INDIAN INSTITUTE OF TECHNOLOGY
HYDERABAD, TELANGANA

EXPLORING CACHE REPLACEMENT POLICIES

A COMPREHENSIVE REVIEW & INNOVATIVE STRATEGY

TERM PAPER

Submitted By:
SHIV TIKOO CS23BTBNRK11001

Submitted To:
Department of Computer Science and Engineering
Indian Institute of Technology
Hyderabad, Telangana

DECEMBER, 2023

DECLARATION

I hereby declare that this term paper titled Exploring Cache Replacement Policies: A Comprehensive Review and Innovative Strategy is my original work. All sources used in this paper have been properly acknowledged and referenced. I have adhered to ethical standards in conducting and presenting this research.

This paper was undertaken as a requirement for the completion of my Advance Computer Architecture Course at Indian Institute of Technology, Hyderabad. The survey for this paper was carried out during the months of October and November, 2023.

I would like to express my gratitude to my course supervisors Dr. Shirshendu Das, Assistant Professor and Rajesh Kedia, Assistant Professor .

I declare that this paper represents my own work, and any assistance received from others has been acknowledged and appropriately referenced.

Shiv Tikoo (CS23BTNRK11001) _____

DATE: DECEMBER, 2023

TABLE OF CONTENTS

1. ABSTRACT

Cache replacement policies play a crucial role in optimizing memory hierarchies to enhance system performance in modern computing architectures. This survey report investigates the evolutionary trajectory of cache replacement strategies by exploring seminal studies and recent advancements in the field. The survey meticulously analyzes six key papers, each presenting novel paradigms and advancements in cache management strategies.

The survey commences by revisiting foundational principles, elucidating the significance of cache replacement policies in mitigating cache pollution and improving overall system efficiency. It navigates through recent innovations, notably focusing on specialized policies such as the Expected Hit Count (EHC) and Deadblock Aware Adaptive Eviction Policy (DAAEP). The analysis showcases EHC's predictive prowess in estimating cache block usefulness and DAAEP's adaptive allocation based on an application's cache-friendliness.

Furthermore, the survey delves into pioneering evaluation metrics, such as Improvement Per Kilo Byte (IPKB), elucidating its role in comprehensively assessing replacement policies. IPKB's emphasis on balancing miss rate improvements against hardware overhead represents a paradigm shift in policy evaluation.

Drawing insights from these diverse approaches, the survey culminates in proposing a novel hybrid cache replacement policy, termed EHC-DAAEP. This hybrid policy aims to synergistically combine the predictive power of EHC and the contextual adaptation of DAAEP. By leveraging Expected Hit Count and Deadblock Rate metrics, the proposed EHC-DAAEP policy seeks to optimize cache utilization by employing a hybrid eviction strategy.

The report underscores the paradigm shift from traditional replacement policies towards multifaceted, context-aware strategies. It highlights the evolution from singular-factor considerations to hybrid models, emphasizing the need for adaptive and application-specific cache management solutions. The proposed EHC-DAAEP policy signifies the contemporary pursuit of predictive, adaptive, and application-aware cache replacement strategies, shaping the landscape of memory hierarchies in modern computing architectures.

2. INTRODUCTION

As a part of the Advance Computer Architecture course, we have entered the realm of contemporary computing. The optimization of cache memory management serves as a linchpin in augmenting overall system performance and efficiency. The intricate orchestration of cache replacement policies stands as a pivotal aspect in this pursuit, wielding substantial influence in shaping computational efficacy. This report, titled **Exploring Cache Replacement Policies: A Comprehensive Review & Innovative Strategy**, endeavors to embark on an extensive exploration of these policies. Through an in-depth analysis of their evolutionary trajectory, efficacy, and potential for transformative innovation, this report endeavors to illuminate the impact of novel strategies.

Setting the foundation for this rigorous investigation, the subsequent sub sections delineate the specific objectives, significance, and extensive scope of this report. It underlines the critical importance of cache replacement policies and outlines the focused goals that propel this comprehensive analysis.

2.1. MOTIVATION

The optimization of cache memory management stands as a linchpin in improving the overall performance of modern computing architectures. This survey report is a culmination of the advanced computer architecture course, aiming to conduct a comprehensive exploration of cache replacement policies. It meticulously examines their evolution, efficacy, and recent advancements.

Cache memory plays a pivotal role in reducing memory access latency, and any inefficiency resulting in a cache miss can lead to significant delays, often spanning hundreds of cycles in modern processors. Memory-intensive workloads operate on massive amounts of data that cannot be captured by last-level caches (LLCs) of modern processors.Consequently, processors encounter frequent off-chip misses, and hence, lose significant performance potential. One of the components of a modern processor that has a prominent influence on the off-chip miss traffic is LLC's replacement policy. Existing processors employ a variation of least recently used (LRU) policy to determine the victim for replacement. Unfortunately, there is a large gap between what LRU offers and that of Belady's MIN, which is the optimal replacement policy. Belady's MIN requires selecting a victim with the longest reuse distance, and hence, is unfeasible due to the need for knowing the future. This underscores the paramount importance of robust cache replacement policies in minimizing cache misses and optimizing

computational efficiency.

Through a meticulous analysis, this report endeavors to shed light on the transformative potential of novel replacement policies. It seeks to offer valuable insights into enhancing system efficiency and computational performance by leveraging innovative cache management strategies.

Furthermore, within the academic context, this survey report serves as the term paper for the Advanced Computer Architecture course. Its comprehensive analysis and critical evaluation of cache replacement policies aim to demonstrate a deep understanding of the intricacies involved in modern computing systems, aligning with the course objectives and academic rigor.

