CSCE 614: HW4 REPORT (SRRIP)

Tex Altstaetter

Texas A&M Computer Science & Engineering

*Abstract*—Write about the replacement techniques you are using in this report in short.

Cache utilization has a great impact on the performance of a computing system. A cache miss can result in hundreds or even thousands of stalls in a program depending on the clock frequency. In this paper we compare the performance of three different cache replacement algorithms: LFU (Least Frequently Used), LRU (Least Recently Used), and SRRIP (Static Re-Reference Interval Prediction). This paper is an analysis of [1] which asserts that SRRIP has an average miss rate 4-10% lower than LRU.

# **Introduction**

A short descriptions of each of the replacement policies and their references. Also compare the pros and cons of each of the techniques.

In the LRU replacement policy, when replacing an entry in the cache, it evicts the entry at the tail and inserts at the head. LRU predicts that more recent entries in the cache will be used in the immediate future and selects the oldest entry for replacement in the cache. ^ Does not consider the frequency.

In the LFU replacement policy, there is a counter associated with each entry that indicates how frequently it has been accessed. Upon adding a new entry, it evicts the entry with the lowest frequency. This has some benefits over LRU in that LRU retains more history. An entry can be the least recently used but will be accessed with high frequency. LFU is more resistant to when there is a mix of new entries in the cache in which infrequent instructions are executed. However, this also biases towards instructions with temporal locality that might not execute in the immediate future and for which other entries are evicted from the cache causing cache misses.

# **SRRIP technique**

## Short description of SRRIP technique

## Implementation details

Computer Science and Engineering, Texas A&M University

Instructor: E. J. Kim, TA: Pritam Majumder

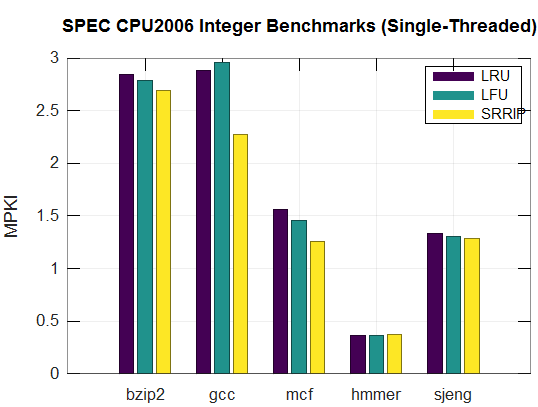
CSCE 614, HW-4 Report, Date: 10/30/2018

# **Methodology**

Write the simulation methodology and configurations considered for the experiments.

# **Evaluation**

In the evaluation section you have to compare the replacement techniques quantitatively. First produce the results. You have to compare three things about the program execution.

1. Number of cycles
2. IPC
3. MPKI

Draw graph (bar chart) for each of these for all the benchmarks across the techniques. Y-axis must be these parameters (#cycles, IPC, MPKI, etc.), and X-axis must be the benchmarks. For each of the benchmarks there should be “replacement technique” number of bars.

Q. How to calculate number of cycles?

*total\_cycles = cycles + cCycles*

For multi-threaded simulations (PARSEC) you have to add across all the processors (westmere-0 to westmere-7). The same thing is valid for all the calculations.

Q. How to calculate IPC?

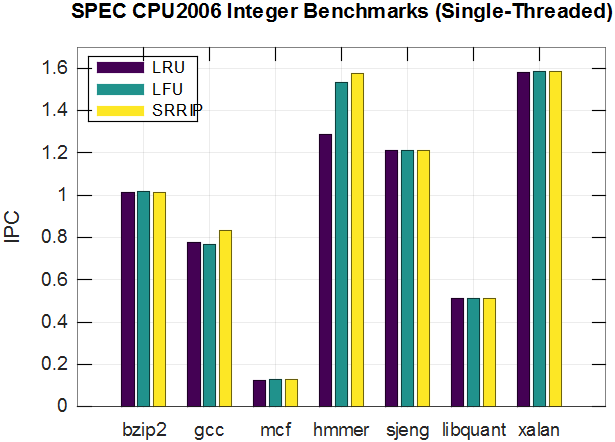
*IPC = #total\_instruction / #total\_cycles*

