ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.11, November 2009, pp.3034−3044
doi: 10.3724/SP.J.1001.2009.03405

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种基于内容的双向共享组播路由协议*

陈继明[1,2+], 潘金贵[2], 鞠时光[1], 贝 佳[2]

[1](江苏大学 计算机科学与通信工程学院,江苏 镇江 212013)

[2](南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

Content-Based Bi-Directional Shared Multicast Routing Protocol

CHEN Ji-Ming[1,2+], PAN Jin-Gui[2], JU Shi-Guang[1], BEI Jia[2]

[1](School of Computer Science and Telecommunications Engineering, Jiangsu University, Zhenjiang 212013, China)

[2](State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

+ Corresponding author: E-mail: cjm_eagle@hotmail.com

Chen JM, Pan JG, Ju SG, Bei J. Content-Based bi-directional shared multicast routing protocol. *Journal of Software*, 2009,20(11):3034−3044. http://www.jos.org.cn/1000-9825/3405.htm

Abstract: In order to improve the scalability of multicast protocol in large-scale distributed interactive systems, a content-based bi-directional shared multicast routing protocol is presented, which is called CBSMRP (content-based bi-directional shared multicast routing protocol). Combined with active routing methods and content-based publish/subscribe pattern, this new protocol supports active routing and bi-directional filtering according to the content of packages in a bi-directional shared multicast tree based on CBT (core-based tree) structure, which cannot only solve the problems in the allocation and maintenance of multicast addresses, but also efficiently reduce the network load. Experimental results and application show that the protocol is scalable enough to meet network communication requirements of large-scale distributed interactive systems.

Key words: active routing; content-based publish/subscribe pattern; bi-directional filtering; scalability; large-scale distributed interactive system

摘 要: 针对组播协议在大规模分布式交互系统应用中面临的可扩展性问题,提出一种基于内容的双向共享组播路由协议 CBSMRP(content-based bi-directional shared multicast routing protocol).该协议结合运用了主动路由思想和基于内容的发布-订购模式,在基于 CBT(core-based tree)结构的双向共享组播树中,根据数据包的内容实现主动路由和双向过滤,不仅解决了组播地址的维护和分配等问题,而且能够有效地减轻系统的网络负载.仿真实验及实际应用表明,该协议具有较好的扩展性,能够满足大规模分布式交互系统的网络通信要求.

关键词: 主动路由;基于内容的发布-订购模式;双向过滤;扩展性;大规模分布式交互系统

中图法分类号: TP393 **文献标识码**: A

随着互联网技术的飞速发展,各种各样的分布式交互系统不断涌现,如分布式虚拟环境系统、分布式协作

* Supported by the National Natural Science Foundation of China under Grant Nos.60473113, 60533080,60773049 (国家自然科学基金)

Received 2008-04-08; Accepted 2008-06-11

系统等.由于IP 组播协议在多对多通信时只在同一条物理链路上传输一份数据拷贝,避免了不必要的通信,因此在分布式交互系统中得到了广泛的应用.目前主要有 3 种方式:① 将交互空间区域和组播地址建立映射关系,通过静态分配或中心服务器来分配地址[1−3];② 为参与交互的每个参与者分配一个组播地址,当两个参与者的空间区域(或预感区域)相交时,参与者加入对方的组播组[4,5];③ 以某种特定的类或内容作为标签,通过标记通信群体的方式进行组播通信[6].但随着分布式交互系统规模的不断增大,组播技术开始面临可扩展性问题,主要体现在以下 3 个方面:第一,由于路由器能够支持的组播地址空间是有限的,对于存在大量对象的交互空间来说,路由器维护每个地址的开销就变得异常庞大;第二,缺乏简单而有效的组播地址动态管理方法,地址分配协议过于复杂,实用性不强;第三,尽管组播技术可以有效降低网络通信量,但当系统中存在大量的对象时,即使每个对象只需要少量带宽,其通信需求累加起来也会超过网络所能承受的限度.

上述可扩展性方面的问题使得组播技术在大规模分布式交互系统中的应用陷入了困境,基于内容的发布/订购通信模式[7,8]和主动路由技术[9]的研究为解决上述问题提供了一个新的思路,即可以将交互系统中对象具有的属性(如空间区域、预感区域、兴趣区域等)作为对象发布和订购的内容,系统则根据数据包的内容进行主动路由和转发,可有效解决组播技术面临的路由地址的维护和分配等问题.因此,本文提出了一种新型的基于内容的双向共享组播路由协议——CBSMRP.该协议不仅能够根据数据包的内容决定其转发的路由地址,还能通过基于内容的路由算法实现数据包的双向过滤,有效地降低了网络中数据信息的通信量,从而进一步减轻网络负载,提高系统的扩展性.

1 相关研究

目前,大多数分布式交互系统采用将交互空间区域和组播地址建立映射关系的方法,一般包括静态分配地址或者通过中心服务器来分配地址.在NPSNET-V系统中[1],将网格与组播地址一一对应,对象进入不同的网格则对应加入不同的组播组.DIVE系统[2]将组织成层次树的场景划分为不同的Light Weight Group,DIVE为每一个Light Weight Group分配一个组播地址,同时场景层次树的根节点也被分配一个组播地址,作为Name Server处理组播地址的查询索引. MASSIVE-3 系统[3]将虚拟空间划分为若干个区域(locale),为每个Locale分配一个组播地址.MASSIVE-2[4]和Community Place[5]系统则为每个参与者分配一个组播地址,遵循SMOI(spatial model of interaction)原则,当两个参与者的预感区域相交时,参与者加入对方的组播组,使得双方的状态能够共享.另外,在一些特殊的应用领域中,还可以将某种特定的类或内容需求作为标签来标记通信群体,如在HLA军事仿真中将属于同一军种的士兵作为一个通信群体而定义相应的组播地址[6].

