

基于语义路由的 P2P 系统综述^{*}

刘 业 杜庆伟 杨 鹏

(东南大学计算机网络和信息集成教育部重点实验室 南京 210096)

摘 要 介绍了 P2P 系统的发展状况,归纳了典型的结构化 P2P 系统模型,对结构化 P2P 系统中的节点拓扑结构、语义路由算法、数据存储和共享、应用层多播等热点问题进行了讨论,介绍了非结构化 P2P 系统的语义路由过程,总结了当前主要的基于语义路由 P2P 系统的优缺点,并据此提出了该领域的发展趋势。

关键词 Peer-to-Peer, 语义路由, 拓扑结构

Survey of the Peer-to-Peer System Based on Semantic Routing

LIU Ye DU Qing-Wei YANG Peng

(The Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, Nanjing 210096)

Abstract The state of arts of Peer-to-Peer system in a large-scale networks is introduced. Some hot topics, such as the topology formed by nodes, semantic routing, global storage, data sharing, and an application level-multicast, are discussed in detail. The advantages and the shortcoming of the P2P system based on semantic routing are summarized. Some trends of development are mentioned as well.

Keywords Peer-to-Peer, Semantic routing, Topology

1 引言

随着网络规模的扩大,基于 C/S 模式的诸类应用对各类服务器的并发处理能力提出了愈来愈高的要求,同时该模式下数据交换都是经由服务器组织进行的,因此网络上的数据流分布呈现很大的不均衡性,如何缓解中心服务器的负载以及带宽瓶颈使得对 P2P 系统的研究得到人们的高度重视。

P2P 系统的提出消除了集中或层次式服务器的概念,系统中每个节点既是客户端、也是服务器端,所有数据的交换都是在节点间完成的。P2P 系统整合了联入网络中的所有硬件的存储资源和计算资源,进而可提供当前任何高性能计算机都无法拥有的计算能力和存储能力。当前比较普及的 P2P 的应用主要是文件共享,比如 Napster, Gnutella^[1]。尽管共享文件的版权问题使得提供基于该模式的文件共享服务的商业应用没有走得更远,但是作为 P2P 技术本身,则有很多值得研究的技术点,比如系统中的完全分布式控制, P2P 系统的自组织特性、系统的可用性和可扩展性等等。近几年来,随着结构化拓扑结构的 P2P 系统的出现,把 P2P 的研究引向了一个新的台阶。P2P 系统最大的特征是:各节点之间相对于集中式和层次结构是平等关系,各节点负载相同、重要性相同,运行于其上的软件的功能相同;健壮性(或鲁棒性);网络上不可能出现 DoS 攻击;但同时 P2P 系统中节点行为的审查变得很难。

通常衡量一个 P2P 系统的性能有四个指标,(1)系统的可扩展性,即系统的容量 N;(2)每个节点所维护的路由表的大小;(3)系统的稳定性,当节点加入和离开系统时,对系统中传递的消息 msg 语义路由的正确性的影响,该影响越小越好;(4)相对距离 RD(Relative Distance),对于系统中的 2 个节点传递一个 msg,设根据 P2P 路由表从源节点路由到目的

节点所经过的 IP 层 hop 数为 Dist-p2p-routing-table,直接从源节点路由到目的节点所经过 IP 层 hops 数为 Dist-complete-routing-table,则 RD 为二者的比值,有的系统称之为 RDP(Relative Delay Penalty);该指标越低,表明该 P2P 系统语义路由的效率越高。

P2P 系统根据节点间的拓扑结构的不同分为两大类,一类是非结构化 P2P 系统,在该类对等网络中,节点间的连接是任意的,数据放置于何处,对其它节点来说是不可知的,在进行搜索数据的过程中所采用的消息前递算法使得消息总数量呈指数增长,存在着报文泛滥的问题,对带宽消耗严重。另一类是结构化 P2P 系统^[2],该类系统中,提供了节点和文件物理存储位置的映射关系,从而保证了查询在有限步内完成,根据系统所提供的路由算法能保证 msg 正确路由到目标节点。对于 Napster 这样的 P2P 系统,由于其采用的是目录服务器来完成对数据的定位,作为目录服务器的节点和其它节点之间的功能是不对等的,该结构存在单点故障问题,因此对此类非完全 P2P 系统本文不作归类。

2 结构化 P2P 系统模型

结构化拓扑结构的 P2P 系统模型如图 1 所示,可分为三层:①连通层,处于模型中的最下层,负责节点间的数据传输,表示节点间的逻辑连接,即逻辑跳数 LH=1(Logic Hops);可使用 Http/XML、TCP、UDP 等报文来传输或转发数据。② P2P 架构层,模型中各节点 node-id(P2P 系统节点的唯一标识符)之间是平等的关系(相对于集中式和层次结构),各节点负载相同、重要性相同,通过节点转发,节点两两之间能相互通讯。整个系统有很强的自组织特性,节点加入退出时,对节点以及在节点上的〈对象定位信息 OLM〉的维护不需要由一

