ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.10, October 2006, pp.2152−2162
DOI: 10.1360/jos172152

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种具有能力约束性能的任意源覆盖多播方法*

陈世平[1,2+], 施伯乐[1]

[1](复旦大学 计算机与信息技术系,上海 200433)
[2](上海理工大学 计算机工程学院,上海 200093)

A Method for Any-Source Capacity-Constrained Overlay Multicast

CHEN Shi-Ping[1,2+], SHI Bai-Le[1]

[1](Department of Computer and Information Technology, Fudan University, Shanghai 200433, China)
[2](Department of Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200433, China)
+ Corresponding author: Phn: +86-21-65687765, Fax: +86-21-65687765, E-mail: chensp@usst.edu.cn

Chen SP, Shi BL. A method for any-source capacity-constrained overlay multicast. *Journal of Software*, 2006,17(10):2152−2162. http://www.jos.org.cn/1000-9825/17/2152.htm

Abstract: Many overlay multicast systems proposed in recent years focus on designing an optimized tree for a single data source. They cannot be extended to any-source multicasting because one tree per source is too costly. The existing P2P (peer-to-peer) systems that allow many data sources have high maintenance overhead and lack the flexibility in supporting host diversity. This paper proposes an any-source capacity-constrained overlay multicast service based on a non-DHT (distributed hash table) overlay network specifically suitable for the purpose of multicast. The nodes have different capacities in supporting different numbers of direct children during a multicast session. No explicit multicast trees are maintained on top of the overlay. This paper presents two distributed multicast algorithms that are able to deliver a multicast message from any source to all nodes in the expected $O(\log_c n)$ hops, which is asymptotically optimal, where c is the average node capacity and n is the number of members in a multicast group.

Key words: overlay multicast; capacity; non-DHT (distributed hash table) ring; hop complexity; communication complexity

摘 要: 近年来提出的许多面向单个数据源设计的多播树并不能简单扩展到任意源多播系统中,因为针对每个源建立一个树代价高昂.而已存在的一些允许多数据源的 P2P(peer-to-peer)系统的维护量大,在体现结点能力差异等方面缺少灵活性.提出一个任意源覆盖多播服务方案,并具有结点能力约束性能.它建立在非 DHT(distributed hash table)覆盖网络上,无须建立显式的多播树.设计了两种分布式多播算法,它们将任意源的多播信息传送到所有结点的期望跳数是 $O(\log_c n)$,其中,c 是平均结点能力,n 是多播组中的结点个数.

关键词: 覆盖多播;能力;非 DHT(distributed hash table)环;跳数复杂性;通信复杂性
中图法分类号: TP393 文献标识码: A

* Supported by the National Natural Science Foundation of China under Grant No.60573142 (国家自然科学基金); the Shanghai Natural Science Foundation of China under Grant No.02ZD14066 (上海自然科学基金)

Received 2005-09-12; Accepted 2005-12-31

多播是一个重要网络应用,它在分布式异构结点动态集合中实现组通信.许多文献提出了应用层的覆盖多播服务,文献[1−6]从不同角度对覆盖多播进行了研究.但现存系统尚不能对需要进行任意源多播且多播组中结点能力不同、组成员动态变化的应用提供有效支持.

文献[7]中证明了构造一个最小直径的有限度生成树是一个 NP 问题.需要说明的是,“度(degree)”与本文提出的“能力(capacity)”是一个类似的概念.集中式启发算法被用于在多播服务结点(MSN)中平衡多播数据流量,以降低端到端响应时间[7,8].不过,算法没有重视动态成员问题,如 MSN 的加入与离开.

一些覆盖多播系统致力于优化单个源的多播树:文献[9]改进了从一个源到一个接受组的数据分发吞吐量,它要求建立一个起始于源结点的覆盖树.分拆的数据目标被从源结点通过树分送到不同的接受者,接受者之间再通信以检索“丢失”的目标.树中的这些动态通信连接与树本身形成网状拓扑(mesh),以提供比单纯的多播树更好的带宽;文献[10]提出的覆盖多播网络基础设施(OMNI)采用具度约束性能的多播树来减少到整个客户端集合的时延;文献[11]提出算法来构造一个针对单源的度约束最小延迟多播树;文献[12]针对动态组提出一个分布式算法以维持有度约束能力的延迟敏感多播树.以上算法均面向单数据源设计,它们适合于分发视频或软件,但这并不适合诸如分布式游戏、电信会议、虚拟教室等多源多播应用,而这是本文面向的应用要求.为每个可能的源建立相应的多播树的代价高昂,而用单个多播树服务于多源同样问题很多.首先,针对某个源的最小延迟树并不见得是其他源的最小延迟树;第二,单树方法将所有流量集中在该树中的连接上,而使其他大多数结点的处理能力闲置不用,这影响了多源多播的总体吞吐能力;第三,由于多播树成员动态变化,结点可以随时加入和离开,单个树可能被切断,而这将使维护多播树成为一个严重问题.

为管理动态多播组并确保可扩展性,已存在其他一些在 P2P 覆盖网上实现多播的方法,如:Bayeux[13], Borg[14]分别是基于 Tapestry,Pastry 等覆盖拓扑结构而实现的;文献[15]研究了在 Chord 网络中进行有效广播的方法,并表明可应用于多播.但是,以上系统均假设每个结点具有相同数量的子结点.而事实上,多播组中结点在体现结点能力的上载带宽、内存、CPU 等方面各不相同.给通信组中每个中间结点指定相同数量子结点的策略显然不够优化:如果子结点数量设置太大,低能力结点将过载,这将导致整个会话速率下降;如果子结点数量太少,则高能力的结点将不能被充分利用.

本文提出一种任意源覆盖多播服务,它建立在随机覆盖网及一个非 DHT(distributed hash table)环的基础上,满足能力约束、动态成员等分布式应用需求,具有完全分布的特点,可以在 Internet 范围内扩展.另外,其结构简单、维护量小.

