基于随机抽样过程的 P2P 集群规模估算方法

王潇斌,李 程,石 碧,杨 哲

(苏州大学计算机科学与技术学院,江苏 苏州 215006)

[摘要] 对 P2P 系统的主动测量,可了解其现状和变化趋势,为系统建模和仿真提供可靠的测量依据. 现有的赠券收集者模型,过度依赖于服务器返回的先验知识,导致测量结果不能反映集群的真实规模. 基于随机抽样过程,本文提出了一种 P2P 集群规模主动估算方法. 根据测量过程中不同时刻获取的节点总数 x 及不重复节点数 u,得到集群规模的估计值. 根据理论分析的结果,分别给出了 3 种不同的实验停止条件. 实验结果表明,对于小于 10^5 的集群,本文的估计方法误差不超过 5%.

[关键词] P2P 集群,主动测量,随机抽样过程,集群规模估计

[中图分类号]TP393 [文献标志码]A [文章编号]1001-4616(2014)01-0076-05

Estimation of P2P Swarm Size Based on Random Sampling Process

Wang Xiaobin, Li Chen, Shi Bi, Yang Zhe

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Active measurement on peer-to-peer system is the best way to understand the current situation and its evolution trends. It also provides the real measurement basis to support the research on system simulation and optimization. The coupon collectors model is too dependent on the prior knowledge returned by server, so the measurement results do not reflect the true scale of the P2P swarm. Based on random sampling process, this paper presents a P2P active swarm size estimation method. Based on the total number of nodes x and the unique nodes u, it can estimate the P2P swarm size. According to the theoretical analysis of the results, we give three different experiments stop condition. The experimental results show that for less than 10^5 clusters, this estimation method error does not exceed 5%.

Key words: P2P swarm, active measurement, random sampling, swarm size estimation

对等网络(Peer-tor-Peer,P2P)技术,由于方便、快捷和开销少的优点,得到了广泛的推广. 文件共享系统是 P2P 技术最为广泛的一个应用,如 BitTorrent(BT)和 eMule/eDonkey 等. 但作为一种上层的覆盖(Overlay)网络,P2P 系统并不了解底层网络拓扑. 它们的节点选择策略都是基于覆盖网络设计的,而没有考虑节点在 AS 上的分布问题,从而导致了大量跨 AS 的流量产生,使得整个网络的性能下降.

近几年来,利用拓扑感知(topology-aware)或位置感知(locality-aware)[1-6]等应用层流量优化技术(Application Layer Traffic Optimization,ALTO),使文件共享系统的流量本地化研究一直倍受关注. ALTO 研究的前提,必须通过测量获得 BT 等系统的集群规模及节点分布,以及流量分布的真实数据. 否则会造成实验参数与真实情况偏差较大,导致优化方法的实际效果不佳. 因此,ALTO 研究需要对 P2P 系统进行准确的测量. 但作为一种大规模、自组织的网络应用,传统的测量技术难以对其进行有效测量.

因此,本文针对 P2P 文件共享系统,提出了一种基于随机抽样过程的 P2P 集群规模主动测量及估算方法.

1 P2P 主动测量研究与存在的问题

现有对 P2P 系统的测量研究,根据其测量方法一般分为主动测量和被动测量[7]. 被动测量主要用于

收稿日期:2013-10-20.

基金项目:国家自然科学基金(61070170)、江苏省高校自然科学基金(11KJB520017)、苏州市科技计划项目(SYG201238,SZS0805).

通讯联系人:杨哲,讲师,研究方向:网络测量与管理、网络与信息安全. E-mail:yangzhe@suda.edu.cn

在特定的观测点上分析 P2P 系统的流量特征,如流量大小、带宽和连接时间等[8]. 主动测量,通过与节点的信息交互,既能测量其微观特征,如延迟、内容可用性、上传/下载比等,也可测量其宏观特征,如拓扑结构. ALTO 研究需要测量的是 P2P 集群的规模以及节点分布等数据,因此主要采用主动测量的方法. 随着 BT 等成为最重要的 P2P 应用[9],它也成为目前的主要测量研究对象.

