



(12)发明专利申请

(10)申请公布号 CN 106850804 A

(43)申请公布日 2017.06.13

(21)申请号 201710066946.1

(22)申请日 2017.02.07

(71)申请人 郑州云海信息技术有限公司

地址 450000 河南省郑州市郑东新区心怡
路278号16层1601室

(72)发明人 李雪生

(74)专利代理机构 济南信达专利事务所有限公
司 37100

代理人 刘继枝

(51)Int.Cl.

H04L 29/08(2006.01)

H04L 29/06(2006.01)

H04L 12/46(2006.01)

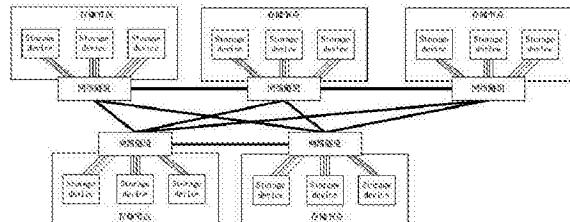
权利要求书1页 说明书6页 附图1页

(54)发明名称

一种海量存储系统减少网络连接数量的方
法

(57)摘要

本发明公开了一种海量存储系统减少网络
连接数量的方法，属于分布式存储系统存储池子
系统的存储节点间的通信框架领域，具体方法如
下：S1、海量存储系统包括n个存储节点，每个存
储节点包括m个存储设备；S2、存储设备管理进
程，所述的进程管理一个存储设备，负责该存储
设备的状态管理、数据管理、数据写入、数据读
出、数据删除；S3、存储节点之间建立TCP连接作
为TCP隧道；S4、存储节点内的存储设备管理进
程，通过TCP over TCP隧道方式实现网络连接；
S5、或者存储节点内的存储设备管理进程，通过
TCP over UDP隧道方式实现网络连接。本发明的
存储设备可以使用原网络接口；采用TCP over
TCP或者TCP over UDP实现网络连接的聚合，减
少资源使用。



1. 一种海量存储系统减少网络连接数量的方法,其特征在于,具体方法如下:

S1、海量存储系统包括n个存储节点,每个存储节点包括m个存储设备;

S2、存储设备管理进程,所述的进程管理一个存储设备,负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除;

S3、存储节点之间建立TCP连接作为TCP隧道;

S4、存储节点内的存储设备管理进程,通过TCP over TCP 隧道方式实现网络连接;

S5、存储节点内的存储设备管理进程,或者通过TCP over UDP隧道方式实现网络连接。

2. 根据权利要求1所述的一种海量存储系统减少网络连接数量的方法,其特征在于,所述的TCP over UDP隧道方式,具体实现方法是在上层的UDP报文封装一层TCP头。

3. 根据权利要求1所述的一种海量存储系统减少网络连接数量的方法,其特征在于,当m为36时,一个存储节点包括36个存储设备及36个存储设备管理进程;每个存储设备管理进程包括100个PG。

4. 根据权利要求3所述的一种海量存储系统减少网络连接数量的方法,其特征在于,所述的网络连接是实现PG与PG之间互联。

5. 根据权利要求1所述的一种海量存储系统减少网络连接数量的方法,其特征在于,具体操作步骤如下:

S1、存储节点之间建立TCP连接作为TCP隧道;

S2、存储设备管理进程内所有需要互联的PG之间,采用UDP与TCP隧道连接;即应用数据的报文先加UDP头,然后加TCP头,通过隧道到达目的节点,解TCP头,解UDP头,到PG;

S3、存储设备管理进程内所有需要互联的PG之间,采用TCP与TCP隧道连接;即应用数据的报文先加TCP头,然后加TCP头,通过隧道到达目的节点,解TCP头,解TCP头,到PG。

6. 一种海量存储系统,其特征在于,包括n个存储节点,每个存储节点包括m个存储设备;

所述的存储节点还包括存储设备管理进程模块,用于管理一个存储设备,负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除;

还包括TCP over TCP 隧道模块,用于实现网络连接。

7. 一种海量存储系统,其特征在于,包括n个存储节点,每个存储节点包括m个存储设备;

所述的存储节点还包括存储设备管理进程模块,用于管理一个存储设备,负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除;

