

A Survey of Miss-Ratio Curve Construction Techniques

Daniel Byrne
Michigan Technological University
Houghton, Michigan
djbyrne@mtu.edu

ABSTRACT

Miss-ratio curve (MRC), or equivalently hit-ratio curve (HRC), construction techniques have recently gathered the attention of many researchers. Recent advancements have allowed for approximating these curves in constant time, allowing for online working-set-size (WSS) measurement. Techniques span the algorithmic design paradigm from classic dynamic programming to artificial intelligence inspired techniques. Our survey produces broad classification of the current techniques primarily based on *what* locality metric is being recorded and *how* that metric is stored for processing.

Applications of these curves span from dynamic cache partitioning in the processor, to improving block allocation at the operating system level. Our survey will give an overview of the historical, exact MRC construction methods, and compare them with the state-of-the-art methods present in today’s literature. In addition, we will show where there are still open areas of research and remain excited to see what this domain can produce with a strong theoretical background.

1 INTRODUCTION

From the working set theory proposed by Denning in 1968 [8] to recent advancements in cache locality theory [9, 14] modeling data locality has been essential to cache design and OS resource allocation choices. In the past decade there has been a large emphasis on using these models to construct plots relating miss rate to cache size, called miss-ratio curves (MRCs) or equivalently hit-ratio curves (HRCs).

Miss ratio curves have proven to be extremely useful in estimating *how much* data is being used by a particular workload, known as the working set size. Knowing the working set size of a workload helps estimate the *utility* of adding more memory or increasing the cache size. An example miss-ratio curve is shown in Figure 1. The knee-point in this particular MRC is just over cache size 3000 for this workload. Notice how that an increase cache size to 4000 yields no significant decrease in miss-rate.

Given the recent theoretic and implementation advancements in miss-ratio curve construction, we find a survey on these methods to be useful update to the community and present some identified open lines of research yet to be fully explored in this domain.

2 TAXONOMY

In this section we present a taxonomy of the current miss-ratio curve construction techniques based on *which* metric is recorded, *how* the data is stored, and the *derivation method* of the miss-ratio curve.

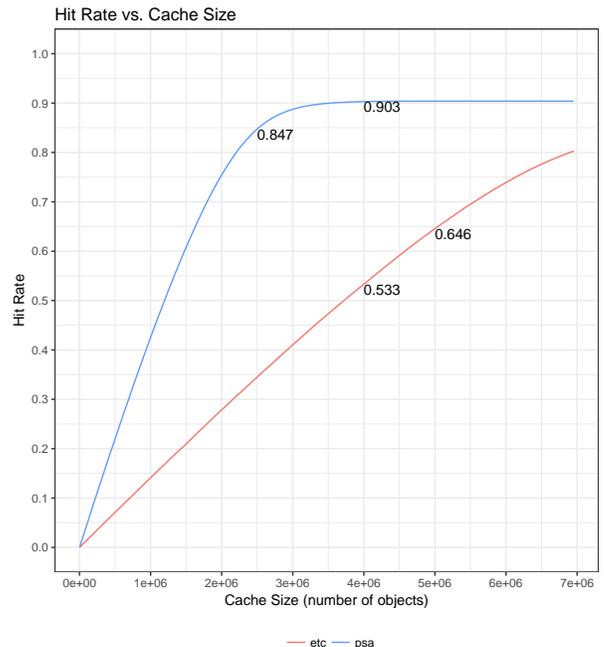


Figure 1: Hit Ratio Curve for two common synthetic web application workloads. ETC models the Facebook workload from [2]. It contains 50 million accesses (GETs) to 7 million unique objects. PSA is the D’Carra workload from [6], it contains 50 million accesses to 7 million unique objects following an Pareto distribution.