Liu 等人[10]最早针对 8893 个种子集群,通过不断请求 Tracker 服务器的方式获取所需的测量数据. 根据 BT 协议,每次请求 Tracker 服务器时,除了获得若干随机选择的节点信息(IP+port),Tracker 服务器还会返回集群当前的做种节点数(seeders)和下载节点数(leechers). 他们将这两者之和作为集群规模的测量结果. 之后,Wang 等人[11]采用类似的方法测量了 70 000 个视频类的种子集群规模. 这些测量的结果严重依赖于服务器返回值,且测量方法缺乏理论模型的支撑,其测量结果的完整性难以保证.

近几年,Hobfeld[12]和 Zhang[13]等人,基于赠券收集者问题(CCP,Coupon Collector′s Problem),设计主动测量模型,并在 PlanedLab、G-Lab 等公共实验平台上搭建测量环境,对 BT 应用进行了规模更大的测量. 根据 CCP 模型,对规模为 n 个节点的 BT 集群,若每次请求 Tracker 服务器可返回 k 个节点,完成 1 次集群快照(snapshot)所需测量次数 X 的数学期望为 $E(X_n^k)\approx n\times H_n/k$,其中 H_n 为 n 的调和级数. 因此根据 CCP 模型,对于规模为 n 个节点的集群,可用期望 $E(X_n^k)$ 来估计所需要的请求次数. 而且实际需要的请求次数不会偏离该期望值太多,而是高度集中在以期望值为中心的极小的区间内. 由于有了理论模型的定量分析,使得根据测量获得的集群快照的完整性得到了保证. 于是,他们在 PlanetLab 和 G-Lab 上分别部署了大量的测量主机,对 BT 集群进行了大规模的主动测量. 他们发现 BT 集群规模符合帕累托法则(Pareto principle),多数集群的规模偏小,其中超过 82% 的集群中节点数少于 10 个.

这些测量结果为 BT 系统的建模和优化提供了可靠的测量数据. 但其采用的 CCP 模型,虽然可以在测量之前精确给出所需的请求次数,从而可以估计主动测量实验的代价. 但还存在以下 3 个问题.

(1)CCP 模型只能用于对 BT 集群,不适用于 eMule/eDonkey 等系统. 运用 CCP 模型测量 BT 集群时,可以通过服务器的返回值,在测量开始前获知集群的规模 n. 而对于 eMule/eDonkey 等应用,由于服务器不返回类似的信息,因此无法运用 CCP 模型对其集群规模进行测量.

(2)由于 BT 应用支持多 Tracker 索引服务,每个节点可以根据网络状况及其所在位置,自动选择合适的 Tracker 服务器进行注册. 因此,不同 Tracker 返回的集群规模大小 n 可能会有较大的偏差,这样导致根据服务器返回值得到的集群规模与实际情况并不一致.

(3)除了 Tracker 索引服务,现有的 P2P 应用还可以通过 DHT(Distributed Hash Tables)和 PEX(Peer EXChange)的方式,获取集群中更多的节点信息. 而现有的主动测量方法,没有很好地利用这些方式,会遗漏集群中的部分节点,导致其测量的集群规模偏小.

总之,现有测量方法过度依赖于服务器,其返回的集群规模信息的完整性和一致性都无法保证,导致测量结果并不能反映集群的真实规模. 同时,由于 CCP 模型依赖的集群规模 n 这一先验知识,并不能适用于 eMule/eDonkey 等 P2P 文件共享系统的主动测量. 为了克服上节所述现有测量方法的局限性,本文在缺乏集群规模 n 这一先验知识的条件下,提出基于一般随机抽样过程,仅根据主动测量获得的节点数观测序列,可以快速准确地完成集群规模的估算.

