Genius: Peer-to-Peer Location-Aware Gossip Using Network Coordinates[1]

Ning Ning, Dongsheng Wang, Yongquan Ma, Jinfeng Hu, Jing Sun,
Chongnan Gao, and Weiming Zheng

Department of Computer Science and Tecohnology, Tsinghua University, Beijing, P.R.C
{nn02, myq02, hujinfeng00, gcn03}@mails.tsinghua.edu.cn
{wds, zwm-dcs}@tsinghua.edu.cn

Abstract. The gossip mechanism could support reliable and scalable communication in large-scale settings. In large-scale peer-to-peer environment, however, each node could only have partial knowledge of the group membership. More seriously, because the node has no global knowledge about the underlying topology, gossip mechanism incurs much unnecessary network overhead on the Internet. In this paper, we present Genius, a novel peer-to-peer location-aware gossip. Unlike many previous location-aware techniques which utilize BGP or other router-level topology information, Genius uses the network coordinates map produced by Vivaldi as the underlying topology information. By utilizing the information, Genius could execute near-preferential gossip, that is, the node should be told the gossip message by nodes as close as possible, through which much unnecessary 'long-range' communication cost could be reduced. Further, the node direction information inherited in the coordinate space is exploited. We present preliminary experimental results which prove the feasibility of our scheme.

1 Introduction

Peer-to-Peer applications have been popularized in the Internet in recent years. Building reliable, scalable, robust and efficient group communication mechanism on top of peer-to peer-overlay is an important research topic. Such mechanism in peer-to-peer environment much meet following three requirements: the first one is scalability; the second one is reliability, robustness and decentralized operation; the third one is efficiency. The gossip communication mechanism pioneered by [2] emerged as a good candidate which has the potential to satisfy the requirements mentioned above.

The work in this paper deals with the unstructured system which is developed in the previous Scamp[4] protocol. However, the Scamp protocol does not take the underlying topology into account, that is, it is not location-aware. This causes much unneces-

[1] This work is supported by National Natural Science Foundation of China (60273006).

V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3516, pp. 163 – 170, 2005.

sary network overhead on internet. The waste of network resources is needless and should be reduced.

Many previous works [6][7][8] address the location-aware issue from various aspects in different settings. Due to shortcomings of these schemes explained at length in Section 4, we propose Genius, a novel peer-to-peer location-aware scheme. In Genius, a network coordinates map which indicates the coordinate of each node is produced by Vivaldi[1] when nodes join the system gradually. Our scheme obtains topology information from this map without the help of network-layer information provided by domain administrators.

In Genius, individual node gains the rather global view of the whole system by exchanging its local information with other nodes in the system. After gathering node information over the whole system, the individual node analyzes them and adjusts its local view so that each individual node has a balanced view of the whole system. Based on these adjusted location-aware local views, Genius could execute location-aware gossip which enables more nodes to be told the gossip message by more nearby nodes, and accordingly reduces the network overhead.

The primary contributions of this paper are the following:

1. It presents Genius, a peer-to-peer location- aware gossip using network coordinates without the help of network-layer information.
2. It proposes an approach to cluster nodes in the coordinate space in a decentralized manner.
3. Genius is the first system to exploit the node direction information in the coordinates map to our knowledge.

The rest of the paper is organized as follows. Section 2 describes the Genius design. Section 3 presents results. Section 4 discusses related work. Finally, Section 5 summarizes our conclusion and future work.

2 Genius Design

There are several steps for node in Genius to obtain location-aware local view.

2.1 Node Arrival and the Network Coordinates Map Construction

In Genius, new node joins the system according to the process specified in the subscription section of Scamp[4]. After collecting latency information from nodes in its *InView*[4], the new node computes good coordinates for itself using Vivaldi[1] which does not depend on the selection of landmarks[3]. When the coordinates converge, one global network coordinates map is formed and Genius enters into nearby nodes exchange stage.

According to [1], coordinates drawn from a suitable model can accurately predict latency between hosts and inter-host RTT is dominated by geographic distance. Thus, the node coordinates reflect its geographic location and it is reasonable to infer the geographic relation of nodes based on the coordinates map.

