



(12)发明专利

(10)授权公告号 CN 103534687 B

(45)授权公告日 2017.04.12

(21)申请号 201280023355.1

C·普尔蒂

(22)申请日 2012.05.15

(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038

(65)同一申请的已公布的文献号

申请公布号 CN 103534687 A

代理人 冯玉清

(43)申请公布日 2014.01.22

(51)Int.Cl.

G06F 9/50(2006.01)

(30)优先权数据

61/486,701 2011.05.16 US

13/239,253 2011.09.21 US

(85)PCT国际申请进入国家阶段日

2013.11.15

(86)PCT国际申请的申请数据

PCT/US2012/037997 2012.05.15

(87)PCT国际申请的公布数据

WO2012/158718 EN 2012.11.22

(73)专利权人 甲骨文国际公司

地址 美国加利福尼亚

(56)对比文件

CN 101442435 A,2009.05.27,

CN 102014137 A,2011.04.13,

US 2008294646 A1,2008.11.27,

Tengjiao Wang 等.Dynamic Data

Migration Policies for Query-Intensive Distributed Data Environments.《Advances in Data and Web Management: Joint International Conferences, APWeb/WAIM 2009, Suzhou, China, April 2-4, 2009, Proceedings》.2009,第63-75页.

审查员 郭妍捷

(72)发明人 R·H·李 G·格莱泽 M·法尔科

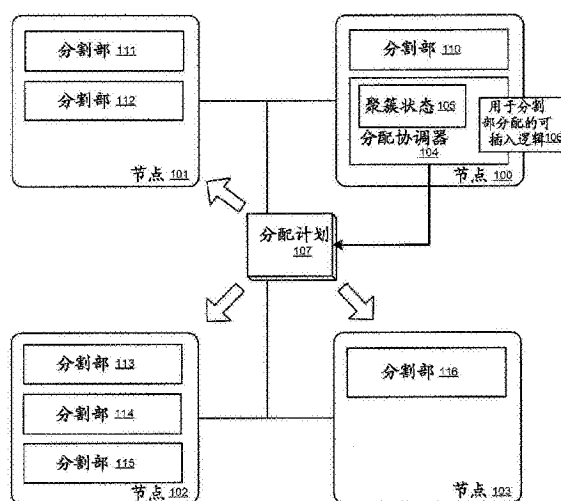
权利要求书4页 说明书9页 附图5页

(54)发明名称

聚簇数据格网中的可扩展集中动态资源分配

(57)摘要

描述了一种集中式资源分配,其中在聚簇节点之间分割数据的决定部分被集中,而实施分割的实际机构保留分布式算法。使用中央分配协调器来产生可扩展中央策略,其控制数据将如何在聚簇上被分割。根据分布式算法,实施该策略的工作由所有成员独立且非同步地执行。中央策略可传输到所有聚簇成员,每个成员能执行分割,只要涉及其本身。例如,根据分布式算法,一个节点可根据中央策略而决定它需要获得特定分割部,并且执行不要的步骤以获得该数据,而另一些节点可以非同步地执行与那些特定节点相关的其他个体分割部的转移。



1. 一种计算机实施的用于提供可扩展集中式动态资源分配的方法,该方法包括:
 储存数据集的多个分割部,所述多个分割部被分配在计算机节点的聚簇上;
 指定所述计算机节点中的一个节点作为中央分配协调器,该中央分配协调器收集指示所述聚簇的全局状态的信息并且提供对所述全局状态的访问;
 通过所述分配协调器周期性地或者响应于聚簇中的成员变动分析所述聚簇的全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变;
 通过所述分配协调器基于对所述分割部的分配的改变产生新分配计划,并且向所述计算机节点的聚簇中的所有计算机节点提供对所述分配计划的访问;以及
 通过如下方式由所述聚簇中的每个计算机节点独立地确定是否执行与所述计算机节点相关联的各个分割部的转移以实施所述新分配计划:
 检查所述新分配计划,并且确定所述新分配计划是否指定了与所述计算机节点相关的改变,并且如果所述新分配计划指定了与所述计算机节点相关的改变,则执行以下之一或两者:
 开始从其它计算机节点到所述计算机节点的分割部的转移,或者
 开始从所述计算机节点到其它计算机节点的分割部的转移。
2. 根据权利要求1的方法,其中通过所述分配协调器周期性地或者响应于聚簇中的成员变动分析所述全局状态以确定是否应对所述分割部的分配进行改变进一步包括:
 所述分配协调器调用可插入分配逻辑模块,其中所述可插入分配逻辑模块能在运行时被切换以调节用于在所述计算机节点的聚簇中分配所述分割部的算法。
3. 根据权利要求1或2的方法,其中采用分布式算法进一步包括:
 在所述计算机节点的聚簇中的两个计算机节点之间点对点地执行各个分割部的转移,所述转移是发送节点和接收节点之间的直接非同步通信的结果,其中所述分配协调器不参与指引所述分割部的转移。
4. 根据权利要求1或2的方法,其中所述聚簇的全局状态包括指示所述分割部中的哪些被分配给所述聚簇中的各个计算机节点的信息。
5. 根据权利要求1或2的方法,其中所述聚簇的全局状态包括指示所述聚簇中的每个计算机节点上的当前处理负载的信息,其中所述当前处理负载通过所述每个计算机节点周期性地发送运行时反馈统计数据到所述分配协调器来确定。
6. 根据权利要求1或2的方法,其中所述聚簇的全局状态包括指示所述聚簇中的每个计算机节点的存储器能力和处理器能力的信息。
7. 根据权利要求1或2的方法,其中所述多个分割部进一步包括一组主分割部和一组备份分割部,其中所述分配协调器确保每个主分割部位于与该主分割部所关联的备份分割部不同的物理节点上。
8. 根据权利要求7的方法,其中所述分配协调器产生的新分配计划实施下列限制:
 对于每个给定计算机节点,允许有限数量的计算机节点包含与位于所述给定计算机节点上的主分割部相关联的备份分割部。
9. 根据权利要求1或2的方法,其中所述分配协调器为所述分割部在所述聚簇的计算机节点中的分配提供单个协调点。
10. 根据权利要求1或2的方法,其中所述新分配计划指明特定分割部应位于所述聚簇

中的指定计算机节点上。

11. 一种用于提供可扩展集中式动态资源分配的装置, 包括一个或多个计算机, 所述一个或多个计算机配置成执行权利要求1至10中的任一项的方法的所有步骤。

12. 一种用于提供可扩展集中式动态资源分配的系统, 该系统包括:

计算机节点的聚簇, 储存数据集的多个分割部, 所述分割部分配在所述计算机节点的聚簇上; 以及

选自所述计算机节点的分配协调器, 所述分配协调器收集指示所述聚簇的全局状态的信息, 周期性地或者响应于聚簇中的成员变动分析所述全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变, 根据所述改变产生新分配计划, 以及向所述计算机节点的聚簇中的所有计算机节点提供对所述分配计划的访问;

其中所述聚簇中的每个计算机节点通过如下方式独立地确定是否执行与所述计算机节点相关联的各个分割部的转移以实施由所述分配协调器产生的所述新分配计划:

