2012 Third International Conference on Networking and Distributed Computing

Comparative Research on Internet Flows Characteristics

Xiaoguo Zhang[1,2], Wei Ding[1]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
[2]Electronic Information Engineering School, Henan University of Science and Technology, Luoyang 471003, China
E-mail: {xgzhang, wding}@njnet.edu.cn

Abstract—**Internet flows characteristics are important reference for network behaviors research. However, there are few latest related studies and comparative analysis on different networks is especially rare. Based on traffic traces from CAIDA and CERNET, this paper presents a detailed and comprehensive comparison on Internet flows characteristics, including the fine-grained distributions of lifetime, size, rate, life stage, port and protocol of flows. Our comparative research and conclusion can provide a latest data support for other studies of network behaviors, traffic classification, network performance and network security etc.**

Keywords-distribution; comparison; CAIDA; CERNET; flow characteristic

I. INTRODUCTION

Based on openness and good expansibility of Internet, adding new applications and expanding topological structures are very easy. As the rapid growth of network bandwidth and the sharp increase in users, all kinds of network applications spring up. And these applications are not just the traditional style that are based on words and pictures, but a mass of vivid and highly interactive network services, such as online video, audio, games, social networks and so on. Meanwhile, a wide variety of new applications protocols are continually appearing, for example new p2p mechanism etc. Moreover, a large number of new management strategies for network, such as excellent firewall policy, routing scheme and so on. All of those things are changing the network behaviors and making the traffic characteristics more and more complex.

In addition, Internet almost covers the global, and regional disparity is obvious in network applications and network behaviors. Because of different political, economic and cultural background, people have different hobbies, life and work routines etc. Moreover, popular applications spread from one part of the world to another over a period of time. These factors will cause some regional and time differences in network behaviors and traffic characteristics.

As we know, flow-based measurements of network traffic can not only gain the inner relationship between the packets and even the higher level information, but also are the needs of network behaviors observation, network optimizations and network security etc. And Internet flows characteristics have the important reference values for researches on strategies of network monitoring, network management and network accounting. Since 2000, some valuable study on Internet flows characteristics appeared [1] [2] [3] [4] [5]. However, with the great development of Internet, all kinds of new technologies, new applications are emerging one after another, the network behaviors and traffic characteristics have been changing. Whether today's Internet flows characteristics have some changes, whether there are some regional differences and what are the regional differences. Unfortunately, the latest research is very few. Aim at these questions, this paper presents a comparative analysis on Internet flows characteristics between CERNET and CAIDA.

The rest of the paper is structured as follows. In section II, we present the background and related works on Internet flows characteristics. In section III, we describe the datasets that we utilize for this study. In section IV, we compare Internet flows characteristics between CAIDA and CERNET. Finally in section V, we conclude our work and discuss possible future research directions.

II. RELATED WORK

Flow is the abstraction of network traffic. In the beginning, derives from different motivation, flow has different definition. Inspired by packet trains model defined by Jain and Routhier, Claffy et al [1] gave an influential flow model based on temporal and spatial locality of traffic and specified a flow is a set of a sequence of packets that have a few same characteristics. This model provides convenient method for describing the network traffic characteristics.

Around 2000, flow-based network traffic characteristics research became a hot issue in the field of network traffic measurement, many research results on flow characteristics successively appeared. Fang, et al [2] researched the traffic characteristics of backbone network and revealed the important phenomenon of mice and elephant in the field of network streams. Brownlee et al [3] studied the stream lifetime and size distributions of network traffic on a campus OC12 link at UC San Diego and showed these distributions from minute to minute over an hour or more. Then Brownlee et al [4] characterized traffic distribution by use of the method for measuring the size and lifetime of Internet streams proposed by them and revealed another important phenomenon of dragonflies and tortoises on network streams. Lan et al [5] studied the correlations of attributes of Internet flows based on traffic of Los Nettos and NLANR and classified the flows as cheetahs or snails by flow rate. After, related works aimed at utilizing flow characteristics to solve the particular problems, such as Zhou et al [6] studied a dynamical timeout strategy of flows for CERNET based on

2165-5006/12 $26.00 © 2012 IEEE
DOI 10.1109/ICNDC.2012.35

IEEE computer society

flow rate, and a recent hot issue of flow-based traffic classifications [7] [8] [9], etc.