随着分布式交互系统规模的不断增大,上述方法开始面临的一个主要的问题即需要大量的组播地址.举例来说,对于 100km×100km×10km的大规模虚拟交互空间,如果每个网格的大小为 1km×1km×1km,就需要 10 万个组播地址.如果路由器维护每个地址的开销是 1Kbytes,那么一个仿真应用就需要路由器维护 100M存储空间,而且同时还需要传播这些地址以及组的信息.为了解决组播地址分配问题,Morse提出了根据通信连接图来分配组播地址的方法[10].该方法为每个应用维护一个路由空间,根据联邦成员的发布区域和订购区域是否相交来确定联邦成员是否需要通信,并据此建立联邦成员之间的有向连接图.Morse把地址分配问题归纳为如何将n个连接分配到m个组播地址中,使得每个数据包从联邦成员发出该数据到其他联邦成员收到该数据的延时都小于某个阈值,并根据贪婪算法的思想提出的最大输出连接(largest outgoing connection,简称LOC)算法和输入限制的最大输出连接(input-restricted LOC,简称IRLOC)算法来实现.这两种算法虽然是分布式算法,但都需要全局的通信连接图,并且由于联邦之间的通信关系一直在不断变化,因此需要动态地收集所有连接信息,这给算法的应用带来了很大的困难.

Adler将组播地址分配问题称为频道化问题,即在大规模数据分发系统中,如果给定数量有限的组播地址,如何找到数据源与组播地址的对应关系,以及组播地址和接收者的对应关系,同时保证每个接收者能够接收到他们感兴趣的数据源的数据,并且使总通信量以及接收者接收到的不必要的数据最少[11].Adler指出频道化问题是个NP-Complete问题,并给出了寻找次优解的 3 种方法:第 1 种是随机分配方案;第 2 种方法称为FBM (flow

based merge),该算法首先假设每个数据源有一个组播地址,然后对这些组进行若干次合并,使得每次合并的代价最小,直到组的个数等于限定的个数;第 3 种方法称为UBM(user based merge),该算法假设每个用户被分配一个组播地址,然后进行同样的合并操作.FBM方法在通常情况下要好于UBM,除非数据源的个数远大于用户数.尽管如此,FBM算法的本质与Morse的LOC算法是相同的,仍然面临着需要收集全局信息的问题,且得到的分配方案所需的通信量也会增加系统的网络负担.另外,文献[12,13]分别提出两种在应用层实现组播技术的方法,但由于缺乏扩展性和难以满足快速动态改变路由信息等方面的原因也未能有效地应用到大规模分布式交互系统中.

2 CBSMRP 协议

现有的组播算法在主机与路由器之间一般采用IGMP协议来维护组信息,但当超出子网的范围时,负责子网与外部网络通信的路由器就需要连接到另一个运行组播协议的路由器上,这些支持组播协议的路由器构成了一个拓扑结构——组播树.组播技术经过多年的发展,已有多个路由协议被提出来.根据其构造组播树算法的不同,大致可以分为两类:一类是构造SBT(source-based tree)结构的组播树,如DVMRP、MOSPF组播协议等;另一类是构造CBT(core-based tree)结构的组播树,如PIM-SM、CBT组播协议等[14].

考虑到 SBT 与 CBT 结构在可扩展性方面的特点,本文选择 CBT 作为基本的通信架构,主要基于以下几个原因:① CBT 结构简单,易于扩展,只需在已有的路由器下连接更多的路由器或者主机就可以达到扩充的目的;② 与SBT相比,能够有效地减少路由器上的存储规模(SBT中每个路由器上的存储规模为 $O(S\times G)$,而在CBT中的存储规模仅为 $O(G)$,其中 S 表示数据源的个数,G 表示分组的数目);③ CBT 与发布/订购模式相结合,对象可以在任何时刻通过简单地向路由器发送新的订购信息,藉此改变通信关系,因此对象间通信关系的改变是完全动态的,可以实现大量对象在不断运动变化的状态下与其他对象进行快速、准确的动态通信.

2.1 相关定义

定义 1. 一个网络可表示为一个图 $G=(V,E)$,其中 V 表示所有的节点集合,E 表示节点间通信链路的集合.

在由n+1 个路由器构成的CBT网络中,路由器节点R^c被组织为树状结构,其中作为根节点的路由器被称为核心路由器 R_0^c,其他路由器 R_i^c $(1\le i\le n)$则作为非叶子节点进行层次状组织,连接作为叶子节点的主机H^c,即 $V=R^c\cup H^c$.

定义 2(虚拟接口(VIF)). 路由器上为通信链路建立的接口可记为二元组⟨*VIFid*,*RHid*⟩.其中,*VIFid* 表示虚拟接口的端口序号,*RHid* 表示通过链路连接的路由器或主机的地址.

在路由器中,如果其 VIF 连接的节点为上游节点,则称为上游 VIF,记为 UVIF;若 VIF 连接的节点为下游节点,则称为下游 VIF,记为 DVIF.对于拥有 m 个 VIF 的路由器 R_i^c $(0\le i\le n)$, $UVIF^{R_i^c}$ 和 $DVIF^{R_i^c}$ 可分别表示为

$$UVIF^{R_i^c}=\{VIF_k^{R_i^c}\mid 1\le k\le m\wedge n_k\in path(R_i^c,R_0^c)\}$$

$$DVIF^{R_i^c}=\{VIF_k^{R_i^c}\mid 1\le k\le m\wedge R_i^c\in path(n_k,R_0^c)\}$$

其中,n_k表示路由器上第k个虚拟接口 $VIF_k^{R_i^c}$ 对应的节点, $path(R_i^c,R_0^c)$ 和 $path(n_k,R_0^c)$ 则分别表示从节点 R_i^c 和n_k到核心路由器 R_0^c 的一条有效路径.