检查所述新分配计划, 并且确定所述新分配计划是否指定了与所述计算机节点相关的改变, 并且如果所述新分配计划指定了与所述计算机节点相关的改变, 则执行以下之一或两者:

开始从其它计算机节点到所述计算机节点的分割部的转移, 或者

开始从所述计算机节点到其它计算机节点的分割部的转移。

13. 一种用于提供可扩展集中式动态资源分配的方法, 该方法包括:

储存数据集的多个分割部, 所述多个分割部被分配在计算机节点的聚簇上;

由分配协调器周期性地或者响应于聚簇中的成员变动分析所述聚簇的全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变, 所述分配协调器是所述计算机节点之一;

由所述分配协调器基于对所述分割部的分配的改变产生新分配计划, 并且向所述计算机节点的聚簇中的所有计算机节点提供对所述分配计划的访问; 以及

通过如下方式由所述聚簇中的每个计算机节点独立地确定是否执行与所述计算机节点相关联的各个分割部的转移以实施所述新分配计划:

检查所述新分配计划, 并且确定所述新分配计划是否指定了与所述计算机节点相关的改变, 并且如果所述新分配计划指定了与所述计算机节点相关的改变, 则执行以下之一或两者:

开始从其它计算机节点到所述计算机节点的分割部的转移, 或者

开始从所述计算机节点到其它计算机节点的分割部的转移。

14. 一种用于提供可扩展集中式动态资源分配的设备, 包括:

用于储存数据集的多个分割部的装置, 所述多个分割部被分配在计算机节点的聚簇上;

用于由分配协调器周期性地或者响应于聚簇中的成员变动分析所述聚簇的全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变的装置, 所述分配协调器是所述计算机节点之一;

用于由所述分配协调器基于对所述分割部的分配的改变产生新分配计划, 并且向所述计算机节点的聚簇中的所有计算机节点提供对所述分配计划的访问的装置; 以及

用于通过如下方式由所述聚簇中的每个计算机节点独立地确定是否执行与所述计算机节点相关联的各个分割部的转移以实施所述新分配计划的装置：

检查所述新分配计划，并且确定所述新分配计划是否指定了与所述计算机节点相关的改变，并且如果所述新分配计划指定了与所述计算机节点相关的改变，则执行以下之一或两者：

开始从其它计算机节点到所述计算机节点的分割部的转移，或者

开始从所述计算机节点到其它计算机节点的分割部的转移。

15. 一种用于提供可扩展集中式动态资源分配的设备，包括：

用于储存数据集的多个分割部的装置，所述多个分割部被分配在计算机节点的聚簇上；

用于指定所述计算机节点中的一个节点作为中央分配协调器的装置，该中央分配协调器收集指示所述聚簇的全局状态的信息并且提供对所述全局状态的访问；

用于通过所述分配协调器周期性地或者响应于聚簇中的成员变动分析所述聚簇的全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变的装置；

用于通过所述分配协调器基于对所述分割部的分配的改变产生新分配计划，并且向所述计算机节点的聚簇中的所有计算机节点提供对所述分配计划的访问的装置；以及

用于通过如下方式由所述聚簇中的每个计算机节点独立地确定是否执行与所述计算机节点相关联的各个分割部的转移以实施所述新分配计划的装置：

检查所述新分配计划，并且确定所述新分配计划是否指定了与所述计算机节点相关的改变，并且如果所述新分配计划指定了与所述计算机节点相关的改变，则执行以下之一或两者：

开始从其它计算机节点到所述计算机节点的分割部的转移，或者

开始从所述计算机节点到其它计算机节点的分割部的转移。

16. 根据权利要求15的设备，其中用于通过所述分配协调器周期性地或者响应于聚簇中的成员变动分析所述全局状态以确定是否应对所述分割部的分配进行改变进一步包括：

用于通过所述分配协调器调用可插入分配逻辑模块的装置，其中所述可插入分配逻辑模块能在运行时被切换以调节用于在所述计算机节点的聚簇中分配所述分割部的算法。

17. 根据权利要求15或16的设备，其中采用分布式算法进一步包括：

在所述计算机节点的聚簇中的两个计算机节点之间点对点地执行各个分割部的转移，所述转移是发送节点和接收节点之间的直接非同步通信的结果，其中所述分配协调器不参与指引所述分割部的转移。

18. 根据权利要求15或16的设备，其中所述聚簇的全局状态包括指示所述分割部中的哪些被分配给所述聚簇中的各个计算机节点的信息。

19. 根据权利要求15或16的设备，其中所述聚簇的全局状态包括指示所述聚簇中的每个计算机节点上的当前处理负载的信息，其中所述当前处理负载通过所述每个计算机节点周期性地发送运行时反馈统计数据到所述分配协调器来确定。

20. 根据权利要求15或16的设备，其中所述聚簇的全局状态包括指示所述聚簇中的每个计算机节点的存储器能力和处理器能力的信息。

21. 根据权利要求15或16的设备，其中所述多个分割部进一步包括一组主分割部和一

组备份分割部,其中所述分配协调器确保每个主分割部位于与该主分割部所关联的备份分割部不同的物理节点上。

22.根据权利要求21的设备,其中所述分配协调器产生的新分配计划实施下列限制:

对于每个给定计算机节点,允许有限数量的计算机节点包含与位于所述给定计算机节点上的主分割部相关联的备份分割部。

23.根据权利要求15或16的设备,其中所述分配协调器为所述分割部在所述聚簇的计算机节点中的分配提供单个协调点。

24.根据权利要求15或16的设备,其中所述新分配计划指明特定分割部应位于所述聚簇中的指定计算机节点上。

聚簇数据格网中的可扩展集中动态资源分配

[0001] 版权声明

[0002] 本专利文献公开的一部分包含受到版权保护的内容。版权所有者不反对任何人对本专利文献或专利公开进行传真复印,只要它出现在专商局专利文件或记录中,但是在其他情况下,保留任何所有版权。

技术领域

[0003] 本发明涉及分布式计算环境中的数据高速缓存技术,更特别地,涉及在聚簇网络的成员之间分割数据。

背景技术

[0004] 分布式计算和分布式算法已经由于增大的性能和承载能力、高可用性和失效转移(failover)、以及更快的数据访问而广泛流行于各种情境中。分布式计算一般涉及彼此通信以解决任务(诸如执行应用,解决复杂的计算问题,或者向用户提供到各种服务的接入)的多个自治计算机(也称为节点)。每个计算机节点一般包括其自身的一个或多个处理器、存储器、以及到其他节点的通信链路。计算机可以位于特定位置(例如,聚簇网络)中,或者可以通过大面积网络(LAN)诸如因特网而被连接。在很多情况中,分布式计算机使用消息来彼此通信以及协调任务处理和数据管理。

[0005] 数据管理是分布式计算中的重要问题。在聚簇网络的情境中,大数据集可以在聚簇的各个节点之间分割。每个节点通常储存多个这样的分割部(整个数据集的子部(subpart))并且对分割部执行事务。在许多情况下,分割部包括在成员之间分配的数据的主拷贝和备份拷贝以用于失效转移。数据以此分割方式的分配能改善易管理性、性能和信息可用性。

