False Rate Analysis of Bloom Filter Replicas in Distributed Systems

Yifeng Zhu
Electrical and Computer Engineering
University of Maine
zhu@eece.maine.edu

Hong Jiang
Computer Science and Engineering
University of Nebraska - Lincoln
jiang@cse.unl.edu

Abstract

Recently, Bloom filters have been widely used in distributed systems where they are replicated to process distributed queries. Bloom filter replicas become stale in a dynamic environment. A good understanding of the impact of staleness on false negatives and false positives can provide the system designers with important insights into the development and deployment of distributed Bloom filters in many distributed systems. To our best knowledge, this paper is the first one that analyzes the probabilities of false negatives and positives by developing analytical models, which take the staleness into consideration. Based on the theoretical analysis, we proposed an updating protocol that directly control the false rate. Extensive simulations validate the analytical models and prove the updating protocol to be very accurate and effective.

1 Introduction

A Bloom filter (BF) [1] is a lossy but succinct and efficient data structure to represent a set $\mathcal{S}$, which processes the membership query "is x in $\mathcal{S}$?" for any given element x with a time complexity of $O(1)$. Its storage requirement falls several orders of magnitude below the lower bounds of error-free encoding structures. This space efficiency is achieved at the cost of allowing a certain (typically nonezero) probability of *false positives*, that is, it may incorrectly return an "yes" although x is actually not in $\mathcal{S}$. Tuning the parameters of a BF can minimize this probability of false-positive to a sufficiently small value so that benefits from the space and time efficiency far outweigh the penalty incurred by false positives in many applications.

In fact, BFs have shown great potentials in many distributed systems where information physically disseminated across the entire system needs to be shared. For example, to reduce the message traffic, Ref. [2] proposes a web cache sharing protocol that employs a BF to represent the content of a cache in a web proxy and then periodically propagates that filter to other proxies. If a cache miss occurs at a local proxy, that proxy checks the BFs replicated from other proxies to see whether they have the desired web objects in their caches. Ref. [3, 4, 5, 6] use BFs to implement the function of mapping logical data identities to their physical locations in distributed storage systems. In such schemes, each storage node constructs a Bloom filter that summarizes the identities of data stored locally and broadcasts it to other nodes. By checking all filters collected locally, a node can locate the requested data without sending massive query messages to other nodes. Similar deployments of BFs have been found in geographic routing in wireless mobile systems [7]), P2P systems [8, 9, 10, 11] and naming services [12].

A common characteristics of distributed applications of BFs, including all those described above, is that *a BF at a local host is replicated to other remote hosts to efficiently process distributed queries*. In such dynamical distributed applications, the information that a BF represents evolves over time. However, the updating processes are usually delayed due to the network latency or the delay necessary in aggregating small changes into single updating message in order to reduce the updating overhead. Accordingly the contents of the remote replicas may become partially outdated. This possible staleness in the remote replicas not only changes the probability of false positive answers to membership queries on the remote hosts, but also brings forth the possibility of *false negatives*. A false negative occurs when a BF replica answers "no" to the membership query for an element while that element actually exists in its host. It is generated when a new element is added to a host while the changes of the BF of this host, including the addition of this new element, have not been propagated to its replicas on other hosts. In addition, this staleness also changes the probability of *false positives*, an event in which an element is incorrectly identified as a member. Throughout the rest of this paper, the probabilities of false negatives and false positives are referred to as the false negative rate and false positive rate, respectively.

While the false negative and false positive rates for a BF

at a local host have been well studied in the context of non-replicated BF [1, 13, 2, 14, 15], very little attention has been paid to the false rates in the Bloom filter replicas in a distributed environment. In the distributed systems considered in this paper, the false rates of the replicas are more important since most membership queries are performed on these replicas. A good understanding of the impact of the false negatives and false positives can provide the system designers with important and useful insights into the development and deployment of distributed BFs in such important applications as distributed file, database, and web server management systems in super-scales. Therefore, the first objective of this paper is to analyze the false rates by developing analytical models and considering the staleness.

Since different application may desire a different tradeoff between false rate (e.g, miss/fault penalty) and update overhead (e.g., network traffic and processing due to broadcasting of updates), it is very important and significant for the systems overall performance to be able to control such a tradeoff for a given application adaptively and efficiently. The second objective is to develop an adaptive control algorithm that can accurately and efficiently maintain a desirable level of false rate for any given application by dynamically and judiciously adjusting the update frequency.

The primary contribution of this paper is its developments of accurate closed-form expressions for the false negative and false positive rates in BF replicas, and the development of an adaptive replica-update control, based on our analytical model, that accurately and efficiently maintains a desirable level of false rate for any given application. To the best of our knowledge, this study is the first of its kind that has considered the impact of staleness of replicated BF contents in a distributed environment, and developed a mechanism to adaptively minimize such an impact so as to optimize systems performance.

The rest of the paper is organized as follows. Section 2 outlines the basic mathematical foundations of BFs. Section 3 presents our analytical models that theoretically derive false negative and false positive rates of a BF replica, as well as the overall false rates in distributed systems. Section 4 validates our theoretical results by comparing them against results obtained from extensive experiments. The adaptive updating protocols based on our theoretical analysis models are presented in Section 5 and Section 6 concludes the paper.

