(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 108632338 B
(45)授权公告日 2022.05.24

(21)申请号 201810161543.X

(22)申请日 2018.02.26

(65)同一申请的已公布的文献号
申请公布号 CN 108632338 A

(43)申请公布日 2018.10.09

(30)优先权数据
62/472,379 2017.03.16 US
15/615,508 2017.06.06 US

(73)专利权人 三星电子株式会社
地址 韩国京畿道

(72)发明人 钦纳克里希南·鲍拉普拉姆

(74)专利代理机构 北京天昊联合知识产权代理有限公司 11112
专利代理师 张帆 赵南

(51)Int.Cl.
H04L 67/60(2022.01)
H04L 67/1097(2022.01)
H04L 67/10(2022.01)
H04L 61/5014(2022.01)
H04L 67/51(2022.01)

(56)对比文件
US 2009100163 A1,2009.04.16
WO 2015194033 A1,2015.12.23
WO 2016139749 A1,2016.09.09
US 2004068561 A1,2004.04.08
US 2014289377 A1,2014.09.25

审查员 冯骥

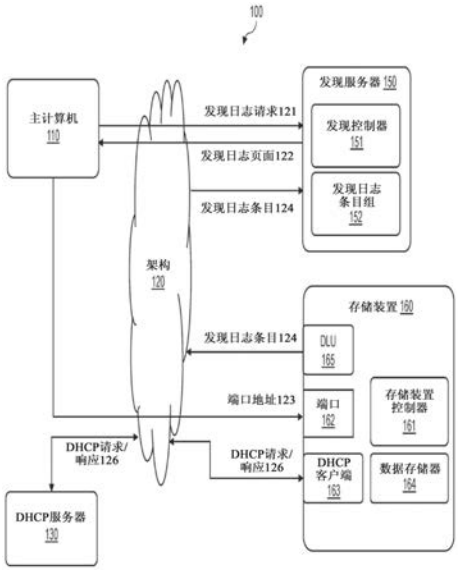
权利要求书2页 说明书8页 附图6页

(54)发明名称

存储系统和用于提供自动存储发现的方法

(57)摘要

提供了存储系统和用于提供自动存储发现的方法。存储系统包括:多个存储装置,每个存储装置运行DHCP客户端和发现日志更新器;主计算机,向多个存储装置提供I/O或计算工作负载;和发现服务器,向主计算机提供发现日志页面。DHCP客户端和发现日志更新器在每个存储装置上运行,将相应存储装置的发现日志条目提供给发现服务器。发现日志页面包含多个存储装置的发现信息,发现信息包括一个或多个发现日志条目,并且每一个发现日志条目包括相应存储装置的被分配用于在主计算机与相应存储装置之间建立通信的地址。主计算机使用包含在发现日志页面中的目标存储装置的发现日志条目来建立与目标存储装置的连接,并且向目标存储装置提供I/O或计算工作负载。



CN 108632338 B

1.一种存储系统,包括:

多个存储装置,所述多个存储装置中的每一个被配置为运行动态主机配置协议客户端;以及

发现服务器,被配置为向主计算机提供发现日志页面,

其中,所述发现日志页面包含所述多个存储装置的发现信息,所述发现信息包括一个或多个发现日志条目,并且所述一个或多个发现日志条目中的每一个包括相应存储装置的被分配用于在所述主计算机与所述相应存储装置之间建立通信的地址,

其中,在所述多个存储装置中的每一个上运行的发现日志更新器自动地更新对应存储装置的发现信息,并自动地将对应存储装置的发现日志条目提供给所述发现服务器,并且

其中,所述主计算机使用包含在所述发现日志页面中的目标存储装置的发现日志条目来建立与所述目标存储装置的连接,并向所述目标存储装置提供输入/输出或计算工作负载。

2.根据权利要求1所述的存储系统,其中,动态主机配置协议服务器为所述多个存储装置中的每一个存储装置分配地址,并且所述多个存储装置中的各发现日志更新器使用所分配的地址来将其各自的包括分配地址的发现日志条目自动发送至所述发现服务器。

3.根据权利要求2所述的存储系统,其中,所述发现日志条目还包括存储装置控制器的端口地址。

4.根据权利要求3所述的存储系统,其中,根据架构上快速非易失性存储器规范来分配所述端口地址。

5.根据权利要求1所述的存储系统,其中,所述发现服务器使用RESTful应用编程接口或GraphQL请求来将所述发现日志页面发送到所述主计算机。

6.根据权利要求1所述的存储系统,其中,所述多个存储装置中的至少一个存储装置是架构上快速非易失性存储器装置,并且所述主计算机向所述发现服务器发送获取日志命令以接收所述发现日志页面。

7.根据权利要求2所述的存储系统,其中,所述发现服务器和所述动态主机配置协议服务器并置在所述存储系统的同一服务器中。

8.根据权利要求1所述的存储系统,其中,所述多个存储装置附接到机架,并且所述机架中的发现服务器收集所述机架专用的机架级发现信息。

9.根据权利要求8所述的存储系统,其中,所述多个存储装置分布在POD中的多个机架之中,负责所述POD的发现服务器收集所述POD专用的POD级发现信息。

10.根据权利要求9所述的存储系统,其中,所述POD的POD管理器基于所述POD级发现信息在所述多个存储装置之间调度工作负载。

11.根据权利要求10所述的存储系统,其中,所述POD管理器与所述存储系统的另一POD管理器交换所述POD级发现信息。

12.一种用于提供自动存储发现的方法,包括:

运行存储装置中的动态主机配置协议客户端;

将所述存储装置的发现日志条目提供给发现服务器;

将发现日志页面从所述发现服务器提供至主计算机,其中所述发现日志页面包含多个存储装置的发现信息,所述发现信息包括一个或多个发现日志条目,并且所述一个或多个

发现日志条目中的每一个包括地址相应存储装置的被分配用于在所述主计算机与对应存储装置之间建立通信的地址；

使用包含在所述发现日志页面中的目标存储装置的发现日志条目来在所述主计算机与目标存储装置之间建立连接;以及

从所述主计算机向所述目标存储装置提供输入/输出或计算工作负载，

其中,将所述存储装置的发现日志条目提供给发现服务器包括:自动地更新所述存储装置的发现信息并自动地将所述存储装置的发现日志条目提供给所述发现服务器。

13.根据权利要求12所述的方法,其中,动态主机配置协议服务器为所述多个存储装置中的每一个分配地址,并且所述多个存储装置中的每个发现日志更新器使用所分配的地址来将其各自的包括分配地址的发现日志条目自动发送到所述发现服务器。

