基于混合式 IP 组播的大规模分布式仿真数据传输分配策略

刘晓建, 钟海荣, 叶超群, 金士尧
(国防科技大学计算机学院, 湖南长沙 410003)

摘 要: 在涉及到/ 一对多0或/ 多对多0数据传输时, 人们通常假定只有广播、IP 组播和多次/ 点对点0方式可用, 研究点对点和IP 组播两种数据传输方式的分配策略. 本文针对这些传输方式的缺陷, 提出了混合式 IP 组播传输方式, 并分析、比较了各种传输方式的性能, 进而提出了基于混合式 IP 组播优先的信息流归并 IP 组播组分配算法. 测试表明, 本算法优于以往相关工作.

关键词: 大规模分布式仿真; IP 组播; 组播组分配; 混合式 IP 组播
中图分类号: TP39119 文献标识码: A 文章编号: 0372-2112 (2003) 11-1678-04

A Hybrid IP Multicast Based Data Dissemination Allocation Strategy in Large Scale Distributed Simulations

LIU Xiao-jian, ZHONG Hai-rong, YIE Chao-qun, JIN Shi-yao
(National University of Defense Technology, School of Computer, Changsha, Hunan 410003, China)

Abstract: When considering one-to-many or many-to-many transfer methods, people used to assume that only flooding, IP multicasting and multiple unicasting (point-to-point) are available. This article brings a new transmission approach: hybrid IP multicast. Based on a detailed analysis of these data transfer methods, we present an allocation strategy, which outperforms its counterparts.

Key words: large scale distributed simulation; IP multicast; multicast group allocation; hybrid IP multicast

1 引言

IP 组播[1]十分适于实现大规模分布式仿真系统中大量存在且相对稳定的/ 一对多0 √ 多对多0信息传输关系, 但由于网络路由器所支持的 IP 组播组数量十分有限, 一般须使多个不相关的信息流使用同一个 IP 组播组传输数据, 这导致信宿收到不相关报文, 从而降低 IP 组播的实际使用效果; 而使用多次/ 点对点0方式进行传输易出现输出阻塞(output throttled)[2])) 在信源 s 向信宿 d 发送报文前, 已经经过的延迟就超过了 d 可容许的消息延迟.

如何分配使用各种数据传输方式就成为实现仿真系统数据传输的一个关键技术.

本文首先介绍了现有分布式仿真数据传输分配策略, 然后提出了混合式 IP 组播传输方式, 分析了其性能, 进而提出了基于混合式 IP 组播优先的分配算法, 并与相关的算法进行了性能比较.

2 研究现状

在相当多的分布式仿真系统中, 被仿真空间(逻辑空间)被划分成多个网格, 每个网格对应一个 IP 组播组, 实体只向自己所在网格对应的组播组发送/接收信息[3,4]. 此时各个组播组的利用率可能相差很大, 存在组播组浪费. 在算法 LOC/IRLOC[2,5]中, 仿真运行过程中的实际信息流动关系及其流量被用来指导 IP 组播组的分配, 文献[6]对此进行了理论上的研究与分析. 此种分配算法的计算代价较大. 文献[7]基于仿真系统的结构特征, 为每个局域网分配一个 IP 组播组用于发送信息, 并设置报文转发过滤器负责进出本局域网的通信, 此方法实际造成了广域网上的广播. 本文所述的策略综合考虑了仿真系统的结构特征和信息流动关系特征.

3 混合式 IP 组播

3.1 依据与原理

与广域网上的 IP 组播相比较, 局域网层次上的组播实现起来要容易得多, 如共享式以太网和 FDDI 本身就支持广播/组播, 且其创建和维护代价极低. 由此我们提出一种新型的数据传输方式))) 混合式 IP 组播(HIPM, Hybrid IP Multicast).