2 Standard (non-replicated) Bloom Filters

To better present our analysis, we begin by introducing the basics of the standard (i.e., non-replicated) BFs, following the analysis and framework of Ref. [15] and [2].

A BF is essentially a bit vector B with m bits that facilitates membership test to a finite set $\mathcal{S} = \{x_1, x_2, \ldots, x_n\}$ of n elements from a universe $\mathcal{U}$. It uses a set $\mathcal{H}(x)$ of k uniform and independent hash functions to map the universe $\mathcal{U}$ to the bit address space $[1, m]$, shown as follows,

$$\mathcal{H}(x) = \{h_i(x) \mid 1 \leq h_i(x) \leq m \text{ for } 1 \leq i \leq k\} \quad (1)$$

Definition 1. *For all $x \in \mathcal{U}$, $B[\mathcal{H}(x)] \equiv \{B[h_i(x)] \mid 1 \leq i \leq k\}$.*

This notation facilitates the description of operations on the subset of B addressed by the hash functions. For example, $B[\mathcal{H}(x)] = 1$ represents the condition in which all the bits in B at the positions of $h_1(x)$, ..., and $h_k(x)$ are "1". "Setting $B[\mathcal{H}(x)]$" means that the bits at these positions in B are set to "1".

Representing the set $\mathcal{S}$ using a BF B is fast and simple. Initially, all the bits in B are set to "0". Then for each $x \in \mathcal{S}$, an operation of setting $B[\mathcal{H}(x)]$ is performed. Given any element x, to check whether x is in $\mathcal{S}$, one only needs to test whether $B[\mathcal{H}(x)] = 1$. If no, then x is said to be out of $\mathcal{S}$; If yes, x is *conjectured* to be in $\mathcal{S}$.

A non-replicated BF has two important properties that are described by the following two theorems respectively.

Theorem 1 (Impossible false negative)**.** *For any $x \in \mathcal{U}$, if $B[\mathcal{H}(x)] \neq 1$, then $x \notin \mathcal{S}$.*

The proof is trivial and is not presented here.

Theorem 2 (Possible false positive)**.** *For any $x \in \mathcal{U}$, if $B[\mathcal{H}(x)] = 1$, then there is a small probability f^+ that $x \notin \mathcal{S}$. This probability is called the false positive rate and $f^+ \approx (1 - e^{-kn/m})^k$. Given a specific ratio of m/n, f^+ is minimized when $k = (m/n)ln2$ and $f^+_{min} \approx (0.6185)^{m/n}$.*

Proof: The proof is based on the mathematical model proposed in Ref. [16, 15]. Detailed proof can be found in Ref. [2] and [14]. For the convenience of the reader, the proof is abbreviated and presented here.

After inserting n elements into BF, the probability that a bit is still not set is given by

$$P_0(n) = \left(1 - \frac{1}{m}\right)^{kn} \approx e^{-kn/m} \quad (2)$$

Thus the probability that k bits are set to 1 is

$$P(\text{k bits set}) = \left(1 - \left(1 - \frac{1}{m}\right)^{kn}\right)^k \approx (1 - e^{-kn/m})^k. \quad (3)$$

Assuming each element is equally likely to be accessed and usually $|\mathcal{S}| \ll |\mathcal{U}|$, then the false positive rate is

$$f^+ = \left(1 - \frac{|\mathcal{S}|}{|\mathcal{U}|}\right) P(\text{k bits set}) \approx (1 - e^{-kn/m})^k. \quad (4)$$

Given a specific ratio of $\frac{m}{n}$, i.e, the number of bits per element, it is easy to prove that the false positive rate f^+ is

minimized when $k = \frac{m}{n} ln2$ and the minimal false positive rate is [14]

$$f^+ \approx 0.5^k = (0.6185)^{m/n} \quad (5)$$

□

Appropriately adjusting m and k can make the false positive rate sufficiently small.

Ref. [2] proposes to use a vector with m counters to facilitate deleting an element x from BF B. More specifically, let $\Gamma = \{\tau_j \mid 1 \leq j \leq m\}$ denote such a counter vector and the counter τ_j represents the difference between the number of settings and the number of unsettings made to the bit $B[j]$. All counters τ_j for $1 \leq j \leq m$ are initialized to zero. When an element x is inserted or deleted, the counters $\Gamma[\mathcal{H}(x)]$ are increased or decreased by one accordingly. If τ_j changes its value from one to zero, $B[j]$ is reset to zero. While this counter vector consumes some memory space, Ref. [2] also shows that 4 bits per counter will guarantee the probability of overflow minuscule even with several hundred million elements in a BF.

3 Bloom Filters in Distributed Systems

In many distributed systems, the information about what data objects can be accessed through a host or where data objects are located usually needs to be shared to facilitate the lookup. To provide high scalability, this information sharing usually takes a decentralized approach, to avoid potential performance bottleneck and vulnerability of a centralized architecture such as a dedicated server. While BFs were initially used in non-distributed systems to save the memory space in the 1980's when memory was considered a precious resource [17, 16], they have recently been extensively used in many distributed systems as a scalable and efficient scheme for information sharing, due to their low network traffic overhead.

