文章编号:1674-2869(2010)09-0090-04

一种应用层组播拓扑的设计

王 立[1],黄俊年[1],余 鑫[1,2],刘 娣[1]
(1.华中师范大学 信息技术系,湖北 武汉 430079;
2.华中科技大学 电子与信息工程系,湖北 武汉 430074)

摘 要:设计了一种网状优先的应用层组播拓扑,将分发节点组成局部 Cluster 后,再将 Cluster 组建为一个层次化树形拓扑结构,并在此层次化树形集群系统基础上建立组播树.为多个组播数据源在此共享拓扑上建立起不同的最短路径组播树,实现应用层组播.

关键词:应用层组播;网状优先;层次拓扑;组播树
中图分类号:TP393 文献标识码:A doi:10.3969/j.issn.1674-2869.2010.09.022

0 引 言

近年来,由于 IPTV 直播、视频监控等应用的需求,支持实时多媒体传输的组播通信技术的研究和方案日益增多.IP 层组播能够在通信时保证在每一条网络链路中只存在一份数据报文,能够极大地节省网络带宽,但由于组播需要底层路由器进行全面的升级支持如 DVMRP[1]和 IGMP[2]等路由协议,在部署方面存在较大的困难[3],同时由于组播路由器要为每一个组播组维护路由状态信息,扩展性不好;另外,组播的可靠性、QOS、拥塞控制等方面目前仍处于研究阶段,因此 IP 组播并没有得到广泛应用.

于是,出现了一些对 IP 层组播的替代方案[4-5].有研究者提出将对组播通信功能的支持从 IP 层的路由器转移到终端系统中,由应用层的终端系统来负责组成员管理、报文的复制和分发,并称之为应用层组播[6]:Application Level Multicast 或者 Overlay Multicast.应用层组播利用现有的网络传输协议,并不需要网络的底层路由和传输结构作调整,不存在部署困难的问题.同时应用层组播的路由路径可以随着网络状况的变化动态地调整,应用程序也可以参与路由策略的制定,能实现 IP 组播所没有的灵活性和扩展性.

1 应用层组播 ALM

IP 应用层组播通常有两种实现方式:一种是将路由和传输功能放在参加组播通信的各个主机中,组成 P2P 的 overlay 网络;另一种则是由分布在网络中的多个组播节点完成组播功能,每个节点可以为多个客户端同时服务.第一种方法可以做到完全分布,第二种方法则可以提高组的规模.

在应用层组播组的管理方面,根据建立应用层组播拓扑结构时采用的方案,可采用两种管理方式:网状优先(mesh first)和树状优先(tree first)[6].网状优先的系统会首先在节点上建立一个网状的拓扑结构,节点间按照某种路由协议来生成路由树.网状优先的模式下,系统的拓扑结构是确定的,但是路由树结构是不确定的.树状优先系统中,节点直接通过某种算法在树形关系中选择其各自的父节点,并检测和避免环路的产生.网状优先的系统能通过重新选择邻居、更改拓扑关系的方式,在很大的范围内更新路由树的结构.因此,在多源的系统中,网状优先系统相对更加稳固,能更加灵活的对不同源建立不同的组播树.

树状优先的方案有 ALMI[7]等,大规模的网状优先应用层组播方案的代表有 NICE[8]和 Zigzag[9],它们对单个数据源组播的问题,都使用了“分层”(Hierarchical)和“分群”(Cluster)的思路.大部分组成员位于分层结构的底层,并只和少量固定数目的节点存在联系,这样就大大降低了大部分组播成员的处理开销.

2 MCR 组播拓扑

本文提出的组播环拓扑 MCR(Multicast Cross Rings)将应用层组播节点以集群的方式管理起来,

收稿日期:2010-05-21
作者简介:王 立(1983-),男,湖北荆州人,助教,硕士研究生.研究方向:复杂系统,信息处理.
指导老师:余 鑫,男,讲师,博士,硕士研究生指导老师,研究方向:多媒体技术.*通讯联系人.

建成以“环”为基本集群单元的层次化管理拓扑结构，并且在此拓扑结构基础上建立组播路由树.

