(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106528288 A
(43)申请公布日 2017.03.22

(21)申请号 201510574287.3

(22)申请日 2015.09.10

(71)申请人 中兴通讯股份有限公司
地址 518057 广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦法务部

(72)发明人 郑鹏飞

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 胡艳华 龙洪

(51)Int.Cl.
G06F 9/50(2006.01)

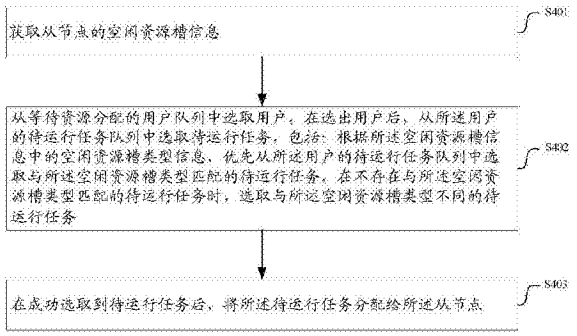
权利要求书3页 说明书8页 附图4页

(54)发明名称

一种资源管理方法、装置和系统

(57)摘要

本发明公开了一种资源管理方法，应用于Hadoop系统的主节点，该方法包括：获取从节点的空闲资源槽信息；从等待资源分配的用户队列中选取用户，在选出用户后，从所述用户的待运行任务队列中选取待运行任务，包括：根据所述空闲资源槽信息中的空闲资源槽类型信息，优先从所述用户的待运行任务队列中选取与所述空闲资源槽类型匹配的待运行任务，在不存在与所述空闲资源槽类型匹配的待运行任务时，选取与所述空闲资源槽类型不同的待运行任务；在成功选取到待运行任务后，将所述待运行任务分配给所述从节点。本发明能够提高Hadoop系统的资源利用率。



CN 106528288 A

1.一种资源管理方法,应用于Hadoop系统的主节点,该方法包括:

获取从节点的空闲资源槽信息;

从等待资源分配的用户队列中选取用户,在选出用户后,从所述用户的待运行任务队列中选取待运行任务,包括:根据所述空闲资源槽信息中的空闲资源槽类型信息,优先从所述用户的待运行任务队列中选取与所述空闲资源槽类型匹配的待运行任务,在不存在与所述空闲资源槽类型匹配的待运行任务时,选取与所述空闲资源槽类型不同的待运行任务;

在成功选取到待运行任务后,将所述待运行任务分配给所述从节点。

2.如权利要求1所述的方法,其特征在于:

所述从等待资源分配的用户队列中选取用户,包括:

从所述等待资源分配的用户队列的头部开始扫描所述用户队列;

每扫描到一个用户,判断所述用户是否满足分配条件,如所述用户满足所述分配条件,则扫描终止,如所述用户不满足所述分配条件,则扫描下一个用户。

3.如权利要求2所述的方法,其特征在于:

所述分配条件包括:所述用户具有满足数据本地性要求的待运行任务。

4.如权利要求3所述的方法,其特征在于:

在所述分配条件包含数据本地性要求时,从所述用户的待运行任务队列中选取待运行任务,还包括:

如扫描完所有的用户后未能选取到待运行任务,则从所述分配条件中去除数据本地性要求,重新从所述等待资源分配的用户队列的头部开始扫描所述用户队列,每扫描到一个用户,判断所述用户是否具有待运行任务,如所述用户具有待运行任务,则扫描终止,从所述用户的待运行任务队列中选取待运行任务,如所述用户没有待运行任务,则扫描下一个用户。

5.一种资源管理方法,应用于Hadoop系统的从节点,该方法包括:

检测到空闲资源槽后,向主节点发送携带空闲资源槽信息的通知消息,所述空闲资源槽信息包括本节点的空闲资源槽的类型信息;

接收所述主节点为本节点的空闲资源槽分配的待运行任务并将接收到的待运行任务放入任务启动队列中;

在所述任务启动队列非空且当前存在空闲资源槽时,从所述任务启动队列中取出待运行任务进行启动。

6.如权利要求5所述的方法,其特征在于:

所述接收所述主节点分配的待运行任务并将接收到的待运行任务放入任务启动队列中,包括:

将接收到的map任务放入map任务启动队列,将接收到的reduce任务放入reduce任务启动队列;

所述在所述任务启动队列非空且当前存在空闲资源槽时,从所述任务启动队列中取出待运行任务进行启动,包括:

如所述reduce任务启动队列非空且当前存在空闲资源槽,则从所述reduce任务启动队列中取出待运行任务进行启动;

如所述reduce任务启动队列为空且所述map任务启动队列非空且当前存在空闲资源

槽，则从所述 map 任务启动队列中取出待运行任务进行启动。

7. 一种资源管理装置，应用于 Hadoop 系统的主节点，包括：

信息接收模块，用于获取从节点的空闲资源槽信息；

任务调度模块，用于从等待资源分配的用户队列中选取用户，在选出用户后，从所述用户的待运行任务队列中选取待运行任务，包括：根据所述空闲资源槽信息中的空闲资源槽类型信息，优先从所述用户的待运行任务队列中选取与所述空闲资源槽类型匹配的待运行任务，在不存在与所述空闲资源槽类型匹配的待运行任务时，选取与所述空闲资源槽类型不同的待运行任务；

