

Canicula: An Improved Hybrid Overlay Networks

Yang Chen

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
chenyang04@mails.tsinghua.edu.cn

Abstract - Hybrid Overlay Network (HONet) is a locality-aware hybrid overlay architecture, which combines the scalability and simplicity of a structured overlay with the connection flexibility of an unstructured overlay. In this paper, we propose Canicula, an improved HONet. First, using a simple and accurate network coordinate system, Canicula achieves more reliable node clustering. Secondly, Canicula uses an enhanced overlay construction method to reduce the impact of guarded hosts. The experimental results on over 450 PlanetLab nodes show that the ARDP in Canicula is only 45% of that in flat bidirectional Chord.

1. INTRODUCTION

To take advantage of both structured overlay and unstructured overlay, Tian *et al.* propose Hybrid Overlay Networks (HONet) [3]. HONet integrates the regularity of structured overlays with the flexibility of unstructured overlays by a hierarchical architecture.

Guarded hosts are hosts which can not accept incoming connections. The existence of *guard hosts* challenges overlay construction because not all hosts are capable of receiving and forwarding requests. The design of HONet does not consider the impact of *guarded hosts*. Moreover, the node distance prediction system in HONet is not accurate. In this paper we propose Canicula, an improved Hybrid Overlay Networks. First, we use triangulated heuristic [4] to predict the network distance. Then nodes self-organize into structured clusters based on our locality-aware clustering algorithm. Secondly, we propose an enhanced overlay construction method to reduce the impact of guarded hosts.

2. CANICULA ARCHITECTURE

2.1 Network Distance Prediction

Just like in HONet, an important step in Canicula is node clustering, which requires the knowledge of nodes location. We employ a simple coordinate system to identify a node's network location, which uses a set of round-trip time from each node to a group of well-known landmark nodes. Any stable nodes which are able to response ICMP ping message can be chosen as landmark nodes.

After getting the network coordinate, we can predict the distance between two nodes. In HONet, Hilbert curves [6] are used to map the coordinates into numbers, called locality number (L number), and then the distance between two nodes is defined as the difference between two locality numbers. Because there is loss of information in mapping from n-

dimensional space to one-dimensional space, two nodes which are close to each other in n-dimensional space may be far from each other after being mapped to one-dimensional space.

Canicula uses triangulated heuristic to predict the network distance for both simplicity and accuracy. In [4], T. S. Eugene Ng and Hui Zhang show by large-scale network measurements that the triangulated heuristic actually achieves good accuracy.

2.2 Node Clustering

After getting the network distance between nodes, we cluster the nodes. In real network applications, nodes can join or leave the network at any time, which will cause a heavy burden on centralized system. Therefore, we use a distributed clustering algorithm. To join the overlay, a node first identifies its coordinates in the network. Then it uses triangulated heuristic to find the closest cluster root. In our design, if the new node cannot locate nearby cluster roots, or its distance to the nearest cluster root is larger than a threshold T , this node joins the backbone network as a new cluster root and announces its coordinates in the backbone network. Otherwise, the node joins the cluster led by its nearest cluster root. There may be some outliers with every dimension of their coordinates bigger than T . We call these nodes "island". Instead of joining in the backbone, these islands join the cluster lead by its nearest cluster root.

Because we design our algorithm in a distributed way, no end-to-end measurement is performed between all pairs of nodes. This design reduces the overhead for large scale network and provides Canicula with high scalability.

2.3 Overlay Construction under Limited End-to-End Reachability

2.3.1. Existence of Guarded Hosts in Internet

Most of current network-overlay construction assumes two-way communication capability: each host can initiate outgoing connections as well as accepting incoming connections. But this assumption is not always true especially due to the use of Network Address Translation (NAT) and firewalls. In [7], experiments on eDonkey and Gnutella file-sharing systems reveal that as many as 36% of the hosts may be guarded.

2.3.2. Overlay Construction in Canicula

Wang *et al.* proposes an overlay optimization called e^* to help existing overlay protocols overcome the reachability problem [8]. In our design, we use similar idea as e^* to construct overlay under limited end-to-end reachability.

A natural approach to integrate guarded hosts into an overlay is to group them into clusters by locality, and then assign an open host as the root of each cluster. To achieve reasonable performance, cluster roots must be carefully selected to meet several criteria, such as the speed of network connection or unicast latencies to other nodes in the cluster.

