

可靠组播研究综述^{*}

The Review on Reliable Multicast

吴文峻 聂树宣

(北京航空航天大学计算机系 国家软件开发环境重点实验室 北京100083)

Abstract The paper presents a review of recent research work on reliable multicast (RM) protocols based on IP multicast. The requirements of diverse RM application and the main design issues of RM protocols are discussed. We describe the behavior of ideal RM algorithm, and introduce two RM protocols: SRM and RMTP. We then give a brief of how to improve the performance of the RM protocol through the methods of local error recovery, RM tree construction and an extended IP multicast forward service. Some open issues in the research and promising directions are also discussed.

Keywords Reliable Multicast(RM), IP Multicast, Local error recovery

一、引言

IP 网络上各种新应用的开展,如:视频会议、远程教学、网络交互式游戏、远程虚拟现实等等,要求互连网络提供可靠的选播服务(Reliable Multicast, RM)。IP 层的 Multicast 通信只提供尽力型服务,不保证选播数据报文的可靠传输,研究高效和可靠的 RM 机制,完成类似单播服务中 TCP 的可靠传输功能,成为最近五年来,IP 选播协议研究中的重要方向。

应当指出的是对 RM 协议机制的研究要比过去对 TCP 可靠传输机制困难得多,其根本的原因是 RM 的应用程序对可靠性的要求差异很大,如果以一个完全满足所有 RM 应用要求的服务模型来设计 RM 协议,最终只会导致协议性能根本无法满足应用的要求,特别是不能适用于大规模组播通信的情况。所以在 RM 研究中,涌现了大量的 RM 协议^[3,4,7,8,10,12],每一种协议往往都是针对一个或一类型的 RM 应用而设计的。

RM 研究的核心问题就是如何根据组播通信过程当中,网络丢失分组报文的情况,迅速、高效地恢复丢失的报文,使接收者接收到正确、有序的报文。RM 设计是在报文恢复的性能和报文恢复的代价之间的折衷。RM 算法主要包括 RM 错误恢复算法和拥塞控制算法两部分。限于本文的篇幅,本文只讨论错误恢复算法。

二、RM 应用需求的讨论

文[6]讨论了 RM 应用的不同需要,归纳为以下

的问题:

1) 应用程序需要了解是否每个接收者都接收到了正确的数据。在很多应用中,应用数据单元(ADU)是指对应用而言,逻辑上可分隔的数据单元。一个 ADU 可由一个或多个数据传输报文组成。可靠数据传输需要对 ADU 的确认,可能存在两个层次的确认:一是对 ADU 的确认,一是对报文的确认;对确认的严格程度随应用的不同而不同。如果应用程序需要知道每个接收者都接收到了报文,则发送者必须从每个接收者那里收到确认。如果应用还需准确了解有哪些接收者没有接收到 ADU,必须对确认的汇集(acknowledgement aggregation)做一些额外的限制。

在类似于 FTP 的文件组播应用中,如果传输尺寸很大的文件,出于数据缓冲区管理的需要,文件服务器需要确切了解接收者是否收到了正确的报文,以便能从缓冲区删除以发送完成的数据,释放缓冲区资源。而在某些交互式应用当中,例如:用于文本和图形信息交互的电子白板(Whiteboard),在整个交互过程中,本文和图形信息一直保留,不会删除,所以数据拥有者可以不必了解是否所有的参与交互的人员都接收到了正确的信息。

2) 应用程序需要扩展到大规模的接收者吗?RM 应用接收者规模的大小决定了 RM 协议所应具有的可扩展性,也限制了它所可能采用的协议机制。传统的单一层次的 ACK 机制可扩展性很差,会导致发送方的确认风暴(Acknowledge implosion)。解决 Ack implosion 的办法主要有四个:一是减少发送确认的频

^{*} 本文得到中国高速互连研究实验网项目 NSFCNet 和 973 海量信息系统项目的资助。吴文峻 博士生,主要研究方向为可靠组播、拥塞控制算法。

率,但这对于实现协议的拥塞控制是很不利的;二是使用否认应答的方式(Nacks)和各种否认应答抑制机制来实现;三是使用 ACK 聚集方式;第四种方法采用前向纠错 FEC 的“开环”方法,由接收者根据接收到的报文自己完成检错和纠错工作,由于取消了差错控制的反馈机制,这种算法的可扩展性比前四种都要好。

