A DNS-aided Application Layer Multicast Protocol

Sze-Horng Lee, Chun-Chuan Yang, and Hsiu-Lun Hsu

Sze-Horng Lee: Department of Computer Science & Information Engineering
National Chi Nan University, Puli, Taiwan, R.O.C.
Email: leeh@ncnu.edu.tw

Chun-Chuan Yang: Department of Computer Science & Information Engineering
National Chi Nan University, Puli, Taiwan, R.O.C.
Email: ccyang@csie.ncnu.edu.tw
Corresponding author

Hsiu-Lun Hsu: Department of Computer Science & Information Engineering
National Chi Nan University, Puli, Taiwan, R.O.C.
Email: leiyate@stu.csie.ncnu.edu.tw

Keywords: ***Application Layer Multicast (ALM), Domain Name Service (DNS)***

Conference: **IAENG International Conference on Communication Systems and Applications (ICCSA'08)**

A DNS-aided Application Layer Multicast Protocol

Sze-Horng Lee, Chun-Chuan Yang, and Hsiu-Lun Hsu

***Abstract*— A variety of issues, both technical and commercial, has hampered the widespread deployment of IP multicast in the global Internet. Application Level Multicast (ALM) approaches using the overlay network have been recently proposed as a viable alternative to IP multicast. In this paper, we proposed a DNS-aided ALM which builts overlay network with the help of the existing Domain Name Service (DNS). The simulation study shows that the proposed protocol have better Relative Delay Penalty and Link Stress for performance, but with lower protocol overhead and Resource Usage for multicast transmission in comparison with two existing ALM solutions, NARADA and NICE.**

***Index Terms*—Application Layer Multicast (ALM), Domain Name Service (DNS).**

I. Introduction

IP multicast (network-layer multicast) [1] is a protocol to deliver information to multiple receivers by using the most efficient strategy. The messages are delivered over each link of the network only once and are only duplicated in the split of a link to different receivers. Bandwidth is significantly saved for a path split which is closer to receivers. Neither a sender nor a middle node has to keep the state of all receivers. It is an efficient mechanism for packet delivery in one-to-many data transfer applications.

Nevertheless, at present, a large part of the Internet is still incapable of supporting native IP multicast. Due to various technical and administrational issues, IP multicast has not been widely deployed after the protocol was developed for more than a decade. The current model [2]-[5] allows for an arbitrary source to send data to an arbitrary group at any time. This induces a serious problem of network vulnerable for flooding attacks by malicious sources. As a result, the network management and provisioning services become too complicated which makes a large number of network administrators to be unwilling to deploy IP multicasting.

Both IP unicast and IP multicast have their individual strengths and weaknesses. Many researchers contributed efforts in *Application-Layer Multicast (ALM)* protocols [6]-[8] which utilize the strength of both IP layer protocols and reduce their weaknesses. ALM protocols do not require the network infrastructure for multicast supporting but using the IP unicast with intelligence. More specifically, instead of relying on the supporting of IP multicast routers, the multicast forwarding functionality of ALM are implemented at end Hosts. Such ALM protocols have been increasingly used to implement efficient commercial content-distribution networks [9], [10]. ALM nodes (end Hosts) participating in a multicast group, or proxies that operate on the behalf of the nodes, are organized into ALM overlay network for multicast data delivery. The network is an overlay in the sense that each link corresponds to a unicast path between two end systems in the underlying Internet.

Drawbacks of application-layer multicast include (1) duplicate packets on physical links, and (2) a larger end-to-end delay than IP Multicast. The key idea for reducing these drawbacks is to build a multicast tree which is as close to the IP multicast tree as possible. Existing approaches for ALM focus on network characteristics (e.g., latency) to construct the multicast distribution tree. From the aspect of network layering concept, it is basically impossible for an application layer mechanism such as ALM to get the real topological information about the physical network [11], [12]. Therefore, most of the existing ALM solutions rely on probing the path status among individual end Hosts and reconfigure the ALM overlay network on a regular basis so as to improve the transmission performance. Major performance concerns of the solutions relying on the probing mechanism are (1) the large amount of probing messages [13]-[15] and (2) the longer time for the overlay network to stabilize [16]-[19].