信息发送模块，用于在成功选取到待运行任务后，将所述待运行任务分配给所述从节点。

8. 如权利要求 7 所述的装置，其特征在于：

所述任务调度模块，用于从等待资源分配的用户队列中选取用户，包括：

从所述等待资源分配的用户队列的头部开始扫描所述用户队列；

每扫描到一个用户，判断所述用户是否满足分配条件，如所述用户满足所述分配条件，则扫描终止，如所述用户不满足所述分配条件，则扫描下一个用户。

9. 如权利要求 8 所述的装置，其特征在于：

所述分配条件包括：所述用户具有满足数据本地性要求的待运行任务。

10. 如权利要求 9 所述的装置，其特征在于：

所述任务调度模块，用于在所述分配条件包含数据本地性要求时，从所述用户的待运行任务队列中选取待运行任务，还包括：如扫描完所有的用户后未能选取到待运行任务，则从所述分配条件中去除数据本地性要求，重新从所述等待资源分配的用户队列的头部开始扫描所述用户队列，每扫描到一个用户，判断所述用户是否具有待运行任务，如所述用户具有待运行任务，则扫描终止，从所述用户的待运行任务队列中选取待运行任务，如所述用户没有待运行任务，则扫描下一个用户。

11. 一种资源管理装置，应用于 Hadoop 系统的从节点，包括：

检测及上报模块，用于向主节点发送携带空闲资源槽信息的通知消息，所述空闲资源槽信息包括本节点的空闲资源槽的类型信息；

接收及处理模块，用于接收所述主节点为本节点的空闲资源槽分配的待运行任务并将接收到的待运行任务放入任务启动队列中；

任务启动模块，用于在所述任务启动队列非空且当前存在空闲资源槽时，从所述任务启动队列中取出待运行任务进行启动。

12. 如权利要求 11 所述的装置，其特征在于：

所述接收及处理模块，用于接收所述主节点分配的待运行任务并将接收到的待运行任务放入任务启动队列中，包括：

将接收到的 map 任务放入 map 任务启动队列，将接收到的 reduce 任务放入 reduce 任务启动队列；

所述任务启动模块，用于在所述任务启动队列非空且当前存在空闲资源槽时，从所述任务启动队列中取出待运行任务进行启动，包括：

如所述 reduce 任务启动队列非空且当前存在空闲资源槽，则从所述 reduce 任务启动

队列中取出待运行任务进行启动；

如所述 reduce 任务启动队列为空且所述 map 任务启动队列非空且当前存在空闲资源槽，则从所述 map 任务启动队列中取出待运行任务进行启动。

13. 一种资源管理系统，包括：

具有权利要求 7-10 中任一项所述的资源管理装置的 Hadoop 系统主节点，和具有权利要求 11-12 中任一项所述的资源管理装置的 Hadoop 系统从节点。

一种资源管理方法、装置和系统

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及的是一种资源管理方法、装置和系统。

背景技术

[0002] Hadoop 系统是目前使用十分广泛的一个分布式系统，用来处理大规模数据。Hadoop 集群由一个主节点和多个从节点组成，每个节点可以是一台计算机或者一台虚拟机。主节点用来管理 Hadoop 分布式文件系统 HDFS(Hadoop Distributed File System，HDFS) 和各个作业的处理过程（即 MapReduce 计算框架），从节点负责数据的存储和对作业数据的处理。Hadoop 采用 Google 公司提出的 MapReduce 并行处理框架。主节点在 MapReduce 中称为 JobTracker，负责作业的处理过程；从节点在 MapReduce 框架中称之为 TaskTracker，负责作业任务的执行。Hadoop 作业的输入数据被划分成很多大小相同的数据块分布在计算机集群中，由多个节点并行处理这些输入数据从而加快作业的处理时间。一个节点可以通过配置同时存储和处理多个数据块，每个数据块对应一个任务。作业的执行分为两个阶段：第一个阶段即 map 阶段，各个节点处理分布在集群中作业的 map 任务；第二个阶段为 Reduce 阶段，即通过 reduce 任务对分布在各个节点的 map 任务处理结果进行汇总，形成最终的作业处理结果。

[0003] 在 Hadoop 集群中，所有的计算资源被抽象为槽，每个槽可以被独占用来处理一个任务，根据计算节点（即从节点）的硬件配置，管理员可以配置不同数目的槽。由于每个作业都由一个 map 任务集合和一个 reduce 任务集合组成，而 map 任务和 reduce 任务对集群资源的需求有所不同，所以将槽划分为 map 槽和 reduce 槽两种类型。其中，map 槽只能运行 map 任务，reduce 槽只能运行 reduce 任务。所以在 Hadoop 中槽是最基本的计算单元，并且槽的数目在集群启动前已被管理员配置完毕，运行过程中不能改变。资源槽也是资源分配的基本单位，每个资源槽占用着本节点上一定的物理资源，比如 CPU、内存、磁盘和网络带宽。图 1 是一个计算节点和资源槽的示意图。

[0004] 在 Hadoop 中，每个作业包括 map 任务集合和 reduce 任务集合，每一个任务对应一个资源槽 (map 任务对应 map 槽，reduce 任务对应 reduce 槽)，对作业任务的执行有两个严格的限制：(1)reduce 任务必须在所有 map 任务完成后才能真正开始；(2)map 任务只能运行在 map 槽上，reduce 任务只能运行在 reduce 槽上。这两个限制带来的结果就是在不同的作业负载和资源槽配置下，集群资源利用率和性能都有较大不同，即使在最优的作业提交顺序和最优的配置资源槽下仍然会严重影响相应资源槽的利用率。由于 map 任务和 reduce 任务的数目随着时间的推移都在不断的变化，分配给 map(或者 reduce) 任务的资源槽数目可能会超过 map(或者 reduce) 任务的数目。所以，在 MapReduce 集群动态负载下，可能会出现一种资源槽负载过重而另一种资源槽却有空闲，从而导致资源浪费

