

基于分群的层状叠加网模型^{*})

刘 丹 李毅超

(电子科技大学计算机学院 成都 610054)

摘 要 为提高叠加网通信效率,提出一种分群自组织层状叠加网模型。以物理距离邻近主机组成的基群为基本自治单位,多个邻近基群自组织为二次群,多个邻近二次群自组织为三次群,依次叠代,形成适应于大规模网络的多次群层状结构。该模型具有逼近物理网结构、自适应网络变化和高效易管理的特点,理论分析证明具有较低通信开销。仿真实验表明,在网络规模较大时其通信负荷只有 Landmark 模型的 30%~50%。

关键词 自组织,叠加网,通信负荷,分群,网络模型

A Hierarchical Overlay Network Model Based on Clustering

LIU Dan LI Yi-Chao

(Computer Institute, University of Electronic Science and Technology of China, Chengdu 610054)

Abstract To improve the communication efficiency of overlay network, a hierarchical self-organizing overlay network model based on clustering is proposed. In this model, basic clusters which are composed of a set of adjacent computers are the mini-autonomous units. Several adjacent basic clusters form a quadric cluster and several adjacent quadric clusters form a cubic cluster and so forth. Then a multi-hierarchical clustering structure which can suit the large-scale network is produced. The model has the following characteristics: manageable, approaching the physical network which is based on, and adapting frequent changes of the physical network. Finally, theoretical analysis prove it has lower communication load. The simulated result shows, while the network is large enough, the communication load in this model is only about 30%~50% of the one in a Landmark model.

Keywords Self-organizing, Overlay network, Communication load, Clustering, Network model

1 引言

网络的逻辑组织通常按照功能来划分和实现,由此现实中产生了很多按服务聚类的应用层叠加网,如网格计算中的资源分布网、多媒体分布网、应用级路由及多播拓扑网等。叠加网中,节点由主机构成,除具有存储转发功能外,还能实现更复杂的算法,如发现服务、编解码、存储服务。但叠加网的路由性能却不如物理网高效。物理网具有物理距离、链路带宽可测的优势,而叠加网只能通过一些外部测试来获得不精确的网络物理拓扑特征。一个高效叠加网的构建应充分反映其所依托的下层物理网特征,以减小将消息转达目标主机的通信距离。如何构建一个良好的叠加网,以赢得高效性和实用性,是本文研究目标。

使叠加网拓扑逼近物理网是提高效率的主要方法,具有代表性的方法通常是通过洪泛搜索、启发式搜索和 Landmark 群集等技术来实现。洪泛搜索通过源节点消息的大规模扩散来寻找物理距离最近目标节点,定义物理距离近的节点为源节点逻辑近邻。为降低洪泛搜索的网络负荷及扩散深度,提出了启发式算法,如 hill climbing^[1]。但启发式算法依赖于局部信息,不能准确定位最近邻节点。Landmark 群集^[2,3]策略中,各节点测量出其对 m 个众所周知的 Landmark 节点的物理距离,形成 m 维距离矢量, m 维距离矢量接近的节点被划为同一集群。该策略是一种粗粒度逼近,在最坏情况下可能

导致绝大部分节点被划分到一个集群中。Xu^[4]提出一种综合 Landmark 策略和启发式算法来判断近邻特征的方法,但它不能很好适应结构动态变化的叠加网,如无线移动网。Zhang^[5]提出一种基于分组的两层拓扑逼近方法,其缺点是对于大尺度网络两层拓扑效率不高且节点加入组的收敛速度较慢。

本文给出一个分群自组织层状叠加网模型。模型以包含 m 个物理距离邻近主机的基群为叠加网基本单位,基群的上级单位为二次群,随网络规模的增加出现三次、四次以至 n 次群。每个 n 次群内最多包含 m 个物理邻近的 $n-1$ 次群,网络规模越大,顶级群数值越高。各次群根据叠加网规模的动态变化及网络物理拓扑的全局和局部距离特征自动产生、分裂与消亡,以动态适应网络变化。各次群以物理距离内聚,使逻辑拓扑逼近物理拓扑,提高网络通信效率。

