# A Futuristic Cache Replacement using Hybrid Regression

**Prakash A. Sapkale**

*Department of MCA, Veermata Jijabai Technological Institute Mumbai, India.*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** *Cache Memory is readily available on every device as it is fast but expensive and small in size. To overcome a situation where we want to use cache memory optimally we researched through varies algorithms like LRU and MRU, even though they are not optimal but we are still using it for our day to day life. In this paper we have studied the several replacement algorithms and adding to that we have a hybrid regression algorithm with which we can get a higher hit rate. The parameters we have used are Memory Region, Memory instruction PC, and Instruction Sequence as it gets higher hit rate as reusable addresses can be found. We updated our model and found that even with random data we get 279% higher hit rate in FIFO and 251% higher hit rate in LRU.*

***Key Words***: Hit/Miss Rate, LRU, FIFO, Belady Algorithm, Cache Memory Simulation, SHiP, Hybrid Regression, SDBP.

## 1. INTRODUCTION

Cache memory is widely used in computers and mobile phones. Cache memory is very fast memory but costly. It includes three types of memory that are L1 (till 64kb), L2 (till 512kb) L3 (2MB or more). Comparatively, it is faster than RAM. The main point of our cache system is to manage this system by inserting addresses to Cache memory and retrieving it also to decide which pages to be removed from cache memory. To manage this cache system traditionally we use LRU or MRU method. But LRU or MRU is not the optimal methods. We will discuss some replacement algorithms in this paper, which are LRU, FIFO, Hawkeye/PA-Hawkeye and SHiP.

LRU is nothing but removing a least used element in our cache memory to free up space for new elements. FIFO is a replacement policy which focuses on a simple idea that which comes first will be removed first. So for freeing up space we will be removing the first element inserted. The SHiP is based on three factors that are memory region, memory instruction PC and instruction sequence as Signature. We use Re-reference pattern of a signature. SHiP decides on the signature and cache hit or misses that the prediction will be immediate, intermediate, far or distant. Hawkeye is based on Beladys MIN algorithm. Beladys MIN algorithm is that when memory is full go to the future find the least used pages and come back to delete those pages to free up the memory. Hawkeye uses the OPTgen algorithm for training the data coming inside the cache memory and Hawkeye Predictor decides which lines are cache friendly according to that it frees space. PA-Hawkeye uses the same Hawkeye predictor with some modifications. In PA-Hawkeye we are using the prefetching technique. The prefetching technique is to make the process fast of adding the pages from slow memory (hard drive) to fast memory(L3). We use modifications of Beladys MIN with Demand-MIN and FLEX-MIN as a differentiator of both of them. We have come across a method called Hybrid Regression method where it can be useful for cache management. A Hybrid Regression Model for Video Popularity-based Cache Replacement in Content Delivery Networks in this paper, on the limited storage of cache they are trying to have more hit ratio. They have created a new prediction model which will use past data. It also uses an offline and online database of videos. And try to improve the algorithm by understanding hit ratio so the more popular of a video more it will stay in cache even with the case of new videos getting popular.

In SHiP algorithm they have used a signature based hit predictor with three factors that is 1) Memory Region 2) Memory Instruction PC 3) Instruction Sequence. With using them they have predicted it should be in memory or it should be evicted. In Hybrid Regression Paper they have used Watch time, Subscription and Age of Video because of this parameter it can predict the video should be stored in the cache system or not. Same we will do in this scenario with our Signature Based Factors.

## 2. Related Work

SHiP have shown that LLC accesses by the same signature have similar re-reference patterns [2]. Namely Memory Region, Memory Instruction PC, and Instruction Sequence. It can use SHCT for Re-Reference Predictions on Hits. The Strengths of this paper, State-of-the-art policies do not fully address scan-resistance. Signatures help improve re-reference predictions to address scans. They have Proposed a Simple and Practical Scan-Resistant Replacement. The SHiP requires less storage. Hardware overhead of SHiP is comparable to LRU.