本文第 1 节定义问题及网络模型.第 2 节提出解决思路,给出一个分布式多播基本算法,并进行性能分析.第 3 节讨论实现中存在的主要问题.第 4 节描述改进算法.第 5 节给出模拟实验结果.最后对全文进行总结.

1 系统模型

针对一个具有 n 个结点的多播组 G,每个结点 $x \in G$ 能力为 c_x,它是 x 向其转发多播信息的直接子结点的最大个数.而 c_x 将与 x 的上载带宽相适应.直观地,当 x 有大的上载带宽时,它就能在多播树中支持更多的直接子结点.在异构环境中,不同结点的能力相差很大.本文的目标是构造一种有弹性的具能力约束的多播服务,它可以满足所有相关结点的能力限制,允许经常性的成员变化,并且可以通过一个动态平衡的多播树实现将多播信息从信息源向多播组成员传递.

为达到这个目标,需针对每个多播组建立一个覆盖网,这样,可以将多播问题转化为在覆盖范围内的广播问题.由于本文面向那些存在许多潜在数据源的应用,一个可能的方法是对每个源采用“隐式”多播树的思路.因为没有真正建立树,因而不会带来维护负荷,当覆盖拓扑发生变化时,隐式树自然地变化且开销为 0.

本文讨论的覆盖网由两部分组成:其中一部分结点间以随机形式相连,仍是一种树形结构;另一部分是一个非 DHT 环,本文称其为“非严格环”,它与所有结点相连.非严格环与大多数结构 P2P 网中基于 DHT 的环结构有根本的不同.在非严格环中,一个结点并不具有一个特定位置,而是可以在环中任意一个位置,环的维护量小.每个结点把其在环中的下一个结点作为其邻居之一,称为后继结点.一个新结点 x 的加入,首先要知道一个已存在

的结点 y.x 通知 y 将自己作为 y 的后继结点,而将 y 原先的后继结点作为 x 的后继结点.除了有后继结点作为邻居以外,x 还将从结点集中随机选取 c_x 或更多结点作为邻居,这方面的详细情况将在第 3 节加以讨论.

2 任意源覆盖多播

2.1 设计思路

随机行走(random walk)、有限洪泛(limited flooding)和概率洪泛(probabilistic flooding)被广泛应用于覆盖网上的共享文件查询.其工作方式效果不错,因为待搜索的目标往往有很多拷贝,这些拷贝分散在整个网络中,只需搜索一小部分结点就找到目标的概率较高,丢失一个存在的目标的概率很小.但这些方法不能简单应用于多播.对多播而言,一个覆盖网构造在多播组所有的结点上,组中每个结点都应当接受多播信息的一个拷贝,而这并不是随机行走或有限洪泛所能保证的.

为了更好地评估多播算法,本文对具有不同能力的结点构成的覆盖多播提出两项性能测评指标:跳数复杂性与通信复杂性.假设 c 为平均结点能力,而 n 为多播组结点集 N 中结点个数.跳数复杂性定义为多播信息到达任一结点所要经过的跳数.通信复杂性是指一个多播信息要被传送到所有结点所需要转发的拷贝份数.较小的跳数复杂性意味着多播树更为均衡,且平均传输时延更小.较小的通信复杂性意味着消耗较少的网络带宽.本文将设计多播方案来平衡两种复杂性,在此之前,先介绍两种相对极端的多播方案.

完全洪泛:在该方案中,每个结点第一次接受到多播信息后,将向其随机邻居发送该信息,这样,跳数复杂性为 $O(\log_c n)$,这是一个较好的结果,因为洪泛沿最短路径实施.通信复杂性是 cn,每个结点 x 发送 c_x 个拷贝;

环行传送:这是另一种方案,它获得的最好通信复杂性是 n,一个信息沿环顺序地与所有结点连接.但是,跳数复杂性是 $n/2$,这是所有方案中最差的结果.

本文提出了一种多播方案:ACOM(any-source capacity-constrained overlay multicast).它在跳数复杂性和通信复杂性之间达成平衡.它能获得一个较好的跳数复杂性 $O(\log_c n)$,而通信复杂性为 $n+O(n/\log_c n)$.由表 1 可以看到几种方案复杂性的对比.

Table 1 Comparison of complexities for some multicast schemes

表 **1** 几种多播方案复杂性比较

Multicast schemes	Hop complexity	Comm. complexity
Fully flooding	$O(\log_c n)$	cn
Ring traversal	$n/2$	n
ACOM	$O(\log_c n)$	$n+O(n/\log_c n)$

ACOM 设计的主要目标是一个多播例程:(1) 它能将一个信息从任一源结点发往每一个结点;(2) 跳数复杂性是 $O(\log_c n)$;(3) 通信复杂性为 $n+O(n/\log_c n)$;(4) 无须创造并维护显式覆盖多播树.

主要思路是执行两阶段多播:第一阶段(随机转发阶段)实行随机分布式传送;而第二阶段(环行转发阶段)实行分段环行传送.所定义的每个源的隐式多播树建立在已存在的覆盖网结构中,避免了在覆盖网上的多播树维护问题.这些隐式树大致均衡并且受能力限制,它们针对的每个源都不一样,并且将结点负荷分散到尽可能多的覆盖连接上.

随机阶段:源结点将信息通过随机邻居向其他结点发送,这部分传递控制在 K 跳之内.K 是一个系统参数,并将在后续讨论中确定其意义.该阶段将信息传递到结点集的一个子集上,该子集结点随机分布于网络上,并被称为随机阶段结点.图 1 中前两个子图给出了示例.假设 K=2,源结点 s 能力是 3,结点 i,j,l 能力分别是 2,2 和 3.这样,在随机阶段,信息被发送到树结构的 10 个结点上.

