

一种改进的单跳 DHT 结构化 P2P 网络的查找算法

杜经纬

(运城学院 计算机科学与技术系, 山西 运城 044000)

摘 要: 本文提出了一种改进后的单跳对等点查找算法. 由于信息传播的改善, 该算法较之其它单跳 DHT 算法, 可更灵活地实现网络负载分布, 很好地节约时间. 实验结果表明, 这种改进后的单跳对等点查找算法通过充当查找效率的一种权衡来实现合理维护, 并且还可以实现系统更公平的网络负载平衡.

关键词: 分布式哈希表; 单跳; 网络负载

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2013)04-0127-04

An Improved Single Hop DHT Structured P2P Network Search Algorithm

DU Jing-wei

(Department of Computer Science and Technology, Yuncheng University, Yuncheng 044000, China)

Abstract: This paper presents an improved single hop peer search algorithm. Due to the improvement of information dissemination, the algorithm compared to other single-hop DHT algorithm, more flexible network load distribution, to save time. The experimental results show that this improved single-hop peer search algorithm to achieve the proper maintenance of a trade-off acts as a lookup efficiency, at the same time, it can achieve a more equitable system, network load balancing.

Key words: distributed hash table; single-hop; network load

1 引言

近年来, 对等(P2P)系统引入互联网, 这种技术已演变成网络领域最流行的应用之一, 并在文件共享、多媒体服务、资源定位等方面发挥了重要作用. 此外, 许多研究人员已开始探讨如何将 P2P 技术引入电信网络, 使得日常沟通费用低、方便、稳定并易维护.

目前, P2P 系统主要分为结构化与非结构化这两种. 结构化 P2P 系统也即分布式哈希表(DHT) P2P 系统, 像 Chord^[1]、Pastry^[2] 和 Kademlia^[3]. 这些 DHT 算法的主要优点在于数据放置和查找过程所产生的流量较小, 但它们总是无法回答所有情况下的这些要求. 例如, 在电信网络方面, 每个查询应满足严格的时间要求. 因此, 为解决这些要求, 有人

提出许多单跳 DHT 查找算法, 因为单跳算法符合快速、准确的查找要求, 且网络流量消耗适当. 由于单跳查找算法的基本理念是在系统的每个对等点储存一个全局路由表, 因此, 在非常短的时间内按照一定的工作量实现对这个全局表的查找就变得着实容易.

对于 DHT 查找算法, 人们关注较多的是数据存储的分布. 不过, 根据文献[4]的研究, 其中提到内部对象是一个 DHT 系统, 路由开销是系统带宽成本的主要部分, 比如在 P2PSIP^[5] 系统中, 只有主机的 IP 地址需要存储, 网络流量负载则主要受路由影响. 而对于单跳查找算法, 如 D1HT^[6], 路由开销的很大部分体现在如何维护全局路由表的操作方面, 因此, 解决这一问题的关键就在于确保这类系统的所有对等点之间的维护操作公平、高效.

收稿日期: 2012-07-08; **修回日期:** 2012-08-13

基金项目: 国家自然科学基金项目(60121731); 山西高校科技研究开发项目(20091150)

大多数的异质 P2P 网络, 它们的容量、实时工作量和网络流量带宽各有不同. 特别是, 有些对等点功能强大但工作量普遍低下, 而一些较弱的对等点则超负荷工作. 同时, 采取 Zipf-like 分布的 P2P 系统发生了一些转变, 这或许引发一个热点话题. 虽然多跳 DHT 查找算法对这个话题进行过不少探讨^[7-8], 但单跳算法文献研究中很少明确提到. 对于单跳算法, 当对等点的变化传播到整个网络时会出现负载均衡不理想的情况, 从而导致系统性能低下. 如, 一个超负荷工作的对等点可能遗漏一条传播消息, 结果在接下来的对等点中就不存在该条更新消息, 所有对等点彼此之间都存在错误, 整个系统陷入异步状态. 更严重的话, 如果这种状态持续, 系统会崩溃, 滞留大量错误.

本文提出了单跳查找算法的路由灵活问题, 分析 D1HT 传播的公平性, 并进一步提出单跳查找算法的一种改进型传播方法, 该算法可实现路由表维护期间的有效传播.

