(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 105069160 A
(43)申请公布日 2015.11.18

(21)申请号 201510530766.5

(22)申请日 2015.08.26

(71)申请人 国家电网公司
地址 100017 北京市西城区西长安街 86 号
申请人 北京许继电气有限公司

(72)发明人 邢艳 张宇 缪燕 刘红超 李海 张学深

(74)专利代理机构 北京立成智业专利代理事务所（普通合伙） 11310
代理人 张江涵

(51)Int.Cl.
G06F 17/30(2006.01)

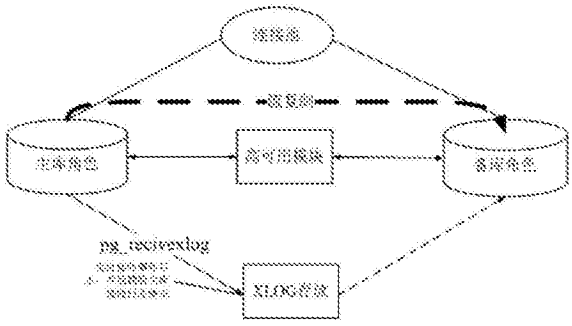
权利要求书1页 说明书4页 附图2页

(54)发明名称

一种基于自主可控数据库的高可用性方法及构架

(57)摘要

本发明提供一种基于自主可控数据库的高可用性方法及构架，该架构是基于国产自主可控的数据库软件产品 PowerDB 设计的，包括连接池、HA 模块、主库、备库组成，其中主库和备库作为主服务器和备用服务器的节点，连接池负责建立应用程序和数据库之间的连接，HA 模块用来做集群的状态监控和主备机自动切换的部分，通过块级别的数据复制技术，以及同步事务提交方式，保证了在灾备工作中，主备库的数据一致性，同时又见减小了复制对性能的影响。



CN 105069160 A

1. 一种基于自主可控数据库的高可用性方法，其特征在于：

该方法具体包括以下步骤：

步骤 1：复制开始后，备库会根据时间线上的时间点向主库发起日志传输请求，主库根据要求将事务日志传送给备库；

步骤 2：主库上每一个事物提交，必须等待日志传入到主库上每一个事务提交，必须等待日志传入到备库上所有节点并写入磁盘后再提交。

2. 根据权利要求 1 所述的一种基于自主可控数据库的高可用性方法，其特征在于：在主库完成一个事务与在备库中看到数据库的变化之间有一个很小的时间差，但远远小于直接对日志的传输。

3. 根据权利要求 1 所述的一种基于自主可控数据库的高可用性方法，其特征在于：只有当主、备库的所有服务器都接收到数据，并写入磁盘上的事务日志后，才能够执行提交或者回滚操作，同时，采用多个备库的同步流复制方案，要求每个备库的存储空间与主节点要保持一致，当同步备用节点异常时，采用角色转嫁给其他备用结点中的一员。

4. 根据权利要求 1 所述的一种基于自主可控数据库的高可用性方法，其特征在于：只有当主、备库的所有服务器都接收到数据，并写入磁盘上的事务日志后，才能够执行提交或者回滚操作，并采用主备两个节点，再增设一个主、备节点都可以访问到的专门存放事务日志的位置。

5. 根据权利要求 1 所述的一种基于自主可控数据库的高可用性方法，其特征在于：重建主库，可以通过原主库生成最新的备份进行恢复，恢复完成后，原主库以备库模式启动，并成为新主库的新备库，将新产生的所有事务日志，进行主备的同步，同步完成后，完成数据库重建。

6. 根据权利要求 5 所述的一种基于自主可控数据库的高可用性方法，其特征在于：主数据库服务器硬件配置优于备数据库服务器。

7. 根据权利要求 1 所述的一种基于自主可控数据库的高可用性方法，其特征在于：连接池将传输过来的 SQL 语句进行解析，将只读性操作分发至备机中的任意数据库，修改操作分发至主数据库。

8. 根据权利要求 7 所述的一种基于自主可控数据库的高可用性方法，其特征在于：当出现一主多备的情况，中间件会采取分布式计算中的动态算法和自适应算法来保证备机间的负载均衡。

