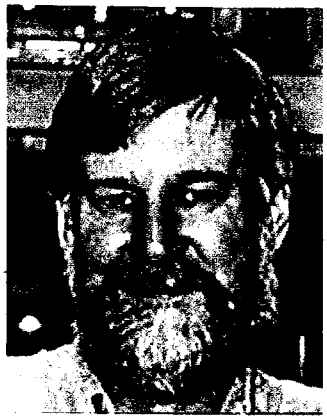
Geo-LPM:
An Efficient Scheme for Locating Nodes in the Internet

Hanh Le, Doan Hoang and Andrew Simmonds, University of Technology, Sydney

Hanh Le

Doan Hoang



Andrew Simmonds

One of the major wea of existing peer-to-pe networks is that their structures do not refl underlying Internet t resulting in unnecess consumption of netw resources. We propos Geographical Longes Matching (Geo-LPM) to self-organise node clusters using IP pre and network metric measurements. Geo-l efficiently locates no produce superior ove while optimising the network resources. I simple, scalable and organising.

Keywords: Peer-to-P Overlay Network, N Location, Internet Infrastructure.

This article is a revi version of a paper o presented to the Aus Telecommunications Networks and Appli Conference (ATNAC Sydney in Decembe

INTRODUCTION

Peer-to-peer (P2P) overlay networks have emerged as highly attractive, decentralised, self-organising and distributed systems. Many useful P2P applications such as distributed file systems [1-4], application-layer multicast [5-8], and event notification services[9, 10] have been developed. P2P overlay networks are normally independent from the Internet infrastructure. Peers communicate with each other regardless of the position or the distance to the other peer. Here the Internet is also called the underlying network.

Some recent efforts have been made to construct overlay networks that are aware of the underlying network infrastructure [11-14]. Others, such as network distance prediction[15, 16], aim at providing physical network topology information to improve performance for wide-scale distributed Internet applications including P2P systems.

One method uses Ping or Traceroute utilities to measure the distance (i.e. propagation delay or latency) between any pair of nodes[17]. This method is not scalable and requires a very high number of measurements (O(N) overhead, where N is the number of nodes).

Another method, called 'landmarking' uses a small number of well-known nodes, called 'landmarks', as a reference frame for other nodes to position them in the Internet. Each node only needs to measure its distance to these landmarks. 'Landmarking' is simple and generates low overhead but might cause hotspots at the landmarks in large-scale systems.

There exist two fundamental problems with the current generation of P2P networks and network distance prediction. Firstly, P2P overlay networks do not adequately take into consideration the physical underlying infrastructure when constructing their overlays. This results in poor utilisation of the underlying network resources and high end-to-end delay for applications. Secondly, present network distance estimation services require a certain level of external information for the setup of landmarks. This makes the resulting systems non-self-organising and prone to the problem of failure due to reliance on the landmarks.

In this paper, we propose a new scheme for quickly and easily locating nodes in the Internet, called **Geographical Longest Prefix Matching** (Geo-LPM). The scheme ingeniously combines two topologically-informative parameters, the IP address and the network distance. It will be shown that Geo-LPM significantly reduces overheads when compared to other schemes that use only the network distance measurement, or even to landmarking schemes [13, 16]. Geo-LPM is distributed, scalable and self-

organising whereas other IP prefix methods are not because of their dependence on centralised servers or external information sources such as Border Gateway Protocol (BGP) routing tables.

LOCATING NODES

Geo-LPM locates nodes to regions (clusters). A cluster consists of nodes that are close to each other in terms of network proximity and IP address prefix (longest common prefix). Clusters are further organised into a hierarchical manner by aggregating IP prefixes.

In Geo-LPM, each cluster has a node that acts as the routing node for the cluster, termed an **'o-router'**. Any node can become an o-router and normally it is the first node that establishes the cluster. After other nodes join the cluster, it is preferable to choose a node which remains online for long periods and has a high bandwidth as the o-router for the cluster.

An o-router does not serve like a 'super-peer' as in a hierarchical P2P system or hybrid P2P system [18, 19] (e.g. to index shared files). Rather, the o-router mainly assists in locating and routing between clusters to optimise the overlay network, a 'super-peer' for a group of peers can be chosen in the formed cluster.

Geo-LPM

We propose that the locating of nodes should be based on the **longest matching prefix** (LPM) and the geography/network proximity/distance between the o-router and other nodes in a cluster. Here, latency is chosen as the distance metric, but depending on the needs of the applications, bandwidth or a combination of different metrics can be substituted for latency as a measure of the distance cost.

A cluster is defined by two essential pieces of information:

1) A node's IP address which provides valuable information about its network membership, and
2) A node's proximity (i.e. latency) which provides information about the node's geographical location.

