I729256

(19)中華民國智慧財產局

(12)發明說明書公告本　**(11)證書號數：TW I729256 B**

(45)公告日： 中華民國 110 (2021) 年 06 月 01 日

(21)申請案號：107100839　(22)申請日： 中華民國 107 (2018) 年 01 月 10 日

(51)Int. Cl.：　*H04L12/24　(2006.01)*　*H04L29/12　(2006.01)*

(30)優先權：2017/03/16　美國　62/472,379

2017/06/06　美國　15/615,508

(71)申請人：南韓商三星電子股份有限公司 (南韓) SAMSUNG ELECTRONICS CO., LTD.　(KR)

南韓

(72)發明人：伯拉普蘭　奇納克里許納 BALLAPURAM, CHINNAKRISHNAN (IN)

(74)代理人：林孟閱；盧姵君；陳怡如

(56)參考文獻：

US　9467513B2　US　20040068561A1

US　20140289377A1　US　20160072761A1

審查人員：黎芷婷

申請專利範圍項數：22 項　圖式數：6　共 37 頁

(54)名稱

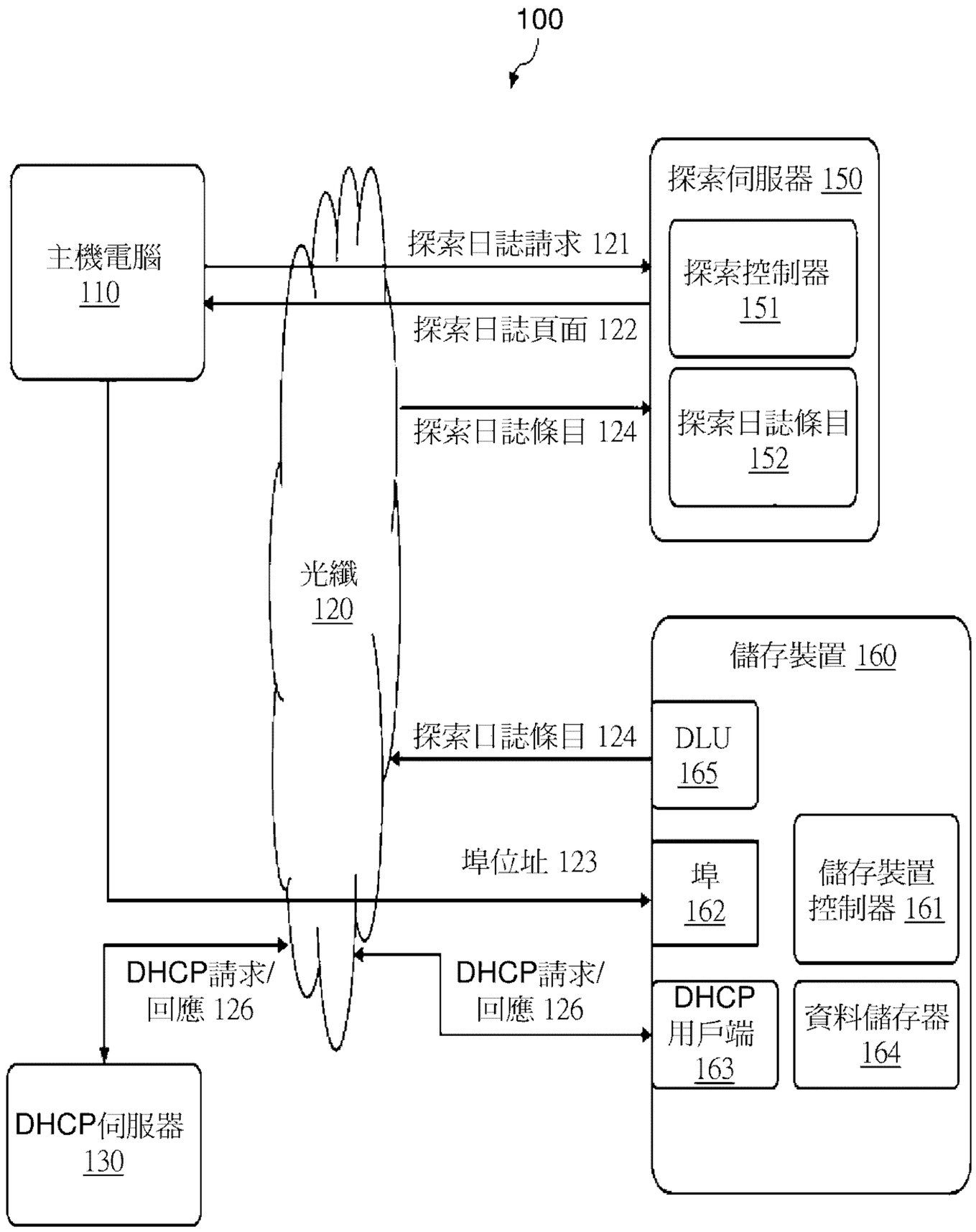
用於自動乙太儲存探索的儲存系統和方法

(57)摘要

一種方法包括：在儲存裝置中運行動態主機設定協定 DHCP 用戶端；向探索伺服器提供儲存裝置的探索日誌條目；將探索日誌頁面自探索伺服器提供至主機電腦，其中探索日誌頁面含有多個儲存裝置的探索資訊，探索資訊包括一或多個探索日誌條目，且一或多個探索日誌條目中的每一者包括對應的儲存裝置的位址，位址被指派以在主機電腦與對應的儲存裝置之間建立通訊；使用探索日誌頁面中所含有的目標儲存裝置的探索日誌條目在主機電腦與目標儲存裝置之間建立連接；以及將輸入/輸出或計算工作負載自主機電腦提供至目標儲存裝置。

A method includes: running a dynamic host configuration protocol (DHCP) client in a storage device; providing a discovery log entry of the storage device to a discovery server; providing a discovery log page from the discovery server to a host computer, wherein the discovery log page contains discovery information of a plurality of storage devices including one or more discovery log entries, and each of the one or more discovery log entries includes an address of the corresponding storage device that is assigned to establish communication between the host computer and the corresponding storage device; establishing a connection between the host computer and a target storage device using the discovery log entry of the target storage device contained in the discovery log page; and providing input/output (I/O) or compute workloads from the host computer to the target storage device.

指定代表圖：



【圖1】

符號簡單說明：

100:分散式儲存系統

110:主機電腦

120:光纖

121:探索日誌請求

122:探索日誌頁面

123:埠位址

124、152:探索日誌條目

126:動態主機設定協定請求/回應

130:動態主機設定協定伺服器

150:探索伺服器

151:探索控制器

160:儲存裝置

161:儲存裝置控制器

162:埠

163:動態主機設定協定用戶端

164:資料儲存器

165:探索日誌更新器

專利案號：107100839

107年4月16日 所提修正

公告本

107-04-16

I729256

【發明摘要】

【中文發明名稱】用於自動乙太儲存探索的儲存系統和方法

【英文發明名稱】STORAGE SYSTEM AND METHOD FOR AUTOMATIC ETHERNET STORAGE DISCOVERY

【中文】一種方法包括：在儲存裝置中運行動態主機設定協定DHCP用戶端；向探索伺服器提供儲存裝置的探索日誌條目；將探索日誌頁面自探索伺服器提供至主機電腦，其中探索日誌頁面含有多個儲存裝置的探索資訊，探索資訊包括一或多個探索日誌條目，且一或多個探索日誌條目中的每一者包括對應的儲存裝置的位址，位址被指派以在主機電腦與對應的儲存裝置之間建立通訊；使用探索日誌頁面中所含有的目標儲存裝置的探索日誌條目在主機電腦與目標儲存裝置之間建立連接；以及將輸入/輸出或計算工作負載自主機電腦提供至目標儲存裝置。

【英文】A method includes: running a dynamic host configuration protocol (DHCP) client in a storage device; providing a discovery log entry of the storage device to a discovery server; providing a discovery log page from the discovery server to a host computer, wherein the discovery log page contains discovery information of a plurality of storage devices including one or more discovery log entries, and each of the one or more discovery log entries includes an

address of the corresponding storage device that is assigned to establish communication between the host computer and the corresponding storage device; establishing a connection between the host computer and a target storage device using the discovery log entry of the target storage device contained in the discovery log page; and providing input/output (I/O) or compute workloads from the host computer to the target storage device.

