A Methodology for Analyzing Backbone Network Traffic at Stream-Level

HE Tao, ZHANG Hui, LI Xing, LI Zhichun
Network Research Center
Tsinghua University
CERNET, Beijing
Email: hetao@ieee.org
Telephone: (086) 010-62784301

***Abstract*— Most of network traffic analysis studies begin at the fine-grained level, such as packet-level and flow-level. Based on accurate measurements and complex analysis methods, these studies could not be carried on widely throughout a backbone network. In this paper, we introduce a methodology for analyzing network traffic at the coarse-grained level by incorporating stream information which could be achieved by simple passive monitoring. A careful study of many traffic traces acquired on an international link between two ISPs validates our work.**

I. Introduction

A number of problems in network operations and engineering call for new methods of traffic analysis. While most existing traffic analysis methods are fine-grained, [7], [10], there is a clear need for the analysis of traffic across a whole network links – that is, for coarse-grained traffic analysis.

We present a methodology for profiling traffic on a link at a higher granularity. Our methodology differs from many previous studies that have concentrated on endpoint definitions of flows in terms of state derived from observing the explicit opening and closing of TCP connections [6], and from detailed packet-level studies that are usually not available throughout a large network.

While stream level data is certainly not as precise as passive measurements of flow or packet level data, we demonstrate that it is sufficient for exposing some different types of unusual traffic across a whole network. It also has the benefit of generating much smaller data sets than other two level measurements, which tends to a significant issue in large, heavily used networks. We will also describe a study carried out using the form of stream data in the repository which we build.

The rest of this article is organized as follows. First of section II describes related work, then define stream and discuss several aspects of stream structure that frame our analysis. Applying the methodology to our measurement yields surprising insights into network traffic anomalies detection, which we review in section III. Further discussion will also be given in section III. In section IV, we give our conclusion.

II. Background and Model

Understanding the variability of Internet traffic in backbone networks is essential to better plan and manage existing networks, as well as to design next generation networks. However, most traffic analyzes that might be used to approach this problem are based on detailed packet or flow level measurements, which are unlikely applicable to a backbone network.

A. Packet-level analysis

In [1], RJ and SAR gave a packet train model of packet arrivals for describing traffic on a token ring local area network. They defined a packet train as a burst of packets arriving from the same source and heading to the same destination. If the spacing between two packets exceeds some inter-train gap, they are said to belong to separate trains. The packet train model reflects the fact that much of network communication involves many packets spaced closely in time between the same two endpoints. After that, many studies have extended the packet train model of flows to the transport or application layers [2], [3], [4], or focused only on TCP traffic flows [6], [7].

B. Flow-level analysis

Inspired by the model [1], K.C. Claffy, H.W. Braun, and G.C. Polyzos presented a parameterizable methodology [5] for profiling Internet traffic flows at a variety of granularities. They described individual Internet flows and explored descriptors of the entire population of flows. And more flow-level research [10] on modeling Internet backbone traffic.

Compared with packet-level analysis, our stream model need not care about the packet arrival by time-stamping every packet, and need not reflect to these flags pointing out some attributes of packets. Our stream model need not care about the timeout parameter and when a flow begins or when it ends either.

The stream only records the host pair and application traffic from source to destination. Any implementation in a router would not be needed by this methodology and removing the complex processing makes low requirement to passive monitoring mechanism.

C. Definition of Stream

In this section, we specify the definition of stream and then discuss some aspects of a stream that structure stream measurements and subsequent analysis.