The inherent nature of such information sharing in almost all these distributed systems, if not all, can be abstracted as a location identification, or mapping problem, which is described next. Without loss of generality, the distributed system considered throughout this paper is assumed to consist of a collection of γ autonomous data-storing host computers dispersed across a communication network. These hosts partition a universe $\mathcal{U}$ of data objects into γ subsets $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_\gamma$, with each subset stored on one of these hosts. Given an arbitrary object x in $\mathcal{U}$, the problem is how to efficiently identify the host that stores x from any one of the hosts.

BFs are useful to solve this kind of problems. In a typical approach, each host constructs a BF representing the subset of objects stored in it, and then broadcasts that filter to all the other hosts. Thus each host keeps $\gamma - 1$ additional BFs, one for every other host. Figure 1 shows an example of a system with three hosts. Note that a filter $\hat{B}_i$ is a replica of B_i from Host i and $\hat{B}_i$ may become outdated if the changes to B_i are not propagated instantaneously. While the solution to the above information sharing problem can implemented somewhat differently giving rise to a number of solution variants [4, 6], the analysis of false rates presented in this paper can be easily applied to these variants.

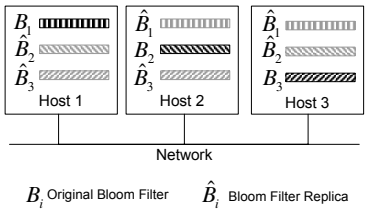

Figure 1. An example of the application of Bloom filters in a distributed system with 3 hosts.

The detailed procedures of the operations of insertion, deletion and query of data objects are shown in Figure 2. When an object x is deleted from or inserted into Host i, the values of the relevant filters $\Gamma_i[\mathcal{H}(x)]$ and bits $B_i[\mathcal{H}(x)]$ are adjusted accordingly. When the fraction of modified bits in B_i exceeds some threshold, B_i is broadcast to all the other hosts to update $\hat{B}_i$. To look up x, Host i performs the membership tests on all the BFs kept locally. If a test on B_i is positive, then x can potentially be accessed locally. If a test in the filter $\hat{B}_j$ for any $j \neq i$ is positive, then x is conjectured to be on Host j with high probability. Finally, if none of the tests is positive, x is considered nonexistent in the system.

We begin the analysis by examining the false negative and false positive rate of a single BF replica in Section 3.1. Then Section 3.2 presents the analysis of the overall false rates of all BFs kept locally on a host. The experimental validations of the analytical models are presented in Section 4.

3.1 False Rates of Bloom Filter Replicas

Let B be a BF with m bits and $\hat{B}$ a replica of B. Let n and $\hat{n}$ be the number of objects in the set represented by B and by $\hat{B}$, respectively. We denote $\triangle_1$ ($\triangle_0$) as the set of all one (zero) bits in B that are different than (i.e., complement of) the corresponding bits in $\hat{B}$. More specifically,

$$\begin{aligned} \triangle_1 &= \{B[i] \mid B[i] = 1, \hat{B}[i] = 0, \forall i \in [1, m]\} \\ \triangle_0 &= \{B[i] \mid B[i] = 0, \hat{B}[i] = 1, \forall i \in [1, m]\}. \end{aligned}$$

Thus, $\triangle_1$ + $\triangle_0$ represent the set of changed bits in B that have not been propagated to $\hat{B}$. The number of bits in this set is affected by the update threshold and update latency. Furthermore, if a nonempty $\triangle_1$ is hit by least one hash function of a membership test on $\hat{B}$ while all other

AddObject(Object x, Host i)
1. Set $(B_i[\mathcal{H}(x)])$ to 1;
2. Increase $\Gamma_i[\mathcal{H}(x)]$ by 1;
3. **if**(the changed portion of B_i is larger than some threshold)
4. Multicast B_i to the other hosts;

DeleteObject(Object x, Host i)
1. Decrease $\Gamma_i[\mathcal{H}(x)]$ by 1;
2. **for**$(j = 1; j \leq k; j++)$
3. **if**$(\Gamma_i[h_j(x)] = 0)$
4. Unset bit $B_i[h_j(x)]$ to 0;

QueryObject(Object x, Host i)
1. $\psi = \emptyset$;
2. /* check the BF of the local host */
3. **if**$(B_i[\mathcal{H}(x)] = 1)$
4. $\psi = \{i\}$;
5. /* check all BF replicas */
6. **for**$(j = 1; j \leq \gamma; j++)$
7. **if**$(j \neq i$ and $\hat{B}_i[\mathcal{H}(x)] = 1)$
8. $\psi = \psi \cup \{j\}$
9. **return** ψ;

Figure 2. Procedures of adding, deleting and querying Object x at Host i

hash functions of the same test hit bits in $\hat{B} - \triangle_1 - \triangle_0$ with a value of one, then a false negative occurs in $\hat{B}$. Similarly, a false positive occurs if the nonempty $\triangle_1$ is replaced by a nonempty $\triangle_0$ in the exact membership test scenario on a $\hat{B}$ described above.