【指定代表圖】圖1。

【代表圖之符號簡單說明】

100：分散式儲存系統

110：主機電腦

120：光纖

121：探索日誌請求

122：探索日誌頁面

123：埠位址

124、152：探索日誌條目

126：動態主機設定協定請求/回應

130：動態主機設定協定伺服器

150：探索伺服器

151：探索控制器

160：儲存裝置

161：儲存裝置控制器

162：埠

163：動態主機設定協定用戶端

164：資料儲存器

165：探索日誌更新器

【特徵化學式】

無

【發明說明書】

【中文發明名稱】用於自動乙太儲存探索的儲存系統和方法

【英文發明名稱】STORAGE SYSTEM AND METHOD FOR AUTOMATIC ETHERNET STORAGE DISCOVERY

【技術領域】

【0001】 本發明大體而言是有關於分散式儲存裝置，更具體而言，是有關於一種用於在超大規模資料中心環境（hyperscale datacenter environment）中提供對分散式儲存裝置的自動探索的系統及方法。

[相關申請案的交叉參考]

本申請案主張於 2017 年 3 月 16 日提出申請的美國臨時專利申請案第 62/472,379 號的權利及優先權，所述美國臨時專利申請案的揭露內容全部併入本申請案供參考。

【先前技術】

【0002】 資料中心中的每一機架（rack）含有用於計算裝置及儲存裝置的滑板（sled）及/或託盤（tray）的組合。超大規模資料中心可包括若干效能最佳化資料中心（performance optimized datacenter，POD），且每一效能最佳化資料中心可包括多個機架以及成百上千個計算裝置及/或儲存裝置。

【0003】 高速非揮發性記憶體（non-volatile memory express，NVMe）界定了供主機軟體經由高速周邊組件互連（peripheral component interconnect express，PCIe）匯流排來與非揮發性記憶體子系統（例如，固態磁碟機（solid-state drive，SSD））進行通訊的暫存器級介面。光纖上高速非揮發性記憶體（NVMe over fabric，NVMeoF）（或簡稱為 NVMf）界定一種藉由廣泛的儲存網路連結光纖（storage networking fabric）（例如乙太網路（Ethernet）、光纖通道（Fibre Channel）、無限頻帶（InfiniBand）、傳輸控制協定（transmission control protocol，TCP）網路、及其他種網路光纖）來支援高速非揮發性記憶體區塊儲存協定的常用架構。

【0004】 在基於光纖上高速非揮發性記憶體（NVMeoF-based）的儲存系統中，每一儲存裝置（亦被稱為與光纖上高速非揮發性記憶體相容的固態磁碟機或乙太固態磁碟機）是藉由其子系統高速非揮發性記憶體定質名字（NVMe Qualified Name，NQN）、網際網路協定（Internet Protocol，IP）位址、埠識別符（identifier，ID）及/或控制器識別符來識別。每一效能最佳化資料中心管理器需要知曉光纖（例如，乙太網路）上的儲存裝置的狀態及可用性，以對工作負載進行分配及排程。儲存裝置中的一些可能停工以進行維護，且工作負載不能被指派給該些儲存裝置。此外，在超大規模資料中心環境中自分散於光纖上的成百上千個儲存裝置收集、維持及更新探索資訊並非是一項輕鬆的任務。

【發明內容】

【0005】 根據一個實施例，一種儲存系統包括：多個儲存裝置，所述多個儲存裝置中的每一者被配置成運行動態主機設定協定（dynamic host configuration protocol，DHCP）用戶端；以及探索伺服器，被配置成向主機電腦提供探索日誌頁面。在所述多個儲存裝置中的每一者上運行的探索日誌更新器將對應的所述儲存裝置的探索日誌條目提供至所述探索伺服器。所述探索日誌頁面含有所述多個儲存裝置的探索資訊，所述探索資訊包括一或多個探索日誌條目，且所述一或多個探索日誌條目中的每一者包括對應的所述儲存裝置的位址，所述位址被指派以在所述主機電腦與對應的所述儲存裝置之間建立通訊。所述主機電腦使用所述探索日誌頁面中所含有的目標儲存裝置的所述探索日誌條目來與所述目標儲存裝置建立連接，並向所述目標儲存裝置提供輸入/輸出（input/output，I/O）或計算工作負載。

【0006】 根據另一實施例，一種方法包括：在儲存裝置中運行動態主機設定協定（DHCP）用戶端；向探索伺服器提供所述儲存裝置的探索日誌條目；將探索日誌頁面自所述探索伺服器提供至主機電腦，其中所述探索日誌頁面含有多個儲存裝置的探索資訊，所述探索資訊包括一或多個探索日誌條目，且所述一或多個探索日誌條目中的每一者包括對應的所述儲存裝置的位址，所述位址被指派以在所述主機電腦與對應的所述儲存裝置之間建立通訊；

使用所述探索日誌頁面中所含有的目標儲存裝置的所述探索日誌條目在所述主機電腦與所述目標儲存裝置之間建立連接；以及將輸入/輸出（I/O）或計算工作負載自所述主機電腦提供至所述目標儲存裝置。

【0007】 現在將參照附圖更具體地闡述並在申請專利範圍中指出以上及其他較佳特徵，包括事件的實施方案及組合的各種新穎細節。應理解，本文所述的特定系統及方法是僅以說明而非限制的方式示出。如熟習此項技術者應理解，在不背離本發明的範圍的條件下，可在各種及眾多實施例中採用本文所述的原理及特徵。

【圖式簡單說明】

【0008】 作為本說明書的一部分被包含的附圖說明目前較佳的實施例，且與上文所給出的大體說明及下文所給出的對較佳實施例的詳細說明一起用於解釋及教示本文所述的原理。

圖 1 示出根據一個實施例的示例性分散式儲存裝置系統的方塊圖。

圖 2 示出根據一個實施例的示例性探索日誌頁面的資料結構。

圖 3 示出根據一個實施例的探索日誌頁面中所含有的示例性日誌條目的資料結構。

圖 4 示出根據一個實施例用於將探索日誌資訊填入的示例性流程圖。

圖 5 示出根據一個實施例用於在主機電腦與目標儲存裝置之

間建立通訊的示例性流程圖。

圖 6 示出根據一個實施例的示例性自動探索過程。

【0009】 各圖未必是按比例繪製，且出於說明性目的，在所有各圖中，相似結構或功能的元件通常由相同的參考編號表示。各圖僅旨在便於對本文所述的各種實施例進行說明。各圖並未闡述本文所揭露教示內容的每一態樣，且並不限制申請專利範圍的範圍。

【實施方式】

【0010】 本文所揭露的特徵及教示內容中的每一者可分別地或者結合其他特徵及教示內容來加以利用，以提供一種用於在超大規模資料中心環境中提供自動乙太儲存探索的系統及方法。將參照附圖更詳細地闡述利用該些附加特徵及教示內容中的諸多特徵及教示內容（分別地及以組合方式）的代表性實例。此詳細說明僅旨在向熟習此項技術者教示用於實踐本教示內容各態樣的其他細節，而不旨在限制申請專利範圍的範圍。因此，在最廣泛意義上來看，以上在詳細說明中所揭露的特徵的組合可並非是實踐教示內容所必需的，而對所述組合的教示僅是為了具體闡述本教示內容的代表性實例。

【0011】 在以下說明中，僅出於解釋的目的，陳述具體術語以提供對本發明的透徹理解。然而，對於熟習此項技術者而言將顯而易見的是，實踐本發明的教示內容並不需要該些具體細節。

【0012】 本文詳細說明的一些部分是依據對電腦記憶體內的資料

位元進行的運算的演算法及符號表示而呈現。該些演算法說明及表示由熟習資料處理技術者用於向其他熟習此項技術者有效地傳達其工作的本質。演算法在此處且通常被設想為得到所期望結果的自相一致的一系列步驟。所述步驟是需要對物理數量進行物理操縱的步驟。通常，雖然未必需要，然而該些數量採取能夠被儲存、傳遞、組合、比較、及以其他方式被操縱的電訊號或磁性訊號的形式。有時，原則上出於常用的原因，將該些訊號稱為位元、值、元素、符號、字元、項、數值等已證明為方便的。