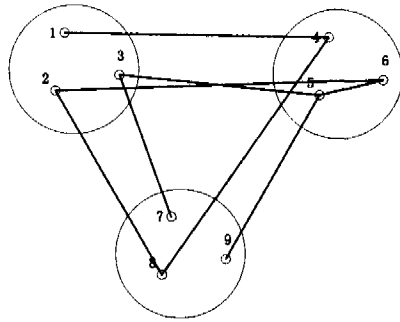
本文首先给出模型描述及形成过程,进而在理论上对模型的效果进行分析和证明,并用 BRITE^[6]生成 BA(Barabasi-Albert)^[7]和 ER(Erdos & Renyi)^[8]模型的物理网,在此基础上进行仿真实验。理论分析和仿真实验表明,分群自组织层状叠加网模型具有良好的本地特征和自治特征,大幅度提高了网络效率。

2 自组织模型

2.1 基本概念和定义

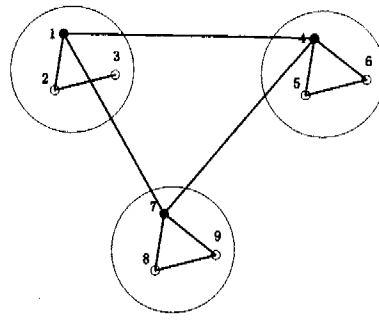
^{*})四川省科技攻关项目(02GG006-018)资助项目。刘 丹 讲师、博士研究生,主要研究方向:分布式计算,计算机网络及网络安全;李毅超 副教授,主要研究方向:计算机网络及网络安全。

叠加网由各终端主机构成。各主机运行应用层协议相互通信,通常各主机维护着一组与其通信的主机集合信息。若两台主机由于应用层通信目的建立了一个应用层连接,即为逻辑邻接点。叠加网中数据的传输不如物理网高效,如随机



(a) 随机组织的叠加网

构造一个叠加网,两个逻辑邻接主机在物理网中可能相距甚远。若在叠加网内互为邻接的主机在物理网上也是邻接点,则叠加网的通信开销会大幅度降低。如图 1(a) 中两主机间的平均距离明显高于(b)。



(b) 逼近物理网拓扑的叠加网

图 1

定义 1 物理距离 d_{ij} 通常节点间网络距离定义可以是往返时间 d_{RTT} 、链路上的最小带宽 d_{band} 或用户自定义距离函数 d_{fun} 。本文以叠加网中两主机间往返时间为距离度量,即 $d_{ij} = d_{ij-RTT}$ 。

定义 2 节点逻辑距离 L_{ij} 两节点 i, j 的逻辑距离定义为 i, j 间逻辑跳数,即 i 发送消息到 j 需要经过叠加网内的主机数+1。

定义 3 网络几何中心节点 在由 n 个节点组成的网络 X 中, $\bar{d}_i$ 表示节点 i 对于其它 $n-1$ 个节点物理距离的几何平均,即 $\bar{d}_i = \frac{1}{n-1} \sum_k d_{ik}$, 其中 $k \in X, k \neq i$, 定义 $\bar{d}_i$ 最小的节点为 X 的几何中心节点。

定义 4 伸展度 R_i 描述叠加网和物理网匹配程度,即叠加网上单位逻辑距离对应于物理网平均距离为 R_i 。当 $R_i = 1$, 则叠加网完全匹配于物理网。

2.2 分群自组织

网络中各主机服务以 WSDL 语言描述^[9], 系统为每种服务类型分配唯一标识(grid service ID), 该标识基于行业标准分类法分类。将相同服务的所有主机划为同一叠加网。各叠加网以包含 m 个主机的基群为基本单位, 基群以特征距离聚类生成。特征距离综合反映节点服务质量、物理距离等。为简化讨论, 设服务质量已按某种方法换算为距离, 下文统一以物理距离 d_{ij} 来形成群。

下面描述某一叠加网形成过程。 m 个距离邻近的主机构成一个基群, m 为群尺寸上限。每个基群有一个群首, 负责和其它基群群首交换信息; 一个二次群包含 $1 \sim m$ 个基群群首, 每个二次群由其二次群群首负责和其它二次群群首交换信息。依次类推, 一个 n 次群包含 $1 \sim m$ 个 $n-1$ 次群群首, 每个 n 次群由其 n 次群群首负责和其它 n 次群群首交换信息。属于不同 n 次群内的两个非 n 次群群首相互无直接消息交换。各次群群首记录该群的群信息, 群 δ 信息主要包括: δ 所包含的节点数即群尺寸 $|\delta|$ 、群内成员及成员间的相互距离信息 $d_{ij}, i, j \in \delta$ 。若 n 次群中的 $n-1$ 次群群首的成员数达到 m , 则该群饱和。设叠加网中有一组主机为信息交换节点(SP), SP 上记录叠加网中各次群群首信息及最高次群群首间的相互距离。所有新主机都知道 SP 位置。

