(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 111316241 A
(43)申请公布日 2020.06.19

(21)申请号 201880071520.8

(22)申请日 2018.10.30

(30)优先权数据
62/579,225 2017.10.31 US

(85)PCT国际申请进入国家阶段日
2020.04.30

(86)PCT国际申请的申请数据
PCT/US2018/058249 2018.10.30

(87)PCT国际申请的公布数据
WO2019/089619 EN 2019.05.09

(71)申请人 起元技术有限责任公司
地址 美国马萨诸塞州

(72)发明人 克雷格·W·斯坦菲尔
约瑟夫·S·沃利三世

(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 9/50(2006.01)
G06F 9/54(2006.01)

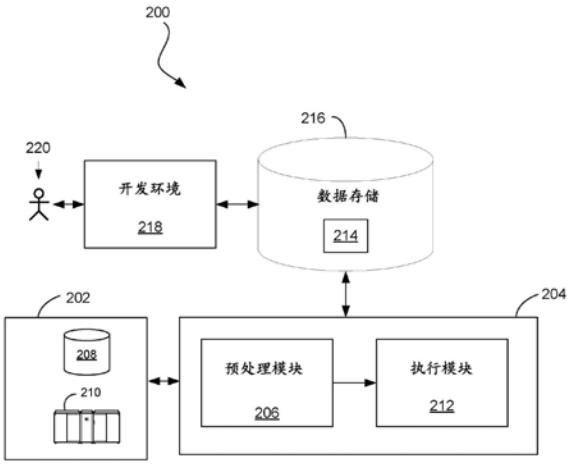
权利要求书8页 说明书27页 附图37页

(54)发明名称

使用复制的任务结果管理计算集群

(57)摘要

一种用于在分布式数据处理系统中处理任务的方法包括处理任务集。该方法包括在第一处理节点处维护多个计数器,该多个计数器包括:工作计数器,该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔;以及复制计数器,该复制计数器指示该多个时间间隔中的某个时间间隔,在该时间间隔内,在该多个处理节点中的数个处理节点处复制以下中的至少一个:(1)与该时间间隔相关联的所有任务以及(2)与该时间间隔相关联的所有相应结果。该方法包括从该第一处理节点向该多个处理节点中的其他处理节点提供消息,该消息包括该工作计数器和该复制计数器。



CN 111316241 A

1.一种用于在包括多个处理节点的分布式数据处理系统中处理任务的方法,该方法包括:

使用该多个处理节点中的两个或更多个处理节点来处理多个任务集,每个任务集被配置用于生成相应的结果集,并与多个时间间隔中的相应时间间隔相关联,

在该多个处理节点中的第一处理节点处维护多个计数器,该多个计数器包括:

工作计数器,该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔,以及

复制计数器,该复制计数器指示该多个时间间隔中的某个时间间隔,在该时间间隔内,在该多个处理节点中的数个处理节点处复制以下中的至少一个:(1)与该时间间隔相关联的所有任务以及(2)与该时间间隔相关联的所有相应结果,以及

从该第一处理节点向该多个处理节点中的其他处理节点提供消息,该消息包括该工作计数器和该复制计数器。

2.如权利要求1所述的方法,其中,该处理包括:在该多个处理节点中的主要处理节点处执行与至少一些任务相关联的计算,以及在该多个处理节点中的一个或多个备用处理节点处执行与该至少一些任务的复制品相关联的计算。

3.如权利要求1或2所述的方法,其中,在该主要处理节点处执行与第一任务相关联的计算包括在该主要处理节点处生成第一结果,并且在该备用处理节点处执行与该第一任务的复制品相关联的计算包括在该备用处理节点处生成该第一结果。

4.如权利要求1至3中任一项所述的方法,其中,在该主要处理节点处执行的与该第一任务相关联的计算和在该备用处理节点处执行的与该第一任务的复制品相关联的计算这两者都在提交操作之后开始,该提交操作指示该第一任务和该第一任务的复制品已被持久地存储。

5.如权利要求1至3中任一项所述的方法,其中,该第一结果包括已在该主要处理节点和该备用处理节点处复制的原始数据的修改后的版本。

6.如权利要求1至3中任一项所述的方法,其中,在该主要处理节点处执行的与该第一任务相关联的计算同在该备用处理节点处执行的与该第一任务的复制品相关联的计算相同。

7.如权利要求6所述的方法,其中,在该主要处理节点处执行的与该第一任务相关联的计算和在该备用处理节点处执行的与该第一任务的复制品相关联的计算是确定性的,并且不依赖于该数个处理节点中的哪一个来执行这些计算。

8.如权利要求2所述的方法,其中,该复制计数器指示该多个时间间隔中的某个时间间隔,在该时间间隔内,在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有任务以及与该时间间隔相关联的所有相应结果。

9.如权利要求1至8中任一项所述的方法,其中,该处理包括:在该多个处理节点中的主要处理节点处执行与至少一些任务相关联的计算,而该任务的复制品在备用处理节点处保持休眠;以及从该主要处理节点向该备用处理节点发送与执行了计算的任务相对应的结果。

10.如权利要求9所述的方法,其中,在该主要处理节点处执行的与该第一任务相关联的计算包括生成第一结果,并且在该第一任务已在该主要处理节点处完成之后从该主要处

理节点向该备用处理节点发送该第一结果。

11.如权利要求9或10所述的方法，其中，该第一结果包括已在该主要处理节点和该备用处理节点处复制的原始数据的修改后的版本。

12.如权利要求9所述的方法，其中，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有任务以及与该时间间隔相关联的所有相应结果。

13.一种以非暂态形式存储在计算机可读介质上的软件，该软件用于在包括多个节点的分布式数据处理系统中处理任务，该软件包括用于使计算系统执行权利要求1至12中任一项所述的所有步骤的指令。

14.一种用于处理数据的装置，该装置包括：

包含多个处理节点的分布式数据处理系统，每个处理节点包括至少一个处理器；以及

通信介质，该通信介质连接该多个处理节点以用于在该多个处理节点中的处理节点之间发送和接收信息；

其中，该分布式数据处理系统被配置用于：

使用该多个处理节点中的两个或更多个处理节点来处理多个任务集，每个任务集被配置用于生成相应的结果集，并与多个时间间隔中的相应时间间隔相关联，

在该多个处理节点中的第一处理节点处维护多个计数器，该多个计数器包括：

工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔，以及

复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制以下中的至少一个：(1)与该时间间隔相关联的所有任务以及(2)与该时间间隔相关联的所有相应结果，以及

从该第一处理节点向该多个处理节点中的其他处理节点提供消息，该消息包括该工作计数器和该复制计数器。

15.一种用于在包括多个处理节点的分布式数据处理系统中处理任务的计算系统，该计算系统包括：

用于使用该多个处理节点中的两个或更多个处理节点来处理多个任务集的装置，每个任务集被配置用于生成相应的结果集，并与多个时间间隔中的相应时间间隔相关联，

用于在该多个处理节点中的第一处理节点处维护多个计数器的装置，该多个计数器包括：

工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔，以及

复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制以下中的至少一个：(1)与该时间间隔相关联的所有任务以及(2)与该时间间隔相关联的所有相应结果，以及

用于从该第一处理节点向该多个处理节点中的其他处理节点提供消息的装置，该消息包括该工作计数器和该复制计数器。

16.一种用于管理包括多个处理节点的分布式数据处理系统的方法，该方法包括：

维护该系统中的多个数据存储装置，该多个数据存储装置中的每个数据存储装置与该

多个处理节点中的相应处理节点相关联，并与多个耐久性等级中的某个耐久性等级相关联，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度；

使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第一时间间隔相关联的第一数据单元集，该处理包括针对每个特定的耐久性等级更新相关联的指示符，以指示与该第一时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储；

使用该多个处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联，该多个请求集包括与该多个时间间隔中的第二时间间隔相关联的第一请求集；

在该多个处理节点中的第一处理节点处维护多个计数器，该多个计数器包括：

工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔，以及

复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求；以及

在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该工作计数器的值和该复制计数器的值。

17.如权利要求16所述的方法，其中，该多个计数器进一步包括：持久性计数器，该持久性计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，与该时间间隔相关联的所有请求均被存储在与该多个处理节点中的至少一个处理节点相关联的持久性存储装置中。

18.如权利要求16所述的方法，其中，针对该第一数据单元集中的每个数据单元，将该数据单元存储在该多个数据存储装置中的与该多个处理节点中的相应处理节点相关联的数据存储装置中，该存储包括将该数据单元存储在该多个数据存储装置中的与该第一耐久性等级相关联的数据存储装置中以及将该数据单元存储在该多个数据存储装置中的与该第二耐久性等级相关联的一个或多个数据存储装置中。

19.一种以非暂态形式存储在计算机可读介质上的软件，该软件用于管理包括多个处理节点的分布式数据处理系统，该软件包括用于使计算系统进行以下操作的指令：

维护该系统中的多个数据存储装置，该多个数据存储装置中的每个数据存储装置与该多个处理节点中的相应处理节点相关联，并与多个耐久性等级中的某个耐久性等级相关联，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度；

使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第一时间间隔相关联的第一数据单元集，该处理包括针对每个特定的耐久性等级更新相关联的指示符，以指示与该第一时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储；

使用该多个处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联，该多个请求集包括与该多个时间间隔中的第二时间间隔相关联的第一请求集；

在该多个处理节点中的第一处理节点处维护多个计数器，该多个计数器包括：

工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔，以及

复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求；以及

在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该工作计数器的值和该复制计数器的值。

20.一种装置，包括：

包含多个处理节点的分布式数据处理系统，每个处理节点包括至少一个处理器；以及

通信介质，该通信介质连接该多个处理节点以用于在该多个处理节点中的处理节点之间发送和接收信息；

其中，该分布式数据处理系统被配置用于：

维护该系统中的多个数据存储装置，该多个数据存储装置中的每个数据存储装置与该多个处理节点中的相应处理节点相关联，并与多个耐久性等级中的某个耐久性等级相关联，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度；

使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第一时间间隔相关联的第一数据单元集，该处理包括针对每个特定的耐久性等级更新相关联的指示符，以指示与该第一时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储；

使用该多个处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联，该多个请求集包括与该多个时间间隔中的第二时间间隔相关联的第一请求集；

在该多个处理节点中的第一处理节点处维护多个计数器，该多个计数器包括：

工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔，以及

复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求；以及

在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该工作计数器的值和该复制计数器的值。

21.一种用于管理包括多个处理节点的分布式数据处理系统的方法，该方法包括：

在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；

将接收到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将

与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；

在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；

在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二时间间隔相关联的指示符；

在该分布式数据处理系统接口组件处，将与该第二时间间隔相关联的指示符与包括在该结果数据中的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；

维护该系统中的多个数据存储装置，该多个数据存储装置中的每个数据存储装置与该多个处理节点中的相应处理节点相关联，并与多个耐久性等级中的某个耐久性等级相关联，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度；以及

使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与该多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第三时间间隔相关联的第一数据单元集，该处理包括针对每个特定的耐久性等级更新相关联的指示符，以指示与该第三时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储。

22.如权利要求21所述的方法，其中，针对该第一数据单元集中的每个数据单元，将该数据单元存储在该多个数据存储装置中的与该多个处理节点中的相应处理节点相关联的数据存储装置中，该存储包括将该数据单元存储在多个数据存储装置中的与该第一耐久性等级相关联的数据存储装置中以及将该数据单元存储在多个数据存储装置中的与该第二耐久性等级相关联的一个或多个数据存储装置中。

23.如权利要求21所述的方法，其中，与该第二时间间隔相关联的指示符被提供给该分布式数据处理系统接口组件。

24.一种以非暂态形式存储在计算机可读介质上的软件，该软件用于管理包括多个处理节点的分布式数据处理系统，该软件包括用于使计算系统进行以下操作的指令：

在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；

将接收到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；

在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；

在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符；

在该分布式数据处理系统接口组件处，将与该第二时间间隔相关联的指示符与包括在该结果数据中的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；

维护该系统中的多个数据存储装置，该多个数据存储装置中的每个数据存储装置与该

多个处理节点中的相应处理节点相关联，并与多个耐久性等级中的某个耐久性等级相关联，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度；以及

使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与该多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第三时间间隔相关联的第一数据单元集，该处理包括针对每个特定的耐久性等级更新相关联的指示符，以指示与该第三时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储。

25.一种装置，包括：

包含多个处理节点的分布式数据处理系统，每个处理节点包括至少一个处理器；以及

通信介质，该通信介质连接该多个处理节点以用于在该多个处理节点中的处理节点之间发送和接收信息；

其中，该分布式数据处理系统被配置用于：

在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；

将接收到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；

在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；

在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符；

在该分布式数据处理系统接口组件处，将与该第二时间间隔相关联的指示符与包括在该结果数据中的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；

维护该系统中的多个数据存储装置，该多个数据存储装置中的每个数据存储装置与该多个处理节点中的相应处理节点相关联，并与多个耐久性等级中的某个耐久性等级相关联，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度；以及

使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与该多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第三时间间隔相关联的第一数据单元集，该处理包括针对每个特定的耐久性等级更新相关联的指示符，以指示与该第三时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储。

26.一种用于管理包括多个处理节点的分布式数据处理系统的方法，该方法包括：

在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；

将接收到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；

在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；

在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符；

在该分布式数据处理系统接口组件处，将与该第二时间间隔相关联的指示符与包括在该结果数据中的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；

使用该多个处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联，该多个请求集包括与该多个时间间隔中的第三时间间隔相关联的第一请求集；

在该多个处理节点中的第一处理节点处维护多个计数器，该多个计数器包括：

工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔，以及

复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求，以及

在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该工作计数器的值和该复制计数器的值。

27.如权利要求26所述的方法，其中，与该第二时间间隔相关联的指示符被提供给该分布式数据处理系统接口组件。

28.如权利要求26所述的方法，其中，该多个计数器进一步包括：持久性计数器，该持久性计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，与该时间间隔相关联的所有请求均被存储在与该多个处理节点中的至少一个处理节点相关联的持久性存储装置中。

29.一种以非暂态形式存储在计算机可读介质上的软件，该软件用于管理包括多个处理节点的分布式数据处理系统，该软件包括用于使计算系统进行以下操作的指令：

在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；

将接收到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；

在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；

在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符；

在该分布式数据处理系统接口组件处，将与该第二时间间隔相关联的指示符与包括在该结果数据中的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；

使用该多个处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联，该多个请求集包括与该多个时间间隔中的第三时间间隔

相关联的第一请求集；

在该多个处理节点中的第一处理节点处维护多个计数器，该多个计数器包括：

工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔，以及

复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求，以及

在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该工作计数器的值和该复制计数器的值。

30.一种装置，包括：

包含多个处理节点的分布式数据处理系统，每个处理节点包括至少一个处理器；以及

通信介质，该通信介质连接该多个处理节点以用于在该多个处理节点中的处理节点之间发送和接收信息；

其中，该分布式数据处理系统被配置用于：

在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；

将接收到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；

在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；

在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符；

在该分布式数据处理系统接口组件处，将与该第二时间间隔相关联的指示符与包括在该结果数据中的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；

使用该多个处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联，该多个请求集包括与该多个时间间隔中的第三时间间隔相关联的第一请求集；

在该多个处理节点中的第一处理节点处维护多个计数器，该多个计数器包括：

工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔，以及

复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求，以及

在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该工作计数器的值和该复制计数器的值。

使用复制的任务结果管理计算集群

[0001] 相关申请的交叉引用

[0002] 本申请要求于2017年10月31日提交的美国申请序列号62/579,225的优先权,该美国申请通过引用并入本文。

背景技术

[0003] 本说明书涉及管理计算集群。

[0004] 数据流计算的一种方法利用了基于图的表示,其中,与图的节点(顶点)相对应的计算组件通过与该图(称为“数据流图”)的链路(有向边)相对应的数据流耦接。通过数据流链路连接到上游组件的下游组件接收有序的输入数据元素流,并按接收到的顺序处理输入数据元素,以可选地生成一个或多个相应的输出数据元素流。在名称为“EXECUTING COMPUTATIONS EXPRESSED AS GRAPHS(执行表示为图的计算)”的在先美国专利5,966,072中描述了一种用于执行此类基于图的计算的系统,该专利通过引用并入本文。在与该在先专利中描述的方法有关的实施方式中,每个组件被实施为托管在多个典型计算机服务器之一上的进程。每个计算机服务器可以在任何一个时间激活多个此类组件进程,并且操作系统(例如,Unix)调度程序在该服务器上托管的组件之间共享资源(例如,处理器时间和/或处理器核)。在这种实施方式中,可以使用操作系统和连接服务器的数据网络的数据通信服务(例如,命名管道、TCP/IP会话等)来实施组件之间的数据流。组件的子集通常用作整个计算(例如,去往和/或来自数据文件、数据库表和外部数据流)的数据源和/或数据接收器。在例如通过协作进程建立了组件进程和数据流之后,数据便流经整个计算系统,该计算系统实施表示为图的计算,该计算通常受每个组件处输入数据的可用性和每个组件的计算资源的调度的支配。因此,至少可以通过使不同的组件能够由不同的进程(托管在相同或不同的服务器计算机或处理器核上)并行执行来实现并行性,其中,在本文中通过数据流图在不同路径上并行执行不同组件称为组件并行性,并且在本文中通过数据流图在同一路径的不同部分上并行执行不同组件称为流水线并行性。

[0005] 这种方法还支持其他形式的并行性。例如,可以例如根据数据集的记录中字段的值的分区来划分输入数据集,其中,每个部分被发送到组件中处理该数据集的记录的单独副本。组件的此类单独副本(或“实例”)可以在单独服务器计算机或服务器计算机的单独处理器核上执行,从而实现本文所称的数据并行性。单独组件的结果可以合并以再次形成单个数据流或数据集。用于执行组件的实例的计算机或处理器核的数量将在开发数据流图时由开发者指定。

[0006] 可以使用各种方法来提高这种方法的效率。例如,组件的每个实例不一定必须托管在其自己的操作系统进程中,例如,使用一个操作系统进程来实施多个组件(例如,形成较大图的连接子图的组件)。

[0007] 上述方法的至少一些实施方式受到与在基础计算机服务器上执行所产生进程的效率有关的限制。例如,这些限制可能与重新配置图的运行实例以更改数据并行性程度、更改托管各种组件的服务器和/或平衡不同计算资源上的负荷的困难有关。现有的基于图的

计算系统还具有启动时间慢的问题，这通常是因为不必要地启动了太多的进程而浪费了大量的内存。一般而言，进程从图执行的启动开始，并且到图执行完成时结束。

[0008] 已使用了用于分布计算的其他系统，其中将整个计算分为较小的部分，并且将这些部分从一个主计算机服务器分布到各个其他(例如，“从”)计算机服务器，这些计算机服务器各自独立地执行计算并且将其结果返回到主服务器。此类方法中的一些称为“网格计算”。然而，此类方法通常依赖于每个计算的独立性，除了经由主计算机服务器调用计算部分之外，不提供用于在那些部分之间传递数据、或者对这些部分的执行进行调度和/或排序的机制。因此，此类方法不能为托管涉及多个组件之间的交互的计算提供直接且高效的解决方案。

[0009] 对大型数据集进行分布式计算的另一种方法利用MapReduce框架(例如，如Apache Hadoop®系统中所实施的)。一般而言，Hadoop具有分布式文件系统，每个命名文件的部分都分布于其中。用户根据两个函数指定计算：在命名输入的所有部分上以分布式方式执行的映射函数和在映射函数执行的输出部分上执行的归约函数。映射函数执行的输出被划分并再次存储在分布式文件系统中的中间部分中。然后以分布式方式执行归约函数以处理中间部分，从而得出整个计算的结果。尽管可以高效地执行可以在MapReduce框架中表示的、并且其输入和输出可修改以存储在MapReduce框架的文件系统内的计算，但是许多计算与此框架不匹配和/或不容易适配用于使其所有输入和输出都包含在分布式文件系统内。

[0010] 在总体方面，一种用于在包括多个处理节点的分布式数据处理系统中处理任务的方法包括使用该多个处理节点中的两个或更多个处理节点来处理多个任务集。每个任务集被配置用于生成相应的结果集，并与多个时间间隔中的相应时间间隔相关联。该方法包括在该多个处理节点中的第一处理节点处维护多个计数器。该多个计数器包括：工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔；以及复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制以下中的至少一个：(1)与该时间间隔相关联的所有任务以及(2)与该时间间隔相关联的所有相应结果。该方法包括从该第一处理节点向该多个处理节点中的其他处理节点提供消息，该消息包括该工作计数器和该复制计数器。

[0011] 各方面可以包括以下特征中的一项或多项。

[0012] 该处理可以包括：在该多个处理节点中的主要处理节点处执行与至少一些任务相关联的计算，以及在该多个处理节点中的一个或多个备用处理节点处执行与该至少一些任务的复制品相关联的计算。在该主要处理节点处执行与第一任务相关联的计算可以包括在该主要处理节点处生成第一结果，并且在该备用处理节点处执行与该第一任务的复制品相关联的计算包括在该备用处理节点处生成该第一结果。

[0013] 在该主要处理节点处执行的与该第一任务相关联的计算和在该备用处理节点处执行的与该第一任务的复制品相关联的计算这两者都可以在提交操作之后开始，该提交操作指示该第一任务和该第一任务的复制品已被持久地存储。该第一结果可以包括已在该主要处理节点和该备用处理节点处复制的原始数据的修改后的版本。在该主要处理节点处执行的与该第一任务相关联的计算可以同在该备用处理节点处执行的与该第一任务的复制品相关联的计算相同。

[0014] 在该主要处理节点处执行的与该第一任务相关联的计算和在该备用处理节点处

执行的与该第一任务的复制品相关联的计算可以是确定性的，并且可以不依赖于该数个处理节点中的哪一个来执行这些计算。该复制计数器可以指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有任务以及与该时间间隔相关联的所有相应结果。

[0015] 该处理可以包括：在该多个处理节点中的主要处理节点处执行与至少一些任务相关联的计算，而该任务的复制品在备用处理节点处保持休眠；以及从该主要处理节点向该备用处理节点发送与执行了计算的任务相对应的结果。在该主要处理节点处执行的与该第一任务相关联的计算可以包括生成第一结果，并且可以在该第一任务已在该主要处理节点处完成之后从该主要处理节点向该备用处理节点发送该第一结果。

[0016] 该第一结果可以包括已在该主要处理节点和该备用处理节点处复制的原始数据的修改后的版本。该复制计数器可以指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有任务以及与该时间间隔相关联的所有相应结果。

[0017] 在另一总体方面，一种用于在包括多个节点的分布式数据处理系统中处理任务的软件以非暂态形式存储在计算机可读介质上。该软件包括用于使计算系统使用该多个处理节点中的两个或更多个处理节点来处理多个任务集的指令。每个任务集被配置用于生成相应的结果集，并与多个时间间隔中的相应时间间隔相关联。

[0018] 这些指令还使该计算系统在该多个处理节点中的第一处理节点处维护多个计数器。该多个计数器包括：工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔；以及复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制以下中的至少一个：(1)与该时间间隔相关联的所有任务以及(2)与该时间间隔相关联的所有相应结果。这些指令还使该计算系统从该第一处理节点向该多个处理节点中的其他处理节点提供消息，该消息包括该工作计数器和该复制计数器。

[0019] 在另一总体方面，一种用于处理数据的装置包括：包含多个处理节点的分布式数据处理系统，每个处理节点包括至少一个处理器；以及通信介质，该通信介质连接该多个处理节点以用于在该多个处理节点中的处理节点之间发送和接收信息。该分布式数据处理系统被配置用于：使用该多个处理节点中的两个或更多个处理节点来处理多个任务集，每个任务集被配置用于生成相应的结果集，并与多个时间间隔中的相应时间间隔相关联；在该多个处理节点中的第一处理节点处维护多个计数器。该多个计数器包括：工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔；以及复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制以下中的至少一个：(1)与该时间间隔相关联的所有任务以及(2)与该时间间隔相关联的所有相应结果。该装置进一步被配置用于从该第一处理节点向该多个处理节点中的其他处理节点提供消息，该消息包括该工作计数器和该复制计数器。

[0020] 在另一总体方面，一种用于在包括多个处理节点的分布式数据处理系统中处理任务的计算系统包括：用于使用该多个处理节点中的两个或更多个处理节点来处理多个任务集的装置，每个任务集被配置用于生成相应的结果集，并与多个时间间隔中的相应时间间

隔相关联;用于在该多个处理节点中的第一处理节点处维护多个计数器的装置。该多个计数器包括:工作计数器,该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔;以及复制计数器,该复制计数器指示该多个时间间隔中的某个时间间隔,在该时间间隔内,在该多个处理节点中的数个处理节点处复制以下中的至少一个:(1)与该时间间隔相关联的所有任务以及(2)与该时间间隔相关联的所有相应结果。该计算系统还包括用于从该第一处理节点向该多个处理节点中的其他处理节点提供消息的装置,该消息包括该工作计数器和该复制计数器。

[0021] 在总体方面,一种用于管理包括多个处理节点的分布式数据处理系统的方法包括:维护该系统中的多个数据存储装置,该多个数据存储装置中的每个数据存储装置与该多个处理节点中的相应处理节点相关联,并与多个耐久性等级中的某个耐久性等级相关联,该多个耐久性等级包括第一耐久性等级和第二耐久性等级,该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度。该方法还包括使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集,每个数据单元集中的每个数据单元与多个时间间隔中的相应时间间隔相关联。该多个数据单元集包括与该多个时间间隔中的第一时间间隔相关联的第一数据单元集。

[0022] 该处理包括针对每个特定的耐久性等级更新相关联的指示符,以指示与该第一时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储。该处理还包括使用该多个处理节点中的两个或更多个处理节点来处理多个请求集,每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新,并与该多个时间间隔中的相应时间间隔相关联,该多个请求集包括与该多个时间间隔中的第二时间间隔相关联的第一请求集。该处理还包括在该多个处理节点中的第一处理节点处维护多个计数器。

[0023] 该多个计数器包括:工作计数器,该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔;以及复制计数器,该复制计数器指示该多个时间间隔中的某个时间间隔,在该时间间隔内,在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求。

[0024] 该方法还包括在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息,该第一消息包括该工作计数器的值和该复制计数器的值。

[0025] 各方面可以包括以下特征中的一项或多项。

[0026] 该多个计数器可以进一步包括:持久性计数器,该持久性计数器指示该多个时间间隔中的某个时间间隔,在该时间间隔内,与该时间间隔相关联的所有请求均被存储在与该多个处理节点中的至少一个处理节点相关联的持久性存储装置中。该方法可以包括:针对该第一数据单元集中的每个数据单元,将该数据单元存储在该多个数据存储装置中的与该多个处理节点中的相应处理节点相关联的数据存储装置中,该存储包括将该数据单元存储在该多个数据存储装置中的与该第一耐久性等级相关联的数据存储装置中以及将该数据单元存储在该多个数据存储装置中的与该第二耐久性等级相关联的一个或多个数据存储装置中。

[0027] 在另一总体方面,一种用于管理包括多个处理节点的分布式数据处理系统的软件以非暂态形式存储在计算机可读介质上。该软件包括用于使计算系统进行以下操作的指令:维护该系统中的多个数据存储装置,该多个数据存储装置中的每个数据存储装置与该

多个处理节点中的相应处理节点相关联，并与多个耐久性等级中的某个耐久性等级相关联，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度。这些指令还使该计算系统进行以下操作：使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第一时间间隔相关联的第一数据单元集，该处理包括针对每个特定的耐久性等级更新相关联的指示符，以指示与该第一时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储。

[0028] 这些指令还使该计算系统进行以下操作：该处理还包括使用该多个处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联，该多个请求集包括与该多个时间间隔中的第二时间间隔相关联的第一请求集。这些指令还使该计算系统在该多个处理节点中的第一处理节点处维护多个计数器。该多个计数器包括：工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔；以及复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求。

[0029] 这些指令还使该计算系统进行以下操作：在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该工作计数器的值和该复制计数器的值。

[0030] 在另一总体方面，一种装置包括：包含多个处理节点的分布式数据处理系统，每个处理节点包括至少一个处理器；以及通信介质，该通信介质连接该多个处理节点以用于在该多个处理节点中的处理节点之间发送和接收信息。该分布式数据处理系统被配置用于维护该系统中的多个数据存储装置，该多个数据存储装置中的每个数据存储装置与该多个处理节点中的相应处理节点相关联，并与多个耐久性等级中的某个耐久性等级相关联，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度。

[0031] 该装置还被配置用于使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第一时间间隔相关联的第一数据单元集，该处理包括针对每个特定的耐久性等级更新相关联的指示符，以指示与该第一时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储。

[0032] 该装置还被配置用于使用该多个处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联，该多个请求集包括与该多个时间间隔中的第二时间间隔相关联的第一请求集。该装置还被配置用于在该多个处理节点中的第一处理节点处维护多个计数器。

[0033] 该多个计数器包括：工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔；以及复制计数器，该复制计数器指示该多个时间间隔

中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求。该装置还被配置用于在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该工作计数器的值和该复制计数器的值。

[0034] 在另一总体方面，一种用于管理包括多个处理节点的分布式数据处理系统的方法包括：在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；将接收到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二时间间隔相关联的指示符；在该分布式数据处理系统接口组件处，将与该第二时间间隔相关联的指示符与该结果数据中包括的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；维护该系统中的多个数据存储装置，该多个数据存储装置中的每个数据存储装置与该多个处理节点中的相应处理节点相关联，并与多个耐久性等级中的某个耐久性等级相关联，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度；以及使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第三时间间隔相关联的第一数据单元集，该处理包括针对每个特定的耐久性等级更新相关联的指示符，以指示与该第三时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储。

[0035] 各方面可以包括以下特征中的一项或多项。

[0036] 针对该第一数据单元集中的每个数据单元，可以将该数据单元存储在该多个数据存储装置中的与该多个处理节点中的相应处理节点相关联的数据存储装置中，该存储包括将该数据单元存储在该多个数据存储装置中的与该第一耐久性等级相关联的数据存储装置中以及将该数据单元存储在该多个数据存储装置中的与该第二耐久性等级相关联的一个或多个数据存储装置中。与该第二时间间隔相关联的指示符可以被提供给该分布式数据处理系统接口组件。

[0037] 在另一总体方面，一种用于管理包括多个处理节点的分布式数据处理系统的软件以非暂态形式存储在计算机可读介质上。该软件包括用于使计算系统进行以下操作的指令：在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；将接收到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符；在该分布式数据处理系统接口组件处，将包括在与该第二时间间隔相关联的指示符与该结果数据中的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第