The idea behind Geo-LPM is that nodes that are in the same physical network and geographically close to each other should belong to the same cluster. The first piece of information is relevant for the following reasons:

1) whenever a node is connected to the Internet, it must have an IP address, either permanently or temporarily for the connection session (dynamic/translated IP address) and
2) nodes belonging to the same (sub-) network have the same (sub-)network address portion. In other words, they share a longest common prefix.

Moreover, the use of Classless Inter-Domain Routing (CIDR) for Internet routing helps arrange the Internet into a hierarchical structure [20, 21]. CIDR was designed to make efficient use of the IP address space and provide efficient Internet routing. CIDR separates networks using a form of variable network masks. We extend CIDR further by using the longest common prefix (LCP) of the IP addresses of joining nodes to partition them into groups. These groups are further examined to determine if they are geographically close enough to form clusters.

The second piece of information expresses the notion of closeness in a physical topological sense, e.g. as measured by latency. We use a predefinced round-trip-time (RTT) threshold between a node and its longest prefix matched cluster to determine whether it can be a member of the cluster.

The essential idea is to employ two parameters to self-organise a cluster: one to efficiently isolate/locate candidate peers for a cluster, and another to make a final decision on whether to join the peers to the cluster by measuring the proximity between the peers and the o-router of the cluster. We believe that latency alone [11-14, 22] is inadequate to create an overlay network that closely matches the topology of the underlying IP network. Nodes that are close in latency may in fact belong to different networks. And if they are classified as neighbours in the overlay network, physical communication network resources will not be used efficiently, since the communication is inter-network rather than intra-network.

On the other hand, nodes with only a LCP may not be close together due to non-contiguous IP address allocation. Geo-LPM addresses these problems and clusters peers appropriately in a self-organising manner.

Geo-LPM locates peers simply and directly, by virtue of CIDR routing and its hierarchical IP address arrangement. Other schemes either require landmarks such as GNP [16], binning [13] and lighthouse [23], or a great number of distant measurements for each new node joining the overlay network to reflect its physical location [11, 23]. TOPLUS [24] and Geo-LPM both make use of the topological information available in the IP

prefix to reduce the overhead. However TOPLUS requires external information input to extract the prefix while Geo-LPM does this by calculating the common prefix of all peers in a cluster. We stress that a peer's IP address is only used for positioning the peer when it first joins the overlay. The Domain Name Service (DNS) plays no part in our overlay construction scheme.

To locate the new node, we propose an IP prefix tree as depicted in Figure 1. The IP prefix tree consists of clusters with their corresponding common prefixes arranged in a CIDR hierarchy. A cluster is a set of nodes that share LCP and are close to each other. The Geo-LPM routing is based on the longest prefix matching (LPM) rule. Clusters at a higher level aggregate the addresses of their child clusters.

Locating a new node

When a node, **X** (step 1 in Figure 1, **X**=00011), joins the overlay network, it contacts any overlay node. This node then forwards the join request from **X** to its o-router. This o-router will check the number of IP common bits between **X** and all IP addresses of other o-routers in its IP tree child list. If there is a cluster that shares longer common bits with **X** than the present cluster, it will forward **X**'s join request to that node.

However, in this case, there is no match between **X**:00011 and cluster:11*, so the o-router at cluster 11* forwards the join request from **X** to its parent (step 2 in Figure 1).

This procedure is repeated until the join request of **X** reaches a cluster that shares the LCP with **X** (step 3 in Figure 1). This is similar to routing in the Internet based on the longest prefix matching (LPM) rule of CIDR[25].

The differentiating feature of Geo-LPM is that the new node will now measure its distance to the o-router of the cluster that it shares the LCP. From this distance value, the new node may join the cluster or create a new cluster, if the distance is smaller or greater respectively than a predefined distance threshold, T.

The PseudoCode of the Geo-LPM can be found in [26].

O-router backup

The departure of a child node will only impact its local cluster. Since an o-router plays a critical role in providing routing functions for the cluster on the overlay, it should be backed up and replicated (at least the routing information part) so that the cluster is vacated only if all nodes in the cluster leave the overlay. It is preferable to choose a node which remains online for long periods and has a high bandwidth as a backup o-router.

When an o-router leaves the overlay, the backup node will replace it and take over the routing functionality for the cluster. The simulation shows that with our clustering method, most of the child nodes still remain in the cluster. In other words, nodes in the cluster share LCP, and are not only close to the o-router but also close to each other.