定义 3(订购区域). 路由器对其下游节点订购内容的表示方式.对于路由器$R_i^c(0\le i\le n)$,它所对应的订购区域记作$SR(R_i^c)$.如果 R_i^c 有 m 个 VIF $(VIF_1^{R_i^c},VIF_2^{R_i^c},...,VIF_m^{R_i^c})$,对于$VIF_k^{R_i^c}$ $(1\le k\le m)$,它所维护的订购区域记作 $SR(VIF_k^{R_i^c})$,则 R_i^c 的订购区域为

$$SR(R_i^c)=\bigcup_{1\le k\le m\wedge VIF_k^{R_i^c}\in DVIF^{R_i^c}} SR(VIF_k^{R_i^c}) \tag{1}$$

定义 4(裁剪区域). 路由器对其上游节点订购内容的表示方式.对于路由器 R_i^c $(1\le i\le n)$,R_i^c必然有一个上游

路由器,记为 R_j^c. 如果 R_j^c 有 p 个下游 VIF,其中第 q 个虚拟接口 $VIF_q^{R_j^c}$ 与 R_i^c 相连,那么 R_i^c 的裁剪区域为

$$CR(R_i^c)=\begin{cases}\left(\bigcup_{1\le s\le q-1\wedge VIF_s^{R_j^c}\in DVIF^{R_j^c}} SR(VIF_s^{R_j^c})\right)\cup\left(\bigcup_{q+1\le s\le p\wedge VIF_s^{R_j^c}\in DVIF^{R_j^c}} SR(VIF_s^{R_j^c})\right), & j=0\\ \left(\bigcup_{1\le s\le q-1\wedge VIF_s^{R_j^c}\in DVIF^{R_j^c}} SR(VIF_s^{R_j^c})\right)\cup\left(\bigcup_{q+1\le s\le p\wedge VIF_s^{R_j^c}\in DVIF^{R_j^c}} SR(VIF_s^{R_j^c})\right)\cup CR(R_j^c), & j\ne 0\end{cases} \tag{2}$$

2.2 路由信息的构造与维护

为了有效地实现对组播树中路由信息的构造和维护,我们设计了以下 4 条原则:

原则 1. 规定路由信息的构造过程由作为接收者的对象发起,通过订购消息(*SUBSCRIBE*)和回复消息(*REPLY*)来完成;对象发出包含订购区域的 *SUBSCRIBE* 消息后,必须等待直到收到 *REPLY* 消息,才能确定本身的订购信息已经成功地被接受,否则不能断定是否建立连接.

原则 2. 在路由器尚未加入组播树的情况下,如果是第 1 次接收到 *SUBSCRIBE* 消息,则将该 *SUBSCRIBE* 消息向上游路由器发送;如果再次接收到新的 *SUBSCRIBE* 消息,若不改变原先已发送的 *SUBSCRIBE* 消息的订购区域的并集,则不再向上发送 *SUBSCRIBE* 消息,否则需要重新向上游路由器发送.

原则 3. 当路由器接收到第 1 个 *REPLY* 消息时,表明路由器已经加入了组播树,可以立即回复收到的 *SUBSCRIBE* 消息;但当下游路由器发送的订购区域发生变化时,则需同时向上游路由器和下游路由器发送更新消息(*UPDATE*)和 *REPLY* 消息.

原则 4. 在路由器已经加入组播树的情况下,当收到 *SUBSCRIBE* 消息时,需要重新计算各虚拟接口的裁剪区域,并根据各下游虚拟接口的裁剪区域的变化情况选择性地发送更新裁剪区域的 *UPDATE* 消息.

根据以上原则,路由器 R_i^c 在第 k 个 VIF($VIF_k^{R_i^c}$)接受订购消息 *SUBSCRIBE*(*SR*(*X*))的路由算法如下:

算法 1.

if $R_i^c.active$ ==false then //如果路由器 R_i^c 尚未加入组播树

wait and sent *SUBSCRIBE* to $UVIF^{R_i^c}$; //等待并向上游转发订购消息

else

if $SR(X)\subseteq SR(VIF_k^{R_i^c})\wedge SR(X)\subseteq SR(R_i^c)$ then

sent *REPLY* to $VIF_k^{R_i^c}$ directly; //直接发送回复消息

if $SR(X)\not\subset SR(VIF_k^{R_i^c})\wedge SR(X)\subseteq SR(R_i^c)$ then

calculate $SR(VIF_k^{R_i^c})$; //计算 $VIF_k^{R_i^c}$ 的订购区域

calculate $CR(VIF_j^{R_i^c})$ and send *UPDATE* to $DVIF^{R_i^c}$; //计算 R_i^c 下游接口裁剪区域并发送更新消息

sent *REPLY* to $VIF_k^{R_i^c}$; //发送回复消息

if $SR(X)\not\subset SR(VIF_k^{R_i^c})\wedge SR(X)\not\subset SR(R_i^c)$ then

calculate $SR(VIF_k^{R_i^c})$; //计算 $VIF_k^{R_i^c}$ 的订购区域

calculate $SR(R_i^c)$ and send *UPDATE* to $UVIF^{R_i^c}$; //计算 R_i^c 的订购区域并发送更新消息

calculate $CR(VIF_j^{R_i^c})$ and send *UPDATE* to $DVIF^{R_i^c}$; //计算 R_i^c 下游接口裁剪区域并发送更新消息

sent *REPLY* to $VIF_k^{R_i^c}$ //发送回复消息

2.3 数据的转发

在传统的组播路由算法中,路由器上数据的转发只能根据数据包中的 IP 地址信息决定数据包的转发地址.

路由器工作简单,实现的功能也极为有限.本文中我们采用了基于内容的订购-发布模式,将对象的订购区域作为路由信息存储在各个路由器上,因而可利用主动路由技术实现数据的主动转发.主动路由技术是在 IP 路由技术的基础上,采用软件的形式实现复杂的路由协议,从而扩充路由器的路由能力.采用主动路由技术,可以在每个路由器上将其存储的订购路由信息与数据包的内容进行匹配运算,并根据匹配结果动态决定该数据包的转发地址.因此,主动路由思想的引入不仅解决了组播技术面临的组播路由地址的维护和分配问题,而且可以实现数据信息的动态过滤,将有效地降低网络通信的信息量.

理论上,CBSMRP 协议中采用的基于内容的订购-发布模式在保证订购过程正确语义的基础上,就可以采用主动路由方式,即根据数据包的内容进行转发.此时考查各个路由器的行为:当路由器在向其下游节点转发数据包时,有精确的订购区域指导转发方向;但向上游节点转发时,却没有任何路由信息提供给路由器以判断其上游路由器是否需要该数据以及需要哪些数据.为此,我们为每个路由器的上游 VIF 关联一个裁剪区域(定义 4 给出了具体的定义),裁剪区域是上游节点订购区域的总和,在数据包向上游节点发送的过程中,路由器可根据其裁剪区域对数据包进行过滤和转发.因此在每个路由器节点上,无论是向下游节点还是上游节点转发数据包时,都能将数据包中的内容与路由器存储的订购区域或裁剪区域进行匹配运算,并根据匹配结果对数据包实现双向过滤和转发.

