ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.5, May 2009, pp.1362−1376
doi: 10.3724/SP.J.1001.2009.03485

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

P2P网络中Churn问题研究*

张宇翔[1,2+], 杨 冬[1], 张宏科[1]

[1](北京交通大学 电子信息工程学院 下一代互联网互联设备国家工程实验室,北京 100044)

[2](中国民航大学 计算机科学与技术学院,天津 300300)

Research on Churn Problem in P2P Networks

ZHANG Yu-Xiang[1,2+], YANG Dong[1], ZHANG Hong-Ke[1]

[1](National Engineering Laboratory for Next Generation Internet Interconnection Devices, School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044, China)

[2](College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)

+ Corresponding author: E-mail: yxzhcn@sina.com

Zhang YX, Yang D, Zhang HK. Research on churn problem in P2P networks. *Journal of Software*, 2009,20(5): 1362−1376. http://www.jos.org.cn/1000-9825/3485.htm

Abstract: Churn is one of the main problems faced by all P2P networks. This paper surveys the existing theories and methods about Churn. On the basis of the generation mechanism of Churn, this study summarizes the major steps of the Churn problem solution: precise measurement of Churn, analysis of the impact of Churn on P2P performance, and provision of specific strategies to handle Churn. Under each step, existing problems, solutions, and the newest recent research are explored. This paper also discusses the prospects of future studies.

Key words: P2P network; Churn; measurement of Churn; performance evaluation of P2P under Churn; strategy to handle Churn

摘 要: Churn 问题是 P2P 网络面临的基本问题之一.通过系统地归纳现有文献,从 Churn 问题产生的机理出发,总结出解决 Churn 问题的主要步骤,依次是准确度量 Churn,分析 Churn 对 P2P 网络性能的影响,给出应对 Churn 的具体策略.以此为主线对 Churn 问题的研究进展进行综述,全面、深入、系统地总结了每个步骤中涉及的关键问题以及解决这些问题的具体方法与最新成果.讨论了存在的问题并指出未来可能的研究方向.

关键词: P2P 网络;Churn;Churn 度量;Churn 下 P2P 网络性能的评估;应对 Churn 策略

中图法分类号: TP393 **文献标识码**: A

为了实现网络资源的分布式共享,P2P 网络的前期研究工作主要集中在 P2P 网络的设计和实现,其中包括对其性能的简单评估,如 Chord[1],Tapestry[2]等协议.随着 P2P 网络的开发应用以及研究的深入,有相当多的工作

* Supported by the National Basic Research Program of China under Grant Nos.2007CB307100, 2007CB307101, 2007CB307106 (国家重点基础研究发展计划(973)); the Cultivation Fund of the Key Scientific and Technical Innovation Project of China under Grant No.706005 (高等学校科技创新工程重大项目培育资金项目); the Program for Cheung Kong Scholars and Innovative Research Team in University of China under Grant No.IRT0707 (长江学者和创新团队发展计划)

Received 2008-04-28; Accepted 2008-10-17

将研究重点转向严重制约 P2P 网络发展与应用的 Churn 问题.在 P2P 网络中,节点可以随时、任意地加入或离开网络,而节点频繁地加入或离开(称为 Churn)会对 P2P 网络的性能造成严重的影响,如导致网络分割、网络带宽消耗增加、查找延迟以及查找正确率降低等[3],故在设计 P2P 网络和评价 P2P 网络性能时必须对 Churn 给予足够的重视,这正是它越来越受到广泛关注的根本原因.

本文首先通过系统归纳现有文献,从 Churn 产生的机理出发,总结出解决 Churn 问题的主要步骤,依次是准确度量 Churn,分析 Churn 对 P2P 网络性能的影响,给出应对 Churn 的具体策略.然后,进一步归纳总结出在每个步骤中分别需要解决的关键问题,以及目前采用的解决这些问题的方法和最新研究成果.其中,在主要步骤中需要解决的问题与采用的方法可概括为:(1) 用什么指标以及如何度量 Churn,主要采用网络测量方法抓取节点的在线、离线以及邻居列表等相关信息,从时间、频度和连接度 3 个方面度量 Churn;(2) Churn 对 P2P 网络的哪些性能指标有怎样的影响,通常采用理论模型分析或系统仿真实验两种方法来研究;(3) 采取什么措施来应对 Churn 对 P2P 网络性能的影响,通常从数据层、路由层、结点层、查找层 4 个层面给出应对 Churn 的具体策略.问题(2)的解决在很大程度上依赖于问题(1)的解决结果,问题(3)的解决很大程度上依赖于问题(2)的解决结果,故问题(1)的解决是解决其他问题的基础,目标是解决问题(3).以上 3 个问题都是当前研究的难点和热点,解决这些问题的最终目标是使 P2P 网络在 Churn 下具有高性能.最后,本文围绕要解决的问题将孤立的、分割的研究有机地整合在一起,在此基础上讨论了各部分存在的问题,并指出了未来可能的研究方向.

本文第 1 节介绍 Churn 的基本概念及其产生的机理.第 2 节详细归纳 Churn 的时间、频度和连接度度量指标.第 3 节从理论分析和系统仿真两个方面深入总结 Churn 对 P2P 网络性能的影响.第 4 节全面概括目前应对 Churn 的主要策略.第 5 节总结全文.在第 2 节~第 4 节中,相应地讨论存在的问题,并指出未来可能的研究方向.

1 Churn的基本概念

1.1 P2P网络基本概念

目前有关 P2P 网络的定义仍未统一,文献[1,4−6]从不同层面给出多种 P2P 网络的定义,包括资源共享与聚合层面[4,5]、通信特征层面[6]以及逻辑拓扑结构层面[1].无论从哪个层面定义 P2P 网络,它们的目标始终是在确保资源高效定位的前提下实现节点之间资源的有效聚合与共享,为此 P2P 网络通常具备以下 3 个特征:(1) P2P 网络是分布式应用系统,它是建立在 Internet 上的覆盖(overlay)网络[4],节点之间没有集中式的控制或者分层组织结构[1];(2) 节点之间可以通过本层路由协议进行对等通信[6];(3) 在大多数 P2P 网络中核心的操作是资源的高效、准确定位[1].

按照网络拓扑结构和资源定位方法,P2P 网络分为无结构 P2P 网络和有结构 P2P 网络[7].在无结构 P2P 网络中,拓扑结构非常松散,常采用泛洪搜索进行资源定位,资源定位效率低、可扩展性差.在基于分布式哈希表(distributed hash table,简称 DHT)的结构化 P2P 网络(简称 DHT 网络)中,拓扑结构按照某种特定的结构(如超立方体[8]、环[1]、Plaxton 树[2,9]等)组织,资源定位方法根据相应的拓扑结构来确定,资源定位效率高、可扩展性好.

1.2 Churn的基本概念

目前有关 Churn 的准确定义未得到统一,现有文献大都是从广义或狭义方面来定义 Churn.从广义上讲,Churn 是指节点频繁地加入或离开 P2P 网络[3,10].从狭义上讲,不同的研究侧重点给出不同的定义:

(1) 侧重于研究 Churn 的度量指标,Churn 被定义为 P2P 网络动态特征的度量值[11],如节点的会话时长、加入率、离开率、变化率等;

(2) 侧重于研究 Churn 对 P2P 网络性能的影响,Churn 被定义为广义 Churn 对 P2P 网络性能影响的集合[12];

(3) 侧重于研究应对 Churn 的策略,Churn 被定义为应对节点突然离开对 P2P 网络性能影响的修复算法[13].

Churn 的这些狭义定义反映出目前关于 Churn 的重点研究方向.在 P2P 存储系统中,为了确定数据的恢复策略,将 Churn 分为 Temporary Churn 和 Permanent Churn[10];为了研究 Gnutella 网络中节点度的变化和节点之间连接关系的变化,将 Churn 分为 Degree Churn 和 Connection Churn[14];此外,根据 Churn 的变化程度将其分为

Ordinary Churn 和 High Churn[15].

无论是无结构 P2P 网络还是 DHT 网络,它们的性能都会受到 Churn 的影响.特别是对于 DHT 网络,Gummadi 等人[16]指出,在 Churn 下 DHT 网络能否高效运行是其面临的关键问题之一,在 Churn 下 DHT 网络为了维护特定的拓扑结构需要不断地修复路由表,导致维护代价显著增加,进而严重影响 DHT 网络的性能.

2 Churn的度量指标

用什么指标来衡量 Churn 的大小呢?目前,Churn 的度量指标尚未统一,现有文献从时间、频度或连接度 3 个方面度量 Churn.时间度量指标用来衡量节点在 P2P 网络中停留的时间长度;频度度量指标用来衡量节点到达或离开 P2P 网络的频率;连接度度量指标用来衡量 P2P 网络中节点间连接关系的变化程度.Churn 的度量指标研究工作特别重要,其原因是,一方面可以通过 Churn 的度量指标来分析 P2P 网络中用户的行为特征,更重要的是,Churn 的度量指标是研究 Churn 对 P2P 网络性能的影响以及提出应对 Churn 的策略的基础.本节系统地总结出 Churn 的各个具体度量指标,详细介绍了相关的最新研究成果,并指出未来可能的研究方向.

