DLMCC:一种动态层次组播拥塞控制机制*)

任立勇 卢显良 吴 劲 周 旭
(电子科技大学计算机学院 成都610054)

摘 要 由于Internet本身固有的异构性,累积层次组播受到广泛关注,但现有的层次组播大多是粗粒度和静态的,不能适应动态变化的网络环境。为解决这一问题,提出了一种动态层次组播拥塞控制机制DLMCC。在该机制中:①给出了一种反馈聚集算法,动态确定层次组播的层次数和层次速率,从而有效地提高了网络的带宽利用率;②采用了一种逐级分组对探测带宽方法,可准确快速测量本地带宽,同时可保证粗粒度的TCP友好;③所有组播层次的数据在同一个组播组中传输,彻底消除了接收者加入新的层次和离开最高层次时需执行复杂和耗时的IGMP操作,以及多个组播组导致的新的异构问题。仿真实验结果表明DLMCC是层次动态变化、TCP友好和可扩展的。

关键词 动态层次组播,拥塞控制,TCP友好,聚集算法

DLMCC:A Congestion Control Mechanism for Dynamic Layered Multicast

REN Li-Yong LU Xian-Liang WU Jin ZHOU Xu
(Department of Computer Science, UEST of China, Chengdu 610054)

Abstract More and more attentions have been paid on layered multicast for Internet intrinsic heterogeneity, but most existing layered multicast protocols cannot adapt to dynamic network conditions because their layers are coarse granularity and static. In this paper, a new congestion control mechanism for dynamic layered multicast (DLMCC)is presented. To implement this mechanism: first, a novel feedback aggregating algorithm is put forward. It can dynamically determine the number of layers and the rate of each layer, and can efficiently improve bandwidth utilization of network. Second, a bandwidth measurement method based on hierarchical packet-pair is introduced, which can infer local bandwidth quickly and accurately. Third, because the data of all the layers is transferred in only one group, the intricate and time-consuming IGMP operations, caused by receiver joining a new layer or leaving the topmost subscribed layer, are thoroughly eliminated. And this mechanism also avoids other problems resulted from multiple groups. Simulation results show that DLMCC is adaptive, TCP friendly and scalable.

Keywords Dynamic layered multicast, Congestion control, TCP friendly, Aggregate algorithm

1 引言

在单点到多点的通信中,组播被认为是一种有效的数据传输方式[1]。然而组播至今无法得到广泛应用,其中一个重要的原因在于组播没有提供合适的拥塞控制机制。近年来,组播拥塞控制得到了广泛的关注,它已成为网络领域中的一个研究热点[2]。研究组播拥塞控制的难点在于既要保证原有的网络协议数据能正常传输,同时也要求具有高可伸缩性,并能有效利用网络带宽。基于层次编码视频源的层次组播机制因其具有良好的可伸缩性和异构性支持,被认为是一种最好的视频组播解决方案[3~7]。

S. McCanne 等人[4]第一次提出了接收者驱动的累积层次组播拥塞控制RLM,将视频数据按其自然属性分割成多个累积层次,并通过不同的组播组发送到接收者。由于不同的接收者可根据其接收链路带宽状况不同而预定不同的层次数,以获取其"最佳"收视效果。但由于RLM采用定时器触发其状态的转移,因此RLM收敛到优化速率的速度较慢。另外,RLM的公平性(主要指RLM协议间和TCP-friendly)较差,并且RLM的周期性预定操作可能导致大量的分组丢失。V. cisano 等人[5]在RLM的基础上提出了一种新的累积层次组播拥塞控制协议RLC。由于RLC采用了按指数分布层次来分割媒体数据(这种方式模拟了TCP的行为),因此在一定程度上达到了TCP友好。但RLC仍然没有解决收敛速度慢的缺陷,同时,与RLM一样,RLC周期性地加入试验仍可能导致大量数据丢失。另外RLM与RLC均采用固定层次数和层次速率,不能很好地适应网络带宽动态变化,带宽利用率不高。

