# A Traffic Sampling Model for Measurement Using Packet Identification

*Cheng Guang, Gong Jian, Ding Wei*

Department of Computer Science & Engineering
Southeast University
Nanjing, Jiangsu Province, 210096, P.R.C.

## Abstract

A new sampling model for measurement using packet identification on IP network is provided in this paper under a principle of PSAMP, a working group of IETF(to be set up), that a good sampling model should work for all purposes of measurement applications at the same time with a simple way. After researching and analyzing huge amounts of packet headers captured randomly on CERNET backbone, the result shows that 16 bits of identification field in IP packet header is enough for matching bits of sampling mask. Randomization and statistical attribute of the sampling are analyzed in the paper, and the experiment also reveals that this sampling way can be used not only in traffic measurement but also for network behavior analysis.

**Key Words:** Sampling Measurement, Packet Identification, Masking Bit, Packet Sampling

## 1. Introduction

With the rapid development of Internet applications, network behavior problems grow quickly and become more and more complex, which makes the study on it a hot focus of relative research field now [1]. Network measurement [2] is the foundation of network behavior research, which includes active measurement and passive measurement. Active measurement [3] injects test traffic into network to measure its behavior. As test traffic always generates additional load on network links and routers, it influences the measurement results significantly. In contrast to this, a passive measurement [4] relies on the traffic which already exists in the network only, but it is difficult to work in a wide bandwidth environment and analyze the traffic on time. Passive measurement is applied to research traffic statistics behavior, such as accounting and traffic management. In recent years the passive measurement technology is also used in network behavior, such as end-to-end network behavior [5] and routing behavior [6] .

In 1993, Claffy [7] processed NSFNET backbone traffic with passive sampling measurement technology for statistical, and RFC2330 [8] suggested a Poisson sampling method for measure high-speed network traffic after analyzed its randomicity, but what they did couldn't work for network behavior. On the other hand, I. COZZANI [5] used bit pattern checksum sampling model for end-to-end QoS in ATM network, and N. DUFFIELD [6] analyzed routing behavior with hash function sampling model. Unfortunately, without guarantee of randomicity, these two methods couldn't be used in sampling traffic and the statistics behavior analyzing on it.

PSAMP [9] suggests sampling model should work for all purposes of measurement applications at the same time with a simple way. The passive measurement is used in two applications mainly, which are traffic behavior analysis and network behavior analysis. The most important issue for the two analyses are insuring randomicity of measurement sampling and keeping consistency of sample at different measuring points respectively. Statistical analyzing of huge amounts of packets on CERNET backbone shows that bits in IP packet identification field possess high randomizing and independent identity distribution, that means by using IP packet identification a sampling measurement model with statistical randomicity and consistency of sample is available.

In the following sections, a sampling measurement model is proposed firstly, then the randomicity of packet bits is compared among packet header fields, after that the paper analyzes the randomicity of measurement sample. The conclusion is given out finally.

## 2. Sampling Measurement Model

### 2.1 Conception

Entropy [10], an important concept of information theory, is being used to measure random degree of various random experiments, which is extended to estimate bit randomicity in this paper. Some concepts about sampling model are defined firstly.

Definition 1: **Bit Entropy**, the entropy value of a bit, is defined as:

$$H(b) = -(p_0 \log_2 p_0 + p_1 \log_2 p_1) \qquad (1)$$

Where $p_0$ is the probability of b=0, and $p_1$ is the probability of b=1.

Theorem 1, **Maximal Bit Entropy Theorem**. In Definition 1, if and only if $p_0 = p_1 = 1/2$, the maximal bit entropy value $H_{max}(b) = 1$ is reached.

Definition 2, **Information Efficiency E of A Bit Entropy**, a metric of bit randomicity, is represented by the ratio between H(b) and $H_{max}(b)$. Due to $H_{max}(b)=1$, $E = H(b)/H_{max}(b) = H(b)$, $0 \le E \le 1$. The randomicity of a bit becomes larger, when E approaches 1, and vice versa.

Definition 3: **Bit Flow Entropy**, is defined as:

$$H(s) = -\sum_{i=0}^{2^s-1} p_i \log_2 p_i \qquad (2)$$

where s is the length of a bit flow which has $n+1=2^s$ events all together, and $p_0, p_1, \cdots, p_n$ are probabilities of each event.

