

Chapter 2

Per-Flow Size Estimators

Abstract This chapter discusses the measurement of per-flow sizes for high-speed links. It is a particularly difficult problem because of the need to process and store a huge amount of information, which makes it difficult for the measurement module to fit in the small but fast SRAM space (in order to operate at the line speed). We provide a novel measurement function that estimates the sizes of all flows. It delivers good performance in tight memory space where other approaches no longer work. The effectiveness of the online per-flow measurement approach is analyzed and confirmed through extensive experiments based on real network traffic traces. The rest of this chapter is organized as follows: Sect. 2.1 discusses the performance metrics. Section 2.2 gives an overview of the system design. Section 2.3 discusses the state of the art. Section 2.4 presents the online data encoding module. Sections 2.5–2.6 present two offline data decoding modules. Section 2.7 discusses the problem of setting counter length. Section 2.8 addresses the problem of collecting flow labels. Section 2.9 presents the experimental results. Section 2.10 extends the estimators for large flow sizes. Section 2.11 gives the summary.

Keywords Per-flow size estimator, Randomized counter sharing

2.1 Performance Metrics

We measure the number of packets in each flow during a measurement period, which ends every time after a certain number (e.g., 10 millions) of packets are processed. The design of per-flow measurement functions should consider the following three key performance metrics.

2.1.1 Processing Time

The per-packet processing time of an online measurement function determines the maximum packet throughput that the function can operate at. It should be made as small as possible in order to keep up with the line speed. This is especially true when multiple routing, security, measurement, and resource management functions share SRAM and processing circuits.

The processing time is mainly determined by the number of memory accesses and the number of hash computations (which can be efficiently implemented in hardware [24]). The counter braids [21, 22] update k counters at the first level for each packet. When a counter at the first level overflows, it needs to update k additional counters at the second level. If $k = 3$, it requires at least 3 hashes and 6 memory accesses to read and then write back after counter increment. In the worse case, it requires 6 hashes and 12 memory accesses. The multi-resolution space-code Bloom filters [19] probabilistically select one or more of its 9 filters and set 3~6 bits in each of the selected ones. Each of those bits requires one memory access and one hash computation.

Our objective is to achieve a constant per-packet processing time of one hash computation and two memory accesses (for updating a single counter). This is the minimum processing time for any method that uses hash operations to identify counters for update.

2.1.2 Storage Overhead

The need to reduce the SRAM overhead has been discussed in Chap. 1. One may argue that because the amount of memory needed is related to the number of packets in a measurement period, we can reduce the memory requirement by shortening the measurement period. However, when the measurement period is smaller, more flows will span multiple periods and consequently the average flow size in each period will be smaller. When we measure the flow sizes, we also need to capture the flow labels [22], e.g., a tuple of source address/port and destination address/port to identify a TCP flow. The flow labels are too large to fit in SRAM. They have to be stored in DRAM. Therefore, in a measurement period, each flow incurs at least one DRAM access to store its flow label. If the average flow size is large enough, the overhead of this DRAM access will be amortized over many packets of a flow. However, if the average flow size is too small, the DRAM access will become the performance bottleneck that seriously limits the throughput of the measurement function. This means the measurement period should not be too small.

2.1.3 Estimation Accuracy

Let s be the size of a flow and $\hat{s}$ be the estimated size of the flow based on a measurement function. The estimation accuracy of the function can be specified by a confidence interval: the probability for s to be within $[\hat{s} \cdot (1 - \beta), \hat{s} \cdot (1 + \beta)]$ is at least a pre-specified value α , e.g., 95%. A smaller value of β means that the estimated flow size is more accurate in a probabilistic sense.

There is a tradeoff between estimation accuracy and storage overhead. If the storage space and the processing time are unrestricted, we can accurately count each packet to achieve perfect accuracy. However, in practice, there will be constraints on both storage and processing speed, which make 100% accurate measurement sometimes infeasible. In this case, one has to settle with imperfect results that can be produced with the available resources. Within the bounds of the limited resources, we must explore novel measurement methods to make the estimated flow sizes as accurate as possible.

2.2 System Design

2.2.1 Basic Idea

We use an example to illustrate the idea behind the new measurement approach. Suppose the amount of SRAM allocated to one of the measurement functions is 2 Mb (2×2^{20} bits), and each measurement period ends after 10 million packets, which translate into about 8 s for an OC-192 link (10+ Gbps) with an average packet size of 1,000 bytes. The types of flows that the online functions may measure include per-source flows, per-destination flows, per-source/destination flows, TCP flows, WWW flows (with destination port 80), etc. Without losing generality, suppose the specific function under consideration in this example measures the sizes of TCP flows.

Figure 2.1 shows the number of TCP flows that have a certain flow size in log scale, based on a real network trace captured by the main gateway of our campus. If we use 10 bits for each counter, there will be only 0.2 million counters. The number of concurrent flows in the trace for a typical measurement period is around 1 million. It is obvious that allocating per-flow state is not possible and each counter has to store the information of multiple flows. But if an “elephant” flow is mapped to a counter, that counter will overflow and lose information. On the other hand, if only a couple of “mouse” flows are mapped to a counter, the counter will be under-utilized, with most of its high-order bits left as zeros.

To solve the above problems, we not only store the information of multiple flows in each counter, but also store the information of each flow in a large number of counters, such that an “elephant” is broken into many “mice” that are stored at different counters. More specifically, we map each flow to a set of l randomly-selected counters and split the flow size into l roughly-equal shares, each of which

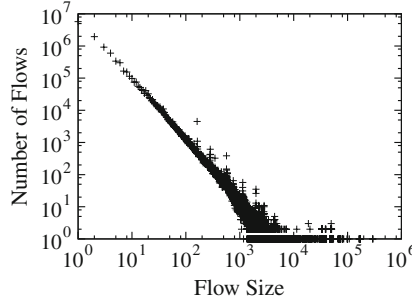


Fig. 2.1 Traffic distribution: each point shows the number (*y coordinate*) of flows that have a certain size (*x coordinate*)

is stored in one counter. The value of a counter is the sum of the shares from all flows that are mapped to the counter. Because flows share counters, they introduce noise to each other’s measurement. The key to accurately estimate the size of a flow is to measure the noise introduced by other flows in the counters that the flow is mapped to.

Fortunately, this can be done if the flows are mapped to the counters uniformly at random. Any two flows will have the same probability of sharing counters, which means that each flow will have the same probability of introducing a certain amount of noise to any other flow. If the number of flows and the number of counters are very large, the combined noise introduced by all flows will be distributed across the counter space about uniformly. The statistically uniform distribution of the noise can be measured and removed. The above scheme of information storage and recovery is called *randomized counter sharing*.

We stress that this design philosophy of “splitting” each flow among a large number of counters is very different from “replicating” each flow in k counters as the counting Bloom filter [7] or counter braids [21] do—they add the size of each flow as a whole to k randomly selected counters. Most notably, the method increments one counter for each arrival packet, while the counting Bloom filter or counter braids increment k counters. We store the information of each flow in many counters (e.g., 50), while they store the information of each flow in a small number of counters.

2.2.2 Overall Design

The online traffic measurement function consists of two modules. The online data encoding module stores the information of arrival packets in an array of counters. For each packet, it performs one hash function to identify a counter and then updates the counter with two memory accesses, one for reading and the other for writing. At the end of each measurement period, the counter array is stored to the disk and then reset to zeros.

The offline data decoding module answers queries for flow sizes. It is performed by a designated offline computer. We present two methods for separating the information about the size of a flow from the noise in the counters. The first one is called the *counter sum estimation method* (CSM), which is very simple and easy to compute. The second one is called the *maximum likelihood estimation method* (MLM), which is more accurate but also more computationally intensive. The two complementary methods provide flexibility in designing a practical system, which may first use CSM for rough estimations and then apply MLM to the ones of interest.