14.根据权利要求13所述的方法,其中,所述发现日志条目还包括存储装置控制器的端口地址。

15.根据权利要求14所述的方法,其中,根据架构上快速非易失性存储器规范来分配所述端口地址。

16.根据权利要求12所述的方法,其中,所述发现服务器使用RESTful应用编程接口或GraphQL请求来将所述发现日志页面发送到所述主计算机。

17.根据权利要求12所述的方法,其中,所述多个存储装置中的至少一个存储装置是架构上快速非易失性存储器装置,并且所述主计算机向所述发现服务器发送获取日志命令以接收所述发现日志页面。

18.根据权利要求13所述的方法,其中,所述发现服务器和所述动态主机配置协议服务器并置在存储系统的同一服务器中。

19.根据权利要求12所述的方法,其中,所述多个存储装置附接到机架,并且所述机架中的发现服务器收集所述机架专用的机架级发现信息。

20.根据权利要求19所述的方法,其中,所述多个存储装置分布在POD中的多个机架之中,负责所述POD的发现服务器收集所述POD专用的POD级发现信息。

21.根据权利要求20所述的方法,其中,所述POD的POD管理器基于所述POD级发现信息来在所述多个存储装置之间调度工作负载。

22.根据权利要求21所述的方法,其中,所述POD管理器与存储系统的另一POD管理器交换所述POD级发现信息。

存储系统和用于提供自动存储发现的方法

[0001] 相关申请的交叉引用

[0002] 本申请要求2017年3月16日提交的美国临时专利申请No.62/472,379的优先权,其公开内容通过引用方式整体合并于此。

技术领域

[0003] 本公开总体而言涉及一种分布式存储装置,更具体地,涉及一种用于提供在超大规模数据中心环境中的分布式存储装置的自动发现的系统和方法。

背景技术

[0004] 数据中心中的每个机架包含用于计算和存储装置的滑轨和/或托盘的组合。超大规模数据中心可以包括若干性能优化的数据中心(POD),并且每个POD可以包括多个机架和成百上千的计算和/或存储装置。

[0005] 快速非易失性存储器(NVMe)定义了用于主机软件(host software)的寄存器级接口,从而经由外围组件互连快速(PCIe)总线与非易失性存储子系统(例如,固态驱动器(SSD))通信。架构上快速非易失性存储器(NVMe over fabrics,简称为NVMeoF)(或简称为NVMf)定义了一种通用体系结构,其在大范围的存储网络架构(fabric)(诸如以太网、光纤通道、InfiniBand、传输控制协议(TCP)网络以及其他网络架构)上支持NVMe块存储协议。

[0006] 在基于NVMeoF的存储系统中,每个存储装置(也被称为NVMeoF兼容SSD或以太网SSD)通过其子系统NVMe限定名(NQN)、互联网协议(IP)地址、端口ID和/或控制器ID来识别。每个POD管理器需要知道架构(例如,以太网)上的存储装置的状态和可用性,来分配和调度工作负载。某些存储装置可能因维护而关闭,并且工作负载不能分配给这些存储装置。此外,在超大规模数据中心环境中对来自分布在架构上的成百上千的存储装置的发现信息的收集、维护和更新并非是一件易事。

发明内容

[0007] 根据一个实施例,一种存储系统包括:多个存储装置,所述多个存储装置中的每一个被配置为运行动态主机配置协议(DHCP)客户端;以及发现服务器,其被配置为向主计算机提供发现日志页面。在多个存储装置中的每一个上运行的发现日志更新器将相应存储装置的发现日志条目提供给发现服务器。所述发现日志页面包含多个存储装置的发现信息,所述发现信息包括一个或多个发现日志条目,并且所述一个或多个发现日志条目中的每一个包括相应存储装置的被分配用于在主计算机与相应存储装置之间建立通信的地址。主计算机使用发现日志页面中包含的目标存储装置的发现日志条目来建立与目标存储装置的连接,并向目标存储装置提供输入/输出(I/O)或计算工作负载。

[0008] 根据另一个实施例,一种方法包括:运行在存储装置中的动态主机配置协议(DHCP)客户端;将存储装置的发现日志条目提供给发现服务器;将发现日志页面从发现服务器提供至主计算机,其中发现日志页面包含多个存储装置的发现信息,所述发现信息包

括一个或多个发现日志条目，并且所述一个或多个发现日志条目中的每一个包括相应存储装置的被分配用于在主计算机与对应存储装置之间建立通信的地址；使用包含在发现日志页面中的目标存储装置的发现日志条目来在主机与目标存储装置之间建立连接；以及从主计算机向目标存储装置提供输入/输出(I/O)或计算工作负载。

[0009] 现在将参照附图更具体地描述并且在权利要求中指出以上和其他优选特征，包括实施方式的各种新颖细节和事件的组合。将理解，本文描述的具体系统和方法仅仅是作为说明，而不是作为限制来示出。如本领域技术人员将理解的，在不脱离本公开的范围的情况下，可以在各种各样的实施例中采用本文所描述的原理和特征。

附图说明

[0010] 作为本说明书的一部分而包括的附图示出了当前优选的实施例，并且与上面给出的一般性描述以及下面给出的优选实施例的详细描述一起用于解释和教导本文描述的原理。

[0011] 图1示出根据一个实施例的示例分布式存储装置系统的框图；

[0012] 图2示出根据一个实施例的示例发现日志页面的数据结构；

[0013] 图3示出根据一个实施例的包含在发现日志页面中的示例日志条目的数据结构；

[0014] 图4示出根据一个实施例的用于提供发现日志信息的示例流程图；

[0015] 图5示出根据一个实施例的用于在主计算机与目标存储装置之间建立通信的示例流程图；以及

[0016] 图6示出根据一个实施例的自动发现的示例过程。

[0017] 附图不一定按比例绘制，并且出于说明的目的，类似结构或功能的元件在整个附图中通常用相同的附图标记来表示。附图仅旨在便于描述本文所述的各种实施例。附图没有描述在此公开的教导的每个方面，并且不限制权利要求的范围。

