(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 114327903 A
(43)申请公布日 2022.04.12

(21)申请号 202111660562.5

(22)申请日 2021.12.30

(71)申请人 苏州浪潮智能科技有限公司
地址 215100 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

(72)发明人 方浩

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 吴磊

(51)Int.Cl.
G06F 9/50(2006.01)
G06F 3/06(2006.01)

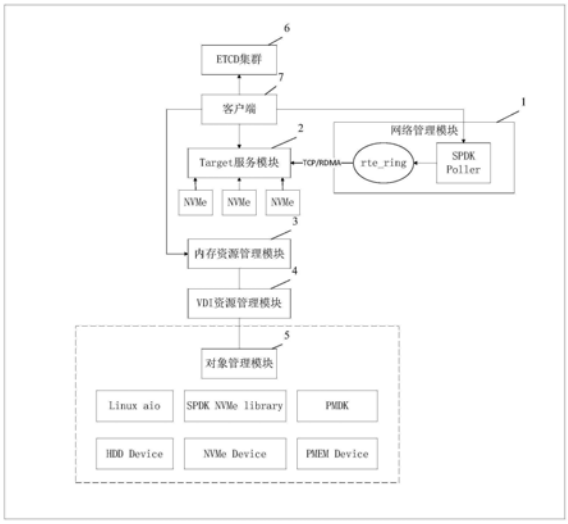
权利要求书2页 说明书6页 附图3页

(54)发明名称

NVMe-oF管理系统、资源配置方法和IO读写方法

(57)摘要

本申请公开了一种NVMe-oF管理系统，应用于分布式存储系统，包括网络管理模块、target服务模块、内存资源管理模块、VDI资源管理模块、对象管理模块、ETCD集群和客户端，由网络管理模块获取客户端的信息，并利用网络管理模块、target访问模块、内存资源管理模块、VDI资源管理模块、对象管理模块和ETCD集群以NVMe-oF协议作为基础提供资源服务，实现了分布式存储领域的NVMe-oF的架构，具有较强的架构优势。相应的，本申请还公开了具有相同技术效果的一种资源配置方法和一种IO读写方法。



CN 114327903 A

1.一种NVMe-oF管理系统,应用于分布式存储系统,其特征在于,包括网络管理模块、target服务模块、内存资源管理模块、VDI资源管理模块、对象管理模块、ETCD集群和客户端,其中:

所述网络管理模块,用于轮询检测连接的所述客户端发送的网络数据资源,并根据所述网络数据资源的target资源将所述网络数据资源发送到相应的NVMe target服务;

所述target服务模块,用于启动单个线程以单独管理每个所述NVMe target服务;

所述内存资源管理模块,用于提供与每个所述NVMe target服务一一对应的NVMe bdev资源,并发送到所述VDI资源管理模块;

所述VDI资源管理模块,用于管理VDI资源,所述VDI资源包括多个内存资源;

所述对象管理模块,用于将分布式系统的存储资源确定为抽象的标准数量值的所述内存资源,还用于管理设备和IO业务数据;

所述ETCD集群,用于存储所述NVMe target服务和NVMe资源信息。

2.根据权利要求1所述NVMe-oF管理系统,其特征在于,所述网络管理模块还用于通过rte_ring机制分别处理不同的连接请求。

3.根据权利要求1所述NVMe-oF管理系统,其特征在于,所述VDI资源管理模块具体用于提供NVMe内存资源或LUN内存资源。

4.根据权利要求1所述NVMe-oF管理系统,其特征在于,所述对象管理模块具体用于:

利用SPDK提供的NVMe驱动NVMe硬盘设备和所述IO业务数据;

或,利用PMDK开发库管理PMEM设备和所述IO业务数据;

或,利用Linux自带的aio管理所述IO业务数据。

5.根据权利要求1至4任一项所述NVMe-oF管理系统,其特征在于,

所述网络管理模块具体用于:利用SPDK提供的poller机制,轮询检测所述客户端发送的网络数据资源,并根据所述网络数据资源的target资源将所述网络数据资源发送到相应的NVMe target服务;

所述target服务模块具体用于,利用所述SPDK提供的target服务管理启动单个线程以单独管理每个所述NVMe target服务;

