



(12)发明专利

(10)授权公告号 CN 105573866 B

(45)授权公告日 2018.11.13

(21)申请号 201510997582.X

(22)申请日 2010.07.13

(65)同一申请的已公布的文献号

申请公布号 CN 105573866 A

(43)申请公布日 2016.05.11

(30)优先权数据

12/502,851 2009.07.14 US

(62)分案原申请数据

201080031812.2 2010.07.13

(73)专利权人 起元技术有限责任公司

地址 美国马萨诸塞州

(72)发明人 B·P·多罗斯 M·D·阿特布里

T·瓦克林

(74)专利代理机构 北京林达刘知识产权代理事

务所(普通合伙) 11277

代理人 刘新宇

(51)Int.Cl.

G06F 11/14(2006.01)

(56)对比文件

US 2001042224 A1,2001.11.15,

US 5966072 A,1999.10.12,

US 2008294937 A1,2008.11.27,

CN 101055538 A,2007.10.17,

CN 101038559 A,2007.09.19,

US 2008005227 A1,2008.01.03,

CN 101271472 A,2008.09.24,

US 2004199812 A1,2004.10.07,

US 2005256826 A1,2005.11.17,

US 7206964 B2,2007.04.17,

US 7412520 B2,2008.08.12,

审查员 李妮

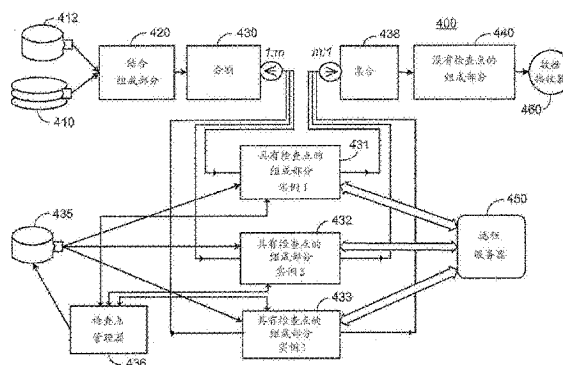
权利要求书3页 说明书7页 附图4页

(54)发明名称

以容错方式处理批量输入数据的方法和系统

(57)摘要

本发明涉及以容错方式处理批量输入数据的方法和系统。处理批量的输入数据包括:读取包括数个记录的批量,并且使所述批量经过数据流程图。至少一个但少于全部的图组成部分包括检查点处理,用于针对与一个或多个记录相关联的数个工作单元中的每一个执行的动作;检查点处理包括:在处理开始时,打开检查点缓存。如果由于对工作单元执行所述动作所产生的结果在之前被保存在检查点缓存中,则使用所保存的结果完成工作单元的处理而不再次执行所述动作。如果由于对工作单元执行所述动作所产生的结果未被保存在检查点缓存中,则执行所述动作以完成工作单元的处理并且将由于执行所述动作所产生的结果保存在检查点缓存中。



1. 一种用于以容错方式处理批量输入数据的方法,所述方法由包括存储器的一个或多个计算机系统执行,所述方法包括:

对所述批量输入数据执行计算,其中所述计算中的至少一个但少于全部的计算包括用于与所述批量输入数据相关联的数个工作单元的检查点处理,

其中,所述检查点处理包括:

对于来自所述批量输入数据的工作单元,

如果对所述工作单元执行动作所产生的结果在之前被保存在存储器中所存储的检查点缓存中,则使用所保存的结果完成对所述工作单元执行所述计算而不再次执行所述动作;或者

如果对所述工作单元执行所述动作所产生的结果未被保存在检查点缓存中,则执行所述动作以完成对所述工作单元执行所述计算,并且将执行所述动作所产生的结果保存在检查点缓存中。

2. 根据权利要求1所述的方法,其中,所述动作包括与远程服务器通信。

3. 根据权利要求2所述的方法,其中,执行所述动作所产生的结果包括所述工作单元的与远程服务器的通信所产生的信息。

4. 根据权利要求1所述的方法,其中,还包括:在完成所述批量输入数据的处理时删除检查点缓存。

5. 根据权利要求2所述的方法,其中,与远程服务器的通信被收费。

6. 根据权利要求2所述的方法,其中,与远程服务器的通信的结果存储在易失性存储器中,并且在发生触发事件时按组保存至检查点缓存。

7. 根据权利要求6所述的方法,其中,所述触发事件是来自检查点管理器的信号。

8. 根据权利要求6所述的方法,其中,所述触发事件是自从上次写至检查点缓存时起数个记录的处理。

9. 根据权利要求6所述的方法,其中,所述触发事件是自从上次写至检查点缓存时起已经经过时间段。

10. 根据权利要求1所述的方法,其中,所述检查点处理并行地在多个处理设备上运行。

11. 根据权利要求10所述的方法,其中,所述批量输入数据包括数据记录,以及所述数据记录在多个并行处理设备之中的分配在所述批量输入数据的运行之间是一致的,并且每个并行处理设备保持独立的检查点缓存。

12. 根据权利要求10所述的方法,其中,所述批量输入数据包括数据记录,以及所述数据记录在多个并行处理设备之中的分配是动态的,并且处理设备共享对于所共享的非易失性存储器中存储的单个检查点缓存的访问,其中对于检查点缓存的写入由检查点管理器控制。

13. 根据权利要求1所述的方法,其中,还包括:

在已经发生错误状况之后,重新开始处理;

读取包括多个记录的所述批量输入数据;以及
处理整个所述批量输入数据。

14. 根据权利要求13所述的方法,其中,所述动作包括与远程服务器通信。

15. 一种用于以容错方式处理批量输入数据的系统,所述系统包括:

用于对所述批量输入数据执行计算的部件,其中所述计算中的至少一个但少于全部的计算包括用于与所述批量输入数据相关联的数个工作单元的检查点处理;

其中,所述检查点处理包括:

对于来自所述批量输入数据的工作单元,

如果对上述工作单元执行动作所产生的结果在之前被保存在存储器中所存储的检查点缓存中,则使用所保存的结果完成对上述工作单元执行所述计算而不再次执行所述动作;或者

如果对上述工作单元执行所述动作所产生的结果未被保存在检查点缓存中,则执行所述动作以完成对上述工作单元执行所述计算,并且将执行所述动作所产生的结果保存在检查点缓存中。

