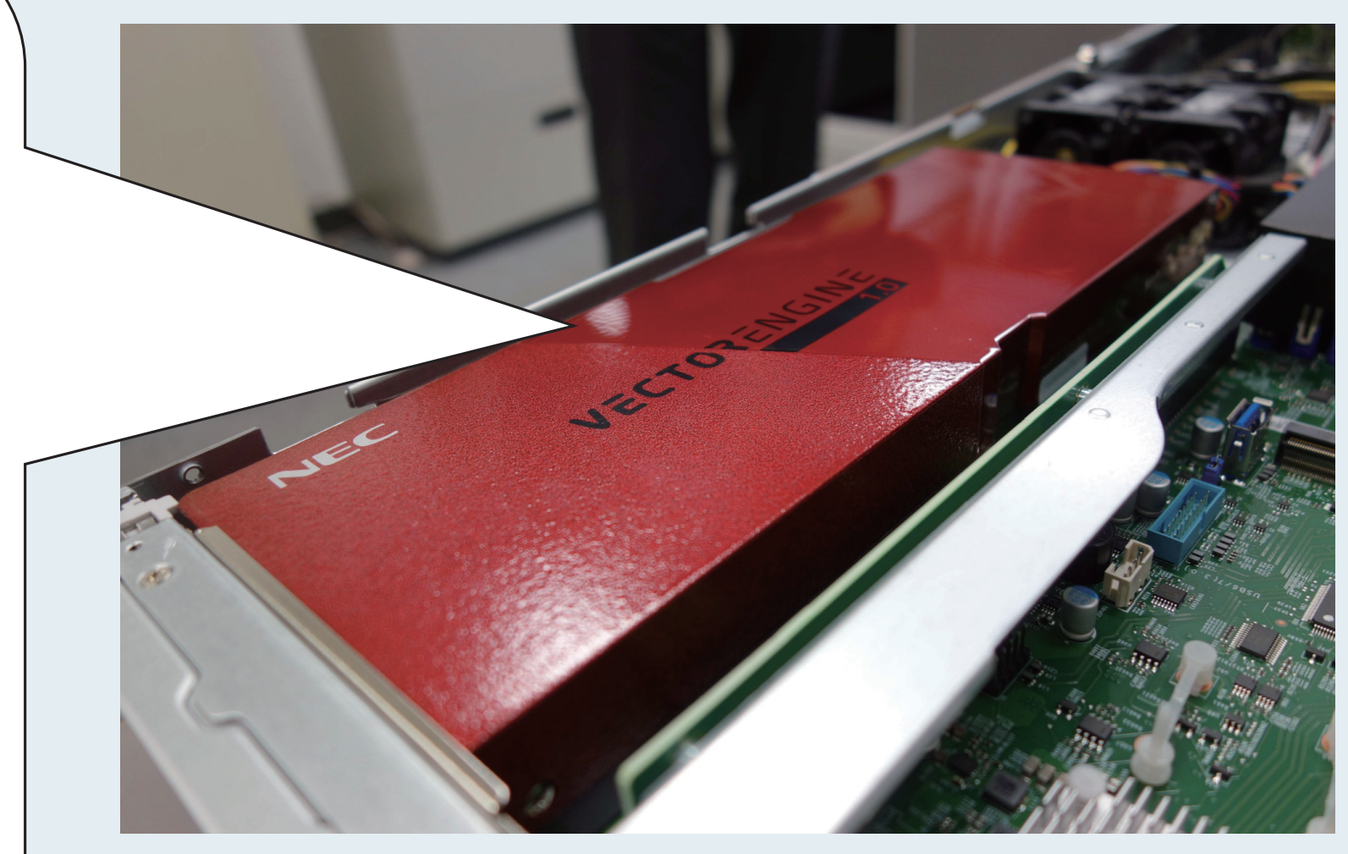
Multiple Vector Cores:
Each core can realize efficient computation by vector instructions

Multi-banked Cache:
Modern vector processors utilize a multi-bank cache for vector data to increase a cache bandwidth