所述内存资源管理模块具体用于,利用所述SPDK提供的用户态NVMe驱动,提供与每个所述NVMe target服务一一对应的NVMe bdev资源,并发送到所述VDI资源管理模块。

6.一种资源配置方法,其特征在于,应用于如权利要求1至6任一项所述NVMe-oF管理系统,该资源配置方法包括:

当接收到客户端的请求,相应执行以下操作:

向ETCD集群提供的数据库写入对应的资源信息,并记录配置状态为配置中;

向正在运行的、对应NVMe target服务的线程发送相应的指令;

创建新的所述NVMe target服务或NVMe资源;

更新所述配置状态为配置完成。

7.根据权利要求6所述资源配置方法,其特征在于,所述向ETCD集群提供的数据库写入对应的资源信息的过程,包括:

向ETCD集群提供的数据库写入对应的资源信息并以key/value的形式存储。

8.一种IO读写方法,其特征在于,应用于如权利要求1至6任一项所述NVMe-oF管理系

统，该IO读写方法包括：

当网络管理模块收到客户端的请求，确定对应的NVMe target服务和待访问NVMe磁盘设备；

根据所述待访问NVMe磁盘设备，按照NVMe协议解析内存资源信息；所述内存资源信息包括IO数据、访问步长和访问长度；

根据所述内存资源信息，确定对应的VDI资源及待访问的内存资源；

将所述IO数据发送到对象管理模块，以使所述对象管理模块根据待访问的所述内存资源将所述IO数据下发到对应的物理磁盘。

9.根据权利要求8所述IO读写方法，其特征在于，还包括：

当所述IO数据下发到对应的物理磁盘，所述对象管理模块收到写入成功信息并返回给内存资源管理模块；

当内存资源管理模块收到所述请求对应的所有副本的所述写入成功信息后，将所述写入成功信息返回给所述网络管理模块；

当所述网络管理模块收到所述写入成功信息，将所述写入成功信息返回到所述客户端。

10.根据权利要求8所述IO读写方法，其特征在于，所述根据所述内存资源信息，确定对应的VDI资源及待访问的内存资源的过程，包括：

根据所述内存资源信息，确定对应的VDI资源；

根据所述VDI资源及所述访问步长和所述访问长度，计算待访问的内存资源。

NVMe-oF管理系统、资源配置方法和IO读写方法

技术领域

[0001] 本发明涉及分布式存储领域，特别涉及一种NVMe-oF的分布式管理系统、资源配置方法和IO读写方法。

背景技术

[0002] 当前，非易失性快速内存(Non-Volatile Memory express，简称NVMe)协议已经较为成熟，属于业内常见的一种协议，作为其扩展，NVMe-oF(NVMe over Fabrics)是一种新的网络协议，实现了NVMe标准在PCIe总线上的扩展，从而能够替代SCSI协议在SAN环境中的应用，包括FC、infiniband、RoCE v2、Iwapp和TCP等不同的实现，据此，部分领域已经出现了NVMe-oF协议的应用，但在分布式存储领域，如何设计一个NVMe-oF的架构方案，是目前本领域技术人员需要解决的问题。

发明内容

[0003] 有鉴于此，本发明的目的在于提供一种NVMe-oF的分布式管理系统、资源配置方法和IO读写方法，以在分布式存储领域提供NVMe-oF的架构。其具体方案如下：

[0004] 一种NVMe-oF管理系统，应用于分布式存储系统，包括网络管理模块、target服务模块、内存资源管理模块、VDI资源管理模块、对象管理模块、ETCD集群和客户端，其中：

[0005] 所述网络管理模块，用于轮询检测连接的所述客户端发送的网络数据资源，并根据所述网络数据资源的target资源将所述网络数据资源发送到相应的NVMe target服务；

