Packet Size Distribution Tendencies in Computer Network Flows

Eimantas Garšva, Nerijus Paulauskas, Gediminas Gražulevičius
Department of Computer Engineering, Faculty of Electronics
Vilnius Gediminas Technical University
Vilnius, Lithuania
Email: {eimantas.garsva|nerijus.paulauskas|gediminas.grazulevicius}@vgtu.lt

***Abstract*—Network flows are easy to get and simple to store network activity data. The challenge is to interpret them efficiently from security and network engineering standpoint as payload and application layer protocol specific information is missing. The article presents the statistical analysis of network flows with the emphasis on packet size distribution. Existing packet size distribution researches were reviewed. Packet size distribution Cumulative Distribution Functions (CDFs) were produced from existing academic computer network data. The CDFs for protocols TCP, UDP, ICMP and popular application layer protocols (HTTP, DNS) were analysed. Network traffic statistics were further visualized using radar graph. Article provides reusable statistical analysis steps and statistical trends for academic computer network.**

Keywords—computer network; packet size distribution; network traffic; NetFlow

I. Introduction

The distribution of packet size in a computer network is one of the main network traffic characteristics which are expected to contribute in determining network usage trends and the normal network state. Network flows are simple to get and to store and that is especially important in high speed computer networks. Effective usage is challenging due to a missing packet payload which provides precise information about the application layer protocol and the query itself. This research aims to provide the findings of network packet size statistical analysis that are useful for traffic analysis.

General analysis of the statistical network data stream properties has shown that the packet size distribution is tri-modal [1]. This was a result of a combination of 40 B size TCP acknowledgments and the existence of distinct default Maximum Transmission Unit sizes, which depend on used applications and network protocols. Internet Protocol version 4 requires network hosts to process the packets of at least 576 bytes and Ethernet MTU is 1500 B. The prominent sizes have varied over the years, as new applications emerge, new protocols replaced the old ones, and the protocol behavior was changing. Subsequent studies [2]–[4] found that the distribution of packet sizes changed from tri-modal to bi-modal 40 B and 1500 B. The following researches found a strong mode around 1300 B as this is the MTU recommended by Virtual Private Network vendors in order not to exceed Ethernet MTU after VPN encapsulation.

Packet size heavily depends on the application. Identification and classification of network applications using data mining and pattern recognition methods [5]–[7] confirm that computer networks, including the Internet, use a wide range of packet-based applications. Traditionally, applications are identified by inspection of the destination port number in packet header. Newer and real-time applications cannot always be detected by such a simple investigation and hence other techniques such as deep packet analysis have been developed. Deep packet analysis however has significant problems such as its inability to operate on encrypted data packets, and its need to capture specific packets from the traffic stream. Alternative approach to the detection of real-time applications is making a statistical fingerprint derivable from the observable traffic streams generated by such applications. This has been found to be the packet size distribution of the application for a range of such applications and network conditions.

This paper presents the research of packet size distribution in network flows and is the progress of the statistical analysis of the NetFlow data in the computer network of Electronics faculty of Vilnius Gediminas Technical University addressed in [8]. All data flows were collected by the main router Cisco 6506 (Supervisor Engine 720, Netflow version 5) during the October of 2013. We consider only working days when network load is the highest. Collected data consists of 27.6×10^6 TCP, 29.6×10^6 UDP and 3.9×10^6 ICMP flows, total of 1.4 GB of raw data. NetFlow data was processed using Nfdump tool (version 1.6.4). The computer network has 253 public IP addresses and connects more than 350 computers including 5 mail servers, 7 web servers and 1 ftp server. IP version 4 is used in the network.