2 基于随机抽样过程的集群规模估算方法

2.1 随机抽样过程

对于任何一种 P2P 文件共享应用,都可通过 Tracker 服务器、DHT 和 PEX 3 种方式,主动获取集群中节点的信息,从而得到集群的 1 个快照. 但受协议限制,无论通过何种方式,每次只能获取部分节点信息,且任意两次请求都可能获得重复的信息. 因此,对 P2P 集群的主动测量是 1 个有放回的随机抽样过程[10].

假设 P2P 集群的规模大小为 n,如果 t 时刻已经获得的节点总数为 x,其中不重复的节点数为 u. 随着时间的推移,u 会随着 x 的增加而增大,且逐渐趋向于 n. 因此,u 可以表示为随 x 变化的函数 $u(x)$. 此时,如果通过测量又获得 1 个节点,即节点总数变为 $x+1$,则该节点为新节点的概率为 $(n-u(x))/n$. 因此,当节点总数变为 $x+1$ 时,其中不重复的节点数 $u(x+1)$ 为:

$$u(x+1)=u(x)+\frac{n-u(x)}{n}, \tag{1}$$

于是:

$$u(x+1)-u(x)=1-\frac{u(x)}{n}\Rightarrow\frac{u(x+1)-u(x)}{1}=1-\frac{u(x)}{n}\Rightarrow u'(x)=1-\frac{u(x)}{n}. \tag{2}$$

该方程的解为:

$$u(x)=Ce^{-\frac{x}{n}}+n, \tag{3}$$

其中,C 是 1 个由初始条件决定的常量. 一般的,当 $x=0$ 时,$u(x)=0$. 于是,可得:

$$u(x)=n\cdot(1-e^{-\frac{x}{n}}), \tag{4}$$

从式(4)可知,存在隐函数关系 $f(n,x,u)=0$,使得随机变量 n,x,u 满足:

$$f(n,x,u)=u-n\cdot(1-e^{-\frac{x}{n}})=0. \tag{5}$$

在对 P2P 集群主动测量过程中,在 n 未知的情况下,在任意时刻 t 均可获得一组 x 和 u 的值,通过求解隐函数(5),从而得到 n 的估计值 $\hat{n}$. 由于隐函数(5)无法进行显化,只能通过二分法、牛顿法等算法求解近似解,存在收敛慢、复杂度高等问题. 而且,通过不同时刻的 x 和 u 得到的 $\hat{n}$ 往往差异较大. 因此,本文先对任意时刻 t 的 x 和 u 的关系进行分析,确定计算 $\hat{n}$ 的最佳时机,再实现对集群规模的估算.

2.2　集群规模估算

对任一个 P2P 集群,在每一个时间周期 t 内可通过 Tracker、DHT 和 PEX 3 种方式分别获得集群中 k_1,k_2,k_3 个节点信息,则一共可获得的节点数为 $k=k_1+k_2+k_3$,其中 k,k_1,k_2,k_3 为 4 个独立随机变量. 为便于问题的讨论,本文假设每 1 个时间周期 t 内可获得的节点数为 k 为 1 常数. 于是,随机变量 x 可表示为关于 t 的函数 $x(t)$,即:

$$x(t)=k\cdot t, \tag{6}$$

将上式代入式(4),可得 u 关于 t 的函数:

$$u(t)=n\cdot(1-e^{-\frac{k\cdot t}{n}}), \tag{7}$$

根据式(6)和(7),随着 t 的增大,x 和 u 都单调递增,且当 $t\to\infty$,存在以下 4 个极限:

$$\lim_{t\to\infty}x(t)=\infty,\lim_{t\to\infty}u(t)=n,\lim_{t\to\infty}x'(t)=k,\lim_{t\to\infty}u'(t)=0. \tag{8}$$

进一步分析 x 和 u 随着 t 的变化曲线,如图 1 所示(图中 $k=50$,$n=2\,000$). 可见,随着时间 t 的推移,获得的节点数 x 越来越多,会远大于集群本身的大小 n,但不重复的节点数 u 逐渐趋向于 n. 而且在这个过程中,x 的增速与 k 有关,而 u 的增速则逐渐趋向于 0.