In light of the fact that two closely related *DNS* (*Domain Name Service*) names (e.g. two DNS names share a same large part) normally imply the proximity in the network distance between the Hosts, the DNS names of participating end Hosts in a group can provide helpful information for constructing the ALM overlay network. In this paper, a DNS-aided Application Layer Multicast protocol is proposed. With the help of DNS, construction of the ALM overlay network in the proposed protocol is faster and more efficient. Simulation study has demonstrated that the DNS-aided ALM outperforms two typical ALM protocols, *NARADA* [6] and *NICE* [16], in terms of lower signaling cost, transmission delay, and the number of links used.

The remainder of the paper is organized as follows. Related works are discussed in section II. The DNS-aided application-level multicast protocol is explained in section III. Performance evaluation for the protocol is presented in section IV. Section V concludes the paper.

Manuscript received November, 2007.

Sze-Horng Lee is with the department of computer science and information engineering, National Chi Nan University, Puli, Taiwan, 545 ROC. (e-mail: leeh@ncnu.edu.tw).

Chun-Chuan Yang is with the department of computer science and information engineering, National Chi Nan University, Puli, Taiwan, 545 ROC. (e-mail: ccyang@csie.ncnu.edu.tw).

Hsiu-Lun Hsu was with the department of computer science and information engineering, National Chi Nan University, Puli, Taiwan, 545 ROC. (e-mail: leiyate@stu.csie.ncnu.edu.tw).

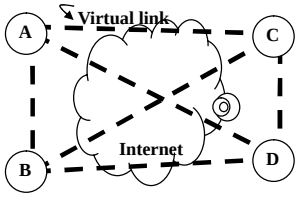
II. RELATED WORKS

All ALM protocols organize the group members into two topologies, namely the control topology and the data topology. Members that are peers on the control topology exchange periodic refresh messages to identify and recover from "ungraceful" departures from the group. The data topology is usually a subset of the control topology and identifies the data path for a multicast packet on the overlay. In fact the data topology is a multicast tree, while the control topology has greater connectivity between members. Therefore, in many protocols the control topology is called a *mesh* and the data topology is called a *tree*. Two of the typical ALM protocols are introduced as follows:

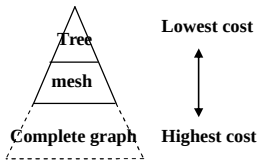
A. NARADA

The *NARADA* protocol [6] was one of the first application layer multicast protocols that demonstrated the feasibility of implementing multicast functionality at the application layer. Regardless of physical links that connecting joined members, a member may have virtual links to all others members that finally formed a complete graph for all members. Take four nodes as an example, their complete graph are shown in Fig. 1-(a) and every member has no idea of what is the network topology. We have to pay highest cost for maintaining virtual links while comparing with mesh and tree that are its links' sub-set which are shown in Fig. 1-(b). *NARADA* employs a two step process. First, a mesh is built among the participating end systems as shown in Fig. 1-(c).

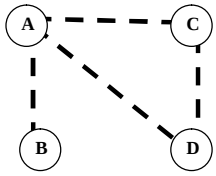
For transport of the actual data, *NARADA* runs a distance vector protocol with latency and bandwidth as the routing metrics on top of the mesh. The resulting tree is a source-specific shortest path tree based on the underlying mesh. The crucial factor in this approach is the quality of the mesh that must balance the number and the characteristics of the used unicast links. If there are too many links in the mesh, the resulting distribution topology will resemble a star of unicast connections from the sender to all receivers. Joining end systems obtain a list of current session members by a bootstrap mechanism and connect to one or more listed nodes. Then, members periodically add links that improve the routing performance and remove links that are rarely utilized by a distribution tree.



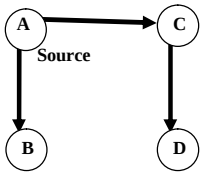
(a) All nodes' relation (complete graph).



(b) The relation and cost among complete graph, mesh and tree.

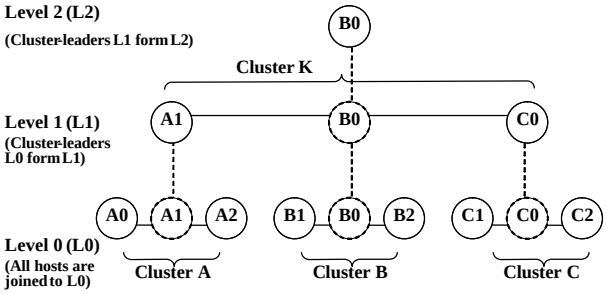


(C) Mesh (a case of NARADA)

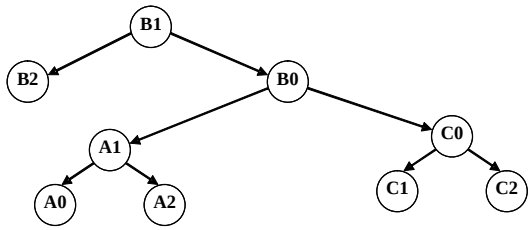


(d) Data delivery tree for source A.