II. Packet Size Distribution

A network flow is a unidirectional sequence of packets between the communicating source and destination. The session between two network hosts is divided in two flows: incoming and outgoing. One Netflow record contains information about the source and destination IP addresses, the source and destination ports, the start and finish timestamps, and the number of bytes and packets transferred. New NetFlow record is produced from the same IP traffic information every 30 minutes when communication is active and the production of the NetFlow is ended after 15 s if

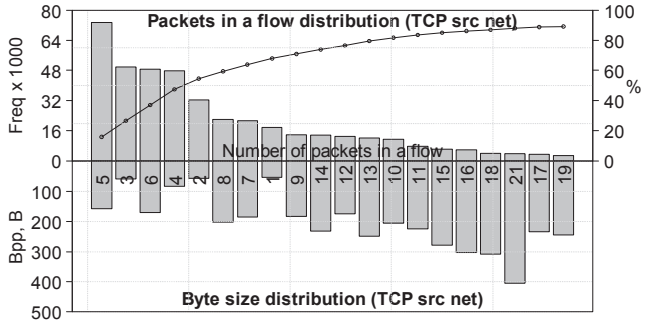
978-1-4673-7445-3/15/$31.00 ©2015 IEEE

inactive or when RST or FIN flag is set, so most likely all the packets from the same communication will fall into the same network flow, but sometimes the communication lasts longer, e.g. when keep-alive packets are used. Internet protocol data is being analysed in the article, so transport layer protocol data units like segments and datagrams are in IP packets thus called TCP packets and similar.

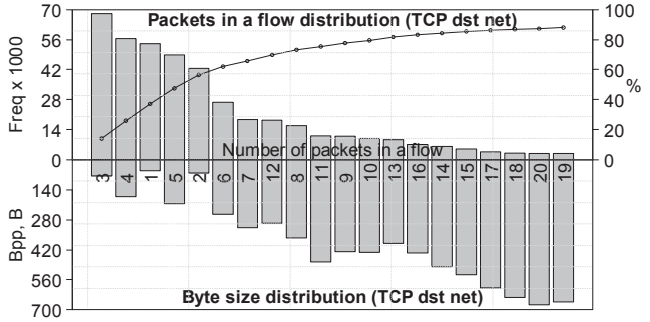
Statistics for transport layer protocols and both traffic directions were graphed in order to determine the number of packets in a flow and how it is related to the size of the packet. Frequency of the flows containing a specific number of packets is the number of times such flow was monitored during the research period of 1 month and averaged for a workday. The frequencies of flows with a specified number of packets paired with an average size of such a packet described in bytes per packet *Bpp* as well as empirical Cumulative Distribution Function – CDF of the flow frequency are presented in Fig. 1 according to the flow direction and the protocol.

The research is centered on the faculty computer network, so when the traffic is originating from the network it is the source network (src net) Fig. 1(a) and when the traffic is destined for it is the destination network (dst net) Fig. 1(b).

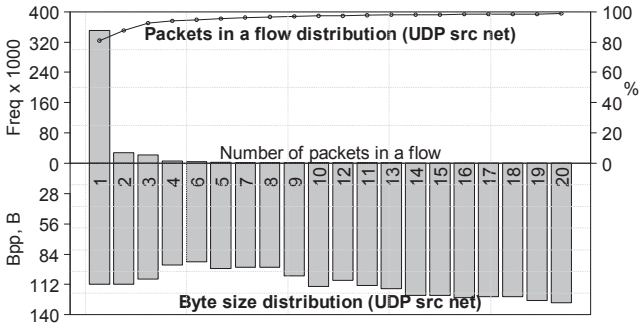
Most of TCP flows consist of a small number of packets (< 10), but the packet size is bigger of those with the higher number of packets in a flow (> 15). Computer network is not heavily loaded (3.1 Mbps of traffic on average) and most of the traffic is queries and replies with low amount of data. Packets bigger in size are used when higher amount of data needs to be sent and their number in a flow is higher. Packets of the incoming Fig. 1(b) traffic are bigger, but the number of packets in the flow is not.