群内各成员周期性探测计算与其它对等成员距离, 各次群尺寸上限 m 的选取以最大限度降低通信开销为前提; m 过

大导致群内维护开销过多, 反之引起叠加网层数过多, 层间、群间组织维护开销变大。下面详细描述各次群的形成及维护。

2.2.1 基群维护

基群形成: 当节点 i 新加入, 设其服务类型为 type- k 。

- 1) i 从 SP 中获取所有服务类型为 type- k 的基群群首集 Ω_i 。
- 2) 若 $\Omega_i = \emptyset$, 转 5; 否则找出 Ω_i 中与 i 物理距离最近的基群 δ_l 的群首 l ,
 $d_{il} = \min(d_{ij}), j \in \Omega_i, j \neq i$ 。
 a) 若 l 所辖基群 $|\delta_l| < m$, 则 i 加入到 δ_l 中, $|\delta_l| = |\delta_l| + 1$, 更新 l 所维护的群信息, 结束。
 b) 若 l 所辖基群 $|\delta_l| = m$, 则该群饱和, i 加入引起群分裂, 转 3。
- 3) l 向群 δ_l 内所有节点发送群分裂报文, 表示群 δ_l 将分裂为两个基群: 基群 A 和基群 B。选择群 δ_l 中物理距离最远的节点对为两个参考点, 不妨设为 a 和 b 。对群 δ_l 中任意节点 c , 若 $d_{ca} < d_{cb}$, 则 c 划归于群 A, 否则归属于群 B。
- 4) 分别选择群 A 和群 B 中的几何中心节点 l_a 和 l_b 为两个新基群群首。更新 SP 中信息及 l_a 和 l_b 中群信息, 结束。
- 5) i 申明为一个新基群 δ_i 的群首, $|\delta_i| = 1$ 。更新 SP 中信息, 结束。

基群消亡: 当节点 i 注销其提供服务 type- k , i 将从其所属基群中退出。

- 1) 若 i 为基群 δ_i 群首,
 a) $|\delta_i| > 1$, 选择 δ_i 中除 i 外的几何中心节点 j 为新群首, $\delta_i = \delta_i - i$ 。将群信息从 i 复制到 j , $|\delta_i| = |\delta_i| - 1$, 结束。
 b) $|\delta_i| = 1$, 基群 δ_i 消亡, 更新 SP 中信息和 δ_i 所属二次群群首中的群信息, 结束。
- 2) 若 i 不为基群群首, 设其属主基群 δ_j , 则 $|\delta_j| = |\delta_j| - 1$, 更新该基群群首的群信息, 结束。

2.2.2 n 次群维护

n 次群的产生与分裂: 随着网络规模的扩大, 出现两种情况:

- 1) 设网络的顶级群为 $n-1$ 次群。当 SP 发现 $n-1$ 次群群首数量超过 m 时, 向所有 $n-1$ 次群群首发送首次生成 n 次群报文, 表示将同时生成两个 n 次群: 群 A 和群 B。选择物理距离最远的 $n-1$ 次群群首对为两个参考点, 不妨设为 a

和 b 。对任意 $n-1$ 次群首 c , 若 $d_{ac} < d_{bc}$, 则 c 划归群 A, 否则划归群 B。分别选择群 A 和群 B 中的几何中心节点 l_a 和 l_b 为两个新 n 次群首。更新 SP 中信息和新群首 l_a 和 l_b 中的群信息。

2) 当某 n 次群首 i 发现其所辖群 δ_i 中 $n-1$ 次群首数量超过 m 时, 向 δ_i 内所有 $n-1$ 次群首发送群分裂报文, 表示 δ_i 将分裂为两个 n 次群: 群 A 和群 B。选择 δ_i 中物理距离最远的 $n-1$ 次群首对为两个参考点, 不妨设为 a 和 b 。对群 δ_i 中任意 $n-1$ 次群首 c , 若 $d_{ac} < d_{bc}$, 则 c 划归群 A, 否则划归群 B。分别选择群 A 和群 B 中的几何中心节点 l_a 和 l_b 为两个新 n 次群首。更新 SP 中信息和新群首 l_a 和 l_b 中的群信息。

n 次群的合并: 随着网络中某应用服务减少, 该叠加网规模相应缩减。若某 $n+1$ 次群首检测到其所包含的两个 n 次群的成员总数 $x < m-r$ 时, r 为一个常量, 且根据两群首距离判断其为距离最近的群, 则这两个群合并为一个 n 次群, 并选择新群成员中的几何中心节点为新 n 次群首。

