International Journal of Business and Technology

Volume 6
Issue 3 *Spring 2018*

Article 3

May 2018

Characteristics and Temporal Behavior of Internet Backbone Traffic

Artan Salihu
University for Business and Technology, artan.salihu@ubt-uni.net

Muharrem Shefkiu
Kosovo Telecom, muharrem.shefkiu@kosovotelecom.com

Arianit Maraj
Kosovo Telecom, arianit.maraj@kosovotelecom.com

Characteristics and Temporal Behavior of Internet Backbone Traffic

Artan Salihu[1], Muharrem Shefkiu[2], Arianit Maraj[2]

[1]UBT – Higher Education Institution, Prishtina, Kosovo
[2]Telecom of Kosovo
artan.salihu@ubt-uni.net[1],
{muharrem.shefkiu, arianit.maraj}@kosovotelecom.com [2]

Abstract. With the rapid increase demand for data usage, Internet has become complex and harder to analyze. Characterizing the Internet traffic might reveal information that are important for Mobile Network Operators (MNOs) to formulate policy decisions, develop techniques to detect network anomalies, help better provision network resources (capacity, buffers) and use workload characteristics for simulations (typical packet sizes, flow durations, common protocols).
In this paper, using passive monitoring and measurements, we show collected data traffic at Internet backbone routers. First, we reveal main observations on patterns and characteristics of this dataset including packet sizes, traffic volume for inter and intra domain and protocol composition. Second, we further investigate independence structure of packet size arrivals using both visual and computational statistics. Finally, we show the temporal behavior of most active destination IP and Port addresses.

Keywords—Internet, traffic, dataset, characteristics, self-similarity, Network Operator.

Introduction

While Internet network complexity continuous to grow and become harder to analyze, contemporary study is required in order to evaluate its underlying structure, protocol composition and temporal behavior. Thus wise, validity of techniques and tools used to manage, solve and avoid network failures can be assessed as well. In general, measurements and analysis of Internet traffic are done at packet or flow level. In this paper, we report on traffic measurements and characteristics of Internet traffic at flow level. We characterize traffic flows over 24 hours and 7 days' time scales in terms of flow volume, packet size arrival, protocol composition, and distribution structure of top port numbers and IP addresses.
A vast amount of work is done in analyzing Internet traffic traces collected at different parts of network. Some of the most seminal work that reported on wide-area network traffic include those in [1] and [2]. Caceres in [1] is well-known for reporting that TCP is responsible for 90% of traffic volume while Thompson et al [2], in addition to introducing a deployment methodology for traffic monitoring, show some important factors on packet size, inter and intra domain traffic share and application usage. Our work is inspired from study in [2], yet it is different in a sense that Thompson et al collected traces in a network dominated by ATM and SONET technologies, a very different application usage (user-behavior) and different bandwidth requirements too. Some other important and recent studies rely on measurements reported by Center of Applied Internet Data Analysis (CAIDA) [3]. CAIDA datasets are a great resource but they contain anonymized passive traffic traces from monitors on specific Internet

backbone-links belonging to more than one service provider which generate very high traffic volumes. Other datasets that have been subject of analysis include Waikato traces [4], LAN Sigcom [5], Berkley dataset[6]. In contrast, our goal is to have a broad picture of network traffic characteristics from single Internet service provider that shares all main properties, yet is smaller in scale and represents a different demographic part of the world. In addition to general backbone traffic characteristics, we utilize our data traffic traces to gain deeper understanding of Internet structure. More specifically, we have initial results for our future work where we look at traffic *bursiness* which was initially introduced by seminal work of Leland et al [7]. Authors in [7] showed that traffic is bursty in different timescales and it cannot be described using few parameters by Poisson processes. Internet and typical voice calls used in circuit switching have dramatically different statistical characteristics from each other. Internet sessions tend to be much more variable and longer in duration than voice calls [8]. Therefore, using self-similar processes as a notion to better understand and model data traffic is vital for Internet traffic-engineering. This effects traffic-engineering in two very fundamental ways; if traffic was bursty, then in order to manage the inevitable peaks that exceed the planned capacity, very sophisticated buffers and packet scheduling would be required. On the other hand, if aggregated traffic is smooth, guaranteeing QoS would be only a function of long-term capacity planning because there would be no queue buildups [9].
In the next section we have presented the data collection and monitoring, section 3 describes the data analyses we used. Metrics and results are shown in section 4 while the conclusions are drawn in section 5.