Fig. 1 Control and data paths in a NARADA overlay network.



(a) Control path in NICE.



(b) An example of data path in NICE. (B1 as source)

Fig. 2 Control and data paths in NICE with fanout=3.

B. NICE

The *NICE* protocol [16] arranges the set of members into a hierarchical control topology. As new members join and existing members leave the group, the basic operation of the protocol is to create and maintain the hierarchy. The hierarchy is created by assigning members to different levels. Members in each layer are partitioned into a set of clusters. Each cluster is of size between *k* and *3k-1*, where *k* is a constant, and consists of a set of members that are close to each other. Further, each cluster has a cluster leader that has the minimum value of the maximum distance to all other members in the cluster. This choice of the cluster leader is important in guaranteeing that a new joining member is quickly able to find its appropriate position in the hierarchy using a very small number of queries to other members. A cluster leader periodically checks the size of its cluster, and appropriately splits or merges the cluster when it detects a size bound violation. As shown in Fig. 2, A1 is selected as leader among A0 and A2 for cluster A. For level 1, A1 further competes with B0 and C0, but B0 wins and becomes the cluster head of clusters K.

NARADA was first proposed as an efficient application layer multicast protocol for small group sizes. Unfortunately, it may introduce heavy overhead and need a longer time to a stabilized status for large group sizes. The *NICE* protocol takes advantage of the hierarchical structure for limiting protocol overhead within a constant number. Round trip time has been used as a tool for guessing a topology relation between members. We believe that a control topology can be constructed faster than the *NICE* protocol if domain names are provided.

III. DNS-AIDED ALM

Each physical link that required by IP multicast is only used by once for a multicast packet. As mentioned in the previous section, most of the ALM protocols tried to guess what the real network topology is like by using different methods. Unfortunately most of them have to pay large overhead or need longer time for stabilization. We believe that a large part of Internet nodes have at least one domain name that can relate to the physical network topology. The goal of our proposal is try to build an overlay network as closer as IP multicast network. As a result, we will enjoy the benefits of application-level multicasting but only paid the cost closer to IP multicasting. It is reasonable to assume that every Internet nodes can easily use domain name service. The service is already there and be part of Internet. So we do not need to pay any extra effort to deal with this service for other applications may also need to use it.

The proposed protocol tries to build an overlay network which is closer to a real network by the help of the DNS name, perhaps not fully matched but be closer when compared with other ALM solutions only using *ping*, *traceroute*, etc. First of all, a DNS-tree is built based on the DNS names of all participants of the group. Additional links are then added among the members that are located in the same DNS-tree node to form a tree-like mesh network. A source-based multicast delivery tree can be easily constructed from the mesh network when a member would like to send multicast packets to all group members. As in other ALM solution, a *Rendezvous Point* (*RP*) is assumed in the protocol for membership management. Moreover, the RP also helps for the construction as well as the information distribution of the DNS-tree and the mesh network. Related mechanisms in the proposed protocol are presented as follows.

A. DNS-tree for a group and domain head selection

In order to build the DNS-tree for a group, the DNS names of all members must be first collected in the RP, which means that each member joining the group must present its DNS name. The DNS-tree of the group is then constructed by extracting different levels of the DNS names of the group and associates each member with the appropriate tree node. For example, for the group with nine members in Table 1, the DNS-tree is displayed in Fig. 3. Note that in the DNS-tree, the non-leaf nodes (rectangle nodes in Fig. 3) represent different levels of DNS domains, and the leaf nodes (circle nodes in Fig. 3) represent the members of the group. Since *Host1*, *Host2*, and *Host3* share the same domain "*ncnu.edu.tw.*", they are put under the same non-leaf node in the DNS-tree.

Table 1: Member lists of a group for an example.

