Mitigating Prefetcher-Caused Pollution Using Informed Caching Policies for Prefetched Blocks

Vivek Seshadri

Samihan Yedkar · Hongyi Xin · Onur Mutlu

Phillip B. Gibbons · Michael A. Kozuch · Todd C. Mowry

Summary

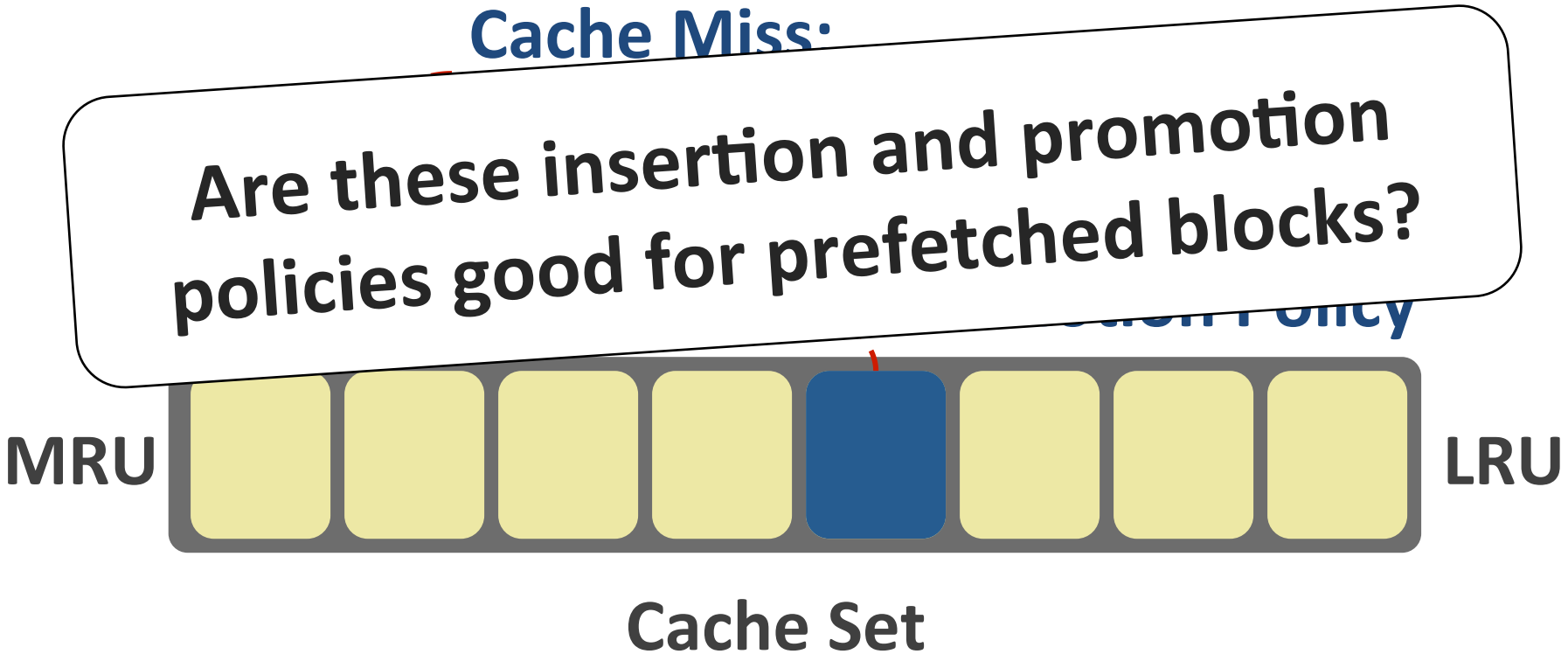
- Existing caching policies for prefetched blocks result in cache pollution

1) Accurate Prefetches (ICP Demotion)
   - 95% of useful prefetched blocks are used only once!
   - Track prefetched blocks in the cache
   - Demote prefetched block on cache hit
2) Inaccurate Prefetches (ICP Accuracy Prediction)
   - Existing accuracy prediction mechanisms get stuck in positive feedback
   - Self-tuning Accuracy Predictor

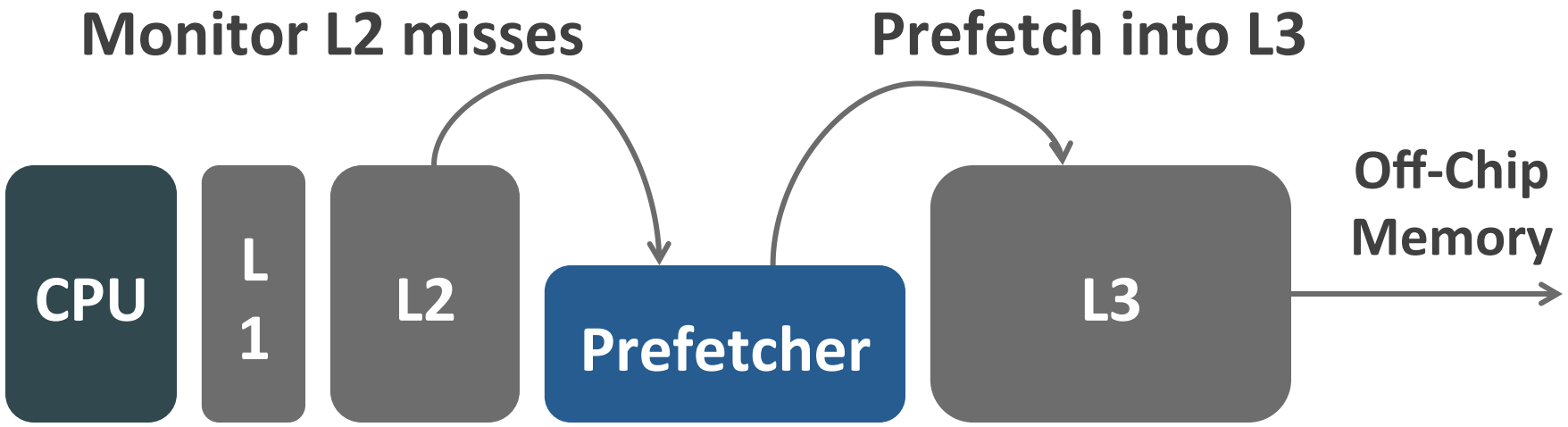
- ICP (combines both mechanisms)
  - Significantly reduces prefetch pollution
  - 6% performance improvement over 157 2-core workloads