2 相关工作

D1HT 的设计应用到与 Chord 一样的拓扑结构. D1HT 系统由 Ring 上排列的对等点构成, 如图 1 所示. 每个对等点和项目分别对应一个识别码 (ID), 基于一致性哈希项目映射到对等点上. K 为 ID 长度, 信息量的单位为 Bit, 对应生成一个身份空格 $ID \in (0, 2^k - 1)$, 通常而言, 一个 ID 即等于其基础网络信息的加密 160BitSHA-1 哈希. 有了这个 ID, 对等点依照 ID 数字顺序排列在环上. 顺着环形拓扑结构, 系统上的每个节点均有一个后续节点和一个前趋节点, 它们分别是环上的顺时针和逆时针的第一个活动对等点. 这个对等点定期向其后续节点和前趋节点发出探测消息以核实它们是否有效. 若一个对等点发现它的后续节点或前趋节点不存在, 会生成一个路由表进行消息更新并将该消息传播到系统的所有对等点.

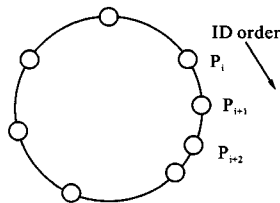


图 1 环形拓扑

与其它单跳 DHT 查找算法一样, D1HT 也采纳全局路由表. 大表将 ID 映射到系统对等点的 IP

地址上, 因为系统的各个对等点内含有全局路由表. 一个对等点内一旦发生查找请求就很容易得到解决, 也不会给每次查找产生网络流量. 因此多数情况下, 查找不会引发任何网络负载问题, 故认为这种设计十分适合查询工作量大而要求快速找到答案的系统.

但这种设计存在一个问题: 如何实时并有效地更新全局路由表. 对此, D1HT 提出用事件探测报告算法 (EDRA) 充当路由表维护的方法. EDRA 将对等点的接入和断开视为事件, 这些事件应认为是系统的所有对等点. 如果一个对等点发现它的直接后续节点或前趋节点无法回答探测消息, 它就会报告该事件, 意味着这个对等点应该将这一事件传播给整个系统. EDRA 算法给每个事件分配一个 TTL 值 ($0 < TTL < \lceil \log_2 N \rceil$, N 指代系统中的对等点总数) 以标记一个事件的生命周期. 一旦收到事件, D1HT 中的对等点会将它转发给它自己的第 2^{TTL} 、 2^{TTL-1} 、 2^{TTL-2} 、……、 2^0 的后续节点并将事件的 TTL 值依次减小 1. 当 TTL 最后等于 0 时, 这个事件消息就会发出, 然后到达系统的每个对等点. 有关 D1HT 的消息传播如图 2 所示.

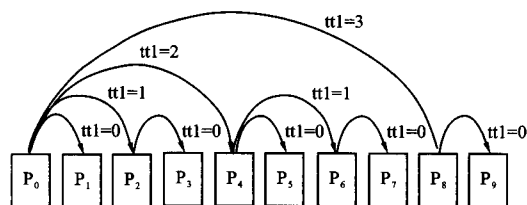


图 2 在 D1HT 中的信息传播

有了这种设计, 这类系统的对等点就可以负担不同的网络流量负荷, 因为维护操作产生大量的网络流量, 而且占据着大多数的网络负载. 因此, EDRA 传播期间每个对等点的网络负载的公平性决定系统网络负载的公平性.

若 D1HT 系统的一个对等点超负荷, 它就可能丢弃传播消息, 从而引发所有对等点的路由表无法进行更新. 一旦这个错误继续存在, 系统就会进入严重状态即系统对等点的路由表各不同. 如图 3 所示, 若对等点 P_4 超负荷, 那么它可能丢掉传播消息, 从而导致对等点 P_5 、 P_6 、 P_7 的路由表无法得到更新. 如此一来, 对等点 P_0 — P_4 的路由表就有别于对等点 P_5 — P_7 的.