2.3 State of the Art

A related thread of research is to collect statistical information of the flows [11, 18], or identify the largest flows and devote the available memory to measure their sizes while ignoring the smaller ones [9, 12, 15, 16]. For example, RATE [17] and ACCEL-RATE [14] measure per-flow rate by maintaining per-flow state, but they use a *two-run sampling* method to filter out small-rate flows so that only high-rate flows are measured.

Another thread of research is to maintain a large number of counters to track various networking information. One possible solution [8, 25] can be statistically update a counter according to the current counter size. This approach is suitable for the applications with loose measurement accuracy. In order to enhance the accuracy performance, Zhao et al. [29] propose a statistical method to make a DRAM-based solution practical. Since large DRAM is involved, this approach is able to achieve decent measurement accuracy.

Also related is the work [26] that measures the number of *distinct* destinations that each source has contacted. Per-flow counters cannot be used to solve this problem because they cannot remove duplicate packets. If a source sends 1,000 packets to a destination, the packets contribute only one contact, but will count as 1,000 when we measure the flow size. To remove duplicates, bitmaps (instead of counters) should be used [5, 13, 27, 28, 30]. From the technical point of view, this represents a separate line of research, which employs a different set of data structures and analytical tools. Attempt has also been made to use bitmaps for estimating the flow sizes, which is however far less efficient than counters, as our experiments will show.

2.4 Online Data Encoding

The flow size information is stored in an array C of m counters. The i th counter in the array is denoted as $C[i]$, $0 \leq i \leq m - 1$. The size of the counters should be set so that the chance of overflow is negligible; we will discuss this issue in details later. Each flow is mapped to l counters that are randomly selected from C through hash functions. These counters logically form a *storage vector* of the flow, denoted

as C_f , where f is the label of the flow. The i th counter of the vector, denoted as $C_f[i]$, $0 \leq i \leq l - 1$, is selected from C as follows:

$$C_f[i] = C[H_i(f)], \quad (2.1)$$

where $H_i(\dots)$ is a hash function whose range is $[0, m)$. We want to stress that C_f is *not* a separate array for flow f . It is merely a logical construction from counters in C for the purpose of simplifying the presentation. In all the formulas, one should treat the notation $C_f[i]$ simply as $C[H_i(f)]$. The hash function H_i , $0 \leq i \leq l - 1$, can be implemented from a master function $H(\dots)$ as follows: $H_i(f) = H(f|i)$ or $H_i(f) = H(f \oplus R[i])$, where $|$ is the concatenation operator, $\oplus$ is the XOR operator, and $R[i]$ is a constant whose bits differ randomly for different indices i .

All counters are initialized to zeros at the beginning of each measurement period. The operation of online data encoding is very simple: When the router receives a packet, it extracts the flow label f from the packet header, randomly selects a counter from C_f , and increases the counter by one. More specifically, the router randomly picks a number i between 0 and $l - 1$, computes the hash $H_i(f)$, and increases the counter $C[H_i(f)]$, which is physically in the array C , but logically the i th element in the vector C_f .

2.5 Offline Counter Sum Estimation

2.5.1 Estimation Method

At the end of a measurement period, the router stores the counter array C to a disk for long-term storage and offline data analysis.

Let n be the combined size of all flows, which is $\sum_{i=0}^{m-1} C[i]$.

Let s be the true size of a flow f during the measurement period. The estimated size, $\hat{s}$, based on the counter sum estimation method (CSM) is

$$\hat{s} = \sum_{i=0}^{l-1} C_f[i] - l \frac{n}{m}. \quad (2.2)$$

The first item is the sum of the counters in the storage vector of flow f . It can also be interpreted as the sum of the flow size s and the noise from other flows due to counter sharing. The second item captures the expected noise. Below we formally derive (2.2).

Consider an arbitrary counter in the storage vector of flow f . We treat the value of the counter as a random variable X . Let Y be the portion of X contributed by the packets of flow f , and Z be the portion of X contributed by the packets of other flows. Obviously, $X = Y + Z$.

Each of the s packets in flow f has a probability of $\frac{1}{l}$ to increase the value of the counter by one. Hence, Y follows a binomial distribution:

$$Y \sim \text{Bino}(s, \frac{1}{l}) \quad (2.3)$$

Each packet of another flow f' has a probability of $\frac{1}{m}$ to increase the counter by one. That is because the probability for the counter to belong to the storage vector of flow f' is $\frac{1}{m}$, and if that happens, the counter has a probability of $\frac{1}{l}$ to be selected for increment. Assume there is a large number of flows, the size of each flow is negligible when comparing with the total size of all flows, and l is large such that each flow's size is randomly spread among many counters. We can approximately treat the packets independently. Hence, Z approximately follows a binomial distribution:

$$Z \sim \text{Bino}(n - s, \frac{1}{m}) \approx \text{Bino}(n, \frac{1}{m}), \text{ because } s \ll n. \quad (2.4)$$

We must have

$$E(X) = E(Y + Z) = E(Y) + E(Z) = \frac{s}{l} + \frac{n}{m}. \quad (2.5)$$

That is,

$$s = l \times E(X) - l \frac{n}{m}. \quad (2.6)$$

From the observed counter values $C_f[i]$, $E(X)$ can be measured as $\frac{\sum_{i=0}^{l-1} C_f[i]}{l}$. We have the following estimation for s :

$$\hat{s} = \sum_{i=0}^{l-1} C_f[i] - l \frac{n}{m}. \quad (2.7)$$

If a flow shares a counter with an “elephant” flow, its size estimation can be skewed. However, the experiments show that CSM works well in general because the number of “elephants” is typically small (as shown in Fig. 2.1) and thus their impact is also small, particularly when there are a very large number of counters and flows. Moreover, the next method based on maximum likelihood estimation can effectively reduce the impact of an outlier in a flow's storage vector that is caused by an “elephant” flow.

2.5.2 Estimation Accuracy

We derive the mean and variance of $\hat{s}$ as follows: Because $X = Y + Z$, we have

$$\begin{aligned}
E(X^2) &= E((Y + Z)^2) = E(Y^2) + 2E(YZ) + E(Z^2) \\
&= E(Y^2) + 2E(Y)E(Z) + E(Z^2) \\
&= \frac{s^2}{l^2} - \frac{s}{l^2} + \frac{s}{l} + 2 \cdot \frac{s}{l} \cdot \frac{n}{m} + \frac{n^2}{m^2} - \frac{n}{m^2} + \frac{n}{m}.
\end{aligned}$$

The following facts are used in the above mathematical process: $E(Y^2) = \frac{s^2}{l^2} - \frac{s}{l^2} + \frac{s}{l}$ because $Y \sim \text{Bino}(s, 1/l)$. $E(YZ) = E(Y)E(Z)$ since Y and Z are independent. $E(Z^2) = \frac{n^2}{m^2} - \frac{n}{m^2} + \frac{n}{m}$ because $Z \sim \text{Bino}(n, 1/m)$.

$$\begin{aligned}
\text{Var}(X) &= E(X^2) - (E(X))^2 \\
&= \frac{s}{l}(1 - \frac{1}{l}) + \frac{n}{m}(1 - \frac{1}{m}).
\end{aligned} \tag{2.8}$$

