分布式系统中的层次式失效检测算法改进

徐光侠[1,3] 陈蜀宇[2]
(重庆大学计算机学院 重庆 400030)[1] (重庆大学软件学院 重庆 400030)[2]
(重庆邮电大学软件学院 重庆 400065)[3]

摘 要 针对在分布式系统中的层次式失效检测方法的检测准确性和检测效率的问题,在层次式失效检测机制的对象级、进程级和主机级的层次思想指导下,基于Chen预测算法提出了一种分布式系统中的层次式失效检测的改进算法。考虑到传统的分布式系统中层次式失效检测方法的单点失效问题、检测时延等因素,在分层时把局域网的检测消息限制在组内,并且使组内的节点承担不同组间的检测。改进算法实现时增设一个信任度变量和修正比例因子,采用向网络中加负载的方式模拟大规模网络的复杂情况以增加网络延迟,完成该算法的实验验证。实验结果表明,改进算法能够提高失效检测的准确性和检测效率,降低误判率,该研究成果也为失效检测方法的进一步优化提供了研究依据。

关键词 分布式系统,层次式,失效检测,检测算法
中图法分类号 TP393 **文献标识码** A

Improvement of Hierarchical Failure Detection Algorithm in Distributed Systems

XU Guang-xia[1,3] CHEN Shu-yu[2]
(College of Computer Science, Chongqing University, Chongqing 400030, China)[1]
(College of Software, Chongqing University, Chongqing 400030, China)[2]
(School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)[3]

Abstract Aiming at the problem of the accuracy and efficiency of hierarchical failure detection method in distributed systems, this paper proposed an improved algorithm of hierarchical failure detection in distributed systems based on Chen prediction algorithm, which is on guidance of hierarchical failure detection mechanism of the object level, process level and host level. In distributed systems, the traditional hierarchical failure detection method always encounters problems such as single point failure, detection delay and so on. They proposed that detection messages in local area network are limited within the group when layering, and different nodes of one group assume different detection among groups. A trust variable and a correction scale factor are added in improved algorithm. In order to increase network delay, increasing the load of network is adopted for simulating the complexity of large-scale networks. Then, the experimental verification of the algorithm is completed. The experimental results demonstrate that the accuracy and efficiency of failure detection are improved and misdiagnosis rate is reduced by adopting the improved algorithm. They offer research base for further optimization of failure detection methods.

Keywords Distributed systems, Hierarchical, Failure detection, Detection algorithm

分布式系统规模越大,失效检测时需要检测的实体对象则越多,心跳检测信息会让系统造成更大的网络负载,针对这个问题,有研究者提出层次式的失效检测构架[1,2]。在单层次失效检测方法的检测系统中所有节点是都对等的,系统中每个节点面向系统中所有其它节点。多层次的失效检测方法对网络负载的影响方面加以改进,解决失效检测消息对网络产生的巨大负载问题[3]。

1 分布式系统的失效检测

相对单层次失效检测而言,多层次的失效检测把失效检测系统划分为多个层次,各节点上的失效检测器[4,5]不再是对等的,各失效检测器所完成的任务和所起作用有差异。在多层次的失效检测方法中,节点分为普通节点和管理节点,每个层次内部的普通节点只能看到本组中的所有节点,而看不到其它组的成员信息,组内的普通节点发送心跳消息时,也只需要向本组中的节点发送,组内节点之间的失效检测采用推模型;多层次系统中的特殊节点称为管理节点,能看到整个系统中的所有节点的状态信息,组间节点的失效检测通知通过管理节点实现。

针对分布式系统的失效检测器(Failure Detector,以下简

到稿日期:2011-01-19 返修日期:2011-03-16 本文受国家自然科学基金项目(60973160),重庆市自然科学基金项目(CSTC,2008BB2307),重庆市教委科学技术研究项目(KJ100506)资助。
徐光侠(1974—),女,博士生,副教授,主要研究方向为可信计算、分布式计算。