一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；维护该系统中的多个数据存储装置，该多个数据存储装置中的每个数据存储装置与该多个处理节点中的相应处理节点相关联，并与多个耐久性等级中的某个耐久性等级相关联，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度；以及使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第三时间间隔相关联的第一数据单元集，该处理包括针对每个特定的耐久性等级更新相关联的指示符，以指示与该第三时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储。

[0038] 在另一总体方面，一种装置包括：包含多个处理节点的分布式数据处理系统，每个处理节点包括至少一个处理器；以及通信介质，该通信介质连接该多个处理节点以用于在该多个处理节点中的处理节点之间发送和接收信息。该分布式数据处理系统被配置用于：在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；将接收到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符；在该分布式数据处理系统接口组件处，将与包括在该第二时间间隔相关联的指示符与该结果数据中的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；维护该系统中的多个数据存储装置，该多个数据存储装置中的每个数据存储装置与该多个处理节点中的相应处理节点相关联，并与多个耐久性等级中的某个耐久性等级相关联，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度；以及使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第三时间间隔相关联的第一数据单元集，该处理包括针对每个特定的耐久性等级更新相关联的指示符，以指示与该第三时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储。

[0039] 在另一总体方面，一种用于管理包括多个处理节点的分布式数据处理系统的方法包括：在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；将接收到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符；在该分布式数据处理系统接口组件处，将与该第二时间间隔相关联的指示符与该结果数据中包括的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；使用该多个

处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联，该多个请求集包括与该多个时间间隔中的第三时间间隔相关联的第一请求集；在该多个处理节点中的第一处理节点处维护多个计数器。该多个计数器包括：工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔；以及复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求。该方法还包括在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该工作计数器的值和该复制计数器的值。

[0040] 各方面可以包括以下特征中的一项或多项。

[0041] 与该第二时间间隔相关联的指示符可以被提供给该分布式数据处理系统接口组件。该多个计数器可以包括：持久性计数器，该持久性计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，与该时间间隔相关联的所有请求均被存储在与该多个处理节点中的至少一个处理节点相关联的持久性存储装置中。

[0042] 在另一总体方面，一种用于管理包括多个处理节点的分布式数据处理系统的软件以非暂态形式存储在计算机可读介质上。该软件包括用于使计算系统进行以下操作的指令：在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；将接收到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符；在该分布式数据处理系统接口组件处，将与包括在该第二时间间隔相关联的指示符与该结果数据中的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；使用该多个处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联，该多个请求集包括与该多个时间间隔中的第三时间间隔相关联的第一请求集；在该多个处理节点中的第一处理节点处维护多个计数器。该多个计数器包括：工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔；以及复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求。该软件还包括用于使该计算系统进行以下操作的指令：在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该工作计数器的值和该复制计数器的值。

[0043] 在另一总体方面，一种装置包括：包含多个处理节点的分布式数据处理系统，每个处理节点包括至少一个处理器；以及通信介质，该通信介质连接该多个处理节点以用于在该多个处理节点中的处理节点之间发送和接收信息。该分布式数据处理系统被配置用于：在与该分布式数据处理系统通信的分布式数据处理系统接口组件处接收输入数据；将接收

到的输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；在该分布式数据处理系统接口组件处从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符；在该分布式数据处理系统接口组件处，将与包括在该第二时间间隔相关联的指示符与该结果数据中的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则从该分布式数据处理系统接口组件释放该结果数据；使用该多个处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联，该多个请求集包括与该多个时间间隔中的第三时间间隔相关联的第一请求集；在该多个处理节点中的第一处理节点处维护多个计数器。该多个计数器包括：工作计数器，该工作计数器在该分布式数据处理系统中指示该多个时间间隔中的当前时间间隔；以及复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求。该装置还被配置用于在第一时间从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该工作计数器的值和该复制计数器的值。

[0044] 在另一方面，总体上，管理包括多个处理节点的分布式数据处理系统包括：将输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的第一时间间隔相关联的指示符分派给该输入数据；从该分布式数据处理系统中接收与该输入数据相关联的结果数据，其中，该结果数据包括与该第一时间间隔相关联的指示符；在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符；将与该第二时间间隔相关联的指示符与该结果数据中包括的与该第一时间间隔相关联的指示符进行比较，并且如果与该第二时间间隔相关联的指示符对应于等于或晚于该第一时间间隔的时间间隔，则释放该结果数据；使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集，每个数据单元集中的每个数据单元与该多个时间间隔中的相应时间间隔相关联，该多个数据单元集包括与该多个时间间隔中的第三时间间隔相关联的第一数据单元集，该处理包括针对多个耐久性等级中的每个特定的耐久性等级更新相关联的指示符，以指示与该第三时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储，该多个耐久性等级包括第一耐久性等级和第二耐久性等级，该第二耐久性等级具有比该第一耐久性等级相对更大的耐久性程度；使用该多个处理节点中的两个或更多个处理节点来处理多个请求集，每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新，并与该多个时间间隔中的相应时间间隔相关联；在该多个处理节点中的第一处理节点处维护复制计数器，该复制计数器指示该多个时间间隔中的某个时间间隔，在该时间间隔内，在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求；以及从该第一处理节点向该多个处理节点中的其他处理节点提供第一消息，该第一消息包括该复制计数器的值。

[0045] 在另一方面，总体上，管理包括多个处理节点的分布式数据处理系统包括：将输入数据提供给该分布式数据处理系统，其中，该分布式数据处理系统将与多个时间间隔中的

第一时间间隔相关联的指示符分派给该输入数据;从该分布式数据处理系统中接收与该输入数据相关联的结果数据,其中,该结果数据包括与该第一时间间隔相关联的指示符;在该多个处理节点中的第一处理节点处确定与该多个时间间隔中的第二间隔相关联的指示符;基于将与该第二时间间隔相关联的指示符与包括在该结果数据中的与该第一时间间隔相关联的指示符进行比较来释放该结果数据;使用该多个处理节点中的两个或更多个处理节点来处理多个数据单元集,每个数据单元集中的每个数据单元与该多个时间间隔中的相应时间间隔相关联,该多个数据单元集包括与该多个时间间隔中的第三时间间隔相关联的第一数据单元集,该处理包括针对多个耐久性等级中的每个特定的耐久性等级更新相关联的指示符,以指示与该第三时间间隔相关联的所有数据单元集均以该特定的耐久性等级存储。使用该多个处理节点中的两个或更多个处理节点来处理多个请求集,每个请求集中的每个请求被配置用于在该多个处理节点中的某个处理节点处引起状态更新,并与该多个时间间隔中的相应时间间隔相关联,其中,在该多个处理节点中的一个或多个处理节点处的状态更新包括更新存储在存储器中的状态,该存储器使用避免自动垃圾回收的存储器管理(例如,显式存储器分配和释放,或自动引用计数)来进行管理;在该多个处理节点中的第一处理节点处维护复制计数器,该复制计数器指示该多个时间间隔中的某个时间间隔,在该时间间隔内,在该多个处理节点中的数个处理节点处复制与该时间间隔相关联的所有请求;以及从该第一处理节点向该多个处理节点中的至少一个其他处理节点提供第一消息,该第一消息包括该复制计数器的值。

[0046] 各方面可以具有以下优点中的一个或多个优点。

[0047] 总体上,与其中组件(或组件的并行执行副本)被托管在不同的服务器上的上述方法相比,本文中描述的一些特征使得能够提高计算(尤其是基础规范是基于图的程序规范的计算)的计算效率(例如,包括多个处理节点的分布式数据处理系统能够增加每单位给定计算资源所处理的记录数量)。例如,调用集群组件被布置在基于图的程序规范中,并用于将该基于图的程序规范与分布式数据处理系统接口连接,使得由该基于图的程序规范中的处理节点以分布式方式执行该基于图的程序规范所需的计算。此外,本文中描述的一些特征提供了适配用于变化的计算资源和计算要求的能力。本文提供了一种计算方法,该计算方法允许适配用于在执行一个或多个基于图的计算期间可用的计算资源的变化,和/或适配用于(例如,由于正在处理的数据的特性导致的)计算负荷的变化或此类计算的不同组件的负荷随时间的变化。例如,各方面能够适配用于处理节点从分布式数据处理系统中添加或移除(或发生故障以及重新在线)。分布式数据处理系统提供适配性的一种方式是管理系统中数据的复制和持久性,包括维护由处理节点发送和接收的消息的计数以及维护所有消息在系统中被复制和/或持久化的时间间隔的指示符。

[0048] 还提供了一种计算方法,该计算方法能够高效地利用具有不同特性的计算资源(例如,使用每个服务器具有不同处理器数量、每个处理器具有不同处理器核数等的服务器)并高效地支持同构以及异构环境两者。本文中描述的一些特征还能够使基于图的计算快速启动。如本文中描述的,提供这种效率和适应性的一方面是提供对处理节点集群的适当管理。

[0049] 各方面还有利地是容错的,因为分布式数据处理系统能够通过及时回滚处理而从发生的任何处理错误中恢复。该系统预见了多个可能的回滚场景,并实施用于在可能的回

滚场景中的每一个中执行回滚的算法。

附图说明

[0050] 图1是用于处理数据的系统的框图。

[0051] 图2是包括计算集群的计算系统的框图。

[0052] 图3是表示各种重复时间间隔的时间的时钟的示意图。

[0053] 图4是操作过程的状态转换图。

[0054] 图5至图12展示了计算系统的正常操作。

[0055] 图13至图15展示了第一回滚过程。

[0056] 图16至图18展示了第二回滚过程。

[0057] 图19至图21展示了第三回滚过程。

[0058] 图22至图25展示了第四回滚过程。

[0059] 图26至图29展示了第五回滚过程。

[0060] 图30至图32展示了第六回滚过程。

[0061] 图33至图35展示了第七回滚过程。

[0062] 图36至图37展示了第八回滚过程。

具体实施方式

[0063] 图1示出数据处理系统200的示例，在该数据处理系统中可以使用计算集群管理技术。系统200包括数据源202，该数据源可以包括一个或多个数据源，诸如存储设备或到在线数据流的连接，数据源中的每一个可以以各种格式(例如，数据库表、电子表格文件、平面文本文件或主机使用的本机格式)中的任一种存储或提供数据。执行环境204包括预处理模块206和执行模块212。例如，在合适的操作系统(诸如UNIX操作系统版本)的控制下，执行环境204可以被托管在一个或多个通用计算机上。例如，执行环境204可以包括多节点并行计算环境，该多节点并行计算环境包括使用多个处理单元(例如，中央处理单元CPU)或处理器核的计算机系统的配置，这些处理单元或处理器核或是本地的(例如，多处理器系统，诸如对称多处理(SMP)计算机)、或是本地分布式的(例如，作为集群或大规模并行处理(MPP)系统耦接的多个处理器)、或是远程的或远程分布的(例如，经由局域网(LAN)和/或广域网(WAN)耦接的多个处理器)、或其任何组合。

[0064] 预处理模块206能够执行在由执行模块212执行程序规范(例如，以下描述的基于图的程序规范)之前可能需要的任何配置。预处理模块206可以配置程序规范以从可以实施数据源202的各种类型的系统(包括不同形式的数据库系统)接收数据。数据可以被组织为具有相应字段(也称为“属性”、“行”或“列”)的值(包括可能的空值)的记录。当首先配置计算机程序(诸如数据处理应用程序)以从数据源读取数据时，预处理模块206通常以关于此数据源中的记录的一些初始格式信息开始。计算机程序可以以如本文中描述的数据流图的形式表示。在一些情况下，数据源的记录结构最初可能不是已知的，而是可在分析数据源或数据之后确定。关于记录的初始信息可包括例如表示不同值的位数、记录内的字段的顺序、以及由位表示的值的类型(例如，字符串、有符号/无符号整数)。

[0065] 提供数据源202的存储设备可以在执行环境204本地，例如，存储在连接到托管执

行环境204的计算机的存储介质(例如,硬盘驱动器208)上,或者可以远离执行环境204,例如,被托管在通过(例如,由云计算基础设施提供的)远程连接与托管执行环境204的计算机通信的远程系统(例如,主机210)上。

[0066] 执行模块212执行由预处理模块206配置和/或生成的程序规范以读取输入数据和/或生成输出数据。输出数据214可以存储回数据源202中或存储在执行环境204可访问的数据存储系统216中、或以其他方式使用。数据存储系统216也可由开发环境218访问,在该开发环境中开发者220能够开发用于使用执行模块212处理数据的应用程序。

[0067] 换言之,数据处理系统200可以包括:

[0068] 耦接到数据存储系统216的可选开发环境218,其中,开发环境218被配置用于构建与数据流图相关联的实施基于图的计算的数据处理应用程序,该基于图的计算是对从一个或多个输入数据集流过处理图组件的图到达一个或多个输出数据集的数据执行的,其中,该数据流图由数据存储系统216中的数据结构指定,该数据流图具有由数据结构指定并表示由一个或多个链路连接的图组件的多个节点,这些链路由数据结构指定并表示图组件之间的数据流;

[0069] 耦接到数据存储系统216并托管在一个或多个计算机上的执行环境212,执行环境212包括预处理模块206,该预处理模块被配置用于读取指定数据流图的已存储数据结构并用于分配和配置诸如进程等的计算资源,以用于执行对由预处理模块206分派给数据流图的图组件的计算;

[0070] 其中,执行环境204包括执行模块212,该执行模块用于调度和控制对所分派计算或进程的执行,以使得能够执行基于图的计算。即,执行模块被配置用于从数据源202读取数据并使用以数据流图的形式表示的可执行计算机程序来处理数据。

[0071] 1计算集群

[0072] 一般而言,用于使用执行模块212处理数据的一些计算机程序(在本文中也称为“应用程序”)包括应用程序用来访问计算集群的调用集群组件。例如,参考图2,在流水线式数据处理的方法中,调用集群组件110与计算机集群120的组件交互以处理在调用集群组件110处从该调用集群组件是其一部分的应用程序(例如,数据流图或其他形式的基于图的程序规范)中的组件接收的记录103,并且将相应的结果105传输到该调用集群组件是其一部分的应用程序中的一个或多个其他组件。对于每个输入记录103,调用集群组件110向集群120发送请求113(例如,用于执行数据处理任务的请求),并且一段时间后,该调用集群组件从集群120中接收对该请求113的响应115。在接收到响应115之后的一段时间,通常是在已知处理请求的结果在集群120中适当地持久化之后,调用集群组件110发送与响应115相对应的结果105。

[0073] 图2中未示出调用集群组件110是其一部分的基于图的程序规范。在图2中,仅示出了单个调用集群组件110,但是应当认识到,通常可以存在许多可以与同一集群120交互的调用集群组件,例如,每个调用集群组件参与相同或不同的应用程序(诸如数据流图)。基于图的程序规范可以被实施为例如如美国专利号5,966,072、美国专利号7,167,850或美国专利号7,716,630中描述的数据流图,或如美国公开号2016/0062776中描述的数据处理图。这种基于数据流图的程序规范通常包括与图的节点(顶点)相对应的计算组件,这些计算组件通过与该图(称为“数据流图”)的链路(有向边)相对应的数据流耦接。通过数据流链路连接

到上游组件的下游组件接收有序的输入数据元素流，并按接收到的顺序处理输入数据元素，以可选地生成一个或多个相应的输出数据元素流。在一些示例中，每个组件被实施为托管在多个典型计算机服务器之一上的进程。每个计算机服务器可以在任何一个时间激活多个此类组件进程，并且操作系统(例如，Unix)调度程序在该服务器上托管的组件之间共享资源(例如，处理器时间和/或处理器核)。在这种实施方式中，可以使用操作系统和连接服务器的数据网络的数据通信服务(例如，命名管道、TCP/IP会话等)来实施组件之间的数据流。组件的子集通常用作整个计算(例如，去往和/或来自数据文件、数据库表和外部数据流)的数据源和/或数据接收器。在例如通过协作进程建立了组件进程和数据流之后，数据便流经整个计算系统，该计算系统实施表示为图的计算，该计算通常受每个组件处输入数据的可用性和每个组件的计算资源的调度的支配。

[0074] 集群120包括由通信网络130(在图2中展示为“云”并且可以具有各种互连拓扑，诸如启动、共享介质、超立方体等)耦接的多个集群组件140、150a至150c。每个集群组件(或简称为“组件”)在集群中都有特定的角色。在一些实施方式中，每个组件都被托管在不同的计算资源(例如，单独计算机服务器、多核服务器的单独核等)上。应当理解，这些组件表示集群内的角色，并且在一些实施例中，多个角色可以被托管在一个计算资源上，并且单个角色可以分布在多个计算资源上。

[0075] 在图2中，根组件140(称为“根”)执行以下全面描述的某些同步功能，但是不直接参与要处理的数据的流或计算。多个工作器组件150a至150c(以下称为“工作器”)处理来自调用集群组件110的请求113。数据165以冗余方式存储在相应工作器150可访问的存储装置160中，并且每个请求113可能需要访问(用于读取和/或写入)由请求113中的密钥标识的存储在存储装置160中的数据的特定部分，该数据的特定部分分布在由密钥确定的工作器的特定子集之中。在持有特定请求所需的密钥数据的那些工作器中，一个工作器被指定为执行请求113的主要工作器(例如，工作器150a)，而其他工作器则被指定为备用工作器，因为这些工作器通常不或者不必执行该请求，但是其数据版本根据主要工作器或以与主要工作器相同的方式进行更新。

[0076] 在图2中，特定输入记录103(其可以被认为是或包括要处理的数据单元)的路径被展示为进入调用集群组件110，然后由组件110将相应的请求113(与数据单元一起)发送到该请求的主要工作器150a(工作器A)，并将来自主要工作器150a的响应115发送回调用集群组件110并发送到该请求的备用工作器150b(工作器B)，并且最后从调用集群组件110输出或发送相应的结果105。通常，每个请求可能有多个备用组件；然而，为了便于说明，在以下许多示例中仅展示了单个备用组件。

[0077] 如下面进一步讨论的，调用集群组件110将请求113缓存在重播缓冲器112中，并且如果必要，可以将请求重新发送到集群120以确保这些请求已被集群120正确地接收和/或处理。组件110还将响应115缓存在代管缓冲器114中，并且在检测到错误状态的情况下可以接收某些响应的冗余副本。通常，组件110“以代管的方式(inescrow)”保存响应，直到集群120通知组件110响应115在集群中适当地持久化(即，存储在具有合适的耐久性等级的数据存储装置中)。

[0078] 根140通过维护时间(间隔)值并将其分配给其他组件以及将某些时间值分配给调用集群组件110来执行同步功能。参考图3，根140的时钟142维护三个时间。时间T1是当前工

作时间或时间间隔，例如表示为整数值，并且被重复更新，例如每秒增加一次。

[0079] 当集群120从调用集群组件110接收到请求113，并且集群生成(或传输)响应115时，这些请求和响应各自与这些请求和响应分别被接收并生成(或传输)的工作时间(T1)相关联(或等效地与时间间隔相关联，在这些时间间隔期间，时间T1具有相同的值，即在T1的增量之间)。根维护并分配第二时间T2，该时间滞后于时间T1。时间T2表示这样的时间(间隔)，其使得在该时间或更早的时间创建的在集群120的组件150a至150c之间发送的所有请求和/或响应已在组件150a至150c中的多处被复制(例如，在易失性存储器中)，从而使得在用于处理错误的操作回滚的情况下不必重新发送这些请求和响应，如下面进行了更详细的描述。在一些示例中，复制(例如，在易失性存储器中)被称为以第一耐久性等级存储在数据存储装置中。根维护并分配第三时间(间隔)T3，该时间滞后于时间T1和T2，该时间表示这样的时间，其使得在该时间或更早的时间创建的所有请求和/或响应在存储数据165的工作器150a至150c中的至少一个或甚至所有处已被存储在持久性存储器中并成为永久的，使得在用于处理集群120中的组件故障的操作回滚的情况下不必重新发送或重新计算这些请求和响应。在一些示例中，被存储在持久性存储器(例如，存储到磁盘)中被称为以第二耐久性等级存储在数据存储装置中，其中该第二耐久性等级比第一耐久性等级相对更耐久。要注意的是，数据存储装置可以与多个不同耐久性等级相关联，这些不同耐久性等级比具有第一耐久性等级的数据存储装置和具有第二耐久性等级的数据存储装置相对更耐久或更不耐久。例如，集群外部的异地数据存储装置可以具有第三耐久性等级，其比第一耐久性等级和第二耐久性等级相对更耐久。在一些示例中，时间间隔T1、T2和T3可替代地称为“状态一致性指示符”。

[0080] 在本说明书中稍后描述根140确定何时增加复制时间(T2)或持久性时间(T3)的机制，以及将时间(T1-T3)的值分配给工作器150a至150c的机制。

[0081] 在正常操作中，由集群120接收的请求113在基于该请求的数据单元的密钥被标识为主要工作器的工作器150处被处理，并且通常在也基于所需数据的密钥标识的一个或多个备用工作器150处被处理。参考图4，该处理可以表示为在调用集群组件110以及主要工作器和备用工作器150处该请求的不同状态之间的转换。注意，不同的请求处于不同的状态并且通常根据引用的数据在不同的工作器处进行处理，并且因此，调用集群组件和任何特定工作器可能具有不同状态下的许多请求。

[0082] 通常，每个密钥与例如基于该密钥(例如，密钥的确定性函数，该函数为每个密钥值以不可预测的方式分配备用工作器)以伪随机方式选择的工作器150的相应子集相关联。更一般地并且优选地，这些子集与其他子集重叠，而不是根据密钥值形成完整的工作器集的分区。

[0083] 当在调用集群组件110处针对每个输入记录103形成具有(或由调用集群组件分派)唯一标识符rid的请求113时，该请求在调用集群组件中进入状态A。在下面的描述中，每个请求113在调用集群组件中处于标记为A-C的三个状态之一，并且在处理该请求的工作器150中的每一个处，处于标记为A-I的九个不同状态之一。在调用集群组件110记录请求113之后，其确定被分派为该请求的主要工作器的工作器150，并将请求113发送到该工作器150(图2中示出为工作器A)。注意，在替代性实施例中，调用集群组件110可能不知道哪个工作器是指定的主要工作器，并且请求113可能在集群120内被内部路由以到达指定的主要工作

器150a。请求113在调用集群组件110处保持状态A，直到从集群120接收回对该请求的响应115为止。

[0084] 当在主要工作器(图2中标记为工作器A)处接收到请求113时，该请求在主要工作器处进入状态A。主要工作器为请求分派请求时间，该请求时间表示为ta，等于(对其已知是)从根140开始分配的当前工作时间T1(认识到根增加T1与工作器知道该增量之间可能存在时间滞后)。在此状态下，请求113与请求id rid、请求时间(在此示例中被表示为ta)相关联地存储在易失性存储器155中，并且被指定为处于等待在主要工作器处执行的状态。在此状态A下，主要工作器将请求113发送到该请求的一个或多个备用工作器150(即，由密钥确定的)。在主要工作器处，例如，基于根据分派给请求的时间(ta)以及(可选地)请求在主要工作器处的到达顺序而对资源的有序分配，最终为请求分派资源以执行该请求。当请求113开始在主要工作器上执行时，该请求在主要工作器处进入状态B。当处理产生响应115时，在此示例中，假设T1工作时间这时为tb，则主要工作器处请求的状态变为状态C。在状态C下，响应115与时间tb相关联地存储在易失性存储器156中。如下面进一步讨论的，响应115和对工作器处的数据存储装置160的任何更新都与时间(此处为时间tb)相关联地进行存储，其方式允许例如使用版本化数据库或其他形式的版本化数据结构消除根据先前的回滚时间的影响。在此状态C下，响应115被传输到调用集群组件110以及(多个)备用组件150。

[0085] 在调用集群组件110处，当从主要工作器接收到响应115时，该请求进入状态B，在该状态下，该响应与主要工作器产生该响应的时间tb相关联地存储。响应115在调用集群组件处被保留在代管缓冲器114中，直到该调用集群组件从根140接收到等于或大于tb的代管时间为止。根据来自该调用集群组件的请求的持久性要求，根可以提供复制时间T2或持久性时间T3作为调用集群组件的代管时间。当调用集群组件110接收到等于或大于tb的代管时间时，该调用集群组件将结果105从调用集群组件中发送出去，并且相应的请求113进入空状态C，在该状态下，不再需要请求113或其响应115的进一步记录(例如，可以将其完全删除)。

[0086] 在(多个)备用工作器150处，当备用工作器接收到来自主要工作器的请求113时，备用工作器进入状态F，在该状态下，该请求与原始请求时间ta相关联(即使当前工作时间T1的增量超过此值也是如此)，并且该请求处于等待来自主要工作器的响应的状态。当备用工作器150b接收到来自主要工作器的响应115并因此将响应115复制在该备用工作器的易失性存储器156中时，该请求进入状态G。

[0087] 主要工作器或备用工作器一具有新生成的响应115，其就可以自由地开始将该响应保存到诸如基于磁盘或基于非易失性存储器的数据库或文件系统等的持久性存储装置160(参见状态D和H)的过程。可以使用基于日志的方法，其中首先在基于易失性存储器的日志中记录对持久性存储器的更新，并且将该日志的部分不时地写入持久性存储装置160。注意，即使当更新日志的一部分被写入持久性存储装置160时，也不会使那些更新成为永久的(即“已提交”)，直到关于被认为是永久的更新程度的显式指示符被写入到持久性存储装置为止。

[0088] 在根140已确定已在所有适当的工作器处复制了与时间tb以及更早时间相关联的所有请求和响应的时间，T2达到或增加至tb。在将时间T2＝tb从根140分配给主要工作器和备用工作器150之后，这些工作器使响应永久地保存在持久性存储装置160中。如果经过该

时间tb的更新日志尚未写入持久性存储器，则它们在那时被写入。更一般地，到时间T2达到或增加到tb为止，经过时间tb的日志已由工作器写入到持久性存储装置160，并且所有必须在此时完成的是通过记录将在持久日志中经过时间tb的更新视为永久的指示符来完成使更新变为永久的任务。在主要工作器正在将日志永久化的可能短时间期间，它处于状态D。当主要工作器已将对图4中所展示的请求的响应保存在持久性存储装置中时，它进入状态E。类似地，当备用工作器使响应永久化时，它处于状态H，并且当备用工作器使响应永久地保存在持久性存储器中时，它进入状态I。当根确定与时间tb(以及更早的时间)相关联的所有响应都永久地保存在持久性存储器中(即，都处于状态E或I)时，根将持久性时间T3增加到tb。如上所述，对于在调用集群组件处的请求的代管时间是持久性时间T3的情况，根140通知调用集群组件110代管时间已等于或大于tb，并且调用集群组件110将针对该请求113和响应115的相应结果105释放到应用程序(例如，图)内的一个或多个其他组件。

[0089] 如上所述，在正常操作中，当在集群中处理来自调用集群组件的连续请求113时，根更新工作时间T1，响应115返回到调用集群组件、并根据代管时间T2或T3的更新从调用集群组件释放到图。通常，对特定请求113的处理可能花费工作时间T1的许多时间“节拍”，例如数以十计或数以百计的节拍，并且因此集群可能有许多进行中的请求、以及与这些请求相关联的许多不同的请求时间。此外，因为数据分布在工作器之间，因此负荷根据那些请求的密钥有效地分布在工作器之间，使得每个工作器可能具有该工作器正在充当主要工作器的多个请求(即，处于状态A-E之一)，并且还具有该工作器正在充当备用工作器的多个请求(即，处于状态F-I之一)。

[0090] 要注意的是，到达集群的用于执行任务的一些请求使用如本文中描述的过程来复制任务并复制执行该任务的相应结果。例如，在任务已被标记并在备用工作器处复制(但不一定使其持久化)之后，该任务将在主要工作器处初始化。如果任务对数据记录进行操作，则初始化可能涉及保留记录的原始版本1。然后，该任务在主要工作器上执行，而在备用工作器上保持休眠。在处理已经完成之后，将存在记录的修改后的版本2。然后，任务的终止可以包括将记录的修改后的版本2从主要工作器发送到备用工作器。然后，主要工作器和备用工作器两者都能够删除记录的原始版本1(以及复制的任务)。这些步骤中的每一个都是合理高效的，但是如果任务持续时间很短，则与这些初始化和终止过程相关联的开销可能会使任务效率降低。

[0091] 可替代地，对于持续时间相对较短的一些任务(“短任务”)，可以使用不同的过程。短任务仍在备用工作器处被标记并复制。但是，初始化不需要保留记录的原始版本1。作为替代，在提交操作指示短任务和短任务的复制品已分别持久地存储在主要工作器和备用工作器上之后，短任务在这两个工作器上执行。在该执行结束时，将在主要工作器和备用工作器两者上都有记录的修改后的版本2的副本，而无需任何通信来传输修改后的记录。这两个工作器都有冗余处理，但是由于任务很短，所以此冗余不会对效率产生很大影响。例如，如果短任务是确定性的，并且不管哪个工作器执行该任务都产生相同的结果，则此替代性过程是有用的。

[0092] 2正常操作的示例

[0093] 参考图5至图12，展示了调用集群组件110和集群120的正常操作的一个示例。在图5中，输入记录103到达调用集群组件110，并且调用集群组件110形成针对输入记录103的请

求113。调用集群组件110将请求113与唯一请求标识符rid相关联，并将其存储在调用集群组件110的重播缓冲器112中。

[0094] 调用集群组件110将请求113传输到集群120，并在时间T1＝ta处在集群120中的主要工作器150a(工作器A)处接收到该请求。请求113被存储在主要工作器150a的易失性存储器155中，并且被分派有等于当前工作时间(T1＝ta)的请求时间。将请求113的请求时间提供给调用集群组件110，该调用集群组件将请求时间(即，ta)与存储在重播缓冲器112中的请求113相关联。存储在调用集群组件110的重播缓冲器112中的请求113处于状态A(参见图4)，以等待来自集群120的响应。存储在主要工作器的易失性存储器155中的请求113处于状态A，以等待分派计算资源以执行请求113。

[0095] 参考图6，主要工作器将请求113发送到备用工作器150b(工作器B)，其中，该请求被存储在备用工作器150b的易失性存储器155中。存储在备用工作器150b的易失性存储器155中的请求113处于状态F，以等待接收来自主要工作器的响应。

[0096] 参考图7，一旦主要工作器105将计算资源(例如主要工作器或集群的另一部分的计算资源)分派给请求113，则请求113在主要工作器105处进入状态B并开始执行。

[0097] 参考图8，在时间T1＝tb，主要工作器105完成了请求113的执行。请求113的执行生成响应115，该响应被存储在主要工作器的易失性存储器156中。响应115与请求113的请求标识符(rid)以及生成该响应的时间(tb)相关联。主要工作器将响应115发送到调用集群组件110和备用工作器150b，并且然后请求113处于状态C，以等待持久性时间T3达到tb。