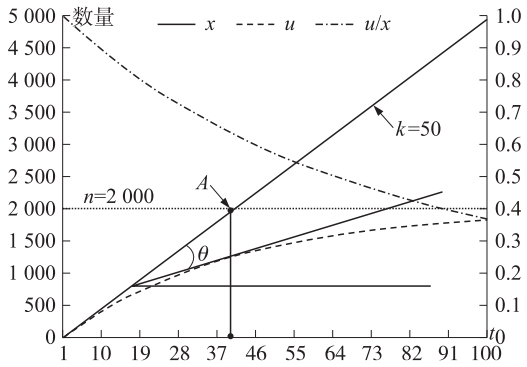

图 1　x,u 和 u/x 随 t 的变化曲线

Fig. 1　The variation curves of x, u and u/x with t

因此,必须选择合适的时机 t,根据 x 和 u 的值估算出 n 的值,在保证估计精度的同时避免大量重复的信息,导致主动测量的代价太大.

(1)条件 e

在图 1 中,曲线 x 和 n 的交点 A,此时满足:

$$u(t)/x(t)=1-e^{-1},\text{且 }u'(t)/x'(t)=e^{-1}. \tag{9}$$

本文中将 A 点定义为以 e 为参数的结束条件,即当 x 和 u 满足式(9)时,计算 n 的估计值 $\hat{n}_e$.

(2)条件 θ

在图 1 中可以看到,随着时间 t 的增大,u 曲线切线与 x 曲线的夹角 θ,从 0 度逐渐增大至 $\arctan k$. 这说明 u 与 x 的距离越来越远,u 的增速越来越小,远小于 x 的增速. 因此,本文定义以 θ 为参数的结束条件,即当 $\theta=\frac{1}{2}\arctan k$ 时,计算 n 的估计值 $\hat{n}_\theta$.

$$\because\theta=\arctan k-\arctan(k\cdot e^{-\frac{kt}{n}})\leqslant\frac{1}{2}\arctan k\Rightarrow\therefore\arctan(k\cdot e^{-\frac{kt}{n}})\geqslant\frac{1}{2}\arctan k, \tag{10}$$

因此,只要满足式(10),即满足条件 θ.

(3)条件 σ

实际上,在任意时刻 t 均可根据 x 和 u 的值,通过二分法求解隐函数(6),从而得到 n 的估计值 $\hat{n}$. 但由于这种情况下,通过不同时刻的 x 和 u 得到的 $\hat{n}$ 往往波动较大,因此本文定义以标准差 σ 为参数的结束条件,即当连续 5 次 $\hat{n}$ 的标准差 σ 小于均值的 5% 时,就以当前的这 5 次 $\hat{n}$ 的均值 $\hat{n}_\sigma$ 作为 n 的估计值.

3 实验与分析

3.1 实验平台

根据上一节给出的集群规模估算方法及估算时机,本文设计并实现了一个对 P2P 集群进行规模估算的实验平台,如图 2 所示. 平台由 40 台主机构成,分成 3 种功能节点,包括 1 台爬虫主机,1 台数据库主机,38 台测量主机. 爬虫及测量主机的机器配置为:双核 P4 2.4 GHz CPU/4 G 内存/250 G 硬盘/1 000 M 网卡/Windows XP SP2. 数据库主机为 Dell PowerEdge 1950 服务器,配置为 1 路 4 核 Xeon CPU/8 G 内存/146 G×2 SAS 硬盘/Windows 7. 所有节点通过千兆局域网相连并接入 CERNET.

