(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 1568467 B
(45)授权公告日 2010.06.16

(21)申请号 02820026.8

(22)申请日 2002.09.05

(30)优先权数据
60/317,718 2001.09.06 US
60/317,566 2001.09.06 US
10/234,693 2002.09.04 US
10/234,597 2002.09.04 US

(85)PCT申请进入国家阶段日
2004.04.09

(86)PCT申请的申请数据
PCT/US2002/028199 2002.09.05

(87)PCT申请的公布数据
WO03/023633 EN 2003.03.20

(73)专利权人 BEA 系统公司
地址 美国加利福尼亚州

(72)发明人 迪安·B·雅各布斯
埃里克·哈尔彭

(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 吕晓章 马莹

(51)Int.Cl.
G06F 15/16(2006.01)
G06F 15/173(2006.01)

(56)对比文件
全文.
全文.
US 5802291 A,1998.09.01, 第3栏第54行至第4栏第23行.
US 6067477 A,2000.05.23, 摘要,第3栏第26行至第5栏第38行,第9栏第30行至第10栏第41行.

审查员 李琼

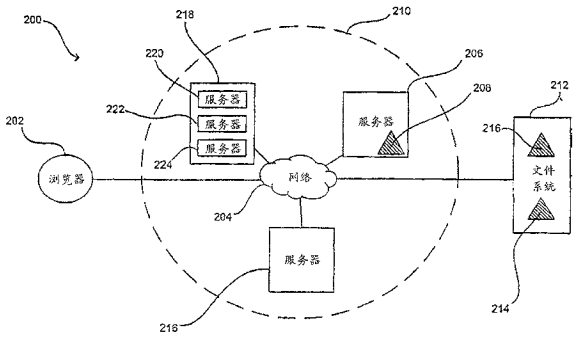
权利要求书 5 页 说明书 10 页 附图 7 页

(54)发明名称

正好一次高速缓存器结构

(57)摘要

一种用于对位于包含一个文件系统的群集网络中的对象进行管理的系统,该文件系统包含至少一个数据对象的拷贝(214)。该系统可以包括与文件系统(212)通信的多个群集服务器。选择一个引导服务器,其包含用于选择主服务器(206)的分布式一致算法,并在执行该算法循环时利用多点传送。所选择的主服务器(206)可以包含诸如本地超高速缓存器中的数据对象(214)的拷贝(208),以便将对本地拷贝(208)的访问提供给群集器中的任一其它服务器。在文件系统(212)中可以更新由主服务器(206)接纳的项的变化。如果主服务器(206)变得不能接纳对象,可以使用分布式一致算法选择新主服务器,并利用多点传送将新主服务器通知给其它的服务器(216、218)。



CN 1568467 B

1. 一种用于管理网上对象的系统，包括：

多个网络服务器，每个网络服务器用于与网络数据源通信；以及

位于所述多个网络服务器中的一个引导服务器，该引导服务器包含用于从所述多个网络服务器中选择一主服务器的一个分布式一致性算法，该主服务器包含与网络数据源中的一个数据项有关的一个对象，从而使需要访问数据项的所述多个网络服务器中的任何一个都可以访问所述主服务器上的所述对象。

2. 根据权利要求 1 的系统，其中，所述网络服务器是从由硬件群集器服务器和软件群集器服务器组成的一个组中选择出来的。

3. 根据权利要求 1 的系统，其中，所述分布式一致性算法包括所述引导服务器与所述多个服务器之间的消息循环，该循环一直继续，直到所述多个网络服务器中的大多数同意所述主服务器为止。

4. 根据权利要求 1 的系统，其中，所述主服务器包含一数据对象，该数据对象包含来自所述网络数据源的数据的拷贝。

5. 根据权利要求 1 的系统，其中，所述主服务器包含一数据对象，该数据对象用作对网络数据源中的数据项进行访问的唯一接入点。

6. 根据权利要求 1 的系统，其中，所述数据项是一事项记录。

7. 根据权利要求 1 的系统，其中，所述分布式一致性算法是 Paxos 算法。

8. 一种用于管理网上对象的系统，包括：

多个网络服务器，每个网络服务器用于与网络数据源通信；以及

位于所述多个网络服务器中的一引导服务器，该引导服务器包含用于从所述多个网络服务器中选择一主服务器的一个分布式一致性算法，该主服务器包含位于网络数据源中的一个数据项的一个拷贝，从而使需要访问所述数据项的所述多个网络服务器中的任何一个都可以访问所述主服务器上的所述拷贝。

9. 一种用于管理网上对象的系统，包括：

多个网络服务器，每个网络服务器用于与网络数据源通信；以及

所述多个网络服务器中的一引导服务器，该引导服务器包含用于从所述多个网络服务器中选择一主服务器的一分布式一致性算法，该主服务器包含对位于网络数据源中的数据项进行访问的唯一接入点，从而使需要访问该数据项的所述多个网络服务器的任何一个都必须经过该主服务器访问该数据项。

10. 一种用于管理网上对象的系统，包括：

一文件系统，该文件系统包括至少一个数据项的拷贝；

与该文件系统进行通信的多个服务器；以及

位于所述多个服务器中的一引导服务器，该引导服务器包含用于从所述多个服务器中选择一主服务器的一分布式一致性算法；以及

位于所述多个服务器中的一主服务器，所述主服务器包含所述数据项的本地拷贝，所述主服务器可使所述多个服务器中的任何一个访问该本地拷贝，并且在更新本地拷贝的任何时间内更新所述文件系统中的所述数据项的所述拷贝。

11. 根据权利要求 10 的系统，其中，所述主服务器进一步用于将本地拷贝存储在本地超高速缓存器中。

12. 根据权利要求 10 的系统，其中，所述多个服务器包括一群集器。

13. 根据权利要求 10 的系统，其中，所述文件系统将数据项复制在多个磁盘上。

14. 一种用于管理网上对象的系统，包括：

一文件系统，该文件系统包括至少一个数据项的拷贝；

与该文件系统进行通信的多个服务器；

一硬件群集器，该硬件群集器包括位于所述多个服务器中的硬件群集器服务器，所述硬件群集器包括用于从所述多个硬件群集器中选择一引导服务器的一算法；

位于所述硬件群集器服务器中的一引导服务器，该引导服务器包括用于从所述多个服务器中选择一主服务器的一分布式一致性算法；以及

位于所述多个服务器中的一主服务器，所述主服务器包括数据项的本地拷贝，所述主服务器可使所述多个服务器中的任何一个可访问该本地拷贝并且只要更新了本地拷贝时就更新文件系统中的数据项的拷贝。

15. 根据权利要求 14 的系统，其中所述主服务器位于所述硬件群集器中。

16. 一种用于管理网上对象的方法，包括：

利用分布式一致性算法从多个网络服务器中选择一主服务器；

将一个数据项的拷贝从文件系统传送给所述主服务器；以及

通知其它网络服务器包含有一个数据项拷贝的所述主服务器将被用在对网络请求的处理中。

17. 根据权利要求 16 的方法，进一步包括当主服务器上的拷贝被修改时，更新文件系统中的数据项的步骤。

18. 根据权利要求 16 的方法，进一步包括限制其它网络服务器经过所述主服务器来访问文件系统的步骤。

19. 根据权利要求 16 的方法，进一步包括保证只有数据项的一个拷贝存在于该文件系统的外部的步骤。

20. 根据权利要求 16 的方法，进一步包括保证数据项的一个拷贝总是存在于该文件系统外部的步骤。

21. 根据权利要求 16 的方法，进一步包括如果主服务器不再接纳该对象，则利用分布式一致性算法从多个网络服务器中选择新的主服务器的步骤。

22. 根据权利要求 16 的方法，进一步包括如果主服务器不再接纳该对象，则将一个数据项的拷贝从所述文件系统传送给一个新的主服务器的步骤，该新的主服务器是使用所述分布式一致性算法选择出的主服务器。

23. 根据权利要求 16 的方法，进一步包括如果主服务器不再接纳该对象，那么通知其它网络服务器一个包含所述数据项的新的主服务器将被用在对网络请求进行的处理中的步骤。