3 算法改进

由于系统的对等点存在不同的实时网络流量负

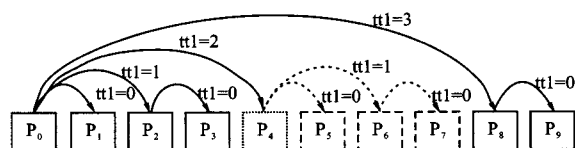


图3 在 D1HT 中一个失败消息的传播发生

载,可用带宽也就不同.一些对等点是自由的,有能力传播更多消息,而有些能力低下,存在超负荷现象.

根据提出的传播算法,每条消息有一个分配号A.如果第一个对等点开始传播,分配号为N(即系统中对等点的总数,可以通过计算全局路由表里面的项目很容易得出),那么这个编号就代表:当一个对等点接收到消息后这条消息应当传播给多少个对等点.因此,当一个对等点P_i接收到一条更新消息,它首先会检查传播数量A以了解此次传播过程中它应负责将消息传播到P_i至P_{i+A}之间的所有对等点.然后该对等点根据以下公式立即计算它能将消息传播给多少个对等点,即所谓的对等点的能力数.

$$C = \frac{B-R}{M} \quad (1)$$

B是该对等点可用的实时带宽,可通过计算实时网络流量负载和总带宽很容易得到. R指其它用途的预备带宽,取决于系统中的用户数. m指传播一条更新消息所需的带宽. 根据这个公式得出这个对等点应当将更新消息传播给系统的随后的至少C个对等点. 然后,从P_i至P_{i+A}之间的这些对等点分成C部分,从每部分中选出一个对等点,要求这些C个对等点负责各个部分.

通过图4说明,P₀检测到一个事件,它将一条更新消息传播给系统的所有对等点. 图4中从P₁至P₉共有9个对等点. P₀首先检查它的实时网络负载,并根据公式(1),它可以将这条消息传播给另两个对等点. 因此P₀在1至9之间选择它的后续和中间对等点,即P₅,作为其对应的消息转发对等点. 然后,P₀将更新消息发给P₁,同时分配给P₁一个任务:对等点P₁应负责P₂至P₄之间的消息传播工作. 当然,如果P₁无法对其负责范围内的其它对等点进行消息传播,它也可以向它们对等点分配任务. P₀发出更新消息给P₅并分配一个任务即P₅应负责将该条消息传播给P₆至P₉之间的各个对等点. 当P₁和P₅接收到来自P₀发出的更新消息,它们会相应地更新自己的路由表,并采取类似对等点P₀的做法继续进行消息传播. 通过这种方式,更新消息就会传播给P₁至P₉之间的所有对等点.

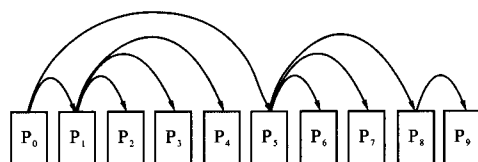


图4 改进单跳算法的信息传播

有了这种改进,我们提出的算法较D1HT能实现更合理的网络负载分布.而且,根据这种算法,功能强大的对等点将承担更多责任.若存在其它可用带宽,这种算法还能更快传播更新消息,并减少查找失败率.

4 实验结果

进行了相关实验,所得结果如下. 评估工作采用到Peersim方面的算法. 实验使用到的参数与D1HT的相同. 同样与D1HT算法一样,在一个 $10^5 - 10^6$ 对等点的P2P系统上进行了实验. 为了描述系统上对等点的变化情况,基于^[9]所做的研究选取三个有代表性的平均访问时间(S_{avg})值60分钟、174分钟和300分钟,代表一个P2P系统里三种典型的对等点运动. 同时,出于统计方便,采用 $m=80\text{bit}$, $t=1$ 秒来表示传播消息.

在图5的模拟过程中,在一个 $10^5 - 10^6$ 对等点的P2P系统上进行了该算法的有关实验,平均访问时间 $S_{avg}=174$ 分钟,每个对等点的平均能力数在2至10之间,如图5所示. 该算法能取得的平均能力数越多,所实现的跳跃就越少,即传播速度更快. 如图5所示,D1HT的能力数一直固定不变,即 $C=2$,无法通过使用可用带宽来改进传播,因此D1HT在图5中只保留了一个点. 当然,如果该算法的能力数C同样为2,得到的结果与D1HT是相同的.