还包括TCP over UDP 隧道模块,用于实现网络连接。

8. 根据权利要求6或7所述的一种海量存储系统,其特征在于,所述的m为12-72;所述的n为20-500。

一种海量存储系统减少网络连接数量的方法

技术领域

[0001] 本发明涉及分布式存储系统存储池子系统的存储节点间的通信框架领域,具体地说是一种海量存储系统减少网络连接数量的方法。

背景技术

[0002] 大数据时代,数据量爆炸式的增长,对海量存储提出了强烈的需求。分布式存储的高可靠、高扩展、高容量和低成本,是实现海量存储的方法之一。分布式存储系统随着扩展规模的增大,海量存储设备互联需要建立海量的网络连接,而系统的资源有限,逐渐成为系统扩展的限制。

[0003] 海量存储系统由大量存储节点,每个节点有一定数量的存储设备组成。由数据的冗余要求,存储池内的存储设备需要建立通信,实现数据的同步。传统的做法是重构系统通信机制,在TCP网络上层构建消息模块进行分发,但该做法使用限制多,工作量巨大。

[0004] TCP (Transmission Control Protocol 传输控制协议) 是一种面向连接的、可靠的、基于字节流的传输层通信协议,由 IETF 的 RFC 793 定义。在简化的计算机网络 OSI 模型中,它完成第四层传输层所指定的功能,用户数据报协议 (UDP) 是同一层内另一个重要的传输协议。在因特网协议族 (Internet protocol suite) 中,TCP 层是位于 IP 层之上,应用层之下的中间层。不同主机的应用层之间经常需要可靠的、像管道一样的连接,但是 IP 层不提供这样的流机制,而是提供不可靠的包交换。

发明内容

[0005] 本发明的技术任务是提供一种海量存储系统减少网络连接数量的方法,本发明采用 TCP over TCP 或者 TCP over UDP 网络隧道技术,减少网络连接瓶颈。

[0006] 本发明的技术任务是按以下方式实现的,一种海量存储系统减少网络连接数量的方法,具体方法如下:

S1、海量存储系统包括n个存储节点,每个存储节点包括m个存储设备;

S2、存储设备管理进程,所述的进程管理一个存储设备,负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除;

S3、存储节点之间建立TCP连接作为TCP隧道;

S4、存储节点内的存储设备管理进程,通过TCP over TCP 隧道方式实现网络连接;

S5、或者存储节点内的存储设备管理进程,通过TCP over UDP隧道方式实现网络连接。

[0007] 进一步的,优选的方法如下:所述的TCP over UDP隧道方式,具体实现方法是在上层的UDP报文封装一层TCP头。

[0008] 进一步的,优选的方法如下:当m为36时,一个存储节点包括36个存储设备及36个存储设备管理进程;每个存储设备管理进程包括100个PG。

[0009] 进一步的,优选的方法如下:所述的网络连接是实现PG与PG之间互联。

[0010] 进一步的,优选的方法如下:具体操作步骤如下:

S1、存储节点之间建立TCP连接作为TCP隧道；

S2、存储设备管理进程内所有需要互联的PG之间，采用UDP与TCP隧道连接；即应用数据的报文先加UDP头，然后加TCP头，通过隧道到达目的节点，解TCP头，解UDP头，到PG；

S3、存储设备管理进程内所有需要互联的PG之间，采用TCP与TCP隧道连接；即应用数据的报文先加TCP头，然后加TCP头，通过隧道到达目的节点，解TCP头，解TCP头，到PG。

[0011] 一种海量存储系统，包括n个存储节点，每个存储节点包括m个存储设备；

所述的存储节点还包括存储设备管理进程模块，用于管理一个存储设备，负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除；

还包括TCP over TCP 隧道模块，用于实现网络连接。

一种海量存储系统，包括n个存储节点，每个存储节点包括m个存储设备；

所述的存储节点还包括存储设备管理进程模块，用于管理一个存储设备，负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除；

还包括TCP over UDP 隧道模块，用于实现网络连接。

[0012] 进一步的，优选的结构如下：所述的m为12-72；所述的n为20-500。

[0013] 现有技术的海量存储系统，由大量节点，每个节点包括几十块存储设备，由于对海量存储空间扩展要求，系统包括上千块存储设备，为了满足数据可靠性要求，存储设备之间需要大量网络通信连接完成数据同步、一致性。系统采用副本或者N+M冗余策略，单个存储设备需要与大量其他存储互联的，而操作系统网络连接资源是有限的。假设每个节点m个存储设备，系统由n个存储节点，存储设备直接都是1对n的互联关系，n与冗余机制相关。