称 FD)设计,其必须保证在完全活动状态下,使分布式系统中的任一成员都能得到失效检测,同时,保证分布式系统中活动成员在失效的情况下,其他成员能够在有效时间内检测到这个成员的失效状态[6]。FD 应避免集中式检测,即避免把检测任务集中在单一节点上而造成的性能瓶颈,还应避免由于某一个检测者失效引发整个失效检测器 FD 的可用性严重降低。

每个 FD 兼有检测者和被检测者两种角色,可以向其它的 FD 发送消息,同时也可接收其它检测器发出的消息。认定分布式系统中节点正常的条件是:只要系统中一个失效检测器认为该节点上的失效检测器正常。反之,则认为节点失效。

2 层次式失效检测的思想

如图 1 所示,层次式失效检测机制将系统实体分为对象级、进程级和主机级共 3 层。失效检测包括对象级失效检测器(Object Failure Detector, OFD),进程级失效检测器(Process Failure Detector, PFD)、主机级失效检测器(Host Failure Detector, HFD)和失效通告器(Failure Annunciator, FA)。每台主机上部署一个或多个对象级失效检测器,使应用对象的失效检测在本机完成,不受网络影响。假如主机上所有对象都失效,包括 OFD 本身,则表示该主机失效,该情形无法在主机本地检测,需要采用部署在其它主机上的 HFD 进行检测。

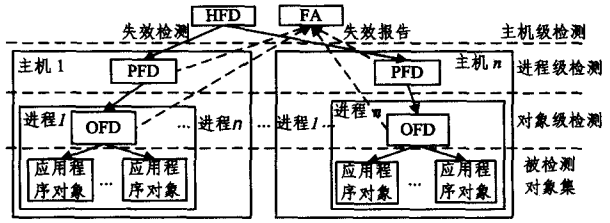

图 1 层次式失效检测结构图

在每台主机上部署一个或多个 OFD,利用每台主机上检测对象失效的 OFD 来代表关键进程,如果主机上运行的 OFD 失效,则此进程被判定为失效。OFD 失效该主机上的应用对象未被监控,此时发生的对象失效状态将不可能被检测到,也不可能会有相应的恢复动作,失效的应用对象无法连续提供服务,达不到容错的目的,这是高可靠和高可用性所不能容忍的行为,所以用 OFD 对象来代表其所在的主机的关键进程,PFD 可以根据对 OFD 的检测来判断进程是否失效。

在每台主机上部署一个 PFD,负责监控进程失效,也就是监控主机中的关键进程,使用 OFD 对象作为关键进程,主机失效就可以用 PFD 失效来代表。

3 算法改进

层次式失效检测算法是一种将分布式系统中每个节点都同等对待的失效检测方法,它对网络负载产生的影响是比较理想的,而且由于组间检测采用的是通知方法,大大降低了组间检测所需要的检测时间。但由于此种方法的组内检测采用直接检测方式,而组间检测的任务由每个组的管理节点负责,管理节点的失效也就意味着其它组与本组的检测中断,因此,单点失效问题会对该失效检测算法产生重大的影响。

3.1 算法思想

课题组所提出的改进算法将在保证失效检测评价标准的及时性与准确性的前提下,针对层次式失效检测方法中的单点失效问题加以改进。改进后的失效检测方法的基本思想是:组内检测负责检测任一组内的节点的状态,组间检测负责不同组的节点之间的检测。文中主要选取两个规则作为改进后的失效检测策略:

规则 1 层次式失效检测方法对网络负载考虑的是整个网络的负载,如果分层时考虑网络拓扑特征,把每个局域网内的节点分为一组,把大量的检测消息限制在局域网之内,就可以有效地缓解整个广域网的网络负载。