[0006] 数据在该情境下的理想分配受到许多限制和问题的影响。例如,将数据从一个服务器移动到另一服务器需要耗费时间和/或处理器能力。出于高可用性的原因,通常优选将数据的主拷贝和备份拷贝定位在物理不同的机器上。此外,出于性能、扩缩性和容量的原因,通常优选在可用的储存服务器之间在一定程度上均等地平衡数据分配,并且当从聚簇增加或去除节点时调整该分配。

[0007] 在一些使用情况下,可能期望进一步的应用特定的偏好。例如,指定特定数据集应位于特定服务器上在某些情况下可能是有用的。此外,可能期望指定分配应使用运行时(runtime)反馈和输入以将数据安置于节点成员中。根据上述内容,需要一种简单有效的方式来解决全部或许多数据分配问题并且在分布式聚簇成员之间优化分割部的平衡。

发明内容

[0008] 根据本发明的各种实施例,在聚簇节点之间分割数据的决策部分可以被集中,而实施分割的实际机构可保留分布式算法。中央分配协调器可执行集中逻辑或算法(策略),其生成分配计划。分配计划控制跨整个聚簇将分割多少数据。根据分布式算法,实施该分配

计划的工作由所有成员独立且非同步地执行。分配计划可以传输到聚簇的所有成员,然后每个成员执行仅涉及其本身的分割。例如,根据分布式算法,一个节点可根据分配计划确定它需要获得特定分割部,并且执行必要的步骤来获得该数据,而另一些节点可以非同步地执行与那些特定节点相关的其他单独分割部的传输。以此方式,可防止单个协同点导致的瓶颈,同时实现数据分配的集中管理和控制。

附图说明

- [0009] 图1是根据本发明各种实施例的在聚簇中实施的可扩展集中资源分配的图示。
- [0010] 图2是根据本发明各种实施例的由聚簇中的节点实施的分配计划的图示。
- [0011] 图3是根据本发明各种实施例的提供集中资源分配的过程的流程图。
- [0012] 图4是根据本发明各种实施例的由分配协调器执行的过程的流程图。
- [0013] 图5是根据本发明各种实施例的由聚簇中的节点执行的过程的流程图。

具体实施方式

[0014] 聚簇数据格网的功能之一在于维护和实现数据储存分割部在各聚簇成员(节点)之间的分配。在此情境下,分配可视为数据的主拷贝和备份拷贝二者在提供数据储存的给定聚簇成员服务器上的安置。

[0015] 跨聚簇实施数据分割的一种方式是通过实施分布式算法,其中每个储存服务器独立地确定是否将其拥有的分割部移动到另一服务器或者从另一服务器获得分割部。作为示例,聚簇的数据集可最初包括256个主分割部(具有对应的256个备份分割部),其最终跨两个聚簇成员(节点)分配,每个成员储存128个主分割部和128个备份分割部。如果第三成员加入到聚簇,则成员将在彼此之间重新分配分割部,使得每个节点将储存85个分割部,节点之一储存86个。例如,新加入的成员节点能请求聚簇中的另外两个节点中的每一个传输43个分割部到新节点。替选地,另外两个节点可以确定新成员已经加入到聚簇并且独立地传输分割部到新节点。

[0016] 这种使用分布式算法来分派分割部的自治且分离的方案提供许多优点,包括更好的扩缩性,消除单个失效点等。然而,也变得更难以实施复杂的分配策略,因为成员节点关于分割部转移作出独立决定。例如,可能期望根据更加动态和复杂的因素来分配分割部,诸如特定分割部可能会被多频繁地访问,特定成员节点上的当前处理负载和每个节点的CPU/存储器能力。此外,可能出于高可用性和失效转移的原因而期望以更复杂和优化的方式来协调主分割部和备份分割部的布置。

[0017] 根据各种实施例,描述一种方案,其中各个分割部的转移被点对点地执行(作为发送节点和接收节点之间的直接但非同步通信的结果),但是允许单个协调点(分配协调器)。该单个协调点提供聚簇的全局状态,包括当前分配(数据拥有者)以及动态运行时反馈,诸如在给定时间点处聚簇中的每个节点上的处理负载。在单个点处具有系统的全局概观的能力允许比分布式算法的比较方案远更具表达性的分配逻辑。

[0018] 根据一实施例,聚簇中的每个服务器节点不经常地将分配状态传输到分配协调器。分配协调器收集该信息并且周期性地(或者作为成员变动的结果)调用所配置的分配策略来分析该分配。此外,可以提供到可插入逻辑(策略)的接口,其提供到当前分配状态的策

略访问,且被策略用于建议分割部在可用服务器节点之间的新分配(布置)。

[0019] 根据一实施例,分配协调器将分配策略作出的建议收集到总分配计划或目标中。计划被传输到聚簇中的所有节点。之后,每个服务器开始各个分割部传输(数据移动)以实现新分配计划。

[0020] 根据各种实施例,可扩展集中分配还能形成各种附加特征的基础。例如,适应性分配允许聚簇动态适应应用负载或数据分配的波动。此外,集中分配能对于WAN安全聚簇是有用的。

[0021] 图1是根据本发明各种实施例的在聚簇中实施可扩展集中资源分配的图示。尽管该图描绘了逻辑分离的组件,但是这种绘示仅用于说明。本领域技术人员将意识到,此图中描绘的组件可以组合或划分为单独的软件、固件和/或硬件。此外,本领域技术人员将意识到,这些组件能在同一计算设备上运行或者能分布在通过一个或更多网络或其他通信手段连接的不同计算设备之间,而与这些组件如何组合或划分无关。

[0022] 如图所示,聚簇包括多个节点(100、101、102、103),其储存遍布其分配的数据分割部(110、111、112、113、114、115、116)。聚簇中的节点之一100被指定为中央分配协调器104。根据一实施例,分配协调器周期性地基于多个因素重新评估分割部跨聚簇的分配(布置),并且确定是否需要对其进行改变。例如,分配协调器能根据每个节点上的当前请求负载,特定分割部可能被多频繁地访问和/或节点的CPU和存储器能力来评估分割部安置。类似地,中央分配协调器104能用于执行更复杂的主分割部相对于备份分割部的分配。例如,对于特定节点上的所有主分割部,分配协调器能确保与那些主分割部相关联的备份分割部不分配在大量节点上。通常优选的是,为了失效转移,备份分割部置于和与之相关联的主分割部不同的物理节点上。然而,当主分割部被更新时,其备份分割部也需要更新。这能导致显著数量的节点跳跃(jump),增加了聚簇中的网络流量。因此,有利的是限制备份分割部位于其上的节点数。中央分配协调器能强制执行该策略以及任何其他分割部布置策略。

[0023] 根据一实施例,分配协调器104维护全局聚簇状态105的概观(view)。全局状态能包括如下信息,包括(但不限于)分割部在节点中的位置,每个节点上的处理负载,对每个分割部中储存的数据的可能需求,每个节点的CPU和/或存储器能力等。根据一实施例,分配协调器采用全局状态来周期性地(或者响应于节点成员变动)重新评估聚簇中的分割部分配。应注意,全局聚簇状态无需一定储存在分配协调器上,能替代地储存在聚簇的其他成员上以及远程储存在其他机器上。