具体实施方式

[0018] 本文公开的每个特征和教导可以单独使用或者与其他特征和教导结合使用，以提供在超大规模数据中心环境中提供自动以太网存储发现的系统和方法。参考附图进一步详细描述了单独和组合地利用这些附加特征和教导中的许多附加特征和教导的代表性示例。该详细描述仅旨在教导本领域技术人员用于实践本教导的方面的更多细节，而不旨在限制权利要求的范围。因此，上面在具体实施方式中公开的特征的组合对以最广泛的含义来实践教导来说可能不是必需的，而是被教导仅仅是为了描述本教导的特定代表性示例。

[0019] 在下面的描述中，仅出于解释的目的，阐述了特定的术语以提供对本公开的透彻理解。然而，对于本领域技术人员显而易见的是，不需要这些具体细节来实践本公开的教导。

[0020] 本文的具体描述中的一些部分是根据对计算机存储器内的数据位的操作的算法和符号表示来呈现的。数据处理领域的技术人员使用这些算法描述和表示来将其工作的实质有效地传达给本领域其他技术人员。在这里，算法通常被构思为产生期望结果的有条理的步骤序列。这些步骤是需要对物理量进行物理操作的步骤。通常，虽然不一定，但是这些量采取能够被存储、传输、组合、比较和以其他方式操作的电信号或磁信号的形式。主要出

于常用的原因，将这些信号称为比特、值、元素、符号、字符、术语、数字等，这已经证实有时是非常方便的。

[0021] 然而，应该记住的是，所有这些和类似的术语都与适当的物理量相关联，并且仅仅是适用于这些量的方便的标签。除非特别另外指出，否则如从下面的讨论中明显的是，应当理解在整个说明书中，利用诸如“处理”、“计算”、“运算”、“确定”、“显示”等术语的讨论指的是计算机系统或类似的电子计算装置的动作和过程，其将在计算机系统的寄存器和存储器内表示为物理(电子)量的数据操作并转换成类似地表示为计算机系统的存储器或寄存器(或者其他这样的信息存储装置、传输装置或显示装置)内的物理量的其它数据。

[0022] 此外，代表性示例和从属权利要求的各种特征可以以未具体和明确列举的方式进行组合，以便提供本教导的另外的有用实施例。还要特别注意，为了原始公开的目的，并且为了限制所要求保护的主题，实体组的所有值范围或指示都公开了每个可能的中间值或中间实体。还要特别注意，图中所示的部件的尺寸和形状被设计成有助于理解如何实践本教导，而不旨在限制示例中所示的尺寸和形状。

[0023] 本公开描述了包括多个数据存储装置的分布式数据存储系统。本分布式数据存储系统采用动态主机配置协议(DHCP)服务来提供、更新和维护多个数据存储装置的动态发现信息。分布式数据存储系统的子系统之间通信所支持的传输协议的示例是NVMeoF协议，并且架构是以太网。子系统包括一个或多个计算节点(例如，远程主计算机或计算节点)和多个存储装置(例如，以太网SSD)。

[0024] NVMeoF标准定义了基于NVMeoF的系统的发现服务，以允许主计算机(在本文中也被称为主机或计算节点)识别和定位主计算机可以访问的子系统(在本文中也被称为存储装置或节点)。提供发现服务的发现服务器是将一个或多个发现控制器公开给基于NVMeoF的系统的主计算机的子系统。发现服务器的每个发现控制器可以向请求主计算机提供发现日志页面。

[0025] 由发现控制器提供的发现日志页面可以包含一个或多个发现日志条目。发现日志页面中包含的每个发现日志条目指定主计算机通过NVMe传输协议连接到子系统所需的信息。例如，发现日志条目指定如下的子系统，该子系统将它的名称空间或引用公开给另一发现服务。根据NVMeoF规范(架构上快速非易失性存储器规范)，最大引用深度可以是八个级别或更少。

[0026] 图1示出了根据一个实施例的示例分布式存储装置系统的框图。分布式存储系统100包括主计算机110、动态主机配置协议(DHCP)服务器130、发现服务器150、存储装置160、以及通过建立的网络提供主计算机110、DHCP服务器130、发现服务器150和存储装置160之间连接的架构120。架构120的示例是以太网。

[0027] 主计算机110可以代表分布式存储系统100的许多主计算机之一，并且存储装置160可以代表分布式存储系统100的许多存储装置之一。根据一个实施例，主计算机110是指分布式存储系统100的计算节点，该计算节点可以向存储装置发布工作负载或者简单地对存储在包括存储装置160的分布式存储系统100的各存储装置中的数据执行读取/写入访问。

[0028] 存储装置160包括存储装置控制器161、端口162、DHCP客户端163、诸如固态驱动器(SSD)的数据存储器164以及发现日志更新器(DLU)165。在一个实施例中，DLU 165是在存储

装置160上运行的服务，以使用架构传输协议和使用RESTful API或GraphQL请求的端口来更新发现服务器150的发现日志条目组152中的存储装置160的发现日志条目。例如，DLU 165可以使用TCP/IP或UDP/IP和TCP/UDP端口来更新发现服务器150的发现日志条目组152中的存储装置160的发现日志条目。DHCP服务器130是分布式存储系统100的子系统，该子系统向请求DHCP客户端提供动态或永久租赁的IP地址。例如，存储装置160的DHCP客户端163向DHCP服务器130发送请求126，并且DHCP服务器130响应于该请求126向DHCP客户端163分配IP地址(动态IP地址或永久IP地址)。

[0029] 发现服务器150向被配置为从发现服务器150获取发现日志信息的任何请求主计算机提供发现服务。发现服务器150包括发现控制器151和发现日志条目组152以及NVMeoF发现服务器规范中描述的与此处的讨论不是特别相关的其他组件。发现控制器151的IP地址或传输地址向包括主计算机110的分布式存储系统100的一个或多个主计算机公开。另外，主机可以在到发现服务的连接命令中使用众所周知的发现服务NQN(例如，“nqn.2014-08.org.nvmexpress.discovery”)。