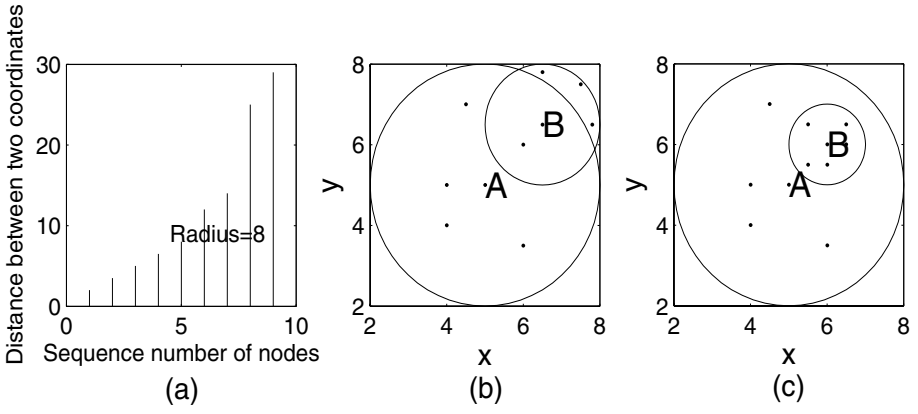

Fig. 1. Illustration of nearby nodes exchange and Clique formation. (a) is the selection of the Clique radius.(b) is the increasing of the Clique radius, and (c) is the decreasing of the Clique radius

2.2 Nearby Nodes Exchange Process and Faraway Nodes Exchange Process

The nearby nodes exchange action is to exchange topology information among the close-by nodes called "Clique" which is formed in the exchanging process simultaneously. The process is actually to cluster nodes in the coordinate space in a decentralized manner. Every node in the system executes the process. The process is described as follows.

Firstly, each node orders the distances between the coordinates of itself and the nodes in its *PartialView* (i.e. the latency between them) in the order of increasing distance. Find the largest difference between the adjacent distances from the smallest one to largest one, which is larger than the preceding difference and the succeeding difference. Define the smaller one of the two distances between which is the largest difference as radius *r*. This is illustrated by Figure 1(a). The largest difference in Figure 1(a) is the one between the distance of node #5 and that of node #6 which is 4(i.e.12-8) and larger than 1.5(i.e. 8-6.5) and 2(i.e. 14-12). The nodes in its *PartialView* whose distance is smaller than *r* are considered to belong to the same Clique as it does and *r* is the radius of the Clique.

According to the Clique radius, there are two sets of nodes in the *PartialView*: nodes belonging to its Clique are called 'Clique contact' and all the nodes except Clique nodes are called 'long-range' contact. The intuition behind the selection of Clique radius *r* is that the nodes in the same Clique should have relative similar distance from the source node and the nodes not belonging to the same Clique should have relative larger distance.

For acquiring more information about the Clique and preventing the Clique radius being excessively inaccurate just depending on its own *PartialView*, the node needs exchange its information about the Clique, that is Clique membership and Clique radius, with other nodes in its Clique. Thus, it will ask other nodes in its Clique about

