202028986

(19)中華民國智慧財產局

(12)發明說明書公開本　(11)公開編號：TW 202028986 A

(43)公開日： 中華民國 109 (2020) 年 08 月 01 日

(21)申請案號：109100029　(22)申請日： 中華民國 109 (2020) 年 01 月 02 日

(51)Int. Cl.：　*G06F12/02　(2006.01)*　*G06F9/312　(2006.01)*

(30)優先權：2019/09/25　美國　16/581,769

2019/01/18　美國　62/794,531

(71)申請人：香港商希瑞科技股份有限公司 (香港地區) SILICON MOTION TECHNOLOGY (HONG KONG) LIMITED　(HK)
香港

(72)發明人：曾國輔 TSENG, GUO-FU (TW) ；張正岳 CHANG, CHENG-YUE (TW) ；邱冠凱 CHIU, KUAN-KAI (TW)

(74)代理人：吳豐任；戴俊彥

申請實體審查：有　申請專利範圍項數：22 項　圖式數：13　共 55 頁

(54)名稱

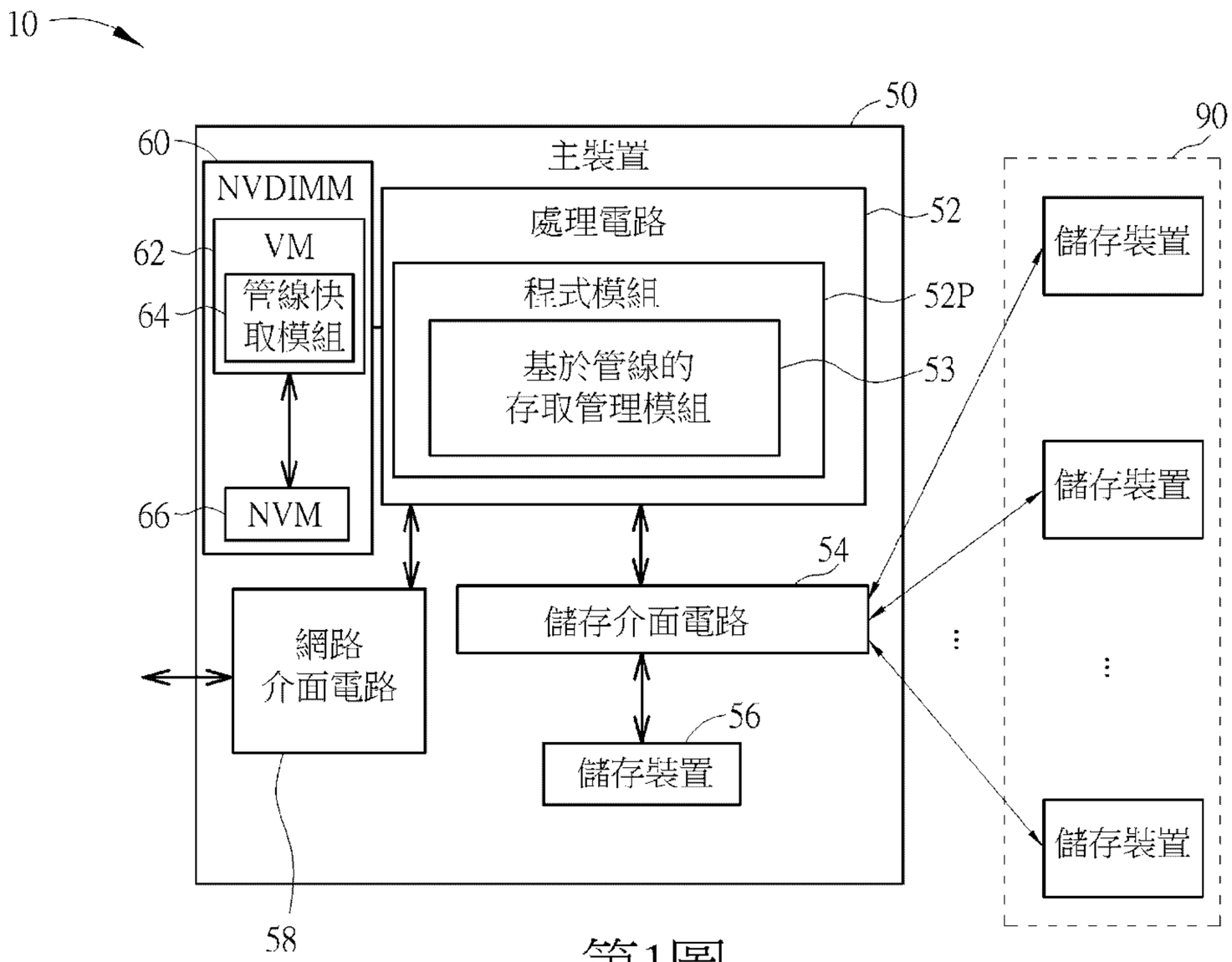
用來在儲存伺服器中進行基於管線的存取管理的方法及設備

(57)摘要

本發明提供一種用來在一儲存伺服器中進行基於管線的存取管理的方法以及設備。該方法包含：因應將使用者資料寫入該儲存伺服器的請求，利用該儲存伺服器中之主裝置將該使用者資料寫入該儲存伺服器之儲存裝置層並且以該儲存伺服器的管線架構開始處理對應於寫入該使用者資料的該請求的物件寫入指令；利用該主裝置將對應於該使用者資料的元資料輸入該管線架構中之管線；以及利用該主裝置以該管線的第一快取模組快取該元資料，以控制該儲存伺服器不產生該元資料的寫入放大即完成該請求，其中該第一快取模組係在該儲存裝置層外部的硬體管線模組。

A method for performing pipeline-based accessing management in a storage server and associated apparatus are provided. The method includes: in response to a request of writing user data into the storage server, utilizing a host device within the storage server to write the user data into a storage device layer of the storage server and start processing an object write command corresponding to the request of writing the user data with a pipeline architecture of the storage server; utilizing the host device to input metadata corresponding to the user data into at least one pipeline within the pipeline architecture; and utilizing the host device to cache the metadata with a first cache module of the pipeline, for controlling the storage server completing the request without generating write amplification of the metadata, wherein the first cache module is a hardware pipeline module outside the storage device layer.