Learning from Beladys algorithm [3,8] by applying it to past cache accesses to inform future cache replacement decisions. Showing that the implementation is surprisingly efficient, as they introduced a new method of efficiently simulating Beladys behavior, and they use known sampling techniques to compactly represent the long history of information that is needed for high accuracy. Occasionally, load instructions will have a low bias, which will result in inaccurate predictions. Their evaluation shows that we can get a small performance gain by augmenting Hawkeyes predictions to include confidence, but the gains are not justified by the additional hardware complexity, so they don’t evaluate this feature in this paper. They have observed that while Beladys MIN algorithm minimizes the total number of cache misses including those for prefetched lines and it does not minimize the number of demand misses. To address this shortcoming, they have introduced Demand-MIN, a variant of Beladys algorithm. It minimizes the number of demand misses at the cost of increased prefetcher traffic. [4] Proposed system requires more memory storage than previous.

According to [5,6], Three types of prediction models using the dataset of articles covering the 4-year period and published by 20 minutes.fr, an important French online news platform. As they perform these methods which are simple linear regression, linear regression on a logarithmic scale model and constant scaling so at last simple linear regression outperforms the other two models. This paper does not give improvement over simple linear regression which was implemented before by them.

In the limited storage of cache, they are trying to have more hit ratio and they have created a new prediction model which will use past data as in [7]. It also uses an offline and online database of videos. And try to improve the algorithm by understanding hit ratio so the more popular of a video more it will stay in cache even with the case of new videos getting popular. When the window in time is small. Small window, such as a day, can incorporate daily user behavior and achieve higher accuracy. So it is one limitation of this system. [9] In Hybrid Regression [7] they have used Watch time, Subscription and Age of Video because of these parameters it can predict the video should be stored in cache system or not. Same we will do in this scenario with our Signature Based Factors. our factors include Memory Region, Memory instruction PC, and Instruction Sequence. Below is the formula for hybrid regression,

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + s_i, \qquad (1)$$

## 3. System Overview

### 3.1 Overall Logic

The flow of Cache will be changed as in each cache system there is cache replacement policy which decides which address to evict from cache system but in our new modification we introduced a new algorithm that is Hybrid Regression and it will work between any of the replacement policy and Last Level Cache. The replacement policy can be LRU, FIFO, Hawkeye, PA-Hawkeye, SDBP or SHiP. After getting evicted by a replacement policy from that frame, if the cache is empty then this predictor will insert an address which is most efficient according to hybrid regression algorithm. It will not affect even if we swapped Replacement policy and Hybrid Regression Predictor in our below Fig. 1.

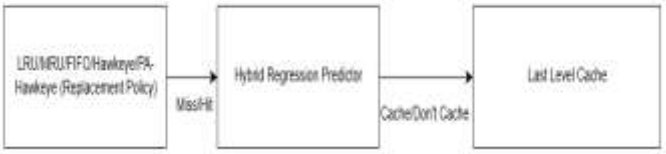

LRU/MRU/FIFO/Hawkeye/PA-Hawkeye (Replacement Policy)
Miss/Hit
Hybrid Regression Predictor
Cache/Don't Cache
Last Level Cache

**Fig -1**: Overall System

### 3.2 Hybrid Regression

The flow of Cache will be changed as in each cache system there is cache replacement policy which decides which address to evict from cache system but in our new modification we introduced a new algorithm that is Hybrid Regression and it will work between any of the replacement policy and Last Level Cache. The replacement policy can be LRU, FIFO, Hawkeye, PA-Hawkeye, SDBP or SHiP. After getting evicted by a replacement policy from that frame, if the cache is empty then this predictor will insert an address which is most efficient according to hybrid regression algorithm. It will not affect even if we swapped Replacement policy and Hybrid Regression Predictor in our below model. B. Hybrid Regression The algorithm is explained in the flowchart (Fig 2,3,4).

At first stage, we have initialized some variables as below, 1. H is when the queued address is found in cache memory. In short, we call it hit. 2. M is when the queued address is missing in cache memory. In short, we call it miss. 3. MR is our Memory Region parameter and we are taking ts predetermined Coefficient Correlation as MR. 4. MIC is our Memory Instruction PC parameter and we are taking its predetermined Coefficient Correlation as MIC. 5. IS is our Instruction Sequence parameter and we are taking its predetermined Coefficient Correlation as IS. 6. E is the current address value from the queue. 7. P is set of our parameters MR, MIC and IS. 8. A is our alpha value we can adjust it for better performance of our algorithm.