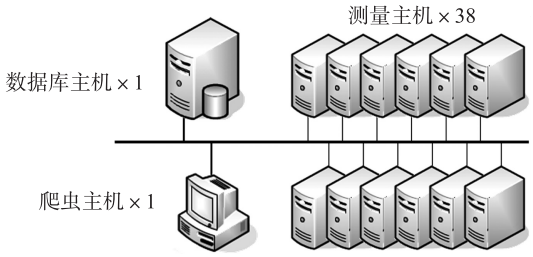

图 2 P2P 集群规模主动估算实验平台

Fig. 2 The platform for active estimation of P2P swarm size

实验首先用爬虫主机从 Torrentz. eu 上自动获取 BT 种子. 测量主机上运行修改源码的 Vuze 4.2.0.2 软件,使其主动加入 P2P 集群,通过 Tracker、DHT 和 PEX 方式获取其他节点的信息. 由于测量主机是分别独立工作的,因此它们会获取很多重复的节点信息. 测量主机并不负责去除重复信息,而只是将各自获得的全部信息写入数据库主机. 数据库主机负责存储测量节点所得到的全部节点信息,并且负责统计各测量主机获得的节点总数 x 以及其中不重复的节点数量 u. 对每个 P2P 集群,如果测量结果满足条件 e,条件 θ 和条件 σ 时,则结束测量. 然后将 n 的估计值 $\hat{n}_e$,$\hat{n}_\theta$ 和 $\hat{n}_\sigma$ 分别于 Tracker 返回值进行比较,以确认 3 个结束条件中的最佳条件.

3.2 集群测量时间

由于实际的 P2P 集群一直处于动态变化中,因此如果本文提出的集群规模估算方法及设定的估算条件,不能在较短时间内对集群规模做出估算,则测量结果将失去意义. 因此,本文首先通过实验,检验本实验平台对不同规模的集群所需的快照时间. 如图 3 所示为不同规模的集群,利用本文实验平台完成快照的实际时间. 其中 k 的取值为 200,这是根据实际测量中每次请求获得的返回值而定. 由于每个 BT 种子中,一般包含多个 Tracker,因此每台测量主机一次请求平均可获得 200 ~ 300 个节点. 本文取其最小值,这样可以得出完成快照的最长时间. 在后续的实验中,k 的取值都设定为 200.

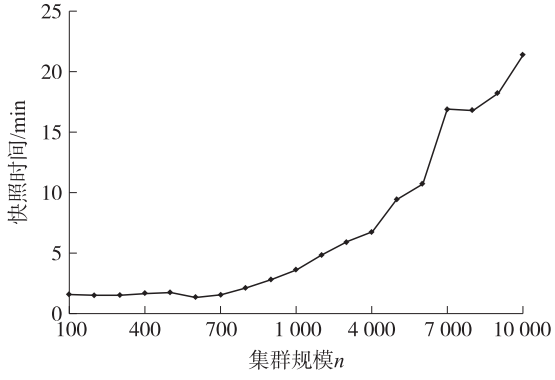

图 3 P2P 集群快照时间(k=200)

Fig. 3 The snapshot time of P2P swarms

实测结果表明,本文测量平台对规模较少的集群(n<1 000),其平均快照时间不超过 5 min. 即便对于规模较大的集群(1 000<n<10 000),其平均快照时间也未超过 20 min. 因此虽然 P2P 集群是动态变化的,但在 20 min 内的变化不会太大. 因此根据本文测量平台得到的测量结果的可信度较高.

3.3 集群规模估算

如图 4 所示为实验中根据条件 e,条件 θ 和条件 σ 分别得到的估计值 $\hat{n}_e$,$\hat{n}_\theta$ 和 $\hat{n}_\sigma$ 与 Tracker 返回的 n 的真实值的误差,其中实验的参数 k 取 200. 可以看到,当集群规模较少时(n<1 000),条件 σ 的估计值 $\hat{n}_\sigma$ 误差最小,不超过 5%. 而另外两个条件下的估计值 $\hat{n}_e$ 和 $\hat{n}_\theta$ 的误差相对较大. 但随着集群规模变大时,$\hat{n}_\sigma$

的误差迅速增大,而且波动较大. 相反的 $\hat{n}_e$ 和 $\hat{n}_\theta$ 的误差则迅速变小,而且相对比较稳定.

