P2P 网络中 Netshot 路由算法的消息通信机制性能分析*)

向学哲

(华南师范大学计算机学院　广州 510631)

摘　要　在基于对等网络的 Netshot 路由模型的构架下，通过选用节点间不同的组织方式来形成不同的网络形态，对这些网络结构中节点的加入、删除、消息通讯等一系列操作的开销进行推导和验证。对在 P2P 方式下，节点间不同的连接方式带来的性能差异进行全面的分析、比较。同时对节点出错时系统可靠性和开销进行分析和讨论。

关键词　P2P，Netshot，路由模型，消息通信

Communication Performance Research about Netshot Routing Algorithm in Peer-to-Peer Network

XIANG Xue-zhe

(College of Computer Science, South China Normal University, Guanzhou 510631, China)

Abstract　According to the basic structure of the Netshot , we organize the nodes of the network in different ways to form several variations of the network. Then we formulate and simulate the costs of some basic operations to have a general analysis . We also discuss the reliability and cost when the peers make failure.

Keywords　Peer-to-Peer, Netshot, Routing model, Message communication

1　引言

P2P 是继 C/S 架构后新兴的网络应用模式。在传统的 C/S 架构应用系统中，客户端与服务端有明确分界，常常发生客户端能力过剩、服务端能力不足或网络拥塞的现象。P2P 系统中的使用者能同时扮演客户端及服务器多重角色，任意两个使用者之间能不通过服务器直接进行信息分享或内容交换，以构建自主、开放、异构特性的分布式网络应用系统。

NetShot 路由算法是 Barnet 系统中所采用的逻辑网络拓扑结构和路由模型。该模型针对 P2P 网络的动态特性为 Barnet 系统提供一个可扩展的动态无冲突节点命名方式、节点间邻接关系、定位以及查找模型。系统为每一个节点指定唯一的逻辑标识，并通过路由表、引入表来确定节点之间的邻接关系，节点利用这种邻接关系向其它节点传递消息，一个消息通过多个节点的协同传递后到达目标节点，实现任意节点之间的通信。

为实现系统中任意节点之间的互相通信，必须为系统中的每一个节点指定一个全局唯一的逻辑标识，实现从逻辑节点到物理节点之间的映射。在 NetShot 路由模型中节点的全局命名空间 GNS(Global Name Space)是大小为 1 的[0,1)数值空间。

当一个新节点要加入 NetShot 时，系统通过将系统中某一个节点的 PNS 分割来为该节点分配一个新的 PNS。新节点首先在[0,1)之间随机均匀产生一个固定长度(例如：128 位)的 0,1 字符串，将这个字符串所对应的数值作为该节点的目标地址。接下来这个新加入的节点至少找到系统中的一个节点作为其引导节点(Bootstrap Node)，并向这个节点发出一个加入请求消息，引导节点根据情况将加入节点请求消息前递(Forward)，直到目标地址所在的节点。目标地址所在的节点将自己的 PNS 均匀地划分为两个部分，自己保留不变界所在的那一部分，并将另一部分作为新加入节点的 PNS。

当一个节点要离开 NetShot 时，首先向与其不变界相邻的节点发送一个离开请求消息，相邻的节点接收到消息后，将离开节点的 PNS 加入自己的 PNS 中。

由于在系统中存在着大量的节点，开销和可扩放性的需要，任何节点都不可能与所有节点保持邻接关系。只能与部分节点保持邻接关系，在消息通信中，必须依靠其它节点的传递来完成不同节点之间大量的通信任务。在 NetShot 中每个节点的 PNS 都和另两个节点的 PNS 邻接，从而每个节点都和相应的那两个节点保持邻接关系。与其它节点之间的邻接关系通过路由表和引入表来维护。

本文对于 NetShot 路由模型的性能进行分析和仿真。在基于对等网络的 Netshot 路由模型基本构架下，通过选用不同的方式来组织不同的网络形态，并针对不同形态下节点的加入、删除、消息通讯等操作的消息通信开销进行推导并进行模拟验证。对在 P2P 方式下，节点之间不同的连接方式带来的性能差异有一个完整的分析和比较。

