一个基于主动网络的大规模可靠组播协议的设计和实现

陈晓林,李 冀,魏明亮,陆桑璐,陈贵海,谢 立
(南京大学软件新技术国家重点实验室,江苏南京 210093)

摘 要: 本文提出一个基于主动网络的大规模可靠组播协议 LARMP(Large - scale Active Reliable Multicast Protocol),它较全面地解决了 NACK/ ACK (Negotiate Acknowledge / Acknowledge Implosion)风暴、选择重发、分布恢复负担、拥塞控制、健壮性这五个 Internet 上的可靠组播面临的关键问题. LARMP 采用一个由四级主动路由器支持的主动组播树结构,利用组播树中的主动路由器实现动态主动组播树的建立和维护、NACK 抑制、NACK/ ACK 聚合、缓存数据报文并为组播树中报文丢失节点恢复报文、及时检测网络拥塞并反馈给发送者以便其调整发送速度等功能. 实验测试表明 LARMP 具有良好性能.

关键词: 组播; 主动路由器; NACK; ACK; 主动报文
中图分类号: TP393 **文献标识码**: A **文章编号**: 0372-2112 (2001) 08-1038-04

Design and Implementation of a Large-Scale Active Reliable Multicast Protocol

CHEN Xiao-lin, LI Ji, WEI Ming-liang, LU Sang-lu, CHEN Gui-hai, XIE Li
(State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, Jiangsu 210093, China)

Abstract: A large-scale active reliable multicast protocol: LARMP is presented. LARMP resolves five crucial problems of reliable multicast in the Internet such as negotiate acknowledge / acknowledge implosion, scope retransmission etc. LARMP employs an active dynamic multicast tree supported by the four-level active routers. The active routers in the tree implement the creation and maintenance of the active dynamic multicast tree, NACK suppression, NACK/ ACK aggregation, datagram buffering, limitation of the delivery of repair packets to the nodes experiencing loss, and timely detection and report of net congestion to allow the sender to adjust its sending speed. The initial experiments have shown good performance.

Key words: multicast; active router; NACK; ACK; active datagram

1 引言

随着 Internet 的迅速发展,越来越多的 Internet 分布式应用需要基于 Internet 的点到多点可靠通信即组播(Multicast)的支持. 然而,Internet 上的大规模可靠组播协议的设计是一个非常困难的问题,必须解决以下五个问题: ①NACK/ ACK 风暴; ②选择重发(Retransmission Scoping): 重传报文应只发给丢失了报文的接收者及相关链路; ③分布恢复负担(Distribute loss burden): 接收者丢失报文后应能尽快恢复(收到重传报文),恢复(发送重传报文)负担不能全由发送者承担; ④拥塞控制; ⑤健壮性: 接收者动态加入或退出组,部分接收者、通信链路、路由器/ 交换机的失效不能影响与之无关的接收者的接收.

然而,目前典型的大规模可靠组播协议大都不能完全解决上述五个问题(见本文第二部分). 大规模组播中发送者、路由器、接收者组成一棵组播树,传统的端到端差错控制和流量控制方法由于无法利用组播树而很难解决上述五个问题. 主动网络[1,2]的出现使组播协议可很好地利用组播树,从而给上述五个问题的解决提供了一条新途径. 设计并实现了一个基于主动网络的大规模可靠组播协议 LARMP,该协议可较好地解决上述五个问题. LARMP 利用组播树中的主动路由器实现差错控制、检测网络拥塞并向发送者报告,发送者通过调整发送速度来控制网络拥塞. LARMP 协议包括五个部分: (1) 体系结构: 它是一个由四级主动路由器支持的层次结构; (2) 动态主动组播树的建立和维护: 提供了无需 IP - Multicast 支持的主动组播路由和接收者动态加入或退出组; (3) 差错恢复: 解决 NACK 风暴,选择重发,分布恢复负担; (4) 拥塞控制; (5) 可靠性和健壮性保证.

本文组织如下: 第二部分简单介绍几种典型的大规模可靠组播协议及比较; 第三部分详细介绍 LARMP 的五个部分; 第四部分介绍 LARMP 的实现和初步的性能评估; 最后是结论.