从整体看,将每个局域网看成是一个"组",组内完成消息的检测,如图 2 所示。每个 HFD 负责检测一定数量的主机进程,HFD 与 PFD 之间形成父子关系,即子 PFD 的检测进程集 Ps 是父 HFD 的检测对象集 Hs 的子集,但是父 HFD 是通过子 PFD 传递来的检测结果来获知 Ps 中服务的状态[7]。优点在于:一个父 HFD 只检测有限范围内(局域网内)的服务,因此可减少检测消息传递的开销,提高检测体系自身的可靠度。改进后的失效检测分层结构易于扩展,能避免集中式结构的检测中心点的瓶颈问题,又能改善分布式分层检测结构的协作效率问题。

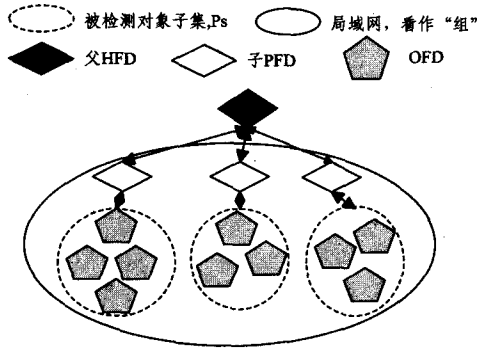

图 2 改进的层次式组内失效检测示意图

规则 2 为了避免单点失效,采用改进的方法使每个组内的各个成员地位不同,即每个组内的节点都承担不同组之间检测的任务,从而使失效检测系统正常提供服务不受单个节点的失效影响。

如图 3 所示,组 1 中节点 p 需要组 2 中节点 q 的状态,首先是组 1 节点 p 向组 2 节点 q 发送检测信息,这时通过组 2 中的某一节点(q 节点或其它节点,如 $k1$ 节点等)通知组 1 中某一节点的方式使 p 知道 q 的状态。如果在规定的时间内,组 1 一直未接收到组 2 中的任何检测消息的响应消息,那么组 1 有理由怀疑组 2 中的所有节点已经失效,将失效报告信息发送至 FA,报告组 2 节点全部失效。如果组 2 未失效,那么此种方法应该可以保证在规定的时间内,会有检测消息的响应消息从组 2 发送到组 1,从而达到组间检测的目的。

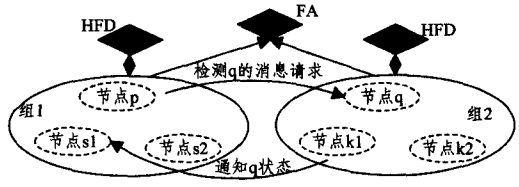

图 3 改进的层次式组间失效检测示意图

3.2 算法描述

基本的失效检测算法中,Chen 预测算法[8]对历史窗口取平均然后加上一个修正值,得到下一个心跳要到达的时间。

其数学描述如下：用 $m_1, m_2, \cdots, m_n$ 表示进程 P 接收到的 n 个最近的心跳消息，用 $A_1, A_2, \cdots, A_n$ 表示这 n 个消息的到达时间，EA 表示理论上的预测到达时间，α 为修正值，α 是为了减少因为网络延迟或者进程负载过重所造成的误报而进行的修正，下一个心跳的预期到达时间延迟 T 由 EA 和修正值 α 构成。则当至少接收到 n 个消息后：

$$EA_{(k+1)}=\frac{1}{n}\sum_{i=k-n}^{k}(A_i-\Delta_i\cdot i)+(k+1)\cdot\Delta_i \quad (1)$$

式中，$EA_{(k+1)}$ 表示第 $k+1$ 个心跳的理论预测到达时间。

$$T_{(k+1)}=\alpha_{(k+1)}+EA_{(k+1)} \quad (2)$$

式中，$T_{(k+1)}$ 表示第 $k+1$ 个心跳的预测到达时间。

Chen 预测算法是第一个自适应失效检测算法，在一定程度上提高了检测的准确性，但因为它的修正值是一个常数，所以会产生导致检测时间过长的问题。同时，针对层次式失效检测方法中的单点失效的问题，本文做了相应的算法改进。主要从两个方面着手，一是添加信任度参数 T_o，二是修正比例因子 β[9]。