[0006] 所述target服务模块，用于启动单个线程以单独管理每个所述NVMe target服务；

[0007] 所述内存资源管理模块，用于提供与每个所述NVMe target服务一一对应的NVMe bdev资源，并发送到所述VDI资源管理模块；

[0008] 所述VDI资源管理模块，用于管理VDI资源，所述VDI资源包括多个内存资源；

[0009] 所述对象管理模块，用于将分布式系统的存储资源确定为抽象的标准数量值的所述内存资源，还用于管理设备和IO业务数据；

[0010] 所述ETCD集群，用于存储所述NVMe target服务和NVMe资源信息。

[0011] 优选的，所述网络管理模块还用于通过rte_ring机制分别处理不同的连接请求。

[0012] 优选的，所述VDI资源管理模块具体用于提供NVMe内存资源或LUN内存资源。

[0013] 优选的，所述对象管理模块具体用于：

[0014] 利用SPDK提供的NVMe驱动NVMe硬盘设备和所述IO业务数据；

[0015] 或，利用PMDK开发库管理PMEM设备和所述IO业务数据；

[0016] 或，利用Linux自带的aio管理所述IO业务数据。

[0017] 优选的，所述网络管理模块具体用于：利用SPDK提供的poller机制，轮询检测所述客户端发送的网络数据资源，并根据所述网络数据资源的target资源将所述网络数据资源发送到相应的NVMe target服务；

[0018] 所述target服务模块具体用于，利用所述SPDK提供的target服务管理启动单个线

程以单独管理每个所述NVMe target服务；

[0019] 所述内存资源管理模块具体用于，利用所述SPDK提供的用户态NVMe驱动，提供与每个所述NVMe target服务一一对应的NVMe bdev资源，并发送到所述VDI资源管理模块。

[0020] 相应的，本申请还公开了一种资源配置方法，应用于如上文任一项所述NVMe-oF管理系统，该资源配置方法包括：

[0021] 当接收到客户端的请求，相应执行以下操作：

[0022] 向ETCD集群提供的数据库写入对应的资源信息，并记录配置状态为配置中；

[0023] 向正在运行的、对应NVMe target服务的线程发送相应的指令；

[0024] 创建新的所述NVMe target服务或NVMe资源；

[0025] 更新所述配置状态为配置完成。

[0026] 优选的，所述向ETCD集群提供的数据库写入对应的资源信息的过程，包括：

[0027] 向ETCD集群提供的数据库写入对应的资源信息并以key/value的形式存储。

[0028] 相应的，本申请还公开了一种IO读写方法，应用于如上文任一项所述NVMe-oF管理系统，该IO读写方法包括：

[0029] 当网络管理模块收到客户端的请求，确定对应的NVMe target服务和待访问NVMe磁盘设备；

[0030] 根据所述待访问NVMe磁盘设备，按照NVMe协议解析内存资源信息；所述内存资源信息包括IO数据、访问步长和访问长度；

[0031] 根据所述内存资源信息，确定对应的VDI资源及待访问的内存资源；

[0032] 将所述IO数据发送到对象管理模块，以使所述对象管理模块根据待访问的所述内存资源将所述IO数据下发到对应的物理磁盘。

[0033] 优选的，所述IO读写方法还包括：

[0034] 当所述IO数据下发到对应的物理磁盘，所述对象管理模块收到写入成功信息并返回给内存资源管理模块；

[0035] 当内存资源管理模块收到所述请求对应的所有副本的所述写入成功信息后，将所述写入成功信息返回给所述网络管理模块；

[0036] 当所述网络管理模块收到所述写入成功信息，将所述写入成功信息返回到所述客户端。

[0037] 优选的，所述根据所述内存资源信息，确定对应的VDI资源及待访问的内存资源的过程，包括：

[0038] 根据所述内存资源信息，确定对应的VDI资源；

[0039] 根据所述VDI资源及所述访问步长和所述访问长度，计算待访问的内存资源。

[0040] 本申请公开了一种NVMe-oF管理系统，应用于分布式存储系统，包括网络管理模块、target服务模块、内存资源管理模块、VDI资源管理模块、对象管理模块、ETCD集群和客户端，由网络管理模块获取客户端的信息，并利用网络管理模块、target访问模块、内存资源管理模块、VDI资源管理模块、对象管理模块和ETCD集群以NVMe-oF协议作为基础提供资源服务，实现了分布式存储领域的NVMe-oF的架构，具有较强的架构优势。

附图说明

[0041] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的实施例,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,还可以根据提供的附图获得其他的附图。

[0042] 图1为本发明实施例中一种NVMe-Of管理系统的结构分布图;

[0043] 图2为本发明实施例中一种资源配置方法的步骤流程图;

