HiMCAN:一种新型的基于 DHT 的 P2P 内容寻址网络[1]

谢瑶，洪佩琳，李津生

（中国科学技术大学电子工程与信息科学系 信息网路实验室 安徽 合肥 230027）

摘要：基于 DHT（Distributed Hash Table）进行内容检索的 P2P 网络存在逻辑网络和物理网络不匹配，查询路由实际网络链路时延大等问题。在分析总结已有改进算法的基础上，本文提出改进 CAN 算法：HiMCAN，层次结合 CAN。该系统按物理网络距离把节点划分为多个组，组内独立建立 CAN，使用 HD 模型和 FN 模型参数以及结合算法将各组节点连接起来，组成全局意义的 HiMCAN，实现查询路由的本地性，优化查询路由参数，适合在广域范围实现结构型的 P2P 网络应用。

关键词：P2P，DHT，层次化，查询路由算法

中图分类号：TP393.04

HiMCAN: a Novel DHT Based P2P CAN Network

XIE Yao, LI Jin-sheng, HONG Pei-lin

(Dept. of Electronic Engineering and Information Science, Infonet Lab, USTC, Hefei 230027, China)

Abstract: DHT (Distributed Hash Table) based peer-to-peer network has the problem of logical network and physical network mismatching. By analyzing existed improved DHT networks, we propose a novel improved CAN: HiMCAN (Hierarchical Merging Content Addressable Network), to divide nodes into groups according to proximity, construct CAN intra-group, then using HD model, FN model and merging algorithm to link all subCANs into global HiMCAN, to gain benefits o routing locality. HiMCAN is suitable to construct large area file-sharing P2P system.

Key words: P2P, DHT, hierarchical, query routing algorithm

1. 引言

在过去的几年，对等网络(peer-to-peer network，P2P 网络)成为网络通信领域关注的热门话题，是下一代网络体系结构的重要组成。P2P 网络构建了一种完全分布式的网络结构，每个节点的地位对等、既可充当服务器为其它节点服务，也可充当客户机消费其它节点提供的服务[1]，打破了传统的 C/S(客户机/服务器)集中模式的限制；以充分利用网络边缘资源、带宽，有效均衡负载。

P2P 网络主要分为结构型 P2P 和非结构型 P2P，后者本质上是基于 DHT(Distributed Hash Table，哈希散列表)的内容寻址网络(下文中简称为 DHT)，由于其定位内容高效率，是如今研究的热点。DHT 网络中内容主要以文件的形式存在，为了定位文件，把文件名或关键字等文件属性信息抽象表示为 key，把文件本身或存储该文件的节点的 IP 地址抽象表示为 value，每个文件有一一对应的<key,value>二元对。现在研究最多的几种 DHT 系统：CAN[2], Chord[3], Pastry[4]等 DHT 系统都是对 key 进行哈希映射运算，将所有的 key 的哈希值（后文通称 key）形成一个 key 空间哈希表；每个节点被赋予全局唯一的标识(NodeID)，负责存储属于该子空间的所有(key, value)对，因而在 DHT 中定位内容就相当于找到负责给定 key 所属子空间的节点。DHT 实际上是把网络中的节点关系按其负责的 key 子空间重新进行组织，在应用层上形成一个逻辑网络；在逻辑网络中寻址，找到所需文件所在的节点地址，然后再通过实际物理网络获取所需文件。

[1] 基金项目：国家自然科学基金重大研究计划，编号：90104011。

DHT 要得到广泛应用必须实现逻辑网络和实际物理网络的良好结合，但目前突出的问题是定位内容花费的实际链路时延较大，缺乏路由本地性 (routing locality)；原因是建立逻辑网络时未考虑节点实际物理距离，导致逻辑邻近的节点可能实际相距甚远。学术界已提出一些考虑网络拓扑的改进 DHT 网络，如改进 d 维 CAN 的 PNS 方法[2]；先离散分级[8]后建立 CAN 的方法；改进 Chord 系统的 HIERAS 系统[5]及 Canon 原则[9]。分析得出 Canon 原则是几种改进中最为有效的：没有过多增加节点存储开销；子区域结合方法继承了平面 Chord 设计的查询路由有效性。但 Canon 方法只适用于对数型网络，如 Chord 和 $\log_2 N$ 维 CAN，无法推广到固定维数的 d 维 CAN。d 维 CAN 不同于 Chord 等，节点 NodeID 不等长且随虚平面划分情况动态变化，所以两个子 CAN 系统中的节点 NodeID 必有重复，无法象其它改进 DHT 制定全局 NodeID 以建立不同子 CAN 系统中节点的联系。但同时 d 维 CAN 存在优于对数性网络之处：由于维数固定为 d (一般 $d<<\log_2 N$)，节点总数 N 的增加不会增加节点邻居表开销；网络可扩性好，新节点加入只需更新 2d 节点的邻居表，而后者需要更新 $2\log_2 N$ 个节点状态；优于 Pastry 和 Chord 的网络容错性(fault tolerance)。总之，d 维 CAN 适合建立节点的加入和离开频繁的大规模 P2P 网络文件共享系统。