环行阶段:每个随机阶段结点将信息发送给其“邻居”.洪泛会造成一些结点收到该信息的多个拷贝,为避免这个问题,本文使用了非严格环.利用随机阶段结点将环划分成段.如图 1(c)所示,每个随机阶段结点负责一个邻近段,它将信息发送给其后继结点,信息进而又被发送给后者的后继结点,…,直到信息到达一个已接收到该信息的结点为止.

两个阶段的信息传送仍然是在一个树结构中,如图 1(d)所示,由于树本身遵从底层覆盖网结构,因而不需要额外的维护.环行阶段的分段环形传送保证每个结点只接受到一个信息拷贝.如果 K 值足够小,随机阶段中的邻居的随机分布会使得信息传送给不同的随机结点的可能性很大,结点接受多于一个拷贝的概率会很小,这也有助于减少通信复杂性.如果 K 值足够大,会使得更多的随机阶段结点将环划分为小的段,这将有助于在环行阶段减少跳数复杂性.由 K 的这两个有冲突的需求产生了一个有趣的问题,即:能否选择一个合适的 K,以使跳数复杂性达到优化的 $O(\log_c n)$,而且通信复杂性接近 n?由于结点有不同数量的邻居,因此使得对这个问题的解决变得更加困难了.下面将首先给出一种基本算法,并将分析其特性.

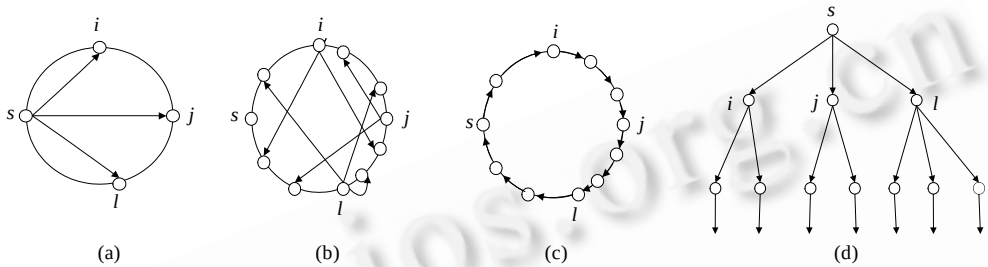

Fig.1 Two-Phase multicast

图 1 两阶段多播

2.2 基本算法(ACOM-1)

一个多播信息表示为 $M(k,id)$,其中,M 是信息;k 是 TTL 属性,其初始值为 K;id 是一个全网唯一的标识符,它由源 ID(如 IP 地址)和一个顺序号组成.传递一个多播信息,源结点 s 首先将 $M(k,id)$发送给 c_s 个随机邻居.而每个被选中的随机邻居或非严格环上后继结点将根据算法继续转发,直至满足终止条件.该算法描述如下.显然,该算法计算复杂度为 $O(c_x)$,这与泛洪算法具相同的计算复杂度.

/*当结点 x 接受到多播信息 $M(k,id)$时,将执行以下算法过程*/

/*阶段 1*/

(1) **if** k>0 and x 尚未向随机邻居发送标识为 id 的多播信息 M **then**

(2) 选择 c_x 个随机邻居

(3) **for** 每一个选中的邻居 y **do**

(4) 发送 $M(k-1,id)$给 y

(5) **if** x 尚未向 $successor(x)$发送标识为 id 的多播信息 M **then**

(6) 发送 $M(0,id)$给 $successor(x)$

/*阶段 2*/

(7) **else if** k=0 and x 尚未向 $successor(x)$发送标识为 id 的多播信息 M **then**

(8) 发送 $M(0,id)$给 $successor(x)$

(9) **else**

(10) 丢弃该多播信息

2.3 性能分析

在随机转发阶段,多播信息是在一个树形结构上传送,称为随机阶段树.树由随机阶段结点组成,树的深度为 K.源结点 s 在第 0 级.设 q_i 是第 i 级的结点数量,q_0=1,q_1=c_s,而 $q_i(i\in[2,\ldots,K])$是一个随机数,因为第 i−1 级及之前各级的结点的随机邻居可能与第 i 级结点的邻居重叠.在随机阶段树中的从第 0 级~第 i 级的结点数量是 $Q_i=\sum_{j=0}^{i} q_j$. 随机阶段树中的结点总数为 Q_K.多播树中间结点个数为 Q_{K-1},这些结点将多播信息传递给它们的随机邻居.而每一个中间结点的子结点数量由其本身能力所限.假设 $c_{i,1},c_{i,2},\ldots,c_{i,q_i}$ 是第 i 级的结点的能力,

则所有中间结点的能力总和为$1+\sum_{i=0}^{k-1}\sum_{j=0}^{q_i}c_{i,j}$.显然,它是树规模 Q_K 的上限.为避免结点对同样信息过量的重复接收,本文设计原则是让随机阶段中信息只传递到结点集 N 中的少量结点上.通过利用随机阶段结点将非严格环划分为段,这样在环行转发阶段,多播信息在各段并行地以环形方式被转发到段中所有其他结点上.因此,可以选择一个 K 使树规模 Q_K 远远小于 n.我们可以将随机阶段树大小按 $\sum_{i=0}^{K}c^i$ 来设定,同时定义一个系统参数 λ,使以下等式成立:

$$\sum_{i=0}^{K}c^i=\lambda n \tag{1}$$

λ表示了“预设”的 N 中结点被随机阶段树覆盖的结点所占的比例.显然,当选择了λ值,K 值可由公式(1)计算得出.因此,确定适当的 K 值问题变成了确定合适λ值的问题.下面将要证明:当选择$\lambda=\Theta(1/\log_c n)$时,跳树复杂性将为 $O(\log_c n)$,而通信复杂性为 $n+O(n/\log_c n)$.