[0030] 主计算机110可以使用发现服务器150的发现服务来查找连接到分布式存储系统100内可用的一个或多个存储装置所需的信息。主计算机110可以以很多方式来获得初始发现信息。例如，没有发现服务器150的IP地址的新附接的主计算机可以从本地主机配置文件或从主计算机的操作系统(OS)的管理程序或对主计算机已知的任何其他子系统中获得初始发现信息。主计算机110还可以使用到发现服务的连接命令中的发现服务NQN(例如，“nqn.2014-08.org.nvmexpress.discovery”)地址。主计算机110用来获得连接到发现服务所需的NVMe传输信息的发现方法基于所使用的架构协议。初始发现服务信息可以包括发现服务器150或具有发现服务器150的信息的另一子系统的IP地址。

[0031] 一旦发现服务器150的身份(例如，发现服务器150的IP地址)已知，主计算机110就发送发现日志请求121(例如，根据NVMeoF规范的获取日志页面请求)到发现服务器150，以根据需要来获得发现信息。作为响应，发现服务器150提供包括发现日志条目组152的发现日志页面122，其中发现日志条目组152包括一个或多个发现日志条目。包含在发现日志页面122中的每个发现日志条目指定主计算机110连接到一个或多个对应的存储装置160所需的信息。例如，每个条目包括对应的存储装置160的一个或多个标识符和连接到存储装置160所支持的传输协议信息。发现日志条目组152可以支持多个传输协议和多个地址类型。

[0032] 根据一个实施例，发现服务器150可以自动地将最新的发现日志条目组152发送到主计算机110。例如，发现服务器150保存主计算机的列表，并且在更新一个或多个发现日志条目时，将更新的发现日志页面122发送到列出的主计算机，或者以预定间隔(例如，每天、每周)将更新的发现日志页面122发送到列出的主计算机，在一种情况下发现服务器150可以是负责POD/POD管理器的发现服务器。在这种情况下，发现日志页面122可以包含仅包括更新的发现日志条目的所存储的发现日志条目组152的子集。

[0033] 基于在发现日志页面122中识别的与存储装置160对应的发现日志条目和传输类型，主计算机110可以识别目标存储装置160的端口地址123，并使用所指示的支持的传输类型和协议、使用端口地址123来建立与目标存储装置160的通信路径。在发现日志页面122中传送主机与目标存储装置之间初始化连接所需的所有信息。存储装置160的存储装置控制器161和主计算机110可以使用建立的端口162彼此通信。

[0034] 在存储装置160上运行的DHCP客户端163向DHCP服务器130发送请求，并且DHCP服务器130将IP地址(或者与在对应架构中的IP地址相对应的相似地址)分配给请求DHCP客户端163。DLU 165在每个目标存储装置上或针对一组目标存储装置运行。例如，对于底架/机架中的全部或部分存储装置160，可以使用一个或几个DHCP客户端163。类似地，一个或几个DLU 165可以更新底架/机架中的一个存储装置160或一组存储装置160的日志。DHCP服务和DLU服务可以在底架/机架中的一个或几个存储装置160上运行，或者在底架/机架中的通用中央处理器(CPU)上运行，或者在底架/机架中的基板管理控制器(BMC)上运行。DHCP客户端163或DLU 165可以将包括存储装置控制器161的被分配的IP地址的发现日志条目124发送到发现服务器150。根据一个实施例，DHCP服务器130和发现服务器150并置在同一台服务器中。在一些实施例中，DHCP服务器130是分布式存储系统100之外的不同于发现服务器150的另一系统。

[0035] 发现服务器150收集并更新包括发现日志条目组152的发现日志信息，使用RESTful API或GraphQL请求从目标存储装置的一个或多个DLU 165接收到所述发现日志条目组152。发现服务器150将发现日志页面提供至请求主计算机110作为对NVMeoF获取日志页面请求的响应。通过在每个目标存储装置或一组目标存储装置上运行的DLU使用由DHCP服务器分配的IP地址来对发现信息进行自动更新可以用连接到驻留在当前分布式存储系统中的目标存储装置所需的最新发现信息将无缝发现服务提供至请求主计算机。

[0036] 由于新存储装置的添加、现有存储装置的替换或退役，作为数据中心环境中的例行操作过程的一部分，可以添加新的存储装置，并且可以关闭一些存储装置。由于租用的存储装置160的各个DHCP客户端163可以更新其自己的当前信息，所以DHCP服务器130可以保持最新的发现日志信息。通过使用在每个存储装置160上运行的瘦DHCP客户端163的发现日志信息的自动更新方案，本分布式存储系统100可以提供可扩展的解决方案，该可扩展的解决方案可以容纳超大规模数据中心环境中的成百上千个存储装置。

[0037] 图2示出根据一个实施例的示例发现日志页面的数据结构。图2中所示的发现日志页面200包含与发现信息有关的各种参数，包括但不限于，生成计数器、记录数量、记录格式、一个或多个发现日志条目300。图3示出了包含在图2的发现日志页面200中的示例日志条目的数据结构。发现日志条目300可以包括与特定日志条目有关的各种参数，包括但不限于，传输类型(例如，远程直接存储器访问(RDMA)、光纤通道、TCP网络)、地址族(例如，IPv4、Ipv6、IB、FC)、端口ID、子系统类型、传输需求、控制器ID、管理最大SQ大小、传输服务标识符、NVM子系统NQN、传输地址(存储装置的IP地址)和传输特定地址子类型。图2中所示的发现日志页面200可以包含一个或多个发现日志条目300。

[0038] 主计算机可以接收多于一个的发现日志页面。在这种情况下，主计算机检查发现日志页面的修订历史，并判断发现日志信息是否改变。根据一个实施例，主计算机可以从一个或多个发现服务器接收多个发现日志页面。主计算机可以以特定的顺序(例如，以增加日志页面偏移值的顺序)来读取接收到的发现日志页面，并对生成计数器计数。生成计数器从值0h开始指示发现信息的版本。对于发现日志中的每次变化，生成计数器都增加1。如果超过最大计数，则生成计数器恢复为初始值0h。如果生成计数器与先前读取的日志信息不匹配，则主计算机丢弃先前读取的日志页面，并相应地更新日志信息。如果在日志页面读取期间日志页面内容改变，则发现控制器可以返回指示发现重新启动的状态。根据发现控制器