Q. How to calculate MPKI for L3?

t*otal\_misses = mGETS + mGETXIM + mGETXSM*

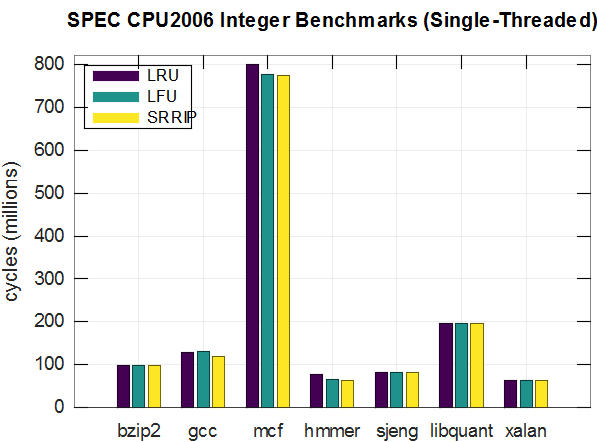
*MPKI = (#total\_misses / #total\_instruction) \* 1000*

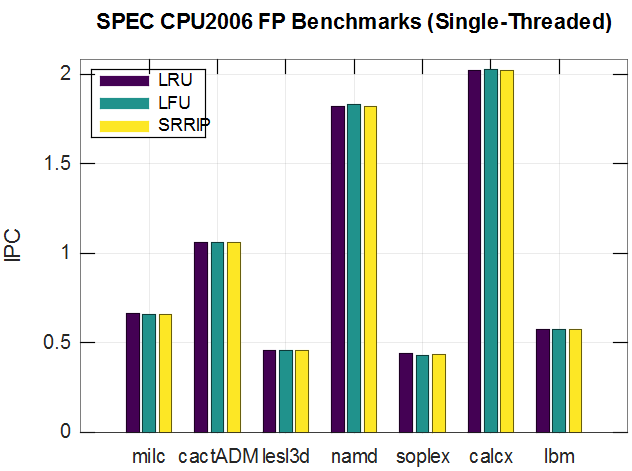
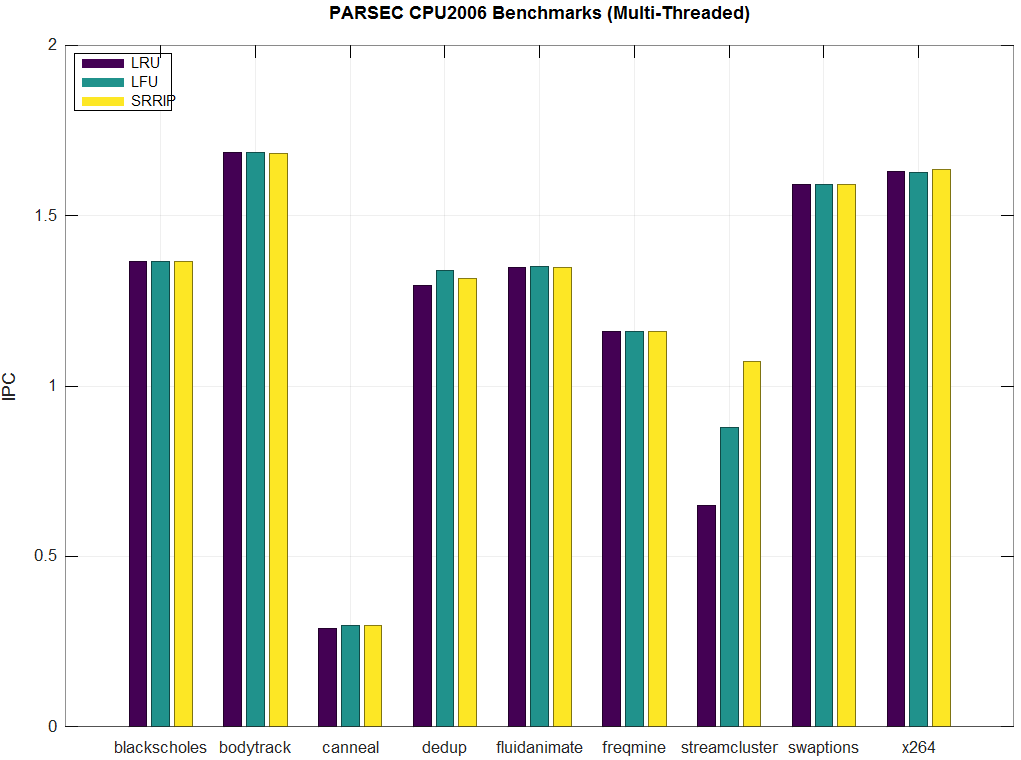
L3 is configured as shared and distributed across multiple cores. So you have sum them up for getting the total number of misses, for multi-threaded workloads (PARSEC).