2.1 树状层次拓扑

MCR 采用的层次化树形结构的管理拓扑，分两个级别管理所有节点间的逻辑关系. 第一级别是树形拓扑，第二级别是环形拓扑(集群子系统). 如图 1 所示，每个树形节点均为一个集群子系统，每个集群子系统保持环状拓扑的多个节点.

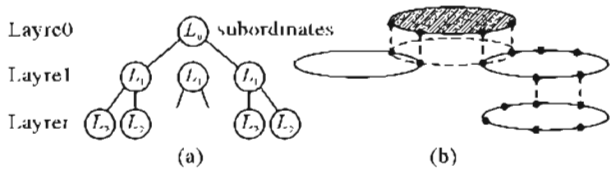

图 1 树状的层次拓扑

Fig. 1 Hierarchy topology tree

MCR 将节点组织为多个 cluster，再以多叉树的方式来组织多个 cluster 形成层次化结构，其中每个 cluster 均为采用环状结构组织的多个节点集合，如图 1(b)所示.

若用 k 表示每个 cluster 的最大下级 cluster 数，则 MCR 是一个由多个 cluster 组成的 k 叉树. MCR 的层次拓扑随着节点的加入而逐渐扩展，节点首先尝试加入顶层 cluster，如果失败则逐级向下寻找一个有空闲的 cluster 加入.

组建层次拓扑的组织基本准则包括：

(1) MCR 是由多个 cluster 组成的 k 叉树($k>=3$)，每个 cluster 最多有 k 个下级分支 cluster；

(2) 最先加入的 $2k$ 个节点组成的 cluster 为 MCR 的顶点 L_0，后续加入的节点寻找 L_0 的一个分支加入；

(3) 每个 cluster 环最多可包含 $2k$ 个节点，其中两个节点为 RP 节点(互为备份的集中点)，其他节点为普通节点(L_0 环没有 RP 节点)；

(4) 任何一个节点最多可以同时属于 2 个 cluster 环，即两个关联(相交)的 cluster 最多有 2 个公共节点；

如图 2 所示，新节点在 L_0 下属子环中寻找一个未满的环如图 2(c)所示，或新建一个环如图 2(b)所法加入，形成共 2 层的拓扑.

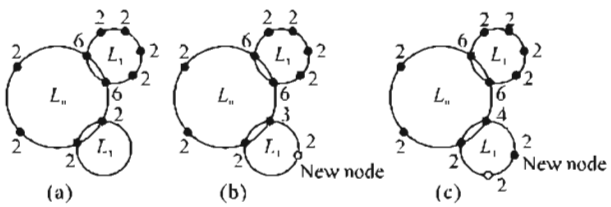

图 2 拓扑组织过程示例($k=3$)

Fig. 2 Contraction example($k=3$)

L_i 表示某环在分层拓扑中的层编号，L_0 是最顶层的环，也是一级树形拓扑的根. 如果要保证一级拓扑树的平衡性，L_0 的位置可能被拓扑管理算法更新.

定义：RS 表示一个环，A_i 为该环上所有节点，$W(A_i)$ 为节点的重量，普通节点的重量为 2，RP 节点 X 的重量计算方式为：

$$W(X)=\sum_{A_i\in RS,A_i\notin RP}W(A_i)/(2+2)$$

如图 3(b)所示，通过计算可知 L_{i+2} 上 RP 节点的 W 为 6，L_{i+1} 上 RP 节点的 W 为 14，L_i 上 RP 节点的 W 为 30. 事实上 W 值也代表了该环及下属环链所包含的总的节点数.

定理：某环的 RP 节点的 W 即为该 RS 环及其下属子环所含总的节点的数量.