John W. Byers 等人在RLC的基础上提出了带动态层次的公平层次增加和减少组播拥塞控制FLID-DL[6],用动态层次组播来模拟静态层次组播,有效地解决了由于IGMP的大离开延迟(leave latency)所带来的对拥塞的响应速度慢的问题。尽管动态层次的速率随时隙(slot)变化,但其模拟的静态层次速率仍然是不变的,因此FLID的带宽利用率也不高。同时,由于FLID-DL每隔固定时间T才改变接收者速率,因此自适应程度不高,协议间公平性也不能很好地满足。Wan Jun-wei 等人提出了一种主动自适应层次组播[7],根据网络动态变化动态调整组播视频层次及各层速率,极大地提高了网络带宽利用率。但遗憾的是,该文并没有详细讨论其核心的归并算法,同时其速率分配算法也没有考虑协议间公平性问题,因此实际应用可能较为困难。

目前研究的大多数层次组播协议均采用不同的层次占用不同的组播组,这种方法存在以下3方面的问题:1)加入/离开层次时,大多需要加入/离开组播组,因此时延大,对拥塞的响应速度慢;2)不同的组播组其路由不一样,这给带宽资源探测

*)本课题得到国防预研基金(51406070201DZ0211)资助,得到电子科技大学青年科技基金(YF020803)资助。任立勇 博士,主要研究方向为可靠组播,网络QoS,无线通信。卢显良 教授,博士生导师,研究方向为分布式操作系统和网络应用技术。吴 劲 博士生,研究方向为计算机网络与移动计算。周 旭 博士生,研究方向为计算机网络与操作系统。

带来极大的困难；3)接收方需预留大量的内存以缓存和重组从不同组播组到达的报文。

针对上述问题，文中提出了一种动态层次组播拥塞控制机制 DLMCC，DLMCC 利用逐级(Hop-by-Hop)分组对来推测本地可用带宽，并与直接下游可用带宽汇聚，实现对可用带宽的准确、快速测量。中间节点和发送方执行反馈聚集算法，并最终由发送方动态调整组播层次数和各层次速率，以适应网络状态的变化。同时，所有层次数据在同一个组播组中以不同的会话形式传输，彻底解决层次组播占用多个组播组所带来的问题。

2 动态层次组播

在累积层次组播中，发送方将数据源分割成 n 层$\{L_1, L_2, \cdots, L_n\}$，各层速率为 r_i，$1 \leqslant i \leqslant n$，各层间互不冗余，$L_1$ 包含数据源里最重要的信息，称为基本层，而其余各层包含增强上一层质量的信息，即 L_i 在 L_{i-1} $(2 \leqslant i \leqslant n)$ 基础上增强，称为增强层。因此，不同的接收者如果预定相同层次数的数据$\{L_1, L_2, \cdots, L_i\}_{i \leqslant n}$，则会有相同的收视效果，并且，预定的层次越多，接收质量就越高。IETF 的可靠组播传输 RMT 工作组提出的异步层次编码[8]算法(ACT，Asynchronous Layered Coding)提供了一种数据分发的可靠层次组播编码方案。

关于媒体数据分层编码算法已超出本文的讨论范围，但层次数目(或层次粒度，layer granularity)和层次速率对层次组播却非常重要。如图1所示，假设数据源完整传输需要4Mb/s，如果采用静态层次数为2，层次速率分别为1Mb/s 和3Mb/s，由于 R_1 和 R_2 的瓶颈带宽分别为3Mb/s 和2Mb/s，因此均只能接收基本层的数据。但如果发送方采用动态层次编码算法，根据当前网络状态产生四个层次(假设基本层次速率不变)，速率均为1Mb/s，则 R_1 能稳定接收基本层和两层增强层，接收速率为3Mb/s，R_2 能稳定接收基本层和一层增强层，接收速率为2Mb/s。显著地提高了网络带宽利用率和总体接收质量。

