

# Can Increasing the Hit Ratio Hurt Cache Throughput?

Ziyue Qiu Juncheng Yang Mor Harchol-Balter  
Carnegie Mellon University

{ziyueqiu,juncheny,harchol}@cs.cmu.edu

## Abstract

Software caches are an intrinsic component of almost every computer system. Consequently, caching algorithms, particularly eviction policies, are the topic of many papers. Almost all these prior papers evaluate the caching algorithm based on its *hit ratio*, namely the fraction of requests that are found in the cache, as opposed to disk. The “hit ratio” is viewed as a proxy for traditional performance metrics like system throughput or response time. Intuitively it makes sense that higher hit ratio should lead to higher throughput (and lower response time), since more requests are found in the cache (low access time) as opposed to the disk (high access time).

This paper challenges this intuition. We show that increasing the hit ratio can actually *hurt* the throughput (and response time) for many caching algorithms. Our investigation follows a three-pronged approach involving (i) queueing modeling and analysis, (ii) implementation and measurement, and (iii) simulation to validate the accuracy of the queueing model. We also show that the phenomenon of throughput decreasing at higher hit ratios is likely to be more pronounced in future systems, where the trend is towards faster disks and higher numbers of cores per CPU.

**Keywords:** caches, hit ratio, performance evaluation, scalability, modification analysis, operational laws, queueing theory, LRU, eviction policies

## 1 Introduction

Software caches are widely deployed in today’s infrastructure. Examples range from simple and small page caches in laptops and mobile phones to large multi-layer distributed and heterogeneous key-value caches and object caches in the data centers, e.g., Meta Cachelib [23], Google CliqueMap [86], and include many types of caches in between [14, 17, 30, 95].

The purpose of the cache is to allow fast data access. This paper focuses on DRAM-based software caches. Typically, cached items can be accessed anywhere from 100 to 10,000 times faster than those on disk [85]. The principle of caching is very simple: Store items that are likely to be accessed soon in the cache. Store everything else on disk.

The goal of utilizing a cache is to improve *throughput*, the average number of requests served per second (RPS). This is particularly important in data processing applications where the goal is to process as many data requests as possible per unit time. Examples include the caches used in big-data systems such as Hadoop and HDFS [55], deep learning systems

such as Pytorch [76] and Alluxio [13], and databases such as RocksDB [44].

### 1.1 Cache eviction algorithms

All caches have a common component: the cache *eviction algorithm*. The eviction algorithm decides which item to evict when the cache is full. The *most common* cache eviction algorithm is Least-Recently-Used (LRU) [2, 10, 49, 72], which evicts the least recently *accessed* item in the cache. Because of its popularity, this paper will focus on LRU cache eviction. LRU is widely used because data access patterns often show locality where recently used data have a higher chance to get reused [42, 43]. Another common algorithm is First-In-First-Out (FIFO), which evicts the least recently *inserted* item, i.e., the oldest item. Besides LRU and FIFO, many advanced eviction algorithms have been designed [18, 21, 26, 45–47, 61–64, 70, 79, 87, 90, 91, 94, 99, 100, 111].

### 1.2 The quest for higher hit ratio

While the overall *performance goal* of a cache is to improve the request throughput and reduce the mean latency for requests, researchers have resorted to using the cache hit ratio as a *proxy* for measuring performance [23, 37, 38, 56, 69, 81, 82, 92, 95, 101, 102, 104, 112]. The *hit ratio* is defined as the fraction of requests that are found in the cache.

Maximizing the hit ratio makes intuitive sense for improving system performance since we want to maximize the fraction of accesses that can be completed quickly from the cache and minimize the fraction of accesses that need to go to the slow disk.

*But what if this intuition is wrong? What if increasing the hit ratio actually hurts performance?*

This is the question investigated in this paper. The cache eviction algorithms we evaluate are summarized in Table 1.

### 1.3 A three-pronged approach to determining if higher hit ratio helps

We take a three-pronged approach to determine if a higher hit ratio, in fact, improves throughput.

#### A. Queueing model for upper bounding throughput.

While many papers have analyzed caching policies in the past, all of these works have focused on analyzing the *hit ratio*, see for example [22, 24, 27, 31, 33, 34, 39, 41, 50–53, 65, 75, 80, 93]. The question of how the hit ratio affects throughput has been overlooked. Perhaps this is because it seems so obvious that increasing the hit ratio can only help.

**Table 1.** Table shows the algorithms we evaluated. Detailed descriptions of the algorithms are in Sec. 4.

Algorithm	Description	Production System	Our Findings: Does increasing the hit ratio always help?
LRU	When item is accessed it moves to front of queue. Evict the item at end of queue.	Alluxio [13], RocksDB [78], LevelDB [6]	no
FIFO	Evicts the oldest item.	ATS [19]	yes
Probabilistic LRU	Only moves accessed item to head of queue with some probability $1 - q$ .	HHVM [4]	depends on $q$
FIFO-Reinsertion a.k.a. CLOCK	Item gets second chance to go through queue before being evicted.	RocksDB [78]	yes
Segmented LRU	Uses two LRU queues to differentiate items that have been accessed twice.	Linux kernel [88]	no
S3-FIFO [106]	Uses a small FIFO queue to evict most new and unpopular objects.	Google [105]	yes

Instead we develop a queueing model of our DRAM-based software cache for a range of popular eviction policies, based on measurements from our implementation. We then use queueing theory to derive an *upper bound on the throughput* of this queueing model. An example of such an upper bound for the case of LRU is given in Figure 1 via the red solid line. While the analysis provides only an *upper bound* on throughput, it clearly shows that throughput first increases with hit ratio, then levels off, and then decreases.

**B. Simulation of the queueing model.** Because exact analysis of the queueing model is not possible, we next simulate the queueing model to obtain its exact throughput. For LRU, the result is shown in Figure 1 via the blue dotted line.

**C. Implementation of the caching system.** Finally we implement our caching system with a range of eviction algorithms. Our implementation is a prototype of Meta’s HHVM Cache [4, 5]. The Meta HHVM Cache is similar to many other in-memory caches, e.g., CacheLib [23], Memcached [7], Intel OCF [9], BCACHE [1], and RocksDB LRU Cache [78].

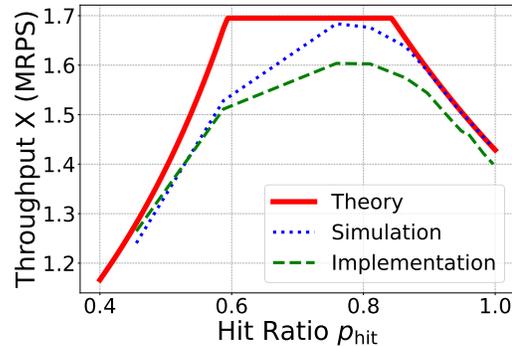
Additionally, to gain insights into evolving trends, our implementation includes a feature to emulate varying disk speeds and Multi-Programming Limits. This functionality is crucial for understanding how the hit ratio will affect throughput in future caching systems.

For LRU, the result of our implementation is shown via the green dashed line in Figure 1. Importantly, the simulation result is within 5% of the implementation result. This tells us the performance predicted by our queueing model provides a very good estimate of system performance.

**1.4 Contributions**

The contributions of this paper are summarized below:

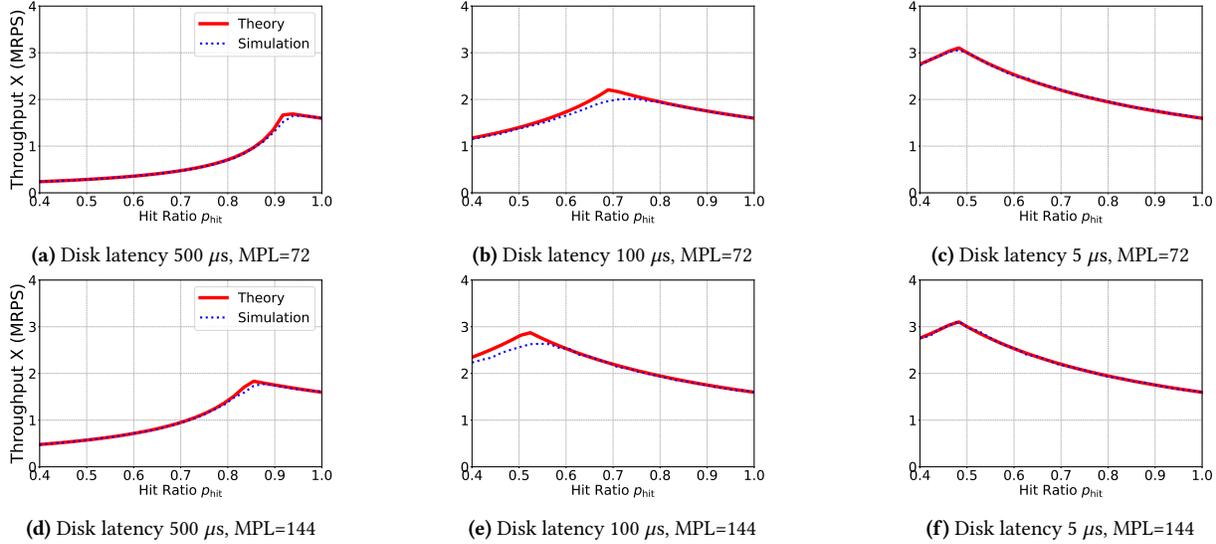
- This paper shows that increasing cache hit ratio can hurt throughput for many LRU-based cache eviction algorithms. We show this via a three-pronged approach, involving queueing theory, simulation, and implementation. This



**Figure 1.** Throughput under LRU (measured in millions of requests per second) increases as the hit ratio increases initially but then drops when the hit ratio gets high.

counterintuitive result will be explained in Section 3.2.1 after we discuss the relevant queueing model.

- We develop *queueing models* that allow us to understand the effect of the cache hit ratio on system throughput. This is done for six different caching policies. The modeling is non-trivial and, to the best of our knowledge, such models do not exist in prior work.
- While our queueing analysis only provides upper bounds, the analysis clearly indicates that throughput initially rises with hit ratio and then drops with hit ratio.
- We also *implement* many caching policies, including LRU, FIFO, Probabilistic LRU, and CLOCK. We validate the correctness of our queueing models, by *simulating* the queueing models and showing that the simulation results match the implementation results within 5% for all caching policies studied.
- We evaluate the effect of changing disk latency as we move from older disk speeds (500  $\mu$ s) to current disk speeds (100  $\mu$ s) to future disk speeds (5  $\mu$ s). An example of the effect of the *disk speed trend* is given in Figure 2, for the case of Segmented LRU (SLRU). As we move to future disk speeds



**Figure 2.** The throughput of a Segmented LRU cache is affected by the hit ratio, but also by the disk latency and the MPL.

- (move from left to right in the figure), the point at which throughput starts to deteriorate moves earlier and earlier.
- We also evaluate the effect of another trend, increasing the number of CPU cores, which allows more requests to be served concurrently – the Multi-Programming Level (MPL) increases. The effect of *increasing the MPL* is shown in Figure 2 when looking from top (MPL = 72) to bottom (MPL = 144). The point at which throughput starts to deteriorate moves earlier for higher MPL.

## 2 Scope of the paper

Our paper is limited to DRAM-based software caches.