[0098] 调用集群组件110接收响应115，并将该响应存储在其代管缓冲器114中。在将响应存储在代管缓冲器114的情况下，结果115在调用集群组件110处于状态B，以等待持久性时间T3(在此示例中为代管时间)达到tb。备用工作器150b接收响应115，并将该响应存储在该备用工作器的易失性存储器156中。请求113在备用工作器150b处进入状态G，以等待持久性时间T3达到tb。

[0099] 尽管在图8中未示出，在将响应115存储(复制)在主要工作器150a和备用工作器150b的易失性存储器156的情况下，复制时间T2被设置为tb。

[0100] 参考图9，一旦响应115被存储在主要工作器150a和备用工作器150b中的一个或两个的易失性存储器156中，主要工作器150a和备用工作器150b就开始将响应115存储到各自的持久性存储装置160，同时还保持存储在各自的易失性存储器155、156中。

[0101] 参考图10，在将响应115存储在主要工作器处并在备用工作器150b处复制响应之后，将持久性时间T3设置为tb。主要工作器150a和备用工作器150b最终将响应115永久存储在持久性存储装置160中。存储在主要工作器处的请求113处于状态D，并且存储在备用工作器150b处的请求113处于状态H，在这些状态下，请求113和响应115仍分别存储在易失性存储器155、156中。

[0102] 参考图11，此示例的代管时间是持久性时间T3，因此在T3更新为tb的情况下，存储在调用集群组件110处的请求113进入状态C，并且从该调用集群组件的代管缓冲器114释放响应115(其与时间tb相关联)。

[0103] 参考图12，在响应115被永久地存储在主要工作器150a的持久性存储装置中的情况下，请求113进入状态E，在该状态下，请求113和响应115均未分别存储在其易失性存储器155、156中。类似地，在响应115被永久地存储在备用工作器150b的持久性存储装置中的情

况下,请求113进入状态I,在该状态下,请求113和响应115均未存储在其易失性存储器155、156中。

[0104] 3回滚场景

[0105] 尽管图4中的状态转换图表示正常操作,但有可能(但很少)未成功接收到工作器之间的消息。此外,工作器有可能在丢失其易失性存储器之后必须重新启动,或者工作器可能完全故障,使得工作器不进一步处理请求(即,担任主要角色或备用角色)。要注意的是,本文中描述的数据处理系统的一些实施例实施了本节中描述的所有回滚场景。还要注意的是,数据处理系统的其他实施例可以实施本节中描述的一个或多个但并非所有回滚场景。

[0106] 3.1场景1:tr<ta

[0107] 首先考虑集群确定存在一些未成功接收到的工作器间消息并且该消息与时间te相关联的情况。一般而言,根通知所有工作器必须将时间"回滚"到te之前的时间tr(即,tr<te),例如,回滚到tr=te-1。即使通过这种回滚,也可以将调用集群组件110提供的结果提供给应用程序或图,就像没有发生回滚一样,并且对分布在工作器之间的数据的更新保持与由调用集群组件提供的结果一致。具体地,直到结果被存储(例如,复制或持久化)在多个节点(例如,工作器)处之后,该结果才从调用集群组件110释放到应用程序或图,从而确保该结果将永远不会被重新调用或变为失效。换言之,发生的任何回滚必然发生在由调用集群组件110将结果提供给应用程序或图之前。

[0108] 当根140确定因为未成功接收到一些工作器间消息而必须执行回滚时,根将回滚时间tr通知给调用集群组件110。当前时间T1增加,并且通常,从时间tr+1直到且包括T1-1的所有活动都被视为这些活动未发生。在调用集群组件110处的影响是,存储在重播缓冲器112中的处于状态B(即,代管时间尚未达到响应时间)的所有请求都返回到状态A,并且代管缓冲器114中的任何相应的响应115被丢弃。然后,处于状态A的请求113(因为这些已经处于状态A或已从状态B返回到状态A)被重新发送到集群120。

[0109] 针对请求的请求时间ta大于回滚时间tr(即,tr<ta)的情况,首先考虑集群中(即,在工作器150处)对尚未开始执行但已在主要工作器和备用工作器之间复制的请求(即,主要工作器处于状态A并且备用工作器处于状态F)的影响。在此展示中,当前工作时间表示为tc。因为ta大于tr,所以调用集群组件无法假定请求已正确复制,并且因此移除了存储在主要工作器和备用工作器的易失性存储器155中的请求的版本。在集群120处从调用集群组件110接收具有相同请求id rid的请求113,并将该请求与新的请求时间tc相关联。当主要工作器接收到请求113时,该主要工作器在状态A下将请求113存储在其易失性存储器155中。主要工作器将请求113发送到(多个)备用工作器150,该备用工作器在状态F下将请求113存储在其易失性存储器155中。然后在主要工作器和备用工作器上以图4所展示的方式进行进一步的处理。

[0110] 注意,如果备用工作器在从主要工作器接收到具有时间tc的已更新请求之前不知道该请求,则备用工作器也将以与现在已正确复制了该请求相同的方式进行。

[0111] 参考图13至图15,示出了第一回滚场景的一个示例。在图13中,在时间ta发布的请求113被存储在调用集群组件110处的重播缓冲器112中并且处于状态A。请求113被存储在主要工作器的易失性存储器155中并且处于状态A,因为该请求尚未开始执行。请求113还存储在备用工作器150b中并且处于状态F。

[0112] 接收到回滚请求以将系统回滚到时间tr<ta。在图14中，在接收到回滚请求之后，从主要工作器150a的易失性存储器155和备用工作器150b的易失性存储器155中移除请求113。由调用集群组件110将同与原始请求113相同的请求标识符(rid)相关联的新请求113'发布到集群120。在时间tc，新请求113'由集群120接收并且与请求时间tc相关联。集群120向调用集群组件110通知与新请求113'相关联的请求时间tc。重播缓冲器112中的新请求113'处于状态A。

[0113] 在集群中，将新请求113'发送到主要工作器。主要工作器150a将新请求113'与请求时间tc一起存储在其易失性存储器155中。存储在主要工作器150a的易失性存储器155中的新请求113'处于状态A。

[0114] 参考图15，主要工作器将新请求113'发送到备用工作器150b。备用工作器150b将新请求113'存储在其易失性存储器155中，并将其与请求时间tc相关联。存储在备用工作器的易失性存储器155中的已更新请求113'处于状态F。

[0115] 然后，集群根据其正常操作进行(如图5至图12所阐述的)。

[0116] 3.2场景2:tr<ta，已开始执行

[0117] 在第二种情况下，先前请求的请求时间ta大于回滚时间tr(即，tr<ta)，但是该请求已开始执行并且尚未在主要工作器上完成执行(即，请求在主要工作器处于状态B，其中可能计算了部分响应115，而在备用工作器处请求处于状态F)。在这种情况下，在主要工作器和备用工作器处终止执行并丢弃部分响应115(或允许执行完成，并且丢弃响应)，并且调用集群组件110将请求113重新发送到集群120。存储在主要工作器和备用工作器处的请求分别返回到状态A和F。主要工作器以与尚未在主要工作器处开始执行请求相同的方式将请求通知给备用工作器。

[0118] 参考图16至图18，示出了第二回滚场景的一个示例。在图16中，在时间ta发布的请求113被存储在调用集群组件110处的重播缓冲器112中并且处于状态A。请求113被存储在主要工作器150a处的易失性存储器155中并且处于状态B，因为该请求已开始执行。该请求还存储在备用工作器150b中并且处于状态F。

[0119] 接收到回滚请求以将系统回滚到时间tr<ta。在图17中，在接收到回滚请求之后，从主要工作器150a的易失性存储器155和备用工作器150b的易失性存储器155中移除请求113。由调用集群组件110将同与原始请求113相同的请求标识符(rid)相关联的新请求113'发布到集群120。在时间tc，新请求113'由集群120接收并且与请求时间tc相关联。集群120向调用集群组件110通知与新请求113'相关联的请求时间tc。重播缓冲器112中的新请求113'处于状态A。

[0120] 在集群中，将新请求113'发送到主要工作器。主要工作器150a将新请求113'与请求时间tc一起存储在其易失性存储器155中。存储在主要工作器150a的易失性存储器155中的新请求113'处于状态A。

[0121] 参考图18，主要工作器150a将新请求113'发送到备用工作器150b。备用工作器150b将新请求113'存储在其易失性存储器155中，并将其与请求时间tc相关联。存储在备用工作器的易失性存储器155中的已更新请求113'处于状态F。

[0122] 然后，集群根据其正常操作进行(如图5至图12所阐述的)。

[0123] 3.3场景3:tr<ta<tb，执行已完成

[0124] 在第三种情况下，先前请求的请求时间ta再次大于回滚时间tr。然而，在这种情况下，我们假设在时间tb完成执行(即，tr<ta≤tb)，并且响应已在备用工作器处复制并在调用集群组件110处接收。即，请求113在调用集群组件110处于状态B，该请求在主要工作器150a处于状态C，并且请求113在备用工作器150b处于状态G。不是仅仅像第二种情况那样必须终止执行正在进行的执行，而是移除已存储在主要工作器和备用工作器处的响应115。如上参考图4所述，在时间tb生成的响应以使得可以将特定时间以及更晚时间的所有更新从数据结构中移除的方式与时间tb相关联地存储在版本化数据结构中。在本情况下，通过移除在晚于时间tr的时间更新的所有数据版本，必须移除在时间tb进行的对所展示请求的更新，并且具有请求时间tc的请求在主要工作器处返回到状态A以等待执行，并且在备用工作器中返回到状态F以等待来自主要工作器的响应。在调用集群组件处，响应被丢弃，并且请求返回到状态A。

[0125] 参考图19至图21，示出了第三回滚场景的一个简单示例。在图19中，在时间ta发布的请求113被存储在调用集群组件110处的重播缓冲器112中。在时间tb生成的对请求的响应115被存储在代管缓冲器114中。因此，请求113在调用集群组件处于状态B。

[0126] 在集群中，请求113和响应115被存储在主要工作器150a处的易失性存储器155、156中。因此，请求113在主要工作器150a处于状态C。请求113和响应115还被存储在备用工作器处的易失性存储器155、156中。因此，请求在备用工作器150b处于状态G。

[0127] 接收到回滚请求以将系统回滚到时间tr<ta<tb。在图20中，在接收到回滚请求之后，将响应115从调用集群组件110的代管缓冲器114中移除。在集群120中，从主要工作器150a的易失性存储器155和备用工作器150b的易失性存储器155中移除请求113和响应115两者。

[0128] 由调用集群组件110将同与原始请求113相同的请求标识符(rid)相关联的新请求113'发布到集群120。在时间tc，新请求113'由集群120接收并且与请求时间tc相关联。集群120向调用集群组件110通知与新请求113'相关联的请求时间tc。重播缓冲器112中的新请求113'处于状态A。

[0129] 在集群中，新请求113'被发送到主要工作器150a。主要工作器150a将新请求113'与请求时间tc一起存储在其易失性存储器155中。存储在主要工作器150a的易失性存储器155中的新请求113'处于状态A。

[0130] 参考图21，主要工作器150a将新请求113'发送到备用工作器150b。备用工作器150b将新请求113'存储在其易失性存储器155中，并将其与请求时间tc相关联。存储在备用工作器的易失性存储器155中的已更新请求113'处于状态F。

[0131] 然后，集群根据其正常操作进行(如图5至图12所阐述的)。

[0132] 3.4场景4:ta<tr，执行尚未开始

[0133] 在第四种情况下，回滚时间tr是在原始请求时间ta处或其之后(即，ta≤tr)，并且原始请求尚未开始执行。该请求被重传到集群120，并且在主要工作器和备用工作器处将该请求排在原始请求(即，{rid,ta})之后执行。主要工作器执行原始请求并生成响应(即，{rid,tb})。然后，主要工作器进行到开始执行重传的请求(即，{rid,tc})，但检测到已经存在与重传的请求的rid相关联的响应，并放弃执行重传的请求。

[0134] 参考图22至图25，示出了第四回滚场景的一个示例。在图22中，在时间ta发布的原

始请求113被存储在调用集群组件110处的重播缓冲器112中并且处于状态A。原始请求113被存储在主要工作器150a的易失性存储器155中并且处于状态A，因为该请求尚未开始执行。原始请求113还存储在备用工作器150b中并且处于状态F。

[0135] 接收到回滚请求以将系统回滚到时间ta<tr。在图23中，由调用集群组件110将同与原始请求113相同的请求标识符(rid)相关联的新请求113'发布到集群120。在时间tc，新请求113'由集群120接收并且与请求时间tc相关联。集群120向调用集群组件110通知与新请求113'相关联的请求时间tc。重播缓冲器112中的请求113保持处于状态A。

[0136] 在集群中，新请求113'被发送到主要工作器150a。主要工作器150a接收新请求113'，并将新请求113'排在原始请求113之后执行。存储在主要工作器150a的易失性存储器155中的原始请求113和新请求113'两者处于状态A。

[0137] 参考图24，主要工作器150a将新请求113'发送到备用工作器150b。备用工作器150b接收新请求113'，并将新请求113'排在原始请求113之后执行。存储在备用工作器150b的易失性存储器155中的原始请求113和新请求113'两者都处于状态F。

[0138] 参考图25，主要工作器150a已经执行了原始请求113以生成响应115，并且响应115在该主要工作器的持久性存储装置160中被持久化。结果，原始请求113在主要工作器150a处于状态D。新请求113'尚未在主要工作器150a处开始执行，并且因此处于状态A。

[0139] 还已将响应115提供给备用工作器150b和调用集群组件110。备用工作器150b已将响应115存储在其易失性存储器156中，并已将响应持久化到其持久性存储装置160。因此，原始请求113在备用工作器处于状态H。调用集群组件110已将响应115存储在其代管缓冲器114中，并且调用集群组件的重播缓冲器112中的请求113处于状态B。

[0140] 当新请求113'在主要工作器150a处开始执行时，主要工作器150a识别出新请求113'同与响应115相同的请求标识符rid相关联，并且因此不执行新请求113'，因为它是复制品。在一些示例中，可以将响应115重传到调用集群组件，该调用集群组件将响应115作为复制品而忽略。

[0141] 然后，集群根据其正常操作进行(如图5至图12所阐述的)。

[0142] 3.5场景5:ta<tr,已开始执行

[0143] 在第五种情况下，回滚时间tr是在原始请求时间ta处或其之后(即，ta≤tr)，并且原始请求已经开始执行，但在主要工作器处尚未完成执行(即，请求在主要工作器处于状态B，而请求在备用工作器处于状态F)。在这种情况下，在主要工作器和备用工作器处终止执行(或允许完成，并且丢弃响应)(即，存储在主要工作器和备用工作器处的请求分别返回到状态A和F)。

[0144] 调用集群组件110将该请求重传到集群120，其中，在主要工作器和备用工作器处将该请求排在原始请求(即，{rid,ta})之后执行。主要工作器执行原始请求并生成响应(即，{rid,tb})。然后，主要工作器进行到开始执行重传的请求(即，{rid,tc})，但检测到已经存在与重传的请求的rid相关联的响应，并放弃执行重传的请求。

[0145] 参考图26至图29，示出了第五回滚场景的一个示例。在图26中，在时间ta发布的原始请求113被存储在调用集群组件110处的重播缓冲器112中并且处于状态A。原始请求113被存储在主要工作器150a处的易失性存储器155中并且处于状态B，因为该请求已开始执行。原始请求113还存储在备用工作器150b中并且处于状态F。

[0146] 接收到回滚请求以将系统回滚到时间ta<tr。在图27中，由调用集群组件110将同与原始请求113相同的请求标识符(rid)相关联的新请求113'发布到集群120。在时间tc，新请求113'由集群120接收并且与请求时间tc相关联。集群120向调用集群组件110通知与新请求113'相关联的请求时间tc。重播缓冲器112中的请求113保持处于状态A。

[0147] 在集群120中，存储在主要工作器150a的易失性存储器155中的原始请求113的执行被终止，并且原始请求113返回到状态A。新请求113'被发送到主要工作器150a。主要工作器150a接收新请求113'，并将新请求113'排在原始请求113之后执行。存储在主要工作器150a的易失性存储器155中的新请求113'处于状态A。

[0148] 参考图28，主要工作器150a将新请求113'发送到备用工作器150b。备用工作器150b接收新请求113'，并将新请求113'排在原始请求113之后执行。存储在备用工作器150b的易失性存储器155中的原始请求113和新请求113'两者都处于状态F。

[0149] 参考图29，主要工作器150a已执行了原始请求113并且已生成了响应115。响应115在该主要工作器的持久性存储装置160中被持久化。结果，原始请求113在主要工作器150a处于状态D。新请求113'尚未在主要工作器150a处开始执行，并且因此处于状态A。

[0150] 还已将响应115复制到备用工作器150b和调用集群组件110。备用工作器150b已将响应115存储在其易失性存储器156中，并已将响应持久化到其持久性存储装置160。因此，原始请求113在备用工作器处于状态H。调用集群组件110已将响应115存储在其代管缓冲器114中，并且请求113'在调用集群组件的重播缓冲器112中处于状态B。

[0151] 当新请求113'在主要工作器150a处开始执行时，主要工作器150a识别出新请求113'同与响应115相同的请求标识符rid相关联，并且因此不执行新请求113'，因为它是复制品。在一些示例中，响应115可以被重传到调用集群组件110，调用集群组件将响应115作为复制品而忽略。

[0152] 然后，集群根据其正常操作进行(如图5至图12所阐述的)。

[0153] 3.6场景6:ta<tb<tr,执行已完成

[0154] 在第六种情况下，回滚时间tr是在请求时间ta处或其之后，并且请求已在时间tb处完成执行，该时间tb也在该回滚时间处或其之前(即，ta≤tb≤tr)。如果响应已成功提供给调用集群组件110(即，该请求在调用集群组件处于状态B)，则回滚请求不会导致重新发送该请求，该回滚请求也不会导致从代管缓冲器114移除任何响应。即，与ta相关联的任何请求和与tb相关联的任何响应均保持不变。

[0155] 但是，如果未将响应成功提供给调用集群组件110，则调用集群组件110将请求重传到集群120。当主要工作器接收到重传的请求时，主要工作器开始执行重传的请求(即，{rid,tc})，但检测到已经存在与请求标识符rid相关联的响应115。因此，不执行重传的请求，并且将通过执行原始请求而生成的响应重传到调用集群组件110。调用集群组件110接收具有响应时间tb的响应，该响应时间用于确定何时可以从调用集群组件处的代管发送响应。

[0156] 参考图30至图32，示出了第六回滚场景的一个示例。在图30中，在时间ta发布的原始请求113被存储在调用集群组件110处的重播缓冲器112中。在时间tb生成了对原始请求113的响应115，但该响应未到达调用集群组件110的代管缓冲器114。因此，请求113在调用集群组件110处于状态A。

[0157] 在集群中，请求113和响应115被存储在主要工作器150a的易失性存储器155、156中。因此，请求113在主要工作器150a处于状态C。请求113和响应115还被存储在备用工作器处的易失性存储器155、156中。因此，请求在备用工作器150b处于状态G。

[0158] 接收到回滚请求以将系统回滚到时间ta<tb<tr。在图31中，由调用集群组件110将同与原始请求113相同的请求标识符(rid)相关联的新请求113'发布到集群120。在时间tc，新请求113'由集群120接收并且与请求时间tc相关联。集群120向调用集群组件110通知与新请求113'相关联的请求时间tc。

[0159] 新请求113'被发送到集群120中的主要工作器150a。主要工作器150a接收新请求113'，并将新请求113'在易失性存储器155中排队以便执行。存储在主要工作器150a的易失性存储器155中的原始请求113保持处于状态C，并且存储在主要工作器150a的易失性存储器155中的新请求113'处于状态A。

[0160] 参考图32，当主要工作器150a开始执行新请求时，主要工作器150a识别出新请求113'具有与原始请求113相同的请求标识符rid并且在主要工作器150a处已经存在与请求标识符rid相关联的响应115。因此，主要工作器150a不执行新请求113'，而是将响应115重传到调用集群组件110。调用集群组件110接收响应115，并将该响应存储在代管缓冲器114中。在响应115被存储在调用集群组件110的代管缓冲器114中的情况下，调用集群组件110处于状态B。

[0161] 然后，集群根据其正常操作进行(如图5至图12所阐述的)。

[0162] 3.7场景7:ta<tr<tb,执行已完成

[0163] 在第七种情况下，回滚时间tr是在请求时间ta处或其之后，并且请求已在回滚时间之后的时间tb处完成执行(即，ta≤tr<tb)，工作器之间的响应的复制可能没有成功。工作器丢弃具有tr之后的时间的所有响应115。存储在备用工作器中的请求113返回到状态F，并且存储在主要工作器中的请求113返回到状态B。调用集群组件110丢弃代管缓冲器114中的所有响应115，将存储在重播缓冲器112中的请求113返回到状态A，并且将请求113重新发送到集群120，该集群重新处理该请求。

[0164] 参考图33至图35，示出了第七回滚场景的一个示例。在图33中，在时间ta发布的请求113被存储在调用集群组件110处的重播缓冲器112中。在时间tb生成的对请求的响应115被存储在代管缓冲器114中。因此，请求113在调用集群组件110处于状态B。

[0165] 在集群120中，请求113和响应115被存储在主要工作器150a处的易失性存储器155、156中。因此，请求113在主要工作器150a处于状态C。请求113还存储在备用工作器105处的易失性存储器155、156中，但是响应115可能已经或可能没有成功地复制到备用工作器150b。因此，请求在备用工作器150b处可能处于或可能不处于状态G。

[0166] 接收到回滚请求以将系统回滚到时间ta<tr<tb。在图34中，将存储在调用集群组件110的代管缓冲器114中的响应115移除。由调用集群组件110将同与原始请求113相同的请求标识符(rid)相关联的新请求113'发布到集群120。在时间tc，新请求113'由集群120接收并且与请求时间tc相关联。集群120向调用集群组件110通知与新请求113'相关联的请求时间tc。重播缓冲器112中的新请求113'处于状态A。

[0167] 在集群120中，备用工作器150b移除存储在其易失性存储器156中的、与tr之后的时间相关联的任何响应，并且因此回到状态F。主要工作器150a返回到状态B。新请求113'被

发送到主要工作器150a。主要工作器接收新请求113’，并将新请求113’排在原始请求113之后执行。存储在主要工作器150a的易失性存储器155中的新请求113’处于状态A。

[0168] 在图35中，主要工作器150a在时间td完成原始请求113的执行并生成新响应115’。主要工作器150a将新响应115’发送到备用工作器150b和调用集群组件110，以使存储在主要工作器150a的易失性存储器中的原始请求113的状态转换为状态C。备用工作器150b接收新响应115’，并将新响应115’存储在其易失性存储器155中，以使存储在备用工作器的易失性存储器155中的原始请求113转换为状态G。调用集群组件110接收新响应115’，并将其存储在代管缓冲器114中，以使存储在重播缓冲器112中的新请求113’转换为状态B。

[0169] 当新请求113’在主要工作器150a处开始执行时，主要工作器150a识别出新请求113’具有与原始请求113相同的请求标识符rid，并且因此不执行新请求113’，因为它是复制品。

[0170] 然后，集群根据其正常操作进行(如图5至图12所阐述的)。

[0171] 3.8场景8:ta<tr<tb，执行已完成

[0172] 最后，在第八种情况下，正在处理请求的作为主要工作器的工作器丢失(例如，已知故障)。一般而言，对于正在等待丢失的主要工作器来提供响应(即，备用工作器处于状态F)的备用工作器处的任何请求，该备用工作器被提升为主要工作器。当根140例如由于未能从该工作器接收到对消息的应答而检测到工作器丢失时，根启动到时间tr的回滚，该时间等于上次复制的时间(即，tr＝T2)。当备用工作器接收了到时间tr的回滚请求时(可能伴随新分区信息以适应丢失的工作器)，备用工作器通过将请求的状态更改为其正在等待资源以执行请求的状态A而开始充当新的主要工作器。

[0173] 参考图36至图37，示出了第八回滚场景的一个示例。在图36中，在时间ta发布的请求113被存储在调用集群组件110处的重播缓冲器112中并且处于状态A。请求113被存储在主要工作器150a处的易失性存储器155中并且处于状态B，因为该请求已开始执行但尚未完成执行。该请求还存储在备用工作器150b中并且处于状态F。在执行请求113期间，主要工作器150a故障或丢失。

[0174] 在图37中，根已请求回滚到时间tr，该时间等于上次复制的时间。此时，备用工作器150b被提升为主要工作器150a，并将其状态更改为状态A。另一工作器150c被分派为处于状态F的备用工作器。

[0175] 然后，集群根据其正常操作进行(如图5至图12所阐述的)。

[0176] 4根节点

[0177] 现在转向根140的操作，如上所述，根周期性地增加当前工作时间(间隔)T1 144。一般而言，当根更新工作时间时，根向所有工作器分配(例如，广播)时间元组(T1，T2，T3)144-146。作为响应，工作器向根提供信息，根可以基于该信息更新T2和/或T3时间。

[0178] 每个工作器维护与特定工作时间相关联的一组计数器151-152。一个计数器151与工作时间t1相关联(称为Sent(t1))，该计数器对已从该工作器发送到备用工作器的、具有请求时间t1的请求的通信数量进行计数，并对已发送到备用工作器的具有响应时间t1的响应的数量进行计数。在图4中，在状态A下，针对发送到备用工作器的具有请求时间ta的每个请求来更新Sent(ta)，并且对于被发送用于在备用工作器处的复制的在时间tb生成的每个响应，增加Sent(tb)。注意，对于从工作器发送到调用集群组件的消息，Sent()计数器不会

增加。另一计数器152Rec(t1)对在工作器处接收到的与时间t1相关联的通信数量进行计数。具体地，当备用工作器在其进入状态F时接收到具有请求时间ta的请求的复制时，备用工作器增加Rec(ta)，并且当备用工作器在进入状态G时接收到在时间tb生成的响应的复制时，备用工作器增加Rec(tb)。每个工作器具有其自己的这些计数器的本地副本，针对工作器w表示为Sentw(t)和Recw(t)。应当明显的是，在与时间t1相关联地发送的所有通信也都在其目的地接收到的程度上，所有工作器w的Sentw(t)的汇总之和等于工作器w的Recw(t)的汇总之和。

[0179] 不时地，例如响应于从根140接收到当前时间(T1,T2,T3)的广播，每个工作器150针对大于复制时间T2的所有时间发送其当前计数Sent(t)151和Rec(t)152。这些计数在根处接收并汇总，使得根针对大于T2的每个时间t确定Sent(t)之和以及Rec(t)之和，并将它们与相应时间相关联地存储在计数器141和142中。如果Sent(T2+1)等于Rec(T2+1)，则从时间T2+1开始的所有传输都已收到，并且增加T2以成为下一个复制时间。重复此过程，直到Sent(T2+1)不等于Rec(T2+1)或T2+1达到T1为止。然后，此增加的T2时间(145)用于下一次从根开始的广播。

[0180] 如上所述，首先在易失性存储器中记录工作器处的数据更新，其中，日志被不时地写入持久性存储装置中。针对复制时间T2之前的更改，每个工作器都可以自由地将持久性存储器中的所记录更改永久化。通常，每个工作器w都已有机会使经过时间T3(w)的所有更改永久化，其中不同的工作器通常已达到了不同的时间。除了响应当前时间的广播而将Rec()和Sent()返回到根之外，每个工作器还返回其T3(w)时间，该时间根据根处或回到根的通信路径上的min()操作进行汇总。即，根确定T3＝minw T3(w)，并且然后根在下一次分配当前时间时分配T3的此新值。

[0181] 在一些实施例中，根在根与每个工作器之间的直接(例如，单播)通信中分配时间元组(T1,T2,T3)。在其他实施例中，元组以另一方式(诸如基于泛洪的广播)分配。在另一实施例中，元组沿着预定的树状结构分发网络分配，其中元组的每个接收者将元组转发给多个其他接收者，使得最终所有工作器都已接收到时间元组。

[0182] 来自工作器的计数的汇总可以通过每个工作器与根节点之间的单播通信来执行，其中，根对所有工作器执行完整的求和。作为一种更高效的解决方案，可以沿与时间元组相同的路径将计数发送回去，其中，路径中的中间节点执行计数总和的部分汇总，从而与根一起分担求和的负担，但仍然获得了关于所有工作器的计数总和。

[0183] 在替代性操作模式中，当复制响应时间而不是持久时间时，可以从调用集群组件中释放响应。以这种方式，可以以较小的延迟将响应提供给图，其中存在该响应可能尚未在集群存储中持久化的可能性。

[0184] 如上所述，将执行请求的响应存储在版本化数据结构中。在一个这种数据结构中，数据项的每次更新都存储为可单独恢复的版本，并利用与该更新相关联的时间标记该版本。例如，可以至少在概念上针对每个访问密钥将数据结构存储为元组(tb值)的列表，其中，tb是值的更新时间。不同时间的值可以共享子结构或使用其他存储优化。在一些示例中，基于时间之间数据值的编辑来存储这些值。作为一个示例，这些值可以表示为基于树的结构，并且每个版本可以存储为足以用于从先前版本创建下一版本的“前向”增量操作，或者存储为足以用于从当前版本重建先前版本的“后向”增量操作。如上所讨论的，这种版本

化数据结构允许在回滚时间之后回滚所有更新。不是维护对数据项的所有更新，而是仅保留相对于更新时间的开始的更新，以便可以完成回滚到任何更新时间的开始。

[0185] 应当认识到，在根增加复制时间T2之后，将不会要求工作器回滚到该时间或其之前的版本。因此，版本化数据结构的优化是可以从数据结构中移除复制时间T2或其之前的版本。

[0186] 在一些实施例中，某些请求在它们的执行时间很小的意义上是“轻量级的”，并且因此，在备用工作器处执行该请求可以比将响应从主要工作器复制到备用工作器消耗更少的资源。在这种实施例中，不执行从主要工作器到(多个)备用工作器的响应复制。每个工作器可以在不同的时间完成处理。为了维护工作器之间的数据同步，主要工作器如上所述分配完成时间tb，而备用工作器则将其本地计算的响应视为这些响就像是在那时计算的。

[0187] 在替代性实施例中，调用集群组件在其从根接收时间元组并且将Sent()和Rec()计数返回到根的意义上参与集群。在此实施例中，调用集群组件为请求分派请求时间，该时间在请求的复制期间由工作器使用。当发生回滚时，因为调用集群组件知道其保存的请求的请求时间，因此仅需在回滚时间之后重新发送请求，而不会丢弃在回滚时间或其之前生成的响应。修改工作器的操作以适应调用集群组件的此操作。

