International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 4 Issue 7, July-2015

A Comprehensive Analysis on Internet Traffic Measurement in Information Technology

Laxmi Devi
MCA, M.Phill, Global Open University, Nagaland, India

ABSTRACT

Measurement and analysis of real traffic is important to gain knowledge about the characteristics of the traffic. Without measurement, it is impossible to build realistic traffic models. It is recent that data traffic was found to have self-similar properties. In this thesis work traffic captured on the network at SICS and on the Supernet, is shown to have this fractal-like behaviour. The traffic is also examined with respect to which protocols and packet sizes are present and in what proportions. In the SICS trace most packets are small, TCP is shown to be the predominant transport protocol and NNTP the most common application. In contrast to this, large UDP packets sent between not well-known ports dominates the Supernet traffic. Finally, characteristics of the client side of the WWW traffic are examined more closely. In order to extract useful information from the packet trace, web browsers use of TCP and HTTP is investigated including new features in HTTP/1.1 such as persistent connections and pipelining. Empirical probability distributions are derived describing session lengths, time between user clicks and the amount of data transferred due to a single user click.

Keywords: Traffic measurement, self-similarity.

INTRODUCTION

Measurement and analysis of real traffic is important to gain knowledge about the characteristics of the traffic. Without measurement, it is impossible to build realistic theoretical traffic models. The traditional telephone network could very successfully be analyzed and modelled using applied mathematics such as stochastic processes. Especially Poisson processes have been used which states that call arrivals are mutually independent and that the call interracial times are all exponentially distributed, with one and the same parameter . Because of the success of voice network modelling and because Poisson processes have some attractive theoretical properties, the same approach have often been used when modelling data network traffic. Packet and connection arrivals have been assumed to be Poisson processes. But several studies have shown that the distribution of packet inter arrivals clearly differs from exponential and Leland et al. Showed that the burstiness on many timescales, observed in real traffic, cannot be described with traditional Poisson-based traffic modeling. Instead they introduced statistically self-similar processes as a better way of modeling LAN traffic. In this thesis work, network traffic on the Supernet and external traffic at SICS is analyzed. The traffic was captured using tcpdump. Figure 1 shows the network at SICS.

The machine running tcp dump (called network monitor in the figure) was listening to the 100 Mbit/sec line connecting all workstations at SICS with the gateway and in the end the SUNET network. This was used to capture all conversations between machines at SICS and the outside Internet world, for 24 hours. The packet trace was taken between 21:36 990414 and 21:40 990415 and includes more than 21 million packets.

BACKGROUND

IP based traffic has dominated the fixed Internet for basically its whole lifetime and it has also been adapted to the mobile Internet. On the other hand, mobile Internet access requires enhanced features from the access network. Thus, data transmission in mobile networks requires also proprietary protocols used with Internet traffic, and measurement traffic at different points provide also different trace data. First, this chapter gives overview on IP traffic characteristics. Second, mobile network architecture is described from a mobile Internet point of view.