Lemma 1. *Suppose that the numbers of bits in $\triangle_1$ and in $\triangle_0$ are $m\delta_1$ and $m\delta_0$, respectively. Then $\hat{n}$ is a random variable following a normal distribution with an extremely small variance (i.e., extremely highly concentrated around its mean), that is,*

$$\mathbb{E}(\hat{n}) = -\frac{m}{k}\ln(e^{-kn/m} + \delta_1 - \delta_0). \tag{6}$$

Proof: In a given BF representing a set of n objects, each bit is zero with probability $P_0(n)$, given in Equation 2, or one with probability $P_1(n) = 1 - P_0(n)$. Thus the average fractions of zero and one bits are $P_0(n)$ and $P_1(n)$, respectively. Ref. [14] shows formally that the fractions of zero and one bits are random variables that are highly concentrated on $P_0(n)$ and $P_1(n)$ respectively.

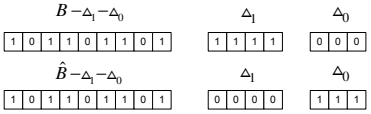

Figure 3. An example of a BF B and its replica $\hat{B}$ where bits are reordered such that bits in $\triangle_1$ and $\triangle_0$ are placed together.

Figure 3 shows an example of B and $\hat{B}$ where bits in $\triangle_1$ and $\triangle_0$ are extracted out and placed together. The expected numbers of zero bits in $B - \triangle_1 - \triangle_0$ and in $\hat{B} - \triangle_1 - \triangle_0$ should be equal since the bits in them are always identical for any given B and $\hat{B}$. Thus for any given n, δ_1 and δ_0, we have

$$P_0(n) - \delta_0 = \mathbb{E}(P_0(\hat{n})) - \delta_1 \tag{7}$$

Substituting Equation 2 into the above equation, we have

$$e^{-kn/m} - \delta_0 = e^{-k\mathbb{E}(\hat{n})/m} - \delta_1 \tag{8}$$

After solving Equation 8, we obtain Equation 6. □

Pragmatically, in any given BF with n objects, the values of δ_1 and δ_0, which represent the probabilities of a bit falling in $\triangle_1$ and $\triangle_0$ respectively, are relatively small. Theoretically, the number of bits in $\triangle_1$ is less than the total number of one bits in B, thus we have $\delta_1 \leq 1 - e^{-kn/m}$. In a similar way, we can conclude that $\delta_0 \leq e^{-kn/m}$.

Theorem 3 (False Negative Rate)**.** *The expected false negative rate $\hat{f}^-$ in the BF replica $\hat{B}$ is $P_1(n)^k - (P_1(n) - \delta_1)^k$, where $P_1(n) = 1 - e^{-kn/m}$.*

Proof: As mentioned earlier, a false negative in $\hat{B}$ occurs when at least one hash function hits the bits in $\triangle_1$ in $\hat{B}$ while the others hit the bits in $\hat{B} - \triangle_1 - \triangle_0$ with a value of one. Hence, the false negative rate is

$$\begin{aligned} \hat{f}^- &= \sum_{i=1}^{k} \binom{k}{i} \delta_1^i \left(P_1(\hat{n}) - \delta_0\right)^{k-i} \\ &= \left(P_1(\hat{n}) - \delta_0 + \delta_1\right)^k - \left(P_1(\hat{n}) - \delta_0\right)^k \end{aligned}$$

Since $P_0(n) = 1 - P_1(n)$ and $P_0(\hat{n}) = 1 - P_1(\hat{n})$, Equation 7 can be rewritten as,

$$\mathbb{E}(P_1(\hat{n})) = P_1(n) + \delta_0 - \delta_1 \tag{9}$$

Hence

$$\begin{aligned} \mathbb{E}(\hat{f}^-) &= \left(\mathbb{E}(P_1(\hat{n})) - \delta_0 + \delta_1\right)^k - \left(\mathbb{E}(P_1(\hat{n})) - \delta_0\right)^k \\ &= P_1(n)^k - \left(P_1(n) - \delta_1\right)^k \end{aligned} \tag{10}$$

□

Theorem 4 (False Positive Rate)**.** *The expected false positive rate $\hat{f}^+$ for the Bloom filter replica $\hat{B}$ is $\left(P_1(n) + \delta_0 - \delta_1\right)^k$, where $P_1(n) = 1 - e^{-kn/m}$.*

