

## 一种提高存储系统可靠性的方法和相关装置

### 技术领域

本申请涉及计算机分布式存储系统技术领域，尤其涉及一种提高存储系统可靠性的方法和相关装置。

### 背景技术

目前的云计算和大数据等平台，分布式数据存储架构的使用越来越广泛，分布式存储技术是指存储设备包括多个存储控制器，在将数据写入存储设备时，可通过多个存储控制器写入数据。

分布式存储系统在写数据时，一般是把数据分散，然后写入到不同的控制器的内存，同时，还为写入每个控制器内存的数据配置多个副本，防止某个控制器失效造成数据不可访问或者丢失。当某个控制器失效时，由于有副本的存在，故数据还是可以访问的。

目前分布式存储系统都是采用存储控制设备和存储器分离的方式。如果存储了目标数据的存储控制器以及配置了该目标数据副本的存储控制器同时发生故障，或者存储了目标数据的存储控制器对应的存储控制设备发生故障导致失效，则会导致目标数据不能访问以及与该存储控制设备连接的存储器中的数据也不能访问，从而导致主机业务中断，存储系统可靠性降低。

因此，如何保证在存储控制设备故障的情况下，主机业务不中断，提高存储系统的可靠性是目前亟待解决的技术问题。

### 发明内容

本申请提供了一种提高存储系统可靠性的方法和相关装置，能够在存储控制设备发生故障导致失效时，保证主机业务不中断，有效提高存储系统的可靠性和可用度。

第一方面，提供了一种提高存储系统可靠性的方法，所述存储系统包括第一控制设备和第二控制设备，所述方法包括：目标控制器接收写请求，该写请求包括待写入数据，该目标控制器属于所述第一控制设备；该目标控制器将所述待写入数据写入该目标控制器的内存；该目标控制器将所述待写入数据写入该目标控制器的镜像控制器的内存，所述镜像控制器中的至少一个控制器属于所述第二控制设备。

可选的，将待写入数据写入到目标控制器和该目标控制器的镜像控制器的内存之后，目标控制器向主机发送写入完成消息，该写入完成消息用于指示待写入数据已经成功写入。

在本申请实施例中，将写请求中携带的待写入数据进行写入时，不仅需要将其写入目标控制器的内存，同时将其写入该目标控制器的镜像控制器的内存，而且镜像控制器中存在与目标控制器不在同一个控制设备的控制器，这样可以保证待写入数据能够写入到不同的控制设备中，进而可以在任意控制设备发生故障的情况下，主机仍能够通过其它的控制设备访问到该待写入数据，保证主机业务的连续性，提高存储系统的可靠性和可用度。

结合第一方面，在一种可能的实现方式中，第一控制设备和第二控制设备连接至存储设备，当第一控制设备故障，主机通过第二控制设备访问该存储设备。

在本申请实施例中，第一控制设备和第二控制设备可以同时访问存储设备，这样可以保证在任意一个控制设备发生故障时，主机仍旧可以通过其它的控制设备访问存储设备，保证主机业务的连续性，提高存储系统的可靠性。

结合第一方面，在一种可能的实现方式中，所述目标控制器及所述目标控制器的镜像控制器构成备份控制器组，所述备份控制器组包括N个控制器，N为大于等于2的整数，当所述备份控制器组中的M个控制器故障，且所述M个控制器包括所述目标控制器时，主控制器从所述目标控制器的镜像控制器中选择一个控制器作为新的目标控制器，所述新的目标控制器用于接收所述写请求，所述主控制器为从所述第一控制设备和第二控制设备中的控制器中选择的用于管理其它控制器的控制器；所述主控制器为所述新的目标控制器建立M个新的镜像控制器，并将所述新的目标控制器中的内存中的数据镜像至所述M个新的镜像控制器中，M为大于等于1且小于N的整数。

在本申请实施例中，在包含目标控制器的多个控制器发生故障时，主控制器可以从目标控制器的镜像控制器中选择新的目标控制器并为其建立新的镜像控制器，保证在多个控制器发生故障时，主机仍能够访问目标数据，保证主机业务不中断，提高存储系统的可靠性。

结合第一方面，在一种可能的实现方式中，所述目标控制器及所述目标控制器的镜像控制器构成备份控制器组，所述备份控制器组包括N个控制器，N为大于等于2的整数，当所述备份控制器组中的M个控制器故障，且所述M个控制器不包括所述目标控制器时，主控制器为所述目标控制器建立M个新的镜像控制器，并将所述目标控制器中的内存中的数据镜像至所述M个新的镜像控制器中，所述主控制器为从所述第一控制设备和第二控制设备中的控制器中选择的用于管理其它控制器的控制器，M为大于等于1且小于N的整数。

在本申请实施例中，在不包含目标控制器的多个控制器发生故障时，主控制器可以为该目标控制器建立新的镜像控制器，保证目标控制器能够维持多个镜像控制器，且在多个控制器发生故障时，主机仍能够访问目标数据，保证主机业务不中断，提高存储系统的可靠性。

结合第一方面，在一种可能的实现方式中，所述N等于3。

结合第一方面，在一种可能的实现方式中，当所述第一控制设备故障时，将所述第二控制设备中的所述目标控制器的其中一个镜像控制器作为新的目标控制器，并为所述新的目标控制器建立镜像控制器，并将所述新的目标控制器的内存中的数据镜像至新的镜像控制器的内存。

在本申请实施例中，每个控制设备中的控制器都在其它控制设备中存在镜像控制器，在控制设备出现故障的情况下，主控制器将其它控制设备中的目标控制器的镜像控制器升级为新的目标控制器，并为其建立镜像控制器，可以保证在控制设备出现故障时，主机仍旧可以访问目标数据，保证主机业务的连续性，提高存储系统的可靠性和可用度。

第二方面，提供了一种存储控制系统，所述存储控制系统包括第一控制设备和第二控

制设备,其特征在于,所述第一控制设备包括目标控制器,所述目标控制器包括:接收模块,用于接收写请求,所述写请求包括待写入数据;处理模块,用于将所述待写入数据写入所述目标控制器的内存;所述处理模块,还用于将所述待写入数据写入所述目标控制器的镜像控制器的内存,所述镜像控制器中的至少一个控制器属于第二控制设备。

结合第二方面,在一种可能的实现方式中,所述目标控制器及所述目标控制器的镜像控制器构成备份控制器组,所述备份控制器组包括N个控制器,N为大于等于2的整数,所述存储控制系统包括主控制器,所述主控制器为从所述第一控制设备和第二控制设备中的控制器中选择的用于管理其他控制器的控制器,所述主控制器包括:故障处理模块,用于当所述备份控制器组中的M个控制器故障,且所述M个控制器包括所述目标控制器时,主控制器从所述目标控制器的镜像控制器中选择一个控制器作为新的目标控制器,所述新的目标控制器用于接收所述写请求;镜像模块,用于为所述新的目标控制器建立M个新的镜像控制器,并将所述新的目标控制器中的内存中的数据镜像至所述M个新的镜像控制器中,M为大于等于1且小于N的整数。

