T2MC: A Peer-to-Peer Mismatch Reduction Technique by Traceroute and 2-Means Classification Algorithm

Guangyu Shi[1], Youshui Long[1], Jian Chen[1], Hao Gong[1] and Hongli Zhang[2]

[1]Huawei Technologies Co., Ltd
Shenzhen, 518129, China
{shiguangyu, longyoushui, jchen, haogong}@huawei.com
[2]Harbin Institute of Technology
Harbin, 150001, China
zhl@pact518.hit.edu.cn

Abstract. The Peer-to-Peer (P2P) technology has many potential advantages, including high scalability and cost-effectiveness. However, most P2P system performance suffers from the mismatch between the overlays topology and the underlying physical network topology, causing a large volume of redundant traffic in the Internet. A lot of research works have been presented to address this issue, but most results still have some drawbacks. In this paper, we propose a quite simple but efficient topology matching technique, T2MC, which uses the peers' Traceroute result to execute 2-Means Classification, thereafter lets peers to build efficient "close" cluster. By performing experiments using the measured realistic Internet data of China, we show that T2MC outperforms the well-known GNP in both aspects of accuracy and maintenance cost.

Keywords: Peer-to-Peer, mismatch, clustering

1 Introduction

Since the emergence of Peer-to-Peer (P2P) file sharing applications, millions of users have started using their home computers for more than browsing the Web and exchanging e-mails. With the development of the networks, the P2P model is quickly emerging as a significant computing paradigm of the future Internet. There are currently several P2P systems in operation and many more are under development. According to the methods of administering peers, they can be broadly classified into four categories, the central directory based P2P system, e.g. Napster [1]; the decentralized unstructured P2P systems, e.g. Gnutella [2] and Kazza [3]; the decentralized structured P2P systems, e.g. Chord [4], CAN [5], Pastry [6], and the reinforced P2P systems such as Bamboo [7], Accordion [8], and PROP [9], which improved the system in Mismatch solution, Churn-Resilient and Heterogeneity-aware.

In a P2P system, an attractive feature 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,

The research is partially supported by the National Information Security 242 Program of China under Grant No.2006A18.

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 logical link latency. Mismatch problem also causes a large volume of redundant traffic in inter-domain between the every ISP, which is also an important reason why ISP may prohibit P2P applications [10, 11, 12]. Studies in [13] show that about 75 percent of the query response paths suffer from the topology mismatch problem in 8000 logical peers on 27000 physical nodes.

The issue of mismatch in P2P systems has been the focus of intensive research in recent years. The research works can be classified into four categories as follows:

The first representative studies to address the mismatch problem are peer coordinate systems such as GNP [14] and Vivaldi [15], which demonstrated the possibility of calculating synthetic coordinates to predict Internet latencies. GNP relies on a small number (18) of "landmark" nodes; other nodes choose coordinates based on RTT measurements to the landmarks. The choice of landmarks significantly affects the accuracy of GNP's RTT predictions. Requiring that certain nodes be designated as landmarks may be a burden on P2P systems. Vivaldi is a simple, light-weight algorithm that assigns synthetic coordinates to hosts such that the distance between the coordinates of two hosts accurately predicts the communication latency between the hosts. Although Vivaldi ensures nodes always decrease the prediction error and has no use for the choice of landmarks, it is so complex that hard to deploy in nowaday Internet. In addition, Vivaldi's convergence of coordinate system is a slow process.

The second study approach is a hierarchical location-based node IDs in P2P systems proposed by [16]. Physical locations and network distances are effectively embedded in the node IDs, and thereby improving routing locality. However, what is the influence on load balance is an important question concerning embedding location prefixes in node IDs.

The clustering algorithm such as Coral [17] takes a third different approach. In Coral system, the peer discovered the router IP cluster by randomly traceroute then chose the first five routers as the close cluster and joined into the cluster. But this scheme is coarse-grained and has difficulty to distinguish relatively close nodes. Why did Coral choose the first five routers as the criterion of close cluster?

Finally, the mechanisms based on measurement have been emerged in the last few years, such as Bamboo [7]. They select proximity neighbor based on monitoring daily DHT operation continuously. The disadvantage of this scheme is that it is hard to measure and learn adequate neighbor information during the peer's limited life time.