[0014] 直接网络连接，那么最差的网络连接数为： $(m*n) * (m*n)$ 。

[0015] 而采用本发明的网络隧道连接，存储节点内的存储设备管理进程，通过隧道连接实现通信互联；

采用TCP over TCP 隧道模式，网络连接数量为： $n*(n-1)$ ；采用TCP over UDP隧道模式，网络连接数量为： n 。

[0016] 本发明的一种海量存储系统减少网络连接数量的方法和现有技术相比，有益效果如下：

1、海量存储系统仍可以使用原网络接口；

2、采用TCP over TCP或者TCP over UDP 实现网络连接的聚合，减少节点系统资源的消耗；

3、使用解决分布式存储系统存储池大规模部署，海量数量的存储设备间网络连接数量指数增长问题。

附图说明

[0017] 附图1为TCP隧道技术的存储设备通信架构图。

具体实施方式

[0018] 海量存储是由数量很大的存储设备组成，一般都是上万块硬盘组成，为了保证数据冗余，硬盘间需要建立网络连接进行数据同步一致、恢复、迁移等。

[0019] UDP为应用程序提供的是一种不可靠的、无连接的分组交付，因此，UDP报文可能会

出现丢失、乱序、重复、延时等问题。传输控制协议TCP是一个面向链接的、可靠的通信协议。1. 在开始传输前,需要进行三次握手建立链接;2. 可靠性:在传输过程中,通信双方的协议模块继续进行通信;3. 通信结束后,通信双方都会使用改进的三次握手来关闭链接。当网络硬件失效或者负担太重时,数据包可能就会产生丢失、重复、延时、乱序的现象。这些都会导致我们的通信不正常。如果让应用程序来担负差错控制的工作,无疑将给程序员带来许多复杂的工作,于是,我们使用独立的通信协议TCP来保证通信的可靠性是非常必要的。

[0020] 适合UDP的环境:

1.在高效可靠的网络环境中(不需要考虑网络不好导致的丢包、乱序、延时、重复等问题),因为UDP是无连接的服务,不用消耗不必要的网络资源(TCP中的协议间通信)和处理时间(预期确认需要的时间),从而效率要高的多。2.在轻权通信中,当需要传输的数据量很小(可以装在一个IP数据包内)时。如果我们使用TCP协议,那么,先建立连接,一共需要发送3个IP数据包,然后数据传输,1个IP数据包,产生一个确认信号的IP包,然后关闭连接,需要传输5个IP数据包。使用TCP协议IP包的利用率为1/10。而使用UDP,只需要发送一个IP数据包。哪怕丢包(服务不成功),也可重新申请服务(重传)。UDP很适合这种客户机向服务器传送简单服务请求的环境。此类应用层协议包括TFTP ,SNMP , DNS ,DHCP等。3.在对实时性要求很强的通信中:在诸如实时视频直播等对实时性要求很高的环境中,从而允许一定量的丢包的情况下,UDP更适合。

[0021] 适合TCP协议的环境:

当网络硬件失效或者负担太重时,数据包可能就会产生丢失、重复、延时、乱序的现象。这些都会导致我们的通信不正常的时候。如果让应用程序来担负差错控制的工作,无疑将给程序员带来许多复杂的工作,于是,我们使用独立的通信协议来保证通信的可靠性是非常必要的。

[0022] 因此直接使用TCP的网络接口,一个存储节点(x86服务器)的一个OSD,会和大量的其他节点的OSD建立网络连接;海量存储系统,从大细分下来分为节点、OSD、PG(置换组),置换组的存在是为了数据迁移、一致、冗余的速度考虑,海量存储系统最小的迁移单元不是硬盘,而是一组数据;

海量存储系统,从大细分下来分为节点、OSD存储设备管理进程模块、PG(置换组),为什么会有置换组的,为了数据迁移、一致、冗余的速度考虑,海量存储系统最小的迁移单元不是硬盘,而是一组数据;

一般1个节点36块盘(即存储设备,Storage Device),及36个OSD存储设备管理进程模块,一个OSD 100个PG(Placement Group)置换组,网络连接是PG和PG间互联,就造成大量的网络连接。

