December 2020

Inferring Duplicate Data Traffic in Backbone Networks

Anonymous

Inferring Duplicate Data Traffic in Backbone Networks

ABSTRACT

This disclosure describes techniques for detection of duplicate data transfers over a backbone network based on time series correlation of the data traffic. Per techniques of this disclosure, a time series of data traffic is generated. Correlation of the time series of traffic flows from different servers is performed to determine duplicate data transfers. A correlation coefficient between different segments of traffic flow time series is determined. Segments of data traffic with a correlation coefficient that meet a threshold correlation coefficient (correlation coefficient close to 1) are identified as likely duplicate transfers. The described techniques can be utilized to detect duplicate data transfers within a backbone network. The techniques enable removal of duplicate data transfers leading to substantial cost savings and can help alleviate network congestion.

KEYWORDS

- Backbone network
- Core network
- Undersea cable
- Backbone traffic
- Data center
- Time series correlation
- Deep packet inspection

BACKGROUND

Enterprises commonly utilize a backbone network to connect clusters of computing resources (e.g., servers) that are spread over data centers in different geographical locations. In

some cases, the backbone network can connect far flung computing resources that are located in different continents. A backbone network can include constrained resources such as under-sea cables, buried cables, etc. Various services and applications utilize the backbone network for data transfers. However, services within the network can be configured in a manner that can sometimes lead to transmission of large volumes of duplicate traffic, thereby increasing infrastructure costs. Additionally, such inefficient network utilization can also lead to service disruptions due to constrained network capacity.

Detecting duplicate transfers using deep packet inspection requires extensive compute resources. Further, since backbone traffic may be encrypted and sourced from different servers, duplicate traffic detection techniques based on deep packet inspection can fail since such traffic may not show up as duplicate when encrypted using different encryption keys. To ensure detection, packet decryption needs to be performed prior to determining correlation of packet payload traces from multiple servers.

DESCRIPTION

This disclosure describes techniques for the detection of duplicate data transfers based on time series correlation of the data traffic. Per techniques of this disclosure, a time series of data traffic, e.g., as measured in bits/second, is generated for different servers from which the traffic originates. Correlation between the different time series of traffic flows is performed to identify duplicate data transfers.

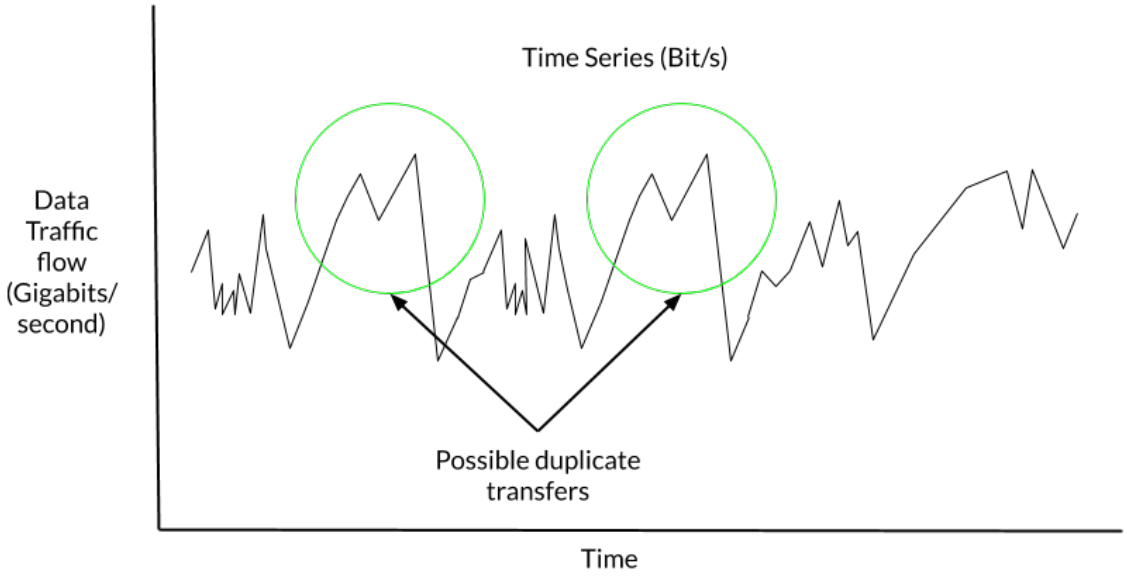

Fig. 1: Time series data of traffic flow is utilized to detect duplicate transfers

Fig. 1 depicts example traffic flow time series data. Per techniques of this disclosure, a correlation coefficient between different segments of traffic flow time series is determined. Segments of traffic with a correlation coefficient that meet a threshold correlation coefficient are identified. For example, segments that meet a threshold coefficient that are approximately 1 (e.g., correlation coefficient close to 1) are identified as possible duplicate flows (data transfers). Based on the time series data, an ordered list of likely duplicate transfers is generated. In this illustrative example, the highlighted segments (indicated by the green circles) are identified as likely duplicate transfers for additional investigation.

The identification of duplicate transfers is based on the principle that different servers that send the same content over the backbone network are driven by the same underlying services or software application(s), and are therefore likely to be synchronous in time. A simple time series correlation of traffic flows can identify such duplicate transfers. In some implementations, a dashboard of time series data is provided for detecting duplicate transfers.

Techniques of this disclosure can be utilized to detect duplicate data transfers within backbone networks. Removal of duplicate data transfers can lead to substantial cost savings and can help alleviate network congestion.

CONCLUSION

This disclosure describes techniques for detection of duplicate data transfers over a backbone network based on time series correlation of the data traffic. Per techniques of this disclosure, a time series of data traffic is generated. Correlation of the time series of traffic flows from different servers is performed to determine duplicate data transfers. A correlation coefficient between different segments of traffic flow time series is determined. Segments of data traffic with a correlation coefficient that meet a threshold correlation coefficient (correlation coefficient close to 1) are identified as likely duplicate transfers. The described techniques can be utilized to detect duplicate data transfers within a backbone network. The techniques enable removal of duplicate data transfers leading to substantial cost savings and can help alleviate network congestion.