[0188] 5替代方案

[0189] 更一般地，在上述回滚场景4-8中，其中，ta<tr，当调用集群组件110重传请求时，调用集群组件不知道(也不在乎)原始请求在时间ta被传输。另一方面，集群120需要考虑原始请求的请求时间，因为集群使用该时间来确定是否回滚。因此，当调用集群组件110将请求(具有请求标识符rid)重新发送到集群120使得ta<tr<tc时，该请求在主要工作器150a处接收并与时间tc相关联。主要工作器150a将请求转发到备用工作器150b。在这种情况下，主要工作器可以在其执行重新发送的请求(即，{rid,tc})之前执行原始请求(即，{rid,ta})。当主要工作器150a进行到执行重新发送的请求(即，{rid,tc})时，因为对原始请求(即，{rid,ta})的响应已经持久化，因此主要工作器将重新发送的请求视为复制品。

[0190] 在一些示例中，请求产生后续任务(有时称为‘任务链’)。在此类示例中，直到产生的任务完成之后，才生成对请求的响应。在一些示例中，如果已存储了对请求{rid,ta}的响应，则工作器将其响应返回到调用集群组件。但是，如果因为请求{rid,ta}尚未完成因此对请求{rid,ta}的响应还不存在，则因为集群知道原始请求最终将完成并生成被返回到调用集群组件的响应，因此具有复制品rid的后续请求{rid,tc}被忽略。

[0191] 在上述示例中，当集群接收到请求时，集群将时间(例如ta)与该请求相关联，并且然后将该时间通知给调用集群组件。调用集群组件将时间与存储在其重播缓冲器中的请求相关联。在回滚的情况下，调用集群组件可以使用与调用集群组件的重播缓冲器中的请求相关联的时间来选择性地重播请求。但是，在一些示例中，集群和调用集群组件都不会将请求与时间相关联。在这些示例中，当在回滚场景的情况下重播请求时，调用集群组件的选择性较低。例如，在回滚请求的情况下，调用集群组件可以系统地重播其重播缓冲器中的所有请求。

[0192] 6实施方式

[0193] 上述计算集群管理方法可以例如使用执行合适的软件指令的可编程计算系统来实施，或者该方法可以在合适的硬件中实施，比如现场可编程门阵列(FPGA)或以某种混合

形式。例如，在程控方法中，软件可以包括一个或多个计算机程序中的在一个或多个程控的或可编程计算系统(其可以是各种体系架构，诸如分布式客户端/服务器、或电网)上执行的过程，该计算系统各自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/或非易失性存储器和/或存储元件)、至少一个用户接口(用于使用至少一个输入设备或端口接收输入，并且用于使用至少一个输出设备或端口提供输出)。软件可以包括较大程序的一个或多个模块，该较大程序例如提供与对数据流图的设计、配置、和执行有关的服务。程序模块(例如，数据流图的元素)可以被实施为数据结构或符合存储在数据储存库中的数据模型的其他经组织的数据。

[0194] 软件可以以非暂态形式存储一段时间(例如，如动态RAM等动态存储器设备的刷新周期之间的时间)，诸如使用介质的物理性质(例如，表面凹坑和岸台、磁畴、或电荷)体现在易失性或非易失性存储介质中、或任何其他非暂态介质中。在准备加载指令时，软件可以提供在如CD-ROM或其他计算机可读介质(例如，由通用或专用计算系统或设备可读)等有形、非暂态介质上，或者可以通过网络的通信介质递送(例如，被编码到传播信号中)到其被执行的计算系统的有形、非暂态介质。该处理的一些或全部可在专用计算机上执行、或使用专用硬件(诸如协处理器或现场可编程门阵列(FPGA)或专用的专用集成电路(ASIC))来执行。该处理可以以分布式方式来实施，其中由软件指定的计算的不同部分由不同的计算元件执行。每一个这种计算机程序优选地存储在或下载到可由通用或专用可编程计算机访问的存储设备的计算机可读存储介质(例如，固态存储器或介质、或磁性介质或光学介质)上，以便当由计算机读取存储设备介质以执行本文中描述的处理时，对计算机进行配置和操作。也可认为本发明的系统被实施为配置有计算机程序的有形、非暂态介质，其中，如此配置的介质致使计算机以指定的且预定义的方式操作以便执行本文中描述的处理步骤中的一项或多项。

[0195] 已经描述了本发明的多个实施例。然而，应当理解，前述描述旨在说明而非限制本发明的范围，本发明的范围由所附权利要求书的范围限定。因此，其他实施例也在所附权利要求书的范围内。例如，在不背离本发明的范围的情况下，可进行各种修改。另外，上述步骤中的一些可以是顺序独立的，并且因此可以以与所描述的顺序不同的顺序来执行。

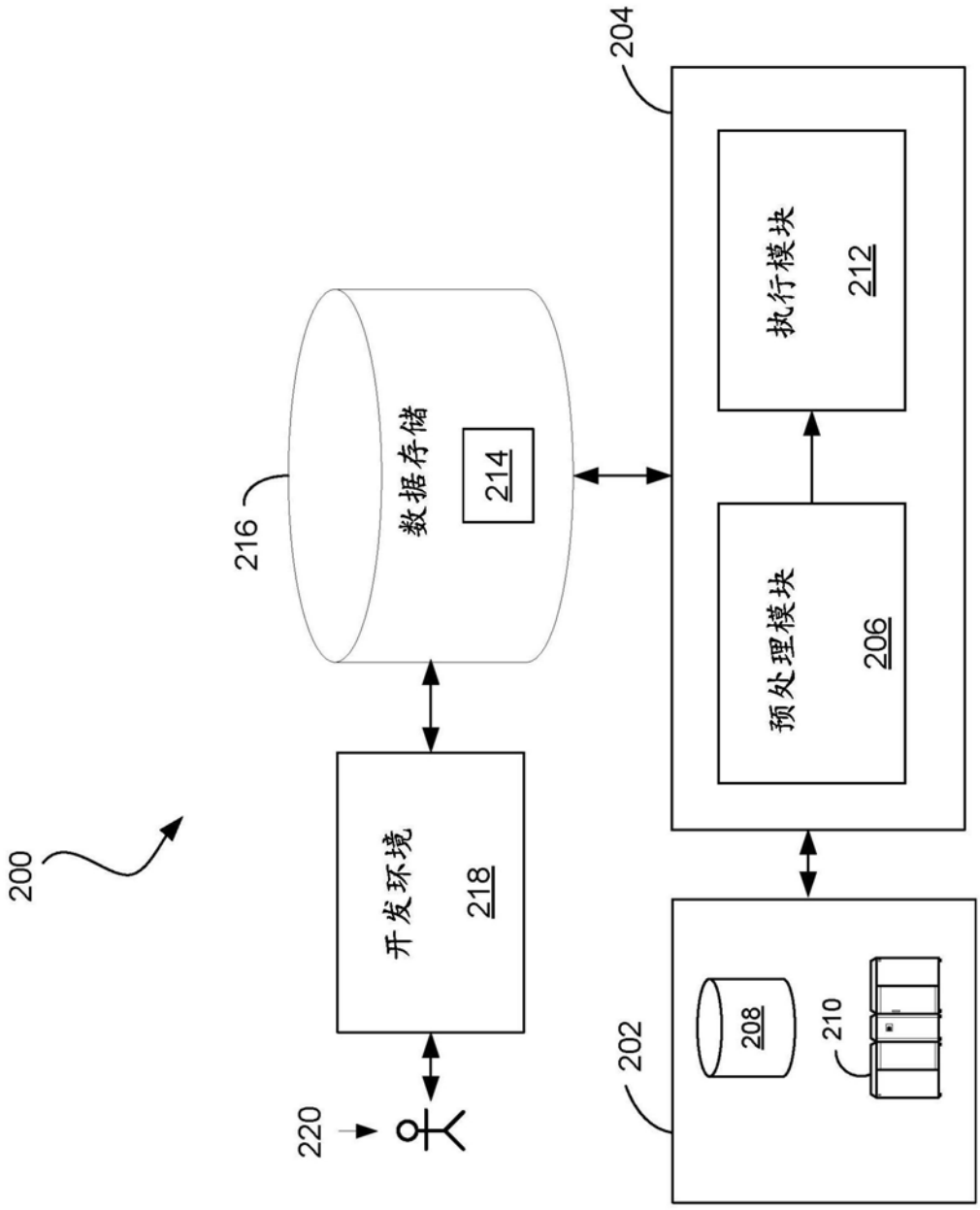

200
开发环境
218
220
216
数据存储
214
204
执行模块
212
预处理模块
206
202
208
210

图1

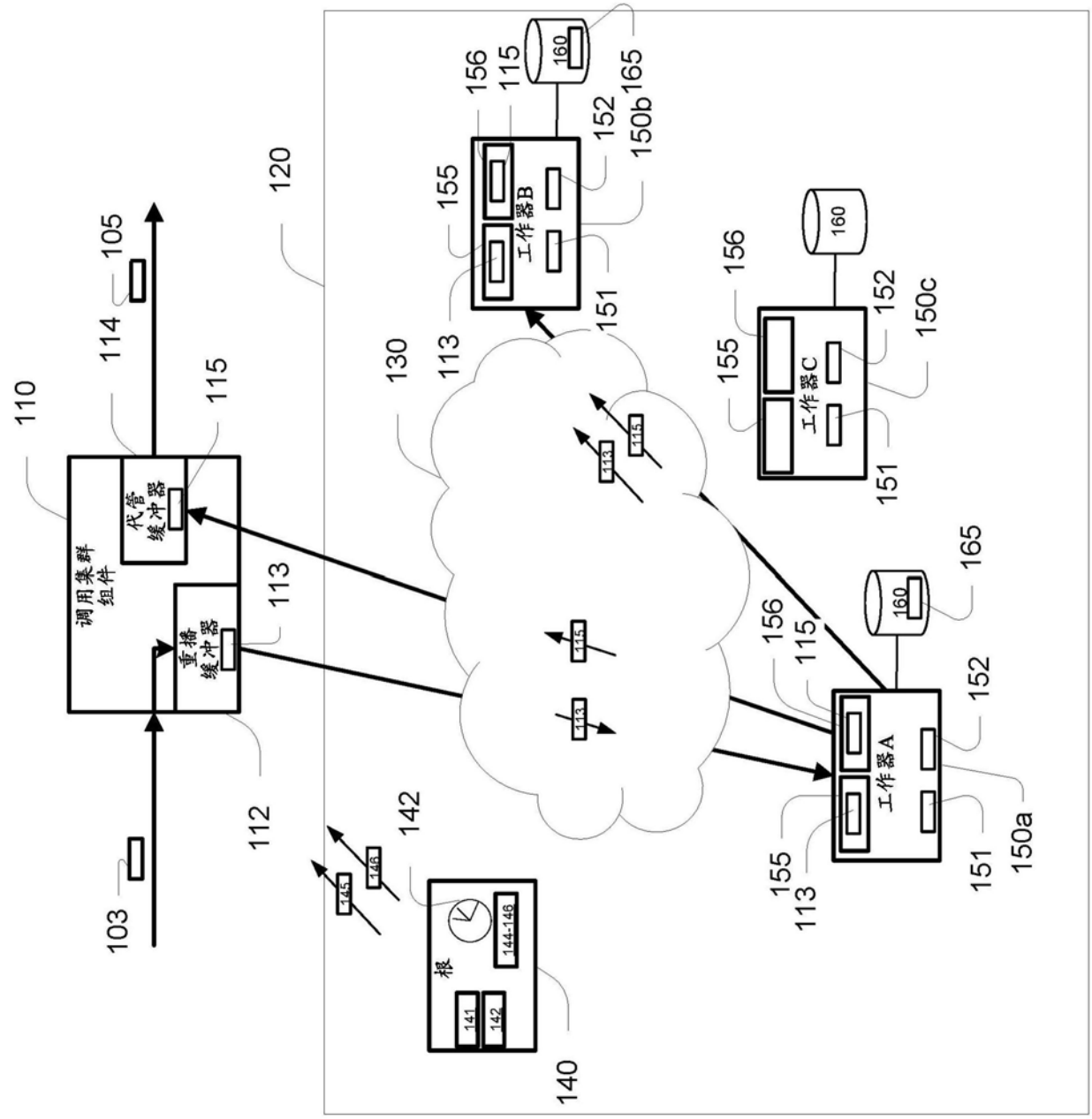

120
110
调用集群组件
代管缓冲器
重播缓冲器
105
114
115
113
112
103
130
142
145
146
根
141
142
144-146
140
工作器B
工作器C
工作器A
150b
150c
150a
151
152
155
156
160
165

图2

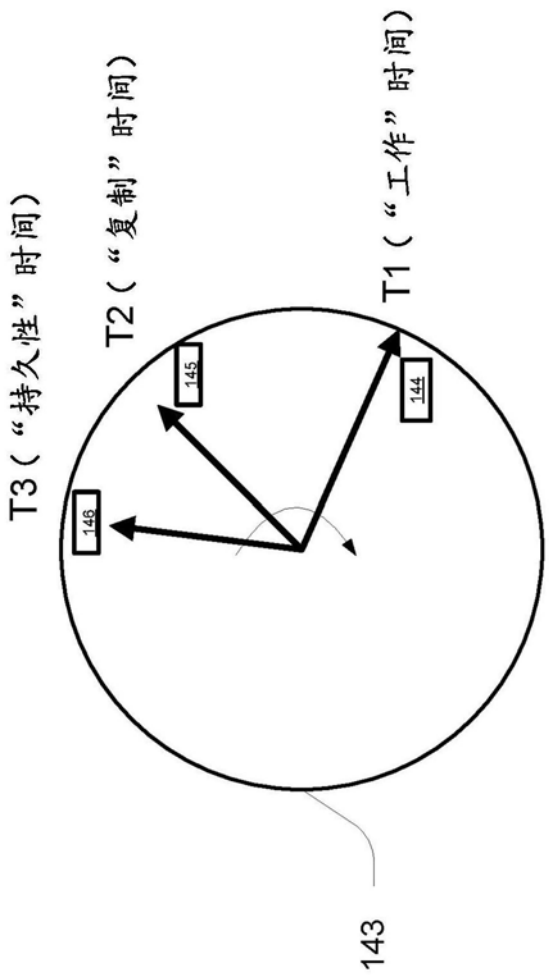


图3

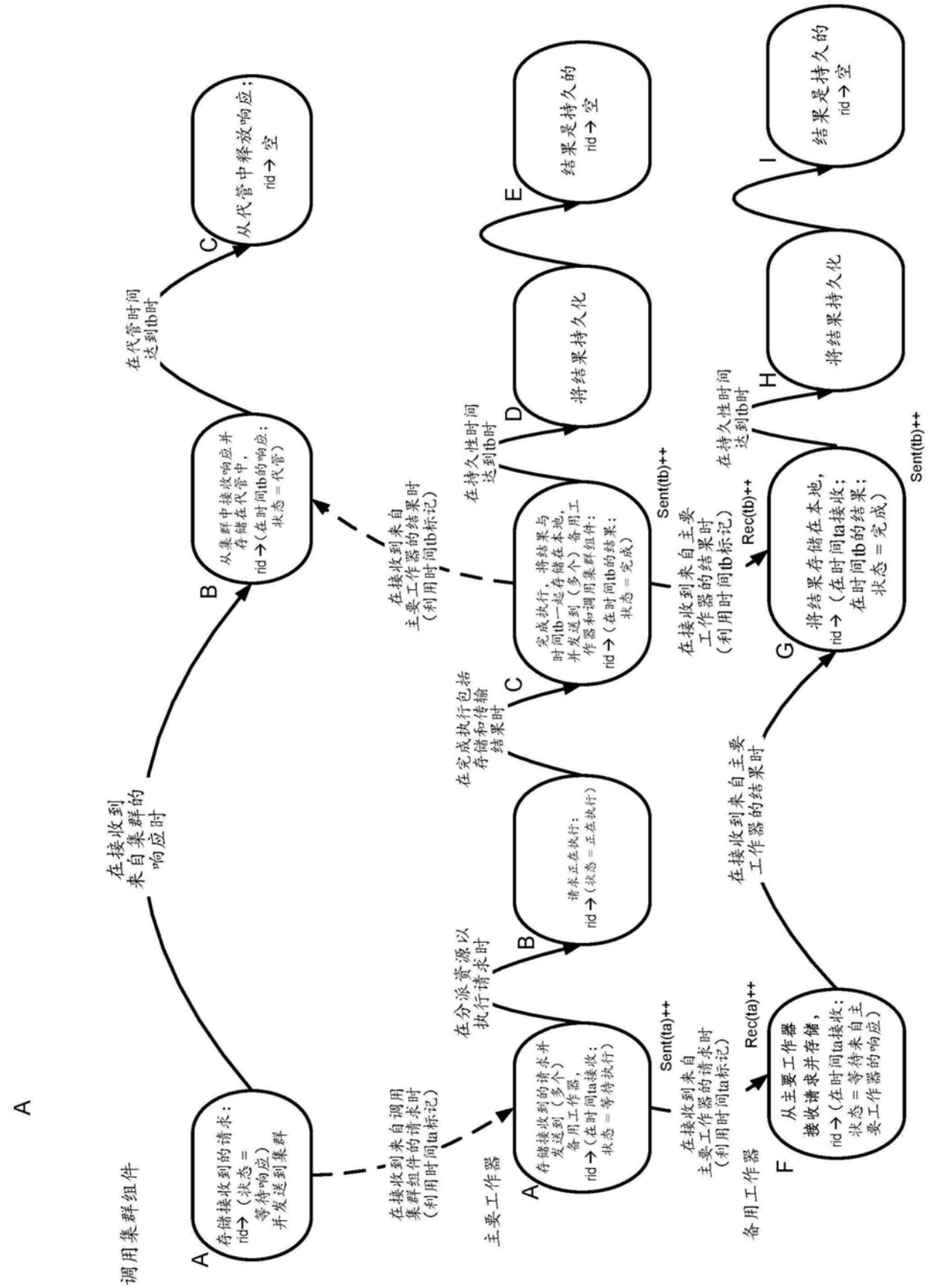

A
调用集群组件
存储接收到的请求：
rid→（状态=
等待响应）
并发送到集群
在接收到
来自集群的
响应时
在接收到来自调用
集群组件的请求时
（利用时间ta标记）
主要工作器
存储接收到的请求并
发送到（多个）
备用工作器，
rid→（在时间ta接收；
状态=等待执行）
Sent(ta)++
在接收到来自
主要工作器的请求时
（利用时间ta标记）
备用工作器
Rec(ta)++
从主要工作器
接收请求并存储，
rid→（在时间ta接收；
状态=等待来自主
要工作器的响应）
F
在分派资源以
执行请求时
B
请求正在执行：
rid→（状态=正在执行）
在完成执行包括
存储和传输
结果时
C
完成执行，将结果与
时间tb一起存储在本地，
并发送到（多个）备用工
作器和调用集群组件：
rid→（在时间tb的结果；
状态=完成）
Sent(tb)++
在接收到来自主要
工作器的结果时
（利用时间tb标记）
Rec(tb)++
在接收到来自主要
工作器的结果时
G
将结果存储在本地，
rid→（在时间ta接收；
在时间tb的结果；
状态=完成）
在接收到
来自集群的
响应时
在代管时间
达到tb时
C
从代管中释放响应：
rid→空
B
从集群中接收响应并
存储在代管中，
rid→（在时间tb的响应；
状态=代管）
在接收到来自
主要工作器的结果时
（利用时间tb标记）
在持久性时间
达到tb时
D
将结果持久化
E
结果是持久的
rid→空
在持久性时间
达到tb时
H
将结果持久化
I
结果是持久的
rid→空

图4

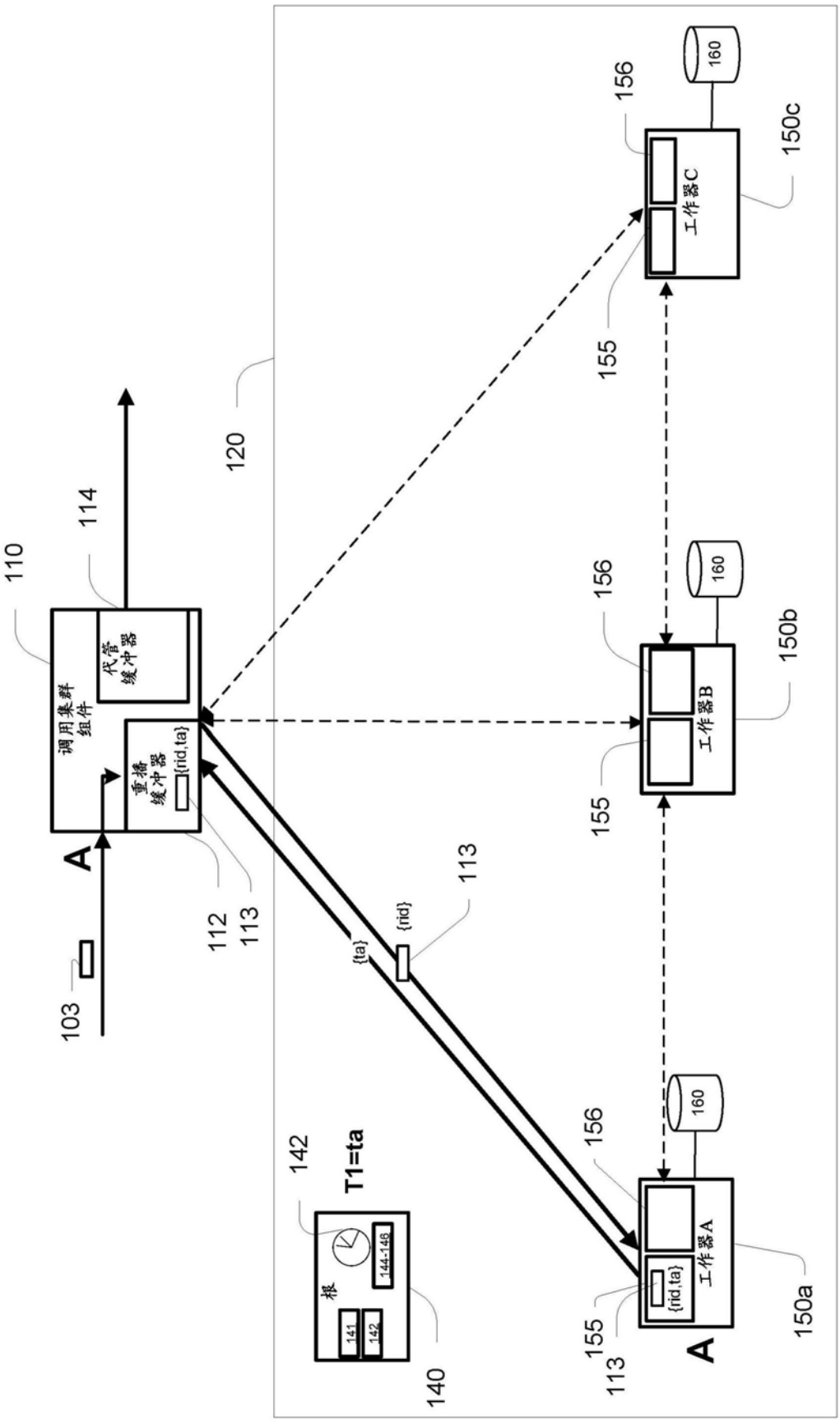

120
110
调用集群组件
代管缓冲器
114
重播缓冲器
{rid,ta}
A
103
112
113
{ta}
{rid}
113
工作器C
150c
156
160
155
工作器B
150b
156
160
155
工作器A
150a
156
160
155
{rid,ta}
113
A
142
T1=ta
根
141
142
144-146
140

图5

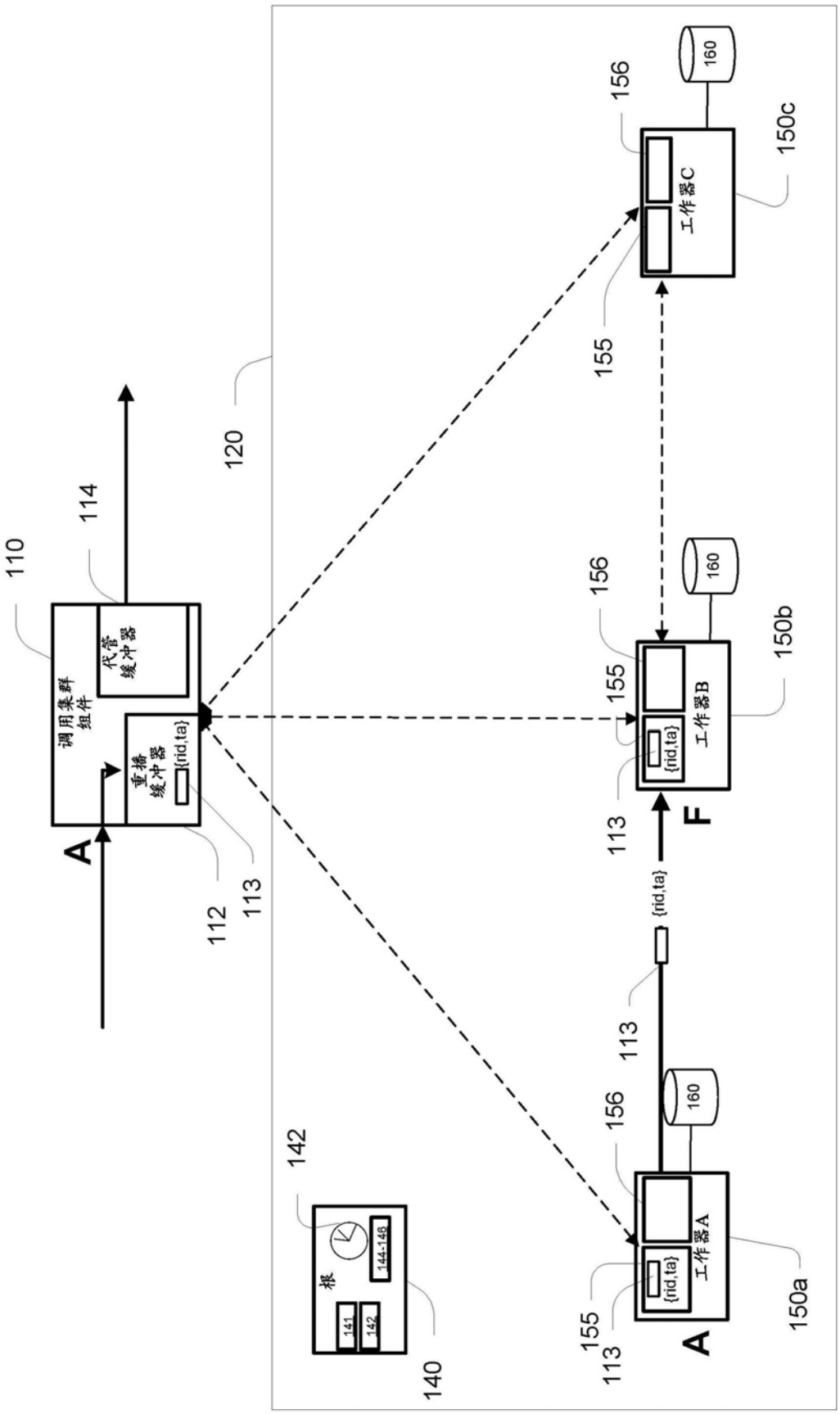

调用集群组件
代管缓冲器
重播缓冲器
{rid,ta}
工作器C
工作器B
工作器A
根
160
156
155
113
112
114
110
120
142
140
150a
150b
150c
141
143
144-146
A
F

图6

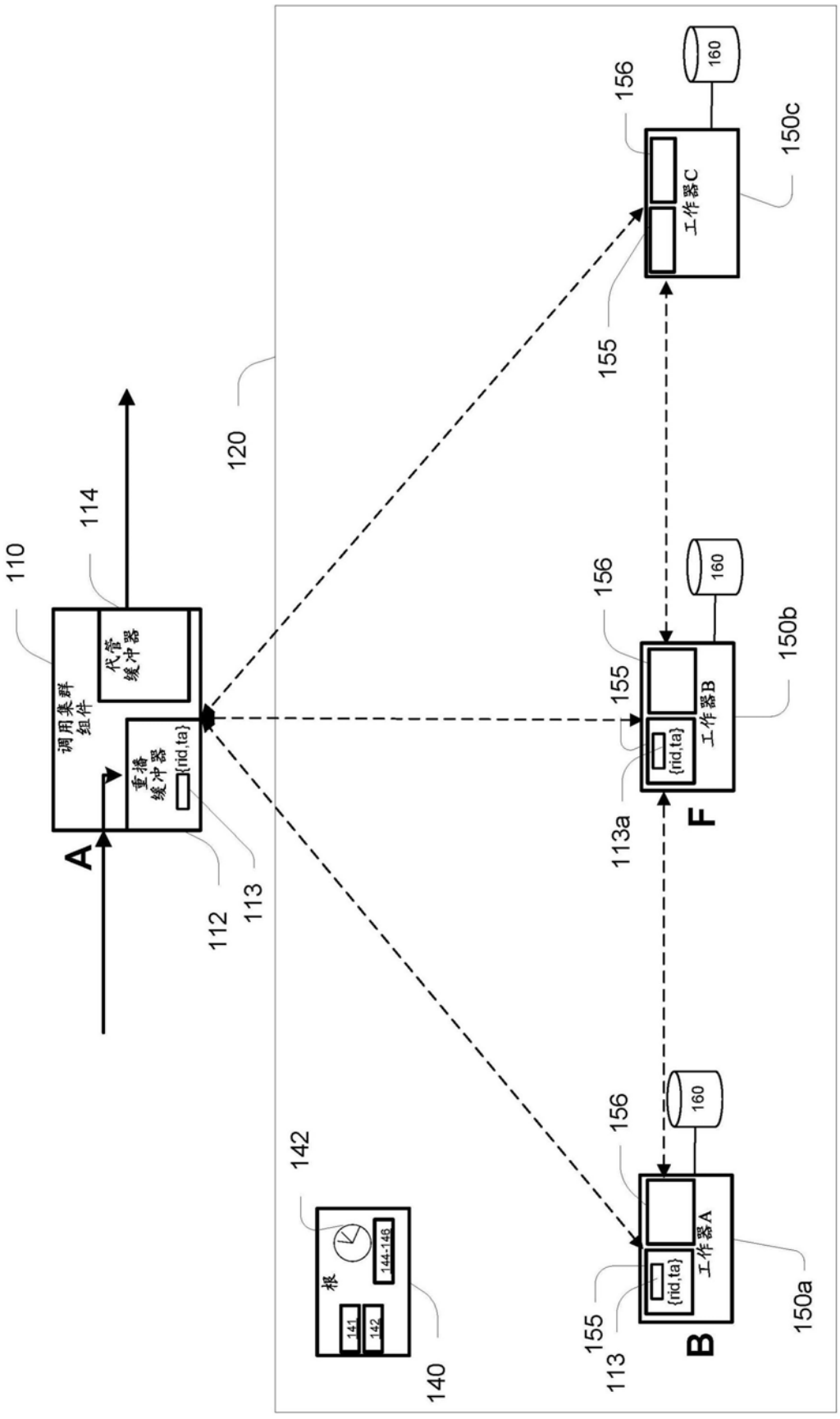

110
114
120
156
160
150c
工作器C
155
调用集群组件
代管缓冲器
重播缓冲器
{rid,ta}
A
112
113
156
160
150b
155
工作器B
{rid,ta}
113a
F
142
根
144-148
141
142
140
156
160
工作器A
{rid,ta}
155
113
B
150a