如图 5 所示为实验中根据条件 e,条件 θ 和条件 σ 分别得到估计值 $\hat{n}_e$,$\hat{n}_\theta$ 和 $\hat{n}_\sigma$ 时,所需的实验次数,其中实验的参数 k 取 200. 可以看到,当集群规模较少时(n<1 000),无论根据何种条件结束实验付出的实验代价都较小. 而随着集群规模变大时(1 000<n<10 000),根据条件 e 和条件 σ 计算估计值 $\hat{n}_e$ 和 $\hat{n}_\sigma$ 时,所需付出的实验代价略有上升. 而当集群规模超过 10 000 时,这两者所需的实验代价急剧增加. 而如果根据条件 θ 计算估计值 $\hat{n}_\theta$,则无论对何种规模的集群,其实验代价基本保持稳定.

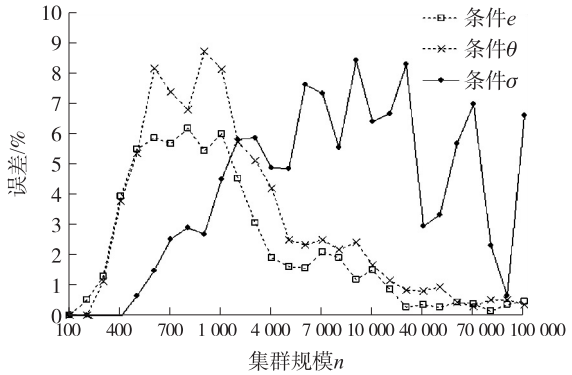

图 4　P2P 集群规模估计值误差(k=200)

Fig. 4　The error of estimation value of P2P swarm size(k=200)

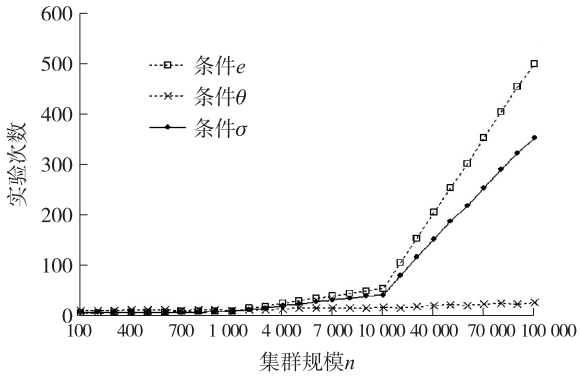

图 5　P2P 集群规模实验代价(k=200)

Fig. 5　The experiment cost of estimation of P2P swarm size(k=200)

综合图 4 和图 5,本文得出以下结论:

(1)当集群规模较小时(n<1 000),应根据条件 σ 计算 $\hat{n}_\sigma$,并以此作为集群规模的估计值,此时误差不超过 5%.

(2)当集群规模变大时(1 000<n<10 000),应根据条件 e 和条件 θ 计算 $\hat{n}_e$ 和 $\hat{n}_\theta$,并以此作为集群规模的估计值,此时的误差不超过 5%.

(3)当面对超大集群时(n>10 000),应根据条件 θ 计算 $\hat{n}_\theta$,此时的实验代价增加不多,且误差小于 1%.

4　结束语

对 P2P 系统的主动测量是对其进行系统建模和行为分析的基础. 现有测量所采用的 CCP 模型,过度依赖于 Tracker 服务器返回的先验知识,导致测量结果并不能反映集群的真实规模. 同时,该方法并不能适用于 eMule/eDonkey 等 P2P 文件共享系统的主动测量. 因此,本文针对 P2P 文件共享系统,提出了一种基于随机抽样过程的 P2P 集群规模主动估算方法,并根据理论分析设定了 3 种不同的实验停止条件. 实验结果表明,本文的估计方法误差不超过 5%. 但面对超大集群时,本文的方法还存在误差较大或实验代价太高的问题,这将是今后的研究方向.

[参考文献]

[1] Karagiannis T, Rodriguez P, Papagiannaki K. Should internet service providers fear peer-assisted content distribution? [C]// Proceedings of IMC 2005, Berkeley, 2005: 6-10.

