Cache Replacement Improvement on Last-Level Caches

Abdullah Abdul Kader
Computer Science and Engineering Department
Texas A&M University
College Station, TX, United States
akader16@tamu.edu

Yerania Hernandez
Computer Science and Engineering Department
Texas A&M University
College Station, TX, United States
hernandez.yerania@tamu.edu

***Abstract*—Our paper implements the Hawkeye cache replacement policy, which learns from Belady's algorithm by using past cache access history to efficiently predict future cache replacement decisions. We demonstrate that the implementation can accurately make predictions when simulating Belady's behaviour to improve cache replacements. We will be evaluating Hawkeye on a single core with 2MB LLC with and without prefetchers using the SPEC2006 benchmark suite through the ChampSim simulator and comparing the results to other existing cache replacement policies.**

***Index Terms*—Cache replacement, Belady's algorithm**

I. Introduction

Caches are important for a number of reasons. The use of caches reduce long latencies in memory accesses, which calls for an effective cache replacement policy. Cache replacement is a difficult problem compared to a problem like branch prediction. The answer to the question "Which cache line to evict?" is unfortunately not as simple as "Will this branch be taken?". Existing replacement algorithms such as Least Recently Used (LRU) and Most Recently Used (MRU) run based on assumptions and work for only certain type of loads. These algorithms fall short in complex scenarios. Back to the future [3] presents the example of a triply nested algorithm used in matrix multiplication to study the performance of different replacement policies as shown in Figure 1 to demonstrate the reuse patterns of A, B and C. In this paper, we implement the algorithm for cache replacement based on Belady's algorithm [3] and include the improvements provided in [4] for prefetching. Even though Belady's algorithm is optimal, it's impractical since it requires the knowledge of the future. Nevertheless, it is possible to apply a variation of the Belady's algorithm by learning the past behavior and using it to predict the future. The decisions made using the Belady's algorithm is referred to as OPT in this paper.

The proposed cache replacement strategy contains two components. The first reconstructs the Belady's solution for past cache accesses and the second is the predictor, which learns OPT's behavior to inform eviction decisions for future loads. The challenge with reconstructing Belady's is that it looks arbitrarily in the future and requires remembering a long history. However, Belady's algorithm has a better performance even with a reuse window of eight times the cache size which leads to the implementation of the Hawkeye replacement policy [3]. Hawkeye is evaluated using the ChampSim simulator that was released by the Second Cache Replacement Championship (CRC) [1]. Additionally, we compare various replacement policies with Hawkeye using a single core configuration with and without a prefetcher. To account for the prefectch requests, Hawkeye ignores redundant prefetches by considering only the prefetches that follow a demand request. Additionally, separate predictors are used to distinguish between the caching behavior associated with demand and the prefetches by the same load instruction.

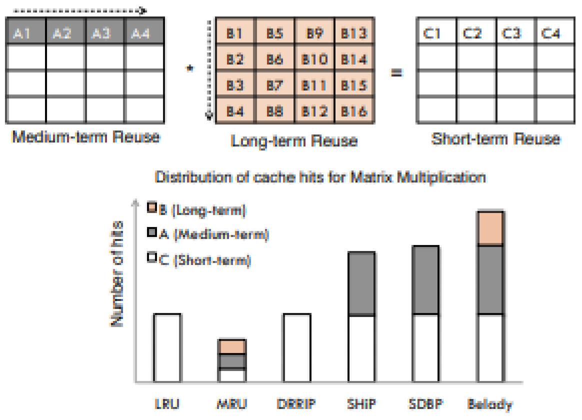

Fig. 1. Existing replacement policies are limited to a few access patterns and are unable to cache the optimal combination of A,B and C

II. Related Work

Some closely related work in the area of cache replacement that is similar to Hawkeye is addressed in this section.

A. Short term history information

Several cache replacement policies rely on short term information to determine which cache line to evict while ignoring the history. One popular example is LRU, which favors the recently used line assuming it will be used again soon in the future. Some other policies rely on access frequencies to decide which line to evict with the assumption that a high frequency access line will soon be accessed again. With the use of re-reference interval prediction, Jaleel et al. [5] enhances recently

