

# Hash & Adjust: Competitive Demand-Aware Consistent Hashing

Arash Pourdamghani  


TU Berlin

Chen Avin 

Ben-Gurion University of the Negev

Robert Sama 

University of Vienna

Maryam Shiran 

TU Berlin

Stefan Schmid 

TU Berlin & Fraunhofer SIT

---

## Abstract

---

Distributed systems often serve dynamic workloads and resource demands evolve over time. Such a temporal behavior stands in contrast to the static and demand-oblivious nature of most data structures used by these systems. In this paper, we are particularly interested in consistent hashing, a fundamental building block in many large distributed systems. Our work is motivated by the hypothesis that a more adaptive approach to consistent hashing can leverage structure in the demand, and hence improve storage utilization and reduce access time.

We initiate the study of demand-aware consistent hashing. Our main contribution is H&A, a constant-competitive online algorithm (i.e., it comes with provable performance guarantees over time). H&A is demand-aware and optimizes its internal structure to enable faster access times, while offering a high utilization of storage. We further evaluate H&A empirically.

**2012 ACM Subject Classification** Theory of computation → Online algorithms; Theory of computation → Data structures design and analysis

**Keywords and phrases** Consistent hashing, demand-awareness, online algorithms

**Digital Object Identifier** 10.4230/LIPIcs...

**Supplementary Material** (*Source Code*): <https://github.com/inet-tub/Hash-And-Adjust>

**Funding** This project has received funding from the European Research Council (ERC) under grant agreement No. 864228 (AdjustNet), 2020-2025.

## 1 Introduction

Access patterns and demand in distributed systems are often dynamic and feature structure over time [5, 7, 12, 18, 21–23, 41, 44, 47]. Such dynamicity and temporal locality of demand have been observed in many application domains and encouraged the design of optimized computer architectures that support such structure, e.g., through a cache hierarchy. Our paper is motivated by the hypothesis that there is still much-untapped potential for adaptive and demand-aware approaches for optimizing the performance of data-centric applications [8].

We are particularly interested in consistent hashing [31], a key component of many distributed systems, which has applications in systems such as Amazon DynamoDB [17, 48] and Apache Cassandra [34]. Consistent hashing aims to provide fast lookups on top of flexible insertion and deletions, making it especially attractive for web applications and distributed databases [33]. However, consistent hashing today is still mostly demand-oblivious and has



shortcomings, e.g., related to the high variance in the load of different servers, which can result in sub-optimal utilization.

This paper envisions a more demand-aware approach to consistent hashing, facilitating the adaptation to—and exploitation of—structure in the demand, with the goal of reducing access costs and improving load balancing (i.e., maximizing storage utilization). Our work is further motivated by a recent work by Mirrokni, Thorup and Zadimoghaddam [39] that showed that bounding server capacities within a factor of the minimum needed capacity (i.e., the number of items divided by the number of servers) is possible given a tolerable access cost increase. Considering access costs is a new dimension in the design of consistent hashing methods that we further develop in this work. We show that our demand-aware approach can improve storage utilization while keeping the access cost competitive. Designing a demand-aware variant of consistent hashing is, however, challenging. This is because the demand is not perfectly known ahead of time. Therefore such a design requires a careful balancing of the benefits of adjustment and the cost of adjustments [46].

Our main contribution is *Hash & Adjust* ( $H\&A$  for short), an online and demand-aware algorithm for consistent hashing. The performance of  $H\&A$  differs only by a constant factor from the optimal offline algorithm (that knows the whole demand in advance), i.e.,  $H\&A$  is constant competitive. Constant competitiveness becomes possible in our system by keeping the recently accessed items close to their original server via self adjustments, in the spirit of strategies known from list update problems [4, 49]. Furthermore, adjustments of  $H\&A$  are local. Local adjustments are crucial for its practical use cases in distributed systems, due to their storage and computation efficiency. To achieve this, we extend list access in a novel direction: *list access with multiple heads and with capacity*. This extension can be of independent interest in other application scenarios.

We complement our theoretical contribution with an empirical comparison between the performance of  $H\&A$  with state-of-the-art algorithms. Our empirical evaluations indicate an average 54% improvement in the access cost (i.e., item access time) compared to the recent consistent hashing method [39] (see details in the Section 6).

## 2 Motivation and Overview of $H\&A$

This section motivates the need to critically reconsider today’s consistent hashing methods. We begin with a short description of the current state of consistent hashing methods, detailed in more depth in Section 7.

**Basic consistent hashing.** In the basic version of consistent hashing [31], servers (e.g., nodes of a system) are arranged over a ring, hosting a set of items (e.g., DNS records). Operations on an item begin by computing the hash value for the item. Then, based on the hash value, the item is assigned to the closest server (in a clockwise manner, details follow in the next section). In practice, the requests are generated in a distributed manner: any server can initiate the request, run the hash function, and find the target server using the network that is built on top of the ring. In this work, following the lead of recent advancements [1, 39], we focus on ring abstraction. With this abstraction, we provide flexibility of choice: we can connect or even better: nodes on the ring can be connected to form a desired network of choice. Hence, we remove the dependence on a particular network and analyze a consistent hashing instance in its most general form.

Consistent hashing operations are *history independent* and *local*. This is because a single shared hash function suffices to determine the server that hosts an item. Furthermore, consistent hashing is preferable to other distributed hashing methods as it allows for online

Algorithm	Access Cost	Memory Utilization
Traditional [31]	Low	Low
WBL [39]	High	Medium
$H\mathcal{E}A$ [this work]	Low	High

■ **Table 1** A comparison between  $H\mathcal{E}A$  and other variants of consistent hashing, in terms of access cost and memory utilization.  $H\mathcal{E}A$  ensures both high memory utilization and low access cost by supporting bounded server loads and on-the-fly self-adjustments, while other algorithms can not support both at the same time. The empirical counterpart of this table can be observed in Figure 3a in Section 6.

insertion and deletions. However, traditional consistent hashing methods suffer from low storage utilization. Therefore, we also evaluate server loads (i.e., the number of items in a server).

**Suboptimal storage utilization.** A high variance in server loads implies a suboptimal storage utilization and is generally undesirable in practice. To this end, we define server capacity as the maximum number of items a server can hold. Since the server with the maximum load can not be predetermined, we need to reserve the capacity equal to the maximum load for *all* servers. As the sum of the load of all servers is equal to the number of items, we define the storage utilization as:

$$\frac{\text{Total Load}}{\text{Total Capacity}} = \frac{\text{Number of Items}}{\text{Maximum Load} * \text{Number of Servers}} = \frac{\text{Average Load}}{\text{Maximum Load}} \quad (1)$$

For a fair comparison, we consider an equal number of items and servers for all algorithms. Hence, given Equation 1, we focus on the maximum load of algorithms.

It is well known that traditional consistent hashing has a non-constant difference between the average and maximum load [25]. This is why a recent work by Mirrokni et al. [39] suggested bounding the load of each server by a constant factor times the minimum load needed (we call their work With Bounded Loads, or “WBL” for short). However, the difference remains limited between *WBL*’s proposal and the case with unbounded capacity. We emphasize that bounding loads aim to adhere to stringent storage constraints of reserved storage resources, that can not be achieved through flexible storage allocation methods [38].

**Increased access time.** The idea of bounding server capacity might seem obvious at first sight. However, a challenge arises: as servers become full, we need to put the items of the oversubscribed server elsewhere, otherwise data may be lost. To model this issue more formally, we define access time as the number of servers that one needs to traverse to find an item. This is a fair consideration, as moving items between servers requires noticeably more time than searching within a server for an item. Also, moving items between servers should be local: without a centralized map of items or separate server assignment functions.

Fortunately, however, real-world request sequences have an inherent *temporal locality*. This means that among items that are assigned to a server, we have intervals of consecutive requests for the same item – see [7] for a study of the temporal locality of real-world instances.

**Overview of  $H\mathcal{E}A$ .** The consistent hashing method proposed in this work relies on an online algorithm,  $H\mathcal{E}A$ .  $H\mathcal{E}A$  allows us to solve the two issues identified above (and summarized in Table 1).

- Our method improves *storage utilization* by considering only an *additive* additional capacity (e.g., considering only 2 more storage slots per server) rather than the multiplicative additional capacity proposed previously. Thus, we can push the storage utilization to its limit.

- To reduce *access time*,  $H\mathcal{E}A$  self-adjusts items between servers after every request – not only upon insertion or deletion. This way, we ensure a decreased overall average cost, especially in the case of a high temporal locality.

### 3 Model and Preliminaries

In this section, we present our theoretical model and introduce the required terminologies and preliminaries.

**Consistent hashing.** We consider a set of  $m$  items  $V$ , and a set  $S$  of  $n$  servers. As in practice, the number of items is much larger than the number of servers; we consider  $m$  to be significantly greater than  $n$ , i.e.,  $m \gg n$ .

We assign a *head* to each item. The head of an item is a server that has the closest hash value to the hash value of the item. In our analysis, we consider a hash function  $hash()$ , that hashes the ID of a server or an item to a value in the range  $[0, 1]$  uniformly at random<sup>1</sup>. Formally, the head  $s$  for an item  $v$  is selected as follows:

$$head(v) = \arg \min_{s \in S} hash(s) - hash(v) \pmod 1$$

In other words, the function  $hash$  maps items and servers on a ring, and the head of an item is a server on the ring closest to the item in a clockwise manner. We emphasize that this process is carried out in a distributed fashion, i.e., the hash function is known globally, not solely to a central entity. Furthermore, we define the server that hosts item  $v$  at time  $t$  as  $host^t(v)$ .

Considering a fixed  $hash$  function, servers maintain a certain order between them. Given a server  $s$ , we denote the server after server  $s$  by  $s^+$  and the server before  $s$  by  $s^-$ .

