面向广域网络的主动发布订阅系统的设计框架

高庆一 逯 鹏 林学练 王 斌

(英国布里斯托大学 计算机科学系, 布里斯托 BS8 1TR) (北京航空航天大学 计算机学院, 北京 100083)

摘 要: 事件空间划分技术改变了传统基于内容发布订阅系统的基本交互模式和相关实现策略,形成了主动发布订阅系统. 为了建立主动发布订阅系统的完整解决方案,在分析系统基本交互模式的基础上,提出了一个主动发布订阅系统的设计框架 ACTPS (Active Publish/subscribe System). 分为 6个设计模型: 事件模型,订阅模型,划分模型,构件模型,路由模型和匹配模型. 详细描述了各模型的关键特征及核心实现策略,以及各个模型之间的联系. ACTPS为一般的主动发布订阅系统设计和实现提供了参考模型,并为进一步深入的研究和分析建立了基础.

关 键 词: 广域网络; 分布式系统; 发布订阅; 设计框架

中图分类号: TP 393

文献标识码: A **文 章 编 号**: 1001-5965 (2007) 02-0238-05

Wide area oriented network active publish/subscribe system design framework

Gao Qingyi

(Department of Computer Science, University of Bristol, Bristol BS8 1TR, United Kingdom)

Lu Peng Lin Xuelian Wang Bin

(School of Computer Science and Technology, Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

Abstract: Event space partition technology changes the traditional basic interaction model of the content-based publish/subscribe system and its related implementation strategies, and enables a class of system which is named active publish/subscribe system. In order to establish the integrity of active publish/subscribe system solutions, based on the analysis of active publish/subscribe system basic interaction model, an active publish/subscribe system design framework ACTPS (active publish/subscribe system) was proposed. ACTPS comprised six design models which were event model, subscription model, partition model, component model, routing model and matching model. Detailed description of key features and the core of implementation strategy of these models, and the linkages among the various models were presented. ACTPS provides the general reference model for designing and realization of active publish/subscribe system. It is also a theory foundation for further in-depth research and analysis of publish/subscribe system.

Key words: wide area networks; distributed computer systems; publish/subscribe; design framework

面向广域网络的发布订阅系统实现了广域网络中异构应用实体之间异步、多对多和松耦合通信,成为当前学术界和商业界共同的研究热点[1]. 基于内容的 PS (Publish/Subscribe)系统是目前的研究重点,路由和匹配技术是其中的关键技术[2],基于内容匹配问题的本质是一种数据处理技术,已有大量的相关算法研究工作[3]. 基于内容路由的难点在于系统中的发布事件并没有唯一标识具体的接收者,需要通过代理在每次转发时的动态匹配确定下一跳的地址,从而进一步将

收稿日期: 2006-03-01

基金项目: 国家 863计划资助项目 (003AA119030); 国家自然科学基金资助项目 (90412011)

作者简介: 高庆一 (1979 -),男,北京人,博士生, qgao@compsci.bristol.ac.uk

事件分发到所有感兴趣的订阅者. 目前两种主要路由算法是泛洪法和匹配优先法[3]. 泛洪法是将事件或者订阅条件在整个代理网络中泛洪,并在客户端完成与订阅条件的匹配. 匹配优先法则是首先在代理网络中泛洪广告消息和订阅条件,两者的匹配结果在代理网络中形成路由状态,到达的事件基于与路由状态的匹配结果以及相关路由协议进行选择性转发. 在 Siena[4], Elvin, JEDI[5]等采用匹配优先路由算法的典型 PS系统中,分别使用了逆向路径转发,"淬火"机制和基于树结构的转发等方法优化路由状态的分布,达到优化路由的目的.

上述方法存在的问题在于泛洪操作限制了 PS系统在广域网络中的可伸缩性. 为了提高面向广域网络 PS系统的可伸缩性,以文献[6-7]为代表,提出了一种基于事件空间划分的方法实现 PS系统. 其基本思想是: 主动将事件内容的逻辑空间分为不同的区域,每个区域划分到一个代理,该代理完成相应区域内的事件发布订阅者之间的通信. 其优点是通过划分而不是泛洪在代理网络建立订阅状态,提高了系统可伸缩性和效率,减少了状态维护、更新和匹配的代价. 称基于该技术的系统为主动 PS系统.