【0013】 然而，應牢記於心，所有該些及類似用語將與適當的物理數量相關聯，且僅為用於該些數量的方便標簽。除非自以下論述顯而易見另有具體陳述，否則應瞭解，在本說明通篇中，利用例如「處理」、「計算（computing、calculating）」、「確定」、「顯示」等用語進行的論述指代電腦系統或類似電子計算裝置的如下動作及過程：將電腦系統的暫存器及記憶體內被表示為物理（電子）數量的資料操縱及變換成電腦系統記憶體或暫存器內或者其他此類資訊儲存裝置、傳輸裝置或顯示裝置內類似地被表示為物理數量的其他資料。

【0014】 此外，代表性實例及附屬請求項的各種特徵可以未具體及明確枚舉的方式進行組合，以提供本教示內容的附加的有用實施例。亦應明確地指出，出於原始揭露內容的目的以及出於限制所主張標的物的目的，所有值範圍或實體群組指示揭露每一可能的中間值或中間實體。亦應明確地指出，各圖所示組件的尺寸及

形狀被設計成有助於理解本教示內容是如何實踐的，而不旨在限制實例中所示的尺寸及形狀。

【0015】 本發明闡述一種包括多個資料儲存裝置的分散式資料儲存系統。本分散式資料儲存系統採用動態主機設定協定（DHCP）服務來提供、更新及維持所述多個資料儲存裝置的動態探索資訊。在分散式資料儲存系統的各子系統之間進行通訊時所支援的傳送協定的實例是光纖上高速非揮發性記憶體協定，且光纖是乙太網路。所述子系統包括一或多個計算節點（例如，遠端主機電腦或計算節點）及多個儲存裝置（例如，乙太固態磁碟機）。

【0016】 光纖上高速非揮發性記憶體標準界定了基於光纖上高速非揮發性記憶體的系統的探索服務，以容許主機電腦（本文中亦被稱為主機或計算節點）識別並定位主機電腦可存取的子系統（本文中亦被稱為儲存裝置或節點）。提供探索服務的探索伺服器是向基於光纖上高速非揮發性記憶體的系統的主機電腦公開一或多個探索控制器的子系統。探索伺服器的探索控制器中的每一者可對作出請求的主機電腦提供探索日誌頁面（discovery log page）。

【0017】 由探索控制器提供的探索日誌頁面可含有一或多個探索日誌條目。探索日誌頁面中所含有的每一探索日誌條目指定主機電腦在藉由高速非揮發性記憶體傳送協定連接至子系統時所必需的資訊。舉例而言，探索日誌條目指定對另一探索服務公開名稱空間（namespace）或轉介（referral）的子系統。根據光纖上高速非揮發性記憶體規範，最大轉介深度（maximum referral depth）

可為八個層級或更少。

【0018】 **圖 1** 示出根據一個實施例的示例性分散式儲存裝置系統的方塊圖。分散式儲存系統 100 包括主機電腦 110、動態主機設定協定（DHCP）伺服器 130、探索伺服器 150、儲存裝置 160、光纖 120，光纖 120 藉由所建立的網路提供主機電腦 110、動態主機設定協定伺服器 130、探索伺服器 150、及儲存裝置 160 之間的連接。光纖 120 的實例是乙太網路。

【0019】 主機電腦 110 可表示分散式儲存系統 100 的諸多主機電腦中的一者，且儲存裝置 160 可表示分散式儲存系統 100 的諸多儲存裝置中的一者。根據一個實施例，主機電腦 110 指代分散式儲存系統 100 的計算節點，所述計算節點可向分散式儲存系統 100 的包括儲存裝置 160 在內的儲存裝置發佈工作負載或者簡單地對分散式儲存系統 100 的包括儲存裝置 160 在內的儲存裝置中所儲存的資料執行讀取/寫入存取。

【0020】 儲存裝置 160 包括儲存裝置控制器 161、埠 162、動態主機設定協定用戶端 163、資料儲存器 164（例如固態磁碟機（SSD））、及探索日誌更新器（discovery log updater，DLU）165。

在一個實施例中，探索日誌更新器 165 是在儲存裝置 160 上運行以使用 RESTful 應用程式設計介面（application programming interface，API）或 GraphQL 請求而使用光纖傳送協定及埠將儲存裝置 160 的探索日誌條目更新於探索伺服器 150 的探索日誌條目 152 中的服務。舉例而言，探索日誌更新器 165 可使用傳輸控制協

定/網際網路協定或使用者資料報協定（User Datagram Protocol，UDP）/網際網路協定及傳輸控制協定/使用者資料報埠來將儲存裝置160的探索日誌條目更新於探索伺服器150的探索日誌條目152中。動態主機設定協定伺服器 130 是分散式儲存系統 100 的向作出請求的動態主機設定協定用戶端提供動態網際網路協定位址或永久租用（permanent lease）網際網路協定位址的子系統。舉例而言，儲存裝置 160 的動態主機設定協定用戶端 163 向動態主機設定協定伺服器 130 發送請求 126，且動態主機設定協定伺服器 130 因應於請求 126 而將網際網路協定位址（動態網際網路協定位址或永久網際網路協定位址）指派給動態主機設定協定用戶端 163。

【0021】 探索伺服器 150 向被配置成自探索伺服器 150 獲取探索日誌資訊的任何作出請求的主機電腦提供探索服務。探索伺服器150包括探索控制器151及探索日誌條目152以及在光纖上高速非揮發性記憶體探索伺服器規範中所述的與本文論述並不特別相關的其他組件。探索控制器 151 的網際網路協定位址或傳送位址被公開給分散式儲存系統 100 的包括主機電腦 110 在內的一或多個主機電腦。另外，主機可在對探索服務的連接命令中使用眾所熟知的探索服務高速非揮發性記憶體定質名字（例如，nqn.2014-08.org.nvmexpress.discovery）。

【0022】 主機電腦 110 可使用探索伺服器 150 的探索服務來查找連接至分散式儲存系統 100 內可用的一或多個儲存裝置所必需的資訊。主機電腦 110 可以若干方式獲得初始探索資訊。舉例而言，

新附裝的並不具有探索伺服器 150 的網際網路協定位址的主機電腦可自本地主機配置檔案或自主機電腦的作業系統（operating system，OS）的超管理器（hypervisor）或自主機電腦已知的任何其他子系統獲得初始探索資訊。主機電腦 110 亦可在對探索服務的連接命令中使用探索服務高速非揮發性記憶體定質名字（例如，nqn.2014-08.org.nvmexpress.discovery）位址。由主機電腦 110 用於獲得連接至探索服務所必需的高速非揮發性記憶體傳送資訊的探索方法是基於所使用的光纖協定。初始探索服務資訊可包括探索伺服器 150 的或具有關於探索伺服器 150 的資訊的另一子系統的網際網路協定位址。

【0023】 一旦知曉探索伺服器 150 的身份（例如，探索伺服器 150 的網際網路協定位址），主機電腦 110 便向探索伺服器 150 發送探索日誌請求 121（例如，根據光纖上高速非揮發性記憶體規範所作的獲取日誌頁面請求（Get Log Page request）），以按照需要獲得探索資訊。作為回應，探索伺服器 150 提供包括一或多個探索日誌條目 152 的探索日誌頁面 122。探索日誌頁面 122 中所含有的每一探索日誌條目指定主機電腦 110 在連接至一或多個對應儲存裝置 160 時所必需的資訊。舉例而言，每一條目包括對應儲存裝置 160 的一或多個識別符及在連接至儲存裝置 160 時所支援的傳送協定資訊。探索日誌條目 152 可支援多個傳送協定及位址類型。

【0024】 根據一個實施例，探索伺服器 150 可自行地向主機電腦 110 發送最新的探索日誌條目 152。舉例而言，探索伺服器 150 保

持主機電腦（在一種情形中，其可為負責效能最佳化資料中心/效能最佳化資料中心管理器的探索伺服器）的清單，且在一或多個探索日誌條目被更新時或以預定的間隔（例如，每天、每週）向所列出的主機電腦發送經更新的探索日誌頁面 122。在此種情形中，探索日誌頁面 122 可含有所儲存的探索日誌條目 152 的子集，所述子集僅包括經更新的探索日誌條目。