In order to address the limitations of the above cited work, we identify several basic characteristics of current networking paradigm, and propose a quite simple but efficient topology matching technique, called T2MC, which uses the peers' Traceroute result to execute 2-Means Classification, thereafter lets peers to build efficient "close" clusters. Using the T2MC technique, we can optimize the overlay topology by identifying and replacing the mismatched connections.

The rest of this paper is organized as follows: Section 2 describes T2MC in detail. Simulation methodology and performance evaluation are discussed in section 3. And we conclude the work in Section 4.

2 The T2MC Technique

Optimizing inefficient overlay topologies can fundamentally improve P2P search efficiency and decrease redundant traffic. T2MC utilizes several basic characteristics of current Internet paradigm, and lets peers belong to same ISP cluster together separately without any centralized control and predefined system parameters. We present two key aspects that need to be addressed in the design of T2MC following.

2.1 The characteristics of Internet paradigm

The Internet presents two kinds of basic characteristic. On one hand, with many terminal peer devices randomly join and leave the network, the topology is dynamic and variable. On the other hand, the routers and switches split the realistic network to different AS domains. The whole topology of Internet presents a strict hierarchy. During a long period to observe, it's clearly that the core routers are more stable than the terminal peers. According to this characteristic, we tried to make use of relationship between peers and routers in a topology location aspect.

In the mean time, the inter-ISP links constitute a considerable small portion of the total links of Internet, and usually such inter-ISP links have a much higher delay than other intra-ISP links. Just like the Fig.1, a mass of inner-routers from different AS domains are connected together by a few "star" edge routers. This inspired us to divide the nodes into "near" and "remote" router clusters. We fulfilled this task with the peers' Traceroute results. Fig.2 is an example of path information by R1 traceroute to R8. As clearly seen from Fig.2, between the R3 and R4, R5 and R6, these links present some huge latency leaps than the others. It usually means those hops across the different AS domains or different ISP ranges. How to find these latency leaps is the core part to distinguish the "near" and "remote" router clusters in T2MC.

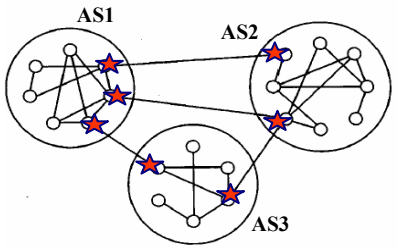


Fig.1 the characteristics of Internet

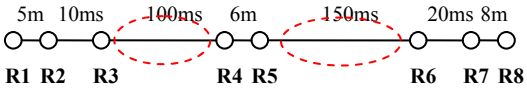


Fig.2 the characteristics of Internet

2.2 Traceroute mechanism

As mentioned before, the stabilized routers are more suitable for making a mark than dynamic terminal peers. We used the routers' location relationship as the judgment criterion of "near" and "remote" routers. T2MC uses the peers' Traceroute result to execute 2-Means Classification, thereafter lets peers to build efficient "closer" cluster.

Firstly, when a peer joining the P2P networks, it randomly picked an Internet IP address and probed it using the traceroute tool. The peer tracked the results into a vector of <IP, Hops, Latency> tuples, where IP, Hops is the ip address and hops of every router in traceroute path and Latency means the link delay between this router and its pervious-hop. In T2MC, the peer chose an IP address with the first bit far

different from itself, if this doesn't work, it will randomly chose another suitable IP address to probe.

Secondly, based on the Latency attributes of the vector tuples, T2MC used the 2-means classification algorithm to classify the Internet routers into "near" and "remote" routers. 2-means classification algorithm is a special example of k-means classification algorithm [18] with k=2. By this tactic, T2MC clustering avoided any presumed hops threshold such as used in Coral [17]. The 2-means classification algorithm in T2MC includes four steps as following.

step1. Peer chose the tuples with minimum or maximum Latency attribute as centroids for two initial sets, "first" and "second".

step2. Peer took the latency attribute of each tuple in the vector to calculate the absolute distance value with the two centroids in turn, and then associated the tuple to the set whose centroid has smaller absolute distance value with it.

step3. Peer calculated the latency mean and variance value of two sets.