24. 一种用于管理网上对象的结构系统，包括：

多个服务器，每个服务器都可高速缓存一数据对象；

一文件系统，该文件系统包括数据项的至少一个拷贝；

一分布式一致性算法，用于从所述多个服务器中选择一主服务器，该主服务器用于高速缓存数据对象的拷贝；以及

一分布系统，用于通知网络上的服务器该主计算机包含所述数据对象的拷贝。

25. 一种用于将对象出租给网络上服务器的方法，包括：

利用分布式一致性算法从多个网络服务器中选择一主服务器；

将数据对象出租给所述主服务器，其中该主服务器在出租时间段内拥有或接纳该数据对象；

从文件系统中取出数据项的拷贝以送至主服务器；以及

向其它网络服务器通知主服务器包含数据项的拷贝以用于对网络请求进行处理。

26. 根据权利要求 25 的方法，进一步包括步骤：在另一出租时间段内，定期地更新将该数据对象出租给主服务器。

27. 根据权利要求 25 的方法，进一步包括步骤：一旦主服务器上的出租时间段到期了，则将该数据对象出租给新的主服务器一个出租时间段。

28. 一种用于将对象出租给网络上的服务器的方法，包括：

从硬件群集器中的多个硬件群集器服务器中选择一引导服务器；

利用所述主服务器上的分布式一致性算法从多个网络服务器中选择一主服务器；

将数据对象指定给主服务器，被指定的主服务器在一个规定的时间周期内可对数据项进行单独的访问；

从文件系统中取出数据项的拷贝以送至主服务器；以及

向其它网络服务器通知主服务器包含数据项的拷贝以用于对网络请求进行处理。

29. 一种用于对网络上的对象的所有权进行分布的方法，包括：

从硬件群集器中的多个硬件群集器服务器中选择一引导服务器；

利用所述引导服务器上的分布式一致性算法从多个网络服务器中选择一主服务器；

将数据对象指定给该主服务器，被指定的主服务器可对网络上的数据对象进行单独访问；

从文件系统中取出数据项的拷贝以送至主服务器；以及

向其它网络服务器通知主服务器包含数据项的拷贝以用于对网络请求进行处理。

30. 一种用于确保在群集器中存在对象的方法，包括：

使用位于多个服务器中的主服务器提供对数据对象的访问；

如果所述主服务器不能提供对所述数据对象的访问，那么，利用分布式一致性算法从多个服务器中选择一新的主服务器；

将信息传送给需要提供对所述数据项提供访问的新的主服务器，和

通知所述多个服务器中的其它服务器新的主服务器可访问该数据对象。

31. 一种用于在群集器中分布对象的方法，包括：

利用分布式一致性算法来从多个网络服务器中选择一主服务器；

将一数据对象指定给该主服务器，被指定的该主服务器可对一个数据项进行单独访问；以及

将一通知多点传送到群集器中的其它服务器以通知包含所述数据项的拷贝的所述主服务器将被用在对网络请求的处理中。

32. 一种用于在群集器中分布对象的方法，包括：

利用分布式一致性算法来从多个网络服务器中选择一主服务器；

连接群集器中的每个服务器以确定所选择的主服务器是否可被那些服务器所接受；以及

如果群集器中的所有服务器都同意所选择的主服务器是可接受的，那么将一通知多点传送到群集器中的其它服务器以通知交由所选择的新主服务器进行负责。

33. 一种用于在群集器中分布对象的系统，包括：

装置，用于利用分布式一致性算法来从多个网络服务器中选择一主服务器；

装置，用于从一文件系统中取出数据项的拷贝以送至主服务器；以及

装置，用于通知其它网络服务器包含该数据项拷贝的该主服务器将被用在对网络请求进行的处理中。

34. 一种用于管理网上对象的方法，包括：

利用 Paxos 算法从软件群集器中的多个网络服务器中选择一主服务器；以及

将数据对象指定给所述主服务器，其中该主服务器提供在网络中对该数据对象的单独访问。

35. 根据权利要求 34 的方法，进一步包括从一文件系统中取出该数据对象数据的步骤。

36. 根据权利要求 34 的方法，进一步包括通知其它网络服务器包含与所述数据项相关的对象的该主服务器将被用在对网络请求进行的处理中的步骤。

37. 根据权利要求 34 的方法，进一步包括将新的主服务器的标识多点传送至另一个网络服务器的步骤。

38. 根据权利要求 34 的方法，其中利用 Paxos 算法来从软件群集器中的多个网络服务器中选择一主服务器的所述步骤包括将循环信息多点传送到另一个网络服务器。

39. 一种用于对群集器中的对象进行分布的系统，包括：

装置，用于利用分布式一致性算法来从多个网络服务器中选择一主服务器；

装置，用于将一 JMS 对象分布给主服务器，所分布的该主服务器提供对 JMS 对象的单独访问；以及

装置，用于通知其网络服务器该主服务器提供对 JMS 的单独访问。

40. 一种用于管理网上对象的方法，包括：

利用 Paxos 算法从软件群集器中的多个网络服务器中选择一主服务器；以及

将 JMS 消息存储器指定给主服务器，该 JMS 消息存储器仅存在于所述主服务器上。

41. 根据权利要求 40 的方法，进一步包括通知其它网络服务器该主服务器接纳所述 JMS 消息存储器的步骤。

42. 根据权利要求 40 的方法，进一步包括步骤：将新的主服务器的标识多点传送至另一个网络服务器。

43. 根据权利要求 40 的方法，其中利用 Paxos 算法来从软件群集器中的多个网络服务器中选择一主服务器的所述步骤包括将循环信息多点传送到另一个网络服务器。

44. 一种用于管理网上 JMS 对象的系统，包括：

多个网络服务器；以及

所述多个网络服务器中的一引导服务器，该引导服务器包括用于从所述多个网络服务器中选择主服务器的一分布式一致性算法，该主服务器包括一 JMS 对象以便其需要访问

JMS 的多个网络服务器的任何一个必须对主服务器上的 JMS 对象进行访问。

45. 根据权利要求 44 的系统，其中，所述网络服务器是从由硬件群集器服务器和软件群集器服务器所组成的这样一组中选择出来的。

46. 根据权利要求 44 的系统，其中，所述分布式一致性算法包括所述主服务器与所述多个服务器之间的循环消息，该循环一直继续直到所述多个网络服务器的大多数同意主服务器。

47. 一种用于将 JMS 对象出租给网络上的服务器的方法，包括：

利用分布式一致性算法来从多个硬件群集器服务器中选择一主服务器；

将 JMS 对象出租给该主服务器，其中该主服务器在出租时间段内拥有或接纳该 JMS 对象；以及

向其它网络服务器通知该主服务器对该 JMS 对象进行接纳。

48. 根据权利要求 47 的方法，进一步包括步骤：在另一出租时间段内，定期地更新将该 JMS 对象出租给主服务器。

49. 根据权利要求 48 的方法，进一步包括步骤：一旦所述主服务器上的出租时间段到期了，则将所述 JMS 对象出租给新的主服务器一个出租时间段。

50. 一种用于将 JMS 对象出租给网络上的服务器的方法，包括：

从硬件群集器中的多个硬件群集器服务器中选择一引导服务器；

利用所述引导服务器上的分布式一致性算法来从多个网络服务器中选择一主服务器；

将 JMS 对象指定给该主服务器，被指定的主服务器能够在规定的一个时间周期内提供对所述 JMS 对象的单独访问；以及

通知其它网络服务器该主服务器接纳所述 JMS 对象。

51. 一种用于对网络上的对象进行分布的方法，包括：

从硬件群集器中的多个硬件群集器服务器中选择一引导服务器；

利用所述引导服务器上的分布式一致性算法来从多个网络服务器中选择一主服务器；

将 JMS 对象指定给该主服务器，被指定的主服务器可以提供对所述 JMS 对象的单独访问；以及

通知其它网络服务器该主服务器接纳所述 JMS 对象。

正好一次高速缓存器结构

[0001] 本申请要求这里所引入的下述申请的优先权：

[0002] 由Dean Bernard Jacobs和EricHalpern于2001年9月6日所申请的申请号为No.60/317,718、发明名称为“EXACTLY ONCE CACHE FRAMEWORK”的美国临时专利申请。由Dean Bernard Jacobs和EricHalpern于2002年9月4日所申请的发明名称为“EXACTLY ONCE CACHE FRAMEWORK”美国专利申请。

[0003] 由Dean Bernard Jacobs和EricHalpern于2001年9月6日所申请的申请号为No.60/317,566、发明名称为“EXACTLY ONCE JMSCOMMUNICATION”的美国临时专利申请。

[0004] 由Dean Bernard Jacobs和EricHalpern于2002年9月4日所申请的发明名称为“EXACTLY ONCE JMS COMMUNICATION”的美国临时专利申请。

[0005] 版权须知

[0006] 该专利文献所公开的部分包括受到版权保护的材料。当它出现在专利与商标局的专利文件或者记录中时，版权所有者不反对任何人传真重现该专利所公开的专利文献，反之，则无论如何都保留所有版权。

[0007] 交叉参考的案例：

[0008] 下述美国专利申请是交叉参考案例并且在这里引入以做参考：

[0009] 由Dean Bernard Jacobs、Reto Kramer、以及Ananthan Bala Srinivasan于2001年7月16日所申请的申请号为No.60/305,986、发明名称为“DATAREPLICATION PROTOCOL”的美国专利申请。

技术领域

[0010] 本发明涉及用于在网络群集器中的服务器当中分布对象的技术。

背景技术

[0011] 在分布式计算机系统中，经常存在这样的情况，即若干个服务器和/或网络节点必须一起工作。当在所述多个设备当中存在需要共享的典型网络信息以便允许它们用做单一实体时，必须对这些服务器和节点进行协调。就资源和效率而言，通常可对设备进行协调的方法是非常昂贵的。

[0012] 通常，由于在多个节点之间存在若干信息传送，因此，为了使这些节点一致，需要某种同步。然而，在群集网络环境中这种同步要求是所不希望的。许多群集环境简单避免了利用任何这种同步要求。然而，在某些应用中，这种一致是必需的。