16. 一种用于以容错方式处理批量输入数据的系统,所述系统包括:

一个或多个计算机;以及

一个或多个存储装置,用于存储指令,所述指令在由所述一个或多个计算机执行的情况下,能够用于使得所述一个或多个计算机进行操作,所述操作包括:

对所述批量输入数据执行计算,其中所述计算中的至少一个但少于全部的计算包括用于与所述批量输入数据相关联的数个工作单元的检查点处理,

其中,所述检查点处理包括:

对于来自所述批量输入数据的工作单元,

如果对上述工作单元执行动作所产生的结果在之前被保存在存储器中所存储的检查点缓存中,则使用所保存的结果完成对上述工作单元执行所述计算而不再次执行所述动作;或者

如果对上述工作单元执行所述动作所产生的结果未被保存在检查点缓存中,则执行所述动作以完成对上述工作单元执行所述计算,并且将执行所述动作所产生的结果保存在检查点缓存中。

17. 根据权利要求16所述的系统,其中,所述操作还包括:

在已经发生错误状况之后,重新开始处理;

获得包括多个记录的所述批量输入数据;以及

处理整个所述批量输入数据。

18. 一种计算机可读硬件存储设备,其存储用于以容错方式处理批量输入数据的计算机程序,所述计算机程序包括使计算机执行以下操作的指令:

对所述批量输入数据执行计算,其中所述计算中的至少一个但少于全部的计算包括用于与所述批量输入数据相关联的数个工作单元的检查点处理,

其中,所述检查点处理包括:

对于来自所述批量输入数据的工作单元,

如果对上述工作单元执行动作所产生的结果在之前被保存在存储器中所存储的检查点缓存中,则使用所保存的结果完成对上述工作单元执行所述计算而不再次执行所述动作;或者

如果对上述工作单元执行所述动作所产生的结果未被保存在检查点缓存中,则执行所述动作以完成对上述工作单元执行所述计算,并且将执行所述动作所产生的结果保存在检

查点缓存中。

19. 根据权利要求18所述的计算机可读硬件存储设备, 其中, 所述动作包括与远程服务器通信。

20. 根据权利要求19所述的计算机可读硬件存储设备, 其中, 执行所述动作所产生的结果包括所述工作单元的与远程服务器的通信所产生的信息。

21. 根据权利要求19所述的计算机可读硬件存储设备, 其中, 与远程服务器的通信被收费。

22. 根据权利要求19所述的计算机可读硬件存储设备, 其中, 与远程服务器的通信的结果存储在易失性存储器中, 并且在发生触发事件时按组保存至检查点缓存。

23. 根据权利要求22所述的计算机可读硬件存储设备, 其中, 所述触发事件是来自检查点管理器的信号。

24. 根据权利要求22所述的计算机可读硬件存储设备, 其中, 所述触发事件是自从上次写至检查点缓存时起数个记录的处理。

25. 根据权利要求22所述的计算机可读硬件存储设备, 其中, 所述触发事件是自从上次写至检查点缓存时起已经经过时间段。

26. 根据权利要求18所述的计算机可读硬件存储设备, 其中, 所述指令还使所述计算机执行: 在完成所述批量输入数据的处理时删除检查点缓存。

27. 根据权利要求18所述的计算机可读硬件存储设备, 其中, 所述检查点处理并行地在多个处理设备上运行。

28. 根据权利要求27所述的计算机可读硬件存储设备, 其中, 所述批量输入数据包括数据记录, 以及所述数据记录在多个并行处理设备之中的分配在所述批量输入数据的运行之间是一致的, 并且每个并行处理设备保持独立的检查点缓存。

29. 根据权利要求27所述的计算机可读硬件存储设备, 其中, 所述批量输入数据包括数据记录, 以及所述数据记录在多个并行处理设备之中的分配是动态的, 并且处理设备共享对于所共享的非易失性存储器中存储的单个检查点缓存的访问, 其中对于检查点缓存的写入由检查点管理器控制。

30. 根据权利要求18所述的计算机可读硬件存储设备, 其中, 所述指令还使所述计算机执行:

在已经发生错误状况之后, 重新开始处理;

获得包括多个记录的所述批量输入数据; 以及

处理整个所述批量输入数据。

31. 根据权利要求30所述的计算机可读硬件存储设备, 其中, 所述动作包括与远程服务器通信。

以容错方式处理批量输入数据的方法和系统

[0001] (本申请是申请日为2010年7月13日、申请号为201080031812.2、发明名称为“容错的批量处理”的申请的分案申请)

技术领域

[0002] 此描述涉及以容错方式处理批量的数据。

背景技术

[0003] 复杂的计算通常可以表示为有向图(称为“数据流程图(dataflow graph)”),其中计算的各组成部分与图的节点(或顶点)相关联并且各组成部分之间的数据流对应于图的各节点之间的链路(link)(或,弧(arc)、边缘(edge))。组成部分包括处理数据的数据处理组成部分和用作数据流的源(source)或接收器(sink)的组成部分。各数据处理组成部分形成可以并发地处理多级的数据的流水线系统。美国专利5,966,072,EXECUTING COMPUTATIONS EXPRESSED AS GRAPHS中描述了实施这种基于图的计算(graph-based computation)的系统。在某些情况下,将基于图的计算配置为接收输入数据的流并且处理连续的数据流以便无限地提供来自一个或更多个组成部分的结果,直到计算关闭为止。在某些情况下,将基于图的计算配置为接收批量的输入数据并且处理批量的数据以提供对于此批量的结果。然后在已经处理了该批量之后关闭或返回到空闲的状态。