Host1.ncnu.edu.tw.	*Host2.ncnu.edu.tw.*
Host3.ncnu.edu.tw.	*Host4.csie.ncnu.edu.tw.*
Host5.csie.ncnu.edu.tw.	*Host6.im.ncnu.edu.tw.*
Host7.ntu.edu.tw.	*Host8.csh.org.tw.*
Host9.hinet.net.	

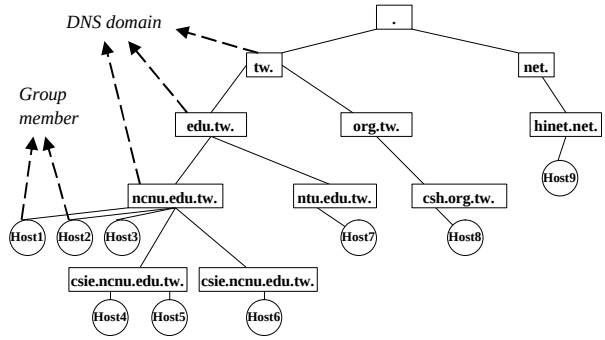


Fig. 3 The DNS-tree for the group in Table 1.

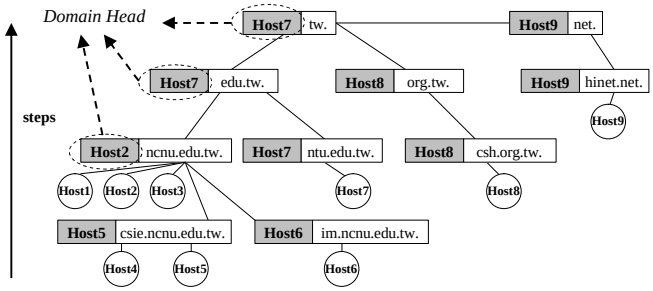


Fig. 4 Assigning domain heads for the DNS-tree in Fig. 3.

To properly connect all the members in the same group, we must select a proper member on behalf of each domain in the DNS-tree. The member on behalf of a domain is called a *domain head* (*DH*) in the paper. The process of selecting the DH for each domain is performed from the bottom level to the top level of the DNS-tree. A member will be selected as a DH if there is no other competitor. For example, *Host6* is the default DH for the domain "*im.ncnu.edu.tw.*" in Fig. 3.

Competition for the DH of a domain occurs when there is more than one candidate. In such case, all candidates measure and compare their *Round Trip Time* (*RTT*) to a randomly selected member that is located outside the sub-tree rooted by the domain name in interest. The member with the smallest *RTT* becomes the DH of the domain name, since the DH should be closer to the outside world than other candidates. For example, *Host4* and *Host5* in Fig. 3 measures the *RTT* to *Host7* in competition for the DH of "*csie.ncnu.edu.tw.*". *Host5* becomes the DH since its *RTT* is smaller as shown in Fig. 4. The process of competition for the DH of each domain continues upwardly until all DHs are selected. For example, *Host1*, *Host2*, *Host3*, *Host5* (the DH of "*csie.ncnu.edu.tw.*"), and *Host6* (the DH of "*csie.ncnu.edu.tw.*") in Fig. 3 compete for the DH of "*ncnu.edu.tw.*". Fig. 4 serves as an example of the result of the DH selection process for Fig. 3.

B. Building the control mesh and the data delivery tree

In the DNS-tree of a group, the communications among the members in the same domain (e.g. *Host1* and *Host3* in Fig. 4) must be relayed via the DH. But sharing the same domain implies the close relationship in the physical network, adding additional links among these members should not cause too much overhead and can reduce the transmission latency in multicast delivery. Therefore, we propose that the members in

the same domain can add more mesh links between each other until reaching the predefined threshold of the maximum number of links to neighbors (fanout). In most of the cases, members in the same domain can form a complete graph, since the number of the members in the same domain usually is not too large. Fig. 5 displays the mesh network for the example in Fig. 4.

As a member of a group sending out a multicast packet to the group, a corresponding source-based multicast delivery tree is constructed on-demand at each node in the control mesh network. Since the members of the group maintain the same mesh network, the multicast delivery tree (rooted from the same source) constructed by all group members is identical which guarantees consistent multicast transmission. Fig. 6 shows some examples of the multicast delivery tree from difference sources for the mesh network in Fig. 5.