### 2.1 DRAM-based software caches versus other caches

In a DRAM-based software cache, DRAM is used to cache the accessed data. This is in contrast to SSD-based software caches where the data are cached in an SSD. In an SSD-based cache, the cache access is 100 times slower and far less concurrent than in the DRAM system. Consequently, results are very different from DRAM-based software caches. We will explain this in more detail in Section 3.2.1.

Hardware caches are also outside the scope of this paper. While software caches are characterized by a global linked list that orders the cache items, hardware caches are typically set-associative caches (e.g., RRIP-based eviction algorithms [60, 96]). This means that the eviction algorithm is limited to evicting data in the same “set.” There is no concept of a global linked list and hence no software operations.

Network-connected caches are also outside the scope of this paper. Here the bottleneck is the time needed to move the data, rather than the time to manage software operations.

### 2.2 Hardware trends

The performance of DRAM-based software caches is highly dependent on the cores on which the software runs and the backend disks. Over the years, the number of cores increases dramatically while the speed and concurrency of backend disks have improved.

Before 2000, CPUs had only one core [8], but a modern CPU has 32-192 cores [15]. The increase in CPU cores enables better performance, allowing more concurrent requests; however, it also presents challenges because the cores need to coordinate with each other.

Meanwhile, the backend disk latency stagnated around 2-10 ms until the wide adoption of SSDs as the backend device. Today’s SSDs exhibit a wide range of latency characteristics. High-end SSDs have latencies around 5  $\mu$ s [58, 83]; low-end SSDs show latencies of a few hundred  $\mu$ s [16, 89], and most commercial SSDs have latency in between [59, 68, 71].

In addition to the reduced latency, SSDs today support massive concurrency [35, 109]. This means that requests at the disk can all be served in parallel without queueing.

## 3 A three-pronged approach to LRU

Because LRU is the most common caching policy, we focus on carefully understanding its performance. The purpose of this section is to explain why increasing the hit ratio can actually lead to decreased throughput for DRAM-based software caches. Our takeaways are summarized in Section 3.5.

### 3.1 A closed-loop queueing model of LRU

As with many caching systems evaluations and benchmarks, e.g., Cachelib [23], YCSB benchmark [40], S3-FIFO [106], and FrozenHot [77], our system is best modeled by a *closed-loop queueing model*, where new requests are triggered by

the completion of previous requests. There is a fixed Multi-Programming Limit (MPL),  $N$ , denoting the number of requests that can be in the system at a time (this is dictated by the number of cores – in our case 72).

### 3.1.1 Modeling concurrency in a closed-loop model.

Each request is handled by a single core. The total number of requests in the system is thus limited by the total number of cores. Throughout, we assume that there is one CPU with 72 cores (this will match our experimental setup), and thus we can process 72 requests concurrently.

We thus model our caching system via a closed-loop queueing model, where a new request is allowed to enter only when some other request completes. The *multi-programming limit* (MPL) for the system is  $N = 72$ . See [57, Chapters 2,6,7] for background on modeling closed-loop queueing models.

### 3.1.2 Modeling disk access and cache access.

Our disk has enough concurrency that it can be accessed simultaneously by all 72 requests. Thus, in queueing speak, we can model the disk as a *think station* (infinite number of simultaneous service stations) with mean think time  $E[Z_{disk}] = 100\mu s$ . Likewise, the cache lookup can also be executed concurrently. Thus, the cache lookup can also be modeled as a “think” station but with a much faster mean think time of  $E[Z_{cache}] = 0.51\mu s$ . Note that a think station is different from a queue station in that there is no queueing at a think station – every request is served immediately and concurrently.

### 3.1.3 Modeling software global list operations.

In an LRU cache, all cached items are stored in a single *global linked list*, where the least-recently-used item is the “tail” item and is the one to evict, while the most-recently-used item is at the “head” of the list. A request for some item  $d$  first goes to the cache to look for  $d$ . Either the request finds  $d$  in the cache (called a “hit”), or it does not (a “miss”). The probability of a hit is denoted by  $p_{hit}$  and the probability of a miss is  $p_{miss}$ , where  $p_{hit} + p_{miss} = 1$ .

If the request is a hit, then two things need to happen:

1. The item  $d$  must be delinked from its position in the global linked list. This is the *delink operation*. The delink time is denoted by the service time random variable  $S_{delink}$ .
2. The item  $d$  needs to be attached to the head of the global linked list. This is called the *cache head update*. The head update time is denoted by the random variable  $S_{head}$ .

Note that the actual “reading time” (the time to read a 4KB block) is not included in our model. The reason is that this is very small compared with cache lookup, is handled concurrently, and is the same across all algorithms.

If the request is a miss, then three things need to happen:

1. The item  $d$  needs to be found on disk.
2. The least-recently-used item (the one at the tail of the global list) needs to be removed. We call this the *cache tail update* and denote it by the random variable  $S_{tail}$ .

3. Item  $d$  needs to be attached to the head of the global linked list. This is the *cache head update* mentioned earlier, denoted by random variable  $S_{head}$ .

We can thus model the LRU Caching system via the queueing model shown in Figure 3.

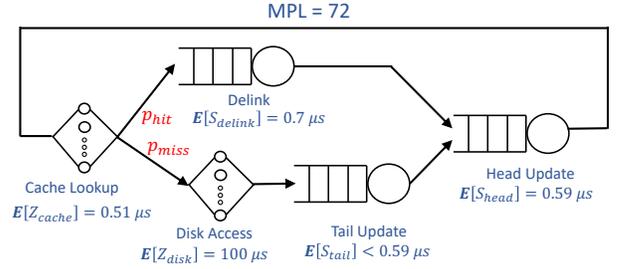


Figure 3. Queueing model of LRU cache.

Understanding how to properly model  $S_{head}$ ,  $S_{tail}$ , and  $S_{delink}$  is important and non-trivial and will come up again when discussing different eviction policies. We will explain these here. We start with  $S_{head}$ , which denotes the time needed for a head update on the global linked list. This has *two components*. The first component is obvious: just the update to the global list. The second component is less obvious: communicating this update to all the other requests in the queue. Specifically, the core performing the head update needs to communicate with all the other cores in the head update queue to alert them that this head update is happening. Obviously, if there are many requests in the head update queue, then this communication will take longer.

In summary,  $S_{head}$  consists of a *constant* head update time and a communication time dependent on *queue length*. When the head update is the bottleneck operation, our measurements show that these two components add up to  $E[S_{head}] = 0.59\mu s$ . When the head update is not the bottleneck operation, the queue length at the head update queue drops, thus  $E[S_{head}]$  decreases. However, as we’ll see in the analysis of LRU (Section 3.2), in the case where an operation is not a bottleneck, its service time does not impact overall throughput much. Specifically, any value in  $0 < E[S_{head}] < 0.59\mu s$  will produce the same throughput if a head update is not the bottleneck operation. We find that  $S_{head}$  follows approximately a Bounded Pareto distribution (with  $\alpha = 0.45$ ) ranging from  $0.1\mu s$  to  $1.2\mu s$ . However, as we’ll see in Section 3.2, the throughput analysis of the queueing model is only influenced by the mean,  $E[S_{head}]$ . Similarly, when the delink queue is the bottleneck,  $E[S_{delink}] = 0.7\mu s$ . When the delink queue is not the bottleneck, its value has little impact on analysis.

*Importantly*, when we look at other eviction algorithms, the number for  $E[S_{head}]$  can change because the queue length at the head update queue will change.

To measure  $E[S_{head}]$ , we create a setting where the head update is the bottleneck operation. This means that the head update queue will be flooded. Consequently the service time,  $S_{head}$ , is simply the inter-departure time from this flooded

queue (namely the time between consecutive departures from the queue), which is easy to measure. Likewise, it is easy to measure  $E[S_{delink}]$ , because that device can also be made into the bottleneck (by setting  $p_{hit}$  appropriately).

The case of  $S_{tail}$  is slightly different because the tail update is *never* the bottleneck operation. This makes it difficult to precisely measure  $E[S_{tail}]$  because we can't keep the tail update server fully utilized as we were able to do for the other servers. Fortunately, again, as shown in Section 3.2, the precise value will not matter.

Before ending the section, we note that there is an alternative implementation possible for LRU, used in systems such as Varnish [11] and Alluxio [13], that combines the delink, tail update, and head update operations into a single queue. However, we chose to use a three-queue implementation because it allows higher concurrency and higher throughput. The three-queue implementation is 33% faster than the single-queue implementation for high hit ratios at  $N = MPL = 72$ .

### 3.2 Analysis of LRU queueing model

The goal of this section is to determine analytically how the system throughput is affected by hit ratio,  $p_{hit}$ . We first illustrate our analysis assuming that mean disk latency is  $E[Z_{disk}] = 100\mu s$  and then explain how the result generalizes to other disk latencies. Our analysis of closed systems is based on [57, Chapters 6,7]) and produces an upper bound on throughput for the queueing network in Figure 3.

The first step is to determine the *mean think time* of the system,  $E[Z]$ , where  $E[Z]$  is the mean time spent on accesses that can be executed concurrently by all cores. This includes cache lookup and disk access:

$$\begin{aligned} E[Z] &= E[Z_{cache}] + p_{miss} \cdot E[Z_{disk}] \\ &= 0.51 + p_{miss} \cdot 100 = 100.51 - 100p_{hit} \end{aligned}$$

For each queue, we now compute the *device demand*, which is the expected service demand on the corresponding server, per request into the system. The device demands are:

$$\begin{aligned} D_{delink} &= p_{hit} \cdot 0.7 \\ D_{tail} &< (1 - p_{hit}) \cdot 0.59 \\ D_{head} &= 0.59 \end{aligned}$$

The *total demand*,  $D$ , is the sum of the device demands:

$$D = D_{delink} + D_{tail} + D_{head}$$

Because  $0 < D_{tail} < 0.59$ , we have upper and lower bounds on  $D$  as follows:

$$\begin{aligned} 0.7p_{hit} + 0.59 &< D < 0.7p_{hit} + 0.59(1 - p_{hit}) + 0.59 \\ 0.7p_{hit} + 0.59 &< D < 0.11p_{hit} + 1.18 \end{aligned}$$

The next step is to determine the *bottleneck device*, which is the device with the highest demand. We can see that the bottleneck device is the delink device if  $p_{hit}$  is sufficiently

high, specifically  $p_{hit} > 0.84$ . Otherwise, the bottleneck device is the head update device. We write this as:

$$D_{max} = \max(0.59, 0.7p_{hit}) = \begin{cases} 0.59 & \text{if } p_{hit} < 0.84 \\ 0.7p_{hit} & \text{if } p_{hit} > 0.84 \end{cases}$$

We use  $X$  to denote system throughput. From [57, Theorem 7.1], we know that  $X$  is upper-bounded by two terms, as follows:

$$X \leq \min\left(\frac{N}{D + E[Z]}, \frac{1}{D_{max}}\right).$$

Substituting in the expressions for  $E[Z]$ , and  $D_{max}$  that we have already derived, as well as the lower bound on  $D$  and the fact that  $N = MPL = 72$ , we have that, for the case of  $E[Z_{disk}] = 100\mu s$ :

$$X_{LRU} \leq \min\left(\frac{72}{101.1 - 99.3p_{hit}}, \frac{1}{\max(0.59, 0.7p_{hit})}\right) \quad (1)$$

Equation (1) represents an *upper bound on throughput*, shown in red in Figure 4(b). This turns out to be a very good bound on the measured throughput from our implementation. When  $p_{hit} < 0.59$ , the first term in (1) is the relevant bound (minimum term). When  $0.59 < p_{hit} < 0.84$ , the second term in (1) is the relevant bound, where the max term in the denominator is 0.59. When  $p_{hit} > 0.84$ , the second term in (1) is again the relevant bound, but the max term in the denominator is  $0.7p_{hit}$ .