step4. If the variance value was larger than the threshold, peer picked two latency mean values as new centroids of "first" and "second" sets, then took a loop from step2 until the two sets achieved the minimum total intra-set variance. This produces a separation of the routers into two groups from which the metric to be minimized can be calculated. At last, two "first" and "second" sets confirmed in the end. The pseudo code of 2-means classification algorithm in T2MC is shown in Fig.3

```
structure Routerinfo
    ip ipaddress                Δ the ip of routers
    double srcdelay             Δ the delay form this to source
    double hopdelay             Δ the link delay

procedure Get2meansCluster(allrouters[1...routenumber])
    firstcenter ← allrouters[min].hopdelay
    secondcenter ← allrouters[max].hopdelay
    oldE ← -1                   Δ the previous variance
    E ← 0                       Δ variance
    while (E - oldE) > 0.000001
        firstlist.clear()
        secondlist.clear()
        foreach i ← 0 to allrouters.size
         if ((allrouters[i].hopdelay-firstcenter) <= (secondcenter
         - allrouters[i].hopdelay))
            then firstlist.add(allrouters[i])
            else secondlist.add(allrouters[i])
        end foreach
        firstcenter ← the mean of firstlist.hopdelay
        firstE ← the variance of firstlist.hopdelay
        secondcenter ← the mean of secondlist.hopdelay
        secondE ← the variance of secondlist.hopdelay
        oldE ← E
        E ← firstE + secondE
        ret ← new Vector[2]
    end while
    Ret[0] ← firstlist          Δ the "close" routers clustering
    Ret[1] ← secondlist         Δ the "far" routers clustering
    Retrun Ret[0…1]
```

Fig.3 pseudo code of 2-means classification

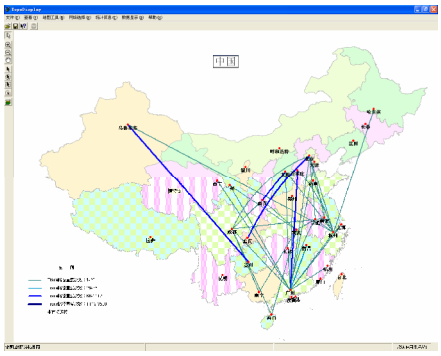


Fig.4 measured Internet topology of China

Finally, peer chose the router with minimum Hops attribute in "second" set as a hop threshold. This router and the other routers whose Hops attribute are larger than the hop are classified as the "remote" router cluster. The remaining routers formed the "near" router cluster. The peer chose the router with maximum Hops attribute in the "near" router cluster as its Edge Gateway, and then registered the Edge Gateway and

the “near” cluster into the P2P overlay, as DHT PUTs. If two peers shared the same Edge Gateway or any members of their “near” router clusters, they would gather together to form a “close” peer clusters. The Edge Gateways presents more valuable information than other members of the “near” router clusters, T2MC were designed to let peers to register/search/save Edge Gateways information with the higher priority.

In T2MC, each peer clusters has a ClusterId generated by consistent hash when the first 2 peers decided to form a cluster. The ClusterId and its member peer Ids will be PUT into the DHT (or other similar information retrieve means), and the peer can find the other members within the same cluster by DHT GET with its ClusterId remembered during its last online life.

So, in T2MC, the judgment criterion for close neighbor peers is based on the much more stable routers in the underlay. The overhead of this technique is the traceroute operation when peer joining the system. On the contrary, the “close” relationship of peers is judged by their Euclidean space in GNP technique. The peer must probe the landmarks to measure the RTT delay as overhead. The simulation results in the next chapter will demonstrate such contrast.

3 Simulations and analysis

In this section, we choose GNP as comparative technique. By performing experiments based on the measured realistic Internet data of China, we get the performance and overhead comparison. Then we discuss some further improvements of T2MC.

3.1 Simulation performance metrics

We mainly focus on three performance metrics: the accuracy of clustering “close” and “far” peers, the average delay in clusters and the average maintenance traffic cost.