2.1 Definitions

In the cache modeling literature there are variations on meaning of terms such as *reuse distance*. To avoid this, we give definitions for the following terms used in our taxonomy:

- Stack Distance - The number of *unique* accesses between two accesses to the same piece of data. This has also been referred to as reuse distance.
- Reuse Time - The *total* number of accesses between two accesses to the same piece of data. This is consistent with [13, 14, 25, 26].
- Buckets - A set of variable sized bins that contain access to a given piece of data.
- Access Counters - Hardware or software monitors that count unique or total accesses to a piece of data.

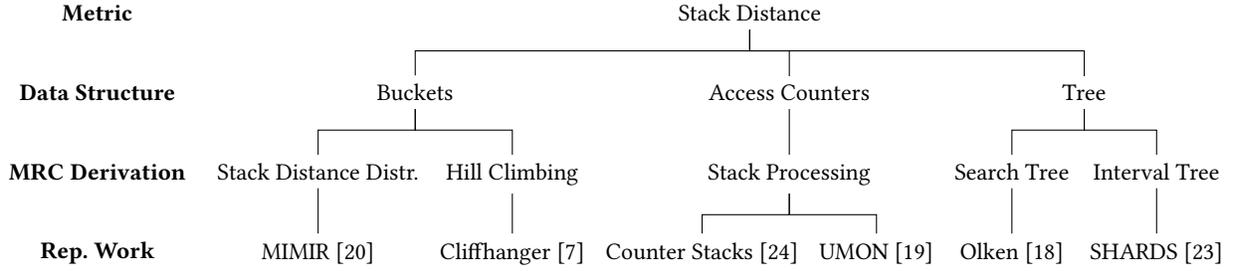


Figure 2: Methods based on Stack Distance

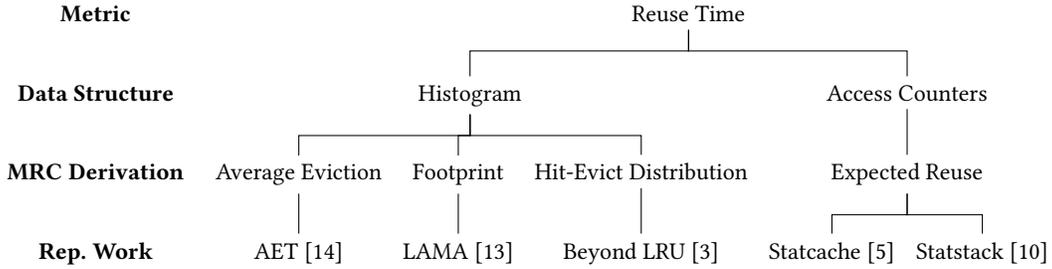


Figure 3: Methods based on Reuse Time

2.2 Taxonomy of MRC Constructions

Figure 2 and 3 classify the techniques that use stack distance and reuse time respectively, by how they store and process the data into a miss-ratio curve. We also give a representative work from each class.

3 STACK DISTANCE ALGORITHMS

This class of algorithms for calculating miss-ratio curves takes advantage of the least recently used (LRU) replacement policy which maintains the stack order. Table 1, demonstrates this for a set of requests.

In order to generate a miss-ratio curve for a given cache size, we create a *distribution* of stack distances. Once we have a distribution of the stack distances, we can calculate the number of hits in a cache of size d as the sum of all the references with stack distances $\leq d$ and equally the number of misses as the sum of all references with stack distance $\geq d$.

Formally, let cache size be d , N is the total number of requests, and M is the total number of unique requests, then the *missrate*(d) is express as:

$$\left(1 - \frac{\sum_{i=1}^d \text{freq}(i)}{N}\right) = \frac{\sum_{i=d}^M \text{freq}(i)}{N}$$

It is important to note that some represent this as an integration of the probability density function given by the distribution. The reason is to highlight the order relationship between stack distance

and miss-rate, that is, miss-rate is an order higher than stack distance. This result was proved in Ding’s Higher Order Theory of Locality (HOTL) [9].

