分层组播拥塞控制策略对组播树稳定性的影响

石 锋, 吴建平, 徐 恪
(清华大学计算机科学与技术系网络技术研究所, 北京 100084)

摘 要: 文章考察在分层组播拥塞控制环境下网络拥塞对组播树(单对多)稳定性的影响, 定义一个稳定性因子来评估和量化这种影响. 并提出一个简单的链路相关拥塞模型, 得到稳定性因子的通用表达式. 模拟结果表明, 即使在链路标记概率较低的情况下, 随着不同链路相关度的减小, 拥塞对组播树稳定性会产生非常大的影响.

关键词: 分层组播; 拥塞控制; 稳定性; TCP 友好
中图分类号: TN915 文献标识码: A 文章编号: 0372-2112 (2003) 12-1639-04

Impact of Congestion on the Stability of Multicast Trees in Cumulative Layered Multicast Schemes

SHI Feng, WU Jian-ping, XU Ke
(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: This paper studies the impact of congestion on the stability of a multicast (one-to-many) tree in the context of cumulative layered multicast congestion control mechanism. A stability factor is defined to evaluate and quantify this impact. For obtaining the general expression of the stability factor, a simple statistical model is developed. Simulation results shows that even in the case of lower link-marking probability, the stability of multicast tree is significantly more sensitive to network congestion when dependency between different links becomes smaller, indicating that more links will leave the tree when congestion occurs.

Key words: layered multicast; congestion control; stability; TCP-friendly

1 引言

设计一种可扩展性好、TCP 友好(TCP-Friendly)的组播拥塞控制协议对组播研究是个重要的挑战. 最近, 分层组播技术得到了较广泛的应用. 它使用了多个组播组以不同的速率发送数据, 满足不同用户群的需求, 从而具有较好的可扩展性. 为了解决视频传输中用户的层次化问题, McCanne 等人提出的 RLM(Receiver-driven Layered Multicast)[1], 是分层组播方面最早的工作之一. RLM 使用累积式分层(cumulative layering), 要求用户按照分层顺序加入和离开对应的组播组. 每订阅一个新层, 用户能得到更好的视频质量. 为了减少编码复杂度、简化层的管理, RLC(Receiver-driven Layered Congestion Control)[2] 提出以下的分层策略: 与基数层对应的组播组使用速率 B_0 发送数据, 基数层以上的层以速率 $B_0 @ 2^{i-1}$ 发送数据.

分层组播通过组管理和路由机制间接实现拥塞控制. 在上述的分层方案中, 订阅新的一层使用户的接收速率加倍; 而离开最新订阅的层, 用户的接收速率减半. 这使得累积式分层中粗粒度的拥塞控制成为可能: 用户定期的执行加入试验和离开试验, 如果加入试验成功, 用户订阅新的层; 如果离开试验失败, 用户取消当前订阅的最高层, 接收速率减半. 与这种粗粒度的拥塞控制机制相反, TCP 通过 AIMD(additive increase/multiplicative decrease)行为实现细粒度的拥塞控制. 研究文[2, 3]显示: 如果分层组播中加入试验和离开试验的频率设计得当, 可以使速率在长期内与 TCP 流量模型近似, 保证组播流量是 TCP 友好的.

在分层组播中, 为了提高拥塞控制的效率, 用户间需要协同工作, 尤其对组播转发树中拥塞瓶颈的下游用户. 一部分用户检测到拥塞, 离开最高层以减少接收速率, 如果其他用户没有同步的离开, 组播转发树不会对拥塞瓶颈的下游进行剪枝, 拥塞依然存在; 如果用户不采取加入操作同步, 可能导致某些组成员无法充分利用带宽. 为了改善接收端之间的同步问题, RLC[2] 和 FLID-DL[4] 要求接收端只能在同步点(synchronization points, 简称为 SP) 加入新层. PLM[5] 也使用了类似的时钟机制来保证接收端的加入/离开同步.

