

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织  
国际局

(43) 国际公布日  
2024 年 5 月 30 日 (30.05.2024)



(10) 国际公布号  
**WO 2024/109239 A1**

- (51) 国际专利分类号:  
**H04L 67/1095 (2022.01) G06F 9/50 (2006.01)**
- (21) 国际申请号: PCT/CN2023/116231
- (22) 国际申请日: 2023 年 8 月 31 日 (31.08.2023)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:  
202211487122.9 2022年11月25日 (25.11.2022) CN
- (71) 申请人: 苏州元脑智能科技有限公司 (SUZHOU METABRAIN INTELLIGENT TECHNOLOGY CO., LTD.) [CN/CN]; 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。

- (72) 发明人: 黄炎 (HUANG, Yan); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。张璐 (ZHANG, Lu); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。王见 (WANG, Jian); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。
- (74) 代理人: 北京康信知识产权代理有限公司 (KANGXIN PARTNERS, P.C.); 中国北京市海淀区知春路甲48号盈都大厦A座16层, Beijing 100098 (CN)。
- (81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ,

(54) Title: CLUSTER DATA SYNCHRONIZATION METHOD AND APPARATUS, DEVICE, AND NONVOLATILE READ-ABLE STORAGE MEDIUM

(54) 发明名称: 集群数据同步方法、装置、设备及非易失性可读存储介质

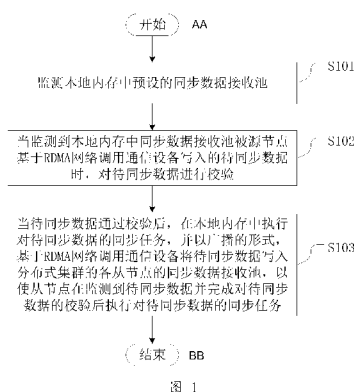


图 1

- S101 Monitor a synchronization data receiving pool provided in advance in local memory
- S102 When detecting that data to be synchronized has been written into the local memory synchronization data receiving pool by a communication device invoked by a source node on the basis of an RDMA network, verify the data to be synchronized
- S103 When the data to be synchronized passes verification, then within the local memory, perform a synchronization task on the data to be synchronized, and use a broadcasting method on the basis of the RDMA network to invoke a communication device to write the data to be synchronized to a synchronization data receiving pool of each secondary node in a distributed cluster, so that when the secondary nodes detect and verify the data to be synchronized, the secondary nodes perform the synchronization task on the data to be synchronized

AA Start  
BB End

(57) Abstract: The present application is applicable in the technical field of data synchronization, and discloses a cluster data synchronization method and apparatus, a device, and a non-volatile readable storage medium. The method comprises: a source node, which generates data to be synchronized, using an RDMA network to invoke a communication device to directly write the data to be synchronized into a synchronization data receiving pool provided in advance in the local memory of a primary node; the primary node, within the local memory, performing a synchronization task on the data to be synchronized, and using a broadcasting method on the basis of the RDMA network to invoke a communication device to directly write the data to be synchronized to a synchronization data receiving pool of a secondary node, causing the secondary node to perform the synchronization task on the data to be synchronized, thereby using a direct writing method to transmit data to be synchronized, and replacing the sending/receiving synchronization mode of the existing Raft protocol, thereby avoiding memory copying in the sending/receiving process, and the primary node likewise not needing to control IO multiplexing during cluster data synchronization operations, thereby reducing the amount of code required for synchronization and reducing the CPU scheduling work of the primary node.

(57) 摘要: 本申请适用于数据同步技术领域, 公开了一种集群数据同步方法、装置、设备及非易失性可读存储介质, 产生待同步数据的源节点基于RDMA网络调用通信设备将待同步数据直接写入主节点的本地内存中预设的同步数据接收池, 主节点在本地内存中执行对待同步数据的同步任务, 并以广播的形式基于RDMA网络调用通信设备将待同步数据直接写入从节点的同步数据接收池, 以使从节点执行对待同步数据的同步任务, 以直接写入的方式进行待同步数据的传递, 代替现有Raft协议的发送/接收的同步模式, 可以避免发送/接收过程中的内存拷贝, 主节点也无需在集群数据同步工作中控制IO复用, 减轻了实现同步的代码量以及主节点的CPU调度工作。

IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ,  
LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN,  
MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA,  
PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD,  
SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ,  
UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区  
保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ,  
NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚  
(AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE,  
BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR,  
HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO,  
PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF,  
CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN,  
TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

## 集群数据同步方法、装置、设备及非易失性可读存储介质

### 相关申请的交叉引用

本申请要求于 2022 年 11 月 25 日提交中国专利局，申请号为 202211487122.9，申请名称为“集群数据同步方法、装置、设备及计算机可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

### 技术领域

本申请涉及数据同步技术领域，特别是涉及一种集群数据同步方法、装置、设备及非易失性可读存储介质。

### 背景技术

分布式环境中需要大量的工作来保证数据一致性，避免不同节点间数据差异而导致的访问结果不同。这个过程需要大量的开发工作，且由于同步流程复杂，通常运行效率并不高。目前主流的数据一致性协议有两种，Paxos 和 Raft，其他所有协议均是这两种协议的变种。其中，Raft 是一种更为简单方便易于理解的分布式算法，在解决分布式中的一致性问题上，相比传统的 Paxos 算法（一种基于消息传递的分布式一致性算法），Raft 将大量的计算问题分解成为了一些简单的相对独立的子问题。

远程直接数据存取（Remote Direct Memory Access，下文简称 RDMA）是一种为解决网络传输中服务器端数据处理的延迟而产生的技术。RDMA 通过网络将资料直接传入计算机的存储区，将数据从一个系统快速移动到远程系统存储器中，而不对操作系统造成任何影响。作为一种全新的网络数据传输机制，RDMA 的实现需要软件、硬件的配合。而相较于传统的传输控制协议（Transmission Control Protocol，下文简称 TCP）/用户数据报协议（User Datagram Protocol，下文简称 UDP）来说，RDMA 可以基于多种底层协议，灵活性较高，无需像 TCP/UDP 协议那样在发送数据时需要进行应用态/内核态切换以及多次内存拷贝，性能更优。且 RDMA 协议自带无损数据网络，可以解决 TCP 协议的粘包、重发问题以及 UDP 协议的丢包问题。且基于 RDMA 协议发送数据无需经过中央处理器（Central Processing Unit，下文简称 CPU）调度，可以节约 CPU 资源，避免中断的产生。

然而，目前 Raft 协议与 RDMA 技术的结合，都是基于基础的发送、接收逻辑，即仅仅是将 RDMA 作为 TCP/UDP 的平行替代方案。这种方案虽然可以实现 Raft 协议的逻辑，但是没有结合 RDMA 网络独有的特性，导致 RDMA 的性能优势没有完全发挥出来，依然存在着数据同步效率低的问题。

对于上述问题，相关技术中并未提出有效的解决方案。

### 发明内容

本申请的目的是提供一种集群数据同步方法、装置、设备及非易失性可读存储介质，被设置为结合 RDMA 网络对 Raft 协议集群数据同步进行优化，充分发挥 Raft 协议集群中 RDMA 的性能优势。

为解决上述技术问题，在本申请的第一个方面，提供一种集群数据同步方法，基于分布式集群中的主节点，包括：

监测本地内存中预设的同步数据接收池；

当监测到本地内存中同步数据接收池被源节点基于 RDMA 网络调用通信设备写入的待同步数据时，对待同步数据进行校验；

当待同步数据通过校验后，在本地内存中执行对待同步数据的同步任务，并以广播的形式，基于 RDMA 网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务；

其中，源节点为分布式集群中的一个从节点。

在一些实施例中，监测到本地内存中同步数据接收池被源节点基于 RDMA 网络调用通信设备写入的待同步数据，可以为：

监测到源节点在写完待同步数据后基于 RDMA 网络的任务完成回调函数在主节点的 CPU 中更新的同步请求字段时，确认源节点写入了待同步数据；

基于 RDMA 网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，可以为：

基于 RDMA 网络调用通信设备将待同步数据写入从节点的同步数据接收池后，基于任务完成回调函数在从节点的 CPU 中更新同步请求字段，以使从节点在监测到更新的同步请求字段后完成对待同步数据的校验并执行对待同步数据的同步任务。

在一些实施例中，监测到本地内存中同步数据接收池被源节点基于 RDMA 网络调用通信设备写入的待同步数据，可以为：

当轮询监测到本地内存中同步数据接收池有新增的待同步数据时，确认源节点写入了待同步数据；

基于 RDMA 网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，可以为：

基于 RDMA 网络调用通信设备将待同步数据写入从节点的同步数据接收池，以使从节点在轮询监测到待同步数据后，完成对待同步数据的校验并执行对待同步数据的同步任务。

在一些实施例中，监测到本地内存中同步数据接收池被源节点基于 RDMA 网络调用通信设备写入的待同步数据，可以为：

接收到源节点在写完待同步数据后调用输入输出端口发送的写完成标志时，确认源节点写入了待同步数据；

基于 RDMA 网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，可以为：

基于 RDMA 网络调用通信设备将待同步数据写入从节点的同步数据接收池后，调用输入输出端口向从节点发送写完成标志，以使从节点确认被写入待同步数据后，完成对待同步数据的校验并执行对待同步数据的同步任务。

在一些实施例中，监测到本地内存中同步数据接收池被源节点基于 RDMA 网络调用通信设备写入的待同步数据，可以为：

监测到源节点在写完待同步数据后在主节点的写完成标志约定地址写入的写完成标志时，确认源节点写入了待同步数据；

基于 RDMA 网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，可以为：

基于 RDMA 网络调用通信设备将待同步数据写入从节点的同步数据接收池后，在各从节点的写完成标志约定地址写入写完成标志，以使从节点确认被写入待同步数据后，完成对待同步数据的校验并执行对待同步数据的同步任务。