[0013] 在某些需要一致的情况下，存在一个群集器试图排除对其进行访问的设备。一种这样的设备是一个事项注册文件系统。只要事项处理在进行中，就存某些必须以持续方式保存的对象，从而如果出现故障，则可恢复持续保存的对象。

[0014] 对于其需要被保存在一个位置上的对象而言，一般存在其运行于群集器或域中每个服务器上的一事项监视器，该事项监视器此后使用本地文件系统来访问对象。每个服务器可以具有其自己的事项管理器，以便在持续性上几乎不存在问题。由于每个服务器都具

有其自己的事项管理器，所以，也不需要协调。

[0015] 例如，可能存在包括三个服务器的群集器，每个服务器具有一事项管理器。这些服务器中的一个可能遇到故障或者由于该服务器不可用于群集器所引起的其它问题。由于有故障的服务器是唯一可访问特定事项处理记录的服务器，所以，在该服务器再次可用于该群集器之前，不可能恢复特定记录上的任何事项。由于服务器必须花费大量的时间来解决这些问题，所以恢复该记录是很困难的或者至少效率很低。重要的服务器问题可能包括诸如服务器上的母板短路或者电源被烧坏这样的事件。

发明内容

[0016] 本发明包括这样一个系统，该系统用于对诸如存储在网络或者群集器上的服务器中的对象进行管理。该系统包括一数据源、应用、或者诸如文件系统或者 Java 消息服务部件这样的位于群集器之内或者群集器之外的服务。该系统包括若干个服务器，这些服务器诸如通过高速网络连接而与该文件系统或者应用进行通信。

[0017] 该系统包括诸如另一个服务器所同意的一引导服务器 (lead server)。该引导服务器包含于硬件或者软件群集器中。该系统包括用于从服务器当中选择引导服务器的一算法，该算法例如可以是内置在硬件群集器设备中的算法。该引导服务器依次包括诸如Paxos算法这样的用于选择主服务器的一分布式一致性算法。用于选择引导服务器的算法可以与用于选择主服务器的算法不同或者与其相同。

[0018] 该主服务器包括诸如存储在本地超高速缓存器中的事项或者对象的拷贝。主服务器提供了可对网络或者群集器中的任何服务器进行访问的本地拷贝。主服务器还可以提供对存储在文件系统中的对象进行访问的唯一接入点，或者可提供对应用或者服务进行访问的唯一接入点。还可以在文件系统、应用、或者服务中更新对主服务器所高速缓冲的、所接纳的、或者所拥有的事项所做出的任何变化。

[0019] 如果主服务器变得不能接纳所述对象，那么使用分布一致性算法选择新的主机，此后从文件系统或者服务中取出该对象所需的数据。群集器中的另一个服务器被通知一个新的服务器正在接纳该对象。通过诸如点到点连接或者通过多点传送这样的适当手段来通知服务器。

附图说明

[0020] 图 1 给出了根据本发明一个实施例的分布式对象系统的示意图；

[0021] 图 2 给出了根据本发明一个实施例的另一个分布式对象系统的示意图；

[0022] 图 3 给出了根据本发明的用于选择主服务器的一方法的流程图；

[0023] 图 4 给出了根据本发明的用于选择新的主服务器的一方法的流程图；

[0024] 图 5 给出了根据本发明的用于使用引导服务器的一方法的流程图；

[0025] 图 6 给出了根据本发明一个实施例的 JMS 消息存储系统的示意图；

[0026] 图 7 给出了根据本发明所使用的计算机系统的部件的方框图。

具体实施方式

[0027] 根据本发明的系统提供了诸如当拥有数据对象的服务器变得不可用于服务器群

集器时发布有效性的解决方案。这样一种解决方案可使该群集器中的另一个服务器拥有该数据对象的所有权。但是，出现了这样一个问题，即在不需对两个服务器上的数据对象进行复制的情况下即可到这两个服务器访问所述数据对象。

[0028] 如果群集器使用文件系统、数据存储器、或者数据库（在下文中被总称为“文件系统”）以持续的存储数据，并且不止一个服务器可访问该文件系统，那么如果拥有那个对象的第一服务器遇到问题，第二服务器即可自动地接管访问数据对象的任务。另外，可利用群集器或者群集器中的服务器所使用的算法来指令服务器接管所述项的所有权。然而，另一个基本问题包括使群集器同意哪一个服务器目前拥有资源或者对象，或者在服务器中间实现“一致同意”。

[0029] 图 1 给出了根据本发明的群集器系统 100 的一个例子，在该例子中将诸如事项注册 114 这样的对象存储在文件系统 112 中。群集器 110 中的所有服务器 106，116，118 都可访问该文件系统 112，但是这些服务器中只有一个每次都可访问注册 114。诸如通过存储注册 114 的拷贝 108 或者通过都可访问文件系统 112 中的注册 114，群集器 110 中的服务器当中的主服务器 106 将“拥有”或者“接纳”该注册 114。群集器 110 中的其它任何服务器 116，118 可以访问该记录的拷贝 108，和 / 或可通过主服务器 106 来访问注册 114。例如，一客户或者浏览器 102 可以对其指群集器 110 中的服务器 116 的网络 104 进行请求。该服务器可通过网络 104 来访问主服务器 106 上的事项记录的拷贝 108。如果必须更新事项记录，那么拷贝 108 与文件系统 112 上的原注册 114 一起被更新。

[0030] 例如当服务器作为对象的存储器时，诸如通过将数据对象的拷贝存储在本地超高速缓存器中并且使该群集器中的其它服务器可使用该拷贝，或者通过使唯一的服务器可随机访问文件系统中的对象，该服务器则可“拥有”或者“接纳”该数据对象，以致该群集器中的所有其它服务器必须通过该主服务器来访问那些对象。这保证了对象“正好一次”存在于服务器群集器中。

[0031] 图 3 给出了一个处理 300，该处理 300 用于建立对一个对象的接纳。利用诸如 Paxos 算法这样的分布式一致性算法 302 来选择主服务器。因为群集器中的服务器通常必须就怎样在群集器服务器中分布对象而达成一般同意或一致同意，所以，这种算法被称为“分布式一致性”算法。

[0032] 如果被接纳的对象例如是被高速缓存在主服务器中，那么，从文件系统中取出的数据对象的拷贝被传送至主服务器并作为一个对象存储在本地超高速缓存器 304 中。此后诸如通过主服务器向网络或者适当群集器中的其它服务器通知该对象的本地拷贝存在于主服务器中并且本地拷贝将被用在对将来的网络要求 306 进行的处理中。

[0033] 在是分布式一致性算法的一个例子的 Paxos 算法中，通过网络服务器来选择一服务器以作为主服务器或者引导服务器，该网络服务器引导了一系列的“一致循环 (consensus rounds)”。在每个一致循环中，建议了新的主服务器或者引导服务器。循环一直继续直到多数或者法定数目的服务器接受所建议的服务器。尽管该系统被配置成总是由引导服务器启动一循环以便选择一主服务器，但是，任何服务器都可通过启动一循环来建议主服务器或者引导服务器。同时可进行用于不同选择的循环。因此，通过循环数或者这样一对值来识别循环选择，至于这一对值，其中的一个与所述循环相关，而另一个与引导所述循环的服务器相关。

[0034] 用于这样一个循环的步骤如下，尽管其它步骤和/或方法可适于某些情况或者应用。首先，通过引导服务器将一“集中”消息传送到群集器中的其它服务器来启动一个循环。集中消息集中了来自群集器中服务器的与在前这些服务器所参与进行的循环有关的信息。如果对于一个特定的选择处理存在在前的一致循环，那么所述集中消息还通知服务器不要提交来自在前循环的选择。例如，一旦引导服务器已集聚了来自至少一半群集器服务器的回应，那么引导服务器就可以决定所述值以建议下一个循环并且将该建议传送到群集器服务器以作为“启动”消息。为了在这种方法中使引导服务器选择一值以提供建议，必须接收来自服务器的初始值信息。

[0035] 一旦服务器接收到来自引导服务器的启动消息，它通过传送一“接受”消息来做出响应，这表明该服务器接受所建议的主/引导服务器。如果该引导服务器接收到多数或者法定数目服务器的接受消息，那么引导服务器将其输出值设置为在循环中所建议的值。如果引导服务器在规定的时段内没有接收到多数或者法定数目的接受（“一致同意”），那么引导服务器可启动新的循环。如果引导服务器接收到一致同意，那么引导服务器可以通知所述群集器或者网络服务器所述服务器将被指定为所选择的服务器。可通过任何适当的广播技术将该通知广播到网络服务器，例如可通过点到点连接或者多点传送。

[0036] 通过建议利用与在前循环有关的信息的选择可保证一致同意方法的一致同意条件。要求该信息来自至少大多数的网络服务器，以便对于任意两个循环来说存在至少一个参与了两个循环的服务器。

[0037] 所述引导服务器可通过向每个服务器询问服务器接受一值的最新循环编号、可能还要询问所接受的值来选择一值。一旦引导服务器从多数或者法定数目的服务器中获得了该信息，它可以选择用于新循环的值，该值等于响应当中最新循环的值。如果没有一个服务器涉及在前的循环，那么引导服务器还可以选择初始值。例如，如果引导服务器接收到上次所接受的循环是 x 的一响应，并且当前循环是 y，那么服务器表示不接受 x 与 y 之间的任何循环，以便保持一致性。