3.1 Mattson’s Stack Distance Algorithm

Mattson et. al. [16] developed the first algorithm to use stack distance as a method for calculating the miss-rate for a given cache size. It has five main steps for each reference x :

- (1) Search stack to find the current location of x (if any)
- (2) Calculate the distance d from the top of the stack
- (3) Update cache miss counters for all sizes $\geq d$
- (4) Update cache hit counters for all sizes $\leq d$
- (5) Push x onto the stack and remove the last location of x (if any)

request	time	stack	stack dist.
a	1	a	∞
b	2	b,a	∞
c	3	c,b,a	∞
d	4	d,c,b,a	∞
a	5	a,d,c,b	3 (d,c,b)
a	6	a,d,c,b	0 (none)
d	7	d,a,c,b	1 (a)
b	8	b,d,a,c	3 (d,a,c)

Table 1: Example stack distance calculation for a set of requests

The algorithm runs in $O(N * M)$ time, where N is total number of references and M is unique number of references. Many works improve on this algorithm by using more efficient data structures, sampling, and estimated stack distances.

Fang et. al. [11] showed that the stack distance distribution of a program’s typical workload could be predicted accurately given a significantly smaller version of the workload. They predicted the stack distance distribution for the *reference* input set of the SPEC2K benchmark using only the *train* and *test* input sets. These results encouraged researchers to move towards sampling the references in order to build the distribution while still achieving accurate stack distance distributions.

3.2 Tree

In tree-based stack distance calculations we exploit the logarithmic search time in balanced tree structure. Methods in this class have developed towards approximation of distances by getting to a leaf node that represents a close enough stack distance.

Partial Sums Tree by Bennet et. al. [4]. Bennet introduces a partial sums tree based on binary vectors representing the stack distance of a reference. But first they introduce the concept of using a hash table to store the *previous* location p of a reference. Second, they maintain a binary vector, B , at time t for a given reference x with the following property:

$$B_t = \begin{cases} 1 & \text{if } x_p \neq x_i (i = p, p + 1, \dots, t) \\ 0 & \text{otherwise.} \end{cases}$$

Now the number of 1s in B_t is the stack distance of reference x . As t increases the idea is to create a hierarchy of partial sums out of B so that we don’t have to scan B each time we calculate the stack distance for x . For example, imagine that every 3 time intervals we add up all the 1s counted so far and store them as a partial sum B^i , then for every 3 partial sums B^i we make a new final sum. To calculate the stack distance now, we do a tree traversal from our current time t to the previous use time p .

Since each lookup in the hash table is $O(1)$ and each stack distance calculation is $O(\log(n))$, then for a trace of N references the running time is now $O(N * \log(n))$.

Interval tree by Almasi et. al. [1]. Almasi interprets Bennet’s binary vectors in a different manner. They use only the 0s from the vector, calling them holes. The calculation for stack distance now becomes:

$$stackdist(x) = t - p - holes(x_p)$$

Where t is current time, p is previous time, and $holes(x_p)$ represents the number of holes (0s) between t and p in B . Now we represent a hole’s index as h_i and let the interval $[h_i, h_j]$ represent the time between two consecutive holes.

Since these intervals of holes are non-overlapping, they can be represented as AVL or red-black trees. For N references this takes $O(N * \log(n))$, similar to Bennet’s asymptotic bound, but in practice it performs significantly better because the tree is well balanced. Although we are still at the point of using a large amount of memory, $O(M)$, where M is unique number of accesses.

Search Tree [18]. Olken proposed ordering the LRU stack as a binary tree based on access times so that an in-order traversal gives the LRU stack position. For a given reference at the root, the right side of the tree represents references used more recently and the left side represents older references. At each node in the tree, we store a count of the number of nodes in the right subtree (more recent) and the number of nodes in the left subtree (older).

In order to calculate a stack distance, on a reference to x , we have to search the tree for its location and traverse our way to the top of the tree, recording the number of nodes in the right subtree of each node that we visited. In order to keep this tree balanced we should use an AVL or Splay tree as suggested by Olken. This algorithm runs in $O(N * \log(n))$ time and needs to store M number of nodes, where M is the number of unique accesses.

