ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.5, May 2006, pp.1107−1114
DOI: 10.1360/jos171107

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

结构化 P2P 网络上可靠的基于内容路由协议*

汪锦岭[1,2+], 金蓓弘[1], 李 京[1]

[1](中国科学院 软件研究所 软件工程技术中心,北京 100080)

[2](中国科学院 研究生院,北京 100049)

Building Reliable Content-Based Routing Protocol over Structured P2P Networks

WANG Jin-Ling[1,2+], JIN Bei-Hong[1], LI Jing[1]

[1](Technology Center of Software Engineering, Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China)

[2](Graduate School, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: Phn: +86-10-62630989 ext 203, Fax +86-10-62562538, E-mail: jlwang@otcaix.iscas.ac.cn, http://www.ios.ac.cn

Wang JL, Jin BH, Li J. Building reliable content-based routing protocol over structured P2P networks. *Journal of Software*, 2006,17(5):1107−1114. http://www.jos.org.cn/1000-9825/17/1107.htm

Abstract: Much work has been done on building content-based publish/subscribe systems over structured P2P networks, so that the two technologies can be combined together to better support large-scale and highly dynamic systems. However, existing content-based routing protocols can only provide weak reliability guarantees over P2P networks. Based on the routing protocols of structured P2P networks, a new type of content-based routing protocol for pub/sub systems is designed, which is called Identifier Range Based Routing (IRBR) protocol. The IRBR protocol guarantees that the subscribing nodes always receive the interested events exactly once as long as the message from publishing nodes to subscribing nodes is arrivable in the P2P network. At the same time, it can also disseminate an event to all interested subscribers with less network traffic. A prototype pub/sub system has been developed on Pastry, and the experimental results demonstrate the fault-tolerance and routing efficiency of the protocol.

Key words: structured P2P network; publish/subscribe; content-based routing

摘 要: 在结构化 P2P 网络上构建基于内容的发布/订阅系统,可以很好地支持大规模、高度动态的分布式应用.然而,现有的基于内容的路由协议在 P2P 网络上只能提供弱的可靠性保证.根据结构化 P2P 网络的路由协议的特点,设计了一种新型的基于内容的路由协议——基于编码区间的路由(identifier range based routing,简称 IRBR)协议.IRBR 协议具有良好的容错性,只要事件的发布者与订阅者之间在 P2P 网络中是可达的,则订阅者一定能够收到它所订阅的事件,且只收到一次.同时,该协议也比现有的协议具有更高的事件路由效率.在 Pastry 上开发了一个原型系统,模拟实验表明了该协议的效率和容错性.

关键词: 结构化 P2P 网络;发布/订阅;基于内容路由

* Supported by the the National Hi-Tech Research and Development 863 Program of China under Grant No.2001AA113010 (国家高技术研究发展计划(863)); National Grand Fundamental Research 973 Program of China under Grant No.2002CB312005 (国家重点基础研究发展规划(973))

Received 2004-10-28; Accepted 2005-09-06

中图法分类号: TP393 文献标识码: A

发布/订阅(publish/subscribe,简称 pub/sub)是一种面向分布式计算环境的松散耦合的通信范型.在 pub/sub 系统中,信息发布者将信息以“事件”的形式发送给事件代理;信息订阅者向事件代理发出一个订阅条件,表示对系统中的特定种类的事件感兴趣;而事件代理则保证将所发布的事件及时、可靠地传送给所有对其感兴趣的订阅者.pub/sub 系统使得信息的生产者和消费者在时间、空间和控制流 3 个方面都被完全解耦合[1],因而能很好地满足大规模、高度动态的分布式系统的需要.

一个大规模的 pub/sub 系统中通常分布着多个事件代理,这些事件代理组织成一定的拓扑结构,每个事件代理为一定数量的客户(发布者或订阅者)服务.这种系统中的一个关键技术问题是消息的路由协议(通常被称为“基于内容路由协议”[2]),即系统中的各种消息按照何种路径从发出节点到达目的节点.虽然人们已经提出了很多种基于内容路由协议,但这些协议一般都建立在固定的网络结构之上,缺乏对节点故障以及拓扑结构变化的适应能力.目前,虽然有一些对 pub/sub 系统的失效恢复和路由重配方面的研究,但是还远未找到令人满意的解决方案.