分层组播拥塞控制本质上是将拥塞控制的复杂性由传输层转移到组管理和组播路由协议交互, 频繁的加入和离开会对组播路由造成较大的负担, 从而影响组播路由的稳定性. 一旦发生拥塞, 同步机制会导致拥塞瓶颈后的用户同时离开组播树, 进一步加剧了这种影响.

以文[6]的链路相关模型为基础, 我们提出了一个新的拥塞统计模型, 用来捕获组播树中发生的拥塞. 模型以 RED[7] 为基础, 考虑了不同链路中拥塞标记的相关性, 更接近于真实网络环境. 在模型中, 我们使用两个基本假设: (1) 所有的组播树都是以源为树根的最短路径树; (2) 如果组播树中发生拥塞, 所有下游节点会立刻离开组播树. 这些假设允许我们简化模型的分析, 同时不会丢失观察目标.

收稿日期: 2002-07-15; 修回日期: 2003-08-15
基金项目: 国家自然科学基金资助项目(No. 90104002, No. 69725003, No. 60203025); 国家 863 高科技发展计划资助项目(No. 2001AA121013)

2 组播树的稳定性

组播路由的目标就是为组播组构建一棵连接所有组成员的无循环树. 有多种方案可以用来构造组播树. 最简单的方法是使用某种最短路径算法(Dijkstra s[8]). 大部分组播路由协议(DVMRP[9], MOSPF[10] 和 PIM2DM[11]) 在实现中使用了最短树策略. 在本文的讨论中, 我们只考虑最短路径树.

组播路由的稳定性是 IP 组播中一个重要的问题. IP 组播允许用户在任意时间加入或离开某个组, 要求组播扩展树支持动态更新. 如果组播树变化太快, 可能导致路由出现短暂的不稳定行为. 文[12]给出了如下定义:

定义 1 一棵最短路径树中分布了 m 个组播用户, 用 $(m)表示一个组成员离开后组播树链路数量的变化. 如果 E($(m)) [1, 组成员的离开被认为不会影响组播树的稳定性.

定义 1 假设在一段时间内, 只允许组播树中一位成员离开. 然而, 在分层组播环境中, 如果在组播树的上游发生拥塞, 下游用户可能同时离开当前的组播树, 定义 1 的假设不再成立. 为了评价和量化这种情况下的组播树稳定性, 我们给出如下的定义:

定义 2 在分层组播环境中, 如果一棵最短路径树(对应分层组播中的某一层) 中分布了 m 个组播用户, 组播树中总的链路数量为 T(m) , $(m)表示发生拥塞时, 组成员(一个或多个) 的离开对组播树链路数量带来的变化(增加或减少) , E($(m)) 表示 $(m)的均值, 我们定义组播树的稳定性因子 F(m)= E($(m))/ T(m).

显然组播树的稳定性因子越高, 它的稳定性越差.

3 统计模型

为了捕获组播树中发生的拥塞, 我们建立如下的组播会话模型: 组内包括唯一的发送端和 m 个接收端, 它们通过最短路径树连接, 发送端位于树的根节点, 接收端都位于树的叶子节点. 发送端和接收端之间使用接收端驱动分层组播拥塞控制机制(例如 RLM) , 如果组播树的某中间节点发现拥塞, 此节点的所有的下游用户同时离开组播树. 为了简化模型分析, 我们只考虑下面这种树型: 有限深度的非平衡二叉树(图 1) . 既然本文主要研究链路离开数量和拥塞的关系, 而拥塞与拥塞瓶颈的深度密切相关, 仅仅考虑二叉树是合理的.