收稿日期: 2000-07-07; 修回日期: 2000-11-26
基金项目: 国家 863 高技术项目(863-306-ZT02-03-01)

2 相关工作

目前的大规模可靠组播协议大致可分为两类:无需主动网络支持和需要主动网络支持.第一类协议典型的有 SRM[3] (Scalable Reliable Multicast)、LBRM[4] (Log-Based Receiver-Reliable Multicast)、RMTP[5] (Reliable Muticast Transport Protocol)等,第二类协议典型的有 ARM[6] (Active Reliable Multicast)等.上述这些协议没有一个能完全解决前述的五个问题. (1)在差错恢复和健壮性方面:SRM 无需路由器,预先指定的 logging server 或接收者支持,但在选择重发,本地恢复方面存在不足. LBRM,RMTP 需预先指定的 logging server 或接收者支持,在选择重发、本地恢复方面优于 SRM,但其健壮性依赖于预先指定的 logging server 或接收者,健壮性较差,也往往不能屏蔽掉瓶颈链路处一些不必要的 NACK 和重传报文. ARM 利用主动路由器实现 NACK 聚合 +选择重发 +本地恢复,在差错恢复方面优于 SRM、LDRM、RMTP.而所设计的 LARMP 在差错恢复方面优于 ARM,具体比较见第四部分. (2)在拥塞控制方面:以上几个协议只有 RMTP 涉及.

3 LARMP 协议

3.1 体系结构

主动路由器由于要执行用户在主动报文中定义的计算而降低了其路由转发报文的速度,所以它一般放置在战略位置(strategic location)如主干网络与边缘网络之间的交界处. Mbone 的实验表明:包丢失大都发生在连接 Mbone 的边缘网络,而不是发生在主干网络内.考虑到 Mbone 中的包丢失特点和主动路由器的特点,认为一个由四级主动路由器支持的主动组播树结构既是有效的,又是现实可行的.其结构如图 1.图 1 中组成员与主动路由器的连接和主动路由器之间的连接均为逻辑连接,组成员均为主动节点,象 ANTS(Active Network Transport System)[7,8]这样的主动网络执行环境提供了主动节点与主动路由器之间通过 UDP 来建立逻辑连接.主动路由器收到主动节点发送的主动报文后执行主动报文中定义的计算,计算完后将其向目的地转发,主动路由器同时提供软状态缓存(soft-state cache,以下简称 cache)给主动报文中定义的计算来存储主动数据/重发数据报文,计算所需的一些状态信息等等.

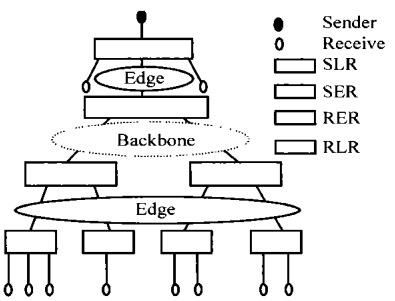

图 1　LARMP 组播树结构. RLR:接收者本地路由器;RER 接收者边缘路由器;SLR:发送者本地路由器;SER:发送者边缘路由器

为了后续说明方便,假设组播树中发送者与任意一个接收者的路径长度为 5;另外,用端口来描述组成员与主动路由器、主动路由器之间的逻辑连接,若某主动路由器有 N 个逻辑连接,则称其有 N 个端口,一个端口对应一个逻辑连接;DATA(K)表示序号为 K 的主动数据报文;NACK(K)表示请求重发 DATA(K)的主动 NACK 报文. ACK(K)表示确认收到所有 DATA(j) (j < = K)的主动 ACK 报文.