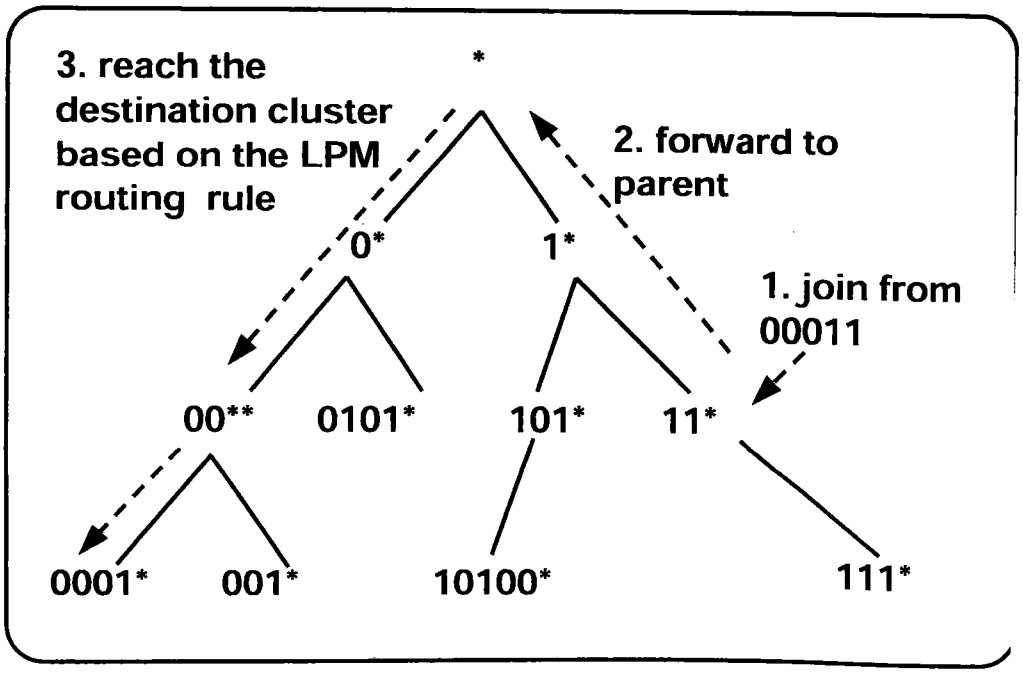


Fig 1 – An example of LPM routing

Cluster size

Geo-LPM reflects closely the underlying network through the proximity metric (geography) and the use of the LPM rule. The size of a Geo-LPM cluster therefore, depends on the number as well as the density of participating nodes in the underlying physical network. To reduce the stress on the o-routers, large clusters can self-divide into smaller clusters by

1) further grouping nodes which have further longer common prefixes; and
2) choosing peers which have high bandwidth connections and low latency as the o-routers of the new clusters when splitting the original cluster.

Significance of Geo-LPM

With the LPM and an appropriate latency threshold, peers in the same cluster often belong to the same physical network. In so doing, our topologically-aware clustering scheme reduces unnecessary inter-physical network communication and hence optimises the use of the underlying network resources by minimising the number of packets travelling over WAN links. As a consequence, communication within a cluster, such as file sharing in SkipNet [27] or a relay multicast server [6-8, 28] to clients within a cluster, does not require packets to cross different physical networks and cause delay unnecessarily.

Geo-LPM ensures that if there is any other live overlay node belonging to the same physical network with the new node, then they will either belong to the same cluster, or become neighbours on the overlay network depending on whether their distance is smaller or greater than the predefined threshold **T**. Geo-LPM also eliminates the migration of a node to different clusters [22]. This is because the node and other clusters do not share a longest common prefix (LCP), which means that they are less likely to be in the same physical network. Therefore, the node should not generate extra traffic (overhead) in measuring its distance to other clusters to which it should not belong.

The threshold value **T** can accommodate a certain fluctuation level of latency measurements. If a node has higher latency to its o-router than **T** because of changes in traffic condition, it may leave the current cluster and form its own cluster. On the other hand, clusters sharing LCP with less than **T** distance can merge together. However, we have not simulated these traffic condition changes in this paper.

It can be seen that our scheme is simpler than existing schemes. It does not require any landmarks or a large number of proximity measurements or external routing information. All o-routers maintain the IP prefix tree by keeping track of their parent and child clusters. In IPv4, the height of the IP prefix tree or the number of join hops will