总结前人的工作，将节点分区域分簇，分别建立子 DHT 系统，再将子 DHT 系统按层次有机结合，是解决 DHT 和实际物理网络结合问题的必经之途。由此本文提出 **HiMCAN**（**Hi**erarchical **M**erging **C**ontent **A**ddressable **N**etwork，层次形结合内容寻址网络），采用区域层次化及 Canon 结合子系统的思想，对 CAN 进行改进以得到层次结构网络的优点；HiMCAN 适合于在广域网络范围建立 DHT。本文第二节简单介绍 CAN 的基本原理，第三节开始从组件模型、基本的查询路由算法等方面详细描述本文提出的 HiMCAN 系统，第五节给出 HiMCAN 的图论理论分析和仿真结果，最后一节是结语和进一步的工作。

2. CAN 简介

DHT 基本算法是在围绕虚平面(virtual coordinate)进行描述的，虚平面本质上是逻辑网络的组织形式，不同 DHT 的虚平面具有不同的拓扑结构。CAN 的虚平面是 d 维环空间(d-torus)，每个节点负责维护邻近的超立方体空间。为叙述方便，一般讨论二维 CAN，所得结论可适用于任何维数 CAN。图 1 给出一个五个节点维护的 CAN 虚平面示意图。其中的节点维护区域代表节点存储的 Hash 函数值的值域范围。CAN 最主要算法的是节点路由算法，即 key 查询算法。如图 2 所示，每个节点存有 2d 个邻居节点信息，如节点 I 提出对具有 key 的内容进行查询，则分别计算 key 的 x 轴哈希函数和 y 轴哈希函数函数值作为<key,value>在虚平面上的坐标点，通过将查询消息不断传递给距离目标坐标点更进的邻居节点，到达负责目标坐标点所在子空间的节点，完成查询路由。

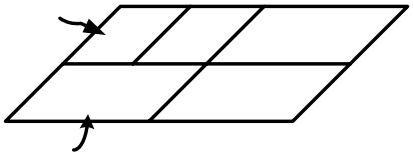


图 1 二维平面 CAN 模型示意图

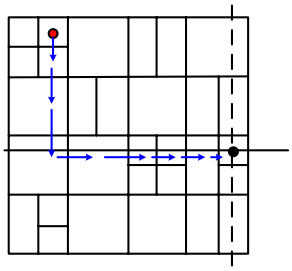


图 2 二维平面 CAN 查询内容路由示意图

3. HiMCAN 模型

从上节的例子中可见 CAN 的缺点：节点 I 到节点 J 之间的寻路跳数(hop)并不代表反映节点间的实际链路时延。I, J 可能是同一网段中的节点，但查询路径却可能经过了位于 Cernet 和欧洲，美国等地的主机，导致不必要的路由时延。针对以上问题， HiMCAN 的基本改进方法是，将地理相近的节点簇独立建立子 CAN(subCAN)，用层次区域模型(HD 模型，

Hierarchical Domain Model)建立 subCAN 间层次关系并得到 subCAN 的标识符(subCAN ID)，计算每个 subCAN 的文件－节点模型(FN 模型，File-Node Model)参数，在此基础上使用结合算法(merging algorithm)建立不同 subCAN 中同位置节点间的有权边(WI 边)，即得 HiMCAN 逻辑网络。其中 HD 模型，FN 模型，和 WI 边的方法等将在下文中详述。

3.1 HD 模型

HD 模型将物理位置相近的节点划分为同一区域。划分方法采用网络自治域(AS 域)划分，或离散分级方法[2]。例如图 3 (a)是 HiMCAN 讨论的一个典型的网络拓扑，由 5 个子网组成，每个子网规模类似于校园网或 MAN，各子网规模的差距不大，但网络链路时延差距较大。HD 模型改变了 CAN 虚平面全局平面结构，能够与实际网络中节点成簇(cluster)、分区域的特性更好结合，提高查询路由效率。

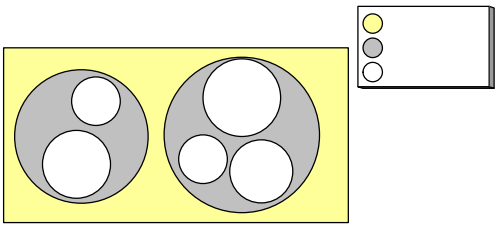


图 3 (a) 集合图表示的区域邻近层次关系

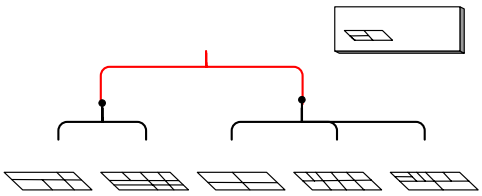


图 3 (b) 对应（1）的，三个 subCAN 组成的二层 HiMCAN

节点分区后需确定区域的邻近关系。可先用集合表示，集合级数(level)表示相应的区域中节点的不同量级时延，如图 3 (a)。在此基础上可建立广义划分树[2]，基本原则是，处于相同子树中的区域比非相同子树中的区域地理越近，且层次越低的子树中的区域距离越近，如图 3 (b)中的树形结构。该树是不完全 k 叉树，k 是每级中的最大分支数。使用反映区域层次关系的类 Huffman 树编码方法确定 subCAN ID：树的每层的分支依次标为 0,1,…,k-1 的整数，subCAN ID 是从根结点到所对应的叶子节点的路径上的标号序列，形式为 $a_0a_1\ldots a_l, a_i \in \{0,1,\ldots,k-1\}$，$l$ 是树的深度。