Proof: If $\hat{B}$ confirms positively the membership of an object while this object actually does not belong to B, then a false positive occurs. More specifically, a false positive occurs in $\hat{B}$ if for any $x \notin B$, all hit bits by hash functions of the membership test for x are ones in $\hat{B} - \triangle_1 - \triangle_0$, or for

any $x \in \mathcal{U}$, all hit bits are ones in $\hat{B}$ but at least one hit bit is in $\triangle_0$. Thus, we find that

$$\begin{aligned}\hat{f}^+ &= \left(1 - \frac{n}{|\mathcal{U}|}\right)(P_1(\hat{n}) - \delta_0)^k + \\ &\quad \sum_{i=1}^{k} \binom{k}{i} \delta_0^i (P_1(\hat{n}) - \delta_0)^{k-i} \\ &= P_1(\hat{n})^k - \frac{n}{|\mathcal{U}|}(P_1(\hat{n}) - \delta_0)^k \qquad (11)\end{aligned}$$

Considering $n \ll |\mathcal{U}|$ and Equation 9, we have

$$\begin{aligned}\mathbb{E}(\hat{f}^+) &= (\mathbb{E}(P_1(\hat{n})))^k - \frac{n}{|\mathcal{U}|}(\mathbb{E}(P_1(\hat{n})) - \delta_0)^k \\ &= (P_1(n) + \delta_0 - \delta_1)^k - \frac{n}{|\mathcal{U}|}(P_1(n) - \delta_1)^k \\ &\approx (P_1(n) + \delta_0 - \delta_1)^k \qquad (12)\end{aligned}$$

□

3.2 Overall False Rates

In there distributed system considered in this study, there are a total of γ hosts and each host has γ BFs, with $\gamma - 1$ of them replicated from the other hosts. To look up an object, a host performs the membership tests in all the BFs kept locally. This section analyzes the overall false rates on each BF replica and each host.

Give any BF replica $\hat{B}$, the events of a false positive and a false negative are exclusive. Thus it is easy to find that the overall false rate of $\hat{B}$ is

$$\mathbb{E}(f_{overall}) = \mathbb{E}(f^-) + \mathbb{E}(f^+) \qquad (13)$$

where $\mathbb{E}(f^-)$ and $\mathbb{E}(f^+)$ are given in Equation 10 and 12 respectively.

On Host i, BF B_i represents all the objects stored locally. While only false positives occur in B_i, both false positives and false negatives can occur in the replicas $\hat{B}_j$ for any $j \neq i$. Since the failed membership test in any BF leads to a lookup failure, the overall false positive and false negative rates on Host i are therefore

$$\mathbb{E}(f^+_{host}) = 1 - (1 - f_i^+) \prod_{j=1, j\neq i}^{\gamma} (1 - \hat{f}_j^+) \qquad (14)$$

and

$$\mathbb{E}(f^-_{host}) = 1 - \prod_{j=1, j\neq i}^{\gamma} (1 - \hat{f}_j^-) \qquad (15)$$

where f_i^+, $\hat{f}_j^-$ and $\hat{f}_j^+$ are given in Theorem 2, 3 and 4 respectively.

The probability that Host i fails a membership lookup can be expressed as follows,

$$\mathbb{E}(f_{host}) = \mathbb{E}(f^+_{host} + f^-_{host} - f^+_{host} f^-_{host}). \qquad (16)$$

In practice, we can use the overall false rate of a BF replica to trigger updating process and use the overall false rate of all BFs on a host to evaluate the whole systems. In a typical distributed environment with many nodes, the updating of a Bloom filter replica $\hat{B}_i$ stored on node j can be triggered by either the home node i or the node j. Since many nodes hold the replica of B_i, it is more efficient to let the home node i to initiate the updating process of all replicas of B_i. Otherwise, the procedure of checking whether an updating is needed would be performed by all other nodes, wasting both network and CPU resources. Accordingly, we can only use the overall false rate of a BF replica $\mathbb{E}(f_{overall})$ as the updating criteria. On the other hand, $\mathbb{E}(f_{host})$ can be used to evaluate the overall efficiency of all BFs stored on the same host.

4 Validation of the Theoretic Models via Experiments

This section validates our theoretical framework developed in this paper by comparing the analytical results produced by our models with experimental results obtained through real experiments.

We begin by examining a single BF replica. Initially the Bloom filter replica $\hat{B}$ is exactly the same as B. Then we artificially change B by randomly inserting new objects into B or randomly deleting existing objects from B repeatedly. For each specific modification made to B, we calculate the corresponding δ_1 and δ_0 and use 100,000 randomly generated objects to test the memberships against $\hat{B}$. Since the actual objects represented in B are known in the experiments, the false negative and positive rates can be easily measured.

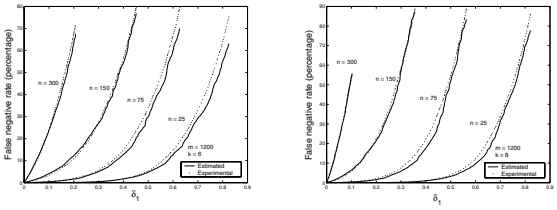

Figure 4. Comparisons of estimated and experimental $\hat{f}^-$ of $\hat{B}$ when k is 6 and 8 respectively. The initial object number in both B and $\hat{B}$ is 25, 75, 150 and 300 ($m = 1200$).

