基于 Region 多层结构 P2P 计算网络模型*

乐光学[1,2+], 李仁发[1], 周祖德[3]

[1](湖南大学 湖南省嵌入式计算及系统重点实验室,湖南 长沙 410082)

[2](怀化学院 计算机系,湖南 怀化 418000)

[3](武汉理工大学,湖北 武汉 430070)

A P2P Network Model with Multi-Layer Architecture Based on Region

YUE Guang-Xue[1,2+], LI Ren-Fa[1], ZHOU Zu-De[3]

[1](Key Laboratory of Embedded Computer and System of Hu'nan Province, Hu'nan University, Changsha 410082, China)

[2](Department of Computer Science, Huaihua University, Huaihua 418000, China)

[3](Wuhan University of Technology, Wuhan 430070, China)

+ Corresponding author: Phn: +86-731-8821544, E-mail: guangxueyue@yahoo.com.cn, http://lecs.hnu.cn/

Received 2004-12-29; Accepted 2005-03-11

Yue GX, Li RF, Zhou ZD. A P2P network model with multi-layer architecture based on region. *Journal of Software*, 2005,16(6):1140−1150. DOI: 10.1360/jos161140

Abstract: Gnutella application layer protocol simply uses flooding algorithm to route peer's querying, which is just implemented on application layer and doesn't use down-layer's information routing of Internet. So it has poor scalability and low efficiency. This paper uses the model of "small world and power law" as the theoretical foundation, and in the light of the users' requirement and a shared organization layer and region of logic manage-unit, a new distributed peer-to-peer network model of RLP2P(region-layer P2P) with multi-layer structure based on region is proposed, and its prototype system is implemented. A new optimized Multilayer Light-Gossip route strategy is implemented. This paper studies the stress and stretch with the model which has two intuitive metrics of goodness defined to evaluate the quality of the data paths. At the same time, the idea is put forward about the tradeoffs between the stress and stretch. Simulation results about RLP2P's protocol show that it could effectively solve the above problems, and the larger the network size is, the more obvious the superiority of its comprehensive performance is. So the model is reasonable and effective.

Key words: peer-to-peer network; layer and region; active nodes/leaders; small world; hierarchical search; stress and stretch

摘 要: 分布式 P2P 网络 Gnutella 模型中共享信息查询的路由协议为"洪泛"算法,其协议机制仅在应用层实现,缺

* Supported by the National Natural Science Foundation of China under Grant No.60273075 (国家自然科学基金); the Key Project in Hu'nan Provincial Department of Education of China under Grant No.03A036 (湖南省教育厅重点项目)

作者简介: **乐光学**(1963−),男,贵州天柱人,副教授,主要研究领域为网络技术与分布式计算,网络安全技术,数控技术;**李仁发**(1957−),男,博士,教授,博士生导师,主要研究领域为网络技术与分布式计算,嵌入式计算;**周祖德**(1946−),男,教授,博士生导师,主要研究领域为数字制造,数控技术,嵌入式计算系统.

乏对 Internet 底层通信子网路由资源的利用,存在可扩展性、性能与效率不高的问题.以“小世界和幂规律”模型为理论基础,以层和域为基本逻辑管理单位,按用户需求和共享目的组织域,提出了基于 Region 多层结构 P2P 网络模型 RLP2P(region-layer P2P),实现了其系统原型;实现了一个优化的 Multilayer Light-Gossip 分级路由策略;量化分析了表征模型数据通道质量指标的压力和伸展率,提出了综合考虑压力和伸展率的思想.模拟分析表明,RLP2P 模型可以有效地解决可扩展性、性能与效率不高问题,且网络规模越大,其综合性能的优越性越明显,因此,模型是合理、有效的.

关键词: 对等网;层和域;主动/中心节点;小世界;分级搜索;压力和伸展率

中图法分类号: TP393 **文献标识码**: A

P2P(peer-to-peer)网络中所有的计算机是对等的,具有相同的责任与能力,兼有客户机和服务器的功能;对等点通过直接互连实现资源的全面共享,无须依赖集中式服务器支持,消除了信息资源孤岛和 C/S 中的服务瓶颈问题[1].随着 Napster[2],Gnutella[3]等 P2P 应用系统的出现,P2P 技术在文件共享和信息搜索等方面的优势受到广泛关注.由于节点可随时动态地加入和退出网络,造成网络带宽和信息存在的不稳定,因此 P2P 网络存在更大的随机性、不确定性、不易管理和资源定位等问题.

1 相关研究

(1) Napster.结构模式为中心文件目录/分布式文件系统,通过中央服务器进行目录管理,实现文件共享,为开放性的共享系统提供了强大的管理能力,数据传输主要在节点之间进行,避免了网络的拥塞.由于 Napster 是一个集中式的系统,还存在着单点瓶颈问题.

(2) Gnutella.协议允许用户与网络中的其他节点直接联系,无须经过中央目录服务器和认证机构,节点是完全对等的.网络中节点接近于绝对的自由,很难进行安全、流量等控制;泛洪式搜索,产生指数级的消息冗余,消耗大量的处理时间和网络带宽,随着这部分节点的失效,Gnutella 网络被分片,使查询访问只能在网络中很小的一部分进行,导致网络的可扩展性下降[4,5].相同的系统还有 Freenet[6].

(3) 基于超级节点的两层结构 P2P 虚拟网络,是 Napster 和 Gnutella 模型的折衷,如 Sun 公司的研究项目 Project JXTA 2.0 Super-Peer Virtual Network[7,8].其思想是将网络划分为 M 个自治域,通过自治域来管理网络,域内节点分为超级节点(super-peers)和普通节点,超级节点由域内功能最强的节点承担,负责域内网络的管理.网络通过分布式文件系统,建立完全开放的可共享文件目录,运用相对的自由来兼顾安全和可管理性.据查阅文献资料显示,这是目前已知的唯一与本文研究模型相似的 P2P 结构模型.

