

A Viable Solution for Large Scale Multicast Support

Jun-Hong Cui¹, Li Lao², Orang Dialameh³

jcui@cse.uconn.edu, llao@cs.ucla.edu, orang@ivisit.com

¹ Computer Science & Engineering Department, University of Connecticut, Storrs, CT 06029

² Computer Science Department, University of California, Los Angeles, CA 90095

³ iVisit LLC, P.O. Box 7639, Santa Monica, CA 90406

Abstract—In this work, we propose a Two-tier Overlay Multicast Architecture (TOMA) to provide scalable and efficient multicast support for a variety of group communication applications. In TOMA, Multicast Service Overlay Network (MSON) is advocated as the backbone service domain, while end users in the access domains form a number of small clusters, in which an application-layer multicast protocol is used for the communication between the clustered end users. Our two-tier architecture is able to provide efficient resource utilization with less control overhead, especially for large-scale applications. It also alleviates the forwarding state scalability problem and simplifies multicast tree construction and maintenance when there are large numbers of groups ongoing in the networks. We conduct simulation studies and the preliminary results demonstrate the promising performance of TOMA. Currently, we are in the process of implementing TOMA in a commercial on-line audio/vedio communication system, iVisit, to check the feasibility of TOMA and to further investigate its performance in real environments.

I. INTRODUCTION

The Internet has overseen more and more emerging group communication applications, such as video conferencing, video on-demand, network games, and distributed interactive simulation (DIS). Over the years, tremendous efforts have been made to provide multicast support, ranging from IP multicast to recently proposed application-layer multicast. IP multicast utilizes a tree delivery structure which makes it fast, resource efficient and scale well to support very large multicast groups. However, even after approximately two decades since the inception of IP multicast, it is still far from being widely deployed in the Internet. This is due to many technical reasons as well as marketing reasons [8, 2]. The most critical ones include: the lack of a scalable inter-domain routing protocol, the state scalability issue with a large number of groups, the lack of support in access control, the requirement of global deployment of multicast-capable IP routers and the lack of appropriate pricing models. These issues make Internet Service Providers (ISPs) reluctant to deploy and provide multicast service.

In recent years, researchers resort to application-layer multicast approach: multicast-related features are implemented at end hosts [10, 7, 14, 4, 12, 15]. Data packets are transmitted between end hosts via unicast, and are replicated at end hosts. Examples are Yoid [10], End System Multicast [7], ALMI [14], and NICE [4], to name a few. These systems do not require infrastructure support from intermediate nodes (such as routers), and thus can be easily deployed. However, application-layer multicast is generally not scalable to support large multicast groups due to its relatively low bandwidth efficiency and heavy control overhead caused by tree maintenance at end hosts. In addition, from the point of view of an ISP, this approach is hard to have an effective service model to make profit: group membership and multicast trees are solely

managed at end hosts, thus it is difficult for the ISP to have efficient member access control and to obtain the knowledge of the group bandwidth usage, and makes an appropriate pricing model impractical, if not impossible.

Then the question is: what is the viable or practical solution for large-scale multicast support? In a multicast service, multiple parties are involved: network service providers (or higher-tier ISPs), Internet Service Providers (or lower-tier ISPs, which we refer to as ISPs for short in this work), and end users. However, which party really cares about multicast? End users do not care as long as they can get their service at a reasonable price. This is why many network games are implemented using unicast. Network service providers do not care as long as they can sell their connectivity service with a good price. This actually contributes as one reason for the delay of IP multicast deployment. Obviously, ISPs in the middle are the ones who really care about multicast: their goal is to use limited bandwidth purchased from network service providers to support as many users as possible, i.e., to make a biggest profit. Therefore, to develop a practical, comprehensive, and profitable multicast service model for ISPs is the critical path to multicast wide deployment.