3.2 FN 模型

FN 模型在虚平面上表示出单个节点上 key 的实际分布。CAN 假设节点上存储有子空间中任何 key 值；HiMCAN 引入 FN 模型是考虑到各个 subCAN 实际存储 key 的集合，及 key 分布的差异，否则，单个 subCAN 即可完成对任何 key 的查询，和 CAN 毫无差别。所以 FN 模型中 key 的查询成功定义与 CAN 有所不同：通过查询路由算法找到 key 哈希值所在的子空间（CAN 到这一步即认为查询成功），且负责的节点确实存储有被查询 key，则查询成功，否则转而查询其它子 CAN 中的节点或认为查询失败。同时，可定义在结合算法中起重要作用的单个节点的 key 分布参数如下：

定义 1 区域文件存储重心(FCG，File Central of Gravity)

$$\overrightarrow{FCG_i} = \frac{\sum_j \overrightarrow{L_{i,j}} \cdot \delta(i,j)}{N_i^{key}},$$

其中 N_i^{key}：i节点上存储的文件总数；

$\overrightarrow{L_{i,j}}$：存储在节点j上的key_i的虚平面位置矢量，即key_i的虚平面二维坐标；

S_j：节点j负责区域的面积。

$$\delta(i,j) = \begin{cases} 1, & \text{文件}i\text{存储在节点}j\text{上；} \\ 0, & \text{其他。} \end{cases}$$

定义 2 节点区域距离(D，Zone Distance)，D 既反映两片区域值域的重叠程度，又反映

两个区域中的 key 在 Cartesian 平面上欧氏距离的统计平均。所以节点 i，j 间距离 D(i,j)的定义应满足：i) D(i,j)=0⇔ 节点 i，j 文件分布情况相同（约为 FCG 重合）；ii) D(i，j)≠0⇔ 节点 i，j 的 FCG 不重合；iii) D(i,j)越大，两区域 FCG 的距离越大（等同于区域中离散点的分布情况差距越大），同时区域值域重合程度越大，本文最终得到最为合理的定义是，（定义的合理性验证见 4.2）

$$D(i,j)=\sigma\left\|\overrightarrow{FCG}_i-\overrightarrow{FCG}_j\right\| \quad (\sigma=\frac{\|S_i\cap S_j\|}{\|S_i\cup S_j\|},$$

$0\le\sigma\le 1$)，其含义是两片区域重心距离的加权值。

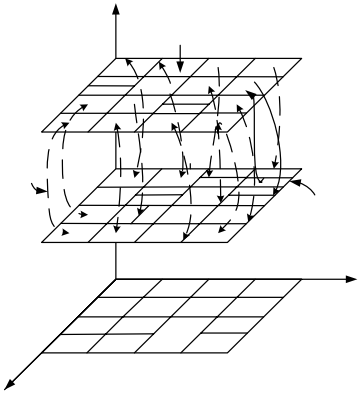


图 4 HiMCAN 层次虚平面示意图

3.3 HiMCAN 的建立

图 4 表示 HiMCAN 的层次虚平面，x、y 轴坐标是[0,1]区间中的有理数，z 轴坐标是 k 进制整数表示的 subCAN ID。其中的各个实体的表示见表 1。各 subCAN 虚平面和 CAN 相同，它们之间使用结合算法)建立的区域间有权边(Weighted Inter-domain link，WI 边)。理论上每个节点需比 CAN 多建立 $O(k^l)$ 条边，但为降低节点存储负担，实际上只需保持权重最大的 $l_{\max}$ 条 WI 边即可达到路由算法的有效性。层次虚平面的性质决定了路由算法的设计，总结如下：

表 1 层次虚平面中实体的表示

	表示方法
key	(Hashx(key), Hashy(key))
节点	Node ID
subCAN	subCAN ID

性质 1 由于节点 Node ID 能完全对应节点在虚平面上的位置[2]，且 Node ID 码长能够确定虚平面的局部划分程度，所以 subCAN ID 不同，Node ID 相同的节点，维护区域的 x-y 值域相同。根据性质 1，可定义两个不同 subCAN 中的同位置节点：

定义 3 subCAN A 中 $VID_A=\{a_1a_2...a_j\,|\,1\le j\le k\}$ 的节点，在 subCAN B 中同位置节点的 VID_B 是：

$$VID_B=\begin{cases}VID_j=\{a_1a_2...a_j\,|\,1\le j\le k\},\ if\quad VID_j\neq\phi\\ VID_j=\{a_1a_2...a_i\,|\,1\le i\le j\},\max i,\min D(VID_A,VID_j),\ if\quad VID_j\neq\phi\\ VID_j=\{a_1a_2...a_i\,|\,j\le i\le k\},\max i,\min D(VID_A,VID_j),\ if\quad VID_j\neq\phi\end{cases}$$