指示的发现重新启动，主计算机停止当前的发现过程，并重新启动新的发现过程。

[0039] 在从发现服务器获取最新的发现日志条目之后，主计算机使用包含在相应日志条目中的目标存储装置的IP地址，并且通过初始化队列对来准备传输级连接从而与目标存储装置进行通信。通过这样做，主计算机发出连接命令来创建控制器队列，以建立与目标存储装置的传输级连接。连接命令可以指定各种参数，包括但不限于，端口ID、队列类型(例如，管理队列或I/O队列)、提交和完成队列的大小、队列属性、主机NQN、NVM子系统NQN和主机标识符。作为响应，目标存储装置向主计算机发送指示连接是否成功建立的连接响应。连接响应可以包含被分配用来与主计算机进行通信的存储装置控制器ID。

[0040] 在超大规模数据中心环境中，在主计算机与目标存储装置之间建立适当连接所需的信息可以经由传输协议所支持的架构(例如图1的架构120)来交换。主计算机110与发现服务器150之间，主计算机110与存储装置160之间以及DHCP服务器130与存储装置160之间的底层架构120可以不同，并且在主计算机110、DHCP服务器130、发现服务器150和存储装置160之间，底层架构120不需要相同。例如，本分布式存储系统可以包括分布在企业网络、数据中心和/或因特网上的成百上千个主计算机和存储装置。本分布式存储系统的一些子系统可以经由第一类型架构来连接，而本分布式存储系统的其他子系统可以经由不同于第一类型架构的第二类型架构来连接。例如，发现服务器可以驻留在公共网络(例如，因特网)上，而主计算机和目标存储装置可以通过专用网络连接。

[0041] 随着主计算机和存储装置的数量扩大和扩展，可以使用基于DHCP服务的本发现服务。根据一个实施例，数据中心中的每个机架可以运行机架级发现服务器，该机架级发现服务器可以自动更新和维护机架专用的存储装置的发现信息。POD可以包括多个机架，并且每个机架级发现服务器可以自动地将机架级发现信息更新到在POD管理器上运行的POD级发现服务器。POD级发现服务器允许POD管理器根据POD级发现信息来调度工作负载。

[0042] 根据一个实施例，本分布式存储系统可以包括多个POD管理器，每个POD有一个POD管理器。可以根据可在POD管理器之间交换的系统级发现信息来在多个POD管理器之间分配工作负载。在一个替代实施例中，POD管理器可以与DHCP服务器通信，并且与其他POD管理器协作以执行工作负载分配。例如，在目标存储装置的停机期间，可以从远程存储装置获得备份数据。在这种情况下，POD管理器可以使用系统级发现信息来将工作负载调度到备份存储装置。在基于NVMeoF的存储系统的情况下，任何POD管理器都可以连接到一个或多个发现服务器，并使用获取日志命令来取回可用存储装置的更新发现信息。一旦使用发现信息建立了通信路径，目标存储装置可以服务于从主计算机接收的I/O请求。因为本发现服务不需要手动更新，所以本发现服务对于超大规模数据中心是可扩展的，该超大规模数据中心可以包括许多机架和POD，且机架中包括许多计算装置和/或存储装置。

[0043] 图4示出根据一个实施例的用于提供发现日志信息的示例流程图。存储装置在启动时运行DHCP客户端，并向DHCP服务器发送请求以租用IP地址(401)。响应于DHCP客户端的请求，DHCP服务器将动态或永久的IP地址分配给存储装置的存储装置控制器(402)。在存储装置上运行的DLU使用RESTful API或GraphQL请求来将包括存储装置控制器的IP地址和端口地址的发现日志条目发送到发现服务器(403)。发现服务器提供包括多个存储装置的一个或多个发现日志条目的发现日志页面(404)。发现服务器将发现日志页面发送到主计算机(405)。

[0044] 图5示出了根据一个实施例的用于在主计算机与目标存储装置之间建立通信的示例流程图。主计算机将发现信息的请求发送到发现服务器(501)。发现服务器存储从多个存储装置的DLU接收到的一个或多个发现日志条目。主计算机从发现服务器接收包括一个或多个发现日志条目的发现日志页面(502)。使用在发现日志页面中接收到的一个或多个发现日志页面条目,主计算机建立与目标存储装置的连接(503),并且将工作负载调度到目标存储装置(504)。

[0045] 图6示出根据一个实施例的自动发现的示例过程。存储装置660的DHCP客户端向DHCP服务器630发送请求以获得IP地址(601),并且DHCP服务器630将IP地址分配并发送到存储装置660的DHCP客户端(602)。存储装置660(或DHCP服务器630)的DLU使用RESTful API或GraphQL请求来自动将存储装置660的发现日志条目发送到发现服务器650(603)。发现服务器650从多个存储装置的每个DLU收集一个或多个发现日志条目,并且准备包括一个或多个发现日志条目的发现日志页面。主计算机610将获取日志页面命令发送到发现服务器650(611),并且作为回应,发现服务器650将发现日志页面发送到主计算机610(612)。使用包含在发现日志页面中的信息,主计算机610建立与存储装置660的连接(613)。

[0046] 根据一个实施例,一种存储系统包括:多个存储装置,所述多个存储装置中的每一个被配置为运行动态主机配置协议(DHCP)客户端;以及发现服务器,其被配置为向主计算机提供发现日志页面。在多个存储装置中的每一个上运行的发现日志更新器将相应存储装置的发现日志条目提供给发现服务器。所述发现日志页面包含所述多个存储装置的发现信息,所述发现信息包括一个或多个发现日志条目,并且所述一个或多个发现日志条目中的每一个包括相应存储装置的被分配用于在主计算机与相应存储装置之间建立通信的地址。主计算机利用发现日志页面中包含的目标存储装置的发现日志条目来建立与目标存储装置的连接,并向目标存储装置提供输入/输出(I/O)或计算工作负载。

[0047] DHCP服务器可以为多个存储装置中的每个存储装置分配地址,并且多个存储装置中的各发现日志更新器可以使用所分配的地址来自动将其各自的包括分配地址的发现日志条目发送到发现服务器。

[0048] 发现日志条目还可以包括存储装置控制器的端口地址。

[0049] 可以根据NVMeoF规范来分配端口地址。

[0050] 发现服务器可以使用RESTful应用编程接口(API)或GraphQL请求来将发现日志页面发送到主计算机。

[0051] 多个存储装置中的至少一个存储装置可以是架构上快速非易失性存储器(NVMeoF)装置,并且主计算机可以将获得日志命令发送到发现服务器以接收发现日志页面。