[0038] 循环引导服务器与网络服务器之间的抽样交互包括以下消息：

[0039] (1)“Collect”——将一消息传送到正在启动一个新循环“r”的服务器。该消息可采取 m = (“Collect”,r) 的形式。

[0040] (2)“Last”——将来自一网络服务器的消息传送给引导服务器，该网络服务器提供了上次循环所接受的“a”以及该循环的值“v”。该消息可采取 m = (“Last”,r,a,v) 的形式。

[0041] (3)“Begin”——将一消息传送给发布与循环 r 相关的所述值的服务器。该消息可采取 m = (“Begin”,r,v) 的形式。

[0042] (4)“Accept”——将一消息从用于接受与循环 r 相关的所述值的服务器传送给所述引导服务器。该消息可采取 m = (“Accept”,r) 的形式。

[0043] (5)“Success”——将一信息传送给用于发布与循环 r 相关的所述值 v 的选择的服务器。该信息可采取 m = (“Success”,r,v) 的形式。

[0044] (6)“Ack”——将一消息从一服务器传送给引导服务器，该服务器承认已经接收到与循环 r 相关的决定。该信息可采取 m = (“Ack”,r) 的形式。

[0045] 存在与其位于硬件群集器或者软件群集器内部或者外部的服务器相分离的一文

件系统。该文件系统诸如通过将记录存储在第一磁盘并将该记录拷贝到位于文件系统之内的第二磁盘来持续的存储事项记录。如果第一磁盘划碰了，那么文件系统可使所述群集器和/或服务器无法察觉所述划碰并且可从第二磁盘中获得记录信息。该文件系统还可以选择以将该记录拷贝到其用作第二磁盘备份的第三磁盘上。

[0046] 从群集器中的服务器的角度来看，该文件系统可以是单一资源。在一个实施例中，该服务器可以只关心单一服务器在任何时间处拥有该文件系统。

[0047] 根据本发明的另一个例子包括位于服务器群集器中的高速缓存器。例如因为网络性能的原因，因此希望在群集环境中使单一高速缓存器群集器中的服务器示出了数据对象。将多个项保存在单一高速缓存器中可能是有利的，因为群集器中的服务器可访问高速缓存器而无需不断的回到永久性存储器。取出已位于存储器中的多个项可极大的增加该系统的利用率，因为命中数据库或者文件系统的时间相对集中。

[0048] 然而，单一高速缓存器所具有的一个问题就是必须保证存储在存储器中的对象与存储在文件系统一磁盘中的对象相同。需要这种一致性的理由是保证在所高速缓存对象上所进行的任何操作或者计算产生了正确结果。另一个原因就是必须对由于高速缓存器划碰或者以别的方式而感染或者不可用所造成的文件系统的高速缓存器进行修复。

[0049] 至少有两个基本方式来处理群集器中的这类高速缓存，虽然其它方式至少对某些应用起作用。一种方式就是在多个地方复制该高速缓存器。该处理是有问题的，因为正被高速缓存项的任何变化都要求拷贝所述高速缓存器的所有服务器一致同意该变化，或者至少知道该变化。就资源以及性能来讲，经证明这是非常昂贵的。

[0050] 根据本发明的一可替换方法分布特定的服务器以使其是群集器中高速缓存器的所有者，并且通过这些特定的服务器都可访问高速缓存器。群集器中的任何服务器接纳这种高速缓存器。每个服务器可以接纳一个、若干个、或者不接纳高速缓存器。在单一服务器上寄主该高速缓存器，或者将其散布到群集器中某些或者所有服务器当中。群集器本身可以是任何适当的群集器，例如硬件群集器或者由处于给定“软件”群集器中的软件应用所指定的一组服务器。

[0051] 还可以考虑作为位于系统上某处的一种对象的事项登录和/或高速缓存器中的一个例子。可以指望保证任何一个这种对象在一个群集器中仅仅存在一次，并且该对象总是可用的。还可以指望保证如果接纳该对象的服务器出现了故障则所述对象可以被恢复到另一个服务器上，并且该对象可用于该群集器。

[0052] 图4给出了用于恢复的一个方法400。在该方法中，确定主服务器是否能够继续接纳对象402，例如确定服务器是否仍可用于该网络。如果不是，利用分布式一致性算法选择新的主机。这个选择可以是根据选择原主机404所使用的方法来执行。从文件系统中取出数据对象的拷贝被提供给新的主机，并且将其存储在本地超高速缓存器406中。网络上或适当群集器中的其它服务器被通知新的主服务器包括所述对象的本地拷贝，并且本地拷贝将用在对将来的任何网络要求408进行处理当中。

[0053] 根据本发明的系统和方法可以定义存在于群集器中正好一个位置中的对象，并且可保证这些对象总是存在的。从服务器的角度来看，是否利用诸如一个文件系统来镜像或者拷贝诸如事项记录这样的对象是无关紧要的。从服务器的角度来看，总有一个可被群集器中任何一个服务器访问的永久性存储器，该系统周期地检查对象的存在，或将对象所有

权指定为很短的时间周期，以便频繁的再指定对象以确保网络上或者群集器中的某个设备上的存在。

[0054] 硬件群集器包括一组设备，每个设备可运行多个服务器。在每个设备之后还存在一文件系统。硬件群集器中的服务器通常是由硬件构成的，所以，它们能更快的做出决定并且对硬件群集器内部的服务器故障进行处理。硬件群集器的大小被限制为包括有所述服务器的设备的实际硬件。硬件群集器中的服务器可以被用作软件群集器中的服务器，并且由于该设备上的单独的服务器可以用于该网络，所以还可以包括网络服务器。

[0055] 用于这些设备中的一个的共享文件系统诸如通过高速网络而可以用于群集器中的所有服务器。文件系统还可以是冗余的。在一个实施例中，通过使用文件系统的多个数据磁盘来实现该冗余。在这样一种冗余的实现中，在将对象写入到该文件系统的任何时候都可通过多个磁盘来拷贝对象。当将该文件系统看作是“黑盒子”时，该文件系统可承受任一磁盘当中的故障并且仍可提供群集器中任何服务器对数据项的访问。

[0056] 假设保存在存储器中的这些对象总是通过可靠的、永久性的存储器机构来恢复，则可建立根据本发明的其被称为“正好一次”结构的结构。例如，存在有表示事项处理记录的一个对象。只要产生了对该对象的调用，则更新对应的事项记录。这包括从数据库中所读取的或者写入到数据库中的一个调用。表示事项记录的对象位于诸如主服务器这样的群集器中的一个服务器上。只要群集器中至少一个服务器启动 (up) 并且运行，那么正好一次结构可确保如果另一个服务器有故障则该服务器可接管该记录的所有权。

[0057] 可能存在一个表示高速缓存器的对象。每当更新高速缓存器时，还可以将该更新写回到永久存储器中。当多个服务器中的一个服务器需要使用数据项时，要求该服务器全面研究 (go through) 该对象。如果用于接纳表示高速缓存器的对象的所述服务器发生故障，则将该对象恢复在另一个服务器上。所恢复的对象可以从永久存储器中取出所有的必要信息。

[0058] 正好一次结构可用作群集器所使用的存储缓冲器。该结构提供了其可提供系统中数据的单个高速缓存器，该系统是通过一可靠的、永久的存储器来恢复的。只要从高速缓存器中读取数据，无需访问永久的存储器即可执行该读取。然而，当将该更新写入到高速缓存器中时，必须通过永久的存储器来写回，以便如果存在一故障则可使该系统恢复。

[0059] 正好一次结构的一个重要方面包括一种方式，可在该方式中抽象出一种根据应用和 / 或实现而变化的方法。创建了诸如被称为“正好一次对象”的新型分布式对象。正好一次对象例如是文件系统中数据项的本地高速缓存拷贝，或者是对群集器中服务器的该数据项进行访问的唯一接入点。实现该抽象的根本技术也很重要。

[0060] 本发明的系统可利用其可用于分布式一致性的多个方法中的任何一个，诸如利用上述 Paxos 算法的一方法。选择这样一种算法，该算法提供了有效的方法以使多节点和 / 或分布式结点同意对象的一个值。即使节点有故障和 / 或在协商处理过程中返回，也可选择该算法以使其起作用。

[0061] 网络群集器的一般方法是利用可靠广播，在该方法中保证将每个消息传送到其所预定的接受者，或者至少将其传送到每个预定作用的服务器。该方法很难使系统并行化，因为可靠广播需要在移到下一个消息或者接受者上之前接受者要肯定应答收到一消息。利用多点传送的分布式算法可降低保证的次数，因为多点传送不能保证所有服务器接收一消

息。然而，多点传送可使该方法简单化以便该系统可参与并行处理，因为单一消息被并发的多点传送到所有群集器服务器，而无需等待来自每个服务器的反应。未接收到多点传送消息的服务器最后可从引导服务器或者另一个群集器服务器或者网络服务器中取出信息。如这里所使用的，网络服务器可参照网络上的任何服务器，无论该网络服务器是在硬件群集器中、软件群集器中、或者任何群集器之外。