3) 应用程序是否需要限制每个接收者的不同差异?有的应用需要限制接收者接收行为的差异,以保证接收数据的特性保持在一定范围之内。如,股票市场信息的发布,如果让一个接收者比其他人很晚接收到消息是不能接受的。解决这类问题通常采用只发送有限数据,直到接收到肯定确认的方法。

4) 应用程序需要保证数据的顺序性吗?数据的顺序性是指应用数据单元 ADU 向上层应用提交的顺序。多数 RM 应用都要求协议提供保证源顺序性的服务,即确保从同一个数据源发出的数据在接收方提交时,其提交顺序和发送方是相同的。在一些 RM 应用中,如:交互式虚拟现实或海量数据库系统中,还要求不同数据源的数据到达接收者的顺序也要一致,满足全局顺序语义或因果顺序语义。

5) 应用程序需要提供保证一定传送时延的服务吗?

6) 应用程序需要很多彼此交互的发送者?

7) 应用数据流是间歇的,还是连续的?

8) 应用程序是否要在公用 Internet 网络上工作?

9) 应用程序是否需要在没有反向消息传输路径下工作?

10) 应用程序是否需要提供安全传送?

在这些问题当中,1)和3)分别回答了应用对传输的可靠性和传输的顺序性的要求,问题5)回答了应用传输的 QoS 问题,问题2)说明了应用的可扩展性。问题6)、7)定义了应用的通信特点。问题8)、9)指明应用所处的网络环境,10)是安全方面的要求。

RM 协议要综合考虑应用的服务要求、应用的可扩展性能和应用所处的网络环境,并利用应用的通信特点,提供高性能、低开销的通信服务。目前 RM 的协议研究主要针对问题1)和问题2)进行,对顺序性的研究则相对较少。

三、RM 错误恢复算法

一个 RM 协议对错误的恢复过程可用图1表示。

当丢失发生后,节点 D、E、F 都将发现丢失了正确的数据,要向拥有正确数据复本的节点 A、B、C 和数据源发送者 S 报告错误,并请求重新发送数据;S、A、B、C 获知发生数据丢失后,四个节点中至少会有一个节

点重传丢失的报文,发送给 D、E、F 节点,这是典型的后向纠错过程。

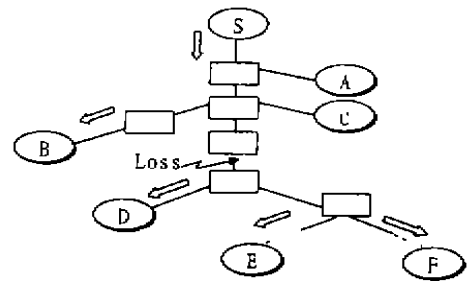


图1 理想 RM 恢复过程示意图

从上面的描述可以看到,设计 RM 正确性维护机制的目标是:减少错误恢复的时延,减少实现错误恢复带来的对节点的处理开销和网络能信的开销。错误恢复的时延由如下几部分时间组成:

$$\text{Loss Recovery Delay} = \text{Loss Detect Delay} + \text{Loss Report Delay} + \text{Retransmission Delay}$$

丢失检测时间同丢失发生点到丢失所影响的接收者的传输路径长度有关,也和错误检测的机制有关。例如:在图1中,节点 D 保留有当前最大接收的报文序列号为 N,它必须知道节点 S 已发出的最大序列号 M,通过比较 $N < M$,才能检测出丢失的发生。节点 D 可通过接收到 S 的新报文 M,或是接收到 S 发出的 Session 消息通报其最新的报文序列号。

丢失报告的时间和重传报文的传输时间则由 D、E、F 和 S、A、B、C 之间的路径长度决定。很显然,假设图中各链路的时延都相同,应当选择离 D、E、F 最近的 C 作为重传的发送者,由它直接向 D、E、F 发送丢失的报文。