3.2.1 增设一个信任度变量 T_o

在 Chen 的预测算法的基础上，为被检测的对象增加一个信任度变量 T_o，每检测出一次失效，信任度变量就减 1，每隔一段时间，所有对象的信任度值都会恢复到系统初始值。在判断对象是否失效时，如果通过失效检测时间难以判断，也可以通过信任度变量 T_o 的值进行判断。如果 T_o 低于阈值，则判定为被检测的对象失效，否则认定没有失效，这样在对象多的情况下，可大大提高失效检测的准确率。

3.2.2 修正比例因子 β

限定比例因子 β 的最小值 β_{min}，当 $\beta<\beta_{min}$ 时，用 $\beta_{min}+EA$ 来确定心跳新鲜点，限定比例因子 β 的最小值，可以在增加少许平均检测时间的基础上，降低误判率。

```
1. 对于主体计算机的对象 P:
2. 每 Δi 秒向 K 发送心跳消息
3. 客体计算机 K:
4. f=-1 //最新点
5. S=nil //S 初始化为空表
6. μ //S 的最大容量(如:1000)
7. β //修正的比例因子
8. To=100 //信任度的初始值
9. 当在 t 时刻收到心跳消息 mj
10. if f=-1
11.     then f=t
12. else
13.     tΔ=t-f
14.     f=t
15.     添加 tΔ 到 S
16.     if size of S>μ
17.         then remove head of S
18.     end if
19. end if
20. 请求得到 t 时 P 的失效概率
21. if β<βmin
22.     then βmin=βmin+EA
23. end if
24. tΔ=t-f
25. |S(tΔ,β)|=S 中 x 小于或等于(tΔ,β)的元素个数
26. |S|=S 中的元素个数
27. Return |S(tΔ,β)|/|S|
```

该算法返回 S 中 x 小于或等于 (t_Δ,β) 的元素个数与 S 中的元素个数的比值，通过比值来判断对象是否已经失效[10]。

4 实验验证

这里提出的分布式系统中的层次式失效检测算法，重点针对层次式失效检测方法中的单点失效和检测时延问题加以改进。因此，实验的重点在于比较算法改进前与改进后各指标的变化，以及改进后的算法对失效检测准确性的影响效果。

(1)实验环境搭建

5 台计算机网络连接，其上部署分布式系统，5 台计算机配置分别为 3 台 Intel Pentium E6700/3.2GHz 与 2 台 Intel Pentium E6300/2.8GHz。网络连接为 100Mb/s，其中前 3 台在一个局域网内，作为组 1；后两台在一个局域网内，作为组 2。每台计算机操作系统都采用 Windows XP。该网络与 IP 多播通讯协议兼容，调试平台为 Visual Studio 2008，算法编程语言采用 C++实现。

该网络中 5 台计算机即 5 个节点，其中，组 1 的一个节点作为被检测节点，主要用于发送心跳报文，另 2 个节点作为检测节点，运行两种失效检测算法。这两种算法分别是最基本的层次式失效检测算法(算法 1)和本文提出的层次式失效检测改进算法(算法 2)。同样，组 2 的两个节点也都作为检测节点，分别运行以上两种失效检测算法，目的是将组 1 和组 2 的两种算法的实现分别进行对比实验。

(2)实验预置环境参数

两个算法的预设参数如下[11]：

$\beta=1$，$delay_o=10$ms，$\Delta_i=5000$ms，$\gamma=0.1$，$\phi=4$，$T_o=100$。

为保证该对比实验的准确性，对网络中各个节点进行检测与对比。目标是保证在网络状况良好的条件下，算法 1 和算法 2 在准确性方面并无太大区别。为此对该实验的验证专门配置了高负载的网络环境：在此 5 台计算机构成的分布式系统内任意节点上运行一个循环发包器，在网络中大量传送报文。同时，在心跳发送过程中，人工设置心跳发送的延迟，主要用于模拟在大规模分布式系统中高负载情形下导致的网络延迟。采用人工设置方式对算法 1 和算法 2 设置相同的延迟时间，目的是为了对比在相同的外界环境下，这两个算法的性能指标。