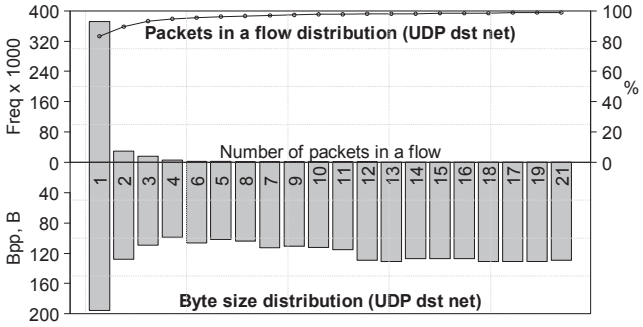
(a) TCP source network



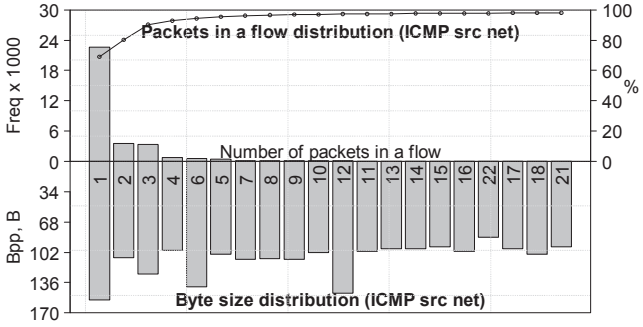
(b) TCP destination network



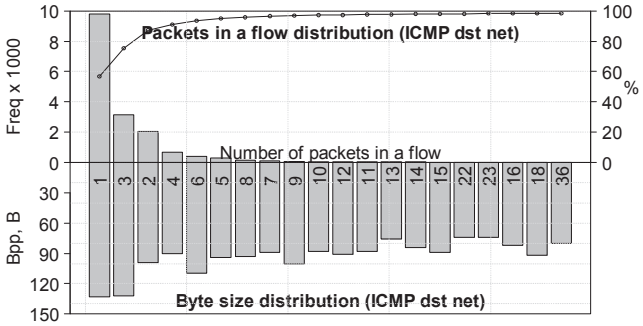
(c) UDP source network



(d) UDP destination network



(e) ICMP source network



(f) ICMP destination network

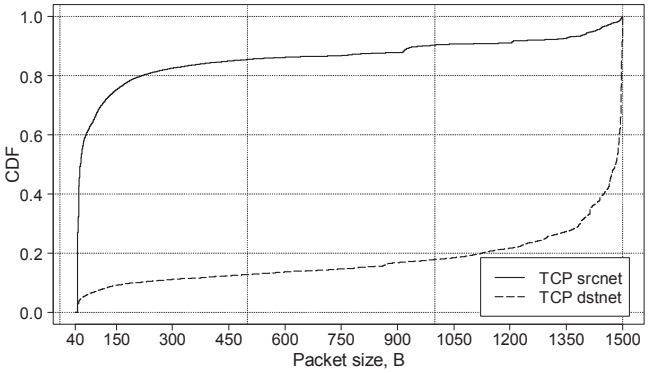
Fig. 1. Frequency of the flows containing a specific number of packets and its CDF, both paired with bytes per packet.

Both UDP and ICMP statistics look similar but significantly differ from the TCP. Most UDP and ICMP flows (average 85 % for UDP and 75 % for ICMP) consists of a single packet and then the size of the packet is slightly bigger than average.

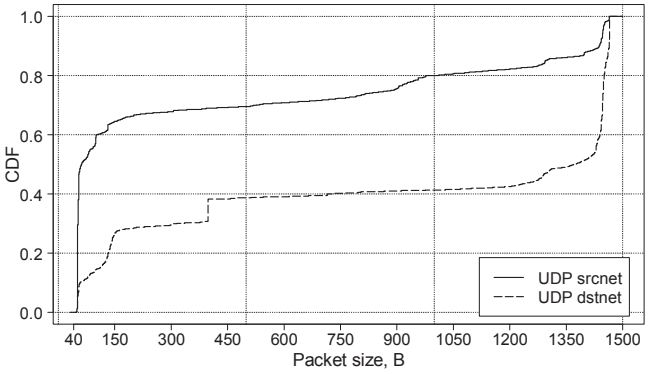
CDF of the TCP grows more consecutively than CDF of

UDP and ICMP and reaches only 90 %, due to the fact that flows with other number of packets exist and despite the fact that the number of flows containing particular number of packets is very low, the variety of such flows is quite high to form 10 %. CDF of UDP and ICMP rise quickly and are further linear (Fig. 1). UDP is used to provide different network services and send data, so more variety was expected, but ICMP primarily used to send informational messages has less traffic therefore the values are lower.