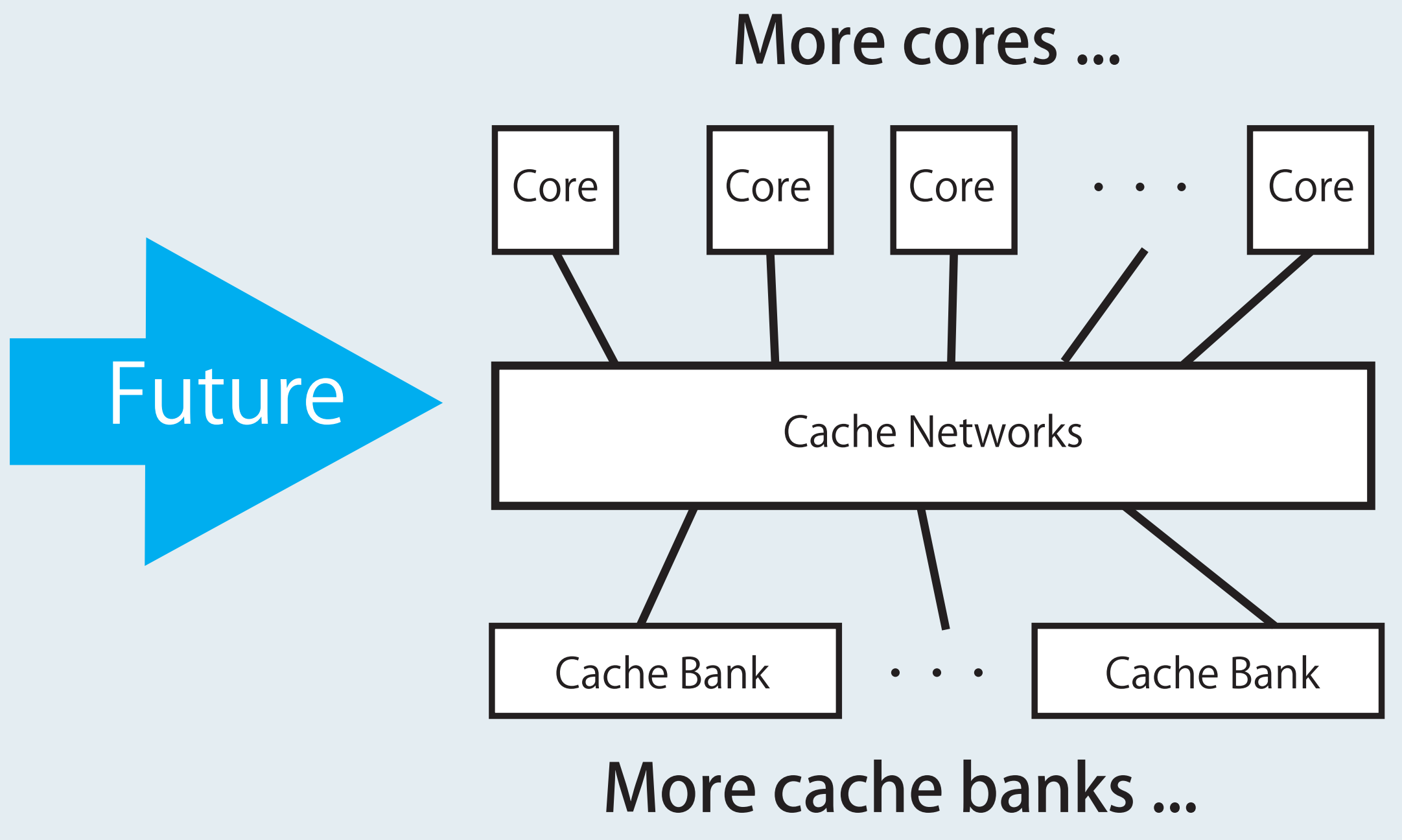
Main Memory:
The memory subsystem achieves the highest memory bandwidth by using 3-D memory modules, HBM2