2　消息通信开销和性能

根据节点之间的邻接关系，Netshot 中的节点与系统中一部分节点保持邻接关系，节点之间的通信基于这种邻接关系采用消息前递的方式进行。当节点需要向某一目标节点发送消息的时候，节点首先检查其路由表以及引入表中是否有目标节点。如果有，就直接发送给目标节点；否则，在路由表项和引入表项中选择与目标数值最接近的路由表项所对应的节点，并将消息发送给该节点，该节点接收到消息后同样根据上述规则继续进行前递。节点间不断接力，直到消息传递到目标节点。

在不同进位制 R 下，节点间消息通信的开销有更细致的认识，从而能对进一步的研究和实际应用有所裨益，就这个问题进行理论推导。

首先进行一个空间映射，将 GNS 从[0,1)映射到[0,R^n)($R^n=N$)，那么，节点之间判定邻接的步长应该从 R^{-k} 变为 R^k。对于最坏情况和平均情况分别进行讨论。

*)广东省自然科学基金资助项目(项目编号 06300907)。向学哲　硕士研究生，讲师，研究方向为计算机网络新技术、分布式计算。

2.1 最坏情况下的消息通信开销

对于单向路由表的情形，设两个节点的不变界 UCB 之差为 Dif，若 Dif 第 I 位的数值是 k，则需经过 k 次距离为 R^{I-1} 的消息传递。当整个 Dif 各个数位取值全是 $R-1$ 时，消息通信开销最大，为 $(R-1)*\log_R N$。

对于双向路由表的情形，由于消除了边的有向性，最大距离不会超过单向路由表情况下的一半，即 $\frac{(R-1)*\log_R N}{2}$。

2.2 平均情况下的消息通信开销

在模型中，是否采用双向路由表，对消息通信的开销影响很大。同时，选用不同的进位制 R，对消息通信的开销也有一定的差异。在单向路由表情况下，消息通信开销性能如下。

为了考察各点间的距离，以标度为 0 的点 O(UCB=0)为基准，所有节点与点 O 通信的平均开销就是整个网络中任意两点间通信的平均开销。

(1)若 $R=2$

设 $N=2^n$，节点 B 到点 A 的距离等于点 B 对应的二进制序列中 1 的个数(若第 I 位是 1，则 A 到 B 的路径上有一条长度为 R^{I-1} 的边)。转而考虑全体 n 位长的二进制序列中，有多少个 1。在 n 位中有 k 个 1 的序列有 C_n^k 个，所有序列中共有 $\sum_{i=1}^{n} iC_n^i$ 个 1，根据组合公式 $\sum_{i=1}^{n} iC_n^i = n*2^{n-1}$，得平均开销为 $\frac{n}{2}$。

(2)$R>2$

设 $N=R^n$，节点 C 到节点 A 的距离等于点 C 对应的二进制序列中 1 的个数(若第 I 位是 k，则 A 到 B 的路径上有 k 条长度为 R^{I-1} 的边)。转而考虑全体 n 位长的 m 进制序列的数码之和，在 n 位序列中任意考察第 j 位，第 j 位的值取 p 的序列共有 R^{n-1} 个，那么易得第 j 位之和是 $\frac{R(R-1)}{2}*R^{n-1}$，那么长度为 n 的序列的所有数码和是 $\frac{R(R-1)}{2}*R^{n-1}*n$，平均开销是 $\frac{(R-1)}{2}*\log_R N$。