结合第二方面,在一种可能的实现方式中,所述目标控制器及所述目标控制器的镜像控制器构成备份控制器组,所述备份控制器组包括N个控制器,N为大于等于2的整数,所述存储控制系统还包括主控制器,所述主控制器为从所述第一控制设备和第二控制设备中的控制器中选择的用于管理其他控制器的控制器,所述主控制器包括:故障处理模块,当所述备份控制器组中的M个控制器故障,且所述M个控制器不包括所述目标控制器时,主控制器为所述目标控制器建立M个新的镜像控制器;镜像模块,用于将所述目标控制器中的内存中的数据镜像至所述M个新的镜像控制器中,M为大于等于1且小于N的整数。

结合第二方面,在一种可能的实现方式中,所述N等于3。

结合第二方面,在一种可能的实现方式中,当所述第一控制设备故障时,将所述第二控制设备中的所述目标控制器的其中一个镜像控制器作为新的目标控制器,并为所述新的目标控制器建立镜像控制器,并将所述新的目标控制器的内存中的数据镜像至新的镜像控制器的内存。

第三方面,提供了一种存储控制系统,所述存储控制系统包括第一控制设备和第二控制设备,其特征在于,所述第一控制设备包括目标控制器,所述目标控制器包括处理器和存储器,所述处理器执行所述存储器存储的计算机指令,使得所述目标控制器执行上述第一方面以及结合上述第一方面中的任意一种实现方式所提供的方法。

结合第三方面,在一种可能的实现方式中,所述存储控制系统包括主控制器,所述主控制器为从所述第一控制设备和第二控制设备中的控制器中选择的用于管理其他控制器的控制器,所述主控制器包括处理器和存储器,所述处理器执行所述存储器存储的计算机指令,使得所述主控制器执行上述第一方面以及结合上述第一方面中的任意一种实现方式所提供的方法。

第四方面,提供了一种计算机存储介质,上述计算机存储介质存储有计算机程序,当该计算机程序被处理器执行时,可以实现上述第一方面以及结合上述第一方面中的任意一

种实现方式所提供的方法的流程。

第五方面，提供了一种计算机程序产品，该计算机程序产品包括计算机指令，当该计算机程序被计算机执行时，使得计算机可以执行上述第一方面以及结合上述第一方面中的任意一种实现方式所提供的方法的流程。

#### 附图说明

- 图 1A 为本申请实施例提供的一种串联模型的示意图；
- 图 1B 为本申请实施例提供的一种并联模型的示意图；
- 图 2A 为本申请实施例提供的一种存储系统的结构示意图；
- 图 2B 为本申请实施例提供的一种数据双副本存储的示意图；
- 图 3 为本申请实施例提供的一种应用场景示意图；
- 图 4 为本申请实施例提供的一种提高存储系统可靠性的方法的流程示意图；
- 图 5 为本申请实施例提供的一种数据多副本写入示意图；
- 图 6A 为本申请实施例提供的一种双控制器同时故障的示意图；
- 图 6B 为本申请实施例提供的一种控制设备故障的示意图；
- 图 6C 为本申请实施例提供的一种控制器依次故障的示意图；
- 图 7 为本申请实施例提供的一种控制设备的可靠性模型的示意图；
- 图 8 为本申请实施例提供的一种存储系统的可靠性模型的示意图；
- 图 9 为本申请实施例提供的一种存储控制系统的结构示意图；
- 图 10 为本申请实施例提供的又一种存储控制系统的结构示意图。

#### 具体实施方式

下面结合附图对本申请实施例中的技术方案进行清楚、完整的描述，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它实施例，都属于本申请保护的范围。

首先，结合附图对本申请中所涉及的部分用语和相关技术进行解释说明，以便于本领域技术人员理解。

高速串行计算机扩展总线标准（peripheral component interconnect express，PCIe）属于高速串行点对点双通道高带宽传输，所连接的设备分配独享通道带宽，不共享总线带宽，数据传输速率高。任何两个 PCIe 设备之间的连接称为链路，由一个或多个通道的集合构建。

小型计算机系统接口（small computer system interface，SCSI）是一种用于计算机和智能设备（例如硬盘、光驱、打印机、扫描仪等）之间系统级接口的独立处理器标准。SCSI 接口是一个通用接口，在 SCSI 母线上可以连接主机适配器和 SCSI 外设控制器，挂在一个

SCSI 母线上的多个外设可以同时工作，SCSI 接口可以同步或异步传输数据。SCSI 接口通过独立的高速的 SCSI 卡来控制数据的读写操作，中央处理器(central processing unit, CPU)不需要浪费时间等待，可以提高系统的整体性能。

串行连接小型计算机系统接口(serial attached SCSI, SAS)是新一代的 SCSI 技术，采用串行技术以获得更高的传输速度，并通过缩短连结先改善内部空间，提供与串行高级技术附件(serial advanced technology attachment, SATA)硬盘的兼容性。SAS 是点到点的结构，可以建立磁盘到控制器的直接连接，此外，SAS 结构有非常好的扩展能力，最多可以连接 16384 个磁盘设备。每个 SAS 电缆有 4 根电缆，2 根输入 2 根输出，可以同时进行数据的读写。

非易失性内存主机控制器接口规范(non-volatile memory express, NVMe)是一个逻辑设备接口规范，是专为闪存类存储设计的类似串行高级主机控制器接口(serial ATA advanced host controller interface, AHCI)的一种协议。其本质是建立了多个计算机与存储设备的通路，多个通路其实就是多个队列，NVMe 协议最多可以有 64k 个队列，并发程度大大提升，性能得到提高。

直接内存访问(direct memory access, DMA)是一种完全由硬件执行 I/O 交换的工作方式。在这种方式中，DMA 控制器从 CPU 完全接管对总线的控制，数据交换不经过 CPU，而直接在内存和 IO 设备之间进行。使用 DMA 可以减少大批量数据传输时 CPU 的开销，采用 DMA 控制器生成访存地址并控制访存过程，传输速度快，且 CPU 基本不干预，仅在初始化和结束时参与，CPU 与外设并行工作，效率较高。

远程直接内存访问(remote direct memory access, RDMA)是一种绕过远程主机操作系统内核访问其内存中数据的技术。由于不经过操作系统，RDMA 通过网络将数据从一个系统快速移动到远程系统存储器中，可以节省大量 CPU 资源，提高了系统吞吐量，消除了外部存储器复制和上下文切换的开销，能够解放内存带宽用于改进应用系统性能。

数据副本是指存储数据的冗余数据，即将同一个数据存储在不同的存储设备上，用来提高系统可用性。当存储主副本数据的存储设备故障失效，可以访问存储了副本数据的其它存储设备来获取数据，保证业务不中断。