In (2.7), $C_f[i]$, $0 \leq i \leq l - 1$, are independent samples of X . We can interpret $\hat{s}$ as a random variable in the sense that a different set of samples of X may result in a different value of $\hat{s}$. From (2.7), we have

$$\begin{aligned}
E(\hat{s}) &= l \times E(X) - l \frac{n}{m} \\
&= l(\frac{s}{l} + \frac{n}{m}) - l \frac{n}{m} = s,
\end{aligned} \tag{2.9}$$

which means our estimation is unbiased. The variance of $\hat{s}$ can be written as

$$\begin{aligned}
\text{Var}(\hat{s}) &= l^2 \times \text{Var}(X) = l^2 \left(\frac{s}{l}(1 - \frac{1}{l}) + \frac{n}{m}(1 - \frac{1}{m}) \right) \\
&= s(l - 1) + l^2 \frac{n}{m}(1 - \frac{1}{m}).
\end{aligned} \tag{2.10}$$

2.5.3 Confidence Interval

We derive the confidence interval for the estimation as follows: The binomial distribution, $Z \sim \text{Bino}(n, 1/m)$, can be closely approximated as a Gaussian distribution, $\text{Norm}(\frac{n}{m}, \frac{n}{m}(1 - \frac{1}{m}))$, when n is large. Similarly, the binomial distribution, $Y \sim \text{Bino}(s, \frac{1}{l})$, can be approximated by $\text{Norm}(\frac{s}{l}, \frac{s}{l}(1 - \frac{1}{l}))$. Because the linear combination of two independent Gaussian random variables is also normally distributed [6], we have $X \sim \text{Norm}(\frac{s}{l} + \frac{n}{m}, \frac{s}{l}(1 - \frac{1}{l}) + \frac{n}{m}(1 - \frac{1}{m}))$. To simplify the presentation, let $\mu = \frac{s}{l} + \frac{n}{m}$ and $\Delta = \frac{s}{l}(1 - \frac{1}{l}) + \frac{n}{m}(1 - \frac{1}{m})$.

$$X \sim \text{Norm}(\mu, \Delta), \tag{2.11}$$

where the mean μ and the variance Δ agree with (2.5) and (2.8), respectively. Because $\hat{s}$ is a linear function of $C_f[i]$, $0 \leq i \leq l-1$, which are independent samples of X , $\hat{s}$ must also approximately follow a Gaussian distribution. From (2.7) and (2.11), we have

$$\hat{s} \sim \text{Norm}(s, s(l-1) + l^2 \frac{n}{m} (1 - \frac{1}{m})). \quad (2.12)$$

Hence, the confidence interval is

$$\hat{s} \pm Z_\alpha \sqrt{s(l-1) + l^2 \frac{n}{m} (1 - \frac{1}{m})}, \quad (2.13)$$

where α is the confidence level and Z_α is the α percentile for the standard Gaussian distribution. As an example, when $\alpha = 95\%$, $Z_\alpha = 1.96$.

2.6 Maximum Likelihood Estimation

In this section, we provide the second estimation method that is more accurate but also more computationally expensive.

2.6.1 Estimation Method

We know from the previous section that any counter in the storage vector of flow f can be represented by a random variable X , which is the sum of Y and Z , where $Y \sim \text{Bino}(s, \frac{1}{l})$ and $Z \sim \text{Bino}(n, 1/m)$. For any integer $z \in [0, n]$, the probability for the event $Z = z$ to occur can be computed as follows:

$$\text{Pr}\{Z = z\} = \binom{n}{z} \left(\frac{1}{m}\right)^z \left(1 - \frac{1}{m}\right)^{n-z}.$$

Because n and m are known, $\text{Pr}\{Z = z\}$ is a function of a single variable z and thus denoted as $P(z)$.

Based on the probability distribution of Y and Z , the probability for the observed value of a counter, $C_f[i]$, $\forall i \in [0, l]$, to occur is

$$\begin{aligned}
& Pr\{X = C_f[i]\} \\
&= \sum_{z=0}^{C_f[i]} (Pr\{Z = z\} \cdot Pr\{Y = C_f[i] - z\}) \\
&= \sum_{z=0}^{C_f[i]} \binom{s}{C_f[i] - z} \left(\frac{1}{l}\right)^{C_f[i] - z} \left(1 - \frac{1}{l}\right)^{s - (C_f[i] - z)} P(z). \tag{2.14}
\end{aligned}$$

Let $y = C_f[i] - z$ to simplify the formula. The probability for all observed values in the storage vector of flow f to occur is

$$\begin{aligned}
L &= \prod_{i=0}^{l-1} Pr\{X = C_f[i]\} \\
&= \prod_{i=0}^{l-1} \left(\sum_{z=0}^{C_f[i]} \binom{s}{y} \left(\frac{1}{l}\right)^y \left(1 - \frac{1}{l}\right)^{s-y} P(z) \right). \tag{2.15}
\end{aligned}$$

The maximum likelihood method (MLM) is to find an estimated size $\hat{s}$ of flow f that maximizes the above likelihood function. Namely, we want to find

$$\hat{s} = \arg \max_s \{L\}. \tag{2.16}$$

To find $\hat{s}$, we first apply logarithm to turn the right side of the equation from product to summation.

$$\ln(L) = \sum_{i=0}^{l-1} \ln \left(\sum_{z=0}^{C_f[i]} \binom{s}{y} \left(\frac{1}{l}\right)^y \left(1 - \frac{1}{l}\right)^{s-y} P(z) \right). \tag{2.17}$$

Because $\frac{d\binom{s}{y}}{ds} = \binom{s}{y}(\psi(s+1) - \psi(s+1-y))$, where $\psi(\dots)$ is the polygamma function [1], we have

$$\begin{aligned}
& \frac{d\left(\binom{s}{y} \left(1 - \frac{1}{l}\right)^{s-y}\right)}{ds} = \\
& \binom{s}{y} \left(1 - \frac{1}{l}\right)^{s-y} \left(\psi(s+1) - \psi(s+1-y) + \ln\left(1 - \frac{1}{l}\right) \right).
\end{aligned}$$

To simplify the presentation, we denote the right side of the above equation as $O(s)$. From (2.17), we can compute the first-order derivative of $\ln(L)$ as follows:

$$\frac{d \ln(L)}{ds} = \sum_{i=0}^{l-1} \frac{\sum_{z=0}^{C_f[i]} \left(O(s) \left(\frac{1}{l} \right)^y P(z) \right)}{\sum_{z=0}^{C_f[i]} \binom{s}{y} \left(\frac{1}{l} \right)^y \left(1 - \frac{1}{l} \right)^{s-y} P(z)}. \quad (2.18)$$

Maximizing L is equivalent to maximizing $\ln(L)$. Hence, by setting the right side of (2.18) to zero, we can find the value for $\hat{s}$ through numerical methods. Because $\frac{d \ln(L)}{ds}$ is a monotone function of s , we can use the bisection search method to find the value $\hat{s}$ that makes $\frac{d \ln(L)}{ds}$ equal to zero.