2.1 Churn的时间度量

Churn 的时间度量指标具体包括会话时长(session time)、当前在线时长(uptime)、生存时长(lifetime)、当前余留时长(remaining uptime)、加入时间间隔(time between joins)、离线时长(offline time)等.目前这些度量指标在文献中都有使用,而且同一度量指标在不同的文献中可能有不同的含义,这就给研究工作和查阅文献带来诸多不便.这些度量指标之间容易混淆的原因是度量指标众多且某些不同的指标在特定情况下是相同的.下面结合节点所处状态进一步规范这些时间度量指标的准确定义,并总结目前关于它们的主要研究成果.

综合现有的理论研究成果,在 P2P 网络中,节点通常有如下 3 个状态(如图 1 所示,其中 T 为观测时段):

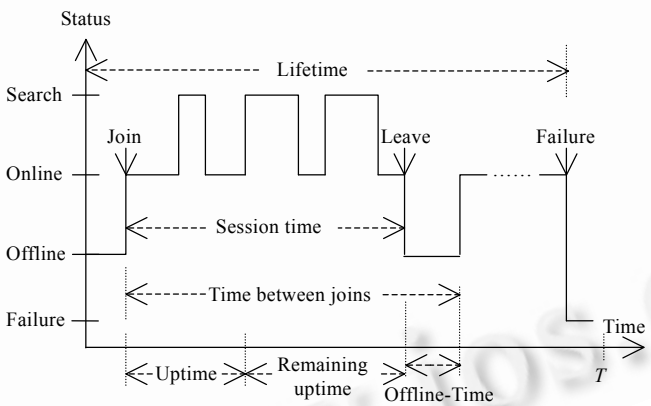

Fig.1 A peer lifetime consists of one or more sessions

图 1 每个节点的生存时长由 1 个或多个会话组成

(1) 在线状态(online state)[3,10,17],是指节点加入 P2P 网络,维护相应的邻居信息并可以参与节点之间的数据交换.在文献[17]中,根据节点是否参与资源查找或数据交换从在线状态中细化出 search 状态(search 状态的节点可以是查找内容或浏览信息或为其他节点提供服务[18]).(2) 离线状态(offline state),是指节点暂时离开 P2P 网络,在一段时间之后会重新加入 P2P 网络.(3) 失败状态(failure state),是指节点永久离开 P2P 网络.

文献[19,20]从网络测量的角度对节点所处状态进行了类似的分类,其目的就是为了准确地度量 Churn.处于在线状态的节点可以转移至离线状态,也可以转移至失败状态;处于离线状态的节点可以转移至在线状态,也可以转移至失败状态;失败状态是吸收态,不再向其他状态转移[10].

2.1.1 会话时长

会话时长是指节点一次加入至离开 P2P 网络所经历的时间段[12,21](如图 1 所示),即 *Session time=time of leave−time of join*,它是衡量 Churn 大小的主要时间度量指标.研究者主要关注它在 P2P 网络中的分布特征和平均值.如何获得实际 P2P 网络(如 Gnutella[22])中节点的较为准确的会话时长是一项极具挑战性的研究工作,是目前研究的难点和热点.在大规模的 P2P 网络中,不可能直接获取到每个节点的准确加入和离开时刻信息,通常采用被动监视(passive monitoring)或主动探测(active probing)的方法来获得节点会话时长的近似值.

被动监视:被动监视方法是指在网络的不同位置部署一定数量的测量点,使用特定的软、硬件设备采集相

关的 P2P 数据流信息[23].Sen 等人[24]采用被动监视方法对 3 个 P2P 文件共享系统 Gnutella,FastTrack[25],Direct Connect[26]进行测量,在一个 ISP(Internet service provider)的多个边缘路由器上布置测量点收集 P2P 抽样流,置会话时长门限为 30 分钟(30 分钟之内到达的 IP 地址相同的数据流作为同一个会话),分析测量数据发现,在 FastTrack 中,60%的节点的会话时长少于 10 分钟,会话时长总和服从重尾而不服从 Zipf 分布.Gummadi 等人[27]同样采用被动监视的方法,在华盛顿大学的出口网关处对 P2P 文件共享系统 KaZaA[25]进行测量,分析测量数据得到 50%的节点的平均会话时长仅为 2.4 分钟,90%的节点的平均会话时长为 28.25 分钟,平均会话时长服从重尾分布.

被动测量固有的局限性限制了测量结果的准确性,其主要的局限性是:(1) 准确、高效地识别 P2P 数据流是非常困难的[28];(2) 测量点监测不到节点加入或离开 P2P 网络的真实时刻,只有在发生交换数据时才能监测到,因此,实测会话时长小于节点在 P2P 网络中实际的停留时间[27];(3) 被动监视只能监测到经过测量点的数据,加之测量点数量有限,因此,实测数据难以代表整个 P2P 网络中节点的行为[12];(4) 实测数据依赖于节点的 IP 地址[12],在测量过程中区分不同节点的唯一标识是其 IP 地址,而动态地址分配协议(DHCP)和点对点协议(PPP)的使用会导致同一个节点被记为多个,网络地址转换(NAT)协议的使用会使得多个不同的节点记为同一个.

主动探测:主动探测方法是使用网络爬虫(crawler)主动加入特定的 P2P 网络,获取 P2P 网络特性以及节点的会话行为.主动探测的优点是探测过程可控性较高,测量结果准确性较好.先后发起探测的时间间隔粒度对测量准确性有较大的影响,特别是在 Churn 下[12].

最初,Chu 等人[19]采用主动探测的方法对 P2P 文件共享系统 Napster[29]和 Gnutella 进行测量,结果发现,31%的节点的会话时长少于 10 分钟,会话时长服从二次对数(log-quadratic)分布.Saroiu 等人[20]同样对 Napster 和 Gnutella 进行主动探测,给出节点会话时长的累积分布函数,发现两个系统的会话时长分布特征非常相似,会话时长的中值大约为 60 分钟,绝大多数的会话非常短暂,会话时长有较重的倾斜.Bustamante 等人[30]通过修改 Gnutella 的客户端软件对 Gnutella 中节点的会话时长进行主动探测,给出会话时长的逆向累积分布函数,发现会话时长服从帕累托(Pareto)分布.Liang 等人[31]采用主动探测方法每隔 5 分钟对 KaZaA 的 965 个超级节点探测一次,给出 KaZaA 中超级节点的会话时长的累积分布函数,发现超级节点的平均会话时长为 2.5 小时.2005 年 Stutzbach 等人[32]对 Gnutella,BitTorrent(内容分发系统)和 Kademlia[33](DHT 系统)进行主动探测,结果发现,这些系统的 Churn 特征非常相似,会话时长服从幂律(power-law)分布.2006 年,Stutzbach 等人[12]经过深入研究后发现,会话时长服从威布尔(Weibull)分布或对数正态(log-normal)分布,而不服从指数分布或者 Pareto 分布.

以上内容基本包括了近年来通过网络测量方法研究会话时长的主要研究成果.会话时长究竟服从什么分布到目前为止尚无定论.学界公认的是在 P2P 网络中大部分节点的会话时长较为短暂[32],通常从几分钟到 1 小时(文献[21]对此进行了详细总结).除了测量研究之外,Yao 等人[13]通过理论分析发现,即使 P2P 网络每个个体节点的会话时长都服从指数分布,整个系统的会话时长也有可能服从重尾分布.目前在 Churn 的理论模型中,会话时长通常采用服从节点行为同质的指数分布或重尾的 Pareto 分布,如文献[18,34−36]等.由于会话时长是认识 P2P 网络 Churn 和预测节点行为的基础,因此关于会话时长的研究将继续进行,一方面继续通过改进测量方法获得更接近实际 P2P 网络的测量数据进行分析,另一方面可以通过建立数学模型从理论上进行分析.

2.1.2 当前在线时长

会话时长不能刻画当前时刻 P2P 网络中存活节点的在线时长,为此引入当前在线时长[20,32].当前在线时长是指 P2P 网络中存活节点从加入至当前为止所经历的时间段(如图 1 所示).若当前时刻为节点离开时刻,则当前在线时长退化为会话时长.当前在线时长主要被用来预测当前余留时长,*Remaining uptime=session time−uptime*.当前余留时长是节点进行邻居选择的重要依据之一(见第 4.3 节).有关当前在线时长的研究主要集中于其分布函数以及预测当前余留时长的理论模型.Saroiu 等人[20]通过分析测量数据给出 Napster 和 Gnutella 的当前在线时长的累积分布函数,并暗示它服从泊松(Poisson)分布.Stutzbach 等人[12,32]同样通过分析测量数据给出 Gnutella,BitTorrent 和 Kademlia 系统的当前在线时长的累积分布函数,指出当前在线时长 $u(t)$不服从泊松分布而是服从幂律分布 $u(t)\propto\int_t^\infty s(t)\mathrm{d}t\propto t\cdot t^{-\alpha}$,其中会话时长 $s(t)$服从幂律分布 $s(t)\propto t^{-\alpha}$,在此基础上,根据测量数据对

当前余留时长进行了简单预测,同时指出,在 Gnutella 和 Kademlia 中,当前时长可以较好地预测当前余留时长,而在 BitTorrent 中则不能.Mickens 等人[37]应用线性预测法来预测当前余留时长.Yao 等人[38]通过理论分析发现:对于生存时长(见第 2.1.3 节)服从指数分布的非结构化 P2P 网络,当前在线时长对其余留时长的影响较小;而对于生存时长服从重尾分布的非结构化 P2P 网络,当前在线时长较长的节点有较长的余留时长.在对当前余留时长进行预测时,已有工作主要考虑当前在线时长的长度.我们认为可能还需考虑当前在线时长的分布特征以及特定节点以往的生存时长等因素.根据这些因素对当前余留时长进行综合预测值得深入研究.