但是主动 PS系统目前并没有一个完整的解决方案[6]. 本文通过对其基本交互过程和关键技术的分析,提出了一个主动 PS系统的设计框架 ACTPS(Active Publish/subscribe System). ACTPS将主动 PS系统分为 6个设计模型: 事件模型,订阅模型,划分模型,构件模型,路由模型和匹配模型. 建立了设计模型的基本概念,描述了其主要特征和实现策略,分析了系统设计模型之间的关系. ACTPS为主动 PS系统的设计和实现提供了基本参考模型,并建立了深入分析和研究的基础.

1　相关工作

目前关于 PS系统设计框架的主要是文献[8-9]提出的模型. 文献[8]从事件生命周期的角度,给出的 7个设计模型是目前多数系统的参考模型,并进行了一定的扩展. 但是这种框架没有考虑主动 PS系统的特殊交互过程,同时没有给出形式化的描述和定义. 文献[9]的模型是一种面向局域网络的基于事件的集成模型,已经不能满足广域网络的分布式特性的发展.

在典型系统中, Siena基于文献[8]的基本设计框架给出了原型实现,并给出了接口和操作语义的部分形式化描述,但是其基本的路由策略是逆向路径转发,需要通过泛洪分布路由状态. 此外,其中缺少对一般的设计框架的研究. Rebeca[10]基于 Siena的基本交互过程和算法给出了基于线性时序逻辑的语法,系统地描述一个 PS系统的动态特征,给出一个自稳定的 PS系统的正确性证明,其主要的目的在于形式推导,并没有给出完整的设计框架.

2　体系结构和基本交互分析

2.1　系统体系结构

面向广域网络的 PS系统体系结构可以表示为三元组 (Π, Θ, Σ),如图 1所示. 其中 $\Pi = \{P_1, P_2, \cdots, P_m\}$表示发布者(发布客户端)的集合, $\Sigma = \{S_1, S_2, \cdots, S_n\}$表示订阅者(订阅客户端)的集合, $\Theta = \{B_1, B_2, \cdots, B_p\}$表示事件代理的集合. 客户端能够通过代理网络中的任意代理接入 PS系统,该代理成为客户端的接入代理. 事件代理组成网络 $G = \langle \Theta, L \rangle$, L是应用层链接集合, G向系统客户端提供了面向广域网络的通知服务.

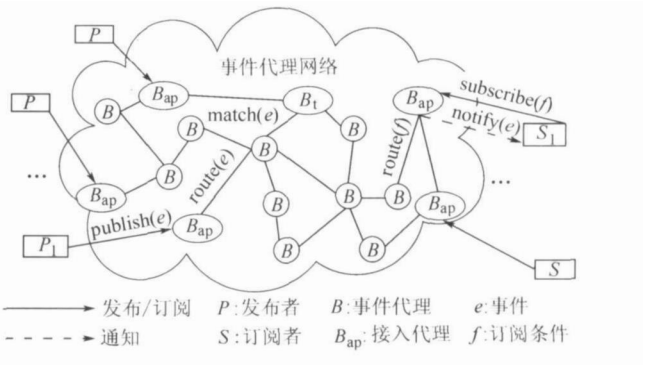

图 1　*PS*系统视图

2.2　*PS*系统的基本交互

面向广域网络的 PS系统的基本通信交互可以分解为以下过程,如图 1所示.

1) publish(*e*): 发布者 P_1通过接入代理 B_{ap}向 G发布事件 e;

2) subscribe(*f*): 订阅者 S_1通过 B_{ap}向 G订阅满足订阅条件 f的事件 e,其反向操作是 unsubscribe(*f*);

3) match(*e*): 代理判断 e的内容是否满足 f;

4) route(*e*): 代理 B基于 e与 f匹配的结果判定路由,最终将 e转发到其订阅者的接入代理 B_{ap};

5) notify(*e*): B_{ap}将通知 e(本文中事件和通知具有相同的内容)以推或拉的方式交付订阅者,完成 PS通信.