2.6.2 Estimation Accuracy

We derive the estimation confidence interval as follows: The estimation formula is given in (2.16). According to the classical theory for MLM, when l is sufficiently large, the distribution of the flow-size estimation $\hat{s}$ can be approximated by

$$Norm\left(s, \frac{1}{\mathcal{I}(\hat{s})}\right), \quad (2.19)$$

where the *fisher information* $\mathcal{I}(\hat{s})$ [20] of L is defined as follows:

$$\mathcal{I}(\hat{s}) = -E\left(\frac{d^2 \ln(L)}{ds^2}\right). \quad (2.20)$$

In order to compute the second-order derivative, we begin from (2.11) and have the following:

$$\begin{aligned} Pr\{X = C_f[i]\} &= \frac{1}{\sqrt{2\pi}\Delta} e^{-\frac{(C_f[i]-\mu)^2}{2\Delta}} \\ \ln(Pr\{X = C_f[i]\}) &= -\ln(\sqrt{2\pi}\Delta) - \frac{(C_f[i] - \mu)^2}{2\Delta}, \end{aligned} \quad (2.21)$$

where $0 \leq i \leq l-1$. Performing the second-order differentiation, we have

$$\begin{aligned} \frac{d^2 \ln(Pr\{X = C_f[i]\})}{ds^2} &= -\frac{\mu'}{l\Delta} + \frac{(\frac{1}{2}(1 - \frac{1}{l}) + \mu - C_f[i])\Delta'}{l\Delta^2} \\ &+ \frac{1}{l\Delta^3} \left(1 - \frac{1}{l}\right) \left((\mu - C_f[i])\mu'\Delta - (\mu - C_f[i])^2\Delta' \right), \end{aligned} \quad (2.22)$$

where $\mu' = \frac{1}{l}$ and $\Delta' = \frac{1}{l}(1 - \frac{1}{l})$. Therefore,

$$\begin{aligned}
& E\left(\frac{d^2 \ln(\Pr\{X = C_f[i]\})}{ds^2}\right) \\
&= -\frac{\mu'}{l\Delta} + \frac{\frac{1}{2}(1 - \frac{1}{l})\Delta'}{l\Delta^2} + \frac{1}{l\Delta^3}(1 - \frac{1}{l})E(\mu - C_f[i])^2\Delta' \\
&= -\frac{1}{l^2\Delta} + \frac{3(1 - \frac{1}{l})^2}{2l^2\Delta^2}, \tag{2.23}
\end{aligned}$$

where we have used the following facts: $E(\mu - C_f[i]) = 0$ and $E(\mu - C_f[i])^2 = \Delta$. Because $L = \prod_{i=0}^{l-1} \Pr\{X = C_f[i]\}$, we have

$$\begin{aligned}
\mathcal{I}(\hat{s}) &= -E\left(\frac{d^2 \ln(L)}{ds^2}\right) = \sum_{i=0}^{l-1} E\left(\frac{d^2 \ln(\Pr\{X = C_f[i]\})}{ds^2}\right) \\
&= \frac{1}{l\Delta} - \frac{3(1 - \frac{1}{l})^2}{2l\Delta^2}. \tag{2.24}
\end{aligned}$$

From (2.19), the variance of $\hat{s}$ is

$$\text{Var}(\hat{s}) = \frac{1}{\mathcal{I}(\hat{s})} = \frac{2l\Delta^2}{2\Delta - 3(1 - \frac{1}{l})^2}. \tag{2.25}$$

Hence, the confidence interval is

$$\hat{s} \pm Z_\alpha \cdot \sqrt{\frac{2l\Delta^2}{2\Delta - 3(1 - \frac{1}{l})^2}}, \tag{2.26}$$

where Z_α is the α percentile for the standard Gaussian distribution.

2.7 Setting Counter Length

So far, our analysis has assumed that each counter has a sufficient number of bits such that it will not overflow. However, in order to save space, we want to set the counter length as short as possible. Suppose each measurement period ends after a pre-specified number n of packets are received. (Note that the value of n is the combined sizes of all flows during each measurement period.)

The average value of all counters will be $\frac{n}{m}$. We set the number of bits in each counter, denoted as b , to be $\log_2 \frac{n}{m} + 1$. Due to the additional bit, each counter can hold at least two times of the average before overflowing. If the allocated memory has M bits, the values of b and m can be determined from the following equations:

$$b \times m = M, \quad \log_2 \frac{n}{m} + 1 = b. \tag{2.27}$$

Due to the randomized counter sharing design, roughly speaking, the packets are distributed in the counters at random. In our experiments, the counter values approximately follow a Gaussian distribution with a mean of $\frac{n}{m}$. In this distribution, the fraction of counters that are more than four times of the mean is very small—less than 5.3 % in all the experiments. Consequently, the impact of counter overflow in CSM or MLM is also very small for most flows. Though it is small, we will totally eliminate this impact later in Sect. 2.10.4.

2.8 Flow Labels

The compact online data structure introduced in Sect. 2.4 only stores the flow size information. It does not store the flow labels. The labels are per-flow information, and it cannot be compressed in the same way we do for the flow sizes. In some applications, the flow labels are pre-known and do not have to be collected. For example, if an ISP wants to measure the traffic from its customers, it knows their IP addresses (which are the flow labels in this case). Similarly, if the system administrator of a large enterprise network needs the information about the traffic volumes of the hosts in the network, she has the hosts' addresses.

In case that the flow labels need to be collected and there is not enough SRAM to keep them, the labels have to be stored in DRAM. An efficient solution for label collection was proposed in [22]. A Bloom filter [2, 3] can be implemented in SRAM to encode the flow labels that have seen by the router during a measurement period, such that each label is only stored once in DRAM when it appears for the first time in the packet stream; storing each label once is the minimum overhead if the labels must be collected.

If we use three hash functions in the Bloom filter, each packet incurs three SRAM accesses in order to check whether the flow label carried the packet is already encoded in the Bloom filter. A recent work on one-memory-access Bloom filters [23] shows that three SRAM accesses per packet can be reduced to one. This overhead is further reduced if we only examine the UDF packets and the SYN packets (which carry the label information of TCP traffic). A recent study shows that UDF accounts for 20 % of the Internet traffic [4] and the measurement of our campus traffic shows that SYN packets accounts for less than 10 % of all TCP traffic. Therefore, the Bloom filter operation only needs to be carried out for less than 28 % of all packets, which amortizes the overhead.

2.9 Experiments

We use experiments to evaluate the estimation methods, CSM (Counter Sum estimation Method) and MLM (Maximum Likelihood estimation Method), which are designed based on the randomized counter sharing scheme. We also compare our

Table 2.1 Number of memory accesses and number of hash computations per packet

	Memory accesses	Hash operations	Constant?
CSM	2	1	Y
MLM	2	1	Y
CB	≥ 6	≥ 3	N
MRSCBF	4.47	4.47	N

methods with CB (Counter Braids) [21] and MRSCBF (Multi-Resolution Space-Code Bloom Filters) [19]. The evaluation is based on the performance metrics outlined in Sect. 2.1, including per-packet processing time, memory overhead, and estimation accuracy.

The experiments use a network traffic trace obtained from the main gateway of our campus. We perform experiments on various different types of flows, such as per-source flows, per-destination flows, per-source/destination flows, and TCP flows. They all lead to the same conclusions. Without losing generality, we choose TCP flows for presentation. The trace contains about 68 millions of TCP flows and 750 millions of packets. In each measurement period, 10 million packets are processed; it typically covers slightly more than 1 million flows.

2.9.1 Processing Time

The processing time is mainly determined by the number of memory accesses and the number of hash computations per packet. Table 2.1 presents the comparison. CSM or MLM performs two memory accesses and one hash computation for each packet. CB incurs three times of the overhead. It performs six memory accesses and three hash computations for each packet at the first counter level, and in the worst case makes six additional memory accesses and three additional hash computations at the second level. MRSCBF has nine filters. The i th filter uses k_i hash functions and encodes packets with a sampling probability p_i , where $k_1 = 3$, $k_2 = 4$, $k_i = 6$, $\forall i \in [3, 9]$, and $p_i = (\frac{1}{4})^{i-1}$, $\forall i \in [1, 9]$. When encoding a packet, the i th filter performs k_i hash computations and sets k_i bits. Hence, the total number of memory accesses (or hash computations) per packet for all filters is $\sum_{i=1}^9 (p_i \cdot k_i) \approx 4.47$.

2.9.2 Memory Overhead and Estimation Accuracy

We provide the estimation accuracies of CSM and MLM under different levels of memory availability. In each measurement period, 10M packets are processed, i.e., $n = 10\text{M}$, which translates into about 8 s for an OC-192 link (10+ Gbps) or about 2 s for an OC-768 link (40+ Gbps) with an average packet size of 1,000 bytes.