The accuracy of clustering is one of the important performance metrics with which topology matching algorithms are seriously concerned. The low accuracy increases the total delay of the response and redundant traffic on backbone Internet. Internet measurement [19] reports that roughly 75%~90% of flows have RTTs less than 200ms. Hence, we define the real “close” peer cluster as those peers whose latency are less than 100ms from the source in the realistic networks. And the accuracy could be measured as an “AND” operation result between the “close” peer cluster formed by T2MC and the real “close” peer cluster.

We define the average latency in cluster as average link latency in each “close” peer cluster. The lower average latency obtained, the more efficient the cluster is.

Traffic cost is another parameter that most seriously concerned by ISPs. Heavy network traffic is always the direct reason why many ISPs dislike P2P. We define the traffic cost as how many network messages will be used in the clustering process.

3.2 Measurement methodology

To evaluate the effectiveness of T2MC, we need an actual topology data which made up of the major ISP routers and most city-level routers in China. We used 9 probing sources to collect the data. The sources are located in Anhui, Gansu, Hainan, Heilongjiang, Hebei, Guangxi, Beijing, Shanghai and Guangzhou, nine mainland provinces. To ensure the effectiveness of probing, we intentionally selected addresses from a large set of IP addresses within China range as targets, each /24 subnet we

chose 8 IP addresses. The sources triggered 20 processes to send the ICMP traceroute messages to the target and calculated the RTT value by monitoring the response messages. If the sequential 10 hops did not response, the probing process terminated. The RTT difference between previous hop and next hop was the link latency.

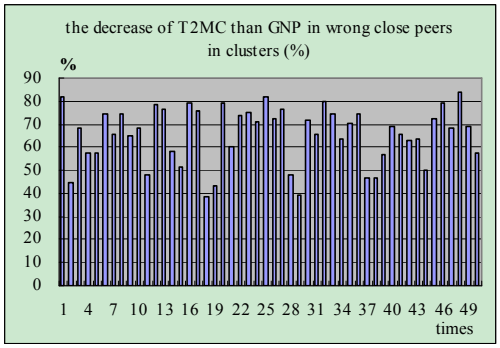
The measurements from all probing sources were conducted roughly around the same time. The whole measurement lasted three days. As a result, 76116 routers' IP addresses, 128083 links and their latency were obtained. They belong to a diverse set of prefixes originating from ASes across the entire Chinese Internet hierarchy. There are 33300 edge routers in whole 76116 routers. The measured routers were spread all over China as shown in Fig.4.

3.3 Experiment parameters

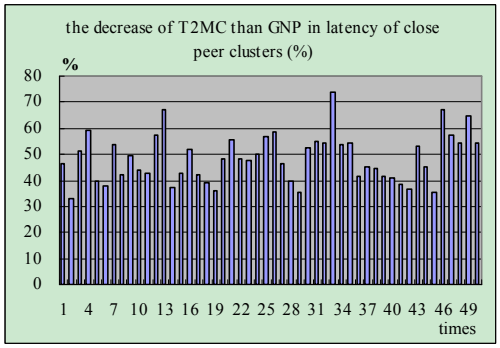
The logical topology represents the overlay topology built on top of the physical topology. We used PlanetSim simulator to generate logical topology, with 10000 peers each round. The peers randomly accessed the edge routes in physical topology. The J-sim simulator provided an interface to connect physical topology and logical topology. On the choice of the landmark number, we considered that in GNP thesis, GNP only set 15 landmarks for the RTT measurement in whole world. So with the China mainland topology, deploying 8 landmarks should be enough to meet the requirement. In our experiments, we set the number of GNP landmarks to 8. And the landmark was chosen randomly in whole nodes of P2P system. For the accuracy each experiment we run 50 rounds, each with different new generated 10000 peers.