指定代表圖：



第1圖

符號簡單說明：

10:儲存伺服器

50:主裝置

52:處理電路

52P:程式模組

53:基於管線的存取管理模組

54:儲存介面電路

56:儲存裝置

58:網路介面電路

60:非揮發性雙列直插式記憶體模組

62:揮發性記憶體

64:管線快取模組

66:非揮發性記憶體

90:儲存裝置

202028986

【發明摘要】

【中文發明名稱】用來在儲存伺服器中進行基於管線的存取管理的方法及設備

【英文發明名稱】METHOD AND APPARATUS FOR PERFORMING PIPELINE-BASED ACCESSING MANAGEMENT IN A STORAGE SERVER

【中文】

本發明提供一種用來在一儲存伺服器中進行基於管線的存取管理的方法以及設備。該方法包含：因應將使用者資料寫入該儲存伺服器的請求，利用該儲存伺服器中之主裝置將該使用者資料寫入該儲存伺服器之儲存裝置層並且以該儲存伺服器的管線架構開始處理對應於寫入該使用者資料的該請求的物件寫入指令；利用該主裝置將對應於該使用者資料的元資料輸入該管線架構中之管線；以及利用該主裝置以該管線的第一快取模組快取該元資料，以控制該儲存伺服器不產生該元資料的寫入放大即完成該請求，其中該第一快取模組係在該儲存裝置層外部的硬體管線模組。

【英文】

A method for performing pipeline-based accessing management in a storage server and associated apparatus are provided. The method includes: in response to a request of writing user data into the storage server, utilizing a host device within the storage server to write the user data into a storage device layer of the storage server and start processing an object write command corresponding to the request of writing the user data with a pipeline architecture of the storage server; utilizing the host device to input metadata corresponding to the user data into at least one pipeline within the pipeline architecture; and utilizing the host device to cache the metadata with a first cache module of the pipeline, for controlling the storage server

completing the request without generating write amplification of the metadata, wherein the first cache module is a hardware pipeline module outside the storage device layer.