在一些实施例中，同步数据接收池与分布式集群中除所在节点外的其余节点一一对应。

在一些实施例中，同步数据接收池设有用于维护所在节点外的其余节点与同步数据接收池的对应关系的 RDMA 映射数组以及用于为待同步数据分配空余空间的原子索引。

在一些实施例中，同步数据接收池中设有多类与不同类型的待同步数据对应的子接收池；

相应的，在本地内存中执行对待同步数据的同步任务，可以为：

按照待同步数据的类型，在本地内存中执行对待同步数据的同步任务；

从节点执行对待同步数据的同步任务，可以为：

从节点按照待同步数据的类型执行对待同步数据的同步任务。

在一些实施例中，同步数据接收池中预先按照各类型待同步数据的数据长度划分为各类型子接收池的数据分区。

在一些实施例中，子接收池的类型包括：小数据接收池、应答数据接收池和大数据接收池。

在一些实施例中，当待同步数据为与小数据接收池对应的数据类型时，

源节点基于 RDMA 网络调用通信设备写入待同步数据，可以为：

源节点将待同步数据组装为第一消息头后，根据本地存储的主节点中同步数据接收池的存储区域信息，基于 RDMA 网络调用通信设备将第一消息头写入对应的存储位置；

基于 RDMA 网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，可以为：

根据本地存储的从节点中同步数据接收池的存储区域信息，基于 RDMA 网络调用通信设备将第一消息头写入对应的存储位置。

在一些实施例中，当待同步数据为与应答数据接收池对应的数据类型时，

源节点基于 RDMA 网络调用通信设备写入待同步数据，可以为：

源节点将待同步数据组装为第二消息头后，根据本地存储的主节点中同步数据接收池的存储区域信息，基于 RDMA 网络调用通信设备将第二消息头写入对应的存储位置；

基于 RDMA 网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，可以为：

根据本地存储的从节点中同步数据接收池的存储区域信息，基于 RDMA 网络调用通信设备将第二消息头写入对应的存储位置。

在一些实施例中，应答数据接收池的第一预设空间用于存放分布式集群中所有节点的心跳信息。

在一些实施例中，还包括：

定时向本地内存的第一预设空间写入当前时间戳；

通过 RDMA 网络将当前时间戳同步至各从节点的第一预设空间；

读取各从节点写入本地内存的响应时间戳；

根据各响应时间戳检查各从节点的连接状态。

在一些实施例中，应答数据接收池的第二预设空间用于存放用户自定义数据。

在一些实施例中，当待同步数据为与大数据接收池对应的数据类型时，

源节点基于 RDMA 网络调用通信设备写入待同步数据，可以为：

源节点将待同步数据的同步任务组装为第三消息头后，根据本地存储的主节点中同步数据接收池的存储区域信息，基于 RDMA 网络调用通信设备将第三消息头写入对应的存储位置；

在本地内存中执行对待同步数据的同步任务，并以广播的形式，基于 RDMA 网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，包括：

根据本地存储的从节点中同步数据接收池的存储区域信息，以广播的形式，基于 RDMA 网络调用通信设备将第三消息头写入对应的存储位置，以使从节点在从节点的大数据接收池的第三预设空间的分页内执行第三消息头所包含的逻辑；

在本地的大数据接收池的第三预设空间的分页内执行第三消息头所包含的逻辑；

在确认本地以及各从节点均完成执行第三消息头所包含的逻辑之后，控制本地以及各从节点将所在节点的第三预设空间的分页拷贝至所在节点的集群内存空间。

在一些实施例中，在确认本地以及各从节点均完成执行第三消息头所包含的逻辑之后，控制本地以及各从节点将所在节点的第三预设空间的分页拷贝至所在节点的集群内存空间，可以为：

在确认本地以及各从节点均完成执行第三消息头所包含的逻辑之后，向各从节点发送确认成功响应码，以使各从节点在接收到确认成功响应码之后与主节点各自执行下述步骤：

将第三预设空间对应的集群内存空间中的第四预设空间的业务逻辑地址交换至第三预设空间，以基于第三预设空间的数据执行下一个处理逻辑，同时将第三预设空间的数据拷贝至第四预设空间；

在拷贝完成后，将业务逻辑地址交换回第四预设空间。

在一些实施例中，当待同步数据为与大数据接收池对应的数据类型且为非事务数据时，第三消息头仅包含待同步数据的同步逻辑；

当待同步数据为与大数据接收池对应的数据类型且为事务数据时，第三消息头包含待同步数据的同步逻辑以及对待同步数据对应的事务回调函数，以使主节点或从节点在执行对待同步数据的同步任务时执行事务回调函数对应的逻辑；

其中，事务回调函数至少包括针对事务数据同步的数据备份函数以及关键数据保护函数。

在一些实施例中，执行事务回调函数对应的逻辑包括：

检查待同步数据的同步任务的合理性；

执行数据备份及关键数据保护流程；

若待同步数据的同步任务合理且成功完成数据备份及关键数据保护流程，则向主节点返回准备逻辑执行成功消息；

在准备逻辑执行成功后，触发执行待同步数据的同步任务的逻辑；

在执行完毕待同步数据的同步任务的逻辑后，向主节点返回同步任务执行完毕消息；

若主节点及各从节点均成功执行待同步数据的同步任务的逻辑，则释放关键数据保护，删除数据备份；否则，执行发生错误后回调操作，释放关键数据保护并基于数据备份恢复集群数据至初始状态。

为解决上述技术问题，在本申请的第二个方面，还提供一种集群数据同步装置，包括：

监测单元，被设置为监测本地内存中预设的同步数据接收池；

校验单元，被设置为当监测到本地内存中同步数据接收池被源节点基于 RDMA 网络调用通信设备写入的待同步数据时，对待同步数据进行校验；

同步单元，被设置为当待同步数据通过校验后，在本地内存中执行对待同步数据的同步任务，并以广播的形式，基于 RDMA 网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务；

其中，源节点为分布式集群中的一个从节点。

为解决上述技术问题，本申请还提供一种集群数据同步设备，包括：

存储器，被设置为存储计算机程序；

处理器，被设置为执行计算机程序，计算机程序被处理器执行时实现如上述任意一项集群数据同步方法的步骤。

为解决上述技术问题，在本申请的第三个方面，还提供一种非易失性可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现如上述任意一项集群数据同步方法的步骤。

本申请所提供的集群数据同步方法，提供了一种结合 RDMA 网络对 Raft 协议优化的集群数据同步方案，产生待同步数据的源节点基于 RDMA 网络调用通信设备直接将待同步数据写入主节点的本地内存中预设的同步数据接收池，主节点在对待同步数据校验通过后，在本地内存中执行对待同步数据的同步任务，并以广播的形式，基于 RDMA 网络调用通信设备将待同步数据直接写入从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，以直接写入的方式进行待同步数据的传递，代替现有 Raft 协议的发送/接收的同步模式，可以避免发送/接收过程中的内存拷贝，主节点也无需在集群数据同步工作中控制 IO（Input/Output，输入/输出）复用，无需流控制，显著减轻了实现同步的代码量以及主节点的 CPU 调度工作量。

本申请还提供一种集群数据同步装置、设备及非易失性可读存储介质，具有上述有益效果，在此不再赘述。

## 附图说明

为了更清楚的说明本申请实施例或相关技术的技术方案，下面将对实施例或相关技术描述中所需要使用的附图作简单的介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图1为本申请实施例提供的一种集群数据同步方法的流程图；

图2为本申请实施例提供的一种集群内存分布示意图；

图3为本申请实施例提供的一种集群数据同步装置的结构示意图；

图4为本申请实施例提供的一种集群数据同步设备的结构示意图。

## 具体实施方式

本申请的核心是提供一种集群数据同步方法、装置、设备及非易失性可读存储介质，被设置为结合RDMA网络对Raft协议集群数据同步进行优化，充分发挥Raft协议集群中RDMA的性能优势。

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

图1为本申请实施例提供的一种集群数据同步方法的流程图。

如图1所示，基于分布式集群中的主节点，本申请实施例提供的集群数据同步方法包括：

S101：监测本地内存中预设的同步数据接收池。

S102：当监测到本地内存中同步数据接收池被源节点基于RDMA网络调用通信设备写入的待同步数据时，对待同步数据进行校验。

S103：当待同步数据通过校验后，在本地内存中执行对待同步数据的同步任务，并以广播的形式，基于RDMA网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务。

其中，源节点为分布式集群中的一个从节点。

在本申请实施例中，主节点（leader）指Raft协议中的主节点，可以为在基于Raft协议进行集群数据同步的分布式集群中通过自动选举产生的仲裁节点。根据Raft协议，所有的从节点（follower）进行状态更新时需要发送消息至主节点，由主节点广播至整个集群。

现有的基于Raft协议进行集群数据同步时还是采用类似传统的TCP/UDP协议的IO收发方式，导致集群同步过程每发送一次数据，都需要将待同步数据在本地拷贝至CPU内核，再通过IO发给目的端，目的端再从CPU内核拷贝到内存中，占用了大量CPU资源、同步效率较低，即使已经应用了Raft协议与RDMA结合的方案，也只是将RDMA当做TCP/UDP的平行替代方案，在进行数据同步时没有结合RDMA网络独有的特性，导致RDMA的性能优势没有完全发挥出来。为发挥RDMA网路的优势，本申请实施例提供的集群数据同步方法提供一种采用更高效率的写/读结合（write/read）的方式代替现有Raft协议的发送/接收（send/recv）的同步模式，比如，是基于RDMA网络下的通信设备，从节点将待同步数据直接写入主节点的内存，主节点再以广播的形式将待同步数据同步至分布式集群中各从节点，主节点和各从节点各自按照待同步数据执行同步任务，完成整个分布式集群的集群数据同步任务。

为实现上述基于RDMA网络的集群数据同步方案，首先要设计集群的内存分布机制。在本申请实施例中，在分布式集群中的主节点的内存和各从节点的内存中，除了集群内存空间（负责存储实际集群数据逻辑，每个节点都有一份，一致性协议目的就是要保证每个节点上集群内存空间的数据一致性）外，还预设有用用于存储待同步数据的同步数据接收池（负责处理其他节点发送的待同步数据）。

为便于管理，在本申请实施例提供的集群数据同步方法中，可以设计同步数据接收池与分布式集群中除所在节点外的其余节点一一对应。

此外，可以在同步数据接收池设有用于维护所在节点外的其余节点与同步数据接收池的对应关系的RDMA映射数组以及用于为待同步数据分配空余空间的原子（atomic）索引。通过原子（atomic）索引，可以对分布式集群中的关键数据进行保护，以保证在待同步数据写入以及执行同步任务时，若有线程访问相关数据，不会访问到未完成同步的数据。

以四节点集群为例，对于任意节点，其内存空间划分为集群内存空间和同步数据接收池，且每个节点上同步数据接收池的数量均为三个，每个同步数据接收池均由一个RDMA映射数组和原子索引构成，用于对该同步数据接收池中的数据进行维护。

为确定要将待同步数据写入的位置，各节点中预先存储有分布式集群中各节点中的内存区域（memory region）信息，即在各节点中预设的同步数据接收池与其他节点的对应关系。

在本申请实施例中，按照Raft协议进行集群数据同步的规则，当某个节点中的集群数据发生变更时，定义该节点为集群数据同步的源节点，由源节点将发生变更的数据作为待同步数据发送至主节点，再由主节点将待同步数据以广播的形式发送给分布式集群中各从节点进行集群数据同步。

则在本申请实施例提供的集群数据同步方法中，对于S101来说，主节点根据预先部署的监测规则，监测本地内存中的同步数据接收池是否被写入待同步数据。例如，主节点可以定时检查同步数据接收池，如发现新增的数据，即确定为待同步数据。或者，源节点也可以采用传统协议的通过IO来发送写完成标志的方式，来告知主节点已将待同步数据写入主节点的同步数据接收池，这相较于完全采用IO的方式发送待同步数据，对IO的占用时间有所减少，但仍然会占用IO。另外，还可以采用源节点在完成待同步数据的写入后，在待同步数据之后的地址或主节点的同步数据接收池中的另一约定地址写入写完成标志，等待主节点定时轮询同步数据接收池中是否出现写完成标志，当主节点发现写完成标志时，确定为待同步数据，但这需要主节点的CPU轮询大量的内存数据。