Overview of vector processor with a shared cache

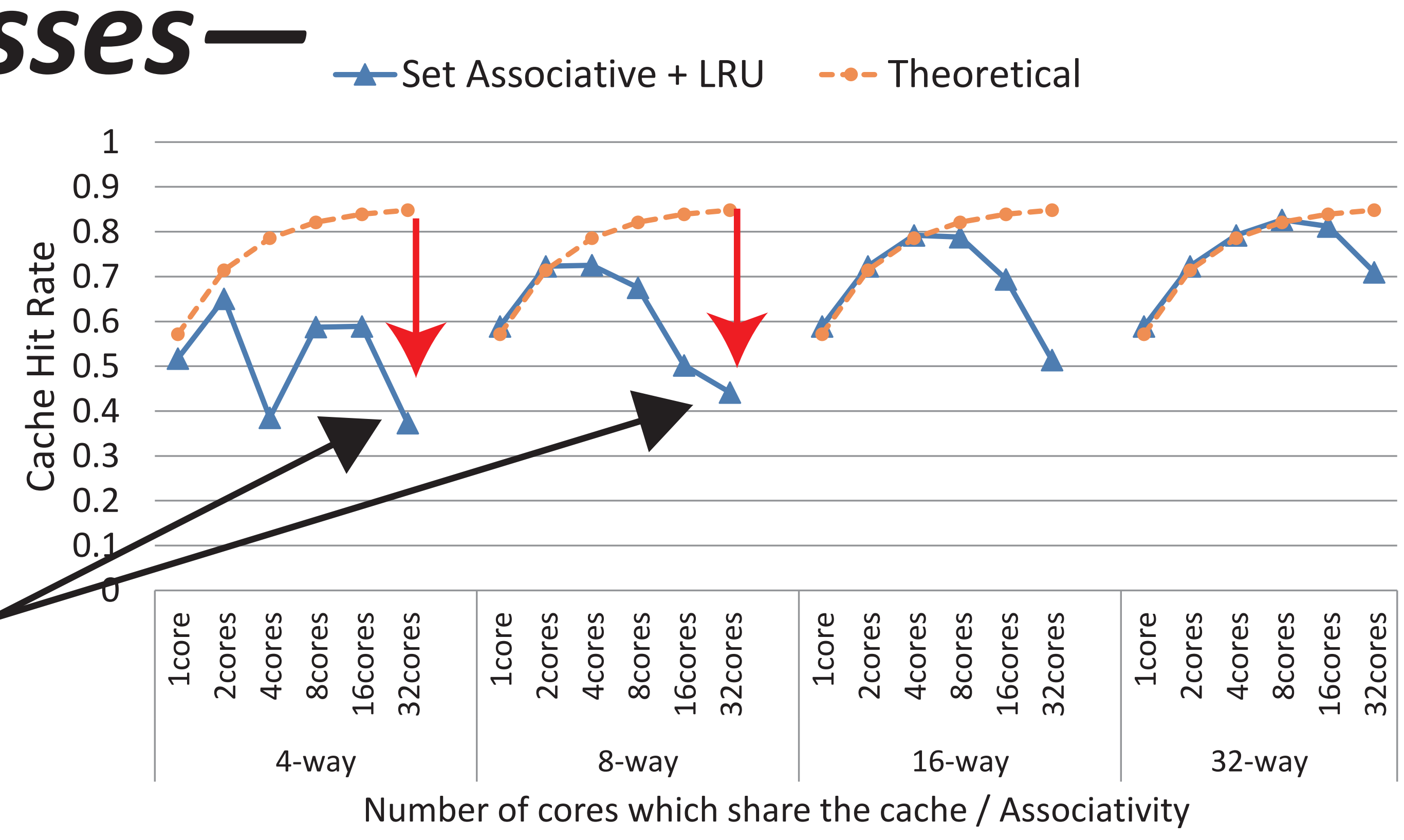


NEC SX-Aurora TSUBASA [1]



Problem —Conflict Misses—

Conflict misses can occur on scientific and engineering applications, because threads access memory with a similar pattern



The cache hit rate on 3-D 7-point stencil calculation **decreases** as increasing the number of cores that share the same cache.