[2] Xie J, Yang Y R, Krishnamurthy A, et al. P4P: provider portal for applications[J]. Computer Communication Review, 2008, 38(4): 351-362.

[3] Aggarwal V, Feldmann A, Scheideler C. Can ISPs and p2p users cooperate for improved performance? [J]. Computer Communication Review, 2007, 37(3): 29-40.

[4] Choffnes D R, Bustamante F E. Taming the torrent: a practical approach to reducing cross-ISP traffic in Peer-to-Peer systems[J]. Computer Communication Review, 2008, 38(4): 363-374.

(下转第 98 页)

5 结语

本文设计并开发了一个透明、相容、一致、易查的虚拟机管理系统.该系统实现了查看主机和虚拟机的简要信息或者XML格式的详细信息,能够对虚拟机进行开机、关机、暂停、恢复、重启、强制关闭、远程操控,系统还实现了日志记录的功能.测试运行表明该系统能有效地监控云计算中物理节点以及虚拟机的运行状态,方便了对虚拟机的管理.

[参考文献]

[1] 田文洪,赵勇.云计算资源调度管理[M].北京:国防工业出版社,2011.
[2] 王金波,《虚拟化与云计算》小组编.虚拟化与云计算[M].北京:电子工业出版社,2009.
[3] Armbrust M,Fox A,Griffith R,et al. A view of cloud computing[J]. Communication of the ACM,2010,53(4):50-58.
[4] 刘晓茜.云计算数据中心结构及其调度机制研究[D].合肥:中国科学技术大学计算机科学与技术学院,2011.
[5] 谭浩宇.多虚拟机管理平台中的监控系统[D].武汉:华中科技大学计算机学院,2008.
[6] Mark Stillwell, David Schanzenbach, Frederic Vivien, et al. Resource allocation algorithms for virtualized service hosting platforms[J]. Journal of Parallel and Distributed Computing,2010,70(9):962-974.
[7] Daniel Warneke,Odej Kao. Exploiting dynamic resource allocation for efficient parallel data processing in the cloud[J]. IEEE Transactions on Parallel and Distributed Systems,2011,22(6):1 045-9 219.
[8] 王忠儒.云环境下的虚拟机监控和服务部署关键技术研究[D].长沙:国防科学技术大学计算机学院,2010.

[责任编辑:丁 蓉]

(上接第80页)

[5] Blond S L,Legout A,Dabbous W. Pushing Bittorrent locality to the limit[J]. Computer Networks,2011,55(3):541-557.
[6] Bindal R,Cao P,Chan W,et al. Improving traffic locality in Bittorrent via biased neighbor selection[C]//Proceedings of ICDCS 2006,Lisboa,2006:66-66.
[7] 刘琼,徐鹏,杨海涛,等.Peer-to-Peer文件共享系统的测量研究[J].软件学报,2006,17(10):2 131-2 140.
[8] Hu C L, Lu Z X. Downloading trace study for Bittorrent P2P performance measurement and analysis[J]. Peer to Peer Networking and Applications,2012,5(4):384-397.
[9] IPOQUE. Internet Study 2008/2009[EB/OL]. http://www.ipoque.com/resources/internet-studies/internet-study-2008_2009,2009-2-18.
[10] Liu J C,Wang H Y,Xu K. Understanding peer distribution in the global internet[J]. IEEE Network,2011,24(4):40-44.
[11] Wang H,Liu J,Xu K. On the locality of bittorrent-based video file swarming[C]//Proceedings of IPTPS 2009,Boston,2009:12-12.
[12] Hobeld T,Lehrieder F,Hock D,et al. Characterization of Bittorrent swarms and their distribution in the Internet[J]. Computer Networks,2011,55(5):1 197-1 215.
[13] Zhang C,Dhungel P,Wu D,et al. Unraveling the Bittorrent ecosystem[J]. IEEE Transactions on Parallel and Distributed Systems,2011,22(7):1 164-1 177.

[责任编辑:顾晓天]