Computer Communications 32 (2009) 257–267

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Fast and proximity-aware multi-source overlay multicast under heterogeneous environment

Zhenyu Li [a,b,*], Zengyang Zhu [a,b], Gaogang Xie [a], Zhongcheng Li [a]

[a] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, PR China
[b] Graduate School of Chinese Academy of Sciences, Beijing 100049, PR China

ARTICLE INFO

Article history:
Received 22 July 2008
Received in revised form 13 October 2008
Accepted 14 October 2008
Available online 25 October 2008

Keywords:
Overlay multicast
Multi-source
Heterogeneous environment
Hierarchical structure
Probabilistic forwarding

ABSTRACT

Overlay multicast has been considered as one of the most important developments for the next generation Internet infrastructure. In this paper, we consider overlay multicast in the scenarios where any participant node is a potential data source. Existing multicast algorithms for single-source always require a long time to deliver messages or have high maintenance overhead when multiple data sources are allowed. There are other algorithms that are designed for multi-source scenarios. But they consume too much network resources and have a long convergence time because of proximity ignorance. To address the issues, we present *FPCast*, which leverages node heterogeneity and proximity information at the same time. Physically close nodes are grouped into clusters and each cluster selects a powerful, stable node as its rendezvous point. The rendezvous nodes form a DHT-based structure. Data messages are replicated and forwarded along implicit, source specific, and heterogeneity-aware multicast trees. We further reduce the delivery delay by introducing probabilistic forwarding scheme. We show the average delivery path length converges to $O(\log n)$ automatically (n is the number of nodes in the overlay). The simulation results demonstrate the superiority of our algorithm in terms of message delivery time and network resource consumption, in comparison with the previous randomized algorithms. The algorithm is also resilient to node failures.

1. Introduction

Overlay multicast has been considered as a promising alternative to the un-widely deployed IP multicast and is one of the most important developments for the next generation Internet. In this paper, we study overlay multicast to support the applications in which there are multiple data sources. The applications include distributed multi-player games, group communication for large-scale systems, teleconferencing, remote collaboration, etc. Since these applications always consist of a large number of dynamic nodes which are heterogeneous in terms of node capacity and spread over the entire Internet, it is a challenge work to design an efficient multi-source overlay multicast algorithm. In our opinion, an efficient multi-source overlay multicast algorithm should meet at least the following basic requirements.

- *Scalable performance.* The algorithm should be decentralized and support node churns (i.e. joins, graceful departures and failures). As the system size grows, the efficiency should degrade gracefully.
- *Low delay.* Some applications (e.g. group communication) have soft-time constraints while others (e.g. distributed games and teleconferencing) are interactive applications. Lower delay is the main design objective.
- *Proximity-aware.* In order to reduce the network resource consumption and enable fast message delivery, the algorithm should take node proximity information into consideration. The proximity of two nodes is measured by the latency in the physical IP topology. The closer two nodes are in the IP topology, the more proximate they are.
- *Heterogeneity awareness.* Nodes have different capacities. The algorithm should account for node heterogeneity in terms of capacity so that the load on a node is proportional to its capacity (i.e. achieving a load balance state). Otherwise, some nodes are overloaded while others are underloaded.
- *Reasonable network resource consumption.* Both the overlay maintenance overhead and the cost for message delivery should be small.

The single-source overlay multicast algorithms can be classified into two categories: tree-based algorithms [5,24] and data-driven algorithms [30,15]. Tree-based algorithms build one tree or several

* Corresponding author. Address: RM. 709, No. 6, South Road, Kexueyuan, Zhongguancun, P.O. Box 2704, Beijing 100190, PR China. Tel.: +86 10 82628446; fax: +86 10 62533449.
E-mail addresses: zyli@ict.ac.cn (Z. Li), antonial@ict.ac.cn (Z. Zhu), xie@ict.ac.cn (G. Xie), zcli@ict.ac.cn (Z. Li).

0140-3664/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2008.10.007

disjoint trees for the specified single-source. The multicast trees that are optimal for one source may be bad for others [7], while building and maintaining one tree for each potential node are too costly. Data-driven algorithms suffer a basic control-overhead-versus latency tradeoff [24]. Thus, they are not suitable for the delay sensitive applications that require fast message delivery, or they bring considerable control-overhead.

In the past, there have been several algorithms for applications with multiple data sources [11,29,9,7]. However, few of them meet above basic requirements. The probabilistic gossip based hybrid push/pull schemes [11] bring lots of duplicate messages and only guarantee probabilistic convergency. Structured P2P network based schemes [9,29] implement multicast service using implicit trees. However, the overlays are too strict to be optimized and the maintenance overhead is a major design concern [6]. ACOM [7], on the other hand, is based on unstructured overlays. Long delay and high redundancy (i.e. the number of duplicated messages) are its two major concerns. A common and major limitation of these algorithms is proximity ignorance. Therefore, they would consume unnecessary network resources (e.g. backbone bandwidth) and have a long convergence time.