本模型中, n 次群首节点仅负责维护 $n-1$ 次群的群信息, 无须维护全局信息及其上、下层信息, 使叠加网维护开销和定位复杂性大幅降低, 同时模型具有分布式技术伸缩性好的特征, 能很好地适应动态变化的网络拓扑。

3 性能分析

定理 1 按上节形成分群自组织层状叠加网, 在一个规模为 N (总节点数为 N) 的网络中, 有 $\log_m N - 1 \leq h \leq \log_m N$ 。 h 为叠加网中顶级群编号, 表示叠加网层数, m 为叠加网群尺寸上限。

证明: 用 N_n 表示 n 次群的数目, 由于顶级群为 h , 有 $1 \leq N_h \leq m$ 。

$$N_1 = \frac{N}{m}, N_2 = \frac{N_1}{m} = \frac{N}{m^2}, \dots, 1 \leq N_h = \frac{N}{m^h} \leq m, \text{ 有}$$

$$\log_m N - 1 \leq h \leq \log_m N \quad (1)$$

证毕。

下面分析比较本分群叠加网和随机生成叠加网的通信效率。

设这两种叠加网依托于相同物理网, 物理网中节点间平均物理距离为 $\bar{d}$, 网络规模为 N 。为便于分析, 设网络中两节点间平均通信量是常量, 即通信开销和主机数对呈正比关系。

对于基群, 大小为 m , 设基群内主机以随机方式建立连接, 节点平均度为 k , 基群内节点平均逻辑距离 $L_1 = \log_m m^{[10,11]}$ 。由定义 4, 平均物理距离 $d_1 = R_s \log_m m$ 。设基群内通信量为 γm^2 , γ 为常量。基群内部通信负荷 $C_1 = r m^2 d_1 = \gamma m^2 R_s \log_m m$ 。

对于包含单个二次群规模的饱和网络, 包含 m 个基群, 群内通信负荷为 $m C_1$ 。相邻基群群首间的平均物理距离为 $2 d_1$, 两基群群首间的平均物理距离为 $d_2 = 2 d_1 \times \log_m m = 2 R_s (\log_m m)^2$, 群间通信量为 γm^2 , 总通信负荷

$$C_2 = m C_1 + d_2 \times \gamma m^2 = R_s \gamma m^3 \log_m m + 2 R_s \gamma m^2 (\log_m m)^2$$

依此类推, 对于包含单个 h 次群规模饱和网络的总通信负荷为:

$$C_h = \gamma R_s (m^{h+1} \beta + 2 m^h \beta^2 + \dots + 2^{h-1} m^2 \beta^h),$$

$$\text{其中 } \beta = \log_m m$$

$$\Rightarrow C_h = \gamma R_s m^{h+1} \beta \left[1 + \frac{2\beta}{m} + \dots + \left(\frac{2\beta}{m} \right)^{h-1} \right]$$

$$\Rightarrow C_h = \frac{R_s (\gamma m^{h+2} \beta - 2^h \gamma m^2 \beta^{h+1})}{m - 2\beta}$$

将 $\beta = \log_m m$ 代入

$$\Rightarrow C_h = R_s \frac{\gamma m^{h+2} \log_m m - 2^h \gamma m^2 \log_m^{h+1} m}{m - 2 \log_m m}$$

在分群自组织层状叠加网形成过程中, 各物理距离邻近的节点才被划分到同一群中, 因此叠加网拓扑基本反映了物理网拓扑结构, 两节点逻辑路径和物理路径基本吻合, 有:

$R_s \rightarrow 1$, 即

$$C_h \approx \frac{\gamma m^{h+2} \log_m m - 2^h \gamma m^2 \log_m^{h+1} m}{m - 2 \log_m m} \quad (2)$$

对于包含 $N = m^h$ 的非分群随机叠加网, 其平均逻辑距离为 $h \log_m m$, 任一逻辑跳对于物理网拓扑是随机的。由于每逻辑跳对应的平均物理距离为 $\bar{d}$, 随机叠加网中两节点的平均物理距离为 $d_{mn} = \bar{d} h \log_m m$, 网内通信量为 $\frac{\gamma}{2} (m^{2h} - m^h)$ 。当 m, h 较大时, 通信负荷主要依赖于首项指标, 有

$$C \approx \frac{\gamma}{2} h \bar{d} m^{2h} \log_m m \quad (3)$$

由 (2)、(3) 两式, 有

$$\frac{C}{C_h} \approx \frac{\bar{d}}{2} \times \frac{h m^{2h} - 2 h m^{2h-1} \log_m m}{m^{h+1} - (2 \log_m m)^h}$$