[0024] 根据一实施例,分配协调器104能调用可插入逻辑组件106以评估分割部分配。可插入逻辑器106能指定应用于聚簇的特定的定制分配策略。分配协调器能根据该策略评估分割部分配并且确定是否需要对其进行改变。

[0025] 如果分配协调器104确定应对分割部分配作出改变,则它能产生分配计划107。该分配计划107然后能用于聚簇中的每个节点。根据一实施例,分配计划107能指定哪些分割部应位于哪个节点。一旦新分配计划可用,各节点能以分布式方式进行适当的分割部转移,如下面将描述的那样。

[0026] 图2是根据本发明各种实施例的由聚簇中的节点实施的分配计划的图示。尽管该图描绘了逻辑分离的组件,但是这种绘示仅用于说明。本领域技术人员将意识到,此图中描绘的组件可以组合或划分为单独的软件、固件和/或硬件。此外,本领域技术人员将意识到,

这些组件能在同一计算设备上运行或者能分布在通过一个或更多网络或其他通信手段连接的不同计算设备之间,而与这些组件如何组合或划分无关。

[0027] 根据所示实施例,一旦分配协调器产生分配计划,它就能使分配计划可用于聚簇中的每个节点。然后节点可以根据分配计划独立执行转移分割部到正确节点所需的决定和步骤。以此方式,分配分割部的机构保留分布式算法,而分割部布置的决定和策略被集中作出。

[0028] 如图所示,新分配计划107指定节点100应储存分割部110和112;节点101应储存分割部111;节点102应储存分割部113和114;节点103应储存分割部115和116。因为分割部112当前位于节点101上,所以节点100能检查新分配计划并从节点101请求分割部112。类似地,节点103能从节点102请求分割部115。根据备选实施例,节点101和102能在接收分配计划之后传输必要的分割部到节点100和103,而无需等候来自节点100和103的任何请求。

[0029] 图3是根据本发明各种实施例的提供集中资源分配的过程的流程图。尽管该图为了说明而以特定顺序描绘了功能步骤,但是该过程不一定局限于该特定步骤顺序。本领域技术人员将意识到,此图描绘的各种步骤可以被改变,重新布置,并行执行或以各种方式变化。此外将理解,某些步骤或步骤序列能被添加或从该过程省略,而不偏离本发明的思想和范围。

[0030] 如步骤300所示,计算机节点的聚簇将数据集储存为多个分割部。这些分割部分布在聚簇中的节点上。在步骤301,计算机节点之一被指定为中央分配协调器。根据一实施例,分配协调器收集指示聚簇的全局状态的数据,并且提供对全局状态的访问。如步骤302进一步所示,分配协调器周期性地分析聚簇的全局状态并且确定是否需要对分割部在聚簇的节点中的分配进行改变。替选地,分配协调器可响应于聚簇中的成员变动(而不是周期性地)来重新评估分割部分配。

[0031] 如果分配协调器确定应进行改变,则它基于那些改变产生新的分配计划,并向聚簇中的所有节点提供对分配计划的访问,如步骤303所示。然后可以利用分布式算法进行各个分割部传输,如步骤304所示。换言之,每个节点能独立确定如何进行涉及其的各个分割部转移,以最优地实施新分配计划。

[0032] 图4是根据本发明各种实施例的由分配协调器执行的过程的流程图。尽管该图为了说明而以特定顺序描绘了功能步骤,但是该过程不一定局限于该特定步骤顺序。本领域技术人员将意识到,此图描绘的各种步骤可以被改变,重新布置,并行执行或以各种方式变化。此外将理解,某些步骤或步骤序列能被添加或从该过程省略,而不偏离本发明的思想和范围。

[0033] 根据所示实施例,过程始于步骤400。一旦开始,分配协调器就从聚簇中的每个节点持续接收运行时反馈和其他数据以汇编聚簇的全局状态(步骤401)。分配协调器能使用该全局状态来周期性地分析分割部在聚簇中的分配,如步骤402所示。如果无需改变(步骤403),则分配协调器能不执行动作,直到下一次需要评估分割部分配。另一方面,如果分配协调器确定应对分配进行改变,那么它能生成包括那些改变的新分配计划,如步骤404所示。在步骤405,分配协调器能向聚簇中的所有节点提供分配计划。

[0034] 图5是根据本发明各种实施例的由聚簇中的节点执行的过程的流程图。尽管该图为了说明而以特定顺序描绘了功能步骤,但是该过程不一定局限于该特定步骤顺序。本领域

域技术人员将意识到,此图描绘的各种步骤可以被改变,重新布置,并行执行或以各种方式变化。此外将理解,某些步骤或步骤序列能被添加或从该过程省略,而不偏离本发明的思想和范围。

[0035] 根据所示实施例,过程始于步骤500。一旦开始,每个节点能周期性地发送运行时反馈、负载统计和其他数据到分配协调器,如步骤501所示。步骤501还包括回到其本身的箭头,表示没有新计划生成的可能情况,成员简单地继续运行,周期性地收集和发送统计数据。

[0036] 在步骤502,节点可以从分配协调器接收新分配计划。此时,节点可检查分配计划并确定该计划是否指定与该特定节点相关的改变(步骤503)。如果新计划不涉及该节点,则节点能不执行转移,并且能继续周期性地发送运行时反馈到分配协调器。另一方面,如果分配计划包括涉及该节点的分割部变动,则节点能根据分配计划从其他节点获得必要的分割部和/或提供必要的分割部到其他节点(步骤504)。

[0037] 贯穿本公开描述的各种情境,本发明的实施例还涵盖计算机装置、计算系统和机器可读介质,其配置为执行前述系统和方法。除了包括具体设计的集成电路或其他电子电路的实施例之外,本发明可以利用常规通用或专用数字计算机或微处理器来便捷地实施,数字计算机或微处理器被根据本公开的技术编程,如计算机领域的技术人员将要意识到的那样。

[0038] 总体上,本发明涉及提供可扩展集中动态资源分配的系统,该系统包括:用于储存分配在计算机节点的聚簇中的数据集的多个分割部的装置;用于指定所述计算机节点之一作为中央分配协调器的装置,该中央分配协调器收集指示所述聚簇的全局状态的信息并且提供对所述全局状态的访问;用于通过所述分配协调器来周期性分析所述聚簇的全局状态以确定是否应对所述分割部在所述计算机节点中的分配作出改变的装置;用于通过所述分配协调器基于对所述分割部的分配的改变产生新分配计划并且向所述计算机节点的聚簇中的所有计算机节点提供对所述分配计划的访问的装置;以及用于采用分布式算法由聚簇中的每个节点独立地确定如何执行与所述节点相关联的分割部的转移以实施所述新分配计划的装置。

[0039] 在所述系统中,所述用于通过所述分配协调器周期性分析所述聚簇的全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变的装置进一步包括:用于由所述分配协调器调用可插入分配逻辑模块的装置,其中所述可插入分配逻辑模块能在运行时切换以调节用于在所述计算机节点的聚簇中分配所述分割部的算法。