失效率(failure rate)也称故障率，是一个工程系统或零件失效的频率，单位通常为每小时的失效次数，用字母  $\lambda$  表示。平均故障间隔时间(mean time between failure, MTBF)是指可修理的设备从本次故障起到下一次故障为止，若干次的时间平均值， $\lambda = 1/\text{MTBF}$ 。平均修复时间(mean time to repair, MTTR)是指设备由故障状态转为工作状态时修理时间的平均值。可用度(availability, A)是指系统、子系统或者设备在开始一项任务时处在指定的可操作或可提交状态的程度，换句话说，可用度就是一个系统处在可工作状态的时间的比例， $A = \text{MTBF} / (\text{MTBF} + \text{MTTR})$ 。若计算一个系统的可用度，可以根据单个单元的失效率计算出单个单元的 MTBF，从而计算出单个单元的可用度，再根据串联和并联模型计算系统整体可用度。参见图 1A，是一种串联模型的示意图。单元 1、单元 2、...、单元 N 是串联的，单元 1 的可用度是  $A_1$ ，单元 2 的可用度是  $A_2$ ，...，单元 N 的可用度是  $A_n$ ，则整个系统的可用度  $A = A_1 * A_2 * \dots * A_n$ 。参见图 1B，是一种并联模型的示意图。并联模型比串联模型复杂一些，并联模型的冗余单元分为主备用冗余单元和负荷分担冗余单元，若图 1B 中的

单元 1、单元 2、...、单元 N 是主备用冗余单元，每个单元的可用度为  $a$ ，则整个系统的可用度  $A=1-\{C*[1-a^N*(N+1-N*a)]/N+(1-C)*(1-a^N)/N\}$ 。其中， $C$  表示倒换率，即成功从主单元倒换到备用单元的概率， $C=C1*C2$ ， $C1$  表示主单元的故障检测率， $C2$  表示备用单元的故障检测率。特别地，若每个单元只有一个冗余单元，即  $N=1$ ，那么公式可以化简为  $A=a+C*a*(1-a)$ 。若图 1B 中的单元 1、单元 2、...、单元 N 是负荷分担冗余单元，每个单元的可用度都为  $a$ ，则整个系统可用度  $A=1-\{C_A*[1-a^N*(N+1-N*a)]/N+(1-C_A)*(1-a^{N+1})/(N+1)\}$ 。其中， $C_A$  表示每个单元的故障检测率。

一般，企业存储系统为了提高系统的可扩展性，通常都是采用互联方式把多控制存储节点互联起来组成一个统一的分布式系统，然后提供存储服务给应用使用。具体地，参见图 2A，图 2A 是一种存储系统的结构示意图。如图 2A 所示，设备 210 是使用存储服务的设备，例如：主机、服务器等，设备 210 与控制设备 220 和控制设备 230 连接。控制设备 220 中包括多个控制器，本实施例中以四个控制器为例说明，这四个控制器分别为控制器 221、控制器 222、控制器 223 和控制器 224，它们之间通过 PCIe 接口相互连接；控制设备 230 也包括多个控制器，本实施例中以四个控制器为例进行说明，这四个控制器分别为控制器 231、控制器 232、控制器 233 和控制器 234，它们之间也通过 PCIe 接口相互连接。为了提高系统可靠性，各个控制器之间都是采用双链路冗余连接。控制设备 220 和控制设备 230 之间也是通过 PCIe 接口进行双链路冗余连接，控制设备 220 和控制设备 230 可以是存储阵列，它们中所包含的控制器可以是中央处理器 CPU 加内存的结构，其中 CPU 调用内存中的程序执行相应的功能，也可以是独立的控制芯片，例如现场可编程阵列 (field-programmable gate array, FPGA)，程序烧录在所述 FPGA 中。控制设备 220 通过 SAS 接口与存储器 240 连接，控制设备 230 通过 SAS 接口与存储器 250 连接。为了提高系统可靠性，控制设备 220 与存储器 240 之间，以及控制设备 230 与存储器 250 之间都是采用双链路冗余连接。存储器 240 和存储器 250 都是存储数据的设备，例如：存储阵列。存储器 240 中包括级联板 241、级联板 242 和  $N$  个存储介质， $N$  为大于 1 的正整数，级联板 241 负责管控该存储器 240，级联板 242 为级联板 241 的冗余级联板，在级联板 241 发生故障时，接替级联板 241 的工作以提高整个系统的可靠性。存储器 250 的结构与存储器 240 一致，包括级联板 251、级联板 252 和  $N$  个存储介质。

在实际应用中，存储系统为了兼顾系统可靠性与控制器的性能，采用了内存双副本技术。即每个控制器的内存中写入的数据，都在属于同一控制设备的另一个控制器的内存中进行了数据备份，如图 2B 所示，数据 A 存储在控制器 221 中，数据 A 的副本数据 A1 存储在控制器 222 中；数据 B 存储在控制器 222 中，数据 B 的副本数据 B1 存储在控制器 221 中；数据 E 存储在控制器 231 中，数据 E 的副本数据 E1 存储在控制器 232 中；数据 H 存储在控制器 234 中，数据 H 的副本数据 H1 存储在控制器 233 中。若控制器 221 发生故障失效，则设备 210 将不能访问控制器 221 中存储的数据，即设备 210 不能访问数据 A 和 B1，由于控制器 222 中存储了数据 A 的副本数据 A1，因此设备 210 可以通过访问控制器 222 进而访问到数据 A1，从而保证业务不中断。

若控制器 221 和控制器 222 同时发生故障，或者控制设备 220 发生故障，设备 210 将

不能访问到数据 A 和数据 A1, 可能会导致业务中断。

此外, 由于每个控制设备中的控制器只能访问与该控制设备相连的存储器, 例如, 控制设备 220 中的控制器 221 只能访问存储器 240 中的数据, 不能访问存储器 250 中的数据, 因为控制设备 220 与存储器 250 没有相连; 同理, 控制设备 230 中的控制器 231 只能访问存储器 250 中的数据, 不能访问存储器 240 中的数据。因此, 当任一控制设备由于掉电或者硬件故障导致失效时, 该控制设备中的控制器上的数据以及与该控制设备连接的存储器上的数据都将不能访问, 从而导致主机上的业务中断或失效。例如, 控制设备 220 发生故障, 那么设备 210 将不能访问控制设备 220 以及存储器 240 中的数据, 导致设备 210 上的业务中断或失效。

需要说明的是, 控制设备与存储器之间可以通过 SAS 链路或 NVMe 链路相连, 每个控制设备还可以与多个存储器相连, 此外, 控制设备与存储器可以部署在同一个物理设备上, 例如, 上述存储器 240 可以和控制设备 220 部署在同一个物理设备上, 当然, 也可以将它们分开部署在不同的物理设备上。