[0023] OSD(Object-based Storage Device)对象存储设备,即存储设备管理进程,通过对象层次的抽象设计,将数据存储的访问、控制、管理等功能重新划分,将文件系统的逻辑结构与物理存储的映射关系隐藏到对象一层,数据的访问操作通过对对象的接口实现。即在OSD结构中,存储不是以块或文件的形式组织,而是对象。例如,一个对象可能包含一条数据库记录或者一个数据库表,甚至整个数据库本身,也可以包含一个文件或者文件的一部分。

[0024] 通常的块存储管理是以固定长度的块为单位实现的,数据的读写也是以块为单位。而OSD是以对象的形式实现数据的组织和访问,在OSD中,存储设备只包含有对象,每个

对象是一系列有序字节的集合,对象具有唯一的ID标示符,对象有自己特定的方法和属性。对象的方法如创建、打开、读、写、关闭对象、设置、获取对象属性等;对象的数据包括两类,即元数据和用户数据,其中,元数据用来描述对象特定的属性,如对象的元数据的大小、总的字节的大小,对象的逻辑大小以及其他参数,用户数据用来保存实际的二进制数据。

[0025] 实施例1:

一种海量存储系统减少网络连接数量的方法,具体方法如下:

S1、海量存储系统包括20个存储节点,每个存储节点包括12个存储设备;

S2、存储设备管理进程,所述的进程管理一个存储设备,负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除;

S3、存储节点之间建立TCP连接作为TCP隧道;

S4、存储节点内的存储设备管理进程,通过TCP over TCP 隧道方式实现网络连接;

S5、或者存储节点内的存储设备管理进程,通过TCP over UDP隧道方式实现网络连接,具体实现方法是在上层的UDP报文封装一层TCP头。

[0026] 一个存储节点包括12个硬盘及12个OSD(存储设备管理进程);每个OSD包括100个PG。所述的网络连接是实现PG与PG之间互联。

[0027] 具体操作步骤如下:

S1、存储节点之间建立TCP连接作为TCP隧道;

S2、OSD内所有需要互联的PG之间,采用UDP与TCP隧道连接;即应用数据的报文先加UDP头,然后加TCP头,通过隧道到达目的节点,解TCP头,解UDP头,到PG;

S3、OSD内所有需要互联的PG之间,采用TCP与TCP隧道连接;即应用数据的报文先加TCP头,然后加TCP头,通过隧道到达目的节点,解TCP头,解TCP头,到PG。

[0028] 根据一种海量存储系统减少网络连接数量的方法,本发明进一步提出了,与之相关的一种海量存储系统,包括20个存储节点,每个存储节点包括12个存储设备;

包括存储设备管理进程模块,用于管理一个存储设备,负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除;

还包括TCP over TCP 隧道模块,用于实现网络连接。

一种海量存储系统,包括20个存储节点,每个存储节点包括12个存储设备;

包括存储设备管理进程模块,用于管理一个存储设备,负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除;

还包括TCP over UDP 隧道模块,用于实现网络连接。

[0029] 直接网络连接,那么最差的网络连接数为: $(12*20)*(12*20)=57600$ 。

[0030] 而采用本发明的网络隧道连接,存储节点内的存储设备管理进程,通过隧道连接实现通信互联;

采用TCP over TCP 隧道模式,网络连接数量为 : $20*(20-1)=380$;采用TCP over UDP 隧道模式,网络连接数量为: $n=20$ 。

[0031] 实施例2:

一种海量存储系统减少网络连接数量的方法,具体方法如下:

S1、海量存储系统包括100个存储节点,每个存储节点包括36个存储设备;

S2、存储设备管理进程,所述的进程管理一个存储设备,负责该存储设备的状态管理、

数据管理、数据写入、数据读出、数据删除；

S3、存储节点之间建立TCP连接作为TCP隧道；

S4、存储节点内的存储设备管理进程，通过TCP over TCP 隧道方式实现网络连接；

S5、或者存储节点内的存储设备管理进程，通过TCP over UDP隧道方式实现网络连接，具体实现方法是在上层的UDP报文封装一层TCP头。

[0032] 一个存储节点包括36个硬盘及36个OSD(存储设备管理进程)；每个OSD包括100个PG。所述的网络连接是实现PG与PG之间互联。

[0033] 具体操作步骤如下：

S1、存储节点之间建立TCP连接作为TCP隧道；

S2、OSD内所有需要互联的PG之间，采用UDP与TCP隧道连接；即应用数据的报文先加UDP头，然后加TCP头，通过隧道到达目的节点，解TCP头，解UDP头，到PG；