即具有相同 Node ID 的节点，或 Node ID 码距最小的节点。

性质 2 节点发布新共享文件只在本地 subCAN 中进行，所以每个 subCAN 存储的 key 集合不同。

性质 3 同一个 key 出现在不同的 subCAN 中，映射到不同 subCAN 虚平面的相同位置上。因为 subCAN 的单位虚平面（即未划分前的整体虚平面）都是 $[0,1]\times[0,1]\times\cdots\times[0,1]$ 形式，且使用相同的 Hash 函数序列 $\{H_1(\cdot),H_2(\cdot),...,H_d(\cdot)\}$。

d

在层次虚平面上可直观描述结合算法（Merging Algorithm）。先讨论两

表 2 仿真参数表

仿真试验 No.	时延模型	节点数	文件 key 数	S-S 时延（ms）	T-S 时延（ms）	T-T 时延（ms）
1	TS 模型 1	100	1000	10	100	1000
3		512	1024	10	100	1000
4		512	4096	10	100	1000
5		1024	2048	10	100	1000
6	TS 模型 2	100	1000	1	10	100
7	TS 模型 3	100	1000	1	5	10

个 subCAN 之间的结合：在两个区域中分别建立二维 subCAN A 和 subCAN B；计算 subCAN A 和 subCAN B 中每个节点 n_i 负责区域的文件重心 FCG_i；从 subCAN A 每个节点 n_i 到 subCAN B 虚平面的同位置节点 n_i' 建立连接，计算距离 $D(n_i, n_i')$ 作为 WI 边的权重。其次，从 subCAN 开始深度优先遍历 HD 层次树，在遍历过程中 subCAN A 中的每个节点选择距离最大的 l_{max} 个节点建立 WI 边连接。最后，每个 subCAN 按照与 subCAN A 相同的方法建立到其它 subCAN 的 WI 边，最终结合成完整的 HiMCAN。

3.4 HiMCAN 路由算法

路由算法是 HiMCAN 中的核心算法；即给定目标 key 坐标，从提出查询的节点到达负责存贮 key 的节点的算法。x-y 方向路由可采用和 CAN 相同的方法，但如果本地 sub-CAN 查找不到 key，根据性质 3，可能在另外一个 subCAN 上找到 key，又根据性质 4，同一 key 可能存储在不同 subCAN 的同位置节点，所以如何在本地查询失败时选择 subCAN 跳转，即在 z 方向进行路由是考虑的关键。最简单的方法是泛洪(Flooding)方式，无选择依次跳转到所有 subCAN 的同位置节点查找，如都失败则认为查询 key 失败，但这样做查询效率显然很低。若能一次恰好跳转到存储 key 的 subCAN 同位置节点，则查询效率最高，但节点无法保存所有其它 subCAN 的 key 分布信息。考虑折中，由于 WI 边定义的本质是节点到同位置节点的区域重叠程度和 key 分布差别程度的衡量，距离越大则另一区域的 key 分布越不同并且是同位置节点。所以在一个节点上查找 key 失败时，跳转到距离大的节点查找成功的可能性更大。

4 合理性验证、性能评估与仿真结果

4.1 HiMCAN 性能分析

subCAN 的节点数目平均为 N_0，参考前文的变量定义，对 HiMCAN 层次虚平面从图论角度可导出以下结论，见表 2。

表 2 HiMCAN 和 CAN 的参数对比

	HiMCAN	CAN
节点的度	$2d + l_{max}$	$2d$
平均成功寻路跳数	$\frac{1}{2} dN_0^{\frac{1}{d}} + l_{max}/2$	$\frac{1}{2} dN_0^{\frac{1}{d}}$

虽然结论中 HiMCAN 的成功寻路平均跳数略大于原型 CAN，节点存储开销略加大，但代价的付出是取得减小路由时延的必需。从应用角度，HiMCAN 的优点还有：(1) 本地路由性，所以有较小的网路时延；(2) subCAN 间查询路径汇聚(convergence)[9]。

4.2 合理性验证与仿真结果

本文在 GT-ITM[10]的 TS 网络拓扑模型上同时搭建 HiMCAN 和 CAN 网络进行比较，仿真程序主要由 C++实现。仿真使用简单的 TS 网络拓扑代表实际网络如图 5，仿真参数见表 2。

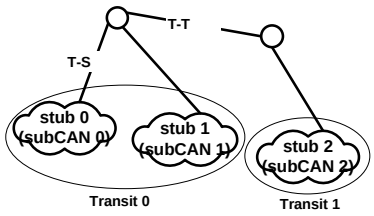


图 5 仿真拓扑图

仿真一：验证区域距离定义 3 的合理性。仿真在两个有 100 个节点的 subCAN 中随机分布 1024 个文件 key，测量所有节点之间的区域距离和它们实际存储的文件 key 的相同率，使用定义 3 和另一种距离定义。根据距离定义应有：区域距离越小，节点文件 key 的相似程度越大；图形靠近原点处越尖锐，定义反映的相关性越大。从图 6(a)定义 2 的相关性明显好于图 6(b)的定义。