对于S102，当主节点的CPU监测到本地内存中同步数据接收池被源节点写入待同步数据后，启动在整个分布式集群中的集群数据同步工作，首先在主节点上对待同步数据进行信息校验。可选的校验方式可以参考现有的分布式集群同步方案，在此不再赘述。

对于S103，按照Raft协议进行集群数据同步的规则，当主节点通过对待同步数据的信息校验后，以广播的形式，基于RDMA网络调用通信设备将待同步数据直接写入各从节点的同步数据接收池，以使各从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务。

其中，主节点将待同步数据写入从节点的同步数据接收池的过程，与源节点将待同步数据写入主节点的同步数据接收池的方式类似，均为通过调用通信设备将待同步数据写入目的端的同步数据接收池。相应的，从节点监测本地内存中的同步数据接收池、当检测到待同步数据后启动校验，并在校验通过后执行对待同步数据的同步任务，与主节点在监测到待同步数据后执行对待同步数据的同步任务的过程类似，可以参考上述S101的说明。

主节点与各从节点在识别到本地内存的同步数据接收池中的待同步数据并通过校验后，各自根据待同步数据执行本地的同步任务。如直接按照待同步数据进行本地的同步任务即完成集群数据同步任务，则从节点无需向主节点返回确认码，由此节约了集群数据同步所需的CPU调度工作。

当产生集群数据变更的节点为主节点，则主节点在完成对待同步数据的校验后，直接执行S103即可。

在本申请实施例提供的集群数据同步方法中，通信设备可以采用RoCE网卡。RoCE全称RDMA over Converged Ethernet，即基于融合以太网的RDMA。

本申请实施例提供的集群数据同步方法，提供了一种结合RDMA网络对Raft协议优化的集群数据同步方案，产生待同步数据的源节点基于RDMA网络调用通信设备直接将待同步数据写入主节点的本地内存中预设的同步数据接收池，主节点在对待同步数据校验通过后，在本地内存中执行对待同步数据的同步任务，并以广播的形式，基于RDMA网络调用通信设备将待同步数据直接写入从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，以直接写入的方式进行

待同步数据的传递，代替现有Raft协议的发送/接收的同步模式，可以避免发送/接收过程中的内存拷贝，主节点也无需在集群数据同步工作中控制IO复用，无需流控制，显著减轻了实现同步的代码量以及主节点的CPU调度工作量。

不论是源节点将待同步数据写入主节点的同步数据接收池，还是主节点将待同步数据写入从节点的同步数据接收池，即不论是源端是源节点、目的端是主节点，还是源端是主节点、目的端是从节点，均需要设置使目的端识别到待同步数据的写入的机制。

在上述实施例中介绍了可以由目的端定时检查本地的同步数据接收池中是否有新增数据的方式来识别到待同步数据。则在本申请实施例提供的集群数据同步方法中，S102中主节点监测到本地内存中同步数据接收池被源节点基于RDMA网络调用通信设备写入的待同步数据，可以为：当轮询监测到本地内存中同步数据接收池有新增的待同步数据时，确认源节点写入了待同步数据。S103中主节点基于RDMA网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，可以为：基于RDMA网络调用通信设备将待同步数据写入从节点的同步数据接收池，以使从节点在轮询监测到待同步数据后，完成对待同步数据的校验并执行对待同步数据的同步任务。即源端在将待同步数据写入目的端后，无需通知目的端写入完成，而是由目的端轮询本地内存，当检查到同步数据接收池中有新增数据时则认为是待同步数据，再启动同步操作。

或者可以采用传统协议的通过IO的方式在源端写完待同步数据后，再通过IO将写完成标志发给目的端，以使目的端知晓本地的同步数据接收池被写入待同步数据。则在本申请实施例提供的集群数据同步方法中，S102中主节点监测到本地内存中同步数据接收池被源节点基于RDMA网络调用通信设备写入的待同步数据，可以为：接收到源节点在写完待同步数据后调用输入输出（IO）端口发送的写完成标志时，确认源节点写入了待同步数据。S103中主节点基于RDMA网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，可以为：基于RDMA网络调用通信设备将待同步数据写入从节点的同步数据接收池后，调用输入输出（IO）端口向从节点发送写完成标志，以使从节点确认被写入待同步数据后，完成对待同步数据的校验并执行对待同步数据的同步任务。即源端在将待同步数据写入目的端后，通过调用目的端的IO的方式发送写完成标志，以使目的端获悉本地同步数据接收池中被写入了待同步数据。

或者可以由源端将待同步数据写入完成后，将写完成标志继续写入待同步数据之后的地址或目的端的同步数据接收池的另一约定地址，等待目的端定时轮询发信写完成标志时确定为待同步数据。则在本申请实施例提供的集群数据同步方法中，S102中主节点监测到本地内存中同步数据接收池被源节点基于RDMA网络调用通信设备写入的待同步数据，可以为：监测到源节点在写完待同步数据后在主节点的写完成标志约定地址写入的写完成标志时，确认源节点写入了待同步数据。S103中主节点基于RDMA网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，可以为：基于RDMA网络调用通信设备将待同步数据写入从节点的同步数据接收池后，在各从节点的写完成标志约定地址写入写完成标志，以使从节点确认被写入待同步数据后，完成对待同步数据的校验并执行对待同步数据的同步任务。可选地，源端可以在基于RDMA网络调用通信设备将待同步数据写入目的端之后，继续基于RDMA网络调用通信设备将写完成标志写入目的端中待同步数据之后或另一约定地址（即写完成标志约定地址），目的端轮询写完成标志约定地址识别到写完成标志后，确认本地同步数据接收池中被写入了待同步数据。

而上述方式，要么需要占用IO来传输写完成标志，要么目的端无法及时发现写完成标志、造成集群数据同步任务无法及时启动。

故在上述实施例的基础上，在本申请实施例提供的集群数据同步方法中，S102中主节点监测到本地内存中同步数据接收池被源节点基于RDMA网络调用通信设备写入的待同步数据，可以为：监测到源节点在写完待同步数据后基于RDMA网络的任务完成回调函数在主节点的CPU中更新的同步请求字段后，确认源节点写

入了待同步数据。

S103中主节点基于RDMA网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，可以为：基于RDMA网络调用通信设备将待同步数据写入从节点的同步数据接收池后，基于任务完成回调函数在从节点的CPU中更新同步请求字段，以使从节点在监测到更新的同步请求字段后完成对待同步数据的校验并执行对待同步数据的同步任务。

在可选实施中，在各节点的CPU的控制程序中增加同步请求字段（可以定义为available字段），通过RDMA技术特有的completion IO回调功能（即上文定义的任务完成回调函数）来标识有同步数据接收池中有数据写入。源端在完成对待同步数据的写入后，利用任务完成回调函数更新目的端的同步请求字段（如+1），表示有数据发送，需要等待目的端处理。在这个过程中，使用RDMA无损数据网络，则可以省略源端检查发送失败的流程，同样减少了代码量。在目的端中，增加一个循环遍历与同步数据接收池对应的同步请求字段的检查线程，遍历循环各源端对应的同步数据接收池对应的同步请求字段，例如当发现某个节点对应的同步请求字段有一个等待处理的请求（available =1），则确认监测到待同步数据，进入处理待同步数据的流程。

当各节点均根据待同步数据完成本地的同步任务后，再将本地的同步数据接收池对应的同步请求字段减1（available -1）。

图2为本申请实施例提供的一种集群内存分布示意图。

在上述实施例中提到，通过设计各节点上同步数据接收池与其他节点一一对应以便于对来自不同源节点的待同步数据的管理。除此以外，还可以针对不同类型的待同步数据设置不同的同步规则，并将不同类型的待同步数据写入不同的存储区域来便于管理。在本申请实施例提供的集群数据同步方法中，同步数据接收池中设有多类与不同类型的待同步数据对应的子接收池。

相应的，S103中主节点在本地内存中执行对待同步数据的同步任务，可以为：按照待同步数据的类型，在本地内存中执行对待同步数据的同步任务。

S103中从节点执行对待同步数据的同步任务，可以为：从节点按照待同步数据的类型执行对待同步数据的同步任务。

在可选实施例中，不论是源端是源节点、目的端是主节点，还是源端是主节点、目的端是从节点，在源端根据本地存储的分布式集群中其他节点中源节点与同步数据接收池的对应关系的基础上，增加对不同类型的待同步数据的处理逻辑，以在源端写入以及目的端读取完成同步任务的过程中均按照待同步数据的类型执行对应的操作。

如可以按照待同步数据的数据量大小和同步频繁程度设计子接收池的类型。如图2所示，子接收池的类型可以包括：小数据接收池（small pool）、应答数据接收池（res pool）和大数据接收池（big pool）。

其中，小数据接收池用于接收短数据包，一般用于更新状态等数据交互量少，操作频繁，且CPU可以固定时间内完成数据处理操作的数据。

应答数据接收池用于接收确认数据和更新集群状态，在主节点状态下负责处理复杂逻辑的响应确认包，例如更新心跳信息。

大数据接收池用于处理大更新量数据，在更新时可能会直接覆盖远端节点的对应内存页，如快照（snapshot，Raft协议概念，指新节点加入时需要将集群状态全部更新至本地节点，操作过程可以近似于memcpy）更新过程，以及处理集群事务对应的过程。

同时，还可以设计同步数据接收池中预先按照各类型待同步数据的数据长度划分为各类型子接收池的数据分区。每个数据分区对应一段连续的内存地址。在写入待同步数据时，将待同步数据写入对应类型的空白数据分区。

基于本申请实施例提供的内存架构设计,结合本申请实施例提供的任务完成回调机制,在各节点的同步数据接收池中,可以设计同步请求字段和与其他节点对应的同步数据接收池中的子接收池的类型一一对应,源端在将待同步数据写入目的端之后,调用任务完成回调函数,根据待同步数据的类型将目的端与源端对应的同步数据接收池中对应数据类型的子接收池的同步请求字段进行更新,以便目的端的CPU在查看到同步请求字段更新后能够直接确定待同步数据的类型。以四节点集群中的主节点为例,主节点设有与其他三个从节点对应的同步数据接收池,每个同步数据接收池中又分为小数据接收池、应答数据接收池和大数据接收池三类子接收池,则一个同步数据接收池对应三个同步请求字段,则主节点中共维护九个同步请求字段。

结合上述设计,设计分布式集群中每个节点的本地内存中设有与其他节点一一对应的同步数据接收池,在每个同步数据接收池内又划分为小数据接收池、应答数据接收池和大数据接收池三类子接收池,且每类子接收池均包含多个该类子接收池的数据分区,实现对同步数据接收池的存储管理。