为了解决上述问题, 本申请提供了一种提高存储系统可靠性的方法和相关装置, 保证在控制设备发生故障时, 主机业务的连续性, 提高存储系统的可靠性和可用度。

参见图 3, 图 3 示出了本申请实施例的一种可能的应用场景。在该应用场景中, 主机 310 与控制设备 320 和控制设备 330 连接, 主机 310 可以通过光纤与控制设备 320 和控制设备 330 连接, 也可以通过网络, 例如以太网 (ethernet), 与控制设备 320 和控制设备 330 连接。此外, 控制设备 320 和控制设备 330 中存在两个交换卡 (图 3 中并未画出), 其中一个为主交换卡, 另一个作为备份。控制设备 320 和控制设备 330 中的所有的控制器都需要连接到主交换卡, 该主交换卡用于支持控制设备中的控制器与其它控制设备中的控制器进行通信, 以及与存储器 340 之间进行通信。当主交换卡发生故障时, 备份交换卡接替主交换卡的工作以保证系统的可靠性。控制设备 320 包括多个控制器, 本实施例中以四个为例进行说明, 例如控制器 321、控制器 322、控制器 323 和控制器 324, 控制设备 330 也包括多个控制器, 本实施例中以四个为例进行说明, 例如控制器 331、控制器 332、控制器 333 和控制器 334。控制设备 320 和控制设备 330 中的各个控制器之间、以及控制设备 320 和控制设备 330 之间都是通过 RDMA 通道相互连接, 控制器之间的数据可以不经过 CPU 直接传输到对端, 当然, 也可以通过 PCIe 接口相互连接。控制设备 320 和控制设备 330 利用 RDMA 通道分别与存储器 340 连接, 即存储器 340 对于控制设备 320 和控制设备 330 来说是共享的, 也就是说控制设备 320 和控制设备 330 可以同时访问存储器 340 中的数据, 存储器 340 可以是一个包括 CPU341 和内存 342 的存储服务器 (即一个存储节点), 其中包括 N 个存储介质, 用来存储数据以供控制器访问, N 为大于 1 的正整数。存储器 340 的内存 342 中存储了程序指令, CPU341 调用该程序指令实现对存储器 340 的管控, 例如对存储器 340 进行相关配置、实现存储器 340 的升级等。所有控制设备中的控制器都可以用于接收主机下发的 IO, 并根据 IO 中携带的地址确定处理所述 IO 的目标控制器, 如果所述目标控制器为接收主机下发的 IO 的控制器, 则所述目标控制器直接处理所述 IO, 如果所述目标控制器不是接收主机下发的 IO 的控制器, 则该控制器将所述 IO 转发至所述目标控制器,

由所述目标控制器处理。

需要说明的是，为了提高系统可靠性，所有的连接都是采用双链路冗余，即控制器之间、控制设备之间以及控制设备与存储器之间都是采用两条 RDMA 通道进行连接。此外，主机 310 还可以连接更多的控制设备，控制设备还可以连接多个共享存储器。此外，每个控制器都包括至少一个镜像控制器，且至少有一个镜像控制器属于另一个控制设备，即每个控制器中存储的数据都在其它的控制器中进行了数据备份，而且至少有一个控制器是属于其它控制设备。示例性的，控制器 321 的镜像控制器包括控制器 322 和控制器 331，若控制器 321 中所存储的数据 A 进行了数据备份，那么控制器 322 和控制器 331 则相应的存储了数据 A 的副本数据 A1 和副本数据 A2，控制器 321 和控制器 322 属于控制设备 320，而控制器 331 属于第二控制设备 330，数据 A 相当于跨控制设备进行了存储。每个控制器的镜像控制器预先设置，每个控制器及其对应的镜像控制器构成本实施例提供的提高存储系统可靠性的方法，如图 4 所示，该方法包括但不限于以下步骤：

容易理解，当某个控制设备中的控制器发生故障或者某个控制设备发生故障，由于数据是跨控制设备进行备份的，所以主机仍可以访问其它的控制设备中的备份数据，从而保证业务的连续性。此外，多个控制设备共享存储器，这样即使某个控制设备发生故障，主机仍可以通过其它控制设备中的控制器访问存储器中的数据，可以保证主机上的业务不会中断或失效，有效的提高了系统的可靠性和可用度。

结合图 3 所示的应用场景，下面将结合图 4 描述本实施例提供的提高存储系统可靠性的方法，如图 4 所示，该方法包括但不限于以下步骤：

S410：目标控制器接收写请求。

具体地，目标控制器可以是与主机连接的任意一个控制设备中的任意一个控制器。示例性的，参见上述图 3 中的相关描述，目标控制器可以是控制设备 320 中控制器 321。特别的，当所述目标控制器是当前接收主机下发的 IO 的控制器时，则所述目标控制器直接接收所述主机发送的写请求，当所述目标控制器不是当前接收主机下发的 IO 的控制器时，则所述目标控制器接收的是该控制器转发的写请求。

进一步的，该写请求中包括主机需要写入的待写入数据。

S420：目标控制器将待写入数据写入该目标控制器的内存。

具体地，目标控制器可以是一个包括 CPU 和内存的控制芯片，目标控制器在接收到写请求之后，解析得到该写请求中携带的待写入数据，并将该待写入数据写入到该目标控制器的内存中。

S430：目标控制器将待写入数据写入该目标控制器的镜像控制器的内存。

如图 3 所述，本发明实施例中，每个控制器都设置有镜像控制器，且所述镜像控制器中至少有一个控制器与所述目标控制器不属于一个控制设备。所以在所述目标控制器在将所述待写数据写入内存后，还会将待写入数据写入所述镜像控制器的内存。由于控制器之间都是通过 RDMA 连接的，所以所述目标控制器可以使用 RDMA 将所述待写数据直接写入目标控制器的镜像控制器的内存。

应理解，任意两个控制器之间都通过 RDMA 通道进行数据传输，数据不会经过 CPU，提高了传输效率和控制器的 CPU 性能。

示例性的，参见图 5，图 5 是本申请实施例提供的一种数据多副本写入示意图。如图 5 所示，控制器 521 的镜像控制器是控制器 522 和控制器 532，因此控制器 521 先将数据 A 写入内存之后，再将数据 A 的副本数据 A1 和 A2 分别写入控制器 522 和控制器 532 的内存中；控制器 523 的镜像控制器是控制器 524 和控制器 534，控制器 523 将数据 B 写入内存之后，将数据 B 的副本数据 B1 写入控制器 524 的内存，将数据 B 的副本数据 B2 写入控制器 534 的内存；同理，控制器 531 的镜像控制器是控制器 522 和控制器 532，控制器 531 将数据 C 写入内存之后，将数据 C 的副本数据 C1 写入控制器 532 的内存，将数据 C 的副本数据 C2 写入控制器 522 的内存；控制器 533 的镜像控制器是控制器 524 和控制器 534，控制器 533 将数据 D 写入内存之后，将数据 D 的副本数据 D1 写入控制器 534 的内存，将数据 D 的副本数据 D2 写入控制器 524 的内存。

特别的，在所有相互连接的控制器中存在一个主控制器，该主控制器可以是所有控制器中选出的任意一个控制器，用于管理所有控制器的镜像关系，即每一个控制器存在哪些镜像控制器，其它控制器中都存在该镜像关系的备份。当主控制器发生故障不能工作时，在剩下的控制器中，会推选出一个新的主控制器，用于接替原来主控制器的工作，管理所有控制器的镜像关系。此外，在主控制器所管理的镜像关系中，会避免将过多控制器的镜像控制器设置为同一个控制器，主控制器可以利用打散算法等进行镜像关系的设置，以实现各个控制器之间的负载均衡。

示例性的，参见上述图 5，控制器 521 为目标控制器，控制器 522 和控制器 532 为目标控制器的镜像控制器，控制器 521 将数据 A 写入内存之后，将数据 A 的副本数据 A1 和 A2 分别写入到控制器 522 和控制器 532 的内存。