3.4 Experiment result

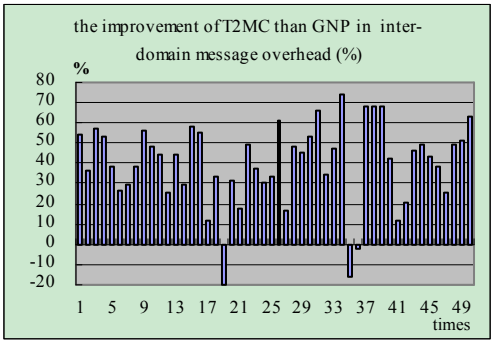
First of all, we need to measure the accuracy of clustering "close" and "far" peers. We choose the "close" peer cluster by topology matching technique to make a comparison with the real "close" peer with latency less than 100ms each other in the realistic network. From Fig.5(a), we find that the number of wrong close peers in close cluster decreases by 65% in T2MC than that of in GNP. It is mainly because in GNP, the landmarks as probing targets are chosen randomly. Every peer calculates own GNP distance and clusters together by probing the landmark, so the number and location of landmarks directly determine the performance of GNP schemes. But in T2MC technique, there is no landmark and the probing target is random. It avoids that the limitation of landmark in peer clustering. Furthermore, through efficiently clustering, T2MC can reduce a large volume of inter-domain redundant traffic than GNP.

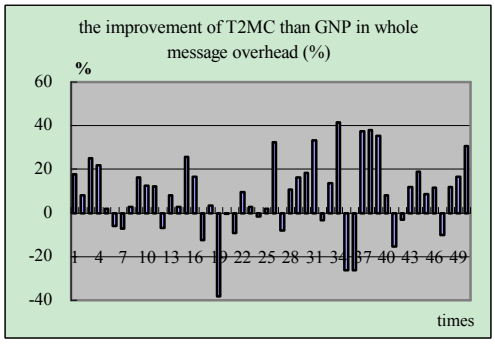


(a) the comparison of wrong close peer in clusters



(b) the comparison in latency of close peer clusters





(c) the comparison of inter-domain message overhead (d) the comparison in whole message overhead

Fig.5 performance comparison of T2MC and GNP in measured realistic Internet data (China)

The accuracy of clustering brings forth the improvement of latency immediately. Because of the "close" peers' number increased and accuracy of "close" peer clusters improved, the average latency in cluster decreases by 48% approximately from the Fig.5(b). The lower average latency means quicker P2P interactions.

As is clear from the Fig.5(c) and Fig.5(d), T2MC achieves another favorable performance metric, the maintenance traffic overhead. T2MC consumes less maintenance bandwidth overhead of clusters construction than GNP both in the inter-domain and whole network level. The reason is that in T2MC, each peer just probes only one peer. On the contrary, the peer must measure the delay of every landmark in GNP scheme, so that the number of landmark determines the maintenance overhead of scheme. As one can see in Fig.5(c) and Fig.5(d), T2MC decreases the average maintenance overhead in inter-domain by 40%, in whole by 8%.

3.5 Further improvements

Previous sections have shown that T2MC can achieve favorable performance. In the practical deployment, there are still some details need to be mentioned.

(1) In our measurement process, we only use ICMP traceroute as the probing method. The result of 76116 routers IPs only represents 50% of whole routers in China mainland probably. In future measurement, if chose multiple probing tools such as TCP traceroute and UDP traceroute, the more routers' IP address will be identified, so as to let T2MC get more excellent performance in topology matching.

(2) In probing process, for that peers could not get any traceroute response, they can't find their "close" routers to form a cluster. The solution is to utilize the IP address and subnet mask address to find out its gateway router.

(3) For neglecting the probing error practically, before the 2-means classification, we modify the Latency attribute less than 10ms to 10ms for each vector tuples, and also modify the Latency attribute more than 200ms to 200ms.

(4) T2MC uses the peers' traceroute result to execute 2-means classification, and lets peers to build their "close" and "far" clusters. For the further improvement, T2MC can perform 2-means classification to pervious "close" and "far" clusters recursively until the far centroid is less than 2 times of the close centroid during the latest round. With our experience, we found that 3 or 4 times recursive classification is quite enough for China network topologies, we suppose the whole world Internet maybe only needs 4 or 5 times.

(5) In T2MC, 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 would decide to combine to a new bigger cluster. The mapping between those original ClusterIds and the new generated ClusterIds would also be registered into the DHT so as to let those peers belonging to the original clusters could find and join the new cluster. This tactic alleviates the problem occurred when peers belonging to same cluster get different Edge Gateways from their traceroute response ("stars" in Fig.1), thus they may not form the ideal cluster.