仿真二：WI 边方法与泛洪方法查询效率比较。图 7 中 WI 边方法的成功寻路跳数小于泛洪方法。同时仿真得出，若要到达最大查询成功率，节点 WI 边表的长度 l_{max} 应至少为 subCAN 数目的 0.45 倍。

仿真三：图 8 给出 HiMCAN 与 CAN 路由算法效率在平均链路时延参数上的对比，可见 CAN 的平均时延约大于 HiMCAN 两个数量级，表明 HiMCAN 路由效率优于原型 CAN。

6 结语和进一步的工作

在分析现有 DHT 路由算法研究发展状况的基础上，本文提出一种新型的针对解决 CAN 路由本地性问题的改进型 CAN 网络：HiMCAN，并全面阐述了其设计目标、组件 HC 模型和 FN 模型、静态路由算法等；从理论分析和仿真实验两方面验证了系统的合理性与有效性。进一步工作可考虑 HiMCAN 的动态特性：节点的加入、离开和失效及优化策略；HiMCAN 是着眼于应用，其改进付出了一定系统的开销，其代价－效率比有多大还需探讨，也希望借此启发新的研究思路。

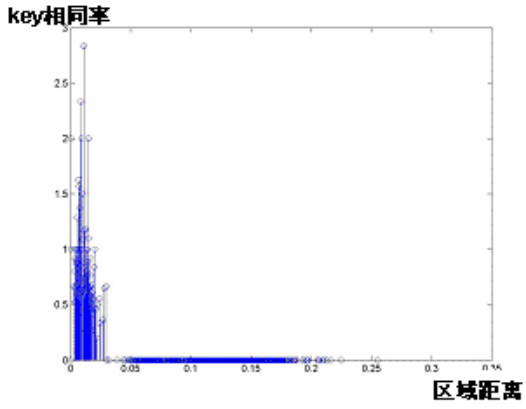


图 6(a)定义 3 的情况

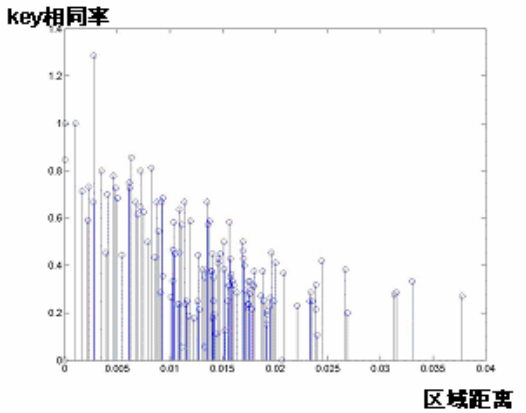


图 6(b)如果定义 3 中采用$\sigma'=1-\sigma$

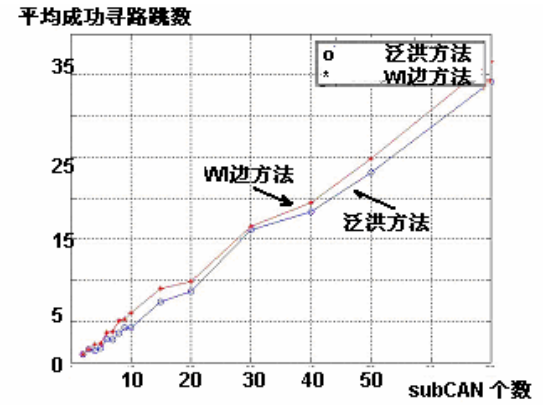


图 7 WI 边方法和泛洪方法效率比较

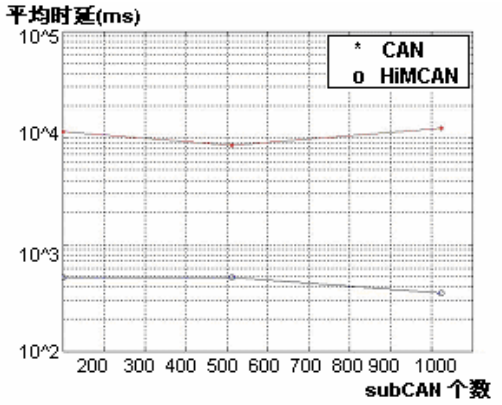


图 8 HiMCAN 和 CAN 查询时延比较

参考文献

［1］李振武，白英彩，“影响 Internet 未来的对等网络”，上海交通大学金桥网络工程中心，技术报告，2003。

[2] Sylvia Paul Ratnasamy: “A Scalable Content-Addressable Network,” PhD Dissertation of *U.C.Berkeley*, 2002

[3] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, Hari Balakrishnan: “Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications,” *SIGCOMM’01*, August 27-31, 2001，CA, USA.