国内外在 P2P 领域的研究基本集中在基于纯 P2P 网络 Gnutella 模型和基于自治域有超级节点的双层结构 P2P 网络模型上展开,对模型进行优化和扩充.据查阅文献资料显示,在对基于自治域的有超级节点的双层结构 P2P 网络模型研究中存在如下问题:(1) 自治域的划分策略不清,均没有给出具体的数学模型;(2) 路由协议缺乏对 Internet 底层通信子网路由资源的利用,对资源定位、搜索策略和访问控制规则表达式研究得较少,存在可扩展性差、性能与效率不高的问题;(3) 没有统一的网络综合性能的评价标准,网络开销大部分是估算或引用传统网络的研究结果来评价;(4) 对表征 P2P 网络数据通道质量评价指标压力和伸展率没有进行详细分析和量化;(5) 节点采用就近加入网络,易造成网络波动,导致管理开销增加.

针对上述问题,受现实社会“物以类聚”规律的启示,本文以“小世界模型和幂规律”为理论基础,以层和域为基本逻辑管理单位,按用户需求和共享目的组织域,将 Gnutella 网络模型抽象层次化,提出了一种基于 Region 多层结构 P2P 计算网络协议模型 RLP2P(region-layer P2P).

2 基于 Region 多层结构 P2P 网络模型 RLP2P 体系结构

2.1 模型假设

RLP2P 网络模型是建立在 Internet 网络上的逻辑网络,假设每个域都有一个 SNMP(simple network

management protocol)网管,以获得域内的静态拓扑结构;具有良好的扩展性和灵活性使客户可加载业务,主动、中心节点可配置,运行 Java 虚拟机的节点处理,可获取动态网络拓扑和链路状态、定位搜索对等机信息、进行相关路由查找.每一个接入网络的计算机均采用主动网络技术,每个节点兼有服务器和客户机的功能.

2.2 基本概念

在具有"小世界和幂规律"特性的网络中,网络拓扑具有高聚集度而低特征路径长的特性[9−11].

定义 1(幂规律). 在一个随机无向图中,其节点"度"为 K 的节点的分布概率满足:$P(k)\propto k^{-\tau}$,其中,$1<\tau<\infty$.在网络中少数节点有较高的"度",通过"度"较高的节点找到待查信息的概率较高[10,11].

定义 2(小世界特性). 网络拓扑具有高聚集度而低特征路径长的特性[9].

定义 3(聚集度 $C(l)$). 已知以顶点 v 为根的深度为 l 的 BFS(breadth first search)树,则该顶点的横向边的数目为 C_v,满足关系:$\mathrm{Max}(C_v)=C_k^2-(k-1)$,$k$ 为 BFS 树的所有顶点,则一个图的聚集度为其所有顶点 v 的 C_v 值的平均值,即 $C(l)=Average(C_v)$[9].

定义 4(特征路径长 L). 已知一个无向图 G,任意两节点(u,v)间最短路径的边数为 $Num(u,v)$,则有 $L=Average[Num(u,v)]$,L 为网络中所有任意两节点间最短路径的边数的平均值[9].

定义 5(域 Region). 按节点对资源需求、共享目的和属性的相似性进行划分的基本逻辑管理单位,由中心节点所辖区域内多个属性相似、性能相对较弱的主机组成的一个闭集.域内节点有很大的相似系数,域文件能够在很大程度上代表域内部提供的服务信息.

定义 6(中心节点 Leader). 由域内性能最高节点充当,逻辑上位于域的中心,与域内节点距离最短,是域间接口,类似于 C/S 结构中央服务器和 JXTA 中的集合点[7,8],服务功能已被弱化,负责域内节点的管理.

定义 7(普通节点 Node). 类似于 C/S 结构中的客户机和 JXTA 中的普通节点[7,8],性能较弱,服务功能已被加强.与域内节点的性能和属性基本相似,提供简单的管道和访问等服务.

定义 8(主动节点 AN(active network node)). 类似于网络中的路由,存有其下一层网络提供的服务和状态信息列表,负责引导节点加入和消息路由.

定义 9(层 Layer). L_i 层的节点由 L_{i-1} 层的中心节点构成,L_i 层是下一层中心节点的集合.

定义 10. 任意时刻节点 h 请求加入网络,h 向主动节点发出加入请求,将 h 的属性和特征状态信息传给主动节点,主动节点在所辖域内确定一个距它最近且属性最相似的响应节点,h 查询该域中的所有节点,以确定 h 的加入点,直到找到 L_0 层中的某一个合适的域加入.

2.3 RLP2P模型体系结构

在研究 P2P 网络体系结构的过程中,发现 Gnutella 网络拓扑节点的分布呈现典型的幂规律,且具有小世界特征,与 Internet 骨干网络节点的拓扑分布规律非常类似,幂指数分别为 $\tau\approx2.3$ 和 $\tau\approx2.2$[10,11].P2P 网络中节点的行为与现实世界中"物以类聚"的现象极为相似.Gnutella 网络模型中共享信息查询的搜索,定位路由协议为"洪泛"算法,在应用层实现,缺乏对 Internet 底层通信子网路由资源的利用,存在可扩展性差、性能与效率不高的问题[12−14].结合主动网络技术,将 Gnutella 网络模型抽象层次化,将骨干网络中的路由交换节点扩展成主动节点,对等节点在加入该网络时,总是找离自己最近的那个主动节点,由其引导加入到属性相识的域中,充分利用原 Internet 网络路由资源和节点的拓扑结构,形成自然的小世界模型和幂规律特性,网络模型中对等点的定位与搜索将主要通过主动节点进行,充分利用主动节点的路由表信息,使得查询定位不再以盲目扩散方式进行,从根本上改善了定位搜索效率和网络的可扩展性;将网络中的中心服务器和客户机抽象为中心节点和普通节点,并赋予其不同的职责.理论上,主动、中心和普通节点在搜索、资源和信息共享上处于同等位置,是完全对等的,但主动、中心节点同属于两个不同的层和域,域与域之间由主动、中心节点连接,并保存有相邻域主动、中心节点的信息,其任务特征和性能比普通节点更广;将聚集在各主动、中心节点周围的主动、中心节点抽象为一个统一的"共享信息"域,与其他"共享信息"域又组成对等连接关系,重复这个过程,就可得到一个基于层域结构的广域 P2P 网络 RLP2P,网络体系结构如图 1 所示.据定义 1~定义 10 对任意网段抽象可以得图 2 所示的网络拓扑结构,网络拓扑由 m 层构成,以 $L_i(i=0,1,\ldots,m)$表示,每层由 $n(n=1,2,...,n)$个域组成,每个域含 $k(k=1,2,\ldots,k)$个节点;网

络中的任意主机必属于一个特定的域,物理上所有节点位于第 0 层不同的域中.从第 1 层开始,第 i 层的节点由第 i–1 层的主动、中心节点组成,以此类推,就构成了 RLP2P 网络拓扑.节点加入须向主动、中心节点注册节点名和地址等信息;节点请求加入或离开网络时,通过向网络发布消息,建立或撤销连接,更新资源和服务目录;节点通过主动、中心节点实现跨域访问.为了保持网络的鲁棒性,对重要的主动、中心节点采用冗余的方法.