[0040] 在所述系统中,所述用于采用分布式算法的装置进一步包括:用于在两个计算机节点之间点对点地执行各个分割部的转移的装置,所述转移是发送节点和接收节点之间的直接非同步通信的结果,其中所述分配协调器不参与指引所述分割部转移。

[0041] 在所述系统中,所述聚簇的全局状态包括指示所述分割部中的哪些被分配给所述聚簇中的各个计算机节点的信息。

[0042] 在所述系统中,所述聚簇的全局状态包括指示聚簇中的每个计算机节点上的当前处理负载的信息,其中所述当前处理负载通过所述每个计算机节点周期性发送运行时反馈统计数据到所述分配协调器来确定。

[0043] 在所述系统中,所述聚簇的全局状态包括指示聚簇中的每个计算机节点的存储器

能力和处理器能力的信息。

[0044] 在所述系统中,所述多个分割部进一步包括一组主分割部和一组备份分割部,其中所述分配协调器确保每个主分割部位于与该主分割部所关联的备份分割部不同的物理节点上。

[0045] 在所述系统中,所述分配协调器产生的新分配计划实施下列限制:对于每个给定计算机节点,限制允许包含位于所述给定计算机节点上的主分割部所关联的备份分割部的计算机节点的数量。

[0046] 在所述系统中,所述分配协调器为所述分割部在所述聚簇中的计算机节点中的分配提供单个协调点。

[0047] 在所述系统中,所述新分配计划指明特定分割部应位于聚簇中的指定计算机节点上。

[0048] 有经验的编程人员可以根据本公开的教导容易地准备适当的软件代码,如软件领域的技术人员将要意识到的那样。本发明还可以通过准备特定用途集成电路或者通过互连常规组件电路的适当网络来实施,如本领域技术人员将容易理解的那样。

[0049] 各种实施例包括计算机程序产品,其是其上/其中储存有指令的储存介质,该指令可用于编程通用或专用计算处理器/设备以执行这里给出的任何特征。储存介质可包括但不限于下列中的一个或多个:任何类型的物理介质,包括软盘、光盘、DVD、CD-ROM、微驱动器、磁光盘、全息储存器、ROM、RAM、PRAMS、EPROM、EEPROM、DRAM、VRAM、闪存器、磁卡或光学卡、纳米系统(包括分子存储器IC);纸张或基于纸张的介质;以及适于储存指令和/或信息的任何类型的介质或设备。计算机程序产品可整体或部分地通过一个或多个公共和/或私有网络来传输,其中传输包括能由一个或多个处理器用于执行这里给出的任何特征的指令。传输可包括多个单独的传输。然而,根据某些实施例,含有指令的计算机储存介质不是临时性的(即,不是在正被传输的过程中),而是在物理设备上持续。

[0050] 1.一种提供可扩展集中式动态资源分配的方法,该方法包括:

[0051] 储存数据集的多个分割部,所述多个分割部分配在计算机节点的聚簇上;

[0052] 指定所述计算机节点中的一个节点作为中央分配协调器,该中央分配协调器收集指示所述聚簇的全局状态的信息并且提供对所述全局状态的访问;

[0053] 所述分配协调器周期性地分析所述聚簇的全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变;

[0054] 所述分配协调器基于对所述分割部的分配的改变产生新分配计划,并且向所述计算机节点的聚簇中的所有计算机节点提供对所述分配计划的访问;以及

[0055] 采用分布式算法来由所述聚簇中的每个节点独立地确定如何执行与所述节点相关联的各个分割部的转移以实施所述新分配计划。

[0056] 2.根据项目1的方法,其中所述分配协调器周期性地分析所述聚簇的全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变进一步包括:

[0057] 所述分配协调器调用可插入分配逻辑模块,其中所述可插入分配逻辑模块能在运行时被切换以调节用于在所述计算机节点的聚簇中分配所述分割部的算法。

[0058] 3.根据项目1的方法,其中采用分布式算法进一步包括:

[0059] 在两个计算机节点之间点对点地执行各个分割部的转移,所述转移是发送节点和

接收节点之间的直接非同步通信的结果,其中所述分配协调器不参与指引所述分割部的转移。

[0060] 4.根据项目1的方法,其中所述聚簇的全局状态包括指示所述分割部中的哪些被分配给所述聚簇中的各个计算机节点的信息。

[0061] 5.根据项目1的方法,其中所述聚簇的全局状态包括指示所述聚簇中的每个计算机节点上的当前处理负载的信息,其中所述当前处理负载通过所述每个计算机节点周期性地发送运行时反馈统计数据到所述分配协调器来确定。

[0062] 6.根据项目1的方法,其中所述聚簇的全局状态包括指示所述聚簇中的每个计算机节点的存储器能力和处理器能力的信息。

[0063] 7.根据项目1的方法,其中所述多个分割部进一步包括一组主分割部和一组备份分割部,其中所述分配协调器确保每个主分割部位于与该主分割部所关联的备份分割部不同的物理节点上。

[0064] 8.根据项目7的方法,其中所述分配协调器产生的新分配计划实施下列限制:

[0065] 对于每个给定计算机节点,限制允许包含与位于所述给定计算机节点上的主分割部相关联的备份分割部的计算机节点的数量。

[0066] 9.根据项目1的方法,其中所述分配协调器为所述分割部在所述聚簇中的计算机节点中的分配提供单个协调点。

[0067] 10.根据项目1的方法,其中所述新分配计划指明特定分割部应位于所述聚簇中的指定计算机节点上。

[0068] 11.一种用于提供可扩展集中式动态资源分配的系统,该系统包括:

[0069] 计算机节点的聚簇,储存数据集的多个分割部,所述分割部分配在所述计算机节点的聚簇上;以及

[0070] 选自所述计算机节点的分配协调器,所述分配协调器收集指示所述聚簇的全局状态的信息,周期性地分析所述全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变,根据所述改变产生新分配计划,以及向所述计算机节点的聚簇中的所有计算机节点提供对所述分配计划的访问;

[0071] 其中所述计算机节点的聚簇采用分布式算法来由所述聚簇中的每个节点独立地确定如何执行与所述节点相关联的各个分割部的转移以实施所述分配协调器产生的新分配计划。

[0072] 12.根据项目11的系统,其中所述分配协调器调用可插入分配逻辑模块以确定是否应进行所述改变,其中所述可插入分配逻辑模块能在运行时被切换以调节用于在所述计算机节点的聚簇中分配所述分割部的算法。

[0073] 13.根据项目11的系统,其中所述计算机节点采用所述分布式算法进一步包括:

[0074] 在两个计算机节点之间点对点地执行各个分割部的转移,所述转移是发送节点和接收节点之间的直接非同步通信的结果,其中所述分配协调器不参与指引所述分割部的转移。

[0075] 14.根据项目11的系统,其中所述聚簇的全局状态包括指示所述分割部中的哪些被分配给所述聚簇中的各个计算机节点的信息。

[0076] 15.根据项目11的系统,其中所述聚簇的全局状态包括指示所述聚簇中的每个计

算机节点上的当前处理负载的信息,其中所述当前处理负载通过所述每个计算机节点周期性地传输运行时反馈统计数据到所述分配协调器来确定。

[0077] 16.根据项目11的系统,其中所述聚簇的全局状态包括指示所述聚簇中的每个计算机节点的存储器能力和处理器能力的信息。

[0078] 17.根据项目11的系统,其中所述多个分割部进一步包括一组主分割部和一组备份分割部,其中所述分配协调器确保每个主分割部位于与所述主分割部所关联的备份分割部不同的物理节点上。