[4] A. Rowstron and P. Druschel: “Pastry: Scalable, distributed object location and routing for large – scale peer-to-peer systems.” In *Proc. IFIP/ACM Middleware* 2001, Heidelberg, Germany, Nov. 2001.

[5] Zhiyong Xu, Rui Min and Yiming Hu, HIERAS: A DHT Based Hierachical P2P Routing Algorithm, *Proc.* of *ICPC*, 2003.

[8] Sylvia P. Ratnasamy, Mark Handley, Richard Karp, Scott Shenker: “Topologicaly-Aware Overlay Construction and Server Selection,” in the *proceeding of INFOCOM*, 2002.

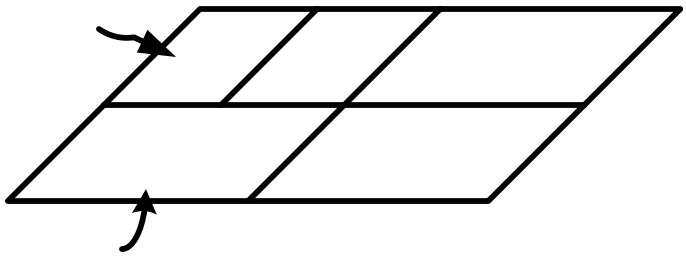
[9] Prasanna Ganesan, Krishna Gummadi, Hector Garcia-Molina: “Canon in G Major: Designing DHTs with

Hierarchical Structure," In *ICDCS*, 2004.

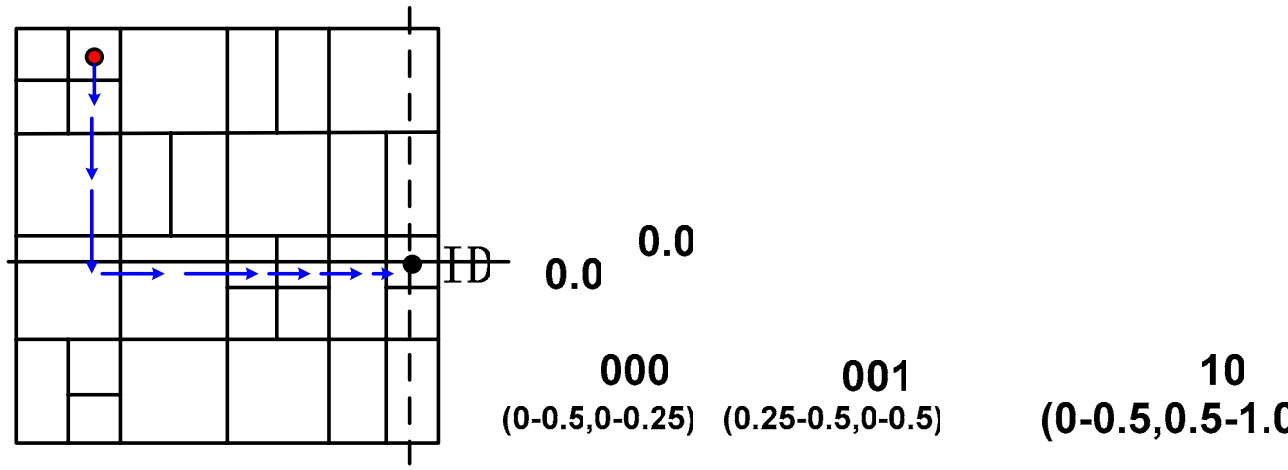
[10]Kenneth L. Calvert, Matthew B. Doar, Ellen W. Zegura: "Modeling Internet Topology",1997。

作者简介：谢瑶（1982－），女，内蒙古呼和浩特人，中国科学技术大学本科毕业，目前在美国佛罗里达大学攻读博士学位；洪佩琳（1961－），女，浙江宁波人，中国科学技术大学教授，博士生导师，主要研究方向为信息通信网和网络安全；李津生（1937－），男，上海人，中国科学技术大学教授，博士生导师，主要研究方向为信息通信网。

放大图：



图一



图二

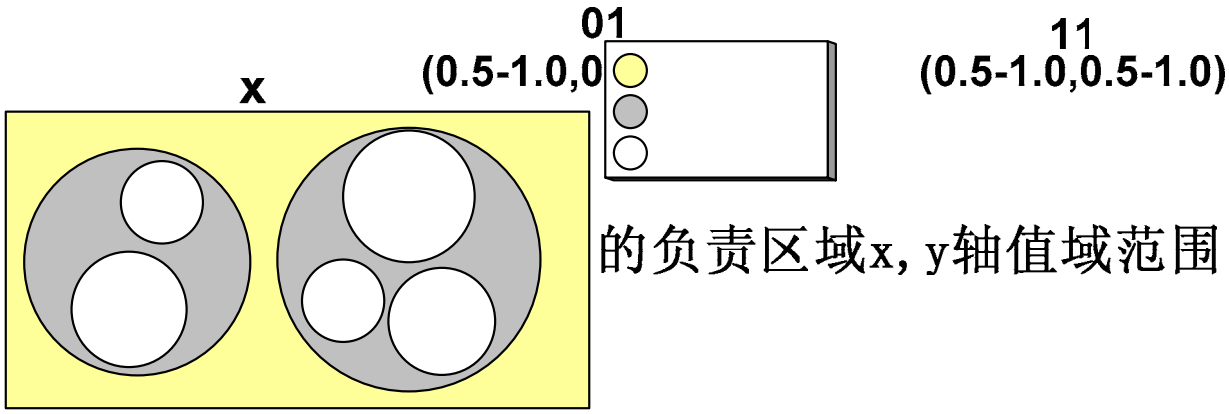


图 3(a)

y (Key,Value)