发明内容

[0004] 在一方面,总体上,用于以容错方式处理批量输入数据的方法包括:从一个或更多个数据源读取包括多个记录的批量输入数据;以及使所述批量经过包括由链路连接的代表组成部分的两个或更多个节点的数据流程图,所述链路代表各组成部分之间的数据的流,其中,至少一个但少于全部的组成部分包括检查点处理,用于针对与一个或更多个记录相关联的数个工作单元中的每一个执行的动作。所述检查点处理包括:在针对批量的处理开始时,打开非易失性存储器中存储的检查点缓存;以及对于来自所述批量的每个工作单元,如果由于对工作单元执行所述动作所产生的结果在之前被保存在检查点缓存中,则使用所保存的结果完成工作单元的处理而不再次执行所述动作,或者如果由于对工作单元执行所述动作所产生的结果未被保存在检查点缓存中,则执行所述动作以完成工作单元的处理并且将由于执行所述动作所产生的结果保存在检查点缓存中。

[0005] 一些方面可以包括下列特征中的一个或更多个。

[0006] 所述动作包括与远程服务器通信。

[0007] 由于执行所述动作所产生的结果包括由于用于工作单元的与远程服务器的通信所产生的信息。

[0008] 所述方法进一步包括:在完成所述批量的处理时删除检查点缓存。

[0009] 与远程服务器的通信被收费。

[0010] 与远程服务器的通信的结果存储在易失性存储器中,并且在发生触发事件时按组

保存至检查点缓存。

[0011] 所述触发事件是来自检查点管理器的信号。

[0012] 所述触发事件是自从上次写至检查点缓存时起记录的数目的处理。

[0013] 所述触发事件是自上次写至检查点缓存时起时间段的经过。

[0014] 包括检查点处理的组成部分并行地在多个处理设备上的运行。

[0015] 数据记录在多个并行处理设备之中的分配在批量的运行之间是一致的,并且每个处理设备保持独立的检查点缓存。

[0016] 数据记录在多个并行处理设备之中的分配是动态的,并且处理设备共享对于所共享的非易失性存储器中存储的单个检查点缓存的访问,其中对于检查点缓存的写入由检查点管理器控制。

[0017] 所述方法进一步包括:在已经发生错误状况时,重新开始数据流程图中的所有组成部分;从一个或多个数据源读取包括多个记录的批量输入数据;以及使整个批量经过数据流程图。

[0018] 所述动作包括与远程服务器通信。

[0019] 在另一方面,总体上,计算机可读介质存储用于以容错方式处理批量输入数据的计算机程序。该计算机程序包括指令,所述指令使得计算机:从一个或多个数据源读取包括多个记录的批量输入数据;以及使所述批量经过包括由链路连接的代表组成部分的两个或多个节点的数据流程图,所述链路代表各组成部分之间的数据的流,其中,至少一个但少于全部的组成部分包括检查点处理,用于针对与一个或多个记录相关联的数个工作单元中的每一个执行的动作。所述检查点处理进一步包括:在针对批量的处理开始时,打开非易失性存储器中存储的检查点缓存;以及对于来自所述批量的每个工作单元,如果由于对工作单元执行所述动作所产生的结果在之前被保存在检查点缓存中,则使用所保存的结果完成工作单元的处理而不再次执行所述动作,或者如果由于对工作单元执行所述动作所产生的结果未被保存在检查点缓存中,则执行所述动作以完成工作单元的处理并且将由于执行所述动作所产生的结果保存在检查点缓存中。

[0020] 在另一方面,总体上,用于以容错方式处理批量输入数据的系统包括:用于从一个或多个数据源读取包括多个记录的批量输入数据的部件;以及用于使所述批量经过包括两个或多个节点的数据流程图的部件,所述节点代表由链路连接的组成部分,所述链路代表各组成部分之间的数据的流,其中,至少一个但少于全部的组成部分包括检查点处理,用于针对与一个或多个记录相关联的数个工作单元中的每一个执行的动作。所述检查点处理包括:在针对批量的处理开始时,打开非易失性存储器中存储的检查点缓存;以及对于来自所述批量的每个工作单元,如果由于对工作单元执行所述动作所产生的结果在之前被保存在检查点缓存中,则使用所保存的结果完成工作单元的处理而不再次执行所述动作,或者如果由于对工作单元执行所述动作所产生的结果未被保存在检查点缓存中,则执行所述动作以完成工作单元的处理并且将由于执行所述动作所产生的结果保存在检查点缓存中。

[0021] 一些方面可以包括下列优点中的一个或多个:可以消除对于数据流程图中不同组成部分之间的一些检查点有关通信的需要。可以选择性地避免在错误恢复期间在多级批量处理中重复复杂和高成本的步骤,而没有实施整个流水线的系统的检查点的复杂性和花

费。例如,通过避免对于收费服务的重复调用,此方法可以用于省钱。

[0022] 本发明的其它特征和优点将根据下面的描述并且根据权利要求书而变得明显。

附图说明

[0023] 图1是具有输入/输出检查点的批量数据处理系统的方框示意图。

[0024] 图2是检查点处理的流程图。

[0025] 图3是具有拥有并行性的输入/输出检查点的批量数据处理系统的方框示意图。

[0026] 图4是具有拥有并行性的输入/输出检查点的批量数据处理系统和检查点管理器的方框示意图。

具体实施方式

[0027] 基于图的数据处理系统可以配置为以容错 (fault tolerant) 方式处理批量的输入数据,其包含:将数据流程图中的一个组成部分的中间结果保存至缓存 (buffer),其中它们可以从该缓存中检索到并且在错误状况强制批量输入数据的处理的重新启动的情况下被重新使用。

[0028] 图1是示例性的数据处理系统100的方框示意图。数据经过数据流程图的数据处理组成部分的序列,所述数据流程图处理从一个或多个数据源 (data source) 至一个或多个数据接收器 (data sink) 的数据流。数据流程图中的各种数据处理组成部分中的任何一个均可以通过在单独的处理设备上运行的处理而加以实施,或者数个数据处理组成部分可以通过在单个处理设备上运行的一个或多个处理而加以实施。可以按批量处理数据,其标识要由系统100处理的一组输入数据记录。

