

文章编号:1001-9081(2005)02-0259-03

一个单源的应用层组播协议的设计和实现

方 奕, 张 卫

(华东师范大学 计算机科学系, 上海 200062)

(frankiefang@citiz.net)

摘 要:提出了一个单源的应用层组播的协议 SSALM。SSALM 协议采用树优先的构造策略,在组成员之间构造基于源的组播分发树,并在源树的基础上建立叠加网。成员可以独立地加入组和退出组,并在组播树中进行切换。给出了该协议的主要算法和实现模型。在现有的 Internet 环境下使用该协议能有效地节省网络带宽。

关键词:应用层组播;组播协议;加入组;退出组

中图分类号: TP393.03 **文献标识码:** A

Design and implementation of a single source application layer multicast protocol

FANG Yi, ZHANG Wei

(Department of Computer Science, East China Normal University, Shanghai 200062, China)

Abstract: This paper proposed an application layer multicast protocol named SSALM (Single Source Application Layer Multicast). SSALM protocol adopted tree-first approach, which constructed source-specific multicast tree among group members and constructed overlay network based on the tree. Group members were able to join group and quit group independently. Besides, members were able to switch in the tree. This paper described the major algorithms and the modules of implementation of this protocol. Using this protocol can save network bandwidth effectively in the existing Internet environment.

Key words: application layer multicast; multicast protocol; join group; quit group

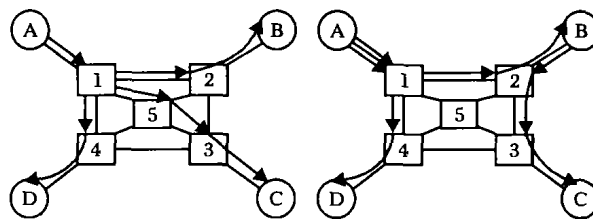
0 引言

近年来,Internet 上产生了许多新的应用,其中不少是高带宽的多媒体应用,譬如网络视频会议、AOD/VOD、股市行情发布、多媒体远程教育、CSCW 协同计算。这些应用都需要占用比较大的网络带宽,会引发带宽的急剧消耗和网络拥挤问题。为了缓解网络瓶颈,人们提出了组播(IP Multicast)技术方案。但是从因特网中 IP 组播的应用现状看,IP 组播并没有取得预期的成功。组播路由器需要为每个活动的组维护路由状态信息,网络中大量的活动组将需要路由器巨大的存储和处理开销;开放的 IP 组播模型在开放的 Internet 环境中难以支持有效的管理和控制机制。以上这些因素都制约了 IP 组播的发展。于是,一些研究者提出将复杂的组播功能放在端系统实现的新思想,也就是使用应用层组播(Application Layer Multicast)技术来替代传统的 IP 组播技术,从而达到节省带宽的目的。

图 1 显示了应用层组播的基本思想及其与 IP 组播的差别,IP 组播的数据沿着物理链路复制和转发,而应用层组播的数据则在主机实现复制和转发,数据报沿着逻辑链路转发,多条逻辑链路可能经过同一条物理链路。

应用层组播和传统的 IP 组播技术相比,可以避开网络层实现组播功能的许多难题:1)应用层组播的状态在主机系统

中维护,不需要路由器保持组的状态,解决了业务的扩展性问题;2)组播应用可以随时部署,不需要网络设备的升级和功能扩展;3)可以简化组播的控制、可靠等功能的实现,建立在网络连接之上的应用层组播可以使用传输层的 TCP、UDP 协议,如可以利用 TCP 的可靠和拥塞控制简化组播的可靠和拥塞控制。当然应用层组播对带宽的节省不如 IP 组播;协议额外的开销比较大。但是因为其开发方便,部署简单,应用层组播还是具有比较好的应用前景。



(a) IP 层组播 (b) 应用层组播
图 1 IP 层组播和应用层组播的比较

基于上述原因,本文提出了一个单源的应用层组播协议(Single Source Application Layer Multicast Protocol, SSALM)的系统设计,并使用 WinSock 在 Windows 平台上实现了该协议。

1 SSALM 的特点和设计目标

SSALM 继承了应用层组播的思想,它是在主机(端系统)

收稿日期:2004-07-23;修订日期:2004-09-28

作者简介:方奕(1980-),男,上海人,硕士研究生,主要研究方向:计算机网络; 张卫(1954-),男,上海人,教授,主要研究方向:计算机网络。

上实现的协议,不需要中间的路由器的参与,从而对协议的部署和实施带来了很大的方便。同时也使得 SSALM 的使用限制条件大大减少,为在该协议之上构建应用带来了很大的方便。