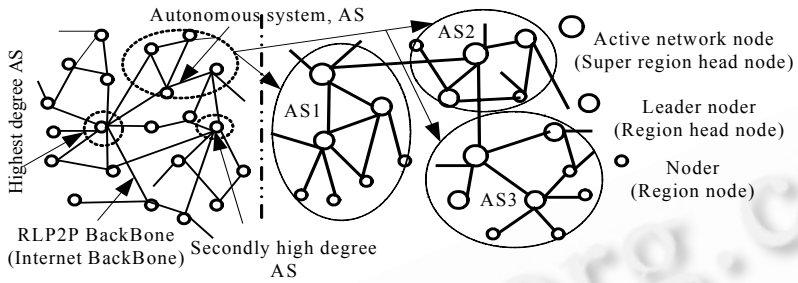

Fig.1 RLP2P architecture

图 1 RLP2P 网络模型体系结构图

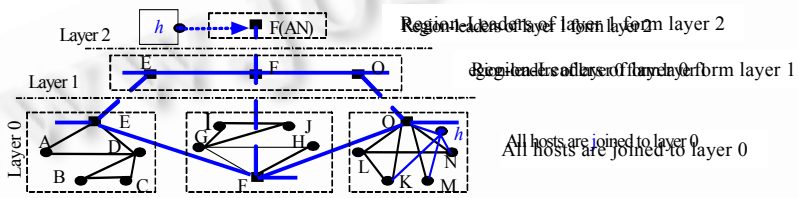

Fig.2 RLP2P topology

图 2 RLP2P 网络拓扑图

由上述可知,RLP2P 网络模型形成自然的小世界模型和幂规律特性,每个结点掌握着大量的域内节点服务信息,节点间有很大的相识系数,域文件能够在很大程度上代表域内部提供的服务信息,使得在域内部找到满足服务请求的概率很大,节点出现跨域请求服务的概率很小,使定位某种服务的工作量、查询范围从网络中的所有节点数降低到域内的节点数,有效地防止了请求洪.因此,RLP2P 模型自然满足构建网络模型应遵循的规则:(1) 检索尽可能少的对等点就能够满足尽可能多的请求;(2) 易于扩展、鲁棒性高、有利于信息传输,系统维护和控制开销尽量小;(3) 有利于数据流向对其更感兴趣的区域(即请求活跃区).

3 协议描述与性能分析

3.1 域的划分策略

网络的规模直接关系到网络的性能,量化域的划分是为了便于对网络的控制和管理,以小的代价实现网络服务资源的快速定位和资源的全面共享.通过设置属性相似系数$\mu(0.6\le\mu\le1)$来决定节点属于哪一个域,以避免节点同属于多个域的情况发生.域的划分和合并算法描述见算法 1 和算法 2.

算法 1. 域的划分操作.

Procedure: RegionSplit(C)

{$|C|\ge3k$}; /*$|C|=RL_i(X)$/

$\partial\Leftarrow\{Q|Q\subset C\cap|Q|,|C-Q|\ge|3k-1|\}$

Let $R(Q)$=max(radius(Q),radius($C-Q$))

Find Q^* s.t. $R(Q^*)\le R(Q)$; /* where Q, $Q^*\in\partial$/

LeaderTransfer(Ldr(C),Q^*,Ldr(Q^*))

LeaderTransfer(Ldr(C),$C-Q^*$,Ldr($C-Q^*$))

算法 2. 域的合并操作.

Procedure: RegionMerge(C)

{$|C|<k$ and L_i is the layer to which C belongs}

$l\Leftarrow$Ldr(C)

Find y s.t.dist(l,y)<dist(h,x),x,$y\in Rl_{i+1}(l)$

RegionMergeRequest(l,y,L_i)

LeaderTransfer(l,C,y)

定理 1. 每层的节点分属于不同的域,域的规模为 $RL_i(X)$,$k \le RL_i(X) \le 3k-1$,其中 $RL_i(X)$为域所含的节点数,k 为常数(k 值由网段的路由、中心服务器的负载能力和带宽确定).当 $RL_i(X)>3k-1$,将域等分为两个域;当 $RL_i(X)<k$,合并域.任意时刻,域至少有一个中心节点.

证明: (1) 设域的上界取一个较小的值,如 $RL_i(X) \le 2k-1$.当 $RL_i(X)>2k-1$ 时,域被等分为两个域,$RL_i(X)=k$;任意时刻,若此两域中任意域有 $n \ge 1$ 个节点离去或失效,使 $RL_i(X)<k$,两个刚划分的域又需合并;若有 $n \ge k$ 个节点加入,使 $RL_i(X)>2k-1$,域又需划分为两个,导致网络波动和管理开销增大;

(2) 设域的上界为 $RL_i(X) \le 3k-1$.当 $RL_i(X)>3k-1$,域被等分为两个域,$RL_i(X)=3k/2$;任意时刻,只有当此两个域中任意一个域有 $n \ge k/2$ 个节点离去或有 $n \ge 3k/2$ 个节点加入时,系统才进行域的合并和划分操作,使网络保持相对的稳定.因此,$RL_i(X) \le 3k-1$ 可避免域的频繁分割和合并;

(3) 设域的上界取一个较大的值,如 $RL_i(X) \le 4k-1$.域的规模增大,其划分和合并操作相对减少,但系统的管理和控制开销急剧增加,加重中心节点的负担,引发信息瓶颈,甚至导致中心节点失效,使系统的交互性和鲁棒性急剧下降.