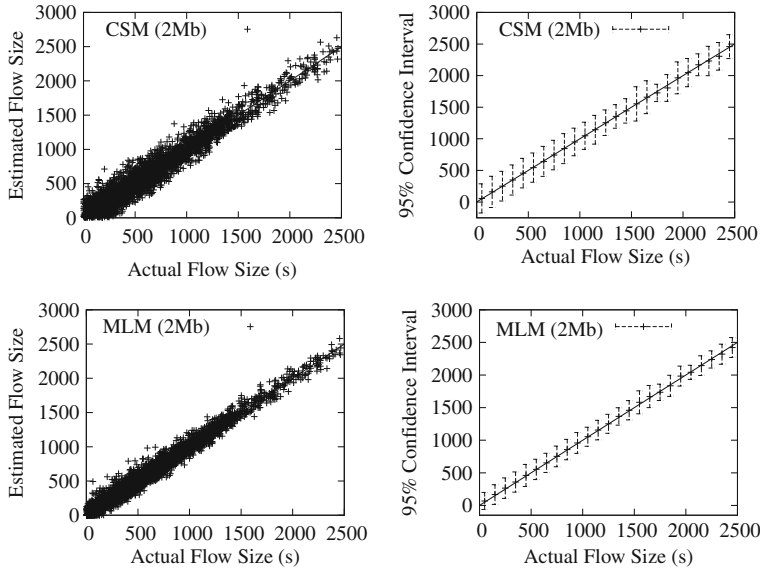


Fig. 2.2 Estimation results by CSM and MLM when $M = 2$ Mb

The memory M allocated to this particular measurement function is varied from 2 Mb (2×2^{20} bits) to 8 Mb. The counter length b and the number of counters m are determined based on (2.27). The size of each storage vector is 50.

When $M = 2$ Mb, the experimental results are presented in Fig. 2.2. The top left plot shows the estimation results by CSM for one measurement period; the results for other measurement periods are very similar. Each flow is represented by a point in the plot, whose x coordinate is the true flow size s and y coordinate is the estimated flow size $\hat{s}$. The equality line, $y = x$, is also shown for reference. An estimation is more accurate if the point is closer to the equality line.

The top right plot presents the 95 % confidence intervals for the estimations made by CSM. The width of each vertical bar shows the size of the confidence interval at a certain flow size (which is the x coordinate of the bar). The middle point of each bar shows the mean estimation for all flows of that size. Intuitively, the estimation is more accurate if the confidence interval is smaller and the middle point is closer to the equality line.

The bottom left plot shows the estimation results by MLM, and the bottom right plot shows the 95 % confidence intervals for the estimations made by MLM. Clearly, MLM achieves better accuracy than CSM. The estimation accuracy shown in Fig. 2.2 is achieved with a memory of slightly less than 2 bits per flow.

We can improve the estimation accuracy of CSM or MLM by using more memory. We increase M to 4 Mb and repeat the above experiments. The results are shown in Fig. 2.3. We then increase M to 8 Mb and repeat the above experiments. The results

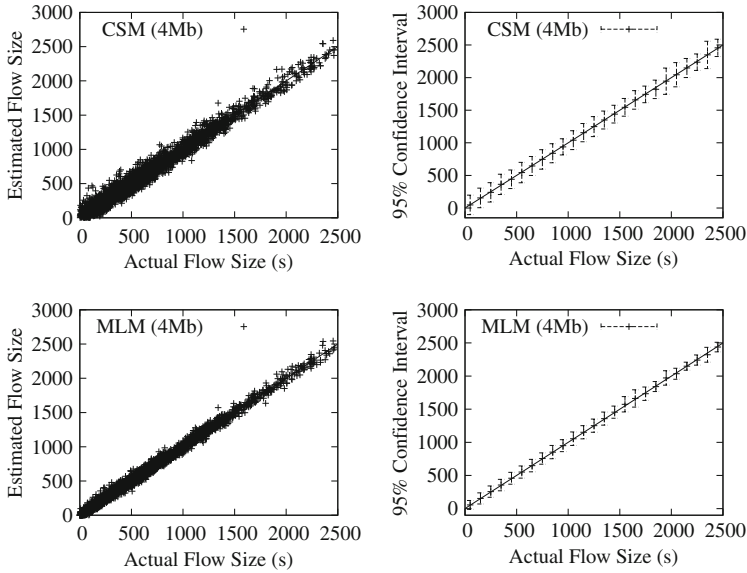


Fig. 2.3 Estimation results by CSM and MLM when $M = 4$ Mb

are shown in Fig. 2.4. The accuracy clearly improves as the confidence intervals shrink when M becomes larger.

We repeat the same experiments on CB, whose parameters are selected according to [21]. The results are presented in Fig. 2.5. The top left plot shows that CB totally fails to produce any meaningful results when the available memory is too small: $M = 2$ Mb, which translates into less than 2 bits per flow. In fact, its algorithm cannot converge, but instead produce oscillating results. We have to artificially stop the algorithm after a very long time. The top right plot shows that CB works well when $M = 4$ Mb. The algorithm still cannot converge by itself, even though it can produce very good results when we artificially stop it after a long time without observing any further improvement in the results. It can be seen that the results carry a small positive bias because most points are on one side of the equality line. The bottom plot shows that CB is able to return the exact sizes for most flows when the memory is $M = 8$ Mb.

Combining the results in Table 2.1, we draw the following conclusion: (1) In practice, we should choose CSM/MLM if the requirement is to handle high measurement throughput (which means low per-packet processing time) or if the available memory is too small such that CB does not work, while relatively coarse estimation is acceptable. (2) We should choose CB if the processing time is less of a concern, sufficient memory is available, and the exact flow sizes are required.

We also run MRSCBF under different levels of memory availability. We begin with $M = 8$ Mb. CSM or MLM works very well with this memory size (Fig. 2.4). The performance of MRSCBF is shown in the top left plot of Fig. 2.6. There are some very

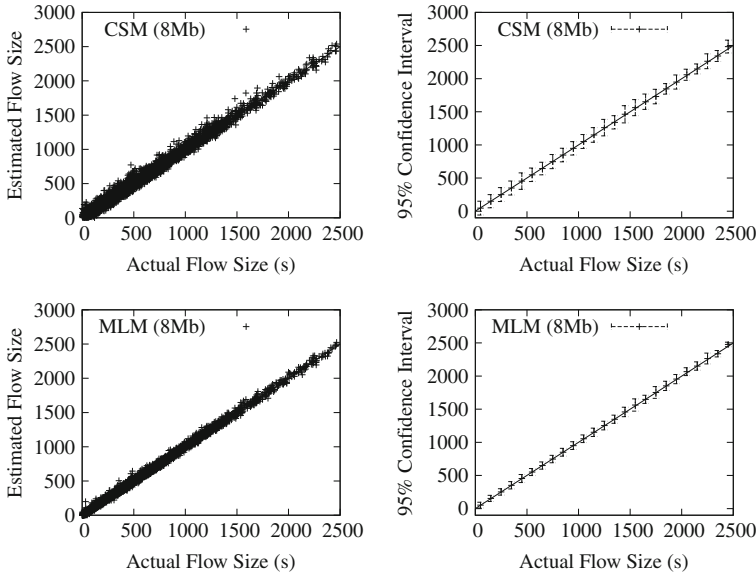


Fig. 2.4 Estimation results by CSM and MLM when $M = 8$ Mb

large estimated sizes. To control the scale in the vertical axis, we artificially set any estimation beyond 2,800 to be 2,800. The results demonstrate that MRSCBF totally fails when $M = 8$ Mb. The performance of MRSCBF improves when we increase the memory. The results when $M = 40$ Mb are shown in the top right plot.¹ In the bottom left plot, when we further increase M to 80 Mb,² no obvious improvement is observed when comparing with the top right plot. A final note is that the original work of MRSCBF uses log scale in their presentation. The bottom left plot in Fig. 2.6 will appear as the bottom right plot in log scale.