2.2. SCOPE

This survey report undertakes a comprehensive exploration of established and emerging cache replacement policies, aiming to dissect their implications on critical performance metrics such as cache hit rates, memory access latencies, and overall system responsiveness. The foundational research for this survey draws from specific papers provided during the course, notably focusing on innovative approaches, including the dynamic insertion policy, reference interval prediction replacement policy, and utility-based cache partitioning. These seminal papers have ignited a profound interest in this specialized domain of computer architecture, serving as the cornerstone for this comprehensive survey.

Building upon the foundation laid by these seminal papers, this report aims to extend its exploration by surveying and analyzing additional novel approaches that derive inspiration from or build upon the concepts introduced in these seminal works. The survey will encompass a meticulous examination of newer approaches that have embraced and expanded upon these foundational ideas, providing a holistic understanding of the evolving landscape of cache replacement policies.

Moreover, this report will delve into an in-depth comparative analysis of these newer approaches. By juxtaposing their strengths, weaknesses, and practical implications, this comparative study aims to offer a comprehensive overview of the diverse spectrum of cache management strategies, highlighting the advancements and potential limitations in contemporary research.

Building on that, the scope extends beyond the survey and comparative analysis to introduce a novel approach derived from a synthesis of insights gleaned from these approaches.

Leveraging the comprehensive understanding garnered from the surveyed literature, this report will propose and delineate a novel cache replacement policy. This proposed approach aims to integrate the knowledge and innovations gleaned from the surveyed methodologies, offering a unique perspective on cache management strategies.

2.3. IMPORTANCE

Cache replacement policies play an instrumental role in optimizing system performance by efficiently managing cache memory utilization. The effective implementation of these policies directly influences crucial performance metrics and is pivotal in reducing memory access latencies, thereby enhancing overall system responsiveness across diverse computing environments.

Efficient cache management directly translates to minimized memory access bottlenecks, significantly impacting system performance. The careful selection and implementation of replacement policies have a profound effect on cache hit rates and resource utilization within computing systems. A well-tailored replacement policy effectively mitigates cache misses, ensuring that frequently accessed data remains readily available in the cache, thus reducing the need for accessing slower main memory.

This subsection aims to underscore the paramount significance of replacement policies in optimizing computational processes. By scrutinizing the advancements and nuances within these policies, this report aims to illuminate their pivotal role in mitigating memory access latencies, enhancing cache hit rates, and ultimately bolstering the overall efficiency and responsiveness of computing systems.

2.4. OVERVIEW

The report is structured into several sections, each addressing distinct aspects crucial to understanding cache replacement policies and their advancements. These sections include:

- **BACKGROUND** In this section, the foundational papers provided in the course serve as the focal point of discussion. A comprehensive analysis and discussion of these seminal papers will be undertaken, elucidating their contributions and significance in the realm of cache replacement policies. This foundational understanding will pave the way for a deeper exploration of subsequent sections.

- **SURVEY** The crux of this report lies within this section. Building upon the foundational papers, recent novel approaches that cite these foundational works will be scrutinized. This section entails a meticulous examination and comparison of these contemporary approaches, extracting valuable insights and identifying key advancements in cache replacement policies derived from each approach.

- **LEARNINGS** Here, the collective insights garnered from the research conducted on various papers will be synthesized and presented. This section aims to dissect and consolidate the key learnings, trends, and patterns identified from the surveyed literature. The culmination of these learnings will lay the groundwork for the subsequent proposed idea section.

- **PROPOSED IDEA** The final segment of the report revolves around the proposition of a novel approach. Based on the comprehensive analysis and learnings obtained from the surveyed papers, this section will introduce and outline a unique and innovative cache replacement policy. The proposed idea aims to integrate and build upon the insights gleaned from the surveyed approaches, offering a distinctive perspective in cache management strategies.

3. BACKGROUND

The landscape of cache replacement policies has witnessed significant advancements in the recent years, revolutionizing the efficiency and performance of modern computing systems. This section delves into the exploration of the foundational papers, which are the papers provided to us in our course that have shaped and redefined cache management strategies, addressing crucial challenges and introducing novel ideas.

3.1. DYNAMIC INSERTION POLICY

This was the first paper which introduced us to the advancements in replacement policy and how the traditional LRU policy can be moidified to handle certain application requirements more efficiently.

- **Problem Addressed**
  Explores the challenge of cache replacement inefficiency for memory-intensive workloads surpassing cache size thresholds, resulting in thrashing under the traditional LRU policy. Discusses the detrimental impact on cache performance and system responsiveness.

- **Proposed Solution & Evaluation**
  Details the innovative insertion policies: LRU Insertion Policy (LIP), Bimodal Insertion Policy (BIP), and Dynamic Insertion Policy (DIP), emphasizing their roles in improving cache efficiency. Provides a comprehensive evaluation using SPEC CPU2000 benchmarks, highlighting DIP's substantial reduction in misses per 1000 instructions (MPKI) by 21.3% and its minimal hardware overhead. The below Table 05 shows the same in a more detailed manner.

Table 5: Comparison of replacement policies

Replacement Policy	%Reduction in MPKI over LRU	Hardware Overhead
SBAR (LRU + MRU-Repl)	8.8	2 kB
SBAR (LRU + NMRU-mid)	5.1	2 kB
SBAR (LRU + Rand)	8.9	2 kB
SBAR (LRU + RLRU-Skew)	6.6	2 kB
SBAR (LRU + RMRU-Skew)	11.3	2 kB
SBAR (LRU + LFU)	14.7	12 kB
DIP	**21.3**	**2 B**
Belady's OPT	32.2	N/A

- **Contribution & Significance**
  The seminal contribution of introducing insertion policies that significantly enhance cache performance for memory-intensive workloads, emphasizing their simplicity and efficacy in mitigating cache thrashing.

3.2. THREAD AWARE DYNAMIC INSERTION POLICY

Further improvements were made in DIP by improving its dynamic nature and making it aware of the nature of the thread to check which policy would be more benefitial for the respected application.