首先给出 Q_K 的上、下限,然后导出算法的跳数复杂性和通信复杂性.

引理 **1**. $E(Q_i)\le\sum_{j=0}^{i}c^j$.

证明:第 i 级的结点数量上限是第 $i-1$ 级所有结点能力之和,即 $q_i\le\sum_{j=1}^{q_{i-1}}c_{i-1,j}$.

在第 $i-1$ 级存在 q_{i-1} 个结点前提下,q_i 期望值为

$$E(q_i|q_{i-1})\le E\left(\sum_{j=1}^{q_{i-1}}c_{i-1,j}\right)=q_{i-1}c.$$

两边针对 q_{i-1} 求期望值,由于 $E(E(q_i|q_{i-1}))=E(q_i)$,我们有

$$E(q_i)\le E(q_{i-1})c,$$

上面不等式不断做迭代,则有

$$E(q_i)\le c^iE(q_0)=c^i \tag{2}$$

那么,

$$E(Q_i)=E\left(\sum_{j=0}^{i}q_i\right)\le\sum_{j=0}^{i}c^j \tag{3}$$

引理得证. □

引理 **2**. 对任一个条件集合 S,若有 $\left\{q_1,...,q_i\middle|Q_i\le\sum_{j=0}^{i}c^j,S\right\}\neq 0$,则以下不等式成立:

$$E(Q_i|S)\ge E\left(Q_i\mid Q_i\le\sum_{j=0}^{i}c^j,S\right).$$

证明:这是显然的,因为当更多的数据项被排斥,期望值更小.

引理 **3**. 对任一个条件集合 S,若有 $\left\{q_1,...,q_i\middle|Q_i\le\sum_{j=0}^{i}c^j,S\right\}\neq 0$,则以下不等式成立:

$$E\left(Q_i\mid Q_i\le\sum_{j=0}^{i}c^j,S\right)\ge c\left(1-\frac{\lambda}{c^{K-i}}\right)E\left(Q_{i-1}\mid Q_i\le\sum_{j=0}^{i}c^j,S\right).$$

证明:假设 T_i 是随机阶段树中包含有从第 0 级~第 i 级结点的一个子树,T_i 大小是 Q_i.在 T_i 形成过程中,每个从第 0 级~第 $i-1$ 级的结点 x 均会尝试将其随机邻居纳入到下一个级别中,由于树中有少于 Q_i 个结点,那么每一次尝试至少有 $\left(1-\frac{Q_i}{n}\right)$ 的成功概率,即:随机邻居在当前树之外的概率不小于 $\left(1-\frac{Q_i}{n}\right)$.树在这样不断增长,从第 0 级~第 $i-1$ 级将一共有 $\sum_{l=0}^{i-1}\sum_{j=1}^{q_l}c_{l,j}$ 次尝试,因此,$Q_i>\sum_{l=0}^{i-1}\sum_{j=1}^{q_l}c_{l,j}\left(1-\frac{Q_i}{n}\right)$.

对于出现在 $\left\{q_1,...,q_i\middle|Q_i\le\sum_{j=0}^{i}c^j,S\right\}$ 中任何集合 q_l,计算 Q_i 的条件期望值,

$$E\left(Q_i \middle| q_l, l\in[1...i-1], Q_i \le \sum_{j=0}^{i} c^j, S\right) > E\left(\sum_{l=0}^{i-1}\sum_{j=1}^{q_l} c_{i,j}\left(1-\frac{Q_i}{n}\right)\right)$$

$$\ge E\left(\sum_{l=0}^{i-1}\sum_{j=1}^{q_l} c_{i,j}\left(1-\frac{\sum_{j=0}^{i} c^j}{n}\right)\right)$$

$$=\left(1-\frac{\sum_{j=0}^{i} c^j}{n}\right)\sum_{l=0}^{i-1}\sum_{j=1}^{q_l} E(c_{l,j})$$

$$=c\left(1-\frac{\sum_{j=0}^{i} c^j}{n}\right)\left(\sum_{l=0}^{i-1}\sum_{j=1}^{q_l} 1\right)$$

$$=c\left(1-\frac{\sum_{j=0}^{i} c^j}{n}\right)Q_{i-1} \ge c\left(1-\frac{\lambda}{c^{K-i}}\right)Q_{i-1}.$$

基于以上不等式,仍在 $Q_i \le \sum_{j=0}^{i} c^j$ 和 S 条件下,求取对所有 q_l 的 Q_i 期望值,有

$$E\left(Q_i \middle| Q_i \le \sum_{j=0}^{i} c^j, S\right) \ge c\left(1-\frac{\lambda}{c^{K-i}}\right)E\left(Q_{i-1} \middle| Q_i \le \sum_{j=0}^{i} c^j, S\right).$$ □

引理 **4**. $E(Q_i) > \left(1-\frac{\lambda}{c^{K-i-1}(c-1)}\right)c^i$.

证明:由引理 2,$E(Q_i) \ge E\left(Q_i \middle| Q_i \le \sum_{j=0}^{i} c^j\right)$.