C. Member Join or Leave

Since join or leave of a member introduces changes in the DNS-tree of the group, some DHs probably should be reassigned to better fit for the new membership. Therefore, after obtaining the updated membership of the group on join or leave of a group member, the RP (1) updates the DNS-tree, (2) re-invokes DH selection process, (3) rebuilds the mesh network, and (4) notifies all members with the changes.

Sometimes a member may leave a group without notifying the RP, or a member may experience network problems that cause the failure of connection to the mesh network. To deal with such cases, nodes in the mesh network should maintain and monitor links (e.g. exchange periodic hello messages) to neighbors and notify the RP when changes occur for proper adjustment in the mesh network.

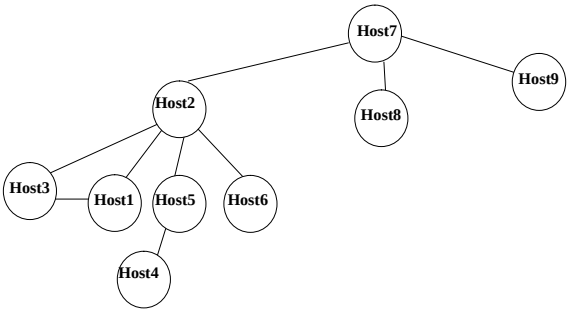
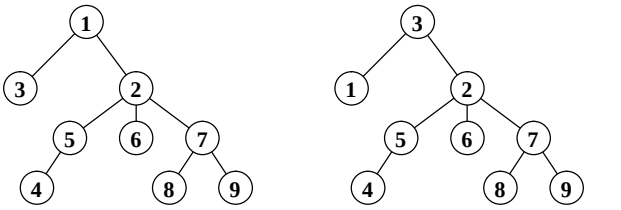
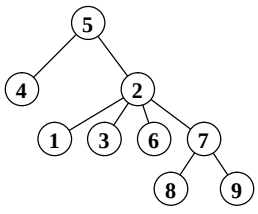


Fig. 5 Building the tree-like mesh for Fig. 4.

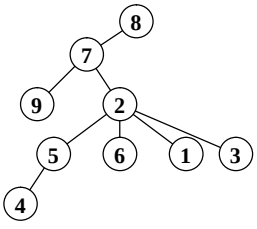


(a) Host1 acts as source.

(b) Host3 acts as source.



(c) Host5 acts as source.



(d) Host8 acts as source.

Fig. 6 Source-based multicast trees for Fig. 5

D. Discussion

Although we suppose that the DNS names present topological information in the network, there are cases that a DNS name should not located in an expected (normal) place in Internet. From the viewpoint of the proposed protocol, the mismatch of the DNS names with the real network topology introduces the misplacement of a member in the DNS-tree. There should be mechanisms to detect and fix such abnormity for performance improvement. Detection of the misplacement of a node in the DNS-tree relies on the comparison of the RTT values measured by the members share the same domain. If a member presents a large gap of RTT comparing with the other members (and the DH) in the same domain, the member is considered as a candidate of misplacement in the DNS-tree. The RP should invoke a finding process for the new position of the candidate in the DNS-tree. Moreover, the candidate of misplacement should not compete for the DH for any domain. Detail of the mechanisms for dealing with the abnormity in the DNS-tree is left as the future work of the research.

IV. PERFORMANCE EVALUATION

A. Simulation Environment and Performance Criteria

GT-ITM [20] was used to generate the hierarchical network topologies with 1024 nodes and 1633 edges (physical links). All nodes were distributed in the 4-layer hierarchy and each node was given a unique domain name. The maximum number of neighbors in the simulated physical network was limited to10. Different group sizes including 8, 16, 32, 64, 128, 256 and 512, are selected for performance evaluation, in which the group members were randomly selected from the nodes. For a given group size, the performance parameters were calculated over 40 different simulation rounds.

Six performance criteria are defined for performance evaluation: (1) *Average Relative Delay Penalty (ARDP),* (2) *Maximum Relative Delay Penalty (MRDP),* (3) *Average Link Stress,* (4) *Maximum Link Stress,* (5) *Protocol Overhead* and (6) *Resource Usage*. ARDP measures the average ratio of the delay from the source to receivers in ALM protocols over the delay of the shortest path in the physical network. A shortest path tree in the physical network has the RDP=1. MRDP is defined as the worst (largest) delay ratio.