**Bounding servers' capacity.** All servers can contain up to  $c = \lceil \frac{m}{n} \rceil + \alpha$  items, i.e., they have *capacity* equal to  $c$ . Variable  $\alpha \geq 1$  is the *additive extra capacity* which is fixed throughout the run time of  $H\mathcal{E}A$ . See Figure 1a for an example.

Considering additive extra capacity results in a tradeoff between storage utilization and access cost. Increasing the value of  $\alpha$  would lower the access cost of the system (as items will be stored closer to their head), but requires more storage per server.

**Online requests sequence.** We consider a significantly large request sequence  $\sigma = (\sigma_1, \dots)$ , in which  $\sigma_t$  is the request at time  $t$ . Requests can either refer to an item or a server, and they might have one of three types:

- *Access* to an item: searching for an item inside the system,
- *Inserting* a server or an item: adding a new server or an item that was not in the system before,
- *Deleting* a server or an item: removing one of the existing items or servers from the system.

It is important to respond to requests locally and immediately, i.e., not requiring a global view or a (estimation of) future knowledge when updating the data structure, as they are critical for distributed settings. Furthermore, we assume that insertion and deletion operations are happening rarely [14], in particular after every  $\sum_{i=1}^{n-1} e^{-\frac{\alpha^2}{2m}(i+1)}$  access operation. While responding to requests, a system might *reconfigure* some items, i.e., moving them from one server to another.

---

<sup>1</sup> This is a common [39] assumption, given that there are hash functions that can provide almost uniformly at random results [52].

**Cost model.** Responding to a request comes at a delay: we might need to search a few servers to find the item or move items around after insertion and deletion requests. Therefore, we consider two types of delay:

- *Search delay:* the delay that occurs while searching and retrieving data from a server. In our abstraction, we consider the access delay as the number of servers traversed to find an item, including the head of an item.
- *Reconfiguration delay:* the delay of moving items between servers. Similar to the access delay, we only consider the number of servers that we move an item between.

We emphasize that all the delays in any operation (access, insertion, and deletion) are essentially either a search delay or a reconfiguration delay. We abstract the delay we observe for a request as the *cost* of that request. The cost of searching or moving items inside a server is negligible compared to the inter-server costs. We point out, that in our model, accessing the head of an item has a cost of 1, for any algorithm.

Given that transferring data costs more than accessing it, we consider a constant factor  $\omega > 1$  for the cost of relocating a single item from a server to a neighboring server compared to the cost of going from one server to another searching for an item (this is another extension of the previous consistent hashing models [1, 39], which only considered  $\omega = 1$ ). Hence, the *total* cost of an algorithm  $ALG$  for all its operations over a request sequence is:

$$C_{ALG}(\sigma) = C_{ALG}^{\text{Search}}(\sigma) + \omega \cdot C_{ALG}^{\text{Reconfiguration}}(\sigma) \quad (2)$$

Our aim is to minimize the total cost of our online algorithm compared to the optimal offline algorithm. We consider that the inputs are generated by an oblivious adversary. The oblivious adversary is not aware of the random bits used by an algorithm. This is a common assumption, as there are opportunities for pseudorandom number generation (for details on different types of adversaries, check [11]). Formally, our objective is to develop an algorithm with a constant competitive ratio, i.e., we want the cost of our algorithm to match the optimal offline algorithm asymptotically.

► **Definition 1 (Competitive ratio).** *Consider an online algorithm  $ALG$ , and an optimal offline algorithm  $OPT$ . Denote the total cost of  $ALG$  over an input sequence  $\sigma$  as  $C_{ALG}(\sigma)$ , and similarly the cost of  $OPT$  as  $C_{OPT}(\sigma)$ . Then the (strict) competitive ratio of  $ALG$  is defined as:  $\max_{\sigma} \frac{C_{ALG}(\sigma)}{C_{OPT}(\sigma)}$ .*

## 4 H&A Algorithms

In this section, we first discuss how access operations are done in  $H\mathcal{E}A$ , and then discuss insertion and deletion operations.

### 4.1 Access in $H\mathcal{E}A$

We start by describing how we handle an access request in  $H\mathcal{E}A$ . The main innovation of our approach is to use self-adjustments after each access request.

**Self-adjustments on a ring.** With self-adjustments, we aim to reduce the access cost by adjusting the position of items inside the data structure. In particular, we look at a self-adjustment procedure that brings back items to their heads step by step, from a server to its neighbor on the ring. We consider self-adjustments on a ring as an abstraction (rather than considering a fixed network used in distributed hash tables). This is a plus: it provides

■ **Algorithm 1** *H&A* access to an item  $v$

---

```
1 set  $s = head(v)$ .
2 while  $v$  is not in  $s$  and  $s$  is full:
3   set  $s$  as  $s^+$ .
4 if  $v$  is not in  $s$  and  $s$  is not full:
5   return “ $v$  is not in the system”.
6 while  $s$  is not  $head(v)$ :
7   name  $u$  as the least recently accessed item in  $s^-$ .
8   swap( $v, u$ ).
9   set  $s$  as  $s^-$ .
```

---

a flexible choice for the network that can be built on top, as each application in peer-to-peer usecases uses a different overlay network (Chord, Kademia, etc.).

**Accessing an item.** After our algorithm (Algorithm 1) receives access to an item  $v$ , it starts looking for  $v$  inside the head assigned to it,  $s = head(v)$ . From then on, for any given server  $s$ , *H&A* checks for the following three possibilities in order:

1. Item  $v$  is not in server  $s$  and  $s$  is full. Then, item  $v$  might have been moved to server  $s^+$ , so we set  $s = s^+$  and start checking again.
2. Item  $v$  is not in server  $s$  and  $s$  is not full. This means that item  $v$  is not in our system.
3. Item  $v$  is in server  $s$ . Then we successfully found the item. If  $s \neq head(v)$ , then we swap  $v$  with the least recently accessed item in server  $s^-$ ,  $u$ , through a  $swap(u, v)$  operation. The swap operation for items  $u$  and  $v$  works for two items with adjacent servers. We repeat swapping until  $v$  reaches  $head(v)$ .

Accessing only involves items that are already in the system. If an item is not already in the system, it is handled by insertion procedure, discussed next.

## 4.2 Insertion and Deletion in *H&A*

We now focus on how the insertion and deletion of items (or servers) are performed. Now that both numbers of items and servers could change, we define  $m^t$  to be the number of items and  $n^t$  to be the number of servers at time  $t$ . Given the dynamicity of  $n^t$  and  $m^t$ , changing the servers' capacity is essential to ensure our capacity constraint on each server.

**Capacity Change.** We perform capacity changes in phases. A phase ends in one of two scenarios:

- if we have a server insertion/deletion,
- if the difference between the number of item insertions and deletions overshoots  $n$  since the start of the phase. We keep this difference in a parameter called  $\delta$ .

At the end of a phase, we need to decrease or increase the capacity of all servers to adhere to the capacity constraint based on the number of items and servers. In case of a capacity decrease, we might have some servers that have more items than their capacity. With the capacity increase, servers that were previously full are not full anymore, affecting access operations. To remedy this issue, we first define *valid* items for a server  $s$ .

► **Definition 2.** *An item is valid for server  $s$  if its head is  $s$  or any of the servers located in front of  $s$  but before the next non-full server.*

Given the definition of valid items, we now detail how *H&A* solves the aforementioned issues:

---

**Algorithm 2** *H&A* Item Insertion/Deletion

---

- 1 use Algorithm 1 to search for the item  $v$ .
  - 2 **if** *the request is item insertion and  $v$  was not found*:
  - 3     Put  $v$  in the first non-full server after  $head(v)$ .
  - 4     increase  $\delta$  and run change capacities if  $\delta \geq n$ .
  - 5 **if** *the request is item deletion and  $v$  was found*:
  - 6     remove  $v$  from its current server.
  - 7     fill the extra capacity.
  - 8     decrease  $\delta$  and change capacities if  $\delta \leq -n$ .
- 

---

**Algorithm 3** *H&A* Capacity Change

---

- 1 Set the capacity of all servers equal to  $\lceil \frac{m^t}{n^t} \rceil + \alpha$ .
  - 2 Set  $\delta$  equal to 0.
  - 3 **for** *all servers  $s$  starting from server 0*:
  - 4     **if**  *$s$  was full in the previous phase and had extra capacity*:
  - 5         **while** *Server  $s$  has extra capacity*:
  - 6             set  $s' = s^+$ .
  - 7             **while**  *$s'$  does not have a valid item*:
  - 8                  $s' = s' + 1$
  - 9             bring the newest valid item from  $s'$  to  $s$ ,
  - 10     **if**  *$s$  has more items than its capacity*:
  - 11         move the least recently accessed item of  $s$  to  $s^+$ .
- 

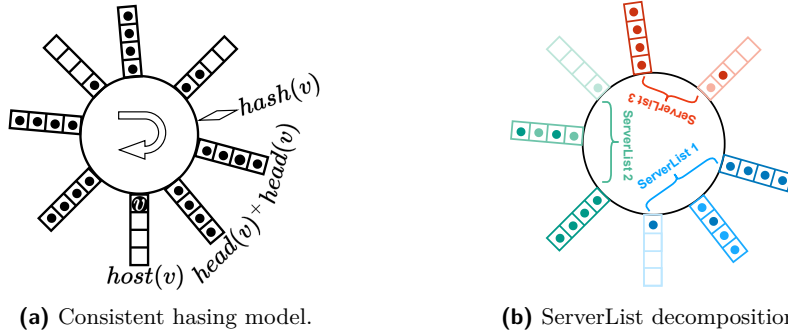
- (i) *Filling unused capacity*: In this case, we fill the additional slot that has been created in  $s$  by moving the newest *valid* item from the back of  $s$  (server  $s^+$  and beyond) to  $s$ . In particular, we first search the server  $s^+$  for a valid item and move it to  $s$ . If there is none, we look into the next servers until reaching a non-full server (given that we have extra capacity for all servers, as we have extra capacity per server, i.e.  $\alpha \geq 1$ , such a non-full server could always be found). In case no such valid item exists until the next non-full server, it means no items from the server in front of  $s$  have been moved over  $s$ , so we keep  $s$  as non-full and stop, while being sure that Invariant 2 is intact.
- (ii) *Moving out extra load*: In this case, we need to move out the extra item  $v$ . For that, we move the least recently accessed item of server  $s$  to server  $s^+$ .