结论 1. 域的边界值取 $k \le RL_i(k) \le 3K-1$ 是正确、合理的.

3.2 节点加入网络、调整及消息转发算法描述

根据定义 10,设节点 h 请求加入网络(如图 2 所示),将 h 的属性和特征状态信息传给 AN,h 在 AN 的引导下搜索 L_1 层的所有成员中,发现其属性与 O 最相似且距离最近,h 向 O 发出加入请求并响应,h 向 O 注册节点名和地址等信息,O 告知 h 在 L_0 层的所辖域中有节点 K,L,M,N,h 向网络发布能提供的服务和特征状态信息,并与节点建立对等连接,系统更新服务和信息列表,h 加入完成;若域中的节点发现网络中另外一个域的特性更适合自己,节点自动进行调整;消息转发采用分布式向前咨询方式完成.

为了减少节点的加入延迟和避免 AN 成为瓶颈,AN 需要维持一个适当的加入请求响应信息表.规定当 AN 收到加入请求时,将请求加入节点视为临时节点,进入当前层,但不属于任何域,由当前层的主动、中心节点向下发出请求.在图 2 中,h 在加入到 K,L,M,N 节点所在的域之前,h 是 F 的一个临时成员.节点请求加入网络、位置转移的算法源语描述见算法 3 和算法 4.

算法 3. 节点请求加入操作.

```
Procdure: BasicJoinLayer(h,i)
RL_j⇐Query(AN,–)
While (j>i)
  Find y s.t dist(h,y)≤dist(h,x),x,y∈RL_j
  RL_{j–1}(y)⇐Query(y,j–1)
  Decrement j,RL_j⇐RL_{j–1}(y)
Endwhile
JoinRegion RL_j
```

算法 4. 节点域间转移操作.

```
Procedure: RegionRefine(h)
{L_i is highest layer to which h belongs}
l⇐Ldr(RL_i(h)): C⇐RL_{i+1}(l)
Find y s.t. dist(h,y)<dist(h,x),x,y∈C
if (y≠l)
 LeaveRegion(h,l,L_i)
 JoinRegion(h,y,L_i)
endif
```

3.3 节点离开、中心节点选取策略和算法描述

若节点 h 正常离开,它把离开信息在域内广播;若 h 因失效或异常离开,h 发不出离开信息,则系统通过 h 的“心率信息”来确认.若正常离开节点 h 是主动、中心节点,则根据域内节点的服务和特征状态信息,直接推举一个节点,承担主动、中心节点之责,将网络服务、资源和状态信息直接传给新的主动、中心节点,并做标记;若主动、中心节点因失效或异常离开,且没有冗余主动、中心节点,则该域暂时与网络断开,系统运行产生主动、中心节点的代理(Agent),在域内选取一个综合性能高的节点为新主动、中心节点,重新建立系统服务、资源、状态信息表和连接,并做标记.因此,依据接入网络性能的高低,域内节点的角色可互换(节点属性可配置),为增强网络的鲁棒性,可在域内增设一个冗余的“准”主动、中心节点.

由上述分析得知,RLP2P 网络模型具有如下特点:(1) 每个层分为若干个域,每个域的规模为 k~$3k-1$ 台主机;每个域至少有一个主动、中心节点,所有主机最初都属于初始层 L_0.(2) 任意主机在任何层只能存在于一个域,

但主动、中心节点同属于相邻层中不同的两个域;如果主机处于 L_i 层的某个域中,它一定是 $L_0,\ldots,L_{i-1}$ 层的某个域中的主动、中心节点;如果一个主机不在 L_i 层中,也必不在 L_j 层中,其中 $j>i$.(3) 网络最多有 $\log_k(N)$层,最高层只有一个成员;(4) 路由搜索充分利用了 Internet 网络路由资源和节点的拓扑结构,具有良好的扩展性;由 AN 引导节点加入到属性相识的域,避免了节点采用就近加入网络时因属性相异而出现节点域间频繁移动,使系统开销增加,鲁棒性变低的情况发生.其协议模型结构[7]见表 1.

Table 1 The hierarchy view of RLP2P protocol model

表 1 RLP2P 协议层次模型

<table>
<tr><td colspan="7">Region group service</td></tr>
<tr><td colspan="2">Discovery service</td><td>Pipe service</td><td colspan="2">Membership service</td><td colspan="2">Peer info service</td></tr>
<tr><td colspan="2">Resolver service</td><td>SRDI</td><td colspan="2">Rendezvous(active) service</td><td>RPV</td><td>Walker</td></tr>
<tr><td colspan="7">Endpoint service</td></tr>
<tr><td colspan="2">Advertisements</td><td colspan="3">XML parser</td><td colspan="2">Virtual messenger</td></tr>
<tr><td rowspan="2">ID</td><td rowspan="2">Cache/Index manager</td><td colspan="3">Message</td><td rowspan="2">Relay</td><td rowspan="2">Router</td></tr>
<tr><td>Http transport</td><td>TCP/IP Transport</td><td>TLS transport</td></tr>
</table>

4 路由协议描述

4.1 主动、中心节点连接策略

理论上 P2P 网络中的对等点是直接互连的,但要在广域环境下实现所有对等点直接互连是不可能的,只能实现局部直接互连.在 RLP2P 模型中,用于实现区域间主动、中心节点通信机制的拓扑结构采用类似蜂窝的正六边形网格,如图 3 所示.主动、中心节点在探测结果中选出符合要求的 3 个邻近节点连接并将节点信息保存,图 3(a)为理想的网络逻辑互连状态,图 3(b)为有主动、中心节点缺省的网络逻辑互连状态.主动、中心节点的加入/离开管理策略与第 3 节相同.