9. 一种基于自主可控数据库的高可用性构架，其特征在于：

构架包括连接池、HA 模块、主库、备库组成，其中主库和备库作为主服务器和备用服务器的节点，连接池负责建立应用程序和数据库之间的连接，HA 模块用来做集群的状态监控和主备机自动切换的部分。

10. 根据权利要求 9 所述的一种基于自主可控数据库的高可用性构架，其特征在于：HA 模块包括监控模块、切换模块及仲裁模块三个模块，其中：监控模块主要负责循环监控主库和备库所有服务器网络情况与数据库健康状况；仲裁模块负责接受监控模块提交的故障信息，对故障进行分类诊断和处理；切换模块负责接收仲裁模块的切换指令，完成主备节点切换。

一种基于自主可控数据库的高可用性方法及构架

技术领域

[0001] 本发明涉及数据库领域，尤其是数据库容灾技术领域。

背景技术

[0002] 近几年来，随着社会信息量的急速增加，数据库产品被越来越广泛的应用，企业和用户对数据库的可用性，实时性和安全性，也提出了更高的要求。用户原有意识中的数据备份已经无法满足要求，大部分企业都需要更高端的容灾备份产品来解决人为误操作、软件错误、病毒入侵等“软”性灾害以及硬件故障、自然灾害等“硬”性灾害。在灾备产品方面，用户比较关注以下几个问题，高可用性，对数据库性能的影响以及主备端数据的一致性。很多主流的数据库，由于实现机制问题，经常会造成备库丢数据，或者主备库数据的不一致的情况。这都造成了产品的使用隐患，亟待解决。

发明内容

[0003] 本发明的要解决的问题是，自主可控数据库（PowerDB 数据库）管理软件，在灾备工作中，主备端由于数据复制造成的性能显著下降，灾备端数据库无法自动切换和回切，无法保证灾备端数据的一致性和可靠性，无法保证备库在介质恢复过程中处于只读模式等问题。

[0004] 本发明提供一种基于自主可控数据库的高可用性方法，该方法是采用一条一条记录的方式，随着主机日志的产生实施传送至备机，该方法具体包括以下步骤：

[0005] 步骤 1：复制开始后，备库会根据时间线上的时间点向主库发起日志传输请求，主库根据要求将事务日志传送给备库；

[0006] 步骤 2：主库上每一个事物提交，必须等待日志传入到主库上每一个事务提交，必须等待日志传入到备库上所有节点并写入磁盘后再提交。

[0007] 作为本发明的进一步改进，事务提交的方式是异步的，即在主库完成一个事务与在备库中看到数据库的变化之间有一个很小的时间差，但远远小于直接对日志的传输；

[0008] 作为本发明的进一步改进，只有当主、备库的所有服务器都接收到数据，并写入磁盘上的事务日志后，才能够执行提交或者回滚操作，同时，采用多个备库的同步流复制方案，要求每个备库的存储空间与主节点要保持一致，当同步备用节点异常时，采用角色转嫁给其他备用结点中的一员；或者，采用主备两个节点，再增设一个专门存放事务日志的位置（主、备节点都可以访问到）。

[0009] 为了保证主备切换的高可用性能，采用高可用性方法构建该构架，构架包括连接池、HA 模块、主库、备库组成，其中主库和备库作为主服务器和备用服务器的节点，连接池负责建立应用程序和数据库之间的连接，HA 模块用来做集群的状态监控和主备机自动切换的部分。

[0010] 进一步地，HA 模块包括监控模块、切换模块及仲裁模块三个模块，其中：监控模块主要负责循环监控主库和备库所有服务器网络情况与数据库健康状况；仲裁模块负责接受

监控模块提交的故障信息，对故障进行分类诊断和处理；切换模块负责接收仲裁模块的切换指令，完成主备节点切换。

[0011] 进一步地，重建主库，可以通过原主库生成最新的备份进行恢复，恢复完成后，原主库以备库模式启动，并成为新主库的新备库，将新产生的所有事务日志，进行主备的同步，同步完成后，完成数据库重建。

[0012] 进一步地，主数据库服务器硬件配置优于备数据库服务器。

[0013] 为了保证在介质恢复过程中，从库能够始终处于可读取状态，连接池将传输过来的 SQL 语句进行解析，将只读性操作分发至备机中的任意数据库，修改操作分发至主数据库。