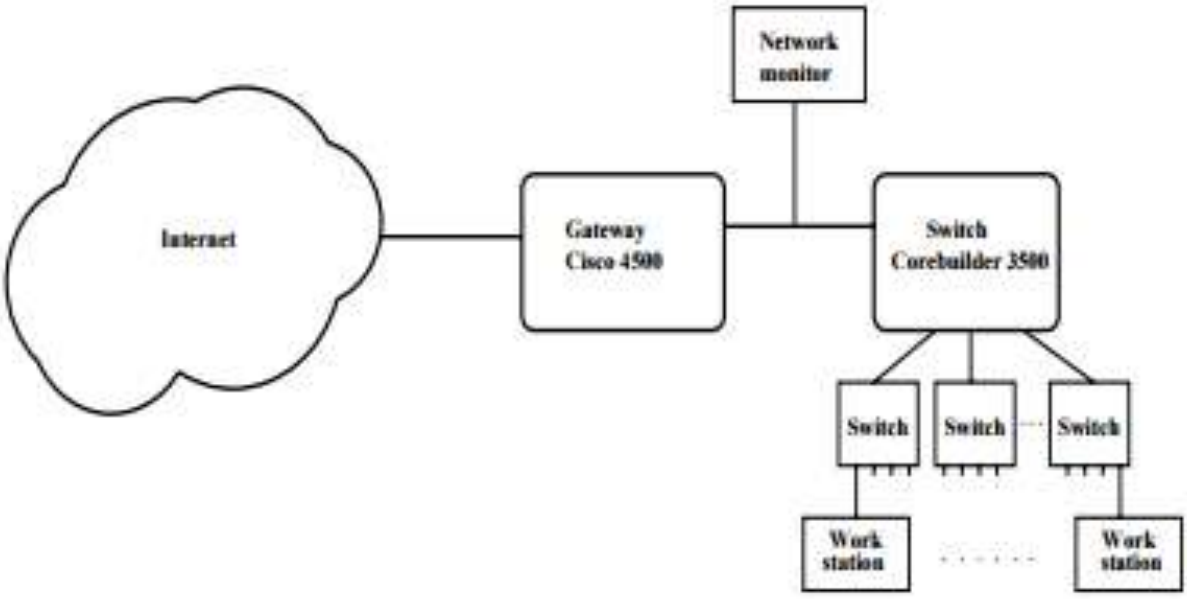

Figure 1. The network at SICS

IP Traffic

The objective of IP is to enable an interconnection between heterogeneous networks and their interoperation (Comer, 2000). IP is based on a layered model where multiple protocols handle different level of connectivity between two hosts.

Protocol Layers

TCP/IP suite describes the different layers of Internet traffic. The layers from bottom-up are the physical layer, data-link layer, network layer, transport layer, and application layer, each including their own protocols that are used for the interconnectivity (Forouzan, 2002). The protocols in different layers encapsulate and decapsulate the traffic when it is being transmitted from a sender to a receiver. Encapsulation means that at the sender side a lower level protocol takes a higher level protocol packet and without changing it adds a standardized header of its own to the beginning. In decapsulation at the receiver end the headers at each layer are removed and the original application layer packet stays untouched. The TCP/IP suite and exemplary protocols are presented in Figure 2. The figure also demonstrates the encapsulation and decapsulation process.

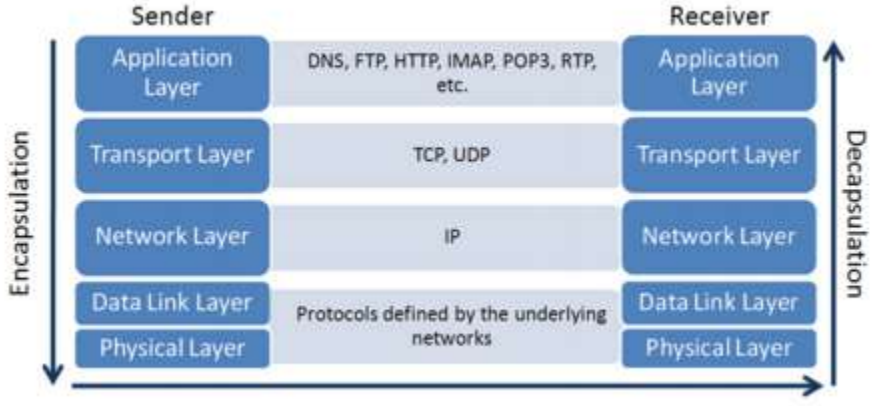

Figure 2: Internet model and encapsulation of traffic (adapted from Forouzan et al., 2002)

On the data-link and physical layers there are no specific protocols defined. These layers support the proprietary protocols of the underlying networks and take care of the link and physical level connectivity.

The IP on the network layer handles the host-to-host traffic by mainly offering the addressing for it. It uses IP-addresses, which are fixed length of 32 bits, and transmits the data unreliably over one or several networks in IP-datagrams. Unreliability means that the protocol does not include error checking or tracking of the datagrams, for instance. As the maximum packet sizes in different network links may vary, the IP datagram may have to be split into smaller packets, which is why IP protocol also offers fragmentation and reassembly (Postel, 1981a).