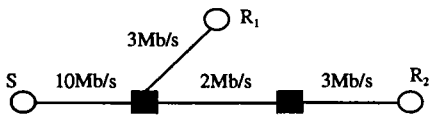

图1 层次组播示例

2.1 层次分割

正如上文所述，层次数目和层次速率对组播效率非常重要。本文讨论的动态层次组播包含两个方面的含义，一是动态变化的层次数目，二是动态变化的层次速率。为讨论方便，定义下列术语。

表1 动态层次组播所用变量

变量	含义
n	接收者数目
L	编码算法允许的最大编码层次数
r_b	基本层最低速率
Q	接收速率向量$\{q_1, q_2, \cdots, q_n\}$，升序排列
(q_i, c_i)	c_i 表示接收速率为 q_i 的接收者数目
Gr	层次编码粒度，即层间最小速率之差
w	组播层次数目(即接收者分组个数)
r_i	第 i 层的速率，$1 < i \leqslant w$
Q_i	预定到第 i 层的接收者速率向量，$1 \leqslant i \leqslant w$
$\|Q_i\|$	预定到第 i 层的接收者个数，$1 \leqslant i \leqslant w$

假设数据源将数据按照一定的参数分割成 w 层，第 i 层速率为 $r_i = q_i - q_{i-1}$，$2 \leqslant i \leqslant n$，基本层速率为 $r_1 = q_1$。

定义1 层次组播的总体接收质量为：

$$\sum_{i=1}^{w} |Q_i| \cdot \min(Q_i) \tag{1}$$

因此，为提高总体接收质量，动态层次组播的目标就转化为：给定接收速率向量 R 求合适的 w，以获得(1)式的最大值，即：

$$T(w) = \max\{\sum_{i=1}^{w} |Q_i| \cdot \min(Q_i)\} \tag{2}$$

很明显，在极端情况下，将 n 个接收者分成 n 个组，此时(1)式获得最大值。但由于组播用户成千上万，因此在实际应用中，这是不可能的。

定义2 如果将源数据分割成 w 层，则定义带宽利用率为：

$$\Gamma_w(n) = T(w)/T(n) \tag{3}$$

文[9]证明了当接收者数目 n 趋于无穷时，层次组播带宽利用率等于 $1-(1/w)$，即 $\lim_{n\to\infty}\Gamma_w(n) = 1-(1/w)$。因此，在实际应用中，无需将源数据分割成太多的层次就可以获得较好的带宽利用率。

为计算最佳的组播层次数 w 和每层速率，本文提出了如下两个聚集算法：

算法1 前提是基本层速率 r_b 不变，则算法可分为如下两部分：

1)每一组接收者用二元组表示(q_i, c_i)，c_i 为接收速率为 q_i 的接收者数目。将所有 $q_i < r_b$ 的项删除，剩下的二元组构成一个以 q_i 升序的二元组排列$\{(r_b, c_b), (q_i, c_i)\}$，$c_b = 0$。如果 $q_1 - r_b < Gr$，则删除(q_1, c_1)，并令 $c_b = c_1$。如果 $q_{i+1} - q_i < Gr$，则删除二元组(q_{i+1}, c_{i+1})，并令 $c_i = c_i + c_{i+1}$。最终形成以 q_i 为升序的二元组排列$\{(q_i, c_i)\}_{1 \leqslant i \leqslant K}$，其中 $K \leqslant n$。

2)如果 $K > L$，则在二元组排列中选择 $c_i * q_i$ 最小的一项，即$\{i \mid \min(c_i * q_i)_{2 \leqslant i \leqslant K}\}$，删除该项，并令 $c_{i-1} = c_{i-1} + c_i$。重复上述操作，直到 $K \leqslant L$。需要注意的是，选择最小的 $c_i * q_i$ 时，需排除(q_1, c_1)，因为该项是反映基本层次的速率。如果 $K \leqslant L$，则不作任何处理。