在整个丢失恢复过程中,除了发送丢失的数据报文之外,还要发送一些控制报文,如:ACK 或 NACK 报文。由于 RM 协议的分布性,重传的数据报文和控制报文很可能会出现冗余的情况。

由上面的分析可以看出,评价 RM 性能的要害有两个方面:一是数据吞吐率,即发送有用数据同冗余数据和控制报文的比例;二是恢复的时延。最理想的 RM 算法应当使吞吐率最大,而恢复的时延最小。

如果不考虑使用 ACK 报文的开销(实际上在 RM 中,对每个报文进行应答是不必要,也是不可能的。有的协议只是定期报告正确接收的情况,以便数据源释放缓冲区。),而只使用 NACK 通知发送者报文的丢失,那么一个理想的 RM 算法行为可描述如下:

当一次丢失发生时,设丢失的报文为 p,丢失发生在链路 L 上,对于参加这次组播的所有成员,可以得到两个集合 Loss-Helper-Set 和 Loss-Requester-Set。

NACK 和 NACK 抑制机制。

从 RMTP 的工作原理可以看出,发生丢失时, Loss-Requester-Set 的成员既有可能通过 ACK 报文向自己的父节点报告,也有可能发出 NACK 显式,及时地报告丢失,无论哪种情况,丢失报文的报告将沿着 ACK 树到达 Loss-helper-Set 的成员,从而该成员将顺着 ACK 树逐级地重发丢失的报文。很显然,同 SRM 相比, RMTP 的协议成员的丢失报告和丢失重发报文的发送都被限制在父-子节点这一局部范围内,不会产生 SRM 中的 exposure 问题。不过重复的 NACK 或 Repair 仍然会出现,虽然较 SRM 为少。但是由于丢失报告是沿着 ACK 树上溯,而 Repair 顺着 ACK 树下行,丢失报告未必能经过组播树的最短路径,到达最近的 Loss-helper-Set 集成员。而 Repair 的报文的路径也未必是最短的。两类报文都有可能需要 RMTP 节点协议部件的转发,这无疑也增加了这些报文的传播时间,因此 RMTP 的丢失恢复时间会较长。

基于树的 RM 协议的关键在于构造和维护 ACK 树。从性能的角度讲,ACK 逻辑树能够和实际的 IP 组播路由树匹配得比较好,其协议性能,尤其是丢失恢复的时延将得到很好的改善。

五、解决 RM 算法问题的思路

针对上述这两类协议的问题,人们又进行了大量的研究,提出了许多新的设想,来逼近理想 RM 算法的行为。这些改进基本上可分为如下两个大的方向:

1) 构造树结构,形成局部组

从前面的分析可以看出,当整个组播树的规模很大,参与组播的成员很多时,把错误的报告和错误恢复——即报文的重传,限制在局部的范围内,有利于降低 RM 协议给网络和节点带来的通信开销,提高网络的吞吐率。如果报文丢失发生的链路不在本地局部组的范围内,应当沿着组播树的方向,选择发生错误概率最大的方向,传播 NACK,尽快地到达 Loss-Helper-Set 集合的成员。而 Repair 的传播也应当尽可能地沿着组播树,同时不扩散到无关的接收者。组播树或组播层次的构成就成为实现上述 Local Recovery 的关键因素。

RM 树构造和维护机制 RM 树算法机制由构造算法和维护算法两部分组成。RM 树构造算法,解决 RM 树如何生成的过程,即如何把分布的组播参加者,形成逐级的父-子节点的关系。而 RM 维护算法是解决网络故障、节点故障以及节点离开等等造成 RM 树断开时,如何重新安排部分节点,修复 RM 树的问题。

RM 树的建立过程可采用自顶向下或自底向上两种方法。在自顶向下的方法中,数据发送节点首先开始接收组播成员成为其子节点,这些子节点又尝试建立

自己的局部恢复组;而在自底向上的过程中,各节点可以先独立地构造自己的局部恢复组,而后再加入到 RM 树当中,一般来说,自顶向下的方式可以保证加入到 RM 树的节点,一定能获得对丢失数据的恢复。而自底向上的方式形成 RM 树的速度比较快,但是不能提供对丢失数据的恢复保证。