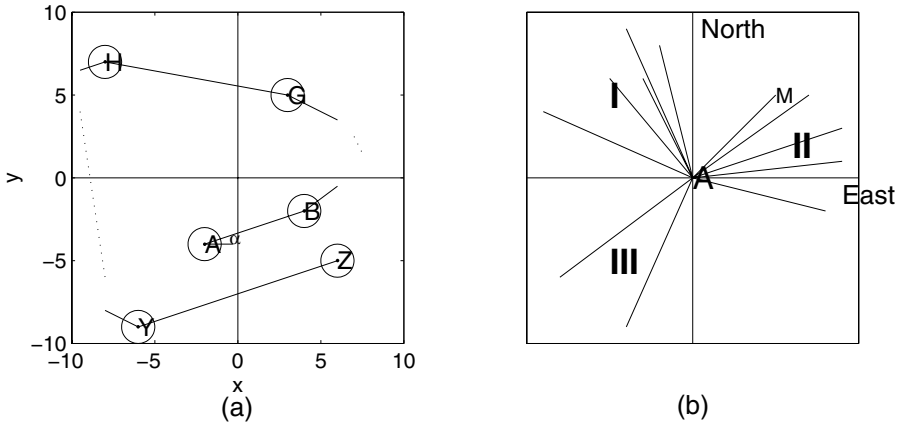

Fig. 2. Illustration of faraway nodes exchange and the 'long−range'contact set adjustment. In (a), node A traverses the whole system from node B to node Z. The circle centered at each node represents its Clique.'long−range' contacts of node A in different directions are shown in (b). They form areas of different angulardensity, such as area I,II,III

which nodes are in their Clique and their Clique radius. For the exchanging process progressing in the right direction and converging to the right state reflecting the real topology, we should force the node exchanging information with nodes whose information is more accurate than that of itself. We define the Clique radius of the node as its credit. Smaller the radius, more credit should be given to the node. The node only trusts nodes whose radius is smaller than its own and exchanges information with them.

Exchanging action is described as follows. As illustrated in Figure 1(b), node B is in the Clique of node A. The radius of Clique of node A is 3 and that of Clique B is 2. So, node A trusts node B and adds the nodes in the Clique of node B into its own 'Clique node' set of its *PartialView*. But the size of original *PartialView* is maintained for future use. This intuition behind the augment of membership is the transitivity property, that is, node A and node B belong to a same Clique, and that, nodes in the Clique of node B belong to the same Clique as node B do, so, nodes in the Clique of node B should belong to the Clique of node A. After adding more nodes into its own 'Clique node' set, node A starts modifying its Clique radius. Then, node A uses similar method described above to determine the new radius based on the result of the augment of Clique membership. This example illustrates the increasing of Clique radius. The example of deceasing of excessively large Clique radius of node A is shown in Figure 1(c). Through repeatedly exchanging information with trusted nodes in its own Clique, the node modifies its Clique radius in every step. The process converges until the variance of radius per step is smaller than one small percent of the radius, say 1%, in several successive steps.

After the node has exchanged information among nodes in its Clique, it needs execute the faraway nodes exchange process which is used to exchange information about 'long-range' contacts to gain a rather complete view of the whole system. We propose a method exploiting the node direction information in the coordinates map.

We observe that nodes with the same latency from the node, but in different direction mean differently to the node. Thus, nodes should not only be classified in the measure of latency and node direction is also an important parameter to be considered. For the node to be able to obtain a summary view of the whole system, our method works as follows. In Figure 2(a), node A wants to exchange information. Firstly, node A selects the longest 'long-range' contact node B which is of longest distance from node A in its 'long-range' contact set. The vector AB makes an angle of α with axis *x*. It then visits node B, gets one copy of 'long-range' contacts of node B. Next, node B finds the set of its 'long-range' contacts which satisfy that vector BC makes an angle of α+β with axis *x* (suppose node C is a random element of the set), then selects a node C randomly from the set as the next stop of the traverse. Similarly, node A visits node C, and gets one copy of its 'long-range' contacts. After that, node C selects a node D randomly from its 'long-range' contacts satisfying that vector CD makes an angle of α+2β with axis *x* as the next stop. Analogically, node E, F, and G are selected as the next stops. The process terminates when the last vector which is formed by the last two stops, such as YZ, makes an angle between α-β and α+β with axis *x*. This termination condition implies that node A traverses a circle on the whole system. When no node in the 'long-range' contacts satisfies the above requirement, the node most approximately meeting the requirement is selected.

The selection of β is determined as follows. Preliminary experimental results from Section 3 show that traversing about 35 nodes are enough to obtain a rather complete view of the system. So, it is suitable to let β be 360/35, approximately 10 degree. Making β positive means counterclockwise traverse and making β negative means clockwise traverse.

When node A traverses the system, it aggregates the copies of 'long-range' contacts of visited nodes in the network by processing over the data when it flows through the nodes, discarding irrelevant data and combining relevant data when possible. The goal of our aggregation of 'long-range' contacts is to identify a set of contacts in which only one contact belongs to every distinct Clique, that is, this set contains just one representative contact for each Clique.

