高技术通讯 2012 年 第22 卷 第11 期:1155 ~ 1160

doi:10.3772/j.issn.1002-0470.2012.11.008

对等网络中的低开销失效检测算法研究①

任 潇② 董 剑③ 左德承 杨孝宗

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 针对当前大规模 P2P 网络失效检测负载对系统可扩展性的影响,对失效检测结果的共享机制展开了研究,提出了一个基于被动订阅机制的低开销失效检测(L-FD)算法。L-FD 算法通过被检测节点来建立检测结果的订阅关系,使每个节点只需保持常数个检测关系即可获得所有邻居节点的失效状态,在节点不发生失效情况下可使检测负载下降为 $O(N)$。该算法的结果共享关系可灵活建立,不受覆盖网拓扑结构及同步时钟等因素的影响,可灵活适应不同的 P2P 系统。仿真实验及分析结果证实了这一新算法的可行性和有效性。

关键词 P2P 网络,失效检测,结果共享,检测负载

0 引 言

对等(peer-to-peer, P2P)网络已成为目前 Internet 上实现资源共享的主要平台之一[1,2]。P2P 网络是 Internet 节点自组织成的覆盖网络,它打破了传统客户/服务器架构,使其应用能够充分利用网络带宽以及每个网络节点的潜力,从而极大提高了网络和资源的利用率。但其高效运行要有快速资源定位和资源高可用性这两个重要前提[3],而这两点在很大程度上依赖覆盖网络中每个节点保持的连接数,即节点路由表的容量。如果覆盖网中的节点都能维持较大的出度,根据图论知识,一方面可以有效降低资源搜索的跳数,提高搜索速度,另一方面可以提供更多冗余通路,在节点或链路发生失效时有效提高系统容错能力,保障系统资源的可用性[4]。例如,在 Chord、Kademlia、SkipNet 等覆盖网络中,每个节点的路由表至少保持 $O(\log N)$ 项[5],即使在节点的失效概率为 50% 的情况下,仍可以保障 $O(\log N)$ 跳的定位效率(接近覆盖网最优定位效率)。因此,如何高效地维护每个节点大量的连接($O(\log N)$),已成为 P2P 系统要解决的一个重要基础问题。

失效检测技术是目前解决P2P系统这一问题的主要手段[6],该技术通过周期性地发送心跳检测和应答消息来探测路由表中节点的状态,从而及时发现失效项并进行修复。然而,这些心跳检测消息所产生的失效检测负载在最初的覆盖网研究中却常被忽略。随着 P2P 应用规模的不断扩大,这种简单的失效检测机制所带来的检测消息负载已对系统的带宽消耗产生了重要影响,极大限制了系统的可扩展性。因此,低开销失效检测机制的研究对提高 P2P 系统性能有重要意义。基于这种考虑,本文提出了一种基于被动订阅机制的失效检测算法——低开销失效检测(low overhead failure detection, L-FD)算法。L-FD 算法可在节点不发生失效的情况下有效降低失效检测负载。同时,该算法通过被检测者建立检测结果的订阅关系,结果的共享关系可灵活建立,因而可灵活适应不同的 P2P 系统。

1 相关知识

失效检测最早由 Chandra 和 Toueg[7] 作为一种增强异步系统计算模型的有效途径提出,并给出了形式化定义,目前已广泛应用到网格计算、集群管理、通信协议等多个相关领域[8,9],在本文中,它是研究的重点。

① 国家自然科学基金(61100029)和高效能服务器和存储技术国家重点实验室开放课题基金(2009HSSA07)资助项目。

② 女,1983 年生,博士生;研究方向:容错计算技术;E-mail:renxiao@hit.edu.cn

③ 通讯作者,E-mail:dan@hit.edu.cn

(收稿日期:2011-11-15)

