# $Highways$: Proximity Clustering for Scalable Peer-to-Peer Network

Eng Keong Lua
University of Cambridge
15 JJ Thomson Avenue
Cambridge CB3 0FD
eng.keong-lua@cl.cam.ac.uk

Jon Crowcroft
University of Cambridge
15 JJ Thomson Avenue
Cambridge CB3 0FD
jon.crowcroft@cl.cam.ac.uk

Marcelo Pias
Intel Research Laboratory
15 JJ Thomson Avenue
Cambridge CB3 0FD
marcelo.pias@intel.com

## Abstract

*The "location-aware" construction of overlay networks requires the identification of nodes that are efficient with respect to network delay and available bandwidth. In this short paper, we propose* $Highways$ *to create clusters of nodes using a novel "location-aware" method, based on a scalable and distributed network coordinate system. This helps to build overlay routing tables to achieve better proximity accuracy, thus, providing a mechanism to boost performance in application overlay routing.*

## 1. Introduction

Peer-to-peer (P2P) overlay networks such as Pastry, Tapestry, Chord and CAN [2] are scalable, decentralized, and self-organizing. The basic functionality of that these Structured P2P systems provide is a Distributed Hash Table (DHT). P2P nodes have routing tables comprised of neighbors chosen based on other criteria in addition to identifiers; most notably, network proximity metrics (i.e. latency) have been used to select neighbors. Given a set of neighbors and a destination, the routing algorithm determines the next hops based on proximity. Despite the elegance from a theoretical perspective, these systems have some practical limitations. Firstly, they rely on application-level routing that largely ignores the characteristics of the underlying network topology resulting in significant routing penalty. Second, number of hops have been used as a mechanism to measure path length (distance). We observed that a few hops do not directly translate to low network latency (and vice-versa). This can lead to routing paths with significant latency and unnecessary waste of bandwidth. Suppose a message needs to be routed from node A (in Berkeley) to node B (in Boston) in the Chord ring, and two paths can be chosen. One has length of 5 hops and the other has length of 3 hops. The optimal choice is to select the path with lower cost (shortest path), in terms of number of hops, thus, the 3 hops path is selected. However, it turns out that due to the random distribution of nodes in the ID space, this path may have intermediate routes in Europe; whereas the path with 5 hops may have direct routes in North America. Therefore, a message has to traverse a transatlantic link (average minimum round trip latency of 60 *ms*). This example illustrates a significant high network delay penalty but low hop count cost. Pastry and Tapestry routing tables are constructed with downstream links pointing to nodes close by, but choices of nearby nodes are limited to a small set of nodes available at a given moment. CAN uses a landmark ordering and binning scheme to cluster nodes that are close in the underlying network during the CAN construction. This approach may cause significant imbalance in the nodes distribution that can lead to hotspots. Chord uses finger table to perform measurements and select the next hop heuristically. Similarly, the choices are restricted to entries in the finger table.

## 2. $Highways$ Network Architecture

The approach we propose is generic and applicable to a number of P2P overlay networks. Our goal is to reduce P2P system maintenance cost, and to provide avenues to adjust the system to suit the application needs and the underlying network conditions. For instance, P2P nodes maintain their neighbor list with *k* entries based on proximity, requires network measurements for each entry and entries in their neighbors' neighbor list. This generates substantial volume of maintenance overheads. To overcome this problem, network coordinate system [1, 3] can be devised to assign each node with synthetic coordinates using a network distance metric space (e.g. latency). These coordinates can be computed and updated by measuring the distance to a small selective set of *Beacons* (Landmarks or Lighthouses). Instead of performing a series of greedy walks in the overlay towards the closest node by computing rough coordinates from a set of random nodes, our contribution is to construct distributed network coordinates by a clustering scheme of *Beacons* and nodes. We argue that this approach achieves better proximity accuracy and reduce overheads through the