【0025】 基於在探索日誌頁面 122 中所識別的與儲存裝置 160 對應的探索日誌條目及傳送類型，主機電腦 110 可識別目標儲存裝置 160 的埠位址 123 並使用所指示的受支援傳送類型及協定而使用埠位址 123 來與目標儲存裝置 160 建立通訊路徑。探索日誌頁面 122 傳遞了在主機與目標儲存裝置之間起始連接所需的所有資訊。儲存裝置 160 的儲存裝置控制器 161 與主機電腦 110 可使用所建立的埠 162 來彼此進行通訊。

【0026】 在儲存裝置 160 上運行的動態主機設定協定用戶端 163 向動態主機設定協定伺服器 130 發送請求，且動態主機設定協定伺服器 130 將網際網路協定位址（或對應光纖中的與網際網路協定位址對應的類似位址）指派給作出請求的動態主機設定協定用戶端 163。探索日誌更新器 165 在每一目標儲存裝置上運行或針對目標儲存裝置群組而運行。舉例而言，對於底盤/機架中的儲存裝置 160 中的全部或一些，可使用一個或幾個動態主機設定協定用戶端 163。類似地，一個或幾個探索日誌更新器 165 可為底盤/機架中的一個儲存裝置 160 或一群組的儲存裝置 160 更新日誌。動

態主機設定協定及探索日誌更新器服務可在底盤/機架中於一個或幾個儲存裝置 160 上或者通用中央處理單元（central processing unit，CPU）上或者底板管理控制器（baseboard management controller，BMC）上運行。動態主機設定協定用戶端 163 或探索日誌更新器 165 可向探索伺服器 150 發送探索日誌條目 124，探索日誌條目 124 包括儲存裝置控制器 161 的被指派網際網路協定位址。根據一個實施例，動態主機設定協定伺服器 130 與探索伺服器 150 並置於同一伺服器中。在一些實施例中，動態主機設定協定伺服器 130 是位於分散式儲存系統 100 之外的與探索伺服器 150 不同的另一系統。

【0027】 探索伺服器 150 收集並更新包括一或多個探索日誌條目 152 的探索日誌資訊，所述一或多個探索日誌條目 152 是使用 RESTful 應用程式設計介面或 GraphQL 請求自目標儲存裝置的一或多個探索日誌更新器 165 接收。作為對光纖上高速非揮發性記憶體獲取日誌頁面請求的回應，探索伺服器 150 將探索日誌頁面提供至作出請求的主機電腦 110。由在每一目標儲存裝置或每一目標儲存裝置群組上運行的探索日誌更新器使用藉由動態主機設定協定用戶端被指派的網際網路協定位址對探索資訊的自動更新可以連接至駐留於本分散式儲存系統中的目標儲存裝置所必需的最新探索資訊向作出請求的主機電腦提供無縫式探索服務（seamless discovery service）。

【0028】 可能會添加新的儲存裝置，且由於新儲存裝置的添加、

對現有儲存裝置的替換或引退（retirement），作為資料中心環境中的常規操作過程的一部分，一些儲存裝置可能停工。由於受租用（on lease）的儲存裝置 160 各自的每一動態主機設定協定用戶端 163 可更新其自身的當前資訊，因此動態主機設定協定伺服器 130 可保持最新的探索日誌資訊。使用藉由在儲存裝置 160 中的每一者上運行的瘦型動態主機設定協定用戶端（thin DHCP client）163 對探索日誌資訊的自動化更新方案，本分散式儲存系統 100 可提供一種能夠適應超大規模資料中心環境中成百上千個儲存裝置的可按比例縮放解決方案。

【0029】 **圖 2** 示出根據一個實施例的示例性探索日誌頁面的資料結構。**圖 2** 所示的探索日誌頁面 200 含有與探索資訊有關的各種參數，包括但不限於產生計數器（generation counter）、記錄數目（a number of records）、記錄格式（record format）、一或多個探索條目 300。**圖 3** 示出**圖 2** 所示探索日誌頁面 200 中所含有的示例性日誌條目的資料結構。日誌頁面條目 300 可包括與特定日誌條目有關的各種參數，包括但不限於傳送類型（transport type）（例如，遠端直接記憶體存取（remote direct memory access，RDMA）、光纖通道（Fibre Channel）、傳輸控制協定網路（TCP network））、位址家族（address family）（例如，IPv4、Ipv6、無限頻帶（IB）、光纖通道（FC））、埠識別符（port ID）、子系統類型（subsystem type）、傳送要求（transport requirement）、控制器識別符（controller ID）、管理最大 SQ 大小（admin max SQ size）、傳送服務識別符（transport

service identifier）、非揮發性記憶體子系統高速非揮發性記憶體定質名字、傳送位址（transport address）（儲存裝置的網際網路協定位址）、及傳送特有的位址子類型（transport specific address sybtype）。**圖 2** 所示的探索日誌頁面 200 可含有一或多個日誌條目 300。

【0030】 主機電腦可接收多於一個探索日誌頁面。在此種情形中，主機電腦檢查探索日誌頁面的修訂歷史，並判斷探索日誌資訊中是否存在變化。根據一個實施例，主機電腦可自一或多個探索伺服器接收多個探索日誌頁面。主機電腦可以特定次序（例如，隨著日誌頁面偏移值（log page offset value）的增加）讀取所接收的探索日誌頁面，並使產生計數器進行計數。產生計數器以值 0h 開始指示探索資訊的版本。對於探索日誌中的每次變化，使產生計數器遞增 1。若超過最大計數，則產生計數器回繞（wrap）至初始值 0h。若產生計數器與先前所讀取的日誌資訊不匹配，則主機電腦摒棄先前所讀取的日誌頁面並相應地更新日誌資訊。若日誌頁面內容在日誌頁面讀取期間發生變化，則探索控制器可傳回指示探索重新開始的狀態。根據探索控制器所指示的探索重新開始，主機電腦停止當前的探索過程並重新開始新的探索過程。

【0031】 在自探索伺服器擷取最新的探索日誌條目之後，主機電腦使用對應日誌條目中所含有的目標儲存裝置的網際網路協定位址，並藉由將佇列對（queue pair）初始化來準備傳送級連接（transport-level）以與目標儲存裝置進行通訊。如此一來，主機

電腦發佈連接命令，以創建控制器佇列來與目標儲存裝置建立傳送級連接。所述連接命令可指定各種參數，包括但不限於埠識別符、佇列的類型（例如，管理佇列或輸入/輸出（I/O）佇列）、提交及完成佇列（submission and completion queue）的大小、佇列屬性（queue attribute）、主機高速非揮發性記憶體定質名字（host NQN）、非揮發性記憶體子系統高速非揮發性記憶體定質名字（NVM subsystem NQN）、及主機識別符（host identifier）。作為回應，目標儲存裝置向主機電腦發送指示是否成功建立連接的連接回應（connect response）。連接回應可含有被分配用於與主機電腦進行通訊的儲存裝置控制器識別符。

【0032】 在超大規模資料中心環境中，可藉由傳送協定所支援的光纖（例如，**圖 1** 所示光纖 120）來交換在主機電腦與目標儲存裝置之間建立恰當連接所需的資訊。基礎的光纖 120 可在主機電腦 110 與探索伺服器 150 之間、在主機電腦 110 與儲存裝置 160 之間以及在動態主機設定協定伺服器 130 與儲存裝置 160 之間為不同的，且不需要在主機電腦 110、動態主機設定協定伺服器 130、探索伺服器 150 及儲存裝置 160 之間為相同的。舉例而言，本分散式儲存系統可包括分散於企業網路、資料中心及/或網際網路上的數百個或數千個主機電腦及儲存裝置。本分散式儲存系統的一些子系統可藉由第一類型的光纖進行連接，而本分散式儲存系統的其他子系統可藉由與第一類型的光纖不同的第二類型的光纖進行連接。舉例而言，探索伺服器可駐留於公用網路（例如，網際網