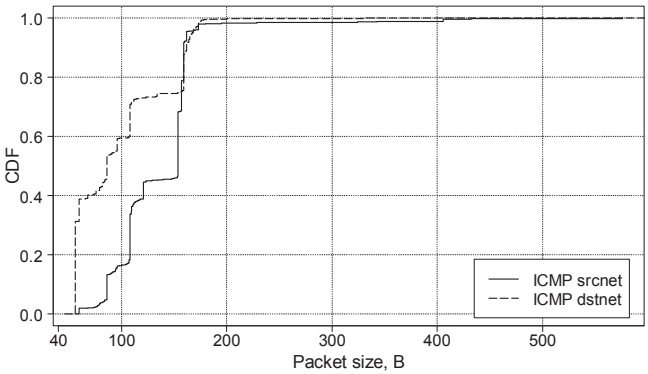
TCP, UDP and ICMP packet size distribution CDF graphs are presented in Fig. 2.



(a) TCP packet size distribution



(b) UDP packet size distribution



(c) ICMP packet size distribution

Fig. 2. Packet size distribution CDF for TCP, UDP and ICMP.

Packet size depicted on horizontal axis has values up to Ethernet MTU, except for ICMP (Fig. 2(c)) as the number of packets higher than 600 B do not introduce significant change to CDF. Average packet size differs for incoming and outgoing traffic. The biggest difference is for TCP packets as average outgoing packet is 247 B and incoming packet is 1245 B. UDP outgoing packets are 414 B in average and the incoming packets – 920 B. The difference is least significant for ICMP as outgoing packets are 136 B and incoming packets are 99 B on averages.

Incoming traffic TCP (Fig. 2(a)-dstnet) and UDP (Fig. 2(b)-dstnet) packets are bigger in size and in number than outgoing ones, except small packets which are more likely to outgoing traffic. 68 % of all TCP outgoing data (Fig. 2(a)-srcnet) is comprised of packets smaller in size than 100 B due to the fact that smaller queries are sent out. TCP incoming traffic with responses has 60 % of packets larger than 1450 B. UDP case is similar to TCP with less expressed polarity for queries and responses (Fig.2(b)). The obtained results proved that distribution of packet sizes in modern computer network is bi-modal. The same results show studies of Center for Applied Internet Data Analysis.

High number of maximal Ethernet MTU packets is monitored despite the fact that TCP flow usually consists of more than a single packet and the packet size is averaged as network flows do not provide the information about the individual packet. If all the packets are similar in size the average value is more accurate. Session initiation is done with small packets and the accuracy is higher if a flow is comprised of small packets only. 80 % of UDP and ICMP flows contain a single packet therefore the packet size value is not impacted by the average value. ICMP protocol graphs are more similar to the application specific graphs as it is used to send the specific messages and encapsulates UDP datagram if needed. 99 % of ICMP packets fall into the 40B-200 B interval (Fig.2(c)).

Distribution is influenced by the available protocol features and its popularity: TCP has biggest variety of packet sizes as most of the applications are using it, UDP also has vast variety of packets and the application dependence is represented by the step-like structure of CDF. ICMP is limited in its application and is not usually used to transfer data, so the variety of packet sizes is the lowest.

III. HTTP AND DNS TRAFFIC STATISTICS

The most popular application layer protocol that uses TCP for transport is Hypertext Transfer Protocol – HTTP, it usually uses port 80 to deliver web content (it comprises 58 % of all TCP flows). Domain Name System protocol – DNS, operating on port 53 is used to resolve numeric addresses from symbolic names was most popular for UDP (it comprises 26.7 % of all UDP flows). DNS using TCP for the transport is not addressed in this research. Traffic which belongs to these protocols was determined solely by the port numbers (Table I).