发明内容

[0005] 本发明所要解决的技术问题是提供一种资源管理方法、装置和系统，能够提高

Hadoop 系统的资源利用率。

[0006] 本发明实施例提供了一种资源管理方法，应用于 Hadoop 系统的主节点，该方法包括：

[0007] 获取从节点的空闲资源槽信息；

[0008] 从等待资源分配的用户队列中选取用户，在选出用户后，从所述用户的待运行任务队列中选取待运行任务，包括：根据所述空闲资源槽信息中的空闲资源槽类型信息，优先从所述用户的待运行任务队列中选取与所述空闲资源槽类型匹配的待运行任务，在不存在与所述空闲资源槽类型匹配的待运行任务时，选取与所述空闲资源槽类型不同的待运行任务；

[0009] 在成功选取到待运行任务后，将所述待运行任务分配给所述从节点。

[0010] 可选地，所述从等待资源分配的用户队列中选取用户，包括：

[0011] 从所述等待资源分配的用户队列的头部开始扫描所述用户队列；

[0012] 每扫描到一个用户，判断所述用户是否满足分配条件，如所述用户满足所述分配条件，则扫描终止，如所述用户不满足所述分配条件，则扫描下一个用户。

[0013] 可选地，所述分配条件包括：所述用户具有满足数据本地性要求的待运行任务。

[0014] 可选地，在所述分配条件包含数据本地性要求时，从所述用户的待运行任务队列中选取待运行任务，还包括：

[0015] 如扫描完所有的用户后未能选取到待运行任务，则从所述分配条件中去除数据本地性要求，重新从所述等待资源分配的用户队列的头部开始扫描所述用户队列，每扫描到一个用户，判断所述用户是否具有待运行任务，如所述用户具有待运行任务，则扫描终止，从所述用户的待运行任务队列中选取待运行任务，如所述用户没有待运行任务，则扫描下一个用户。

[0016] 本发明实施例还提供了一种资源管理方法，应用于 Hadoop 系统的从节点，该方法包括：

[0017] 检测到空闲资源槽后，向主节点发送携带空闲资源槽信息的通知消息，所述空闲资源槽信息包括本节点的空闲资源槽的类型信息；

[0018] 接收所述主节点为本节点的空闲资源槽分配的待运行任务并将接收到的待运行任务放入任务启动队列中；

[0019] 在所述任务启动队列非空且当前存在空闲资源槽时，从所述任务启动队列中取出待运行任务进行启动。

[0020] 可选地，所述接收所述主节点分配的待运行任务并将接收到的待运行任务放入任务启动队列中，包括：

[0021] 将接收到的 map 任务放入 map 任务启动队列，将接收到的 reduce 任务放入 reduce 任务启动队列；

[0022] 所述在所述任务启动队列非空且当前存在空闲资源槽时，从所述任务启动队列中取出待运行任务进行启动，包括：

[0023] 如所述 reduce 任务启动队列非空且当前存在空闲资源槽，则从所述 reduce 任务启动队列中取出待运行任务进行启动；

[0024] 如所述 reduce 任务启动队列为空且所述 map 任务启动队列非空且当前存在空闲

资源槽，则从所述 map 任务启动队列中取出待运行任务进行启动。

[0025] 本发明实施例还提供了一种资源管理装置，应用于 Hadoop 系统的主节点，包括：

[0026] 信息接收模块，用于获取从节点的空闲资源槽信息；

[0027] 任务调度模块，用于从等待资源分配的用户队列中选取用户，在选出用户后，从所述用户的待运行任务队列中选取待运行任务，包括：根据所述空闲资源槽信息中的空闲资源槽类型信息，优先从所述用户的待运行任务队列中选取与所述空闲资源槽类型匹配的待运行任务，在不存在与所述空闲资源槽类型匹配的待运行任务时，选取与所述空闲资源槽类型不同的待运行任务；

[0028] 信息发送模块，用于在成功选取到待运行任务后，将所述待运行任务分配给所述从节点。

[0029] 可选地，所述任务调度模块，用于从等待资源分配的用户队列中选取用户，包括：

[0030] 从所述等待资源分配的用户队列的头部开始扫描所述用户队列；

[0031] 每扫描到一个用户，判断所述用户是否满足分配条件，如所述用户满足所述分配条件，则扫描终止，如所述用户不满足所述分配条件，则扫描下一个用户。

[0032] 可选地，所述分配条件包括：所述用户具有满足数据本地性要求的待运行任务。

[0033] 可选地，所述任务调度模块，用于在所述分配条件包含数据本地性要求时，从所述用户的待运行任务队列中选取待运行任务，还包括：如扫描完所有的用户后未能选取到待运行任务，则从所述分配条件中去除数据本地性要求，重新从所述等待资源分配的用户队列的头部开始扫描所述用户队列，每扫描到一个用户，判断所述用户是否具有待运行任务，如所述用户具有待运行任务，则扫描终止，从所述用户的待运行任务队列中选取待运行任务，如所述用户没有待运行任务，则扫描下一个用户。

[0034] 本发明实施例还提供了一种资源管理装置，应用于 Hadoop 系统的从节点，包括：

[0035] 检测及上报模块，用于向主节点发送携带空闲资源槽信息的通知消息，所述空闲资源槽信息包括本节点的空闲资源槽的类型信息；

[0036] 接收及处理模块，用于接收所述主节点为本节点的空闲资源槽分配的待运行任务并将接收到的待运行任务放入任务启动队列中；

[0037] 任务启动模块，用于在所述任务启动队列非空且当前存在空闲资源槽时，从所述任务启动队列中取出待运行任务进行启动。

[0038] 可选地，所述接收及处理模块，用于接收所述主节点分配的待运行任务并将接收到的待运行任务放入任务启动队列中，包括：

[0039] 将接收到的 map 任务放入 map 任务启动队列，将接收到的 reduce 任务放入 reduce 任务启动队列；

[0040] 所述任务启动模块，用于在所述任务启动队列非空且当前存在空闲资源槽时，从所述任务启动队列中取出待运行任务进行启动，包括：

[0041] 如所述 reduce 任务启动队列非空且当前存在空闲资源槽，则从所述 reduce 任务启动队列中取出待运行任务进行启动；

[0042] 如所述 reduce 任务启动队列为空且所述 map 任务启动队列非空且当前存在空闲资源槽，则从所述 map 任务启动队列中取出待运行任务进行启动。