设 S_0=0,由引理 3,有

$$E(Q_i) > c\left(1-\frac{\lambda}{c^{K-i}}\right)E\left(Q_{i-1} \middle| Q_i \le \sum_{j=0}^{i} c^j, S_0\right).$$

设 $S_1 = S_0 + \left\{Q_i \le \sum_{j=0}^{i} c^j\right\}$,由引理 2 及引理 3,可得

$$E(Q_i) > c\left(1-\frac{\lambda}{c^{K-i}}\right)E(Q_{i-1} \mid S_1)$$

$$\ge c\left(1-\frac{\lambda}{c^{K-i}}\right)E\left(Q_{i-1} \middle| Q_{i-1} \le \sum_{j=o}^{i-1} c^j, S_1\right)$$

$$> c^2\left(1-\frac{\lambda}{c^{K-i}}\right)\left(1-\frac{\lambda}{c^{K-i+1}}\right)E\left(Q_{i-2} \middle| Q_{i-1} \le \sum_{j=o}^{i-1} c^j, S_1\right).$$

应用引理 2、引理 3,重复进行迭代,我们有

$$E(Q_i) > c^i \prod_{j=1}^{i}\left(1-\frac{\lambda}{c^{K-j}}\right)E(q_0) = c^i \prod_{j=1}^{i}\left(1-\frac{\lambda}{c^{K-j}}\right).$$

通过归纳,$\prod_{j=1}^{i}\left(1-\frac{\lambda}{c^{K-j}}\right) > 1-\frac{\lambda}{c^{K-i-1}(c-1)}$,因此,$E(Q_i) > \left(1-\frac{\lambda}{c^{K-i-1}(c-1)}\right)c^i$. □

引理 **5**. $\left(1-\frac{1}{c}\right)\left(1-\frac{\lambda c}{c-1}\right)\lambda n < E(Q_K) \le \lambda n$.

证明:由引理 1, $E(Q_K) \le \sum_{i=0}^{K} c^i = \lambda n$. 又经引理 4,

$$E(Q_i) > \left(1-\frac{\lambda c}{c-1}\right)c^K \tag{4}$$

由公式(1)得 $c^K=\frac{\lambda n(c-1)+1}{c}>\frac{\lambda n(c-1)}{c}$,将其代入公式(4)有 $E(Q_i)>\left(1-\frac{1}{c}\right)\left(1-\frac{\lambda c}{c-1}\right)\lambda n$. □

定理 **1**(跳数复杂性). 一个多播信息到达任意结点的跳数期望值不超过 $\log_c(\lambda n)+\frac{1}{2\left(1-\frac{1}{c}\right)\left(1-\frac{\lambda c}{c-1}\right)\lambda}$.

证明:由公式(1),$\sum_{i=0}^{K}c^i=\lambda n$.因此 $c^K<\lambda n$,则有 $K<\log_c(\lambda n)$,意味着随机阶段在不超过 $\log_c(\lambda n)$跳内终止.而由引理 5 可知:在随机阶段后,环被划分为不少于 $\left(1-\frac{1}{c}\right)\left(1-\frac{\lambda c}{c-1}\right)\lambda n$ 段数,环行阶段以并行方式将多播信息在这些段中传送.而一个段的长度的期望值上限为 $\frac{n}{\left(1-\frac{1}{c}\right)\left(1-\frac{\lambda c}{c-1}\right)\lambda n}=\frac{1}{\left(1-\frac{1}{c}\right)\left(1-\frac{\lambda c}{c-1}\right)\lambda}$.

在环行阶段,信息到达一个结点的平均跳数是段长度的一半.这样,结合随机阶段和环行阶段,一个多播信息到达任何结点的跳数的期望值上限是 $\log_c(\lambda n)+\frac{1}{2\left(1-\frac{1}{c}\right)\left(1-\frac{\lambda c}{c-1}\right)\lambda}$. □

推论 **1**(跳数复杂性). 如果 $\lambda=\Theta\left(\frac{1}{\log_c n}\right)$,那么,一个多播信息到达任意一个结点的跳数期望值为 $O(\log_c n)$.

证明:由定理 1 可以推导出该结论.限于篇幅,此过程略. □

定理 **2**(通信复杂性). 一个多播信息传输过程中的拷贝份数期望值不超过$(1+\lambda)n$.

证明:随机阶段中信息在传输中的拷贝份数 X 等于从第 0 级~第 $K-1$ 级所有结点的能力之和.

$$X=\sum_{i=0}^{K-1}\sum_{j=1}^{q_i}c_{i,j},$$

X 是反映 Q_{K-1} 的一个独立随机变量,Q_{K-1} 自己也是一个随机变量.若给定 Q_{K-1} 一个值,则 X 的条件期望值是

$$E(X\mid Q_{K-1})=\sum_{i=0}^{K-1}\sum_{j=1}^{q_i}E(c_{i,j})=\sum_{i=0}^{K-1}\sum_{j=1}^{q_i}c=cQ_{K-1}.$$

针对 Q_{K-1} 在等式两边求取期望值,由公式(3)和公式(1),有

$$E(X)=cE(Q_{K-1})\le c\sum_{i=0}^{K-1}c^i<\lambda n \tag{5}$$

因此,传输一个多播信息的信息拷贝份数期望值是$(1+\lambda)n$. □

推论 **2**(通信复杂性). 如果 $\lambda=\Theta\left(\frac{1}{\log_c n}\right)$,那么,一个多播信息在传输过程中的信息拷贝份数期望值不超过 $n+O(n/\log_c n)$.

证明:直接从定理 2 得出. □

3 实现中的主要问题

3.1 建立随机邻居

当一个新结点 x 加入环时,它可以请求一个多播源 s 帮助以找到自己的 c_x 个随机邻居.当 s 多播一个信息时,它将附上 c_x 个令牌,其中含 x 的地址.每个令牌会独立地选择在随机阶段树中的一个没有携带令牌信息的随机路径传递.当一个令牌到达了叶结点,它将沿环行阶段的非严格环中的某一段进行传递.当其开始传递时,首先计算各段长度 L,当它到达了段尾,从$[0\ldots L-1]$中取一随机数 r,令牌将反向行进 r 跳,所到达的结点将自己报告给 x.由于多播组中每一结点均将收到该多播信息,这样,每个结点都有机会成为 x 的邻居.为实现该方案,环中每个结点要知道它的前驱结点.在随机阶段中会发生一种称为小概率的事件,即携带令牌的多播信息传递给了一个

已收到多播信息的结点,该结点将丢弃该多播信息,但仍将自己报告给 x.由于多播信息传输到达任一个结点的跳数期望值为 $O(\log_c n)$,则显然,c_x 个随机邻居选取的通信复杂性为 $O(c_x\log_c n)$.