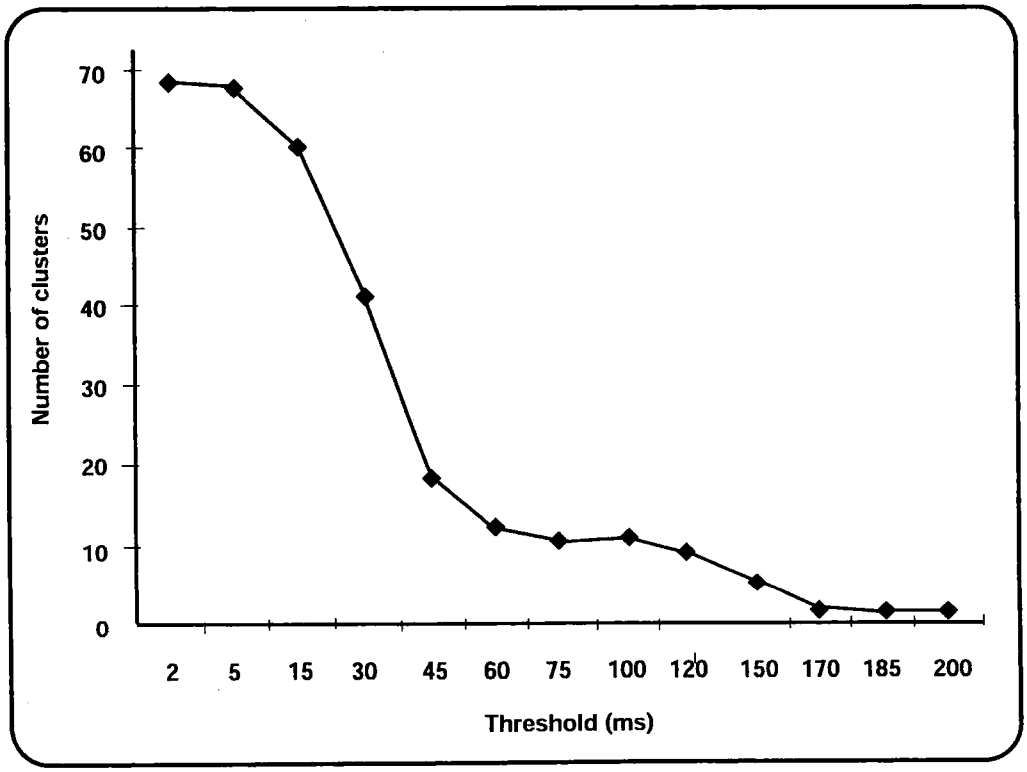


Fig 2 – Number of clusters vs. Threshold

be $32/\log_2 k$, where k is a degree of k-ary tree of IP addresses in the CIDR hierarchy. For example, if k = 16, the join hop order is of 8.

PERFORMANCE EVALUATION

Simulation Setup

We have used the J-Sim [29] network simulator to evaluate Geo-LPM. To simulate the underlying network, we imported transit-stub network topologies generated by GT-ITM network generator [30]. All the underlying network communication is driven by the J-Sim mechanism. Geo-LPM operates at the application layer; each node joining the overlay has an attached application component on top of the basic node class via a UDP port.

All the simulation results were averaged over 20 different runs. For each run, a random bootstrapping node is chosen; nodes join the overlay network in random order.

Simulation Results

Number of clusters vs. different value of latency threshold

Figure 2 shows the relationship between the number of clusters and the latency threshold as a parameter to decide whether the new node should join its LCP cluster or not. We used 68 nodes in the simulation. The threshold, **T**, varies from 2ms to 200ms. As expected, when **T** is negligible, every node forms its own cluster and the number of clusters equals the number of nodes on the overlay. On the other hand, when T is very large, every node will belong to just one cluster.

Effects of CIDR and Geo-LPM on the average distance (latency) between peers

Figure 3 shows the average latency/distance (D_{Avg}) between nodes against the different threshold, **T**. The average distance between nodes is defined as the ratio of the sum of the distances between all peers and their local o-routers as well as between o-routers along CIDR tree to the number of nodes on the overlay.

$$D_{Avg} = \left(\sum_{i=1}^{M} \sum_{j=0}^{ClusterSize} d_{ij} + \sum_{i=1;j=1}^{i=M;j=k} D_{ij} \right) \Big/ N$$

where:

N	the number of nodes;
M	the number of clusters
k	the degree of k-ary tree of IP addresses
d_{ij}	the distance between o-router i to peer j in the cluster i
D_{ij}	the distance between o-router i to all its IP child o-routers (j) along the IP tree
$\sum_{i=1}^{M} \sum_{j=0}^{ClusterSize} d_{ij}$	the sum of the total distance between an o-router and all other peers in the cluster, summed over all clusters.
$\sum_{i=1;j=1}^{i=M;j=k} D_{ij}$	the total distance between o-routers and their IP child o-routers along the IP tree.