2.1.3 生存时长

生存时长[28]是指 P2P 网络中节点从加入至永久失效为止所经历的时间段(如图 1 所示),*Lifetime=time of failure−time of join*.若每个节点只会话一次或将同一节点的每次会话作为不同节点的会话,则生存时长退化为会话时长,因此理论文献中常使用生存时长作为 Churn 的度量指标,在一些理论研究文献(如文献[13,18,34,38]等)中生存时长与会话时长相一致.

除上述 Churn 的主要时间度量指标之外,Churn 的时间度量指标还包括离线时长和加入时间间隔.离线时长[13]是指 P2P 网络中节点从离开至下次加入为止所经历的时间段(如图 1 所示).加入时间间隔是指 P2P 网络中节点加入至下一次加入所经历的时间段(如图 1 所示),*Time between joins=Session time + Offline time*,文献[32]通过分析测量数据发现它服从泊松分布.

这些众多的时间度量指标之间有什么内在联系是值得深入研究的方向.Yao 等人[13]抓住非结构化 P2P 网络中节点异质性的特征,以单个节点的会话时长和离线时长为变量计算所有新加入节点的 *Lifetime* 的分布函数 $F(x)$、系统中任意存活节点的 *Residual lifetime* 的分布函数 $H(x)$和系统中所有存活节点的 *Lifetime* 的分布函数 $J(x)$.设节点 i 的会话时长为 L_i,分布函数为 $F_i(x)$;离线时长为 D_i,分布函数为 $G_i(x)$.假设每个节点的会话时长、离线时长的分布各异,利用交替更新过程建立数学模型:(1) $F(x)=\sum_{i=1}^{n}b_iF_i(x)$,其中 $b_i=\left(\sum_{j=1}^{n}\frac{E[L_i]+E[D_i]}{E[L_j]+E[D_j]}\right)^{-1}$,$n$ 是节点数目,权重 b_i 偏向频繁加入和离开系统的节点.该式表明,即使 $F_i(x)$不服从重尾分布,如服从指数分布或指数分布与重尾分布的混合,$F(x)$也可能服从重尾分布,这从理论上验证了“会话时长服从重尾分布[27,30,39,40]”结论的正确性.(2) $H(x)$的计算公式十分复杂,它与邻居节点选择策略有关,假如采用文献[41]中提及的均匀选择策略,简化为 $H(x)=\frac{1}{E[L]}\int_0^x(1-F(u))\mathrm{d}u$,其中 $E[L]=\sum_{i=1}^{n}b_iE[L_i]$.(3) $J(x)=\frac{1}{E[L]}\left(xF(x)-\int_0^xF(u)\mathrm{d}u\right)$,可利用随机理论证明存活节点的 *Lifetime* 比新加入节点要长,这验证了“在 Gnutella 和 BitTorrent 中当前节点的平均会话时长较新加入节点的长[12]”这一结论的正确性.

2.2 Churn的频度度量

Churn 的频度度量指标用来衡量节点到达或离开 P2P 网络的频繁程度,包括节点的加入率、离开率及其他度量指标.加入率和离开率常用于理论研究(如文献[34,36,42])和仿真实验(如文献[12,21])中.Kumar 等人[43]定义 Churn 变化率 $r=(N_{arr}-D_{dep})/N_T$,其中 N_{arr} 是观测时段 T 内加入节点的总数,D_{dep} 是 T 内离开节点的总数,N_T 是开始时刻节点的总数.具有代表性的是 Godfrey 等人[3]定义的 Churn 度量 $C=\frac{1}{T}\sum_{enent\ i}\frac{|U_i-U_{i-1}|}{\max\{|U_{i-1}|,|U_i|\}}$,其中,$U_i$ 表示第 i 次变化(指节点的加入或离开)发生后系统中存活节点的集合,U_0 为 P2P 网络初始存活节点集合,T 为整个系统运行的时间.C 反映了单位时间内 P2P 网络中节点的变化程度,然而该公式没有对节点加入和离开操作进行区分.怎样定义 Churn 的频度度量指标才能使其较为准确地刻画 Churn 值得进一步研究.

2.3 Churn的连接度度量

P2P 网络中节点的动态加入和离开会导致节点之间连接关系的动态变化,主要包括节点度的变化和节点之间连接关系的变化.目前有关 Churn 连接度度量的研究文献很少.Li 等人[14]采用连接的平均生存时长(connection lifetime)来度量节点之间连接关系的变化程度,通过主动网络探测的方法研究发现:在 Gnutella 网络

中,节点之间连接关系的变化比节点度的变化更为厉害(其中,连接的平均生存时长大约为 2~3 小时,而节点的平均生存时长大约为 8~11 小时);连接的平均生存时长服从指数分布,连接失效率服从重尾的泊松分布;进而利用 M/M/m/m 排队理论解释了节点度的稳定是靠大量动态变化的连接来保证的.

本节对目前关于 Churn 的度量研究进行了总括,从时间、频度和连接度 3 个不同方面全面地给出有关 Churn 度量的研究现状.Churn 的这 3 个方面度量指标之间有哪些内在联系有待于进一步研究.Stutzbach 等人[44]通过网络探测方法发现,在非结构化 P2P 网络中,Churn 导致洋葱状的节点连接态势,因为每个节点都喜欢连接到当前在线时长较长的节点,这样生存时长较长的节点成为稳定的核,核节点保证了在 Churn 下非结构化 P2P 网络的连通.Yao 等人[13]通过理论分析方法同样发现,在非结构化 P2P 网络中,停留时间较长的节点其连接度较大.

3 Churn对P2P网络性能的影响

P2P 网络性能的优劣直接关系到它未来的发展前景,因大部分 P2P 网络常处于 Churn 环境中,故在 Churn 下评估 P2P 网络的性能已成为目前非常重要的研究方向.通过系统地归纳文献发现:相关研究集中于在 Churn 下对 P2P 网络(包括非结构化和 DHT)的性能进行理论分析和对 DHT 网络的性能进行系统仿真实验评估.理论分析是指,通过合理假设,利用相关理论知识对 Churn 下 P2P 网络的性能问题建立数学模型,分析 Churn 对 P2P 网络性能的影响及 Churn 下 P2P 网络性能的演化趋势.该方法的优点在于,可对某些具体细节放大并进行中立分析.系统仿真实验是指在 DHT 网络系统仿真平台上进行大量实验后获得实验数据,通过对实验数据的分析掌握 Churn 对 DHT 网络性能的影响.本节详细总结了上述个两方面的最新研究成果,并指出未来可能的研究方向.

3.1 在Churn下P2P网络性能的理论分析

关于在 Churn 下 P2P 网络性能的理论分析研究已经成为当前热门的研究领域,其主要关注:(1) 在 Churn 下 P2P 网络仍然保持连通(连通是指 P2P 网络中的任意两个节点之间都有通路)的能力;(2) 在 Churn 下 P2P 网络的查找性能,如平均查找路径长度、平均查找成功率等.

3.1.1 在 Churn 下 P2P 网络连通性的理论分析

Pandurangan 等人[45]最先对 Churn 下非结构化 P2P 网络的连通性进行研究.文中假设节点的加入率服从泊松分布,生存时长服从指数分布,利用 M/M/1 排队论对常度数(常度数是指 P2P 网络中节点的度数接近常数,假设节点的邻居失效后立即被新邻居取代)的非结构化 P2P 网络建立模型,得到结论:存在一个正常数 c,满足对于任意给定时刻 $t>c\times\log N$,有 $P(G_t \text{ is connected})\geq 1-O((\log^2 N)/N)$,其中 G_t 表示 t 时刻的非结构化 P2P 网络,N 表示节点的数目.上式表明,常度数的非结构化 P2P 网络经过短期的初始化之后其保持连通的概率很高.