Scale Tree [26]. In Zhong et. al. they produce a modified version of Olken’s search tree by modifying it to support a *range* of times at each node. This maps several references to one node and allows for much faster searches. The amount of error incurred by using a range of times at each node is bounded by the range. The overall run time is reduced to $O(N * \log(\log(M)))$, but we still require $O(M)$ space.

SHARDS [23]. Spatially hashed reuse distances (stack distances), or SHARDS, is a recent advancement to constant space use. They achieve this by sampling accesses based on the hashed value of their location, hence spatially hashed. These samples are then put into an interval tree to compute their stack distance. After computing stack distance, it is recorded in the stack distance histogram so that we can solve for the miss-ratio curve. The achievement in this work is running in $O(1)$ through sampling and linear run time since it only requires one scan through the set of accesses.

3.3 Buckets

These algorithms employ some variant on reducing the number of stack positions that exist in the LRU stack. The key insight is to reduce n stack positions into m buckets while still maintaining the LRU stack order.

MIMIR by Saemundsson et. al. [20]. The MIMIR algorithm created fixed number of buckets to store references. Each bucket B maintained the LRU stack order property, that is B_i has lower stack distance than B_j since $i < j$. While buckets can vary in size, they introduce an *aging* procedure to ensure that buckets stay balanced. A common metric used to resize the buckets is the *average* stack distance.

The miss-ratio curve is generated from the estimated stack distance distribution. For any bucket we know that the actual stack distance lies somewhere between the sum of the size of the buckets before B_i , call it n and $n + sizeof(B_i)$. Now we create the normal stack distance histogram from the intervals and perform the integration to get the miss rate at for a cache size (stack distance) d .

Cliffhanger by Cidon et. al. [7]. The Cliffhanger approach uses the same MIMIR variable bucket algorithm, but their insight is to use *shadow queues*. Shadow queues can be thought of victim caches but without the value of the data, just the *key*. For example, if we

were to a miss in a cache with size 10, but then hit in the shadow queue of size 20 then that would correspond to a hit with cache size 30. Therefore, in the MIMIR algorithm each bucket corresponds to a shadow queue of size n .

In order to solve for the miss-ratio curve, they use *hill climbing* to incrementally build a miss-ratio curve. Hill climbing is a technique to find the local slope of a curve. Recall that a hit in the shadow queue corresponds to a miss in the cache, hence and increase in the miss-rate for that given cache size. Therefore, we can increase the size of this cache a small amount in order to combat future misses. In resource partitioning, this also means decreasing the size of a randomly chosen *different* cache. This process continues until there is no overall improvement in the total hit rate of the system.

3.4 Access Counters

These methods either use: (1) a set of counters to represent stack distances at each cache size or (2) a set of counters to represent *total* accesses and *unique* accesses.

Powers of 2 by [15]. First purposed by Kim et. al. in 1991, they use Mattson’s algorithm with two key differences:

- Counters are used representing cache sizes of 2^k
- Use a hash table to record previous stack location and stack distance for each reference

Therefore on reference x , we lookup its position and stack distance i . Then increment the counter corresponding to cache size i , this means that x would hit in all caches size $> i$. Now we have to push x on the top of the stack and update stack distances for all references in the stack, the worst case on this is M . Therefore the running time is in the same asymptotic bound, $O(N * M)$ as Mattson’s, but in practice it runs faster since it is insensitive to a program’s average stack distance.

This algorithm has been applied to build miss-ratio curves in Zhou et. al. [27]. They present a hardware and software implementation for calculating page miss-ratio curves to guide memory management in a multi-program environment, a known NP-hard problem.

At the hardware level, Tam et. al. [22] have used this algorithm to manage L2 cache partitioning. Here they create an access trace to L2 by recording the misses of L1, then they send the access trace log into the algorithm given by Kim. Their miss-ratio curve is then used to guide how much of the L2 cache should an application receive.