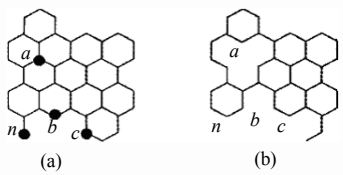

Fig.3 Active nodes/Leaders communicate catenation

图 3 主动/中心节点通信连接关系示意图

4.2 路由表

与纯 P2P 网络结构相比,尽管 RLP2P 系统需要进行远距离信息交互的节点数目大大降低,但要保存整个网络内的所有主动、中心节点间的邻接关系也是不现实的.为了实现高效的数据搜索和最小冗余扩散,减少控制开销,规定每个主动、中心节点拥有一个聚类中心,节点根据搜索表现更新路由表,使内容向着聚类中心进行聚类.各个主动、中心节点的聚类中心在 key 空间中均匀分布,主动、中心节点根据近期 query 的分布分配聚类中心以保证整个系统的索引均衡性.主动、中心节点维护一张长度为 L 的本地路由表,由层域表(LRB)和域内表(RB)组成,长度分别为 L_L 和 $L_R(L=L_L+L_R)$,LRB 表记录节点 ID 及其聚类中心,RB 表记录节点 ID 和代表其文件特征的 key,二者按相应的聚类策略进行更新.每个主动、中心节点维护的路由表除了保存与其邻居节点的连接关系信息以外,还必须保存与其相邻节点的邻居节点的节点聚集连接关系列表,即节点必须了解其邻居节点的邻居连接关系信息,节点通过与邻居定期交换信息实现网络服务和状态更新.

4.3 消息路由策略

在 LRP2P 网络模型中,针对域文件能够在很大程度上代表域内部提供的服务信息,在域内部找到满足服务请求的概率很大,节点出现跨域请求服务的概率很小以及节点个体自私、动态演化的特征,我们提出了一种级间消息扩散的动态预分组和节点最优路径动态预测策略,使消息能够自适应地沿着一条在时间度量上尽量短的路径前进,以提高每次路由的效率;结合“泛洪”(高效、稳健、简单,指数级冗余)和“最小生成树”(鲁棒性差、

未用冗余)搜索算法的特点,运用“最大聚集度优先 MCF(maximum clustering-coefficient first)”的原则,实现一个高效的分级路由策略 Multilayer Light-Gossip 和基于 F-measure 算法的网络域间节点连接管理策略.其核心思想为:以域作为搜索包扩散基本逻辑单位,消息分级扩散;大范围搜索(如层和域间)采用 Light-Flooding 算法,小范围搜索(如域内)采用 Gossip[15]算法,以有限冗余实现高效搜索.在域间,由主动、中心节点将消息扩散到邻域中的主动、中心节点,实现节点跨域搜索;在域内,主动、中心节点将接收到的消息在所辖域内节点中进行扩散.每条消息报头加入 1 bit 的标志位,以标识消息扩散级别,域级主动、中心节点间扩散标志位置 1,域内节点级扩散标志位置 0.若主动、中心节点接收到的消息标志为 1 时,需对消息转发 2 次,向邻域的主动、中心节点转发,标志位不变;向所辖域内节点转发,标志位置 0.若主动、中心节点接收到消息的标志位为 0,则只在域内节点间转发.由于篇幅有限,路由策略的具体算法描述、性能分析和仿真另文介绍.

5 RLP2P 网络模型性能分析与仿真

5.1 RLP2P网络模型和路由算法的性能分析

在 RLP2P 网络模型中,每个域的节点数为 $k \le RL_i(X) \le 3k-1$,网络规模为 N.一般地,L_i 层中任意节点与 $L_0,\ldots,L_i$ 层的域中其他节点控制信息的开销为 $O(k)$,域内节点控制开销的上界为 $O(N/k^i)$;最高层节点的控制总开销为 $O(k\log N)$,如图 2 中的 F.当 N 逐渐增大时,网络的平均控制开销可近似表示为

$$ControlOverhead \le \frac{1}{N}\sum_{i=0}^{\log N}\frac{N}{k^i}k \times i = O(K) + O\left(\frac{\log N}{N}\right) + O\left(\frac{1}{N}\right) \Rightarrow O(k) \quad (1)$$

因此,对于任意节点,其控制开销的平均值为 $O(k)$,最坏情况为 $O(k\log N)$.任意两对等点间的交互开销为 $O(\log N)$,当节点交互开销达到 $O(k\log N)$时,则需减少域规模 k 的值.

文献[9]已论证,在具有小世界(small-world)特性的分布式网络中,其搜索算法路由链上界为 $O\left(\log\sqrt{n}\right)^2$,其中 n 为网络节点规模;文献[15]已论证,设网络系统中节点和服务是平均分布的,则 Gossip 算法在高概率下最多经过 $O\left(\log\sqrt{n}\right)^{1+\varepsilon}$ 次消息扩散到达距离其 $\sqrt{n}$ 的任意其他结点.在 RLP2P 网络协议模型中,设域规模相同(节点数为 m),同理可得:消息经 Multilayer Light-Gossip 算法分级扩散,算法路由链上界为 $O\left(\log\sqrt{n/m}\right)^{1+\varepsilon}$.

5.2 协议数据通道性能指标:压力和扩展率分析

5.2.1 模型假设

设分布式网络节点众多,聚集且同构,协议模型由 3 层构成,如图 4 所示.A 为 L_0 层的某一域内的普通节点,B 为该域的中心节点,C 为 L_0 层的另一域的中心节点;在 L_1 层,B,C 属于同一域,且 C 是中心节点,因此,C 位于 L_2 层.对于任意节点 u,与域内其他节点的距离为 $r,r\ge 1$,记为 $V(u,r)$,存在常数 C_1,C_2,且 $C_2 \ge C_1 > 1$,则有:$C_1V(u,r) \le V(u,2r) \le C_2V(u,r)$[16,17].设节点间的相互距离符合 Euclidean 定理,对于一个大规模同构分布式网络节点,RLP2P 协议创建的层的特性相似,域的规模相同,域内节点具有相似的性能和属性,作用半径相同.