- **Problem Addressed**
  Explores the innovative technique of dynamically managing cache resources among multiple applications within a chip multiprocessor (CMP) environment. Highlights the challenge of traditional insertion policies failing to account for individual application characteristics in shared cache environments.
- **Proposed Solution & Evaluation**
  Provides an extensive performance evaluation across various core systems and workload mixes, showcasing TADIP's improvements in throughput, speedup, and fairness metrics compared to the baseline LRU policy.
- **Contribution & Significance**
  Emphasizes the novel extension of adaptive insertion policies to shared caches, introducing TADIP, which adapts to individual application characteristics and outperforms previous schemes.

3.3. RE REFERENCE INTERVAL PREDICTION

This paper takes up a novel approach by bringing up the idea of assigning blocks a re reference period value rather than least recently used and most recently used tags.

- **Problem Addressed**
  Investigates the inefficiencies of traditional LRU policies in handling mixed access patterns, especially for blocks with distant re-reference intervals. Emphasizes the limitations of existing policies in distinguishing between various access patterns.

- **Proposed Solution & Evaluation**

  Explores the architecture and design of Re-reference Interval Prediction (RRIP) policies - Static RRIP (SRRIP) and Dynamic RRIP (DRRIP), highlighting their minimal storage requirements and efficacy in improving cache performance.The main idea behind the scheme is to evict the highest RRPV block and promote a block upon cache hit by making its value 0. The paper provides experimental evidence showcasing RRIP's outperformance of LRU and other state-of-the-art algorithms on multi-core systems across diverse workloads.

- **Contribution & Significance**

  The significance of RRIP in mitigating the limitations of traditional policies, emphasizing its scalability, minimal hardware requirements, and superiority in handling mixed access patterns.The below Figure 10 shows a graph drawing comparison between our earlier discussed work DIP and RRIP.

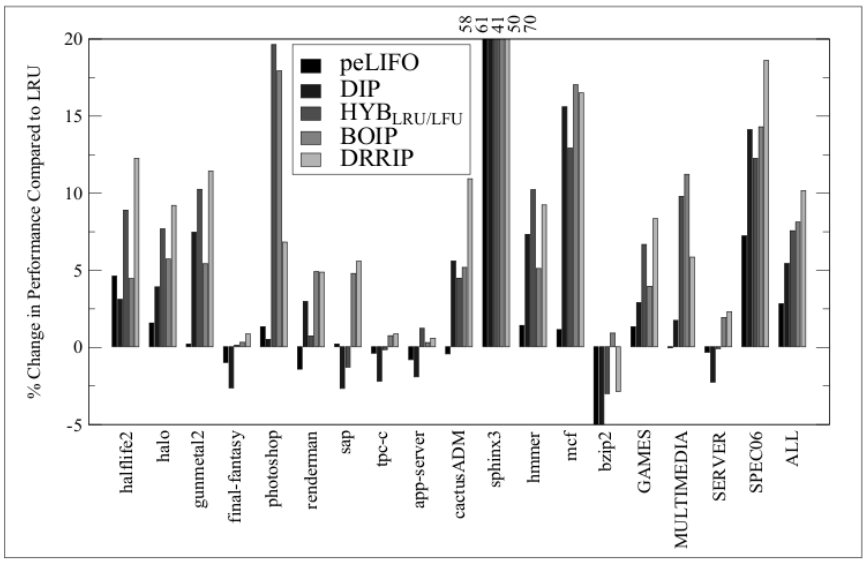


Figure 10: Comparison of Replacement Policies.

3.4. UTILITY BASED CACHE PARTITIONING

Moving from modifications in replacement policies, this paper introduced us to the idea of cache partitioning.

- **Problem Addressed**

  Investigates cache partitioning based on utility rather than access demand, highlighting the limitations of conventional demand-based partitioning schemes in optimizing cache utilization.

- **Proposed Solution & Evaluation**
  The paper introduces the Utility-Based Cache Partitioning (UCP) scheme, showcasing its utility monitor (UMON) and performance enhancements over LRU on multiprogrammed workloads.

Building upon these groundbreaking concepts, the subsequent section embarks on a survey of recent novel approaches that have drawn inspiration from and cited these seminal works. These contemporary advancements further enrich our understanding of cache management strategies, showcasing innovative adaptations and extensions derived from these foundational studies.

4. SURVEY

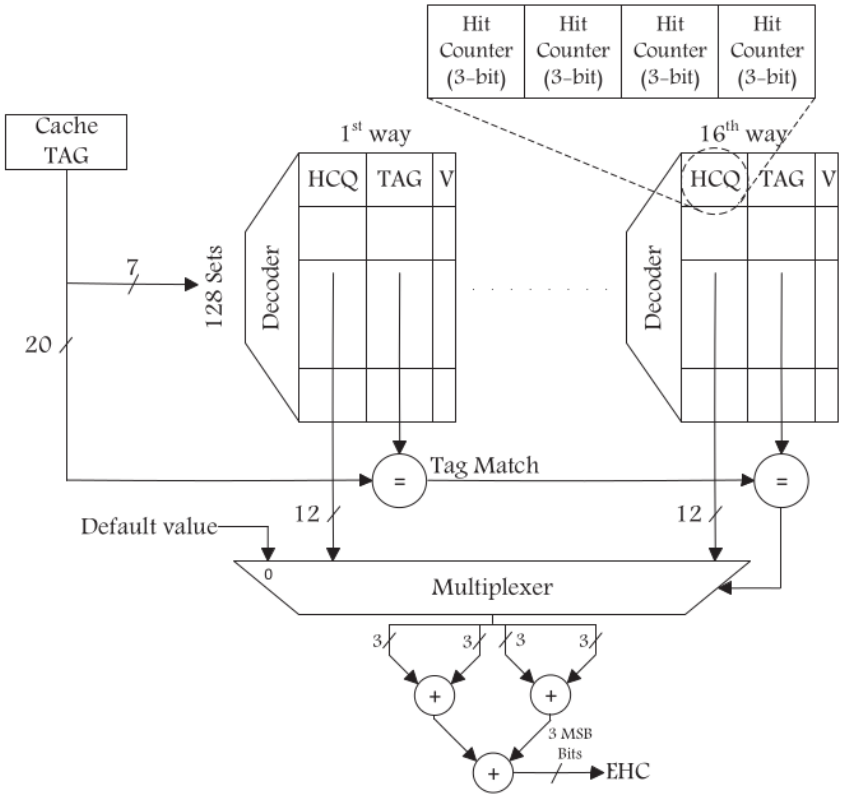
As seen in the previous section we have established a robust foundation by examining seminal works in cache replacement policies, this survey embarks on an exploration of six recent papers that have cited and expanded upon the foundational studies. These selected papers,follow the chronology based on their publication dates, represent evolving paradigms and innovations in cache management strategies. Each paper builds upon the principles and concepts elucidated in the foundational studies, contributing novel perspectives and enhancements to the cache replacement domain.