Caching Policies for Prefetched Blocks

Problem: Existing caching policies for prefetched blocks result in significant cache pollution



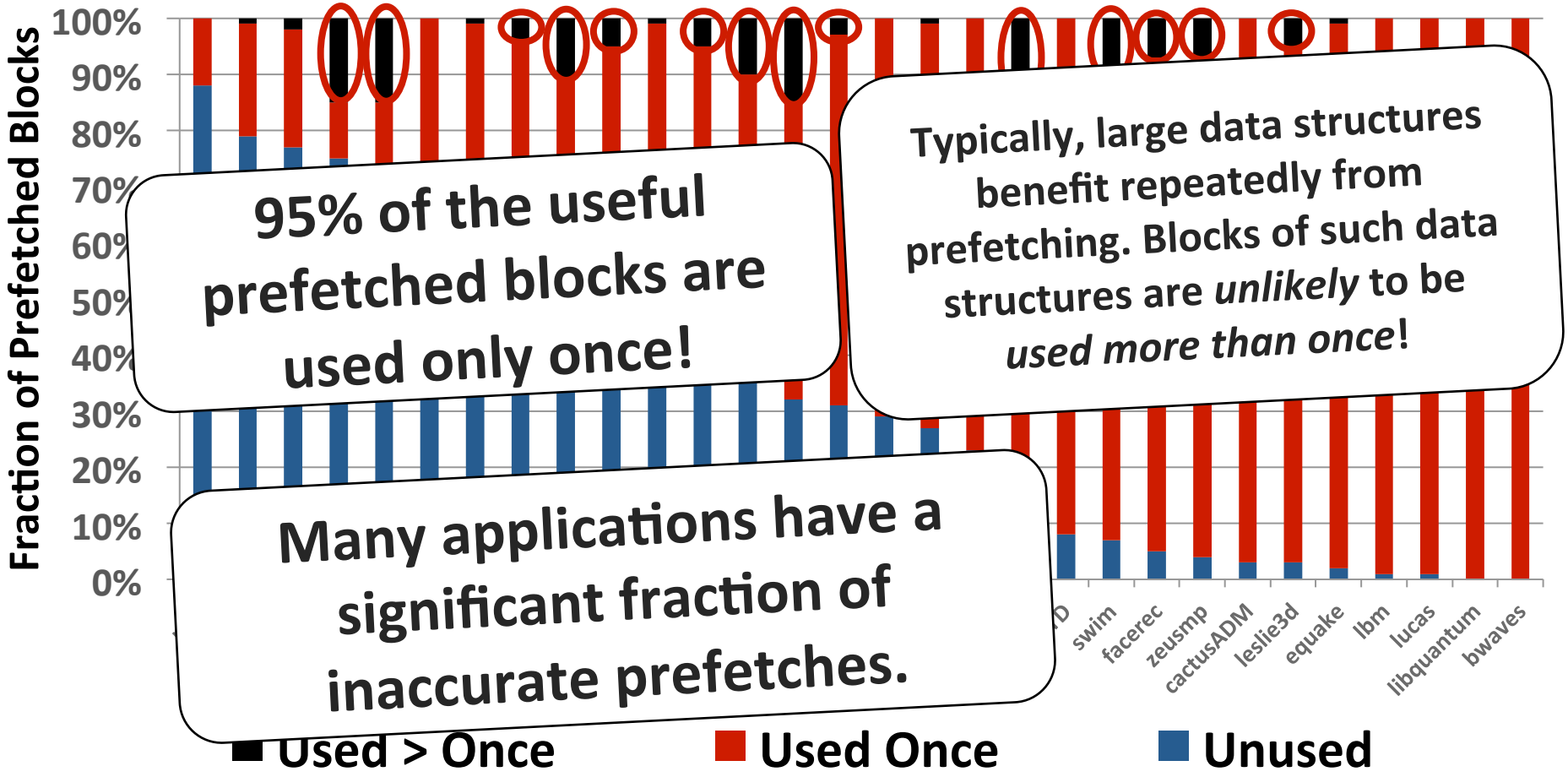
Prefetch Usage Experiment



Classify prefetched blocks into three categories

1. Blocks that are unused
2. Blocks that are used exactly once before evicted from cache
3. Blocks that are used more than once before evicted from cache

