SAN-EBON：一种基于结构化对等网的 P2P 工作流系统节点定位网络

高 磊 曾广周

（山东大学计算机科学与技术学院 济南 250101）

摘 要 基于 P2P 的工作流系统符合工作流去中心化的发展趋势. 目前，P2P 工作流系统主要是基于非结构化 P2P 网络构建的. 然而，非结构化 P2P 网络提供的泛洪或基于超级节点的中心化发现策略和中心化的负载分配机制无法满足大型 P2P 工作流系统在动态环境下的需求. 因此，在基于非结构化 P2P 网络构建的工作流系统中，节点发现和任务负载均衡成为制约系统性能的关键因素. 文中提出一种新的基于结构化 P2P 网络的工作流系统节点定位网络——SAN-EBON. 该系统采用分层逐步求精的节点发现策略，外层在服务聚类的基础上首次在工作流系统中引入服务定位网络组织服务联盟，构建一种新的多层结构化 P2P 网络 SAN，实现服务的快速发现；内层构建一种新的负载均衡网络 EBON，使用基于随机图的增强算法实现服务联盟内部实时的去中心化负载均衡，与 SAN 结合，从而达到提高发现效率和精度、降低通信带宽的目的.

关键词 P2P 工作流系统；结构化 P2P 网络；服务联盟；服务寻址网络；平衡覆盖网

中图法分类号 TP311 **DOI 号**：10.3724/SP.J.1016.2010.02353

SAN-EBON: A Novel P2P Location Network Based on Structured Overlay Network for P2P Workflow System

GAO Lei ZENG Guang-Zhou

(*School of Computer Science & Technology, Shandong University, Jinan* 250101)

Abstract The P2P workflow systems which are so far based on unstructured P2P network meet the development trend of workflow systems. Both discovery of system peers and decentralized load balancing are key factors having a great impact on the performance of this kind of systems, especially the running time of workflow instance. Unstructured P2P systems have exhibited common weakness such as flood routing and centralized load distribution which limit system application in large-scale and dynamic environments. This paper presents a novel location network based on structured P2P network named SAN-EBON for P2P workflow system, which uses a hierarchical step-wise refinement strategy. Be the first to structure network of service alliances using SAN which is an innovative structured P2P network based on the services cluster in workflow system, and encode the information about each node's available computational resources in structure of an enhanced random graph in the alliance, which is named EBON, to achieve decentralized real-time load balancing. The combination of SAN and EBON raises the efficiency and precision of peers location and lower communication bandwidth and network fluctuation.

Keywords P2P workflow system; structured P2P network; services alliance; services addressed network; balanced overlay networks

收稿日期：2008-08-19；最终修改稿收到日期：2010-01-30. 本课题得到国家自然科学基金(60573169)资助.**高 磊**，男，1979 年生，博士研究生，主要研究方向为移动计算及其应用技术、对等计算、CSCW、工作流管理. E-mail：aicscw2006@yahoo.cn. **曾广周**(通信作者)，男，1947 年生，教授，博士生导师，主要研究领域包括智能计算与移动计算、CSCW、工作流管理等. E-mail：gzzeng@sdu.edu.cn.

1 引 言

基于 P2P 的工作流系统是一种新的工作流管理系统,符合下一代工作流管理系统去中心化的发展趋势[1]. 对等计算模式的引入使工作流节点在运行时阶段直接通信,尤其实现了控制流在各节点中的分布,更好地反映了工作流的分布特性和群组协作的社会属性.

目前对工作流的研究主要集中在工作流模型及其执行语义和时序约束等方面[2],对于运行时阶段的流程控制主要依赖于工作流中心引擎. 但 P2P 工作流系统更加强调分布式的控制结构和对等点之间的协同效率,因此高效的节点定位机制,即服务请求消息在参与节点之间的路由,成为影响这类系统性能的重要因素. 文献[3]通过业务流程组织服务工作流,根据全局服务信息搜索最优执行路径,但在缺少全局信息的对等环境中无法有效展开. 文献[4-5]均采用分布式控制结构,其中文献[4]使用一种静态路由结构,在建模时阶段静态绑定了任务前驱和后继节点的消息队列标识;文献[5]则采用类似于 UDDI 的中心路由结构实现对执行节点的动态查询;静态路由无法适应动态变化的应用环境,而中心化的动态定位机制容易形成路由瓶颈. SwinDeW 及其扩展系统[6-9]是基于 JXTA[10]构造的非结构化 P2P 工作流系统,采用了动态非中心化的路由机制避免了上述问题. 但该系统通过泛洪实现节点定位引入较大的通信负载,且定位效率和准确性不高;通过群组内节点之间的协商机制实现任务分配要求每个节点必须在线维护本群组内所有节点状态,在动态环境中容易引发网络波动.

针对上述问题,本文采用分层逐步求精的路由策略,按服务对节点进行聚类,首次将结构化 P2P 网络引入工作流系统,提出一种新的基于结构化对等网的 P2P 工作流节点定位系统 SAN-EBON. 结构化 P2P 网络采用基于 DHT(Distributed Hash Table)的分布式发现和路由算法[11],有着良好的定位准确性、可扩展性和自组织性. 基于这一特性,本系统外层结合 CAN 算法[12]和服务聚类构建了一种新的结构化 P2P 网络 SAN(Service Addressed Network),该网络可视作多层结构化 P2P 网络的叠加,实现了对服务联盟的快速精确定位;在联盟中,内层将工作负载映射为随机图拓扑,采用基于随机漫步的 EBON(Enhanced Balanced Overlay Network)网络,可以较小的连接度数在联盟节点中实现去中心化的实时任务负载分配,最终精确定位执行节点. 两者结合提高了路由精度和效率,降低了定位最优执行节点时的通信带宽,达到缩短流程实例在 P2P 工作流系统中非人为因素执行时间的目的,适用于节点数目多且动态变化范围大的系统.

2 SAN-EBON 系统结构

系统主要由工作流模型映射(Workflow Modeling Mapping,WMM)、流程模型发布(Workflow Modeling Publishing,WMP)和 SAN-EBON 网络三部分构成(如图 1 所示).

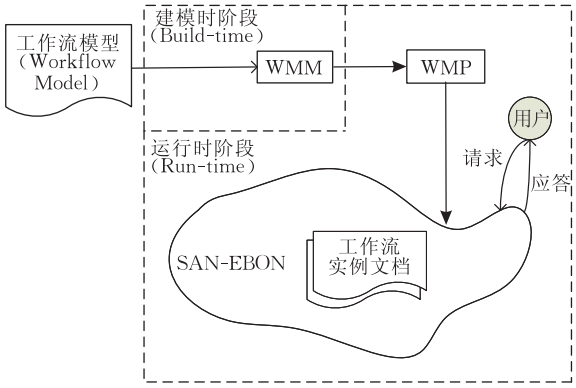

图 1 SAN-EBON 系统结构

WMM 在建模时阶段进行,将已有的业务流程映射为 SAN 网络中的逻辑节点,通过流程发布与相应的执行节点绑定,从而实现流程模型在 SAN 网络中的部署.