【指定代表圖】第（1）圖。

【代表圖之符號簡單說明】

10・・・儲存伺服器

50・・・主裝置

52・・・處理電路

52P・・・程式模組

53・・・基於管線的存取管理模組

54・・・儲存介面電路

56・・・儲存裝置

58・・・網路介面電路

60・・・非揮發性雙列直插式記憶體模組

62・・・揮發性記憶體

64・・・管線快取模組

66・・・非揮發性記憶體

90・・・儲存裝置

【特徵化學式】

無

【發明說明書】

【中文發明名稱】用來在儲存伺服器中進行基於管線的存取管理的方法及設備

【英文發明名稱】METHOD AND APPARATUS FOR PERFORMING PIPELINE-BASED ACCESSING MANAGEMENT IN A STORAGE SERVER

【技術領域】

【0001】　本發明係關於資料儲存，尤指一種用來在一儲存伺服器中進行基於管線的（pipeline-based）存取管理的方法及設備，其中該設備的例子可包含（但不限於）：該儲存伺服器的整體、該儲存伺服器中之一主裝置、該主裝置中之一處理電路、以及在該處理電路中運行著對應於該方法的一或多個程式模組的至少一處理器／處理器核心（processor core）諸如中央處理單元（Central Processing Unit, CPU）／中央處理單元核心（CPU core）。

【先前技術】

【0002】　近年來由於記憶體的技術不斷地發展，各種可攜式以及非可攜式記憶裝置（例如：符合SD/MMC、CF、MS、XD或UFS標準的記憶卡；又例如：固態硬碟（solid state drive, SSD）；又例如：符合UFS或EMMC規格的嵌入式（embedded）儲存裝置）被廣泛地實施於諸多應用中。因此，這些記憶裝置中之記憶體的存取控制遂成為相當熱門的議題。

【0003】　以常用的NAND型快閃記憶體而言，其主要可包含單階細胞（single level cell, SLC）與多階細胞（multiple level cell, MLC）兩大類之快閃記憶體。單階細胞快閃記憶體中之每個被當作記憶細胞（memory cell）的電晶體只有兩種電荷值，分別用來表示邏輯值0與邏輯值1。另外，多階細胞快閃記憶體中之每個被當作記憶細胞的電晶體的儲存能力則被充分利用，其採用較高的電壓來驅動，以透過不同的電壓位準在一個電晶體中記錄至少兩位元的資訊（諸如00、01、11、10）。理論上，多階細胞快閃記憶體的記錄密度可以達到單階細胞快閃

記憶體的記錄密度之至少兩倍，這對於曾經在發展過程中遇到瓶頸的NAND型快閃記憶體之相關產業而言，是非常好的消息。

【0004】　相較於單階細胞快閃記憶體，由於多階細胞快閃記憶體之價格較便宜，並且在有限的空間裡可提供較大的容量，故多階細胞快閃記憶體很快地成為市面上之記憶裝置競相採用的主流。然而，多階細胞快閃記憶體的不穩定性所導致的問題也一一浮現。為了確保在記憶裝置中對快閃記憶體之存取控制能符合相關規範，快閃記憶體的控制器通常備有某些管理機制以妥善地管理資料之存取。

【0005】　然而，具備上列管理機制的記憶裝置仍有不足之處。尤其，當記憶裝置被設置在具有Ceph控制架構的儲存系統上時，會發生某些問題。例如，記憶裝置會因為Ceph相關（Ceph-related）控制而造成嚴重的寫入放大。寫入放大通常來自於儲存系統中的日誌（journal）與元資料（metadata）運作，這會造成在儲存系統中額外的記憶體複製與資料流量（data traffic），因此會增加輸入／輸出的延遲並且降低儲存系統的效能。由於區塊裝置可具有一特定區塊大小（例如4096位元組（byte）或4千位元組（kilobyte；KB）），對於小於該區塊大小的任一更新（例如元資料的更新），它會被放大為該區塊大小。儲存系統的效能會因為來自小量更新（諸如日誌、元資料等的更新）以及固定的區塊大小的寫入放大而大幅地降低。另外，記憶裝置會因為對應於該Ceph相關控制之額外的寫入放大而具有較短的預期壽命。因此，需要一種新穎的方法以及相關架構，已在沒有副作用或較不會帶來副作用的情況下改善整體效能。