4.1. EXPECTED HIT COUNT POLICY

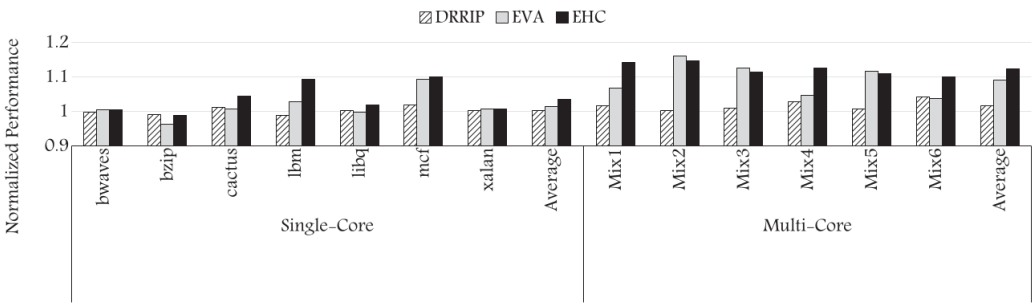
The paper titled ”Cache Replacement Policy on expected Hit count” presents a pioneering cache replacement policy, the idea behind it is to build a deadblock predictor using the average of the 4 most recently used blocks with the tag address. It is able to predict woth an accuracy of 69% using this logic.

4.1.1. COMPARISON

The HCR policy extends the concept of adaptive insertion policies, it works on the logic that the expected hit count is inversely proportional to the reuse distance. It is able to predict the hit count with an accuracy of 69%. The functionality can be better understood by understanding the structure of the Hit History Table. The below figure represents the structure of the same.



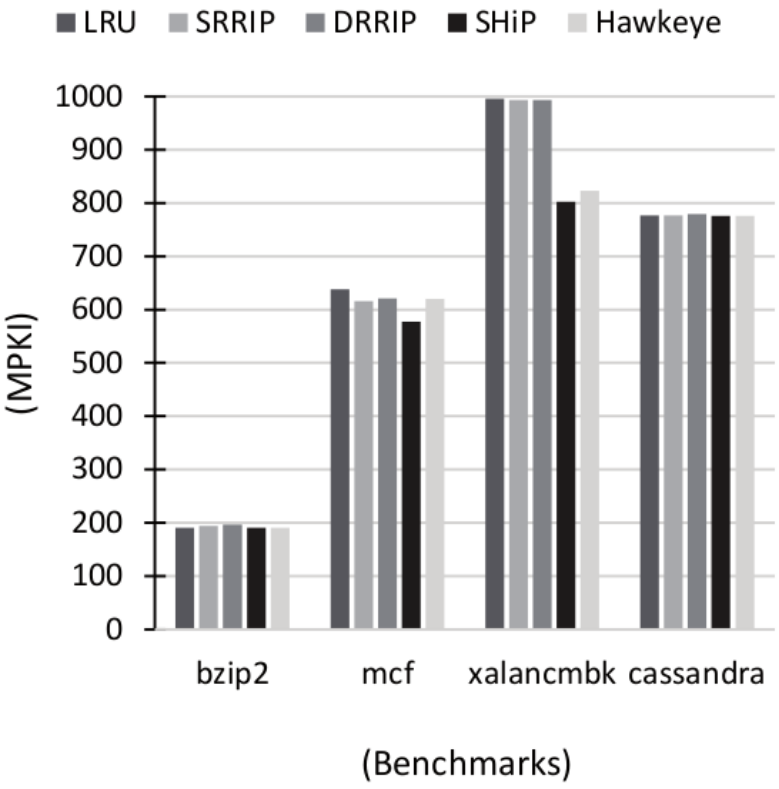
It is a data structure to store the EHC values of cache lines. It is organized as a set-associative table with a fixed number of entries per set. Each entry consists of a tag, an EHC value, and a valid bit. The HHT is updated on every cache access (hit or miss) using a simple update algorithm. This demonstrates a clear evolution from the foundational principles towards more specialized and nuanced approaches in cache replacement strategies, emphasizing factors beyond mere access patterns and recency. Its' performance can be seen in comparison to other common replacement policies in the figure below.