定义 1. 业务流程可表示为一个有向图 $D=\langle T,R\rangle$,其中 T 对应有向图中的顶点集,表示流程中所有任务(包括原子任务和复合任务)的集合,称为任务集;$R\subseteq T\times T$ 对应有向图中的边集,为定义在任务集上的逻辑关系.

定义 2. 任务 $t_i\in T, i=1,2,\cdots,n$ 由二元组表示($Task\_id$,$Task\_des$),其中,$Task\_des=(input, output, pre\_condition, post\_condition, task\_semantic, resource, QoS)$ 为任务描述;$input$,$output$,$pre\_condition$,$post\_condition$ 分别为输入数据、输出数据、入口约束条件、出口约束条件;$task\_semantic$ 为任务语义描述(语义描述语言和模型在此不作讨论);$resource$ 为任务所需的计算资源,QoS 为完成该任务必须提供的服务质量. $Task\_id=(tp\_id, san\_id)$ 为任务标识,tp_id 为任务模型标识,在流程模型中唯一标识一个任务;$san\_id=(x_1,x_2,\cdots,x_d)$

是 d 维笛卡儿坐标，在 SAN 逻辑坐标空间中唯一标识一个任务以及部署了该任务的节点，称为 SAN 坐标.

定义 3.　$SanHash$：$Task\_des \rightarrow san\_id$ 是定义在 T 上的映射，对于任意一个 $Task\_des$，唯一确定一个 san_id 与之对应，该映射通过一致性散列函数[11]实现，记作 $(x_1, x_2, \cdots, x_d) = SanHash(Task\_des)$.

定义 4.　Td：$R \rightarrow R_{lo}$ 是定义在 R 上的映射，R_{lo} 是当前任务的局部控制关系，由二元组 $(Tset_{\text{pre}}, Tset_{\text{post}})$ 表示，其中，$Tset_{\text{pre}} = \{(san\_id_j, L_{\text{pre}}) \mid t(san\_id_j) \in T \wedge \langle t(san\_id_j), t(san\_id_i)\rangle \in R, i, j = 1, 2, \cdots, n\}$ 是任务 $t(san\_id_i)$ 的前驱任务集，L_{pre} 是前驱任务与当前任务之间的逻辑关系；$Tset_{\text{post}} = \{(san\_id_j, L_{\text{post}}) \mid t(san\_id_j) \in T \wedge \langle t(san\_id_i), t(san\_id_j)\rangle \in R, i, j \in 1, 2, \cdots, n\}$ 是任务 $t(san\_id_i)$ 的后继任务集，L_{post} 是后继任务与当前任务之间的逻辑关系.

通过 $SanHash$ 和 Td 映射，业务流程中的原子任务被打包为相应的流程服务并映射到 SAN 空间中的逻辑节点上，即定义 5.

定义 5.　SAN 流程服务 $S_i, i = 1, 2, \cdots, n$ 由二元组 (t_i, R_{lo}) 表示.

由上述定义，每个 SAN 逻辑节点由 san_id 唯一标识，即 SAN 坐标空间中的不同逻辑节点对应了工作流模型中的不同流程服务，从而实现了工作流模型到 SAN 坐标空间的映射.

WMP 将经过散列后的业务流程发布到相应的执行节点，使流程服务与具体的执行节点绑定，从而将物理节点映射到 SAN 空间. 该过程在流程运行时阶段进行，采用自顶向下由物理节点主动拉动的方式实现. 流程模型的定义及映射后的流程服务可由系统的注册服务器统一管理，模型发布则在运行阶段由注册节点主动发起.

1. 注册节点从注册服务器下载希望承担的任务描述 $Task\text{-}des$；

2. 与本地服务描述进行匹配，评估本地资源，若达到 QoS 要求，则在本地部署该任务（包括加载相应服务和资源）；

3. 向注册服务器注册本地部署任务和加载服务；

4. 注册服务器将部署任务的 san_id 和 R_{lo} 绑定到注册节点；

5. 注册节点根据 san_id 加入系统，确定其在 SAN-EBON 网络中的位置，根据 R_{lo} 明确需要接收什么数据并将自身的处理结果传递给什么任务，实现流程模型在所有参与节点中的数据和控制分布.

SAN-EBON 网络（见第 3 节）是系统数据流和控制流运行的自组织环境，构成系统的基础结构. 系统执行节点的自组织、有效定位、加入、离开以及任务实例的负载分配均通过该网络实现. 该网络由 SAN 和 EBON 两层网络组成，其中 SAN 以一种新的结构化 P2P 网络组织外层定位，较已有的 P2P 网络（结构化和非结构化）具有更高的路由效率和更小的通信负载；EBON 以随机图映射节点负载状况，实现了内层的去中心化实时负载均衡，较已有的 P2P 工作流系统（均采用中心化任务调度）避免了网络负载波动和路由瓶颈.

3　SAN-EBON 网络结构

3.1　SAN 网络

可扩展内容寻址网络（CAN）在结构化 P2P 网络中引入了多维标识符空间的概念，与单维空间中线性遍历相比，路由效率得到了极大的提高. CAN 将覆盖网映射到一个 d 维笛卡尔坐标空间，每个节点的位置由其自身 IP 经散列后得到的 d 维坐标决定，坐标空间被不同节点分割成互不相交的区域，且每个节点占有一个区域. 但是 CAN 网络中每个区仅包含一个节点，在节点变化频繁的动态环境中，由于区域的合并拆分将引入较大的拓扑维护开销；同时像所有基于 DHT 的结构化 P2P 网络一样，CAN 网络中的节点上存储的是指向其它节点数据源的索引，因此 CAN 中的目标节点并不是数据源的存储节点，还需进行二次映射. 针对上述问题，在 CAN 的基础上，我们首先用服务对节点进行聚类，形成服务联盟，将服务联盟映射到多层 CAN 坐标空间中，再通过 EBON 网络将多层逻辑空间压缩到一起，从而构成 SAN（Service Addressed Network）网络.

定义 6.　承担相同任务的节点构成的集合称为服务联盟，一个服务联盟对外提供同一类服务.

定义 7.　任务的 SAN 坐标唯一标识坐标空间中的一个逻辑节点，称为集结点，集结点唯一标识一个服务联盟以及联盟中承担该任务的所有物理节点.