研究证明,P2P 系统具有高扰动性[10](大量节点的加入或者离开),且大部分节点停留时间较短,这对系统的运行性能和资源的可用性产生了极大的影响。针对这一问题,作为系统基础组件的失效检测机制,可通过周期性探测系统中节点的状态,为节点失效后路由表的及时恢复及更新提供支持,以维护系统路由的一致性和可靠性,解决 P2P 网络中相对严格的拓扑要求对系统容错性能的影响。例如,在 Gnutella、Edonkey、Chord、Tapestry 等系统中,可通过心跳检测机制(Keep-Alive 算法)来完成系统成员管理。但要注意到,失效检测在 P2P 系统中发挥重要作用的同时,其产生的负载随着系统规模的不断扩大而增大,这已成为系统带宽消耗的主要来源。一方面高扰动性带来的节点突然离开使得心跳检测的频率非常频繁,另一方面,为保持系统的运行效率及可用性,每个节点需要维持较大的连接数($O(\log N)$),这使失效检测负载将达到 $O(N\log N)$,这极大地影响了整个 P2P 系统的可扩展性。由此可见,研究低开销的失效检测机制对提高 P2P 系统性能有着重要意义。

层次式失效检测[11]和 Gossip[12]式失效检测是降低失效检测开销的两种传统方法。层次式方法通过对节点进行分组,将网络拓扑组织成某种层次式结构(如树、森林等),以达到降低系统检测消息的复杂度,提高可扩展性的目的。在一些无结构 P2P 系统中,如 Gnutella0.6、emule 等,利用超级节点(emule 中称为服务器)作为 Leader[13],形成了一个两层的失效检测机制,有效降低了检测消息数量。但是,这种方式必须依赖覆盖网络原有的层次化拓扑结构。在其它系统中,如 Chord、Pastry 等,覆盖网拓扑本身不具备层次化结构,若是采用层次化方法,在系统拓扑变化较快的 P2P 系统中,建立层次式检测架构、选举 Leader 节点等开销将十分巨大。Gossip 方法是一种适合开放式分布式系统使用的检测方法,其开销接近常数,但是受限于算法本身随机性的特点,较大的检测负载将影响路由更新的速度,使路由表中产生大量的坏项。P2P 系统 Kademlia[14]采用的"捎带确认"方式充分利用上层应用之间传递的消息来完成失效检测,几乎不需要发送额外的检测消息,但是,这种方式依赖上层应用的类型及 Kademlia 特殊的对称结构,并不适用其它系统。P2P 系统失效检测主要针对邻居节点(路由表项)的特点,如 Chord 针对后继表,Tapestry 针对路由表中的邻接点,Pastry 针对叶节点,针对这种情况,Zhuang 等[15]提出了节点之间共享检测结果的思路,并对三种共享方式进行了讨论,但是由于其研究目的只是为了降低检测延迟,并未考虑结果共享对降低负载的影响,但这一思想为低开销失效检测机制的研究提供了一种新的解决方法。Castro 等[16]基于 Pastry 系统提出了一个兼顾性能和高可用性的 P2P 系统 MSPastry,其所采用的失效检测机制的负载与叶集的大小($O(\log N)$)无关,每个节点在每一个检测周期中只会对直接左邻居发起检测,只有在发现失效后,才会通知叶集中其它节点。通过叶集节点之间对检测结果的共享,MSPastry 的失效检测负载低于 Pastry 系统的 50%,而检测延迟不超过 Pastry 的两倍,但是,这种检测机制依赖 Pastry 的特殊叶集结构,无法应用到其它系统,适用性较差。针对这种情况,Dedinski 等[4]提出了合作检测(cooperative keep-alive, CKA)算法——一个可应用各种网络(无结构或结构化网络)的低开销检测算法。*CKA* 算法通过一个控制算法控制检测者发出 ping 包的时间,使其可以保证在 K(常数)时间内只收到一个检测消息。为了进一步降低负载,检测者每次只选择一半邻居进行检测,发现节点 X 失效后,通过一个简单的 flooding 过程来通知 X 的邻居。在正常情况下,CKA 法只需 $O(N)$ 的维护成本,且检测延迟以较高的概率保持在 K 以下。但是,CKA 算法的关键在于被检测者可以反向控制检测者发起检测的时刻,这不仅需要通讯链路双向的支持,而且需要同步时钟的支持,而 P2P 这类大规模分布式系统的网络条件复杂,节点异构性强,显然其具有较强的异步性,在这种环境下实现同步时钟的开销是巨大的。