Theorem 2, **Maximal Bit Flow Entropy Theorem**. In Definition 3, if and only if the $2^s$ events of a bit flow have the same probability, that means, $p_0=p_1=\cdots=p_n= 1/2^s$, the maximal bit flow entropy

$$H_{\max}(s) = -\sum_{i=0}^{2^s-1} \frac{1}{2^s} \log_2 \frac{1}{2^s} = s \qquad (3)$$

is reached.

Definition 4, **Information Efficiency E of A Bit Flow**, a metric of bit flow randomicity, is the ratio between H(s) and $H_{max}(s)$ : $E = H(s)/H_{\max}(s) = H(s)/s$ .

## 2.2 Sampling Measurement Model

PSAMP suggests that sampling measurement model should be simple and suffice for a wide range of measurement applications that include traffic statistics analysis and network behavior research. Measurement sample should be random for traffic statistics analysis, but it should be kept consistency at different points for network behavior research on the other hand. By choosing mask bits from an IP packet, a sampling measurement model on statistics analysis is provided in this section, which can assure not only randomicity of the sample but also its consistency at different measuring points.

Sampling measurement on high-speed IP traffic for statistical data is aimed at estimating total traffic information by selected samples. Sampling theory is based on randomicity in this case. The veracity of estimation depends on the randomicity of samples. On the other hand, for network behavior analysis, a set of samples is chosen from heavy traffic, and the result is worked out by the research on it only, so keeping consistency of selected samples at all measuring points is the most important, that means a measuring information is available if and only if it is obtained from the samples which are measured at each needed point.

Choosing some fixed bits in an IP packet as the measuring sample, consistency is obtained simply. Suppose these chosen bits can be proved to assure statistical randomicity, then they can be used as the mask bits of sampling in measuring model for traffic statistical analysis. Figure 1 shows an example of the sampling measurement model. If both probabilities that each masking bit appears as 0 or 1 are 0.5 separately and the bits in mask obey independency identity distributing, the theoretic sampling ratio is settled by the mask length n, which equals $1/2^n$. Actually, it is impossible to suppose each bit in mask has the equal probability and keep independent identity distributing, so sampling ratio is settled by masking bits directly. As Information Efficiency E of Bit Entropy approaches 1, the sampling ratio approaches theoretic one. Following factors should be satisfied when choosing masking bits from packet. Firstly, the masking bits can't be changed during the whole transportation to insure the consistency of sample. Secondly, the masking bits should have high randomicity for sampling ratio equals theoretic ratio statistically, and finally the masking ratio should be independent of statistical content of packet.

It is easy to assure measuring consistency with this sampling model, but randomicity can't be proved by mathematics method, and can only be analyzed from network traffic statistically. In the following sections, bit entropy of CERNET backbone traffic is analyzed, and bits with maximal bit entropy are chosen as the masking bits.

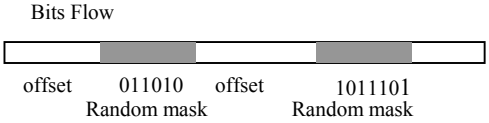


**Figure.1** Sampling Model

# 3. Statistical Analysis of IP Packet Header

## 3.1 Bit Entropy Analysis of the Header

Study in following sections shows that the entropy of identification field in IP packet header is quite large and its content is not changed during transportation. This result comes from statistical analysis to the first 20 bytes or 160 bits in 100000000 IP packets captured from CERNET backbone link directly by a system which is developed for measuring backbone traffic with hardware of a1000Mbps network card, PIII 1G CPU, and Red Hat Linux6.2 operating system.

As all captured packets are Ipv4 and their IHL are 5, information efficiency of bits entropy in these two fields(version field and IHL field) are both 0.

For TOS field, there are only 0.021% packets which use the 1th to 3th bits in this field to express PRI. 2.55% packets use the 4th bit, 2.98% use the 5th, and 0.03% use the 6th. The bit entropies from the 4th to the 6th bit are 0.171, 0.193 and 0.004 separately. The last two bits aren't defined, and there are only 40 packets use them within all 100000000. From these statistical figures, the bit entropy of TOS field is too little, and some bits in this field may be changed when packets pass through routers, so it isn't suitable to be used as matching bit flow of sampling mask.

