CHROME: Concurrency-Aware Holistic Cache Management Framework with Online Reinforcement Learning

Xiaoyang Lu[1], Hamed Najafi[2], Jason Liu[2], Xian-He Sun[1]
[1]Illinois Institute of Technology, [2]Florida International University

Cache Management

Essential for bridging the performance gap between fast CPU and slower main memory:

Cache Replacement: Determines which cache blocks to evict when new data needs to be loaded

Cache Bypassing: Decides whether incoming data should be stored in the cache

Prefetching: Predictively loads data into the cache before it is actually requested by the CPU

Limitations of Current Cache Management Schemes

We observe there are **two common limitations** faced by traditional cache management techniques:

Lack of Holistic View: Current schemes often examine cache replacement, bypassing, and prefetching in isolation, overlooking the potential benefits that could arise from a joint optimization strategy

Lack of Adaptability: Current schemes often rely on fixed heuristics that don't account for the changing access patterns of modern applications and system configurations

Our Solution: CHROME

A **holistic** cache management framework that **dynamically adapts** to various workloads and system configurations**:**

Holistic Integration: Integrates **cache bypassing** and **replacement** with pattern-based **prefetching**

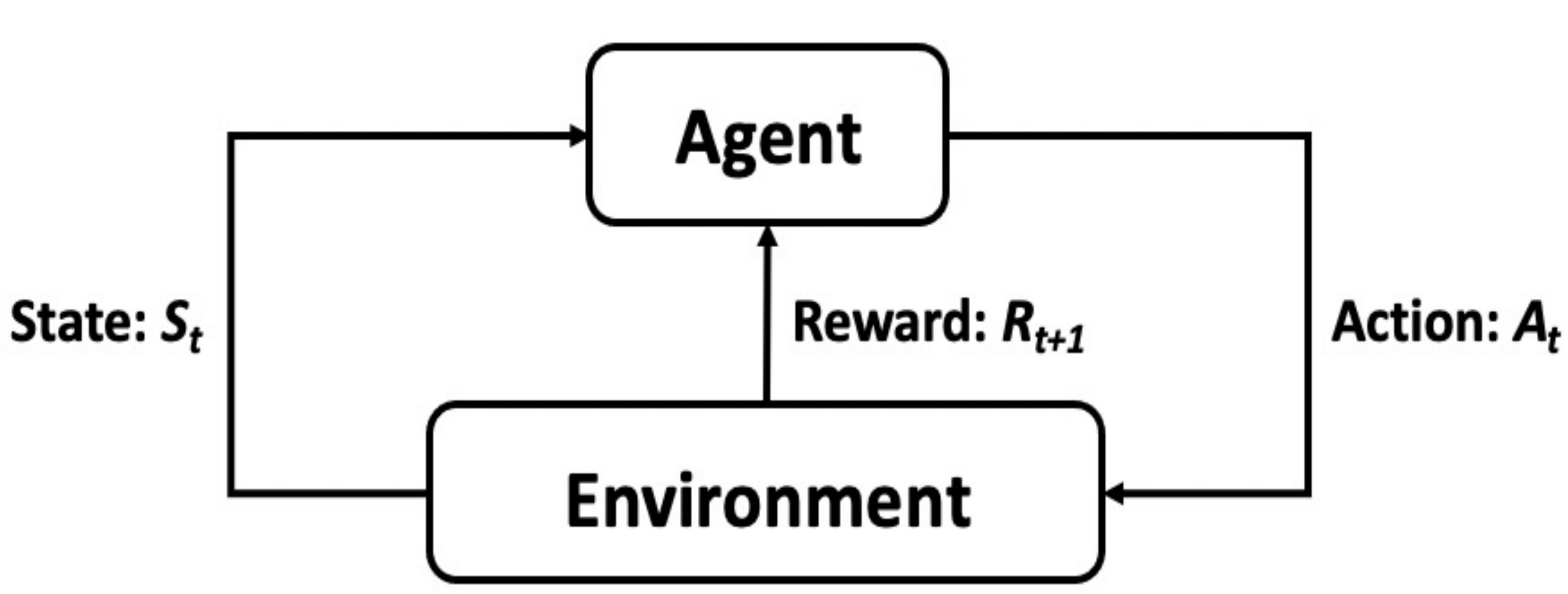
Dynamic Online Learning: Utilizes **online reinforcement learning** to adapt cache management to varying workloads and system configurations

Multiple Program Features: Employs **multiple program features** to achieve a thorough understanding of memory access patterns

Concurrency-Aware Rewards: Implements a reward system that is **aware of concurrent accesses**, factoring in system-level feedback for decision evaluation

Efficient Design: **Minimal** hardware overhead

Reinforcement Learning (RL)



Formulating Cache Management as an RL Problem

State: A vector of features for each access

S = (PC, page number)