Combining the above procedures, we end up with the algorithm to keep the capacity of servers consistent, Algorithm 3. Now, we discuss how insertion and deletions are designed on top of capacity change operations.

**Inserting an item.** To insert  $v$ , after searching for it, we assume that we have not found it, otherwise, we do not need to do anything, to avoid duplicates of the same item. Then, we place  $v$  in  $s^{nf}$ , the first non-full server after  $head(v)$ . As  $s^{nf}$  is non-full, we can place  $v$  in it, without moving other items.

**Deleting an item.** To delete  $v$ , if we could not find it between  $head(v)$  and the next non-full server, it is nowhere else to be found (based on Invariant 2), and nothing else is needed to be done. If the item was found, we would remove it from the server that currently contains  $v$ ,  $host(v)$ . After that, our algorithm needs to fill the extra capacity created. If we can not find a valid item to fill the extra capacity, we do not take any further action.

**Inserting a server.** We first determine the position of the server using our hash function. We then tag the new server as a server that was previously full and run Algorithm 3 to bring valid items back to this server. Keep in mind that there might not be enough valid items,



■ **Figure 1** Figure 1a shows an example of our model with 8 servers, each with capacity 4, and 24 items. In this example, item  $v$  was initially inserted into server  $head(v)$  (it had the closest hash value); however, because  $head(v)$  and  $head(v)^+$  were full, it is moved to  $host(v)$ . Figure 1b depicts decomposition of the previous example into ServerLists. Each ServerList is shown by a different color, and servers have different gradients of the color of their ServerList as they are different heads on their own. Items are colored by the color of their host.

and this server can remain non-full.

**Deleting a server.** To delete server  $s$ , we move all of its items to the server  $s^+$  temporarily, and then remove it from the system. After running Algorithm 3 to adjust the capacity of all servers, we move the extra items that might be accumulated on top of server  $s^+$  to the next servers, giving priority to the oldest items in each server to stay in that server.

An important invariant of discussed insertion and deletion operations is that they keep the relative order of items.

▶ **Invariant 1.** *After an item (or a server) insertion (or deletion), our algorithm keeps the order of items that were previously in the system, the same.*

## 5 Competitiveness

In this section, we provide the competitiveness analysis of our algorithms. We first start by stating a general method, and then discuss the competitiveness of access and insertion/deletion operations separately.

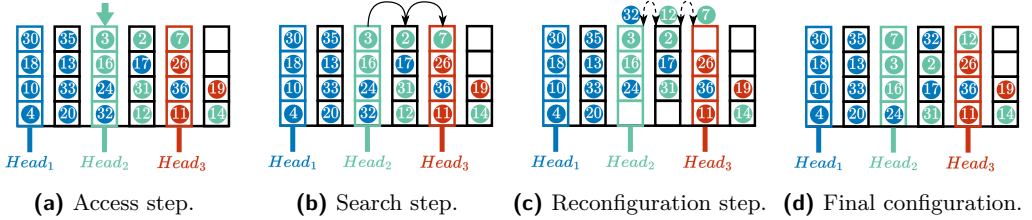
**Decomposition to ServerLists.** For analytical purposes, we decompose the ring used by  $H\&A$  into a set of full servers, except their last server: a set of ServerLists. Informally, a ServerList is a consecutive subsequence of servers on the ring (a formal and more general definition follows, see Figure 1b for an example). Such a decomposition is possible because of additional extra capacity, which results in non-full servers. Those non-full servers are at the end of a ServerList, and the server after them is the start of another ServerList. Our algorithm is designed such that no item moves over the non-full servers; hence the interactions inside a ServerList do not affect other ServerLists.

▶ **Invariant 2.** *For any given item  $v$  at any given time  $t$  in a ServerList of  $H\&A$ , there is no non-full server between  $head(v)$  and  $host^t(v)$ .*

### 5.1 Access Operations

The traditional list access problem [49] considers a list of servers (with capacity one) on a *line*, and all items share the same head. A ServerList is an extension of a list in two directions:





■ **Figure 2** An example of the ServerList access problem with three heads and a capacity of four. In this figure, the relation between items and their original host is shown by using the same color. Assume access to the item 7 with  $Head_2$ . As it is a green item, that access starts from the  $Head_2$ , the head for green items (Figure 2a). Then we search server  $Head_2$  and the servers that come after it for item 7 (Figure 2b). After accessing item 7 we swap it with the oldest items in servers between its current server and its host server (Figure 2c), to reach the final configuration (Figure 2d).

- We allow the list to have multiple heads, i.e., items can be assigned to one of the servers from a subset of  $0 < b \leq n$  servers, which we call heads of the ServerList. We denote these heads by  $\{head_1, \dots, head_b\}$ <sup>2</sup>.
- We allow servers to have capacities larger than one and equal to  $c$ , and we consider all of the servers to be full except the last one.

In the following, we will show that a competitive result can be achieved for ServerLists. The algorithm that we use for access in a ServerList is the same as Algorithm 1. We can see the steps of the access algorithm on a ServerList in Figure 2.

In the ServerList maintained by Algorithm 1, between two items with the same head, the item that has been accessed more recently is closer to the head. It is because after an access request, the least recently accessed item is moved to its head, and we have chosen the oldest items from servers in between to move back.

► **Invariant 3.** *For any given pair of items  $u$  and  $v$  with the same head  $H$ , if the last access to  $u$  was sooner than  $v$ , item  $u$  is closer to head  $H$ .*

Our algorithm achieves a low access cost, and as we show in the following, a low competitive ratio. To show this, we compare the amortized total cost of our algorithm with the optimal offline algorithm  $OPT$  over the whole request sequence. We consider the optimal algorithm that is offline, i.e., it is aware of the input sequence in advance, hence this proof shows that Algorithm 1 performs almost optimal, even without any additional information about possible future requests or patterns in the input. Furthermore, among such optimal offline algorithms, we select an algorithm that only reconfigures items between servers, i.e., keeps items in each server in the order that they have been added. Such an optimal offline algorithm exists, as the order of items inside a server does not affect the cost of an algorithm, since only movements between servers are costly. Proof of the following theorem uses an argument based on counting the number of inversions between the optimal offline algorithm and our algorithm, to show competitiveness of our algorithm.<sup>3</sup>

► **Theorem 1.**  $(\star)$  *Algorithm 1 is  $2 \cdot (1 + \omega)$  competitive, where  $\omega$  is the cost of moving items between two adjacent servers.*

<sup>2</sup> This definition generalizes definition needed to prove the competitiveness of  $H\mathcal{E}A$ , in which every server would be a head.

<sup>3</sup> The proofs of statements marked by  $\star$  are deferred to Appendix A.

## 5.2 Insertion and Deletions

Now we prove that  $H\mathcal{E}A$  is constant competitive even under insertions and deletions.

While our algorithm works in general, as usual in the literature [4, 11, 14], for analytical purposes, we consider the following assumptions for any algorithm. This assumption is practically motivated.

► **Assumption 1.** *Accesses are more frequent than insertions and deletions: in particular, we assume that each phase has at least length  $\sum_{i=1}^{n-1} e^{-\frac{\alpha^2}{2m}(i+1)}$  (in which  $\alpha$  is the additive capacity of each server). We call such a request sequence well-behaved.*

In order to justify constant competitiveness of  $H\mathcal{E}A$ , we first analyze the maximum expected length among all *maximal sequence of consecutive full servers* (by maximal we mean a sequence of servers that has non-full servers before and after), or a MSCFS for short. In doing so, we first analyze the probability of having a MSCFS of length at least  $\ell$ .

Let random variable  $L$  be the length of a MSCFS. Now we compute the probability of  $L \geq \ell$ , i.e., having a MSCFS with length at least  $\ell$ .

► **Lemma 1.** *The probability of having a maximal sequence of consecutive full servers with length at least  $\ell$  is less than or equal to  $e^{-\frac{\alpha^2}{2m} \cdot (\ell+1)^2}$ .*

**Proof.** Consider  $s$  to be the immediate full server after a non-full server  $s^-$  (only such servers can be a start of a MSCFS). Now, let us consider  $a_k$  as the number of items such that their head (the server with the closest hash in a clockwise manner) is the  $k - 1$  server after from  $s$ .

Having a MSCFS with length  $\ell$  from server  $s$  requires all servers starting from  $s$  should have been head of "enough" items. Formally, it is equivalent to having at least  $k \cdot c$  items with their head in any of  $1 \leq k \leq \ell$  consecutive servers starting from  $s$  (remember that  $c$  is the capacity of each server). In other words,

$$P(L \geq \ell) = P(\forall 1 \leq k \leq \ell, \sum_{j=1}^k a_j \geq k \cdot c) \quad (3)$$

However, we only need a weaker assumption in this proof, which is  $\sum_{k=1}^{\ell} \sum_{j=1}^k a_j \geq \sum_{k=1}^{\ell} k \cdot c$ , therefore we can rewrite Inequality 3 as:

$$P(L \geq \ell) \leq P\left(\sum_{k=1}^{\ell} \sum_{j=1}^k a_j \geq \frac{\ell \cdot (\ell + 1)}{2} \cdot c\right)$$

Now, to simplify the right-hand side, let us assign weight  $\ell - k$  to all items that their head have distance  $k$  from  $s$ . To formalize this, let us consider  $X_1, X_2, \dots, X_m$  as random variables for all items. If the head of an item  $v$  is not in the range of  $\ell$  servers including and after  $s$ , we set  $X_v$  equal to 0. Otherwise,  $X_v$  is  $\ell$  minus distance between the head of  $v$  and  $s$ . Hence,  $0 \leq X_v \leq \ell$ .