CounterStacks [24]. This recent development by Wires et. al. uses probabilistic counters (HyperLogLogs) and Bloom filters (for set inclusion) in order to maintain a list of unique accesses over a period of time. Specifically, they build a list for time t that contains the number of unique accesses. This allows us to keep track of how much a counter increased (stack distance) over a number of non-distinct accesses. An example is given in Table 2.

In order to derive the stack distance, we need to look at the *intra-counter* change, that is to find the last reference, we must find the newest counter. In Table 2 we find the last reference to a is at position 1, hence the stack distance of a lies at position (1,4) = 3.

CounterStacks uses $O(N * \log(M))$ to process a set of references, since we still need to find the last reference to each element. But

Table 2: Counter Stacks

a	b	c	a
1	2	3	3
		1	2
			1
			1

CounterStacks achieves a great reduction in space complexity, $O(\log(M))$ through its use of probabilistic counters which set a small bound on the overall error in stack distance calculation.

UMONs [19]. Qureshi et. al. developed utility monitor circuits (UMONs) that approximate hit rate vs. number of sets in an LRU cache. They use access counters for a fixed number of stack positions for each set. Since the cache obeys the LRU policy, we can get the hit rate for a given number of sets by reading the counters for n sets.

This process is used in partitioning caches by set per application. If application A has a higher utility for a number of sets (i.e. its hit rate curve is not as steep) then it will be assigned that number of sets. Application B will be left with the remaining number of sets.

Fractals [12]. L. He et. al. use a fractal equation to calculate the miss-rate of a given cache size from the time between two consecutive misses. This is known as the inter-miss gap described by Denning and related to LRU miss-rate by Ding in HOTL [9]. From the distribution of these inter-miss gaps we can get the LRU miss rate for a given cache size, by the following:

$$intermiss(c) = \frac{1}{missrate(c)}$$

Their collection inter-miss gaps used hardware counters to record a cache miss, the number of references n , and a new cache miss. n is then saved into a buffer. This model allows for online processing at the cost of a 2% slowdown.

Unfortunately, their model ignores cold misses which leads to inaccuracy (the cold miss equation is not fractal). On average in the SPEC2006 benchmark they report 76% percent miss-ratio curve accuracy.

PARDA [17]. Q. Niu et. al introduce the first explicit parallel stack distance algorithm based on Mattson’s stack processing. PARDA exploits the independences of the *define-use* chains in references. That is, since we only care about the reference to the last location, all previous references are independent. Therefore, we can create chunks that contain all references between the last use and the current use for a given access. The algorithm is outlined below:

- (1) Start at the end of the trace with reference x
- (2) Scan until you see another x
- (3) Send that interval off for processing via Mattson’s algorithm
- (4) Repeat with the next reference after x

request	time	last use	reuse time
a	1	∞	∞
b	2	∞	∞
c	3	∞	∞
d	4	∞	∞
a	5	1	4 (5-1)
a	6	5	1 (6-5)
d	7	4	3 (7-4)
b	8	2	6 (8-2)

Table 3: Example reuse time calculation for a set of requests

4 REUSE TIME ALGORITHMS

More recently, reuse time has emerged as a metric used to approximate miss-ratio curves. In 1968, Denning proposed that the reuse time distribution can yield the working set size at time t , but until recently it was hard to achieve accurate results with the analytical models available.

4.1 Reuse Time

Table 3 gives an example of the reuse time calculation. Note we do not need to maintain the LRU stack, but as a result the miss-rate for a given cache size is no longer an exact integration over the distribution of reuse times.

4.2 Histogram

These methods bin the data recorded into a histogram, and build a miss-ratio curve from that histogram by integrating the probability density function or summing the total frequencies over a given range.

Average Eviction Time [14]. Hu et. al. presents a series of kinetic equations related to average data eviction in the cache. Based on the reuse time histogram, they estimate the probability that reference x has reuse time greater than t , which is then related to a stack movement. From here we can solve for the *average* eviction time for a given cache size. The average eviction time for a cache now relates miss-rate to time, so we have the following equation:

$$\int_0^{AET(c)} P(t)dt = c$$

Where $P(x)$ is the probability that a reference has reuse time greater than t .