[0052] 发现服务器和DHCP服务器可以并置在存储系统的同一服务器中。

[0053] 多个存储装置可以附接到机架,并且机架中的发现服务器可以收集机架专用的机架级发现信息。

[0054] 多个存储装置可以分布在POD中的多个机架之中,并且负责POD的发现服务器可以收集POD专用的POD级发现信息。

[0055] POD的POD管理器可以基于POD级发现信息来在多个存储装置之间调度工作负载。

[0056] POD管理器可以与存储系统的另一个POD管理器交换POD级发现信息。

[0057] 根据另一个实施例，一种方法包括：运行在存储装置中的动态主机配置协议(DHCP)客户端；将存储装置的发现日志条目提供给发现服务器；将发现日志页面从发现服务器提供至主计算机，其中发现日志页面包含多个存储装置的发现信息，所述发现信息包括一个或多个发现日志条目，并且所述一个或多个发现日志条目中的每一个包括相应存储装置的被分配用于在主计算机与对应存储装置之间建立通信的地址；使用包含在发现日志页面中的目标存储装置的发现日志条目来建立主计算机与目标存储装置之间的连接；以及将输入/输出(I/O)或计算工作负载从主计算机提供到目标存储装置。

[0058] DHCP服务器可以为多个存储装置中的每个存储装置分配地址，并且多个存储装置中的各发现日志更新器可以使用所分配的地址来将其各自的包括分配地址的发现日志条目自动发送至发现服务器。

[0059] 发现日志条目还可以包括存储装置控制器的端口地址。

[0060] 可以根据NVMeoF规范来分配端口地址。

[0061] 发现服务器可以使用RESTful应用编程接口(API)或GraphQL请求来将发现日志页面发送到主计算机。

[0062] 多个存储装置中的至少一个存储装置可以是架构上快速非易失性存储器(NVMeoF)装置，并且主计算机可以将获取日志命令发送至发现服务器以接收发现日志页面。

[0063] 发现服务器和DHCP服务器可以并置在存储系统的同一服务器中。

[0064] 多个存储装置可以附接到机架，并且机架中的发现服务器可以收集机架专用的机架级发现信息。

[0065] 多个存储装置可以分布在POD中的多个机架中，并且负责POD的发现服务器可以收集POD专用的POD级发现信息。

[0066] POD的POD管理器可以基于POD级发现信息来在多个存储装置之间调度工作负载。

[0067] POD管理器可以与存储系统的另一个POD管理器交换POD级发现信息。

[0068] 以上已经描述了上述示例实施例以说明实施用于在超大规模数据中心环境中提供自动以太网存储发现的系统和方法的各种实施例。本领域的普通技术人员将会想到对所公开的示例实施例的各种修改和偏离。在以下权利要求中阐述了旨在落入本发明范围内的主题。