当 m, h 较大时, 有

$$\frac{C}{C_h} \approx \bar{d} \frac{h m^{2h}}{m^{h+1}} = \bar{d} h m^{h-1}$$

由 (1) 式得

$$\frac{C}{C_h} \approx \bar{d} h m^h \leq \bar{d} m N \log_m N \quad (4)$$

可见, 本分群自组织层状叠加网拓扑比随机生成的叠加网拓扑提高了近 $\bar{d} m N \log_m N$ 倍的通信效率。对于规模很大的网络, 分群自组织层状叠加网的平均物理距离比随机叠加网大幅缩短, 而平均物理距离反映了拓扑维持开销和洪泛搜索开销等多种操作的性能, 即网络负荷的大幅降低。

4 仿真实验

BRITE 生成常用来仿真大型网的二种网络模型——BA^[7] 和 ER^[8] 模型, 作为叠加网所依托的物理网进行对比实验。BA 模型具有小平均距离、大簇系数和 Power-law 度分布三个特征, 是现实网络至少应具备的, 而 ER 模型仿真随机网络拓扑。两种拓朴生成参数, BA 模型中 $m=2$, ER 模型中 $\alpha = \beta = 0.2$, 具体参见文[6]。

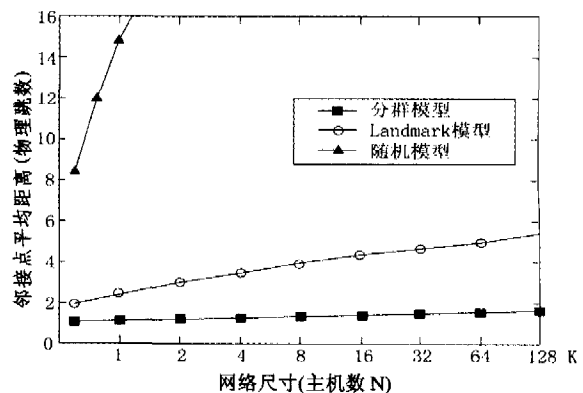


图 2 物理网络为 ER 拓朴模型

在相同物理网基础上 (BA 或 ER), 按照分群自组织模型

(下转第 34 页)

性要求,用 QoRS 指标 D_k 约束,使得节点 $NODE_i$ 的周期性消息流 CM_k 的实时性得以保证的充分必要条件为:

$$N \times T_{\max} \leq D_k \quad (5)$$

整个测控网络系统周期性通信的实时性得以保证的充分条件为:

$$N \times T_{\max} \leq D_{\min} \quad (6)$$

(6)式中 $D_{\min} = \{D_i, i = 1, 2, \dots, k; k \in N; \text{并且 } D_{\min} \text{ 为最小值}\}$,根据(6)式可以进一步确定整个测控网络系统中的最大节点数 $N_{\max}$,即:

$$N_{\max} = D_{\min} / T_{\max} \quad (7)$$

从(5)~(7)式可以看出:用 QoRS 来分析测控网络实时系统时,可以分析现有的系统的实时性,评价其是否满足系统的实时性要求;这种分析方法也可以用于测控网络实时系统的设计中,通过系统的实时性要求来选择通信网络及协议,通信速率和系统的节点数。

结论 本文提出的实时服务质量概念及指标体系对分布式测控网络这一类实时系统的分析与实现具有指导意义,

建立的实时服务质量函数 QoRS(E)的数学模型是对实际工程应用中实时系统的服务质量的高度概括和抽象。通过对 QoRS 的分析研究,阐述了 QoRS 在测控网络分析中的实际应用。今后的研究工作是进一步研究 QoRS 指标体系的细节问题,用于分析研究实时系统中的实时消息调度,以研究解决实时系统的关键技术。

参考文献

- 1 Chou C C, Shin K G. Statistical Real-Time Channels on Multi-access Bus Networks. IEEE Transaction on Parallel and Distributed Systems, 1997, 7(8): 769~780
- 2 Bernat G, Burns A. Weakly Hard Real-Time Systems. IEEE Transaction on Computer, 2001, 50(4): 308~33
- 3 Crawley E, Nair R, Rajagopalan B, Sandick H. A framework for QoS-based routing in the internet. IETF RFC2386, August 1998
- 4 Krishna C M, Shin K G. Real-time Systems. McGraw-Hill Companies, Inc. 1997
- 5 Bernat G, Burns A. New Results on Fixed Priority Aperiodic Servers. Procs. of IEEE Real-Time Systems Symposium, 1999, 68~78