Clearly, the bitmap-based MRSCBF performs worse than CB, CSM or MLM. To measure flow sizes, counters are superior than bitmaps.

2.10 Extension of Estimation Range

We set the upper bound on the flow size that CSM and MLM can estimate in Sect. 2.9 to 2,500. However, in today's high-speed networks, the sizes of some flows are much larger than 2,500. In order to extend the estimation range to cover these large flows, we present four approaches that increase the estimation upper bound, and

¹ At the end of each measurement period, about half of the bits in the filters of MRSCBF are set to ones.

² At the end of each measurement period, less than half of the bits in the filters of MRSCBF are set to ones.

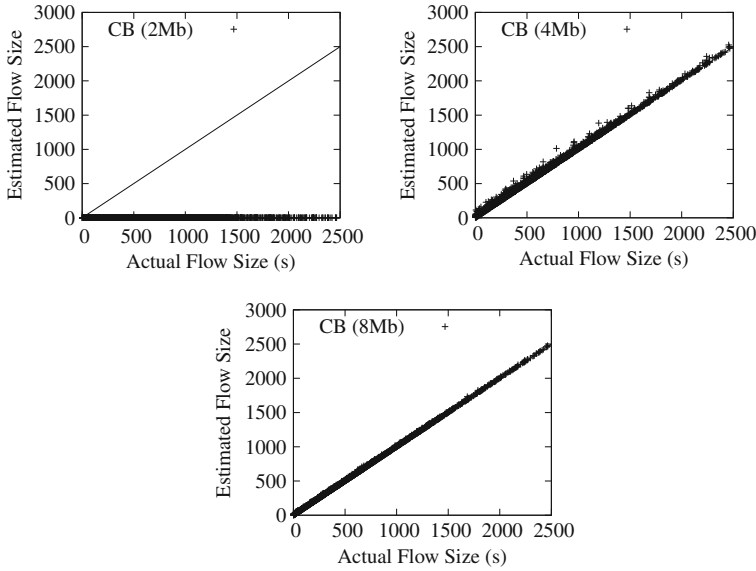


Fig. 2.5 Estimation results by CB when $M = 2, 4$, and 8 M

present extensive experimental results to demonstrate their effectiveness. Since MLM generally performs better than CSE, we only discuss how to extend the estimation range of MLM. CSE can be easily enhanced by similar approaches.

According to Sect. 2.4, each flow is assigned a unique storage vector. A flow's storage vector consists of l counters and each counter has b bits. Therefore, the maximum number of packets that the storage vector can represent is $l \times (2^b - 1)$. If we increase b by one, the number of packets that the vector can represent will be doubled. Similarly, if we increase l by a certain factor, the number of packets that the vector can represent will be increased by the same factor. Based on these observations, we extend the estimation range of MLM by increasing the value of b and l , respectively. In addition, we add a sampling module to MLM and consider hybrid SRAM/DRAM implementation to extend the estimation range.

2.10.1 Increasing Counter Size b

The first approach to extend the estimation range is to enlarge the counter size b . We repeat the same experiment on MLM presented in the bottom left plot of Fig. 2.3 (Sect. 2.9.2), where $M = 4\text{ Mb}$, $l = 50$, and $n = 10\text{ M}$. This time, instead of computing b from (2.27), we vary its value from 6 to 9. The new experimental results are shown in Fig. 2.7. In the top left plot, the maximum flow size that MLM can estimate is about 1,400 when $b = 6$. In the top right plot, where $b = 7$, the

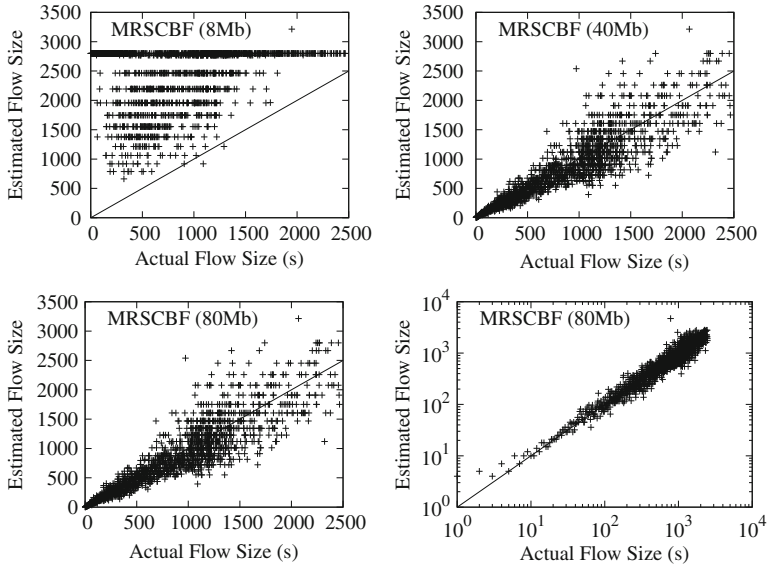


Fig. 2.6 Estimation results by MRSCBF

maximum flow size is about 2,800, which is twice of the maximum flow size that the top left plot can achieve. When b is set to 8, the bottom left plot shows that the estimation range of MLM is further extended. The bottom right plot shows that, when $b = 9$, the maximum flow size that MLM can estimate does not increase any more when comparing with the bottom left plot, which we will explain shortly.

The estimation accuracy of the above experiments is presented in Fig. 2.8, where the first plot shows the estimation bias and the second plot shows the standard deviation of the experimental results in Fig. 2.7. Generally speaking, both bias and standard deviation increase slightly when b increases.

Since flows share counters in MLM, the size information of one flow in a counter is the noise to other flows that share the same counter. When the amount of memory allocated to MLM is fixed ($M = 4$ Mb in these experiments), a larger value for b will result in a smaller value for m , i.e., the total number of counters is reduced. Hence, each counter has to be shared by more flows, and the average number of packets stored in each counter will increase. That means heavier noise among flows, which degrades the estimation accuracy, as is demonstrated by Fig. 2.8. Moreover, although a counter with a larger size b can keep track of a larger number of packets, since it also carries more noise, MLM has to subtract more noise from the counter value during the estimation process. As a result, the estimation range cannot be extended indefinitely by simply increasing b , which explains the fact that the maximum flow size that MLM can estimate does not increase when b reaches 9 in Fig. 2.7.

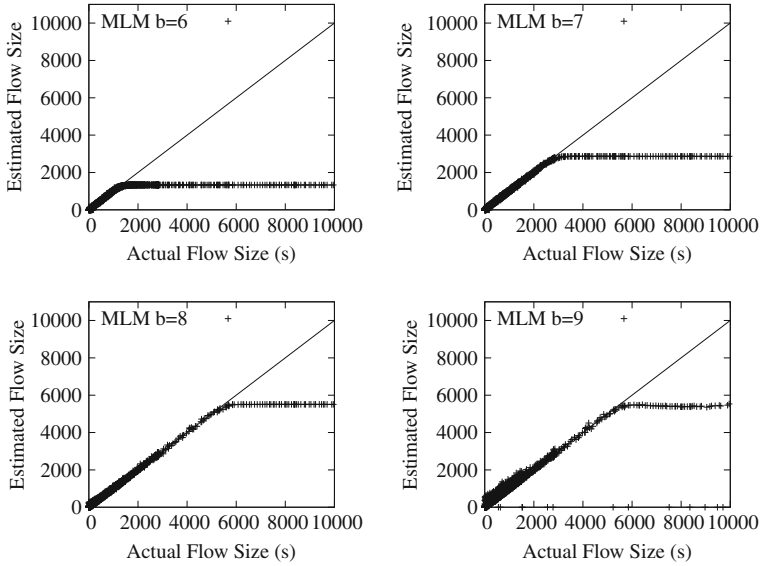


Fig. 2.7 Estimation results by MLM when $b = 6, 7, 8$, and 9 . In these experiments, $n = 10M$, $M = 4Mb$