算法2 基本层速率可大于 r_b，则算法可分为如下两部分：

1)每一组接收者用二元组表示(q_i, c_i)，将所有 $q_i < r_b$ 的项删除，剩下的二元组构成一个以 q_i 为升序的二元组排列$\{(q_i, c_i)\}$。如果 $q_{i+1} - q_i < Gr$，则删除二元组(q_{i+1}, c_{i+1})，并令 $c_i = c_i + c_{i+1}$。最终形成以 q_i 为升序的二元组排列$\{(q_i, c_i)\}_{1 \leqslant i \leqslant K}$，其中 $K \leqslant n$。

2)如果 $K > L$，则在二元组排列中选择 $c_i * q_i$ 最小的一项，即$\{i \mid \min(q_i * c_i)_{2 \leqslant i \leqslant K}\}$，删除该项，并令 $c_{i-1} = c_{i-1} + c_i$。重复上述操作，直到 $K \leqslant L$。需要注意的是，选择最小的 $c_i * q_i$ 时，需排除(q_1, c_1)，因为该项是反映基本层次的速率。如果 $K \leqslant L$，则不作任何处理。

上述两个算法的主要区别在于算法1在最终形成的二元组排列$\{(q_i, c_i)\}_{1 \leqslant i \leqslant w}$ $(w \leqslant L)$ 中 $q_1 = r_b$，而算法2中 $q_1 \geqslant r_b$。经过上述算法处理，在满足限定的条件下(1)式取得最大值，此时 $\min(R_i) = q_i$，$|R_i| = c_i$。

由此，可计算累积层次组播各层速率分别为 $r_1 = q_1$，$r_i = q_i - q_{i-1}$(其中 $2 \leqslant i \leqslant w$)。

为避免反馈爆炸，DLMCC 采用分级反馈聚集，即在组播树中的中间结点也执行上述算法，对下游的带宽反馈执行上述聚集运算和向上一级反馈。

2.2 层次组播会话

在大多数层次组播拥塞控制机制中，一般采用不同的层次占用不同的组播组，这种层次组播机制将会带来诸多问题(如引言所述)。为解决上述问题，本文采用所有层次数据在一个组播组内传输，在报文中增加一个会话号字段表示该数据在组播中所在的层次。通过这种方法，每个接收者预定的所有层次的数据均沿同一路经到达，因此带宽探测就更为准确和有效，同时也避免了接收者需要预留大量的内存来缓存和重组来自不同层次的数据。更为重要的是，彻底消除了接收者加入新的层次和离开最高层次时需执行复杂和耗时的 IGMP 操作，从而对拥塞的响应更为迅速。另外，这种方法还有助于中间节点实现优先级丢弃策略。

本文提出的 DLMCC 采用了分级层次预定策略，中间节点需要维护其直接下游节点的带宽和层次预定信息，并根据这些信息对到达的报文进行过滤和转发。

3 逐级分组对探测网络带宽

对网络拥塞做出正确响应的前提是必须首先推测网络的状况(如可用带宽)。现有的方法主要分为两类：一类是根据网络的丢包情况来增加和降低发送速率(如 TCP)；另一类方法是收集接收方的反馈信息(如丢失率，往返时延等)来计算网络的可用带宽(实时媒体单播)。上述方法能比较好地应用在单播环境下，如应用在组播模式下，则有可能导致反馈爆炸问题。S. Keshav[10]在其博士论文中首次提出了发送方利用分组对推测网络带宽的理论。这种基于发送方的分组对在单播模式下能有效推测网络可用带宽，但却存在两个问题：第一，接收者收到分组对后原样返回给发送方，发送方对收到的分组对进行处理，因此在组播模式下基于发送方的分组对可能导致反馈爆炸问题；第二，现有的网络链路往往具有不对称性(如卫星链路、xDSL 等)，因此这种端到端双向测量方法不能准确反映从发送方到接收者间的可用链路带宽。为此，DLMCC 采用单向的基于接收者的分组对推测网络带宽的方法，即发送方逐级向下定期发送分组对，中间节点和接收者收到完整的分组对后，根据式(4)推测本地可用链路带宽：