First of all, we must check whether a host is guarded or not. A host sends query messages to selected overlay nodes with a callback bit set in each message. Upon receiving such messages, an overlay node attempts to connect to the caller. If any of these callbacks succeeds, the callback bit will be cleared in the following queries, and this host considers itself as open. If all of the requested callbacks fail to return, the host recognizes itself as guarded. Also, whenever one host is connected by another host, it recognizes itself as open.

Then, we employ a Root Election Protocol. Root election in a cluster is based on a Root Rank Vector (RRV). Each overlay node has its own RRV and each element in the RRV is a test condition for root election. In Canicula, a typical RRV is as follows.

$$RRV = \langle open, lifetime, cluster \ dist \rangle \quad (4)$$

In RRV, *open* represents whether one host is guarded; *lifetime* represents how long the host has stayed in the overlay; *cluster dist* represents the summation of latencies to all the other members in the cluster.

Each overlay node is responsible for keeping its RRV up-to-date. A node periodically updates its RRV by active and passive probing. The computation of cluster dist requires latencies to other cluster members, which can be measured by the periodically exchanged “heartbeat” messages without extra measurement overhead. Each cluster member includes RRVs as part of its “heartbeat” messages to others within the same cluster. The received RRVs are then sorted in the order of the three elements with *open* having the highest priority and *cluster dist* the lowest. The node with the top ranked RRV is elected as the root.

2.4 Basic Structured Overlay and Message Routing

In Canicula, we use bidirectional Chord as our basic structured overlay. An important reason for choosing bidirectional overlay is the ubiquity of the guarded hosts. Without using bidirectional overlay, the guarded hosts will be unreachable. The overlay routing of bidirectional Chord is presented in [5].

If a message is destined for a local cluster node, normal structured overlay routing presented in [8] is used. Otherwise, we will use two approaches, hierarchical routing and fast routing. In hierarchical routing, messages are delivered from

one cluster to another through the backbone network. Fast routing utilizes the random connections between clusters as inter-cluster routing shortcuts. To implement fast routing, each cluster nodes announces information about its inter-cluster links in the local cluster DHT.

3. SIMULATION AND EXPERIMENT

The main performance metric we adopt is the Average Relative Delay Penalty (ARDP). Here we distinguish the latencies in an overlay, which we call overlay latencies, from the unicast latencies, which are the latencies in the underlying physical network. Smaller ARDP indicates that most overlay latencies are close to the respective unicast latencies.

3.1 Setup of Simulation and Experiment

In simulation, a transit-stub network topology with 287 transition nodes and 3,000 stub nodes is generated for our simulation. 300 PlanetLab nodes are used to help topology construction. We assign different distances to the edges in the topologies: the distance of intra-stub edges is 1; the distance of the edges between transition node and stub node is a random integer within [2; 15]; and the distance between transition nodes is measured from the distance matrix. T is set as 40 ms in our simulation.

We use about 450 PlanetLab nodes for our experiment.

3.2 Performance Evaluation on the Generated Topology

Since the random connections are an import factor affecting the performance of Canicula, we compare the performance of Canicula by varying the number of random connections on each node (RC = 1 or 2).

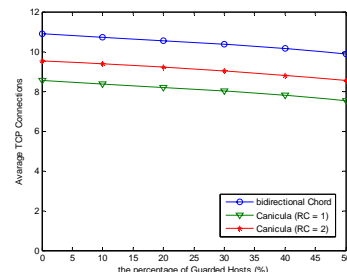


Fig. 3. Average TCP Connections

Fig.3 shows the average number of the TCP connections in each node with the increasing guarded hosts. Generally, an overlay with a larger number of links produces lower ARDP but consumes more network resources. On the other hand, as shown later, with the same number of links, the ARDP of an overlay depends on its path optimization algorithm. From Fig.3 we find that in our simulation, when the number of random connections is 1 or 2, the average number of the TCP connections of each node in the Canicula overlay is less than that in the flat bidirectional Chord.

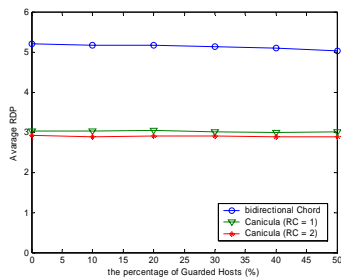


Fig. 4. Average RDP

Fig.4 shows the comparison of ARDP between Canicula and flat bidirectional Chord. The ARDP in Canicula is much smaller than that in flat bidirectional Chord because Chord does not consider the network locality, though Canicula uses less TCP connections as shown in Fig.3. When we set RC to 2 in Canicula, the ARDP is only 55% of that in the flat bidirectional Chord. Our simulation results show that hierarchical structured overlays outperform flat structures.