Recall that  $0 < E[S_{tail}] < 0.59$ . In the above analysis, we assumed that  $E[S_{tail}] = 0$  because we wanted an upper bound on  $X$ . If instead, we had used any value of  $E[S_{tail}]$  in the range between 0 and 0.59, the impact on our result for  $X$  would be very small ( $< 0.5\%$ ). The reason is that changing  $E[S_{tail}]$  would only change  $D$ , not  $D_{max}$ . Hence only the first term in (1) would change, and we only care about this first term when  $p_{hit}$  is low, specifically  $p_{hit} < 0.59$ . Within this first term,  $D + E[Z]$  would change from  $101.1 - 99.3p_{hit}$  to  $101.69 - 99.89p_{hit}$ .

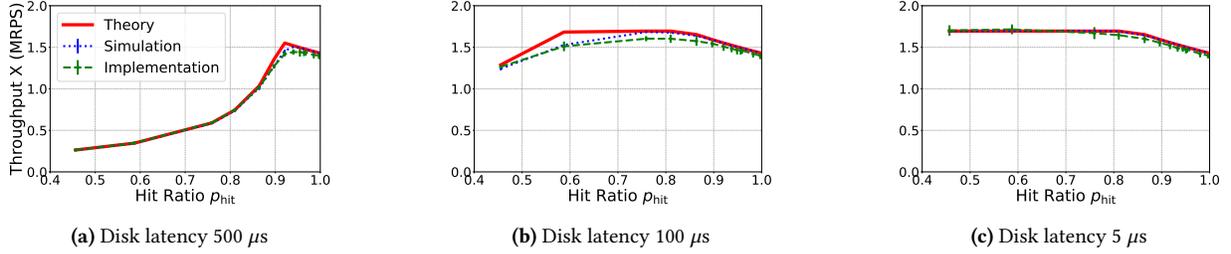
The above analysis assumed that  $E[Z_{disk}] = 100\mu s$ . If we redo the analysis for the case where  $E[Z_{disk}] = 5\mu s$ , the throughput is (2), as shown in red in Figure 4(c).

$$X_{LRU} \leq \min\left(\frac{72}{6.1 - 4.3p_{hit}}, \frac{1}{\max(0.59, 0.7p_{hit})}\right) \quad (2)$$

If we repeat the analysis for the case where  $E[Z_{disk}] = 500\mu s$ , we get (3) shown in red in Figure 4(a).

$$X_{LRU} \leq \min\left(\frac{72}{501.1 - 499.3p_{hit}}, \frac{1}{\max(0.59, 0.7p_{hit})}\right) \quad (3)$$

**3.2.1 Discussion of the analytic results.** We've seen that increasing the hit ratio leads to lower throughput when the hit ratio is high. We have shown via a queueing analysis why this happens. From a more intuitive perspective, when the hit ratio is high, we see that the *delink* operation becomes



**Figure 4.** Results for theory, implementation, and simulation under an LRU cache. For all three curves, the throughput of the LRU cache decreases at higher hit ratios. This trend becomes more pronounced as we move towards lower disk latencies – from (a) to (b) to (c).

the bottleneck. Hence almost all requests are queued behind the delink server in Figure 3. Thus, while it seems that we are saving time by not going to disk, we instead are wasting time by having to queue up at the delink device. Thus requests can actually take *longer*. This longer response time translates to a drop in throughput.

In all the curves of Figure 4, we find that there is some point,  $p_{hit}^*$ , after which increasing the hit ratio only hurts. This point  $p_{hit}^*$  decreases as the mean disk latency decreases. When the mean disk latency is really low (Figure 4(c)), increasing the hit ratio never helps! Thus our message about not blindly increasing the hit ratio will become more and more valid as we move to faster disks.

Throughout, we have looked at throughput, but we could instead have looked at *mean response time*, namely the time from when a request is submitted until it completes. Given that we have a closed-loop setting, mean response time and throughput are inversely related (see [57, Chapters 6,7]). Hence mean response time *increases* for higher hit ratios.

**3.2.2 How the analysis changes for SSD caches.** In an SSD cache, the queueing network looks similar to Figure 3, but cache lookup and disk access can no longer be modeled as think stations. Because the SSD hardware has much lower concurrency than DRAM, cache lookup is better modeled via a queue of waiting requests. Likewise, the backend disk should be modeled as a queue due to its low concurrency. The device demands on the cache lookup and the disk access are now much higher than that of delink, tail update, and head update. Hence throughput is dominated by the disk access time. Thus increasing the hit ratio never hurts throughput.

### 3.3 Simulation evaluation

Recall that our queueing analysis from Section 3.2 only provides upper bounds on throughput. To get the exact throughput of the queueing network, we turn to simulation.

We use an event-driven simulation. We model the service time for each operation as exponentially distributed with the appropriate mean (where the mean comes from the implementation measurements). While we chose the service time distribution arbitrarily, we note that we tried other distributions and found that the results were insensitive to the

particular service time distribution used. This is consistent with the findings in [84] for closed-loop models.

The results of our simulation are shown in Figure 4 via dotted blue lines. We also show 95% confidence intervals but they are typically too small to be visible. As expected, the results of simulation lie below the red theory lines, which represent the theoretical upper bound.

### 3.4 Implementation Setup and Results

In both the analysis (Section 3.2) and the simulation (Section 3.3), we were evaluating the queueing network in Figure 3. We now study LRU via our implementation that is independent of any queueing network.

Our implementation is a prototype based on Meta’s HHVM Cache. Figure 4 shows the results of our LRU implementation via green dashed lines with 95% confidence intervals.

**3.4.1 Implementation and experimental setup.** Our experiments use dual-socket servers with Intel Xeon Platinum 8360Y 36-core processors (Ubuntu 20) at 2.4GHz, running on CloudLab platform [3]. To avoid Non-Uniform Memory Access (NUMA) impacts, our evaluations only use a single socket with hyperthreading enabled. To provide consistent results, we disable turbo-boosting and fix the per-core frequency at 3.1 GHz.

Our Intel Xeon Platinum CPU allows for 72 cores to run concurrently. This limits the number of requests that can be in the system at once to 72. Each request accesses a 4KB block of data, the common size in database block caches. We emulate three different disk speeds: 500 $\mu s$ , 100 $\mu s$ , and 5 $\mu s$ .

In our experiments, we consider  $p_{hit}$  in the range of [0.4, 1], with a step size of 0.05 in most cases, but a step size of 0.02 for higher  $p_{hit}$  values. Each experiment is run 20 times, and we determine 95% confidence intervals in each case.

**3.4.2 Workload creation.** For request generation, we employ 72 client threads, where each thread is assigned to a single CPU core. We use a synthetic popularity distribution following the Zipfian parameter  $\theta = 0.99$ , representative of cache accesses from e-Commerce websites [36] and social networks [103]. Recognizing that our goal is to assess the impact of hit ratio on throughput, and that the popularity distribution only affects the hit ratio, we determine that it

is sufficient to test with this well-established Zipfian distribution without the need for a broader range of models or real-world access traces. Throughput measurements are conducted after some warmup period, when the cache is full.

**3.4.3 Results of implementation.** Our implementation results and simulation results are always within 5%.

### 3.5 Takeaways

We started the section by presenting a queueing model of our LRU caching system (Section 3.1). We were able to derive an *upper bound* on throughput in our model as a function of the hit ratio (Section 3.2). Our analysis elucidated that when the hit ratio gets high, the delink operation becomes the bottleneck in the queueing network, resulting in longer delays and lower throughput. We next simulated the queueing network (Section 3.3), which allowed us to determine the exact throughput as a function of hit ratio. Finally, in Section 3.4, we implemented our LRU caching system and showed that the results are always within 5% of the simulation.

There are two takeaways. First, because the implementation matches the simulation, we conclude that our queueing model is an excellent representation of the real system, at least with respect to understanding system throughput as a function of hit ratio. Second, we see that a very simple queueing analysis enables us to easily predict the point at which increasing the hit ratio starts to hurt the throughput. This foreshadows a *theme of this paper* – queueing analysis alone suffices to predict the behavior of throughput as a function of hit ratio for many eviction policies.

## 4 Evaluation of other policies

In this section, we will apply our three-pronged approach to the remaining algorithms in Table 1. Our goal is to determine how the throughput is affected by the hit ratio,  $p_{hit}$ .

### 4.1 FIFO

In the *FIFO policy* we store all cached objects into a single global linked list. The list maintains the same ordering the entire time, with new items being added to the list head and the oldest item dropping off the tail.

If a request for item  $d$  is a cache hit, *nothing* happens to the global linked list. When the request is a miss, the list must be updated. Specifically:

1. The item,  $d$ , needs to be read from disk.
2. The oldest item needs to be removed from the global list, requiring a cache tail update.
3. Item  $d$  needs to be attached to the head of the list, requiring a cache head update.

We can thus model the FIFO caching system via the closed-loop queueing model in Figure 5. It may seem strange that  $E[S_{head}]$  is higher in the FIFO system (Figure 5) than in the LRU system (Figure 3). To understand why this happens, recall from Section 3 that there are two components

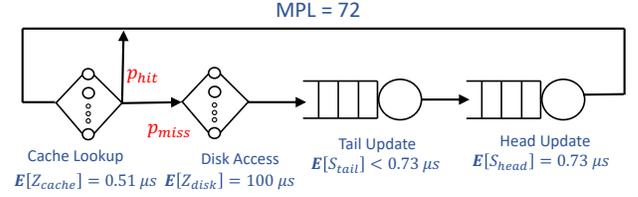


Figure 5. Queueing model of FIFO cache.

to  $E[S_{head}]$ : (i) a constant time needed for a head update to the global linked list, and (ii) a communication time proportional to the queue length at the head update queue. Now, in LRU there are 72 jobs in the system, split among *three* queues, but in FIFO, the 72 jobs are split among only *two* queues. Consequently the expected queue length of each of FIFO's two queues is larger than each of LRU's three queues. This in turn means that component (ii) is higher under FIFO, explaining why  $E[S_{head}]$  is larger under FIFO.

**Analysis of the queueing network model.** We now follow the same approach that we used for LRU to analyze our closed queueing network.