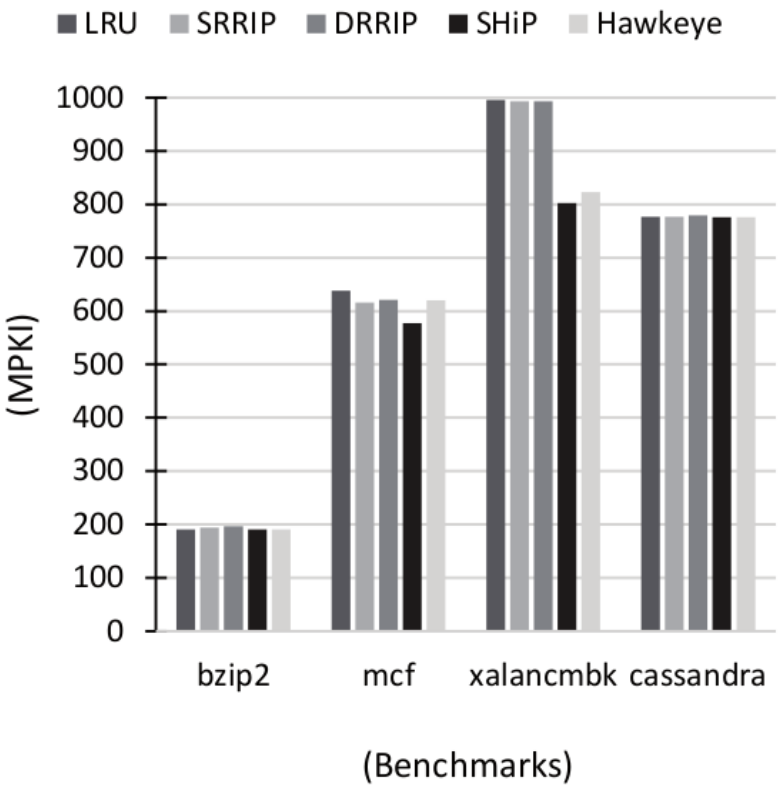


4.2. IMPROVEMENT PER KILO BYTE METRIC

Stepping into the world of replacement policies, it is a necessary step to understand how well a policy works; this paper introduces us to a metric to do exactly the same. The paper introduces a novel evaluation metric, Improvement Per Kilo Byte (IPKB), aimed at comprehensively assessing cache replacement policies based on miss rate improvement and hardware overhead. Existing metrics like Miss per Kilo Instruction (MPKI) or Instruction Per Cycle (IPC) fall short as they neglect the crucial consideration of hardware costs associated with implementing intricate replacement policies.

The IPKB metric offers a holistic view by comparing five prevalent replacement policies: LRU, SRRIP, DRRIP, SHiP, and Hawkeye. It formulates the IPKB metric to quantify miss rate improvement per kilobyte of additional hardware overhead in the cache structure.This helps us understand that the policy with the lowest MPKI is not necessarily the best policy. The below graphs can be used to see a comparison of the two metrics.





The IPKB formula elucidates the calculation of miss rate improvement over LRU concerning the hardware overhead difference between LRU and the policy under evaluation. It normalizes the improvement by considering the trade-off between hardware overhead and miss rate enhancement. Additionally, the paper introduces a weighted IPKB variation to accommodate scenarios where the evaluated policy exhibits lower hardware overhead than LRU. The figure below provides the weighted IPKB formulation, where m_improvement stands for improvement in the MPKI value compared to LRU.

$$IPKB = \frac{LRU\ overhead}{new\ hardware\ overhead} \times m\_improvement$$

Hardware overhead plays a pivotal role in cache design, influencing power consumption, area, and overall complexity. The paper emphasizes the significance of this factor in evaluating replacement policies. It elucidates that a policy with lower miss rates might not necessarily be ideal if it incurs substantial hardware overhead. IPKB encapsulates this trade-off by providing a normalized assessment that factors in hardware costs.

4.2.1. COMPARISON

The IPKB metric extends the evaluation criteria beyond traditional miss rate metrics like MPKI, aligning with the foundational studies' emphasis on enhancing cache efficiency. By integrating hardware overhead considerations, it reflects a more comprehensive and nuanced evaluation approach, aligning with the pursuit of optimal cache management strategies discussed in the foundational works. This delineates a shift towards a more comprehensive evaluation paradigm that transcends mere miss rate improvements, aligning with the foundational papers' aspirations to refine cache management strategies holistically.

4.3. PV AWARE REPLACEMENT POLICY

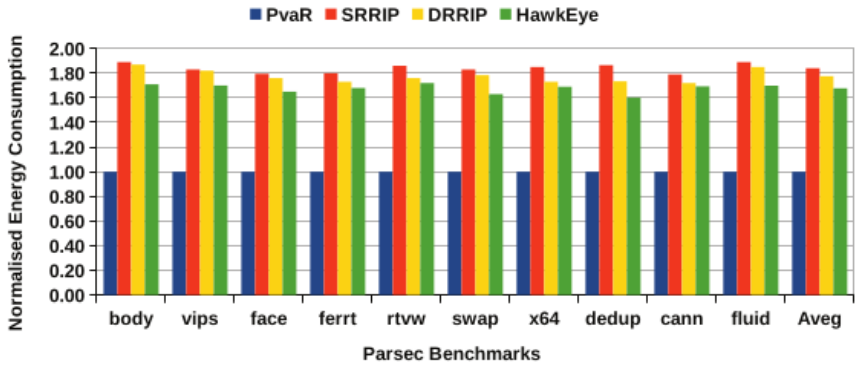
This paper introduces a pioneering cache replacement policy designed for the L3 SRAM cache within a 3D TCMP architecture, featuring the L4 DRAM cache functioning as the Last Level Cache (LLC) and susceptible to process variation (PV).

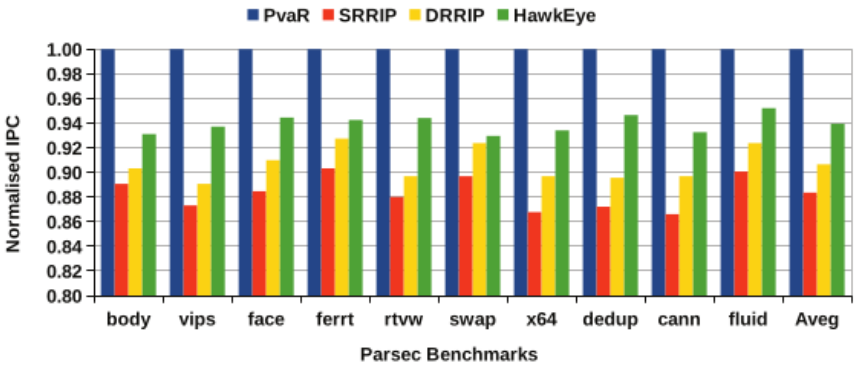
The paper elucidates how process variation impacts different L4 banks, causing variations in latency and energy consumption. Moreover, it highlights the correlation between diverse cache access patterns and the potential benefits derived from employing specific replacement policies, emphasizing the need for a tailored approach due to PV-induced disparities.