S440：目标控制器向主机发送写入完成消息。

具体地，目标控制器将待写入数据写入该目标控制器以及该目标控制器的镜像控制器的内存之后，再向主机发送写入完成消息以通知主机，待写入数据已经成功写入目标控制器和目标控制器的镜像控制器的内存。

在一种可能的实现方式中，目标控制器及目标控制器的镜像控制器构成本备份控制器组，该备份控制器组包括 N 个控制器，N 为大于等于 2 的整数。当该备份控制器组中的 M 个控制器故障，且该 M 个控制器包括目标控制器，主控制器从目标控制器的镜像控制器中选择一个控制器作为新的目标控制器，该新的目标控制器用于接收写请求，该主控制器用于从主机接收写请求，并将该写请求转发至新的目标控制器；该主控制器为新的目标控制器建立 M 个新的镜像控制器，并将新的目标控制器中的内存中的数据镜像至该 M 个新的镜像控制器中，M 为大于等于 1 且小于 N 的整数。

具体地，所有控制设备中的控制器两两之间周期性的发送心跳报文，用于感知相互的状态。若某个控制器超过预设时间没有接收到另一个控制器发送的心跳报文，则可以确定该控制器发生了故障，这时，主控制需要重新调整每个控制器的镜像关系。其中，心跳周期以及预设时间可以根据实际需要进行设置，本申请对此不作限定。例如，在上述图 5 中，控制设备 520 和控制设备 530 中的八个控制器之间相互发送心跳报文，若控制器 521 在预设时间内没有接收控制器 522 发送的心跳报文，则控制器 521 可以确定控制器 522 发生故障。

可选的,主控制器可以对管理的控制器的镜像控制器设置优先级,即在目标控制器发生故障时,根据优先级的高低,从目标控制器的镜像控制器中选择优先级高的镜像控制器作为新的目标控制器。可选的,主控制器可以将与目标控制器在同一个控制设备中的镜像控制器设置为高优先级的控制器,当然,也可以通过其它条件进行设置,例如控制器的物理状态、备电关系等,本申请对此不作限定。例如在图 5 中,控制器 522 和控制器 532 都是控制器 521 的镜像控制器,但控制器 522 的优先级高于控制器 532,因此,在控制器 521 发生故障时,主控制器将选择控制器 522 作为新的目标控制器。或者是,主控制器从目标控制器的镜像控制器中随机选择一个控制器作为新的目标控制器,本申请对主控制器选用何种方式确定新的目标控制器不作限定。

进一步的,当包括目标控制器在内的  $M$  个控制器发生故障,主控制器将从剩下的镜像控制器中确定新的目标控制器,并且为该新的目标控制器建立  $M$  个新的镜像控制器,保证备份控制器组中的控制器个数仍旧与之前一样(即  $N$  个控制器),并将该新的目标控制器中的数据镜像至  $M$  个新的镜像控制器中。此外,主机控制器可以从主机接收写请求,然后将该写请求转发至新的目标控制器,新的目标控制器将该写请求中的待写入数据写入新的目标控制器以及该新的目标控制器的镜像控制器(包括  $M$  个新的镜像控制器)的内存。

可选的, $N$  的取值为 3,即目标控制器包括两个镜像控制器,且其中至少有一个镜像控制与目标控制器部署在不同的控制设备。

在一种可能的实现方式中,目标控制器及目标控制器的镜像控制器构成本备份控制器组,该备份控制器组包括  $N$  个控制器, $N$  为大于等于 2 的整数。当该备份控制器组中的  $M$  个控制器故障,且该  $M$  个控制器不包括目标控制器,主控制器为目标控制器建立  $M$  个新的镜像控制器,并将目标控制器中的内存中的数据镜像至该  $M$  个新的镜像控制器中, $M$  为大于等于 1 且小于  $N$  的整数。

具体地,当备份控制器组中不包括目标控制器在内的  $M$  个控制器发生故障,主控制器更新全局镜像关系,即主控制器直接为目标控制器建立  $M$  个新的镜像控制器,保证备份控制器组中的控制器个数仍旧为  $N$ ,并将目标控制器中的数据镜像至  $M$  个新的镜像控制器中。

可选的, $N$  的取值为 3,即目标控制器包括两个镜像控制器,且其中至少有一个镜像控制与目标控制器部署在不同的控制设备。

示例性的,参见图 6A,如图 6A 所示,在控制设备 520 和控制设备 530 正常运行时,控制器 521、控制器 522 和控制器 532 构成一个备份控制器组,它们分别存储了数据 A 以及数据 A 的副本数据 A1 和 A2;控制器 522、控制器 531 和控制器 532 构成另一个备份控制器组,它们分别存储了数据 C 和数据 C 的副本数据 C1 和 C2。当控制器 521 和控制器 522 出现故障时,对于数据 A 来说,其目标控制器为控制器 521,由于控制器 522 和控制器 532 存储了数据 A 的副本数据 A1 和 A2,所以在控制器 521 和控制器 522 同时发生故障时,主控制器将选择控制器 532 作为新的目标控制器,相当于控制器 532 中存储的数据变为数据 A。主控制器重新建立新的镜像关系,在控制器 532 确定为新的目标控制器之后,将数据 A 的副本数据 A1 和 A2 分别镜像至控制器 531 和控制器 523,即控制器 523 和控制器 531 为控制器 532 的镜像控制器。同理,对于数据 C 来说,其目标控制器为控制器 531,由于控制器 522 存储了数据 C 的副本数据 C2,所以在控制器发生故障时,主控制器为目标控制器

531 建立新的镜像控制器以建立新的镜像关系。主控制器将数据 C 的副本数据镜像至控制器 524 中，即控制器 524 和控制器 532 为目标控制器 531 的镜像控制器。

应理解，上述只是以每个目标控制器仅包括两个镜像控制器为例进行说明的，当然，每个目标控制器还可以包括更多个镜像控制器，其具体实现逻辑以及过程与上述一致，为了简洁，在此不再赘述。

可以看出，在两个控制器同时发生故障的情况下，由于目标数据（例如数据 A）进行了数据备份，目标数据的副本数据存储在了三个不同的控制器中，主机仍旧可以访问目标数据，从而保证主机业务的连续性，提高整个存储系统的可靠性和可用度。此外，控制设备中的主控制器还可以重新建立镜像关系，在未发生故障的控制器中建立新的目标控制器或镜像控制器，保证目标数据和该目标数据的副本数据仍旧可以存储在三个不同的控制器中。

在一种可能的实现方式中，当第一控制设备故障时，将第二控制设备中的目标控制器的其中一个镜像控制器作为新的目标控制器，并为该新的目标控制器建立镜像控制器，并将该新的目标控制器的内存中的数据镜像至新的镜像控制器的内存。

具体地，当第一控制设备发生故障时，主控制器需要重新更新全局镜像关系，由于第一控制设备中的每一个控制器在第二控制设备中都存在至少一个镜像控制器，所以主控制器可以从第二控制设备中的镜像控制器中选择一个作为新的目标控制器，然后为该新的目标控制器建立新的镜像控制器，并将新的目标控制器的内存中的数据镜像至新的镜像控制器的内存。在本发明实施例中，可从所述第一控制设备的控制器和第二控制设备中的控制器中选择其中一个作为主控制器，关于分布式系统中选择主控制器的方法为现有技术，在此不再赘述。