This result demonstrates one of the advantages in using Geo-LPM. Geo-LPM

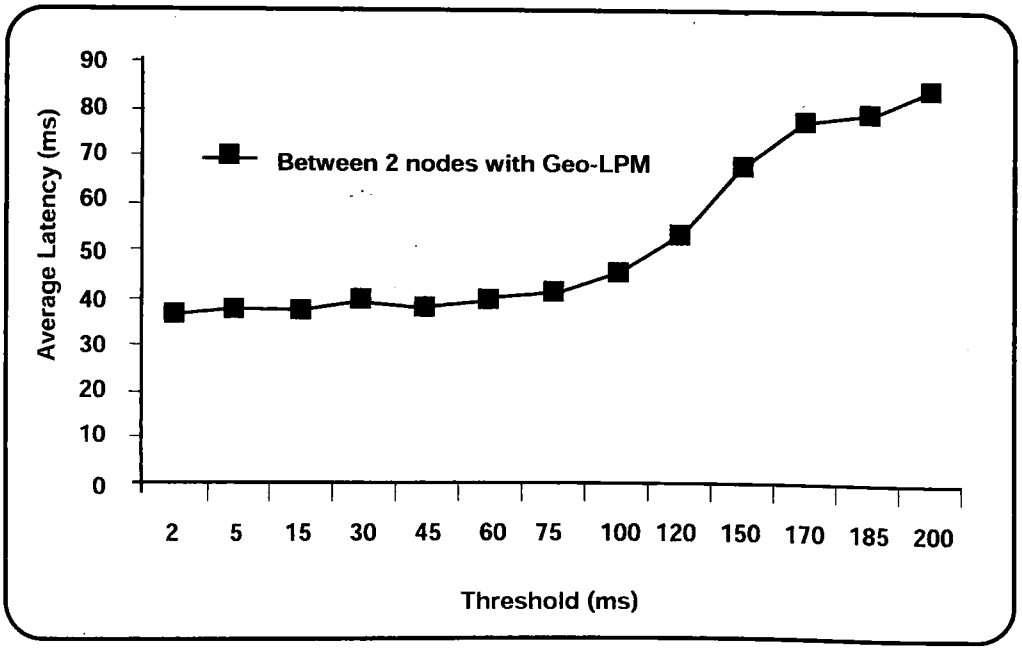


Fig 3 – Effects of Geo-LPM and CIDR on the average distance (latency) between peers

reduces the average distance between nodes to less than 50ms, whereas the average distance between two nodes in many existing P2P networks using pure Distributed Hash Tables [31, 32] is of the order of few hundreds milliseconds.

Effectiveness of Geo-LPM

This measure is defined as the number of peers that can be located correctly over the total number of clustered peers against the threshold **T**. By correct location, we mean all peers in a cluster belong to the same physical network, so that clusters match with physical networks.

As expected, when **T** is too small, there will not be any clustered peers. Routing of Geo-LPM becomes LPM in a CIDR hierarchy. If **T** is too high, one cluster could cover multiple physical networks and incorrect location peers would exist. From Figure 4, when the threshold **T** is around 60 ms, Geo-LPM achieves highest effectiveness with correctly located peers of more than 80% of the total number of nodes, and no incorrect location peers.

Comparisons with the binning technique

Figure 5 shows the effectiveness of the binning method [13] against the number of landmarks by counting the number of peers which have identical bins and are in the same physical networks. To achieve 50 correctly located peers as Geo-LPM, binning requires more than 7 landmarks. However the overhead of Geo-LPM is equal to the one of binning with only two landmarks as indicated in Figure 6. The overhead of binning is calculated by the number of nodes (N) times the number of landmarks (H) times 2 for the ping and pong messages. The overhead of Geo-LPM is defined as the total number of messages which are generated to locate nodes.

We also measure levels of physical matching of binning by calculating the average latency between pairs of nodes which have identical bins. Table 2 describes binning with different numbers of landmarks, number of bins and average distance.

# landmarks	# bins	Avg. Distance
2	2	128.461
3	4	92.923
5	7	73.211
7	11	57.580
9	14	53.432

Table 1: Binning result summary

Geo-LPM reduces the average distance between nodes whilst keeping the overhead very low (as indicated in Figure 6).

The effectiveness of binning depends on the number of landmarks. However landmark setup and maintenance is a restrictive point