另一方面,以 Pastry[3],Tapestry[4],Chord[5]和 CAN[6]为代表的结构化 P2P 网络,近年来得到人们的很大关注.这种网络具有一系列优点,例如分散控制(整个系统没有一个集中控制点)、自组织(各成员可以动态地加入和退出)以及较强的容错能力等.因此,很多人试图在结构化 P2P 网络上构建基于内容的 pub/sub 系统,希望利用 P2P 网络的优点,提高 pub/sub 系统对节点故障和拓扑结构变化的适应能力.但是,现有的这些系统都弱化了事件传输的可靠性,即不能保证所有当前活动节点都能收到它所感兴趣的事件.同时,很多系统还要求在 P2P 网络中有若干个特殊的“汇合点”,这些汇合点提供集中式服务,其负载远远超过其他节点,从而也丧失了 P2P 网络的分散控制和均衡负载的优点.

我们根据结构化 P2P 网络的特点,设计了一种新型的基于内容路由协议——基于编码区间的路由(identifier range based routing,简称 IRBR)协议.它能够较为自然地与 P2P 网络本身的路由协议集成在一起,并利用 P2P 网络路由协议的容错机制来提高事件传输的可靠性.只要事件的发布者与订阅者之间在 P2P 网络中是可达的,则订阅者一定能够收到它所订阅的事件,且只收到一次.与此同时,本协议也具有较高的路由效率.与现有的结构化 P2P 网络上的基于内容的 pub/sub 系统相比,本协议只需更少次消息转发就能将事件传送到各订阅者.

根据 IRBR 协议,我们在 Pastry 之上开发了一个 pub/sub 原型系统.模拟实验表明,本协议具有较好的路由效率和容错能力.

本文第 1 节介绍相关的研究工作.第 2 节介绍系统模型.第 3 节介绍 IRBR 协议的主要内容.第 4 节介绍对 IRBR 协议的模拟实验.第 5 节对全文进行总结.

1 相关工作

Pub/sub 系统可以被分为基于主题和基于内容两大类:在基于主题的系统中,事件被划分为若干个固定的主题,每个事件都只能属于其中的某一个主题,订阅者则对某一主题下的所有事件进行订阅;而在基于内容的系统中,订阅者根据事件的内部结构设置一个订阅条件,所有满足该条件的事件都将被传送给该订阅者.与基于主题的系统相比,基于内容的系统提供了更强的表达能力,因此本文重点考虑基于内容的 pub/sub 系统.

在固定网络上的 pub/sub 系统中使用的路由协议大多数是基于逆向路径转发[7]的思想.为了简化路由协议,有的系统[2,8,9]事件代理网络组织成树型结构或无向无环图结构,从而容错能力较差;也有的系统[10−12]将代理网络组织成有向无环图,并增加一些冗余边,以提高容错能力.但是,这些系统限制了客户的发布和订阅事件的能力:某些客户只能发布事件;其他客户只能订阅事件.文献[13–15]研究了当代理网络中出现链接失败时如何重配系统以及提高消息传输的可靠性.然而,这些方案都不能像 P2P 网络那样提供很好的容错能力和自组织特性.

另一方面,人们已经在结构化 P2P 网络上构建了一些 pub/sub 系统.这些系统所使用的路由协议多数是基于 Scribe[16]中提出的思想.Scribe 是一个建立在 Pastry 上的基于主题的 pub/sub 系统,每个事件主题在网络中都对

应一个特殊的节点,称为汇合点.各节点在订阅事件时,将订阅消息发送到相应主题的汇合点.汇合点根据各订阅消息的逆向路径,构建一棵以自己为根的事件分发树.各节点在发布事件时,也将它发送到相应主题的汇合点,再由汇合点通过事件分发树转发出去.树中的每个节点都定期向其孩子发送心跳消息,如果某节点在一定时期内未收到其父节点的心跳消息,它就认为父节点已失败,然后开始修复多播树.Scribe 只提供了弱的可靠性保证,在多播树的断裂和修复期间,汇合点所发出的事件可能会丢失.此外,它还存在树根负载较大和单点失败的问题.

Hermes[17]和 P2P-ToPSS[18]都是以 Scribe 的路由机制为基础,在此之上加以扩展,以构建基于内容的 pub/sub 系统.其思路是,首先将其全局事件空间分成很多个子空间,将每个子空间看成一个主题,一个订阅可能会涉及多个主题;然后,再用 Scribe 的方法去进行订阅和事件的转发.但是,这些系统不能提供比 Scribe 更强的可靠性保证.