However, these works are mostly based on one certain network and there are few comparative researches among different networks, especially almost no comparative analysis on Internet flows characteristics between CERNET and CAIDA. So, this paper compares distributions of flows for them on size, lifetime, rate, port and protocol of flow. Our works are more detailed and comprehensive than previous works.

III. Traces

The traffic traces we analyze in this study are from two different sources. The first set of traces was provided by CAIDA. It was collected by the equinix-chicago Internet data collection monitor that is located at an Equinix datacenter in Chicago, IL, and is connected to a backbone link of a Tier1 ISP between Chicago, IL and Seattle, WA, and prior to December 2010 this was an OC192 link and currently it is a 10GigE link. This set of traces contains a number of trace files and these files have been labeled A (Seattle to Chicago) and B (Chicago to Seattle). The second set of traces was provided by CERNET. It was collected from the main channel of CERNET in Jiangsu with passive measurement methods which uses flow sampling methods to collect data with the default sampling ratio 1/4. This link covers over 100 universities and high schools and its bandwidth has been updated to 10G from 2.5G since 2006. Similarly, the second set of traces consists of a number of trace files and these files have been labeled In (enter into the Jiangsu network) and Out (leave the Jiangsu network).

TABLE I. Traffic traces information

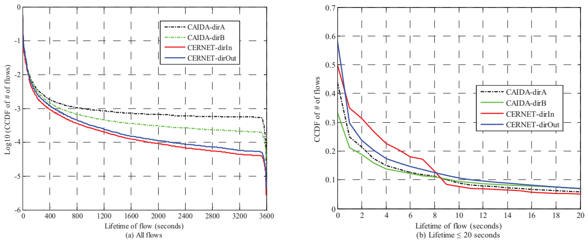
No.	Source	Date	Local time	Duration	Size	# of flows	# of packets
1	CAIDA	March 24, 2011	8:00-9:00	1 hour	115GB	107,229,781	2,166,660,064
2	CERNET	March 11, 2011	8:00-9:00	1 hour	35GB	58,420,427	548,868,210

IV. Flow Characteristics Comparison

In general, a flow is identified as a unidirectional stream of packets between a given source and destination—both defined by a network-layer IP address and a transport-layer port number. In this paper, a flow is the combination of five key fields (source IP address, destination IP address, source port number, destination port number and transport-layer protocal) and satisfies timeout strategy of 64 second. Specifically, TCP flows which have reached the end of byte stream (FIN) or which have been reset (RST) will be terminated.

A. *Flow Lifetime*

Lifetime is the duration time of a flow, i.e. arrival time of last packet minus the arrival time of the first packet. The arrival time is measured in microsecond. For our traces cover 1 hour, we use time granularity of 1 second and set 3600 intervals as (0s,1s),[1s,2s),..., [3599s,3600s], specially add [0s,0s] for the flows that lifetime is 0 second. We calculate the number of flows for all time slots for CAIDA and CERNET and then plot in Figure 1.



(a) All flows

(b) Lifetime ≤ 20 seconds

Figure 1. Distribution of flow lifetime

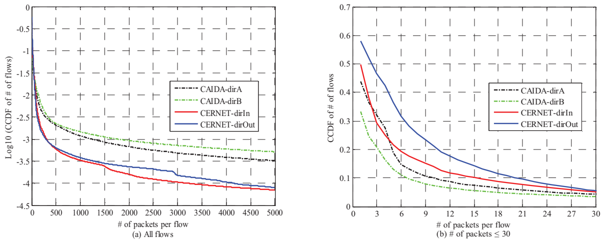
From the measurement results of flow lifetime, we can find that the proportion of same lifetime flows of CAIDA is generally higher than CERNET, and flow lifetime distributes more evenly. It is obvious that lifetime of 80% of CERNET flows have not longer than 6 seconds, while 80% of CAIDA flows have lifetime of 3 seconds or less. Specially, the proportion of flows that lifetime equals zero second, CAIDA is larger than CERNET.