The proposed PV-aware Replacement (PVaR) policy fundamentally considers the health status of L4 banks to optimize cache utilization. PVaR strategically favors retaining blocks from less healthy L4 banks within the L3 cache, thereby reducing accesses to the L4 cache and in turn helps improving the performance of the replacement policies. This approach aims to mitigate the adverse effects of PV on performance and energy consumption by intelligently managing cache block migrations between L3 and L4 caches. PVaR uses a predicted re-reference value (RRPV) to estimate the reuse interval of each block in L3. PVaR also uses a chance counter to indicate how many extra chances a block from an unhealthy L4 bank should get to stay in L3. The chance counter is initialized based on the health index of the block's home bank in L4. PVaR selects a victim block based on its RRPV and chance counter values.

The paper showcases substantial performance enhancements compared to prevalent policies. PVaR demonstrates a remarkable performance boost, surpassing SRRIP, DRRIP, and HawkEye by 13%, 10%, and 6% respectively. Additionally, it significantly curtails dynamic energy consumption by 46%, 44%, and 40% compared to the aforementioned policies, further validating its efficacy in optimizing energy efficiency.The same can be seen in the graphs

provided below, the first graph shows the energy comparison while the second graph shows the ipc comparison.





4.3.1. COMPARISON

PVaR's tailored approach aligns with the foundational studies' exploration of adaptive cache management strategies. It addresses specific architecture nuances, akin to the emphasis on dynamic cache adaptation in shared environments discussed in prior foundational works. The focus on mitigating PV-induced disparities reflects a concerted effort towards more sophisticated cache policies, in line with the foundational papers' quest for enhanced cache performance in complex computing environments. This delineates a notable evolution from generic cache replacement policies to more nuanced, context-aware strategies, echoing the foundational papers' call for adaptive and tailored cache management solutions. .

4.4. RECENCY TIME RE REFERENCE INTERVAL PREDICTION

This paper targets the optimization of cache replacement policies by integrating both recency and frequency of data accesses.

The primary focus revolves around introducing a novel cache replacement policy termed Recency Time Re-Reference Interval Prediction (RT-RRIP). RT-RRIP is designed to amalgamate two crucial aspects of data access—recency (recent access time of a data item) and re-reference (the anticipated time for a data item to be accessed again). This fusion aims to enable more informed decisions regarding cache block retention or replacement.

The RT-RRIP policy operates through a dual-component mechanism, incorporating a recency time prefilter alongside a re-reference interval prediction model. This technique filters out cache blocks that have been recently accessed and are likely to be accessed again soon. It uses a threshold based on the average recency time of all blocks to select a subset of blocks for eviction. The recency time prefilter acts by filtering out cache blocks exhibiting low recency time, whereas the re-reference interval prediction model prognosticates the future re-reference intervals of cache blocks. This symbiotic relationship between the two components significantly curtails cache misses and miss latency, thereby augmenting cache performance.

The paper substantiates its claims with simulations employing SPEC CPU2006 benchmarks. The results affirm the superiority of the RT-RRIP policy by showcasing reduced cache misses and miss latency in comparison to prevalent policies like Least Recently Used (LRU), Least Frequently Used (LFU), and Re-Reference Interval Prediction (RRIP). The results can be seen below, which shows the miss rate in L1 cache.

	16 2	8 4	16 4	4 8	8 8	16 8	4 16	8 16	16 16	4 32
RTRRIP	0.0757	0.0689	0.0686	0.0426	0.0389	0.0382	0.011	0.0024	0.0013	0.0008
RRIP	0.0767	0.0661	0.0663	0.0407	0.0408	0.0402	0.0112	0.0046	0.0033	0.0015
LRU	0.0759	0.0718	0.0721	0.0476	0.0462	0.0513	0.0125	0.0035	0.0013	0.0008
LFU	0.1433	0.0834	0.1019	0.0698	0.0717	0.0761	0.0376	0.0406	0.0445	0.0063

The significance of this paper lies in its proposition of a hybrid approach that amalgamates recency and re-reference intervals. This novel fusion has the potential to revolutionize cache replacement policies, promising enhanced system performance and energy efficiency. The integration of these dual aspects presents a pathway towards more sophisticated and efficient cache utilization.

4.4.1. COMPARISON

The integration of both recency and re-reference intervals aligns with the foundational papers' exploration of adaptive cache management strategies. While prior works primarily focused on singular aspects, such as recency or re-reference intervals, this paper extends the horizon by combining both. This underscores an evolution towards more comprehensive cache replacement strategies that consider multiple facets of data access dynamics, echoing the foundational studies' pursuit of adaptive cache management solutions.

4.5. DEADBLOCK AWARE ADAPTIVE EVICTION POLICY

The paper addresses the challenge of managing shared Last-Level Cache (LLC) in multicore systems, where varied application types often impede cache-friendly applications, resulting in performance degradation.

In these multicore systems, diverse applications interfere within the shared LLC, impacting cache performance. Deadblock Aware Adaptive Eviction Policy (DAAEP) aims to optimize LLC space allocation by utilizing the Deadblock Rate (DR) of each application. Here, a block is labeled "dead" if it won't be accessed in the future. The DR of an application is computed as the ratio of dead blocks to the total number of blocks it brought into the cache. DAAEP strategically evicts blocks from streaming applications with a high DR, making way for cache-friendly applications with lower DR. The policy dynamically adjusts the DR threshold for eviction based on the real-time cache behavior of running applications.