S3、OSD内所有需要互联的PG之间，采用TCP与TCP隧道连接；即应用数据的报文先加TCP头，然后加TCP头，通过隧道到达目的节点，解TCP头，解TCP头，到PG。

[0034] 根据一种海量存储系统减少网络连接数量的方法，本发明进一步提出了，与之相关的一种海量存储系统，包括100个存储节点，每个存储节点包括36个存储设备；

包括存储设备管理进程模块，用于管理一个存储设备，负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除；

还包括TCP over TCP 隧道模块，用于实现网络连接。

一种海量存储系统，包括100个存储节点，每个存储节点包括36个存储设备；

包括存储设备管理进程模块，用于管理一个存储设备，负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除；

还包括TCP over UDP 隧道模块，用于实现网络连接。

[0035] 直接网络连接，那么最差的网络连接数为： $(36*100)*(36*100)=12960000$ 。

[0036] 而采用本发明的网络隧道连接，存储节点内的存储设备管理进程，通过隧道连接实现通信互联；

采用TCP over TCP 隧道模式，网络连接数量为： $100*(100-1)=9900$ ；采用TCP over UDP隧道模式，网络连接数量为： $n=100$ 。

[0037] 实施例3：

一种海量存储系统减少网络连接数量的方法，具体方法如下：

S1、海量存储系统包括500个存储节点，每个存储节点包括72个存储设备；

S2、存储设备管理进程，所述的进程管理一个存储设备，负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除；

S3、存储节点之间建立TCP连接作为TCP隧道；

S4、存储节点内的存储设备管理进程，通过TCP over TCP 隧道方式实现网络连接；

S5、或者存储节点内的存储设备管理进程，通过TCP over UDP隧道方式实现网络连接，具体实现方法是在上层的UDP报文封装一层TCP头。

[0038] 一个存储节点包括72个硬盘及72个OSD(存储设备管理进程)；每个OSD包括100个PG。所述的网络连接是实现PG与PG之间互联。

[0039] 具体操作步骤如下：

S1、存储节点之间建立TCP连接作为TCP隧道；

S2、OSD内所有需要互联的PG之间，采用UDP与TCP隧道连接；即应用数据的报文先加UDP头，然后加TCP头，通过隧道到达目的节点，解TCP头，解UDP头，到PG；

S3、OSD内所有需要互联的PG之间，采用TCP与TCP隧道连接；即应用数据的报文先加TCP头，然后加TCP头，通过隧道到达目的节点，解TCP头，解TCP头，到PG。

[0040] 根据一种海量存储系统减少网络连接数量的方法，本发明进一步提出了，与之相关的一种海量存储系统，包括500个存储节点，每个存储节点包括72个存储设备；

包括存储设备管理进程模块，用于管理一个存储设备，负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除；

还包括TCP over TCP 隧道模块，用于实现网络连接。

一种海量存储系统，包括500个存储节点，每个存储节点包括72个存储设备；

包括存储设备管理进程模块，用于管理一个存储设备，负责该存储设备的状态管理、数据管理、数据写入、数据读出、数据删除；

还包括TCP over UDP 隧道模块，用于实现网络连接。

[0041] 直接网络连接，那么最差的网络连接数为： $(72*500)*(72*500)=1,296,000,000$ 。

[0042] 而采用本发明的网络隧道连接，存储节点内的存储设备管理进程，通过隧道连接实现通信互联；

采用TCP over TCP 隧道模式，网络连接数量为： $500*(500-1)=249500$ ；采用TCP over UDP隧道模式，网络连接数量为： $n=500$ 。

[0043] 其中采用UDP over TCP，网络隧道技术是把上层的报文直接再封装一层TCP头，意味着上层UDP的报文直接作为TCP的数据，直接发/收；

这样一个节点只要一个TCP连接，上层通信报文采用UDP进行分发，到各个OSD的PG；UDP是数据报协议，直接采用ip，端口就能分发，类似一个内部的小总线网络直接互联，存在的问题就是UDP不是可靠性传输，但由于是一台机器内部的传输，可靠性能够得到一定保证。

[0044] 节点间先建立TCP连接作为TCP隧道，OSD内所有需要互联的PG，采用UDP与TCP隧道连接，应用数据报文先加UDP头，加TCP头，再通过隧道，到目的节点，解TCP，解UDP，到PG。