在 HIPM 中, 每个信息流都对应一个全局统一命名的/ 内部0IP 组播地址, 此地址类似于 Internet 上的内部 IP 地址, 仅在局域网内有效, 而在局域网外无意义(即: 内部 IP 组播的报

收稿日期: 2003-02-10; 修回日期: 2003-06-09
基金项目: 国家自然科学基金(No. 60073033); 国防预研基金(No. 51404020201 KG 01 1.7)

文不会被传输到局域网外)(相应的,我们把原来意义上的 IP 组播称为全局 IP 组播.若无特殊说明,IP 组播指全局 IP 组播),这样可以有效减少路由器所需支持的组播组数量,降低组播组的创建、维护代价.

HIPM 的信宿分为两类:组长信宿和组员信宿.一般来说,对于存在信宿的局域网,有且只有一个组长信宿.信源通过点对点方式发送信息到各局域网内的组长信宿,后者在收到报文后,(视需要)将其转发到信息流对应的内部 IP 组播组中.组员信宿则通过加入该内部组播组来获取相应的信息.对于信源所在的局域网,不存在组长信宿,由信源负责本局域网内的 IP 组播.图 1 为混合式组播的示意图,表 1 为其总结.

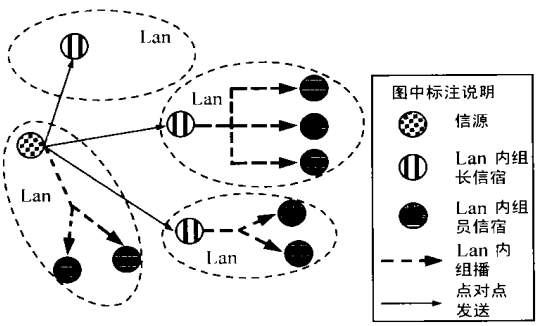


图 1　混合式组播示意图

表 1　HIPM 总结

报文转发者		传输方式	
确定方法	确定时机	广域网	局域网
同信息流对应	运行时动态确定	点对点	IP 组播

3.2　混合式 IP 组播的时间延迟分析

定义 1　信源 s 到信宿 d 的报文传输延迟 $t_{px}(s,d)$ 为:从报文出现在 s 主机的网络接口,到报文到达 d 主机的网络接口,中间所经过的时间延迟.

实际上,若将报文视作等大小,则报文传输延迟 $t_{px}(s,d)$ 由通信网络决定.设 s 所在的局域网为 S,d 所在的局域网为 D,则 $t_{px}(s,d)$ 可以表示为 $t_n(S,D)$.

定义 2　信源 s 到信宿 d 的信息传输延迟 $t_{dx}(s,d)$ 为:从 s 产生信息,到 d 开始处理该信息,中间所经过的时间延迟.

$t_{dx}(s,d)$ 与信源负载、网络状况、信宿负载等因素有关.

易知 $t_{px}(s,d)$ 是 $t_{dx}(s,d)$ 的一部分.若信源产生了信息,但并未立即发送出去(如先去发送别的报文),或信宿频繁受到网络中断,则 $t_{px}(s,d)$ 可能与 $t_{dx}(s,d)$ 有较大的差别.在大部分场合下,应用系统真正关心的是信息传输延迟 $t_{dx}(s,d)$,而不是报文传输延迟 $t_{px}(s,d)$.

关于报文的发送和接收时间开销,我们采用文献[2,5]的假设,即:所有仿真应用的报文发送开销 t_s 大致相同;仿真应用丢弃一个不相关报文的开销 t_r 同接收一个相关报文的开销相同,且与 t_s 大致相同.设局域网 S 内任意实体发送报文到局域网 D 内任意实体的报文传输延迟均为 $t_n(S,D)$,且设在报文到达 d 的网络接口后,需在 d 的报文队列中等待 $t_q(d)$(t_q 与进程切换、网络中断、主机报文队列长度等因素有关)时间才能得到处理.图 2(a)为使用 HIPM 时,组员信宿在一般情况下的最小信息传输延迟分析,其中 t_n 为从信源传输到信宿的报文传输延迟,tc_n 为同一个局域网内的报文传输延迟($tc_n(D)$ 与网络物理结构和网络负载状况相关.可假定在正常负载情况下,$tc_n(D)$ 不大于某常数),图中的阴影部分为由于使用 HIPM 而带来的额外时间开销 $t_c(m,D)=t_q+t_r+t_s+tc_n$,其中 m 为局域网 D 内的组长信宿.