A physical link for multicast transmission of a packet is used only once in IP multicasting, but the physical link may be used more than once for ALM. Link Stress for a physical link is defined as the number of packet transmission to accomplish one multicast transmission. The average value of Link Stress can demonstrate how close for an ALM protocol to IP multicast from the aspect of physical link usage. Moreover, a large value of Link Stress for a physical link may imply that the link is a bottleneck in the network.

Protocol Overhead is defined as the overhead to construct

and maintain the control mesh network. In ALM, a node needs to exchange routing information or refresh (hello) messages with its neighbors in the mesh network. Since a larger number of mesh links implies a larger Protocol Overhead, we defined Protocol Overhead of an ALM protocol as the number of mesh links in the control mesh network.

Resource Usage is the total number of hop-wise transmission for a single multicast transmission. For better comparison with the optimal solution of IP multicast, we define Resource Usage in the paper as the normalized term by IP multicast (i.e. Resource Usage of IP multicast is set to 1).

B. Simulation results

ARDP and MRDP of NARADA, NICE, and the proposed DNS-aided protocol are displayed in Fig. 7 and Fig. 8 respectively. The figures show that for large group sizes (128 and above), the delay performance of NICE drops seriously in contrast to the other schemes. The reason is the multi-level hierarchical structure in NICE is aiming to reduce the cost of mesh construction, but the tree-like mesh presents the longer path for multicast transmission. Also adopting a tree-like structure for the overly mesh network, but the proposed DNS-aided scheme does not present the poor delay performance as in NICE, which demonstrates the mesh network in the DNS-aided scheme is close to the underlying physical topology.

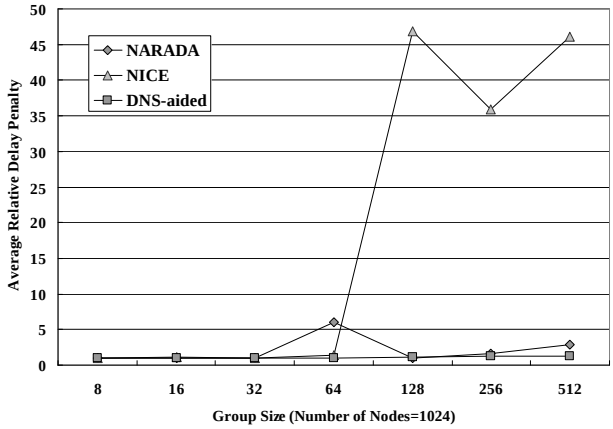


Fig. 7 ARDP for Narada, NICE and DNS-aided ALM under different group sizes.

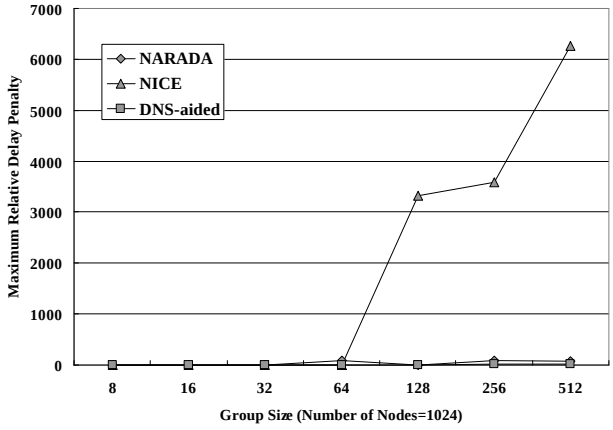


Fig. 8 MRDP for Narada, NICE and DNS-aided ALM under different group sizes.

Average and Maximum Link Stress for the three schemes are displayed in Fig. 9 and Fig. 10 respectively. Note that Link Stress of IP multicast is 1, therefore the figures demonstrate that the proposed scheme is closer to IP multicast in the aspect of link usage. Higher Link Stress (avg. and max.) of NARADA is due to the mechanism of adding links to the mesh network. The mechanism can help to reduce delay as shown in Fig. 7 and Fig. 8, but it introduces larger Link Stress, Protocol Overhead, and Resource Usage as displayed in Fig. 9 – Fig. 12.

NICE and the DNS-aided scheme take advantage of the hierarchical characteristic in mesh network construction, therefore the two schemes save much Protocol Overhead for large group sizes in contrast to NARADA as shown in Fig. 11. Fig. 12 shows that the hierarchical schemes, NICE and DNS-aided, requires less network resource for multicast transmission especially for large group sizes. Moreover, the DNS-aided scheme even outperforms NICE in terms of Resource Usage. The simulation result presents the gain of the DNS-aided scheme over NICE in Resource Usage is 122% for group size 8, 65% for group size 64, and 35% for group size 512.