形成父-子关系的过程,实际上是一个双向选择的过程,希望形成自己局部恢复组的节点可以广播自己的状况,吸引未加入 RM 树的节点,加入其局部恢复组。节点也可以发出加入选择组播组请求,已加入 RM 树并接收到该请求的节点,可返回回答,告知该节点自己的情况。考虑到发送这两种请求的代价,可使用 Expanding Ring Search (ERS) 算法^[2],不断扩大发送该请求的范围。

维护机制的思想是通过定期发送联系报文,保持父-子节点关系;当父-子节点关系破坏后,尤其是父节点不存在,子节点要重新进行父节点的查找工作,重新修复 RM 树。

子节点从多个父节点的应答或广播中选择正确的父节点,决定了 RM 树的性质。

RM 树生成策略:使 RM 树同 IP 组播路由树匹配子节点如何选择父节点是 RM 树生成策略的核心问题。有许多协议^[1,2,11],提出了多种选择的标准。主要应当从两个方面来考虑,一是距离,二是报文丢失的情况。优化的 RM 树应当根据实际的 IP 组播路由树的情况,把距离相距比较近,拥有较多公共链路的成员划分到一个局部组中,应选择最有可能得到正确数据的节点作为局部组播组的父节点,这样才能逼近理想 RM 算法行为。

由于 Internet 的协议结构,IP 组播路由树对 RM 协议而言是透明的。如果不改变这种体系结构,不扩充 IP 组播的功能, RM 只能采取一些推测的办法,根据端节点能够获得的拓扑信息或报文丢失的统计信息,推测组播树结构,实现 RM 树生成策略。文[2]给出了利用 ICMP 报文,采用类似 Mtrace 的办法,获取部分 IP 组播路径信息,并构造 RM 树的算法。文[11]则讨论了如何根据端节点看到的丢失模式,推断 IP 组播树结构,并构造 RM 树的方法。

2) 增强 IP Multicast 的服务模型,为 RM 协议提供更多支持

如果只是基于文[8]提出的 IP Multicast 服务,让端节点负责完成可靠组播,前面所论述的一些问题是难以避免的, RM 的协议不大可能实现理想的 RM 算法行为。有的研究就提出,应当考虑扩充现有的 IP Multicast 服务模型,使得 IP 路由器直接参与 RM 协议的工作中。

总的来说,IP Multicast 可以在以下两个方面为 RM 协议提供支持:1) 为端节点提供关于组播树的拓扑结构信息,帮助端节点划分局部组,建立 RM 树;2) 直接参与 RM 通信过程,帮助实现 NACK 和 Repair 的传播和抑制

在理想 RM 算法行为描述中,我们可以看到 RM 的过程可分成两个主要步骤:首先是 Loss-Requester-Set 的成员,传播 NACK 请求到 Loss-Helper-Set 的成员;第二步是由 Loss-Helper-Set 中的成员把请求的报文传播到 Loss-Requester-Set 的成员。在第一步中,路由器可以负责把 NACK 沿着 IP 组播树向数据源方向传播,对于重复收到的 NACK,则禁止它继续向数据源传播。对每个路由器,都动态或静态地指定可能 Loss-Helper-Set 成员,如果这些成员没有正确的数据,则继续向上传播 NACK,直到遇到 Loss-Helper-Set 的成员为止。第二步,由接收到 NACK 通知的 Loss-Helper-Set 成员,先向上传播正确的数据报文,到达每个设定号的“Turning Point”路由器后,由该路由器再向下组播该数据报文,这个路由器组播的范围是 Loss-Requester-Set 成员所在的组播路由于树,这个过程可称为 Subcast,整个过程可参考图3。