^{*} 基金项目:江苏省高技术研究项目(BG2001032)。刘 业 博士生,主要研究方向为分布对象计算、对等网络计算。

个集中的协调者统一控制。目前,结构化的 P2P 系统均是在此基础上的扩展和细化。不同的系统主要是 P2P 架构层中节点 node_id 的组织方式的不同。③资源层,负责将网络中物理节点上的对象(例如:文件)的定位信息 OLM 散布到整个 P2P 系统中,使得在完全分布式环境下实现通过语义路由定位目标对象。

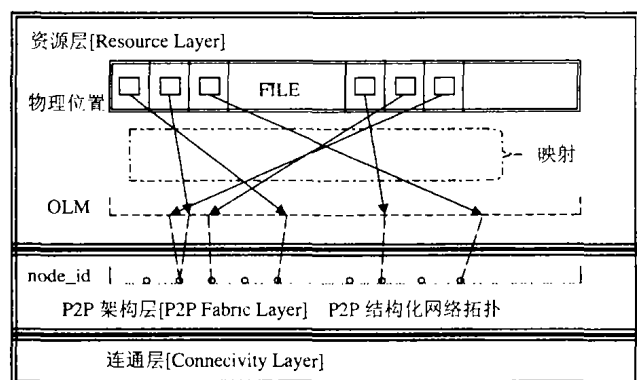


图1 结构化 P2P 系统模型

结构化拓扑的 P2P 系统实现语义路由的方式就是:(1)将具体的对象通过一定的映射算法得到它在系统中的唯一标志号,通常有 2 种方式,一是随机产生,二是根据 SHA-1 或 MD5 哈希函数得到一个 object_id;(2)将 object_id 跟节点 node_id 之间建立 assign 关系;(3)将 node_id 组织成一个结构化的拓扑,每个节点通过一定的规则跟系统中其它一小部分节点建立连通关系。这样当查询报文从源节点到达目的节点的时候,我们称之为完成了语义路由,或者说是应用层路由。

3 结构化 P2P 系统分类

3.1 弦环型拓扑结构

弦环拓扑结构的典型代表是 MIT 提出来的基于 P2P 的信息资源定位、查找模型 Chord^[3],模型中节点与其它节点通过 node_id 的值维持一种幂(2ⁱ)相邻关系来建立路由表;例如系统中节点为 {N1, N8, N14, N21, N32, N38, N42, N48, N51, N58};则 N8 节点维护的路由表如图 2 所示。查找文件 123.mp3 的定位信息的语义路由步骤如下:OLM = Hash(123.mp3) = K50;K50 在 Chord 模型中放置于节点 N51 上,则根据路由表通过幂相邻关系逐步逼近的语义路由过程为 [当前节点 N8 ==> (N8 + 32, N42) ==> (N42 + 8, N51)];从源端到目的端的逻辑跳数为 O(log₂N)。

N8+1	N14
N8+2	N14
N8+4	N14
N8+8	N21
N8+16	N32
N8+32	N42

图2 Chord 模型中节点 N8 的路由表

文[9]针对新节点并发加入和节点不可预测离开的情形提出来的路由表维护算法,缓解了 P2P 系统中新节点并发加入的 node_id 命名冲突问题以及系统中节点加入离开时系统路由正确性的问题。

3.2 树型拓扑结构

0642	*042	**02	***0	----
1642	*142	**12	***1	----
2642	*242	**22	***2	----
3642	*342	**32	***3	----
4642	*442	**42	***4	----
5642	*542	**52	***5	----
6642	*642	**62	***6	----
7642	*742	**72	***7	----
链接 1	链接 2	链接 3	链接 4	----
反向链接 (Back Pointers)				

图3 Tapestry 模型中节点 3642 的路表由

Tapestry^[4]、Pastry^[5,6]都是基于 Plaxton 算法的 P2P 语义路由算法。以 Tapestry 为例,该模型的路由表的构成模式如图 3 所示,取后缀相同位构造路由表。一次从源端到目的端的路由过程如下(当前节点 * * * * ==> * * * 8 ==> * * 98 ==> * 598 ==> 4598 其中 * 表示统配符)。当前节点可以看成是树的根节点,其最大出度为 b,路由过程中的每一跳可以看成是从父节点到其子节点的过程,从源端到目的端的逻辑跳数小于 $\lceil \log_b N \rceil$,此处 b=8。