[0014] 进一步地，当出现一主多备的情况，中间件会采取分布式计算中的动态算法和自适应算法来保证备机间的负载均衡。

附图说明

[0015] 图 1 为本发明数据库高可用技术架构图；

[0016] 图 2 为本发明应用同异步流复制工具的高可用架构图；

[0017] 图 3 为本发明高可用读写分离架构图；

具体实施方式

[0018] 以下结合说明书附图对本发明进一步详细说明。应当理解为，此处所描述的实施例仅用于解释本发明，但并不限定本发明。

[0019] 本发明的高可用技术架构是基于 PowerDB 数据库软件产品提出的。PowerDB 数据库是一款基于开源数据库 PostgreSQL 二次开发并封装的数据库管理软件，其内核延续了 PostgreSQL 的架构和设计。该数据库功能全面，可以应用于政府、科研、互联网、工业企业等多种场景。

[0020] 本发明的高可用架构由连接池、HA 模块、主库、备库组成，应用了同步 / 异步的流复制技术，其中主库和备库作为主服务器和备用服务器的节点（备库可以采取多个节点也可以采取级联架构），连接池负责建立应用程序和数据库之间的连接，同时具有主备库读写分离和负载均衡的用途。HA 模块用来做集群的状态监控和主备机自动切换的部分。

[0021] HA 模块可以划分以下三个模块：监控模块、切换模块，仲裁模块。

[0022] a) 监控模块

[0023] 监控模块主要负责循环监控主库和备库所有服务器网络情况与数据库健康状况，具体包括：监控主 / 备库网关，服务器网络状况；监控主 / 备库心跳，数据库健康状况；监控主库角色对应 IP；监听主 / 备库监听是否启动；监控主备库虚拟 IP，端口和数据库心跳；检查备机延迟情况，判断是否允许切换。

[0024] b) 仲裁模块

[0025] 负责接受监控模块提交的故障信息，对故障进行分类诊断和处理。通过仲裁模块的相关判断，来决定是否进行主备机进行的切换。这样，就可以避免出现资源争用，即所谓的“脑裂”现象。

[0026] c) 切换模块

[0027] 负责接收仲裁模块的切换指令，完成主备节点切换。具体工作包含主备节点的切换，主节点接管备库的虚拟 VIP，主备角色转换，备节点升级为主节点，备节点停库，备节点备份配置文件，备节点启动数据库，备节点激活数据库，备节点接管主库虚拟 VIP，备节点转换主备角色。

[0028] 一旦主库发生故障，监控模块采集后，通过仲裁模块进行判断与处理，将指令发送给切换模块，完成切换操作。

[0029] 在主库故障短时间内无法排除的情况下，为了避免备库单点运行，降低业务终端风险，需对主库进行快速重建，将容灾系统恢复到可提供故障自动切换的高可用状态。重建主库，可以通过原主库生成最新的备份进行恢复，恢复完成后，原主库以备库模式启动，并成为新主库的新备库。同样利用流复制的原理，将新产生的所有事务日志，进行主备的同步，同步完成后，完成数据库重建。快速重建技术要应用归档日志，同样利用流复制的原理，采用数据库并行恢复技术，加快重建速度。

[0030] 通常情况下，主数据库服务器硬件配置优于备数据库服务器。这样在主库故障排除或者重建后，为了保证服务的高性能，需进行回切操作，恢复原来的主备状态。

[0031] 这时原主库需要以备库模式启动，启动后成为新主库的新备库。新主库需要将故障发生时刻至当前这一时间段里产生的事务日志复制到新备库（原主库）上，并在数据同步完成后，执行切换操作，恢复容架构的原有模式。

[0032] 在 PowerDB 的架构上，对主备两端的数据库块变化进行同步。但有别于传统的日志传输模式，本发明采用的是流复制的方式进行日志传输。此外，在事务提交方式上也采取了新的架构模式，良好的解决了数据库复制性能问题以及主备端数据不完整，不一致的问题。

[0033] 本发明采用的是是一种基于日志的传输技术，在传输过程中不需要等待主库日志填写完整后再传送到备库，而是采用一条一条记录的方式，随着主机日志的产生实时传送至备机。流复制分为同步流复制和异步流复制两种方式，同步流复制是指主库上每一个事务提交，必须等待日志传入到备库上所有节点并写入磁盘后再提交；异步流复制是指复制开始后，备库会根据时间线上的时间点向主库发起日志传输请求，主库根据要求将事务日志传送给备库，因此备库会和主库有一个微小的时间差。