Using PC signature to distinguish between **demand accesses** and **prefetch accesses**

Action: Using EPV to reflect the eviction priorities of the cache block

Cache miss (4 optional actions):

- Bypass LLC
- Insert the corresponding block in LLC with an EPV of low, moderate, or high

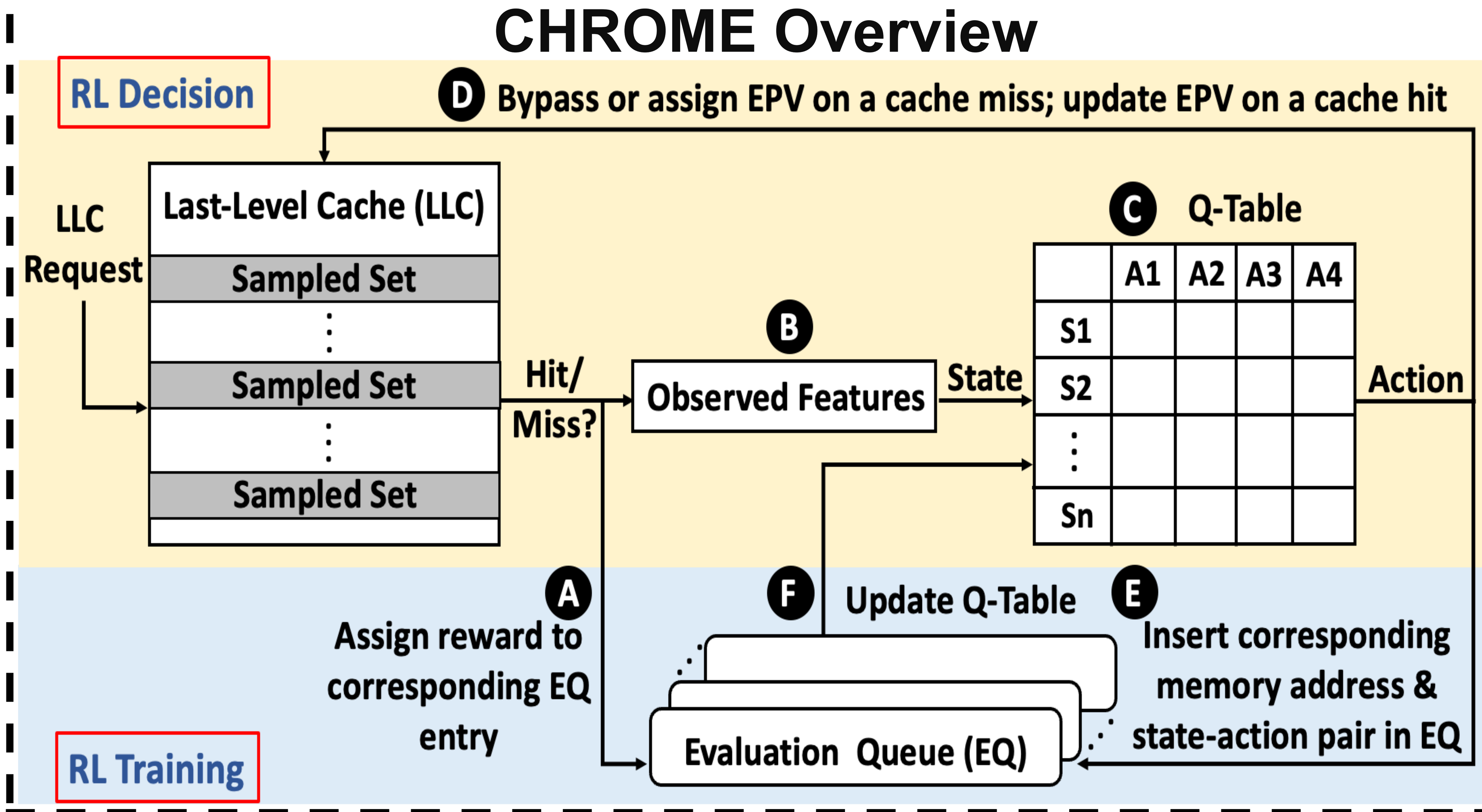
Cache hit (3 optional actions):