图7

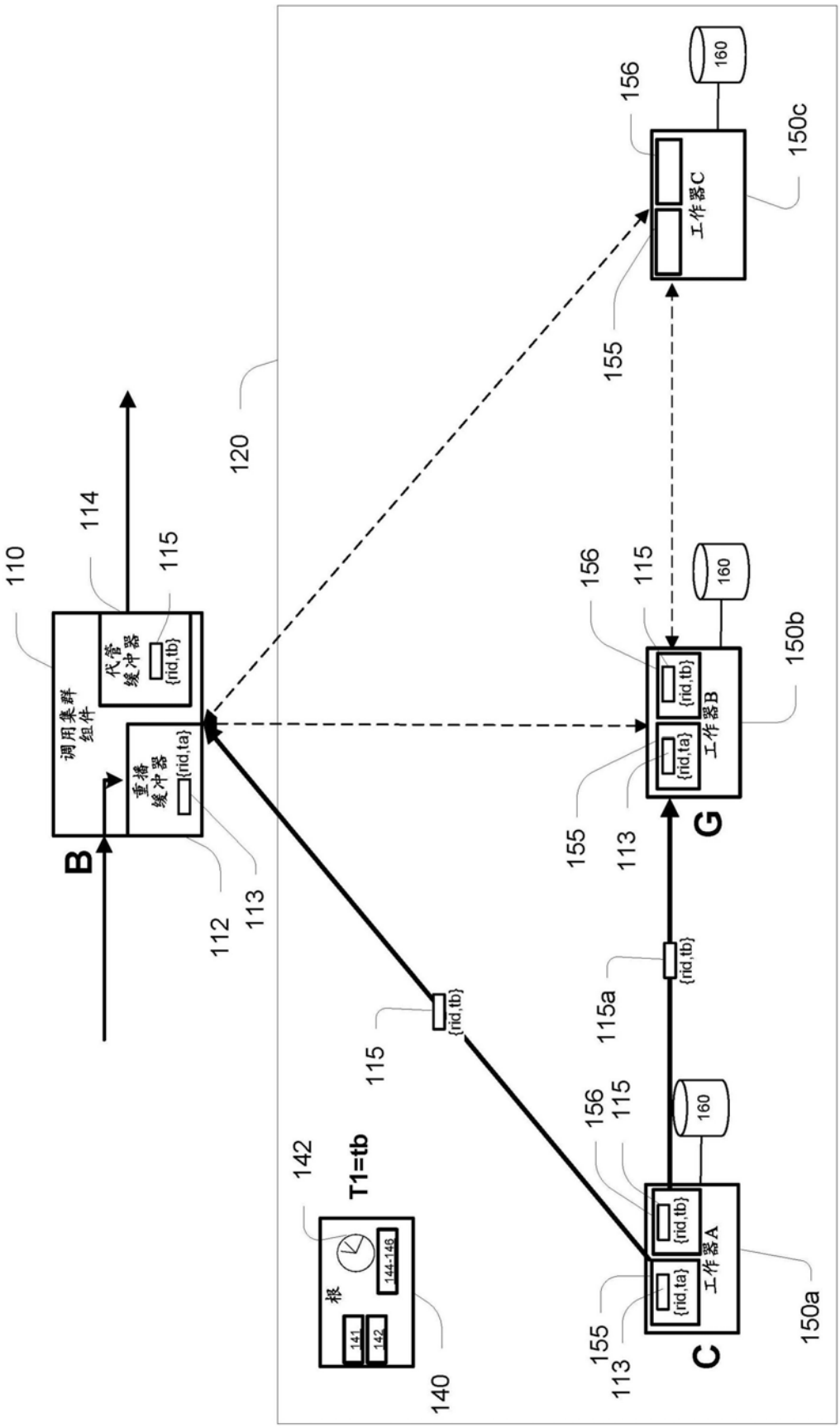

120
110
调用集群组件
代管缓冲器
{rid,tb}
重播缓冲器
{rid,ta}
B
114
115
112
113
{rid,tb}
115
142
T1=tb
根
141
142
144-146
140
工作器A
{rid,ta}
{rid,tb}
C
150a
155
113
156
115
160
115a
{rid,tb}
工作器B
{rid,ta}
{rid,tb}
G
150b
155
113
156
115
160
工作器C
150c
155
156
160

图8

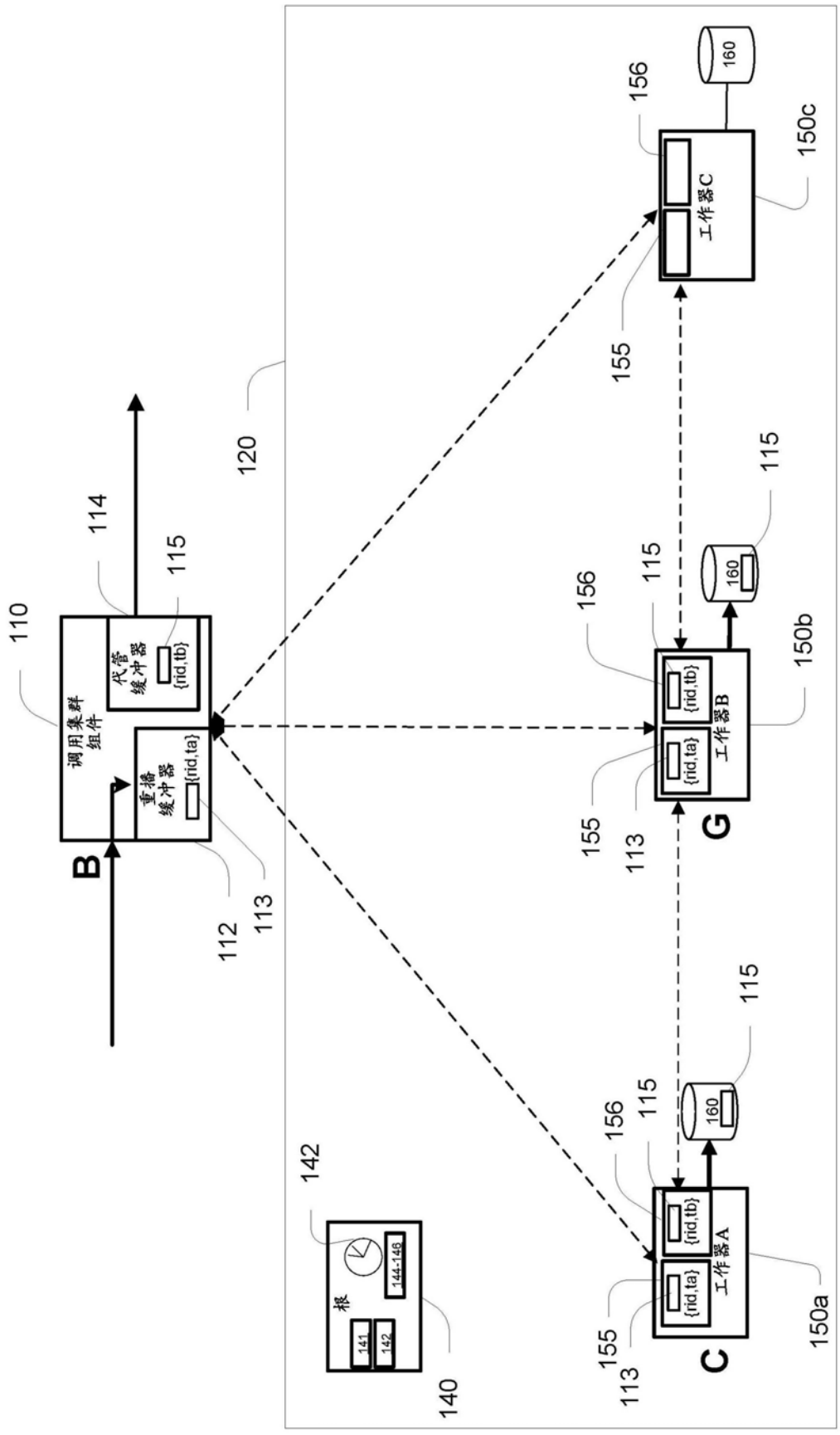

120
110
调用集群组件
代管缓冲器
{rid,tb}
重播缓冲器
{rid,ta}
114
115
112
113
B
工作器C
工作器B
工作器A
156
155
160
150c
150b
150a
G
C
142
根
141
142
144-148
140

图9

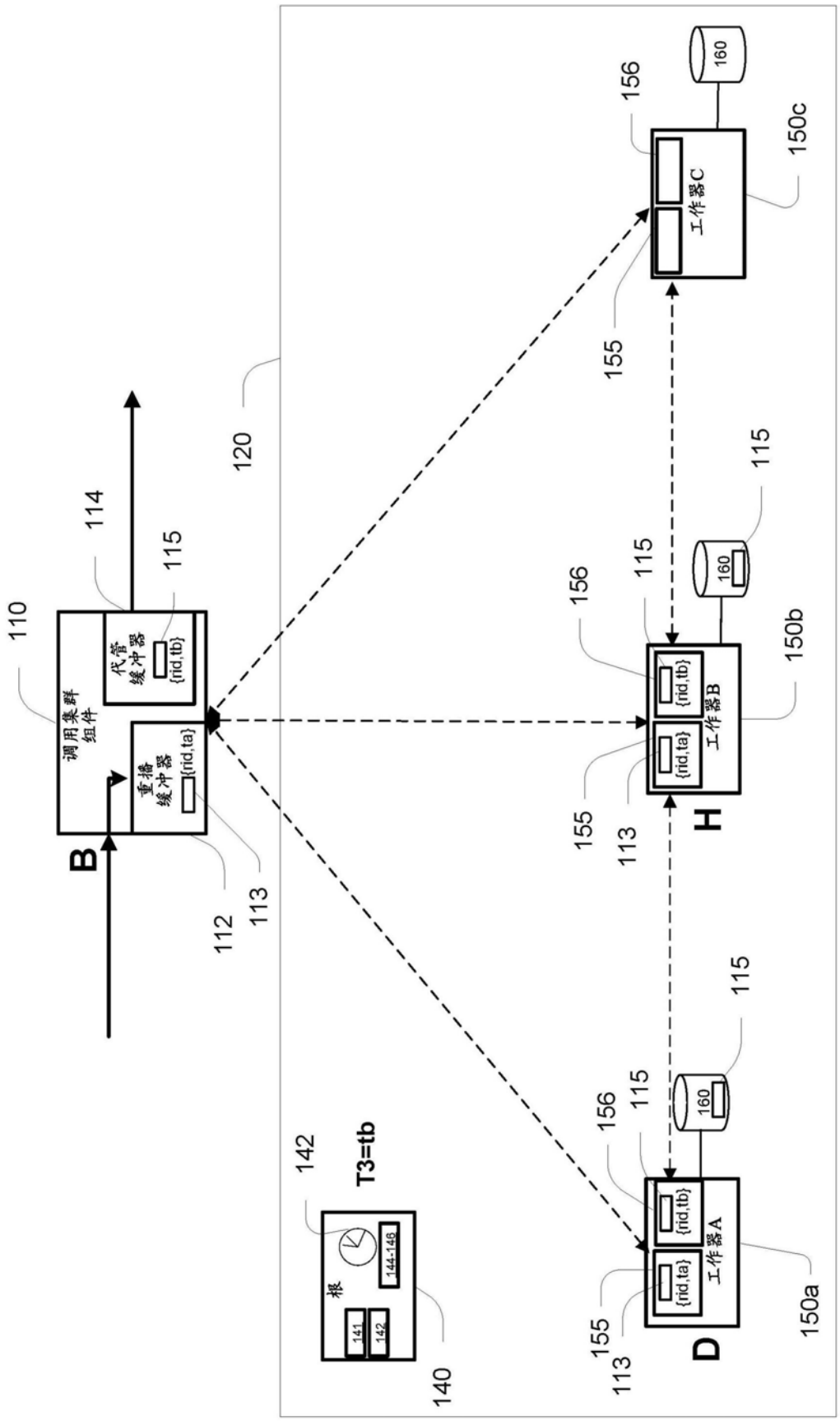

120
110
调用集群组件
代管缓冲器
{rid,tb}
重播缓冲器
{rid,ta}
114
115
112
113
B
工作器C
150c
工作器B
150b
工作器A
150a
156
155
160
H
D
T3=tb
142
根
141
142
144-146
140

图10

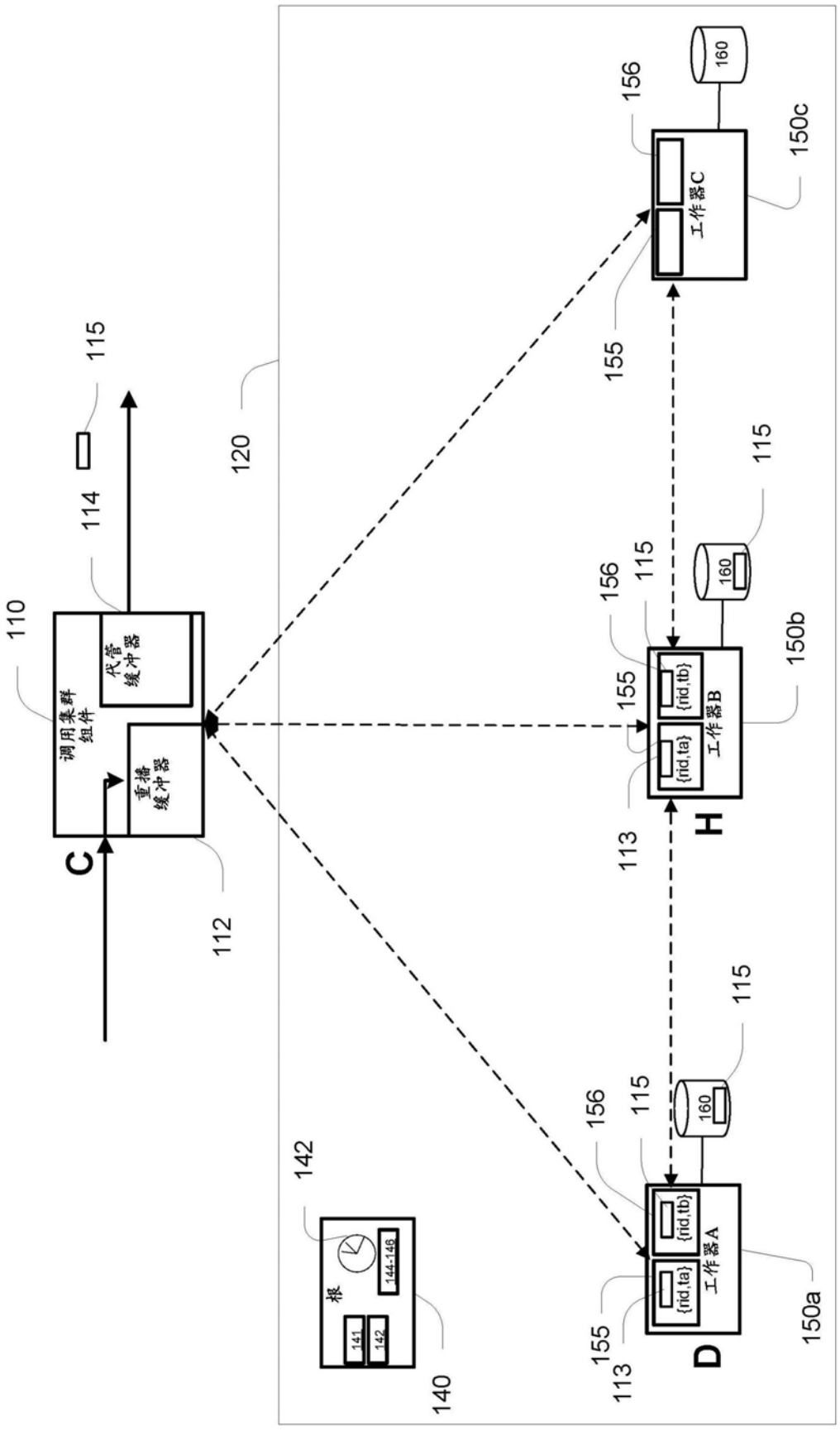

120
110
调用集群组件
代管缓冲器
重播缓冲器
C
112
114
115
工作器C
150c
工作器B
150b
工作器A
150a
{rid,ta}
{rid,tb}
160
155
156
113
H
D
140
根
141
142
144-148

图11

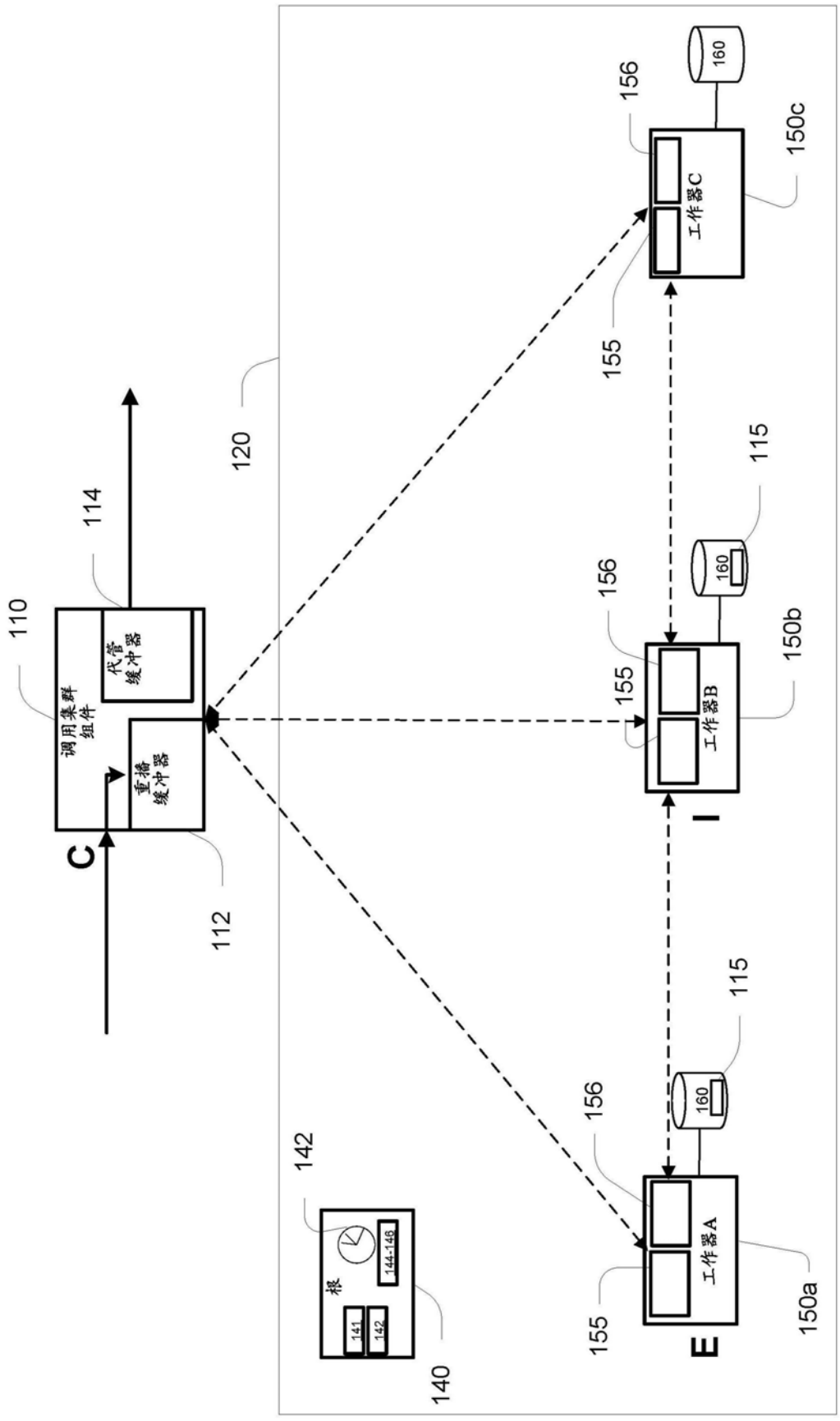

120
110
调用集群组件
代管缓冲器
重播缓冲器
114
112
C
156
160
工作器C
150c
155
115
工作器B
150b
I
工作器A
150a
E
142
根
144-146
141
142
140

图12

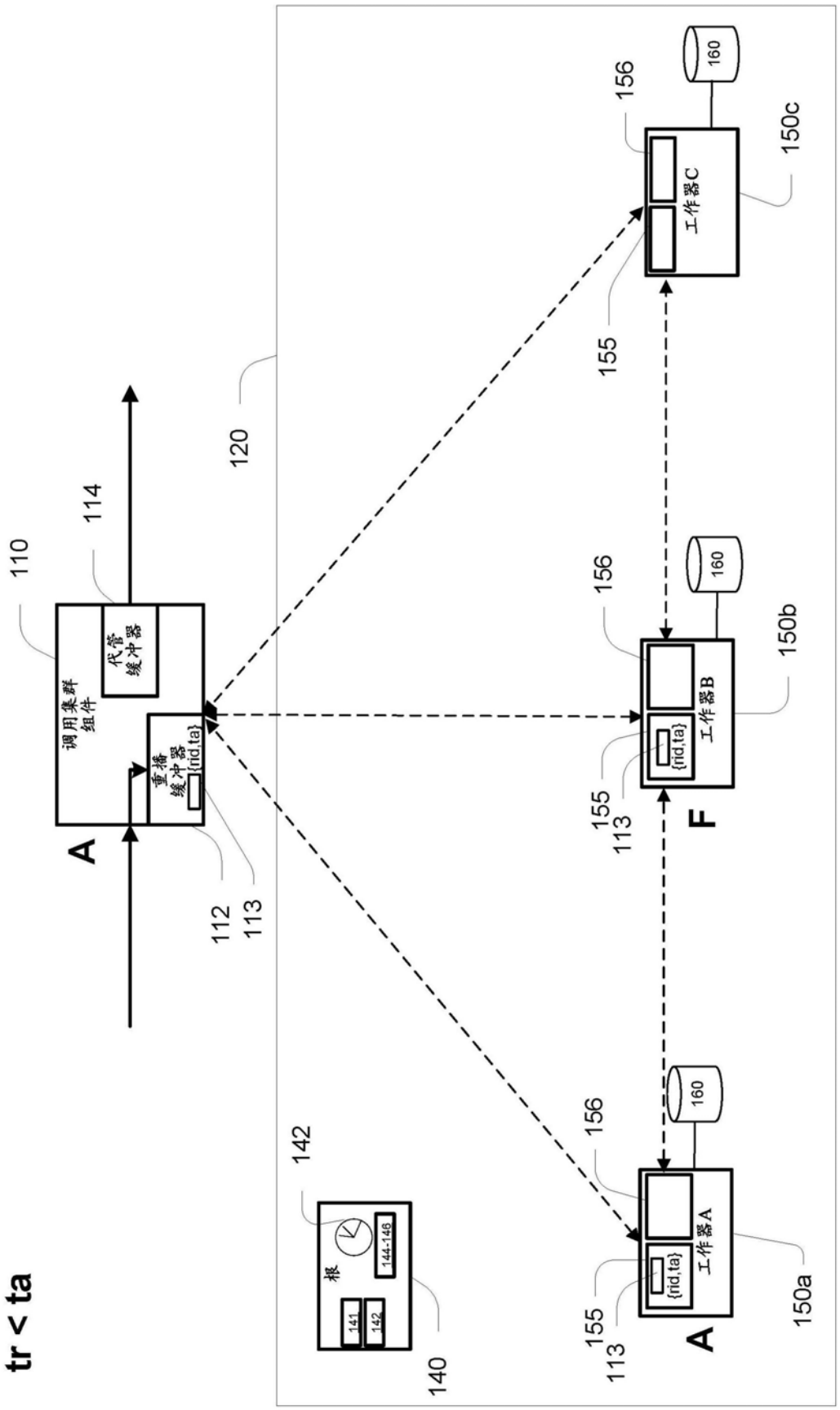

tr < ta
110
114
120
调用集群组件
代管缓冲器
重播缓冲器
{rid,ta}
A
112
113
142
根
144-148
141
142
140
156
160
150c
工作器C
155
150b
工作器B
F
150a
工作器A

图13

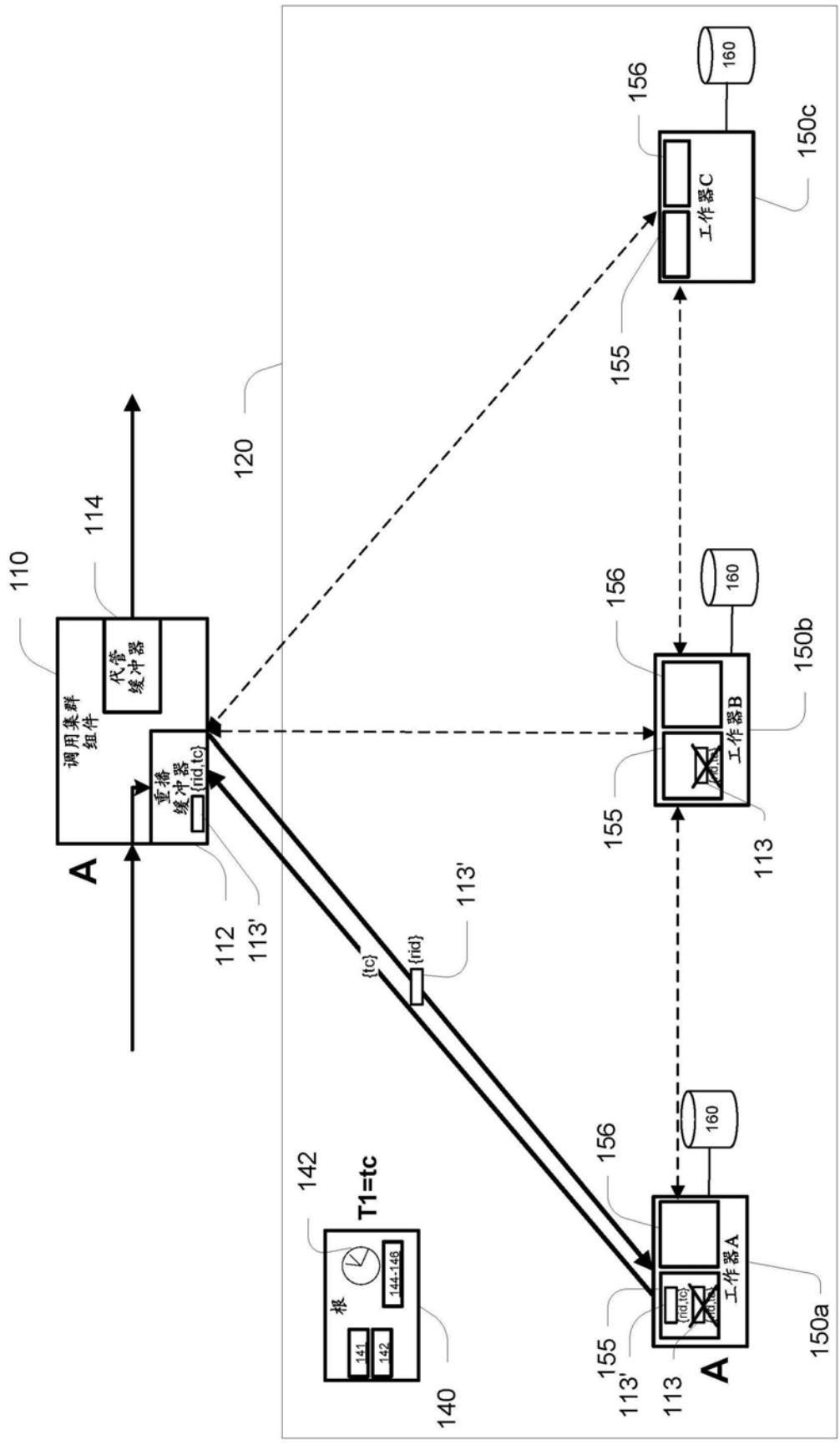

110
114
120
调用集群组件
代管缓冲器
重播缓冲器{rid,tc}
A
112
113'
{tc}
{rid}
113'
140
142
T1=tc
根
141
142
144-146
156
160
150c
工作器C
155
150b
工作器B
113
150a
工作器A
{rid,tc}
{tc}
155
113'
113
A