3.3 平面拓扑结构

CAN^[10]是 Berkeley 提出来的基于平面空间划分来建立节点的相邻关系,节点间的路由过程是按照节点的几何空间相邻关系完成的。该模型的缺点是节点并发加入或节点不可预测失效时,维护系统稳定性的管理十分复杂。

文[7]提出了一种概然性定位和路由算法。该种算法能极大地提高当前的 P2P 系统的对象定位和语义路由的效率。对于近处的对象,语义路由效率比较高,若对象在较远的地方不在近处,即便对象存在,语义路由亦失败。该算法在每个节点中增加了一个过滤器(attenuated Bloom filter),以过滤器来决定语义路由的下一跳,过滤器中存储近处的对象哈希值记录。这种路由算法是与物理位置相关的。

结构化 P2P 系统的缺点是对于对象和它在系统中的唯一标志符 object_id 的映射关系,所做的处理还比较简单,对于散列算法中的输入集并没有很复杂的考虑。

4 完全分布式 P2P 系统

在完全分布式 P2P 系统模型中,节点无需维护一个结构化 P2P 模型中的路由表。Freenet^[11]在对系统中的文件进行发布、复制、使用的时候,对用户和所有者都提供了匿名机制;如图 4 所示,在进行查找的时候,节点间是采用回溯的方法;文件以与位置无关的方式在系统中动态地流动着,在文件需求者附近的节点增加文件的复制,而对于节点上一段较长时间不用的数据,将自动删除它。这样使得 P2P 系统中对无用数据有自动过滤的功能。在图 4 中,查询结果返回的过程中,节点总是发送给向它发送报文的上一个节点,它并不知道上一节点是需求的发起者,还是转发者,因此实现了 Freenet 网络中的查询者的匿名的特点。不过回溯法将造成查询过程的延迟很大。

对于另一类完全分布 P2P 系统 Gnutella 来说,它与 Freenet 不同的是在查询消息前的过程中不是回溯,而是扩散法(flooding),扩散过程本身产生的开销非常大,该 P2P 系统仅仅适合规模小的 P2P 应用。

结构化拓扑的 P2P 网络,节点的动态加入和退出需要额外的维护工作;而对于无结构化拓扑的 P2P 网络中,节点加

入和退出的维护是不需要做额外的工作。

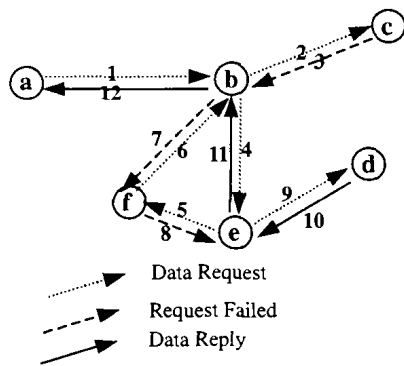


图4 Freenet 语义路由模型

5 结构化 P2P 系统中映射算法

结构化 P2P 系统实现从对象到对象名映射的过程中,根据输入对象名、对象拥有者、对象属性等信息映射得到 object-id,为了使得 object-id 可以负载均衡地分布在 P2P 系统中所有的节点 node-id 上,需要产生均匀分布的映射值(即哈希散列值),因此大多 P2P 系统选用了报文摘要的方式。报文摘要又称单向散列函数值,它是现代密码学的重要研究内容,报文摘要作为文件的指纹,起到惟一的标定文件的作用^[6]。报文摘要中应用最广的是 MD5(Rivest 1992)和 SHA(NIST 1993)。MD5 的输出是一个 128 比特的 2 进制数,SHA/SHA-1 的输出为 160 比特。同样道理,在 P2P 系统中使用的 MD5 或 SHA-1 根据输入节点 IP 地址等信息得到 node-id。Pastry 中 node-id 为 128 位,object-id 为 160;Chord 中 node-id、object-id 都为 128 位。

6 结构化 P2P 应用

6.1 海量数据存储及信息共享

海量数据的存储和信息共享是互联网中最基本的应用,不少大学和研究机构开展了基于结构化拓扑的 P2P 系统的数据存储和信息共享的研究项目。PAST^[8,12]是美国 Rice 大学联同微软研究院开展的提供面向整个互联网的存储服务项目,其路由算法是 Pastry。CFS^[13]是由 MIT 提出的基于文件内容进行定位的分布式信息服务系统,其基本路由算法是 Chord。UC Berkeley 开展的 OceanStore^[14]项目是提供面向整个互联网的存储服务,该项目中的路由算法采用的是 Tapestry。该类应用不仅可以提供海量数据的存储以及文件共享服务,而且具有很强的鲁棒性,对于当前对网络造成较大危害的 DoS(服务失效攻击)、DDoS(分布式服务失效攻击)几乎难以构成威胁。