Let us then consider  $X = \sum_{v \in V} X_v$ . By definition of our random variables, we have the following which essentially comes from a double counting method:

$$P\left(\sum_{k=1}^{\ell} \sum_{j=1}^k a_j \geq \frac{\ell \cdot (\ell + 1) \cdot c}{2}\right) = P\left(\sum_{v \in V} X_v \geq \frac{\ell \cdot (\ell + 1) \cdot c}{2}\right) = P\left(X \geq \frac{\ell \cdot (\ell + 1) \cdot c}{2}\right)$$

To go one step further, we now use Hoeffding's inequality [26]. As the heads of items are selected uniformly at random,  $X_v$  is independent from other items' random variables.

To compute  $\mu = E[X]$ , using linearity of expectations, we have  $E[X] = \sum_{v \in V} E[X_v]$ . As in expectation  $\frac{1}{n}$  of items are assigned to a server (due to the uniformity of the hash function), we have:

$$\sum_{v \in V} E[X_v] = \sum_{v \in V} \frac{\ell \cdot (\ell + 1)}{2 \cdot n} = \frac{m \cdot \ell \cdot (\ell + 1)}{2 \cdot n}$$

By setting  $\tau = (c - \frac{m}{n}) \cdot \frac{\ell \cdot (\ell + 1)}{2}$ , and using Hoeffding's inequality, and as  $\alpha \leq (c - \frac{m}{n})$ , we have:

$$P(X \geq \frac{\ell \cdot (\ell + 1) \cdot c}{2}) = P(X \geq \mu + \tau) \leq e^{-\frac{2 \cdot ((c - \frac{m}{n}) \cdot \frac{\ell \cdot (\ell + 1)}{2})^2}{m \cdot \ell^2}} \leq e^{-\frac{\alpha^2}{2 \cdot m} (\ell + 1)^2}$$

Putting everything together, we have:

$$P(L \geq \ell) \leq P(X \geq \frac{\ell \cdot (\ell + 1) \cdot c}{2}) \leq e^{-\frac{\alpha^2}{2 \cdot m} (\ell + 1)^2}$$

◀

Given Lemma 1, we now compute the maximum length among all MSCFSs.

► **Lemma 2.** *The expected maximum length among all of the maximal sequence of consecutive full servers is less than or equal to  $\sum_{i=1}^{n-1} e^{-\frac{\alpha^2}{2m}(i+1)^2}$ .*

**Proof.** Consider  $L_{max}$  is a random variable indicating the length of the maximum MSCFS, and  $L$  is a random variable that determines the length of any arbitrary MSCFS. Hence:

$$E[L_{max}] = \sum_{i=1}^n i \cdot P(L_{max} = i) \leq \sum_{i=1}^n i \cdot P(L = i)$$

The above inequality is true because event  $L_{max} = i$  means that we have an MSCFS with length  $i$  and the length of all other MSCFS is less than equal to  $i$ . We then replace  $P(L = i)$  with  $P(L \geq i) - P(L \geq i + 1)$ :

$$E[L_{max}] \leq \sum_{i=1}^n i \cdot (P(L \geq i) - P(L \geq i + 1)) = \sum_{i=1}^n i \cdot P(L \geq i) - \sum_{i=1}^n (i - 1) \cdot P(L \geq i) = \sum_{i=1}^{n-1} P(L \geq i)$$

The last step is true given the extra capacity per server, we can not have all servers full, hence  $P(L \geq n)$  and therefore  $n \cdot P(L \geq n)$  is 0. Now, based on Lemma 1 we get:

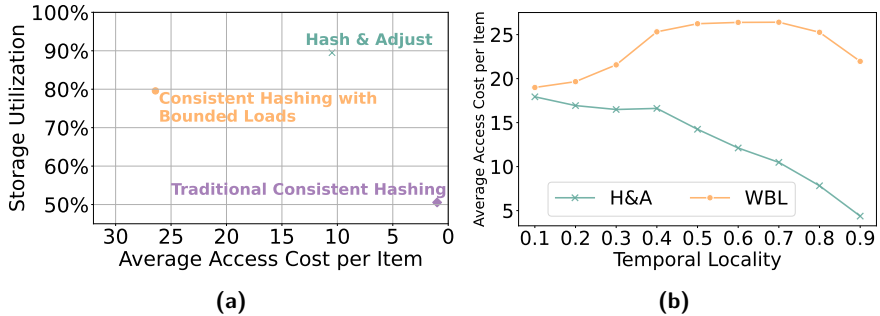
$$E[L_{max}] \leq \sum_{i=1}^{n-1} P(L \geq i) \leq \sum_{i=1}^{n-1} e^{-\frac{\alpha^2}{2 \cdot m} (i+1)^2}$$

Hence, the expected maximum length of a MSCFS is less than or equal to  $\sum_{i=1}^{n-1} e^{-\frac{\alpha^2}{2m}(i+1)^2}$ .

◀

Recall that insertions and deletions require capacity changes at the end of the phase, and in the following, we first discuss the effects of such changes on the competitive ratio using the result of Lemma 2.

► **Lemma 3.** *(\*) A capacity change operation at the end of each phase increases the competitive ratio by an additive constant, in expectation.*



**Figure 3** Figure 3a compares the average cost and memory utilization of the  $H\&A$  and the  $WBL$  algorithm [39] and Traditional method [31]. We normalized the access cost by the number of items. This figure considers 100,000 requests, 10,000 items, and 20 servers, and uses an instance generated by temporal locality 0.75. Figure 3b compares the access cost of  $WBL$  with  $H\&A$ , by varying temporal locality of the input. For this figure, we consider same setup.

Theorem 2 builds upon Lemma 3 and Theorem 1, given realistic assumptions on input, mentioned at the beginning of this section. This theorem shows, that in expectation, how we can conclude that the amortized cost of each operation is constant, given that access operations have a constant cost, and insertion and deletions are rare enough to match our assumption.

► **Theorem 2.** ( $\star$ ) *Considering a well-behaved request sequence,  $H\&A$  is constant competitive, in expectation.*

## 6 Experimental Evaluation

In this section, we complement our theoretical analysis of  $H\&A$  (the main contribution of this paper) by providing first insights into the empirical performance of  $H\&A$ .

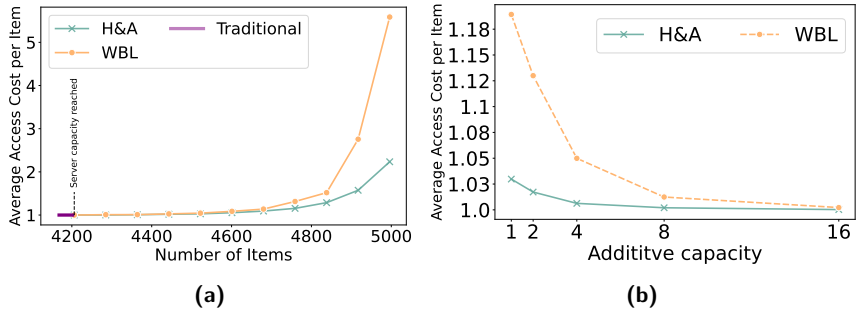
We use our own benchmarking tool to compare consistent hashing algorithms in terms of cost and memory utilization and then show the effect of our parameters on them. Our tool can create a variety of input sequences, both based on real-world clicks datasets [6] and also synthetic inputs with various *temporal localities*. Temporal locality measures the probability of an item being repeated consecutively.

### 6.1 Evaluation Setup

Our experimental results are based Python 3.6 implementation. For visualizations, we used seaborn 0.11 [53] and Matplotlib 3.5 [27] libraries. The code was executed on a machine with 2x Intel Xeons E5-2697V3 SR1XF with 2.6 GHz, 14 cores each, and a total of 128 GB DDR4 RAM. A summary of parameters used in our benchmarking is in Table 2 in the appendix.

**Input generation.** We create a range of input sequences, both based on real-world and also synthetic inputs. For real-world inputs, we considered a click dataset [6] and the CAIDA Anonymized Internet Traces Dataset [13].

Our synthetic data set ranges over with various *temporal localities*. By varying temporal locality, we can adjust the number of consecutive repetitions of items that are assigned to a server. Such temporal bursts are a typical pattern in communication traffic To formalize temporal locality, we consider the probability of an item being accessed consecutively. The temporal locality parameter has range of [0.1, 0.9], and has the value 0.7 by default to



■ **Figure 4** Comparing average access cost of different algorithms considering the same capacity for all of the algorithms. Figure 4a shows how the cost changes for an increasing number of items. This figure considers 50 servers and 150,000 requests. Figure 4a shows how the cost changes for an increasing number of items. Here, the traditional algorithm stops when a server of it becomes full, and for other items, we run the experiment from scratch. The second figure considers 5,000 items and the rest is similar to previous figure. Both figures consider the CAIDA [13] dataset.

simulate bursts in the input. We use 10,000 and 1,000,000 by default for the number of items and servers, respectively.