3.2 动态主动组播树的建立和维护

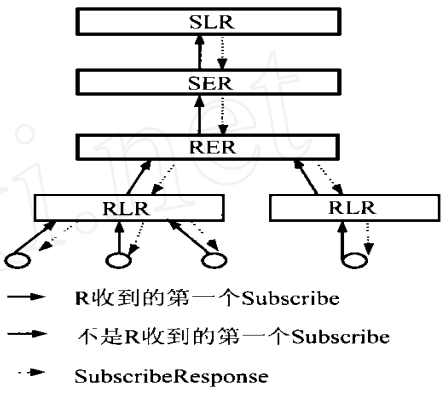

图 2　接收者动态加入组

接收者通过发送 Subscribe/ Unsubscribe 主动报文来申请加入/退出组,主动路由器收到它后建立或更新动态组播树. (1)接收者动态加入组的示意如图 2 所示:主动路由器 R 从端口 P 收到 Subscribe 主动报文(以下简称 Subscribe),①若 P 不在 R 的组下行端口列表中,则把 P 加入到 R 的组下行端口列表中,并为 P 初始化一个状态对象;②若 R 为 RLR/ RER/ SER 且该 Subscribe 是 R 收到的第一个 Subscribe,则将该 Subscribe 向目的地转发;③若 R 为 SLR,则向 P 发 SubscribeResponse;④若 R 已加入动态组播树,则向 P 发 SubscribeResponse;⑤若该 Subscribe 是 R 收到的第一个重发次数为 K 的 Subscribe(接收者等待 SubscribeResponse 超时,将重发 Subscribe),则将该 Subscribe 向目的地转发.主动路由器或接收者收到 SubscribeResponse,表示主动路由器或接收者已加入动态组播树.若主动路由器 R 收到 SubscribeResponse,则在 cache 中记录 R 已加入动态组播树,并向所有组下行端口发 SubscribeResponse. (2)接收者动态退出组算法类似于加入组算法,由于篇幅有限,具体算法略.可以看出,主动组播树建立和维护算法有良好的可扩展性.

3.3 差错恢复

3.3.1　报文丢失检测和 NACK 抑制.报文丢失检测由接收者和主动路由器负责. (1)组播树中某主动路由器 R 检测到 DATA(J)丢失后,①在 cache 中登记一个 NACK 记录,记录其所有子节点丢失 DATA(J),并通知所有子节点:R 丢失 DATA(J);②如果丢失报文在上一级节点没有丢失,则向上一级节点发 NACK 请求恢复. (2)接受者检测到 DATA(J)丢失后,如果 DATA(J)在上一级节点没有丢失,则向上一级节点发 NACK 请求恢复.

3.3.2　NACK 聚合.主动路由器 R 处理 NACK 聚合的方法类似于对 Subscribe 的处理,当 R 从 M 个不同端口(Pi, i = 1, …, M)收到 NACK(J)且不能本地恢复,则将需上传的 M 个 NACK(J)聚合为一个上传,并在 cache 中保存一个 NACK 记录,在

NACK 记录中记录下 Pi, $i = 1, \cdots, M$.

3.3.3 选择重发. 主动路由器收到 DATA(J),如果 cache 中有空间存储 DATA(J),则存之. 若 DATA(J)是恢复报文,查找 cache 中的 NACK 记录,找出哪些下行端口曾发过 NACK 请求 DATA(J)并向这些端口转发 DATA(J),然后清除 cache 中的相应 NACK 记录.

3.3.4 分布恢复负担. 主动路由器在端口 Pi 收到 NACK(J),若 DATA(J)在 cache 中,则将它从 cache 中取出并设置为恢复数据报文,然后向 Pi 端口转发. 考虑到除网络拥塞外接收者接收缓冲满也会导致接收者丢失报文,当接收者 Ri 发的 NACK(J)到 RLR,若 DATA(J)不在 RLR 的 cache 中但接收者 Rj 已收到 DATA(J),则将 NACK(J)发给 Rj,由 Rj 给 Ri 恢复.

LARMP 的差错恢复机制类似于 ARM,但做了两点改进: (1) NACK 抑制. ARM 的差错检测只由接收者负责,若 SLR 丢失 DATA(J),所有接收者都会检测到 DATA(J)丢失并发 NACK(J)给 RLR,不但浪费网络带宽,而且 RLR 将处理大量的不必要的 NACK. 另外,增加了接收者的恢复时延. (2) 快速接收者给慢速接收者恢复. ARM 没有考虑接收者接收缓冲满而丢失报文的情况,当主动路由器 cache 满而不能存储 DATA,即使丢失 DATA 的接收者所在子网中只有它一个丢失该报文,丢失报文也只能由发送者恢复.