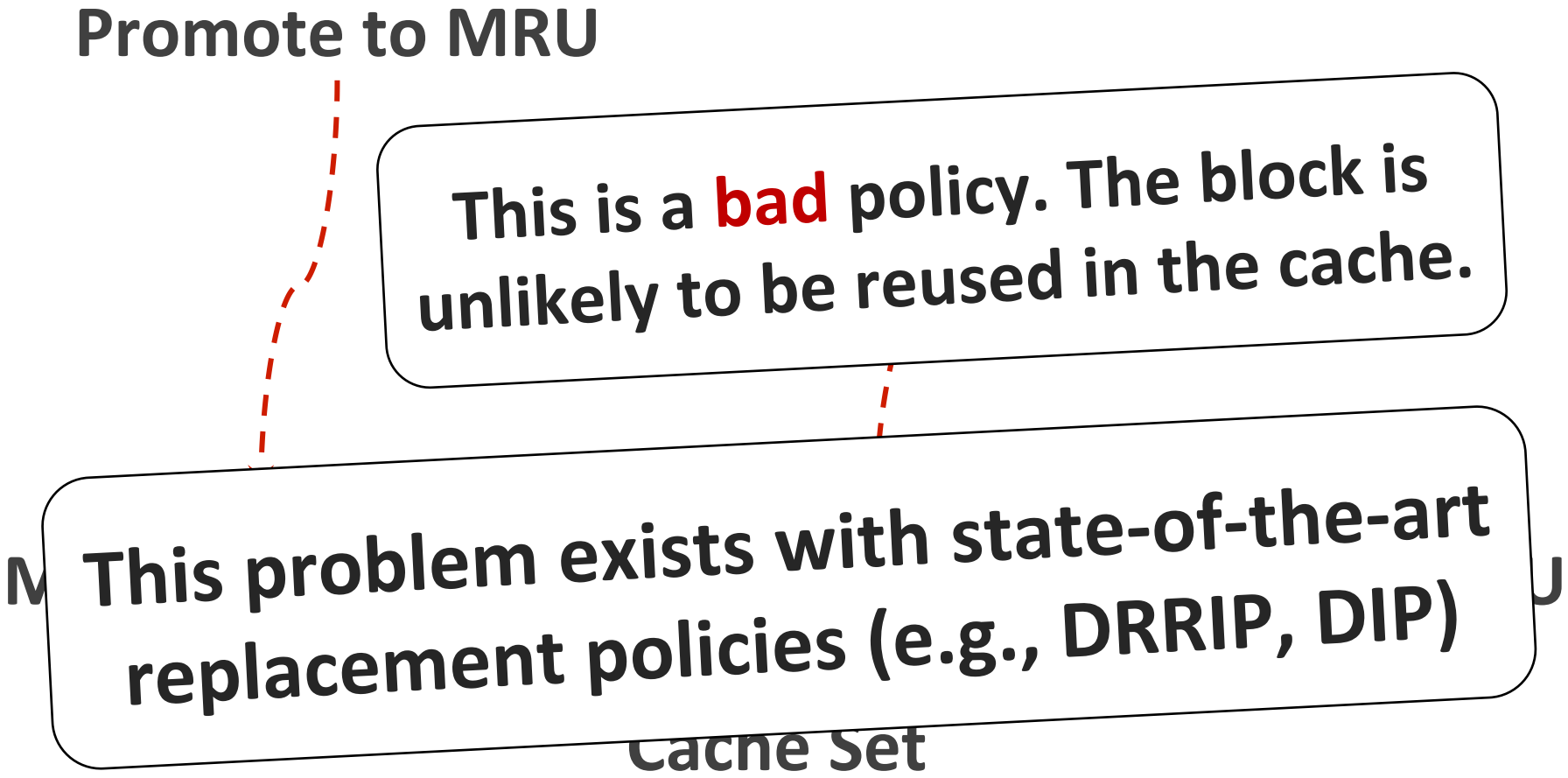
Usage Distribution of Prefetched Blocks



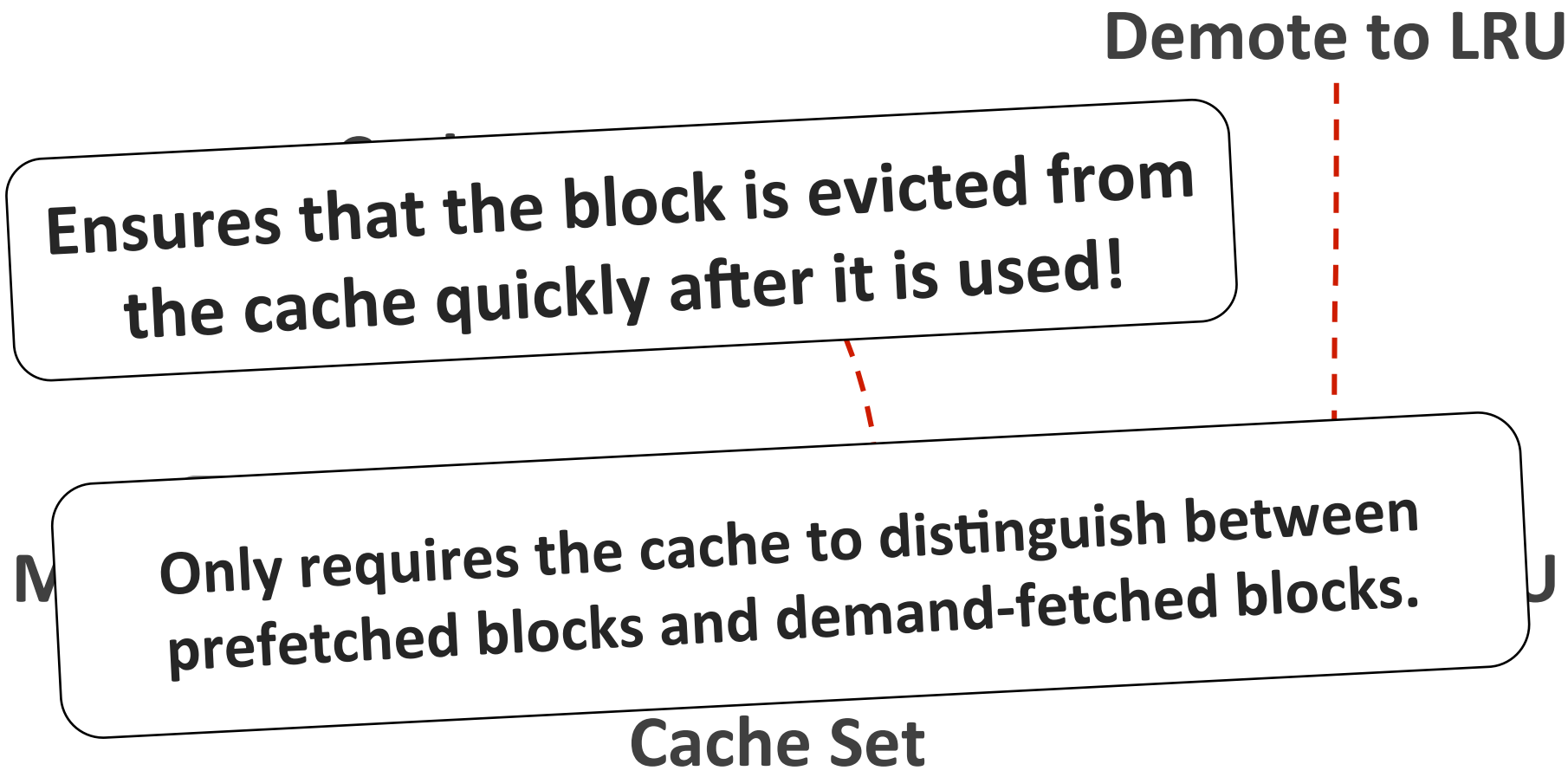
Outline

- ✓ Introduction
- **ICP – Mechanism**
  - ICP promotion policy
  - ICP insertion policy
- Prior Works
- Evaluation