A **high-associative cache** can prevent conflict misses, but **will increase power consumption and area overheads.**

A low-cost multi-bank cache without conflict misses is mandatory

Evaluation Environment

Core	Architectural setting is similar to NEC SX-ACE
Number of cores	32
Number of cores sharing a cache	1, 2, 4, 8, 16, 32
Associativity	4, 8, 16, 32
Cache capacity	1MB / core
Number of banks	16banks / core
Kernel	3-D 7-point stencil calculation

Proposal —Skewed Multi-bank Cache—

Basic Idea

Realize conflict-tolerant caches without increasing associativity by **adopting skewed-associativity** [2] on the shared cache of vector processors

What Is Skewed-associativity?

Different hashing functions are used for the distinct cache ways to **determine accessed sets from a memory address**, whereas a conventional set associative uses a single function for all the ways

Design of a Skewed Multi-bank Cache

Hashing Function:

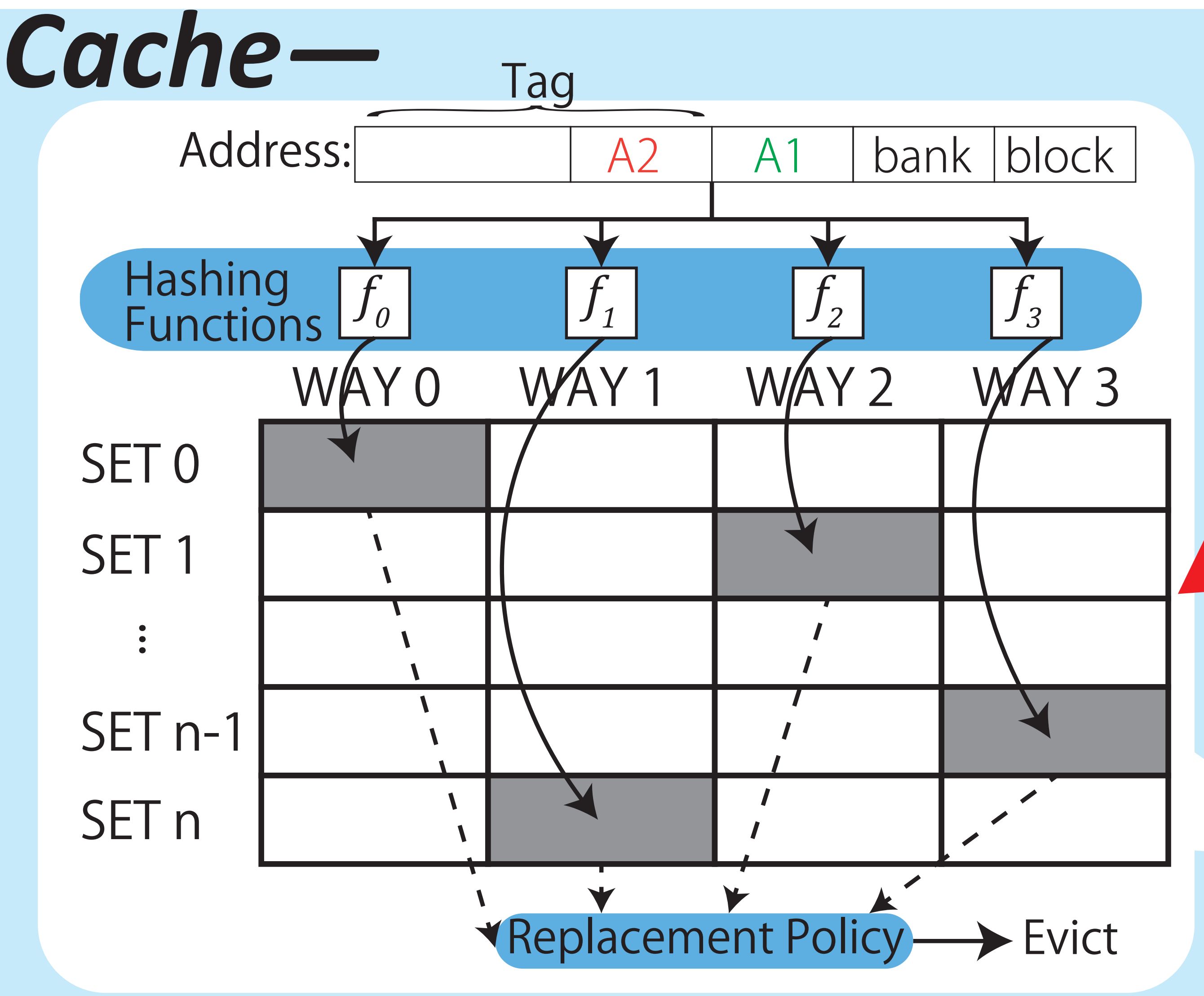
Odd-multiplier displacement [3]