[0044] 图3为本发明实施例中一种IO读写方法的步骤流程图。

具体实施方式

[0045] 下面将结合本发明实施例中的附图,对本发明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发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,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,都属于本发明保护的范围。

[0046] 部分领域已经出现了NVMe-oF协议的应用,但在分布式存储领域,如何设计一个NVMe-oF的架构方案,是目前本领域技术人员需要解决的问题。

[0047] 本申请公开了一种NVMe-oF管理系统,应用于分布式存储系统,由网络管理模块获取客户端的信息,并利用网络管理模块、target访问模块、内存资源管理模块、VDI资源管理模块、对象管理模块和ETCD集群以NVMe-oF协议作为基础提供资源服务,实现了分布式存储领域的NVMe-oF的架构,具有较强的架构优势。

[0048] 本发明实施例公开了一种NVMe-oF管理系统,应用于分布式存储系统,包括网络管理模块1、target服务模块2、内存资源管理模块3、VDI资源管理模块4、对象管理模块5、ETCD集群6和客户端7,其中:

[0049] 网络管理模块Net Manager 1,用于轮询检测连接的客户端7发送的网络数据资源,并根据网络数据资源的target资源将网络数据资源发送到相应的NVMe target服务;

[0050] target服务模块NVMe Target Service 2,用于启动单个线程以单独管理每个NVMe target服务;

[0051] 内存资源管理模块3,用于提供与每个NVMe target服务一一对应的NVMe bdev资源,并发送到VDI资源管理模块4;

[0052] VDI资源管理模块VDI Resource Manager 4,用于管理VDI(Virtual Desktop Infrastructure,虚拟桌面基础架构)资源,VDI资源包括多个内存资源;

[0053] 对象管理模块Object Storage Manager 5,用于将分布式系统的存储资源确定为抽象的标准数量值的内存资源,还用于管理设备和IO业务数据;

[0054] ETCD集群6,用于存储NVMe target服务和NVMe资源信息。

[0055] 可以理解的是,ETCD集群作为分布式的KEY/VALUE数据库,存储的资源主要包括target服务和NVMe资源信息,NVMe资源信息包括NVMe盘大小、名称等NVMe磁盘的属性特征。

[0056] 可以理解的是,VDI资源管理模块4是底层分布式存储资源的对外表现形式,具体可提供NVMe内存资源或LUN内存资源,当它被赋予NVMe的相关属性也即NVMe SPEC标准时,提供NVMe内存资源,当它被赋予LUN(Logical Unit Number,逻辑单元号)的相关属性,则可

向外提供LUN内存资源。

[0057] 进一步的,对象管理模块5具体用于:

[0058] 利用SPDK提供的NVMe驱动NVMe硬盘设备和IO业务数据;

[0059] 或,利用PMDK开发库管理PMEM设备和IO业务数据;

[0060] 或,利用Linux自带的aio管理IO业务数据,相应管理HDD设备。

[0061] 可以理解的是,本实施例中的部分服务或机制由SPDK(Storage Performance Development Kit,存储高性能开发组件)或DPDK(Data Plane Development Kit,数据平面开发套件)提供,SPDK是由Intel发起的、将NVMe SSD作为存储后端的应用软件加速库,它的主要核心目的是实现用户态的、异步、无锁、轮询的方式的NVMe驱动,同时提供NVMe-oF的Target服务的实现。具体的:

[0062] 网络管理模块1具体用于:利用SPDK提供的poller机制,轮询检测客户端7发送的网络数据资源,并根据网络数据资源的target资源将网络数据资源发送到相应的NVMe target服务;该轮询动作需要占用一个CPU;可以理解的是,poller机制能够进行不同客户端

[0063] target服务模块2具体用于,利用SPDK提供的target服务管理启动单个线程以单独管理每个NVMe target服务;

[0064] 内存资源管理模块3具体用于,利用SPDK提供的用户态NVMe驱动,提供与每个NVMe target服务一一对应的NVMe bdev资源,并发送到VDI资源管理模块4。

[0065] 网络管理模块1还用于通过rte_ring机制分别处理不同的连接请求,该rte_ring机制由DPDK提供。

[0066] 可以理解的是,本实施例中NVMe-oF管理系统的应用包括资源配置和IO读写,客户端的请求或指令将以命令行的形式呈现,资源配置包括但不限于target服务的创建、删除和权限控制等功能、NVMe虚拟磁盘的创建、删除、扩容等功能。