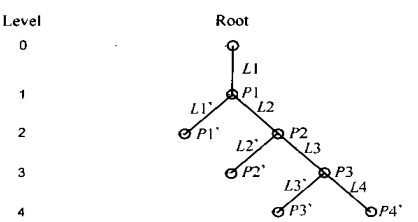

图 1 非平衡二叉树: 深度 m= 4

链路 L_i 处 $x_i= 1$ 的概率: $Pr\{X_i= 1\}= p_i$;

链路 L_i 处 $x_i= 0$ 的概率: $Pr\{X_i= 0\}= 1- p_i$;

链路 Lc_i 处 $xc_i= 1$ 的概率: $Pr\{Xc_i= 1\}= 0$;

链路 Lc_i 处 $xc_i= 0$ 的概率: $Pr\{Xc_i= 1\}= 1$.

311 一个新的链路相关模型

为了捕获组播树中的拥塞瓶颈, 我们引入下面的定义.

定义 3 一棵深度为 m(m<]) 的非平衡二叉树, 由中间节点集合 P、用户节点集合 Pc、中间链路集合 L、用户链路结合 Lc 组成. 它满足下面 3 个条件:

条件 1 所有的节点和链路都在图 1(m= 4) 中有标记, Pc 代表 m 个接收端分布的节点.

$$P= \{P_1, P_2, P_3, \ , P_{m-1}\};$$

$$P= \{Pc_1, Pc_2, Pc_3, \ , Pc_{m-1}, Pc_m\};$$

$$L= \{L_1, L_2, L_3, \ , L_{m-1}\};$$

$$L= \{Lc_1, Lc_2, Lc_3, \ , Lc_{m-1}, Lc_m\}.$$

条件 2 用随机变量 $X_i(Xc_i)$ 表示链路 $L_i(Lc_i)$ 的拥塞状态, $X_i(Xc_i)$ I {0, 1}(参考图 1) .

$$P_r\{X_i= x_i\}= \begin{cases} p_i, x_i= 1 \\ 1- p_i, x_i= 0 \end{cases}, P_r\{Xc_i= xc_i\}= \begin{cases} pc_i, xc_i= 1 \\ 1- pc_i, x_i= 0 \end{cases}$$

其中, $0< p_i, pc_i< 1$, 是链路 $L_i(Lc_i)$ 的拥塞标记. 为了简化模型说明, 我们假设与用户直接相连的链路 $Lc_i(Lc_i$ I $Lc)$ 上不会发生拥塞.

条件 3 在组播树中, 上游链路与下游链路的拥塞状态相关, 且满足下面的表达式:

$$P_r\{X_i = x_i| X_{i-1}= x_{i-1}, X_{i-2}= x_{i-2}, \ , X_1= x_1\}$$
$$= Pr\{X_i= x_i| X_{i-1}= x_{i-1}\}$$

312 组播树拥塞瓶颈的概率分布

根据上面给出的假设, 拥塞发生后, 为了降低拥塞, 拥塞节点下游的用户应该离开组播树. 在所有的瓶颈节点中, 距离源最近的拥塞节点对整个组播树的影响最大. 为了更清楚的描述这种特征, 我们引入下面的定义.

定义 4 对定义 3 定义的一棵组播树, 如果发生拥塞, 在中间节点集合 P 中, 距离源节点最近的检测到拥塞的中间节点 $P_k(P_k$ I $P)$ 被称为组播树的拥塞瓶颈.

根据定义 3 和定义 4, 我们可以推导出下面的命题.

命题 1 在定义 3 中, 随机变量序列 $\{X_1, X_2, \ , X_{m-1}, X_m\}$ 定义了一个 2 态的离散马尔可夫链.

证明 可由定义 3 直接推出.

由定义 3 和命题 1, 我们可以得到组播树拥塞瓶颈的概率分布.

定理 1 如果 m<] , 那么在定义 1 中定义的组播树中, 最多存在一个拥塞瓶颈, 且中间节点 P_k 成为拥塞瓶颈的概率用 $7_d(P_k, m)$ 表示, 那么

$$7_d(P_k, m)= \begin{cases} Pr\{X_1= 1\}, \ \text{if} \ \ k= 1 \\ Pr\{X_1= 0\} @ Pr\{X_k= 1| X_{k-1}= 0\} \\ \quad @ \prod_{i=1}^{k-2} Pr\{X_{i+1}= 0| X_i= 0\}, \\ \qquad \text{if} \ \ 2[k[m- 1; \end{cases} \quad (1)$$

证明 见文[13] 中定理 1 证明.