在双向路由表情况下，仍以标度为 0 的点 O 为基准，与 a 中情形相似，所有节点与点 O 通信的平均开销就是整个网络中任意两点间通信的平均开销。

(1)若 $R>2$

① $R=2k(k\geqslant 2)$。对任意给定的 n 位 R 进制序列进行分析，对于第 I 位数字 p，在最终分拆中，或者包含 p 条长度为 R^{I-1} 的边，或者包含 $(R-p)$ 条长度为 R^{I-1} 的边和 1 条额外的长度为 R^I。因此，对所有的数字做分类，定义函数 $f(t)$，$(0\leqslant t\leqslant R-1)$，$f(t)=\begin{cases} t, t\leqslant k \\ R+1-t, t>k \end{cases}$，函数 $F(A)$，A 为一个给定的 R 进制序列，$F(A)=\sum f(t)$，$(t\in A)$，$F(A)$ 的值是序列 A 所代表的点到点 0 开销的近似解。事实上，由于第 $I-1$ 位的值可能对第 I 位的值产生影响(如当第 I 位的值是 k，而第 $I-1$ 位的值大于 k)，所以得出的只是近似解，下面进行修正。

若第 I 位的取值是 k，且后一位的值大于 k，此时应将值从 $k+1$ 修正为 k；

若第 I 位的取值是 $2k-1$，且后一位的值大于 k，应将值减去 $2k$。

$$F'(A)=\sum f(t)-\frac{(n-1)*R^{n-1}}{2N}-\frac{(n-1)*R^{n-1}*R}{2N},\quad (t\in A)$$

遍取所有的序列 A_i，求得 $F(A_i)$ 的平均值再减去修正项即为所求距离。在 n 位序列中任意考察第 j 位，第 j 位的值取 p 的序列共有 R^{n-1} 个，那么第 j 位之和是 $R^{n-1}*(\sum_{i=0}^{k} i+\sum_{i=k+1}^{2k-1}(R+1-i))=R^{n-1}*(k^2-1)\approx\frac{RN}{4}$，所有 n 位之和是 $n*\frac{RN}{4}$，平均距离是 $\frac{nR}{4}-\frac{n-1}{2R}-\frac{n-1}{2}$。

② $R=2k-1$。可以完全平行地做出推导，结论同上，平均距离约为 $\frac{nR}{4}-\frac{n-1}{2R}-\frac{n-1}{2}$。

(2)若 $R=2$

定义函数 $g(A)=2x+y$，其中 x 是序列 A 中长度大于 1 的全 1 串，y 是序列中单独的 1 的个数。对于在第 I 位的单独的 1，那么在最短路径中必定存在长度为 2^{i-1} 的边；相应地，对于第 j 位到第 k 位的全 1 串，可以通过 $2^{j-1}-2^k$ 来得到，即在最短路径中只占了两条边。综上所述，函数 $g(A)$ 的值就是序列 A 所对应的点到点 O 的距离。定义函数 $G(n)$，$G(n)=\sum f(A_n)$，即 $G(n)$ 是所有长度为 n 的二进制序列 A 对应 $f(A)$ 值的总和，$\frac{F(n)}{2^n}$ 是平均水平下二点的距离，下转求 $F(n)$。按照序列中第 1 个“10”子串的位置进行分类。根据分类讨论可得到递推公式：

$$F(n)=\sum_{i=2}^{n}((2i-3)*2^{n-i})+(i-1)*F(n-i))+2n-1$$

$$=\frac{3n*2^n}{8}+\frac{2^n}{4},\frac{F(n)}{2^n}=\frac{1}{4}+\frac{3n}{8}$$

即当 $R=2$ 时，平均开销是 $\frac{3}{8}\log_2 N$。

3 节点出错时系统可靠性分析和消息通信开销

假设每个节点的出错概率为 p，NetShot 系统中节点的总数量为 N，当任意一个节点的所有路由表和引入表中的节点都出错的时候，整个 NetShot 系统就被分成了两个不可达的部分。某个节点的所有路由表和引入表中的节点都出错的概率是 $pr+i$。令 F 为 NetShot 系统出错的概率，$F<\sum pr+i<Npr+i$，整个 NetShot 路由系统的可靠性为 $1-Npr+i$。