**Server insertion and deletion rate.** Given that we aim to utilize most of the servers’ storage, we consider the hardware failure and replacement of servers to be independent of each other. As a result, in our evaluations, we used a Poisson distribution to simulate server (i.e., hard disk) insertion and deletion processes [56].

**Choice of hash function.** The best choice of a hash function for the consistent hashing method has been a long long-held debate [2]. For our plots, we used more practically safe and secure SHA-512 [19]. However, our implementation also supports 5-independent hash functions [15, 52].

**Extra capacity.** Our algorithm,  $H\&A$ , and  $WBL$ ’s algorithm rely on having extra capacity per each server. In our simulations, the default  $\alpha$  value for our algorithm is 4, i.e., four extra slots added to the minimum capacity of each server ( $\lceil \frac{m}{n} \rceil$ ). In our evaluations, we considered the constant factor of 1.25 for  $WBL$ , based on their recommendation [39].

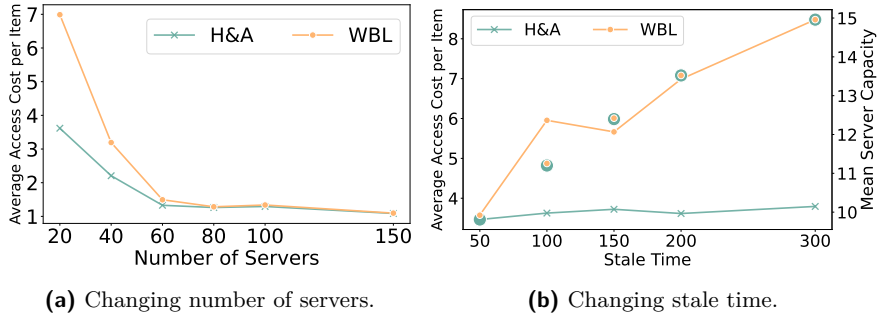
**Number of initial servers.** To bootstrap an instance of consistent hashing, we need to set up a number of servers initially. We consider this value to be equal to 20 by default.

**Stale time.** With *stale time*, we aim to simulate the time-out option for each item stored in an instance of consistent hashing. To do so, we delete an item after a fixed period of time, which has been set to 20 minutes by default in our simulations.

## 6.2 Experimental Results

**Comparison of algorithms.** We compare our three main algorithms in terms of storage utilization and amortized access cost per request, given the real-world dataset. Regarding storage utilization, using  $WBL$  algorithm with a multiplication factor of 1.25, the storage utilization only increases by 40 percent from 50 of the traditional algorithm (cf. Figure 3a). However, as also shown in the figure, this better storage utilization comes with an access cost of 12 per item on average. Using  $H\&A$ , we can achieve closer to 100% storage utilization, 90% to be exact. Furthermore, we require more than 61% less access cost compared to  $WBL$ .

**Effect of temporal locality.** When we increase temporal locality (i.e., the probability of an item being requested consecutively), we can see a decrease in the cost of  $H\&A$ , as shown in Figure 3b. That is because the item that has been moved to its head is going to be



**Figure 5** Comparing the average cost of the *H&A* and the "With Bounded Load" (*WBL*) algorithm [39] based on two parameters of our benchmarking tool. We normalized the access cost by the number of items. The dots in Figure 5b shows the average server capacity by colored dots. First figure considers 1029 items, and 100,000 requests, and in second figure we consider 20 servers. Both figures consider the click dataset [6].

accessed again soon. The increase in the *WBL*'s access cost is because an item from further away might be accessed more often.

**Performance under same capacity.** In this experiment, we consider that all algorithms have *the same* capacity on the servers. As expected, *H&A* performs much better when the additional capacity is small, as shown in Figure 4b. Furthermore, for such a fixed capacity we can see our algorithm can tolerate more items with a lower cost, see Figure 4a.

**Effect of changing the number of servers.** By increasing the number of servers (see Figure 5a), our algorithm and *WBL* both observe a reduced access cost. We can report from our observation that the sharp drop in access cost is because as we increase the number of servers, there is a higher chance that more non-full servers appear between a series of full servers, i.e., the maximum length of consecutive full servers decreases. This result is consistent with the outcome of Lemma 2 that shows that the number of consecutive full servers decreases as the number of servers grows.

**Effect of other parameters.** We consider the effects of two further benchmarking parameters: the number of servers and stale time. By increasing the number of servers (see Figure 5a), both algorithms observe a reduced access cost. We can report from our observation that the sharp drop in access cost is because as we increase the number of servers, there is a higher chance that more non-full servers appear between a series of full servers, i.e., the maximum length of consecutive full servers decreases.

Considering the stale time (Figure 5b), we saw an increase in the maximum capacity needed for servers, which was predictable, as more items will be in the system with a higher stale time. However, the increase in the access cost of *WBL* means that it performs worse in terms of both access cost and memory utilization for an increased stale time.

## 7 Further Related Work

Our work builds upon literature in the context of consistent hashing and self-adjusting data structures. We review them in turn, showing how previous consistent hashing ideas result in increasing storage utilization, and how self-adjustments can lead to decreasing access time.

**Consistent hashing and storage utilization.** Consistent hashing was first introduced by Karger et al. [31], building on top of the classical problem of "balls into bins" [9]. The concept quickly became very important for distributed and peer-to-peer systems [35,42]. The classic use case of consistent hashing is in the design of distributed hash tables in peer-to-peer

systems like Chord [50] or Kademlia [37] and more [36, 40]. The second use case of consistent hashing is in web caching [14, 33] in cloud systems (our experimental results explore this second scenario).

Recently, a team at Google implemented consistent hashing with bounded loads, showing the effects of bounding load in their content hosting system [39]. Inspired by the work of Google, the team behind Viemo load balancer implemented a similar algorithm as part of HAProxy [51] that resulted in 8-fold improvement in their cache bandwidth [45].

Traditionally, consistent hashing approaches abstract the routing protocol, i.e., allowing for flexible routing, as we considered in our paper. Some of the related works suggest certain approaches for routing. One of the proposals is using *virtual servers* [1, 16, 24, 32, 55], in which a server is not responsible only for one hash value; but two or more. On top of complicated routing, this method requires more storage to store the relation of each server with each hash function. Another idea is to change the assignment of the next server for an item based on the number of full servers it observed so far [14]. This approach requires dedicated routing tables, and can only support insertion and deletion in specific scenarios.

Furthermore, we point out two differences of our model with traditional caching models [20, 46]. Firstly, in such models cache hits come with zero cost, while in our model, accessing the first server has a cost of one. Secondly, in our model, all servers provide the same speed in terms of accessing items, however, in the caching hierarchies, usually we see a slowdown when we get further away from processing units.

**Self-adjusting data structures and decreasing access time.** Self-adjusting data structures have been explored for almost half a century now [29]. The pioneering work of Sleator and Tarjan [49] used the amortized analysis for the online algorithm of list update. To the best of our knowledge (check [3] for an updated list of papers), there are no results on the list update problem with capacity or multiple heads.

The notion of a self-adjusting hash table was first suggested by Pagli in 1985 [43] and then by Wogulis [54], both as heuristics. Self-adjusting hashing gained attention recently [28], especially when Microsoft’s team proposed VIP Hashing [30].

However, there are significant differences between our work, and the other suggested self-adjusting hash tables. First of all, we emphasize that comparing the performance of our algorithm with methods of [28, 30, 43, 54] is not possible: they aim to optimize *intra*-server costs, but we aim to minimize the *inter*-server costs. On top of that, our work provides additional benefits such as: (1) proving competitive guarantees formally, (2) balancing between storage utilization and access cost, and (3) analyzing the effects of temporal locality. Furthermore, compared to VIP Hashing, our algorithm does not require a complicated learning phase and can be implemented with the change of a few lines in existing systems.

## 8 Conclusion & Future work

Motivated by the dynamic temporal structure of demands in networked and distributed systems, we have introduced an adaptive approach to improve the performance of distributed hash tables. Our main contribution is a constant-competitive algorithm for self-adjustments guaranteeing bounded loads.

In our future research, we aim to improve our approximation bounds and explore opportunities to render other distributed datastructures self-adjusting. From a practical point of view, we are working on incorporating our algorithms into real-world applications. In particular, we are working on an updated code for HAProxy, and also to improve load balancing for virtual IP address assignment (e.g., as proposed by Google [10]).