$$T=S/\Delta t \tag{4}$$

式中 S 表示分组对的报文长度，Δt 表示分组对中两个分组接收时间间隔。需要注意的是，中间节点收到完整的分组对后，需对该分组对进行重新定时，然后为每个分枝复制一个分组对并发送出去。这种基于接收者的分组对方法既能过滤双向测量带来的估算噪声，同时也能在路由器发生拥塞丢包前就推测出网络的拥塞情况。需要注意的是，式(4)推测出的可用带宽可能会因为偶然原因(如 TCP 突发业务流等)而造成振荡现象，因此，我们采用指数加权滑动平均(EWMA)对带宽估计值进行了平滑：

$$T=\eta * T_{estimate}+(1-\eta) * T_{old} \tag{5}$$

当接收者推断出本地带宽后，产生一个反馈报文$(T,1)$并向其上一级节点发送。当中间节点收到所有其直接下游节点的反馈报文后，除为每个分枝保存其当前带宽外，还需要执行2.1节提出的聚集算法，但由于上级链路有可能成为下游分枝的瓶颈，因此，需要对聚集算法进行调整，在聚集前将所有反馈报文中速率小于本地带宽的项删除，即假设某中间节点的本地带宽为 T_l，则删除集合 C 中的所有二元组项。

$$C=\{(q_i,c_i)\mid q_i>T_l\}_{0<i\leqslant n} \tag{6}$$

为保证共享瓶颈链路的接收者同步加入或离开最高层次，并在粗粒度上实现 TCP-like，DLMCC 中，发送方将每隔一定的时间发送探测分组对，探测报文中包含源数据的层次数和每层的速率。分组对之间的发送间隔时间随数据发送速率变化而变化[5]，即每隔$1/R_i$发送一对探测分组对，其中 $R_i=\sum_{k=1}^{i}r_k, i$ 为发送的最高层次。

4 仿真实验

DLMCC 的设计目标是能针对网络的动态变化，动态调整组播源数据的分割层次数和层次速率，以提高网络的带宽利用率和总体接收质量。另外，组播拥塞控制的可扩展性和 TCP 友好(TCP-friendly)也是拥塞控制协议的主要评价目标。为此，我们在 ns2[11]中对 DLMCC 进行了仿真实验。

4.1 动态层次特性

为验证 DLMCC 对网络动态响应的动态层次特性，我们采用图2作为实验拓扑图，链路参数如图所示。实验中，我们假定发送源为无限速率数据源，同时层次划分粒度 $Gr\geqslant$1Mb/s，最大层次数为10层。在中间节点 I_1 和 I_2 之间0～3s 时设置恒定速率为6Mb/s 的干扰源，在3～7s 时设置方波干扰源，其每隔1s 在8Mb/s 和6Mb/s 之间跳变。分别对2.1节的两种聚集算法进行了实验，其中基本层速率 r_b=1Mb/s。