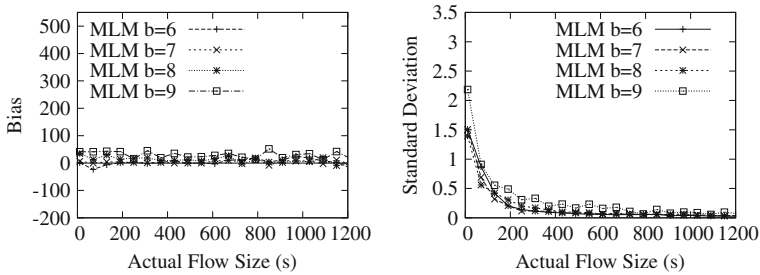


Fig. 2.8 Estimation bias and standard deviation in the experimental results shown in Fig. 2.7

2.10.2 Increasing Storage Vector Size l

The second approach for extending the estimation range is to increase the storage vector size l . We repeat the experiments in the previous subsection for MLM with $M = 4Mb$, $b = 7$, and $n = 10M$. We vary l from 50 to 1,000. Figure 2.9 presents the experimental results. The top left plot shows that the maximum flow size that MLM can estimate is about 5,800 when $l = 50$. As we increase the value of l , MLM can estimate increasingly larger flow sizes. However, when l becomes too large, estimation accuracy will degrade, which is evident in the bottom right plot. The reason is that each flow shares too many counters with others, which results

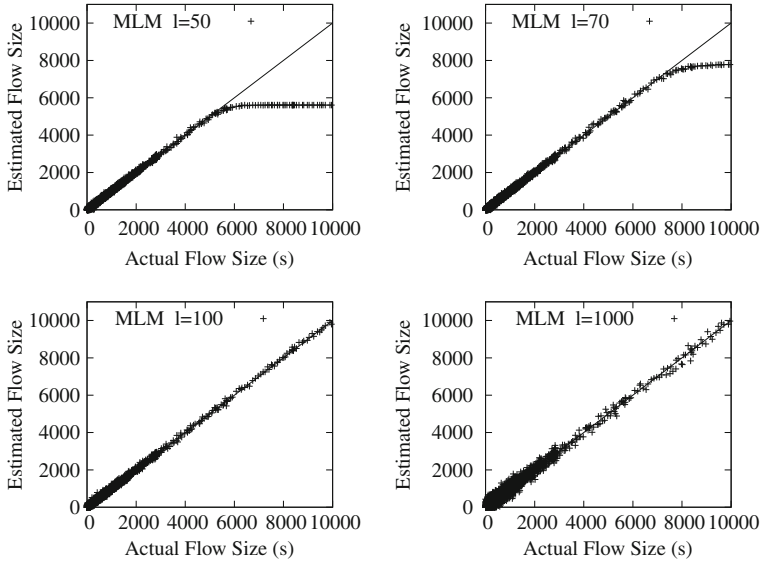


Fig. 2.9 Estimation results by MLM when $l = 50, 70, 100$, and $1,000$. In these experiments, $n = 10\text{M}$, $M = 4\text{M}$

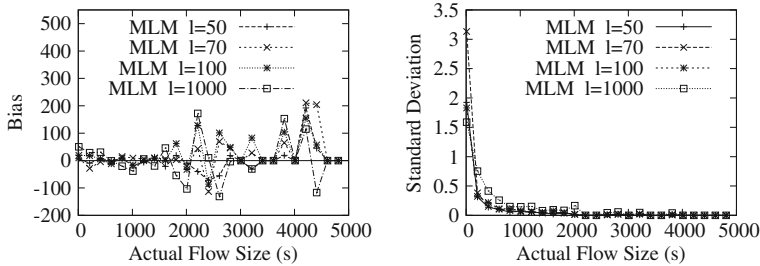


Fig. 2.10 Estimation bias and standard deviation in the experimental results shown in Fig. 2.9

in excessive noise in the counters and consequently introduce inaccuracy in the estimation process.

The estimation accuracy of the above experiments is presented in Fig. 2.10, where the first plot shows the estimation bias and the second plot shows the standard deviation of the experimental results in Fig. 2.9. Generally speaking, both bias and standard deviation increase slightly when l increases. Clearly, the value of l should not be chosen too large (such as $l = 1,000$) in order to prevent estimation accuracy to degrade significantly.

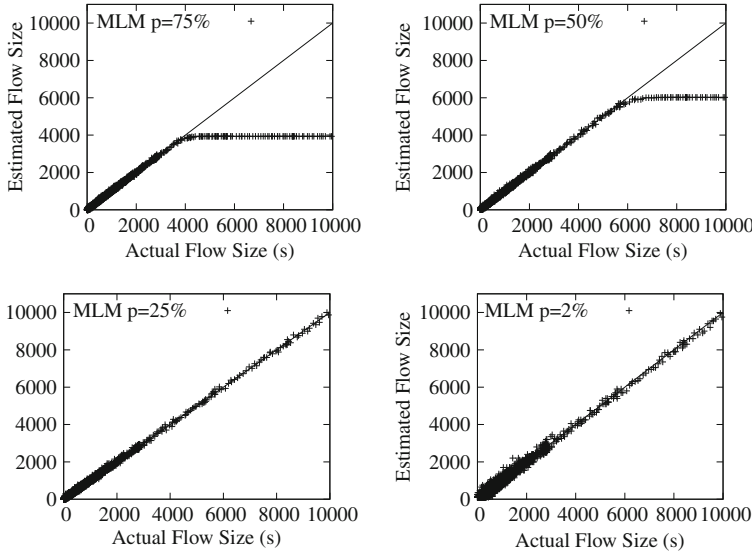


Fig. 2.11 Estimation results by MLM when $p = 75\%$, 50% , 25% , and 2% . In these experiments, $n = 10\text{M}$, $M = 4\text{Mb}$

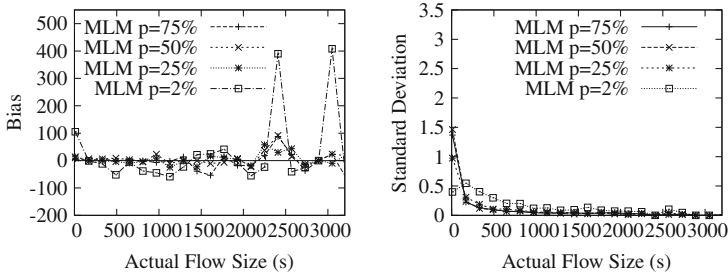


Fig. 2.12 Estimation bias and standard deviation in the experimental results shown in Fig. 2.11

2.10.3 Employing Sampling Module

In the third approach, we add a sampling module to MLM to enlarge the estimation range. The sampling technique has been widely used in network measurement [5, 10, 11, 19, 30]. We show that it also works for MLM. Let p be the sampling probability. For each packet that the router receives in the data encoding phase, the router generates a random number r in a range $[0, N]$. If $r < p \times N$, the router processes the packet as we describe in Sect. 2.4. Otherwise, it ignores the packet without encoding it in the counter array. In the data decoding phase, the estimated flow size should be $\frac{\hat{s}}{p}$, where $\hat{s}$ is computed from (2.18). The estimation range is expanded by a factor of $\frac{1}{p}$.