定义 8.　集结点将坐标空间划分为不同的区域，设空间维数是 d，若两个集结点坐标中有 $d-1$ 维是相等的，剩余的一维相邻，则称两个集结点相邻，相应区域为邻接区域. 集结点相邻是其对应联盟中节点相邻的必要条件.

由上述定义，SAN 网络是由各集结点共同构成的逻辑坐标空间，两个相邻集结点沿相邻维度划分坐标区域（具体划分算法在 4.2 节给出），每个集结点占有一个区域，由 san_id 唯一标识. 根据定义 2、6，因为 san_id 在逻辑空间中唯一标识一个任务及

标识承担该任务的所有物理节点，所以承担相同任务的物理节点具有相同的 SAN 坐标，将被映射为同一个集结点，从而将服务联盟映像到 SAN 空间中. 这样在一个坐标区域中实现了分配多个物理节点，使得 SAN 网络中的逻辑节点(集结点)数目等于流程模型中的原子任务数，相对于 CAN 等典型的结构化 P2P 网络以及非结构化 P2P 网络极大减少了坐标空间中的节点数目，从而有效提高了路由效率. 同时在一个 SAN 区域中分配多个物理节点提高了区域的有效性，即只要有一个有效节点存在，区域拓扑就保持不变，从而避免了 CAN 网络中由于节点频繁加入、离开给系统带来的区域分裂、合并等巨大的结构维护开销；当联盟中的节点状态发生变化时，SAN 的逻辑拓扑保持稳定，使其更加适用于动态环境. 但是在每个 SAN 区域中并不存在超级节点(Super Node，SN)，域内节点不依赖联盟内的其它节点独立路由，即每个节点独立与相邻区域的一个或多个节点形成域间邻居关系，由此形成一层类 CAN 路由子网，并依赖其所在子网转发路由请求，因此可以将 SAN-EBON 网络看作多层 CAN 网络的叠加，每一条路由消息只需通过任意子网转发到目标区域，而各层子网通过区域内的 EBON 网络连结在一起. 这一结构与 CAN 中的过载区域(Overloading Coordinate Zones)[12] 相比有效避免了中心节点带来的区域路由瓶颈及由多节点分配带来的区域维护开销. 根据上述结构，每个物理节点在 SAN 网络中的映像位置是由其所承担任务的*san_id*决定和唯一标识的，因此在 SAN 中，对服务的请求即是对能够提供该服务的目标区域节点的查找，无须经二次映像，从而将业务流程映射到 SAN 坐标空间中.

图 2 给出了一个二维 SAN 网络映射的例子，图中空心点代表物理节点，实心点代表集结点，不同形状表示加载了不同任务，承担相同任务的物理节点由 EBON 网络连结为服务联盟. 由图中可见共有 3 个服务联盟，分别用圆、三角、正方形不同形状表示. 3 个集结点将二维坐标空间分为 3 个互相邻接的区域，每个区域中分别包含 2、3、4 个物理节点，相同形状的空心点之间的细虚线表示该联盟内的 EBON 网络. 由图中可见，左侧的 SAN 网络可被分解为右侧的两个子网，处于同一子网中的物理节点逻辑上为域间邻接节点，并通过 EBON 网络联结在一起.

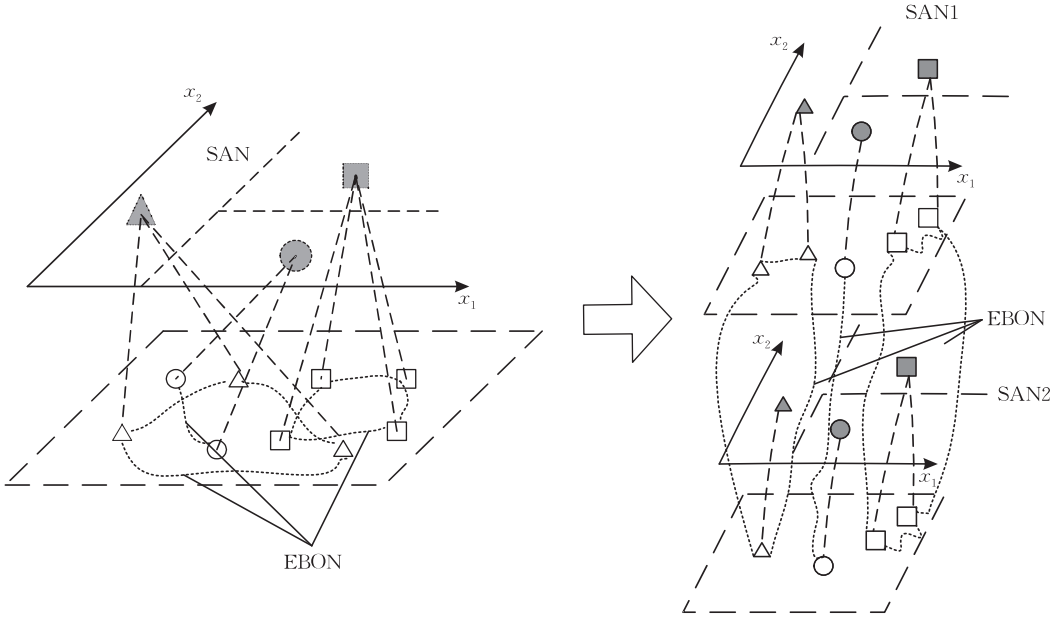

图 2 二维 SAN 网络

3.2 EBON 网络

EBON(Enhanced Balanced Overlay Network)网络用于在服务联盟内均衡任务负载，实现联盟内最优执行节点的定位. BON 网络[13] 是一个有向随机图；图中的每个顶点代表一个物理节点，每条有向边代表一个任务实例负载，节点权值 w_i 表示结点 i 所能承担实例负载的个数，节点的入度 $in\_degree(i)$ 表示节点 i 空闲实例负载的个数，这样节点的负载状况被映射为随机图的拓扑. 每个节点的入度数反比于节点的任务负载，通过随机采样选择空闲负载最大的节点接受任务. 随机采样中请求消息的随机游走遵循由节点 i 向其相邻节点 j 的转移概率为

$$p_{ij}=1/out\_degree(i) \tag{1}$$

节点 k 被采样的概率为

$$p_k = in\_degree(i) \Big/ \sum_{i \in V} in\_degree(i) \quad (2)$$

由此可见，入度数越大说明空闲资源越多，被采样的概率也就越大，从而通过在图上的随机游走指引任务向轻负载节点流动，且负载的概率分布正比于节点的空闲资源. 这样就以较小的连接度数实现了节点之间的去中心化动态负载分配，消除了中心化负载分配带来的性能瓶颈，减弱了因全连接结构引起的网络波动.