[0029] 系统100对于批量数据的处理可以通过用户输入或者某种其它事件 (如,定时器的到期) 而启动。当开始批量数据的处理时,从一个或多个输入数据源读取输入数据记录。例如,从诸如数据存储组成部分110表示的计算机可读存储器件上存储的一个或多个文件读取输入数据。也可以从诸如数据存储组成部分112表示的服务器上运行的数据库读取输入数据记录。结合组成部分120按顺序从数个数据源读取数据 (例如,记录),并且将输入数据安排为一序列的离散工作单元 (work unit)。工作单元例如可以代表基于输入记录以预定格式存储的记录,或者例如可以代表要处理的事务 (transaction)。在某些实施方案中,每个工作单元可以由批量内唯一的编号 (如,处理的工作单元的计数) 来标识。然后按顺序将工作单元传给数据流程图中的下一组成部分。

[0030] 实施系统100的示例性数据流程图还包括数据处理组成部分130和140。数据处理组成部分130包括检查点处理,其在批量处理过程期间将关于其处理的状态信息有规律地保存至非易失性存储器。当出现错误状况并且必须重新启动批量时,检查点的组成部分130访问所存储的状态信息以降低在批量的重复运行期间必须重复的处理的量。由此,检查点以使用非易失性存储器资源和向数据处理组成部分130添加复杂性为代价提供了容错。数据处理组成部分140是没有检查点的组成部分。其它数据流程图可以包括更多或更少的数据处理组成部分。可以将所需要那么多的数据处理组成部分配置为包括检查点处理。典型地,将在延迟或某些其它度量方面具有高成本的组成部分配置为包括检查点处理,从而在错误状况的情况下,无需针对批量中的所有工作单元重复系统100中的高成本处理步骤。

[0031] 数据处理组成部分130包括访问远程服务器150的步骤。对于处理的每个工作单元,第一处理组成部分130将把请求发送至远程服务器150并且接收来自该远程服务器的结果(例如,来自数据库的数据)。这种操作出于各种原因而可能是高成本的,所述各种原因包括:在与远程服务器通信时经历的网络延迟或者远程控制器提供的服务的收费。在接收到结果后,组成部分130为下一数据处理组成部分140产生输出。由于此组成部分130已配置为包括检查点处理,其通过将工作单元的输出传递至下一数据处理组成部分140并且开始下一工作单元的处理,以保存来自远程服务器150的结果作为完成处理之前的处理状态信息的一部分。处理状态信息可以暂存在运行检查点处理的处理设备上的易失性存储器中。以有规律的时间将一个或更多个工作单元的处理状态信息写入至非易失性存储器中所存储的检查点缓存,从而其稍后在错误状况的情况下是可用的。

[0032] 随着工作单元通过数据流程图的各数据处理组成部分,与每个工作单元相关联的最终结果被传输至数据接收器160。可以单独地传输工作单元,或者在某些实施方案中,工作单元可以用于递增地更新最终结果,或者可以在最终结果传输至数据接收器160之前(例如,按队列)予以累积。数据接收器160可以是存储工作单元或者基于工作单元的一些累积输出的数据存储组成部分,例如,或者数据接收器160可以是工作单元被公布到的队列或者用于接收最终结果的某些其它类型的接收器。批量处理在批量中的所有工作单元的结果已被传输至数据接收器160时结束。此时,可终止数据流程图中的组成部分。与检查点的组成部分相关联的检查点处理可以删除其检查点缓存作为其终止例程的一部分。

[0033] 图2是用于对检查点的组成部分进行检查点的示例性处理200的流程图。例如,一旦来自于通过数据流程图实施批量处理的软件的外部调用时,处理200启动201。启动可以包括:为检查点的组成部分在其上运行的处理设备上的处理200分配易失性存储器,并且保留任何其它所需要的资源。处理200接着检查205是否已经将与此处处理相关联的检查点缓存保存在非易失性存储器中。如果不存在检查点缓存,则在非易失性存储器中创建207新的检查点缓存。如果之前存储了检查点缓存,则其被打开208。打开208检查点缓存可以包括:发现非易失性存储器中检查点缓存的位置,或者可能将检查点缓存的所有或部分复制到处理设备上的易失性存储器。

[0034] 在用于应对每个工作单元的循环开始时,从数据流程图中的之前组成部分或者从源接收210与工作单元相关联的输入数据。可选地,对工作单元进行预处理220。预处理220例如可包括:重新格式化数据记录,或者确定可用于针对与工作单元相关联的结果搜索检查点缓存的值。检查点处理200的检查点缓存被检查225以(例如,根据被中断的批量的之前处理)确定此工作单元的结果是否存储在检查点缓存中。

[0035] 如果相关联的结果未存储在检查点缓存中,则对工作单元进行包括高成本动作的处理230。高成本动作的示例可包括:通过网络访问远程服务器上的资源并且遭遇显著的延迟或收取费用。然后将此处理的结果存储240在检查点缓存中。例如可以使用递增计数器将此结果与正在处理的工作单元相关联,所述递增处理器可以通过相同的计数器值标识工作单元和其相关联的结果。该结果可以直接写入至非易失性存储器,或者可以暂时缓存在易失性存储器中,直到触发事件使得其被复制至非易失性存储器为止。示例性的触发事件包括:处理固定数量的工作单元、经过的时间段或者来自外部处理的信号。

[0036] 如果相关联的结果存储在检查点缓存中,则从检查点缓存中检索250该结果。

[0037] 可选地,进行后处理260以完成工作单元的处理。后处理260例如可以包括:重新格式化数据或者将数据传递至数据流程图中的下一组成部分。在工作单元的处理完成后,检查点处理200接着检查270是否另一工作单元有待处理。如果另一工作单元是可用的,则检查点处理200循环回至读取与下一工作单元相关联的输入数据。当没有更多的工作单元有待处理时,工作点处理200等待280指示批量处理已经完成并且指令其终止的外部信号。当接收到终止信号时,检查点处理200在完成其终止序列290之前从非易失性存储器中删除285其检查点缓存。完成终止序列290可以包括:释放处理设备或其它保留的资源上的易失性存储器。

[0038] 图3是示例性的数据处理系统300的方框示图,其中实施系统300的数据流程图包括具有分布式检查点处理的并行组成部分。数据流程图中的一个或多个组成部分可以并行地运行在数个处理设备(例如,数个计算机或并行处理器的处理器核或数个处理器)上。在此示例中,明确地描绘了检查点并行组成部分的数个实例331、332和333。并行组成部分的实例运行在每个处理设备上,并且每个实例处理批量的工作单元的子集。在分布式检查点办法的此示例中,针对并行组成部分的三个实例中的每一个运行不同的检查点处理。