当 $N=1,024$，$r=i=\log N$，$p=1/2$ 时，NetShot 路由系统的可靠性为 99.9023%。

当 $N=1,000,000$，$r=i=\log N$，$p=1/2$ 时，NetShot 路由系统的可靠性为 99.9999%。

假如每个节点维护着一个大小为 r 的路由表和一个大小为 i 的引入表。每个查询操作将访问 $O(\log N)$ 个节点。所以每秒用于修复所需要的开销 Overhead $=(r+i)\log N/60$ packets。

如果 $r=i=\log N$；$N=100,000$，则 Overhead $=9.196$ packets。

假设一个 packet 由 64 个字节组成，则 Overhead $=0.6$ kb/Sec。这样的开销是不可忽略的，但是对于任何一个节点还是可以容忍的。

4 消息通信性能分析

4.1 程序模拟方法的描述

下面是程序模拟的结果。具体方法是按照 Netshot 路由算法，实现节点的静态加入，形成一个有多个节点参与的网络结构，然后随机选择两点进行消息通信，统计路由开销。重复多次，消除由于节点选取过于特殊带来的干扰，得到一个平均意义上消息通信的路由开销。

4.2 模拟结果的图形表示

选择不同的进位制 R，在单向路由表的连接方式下，得到

如下结果(采用SAS统计软件得到图1-图4)。

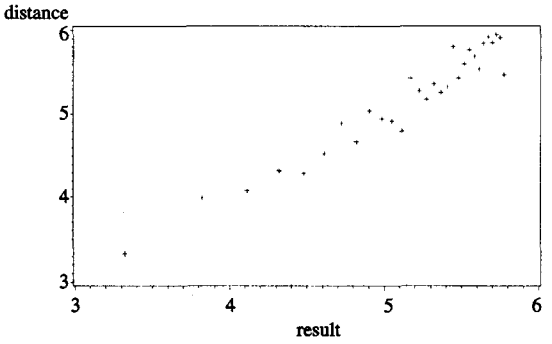

图1 Radix=2

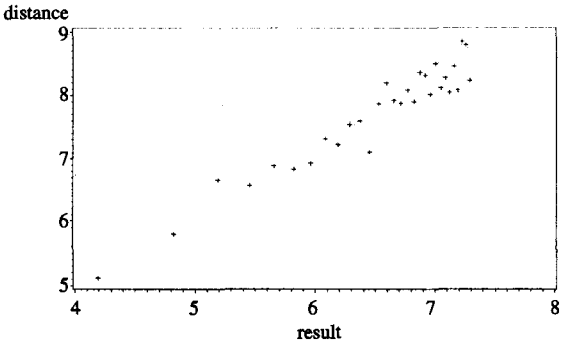

图2 Radix=3

(X轴取模拟的实际开销,Y轴取推导得到的理论值,节点数取100到3000,下同)

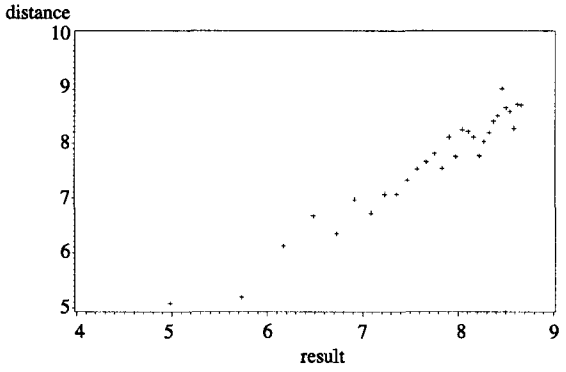

图3 Radix=4

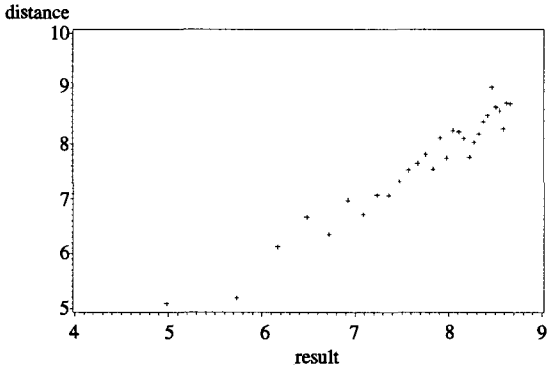

图4 Radix=5

从图中可以看出,当进位制R不断改变时,节点间通信的开销(图中由X轴表示)与理论值$\frac{(R-1)}{2} * \log_R N$吻合,在图形中表现为模拟结果对应的点$(x,y)$分布在45°线附近。