图14

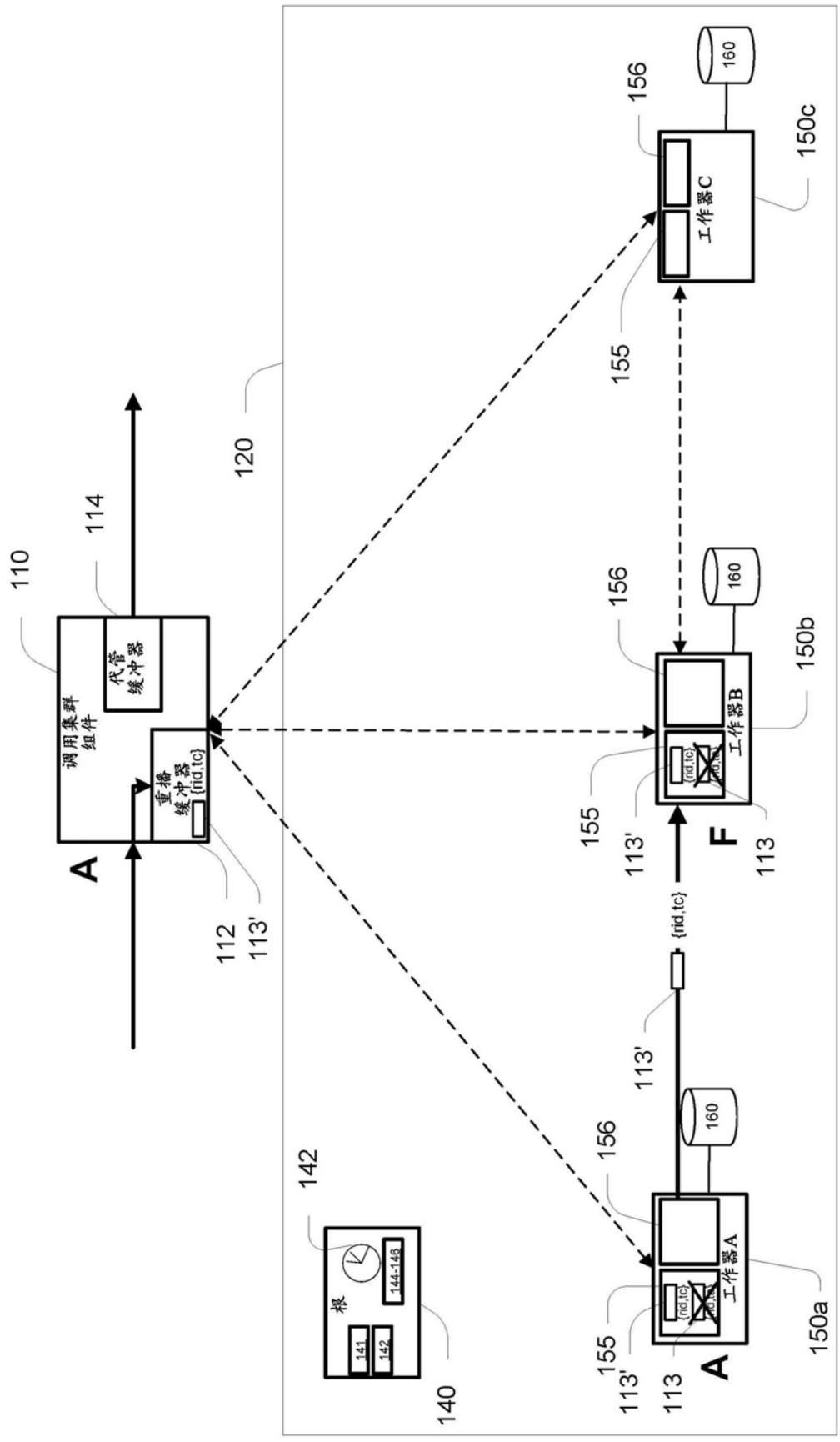

120
110
114
调用集群组件
代管缓冲器
重播缓冲器 {rid,tc}
A
112
113'
156
160
150c
工作器C
155
156
160
150b
工作器B
155
113'
F
113
{rid,tc}
113'
156
160
142
根
144-146
141
142
140
工作器A
155
113'
113
A
150a

图15

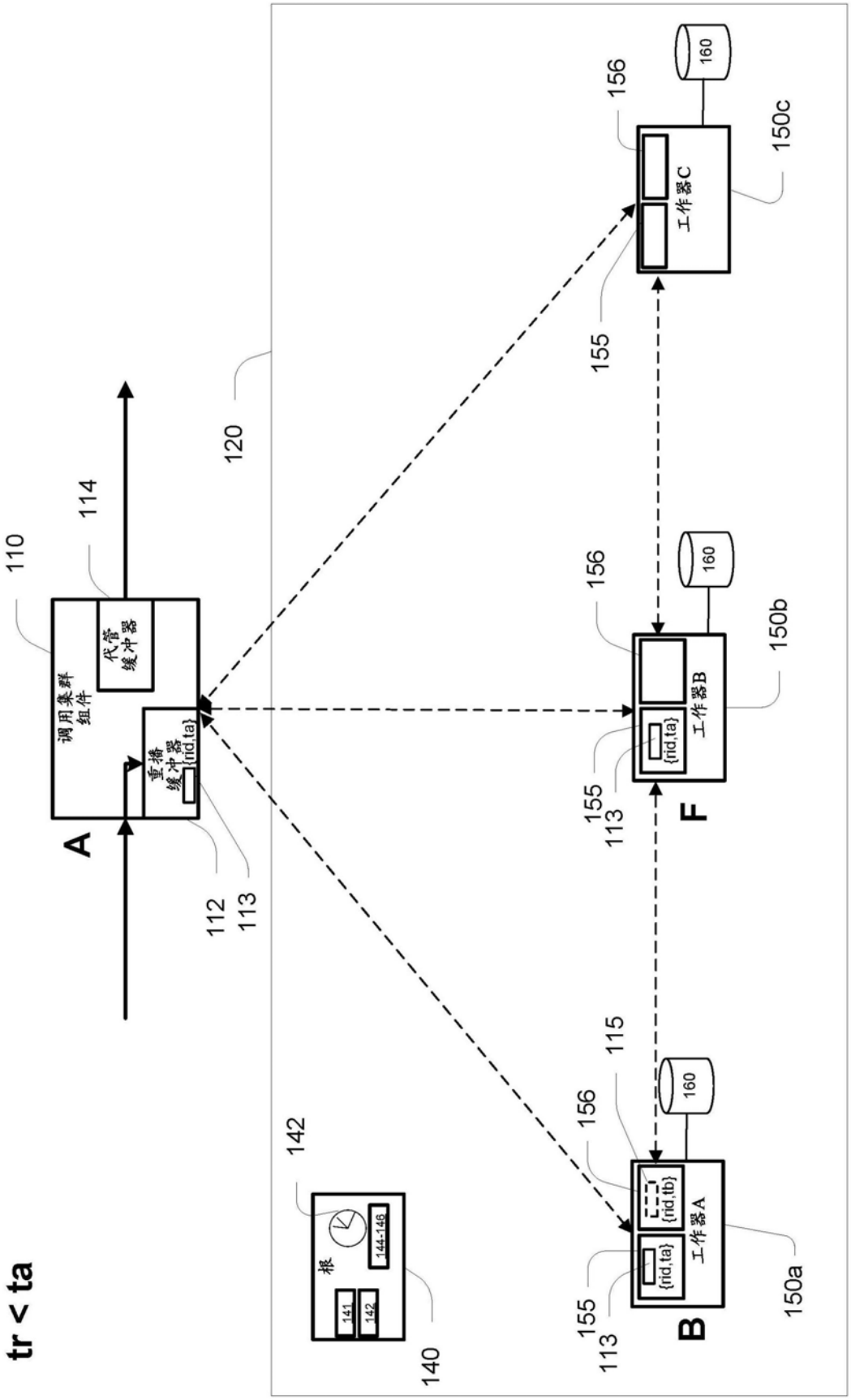

tr < ta
A
B
F
调用集群组件
代管缓冲器
重播缓冲器
{rid,ta}
工作器A
工作器B
工作器C
{rid,tb}
根
110
112
113
114
115
120
140
141
142
144-148
150a
150b
150c
155
156
160

图16

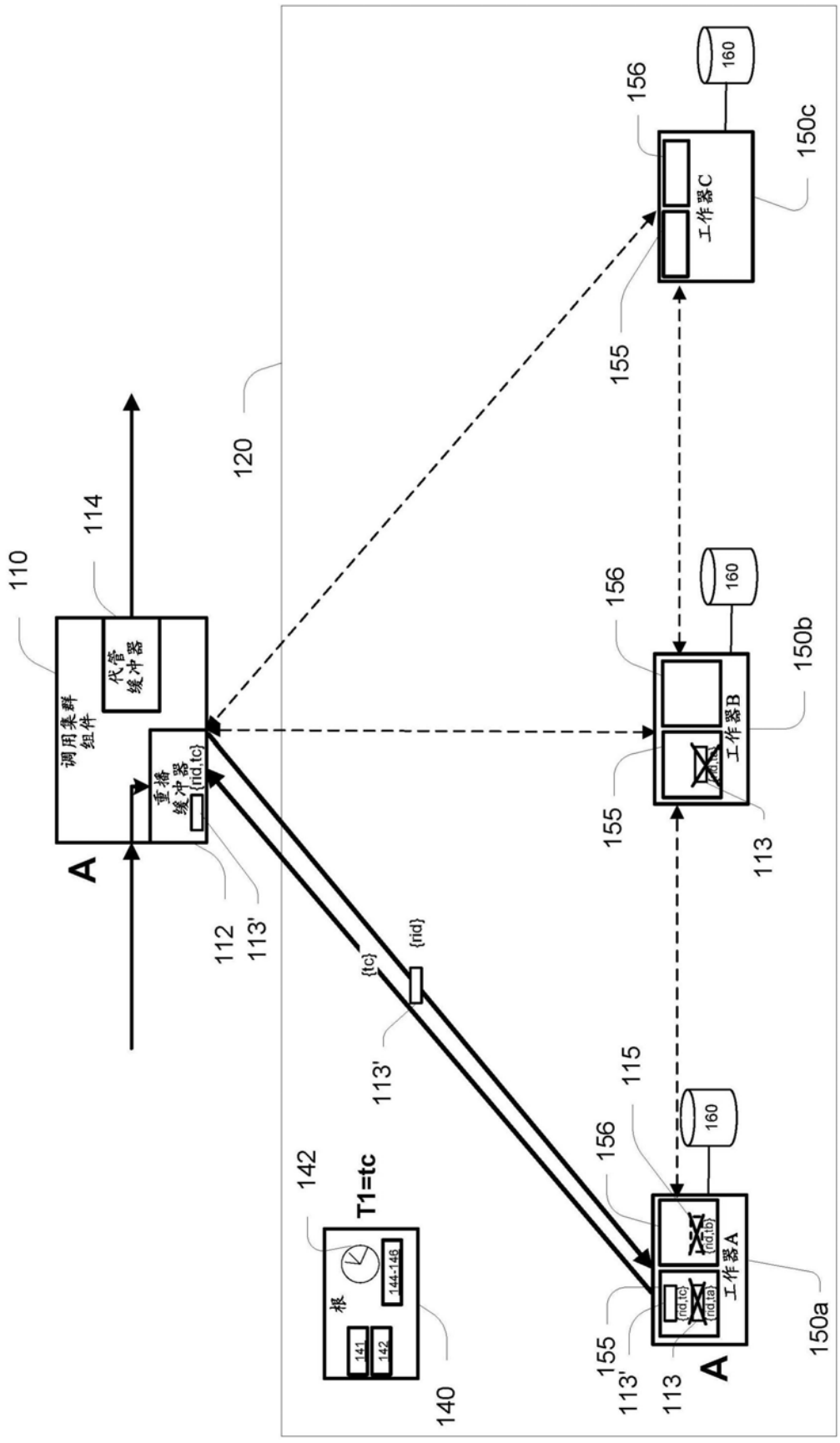

调用集群组件
代管缓冲器
重播缓冲器{rid,tc}
工作器C
工作器B
工作器A
根
T1=tc
{tc}
{rid}
A
A
110
114
120
112
113'
113
115
140
141
142
144-146
150a
150b
150c
155
156
160

图17

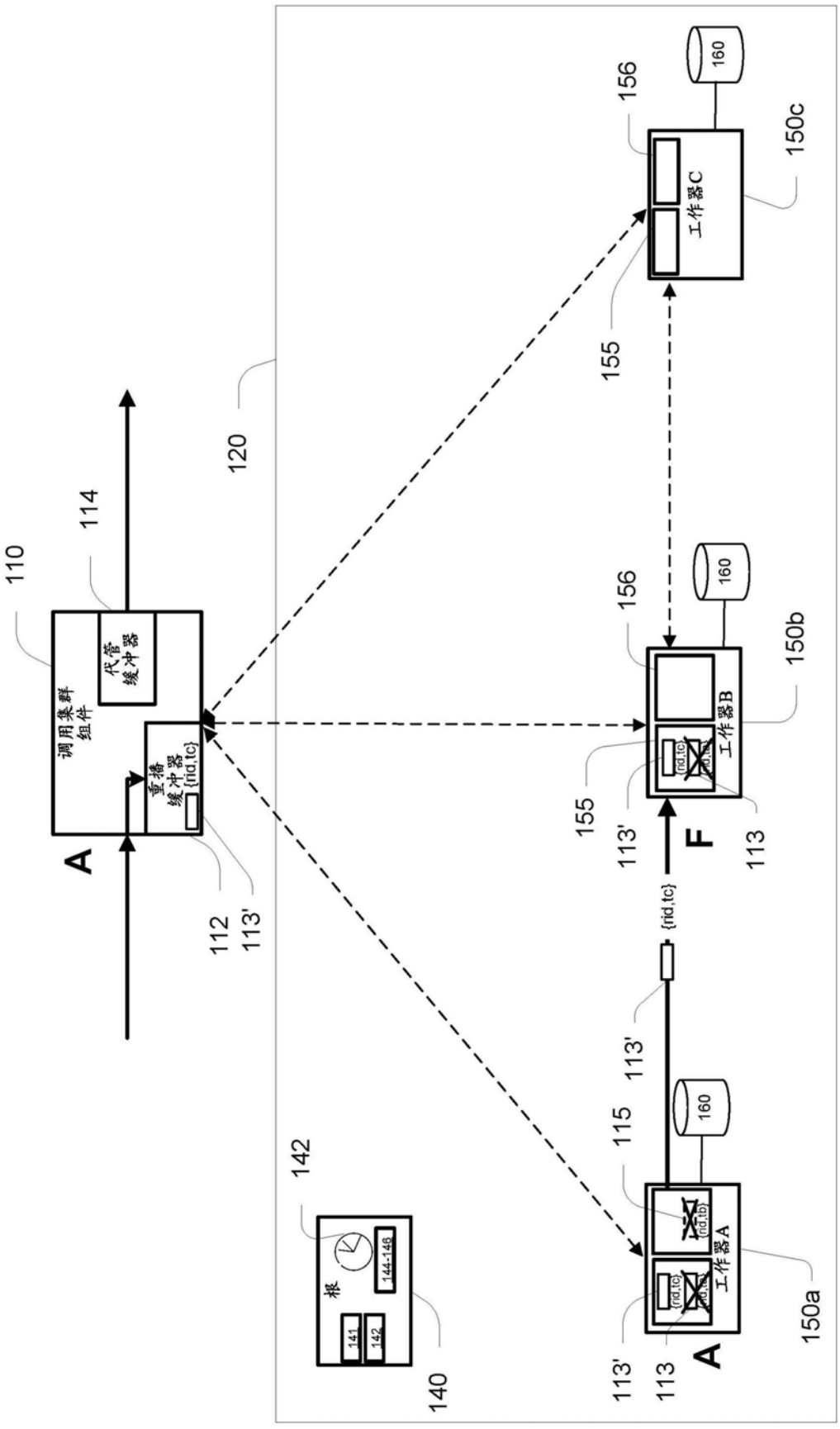

120
110
114
调用集群组件
代管缓冲器
重播缓冲器{rid,tc}
A
112
113'
156
160
150c
工作器C
155
156
160
150b
工作器B
155
113'
F
113
{rid,tc}
113'
115
160
142
根
144-146
141
142
140
150a
工作器A
113'
113
A

图18

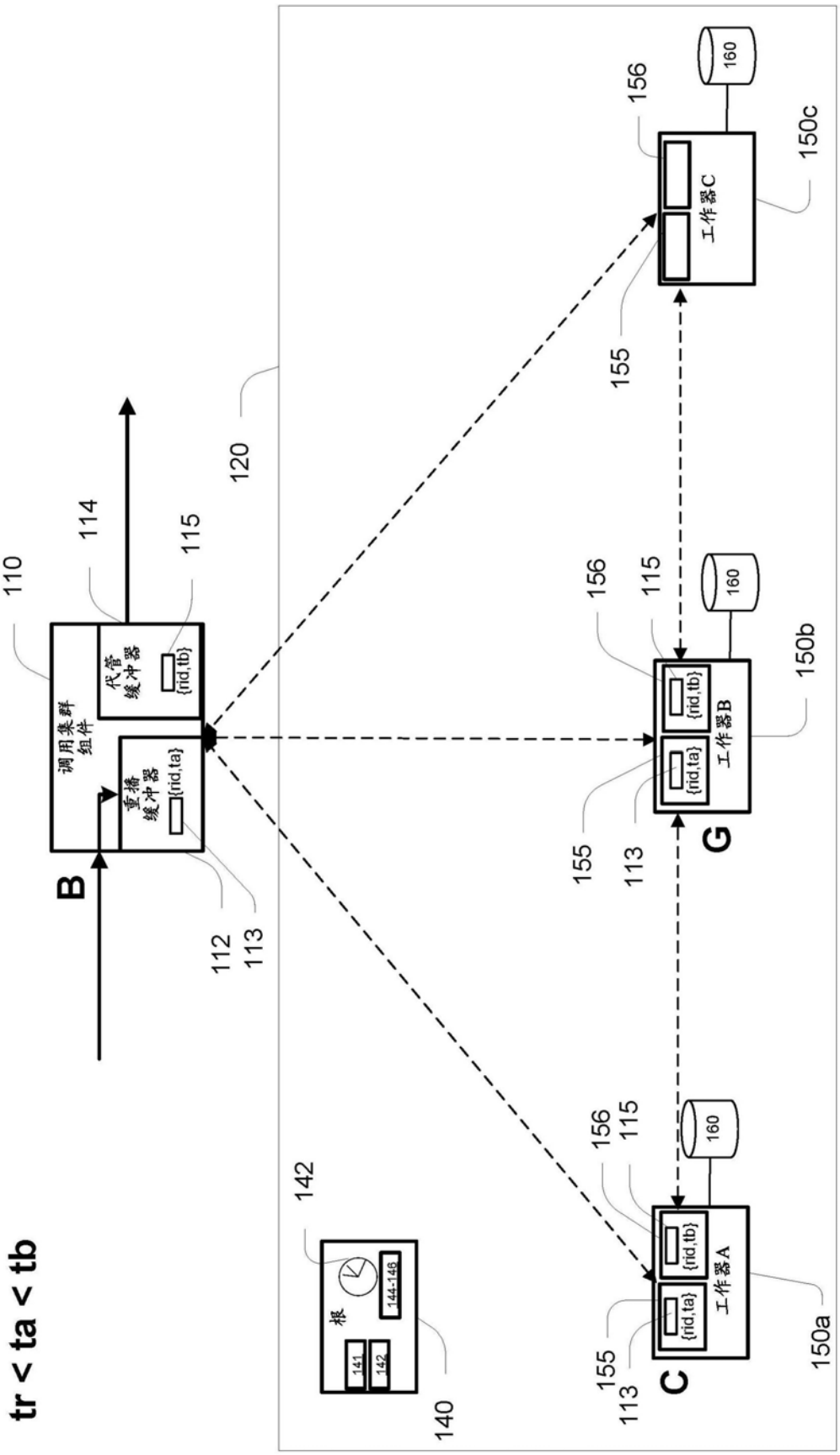

tr < ta < tb
B
G
C
调用集群组件
代管缓冲器
{rid,tb}
重播缓冲器
{rid,ta}
工作器A
工作器B
工作器C
根
141
142
144-146
110
112
113
114
115
120
140
142
150a
150b
150c
155
156
160

图19

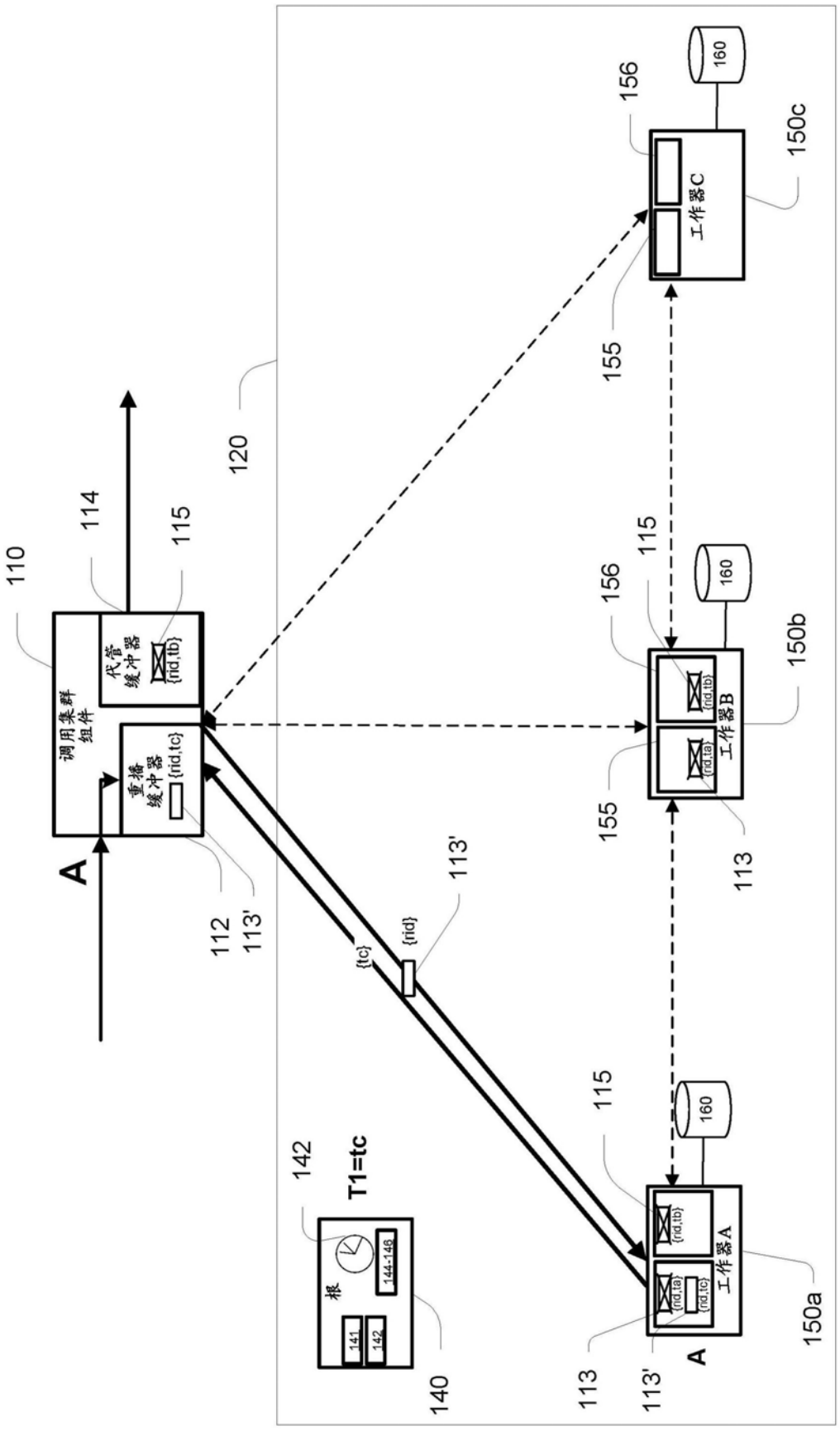

120
110
调用集群组件
代管缓冲器
{rid,tb}
重播缓冲器
{rid,tc}
114
115
112
113'
A
{tc}
{rid}
113'
142
T1=tc
根
144-146
141
142
140
113
113'
A
工作器A
{rid,ta}
{rid,tb}
{rid,tc}
150a
115
160
156
155
工作器B
150b
工作器C
150c

图20

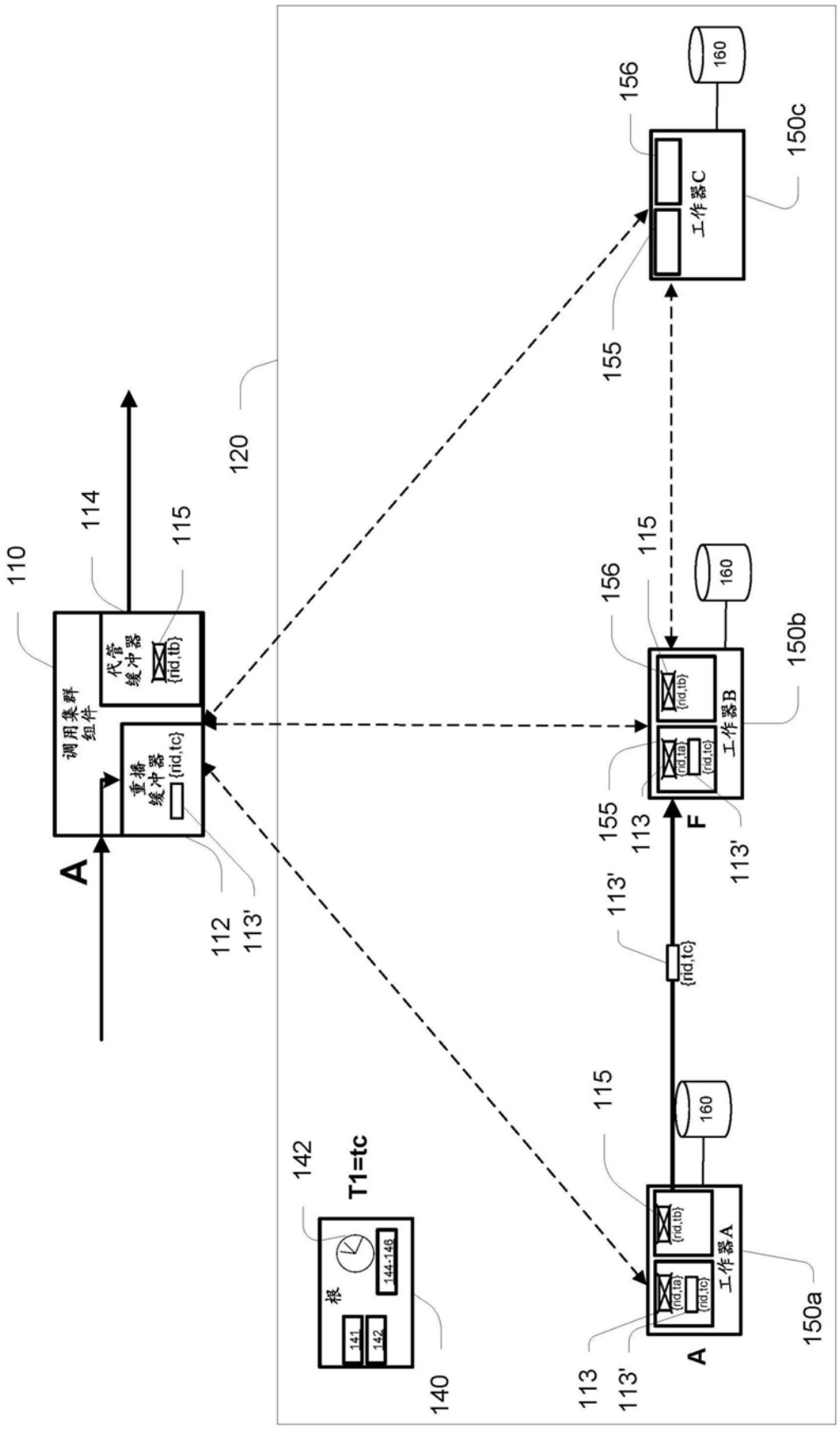

调用集群组件
代管缓冲器 {rid,tb}
重播缓冲器 {rid,tc}
110
114
115
120
112
113'
A
156
160
150c
工作器C
155
115
156
160
150b
工作器B
{rid,tb}
{rid,tc}
155
113
113'
F
113'
{rid,tc}
115
160
工作器A
150a
113
113'
A
142
T1=tc
根
141
142
144-146
140

图21

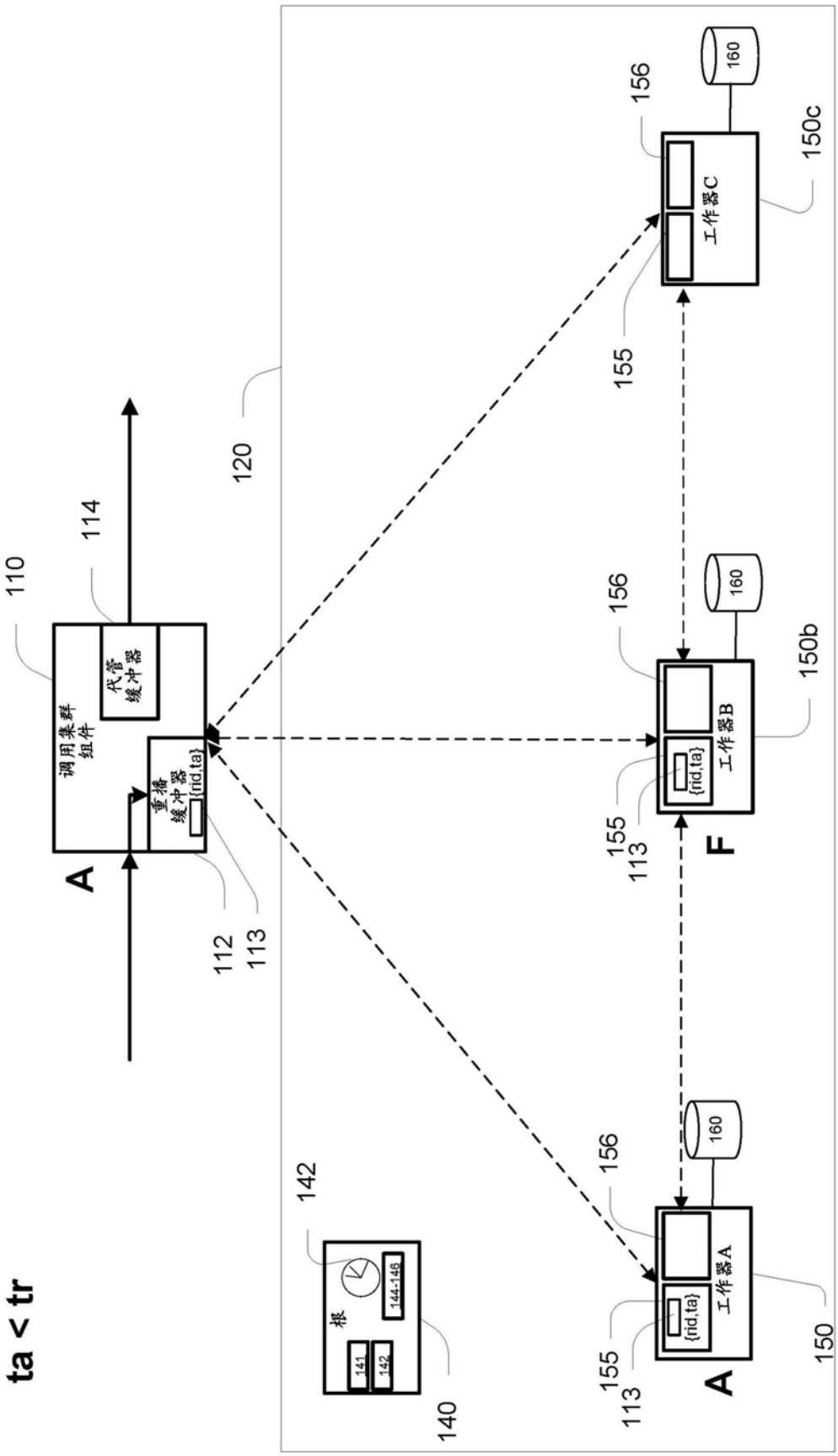

ta < tr
A
110
114
调用集群组件
代管缓冲器
重播缓冲器
{rid,ta}
112
113
120
156
160
150c
工作器C
155
156
160
150b
工作器B
{rid,ta}
155
113
F
156
160
工作器A
{rid,ta}
155
113
A
150
142
根
144-146
141
142
140