---

## References

- 1 Anders Aamand, Jakob Bæk Tejs Knudsen, and Mikkel Thorup. Load balancing with dynamic set of balls and bins. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 1262–1275. ACM, 2021. doi:10.1145/3406325.3451107.
- 2 Anders Aamand and Mikkel Thorup. Non-empty bins with simple tabulation hashing. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2498–2512. SIAM, 2019. doi:10.1137/1.9781611975482.153.
- 3 Vamsi Addanki, Maciej Pacut, Arash Pourdamghani, Gábor Rétvári, Stefan Schmid, and Juan Vanerio. Self-adjusting partially ordered lists. In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications, New York City, NY, USA, May 17-20, 2023*, pages 1–10. IEEE, 2023. doi:10.1109/INFOCOM53939.2023.10228937.
- 4 Susanne Albers. Online algorithms: a survey. *Math. Program.*, 97(1-2):3–26, 2003. doi:10.1007/s10107-003-0436-0.
- 5 Berk Atikoglu, Yuehai Xu, Eitan Frachtenberg, Song Jiang, and Mike Paleczny. Workload analysis of a large-scale key-value store. In Peter G. Harrison, Martin F. Arlitt, and Giuliano Casale, editors, *ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS '12, London, United Kingdom, June 11-15, 2012*, pages 53–64. ACM, 2012. doi:10.1145/2254756.2254766.
- 6 Avazu clickthrough rate prediction. URL: <https://www.kaggle.com/c/avazu-ctr-prediction>.
- 7 Chen Avin, Manya Ghobadi, Chen Griner, and Stefan Schmid. On the complexity of traffic traces and implications. In *ACM SIGMETRICS*, 2020. doi:10.1145/3393691.3394205.
- 8 Chen Avin and Stefan Schmid. Toward demand-aware networking: a theory for self-adjusting networks. *Comput. Commun. Rev.*, 48(5):31–40, 2018. doi:10.1145/3310165.3310170.
- 9 Yossi Azar, Andrei Z. Broder, Anna R. Karlin, and Eli Upfal. Balanced allocations (extended abstract). In Frank Thomson Leighton and Michael T. Goodrich, editors, *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, 23-25 May 1994, Montréal, Québec, Canada*, pages 593–602. ACM, 1994. doi:10.1145/195058.195412.
- 10 Betsy Beyer, Chris Jones, Jennifer Petoff, and Niall Richard Murphy. *Site reliability engineering: How Google runs production systems*. " O'Reilly Media, Inc.", 2016.
- 11 Allan Borodin and Ran El-Yaniv. *Online computation and competitive analysis*. Cambridge University Press, 1998.
- 12 Lee Breslau, Pei Cao, Li Fan, Graham Phillips, and Scott Shenker. Web caching and zipf-like distributions: Evidence and implications. In *Proceedings IEEE INFOCOM '99, The Conference on Computer Communications, Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, The Future Is Now, New York, NY, USA, March 21-25, 1999*, pages 126–134. IEEE Computer Society, 1999. doi:10.1109/INFOCOM.1999.749260.
- 13 P CAIDA. The caida ucsd anonymized internet traces 2016, 2019.
- 14 John Chen, Benjamin Coleman, and Anshumali Shrivastava. Revisiting consistent hashing with bounded loads. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3976–3983. AAAI Press, 2021. doi:10.1609/aaai.v35i5.16517.
- 15 Tobias Christiani and Rasmus Pagh. Generating k-independent variables in constant time. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pages 196–205. IEEE Computer Society, 2014. doi:10.1109/FOCS.2014.29.
- 16 Frank Dabek, M. Frans Kaashoek, David R. Karger, Robert Tappan Morris, and Ion Stoica. Wide-area cooperative storage with CFS. In Keith Marzullo and Mahadev Satyanarayanan, editors, *Proceedings of the 18th ACM Symposium on Operating System Principles, SOSP 2001*,



- Chateau Lake Louise, Banff, Alberta, Canada, October 21-24, 2001, pages 202–215. ACM, 2001. doi:10.1145/502034.502054.
- 17 Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Voshall, and Werner Vogels. Dynamo: amazon’s highly available key-value store. In Thomas C. Bressoud and M. Frans Kaashoek, editors, *Proceedings of the 21st ACM Symposium on Operating Systems Principles 2007, SOSP 2007, Stevenson, Washington, USA, October 14-17, 2007*, pages 205–220. ACM, 2007. doi:10.1145/1294261.1294281.
  - 18 Peter J Denning. The working set model for program behavior. *Commun. ACM*, 1968. doi:10.1145/363095.363141.
  - 19 Christoph Dobraunig, Maria Eichlseder, and Florian Mendel. Analysis of SHA-512/224 and SHA-512/256. *IACR Cryptol. ePrint Arch.*, 2016. URL: <http://eprint.iacr.org/2016/374>.
  - 20 Amos Fiat, Richard M. Karp, Michael Luby, Lyle A. McGeoch, Daniel Dominic Sleator, and Neal E. Young. Competitive paging algorithms. *J. Algorithms*, 1991. doi:10.1016/0196-6774(91)90041-V.
  - 21 Aleksander Figiel, Janne H. Korhonen, Neil Olver, and Stefan Schmid. Efficient algorithms for demand-aware networks and a connection to virtual network embedding. In *International Conference on Principles of Distributed Systems (OPODIS)*, 2024. doi:10.4230/LIPIcs.CVIT.2016.23.
  - 22 Aleksander Figiel, Darya Melnyk, Andre Nichterlein, Arash Pourdamghani, and Stefan Schmid. Spiderdan: Matching augmentation in demand-aware networks. In *SIAM Symposium on Algorithm Engineering and Experiments (ALENEX)*, 2025.
  - 23 Alexander Fuerst and Prateek Sharma. Locality-aware load-balancing for serverless clusters. In Jon B. Weissman, Abhishek Chandra, Ada Gavrilovska, and Devesh Tiwari, editors, *HPDC ’22: The 31st International Symposium on High-Performance Parallel and Distributed Computing, Minneapolis, MN, USA, 27 June 2022 - 1 July 2022*, pages 227–239. ACM, 2022. doi:10.1145/3502181.3531459.
  - 24 Brighten Godfrey, Karthik Lakshminarayanan, Sonesh Surana, Richard M. Karp, and Ion Stoica. Load balancing in dynamic structured P2P systems. In *Proceedings IEEE INFOCOM 2004, The 23rd Annual Joint Conference of the IEEE Computer and Communications Societies, Hong Kong, China, March 7-11, 2004*, pages 2253–2262. IEEE, 2004. doi:10.1109/INFCOM.2004.1354648.
  - 25 Nicholas J. A. Harvey. A first course in randomized algorithms. *Book draft*, 2022. doi:10.48550/arXiv.cs/0601026.
  - 26 Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *JASA*, 1963.
  - 27 John D. Hunter. Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.*, 2007. doi:10.1109/MCSE.2007.55.
  - 28 Kaiyi Ji, Guocong Quan, and Jian Tan. Asymptotic miss ratio of LRU caching with consistent hashing. In *IEEE INFOCOM*. IEEE, 2018. doi:10.1109/INFOCOM.2018.8485860.
  - 29 Edward G. Coffman Jr. and Peter J. Denning. *Operating Systems Theory*. Prentice-Hall, 1973.
  - 30 Aarati Kakaraparthi, Jignesh M. Patel, Brian Kroth, and Kwanghyun Park. VIP hashing - adapting to skew in popularity of data on the fly. *VLDB*, 2022. doi:10.14778/3547305.3547306.
  - 31 David R. Karger, Eric Lehman, Frank Thomson Leighton, Rina Panigrahy, Matthew S. Levine, and Daniel Lewin. Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the world wide web. In Frank Thomson Leighton and Peter W. Shor, editors, *Proceedings of the Twenty-Ninth Annual ACM Symposium on the Theory of Computing, El Paso, Texas, USA, May 4-6, 1997*, pages 654–663. ACM, 1997. doi:10.1145/258533.258660.
  - 32 David R. Karger and Matthias Ruhl. Simple efficient load balancing algorithms for peer-to-peer systems. In Phillip B. Gibbons and Micah Adler, editors, *SPAA 2004: Proceedings of the*

- Sixteenth Annual ACM Symposium on Parallelism in Algorithms and Architectures, June 27-30, 2004, Barcelona, Spain*, pages 36–43. ACM, 2004. doi:10.1145/1007912.1007919.
- 33 David R. Karger, Alex Sherman, Andy Berkheimer, Bill Bogstad, Rizwan Dhanidina, Ken Iwamoto, Brian Kim, Luke Matkins, and Yoav Yerushalmi. Web caching with consistent hashing. *Comput. Networks*, 1999. doi:10.1016/S1389-1286(99)00055-9.
  - 34 Avinash Lakshman and Prashant Malik. Cassandra: a decentralized structured storage system. *ACM SIGOPS Oper. Syst. Rev.*, 2010. doi:10.1145/1773912.1773922.
  - 35 John Lamping and Eric Veach. A fast, minimal memory, consistent hash algorithm. *CoRR*, 2014. doi:10.48550/arXiv.1406.2294.
  - 36 Eng Keong Lua, Jon Crowcroft, Marcelo Pias, Ravi Sharma, and Steven Lim. A survey and comparison of peer-to-peer overlay network schemes. *IEEE Commun. Surv. Tutorials*, 2005. doi:10.1109/COMST.2005.1610546.
  - 37 Petar Maymounkov and David Mazières. Kademlia: A peer-to-peer information system based on the XOR metric. In *IPTPS*. Springer, 2002. doi:10.1007/3-540-45748-8\_5.
  - 38 Seyedehmehrnaz Mireslami, Logan Rakai, Mea Wang, and Behrouz Homayoun Far. Dynamic cloud resource allocation considering demand uncertainty. *IEEE Trans. Cloud Comput.*, 2021. doi:10.1109/TCC.2019.2897304.
  - 39 Vahab S. Mirrokni, Mikkel Thorup, and Morteza Zadimoghaddam. Consistent hashing with bounded loads. In Artur Czumaj, editor, *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 587–604. SIAM, 2018. doi:10.1137/1.9781611975031.39.
  - 40 Moni Naor and Udi Wieder. Novel architectures for P2P applications: the continuous-discrete approach. In *SPAA*. ACM, 2003. doi:10.1145/777412.777421.
  - 41 Petros Nicosopolitidis, Georgios I Papadimitriou, and Andreas S Pomportsis. Exploiting locality of demand to improve the performance of wireless data broadcasting. *IEEE Trans. Veh. Technol.*, 2006. doi:10.1109/TVT.2006.877464.
  - 42 M. Tamer Özsu and Patrick Valduriez. *Principles of Distributed Database Systems, 4th Edition*. Springer, 2020. doi:10.1007/978-3-030-26253-2.
  - 43 Linda Pagli. Self-adjusting hash tables. *Inf. Process. Lett.*, 1985. doi:10.1016/0020-0190(85)90103-6.
  - 44 Arash Pourdamghani, Chen Avin, Robert Sama, and Stefan Schmid. Seedtree: A dynamically optimal and local self-adjusting tree. In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications, New York City, NY, USA, May 17-20, 2023*, pages 1–10. IEEE, 2023. doi:10.1109/INFOCOM53939.2023.10228999.
  - 45 Andrew Rodland. Improving load balancing with a new consistent-hashing algorithm. *Vimeo Engineering Blog, Medium*, 2016.
  - 46 Tim Roughgarden. Beyond worst-case analysis. *Commun. ACM*, 2019. doi:10.1145/3232535.
  - 47 Arjun Roy, Hongyi Zeng, Jasmeet Bagga, George Porter, and Alex C. Snoeren. Inside the social network’s (datacenter) network. In Steve Uhlig, Olaf Maennel, Brad Karp, and Jitendra Padhye, editors, *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication, SIGCOMM 2015, London, United Kingdom, August 17-21, 2015*, pages 123–137. ACM, 2015. doi:10.1145/2785956.2787472.
  - 48 Swaminathan Sivasubramanian. Amazon dynamodb: a seamlessly scalable non-relational database service. In K. Selçuk Candan, Yi Chen, Richard T. Snodgrass, Luis Gravano, and Ariel Fuxman, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 729–730. ACM, 2012. doi:10.1145/2213836.2213945.
  - 49 Daniel Dominic Sleator and Robert Endre Tarjan. Amortized efficiency of list update rules. In Richard A. DeMillo, editor, *Proceedings of the 16th Annual ACM Symposium on Theory of Computing, April 30 - May 2, 1984, Washington, DC, USA*, pages 488–492. ACM, 1984. doi:10.1145/800057.808718.