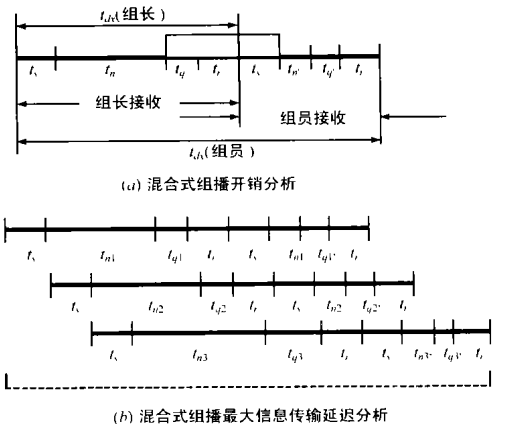


图 2　混合式 IP 组播的时间延迟性能分析

设信息流 s 的信源位于局域网 S 内,s 有 N 个信宿,它们分布在 K 个局域网中,将这些信宿局域网从 1 开始编号;信宿 d 位于 s 的第 k 个局域网,即局域网 D 中;m 为 D 内关于信息流 s 的组长信宿.若不考虑 S 和 D 是同一个局域网的情形(当 S=D 时不需转发报文,从而 $t_{dx}(s,d)$ 较小),令 $t_0=t_s+t_n(S,D)+tc_q(e)+t_r$,其中 e 为局域网 D 内 s 的信宿,则采用 HIPM 时,$t_{dx}(s,d)$ 为:

$$t_{dx}(s,d)=k\#t_s+t_n(S,D)+t_c(m,D)+tc_q(d)+t_r$$
$$=(k-1)\#t_s+t_0+t_c(m,D),\quad k\geq 1$$

类似地,可计算出在 P2P 和 IP 组播方式中,$t_{dx}(s,d)$ 分别为:

$(x-1)\#t_s+t_0,\ (x\geq 1)$,x 为 d 在 s 的信宿实体队列中的位置(P2P).t_0(IP 组播)

由于 s 可能与其它不相关信息流共用一个 IP 组播组(记此时的 IP 组播为 IP/M,即:使用信息流归并的 IP 组播),s 的信宿会接收到部分不相关报文.设 s 的流量为 K,s 所在的全局组播组中其它信息流的总流量为 Kc,且 t_q 不会因接收不相关报文而变大.

设不相关报文与 s 的报文等间隔到来,令 r=Kc/K,则 s 的信宿 d 在接收 s 的报文前,平均需先丢弃 r 个不相关报文,这导致信息传输延迟 $t_{dx}(s,d)$ 事实上增加了 $r\#t_s$,即实际使用 IP 组播的信息传输延迟为:$t_{dx}(s,d)=t_0+r\#t_s$.表 2 为各种一对多数据传输方式的时间延迟性能比较.

令 $L=t_c(m,D)/t_s$,根据表 2,可得:

结论 1　从信息传输延迟角度看,信息流 s 使用某 IP 组播组(而不是使用 HIPM)的必要条件是:$K>1+2r-2L$ 依据平均传输延迟或 $K>1+r-L$(依据最大传输延迟)

表 2 各种一对多传输方式比较

传输方式	发送次数	全局?	最大 $t_{dx}(s,d)$	平均 $t_{dx}(s,d)$
P2P	N	否	$t_0+(N-1)\#t_s$	$t_0+(N-1)t_s/2$
IP	1	是	t_0	t_0
IP/M	1	是	$t_0+r\#t_s$	$t_0+r\#t_s$
HIPM	K	否	$t_0+(K-1)\#t_s+t_c(m,D)$	$t_0+(K-1)t_s/2+t_c(m,D)$

结论 2 从负载平衡角度看, 在各主机负载大致相同的情况下, 信息流 s 可以被归并到某组播组中去的必要条件是: $K>1+r$.