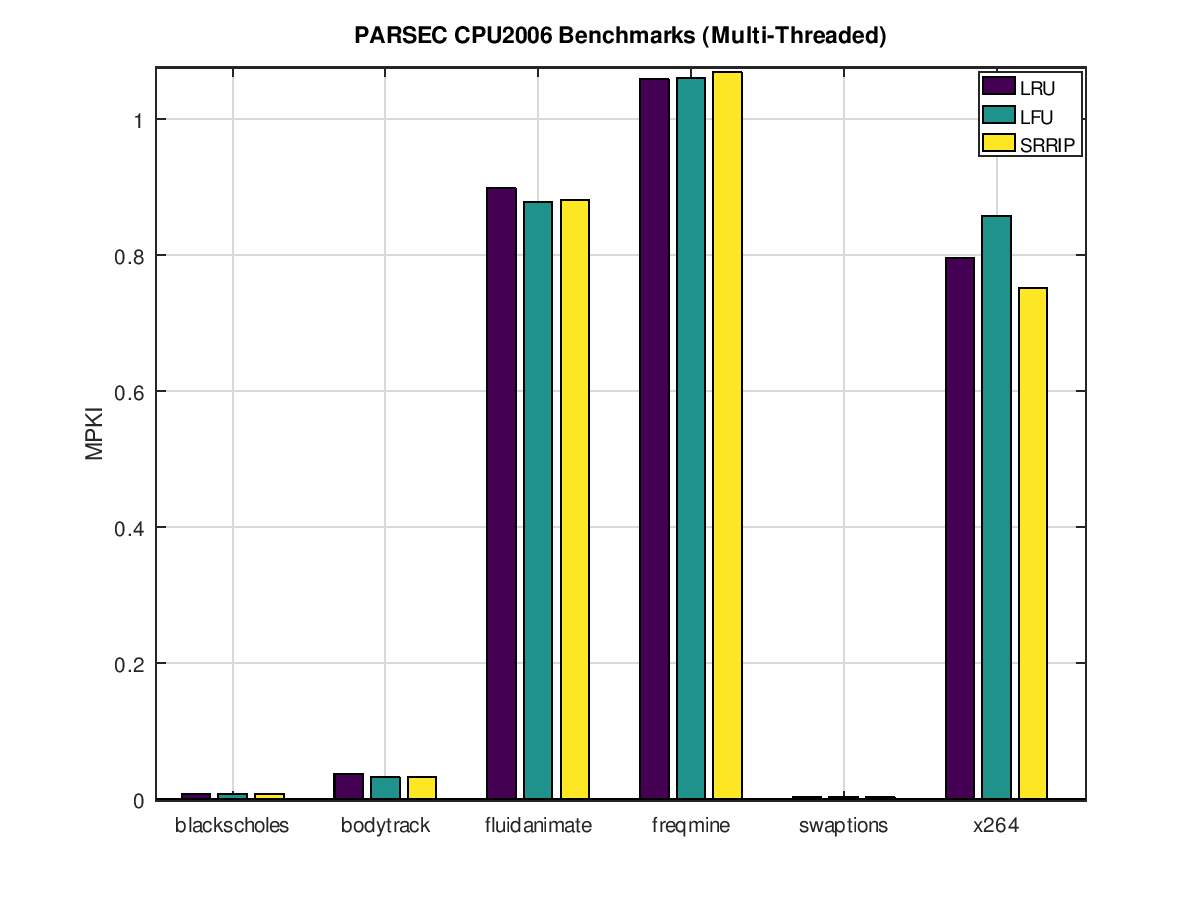


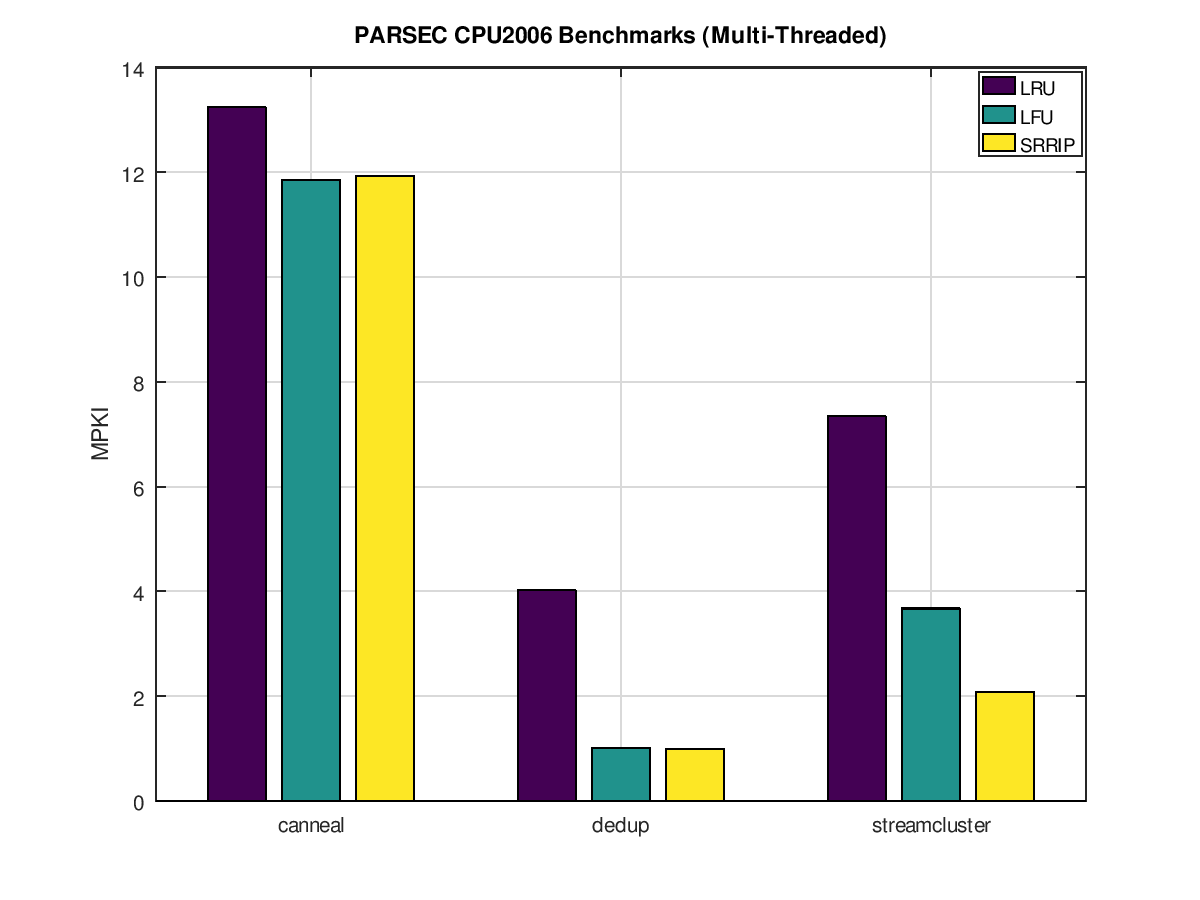
##### 

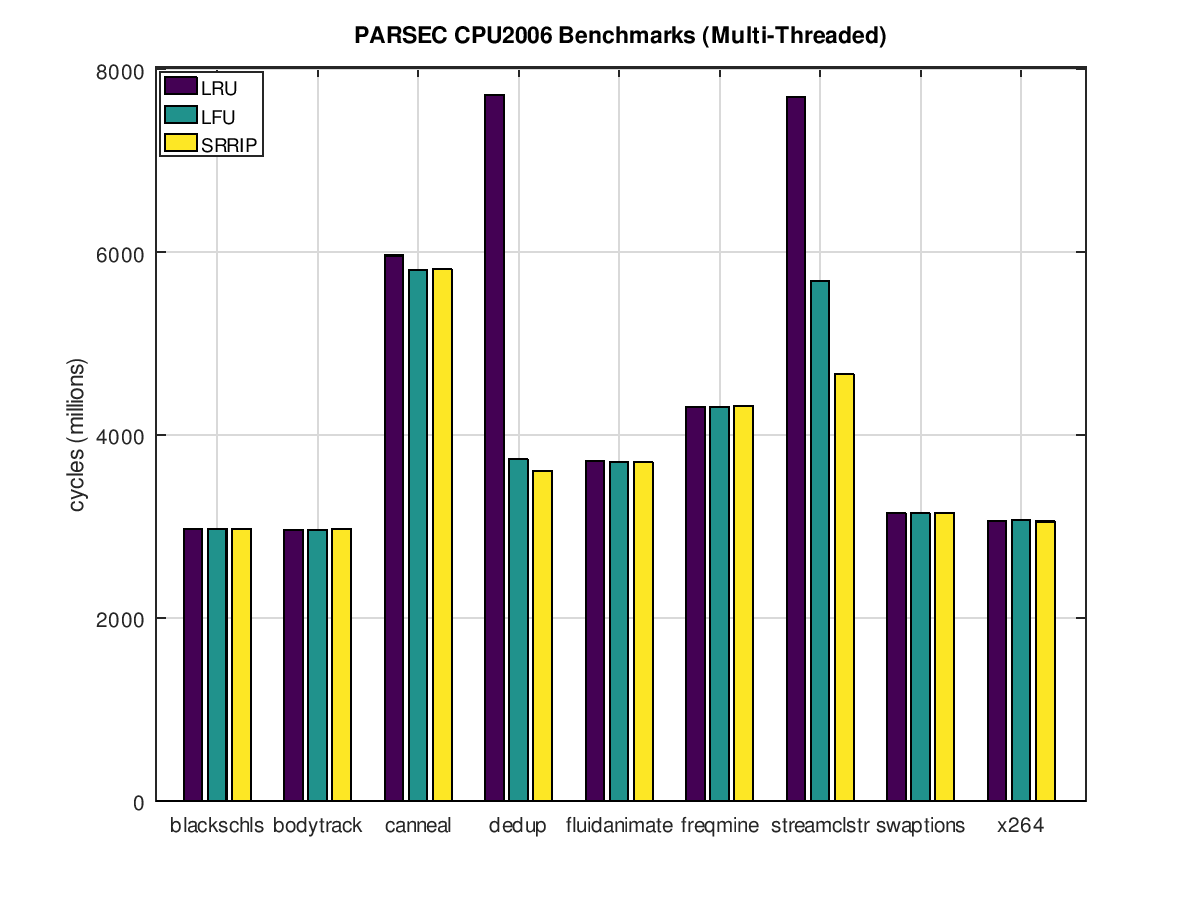
##### 











we ran SRRIP-HP set to zero after a rereference

SRRIP-FP decrements a re-ref instead of predicting

its a near-immed re-ref.

the PARSEC sims have a L3 cache that is 4X larger

8MB vs 2MB

We executed our sims with M=2 and the

insertion value = (2^M)-1 which actually

has the worst performance of the the SRRIP

configurations due to it quickly aging out

It is shown (as indicated in the paper) that

a long re-ref prediction has the best perf

long re-ref t >= 2^(M-1), whereas distant

re-ref t = (2^M)-2

we only evaluated at the L3 cache bc:

6.6. RRIP at Different Cache Levels

At the L1 cache,

SRRIP provides no opportunity to improve performance because the

cache size is too small and the temporal locality is too high. At the L2

cache, SRRIP provides no significant performance gains because the