[0079] 18.根据项目17的系统,其中所述分配协调器产生的新分配计划实施下列限制:

[0080] 对于每个给定计算机节点,限制允许包含与位于所述给定计算机节点上的主分割部相关联的备份分割部的计算机节点的数目。

[0081] 19.根据项目11的系统,其中所述新分配计划指明特定分割部应位于所述聚簇中的指定计算机节点上。

[0082] 20.一种非暂时性计算机可读储存介质,存储一个或多个指令序列,所述指令序列可由一个或多个处理器运行以执行一组步骤,所述步骤包括:

[0083] 储存数据集的多个分割部,所述多个分割部分配在计算机节点的聚簇上;

[0084] 指定所述计算机节点中的一个节点作为中央分配协调器,该中央分配协调器收集指示所述聚簇的全局状态的信息并且提供对所述全局状态的访问;

[0085] 所述分配协调器周期性地分析所述聚簇的全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变;

[0086] 所述分配协调器基于对所述分割部的分配的改变产生新分配计划,并且向所述计算机节点的聚簇中的所有计算机节点提供对所述分配计划的访问;以及

[0087] 采用分布式算法来由所述聚簇中的每个节点独立地确定如何执行与所述节点相关联的各个分割部的转移以实施所述新分配计划。

[0088] 21.一种用于提供可扩展集中式动态资源分配的系统,所述系统包括:

[0089] 用于储存数据集的多个分割部的装置,所述多个分割部分配在计算机节点的聚簇上;

[0090] 用于指定所述计算机节点中的一个节点作为中央分配协调器的装置,该中央分配协调器收集指示所述聚簇的全局状态的信息并且提供对所述全局状态的访问;

[0091] 用于由所述分配协调器周期性地分析所述聚簇的全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变的装置;

[0092] 用于由所述分配协调器基于对所述分割部的分配的改变产生新分配计划,并且向所述计算机节点的聚簇中的所有计算机节点提供对所述分配计划的访问的装置;以及

[0093] 用于采用分布式算法来由所述聚簇中的每个节点独立地确定如何执行与所述节点相关联的各个分割部的转移以实施所述新分配计划的装置。

[0094] 22.根据项目21的系统,其中用于由所述分配协调器周期性地分析所述聚簇的全局状态以确定是否应对所述分割部在所述计算机节点中的分配进行改变的装置进一步包括:

[0095] 用于由所述分配协调器调用可插入分配逻辑模块的装置,其中所述可插入分配逻辑模块能在运行时被切换以调节用于在所述计算机节点的聚簇中分配所述分割部的算法。

[0096] 23. 根据项目21的系统,其中用于采用分布式算法的装置进一步包括:

[0097] 用于在两个计算机节点之间点对点地执行各个分割部的转移的装置,所述转移是发送节点和接收节点之间的直接非同步通信的结果,其中所述分配协调器不参与指引所述分割部的转移。

[0098] 24. 根据项目21的系统,其中所述聚簇的全局状态包括指示所述分割部中的哪些被分配给所述聚簇中的各个计算机节点的信息。

[0099] 25. 根据项目21的系统,其中所述聚簇的全局状态包括指示所述聚簇中的每个计算机节点上的当前处理负载的信息,其中所述当前处理负载通过所述每个计算机节点周期性地发送运行时反馈统计数据到所述分配协调器来确定。

[0100] 26. 根据项目21的系统,其中所述聚簇的全局状态包括指示所述聚簇中的每个计算机节点的存储器能力和处理器能力的信息。

[0101] 27. 根据项目21的系统,其中所述多个分割部进一步包括一组主分割部和一组备份分割部,其中所述分配协调器确保每个主分割部位于与该主分割部所关联的备份分割部不同的物理节点上。

[0102] 28. 根据项目27的系统,其中所述分配协调器产生的新分配计划实施下列限制:

[0103] 对于每个给定计算机节点,限制允许包含与位于所述给定计算机节点上的主分割部相关联的备份分割部的计算机节点的数量。

[0104] 29. 根据项目21的系统,其中所述分配协调器为所述分割部在所述聚簇中的计算机节点中的分配提供单个协调点。

[0105] 30. 根据项目21的系统,其中所述新分配计划指明特定分割部应位于所述聚簇中的指定计算机节点上。

[0106] 已经为了示范和说明而给出了前面对本发明的优选实施例的描述。其无意是详尽无遗的,也无意将本发明限制到所公开的精确形式。许多修改和变型对本领域技术人员而言是显而易见的。选择和描述了这些实施例是为了最佳地说明本发明的原理和其实际应用,由此使相关领域的技术人员能够理解本发明。本发明的范围旨在由所附权利要求及其等价物定义。