Strongly motivated, in this work, we propose a **Two-tier Overlay Multicast Architecture** (called **TOMA**) to provide scalable, efficient, and practical multicast support for a variety of group communication applications. In this architecture, we advocate the notion of **Multicast Service Overlay Network** (referred to as **MSON**) as the backbone service domain. MSON consists of service nodes or proxies which are strategically deployed by an MSON provider (ISP). The design of MSON relies on well-defined business relationships between the MSON provider, network service providers (i.e., the underlying network domains), and group coordinators (or initiators): the MSON provider dimensions its overlay network according to end user requests (provided by long-term measurement), purchases bandwidth from the network service providers based on their service level agreements (SLAs), and sells its multicast services to group coordinators via service contracts. Outside MSON, end hosts (group members) subscribe to MSON by transparently connecting to some special proxies (called *member proxies*) advertised by the MSON provider. Instead of communicating with its member proxy using unicast, an end host could form a cluster with other end hosts close by. In the cluster, application-layer multicast is used for efficient data delivery between the limited number of end users. The end users participating in the groups only need to pay for their regular network connection service outside MSON. A high level

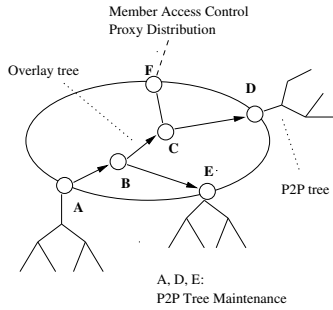


Fig. 1. A big picture of TOMA, where F is the group registry server/DNS server, and A, B, C, D and E are overlay proxies.

picture of TOMA is illustrated in Fig. 1. F is the group registry server, containing information about groups registered in the MSON, thus helps access control. A DNS server can also reside in F, and it is used by the MSON provider to advertise member proxies. In the network, there is one multicast group with (source) member proxy A and (destination) member proxies D and E. An overlay multicast tree is built in MSON for this group. Three P2P trees (one for each member proxy) are established outside MSON.

The proposed TOMA architecture not only provides scalable and efficient multicast support as well as a practical pricing model for ISPs, it also brings many other advantages. First, an MSON provider can support a variety of group communication applications simultaneously, unlike some other existing multicast overlays (such as [7], [10], [6], and [11]), with each overlay only supporting one group. This provides an additional incentive for ISPs to adopt TOMA. Second, it is relatively easy for ISPs to manage resources since MSON is based on well-defined business relationships via SLAs with network service providers and service contracts with group coordinators. They can put major efforts on planning and managing their overlay networks. Third, the notion of MSON significantly simplifies the management of underlying networks: network service providers only need to provide services to limited numbers of MSON providers instead of millions or billions of individual end users. This level of traffic aggregation, in the long run, will make IntServ practical. A good analogy to this scenario is the relationship between manufacturers, dealers, and consumers in our daily life. Lastly, MSON can be further extended to support virtual group services, such as web content distribution. Web clients who share similar interests can form a virtual group managed by MSON. Data transmission inside MSON can be reorganized in order to provide better services.

To make TOMA a reality, we face many challenges:

- **Efficient management of MSON:** It is anticipated that an MSON will accommodate a large number of multicast groups. How does an MSON provider efficiently establish and manage numerous multicast trees?
- **Cluster formation outside MSON:** End users of multicast groups might disperse around the country (or even the world). They first need to subscribe to MSON and connect to some member proxies. How should appropriate member proxies be selected? And how are efficient clusters formed among end users?

- **MSON dimensioning:** Given that the MSON provider has an overlay architecture, how should it dimension the overlay network? In other words, where should the overlay proxies be placed, which links are used to transmit data, and how much bandwidth on each link should be reserved?

- **Pricing:** The lack of mechanisms to measure resource usage and bill multicast users is one of the main barriers that delay the deployment of IP multicast. Therefore, how to charge the multicast group users of MSON is an important factor to decide whether MSON is a practical solution.

In this work, we tackle all these issues. To address the efficient management of MSON, we propose an efficient and scalable protocol, called OLAMP (OverLay Aggregated Multicast Protocol), in which we adopt *aggregated multicast* approach [9], with multiple groups sharing one delivery tree. Outside MSON, we develop efficient member proxy selection mechanisms, and choose a core-based application-layer multicast routing approach for data transmission inside clusters. Besides, we suggest several effective algorithms to dimension overlay networks: locating overlay proxies, identifying overlay links, and dimensioning bandwidth. We also propose a cost-based pricing model for the overlay ISPs who adopt the TOMA architecture to charge multicast users. We believe this pricing model provides incentives for both service providers and clients to adopt our multicast service scheme. We have conducted simulation studies and our preliminary results have shown the promising performance of TOMA as well as the effectiveness of our dimensioning algorithms.