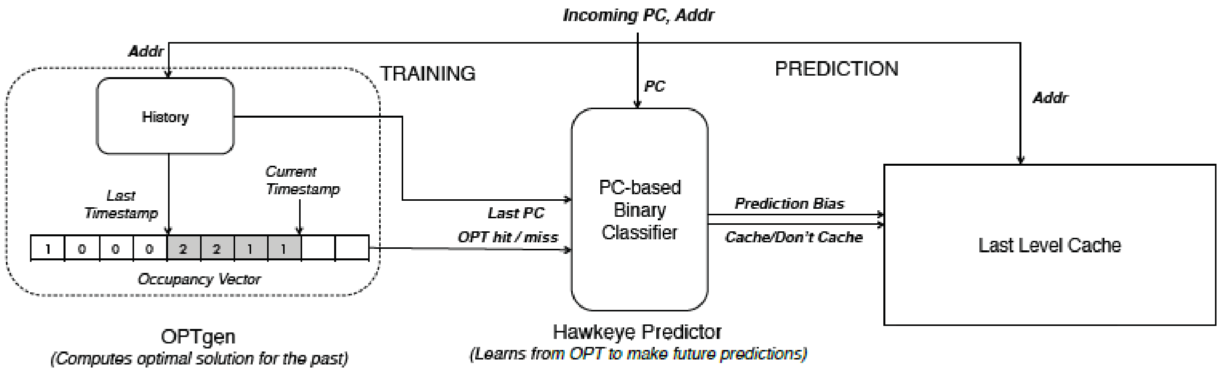

Fig. 2. Overall structure of Hawkeye replacement algorithm

friendly policies by using 2-bits for prediction. Hawkeye uses RRIPs idea of aging to adapt to changes in phase behavior. Other policies include hybrid replacements which allow the flexibility of choosing a policy for different cache lines and Hawkeye exploits this approach by using different policies for different load instruction.

B. Long term history information

Contrast to the short term information approach, some replacement policies exploit long term information to inform their eviction decisions. For instance, some policies use an expensive approach of reuse distances to predict their decisions. On the contrary, Hawkeye heavily relies on liveliness interval which is considerably inexpensive. SHiP [2] uses a predictor to identify instructions that load stream accesses. Hawkeye builds on the prior work done by SHiP to inform future decisions based on past load instructions. Hawkeye uses a longer history (8x the size of cache) compared to SHiP to inform evict decisions.

III. Design Overview

The purpose of Hawkeye is to be able to predict if incoming lines are cache-friendly or cache-adverse. If a line is predicted to be cache-friendly, it will be inserted with high priority in the cache, while if a line is predicted to be cache-averse, it will be inserted with low priority in the cache. In order to provide this prediction, Hawkeye attempts to create a Belady's optimal solution based on the past accesses and applies these solutions to predicting the caching behavior of future loaded instructions.

Figure 2 summarizes the structure of the Hawkeye replacement algorithm. The two main components are the OPTgen algorithm, which is used to compute the optimal solution from past access, and the Hawkeye Predictor, which uses OPTgen's solutions as input in order to train a PC-based predictor to provide an eviction decision. We will describe each component of the Hawkeye in further detail.

A. OPTgen

OPTgen is the algorithm used to compute the OPT of past history caches. OPT is the decision made by Belady's algorithm by relying on history to predict if a load instruction, for example, will bring a line in the future based on it's past behavior. This helps inform eviction decisions for future load instructions. The terms *usage interval* and *livelines interval* are used to further define the functions of OPTgen. *Usage interval* is the time period that starts with a reference to X and leads up to but not including it's next reference X'. Alternatively, *usage interval* can also be defined as the demand of X in a cache. *Livelines interval* is the time the line resides in the cache under the OPT policy. Figure 3 demonstrates the usage interval of X which proceeds up to it's next reference. The liveliness intervals of A, B and C start at their respective first reference and end immediately at their second reference. These liveliness intervals do not overlap, therefore the cache capacity is not reached leading to a cache hit for X'. The question of whether a line would be a hit or miss is based on both the reuse distances and their overlap.

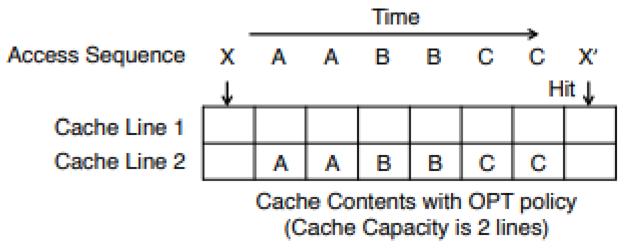

Fig. 3. Intuition behind OPTgen.