如图2所示,在一个四节点的分布式集群中,对于任意节点201(例如节点1),其本地内存的内存空间划分为集群内存空间和与其余节点201(节点2、节点3、节点4)对应的同步数据接收池,在节点1中,同步数据接收池与节点2、节点3、节点4一一对应,每个同步数据接收池都由RDMA映射数组和原子索引构成,每个同步数据接收池中又划分为多个小数据接收池、多个应答数据接收池和多个大数据接收池,用于接收不同类型的待同步数据的写入以及在目的端上采用对应的处理方式执行同步任务。其中,节点1、节点2、节点3、节点4为各节点在分布式集群中的序号,而在进行集群数据同步时,在Raft协议网络中,若选举节点1为主节点,则节点2、节点3、节点4可以依次对应为从节点1、从节点2、从节点3。当其中任意一个从节点产生待同步数据时,则该从节点即为源节点,并进入源节点将待同步数据发送至主节点、由主节点将待同步数据广播至各从节点以完成集群数据同步任务的流程。

在上述实施例的基础上,本申请实施例进一步对不同类型的待同步数据的同步方法进行说明。

在本申请实施例提供的集群数据同步方法中,当待同步数据为与小数据接收池对应的数据类型时,

S102中源节点基于RDMA网络调用通信设备写入待同步数据,可以为:源节点将待同步数据组装为第一消息头后,根据本地存储的主节点中同步数据接收池的存储区域信息,基于RDMA网络调用通信设备将第一消息头写入对应的存储位置;

S103中主节点基于RDMA网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池,可以为:根据本地存储的从节点中同步数据接收池的存储区域信息,基于RDMA网络调用通信设备将第一消息头写入对应的存储位置。

在可选实施中,当待同步数据为与小数据接收池对应的数据类型时,本申请实施例提供一种集群数据同步过程如下:

源节点(例如上述实施例中举例的从节点1)获取本地集群内存空间,根据写入请求组织第一消息头(如Raft协议数据、对较复杂数据的序列化处理等);

源节点根据本地存储的主节点中源节点与同步数据接收池的对应关系,确定自己在主节点的本地内存中对应的同步数据接收池。例如,从节点1为图2所示的节点2,则从节点1根据主节点的内存区域(memory region)信息,将第一消息头写入主节点的内存中与节点2对应的同步数据接收池的小数据接收池中(如node 2 small pool [index =0]),并更新主节点与节点2对应的同步数据接收池的小数据接收池可用消息数加1;

在源节点的写入过程中,源节点调用写(write)接口,直接通过通信设备(如RoCE网卡)将第一消息头直接写入主节点的对应内存区域中,而非调用类似TCP协议的发送/接收(send/recv)功能,避免在发送/接收(send/recv)中的内存拷贝,主节点也无需维护IO复用数据,显著减少了代码量与主节点的CPU调度;

源节点在完成写入后,通过RDMA技术特有的任务完成回调函数(completion IO),在主节点更新同步

请求字段 (available)，表示有数据发送，需要等待主节点处理；在这个过程中，使用RDMA无损数据网络，则可以省略源节点检查发送失败的流程，同样减少了代码量；

在主节点中，通过对同步请求字段的检查线程遍历本地与其余各节点对应的同步数据接收池时，发现与源节点对应的同步数据接收池中有一个待处理请求 (available=1)，则进行信息校验；当对第一消息头校验合法后，更新主节点集群内存空间中的对应数据，并将第一消息头广播至所有从节点；同时，由于采用无损数据网络，在主节点广播完成后，将Raft协议中定义的集群间消息编号logentry加1；

主节点广播至所有从节点的过程与源节点将待同步数据写入主节点的过程类似；

在各从节点接收到待同步数据时，各自更新本地集群内存空间的数据，无需返回确认码。

当待同步数据为与应答数据接收池对应的数据类型时，与小数据接收池类似，即同样是通过任务完成回调函数处理返回值并更新同步请求字段。则在本申请实施例提供的集群数据同步方法中，当待同步数据为与应答数据接收池对应的数据类型时，

S102中源节点基于RDMA网络调用通信设备写入待同步数据，可以为：源节点将待同步数据组装为第二消息头后，根据本地存储的主节点中同步数据接收池的存储区域信息，基于RDMA网络调用通信设备将第二消息头写入对应的存储位置；

S103中基于RDMA网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，可以为：根据本地存储的从节点中同步数据接收池的存储区域信息，基于RDMA网络调用通信设备将第二消息头写入对应的存储位置。

应答数据接收池的写入方法可以参考小数据接收池的写入方法。不同的是，针对不同类型的应答数据，可以设置额外的同步规则。如在应答数据接收池中可以包括各节点的心跳信号，以完成分布式集群中各节点的状态监控。例如，可以设计应答数据接收池的第一预设空间 (如res pool[0]) 用于存放分布式集群中所有节点的心跳信息。可选地，各节点之间可以定时通过本申请实施例提供的集群数据同步方法进行心跳信息的同步。

则在本申请实施例中，本申请实施例提供的集群数据同步方法还可以包括：

定时向本地内存的第一预设空间写入当前时间戳；

通过RDMA网络将当前时间戳同步至各从节点的第一预设空间；

读取各从节点写入本地内存的响应时间戳；

根据各响应时间戳检查各从节点的连接状态。

在可选实施中，主节点会调度一个线程 (如定义为心跳线程) 每隔预设时间向特定RDMA内存写入当前时间戳，则当前时间戳会通过RDMA无损网络写入所有从节点的应答数据接收池的第一预设空间，各从节点再通过RDMA无损网络将响应时间戳写入主节点的应答数据接收池的第一预设空间，主节点检查所有从节点的响应时间戳即可。

对比传统阻塞方式发送/接收 (send/recv) 写入，本申请实施例提供的心跳包处理方式对于上层应用来说类似于完全读写本地内存，整个过程是非阻塞的。通常来说，这意味着更快的错误检测。发送/接收 (send/recv) 在集群状态发生错误的时候，系统底层发送 (send) 过程会阻塞，携带消息头的信息包与心跳信息混杂在一起，在超时发生后一起向上报告。对主节点来说，这意味着突发性的错误状态需要恢复，逻辑复杂后耗时较长。而通过非阻塞的写 (write) 操作，主节点可以在其他消息发生错误前就检测到心跳包超时，于是直接调度重构集群操作，对于集群恢复时间有相当的帮助。

应答数据接收池除了用于存储确认数据和更新集群状态数据外，还可以设置有用于存放用户自定义数据的第二预设空间 (如res pool[1]) 用于存放用户自定义数据，即用于用户扩展的待同步数据类型。

在上述实施例的基础上，本申请实施例进一步对不同类型的待同步数据的同步方法进行说明。

在本申请实施例提供的集群数据同步方法中，当待同步数据为与大数据接收池对应的数据类型时，

S102中源节点基于RDMA网络调用通信设备写入待同步数据，可以为：

源节点将待同步数据的同步任务组装为第三消息头后，根据本地存储的主节点中同步数据接收池的存储区域信息，基于RDMA网络调用通信设备将第三消息头写入对应的存储位置；

S103中在本地内存中执行对待同步数据的同步任务，并以广播的形式，基于RDMA网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，包括：

根据本地存储的从节点中同步数据接收池的存储区域信息，以广播的形式，基于RDMA网络调用通信设备将第三消息头写入对应的存储位置，以使从节点在从节点的大数据接收池的第三预设空间的分页内执行第三消息头所包含的逻辑；

在本地的大数据接收池的第三预设空间的分页内执行第三消息头所包含的逻辑；

在确认本地以及各从节点均完成执行第三消息头所包含的逻辑之后，控制本地以及各从节点将所在节点的第三预设空间的分页拷贝至所在节点的集群内存空间。

针对大更新量数据，各节点在执行同步任务时往往需要耗费较长时间，将对分布式集群提供的业务服务造成较长时间的中断。故针对大更新量数据，本申请实施例提供集群数据同步方法设计大数据接收池来接收待同步数据，并在更新过程中，先接收待同步数据，并在大数据接收池中以待同步数据在集群内存空间中对应数据页为单位进行更新后，再将大数据接收池中更新后的数据页替换集群内存空间中对应的数据页，从而在更新过程中缩短对业务服务的中断。

在可选实施中，本申请实施例提供一种集群数据同步过程如下：

源节点根据本地存储的主节点中源节点与同步数据接收池的对应关系，确定自己在主节点的本地内存中对应的同步数据接收池后，源节点发起对主节点的大数据接收池的写入请求，并组装的第三消息头，将第三消息头写入主节点中与源节点对应的同步数据接收池（例如上述实施例中举例的四节点集群中，节点2/从节点1对应的node 2 big pool [index =0]）中，并更新主节点中该同步数据接收池的可用消息数加1；

源节点将第三消息头写入主节点的过程与对小数据接收池的写入同理；

主节点通过对同步请求字段的检查线程遍历本地与其余各节点对应的同步数据接收池，发现与源节点对应的同步数据接收池中有一个待处理请求（available=1），则进行信息校验；当对第三消息头校验合法后，以广播形式基于RDMA网络调用通信设备将第三消息头写入各从节点对应主节点设置的同步数据接收池中的大数据接收池，同时在本节点开始对第三消息头中包含的同步任务的逻辑进行处理；

各从节点在接收到主节点发送的第三消息头后，分别在本地开始对第三消息头中包含的同步任务的逻辑进行处理；

通常来说，设计由大数据接收池接收的待同步数据均涉及到较为复杂的逻辑，处理相对更加耗时；在这个过程中，在大数据接收池中，按照第三消息头涉及到的数据，优先在大数据接收池的第三预设空间的分页内（例如预先已经将同步数据接收池中划分了多个大数据接收池的数据分区，则在写入第三消息头的数据分区内）处理第三消息头的同步任务的相关逻辑，并在执行完毕后，向主节点发送确认成功响应码；

确认成功响应码属于一种确认数据，故各从节点向主节点发送确认成功响应码，可以为各从节点基于RDMA网络调用通信设备将确认成功响应码写入主节点中与各从节点对应的同步数据接收池中的应答数据接收池；

主节点检测到本地同步数据接收池中的应答数据接收池中包含所有从节点的确认成功响应码，并确认本地已经执行完毕第三消息头的同步任务的相关逻辑，则将确认执行消息广播至所有从节点；

各从节点收到确认执行消息后，将第三预设空间内的分页拷贝至所在节点的集群内存空间中的对应分页，完成本地的同步任务；

各节点均完成分页替换后，完成本次分布式集群中的集群数据同步任务，各节点将本地的大数据接收池对应的同步请求字段减1（big pool available -1）。

由于在进行分页替换时，也会对业务服务造成中断，为进一步减少对业务服务的影响，在本申请实施例提供的集群数据同步方法中，在确认本地以及各从节点均完成执行第三消息头所包含的逻辑之后，控制本地以及各从节点将所在节点的第三预设空间的分页拷贝至所在节点的集群内存空间，可以为：

在确认本地以及各从节点均完成执行第三消息头所包含的逻辑之后，向各从节点发送确认成功响应码，以使各从节点在接收到确认成功响应码之后与主节点各自执行下述步骤：

将第三预设空间对应的集群内存空间中的第四预设空间的业务逻辑地址交换至第三预设空间，以基于第三预设空间的数据执行下一个处理逻辑，同时将第三预设空间的数据拷贝至第四预设空间；