证明 使用全局 IP 组播时, 信宿将多接收 $r\#K$ 个报文; 而在 HIPM 中, 信源主机将多发送 $(K-1)\#K$ 个报文. 当 $r<(K-1)$ 时, 使用 IP 组播更易导致全系统的负载加重.

结论 3 同 P2P 方式相比, HIPM 减少了信源发送报文的次数, 降低了信源主机的负载. 设 $c=N/K$, 为信宿局域网平均包含的信宿实体个数, 则从信息传输延迟角度看, 使用 HIPM 而不是点对点方式的必要条件是: $c<1+2L/K$(根据平均传输延迟)或 $c<1+L/K$(根据最大传输延迟)

结论 4(特例) 若信息流 s 的信宿全部处于某局域网 D 中, 则(1)若信源实体亦位于 D 中, 则 HIPM 实际上就是局域网 D 内的 IP 组播; (2)若信宿实体所在的局域网为 S, 且 $S\neq D$, 则在不考虑负载平衡的情况下, 当 D 内信宿实体个数 M 满足: $M>1+L$(最大传输延迟)或 $M>1+2L$(平均传输延迟)时, HIPM 传输方式优于 P2P.

以上结论均很直观, 故此处略去其证明过程.

说明 L 表示混合式 IP 组播导致的额外时间延迟内所能发送/接收的报文数量. 对于仿真应用而言, 此额外时间延迟等效于使用 IP 组播, 但接收了 L 个不相关报文. 可以通过监测实际系统的局域网 D 报文传输延迟 $t_{cn}(D)$, 组长信宿报文队列等待时间 $t_q(m)$ 和报文发送接收延迟 t_s 等以获得 L 的值.

4 数据传输方式分配算法

4.1.1 可用网络带宽不受限

结合 HIPM, 本文提出混合式 IP 组播优先的信息流归并 FM/HMF(Flow Merge with Hybrid IP Multicast First)组播组分配算法. FM/HMF 没有考虑网络可用带宽问题, 其主要思想为:

(1)混合式 IP 组播优先(HMF)(因点对点方式灵活性较差, 信宿集的任何改变都需通知信源更新信宿列表, 故虽在局域网内信宿数较少时, 其性能优于 HIPM, 仍不采用点对点方式).

对于所有信息流均优先考虑使用混合式 IP 组播, 若 HIPM 不易满足信宿对信息传输延迟要求, 则将此信息流标记为"需使用全局 IP 组播".

(2)信息流归并

对每个需使用全局 IP 组播的信息流 s, 计算其权值 $W(s)=K\#K_s$, 其中 K_s 为 s 的信宿局域网数量, K 为其流量. 按权值大小顺序, 获取一个信息流 s_0, 根据上节的结论 1、2 判断它能否放入某(可能已含有信息流的)组播组中. 若 s_0 不能成功放入任何一个组播组, 则使用 HIPM 传输该信息流.

在我们的信息流传输方式分配算法中, 涉及如下链表:

L_{mc}: 所有未分配传输方式、且"期望"(由其权值 W(s)和信宿局域网数量 K 决定.)使用全局 IP 组播传输数据的信息流集合;

L_{hm}: 所有未分配传输方式、且"期望"(由其权值 W(s)和信宿局域网数量 K 决定.)使用 HIPM 传输数据的信息流集合;

$L_{neutral}$: 所有未分配传输方式、且未对传输方式提出要求的信息流集合;

$L_{candidate}$: 算法执行过程中"缓存的"未分配传输方式的信息流集合.