The aggregation at each traversed node is described as follows. When node A visits node B, if the distance between node A and the node in the 'long-range' contact set of node B (for instance, node s) is smaller than the Clique radius of node s, node s is considered to belong to the same Clique of the 'long-range' contact of node A. When the traverse is complete, node A obtains the summary knowledge about other Cliques in the system. Every node in the system traverses the system and gets its own 'long-range' contact set. Although working in this way seems too costly, this kind of traverse happens only once when the system initializes.

2.3 The *PartialView* Adjustment and Maintenance in Dynamic Environment

After the detailed knowledge about its own Clique and a summary knowledge about other Cliques in the system are obtained by each node, the *PartialView* of each node should be adjusted so as to the size of the *PartialView* is *O(log(N))*, where N is the size of the system.

Firstly, we determine the size of the result *PartialView* of node A as the mean size of all original *PartialView* size of visited nodes in the previous traverse. This size is

O(*log*(N)). Secondly, we compute the proportion between the size of the Clique of node A and the size of the 'long-range' contact set. Then, we could determine the size of the result Clique contact set and that of the result 'long-range' contact set.

The next job is to select determined size of nodes from the two former sets. Basically, the selection strategy lies on the most concern of the system designer. There is an obvious tradeoff between network overhead (i.e. message redundancy) and reliability in gossip mechanism. For the Clique contacts, if the designer most concerns about reducing network overhead, those nodes which are nearest should be kept in the result Clique contact set. On the contrary, if the designer most concerns about gossip reliability, the result Clique contacts should be selected randomly from the former Clique contact set of node A. For the 'long-range' contacts (see Figure 2(b)), suppose a 'long-range' node M, node A computes the angle formed by vector AM and axis *x*. Node A computes the angle for each node in the 'long-range' contact set. If reducing network overhead is the most concern, contacts are selected to be kept according to the density of distribution of angles of contacts. If the most concern is gossip reliability, contacts in sparse area, such as area III, encompassing few nodes should be given priority to be selected. Given the location-aware *PartialView*, more nodes could be told the gossip message by closer neighbors. Thus, location-aware gossip is achieved.

Nodes leaves the system according to the process specified in the unsubscription section of Scamp[4].Because Vivaldi[1] naturally adapts to network changes, the node coordinates remain accurate.

3 Results

Preliminary experiments mostly concentrate on the faraway nodes exchange process. We use a data set which involves 1740 DNS servers mentioned in [1] as the input to

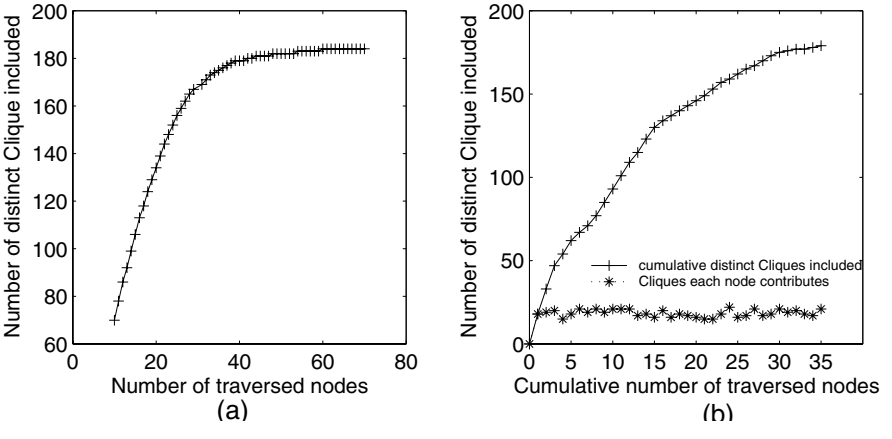

Fig. 3. Preliminary experimental results focusing on faraway nodes exchange process. (a) is the number of distinct Cliques included as a function of number of traversed nodes. (b) is the number of distinct Cliques included as a function of cumulative number of traversed nodes in one experiment in which 35 nodes are traversed

Vivaldi. The node coordinates output by Vivaldi is taken as the input to our simulator. Our simulator implements the logic in the paper.