Again the *mean think time* of the system is:

$$E[Z] = E[Z_{cache}] + p_{miss} \cdot E[Z_{disk}] = 100.51 - 100p_{hit}$$

For each queue, we now compute the *device demand*:

$$D_{tail} < (1 - p_{hit}) \cdot 0.73 \quad D_{head} = (1 - p_{hit}) \cdot 0.73$$

The *total demand*,  $D$ , is the sum of the device demands:

$$D = D_{tail} + D_{head}$$

Because  $0 < D_{tail} < (1 - p_{hit}) \cdot 0.73$ , we have upper and lower bounds on  $D$  as follows:

$$(1 - p_{hit}) \cdot 0.73 < D < (1 - p_{hit}) \cdot 0.73 + (1 - p_{hit}) \cdot 0.73$$

$$0.73 - 0.73p_{hit} < D < 1.46 - 1.46p_{hit}$$

The *bottleneck device* is always the head update, so:

$$D_{max} = 0.73 - 0.73p_{hit}.$$

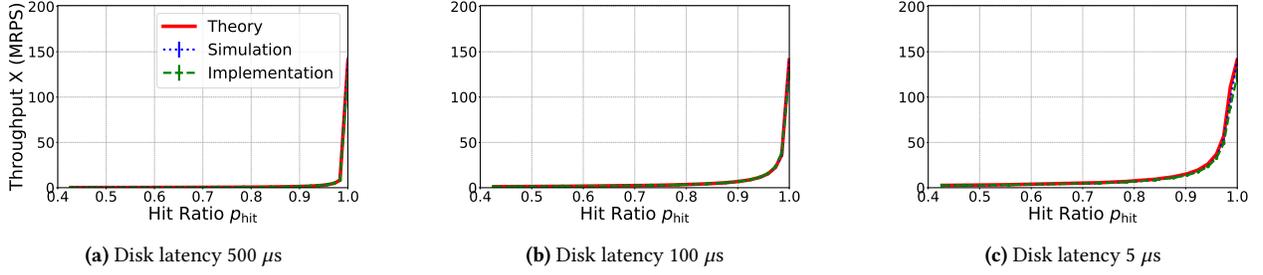
Given all the above terms, where  $E[Z_{disk}] = 100\mu s$ , from [57, Theorem 7.1] our upper bound on throughput  $X$  is:

$$X_{FIFO} \leq \min \left( \frac{72}{101.24 - 100.73p_{hit}}, \frac{1}{0.73 - 0.73p_{hit}} \right) \quad (4)$$

Equation (4) represents an *upper bound on throughput*, shown in red Figure 6(b). Recall that  $0 < E[S_{tail}] < 0.73$ . In the above analysis, we assumed that  $E[S_{tail}] = 0$  because we wanted an upper bound on  $X$ . If instead, we had used any value of  $E[S_{tail}]$  in the range between 0 and 0.73,  $X$  would only change by  $< 0.5\%$ .

What's interesting about (4) is that for both terms in the bound of  $X$ , increasing  $p_{hit}$  increases the throughput. This contrasts with the expressions of throughput for LRU, where increasing  $p_{hit}$  decreased the second term in the bound of  $X$ .

The above analysis assumed that  $E[Z_{disk}] = 100\mu s$ . We can likewise redo the analysis for the case where  $E[Z_{disk}] = 5\mu s$ , obtaining:



**Figure 6.** Results for theory, implementation and simulation under a FIFO cache. For all three curves, the throughput of the FIFO cache always increases at higher hit ratios under different disk latencies.

$$X_{\text{FIFO}} \leq \min \left( \frac{72}{6.24 - 5.73p_{\text{hit}}}, \frac{1}{0.73 - 0.73p_{\text{hit}}} \right) \quad (5)$$

Likewise, when  $E[Z_{\text{disk}}] = 500\mu\text{s}$ , we obtain:

$$X_{\text{FIFO}} \leq \min \left( \frac{72}{501.24 - 500.73p_{\text{hit}}}, \frac{1}{0.73 - 0.73p_{\text{hit}}} \right) \quad (6)$$

The upper bounds in (4), (5), and (6) are shown via the red solid lines in Figure 6.

**Results of the three-pronged approach.** Our analysis for FIFO shows that increasing the hit ratio *always* leads to higher throughput, regardless of the mean disk latency. The queueing theory shows this mathematically. More intuitively, the bottleneck device (the head update) is now in the *miss* path, not in the hit path. Therefore, increasing the hit ratio does not result in increased demand on the bottleneck device, and hence does not lead to deleterious effects on throughput. Our simulation of the queueing network follows the same process as Section 3.3, and our implementation follows the process of Section 3.4. Results of simulation are shown in blue dotted lines and results of implementation are shown via green dashed lines in Figure 6. These agree within 5%.

#### 4.2 Probabilistic LRU

Under Probabilistic LRU there is an additional parameter  $q$  that controls how close the algorithm is to LRU (lower  $q$ ) versus FIFO (higher  $q$ ). As always, there is a global linked list; however, now the ordering of items in the linked list is a mixture of FIFO and LRU.

If a request for item  $d$  is a hit, then, with probability  $q$  *nothing* happens to the global linked list (as in FIFO). However, with probability  $1 - q$ , we follow LRU, where:

1. The item  $d$  needs to be unlinked from its position in the global linked list (a delink operation).
2. The item  $d$  needs to be attached to the head of the global linked list (a head update operation).

If the request is a miss, then three things need to happen:

1. The item  $d$  needs to be read from disk.
2. The item at the tail of the global linked list needs to be removed (a tail update operation).

3. Item  $d$  needs to be attached to the head of the global linked list (a head update operation).

The service times for the usual operations turn out to be slightly affected by the value of  $q$ . This is because  $q$  affects the queue lengths, hence affecting the communication overhead as explained in Sections 3.1.3. The queueing network for Probabilistic LRU is shown in the case of  $q = 0.5$  (Figure 7(a)) and  $q = 1 - \frac{1}{72}$  (Figure 7(b)). We chose these values because it turns out that  $q$  has to be extremely high,  $\geq 1 - \frac{1}{N}$ , to show FIFO-like behavior. For all other values of  $q$ , our analysis (and implementation), shows LRU-like behavior. This is an interesting finding in its own right.

**Analysis of the queueing network model.** Analysis is deferred to the supplemental document.

**Results of the three-pronged approach.** Figures 8 and 9 show the results of analysis (red solid lines), simulation (blue dotted lines) and implementation (green dashed lines). Again simulation and implementation agree within 5% and the analytic upper bound provides an excellent indication of the trends.

The behavior of Probabilistic LRU is highly dependent on the  $q$  parameter. When  $q$  is *not very high*, we see that the throughput starts decreasing beyond hit ratio  $p_{\text{hit}}^*$ . The queueing theory shows this fact mathematically. More intuitively, given that  $q$  is not high, many requests need to go through the delink queue, which becomes the bottleneck, resulting in higher queueing times, and reduced throughput.

By contrast, when  $q$  is *very high*, we basically skip the delink operation. Hence our network behaves very similarly to FIFO. Now the bottleneck device is on the *miss* path, and hence is not affected by increasing the hit ratio.

#### 4.3 CLOCK

The CLOCK cache eviction policy operates a global linked list, ordered like FIFO, however every object that is accessed gets a “second chance” to live before eviction. As in FIFO, each object  $d$  is appended to the head of the linked list and moves down towards the tail of the list as new objects are appended to the head. Each object is given a special bit set to 0. If object  $d$  gets to the tail of the list, and its bit is still 0, then it is removed from the tail. However, if  $d$  is accessed before it gets to the tail, then  $d$ 's bit is set to 1. Note that  $d$ 's

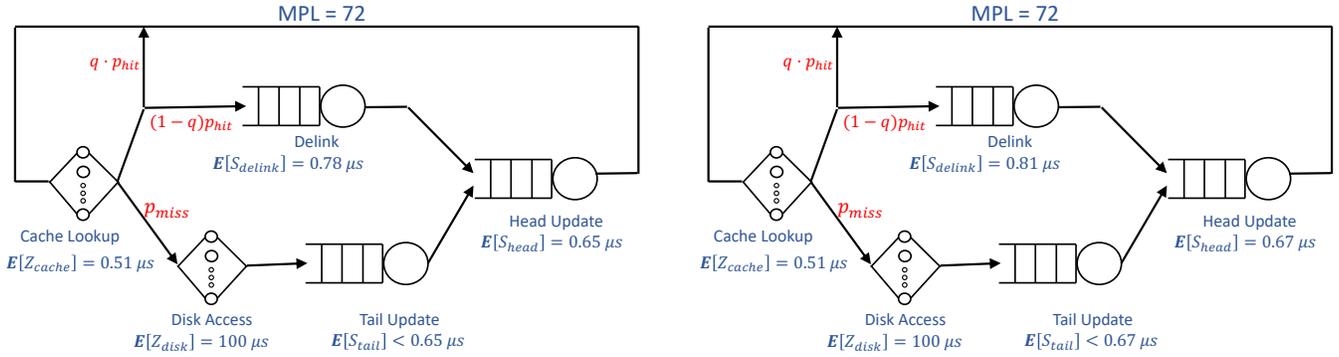


Figure 7. (Left) queueing model of Probabilistic LRU cache with  $q = 0.5$ ; (Right) queueing model with  $q = 1 - \frac{1}{72}$ .

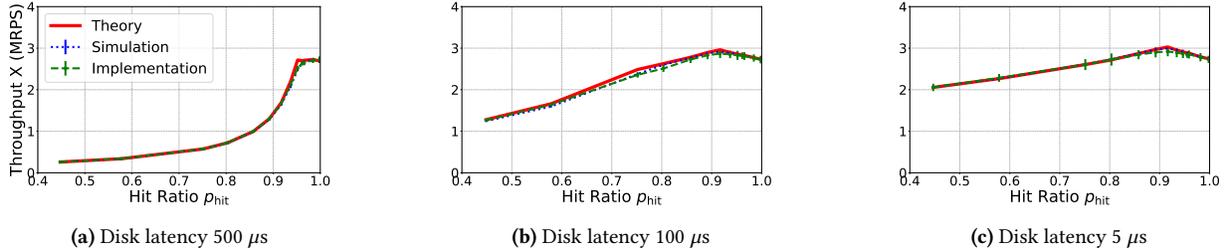


Figure 8. Results for theory, implementation and simulation under a Probabilistic LRU cache with  $q = 0.5$ . For all three curves, throughput decreases at higher hit ratios, under different disk latencies.

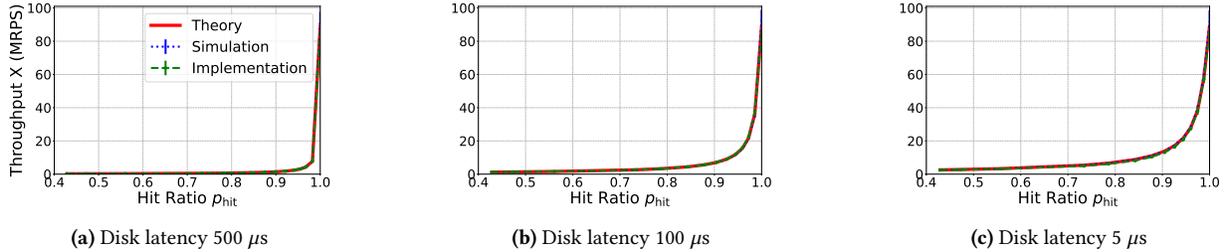


Figure 9. Results for theory, implementation and simulation under a Probabilistic LRU cache with  $q = 1 - \frac{1}{72} = 0.986$ . For all three curves, throughput always increases with a hit ratio, under different disk latencies.

position in the list does not change. When  $d$  gets to the tail of the list, because its bit is 1, it is skipped over for eviction, provided that there are other candidates with a 0 bit. The CLOCK queueing network is shown in Figure 10.