路）上，且主機電腦及目標儲存裝置可藉由私用網路進行連接。

【0033】 隨著主機電腦及儲存裝置的數目擴展及按比例放大，可使用基於動態主機設定協定服務的本探索服務。根據一個實施例，資料中心中的每一機架可運行機架級（rack-level）探索伺服器，所述機架級探索伺服器可自動地更新及維持機架特有的儲存裝置的探索資訊。效能最佳化資料中心可包括多個機架，且每一機架級探索伺服器可自動地將機架級探索資訊更新至在效能最佳化資料中心管理器上運行的效能最佳化資料中心級探索伺服器。效能最佳化資料中心級探索伺服器容許效能最佳化資料中心管理器基於效能最佳化資料中心級探索資訊來對工作負載進行排程。

【0034】 根據一個實施例，本分散式儲存系統可包括多個效能最佳化資料中心管理器，每一效能最佳化資料中心有一個效能最佳化資料中心管理器。可基於可在各效能最佳化資料中心管理器之間交換的系統級探索資訊在多個效能最佳化資料中心管理器中分散工作負載。在替代實施例中，效能最佳化資料中心管理器可與動態主機設定協定伺服器進行通訊並與其他效能最佳化資料中心管理器協作來執行工作負載分散。舉例而言，在目標儲存裝置的停工時間（downtime）期間，可自遠端儲存裝置得到備份資料（backup data）。在此種情形中，效能最佳化資料中心管理器可使用系統級探索資訊將工作負載排程至備份儲存裝置。在基於光纖上高速非揮發性記憶體的儲存系統的情形中，任何效能最佳化資料中心管理器可連接至一或多個探索伺服器且使用獲取日誌命令

（get log command）來擷取可用儲存裝置的經更新探索資訊。一旦使用探索資訊建立了通訊路徑，目標儲存裝置便可服務於自主機電腦接收的輸入/輸出請求。由於本探索服務並不需要手動更新，因此本探索服務可針對可能包括諸多機架及效能最佳化資料中心且機架中包括諸多計算裝置及/或儲存裝置的超大規模資料中心而按比例放大。

【0035】 **圖 4** 示出根據一個實施例用於將探索日誌資訊填入的示例性流程圖。儲存裝置在啟動後運行動態主機設定協定用戶端，並向動態主機設定協定伺服器發送欲租用網際網路協定位址的請求（401）。動態主機設定協定伺服器因應於動態主機設定協定用戶端的請求而將動態網際網路協定位址或永久網際網路協定位址指派給儲存裝置的儲存裝置控制器（402）。在儲存裝置上運行的探索日誌更新器使用 RESTful 應用程式設計介面或 GraphQL 請求向探索伺服器發送探索日誌條目（403），所述探索日誌條目包括儲存裝置控制器的網際網路協定位址及埠位址。探索伺服器對探索日誌頁面進行填入（404），所述探索日誌頁面包括多個儲存裝置的一或多個探索日誌條目。探索伺服器將探索日誌頁面發送至主機電腦（405）。

【0036】 **圖 5** 示出根據一個實施例用於在主機電腦與目標儲存裝置之間建立通訊的示例性流程圖。主機電腦向探索伺服器發送探索資訊請求（501）。探索伺服器儲存有自多個儲存裝置的探索日誌更新器接收的一或多個探索日誌條目。主機電腦自探索伺服器

接收包括所述一或多個探索日誌條目的探索日誌頁面（502）。使用在探索日誌頁面中接收的所述一或多個探索日誌頁面條目，主機電腦與目標儲存裝置建立連接（503）並將工作負載排程至目標儲存裝置（504）。

【0037】 **圖 6** 示出根據一個實施例的示例性自動探索過程。儲存裝置 660 的動態主機設定協定用戶端向動態主機設定協定伺服器 630 發送欲獲得網際網路協定位址的請求（601），且動態主機設定協定伺服器 630 將網際網路協定位址指派並發送至儲存裝置 660 的動態主機設定協定用戶端（602）。儲存裝置 660 的探索日誌更新器（或動態主機設定協定伺服器 630）使用 RESTful 應用程式設計介面或 GraphQL 請求自動地向探索伺服器 650 發送儲存裝置 660 的探索日誌條目（603）。探索伺服器 650 自多個儲存裝置的每一探索日誌更新器收集一或多個探索日誌條目並準備包括所述一或多個探索日誌條目的探索日誌頁面。主機電腦 610 向探索伺服器 650 發送獲取日誌頁面命令（611），且作為回復，探索伺服器 650 將探索日誌頁面發送至主機電腦 610（612）。使用探索日誌頁面中所含有的資訊，主機電腦 610 建立通往儲存裝置 660 的連接（613）。

【0038】 根據一個實施例，一種儲存系統包括：多個儲存裝置，所述多個儲存裝置中的每一者被配置成運行動態主機設定協定（DHCP）用戶端；以及探索伺服器，被配置成向主機電腦提供探索日誌頁面。在所述多個儲存裝置中的每一者上運行的探索日誌

更新器將對應的所述儲存裝置的探索日誌條目提供至所述探索伺服器。所述探索日誌頁面含有所述多個儲存裝置的探索資訊，所述探索資訊包括一或多個探索日誌條目，且所述一或多個探索日誌條目中的每一者包括對應的所述儲存裝置的位址，所述位址被指派以在所述主機電腦與對應的所述儲存裝置之間建立通訊。所述主機電腦使用所述探索日誌頁面中所含有的目標儲存裝置的所述探索日誌條目來與所述目標儲存裝置建立連接，並向所述目標儲存裝置提供輸入/輸出（I/O）或計算工作負載。

【0039】 動態主機設定協定伺服器可為所述多個儲存裝置中的每一者指派所述位址，且所述多個儲存裝置的每一所述探索日誌更新器可使用被指派的所述位址來自動地將其包括被指派的所述位址的各自的探索日誌條目發送至所述探索伺服器。

【0040】 所述探索日誌條目可更包括儲存裝置控制器的埠位址。

【0041】 所述埠位址可根據光纖上高速非揮發性記憶體規範而指派。

【0042】 所述探索伺服器可使用 RESTful 應用程式設計介面（API）或 GraphQL 請求向所述主機電腦發送所述探索日誌頁面。

【0043】 所述多個儲存裝置中的至少一者可為光纖上高速非揮發性記憶體（NVMeoF）裝置，且所述主機電腦可向所述探索伺服器發送獲取日誌命令以接收所述探索日誌頁面。

【0044】 所述探索伺服器與所述動態主機設定協定伺服器可共置於所述儲存系統的同一伺服器中。

【0045】 所述多個儲存裝置可被附裝至機架，且所述機架中的所述探索伺服器可收集所述機架特有的機架級探索資訊。

【0046】 所述多個儲存裝置可分散於效能最佳化資料中心中的多個機架中，且負責效能最佳化資料中心的所述探索伺服器可收集所述效能最佳化資料中心特有的效能最佳化資料中心級探索資訊。

【0047】 所述效能最佳化資料中心的效能最佳化資料中心管理器可基於所述效能最佳化資料中心級探索資訊在所述多個儲存裝置中對工作負載進行排程。

【0048】 所述效能最佳化資料中心管理器可與所述儲存系統的另一效能最佳化資料中心管理器交換所述效能最佳化資料中心級探索資訊。

【0049】 根據另一實施例，一種方法包括：在儲存裝置中運行動態主機設定協定（DHCP）用戶端；向探索伺服器提供所述儲存裝置的探索日誌條目；將探索日誌頁面自所述探索伺服器提供至主機電腦，其中所述探索日誌頁面含有多個儲存裝置的探索資訊，所述探索資訊包括一或多個探索日誌條目，且所述一或多個探索日誌條目中的每一者包括對應的所述儲存裝置的位址，所述位址被指派以在所述主機電腦與對應的所述儲存裝置之間建立通訊；使用所述探索日誌頁面中所含有的目標儲存裝置的所述探索日誌條目在所述主機電腦與所述目標儲存裝置之間建立連接；以及將輸入/輸出（I/O）或計算工作負載自所述主機電腦提供至所述目標