[0067] 可以理解的是,本实施例中NVMe-oF管理系统在分布式存储系统可支持NVMe-oF协议和iSCSI协议,这两种协议使用同一个分布式存储架构,利用SPDK提供的组件将所有模块联结起来,实现整体NVMe-oF的架构优势。

[0068] 本申请公开了一种NVMe-oF管理系统,应用于分布式存储系统,包括网络管理模块、target服务模块、内存资源管理模块、VDI资源管理模块、对象管理模块、ETCD集群和客户端,由网络管理模块获取客户端的信息,并利用网络管理模块、target访问模块、内存资源管理模块、VDI资源管理模块、对象管理模块和ETCD集群以NVMe-oF协议作为基础提供资源服务,实现了分布式存储领域的NVMe-oF的架构,具有较强的架构优势。

[0069] 相应的,本申请实施例还公开了一种资源配置方法,应用于如上文任一项所述NVMe-oF管理系统,参见图2所示,该资源配置方法包括:

[0070] 当接收到客户端的请求,相应执行以下操作:

[0071] S11:向ETCD集群提供的数据库写入对应的资源信息,并记录配置状态为配置中;

[0072] S12:向正在运行的、对应NVMe target服务的线程发送相应的指令;

[0073] S13:创建新的NVMe target服务或NVMe资源;

[0074] S14:更新配置状态为配置完成。

[0075] 进一步的,步骤S11中向ETCD集群提供的数据库写入对应的资源信息的过程,包

括：

[0076] 向ETCD集群提供的数据库写入对应的资源信息并以key/value的形式存储。

[0077] 可以理解的是，本实施例中客户端的请求以命令行的形式呈现，资源配置的实现包括但不限于target服务的创建、删除和权限控制等功能、NVMe虚拟磁盘的创建、删除、扩容等功能。

[0078] 可以理解的是，本实施例中有关NVMe-oF管理系统的细节内容，可以参照上文实施例中的相关描述，此处不再赘述。

[0079] 其中，本实施例中资源配置方法具有与上文实施例中NVMe-oF管理系统相同的技术效果，此处不再赘述。

[0080] 相应的，本申请实施例还公开了一种IO读写方法，应用于如上文任一项NVMe-oF管理系统，参见图3所示，该IO读写方法包括：

[0081] S21：当网络管理模块收到客户端的请求，确定对应的NVMe target服务和待访问NVMe磁盘设备；

[0082] S22：根据待访问NVMe磁盘设备，按照NVMe协议解析内存资源信息；内存资源信息包括IO数据、访问步长和访问长度；

[0083] S23：根据内存资源信息，确定对应的VDI资源及待访问的内存资源；

[0084] S24：将IO数据发送到对象管理模块，以使对象管理模块根据待访问的内存资源将IO数据下发到对应的物理磁盘。

[0085] 其中，步骤S23具体包括：

[0086] 根据内存资源信息，确定对应的VDI资源；

[0087] 根据VDI资源及访问步长和访问长度，计算待访问的内存资源。

[0088] 可以理解的是，至步骤S24已完成IO数据的下发，下发后还可向客户端返回相应的写入成功信息，因此，IO读写方法还包括：

[0089] S25：当IO数据下发到对应的物理磁盘，对象管理模块收到写入成功信息并返回给内存资源管理模块；

[0090] S26：当内存资源管理模块收到请求对应的所有副本的写入成功信息后，将写入成功信息返回给网络管理模块；

[0091] S27：当网络管理模块收到写入成功信息，将写入成功信息返回到客户端。

[0092] 可以理解的是，本实施例中有关NVMe-oF管理系统的细节内容，可以参照上文实施例中的相关描述，此处不再赘述。

[0093] 其中，本实施例中IO读写方法具有与上文实施例中NVMe-oF管理系统相同的技术效果，此处不再赘述。

[0094] 最后，还需要说明的是，在本文中，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，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0095] 以上对本发明所提供的一种NVMe-oF的分布式管理系统、资源配置方法和IO读写方法进行了详细介绍，本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本发明的限制。