Leonard 等人[18,46]提出了在 Churn 下 P2P 网络中节点被孤立(节点被孤立是指它的所有邻居都处于失效状态)的理论模型,对节点被孤立之前在系统中的停留时长(time to isolation,简称 TTI)和节点被孤立的概率进行研究.文中假设 P2P 网络的规模较大且达到稳态,当邻居节点失效后,立即采用随机选择策略从存活节点中选取节点来取代它.设任意存活节点 v 的邻居节点 $i(1\leq i\leq k)$的余留时长为随机变量 R_i,邻居节点 i 失效后被存活节点取代所需时长为随机变量 S_i,在 S_i 相对小的假设下,利用更新过程理论得到:节点被孤立之前,在系统中停留时长的期望 $E[TTI]\approx\frac{E[S_i]}{k}\left[\left(1+\frac{E[R_i]}{E[S_i]}\right)^k-1\right]$,节点被孤立的概率 $\pi\leq\frac{\rho k}{(1+\rho)^k-1}$,其中$\rho=E[L_i]/E[S_i]$.对上述公式进行数值仿真计算与分析后发现:$E[TTI]$对节点的生存时长和邻居取代时长分布不敏感;生存时长服从重尾分布的 P2P 网络的连通能力比服从轻尾分布的要强;只要没有节点的邻居全部同时失效,在 Churn 下 P2P 网络连通的概率就很高.

文献[18,46]采用随机邻居选择策略(参见第 4.3 节)从存活节点中选取节点取代失效的邻居,而 Yao 等人[38]将其改进为根据节点的当前在线时长进行邻居的选择,在此基础上研究了在 Churn 下非结构化 P2P 网络中节点被孤立的概率.文中假设邻居节点的余留时长服从超指数分布,查找延时也服从超指数分布(这样假设的原因是:由文献[47]可知,对于任意单调的概率密度函数都可以用超指数概率密度函数以任意精确度来近似).利用随

机过程理论得到了计算节点被孤立的概率公式$\phi=\pi(0)VBV^{-1}r$(因公式中各个参数的表达式较为复杂,这里不作说明,参见文献[38]),该公式用来计算在根据节点的当前在线时长的邻居选择策略下生存时长服从指数分布或重尾分布的非结构化 P2P 网络中节点被孤立的概率,结果表明,对于生存时长服从重尾分布的非结构化 P2P 网络,采用选择当前在线时长较长的邻居相对于随机邻居选择策略进一步降低了节点被孤立的概率.

Loguinov 等人[48]运用图论知识对一些 DHT 网络的抗分割性进行了研究.文中采用对剖宽度(对剖宽度是指对分网络各半所必须移去的最少的边数[49])指标来衡量 DHT 网络的连通性,它们的对剖宽度值见表 1,其中 N 表示 DHT 网络中节点的个数,k 表示 DHT 网络中每个节点的度数,d 表示 CAN(content addressable network)[8]中标识空间维数(表 1 中 d 为偶数).文中进一步分析表明,De Bruijn 的对剖宽度比其余三者要大,也即在 Churn 下其连通能力比其余三者要强.

Table 1 Bisection width of the DHT networks

表 1 DHT 网络的对剖宽度值

	Chord	CAN	Butterfly	de Bruijn
Bisection width ($bw(G)$)	N	$2N^{\frac{d-1}{d}}$	$\approx\frac{kN}{2\log_k N}$	$\frac{kN}{2\log_k N}(1-O(1))\le bw(G)\le\frac{2kN}{\log_k N}(1+O(1))$

Krishnamurthy 等人[36]利用统计物理学中的主方程方法来分析 Chord 网络断裂的概率.文中设每个节点的加入率为 λ_j、离开率为 λ_f、稳定率为 λ_s.假设系统的规模是稳定的 $\lambda_j=\lambda_f, r=\lambda_s/\lambda_f$(如 r=50,其含义为:如果每隔半小时节点加入或离开发生一次,那么每隔 36 秒节点需要对路由表进行稳定维护一次).设 α 为稳定因子,则节点的后继稳定率为 $\alpha\lambda_s$,节点的 finger 稳定率为$(1-\alpha)\lambda_s$.假设节点的加入率服从泊松分布,生存时长服从指数分布,得到 Chord 网络断裂的概率 $P_{bu}(S)=(S+1)!/2(\alpha r)^S$,其中 S 为后继节点的数目.

以上文献集中于研究 P2P 网络的连通能力,Liben-Nowell 等人[34]给出在 Churn 下 P2P 网络维持连通所需要消耗带宽下界的理论模型.文中用 *half-life*(*half-life*=min(*doubling time*,*halving time*),在时刻 t 有 N 个节点的 P2P 网络中,*doubling time* 是指从 t 开始至又有 N 个节点加入 P2P 网络时刻之间的时间段,*halving time* 是指从 t 开始至有 $N/2$ 个节点离开 P2P 网络时刻之间的时间段)来粗略地衡量 Churn,假设节点到达服从泊松分布,生存时长服从指数分布,得到结论:在 Churn 为 *half-life* τ 的环境下,对于任何有 N 个节点的 P2P 网络,要以 $1-O(1/N)$高概率保持连通,必须以平均每τ 时间通知$\Omega(\log N)$个节点的速度通知每个节点,以便于在 Churn 下节点更新邻居信息来确保 P2P 网络的连通.

以上内容基本包括了近年来关于在 Churn 下 P2P 网络连通性方面所涉及的理论文献,为了简化模型,这些文献普遍采用了较强的假设条件,且涉及较少的参数(从第 3.2 节可知 P2P 网络涉及的参数相当多),如何在弱假设的条件下给出更接近于真实网络环境的连通性理论模型可能成为将来重要的研究方向.

3.1.2 在 Churn 下 P2P 网络查找性能的理论分析

在理论分析中,衡量 P2P 网络查找性能的主要指标是节点之间的平均距离,它是指在 Churn 下进行查找所经过的平均跳数,决定 P2P 网络的查找反应能力和容量[48],不同于网络直径[50].

Loguinov 等人[48]在不考虑 Churn 的情况下利用图论知识对一些 DHT 网络中节点之间的平均距离进行了分析,平均距离值见表 2,其中,N 表示 DHT 网络中节点的个数,k 表示 DHT 网络中每个节点的度数,D 表示 DHT 网络的直径.文中通过分析表明 De Bruijn 的平均距离比其余三者要小.

Table 2 Average distance between all pairs of nodes in DHT networks

表 2 DHT 网络中节点之间的平均距离

	Chord	CAN	Butterfly	De Bruijn
Average distance (μ_d)	$\approx\frac{D}{2}$	$\approx\frac{D}{2}$	$\approx\frac{3\log_k N}{2}$	$\approx D-\frac{1}{k-1}$

在系统中节点数量稳定且即使一些节点失效,系统的路由功能仍正常的假设下,Wang 等人[35]分析了在节点失效情况下,DHT 网络中两种不同类型的邻居对其查找性能的影响.一类是 Special 邻居,它在失效后立即得

到重建(被其他节点所取代),如 Chord 协议中的后继邻居列表;另一类是 Finger 邻居,它在失效后很长时间后才重建或者不再重建,如 Chord 协议中的 Finger 邻居列表.文中利用 Markov 链对处于节点失效环境下的查找过程进行建模,得到平均查找距离 $\bar{m}=(\pi^0 M)/n$ 和平均命中成功率 $\bar{a}=(\pi^0 A+1)/n$ (因公式中各个参数的表达式较为复杂,这里不作说明,参见文献[35]),通过分析公式发现,Finger 邻居对平均查找距离影响显著,而 Special 邻居对平均查找命中率影响显著.

Krishnamurthy 等人[36]利用随机微分方程对 Churn 下 Chord 协议的查找性能进行分析(模型假设及相关说明见第 3.1.1 节),得到结论:(1) 节点间隔长度 x 的分布为 $P(x)=\rho^{x-1}(1-\rho)$,其中,$\rho=(K-N)/K$,K 为 Hash 空间大小,N 为节点总数;(2) 在查找过程中假设到达失效节点所用时间等价于查找 1 跳所用时间,平均查找距离为 $L(r,\alpha)=\sum_{i=1}^{K-1}C_i(r,\alpha)/K$,其中 $C_i(r,\alpha)$表示节点查找关键字距离自己 i 远的节点所需的跳数,具体计算公式见文献[36].

对于非结构化 P2P 网络,Pandurangan 等人[45]通过理论分析(模型假设及相关说明见第 3.1.1 节)得到,对于任意时刻 t 满足 $t/N\to\infty$,在 Churn 下非结构化 P2P 网络拥有 $O(\log N)$规模直径的概率为 $1-O((\log^2 N)/N)$.结果表明,常度数非结构化 P2P 网络拥有对数规模直径的概率很高.

目前有关 Churn 下 P2P 网络查找性能的理论分析工作相对甚少,且模型中假设条件较强.在现有模型的基础上,弱化假设条件并且加入邻居选择策略和路由策略值得进一步深入研究.

3.2 在Churn下DHT网络性能的实验评估

理论分析方法是在一定的假设下对 DHT 网络中的某些特定细节进行中立分析,涉及参数较少.而系统仿真实验方法将 DHT 网络涉及的众多参数编写到同一个系统仿真平台中,以综合考察 Churn 与系统性能及系统中各参数之间的关系.

3.2.1 参数及仿真方法

为了理清 DHT 网络涉及的众多参数,根据 DHT 网络的构件将参数分为 DHT 协议参数、IP 层参数、节点相关参数、查找相关参数、共享资源参数以及输出的性能参数.表 3 列出了 DHT 网络涉及的主要参数.