【發明內容】

【0006】　因此，本發明之一目的在於提供一種用來在一儲存伺服器中進行基於管線的（pipeline-based）存取管理的方法，並提供相關設備諸如該儲存伺服器、

該儲存伺服器中之一主裝置等，以解決上述問題。

【0007】　本發明之另一目的在於提供一種用來在一儲存伺服器中進行基於管線的存取管理的方法，並提供相關設備諸如該儲存伺服器、該儲存伺服器中之一主裝置等，以在沒有副作用或較不會帶來副作用的情況下達到最佳化的效能。

【0008】　本發明至少一實施例提供一種用來在一儲存伺服器中進行基於管線的存取管理的方法，其中該方法係應用於該儲存伺服器。該方法可包含：因應將使用者資料寫入該儲存伺服器的一請求，利用該儲存伺服器中之一主裝置將該使用者資料寫入該儲存伺服器之一儲存裝置層並且以該儲存伺服器的一管線（pipeline）架構開始處理對應於寫入該使用者資料的該請求的一物件寫入指令，其中該儲存伺服器包含該主裝置以及該儲存裝置層，該儲存裝置層包含耦接至該主裝置的至少一儲存裝置，該主裝置係用來控制該儲存伺服器的運作，以及所述至少一儲存裝置係用來為該儲存伺服器儲存資訊；在以該管線架構處理該物件寫入指令的期間，利用該主裝置將對應於該使用者資料的元資料（metadata）輸入該管線架構中之至少一管線，其中該管線架構穿過運行於該主裝置之一處理電路上的複數個程式模組中之一或多個中介層（intermediate layer）的程式模組，並且達到該儲存裝置層；以及在以該管線架構處理該物件寫入指令的期間，利用該主裝置以所述至少一管線的一第一快取（cache）模組快取該元資料，以控制該儲存伺服器在不產生該元資料的針對該一或多個中介層的程式模組之寫入放大的情況下完成該請求，其中該第一快取模組係在該儲存裝置層外部的一硬體管線模組。

【0009】　除了上述方法以外，本發明亦提供一種主裝置。該主裝置可包含一處理電路，用來控制該主裝置在一儲存伺服器中進行基於管線的存取管理，其中該儲存伺服器包含該主裝置以及一儲存裝置層，該儲存裝置層包含耦接至該主裝置的至少一儲存裝置，該主裝置係用來控制該儲存伺服器的運作，以及所

述至少一儲存裝置係用來為該儲存伺服器儲存資訊。例如，因應將使用者資料寫入該儲存伺服器的一請求，該儲存伺服器中之該主裝置將該使用者資料寫入該儲存伺服器之該儲存裝置層並且以該儲存伺服器的一管線架構開始處理對應於寫入該使用者資料的該請求的一物件寫入指令；在以該管線架構處理該物件寫入指令的期間，該主裝置將對應於該使用者資料的元資料輸入該管線架構中之至少一管線，其中該管線架構穿過運行於該主裝置之該處理電路上的複數個程式模組中之一或多個中介層的程式模組，並且達到該儲存裝置層；以及在以該管線架構處理該物件寫入指令的期間，該主裝置以所述至少一管線的一第一快取模組快取該元資料，以控制該儲存伺服器在不產生該元資料的針對該一或多個中介層的程式模組之寫入放大的情況下完成該請求，其中該第一快取模組係在該儲存裝置層外部的一硬體管線模組。

【0010】　除了上述方法以外，本發明亦提供一儲存伺服器。該儲存伺服器可包含一主裝置以及一儲存裝置層，其中該主裝置係用來控制該儲存伺服器的運作。例如，該主裝置可包含一處理電路，用來控制該主裝置在該儲存伺服器中進行基於管線的存取管理。另外，該儲存裝置層可包含耦接至該主裝置的至少一儲存裝置，並且所述至少一儲存裝置係用來為該儲存伺服器儲存資訊。例如，因應將使用者資料寫入該儲存伺服器的一請求，該儲存伺服器中之該主裝置將該使用者資料寫入該儲存伺服器之該儲存裝置層並且以該儲存伺服器的一管線架構開始處理對應於寫入該使用者資料的該請求的一物件寫入指令；在以該管線架構處理該物件寫入指令的期間，該主裝置將對應於該使用者資料的元資料輸入該管線架構中之至少一管線，其中該管線架構穿過運行於該主裝置之該處理電路上的複數個程式模組中之一或多個中介層的程式模組，並且達到該儲存裝置層；以及在以該管線架構處理該物件寫入指令的期間，該主裝置以所述至少一管線的一第一快取模組快取該元資料，以控制該儲存伺服器在不產生該元