SSALM 是针对单个源的应用层组播协议,也就是说在一个特定的组播组中同时只能有一个发送者,这样的规定在一定程度上降低了协议的复杂度,便于协议的实现。同时,由于一个组中只有一个发送者,该协议的过程得到了大大的简化,因此有较好的性能。例如,假设一个提供在线电台广播服务的服务器所能提供的带宽是 2M 字节/秒,它需要为每个连接提供至少 8K 字节/秒的带宽,那么如果采用单播方案的话,则最多同时可以为 25 个用户提供服务,而如果采用应用层组播方案的话,已经收到广播内容的结点可以向其他结点转发内容数据,从而使得更多的用户能够获取服务。SSALM 协议特别适用于那些需要高带宽的单个数据源的应用(例如:文件分发系统、视频/音频广播系统、证券行情发布等)。

2 SSALM 的主要算法

SSALM 在组成员之间建立以服务器为根的组播树,数据从组播树的根结点发出,沿着树的分支不断转发,到达叶子结点,这种树被称为组播分发树。除了组播分发树的父子关系以外,SSALM 结点还维护和其他部分结点的关系,以此来提高组播分发树的健壮性。SSALM 中的结点就是加入组的主机的一个网络接口,它必须支持 TCP/IP 协议。而链路则是对两个结点之间的物理通路的逻辑描述。下面根据成员结点参与组播的过程,对 SSALM 中的主要算法进行说明。

2.1 链路好坏的判断依据

SSALM 里定义了两个重要的度量:时延(Delay)和带宽(Bandwidth)。时延是指从源发出的应用层数据到达某一结点所花费的时间。对于实时的应用,各个结点的时延的相异性不能太大,否则将导致一定的信息的不一致。而带宽指的是单位时间内可以从父结点接收到的数据量,以 K 字节/秒为单位。对应用层组播来说,它在结点之间构建组播树来传输数据。如果某一个结点有较多的子结点,则较多的复制报文会出现在物理线路上,使该链路上的可用带宽下降。如果带宽不能达到应用的要求的话,则不能将新的结点作为其子结点。此外,对于带宽优先的应用(例如,大文件传输),带宽这个度量可以作为衡量链路优劣的主要标准。

在 SSALM 中以时延和带宽的值作为链路好坏的判断依据,加入组的时候父结点的选取的主要依据就是到候选父结点的链路好坏的比较。我们定义如下公式:

$$Goodness = ((sDelay - Delay)/sDelay) \times t1 + ((Bandwidth - sBandwidth)/sBandwidth) \times t2 \quad (0 < t1 < 1, 0 < t2 < 1, t1 + t2 = 1)$$

其中 Goodness 表示链路的好坏情况,Delay 表示度量时延的值,Bandwidth 表示度量带宽的值,sDelay 表示标准的时延值,sBandwidth 表示标准的带宽值,sDelay 和 sBandwidth 的值是事先由上层的应用根据实际情况指定的,t1 和 t2 是权值,由具体的应用定义。Goodness 的值越大,表明链路的状况越好。在 SSALM 中有专门的测量 Delay 和 Bandwidth 的值的方法。

2.2 加入组

加入组的成员必须以某种方式知道要加入组的根的 IP 地址,通常可以通过网页的形式来发布。当结点 M 要加入组时,执行如下的父结点选择算法:

1) M 向根结点发送加入请求。

2) 根结点收到加入请求后,向其所有子结点通告 M 的信息。所有子结点收到后,继续将该通告信息转发给它们各自的子结点,此过程一直继续直到所有的叶子结点收到通告信息为止。