2.3　主动 PS系统的基本交互

在主动 PS系统中,事件内容的 N个属性分

别代表事件空间 Ω空间中的一维. 则 $e \in \Omega$成为 Ω中的一个点, f是半开半闭区间,例如 $(100 < X_{price} \leq 150)$,构成的超矩形区域 $z \subseteq \Omega$. 基于上述假设,主动 PS系统的基本交互,除客户端的 publish(e)和 subscribe(f)操作之外,分为 5个步骤,如图 2所示.

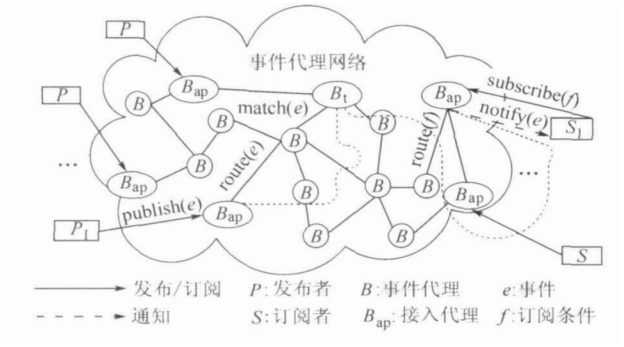

图 2 ACTPS交互示意图

1) 事件空间划分和委派: 将 Ω划分为多个区域并分别委派 $\Theta = \{B_1, B_2, \cdots, B_p\}$,如图 3所示. 相应代理负责区域内 e和 f的匹配和通知,成为其目标代理 B_t;

2) 订阅条件路由: 订阅者的接入代理 B_{ap}将 f单播路由到其目标代理 B_t,并复制;

3) 事件路由: 发布者的接入代理 B_{ap}解析 e的路由目标代理 B_t,并通过单播路由 e到 B_t;

4) 事件匹配: B_t基于当前维护的 f和 e,匹配 e的订阅者子集 S_{ub};

5) 通知路由: B_t作为 e的源端,以 S_{ub}的接入代理集合 $A(S_{ub})$作为目的端,通过单播或多播将 e分发到其所有的订阅者.

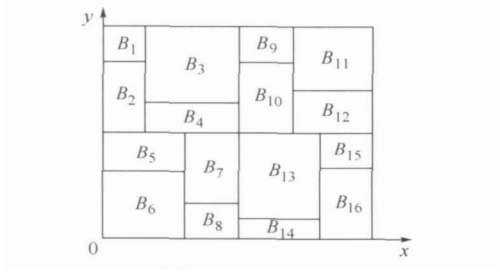

图 3 事件空间的划分示意图

主动发布订阅系统特点在于: ①事件空间划分使 e和 f具有了明确的目的地址 B_t; ②使用单播完成 e和 f的路由; ③主动的事件空间划分和委派取消了建立路由状态的泛洪操作,并且能够进行动态的调整和更新区域的规模; ④划分限制了代理事件匹配操作的范围和数量,提高了基于内容匹配的效率; ⑤使用通知路由的过程完成通知交付,使通知操作更加灵活. 主动 PS系统与其它路由方法实现策略的比较如表 1所示.

表 1 3种基于内容路由实现策略的比较

路由阶段/方法	泛洪法	匹配优先法	划分法
订阅路由	泛洪/NO	泛洪	单播
事件路由	NO/泛洪	广播/多播	单播
通知路由	NO	NO	单播/多播
事件匹配	客户端匹配	多代理匹配	目标代理匹配

注: "NO"表示没有设计相关的路由算法; "/"表示两种选择.

3 ACTPS的设计框架

基于上述分析,将主动 PS系统抽象为 6个设计模型. 这些设计模型及其关系形成主动 PS系统的设计框架,称为 ACTPS.