Table 3 Parameters of the DHT networks

表 3 DHT 网络中的参数

DHT network protocol	IP level	Peer	Lookup	Resource	Performance of DHT network
Overlay geometry	IP topology	Number of peers	Lookup frequency	Resource space	Bandwidth consumption
ID space size	Round-Trip time	**Average session time**	Lookup timeout	Resource distribution	Lookup latency
Routing table size		**Join rate**			Lookup success rate
Stabilization interval		**Leave rate**			
Routing algorithm					

协议参数:尽管不同的 DHT 协议有各自不同的参数,但在仿真实验中主要关注的共同参数有:逻辑拓扑几何结构、节点标识空间大小、路由表的大小(路由表中邻居信息条目数)、路由表的稳定时长以及协议采用的路由算法.通常逻辑拓扑几何结构决定着路由表的邻居信息条目数.文献[16]详细总结了目前主要 DHT 协议的逻辑拓扑结构及其路由表的大小.

IP 层参数:IP 层参数主要包括 IP 层拓扑数据集和 IP 层的往返时延两个参数.仿真实验中常用的 IP 层拓扑数据集是 King Data[51],它是利用 DNS 服务器之间的延时构造的 IP 层拓扑环境.

节点相关参数:节点相关参数主要用来描述 DHT 网络中节点的行为,主要包括节点的数目、节点的平均会话时长、节点的加入率与节点的离开率,通常将后三者(表 3 中加粗部分文字)称为 Churn 参数.

查找相关参数:查找相关参数主要包括查找频率和查找超时.其中,查找频率是指每个节点单位时间内发出的平均查找次数[52],用来描述 DHT 网络的查找繁忙程度;查找超时是查找结束的时间窗口,这是一个非常重要的参数,它的取值要适中[21,52,53](见第 4.4 节).

资源参数:资源参数用来描述 DHT 网络中资源的特征,包括资源空间大小、分布特征等[11].

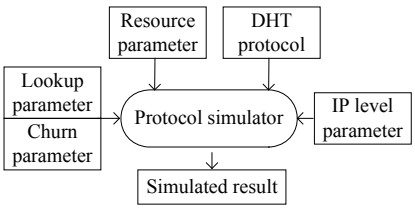

Fig.2 Simulator and work parameters

图 2 仿真系统及工作参数

性能参数:性能参数是指在仿真系统上运行具体的 DHT 协议后得到的性能指标,主要包括:(1) 带宽消耗,是指每个存活节点每秒接收和发送消息(主要包括路由表更新产生的流量、查找产生的流量和加入产生的流量)的字节数,用来衡量通信代价;(2) 查找延时,是指查找请求发出至查找结果返回之间的时间间隔,主要用来衡量路由效率;(3) 查找成功率,是指在查找超时窗口内查找返回的正确结果数量与所有返回结果数量之比,主要用来衡量路由的正确程度.

对于给定的仿真系统,如 P2Psim[54]等(文献[55]详细总结了现有大部分仿真系统),向其中加载 DHT 网络协议、资源参数、IP 层参数、查找参数和 Churn 参数,运行仿真系统后得到相应的结果数据.通常将查找参数和 Churn 参数合称为工作任务参数[52].工作原理如图 2 所示,形式化描述如下:

DHT 网络性能结果集=*Simulator*(DHT 协议参数集,IP 层参数集,资源参数集,工作任务集).

在进行仿真实验时,改变要考察参数(如 Churn 参数)的取值,固定其余参数的取值,通过大量实验得到考察参数与不同性能指标之间的关系,以此来研究考察参数对网络性能指标的影响.

3.2.2 在 Churn 下实验评估 DHT 网络性能

Rhea 等人[21]利用仿真平台对 Bamboo 等协议进行实验,着重研究了会话时长与查找一致性、查找延时之间的关系.查找一致性是指从不同节点出发查找同一个关键字,查找结果大部分是一致的.实验结果表明:当平均会话时长达到 47 分钟时,对于 Bamboo 与 Chord 协议,其查找结果绝大部分是一致的;查找延时随着平均会话时长的增加而降低,并趋于平稳.Herrera 等人[11,56]对 Chord,Tapestry 协议进行实验仿真研究,通过改变平均会话时长的取值得到它与网络性能之间的关系图,图中清楚地反映出,平均会话时长对网络性能影响显著,平均会话时长缩短导致平均查找延时增加、平均带宽消耗增加以及平均查找成功率降低.

目前已有文献除了研究 Churn 对 DHT 网络性能的直接影响之外,还研究在 Churn 下性能指标之间的关系.Li 等人[52,57]利用 P2Psim[54]仿真平台对 Chord,Kademlia 等 5 个 DHT 协议进行实验,着重研究了在 Churn 下查找延时与带宽消耗之间的关系.设每个节点加入、离开的时间间隔服从指数分布,其均值都为 1 小时,平均查找时间间隔为 10 分钟.实验结果表明,平均查找延时与带宽消耗之间存在折衷关系,小的查找延时需要以大的带宽消耗为代价,而小的带宽消耗会导致大的查找延时.此外,在以平均带宽消耗为 X 轴和平均查找延时为 Y 轴的二维空间内,文献还研究了具体参数的不同取值(如 Chord 协议中,*Routing table size* 分别取 10,23,62)对 DHT 网络性能的影响,并在最佳参数取值下比较了不同 DHT 协议的性能.其中,在 Churn 下比较不同 DHT 协议的性能是目前研究的热点和难点,热是因为研究者非常希望明确现有 DHT 协议的性能孰好孰坏,难是因为每个 DHT 协议涉及参数众多,在实验中取不同的协议参数值会得到不同的性能值数据,由于缺乏一个不同 DHT 协议公认的比较标准,造成不同 DHT 协议性能的可比性不强.有关该问题的研究可能仍将继续进行下去.

虽然实验仿真平台工作原理与真实 DHT 网络工作原理一致,但它仍有相当大的局限性.仿真平台与真实世界的 DHT 网络存在着不一致之处,尤其是在 IP 层,仿真平台不考虑 IP 层的数据传输问题[52].

4 降低Churn对P2P网络影响的策略

最后一个要解决的问题是采取什么措施来应对 Churn 对 P2P 网络性能的影响.特别是对于 DHT 网络,文献[58]指出,在高 Churn 下如何使 DHT 网络提供高性能目前是一个开放的问题,也是一个非常活跃的研究领域.通过第 3 节对 Churn 危害的分析可知,目前主要从数据层、路由层、节点层和查找层来应对 Churn.在数据层采用数据冗余策略;在路由层采用路由维护策略;在节点层采用节点选择策略;在查找层设定合适的查找超时以及采用并行查找算法.本节从上述 4 个方面详细、深入地归纳了应对 Churn 的策略,并指出未来可能的研究方向.

4.1 数据冗余策略

为了应对节点失效导致其上存储的数据丢失问题,最基本的解决方法是对存储数据进行一定的冗余.目前数据冗余策略已成功地嵌入 P2P 文件存储系统(如 MIT 的 CFS[59],UC Berkeley 的 OceanStore[60]等).由于 P2P 文件存储系统对加入节点进行认证管理,系统中的节点较为稳定,故数据冗余策略容易部署.而对于 P2P 文件共享系统,由于节点随意加入或离开系统,故 Churn 问题使得在其中部署数据冗余策略变得非常困难,在目前流行的 P2P 文件共享系统中还没有嵌入数据冗余策略,该问题尚处于研究阶段,且有待于进一步深入研究.目前在 P2P 文件共享系统中,文件的可用性主要依赖于其流行度(popularity).流行度高的文件在 P2P 网络中有较多的副本,从而保证了其高可用性.相反地,对于流行度低的文件,其可用性较差[61].

无论是 P2P 文件存储系统还是文件共享系统,我们通过系统归纳已有文献发现数据冗余策略研究主要集中于 4 个方面.(1) 数据冗余方法.目前主要采用副本(replication)和纠删码(erasure codes)两种冗余方法.副本是指在多个节点上存储同一个文件,该方法适用于冗余小文件和访问频繁的文件.纠删码[62]是指将要存储的数据先切分为 m 个部分,然后通过编码算法生成 $n(n>m)$个部分,使用其中任意 $r(r\geq m)$个部分,通过解码算法来恢复原始数据.采用纠删码冗余方法会由于编码和解码操作增加一定的计算量,但是该方法可以在较少冗余数据的情况下提供与副本方法相同的效果.(2) 对部分数据进行适度冗余.适度冗余是保证数据可用性的前提,而过量的数据冗余会增加系统的维护负担,降低系统的性能,特别是对于 P2P 文件存储系统.(3) 确定冗余数据的合适存储位置.合适的存储位置(如将数据存储至在线时长较长的节点上)能够增强数据的可用性.(4) 冗余数据的错误检测与维护策略.由于存储冗余数据的节点同样会失效,因此需要有效的机制来发现失效节点并修复丢失的冗余数据.在 P2P 文件存储系统中常采用定期心跳法和失效事件广播法来发现失效节点,采用立即修复和延迟修复策略[63]来修复丢失的冗余数据.Datta 等人[64]利用 Markov 模型对 Churn 下的修复策略进行比较,并提出了随机延迟修复策略.

