Research on the network data stream conversion based on different specifications of Flow

DONG Shi
Southeast University,Jiangsu Institute of Computer Science and Engineering
Jiangsu Province Key Laboratory of Computer Network Jiangsu Nanjing
shdong@njnet.edu.cn

DING Wei
Southeast University,Jiangsu Institute of Computer Science and Engineering
Jiangsu Province Key Laboratory of Computer Network Jiangsu Nanjing
wding@njnet.edu.cn

DING Feng
Southeast University,Jiangsu Institute of Computer Science and Engineering
Jiangsu Province Key Laboratory of Computer Network Jiangsu Nanjing
fding@njnet.edu.cn

***Abstract*—Compared to IP Trace,as a network form the data stream has a small storage and high morphology content advantage,different stream will causes different result when it act on the same data.This paper lead the compression ratio mold and combined with the analysis which is based on the common data stream conversion systerm for different flow parameter IP Trace.The result show that transforming for certain regular data stream can accept a idealized compression ratio and save the much storage space.**

Keywords- Data flow; flow specification; network measurement

I. INTRODUCTION

As monitoring, understanding and knowledge of network behavior, network measure and analysis become the more improtant way to improve the quality of the network and be paid more attention by researchers and operators. Data flow is the basis of network behavior analysis. The data flow(flow for short in following) is refer to the series data packet of special flow specification and flow timeout.[1] For the same data packet sequence(IP Trace),different flow specification will has different sequence in order to meet the needs of different researchers. On the other hand, these flow specification has certain relevance,this thesis works around this association.this paper design and realize the original system of data flow conversion which is based on analysis of these flow conversion.Then we use this systerm and a measured IP Trace data to analyse the different flow compression ratios.the result shows that the high cost-effective has a certain reference value for the data management.Compression ratio of the original IP Trace are from the collector of Jiangsu network CERNET 2.5G. the collect date is between 14 pm to 15 pm on March 16th 2008,Jan 25th 2009.Aug 20th 2009, the totally data quantity is 65G.these data are all store by 200MB unit which about 330 pieces file.

II. DATA FLOW CONVERSION

Research and analysis (such as Ryu who propose the adaptive timeout algorithm [2]) presents the norm of different flow sepcification and flow timeout. On account of the F1 achieve the good equilibrium ponit during the test of router management and RAM consumption[3], so that it is consider as the appropriate point during the backbone flow timeout. So this thesis choose the six regular flow specfciation from F1 to F6 as our analysis object. The specification as the following chart 1.

This six flow exsist the certain conversion relationship each other, for example.F1 can change to the relevant IP Trace F2 directly according the flow sequence. As well F2 can change to F3 directly.this conversion is single side, the element grit change from big to small,the reference change from small to big.The conversion between F1 to F6 as the chart show.

sign	F1	F2	F3	F4	F5	F6
specification	③ 16sec	③ 64sec	② 64sec	① 64sec	① 16sev	② 16sec

①two tuples:source ip,dest ip;②three tuples:source ip,dest ip,proto；
③five tuples：source ip，dest ip,source port，Dest port，proto

Table 1 Flow specification ID

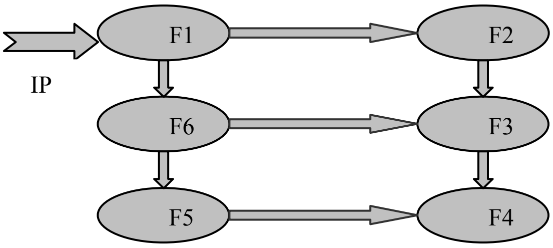

Chart 1 Transform relationship between the flow specificati

Figure 1 represents the conversion between the flows,this conversion has transitivity and exsists the relationship between the accessibility with the figure,from which we can know that F1 can reach every point,so that we can get the result is that F1 flow record has the list of high conservation values.according this theory we design and realize the systerm which froming by the data flow sequence.this systerm has two functions,one is that the original IP Trace can sequencing and store by F1 team flow and flow time starting.the second is that the systerm can change the matchable F1 flow sequence to the F2-F6 flow sequence.there is one explain is that every conversion is finished directly in this trial systerm.