An *occupancy vector* is used to keep a track of the occupied cache capacity over time. The number of liveliness intervals that overlap at any time is recorded in each entry of the vector. Figure 4 shows the use of an occupancy vector through the view of access stream in Figure 4(a), the optimal solution to the access stream is presented in Figure 4(b) and the Figure 4(c) shows how the occupancy vector gets calculated and modified over time. For each time starting at 0 (most recent entry) when line X is not loaded for the first time, the OPT checks to see if every element corresponding to the usage interval is less than the cache capacity. If yes, then X is placed

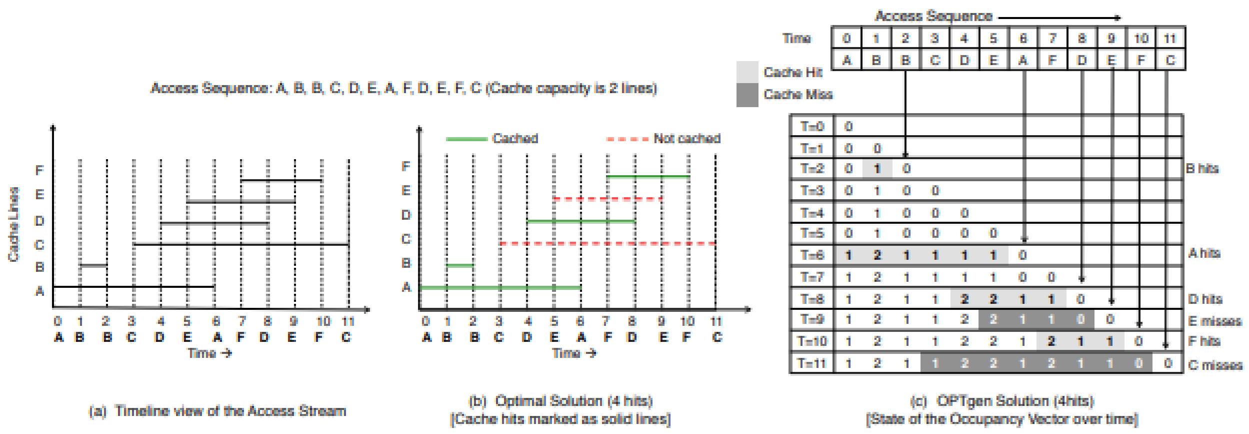
Access Sequence: A, B, B, C, D, E, A, F, D, E, F, C (Cache capacity is 2 lines)
Cache Lines
Time →
(a) Timeline view of the Access Stream
Cached
Not cached
(b) Optimal Solution (4 hits)
[Cache hits marked as solid lines]
Access Sequence
Time
Cache Hit
Cache Miss
B hits
A hits
D hits
E misses
F hits
C misses
(c) OPTgen Solution (4hits)
[State of the Occupancy Vector over time]

Fig. 4. Example to illustrate OPTgen

in the cache.The occupancy vector is incremented and it is shaded. Consider the time T=8, the usage interval starts at 4 and ends at 7 shown by the shaded portion. At time T=7 the cache capacity at all entries starting from 4 to 7 are less than cache capacity (2). This results in a hit for D as shown in T=8 while E is a miss in time T=9 because the an element in the usage interval for E has a value of 2, already at the cache capacity. See Appendix A for the code implementation of OPTgen.

B. *Hawkeye Predictor*

The second major component of this design is the Hawkeye Predictor, which classifies loaded instructions as either cache-friendly or cache-averse. It learns from the OPT policy whether a loaded instruction would have caused a hit or miss. If the OPTgen determines that there was a hit for a specific line, then the PC of that recently accessed line is trained positively under the predictor. Otherwise, if there was miss instead, then the PC of that recently accessed line is trained negatively. Overall, the Hawkeye Predictor will consist of 2K entries, 5-bit counters to train, and indexed by an 11-bit hashed PC. See Appendix B for the code implementation of the Hawkeye Predictor and Appendix C for the code implementation of the hash algorithm.

C. *Cache Replacement*

To efficiently implement Hawkeye, each of the cache lines are associated to a 3-bit RRIP counter. This counter maintains the eviction priorities, where a high RRIP value (RRPV) represents a high eviction priority and a low RRPV represents a low eviction priority. As a result, the RRIP counter is maintaining information on both the Hawkeye's prediction and the age for each line. The Hawkeye predictor will generate a prediction on every cache access. If the line is predicted to be a cache-friendly line, then the RRIP counter will update with an RRPV of 0. However, for a newly inserted cache-friendly line, the RRIP counters for all the other cache-friendly lines are also updated by increasing their value by one (aging). If the line is predicted to be a cache-averse line, then the RRIP counter will update with an RRPV of 7. Figure 5 summarizes how the RRIP counter is updated with Hawkeye's predictions.

Hawkeye Prediction \ Hit or Miss	Cache Hit	Cache Miss
Cache-averse	RRIP = 7	RRIP = 7

Cache-friendly	RRIP = 0	RRIP = 0;
		Age all lines:
		if (RRIP < 6)
		RRIP++;

Fig. 5. Hawkeye's policy based on predictor.