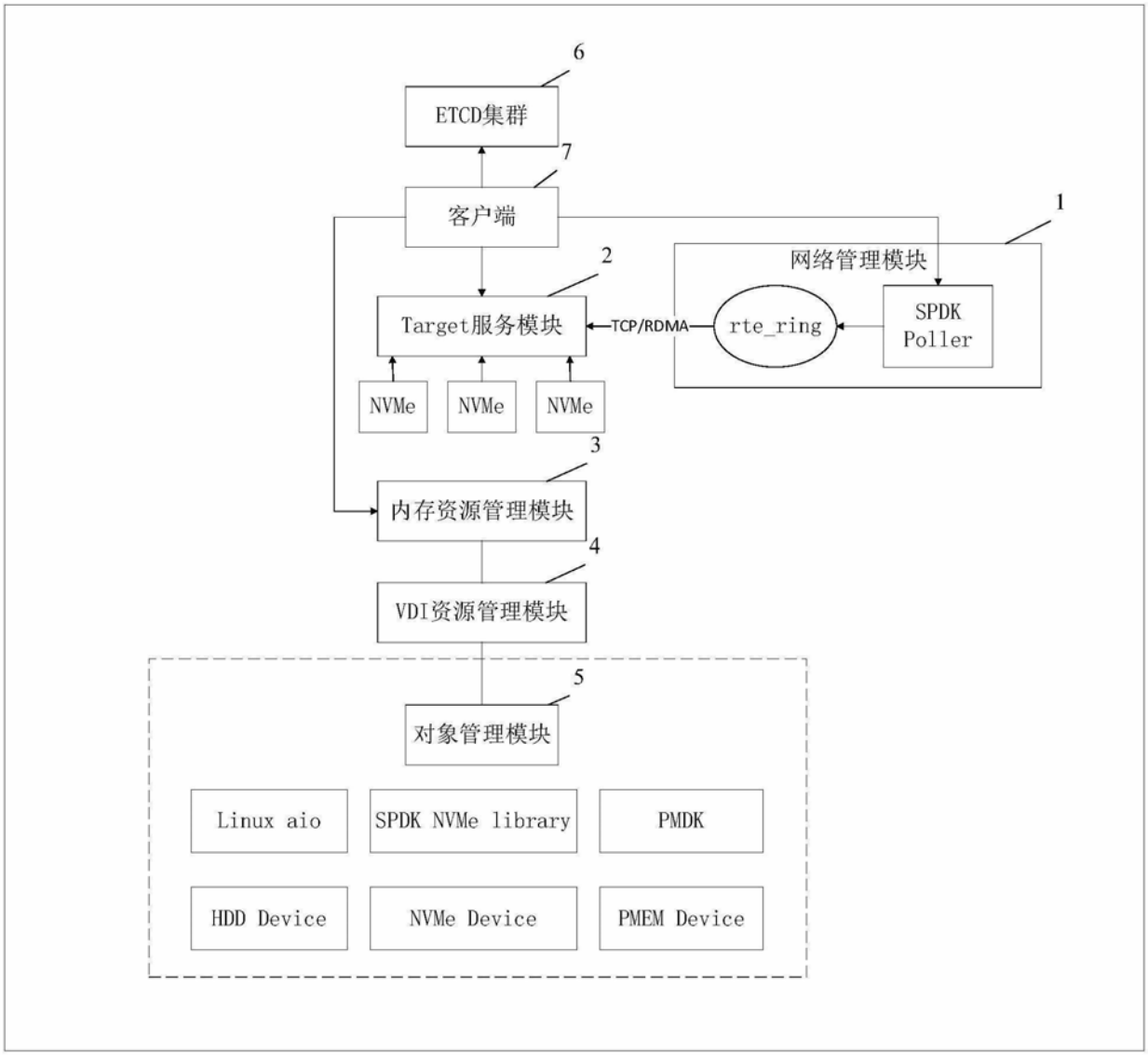

6
ETCD集群
7
客户端
1
网络管理模块
2
Target服务模块
TCP/RDMA
rte_ring
SPDK
Poller
NVMe
NVMe
NVMe
3
内存资源管理模块
4
VDI资源管理模块
5
对象管理模块
Linux aio
SPDK NVMe library
PMDK
HDD Device
NVMe Device
PMEM Device