[0062] 正好一次结构的一个重要方面是降低了一致的困难性。根据本发明，通过利用分布式一致性算法来多点传送消息可改善分布式一致性实现的性能。该方法可为使所有服务器相互一致同意而所需的消息交换和 / 或网络通信量最小化。

[0063] 当多点传送时，可采用若干方法中的一个。在其被称为“单相分布”的第一个方法中，引导服务器将一消息多点传送到位于网络上的所有其它服务器上，例如用在 Paxos 算法的循环中或者用于这样一种状态，即已为一对象选择了新的主机。在该方法中，主服务器只须传送一个消息，该消息被传送到网络中任何可用服务器上。如果一服务器临时的断开该网络，那么该服务器要求在回到该网络上之后对新的主机进行识别。

[0064] 利用另一个其被称为“双相分布”的多点传送方法，引导服务器可利用适当的算法的来预选择一主服务器。然而，在将一对象分布给该主机之前，该引导服务器可与群集器中的所有其它服务器相连接以确定服务器是否同意所选择的新的主服务器。引导服务器可通过点到点连接而与每个服务器相连接，或者可传送多点传送要求并且此后等待每个服务器以应答。如果服务器不同意所选的主机，那么引导服务器利用该算法来预选择新的主机。引导服务器此后在另一个循环中传送另一个多点传送请求以及所重新预选择的主机的标识。

[0065] 如果每个服务器都同意预选择的主机，那么引导服务器将该对象分布给主服务器。此后引导服务器多点传送一提交信息以通知服务器将新的变化已生效并且因此服务器将更新其信息。

[0066] 正好一次结构还可以利用“出租”机构。在利用这种机构的过程中，使用一种算法以使群集器服务器与引导服务器一致，例如通过利用分布式一致性算法。一旦选择了，该引导服务器担负将对象正好一次分布给群集器中的各种服务器。建立该系统以便如果现有的主服务器出现了故障，那么群集器服务器总是与新的引导服务器一致。

[0067] 在所述引导服务器被激活的同时，该引导服务器可以知道需要存在于该系统之中的所有正好一次对象。该引导服务器可决定哪个服务器将接纳每个对象，并且因此可将该对象“出租”给所选择的服务器。当一对象被出租给一服务器时，该服务器可拥有或者接纳该对象某一段时间，例如出租时间段。将该引导服务器配置成定期更新这些出租。该方法可提供一方式以保证如果一服务器有故障或者以任何一种方式而断开了或者否则不能在群集器内正常的操作，那么该服务器不会使其出租被更新。

[0068] 在出现故障的情况下分布式系统所具有的最大问题是很难分辨出具有故障的服务器与仅是未做出响应的一个服务器之间的不同。由于某种原因而与该网络断开的任何服务器不再接纳对象。即使该服务器不可以用于该群集器，但是其仍然知道在出租时间之后它将结束对所有对象的接纳。当该服务器不能被用于该群集器时，其出租将不会被更新。

[0069] 引导服务器还知道，如果引导服务器不能在一定量的时间之内到达主服务器，那么该主服务器将放弃该对象的所有权。出租周期可以是任一适当的时间，例如几秒。出租周期对群集器中的所有对象都相同，或者可在对像之间变化。

[0070] 利用正好一次结构的系统还可以更加紧凑。操作系统经常提供更加接近硬件并可提供更多控制的特定机制。然而，这种方法的一个问题就是它受到可用硬件的限制。例如，服务器的硬件群集器具有按照顺序的 16 个服务器。由于这些系统需要某种紧密的硬件耦合，因此对可以包含在群集器中的服务器数目做出了限制。

[0071] 另一方面，与专用硬件群集器所处理的群集相比，正好一次结构可处理更多的群集。该结构允许从一个专用群集器中所获得的服务质量的某种杠杆作用，从而允许更大的群集器。不同的服务质量可包括例如是否通过诸如点到点连接的可靠协议来传送消息，或者通过诸如多点传送的可靠性稍差但多友好资源的协议来传送消息。利用正好一次结构的优点是能够平衡可量测性与容错性，以便用户可使该系统适应于特定应用的需要。

[0072] 诸如硬件群集器机制的现有系统通过具有（将被提供给所述群集器的）由第二机制所支持的单一机制来尝试高实用性解决方案。如果第一机制失败，则存在执行接管的一个“伙伴”，并且原来运行于第一机制的任何软件被转移到第二机制。

[0073] 根据本发明的正好一次结构可将引导服务器分布给位于这些硬件群集器的一个中的服务器，以便对引导服务器故障进行处理变得比对软件群集器中的故障进行处理更快。然而，该引导服务器将少量的出租发放到服务器，而不管那些服务器是否位于硬件群集器或者软件群集器中。该配置可使引导服务器更快地恢复，虽然允许软件群集器大于硬件群集器，但是软件群集器仍包括硬件群集器。

[0074] 一个这种系统 200 如图 2 所示。硬件群集器 218 包括其包含有多个服务器 220，222，224 的单一机构。例如为提高效率，该硬件群集器可被用于从该机构上的服务器中选择一引导服务器 220。一旦选择了引导服务器 220，该引导服务器就可以在文件系统 212 中选择用于对象 214 的主机 206，该文件系统 212 可能位于软件群集器 210 内部或者外部。文件系统 214 本身可将对象 214 拷贝为位于文件系统的另一磁盘上的第二对象 216，从而提供持久性。新的主机 206 利用高速缓存在所述主机 206 上的所述对象的拷贝 208 从文件系统 212 中取出该对象 214。当经过网络 204 从浏览器或客户机 202 接收到诸如服务 206、216 和 220 的请求时，如果该服务器需要访问对象 208 的高速缓存拷贝，那么该服务器知道要与主服务器 206 连接。

[0075] 利用这种系统的一种方法 500 示于图 5。使用硬件群集器 502 的算法选择引导服务器。该算法例如可以是硬件群集器设备的专用算法，或者可以是其只要求硬件群集器服务器一致的分布式一致性算法。此后利用引导服务器 504 所具有的诸如 Paxos 算法这样的分布式一致性算法来预选择主服务器。此后将所预选择的主机的标识多点传送到位于其包括有硬件群集器的软件群集器中的另一个服务器上。该引导服务器接收来自每个服务器的赞同或者不赞同，这些服务器不久将进行操作并且与群集器 508 相连。如果服务器与所预选择的主服务器一致，那么将一提交信息多点传送到群集器服务器以向该服务器通知所预选择的主机现在接纳该对象；否则，如果服务器不赞同新的主机是预选择的并且再次启动该处理。

[0076] 正好一次结构例如用于对事项记录进行处理或者高速缓存。这种结构例如还可以用于将管理服务器定义为正好一次对象并且出租管理服务器以便该管理服务器决不会停止工作。

[0077] 图 6 给出了根据本发明的群集系统 600 的另一个例子，其中，对象 608 用作与 Java

消息服务 612(JMS) 相关的消息存储器。群集器 610 中的所有服务器 606,614,616 使用 JMS,但是它们必须将消息传送到消息存储器 608 并且通过网络 604 而获得来自该消息存储器 608 的任何消息。群集器 610 中服务器的主服务器 606 将“拥有”或者“接纳”消息存储器 608。客户或者浏览器 602 可以向网络 604 提出直接指向群集器 610 中服务器 616 的一个请求。服务器 616 仅通过经网络 604 将一消息传送到位于主服务器 606 上的消息存储器 608 中来访问 JMS。

[0078] 图 7 给出了可用做本发明部件或者能够实现本发明方法的计算机系统的方框图 700。图 7 的计算机系统包括一处理单元 704 和主存储器 702。处理单元 704 包括单一微处理器,或者可包括多个微处理器以将计算机系统配置为多处理器系统。主存储器 702 部分的存储由处理器单元 704 所执行的指令和数据。如果本发明整个或者部分是以软件实现的,那么当进行操作时主存储器 702 可存储可执行的程序代码。主存储器 702 包括动态随机存取存储器组 (DRAM)、高速超高速缓存器、以及现有技术中的所大家所熟知的其它类型的存储器。

[0079] 图 7 所示的系统进一步包括一大容量存储器 706、一外围设备 708、一用户输入装置 712、一便携式存储介质驱动器 714、一图形子系统 718 以及一输出显示 716。为了简单起见,图 7 所示的部件通过单一总线 720 相连。然而,对于本领域普通技术人员来说显而易见的是,该部件可通过一个或多个数据传送装置相连。例如,处理器单元 704 和主存储器 702 通过本地微处理器总线而连接,并且大容量存储器 706、外围设备 708、便携式存储介质驱动器 714、以及图形子系统 718 通过一个或多个输入 / 输出 (I/O) 总线相连。由一磁盘驱动器、光盘驱动器、以及本领域所公知的其它驱动器所实现的大容量存储器 706 是一非易失性存储设备以用于存储处理器单元 704 所使用的数据和指令。在一个实施例中,大容量存储器 706 存储用于实现本发明的软件以便将其装入主存储器 702 中。