在拷贝完成后，将业务逻辑地址交换回第四预设空间。

在可选实施中，各从节点在接收到主节点发送的确认执行消息后进行第三预设空间的分页与集群内存空间内对应分页的替换时，以及主节点在本地进行第三预设空间的分页与集群内存空间内对应分页的替换时，均通过原子（atomic）逻辑将集群内存空间中的对应地址（假设对应第四预设空间）交换至第三预设空间，以基于第三预设空间处理集群内存空间中该部分分页提供的服务，执行下一个处理逻辑；与此同时，各节点将第三预设空间的分页拷贝至第四预设空间；在拷贝完成后，短暂中断该部分数据提供的业务服务，再利用（atomic）逻辑交换回原地址后，恢复业务服务。

以集群内存空间有四个分页为例，假设第三预设空间为同步数据接收池中的大数据接收池中的big pool[0]分页，第四预设空间为集群内存空间的分页2，则在big pool[0]分页中执行同步任务后，交换big pool[0]分页的指针与分页2的指针，即以big pool[0]分页提供原本由分页2提供的服务，并将big pool[0]分页的数据拷贝至分页2，在拷贝完成后，再将指针交换回来，而后释放big pool[0]分页的空间。

本申请上述实施例对分布式集群中大数据的同步过程进行了说明。目前在业界中，Raft协议没有与事务结合起来，对于某些节点数据执行失败只能重置节点状态，这同样影响分布式集群的集群数据同步效率和同步失败后的数据恢复效率。对于集群事务来说，要求全部节点保持数据一致性，但对于Raft协议来说，数据只需要传输到过半节点即可保证集群数据唯一性，而未执行成功的少数节点会直接退出集群，重新刷新日志信息。即是说，对同步任务执行失败的节点需要清除所有数据，再从主节点重新获取。

而对大更新量数据的处理相较于小数据的同步会更加耗时和复杂，各节点在执行同步任务的过程中更容易执行失败，则采用现有的Raft协议的同步办法即会出现同步失败后集群恢复时间较长的问题。

事务（Transaction）是一种机制、一个操作序列，包含了一组操作命令。在对于事务数据进行同步时之所以会出现上述问题，是事务把所有的命令作为一个整体一起向系统提交或撤销操作请求，即这一组数据库命令要么都执行，要么都不执行，因此事务是一个不可分割的工作逻辑单元。

故在上述实施例的基础上，在本申请实施例提供的集群数据同步方法中，当待同步数据为与大数据接收池对应的数据类型且为非事务数据时，第三消息头仅包含待同步数据的同步逻辑；

当待同步数据为与大数据接收池对应的数据类型且为事务数据时，第三消息头包含待同步数据的同步逻辑以及有待同步数据对应的事务回调函数，以使主节点或从节点在执行对待同步数据的同步任务时执行事务回调函数对应的逻辑；

其中，事务回调函数至少包括针对事务数据同步的数据备份函数以及关键数据保护函数。

需要说明的是，在本申请实施例中，非事务数据指需要同步的数据量较大、较为集中，但允许部分数据同步失败的待同步数据，例如针对部分节点上区分与其他节点的独有配置的数据同步，则无需使其他所有节点均对这一部分数据进行同步，也能视为完成了分布式集群的集群同步任务。事务数据则指上文提到的事务过程的待同步数据，在执行过程中只能全部执行成功，若其中部分数据执行失败，则整个待同步数据的同步任务均执行失败。

针对非事务数据的同步过程的说明请参考上述实施例，在此不再赘述。

而针对事务数据的同步过程，在进行待同步数据的传输、在各节点上执行同步任务的相关逻辑以及分页替换的步骤与非事务数据的步骤，但要求事务数据的第三消息头中提供对应功能的事务回调函数。

可选地，事务过程均有唯一对应的标识（ID），预先针对分布式集群中的事务过程设计事务回调函数，以使各节点在发起对事务数据的同步时在注册的第三消息头中提供对应功能的事务回调函数。要提供的事务回调函数至少包括各节点在执行同步任务之前进行数据备份，并对关键数据进行保护，如加锁以暂停对外的业务服务，避免在同步过程中产生错误。这些准备工作是为了在对事务数据同步失败后可以回到同步前的状态，而不是像相关技术中那样一旦同步失败就要将节点退出集群再重新加入。其中，关键数据可以为分布式集群所面向的用户的用户数据，在进行事务性数据的同步时，需要对用户数据进行保护。例如，用户对分布式集群的命名需要同步至分布式集群中的各个节点，而当用户修改集群名称时，就需要对所有节点配置的集群名称先进行加锁保护，避免在同步过程中其他线程访问到集群名称为不确定的集群名称（有的访问到修改前的集群名称，有的访问到修改后的集群名称）。通过对关键数据进行保护，避免在分布式集群数据同步过程中接收业务访问时数据错乱。

在此基础上，本申请实施例提供一组事务回调函数的类型如下：

事务检查函数（`evt_transaction_check`），用于检查该数据执行事务的前提条件是否合理；

数据备份函数（`evt_transaction_prepare`），用于在准备执行事务时，备份原有数据，关键数据保护等流程；

以上两个操作成功或失败时，从节点会向主节点返回逻辑成功消息；

事务执行函数（`evt_transaction_commit`），用于执行事务的实际逻辑，该函数回调执行后，事务数据应当准备完毕，并且可以正常读取；在待同步数据的同步任务执行完毕后，所有从节点会向主节点返回同步任务执行完毕消息；若主节点检查所有从节点均成功执行完了待同步数据的同步任务，则对所有节点的调度事务完成函数（`evt_transaction_finish`）；否则，调度事务修复函数（`evt_transaction_restore`）；

事务完成函数（`evt_transaction_finish`），用于在待同步数据的同步任务执行成功后回调，释放关键数据保护和删除备份操作等过程；

事务修复函数（`evt_transaction_restore`），用于在意外执行发生错误后回调，释放数据保护，恢复数据至原始状态。

则基于上述事务回调函数，本申请实施例提供一种执行事务回调函数对应的逻辑包括：

检查待同步数据的同步任务的合理性；

执行数据备份及关键数据保护流程；

若待同步数据的同步任务合理且成功完成数据备份及关键数据保护流程，则向主节点返回准备逻辑执行成功消息；

在准备逻辑执行成功后，触发执行待同步数据的同步任务的逻辑；

在执行完毕待同步数据的同步任务的逻辑后，向主节点返回同步任务执行完毕消息；

若主节点及各从节点均成功执行待同步数据的同步任务的逻辑，则释放关键数据保护，删除数据备份；否则，执行发生错误后回调操作，释放关键数据保护并基于数据备份恢复集群数据至初始状态。

上文详述了集群数据同步方法对应的各个实施例，在此基础上，本申请还公开了与上述方法对应的集群数据同步装置、设备及非易失性可读存储介质。

图3为本申请实施例提供的一种集群数据同步装置的结构示意图。

如图3所示，本申请实施例提供的集群数据同步装置包括：

监测单元301被设置为监测本地内存中预设的同步数据接收池；

校验单元302被设置为当监测到本地内存中同步数据接收池被源节点基于RDMA网络调用通信设备写入的待同步数据时，对待同步数据进行校验；

同步单元303被设置为当待同步数据通过校验后，在本地内存中执行对待同步数据的同步任务，并以广播的形式，基于RDMA网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使

从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务；

其中，源节点为分布式集群中的一个从节点。

可选的，监测单元301监测到本地内存中同步数据接收池被源节点基于RDMA网络调用通信设备写入的待同步数据，可以为：

监测到源节点在写完待同步数据后基于RDMA网络的任务完成回调函数在主节点的CPU中更新的同步请求字段后，确认源节点写入了待同步数据；

同步单元303基于RDMA网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，可以为：

基于RDMA网络调用通信设备将待同步数据写入从节点的同步数据接收池后，基于任务完成回调函数在从节点的CPU中更新同步请求字段，以使从节点在监测到更新的同步请求字段后完成对待同步数据的校验并执行对待同步数据的同步任务。

可选的，同步数据接收池与分布式集群中除所在节点外的其余节点一一对应；

同步数据接收池设有用于维护所在节点外的其余节点与同步数据接收池的对应关系的RDMA映射数组以及用于为待同步数据分配空余空间的原子索引。

可选的，同步数据接收池中设有多类与不同类型的待同步数据对应的子接收池；

相应的，在本地内存中执行对待同步数据的同步任务，可以为：

按照待同步数据的类型，在本地内存中执行对待同步数据的同步任务；

从节点执行对待同步数据的同步任务，可以为：

从节点按照待同步数据的类型执行对待同步数据的同步任务。

可选的，同步数据接收池中预先按照各类型待同步数据的数据长度划分为各类型子接收池的数据分区。

可选的，子接收池的类型包括：小数据接收池、应答数据接收池和大数据接收池。

可选的，当待同步数据为与小数据接收池对应的数据类型或应答数据接收池对应的数据类型时，

源节点基于RDMA网络调用通信设备写入待同步数据，可以为：

源节点将待同步数据组装为第一消息头后，根据本地存储的主节点中同步数据接收池的存储区域信息，基于RDMA网络调用通信设备将第一消息头写入对应的存储位置；

同步单元303基于RDMA网络调用通信设备将待同步数据写入分布式集群的各从节点的同步数据接收池，可以为：

根据本地存储的从节点中同步数据接收池的存储区域信息，基于RDMA网络调用通信设备将第一消息头写入对应的存储位置。

可选的，应答数据接收池的第一预设空间用于存放分布式集群中所有节点的心跳信息。

可选的，本申请实施例提供的集群数据同步装置还包括：

时间戳写入单元，被设置为定时向本地内存的第一预设空间写入当前时间戳；

时间戳同步单元，被设置为通过RDMA网络将当前时间戳同步至各从节点的第一预设空间；

时间戳读取单元，被设置为读取各从节点写入本地内存的响应时间戳；

状态检查单元，被设置为根据各响应时间戳检查各从节点的连接状态。

可选的，应答数据接收池的第二预设空间用于存放用户自定义数据。

可选的，当待同步数据为与大数据接收池对应的数据类型时，

源节点基于RDMA网络调用通信设备写入待同步数据，可以为：

源节点将待同步数据的同步任务组装为第三消息头后，根据本地存储的主节点中同步数据接收池的存储区域信息，基于RDMA网络调用通信设备将第三消息头写入对应的存储位置；

同步单元303在本地内存中执行对待同步数据的同步任务，并以广播的形式，基于RDMA网络调用通信设

备将待同步数据写入分布式集群的各从节点的同步数据接收池，以使从节点在监测到待同步数据并完成对待同步数据的校验后执行对待同步数据的同步任务，包括：

根据本地存储的从节点中同步数据接收池的存储区域信息，以广播的形式，基于RDMA网络调用通信设备将第三消息头写入对应的存储位置，以使从节点在从节点的大数据接收池的第三预设空间的分页内执行第三消息头所包含的逻辑；