In order to gain constant space complexity, they perform sampling over random random intervals, keeping the amount of data in the reuse time histogram constant with respect to the number of requests.

Footprint [25]. Xiang et. al. introduces the *footprint* function as a mapping from an execution window to the volume of references during that time. For a given number of references n there are $\binom{n}{2}$ number of execution windows. Therefore, the *average* footprint is measured for a set of references. This can be done by measuring the following:

- Reuse Time Histogram
- The first use of every reference

- The last use of every reference

The miss-ratio curve under the footprint metric is constructed for a given cache size d as the fraction of reuse times that have an average footprint smaller than d . In HOTL, it was shown that the footprint metric is the cache locality theory equivalent of Denning’s working set size.

The footprint algorithm runs in $O(N)$ and $O(M)$ where N is the total number of accesses and M is the total unique accesses. The algorithm was recently applied in LAMA (Locality Aware Memory Allocation) [13], to guide memcached’s slab reallocation problem.

Hit-Evict Distribution [3]. Beckmann et. al. introduces two new histograms of information, hit and evict. The hit distribution is the frequency of hits for given reuse times. Likewise, the evict distribution is the frequency of evicts for given reuse times. The motivation behind this work is to reduce the dependence on the LRU stack property that many analytic models assume. In their work they create an *age* distribution based off of the reuse times. This age distribution when combined with the hit and evict distribution can give an accurate LRU model for cache miss rates.

4.3 Access Counters

These methods use counters to record reuse times for a particular set of references, these counters are then aggregated to form a reuse time distribution.

Statcache [5]. Statcache by Berg et. al. uses random sampling to select some memory references. For each memory reference it does the following:

- (1) Set a (hardware) watchpoint on the address and record current number of memory accesses at that time n .
- (2) On the watchpoint trap read the new current number of memory access N . The reuse time of this address is now $N - n$.
- (3) Update the reuse time distribution with the new reuse time.

The Statcache model assumes a cache with random replacement, because of this we can assume that the probability that a cache line is still in a cache after a cache miss is uniform. Since the model calculates the probability that a reference will cause a cache miss, by assuming miss-ratio does not change over time, we can fix miss-ratio and derive the probability that a number of references cause a miss from its reuse distance from their reuse time.

Statstack [10]. The Stackstack model builds off of the Statcache work, this time modeling an LRU cache. They define the *expected* stack distance of a reference with reuse time t to be the *average* stack distance of all references with reuse time t . The miss-ratio curve is then constructed by computing the expected stack distance of each reuse time weighted by their frequency, this gives us a stack distance distribution.

If one were to interpret the Statstack model’s relationship between average stack distance vs. time, then we would arrive at the kinetic equation present in the average eviction model (AET). Stackstack performs in the same time bound as the AET model, $O(N)$ time and $O(1)$ space due to their sampling techniques.

5 OPEN PROBLEMS

We have described the history of the miss-ratio curve construction techniques that rely on reuse time and stack distance measurements. In a broader scope, we have the enormous amount of research done in the cache behavior modeling domain. Fixing our limit on constructing practical miss-ratio curves and their applications, we find that results in the following areas would excite many researchers in this field.

- **To what extent can we relate logical access time to physical clock time?**

Currently, many models assume logical access time, that is each reference is considered a unique point in time and there are no inactive periods. However, in real systems it is often the case where the number of references per unit of time vary significantly. An investigation into how much error is produced when using physical clock time vs. the overhead caused by measuring logical time would give insights on how to better formulate these models for use on real systems.

- **How to extend these models to non-LRU caches?**

While some models could be used for alternative cache policies, there is a large reliance on the nice LRU stack property, this is especially painful since Intel's switch to RRIP policy has made cache modeling significantly more difficult. In latest work by Beckmann et. al. [3], they claim a general model, and that policy specific models will be the subject of future work.