[0039] 当开始批量数据的处理时,从一个或多个输入数据源读取输入数据记录。例如,可以从诸如数据存储组成部分310表示的计算机可读存储设备上存储的一个或多个文件读取输入数据。也可以从在诸如数据存储组成部分312表示的服务器上运行的数据库读取输入数据记录。结合组成部分320按顺序从数个数据源读取数据,并且将输入数据安排到一序列的离散工作单元。工作单元按顺序传递至数据流程图中的下一组成部分。

[0040] 由于数据流程图中的下一数据处理组成部分是并行组成部分,因此工作单元由数据单元分割组成部分330分割并分配至数个组成部分实例。在此示例中,工作单元在各实例之中的分配在不同的批量处理运行之间一致(consistent),从而各实例无需访问被分配给其它实例的工作单元的状态信息。在出现错误状况并且需要再次运行批量时,工作单元分割组成部分330基于利用一致的结果可重复的一致算法,将工作单元分配至特定的实例。例如,工作单元分配分割组成部分330继而可以只是将工作单元每次一个地分配给每个组成部分实例,在工作单元计数超过并行实例的数目时循环至第一实例。在另一示例中,工作单元分割组成部分330可以应用这样的分割算法:其未保证在运行之间产生一致分配并且将分配信息保存至非易失性存储器,从而在需要批量的重复运行的情况下可以重复相同的分配。

[0041] 检查点并行组成部分的每个实例331、332和333使用结合图1的检查点组成部分130所描述的方法,独立地处理向其分配的工作单元。每个实例331、332和333在非易失性存储器中创建并保持其自身的检查点缓存。当处理了工作单元时,实例检查其自身的检查点缓存以确定之前是否已经在批量的先前运行期间处理了工作单元。在示例性的系统300中,检查点的并行组成部分包括与远程服务器350通信以获取每个工作单元的信息的动作。在其它示例中,检查点并行组成部分可以包括具有与其相关联的高成本的其它动作,其为容错判断检查点缓存的维护(maintenance)。

[0042] 当工作单元的处理结束时,将结果传递至聚集组成部分338,该聚集组成部分338收集来自数个实例的结果并且将它们传递至数据流程图中的下一数据处理组成部分。

[0043] 数据处理组成部分340是没有检查点的组成部分。在其它示例中,数据流程图中的

任何数目的组成部分均可以包括检查点。在某些情况下,将检查点处理限制到进行高成本动作的组成部分是有益的。其它数据流程图可以包括更多或更少的数据处理组成部分,其对于任何给定的数据处理组成部分具有或没有并行性(parallelism)。

[0044] 随着工作单元通过数据流程图的各组成部分,与每个工作单元相关联的最终结果被传输至数据接收器360。批量处理在批量中的所有工作单元的结果已被传输至数据接收器360时结束。此时,可终止数据流程图中的各组成部分相关联的处理。给定实例的检查点处理可以删除其检查点缓存作为其终止例程的一部分。

[0045] 图4是示例性的数据处理系统400的方框示图,其中实施系统400的数据流程图包括具有集中式检查点处理的并行组成部分。在此示例中,明确地描绘了检查点并行组成部分的数个实例431、432和433。并行化的组成部分的实例运行在每个处理设备上,并且每个实例处理批量的工作单元的子集。在集中式检查点办法的此示例中,检查点管理器436应对与并行组成部分的三个实例中的每一个通信的至少一些检查点处理。可以在运行并行组成部分的实例的各处理设备中的一个处理设备上或者在单独的处理设备上运行检查点管理器436。

[0046] 当开始批量数据的处理时,从数据存储组成部分410和412读取输入数据记录。结合组成部分420按顺序从数个数据源读取数据,并且将输入数据安排到所存储的一序列的离散工作单元。然后将工作单元按顺序传递至数据流程图中的下一组成部分,其在此示例中是检查点并行组成部分。

[0047] 在图4的示例中,检查点管理器436控制对于由实例431、432和433(其每一个均运行在不同的处理设备上)共享的单个检查点缓存的访问。对于批量中的所有工作单元共享单个检查点缓存允许将工作单元动态地分配至各实例,而不需要从批量的之前运行匹配分配。共享的检查点缓存存储在所有的实例可以经由总线或通信网络直接地、或者经由与检查点管理器436的通信间接地访问的所共享的非易失性存储器435上。实例431、432、433可以在它们处理工作单元时读取所共享的非易失性存储器435以检查检查点缓存。如果在检查点缓存中发现当前工作单元的结果,则所存储的结果用以避免重复高成本动作。如果在检查点缓存中未发现当前工作单元的结果,则执行工作单元的动作并且将结果存储在检查点缓存中。为了写入至检查点缓存,实例431、432、433将写请求消息发送至检查点管理器436。检查点管理器436然后向共享的非易失性存储器435写入以更新检查点缓存。在可替换的实施例中,检查点管理器436将记号(token)发送至给出向共享的非易失性存储器435写入的许可的请求实例,以便更新检查点缓存。

[0048] 由于所有的实例431、432、433使用共享的检查点缓存,因此在批量数据的每一运行期间,工作单元分割组成部分430可以以不同的方式在各实例之间动态地分配工作单元。例如,工作单元分割组成部分430可以基于运行时在每个处理设备上的可用容量(其可能在运行之间变化)动态地分配每个工作单元。这种方法还允许工作单元分割组成部分430使用不同数量的并行实例。例如,在错误状况之后,运行并行组成部分的实例(例如,实例433)的处理设备之一可以被禁止或者不可用。在批量重新启动的这种情况下,工作单元分割组成部分430可以将所有的工作单元分配至剩余的实例431、432,其可以无缝地访问由禁止的实例433之前处理的工作单元的检查点缓存入口。

[0049] 检查点管理器436可以由在单独的处理设备上运行的处理来实施,或者其可以由

在运行并行组成部分的实例的处理设备之一上运行的处理来实施。在检查点缓存更新事件之间,实例431、432、433可以将检查点缓存更新缓存在本地的易失性存储器中。检查点管理器436可以将信号发送至实例,其触发实例以易失性存储器中缓存的任何信息开始检查点缓存更新。