在本地的大数据接收池的第三预设空间的分页内执行第三消息头所包含的逻辑；

在确认本地以及各从节点均完成执行第三消息头所包含的逻辑之后，控制本地以及各从节点将所在节点的第三预设空间的分页拷贝至所在节点的集群内存空间。

可选的，同步单元303在确认本地以及各从节点均完成执行第三消息头所包含的逻辑之后，控制本地以及各从节点将所在节点的第三预设空间的分页拷贝至所在节点的集群内存空间，可以为：

在确认本地以及各从节点均完成执行第三消息头所包含的逻辑之后，向各从节点发送确认成功响应码，以使各从节点在接收到确认成功响应码之后与主节点各自执行下述步骤：

将第三预设空间对应的集群内存空间中的第四预设空间的业务逻辑地址交换至第三预设空间，以基于第三预设空间的数据执行下一个处理逻辑，同时将第三预设空间的数据拷贝至第四预设空间；

在拷贝完成后，将业务逻辑地址交换回第四预设空间。

可选的，当待同步数据为与大数据接收池对应的数据类型且为非事务数据时，第三消息头仅包含待同步数据的同步逻辑；

当待同步数据为与大数据接收池对应的数据类型且为事务数据时，第三消息头包含待同步数据的同步逻辑以及有待同步数据对应的事务回调函数，以使主节点或从节点在执行对待同步数据的同步任务时执行事务回调函数对应的逻辑；

其中，事务回调函数至少包括针对事务数据同步的数据备份函数以及关键数据保护函数。

进一步，同步单元303执行事务回调函数对应的逻辑包括：

检查待同步数据的同步任务的合理性；

执行数据备份及关键数据保护流程；

若待同步数据的同步任务合理且成功完成数据备份及关键数据保护流程，则向主节点返回准备逻辑执行成功消息；

在准备逻辑执行成功后，触发执行待同步数据的同步任务的逻辑；

在执行完毕待同步数据的同步任务的逻辑后，向主节点返回同步任务执行完毕消息；

若主节点及各从节点均成功执行待同步数据的同步任务的逻辑，则释放关键数据保护，删除数据备份；否则，执行发生错误后回调操作，释放关键数据保护并基于数据备份恢复集群数据至初始状态。

可选的，通信设备可以为RoCE网卡。

由于装置部分的实施例与方法部分的实施例相互对应，因此装置部分的实施例请参见方法部分的实施例的描述，这里暂不赘述。

图4为本申请实施例提供的一种集群数据同步设备的结构示意图。

如图4所示，本申请实施例提供的集群数据同步设备包括：

存储器410被设置为存储计算机程序411；

处理器420被设置为执行计算机程序411，该计算机程序411被处理器420执行时实现如上述任意一项实施例集群数据同步方法的步骤。

其中，处理器420可以包括一个或多个处理核心，比如3核心处理器、8核心处理器等。处理器420可以采用数字信号处理DSP（Digital Signal Processing）、现场可编程门阵列FPGA（Field-Programmable Gate Array）、可编程逻辑阵列PLA（Programmable Logic Array）中的至少一种硬件形式来实现。处理器420也可以包括主处理器和协处理器，主处理器是被设置为对在唤醒状态下的数据进行处理的处理单元，也称

中央处理器CPU (Central Processing Unit)；协处理器是被设置为对在待机状态下的数据进行处理的低功耗处理器。在一些实施例中，处理器420可以集成有图像处理器GPU (Graphics Processing Unit)，GPU被设置为负责显示屏所需要显示的内容的渲染和绘制。一些实施例中，处理器420还可以包括人工智能AI (Artificial Intelligence) 处理器，该AI处理器被设置为处理有关机器学习的计算操作。

存储器410可以包括一个或多个非易失性可读存储介质，该非易失性可读存储介质可以是非暂态的。存储器410还可包括高速随机存取存储器，以及非易失性存储器，比如一个或多个磁盘存储设备、闪存存储设备。本实施例中，存储器410至少被设置为存储以下计算机程序411，其中，该计算机程序411被处理器420加载并执行之后，能够实现前述任一实施例公开的集群数据同步方法中的相关步骤。另外，存储器410所存储的资源还可以包括操作系统412和数据413等，存储方式可以是短暂存储或者永久存储。其中，操作系统412可以为Windows。数据413可以包括但不限于上述方法所涉及到的数据。

在一些实施例中，集群数据同步设备还可包括有显示屏430、电源440、通信接口450、输入输出接口460、传感器470以及通信总线480。

本领域技术人员可以理解，图4中示出的结构并不构成对集群数据同步设备的限定，可以包括比图示更多或更少的组件。

本申请实施例提供的集群数据同步设备，包括存储器和处理器，处理器在执行存储器存储的程序时，能够实现如上的集群数据同步方法，效果同上。

需要说明的是，以上所描述的装置、设备实施例仅仅是示意性的，例如，模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个模块或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或模块的间接耦合或通信连接，可以是电性，机械或其它的形式。作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理模块，即可以位于一个地方，或者也可以分布到多个网络模块上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能模块可以集成在一个处理模块中，也可以是各个模块单独物理存在，也可以两个或两个以上模块集成在一个模块中。上述集成的模块既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。

集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，可以存储在一个非易失性可读存储介质中。基于这样的理解，本申请的技术方案本质上或者说对相关技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，执行本申请各个实施例方法的全部或部分步骤。

为此，本申请实施例还提供一种非易失性可读存储介质，该非易失性可读存储介质上存储有计算机程序，计算机程序被处理器执行时实现如集群数据同步方法的步骤。

该非易失性可读存储介质可以包括：U盘、移动硬盘、只读存储器ROM (Read-Only Memory)、随机存取存储器RAM (Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

本实施例中提供的非易失性可读存储介质所包含的计算机程序能够在被处理器执行时实现如上的集群数据同步方法的步骤，效果同上。

以上对本申请所提供的一种集群数据同步方法、装置、设备及非易失性可读存储介质进行了详细介绍。说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。对于实施例公开的装置、设备及非易失性可读存储介质而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以对本申请进行若干改进和修饰，这些改进和修饰也落入本申请权利要求的保护范围内。

还需要说明的是，在本说明书中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括上述要素的过程、方法、物品或者设备中还存在另外的相同要素。

## 权 利 要 求 书

1. 一种集群数据同步方法，其特征在于，基于分布式集群中的主节点，包括：

监测本地内存中预设的同步数据接收池；

当监测到本地内存中所述同步数据接收池被源节点基于RDMA网络调用通信设备写入的待同步数据时，对所述待同步数据进行校验；

当所述待同步数据通过校验后，在本地内存中执行对所述待同步数据的同步任务，并以广播的形式，基于所述RDMA网络调用通信设备将所述待同步数据写入所述分布式集群的各从节点的所述同步数据接收池，以使所述从节点在监测到所述待同步数据并完成对所述待同步数据的校验后执行对所述待同步数据的同步任务；

其中，所述源节点为所述分布式集群中的一个所述从节点。

2. 根据权利要求1所述的集群数据同步方法，其特征在于，监测到本地内存中所述同步数据接收池被所述源节点基于RDMA网络调用通信设备写入的所述待同步数据，包括：

监测到所述源节点在写完所述待同步数据后基于所述RDMA网络的任务完成回调函数在所述主节点的CPU中更新的同步请求字段时，确认所述源节点写入了所述待同步数据；

所述基于所述RDMA网络调用通信设备将所述待同步数据写入所述分布式集群的各从节点的所述同步数据接收池，以使所述从节点在监测到所述待同步数据并完成对所述待同步数据的校验后执行对所述待同步数据的同步任务，包括：

基于所述RDMA网络调用通信设备将所述待同步数据写入所述从节点的所述同步数据接收池后，基于所述任务完成回调函数在所述从节点的CPU中更新所述同步请求字段，以使所述从节点在监测到更新的所述同步请求字段后完成对所述待同步数据的校验并执行对所述待同步数据的同步任务。

3. 根据权利要求1所述的集群数据同步方法，其特征在于，监测到本地内存中所述同步数据接收池被所述源节点基于RDMA网络调用通信设备写入的所述待同步数据，包括：

当轮询监测到本地内存中所述同步数据接收池有新增的所述待同步数据时，确认所述源节点写入了所述待同步数据；

所述基于所述RDMA网络调用通信设备将所述待同步数据写入所述分布式集群的各从节点的所述同步数据接收池，以使所述从节点在监测到所述待同步数据并完成对所述待同步数据的校验后执行对所述待同步数据的同步任务，包括：

基于所述RDMA网络调用通信设备将所述待同步数据写入所述从节点的所述同步数据接收池，以使所述从节点在轮询监测到所述待同步数据后，完成对所述待同步数据的校验并执行对所述待同步数据的同步任务。

4. 根据权利要求1所述的集群数据同步方法，其特征在于，监测到本地内存中所述同步数据接收池被所述源节点基于RDMA网络调用通信设备写入的所述待同步数据，包括：

接收到所述源节点在写完所述待同步数据后调用输入输出端口发送的写完成标志时，确认所述源节点写入了所述待同步数据；

所述基于所述RDMA网络调用通信设备将所述待同步数据写入所述分布式集群的各从节点的所述同步数据接收池，以使所述从节点在监测到所述待同步数据并完成对所述待同步数据的校验后执行对所述待同步数据的同步任务，包括：

基于所述RDMA网络调用通信设备将所述待同步数据写入所述从节点的所述同步数据接收池后，调用输入输出端口向所述从节点发送写完成标志，以使所述从节点确认被写入所述待同步数据后，完成对所述待同步数据的校验并执行对所述待同步数据的同步任务。

5. 根据权利要求1所述的集群数据同步方法，其特征在于，监测到本地内存中所述同步数据接收池被所述源节点基于RDMA网络调用通信设备写入的所述待同步数据，包括：

监测到所述源节点在写完所述待同步数据后在所述主节点的写完成标志约定地址写入的写完成标志时，确认所述源节点写入了所述待同步数据；

所述基于所述RDMA网络调用通信设备将所述待同步数据写入所述分布式集群的各从节点的所述同步数据接收池，以使所述从节点在监测到所述待同步数据并完成对所述待同步数据的校验后执行对所述待同步数据的同步任务，包括：

基于所述RDMA网络调用通信设备将所述待同步数据写入所述从节点的所述同步数据接收池后，在各所述从节点的写完成标志约定地址写入写完成标志，以使所述从节点确认被写入所述待同步数据后，完成对所述待同步数据的校验并执行对所述待同步数据的同步任务。

6. 根据权利要求1所述的集群数据同步方法，其特征在于，所述同步数据接收池与所述分布式集群中除所在节点外的其余节点一一对应。

7. 根据权利要求1所述的集群数据同步方法，其特征在于，所述同步数据接收池设有用于维护所在节点外的其余节点与所述同步数据接收池的对应关系的RDMA映射数组以及用于为所述待同步数据分配空余空间的原子索引。