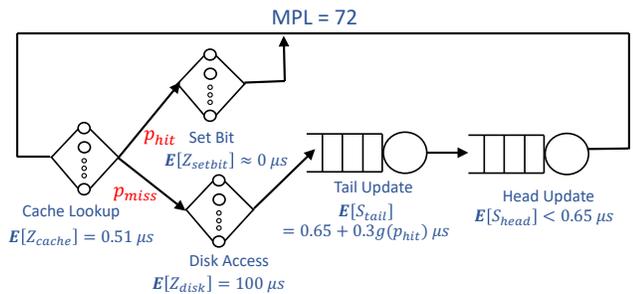


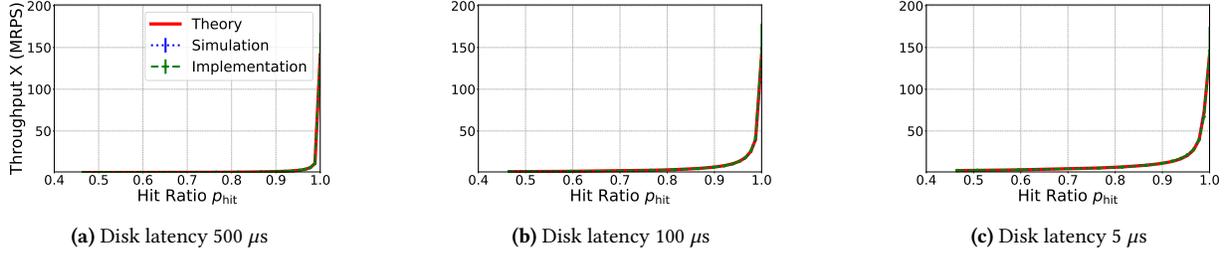
Figure 10. Queueing model of CLOCK cache.

**Analysis of the queueing network model.** Analysis is deferred to the supplemental document.

**Results of the three-pronged approach.** Figure 11 shows the result of the theory, simulation, and implementation of CLOCK. We see that increasing the hit ratio,  $p_{hit}$ , always leads to higher throughput, regardless of the mean disk latency. The queueing theory shows this fact mathematically. Intuitively, the bottleneck device (the tail update) is in the miss path, not in the hit path. Therefore, increasing the hit ratio this does not result in increased demand on the bottleneck device, and hence does not impinge on throughput.

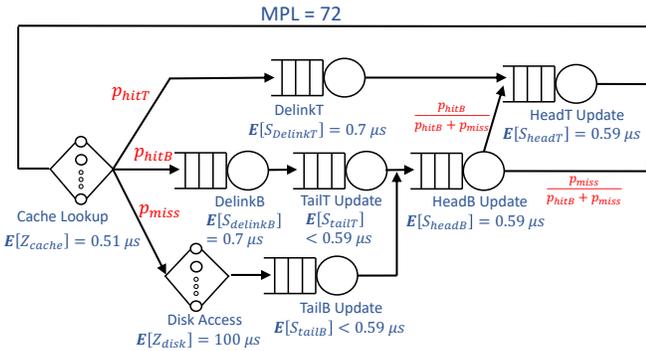
4.4 Segmented LRU

The Segmented LRU (SLRU) policy is one of the more advanced policies that researchers and practitioners use. The high-level idea in SLRU is that the global linked list of all objects in the cache is divided into two lists: the Probationary list (denoted B for bottom) and the Protected list (T for top). Objects initially enter the B list. If an object is never accessed after being put on the B list, it will eventually leave



**Figure 11.** Results for theory, implementation and simulation under a CLOCK cache. For all three curves, the throughput of the CLOCK cache always increases with hit ratio under different disk latencies.

the cache. However, if an object on the B List is accessed, then the object will be delinked from the B list and moved onto the T list. The T list is an LRU list in the sense that it is sorted from most recently accessed to least recently accessed. When an object on the T list is accessed, the object moves to the head of the T list. When an object leaves the T list, it is moved back to the B list and the process repeats.



**Figure 12.** Queueing model of Segmented LRU cache.

The queueing network for SLRU is shown in Figure 12. When an object  $d$  is first accessed, we do a cache lookup. There are now three cases of what might happen. If  $d$  is currently in the T list (which runs LRU), then  $d$  needs to be delinked from its current location and moved to the head of the B list (thus we have a delinkT operation followed a headT update). If  $d$  is currently in the B list, then  $d$  is delinked from the B list and moved to the head of the T list. When this happens, we need to remove the object that is at the tail of the T list. That object is moved to the head of the B list. Finally, if  $d$  is not in the cache (a “miss”), then we need to find  $d$  in the disk. After finding  $d$ , we put it on the head of the B list and remove the object at the tail of the B list.

To do a queueing analysis of the network in Figure 12, we need two more things: First, we need the service times of potential bottleneck operations. These are the same as the numbers in the LRU network. Second, and more challenging, we need to understand the fraction of time that an object is found on the T list as opposed to the B list. This is obviously a function of  $p_{hit}$ . We used our simulation to estimate this function. The details are in the supplementary document.

**Results of analysis and simulation.** Figure 2 (see the front of the paper) shows the results of analysis (red solid lines)

and simulation (blue dotted lines) for Segmented LRU. We have not implemented SLRU.

Like LRU, SLRU also experiences reduced throughput with high hit ratios. The queueing theory shows this mathematically. More intuitively, the bottleneck here is often the delinkT operation. Therefore, increasing hit ratio makes more requests queue behind the delinkT server which is already bottlenecked, hence lowering the throughput. Note that this same behavior would happen if the DelinkB operation were the bottleneck. Thus using a more complex policy, like SLRU, which has more queues, does not alleviate the problem of having a bottleneck on the hit path.

In all the curves of Figure 2, we find that there is some point,  $p_{hit}^*$ , after which increasing the hit ratio only hurts. This point  $p_{hit}^*$  decreases as the mean disk latency decreases.

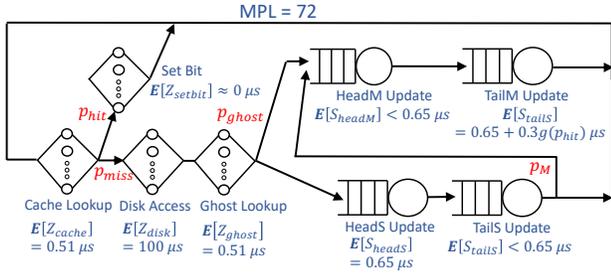
#### 4.5 S3-FIFO

The S3-FIFO policy is a new algorithm that claims to have state-of-the-art hit ratio. We have not implemented this policy, but will study it via analysis and simulation. S3-FIFO divides the global list of all elements in the cache into two global FIFO-ordered lists. The first global list is called the Small List, because it has a fixed short length, while the second global list is called the Main List. Typically, the Small List contains 10% of the items while the Main List contains the rest (we used these numbers in our model). The S3-FIFO algorithm is very similar to CLOCK in that only a miss causes work to be done, while a hit only involves setting a bit.

There is also a Ghost element which determines whether the requested item should be stored in the Small list or the Main list. The Ghost’s decision is generally based on whether the object was accessed within the last  $x$  misses of the cache, where  $x$  is the number of items in the Main List. Ghost lookup is similar in speed to a cache lookup, so  $E[Z_{ghost}] = 0.51\mu s$ .

As in CLOCK, every object in the cache has a special bit. When the object first enters the cache, the object is assigned a bit of 0. If the object is accessed while it is in the cache, then its bit is changed from 0 to 1.

When an object  $d$  is requested, if  $d$  is in the cache (regardless of which list), we get  $d$  and set  $d$ ’s bit to 1 (if it’s not already 1), completing the request. If, on the other hand,  $d$  is not in the cache (a “miss”), then we need to get  $d$  from disk. We now use the Ghost to figure out in which list to store  $d$ .



**Figure 13.** Queuing model of S3-FIFO cache.

If  $d$  was “missed recently” (meaning it was missed sometime within the last  $x$  misses of the cache), then we store  $d$  in the Main List,  $M$ . Otherwise, we store  $d$  in the Small List,  $S$ . Since both lists are FIFO lists, it is advantageous to  $d$  to be stored in  $M$  (which is a lot longer). Either way, we still need a Head Update (to append  $d$ ) and also a Tail Update (to remove the object at the tail of the list), thus keeping both lists at their original size.

Before the object at the tail of the  $S$  list is thrown out, it has the opportunity to move to the  $M$  list. This happens if and only if its bit is 1. Otherwise it is removed from the cache. Objects in the  $M$  list never move to the  $S$  list.

To do a queuing analysis of the network in Figure 13, we need three more things: First, we need to know the service times for all potential bottleneck operations. These are the same as the numbers in the CLOCK network. Second, and more challenging, we need to understand the fraction of requests that the Ghost sends to the  $M$  list (denoted  $p_{ghost}$ ) as opposed to the  $S$  list. Third, we need to know the fraction of items,  $p_M$ , at the tail of the  $S$  list that have a 1 bit associated with them as opposed to a 0 bit. We used our simulation to estimate both  $p_{ghost}$  and  $p_M$  as functions of  $p_{hit}$ . Our results are expressible in terms of the  $\chi^2$  function. See the supplementary document for details.

**Results of analysis and simulation.** Figure 14 shows the results of analysis (red solid lines) and simulation (blue dotted lines) for S3-FIFO. We see that increasing the hit ratio,  $p_{hit}$ , always leads to higher throughput, regardless of the mean disk latency,  $E[Z_{disk}]$ . Intuitively, this holds because the bottleneck device (no matter whether it is the headS update or tailM update) is in the miss path, not in the hit path. Therefore, increasing the hit ratio does not result in increased demand on the bottleneck device, and hence does not hurt throughput.

## 5 Discussion

### 5.1 A classification of eviction algorithms

Throughout our model analysis, simulation, and implementation study, we find that common eviction algorithms can be classified into two categories: LRU-like algorithms and FIFO-like algorithms. The throughput of LRU-like algorithms (LRU, Probabilistic LRU with lower  $q$ , and SLRU) typically

**Table 2.** Conjecture of algorithms we do not evaluate.

Classification	More Algorithms
LRU like	ARC [70], LIRS [62], TinyLFU [47], LeCaR [94], CACHEUS [79], LFU
FIFO like	CLOCK variants [32], SIEVE [110], QDLP [100], Hyperbolic [29], Random, LHD [21], LRB [90]

drops when the hit ratio becomes high. By contrast, the throughput of FIFO-like algorithms (FIFO, Probabilistic LRU with very high  $q$ , CLOCK, and S3-FIFO) always increases with the hit ratio.

While we have only studied six eviction algorithms, we can conjecture the behavior of other algorithms based on what we’ve learned. In Table 2, we have classified recent caching algorithms into LRU-like versus FIFO-like. The LRU-like algorithms all have the common feature that they perform a delink operation upon a cache hit. This delink operation becomes the bottleneck, meaning that increasing the hit ratio will increase the queuing time at the delink queue, which will lead to lower throughput.

The FIFO-like algorithms have the common feature that they do not update the global data structure upon cache hits. Consequently, increasing the hit ratio will not increase the queue length at the bottleneck device, and therefore will not lead to lower throughput. These FIFO-like algorithms can be further divided into two types. Algorithms in the first type use FIFO as the basic building block, e.g., CLOCK variants [32], S3-FIFO [106], QDLP [100], and SIEVE [110]. The second type of FIFO-like algorithms does not have a global data structure, but rather uses random sampling to choose eviction candidates. Examples include LHD [21], LRB [90], and Hyperbolic [29]. Because these algorithms do not maintain a global data structure, they are never bottlenecked by cache hits, so throughput always increases with hit ratio.