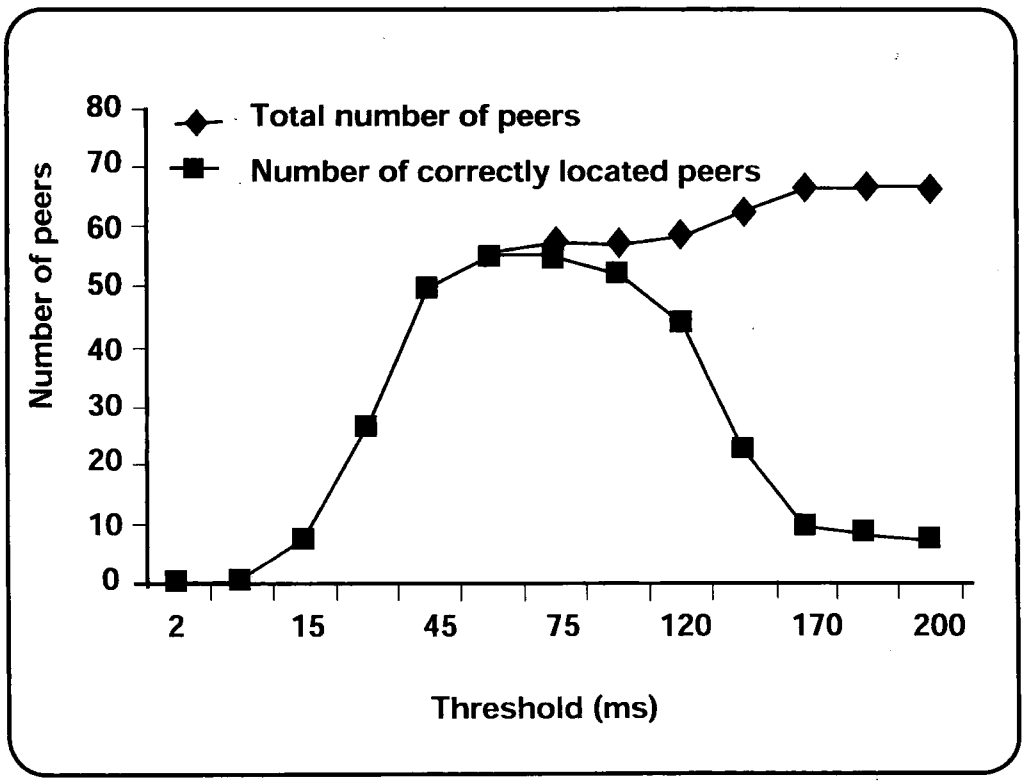


Fig. 4 – Effectiveness of Geo-LPM

of the landmarking technique because the system is prone to the problem of landmark failure. It is remarkable that Geo-LPM has high effectiveness but does not require any external information sources or landmark setup, therefore Geo-LPM is self-organising and adaptive to topology changes.

LPM routing of TOPLUS [24] is a special case of Geo-LPM when the threshold T is very small, every node forms its own cluster and Geo-LPM routing becomes LPM routing.

The major advantages of Geo-LPM over TOPLUS include:

1) Geo-LPM does not use any external information sources such as BGP routing tables;
2) Routing tables of nodes of Geo-LPM are small.
3) Nodes in a cluster can form a 'virtual' node to backup each other and to accommodate a certain level of dynamic node participation and departure. Peers leaving and joining overlays produce only local effects. When T is 60 ms, clusters have an average membership of 5.5.