8. 根据权利要求1所述的集群数据同步方法，其特征在于，所述同步数据接收池中设有多类与不同类型的所述待同步数据对应的子接收池；

相应的，所述在本地内存中执行对所述待同步数据的同步任务，包括：

按照所述待同步数据的类型，在本地内存中执行对所述待同步数据的同步任务；

所述从节点执行对所述待同步数据的同步任务，包括：

所述从节点按照所述待同步数据的类型执行对所述待同步数据的同步任务。

9. 根据权利要求8所述的集群数据同步方法，其特征在于，所述同步数据接收池中预先按照各类型所述待同步数据的数据长度划分为各类型所述子接收池的数据分区。

10. 根据权利要求8所述的集群数据同步方法，其特征在于，所述子接收池的类型包括：小数据接收池、应答数据接收池和大数据接收池。

11. 根据权利要求10所述的集群数据同步方法，其特征在于，当所述待同步数据为与所述小数据接收池对应的数据类型时，

所述源节点基于所述RDMA网络调用通信设备写入所述待同步数据，包括：

所述源节点将所述待同步数据组装为第一消息头后，根据本地存储的所述主节点中所述同步数据接收池的存储区域信息，基于所述RDMA网络调用通信设备将所述第一消息头写入对应的存储位置；

所述基于所述RDMA网络调用通信设备将所述待同步数据写入所述分布式集群的各从节点的所述同步数据接收池，包括：

根据本地存储的所述从节点中所述同步数据接收池的存储区域信息，基于所述RDMA网络调用通信设备将所述第一消息头写入对应的存储位置。

12. 根据权利要求10所述的集群数据同步方法，其特征在于，当所述待同步数据为与所述应答数据接收池对应的数据类型时，

所述源节点基于所述RDMA网络调用通信设备写入所述待同步数据，包括：

所述源节点将所述待同步数据组装为第二消息头后，根据本地存储的所述主节点中所述同步数据接收池的存储区域信息，基于所述RDMA网络调用通信设备将所述第二消息头写入对应的存储位置；

所述基于所述RDMA网络调用通信设备将所述待同步数据写入所述分布式集群的各从节点的所述同步数据接收池，包括：

根据本地存储的所述从节点中所述同步数据接收池的存储区域信息，基于所述RDMA网络调用通信设备将所述第二消息头写入对应的存储位置。

13. 根据权利要求10所述的集群数据同步方法，其特征在于，所述应答数据接收池的第一预设空间用

于存放所述分布式集群中所有节点的心跳信息。

14. 根据权利要求13所述的集群数据同步方法，其特征在于，还包括：

定时向本地内存的所述第一预设空间写入当前时间戳；

通过所述RDMA网络将所述当前时间戳同步至各所述从节点的所述第一预设空间；

读取各所述从节点写入本地内存的响应时间戳；

根据各所述响应时间戳检查各所述从节点的连接状态。

15. 根据权利要求10所述的集群数据同步方法，其特征在于，所述应答数据接收池的第二预设空间用于存放用户自定义数据。

16. 根据权利要求10所述的集群数据同步方法，其特征在于，当所述待同步数据为与所述大数据接收池对应的数据类型时，

所述源节点基于所述RDMA网络调用通信设备写入所述待同步数据，包括：

所述源节点将所述待同步数据的同步任务组装为第三消息头后，根据本地存储的所述主节点中所述同步数据接收池的存储区域信息，基于所述RDMA网络调用通信设备将所述第三消息头写入对应的存储位置；

所述在本地内存中执行对所述待同步数据的同步任务，并以广播的形式，基于所述RDMA网络调用通信设备将所述待同步数据写入所述分布式集群的各从节点的所述同步数据接收池，以使所述从节点在监测到所述待同步数据并完成对所述待同步数据的校验后执行对所述待同步数据的同步任务，包括：

根据本地存储的所述从节点中所述同步数据接收池的存储区域信息，以广播的形式，基于所述RDMA网络调用通信设备将所述第三消息头写入对应的存储位置，以使所述从节点在所述从节点的所述大数据接收池的第三预设空间的分页内执行所述第三消息头所包含的逻辑；

在本地的所述大数据接收池的所述第三预设空间的分页内执行所述第三消息头所包含的逻辑；

在确认本地以及各所述从节点均完成执行所述第三消息头所包含的逻辑之后，控制本地以及各所述从节点将所在节点的所述第三预设空间的分页拷贝至所在节点的集群内存空间。

17. 根据权利要求16所述的集群数据同步方法，其特征在于，所述在确认本地以及各所述从节点均完成执行所述第三消息头所包含的逻辑之后，控制本地以及各所述从节点将所在节点的所述第三预设空间的分页拷贝至所在节点的集群内存空间，包括：

在确认本地以及各所述从节点均完成执行所述第三消息头所包含的逻辑之后，向各所述从节点发送确认成功响应码，以使各所述从节点在接收到所述确认成功响应码之后与所述主节点各自执行下述步骤：

将所述第三预设空间对应的所述集群内存空间中的第四预设空间的业务逻辑地址交换至所述第三预设空间，以基于所述第三预设空间的数据执行下一个处理逻辑，同时将所述第三预设空间的数据拷贝至所述第四预设空间；

在拷贝完成后，将所述业务逻辑地址交换回所述第四预设空间。

18. 根据权利要求16所述的集群数据同步方法，其特征在于，当所述待同步数据为与所述大数据接收池对应的数据类型且为非事务数据时，所述第三消息头仅包含所述待同步数据的同步逻辑；

当所述待同步数据为与所述大数据接收池对应的数据类型且为事务数据时，所述第三消息头包含所述待同步数据的同步逻辑以及与所述待同步数据对应的事务回调函数，以使所述主节点或所述从节点在执行所述对所述待同步数据的同步任务时执行所述事务回调函数对应的逻辑；

其中，所述事务回调函数至少包括针对事务数据同步的数据备份函数以及关键数据保护函数。

19. 根据权利要求18所述的集群数据同步方法，其特征在于，执行所述事务回调函数对应的逻辑包括：

检查所述待同步数据的同步任务的合理性；

执行数据备份及关键数据保护流程；

若所述待同步数据的同步任务合理且成功完成所述数据备份及关键数据保护流程，则向所述主节点返

回准备逻辑执行成功消息；

在准备逻辑执行成功后，触发执行所述待同步数据的同步任务的逻辑；

在执行完毕所述待同步数据的同步任务的逻辑后，向所述主节点返回同步任务执行完毕消息；

若所述主节点及各所述从节点均成功执行所述待同步数据的同步任务的逻辑，则释放关键数据保护，删除数据备份；否则，执行发生错误后回调操作，释放所述关键数据保护并基于数据备份恢复集群数据至初始状态。

20. 一种集群数据同步装置，其特征在于，包括：

监测单元，被设置为监测本地内存中预设的同步数据接收池；

校验单元，被设置为当监测到本地内存中所述同步数据接收池被源节点基于RDMA网络调用通信设备写入的待同步数据时，对所述待同步数据进行校验；

同步单元，被设置为当所述待同步数据通过校验后，在本地内存中执行对所述待同步数据的同步任务，并以广播的形式，基于所述RDMA网络调用通信设备将所述待同步数据写入分布式集群的各从节点的所述同步数据接收池，以使所述从节点在监测到所述待同步数据并完成对所述待同步数据的校验后执行对所述待同步数据的同步任务；

其中，所述源节点为所述分布式集群中的一个所述从节点。

21. 一种集群数据同步设备，其特征在于，包括：

存储器，被设置为存储计算机程序；

处理器，被设置为执行所述计算机程序，所述计算机程序被所述处理器执行时实现如权利要求1至19任意一项所述集群数据同步方法的步骤。

22. 一种非易失性可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至19任意一项所述集群数据同步方法的步骤。