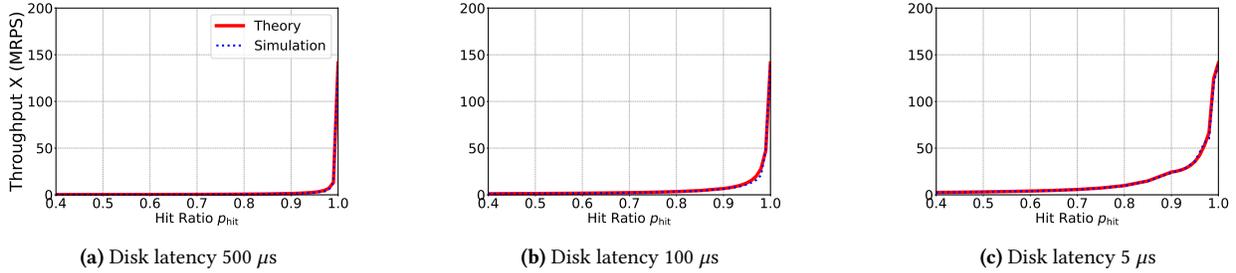
### 5.2 Trends for future caching systems

There are two trends in computing and storage hardware. First, newer CPUs have more and more cores, increasing their concurrency. At the same time, disks are getting faster, with high-end disks having single-digit microseconds of latency. Figure 2 shows the effects of these two trends. We see that the hit ratio threshold  $p_{hit}^*$  (the hit ratio point where throughput starts dropping) will shift to smaller values in the future.

### 5.3 Implication of future trends

As we move to a higher number of cores per CPU and lower disk times, it is evident the LRU-like caching algorithms will start to experience more problems. Given that LRU is the predominant caching algorithm used today, we need to figure out how we can adapt LRU-like algorithms so that their throughput doesn’t deteriorate with increasing hit ratio.

The easiest mitigation is to reduce the number of delink operations. This can be achieved by performing the delink



**Figure 14.** Results for theory, implementation and simulation under a S3-FIFO cache. For all three curves, throughput of the S3-FIFO cache always increases with a hit ratio, under different disk latencies.

operation probabilistically (probabilistic LRU). As we have shown in Section 4, when  $q$  is very high, probabilistic LRU behaves like FIFO, where increasing the hit ratio only helps.

Another solution is to simply reduce the hit ratio when it exceeds  $p_{hit}^*$  by reducing the cache size. This helps with both throughput and cost because the cache is expensive.

Another approach is to send some of the requests to the disk directly without touching the cache. This solution has been proposed recently [97]. In such a system, a request is either sent to the cache or to the disk, depending on the load of the cache.

## 6 Prior Work

DRAM-based caches are increasingly deployed in today’s system stack and have been the topic of many research papers [12, 20, 25, 28, 29, 38, 48, 54, 64, 66, 72, 74, 79, 91, 94, 98, 99, 101, 107, 108].

As mentioned in Section 1, most of the above works use the *hit ratio* as a proxy for system performance. The others that do look at throughput only consider throughput or latency under a single thread (i.e., MPL=1), e.g., DistCache [67], p-redis [73], LRB [90], GL-Cache [99]. By contrast, our paper models modern systems with MPL = 72 concurrent threads. Moreover, none of these works uses queueing-theoretic modeling to understand performance.

There are two papers that deserve a bit more attention because, like our paper, both are aimed at increasing throughput and both assume a high number of threads. In the first paper, FrozenHot [77], the authors fix the hit ratio at 99% and do not study throughput as a function of changing hit ratio. The second paper, NHC [97], is quite different from ours because it uses a single-queue implementation of a cache (rather than a 3-queue implementation). Consequently, there is no separate path for cache hits versus misses. Finally, neither paper has any analytic modeling component.

Non-DRAM-based caches are outside this paper’s scope.

## 7 Conclusion

The goal of this paper is to understand the effect of the cache hit ratio on the request throughput of DRAM-based software caching systems. Such caching systems manage

the cached items through a global linked list, where the particular eviction policy dictates the organization of this list. Our investigation uses a *three-pronged approach*. First we create a queueing model of the particular eviction policy, which we evaluate analytically to obtain an upper bound on the throughput as a function of the hit ratio. We next use simulation to exactly evaluate the queueing network. Finally we use implementation to study the eviction policy outside of a queueing network.

For all policies studied, our simulation and implementation agree within 5%, indicating that our queueing network models are quite accurate. Another takeaway is that our analysis, while only providing an upper bound, is an excellent indicator of the point at which throughput stops increasing with hit ratio. The message is that for the evaluation of future eviction policies, it suffices to use analytic upper bounds to understand whether increasing hit ratio is a good idea.

While it intuitively makes sense that higher hit ratio should lead to higher throughput, we find that this is not always true. In fact, we find that for a broad class of LRU-like algorithms (including LRU, most settings of Probabilistic LRU, and Segmented LRU), when the hit ratio gets high, the throughput actually decreases. By contrast, the throughput of FIFO-like algorithms (including FIFO, extreme settings of Probabilistic LRU, CLOCK, S3-FIFO) always increases with the hit ratio.

Our queueing analysis clearly shows why throughput drops with higher hit ratio for LRU-like algorithms. In a nutshell, LRU-like algorithms have a common feature: a cache hit requires updating the global linked list. Our analysis shows that this update operation becomes the system bottleneck, meaning that many requests queue up there. As a consequence, increasing the hit ratio increases the queueing time at this bottleneck, which reduces system throughput. Based on this intuition, we conjecture that many algorithms that we have not yet studied (ARC, LIRS, LFU, CACHEUS, LeCAR) should also exhibit this perverse behavior. By contrast, the FIFO-like algorithms all have the common feature that they do not update the global linked list upon cache hits, so increasing the hit ratio does not diminish throughput.

We also study two major trends in computing and storage: newer CPUs have an increasing number of cores, while disk latencies are steadily decreasing. We show that both these

two trends imply that LRU-like algorithms will only behave worse in the future; i.e., throughput will start to decrease at lower and lower values of the hit ratio.

**Funding:** This work was supported by NSF-IIS-2322973 and NSF-CMMI-1938909.

## References

- [1] bcache. <https://www.kernel.org/doc/Documentation/bcache.txt>. Accessed Jan 20, 2024.
- [2] Cachelib eviction policies. [https://cachelib.org/docs/Cache\\_Library\\_User\\_Guides/eviction\\_policy](https://cachelib.org/docs/Cache_Library_User_Guides/eviction_policy). Accessed: 2024-04-22.
- [3] Cloudlab. <https://www.cloudlab.us/>. Accessed Jan 20, 2024.
- [4] HHVM project. <https://github.com/facebook/hhvm.git>. Accessed Jan 20, 2024.
- [5] HHVM Scalable Concurrent Cache. <https://github.com/facebook/hhvm/blob/master/hphp/util/concurrent-scalable-cache.h>. Accessed Jan 20, 2024.
- [6] Leveldb. <https://github.com/google/leveldb>. Accessed: 2024-04-22.
- [7] Memcached. memcached - a distributed memory object caching system. <http://memcached.org/>. Accessed Jan 20, 2024.
- [8] Multi-core Processor. [https://en.wikipedia.org/wiki/Multi-core\\_processor](https://en.wikipedia.org/wiki/Multi-core_processor). Accessed Jan 30, 2024.
- [9] Open CAS. Open Cache Acceleration Software. <https://open-cas.github.io/>. Accessed Jan 20, 2024.
- [10] Redis eviction policies. <https://redis.io/blog/cache-eviction-strategies/>. Accessed: 2024-04-22.
- [11] Varnish cache. <https://varnish-cache.org/>. Accessed: 2024-01-20.
- [12] Zahaib Akhtar, Yaguang Li, Ramesh Govindan, Emir Halepovic, Shuai Hao, Yan Liu, and Subhabrata Sen. AViC: a cache for adaptive bitrate video. In *Proceedings of the 15th International Conference on Emerging Networking Experiments And Technologies*, pages 305–317, Orlando Florida, December 2019. ACM.
- [13] Alluxio. Alluxio - data orchestration for the cloud. <https://www.alluxio.io/>. Accessed: 2024-01-10.
- [14] Jussara M Almeida, Derek L Eager, and Mary K Vernon. Hybrid caching strategy for streaming media files. In *Multimedia Computing and Networking 2001*, volume 4312, pages 200–212. SPIE, 2000.
- [15] AMD. Amd epyc 4th gen cpu list. <https://www.amd.com/en/products/processors/server/epyc/4th-generation-9004-and-8004-series.html>. Accessed: 2024-01-10.
- [16] Anadtech. Intel ssd dc p3700 review: The pcie transition begins with nvme. <https://download.semiconductor.samsung.com/resources/brochure/Ultra-Low%20Latency%20with%20Samsung%20Z-NAND%20SSD.pdf>. Accessed: 2024-01-10.
- [17] Ganesh Ananthanarayanan, Ali Ghodsi, Andrew Warfield, Dhruva Borthakur, Srikanth Kandula, Scott Shenker, and Ion Stoica. {PACMan}: Coordinated memory caching for parallel jobs. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 267–280, 2012.
- [18] Ganesh Ananthanarayanan, Ali Ghodsi, Andrew Warfield, Dhruva Borthakur, Srikanth Kandula, Scott Shenker, and Ion Stoica. PAC-Man: Coordinated Memory Caching for Parallel Jobs. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 267–280, 2012.
- [19] Apache. Apache traffic server. <https://trafficserver.apache.org/>. Accessed: 2024-01-10.
- [20] Tahir Azim, Manos Karpathiotakis, and Anastasia Ailamaki. Re-Cache: reactive caching for fast analytics over heterogeneous data. *Proceedings of the VLDB Endowment*, 11(3):324–337, November 2017.
- [21] Nathan Beckmann, Haoxian Chen, and Asaf Cidon. LHD: Improving cache hit rate by maximizing hit density. In *15th USENIX symposium on networked systems design and implementation*, NSDI’18, pages 389–403, 2018.
- [22] Hamza Ben-Ammar, Yassine Hadjadj-Aoul, Gerardo Rubino, and Soraya Ait-Chellouche. On the performance analysis of distributed caching systems using a customizable markov chain model. volume 130, pages 39–51, 2019.
- [23] Benjamin Berg, Daniel S. Berger, Sara McAllister, Isaac Grosf, Sathya Gunasekar, Jimmy Lu, Michael Uhlar, Jim Carrig, Nathan Beckmann, Mor Harchol-Balter, and Gregory R. Ganger. The CacheLib caching engine: Design and experiences at scale. In *14th USENIX symposium on operating systems design and implementation*, OSDI’20, pages 753–768. USENIX Association, November 2020.
- [24] Daniel Berger, Ramesh Sitaraman, and Mor Harchol-Balter. Adaptsize: Orchestrating the hot object memory cache in a cdn. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2017)*, pages 483–498, Boston, MA, March 2017.
- [25] Daniel S. Berger. Towards Lightweight and Robust Machine Learning for CDN Caching. In *Proceedings of the 17th ACM Workshop on Hot Topics in Networks*, Hotnets’18, pages 134–140, Redmond WA USA, November 2018. ACM.
- [26] Daniel S. Berger, Benjamin Berg, Timothy Zhu, Siddhartha Sen, and Mor Harchol-Balter. RobinHood: Tail latency aware caching – dynamic reallocation from Cache-Rich to Cache-Poor. In *13th USENIX symposium on operating systems design and implementation*, OSDI’18, pages 195–212, Carlsbad, CA, October 2018. USENIX Association.
- [27] Daniel S. Berger, Sebastian Henningsen, Florin Ciucu, and Jens B. Schmitt. Maximizing cache hit ratios by variance reduction. *SIGMETRICS Perform. Eval. Rev.*, 43(2):57–59, September 2015.
- [28] Daniel S Berger, Ramesh K Sitaraman, and Mor Harchol-Balter. Adapt-Size: Orchestrating the hot object memory cache in a content delivery network. In *14th USENIX symposium on networked systems design and implementation*, NSDI’17, pages 483–498, 2017.
- [29] Aaron Blankstein, Siddhartha Sen, and Michael J. Freedman. Hyperbolic caching: Flexible caching for web applications. In *2017 USENIX annual technical conference*, ATC’17, pages 499–511, Santa Clara, CA, July 2017. USENIX Association.
- [30] Sem Borst, Varun Gupta, and Anwar Walid. Distributed caching algorithms for content distribution networks. In *2010 Proceedings IEEE INFOCOM*, pages 1–9. IEEE, 2010.
- [31] Niklas Carlsson and Derek Eager. Ephemeral content popularity at the edge and implications for on-demand caching. *IEEE Transactions on Parallel and Distributed Systems*, 28(6):1621–1634, 2017.
- [32] Richard W. Carr and John L. Hennessy. WSCLOCK: a simple and effective algorithm for virtual memory management. In *Proceedings of the eighth ACM symposium on Operating systems principles*, SOSP ’81, pages 87–95, New York, NY, USA, December 1981. Association for Computing Machinery.
- [33] Giuliano Casale and Nicolas Gast. Performance analysis methods for list-based caches with non-uniform access. *IEEE/ACM Transactions on Networking*, 29(2):651–664, 2021.
- [34] Hao Che, Ye Tung, and Zhijun Wang. Hierarchical web caching systems: modeling, design and experimental results. *IEEE Journal on Selected Areas in Communications*, 20(7):1305–1314, 2002.
- [35] Feng Chen, Binbing Hou, and Rubao Lee. Internal parallelism of flash memory-based solid-state drives. *ACM Trans. Storage*, 12(3), may 2016.
- [36] Jiqiang Chen, Liang Chen, Sheng Wang, Guoyun Zhu, Yuanyuan Sun, Huan Liu, and Feifei Li. HotRing: A Hotspot-Aware In-Memory Key-Value Store. In *18th USENIX Conference on File and Storage Technologies*, FAST’20, pages 239–252, 2020.
- [37] Kerhong Chen, Richard B Bunt, and Derek L Eager. Write caching in distributed file systems. In *Proceedings of 15th International Conference on Distributed Computing Systems*, pages 457–466. IEEE, 1995.
- [38] Audrey Cheng, David Chu, Terrance Li, Jason Chan, Natacha Crooks, Joseph M. Hellerstein, Ion Stoica, and Xiangyao Yu. Take Out the