L2 cache is small (256KB in our study). SRRIP did not degrade

performance of the L1 or L2 caches. To ensure that SRRIP performs

well at the LLC, we modified our hierarchy from a 3-level to a 2-

level hierarchy by removing the L2 cache. For this 2-level hierarchy,

both SRRIP and DRRIP outperform LRU by 4.8% and 10%

respectively. Thus, RRIP is most applicable at the LLC where the

temporal locality is filtered by smaller levels of the hierarchy7.

SRRIP works better when there is not as robust to temporal locality

due to its small cache size. WHen used in the larger L3 cache however,

it does perform relatively better by comparison. The tradeoff is that

scanning the cache for a hit takes longer for more entries, but it

is a low overhead. It is not affected considerably on an insertion bc

an it can replace any element that is of age to be replaced and finds

the first match and is "greedy" in selecting its victim.

8 KB Memory Page Size

The L1-I:

4-way Set Associative

LRU

32 KB

L1-D:

8-way Set Associative

LRU

32 KB

L2:

8-way Set Associative

LRU

256 KB

L3:

16-way Set Associative

SRRIP

RPV\_Max = 3

2 MB

##### **Acknowledgment**

Write about the discussions you have done with your friends (mention names), and also acknowledge their contributions. Also remember it is individual project, where you are allowed to discuss the concepts, but NOT allowed to share your implementations.

##### **Conclusions**

Your conclusion about the techniques.

##### **References**

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first …

1. A. Jaleel, K. Theobald, S. Steely Jr. and J. Emer, “High Performance Cache Replacement Using Re-Reference Interval Prediction (RRIP)” ISCA 2010 Proceedings of the International Symposium on Computer Architecture, vol. 37, pp. 60-71, June 2010.
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper/report may result in your report not being accepted.**