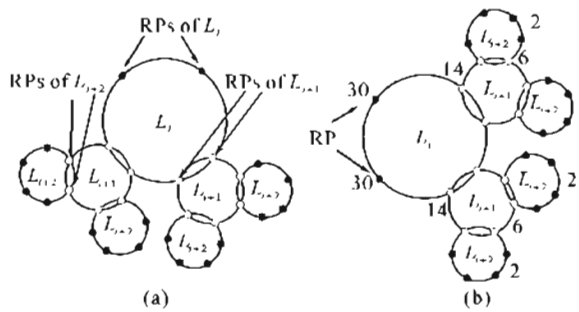

图 3 节点与环的关系

Fig. 3 Relationship between nodes and clusters

2.2 层数与节点数的关系

若用 N 表示拓扑中总的节点数，H 表示层次拓扑的层数. 如图 3(a)所示，满环情况下，将每个环的 RP 节点仅在主环上统计一次，某顶层环节点数为 $2k$，其下属各个环的节点数均为 $2k-2$，第 i 层的总环数为 $(k-1)^i$. 则满环情况下，H 层拓扑容纳的最大节点数为：

$$N_{\max}=\sum_{i=0}^{H-1}\left(\sum_{j=0}^{(k-1)^i}R_{ij}\right)=2k+(k-1)(2k-2)+\cdots+(k-1)^{H-1}(2k-2)=2\sum_{i=1}^{H}(k-1)^i+2=2(k-1)\frac{(k-1)^H-1}{k-2}+2$$

令 $j=k-1$，$k>=3$，即 $j>=2$，则：

(1) 层数确定时的最大节点数为 $N_{\max}=2\dfrac{j^{H+1}-1}{j-1}$；

(2) 节点数确定时的最小层数为 $H_{\min}=\log_j\left(\dfrac{j-1}{2}N+1\right)-1$.

完整的 MCR 拓扑在 L_0 层实际可以由 k 个图

3(a)所示的 H 层子环拓扑组成，因此，MCR 的总层数为 $H^O = H + 1$，全网最大节点数为 $N^O_{\max} = kN_{\max}$，即：

$$N^O_{\max} = 2\frac{j^{H+1}-1}{j-1}k = 2\frac{j+1}{j-1}(j^{H^O}-1)$$

于是，节点数确定时的 MCR 最小层数为：

$$H^O_{\min} = \log_j\left(\frac{j-1}{2(j+1)}N+1\right)$$

3 MCR 拓扑维护

根据前面描述的拓扑组建原则，新节点加入时，首先找到 L_0 环，请求加入该环，当 L_0 环上的节点数已达到 $2k$ 时，如果 L_0 没有子环则新建一个，否则通过子环选择算法在其 k 个下属子环中选择一个加入.

3.1 子环选择

节点加入时，应保证拓扑树的总层数最小、枝的数量最少，即树最矮，从而使组播路由的跳数最小. 此算法的节点加入拓扑的顺序图 4 所示，层次结构中将依次填满 L_0、所有 L_1、所有 L_2……. 拓扑包含的所有环仅有一个环的节点数小于 $2k$，每次节点加入的算法就是找到层次拓扑中节点数小于 $2k$ 的这个环，并加入该环.

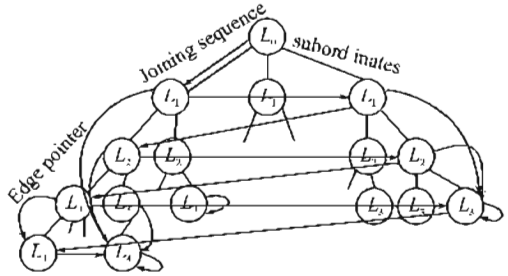

图 4 拓扑扩展方式

Fig. 4 Topology growing path

子环选举算法如下：

(1)遍历 L_0 环的节点重量集合 W_0，计算系统总节点数 $N = \sum_{w_i \in W_0} w_i/2$；

(2)计算新增一个节点后的总层数最小值 $H^O_{\max} = \log_j\left(\frac{j-1}{2(j+1)}(N+1)+1\right)$；

(3)计算子环所内包含的最大节点数 $N_{\max} = 2\frac{j^{H+1}-1}{j-1}$，其中 H 为子环层数，所以 $H = H^O - 1$；

(4)找到 L_0 环上 w 最大且小于 $N_{\max}$ 的节点 X_m，新节点向该节点(即以其为 RP 的环)发起加入子环的流程.(注：若所有 w 均为 2 将建立新环.)