Tian 等人[65]从以上 4 个方面对 P2P 文件存储系统进行了详细的总结与分析.针对 P2P 文件共享系统的数据冗余策略研究较少,最典型的是文献[66].Cohen 等人[66]以平均查找距离(expected search size,简称 ESS)为度量指标对非结构化 P2P 文件共享系统建立线性规划模型,假设每个节点的容量相同,且文件被分发到每个节点上的概率相等,目标函数: $ESS_q(r)=\text{Minimize}\ \frac{R}{\rho}\sum_{i=1}^{m}\frac{q_i}{r_i}$, 约束条件 $R=\sum_{i=1}^{m}r_i$ 和 $\sum_{i=1}^{m}q_i=1$,其中 m 为待复制的不同文件的个数,r_i 为文件 i 的副本数目,R 为系统的总存储容量,ρ为每个节点的存储容量,q_i 为文件 i 被查找的概率.求解后得到,当 $r_i=R\sqrt{q_i}\Big/\sum_{i=1}^{m}\sqrt{q_i}$ 时,最小平均查找距离值 $ESS=\frac{\left(\sum_{i=1}^{m}\sqrt{q_i}\right)^2}{\rho}$,在实际应用中还不知道如何部署达到该理论最小值的复制策略.此外,还利用该模型对均匀复制($r_i=R/m$)与按比例复制(按照查询概率进行复制,$r_i=R\times q_i$)进行比对,发现两种复制策略具有相同的平均查找距离 $ESS=m/\rho$,在按比例复制策略中流行度较高的文件很容易被查到,而流行度较低的文件则恰好相反.

4.2 路由维护策略

通过优化 DHT 网络路由表的配置来应对 Churn 带来的危害.DHT 网络路由表的配置主要关注邻居信息的组成、路由表的大小和路由表的更新.对于特定的 DHT 网络,对于特定的 DHT 网络,其邻居信息的组成是根据逻辑拓扑结构来确定,从 Churn 角度出发,有关它的研究很少.目前研究主要集中于路由表的大小和路由表的更新.

路由表的大小对平均查找距离与带宽消耗有着显著的影响.目前认为[40,50]较小的路由表适合于高 Churn 环境,如 Chord,Kademlia 等协议;而较大的路由表适合低 Churn 环境,如单跳路由(one-hop)协议.为了降低查找延时,Li 等人[40]提出节点根据对 DHT 网络运行状况观测的结果调节路由表的大小.在 DHT 网络运行状况良好的情况下,路由表可增至很大,直至存储所有节点信息;在运行状况较差(如 Churn 高)的情况下,路由表的大小减少

到 $O(\log_2 N)$(N 是节点数目).Li 等人在文献[52]中具体提出,根据 DHT 网络的额外带宽和 Churn 大小自适应地调节路由表的大小.

路由表的更新时长与更新内容对查找延时影响显著,一些 DHT 协议(如 Chord)周期性地检测邻居节点的有效性并更新路由表中无效的邻居信息,另一些 DHT 协议(如 Kademlia)采用触发式方法更新邻居信息.所谓触发式更新是指在路由过程中,当发现邻居节点不可达时立即更新对应的邻居信息.Rhea 等人[21]对 Bamboo 协议进行实验仿真,从带宽消耗和查找延时性能评价指标方面比对了触发式更新与周期更新,发现周期更新优于触发式更新.Herrera 等人[56]对 Chord 和 Tapestry 协议进行实验仿真,对固定更新周期与根据 Churn 大小动态调整更新时长进行比对,结果表明,根据 Churn 动态调整更新时长可以有效地提高上述协议的性能.此外,Kersch 等人[58]对路由表中的邻居进行区分,根据在散列空间上距离当前节点散列值的远近将路由表中的邻居分为局部邻居和远距离邻居.其中,对于局部邻居信息采用周期主动探测更新方法,一旦发现错误邻居信息则立即进行更新,并按照某种确定性的策略选择新邻居(邻居选择策略见第 4.3 节);而对于远距离邻居信息则采用触发式更新方法,随机选择新邻居.该文献利用 Markov 链建立模型对文中提及的路由更新策略进行理论分析,结果表明,该策略可以有效地降低高 Churn 下路由表的维护代价,可在未降低 DHT 网络路由性能的前提下获得理论上路由表的维护代价下界.

此外,Castro 等人[53]在 Pastry 协议中引入路由表的连续错误检测与修复技术,使得在低开销的情况下提高了 Pastry 协议的查找可靠性.Lam 等人[42]对路由表中的路由条目的组成作了相应的限定(K-consistency[42]),以加强在 Churn 下 DHT 网络的连通性.

4.3 节点选择策略

为了降低 Churn 对 P2P 网络性能的影响,常根据特定的选择策略从系统的可用节点集中选择部分节点来使用[3].节点选择策略可分为随机选择和确定性选择两大类,其中随机选择策略是指不考虑节点的特征随机选择节点,确定性选择策略是指根据节点的某种特征选择满足一定标准的节点,如根据节点的当前在线时长、带宽吞吐能力、邻近地理位置(如文献[16,21])等选择节点.下面从应对 Churn 方面出发,对随机选择和根据当前在线时长选择策略进行总结.

由于节点的当前在线时长对其余留时长有一定的预见性(参见第 2.1.2 节),因此选择最长当前在线时长的策略用于 DHT 网络中邻居的选择[67]和超级节点的选择[68],以及覆盖层多播树中双亲节点的选择[69].Li 等人[40]针对 Accordion 协议(DHT 网络)通过计算生存时长的条件概率来选择邻居节点,在生存时长服从 Pareto 分布的假设下,计算条件概率 $P(lifetime>(\Delta t_{alive}+\Delta t_{since})|lifetime>\Delta t_{alive})$,其中 Δt_{alive} 为邻居节点的当前在线时长,Δt_{since} 为邻居节点的上次确认存活时刻至当前的时长.Godfrey 等人[3]将不同的选择策略应用于 5 个当前广泛使用的 P2P 网络中,在其自定义的 Churn 频度度量指标(见 2.2 节)下比较这些选择策略的优劣,实验结果表明,与其他选择策略相比,随机选择策略可以有效地降低 P2P 网络的 Churn,这样的结果是研究者所未预料到的,文中从直觉和理论分析两个方面对这种现象进行了解释.由于随机选择策略具有简单、易实现等特点,因而被广泛地应用于 P2P 网络中,如 P2P 文件存储系统 Total Recall[63]等.

4.4 查找配置策略

在查找层主要通过选择合适的查找方法、控制查找超时、采用并行查找算法等技术来应对 Churn 对 P2P 网络性能的影响.在 DHT 网络中常用的查找方法包括迭代查找和递归查找.迭代查找是指当前节点直接向其每个后继节点发送查找消息,使用该查找方法的有 Chord,Kademlia 协议.递归查找是指当前节点向其直接后继发送查找消息,然后将后继节点作为当前节点重复上述操作,使用该查找方法的有 Tapestry 协议.迭代查找比递归查找消耗较大的带宽,该方法可以直接发现失效的节点,而递归查找却不能.Dabek 等人[70]针对 Chord 协议以查找延时为衡量指标实验对比了这两种查找方法,实验结果表明,递归查找方法优于迭代查找方法.Wu 等人[71]应用概率论知识建立数学模型,对 Churn 下的这两种查找方法进行了理论分析比对.模型假设网络中物理链路处于理想状态且在数据传输中无丢包,查找失败仅仅是因 Churn 所致,并设节点的生存时长服从特定的分布.文中

讨论了指数和 Pareto 分布,理论分析结果表明,在 Churn 下以查找延时和带宽消耗两个指标来衡量,迭代查找优于递归查找(该理论结果与文献[70]的实验结果相矛盾,有待于进一步深入研究).在此基础上,本文提出两阶段的查找策略:首先使用递归查找进行快速查找,一旦查找失败,则转向迭代查找.还提出递归查找加 ACK 确认机制查找策略.分析结果表明,这两种查找方法比递归查找更优.

查找超时的取值对 Churn 下 DHT 网络性能的影响显著[21].在 Churn 下的 DHT 网络中,如果查找超时设置太短,将会导致查找仍在正常处理中就过早地结束;相反,查找超时设置太长,将会导致查找发起节点浪费时间等待因邻居节点失效所致的不能返回的查找消息.Rhea 等人[21]讨论了定值查找超时、将直接测量所得的查找响应时间值代入 TCP 协议中计算 RTO 的公式中计算查找超时、基于 Vivaldi 坐标计算查找超时[70],实验结果表明,在 Churn 下采用递归查找方法,以 TCP 风格设定查找超时,可以有效地降低查找延时.

在 Churn 下并行查找可以有效地提高 DHT 网络的查找效率,但同时也增加了带宽的消耗[52,71].Li 等人[52]通过仿真实验对采用并行查找提高 DHT 网络的查找效率进行了简单的分析.Wu 等人[71]从理论上对并行查找度(查找发起节点同时发出同样的查找请求的数目)进行了分析,在给定的假设下(见本节第 1 段),采用迭代方法查找,研究得到最优并行查找度为 2 或 3.Stutzbach 等人[72]在 Kademlia 协议中研究了并行查找度数,得到最优并行查找度为 3.上述研究尽管采用了不同的方法,研究结果却非常接近.