[0050] 当工作单元的处理完成时,将结果传递至聚集组成部分438,该聚集组成部分438收集来自数个实例的结果并且将它们传递至数据流程图中的下一数据处理组成部分。

[0051] 数据处理组成部分440是没有检查点的组成部分。在其它示例中,数据流程图中的任何数量的组成部分可以包括检查点。在某些情况下,将检查点处理限制到要进行高成本动作的组成部分是有益的。其它数据流程图可以包括更多或更少的处理组成部分,其对于任何给定的数据处理组成部分具有或没有并行性(parallelism)。

[0052] 随着工作单元通过数据流程图的各组成部分,与每个工作单元相关联的最终结果被传输至数据接收器460。批量处理在批量中的所有工作单元的结果已被传输至数据接收器460时结束。此时,可终止数据流程图中的组成部分。检查点管理器436可以删除检查点缓存作为其终止例程的一部分。

[0053] 上面描述的容错批量处理办法可以通过使用用于在计算机上运行的软件来实施。例如,软件将过程形成在一个或更多个计算机程序中,所述程序运行在一个或更多个编程的或可编程的计算机系统(其可以是诸如分布式、客户端/服务器或网格(grid)之类的各种架构)上,每个计算机系统包括至少一个处理器、至少一个数据存储系统(包括易失性和非易失性存储器和/或存储元件)、至少一个输入设备或端口和至少一个输出设备或端口。软件例如可以形成更大程序的一个或更多个模块,其提供计算图(computation graph)的配置和设计有关的其它服务。图的节点和要素可以实施为计算机可读介质中所存储的数据结构或者遵循数据仓库(data repository)中存储的数据模型的其它有组织的数据。

[0054] 可以将软件提供在可由通用或专用可编程计算机读取的CD-ROM之类的存储介质上,或者经由网络的通信介质输送(以传播的信号编码)至其被运行的计算机。可以在专用的计算机上或者使用专用硬件(如,协处理器)执行所有的功能。可以以由不同的计算机执行软件所指定的计算的不同部分的分布式方式实施软件。最好将每个这种计算机程序存储于或者下载至可由通用或专用可编程计算机读取的存储介质或器件(例如,固态存储器或介质,或者磁或光介质),用于在计算机系统读取存储介质或器件时配置和操作计算机以执行在此描述的过程。还可以考虑将本发明的系统实施为以计算机程序配置的计算机可读存储介质,其中如此配置的存储介质使得计算机系统以特定和预定义的方式工作以执行在此描述的功能。

[0055] 已经描述了本发明的多个实施例。然而,要理解的是,在不脱离本发明精神和范围的情况下可以进行各种变型。例如,上面描述的步骤中的一些可以是有序独立的,由此可以以与描述的顺序不同的顺序加以执行。

[0056] 要理解,前面的描述旨在说明而不是要限制本发明的范围,本发明的范围由所附权利要求书的范围限定。例如,可以在实质上不影响总体处理的情况下以不同的顺序执行上面描述的多个功能步骤。其它实施例在所附权利要求书的范围内。