The evaluation of DAAEP employs ChampSim, a trace-based simulator modeling out-of-order cores. Using 14 memory-intensive SPEC CPU2006 applications exhibiting diverse behaviors, the paper illustrates that DAAEP enhances weighted speedup by 7.5% and 0.9% over Static Re-Reference Interval Prediction (SRRIP) and Deadblock Aware Adaptive Insertion Policy (DAAIP), respectively. These results underscore the efficacy of DAAEP in managing shared LLC in multicore systems. The same can be observed in the graph below.

The paper concludes by emphasizing the performance benefits of DAAEP in mitigating cache interference, thus improving system performance. The authors also acknowledge the funding support for their research.

4.5.1. COMPARISON

This paper further advances the domain of adaptive cache replacement policies outlined in foundational studies. While earlier works explored adaptive policies, this paper specifically targets shared LLC in multicore systems, a critical concern in modern computing. The emphasis on DR-based eviction strategies adds a nuanced dimension to cache management, aligning with the foundational studies' call for more intricate approaches to handle diverse applications' interference within shared caches.

This emphasizes the evolution toward more context-aware cache replacement strategies, aligning with the foundational studies' intent to address modern computing challenges through adaptive cache management solutions.

4.6. EMISSARY

The paper introduces EMISSARY, a novel cache replacement policy specifically designed for L2 instruction caching, focusing on prioritizing lines whose misses lead to decode starvations. Based on the insight that modern processors can withstand numerous instruction cache misses without performance degradation if they don't starve the decode stage, EMISSARY—short for Enhanced MISS-Awareness Replacement Policy—is devised.

EMISSARY adopts a bimodal policy that categorizes cache lines as high or low priority, depending on whether their misses cause decode starvation. High-priority lines, tagged with a priority bit, are safeguarded from eviction by up to N ways per set in the L2 cache. Impressively, EMISSARY necessitates minimal hardware modifications—merely two bits per line—without involving historical tracking, coordination with prefetchers, predictions, or complex calculations.

The performance evaluation demonstrates EMISSARY's remarkable accomplishments. It achieves a geomean speedup of 3.24% and energy savings of 2.12% compared to the baseline policy (TPLRU) across 13 server applications with extensive code footprints. Additionally, it outperforms various cache replacement policies, including both cost-aware and cost-unaware ones. Furthermore, EMISSARY demonstrates a substantial reduction in the L2 instruction MPKI by 13.55% and the L2 data MPKI by 8.8%, indicating a more balanced utilization of cache resources between instructions and data. Notably, EMISSARY achieves a significant portion (21.6%) of the speedup achievable by an unattainable L2 cache with zero-cycle miss latency for all capacity and conflict instruction misses.

EMISSARY's notable contributions are marked by being the first cost-aware cache replacement policy explicitly tailored for instruction caching. Leveraging existing signals of decode starvation and issue queue emptiness, it effectively identifies costly instruction misses and prioritizes them in the cache. Simple, effective, and energy-efficient, EMISSARY seamlessly integrates into existing processor designs, leaving scope for further enhancements through amalgamation with other techniques like dead-block prediction, profile-guided optimization, and specialized caching solutions. Moreover, its potential extension to different cache hierarchy levels or other cache types is also highlighted.

Moving forward, the ensuing section explores a proposed idea for enhancing cache management strategies, building upon the insights gleaned from the surveyed papers and foundational studies.

5. PROPOSED IDEA

The objective of our proposed idea is to design a hybrid cache replacement policy, coined as EHC-DAAEP, synergizing the strengths of the Expected Hit Count (EHC) policy and the Deadblock Aware Adaptive Eviction Policy (DAAEP).

5.1. MOTIVATION

Our motivation for proposing this hybrid policy stems from the distinct advantages offered by both EHC and DAAEP:

1. **EHC's Advantage**: EHC, surpassing conventional policies like Static Re-Reference Interval Prediction (SRRIP), leverages the Expected Hit Count metric, enabling informed decisions regarding block eviction based on anticipated usefulness.

2. **DAAEP's Strength**: DAAEP, rooted in SRRIP, effectively employs the Deadblock Rate to differentiate cache-friendly applications from streaming ones, optimizing cache space allocation.

The proposed fusion aims to harness the predictive capabilities of EHC and the application-aware eviction strategy of DAAEP for enhanced cache management. The idea to move forward with this makes sense because DAAEP is based on the SRRIP and when EHC was coined it had been put up against SRRIP and provided better results against it.

5.2. METHODOLOGY

The EHC-DAAEP policy will amalgamate these strategies as follows:

1. **Expected Hit Count Computation**: Similar to EHC, the policy will calculate the Expected Hit Count for each cache block, estimating their future utility based on access patterns.

2. **Deadblock Rate Evaluation**: Like DAAEP, the policy will assess the Deadblock Rate for individual applications, distinguishing cache-friendly applications from those causing cache contention.

3. **Hybrid Eviction Approach**: Our policy's eviction strategy will consider both the Ex-

pected Hit Count of blocks and the Deadblock Rate of applications. Blocks with low Expected Hit Count and belonging to applications with high Deadblock Rates will be prioritized for eviction, optimizing cache space allocation.

5.3. EXPECTED BENEFITS & FUTURE WORKS

The amalgamation of EHC and DAAEP techniques is anticipated to yield several advantages:

- Enhanced Cache Performance: EHC-DAAEP is expected to reduce cache miss rates and elevate system performance by selectively evicting less valuable blocks and favoring cache-friendly applications. - Optimized Cache Utilization: The hybrid policy aims to strike a balance between predictive block retention and application-aware eviction, leading to more efficient cache utilization.