但 BON 算法假设所有节点的处理能力是相同的，采样时只考虑空闲资源，而实际系统中参与节点的处理能力和能够提供的服务质量都存在较大差异，因此我们在 BON 基础上提出 EBON 算法，同时考虑节点的处理能力、空闲资源和服务质量的差异.

(1) 考虑节点处理能力和拥有资源的差异，具有较多空闲资源的节点其处理队列也可能较长，无法第一时间处理当前任务，因此应以节点中空闲资源所占比例作为度量. 定义如下.

定义 9.　以单个任务实例为单位，称节点 i 所能承担的最大任务实例数为节点 i 的权值，记作 w_i；设当前节点 i 处理队列长度为 λ_i，则

$$\rho_i = 1 - \lambda_i / w_i \quad (3)$$

为节点 i 的占空比；ρ_i 越大表明当前节点有效计算资源越多.

(2) 考虑承担相同任务的节点其所能提供服务的 QoS 可能差异很大，根据文献[3]提出的工作流服务质量调度算法，QoS 调度要考虑的主要因素有 $QoS(cap)$、$QoS(tim)$、$QoS(co)$、$QoS(ava)$、$QoS(rep)$，通常要求计算能力尽量大，执行时间尽量短，开销费用尽量小，服务可用性尽量高，服务可靠性尽量大. 根据上述优化目标，并将极小约束转变为极大约束，有如下定义.

定义 10.　设节点 i 的 QoS 权值向量为 $[\omega_1, \omega_2, \cdots, \omega_5]$，则

$$\alpha_i = [\omega_1, \omega_2 \cdots, \omega_5] \cdot [QoS(cap), (1-QoS(tim)), (1-QoS(co)), QoS(ava), QoS(rep)]^{\mathrm{T}} \quad (4)$$

为节点 i 的 QoS 因子；α_i 越大表明节点提供服务质量越高.

定义 11.　设节点 i 的占空比为 ρ_i，QoS 因子为 α_i，则

$$\Phi_i = \alpha_i \times \rho_i \quad (5)$$

为节点 i 的采样值.

根据上述定义，EBON 算法在随机采样时记录并优先选择随机路径上具有最大采样值的节点分配任务，即 $Target = \arg\max_{i \in V} \Phi_i$，其中 $Target$ 为选定的目标节点. 图 3 给出了一个通过 EBON 网络分配负载的例子. 图中每个节点的标号为该节点的权值，节点旁的小圆为分配给该节点的任务负载，图 3(b) 中的虚线为新负载分配时随机采样的路径，图 3(c) 为负载分配后新的拓扑结构. 设 n_i 为图中权值为 w_i 的节点，$\alpha_i = 1$，按顺时针方向 $i = \{6,2,3,4,1,5,4\}$，则采样路径上各节点的采样值依次为 $\Phi_{n_1} = 0$，$\Phi_{n_4} = 1/4$，$\Phi_{n_6} = 1/3$，$\Phi_{n_2} = 1/2$，$\Phi_{n_3} = 2/3$，$\Phi_{n_4} = 1$，根据对图中随机漫步路径上各节点采样值的比较选择具有最大采样值的节点，目标节点为 $\Phi_{n_4} = 1$ 的 n_4，因此将新任务分配给 n_4，同时剪除一条指向 n_4 的有向边 $\langle n_3, n_4 \rangle$.

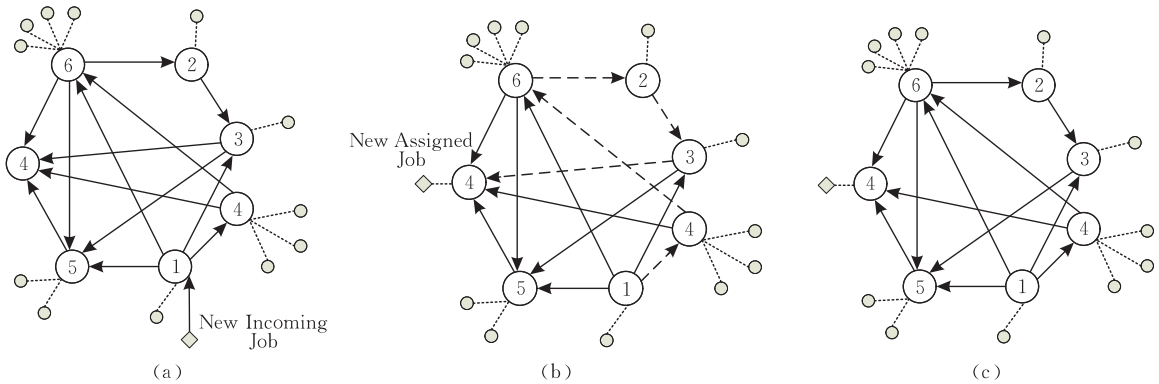

图 3　EBON 网络负载分配实例

4　SAN-EBON 网络功能

4.1　SAN-EBON 路由算法

SAN-EBON 路由是指对服务请求消息的路由，即节点完成本地任务实例后定位下一步最优执行节点的过程. 这个过程在 SAN-EBON 网络的基础上分服务联盟定位和区域负载均衡两个阶段完成.

4.1.1　服务联盟定位

服务联盟定位实现对目标域的快速定位，遵循

SAN 坐标空间中从源坐标到目的坐标的一条直线转发路由请求. 每条 SAN 路由消息都包含目的域坐标,如果本地节点不拥有包含这些坐标的区域,则它将消息转发到拥有离目标域最近坐标的邻居节点. 最近转发邻居节点的确定需要考虑两个因素:确定最近转发邻域和选择最小延时节点. 最近转发邻域具有到目标域的最短直线距离,由 SAN 空间中本地域到目标域的欧式距离给出:设当前节点第 i 个邻接区域 Σ_i 的 SAN 坐标为 $(x_1^i, x_2^i, \cdots, x_d^i)$, $i \in [1,d]$,其中 d 是当前 SAN 空间的维数;目标域的 SAN 坐标为 $(x_1^s, x_2^s, \cdots, x_d^s)$,则

$$\Sigma_{\min} = \arg\min_{i\in[1,d]} \sqrt{(x_1^i - x_1^s)^2 + (x_2^i - x_2^s)^2 + \cdots + (x_d^i - x_d^s)^2} \tag{6}$$

其中 $\Sigma_{\min}$ 为最近转发邻域. 根据 SAN-EBON 结构,当前节点在 $\Sigma_{\min}$ 中可拥有多个邻居节点,并记录到每个 $\Sigma_{\min}$ 邻居节点的延时,从而选择延时最小的节点作为最近转发邻居节点. 这一过程直到找到目标区域为止,而目标区域的转发节点作为区域接入节点发起区域负载均衡.