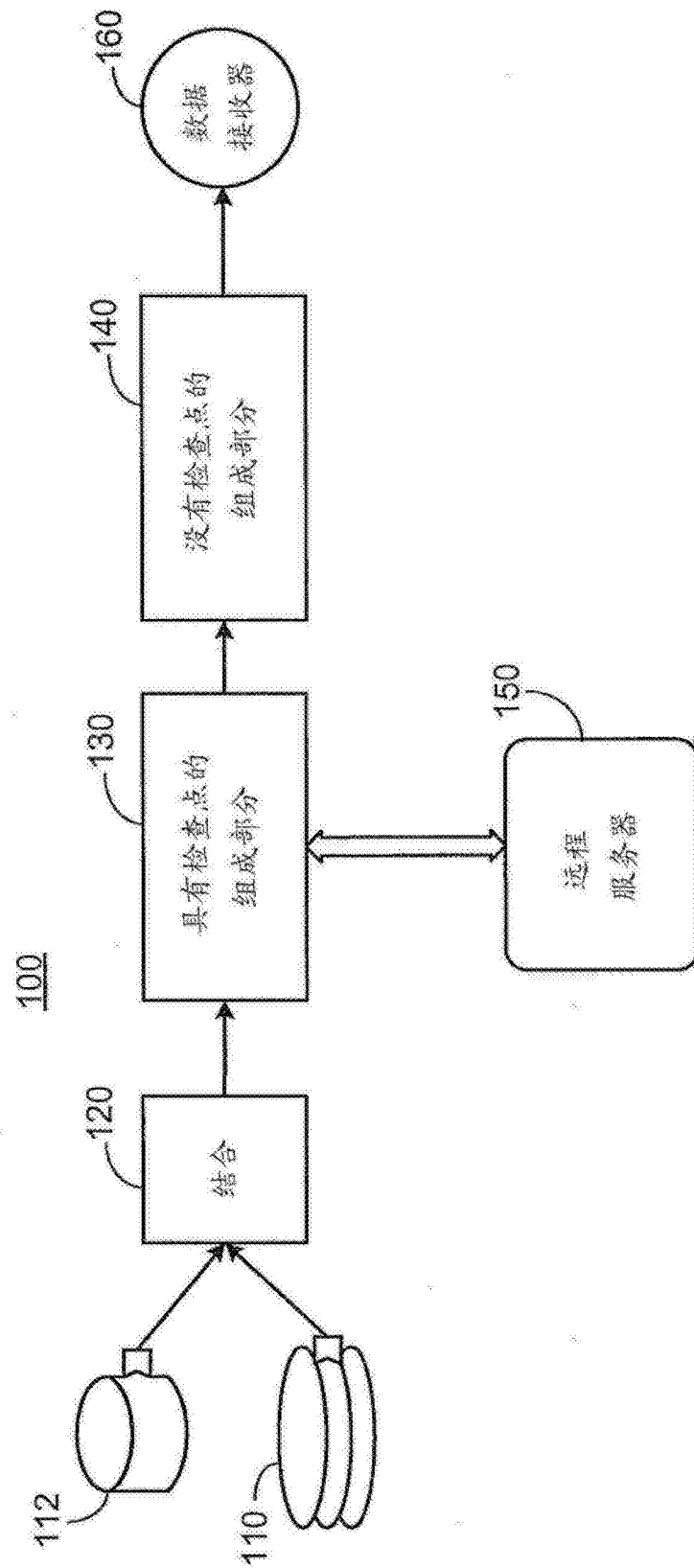


图1

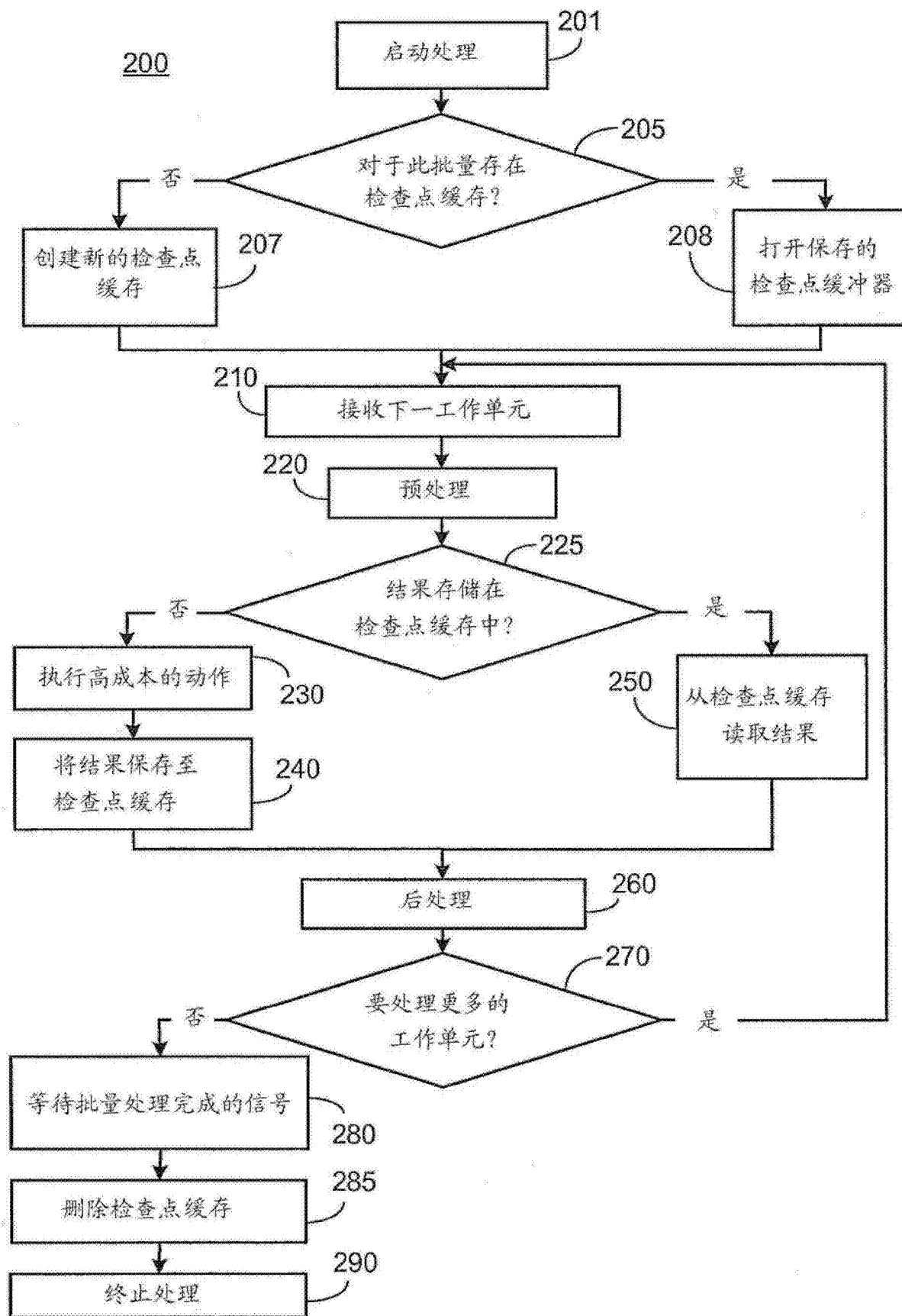


图2