II. PRELIMINARY RESULTS

In this section, we briefly present some results with regards to the performance of multicast trees built by different multicast schemes: TOMA, NICE [4] (a scalable application-layer multicast protocol), an IP multicast protocol (Core-Based Tree [3]), and unicast protocol (we include unicast as a reference point for some metrics).

We use following metrics to compare multicast tree performance. *Multicast tree cost* is measured by the number of links in a multicast distribution tree. It quantifies the efficiency of multicast routing schemes. To assess the quality of data paths, we measure link stress and path length when data are transmitted from a randomly selected source to all members. *Link Stress* is defined as the number of identical data packets delivered over each link. *Path Length* is the number of links on the path from the source to a member.

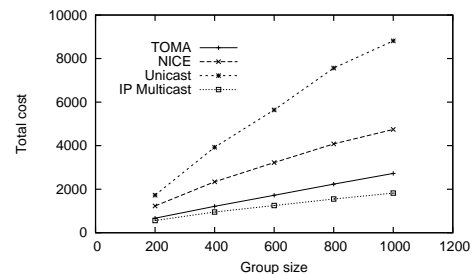


Fig. 2. Tree cost vs. group size.

We focus on large group sizes of 200 to 1000 members to test the scalability of different protocols. In simulation experi-

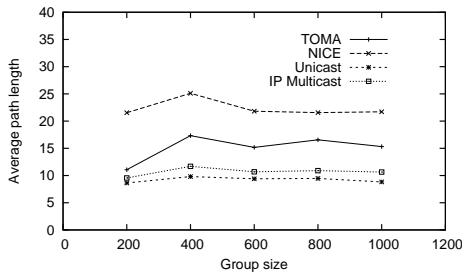


Fig. 3. Average path length vs. group size.

ments, end hosts join the multicast group during an interval of 400 seconds, and the session ends at 1000 seconds. We collect the metrics after the multicast tree has stabilized. We use various network topologies in our simulations, and different topologies yield similar results [13]. Due to space limitation, we only present the results for a set of Transit-Stub topologies [5].

Multicast tree cost In Fig. 2, we plot the average tree cost of TOMA, NICE, and CBT as group size increases. As a reference, we also include the total link cost for unicast. Compared with the cost of unicast paths, NICE trees reduce the cost by 30 – 46%, TOMA trees reduce the cost by approximately 61 – 70%, and CBT trees save the cost by 68 – 80%. Clearly, the performance of TOMA is comparable to IP multicast. In addition, TOMA outperforms NICE in all cases, and their difference magnifies as group size is increased. This efficiency gain of TOMA vs. NICE is due to two reasons. First, TOMA takes advantage of the carefully dimensioned overlay networks which resemble the underlying network topology, whereas NICE relies on path probing techniques and constructs overlay topologies with degraded quality. Second, by using proxies as intermediate nodes, TOMA allows the packets to be replicated at proxies and hence decreases the redundant packets sent over the physical links. We also found out that average link stress shows a similar trend [13].

Average path length The results for average path length are shown in Fig. 3. As expected, unicast and IP multicast have the shortest end-to-end paths. Additionally, the path lengths of TOMA trees are much shorter than those of NICE trees on average. For instance, at group size of 1000, the average path lengths of TOMA and NICE trees are 15.32 and 21.72, respectively. Again, the performance improvement of TOMA over NICE is gained through efficient overlays: in TOMA, overlay links are constructed based on the shortest paths in network layer; as a result, data packets can avoid going through unnecessarily long paths.

We also obtained results for the metrics of control overhead, robustness, and state scalability. Our observations through simulation experiments can be summarized as follows: TOMA creates multicast distribution trees with tree cost and average link stress almost comparable to those of IP multicast; the data paths of TOMA trees have lower latency than that of NICE trees; the control overhead of TOMA is significantly less than NICE for large groups; TOMA is more robust than NICE when there are ungraceful user leaves; TOMA is scalable to large numbers of groups in terms of control overhead and multicast state.