算法开始时, 所有的链表均为空. 在算法执行过程中, 除 L_{hm} 外, 所有其它链表中的信息流都必须呈权值递降顺序排列.

具体信息流传输方案的分配算法如下:

① 对每一条信息流 s, 根据其信宿局域网的数量 K 和报文发送频率 K, 将其放入 L_{hm}, L_{mc}, 和 $L_{neutral}$ 三个链表之一, 并将 L_{hm} 中的所有信息流均标记为"已设置";

② 获取一个空闲 IP 组播组 mcg; 若失败, 则转入⑧;

③ 按照 $L_{candidate}$, L_{mc} 和 $L_{neutral}$ 的顺序, 获取第一个未尝试过归并到此 mcg 中去的信息流, 记其为 s_0, 并将其标记为"已被 mcg 尝试过". 若不能找到这样的 s_0, 则转到⑦;

④ 若根据上节结论, s_0 不能满足放入 mcg 中, 则转入③;

⑤ 将 s_0 从其所在的链表中取出, 指定其通过组播组 mcg 传输, 并把 L_{mc} 和 $L_{neutral}$ 中所有与 s_0 具有相同的信宿主机、且未指定传输方式的信息流取出, 放入链表 $L_{candidate}$. 注意: $L_{candidate}$ 需保持权值降序的性质.

⑥ 重复③~⑤, 直到所有的信息流均已被尝试, 或组播组中信息流总流量已到达了容许的最大值;

⑦ 将 $L_{candidate}$ 中所有信息流均放回原来所在的链表(此步完成后, $L_{candidate}$ 将变为空);

⑧ 重复②~⑦, 直到无可用的信息流, 或无可用的组播组;

⑨ 将 L_{hm} 中的信息流和剩余的未分配传输方式的信息流采用 HIPM 传输方式.

⑩ 算法结束.

4.1.2 可用网络带宽受限

实际网络环境的可用带宽总是有限的, 由于广域网上实体较多, 但可用网络带宽却比局域网的小, 从而在仿真运行过程中, 广域网链路更易超载. 而使用点对点或 HIPM 时, 广域网链路上可能会出现冗余报文, 加剧了网络链路带宽不足的问题, 故在传输方式分配算法中必须考虑广域网可用带宽因素. 因网络链路超载造成的后果一般比 IP 组播组超载要严重, 所以必要时应牺牲组播组来换取网络链路不超载. 对上述算法中步骤⑨的改进方法有分配前预估和分配后调整两种.

4.2.1.1　分配前预估

在确定使用 HIPM 传输某信息流 s 前,首先考察它所需的广域网络带宽.若相关网络链路的实际使用量接近可用带宽总量,则 s 以较大概率被归并到流过该链路且现有流量最小的组播组中去(即使该组播组已经满负荷).此算法将产生较保守的分配结果:IP 组播组可能会不必要的超载.它适于动态分配传输方式.

4.2.1.2　分配后调整

在步骤Á 结束后,检查分配方案是否会造成某些广域网络链路超载.若发生超载,则逐个将占用该链路带宽较大的信息流归并到流过该链路且当前总流量最小的组播组中进行传输.此时能充分利用可用带宽资源,较适于静态传输方式分配.

5　相关算法性能比较

我们将 FM/HMF 与同样基于信息流动关系特征的 IRLOC 全局组播组分配算法进行了性能比较.

在 LOC 算法中,定义信息流 s 的权值为 k#K,其中 k 为 s 的信宿数量,K 为 s 的流量.其算法的主要思想为:循环将具有最大权值的信息流放入 IP 组播组中(即:使用该组播组传输数据),直到通过此组播组传输的报文总量达到容许的最大值;不能放入任何组播组中去的信息流将通过多次点对点方式进行传输.由于多个不相关的信息流可能使用同一个 IP 组播组传输信息,信息流的信宿会接收到不相关报文,而 LOC 算法没有考虑不相关报文对组播组信宿带来的负面影响.因此 Morse 等人又提出了 LOC 算法的改进:IRLOC 算法.此时,/ 将信息流归并入 IP 组播组0 的一个必要条件是/ 使用 IP 组播的收益大于它带来的性能损失0.