[0043] 本发明实施例还提供了一种资源管理系统，包括：

[0044] 具有上述资源管理装置的Hadoop系统主节点，和具有上述资源管理装置的Hadoop系统从节点。

[0045] 与现有技术相比，本发明提供的一种资源管理方法、装置和系统，通过对主节点上的调度器和从节点上的任务跟踪器进行改进，打破了Hadoop系统中map槽只能运行map任务，reduce槽只能运行reduce任务的限制，尽可能使所有的资源槽都保持忙碌，从而提高Hadoop系统的资源利用率。

附图说明

[0046] 图1为计算节点和资源槽的示意图。

[0047] 图2为本发明实施例用户资源池内部资源槽借用示意图。

[0048] 图3为本发明实施例用户间资源池资源槽借用示意图。

[0049] 图4为本发明实施例一种资源管理方法的示意图（主节点）。

[0050] 图5为本发明实施例一种资源管理方法的示意图（从节点）。

[0051] 图6为本发明实施例一种资源管理装置示意图（主节点）。

[0052] 图7为本发明实施例一种资源管理装置示意图（从节点）。

[0053] 图8为本发明实施例一种资源管理系统示意图。

具体实施方式

[0054] 为使本发明的目的、技术方案和优点更加清楚明白，下文中将结合附图对本发明的实施例进行详细说明。需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互任意组合。

[0055] 多个作业在从map阶段进入reduce阶段的过程中，在不同的时间段可能会出现一类资源槽空闲，而另一类却负载过重。对于这些空闲的reduce槽（或者map槽），可以借给负载过重的map（或者reduce）任务使用，从而提高Hadoop系统的资源利用率。

[0056] 调度器（主节点上）会负责在用户级进行用户选择，在选定用户后，再从该用户的作业队列内选取合适的作业，最后将作业任务的启动交给MapReduce框架中的任务跟踪器TaskTracker（从节点上）。

[0057] 对Hadoop系统的调度器和MapReduce框架进行改进，打破MapReduce并行计算框架中对于map资源槽和reduce资源槽的限制，在保证用户之间公平性的同时，分别在用户资源池内部和用户资源池之间进行资源槽借用。如图2所示，用户资源池内部的借用，即借用用户资源池内的空闲资源槽给该用户负载过重的资源槽。如图3所示，用户资源池之间的借用，即用户可以借用其他用户资源池的空闲资源槽。资源槽借用减少了资源槽空闲现象，尽可能使所有资源槽保持忙碌，从而提高了Hadoop集群的资源利用率。

[0058] 如图4所示，本发明实施例提供了一种资源管理方法，应用于Hadoop系统的主节点，该方法包括：

[0059] S401，获取从节点的空闲资源槽信息；

[0060] 其中，所述获取从节点的空闲资源槽信息，包括：

[0061] 接收到从节点发送的请求分配任务的心跳消息后，根据所述心跳消息中携带的空闲资源槽信息获知所述从节点有空闲资源槽；

[0062] 其中，所述空闲资源槽信息包括：资源槽类型信息；

[0063] 其中，所述资源槽类型包括：map 资源槽或 reduce 资源槽；

[0064] S402，从等待资源分配的用户队列中选取用户，在选出用户后，从所述用户的待运行任务队列中选取待运行任务，包括：根据所述空闲资源槽信息中的空闲资源槽类型信息，优先从所述用户的待运行任务队列中选取与所述空闲资源槽类型匹配的待运行任务，在不存在与所述空闲资源槽类型匹配的待运行任务时，选取与所述空闲资源槽类型不同的待运行任务；

[0065] 其中，所述从等待资源分配的用户队列中选取用户，包括：

[0066] 从所述等待资源分配的用户队列的头部开始扫描所述用户队列；

[0067] 每扫描到一个用户，判断所述用户是否满足分配条件，如所述用户满足所述分配条件，则扫描终止，如所述用户不满足所述分配条件，则扫描下一个用户；

[0068] 其中，所述分配条件包括：所述用户具有满足数据本地性要求的待运行任务；

[0069] 其中，所述数据本地性要求是指：任务要处理的数据块与分配给该任务的资源槽在同一个节点或同一机架上；

[0070] 其中，在所述分配条件包含数据本地性要求时，从所述用户的待运行任务队列中选取待运行任务，还包括：

[0071] 如扫描完所有的用户后未能选取到待运行任务，则从所述分配条件中去除数据本地性要求，重新从所述等待资源分配的用户队列的头部开始扫描所述用户队列，每扫描到一个用户，判断所述用户是否具有待运行任务，如所述用户具有待运行任务，则扫描终止，从所述用户的待运行任务队列中选取待运行任务，如所述用户没有待运行任务，则扫描下一个用户；

[0072] 其中，所述等待资源分配的用户队列按照公平性算法对用户进行排序；

[0073] 其中，在同类型的待运行任务存在多个时，优先选取等待时间长的待运行任务；

[0074] 其中，用户内部资源槽的借用是指：用户在分配到资源槽后，分四种情况进行分析：1) 第一种情况：判断空闲资源槽是否为 map 槽且用户有满足本地性的 map 任务，若满足该条件则会将 map 资源槽分配给 map 任务；2) 第二种情况：判断空闲资源槽是否为 reduce 槽且用户有待执行的 reduce 任务，若满足该条件则将 reduce 资源槽分配给 reduce 任务；3) 第三种情况：判断空闲资源槽是否为 map 槽且用户有待执行 reduce 任务，若满足该条件则借用 map 槽执行 reduce 任务；4) 第四种情况：判断空闲资源槽是否为 reduce 槽且有满足本地性的 map 任务，若满足该条件则将 reduce 槽借用给 map 任务。可以看出，用户内部资源槽借用发生在上述第三种和第四种情况。