示例性的，参见图 6B，如图 6B 所示，在控制设备 520 和控制设备 530 正常运行时，对于数据 A 来说，其目标控制器为控制器 521，控制器 532 存储了数据 A 的副本数据 A2，控制器 532 为控制器 521 的镜像控制器；对于数据 B 来说，其目标控制器为控制器 523，控制器 534 存储了数据 B 的副本数据 B2，控制器 534 为控制器 523 的镜像控制器。当控制设备 520 出现故障时，对于数据 A，主控制器将控制器 532 作为新的目标控制器，控制器 532 中存储的数据变为数据 A，主控制器重新建立新的镜像关系，将数据 A 的副本数据 A1 和 A2 分别镜像至控制器 531 和控制器 533；对于数据 B，主控制器将控制器 534 作为新的目标控制器并重新建立新的镜像关系，将数据 B 的副本数据 B1 和 B2 分别镜像至控制器 531 和控制器 533；对于数据 C，主控制器直接建立新的镜像关系，将数据 C 的副本数据 C2 镜像至控制器 534；对于数据 D，主控制器也是直接建立新的镜像关系，将数据 D 的副本数据 D2 镜像至控制器 532。

应理解，上述只是以两个控制设备为例进行说明的，对于存在更多个控制设备的情况，其具体实现逻辑与上述一致，在此不再赘述。

可以看出，在控制设备出现故障的情况下，由于控制设备中的所有控制器都在另一个控制设备中存在至少一个镜像控制器，即该控制设备中的所有数据都在另一个控制设备中进行了备份，所以主机仍可以通过访问其它的控制设备进而访问到发生故障的控制设备中的数据，保证主机业务的连续性，提高存储系统的可靠性和可用度。

在第一控制设备包括 4 个控制器，第二控制设备包括 4 个控制器的场景下，采用本发明提供的方法可以支持 7 个控制器发生故障，而业务不会中断。具体请参见图 6C 所示的实施例。如图 6C 所示，在控制设备 520 和控制设备 530 正常运行时，控制器 521 存储了数据 A，控制器 522 和控制器 532 分别存储了数据 A 的副本数据 A1 和 A2；控制器 523 存储了数据 B，控制器 524 和控制器 534 分别存储了数据 B 的副本数据 B1 和 B2；控制器 531 存储了数据 C，控制器 532 和控制器 522 分别存储了数据 C 的副本数据 C1 和 C2；控制器 533 存储了数据 D，控制器 534 和控制器 524 分别存储了数据 D 的副本数据 D1 和 D2。当控制器 521 出现故障时，主控制器建立新的镜像关系，将控制器 522 作为数据 A 的目标控制器，并将数据 A 的副本数据 A1 镜像至第三控制器 523，其它控制器的镜像关系保持不变。进一步的，控制器 522 也发生故障，主控制器将控制器 523 作为数据 A 的目标控制器，并将数据 A 的副本数据 A1 镜像至控制器 524；将数据 C 的副本数据 C2 重镜像至控制器 524，其它控制器的镜像关系保持不变。进一步的，控制器 523 也发生故障，主控制器将控制器 524 作为数据 A 的目标控制器，并将数据 A 的副本数据 A1 镜像至控制器 534；主控制器将控制器 524 作为数据 B 的目标控制器，并将数据 B 的副本数据 B1 镜像至控制器 532，其它控制器的镜像关系保持不变。进一步的，控制器 524 也发生故障，主控制器将控制器 534 作为数据 A 的目标控制器，并将数据 A 的副本数据 A1 镜像至控制器 531；主控制器将控制器 532 作为数据 B 的目标控制器，并将数据 B 的副本数据 B1 镜像至控制器 533；主控制器将数据 C 的副本数据 C2 重镜像至控制器 534；主控制器将数据 D 的副本数据 D2 重镜像至控制器 532，其它的镜像关系保持不变。更进一步的，控制器 531 也发生故障，主控制器将数据 A 的副本数据 A1 重镜像至控制器 533；主控制器将数据 B 的副本数据 B1 重镜像至控制器 533；主控制器将控制器 532 作为数据 C 的目标控制器，并将数据 C 的副本数据 C1 镜像至控制器 533，其它镜像关系保持不变。更进一步的，控制器 532 也发生故障，由于这时只剩下两个控制器，即控制器 533 和控制器 534，所以这时主控制器所管理的镜像关系中，每一个目标控制器仅包括一个镜像控制器。主控制器将控制器 533 作为数据 B 的目标控制器，将控制器 533 同时也作为数据 C 和数据 D 的目标控制器，由于控制器 534 中存储有数据 B 的副本数据 B1、数据 C 的副本数据 C1 和数据 D 的副本数据 D1，因此，主控制器不必再建立新的镜像关系。更进一步的，控制器 533 也发生故障，所有的控制器中仅剩控制器 534，所有的数据只能存储在控制器 534 中，而且不能进行数据备份，即主控制器不能再建立新的镜像控制器，控制器 534 作为数据 A、数据 B、数据 C 和数据 D 的目标控制器。

可以看出，在控制器依次发生故障时，通过持续镜像，不断建立更新镜像关系，可以保证主机能够正常访问到目标数据，即使只剩一个控制器工作，仍旧可以支持主机访问目标数据，保证主机业务的连续性。

需要说明的是，上述只是以每个控制器依次故障为例，阐述了控制器的镜像关系的变化，当然，也存在多个控制器同时故障以及多个控制器同时故障和依次故障相结合的应用场景，其具体的镜像关系的变化逻辑与上述一致，为了简洁，在此不再赘述。

另外，为了更好的说明本申请对存储系统可用度的影响，以存储系统包括两个控制设

备, 8 个控制器为例进行说明。请参阅图 7, 图 7 是一种控制设备的可靠性模型的示意图, 如图 7 所示, 该模型中包括 4 个并联的控制器、2 个并联的备电单元、2 个并联的电源模块、2 个并联的风扇、一个背板、2 个并联的管理模块, 它们之间相互串联。分别计算上述图 2B 和图 5 所示场景下的每个控制设备的可用度, 需要说明的是, 在上述两种场景下, 其它器件都是一致的, 连接关系也是一致的, 所以可以仅根据控制器的相关冗余度计算控制设备的可用度。假设每个控制器的可用度为 0.999989496, 利用公式 1 计算图 2B 中每个控制设备的可用度, 公式 1 如下所示:

$$A = (a + C * a * (1 - a)) * (a + C * a * (1 - a)) \quad \text{公式 1}$$

其中,  $a$  表示每个控制器的可用度,  $C$  表示倒换率, 计算得到控制设备的可用度为 0.999998950。

利用公式 2 计算图 5 中每个控制设备的可用度, 公式 2 如下所示:

$$A = 1 - \{C * [1 - a^N * (N + 1 - N * a)] / N + (1 - C) * (1 - a^N) / N\} \quad \text{公式 2}$$