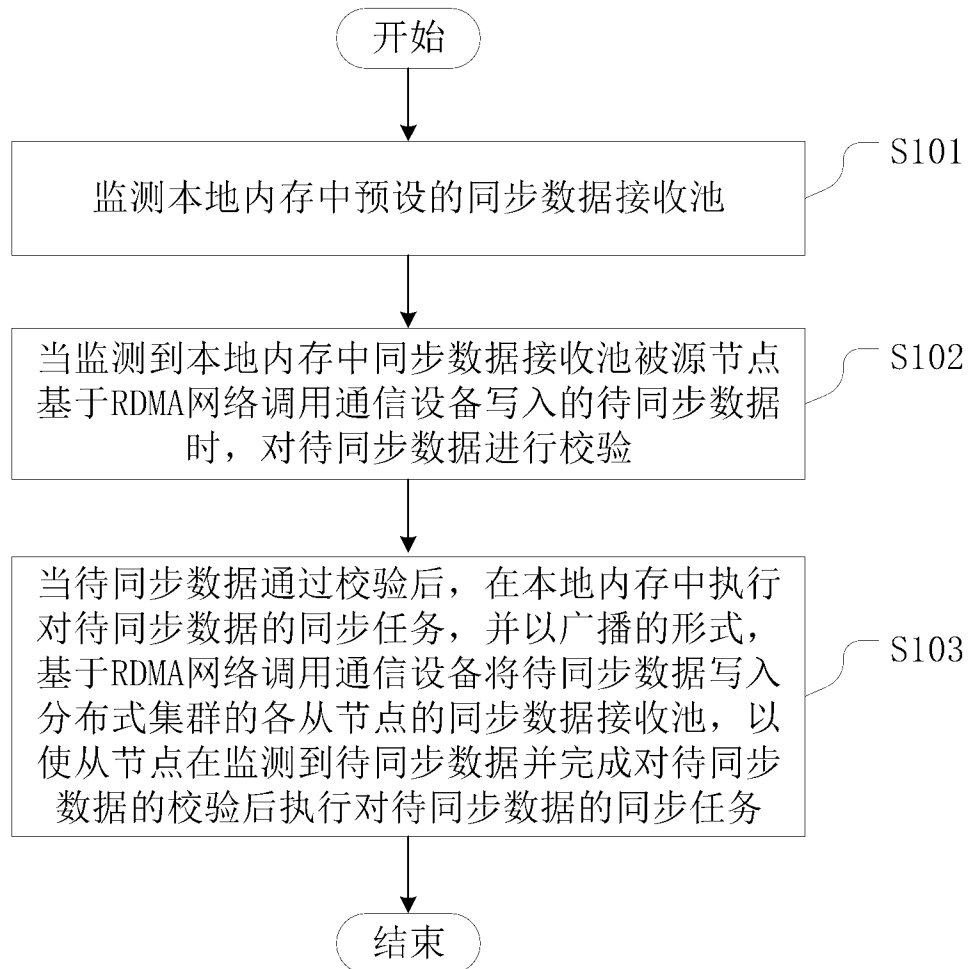


图 1

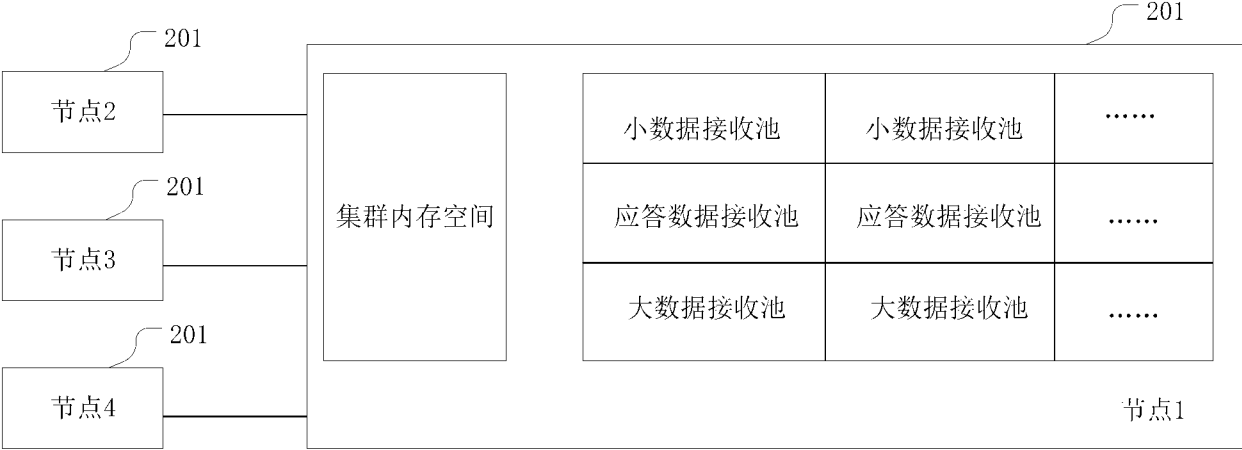


图 2

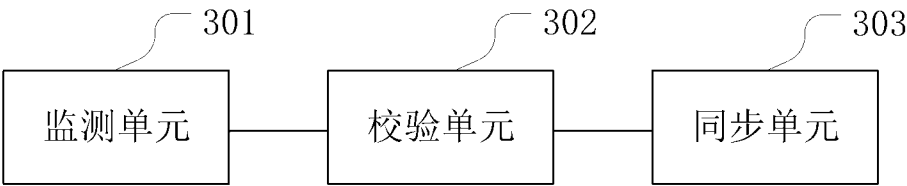


图 3

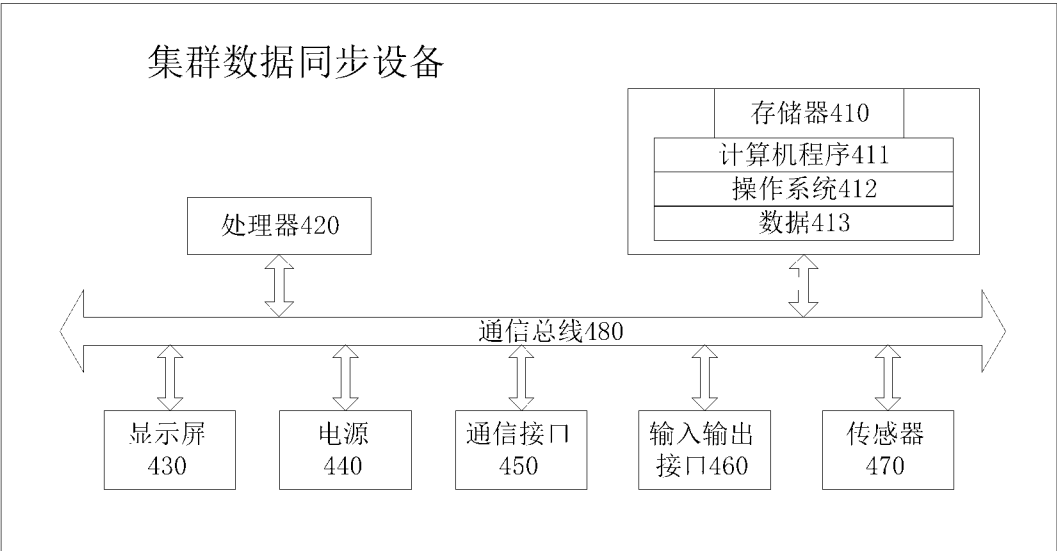


图 4

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/116231

**A. CLASSIFICATION OF SUBJECT MATTER**

H04L67/1095(2022.01)i; G06F9/50(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

IPC: H04L G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT, ENTXT, ENTXTC, VEN, WPABSC, CNKI: 同步, 主, 从, 远程直接数据存储, 广播, 校验, 验证, synchornize, main, master, secondary, remote direct memory access, broadcast, check, verify

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                 | Relevant to claim No. |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| PX        | CN 115550384 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 30 December 2022 (2022-12-30)<br>claims 1-22                                       | 1-22                  |
| X         | CN 112597251 A (TIANJIN NANKAI UNIVERSITY GENERAL DATA TECHNOLOGIES CO., LTD.) 02 April 2021 (2021-04-02)<br>description, paragraphs [0046]-[0080] | 1-7, 20-22            |
| A         | CN 112597251 A (TIANJIN NANKAI UNIVERSITY GENERAL DATA TECHNOLOGIES CO., LTD.) 02 April 2021 (2021-04-02)<br>entire document                       | 8-19                  |
| A         | CN 113360577 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 07 September 2021 (2021-09-07)<br>entire document                                  | 1-22                  |
| A         | CN 110691062 A (ZHEJIANG UNIVERSITY) 14 January 2020 (2020-01-14)<br>entire document                                                               | 1-22                  |
| A         | US 2014047263 A1 (COATNEY SUSAN et al.) 13 February 2014 (2014-02-13)<br>entire document                                                           | 1-22                  |



Further documents are listed in the continuation of Box C.



See patent family annex.

\* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

16 October 2023

Date of mailing of the international search report

18 October 2023

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/  
CN)  
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,  
Beijing 100088

Authorized officer

Telephone No.

**INTERNATIONAL SEARCH REPORT**  
**Information on patent family members**

International application No.

**PCT/CN2023/116231**

| Patent document<br>cited in search report |            |    | Publication date<br>(day/month/year) | Patent family member(s) |            |    | Publication date<br>(day/month/year) |
|-------------------------------------------|------------|----|--------------------------------------|-------------------------|------------|----|--------------------------------------|
| CN                                        | 115550384  | A  | 30 December 2022                     | None                    |            |    |                                      |
| CN                                        | 112597251  | A  | 02 April 2021                        | None                    |            |    |                                      |
| CN                                        | 113360577  | A  | 07 September 2021                    | None                    |            |    |                                      |
| CN                                        | 110691062  | A  | 14 January 2020                      | None                    |            |    |                                      |
| US                                        | 2014047263 | A1 | 13 February 2014                     | WO                      | 2014026025 | A1 | 13 February 2014                     |
|                                           |            |    |                                      | US                      | 8904231    | B2 | 02 December 2014                     |
|                                           |            |    |                                      | EP                      | 2883147    | A1 | 17 June 2015                         |
|                                           |            |    |                                      | EP                      | 2883147    | A4 | 04 May 2016                          |
|                                           |            |    |                                      | EP                      | 2883147    | B1 | 28 June 2017                         |

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                                                                                    |                                            |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|--------------------------------------------|
| A. 主题的分类<br>H04L67/1095(2022.01)i; G06F9/50(2006.01)i<br><br>按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类                                                                                                                                                                                                                                                                                                                                                                              |                                                                                    |                                            |
| B. 检索领域<br>检索的最低限度文献(标明分类系统和分类号)<br>IPC: H04L G06F<br><br>包含在检索领域中的除最低限度文献以外的检索文献<br><br>在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))<br>CNTXT, ENTXT, ENTXTC, VEN, WPABSC, CNKI:同步, 主, 从, 远程直接数据存储, 广播, 校验, 验证, synchornize, main, master, secondary, remote direct memory access, broadcast, check, verify                                                                                                                                                                      |                                                                                    |                                            |
| C. 相关文件                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                    |                                            |
| 类 型*                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 引用文件, 必要时, 指明相关段落                                                                  | 相关的权利要求                                    |
| PX                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | CN 115550384 A (苏州浪潮智能科技有限公司) 2022年12月30日 (2022 - 12 - 30)<br>权利要求1-22             | 1-22                                       |
| X                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | CN 112597251 A (天津南大通用数据技术股份有限公司) 2021年4月2日 (2021 - 04 - 02)<br>说明书第[0046]-[0080]段 | 1-7, 20-22                                 |
| A                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | CN 112597251 A (天津南大通用数据技术股份有限公司) 2021年4月2日 (2021 - 04 - 02)<br>全文                 | 8-19                                       |
| A                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | CN 113360577 A (苏州浪潮智能科技有限公司) 2021年9月7日 (2021 - 09 - 07)<br>全文                     | 1-22                                       |
| A                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | CN 110691062 A (浙江大学) 2020年1月14日 (2020 - 01 - 14)<br>全文                            | 1-22                                       |
| A                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | US 2014047263 A1 (COATNEY SUSAN等) 2014年2月13日 (2014 - 02 - 13)<br>全文                | 1-22                                       |
| <input type="checkbox"/> 其余文件在C栏的续页中列出。 <input checked="" type="checkbox"/> 见同族专利附件。                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                    |                                            |
| <p>* 引用文件的具体类型:</p> <p>“A” 认为不特别相关的表示了现有技术一般状态的文件</p> <p>“D” 申请人在国际申请中引证的文件</p> <p>“E” 在国际申请日的当天或之后公布的在先申请或专利</p> <p>“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)</p> <p>“O” 涉及口头公开、使用、展览或其他方式公开的文件</p> <p>“P” 公布日先于国际申请日但迟于所要求的优先权日的文件</p> <p>“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件</p> <p>“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性</p> <p>“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性</p> <p>“&amp;” 同族专利的文件</p> |                                                                                    |                                            |
| 国际检索实际完成的日期<br>2023年10月16日                                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                                                    | 国际检索报告邮寄日期<br>2023年10月18日                  |
| ISA/CN的名称和邮寄地址<br>中国国家知识产权局<br>中国北京市海淀区蓟门桥西土城路6号 100088                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                    | 授权官员<br>李美丽<br><br>电话号码 (+86) 010-62089957 |

国际检索报告  
关于同族专利的信息

国际申请号

PCT/CN2023/116231

| 检索报告引用的专利文件 |            |    | 公布日<br>(年/月/日) | 同族专利 |            |    | 公布日<br>(年/月/日) |
|-------------|------------|----|----------------|------|------------|----|----------------|
| CN          | 115550384  | A  | 2022年12月30日    | 无    |            |    |                |
| CN          | 112597251  | A  | 2021年4月2日      | 无    |            |    |                |
| CN          | 113360577  | A  | 2021年9月7日      | 无    |            |    |                |
| CN          | 110691062  | A  | 2020年1月14日     | 无    |            |    |                |
| US          | 2014047263 | A1 | 2014年2月13日     | WO   | 2014026025 | A1 | 2014年2月13日     |
|             |            |    |                | US   | 8904231    | B2 | 2014年12月2日     |
|             |            |    |                | EP   | 2883147    | A1 | 2015年6月17日     |
|             |            |    |                | EP   | 2883147    | A4 | 2016年5月4日      |
|             |            |    |                | EP   | 2883147    | B1 | 2017年6月28日     |