At Second Stage, we process addresses from our queue until there are addresses in the system queue. At Third Stage, if Cache is hit then we Hit counter incremented and if not then Miss counter incremented. At Fourth Stage we send our parameters to predict function here we set a default beta value and Calculate New Coefficient Correlation of MR, MIC and IS as show in Predict Diagram. To Calculate the predictor of it we have a formula(Refer formula 1),

$$V = \beta_0 + MR * MR2 + MIC * MIC2 + IS * IS2 \quad (2)$$

The V is returned back which is stored in array P [_]. At Fifth Stage for our P that is MR, MIC and IS, for all of the three we apply update function so the value of each is updated and stored back in their respective variables. For Update function, we have a simple formula,

$$V = (A * P) + (1 - A) * E * E \qquad (3)$$

The V is returned for each parameter and Stored in MR, MIC and IS. So these variables are updated for the next iteration. At the Sixth stage, we sort our P array from the fifth stage and then find the highest value from it. Now we insert the address of highest prediction value in cache for more cache hits. After this, it will again go to the second stage and continue.

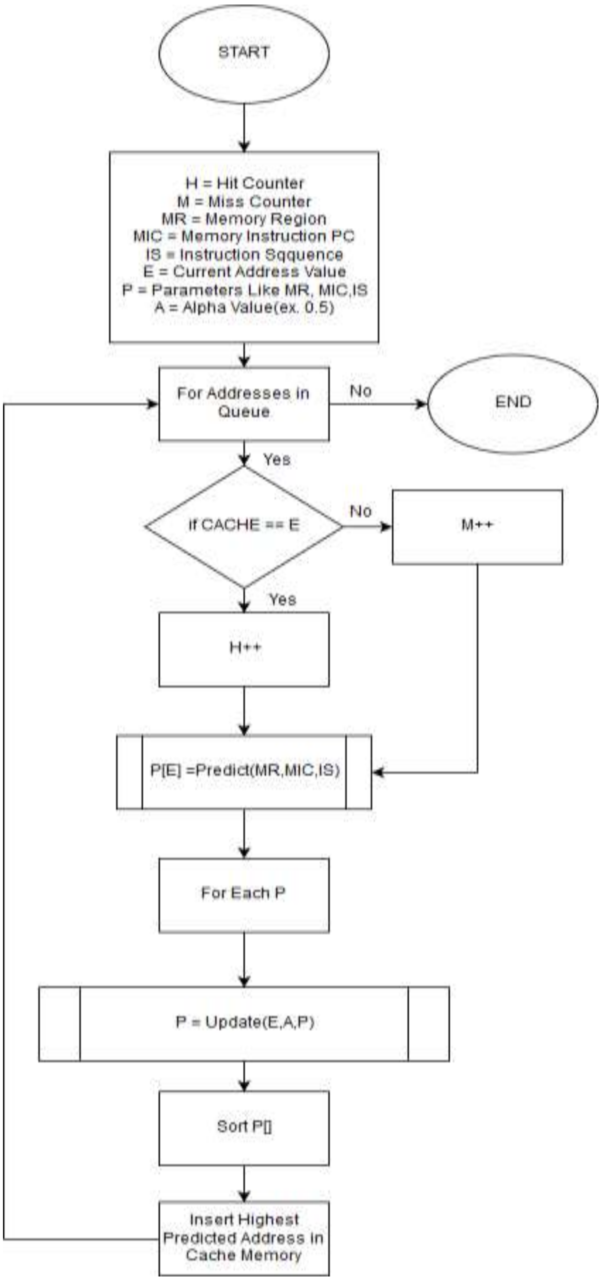


**Fig -2**: Hybrid Regression Flow Chart

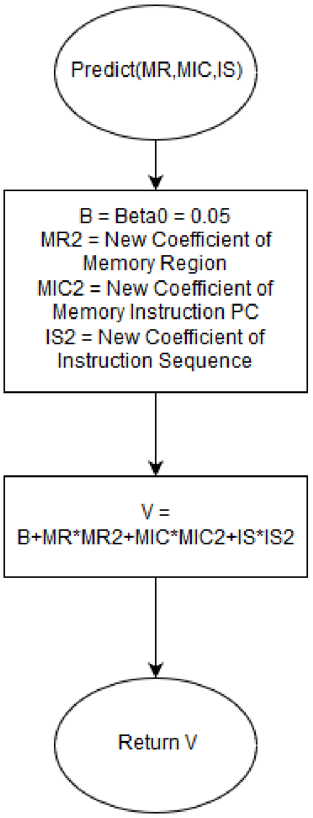

Predict(MR,MIC,IS)
B = Beta0 = 0.05
MR2 = New Coefficient of Memory Region
MIC2 = New Coefficient of Memory Instruction PC
IS2 = New Coefficient of Instruction Sequence
V = B+MR*MR2+MIC*MIC2+IS*IS2
Return V