Figure 3(a) shows that when more than 35 nodes are traversed, the number of distinct Cliques collected increases very slowly during the faraway nodes exchange process. The experiment suggests suitable number of nodes required to be traversed for obtaining a rather complete view of the system is 35.

Figure 3(b) shows that in one experiment in which 35 nodes are traversed, how the number of distinct Cliques included increases along with the progression of the traverse.

4 Related Work

Many location-aware techniques are proposed to achieve location-aware operations in different environment. LTM[6] is a measurement-based technique designed to do location-aware query in unstructured peer to peer systems. It builds a more efficient overlay by cutting low productive connections which is detected by TTL2 detector and choosing physically closer nodes as logical neighbors. Routing underlay[8] provides underlying network topology for overlay services by aggregating BGP routing information from several nearby routers. However, this approach is somewhat too ideal because not all the administrators of domain are willing to provide a feed from their BGP router to the underlay. In contrast, our network coordinates-based approach will not encounter this problem since coordinates are computed by participants cooperatively with no relation to BGP router.

As to the clustering of nodes in coordinate space, Sylvia Ratnasamy *et. al*.[9] proposed a binning scheme. A node measures its round-trip-time to each of d landmarks and orders the landmarks in order of increasing RTT. Thus, every node has an associated ordering of landmarks. The nodes having the same ordering of landmarks belong to the same bin. But, unfortunately, the approach corresponds to halve the right angle in the d-D Euclidean space, thus it does not reflect the node partition based on distance. Some research works also deal with the location-aware gossip. Localiser [7] chooses an energy function over all the nodes and edges in the system incorporating its locality-related objectives and then the function is minimized by simulated annealing through decentralized operation. However, Localiser depends on the choosing of the parameter sensitive energy function too heavily and it is possible for Localiser to be trapped at a local minimum.

5 Conclusions and Future Work

In this paper, we present Genius, a peer to peer location-aware gossip using network coordinates. In contrast to other approaches, it does not rely on any kind of network-layer information which is probably not available. It also does not depend on certain energy function very much. In the future, more experiments about Clique formation and gossip overhead, reliability should be conducted.

References

[1] Frank Dabek, Russ Cox, Frans Kaashoek, Robert Morris Vivaldi: A Decentralized Network Coordinate System In Proceedings of ACM SIGCOMM'04, Portland, Oregon, Aug, 2004
[2] A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Stuygis, D. Swinehart, and D. Terry. Epidemic algorithms for replicated database maintenance. In Proceedings of 7th ACM Symp. on Operating Systems Principles, 1987.
[3] T. S. Eugene Ng and Hui Zhang Predicting Internet Network Distance with Coordinates-Based Approaches In Proceedings of IEEE INFOCOM'02 New York, June 2002.
[4] A. Ganesh, A.-M. Kermarrec, and L. Massouli´e. Peer to-peer membership management for gossip-based protocols. In IEEE Transactions on Computers, 52(2), February 2003.
[5] A.-M.Kermarrec, L.Massouli'e, and A.J. Ganesh Probabilistic reliable dissemination in large-scale systems. IEEE Transactions on Parallel and Distributed Systems, 14(3), March 2003
[6] Yunhao Liu, Xiaomei Liu, Li Xiao, Lionel Ni, Xiaodong Zhang Location-aware Topology Matching in P2P Systems. In Proceedings of IEEE INFOCOM'04 Hong Kong, March, 2004
[7] Laurent Massouli´, Anne-Marie Kermarrec, Ayalvadi J. Ganesh Network Awareness and Failure Resilience in Self-Organising Overlay Networks In IEEE Symposium on Reliable and Distributed Systems, Florence, 2003
[8] Akihiro Nakao, Larry Peterson and Andy Bavier A Routing Underlay for Overlay Networks In Proceedings of ACM SIGCOMM'03 Karlsruhe, Germany, August, 2003
[9] Sylvia Ratnasamy, Mark Handley, Richard Karp, Scott Shenker Topologically-Aware Overlay Construction and Server Selection. In Proceedings of IEEE INFOCOM'02 New York, June 2002.