基于上述分析,我们提出了一个针对邻居节点检测的低开销检测算法——L-FD 算法,该算法通过由被检测节点管理的检测结果被动订阅机制,可根据当前检测需求灵活建立结果共享关系,不需同步时钟支持,不依赖任何覆盖网结构,可灵活应用于不同的 P2P 系统。L-FD 算法使每个节点只需保持常数个检测关系,在节点不发生失效的情况下只需要 $O(\log N)$ 的检测负载,可有效降低失效检测负载,对大规模 P2P 系统的可扩展性提供有力的支持。

2 低开销失效检测 *L-FD* 算法

2.1 系统假设

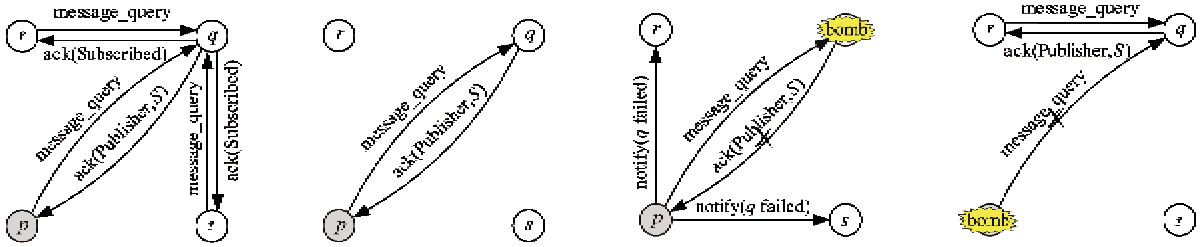
我们假设一个 P2P 网络包含 n 个节点,任意一个节点 X 通过自身的路由表可以访问到 d 个其它

的节点,称为邻居节点,记为 $N(A)$。同大多数 P2P 失效检测算法一样,L-FD 算法将节点未经通知离开系统的行为视为失效。系统失效模型采用 fail-stop 模型,在大多数 P2P 系统中,节点一旦离开系统,再回来时将作为新节点加入。鉴于本文研究重点在于如何通过改变检测架构来降低负载,为了便于描述,我们将节点间的链路视为完美链路,不存在消息的丢失及传输的延迟。对于实际系统中因消息丢失及传输延迟对检测结果所产生的影响,可通过重传(resent)等方法予以解决,这并不是本文的研究重点。

2.2　L-FD 算法

L-FD 算法的基本失效检测策略采用了 PULL 模式,即检测节点 p 首先向被检测节点 q 发出状态查询消息 message _ query(mq),节点 q 在收到后以应答消息 ack 回复 p,以表明自己处于工作状态。若 p 在规定时间内没有收到 ack 消息,则判定节点 q 失效。这种检测方式由检测者主动发起检测,且不需要时钟同步的假设,非常适合 P2P 这类拓扑经常发生变化的复杂分布式系统。

基于 PULL 模式,我们提出了可有效降低检测负载的 L-FD 算法,其基本思想如图 1 所示。系统中每个节点 X 只对自己路由表中所记录的邻居节点集 $N(X)$ 发起检测。作为一个新加入系统的新节点 q 来讲,在收到查询消息 mq 后,其应答消息 ack 有两种类型。如图 1(a)所示,节点会对最初收到的查询消息返回 ack(Publisher, S),将这个查询消息的发出者分配为检测结果的发布者(图 1 中的节点 p)。被检测者将此后对其发起检测的节点加入到订阅者集合 S 中,并通过应答消息 ack(Subscribed)通知它们处于订阅状态(如图 1 中的节点 r 和 s)。在系统无失效发生时,处于订阅态的节点将不再对节点 q 发起任何检测,只有发布者会继续对 q 以周期 Δ 发送查询消息 mq,而被检测节点 q 会将不断更新的订阅者集合附在 ack(Publisher, S)消息中发送给 p,以对 p 所保存的 S 集合进行更新,为降低算法开销,采用增量备份方式,见图 1(b)。