Web traffic IP packet size is limited by the Ethernet Maximum Transmission Unit, which is 1500 bytes, despite the

fact, that the maximum IPv4 packet size is 65535 bytes. DNS message size using UDP is restricted to 512 bytes, maximum IP header length is 60 bytes and UDP header is 8 bytes and Ethernet frame header as well as trailing Cyclical Redundancy Check are not accounted in a flow, so maximal DNS packet size when UDP is used – 580 bytes.

TABLE I. HTTP AND DNS TRAFFIC SUMMARY

	HTTP				DNS			
Transport	TCP				UDP			
Port no.	80				53			
Max. theor. packet size, B	1500				580			
Direction	**Incoming**		**Outgoing**		**Incoming**		**Outgoing**	
Port type	**Src.**	**Dst.**	**Src.**	**Dst.**	**Src.**	**Dst.**	**Src.**	**Dst.**
Size, %	50.1	1.2	1.2	47.5	49.7	0.4	0.1	49.8
Packet no., %	63.7	1.1	2.0	33.2	49.1	0.9	0.6	49.5
Avg. packet size, B	1373	65	1403	101	209	141	70	66
Avg. packet no. in a flow	63.2	47.6	81.4	34.7	1.1	2.3	10.2	1.1

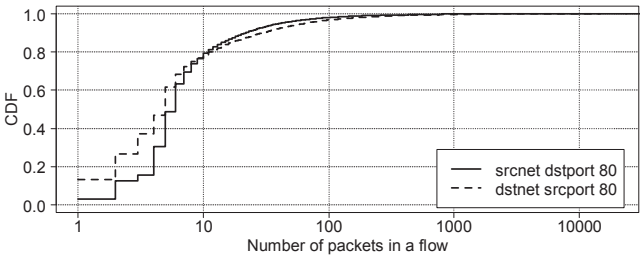
Table I data confirms, that the network addressed in this research (faculty network) is the user of the network services and not the provider: incoming HTTP and DNS traffic comprise the biggest part of downloaded TCP and UDP data, the number and size of the packets is the biggest. There is 630 times more HTTP traffic than DNS.

Content is usually requested from the external networks, but it can be requested from the faculty network too. The query from the faculty network is represented by outgoing traffic to the analyzed destination port (srcnet dstport) and the response to the query sent is the incoming traffic where the analyzed port is a source port (dstnet srcport). When the content is hosted on the faculty network and is requested from the outside networks the query is incoming traffic to the analyzed port as destination (dstnet dstport) and the response is outgoing traffic from the analyzed port as source (dstnet srcport).

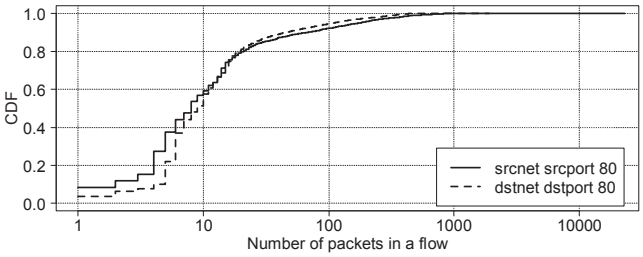
HTTP responses received from the external networks are 26 times bigger than the queries sent. HTTP responses sent from the faculty network are 39 times bigger than the requests received from the external networks and network flow has the biggest number of packets in it on average – 81. Faculty network provides some content intense HTTP services to the external networks and received HTTP content is less intense and more distributed due to the fact that there is 31 times more of incoming content than the served one.

DNS responses received from the external network are 3 times bigger than the requests, but the DNS queries coming from the external networks comprise 3 times bigger amount of data than the replies. One query packet is usually answered with a single reply packet. Situation with the incoming DNS queries is not normal and shows that there is some misconfiguration and some DNS server records point to the non-existent DNS servers on the faculty network. Also, DNS replies from the faculty network have the unusually high average number of packets in a flow (10), this can indicate that some computer in a faculty network is infected with malware and the covert communication channel is set over DNS port.