3.3 Experiment on PlanetLab

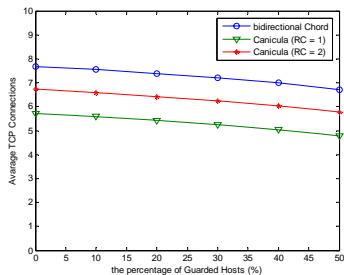


Fig. 5. Average TCP Connections

Fig.5 shows the average number of TCP connections in each node with the increasing guarded hosts. When the number of random connections is 1 or 2, the average number of TCP connections of each node in the Canicula overlay is less than that in the flat bidirectional Chord.

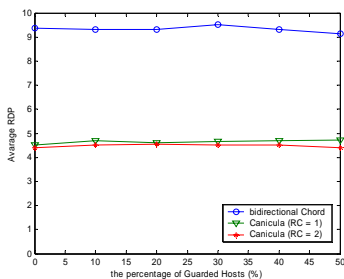


Fig. 6. Average RDP

Fig.6 shows the comparison of ARDP between Canicula and flat bidirectional Chord. When RC is set to 2 in Canicula, the ARDP is only 45% of that in the flat bidirectional Chord. Our experiment results show that hierarchical structured overlays perform better than flat structures.

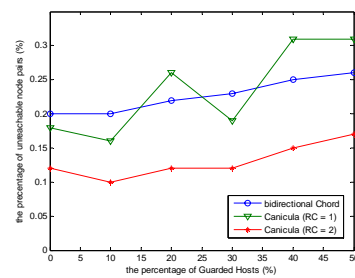


Fig. 7. Unreachable node pairs

Fig.7 shows that given the percentage of guarded hosts below 50%, the percentage of the unreachable node pairs is less than 0.32%. It means that most node pairs are reachable between each other through the overlay. And Canicula (RC = 2) achieves better reachability than flat bidirectional Chord.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose Canicula, an improved Hybrid Overlay Network. According to our simulation and our experiment on PlanetLab, Canicula performs well in Internet, even when up to 50% of the hosts are guarded. It achieves much lower ARDP than flat bidirectional Chord, without consuming more network resources. Canicula also has better overall reachability with less than 0.32% node pairs unreachable between each other through the overlay.

5. REFERENCES

- [1] Ion Stoica, Robert Morris, David Karger and M. Frans Kaashoek, Hari Balakrishnan. "Chord: A scalable peer-to-peer lookup service for internet applications". In *Proc of ACM SIGCOMM'01*, 2001
- [2] Xin Yan Zhang, Qian Zhang, Zhensheng Zhang, *et al.* "A Construction of Locality-Aware Overlay Network: mOverlay and Its Performance". *IEEE Journal on Selected Areas in Communications*, Vol.22, No.1, Pages 18-28, 2004.
- [3] Tian Ruixiong, Xiong Yongqiang, Zhang Qian, Li Bo, Zhao Ben Y., Li Xing, "Hybrid Overlay Structure Based on Random Walk". In *Proc of the 4th International Workshop on Peer-To-Peer Systems (IPTPS'05)*, 2005.
- [4] T. S. Eugene Ng and Hui Zhang. "Predicting Internet network distance with coordinates-based approaches". In *Proc of IEEE INFOCOMM'02*, 2002.
- [5] Prasanna Ganesan, Gurmeet Singh Manku. "Optimal Routing in Chord". In *Proc of 15th Annual ACM-SIAM Symposium on Discrete Algorithms(SODA'04)*, 2004.
- [6] Tetsuo Asano, Desh Ranjan, Thomas Roos, *et al.* "Space-filling curves and their use in the design of geometric data structures". *Theoretical Computer Science*, v181n1, Pages 3-15, 1997.
- [7] Wenjie Wang, Hyunseok Chang, Amgad Zeitoun and Sugih Jamin. "Characterizing Guarded Hosts in Peer-to-Peer File Sharing Systems". In *Proc of IEEE GLOBECOM'04*, 2004.
- [8] Wenjie Wang, Cheng Jin and Sugih Jamin. "Network Overlay Construction under Limited End-to-End Reachability". In *Proc of IEEE INFOCOMM'05*, 2005.