Structural levels of traffic

In addition to the different layers of traffic protocols, there are three principal structural levels of traffic in the Internet (Crovella & Krishnamurthy, 2006). From lowest to highest they are the packet level, train or flow level, and session level, as presented in Figure 3.

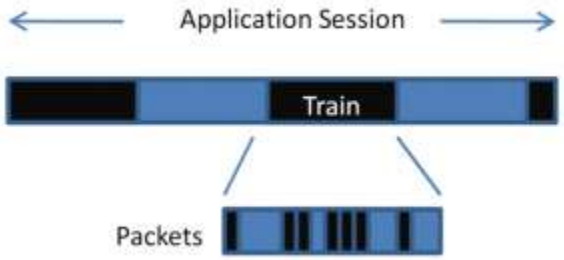

Figure 3: Levels of structure in Internet traffic (adapted from Crovella & Krishnamurthy, 2006)

The lowest level includes packets that are either transferred in the network or not. The second level includes trains of packets, often called as flows, as rather rarely only one packet is sent between a source and destination. A flow or a packet train can, for example, represent a download of one file or a web page. At its simplest a packet train or a flow is a burst of packets from a certain source arriving to a certain destination. If a timeout between two packets exceeds a certain interval, the packets belong to a different packet train or flow. One of the most popular ways of defining a flow is a five-tuple, which includes source and destination IP addresses, source and destination port numbers, and the protocol number.

The highest structural level of IP traffic is the session level which describes a single execution of an application by including a set of flows. Often traffic in the Internet has a clear beginning and an end that correspond with the usage time that a human is spending with the application. However, the increased P2P (peer-to-peer) traffic, for instance, has changed this behavior during the last years, as many P2P applications are used also while the end user himself or herself is inactive. In these cases the application session times last longer and do not represent a user session anymore.

INTERNET MEASUREMENT METHODOLOGIES

The most common way to classify traffic measurement methods is to distinguish between active and passive approaches. Active measurement involves injection of traffic into the network in order to probe certain network devices (e.g. PING) or to measure network properties such as round-trip-times (RTT) (e.g. traceroute). Pure observation of network traffic, referred to as passive measurement, is non-intrusive and does not change the existing traffic. Network traffic is tapped at a specific location and can then be recorded and processed at different levels of granularity, from complete packet-level traces to statistical figures. Even though active measurement offers some possibilities that passive approaches cannot provide, in this theses only passive measurement is considered, since it is best suitable for analysis of Internet backbone traffic properties. Passive traffic measurement methods can be further divided into software-based and hardware-based approaches. Software-based tools modify operating systems and device drivers on network hosts in order to obtain copies of network packets (e.g. BSD packet filter). While this approach is inexpensive and offers good adaptability, its possibilities to measure traffic on high speed networks are limited. In contrast, hardware-based methods are designed specifically for collection and processing of network traffic on high speed links such as an Internet backbone. Special traffic acquisition hardware is used to collect traffic directly on the physical links (e.g. by using optical splitters) or on network interfaces (e.g. mirrored router ports). Since highly specialized, such equipment is rather expensive and offers limited versatility.

The measurements described in this thesis are performed by the use of optical splitters feeding Endace DAG cards, currently the most common capture cards for high-speed network measurements on the market. Data gathered on different

protocol layers can present different levels of granularity. The most coarse granularity is provided by cumulated traffic summaries and statistics, such as packet counts or data volumes, as typically provided by SNMP. Another common practice is to condense network data into network flows. A flow can be described as a sequence of packets exchanged between common endpoints, defined by certain fields within network and transport headers. Instead of recording each individual packet, flow records are stored, containing relevant information about the specific flow. Such flow records can be unidirectional, as in the case of Net Flow, or bidirectional, as used in Papers II-IV included into this thesis. The most fine grained level of granularity is provided by packet level traces. Packet-level traces can include all information of each packet observed on a specific host or link.