当 x 加入并作为结点 z 的后继结点时,如果没有多播源信息,它可以请求 z 来模拟以上过程,由 z 送出 c_x 个令牌.每个令牌独立执行 $\log_c n$ 量级跳数的随机行走(random walk)过程.每一跳均选择一个随机邻居继续前进,在随机行走结束后,令牌将沿环传送介于 0~$\log_c n$ 之间的随机跳,而最终接受到令牌的结点将自己报告给 x.由此可知:令牌在经平均 $1.5\log_c n$ 跳传递后,可以找到 c_x 个随机邻居.整个随机邻居选取过程通信复杂性的期望值为 $O(1.5c_x\log_c n)$.另外需要说明的是,c_x 个随机邻居的选取是在新结点加入多播组时发生,并不需要伴随结点的每一次多播而发生.

3.2 测量n和c

由于多播信息将到达每一个结点,n 的计算可以在多播过程中实现.多播源将定期通过发送附带一个标记的多播信息来帮助实现.当附带标记的信息被分发,则每个结点将暂时记忆其父结点,结点计数信息将分别从环行阶段中非严格环的各分段的最后结点反向传回源结点,这样,n 值就确定了.源结点然后通过多播信息将 n 分送给所有其他结点.

考虑到在此过程中一些结点会离开网络,当结点 x 发现其到子结点的连接断了或子结点的结点计数回送超时,则该结点将自动赋予该子结点回送计数值为来自其他子结点返回的结点计数平均值,然后将接收到的所有结点计数加 1 后回送该结点之父结点.为了改进准确性,源结点可以在一段时间内执行若干次 n 计数测量,并将这一时段中测量的 n 最大值作为 n 最新的测量值.c 值与 n 同时测量,平均结点能力和 n 计数一起将被回送到源结点.

4 改进算法(ACOM-2)

依据算法 ACOM-1 的设计,源结点首先选择一个具有 $1/\log_c n$ 量级的λ值,对于一个具有λn 量级结点规模的随机阶段树,可以从公式(1)计算 K 值

$$K=\log_c(\lambda n(c-1)+1)-1 \tag{6}$$

由于 K 可能不是一个整数,这样,源结点可能通过发送 $M(\lfloor K\rfloor,id)$或 $M(\lceil K\rceil,id)$来初始化分布式算法 ACOM-1,这可能导致算法结果的不准确.如果使用 $M(\lfloor K\rfloor,id)$,随机阶段树的规模就不是λn 而是 $\sum_{i=0}^{\lfloor K\rfloor}c^i$,而这可能导致随机阶段树大小为$(1/c)\lambda n$;另一方面,如果 $M(\lceil K\rceil,id)$被采纳,随机阶段树就有了 $c\lambda n$ 的规模.为了保持树规模在λn 左右,随机阶段树应当控制在第 0 级~第$\lfloor K\rfloor$+1 级,只不过第$\lfloor K\rfloor$+1 级只有部分结点有多播信息到达.即,这一级的每个结点按概率 p 被传送,p 可如下计算:

$$\left(\sum_{i=0}^{\lfloor K\rfloor}c^i\right)+c^{\lfloor K\rfloor+1}p=\lambda n,$$

$$P=\frac{\lambda n-\dfrac{c^{\lfloor k\rfloor+1}-1}{c-1}}{c^{\lfloor k\rfloor+1}}.$$

源结点 s 将增加了一个域 p 的 $M(\lfloor K\rfloor,p,id)$发出,当一个结点 x 收到一个信息 $M(0,p,id)$时,它将按概率 p 将该信息转发给其 c 个随机邻居.该算法描述如下:

/*当结点 x 接受到多播信息 $M(k,p,id)$时,将执行以下算法过程*/

/*阶段 1*/

(1) **if** k>0 and x 尚未向随机邻居发送标识为 id 的多播信息 M **then**

(2) 选择 c_x 个随机邻居

(3) **for** 每一个选中的邻居 y **do**

(4) 发送 $M(k-1,p,id)$给 y

(5) **if** x 尚未向 $successor(x)$发送标识为 id 的多播信息 M **then**

(6) 发送 $M(0,0,id)$给 $successor(x)$

(7) **else if** k=0 and p>0 and x 尚未向随机邻居发送标识为 id 的多播信息 M **then**

(8) 选择 c_x 个随机邻居

for 每一个选中的邻居 y **do**

以概率 p 选择将 $M(0,0,id)$发送给 y

(11) **if** x 尚未向 $successor(x)$发送标识为 id 的多播信息 M **then**

(12) 发送 $M(0,0,id)$给 $successor(x)$

/* 阶段 2*/

(13) **else if** k=0 and x 尚未向 $successor(x)$发送标识为 id 的多播信息 M **then**

(14) 发送 $M(0,0,id)$给 $successor(x)$

(15) **else**

(16) 丢弃该多播信息

5 实 验

覆盖网按 10 000 个结点的规模建立.由于每个结点都是潜在的数据源,我们不对每个结点维护一个显式多播树,而是让多播直接在覆盖网上执行,该覆盖网由上面叙述的随机树和非严格环组成.本实验同时实现了 5 个多播算法,包括完全洪泛、概率洪泛、有限度(LD)洪泛、ACOM-1(K 按$\lfloor K \rfloor$取值)、ACOM-2.完全洪泛算法在 IP 网中是非常流行的广播算法,在本实验中,每个结点第一次接受到多播信息后将向所有随机邻居发送一个拷贝;而概率洪泛则是为了控制网络负载,按一定概率向邻居转发多播信息;在 LD 洪泛算法中,结点只向固定数目的随机邻居转发多播信息;ACOM-1 算法中,K 取值来源于公式(6),K 最终取值为$\lfloor K \rfloor$.由于 ACOM-2 比 ACOM-1 效果更好,在一些实验中,只将 ACOM-2 参与了比较.