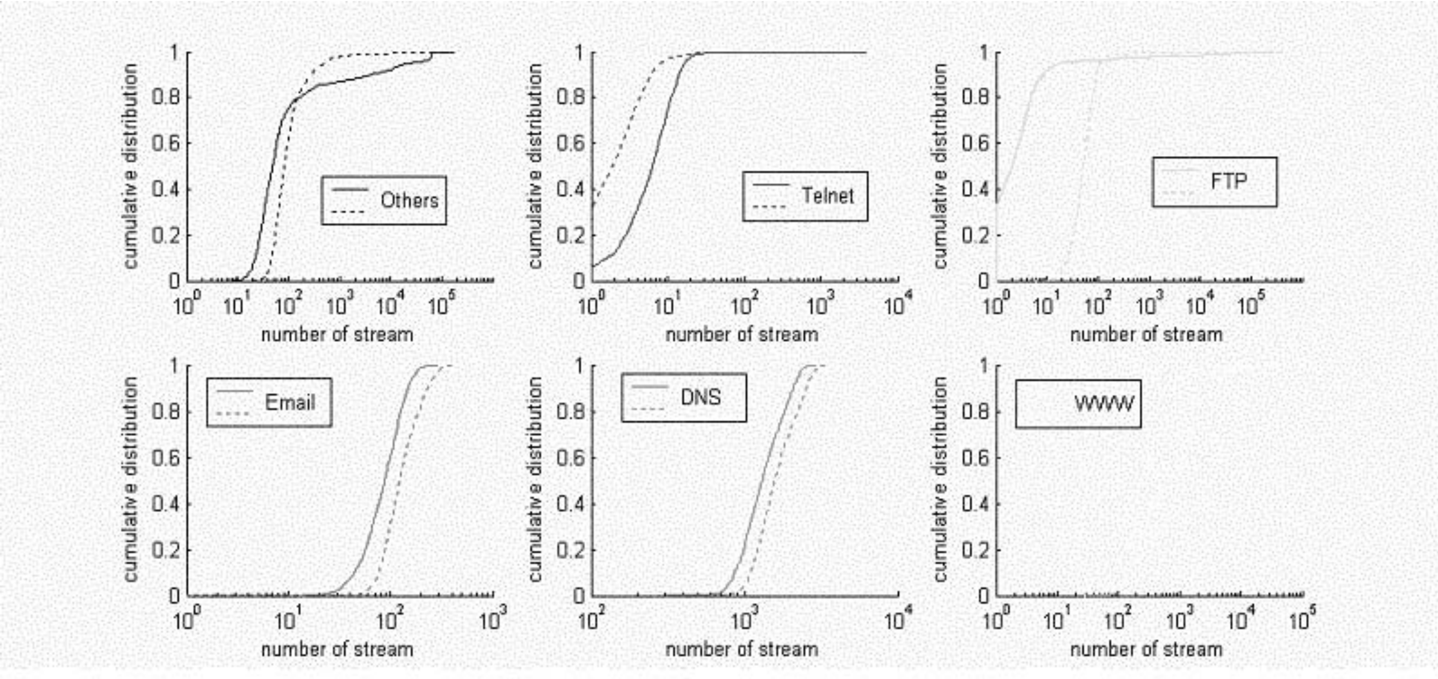


Fig. 1. Cumulative distribution of stream numbers per application/protocol

We ground our model of a stream on actual traffic activity from both of its transmission endpoints as perceived at a given network measurement point. The model isn't involved in the definition of timeout that used in studies of timeout-based traffic behavior [5]. Stream will exist during a specified time interval when the traffic activity happens. This interval-based stream definition allows flexibility in how one further specifies a stream. In particular, we describe three aspects of a stream that structure a stream specification: directionality, endpoint granularity and functional layer.

1) Stream directionality: First, we define a stream as unidirectional, i.e., bidirectional traffic between A and B show up as two bidirectional stream: traffic from A to B, and traffic from B to A. Unidirectional streams can be transformed into bidirectional streams during the analysis processing, based on actual requirements at that time.

2) Stream endpoint granularity: The second aspect of a stream is the endpoint granularity, or the extent of the communicating entities. Potential granularity include aspects which could be achieved easily such as host, IP block, and network number. For example, a stream may be defined to contain all traffic flowing between two IP hosts (i.e., host-to-host streams). Alternatively, a stream may contain traffic flowing between two IP blocks or networks. In order to expose unusual behavior involved to hosts, we assume stream by host pair. And this selection could be carried on easier than others.

3) Functional layer: Finally, there is the functional, or protocol, layer of the network stream. In order to maintain applicability across all traffic, we define streams based on packet transmission protocol which separated by well-known ports in source and/or destination endpoints.

In the real measurements, we use the following definitions of ”stream”: ”Stream” defined by 3-tuple, which is a stream of packets having the same source and destination IP addresses, and same protocol number. In that case, the size of a stream is measured in bytes in two directions, while the stream duration is not equal to the time interval between the first and last packet of the flow. In order to identify the end of a stream, we use a fixed interval of 1 hour: if the interval encounters, the stream is considered completed.

D. Metrics of Stream

Now, we will give several metrics, with details of the data collecting being described in section III:

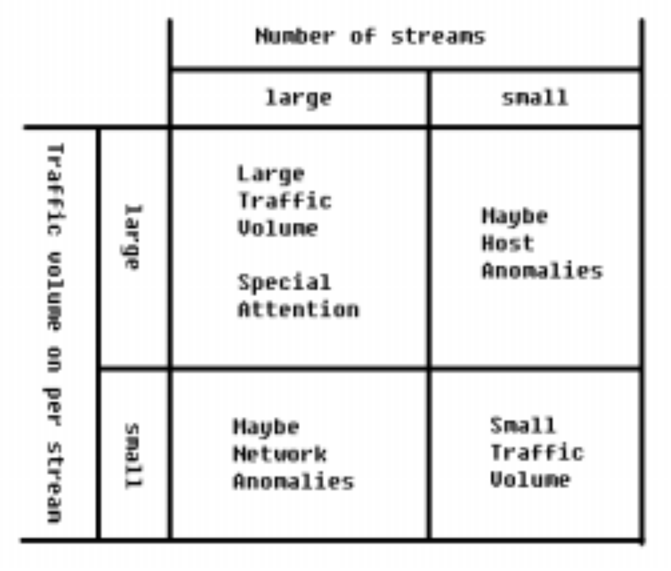
Inbound and outbound streams which described in figure I.