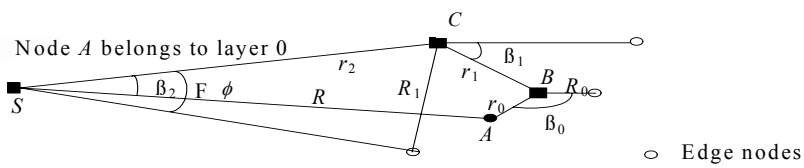

Fig.4 Node catenation for RLP2P model

图 4 RLP2P 网络模型节点连接图

5.2.2 伸展率(stretch)和压力(stress)

如图4所示,设S为源点,从S到节点A的直接单播路径长度为R,而应用层多播路径长度为$r_0+r_1+r_2$,则A的伸展

率为$(r_0+r_1+r_2)/R$,显然,$R=\sum_{i=0}^{2} r_i\times\cos\beta_i$.

一般地,设网络规模为 N,最多分 $M=\log_k N$ 层,任意节点 X,属于 $L_0,L_1,\ldots,L_j$ 层,其伸展率 S_X 为

$$S_X=\sum_{i=j}^{M} r_i \Big/ \sum_{i=j}^{M} r_i\times\cos\beta_i \tag{2}$$

对于“远节点”,在应用层多播数据路径上距离源点 S 第 1 跳段数据通道的最小距离为μR_{M-1}.如图 4 所示,$r_2\geq\mu R_1$,其中$\mu\geq 2$.

证明:设节点 X 属于 L_0 层,且不是中心节点,视为远节点,其伸展率为 $S_{x,f}$.

因为 $0\leq i<M, r_i\leq R_i, \min(\cos\beta_i)=-1; \min(\cos\beta_M)=\cos\Phi, \Phi=\sin^{-1}R_{M-1}/r_M=\sin^{-1}1/\mu, r_M\leq\mu R_{M-1}=2R_0\times k^{(M-1)/2}$,则有

$$\max\left(\sum_{i=j}^{M} r_i\right)=r_M+\sum_{i=0}^{M-1}R_i,\ \min\left(\sum_{i=j}^{M} r_i\times\cos\beta_i\right)=r_M\times\cos\Phi-\sum_{i=0}^{M-1}R_i .$$

所以

$$S_{x,f}\leq\max\left(\sum_{i=0}^{M} r_i\right)\Big/\min\left(\sum_{i=0}^{M} r_i\times\cos\beta_i\right)=\left(r_M+\sum_{i=0}^{M-1}R_i\right)\Big/\left(r_M\times\cos\Phi-\sum_{i=0}^{M-1}R_i\right) \tag{3}$$

取$\mu=2, \Phi=\pi/6$, $R_i=R_{i-1}\sqrt{k}$,则有

$$\sum_{i=0}^{M-1}R_i=R_0(\sqrt{N}-1)\big/(\sqrt{k}-1),$$

带入式(3)化简,得到

$$S_{xf}\leq_{N\Rightarrow P(P\text{是一个很大的数})} 2\left(k-\sqrt{k}+1\right)\Big/\left(k\sqrt{3}-\sqrt{3k}-2\right) \tag{4}$$

显然,当μ取一个大值,使Φ变小,$\cos\Phi\propto 1$,那么其作用范围就会缩小.

对于近节点,其距源点的第1跳段数据通路的距离$\leq\mu R_{M-1}$,用δ表示这些节点的最大伸展率,则网络成员的最大伸展率为$\max(\delta,S_{x,f})$.设N_n,N_f 分别表示“近节点”和“远节点”,其伸展率分别表示为$S_{x,n},S_{x,f}$,则其伸展率的平均值为

$$\overline{S}\leq(N_n\times S_{x,n}+N_f\times S_{x,f})\ /\ N \tag{5}$$

其中$S_{x,n}\leq\delta, N_n/N\leq 1, N_f/N\leq 1$.

同理,平均压力和最大压力分别为

$$\overline{\lambda}\leq\left((1/Q)\times\sum_{i=0}^{M=\log_k Q} Q\times k\times i/k^i\right)=k^2/(k-1)^2+O(\log Q/Q)\leq k^2/(k-1)^2 \tag{6}$$

$$\lambda_{\max}=k\log_k N \tag{7}$$

5.3 数据通道性能模拟仿真分析

网络节点或链路压力和伸展率的关系曲线如图 5 所示.取不同的 k 值,根据式(4)、式(7)得到不同的 $S_{x,f},\lambda_{\max}$ 值.由图 5 可知,伸展率与域的规模 K 或网络规模 N 无关,为一个常数,而最大压力却与域的规模 K 或网络规模 N 成正比,因此,在实际构造网络拓扑时,应根据网络的压力和伸展率进行折衷考虑.

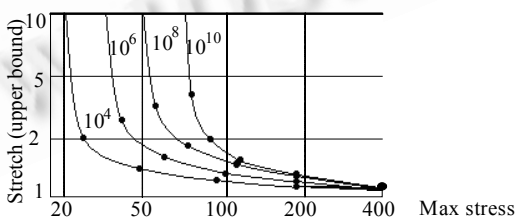

Fig.5 Stretch vs. stress for the RLP2P protocol as the region size is varied

图 5 不同 k 值下网络压力和伸展率关系曲线图

文献[17]对同构集中环境下网络节点或链接的伸展率和压力进行了详细的分析,在域内任意两节点间的伸展率为 $O(1)$,但压力却是一个很大的值 $O(N)$,控制开销为 $O(N^2)$.在 RLP2P 模型中,对网络的压力和伸展率进行综合考虑,使任意两节点间的伸展率为 $O(1)$,压力为 $O(\log N)$,其综合性能优于文献[17]中的算法.

5.4 协议模型仿真

为了分析协议的综合性能,从数据通道质量、鲁棒性和控制开销,将 RLP2P 与 Narada[18]应用层多播协议进

行对比分析.Narada 协议是一个基于 Overlay 技术的分布式应用层多播协议,文献[18]研究表明,该协议具有良好的自组织和负载平衡能力,特别是在多播组的规模较小时,其综合性能相当优秀,是经典的分布式应用层多播协议之一.