B. *Flow Size*

The number of packets, bytes and mean length of packets of a flow are the main attributes of flow size. These attributes are not only the important metrics for traffic classification [7] [8] [9], but also useful for identifying the heavy hitter flows and network traffic charging.

We measure the packets distribution of flows by setting 5001 counters, if the number of packets of a flow is among 1 to 5000, the corresponding counter add 1 and the counter 5001 record the number of flows that the number of packets is bigger than 5000. Similarly, we measure the bytes distribution of flows use granularity of 500 bytes by setting 5001 intervals as [0*500B, 1*500B), [1*500B, 2*500B), ..., [5000*500B, ∞). In the same way, we set 1500 counters correspond to mean packet length among 1 to 1500 bytes and add counter 1501 record the number of flows that mean packet length are bigger than 1500 bytes.

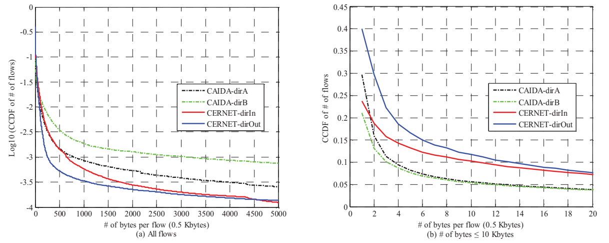
We can find that the proportion of same size flows of CAIDA is generally higher than CERNET, and flow size distributes more evenly. It is obvious that 80% of CERNET flows' packet number is not larger than 10 packets and 90% of CERNET flows' byte number is not larger than 7 Kbytes, while 80% of CAIDA flow' packet number is not larger than 5 packets and 90% of CAIDA flows' byte number is not larger than 2 Kbytes. Specially, the proportion of single packet flows, CAIDA is larger than CERNET. From the mean length of packets data, we can find that 90% of CAIDA flows' packet length is not more than 350 bytes while CERNET flows' packet length is not more than 650 bytes. Both flows sizes and mean packet length of CAIDA distribute more evenly than CERNET.



(a) All flows

(b) # of packets ≤ 30

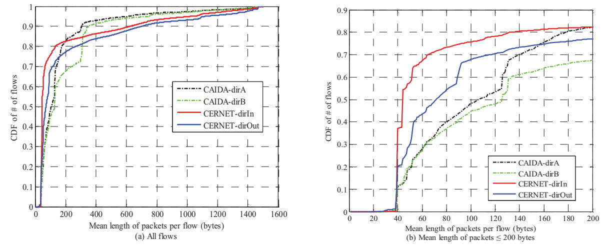
Figure 2. Distribution of flow size measured in packets



(a) All flows

(b) # of bytes ≤ 10 Kbytes

Figure 3. Distribution of flow size measured in bytes



(a) All flows

(b) Mean length of packets ≤ 200 bytes

Figure 4. Distribution of mean length of packets per flow

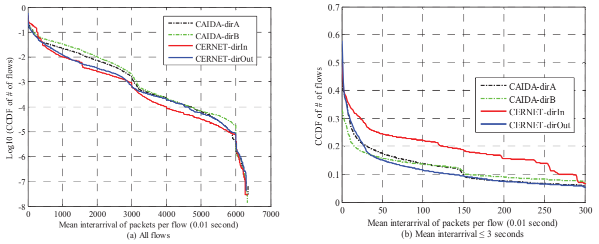
C. *Flow Rate*

Flow rate is an important metric of network performance. In general, we can measure flow rate in packets per second or bytes per second. In addition, interarrival of packets of a flow is another important metric of flow rate and its distribution can help to choose the suitable timeout strategy of flow [6]. We can obtain the mean interarrival of packets of a flow by using its lifetime divide by its number of packets. The lifetime is accurate to microsecond when we calculate the flow rate.