3.2 组播路由

应用层组播分发的关键问题是节点度与路由深度的平衡. 节点度的表示节点分发数据流时的输出端连接数，路由深度表示数据流从源端到达接收端经过的转发次数. 组播路由要保证数据报文通过组播数据源到达加入该组播组的各个节点，同时要保证达到不同的节点时的转发的跳数受控，从而提高转发效率.

假设环中的各个节点之间通过直接或间接的网络层路由互相可达，因此环内的节点通信距离为 1. 若规定环间节点的通信通过 RP 节点分发，则环间节点的通信距离与通信通过的环的数量相同. 根据 MCR 的组网算法，在节点加入流程中，采用最短树算法保持网络的总环数最少时，MCR 的拓扑树高度被算法控制为 $H^O_{\min} = \log_j\left(\frac{j-1}{2(j+1)}N+1\right)$，因此在此拓扑基础上建立的路由转发数可以保证最差情况下的路由跳数能够受到限制.

如图 5(b)所示，一个组播源 source 节点发送组播数据，有两个 destination 节点希望接收该组播数据，即加入该组播组. 其中一个 destination 节点与 source 处于同一子环内，另一个 destination 节点位于另一个子环. 此时组播分发路径为：source 节点将数据流发送给其所在 RM 环的 RP1 节点，该 RP1 节点将流转发给同一 RS 环内的 destination1 节点，同时将流分发给上级 RM 环的 RP2 节点. RP2 将数据流转发给 RP3，RP3 将数据转发给 RP4，RP4 将数据转发给最终目的 destination2. 分发采用的路由树如图中箭头所示，对于图 5 中所示的 3 层拓扑，路由树的最大深度为 5.

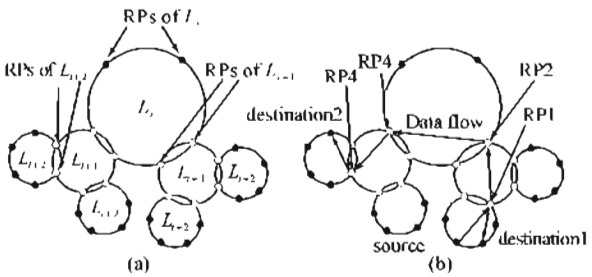

图 5 组播分发路径

Fig. 5 Multicast tree

4 MCR 的性能分析

MCR 的复杂度分析如下：

(1)每个节点可能存在于两个环中，每个环有 $2k-1$ 个邻接点，所以节点的度最大值为

$$2(2k-1)-1 = 4k-3$$

平均度为

$$2k+1$$

而在满环配置情况下，中间节点的度均为

$$4k-3$$

最外环节点的度均为

$$2k-1$$

因此总度和平均度分别为：

总度：

$$\begin{aligned}&\lceil 2k+(k-1)(2k-2)+\cdots+\\&(k-1)^{H-2}(2k-2)\rceil(4k-3)+\\&(k-1)^{H-1}(2k-2)(2k-1)=\\&2\frac{j^{H+1}-1}{j-1}(4k-3)-2j^{H}(2k-2)\end{aligned}$$

平均度：

$$\frac{2\frac{j^{H+1}-1}{j-1}(4k-3)-2j^{H}(2k-2)}{2\frac{j^{H+1}-1}{j-1}}=$$

$$4k-3-\frac{2j^{H+1}(j-1)}{j^{H+1}-1}\Rightarrow$$

$$4k-3-2(j-1)=2k+1$$

(2)路由的跳数即转发次数受控，因此组播分发的端到端延时受控为 $O(\log N)$ 级别.

因为总层数为

$$\log_j\left(\frac{j-1}{2(j+1)}N+1\right)$$

因此最坏情况的组播转发跳数为

$$2\log_j\left(\frac{j-1}{2(j+1)}N+1\right)$$

(3)加入和离开拓扑的控制开销为拓扑的层数，即

$$O(\log N)$$

5 结 语

MCR 采用网状优先的拓扑组织形式，将节点组成局部 cluster 后，再将 cluster 建为一个层次化树形拓扑结构，并在此层次化树形集群系统基础上建立组播树. 这种拓扑结构适合稳定节点的组网，大量组播源数据源在共享拓扑上建立起不同的源组播最短路径树. 拓扑管理中节点的离开和加入仅仅影响局部 cluster 内的节点和少量上级节点，对网络全局拓扑不产生影响. 并且通过设置两个 RP 节点的方式保证节点失效时的业务连续和快速恢复.