313 相关度模型

为了对公式 1 求解, 我们需要得到所有的一步转移概率 $Pr\{X_i= x_i| X_{i-1}= x_{i-1}\}$ 的表达式. 然而, 要准确的计算两个随机变量 X_i 与 X_{i-1} 之间的相关度是非常困难的, 我们引入文[6]中相关度因子 A(AI [0, 1]) 来量化一步转移概率中两个随机变量的相关程度. 我们引入下面的定义和定理:

定义 5 对两个相关的链路标记状态 X_i 与 X_{i+1}, 如果 $Pr\{X_{i+1}= x| X_i= x\}> Pr\{X_{i+1}= x| X_i= x\}$, 那么说 X_i 与 X_{i+1} 是正相关的; 如果 $Pr\{X_{i+1}= x| X_i= x\}< Pr\{X_{i+1}= x| X_i=$

x}, 那么说 X_i 与 X_{i+1} 是负相关的; 其中 x ∈ {0, 1}.

定理 2　考虑定义 1 中定义的马尔可夫链 $\{X_i\}$, 如果 $\{X_i\}$ 是正相关的, 且组播树中各中间链路的拥塞标记概率为矢量 $p=(p_1, p_2, p_3, \cdots, p_m)$. 那么下面的结论成立:

结论 1　$\forall A_i \in [0, 1]$, 随机变量 X_i 与 X_{i+1} 之间的相关程度可以通过相关度因子 A_i 计算, 且

$$\begin{cases} A_i = 0, \text{如果 } X_i \text{ 与 } X_{i+1} \text{之间是独立的} \\ A_i = 1, \text{如果 } X_i \text{ 与 } X_{i+1} \text{之间是完全相关的} \end{cases}$$

$\forall A_i \in [0, 1]$, A_i 是随机变量 X_i 与 X_{i+1} 之间的相关度因子, 那么条件概率 $\Pr\{X_{i+1} = x_{i+1} \mid X_i = x_i\}$ 的表达式如下, 其中 $x_i, x_{i+1} \in \{0, 1\}$:

$$\Pr\{X_{i+1}=0 \mid X_i=0\} = \begin{cases} 1-(1-A_i)p_{i+1}, & \text{if } p_i \geqslant p_{i+1}; \\ (1-A_i)(1-p_{i+1}) + A_i\left(\dfrac{1-p_{i+1}}{1-p_i}\right), & \text{if } p_i < p_{i+1}; \end{cases} \tag{2}$$

$$\Pr\{X_{i+1}=1 \mid X_i=0\} = \begin{cases} (1-A_i)p_{i+1}, & \text{if } p_i \geqslant p_{i+1}; \\ (1-A_i)p_{i+1} + A_i\left(\dfrac{p_{i+1}-p_i}{1-p_i}\right), & \text{if } p_i < p_{i+1}; \end{cases} \tag{3}$$

证明　见文[6] 中定义 3 和定理 3 证明.

将定理 2 中的公式 2 和公式 3 代入公式 1, 我们可以得到组播树中拥塞瓶颈的概率分布的表达式.

推论 1　对定义 1 中定义的深度为 m 的组播树, 用矢量 $A=(A_1, A_2, A_3, \cdots, A_{m-1})$ 表示定理 4 中的相关度因子, 用矢量 $p=(p_1, p_2, p_3, \cdots, p_m)$ 表示集合 L 中各中间链路的拥塞标记, 那么中间节点 P_k 成为拥塞瓶颈的概率用下式表示:

$$\gamma_d(P_k, A, p, m) = \begin{cases} p_1, & \text{if } k=1; \\ (1-p_1)(1-A_{k-1})p_k \prod_{i=2}^{k-2}[1-(1-A_i)p_{i+1}], & \text{if } 2 \leqslant k \leqslant m-1; \end{cases} \tag{4}$$

3.4　组播树的稳定性

定理 3　对定义 3 中定义的深度为 m 的组播树, 如果中间节点 P_k 是拥塞瓶颈, 那么用户的离开对组播树中链路数量造成的影响为 $\Phi(m, P_k)$, 表达式如下:

$$\Phi(m, P_k) = 2(m-k)+1 \tag{5}$$

证明　参考定义 3 和图 1.