除了上述应对 Churn 的策略之外,还有通过激励机制使用户在线时间尽可能长、采用抗 Churn 的逻辑拓扑结构(如 De Bruijn 图[48]、k 规则图[18])等策略.这些应对 Churn 的策略来自于 Churn 对 P2P 网络性能影响的深刻认识,因此只有进一步全面、深入地分析 Churn 对 P2P 网络性能的影响,才能发现应对 Churn 的更好策略.

以上总结的这些策略都可以有效地降低 Churn 对 P2P 网络性能的负面影响.其中,数据冗余策略是在数据层以牺牲 P2P 系统的存储空间为代价来应对 Churn 下因节点离开而导致的数据丢失问题,以及提高非结构化 P2P 网络中数据的查询效率.路由表的规模直接影响查找跳数和维护代价(二者是一对矛盾),为了使 P2P 网络性能提高,路由维护策略要求根据 Churn 的大小调节路由表的大小及其更新频度.节点选择策略需要对节点的余留时长有预见性,然而预判节点的余留时长是很困难的,因此实施难度大.查找配置策略是查找层的优化问题,需要根据 Churn 的实际大小优化查找超时和并行查找度.

5 总结及展望

本文首先通过系统地归纳现有文献,从 Churn 产生的机理出发,总结出解决 Churn 问题的主要步骤,它们依次是准确度量 Churn,分析 Churn 对 P2P 网络性能的影响,给出应对 Churn 的具体策略.然后,进一步归纳总结出在每个步骤中分别需要解决的关键问题,以及目前采用的解决这些问题的方法和最新研究成果.最后,围绕要解决的问题将孤立的、分割的研究有机地整合在一起,在此基础上讨论了各部分存在的问题,并指出了未来可能的研究方向.

从目前的研究状况来看,尽管取得了一定成果,但距离认清 Churn 对 P2P 网络性能的影响、提出更加有效的应对 Churn 的策略以及建立完整的理论模型对该问题进行分析还很远,仍然需要继续深入研究.目前相关的研究都集中于 Churn 对 P2P 网络性能造成的负面影响,Churn 对 P2P 网络性能有无积极的影响,特别是对非结构化 P2P 网络,是非常值得研究的.

致谢 在此,我们向本文的审稿老师表示感谢,感谢他们认真、细致的工作以及提出的宝贵意见,并向对本文的工作给予支持和建议的老师和同学表示感谢,尤其是林福宏、吴恒奎同学.

References:

[1] Stoica I, Morris R, Liben-Nowell D, Karger DR, Kaashoek MF, Dabek F, Balakrishnan H. Chord: A scalable peer-to-peer lookup service for Internet applications. IEEE/ACM Trans. on Networking, 2003,11(1):17−32.

[2] Zhao BY, Huang L, Stribling J, Rhea SC, Joseph AD. Tapestry: A resilient global-scale overlay for service deployment. IEEE Journal on Selected Areas in Communications, 2004,22(1):41−53.

[3] Godfrey BP, Shenker S, Stoica I. Minimizing churn in distributed systems. In: Proc. of the ACM SIGCOMM 2006. New York: ACM Press, 2006. 147−158.

[4] Milojicic DS, Kalogeraki V, Lukose R, Nagaraja K, Pruyne J, Richard B, Rollins S, Xu ZC. Peer-to-Peer computing. Technical Report, HPL-2002-57, Palo Alto: HP Labs, 2002.

[5] Andy O. Peer-to-Peer: Harnessing the Power of Disruptive Technologies. Cambridge: O'Reilly & Associates, Inc., 2001. 3−159.

[6] Wang CG, Li B. Peer-to-Peer overlay networks: A survey. Technical Report, Hong Kong: Hong Kong University of Science & Technology, 2003.

[7] Lua EK, Crowcroft J, Pias M, Sharma R, Lim S. A survey and comparison of peer-to-peer overlay network schemes. Journal of IEEE Communications Survey and Tutorial, 2005,7(2):72−93.

[8] Ratnasamy S, Francis P, Handley M, Karp R, Shenker S. A scalable content-addressable network. In: Govindan R, ed. Proc. of the ACM SIGCOMM. New York: ACM Press, 2001. 161−172.

[9] 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 Int'l Conf. on Distributed Systems Platforms (Middleware 2001). Berlin: Springer-Verlag, 2001. 329−350.

[10] Wu D, Tian Y, Ng KW, Datta A. Stochastic analysis of the interplay between object maintenance and churn. Computer Communications, 2008,31(2):220−239.

[11] Herrera O, Znati T. Modeling churn in P2P networks. In: Proc. of the 40th Annual Simulation Symp. Washington: IEEE Press, 2007. 33−40.

[12] Stutzbach D, Rejaie R. Understanding churn in peer-to-peer networks. In: Proc. of the 6th ACM SIGCOMM on IMC. New York: ACM Press, 2006. 189−202.

[13] Yao Z, Leonard D, Wang X, Loguinov D. Modeling heterogeneous user churn and local resilience of unstructured P2P networks. In: Proc. of the IEEE Int'l Conf. on Network Protocols (ICNP 2006). Washington: IEEE Press, 2006. 32−41.

[14] Li CX, Chen CJ. On Gnutella topology dynamics by studying leaf and ultra connection jointly in phase space. Computer Networks, 2008,52(3):695−719.

[15] Liu Z, Yuan R, Li Z, Li H, Chen G. Survive under high churn in structured P2P systems: Evaluation and strategy. In: Proc. of the ICCS 2006. Berlin: Springer-Verlag, 2006. 28−31.

[16] Gummadi K, Gummadi R, Gribble S, Ratnasamy S, Shenker S, Stoica I. The impact of DHT routing geometry on resilience and proximity. In: Proc. of the ACM SIGCOMM 2003. New York: ACM Press, 2003. 381−394.

[17] Kunzmann G, Binzenhöfer A, Henjes R. Analyzing and modifying chords stabilization algorithm to handle high churn rates. In: Proc. of the MICC & ICON 2005. Piscataway: IEEE Press, 2005. 885−890.

[18] Leonard D, Yao Z, Rai V, Loguinov D. On lifetime-based node failure and stochastic resilience of decentralized peer-to-peer networks. IEEE/ACM Trans. on Networking, 2007,15(3):644−656.

[19] Chu J, Labonte K, Levine B. Availability and locality measurements of peer-to-peer file systems. In: Firoiu V, Zhang ZL, eds. Proc. of the Scalability and Traffic Control in IP Networks II. SPIE 4868, Boston: SPIE, 2002. 310−321.

[20] Saroiu S, Gummadi PK, Gribble SD. Measuring and analyzing the characteristics of Napster and Gnutella hosts. Multimedia Systems Journal, 2003,8(5):170−184.

[21] Rhea S, Geels D, Roscoe T, Kubiatowicz J. Handling churn in a DHT. In: Proc. of the USENIX Annual Technical Conf. Boston: USENIX Association Press, 2004. 127−140.

[22] Gnutella. 2005. http://rfc-gnutella.sourceforge.net/

[23] Liu Q, Xu P, Yang H, Peng Y. Research on measurement of peer-to-peer file sharing system. Journal of Software, 2006,17(10): 2131−2140 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/2131.htm

[24] Sen S, Wang J. Analyzing peer-to-peer traffic across large networks. IEEE/ACM Trans. on Networking, 2004,12(2):219−232.

[25] KaZaA. 2003. http://www.kazaa.com

[26] Direct connect. 2001. http://www.neo-modus.com

[27] Gummadi KP, Dunn RJ, Saroiu S, Gribble SD, Levy HM, Zahorjan J. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In: Proc. of the 19th ACM Symp. on Operating Systems Principles (SOSP 2003). New York: ACM Press, 2003. 314−329.

[28] Karagiannis T, Broido A, Brownlee N, Claffy KC, Faloutsos M. Is P2P dying or just hiding. In: Proc. of the IEEE Globecom 2004. Washington: IEEE Press, 2004. 1532−1538.

[29] Napster. 2001. http://www.napster.com

[30] Bustamante F, Qiao Y. Friendships that last: Peer lifespan and its role in P2P protocols. In: Proc. of the 8th Int'l Workshop on Web Content Caching and Distribution (WCW 2003). Norwell: Kluwer Academic, 2003. 233−246.

[31] Liang J, Kumar R, Ross KW. The KaZaA overlay: A measurement study. Computer Networks, 2006,50(6):842−858

[32] Stutzbach D, Rejaie R. Characterizing churn in peer-to-peer networks. Technical Report, CIS-TR-2005-03, University of Oregon, 2005.

[33] Maymounkov P, Mazieres D. Kademlia: A peer-to-peer information system based on the XOR metric. In: Proc. of the 1st Int'l Workshop on Peer-to-Peer Systems (IPTPS 2002). Berlin: Springer-Verlag, 2002. 53−65.

[34] Liben-Nowell D, Balakrishnan H, Karger D. Analysis of the evolution of peer-to-peer systems. In: Proc. of the 21st Annual ACM Symp. on Principles of Distributed Computing (PODC 2002). New York: ACM Press, 2002. 233−242.