Statistical data is averaged for a workday of the one month research period data and presented as Cumulative Distribution Function graphs in Fig. 3 and Fig. 4. CDFs show the percentage of flows with a defined number of packets in it.

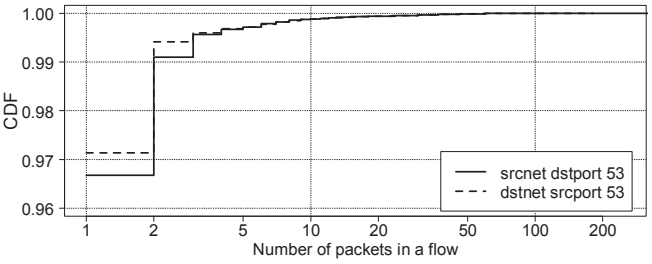


(a)

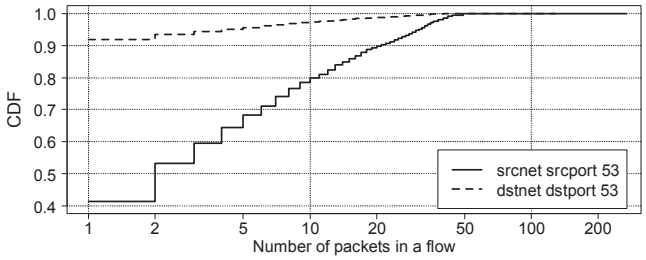


(b)

Fig. 3. Distributios of packets number in a flow for HTTP traffic.



(a)



(b)

Fig. 4. Distributios of packets number in a flow for DNS traffic.

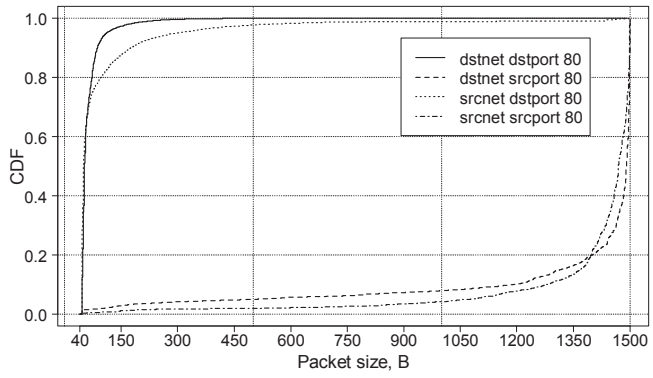
HTTP data is graphed on Fig. 3 and shows that flows have less than 10 packets for 80 % of requested content (Fig. 3(a)) and 60 % of provided content (Fig. 3(b)). Outgoing HTTP traffic (Fig. 3(b)) where port is the source port has big amount of packets and those packets are big, the size of a packet is bigger when the amount of packets in a flow is higher. This traffic is the response of the internal web servers to the queries

of the external users. Incoming HTTP traffic where HTTP port is destination port has small amount of packets, packets themselves are small. This traffic originating from external clients is destined to internally hosted web servers. Outgoing queries receive more responses with less than 10 packets in a flow than incoming ones. This shows that the content hosted internally is denser: documents are hosted instead of advertisements.

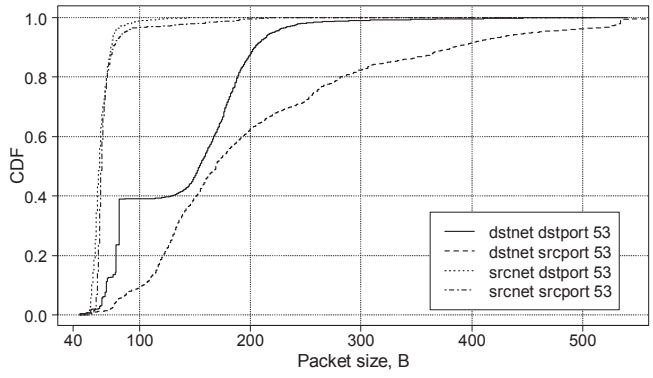
Outgoing DNS queries (Fig. 4(a)-srcnet dstport) are more frequent than the incoming ones (Fig. 4(b)-dstnet dstport) as well as DNS replies. Graph shows that outgoing DNS queries and replies have small packets and incoming queries and replies have more than 4 times bigger packets (Fig. 5(b)). For incoming queries this is an anomaly.