利用推论 1 中的概率分布公式和定理 3 中的公式 5, 可以获得定义 3 中组播树的稳定性因子的表达式.

定理 4　对定义 3 中定义的一棵深度为 m($2 \leqslant m < \infty$) 的组播树, 有 m 个用户分布在组播树中, 如果对 $0 < p_i = p < 1$ 且 $0 < A_i = A < 1$, $\forall i$, 那么下面的结论成立:

结论 1　$\gamma_d(P_k, A, p, m)$ 表示中间节点 P_k 成为拥塞瓶颈的概率分布, 那么

$$\gamma_d(P_k, A, p, m) = \begin{cases} p, & \text{if } k=1; \\ (1-A)(1-p)p[1-(1-A)p]^{k-2}, & \text{if } 2 \leqslant k \leqslant m-1; \end{cases} \tag{6}$$

结论 2　对每个中间节点 P_k 和一个给定的 A, 当 p 满足下列表达式:

$$p^* = \begin{cases} 1, & \text{if } k=1; \\ \dfrac{(1-A)k + A + 1 - \sqrt{(1-A)^2(k-1)^2 + 4A}}{2k(1-A)}, & \text{if } 2 \leqslant k \leqslant m-1; \end{cases} \tag{7}$$

$\gamma_d(P_k, A, p, m)$ 取唯一最大值.

结论 3　对每个中间节点 P_k 和一个给定的 p, 当 A 满足下列表达式:

$$A^* = 1 - \frac{1}{p(k-1)}, \quad \text{if } 2 \leqslant k \leqslant m-1, \text{and } k \geqslant 1 + (1/p); \tag{8}$$

$\gamma_d(P_k, A, p, m)$ 取唯一最大值.

结论 4　如果发生拥塞时组播树的链路变化的均值用 $E(\Phi(A, p, m))$ 表示, 那么

$$E(\Phi(A, p, m)) = \begin{cases} p(2m-1), & \text{if } A=1; \\ (2m-1) - \dfrac{1-p}{p(1-A)}\{2+[1-(1-A)p]^{m-2} \\ \quad -3[1-(1-A)p]^{m-1}\}, & \text{if } 0 < A < 1; \end{cases} \tag{9}$$

结论 5　组播树的稳定性因子 $F(m)$ 有如下表达式:

$$F(m) = \begin{cases} p, & \text{if } A=1; \\ 1 - \dfrac{1-p}{p(1-A)(2m-1)}\{2+[1-(1-A)p]^{m-2} \\ \quad -3[1-(1-A)p]^{m-1}\}, & \text{if } 0 < A < 1; \end{cases} \tag{10}$$

证明　见文[13] 中定理 4 证明.

4　数学评价

4.1.1　组播树拥塞瓶颈的概率分布 $\gamma_d(P_k, A, p, m)$

图 2(a) 给出了当相关度因子 A 改变时, $\gamma_d(P_k, A, p, m)$ 相对于 k 的变化曲线. 我们发现, $\gamma_d(P_k, A, p, m)$ 是一个严格单调递减函数, 无论 A= 0(独立) 还是 A> 0(相关). 这种属性是我们期望得到的, 因为拥塞瓶颈更容易出现在距离源较近的中间节点中.