978-1-4244-5392-4/10/$26.00 ©2010 IEEE

III. MODEL OF COMPRESSION RATIO

The first definition, compression ratio: different flows act on the same IP Trace which can achieve the store room ratio of sequence record.as following

$$Cr(Fk,Fj,IPTrace) = {|Fj(IPTrace)|} \Big/ {|Fk(IPTrace)|} \quad (1)$$

The same Fk and Fj act on the N pieces IP Trace which can achieve the assemblage that consist of Cr1-Crn. The caculation formula is that $AveCr = \overline{Cr} = \sum_{i=1}^{n} Cri / n$,n is the specimen quantity.for the different stage of IP Trace,we can use the compression ratio caculation formula to work and take the uncertain caculation into the account.

$$U = \sqrt{\left(\sqrt{\frac{\sum_{i=1}^{n} \left(Cri - \overline{Cr}\right)\left(Cri - \overline{Cr}\right)}{n(n-1)}} \right)} \quad (2)$$

The formula based on the compression ratio

- the forming of flow record sequence: Through the transform the systerm to achieve the flow sequence under the different flow specification.
- The compression ratio of the flow record: according the compression ratio model to caculate the compression ratio between different flow record sequences.
- Picking up the compression ratio of different flow record sequence from the conversion systerm and making the descending order according the compression raito magnitude.
- .Picking the maxmum compression raito from the different flow record sequence.
- Statistic analysis: when every data has finished the conversion,we will account and analyse it.

IV. EXPERIMENTAL METHODOLOGY

According the formula,the systerm and trail data we had introduced.we choose some flow specification team to test.

structure	bytes
first	8
last	8
srcaddr	4
dstaddr	4
srcport	2
dstport	2
proto	1
dPkts	4
dOctets	4

Table 2 Flow record data structure

From the data record of chart 2, we can know that the data structure field has 37 segment.so from F1 to F6, the record segment is 37,37,33,32,32,33. the result is showed in table 3.Table 4 is the maximum compression ratio which is based on the caculation of data structure.

accroding this result we can get a conclusion is as following

Some Conclusion

- compare with the F1 flow specification, the data flow record only needs 20% store room when adopt F4 flow specification store .
- The theoretically speaking, the compression raito of accessiblity flow specification has transtivity.that is $Com(Fi - Fk - Fj) = Com(Fi - Fk) * Com(Fk - Fj)$ .the charateristic is showed as Table 2. $Com(F1 - F2 - F3)$, $Com(F1 - F6 - F3)$, $Com(F1 - F2 - F3 - F4)$, $Com(F1 - F6 - F5 - F4)$, $Com(F1 - F6 - F5)$. According this theory and data of Table 2, we can find that the compression ratio has the certain deviation which achieved by transitivity caculation and systerm directly.the detail information is show as Table 5.the reason of this deviation we should have research in the future.
- F6-F5,F3-F4 this two kind of conversion raito is on the 95% more or less which is higher than other flow sequence ratio and approach to the maximum 0.9696.this result maybe due to that if we don’t take the protocol into account, the flow which has the same original address and can satisfy the timeout condition can merge one flow record during the triad flow change to the two-tuples. In the triad flow record the same original address recorded quantity is small. So when it change to the two-tuples the merge flow record has little proportion.In this situation, it is a possible happen in the internet data flow.This result shows that the flow record sequence of triple and two-tuples in same timeout has the more probablity in similarity.It has inferior effective when the triad flow record change into the two-tuples record sequence in saving the store room[4].On the same principle, it has a better effective when the five flow record sequence change into the two-tuples in saving store room.