(a) 订阅过程　(b) 运行态　(c) 发现失效　(c) 发布者失效

图 1　L-FD 检测算法的基本思想

在图 1(c)中可以看到,当节点 p 在 Δ 时间内未收到 q 返回的应答消息,则判定 q 已经发生了失效,修改自己的检测结果输出的同时,根据订阅者集合 S 发起广播,通知节点 q 的所有检测者,以实现检测结果的共享。图 1(d)指出了发布者节点发生失效的情况。采用 L-FD 算法时,发布者与被检测节点之间采用一种双向检测机制,节点 q 在发出应答消息 ack 之后会启动一个定时器,由于已知发布者 p 的检测消息发送间隔为 Δ,为此,当 Δ 时间内未收到新的检测消息 mq,则 q 将会发现发布者失效(因为我们假设链路不存在传输延迟,q 因此应该在 Δ 时间内收到下一个 mq,而在实际系统中,q 需要评估双向的传输延迟 δ,定时器的值应大于 $\Delta+\delta$。但是,这个值的设定并不需要 p 的时钟,这与 CKA 算法不同)。此时,节点 q 将从集合 S 中选择一个在线时间较长的节点 r(根据一些研究结果,在线时间越长的节点继续留在系统中的概率越大),通过发送消息 ack(Publisher, S)将其分配为新的发布者。

但是,从上面的分析可以发现,当发布者与被检测节点同时失效时,会导致订阅者节点无法获取这一失效信息,发布者节点是算法的一个可靠性瓶颈。为此,L-FD 算法设置了一个发布者集合 M($|M|=c, c\geqslant 1$),以提高发布者的可靠性,保障发布者节点集合比被检测者具有更长的在线时间。假设节点的失效率为 p,则发布者集合的平均在线时间将提高 p^{1-c} 倍。考虑到 c 的取值会对检测负载产生一定的影响,在大多数情况下,c 一般取 2 即可。例如,假设节点平均在线时间达到 10h,当 $c=2$ 时,发布者集合 M 的平均在线时间将达到 100h,已远高于被检测节点,只有在网络条件非常恶劣的环境下,才需要适当调高 c 的取值。同时,由于 c 为常数,由此产生的负载对系统可扩展性不会产生影响。从后面所列

出的实验结果可以看出，当 c 取值为 2 时，增加的检测负载非常小。

基于上述思想，L-FD 算法描述如图 2 所示。为了使算法描述更加直观，我们从检测节点 p 和被检测节点 q 的角度进行描述。而在实际的 P2P 系统中，每个节点在检测其它节点的同时，也会成为其它节点的检测对象，因此，图中算法的两个模块同时运行在每一个系统节点上。低开销失效检测中，订阅者集合 S 为 FIFO 队列，当 M 中节点发生失效时，直接以 S 的首节点作为备选发布者。图中的 $F(p)$ 用来表示节点 p 上失效检测模块的输出结果，对于 $F(p)$ 中的节点，路由表维护算法会启动路由恢复过程，找到新的节点替换失效项。从 L-FD 算法中可以看出，当系统正常运行时，每个节点至多有 c 个节点对其发起检测，因此每一个检测周期内检测消息的数量为 $O(c \cdot N)$。当发生失效时，检测负载将上升为 $O(N\log N)$，但是广播过程可以采用一些优化算法进一步降低负载。我们将通过实验对 L-FD 算法的性能进行分析。

```
Algorithm L-FD:
1    node q:  //被检测节点
2    upon receive mq_p form p do
3      if  p ∈ M  then
4        send ack(Publisher,S) to p;
5        t_p=t_current;
6      else if  p ∉ S  then
7            if  |M| < c  then
8              send ack(Publisher,S) to p; M.add(p);
9            else send ack(Subscribed) to p; S.add(p);
10   for any node  p ∈ M :
11     if not receive mq_p from p until t_p+Δ  then
12       M.remove(p);
13       send ack(Publisher,S) to r; //r is the first node of S
14       M.add(r);
15
16   node p: //检测节点
17     D(p)=N(p); F(p)=∅;
18     for any  q ∈ D(P), at time  i·Δ, send mq_p to q;
19       upon receive ack(Publisher,S_q) from q do
20         if  q ∉ D(p)  then D(p).add(q);
21       upon receive ack(Subscribed) from q do
22         D(p).remove(q);
23       if not receive ack from  q ∈ D(p)  until  (i+1)·Δ  do
24         D(p).remove(q);  F(p).add q;
25         notify (q failed) to every node of S_q;
```