Terpstra 等人[19]提出了一种“many trees”方法,以在 Chord 上构造基于内容的 pub/sub 系统.其基本思想是,首先为每个节点构建一棵以其为根的生成树,并将该树作为事件分发树;对于树中每个节点,它向父节点提出的订阅请求是其子树中各节点的订阅请求之和.该方法也弱化了事件传输的可靠性,不能保证事件能到达所有的订阅者.同时,该方法要求每个节点不仅知道其输出边指向的各节点,还应知道哪些节点的输出边指向自己,这将大大增加网络自组织的成本.而且在现有的结构化 P2P 网络中,大多数并不具备该功能.

2 系统模型

一个构建于结构化 P2P 网络之上的 pub/sub 系统可以分为 P2P 层、事件通知层和应用层 3 个层次,如图 1 所示.P2P 层用于将各事件代理组成一个自组织的 P2P 网络,事件通知层负责事件在各事件代理之间的转发;应用层是负责发布和接收事件的应用系统.我们所提出的 IRBR 协议是事件通知层的路由协议,它建立在 P2P 层路由协议上.为简化起见,在下面的讨论中,我们假定 pub/sub 系统的每个节点上只有 1 个应用层实例.

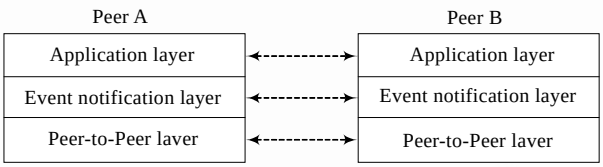


Fig.1 Layered structure of the system

图 1 系统的层次结构

事件通知层与另外两个层次之间通过“操作”进行交互,每个操作可以带有若干参数.不同节点的事件通知层之间则通过“消息”进行交互,每个消息也可以带有若干参数.

一般而言,事件通知层至少应完成应用层提出的如下几种操作:

- *subscribe*(*filter*):应用层对所有满足 *filter* 条件的事件感兴趣;
- *unsubscribe*(*filter*):应用层不再对满足 *filter* 条件的事件感兴趣;
- *publish*(*event*):应用层发布一个新的事件.

同时,事件通知层向应用层提出如下操作:

- *notify*(*event*):事件通知层告诉应用层,到达了一个它感兴趣的事件.

我们借鉴了结构化 P2P 网络的路由协议的思想来构建 IRBR 协议.虽然每种结构化 P2P 网络的路由协议有所不同,但总的来说,它们都是采用基于键的路由(key-based routing)协议[20].每个节点都有一个编码,节点之间通过有向的边相连,节点的每个输出边负责一定的编码区间.对于每个节点而言,它的各输出边和其所负责的编码区间构成了该节点的路由表.下面,我们以 Pastry 为例来介绍结构化 P2P 网络的路由协议.

在 Pastry 中,每个节点的编码是一个长度为 L 的 k 进制的数.Pastry 按照编码前缀来对全局编码区间进行分段,将其分成不同的子区间.设某节点的编码为 n,该节点的路由表为 RT^n.为了简化起见,我们可以将 RT^n 理解为如下集合

$$RT^n=\{(prefix,nodeId,address)\}.$$

其中的每个路由项表示:对于所有编码,以 *prefix* 为前缀的节点;本节点将以 *nodeId* 作为下一跳向其转发消息,

节点 *nodeId* 的 IP 地址为 *address.NodeId* 的编码一定以 *prefix* 为前缀,它相当于这一区间中的各节点的“代表”.例如,设 k=4,L=3,则编码为 213 的节点的路由表见表 1.如果整个网络中不存在以某编码为前缀的节点,则在路由表中不存在相应的项.

Table 1 Routing table of peer 213 in Pastry

表 1 Pastry 中节点 213 的路由表

Prefix	Node Id	Address
0	031	192.168.2.1
1	102	192.168.5.5
3	320	…
20	202	…
22	221	…
23	233	…
210	210	…
211	211	…
212	212	…

假设节点 213 要发送一个消息给节点 201.节点 213 在自己的路由表中查以“20”为前缀的路由项,得到下一跳 202,然后将此消息发给 202;节点 202 再在自己的路由表中查以“201”为前缀的路由项,将消息发给相应的地址.