- {TraChe}: Maximizing (Tra)nsactional Ca(che) Hit Rate. pages 419–439, 2023.
- [39] Nicaise Choungmo Fofack, Philippe Nain, Giovanni Neglia, and Don Towsley. Performance evaluation of hierarchical ttl-based cache networks. *Computer Networks*, 65:212–231, 2014.
- [40] Brian F. Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan, and Russell Sears. Benchmarking cloud serving systems with YCSB. In *Proceedings of the 1st ACM symposium on Cloud computing - SoCC '10*, page 143, Indianapolis, Indiana, USA, 2010. ACM Press.
- [41] Asit Dan and Don Towsley. An approximate analysis of the lru and fifo buffer replacement schemes. In *Proceedings of the 1990 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '90, page 143–152, New York, NY, USA, 1990. Association for Computing Machinery.
- [42] Peter J Denning. The working set model for program behavior. *Communications of the ACM*, 11(5):323–333, 1968.
- [43] P.J. Denning. Working Sets Past and Present. *IEEE Transactions on Software Engineering*, SE-6(1):64–84, January 1980.
- [44] Siying Dong, Andrew Kryczka, Yanqin Jin, and Michael Stumm. RocksDB: Evolution of Development Priorities in a Key-value Store Serving Large-scale Applications. *ACM Transactions on Storage*, 17(4):26:1–26:32, October 2021.
- [45] Donghee Lee, Jongmoo Choi, Jong-Hun Kim, S.H. Noh, Sang Lyul Min, Yookun Cho, and Chong Sang Kim. LRFU: a spectrum of policies that subsumes the least recently used and least frequently used policies. *IEEE Transactions on Computers*, 50(12):1352–1361, December 2001.
- [46] Gil Einziger, Ohad Eytan, Roy Friedman, and Benjamin Manes. Lightweight robust size aware cache management. *ACM Transactions on Storage*, 18(3), August 2022.
- [47] Gil Einziger, Roy Friedman, and Ben Manes. TinyLFU: A Highly Efficient Cache Admission Policy. *ACM Transactions on Storage*, 13(4):1–31, December 2017.
- [48] Assaf Eisenman, Asaf Cidon, Evgenya Pergament, Or Haimovich, Ryan Stutsman, Mohammad Alizadeh, and Sachin Katti. Flashield: a hybrid key-value cache that controls flash write amplification. In *16th USENIX symposium on networked systems design and implementation*, NSDI'19, pages 65–78, Boston, MA, February 2019. USENIX Association.
- [49] Ohad Eytan, Danny Harnik, Effi Ofer, Roy Friedman, and Ronen Kat. It's time to revisit LRU vs. FIFO. In *12th USENIX workshop on hot topics in storage and file systems*, hotStorage'20. USENIX Association, July 2020.
- [50] Christine Fricker, Philippe Robert, and James Roberts. A versatile and accurate approximation for lru cache performance. In *2012 24th International Teletraffic Congress (ITC 24)*, pages 1–8, 2012.
- [51] Michele Garetto, Emilio Leonardi, and Valentina Martina. A unified approach to the performance analysis of caching systems. *ACM Trans. Model. Perform. Eval. Comput. Syst.*, 1(3), may 2016.
- [52] Nicolas Gast and Benny Van Houdt. Transient and steady-state regime of a family of list-based cache replacement algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '15, page 123–136, 2015.
- [53] Nicolas Gast and Benny Van Houdt. Ttl approximations of the cache replacement algorithms lru(m) and h-lru. *Performance Evaluation*, 117:33–57, 2017.
- [54] Binny S Gill and Luis Angel D Bathen. AMP: Adaptive multi-stream prefetching in a shared cache. In *FAST*, volume 7, pages 185–198, 2007.
- [55] Hadoop. Hdfs architecture guide. [https://hadoop.apache.org/docs/r1.2.1/hdfs\\_design.html](https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html). Accessed: 2024-01-10.
- [56] Ubaid Ullah Hafeez, Muhammad Wajahat, and Anshul Gandhi. Elmem: Towards an elastic memcached system. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pages 278–289. IEEE, 2018.
- [57] Mor Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queuing Theory in Action*. Cambridge University Press, 2013.
- [58] Intel. Accelerate data center storage and memory performance. <https://www.intel.com/content/www/us/en/products/docs/memory-storage/solid-state-drives/data-center-ssds/optane-ssd-p5800x-p5801x-brief.html>. Accessed: 2024-01-10.
- [59] Intel. Intel solid-state drive dc s3700. [https://download.intel.com/newsroom/kits/ssd/pdfs/Intel\\_SSD\\_DC\\_S3700\\_Product\\_Specification.pdf](https://download.intel.com/newsroom/kits/ssd/pdfs/Intel_SSD_DC_S3700_Product_Specification.pdf). Accessed: 2024-01-10.
- [60] Aamer Jaleel, Kevin B. Theobald, Simon C. Steely, and Joel Emer. High performance cache replacement using re-reference interval prediction (RRIP). In *Proceedings of the 37th annual international symposium on Computer architecture*, ISCA '10, pages 60–71, New York, NY, USA, June 2010. Association for Computing Machinery.
- [61] Song Jiang, Feng Chen, and Xiaodong Zhang. CLOCK-Pro: an effective improvement of the CLOCK replacement. In *Proceedings of the annual conference on USENIX Annual Technical Conference*, ATC'05, page 35, USA, April 2005. USENIX Association.
- [62] Song Jiang and Xiaodong Zhang. LIRS: an efficient low inter-reference recency set replacement policy to improve buffer cache performance. In *ACM SIGMETRICS Performance Evaluation Review*, volume 30 of *SIGMETRICS'02*, pages 31–42, June 2002.
- [63] Theodore Johnson and Dennis Shasha. 2Q: A Low Overhead High Performance Buffer Management Replacement Algorithm. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB'94, pages 439–450, San Francisco, CA, USA, September 1994. Morgan Kaufmann Publishers Inc.
- [64] Vadim Kirilin, Aditya Sundarajan, Sergey Gorinsky, and Ramesh K. Sitaraman. RL-Cache: Learning-Based Cache Admission for Content Delivery. In *Proceedings of the 2019 Workshop on Network Meets AI & ML - NetAI'19*, pages 57–63, Beijing, China, 2019. ACM Press.
- [65] Jim Kurose. Information-centric networking: The evolution from circuits to packets to content. *Computer Networks*, 66:112–120, 2014. Leonard Kleinrock Tribute Issue: A Collection of Papers by his Students.
- [66] Zhenmin Li, Zhifeng Chen, Sudarshan M. Srinivasan, and Yuanyuan Zhou. C-Miner: mining block correlations in storage systems. In *Proceedings of the 3rd USENIX conference on File and storage technologies*, FAST'04, page 13, USA, March 2004. USENIX Association.
- [67] Zaoying Liu, Zhihao Bai, Zhenming Liu, Xiaozhou Li, Changhoon Kim, Vladimir Braverman, Xin Jin, and Ion Stoica. DistCache: Provable load balancing for Large-Scale storage systems with distributed caching. In *17th USENIX conference on file and storage technologies*, FAST'19, pages 143–157, Boston, MA, February 2019. USENIX Association.
- [68] Louwrentius. Understanding storage performance - iops and latency. <https://louwrentius.com/understanding-storage-performance-iops-and-latency.html>. Accessed: 2024-01-10.
- [69] Sara McAllister, Benjamin Berg, Julian Tutuncu-Macias, Juncheng Yang, Sathya Gunasekar, Jimmy Lu, Daniel S. Berger, Nathan Beckmann, and Gregory R. Ganger. Kangaroo: Caching billions of tiny objects on flash. In *Proceedings of the ACM SIGOPS 28th symposium on operating systems principles*, SOSP '21, pages 243–262, New York, NY, USA, 2021. Association for Computing Machinery.
- [70] Nimrod Megiddo and Dharmendra S Modha. ARC: A self-tuning, low overhead replacement cache. In *2nd USENIX conference on file and storage technologies*, FAST'03, 2003.
- [71] Micron. Micron 7450 ssd with nvme. [https://media-www.micron.com/-/media/client/global/documents/products/product-flyer/7450\\_nvme\\_ssd\\_product\\_brief.pdf](https://media-www.micron.com/-/media/client/global/documents/products/product-flyer/7450_nvme_ssd_product_brief.pdf). Accessed: 2024-01-10.
- [72] Elizabeth J. O'Neil, Patrick E. O'Neil, and Gerhard Weikum. The LRU-K page replacement algorithm for database disk buffering. *ACM*