Traffic Monitoring and Data Collection

In this paper, we have collected traffic traces from a Service Provider, Telecom of Kosovo (TK). Traffic flows were captured using nfcapd (Netflow capture daemon) where the machine collecting the data listens to netflow packets. IP Flow feature and NetFlow protocol was enabled in Internet Border Gateways (IBGs). Two IBG routers are responsible for routing the whole traffic as shown in Figure 1.

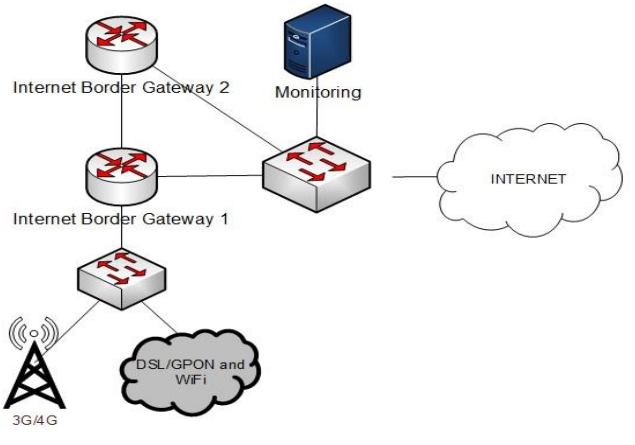

Figure 1. TK Network and measurement infrastructure to collect data from all regions in Kosovo

DOI: 10.33107/ijbte.2018.6.3.03

This means that flow collector captures all conversations between machines at TK network and outside it. We have been collecting data for a period of approximately three months. All these series of packets that share same quintet, or a flow, is saved on a IBGs either until buffer/cache is filled or every five-minute interval. All the files stored in local folders can be read with nfdump using its syntax and Boolean expressions to filter the data. For example, the command: nfdump -r nfcapd.201612291520 "proto tcp" which selects all the flows in the file nfcapd.201612291520 and prints out ones that use TCP protocol and other information showing the timestamp, duration of the session in seconds, protocol, source IP address and the port number, destination IP address and port number, number of packets sent during this session, bytes and the flows.
General definition of an IP Flow is a quintet made up of source and destination IP address, source and destination port number and the protocol. In Table 1 a summary information from one flow is depicted that can be found in a typical row of NetFlow.For the purpose of this paper, not all the fields are used.

Table 1: Information from one flow

Field	Description	Bytes
%ts	Unix start time of the flow	4
%te	Unix end time of the flow	4
%sa	IPv4 source address where flow is originated from	4
%da	IPv4 destination address where flow is destined to	4
%sp	IPv4 source port where is originated from	2
%dp	IPv4 destination port where flow is destined to	2
%pr	Protocols up to Layer 4: UDP, TCP, ICMP, ESP, etc.	1
%flg	TCP flag	1
%tos	IP type of service	1
%ipkt	Number of packets in one session or flow	4
%ibyt	Number of bytes for one session or flow	4
%in	Interface number where incoming flow is processed by	4
%out	Interface number where outgoing flow is processed by	4
%smk	Source subnet prefix based on BGP router's table	1
%dmk	Destination subnet prefix based on BGP router's table	1
%sas	Autonomous System where flow was originated from	4
%das	Autonomous System where flow is destined to	4
%nh	Next destination address	4
Total	Total Byte per Typical Flow	53

Data Analysis

It is important to know general characteristics of the data set such as distribution of packet sizes, traffic volume for inter and intra domain and what protocols are present and their proportion. In this Section, main characteristics of the dataset that are used for the purpose of this study are shown in Table 2. One can see an overview of the processed daily data that are used for analysis. In total we have worked with 31,602,250 flows, when aggregated. They generated $1{,}4752 \cdot 10^{11}$ packets during seven-day interval. In addition, it shows the average number of packets per second, average bytes per packet as well as the number of unique connections during busy hours (11 AM to 3PM).