根据结构化 P2P 网络的特点,我们可以根据路由表为每个节点创建一棵以其为根的生成树.其基本思想是“责任委派”[21]:生成树中的每个节点负责一个编码区间,根节点负责全局编码区间,其他每个节点所负责的编码区间是其父节点所负责的编码区间的一部分.因此,在 P2P 网络中比较简单易行的广播算法是“基于源转发(source based forwarding)”[7]广播算法.IRBR 协议即以该广播算法为基础.

3 基于编码区间的路由协议

3.1 过滤条件表

在传统的基于内容的路由协议中,每个节点都维护一个过滤条件表,记录了其每个邻居(以及通过该邻居所能到达的其他节点)的订阅请求.而在结构化 P2P 网络中,每个节点的邻居是经常变化的,因而不能按照邻居来组织过滤条件表.但是,对于结构化 P2P 网络中的每个节点而言,它对全局编码区间的划分是固定的,即路由表中的总项数以及每一项的编码区间是固定的,只不过每个编码区间中的“代表”可能经常变化.所以,我们可以按照编码区间来组织各节点的过滤条件表.

对于过滤条件 f,令 $f(e)$表示“事件 e 满足 f”,令 E_f 表示集合$\{e|f(e)\}$,即所有满足该过滤条件的事件.对于过滤条件 f_1,f_2,令 $f_1 \sqsupseteq f_2$ 表示 f_1 覆盖 f_2,即

$$f_1 \sqsupseteq f_2 \Leftrightarrow E_{f1} \supseteq E_{f2}.$$

每个节点的事件通知层都维护着一个过滤条件表.设节点 n 的过滤条件表为 FT^n,我们可以抽象地将它表示为如下集合:

$$FT^n=\{(prefix,filter)\},$$

其中每一项称为一个过滤项,表示在以 *prefix* 为前缀的编码区间中至少有一个节点对 E_{filter} 感兴趣.一个 *prefix* 可以对应多个过滤项.

过滤条件表中的 *prefix* 可以等于本节点自身的编码,表示本节点的应用层对哪些事件感兴趣.除本节点编码之外,其他各 *prefix* 都应在 Pastry 路由表中有相应的项.

3.2 Publish操作的处理

在 IRBR 协议中,事件的传播以“基于源转发(source based forwarding)”广播算法为基础并进行优化,以避免不必要的消息转发.当某节点发布事件时,该事件将沿着以该节点为根的生成树转发到其他节点.在事件转发过程中,如果生成树中的某节点所负责的编码区间中没有节点对此事件感兴趣,则该事件将不被发送到该子树.例

如,设节点 213 的应用层发布了事件 e,它满足过滤条件 f_a 和 f_b,则其事件分发树如图 2 所示.其中,每个节点旁的方框为其过滤条件表的内容,方框下的文字表示该节点所负责的编码区间.

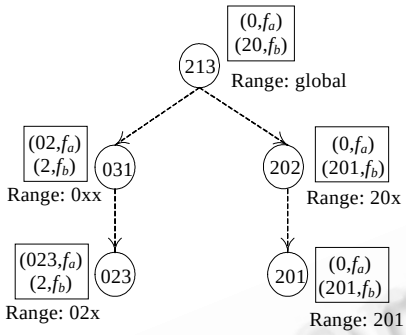


Fig.2 Dissemination tree for event e when $f_a(e)$ and $f_b(e)$ hold

图 2 当 $f_a(e)$和 $f_b(e)$成立时事件 e 的分发树

由于事件分发树是在事件发布时动态创建的,因而从订阅操作到发布操作之间的邻居节点的变化并不影响事件的发送.此外,一个消息只需要被发送到目标编码区间的任一节点即可,而并不依赖于某些固定的节点,从而进一步提高了本协议的容错性.

3.3 Subscribe操作的处理

在 IRBR 协议中,订阅消息的传播也以“基于源转发”广播算法为基础.当一个节点收到应用层的订阅请求后,它就将订阅消息按照以自己为根的生成树发送到其他节点.对于收到订阅消息的节点,它不关心该消息来自哪个输入边,而关心订阅者的编码相对于自己而言的编码区间(即本节点沿着哪个输出边可以到达该订阅者),然后将在自己的过滤条件表中加入相应的项.