- 50 Ion Stoica, Robert Tappan Morris, David Liben-Nowell, David R. Karger, M. Frans Kaashoek, Frank Dabek, and Hari Balakrishnan. Chord: a scalable peer-to-peer lookup protocol for internet applications. *IEEE/ACM Trans. Netw.*, 2003. doi:10.1109/TNET.2002.808407.
- 51 Willy Tarreau et al. Haproxy-the reliable, high-performance tcp/http load balancer. <https://www.haproxy.org>, 2012.
- 52 Mikkel Thorup and Yin Zhang. Tabulation-based 5-independent hashing with applications to linear probing and second moment estimation. *SIAM J. Comput.*, 2012. doi:10.1137/100800774.
- 53 Michael L. Waskom. seaborn: statistical data visualization. *J. Open Source Softw.*, 6(60):3021, 2021. doi:10.21105/joss.03021.
- 54 James Wogulis. Self-adjusting and split sequence hash tables. *Inf. Process. Lett.*, 1989. doi:10.1016/0020-0190(89)90210-X.
- 55 Min Xiang, Yuzhou Jiang, Zhong Xia, and Chunmei Huang. Consistent hashing with bounded loads and virtual nodes-based load balancing strategy for proxy cache cluster. *Clust. Comput.*, 2020. doi:10.1007/s10586-020-03076-4.
- 56 Zhisheng Ye, Min Xie, and Loon-Ching Tang. Reliability evaluation of hard disk drive failures based on counting processes. *Reliab. Eng. Syst. Saf.*, 2013. doi:10.1016/j.ress.2012.07.003.

## A Omitted Proofs

In this section, we present the proofs omitted in the main body of the paper.

### A.1 Proof of Theorem 1

Before going into details of the proof, we define two concepts: *position*, and *inversion*. We first look at the position of an item inside the server that currently hosts it at time  $t$ . To define the position of an item, we focus on algorithms that maintain a pre-determined ordering of items inside all of their servers. We developed  $H\mathcal{E}A$  such that it has this property, and as discussed before, we consider such an  $OPT$  algorithm.

► **Definition 3** (Inside position). *The inside position of an item  $v$  at time  $t$  is the number of items that are in front of it in the ordered list of its host,  $host^t(v)$ .*

The server that contains an item  $v$  might change over time. This might be because the  $head(v)$  was full during the insertion of  $v$ , or due to reconfigurations of other items. In the following, we define the relative position for all items at any given time compared to any given head in a ServerList.

► **Definition 4** (Headed relative position). *The headed relative position of an item  $v$  that is in front of the head  $H$  in an algorithm  $ALG$  at time  $t$ ,  $pos_{ALG}^t(v, H)$ , is its inside position at time  $t$ , plus  $c$  times the number of servers between  $host^t(v)$  and  $head(v)$ .*

The position for an item  $v$  at time  $t$  is its relative position to its head, i.e.,  $pos_{ALG}^t(v) = pos_{ALG}^t(v, head(v))$ .

We now define an *asymmetric inversion*. We say there is an inversion between a pair of items  $u$  and  $v$  if they are in different orders in the  $H\mathcal{E}A$  and  $OPT$  ServerLists. The definition of inversion helps us determine “how far”  $H\mathcal{E}A$ ’s ServerList is from  $OPT$ ’s ServerList, which is key in comparing the cost of  $H\mathcal{E}A$  and  $OPT$ .

► **Definition 5** (Inversion). *Considering an algorithm  $OPT$ , there exists an asymmetric inversion between items  $u$  and  $v$  at time  $t$  if and only if  $u$  and  $v$  are in front of  $head(u)$  and  $pos_{OPT}^t(v, head(u)) < pos_{OPT}^t(u)$  but  $pos_{H\mathcal{E}A}^t(v, head(u)) > pos_{H\mathcal{E}A}^t(u)$ .*

As an example, let us consider ServerList in Figure 2a as the  $OPT$ ’s ServerList and the one in Figure 2d as  $H\mathcal{E}A$ ’s ServerList. Then there exists an inversion between items 32 and 7, as item 32 is closer to its head ( $Head_1$ ) than 7 in  $OPT$ ’s ServerList, but vice versa in  $H\mathcal{E}A$ ’s ServerList.

We emphasize that inversions are asymmetric, i.e., in one direction:  $u$  might have an inversion with  $v$  but not the other way around. Furthermore, as we considered an  $OPT$  algorithm that orders items inside a server as they have been moved to that server, similar to  $H\mathcal{E}A$ , ordering of items does inside a server not create an inversion.

We consider an indicator variable  $I^t(u, v)$ , which equals 1 if there is an inversion between  $u$  and  $v$ , 0 otherwise. Then we define the number of inversions in the ServerList of  $OPT$  at time  $t$ , as the sum of inversions for all items, and call it  $I^t$ . In other words, we have  $I^t = \sum_{u, v \in V} I^t(u, v)$ .

Using the above-mentioned definitions, next we show that our algorithm is constant competitive, using a potential function analysis for our model that considers a large enough request sequence.

**Proof of Theorem 1.** In the following proof, we consider the ServerList of  $OPT$  at time  $t$ , and see the effects of an access request at time  $t + 1$ . To compute the amortized cost of  $H\mathcal{E}A$ , we consider the following potential function:  $\Phi^t = \frac{(1+\omega)}{c} \cdot I^t$ . Consider  $C_{H\mathcal{E}A}^t$  as the cost of system at time  $t$ , and  $\Delta\Phi^{t \rightarrow t+1}$  as the change in potential from time  $t$  to  $t + 1$ . We focus on the changes in the potential function, as  $OPT$  and  $H\mathcal{E}A$  start from the same configuration, hence  $\Phi^0 = 0$ . Furthermore, given that the number of inversions is always non-negative ( $I^t \geq 0, \forall t \geq 0$ ), the potential function is always larger than 0 ( $\Phi^t \geq 0, \forall t \geq 0$ ) by definition.

Then, the amortized cost of  $H\mathcal{E}A$  at time  $t$  ( $A_{H\mathcal{E}A}^t$ ) is defined as the cost of  $H\mathcal{E}A$  at  $t$  plus the change in the potential:  $A_{H\mathcal{E}A}^t = C_{H\mathcal{E}A}^t + \Delta\Phi^{t \rightarrow t+1}$ .

Now, we discuss changes in the potential function, i.e., changes in the number of inversions. We divide these changes into two cases: right after  $v$  was accessed in  $H\mathcal{E}A$ , and after possible changes by  $OPT$ , and conclude by total changes in potential function.

**Right after  $v$  was accessed.** We start by considering inversions that include  $v$ .

1. *Increase in the number of inversions.* Right after  $v$  was accessed, its position in the ServerList of  $H\mathcal{E}A$  is 0, as we move  $v$  to its head. At this point, some inversions might be created for items between  $v$  and its head in ServerList of  $OPT$ . There are at most  $pos_{OPT}^t(v)$  items between  $v$  and its head in the  $OPT$  ServerList; hence the number of inversions created is at most  $pos_{OPT}^t(v)$ .
2. *Decrease in the number of inversions.* On the other hand, all the previous inversions of the  $v$  get destroyed. Those items were before  $v$  in the  $H\mathcal{E}A$  ServerList (there are  $pos_{H\mathcal{E}A}^t(v)$  items), and are positioned after  $v$  in  $OPT$  ServerList (at most  $pos_{OPT}^t(v)$  of those items can be positioned before  $v$  in  $OPT$ ). Hence at least  $pos_{ALG}^t(v) - pos_{OPT}^t(v)$  inversions are destroyed.

As the ordering of other items does not change in  $H\mathcal{E}A$ 's ServerList, their relative position does not change (keep in mind that the position consists of *inside position*). Therefore, inversions that do not include  $v$  are not created nor removed.

**After possible changes by  $OPT$ .** Each reconfiguration of  $OPT$  creates at most  $c$  new inversions, besides some inversions that might be removed. This is because, at most, one item might move from one server to the other, and might cause inversion with all items that were in its previous server. Assuming that  $OPT$  has moved  $r$  times,  $OPT$  movements result in maximum  $c \cdot r$  change in the potential function.