- Conclusion

Shortcoming of Traditional Promotion Policy



ICP Demotion



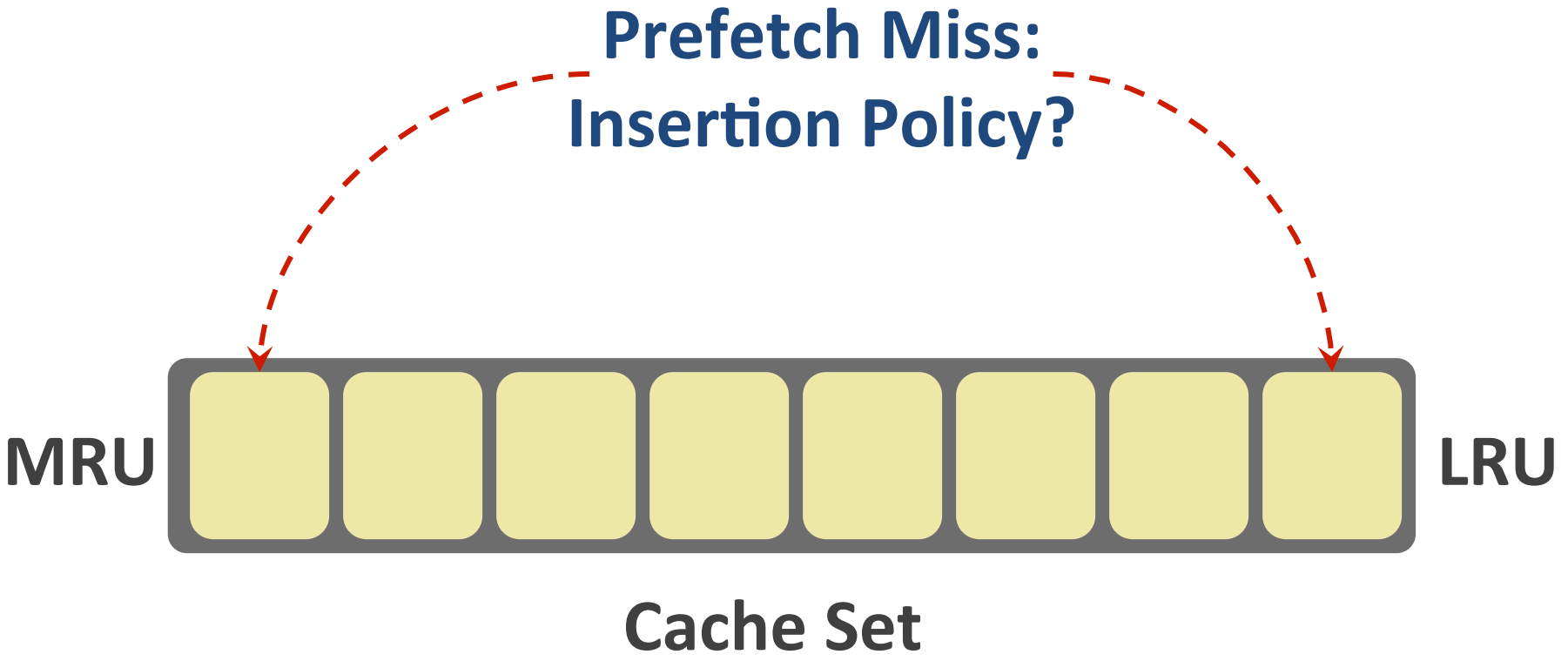
Outline

- ✓ Introduction
- ICP – Mechanism
  - ICP promotion policy
  - ICP insertion policy
- Prior Works
- Evaluation
- Conclusion