图 2 L-FD 算法

3 实验验证及分析

为了对 L-FD 算法的性能进行验证，基于商用网络仿真工具包 OPNET 设计了 L-FD 算法的实验方案。在不同的试验中，利用 OPNET 按照规模 N、连接度 $\log N$ 生成一个随机拓扑。假设节点存在一定的故障率，实验过程中一些节点会按照一定概率离开系统，而节点之间的链接不会发生失效，传输延迟则参考在 Internet 上所得到的数据为参数进行设置，以使实验结果更加接近实际系统。在这个实验环境下，我们对 L-FD 算法的可扩展性和检测延迟进行了验证和对比分析。对比实验参照对象选择了大多数当前 P2P 应用中常用的标准检测（standard keep-alive，SKA）算法，而对于 MSPastry 的检测算法及 CKA 算法，二者与 L-FD 算法都是对 SKA 算法在检测架构上的改进，在可扩展性上有三者近似的结果，而与 L-FD 算法的主要区别是在拓扑要求和系统假设上，因此未将二者作为比较对象。

（1）L-FD 算法的可扩展性

L-FD 算法通过被动订阅机制，可以有效降低检测消息数量，提高系统的可扩展性。在实验中，我们在多种不同规模的网络环境中将 L-FD 算法与 SKA 算法进行了对比。在图 3 显示了在一个执行片段中，这两种算法的检测负载的对比。可以看出，在同样的网络规模（$N=1000$），传统的 SKA 算法要比 L-FD 算法高出 50% 的负载。

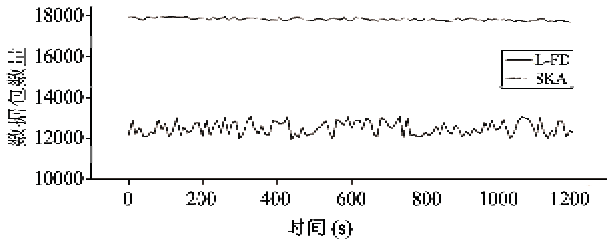
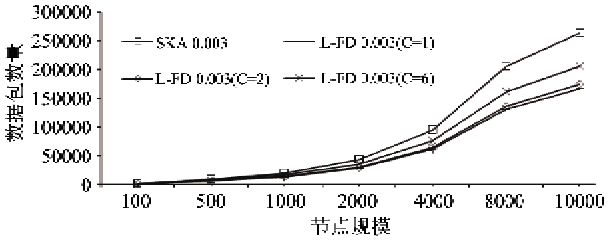


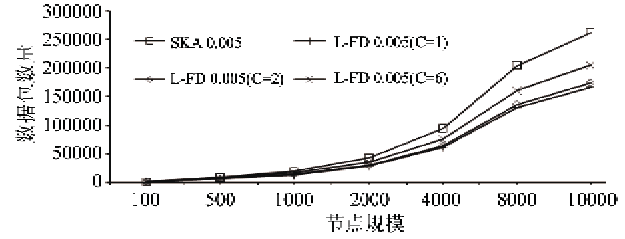
图 3 L-FD 算法与 SKA 算法的检测负载对比

根据前面的分析，在节点不发生故障时，L-FD 算法可以将检测负载由 $O(N\log N)$ 降为 $O(N)$，可见，随着网络规模的增大，L-FD 算法对负载的降低更加明显，这种趋势在图 4 中可以得到验证。考虑到节点故障对 L-FD 算法的检测负载存在一定影响，我们在不同的节点故障率下（3‰和 5‰）对算法进行了实验验证。可以看出，高故障率会增加 L-FD 算法一定的负载，但是，与 SKA 算法相比，L-FD 算