定义 1 ACTPS = (ET, SN, CT, ES, RT, MH)

其中, ET(Event)表示事件模型,描述了系统中事件的表达方式; SN(Subscription)表示订阅模型,描述了事件的订阅模式; ES(Event Space)表示空间划分的模型,描述了事件空间的划分和委派策略; CT(Component)表示构件模型,描述了系统中的通信构件及其关系; RT(Route)表示路由模型,描述了系统中的路由策略; MH(Match)表示匹配模型,描述了系统的匹配策略.

ACTPS涵盖了主动 PS系统的基本组成,同时基于不同的应用场景可以扩展其它设计模型,例如安全性和移动性等.

3.1 事件模型

在基于内容的 PS系统中,事件通常使用多个任意的"属性=值"的表达式组成的"事件头"对"事件体"(内容)进行描述. 这种随机的属性名/值对很难建立结构化的事件空间,不支持基于区间的订阅机制,降低了事件的可表达能力. 因此, ACTPS ET采用了类似关系数据库模式的方式描述事件内容,并基于此模式构造结构化的事件空间.

定义 2 事件模型

事件模型 ET = (Type, attribute, value). 其中, Type是数据类型集合,且是具有连续区间的数据类型; attribute是属性名集合; value是属性值域. 事件 $e = (v_1, \cdots, v_i, \cdots, v_n)$是该模式的一个数据.

基于上述事件模型,事件的 N个属性分别代表事件空间 Ω中的一维,则事件 $e = (v_1, \cdots, v_i, \cdots, v_n)$成为 Ω中的点 $e \in \Omega$. 非连续区间属性,例如名称,可以通过索引方式将其线性化为连续的区间. 事件模型 ET是主动 PS系统设计的基础性问题,其它模型均依赖于 ET.

3.2 订阅模型

订阅模型解决的问题是订阅者如何描述对事件的订阅兴趣,即订阅条件.

定义 3　订阅模型

订阅模型 SN = (Type, attribute, OP, value). 其中, Type是属性的类型; attribute是属性名称; value是属性值域; OP是操作符号,例如"<","=","≤"等. 订阅条件 $f = (\sigma_1 \wedge \sigma_2 \wedge \cdots \wedge \sigma_i \wedge \cdots \wedge \sigma_n)$是订阅模型的一个实例, σ_i是一个属性的原子约束,限定了一个半开半闭区间.

ACTPS SN的特征在于: f构成了 Ω中的一个超矩形区域 $z \subseteq \Omega$,例如 $(100 < x \leq 150, 1 < y \leq 5)$. e匹配 f等价于点 e属于区域 z

3.3 划分模型

划分模型 ES解决的问题是划分事件空间,并将划分区域委派到系统中的代理.

定义 4　划分模型

$U = \{u_1, \cdots, u_i, \cdots, u_n\}$是事件空间 Ω的一个划分,其中 $u_i \subseteq \Omega$, $\bigcup_{i=1}^{n} u_i = \Omega$, $u_i \cap u_j = \Phi$, $i \neq j$,且满足映射 $g: U \rightarrow \Theta$,称 g为划分模型,表示为 ES.

文献[6]已经给出 Ω的多种划分策略. 在具体的设计实现过程中,自组织的划分模型是一个重要的研究问题. 即当 Θ发生代理的加入、退出、过载和失效等动态变化,如何改变 U和 g以保证系统的负载均衡和可靠性.

基于划分模型 ES, e, f, U之间的关系表示为映射 R和 I

$$R: \Omega \rightarrow U$$

即 $\forall e (e \in \Omega), R(e) \subseteq U$. g基于 R解析出 e的目标代理 B_t.

假定 S是系统当前所有 f的集合:

$$I: S \rightarrow 2^U$$

即 $\forall f (f \in S), I(f) \subseteq U$. g基于 I解析出 f的目标代理 B_t.

3.4 构件模型

如图 1所示的 PS系统体系结构,构件模型 CT的包括事件代理和客户端两种类型的构件.