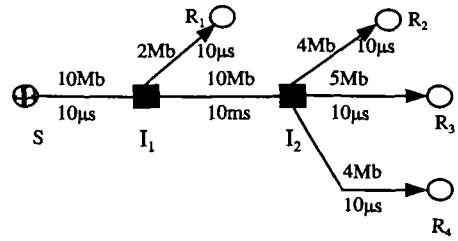

图2 动态层次实验拓扑图

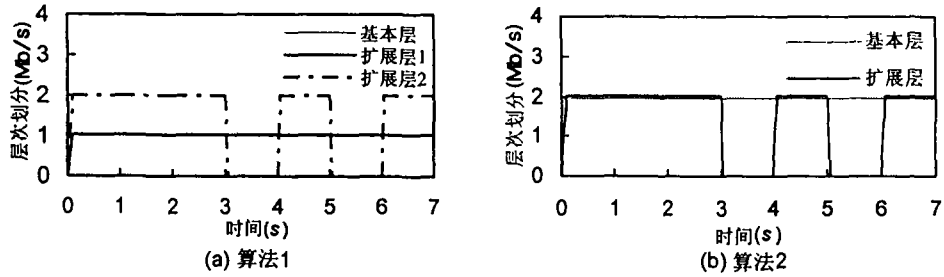

图3 DLMCC 的动态层次特性

图3给出了当网络状况动态变化时，DLMCC 对层次数和层次速率的调整，以响应网络拥塞。在实验(a)中，发送方和中间节点中采用了聚集算法1，即基本层速率 r_b 不变。R_1的接收速率恒定为2Mb/s，R_2、R_3、R_4在0～3s 时，由于受共享瓶颈链路 I_1-I_2限制，接收速率均为4Mb/s，因此数据源共产生两个扩展层次，速率分别为1Mb/s、2Mb/s。在3～7s 时，受方波干扰

源影响，共享瓶颈链路 I_1-I_2 带宽在2Mb/s 和4Mb/s 之间跳变，因此数据源产生基本层和扩展层1速率不变，而周期性地删除和产生扩展层2，其速率为2Mb/s。在实验(b)中，发送方和中间节点采用了聚集算法2，即基本层速率 r_b 可调整。R_1 的接收速率恒定为2Mb/s，R_2、R_3、R_4 在0～3s 时，由于受共享瓶颈链路 I_1-I_2 限制，接收速率均为4Mb/s，因此数据源共产生两个层次，基本层和扩展层速率均为2Mb/s。在3～7s 时，受方波干扰源影响，共享瓶颈链路 I_1-I_2 带宽在2Mb/s 和4Mb/s 之间跳变，因此数据源产生的基本层速率不变，而周期性地删除和产生扩展层，其速率为2Mb/s。从实验中可以看出，尽管两个算法产生的总体接收质量完全一样，但由于算法2可以动态调整基本层速率，因此当所有接收者中的最小速率大于基本层速率时，数据源端只需要划分更少的层次。

4.2 可伸缩性

组播拥塞控制的一个最重要的性能参数就是协议的可伸缩性，可以毫不夸张地说，可伸缩性的好坏将成为决定该组播拥塞控制协议最终是否得以应用关键。为此，我们对 DLMCC 的可伸缩性进行了评估，实验拓扑如图4所示。实验中，我们假定有一个 DLMCC 层次组播源，为无限速率数据源，N 个接收者，其接收速率分布为平均分布函数：$T_i = i * \frac{T}{N}$，$i \leqslant N$。其中 T 为中间链路带宽，可以看出本实验的瓶颈链路在组播树的叶子链路上。同时，假设最大组播层次为 w，基本层速率为 r_b，$N \gg w$，并且始终保证所有接收者均能接收基本层的数据，即 $r_b \leqslant T/N$。图5给出了最大组播层次 w 取不同值时的实验结果。可以看出实验结果的带宽利用率 $\Gamma_w(n)$ 与2.1节中的分析是相吻合的，例如当 $w=4$ 时，其带宽利用率位于60%～75%之间。随着层次数的显著增加，带宽利用率并没有显著增加，尤其是当组播层次超过10层后，带宽利用率几乎很接近。另外，随着接收者数目的增加，带宽利用率的下降极为平缓。从实验结果和上述分析可以看出本文提出的动态层次组播拥塞控制机制 DLMCC 具有良好的可伸缩性和较高的带宽利用率。