法所产生的检测负载与节点规模 N 更接近线性关系。在图 4 的实验中还对 c 值对检测负载的影响进行了讨论,可以看出,在两种节点失效率下,当 $c=2$ 时,检测负载的变化都非常小,而随着 c 值逐渐增大,所产生的额外检测负载也越来越明显(如图中 $c=6$)。结合前面讨论的关于 c 值对发布者集合可靠性的影响,c 的取值不宜过高,2 或 3 是较好的选择。



(a) 节点失效率为 3‰



(b) 节点失效率为 5‰

图 4　不同网络规模下检测负载的对比

(2) L-FD 算法的失效检测延迟

检测延迟是失效检测算法的一个重要评价指标,过高的检测延迟引起的正常数据包丢失(发向已经失效的节点)对高层应用程序的性能如任务完成时间、网络吞吐率、流媒体帧丢失率等,有着极大影响。而 L-FD 算法的检测延迟需要考虑订阅者节点与被订阅者节点之间的传输延迟,这会对算法的检测延迟产生一定影响,因此,在图 5 中可以看出,在算法的一个执行片段(1000s)中,L-FD 算法的检测延迟在大多数情况下要高于 SKA 算法。但是,按照文献[15]的分析,在结果共享机制下,由于各节点检测周期的不同步,通过共享获得的失效通知可能会早于本地发起的检测,这也就是在图 5 的一些检测周期中,L-FD 算法的检测延迟反而会优于 SKA 算法。

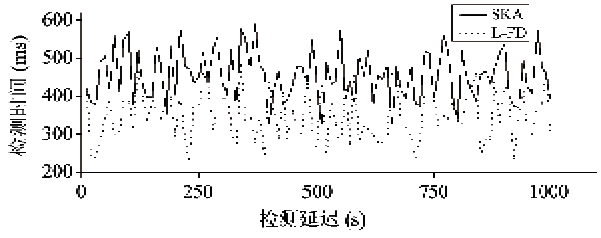


图 5　L-FD 算法与 SKA 算法检测延迟对比

通过长时间实验,对 L-FD 算法的检测延迟的概率分布进行了分析,结果如图 6 所示。实验选择了两种检测周期($\Delta=300$ms 和 $\Delta=500$ms),图中纵坐标为失效检测时间 Td 小于某个值 X 的概率。通过与 SKA 算法对比发现,L-FD 算法对检测延迟的影响并不大,而且从图中可以看出,L-FD 算法的检测延迟未超过 SKA 算法检测延迟的上限 2Δ。可见,L-FD 算法引入的结果共享机制对系统失效检测能力的影响可以为大多数应用所接受。

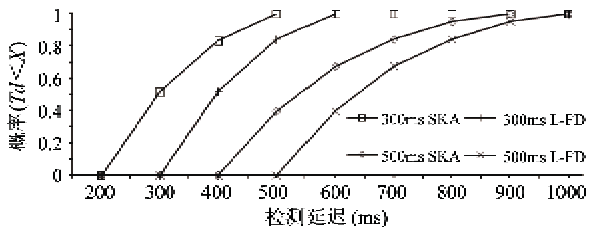


图 6　检测延迟的概率分布

4　结　论

针对 P2P 系统在规模不断增大时对失效检测算法可扩展性的要求,本文提出了一个可有效降低检测负载的失效检测算法——L-FD 算法。面对大量邻居节点的检测需求,L-FD 算法通过由被检测节点管理的检测结果被动订阅机制,根据当前检测需求可灵活建立结果共享关系,不需同步时钟支持,不依赖任何覆盖网结构,可灵活应用于不同的 P2P 系统,在节点不发生失效的情况下可将检测负载由 $O(M\log N)$ 降为 $O(N)$,有效地降低失效检测负载,对大规模 P2P 系统的可扩展性提供了有力的支持。同时,在仿真实验中发现,与当前 P2P 应用中常用的 SKA 算法相比,L-FD 算法在有效降低检测负载的同时,其结果传输时间对检测延迟指标的影响并不大,在大多数 P2P 应用的可接受的范围内。

参考文献

[1] Eng K L, Crowcroft J, Pias M, et al. A survey and comparison of peer-to-peer overlay network schemes. *Communications Surveys & Tutorials, IEEE*, 2005, 7(2): 72-93

[2] Kurian J, Sarac K. A survey on the design, applications, and enhancements of application-layer overlay networks.