图1

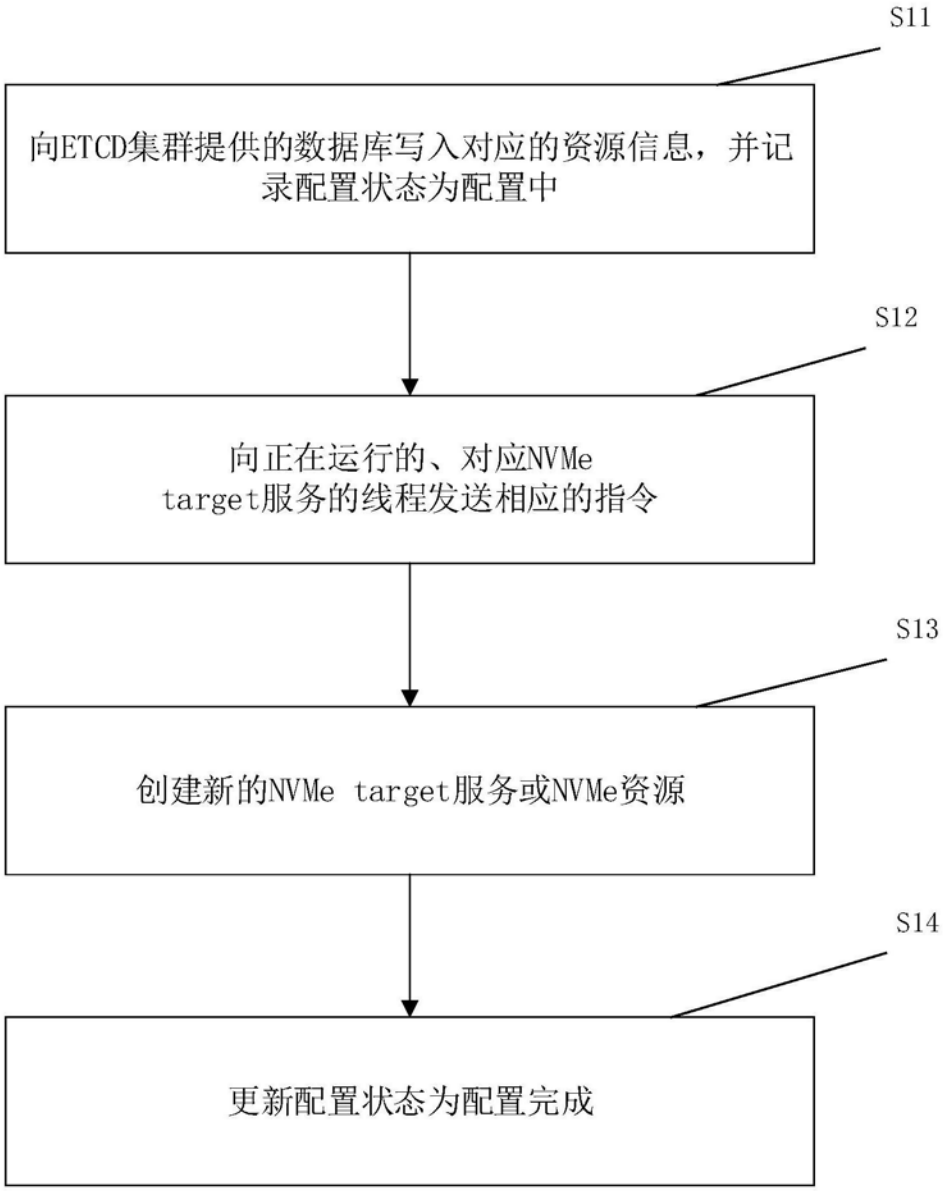

S11
向ETCD集群提供的数据库写入对应的资源信息，并记录配置状态为配置中
S12
向正在运行的、对应NVMe target服务的线程发送相应的指令
S13
创建新的NVMe target服务或NVMe资源
S14
更新配置状态为配置完成