- Despite the above remarks, the theory remains mature and well understood. **How can we leverage these new models in day-to-day systems?**

The implementation of these models leaves much to be desired, rather high overhead has stopped many of these miss-ratio curve guided techniques from making it to practice. What improvements can we expect with miss-ratio curve constructions being done in $O(1)$ space in an online and dynamic workload environment?

6 CONCLUSION

Our survey has covered the two main metrics, stack distance and reuse time, and their associated cache models from a miss-ratio curve construction standpoint. In Denning's 1968 work [8], he defined page residency as how long a page will stay in main memory. In the cache modeling field, we have called it the stack distance. More recently, it has been called the average eviction time of a cache. You could say that modeling locality has been reused over and over again.

REFERENCES

- [1] George Almási, Călin Caşcaval, and David A. Padua. 2002. Calculating Stack Distances Efficiently. *SIGPLAN Not.* 38, 2 supplement (June 2002), 37–43. <https://doi.org/10.1145/773039.773043>
- [2] Berk Atikoglu, Yuehai Xu, Eitan Frachtenberg, Song Jiang, and Mike Paleczny. 2012. Workload analysis of a large-scale key-value store. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 40. ACM, 53–64.
- [3] Nathan Beckmann and Daniel Sanchez. 2016. Modeling cache performance beyond lru. In *High Performance Computer Architecture (HPCA), 2016 IEEE International Symposium on*. IEEE, 225–236.
- [4] Bryan T Bennett and Vincent J. Kruskal. 1975. LRU stack processing. *IBM Journal of Research and Development* 19, 4 (1975), 353–357.
- [5] Erik Berg and Erik Hagersten. 2004. StatCache: a probabilistic approach to efficient and accurate data locality analysis. In *Performance Analysis of Systems and Software, 2004 IEEE International Symposium on-ISPASS*. IEEE, 20–27.
- [6] Damiano Carra and Pietro Michiardi. 2014. Memory partitioning in memcached: An experimental performance analysis. In *Communications (ICC), 2014 IEEE International Conference on*. IEEE, 1154–1159.
- [7] Asaf Cidon, Assaf Eisenman, Mohammad Alizadeh, and Sachin Katti. 2016. Cliffhanger: scaling performance cliffs in web memory caches. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*. 379–392.
- [8] Peter J Denning. 1968. The working set model for program behavior. *Commun. ACM* 11, 5 (1968), 323–333.
- [9] Chen Ding and Xiaoya Xiang. 2012. A higher order theory of locality. In *Proceedings of the 2012 ACM SIGPLAN Workshop on Memory Systems Performance and Correctness*. ACM, 68–69.
- [10] David Eklov and Erik Hagersten. 2010. StatStack: Efficient modeling of LRU caches. In *Performance Analysis of Systems & Software (ISPASS), 2010 IEEE International Symposium on*. IEEE, 55–65.
- [11] Changpeng Fang, S Can, Soner Onder, and Zhenlin Wang. 2005. Instruction based memory distance analysis and its application to optimization. In *Parallel Architectures and Compilation Techniques, 2005. PACT 2005. 14th International Conference on*. IEEE, 27–37.
- [12] L. He, Z. Yu, and H. Jin. 2012. FractalMRC: Online Cache Miss Rate Curve Prediction on Commodity Systems. In *2012 IEEE 26th International Parallel and Distributed Processing Symposium*. 1341–1351. <https://doi.org/10.1109/IPDPS.2012.121>