Figure 4 compares analytical and real false negative rates, obtained from the theoretic models and from experiments respectively, by plotting the false negative rate in $\hat{B}$ as a function of δ_1, a measure of update threshold, for different numbers of hashing functions ($k = 6$ and $k = 8$) when

Table 1. False positive rates comparisons when k is 6 and 8 respectively ($m = 1200$).

				f^+ (percentage)	
k	$\hat{n}$	δ_0	δ_1	Estimated	Experimental
6	25	0.0942	0.2042	0.0002	0
6	25	0.0800	0.3650	0.0002	0
6	25	0.0600	0.4875	0.0001	0
6	75	0.0800	0.1608	0.0934	0.1090
6	75	0.0600	0.2833	0.0794	0.1090
6	75	0.0483	0.3758	0.0799	0.1090
6	150	0.0533	0.1042	2.2749	2.6510
6	150	0.0400	0.1800	2.3540	2.6510
6	150	0.0325	0.2508	2.1872	2.6530
6	300	0.0250	0.0417	23.6555	25.4790
6	300	0.0183	0.0692	25.4016	25.4710
6	300	0.0117	0.1000	24.7241	25.4750
8	25	0.1083	0.2425	0.00002	0
8	25	0.0792	0.4192	0.00002	0
8	25	0.0550	0.5425	0.00002	0
8	75	0.0792	0.1767	0.0525	0.0540
8	75	0.0550	0.3000	0.0504	0.0540
8	75	0.0425	0.3917	0.0506	0.0540
8	150	0.0475	0.1050	2.5163	2.5770
8	150	0.0350	0.1758	2.6783	2.5780
8	150	0.0283	0.2367	2.5384	2.5790
8	300	0.0192	0.0333	33.2078	33.2580
8	300	0.0133	0.0558	34.4915	33.2550
8	300	0.0083	0.0817	32.1779	33.2550

Table 2. Overall false rate comparisons under optimum initial operation state when k is 6 and 8 respectively. 100 new objects are added on each host and then a set of existing objects are deleted from each host. The number of deleted objects increases from 10 to 100 with a step size of 10. ($m = 1200$) In the first group, initially Initially $n = 150$ and $m/n = 8$; in the second group, $n = 100$ and $m/n = 12$ initially.

			$f_{overall}$ (percentage)	
k	δ_0	δ_1	Estimated	Experimental
6	0.0100	0.1705	46.2259	45.2200
6	0.0227	0.1657	42.4850	40.6880
6	0.0347	0.1627	38.7101	37.2420
6	0.0458	0.1582	34.9268	33.8460
6	0.0593	0.1545	31.3748	30.4540
6	0.0715	0.1497	27.8831	27.3700
6	0.0837	0.1445	24.5657	24.8000
6	0.0938	0.1392	21.2719	22.5560
6	0.1045	0.1340	18.2490	20.4520
6	0.1165	0.1300	15.5103	18.7540
8	0.0123	0.2375	30.9531	29.6280
8	0.0255	0.2275	25.7946	23.6280
8	0.0413	0.2180	21.0943	18.0000
8	0.0552	0.2123	16.7982	14.6720
8	0.0658	0.2043	12.9800	12.0040
8	0.0772	0.1965	9.7307	9.7320
8	0.0920	0.1900	7.1016	7.7520
8	0.1075	0.1848	4.9936	6.1280
8	0.1237	0.1788	3.4031	4.8400
8	0.1377	0.1732	2.2034	3.8160

the initial number of objects in B are 25, 75, 150 and 300 respectively. Since the false negative rates are independent of δ_0, only object deletions are performed in B.

Table 1 compares the analytical and real false positive rates of $\hat{B}$ when k is 6 and 8 respectively. In these experiments, both object deletions and additions are performed in B while $\hat{B}$ remains unaltered. It is interesting that the false positive rates of $\hat{B}$ is kept around some constant for a specific $\hat{n}$ although the objects in B changes in the real experiments. It is true that if the number of objects in B increases or decreases, the false positive rate in $\hat{B}$ should decrease or increase accordingly before the changes of B is propagated to $\hat{B}$. However, due to the fact that n is far less than the total object number in the universe $\mathcal{U}$, the change of the false positive rate in $\hat{B}$ is too small to be perceptible. These tests are made accordant with the real scenarios of BF applications in distributed systems. In such real applications, the number of possible objects is usually very large and thus BFs are deployed to efficiently reduce the network and network communication requirements. Hence, in these experiments the number of objects used to test $\hat{B}$ is much larger than the number of objects in B or $\hat{B}$ (100,000 random objects are tested). Under such large size of testing samples, the influence of the modification in B on the false positive rate of $\hat{B}$ is difficult to be observed.