图2

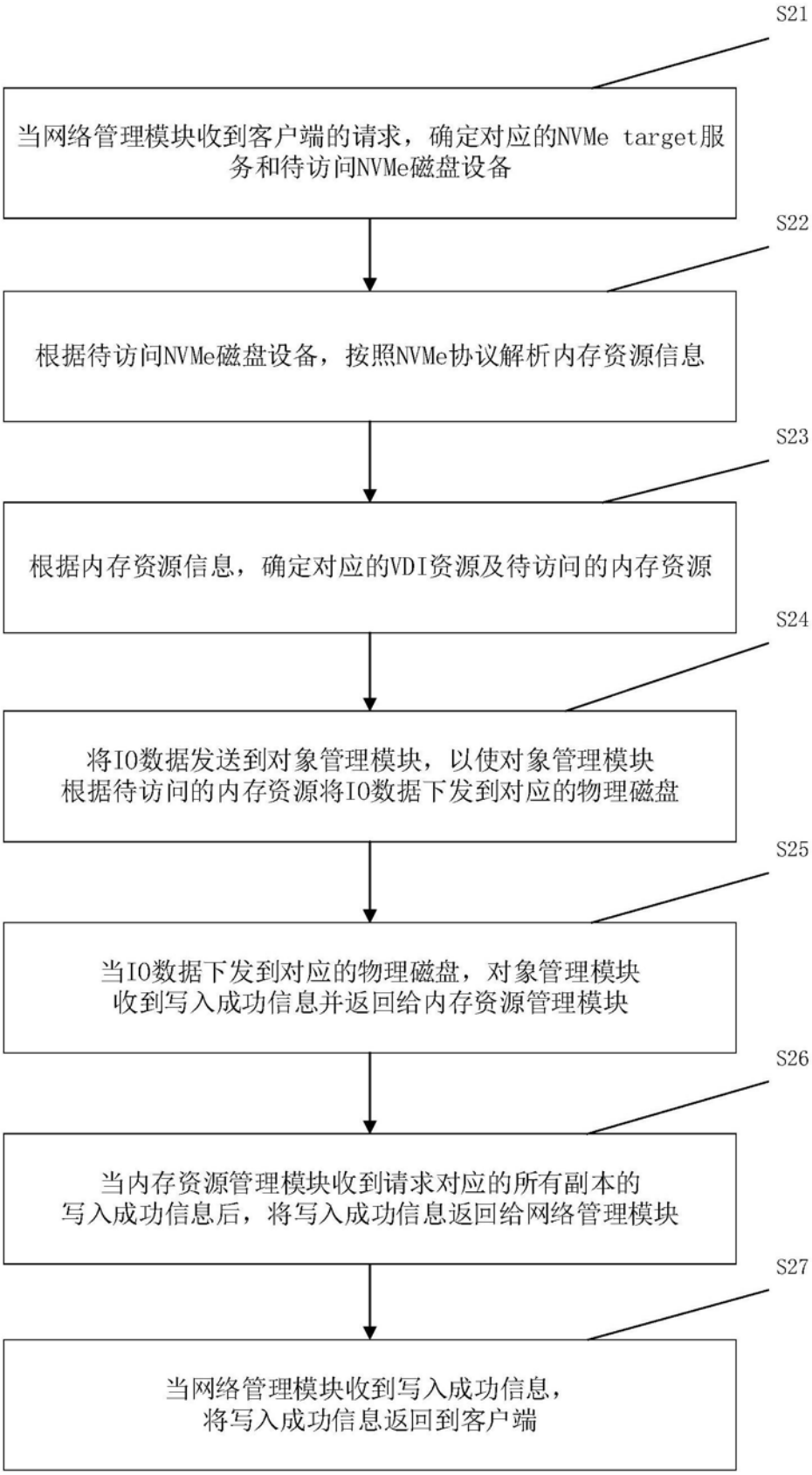

S21
当网络管理模块收到客户端的请求，确定对应的NVMe target服务和待访问NVMe磁盘设备
S22
根据待访问NVMe磁盘设备，按照NVMe协议解析内存资源信息
S23
根据内存资源信息，确定对应的VDI资源及待访问的内存资源
S24
将IO数据发送到对象管理模块，以使对象管理模块根据待访问的内存资源将IO数据下发到对应的物理磁盘
S25
当IO数据下发到对应的物理磁盘，对象管理模块收到写入成功信息并返回给内存资源管理模块
S26
当内存资源管理模块收到请求对应的所有副本的写入成功信息后，将写入成功信息返回给网络管理模块
S27
当网络管理模块收到写入成功信息，将写入成功信息返回到客户端

图3