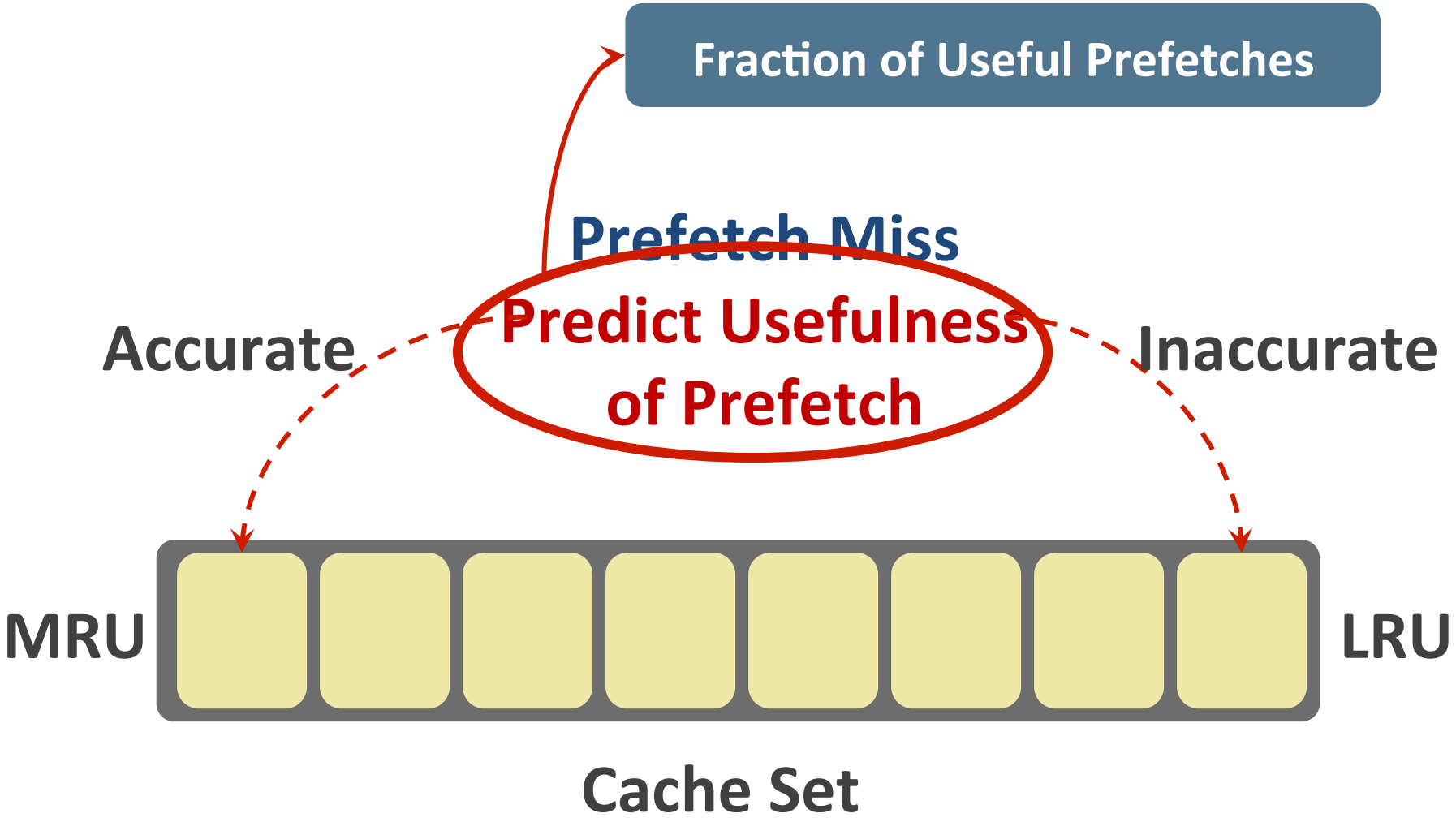
Cache Insertion Policy for Prefetched Blocks

Good (Accurate prefetch)
Bad (Inaccurate prefetch)

Good (Inaccurate prefetch)
Bad (accurate prefetch)

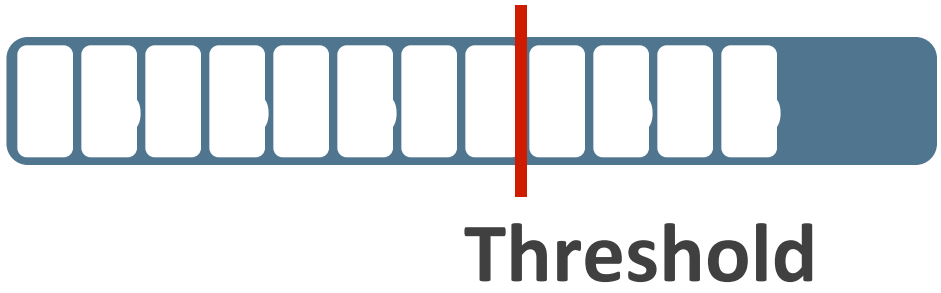


Predicting Usefulness of Prefetch



Shortcoming of "Fraction of Useful Prefetches"