(6) T2MC is presented as a novel mismatch reduction technique which outperforms existing schemes. But it should be mentioned that T2MC also can be easily applied to existing equivalent techniques without the loss of their primary characteristics, such as utilize T2MC to choose the landmark peers in GNP scheme, choose the replacement of cache points in IPTV networks, or help to construct the ALM tree.

4 Conclusion

In this paper we presented a simple but effective P2P mismatch reduction technique called T2MC. T2MC uses the peers' Traceroute result to execute 2-Means Classification, thereafter lets peers to build efficient "close" cluster. T2MC utilizes several basic characteristics of current Internet paradigm, which can efficiently and accurately classify the peers belong to same ISP into a sort of cluster. Furthermore, T2MC is completely decentralized, scalable, reliable and self-organizing, without any presumed hops threshold parameters. By performing experiments using the measured realistic Internet data of China, we show that T2MC outperforms GNP scheme in both aspects of accuracy and maintenance cost.

References

[1] Napster, http://www.napster.com, 2007.
[2] Gnutella, http://gnutella.wego.com, 2007.
[3] KaZaA, http://www.kazaa.com, 2007.
[4] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan, "Chord: A Scalable Peer-to-Peer Lookup Ser-vice for Internet Applications," In Proc of. ACM SIGCOMM, 2001.
[5] S. Ratnasamy, P. Francis, and S. Shenker, "A Scalable Content-Addressable Network," In Proc of. ACM SIGCOMM, 2001.
[6] A. Rowstron and P. Druschel, "Pastry: Scalable, Distributed Object Location and Routing for Large-Scale Peer-to-Peer Systems," In Proc of. Int'l Conf. Distributed Systems Platforms, 2001.
[7] Sean Rhea, Dennis Geels, Timothy Roscoe, and John Kubiatowicz, "Handling Churn in a DHT," In Proc of. USENIX, 2004.
[8] Jinyang Li, Jeremy Stribling, Robert Morris, and M. Frans Kaashoek, "Bandwidth-efficient Management of DHT Rout-ing Tables," In Proc of. NSDI, 2005.
[9] Tongqing Qiu, Guihai Chen, Mao Ye, Edward Chan and Ben Y. Zhao, "Towards Location-aware Topology in both Unstructured and Structured P2P Systems," In Proc of. IEEE ICPP, 2007.
[10] Mudhakar Srivatsa, Bugra Gedik, and Ling Liu, "Large Scaling Unstructured Peer-to-Peer Networks with Heterogeneity-Aware Topology and Routing," IEEE Transactions on Parallel and Distributed Systems, 2006, 17(11), pp.1277-1293.
[11] Vinay Aggarwal, Anja Feldmann and Christian Scheideler, "Can ISPs and P2P Systems Cooperate for Improved Performance?" ACM SIGCOMM Computer Communications Review, 2007, 37(3), pp.29-40.
[12] G. Shen, Y. Wang, Y. Xiong, B. Zhao, and Z. Zhang, "HPTP: Relieving the Tension between ISPs and P2P," IPTPS, 2007.
[13] Yunhao Liu, Li Xiao and Lionel M. Ni, "Building a Scalable Bipartite P2P Overlay Network," IEEE Transactions on Parallel and Distributed Systems, 2007, 18(9), pp.1296-1306.
[14] T. S. Eugene Ng and Hui Zhang, "Predicting Internet Net-work Distance with Coordinates-Based Approaches," In Proc of. IEEE INFOCOM, 2002.
[15] F. Dabek, R. Cox, F. Kaashoek, and R. Morris, "Vivaldi: A Decentralized Network Coordinate System," In Proc of. ACM SIGCOMM, 2004.
[16] Shuheng Zhou, Gregory R. Ganger, and Peter Steenkiste, "Location-based Node IDs: Enabling Explicit Locality in DHTs," Technical Report CMU-CS-03-171, 2003.
[17] Michael J. Freedman and David Mazieres, "Sloppy Hashing and Self-organizing Clusters," In Proc of. IPTPS, 2003.
[18] Vance Faber, "Clustering and the Continuous k-Means Al-gorithm," Los Alamos Science, 1994, 22, pp.138-144.
[19] H. Jiang and C. Dovrolis, "Passive Estimation of TCP Round-Trip Times," ACM Computer Communications Review, 2002, 32(3), pp.75-88.