Flows with 5 packets on average are the most popular for HTTP despite the fact that overall there is 57 packets in a flow on average due to the smaller number of flows containing high number of packets. DNS flows usually have one packet in a flow, despite of the destination and the number of packets increases evenly when the flows frequency decreases.

Packet size distribution curves (Fig. 5) show more network specific information as they depend on the hosted content. Application specific traffic depends on the application and the content hosted. HTTP queries (dstport 80) tend to have small packets (70 % of packets are smaller than 64 B) and responses are distributed with a slight rise when the packets get bigger (80 % of all packets are bigger than 1400 B) (Fig. 5(a)).



(a) HTTP packet size distribution



(b) DNS packet size distribution

Fig. 5. Packet size distribution CDF for HTTP and DNS.

TCP tendencies are true for HTTP there is high number of small packets and ones with the size close to maximal Ethernet MTU.

DNS packets are smaller than HTTP ones (Fig. 5(b)). The number of internally hosted DNS zones is low and the incoming traffic is low, that is the cause of the distinct curve of the incoming DNS queries (dstnet dstport 53) where smaller than 83 B packets comprise 38 % and 97 % of outgoing response (srcnet srcport 53) packets are smaller than 110 B.

IV. Network Traffic Statistics Visualization

In order to aggregate NetFlow records as bidirectional flows we write own scripts using R programming language as aggregation function in nfdump is not very accurate. When source and destination ports are > 1024 or < 1024, the flows are taken as is in nfdump. In our case we sorted all flows by time and the flow which appears first is treated as originated from client and its pair (if exists) is treated as flow from server. Flow with higher source port number is placed first in the case where appearance times of two or more flows are identical.

R programming language is best to process vectorized information and is slow when all the entries in the array are accessed individually. This was problematic because the aggregation function uses indexing and aggregation in R is slow. Alternatively Perl script was prepared to process the data, it was faster than R, but not fast enough. Final NetFlow data processing tool is based on R with the data aggregation function written in C++. The tool is able to process the data of the day in less than 7 minutes on midrange office workstation. The performance example of NetFlow statistical processing is shown in Table II.

Table II. NetFlow Processing Performance

N_F in a file	Perl script	R script	R script with C++ function
21453	14.612 s	489.754 s	2.885 s

ICMP messages were mostly the result of earlier queries using other protocol, so ICMP flows were accounted analyzing type and code of the flow and only outgoing traffic was known. Only the traffic where local client (LC) was communicating with remote server (RS) and remote client (RC) was communicating to local server (LS) was monitored.

Network traffic statistical data analysis needed for traffic engineering and possible attack vector analysis can benefit from the visual network traffic graphs. Radar graph with a traffic property on the axis (Fig. 6) outlines their values and gives better visual information allowing the comparison for the communication parties.

There is more incoming traffic presented by number of packets and the size of the packet than outgoing. The number of unique hosts N_{UH} can represent clients or servers, the amount of traffic in bytes is divided based on direction to incoming – N_{Bin} and outgoing – N_{Bout}, as well as the number of packets: N_{Pin} – incoming and N_{Pout} – outgoing.

100 local TCP clients were communicating to 170 times more of remote servers. There are 300 times more UDP

remote clients than local ones and 700 more remote servers than local ones. Local clients get 3.6 times more content that the remote ones.