定义 5　事件代理模型

事件代理模型 CT = $(C_p, C_s, N_b, R_s, E_p)$. 其中, C_p是接入的发布者的集合; C_s是接入的订阅者的集合; N_b是相邻代理的集合; R_s是路由状态; E_p是划分到代理的区域的集合.

CT的特征是: ①通过 E_p的动态更新和调整,能够改变相关 e, f的目的地址,并改变代理的静态负载; ②为了避免维护全局路由信息,提高系统路由效率和可伸缩性, CT维护了相邻代理的集合 N_b; ③R_s是属于 E_p的 f组成的过滤器,用于通知 e的路由.

客户端构件通常在资源和软硬件环境异构设备上部署,基于同步或异步方式接入 B_{ap}获得 G的通知服务. 为了提高系统的灵活性,其功能设计抽象比较简单.

定义 6　发布者 (发布客户端)

ACTPS中的发布者可以是两个元素的抽象: 即 $P = (A_p, A_s)$. 其中, A_p是其接入代理; A_s是发布者当前的链接状态.

定义 7　订阅者 (订阅客户端)

ACTPS中的订阅客户端可以是两个元素的抽象: $S = (A_p, A_s)$. 此时, A_s是订阅者当前的链接状态.

3.5 路由模型

路由模型 RT反映了主动 PS系统基本交互过程中的 3种路由过程,即事件路由、订阅路由、通知路由,表示为集合 RT = $\{S_{route}, E_{route}, N_{route}\}$.

S_{route}的初始假设是订阅者的接入代理能够实现映射 $I: S \rightarrow 2^U$和 g. f可能与多个区域相交而解析生成多个目标代理. S_{route}路由项可以抽象为元组 (E_p, N_b),即路由判定只与代理的区域属性和相邻代理集合相关.

S_{route}的使用单播代替了泛洪操作,并在 B_t复制 f,完成路由状态的分布.

E_{route}, N_{route}是事件在发布者与订阅者之间分发的两个阶段. 可以抽象为映射

$$C: \Omega \times \Theta \rightarrow \text{Boolean}$$

即代理基于当前状态判断是否接收到达的事件 e其中 E_{route}表示为

$$(\forall e \in \Omega, \exists B_t \in \Sigma, \exists u_i \in U)(I(B_t) = u_i) \rightarrow C(e, B_t) = \text{true}$$

E_{route}使用 B_t作为明确的目的地址,避免了泛洪或者每跳事件匹配判断路由的降低系统效率的操作.

基于事件空间的最短距离的贪婪路由是一种典型的 E_{route}单播算法,如算法 1所示. 算法将代理 B_i, B_j间的距离抽象为 Ω中区域 u_i, u_j之间的路由,并使用贪婪算法实现路由.

算法 1　事件路由算法

输入: 事件 e

```
if e.target∈B.E_p then {
    receive(e);
}else{
    Nexthop = Lookup B.N_b for Nearest dis-
```

tance to e. target;

send e to Nexthop;

}

算法首先判断事件 e的目标是否为本地,是则接收;否则,查找并计算 N_b. E_p距 e target最短的代理转发.算法也可以应用于 S_{route}和 N_{route}的单播路由,并在此基础上进行多种策略的优化.

N_{route}是以 B_t为源端,将 e分发到其订阅者的接入代理的过程.表示为

$$\forall e \in S_{ub},\ \forall B \in A(S_{ub}) \rightarrow C(e,B) = \text{true}$$

其中 $A(S_{ub})$表示匹配 e的订阅者集合 S_{ub}的接入代理集合.

设 t为代理所维护 f的平均值,则 N_{route}基于参数 t在多播效率和构造多播树的系统开销之间取得权衡. $|B_i. R_s| < t$则采用类 E_{route}的单播路由; $|B_i. R_s| > t$,则构造以当前代理 B_t为根的多播树 NHST(B_t)指向下一跳,NHST(B_t)算法的实现是多播的关键研究问题.