其中,  $N$  表示每个控制器的冗余控制器的个数 (即镜像控制器的个数), 计算得到控制设备的可用度为 0.999999475。

可以看出, 图 5 中的控制设备的可用度要高于图 2B 中的控制设备的可用度, 其控制设备的可靠性更高。

进一步的, 参见图 8, 图 8 是一种存储系统的可靠性模型的示意图。如图 8 所示, 该模型包括 2 个并联的控制设备、 $X$  个串联的硬盘和  $Y$  个串联的盘组, 它们之间相互串联。利用图 8 所示模型计算图 2B 和图 5 中整个存储系统的可用度, 应理解, 图 2B 和图 5 所示的场景中, 其它器件都是一致的, 连接关系也是一致的, 所以可以仅根据控制设备来计算整个存储系统的可用度。值得说明的是, 在图 2B 所示的场景中, 由于每个控制设备之间是相互独立的, 即一个控制设备发生故障, 整个存储系统都将不能工作, 因此控制设备之间应该属于串联模型, 利用公式 3 计算图 2B 中存储系统的可用度, 公式 3 如下所示:

$$A = A1 * A2 \quad \text{公式 3}$$

其中,  $A1$  和  $A2$  表示每个控制设备的可用度, 因为每个控制设备都是一样的, 所以  $A1$  与  $A2$  的值是一样的, 都为 0.999998950, 因此整个存储系统的可用度为 0.999997900。

而在图 5 所示的场景中, 每个控制设备之间是有关联的, 即使一个控制设备发生故障, 另一个控制设备仍能支持主机进行访问, 整个存储系统能够正常工作, 因此控制设备之间应该属于并联模型, 利用公式 4 计算图 5 中存储系统的可用度, 公式 4 如下所示:

$$A = a + C * a * (1 - a) \quad \text{公式 4}$$

其中,  $a$  表示每个控制设备的可用度,  $C$  表示倒换率, 计算得到整个存储系统的可用度为 0.999999475。

可以看出, 图 5 中存储系统的可用度要大大高于图 2B 中存储系统的可用度, 也就是说, 实施本申请实施例所提供的方案, 能够有效提高控制设备以及存储系统的可用度, 提高存储系统的可靠性, 保证主机业务的连续性。

上述详细阐述了本申请实施例的方法, 为了便于更好地实施本申请实施例的上述方案, 相应地, 下面还提供用于配合实施上述方案的相关装置。

参见图 9, 图 9 为本申请实施例提供的一种存储控制系统的结构示意图。如图 9 所示, 存储控制系统 900 包括第一控制设备 9100 和第二控制设备 9200, 所述第一控制设备 9100 包括目标控制器 9110, 目标控制器 9110 包括接收模块 9111 和处理模块 9112。其中,

接收模块 9111, 用于接收写请求, 所述写请求包括待写入数据。

处理模块 9112, 用于将所述待写入数据写入所述目标控制器 9110 的内存;

所述处理模块 9112, 还用于将所述待写入数据写入所述目标控制器 9110 的镜像控制器的内存, 所述镜像控制器中的至少一个控制器属于第二控制设备 9200。

在一种可能的实现中, 所述目标控制器 9110 及所述目标控制器 9110 的镜像控制器构成备份控制器组, 所述备份控制器组包括 N 个控制器, N 为大于等于 2 的整数, 所述存储控制系统 900 包括主控制器 9210, 所述主控制器 9210 为从所述第一控制设备 9100 和第二控制设备 9200 中的控制器中选择的用于管理其他控制器的控制器, 所述主控制器 9210 包括: 故障处理模块 9211, 用于当所述备份控制器组中的 M 个控制器故障, 且所述 M 个控制器包括所述目标控制器时, 主控制器 9210 从所述目标控制器的镜像控制器中选择一个控制器作为新的目标控制器, 所述新的目标控制器用于接收所述写请求; 镜像模块 9212, 用于为所述新的目标控制器建立 M 个新的镜像控制器, 并将所述新的目标控制器中的内存中的数据镜像至所述 M 个新的镜像控制器中, M 为大于等于 1 且小于 N 的整数。

在一种可能的实现中, 所述 N 等于 3。

在一种可能的实现中, 当所述第一控制设备 9100 故障时, 将所述第二控制设备 9200 中的所述目标控制器 9110 的其中一个镜像控制器作为新的目标控制器, 并为所述新的目标控制器建立镜像控制器, 并将所述新的目标控制器的内存中的数据镜像至新的镜像控制器的内存。

应理解, 上述目标控制器和主控制器的结构以及提高存储控制系统可靠性的处理过程仅作为一种示例, 不应构成具体的限定, 可以根据需要对目标控制器和主控制器中的各个模块进行增加、减少或合并。此外, 目标控制器和主控制器中的各个模块的操作和/或功能分别为了实现上述图 4 所描述的方法的相应流程, 为了简洁, 在此不再赘述。

参见图 10, 图 10 为本申请实施例提供的又一种存储控制系统的结构示意图。如图 10 所示, 存储控制系统 100 包括第一控制设备 1100 和第二控制设备 1200, 所述第一控制设备 1100 包括目标控制器 1110, 目标控制器 1110 包括: 处理器 1111, 通信接口 1112 以及存储器 1113, 所述处理器 1111、通信接口 1112 以及存储器 1113 通过内部总线 1114 连接。

在一种可能的实现中, 所述存储控制系统 100 包括主控制器 1210, 所述主控制器 1210 为从所述第一控制设备 1100 和第二控制设备 1200 中的控制器中选择的用于管理其他控制器的控制器, 所述主控制器 1210 包括: 处理器 1211, 通信接口 1212 以及存储器 1213, 所述处理器 1211、通信接口 1212 以及存储器 1213 通过内部总线 1214 连接。

应理解, 接收模块 9111 所实现的功能可以用通信接口 1112 具体实现, 处理模块 9112 所实现的功能可以由处理器 1111 具体实现, 故障处理模块 9211 和镜像模块 9212 所实现的功能可以由处理器 1211 具体实现。

存储器 1113 可以用于存储程序代码和数据, 以便于处理器 1111 调用存储器 1113 中存

储的程序代码和数据以实现上述接收模块 9111 和/或处理模块 9112 的功能。处理器 1111 可以调用存储器 1113 存储的程序代码和数据以实现上述图 4 所描述的方法的相应流程。

存储器 1213 可以用于存储程序代码和数据,以便于处理器 1211 调用存储器 1213 中存储的程序代码和数据以实现上述故障处理模块 9211 和/或镜像模块 9212 的功能。处理器 1211 可以调用存储器 1213 存储的程序代码和数据以实现上述图 4 所描述的方法的相应流程。

需要说明的是,目标控制器 1110 和主控制器 1210 可以是单独采用内置处理逻辑的逻辑器件来实现,例如 FPGA 或数字信号处理器(digital signal processor, DSP)等,此外,目标控制器 1110 和主控制器 1210 可能包含相比于图 10 展示的更多或者更少的组件,或者有不同的组件配置方式。

应理解,上述图 10 所示的目标控制器 1110 和主控制器 1210 提高存储控制系统可靠性的处理过程可以参照上述图 1 至图 8 的相关描述,为了简洁,在此不再赘述。

本申请实施例还提供了一种计算机可读存储介质,该计算机可读存储介质中存储有指令,当其在计算机或处理器上运行时,使得计算机或处理器执行上述实施例中记载的任意一种的部分或全部步骤。