資料的針對該一或多個中介層的程式模組之寫入放大的情況下完成該請求，其中該第一快取模組係在該儲存裝置層外部的一硬體管線模組。

【0011】　本發明的方法及相關設備能確保整個儲存伺服器（例如該主裝置以及儲存裝置）可妥善地運作以避免相關技術中的問題，諸如因為Ceph相關控制的儲存裝置的寫入放大問題、成本增加的問題等。另外，本發明的實施例不會大幅地增加額外的成本，尤其，儲存裝置中的控制器（諸如記憶體控制器）的製造商不需要實現控制器的新硬體架構，而設計與製作對應於這個新硬體架構的新積體電路（integrated circuit, IC）所需的成本能得以節省。因此，相關技術的問題能被解決，而整體成本不會增加太多。相較於相關技術，本發明的方法與相關設備能在沒有副作用或較不會帶來副作用的情況下改善整體效能。

【圖式簡單說明】

【0012】

第1圖為依據本發明一實施例之一儲存伺服器的示意圖。

第2圖依據本發明一實施例繪示第1圖所示之儲存伺服器中之一儲存裝置的某些實施細節。

第3圖為依據本發明一實施例之用來在一儲存伺服器中進行基於管線的存取管理的方法的基於管線的快取控制方案。

第4圖依據本發明一實施例繪示對應於第3圖所示之方法的效能增強。

第5圖繪示用於每一寫入運作（例如一物件寫入運作）的物件儲存裝置（Object Storage Device，以下簡稱OSD）元資料，諸如將被儲存的相關資訊。

第6圖繪示一原始OSD日誌控制方案的一系列運作。

第7圖依據本發明一實施例繪示在第3圖所示之方法的一專門OSD日誌控制方案中的複數個每執行緒日誌環（per thread journal ring）以及在該專門OSD日誌

控制方案中的相關運作的一部分。

第8圖繪示在第7圖所示之專門OSD日誌控制方案中的相關運作的另一部分。

第9圖繪示在第7圖所示之專門OSD日誌控制方案中的相關運作的另一部分。

第10圖繪示在第7圖所示之專門OSD日誌控制方案中的相關運作的另一部分。

第11圖為依據本發明一實施例之第3圖所示之方法的工作流程。

第12圖依據本發明一實施例繪示第1圖所示之管線快取模組的某些子模組。

第13圖依據本發明一實施例繪示第1圖所示之主裝置的某些非揮發性雙列直插式記憶體模組。

【實施方式】

【0013】　本發明的實施例提供了一種用來在一儲存伺服器中進行基於管線的（pipeline-based）存取管理的方法以及設備，而該設備可包含該儲存伺服器的至少一部分（例如一部分或全部）。例如，該設備可包含該儲存伺服器的一部分，諸如該儲存伺服器中之一主裝置或位於該主裝置中的相關控制電路（例如運行著對應於該方法的一或多個程式模組的處理電路、以及相關電路）。又例如，該設備可包含該儲存伺服器的整體。

【0014】　第1圖為依據本發明一實施例之儲存伺服器10的示意圖。儲存伺服器10可包含一主裝置50，並且可包含至少一儲存裝置（例如一或多個儲存裝置）諸如複數個儲存裝置90，其中複數個儲存裝置90耦接至主裝置50。依據本實施例，主裝置50可用來控制儲存伺服器10的運作，而複數個儲存裝置90可用來為儲存伺服器10儲存資訊。如第1圖所示，主裝置50可包含一處理電路52（例如至少一處理器／處理器核心以及相關電路諸如隨機存取記憶體（Random Access Memory, RAM）、匯流排等）以控制主裝置50的運作，且可包含至少一儲存介面電路54以將複數個儲存裝置90以及多個儲存或記憶裝置（例如一或多個硬式磁