定义 5(匹配). 路由器上根据订购区域或裁剪区域来判断数据包是否通过检测的过程.如果某个数据包通过 $VIF_k^{R_i^c}$ 到达拥有 m 个 VIF 的路由器 R_i^c $(0\le i\le n)$,若将该数据包记作 $D_k^{R_i^c}$,它所关联的发布内容记作 $PR(D_k^{R_i^c})$,则可以定义 $D_k^{R_i^c}$ 与 $VIF_j^{R_i^c}$ $(1\le j\le m)$之间的匹配函数 $match: D_k^{R_i^c}\times VIF_j^{R_i^c}\to\{\text{true,false}\}$.

$$match(D_k^{R_i^c},VIF_j^{R_i^c})=\begin{cases}\text{true}, & ((j\neq k)\wedge(PR(D_k^{R_i^c})\cap SR(VIF_j^{R_i^c})\neq\varnothing))\wedge VIF_j^{R_i^c}\in DVIF^{R_i^c}\\ \text{false}, & ((j=k)\vee(PR(D_k^{R_i^c})\cap SR(VIF_j^{R_i^c})=\varnothing))\wedge VIF_j^{R_i^c}\in DVIF^{R_i^c}\\ \text{true}, & ((j\neq k)\wedge(PR(D_k^{R_i^c})\cap CR(R_i^c))\neq\varnothing))\wedge VIF_j^{R_i^c}\in UVIF^{R_i^c}\\ \text{false}, & ((j=k)\vee(PR(D_k^{R_i^c})\cap CR(R_i^c))=\varnothing))\wedge VIF_j^{R_i^c}\in UVIF^{R_i^c}\end{cases}\tag{3}$$

假设路由器 R_i^c 从第 k 个 VIF($VIF_k^{R_i^c}$)接收到发布信息 $D_k^{R_i^c}$,具体算法如下:

算法 2.

```
if VIF_k^{R_i^c} ∈ UVIF^{R_i^c}   //如果数据包从上游接口到达
    for each VIF_j^{R_i^c} in DVIF^{R_i^c}
        if match(D_k^{R_i^c}, VIF_j^{R_i^c}) ==true   //根据式(3)进行匹配运算
            forward it;   //转发数据包
        else
            filter it;   //过滤数据包
if VIF_k^{R_i^c} ∈ DVIF^{R_i^c}   //如果数据包从下游接口到达
    for each VIF_j^{R_i^c} in (UVIF^{R_i^c} ∪ DVIF^{R_i^c})
        if match(D_k^{R_i^c}, VIF_j^{R_i^c}) ==true   //根据式(3)进行匹配运算
            forward it;   //转发数据包
        else
            filter it   //过滤数据包
```

3 节点的加入与退出

当有新的网络节点加入时,该节点可同时作为发布者和接收者通过 CBSMRP 协议向系统发送订购和发布信息.根据原则 1,新加入的节点首先应向网络系统发布包含订购区域的 *SUBSCRIBE* 消息并等待,直到收到系统的 *REPLY* 消息,才能确定本身的订购信息已经成功地被接受,然后再向系统发送发布信息.下面我们将通过实例

来说明新节点加入的过程及各个网络路由节点上路由信息的变化处理情况,如图 1 所示.