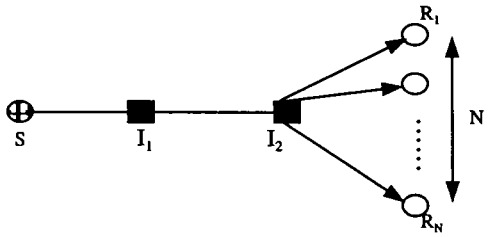

图4 可伸缩性实验拓扑图

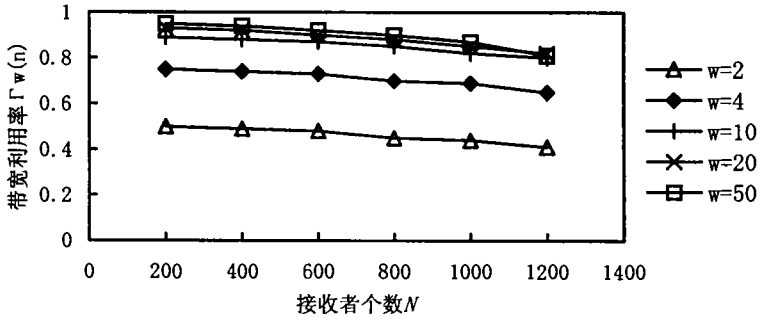

图5 DLMCC 的可伸缩性

结束语 累积层次组播在解决组播接收者异构问题方面具有独特的优势，因此广受关注。但现有的层次组播大多采用静态分层的方法，不能有效地利用网络带宽和有效地提高总体接收质量。同时，现有的层次组播大多将不同层次的数据在不同的组播组中传输，由此带来诸多问题。为此，本文提出了一种动态层次组播拥塞控制机制，通过定期发送分组对逐级探测网络带宽，在中间节点实施反馈聚集算法，并最终由发送方动态调整层次组播的层次数和层次速率，从而有效地提高了网络的带宽利用率。同时，所有组播层次的数据在同一个组播组中传输，彻底消除了接收者加入新的层次和离开最高层次时需执行复杂和耗时的 IGMP 操作，以及多个组播组带来的其他问题。文章中给出的仿真实验验证了 DLMCC 是 TCP 友好、可扩展和高带宽利用的。

在目前 DLMCC 工作的基础上，我们下一步的工作主要集中在如下几个方面：(1)进一步完善 DLMCC 的实现，包括在实际互联网上进行 DLMCC 实验；(2)研究无最大层次数限定条件下的动态层次计算算法；(3)研究层次组播与主动队列管理机制和优先级丢弃策略相结合的层次组播拥塞控制机制，这将有效地提高层次组播的公平性和响应拥塞的能力。

参 考 文 献

1 Deering S. Multicast routing in a datagram internetwork:[Ph. D. Thesis]. Stanford University, 1991

2 Shi Feng, Wu Jian-ping. A Survey on Multicast Congestion Control. Journal of Software, 2002,13(8):1441～1449

3 Byers J, Luby M, Mitzenmacher M. Fine-grained layered multicast. In:Proc. IEEE INFOCOM, Apr. 2001. 275～283

4 McCanne S, Jacobson V, Vetterli M. Receiver-Driven Layered Multicast. In SIGCOMM'96, Aug. 1996. 117～130

5 Vicisano L, Rizzo L, Crowcroft J. TCP-like Congestion Control for Layered Multicast Data Transfer. In: Proc. of IEEE INFOCOM, San Francisco, CA, USA, March 1998. 996～1003

6 Byers J W, Gavin H, Michael L, et al. FLID-DL: Congestion Control for Layered Multicast. IEEE Journal on Selected Areas in communications, 2002, 20(8): 1558～1570

7 Wan Jun-wei, Lu Xi-cheng. Active and Adaptive Multicast of Layered video. Journal of Computer Research & Development, 2000, 37(8): 994～999

8 Luby M, Gemmell J, Vicisano L, et al. Asynchronous Layered Coding Protocol Instantiation. Feb. 2002, IETF Internet Draft draft-ietf-rmt-pi-alc-06. txt.

9 Gau Rung-hung, Hass A J, Bhaskar K. On Multicast Flow Control for Heterogeneous Receivers. IEEE Journal on Selected Areas in communications, 2002, 10(1): 86～101

10 Keshav S. Congestion Control in Computer Networks. PhD thesis, EECS, University of Berkeley, CA 94720, USA, Sep. 1991

11 NS: Network Simulator. [Online]. Available at: http://www.isi.edu/nsnam/ns