III. ON-GOING WORK

To further check the feasibility of TOMA and evaluate its performance in real environments, we are collaborating with iVisit¹, to implement and test TOMA in the existing iVisit system.

The current iVisit audio/video communication system explores a hybrid peer-to-peer/central-server approach. Multi-party communications can adaptively switch between central server mode and peer-to-peer mode based on the criteria such as network congestions, network failures, and participant dynamics, etc. In this way, Ivy can take advantages of both the robustness of central servers and the cost-efficiency of peer-to-peer services, achieving better performance compared with pure peer-to-peer or central server mode. However, this approach is not resource-efficient and scalable enough to support large numbers of users. Thus, the company is seeking for new solutions. MSON perfectly fits in the iVisit system, replacing the central-server. Proxies will be deployed by iVisit, and each of the member proxies will be responsible for end user cluster formation, as well as the multicast routing in MSON. In this way, TOMA provides an effective solution to scalable multimedia group communications for iVisit.

REFERENCES

- [1] iVisit LLC. <http://www.िवisit.com>.
- [2] K. Almeroth. The evolution of multicast: From the Mbone to inter-domain multicast to Internet2 deployment. *IEEE Network*, Jan./Feb. 2000.
- [3] A. Ballardie. Core Based Trees (CBT version 2) multicast routing: protocol specification. *IETF RFC 2189*, Sept. 1997.
- [4] S. Banerjee, C. Kommareddy, and B. Bhattacharjee. Scalable application layer multicast. In *Proceedings of ACM SIGCOMM*, Aug. 2002.
- [5] K. Calvert, E. Zegura, and S. Bhattacharjee. How to model and inter-network. In *Proceedings of IEEE INFOCOM*, Mar. 1996.
- [6] Y. Chawathe, S. McCanne, and E. A. Brewer. *An Architecture for Internet Content Distributions as an Infrastructure Service*, 2000. Unpublished, <http://www.cs.berkeley.edu/yatin/papers/>.
- [7] Y.-H. Chu, S. G. Rao, and H. Zhang. A case for end system multicast. In *Proceedings of ACM Sigmetrics*, June 2000.
- [8] C. Diot, B. Levine, J. Lyles, H. Kassem, and D. Balensiefen. Deployment issues for the IP multicast service and architecture. *IEEE Network*, Jan. 2000.
- [9] A. Fei, J.-H. Cui, M. Gerla, and M. Faloutsos. Aggregated Multicast: an approach to reduce multicast state. *Proceedings of Sixth Global Internet Symposium (GI2001)*, Nov. 2001.
- [10] P. Francis. *Yoid: Extending the Multicast Internet Architecture*. White paper, <http://www.aciri.org/yoid/>.
- [11] J. Jannotti, D. K. Gifford, K. L. Johnson, M. F. Kaashoek, and J. W. O. Jr. Overcast: Reliable multicasting with an overlay network. In *Proceedings of USENIX Symposium on Operating Systems Design and Implementation*, Oct. 2000.
- [12] M. Kwon and S. Fahmy. Topology aware overlay networks for group communication. In *Proceedings of NOSSDAV'02*, May 2002.
- [13] L. Lao, J.-H. Cui, and M. Gerla. A scalable overlay multicast architecture for large-scale applications. Technical report, UCLA CSD TR040008. <http://www.cs.ucla.edu/NRL/hpi/papers.html>, Feb. 2004.
- [14] D. Pendarakis, S. Shi, D. Verma, and M. Waldvogel. ALMI: An application level multicast infrastructure. In *Proceedings of the 3rd USNIX Symposium on Internet Technologies and Systems*, Mar. 2001.
- [15] A. Sobeih, W. Yurcik, and J. C. Hou. Vring: a case for building application-layer multicast rings (rather than trees). In *Proceedings of the IEEE Computer Society's 12th Annual International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems (MASCOTS'04)*, Oct. 2004.

¹iVisit LLC (“iVisit”) [1] is a provider of scalable online video/audio communications and collaboration technologies and services. It has over 200,000 current active users from 132 countries. Over the past few years, iVisit has hosted over 10 billion minutes of video conferencing and over 2.0 mil downloads of the software.