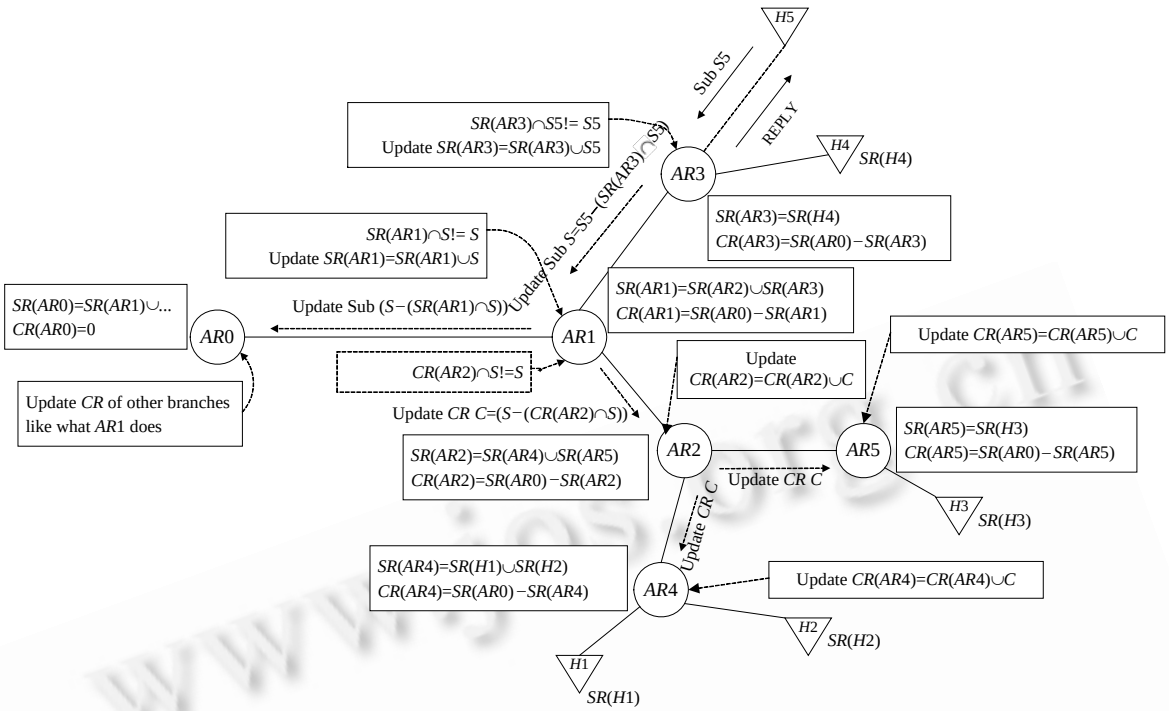

Fig.1 Update procedure of a new subscription

图 1 新订购引起的更新过程

在图 1 中,圆圈表示主动路由器节点,三角表示主机节点;实线方框部分表示当前主动路由器的路由信息,虚线方框表明路由器上的判断和处理;实线箭头表示实际消息的发送,虚线表示有条件的消息发送.当图 1 最上方的主机节点 *H*5 加入到已有的网络系统时,首先向其直接相连的上游路由器 *AR*3 发送包含订购区域为 *S*5 的订购消息(*SUBSCRIBE*(*S*5)).该消息到达 *AR*3 后,由于 *AR*3 已经加入系统(尚未加入的情况根据原则 1、原则 2 以及算法 1 中第 1 种情况处理),计算 *SR*(*AR*3)与 *S*5 的相交情况.当 *SR*(*AR*3)∩*S*5!=S5 时,表明 *AR*3 下游订购区域发生变化,根据原则 3 向上游 *AR*1 发送 *UPDATE* 消息并更新本地 *SR*(*AR*3),同时直接向下游发送 *REPLY* 消息.由于 *AR*3 没有下游主动路由器,因而无须判断下游是否需要进行裁剪区域的更新.若上游 *AR*1 接收到订购更新消息,则更新本地订购并判断是否需继续向其上游发送订购更新,同时由于 *AR*1 拥有其他下游主动路由器,因而需要判断下游裁剪区域是否需要更新,当 *CR*(*AR*2)∩*S*!=*S* 时,表明下游主动路由器 *AR*2 的裁剪区域并没有完全包含 *S* 即存在裁剪区域的变化,则根据原则 4,*AR*1 向下游主动路由器 *AR*2 发送裁剪区域更新消息.若 *AR*1 仍需要向上更新订购,则订购更新消息将到达核心路由器 *AR*0. *AR*0 根据原则 3 更新本地订购之后,进行与 *AR*1 相同的下游裁剪区域判断过程,进一步更新其他分支的裁剪区域.若 *AR*2 接收到裁剪区域更新消息,则更新本地裁剪区域,同时由于 *AR*2 下游还存在主动路由器(*AR*4 和 *AR*5),那么还需向下游转发裁剪区域更新消息(*AR*4 和 *AR*5 上裁剪区域的更新与 *AR*2 的更新相类似).上述过程完成之后,不仅系统内每个主动路由器上都包含了新节点 *H*5 的订购信息,而且 *H*5 也接收到系统的 *REPLY* 消息,可以进行发布信息的发送.

当有网络节点退出时,首先向本地主动路由器发送 *QUIT* 消息,主动路由器在接收到 *QUIT* 消息后,将对应的虚拟接口删除,并更新其他虚拟接口的裁剪区域和订购区域.当主动路由器的最后一个下游虚拟接口收到 *QUIT* 消息后,主动路由器向上游路由器发送 *QUIT* 消息,并删除上游虚拟接口,表示此时主动路由器从组播树中退出.

值得一提的是,在分布式交互系统中,参与对象通常订购的更新频率不高,且订购的内容在网络拓扑结构上

与地域相关,因而路由器存储的订购区域的更新不会十分频繁;另一方面,路由器存储的裁剪区域是由多个表示订购信息的兴趣表达式合并的,且原始表达式越多,合并后的范围也越大,当某一个原始表达式发生变化时,导致裁剪区域表示范围发生变化的概率通常很小.因此,当分布式交互系统中有对象加入或者退出时不会给各个主动路由器上路由信息的维护带来太大的开销.

4 仿真及性能分析

4.1 仿真模型

我们设计并实现了一个基于内容发布-订购模式的、事件驱动的CBSMRP协议仿真平台.该平台底层网络拓扑是基于乔治亚理工大学(Georgia Institute of Technology)的网络拓扑生成器GT-IMT生成器[15]生成的.利用生成器的Transit-Stub图形模块生成CBT结构的树形网络,其中核心路由器节点 1 个,其他路由器节点 1 050 个,每个节点的最大度设置为 12.在实验中,每个端节点(主机节点)同时作为发布者和接收者通过CBSMRP协议向系统发送订购和发布信息.主机对对象的空间活动行为进行仿真,用于实验的空间大小固定在 720m×720m×400m的空间范围,主机维护的活动对象基本上呈均匀分布,并在各自的均衡位置附近垂直方向上做简谐振动,对象的最大订购区域是以各自位置为中心,尺度为 220cm×220cm×200cm的范围内.

在本文中,我们主要采用数据包的平均接收率和树的总代价这两个性能指标对 CBSMRP 协议的相关性能进行评价,其具体定义如下:

定义 6(数据包的平均接收率(ρ_{arp})). $\rho_{arp}=\frac{P}{T}$, 其中,T表示时间,P表示在时间T内接收到的数据包的数量.ρ_{arp}反映了网络节点的处理量与通信量的大小.

定义 7(树的总代价(*Tcost*)). $Tcost=\sum_{v\in T}cost_T(v)$, 其中,$v$表示组播树$T$上的一条链路,$cost_T(v)$表示链路$v$的消息量.*Tcost* 是指整个网络中总的信息量,反映了系统的网络负载情况.

4.2 实验结果及性能分析

4.2.1 正确性

下面我们通过几个定理的非形式化证明来简要分析 CBSMRP 协议的正确性.

定理 1. CBSMRP 能够保证订购端接收到的发布消息必定满足其订购.

证明:(反证法)设某个订购端 $H_j^{R_i^c}$ 从其上游路由器 R_i^c $(0\le i\le n)$的虚拟端口 $VIF_j^{R_i^c}$ 接收到数据包 $D_k^{R_i^c}$. 假设 $PR(D_k^{R_i^c})\cap SR(H_j^{R_i^c})=\varnothing$, 即发布消息不满足订购,由于 $VIF_j^{R_i^c}$ 与 $H_j^{R_i^c}$ 直接相连,则有 $SR(H_j^{R_i^c})=SR(VIF_j^{R_i^c})$, 由此得到 $PR(D_k^{R_i^c})\cap SR(VIF_j^{R_i^c})=\varnothing$. 根据定义 5 与算法 2,易得 $PR(D_k^{R_i^c})\cap SR(VIF_j^{R_i^c})\neq\varnothing$. 与假设矛盾,故假设不成立,原命题成立. □

定理 2. 若存在满足某订购的发布消息,则 CBSMRP 能将该发布消息发送到订购端.