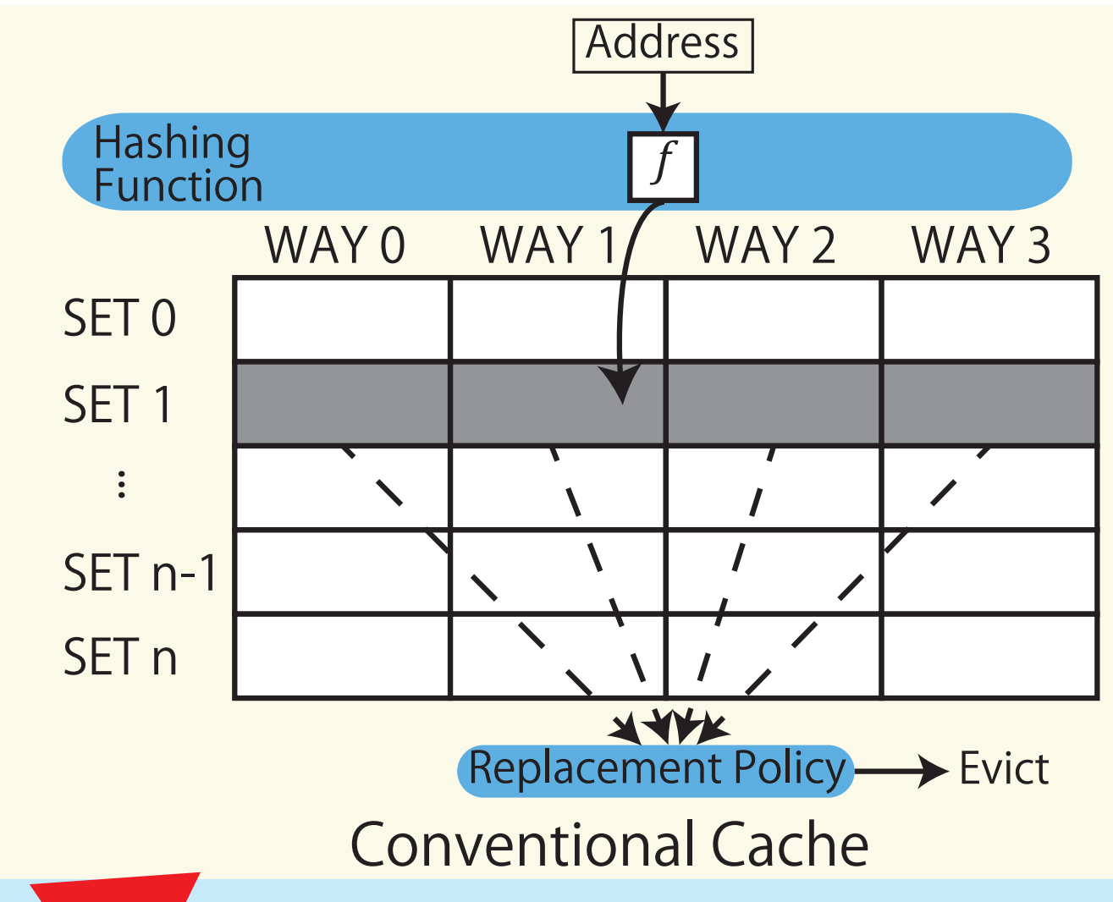
$$f(A1,A2)=(o * A2 + A1) \bmod n_{set}$$