3) 所有收到 M 通告信息的结点,如果它能够(当前子结点数小于最大结点数)接收子结点,那么它向 M 发出加入邀请。

4) M 会陆续收到这些加入邀请,它将发出这些邀请的结点按先来后到的次序排序。由于树的转发特性,M 收到第一个加入邀请的发送者,是具有最小时延的候选父结点。

5) M 选取具有较好时延的结点(步骤 d 中排序在前的结点),测量到该结点的带宽值,并计算链路 Goodness 值。

6) 选取一个具有最佳 Goodness 值的结点,作为父结点,加入组播树。

在该算法中,父结点的选取范围是整棵组播树,由于采用被动寻找的方式,而非主动寻找的方式,使得选取父结点的时间比较短,结点能够比较快地加入到组播组。

2.3 数据传输

当结点加入组播组后,就能够收到从根结点发送的数据了。在一个组播组中,数据是沿着分发树的根,自顶向下到达叶子结点。每个结点从其父结点处接收数据,并把它转发给其所有的子结点。若收到的数据报不是发自父结点的,则丢弃该数据报。根据上层应用的不同可以选择采用 TCP 或者 UDP 来传输数据。若上层为文件分发系统,则可以采用可靠的 TCP 协议,若为音频/视频广播系统,则可采用开销较小的 UDP 协议。

2.4 离开组

为了维持组播树的可用性,成员离开组时,必须不能影响组播树的正常运作。如果成员没有子结点,那么就简单地退出组,并报告父结点。如果该成员是包含了若干子结点的中间结点,那么成员离开组后,必须使得其子结点还是能够照常接收数据。若某一中间结点离开组,最理想的情况是将其所有的子结点连接到它们的祖父结点上,这样在最小的局部范围内能够恢复组播树,但是如果这时祖父结点接收子结点的能力有限,也就是说并不是所有的结点都能够作为它的子结点的话,则需要进一步处理。为此包含子结点的组成员结点 M 离开组的时候,执行以下的离开组算法:

1) M 获取其父结点的信息。根据父结点的最大载荷情况和子结点的个数,计算出可容纳子结点的个数 n。

2) M 获取其所有子结点的它们各自的子结点数,选出具有最多子结点个数的 n 个子结点。通告这 n 个节点连接到 M 的父结点上(对于这些子结点来说是它们的祖父结点)。然后,断开与它们的数据通路。

3) M 通告除以上 n 个结点以外的其他所有子结点重新加入组。

4) 这些结点像新加入组的成员那样,自由地加入到分发树上。一旦加入到分发树上,结点就通告其原来的父结点 M,并断开到 M 的连接。

5) 一旦所有的这些子结点都断开了到 M 的连接后,M 才退出组。

该算法保证子结点在断开到原来父结点连接之前,已经建立了到父结点的连接,从而保证了不间断的数据传输。但是,这样也会带来问题。这些更改父结点的成员结点可能会收到重复的数据报文,一份拷贝来自原来的父结点,另一份拷贝来自新的父结点。所以,SSALM 的结点应该能在应用上具

有处理重复报文的能力。

2.5 网络断开的检测和恢复

当组播分发树建立以后,为了维护结点之间的连通性,需要定期地在子结点和父结点之间交换信息。在 SSALM 中,子结点定期向父结点发送保活报文以报告它的存在,同时父结点在收到子结点的保活报文后,返回给该子结点相应的应答。

若在一段时间内(例如:10 秒),父结点没有收到其子结点发送的保活报文,则认为该子结点已经从组播树上断开。父结点停止向该子结点转发数据,并撤除到该子结点的链路。

若组成员连续发送多个(例如:5 个)保活报文后,一直没有收到父结点的应答报文,则认为父结点不可达。此时该成员结点需要执行断开恢复算法,重新连接到组播树上:

1) 组成员结点试图连接到组父结点上,若连接成功,则操作成功结束;否则,执行 2)。

2) 组成员判断执行父结点选择算法,重新加入组。

3 SSALM 的实现

3.1 报文定义

SSALM 协议的报文分为两大类:一类是控制报文,它用来维护控制所有的成员结点的动作;另外一类是数据报文,它用来传输应用上的数据。由于控制报文会影响协议的行为和改变协议的状态,它对可靠性有一定的要求,所以 SSALM 中采用 TCP 协议来承载控制报文。而对于数据报文,可以根据应用的不同而采用 TCP 或者是 UDP。无论是控制报文,还是数据报文,它们都有一个固定的报文头,该头部的格式定义如图 2 所示。

0	4	8	16	31
版本	标志	类型	保留	
序号				
发送端的成员编号				
组号				
报文长度		报文数据		
报文数据...				