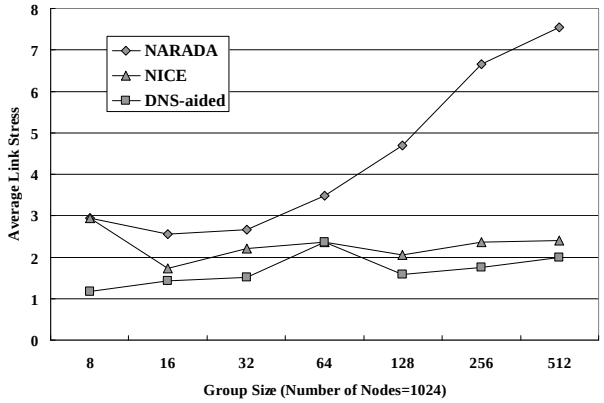


Fig. 9 Average Link Stress for Narada and DNS-aided ALM under different group sizes.

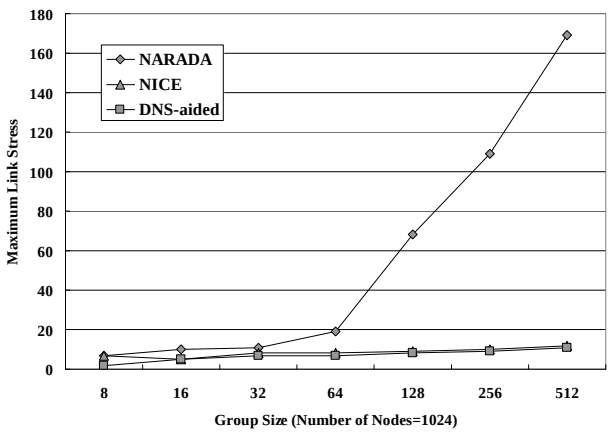


Fig. 10 Maximum Link Stress for Narada, NICE and DNS-aided ALM under different group sizes.

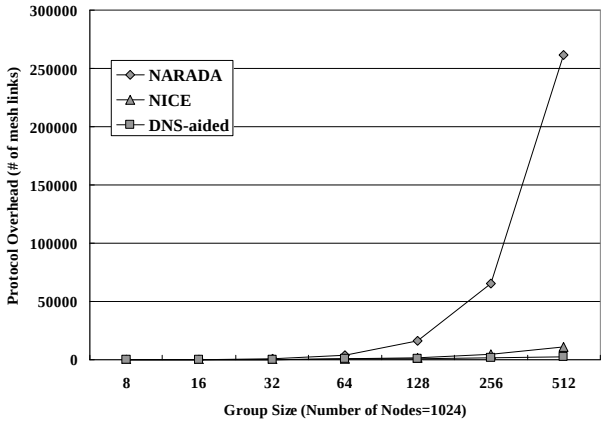


Fig. 11 Protocol Overhead for Narada, NICE and DNS-aided ALM under different group sizes.

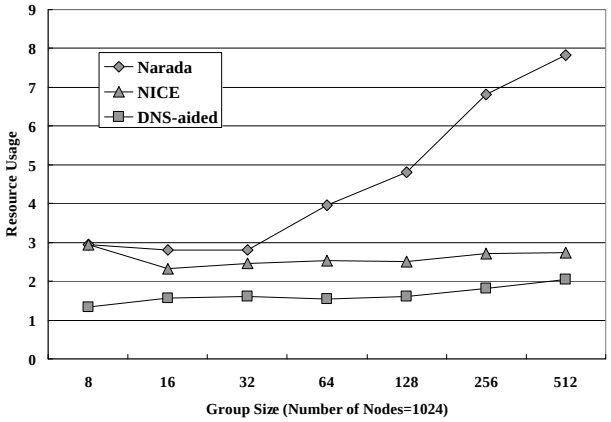


Fig. 12 Resource Usage for Narada, NICE and DNS-aided ALM under different group sizes.