identification of the cluster areas that are close using *Beacon's* coordinates, thus, giving a choice of close by nodes in these cluster areas. Entries in a node's routing table are updated to a set of chosen nodes, among all live P2P overlay nodes in close by cluster areas using their *Beacons* as pointers. These *Highways* metrics are accessible through *cross-layer* abstractions and interactions. The *Beacons* serve two purposes: (1) to act as clusters landmarks and resolve the routing destinations and (2) to propagate routing information when nodes join or leave the network. Hence, *Highways* key concepts: (1) Construct *Beacons* and nodes clusters and topologies, (2) Implement flexibility, such as load balancing mechanism and distributed network coordinate system, (3) Archive relevant system information, that are easy to update and retrieve, for building of overlay routing table efficiently, (4) Adjust dynamically with (2) and (3), according to network conditions and application behavior. In *Highways*, a cluster defines an embedding space (i.e. coordinate system) that can be easily (linearly) mapped to another embedding space derived from a different cluster. We define a metric space: A pair (X,D) where X is a set of points and D:XxX$\rightarrow$[0,$\infty$) is a distance function satisfying the following conditions for all x,y$\in$X: (1) D(x,y)=0 iff x=y; (2) D(x,y)=D(y,x) (Symmetry); (3) D(x,y) + D(y,z) = D(x,z) (Triangle Inequality). Isometry is defined as a mapping f :X$\rightarrow$ $X^{'}$, where (X,D) and ($X^{'}$,$D^{'}$) are metric spaces, with $D^{'}$(f(x),f(y)) = D(x,y) for all x,y$\in$X. To achieve the desired performance, we propose an integrated metric classification function $x_{metric}(s_i)$ that combines the network latency and available bandwidth. This metric function can be defined as: $x_{metric}(s_i) = f(w_d.d_i * w_{bw}.bw_i)$, with $d_i$ is the measured minimum or median network latency from node $i$ to *Beacons*, $bw_i$ is the measured maximum available bandwidth for node $i$, $w_d$ and $w_{bw}$ are the respective weights, where $*$ denotes combined relationship (e.g. to have an optimized criteria of minimum network latency and maximum available bandwidth, we will have $d_i/bw_i$, which will give the smallest value for minimum network latency and maximum available bandwidth). Thus, node $i \in \{1, ..., m\}$, has a $n$-dimensional distance vector $D_i^{'} = [x_{metric}(s_{i1}), ..., x_{metric}(s_{in})]^T$. The overall coordinate system is represented by $m$ x $n$ distance matrix where **A** is symmetric with zero diagonal entries. Dimensionality reduction to $k$ is done using Singular Value Decomposition (SVD) or QR factorization. For a geometric space of dimensionality $k$, this should have **at least** $k+1$ *Beacons*. SVD factors $k$ x $k$ matrix **A** into the product of three matrices U, W and $V^T$. By using the first $k$ columns of U denoted by $U_k$, we project the $n$-dimensional space into a new $k$-dimensional space: where $x_i = U_k^T.D_i^{'}$ is the new coordinates of node $i$ after dimensionality reduction. To minimize the discrepancy between the distance represented in the coordinates system and the measured distance between $n$ *Beacons*, we defined and used a scaling factor [1]: $\alpha_k = \frac{\sum_i^n \sum_j^n x_{metric}(s_{ij}).L_{ij}}{\sum_i^n \sum_j^n L_{ij}^2}$ where $L_{ij} = l_2(U_k^T.D_i^{'}, U_k^T.D_j^{'})$, $l_2$ is Euclidean norm. We conducted experiments to analyze the estimation accuracy of clustering nodes and *Beacons*. We used minimum RTT and maximum available bandwidth data between all pairs of PlanetLab[1] sites. These data were aggregated over a 7-day period from March 22-28, 2004, and there were 325 nodes distributed in 139 sites. Experimental parameters include: $m$=$\{$139,60,30$\}$ nodes, these nodes $RTT$ <$\{$320,60$\}$ *ms*, sets of *Beacons* ($n$=$\{$139,60,40,30,15$\}$), based on various proximity RTT bins ($Proximity$ <$\{$320,200,100,40,30$\}$) between nodes and *Beacons* for each cluster, and $k$=10 (dimensionality reduction). We defined the relative estimation error $\epsilon$ is defined as $\sum_{i,j \in 1,...,m} \frac{|x_{metric}(s_{ij}) - l_2(x_i,x_j)|}{x_{metric}(s_{ij})}$. Preliminary results are shown in Figure 1, illustrating the CDF of the relative estimation error. This results suggest that *Highways* clustering of *Beacons* and nodes with *close proximity bins*, give improved accuracy for efficient P2P network routing. Available bandwidth metric results show similar trends. Further experiments to evaluate the proposed integrated metric classification are required and carried out.

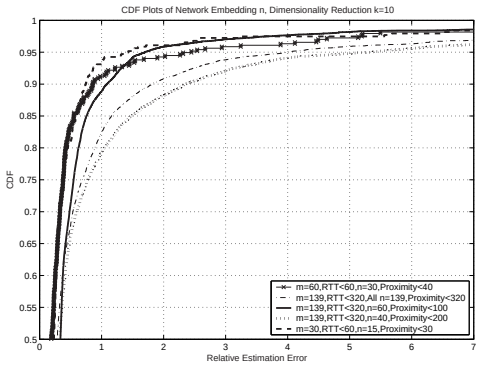


**Figure 1. Proximity Clustering Accuracy for *Highways*.**

## References


[1] H. Lim, J. Hou, and C. Choi. Constructing internet coordinate system based on delay measurement. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, Miami, Florida, USA, October 2003.

[2] E. K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim. A survey and comparison of peer-to-peer network schemes. *In Submission to IEEE Communications Survey*, March 2004.

[3] M. Pias, J. Crowcroft, S. Wilbur, T. Harris, and S. Bhatti. Lighthouses for scalable distributed location. In *2nd Int. Workshop on Peer-to-Peer Systems*, October 2003.


1 PlanetLab is an open, globally distributed platform for developing, deploying and accessing planetary-scale network services. Latency and bandwidth data used are available in http://www.planet-lab.org