图22

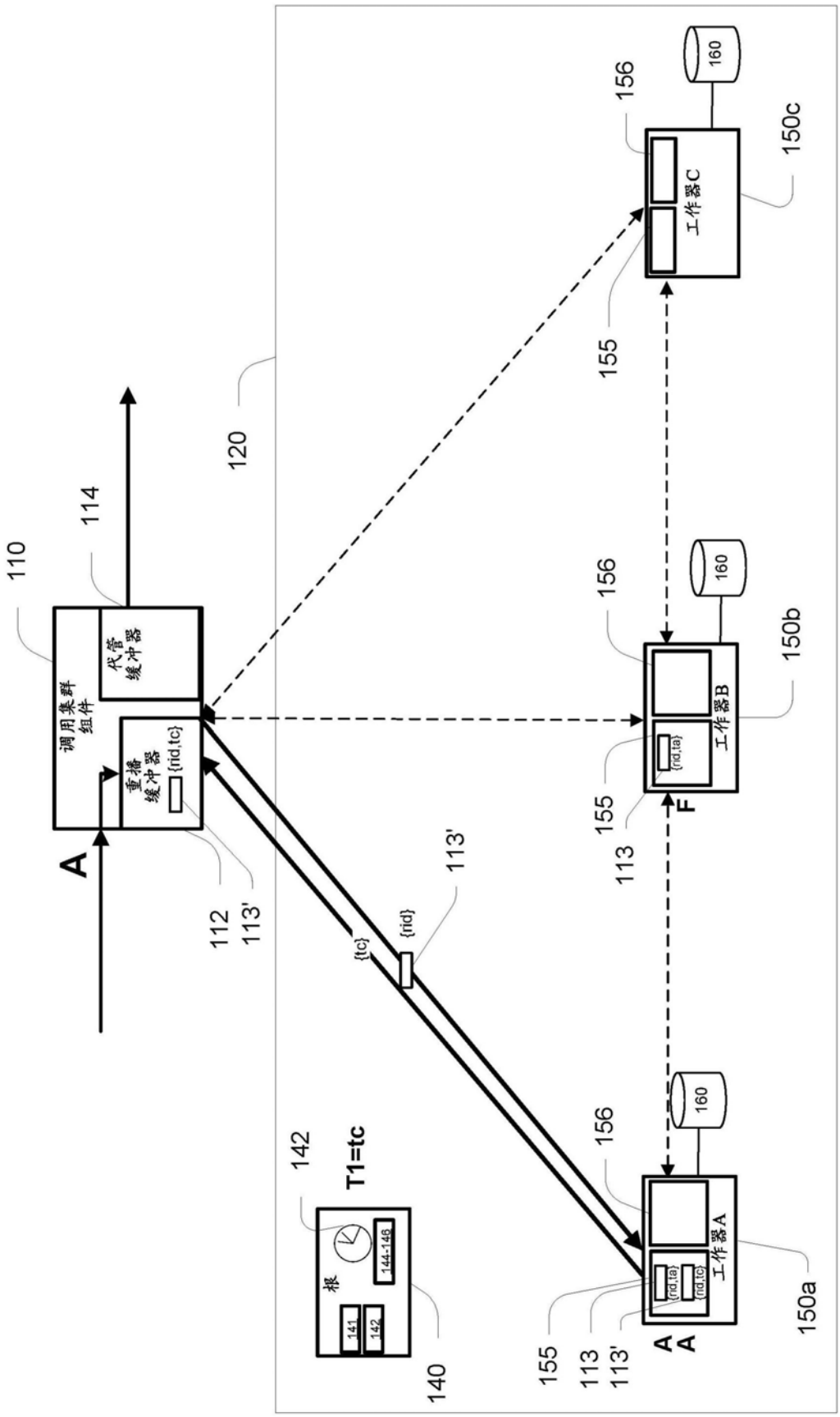

调用集群组件
代管缓冲器
重播缓冲器
{rid,tc}
110
114
120
112
113'
A
{tc}
{rid}
113'
T1=tc
142
根
141
142
144-146
140
工作器A
工作器B
工作器C
150a
150b
150c
156
160
155
113
113'
A
A
F
{rid,ta}
{rid,tc}

图23

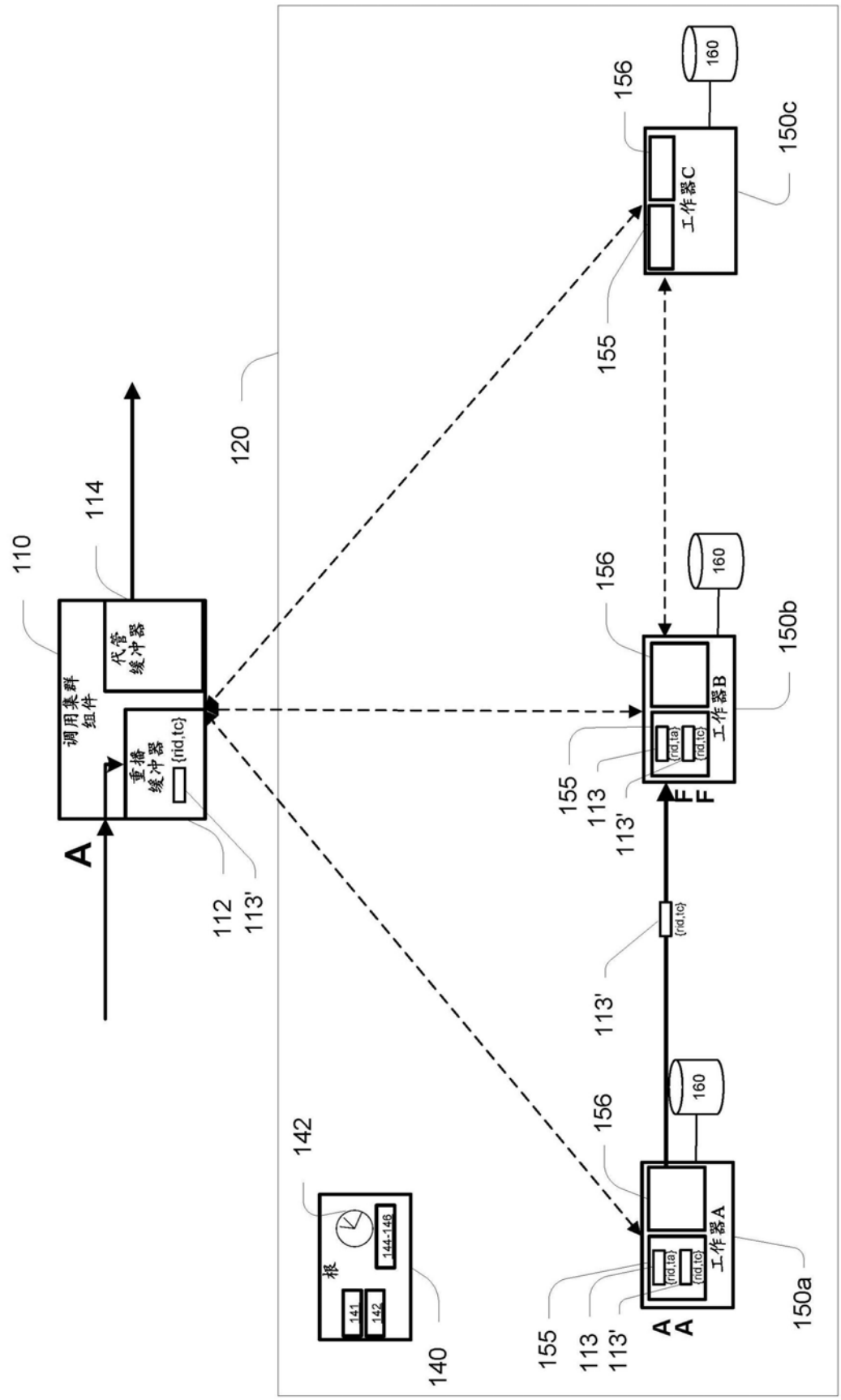

110
114
120
调用集群组件
代管缓冲器
重播缓冲器
{rid,tc}
A
112
113'
156
160
150c
工作器C
155
150b
工作器B
{rid,ta}
{rid,tc}
155
113
113'
F
F
{rid,tc}
113'
156
160
142
根
144-146
141
142
140
工作器A
150a
155
113
113'
A
A

图24

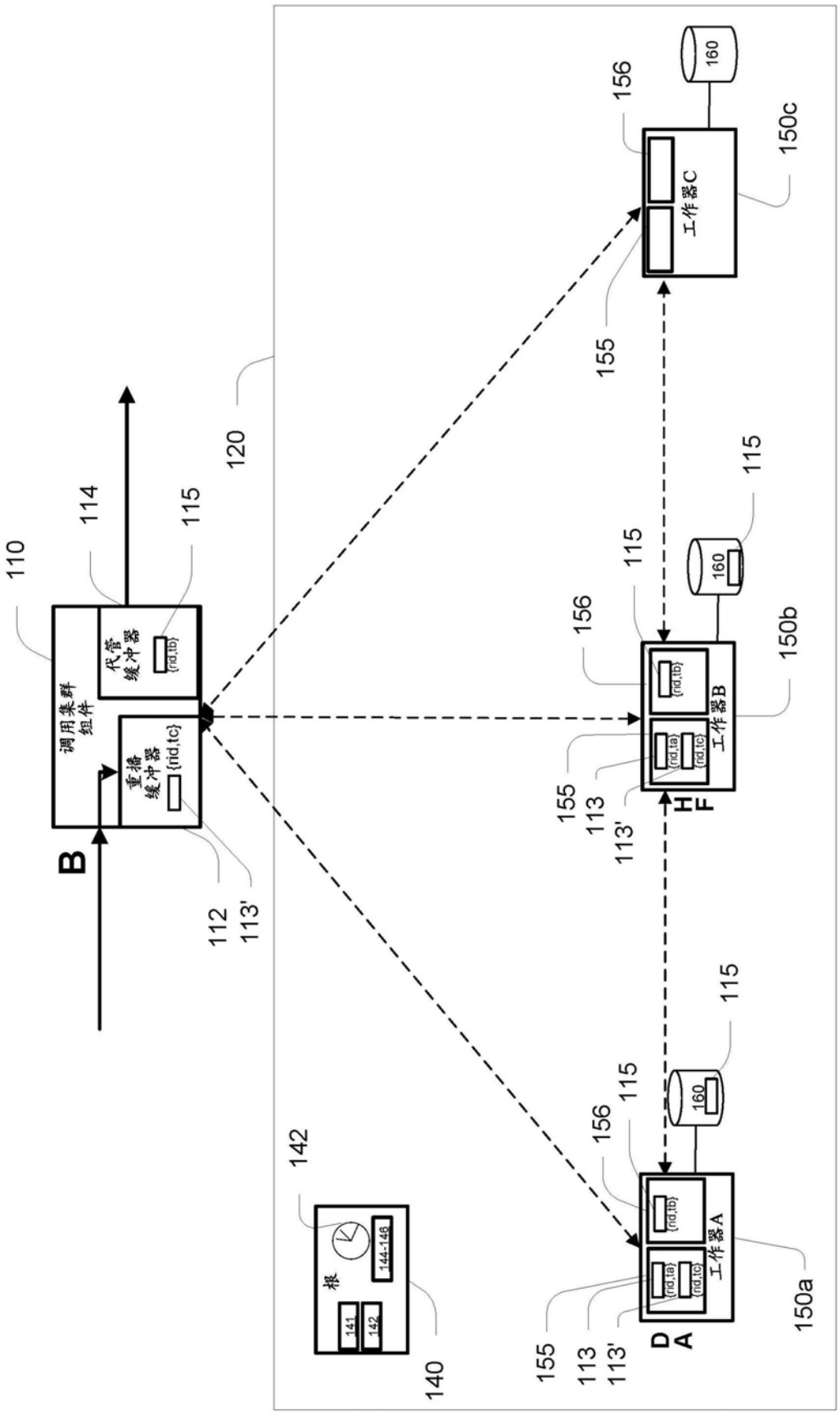

110
114
115
120
调用集群组件
代管缓冲器
{rid,tb}
重播缓冲器
{rid,tc}
B
112
113'
156
160
150c
工作器C
155
115
150b
工作器B
{rid,ta}
{rid,tc}
H
F
142
根
141
142
144-148
140
150a
工作器A
D
A

图25

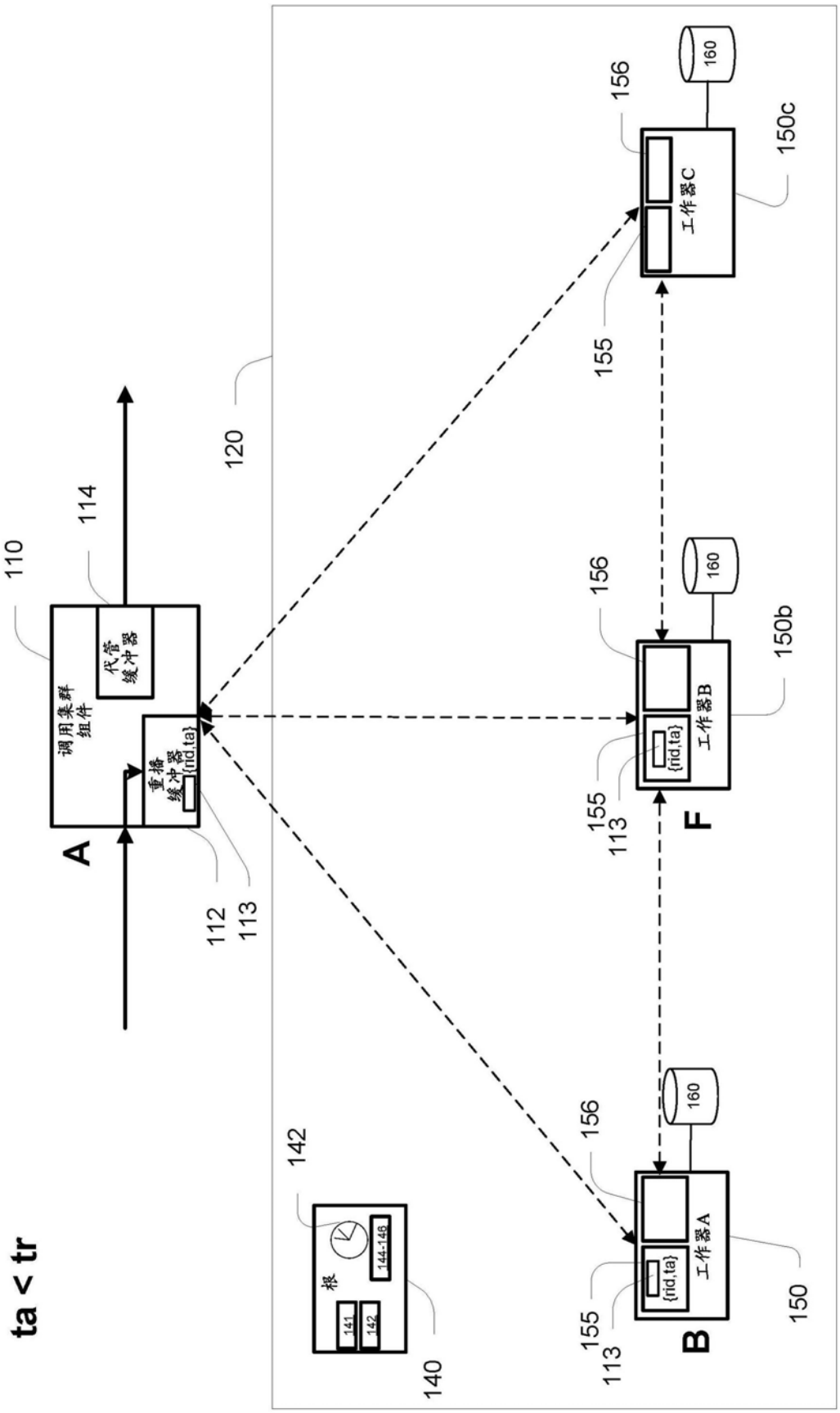

ta < tr
A
B
F
调用集群组件
代管缓冲器
重播缓冲器
{rid,ta}
工作器A
工作器B
工作器C
160
156
155
113
112
114
110
120
142
140
141
142
144-146
根
150
150b
150c

图26

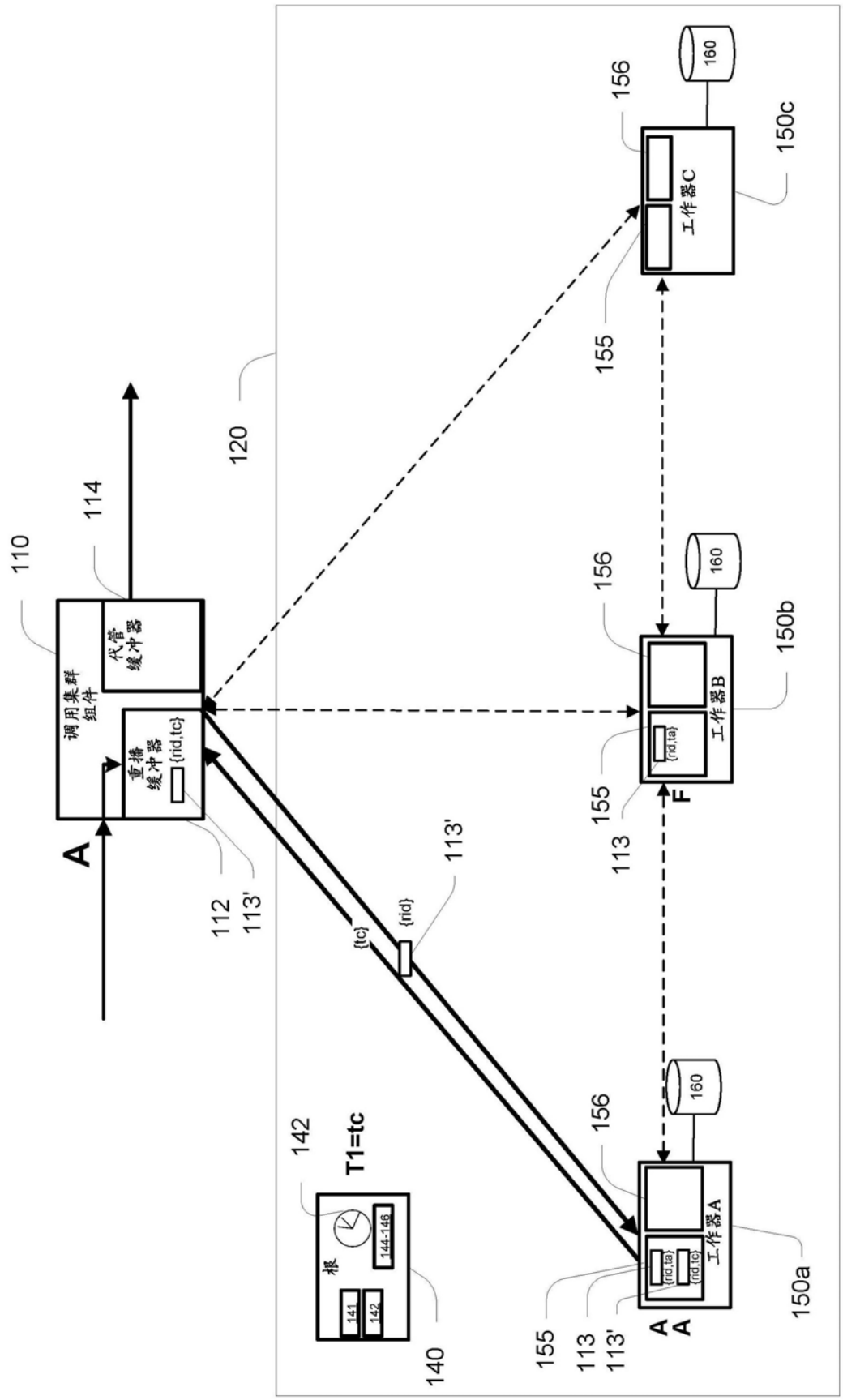

120
110
114
调用集群组件
代管缓冲器
重播缓冲器
{rid,tc}
A
112
113'
{tc}
{rid}
113'
142
T1=tc
根
141
142
144-146
140
155
113
113'
A
A
工作器A
{rid,ta}
{rid,tc}
156
160
150a
工作器B
{rid,ta}
155
113
F
156
160
150b
工作器C
156
160
155
150c

图27

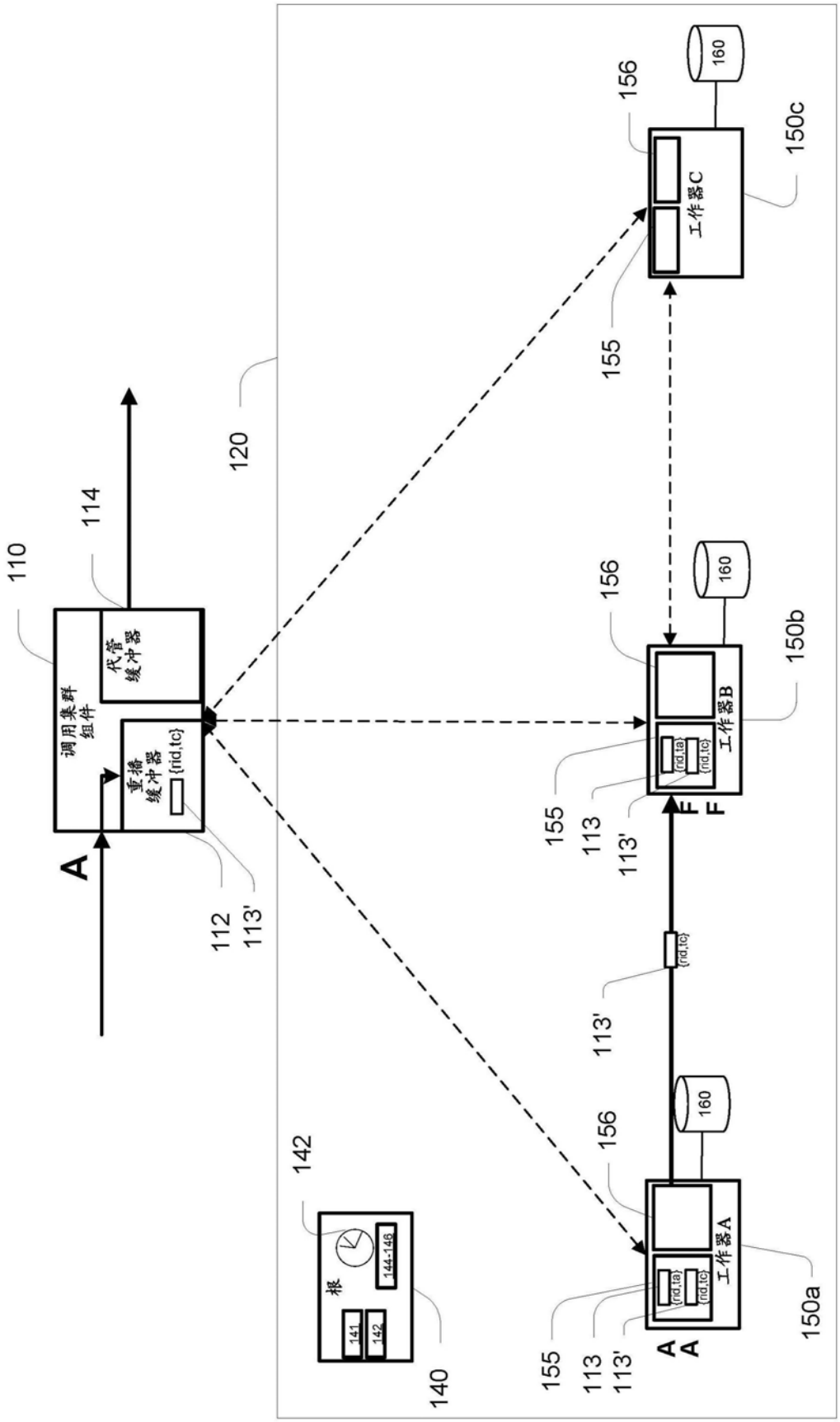

调用集群组件
代管缓冲器
重播缓冲器
{rid,tc}
110
114
120
112
113'
A
156
160
150c
工作器C
155
150b
工作器B
{rid,ta}
{rid,tc}
113
F
150a
工作器A
A
142
根
141
142
144-146
140

图28

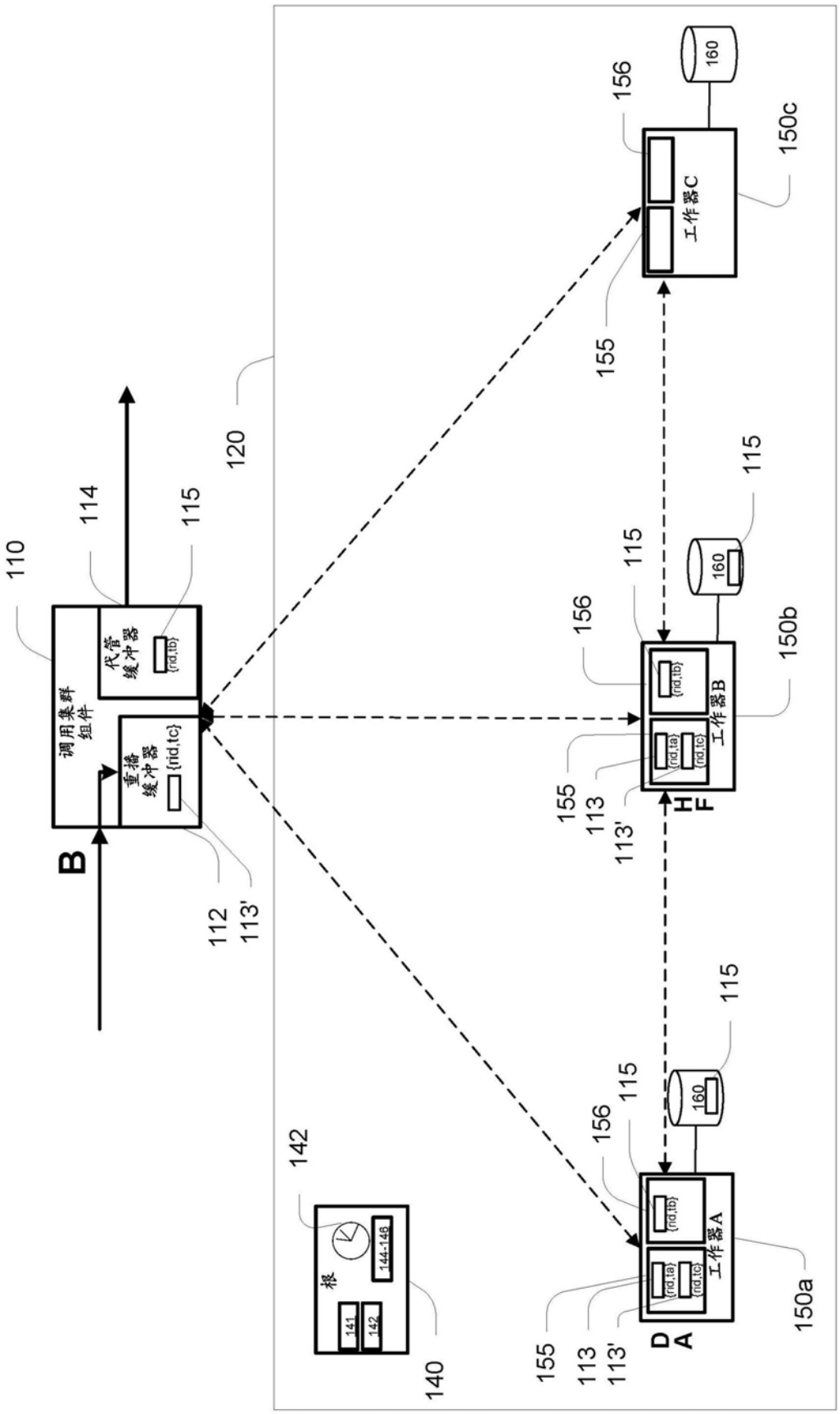

调用集群组件
代管缓冲器
{rid,tb}
重播缓冲器
{rid,tc}
工作器C
工作器B
工作器A
{rid,ta}
{rid,tb}
{rid,tc}
160
156
155
115
113
113'
112
120
114
110
B
H
F
D
A
142
144-148
141
142
140
150a
150b
150c

图29

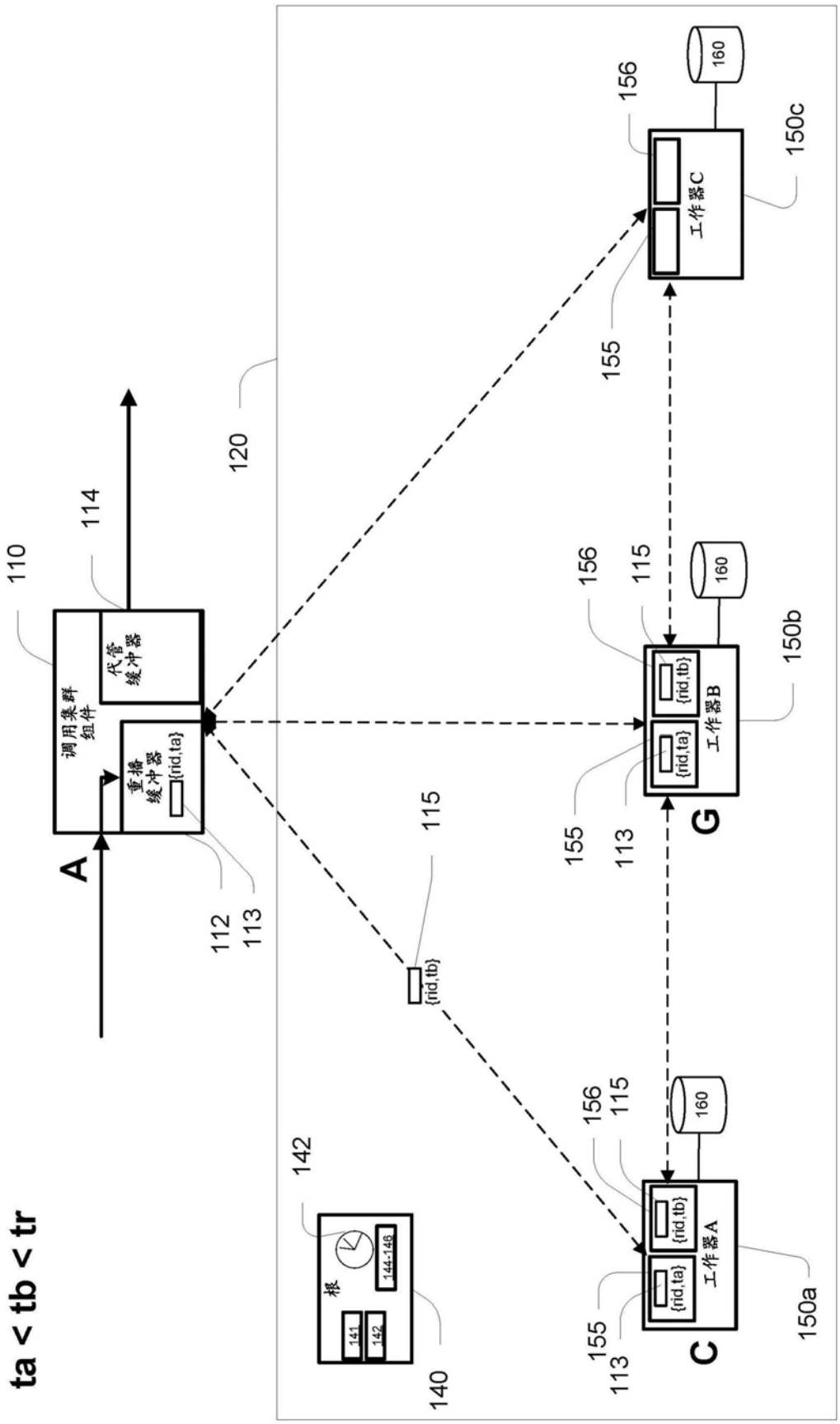

ta < tb < tr
A
G
C
调用集群组件
代管缓冲器
重播缓冲器
{rid,ta}
{rid,tb}
工作器A
工作器B
工作器C
根
141
142
144-146
110
112
113
114
115
120
140
142
150a
150b
150c
155
156
160