碟機（Hard Disk Drive）及／或一或多個固態硬碟（Solid State Drive, SSD））耦接於主裝置50，又可包含一網路介面電路58已將主裝置50耦接至至少一網路。該多個儲存或記憶裝置可包含至少一儲存裝置諸如一或多個儲存裝置，其可統稱為儲存裝置56。例如，儲存裝置56可包含一組儲存裝置，其中的一儲存裝置可用來當作主裝置50的系統磁碟，而其它的則可用來為主裝置50儲存使用者資料，但本發明不限於此。又例如，該儲存裝置56可包含一個儲存裝置，而這個儲存裝置可用來當作主裝置50的系統磁碟。另外，主裝置50可另包含至少一非揮發性雙列直插式記憶體模組（non-volatile dual in-line memory module, NVDIMM）（例如一或多個非揮發性雙列直插式記憶體模組）諸如非揮發性雙列直插式記憶體模組60（在第1圖中標示為「NVDIMM」以求簡明）。例如，非揮發性雙列直插式記憶體模組60可包含一揮發性（volatile）記憶體62以及一非揮發性（non-volatile, NV）記憶體66（在第1圖中分別標示為「VM」以及「NVM」以求簡明），其可分別用來為處理電路52儲存（例如緩衝或快取（cache））資訊以及為處理電路52保存在揮發性記憶體62中的該資訊，並且可耦接至處理電路52以供處理電路52存取。處理電路52可存取揮發性記憶體62中之管線快取模組64，諸如揮發性記憶體62中之一儲存區。

【0015】　依據本實施例，運行著程式模組52P（尤指對應於該方法的基於管線的存取管理模組53）的處理電路52可用來依據該方法控制主裝置50的運作，例如控制主裝置50在儲存伺服器10中進行基於管線的存取管理。儲存介面電路54可符合一或多個標準（例如串行高技術附件（Serial Advanced Technology Attachment，簡稱「SATA」）標準、外設組件互聯（Peripheral Component Interconnect，簡稱「PCI」）標準、快捷外設互聯（Peripheral Component Interconnect Express，簡稱「PCIe」）標準、快捷非揮發性記憶體（Non-Volatile Memory Express，簡稱「NVMe」）標準、快捷非揮發性記憶體外接存取

（NVMe-over-Fabrics，簡稱「NVMeoF」）標準、小型電腦系統介面（Small Computer System Interface，簡稱「SCSI」）標準、通用型快閃記憶體儲存（Universal Flash Storage，簡稱「UFS」）標準等之中的一或多者），並且可依據該一或多個標準進行通訊以容許運行著程式模組52P的處理電路52透過儲存介面電路54存取儲存裝置56以及複數個儲存裝置90。另外，網路介面電路58可用來提供有線或無線網路連接，而對應於一或多個使用者的一或多個用戶端裝置可透過該有線或無線網路連接來存取在儲存伺服器10（例如其內的儲存裝置56及複數個儲存裝置90）中的使用者資料。為便於理解，第1圖左手邊所示之架構中的主裝置50以及相關電路／模組／裝置（例如運行著程式模組52P的處理電路52、儲存介面電路54以及非揮發性雙列直插式記憶體模組60）可分別作為上述主裝置及其相關電路／模組／裝置（例如運行著對應於該方法的該一或多個程式模組的該處理電路、及相關電路）的例子。

【0016】　在第1圖所示之架構中，儲存伺服器10可被繪示為包含有主裝置50以及耦接至主裝置50的複數個儲存裝置90，但本發明不限於此。例如，主裝置50可另包含一機殼（case）（例如一電腦機殼，其可由金屬及／或一或多個其它材料製作而成）以供設置主裝置50的元件諸如第1圖所示之元件（例如處理電路52、儲存介面電路54、網路介面電路58、非揮發性雙列直插式記憶體模組60等）以及複數個儲存裝置90的至少一部分（例如一部分或全部）。又例如，儲存伺服器10可另包含耦接於主裝置50與複數個儲存裝置90中之至少一部分（例如一部分或全部）之間的至少一開關電路（例如一或多個開關電路），以進行主裝置50與複數個儲存裝置90中之上述至少一部分之間的訊號切換。

【0017】　依據某些實施例，運行著程式模組52P的處理電路52、或儲存介面電路54可組態（configure）複數個儲存裝置90的至少一部分（一部分或全部）以形成一儲存池（storage pool）架構，但本發明不限於此。依據某些實施例，運行著

程式模組52P的處理電路52、或儲存介面電路54可組態複數個儲存裝置90的至少一部分（一部分或全部）以形成儲存伺服器10的一容錯式磁碟陣列（Redundant Array of Independent Disks, RAID），諸如一全快閃記憶體陣列（All Flash Array, AFA）。