MCR 采用与 NICE 和 Zigzag 类似的层次化拓扑，将节点组成多个小集群系统，将小集群系统构建成层次化拓扑，并在拓扑基础上建立组播路由树. 与 NICE 和 Zigzag 不同的是，MCR 中的节点自上而下加入层次化拓扑，而且节点最多同时出现在两个相邻的层中，因此最多需要维护的邻接关系为 $4k-3$. NICE 和 Zigzag 中，节点自下而上加入拓扑，并且节点可以同时出现在多个层中.

MCR 能提供基于应用层组播的 S2S 服务对等网络，网络的节点都是服务器. 客户端可以通过任何一个节点接入，从而可以访问到整个网络的资源，即资源是由整个网络提供的.

参考文献：

[1] Waitzman D, Partridge C, Deering S. Distance Vector Multicast Routing Protocol [S]. RFC 1075, Internet Engineering Task Force, 1988.

[2] Fenner W. Internet Group Management Protocol, version 2 [S]. IETF RFC 2236, Internet Engineering Task Force, 1997.

[3] Diot C, Levine B N, Lyles B, et al. Deployment Issues for the IP Multicast Service and Architecture [J]. IEEE Network Mag, 2000, 14(1): 78－88.

[4] El-Sayed A, Roca V, Mathy L. A Survey of Proposals for an Alter-native Group Communication Service [J]. IEEE Network Mag, 2003, 17(1): 46－51.

[5] 刘书，潘成胜，张德育. Mobile Agent 在网络系统监控中数据采集的设计与应用[J]. 武汉工程大学学报，2010, 31(3): 77－80.

[6] Hosseini M, Ahmed D T, hirmohammadi S, et al. A Survey of Application-Layer Multicast Protocols [J]. IEEE Communications Surveys & Tutorials, 2007, 9(3): 58－74.

[7] Pendarakis D, Shi S, Verma D, et al. ALMI: An Application Level Multicast Infrastructure [A]. Proceedings of the 3rd conference on Usenix Symposium on Internet Technologies and Systems [C]. San Francisco: USENIX Association, 2001: 49－60.

[8] Tran D A, Hua K A, Do T T. A Peer-to-Peer Architecture for Media Streaming [J]. IEEE JSAC, 2004, 22(1): 121－33.

[9] Banerjee S, Bhattacharjee B, Kommareddy C. Scalable Application Layer Multicast [A]. Proceedings of the 2002 conference on Applications, technologies, architectures, and protocols for computer communications [C]. College Park: ACM, 2002: 205－17.

(下转第 97 页)

ZHANG Yan

(School of Chemical and Environmental Engineering, Jianghan University, Wuhan 430056, China)

Abstract: Due to the subjective and objective reasons, graduation design quality of engineering students has decreased. We think that a scientific and reasonable evaluation method is an important guide to improve the quality of graduation design. In this paper, an indices system for evaluation was stated and systemic model were proposed for evaluating graduation design quality of engineering students.

Key words: engineering students; graduation design quality; evaluation indices; evaluation methods

本文编辑:陈小平

☆

(上接第 93 页)

A topology design of application level multicast

WANG Li[1], *HUANG Jun-nian*[1], *YU Xin*[2], *LIU Di*[1]

(1. Department of Information Technology, Huazhong Normal University, Wuhan 430079, China;
2. Department of Electronics and Information Engineering,
Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: A mesh-first application level multicast topology for live-streaming CDN is proposed to support multiple source specific trees with one single topology. This topology organizes the multicasting nodes into an administrative hierarchy of clusters, and it builds routing trees atop this hierarchy.

Key words: application level multicast; mesh first; hierarchy topology; multicast tree

本文编辑:陈小平