Per protocol and per application streams which could be seen on table II.

E. Analysis

As shown in figure I, for an interval value of 1 hour, the cumulative distributions of stream numbers for five common applications on the Internet, specifically telnet, ftp, email, dns, and www. We aggregate all other stream types into an ”others” category. All six graphs in the figure use the host pair stream granularity, i.e., aggregate all host pair streams with regard to the above six types, while using the data set which is got from the international link between CERNET and JANET.

Observing the quantification of each graph, one can learn that large majority of the stream numbers is from the dns protocol, and then www-http protocol. We also could get the



		Number of streams	
		large	small
Traffic volume on per stream	large	Large Traffic Volume Special Attention	Maybe Host Anomalies
	small	Maybe Network Anomalies	Small Traffic Volume

Fig. 2. Two dimension profile using traffic and stream

conclusion that stream numbers are close quantitatively in both directions, and that it is little probability of seeing very large and very small stream numbers.

In conventional networks, peak traffic is typically an aggregation of a very large number of individual burst traffic streams; in a high speed network peak traffic may represent one or a relatively small number of traffic sources, each requiring high bandwidth. In view of this we focus on not only the stream numbers, but also the traffic volume per stream carry.

To explore the interaction between the number of streams and the traffic volume which stream carried, we depict a two-dimensional profile, using the number of active streams per hour in one dimension, and the mean traffic volume on streams in the other dimension, as shown in figure 2.

III. Experimental Results

Our goal is to verify traffic metrics for a link between two large scale research network that is simple enough to be used in network operation, and that is packet agnostic in order to be as general as possible.

We collected three month traces on the link between CERNET and JANET.

A. Network Environment

The data used for this traffic study was taken from an international link between CERNET and JANET. The experimental research prototype monitoring system captures packets from the network and provides stream records, each of which consists of a 4 byte source address, a 4 byte destination address, a 2 byte protocol type, a 4 type traffic volume from source to destination, and a 4 type traffic volume from destination to source per host pair in fixed interval (here is 1 hour) from the network.

B. Traffic analysis of an international link between CERNET and JANET

TABLE I
PERCENTAGE OF STREAM IN A WEEK

	Traffic Volume	Mean Streams
Sunday	11.80%	11.69%
Monday	14.00%	12.76%
Tuesday	16.40%	19.80%
Wednesday	15.80%	18.52%
Thursday	14.00%	11.09%
Friday	16.40%	11.34%
Saturday	11.60%	14.79%

In this section, we present the analysis of the characteristics, which are fundamental for the stream-level traffic analysis approach of Section 2. First of all, we consider the usage fraction of link partitioned in the day of a week Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, and Saturday. Table I shows these fractions with respect to the overall data volume. We observe that the data volume at the weekend is dominated by other data volume portions, e.g., data volume in Saturday and Sunday added together only occupied 23.4% of total.

Compared with day fraction of traffic volume on a link we monitored, the mean stream faction does not show a strong week periodicity. After analyzing three month traffic traced, we noticed that stream numbers are close to network utilization:

1) Small stream number and small traffic volume won't give too much on our network.
2) When traffic volume per stream increase, we classified it ”burst” traffic. That's may be caused by some application.
3) When stream number increase, we pay more attention on it. It almost gives us an alarm that some unusual traffic emerged.

TABLE II
PERCENTAGE OF STREAM PER APPLICATION

	Traffic volume	Mean Streams
Telnet	0.80%	0.14%
FTP	18.30%	9.88%
Email	2.80%	2.02%
DNS	0.70%	26.97%
WWW	51.20%	18.44%
ICMP	1.90%	4.21%
Others	24.30%	38.34%

Furthermore, we analyze the application usage pattern of current link. Table II shows the application usage pattern with respect to the overall streams broken down in WWW, FTP, Email, DNS, Telnet, other TCP/UDP application, and ICMP. The surprising portion is ”Others” partition with 38.34%, which means that lots of types of stream exists on the link which we haven't realized it. However its traffic volume (24.30%) follows WWW (51.20%). Obviously, DNS is the secondary application with a fraction of 26.97%, but its traffic