【0018】　第2圖依據本發明一實施例繪示第1圖所示之儲存伺服器10中之一儲存裝置的某些實施細節，其中儲存裝置100可作為複數個儲存裝置90中之任一者的例子。尤其，複數個儲存裝置90的每一者均可依據儲存裝置100的架構來實施，但本發明不限於此。儲存裝置100可用來提供儲存空間給主裝置50。在主裝置50的控制下，該一或多個用戶端裝置可存取（例如讀取或寫入）在該儲存空間的使用者資料。主裝置50的例子可包含（但不限於）：個人電腦諸如桌上型電腦及膝上型電腦。儲存裝置100的例子可包含（但不限於）：固態硬碟、以及各種類型的嵌入式記憶裝置諸如符合UFS或EMMC等標準的嵌入式記憶裝置。依據本實施例，儲存裝置100可包含一控制器諸如記憶體控制器110，並且可另包含一非揮發性記憶體120，其中該控制器係用來控制儲存裝置100的運作以及存取非揮發性記憶體120，以及非揮發性記憶體120係用來儲存資訊。非揮發性記憶體120可包含至少一非揮發性記憶體元件（例如一或多個非揮發性記憶體元件），諸如複數個非揮發性記憶體元件122-1、122-2、...、及122-N，其中「N」可代表大於1的正整數。例如，非揮發性記憶體120可為一快閃記憶體，而複數個非揮發性記憶體元件122-1、122-2、...、及122-N可為複數個快閃記憶體晶片或複數個快閃記憶體裸晶，但本發明不限於此。

【0019】　依據本實施例，記憶體控制器110可用來控制非揮發性記憶體120的存取，以容許主裝置50透過記憶體控制器110存取非揮發性記憶體120，以在儲存伺服器10中進行存取管理。如第2圖所示，記憶體控制器110可包含一處理電路諸如微處理器112、一儲存單元諸如唯讀記憶體（read-only memory, ROM）

112M、一控制邏輯電路114、一隨機存取記憶體116以及一傳輸介面電路118，其中以上元件可透過匯流排互相耦接。隨機存取記憶體116可由一靜態隨機存取記憶體（Static RAM, SRAM）來實施，但本發明不限於此。隨機存取記憶體116可用來提供內部儲存空間給記憶體控制器110，例如，隨機存取記憶體116可用來當作一緩衝記憶體以供緩衝資料。另外，本實施例的唯讀記憶體112M係用來儲存一程式碼112C，而微處理器112係用來執行程式碼112C以控制記憶體控制器110的運作來控制非揮發性記憶體120的存取，以容許主裝置50透過記憶體控制器110存取非揮發性記憶體120。請注意，在某些例子中，程式碼112C可被儲存在隨機存取記憶體116或任何類型的記憶體中。另外，控制邏輯電路114可用來控制非揮發性記憶體120，並且可包含一資料保護電路（未顯示）以保護資料及／或進行錯誤更正，但本發明不限於此。傳輸介面電路118可符合一特定通訊標準（例如SATA標準、PCI標準、PCIe標準、NVMe標準、NVMeoF標準、SCSI標準、UFS標準等），並且可依據該特定通訊標準進行通訊，例如，為儲存裝置100，和主裝置50進行通訊，其中儲存介面電路54可符合該特定通訊標準，以供為主裝置50，和儲存裝置100進行通訊。

【0020】　依據某些實施例，主裝置50可傳送主裝置指令以及對應的邏輯位址至記憶體控制器110以存取儲存裝置100。記憶體控制器110接收主裝置指令及邏輯位址，並且將主裝置指令轉譯為記憶體操作指令（其可簡稱為操作指令），再以這些操作指令控制非揮發性記憶體120以讀取、寫入／編程非揮發性記憶體120中之具有實體位址的記憶單元（例如資料頁），其中實體位址可對應於邏輯位址。當記憶體控制器110對複數個非揮發性記憶體元件122-1、122-2、...、及122-N中之任一非揮發性記憶體元件122-n（「n」可代表在區間[1, N]內的任一整數）進行一抹除運作時，非揮發性記憶體元件122-n的多個區塊的至少一區塊可被抹除，其中該多個區塊的每一區塊可包含多個頁面（例如資料頁面），並且可