**Fig -3**: Predict Function Flow Chart

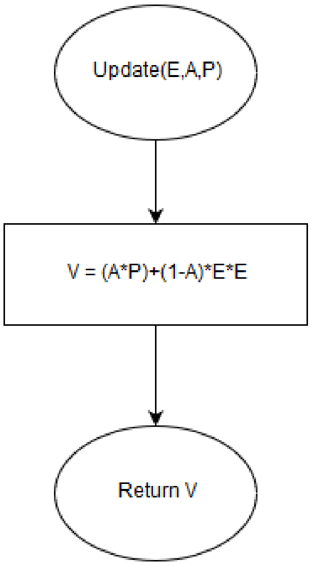


**Fig -4**: Update Function Flow Chart

### 3.3 Problems

In our simulation, to get the better results we needed a good dataset with Memory Region, Memory Instruction PC, Instruction Sequence Addresses, Hit or Miss. But finding a good dataset with all the elements is impossible as this type of data is not available for the public domain. And for each device, it’s different for each user so finding a correlation coefficient for our parameters was a very difficult task.

### 3.4 Proposed Solution

To overcome this situation, we have taken random data for our simulation. But in cache scenario finding coefficient correlation easily, we are suggesting to use Rank Method as it is better and will need less computational power. Also, we are open to suggestions and believe there might be a way to optimize this algorithm.

## 4. Results and Discussions

We have created our simulation with two replacement policies as of now both are well known that is FIFO and LRU. First, we have checked the performance of simple FIFO and FIFO with Hybrid Regression Algorithm. Below Fig. 5 shows the result. It has improved the result of a simple FIFO of 2.88% hit rate to 8.04% hit rate.

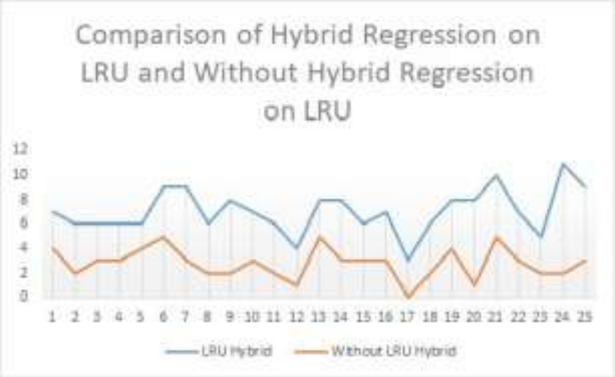
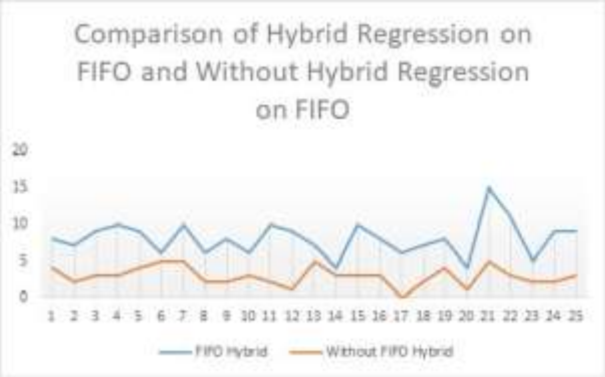


**Fig -5**: Comparison of Hybrid Regression on LRU and without Hybrid Regression on LRU

Second we have checked the performance of simple LRU and LRU with Hybrid Regression Algorithm. Below Fig. 6 shows the result. It has improved the result of simple LRU of 2.8% hit rate to 7.04% hit rate.

**Fig -6**: Comparison of Hybrid Regression on FIFO and without Hybrid Regression on FIFO



Third we have shown the comparisons of simple FIFO, simple LRU, with Hybrid Regression FIFO and with Hybrid Regression LRU. At the Fig. 7, we can see that any replacement algorithm gives better performance with our Hybrid Regression algorithm.