Table 4. Characteristics of Dataset used in this study

Days	Number of Flows	Number of Packets	Average Packets per second	Average Bytes per packets	Unique Connections during busy hour (Upstream/Downstream)
Day1	9,467,477	25,259,807,842	1947494	1068	160911 / 1864484
Day2	4,261,459	20,526,364,391	2174049	1054	167047 / 1945073
Day3	2,929,059	16,161,126,666	1496402	986	334513 / 3585167
Day4	4,164,621	24,641,601,294	2085229	1058	227875 / 2705763
Day5	4,150,023	23,951,432,184	2094708	1064	212447 / 2277633
Day6	2,799,257	14,815,754,013	1371831	1065	183080 / 2122497
Day7	3,830,354	22,163,825,619	2052211	1043	235895 / 1705700
	31,602,250	147,519,912,009	13221924	1047	

Figure 2 shows the number of packets that arrive each minute for 24 hours for the local traffic (intra-domain). To filter only local traffic, knowledge about subnetworks from IP-Plan is used.

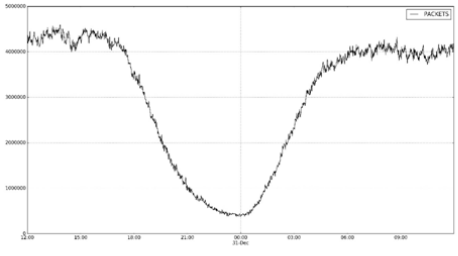

Figure 2. Local Packet Traffic

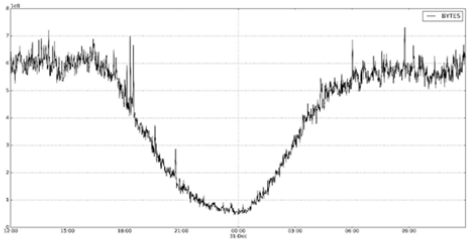

Figure 3. Local Byte Traffic

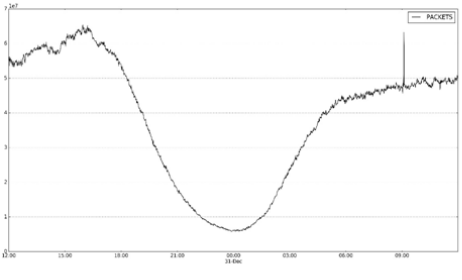

Figure 4. International Downstream Packets

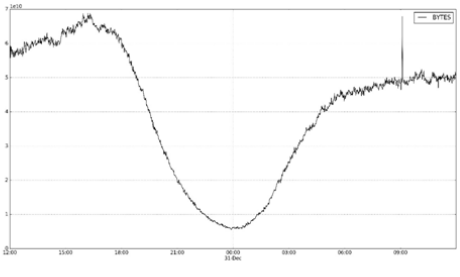

Figure 5. International Downstream Bytes

As presented in Figure 2 the maximum of 4598000 packets in one minute is at 13:51 and minimum of 372000 packets at 23:59. The maximum number of bytes is 667001000 bytes at 08:47 which surprisingly is not contained by the maximum number of packets during this interval but corresponds to 4321000 packets. This suggests that despite the fewer packet arrivals at 08:47, they are greater in size. The minimum number of bytes is 45597000 at 23:47. The local byte traffic during this interval is shown in Figure 3.

While the local traffic does not exceed more than 0.6 GB in one minute, downstream traffic toward users within TK network reaches a peak of 68.85 GB in one minute which corresponds to time interval around 16:19 and minimum of 0.045 GB at 23:47. One can observe that the peak interval is not the same for local and international downstream traffic. In addition, if graphs for local traffic and downstream international traffic are compared (Fig 3 versus Fig 4

DOI: 10.33107/ijbte.2018.6.3.03

and 5), it can be observed that international downstream traffic is much smoother than local traffic when aggregated at one minute.
Another way of expressing the load in a more common unit is by counting or computing bits per second or packets per second within interval of the interest. One can read the statistics about bits and packets per second during this interval. In this case, a 2-hour interval was taken into consideration which corresponds to the peak hours. It shows that average bits in one second is 5347214353 (around 5.3 Gbps) and in average 607143 packets in one second.

Protocols

The table 3 presents statistics about the proportion IP protocols which appears in the data collected during a one-day interval, it shows six present protocols ordered by the volume (bytes). TCP continues to be the transport layer protocol that dominates the traffic with approximately 93.3% of the bytes and 82% of the packets. Nearly 6.5% of bytes and 17.8% of packets in one-day interval are UDP. TCP together with UDP comprise 99.9% of bytes and 99.8% of packets in TK network. Note that other IP protocols, appear with negligible probability.