With this setup, cache replacement is decided based on the RRPV. Any line with the RRPV of 7 is selected as an eviction candidate. However, if there is no line that contains an RRPV of 7, then the Hawkeye selects the line that contains the next highest RRPV as the eviction candidate. The implementation of this cache replacement aspect is still in progress considering it is integrating the OPTgen and Hawkeye Predictor. However, the initial setup of the replacement state has been created and we only have the update replacement state left to develop.

D. *Prefetch Awareness*

In order to accommodate prefetch requests, the OPT gen and the Hawkeye predictor need to be slightly modified. In the case of no prefetching as shown in Figure 3, the start and end of liveliness intervals are both demand accesses. However, when prefetching is introduced, it essentially adds up 4 different intervals: Prefetch to Prefetch P-P, Prefetch to Demand P-D, Demand to Prefetch D-P and Demand to Demand D-D. Treating all these intervals similarly would result in poor performance since they contribute to redundant caches as shown in Figure 6. Therefore, we modify the

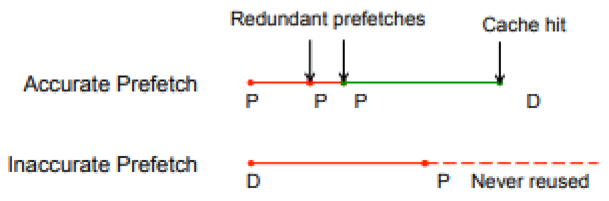

Fig. 6. Liveness intervals that end with a prefetch are not cached by prefetch-aware OPTgen.

liveliness interval to consider only the intervals that end with a demand access P-D or D-D. As seen in Figure 6, P-P and D-P result in inaccurate fetches when they are not followed by a demand access. The implementation of prefetch awareness in Hawkeye is still in progress considering it will take place in the integration of OPTgen and Hawkeye Predictor, as described in the previous subsection. Once we have developed the cache replacement aspect of Hawkeye, we will easily be able to accommodate the modifications for prefetch requests.

IV. Evaluation Methodology

We will be evaluating our algorithm using the ChampSim simulator that was released by the Second Cache Replacement Championship (CRC) [1]. This simulator is designed to evaluate replacement algorithms for private and shared last level caches (LLCs) with and without a data prefetcher. We will evaluate Hawkeye using the two different configurations that the CRC provides:

- Single core with 2 MB LLC without a prefetcher.
- Single core with 2 MB LLC with L1/L2 data prefetchers.

In order to compare the results of the Hawkeye algorithm with that of the original paper, we will also use the SPEC2006 benchmark suite with the traces that demonstrate more than a 2% improvement with the OPT policy. In order to generate the samples from each benchmark, we used Pin 3.2 with the CRC, creating a total sample of 250 million instructions. This sample is divided between the warm-up stage and behavior stage, where 50 million instructions are used to warm the cache and the other 200 million instructions are used to measure the behavior.

To evaluate the performance, we will be reporting the miss rate and the speedup for each benchmark combination. In addition, we will compare our implementation of Hawkeye to other caching systems. We will compare against LRU (Least Recently Used), SRRIP (Static Re-Reference Interval Prediction), and SHiP (Signature-based Hit Predictor). As noted earlier, LRU is a commonly used replacement policy always predicting a "near-immediate re-reference interval on cache hits and misses" [5]. Further improvements have been made on the policy by using SRRIP for a scan-resistant approach. SHiP is a cache replacement algorithm that learns caching priorities on each PC load and uses a 2-bit re-reference counter for each cache line. For each of these implementations, we used the code provided on the CRC website and further tuned the parameters to provide an adequate comparison to the Hawkeye.

V. Results

Once we have completed our simulations, we will evaluate the performance of Hawkeye for the different configurations we have listed in the *Evaluation* section.

A. Single-core, No Prefetching Configuration

B. Single-core, With Prefetching Configuration

VI. Conclusion

In this paper, we explained the original Hawkeye cache replacement policy that was developed for the CRC in 2016 [3] and implemented this algorithm with the improvements presented in the 2017 paper [4]. This algorithm consisted of three major advantages compared to current work in cache replacement. From one aspect, the Hawkeye is able to demonstrate that by looking at a long history of memory accesses, it can learn and predict the optimal behavior for the instructions loaded. Additionally, unlike other policies, the decisions provided by the Belady's algorithm reuses all types of workloads, not solely based on recently used or mostly used. The last advantage of using Belady's algorithm is that it is able to consider both reuse distance and cache demand. With further analysis and simulations on Hawkeye, we will demonstrate performance improvements of this algorithm compared to other policies in miss predictions and speedup.