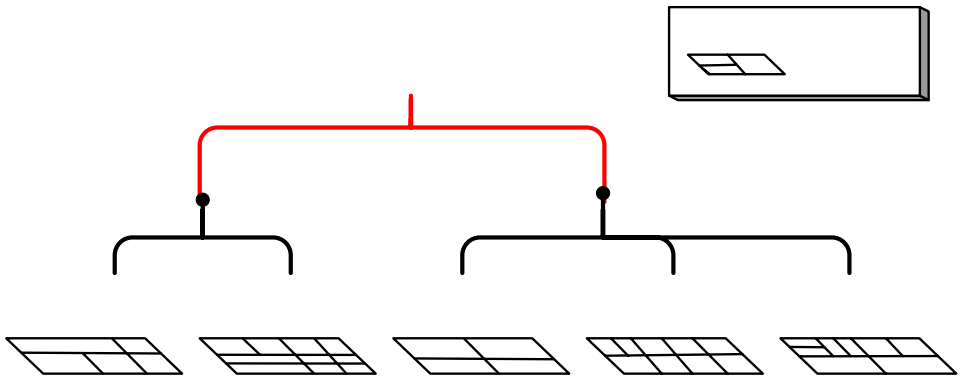
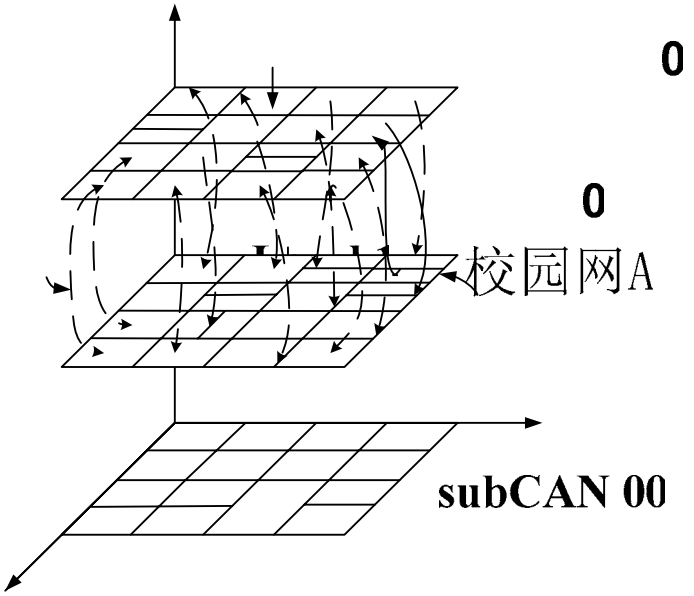


图 3(b)

Level 0



0

0 1 0

校园网A 城域网B 城域网

subCAN 00 subCAN 01 subC

图 4

图 6(a)

z

subCAN ID

节点维护
区域

011

010

WI 边

subCAN

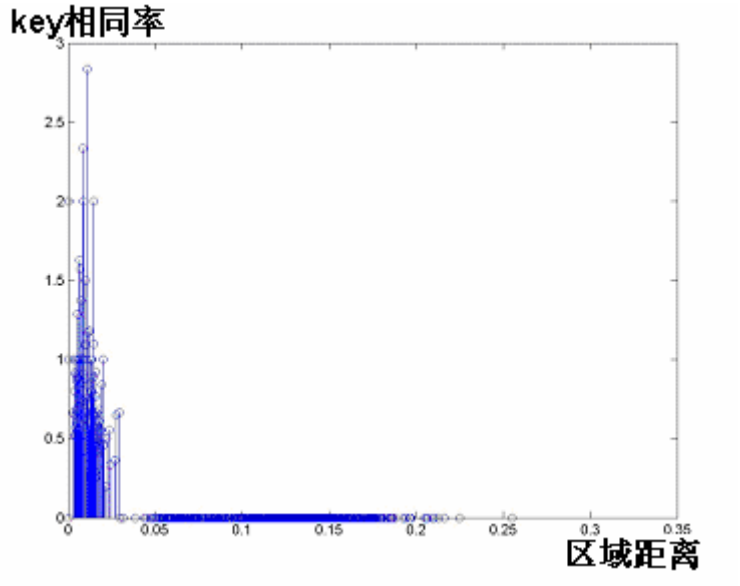


图 6(b)