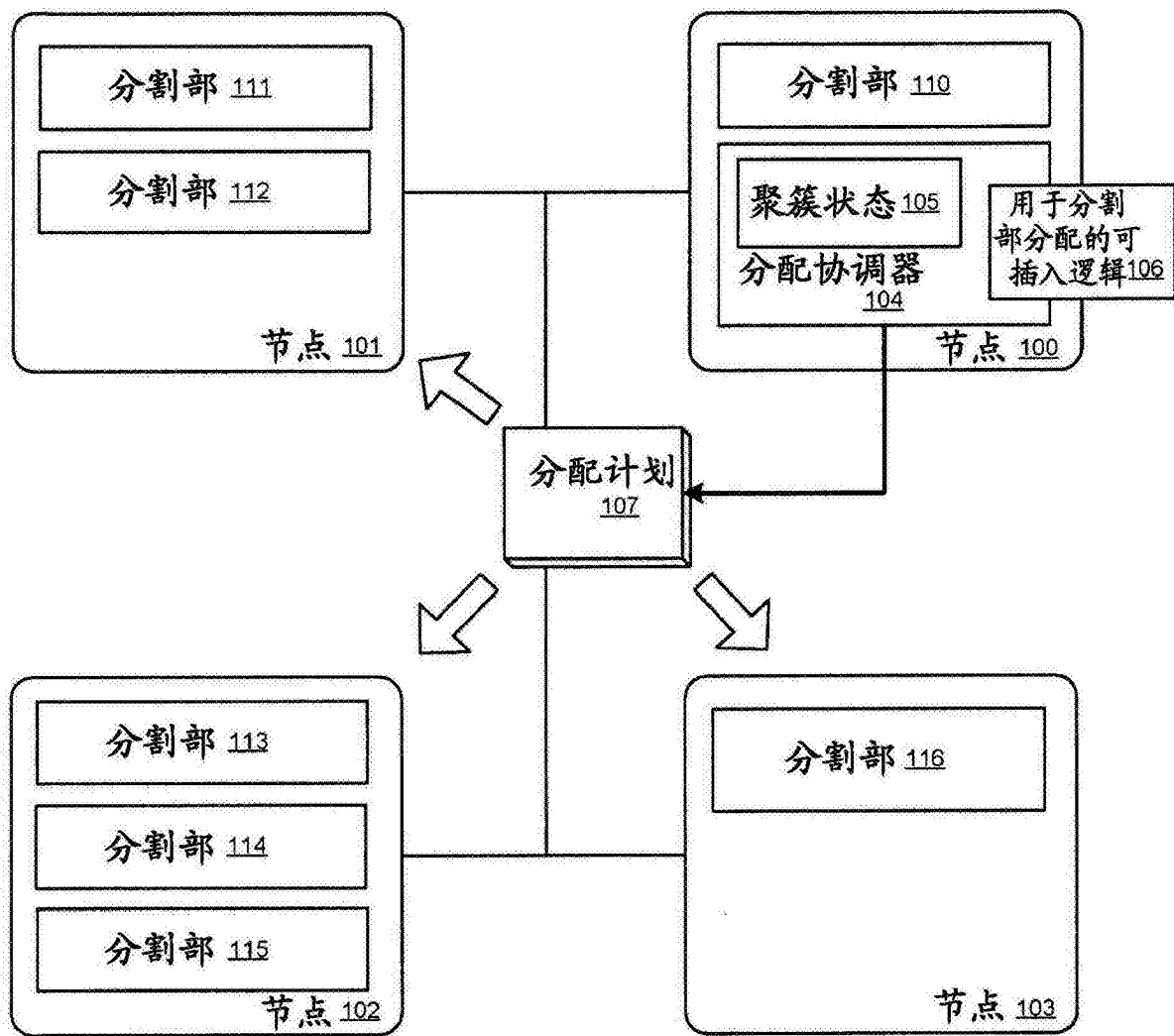


图1

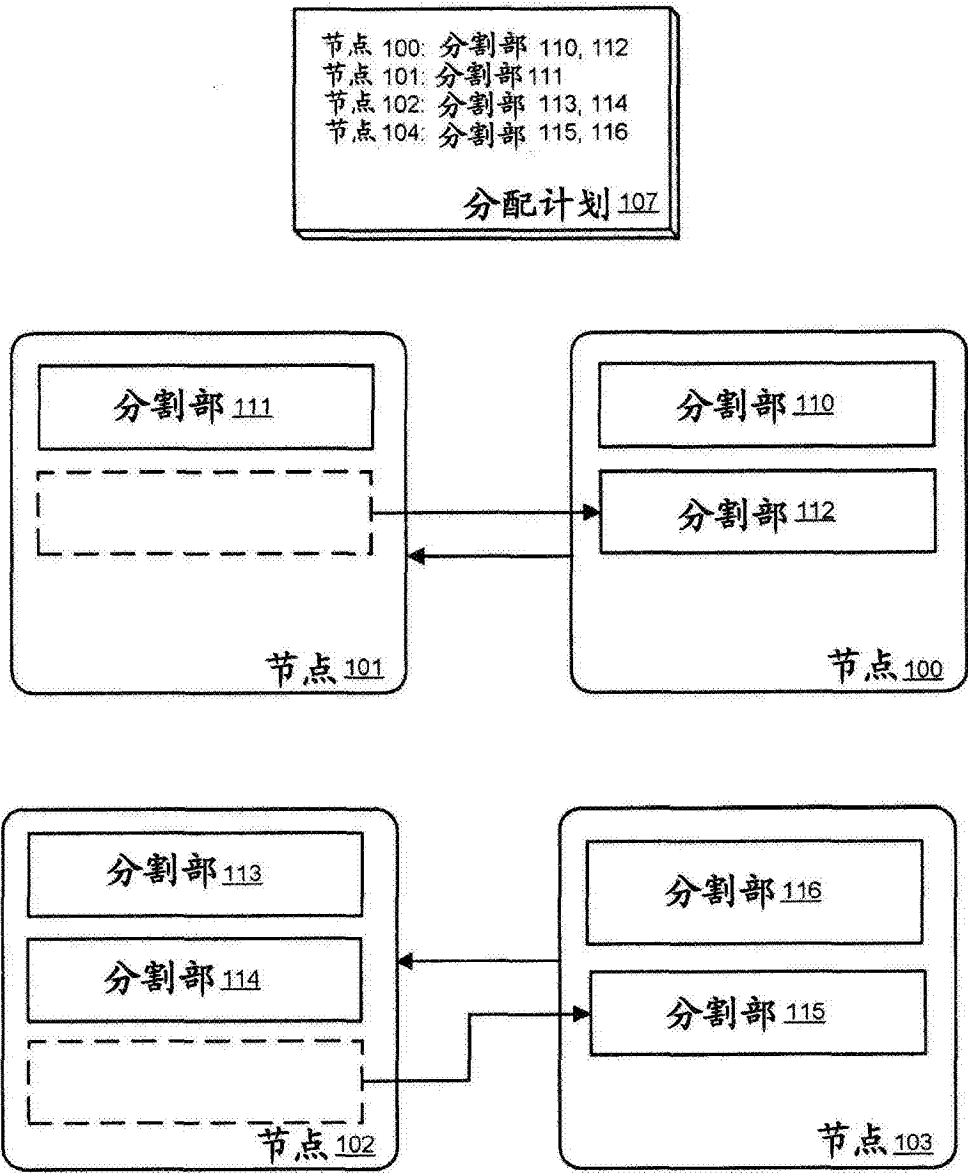


图2

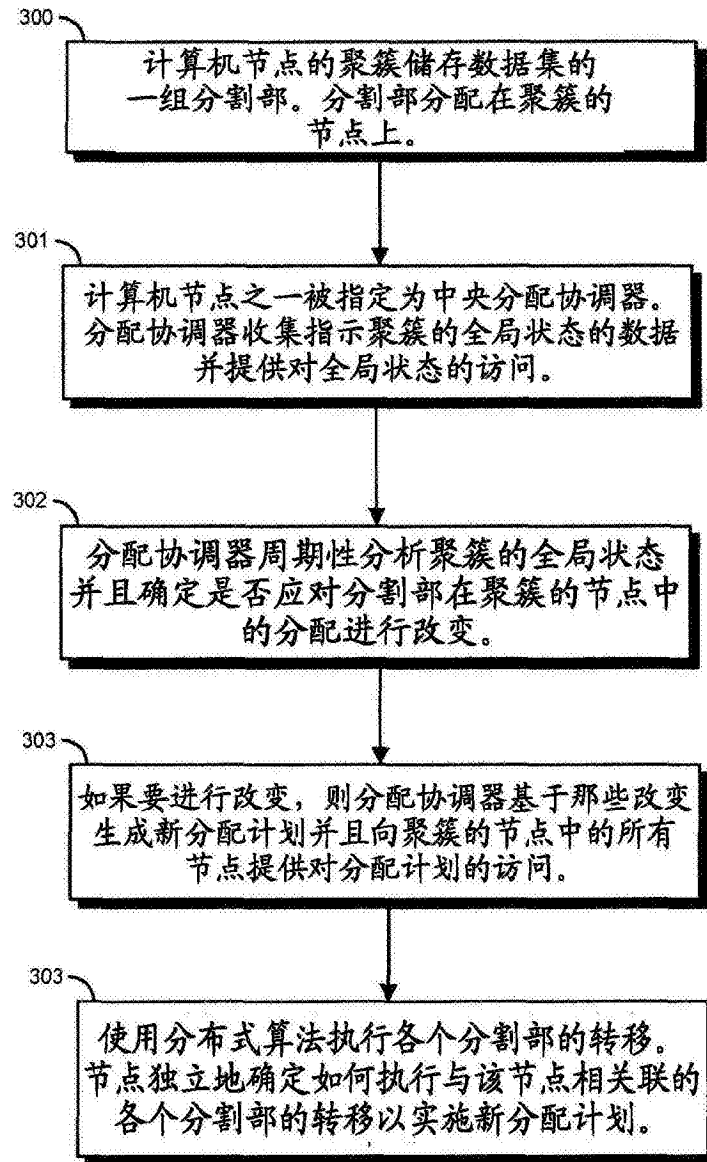