Table 5. IP Protocols Proportion

Proto	Bytes (%)	Packets (%)	pps	bps	bpp
TCP	48.7 T(93.3)	42.6 G(82.0)	492493	4.5 G	1144
UDP	3.5 T(6.6)	9.2 G(17.8)	106691	320.1 M	375
ESP	33.0 G(0.1)	70.4 M(0.1)	814	3.1 M	468
ICMP	5.4 G(0.0)	55.1 M(0.1)	637	502554	98
GRE	4.2 G(0.0)	8.9 M(0.0)	103	387946	469
IPv6	24.2 M(0.0)	53000(0.0)	1	3692	457

Packet size

The frequency of packet sizes resulted from a dataset extracted between 15:00 and 18:00 is shown in the Figure 6. The smallest and largest packet sizes are 21 bytes and 1500 bytes respectively. The ten most frequent packet sizes are 1430, 1450, 32 52, 1472, 1492, 21, 40, 1480 and 1500.

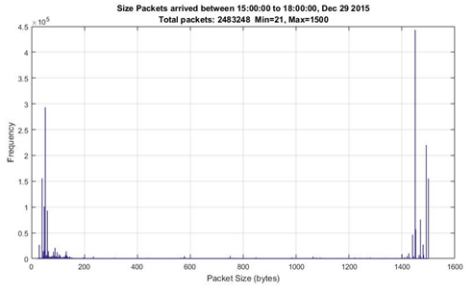

Figure 6. Frequency of Packet Sizes in PTK Network

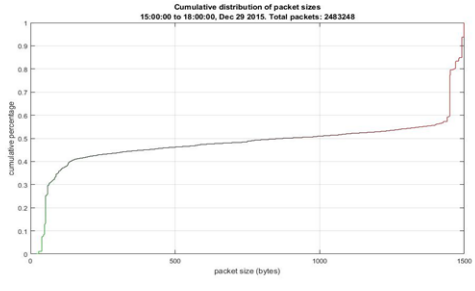

Figure 7. Empirical Cumulative Distribution of Packet Sizes

Cumulative distribution of packet sizes is shown in Figure 7. It shows that 30% of packets are small and less than 100 bytes, most of the packet sizes are distributed between 100 and 1400 and almost 30% of packets are between 1430 and 1500.

Metrics and results

One of the most important goals of this analysis was to check the temporal properties of the network under study. We show that distribution of most frequent port numbers and IP addresses do not change over time. To do so, we evaluate daily measurements for most active destination IP addresses and port numbers for the upstream traffic and calculate parameter α where data can fit an exponential form of $y \sim C x^{-\alpha}$. We show that the number of days have very little impact in the distribution of most visited port numbers and IP addresses. Certainly, our calculations are for only seven days, but we have daily observations from a very rich number of sources and we do not expect that it changes if we calculate for longer period. Instead variation is more related to daily periodicity rather than number of days itself.

Parameter α during 7-day observations for most visited port numbers is $\alpha = 1.9420 \pm 0.0592$ with 90% confidence level, using t-test. And, α for most visited IP addresses is $\alpha = 0.6961 \pm 0.0627$. It would be interesting to see for spatial diversity and variation of parameter α. A study in [11] claims that distribution of port numbers and IP addresses varies from place to place. We have shown the fitted data to the exponential distribution of the form discussed above and have checked the accuracy using:Visual inspection of distribution fit and Coefficient of determination r^2.

Visual inspection of distribution fit: Figures8 and Figure 9 show the best-fit parameter together with least square in logarithmic form and cumulative distribution. In log-log plot we can observe a piecewise linear line, which is an indication of heavy-tailed distribution.

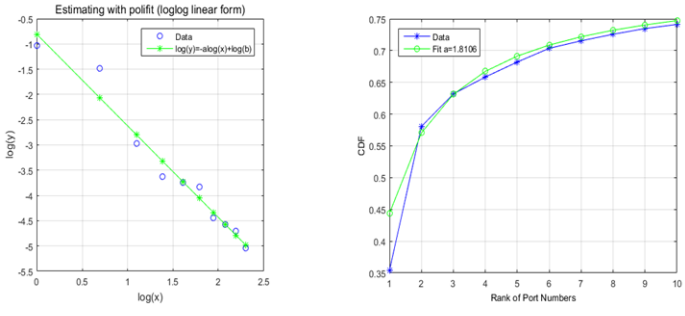

Figure 8. Cumulative distribution function of most popular port numbers

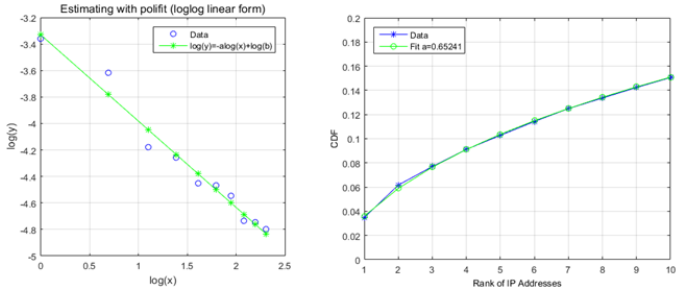

Figure 9. Cumulative distribution function of most popular IP addresses for upstream traffic in one day.