证明:因为系统的通信架构为树形结构,因此,作为叶节点的发布端 H_0^c 和订购端 H_m^c 之间有且仅有 1 条路径,设该路径上的路由器表示为 $R_1^c,...,R_k^c$.

根据定义 4 与算法 1 可得: $\forall 1\le i\le k\ (SR(H_m^c)\subseteq SR(R_i^c)\vee SR(H_m^c)\subseteq CR(R_i^c))=\text{true}$.

又因为存在满足某订购的发布消息(数据包 D),可得 $(PR(D)\cap SR(H_m^c)\neq\varnothing)$, 所以

$$\forall 1\le i\le k(PR(D)\cap SR(R_i^c)\vee PR(D)\cap CR(R_i^c))\neq\varnothing.$$

再根据算法 2,数据包 D 可在路径 $R_1^c,...,R_k^c$ 上逐个进行转发从而到达端点 H_m^c, 即若某数据包 D 满足某订购 $SR(H_m^c)$, 则 H_m^c 能够接收到 D,故命题成立. □

定理 3. CBSMRP 能够保证满足订购的发布消息最多到达订购端 1 次.

证明:因为数据包 D 的转发过程总是从某一个端口到达,向主动路由器的其他端口转发,因而是一个单向的

转发过程;又因为CBSMRP采用的通信构架是一种树形结构,任意两个叶节点之间有且仅有1条通路,所以订购端最多只会收到1个数据包,故命题成立. □

4.2.2 可扩展性

协议的扩展性主要可从以下3个方面来衡量:① 路由器存储信息量随系统规模的变化情况;② 协议构造是否易于动态扩充;③ 主机接收数据包的速率与系统的网络负载.

第一,采用双向共享组播树作为基本的拓扑结构,每个主动路由器存储的路由信息主要是虚拟接口以及维护这些虚拟接口所需的信息.除了组播地址对应的虚拟接口以外,每个虚拟接口的存储量都是固定的.在组播树中,每个主动路由器的下游路由器数量是固定的,如果没有主机直接与该路由器连接,则存储量为常数.

第二,协议构造非常易于扩充,只需在已有的主动路由器下连接更多的路由器或者主机就可以达到扩充的目的.例如,在增加新的主动路由器时,可以将其上游路由器的地址设置为某个已在组播树中的主动路由器的地址,当主机向新的路由器发送订购消息时,新路由器就会尝试加入组播树;同时,主机也可以直接连接到某个主动路由器上,只要该路由器已在组播树中,主机就可以顺利地加入系统.

第三,主机接收数据包的速率和系统的网络负载是衡量扩展性的最主要指标之一,它反映了主机处理量的大小和系统冗余通信量的大小.由于CBSMRP协议采用了主动路由技术,系统内的每个路由器都是根据主机的订购区域转发数据包,可有效地对数据包实现过滤,因此每个主机几乎不会收到冗余的数据量,同时系统的网络负载压力也会得到缓解.我们在第4.1节给出的仿真平台下对同样基于CBT结构的PIM-SM和CBT协议进行了仿真实验,下面分别通过两组实验结果作进一步分析和说明.

实验 1. 主机数据包的平均接收率(ρ_{arp}).图2描述了在系统内随机抽取的若干台主机上ρ_{arp}的平均变化情况,当采用PIM-SM和CBT协议时,主机上的ρ_{arp}随着用户(主机)的不断加入而迅速增长;而当采用CBSMRP协议时,由于数据包在转发的过程中能够根据发布内容和订购区域(或裁剪区域)的匹配结果实现精确过滤,主机几乎不会收到冗余的数据量,用户的增加对主机上的ρ_{arp}影响较小.

实验 2. 系统的网络负载.根据 *Tcost* 的计算方法,我们对系统运行时网络中的所有链路的消息量进行了记录,如图3所示.通过比较PIM-SM,CBT和CBSMRP协议的 *Tcost* 的变化情况来看,随着用户对象不断加入,CBSMRP的 *Tcost* 值的增长速度明显低于PIM-SM和CBT.分析可知,由于路由器可以通过匹配来决定是否需要转发外部发送来的数据,大量的"无效"数据信息都被路由器过滤了,有效地减少了系统中每条通信链路上的消息量,因而引起系统的网络负载的增加也是有限的.

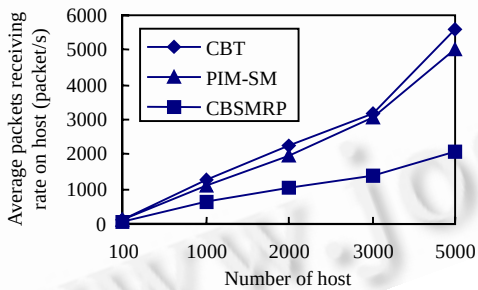

Fig.2 Average packets receiving rate on host

图2 主机数据包平均接收率

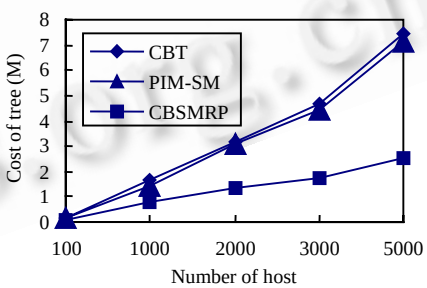

Fig.3 *Tcost* of system

图3 系统的 *Tcost* 值

4.2.3 核心路由器的瓶颈问题

采用CBT结构在构建组播树时面临的一个至关重要的问题就是如何减轻核心路由器的负担.如果所有的组播组都共享同一个组播树,则所有的通信量都要经过核心路由器进行转发,核心路由器必将成为通信瓶颈[16].而本文提出的CBSMRP协议采用了订购/发布机制,明显的通信组概念已经不存在了,主机之间通过发布和订购形成通信关系,并且由路由器作出如何转发的决定,核心路由器几乎不需要承担通信量的转发,其他主动路由器已经能够处理各自范围内的转发工作,因此可有效解决CBT结构中核心节点的瓶颈问题.