我们采取如下思路来对订阅消息的转发进行优化:考虑到一个节点所在编码区间能够和其相邻的区间构成一个更大的编码区间,一个节点在处理订阅请求时,可以利用相邻区间的过滤条件来缩小订阅消息的发送范围.例如,设节点 213 的应用层执行操作 *subscribe*(f_1).在节点 213 的过滤条件表中,如果已有一个长度为 3 的编码前缀(如 210)设置了一个过滤条件 f_2,$f_1 \sqsubseteq f_2$,则网络中的各节点都已经知道在以“21”为前缀的区间中有节点对 E_{f2} 感兴趣,而 $E_{f1} \subseteq E_{f2}$,所以节点 213 只需发出订阅消息到路由表中编码前缀长度为 3 的各项(即 210,211,212),而无须发送订阅消息到路由表中编码前缀长度小于 3 的各项.该操作所导致的各节点的过滤条件表变化如图 3 所示.通过这种方式,我们可以大大减少订阅消息在网络中传播的数量.随着整个系统中订阅数量的增加,一个订阅消息所要到达的目标节点将会越来越少.

Node Id	Filter table
011	(2,f_2)
…	…
201	(21,f_2)
…	…
210	(210,f_2)
211	(210,f_2)
212	(210,f_2)
213	(210,f_2)
…	…

(a) Initial state: Peer 210 defined a filter f_2

(a) 初始时,节点 210 定义了 f_2

Node Id	Filter table
011	(2,f_2)
…	…
201	(21,f_2)
…	…
210	(210,f_2), (213,f_1)
211	(210,f_2), (213,f_1)
212	(210,f_2), (213,f_1)
213	(210,f_2), (213,f_1)
…	…

(b) Peer 213 executing *subscribe*(f_1), $f_1 \sqsubseteq f_2$

(b) 节点 213 执行 *subscribe*(f_1),$f_1 \sqsubseteq f_2$

Fig.3 Change of filter tables caused by a subscribing operation

图 3 订阅操作导致的过滤条件表的变化

3.4 Unsubscribe操作的处理

当某节点的应用层执行 *unsubscribe*(*filter*)操作时,该节点的事件通知层一方面要通知其他节点,自己对 E_{filter} 不再感兴趣;另一方面,还应考虑到该 *filter* 可能覆盖了相邻编码区间中的其他节点的过滤条件,因此要对这些被覆盖的过滤条件进行特别处理.消息转发的思路以及优化措施与 subscribe 操作类似,在此不再详述.

3.5 其他应考虑的事项

为了实现上述思路,还必须要解决其他一些问题,包括:

1. 消息顺序问题.在 P2P 网络中,一个节点向另一个节点发送的不同消息可能是沿着不同的路径到达的.因此,不能根据消息接收的顺序来判断消息发送的顺序.如果一个节点先后发出"订阅"消息和"取消订阅"消息,有可能"取消订阅"消息会先于"订阅"消息到达目的地,从而造成系统状态错误.IRBR 协议采用了一种类似于 Lamport 逻辑时钟[22]的机制来解决这个问题;

2. 并发正确性问题.当某节点在处理订阅操作时,它是根据自己的过滤条件表中所记录的相邻区间的订阅信息来缩小订阅消息的发送范围的.而在并发操作的情况下,相邻区间的订阅信息可能已经发生变化,从而使得本节点的决策变得不正确.为了解决这个问题,IRBR 协议要求当一个节点执行 *unsubscribe* 操作后,监控此后一段时间内到达的消息,并采取措施防止系统状态出现不一致;

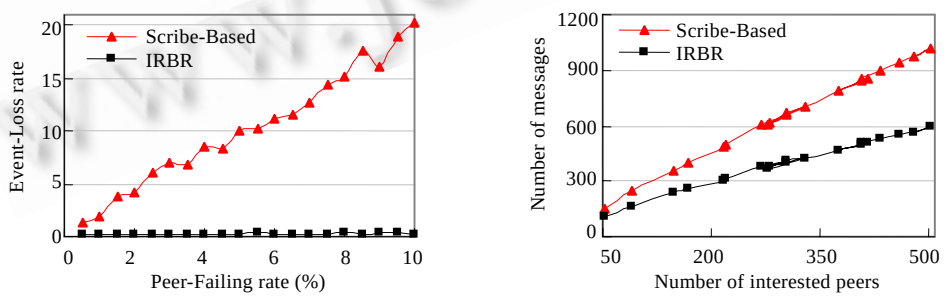
3. 自组织问题.在 P2P 网络中,节点可以自由地增加和退出.当节点加入网络中时,该节点的事件通知层能够根据其他节点的信息初始化自己的过滤条件表;当节点退出时,应尽量不影响系统中的消息的传播.IRBR 协议中提供了相应的机制,以支持节点的加入和退出.