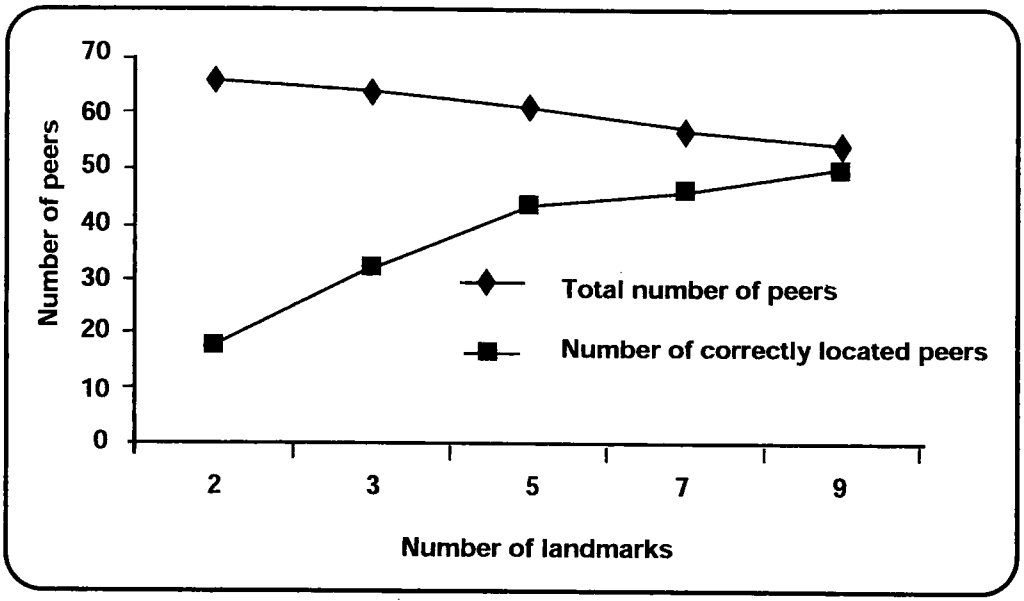


Fig 5 – Effectiveness of Binning

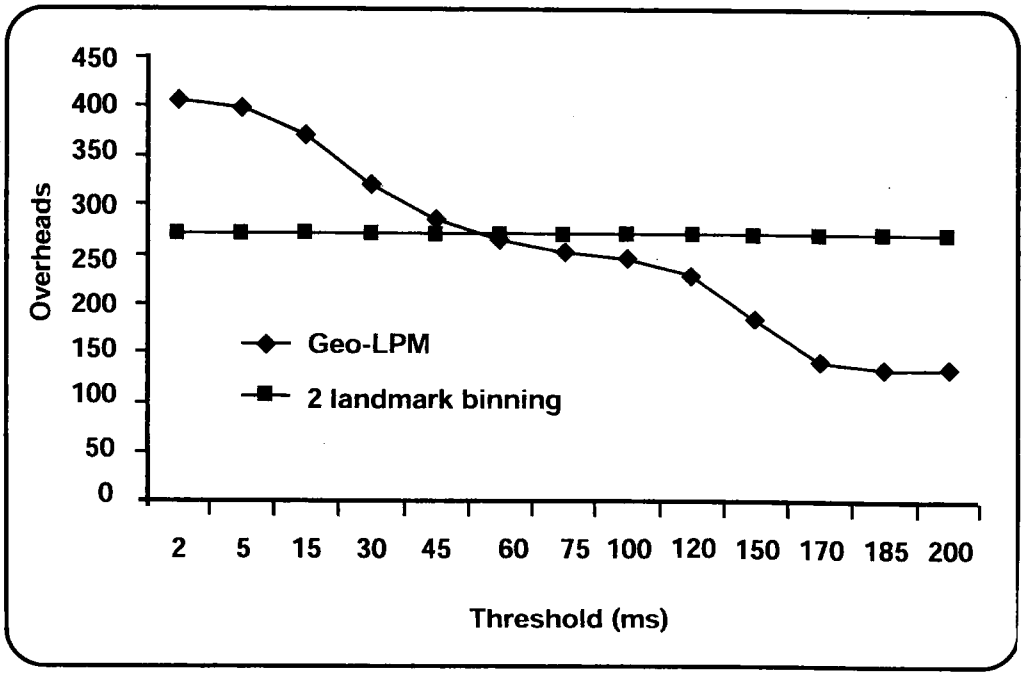


Fig 6 – Overheads

FURTHER WORK

Because of non-contiguous IP address allocation, clusters that share a longest common prefix may not be geographically close to each other. In a future paper, we will present how an o-router of a new cluster could explore the underlying network infrastructure to find its location. All other peers in the cluster will follow the local o-router, so this would be a scalable way to geographically locate nodes.

CONCLUSION

Current P2P overlay networks are faced with the underlying network mismatching problem that causes high end-to-end latency and inefficient network resource usage. We responded to this challenging performance issue by proposing the simple Geo-LPM scheme for effectively locating nodes in the Internet.

Geo-LPM creates clusters that closely match the physical network by considering both the addressing/routing scheme of the Internet, and the network proximity/geography. As a result, a new node can locate itself quickly without using landmarks or a large number of proximity measurements.

Our Node Location Scheme of Geo-LPM is shown to be a simple and effective way of locating nodes in the Internet with respect to their geographical location and their network membership. Geo-LPM is decentralised and self-organising with low overhead.

REFERENCES

[1]. F. Dabeck, M. F. Kasshoek, D. Karger, R. Morris, and I. Stoica, "Wide-are cooperative storage with cfs," In *18th ACM Symposium on Operating Systems Principles*, 2001.

[2]. P. Druschel and A. Rowstron, "PAST: a large-scale, persistent peer-to-peer storage utility" *Proceedings of the Eighth Workshop on Hot Topics in Operating Systems*, 2001.

[3]. J. Kubiatowicz, "Oceanstore: An architecture for global-scale persistent storage," *Proc. Of ASPLOS*, 2000.

[4]. S. Q. Zhuang, B. Y. Zhao, A. D. Joseph, R. Katz, and J. Kubiatowicz, "Bayeux: An architecture for scalable and fault-tolerant wide-area data dissemination," *11th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV 2001)*, 2001.

[5]. S. Ratnasamy, M. Handley, R. Karp, and S. Shenker, "Application-level multicast using content-addressable networks," *Proc. 3rd International Workshop on Networked Group Communication*, 2001.

[6]. S. Banerjee, B. Bhattacharjee, and C. Kommareddy, "Scalable application layer multicast," *Proc. ACM Sigcomm*, 2002.

[7]. J. Jannotti, D. Gifford, K. Johnson, M. Kaashoek, and J. O'Toole, "Overcast: Reliable Multicasting with an Overlay Network," *Proc. 4th Symposium on Operating Systems Design and Implementation*, 2000.

[8]. D. Pendarakis, D. V. S. Shi, and M. Waldvogel, "ALMI: An Application Level Multicast Infrastructure," *Proc. 3rd Usenix Symposium on Internet Technologies & Systems*, 2001.

[9]. A. Rowstron, A.-M. Kermarrec, M. Castro, and P. Druschel., "Scribe: The Design of a Large-Scale Event Notication Infrastructure," *Proc. of the 3rd Int. Workshop on Networked Group Communication (NGC'01)*, 2001.

[10]. L. F. Cabrera, M. B. Jones, and M. Theimer, "Herald: Achieving a Global Event Notication Service," *Proc. of the 8th Workshop on Hot Topics in Operating Systems*, Elmau, Germany, 2001.

[11]. M. Waldvogel and R. Rinaldi, "Efficient Topology-Aware Overlay Network," in *Proceedings of ACM HotNets-I*, 2002.

[12]. B. Y. Zhao, J. Kubiatowicz, and A. Joseph, "Tapestry: An infrastructure for fault-tolerant wide-area location and routing," *University of California at Berkeley, Computer Science Department Tech. Rep. UCB/CSD-01-1141*, 2001.