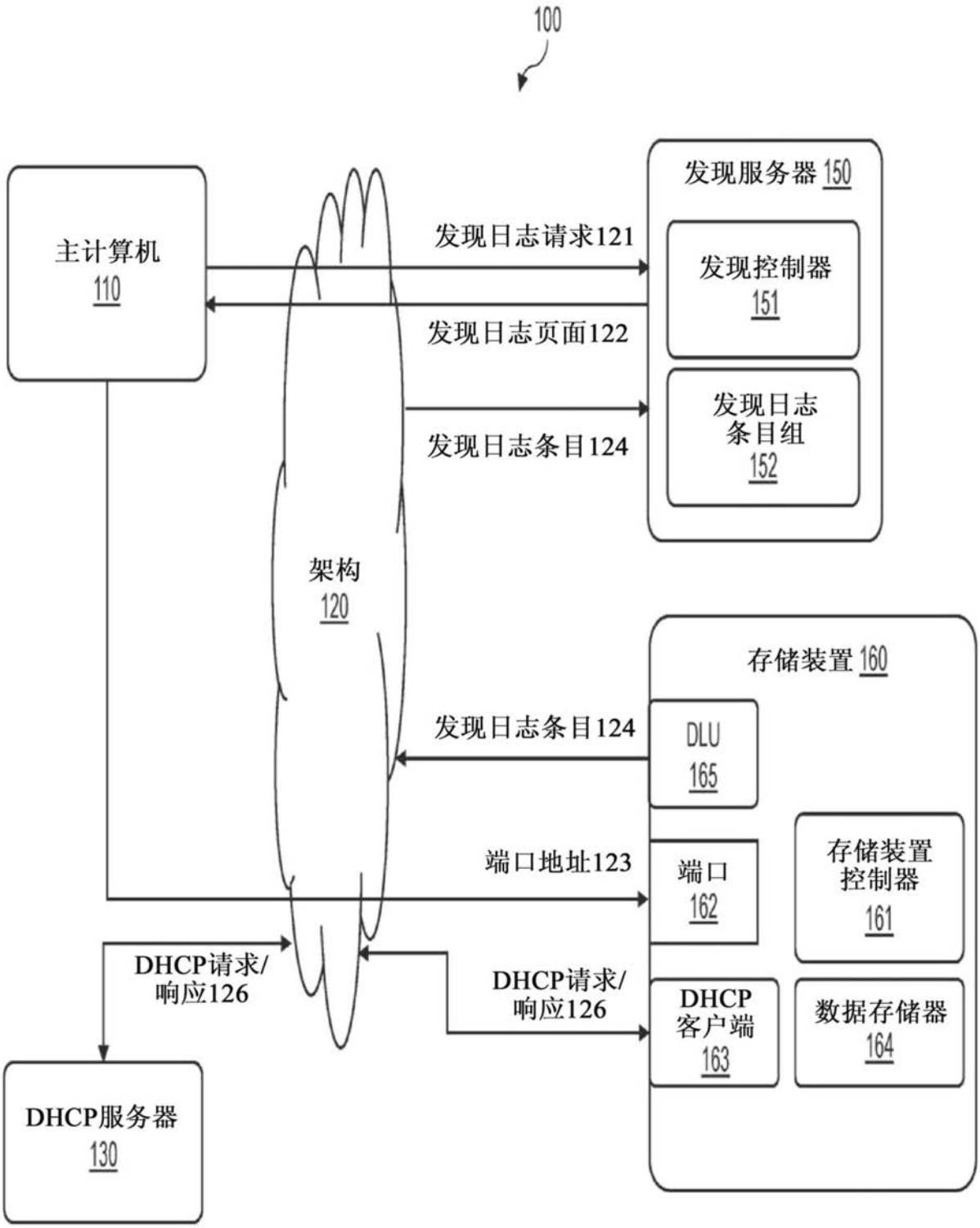
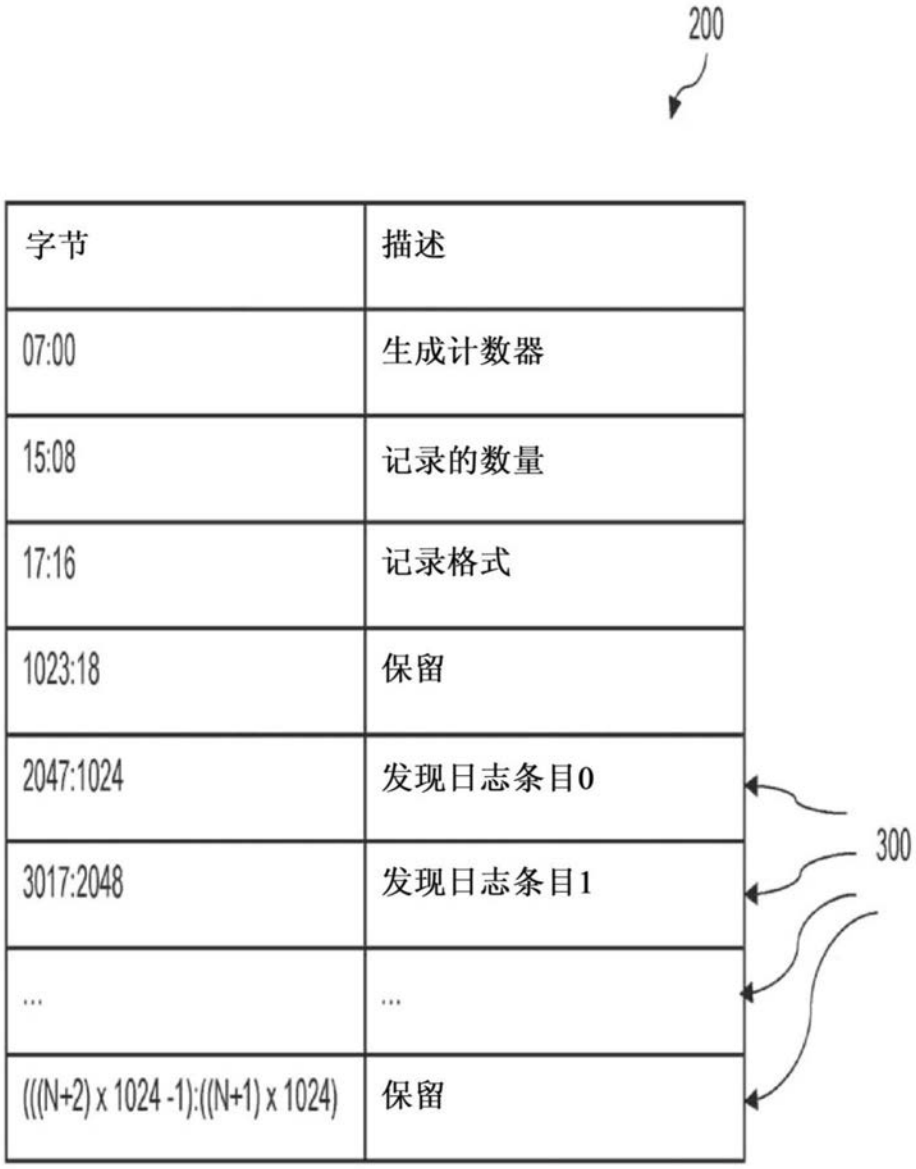

100
主计算机
110
发现服务器 150
发现控制器
151
发现日志
条目组
152
发现日志请求121
发现日志页面122
发现日志条目124
架构
120
存储装置 160
DLU
165
发现日志条目124
端口
162
端口地址123
存储装置
控制器
161
DHCP请求/
响应126
DHCP请求/
响应126
DHCP
客户端
163
数据存储器
164
DHCP服务器
130

图1



字节	描述
07:00	生成计数器
15:08	记录的数量
17:16	记录格式
1023:18	保留
2047:1024	发现日志条目0
3017:2048	发现日志条目1
...	...
(((N+2) x 1024 -1):((N+1) x 1024)	保留

图2

300

字节	描述
00	传输类型
01	地址族
02	子系统类型
03	传输需求
05:04	端口ID
07:06	控制器ID
09:08	管理最大SQ大小
31:10	保留
63:32	传输服务标识符
255:64	保留
511:256	NVM子系统限定名（NQN）
767:512	传输地址
1023:768	传输特定地址子类型

图3

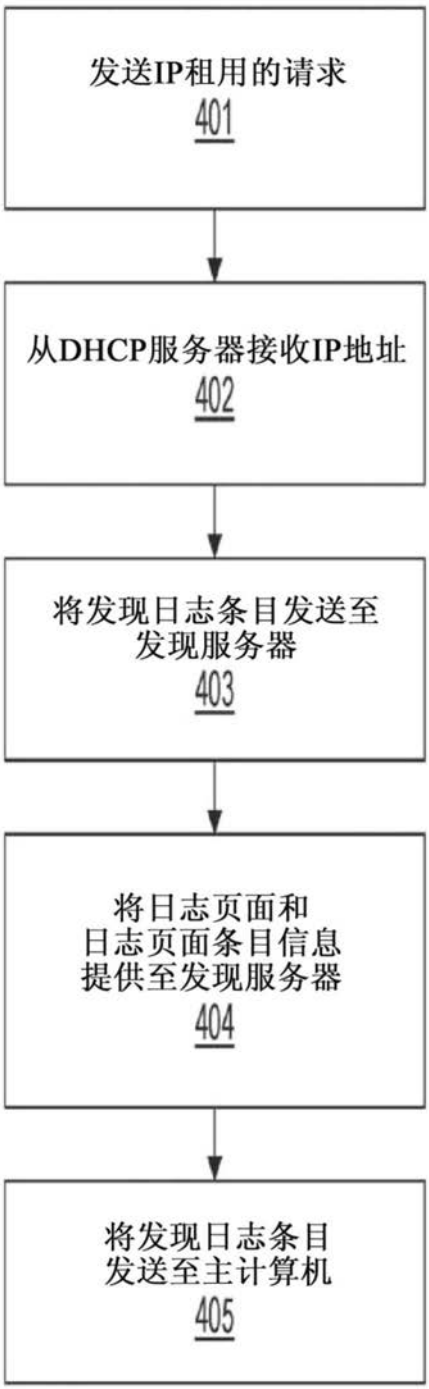

发送IP租用的请求
401
从DHCP服务器接收IP地址
402
将发现日志条目发送至
发现服务器
403
将日志页面和
日志页面条目信息
提供至发现服务器
404
将发现日志条目
发送至主计算机
405