ACM Computing Surveys (*CSUR*), 2010, 43(1): 1-34

[3] 陈贵海,李振华. 对等网络. 北京: 清华大学出版社, 2007. 11-19

[4] Dedinski I, Hofmann A, Sick B. Cooperative keep-alives: an efficient outage detection algorithm for P2P overlay networks. In: Proceedings of the 7th IEEE International Conference on Peer-to-Peer Computing, Galway, Ireland, 2007. 140-150

[5] Rao W, Chen L, Fu A W C, et al. Optimal resource placement in structured peer-to-peer networks. *IEEE Transactions on Parallel and Distributed Systems*, 2010, 21(7): 1011-1026

[6] Price R, Tino P. Still alive: Extending keep-alive intervals in P2P overlay networks. In: Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing. Washington, D C, USA, 2009. 1-10

[7] Chandra T D, Toueg S. Unreliable failure detectors for reliable distributed systems. *Journal of the ACM* (*JACM*), 1996, 43(2): 225-267

[8] 皿东,陈蜀宇,陈峰. 一种网格环境下的动态故障检测算法. 计算机研究与发展, 2006,(11): 1870-1875

[9] Lavinia A, Dobre C, Pop F, et al. A Failure detection system for large scale distributed systems. In: Proceedings of 2010 International Conference on Complex, Intelligent and Software Intensive Systems, Krakow, Poland, 2010. 482-489

[10] Ohzahata S, Kawashima K. An experimental study of peer behavior in a pure P2P network. *Journal of Systems and Software*, 2011, 84(1): 21-28

[11] Stelling P, Foster I, Kesselman C, et al. A fault detection service for wide area distributed computations. In: Proceedings of the 7th International Symposium on High Performance Distributed Computing, Chicago, USA, 1998. 268-278

[12] Van Renesse R, Minsky Y, Hayden M. A gossip-style failure detection service. In: Proceedings of IFIP International Conference on Distributed Systems Platforms and Open Distributed Processing Middleware, The Lake District, UK, 2009. 55-70

[13] Xie C, Chen G, Vandenberg A, et al. Analysis of hybrid P2P overlay network topology. *Computer Communications*, 2008, 31(2): 190-200

[14] Maymounkov P, Mazières D. Kademlia: A peer-to-peer information system based on the XOR metric. In: Proceedings of the International Workshop on Peer-to-Peer Systems, Cambridge, USA, 2002. 53-65

[15] Zhuang S Q, Geels D, Stoica I, et al. On failure detection algorithms in overlay networks. In: Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies, Miami, USA, 2005. 2112-2123

[16] Castro M, Costa M, Rowstron A. Performance and dependability of structured peer-to-peer overlays. In: Proceeding of the International Conference on Dependable Systems and Networks, Florence, Italy, 2004. 9-18

A low overhead failure detection algorithm for peer-to-peer networks

Ren Xiao, Dong Jian, Zuo Decheng, Yang Xiaozong

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract

A study of the failure-detection-result sharing in peer-to-peer (P2P) networks was performed to reduce the impact of significant detection overheads on the scalability of large scale P2P systems, and on this basis, a low overhead failure detection (L-FD) algorithm based on the passive subscribing mechanism was proposed. The L-FD algorithm can establish the relations of detection results sharing by the monitored nodes. Each node in the system only needs detecting invariable nodes to achieve the status of all neighbors. The L-FD algorithm can reduce the detection overhead complexity to $O(N)$ without failure, in addition it can not be limited by the factors of overlay topology and synchronization when establishing detection-result sharing relations, thus, it can be rapidly and flexibly applied to different P2P systems. The experimental results and the corresponding analysis show that the new L-FD algorithm is feasible and effective.

Key words: peer-to-peer networks, failure detection, results sharing, detection overhead