图3

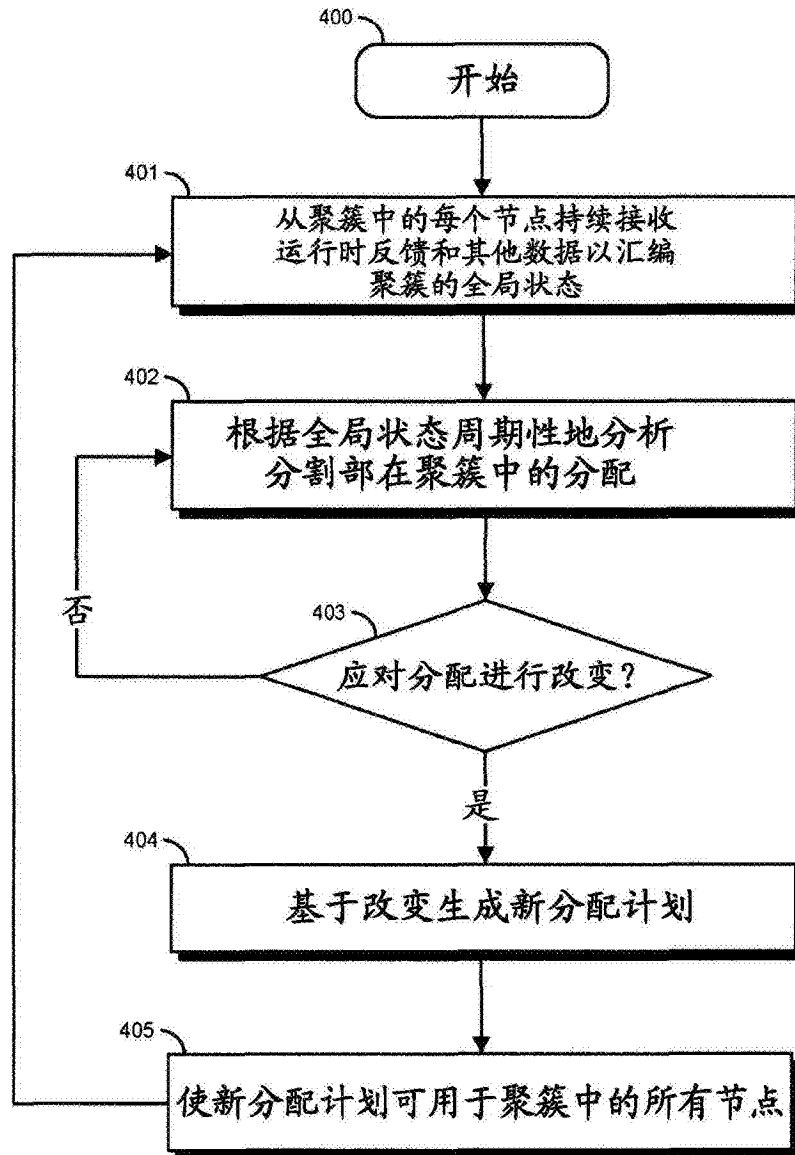


图4

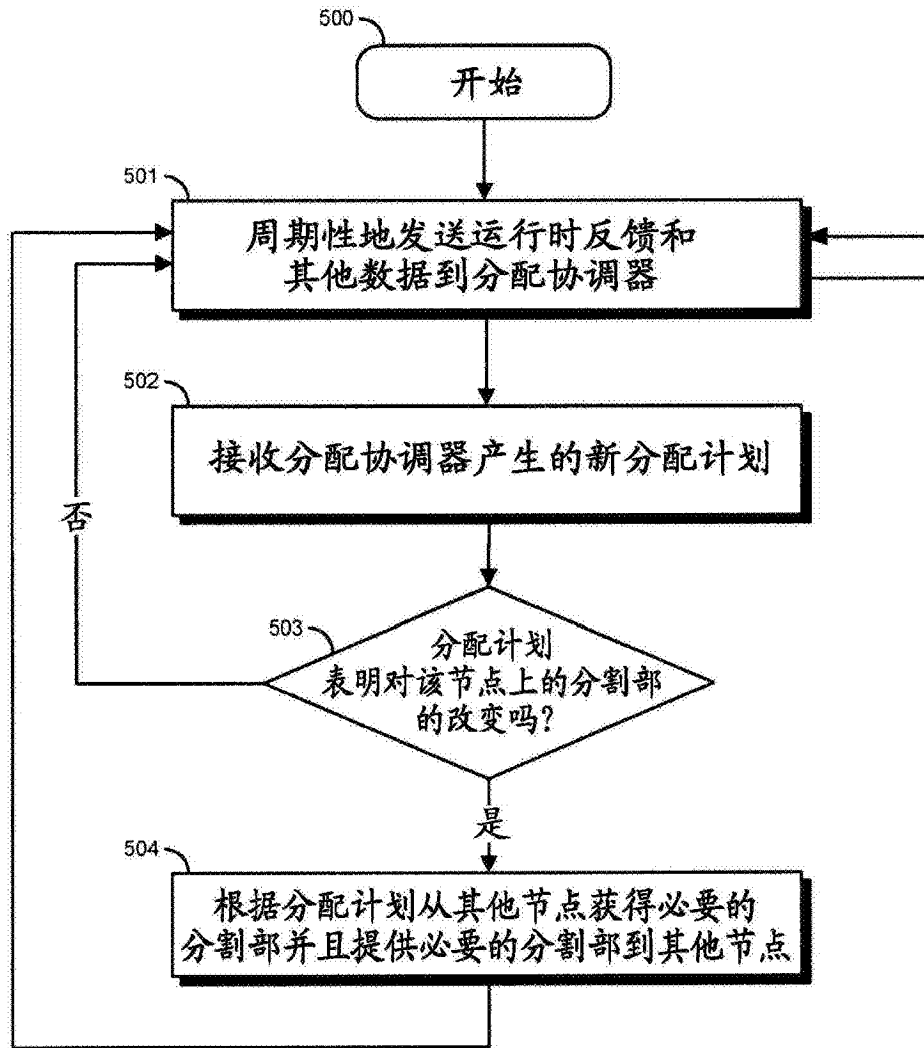


图5