图2 SSALM 协议头部格式

其中,版本字段指明了当前 SSALM 协议的版本号。标志字段目前没有使用。类型字段指明了 SSALM 报文的类型。序号字段只在类型字段是数据报文时才有意义。对于控制报文来说,该序号字段始终填 0,当结点处理重复报文时可以根据报文的序号来进行判别。在 SSALM 协议中,每个成员结点都有一个全局唯一的编号,在发送报文时,结点将自己的编号填在发送端的成员编号字段。每一个组也有一个全局唯一的编号,由根结点在创建组的时候指定,组号字段指明了该报文是发给哪个组的。报文长度字段指出了包括头部长度在内的整个报文的长度。报文数据字段是实际的协议报文的内容,如果是控制报文的话,其数据就是控制报文,如果该报文是数据报文,则携带的是应用数据。

SSALM 协议中还定义了 20 种不同类型的控制报文,用来完成不同的功能。例如,加入请求报文是当结点想要加入某个组时,向该组的根结点所发送的控制报文。每种控制报文还有其自己的头部格式定义,它们都作为报文数据来传输。

3.2 协议实现模型

根据上面提出的协议设计,在 Windows 平台上利用 WinSock 编程接口和多线程的机制实现了该协议。图 3 是该协议的控制结构,用来处理控制报文。当结点启动时,首先开启一个侦听线程,该线程的主要工作是负责接收连接请求,一

旦接受了一个外来的请求后,就启动一个子线程来处理该连接。子线程负责接收连接上传过来的报文并对报文进行正确性检查,如果检查通过,则将该报文置于处理队列中,同时将该报文的套接字一并进队。因为同一时刻可能有来自多个结点的请求报文,保存套接字可以区分来自不同结点的连接。在处理队列中的报文按照先进先出的规则进行处理,出队的报文被送到报文处理线程中进行判断处理。报文处理线程是实现中最为复杂的一个部分,它要根据当前的协议状态对各种各样的控制报文进行处理,从而调整协议行为,改变协议状态。当一个报文处理完毕后,大多数情况下都会产生一个应答报文来响应发送方结点,该应答报文在同一个连接上返还给发送结点。

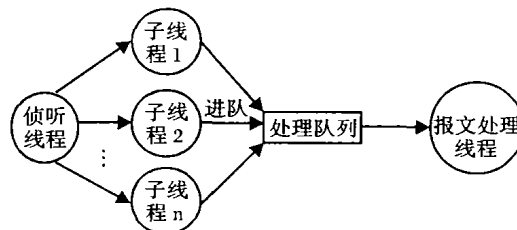


图3 协议控制结构

上述的控制结构用来处理协议的控制报文。在 SSALM 的实现中,控制报文和数据报文使用不同的端口。这样做的好处是在有大量数据报文传输的时候,不会影响到控制报文的接收。当成员结点加入组后,它就会建立到其父结点的连接,该连接专门负责传输数据报文。当成员结点退出组后,该连接会被释放。

4 结语

SSALM 协议已经在局域网内部进行了小规模(10 个结点)的实际测试,测试结果表明在局域网内 SSALM 有较好的表现。结点加入组和退出组的平均时间不到 1s,结点间的数据转发的延迟也在 100ms 之内。接下来的测试将从局域网转向广域网,在 Internet 上进行实际的测试,来验证协议的正确性和观测协议的表现。此外,协议的测试还将在组的规模上进行扩展,将使更多的结点加入组,并模拟一些网络故障,来测试协议的性能和自主恢复能力。当协议的各项测试完成后,还将在该 SSALM 协议之上开发一个多媒体的应用(例如:音频广播系统),来验证 SSALM 应用层组播思想的正确性和可行性。

参考文献:

- [1] CHU Y-H, RAO SG, SESHAN S, *et al.* A Case for End System Multicast[A]. Proc of ACM SIGMETRICS'00[C], 2000.1-12.
- [2] PENDARAKIS D, SHI D, VERMA D, *et al.* Waldvogel. ALMI: An Application Level Multicast Infrastructure[A]. In Proceedings of 3rd Usenix Symposium on Internet Technologies & Systems[C], March 2001.
- [3] JANNOTTI J, GIFFORD D, JOHNSON K, *et al.* Overcast: Reliable Multicasting with an Overlay Network[Z]. The 4th Symposium on Operating Systems Design and Implementation, October 2000.
- [4] BANERJEE S, BHATTACHARJEE B. Analysis of the NICE Application Layer Multicast Protocol[R]. Technical report, UMIACSTR 2002-60 and CS-TR 4380, Department of Computer Science University of Maryland, College Park, June 2002.
- [5] Yoid: [http://www.aciri.org/yoid/\[EB/OL\]](http://www.aciri.org/yoid/[EB/OL]), 2004-07.