- [13] Xiameng Hu, Xiaolin Wang, Yechen Li, Lan Zhou, Yingwei Luo, Chen Ding, Song Jiang, and Zhenlin Wang. 2015. LAMA: Optimized locality-aware memory allocation for key-value cache. In *2015 USENIX Annual Technical Conference (USENIX ATC 15)*. 57–69.
- [14] Xiameng Hu, Xiaolin Wang, Lan Zhou, Yingwei Luo, Chen Ding, and Zhenlin Wang. 2016. Kinetic Modeling of Data Eviction in Cache. In *2016 USENIX Annual Technical Conference (USENIX ATC 16)*.
- [15] Yul H. Kim, Mark D. Hill, and David A. Wood. 1991. Implementing Stack Simulation for Highly-associative Memories. In *Proceedings of the 1991 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '91)*. ACM, New York, NY, USA, 212–213. <https://doi.org/10.1145/107971.107995>
- [16] Richard L. Mattson, Jan Gecsei, Donald R. Slutz, and Irving L. Traiger. 1970. Evaluation techniques for storage hierarchies. *IBM Systems journal* 9, 2 (1970), 78–117.
- [17] Q. Niu, J. Dinan, Q. Lu, and P. Sadayappan. 2012. PARDA: A Fast Parallel Reuse Distance Analysis Algorithm. In *2012 IEEE 26th International Parallel and Distributed Processing Symposium*. 1284–1294. <https://doi.org/10.1109/IPDPS.2012.117>
- [18] Frank Olken. 1981. *Efficient methods for calculating the success function of fixed-space replacement policies*. Technical Report. Lawrence Berkeley Lab., CA (USA).
- [19] Moinuddin K. Qureshi and Yale N. Patt. 2006. Utility-Based Cache Partitioning: A Low-Overhead, High-Performance, Runtime Mechanism to Partition Shared Caches. In *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 39)*. IEEE Computer Society, Washington, DC, USA, 423–432. <https://doi.org/10.1109/MICRO.2006.49>
- [20] Trausti Saemundsson, Hjortur Bjornsson, Gregory Chockler, and Ymir Vigfusson. 2014. Dynamic Performance Profiling of Cloud Caches. In *Proceedings of the ACM Symposium on Cloud Computing (SOCC '14)*. ACM, New York, NY, USA, Article 28, 14 pages. <https://doi.org/10.1145/2670979.2671007>
- [21] Xudong Shi, Feiqi Su, Jih-Kwon Peir, Ye Xia, and Zhen Yang. 2009. Modeling and stack simulation of CMP cache capacity and accessibility. *IEEE Transactions on Parallel and Distributed Systems* 20, 12 (2009), 1752–1763.
- [22] David K. Tam, Reza Azimi, Livio B. Soares, and Michael Stumm. 2009. RapidMRC: Approximating L2 Miss Rate Curves on Commodity Systems for Online Optimizations. In *Proceedings of the 14th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS XIV)*. ACM, New York, NY, USA, 121–132. <https://doi.org/10.1145/1508244.1508259>
- [23] Carl A Waldspurger, Nohhyun Park, Alexander T Garthwaite, and Irfan Ahmad. 2015. Efficient MRC Construction with SHARDS. In *FAST*. 95–110.
- [24] Jake Wires, Stephen Ingram, Zachary Drudi, Nicholas JA Harvey, Andrew Warfield, and Coho Data. 2014. Characterizing Storage Workloads with Counter Stacks. In *OSDI*. 335–349.
- [25] X. Xiang, B. Bao, C. Ding, and Y. Gao. 2011. Linear-time Modeling of Program Working Set in Shared Cache. In *2011 International Conference on Parallel Architectures and Compilation Techniques*. 350–360. <https://doi.org/10.1109/PACT.2011.66>
- [26] Yutao Zhong, Xipeng Shen, and Chen Ding. 2009. Program Locality Analysis Using Reuse Distance. *ACM Trans. Program. Lang. Syst.* 31, 6, Article 20 (Aug. 2009), 39 pages. <https://doi.org/10.1145/1552309.1552310>
- [27] Pin Zhou, Vivek Pandey, Jagadeesan Sundaresan, Anand Raghuraman, Yuanyuan Zhou, and Sanjeev Kumar. 2004. Dynamic tracking of page miss ratio curve for memory management. In *ACM SIGOPS Operating Systems Review*, Vol. 38. ACM, 177–188.