第 1 组模拟实验比较了不同算法在传送多播信息时的平均性能,图 2 显示了这些算法针对平均结点能力产生的网络流量情况.如果平均结点能力为 c,结点能力从均匀分布的取值范围[2,2(c−1)]中选取.在概率洪泛中,结点向邻居转发信息的概率为 30%;而在 LD 洪泛中,每个结点向两个邻居转发信息;在 ACOM-2 算法中,λ取值 0.25.从图 2 中可知:完全洪泛算法产生了很大的网络流量;概率洪泛则在平均结点能力小时产生较小流量;而 LD 洪泛产生的流量适中,并且对结点能力不敏感.图 3 显示概率洪泛和 LD 洪泛算法不能保证多播信息到达多播组中所有结点,一种解决办法是结合本文提出的非严格环实现多播信息全组到达.当一个结点首次接受到一个多播信息时,它除了向随机邻居转发外,还向邻接的环上结点传送,该环上传送将一直持续到一个已收到该信息的结点为止.但是,该分段进行的环上传送也导致转发 n(本实验中,n=10000)个多播信息拷贝,相应效果可如图 4 所示.从中可以看出,ACOM-2 算法的网络流量相对较少.

图 5 中比较了各算法完成多播的平均跳数.洪泛算法显然平均跳数最小;ACOM-2 算法效果比 LD 洪泛算法要好,但比完全洪泛平均跳数要多;而概率洪泛在平均结点能力大时有较好效果.但如图 4 所示,此时其负面因素是网络流量大.

第 2 组模拟主要研究 ACOM 算法在面向λ取值变化时的性能变化情况,所使用的平均结点能力设为 6,而结点能力取值来源于一个取值范围为[2,10]的满足均匀分布的数据集合.图 6、图 7 显示出 ACOM 算法跳数复杂性与通信复杂性之间的关系.λ增加,ACOM 减少时延,但显然网络流量增大了.由图 6 可知,较大的λ意味着随机阶段多播树大,这将导致网络流量的增加.环行阶段网络流量始终是 n 次转发拷贝.另一方面,从图 7 反映出:较大的随机阶段多播树意味着非严格环每个分段更小,这减少了跳数及环行阶段时延,也使得总体的跳数、时延减少.ACOM-1 在λ发生变化时,性能表现较为稳定,主要原因是,根据公式(6),λ的改变并不是总能导致$\lfloor K \rfloor$的变化.

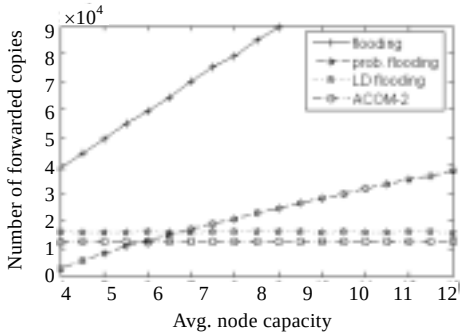

Fig.2 Comparison of different algorithms on traffic volume for delivering one multicast message

图 2 不同算法传送一个多播信息的网络流量比较

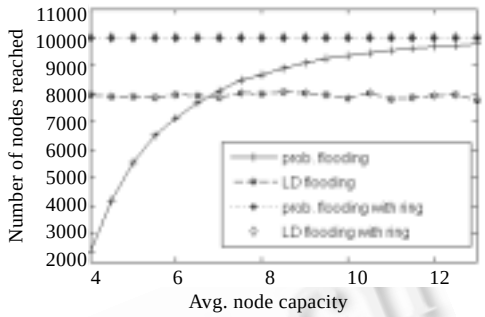

Fig.3 Number of nodes reached by a multicast message, not all algorithms deliver the message to all nodes

图 3 一个多播信息到达的结点数,并非所有算法可以将信息传送给所有结点

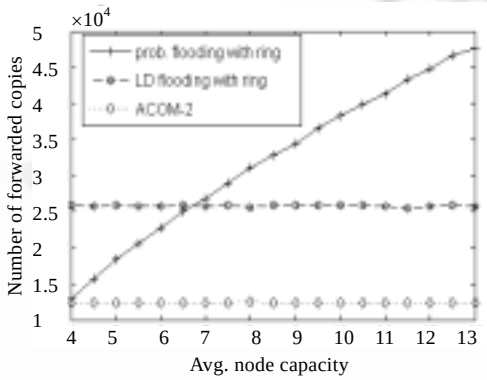

Fig.4 Comparison on traffic volume for delivering one message. All algorithms perform segmented ring traversal