[0075] 用户之间资源槽的借用是指：在进行资源分配时首先会对用户队列按照优先级排序。按照优先级原则，该资源槽应该分配给优先级最高的用户。但是，很可能的情况是：该用户可能没有符合条件的任务，例如没有 reduce 任务且没有满足数据本地性的 map 任务，所以可以将该资源槽借用给其他用户。借用的实现就是通过扫描用户队列中下一个用户，判断下一个用户是否具有满足条件的任务，如果有则将该资源槽借给这个用户，否则继续依次扫描用户队列中的其他用户。

[0076] S403，在成功选取到待运行任务后，将所述待运行任务分配给所述从节点；

[0077] 其中，所述将所述待运行任务分配给所述从节点，包括：

[0078] 向所述从节点返回所述心跳消息的响应消息，其中携带所述待运行任务的信息；

[0079] 其中，主节点上的调度器负责调度并分配 map 任务或 reduce 任务给从节点；

[0080] 如图 5 所示，本发明实施例提供了一种资源管理方法，应用于 Hadoop 系统的从节点，该方法包括：

[0081] S501，检测到空闲资源槽后，向主节点发送携带空闲资源槽信息的通知消息，所述空闲资源槽信息包括本节点的空闲资源槽的类型信息；

[0082] 其中，所述向主节点发送携带所述空闲资源槽信息的通知消息，包括：

[0083] 向主节点发送请求分配任务的心跳消息，其中携带空闲资源槽信息；

[0084] 其中，所述资源槽类型包括：map 资源槽或 reduce 资源槽；

[0085] S502，接收所述主节点为本节点的空闲资源槽分配的待运行任务并将接收到的待运行任务放入任务启动队列中；

[0086] 可选地，所述接收所述主节点分配的待运行任务并将接收到的待运行任务放入任务启动队列中，包括：

[0087] 将接收到的 map 任务放入 map 任务启动队列，将接收到的 reduce 任务放入 reduce 任务启动队列；

[0088] 可选地，也可以将接收到的 map 任务和 reduce 任务放入同一个任务启动队列中；

[0089] S503，在所述任务启动队列非空且当前存在空闲资源槽时，从所述任务启动队列中取出待运行任务进行启动；

[0090] 可选地，所述在所述任务启动队列非空且当前存在空闲资源槽时，从所述任务启动队列中取出待运行任务进行启动，包括：

[0091] 如所述 reduce 任务启动队列非空且当前存在空闲资源槽，则从所述 reduce 任务启动队列中取出待运行任务进行启动；

[0092] 如所述 reduce 任务启动队列为空且所述 map 任务启动队列非空且当前存在空闲资源槽，则从所述 map 任务启动队列中取出待运行任务进行启动；

[0093] 先启动 reduce 任务有利于尽快结束作业，释放作业占用的资源。

[0094] 可选地，也可以采用其他的策略启动待运行任务；比如，如果将接收到的 map 任务和 reduce 任务放入同一个任务启动队列中，则可以顺序从所述任务启动队列中取出待运行任务进行启动；

[0095] 其中，从节点上的任务跟踪器 (TaskTracker) 负责启动 map 任务或 reduce 任务；

[0096] 如图 6 所示，本发明实施例提供了一种资源管理装置，应用于 Hadoop 系统的主节点，包括：

[0097] 信息接收模块 601，用于获取从节点的空闲资源槽信息；

[0098] 任务调度模块 602，用于从等待资源分配的用户队列中选取用户，在选出用户后，从所述用户的待运行任务队列中选取待运行任务，包括：根据所述空闲资源槽信息中的空闲资源槽类型信息，优先从所述用户的待运行任务队列中选取与所述空闲资源槽类型匹配的待运行任务，在不存在与所述空闲资源槽类型匹配的待运行任务时，选取与所述空闲资源槽类型不同的待运行任务；

[0099] 信息发送模块 603，用于在成功选取到待运行任务后，将所述待运行任务分配给所述从节点。

[0100] 其中，所述任务调度模块602，用于从等待资源分配的用户队列中选取用户，包括：

[0101] 从所述等待资源分配的用户队列的头部开始扫描所述用户队列；

[0102] 每扫描到一个用户，判断所述用户是否满足分配条件，如所述用户满足所述分配条件，则扫描终止，如所述用户不满足所述分配条件，则扫描下一个用户。

[0103] 其中，所述分配条件包括：所述用户具有满足数据本地性要求的待运行任务。

[0104] 其中，所述任务调度模块602，用于在所述分配条件包含数据本地性要求时，从所述用户的待运行任务队列中选取待运行任务，还包括：如扫描完所有的用户后未能选取到待运行任务，则从所述分配条件中去除数据本地性要求，重新从所述等待资源分配的用户队列的头部开始扫描所述用户队列，每扫描到一个用户，判断所述用户是否具有待运行任务，如所述用户具有待运行任务，则扫描终止，从所述用户的待运行任务队列中选取待运行任务，如所述用户没有待运行任务，则扫描下一个用户。