图6给出的是系统消耗的平均外出带宽情况. 对于D1HT, $C=2$,消耗的外出带宽最少但同时传播效率也低下. 本文算法可以超出这个数,且随着带宽消耗越多,算法性能越好.

与此同时,系统中对等点的数量也是影响带宽消耗的一个因素. 较大型的系统外出带宽消耗更高. 图6中当对等点数量逐渐增加,平均外出带宽呈现线性走势,但仍被许多系统所接受. 如,对于 $N=10^6$,平均对等点的能力数 $C=3$,当 $S_{avg}=174$ 分钟时,系统的对等点需要30KB带宽来维持全路由表. 这对许多系统是可以接受的,如电信网络.

由于大多数查找只是根据对等点的全路由表来进行,查找失败率是评价查找效率的一个关键因素. 文中将查找失败率界定为:在一个时段间隔T内,

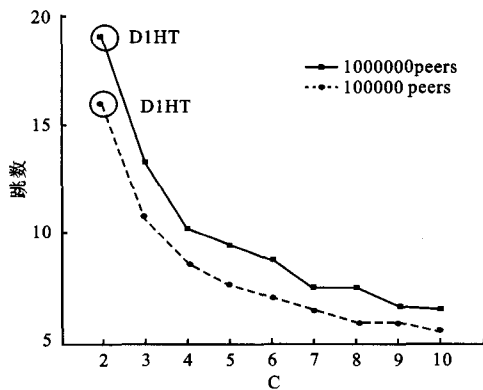


图5 C能力数不同平均总跳数

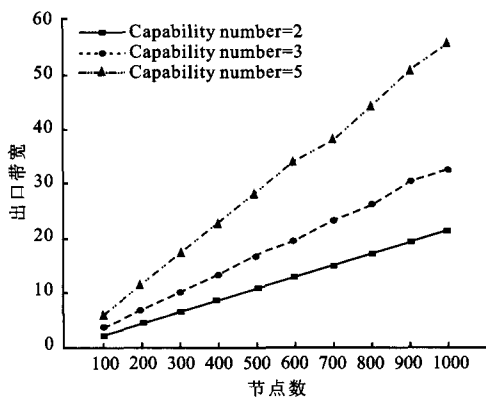


图6 系统节点数的出口带宽消耗

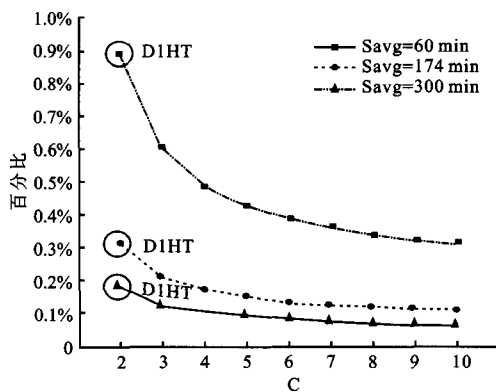


图7 查找故障率

如果查找目标未得到更新,将导致出错.因此,对于传播时间间隔 T 和系统变化 r ,在 N 个项目的全路由表中会有 $r * T$ 个项目未更新,故失败率是:

$$F = \frac{r * T}{N} \quad (2)$$

而且,根据^[10],得出

$$r = \frac{2 * N}{S_{avg}} \quad (3)$$

因此得到:

$$F = \frac{r * T}{N} = \frac{2 * N * T}{S_{avg} * N} = \frac{2 * T}{S_{avg}} \quad (4)$$

在一个 10^5 对等点系统中进行了这次模拟实

验,如图7所示. D1HT 仍然只保留一个点.但根据本文算法,如果一个事件传播给整个系统所需的总时间少,查找失败率往往也低,可用带宽较高,查找失败率也就低.但很多时候,本文的算法传播一条消息所需的时间可以少于 20 秒,因此,系统的查找失败率也就不足 1%.

5 结束语

本文针对 P2P 覆盖提出了一种改进型的单跳 DHT 查找算法.相比于多跳查找算法,本文的算法可显著减少 P2P 应用环境下的查找跳跃,在大多数情况下实现单跳.此外,本文还得出路由表维护算法可以通过一种更灵活的方式将系统的更新消息传达给所有对等点,传播速度快,同时还比其它单跳算法更有效地使用对等点的可用带宽.