[35] Wang S, Xuan D, Zhao W. Analyzing and enhancing the resilience of structured peer-to-peer systems. Journal of Parallel and Distributed Computing. 2005,65(2):207−219.

[36] Krishnamurthy S, El-Ansary S, Aurell E, Haridi S. An analytical study of a structured overlay in the presence of dynamic membership. IEEE/ACM Trans. on Networking, 2008,16(4):814−825.

[37] Mickens JW, Noble BD. Predicting node availability in peer-to-peer networks. In: Proc. of the ACM SIGMETRICS Int'l Conf. on Measurement and Modeling of Computer Systems. New York: ACM Press, 2005. 378−379.

[38] Yao Z, Wang X, Leonard D, Loguinov D. On node isolation under churn in unstructured P2P networks with heavy-tailed lifetimes. In: Proc.of the IEEE INFOCOM 2007. Piscataway: IEEE Press, 2007. 2126−2134.

[39] Bustamante F, Qiao Y. Designing less-structured P2P systems for the expected high churn. IEEE/ACM Trans. on Networking, 2008,16(3):617−627.

[40] Li J, Stribling J, Morris R, Kaashoek F. Bandwidth-Efficient management of DHT routing tables. In: Proc. of the 2nd Conf. on Symp. on Networked Systems Design & Implementation (NSDI 2005). Berkeley: USENI Press, 2005. 99−114.

[41] Zhong M, Shen K, Seiferas J. Non-Uniform random membership management in peer-to-peer networks. In: Proc. of the IEEE INFOCOM. Piscataway: IEEE Press, 2005. 1151−1161.

[42] Lam S, Liu H. Failure recovery for structured P2P networks: Protocol design and performance under churn. Computer Networks, 2006,50(16):3083−3104.

[43] Kumar P, Sridhar G, Sridhar V. Bandwidth and latency model for DHT based peer-to-peer networks under variable churn. In: Proc. of the IEEE Systems Communications (ICW 2005). Piscataway: IEEE Press, 2005. 320−325.

[44] Stutzbach D, Rejaie R, Sen S. Characterizing unstructured overlay topologies in modern P2P file-sharing systems. IEEE/ACM Trans. on Networking, 2008,16(2):267−280.

[45] Pandurangan G, Raghavan P, Upfal E. Building low-diameter peer-to-peer network. IEEE Journal on Selected Areas in Communications (JSAC), 2003,21(6):995−1002.

[46] Leonard D, Yao Z, Wang X, Loguinov D. On static and dynamic partitioning behavior of large-scale networks. In: Proc. of the 13th IEEE Int'l Conf. on Network Protocols (ICNP 2005). Piscataway: IEEE Press, 2005. 345−357.

[47] Feldmann A, Whitt W. Fitting mixtures of exponentials to long-tailed distributions to analyze network performance models. Performance Evaluation, 1998,31(3-4):245−279.

[48] Loguinov D, Casas J, Wang X. Graph-Theoretic analysis of structured peer-to-peer systems: Routing distances and fault resilience. IEEE/ACM Trans. on Networking, 2005,13(5):1107−1120.

[49] Leighton FT. Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes. New York: Academic/Morgan Kaufmann Publishers, 1991.

[50] Xu J, Kumar A, Yu XX. On the fundamental tradeoffs between routing table size and network diameter in peer-to-peer networks. IEEE Journal on Selected Areas on Communications (JSAC), 2004,22(1):151−163.

[51] King Data Set. 2006. http://pdos.csail.mit.edu/p2psim/kingdata/

[52] Li J, Stribling J, Morris R, Kaashoek F, Gil T. A performance vs. cost framework for evaluating DHT design tradeoffs under churn. In: Proc. of the 24th IEEE INFOCOM 2005. Piscataway: IEEE Press, 2005. 225−236.

[53] Castro M, Costa M, Rowstron A. Performance and dependability of structured peer-to-peer overlays. In: Proc. of the Dependable Systems and Networks (DSN 2004). Los Alamitos: IEEE Press, 2004. 9−18.

[54] Gil TM, Kaashoek F, Li J, Morris R, Stribling J. P2Psim: A simulator for peer-to-peer protocols. 2006. http://pdos.csail.mit.edu/p2psim/

[55] Naicken S, Basu A, Livingston B, Rodhetbhai S. A survey of peer-to-peer network simulators. In: Proc. of the 7th Annual Postgraduate Symp. Liverpool: Liverpool John Moores University, 2006. 1−8.

[56] Herrera O, Znati T. Static resiliency vs. churn-resistance capability of DHT-protocols. In: Proc. of the18th Int'l Conf. on Parallel and Distributed Computing Systems (PDCS 2005). Phoenix: IASTED/ACTA Press, 2005. 141−147.

[57] Li J, Stribling J, Gil T, Morris R, Kaashoek MF. Comparing the performance of distributed hash tables under churn. In: Proc. of the

3rd Int'l Workshop on Peer-to-Peer Systems (IPTPS 2004). Berlin: Springer-Verlag, 2004. 87−99.

[58] Kersch P, Szabo R, Cheng L, Jean K, Galis A. Stochastic maintenance of overlays in structured P2P systems. Computer Communications, 2008,31(3):603−619.

[59] Dabek F, Kaashoek M, Karger D, Morris R, Stoica I. Wide-Area cooperative storage with CFS. In: Proc. of the 18th ACM Symp. on Operating Systems Principles. New York: ACM Press, 2001. 202−215.

[60] Wells C. The oceanstore archive: Goals, structures, and self-repair [MS. Thesis]. Berkeley: UC Berkeley, 2002.

[61] Knezevic P, Wombacher A, Risse T. Managing and recovering high data availability in a DHT under churn. In: Proc. of the 2nd Int'l Conf. on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2006). Piscataway: IEEE Press, 2006. 45−55.

[62] Rizzo L. Effective erasure codes for reliable computer communication protocols. Computer Communication Review, 1997,27(2): 24−36.

[63] Bhagwan R, Tati K, Cheng Y, Savage S, Voelker G. Total recall: System support for automated availability management. In: Proc. of the 1st ACM/Usenix Symp. on Networked Systems Design and Implementation (NSDI 2004). San Francisco: USENI Press, 2004. 337−350.

[64] Datta A, Aberer K. Internet-Scale storage systems under churn—A study of the steady-state using Markov models. In: Proc. of the 6th IEEE Int'l Conf. on Peer-to-Peer Computing (P2P 2006). Piscataway: IEEE Press, 2006. 133−144.

[65] Tian J, Dai YF. Study on durable peer-to-peer storage techniques. Journal of Software, 2007,18(6):1379−1399 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/1379.htm

[66] Cohen E, Shenker S. Replication strategies in unstructured peer-to-peer networks. In: Proc. of the ACM SIGCOMM. New York: ACM Press, 2002. 177−190.

[67] Ledlie J, Shneidman J, Amis M, Mitzenmacher M, Seltzer M. Reliability- and capacity-based selection in distributed hash tables. Technical Report, Harvard University Computer Science, 2003. 1−14.

[68] Garces-Erice L, Biersack EW, Ross KW, Felber PA, Urvoy-Keller G. Hierarchical Peer-to-Peer systems. In: Proc. of the ACM/IFIP Int'l Conf. on Parallel and Distributed Computing (Euro-Par). Berlin: Springer-Verlag, 2003. 1230−1239.

[69] Sripanidkulchai K, Ganjam A, Maggs B, Zhang H. The feasibility of supporting large-scale live streaming applications with dynamic application end-points. In: Proc. of the ACM SIGCOMM. New York: ACM Press, 2004. 107−120.

[70] Dabek F, Li J, Sit E, Robertson J, Kaashoek MF, Morris R. Designing a DHT for low latency and high throughput. In: Proc. of the 1st USENIX Symp. on Networked Systems Design and Implementation (NSDI 2004). San Francisco: USENI Press, 2004. 85−98.

[71] Wu D, Tian Y, Ng KW. Analytical study on improving DHT lookup performance under churn. In: Proc. of the 6th IEEE Int'l Conf. on Peer-to-Peer Computing (P2P 2006). Piscataway: IEEE Press, 2006. 249−258.

[72] Stutzbach D, Rejaie R. Improving lookup performance over a widely-deployed DHT. In: Proc. of the IEEE INFOCOM. Washington: IEEE Press, 2006. 1−12.

附中文参考文献:

[23] 刘琼,徐鹏,杨海涛,彭芸.Peer-to-Peer 文件共享系统的测量研究.软件学报,2006,17(10):2131−2140. http://www.jos.org.cn/1000-9825/17/2131.htm

[65] 田敬,代亚非.P2P 持久存储研究.软件学报,2007,18(6):1379−1399. http://www.jos.org.cn/1000-9825/18/1379.htm

张宇翔(1975－),男,山西五寨人,博士生,讲师,主要研究领域为分布式网络理论与技术,下一代互联网络服务架构体系.

张宏科(1957－),男,博士,教授,博士生导师,主要研究领域为下一代互联网关键理论与技术.

杨冬(1980－),男,博士,讲师,主要研究领域为分布式网络理论与技术,下一代互联网络服务架构体系.