儲存裝置。

【0050】 動態主機設定協定伺服器可為所述多個儲存裝置中的每一者指派所述位址，且所述多個儲存裝置的每一探索日誌更新器可使用被指派的所述位址來自動地將其包括被指派的所述位址的各自的探索日誌條目發送至所述探索伺服器。

【0051】 所述探索日誌條目可更包括儲存裝置控制器的埠位址。

【0052】 所述埠位址可根據光纖上高速非揮發性記憶體規範而指派。

【0053】 所述探索伺服器可使用 RESTful 應用程式設計介面（API）或 GraphQL 請求向所述主機電腦發送所述探索日誌頁面。

【0054】 所述多個儲存裝置中的至少一者可為光纖上高速非揮發性記憶體（NVMeoF）裝置，且所述主機電腦可向所述探索伺服器發送獲取日誌命令以接收所述探索日誌頁面。

【0055】 所述探索伺服器與所述動態主機設定協定伺服器可共置於儲存系統的同一伺服器中。

【0056】 所述多個儲存裝置可被附裝至機架，且所述機架中的所述探索伺服器可收集所述機架特有的機架級探索資訊。

【0057】 所述多個儲存裝置可分散於效能最佳化資料中心中的多個機架中，且負責效能最佳化資料中心的所述探索伺服器可收集所述效能最佳化資料中心特有的效能最佳化資料中心級探索資訊。

【0058】 所述效能最佳化資料中心的效能最佳化資料中心管理器

可基於所述效能最佳化資料中心級探索資訊在所述多個儲存裝置中對工作負載進行排程。

【0059】 所述效能最佳化資料中心管理器可與所述儲存系統的另一效能最佳化資料中心管理器交換所述效能最佳化資料中心級探索資訊。

【0060】 上文中已闡述了以上示例性實施例，以說明實作用於在超大規模資料中心環境中提供自動乙太儲存探索的系統及方法的各種實施例。此項技術中具有通常知識者將會聯想到相對於所揭露示例性實施例的各種潤飾及偏差形式。在以下申請專利範圍中陳述旨在處於本發明範圍內的標的物。

【符號說明】

【0061】

100：分散式儲存系統

110、610：主機電腦

120：光纖

121：探索日誌請求

122、200：探索日誌頁面

123：埠位址

124、152：探索日誌條目

126：動態主機設定協定請求/回應

130、630：動態主機設定協定伺服器

150、650：探索伺服器

151：探索控制器

160、660：儲存裝置

161：儲存裝置控制器

162：埠

163：動態主機設定協定用戶端

164：資料儲存器

165：探索日誌更新器

300：探索條目/日誌頁面條目/日誌條目

401、402、403、404、405、501、502、503、504、601、602、603、611、612、613：步驟

【發明申請專利範圍】

【第1項】 一種儲存系統，包括：

多個儲存裝置，所述多個儲存裝置中的至少一個儲存裝置被配置成運行動態主機設定協定（DHCP）用戶端；以及

探索伺服器，被配置成向主機電腦提供探索日誌頁面，

其中所述多個儲存裝置中的所述至少一個儲存裝置運行探索日誌更新器，所述探索日誌更新器基於所述 DHCP 用戶端的請求獲得的對應的儲存裝置的網際網路協定（IP）位址的更新，自動地將所述對應的儲存裝置的探索日誌條目提供至所述探索伺服器，

其中所述探索日誌頁面含有所述多個儲存裝置的探索資訊，所述探索資訊包括一或多個探索日誌條目，且

其中所述主機電腦參照所述探索日誌頁面中所含有的目標儲存裝置的所述探索日誌條目，使用所述目標儲存裝置的 IP 地址來與所述多個儲存裝置中的所述目標儲存裝置建立連接，並向所述目標儲存裝置提供輸入/輸出（I/O）或計算工作負載。

【第2項】 如申請專利範圍第1項所述的儲存系統，其中動態主機設定協定伺服器為所述多個儲存裝置中的所述至少一個儲存裝置指派所述IP位址，且所述多個儲存裝置中的所述至少一個儲存裝置的所述探索日誌更新器使用被指派的所述IP位址來自動地將其包括被指派的所述IP位址的各自的探索日誌條目發送至所述探索伺服器。

【第3項】 如申請專利範圍第2項所述的儲存系統，其中所述探索日誌條目更包括儲存裝置控制器的埠位址。

【第4項】 如申請專利範圍第3項所述的儲存系統，其中所述埠位址是根據光纖上高速非揮發性記憶體規範而指派。

【第5項】 如申請專利範圍第1項所述的儲存系統，其中所述探索伺服器使用RESTful應用程式設計介面（API）或GraphQL請求向所述主機電腦發送所述探索日誌頁面。

【第6項】 如申請專利範圍第1項所述的儲存系統，其中所述多個儲存裝置中的至少一者是光纖上高速非揮發性記憶體（NVMeoF）裝置，且所述主機電腦向所述探索伺服器發送獲取日誌命令以接收所述探索日誌頁面。

【第7項】 如申請專利範圍第2項所述的儲存系統，其中所述探索伺服器與所述動態主機設定協定伺服器共置於所述儲存系統的同一伺服器中。

【第8項】 如申請專利範圍第1項所述的儲存系統，其中所述多個儲存裝置被附裝至機架，且所述機架中的所述探索伺服器收集所述機架特有的機架級探索資訊。

【第9項】 如申請專利範圍第8項所述的儲存系統，其中所述多個儲存裝置分散於效能最佳化資料中心中的多個機架中，且負責所述效能最佳化資料中心的所述探索伺服器收集所述效能最佳化資料中心特有的效能最佳化資料中心級探索資訊。

【第10項】 如申請專利範圍第9項所述的儲存系統，其中所述效能最佳化資料中心的效能最佳化資料中心管理器基於所述效能最佳化資料中心級探索資訊在所述多個儲存裝置中對工作負載進行排程。

【第11項】 如申請專利範圍第10項所述的儲存系統，其中所述效能最佳化資料中心管理器與所述儲存系統的另一效能最佳化資料中心管理器交換所述效能最佳化資料中心級探索資訊。

【第12項】 一種提供自動乙太儲存探索的方法，包括：

在儲存裝置中運行動態主機設定協定（DHCP）用戶端；

基於所述 DHCP 用戶端的請求獲得的所述儲存裝置的網際網路協定（IP）位址的更新，自動地向探索伺服器提供所述儲存裝置的探索日誌條目；

將探索日誌頁面自所述探索伺服器提供至主機電腦，其中所述探索日誌頁面含有多個儲存裝置的探索資訊，所述探索資訊包括一或多個探索日誌條目；

參照所述探索日誌頁面中所含有的目標儲存裝置的所述探索日誌條目，使用所述目標儲存裝置的 IP 位址在所述主機電腦與所述目標儲存裝置之間建立連接；以及

將輸入/輸出（I/O）或計算工作負載自所述主機電腦提供至所述目標儲存裝置。

【第13項】 如申請專利範圍第12項所述的方法，其中動態主機設定協定伺服器為所述多個儲存裝置中的至少一個儲存裝置指派所

述IP位址，且所述多個儲存裝置中的所述至少一個儲存裝置的探索日誌更新器使用被指派的所述IP位址來自動地將其包括被指派的所述IP位址的各自的探索日誌條目發送至所述探索伺服器。

【第14項】 如申請專利範圍第13項所述的方法，其中所述探索日誌條目更包括儲存裝置控制器的埠位址。

【第15項】 如申請專利範圍第14項所述的方法，其中所述埠位址是根據光纖上高速非揮發性記憶體規範而指派。

【第16項】 如申請專利範圍第12項所述的方法，其中所述探索伺服器使用RESTful應用程式設計介面（API）或GraphQL請求向所述主機電腦發送所述探索日誌頁面。

【第17項】 如申請專利範圍第12項所述的方法，其中所述多個儲存裝置中的至少一者是光纖上高速非揮發性記憶體（NVMeoF）裝置，且所述主機電腦向所述探索伺服器發送獲取日誌命令以接收所述探索日誌頁面。