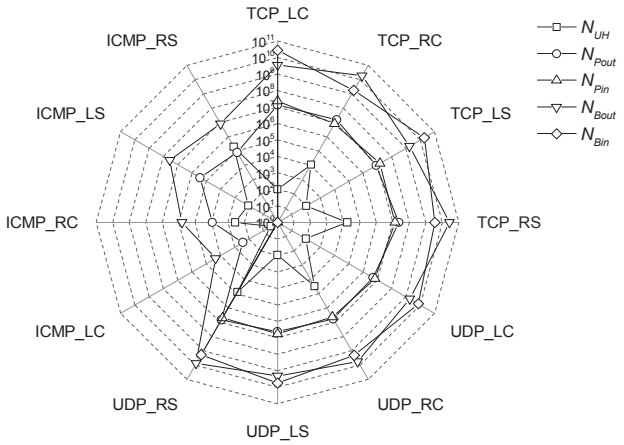


Fig. 6. Network traffic statistic visualization.

ICMP brings even more information about the network. Incoming Echo request queries had most of the packets, while for the response most of the packets were for Host unreachable. Largest packets were for Fragmentation Needed and Don't Fragment was Set message. Most of the packets per flow as well as destination IPs had incoming Echo Replies.

Port Unreachable messages were most popular for outgoing responses and the single type of outgoing queries was Echo request. Biggest packets were for outgoing Protocol Unreachable messages.

ICMP has the information from the packet which caused the response encapsulated, so the size of the message is impacted by it directly.

V. Conclusions

TCP has biggest variety of packet sizes as most of the applications are using it, UDP also has a variety of packets and application dependence is represented by the step-like structure of CDF. ICMP is limited in its application and is not usually used to transfer data, so the variety of packet sizes is the lowest.

Flows with 5 and 3 packets on average for outgoing and incoming cases respectively are the most popular for TCP protocol. 71 % of outgoing flows consists of less than 10 packets with average 123 B sizes. 64 % of incoming flows consists of less than 8 packets with average 154 B sizes.

UDP and ICMP statistics look similar: most of the flows (average 85 % for UDP and 75 % for ICMP) consist of a single packet and the size of the packet is bigger than average. Packets of the incoming traffic are bigger, but the number of packets in the flow is not.

References

[1] C. Shannon, D. Moore, and K. C. Claffy, "Beyond Folklore: Observations on Fragmented Traffic," *IEEE/ACM Trans. Networking*, vol. 10, no. 6, pp. 709–720, Dec. 2002.

[2] P. Hurtig, W. John, and A. Brunstrom, "Recent Trends in TCP Packet-Level Characteristics," in *Proc. of the 7th International Conference on Networking and Services (ICNS 2011)*, Venice/Mestre, Italy, May. 2011, pp. 49–56.

[3] D. Murray and T. Koziniec, "The State of Enterprise Network Traffic in 2012," in *Proc. of the 18th Asia-Pacific Conference on Communications (APCC)*, Jeju Island, South Korea, 2012, pp. 179–184.

[4] R. Sinha, C. Papadopoulos, and J. Heidemann, "Internet Packet Size Distributions: Some Observations," *Technical Report ISI-TR-2007-643*, USC/Information Sciences Institute, May. 2007.

[5] D. J. Parish, K. Bharadia, A. Larkum, I. W. Phillips, and M. A. Oliver, "Using Packet Size Distribution to Identify Real-time Networked Applications," in *Proc. of the IEEE Communications*, vol. 150, Aug. 2003, pp. 221–227.

[6] Ying-Dar Lin, Chun-Nan Lu, Yuan-Cheng Lai, Wei-Hao Peng, and Po-Ching Lin, "Application Classification Using Packet Size Distribution and Port Association," *Journal of Network and Computer Applications*, vol. 32, pp. 1023–1030, Sept. 2009.

[7] Chun-Nan Lu, Ying-Dar Lin, Chun-Ying Huang, and Yuan-Cheng Lai, "Session Level Flow Classification by Packet Size Distribution and Session Grouping," in *Proc. of the 26th International Conference on Advanced Information Networking and Applications (AINA)*, Fukuoka, Japan, 2012, pp. 221–226.

[8] E. Garšva, N. Paulauskas, G. Gražulevičius, and L. Gulbinovič, "Packet Inter-arrival Time Distribution in Academic Computer Network", *Elektronika ir elektrotechnika (Electronics and Electrical Engineering)*, no. 3, 2014, pp. 87–90.