4.3 节点出错时消息通信开销

由于在P2P系统中任意节点都是不确定的,节点可以随时离开也可能随时出错,而不可访问,所以消息在向下一个节点发送的过程中可能会出现错误。当部分节点出现错误时,NetShot路由算法能够合理地选择未出错的节点来继续传递消息。图5是系统中节点的总数为1024且节点出错概率P分别为1/16, 1/12, 1/10, 1/8, 1/6, 1/5的情况下新节点加入所需的逻辑路径长度Hop数的模拟结果。采取的简单错误恢复策略是路由消息前递到一个出错节点时,采用强制方法迫使该出错节点离开系统。图6是节点以一定概率出错时系统中加入节点所需要的时间。

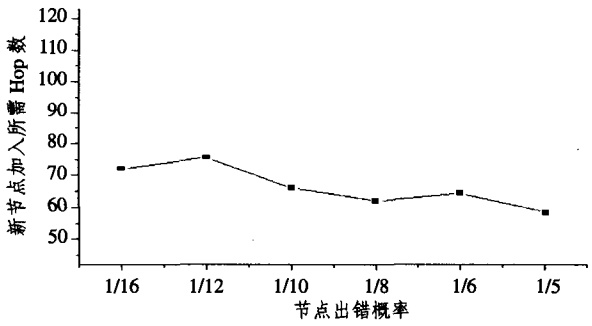

图5 节点出错时节点加入所需平均Hop数

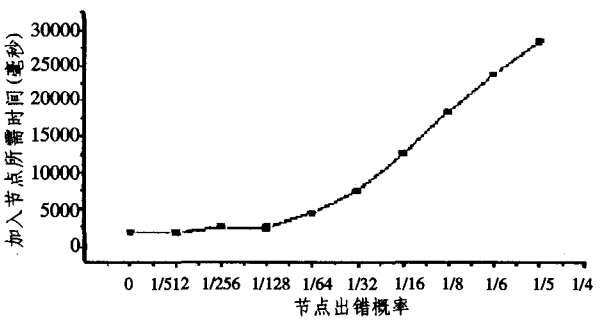

图6 节点以一定概率出错加入节点所需时间

从图中可以发现,当节点发生错误时,系统修正错误,维护正常路由所需的时间随着出错概率的增大而迅速增大。

结束语 NetShot是一个基于移位弦环空间分割的P2P路由结构,每个节点具有路由表和引入表来维护它与系统中部分节点的邻接关系,并根据节点间的邻接关系前递请求。理论分析和模拟结果表明该结构能够保证节点的加入、离开以及节点之间传递消息的数量都保持在很小的数量级,具有很好的可用性和可靠性。

参考文献

[1] Zhao B, Joseph A, Kubiatowicz J. Tapestry: An infrastructure for fault-tolerant wide-area location and routing. Tech. Rep. UCB//CSD-01-1141. University of California, Berkeley Computer Science Division, April 2001

[2] Berners-Lee T, Masinter L, McCahill M. RFC1738-Uniform Resource Locators (URL), December 1994

[3] Francis P, Handley M, Karp R, et al. A scalable content-addressable network // Proceedings of SIGCOMM. ACM, August 2001

[4] Rabin M O. Efficient Dispersal of Information for Security, Load Balancing, and Fault Tolerance. Journal of the Association for Computing Machinery, 1989, 36(2): 335-348