值得一提的是,CBSMRP协议中采用了裁剪区域的概念,可进一步有效地缓解核心路由器的压力.下面我们通过实验加以说明.图 4 给出在系统稳定运行的情况下,核心路由器上的ρ_{arp}的变化情况.从图 4 可以看出,在未采用裁剪区域时,核心路由器将接收所有主机和下游路由器发送的数据包,其ρ_{arp}约为 3 000packets/s,而采用裁剪区域时,每个下游的路由器在向上游路由器转发数据包时,将通过裁剪区域对其实现过滤,作为最上层的核心路由器接收到的数据包的数量得到了极大的控制,其ρ_{arp}下降约为 1 980packets/s.因此,裁剪区域可以有效地减小下游节点向上游节点转发的信息量,对缓解核心路由器的压力起着明显的作用.

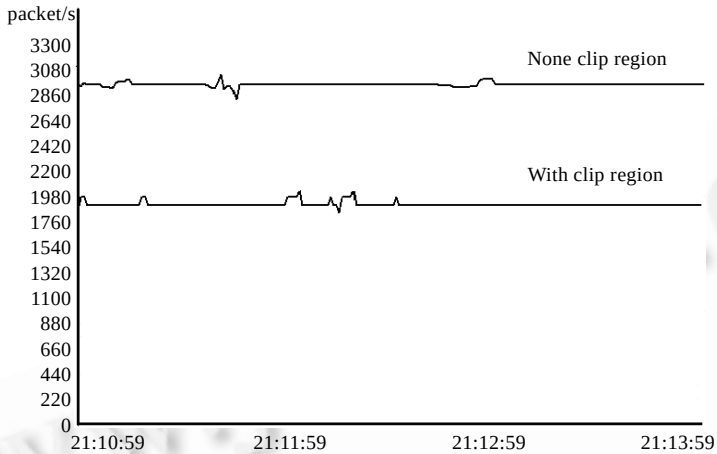

Fig.4 Average packets receiving rate on core router

图 4 核心路由器上数据包的平均接收率

5 CBSMRP 协议在 AIMNET 系统中的实现

AIMNET 是我们自主开发的一个基于发布/订购通信模式的分布式虚拟环境系统.该系统采用了主动兴趣管理技术,即将主动路由技术应用于兴趣管理领域,采用发布/订购模式,实现了数据转发和兴趣管理相结合,根据用户的兴趣,通过信息订购的方式来过滤分布式虚拟环境中对象发布的数据,有效地降低了系统的整体通信量,从而提高系统的可扩展性,进一步满足了系统中大量对象之间多对多的动态通信的要求.

在 AIMNET 系统中主要采用的协议包括 CBSMRP,STMP(stream transport multicast protocol)和 GTP (geometry transport protocol).CBSMRP 作为核心协议配置在路由器的应用层,直接利用现有 UDP 进行传输,这样不仅实现方便,而且不受兴趣表达式长度的限制,在传输时不需要实现数据包的重组.CBSMRP 协议主要的作用是构造组播树,传播和维护路由信息以及数据包的转发等.该协议从网络通信架构上保证了主动兴趣管理技术在分布式虚拟环境中实现.STMP 协议用于可靠的数据传输,通过 TCP 保证点到点的可靠性,该协议是为了满足今后网络会议和桌面共享应用对数据传输可靠性的要求.GTP 协议规定了主机与数据服务器之间的传输协议,主要用来传输分布式虚拟环境中的静态场景数据,实现一种可扩展的远程渲染体系结构.

图 5 描述了 CBSMRP,STMP 和 GTP 协议的基本结构以及在 AIMNET 系统中的配置情况.在主动路由器上配置协议包括 CBSMRP 和 STMP 两个协议,用于维护路由空间以及数据包的转发;数据服务器上一般只需采用 GTP 协议即可,即通过配置了 GTP Server 来响应和处理主机上 GTP Client 发送的请求,同时考虑到服务器有时也需像普通主机一样对活动对象的行为进行模拟,因此增加了 CBSMRP 和 STMP 协议.

在主动路由器上,CBSMRP 协议的主要功能有两个:一是当接收到订购信息时实现路由信息的维护和传播;二是当接收到发布信息时实现数据包的过滤和转发.在处理路由维护事件时,CBSMRP 协议由 4 个线程组成:输入线程(InputThread)、CBSMRP 处理线程(CBSMRPThread)、输出线程(CBSMRPOutputThread)和定时线程(TimerThread).其中,CBSMRPThread 和 CBSMRPOutputThread 都采用了 ActiveObject 设计模式,内置一个消息队列.路由器首先设置 InputThread 和 TimerThread 的回调对象,然后启动这 4 个线程.当订购消息到达时,

InputThread 调用先前设置的回调对象,该对象调用 CBSMRP 线程的 PostEvent 函数,直接根据数据包的类型(如 SUBSCRIBE,UPDATE 等)分别放入不同的队列中,最后唤醒 CBSMRPThread.CBSMRPThread 线程一直在等待各个队列中的事件,当任何一个队列中有事件发生时,取出同一个队列中的所有事件,并执行该类型事件的处理方法(具体见第 2.2 节).