In this paper, we present *FPCast*, a fast and proximity-aware multi-source overlay multicast algorithm. FPCast is motivated by two facts. First, although the structured overlay networks need considerable maintenance overhead and have relative rigid structures, their predefined structures (e.g. ring and hypercube) give us useful information to extract multicast trees. Second, measurement results in [25,18] have shown that there exist some relative powerful and stable nodes in large-scale wide area distributed applications.

Therefore, we organize nodes in a two-layer structure: the upper layer is based on DHT protocol and composed of powerful and stable nodes (denoted as *Dnodes*), while the nodes at the lower layer (denoted as *Onodes*) attach to physically close Dnodes. Obviously, a Dnode and the Onodes attaching to it constitute a cluster and the Dnode acts as the rendezvous point of that cluster. Data items are replicated and forwarded along implicit, source specific, and heterogeneity-aware multicast trees. We further reduce the delivery delay by introducing probabilistic forwarding scheme in the DHT structure. The analysis results show that the multicast message from any source can been delivered to all the other nodes within $O(\log n)$ hops, where n is the number of member nodes. We evaluated the performance of the FPCast algorithm via comprehensive simulations. The results demonstrate the superiority of our algorithm in terms of message delivery time and network resource consumption, in comparison with the previous randomized algorithms. The algorithm is also resilient to node failures.

The rest of the paper is organized as follows: Section 2 provides a survey of related work. Section 3 describes the construction of hierarchical structure in detail. Section 4 gives the details of the basic multicast algorithm, followed by the enhanced algorithm in Section 5. We evaluate our scheme through simulation experiments and show the results in Section 6. Finally, we conclude our work in Section 7.

2. Related work

Tree-based multicast algorithms [5,24] designed for single data source are not applicable for the applications with multiple sources. Building and maintaining one tree for each potential node are too costly, while using single tree or fixed small number of trees is not a good choice for that an optimal tree for one source may not be an optimal one for others and the traffic would be concentrated on the tree links [7]. Data-driven multicast algorithms [30,15] suffer a control-overhead-versus latency tradeoff [24]. Thus, they are not suitable for the delay sensitive applications that require fast message delivery, or they bring considerable control overhead.

Gossip based schemes [11] are scalable and probabilistic reliable. Their major limitation is considerable duplicate messages. NICE [2] forms nodes in a hierarchical structure to achieve high scalability. It is heterogeneity ignorant and fragile due to lack of redundancy in overlay networks. Narada [8] extracts source specific distribution trees from overlay network. It requires each node to maintain the information of all other nodes. Thus, it is only suitable for small scale systems.

The predefined structures in structured overlays provide us with useful information to build multicast trees. [9,17] are based on Chord [22] and CAN [16], respectively and can support the applications that have multiple data sources. However, both of them assume each nodes have same capacities and, therefore, is not suitable for heterogeneous systems. CAM [29] takes node heterogeneity into account and implements multicast service on capacity-aware DHT structures. A disadvantage of structured overlay based schemes is the considerable maintenance overhead [6], especially in capacity-aware structures in which each node has $O(c\log_c n)$ neighbors on average (c is average node capacity and n is the number of overlay nodes). Although our scheme also leverages DHT protocols to organize the upper layer nodes, the maintenance overhead is relative small because upper layer only consists of small portion of nodes which are more stable.

ACOM [7] and REM [12], on the other hand, implement overlay multicast on top of unstructured overlay networks. ACOM broadcasts the message on a random graph, and then transmits it down to the ring from the aware nodes. It transfers messages from any source to all the other nodes within $O(2\log_c n)$ hops, where c is the average node capacity and n is the number of overlay nodes. But this is at the cost of non-negligible redundancy: the average number of duplicate messages is $O(n/\log_c n)$. REM combines the advantages of flooding based scheme and tree-based scheme to transfer messages on top of a reliable and adaptive ring structure overlay. It needs a relative long time (e.g. 100 s) to adaptively optimize the overlay network and requires periodical tree reconstructions.

Cluster based scheme is proposed in [23]. Both the upper layer and second layer are organized in structured P2P fashions and node heterogeneity is not considered. Shen and Xu in [19] propose a landmark clustering based hierarchical structure to account for proximity information. They directly use node landmark numbers as the logical node IDs in the auxiliary DHT-based structure. Thus, the identifier randomness in the structure does not exist any more. Although this geographic layout provides network locality, it sacrifices the diversity of neighboring nodes in the ID space, which has adverse impact on failure resilience [4]. We also use hierarchical structure in [13]. But the upper bound of the number of nodes within a cluster is fixed, which limits the scalability. Besides, the ordinary nodes, which constitute the major part of an overlay, contribute nothing.

The authors in [10] point out that a proximity optimization can be imposed on Chord ring without affecting the query complexity. Yao and Loguinov in [27] derive closed-form models for the probability that Chord remains connected under node failures. These works are largely complementary to the work presented in this paper.

Our FPCast algorithm delivers messages along source specific multicast trees which are implicitly extracted from a proximity-aware hierarchical structure. The idea was first presented at the IFIP Networking 2008 [14]. This paper is significantly extended from the conference version. The idea of using probabilistic forwarding to reduce the delivery delay is newly added. The thorough theoretical results are added with complete proofs. Finally, the