仿真环境:联想万全服务器 T100 一台;16 台 PC 机 IBM8624;路由器:HUAWEI Quidway R2501E;交换机:Fast Ethernet ES-3124RL 24-port 10/100Mbps;软件:Linux9.0,Windows2000,JAXT2.0;用 nem[19]产生仿真网络拓扑,拓扑生成算法为 PLOD[20];单独 1 台 PC 用于网络协议和流量分析.提供的服务类型为:新闻、体育、娱乐、音乐、文学、财经 6 类,服务只有标志,没有内容;域的规模 $RL_i(X)$=8~512,k≤256,节点的度为 3~5,网络状态信息更新频率为 5s,TTL=7s,为 3 层逻辑拓扑结构,连接相对稳定,数据源端节点以固定的速率连续产生数据流,数据包高效、有序地在网络中传输,且在给定的生命周期内均能被捕获,节点可随机加入和非正常离开网络.

当图 6、图 7 为 $RL_i(X)$=256 时,RLP2P 协议在链路上产生的压力比 Narada 协议减少 25%,数据通道的伸展率增加 5%;当 Link≥420 时,Narada 协议的链路压力呈直线增加,在 RLP2P 协议中,当 Link≥580 时,链路压力急剧增加.

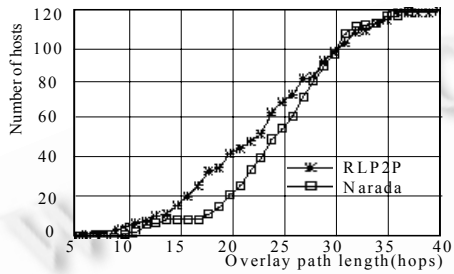

Fig.6 Path length distribution curve

图 6 通道长度分布曲线

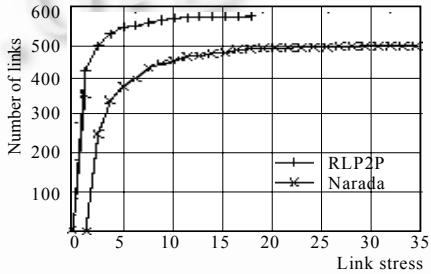

Fig.7 Stress distribution curve

图 7 压力分布曲线

图 8~图 10 是 128 个节点内随机请求加入网络时产生的压力和伸展率,Narada 和 RLP2P 协议分别在 t=400s 和 t=350s 时完全稳定,压力分别为 2.2 和 1.6,通道平均长度分别为 24 和 23;当 t≤200s 时,Narada 协议的压力和伸展率优于 RLP2P 协议;节点成功加入网络,RLP2P 所需带宽≤2Kbps,Narada 所需带宽≤6.5Kbps.

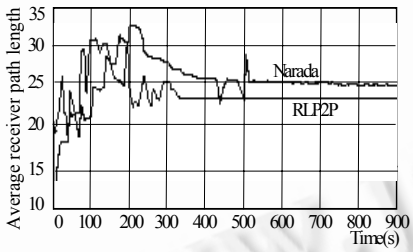

Fig.8 Average path length(stretch)

图 8 平均通道长度(伸展率)

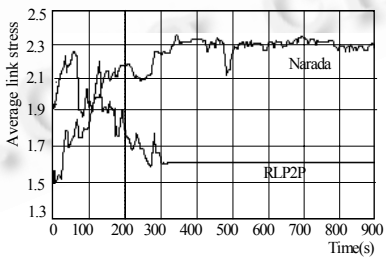

Fig.9 Average link stress distribution curve

图 9 平均压力分布曲线

图 11,图 12 是网络控制开销和延迟,RLP2P 协议在 $RL_i(X)$=128 时,控制开销为 2.8s;Narada 协议在节点数≤200 时,其控制开销优于 RLP2P,当节点数 200~650 时,两者的控制开销相当;当节点数≥650 时,RLP2P 协议的控制开销仅为 Narada 协议的 5%,减少延迟达 30%,Query 消息平均延迟为 8.5×10^{-2}s.图 13 为 N=640,分为 3 层和 8 个域,$RL_i(X)$=80128 个节点随机请求加入和失效离开网络,在节点连续失败时,目的节点正确接收到数据包的模拟曲线图,RLP2P 协议的平均有效性比 Narada 协议优 35%,RLP2P 协议的最大丢包率≤1%.

实验结果表明,当域的规模 $RL_i(X)$≤256 时,Narada 协议模型的性能优于 RLP2P 模型,这与 Narada 协议适用于小规模多播组的特性相符;当域的规模 $RL_i(X)$>256 时,RLP2P 协议的各项性能指标优于 Narada 协议,且网络规模越大,RLP2P 协议越有效.由表 2 可知,网络的各项性能指标处于优良状态,且网络规模越大,其综合性能的优越性越明显.这仅为仿真结论,在实际应用中其各项性能指标可能会有些差异.

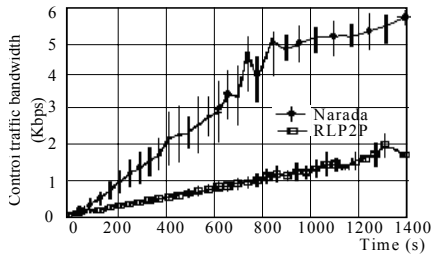

Fig.10 Average control bandwidth required node access

图 10 节点成功加入平均带宽

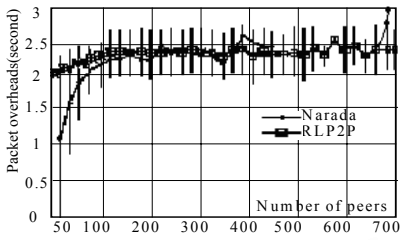

Fig.11 Average control overheads at the nodes

图 11 节点平均控制开销

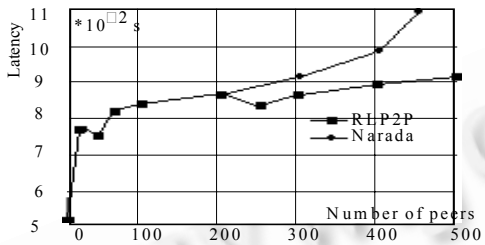

Fig.12 Query latency (varying join/leave rates)