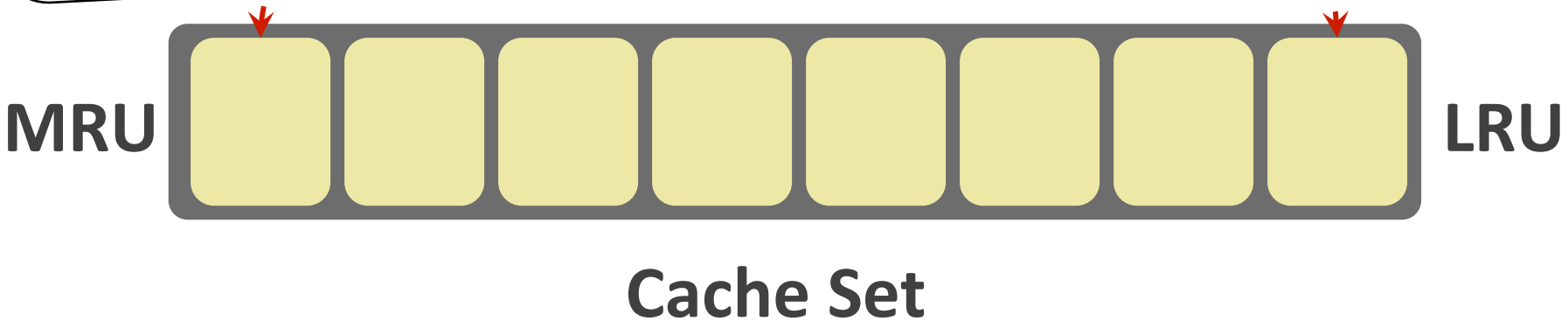
Accurate prefetches predicted as inaccurate and evicted before use



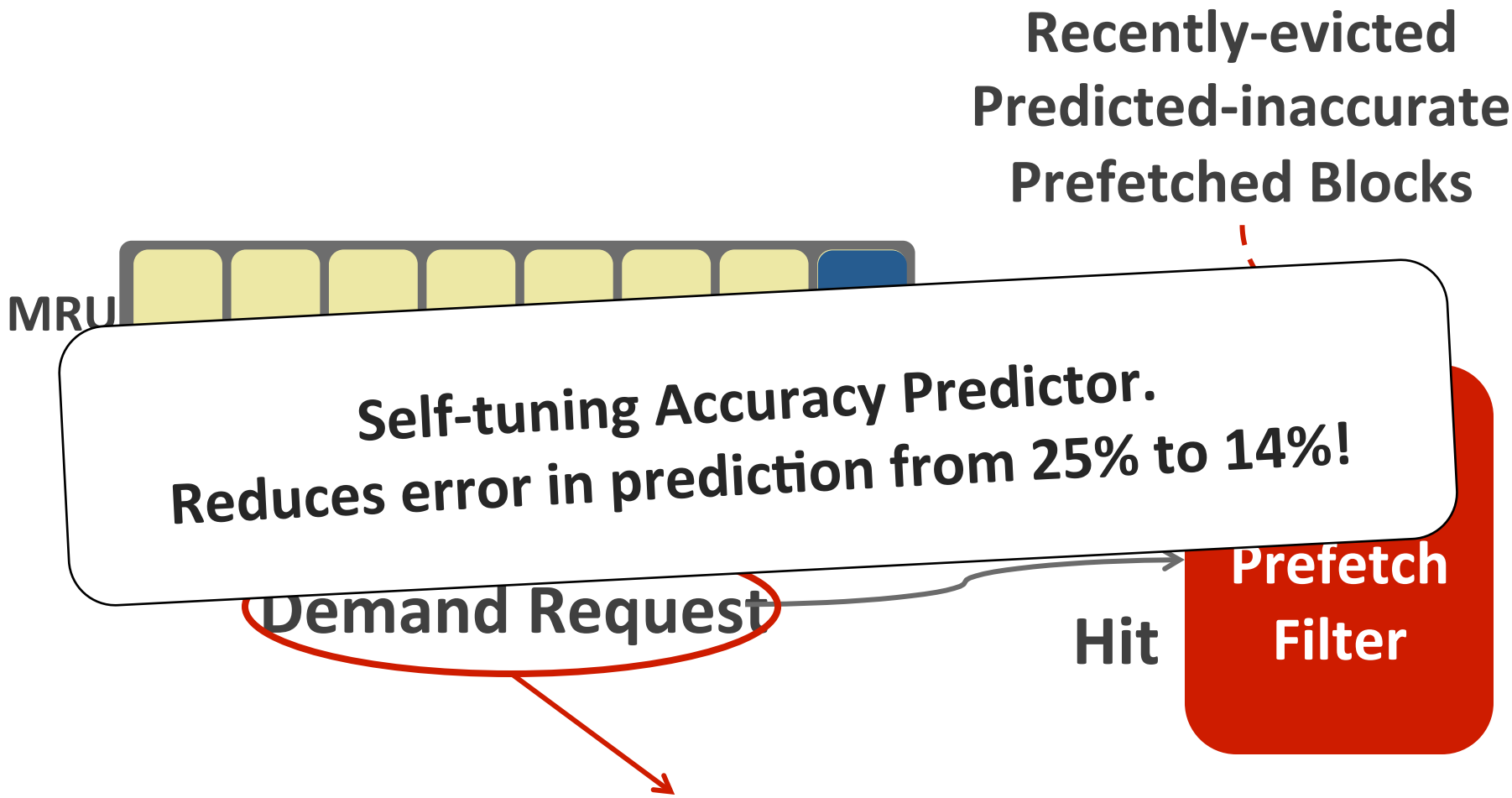
The predictor may get stuck in a state where all prefetches are predicted to be inaccurate!