[0045] 采用TCP over TCP，UDP是面向存在连接不可靠的问题，即UDP不重发报文。这里提TCP就是为了更高的网络传输可靠性和网络拥塞控制，思路和UDP over TCP类似，OSD内部(PG—Placement Group)置换组和隧道建立连接，利用隧道到其他节点，过程是类似的。

[0046] TCP隧道部署的意义：1. UDP和TCP全部封装在TCP隧道，因为TCP拥塞控制具有公平性，公平性；2. 带宽的适配。本发明涉及到海量存储系统存储池子系统的通信框架，主要提出了利用网络隧道技术，解决分布式存储系统存储池大规模部署，海量数量的存储设备间网络连接数量指数增长问题。

[0047] 通过上面具体实施方式，所述技术领域的技术人员可容易的实现本发明。但是应当理解，本发明并不限于上述的几种具体实施方式。在公开的实施方式的基础上，所述技术领域的技术人员可任意组合不同的技术特征，从而实现不同的技术方案。

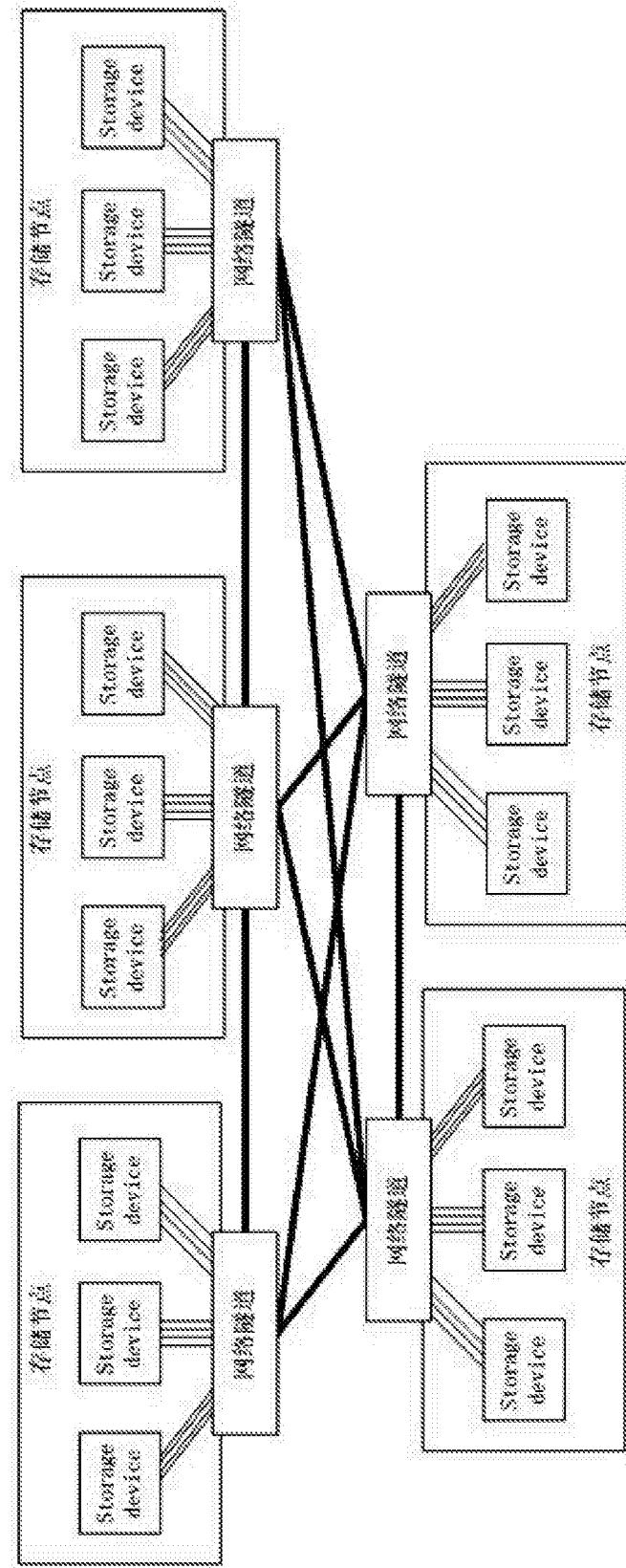


图1