3.6 匹配模型

MH与传统基于内容的匹配均可以抽象为映射

$$M: \Omega \times S \rightarrow 2^{\Theta}$$

ACTPS的匹配模型 MH通过划分 $U = \{u_1, \cdots, u_n\}$将代理维护的 f数量平均缩小为 $1/n$,同时,相关路由算法减少了每跳的匹配操作.

$\omega, \mu \subseteq E_s$分别表示属于当前区域的 e, f的集合. $S \subseteq A(\mu)$是匹配 e的接入代理的集合.

e, f的匹配关系如定义 8所示.

定义 8 匹配关系

如果 $e = (v_1, \cdots, v_i, \cdots v_n)$中每个值 v_i均满足 $f = (\sigma_1, \cdots, \sigma_i, \cdots \sigma_n)$中的原子约束 σ_i,则称 e匹配 f,记为 $e \subseteq f$.

定理 1 $\forall f_1, f_2$,如果 $e \subseteq f_1 \rightarrow e \subseteq f_2$,则 $f_1 \subseteq f_2$,即 $\forall e, (e \subseteq f_1 \rightarrow e \subseteq f_2) \rightarrow f_1 \subseteq f_2$.

证明略.

推论 1 $\forall f_1, f_2, f_1 \subseteq f_2 \wedge f_2 \subseteq f_1 \rightarrow f_1 = f_2$.

证明略.

主动 PS系统匹配关系的特征是:设 z是 f在事件空间 Ω中的区域,则

$$e \subseteq f \Leftrightarrow e \in z$$

即点和区域的包含关系与匹配关系等价.

4 结 束 语

本文详细描述了一个面向广域网络的主动 PS系统的基本设计框架 ACTPS.它是主动 PS系统的设计和实现的基础性的工作.对于深入分析和设计主动 PS系统具有重要的作用.后续的工作是以 ACTPS为基本设计框架,研究主动 PS系统的实现策略和关键算法,并基于自主开发的 SOAP(Simple Object Access Protocol)消息中间件 ACT-XLink设计和实现一个基于内容的主动 PS信息交换中间件系统.

参考文献 (References)

[1] Eugster Patrick Th, Felber P, Guerraoui R, et al. The many faces of publish/subscribe[J]. ACM Computing Surveys, 2003, 35(2): 114 - 131

[2] 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[J].软件学报,2005,16(2): 251 - 259
Xue Tao, Feng Boqin. Research on routing algorithm and self-configuration in content-based publish-subscribe system [J]. Journal of Software, 2005, 16(2): 251 - 259 (in Chinese)

[3] Banavar G, Chandra T, Mukherjee B, et al. An efficient multicast protocol for content-based publish-subscribe systems[C]// International Conference on Distributed Computing Systems. Piscataway, NJ: IEEE Computer Society, 1999: 262 - 272

[4] Carzaniga A, Rosenblum D S, Wolf A L. Design and evaluation of a wide-area event notification service [J]. ACM Trans. on Computer Systems, 2001, 19(3): 332 - 383

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

[6] Riabov A, Liu Zhen, Wolf J L, et al. Clustering algorithms for content-based publication/subscription systems [C]// Dahlia Malkhi. Proceedings 22nd International Conference on Distributed Computing Systems (ICDCS'02). Berlin: Springer-Verlag, 2002: 133 - 142

[7] Wang Y M, Qiu L, Achlioptas D, et al. Subscription partitioning and routing in content-based publish/subscribe networks [C]// Dahlia Malkhi. 16th International Symposium on Distributed Computing. Berlin: Springer-Verlag, 2002: 28 - 30

[8] Rosenblum D S, Wolf A L. A design framework for internet-scale event observation and notification [C]//6th European Software Engineering Conference. Berlin: Springer-Verlag, 1997: 344 - 360

[9] Barrett D J, Clarke L A, Tarr P L, et al. A framework for event-based software integration [J]. ACM Transactions on Software Engineering and Methodology, 1996, 5(4): 378 - 421

[10] Mühl G. Large-scale content-based publish/subscribe systems [D]. Germany Darmstadt: Electrical Engineering and Information Technology, Darmstadt University of Technology, 2002

(责任编辑:张 婧)