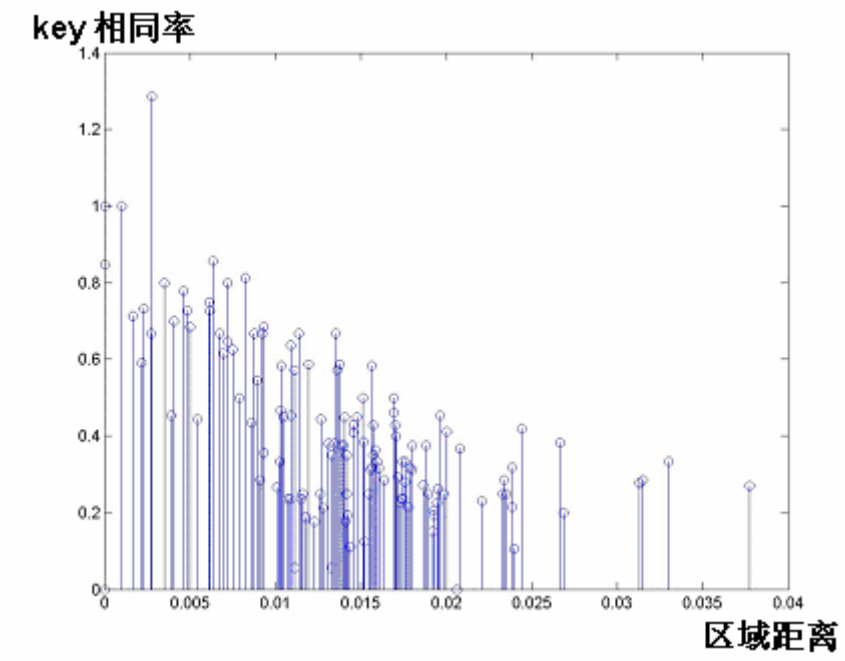
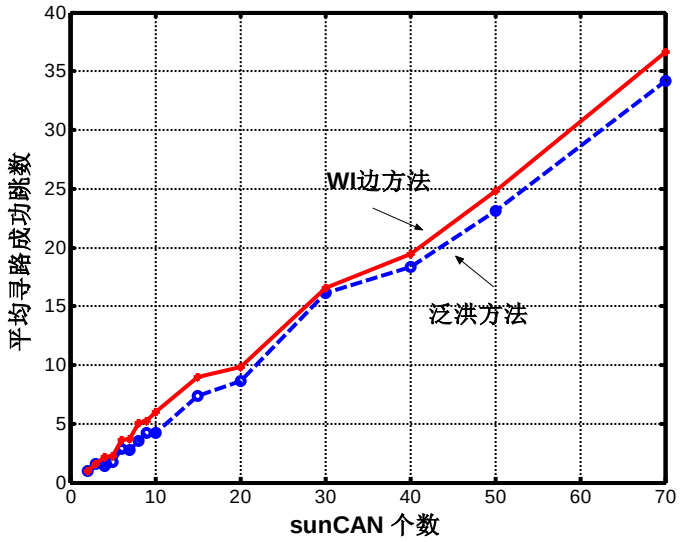
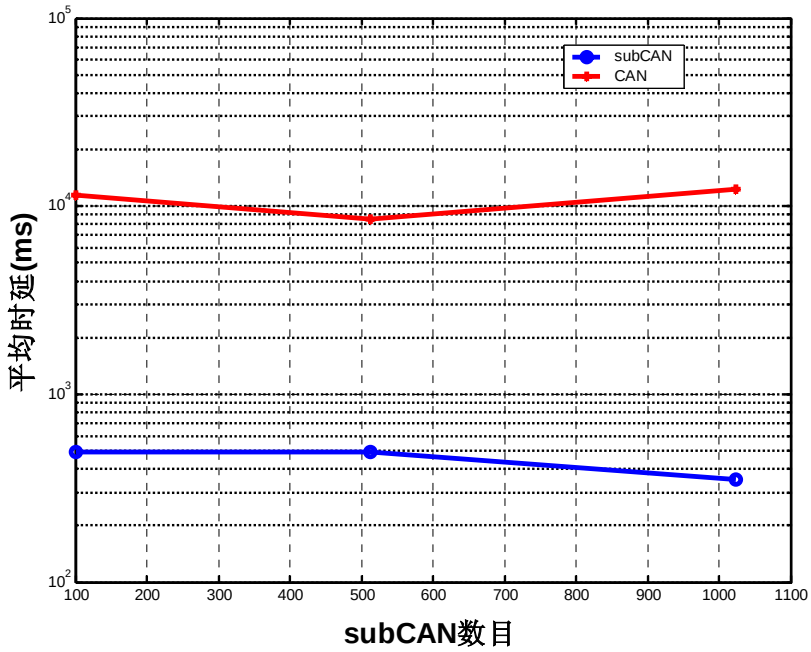


图 7，8


平均寻路成功跳数
sunCAN 个数
WI边方法
泛洪方法
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50
60
70


平均时延(ms)
subCAN数目
subCAN
CAN
10^2
10^3
10^4
10^5
100
200
300
400
500
600
700
800
900
1000
1100