[0105] 其中，所述信息接收模块601，用于获取从节点的空闲资源槽信息，包括：

[0106] 接收到从节点发送的请求分配任务的心跳消息后，根据所述心跳消息中携带的空闲资源槽信息获知所述从节点有空闲资源槽；

[0107] 其中，所述资源槽类型包括：map资源槽或reduce资源槽。

[0108] 如图7所示，本发明实施例提供了一种资源管理装置，应用于Hadoop系统的从节点，包括：

[0109] 检测及上报模块701，用于向主节点发送携带空闲资源槽信息的通知消息，所述空闲资源槽信息包括本节点的空闲资源槽的类型信息；

[0110] 接收及处理模块702，用于接收所述主节点为本节点的空闲资源槽分配的待运行任务并将接收到的待运行任务放入任务启动队列中；

[0111] 任务启动模块703，用于在所述任务启动队列非空且当前存在空闲资源槽时，从所述任务启动队列中取出待运行任务进行启动。

[0112] 其中，所述接收及处理模块702，用于接收所述主节点分配的待运行任务并将接收到的待运行任务放入任务启动队列中，包括：

[0113] 将接收到的map任务放入map任务启动队列，将接收到的reduce任务放入reduce任务启动队列；

[0114] 所述任务启动模块703，用于在所述任务启动队列非空且当前存在空闲资源槽时，从所述任务启动队列中取出待运行任务进行启动，包括：

[0115] 如所述reduce任务启动队列非空且当前存在空闲资源槽，则从所述reduce任务启动队列中取出待运行任务进行启动；

[0116] 如所述reduce任务启动队列为空且所述map任务启动队列非空且当前存在空闲资源槽，则从所述map任务启动队列中取出待运行任务进行启动。

[0117] 其中，所述检测及上报模块701，用于向主节点发送携带所述空闲资源槽信息的通知消息，包括：

[0118] 向主节点发送请求分配任务的心跳消息，其中携带空闲资源槽信息；

[0119] 其中，所述资源槽类型包括：map资源槽或reduce资源槽。

[0120] 如图8所示，本发明实施例提供了一种资源管理系统，包括：具有上述资源管理装

置的 Hadoop 系统主节点，和具有上述资源管理装置的 Hadoop 系统从节点。

[0121] 上述实施例提供的一种资源管理方法、装置和系统，通过对主节点上的调度器和从节点上的任务跟踪器进行改进，打破了 Hadoop 系统中 map 槽只能运行 map 任务，reduce 槽只能运行 reduce 任务的限制，尽可能使所有的资源槽都保持忙碌，从而提高 Hadoop 系统的资源利用率。

[0122] 本领域普通技术人员可以理解上述方法中的全部或部分步骤可通过程序来指令相关硬件完成，所述程序可以存储于计算机可读存储介质中，如只读存储器、磁盘或光盘等。可选地，上述实施例的全部或部分步骤也可以使用一个或多个集成电路来实现，相应地，上述实施例中的各模块 / 单元可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。本发明不限制于任何特定形式的硬件和软件的结合。

[0123] 需要说明的是，本发明还可有其他多种实施例，在不背离本发明精神及其实质的情况下，熟悉本领域的技术人员可根据本发明作出各种相应的改变和变形，但这些相应的改变和变形都应属于本发明所附的权利要求的保护范围。