每个 SAN-EBON 节点均维护域间路由表以支持上述域间路由过程. 域间路由表记录 SAN 网络中邻接区域的邻居节点信息,每个表目对应一个域间邻居节点,可描述为一个四元组(节点 *san_id*,节点 IP,节点区域范围,节点延迟),若节点离开网络,相应表目将被删除. 域间路由表的生成基于 EBON 网络上的随机漫步,对每一个邻接区域,采样随机路径上节点的域间路由表,选择拥有该区域邻居最多的节点,均分其在该区域的邻居节点,形成本地对该区域的相邻节点表目.

定理 1. 设 SAN 空间维数是 d, m_i 为第 i 个邻接区域 Σ_i 的节点数目,若当前区域的节点数为 n,则该区域中节点的域间路由表空间复杂度上限为 $O\left(\sum_{i=1}^{2d} m_i / 2^{\lfloor \log_2 n \rfloor}\right)$.

证明. 对 Σ_i 邻居表的裂变过程可表示为一棵完全二叉树(如图 4 所示).

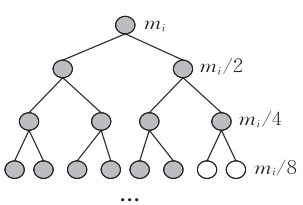

图 4 域间路由表裂变树

因为树中的每个节点对应一个新的 Σ_i 邻居表(灰色节点为已扩展节点,白色节点为尚未加入的节点),当节点加入域时,将分割父节点 Σ_i 邻居表生成两个子节点,其 Σ_i 邻居数为父节点的 1/2;

设根节点的层数为 0,则第 h 层节点的 Σ_i 邻居数为 $m_i/2^h$;

因此,当 $n = 2^h$ 时,节点 Σ_i 邻居数为 $m_i/2^h$;当 $2^{h-1} < n < 2^h$ 时,节点 Σ_i 邻居数 $\in [m_i/2^h, m_i/2^{h-1}]$.

又由于当前节点的最小层次为 $\lfloor \log_2 n \rfloor$,

所以 Σ_i 区域路由表复杂性上限为 $O(m_i/2^{\lfloor \log_2 n \rfloor})$;

又因在 d 维空间中最多有 $2d$ 个邻接区域,

因此,域间路由表复杂性上限为 $O\left(\sum_{i=1}^{2d} m_i / 2^{\lfloor \log_2 n \rfloor}\right)$. 证毕.

根据 CAN 算法的期望路由跳数为 $O(dn^{1/d})$(其中 n 为系统节点数)[12]可知 SAN 网络的域间期望路由复杂度为 $O(dN^{1/d})$,其中 d 为 SAN 空间维数,N 为系统中的联盟数. 由于服务联盟数 N 等于流程模型中的任务数,远远小于系统中的物理节点数 n,从而较 CAN 网络和泛洪的 $O(n^2)$ 有效降低了域间路由复杂度;由式 $\partial dN^{1/d}/\partial d = 0$ 可得,当 $d = \ln N$ 时,有极小值 $O((\ln N)^{1/\ln N})$. 设服务请求消息带宽为 β,则域间路由带宽为 $O(dN^{1/d}\beta)$,明显低于泛洪且其定位精度为 100%.

4.1.2 区域负载均衡

区域负载均衡基于 EBON 网络实现服务联盟内的去中心化负载调度,确定任务最终执行节点. 根据式(1),通过在 EBON 网络上的随机漫步转发路由请求,并由式(3)~(5)在采样路径上确定最优执行节点分配任务负载. 每个节点除维护域间路由表外同时维护域内路由表以支持上述域内路由过程.

域内路由表记录 EBON 网络中的邻居节点信息(两个域内节点之间存在有向边即存在邻居关系),每个表目对应一个域内邻居节点,可分为入度节点和出度节点两类,用四元组(节点 IP,节点权值,节点负载,连接方向)表示. 当节点被分配新任务时,将通过随机漫步选择一个入度邻居,删除其相应表目;被选中节点删除本地指向当前节点的出度表目. 当节点完成当前任务时,通过随机漫步选择一个不相邻节点,添加为入度邻居;被选中节点添加本地指向当前节点的出度表目. 根据规则随机图特性[14]及文献[13]实验表明,当 BON 网络最小入度数不小于 4 时,其对应的无向随机图为强连通图. 因此可

确定节点 i 域内路由表入度复杂性 O_i 在正常负载下有 $O_i \propto w_i$；在过载情况下入度应为 0，但为保持网络的连通性，有最小入度复杂性 $O_i^{\min}=4$.

根据 BON 网络期望直径为 $O(\log m)$（其中 m 为网络节点数）[13] 可知域内期望路由复杂度为 $O(\log \overline{M})$，其中 $\overline{M}$ 为各服务联盟节点数目的期望值.

4.1.3　算法描述

基于以上描述，SAN-EBON 路由算法如下.

1. 初始化. 若节点 $R_{to}=\varnothing$，即节点无后继任务集，转步 6；否则节点根据 R_{to} 将后继任务 *san_id*、本地 IP 和 *san_id* 封装为服务请求消息 *Request*；

2. 若 *Request* 中的目的坐标与本地坐标相同或落在本地区域范围内，则转步 4；否则转步 3；

3. 查找域间路由表，根据式(6)计算邻接区域与目标域之间的空间距离，选择距离最小的区域作为转发区域，记为 Π_{post}，在 Π_{post} 邻居节点中选择延迟最小的节点转发 *Request*，转步 2；

4. 当前节点为区域接入节点，在本联盟内通过 EBON 算法，根据式(1)～(5)找到最优执行节点，转发 *Request*；

5. 检查当前节点 R_{to}，若源 *san_id* $\in$ 本地 $Tset_{\text{pre}}$，则通过源 IP 与 *Request* 发起节点建立数据连接，接受新任务，同时根据 EBON 算法更新本地相关节点的域内路由表，转步 1；

6. 任务分支结束.

4.2　SAN-EBON 网络动态维护

SAN-EBON 网络维护主要处理节点加入离开时 SAN-EBON 坐标空间的差分、合并以及相关节点路由表的刷新.

新节点加入系统时分两种情况：

(1) 节点加入时对应集结点已存在，根据 SAN 路由算法找到其坐标所在区域接入节点，通过 EBON 算法构建自身的域间和域内路由表，并通知其邻居节点更新自身的域间和域内路由表.

(2) 节点是其服务联盟的第 1 个加入点. 首先根据 SAN 路由算法找到其坐标所在区域接入节点，由于坐标不同于接入节点，因此需要差分接入节点所在区域，此时选择两个节点坐标相差最大的维度，沿该维度平分接入节点所辖区域，两者建立域间邻居关系，接入节点更新自身的区域范围和域间路由表；同时通知其域间和域内邻居新节点的 *san_id*、所辖区域范围、节点 IP 以及自己的新区域范围；收到通知的域间和域内邻居节点根据通知更新自身的域间路由表，同时通知其他相邻的域内邻居并与新节点建立邻居关系；新节点根据收到的邻接请求建立自身的域间路由表.