由于单跳非常适合电信网络,今后将深入研究电信网络的典型特征,如功能强大的 ISP 部署对等点,它们稳定性好,性能佳,比普通的网络对等点有很多优点,因此今后将在电信网络方面引入这一算法,并结合这些特征以改善电信网络方面 P2P 应用的性能.

参考文献:

- [1] Ion Stoica, Robert Morris, David Karger. Chord: a scalable peer-to-peer lookup service for internet applications[C]// Proceedings of the Conference on Applications, Technologies, Architectures Computer Communications, 2001:345-349.
- [2] Antony Rowstron, Peter Druschel. Pastry: scalable, decentralized object location, and routing for large-scale peer-to-peer systems. IFIP/ACM[C]// International Conference on Distributed Systems Platforms. Redmond, WA, USA; ACM, 2001:123-128.
- [3] Petar Maymounkov, David Mazières. Kademlia: a peer-to-peer information system based on the XOR Metric. Proceedings[J]. IPTPS, 2002:23-29.
- [4] B Godfrey, I Stoica. Heterogeneity and load balance in distributed hash tables[C]// Proc of IEEE Infocom[s. l.]; IEEE, 2005:124-127.
- [5] C Jennings, B Lowekamp, E Rescorla, et al. REsource LOcation And Discovery [J]. RELOAD, 2009:89-96.
- [6] L R Monnerat, C L Amorim. D1HT: A distributed one hop hash table[C]// Proc of the 20th IEEE Intl Parallel & Distributed Processing Symp. Cambridge, MA, USA; IEEE, 2006:56-62.

(下转第 134 页)

```

Starting XPS Central DMA Transfer(s):
XPS Central DMA 83.220846 MB/s
Starting HDMA Transfer(s) to DDR2:
HDMA Tx 225.518182 MB/s
HDMA Rx 225.479983 MB/s
Test Completed!

```

图 7 XPS Central DMA 和 HDMA 数据传输速率

5 结束语

本文针对基于 Xilinx PowerPC 的光纤通道 SOPC 设计,研究了 XPS Central DMA 和 HDMA 两种 DMA 的特点,对比了它们的优缺点,并在开发板上做了硬件实现.实验结果表明 HDMA 收发通道传输速率均可以达到 225MB/s,满足光纤通道的传输速率(200MB/s)要求.

参考文献:

- [1] Fibre Channel Framing and Signaling-3 (FC-FS-3), Rev0.94[S]. 2010.
- [2] 陈俊,基于 FPGA 的光纤通道系统设计与实现[D].成都:电子科技大学,2012.
- [3] PLB v3.4, OPB to PLB v4.6 System and Core Migration User Guide, REV1.4[S]. 2008.
- [4] XILINX XPS Central DMA Controller, REV2.01a[S]. 2009.
- [5] 张宇,冯丹.针对 Xilinx 可编程片上系统的硬件加速方案的研究[J].小型微型计算机系统,2010,31(6):1147-1151.
- [6] XILINX Embedded Processor Block in Virtex-5 FPGAs, Rev1.8[S]. 2010.
- [7] James Lucero, XILINX Reference System: Optimizing Performance in PowerPC 440 Systems, REV1.0[S]. 2008.
- [8] 张雷峰 男,(1987-),硕士研究生.研究方向为光纤通道、片上可编程系统.
- [9] 王忆文 男,(1966-),教授.研究方向为 SOC/SOPC 的设计与测试、光纤通道和无线网络.
- [7] 杨继萍,王颖.一种基于节点交换的 DHT 优化方法[J].计算机工程,2011:34-38.
- [8] 郝结.基于 DHT 的结构化 p2p 路由协议 chord 的研究与改进[D].北京:北京邮电大学,2010.
- [9] A Gupta, B Liskov and, Rodrigues. One hop lookups for peer-to-peer overlays[C]//Proceedings of the 9th Conference on Hot Topics in Operating Systems, 2007: 34-41.
- [10] S Saroiu, P Gummadi, S Gribble. A measurement study of peer-to-peer file sharing systems[J]. Of SPIE/ACM MMCN, 2008: 35-40

作者简介:

杜经纬 男,(1978-),硕士,讲师.研究方向为计算机网络与信息安全.

(上接第 130 页)