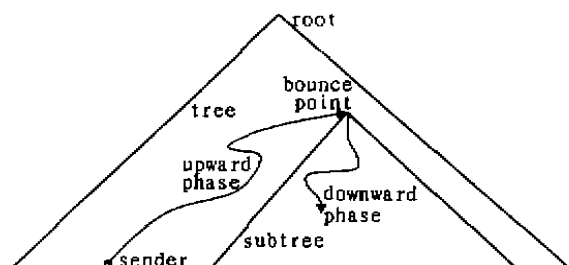


图3 Subcast 示意图

已经有一些 RM 协议方案中提出了具体的方案,如文[4]的 LMS 和文[12]的 PGM,CISCO 公司的一些产品甚至开始支持 PGM 协议。LMS 的设计思想同 PGM 是类似的,只是 LMS 要求路由器处理的负担较轻,路由器不要为每次丢失事件维护相关的控制;相比较而言,PGM 对路由器要求较多的处理开销,路由器要为每次丢失事件的维护控制状态,所以 PGM 可以更好地限制报文的重传范围。

从上述分析可以看到这两个方面已经有了一定的研究结果,但是对几个关键性的问题,仍然没有定论,如:什么样的 RM 树是能使丢失恢复性能最好的树?如何以最小的代价构造和维护这样的树?能否有比现有 RM 路由器方案更简单的方法,使主机和路由器承担的处理比较合理?

结论 RM 协议的设计必须充分考虑到它所针对

的应用需求,不同的需求决定了 RM 协议的控制机制,也决定了 RM 协议所能达到的性能。在解决不同应用问题的过程中,人们提出了许多 RM 协议,并逐渐认识到 RM 协议性能的关键问题——错误恢复性能和恢复开销之间的内在矛盾,以及 RM 性能与组播路由拓扑结构的密切关系和基于 IP 组播模型实现 RM 之间的矛盾。虽然从层次化的局部恢复和新的 IP 组播服务两个方向,又提出了许多新的思路,但这些研究之间的定量比较和分析,以及各种机制可能的综合方案,还有待进一步的深入研究。

感谢李未老师给我们指定了论文的研究方向,并在研究方法方面给予了指导。

参考文献

- 1 Levine B, Garcia-Luna-Aceves J J. A Comparison Reliable Multicast Protocols. *Multimedia Systems Journal (ACM/Springer)*, 1998, 6(5)
- 2 Levine B, et al. Organizing Multicast Receivers Deterministically by Packet-Loss-Correlation. In: *Proc. of ACM Multimedia'98*. September 1998. Available at: <http://www.cse.ucsc.edu/~brian/pubs/>
- 3 Whetten B, et al. The RMTP-II Protocol, draft-whetten-rmtp-II-00.txt, Internet Draft, IETF, April, 1998
- 4 Papadopoulos C, et al. An Error Control Scheme for Large-Scale Multicast Applications. In: *Proc. of Infocom'98*, San Francisco, CA, March 1998
- 5 Papadopoulos C, et al. Performance Comparison of LMS and PGM using Simulation. Available at: <http://dworkin.wustl.edu/~christos>
- 6 Handley M, Whetten B, Kermode R. The reliable multicast design space for bulk data transfer, draft-ietf-rmt-design-space
- 7 Yavatkar R, Griffioen J, Sudan M. A reliable dissemination protocol for interactive collaborative applications. In: *Proc. ACM Multimedia*. November 1995. 333~344
- 8 Deering S, Cheriton D. Multicast routing in datagram internetworks and extended LANS. *ACM Transactions on Computer Systems (TOCS)*, 1990, 8(2): 85~110
- 9 Floyd S, et al. A Reliable Multicast Framework for Lightweight Sessions and Application Level Framing. In: *Proc. Of ACM Sigcomm'95*. 342~356
- 10 Paul S, et al. RMTP: A Reliable Multicast Transport Protocol for High-Speed Networks. In: *Proc. of the Tenth Annual IEEE Workshop on Computer Communications*. September 1995
- 11 Ratnasamy S, McCanne S. Inference of multicast routing tree topologies and bottleneck bandwidths using end-to-end measurements. In: *INFOCOM'99* [5]. Available at: <http://www.cs.berkeley.edu/~sylvia/>
- 12 Spakman T, et al. Pragmatic General Multicast (PGM) Transport Protocol Specification, draft-speakman-pgm-spec-01.txt.