节点离开系统时也分为两种情况：

(1) 离开节点不是联盟中的最后一个节点，只需通知路由表中的域间和域内邻居节点更新相应的域间和域内路由表，同时调整其出度边所指节点的拓扑结构，即该出度边所指节点通过随机漫步在联盟内采样新节点生成新的入度边代替离开节点的出度边.

(2) 离开节点是联盟中的最后一个节点，此时撤销节点对其区域的管辖，进行区域合并. SAN 坐标区域的重分割可按一棵二分树[15] 展开，树中每个节点代表一个集结点，两个子节点分割父节点所辖区域. 区域合并可分两种情况讨论：① 撤销区域的兄弟节点（即邻接区域）为叶节点，即两者共同分割了父节点区域，则由兄弟节点接管撤销区域；离开节点将自身域间路由表移交兄弟节点，据此信息，兄弟节点在域内更新自身区域范围和域间路由表，同时通知其域间邻居节点更新本地的区域范围和域间路由表. ② 撤销区域的兄弟节点为中间节点，则称其父节点所辖区域为裂变区域，裂变区域的邻接区域为非裂变区域. 当前区域撤销后，由以兄弟节点为根的子树重新划分裂变区域，划分过程从子树的根部开始逐层向下进行. 相应的，离开节点将自身的域间路由表移交给子树中的邻接区域，并通过区域相邻传递给裂变区域中的所有子区域，各区域节点根据子树结构自顶向下重新计算自身的区域范围，更新域间路由表，并通知邻接的非裂变区域节点更新域间路由表.

从以上讨论可以看出，当网络动态维护涉及区域的差分、合并时牵涉节点较多，运算代价较大，但根据 SAN-EBON 结构，服务联盟中只要存在一个节点，就不需要进行区域调整，在系统运行的大部分时间内只需处理第一种情况，网络结构保持相对稳定，因此可以较小的维护代价适应节点状态变化频繁的环境.

4.3　SAN-EBON 路由实例

文献[6]中的学生注册服务将在较短时间段内接受大量请求，从而对系统运行效率有较高要求. 我们构造 SAN-EBON 网络以支持该业务流程，并与 SwinDeW 系统在路由效率和通信负载方面针对该实例进行了对比. 学生注册服务可表示为图 5 所示的流程[6]，其中略去了与路由无关的业务描述，着重讨论其节点定位过程.

任务描述（只给出名称）、SAN 坐标和局部逻辑关系如表 1 所示. 表中信息在节点注册前通过工作流模型映射过程在建模时阶段生成，并保存在注册服务器中，节点注册后绑定到执行节点，并在执行时阶段控制节点的局部逻辑关系.

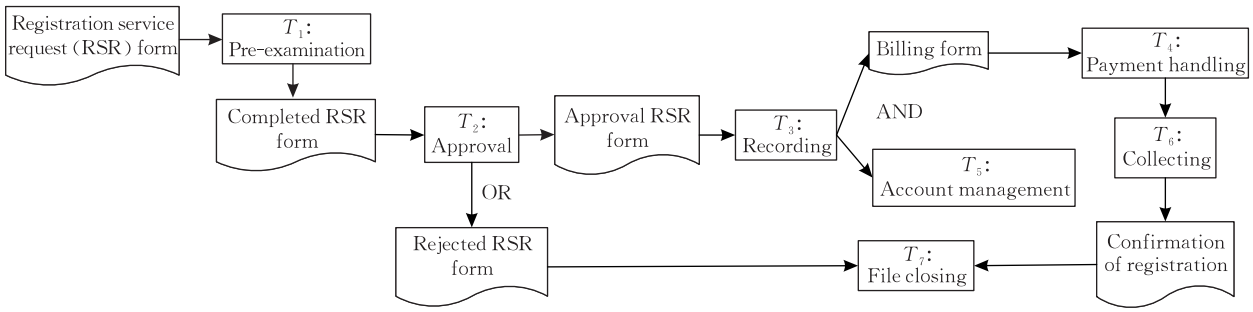

图 5 学生注册服务业务流程图

表 1 学生注册服务流程映射

	Task_des	*san_id*= *SanHash*(*Task_des*)	R_{lo}	
			前驱任务集 $Tset_{pre}$	后继任务集 $Tset_{post}$
T_1	Preexamination	(5,17)	—	((11,3),straight)
T_2	Approval	(11,3)	((5,17),straight)	((10,18),Approved RSR form), ((21,13),Rejected RSR form)
T_3	Recording	(10,18)	((11,3),Approved RSR form)	((16,20),AND),((4,5),AND)
T_4	Payment handing	(16,20)	((10,18),AND)	((15,14),straight)
T_5	Account management	(4,5)	((10,18),AND)	—
T_6	Collecting	(15,14)	((16,20),straight)	((21,13),straight)
T_7	File closing	(21,13)	((15,14),straight),((11,3), Rejected RSR form)	—

任务发布后各节点构成的 SAN-EBON 网络(此处映射为二维空间)如图 6 所示,其中实心圆点表示相应服务联盟对应的集结点,空心圆点表示服务联盟中的物理节点,域间相邻节点用虚线连接. 由图可见坐标空间分为 7 个区域,每个区域中包含多个承担相同任务的物理节点,如 T_1、T_2、T_5 所辖区域分别包含 4 个、5 个、2 个节点.

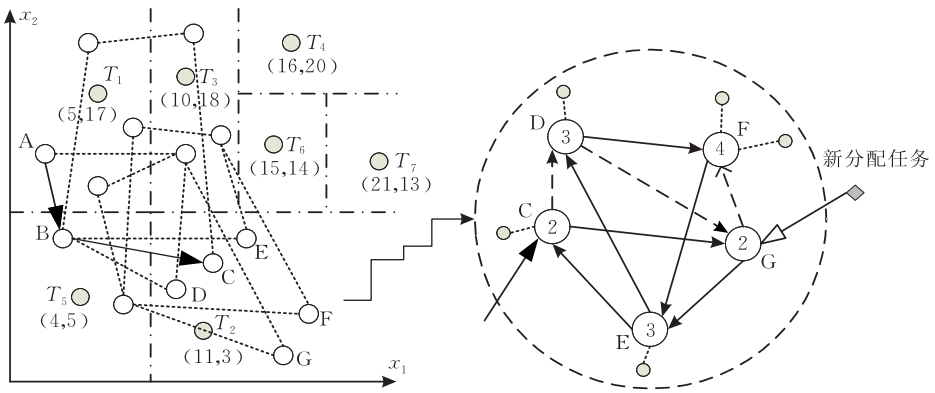

图 6 学生注册系统 SAN-EBON 网络拓扑

表 2～5 分别给出了节点 A、B、C 的域间路由表和节点 C 的域内路由表,其它节点路由表有相同的形式,其中省略了邻居节点的 IP 地址(用节点标识代替),同时省略了与本例转发路径无关的节点标识,并假设到各邻居节点的延迟均为 1(此参数需要通过心跳信息由节点实时刷新).