[0034] 在默认情况下，事务提交的方式是异步的，即在主库完成一个事务与在备库中看到数据库的变化之间有一个很小的时间差（远远小于直接对日志的传输），用户也可以将事务提交模式更改为同步，即只有当主、备库的所有服务器都接收到数据，并写入磁盘上的事务日志后，才能够执行提交或者回滚操作，只有其中一方完成，其它数据库没有完成操作，事务无法提交。这种传输模式虽然保证了灾备端数据的完整性和一致性，但是，一旦备库出现问题，势必造成很大的影响。因此可以采用以下两种方式解决。第一，如果预算可以，建议采用多个备库的同步流复制方案，要求每个备库的存储空间与主节点要保持一致，当同步 standby 节点异常时，采用角色转嫁给其他 standby 中的一员，这样做可以降低了 standby 异常带来的风险。还有一种架构模式，采用主备两个节点，再增设一个专门存放事务日志的位置（主、备节点都可以访问到），以此来保证事务日志的安全，同时在备库端增加一个实时接收日志工具，这个工具的作用是制造一个虚拟的 standby 节点，使得数据库在切换之前，应用高可用模块 fence 掉主节点，同时判断当前 standby 节点的恢复进度是否

比 pg_recivexlog 新，如果不是，则从 pg_recivexlog 中将事务日志拷贝过来，应用后激活 standby，达到数据一致，不丢失的效果。

[0035] 灾备工作中，备库应用主库传过来的日志，进行介质恢复，达到和主数据库同步的效果。如果在介质恢复过程中，从库能够始终处于可读取状态，就可以来处理查询、报表和统计等业务工作。这样做，实际上是应用了读写分离的工作模式，可以有效的减轻主数据库的压力和 IO。

[0036] 读写分离操作是通过连接池来完成，连接池将传输过来的 SQL 语句进行解析，并分为 SELECT 和修改操作两大类，将只读性操作分发至备机中的任意数据库，修改操作分发至主数据库。因此在主 / 备模式中，DDL 和 DML 操作都在主节点上完成，SELECT 操作可以在备节点上执行，当然也可以在主节点上执行。同时，主库通过流复制技术，将日志同步到备库中，以保证所有的数据块改变是一致的。同时，由于 PowerDB 采取了写新数据时，旧数据不删除，只是将新数据插入的方式进行的多版本并发控制，因此备库完全可以只读取当前生效的数据。从而确保查询与恢复的同时进行，即实现了在介质恢复过程中，可以对备机进行统计和查询操作。同时，当出现一主多备的情况，中间件会采取分布式计算中的动态算法和自适应算法来保证备机间的负载均衡；也会考虑级联复制的部署方式来解决主机的性能问题。

[0037] 以上内容是结合具体的优选实施方式对本发明所做的进一步详细说明，不能认定本发明的具体实施只局限于该说明。对于本发明所属技术领域的普通技术人员来说，在不脱离本发明构思及精神的前提下，通过若干简单推演或替换，都应视为属于本发明的保护范围。