图 2(b) 显示函数 $\gamma_d(P_k, A, p, m)$ 与拥塞瓶颈的深度 k 成反比. 图 2(b) 也显示, 对任意 k, 存在一个唯一最大值 $\gamma_d^*(P_k, A, p^*, m)$, 验证了定理 4 的结论 2. 图 2(c) 显示对任意 A, k 越大, $\gamma_d(P_k, A, p, m)$ 越小. 另外, $\gamma_d(P_k, A, p, m)$ 不是对 A 的单调函数. 但是当 A 和 k 满足式(8) 时, 函数有唯一最大值 $\gamma_d^*(P_k, A^*, p, m)$. 当 k 变大时, A^* 随之增加. 验证了定理 4 中的结论 3.

4.1.2　分层组播拥塞控制对组播树的影响

图 3(a) 给出对不同的 A, 组播树稳定性因子 $F(m)$ 对 m 的变化曲线. 我们观察到, 链路标记相关程度(A) 直接影响到组播树的稳定性. A 越大, 组播数的稳定性因子越低, 拥塞发生时, 组播树受的影响越小. 另外, 我们还观察到, 随着组播树中用户数量 m 的增加, 组播树变得越来越不稳定. 最重要的是, 我们观察到在分层组播环境下, 即使在比较低的链路标记概率下(p = 0.1), 拥塞对组播树的影响也可能很大(超过

50% 的链路离开组播树).

利用式(10) 图 3(b)给出对不同的 p, 组播树稳定性因子 F(m) 对 m 的变化曲线. 显然, 在高拥塞网络中(p 比较大), 组播树的稳定性更差. 图 3(c) 给出 m 固定的情况下, A 和 p 对组播树稳定性因子 F(m) 的影响. 进一步验证了图 3(a) 和图 3(b) 的结论.