表 2 节点 A 域间路由表

序号	节点标识	SAN 坐标	区域范围	延迟
1	B	(4,5)	$x\in[0,8],y\in[0,10]$	1
2	—	(10,18)	$x\in[8,13],y\in[10,\infty)$	1

表 3 节点 B 域间路由表

序号	节点标识	SAN 坐标	区域范围	延迟
1	A	(5,17)	$x\in[0,8],y\in[10,\infty)$	1
2	—	(5,17)	$x\in[0,8],y\in[10,\infty)$	1
3	C	(11,3)	$x\in[8,\infty),y\in[0,10]$	1
4	D	(11,3)	$x\in[8,\infty),y\in[0,10]$	1
5	E	(11,3)	$x\in[8,\infty),y\in[0,10]$	1

表 4 节点 C 域间路由表

序号	节点标识	SAN 坐标	区域范围	延迟
1	B	(4,5)	$x\in[0,8],y\in[0,10]$	1
2	—	(10,18)	$x\in[8,13],y\in[10,\infty)$	1

表 5　节点 C 域内路由表

序号	节点标识	节点权值	节点负载	连接方向
1	E	3	1	入
2	D	3	1	出
3	G	2	0	出

节点 A 完成本地任务后将按下述步骤定位下一步执行节点：(1) 根据 $Tset_{post}$ 可知后继任务坐标为(11,3)，将其加入查询消息；(2) 查询本地域间路由表，发现(11,3)在邻居节点中没有命中且不在邻接区域内；(3) 计算邻接区域 T_3(区域坐标为(10,18))，T_5(区域坐标为(4,5))与(11,3)的空间距离分别为 15.03 和 7.28，选择空间距离最小邻域 T_5 为转发区域，在该区域内节点 A 只有一个域间邻居节点 B，则 B 为转发节点转发查询消息；(4) 节点 B 收到消息后查询本地域间路由表，命中目标节点 C、D、E；(5) 选择延迟最小节点，当延迟相同时按均匀概率随机选取目标节点转发查询消息，此处选 C；(6) 节点 C 作为目标区域的接入节点查询本地域内路由表，根据 EBON 算法选择随机路径上采样值最大的节点(图 6 右图为区域 T_2 的局部放大，其中虚线给出了域内随机路径)G(计算方法同图 3 实例)；(7) 节点 G 为最终执行节点，并与节点 A 直接建立连接(图 6 中实心箭头标出了域间转发路径，空心箭头标出了新分配的任务)接受相关任务数据.

由上述实例可见，在服务联盟内，任务分配通过 EBON 网络实现了去中心化. 节点的负载状况通过在 EBON 网络上的随机漫步采样得到，而不需要由中心节点统一管理，从而每个节点只需维护与其所能承担的负载数量成正比的邻居数. 如 T_2 区域中(图 6 中右图)节点的最大域内度数为 4，最小为 2，而传统的中心化负载分配需要域内节点实时维护本区域所有节点，则每个节点的邻居数为 5. 并且当域内节点数目增大时，中心化结构的域内邻居数将随 N 的增大而增大，其复杂度为 $O(N)$，而 EBON 网络的域内邻居数仍保持稳定，即若 T_2 中的节点数增大到 10000，节点 C,D,E,F,G 的域内最大度数仍保持不变，而中心化结构每个节点则需维护 10000 个域内节点. 因此在节点数目较大且动态变化频繁的系统中，EBON 的优势比较明显.

对于服务联盟间的服务发现效率，基于多层结构化对等网的 SAN 网络较 SwinDeW 系统中的洪泛策略有了明显的提高. 定义服务发现请求在到达目标区域前跳转的次数为服务请求路由步数，记为 H_{ij}，表示由区域 T_i 向区域 T_j 跳转的步数，即加载任务 T_i 的节点完成任务后在网络中搜索加载后继任务 T_j 的节点的步数. 同时定义在路由过程中，到达目标节点前服务请求传递到的节点数目为扩展节点数，记为 N_{ij}，表示由区域区域 T_i 向区域 T_j 跳转过程中除起始和目标节点外扩展的节点数. 例如，在上述实例中，由区域 T_1 向区域 T_2 的路由路径有节点 A→B→C，则 $H_{12}=2$，$N_{12}=1$. 基于上述定义和实例，我们比较了 SAN 和 SwinDeW 对应于学生注册系统的路由步数和扩展节点数，列于表 6 和表 7 中，其中 SwinDeW 系统的结构见文献[6].

表 6　SAN-EBON 与 SwinDeW 服务请求路由步数比较

	H_{12}	H_{23}	H_{27}	H_{34}	H_{35}	H_{46}	H_{67}	$E(H)$
SAN-EBON	2	1	1	1	2	1	1	1.286
SwinDeW	2	2	1	6	4	6	1	3.143

表 7　SAN-EBON 与 SwinDeW 扩展节点数比较

	N_{12}	N_{23}	N_{35}	$E(N)$
SAN-EBON	1	0	1	0.667
SwinDeW	7	6	8	7

上表中 $E(H)$ 和 $E(N)$ 分别表示路由步数和扩展节点数的平均值，其中表 6 中列出了实例中所有可能路径的路由步数，表 7 中仅列出了与实例中相对应的 T_1，T_2，T_3，和 T_5 区域的扩展节点数，但已具代表性. 从上述比较结果可以看出，SAN 网络的路由步数和扩展节点数明显优于 SwinDeW 系统. 并且，随着系统中节点数目的增长，SwinDeW 系统的路由步数和扩展节点数将按节点数目的幂率增长(4.1 节中有具体分析)，而 SAN 网络的路由步数仅与工作流模型中的任务数有关，只要工作流模型不发生变化，其路由步数不随节点数目的增长而增大，而扩展节点数等于路由步数减 1，所以也保持稳定. 因此在节点数较大的系统中 SAN 网络具有明显优的路由效率且引入较小的通信负载.