o: arbitrary odd value (different for each ways)
n_{set}*:number of sets*

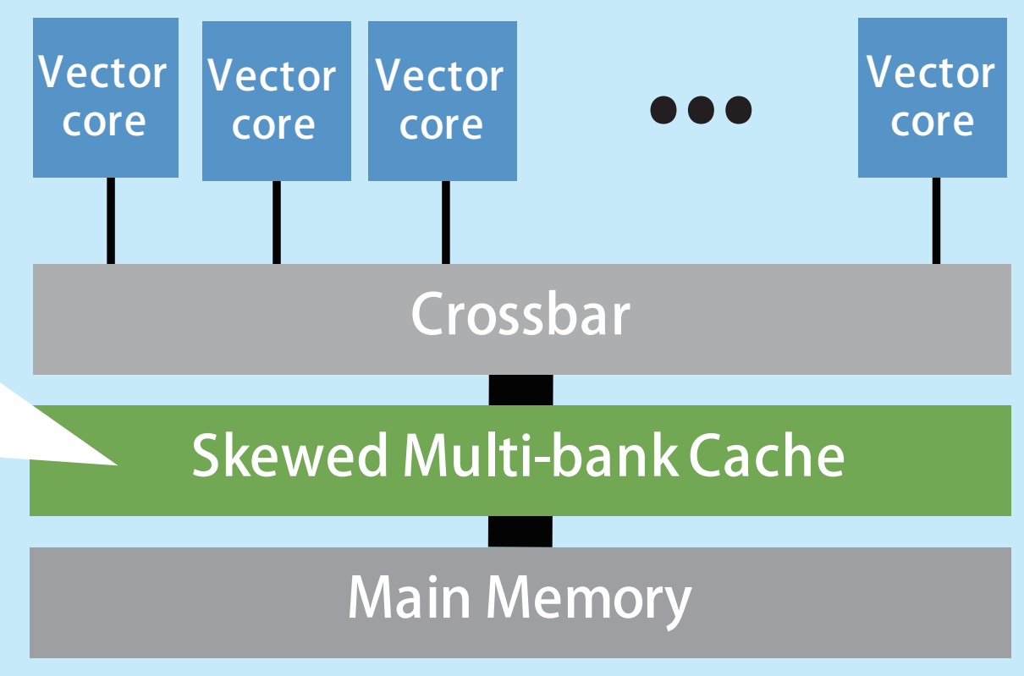
Replacement Policy:

Static re-reference interval prediction (SRRIP) [4]
SRRIP can obtain a high hit rate and be applied for a high-associative skewed cache compared with the LRU policy.