[13]. S. Ratnasamy, M. Handley, R. Karp, and S. Shenker, "Topologically-aware overlay construction and server selection," *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, 2002.

[14]. M. Castro, P. Druschel, Y. C. Hu, and A. Rowstron, "Topology-aware routing in structured peer-to-peer overlay networks," *FuDiCo 2002: International Workshop on Future Directions in Distributed Computing*, University of Bologna Residential Center Bertinoro (Forli), Italy, 2002.

[15]. T. S. E. Ng and H. Zhang, "Predicting Internet network distance with

coordinates-based approaches," INFOCOM 2002. *Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, 2002.

[16]. T. S. E. Ng and H. Zhang, "Towards global network positioning," *Proceedings of the First ACM SIGCOMM Workshop on Internet Measurement*, 2001.

[17]. R. Govindan and H. Tangmunarunkit, "Heuristics for Internet Map discovery," *Proceedings of INFOCOM*, 2000.

[18]. B. Yang and H. Garcia-Molina, "Designing a super-peer network," *Proceedings of 19th International Conference on Data Engineering*, 2003.

[19]. B. Yang and H. Garcia-Molina, "Comparing Hybrid Peer-to-Peer Systems," *Proceedings of the 27th International Conference on Very Large Data Bases*, 2001.

[20]. RFC1518, "An Architecture for IP Address Allocation with CIDR," Web page **http://www.faqs.org/rfcs/rfc1518.htm**, 1993.

[21]. RFC1519, "Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy.," Web page **http://www.faqs.org/rfcs/rfc1519.htm**, 1993.

[22]. S. Jain, R. Mahajan, D. Wetherall, and G. Borriello, "Scalable Self-Organizing Overlays," *Computer Science and Engineering, University of Washington UW-CSE 02-06-04*, 2002.

[23]. M. Pias, J. Crowcroft, S. Wilbur, T. Harris, and S. Bhatti, "Lighthouses for Scalable Distributed Location," in *Proc IPTPS '03*, 2003.

[24]. L. Garces-Erce, K. W. Ross, E. Biersack, P. Felber, and G. Urvoy-Keller, "TOPLUS: Topology Centric Lookup Service," *Fifth International Workshop on Networked Group Communications (NGC'03)*, Munich, 2003.

[25]. S. Halabi and D. McPherson, Internet Routing Architectures, Second Edition ed: Cisco Press, 2000.

[26]. H. Le, D. Hoang, and A. Simmonds, "An Efficient Scheme for Locating Nodes in the Internet," *Australian Telecommunications Networks & Applications Conference (ATNAC 2004)*, Sydney, Australia, 2004.

[27]. N. Harvey, M. Jones, S. Saroiu, M. Theimer, and A. Wolman, "SkipNet: A Scalable Overlay Network with Practical Locality Properties," *USITS 2003*, Seattle WA, 2003.

[28]. D. A. Tran, K. A. Hua, and T. T. Do, "ZIGZAG: An Efficient Peer-to-Peer Scheme for Media Streaming," in *Proceedings of the 22nd Conference of IEEE Communications Society (INFOCOM 2003)*, San Francisco, USA, 2003.

[29]. H.-y. Tyan, "J-Sim network simulator," The Ohio State University. Web page. **http://www.j-sim.org/**, 2002.

[30]. GT-ITM, "Georgia Tech Internetwork Topology Models," Web page: **http://www.cc.gatech.edu/projects/gtitm/** 1997.

[31]. I. Stoica, R. Morris, D. Liben-Nowell, D. R. Karger, M. F. Kaashoek, F. Dabek, and H. Balakrishnan, "Chord: a scalable peer-to-peer lookup protocol for Internet applications," *IEEE/ACM Transactions on Networking, vol. 11, pp. 17-32*, 2003.

[32]. S. Ratnasamy, P. Francis, M. Handley, and R. Karp, "A Scalable Content-Addressable Network," *ACM SIGCOMM*, 2001.

GLOSSARY

BGP: Border Gateway Protocol
CIDR: Classless Inter-Domain Routing
IP: Internet Protocol
LCP: Longest Common Prefix
LPM: Longest Prefix Matching
P2P: Peer-to-Peer
RTT: Round Trip Time

THE AUTHORS

Hanh Le **is currently a PhD student at University of Technology, Sydney (UTS), in the faculty of Information Technology. She holds a first class honours degree in Information Technology from Post and Telecom Institute of Technology, Vietnam. Her research interests are Peer-to-Peer Overlay Networks, Internet Node Positioning and Mobile Networks.**

Doan B. Hoang **is a Professor of Computer Networks in the Department of Computer Systems, Faculty of Information Technology, UTS. He received his PhD in electrical and computer engineering from the University of Newcastle. Before joining UTS, he was with Basser Department of Computer Science, the University of Sydney. He has also held various visiting**

Continued page 37