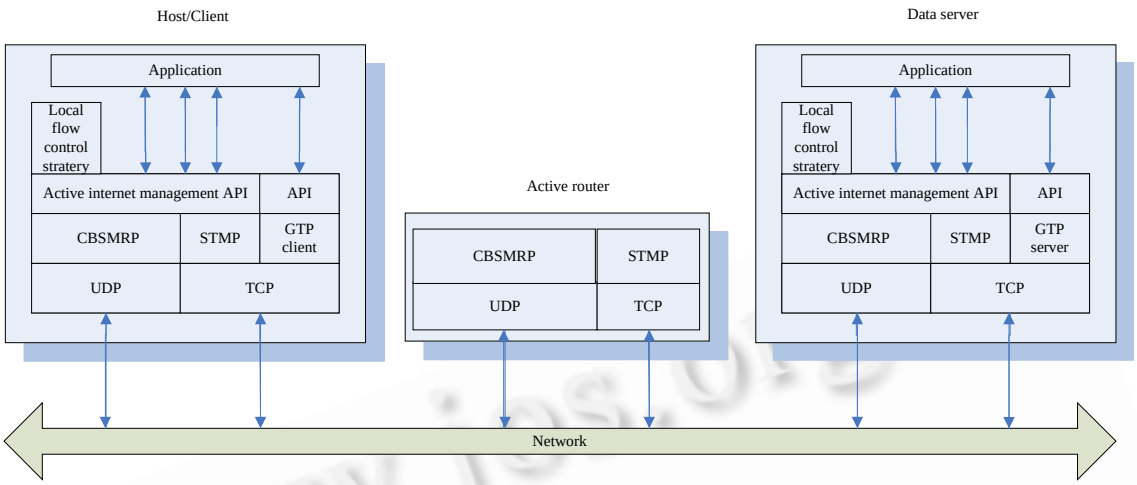

Fig.5 Structure and configuration of protocol

图 5 协议结构及其配置

在处理数据包转发事件时,考虑到 CBSMRP 利用 UDP 接收数据包,当 InputThread 来不及处理现有数据包从而导致 UDP 对应的 socket 缓存不够时,后继的 UDP 数据包很可能被丢弃.为了避免这种情况,CBSMRP 没有采用单个线程来处理输入的数据包,而是采用了基于 Proactor 设计模式的线程池(thread pool)技术.CBSMRP 协议在启动时创建一个可以进行重叠 I/O 的 socket,并将该 socket 与一个 I/O 端口相关联,然后创建固定数量的线程,每个线程分配的协议数据包作为接收缓冲区,等待数据到达.一旦有数据到达,调用该对象的 OnRecv 方法,即在路由空间中查找订购该数据包的虚拟接口,然后根据查找结果,直接通过调用同步 WSASentTo 函数向这些虚拟接口转发该数据包.

6 结 论

CBSMRP 作为一种新型的组播路由协议,不仅能够解决传统 IP 组播技术所面临的路由地址的维护及分配问题而且能够根据路由器上的订购区域和裁剪区域实现数据包的双向过滤,有效地提高了组播协议在大规模分布式交互系统中的可扩展性.目前,CBSMRP 协议的相关技术已经应用到我们自主开发的发布式虚拟环境系统 AIMNET 中,采用描述虚拟对象属性的兴趣表达式作为发布和订购的内容,并根据数据包的兴趣区域进行主动地过滤和转发,实际应用效果进一步验证了该方法的可行性和有效性.

致谢 在此,向对本文的工作给予支持和建议的同行,尤其是南京大学计算机科学与技术系潘金贵教授领导的 VR 研究小组的成员表示感谢.

References:

[1] Capps M, McGregor D, Brutzman D, Zyda M. NPSNET-V: A new beginning for dynamically extensible virtual environments. IEEE Computer Graphics and Applications, 2000,20(5):12−15.

[2] Frécon E, Stenius M. DIVE: A scaleable network architecture for distributed virtual environments. Distributed Systems Engineering Journal, 1998,5(3):91−100.

[3] Purbrick J, Greenhalgh C. Extending locales: Awareness management in MASSIVE-3. In: Thalmann D, Feiner S, eds. Proc. of the IEEE Virtual Reality 2000 Conf. Washington: IEEE Computer Society, 2000. 287−294.

[4] Greenhalgh C. Large scale collaborative virtual environments [Ph.D. Thesis]. Nottingham: University of Nottingham, 1997.

[5] Lea R, Honda Y, Matsuda K, Matsuda S. Community place: Architecture and performance. In: Proc. of the VRML'97 Symp. New York: ACM Press, 1997. 41−50.

[6] U.S. Department of Defense. High Level Architecture Rules, Version 1.3. DMSO, 1998. http://hla.dmso.mil/

[7] Xue T, Feng BQ. Research on routing algorithm and self-configuration in content-based publish-subscribe system. Journal of Software, 2005,16(2):251−259 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/251.htm

[8] Yuan HL, Shi DX, Wang HM, Zou P. Research on routing algorithm based on subscription covering in content-based publish/subscribe. Chinese Journal of Computers, 2006,29(10):1804−1812 (in Chinese with English abstract).

[9] Calvert K. Reflections on network architecture: An active networking perspective. ACM SIGCOMM Computer Communication Review, 2006,36(2):27−30.

[10] Morse KL. An adaptive distributed algorithm for interest management [Ph.D. Thesis]. Irvine: University of California at Irvine, 2000.

[11] Adler M, Ge ZH, Kurose J, Towsley D, Zabele S. Channelization problem in large scale data dissemination. In: Satish K, ed. Proc. of the 9th Int'l Conf. on Network Protocols. Washington: IEEE Computer Society, 2001. 100−109.

[12] Yang HC, Sanjay GR, Zhang H. A case for end system multicast. ACM SIGMETRICS Performance Evaluation Review, 2000, 28(1):1−12.

[13] Rowstron A, Kermarrec A M, Castro M, Druschel P. SCRIBE: The design of a large-scale event notification infrastructure. In: Crowcroft J, Hofmann M, eds. Proc. of the 3rd Int'l Workshop on Networked Group Communication. London: Springer-Verlag, 2001. 30−43.

[14] Koh S, Kong S, Park K. Enhanced core based tree for many-to-many IP multicasting. Telecommunications Review, 2001,11(3): 485−493.

[15] Zegura EW, Calvert K, Bhattacharjee S. How to model an Internetwork. In: Sohraby K, ed. Proc. of the IEEE Infocom'96. San Francisco: IEEE Computer Society Press, 1996. 594−602.

[16] Stardust.com, Inc. A survey of the history of Internet multicast. MCAST White Paper, 2000. http://citeseer.ist.psu.edu/stardust00survey.html

附中文参考文献:

[7] 薛涛,冯博琴.内容发布订阅系统路由算法和自配置策略研究.软件学报,2005,16(2):251−259. http://www.jos.org.cn/1000-9825/16/251.htm

[8] 苑洪亮,史殿习,王怀民,邹鹏.内容发布订阅中支持订阅覆盖的路由算法研究.计算机学报,2006,29(10):1804−1812.

陈继明(1977－),男,江苏镇江人,博士生,讲师,主要研究领域为分布式网络,虚拟现实技术.

鞠时光(1955－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为空间数据库系统,网络信息安全技术.

潘金贵(1952－),男,教授,博士生导师,主要研究领域为多媒体信息处理技术,虚拟现实技术.

贝佳(1979－),男,博士,讲师,主要研究领域为分布式虚拟环境,信息交换技术.