[0080] 便携式存储介质驱动器 714 与诸如软盘的便携式非易失性存储介质相连以便向 / 从图 7 所示的计算机系统输入 / 输出数据和程序代码。在一个实施例中,用于实现本发明的系统软件被存储在这种便携式介质上,并经过该便携式存储介质驱动器 714 输入给所述计算机系统。外围设备 708 可包括诸如输入 / 输出 (I/O) 接口的任何一种将辅助功能添加到所述计算机上的计算机支援设备。例如,外围设备 708 可包括其可使计算机系统与网络相连一网络接口以及诸如调制解调器、路由器、或者本领域所公知的其它硬件这样的其它网络硬件。

[0081] 用户输入装置 712 提供部分用户接口。用户输入装置 712 可包括用于输入阿尔法数字及其它信息的一阿尔法数字小键盘或者诸如鼠标、跟踪笔、触笔、或者光标方向键这样的指示设备。为了显示文本和图形信息,图 7 的计算机系统包括图形子系统 718 以及输出显示器 716。输出显示 716 包括一阴极射线管 (CRT) 显示器、液晶显示器 (LCD) 或者其它合适的显示设备。图形子系统 718 接收文本和图形信息,并且对该信息进行处理以输出到显示 716 器。另外,图 7 的系统包括输出装置 710。合适的输出装置的例子包括扩音器、打印机、网络接口、监视器、及本领域所公知的其它输出装置。

[0082] 一般在适用于本发明某些特定实施例的计算机系统中可找到图 7 的计算机系统所包含的部件,并且其表示本领域所公知的一大类这种计算机元件。因此,图 7 的计算机系统可以是一个人计算机、工作站、服务器、微型计算机、大型计算机、或者任何其它计算机。

计算机系统700还可以采用不同的总线结构、网络平台、多处理器平台等等。可以使用包括Unix、Linux、Windows、Macintosh OS、Palm OS及其它合适的操作系统的任何操作系统。

[0083] 为了说明和描述而提出了本发明的优选实施例。这并不是对所公开的精确结构的详述或者限制。显然，对于本领域普通技术人员来说显而易见的是可做出各种修改和变化。为了更好的说明本发明的原则以及其实际应用，可选择并且描述该实施例，从而使本领域普通技术人员可知道本发明可用于各种实施例并且可做出其适合于特定使用目的的修改。本发明的范围由下述权利要求以及其等效体所规定。

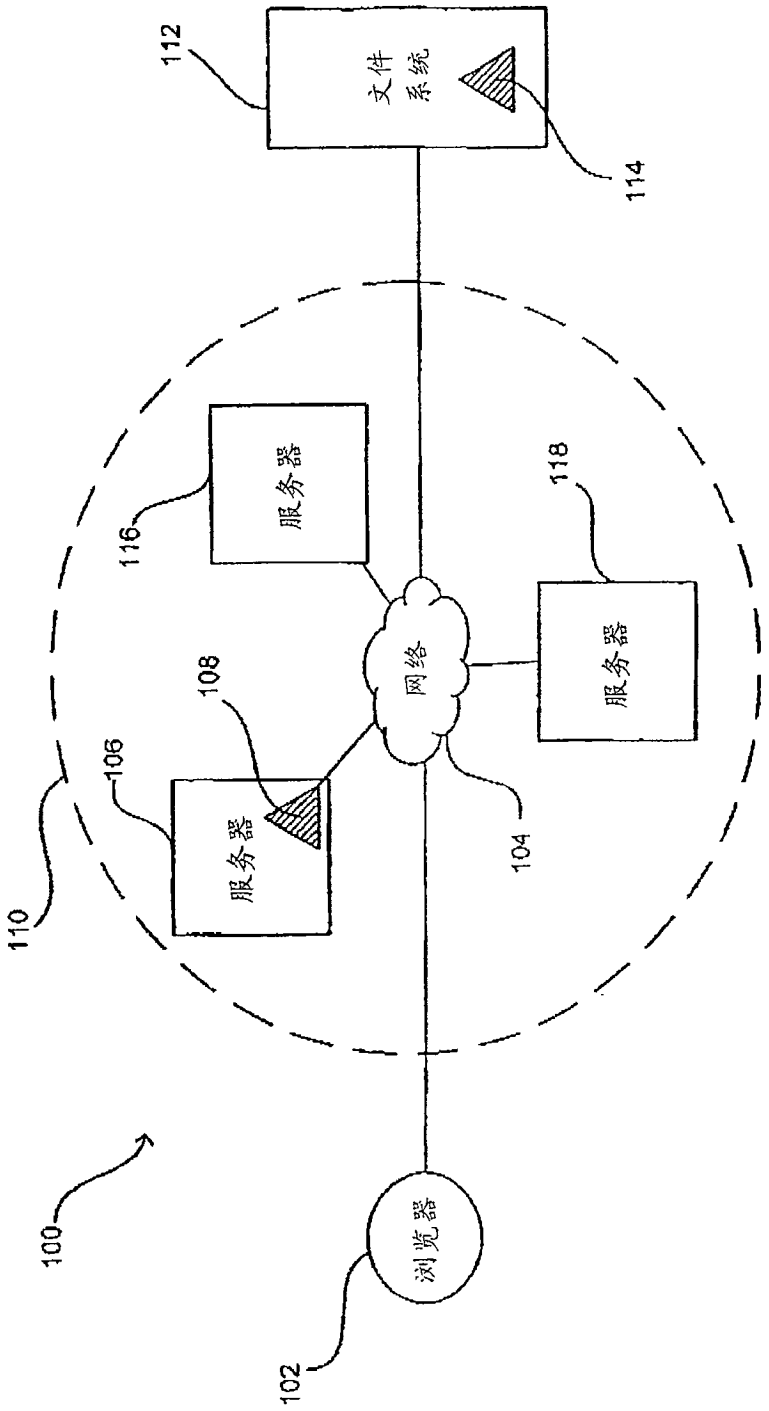


图 1

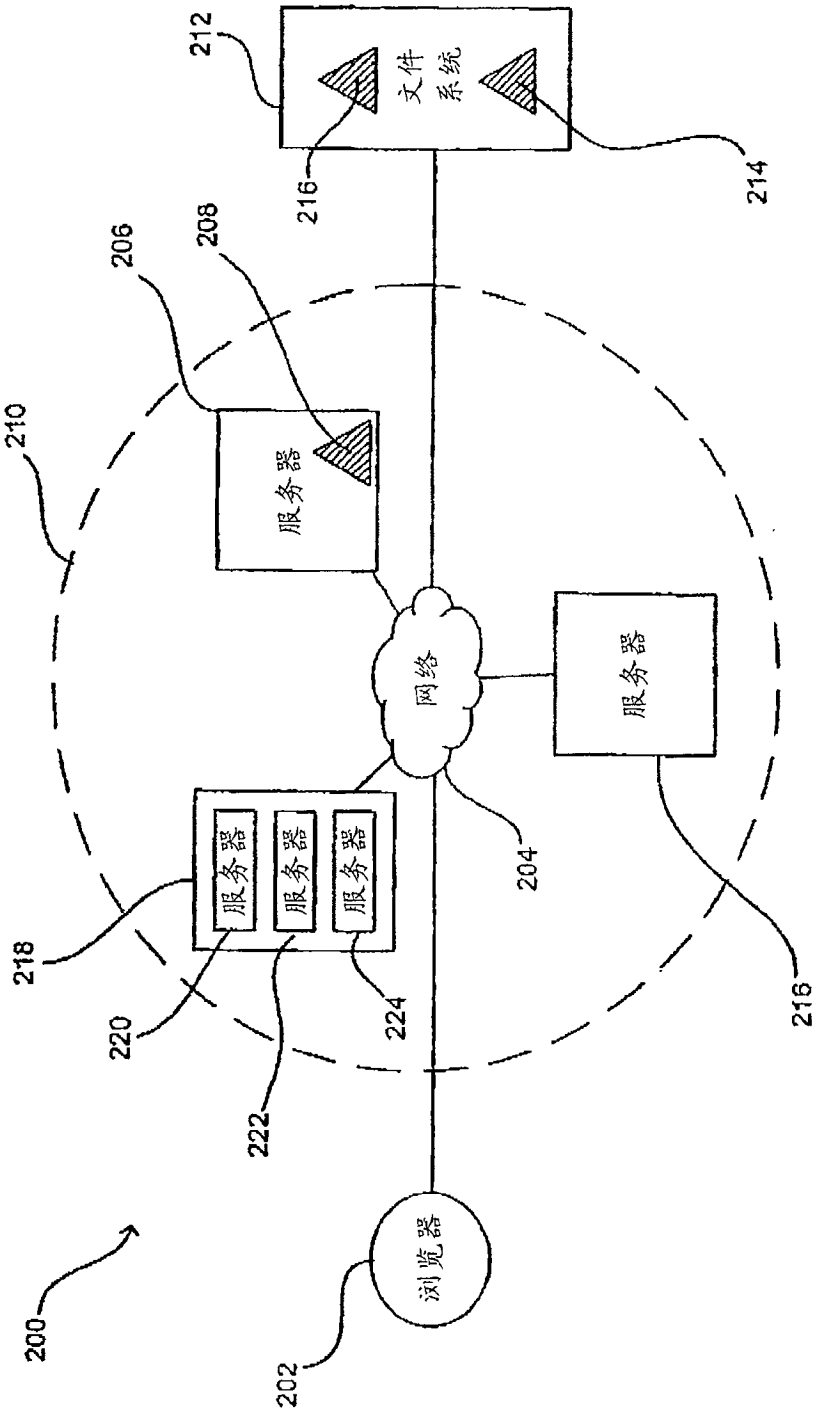

212
文件系统
214
216
208
206
210
服务器
网络
服务器
204
218
服务器
服务器
服务器
220
222
224
216
200
浏览器
202

