History-based classification of encrypted traffic into QoS class with self-update

Chi-Jiun Su
Hughes Network Systems, LLC
11717 Exploration Lane
Germantown, MD 20876
chi-jiun.su@hughes.com

Kaustubh Jain
Hughes Network Systems, LLC
11717 Exploration Lane
Germantown, MD 20876
Kaustubh.Jain@hughes.com

Sriram Vasudevan
Hughes Network Systems, LLC
11717 Exploration Lane
Germantown, MD 20876
sriram.vasudevan@hughes.com

Abstract— Traffic flows needs to be classified into QoS classes to prioritize traffic according to their QoS requirement. Without proper provision of differentiated service based on QoS requirement of each application, a user's Quality of Experience (QoE) will be poor and communication system resources may not be efficiently utilized. Since encrypted traffic now constitutes most of the Internet traffic, classification of encrypted traffic is necessary for a satellite broadband system. We propose a novel approach to classify encrypted traffic instantaneously at the start of a traffic flow based on actual measured traffic characteristics. The classification is based on the history which is self-updated when a flow is finished. The approach is also capable of updating classification type at the middle of the flow when the measured traffic characteristics do not match with the history-based classification.

Keywords—Encrypted traffic, QoS, Traffic Classification

I. INTRODUCTION

In recent years, more and more traffic use a few popular port numbers such as TCP port 80 and TCP port 443, to avoid blocking by firewalls and ISPs (Internet Service Providers). As a result, port numbers no longer carry useful information about application and traffic types, and it is not possible to achieve meaningful classification of traffic using port numbers. The alternative approaches for traffic classification are based on signatures in packet content, on traffic statistics and on keywords. Most of them examine several packets to classify traffic type. Keyword-based approaches usually make use of a FQDN (Fully Qualified Domain Name) of a flow and the information may no longer be available as new protocol standards start to encrypt not only the packet payload but also the packet header. Moreover, they do not consider the traffic characteristics for classification and this may lead to incorrect classification when key words used do not match the actual traffic type.

Most traffic classification techniques, including commercial products, classify traffic into an application or an application class. To provision differentiated service according to QoS classes, the list of applications or application classes often need to be manually categorized into QoS classes. Every time a new application appears, human inspection is required to map it onto a QoS class. To eliminate the human-in-the-loop problem, our approach directly classifies applications with different QoS requirement into different queues of QoS classes. Examples of QoS classes include interactive class, streaming classes, real-time class, and bulk class.

II. HISTORY-BASED CLASSIFICATION SYSTEM

A. Location of a Classifier

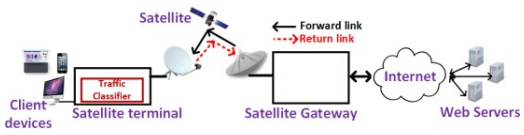

Figure 1 A traffic classifier located inside a satellite terminal.

Since a client usually initiates a traffic flow, a natural place to have a traffic classifier is at a Customer Premise Equipment (CPE) device such as a satellite terminal or a terrestrial broadband modem/router.

B. History based Classification system with Self-update

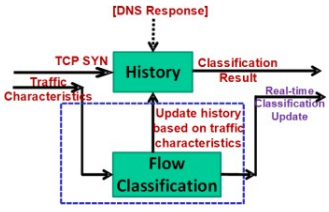

Figure 2 History based Classification system with Self-update.

Since Internet traffic is predominantly TCP, the whole discussion of our approach will focus on TCP traffic as a use case.

Figure 2 shows a history-based classification system with self-update mechanism. When a TCP SYN arrives as the first packet of a flow, classification type of the flow can be read from the history using a lookup key and the flow is classified instantaneously. The lookup key can be either a FQDN or an IP address. History is built from classification type update from flow classification. The traffic characteristics of the flow are also fed into the flow classification as the flow progresses.

At the middle of the flow, if real-time classification result of the flow does not match the classification type in the history, classification type of the flow will be updated to take appropriate action with respect to the treatment of the flow since QoS class of the flow is incorrect. For example, the flow was initially classified as interactive class from the history at the beginning of the flow but real-time classification from the flow classification indicates that it is bulk. If it is a big file download which may last for a long time, it may be better to downgrade the flow to bulk class and move the flow from interactive class queue to bulk class queue to avoid adding unnecessary delay to packets of other interactive class flows.

At the end of the flow, all the information on traffic characteristics of the entire flow is used to perform off-line classification. Then, history is updated with classification type of the flow from off-line classification. Note that real-time classification tries to classify the flow as soon as possible and makes use of partial information of traffic characteristics of the flow, whereas off-line classification tries to classify the flow as accurately as possible and uses all the information of the entire flow.

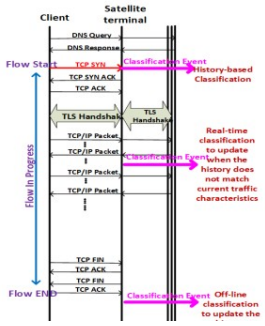

Figure 3 History based classification with self-update during a life-time of a TCP flow.

Figure 3 depicts the timeline of classification events during a lifetime of a traffic flow.

C. *Flow Classification*

The observable to a flow classification is information on IP packets of a flow in both directions. One approach to classify a flow is to characterize application from information of IP packets in both directions as shown in Figure 4. Application layer segments are explicitly or implicitly reconstructed from information on IP packets to derive application layer segment size and inter arrival time of application layer segments. Application layer characterization can be performed using TCP level statistics. The TCP level statistics can be readily available by reading TCP state machine statistics. By using application layer characteristics, rule-based or machine learning based approaches can be employed for flow classification.

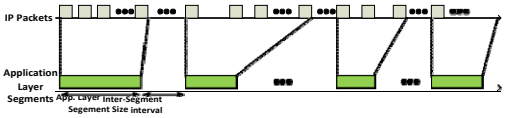

Figure 4 Application characterization from IP packets.

D. *FQDN/IP Address Level Classification*

There can be more than one flow to a FQDN/IP Address and flow classification types of the flows can be different. A weight can be assigned to each flow. The weight can be flow classification confidence level, total byte of a flow, flow duration or some combination of characteristics of a flow One candidate for classification type of a FQDN/IP Address is to select the classification type with maximum weight if total weight of the dominant type flow exceeds a predefined threshold.

III. PERFORMANCE EVALUATION

Experiments were performed to evaluate the performance of the history-based classification system. A simple rule-based algorithm using TCP statistics was used for off-line flow classification. Traffic traces were collected for different QoS traffic classes such as multimedia, interactive and bulk from Alexa top 50 HTTPS sites and popular multimedia and bulk web sites. Overall accuracies of classification achieved for flow level and for FQDN/IP Address level are 90% and 95% respectively. Aggregation of flow level classification into FQDN level classification improves the overall accuracy of the history-based classification system.