(上接第 30 页)

($m=40$)、Landmark 模型和随机模型构造叠加网,每个叠加网的主机数和边数相同。图 2 和 3 是物理网分别是 ER 和 BA 模型的叠加网平均物理距离比较。为简化实验,叠加网中两逻辑邻接主机距离以两主机间物理网的连接跳数代表。

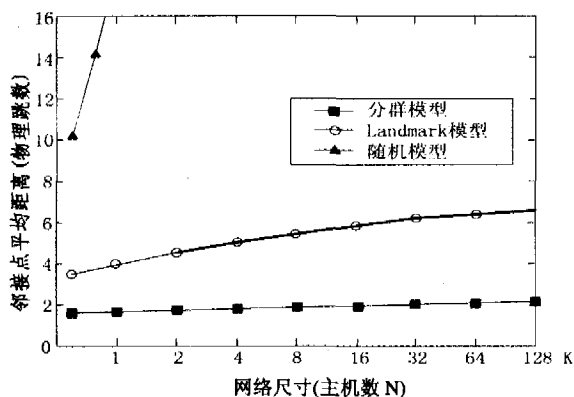


图 3 物理网络为 BA 拓扑模型

由仿真结果可以看出,无论在 BA 或 ER 物理网拓扑下,随机模型的平均物理距离都大大高于分群模型和 Landmark 模型,而分群模型的平均物理距离大约为 Landmark 模型的 30%~50%。可见,分群模型赢得更短的平均物理距离,降低了网络开销。

结论 本文研究了如何自组织一个良好的叠加网,以赢得高效性和实用性。提出了一种分群自组织层状叠加网模型,模型的形成过程以节点间 RTT 作为成群的距离标准,以逼近物理网结构、减小叠加网通信开销。动态的群生成、合并、分裂与消亡使该模型能很好地适应网络变化。模型本质上采用了分层、分布技术,结构上高效且易管理。本文从理论上分析得出该模型比随机模型的通信效率高出近 $d \log_m N$ 倍。通过在以 BA 和 ER 模型物理网基础上的仿真实验结果表明,在相同物理网前提下,该模型的通信负荷只有 Landmark 模型的 30%~50%。

许多基于广域网的大规模网络应用服务,如流媒体服务、

应用层多播服务、分布式搜索及定位等,都可以基于此模型构建相应叠加网,以提高网络效率。针对不同应用网络,可设计不同物理距离函数,并非单纯以 RTT 作为距离度量尺度。

参考文献

- 1 Waldvogel M, Rinaldi R. Efficient Topology-Aware Overlay Network [C]. ACM HotNets 2002, SIGCOMM/CCR, 2003
- 2 Ratnasamy S, Handley M, Karp R, et al. Topologically-aware overlay construction and server selection [C]. INFOCOM 2002, Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. In: Proceedings, IEEE, Volume 3, 23-27, 2002. 1190~1199
- 3 Ratnasamy S, Francis P, Handley M, et al. A Scalable Content-Addressable Network [C]. In: Proceedings of SIGCOMM, 2001. 161~172
- 4 Xu Zhichen, Tang Chunqiang, Zhang Zheng. Building topology-aware overlays using global soft-state [C]. Distributed Computing Systems, 2003. In: Proceedings 23rd International Conference on, 2003. 500~508
- 5 Zhang Xin-Yan, Zhang Qian, Zhang Zhen-sheng, et al. A constellation of locality-aware overlay network: mOverlay and its performance Selected Areas in Communications [J]. IEEE Journal on, 2004, 22(1): 18~28
- 6 BRITE. [Online]. Available: <http://www.cs.bu.edu/brite/>
- 7 Barabasi A-L, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286: 509~512
- 8 Erdos P, Renyi A. On random graphs [M]. I Publ Math, 1959, 6: 290~297
- 9 Christensen E, Curbera F, Meredith G, et al. Web services description language (WSDL) 1.1. W3C (World Wide Web Consortium) [Online]. Cambridge, MA, 2001 <http://www.w3.org/TR/wsdl>
- 10 Béla B. Random graphs. London-New York: Academic Press Inc, 1985
- 11 Watts D J, Strogatz S H. Collective dynamics of "small world" networks [J]. Nature, 1998, 393: 440~442