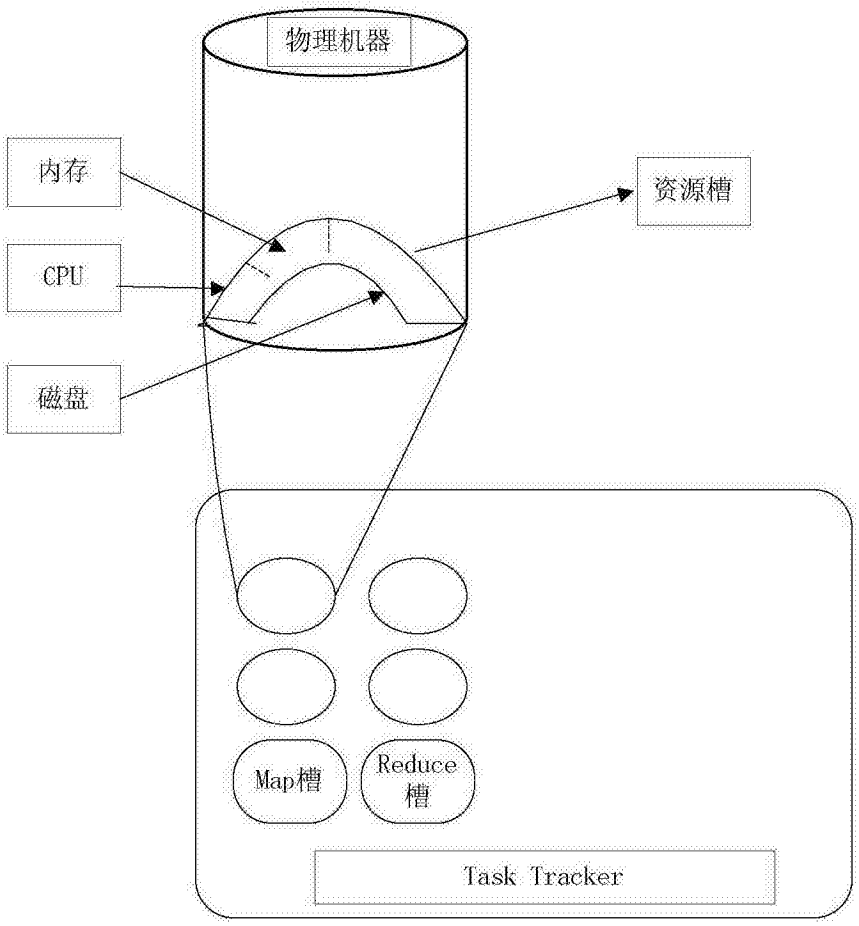

物理机器
内存
CPU
磁盘
资源槽
Map槽
Reduce
槽
Task Tracker

图 1

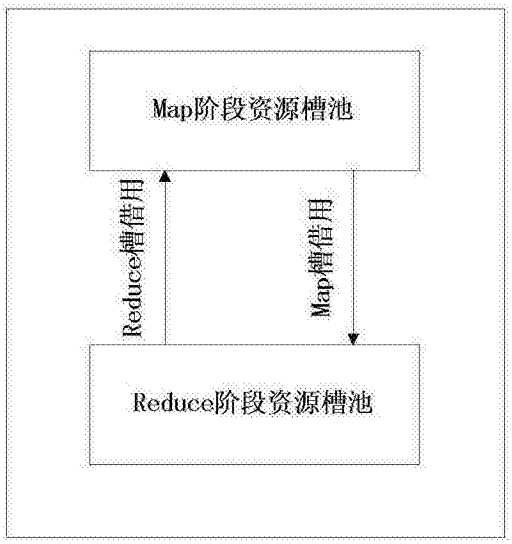

Map阶段资源槽池
Reduce槽借用
Map槽借用
Reduce阶段资源槽池

图 2

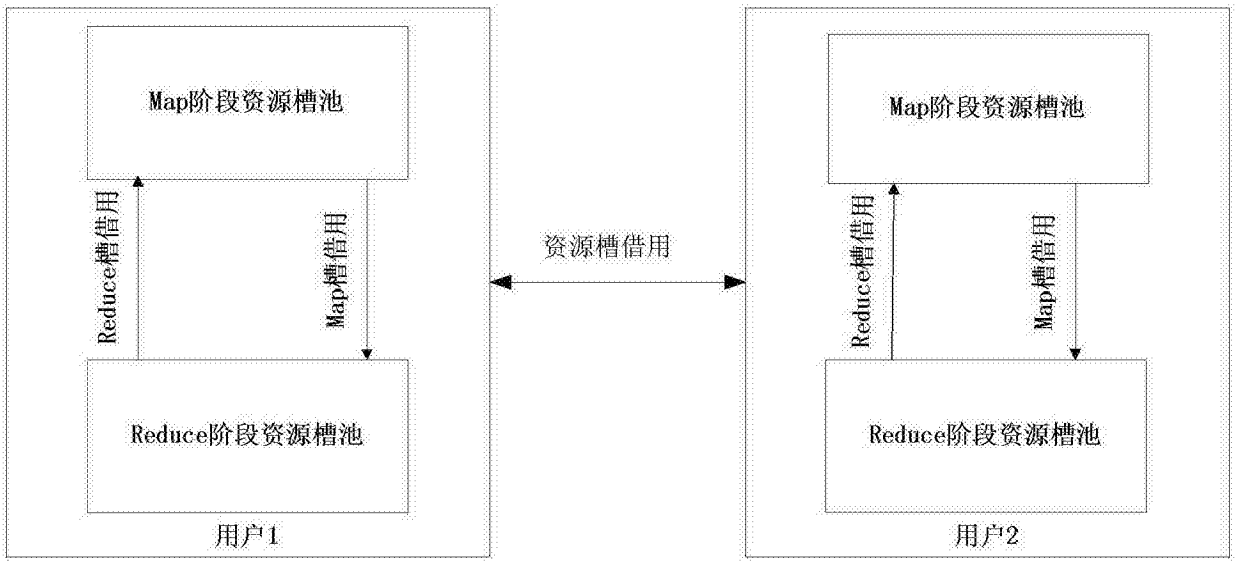

Map阶段资源槽池
Reduce槽借用
Map槽借用
Reduce阶段资源槽池
用户1
资源槽借用
Map阶段资源槽池
Reduce槽借用
Map槽借用
Reduce阶段资源槽池
用户2