本申请实施例还提供一种计算机程序,该计算机程序包括指令,当该计算机程序被计算机执行时,使得计算机可以执行任意一种提高存储系统可靠性方法的部分或全部步骤。

在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中未详述的部分,可以参见其它实施例的相关描述。

需要说明的是,对于前述的各方法实施例,为了简单描述,故将其都表述为一系列的动作组合,但是本领域技术人员应该知悉,本申请并不受所描述的动作顺序的限制,因为依据本申请,某些步骤可能可以采用其它顺序或者同时进行。其次,本领域技术人员也应该知悉,说明书中所描述的实施例均属于优选实施例,所涉及的动作和模块并不一定是本申请所必须的。

在本申请所提供的几个实施例中,应该理解到,所揭露的装置,可通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如上述单元的划分,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,例如多个单元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,装置或单元的间接耦合或通信连接,可以是电性或其它的形式。

上述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外,在本申请各实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用软件功能单元的形式实现。

## 权 利 要 求

1、一种提高存储系统可靠性的方法，所述存储系统包括第一控制设备和第二控制设备，其特征在于，所述方法包括：

目标控制器接收写请求，所述写请求包括待写入数据，所述目标控制器属于所述第一控制设备；

所述目标控制器将所述待写入数据写入所述目标控制器的内存；

所述目标控制器将所述待写入数据写入所述目标控制器的镜像控制器的内存，所述镜像控制器中的至少一个控制器属于所述第二控制设备。

2、如权利要求1所述的方法，其特征在于，所述第一控制设备和所述第二控制设备连接至存储设备，所述方法还包括：

当所述第一控制设备故障，通过所述第二控制设备访问所述存储设备。

3、如权利要求1或2所述的方法，其特征在于，所述目标控制器及所述目标控制器的镜像控制器构成备份控制器组，所述备份控制器组包括N个控制器，N为大于等于2的整数，

所述方法还包括：

当所述备份控制器组中的M个控制器故障，且所述M个控制器包括所述目标控制器时，主控制器从所述目标控制器的镜像控制器中选择一个控制器作为新的目标控制器，所述新的目标控制器用于接收所述写请求，所述主控制器为从所述第一控制设备和第二控制设备中的控制器中选择的用于管理其它控制器的控制器；

所述主控制器为所述新的目标控制器建立M个新的镜像控制器，并将所述新的目标控制器中的内存中的数据镜像至所述M个新的镜像控制器中，M为大于等于1且小于N的整数。

4、如权利要求1或2所述的方法，其特征在于，所述目标控制器及所述目标控制器的镜像控制器构成备份控制器组，所述备份控制器组包括N个控制器，N为大于等于2的整数，

所述方法还包括：

当所述备份控制器组中的M个控制器故障，且所述M个控制器不包括所述目标控制器时，主控制器为所述目标控制器建立M个新的镜像控制器，并将所述目标控制器中的内存中的数据镜像至所述M个新的镜像控制器中，所述主控制器为从所述第一控制设备和第二控制设备中的控制器中选择的用于管理其它控制器的控制器，M为大于等于1且小于N的整数。

5、如权利要求3或4所述的方法，其特征在于，所述N等于3。

6、如权利要求 1 或 2 所述的方法，其特征在于，当所述第一控制设备故障时，将所述第二控制设备中的所述目标控制器的其中一个镜像控制器作为新的目标控制器，并为所述新的目标控制器建立镜像控制器，并将所述新的目标控制器的内存中的数据镜像至新的镜像控制器的内存。

7、一种存储控制系统，所述存储控制系统包括第一控制设备和第二控制设备，其特征在于，所述第一控制设备包括目标控制器，所述目标控制器包括：

接收模块，用于接收写请求，所述写请求包括待写入数据；

处理模块，用于将所述待写入数据写入所述目标控制器的内存；

所述处理模块，还用于将所述待写入数据写入所述目标控制器的镜像控制器的内存，所述镜像控制器中的至少一个控制器属于所述第二控制设备。

8、如权利要求 7 所述的存储控制系统，其特征在于，所述目标控制器及所述目标控制器的镜像控制器构成本备控制器组，所述本备控制器组包括 N 个控制器，N 为大于等于 2 的整数，所述存储控制系统包括主控制器，所述主控制器为从所述第一控制设备和第二控制设备中的控制器中选择的用于管理其他控制器的控制器，所述主控制器包括：

故障处理模块，用于当所述本备控制器组中的 M 个控制器故障，且所述 M 个控制器包括所述目标控制器时，主控制器从所述目标控制器的镜像控制器中选择一个控制器作为新的目标控制器，所述新的目标控制器用于接收所述写请求；

镜像模块，用于为所述新的目标控制器建立 M 个新的镜像控制器，并将所述新的目标控制器中的内存中的数据镜像至所述 M 个新的镜像控制器中，M 为大于等于 1 且小于 N 的整数。

9、如权利要求 7 所述的存储控制系统，其特征在于，所述目标控制器及所述目标控制器的镜像控制器构成本备控制器组，所述本备控制器组包括 N 个控制器，N 为大于等于 2 的整数，所述存储控制系统还包括主控制器，所述主控制器为从所述第一控制设备和第二控制设备中的控制器中选择的用于管理其他控制器的控制器，所述主控制器包括：

故障处理模块，当所述本备控制器组中的 M 个控制器故障，且所述 M 个控制器不包括所述目标控制器时，主控制器为所述目标控制器建立 M 个新的镜像控制器；

镜像模块，用于将所述目标控制器中的内存中的数据镜像至所述 M 个新的镜像控制器中，M 为大于等于 1 且小于 N 的整数。

10、如权利要求 8 或 9 所述的存储控制系统，其特征在于，所述 N 等于 3。

11、如权利要求 7 所述的存储控制系统，其特征在于，当所述第一控制设备故障时，将所述第二控制设备中的所述目标控制器的其中一个镜像控制器作为新的目标控制器，并为所述新的目标控制器建立镜像控制器，并将所述新的目标控制器的内存中的数据镜像至

新的镜像控制器的内存。

12、一种存储控制系统，所述存储控制系统包括第一控制设备和第二控制设备，其特征在于，所述第一控制设备包括目标控制器，所述目标控制器包括处理器和存储器，所述处理器执行所述存储器存储的计算机指令，使得所述目标控制器执行权利要求 1 至 6 任一项所述的方法。

13、如权利要求 12 所述的存储控制系统，其特征在于，所述存储控制系统包括主控制器，所述主控制器为从所述第一控制设备和第二控制设备中的控制器中选择的用于管理其他控制器的控制器，所述主控制器包括处理器和存储器，所述处理器执行所述存储器存储的计算机指令，使得所述主控制器执行权利要求 1 至 6 任一项所述的方法。

14、一种计算机存储介质，其特征在于，所述计算机存储介质存储有计算机程序，当所述计算机程序被处理器执行时实现权利要求 1 至 6 任一项所述的方法。

15、一种计算机程序产品，所述计算机程序产品包括计算机指令，当该计算机程序被计算机执行时，使得计算机可以执行权利要求 1 至 6 任一项所述的方法。