Network Working Group
Internet Draft
Group
Intended status: Informational
Expires: February 1, 2009

Liufei. Wen
Huawei Technologies Network Working

Yunfei. Zhang
China Mobile

July 4, 2008

P2P Traffic Localization by Traceroute and 2-Means Classification
draft-zhang-alto-traceroute-00.txt

Status of this Memo

Copyright Notice

Abstract

Most P2P system performance suffers from the mismatch between the randomly constructed overlays topology and the underlying physical network topology, causing a large burden in the ISP and a long RTT time. This document

describes how DHT overlay peers can interact with the routers by traceroute to get the path information, and execute 2-Means Classification, thereafter peers leverage the DHT itself to build efficient "closer" cluster. This scheme only requires the infrastructure to enable traceroute queries.

Conventions used in this document

In examples, "C:" and "S:" indicate lines sent by the client and server respectively.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119.

Table of Contents

1. Introduction

This document describes how DHT overlay peers get the topology information and reduce the mismatch by traceroute and 2-Means classification. In particular, an assumption is made about the infrastructure routers support peers' traceroute requests, no matter what specific means(ICMP traceroute or TCP traceroute).

In a P2P system, each end node provides services to other

participating nodes as well as receives services from them. An

attractive feature of P2P is that peers do not need to directly interact with the underlying physical network, providing many new opportunities for user-level development and applications. Nevertheless, the mechanism for a peer to randomly choose logical neighbors, without any knowledge about the physical topology, causes a serious topology mismatch between the P2P overlay networks and the physical networks.

The mismatch between physical topologies and logical overlays is a major factor that delays the lookup response time, which is determined by the product of the routing hops and the link latencies. Mismatch problem also causes a large volume of redundant traffic in inter-domain between the every ISP. These has constituted the motivation to the topology-aware P2P, which implies to mitigate such drawbacks.

The purpose of this document is to specify a way to efficient topology matching technique. The DHT overlay peers' Traceroute result are used to get "near" clusters and Edge Gateway by execute 2-Means classification. This information will be put into the DHT. Two peers will be considered as to hava a close neighbor relationship, if they have at least one coomon router among their "near" clusters and Edge Gateways.

1.1. Terminology

In this document, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described RFC 2119 [RFC2119].

This section defines some key concepts using in this document.

DHT - distributed hash table (DHT).

DHT Overlay : An overlay network is a computer network which is built on top of another network. The peer-to-peer networks are overlay networks because they run on top of the Internet. And the peer-to-peer network which build with DHT is DHT Overlay.

2-means classification: k-means classification is an algorithm to classify objects based on their attributes into K number of group. 2-means classification algorithm is a special example of k-means

classification algorithm with k=2.

2. Overview

Usually, to solve the mismatch problem, it needs three steps, first to estimate the physical network distance between two overlay peers through network probing or prediction. then, based on this proximity information , to cluster "near" peers, so as to let peers can find a better candidate than the randomly chosen result. At last, such results are utilized to optimize P2P alogrithm. The "near" peers are the preferential choice in P2P lookup and maintenance, and these will impel the access and traffic localization and reduce delay.

Figure 1 Basic Steps of P2P Traffic Localization

In this document, an efficient topology matching technique is specified. First, before a peer joining the P2P networks, it randomly picks an Internet IP address and probes it using the traceroute tools. According to the measured data, the peer tracked the return information to a vector data. Then 2-means classification algorithm is used to classify the Internet routers into "near" and "remote" routers. Finally, peer chose the router with maximum Hops item in "near" set as a the Edge Gateway. The peer registers into the DHT overlays with the Edge Gateway the Key, then do same to the "near" set. Through shared vector cluster information such as "near" routers cluster and Edge Gateway, two peers were considered as a close neighbor relationship when their "near" routers cluster both had a same router's IP address at least and then gathered together to form a "close" peer clusters.

3. Traceroute and Clustering

3.1. Peer Traceroute

As a rapid developed network, the Internet presents two kinds of basic characteristic. On one hand, with many end user hosts randomly join and leave the network, the topology of Internet is dynamic and variable, but the routers constitute a much more stable

infrastructure topology. On the other hand, regarding Internet as a graph (V=routers, E=direct link between routers), we found that the edges between ASes constitutes a very small portion among the total edges, and usually the delay between ISP routers is more large than the delay between routers in the same AS domains.

A peer need randomly picked an Internet IP address and probed it using the traceroute tools. The peer tracked the return information to a vector data, with the data structure <IP, Hops, Latency>.

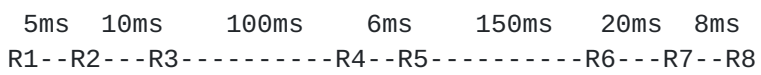

Fig.2 The Traceroute Result

Fig.2 is an example of path information by R1 traceroute to R8. As is clear from Fig.2, between the R3 and R4, R5 and R6, there are some huge latency leaps than the others. It possibly means that traceroute message across the different AS domains or different ISP ranges.