3.4 拥塞控制

3.4.1 拥塞检测. 主动路由器和接收者检测到数据报文丢失后,用 NACK 向发送者报告拥塞,主动路由器将丢弃不必要的报告拥塞的 NACK,这样不但能及时检测出拥塞,而且所花代价较小. (1) SLR/ SER/ RER:主动路由器检测出 DATA(J)丢失且其上一级节点没丢失 DATA(J),立即向上一级节点发 NACK(J),上一级节点若不能恢复就做 NACK 聚合,否则恢复后判断是否满足条件: J 大于发送者最近降速发送的第一个 DATA 序号并且 cache 中的申请降速标志为 FALSE,满足则将此 NACK(J)(设置为无需恢复)上传给发送者报告拥塞并置申请降速标志 = TRUE;不满足则丢弃此 NACK(J). 例如,某主动路由器收到 DATA 0,2,3,5,7,8,9,10,11,13,向上一级主动路由器发 NACK(1,4,6,12),NACK(4,6,12)能在上一级主动路由器恢复,NACK(1)要在发送者处才能恢复且 NACK(1)到达发送者时发送者准备发 DATA(10). 上一级主动路由器将 NACK(4)上传报告拥塞而丢弃 NACK(6). 发送者恢复 NACK(1)后,在 DATA(J) ($J >= 10$)中设置发送者最近降速发送的第一个 DATA 序号 = 10,主动路由器收到 DATA(J) ($J >= 10$)后,在 cache 中保存:发送者最近降速发送的第一个报文序号 = 10 和申请降速标志 = FALSE, 所以上一级主动路由器将 NACK(12)上传给发送者让其再次降速. (2) RLR:若 RLR 能恢复接收者发的 NACK,不将该 NACK 上传. 其余处理同(1).

3.4.2 发送者速度调整算法. 与 TCP 的调整算法类似. W 为拥塞窗口大小,初始化为 1(慢启动), threshold 为可变快增速门限,速度调整算法简述如下:

(1) 收到 ACK(J) If ($W <$ threshold) $W = W + J$; Else $W = W + J/W$;

(2) 收到 NACK(J) 且 $J >$ 最近降速发送的第一个 DATA 序号

拥塞避免: If ($W >= 2$) {threshold = $W/2$; $W = W/2$;}

3.4.3 ACK 聚合. 主动路由器在端口 Pi 收到 ACK(J),令 Pi.ackSeq = J 并将 Pi.ackSeq 保存在 cache 中. 接收者 Ri 发 ACK(Ri.ackSeq),RLR 计算 MinAck = ($P1$.ackSeq, $\cdots$, Pi.ackSeq, $\cdots$)并保存在 cache 中,然后向上一级路由器转发 ACK(MinAck),RER/ SER/ SLR 从的处理同 RLR,最后 SLR 将 ACK(MIN($R1$.ackSeq, $\cdots$, Ri.ackSeq, $\cdots$, Rn.ackSeq)交给发送者.

3.4.4 发送者指定需要确认的 DATA. 为减少 ACK 的数量,发送者每调整一次发送窗口,计算一次哪两个 DATA 需接收者确认.

LARMP 的拥塞检测机制一方面能以较小代价及时检测出拥塞,另一方面由于主动路由器丢弃了不必要的拥塞报告报文而有较好的可扩展性;类似 TCP 的发送速度调整算法能较好地控制拥塞;ACK 聚合和发送者指定需要确认的 DATA 有效控制了 ACK 风暴.

3.5 可靠性与健壮性保证

若 ACK 丢失,接收者通过 NACK 将接收者确认的 DATA 的最大序号上传给丢失 ACK 的主动路由器;若部分接收者、通信链路、路由器/ 交换机失效,将把与故障有关的接收者和主动路由器清除出组. 由于篇幅有限,具体算法略.

4 系统实现和性能评估

在 ANTS 上实现了 LARMP,用 6 台 RS6000 工作站,1 台 SUN Ultra30 工作站,1 台 Windows PC 来模拟 LARMP 的体系结构. RS6000 配置为:200MHz CPU,128M 内存,155M ATM 网卡; SUN Ultra30 配置为:233MHz CPU,128M 内存,100M 以太网卡; PC 配置为:333MHz CPU,64M 内存,10M 以太网卡. 4 台 RS6000 分别模拟一个 SLR/ SER/ RER/ RLR,1 台 RS6000 模拟一个发送者,1 台 RS600,1 台 SUN,Windows PC 每台机器模拟 10 个接收者. SLR 的丢包率为 0.1,每个 DATA 主动报文携带 1kB 数据,每次测试发送者发送 1k 个 DATA.