For future work, we expect to see improvements in cache predictors due to the trend of sophisticated branch predictors. There is a strong potential of improving the Hawkeye predictor while using the OPTgen in order to model the optimal OPT's behavior. Additionally, considering Hawkeye is able to provide long-periods of history, the information provided could be useful in other fields other than caching, such as optimizing shared memory systems. Finally, further evaluation on the relationship between replacement policies and the presence of prefetchers will demonstrate that Hawkeye could provide an optimal solution.

References

[1] "Champsim Source." THE 2ND CACHE REPLACEMENT CHAMPIONSHIP, 2017, crc2.ece.tamu.edu/?page_id=41.

[2] C.-J. Wu, A. Jaleel, W. Hasenplaugh, M. Martonosi, S. C. Steely, Jr., and J. Emer. SHiP: Signature-based hit predictor for high performance caching. In 44th IEEE/ACM International Symposium on Microarchitecture (MICRO), pages 430441, 2011.

[3] J. Akanksha, and C. Lin, "Back to the Future: Leveraging Belady's Algorithm for Improved Cache Replacement", 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), 2016, doi:10.1109/isca.2016.17.

[4] J. Akanksha, and C. Lin, "Hawkeye: Leveraging Belady's Algorithm for Improved Cache Replacement," CS-University of Texas, 2017, www.cs.utexas.edu/ lin/papers/crc17.pdf.

[5] J. Aamer, et al, "High Performance Cache Replacement Using Re-Reference Interval Prediction (RRIP)," Proceedings of the 37th Annual International Symposium on Computer Architecture - ISCA 10, 2010, doi:10.1145/1815961.1815971.

Appendix A
OPTgen

```
struct optgen{
    vector<unsigned int> liveness_intervals;
    uint64_t num_cache;
    uint64_t access;
    uint64_t cache_size;

    //Initialize values
    void init(uint64_t size){
        num_cache = 0;
        access = 0;
        cache_size = size;
        liveness_intervals.resize(OPTGEN_SIZE, 0);
    }

    //Return number of hits
    uint64_t get_optgen_hits(){
        return num_cache;
    }

    void set_access(uint64_t val){
        access++;
        liveness_intervals[val] = 0;
    }

    void set_fetch(uint64_t val){
        liveness_intervals[val] = 0;
    }

    //Return if hit or miss
    bool is_Cache(uint64_t val, uint64_t endVal){
        bool cache = true;
        unsigned int count = endVal;
        while (count != val){
            if(liveness_intervals[count] >= cache_size){
                cache = false;
                break;
            }
            count = (count+1) % liveness_intervals.size();
        }

        if(cache){
            count = endVal;
            while(count != val){
                liveness_intervals[count]++;
                count = (count+1) % liveness_intervals.size();
            }
            num_cache++;
        }
        return cache;
    }
};
```

Appendix B
Hawkeye Predictor

```
class Hawkeye_Predictor{
private:
  map<uint64_t, int> PC_Map;

public:
  //Return prediction for PC Address
  bool get_prediction(uint64_t PC){
    uint64_t result = CRC(PC) % PCMAP_SIZE;
    if(PC_Map.find(result) != PC_Map.end() && PC_Map[result] < ((MAX_PCMAP+1)/2)){
      return false;
    }
    return true;
  }

  void increase(uint64_t PC){
    uint64_t result = CRC(PC) % PCMAP_SIZE;
    if(PC_Map.find(result) == PC_Map.end()){
      PC_Map[result] = (MAX_PCMAP + 1)/2;
    }

    if(PC_Map[result] < MAX_PCMAP){
      PC_Map[result] = PC_Map[result]+1;
    }
    else{
      PC_Map[result] = MAX_PCMAP;
    }
  }

  void decrease(uint64_t PC){
    uint64_t result = CRC(PC) % PCMAP_SIZE;
    if(PC_Map.find(result) == PC_Map.end()){
      PC_Map[result] = (MAX_PCMAP + 1)/2;
    }
    if(PC_Map[result] != 0){
      PC_Map[result] = PC_Map[result] - 1;
    }
  }

};
```

Appendix C
CRC Hash Algorithm

```
uint64_t CRC(uint64_t address){
  unsigned long long crcPolynomial = 3988292384ULL;  //Decimal value for 0xEDB88320
    hex value
    unsigned long long result = address;
    for(unsigned int i = 0; i < 32; i++ )
      if((result & 1 ) == 1 ){
        result = (result >> 1) ^ crcPolynomial;
      }
      else{
        result >>= 1;
      }
    return result;
}
```