【第18項】 如申請專利範圍第13項所述的方法，其中所述探索伺服器與所述動態主機設定協定伺服器共置於儲存系統的同一伺服器中。

【第19項】 如申請專利範圍第12項所述的方法，其中所述多個儲存裝置被附裝至機架，且所述機架中的所述探索伺服器收集所述機架特有的機架級探索資訊。

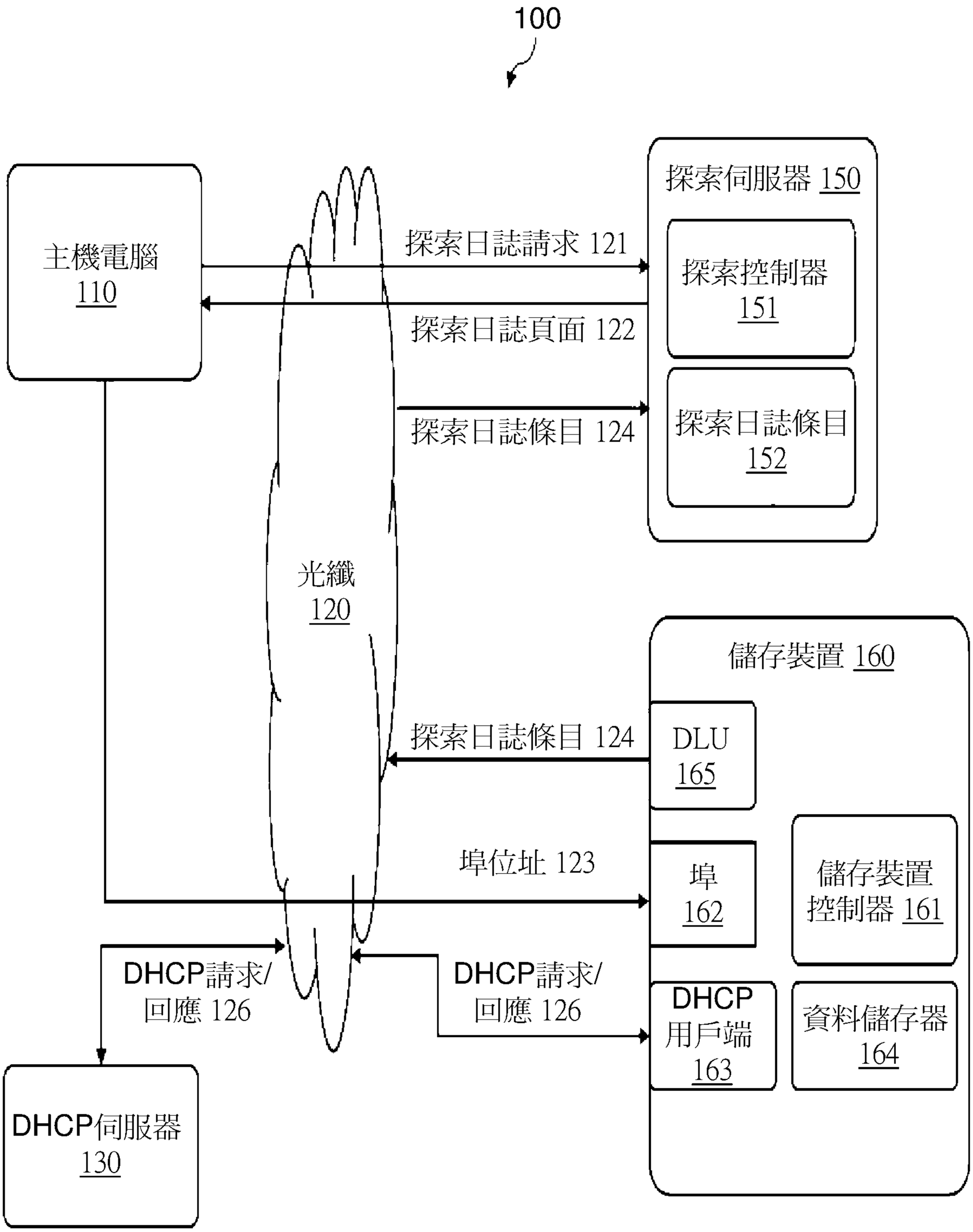
【第20項】 如申請專利範圍第19項所述的方法，其中所述多個儲存裝置分散於效能最佳化資料中心中的多個機架中，且負責所述

效能最佳化資料中心的所述探索伺服器收集所述效能最佳化資料中心特有的效能最佳化資料中心級探索資訊。

【第21項】 如申請專利範圍第20項所述的方法，其中所述效能最佳化資料中心的效能最佳化資料中心管理器基於所述效能最佳化資料中心級探索資訊在所述多個儲存裝置中對工作負載進行排程。

【第22項】 如申請專利範圍第21項所述的方法，其中所述效能最佳化資料中心管理器與儲存系統的另一效能最佳化資料中心管理器交換所述效能最佳化資料中心級探索資訊。

【發明圖式】


100
主機電腦
110
探索日誌請求 121
探索日誌頁面 122
探索日誌條目 124
探索伺服器 150
探索控制器
151
探索日誌條目
152
光纖
120
儲存裝置 160
探索日誌條目 124
DLU
165
埠位址 123
埠
162
儲存裝置
控制器 161
DHCP請求/
回應 126
DHCP請求/
回應 126
DHCP
用戶端
163
資料儲存器
164
DHCP伺服器
130

【圖1】

200

位元組	說明
07:00	產生計數器
15:08	記錄數目
17:16	記錄格式
1023:18	被保留
2047:1024	探索日誌條目0
3017:2048	探索日誌條目1
...	...
(((N+2) x 1024 -1):((N+1) x 1024)	被保留