refetch Miss

Inaccurate



ICP Accuracy Prediction



Accurate prefetch mispredicted as inaccurate

ICP – Summary

- **ICP Demotion (ICP-D)**
  - Track prefetched blocks in the cache
  - Demote prefetched block to LRU on cache hit
- **ICP Accuracy Prediction (ICP-AP)**
  - Maintain accuracy counter for each prefetcher entry
  - Evicted Prefetch Filter (EPF): tracks recently-evicted predicted-inaccurate prefetches
  - Bump up accuracy counter on cache miss + EPF hit
- **Hardware Cost: only 12KB for a 1MB cache**

Outline

- ✓ Introduction
- ✓ ICP – Mechanism
  - ICP promotion policy
  - ICP insertion policy
- Prior Works
- Evaluation
- Conclusion

Prior Works

- **Feedback Directed Prefetching (FDP)** (Srinath+ HPCA-07)
  - Use pollution filter to determine degree of prefetch pollution
  - Insert all prefetches at LRU if pollution is high
  - Can insert accurate prefetches at LRU

- **Prefetch-Aware Cache Management (PACMan)** (Wu+ MICRO-11)
  - Insert prefetches both into L2 and L3
  - Accesses to L3 filtered by L2 (directly insert at LRU in L3)
  - Does not mitigate pollution at L2!

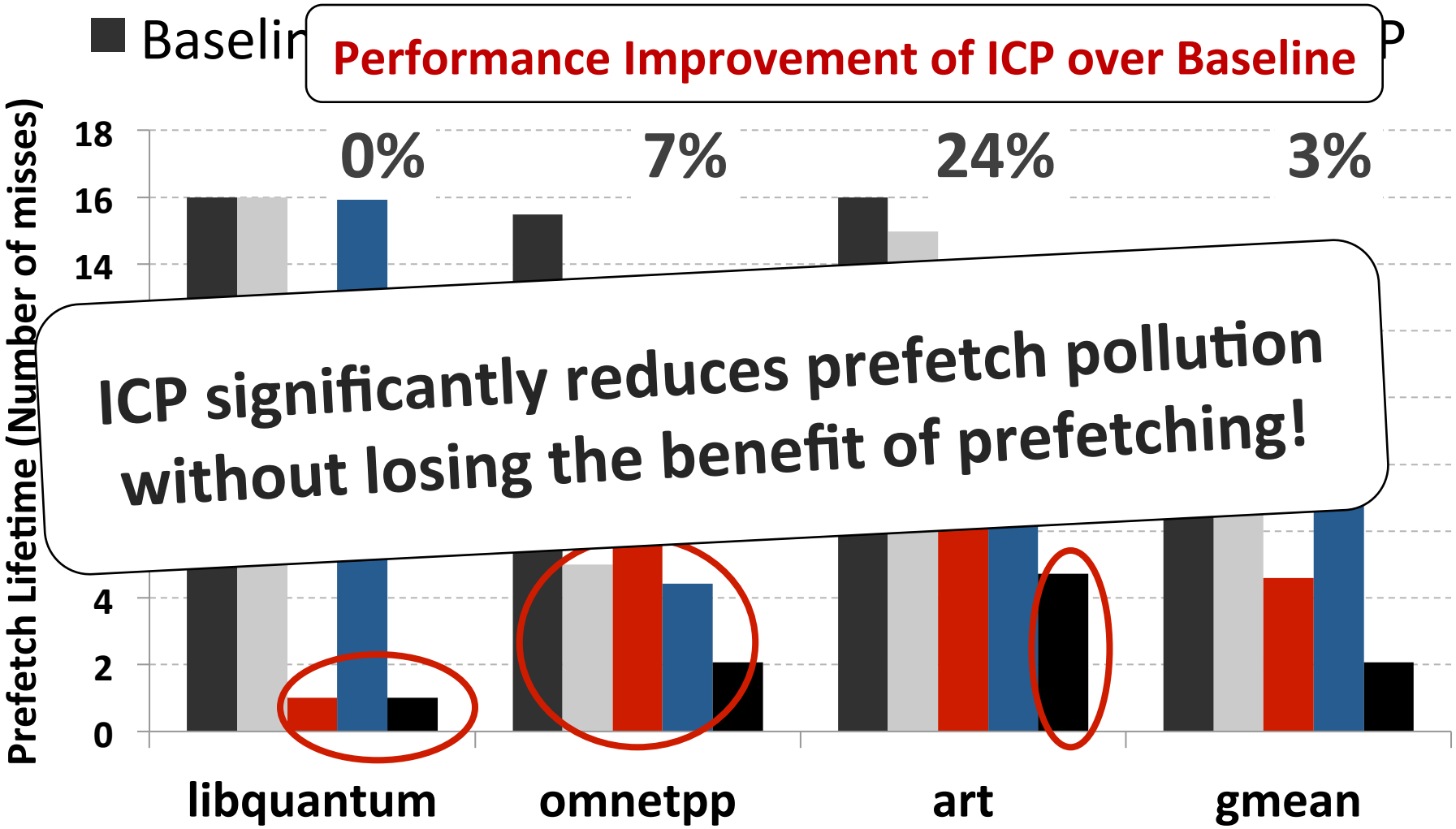
Outline

- ✓ Introduction
- ✓ ICP – Mechanism
  - ICP promotion policy
  - ICP insertion policy
- ✓ Prior Works
- Evaluation
- Conclusion