- SIGMOD Record*, 22(2):297–306, June 1993.
- [73] Cheng Pan, Yingwei Luo, Xiaolin Wang, and Zhenlin Wang. pRedis: Penalty and Locality Aware Memory Allocation in Redis. In *Proceedings of the ACM Symposium on Cloud Computing, SOCC'19*, pages 193–205, Santa Cruz CA USA, November 2019. ACM.
- [74] David Plonka and Paul Barford. Context-aware clustering of DNS query traffic. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement conference - IMC '08*, page 217, Vouliagmeni, Greece, 2008. ACM Press.
- [75] Ioannis Psaras, Richard G Clegg, Raul Landa, Wei Koong Chai, and George Pavlou. Modelling and evaluation of ccn-caching trees. In *NETWORKING 2011: 10th International IFIP TC 6 Networking Conference, Valencia, Spain, May 9-13, 2011, Proceedings, Part I 10*, pages 78–91. Springer, 2011.
- [76] Pytorch. Pytorch. <https://pytorch.org/>. Accessed: 2024-01-10.
- [77] Ziyue Qiu, Juncheng Yang, Juncheng Zhang, Cheng Li, Xiaosong Ma, Qi Chen, Mao Yang, and Yinlong Xu. FrozenHot Cache: Rethinking Cache Management for Modern Hardware. In *Proceedings of the Eighteenth European Conference on Computer Systems, EuroSys '23*, pages 557–573, New York, NY, USA, May 2023. Association for Computing Machinery.
- [78] RocksDB. Block cache. <https://github.com/EighteenZi/rocksdb/wiki/blob/master/Block-Cache.md>. Accessed: 2024-01-10.
- [79] Liana V. Rodriguez, Farzana Yusuf, Steven Lyons, Eysler Paz, Raju Rangaswami, Jason Liu, Ming Zhao, and Giri Narasimhan. Learning Cache Replacement with CACHEUS. In *19th USENIX Conference on File and Storage Technologies, FAST'21*, pages 341–354. USENIX Association, February 2021.
- [80] Elisha J. Rosensweig, Jim Kurose, and Don Towsley. Approximate models for general cache networks. In *2010 Proceedings IEEE INFOCOM*, pages 1–9, 2010.
- [81] Anirudh Sabnis and Ramesh K. Sitaraman. TRAGEN: a synthetic trace generator for realistic cache simulations. In *Proceedings of the 21st ACM Internet Measurement Conference, IMC'21*, pages 366–379, Virtual Event, November 2021. ACM.
- [82] Anirudh Sabnis and Ramesh K. Sitaraman. JEDI: model-driven trace generation for cache simulations. In *Proceedings of the 22nd ACM Internet Measurement Conference*, pages 679–693, Nice France, October 2022. ACM.
- [83] Samsung. Ultra-low latency with samsung z-nand ssd. <https://download.semiconductor.samsung.com/resources/brochure/Ultra-Low%20Latency%20with%20Samsung%20Z-NAND%20SSD.pdf>. Accessed: 2024-01-10.
- [84] Bianca Schroeder, Adam Wierman, and Mor Harchol-Balter. Open versus closed: a cautionary tale. In *Proceedings of the 3rd conference on Networked Systems Design & Implementation-Volume 3*, pages 18–18, 2006.
- [85] Colin Scott. Latency numbers every programmer should know. [https://colin-scott.github.io/personal\\_website/research/interactive\\_latency.html](https://colin-scott.github.io/personal_website/research/interactive_latency.html). Accessed: 2024-01-10.
- [86] Arjun Singhvi, Aditya Akella, Maggie Anderson, Rob Cauble, Harshad Deshmukh, Dan Gibson, Milo M. K. Martin, Amanda Strominger, Thomas F. Wenisch, and Amin Vahdat. CliqueMap: productionizing an RMA-based distributed caching system. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference, SIGCOMM'21*, pages 93–105, Virtual Event USA, August 2021. ACM.
- [87] Yannis Smaragdakis, Scott Kaplan, and Paul Wilson. EELRU: simple and effective adaptive page replacement. *ACM SIGMETRICS Performance Evaluation Review*, 27(1):122–133, May 1999.
- [88] sobyete. Linux kernel page replacement algorithms. <https://www.sobyte.net/post/2022-01/linux-multi-lru/>. Accessed: 2024-01-10.
- [89] Solidigm. Solidigm p44 pro series – exceptional performanc. [https://www.bhphotovideo.com/lit\\_files/1022190.pdf](https://www.bhphotovideo.com/lit_files/1022190.pdf). Accessed: 2024-01-10.
- [90] Zhenyu Song, Daniel S Berger, Kai Li, Anees Shaikh, Wyatt Lloyd, Soudeh Ghorbani, Changhoon Kim, Aditya Akella, Arvind Krishnamurthy, Emmett Witchel, and others. Learning relaxed belady for content distribution network caching. In *17th USENIX symposium on networked systems design and implementation, NSDI'20*, pages 529–544, 2020.
- [91] Zhenyu Song, Kevin Chen, Nikhil Sarda, Deniz Altınbüken, Eugene Brevdo, Jimmy Coleman, Xiao Ju, Pawel Jurczyk, Richard Schooler, and Ramki Gummadi. HALP: Heuristic Aided Learned Preference Eviction Policy for YouTube Content Delivery Network. In *20th USENIX Symposium on Networked Systems Design and Implementation, NSDI'23*, pages 1149–1163, 2023.
- [92] Aditya Sundarrajan, Mingdong Feng, Mangesh Kasbekar, and Ramesh K. Sitaraman. Footprint Descriptors: Theory and Practice of Cache Provisioning in a Global CDN. In *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies, CoNEXT'17*, pages 55–67, Incheon Republic of Korea, November 2017. ACM.
- [93] Saran Tarnoi, Vorapong Suppakitpaisarn, Wuttipong Kumwilaisak, and Yusheng Ji. Performance analysis of probabilistic caching scheme using markov chains. In *2015 IEEE 40th Conference on Local Computer Networks (LCN)*, pages 46–54, 2015.
- [94] Giuseppe Vietri, Liana V. Rodriguez, Wendy A. Martinez, Steven Lyons, Jason Liu, Raju Rangaswami, Ming Zhao, and Giri Narasimhan. Driving cache replacement with ML-based LeCaR. In *10th USENIX workshop on hot topics in storage and file systems, hotStorage'18*, Boston, MA, July 2018. USENIX Association.
- [95] Darryl L Willick, Derek L Eager, and Richard B Bunt. Disk cache replacement policies for network filesystems. In *[1993] Proceedings. The 13th International Conference on Distributed Computing Systems*, pages 2–11. IEEE, 1993.
- [96] Carole-Jean Wu, Aamer Jaleel, Will Hasenplaugh, Margaret Martonosi, Simon C. Steely, and Joel Emer. SHiP: signature-based hit predictor for high performance caching. In *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture - MICRO-44 '11*, page 430, Porto Alegre, Brazil, 2011. ACM Press.
- [97] Kan Wu, Zhihan Guo, Guanzhou Hu, Kaiwei Tu, Ramnathan Alagappan, Rathijit Sen, Kwanghyun Park, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. The storage hierarchy is not a hierarchy: Optimizing caching on modern storage devices with orthus. In *19th USENIX conference on file and storage technologies, FAST'21*, pages 307–323. USENIX Association, February 2021.
- [98] Juncheng Yang, Reza Karimi, Trausti Sæmundsson, Avani Wildani, and Ymir Vigfusson. Mithril: mining sporadic associations for cache prefetching. In *Proceedings of the 2017 Symposium on Cloud Computing, SoCC '17*, pages 66–79, New York, NY, USA, September 2017. Association for Computing Machinery.
- [99] Juncheng Yang, Ziming Mao, Yao Yue, and K. V. Rashmi. GL-Cache: Group-level learning for efficient and high-performance caching. In *21st USENIX Conference on File and Storage Technologies (FAST 23), FAST'23*, pages 115–134, 2023.
- [100] Juncheng Yang, Ziyue Qiu, Yazhuo Zhang, Yao Yue, and K. V. Rashmi. FIFO can be Better than LRU: the Power of Lazy Promotion and Quick Demotion. In *Proceedings of the 19th Workshop on Hot Topics in Operating Systems, HOTOS '23*, pages 70–79, New York, NY, USA, June 2023. Association for Computing Machinery.
- [101] Juncheng Yang, Anirudh Sabnis, Daniel S. Berger, K. V. Rashmi, and Ramesh K. Sitaraman. C2DN: How to harness erasure codes at the edge for efficient content delivery. In *19th USENIX symposium on networked systems design and implementation, NSDI'22*, pages 1159–1177, Renton, WA, April 2022. USENIX Association.
- [102] Juncheng Yang, Yao Yue, and K. V. Rashmi. A large scale analysis of hundreds of in-memory cache clusters at Twitter. In *14th USENIX symposium on operating systems design and implementation, OSDI'20*,

- pages 191–208. USENIX Association, November 2020.
- [103] Juncheng Yang, Yao Yue, and K. V. Rashmi. A Large-scale Analysis of Hundreds of In-memory Key-value Cache Clusters at Twitter. *ACM Transactions on Storage*, 17(3):17:1–17:35, August 2021.
- [104] Juncheng Yang, Yao Yue, and Rashmi Vinayak. Segcache: a memory-efficient and scalable in-memory key-value cache for small objects. In *18th USENIX Symposium on Networked Systems Design and Implementation*, NSDI’21, pages 503–518. USENIX Association, April 2021.
- [105] Juncheng Yang, Yazhuo Zhang, Ziyue Qiu, Yao Yue, and K. V. Rashmi. S3-fifo: Simple scalable caching with three static fifo queues. <https://s3fifo.com>. Accessed: 2024-01-20.
- [106] Juncheng Yang, Yazhuo Zhang, Ziyue Qiu, Yao Yue, and Rashmi Vinayak. FIFO queues are all you need for cache eviction. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 130–149. ACM, 2023.
- [107] Tzu-Wei Yang, Seth Pollen, Mustafa Uysal, Arif Merchant, and Homer Wolfmeister. CacheSack: Admission Optimization for Google Data-center Flash Caches. In *2022 USENIX Annual Technical Conference, ATC’22*, pages 1021–1036, Carlsbad, CA, July 2022. USENIX Association.
- [108] Lei Zhang, Juncheng Yang, Anna Blasiak, Mike McCall, and Ymir Vigfusson. When is the Cache Warm? Manufacturing a Rule of Thumb. 2020.
- [109] Xiangqun Zhang, Shuyi Pei, Jongmoo Choi, and Bryan S. Kim. Excessive ssd-internal parallelism considered harmful. In *Proceedings of the 15th ACM Workshop on Hot Topics in Storage and File Systems, HotStorage ’23*, page 65–72, New York, NY, USA, 2023. Association for Computing Machinery.
- [110] Yazhuo Zhang, Juncheng Yang, Yao Yue, Ymir Vigfusson, and K. V. Rashmi. Sieve is simpler than lru: an efficient turn-key eviction algorithm for web caches. In *21th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2024)*, 2024.
- [111] Chen Zhong, Xingsheng Zhao, and Song Jiang. LIRS2: an improved LIRS replacement algorithm. In *Proceedings of the 14th ACM International Conference on Systems and Storage, SYSTOR’21*, pages 1–12, Haifa Israel, June 2021. ACM.
- [112] Timothy Zhu, Anshul Gandhi, Mor Harchol-Balter, and Michael A Kozuch. Saving cash by using less cache. In *4th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 12)*, 2012.