**Total changes in potential function.** Combining all cases, the total change in the potential function from time  $t$  to  $t + 1$  is:

$$\begin{aligned} \Delta\Phi^{t \rightarrow t+1} &= \frac{(1+\omega)}{c} \cdot (I^{t+1} - I^t) \leq \frac{(1+\omega)}{c} \cdot (pos_{OPT}^t(v) - [pos_{H\mathcal{E}A}^t(v) - pos_{OPT}^t(v)] + c \cdot r) = \\ &= \frac{(1+\omega)}{c} \cdot (2 \cdot pos_{OPT}^t(v) - pos_{H\mathcal{E}A}^t(v) + c \cdot r) \end{aligned}$$

Considering Equation 2 from Section 3, the total cost of access request to an item  $v$  by  $H\mathcal{E}A$  ( $C_{H\mathcal{E}A}^t$ ) is  $\frac{(1+\omega)}{c} \cdot pos_{H\mathcal{E}A}^t(v)$ . That is because:

- $H\mathcal{E}A$  immediately moves back the item  $v$  from its host to its head (as described in Algorithm 1). Given Invariant 3, the access cost of our algorithm equals its position at the ServerList of  $H\mathcal{E}A$   $\frac{pos_{H\mathcal{E}A}^t(v)}{c}$ ,
- given that each reconfiguration cost  $\omega$  times the access, the reconfiguration cost is  $\omega \cdot \frac{pos_{H\mathcal{E}A}^t(v)}{c}$

Now we go back to amortized cost of  $H\mathcal{E}A$  at time  $t$  ( $A_{H\mathcal{E}A}^t$ ):

$$A_{H\mathcal{E}A}^t = C_{H\mathcal{E}A}^t + \Delta\Phi^{t \rightarrow t+1} \leq \frac{(1+\omega)}{c} \cdot \text{pos}_{H\mathcal{E}A}^t(v) + \frac{(1+\omega)}{c} \cdot (2 \cdot \text{pos}_{OPT}^t(v) - \text{pos}_{H\mathcal{E}A}^t(v) + c \cdot r) =$$

$$\frac{(1+\omega)}{c} \cdot (2 \cdot \text{pos}_{OPT}^t(v) + c \cdot r)$$

Then we compare the cost of *OPT* with the amortized cost of the system. We know the cost of *OPT* is the access cost due to the position of  $v$  in *OPT*, i.e.,  $\frac{\text{pos}_{OPT}^t(v)}{c}$ , plus the reconfiguration cost,  $r \cdot \omega$ , hence the total cost of *OPT* at time  $t$  ( $C_{OPT}^t$ ) is  $\frac{\text{pos}_{OPT}^t(v)}{c} + r \cdot \omega$ . Then we have:

$$A_{H\mathcal{E}A}^t \leq \frac{(1+\omega)}{c} \cdot (2 \cdot \text{pos}_{OPT}^t(v) + c \cdot r) < 2 \cdot \frac{(1+\omega)}{c} (\text{pos}_{OPT}^t(v) + c \cdot r) \leq$$

$$2 \cdot (1+\omega) \left( \frac{\text{pos}_{OPT}^t(v)}{c} + r \cdot \omega \right) \leq 2 \cdot (1+\omega) \cdot C_{OPT}^t$$

The second to last step is possible as we assume  $\omega \geq 1$ . Given the last inequality, we can ensure that the Algorithm 1 is  $2 \cdot (1+\omega)$  competitive.

Given that the potential function is initially 0 (both our algorithm and the optimal algorithm start from the same state) and is always positive, by summing up costs at each time step, we have:

$$A_{H\mathcal{E}A} \leq 2 \cdot (1+\omega) \cdot C_{OPT}$$

◀

## A.2 Proof of Lemma 3

**Proof.** We start this proof by considering the case that the capacity of all servers only changes by one. In the end, we discuss the effect of changes by a higher amount (which might happen due to server insertion/deletion). Recall that for a previously full  $s$ , we take the item to the next non-full server in case of increased capacity, or move the item to the next non-full server in case of decreased capacity. In the rest we focus on capacity increase, capacity decrease follows similarly (details at the end).

Let us call the length if  $j$ -th MSCFS  $L_j$ . What we want to compute is  $E[\sum \binom{L_j}{2}]$ , since in a server list with length  $L_j$ , we need to move one item from the non-full server  $L_j$  times to the head server, the item afterwards  $L_j - 1$  times to the server after head and so on. We know that

$$E\left[\sum \binom{L_j}{2}\right] = \frac{1}{2} \cdot E\left[\sum L_j \cdot (L_j - 1)\right] \leq \frac{1}{2} E\left[\sum L_j^2\right]$$

We know that  $\sum L_j \leq n$ . Let us define  $L_{max}$  as the length of MSCFS with maximum length. We know  $L_j \leq L_{max}$  holds for all random variables  $L_j$ , hence we have:

$$E\left[\sum L_j^2\right] \leq E[L_{max} \cdot \sum L_j] \leq E[L_{max} \cdot n] = n \cdot E[L_{max}]$$

And we have the value  $E(L_{max}) = \sum_{i=1}^{n-1} e^{-\frac{\alpha^2}{2m}(i+1)^2}$  from 2. On the other hand, we know that *OPT* needs to maintain search property, so it needs to at least move  $\frac{n}{2}$  items. Furthermore, Assumption 1 about well-behaved request sequence states that insertion deletions happen only  $\sum_{i=1}^{n-1} e^{-\frac{\alpha^2}{2m}(i+1)^2} = E(L_{max})$  times. Hence, the amortized cost of capacity increase is  $\frac{n \cdot E(L_{max})}{\frac{n}{2} \cdot E(L_{max})}$ , i.e. 2 in expectation.

In case of capacity decrease, we know there is always a non-full server in  $H\mathcal{E}A$  (due to the additive extra capacity, on top of the minimum capacity required). Hence, there is no wrap-around (i.e., we do not go over the server  $s$  again). Furthermore, since  $m \gg n$ , the capacity of the last server is increased enough to cover the incoming items. Therefore, the cost of the algorithm depends on the length of intervals of the full servers. In case of capacity change by more than one, the cost of  $OPT$  also increases, hence the constant increase competitive ratio remains. ◀

### A.3 Proof of Theorem 2

**Proof.** We start this proof by looking back at the competitiveness of access operations *given* item (or server) insertion (or deletion), and then focus on the competitiveness of those operations themselves. We emphasize that we compare  $H\mathcal{E}A$  with an optimal offline algorithm that has the same capacity for all servers. Furthermore, among many possible optimal offline algorithms, we consider that the optimal offline algorithm also respects Invariant 2, otherwise, its search mechanism would not be compatible with the standard search mechanism [39].

**Competitiveness of access operations.** We have concluded in Theorem 1 that an access operation inside a ServerList is constant competitive. We now detail why that remains true considering items (or server) insertion (or deletion).

Given Invariant 1 from the previous section, the relative order of items that were previously in the system remains the same after insertions or deletions; in other words, the least recently used items are closer to their own head in each ServerList. Furthermore, the newly added item itself is always added at the end of a ServerList; hence it does not affect the order of items in that or any given ServerLists.

After an insertion or deletion, there is a possibility that a ServerList splits (or a newly added server becomes its own server), or two ServerLists might join together:

- In case of a ServerList splitting, there would be no item left of the first ServerList in the other ServerList. Hence, there would not be any shared inversions between the two split ServerLists, i.e., there would not be an inversion that consists of an item in one ServerList and another item in the other. Hence, our potential function which is based on inversions can be computed considering inversions of two ServerLists separately.
- In case of two ServerLists merging, as two ServerLists lists did not have a shared item before, there were no shared inversions. Therefore, the potential function after the merge does not differ from when two ServerLists were split, i.e., the sum of potential functions of previously split ServerLists.

Therefore, in both cases, the arguments in Theorem 1 still hold and the constant competitiveness of  $H\mathcal{E}A$  holds.

**Competitiveness of other operations.** We discussed in Lemma 3 why capacity changes are competitive. As the further cost of item insertion (or deletion) is to move an item over a server list to the next non-full server (or bring the item from the non-full server in case of deletion), the extra cost depends on the length of consecutive full-server. Given the Assumption 1, we can offset this cost, since we assume insertion and deletion happens at a rate equal to the expected length of consecutive full servers.

Server deletion and insertion do not introduce an extra cost themselves, but rather the movement of items is the costly part. In other words, we can consider a server insertion as a series of item deletions from other servers first and then item insertions to that server. Given that any  $OPT$  algorithm needs to pay at least 1 for each movement of items, we can simply apply what we discussed before for individual item insertion and deletions. ◀

## B Parameters of the benchmarking.

In the Table 2, you can see a summary of the parameters that we used in our benchmarking.

Parameter	Range of values	Default
Input generation	Click <sup>1)</sup> or Temporal <sup>2)</sup>	Click
Locality parameter	[0.1, 0.9]	0.70
Number of Items (m)	$[10^3, 10^9]$ items	10,000
Number of Requests	$[10^4, 10^{13}]$ requests	1,000,000
Server insertion frequency	[100, 1000] minutes	200
Server deletion frequency	[100, 1000] minutes	200
Placement Algorithm	<i>H&amp;A</i> , WBL and Trad. <sup>3)</sup>	<i>H&amp;A</i>
Hash Function	5-Ind. or SHA-512	SHA-512
Additive extra capacity	$[+1, +\lceil \frac{m}{n} \rceil]$	+4
Multiplicative extra capacity	[1, 2]	+1.25
Number of Initial Servers (n)	[20, 10000] servers	20
Stale time	[20, 200] minutes	200

- 1) Refers to the real-world click dataset [6],
- 2) is the temporal locality generation method.
- 3) "Traditional" method from [31].

■ **Table 2** Parameters and system that can be used in our benchmarking tool.