4 性能评价

我们在 Pastry 的一个开放源码版本 FreePastry 1.3.2[23]上开发了一个原型系统,并对原型系统进行了模拟实验,以观测不同负载条件下 IRBR 协议的路由效率和容错能力.FreePastry 中提供了一个网络模拟程序,能够在一台机器上模拟大量的 P2P 网络节点.在我们的实验中,P2P 网络中节点数量为 1 000.

为了评价 IRBR 协议的路由效率,我们还实现了 Scribe-based 协议,以比较这两种协议在相同环境、相同负载条件下的表现.在对 Scribe-based 协议的实验中,假定网络中只有 1 个汇合点.

为了考察各协议的容错情况,我们让网络中的一部分节点同时失败,然后观测"事件丢失率",即对于一个被发布的事件,在当前所有对该事件感兴趣的活动节点中,有百分之多少的节点无法收到该事件.在图 4(a)中,X 轴表示有百分之多少的节点同时失效,Y 轴表示平均的丢失事件率.从图 4(a)可以看出,当节点失败率从 0.5%增加到 10%时,Scribe-based 协议的事件丢失率从 1.35%增加到 20.27%,而 IRBR 协议的事件丢失率仅从 0.11%增加到 0.28%.



(a) Fault-Tolerance capability (b) Routing efficiency for events

(a) 容错能力 (b) 事件路由效率

Fig.4 Comparison of fault-tolerance capability and efficiency of two protocols

图 4 两个协议的容错能力和效率比较

下面我们考虑各协议的事件转发效率.在图 4(b)中,X 轴表示对于一个给定的事件,有多少个节点对之感兴趣,Y 轴表示为了将该事件送到这些节点,所需要的多消息转发次数.从图中我们可以看出,IRBR 协议比 Scribe-based 协议具有更高的事件转发效率.

5 结 论

本文提出了一种面向结构化 P2P 网络的基于内容路由协议——IRBR 协议.它能够较为自然地与 P2P 网络的路由协议结合在一起,从而能够提供更高的可靠性保证,同时也具有更高的事件路由效率.该协议目前只适用于基于前缀的结构化 P2P 网络,包括 Pastry 和 Tapestry.我们下一步的工作将是研究如何扩展 IRBR,以使它也能适用于其他 P2P 网络.

References:

[1] Eugster PT, Felber PA, Guerraoui R, Kermarrec AM. The many faces of publish/subscribe. ACM Computing Surveys, 2003,35(2): 114−131.

[2] Carzaniga A, Rosenblum DS, Wolf AL. Design and evaluation of a wide-area event notification service. ACM Trans. on Computer Systems, 2001,19(3):332−383.

[3] Rowstron A, Druschel P. Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. In: Guerraoui R, ed. Proc. of the IFIP/ACM Int'l Middleware Conf. London: Springer-Verlag, 2001. 329−350.

[4] Zhao B, Kubiatowicz J, Joseph A. Tapestry: An infrastructure for fault-tolerant wide-area location and routing. Technical Report, No.UCB/CSD-01-1141, Berkeley: Computer Science Division, University of California, 2001.

[5] Stoica I, Morris R, Karger D, Kaashoek F, Balakrishnan H. Chord: A scalable peer-to-peer lookup service for Internet applications. In: Cruz R, Varghese G, eds. Proc. of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SigComm). New York: ACM Press, 2001. 149−160.

[6] Ratnasamy S, Francis P, Handley M, Karp R, Shenker S. A scalable content-addressable network. In: Cruz R, Varghese G, eds. Proc. of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SigComm). New York: ACM Press, 2001. 161−172.

[7] Dalal YK, Metcalfe R. Reverse path forwarding of broadcast packets. Communications of the ACM, 1978,21(12):1040−1048.

[8] Muhl G. Large-Scale content-based publish/subscribe systems [Ph.D. Thesis]. Germany: Darmstadt University of Technology, 2002.