5　结束语

本文提出一种新的基于结构化 P2P 网络的 P2P 工作流系统节点定位网络 SAN-EBON. 讨论了其网络结构、路由算法以及系统的动态维护机制. 主要贡献有：(1) 以任务关系对参与节点进行聚类，较基于角色的方法使联盟边缘更加清晰；(2) 首次在工作流系统中引入结构化 P2P 网络组织服务联盟，提出 SAN 网络，相比于非结构化 P2P 网络提供了更高的服务发现效率和更小的通信负载；(3) 以节点承载服务标识该节点，实现了所查即所得；(4) 在联盟内部基于 BON 网络，考虑节点的异质性改进其采样机制，提出了 EBON 算法，实现了去中心化

实时定位最优执行节点.

SAN-EBON 网络在控制流上为相对封闭的自组织运行环境,运行时阶段正常的业务流程无需任何中心服务器介入.注册服务器在系统运行时阶段提供主动性监控,即只有当系统管理者需要了解系统参数及流程实例或者工作位置运行情况时,由注册服务器主动向工作节点查询(其相关协议,数据结构及策略属相对独立的主题,将在后续工作中单独讨论,本文不再展开描述).对于系统异常我们将其分为结构性和非结构性两种.结构性异常影响系统连通性,导致系统产生多个强连通块,此时,被独立的工作节点将通过注册服务器作为新节点重新加入系统,新节点加入实现过程见 4.2 节.非结构性异常不影响系统的连通性和控制流路由,此时异常节点的邻居节点将按异常节点退出自组织修复系统,节点退出实现过程见 4.2 节.在后续工作中我们将考虑利用工作流模型信息构造结构化对等网,以更小的通信负载获得更高的节点定位效率,进一步缩短系统实例的执行时间.

参 考 文 献

[1] Aberer K, Hauswirth M. Peer-to-peer information systems: Concepts and models, state-of-the-art, and future systems//Proceedings of the 9th ACM SIGSOFTSymp, Foundation Software Engineering(FSE-9). Vienna, Austria, 2001: 326-327

[2] Zeng Wei, Yan Bao-Ping. Survey on workflow module. Application Research of Computers, 2005, 5: 11-22(in Chinese)
(曾炜, 阎保平. 工作流模型研究综述. 计算机应用研究, 2005, 5: 11-22)

[3] Hu Chun-Hua, Wu Min, Liu Guo-Ping, Xu De-Zhi. An approach to constructing Web service workflow based on business spanning graph. Journal of Software, 2007, 18(8): 1870-1882(in Chinese)
(胡春华, 吴敏, 刘国平, 徐德智. 一种基于业务生成图的 Web 服务工作流构造方法. 软件学报, 2007, 18(8): 1870-1882)

[4] Alonso G, Mohan C. Exotica/FMQM: A persistent message-based architecture for distributed workflow management//Proceedings of the IFIP WG8.1 Working Conference on Information Systems for Decentralized Organizations. Trondheim, 1995: 1-17

[5] Georgios John Fakas, Bill Karakostas. A Peer to Peer (P2P) architecture for dynamic workflow management. Information and Technology, 2003, 46(6): 423-431

[6] Yan J, Yang Y, Raikundalia G K. SwinDeW—A P2P-based decentralized workflow management system. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2006, 36(5): 922-935

[7] Shen J, Yan J, Yang Y. SwinDeW-S: Extending P2P workflow systems for adaptive composite Web services//Proceedings of the 2006 Australian Software Engineering Conference (ASWEC 2006). Sydney, Australia, 2006: 61-69

[8] Yan Jun, Yang Yun, Raikundalia G K. Critical issues in extending P2P-based SwinDeW system for incomplete process support//Proceedings of the 8th International Conference on Computer Supported Cooperative Work in Design. International Academic Publishers, 2004: 312-317

[9] Shen J, Yang Y, Yan J. Adapting P2P based decentralised Workflow system SwinDeW-S with Web service profile support//Proceedings of the 9th International Conference on Computer Supported Cooperative Work in Design. Coventry University School of Mathematical and Information Sciences, 2005: 535-540

[10] Oaks S, Traversat B, Gong L. JXTA Technology Handbook. Beijing: Publishing House of Tsinghua University, 2004

[11] David Karger, Eric Lehman, Tom Leighton et al. Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the World Wide Web//Proceedings of the 29th Annual ACM Symposium on Theory of Computing. El Paso, Texas, United States, 1997: 654-663

[12] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp. A scalable content-addressable network//Proceedings of the SIGCOMM'01. San Diego, California, USA, 2001: 161-171

[13] Bridgewater J S A, Boykin P O, Roychowdhury V P. Balanced overlay networks (BON): An overlay technology for decentralized load balancing. IEEE Transactions on Parallel and Distributed Systems, 2007, 18(8): 1122-1133

[14] Rucinski A, Wormald N. Random graph processes with degree restrictions. Combinatorics, Probability and Computing, 1992, 1(2): 169-180

[15] Shaffer Clifford A. A Practical Introduction to Data Structures and Algorithm Analysis. Beijing: Publishing House of Electronics Industry, 2002

GAO Lei, born in 1979, Ph. D. candidate. His research interests include mobile computing, peer-to-peer computing, CSCW and workflow technology.

ZENG Guang-Zhou, born in 1947, professor, Ph. D. supervisor. His research interests include CSCW, intelligent computing, mobile computing and workflow technology.

Background

The heart of the decentralized workflow systems based on peer-to-peer (P2P) infrastructure is a formalized paradigm of coordination works in which work activities distributed in the network terminals are organized to provide a unified service. From this perspective, an effective coordination framework routing work instance (WI) in the proper sequence in a decentralized distributed environment is the key element restricting the performance of P2P workflow systems, that is, the distributed running path optimization problem. As we know, the running cost mainly consists of the discovery time and execution time in P2P environment, corresponding to the peer discovery and workload balancing problems. At present, the most P2P workflow systems are constructed based on unstructured P2P network, which provide a flood routing for peer discovery and centralized load scheduling for workload balancing. These approaches just realize the basic control functions with higher extra-bandwidth, lower precision of peers location and frequent network fluctuation, instead of the running time optimization.

In this paper, an innovative structured P2P network named SAN-EBON are constructed to provide an optimized running path in the decentralized dynamic environments. The main contributions are optimizing the service alliances discovery time and precision by constructing a novel superimposed structured P2P network (SAN), realizing decentralized load balancing in the alliance by encode load information in structure of an enhanced random graph (EBON), and realizing full self-organization in P2P workflow systems. The combination of SAN and EBON raises the efficiency and precision of peers location, lower communication bandwidth and network fluctuation.

This work is supported by the National Natural Science Foundation of China under grant No. 160573169 and the Shandong Province Project under grant No. 1031110123 for migrating workflow research and cooperative product commerce application in mobile computing paradigm. Within a P2P workflow system, the traveling path is consisting of work places which represent enterprises or organizations, and the participations allocated tasks must be efficiently coordinated for the optimized running time.