We measure the distribution of mean interarrival of packets of flows by using 6401 time intervals as [0.00s, 0.00s], (0.00s, 0.01s), [0.01s, 0.02s), ..., [63.99s, 64.00s]. If the interarrival value is in an interval, the number of flows of this interval adds 1. We calculate the number of flows of all time intervals for CAIDA and CERNET and then plot the distribution in Figure 5.

Meanwhile, we measure the distribution of flow rate in packets per second. We set 10001 counters correspond to the rate intervals as [0, 1), [1, 2), ..., [9999, 10000), [10000, ∞). If the rate value of a flow is in an interval, the number of flows of this interval adds 1. Similarly, we measure the distribution of flow rate in bytes per second by using 100001 counters correspond to the rate intervals as [0, 1), [1, 2), ..., [99999, 100000), [100000, ∞). Note that we only calculate the rate for flows that their lifetimes do not equal 0. We present the distribution of flow rate in Figure 6 and Figure 7 for CAIDA and CERNET.

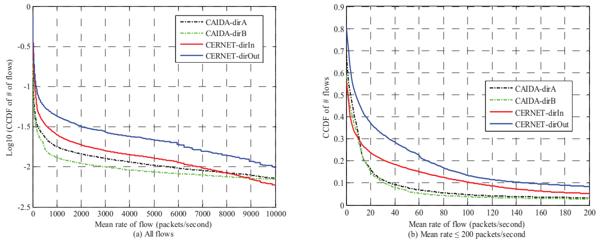
From flow rate data we can find that 80% of CAIDA flows' rate is not larger than 15 packets/second and not larger than 2000 bytes/second, while 80% of CERNET flows' rate in not larger than 70 packets/second and not larger than 8000 bytes/second. And we can find that the proportion of same rate flows of CERNET is generally higher than CAIDA, and distributes more evenly.



(a) All flows

(b) Mean interarrival ≤ 3 seconds

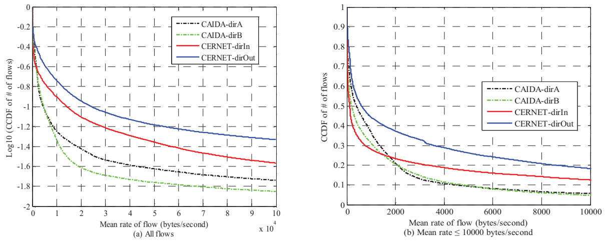
Figure 5. Distribution of mean interarrival of packets per flow



(a) All flows

(b) Mean rate ≤ 200 packets/second

Figure 6. Distribution of mean rate of flow measured in packets/second



(a) All flows

(b) Mean rate ≤ 10000 bytes/second

Figure 7. Distribution of mean rate of flow measured in bytes/second

D. *Flow Life Stage*

Every flow has a life cycle including three stages of starting, active and terminate. Corresponding to three life stages we define starting flow, active flow and terminate flow as follow. Starting flows of a time interval are the flows that are born in this time interval. Terminate flow of a time interval refer to the flows that are terminate in this time interval. Active flows of a time interval are all the flows that lifetimes overlay with this time interval. The number of these flows can provide a reference for network security [6], such as DDOS attack and network worm will increase the number of starting flow drastically and network fault may increase the number of terminate flows.

We use time granularity of 1 second and set 3600 time intervals as [0s, 1s), [1s, 2s), ..., [3599s, 3600s) to count life stage distributions of starting flows, active flows and terminate flows for CAIDA and CERNET. And we present the results in Figure 8, Figure 9 and Figure 10 separately.