图 4 传送一个多播信息的网络流量比较,所有算法均实施了分段环形传送

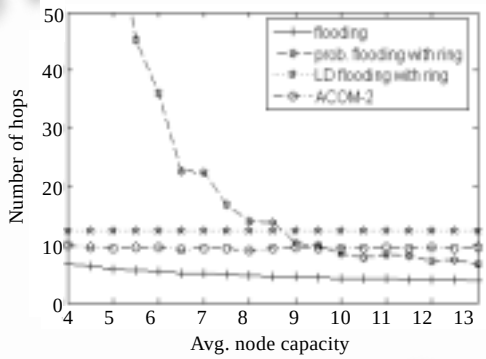

Fig.5 Comparison of the algorithms on average length of delivery paths

图 5 不同算法传送信息平均路径长度比较

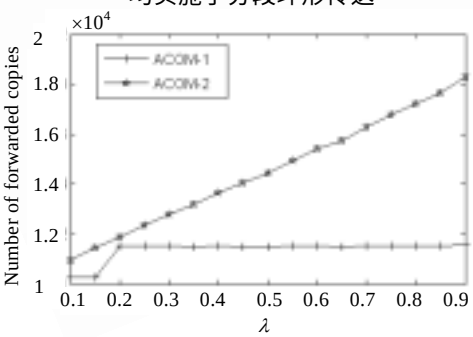

Fig.6 Comparison of ACOM algorithms on traffic volume with respect to λ

图 6 不同 ACOM 算法网络流量受λ影响的比较

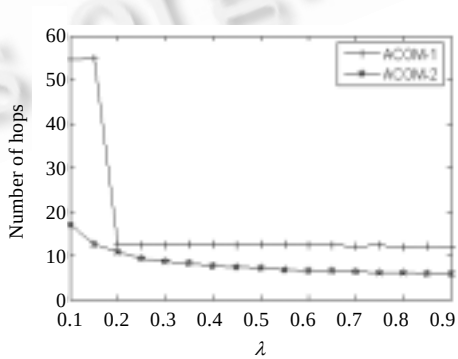

Fig.7 Comparison of ACOM algorithms on average length of delivery paths with respect to λ

图 7 不同 ACOM 算法传送信息平均路径长度受λ影响的比较

6 结 论

本文在覆盖网和非严格环基础上,提出了一种任意源覆盖多播服务.在运行过程中,算法的分布式执行将自然产生一棵负载大致均衡、受结点能力约束的隐式多播树,可供任意源结点传输多播信息,并且树中不同能力结点可以支持不同数量的转发子结点.我们对该系统性能评估给出了严格的分析,并给出了模拟实验.本文提出了两种分布式多播算法,并着重说明了实现中的一些关键技术,包括覆盖维护、参数测量等.

References:

[1] Kwon GI, Byers JW. ROMA: Reliable overlay multicast with loosely coupled TCP connections. In: Proc. of the INFOCOM 2004. 2004. http//www.cs.bu.edu/techreports/pdf/2003-015-roma.pdf

[2] Shavitt Y, Tankel T. On the curvature of the Internet and its usage for overlay construction and distance estimation. In: Proc. of the INFOCOM 2004. 2004. http://www.ieee-infocom.org/2004/Papers/09_1.pdf

[3] Baccelli F, Chaintreau A, Liu Z, Riabov A, Sahu S. Scalability of reliable group communication using overlays. In: Proc. of the INFOCOM 2004. 2004. http://www.ieee-infocom.org/2004/Papers/09_5.pdf

[4] Yue GX. A routing search algorithm on P2P networks based on Gnutella protocol: Light-Flooding. Computer Engineering, 2005, 31(11):112−114 (in Chinese with English abstract).

[5] Wu JG, Ye XG, Jiang AQ. A routing algorithm in heterogeneous overlay multicast networks. Journal of Software, 2005,16(6): 1112−1119 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/1112.htm

[6] Li Z, Mohapatra P. Impact of topology on overlay routing service. In: Proc. of the INFOCOM 2004. 2004. http://www.ieee-infocom.org/2004/Papers/09_4.pdf

[7] Shi S, Turner J, Waldvogel M. Dimensioning server access bandwidth and multicast routing in overlay networks. In: Proc. of the NOSSDAV 2001. 2001. http://www.inf.uni-konstanz.de/disy/publications/waldvogel/shi01dimensioning.pdf

[8] Shi S, Turner J. Routing in overlay multicast networks. In: Proc. of the INFOCOM 2002. New York: IEEE, 2002.

[9] Dejan Kosti JA, Rodriguez A, Vahdat A. Bullet: High bandwidth data dissemination using an overlay mesh. In: Proc. of the SOSP 2003. 2003. http://www.cs.rochester.edu/sosp2003/papers/p183-kostic.pdf

[10] Banerjee S, Kommareddy C, Kar BBK, Khuller S. Construction of an efficient overlay multicast infrastructure for real-time applications. In: Proc. of the INFOCOM 2003. 2003. http://www.ieee-infocom.org/2003/papers/37_03.pdf

[11] Riabov A, Liu Z, Zhang L. Overlay multicast trees of minimal delay. In: Proc. of the ICDCS 2004. Tokyo: IEEE Computer Society, 2004. 654−661.

[12] Yamaguchi H, Hiromori A, Higashino T, Taniguchi K. An autonomous and decentralized protocol for delay sensitive overlay multicast tree. In: Proc. of the ICDCS 2004. Tokyo: IEEE Computer Society, 2004. 662−669.

[13] Zhuang S, Zhao B, Joseph A, Katz R, Kubiatowicz J. Bayeux: An architecture for scalable and fault-tolerant WideArea data dissemination. In: Proc. of the 11th Int'l Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV 2001). New York: ACM, 2001. 11−20.

[14] Zhang R, Hu YC. Borg: A hybrid protocol for scalable application-level multicast in peer-to-peer networks. In: Proc. of the NOSSDAV 2003. Monterey: ACM, 2003. 172−179.

[15] El-Ansary S, Alima LO, Brand P, Haridi S. Efficient broadcast in structured P2P networks. In: Proc. of the IPTPS 2003. London: Springer-Verlag, 2003. 304−314.

附中文参考文献:

[4] 乐光学.基于 Gnutella 协议的 P2P 网络路由搜索算法:Light-Flooding.计算机工程,2005,31(11):112−114.

[5] 吴家皋,叶晓国,姜爱全.一种异构环境下覆盖多播网络路由算法.软件学报,2005,16(6):1112−1119. http://www.jos.org.cn/1000-9825/16/1112.htm

陈世平(1964－),男,浙江绍兴人,博士生,教授,主要研究领域为信息检索,数据库与知识库,计算机网络通信.

施伯乐(1935－),男,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,知识库.