Flow specification	compression ratio
F1-F2	0.8633± 0.0152
F1-F3	0.2123± 0.0358
F1-F4	0.2006± 0.0336
F1-F5	0.3344± 0.0655
F1-F6	0.3502± 0.0358
F6-F5	0.9512± 0.0053
F2-F3	0.2229± 0.0016
F3-F4	0.9422± 0.0057
F5-F4	0.6235± 0.0014
F6-F3	0.6282± 0.0108
F5-F3	0.6377± 0.0316
F2-F6	0.4073± 0.0802

Table 3 Flow between the various compression ratio

Flow specification	Maximum compression ratio
F1-F2	1.0000
F1-F3	0.8919
F1-F4	0.8649
F1-F5	0.8649
F1-F6	0.8919
F6-F5	0.9696
F2-F3	0.8919
F3-F4	0.9696
F5-F4	1.0000
F6-F3	1.0000
F5-F3	1.0312
F2-F6	0.8919

Table 4 Maximum compression ratio

	Compression ratio estimation	Measured the compression ratio
F1-F2-F3	0.1925	0.2123
F1-F6-F3	0.2199	0.2123
F1-F2-F3-F4	0.1813	0.2006
F1-F6-F5-F4	0.2076	0.2006
F1-F6-F5	0.3331	0.3344

Table 5 Calculated from the measured compression ratio and compression ratio

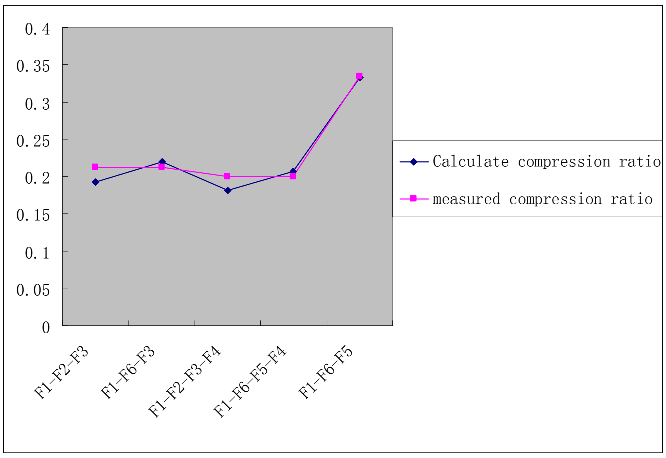

Chart 2 Calculate compression ratio compared with the measured compression ratio curve

V. CONCLUSION

Data flow has advantage in samll storage and high semantic content. Through the combination with the different description of flow semantic content and compression ratio, we obtain the high cost performance.it has a certain research significance in storage and management of mass data flow[5].During this paper we adopt the statistic analysis research method,we want to finish the statistic analysis [6]under prolong the length of IP Trace in the future work, expand the searching method from more angle and establish a more comprehensive math model.From the development of the future network technology,this research can supply the necessary technology support for the network application,such as checking the network attack[7],optimize the router[8] and so on.

REFERENCES

[1] CLAFFY K C. Internet Traffic Characterization. Dissertation for the degree Doctor of Philosophy[D]. University of California, San Diego. 1994.

[2] Ryu B, Cheney D, Braun hW. Internet flow characterization: adaptive timeout strategy and statistical modeling[C] Workshop on Passive and Active Measurement (PAM). 2001: 95 – 105.

[3] 王远, 丁伟， 龚俭. TCP 数据流超时研究[J]. 厦门大学学报, 2007, 46(2):192-195.

[4] 周明中，丁伟，龚俭. 网络流超时策略研究[J].通信学报，2005,26(4):88-93

[5] SHAIKH A, REXFORD J, SHIN K G Load-sensitive routing of long-lived IP flows[A]. Proceedings of SIGCOMM[C]. 1999.215-266.

[6] DUFFIELD N, LUND C, THORUP M. Properties and prediction of flow statistics from sampled packet streams[A]. IMW'02[C].Marseille, France. 2002.159-171.

[7] S. Templeton, K. Levitt, A Requires/Provides Model for Computer Attacks, in Proceedings of the New Security Paradigms Workshop, Cork, Ireland, September 2000.

[8] R. Teixeira, A. Shaikh, T Griffin, and J. Rexford. Dynamics of hot-potato routing in IP networks. In Proc. ACM SIGMETRICS, June 2004a