图 12 Query 消息延迟

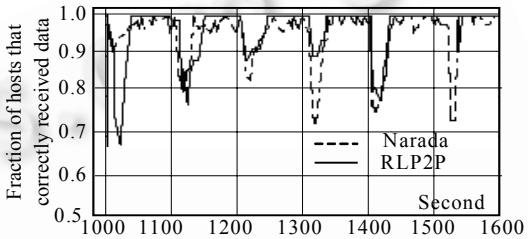

Fig.13 Nodes that received data packets over the duration of member failures

图 13 目的节点正确接收包的概率分布曲线

Table 2 Parameter simulation value of performance

表 2 性能参数模拟值

Group size	Stress		Stretch		Control overhead (Kbps)	
	Mean	Max	Mean	Max	Mean	Max
32	1.85	8.0	1.08	1.61	0.84	2.34
64	1.73	8.0	1.14	1.67	0.77	2.70
96	1.86	9.0	1.04	4.63	0.73	2.65
128	1.90	9.0	1.20	4.85	0.75	2.73

6 总结和进一步的工作

本文提出的 RLP2P 网络模型,在充分利用原 Internet 网络路由节点拓扑结构的基础上,以域和层为基本逻辑管理单位,其域构成自然的“小世界”,整个网络拓扑又符合幂规律特性,因此,RLP2P 网络模型能主动利用小世界特性和幂规律.按用户需求和共享目的组织域,每个结点掌握大量的域内节点服务信息,域内节点间有很大的相识系数,在域内部找到满足服务请求的概率很大,节点出现跨域请求服务的概率很小,使定位某种服务的工作量、查询范围从网络中的所有节点数降低到域内的节点数,并给出了域和层的划分策略和数学模型.由汇聚点 AN 引导节点加入到属性相识的域,避免了节点采用就近加入网络时因属性相异而出现节点域间频繁移动;优化的 Multilayer Light-Gossip 路由算法,消息扩散以分级(域级和域内)方式进行,使得 P2P 网络信息查询定位不再以泛洪扩散方式盲目进行,因而从根本上改善了查询信息的定位搜索效率和网络的可扩展性.量化了压力和伸展率两个重要指标,提出了综合考虑压力和伸展率的思想.当然 RLP2P 网络模型还需要进一步完善和研究,如域的边界 K 的取值范围及相应的约束条件、搜索预分组的开销、访问规则表达式、安全等.

References:

[1] Parameswaran M, Susarla A, Whinston AB. P2P networking: An information-sharing alternative. Computing Practices, 2001,34(7): 31−38.

[2] Napster.http://www.napster.com

[3] Gnutella.http://www.gnutella.com

[4] Jose S. The emergence of distributed content management and peer-to-peer content networks. Gartner GroupInc 2001. http://marketplacena.gartner.com/010022501oth-NextPage.PDF

[5] Zeinalipour-Yazti D, Folias T. A quantitative analysis of the gnutella network traffic. April 2002. http//www.cs.ucr.edu/~csyiazti/courses/cs204/project/html/final.html/

[6] Clarke I, Sandberg O, Wiley B, Hong TW. Freenet: A distributed anonymous information storage and retrieval system. 2004-08-10. http://www.doc.ic.ac.uk/~twh1/academic/papers/icsi-revised.pdf/

[7] Traversat B, Arora A, Abdelaziz M, Duigou M, Haywood C, Hugly J-C, Pouyoul E, Yeager B. Project JXTA 2.0 Super-Peer Virtual Network. 2004-09-20. http://www.jxta.org/project/www/docs/JXTA2.0protocols1.pdf/

[8] Super-Peer Architectures for Distributed Computing. 2004-09-20. http://www.fiorano.com/whitepapers/superpeer.pdf/

[9] Kleinberg J. The small-world phenomenon: An algorithmic perspective. ACM Symp. on Theory of Computing, 2000. 820−828. http://nicomedia.math.upatras.gr/courses/mnets/mat/Kleinberg_SW_algorithmic.pdf/

[10] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the Internet topology. In: Chapin L, Sterbenz JPG, Parulkar G, Turner JS, eds. Proc. of the ACM SIGCOMM'99. New York: ACM Press, 1999. 251−262.

[11] Siganos G, Faloutsos M, Faloutsos P, Faloutsos C. Power-Laws and the AS-level internet topology. 2004-10-05. http://www.cs.ucr.edu/~siganos/papers/SFFF.pdf/

[12] Yang B, Garcia-Molina H. Improving search in peer-to-peer networks. In: Proc. of the 22nd Int'l Conf. Distributed Computing Systems. IEEE Computer Society, 2002. 5−14.

[13] Balakrishnan H, Kaashoek MF, Karger D, Morris R, Stoica I. Looking up data in P2P systems. Communications of the ACM, 2003,46(2):43−48.

[14] Stoica I, Morris R, Karger D, Kaashoek MF, Balakrishnan H. Chord: A scalable peer-to-peer lookup service for Internet applications. In: Proc. of the ACM SIGCOMM (SIGCOMM'01). 2001.149−160.

[15] Kempe D, Kleinberg J. Demers A. Spatial gossip and resource location protocols. In: Proc. of the 33rd ACM Symp. on Theory of Computing (STOC'01). 2001. 163−172.

[16] Plaxton CG, Rajaraman R, Richa AW. Accessing nearby copies of replicated objects in a distributed environment. In: ACM Symp. on Parallel Algorithms and Architectures, 1997. 311−320. http://citeseer.ist.psu.edu/plaxton97accessing.html/

[17] Gupta A. Steiner points in tree metrics don't (really) help. In: Symp. of Discrete Algorithms. 2001. 220−227.

[18] Chu Y-H, Rao SG, Zhang H. A case for end system multicast. IEEE Journal on Volume 20, 2002. 1456−1471.

[19] Magoni D. nem: A software for network topology analysis and modeling. In: Proc. of the 10th IEEE Int'l Symp. on Modeling, Analysis, & Simulation of Computer & Telecommunications Systems (MASCOTS'02). 2002. 364−371.

[20] Palmer CR, Steffan JG. Generating network topologies that obey power laws. In: Proc. of the Global Telecommunications Conf. (GLOBECOM 2000). IEEE, 2000. 434−438.