Finally, packet-level network traces can be stored in different trace formats. Unfortunately, there is no standardized trace format, so developers of trace collection tools historically defined their own trace formats. The most popular trace format, especially common for traces from local area networks (LAN), is the PCAP format, the format of the BSD Packet Filter and TCP dump. For traces of wide area networks (WAN), an often used format was defined by Endace, the Endace record format (ERF), formerly also known as DAG format. The traces analyzed in this thesis have been recorded in ERF format. Other trace formats seen in the Internet measurement community include CAIDA's CORAL Reef format CRL or NLANR's formats FR, FR+ and TSH. This diverseness in trace formats introduces some problems, since public available analysis tools usually do not recognize all of these formats, making conversion of traces from one format to another necessary. Even tools for direct conversion often do not exist, so it might be necessary to convert traces into PCAP format first, which can be seen as the de-facto standard. Thus almost all conversion tools are able to convert their own format to or from PCAP format. Conversion however is usually not without costs. Different timestamp conventions within the trace formats often lead to loss of timestamp precision, which should be considered when performing timing sensitive operations, such as merging of trace files, or calculation of packet delays or inter-arrival times.

INTERNET TRAFFIC MEASUREMENTS

Internet traffic measurements are conducted for a variety of reasons and objectives. They are often performed to characterize, to monitor, or to control the network. For a network operator, monitoring and controlling the network are essential parts of its operation and maintenance. On the other hand, longer time-scale traffic characterization is often related to network research activities. (Peuhkuri, 2003) These characterizations can bring both, technical and commercial information on how to run the network and what kind of traffic is transferred in the network. User behavior and usage characteristics interest many entities, including academic, regulative, as well as corporate entities, but these issues have not been studied academically as much as the more technical aspects. First in this chapter, different measurement methods are classified. Second, different user, device, and application identification approaches are discussed. Third, traffic measurement research in the mobile context is reviewed and measurement points in mobile networks are discussed. In addition, extractable information from each measurement point is described.

Measurement types

The measurement type or setup is mainly affected by the measurement objectives. As traffic measurements can be conducted in multiple parts of a network, the measured trace data has to be representative for the phenomenon that is studied. Thus, the choices of the measurement location and granularity of the data are important. Location describes the part of the network and the traffic that is measured. In general, measurements can be conducted at different nodes or links in the network and in single or multiple locations. On the other hand, granularity of the data, i.e., the level of detail, defines the metrics that can be used. (Peuhkuri, 2003) If Internet usage characteristics are measured, utilization metrics, such as byte count and application distribution, can be used. However, as traffic volume as such, for example, is not always the best metric to describe usage, new metrics may have to be created to match the objectives better. Time scale of a measurement means both, the scale of the measurement period, as well as the continuity of it. Some measurements are sample based by nature, whereas sometimes sampling is conducted to save resources, for instance. However, Smith et al. (2001) argue that web traffic measurements, for example, have to be at least hours long to capture fully the long tail of distribution.

Trace data & post-processing

The measurement analysis can be done both, online or offline (Williamson, 2001), affecting whether the result of the measurement is trace data for further analysis, or a direct report of the real time analysis. As the latter requires high processing power, the traffic is often recorded on-location and the analysis is made afterwards. From the measured packets different information can be recorded. In general, the measurement can record protocol headers, protocol payload, certain fields in the headers or payload, or simply all the data. Packet level data can be provided by, for instance, network and

transport level header measurements. One reason for recording only headers is the huge amounts of data traffic in high speed network links. However, from high bandwidth links, even the headers produce high amounts of trace data. Thus, one way of post-processing packet level traces is to aggregate them to a flow level to compress the amount of data. There are multiple free tools to do the aggregation. For instance, a tool called CoralReef3 3 Coral Reef: takes de facto pcap-format (Packet Capture) packet trace data as input and outputs five-tuple flows with a wanted timeout and an interval period. If application level traces are needed, also the transport protocol payload, in other words the application level data, is measured. As application level protocol headers may be of highly varying lengths, either only certain amount of the application level data is recorded from the beginning, i.e. the headers, or all data is saved. If real time processing is conducted, only the significant fields can be saved for later use, decreasing the amount of trace data.