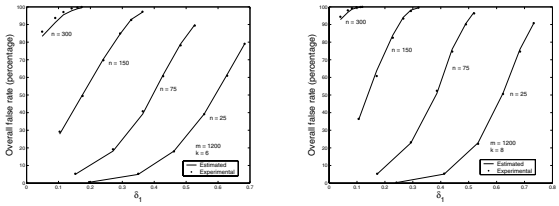

Figure 5. Comparisons of estimated and experimental $f_{overall}$ in a distributed system with 5 hosts when k is 6 and 8 respectively. The initial object number n on each host is 25, 75, 150 and 300 respectively. Then each host adds a set of new objects. The number of new objects on each host increases from 50 to 300 with a step size of 50. ($m = 1200$)

We also simulated the lookup problem in a distributed system with 5 hosts. Figure 5 shows the comparisons of the analytical and experimental average overall false rates on each host. In these experiments, we only added new objects without deleting any existing items so that δ_0 is kept zero. The experiments presented in Table 2 considers both the deletion and addition of objects on each host when the initial state of BF on each host is optimized, this is, the number of hash functions is the optimal under the ratio between m and the initial number of objects n. This specific setting aims to emulate the real application where m/n and k are usually optimally or sub-optimally matched by dynamically adjusting the BF length m [3] or designing the BF length according to the average number of objects [12, 6, 2, 4, 5]. All the analytical results have been very closely matched by their real (experimental) counterparts consistently, strongly validating our theoretical models.

5 Updating Protocol

To reduce the false rate caused by staleness, the remote Bloom filter replica needs to be periodically updated. The update process are typically triggered if the percentage of dirty bits in a local BF exceeds some threshold. While a small threshold causes large network traffic and a large threshold increases the false rate, this tradeoff is usually reached by a trial-and-error approach that runs numerous (typically a large number of) trials in real experiments or simulations [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12].

For example, in the summery cache study [2], it is recommended that if 10 percent of bits in a BF are dirty, then the BF propagates its changes to all replicas. However, this approach has the following disadvantages.

1. It cannot directly control the false rate. To keep the false rate under some target value, complicated simulations or experiments have to be conducted to adjust the threshold for dirty bits. If the target false rate changes, this tedious process has to be repeated to find a "golden" threshold.
2. It treats all dirty bits equally and does not distinguish the zero-dirty bits from the one-dirty bits. In fact, as shown in previous sections, the dirty one bits and the dirty zero bits exert different impacts on the false rates.
3. It does not allow flexible update control. In many applications, the penalty of a false positive and a false negative are significantly different. For example, in summery cache [2], a false positive occurs if a request is not a cache hit on some web proxy when the corresponding Bloom filter replica confirms so. The penalty of a false positive is a waste of query message to this local web proxy. A false negative happens if a request can be hit in a local web proxy but the Bloom filter replica mistakenly indicates otherwise. The penalty of a false negative is a round-trip delay in retrieving information from a remote web server through the Internet. Thus, the penalty of a false negative is much larger than that of a false positive. The updating protocols based on the percentage of dirty bits do not allow one to place more weight on the false negative rate, thus limiting the flexibility and efficiency of the updating process.

Based on the theoretic models presented in the previous sections, an updating protocol that directly control the false rate is designed in this paper. In a distributed system with γ nodes where each node has a local BF to represent all local elements, each node is responsible for automatically updating its BF replicas. Each node estimates the false rate of its remote BF replica and if the false rate exceeds some desire false rate, as opposed to a predetermined threshold on the percentage of dirty bits in the conventional updating approaches, a updating process is triggered. To estimate the false rate of remote BF replica $\hat{B}$, each node has to record the number of elements stored locally (n), in addition to a copy of remote BF replica $\hat{B}$. This copy is essentially the local BF B when the last updating is made. It is used to calculate the percentage of dirty one bits (δ_1) and the dirty zero bits (δ_0). Compared with the conventional updating protocols based on the total percentage of dirty bits, this protocol only needs to record one more variable (n), thus it does not significantly increase the maintenance overhead.

This protocol allows more flexible updating protocols that considers the penalty difference between a false positive and a false negative. The overall false rate can be a weighted sum of the false positive rate and the false negative rate, shown as follows:

$$\mathbb{E}(f_{overall}) = w^+\mathbb{E}(f^+) + w^-\mathbb{E}(f^-) \qquad (17)$$

where w^+ and w^- are the weights. The values of w^+ and w^- depends on the applications and also the application environments.

We prove the effectiveness of this update protocol through event driven simulations. In this simulation, we made the following assumptions.

1. Each item is randomly accessed. This assumption may not be realistic in some real workloads, in which an item has a greater than equal chance of being accessed again once it has been accessed. Though all previous theoretic studies on Bloom filters assume a workload with uniform access spectrum, further studies are needed to investigate the impact of this assumption.
2. Each local node deletes or adds items at a constant rate. In fact, the deletion and addition rate changes dynamically throughout the lifetime of applications. This simplifying assumption is employed just to prove our concept while keeping our experiments manageable in the absence of a real trace or benchmark.
3. The values of w^+ and w^- are 1. Their optimal values depends on the nature of the applications and environments.

We simulate a distributed system with two nodes where each node keeps a BF replica of the other. We assume the addition and deletion are 5 and 2 per time unit respectively and our desired false rate is 10%. Figure 6 shows the estimated false rate and the measured false rate of node 1 throughout the deletion, addition and updating processes. Due to the space limitation, the false rate on node 2, which is similar to node 1, is not shown in this paper. In addition, we have changed the addition rate and deletion rates. Simulation results consistently indicate that our protocol is accurate and effective in control the false rate.