[9] Cugola G, Nitto ED, Fuggetta A. The JEDI event-based infrastructure and its application to the development of the OPSS WFMS. IEEE Trans. on Software Engineering, 2001,27(9):827−850.

[10] Bhola S, Strom R, Bagchi S, Zhao Y, Auerbach J. Exactly-Once delivery in a content-based publish-subscribe system. In: Lala J, ed. Proc. of the Int'l Conf. on Dependable Systems and Networks (DSN 2002). Washington: IEEE Computer Society Press, 2002. 7−16.

[11] Snoeren AC, Conley K, Gifford DK. Mesh-Based content routing using XML. In: Marzullo K, ed. Proc. of the 18th ACM Symp. on Operating Systems Principles (SOSP). New York: ACM Press, 2001. 160−173.

[12] Chand R, Felber PA. A scalable protocol for content-based routing in overlay networks. In: Avresky D, ed. Proc. of the 2nd IEEE Int'l Symp. on Network Computing and Applications. Washington: IEEE Computer Society Press, 2003. 123−130.

[13] Cugola G, Picco GP, Murphy AL. Towards dynamic reconfiguration of distributed publish-subscribe middleware. In: Proc. of the 3rd Int'l Workshop on Software Engineering and Middleware. London: Springer-Verlag, 2002. 187−202.

[14] Shen Z, Tirthapura S. Self-Stabilizing routing in publish-subscribe systems. In: Carzaniga A, Fenkam P, eds. Proc. of the 3rd Int'l Workshop on Distributed Event-Based Systems (DEBS 2004). Washington: IEEE Computer Society Press, 2004. 92−97.

[15] Costa P, Migliavacca M, Picco GP, Cugola G. Epidemic algorithms for reliable content-based publish-subscribe: An evaluation. In: Liu M, Matsushita Y, eds. Proc. of the 24th Int'l Conf. on Distributed Computing Systems. Washington: IEEE Computer Society Press, 2004. 552−561.

[16] Castro M, Druschel P, Kermarrec AM, Rowstron A. SCRIBE: A large-scale and decentralised application-level multicast infrastructure. IEEE Journal on Selected Areas in Communications, 2002,20(8):100−110.

[17] Pietzuch P, Bacon J. Hermes: A distributed event-based middleware architecture. In: Wagner R, ed. Proc. of the 22nd International Conference on Distributed Computing Systems Workshops (ICDCSW 2002). Washington: IEEE Computer Society Press, 2002. 611−618.

[18] Tam D, Azimi R, Jacobsen HA. Building content-based publish/subscribe systems with distributed Hash tables. In: Proc. of the 1st Int'l Workshop on Databases, Information Systems, and Peer-to-Peer Computing (DBISP2P 2003). London: Springer-Verlag, 2003. 138−152.

[19] Terpstra WW, Behnel S, Fiege L, Zeidler A, Buchmann AP. A peer-to-peer approach to content-based publish/subscribe. In: Jacobsen HA, ed. Proc. of the 2nd Int'l Workshop on Distributed Event-Based Systems (DEBS 2003). New York: ACM Press, 2003.

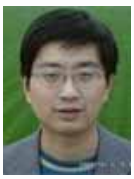
[20] Dabek F, Zhao B, Druschel P, Kubiatowicz J, Stoica I. Towards a common API for structured peer-to-peer overlays. In: Kaashoek MF, Stoica I, eds. Proc. of the 2nd Int'l Workshop on Peer-to-Peer Systems (IPTPS 2003). London: Springer-Verlag, 2003. 33−44.

[21] El-Ansary S, Alima LO, Brand P, Haridi S. Efficient broadcast in structured P2P networks. In: Kaashoek MF, Stoica I, eds. Proc. of the 2nd Int'l Workshop on Peer-to-Peer Systems (IPTPS 2003). London: Springer-Verlag, 2003. 304−314.

[22] Lamport L. Time, clocks, and the ordering of events in a distributed system. Communications of the ACM, 1978,21(7):558−565.

[23] Rice University. FreePastry project. 2004. http://freepastry.rice.edu/FreePastry.

汪锦岭(1974－),男,安徽庐江人,博士,主要研究领域为分布式计算,中间件技术.



李京(1966－),男,博士,研究员,博士生导师,主要研究领域为组合软件技术,分布式计算,移动计算.

金蓓弘(1967－),女,博士,副研究员,CCF高级会员,主要研究领域为分布式计算,软件工程技术.