V. CONCLUSION

Due to the security concern as well as limited deployment, IP multicast has not been widely used in Internet. On the other hand, Application-Layer Multicast (ALM), which does not require the support of network layer multicast but implements multicast forwarding functionality in the end hosts, is becoming a good alternative for multicast network applications. Most of the existing ALM protocols build the overlay network in a random manner and enhance the network by probing mechanisms such as Round Trip Time (RTT) measurement. By taking advantage of Domain Name Service (DNS), a DNS-aided ALM protocol is proposed in the paper. Associated mechanisms such as mesh network construction, member join/leave, and multicast transmission are presented. Simulation study has demonstrated that the proposed DNS-aided ALM protocol outperforms two prestigious ALM protocols, NARADA and NICE, in terms of transmission delay, link stress, protocol overhead, and resource usage.

REFERENCES

[1] S. E. Deering, "Multicast routing in internetworks and extended LANs." *Proc. ACM SIGCOMM*, vol.18, issue 4, Aug 1988, pp.55-64.

[2] J. Moy, "Multicast extensions to OSPF," RFC-1584, Mar 1994. Available: http://www.ietf.org/rfc/rfc1584.txt

[3] D.Waitzman et al.,"DistanceVector Multicast Routing Protocol ", RFC-1075, 1998. Available: http://www.ietf.org/rfc/rfc1075.txt

[4] C. Liu, and L.Wei, "An architecture for wide-area multicast routing." *Proc. ACM SIGCOMM*, vol.24, issue 4, Aug 1994, pp.126-135.

[5] D. Estrin et al., "Protocol independent multicast-sparse mode (PIM-SM): Protocol specification", RFC-2117, 1997. Available: http://www.ietf.org/rfc/rfc2117.txt

[6] Y. Chu et al., "A case for end system multicast." *Proc. ACM Sigmetrics*, Jun 2000, pp.1-12.

[7] J. Jannotti et al., "Overcast: Reliable multicasting with an overlay network." *Proc. 4th Symposium Operating System Design & Implementation*, vol. 4, 2000, pp.14-14.

[8] Y. Chawathe, "Scattercast: An architecture for internet broadcast distribution as an infrastructure service," *Ph.D. thesis, Univ. California, Berkeley*, 2000.

[9] S. Ratnasamy *et al.*, "Application-level multicast using content-addressable networks," in *Proc. 3rd International Workshop on Networked Group Communication*, 2001, vol. 2233, pp.14-29.

[10] S. Zhuang et al., "An architecture for scalable and fault-tolerant wide-area data dissemination." *Proc. 11th International Workshop Network and Operating Systems Support for Digital Audio and Video*, 2001, pp.11-20.

[11] P. Francis, "Yoid: Your own internet distribution" 2000. Available: http://www.icir.org/yoid/

[12] D. Pendarakis et al.,"ALMI: An application level multicast infrastructure." *Proc. 3rd Usenix Symp. Internet Technologies Systems, 2001*. Available: http://www.usenix.org/events/usits01/full_papers/shi/shi.pdf

[13] Z. Li and P. Mohapatra, "HostCast: a new overlay multicasting protocol." *Proc. IEEE International Conference on Communications, May 2003*, vol. 1, pp.702-706.

[14] M. Kwon and S. Fahmy, "Topology Aware Overlay Networks for Group Communication," *Proc. 12th International Workshop Network and Operating Systems Support for Digital Audio and Video*, 2002, pp. 127-136.

[15] M. Castro *et al.*, "Scribe: A large-scale and decentralized application-level multicast infrastructure." *IEEE Journal on Selected Areas in Communications*, vol. 20, Issue 8, Oct. 2002, pp. 1489– 1499.

[16] S. Banerjee et al.,"Scalable application layer multicast." *Proc. ACM SIGCOMM*, 2002. vol 31, pp. 205-217.

[17] V. Padmanabhan et al., "Resilient peer-to-peer streaming." *Proc. IEEE International Conference on Network Protocols*, 2003. pp.16.

[18] W. Wang et al., "Overlay optimizations for end-Host multicast.", *Proc. 4th International Workshop Networked Group Communication*, 2002.

[19] K.-W. R. Cheuk et al." Island multicast: the combination of IP multicast with application-level multicast." *Proc. IEEE International Conference on Communications*, 2004. Communications, 2004 IEEE International Conference on, 20-24 June 2004, vol. 3, pp. 1441-1445

[20] Y GT-ITM: Modeling topology of large internetworks. Available: http://www.cc.gatech.edu/projects/gtitm/