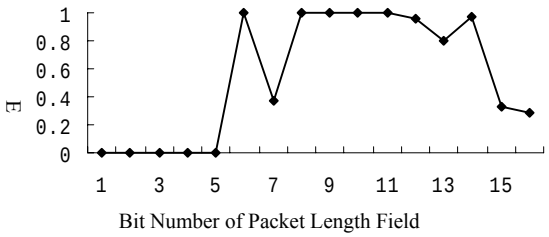


**Figure 2** E of Packet Length Field

Values of packet length field are always 40, 552, 576 and 1500. Entropy efficiency E of this field is shown as Figure 2. It is quite small in the first byte, but is larger than 90% in the second one. As what has happened in TOS field, the bits may be changed in this field through network, it couldn't be used as matching bits either.

Among all fields in IP packet header, identification field is the most suitable one for being used as matching bits of the sampling mask because of its statistical value of E shown as Figure.3 on one hand, and its consistency which means no bit changes its value during the whole transportation on the other hand.

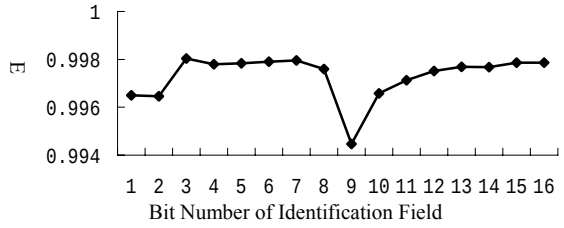


**Figure 3.** E of Identification Field

The first bit of Flag field isn't used. The value of E for DF bit is 88.7% and MF is 0.31% statistically. The other 13 bits E value of fragment field is shown in Figure 4.

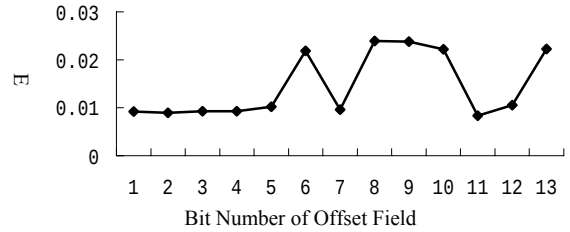


**Figure 4.** E of Offset Field

TTL is decremented per hop. Protocol field has low bit entropy because 93.04% packets are identified as TCP(6), 6.37% of them are UDP(17), and others only 0.59%.

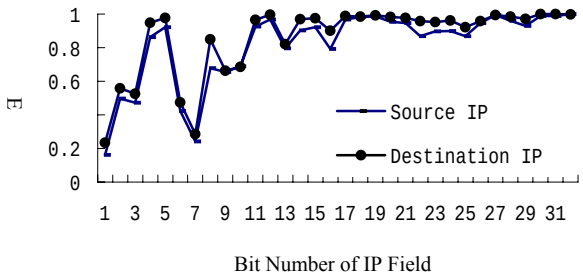


**Figure5.** E of IP Field

The entropy efficiency E of IP address fields both for source and destination are in Figure.5 which shows that the last 16 bits' E value in these fields are larger than 90%, so they can be operated as matching bits also.

According to bit entropies statistical analysis for each field in IP packet header, 16 bits of identification field, and the last 16 bits of source or destination IP address field, can be used as matching bits for they are not changed during transportation and have high efficiency of bit entropy information.

In the next section, interrelation and correlation between bits within a matching bits field is analyzed independently for the each 16 bits field being selected above.

## 3.2 Bit Flow Entropy Analysis

At different time segment, IP traffic on CERNET backbone is captured 10 times, 1,000,000 packets each time. The information efficiency of the three fields bits entropy is calculated and the results are shown in Figure 6. For identification field, the minimal information efficiency E of 16 bits is 0.901, the maximum one is 0.915 and wave range is 0.014. The minimal E of the last 16 bits of source and destination IP field are 0.648 and 0.544, the maximum one are 0.668 and 0.556, and waves ranges are 0.020 and 0.012 separately.

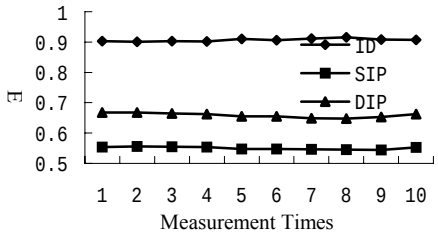


**Figure 6.** Comparison of E

Identification field is chosen as the random sampling matching bits finally for the more analysis because it has more strong stability than the other two.