(3)实验结果比较

两种算法在网络状况良好的条件下，进行的 100 次失效检测中均无误判；但针对大规模分布式系统中高负载情形下导致的网络延迟，两种算法在失效检测准确性出现差异。实验数据对比如表 1 所列。

表 1 算法 1 和算法 2 的准确性对比

	算法	网络延迟/s	检测次数	误判次数	失效次数	未失效次数	对未失效的误判次数
组 1	算法 1	2	100	4	0	100	4
	算法 2	2	100	1	0	100	3
组 2	算法 1	2	100	5	0	100	5
	算法 2	2	100	2	0	100	4
组 1	算法 1	5	100	7	0	100	8
	算法 2	5	100	2	0	100	4
组 2	算法 1	5	100	8	0	100	9
	算法 2	5	100	3	0	100	5

实验结果表明：(1)被检测节点的真实情况均未失效，但

(下转第 99 页)

数为 n，普通节点数为 N。本文所提方案与文献[8]方案的对比结果如表 1 所列。

如表 1 所列，本文提出的方案可增强系统抵御消息篡改攻击能力，在申请节点私钥更新阶段，与文献[8]方案相比，申请节点使用其私钥额外进行 h 次解密运算。为抵御消息篡改攻击和拥有合法身份的恶意节点发起的拒绝服务攻击，DP-KG 节点增加的运算量为 1EP+1CL。

由以上分析可知，本文方案在满足文献[7,8]方案所具有的安全特性基础上，申请节点仅增加 hDS 的计算开销、应答节点增加 1EP 的计算开销，即可使系统具有抗消息篡改攻击能力；应答节点仅增加 1CL 计算开销，即能够抵御拥有合法身份恶意节点发起的拒绝服务攻击。

结束语 构建合理的卫星网络密钥管理方案，关键在于怎样平衡效率与安全需求。提出的改进方案具有如下特点：在增加较小计算开销基础上，能够抵御消息篡改攻击行为，能有效检测拥有合法身份的恶意节点发起的频繁更新私钥分量申请，抵御恶意节点发起的拒绝服务攻击。提出的方案为设计与构建卫星网络安全密钥管理方案提供了一定的思路。

参 考 文 献

[1] Shamir A. Identity-based Cryptosystems and Signature Schemes [C]// Proc. of CRYPTO'84. New York, USA: [s. n]. 1984: 47-53

[2] Boneh D, Franklin M. Dentity-based encryption from the Weil pairing [C]// Advances in Cryptology CRYPTO 2001. Berlin: Springer-Verlag, 2001: 213-229

[3] Zhou L, Hass Z J. Securing Ad-hoc networks [J]. IEEE Networks, 1999, 13(6): 24-30

[4] 吴平，王保云，徐开勇. 基于身份的 Ad-hoc 网络密钥管理方案[J]. 计算机工程，2008，34(24)：143-145

[5] 杨德明，慕德俊，许钟. Ad hoc 空间网络密钥管理与认证方案[J]. 通信学报，2006，27(8)：104-107

[6] 彭长艳，沈亚敏，王剑，等. 基于身份的空间网络安全研究[J]. 飞行器测控学报，2008，27(3)：56-62

[7] 李伟，罗长远，初晓. 分布式网络中基于 IDPKC 的私钥更新方案[J]. 计算机应用，2009，29(7)：1825-1827

[8] 李伟. 空间信息网络密钥管理研究[D]. 郑州：解放军信息工程大学，2009

(上接第 74 页)

因为网络延迟而造成对未失效节点的误判；(2)在检测节点与被检测节点同处一个局域网内时，即在组 1 内，算法 1 和算法 2 对未失效的误判次数要低于在组 2 内的。(3)在网络延迟时间很短的情况下，算法 1 和算法 2 在误判次数上并无太大差异，但在网络延迟时间较长的情况下，由于算法 2 增加了信任度变量和修正比例因子，因此算法 2 的误判次数明显低于算法 1。(4)算法不同对未失效节点的误判次数也不相同，无论网络延迟的设置为多少，算法 2 在实验中对失效的误判率均低于算法 1，特别是当网络延迟越长的情形表现得越明显。