6 Conclusions

In this paper, we have presented the theoretical analysis of the impact of staleness existing in many distributed BF

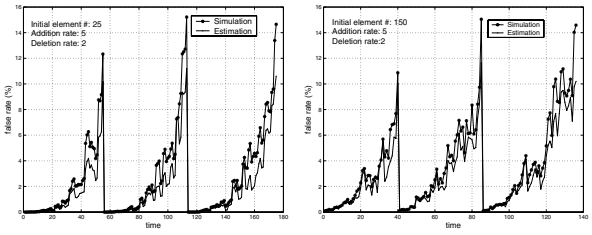

Figure 6. In an environment of two servers, the figures show the overall false rate on one server when the initial number of elements in one server are 25 and 150 respectively. The ratio of bits per element is 8 and 6 hash functions are used. The rate for element addition and deletion are respectively 5 and 2 per time unit on each server.

applications on the false negative and false positive rates, and developed an adaptive update control mechanism that accurately and efficiently maintains a desirable level of false rate for a given application. To the best of our knowledge, we are the first to derive accurate closed-form expressions that incorporate the staleness into the analysis of the false negative and positive rates of a single BF replica, the first to develop the analytical models of the overall false rates of BF arrays that have been widely used in many distributed systems, and the first to develop an adaptively controlled update process that accurately maintains a desirable level of false rate for a given application. We have validated our analysis by conducting extensive experiments. The theoretical analysis presented not only provides system designers with significant theoretical insights into the development and deployment of BFs in distributed systems, but also are useful in practice for accurately determining when to trigger the processes of updating BF replicas in order to keep the false rates under some desired values, or, equivalently, minimize the frequency of updates to reduce update overhead.

7 Acknowledgments

This work was supported by an UMaine Startup Fund and Chinese Government 973 Project (No.2004cb318201).

References

[1] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422–426, 1970.

[2] L. Fan, E. Cao, J. Almeida, and A.Z. Broder, "Summary cache: A scalable wide-area web cache sharing protocol," *IEEE Trans. Commun.*, vol. 8, no. 3, pp. 281–293, 2000.

[3] H. Tang and T. Yang, "An efficient data location protocol for self-organizing storage clusters," in *Proceedings of ACM/IEEE SuperComputing*, Nov. 2003.

[4] J. Ledlie, L. Serban, and D. Toncheva, "Scaling filename queries in a large-scale distributed file systems," Tech. Rep., Harvard University.

[5] M. Ripeanu and I. Foster, "A decentralized, adaptive, replica location service," in *11th IEEE International Symposium on High Performance Distributed Computing*, Edinburgh, Scotland, July 2002.

[6] Y. Zhu, H. Jiang, and J. Wang, "Hierarchical bloom filter arrays (HBA): A novel, scalable metadata management system for large cluster-based storage," in *Proceedings of 2004 IEEE International Conference on Cluster Computing*, California, Sept. 2004.

[7] P.H. Hsiao, "Geographical region summary service for geographical routing," *Mobile Computing and Communications Review*, vol. 5, no. 4, 2001.

[8] H. Cai and J. Wang, "Foreseer: A novel, locality-aware peer-to-peer system architecture for keyword searches," in *ACM/IFIP/USENIX 5th International Middleware Conference*, Toronto, Canada, 2004.

[9] J. Kubiatowicz et al., "Oceanstore: an architecture for global-scale persistent storage," in *Proceedings of the* 9^{th} *international conference on Architectural Support for Programming Languages and Operating Systems*, Cambridge, Massachusetts, 2000, pp. 190–201.

[10] A. Mohan and V. Kalogeraki, "Speculative routing and update propagation: A kundali centric approach," in *Proceedings of IEEE 2003 International Conference on Communications*, Anchorage, AK, May 2003.

[11] S. Rhea and J. Kubiatowicz, "Probabilistic location and routing," in *Proceedings of The 21st Annual Joint Conference of the IEEE Computer and Communications Societies*, New York, 2002, pp. 1248–1257.

[12] M. C. Little, S. K. Shrivastava1, and N. A. Speirs, "Using bloom filters to speed-up name lookup in distributed systems," *The Computer Journal*, vol. 45, no. 6, pp. 645–652, 2002.

[13] A. Broder and M. Mitzenmacher, "Network applications of bloom filters: A survey," in *Proceedings of 40th Annual Allerton Conference on Communication, Control and Computing*, Illinois, Oct. 2002.

[14] M. Mitzenmacher, "Compressed bloom filters," *IEEE Trans. Netw.*, vol. 10, no. 5, pp. 604–612, 2002.

[15] J. K. Mullin, "A second look at bloom filters," *Commun. ACM*, vol. 26, no. 8, pp. 570–571, 1983.

[16] M. D. McIlroy, "Development of a spelling list," *IEEE Trans. Commun.*, vol. 30, pp. 91–99, 1982.

[17] L.L. Gremillion, "Designing a bloom filter for differential file access," *Commun. ACM*, vol. 25, no. 9, pp. 600–604, 1982.