Other Advanced Methods

In addition to the approaches presented above, also other advanced approaches exist, often combining different methods. One recent combinatorial approach presented by Canini et al. (2009) is called GTVS (Ground Truth Verification System) and has multiple iteration rounds with different methods. The identification is made on application level information, flow statistics, host-level connection statistics, host name information, and transport-layer behavior level, thus including basically all the aforementioned methods. Canini et al. (2009) have used their application identification method GTSV also for HTTP traffic classification, including 14 categories that are presented in Table 1. Motivation for this is the expanded role of HTTP protocol in other than plain web browsing use. The advantage of this method is the added information about web traffic, which often is interpreted as web browsing as analysis methods do not have the functionality to do more accurate identification. Different classes are identified with combinations of different HTTP header fields, such as user agent, host name, and content type. An interesting detail about the research by Canini et al. (2009) is that they have planned to publish the tool and the rule sets to be used by other researchers in the future.

Table 1: Web traffic classification categories (adapted from Canini et al., 2009)

Class	Activities
Web Browsing	visiting web pages using a web browser
Web App	applications via web interface: e.g. Java applets, web gadgets
Crawler	bots crawling web pages
File Download	file downloading over HTTP
Webmail	web based e-mail services
Advertising	advertisement on a web page or embedded in software
Multimedia	streaming media or viewing media files on web pages
SW update	software update over HTTP
News Feeds	RSS feeds
Link Validator	automated link validators
Calendar	calendar application based on web, e.g., ical, gcal
Attack	malicious traffic over HTTP
IM	Windows Live Messenger
Monitoring	network monitoring

In addition to general application identification, also more specific research on certain application categories has been conducted, concentrating rather heavily on P2P applications. For instance, Svoboda et al. (2009) have used a cross layer method in VoIP (Voice over IP) classification, Karagiannis et al. (2004) used transport level information and the first 16 bytes from the payload to identify P2P applications, and Sen et al. (2004) used an application level P2P classifier to identify five common P2P applications.

RELATED WORK ON PASSIVE INTERNET MEASUREMENT

Even though large-scale Internet measurement is still rare, there has been a significant amount of effort expended on different Internet measurement activities in recent years. These activities include development of active and passive measurement methodologies and tools, targeting aspects such as network performance, traffic classification and

quantification, reliability and security. A complete survey of these measurement activities is outside the scope of this thesis. However, the following paragraphs will give an overview of measurement projects dealing with passive collection of Internet traces and packet-level analysis thereof, which is in close relation to the topic of this thesis. Generally, access to packet-level backbone traces is very uncommon, and a lot of research is performed on relatively small set of publicly available, but somewhat outdated network traces. This overview first presents the most prominent passive measurement projects, which have the possibilities to overcome the challenges of Internet measurement and therefore have access to own measurement facilities and resulting packet-level traces. Second, some smaller traffic analysis projects are pointed out, which are typically depending on shared, thus outdated datasets or flow-level data from cooperating service providers.

WAND Network Research Group

The WAND network research group is located at the University of Waikato Computer Science Department. WAND is a real network measurement research group, performing among other things collection of very long trace sets, network analysis, development of analysis software and network simulation and visualization. In the field of passive network measurements, WAND is best known for the WITS archive and the development of the DAG measurement cards. The Waikato Internet Traffic Storage archive (WITS archive) contains about 200GB of traces taken on different locations starting in 1999. Currently, only statistical summaries of the traces are publicly available, but the traces are planned to be shared in the near future. The DAG measurement cards have been developed at WAND as flexible and efficient hardware solutions for network measurements. Nowadays, support and development of DAG equipment is done by Endace , founded in 2001 as spin-off company. Except publications describing the development of DAG cards, WAND also contributed scientific measurement results based on WITS data traces, like the analysis of long duration traces by Nelson et al.