From the data of flow life stage we can find that CAIDA flows distribute more evenly than CERNET whether starting flows, terminate flows or active flows.

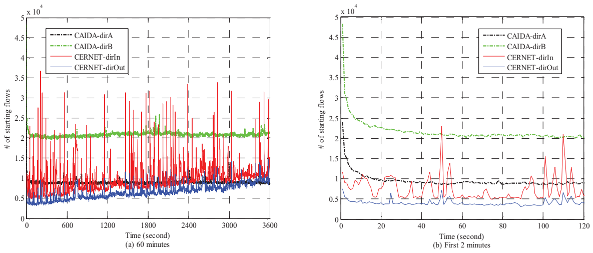


Figure 8. Distribution of starting flows per second

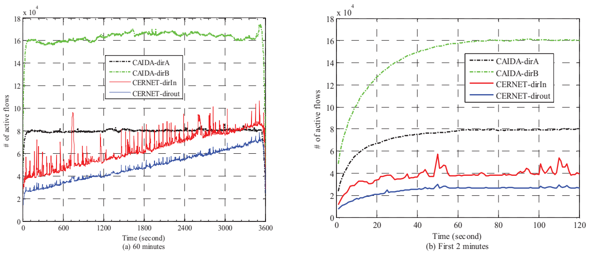


Figure 10. Distribution of active flows per second

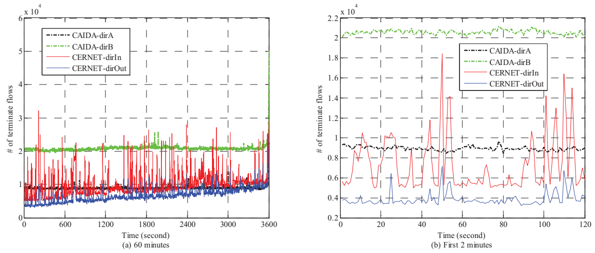


Figure 9. Distribution of terminate flows per second

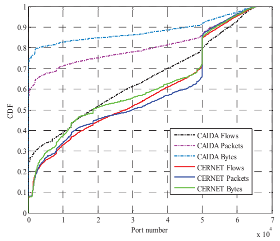


Figure 11. Distribution of ports of flows

E. Flow Protocal and Port

We calculate the number of flows, packets and bytes of every protocal, and then describe their distributions in TABLE II and TABLE III.

TABLE II. FLOW PROTOCAL DISTRIBUTION OF CAIDA AND CERNET

Protocal	CAIDA-dirA			CAIDA-dirB			CERNET-dirIn			CERNET-dirOut		
	Flows	Packets	Bytes	Flows	Packets	Bytes	Flows	Packets	Bytes	Flows	Packets	Bytes
ICMP	2.09%	0.29%	0.06%	2.11%	0.19%	0.02%	0.88%	0.42%	0.06%	1.05%	0.71%	0.12%
TCP	48.61%	84.30%	84.95%	39.01%	86.91%	93.90%	67.61%	53.01%	66.03%	39.22%	46.79%	46.69%
UDP	48.89%	14.53%	14.45%	58.72%	12.38%	5.56%	31.49%	46.51%	33.86%	59.68%	52.43%	53.16%
OTHERS	0.41%	0.88%	0.54%	0.16%	0.52%	0.52%	0.02%	0.06%	0.05%	0.05%	0.07%	0.03%

TABLE III. OTHERS PROTOCALS LIST OF CAIDA AND CERNET

Protocal	CAIDA-dirA	CAIDA-dirB	CERNET-dirIn	CERNET-dirOut
OTHERS	HOPOPT, IPv6, RSVP, GRE, ESP	HOPOPT, IPv6, RSVP, GRE, ESP, IP, AH	IGMP, DCN-MEAS, MERIT-INP, DCCP, IPv6, IPv6-Frag, RSVP, GRE, OSPFIGP, PIM, SM, PTP, Unassigned (158\211\226\233\242), Reserved	IPv6, GRE, ESP