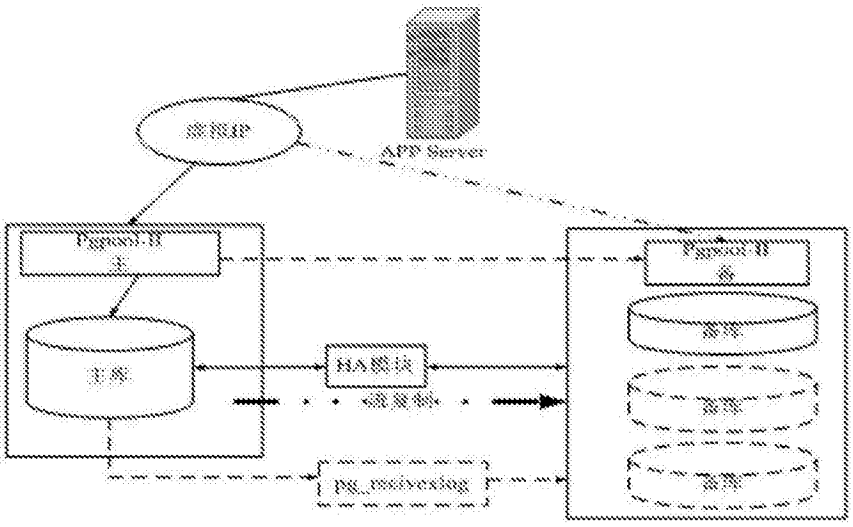

APP Server
HA模块
pg_recivexlog

图 1

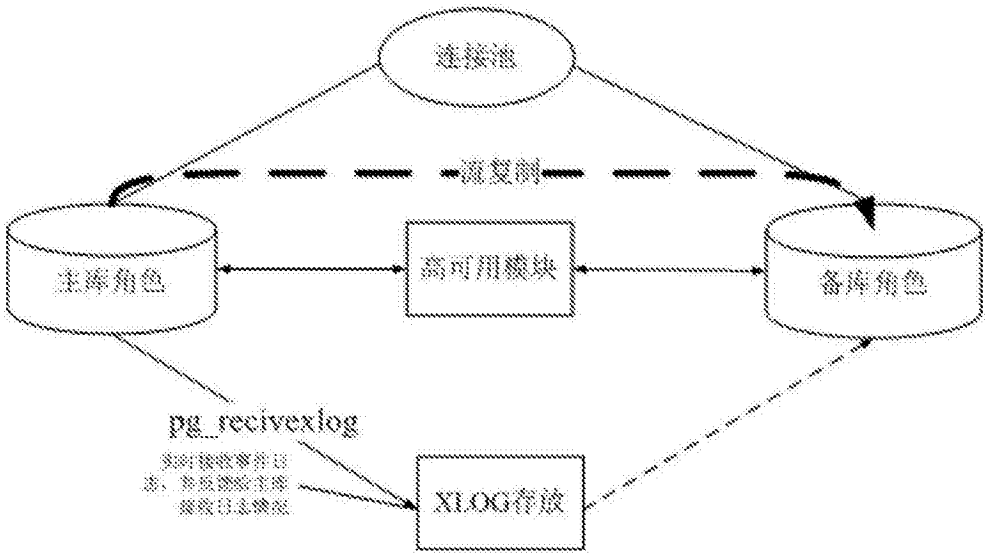

连接池
流复制
主库角色
高可用模块
备库角色
pg_recivexlog
XLOG存放

图 2

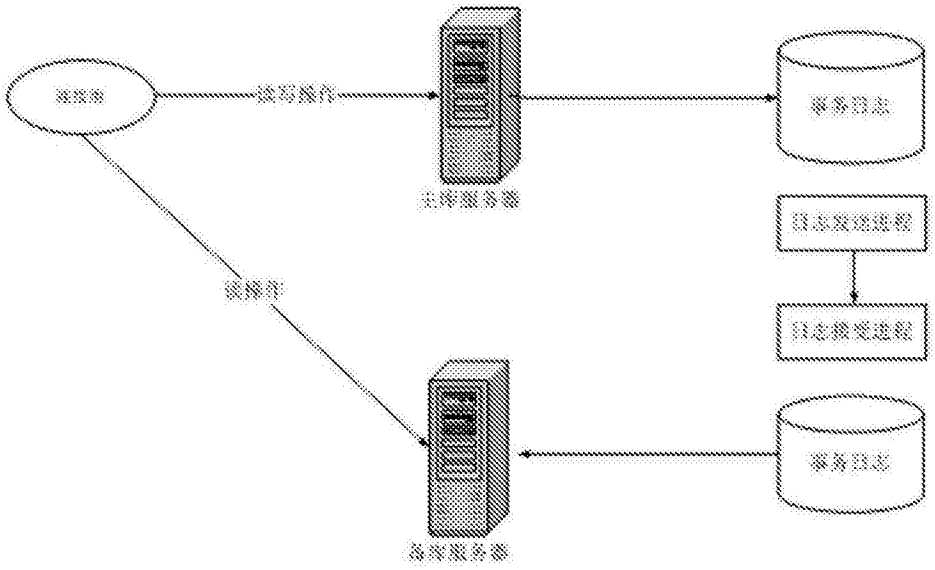


图 3