# 4. Performance Analysis of Sampling Measurement Model

According to above analysis, the matching bits is choose from 1st to 16th bit in identification field, so ratio of sampling can be from 1 to $2^{16}$, the maximal sampling ratio can realize 65536, and this sampling model can work with traffic up to 640Gbps in theory. As a common PC can only deal with and store 10Mbps traffic now, the sampling model is much more than enough to support a system which uses PC measuring traffic. On the other hand, as there are bit operation in the sampling measurement model mainly which can be carried out through hardware easily, the sampling measurement is carried out in network card actually. The randomicity of the model and its statistical character of traffic sample will be analyzed below.

## 4.1 Randomicity Analysis of the Sampling Model

Figure 7 gives out the statistics analysis result of bits E from 1st to 16th bits in identification field with the 10,000,000 packets. In the figure, the minimal bits entropy is larger than 0.9, that means

the bits in identification field are independent and the autocorrelation between them are very small.

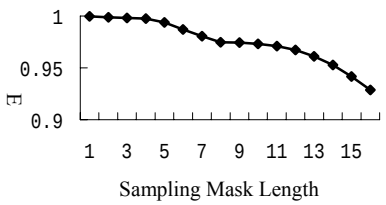


**Figure 7.** Comparison of the Bits E

The correlation between sampling mask length and sampling ratio is shown as Figure 8. If the sampling masking length is n bits, then the theory sampling ratio is $1/2^n$, and there are $2^n$ masks corresponding to sampling ratios. The maximal ratio, minimal ratio, median ratio, 95% ratio, and 5% ratio are listed separately. Except the maximal ratio, the others sampling ratio curves are quite near theoretical one. On the other hand, it is also proved by the statistics that value 0 is not suitable to be a sampling mask for its sampling ratio is much higher than any other values. From figure.8 we can also find that the identification field owns a good randomicity and is fit for matching bits.

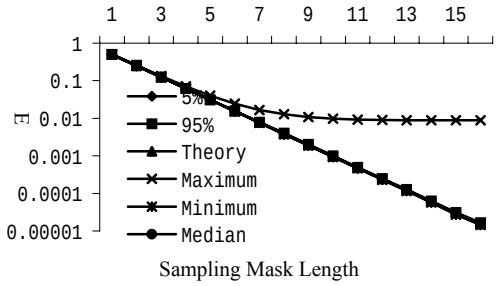


**Figure 8.** Relation between Sampling Ratio and Mask Length

## 4.2 Statistics Character of Traffic Sample

In this section, $\chi^2$ distribution [11] is used to examine the independence hypothesis of packet length, source IP from the samples and the full traffic. Suppose that statistical distribution of full traffic is $F_0(x)$, and the sample statistics is $F(x)$. For a given confidence level $\alpha$ ( $\alpha$ =0.05 or 0.01), independence hypothesis $H_0$:$F(x)=F_0(x)$ is tested by $\chi^2$ distribution.

After dividing the range value of full traffic into I bins according to characters (packet length, and source IP), when $n_i$ packets fall into bin i, the number of full packets is $n=\sum_{i=1}^{I} n_i$ . If there are $m_i$ packets in bins i, the number of sampling packets are $m=\sum_{i=1}^{I} m_i$ . The $\chi^2$ distribution statistics is

$$\chi^2=\sum_{i=0}^{I-1}\frac{(m_i-n_i p)}{n_i p}\sim\chi^2(I-1) \qquad (4)$$

where p is the sampling ratio, $n_ip=n_i\times m/n$ is the number of sample packets in theory. For a given confidence level $\alpha$ =0.05 or 0.01, if $\chi^2<\chi^2_\alpha$, accept the hypothesis. $\chi^2_\alpha$ is the $\alpha^{th}$ quantile of $\chi^2$ distribution with I-1 degrees of freedom.

The statistical attributes to the prefix of source IP and packet length in identification field are tested as matching bits from 1st to 16th bit in that field on 10000000 packets. In the test, 1, 10, 101, 0110, 10111, 101011, 1010111, 10100100, 110011101, 1010111000, 11110000111, 000011110000, 1010101010101, 10000001110010, 011000101110000, 0010101011101101 are selected as sampling masks, which are the 1st, 1st-2nd,--------,1st-16th bits in identification field respectively.

In the experiment, the prefix of source IP is tested with $I=2^5$. Due to the packet length between 40 bytes and 1500 bytes, a bin per 30 bytes, so 49 bins are built. In order to reduce the estimated error, when the theory sampling number in bin i is less than 5, it will be united. The results of test hypothesize for distribution of packet length and source IP prefix are shown in figure.9 and figure.10. From these figure, for a given confidence level $\alpha$ ( $\alpha$ =0.05 or 0.01), $\chi^2<\chi^2_\alpha$, the hypothesis $H_0$ can be accepted that means the sampling traffic has the same statistics distribution as the full traffic.