3.2. 2-Means Classification

When the overlay peer gets the traceroute result through randomly probe, a 2-means classification algorithm is used to classify the Internet routers based on the Latency attribute in these traceroute result. The 2-means classification algorithm includes four steps as following:

step1. Peer chooses the minimum latency item and maximum latency item in whole vectors as centroids for two initialization sets "first" and "second".

As a example in Fig.2, <R1, 1, 5ms> and <R5, 5, 150ms> is selected as centroids for sets "first" and "second".

step2. Peer takes the latency item in vector to make an absolute distance value with two centroids in turn, and then separately associated the corresponding vector to one vector cluster that has smaller absolute distance value.

So for the Fig2. example, the vector of R2 is <R2, 2, 10ms>, and the distance between R2 and R1 is 10-5 = 5, and the distance between R2 and R5 is 150-10 = 140. So <R2, 2, 10ms> belongs to the "first" set.

step3. Peer calculates the latency mean and variance value of two vector clusters. As in in Fig.2, the R1,R2,R4,R6,R7 belong to the "first" set, and R3,R5 belong to the "second" set and the mean and variance can be calculated.

step4. If the variance value was larger than the threshold, peer picks two latency mean values as new centroids of "first" and "second" sets, then goto step2, otherwise finishes the classification and gets two "first" and "second" sets.

Finally, peer chooses the router with minimum Hops item in "second" set as a hop threshold. This router and the other routers whose Hops item are larger than the hop threshold all divided into a "remote" router cluster. And then the remaining routers are gathered into another "near" cluster. and the router with maximum hops of the "near" cluster is regarded as overlay peer's Edge Gateway.

So for the Fig.2 example, R1,R2,R3 are the "near" routers of overlay peer and others routers are the "remoter" routers of the overlay peer. R3 will be the Edge Gateway in this case.

3.3. Form the Cluster

The peer registers into the P2P overlays with their Edge Gateway and "near" routers as the Key, and the DHT ID and IP of itself as the value.Due to the essence of DHT, if two item have the same Key, they will be routed to the same DHT peer. Thus it makes possible to let several peers to know each other, if they have some common elements of their "near" router set. Edge Gateway is more useful, so if the hosting DHT peers become overloaded, it will firstly deletes such non Edge Gateway routers item. The hosting DHT peer will notify those peers to form a "close" cluster, or join an existed cluster.

Each peer clusters has a ClusterId generated by consistent hashing when the first two peers decided to form the cluster. The ClusterId and its member peer Ids will be PUT into the DHT, and the peer can find the other members within the same cluster by DHT GET with its ClusterId remembered during its last online life.

3.4. Update

During normal peer-to-peer interactions such as DHT lookup or maintenance, if peers belonging to different clusters found the delay between them were relatively low, then these two clusters should decide to combine a new bigger cluster. The mapping between those original ClusterIds and the new generated ClusterIds should also be registered into the DHT so as to let those peers belonging to old clusters could find and join the new cluster. This technique alleviates the problem occurred when peers belonging to same cluster get different Edge Gateways from their traceroute response, thus they can not form the ideal bigger cluster.

4. Enhancement Examples

4.1. Find the proximate candidates

Peers register their resources to be shared as <Key, Value> pairs into the DHT. Usually the Key is generated by consistent hashing some information like the file name, and the Value is the IP address of the sharing peer. When use our scheme, we will also include the sharing peer's ClusterID in the Value.

When a overlay peer wants to find a resource, it will raise a DHT get with the hashed Key and piggyback its ClusterID. When getting the request, the peer hosting the Key k will check the list of <Key, Value> pairs registered by all the sharing candidates, then it will choose the sharing peer with the same ClusterId to be the preference result.

4.2. More Efficient Overlay Routing

The delay between the ISP is larger more than the inner-domain delay. And if AS domain is big enough many resource can be found in the same AS domains. So a more efficient hierarchical P2P network is feasible. Each low layer (local) DHT is composed by the peers with same ClusterID. All the peers or some candidate peers from each local DHT will join the global DHT. Every peer firstly search in its local DHT for its desired resource, then it may switched to the glocal DHT only if the resource not available locally.

4.3. Placement of Cache

In order to reduce the inter-domain traffic and delay, cache is

always considered in the P2P network. The placement strategy takes the cluster info into account. It will place caches to cover the peer clusters in the order of its population, the number of peers participating the cluster.

5. Security Considerations

This document does not currently introduce security considerations.

6. IANA Considerations

This document does not specify IANA considerations.

References

[1]RFC2119, Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[2]IFIP Networking 2008, Guangyu Shi, Youshui Long. "T2MC: A Peer-to -Peer Mismatch Reduce Technique by Traceroute and 2-Means Classification Algorithm."

Author's Addresses

Yunfei Zhang
China Mobile Communications Corporation

Phone: +86 10 66006688
Email: zhangyunfei@chinamobile.com

Wen liufei
Huawei Technologies Co. Ltd.

Phone: +86 755 28977571
Email: wenliufei@huawei.com

Intellectual Property Statement

this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

Copyright Statement

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.