为验证算法 1 和算法 2 对实际上已经失效的进程做出错误判断的情况，在实验中，程序中设置一个计时器，专门记录算法对已经失效进程的判断时间，记录周期为 5ms。两种算法都自动设置进行了 100 次失效检测后，手动结束各节点上的被检测进程，模拟被检测进程已经完全失效。然后手动对两种算法的第 101 次失效进行检测，实验中的数据记录如表 2 所列。

表 2 算法 1 和算法 2 对失效进程的判断时间比较

算法	延迟时间/ms	判断时间 t_Δ/ms
算法 1	2000	480
算法 2	356	62

从表 2 中可知，对于被检测进程已真正失效的情况，算法 1 将过去的误判次数直接转换为下次检测延迟时间的算法，实际增加了延迟时间，造成判断时间的加长；算法 2 在增加了信任度变量和修正比例因子的情形下，能更快判定出已失效进程的状态，提高了检测效率。

结束语 综上所述，本文在对象级、进程级和主机级的层次式失效检测方法的思想指导下，以 Chen 预测算法为基础，通过增设一个信任度变量和修正比例因子，提出了一种分布式系统中的层次式失效检测的改进算法。改进算法考虑到传统的分布式系统中层次式失效检测方法中的单点失效问题和检测时间过长等因素，在分层时考虑了网络拓扑结构，把局域网的检测消息限制在组内，减少了广域网的负载和检测时间；同时，该方法使组内的节点都承担不同组之间检测的任务，从而避免了失效检测系统受单点失效的影响。

实验结果表明，改进后的方法其理论上更趋于合理性，综合及时性与准确性两个标准因素考虑，改进算法能够提高失效检测的准确性和检测效率，其性能优于传统的分布式系统中的层次式失效检测方法。同时，该研究成果也为下一步的优化检测方法研究提供了参考和有用的理论依据。

参 考 文 献

[1] Foster I, Kesselman C, Tuecke S. The anatomy of the grid[J]. Int'l Journal of High Performance Computing Applications, 2001, 15(3): 200-222

[2] Stelling P, Foster I, Kesselman C, et al. A fault detection service for wide area distributed computations [C]// Schmidt D, ed. Proc. of the 7th IEEE Symposium on High Performance Distributed Computing. Chicago: IEEE Computer Society Press, 1998: 268-278

[3] 云晓春，余翔湛. 基于确认度失效检测算法的研究与设计[J]. 北京邮电大学学报，2005，28(3)：10-13

[4] Hayashibara N, Défago X, Yared R, et al. The φ-accrual failure detector[C]// Titsworth FM, ed. IEEE Int'l Symp. on Reliable Distributed Systems (SRDS 2004). Florianopolis: IEEE Computer Society Press, 2004: 66-78

[5] Sotoma I, Madeira E R M. Adaptation—Algorithms to adaptive fault monitoring and their implementation on CORBA [C]// Blair G, Schmidt D, Tari Z, eds. Int'l Symposium on Distributed-objects and Applications (DOA 2001). Rome: IEEE Computer Society Press, 2001: 219-228

[6] 董剑，左德承，刘宏伟，等. 一种基于 QoS 的自适应网格失效检测器[J]. 2006，17(11)：2362-2372

[7] 陈宁江. 面向 Web 服务的层次型动态失效检测体系[J]. 计算机工程，2009，35(17)：94-96

[8] Chen W, Toueg S, Aguilera M K. On the quality of service of failure detectors[J]. IEEE Trans. on Computers, 2002, 51(5): 561-580

[9] 刘鹏. 分布式系统中容错中间件冗余服务的设计与实现[D]. 长沙：中南大学，2009

[10] 栾兰，刘淑芬，张欣佳. 基于检测点失效检测算法的研究与改进[J]. 吉林大学学报，2008，46(4)：681-686