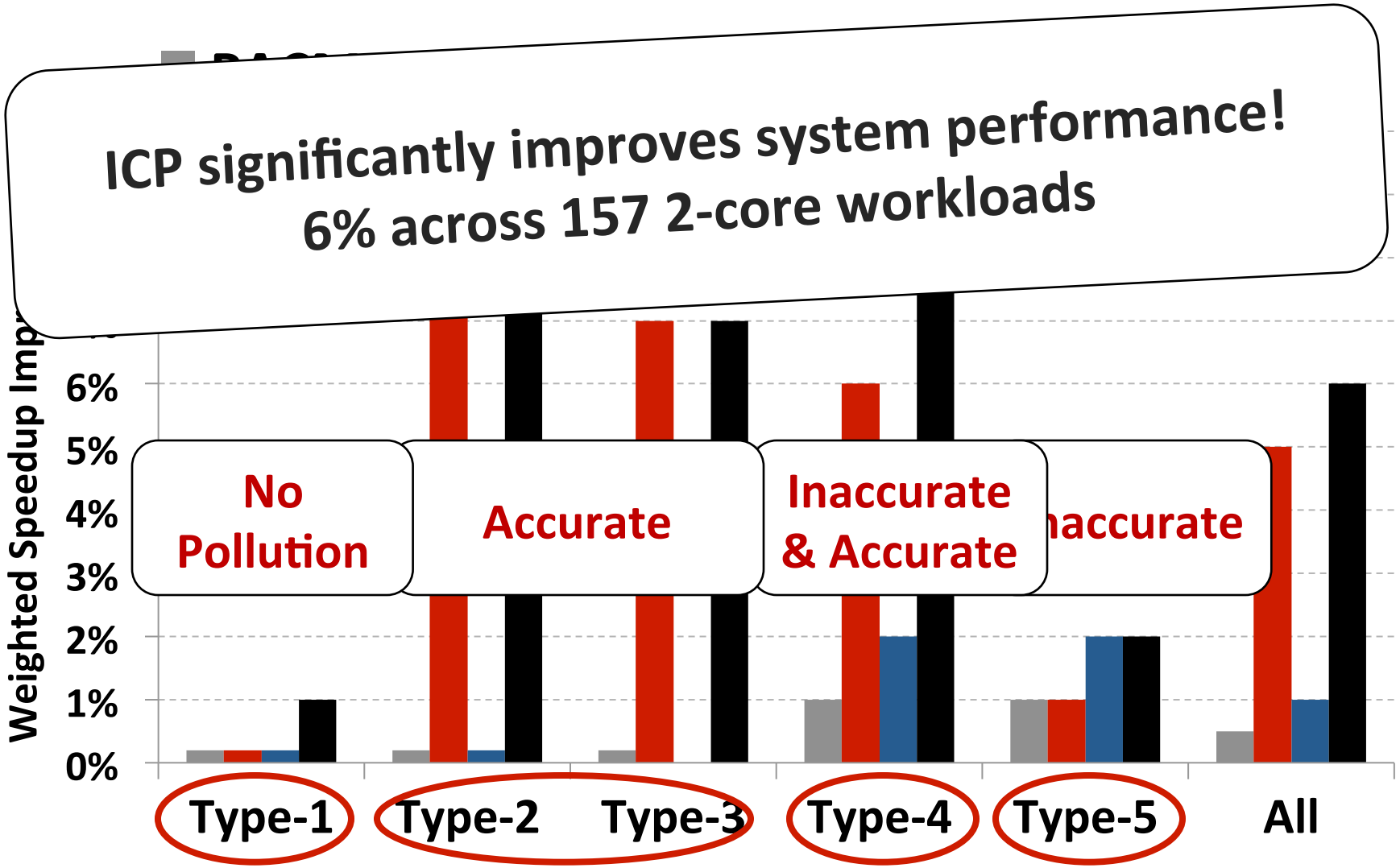
Methodology

- **Simulator** (released publicly) http://www.ece.cmu.edu/~safari/tools/memsim.tar.gz
  - 1-8 cores, 4Ghz, In-order/Out-of-order
  - 32KB private L1 cache, 256KB private L2 cache
  - Aggressive stream prefetcher (16-entries/core)
  - Shared L3 cache (1MB/core)
  - DDR3 DRAM Memory
- Workloads
  - SPEC CPU2006, TPCC, TPCH, Apache
  - 157 2-core, 20 4-core, and 20 8-core workloads
- Metrics
  - Prefetch lifetime (measure of prefetch pollution)
  - IPC, Weighted Speedup, Harmonic Speedup, Maximum Slowdown

Single Core – Prefetch Lifetime



2-Core Performance



Other Results in the Paper

- Sensitivity to cache size and memory latency
- Sensitivity to number of cores
- Sensitivity to cache replacement policy (LRU, DRRIP)
- Performance with out-of-order cores
- Benefits with stride prefetching
- Comparison to other prefetcher configurations

Conclusion

- Existing caching policies for prefetched blocks result in cache pollution

1) Accurate Prefetches (ICP Demotion)
   - 95% of useful prefetched blocks are used only once!
   - Track prefetched blocks in the cache
   - Demote prefetched block on cache hit
2) Inaccurate Prefetches (ICP Accuracy Prediction)
   - Existing accuracy prediction mechanisms get stuck in positive feedback
   - Self-tuning Accuracy Predictor

- ICP (combines both mechanisms)
  - Significantly reduces prefetch pollution
  - 6% performance improvement over 157 2-core workloads

Mitigating Prefetcher-Caused Pollution Using Informed Caching Policies for Prefetched Blocks

Vivek Seshadri

Samihan Yedkar · Hongyi Xin · Onur Mutlu

Phillip B. Gibbons · Michael A. Kozuch · Todd C. Mowry