- Update the EPV of the corresponding block to low, moderate, or high

Reward: Considering

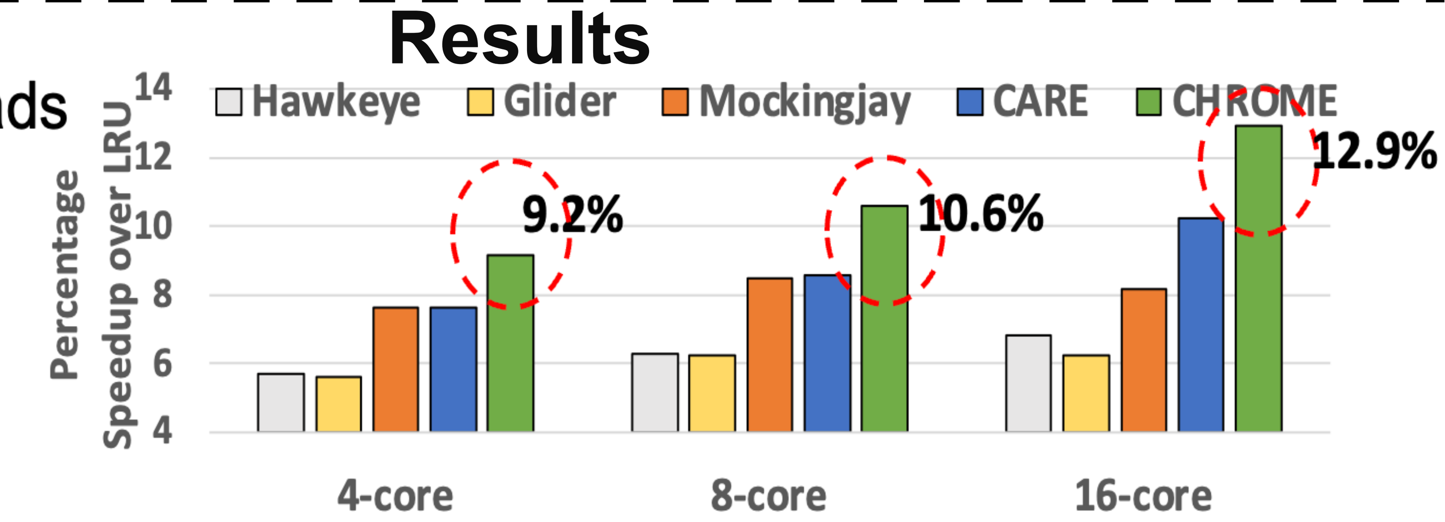
- **Accuracy** of each action
- Distinguish between actions triggered by **demand** or **prefetching**
- **Concurrency-Aware System Feedback**

CHROME Overview

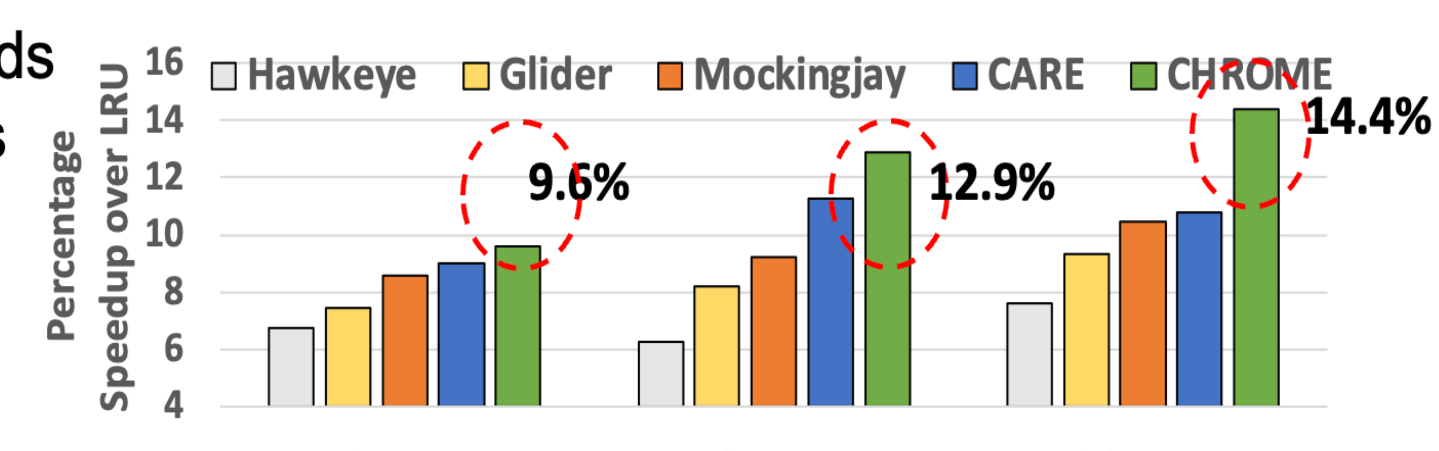


Results

SPEC workloads homogeneous



SPEC workloads heterogeneous



- **The holistic view provides a performance guarantee**
- **Online RL provides good adaptability and scalability**
- **CHROME can accurately provide cache management for different workloads**
- **CHROME outperforms all other schemes across all system configurations**
- **Performance advantage of CHROME over others increases with more cores**

U.S. National Science Foundation
ILLINOIS TECH
FIU FLORIDA INTERNATIONAL UNIVERSITY