Other Related Work

Besides these big measurement projects, other relevant studies based on passive network measurement have been carried out by various researchers. Even if the available datasets did not reflect behavior of large parts of the Internet, some results are very significant and relevant. Allman studied deployment of TCP options within traffic from one particular webserver in a one and a half year period. Also Medina et al. used passive measurements of two weeks duration from a local webserver to present usage of specific TCP features . Other contributions had possibilities to record campus wide traffic for network analysis purposes. Arlitt and Williamson took a year long packet-level trace on the 100Mbit/s Ethernet campus network at the University of Calgary in order to analyse TCP reset behavior . Also Moore and Papagiannaki used packet-level data collected on a campus network based on Gbit-Ethernet to compare network application identification methods. These measurements were taken with Nprobe, a passive measurement architecture to perform traffic capturing and processing at full line-rate without packet loss. Finally, Mori et al. collected packet-level traces on the external 100 Mbit/s links of an University during a one month period to compare flow characteristics between WWW and P2P traffic. Measurements from networks with higher aggregation are usually only available in form of flow data. Gerber et al. e.g. had access to ten months of flow level data collected on several broadband ISPs, which was used to quantify P2P traffic in the Internet during the year 2002.

CONCLUSIONS

In this paper, network traffic c measurement have been studied and analysed. The external traffic at SICS was shown to be self-similar with the Hurst parameter estimated to. The traffic was also examined with respect to which protocols and packet sizes are present and in what proportions. In the SICS trace most packets are small, TCP was shown to be the predominant transport protocol and NNTP the most common application. In contrast to this, large UDP packets sent between not well known ports dominates the Supernet traffic. Finally, characteristics of the client side of the WWW traffic was examined more closely in order to create a simple model of WWW-sessions. Empirical probability distributions were derived describing session lengths, time between user clicks and the amount of data transferred due to a single user click.

REFERENCES

[1] Jain, R & Routhier, Shawn A., 1986. Packet Trains – Measurements and a New Model for Computer Network Traffic. IEEE Journal on Selected Areas in Communications (JSAC), 4(6), pp. 986-995, September 1986.

[2] Kalden R., Varga T., Wouters B., Sanders B. (2003). Wireless service usage and traffic characteristics in GPRS networks. In: Proceedings of the 18th International Teletraffic Congress – ITC18, Vol. 2, pp. 981-990, 31 August – 5 September 2003, Berlin,Germany.

[3] Kalden, R., 2004. Mobile internet traffic measurement and modeling based on data from commercial GPRS networks. Doctoral dissertation. University of Twente, The Netherlands.

[4] S. Coull, C. Wright, F. Monrose, M. Collins, and M. Reiter, "Playing devil's advocate: Inferring sensitive information from anonymized network traces," in Proceedings of the Network and Distributed Systems Security Symposium, San Diego, CA, USA, 2007.
[5] Ruoming Pang, Mark Allman, Vern Paxson, and Jason Lee, "The devil and packet trace anonymization," SIGCOMM Comput. Commun. Rev., vol. 36, no. 1, pp. 29–38, 2006.
[6] Jun Xu, Jinliang Fan, Mostafa H. Ammar, and Sue B. Moon, "Prefix-preserving ip address anonymization: Measurement-based security evaluation and a new cryptography-based scheme," in ICNP '02: Proceedings of the 10th IEEE International Conference on Network Protocols, Washington, DC, USA, 2002, pp. 280–289.
[7] Miquel Carsi Caballer and Lei Zhan, "Compression of internet header traces," Tech. Rep., Master Thesis, Chalmers University of Technology, Department of Computer Science and Engineering, 2006.
[8] Vern Paxson, "Strategies for sound internet measurement," in IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, 2004, pp. 263–271.
[9] J. Cleary, S. Donnelly, I. Graham, A. McGregor, and M. Pearson., "Design principles for accurate passive measurement," in PAM '00: Proceedings of the Passive and Active Measurement Workshop, 2000.
[10] Chuck Fraleigh, Sue Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and Christophe Diot, "Packet-level traffic measurements from the sprint ip backbone," IEEE Network, vol. 17, no. 6, pp. 6–16, 2003.