6.2 应用层多播

Rice 大学的 SCRIBE^[8,15]项目提供了一个基于 Publisher/Scriber 通知服务模型的应用层多播系统,订阅者共享相同主题的多播报文负载,该系统提供了很好的匿名性,同时所构建的多播生成树在可靠性、可扩放性、效率等方面都具备良好的性能。

总结和展望 结构化拓扑结构的 P2P 系统最大的优势就是强大的可扩放性,在没有中心节点控制下提供整个分布式系统范围内的查询服务。提供了基于名字的语义路由,可以将对一个对象的请求路由到对象所在节点,实现了大规模的分布式哈希表 DHT(Distributed Hash Table)^[17]。基于这类 P2P 系统之上的 Internet 应用的性能完全依赖于 P2P 系统的

可靠性和语义路由的正确性。此类 Overlay 网络是在物理网络拓扑基础上增加了一层逻辑层,而这类系统的构建过程中假定的前提是网络上的各节点在网络中的资源是唯一的,对于现实网络中的资源的多个备份在 P2P 系统中很少被考虑到,并且在进行语义路由的时候没有考虑结合节点所在的物理位置。美国 Berkeley 大学的 Brocade^[15]项目考虑在当前 P2P 的一层 Overlay 网络的基础上再增加第二层 Overlay 网络,对提高语义路由效率作出了一定的探索。将 Chord、CAN、Pastry、Tapestry 作为构建大规模 P2P 系统的基本架构得到了越来越广泛的应用,因此通过何种机制提高现有 P2P 系统语义路由效率,将是近阶段 P2P 系统的研究重点。

随着网络带宽和节点处理能力的进一步提高以及 IPv6 的推广使用将使摒弃了中心节点瓶颈的 P2P 系统得到愈来愈广的应用。当前基于 P2P 的应用比较成熟的技术是文件共享和文件存储,而 Peer 端间的协同工作、P2P 系统的安全性、P2P 系统的容错性、实时可调度性等都是有待解决的关键问题。对于此类技术难点的突破性解决方案将会进一步推动 P2P 网络的演化。

参考文献

- Adar E, Huberman B A. Free Riding on Gnutella. First Monday, 2000, 5(10)
- Castro M, Druschel P, Hu Y C, Rowstron A. Topology-aware routing in structured peer-to-peer overlay networks: [Technical Report MSR-TR-2002-82]. Microsoft Research, 2002
- Stoica I, et al. Chord: A scalable peer-to-peer lookup service for internet applications. In: Proc. of the ACM SIGCOMM '01 Conf. San Diego, California, Aug. 2001
- Zhao B Y, Kubiawicz J D, Joseph A D. Tapestry: An infrastructure for fault-tolerant wide-area location and routing: [Technical Report UCB/CSD-01-1141]. UC Berkeley, Apr. 2001
- Rowstron A, Druschel P. Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. In: Proc. of the 18th IFIP/ACM Intl. Conf. on Distributed Systems Platforms (Middleware 2001). Heidelberg, Germany, Nov. 2001
- Schneier B. 著, 吴世忠等译. 应用密码学. 机械工业出版社, 2000
- Rhea S, Kubiawicz J. Probabilistic location and routing. In: Proc. of INFOCOM. IEEE, June 2002
- The Pastry Project. <http://www.cs.rice.edu/CS/Systems/Pastry/>
- Liben-Nowell D, Balakrishnan H, Karger D. Analysis of the evolution of peer-to-peer systems. In: PODC, 2002
- Ratnasamy S, et al. A scalable content-addressable network. In: SIGCOMM, Aug. 2001
- Clarke I, et al. Freenet: A distributed anonymous information storage and retrieval system. In: Proc. Intl. Workshop on Design Issues in Anonymity and Unobservability, volume 2009 of LNCS, Springer-Verlag, 2001. 46~66
- Druschel P, Rowstron A. PAST: A large-scale persistent peer-to-peer storage utility. In: Proc. HOTOS Conf. 2001
- Dabek F, et al. Wide-area cooperative storage with CFS. SOSP '01, Oct. 2001
- Kubiawicz J, et al. Oceanstore: An architecture for global-scale persistent storage. In: Proc. of the Ninth Intl. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2000
- Rowstron A, et al. Scribe: The design of a large-scale event notification infrastructure. In: Third Intl. Workshop on Networked Group Communications, Nov. 2001
- Zhao B Y, et al. Brocade: Landmark routing on overlay networks. In: IPTPS02 [7]
- Joseph D, Kubiawicz J D. Routing Algorithms for DHTs: Some Open Questions. In: Electronic Proc. for the 1st International Workshop on Peer-to-Peer Systems (IPTPS '02) 2002 - MIT Faculty Club, Cambridge, MA, USA