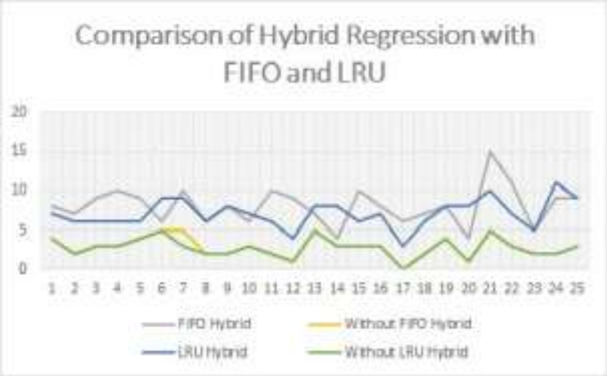


**Fig -7**: Comparison of Hybrid Regression with FIFO and LRU

Fourth we have compared the performance of LRU and FIFO (Chart 1), in this example as per this random scenario FIFO gives better performance than LRU. As Hybrid LRU is improved over simple LRU by 251% and Hybrid FIFO is improved over simple FIFO by 279%. We compare Hybrid LRU and Hybrid FIFO and determine that FIFO is performing better than LRU for in this situation but for each behavioral of user it will be different so propose adaptable replacement policy implementation and it is research topic for future. We also propose the idea of using Simple Dead Block Prediction Technique [1] along with it to know the dead blocks and fill it with our predictor for better results.

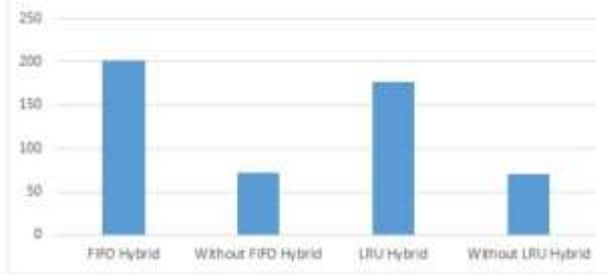


**Chart -1:** Comparison of Hybrid Regression with FIFO and LRU

## 5. CONCLUSION

Our Problem of getting a higher hit rate can be solved with this new algorithm, Overall the speed of cache will be improved and instead of using traditional methods alone we can use those traditional methods plus our Hybrid Regression for better results. Even though we have used random dataset for our simulation but we can conclude that if our algorithm can improve a random scenario with more than 250% then what it will do in real life situation where a user works on similar software again and again and even if user suddenly uses the new application, again and again, our algorithm will understand it and get higher rates accordingly.

## ACKNOWLEDGEMENT

The author is thankful to Dr. A. A. Pandit, HOD, MCA Department, VJTI, Mumbai (India) for useful discussions and support to carry out presented work.

## REFERENCES

[1] S. Khan, Y. Tian, and D. A. Jimenez, Sampling dead block prediction for last-level caches, in 43rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), pp. 175186, 2010.

[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), pp. 430441, 2011.

[3] A. Jain and C. Lin, Back to the future: Leveraging Beladys algorithm for improved cache replacement, in 43rd International Symposium on Computer Architecture (ISCA), June 2016

[4] A. Jain and C. Lin, Rethinking Beladys Algorithm to Accommodate Prefetching in 43rd International Symposium on Computer Architecture (ISCA), June 2018

[5] A. Tatar, P. Antoniadis, M. D. de Amorim and S. Fdida," Ranking News Articles Based on Popularity Prediction," 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, 2012, pp. 106-110.

[6] Tatar, Alexandru, et al." Predicting the popularity of online articles based on user comments." Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011.

[7] E. Ben Abdelkrim, M. A. Salahuddin, H. Elbiaze and R. Glitho," A Hy- brid Regression Model for Video Popularity-Based Cache Replacement in Content Delivery Networks," 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, 2016, pp. 1-7.

[8] L. A. Belady. A study of replacement algorithms for a virtual-storage computer. IBM Systems Journal, 5(2):78101, 1966.

[9] Emira Ben abdelkrim, Mohammad A. Salahuddin, Halima Elbiaze and Roch Glitho," A Hybrid Regression Model for Video Popularity-based Cache Replacement in Content Delivery Networks", 2016 IEEE Global Communications Conference (GLOBECOM)

## BIOGRAPHIES

**Prakash A. Sapkale** is pursuing M.C.A. degree from Veermata Jijabai Technological Institute, Mumbai, Maharashtra, India and completed B.C.A. degree from North Maharashtra University, Maharashtra, India in 2016. His area of interest includes Cache System and Algorithms.