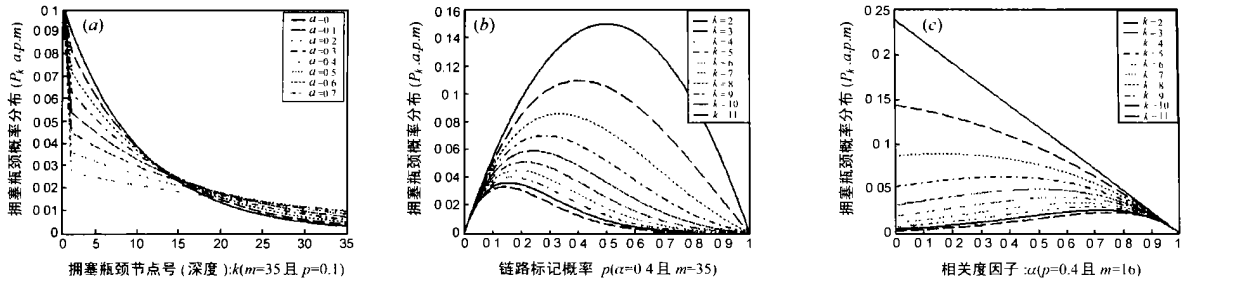

图 2

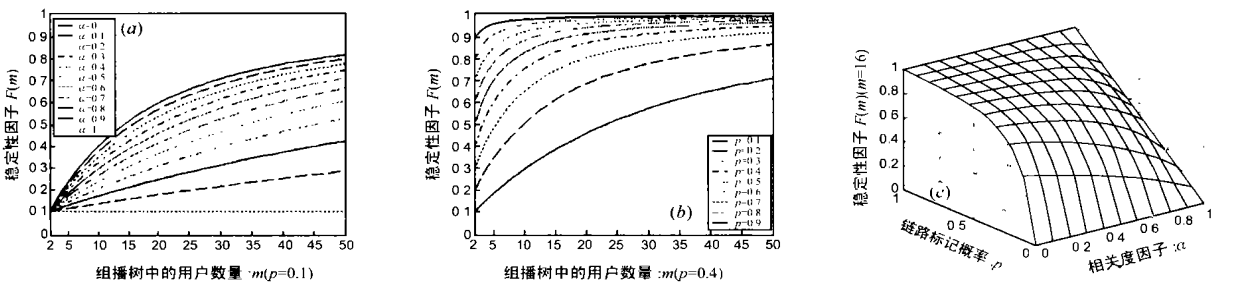

图 3

5 结论

本文考察分层组播环境中拥塞对组播树稳定性的影响. 通过一个新的统计模型, 我们能捕获在链路拥塞状况相关的情况下的组播树的拥塞瓶颈, 从而得到拥塞瓶颈的概率分布表达式. 我们又引入了相关度因子模型, 估计和评价了不同链路拥塞标记的相关程度. 利用上述模型, 我们得到了在分层组播中组播树稳定性因子的通用表达式. 数学评价进一步确认了相关度性模型对的必要性. 特别值得注意的是, 在链路标记概率较低的情况下(p = 0.1), 随着用户数量的增加(m \ 15), 超过 40% 的链路会离开组播树(链路的相关度因子 A[0.4).

参考文献:

[1] McCanne S, Jacobson V, Vetterli M. Receiver2driven layered multicast [A]. In Proceedings of ACM SIGCOMM(ACM SIGCOMM. 96) [C]. Stanford, CA, USA: ACM Press, 1996. 117- 130.

[2] Vicisano L, Rizzo L, Crowcroft J. TCP2like congestion control for lay2 ered multicast data transfer[A]. In Proceedings of IEEE INFOCOM [C]. San Francisco, USA: IEEE Communications Society, 1998. 996- 1003.

[3] Byers J, Frumin M, Horn G, et al. Improved congestion control for IP multicast using dynamic layers[R]. USA: Digital Fountain Technical Company, 2000.

[4] Byers J, Frumin M, Horn G, et al. FLID2DL: congestion control for lay2 ered multicast[A]. In Proceedings of Second Int. l Workshop on Net2 worked Group Communication(NGC 2000) [C]. Palo Alto, CA, USA: ACM Press, 2000. 71- 81.

[5] Legout A, Biersack E W. PLM: fast convergence for cumulative layered multicast transmission schemes[A]. In Proceedings of ACM Sigmetrics [C]. Santa Clara, CA, USA: ACM Press, 2000. 13- 22.

[6] Zhang Xi, Shin K G. Statistical analysis of feedback2synchronization signaling delay for multicast flow control[A]. In Proceedings of IEEE INFOCOM[C]. Anchorage, Alaska, USA: IEEE Communications Soci2 ety, 2001. 1133- 1142.

[7] S Floyd, V Jacobson. Random early detection gateways for congestion avoidance[J]. IEEE/ACM Trans on Networking, 1993, 1(4): 397- 413.

[8] Cormen T H, Leiserson C E, Rivest R L. Introduction to Algorithms [M]. McGraw2Hill, USA: The MIT Press, 1995. 95- 121.

[9] D Waitzman, S Deering, C Partridge. Distance vector multicast routing protocol[S]. RFC 1075, 1988.

[10] J Moy. Multicast extensions to OSPF(MOSPF)[S]. RFC 1584, 1994.

[11] S Deering, D Estrin, D Farinacci, et al. Protocol independent multicast version 2 dense mode specification[S]. Internet Draft, Internet Engi2 neering Task Force, 1999. Work in progress.

[12] Mieghem P V, et al. Stability of a multicast tree[A]. In Proceedings of IEEE INFOCOM[C]. Anchorage, Alaska, USA: IEEE Communications Society, 2002. 1133- 1142.

[13] 石锋, 等. 分层组播拥塞控制策略对组播树稳定性的影响[R]. 北京: 清华大学网络研究所, 2002.

作者简介:

石 锋 男, 1976 年生于湖北沙市, 博士研究生, 主要研究方向是计算机网络体系结构.

吴建平 男, 1953 年生于山东巨野, 博士, 教授, 博士生导师, 主要研究领域为计算机网络体系结构, 协议工程学, 互联网络.

徐 恪 男, 1974 年生于江苏洪泽, 博士, 讲师, 主要研究领域为计算机网络体系结构, 高性能路由器操作系统.