因系统规模较大,故本文在单机上用仿真方法进行了算法的评测.在评测过程中考虑了网络带宽、主机、实体和信息流的数量、发送频率及其比例、信宿实体的分布、信息流对传输延迟的要求等各种因素对分配算法的影响.评测的性能指标为信息传输延迟超过容忍值的数据传输链路的个数、超载主机数量及系统最大负载率($LOAD_{real}$/Capacity)和分配算法自身的执行时间.由于篇幅限制,现只取在某种配置下的评测结果.

表 3　某配置下的算法性能比较

	IRLOC	IRLOC/HM*	FM/HMF
分配算法执行时间(秒)	239	239	12
分配后系统代价	1741	925	653
超载主机个数	18	2	1
系统最大负载率	1.152	1.072	1.092

* IRLOC/HM 采用 HIPM 来实现 IRLOC 算法中要求的点对点传输,其分配策略同 IRLOC 相同.

因 FM/HMF 采用了混合式 IP 组播优先,极大地减少了可能要归并到全局 IP 组播组中去的信息流数量,使得算法执行速度明显快于 IRLOC 和 IRLOC/HM.另外 HIPM 的采用也减少了信源的报文发送次数和输出阻塞的情况,降低了系统的代价和主机的负载率.各种情况下的测试验证了以上分析的正确性.

6　总结

针对当前大规模分布式仿真对数据传输的需求,本文提出了混合式 IP 组播的信息发送方法,并进行了相关的性能分析,得出了各种数据传输方式的使用条件,最后给出了一种混合式 IP 组播优先的信息流归并分配算法))) FM/HMF 及其与相关研究的性能比较结果.在仿真系统运行过程中如何根据实际信息流的变动而对数据传输方式进行动态调整将是我们进一步研究的重点.

参考文献:

[1] Banikazemi M. IP multicasting: concepts, algorithms, and protocols [DB/OL]. http://www.cis.ohio-state.edu/jain/cis788-97/ip_multicast/, 1997.

[2] Morse K L, Zyda M. Multicast grouping for data distribution management [J]. Simulation Practice and Theory, Elsevier, 2002, 9(3-5): 121-141.

[3] Macedonia M R, Zyda M, et al. Exploiting reality with multicast groups: a network software architecture for large scale virtual environments [J]. IEEE Computer Graphics and Applications, 1995, 15(5): 38-45.

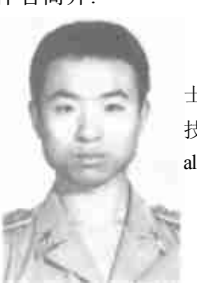
[4] Saville J. Interest management: dynamic group multicasting using mobile Java policies [A]. 1997 Fall, Simulation Interoperability Workshop [C]. 97-fsiw-020, 1997.

[5] Morse K L. An Adaptive, Distributed Algorithm for Interest Management [D]. USA: University of California, Irvine, 2000.

[6] Adler M, Kurose J F, et al. Channelization problem in large scale data dissemination [A]. Proceedings of ICNP 2001 [C]. Riverside, California, USA. Nov. 2001.

[7] 史扬. 新一代仿真框架 HLA/RTI 中数据过滤技术的研究与实现 [D]. 长沙:国防科技大学, 1999.

作者简介:



刘晓建　男,1974 年 11 月生于河北定州,博士生,现主要从事分布式交互仿真平台及其相关技术的研究.Email:tomorrowanewday@yahoo.com;aleck_liu@21cn.com.

钟海荣　男,1971 年 9 月生于湖南耒阳,博士生,现主要从事分布式仿真中的动态时空一致性研究.