TABLE IV. TRAFFIC PROPORTIONS OF DIFFERENT PORT CLASSES

Port Classes	Proportion of Flows		Proportion of Packets		Proportion of Bytes	
	CAIDA	CERNET	CAIDA	CERNET	CAIDA	CERNET
WKP-WKP	0.17%	0.39%	0.17%	0.06%	0.07%	0.01%
WKP-NWKP	44.49%	54.00%	71.25%	41.46%	82.49%	46.83%
NWKP-NWKP	55.34%	45.61%	28.58%	58.48%	17.44%	53.16%

Usually, network ports are divided into well-known ports (0-1023), registered ports (1024-49151) and dynamic ports (49152-65535). However, in many systems dynamic ports and registered ports are not distinguished strictly, so we divide ports into WKP short for well known ports and NWKP short for non-well-known ports. Distribution of flows based on ports can help to design traffic classification algorithms and suggest some especial ports that should be focused on.

We calculate the number of flows, packets and bytes of every port, and then plot their distributions in Figure 11. Meanwhile, we calculate the number of flows, packets and bytes of different port classes and then describe the proportions in TABLE IV.

From the flow protocal distribution we find that protocal was used irregularly of CERNET-dirIn and some packets whose length greater than 1500 bytes were detected. Form flow port distribution we see flow proportion between well-known ports is very small, this will affect the precision of traffic classification algorithm.

V. Conclusions

From our measured data we can find that with the great development of Internet all kinds of new technologies, new applications spring up, Internet flows characteristics have some new changes. And our comparison also shows some regional differences exist in different networks. And it's not hard to imagine an algorithm that is suitable for one network does not always work well in another network. In the future, we will complete more comparisons on different traces of more sites and our ultimate goal is comparative research on the evolution of Internet flows characteristics.

Acknowledgment

We thank the anonymous reviewers for their valuable comments. This work is supported by the National Key Technology Support Program of China under Grant No. 2008BAH37B04 and the State Key Development Program for Basic Research of China under Grant No. 2009CB320505.

References

[1] K. C. Claffy, H. W. Braun and G. C. Polyzos, "A parameterizable methodology for Internet traffic flow profiling," Selected Areas in Communications, vol. 13, no.8, pp.1481-1494, Oct. 1995.

[2] W. Fang and L. Peterson, "Inter-As traffic patterns and their implications," Proc. GLOBECOM'99, Dec. 1999, pp.1859-1868.

[3] N. Brownlee and K. C. Claffy, "Internet Stream Size Distributions," Proc. SIGMETRICS 2002, Jun. 2002, pp. 282-283.

[4] N. Brownlee and K. C. Claffy, "Understanding internet traffic streams: Dragonflies and tortoises," IEEE Communications Magazine, vol. 40, no. 10, pp.110-117, Oct. 2002.

[5] K. C. Lan and J. Heidemann, "A measurement study of correlations of Internet flow characteristics," Computer Networks, vol. 50, no. 1, pp.46-62, Jan. 2006.

[6] M. Z. Zhou, J. Gong, and W. Ding, "High-Speed Network Flows' Dynamical Timeout Strategy Based on Flow Rate Metrics," Journal of Software, vol. 17, no.10, pp.2141−2151, Oct. 2006.

[7] A. W. Moore and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques," Proc. SIGMETRICS 2005, Jun. 2005, pp.50-60.

[8] L. Chen and J. Gong, "Fast Application-level Traffic Classification using NetFlow Records," Journal on Communications, vol. 33, no. 1, pp.145-152, Jan. 2012.

[9] A. Dainotti, A. Pescape, and K. C. Claffy, "Issues and future directions in traffic classification," IEEE Network, vol. 26, no. 1, pp.35-40, Jan.-Feb. 2012.