图30

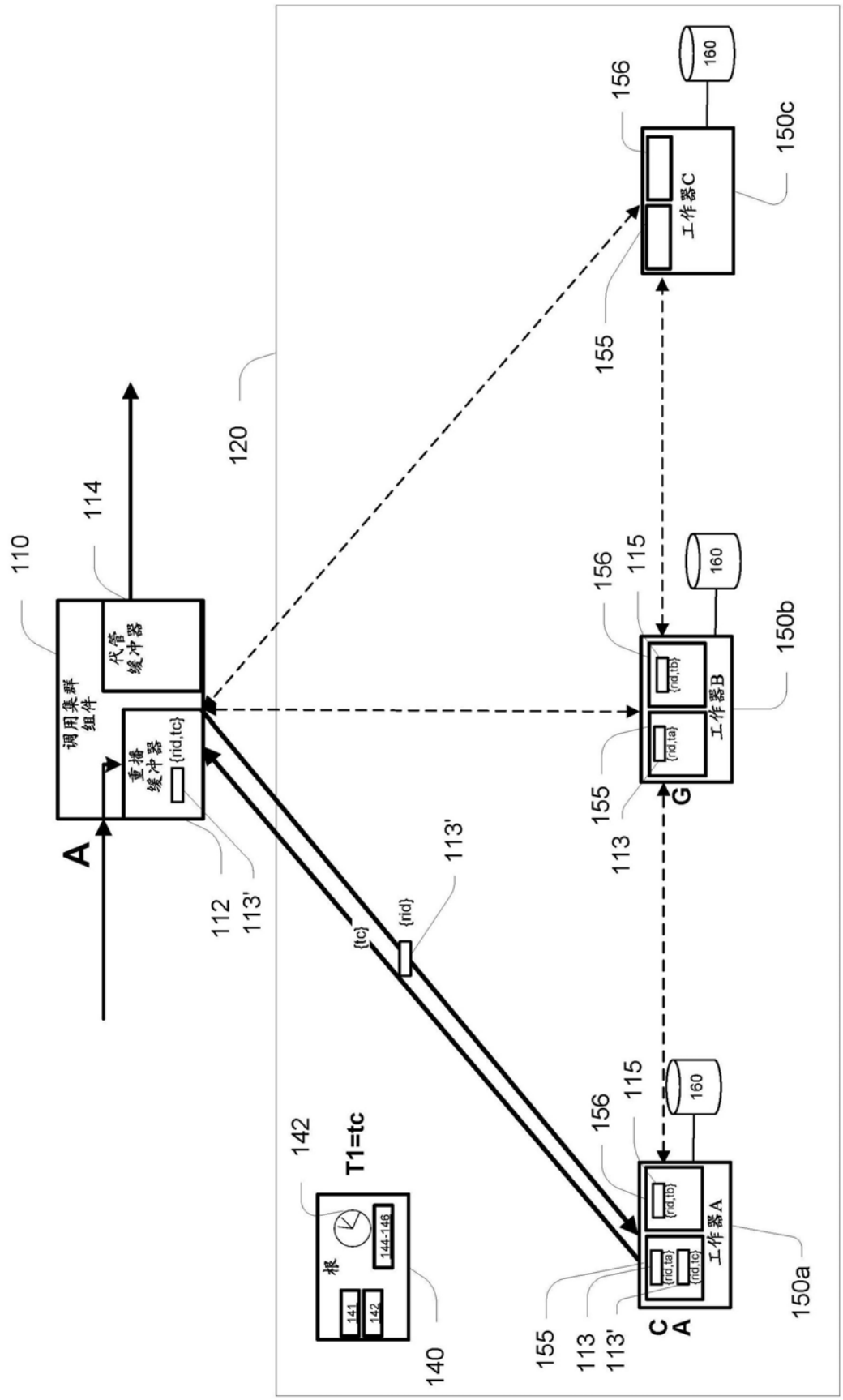

110
114
120
调用集群组件
代客缓冲器
重播缓冲器
{rid,tc}
A
112
113'
{tc}
{rid}
113'
142
T1=tc
根
141
142
144-146
140
155
113
113'
C
A
156
115
160
工作器A
{rid,tb}
{rid,ta}
{rid,tc}
150a
工作器B
150b
G
工作器C
150c

图31

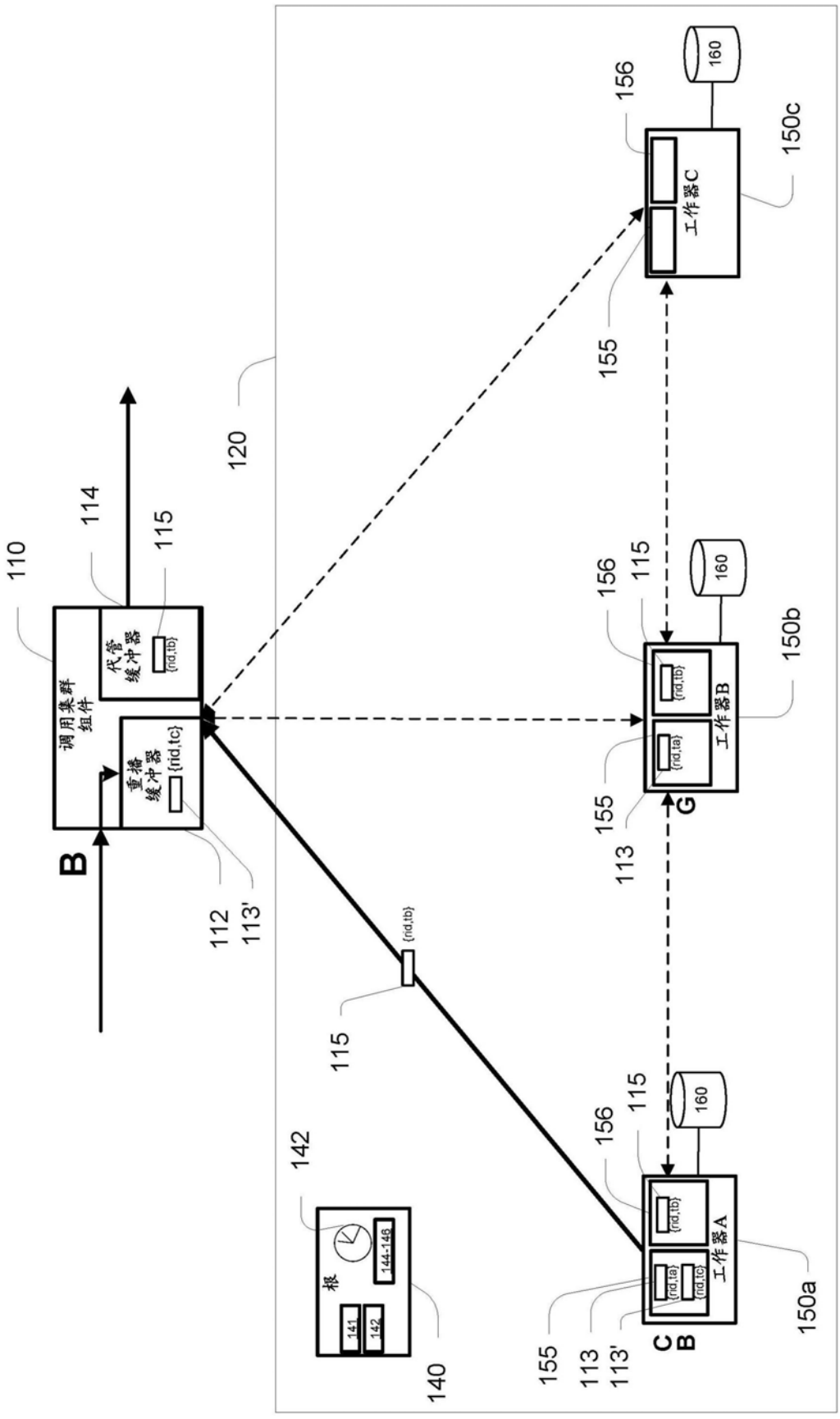

调用集群组件
代管缓冲器
{rid,tb}
重播缓冲器
{rid,tc}
110
114
115
120
112
113'
B
{rid,tb}
115
142
根
141
142
144-146
140
工作器A
{rid,ta}
{rid,tc}
{rid,tb}
150a
C
B
155
113
113'
156
115
160
工作器B
{rid,ta}
{rid,tb}
150b
G
155
113
156
115
160
工作器C
150c
155
156
160

图32

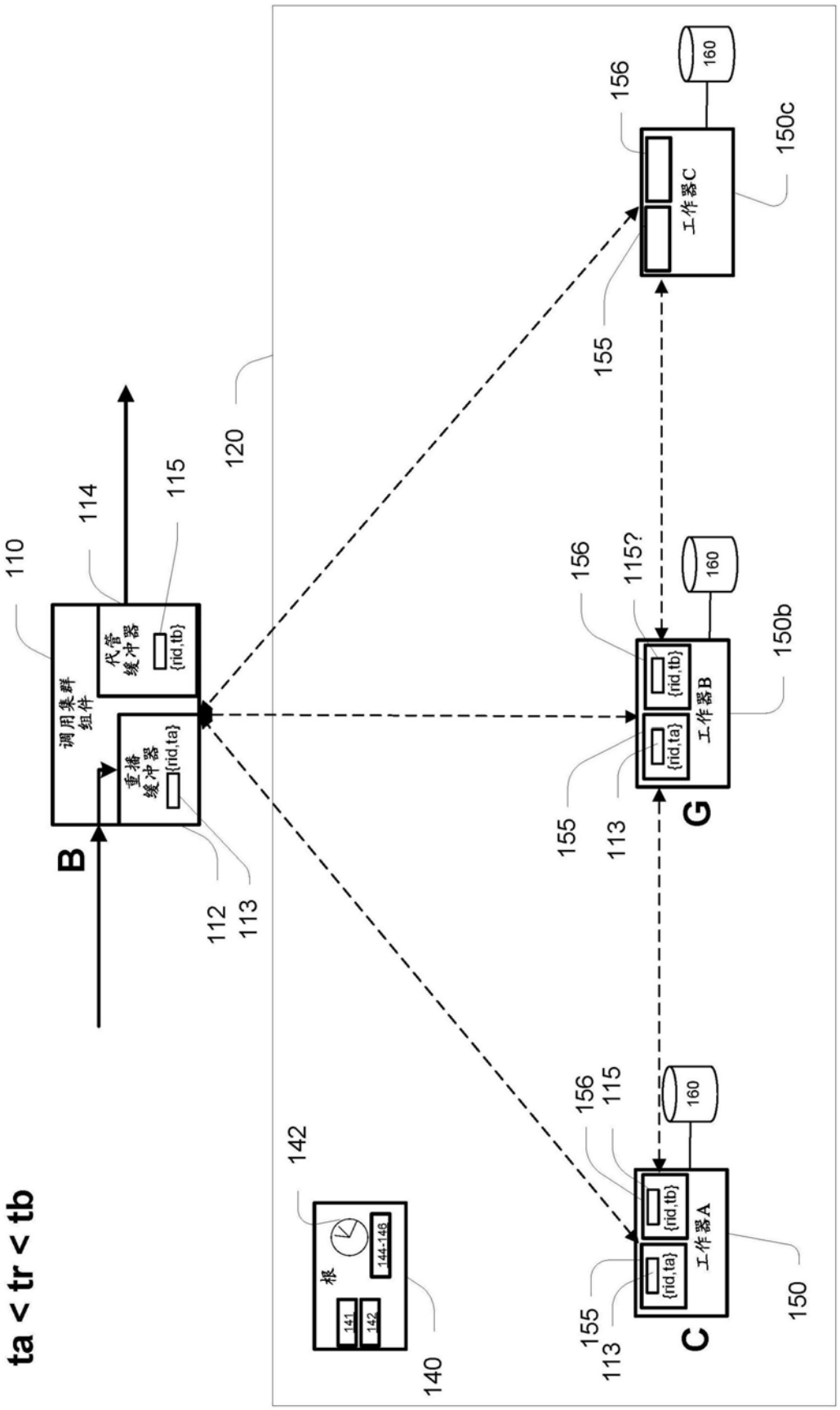

ta < tr < tb
调用集群组件
代管缓冲器
{rid,tb}
重播缓冲器
{rid,ta}
B
110
114
115
120
112
113
142
根
141
142
144-146
140
工作器A
工作器B
工作器C
150
150b
150c
160
155
156
115?
C
G

图33

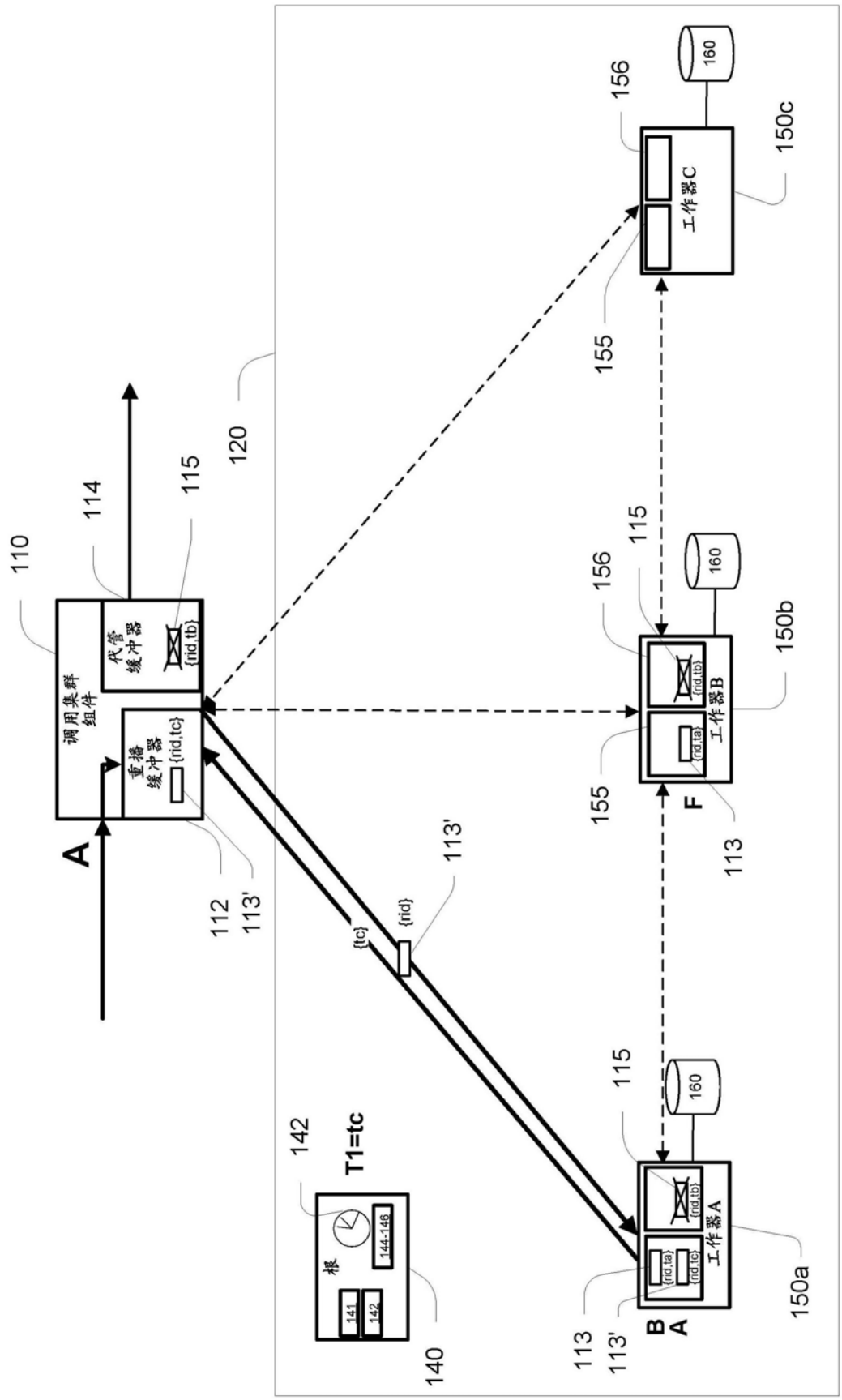

110
调用集群组件
代管缓冲器
{rid,tb}
重播缓冲器
{rid,tc}
114
115
120
112
113'
A
{tc}
{rid}
113'
142
T1=tc
根
144-146
141
142
140
156
160
150c
工作器C
155
115
156
160
150b
工作器B
{rid,tb}
{rid,ta}
155
F
113
115
160
工作器A
{rid,tb}
{rid,ta}
{rid,tc}
150a
B
A
113
113'

图34

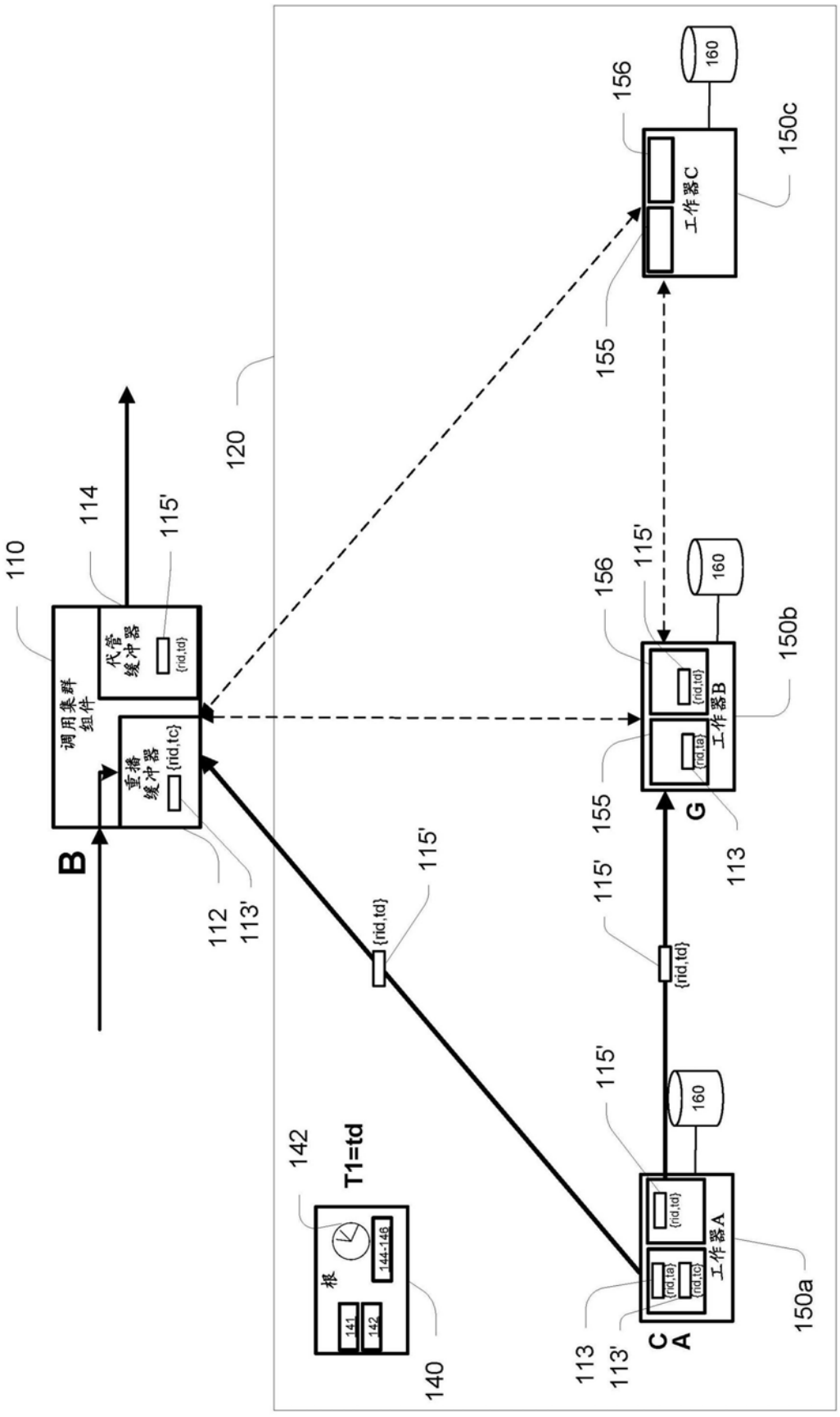

120
110
114
115'
调用集群组件
代管缓冲器
{rid,td}
重播缓冲器
{rid,tc}
B
112
113'
115'
{rid,td}
142
T1=td
根
141
142
144-146
140
156
160
150c
工作器C
155
115'
156
160
150b
工作器B
{rid,td}
{rid,ta}
155
G
113
115'
115'
{rid,td}
115'
160
工作器A
{rid,td}
{rid,ta}
{rid,tc}
150a
113
113'
C
A

图35

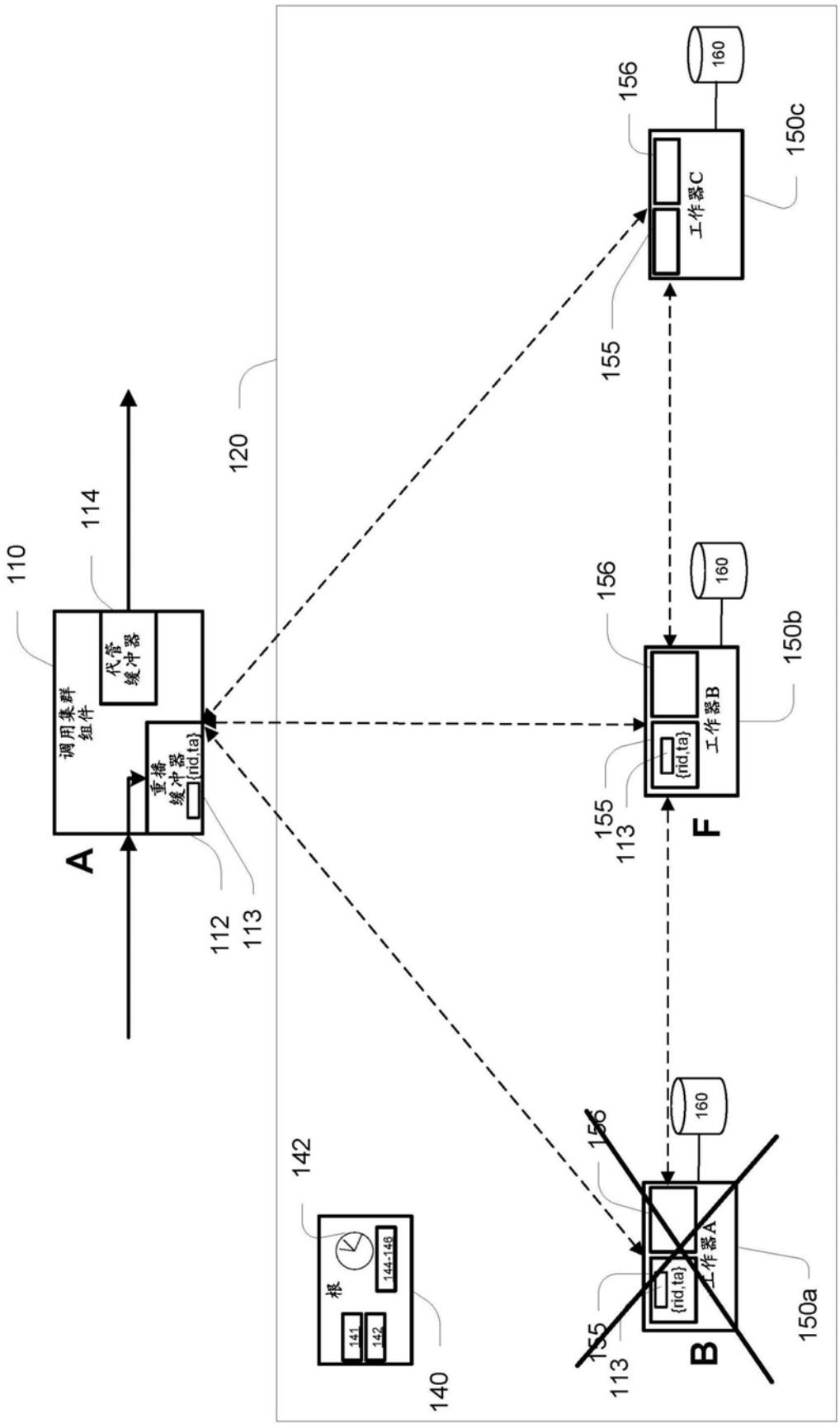

120
110
114
调用集群组件
代管缓冲器
重播缓冲器{rid,ta}
A
112
113
156
160
150c
工作器C
155
156
160
150b
工作器B
{rid,ta}
155
113
F
156
160
工作器A
{rid,ta}
155
113
B
150a
142
根
144-146
141
142
140

图36

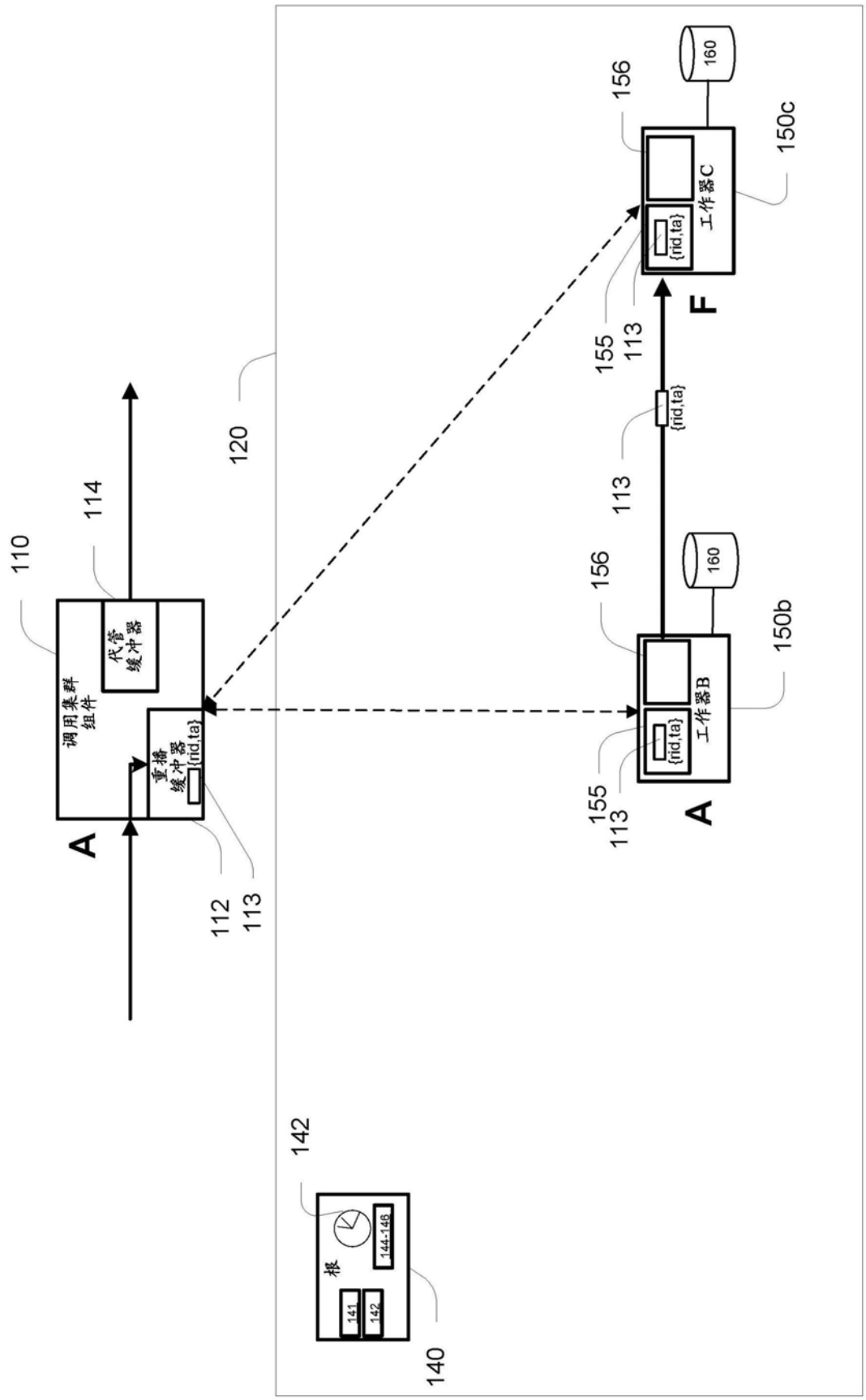

120
110
114
调用集群组件
代管缓冲器
重播缓冲器
{rid,ta}
A
112
113
142
根
144-146
141
142
140
156
160
150c
工作器C
{rid,ta}
155
113
F
113
{rid,ta}
156
160
150b
工作器B
{rid,ta}
155
113
A

图37