图 2

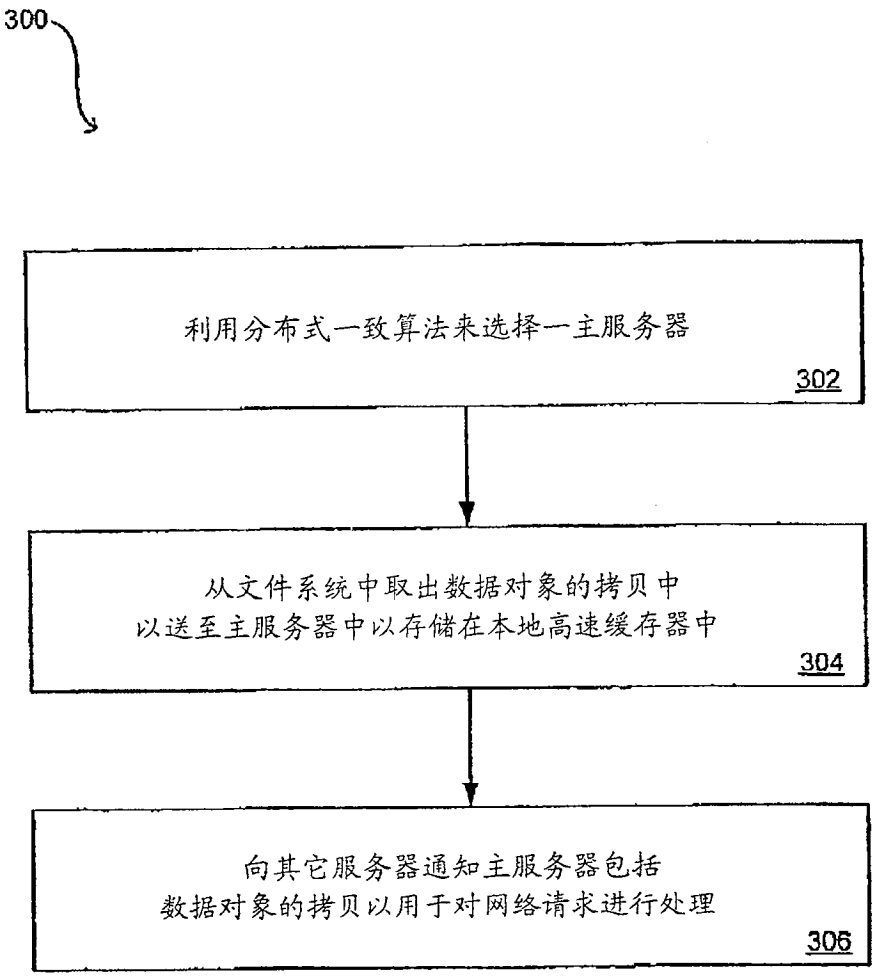

300
利用分布式一致算法来选择一主服务器
302
从文件系统中取出数据对象的拷贝中
以送至主服务器中以存储在本地高速缓存器中
304
向其它服务器通知主服务器包括
数据对象的拷贝以用于对网络请求进行处理
306

图 3

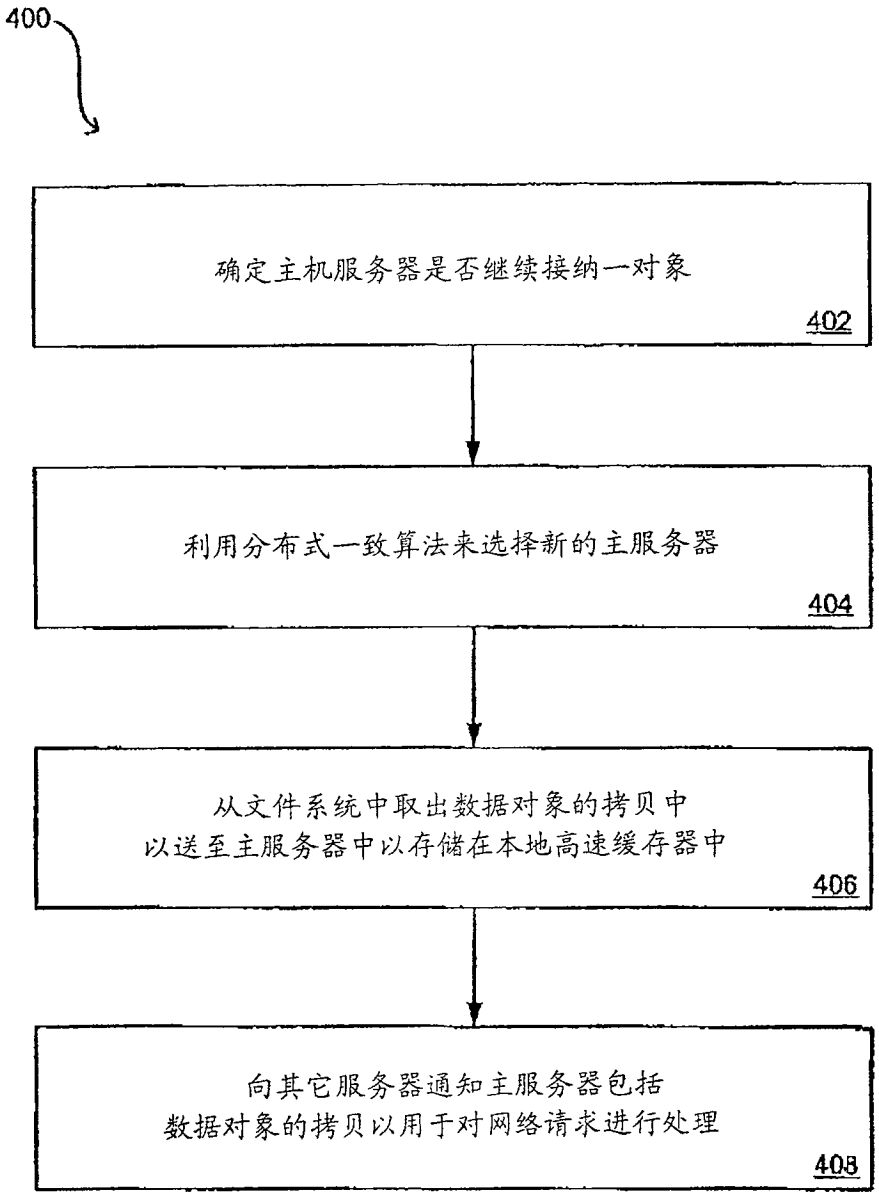

400
确定主机服务器是否继续接纳一对象
402
利用分布式一致算法来选择新的主服务器
404
从文件系统中取出数据对象的拷贝中
以送至主服务器中以存储在本地高速缓存器中
406
向其它服务器通知主服务器包括
数据对象的拷贝以用于对网络请求进行处理
408