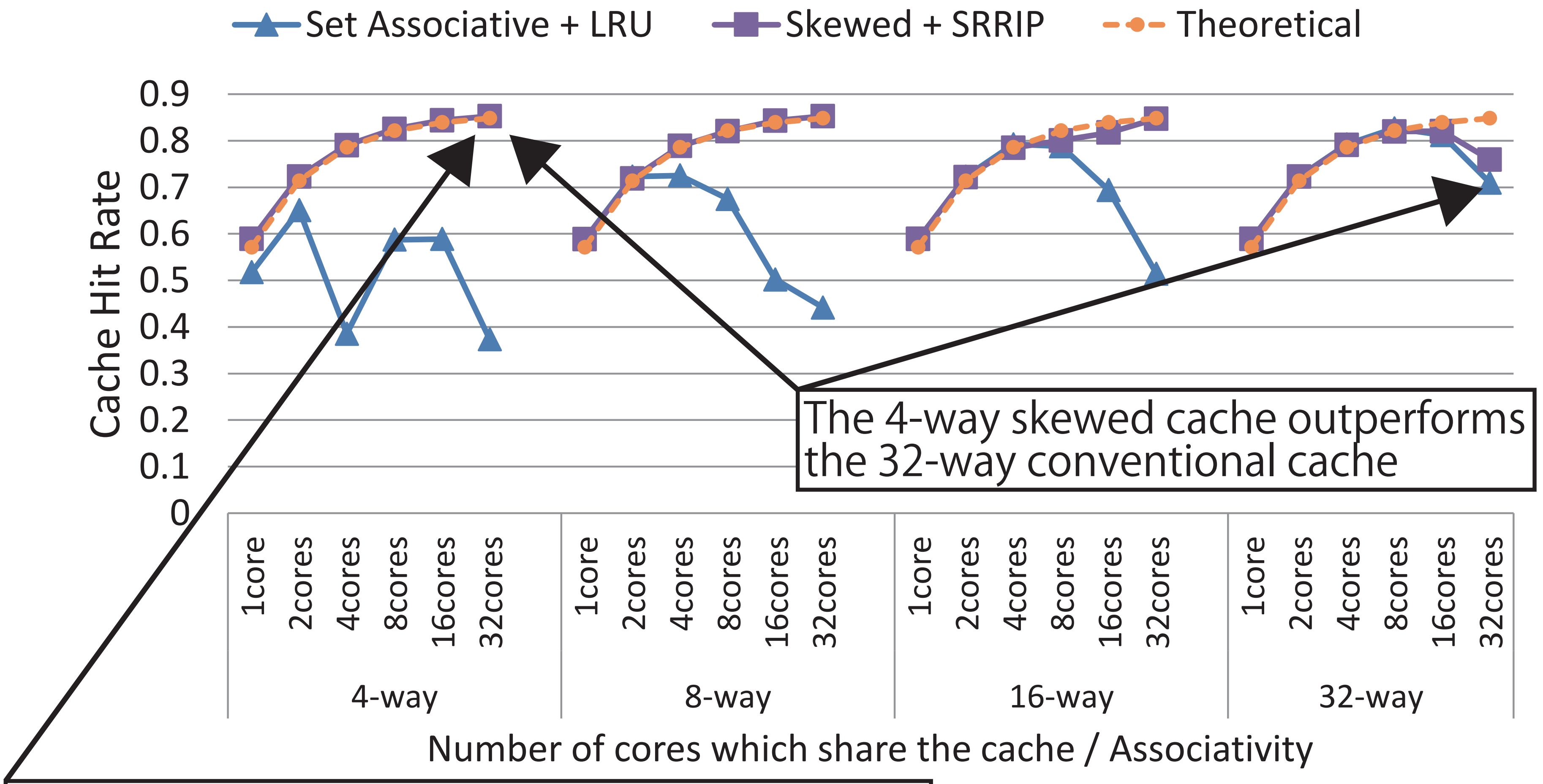


Conventional Cache



Experimental Results

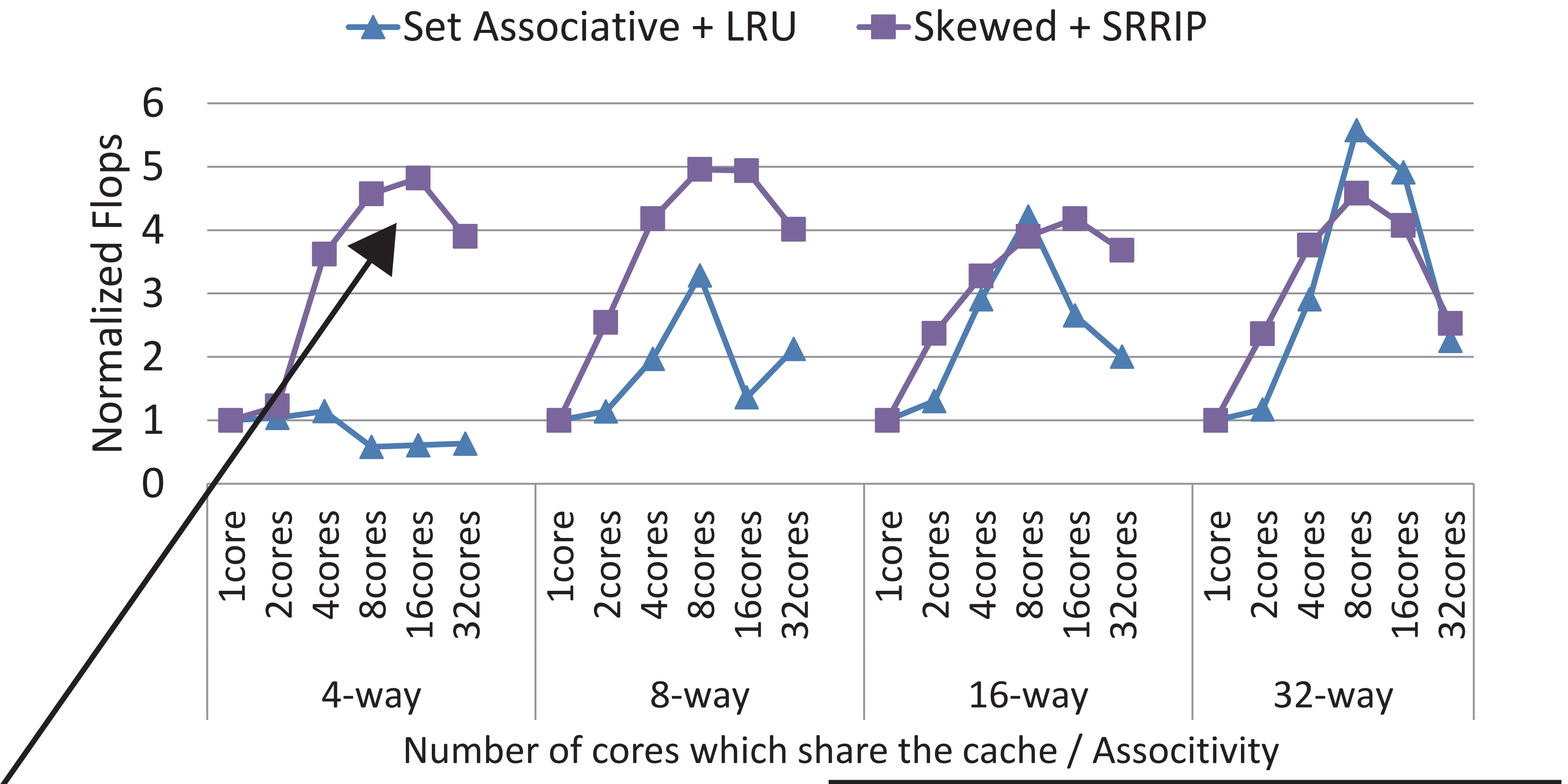
Cache Hit Rate



The 4-way skewed cache outperforms the 32-way conventional cache

The cache hit rate increases on the 4-way cache from 37% to 85% by the skewed cache

Performance



4x performance improvement in the 4-way skewed cache

The 4-way skewed cache is comparable to the 16-way & 32-way set-associative caches

Conclusions & Future Work

- Future vector processors are facing an increase in conflict misses by further increasing the number of vector cores
- A skewed multi-bank cache can reduce conflict misses and improve performance
- More detailed implementations and their hardware costs of the skewed multi-bank cache should be considered in the future

References

[1] Yohei, Y. et al.: Vector Engine Processor of NEC' s Brand-New supercomputer SX-Aurora TSUBASA, Hot Chips 2018, 2018
[2] Seznec, A. : A Case for Two-way Skewed-associative Caches, ISCA '93, 1993
[3] Kharbutli, M. et al. : Eliminating Conflict Misses Using Prime Number-Based Cache Indexing,IEEE Trans. Comput., 2005
[4] Jaleel, A. et al.: High Performance Cache Replacement Using Re-reference Interval Prediction (RRIP), SIGARCH, 2010