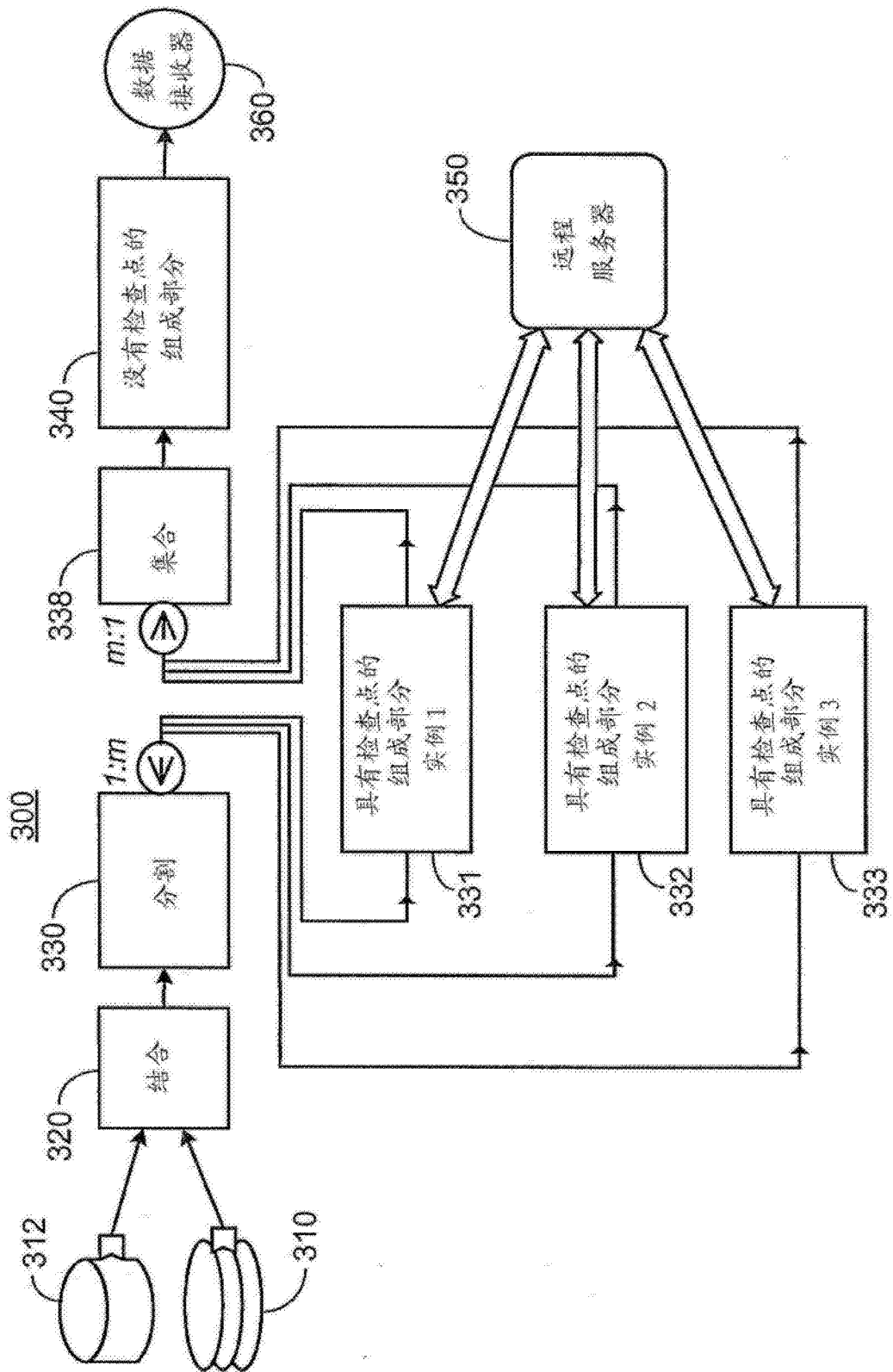


图3

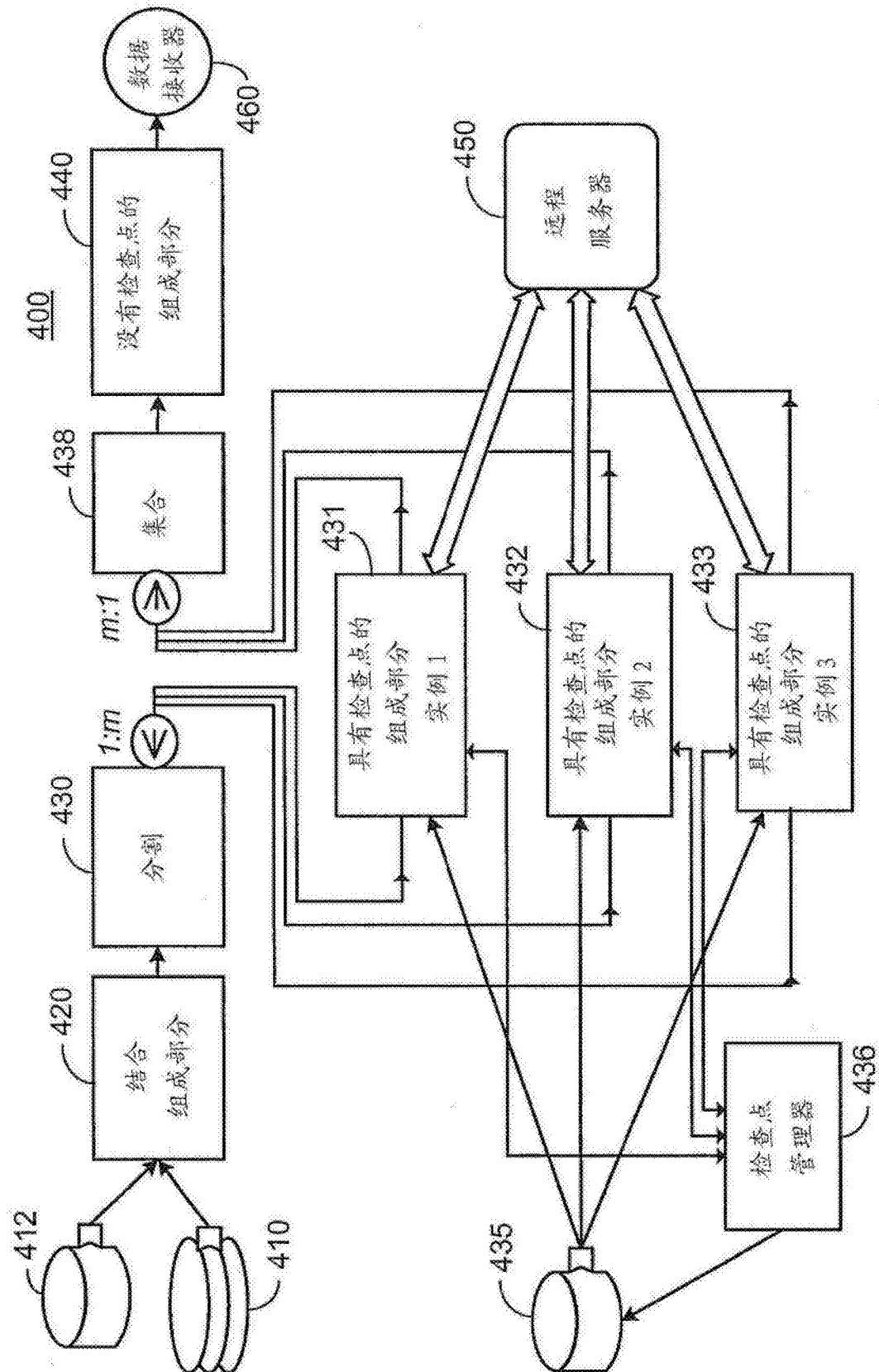


图4