We again repeat the experiments in the previous sections for MLM with $M = 4\text{ Mb}$, $l = 50$, and $n = 10\text{ M}$. The value of b is computed from (2.27). This time, we introduce a sampling probability p and varies its value. Figure 2.11 presents the experimental results of MLM with $p = 75\%$, 50% , 25% , and 2% , respectively. It demonstrates that when the sampling probability decreases, the estimation range increases. However, it comes with a penalty on estimation accuracy. Figure 2.12 shows the estimation bias and standard deviation of the estimation results in Fig. 2.11. If the sampling probability is not decreased too small, e.g., when $p \geq 25\%$, the increase in bias and standard deviation is insignificant. However, if the sampling probability becomes too small such as 2% , the degradation in estimation accuracy also becomes noticeable.

2.10.4 Hybrid SRAM/DRAM Design

Can we extend the estimation range without any limitation and do so without any degradation in estimation accuracy? This will require a hybrid SRAM/DRAM design. In SRAM, we still choose the value of b based on (2.27). The limited size of each counter means that a counter may be overflowed during the data encoding phase even though the chance for this to happen is very small (Sect. 2.7). To totally eliminate the impact of counter overflow, we keep another array of counters in DRAM, each of which has a sufficient number of bits. The counters in DRAM are one-to-one mapped to the counters in SRAM. When a counter in SRAM is overflowed, it is reset to zero and the corresponding counter in DRAM is incremented by one. During offline data analysis, the counter values are set based on both SRAM and DRAM data. Because overflow happens only to a small fraction of SRAM counters and a DRAM access is made only after an overflowed SRAM counter is accessed 2^b times, the overall overhead of DRAM access is very small.

2.11 Summary

Per-flow traffic measurement provides real-world data for a variety of applications on accounting and billing, anomaly detection, and traffic engineering. Online data collection methods must meet the requirements of being both fast and compact. This chapter presents a novel data encoding/decoding scheme, which mixes per-flow information randomly in a tight SRAM space for compactness. Its online operation only incurs a small overhead of one hash computation and one counter update per packet. Two offline statistical methods—the counter sum estimation and the maximum likelihood estimation—are used to extract per-flow sizes from the mixed data structures with good accuracy. Due to its design philosophy that is fundamentally

different from the prior art, the new measurement function is able to work in a tight space where exact measurement is no longer possible, and it does so with the minimal number of memory accesses per packet.

References

1. Abramowitz, M., Stegun, I.: Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables. Dover Publications, New York (1964)
2. Bloom, B.H.: Space/Time trade-offs in hash coding with allowable errors. *Commun. ACM* **13**(7), 422–426 (1970)
3. Broder, A., Mitzenmacher, M.: Network applications of bloom filters: a survey. *Internet Math.* **1**(4), 485–509 (2002)
4. CAIDA: Analyzing UDP Usage in Internet Traffic. <http://www.caida.org/research/traffic-analysis/tcpudpratio/> (2009)
5. Cao, J., Jin, Y., Chen, A., Bu, T., Zhang, Z.: Identifying high cardinality internet hosts. In: Proceedings of IEEE INFOCOM (2009)
6. Casella, G., Berger, R.: Statistical Inference. Duxbury Press, Pacific Grove (2001)
7. Cohen, S., Matias, Y.: Spectral bloom filters. In: Proceedings of ACM SIGMOD (2003)
8. Cvetkovski, A.: An algorithm for approximate counting using limited memory resources. *Proceedings of ACM SIGMETRICS* (2007)
9. Demaine, E., Lopez-Ortiz, A., Munro, J.: Frequency estimation of internet packet streams with limited space. In: Proceedings of 10th ESA Annual European Symposium on Algorithms (2002)
10. Dimitropoulos, X., Hurley, P., Kind, A.: Probabilistic lossy counting: an efficient algorithm for finding heavy hitters. *ACM SIGCOMM Comput. Comm. Rev.* **38**(1), 7–16 (2008)
11. Duffield, N., Lund, C., Thorup, M.: Estimating flow distributions from sampled flow statistics. In: Proceedings of ACM SIGCOMM (2003)
12. Estan, C., Varghese, G.: New directions in traffic measurement and accounting. In: Proceedings of ACM SIGCOMM (2002)
13. Estan, C., Varghese, G., Fish, M.: Bitmap algorithms for counting active flows on high-speed links. *IEEE/ACM Trans. Netw. (TON)* **14**(5), 925–937 (2006)
14. Hao, F., Kodialam, M., Lakshman, T.V.: ACCEL-RATE: A faster mechanism for memory efficient per-flow traffic estimation. In: Proceedings of ACM SIGMETRICS/Performance (2004)
15. Kamiyama, N., Mori, T.: Simple and accurate identification of high-rate flows by packet sampling. In: Proceedings of IEEE INFOCOM (2006)
16. Karp, R., Shenker, S., Papadimitriou, C.: A simple algorithm for finding frequent elements in streams and bags. *ACM Trans. Database Syst.* **28**(1), 51–55 (2003)
17. Kodialam, M., Lakshman, T.V., Mohanty, S.: Runs based traffic estimator (RATE): a simple, memory efficient scheme for per-flow rate estimation. In: Proceedings of IEEE INFOCOM (2004)
18. Kumar, A., Sung, M., Xu, J., Wang, J.: Data streaming algorithms for efficient and accurate estimation of flow size distribution. In: Proceedings of ACM SIGMETRICS (2004)
19. Kumar, A., Xu, J., Wang, J., Spatschek, O., Li, L.: Space-code bloom filter for efficient per-flow traffic measurement. In: Proceedings of IEEE INFOCOM (2004)
20. Lehmann, E., Casella, G.: Theory of Point Estimation. Springer, New York (1998)
21. Lu, Y., Montanari, A., Prabhakar, B., Dharmapurikar, S., Kabbani, A.: Counter braids: a novel counter architecture for per-flow measurement. In: Proceedings of ACM SIGMETRICS (2008)
22. Lu, Y., Prabhakar, B.: Robust Counting via counter braids: an error-resilient network measurement architecture. In: Proceedings of IEEE INFOCOM (2009)
23. Qiao, Y., Chen, S., Li, T., Chen, S.: Energy-efficient polling protocols in RFID systems. In: Proceedings of ACM MobiHoc (2011)

24. Ramakrishna, M., Fu, E., Bahcekapili, E.: Efficient hardware hashing functions for high performance computers. *IEEE Trans. Comput.* **46**(12), 1378–1381 (1997)
25. Stanojevic, R.: Small active counters. In: *Proceedings of IEEE INFOCOM* (2007)
26. Venkatataman, S., Song, D., Gibbons, P., Blum, A.: New streaming algorithms for fast detection of superspreaders. In: *Proceedings of NDSS* (2005)
27. Yoon, M., Li, T., Chen, S., Peir, J.: Fit a spread estimator in small memory. In: *Proceedings of IEEE INFOCOM* (2009)
28. Yoon, M., Li, T., Chen, S., Peir, J.: Fit a compact spread estimator in small high-speed memory. *IEEE/ACM Trans. Netw.* **19**(5), 1253–1264 (2011)
29. Zhao, H., Wang, H., Lin, B., Xu, J.: Design and performance analysis of a DRAM-based statistics counter array architecture. In: *Proceedings of ACM/IEEE ANCS* (2009)
30. Zhao, Q., Kumar, A., Xu, J.: Joint data streaming and sampling techniques for detection of super sources and destinations. In: *Proceedings of USENIX/ACM Internet Measurement Conference* (2005)



<http://www.springer.com/978-1-4614-4850-1>

Traffic Measurement on the Internet

Li, T.; Chen, S.-S.

2013, IX, 82 p. 24 illus., Softcover

ISBN: 978-1-4614-4850-1