从三个方面测试比较了 LARMP 的性能:

(1) NACK 抑制效果:LARMP 中接收者发送的 NACK 数目占接收者丢失报文总数的 20 %以内,而 ARM 中,由于接收者每检测到一个报文丢失就发一个 NACK 请求,接收者发送的 NACK 数目大于等于接收者丢失报文总数.

(2) 恢复时延:定义恢复时延为接收者 Ri 检测到报文 J 丢失到 Ri 收到恢复报文 J 这段时间;平均恢复时延为 Ri 所有丢失报文的恢复时延总和除于丢失报文的总数;最大恢复

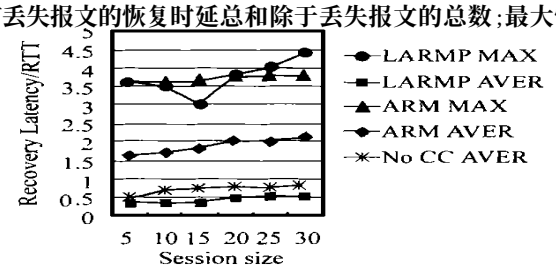

图 3 恢复时延比较

时延为在所有接收者的恢复时延中,恢复时延的最大值.如图3所示,LARMP的最大恢复时延与ARM相同,但LARMP的平均恢复时延小于0.6RTT,而ARM的平均恢复时延小于2RTT.

(3)拥塞控制效果:如图4所示,无拥塞控制控制情况下(off),接收者的平均恢复时延比有拥塞控制情况下大0.2RTT,而且随着接收者数目增加到一定程度,接收者发送的NACK请求数目(最大值)急剧上升;而在有拥塞控制情况下(on),随着接收者数目增加,接收者发送的NACK请求数目变化很小,表明拥塞控制算法有较好的可扩展性.

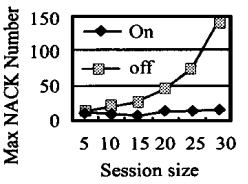

图4　NACK数量比较

5　结论

LARMP较全面地解决了基于Internet的大规模可靠组播协议必须解决的五个主要问题,性能测试的结果也表明了LARMP的良好性能.在今后的工作中,我们还将对LARMP做进一步改进,包括ACK抑制,针对主动路由器的有效拥塞控制算法.

参考文献:

[1] Smith J M,et al. Activating networks:a progress report [J]. Computer, April 1999,32(4):32 - 41.

[2] D L Tenenhouse,D Wetherall. Towards an active network architecture [A]. Proc. Multimedia Comp and Networking 96 [C], San Jose, CA, Jan 1996.

[3] S Floyd,et al. A reliable multicast framework for light-weight sessions and application level framing [J]. IEEE/ACM Transactions on Networking,December 1997,5(6):784 - 803.

[4] H W Holbrook,S K Singhal,D R Cheriton. Log-based receiver-reliable multicast for distributed interactive simulation [A]. Proceedings of ACM SIGCOMM [C],August 1995:328 - 341.

[5] J Lin,et al. RMTP:A reliable muticast transport protocol [A]. Proceedings of IEEE Infocom [C],1996.

[6] L lehman,et al. Active reliable multicast [A]. Proc. IEEE INFOCOM'98 [C], San Francisco, CA, Mar 1998.

[7] D Wetherall,J Guttag,D L Tenenhouse. ANTS:A toolkit for building and dynamically deploying network protocols [A]. Proc. IEEE OPENARCH'98 [C], San Francisco, Apr 1998.

[8] D Wetherall. Developing network protocols with the ANTS toolkit [A]. http://www.sds.lcs.mit/activeware.

作者简介:

陈晓林　男,1973年生于云南省楚雄市,1994年7月毕业于华中理工大学自控系,获工学学士,现为南京大学计算机系在职硕士生.研究领域为分布式处理和并行计算.

李　冀　男,1975年生于江西省湖口县,南京大学计算机系硕士研究生,研究领域为分布式处理和并行计算.