The proposed idea will be realized through the following future steps:

- Implementation and Evaluation: Implementing the EHC-DAAEP policy and evaluating its performance using suitable cache simulation tools and comprehensive benchmark suites. - Comparative Analysis: Comparative assessment of EHC-DAAEP against EHC, DAAEP, and other established cache replacement policies to ascertain its efficiency and superiority in diverse computing scenarios.

6. CONCLUSION

The survey embarked on a comprehensive exploration of diverse cache replacement policies, spanning seminal works and recent advancements, to unravel evolving paradigms and innovations in cache management strategies. Through an in-depth analysis of six selected papers, each building upon foundational studies, this survey delineated the progression and enhancements in cache replacement policies.

The survey's notable findings encompass a spectrum of innovative strategies and evaluation metrics, illuminating the domain's evolution:

1. **Emergence of Specialized Policies**: The survey unveiled the evolution from traditional replacement policies to specialized ones, such as the Expected Hit Count (EHC) and Dead-block Aware Adaptive Eviction Policy (DAAEP), emphasizing tailored and application-aware cache management strategies.

2. **Comprehensive Evaluation Metrics**: The introduction of metrics like Improvement Per Kilo Byte (IPKB) and hybrid evaluation paradigms like EHC-DAAEP underscored the shift towards more nuanced, holistic, and hardware-conscious assessment criteria for cache replacement policies.

3. **Contextual Adaptation and Prediction**: The progression from singular factors like recency or re-reference intervals to hybrid models like Recency Time Re-Reference Interval Prediction (RT-RRIP) indicated a move towards sophisticated predictive models, reflecting a contextual adaptation in cache policies.

4. **Multifaceted Cache Utilization**: Strategies like PV-aware Replacement (PVaR) and DAAEP showcased an inclination towards multifaceted cache utilization, addressing nuanced concerns like process variation and application interference in shared caches.

The survey's culmination in proposing a hybrid cache replacement policy, EHC-DAAEP, amalgamating the predictive strengths of EHC and the application-aware eviction strategy of DAAEP, underlines the quest for enhanced cache performance and optimal resource utilization.

In essence, this report illuminates the trajectory of cache replacement policies, from foundational principles to specialized and context-aware strategies. It underscores the domain's evolution towards adaptive, predictive, and contextually sensitive cache management solutions, poised to meet the diverse challenges of modern computing architectures.

A. APPENDIX

A.1. References

1.Moinuddin K. Qureshi, Aamer Jaleel, Yale N. Patt, Simon C. Steely, and Joel Emer. 2007. Adaptive insertion policies for high performance caching. SIGARCH Comput. Archit. News 35, 2 (May 2007), 381–391. https://doi.org/10.1145/1273440.1250709

2.Aamer Jaleel, William Hasenplaugh, Moinuddin Qureshi, Julien Sebot, Simon Steely, and Joel Emer. 2008. Adaptive insertion policies for managing shared caches. In Proceedings of the 17th international conference on Parallel architectures and compilation techniques (PACT '08). Association for Computing Machinery, New York, NY, USA, 208–219. https://doi.org/10.1145/1454115.1454145

3.Aamer Jaleel, Kevin B. Theobald, Simon C. Steely, and Joel Emer. 2010. High performance cache replacement using re-reference interval prediction (RRIP). SIGARCH Comput. Archit. News 38, 3 (June 2010), 60–71. https://doi.org/10.1145/1816038.1815971

4.M. K. Qureshi and Y. N. Patt, "Utility-Based Cache Partitioning: A Low-Overhead, High-Performance, Runtime Mechanism to Partition Shared Caches," 2006 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'06), Orlando, FL, USA, 2006, pp. 423-432, doi: 10.1109/MICRO.2006.49.

5.A. Vakil-Ghahani, S. Mahdizadeh-Shahri, M. -R. Lotfi-Namin, M. Bakhshalipour, P. Lotfi-Kamran and H. Sarbazi-Azad, "Cache Replacement Policy Based on Expected Hit Count," in IEEE Computer Architecture Letters, vol. 17, no. 1, pp. 64-67, 1 Jan.-June 2018, doi: 10.1109/LCA.2017.2762660.

6.A. Haddadi, M. Rezaei and H. Nikmehr, "IPKB: A New Metric for Evaluating Cache Replacement Policies," 2019 27th Iranian Conference on Electrical Engineering (ICEE), Yazd, Iran, 2019, pp. 1940-1945, doi: 10.1109/IranianCEE.2019.8786446.

7.B. Agarwalla, N. Sahu and S. Das, "PV-aware Replacement Policy for Two-level Shared Cache," 2022 IEEE International Symposium on Smart Electronic Systems (iSES), Warangal, India, 2022, pp. 459-464, doi: 10.1109/iSES54909.2022.00100.

8.C. R. Athni, V. Vinod Chippalkatti, A. Nandakumar, A. V. Nandana and Y. J. Pavitra, "Improved Cache Replacement Policy based on Recency Time Re-Reference Interval Pre-

diction," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/I2CT54291.2022.9824298.

9.Z. Wu, W. Chen, B. Li, Z. Liu and S. Long, "Deadblock Aware Adaptive Eviction Policy for Shared Last-Level Cache," 2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Beijing, China, 2022, pp. 1498-1501, doi: 10.1109/IAEAC54830.2022.9930070.

10.Nayana Prasad Nagendra, Bhargav Reddy Godala, Ishita Chaturvedi, Atmn Patel, Svilen Kanev, Tipp Moseley, Jared Stark, Gilles A. Pokam, Simone Campanoni, and David I. August. 2023. EMISSARY: Enhanced Miss Awareness Replacement Policy for L2 Instruction Caching. In Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA '23). Association for Computing Machinery, New York, NY, USA, Article 62, 1–13. https://doi.org/10.1145/3579371.3589097

A.2. Timeline

The timeline of this survey was from Oct,2023 to Nov,2023.