图 4

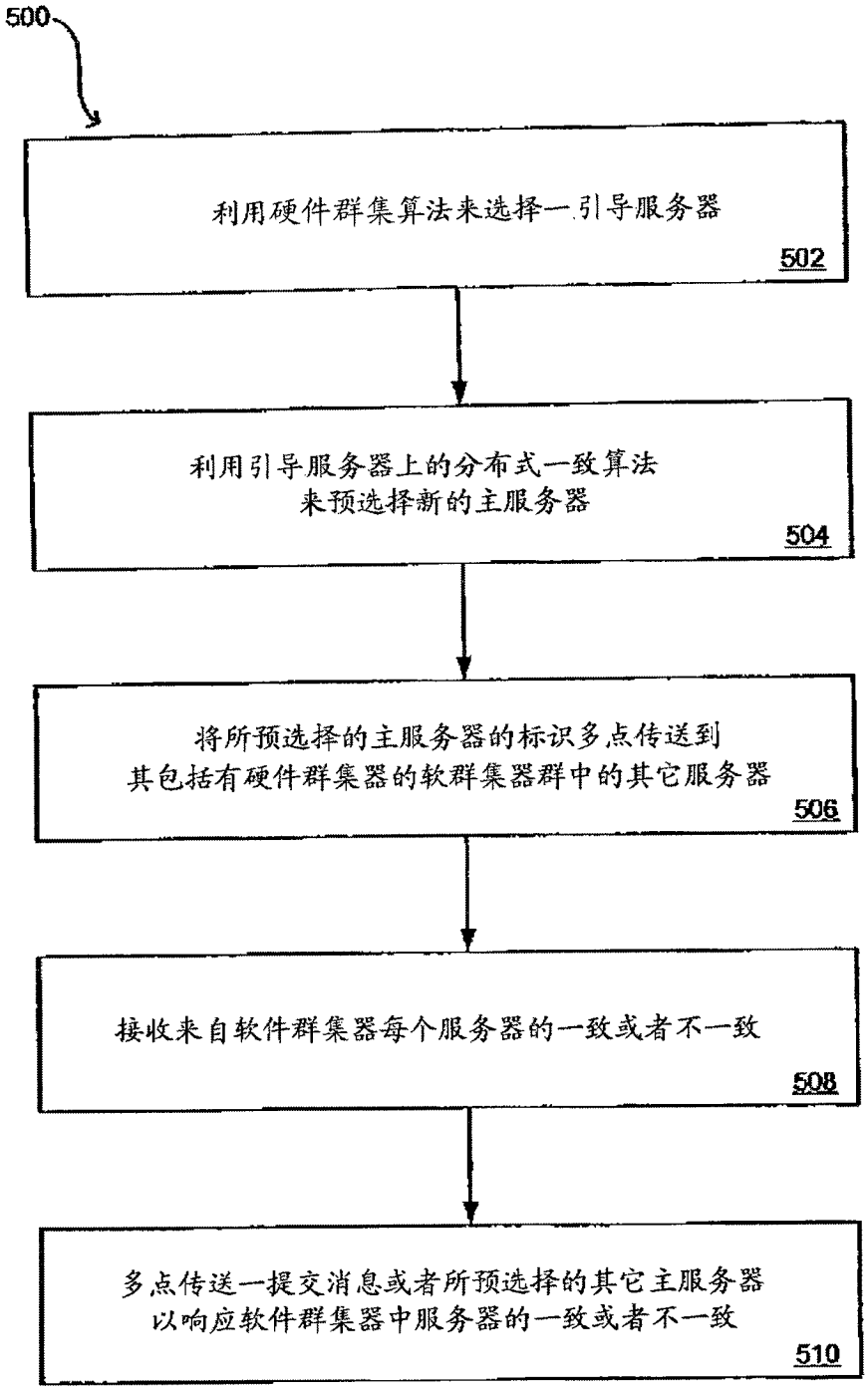

500
利用硬件群集算法来选择一引导服务器
502
利用引导服务器上的分布式一致算法
来预选择新的主服务器
504
将所预选择的主服务器的标识多点传送到
其包括有硬件群集器的软群集器群中的其它服务器
506
接收来自软件群集器每个服务器的一致或者不一致
508
多点传送一提交消息或者所预选择的其它主服务器
以响应软件群集器中服务器的一致或者不一致
510

图 5

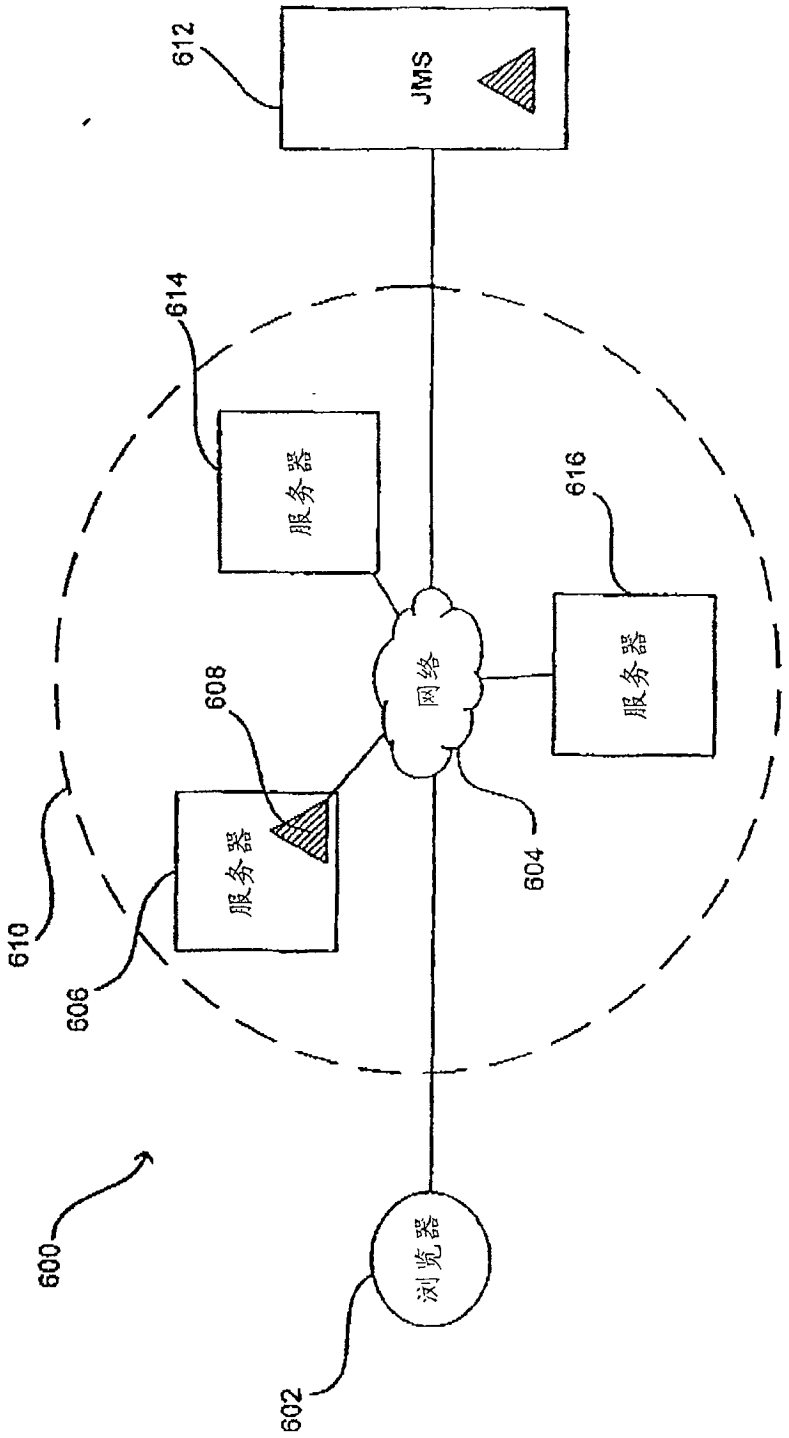


图 6

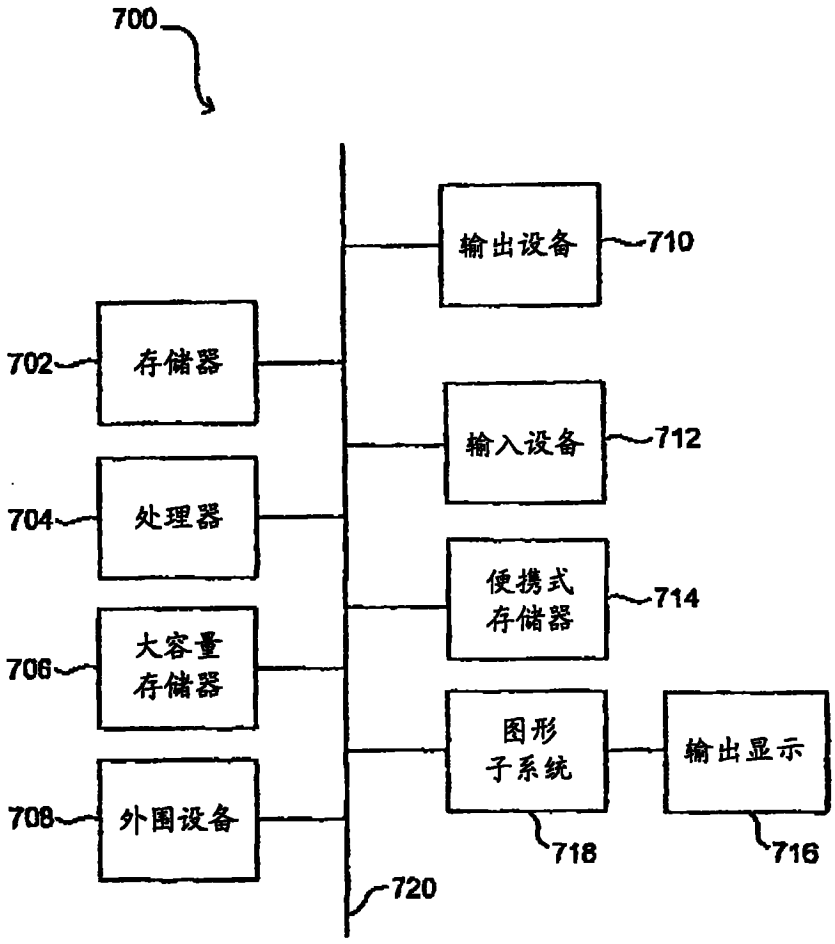

700
输出设备
710
存储器
702
输入设备
712
处理器
704
便携式
存储器
714
大容量
存储器
706
图形
子系统
输出显示
外围设备
708
718
716
720

图 7