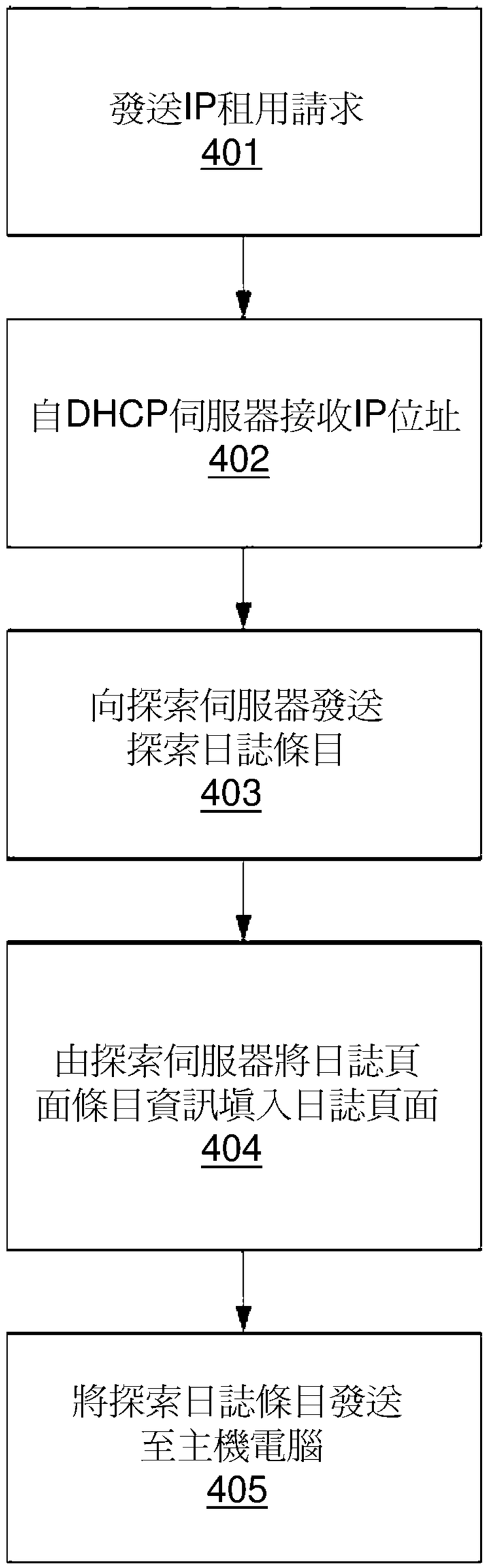
300

【圖2】

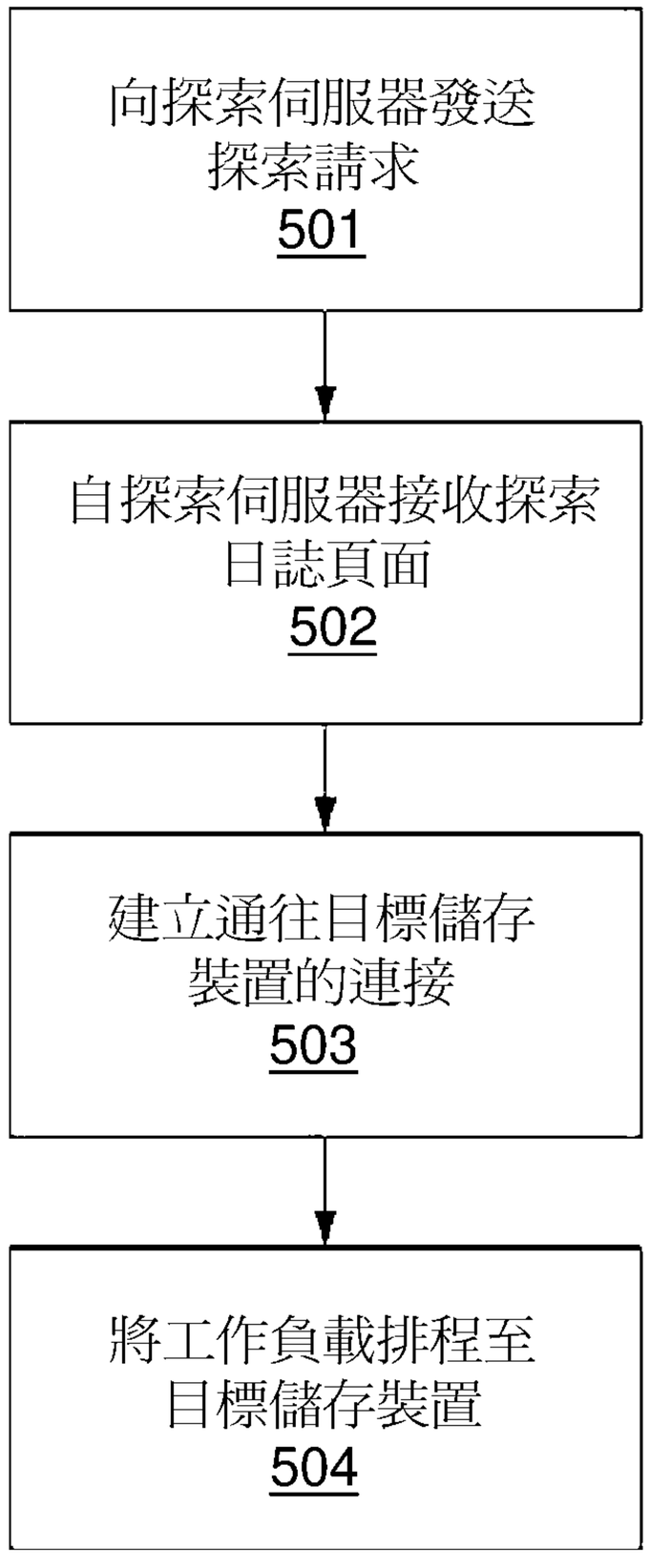
300

位元組	說明
00	傳送類型
01	位址家族
02	子系統類型
03	傳送要求
05:04	埠ID
07:06	控制器ID
09:08	管理最大SQ大小
31:10	被保留
63:32	傳送服務識別符
255:64	被保留
511:256	NVM子系統定質名字（NQN）
767:512	傳送位址
1023:768	傳送特有的位址子類型

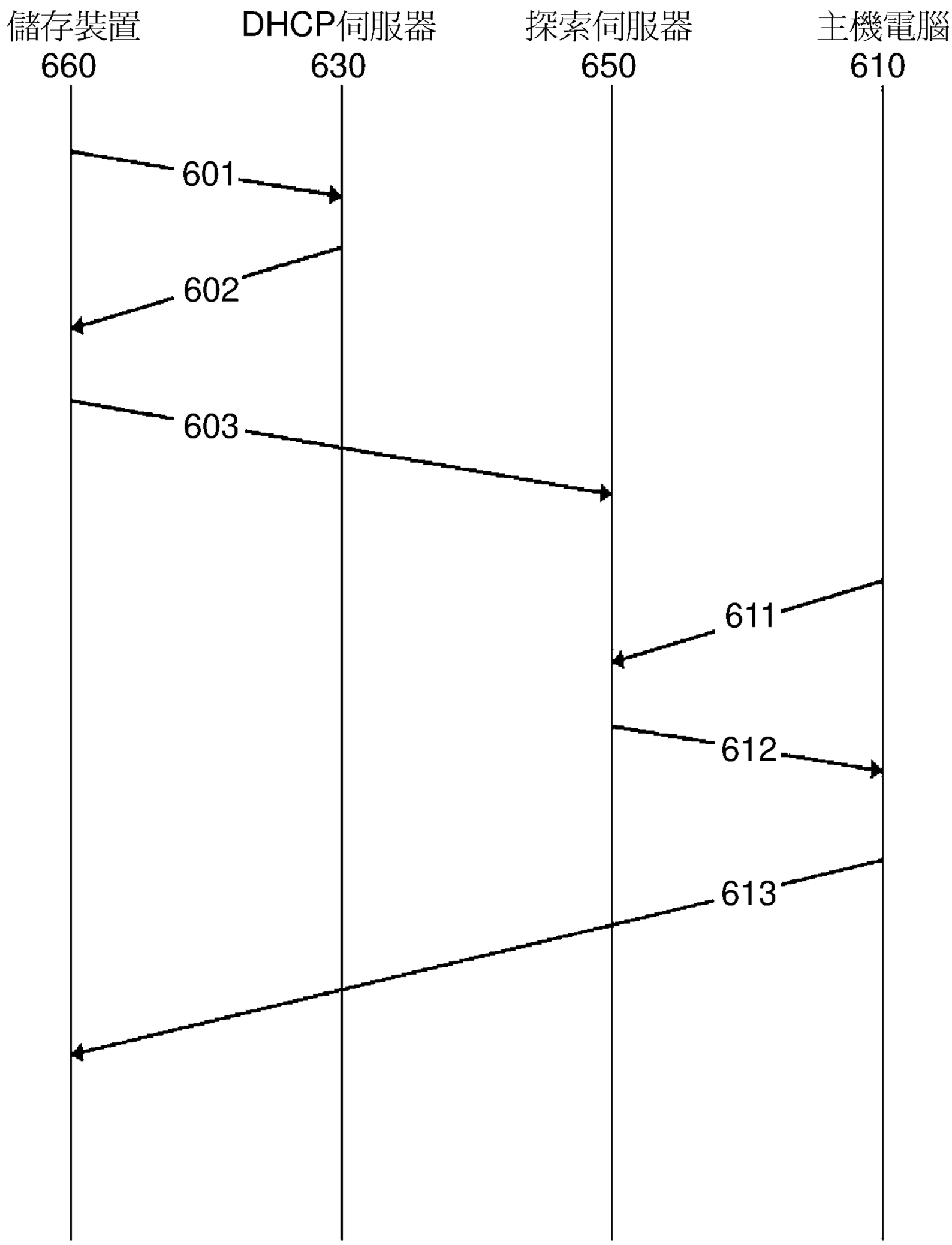
【圖3】


發送IP租用請求
401
自DHCP伺服器接收IP位址
402
向探索伺服器發送
探索日誌條目
403
由探索伺服器將日誌頁
面條目資訊填入日誌頁面
404
將探索日誌條目發送
至主機電腦
405

【圖4】


向探索伺服器發送
探索請求
501
自探索伺服器接收探索
日誌頁面
502
建立通往目標儲存
裝置的連接
503
將工作負載排程至
目標儲存裝置
504

【圖5】


儲存裝置
660
DHCP伺服器
630
探索伺服器
650
主機電腦
610
601
602
603
611
612
613

【圖6】