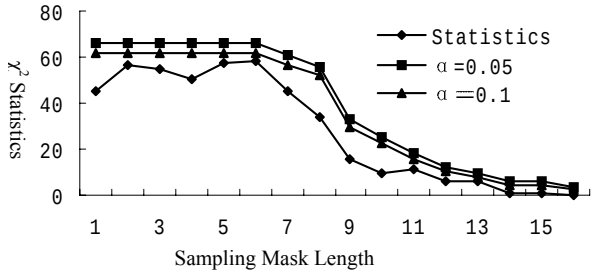


**Figure 9.** Hypothesis Distribution of Packet Length

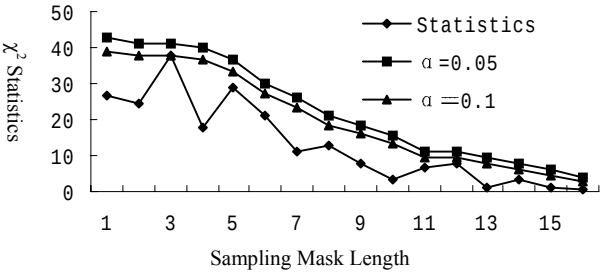


**Figure 10.** Hypothesis Distribution of IP Address Prefix

## 5. Conclusions

The sampling measurement on network traffic is a hot focus in network behavior research field. After a huge amounts of traffic passing through CERNET backbone is analyzed, we find that bits in packet identification of IP packet header aren't changed during the whole transport process and the randomicity of them are quite large. A sampling measurement model behind the fact is put forward in the paper, and efficiency of it is proved by randomicity and independency analysis to measuring sample. It is very easy also to apply the sampling model in a router or measurer. Because packet identification isn't changed while it is transported, so the consistency of sample in different points can be assured, that means the sampling model is used not only for traffic behavior analysis, but also in network behavior research.

## 6. Acknowledgments

The project is supported by the National Natural Science Foundation of China under grant No. 90104031, and the 863 program of China under grant No. 2001AA112060.

## 7. References

[1] Kevin Thompson, Gregory J. Miller, and Rick Wilder, Wide-Area Internet Traffic Patterns and Characteristics (Extended Version), *IEEE Network*, November/December 1997.

[2] CAIDA Homepage, http://www.caida.org.

[3] I. D. Graham, S. F. Donnelly, S. Martin, J. Martens, and J. G. Cleary, "Nonintrusive and Accurate Measurement of Unidirectional Delay and Delay Variation on the Internet," Proc. *INET '98*, Jul. 1998.

[4] B. Huffaker, Marina Fomenkov, David Moore, Evi Nemeth, K. Claffy, Measurements of the Internet topology in Asia-pacific Region, 2000, http://www.caida.org/outreach/papers/asia_paper/

[5] Cozzani, I.; Giordano, S, A passive test and measurement system: traffic sampling for QoS evaluation, *Global Telecommunications Conference, 1998. GLOBECOM 1998. The Bridge to Global Integration. IEEE* , Page(s): 1236 – 1241, Volume: 2 , 1998

[6] Nick Duffield, Matthias Grossglauser, Trajectory Sampling for Direct Traffic Observation, *Proceedings of ACM SIGCOMM 2000*, Stockholm, Sweden, August 28 – September 1, 2000

[7] K. Claffy, G. Polyzos, H. Braun, Application of Sampling Methodologies to Network Traffic Characterization, May 1993, *Proceedings of ACM SIGCOMM '93*.

[8] V. Paxson, G. Almes, J. Mahdavi, M. Mathis, Framework for IP Performance Metrics, *IETF* RFC 2330, 1998.

[9] Packet Sampling (psamp) Bof, Minutes of the Packet Sampling (PSAMP) BOF *IETF* 53, Minneapolis, Tuesday March 19, 2002; 9:00-11:30, http://www.ietf.org/proceedings/02mar/164.htm

[10] Jin Zhenyu, Information Theory, *Beijing University of Science and Technology Press*, pp： 11 ─ 47, 1991.12, BeiJing. (in Chinese)

[11] Tang Xiangneng, Dai Jianhua, Mathematics Statistics, *Mechanism Technology Press* pp： 140 ─ 151, 1994.5, BeiJing (in Chinese)