图4

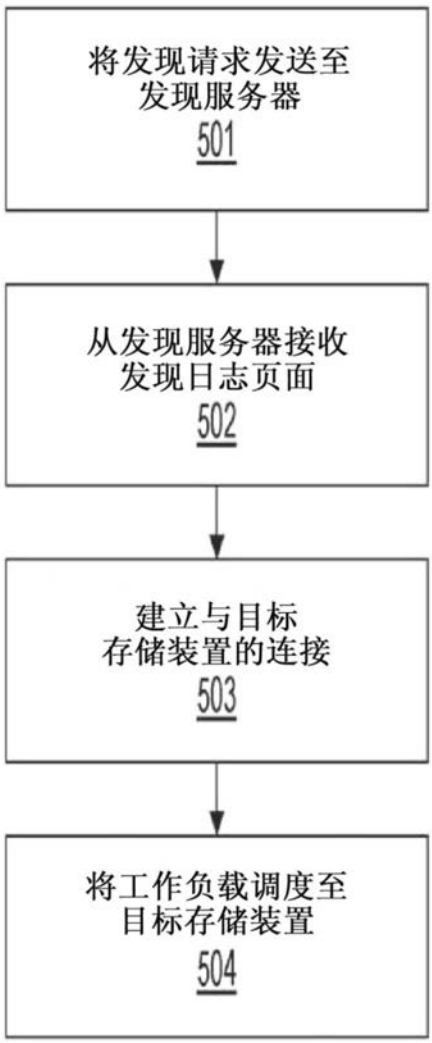

将发现请求发送至
发现服务器
501
从发现服务器接收
发现日志页面
502
建立与目标
存储装置的连接
503
将工作负载调度至
目标存储装置
504

图5

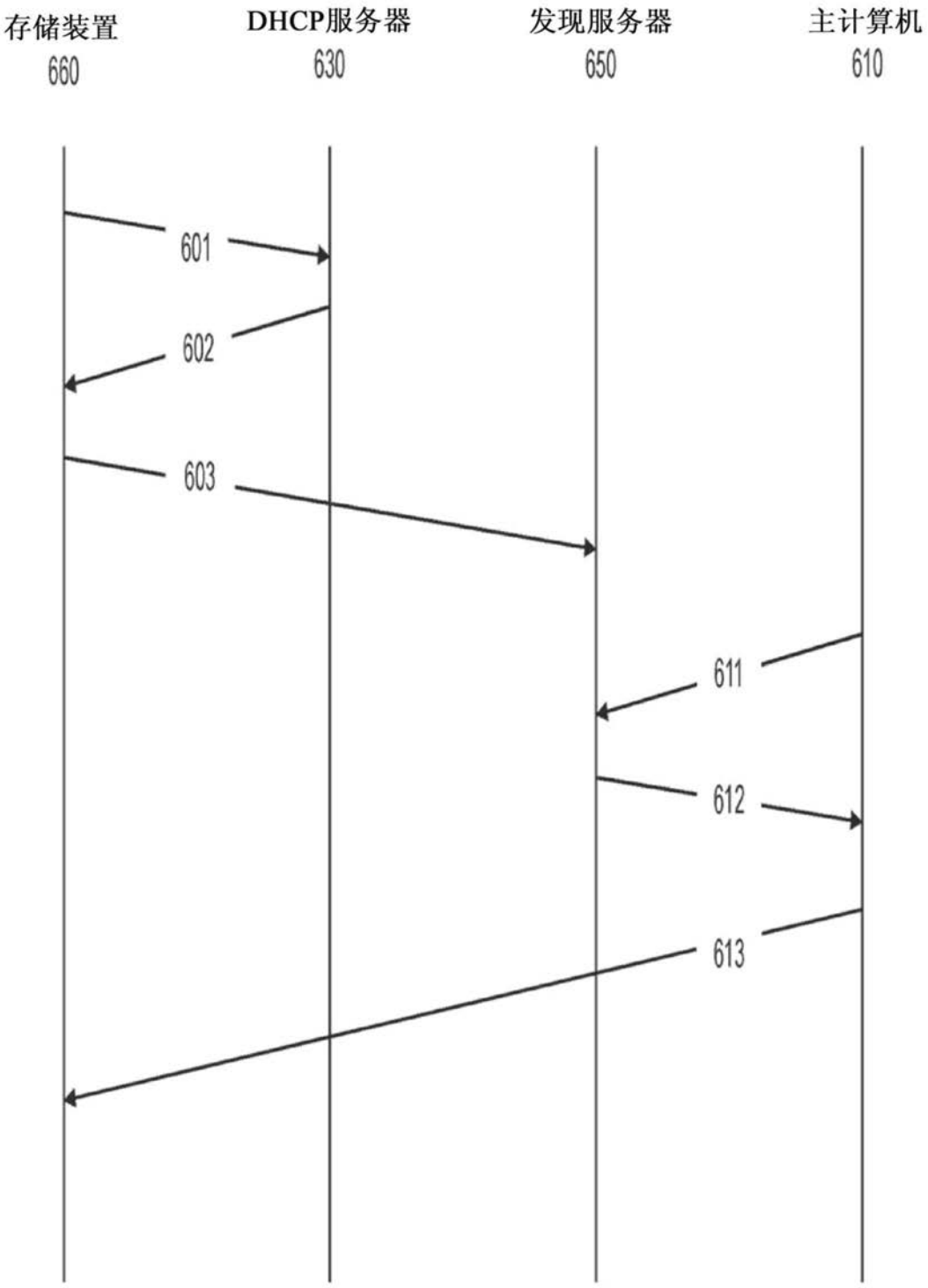

存储装置
660
DHCP服务器
630
发现服务器
650
主计算机
610
601
602
603
611
612
613

图6