图 3

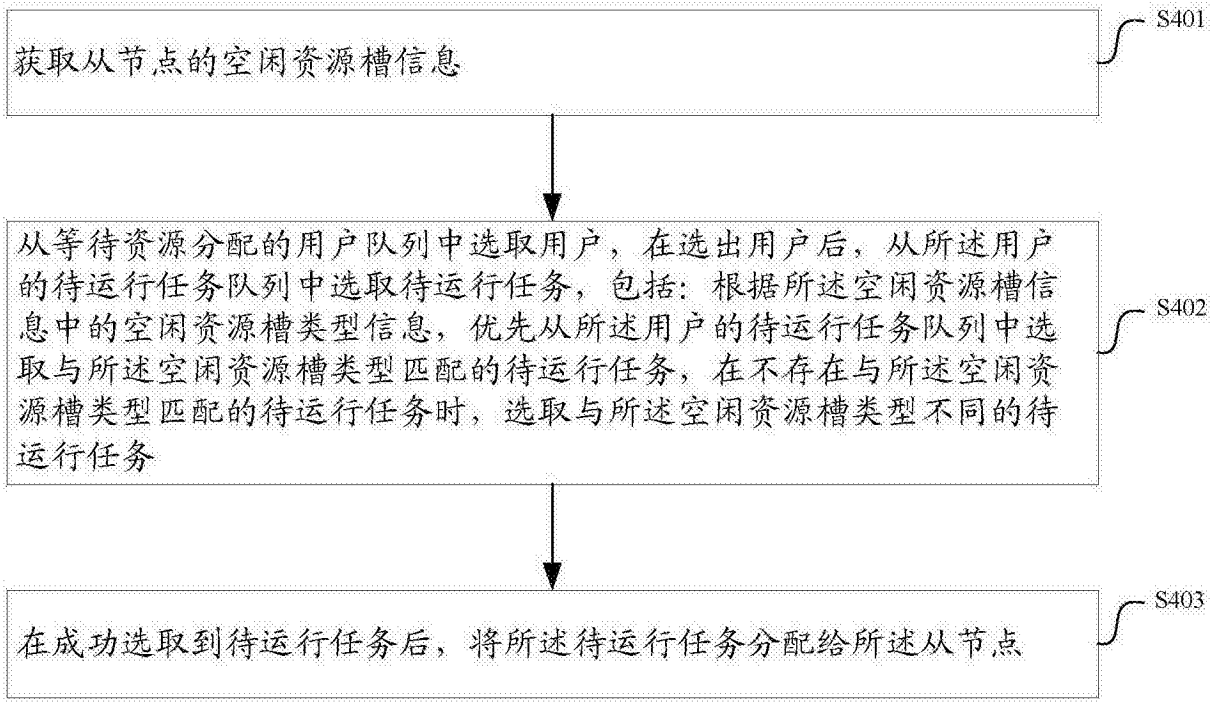

获取从节点的空闲资源槽信息
S401
从等待资源分配的用户队列中选取用户，在选出用户后，从所述用户的待运行任务队列中选取待运行任务，包括：根据所述空闲资源槽信息中的空闲资源槽类型信息，优先从所述用户的待运行任务队列中选取与所述空闲资源槽类型匹配的待运行任务，在不存在与所述空闲资源槽类型匹配的待运行任务时，选取与所述空闲资源槽类型不同的待运行任务
S402
在成功选取到待运行任务后，将所述待运行任务分配给所述从节点
S403

图 4

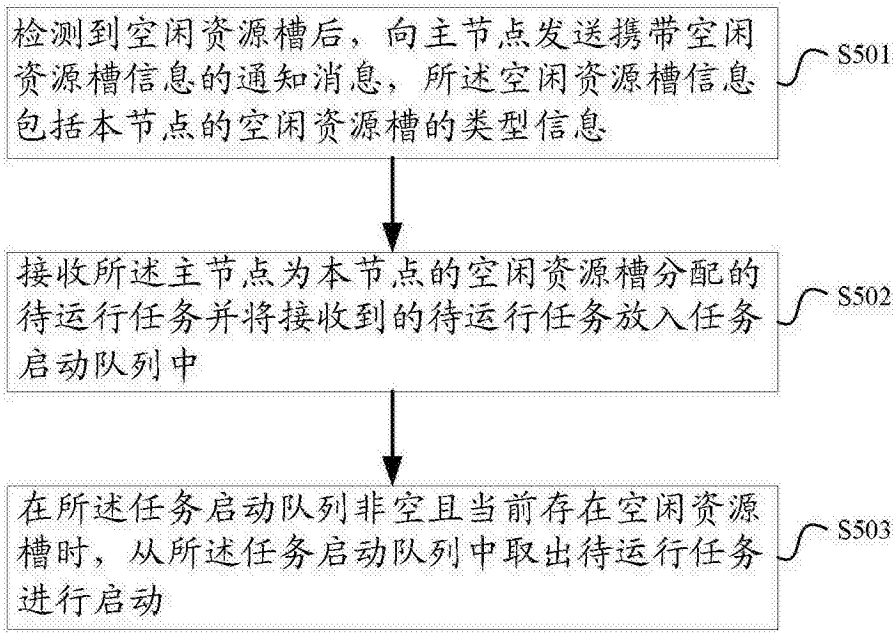

检测到空闲资源槽后，向主节点发送携带空闲资源槽信息的通知消息，所述空闲资源槽信息包括本节点的空闲资源槽的类型信息
S501
接收所述主节点为本节点的空闲资源槽分配的待运行任务并将接收到的待运行任务放入任务启动队列中
S502
在所述任务启动队列非空且当前存在空闲资源槽时，从所述任务启动队列中取出待运行任务进行启动
S503

图 5

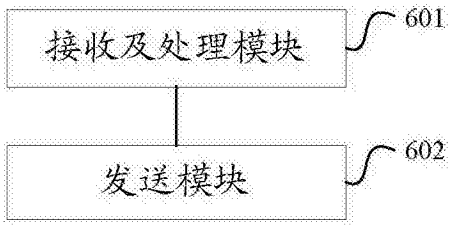

接收及处理模块
601
发送模块
602

图 6

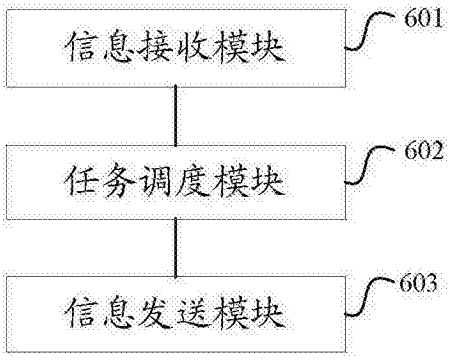

信息接收模块
601
任务调度模块
602
信息发送模块
603

图 7

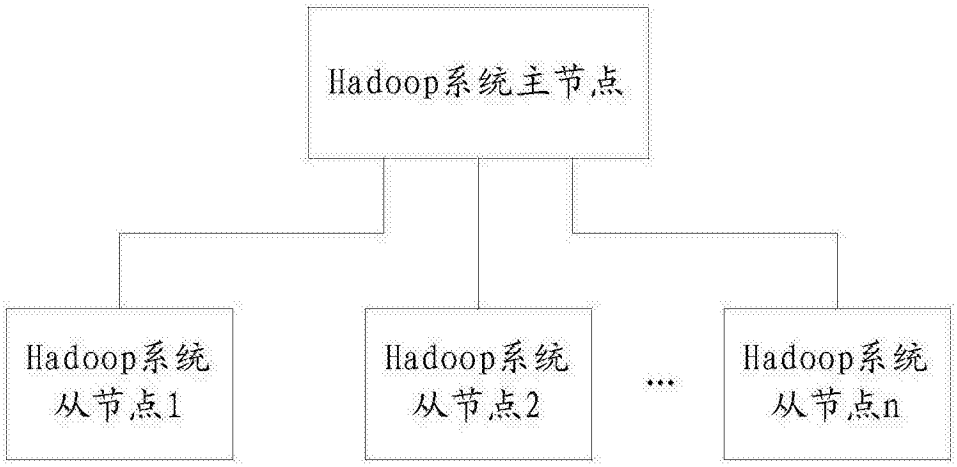


图 8