DOI: 10.33107/ijbte.2018.6.3.03

Coefficient of determination r^2*:* Coefficient of determination measures the proportion of variation explained by the independent variable in a regression model and is given by $r^2 = \frac{S_{\hat{y}}^2}{S_{\bar{y}}^2}$where $S_{\hat{y}}^2$ is the variance of the predicted values and $S_{\bar{y}}^2$ is the variance of the calculated values from the measurements. For one day measurements, coefficient of determination is $r^2 = 0.9657$ for port numbers and$r^2 = 0.9235$ for IP addresses.

Conclusions

In this paper, we have reported on data collected and shown general characteristics of internet traffic captured at TK. We show that peak intervals for byte counts are not symmetric for local and international downstream traffic and despite the fewer packet arrivals they are greater in size. This suggests that traffic volume for packet counts is negatively correlated with the size of packets. Also, when aggregating traffic at 1-minute granularity, we see that international downstream traffic becomes smoother and thus can be easily modeled using traditional *Teletraffic* models.

TCP continues to be the transport layer protocol that dominates the traffic with approximately 93.3% of the bytes and 82% of the packets. Similar results are reported from studies in [2] and [10]. The smallest and largest packet sizes are 21 bytes and 1500 bytes respectively. Few packets are detected with size larger than 1500 bytes in size and 30% of packets are between 1430 and 1500, indicating typical Ethernet MTU implementation.

Finally, we show that distribution of most frequent port numbers and IP addresses do not change over time. Top 10 IP addresses and Port numbers are stable and do not change during the course of 7 days in terms of distribution. Instead variation is more related to daily periodicity rather than number of days itself. When fitted to a power-law distribution of an exponential form, one can use $\alpha = 1.9420 \pm 0.0592$ with 90% confidence level in order to model such network characteristics.

References

1. R. Caceres, “Measurements of Wide Area Internet Traffic”, UCB/CSD, Univ. CA, Berkley, Dec, 1989.
2. K. Thompson, G. J. Miller and R. Wilder, “Wide-Area Internet Traffic Patterns and Characteristics”, IEEE Network, Nov/Dec 1997, pp.10-23
3. The Cooperative Association for Internet Data Analysis - CAIDA, [Online]. Available: http://www.caida.org. [Accessed 10 2017].
4. WITS: Waikato Internet Traffic Storage, [Online]. Available: https://wand.net.nz/wits/. [Accessed 10 2017]
5. Wireless LAN Traces from ACM SIGCOMM'01, [Online]. Available: http://www.sysnet.ucsd.edu/pawn/sigcomm-trace/. [Accessed 10 2017]
6. Internet Traffic Archive, [Online]. Available: http://ita.ee.lbl.gov/html/traces.html. [Accessed 10 2017]
7. W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "On the Self-Similar nature," IEEE/ACM Transactions o Networking, 1994.
8. K. Park and W. Willinger, "Self-similar network traffic: An overview," 2000. [Online]. Available: https://www.cs.purdue.edu/nsl/intro-ss-chap.pdf. [Accessed 4 2016].

9. A. Salihu, “Internet Traffic and Topology Characteristics From a National ISP Perspective”, Master thesis, School Of Information Sciences, University of Pittsburgh, 2016
10. S. McCreary and K. Claffy, “Trends in wide area IP traffic patterns - A view from Ames Internet Exchange”, ITC Specialist Seminar, Monterey, CA, Sep 2000.
11. J. L. Garcia-Dorado, J. A. Hernandez, J. Aracil, J. E. d. Vergara, F. J. Monserrat and T. P. d. M. E. Robles, "On the duration and spatial characteristics of internet traffic measurement experiments," IEEE/ACM Communications Magazine, vol 46, no. 11, 2008.

DOI: 10.33107/ijbte.2018.6.3.03