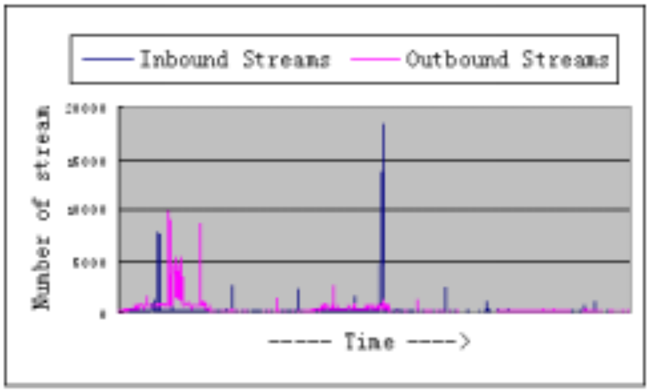


Fig. 3. ICMP peak streams which could be observed from the figure indicate ICMP traffic anomalies

volume occupied only 0.70% of total. The WWW application (i.e., 18.44%) is followed by the FTP application (i.e., 9.88%) and ICMP traffic (i.e., 4.21%). Email application with 2.02% and Telnet application with 0.14% do not contribute a significant amount of streams as same as they in traffic volume.

We analyze the accuracy of the metrics with three month traces collected on the link between CERNET and JANET. Despite its simplicity, the metrics provides a good assessment of the real traffic and of its anomalies observed on the link (e.g., ICMP traffic anomalies illustrated in figure 3).

The stream-level analysis methodology could be used in various applications such as detection of anomalies (e.g., denial of service attacks or probing attacks), prediction of traffic growth, or assessment of the impact on the network traffic of a new customer or of a new application. Consequently, a second important characteristic of the model we want to design is to be packet and flow agnostic: it needs to be general enough to evaluate link throughput independently of the packet nature and of the flow behavior.

IV. Conclusion and Future Work

In this paper, we present a methodology that relies on stream-level information observed on an IP backbone link. We are interested in capturing the dynamics of the traffic at fixed long timescales (i.e., in the order of one hour). We believe that effective network engineering must consider both the traffic properties on the network's links. Thus our approach is intended to support a whole-network view of data traffic.

We define a stream as the unique 3-tuple comprising a source IP address, destination IP address, and protocol of traffic. Stream-level information enables us to conduct a coarse analysis of traffic bursts. From a simplicity standpoint, it is much easier to monitor streams than to monitor packets or flows in a router.

While this level of reporting is not fine-grained enough so that short time scale behavior will be missed, it is sufficient for observing traffic flow anomalies macroscopically.

Our anomaly analysis process based on stream is only to aggregate anomalies with simple statistical features but not actually to characterize the features of anomaly groups rigorously.

Acknowledgment

The work presented in this paper is currently funded by National Natural Science Foundation of China under contract number 60103005 and by National High Technology Research and Development Program of China under contract number 2001AA142080.

References

[1] R. Jain and S. A. Routhier, ”Packet trains-measurement and a new model for computer network traffic”, IEEE Journal on Selected Areas in Communications, vol. 4, no. 6, pp. 986-995, September 1986.

[2] J. Mogul, ”Network locality at the scale of process”, in Proceedings of ACM SIGCOMM ’91, September 1991, pp. 273-285.

[3] M. Acharya, R. Newman-Wolfe, H. Latchman, R. Chow, and B. Bhalla, ”Real-time hierarchical traffic characterization of a campus area network”, in Proceedings of the Sixth International Conference on Modeling Techniques and Tools for Computer Performance Evaluation, 1992, University of Florida.

[4] M. Acharya and B. Bhalla, ”A flow model for computer network traffic using real-time measurements”, in Second International Conference on Telecommunications Systems, Modeling and Analysis, March 24-27 1994.

[5] Kimberly C. Claffy, Hans-Werner Braun, and George C. Polyzos, ”A parameterizable methodology for Internet traffic flow profiling”, IEEE Journal on Selected Areas in Communications 13, 1481-1494.

[6] R. Caceres, P. Danzig, S. Jamin, and D. Mitzel, ”Characteristics of wide-area TCP/IP conversations”, in Proceedings of ACM SIGCOMM ’91, September 1991, pp. 101-112.

[7] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, ”Modeling TCP Throughout: a Simple Model and its Empirical Validation”, in Proceedings of ACM SIGCOMM ’98 conference.

[8] R. Caceres, ”Measurements of wide-area Internet traffic”, Tech. Rep. UCB/CSD 89/550, Computer Science Department, University of California, Berkeley, 1989.

[9] V. Paxson and S. Floyd, ”Wide-area traffic: The failure of Poisson modeling”, IEEE/ACM Transactions on Networking, vol. 3(3), pp. 226-244, June 1995.

[10] C. Barakat, P. Thiran, G. Iannaccone, C. Diot, and P. Owezarski, ”A flow-based model for Internet backbone traffic”, ACM Internet Measurement Workshop, Marseille, November 2002.