對一或多個頁面進行一存取運作（例如讀取或寫入）。

【0021】　依據某些實施例，儲存伺服器10（例如主裝置50）可依據一Ceph解決方案來運作，以透過軟體定義的方式使儲存伺服器10成為一分散式儲存系統（distributed system）的一部分。複數個儲存系統（例如多個儲存伺服器{10}諸如儲存伺服器10）較佳可用來形成該分散式儲存系統的一資源池（resource pool），以支援各種類型的存取，諸如一區塊裝置（block device）類型（例如用來存取該分散式儲存系統中之一模擬區塊裝置）、一檔案系統類型（例如用來存取該分散式儲存系統中之一檔案系統）、以及一物件類型（例如用來存取於該分散式儲存系統中之物件命名空間（object namespace）內的一物件），例如，可附帶某些特徵諸如容錯（fault tolerance）、自動故障轉移（failover）控制等，但本發明不限於此。例如，一檔案可被儲存為在該分散式儲存系統中之該物件命名空間內的一物件。又例如，對應於一特定位址的一資料區塊可被儲存為在該分散式儲存系統中之該物件命名空間內的一物件。為了實現該分散式儲存系統的特徵，除了資料（例如使用者資料）以外，儲存伺服器10可用來儲存額外資訊諸如元資料（metadata）及日誌（journal）。

【0022】　第3圖為依據本發明一實施例之用來在該儲存伺服器中進行基於管線的存取管理的方法的基於管線的快取控制方案。該方法能被應用於儲存伺服器10，尤其，能被應用於包含有運行著程式模組52P（例如對應於該方法的基於管線的存取管理模組53）的處理電路52的主裝置50，連同複數個儲存裝置90的每一者，諸如包含有記憶體控制器110的儲存裝置100，其中該控制電路諸如運行著基於管線的存取管理模組53的處理電路52可控制主裝置50依據該方法來運作。例如，儲存伺服器10可依據Ceph解決方案來運作，並且本發明的方法及相關設備（例如主裝置50，尤其，運行著基於管線的存取管理模組53的處理電路52）能以一種新穎的管線架構來進行存取控制，以在沒有副作用或較不會帶來

副作用的情況下達到最佳效能。該管線架構可利用上述至少一非揮發性雙列直插式記憶體模組（例如一或多個非揮發性雙列直插式記憶體模組）諸如非揮發性雙列直插式記憶體模組60作為儲存日誌（storage journal）的媒介（medium），並且可利用上述至少一非揮發性雙列直插式記憶體模組諸如非揮發性雙列直插式記憶體模組60作為儲存元資料（storage metadata）的媒介。另外，非揮發性雙列直插式記憶體模組60可包含一控制器（例如與儲存裝置100的記憶體控制器110類似的記憶體控制器）以控制非揮發性雙列直插式記憶體模組60的運作，並且非揮發性記憶體66可具有與非揮發性記憶體120類似的架構並且可包含對應於一非揮發性記憶體元件數量N'的一或多個非揮發性記憶體元件，其中非揮發性記憶體元件數量N'可代表一正整數，但本發明不限於此。例如，揮發性記憶體62可為一動態隨機存取記憶體（Dynamic RAM, DRAM），而非揮發性記憶體66可為包含有N'個快閃記憶體晶片或N'個快閃記憶體裸晶的一快閃記憶體。另外，非揮發性雙列直插式記憶體模組60可在儲存伺服器10的一常態模式中透過匯流排自儲存伺服器10的電源供應器取得常態電源（例如一或多個驅動電壓），並且可在儲存伺服器10的一故障轉移模式中自一備用電源（例如儲存伺服器10中之一電池、或非揮發性雙列直插式記憶體模組60中之一超級電容器（supercapacitor））取得緊急電源。非揮發性雙列直插式記憶體模組60的該控制器可利用非揮發性記憶體66依據來自處理電路52的一清理（flush）指令保存於揮發性記憶體62（例如管線快取模組64）中所儲存或緩衝的該資訊。例如，處理電路52可在該常態模式或該故障轉移模式中發送該清理指令。

【0023】　為便於理解，假設基於管線的存取管理模組53的一寫入放大減少機制可被暫時禁用（disable），且與本發明相關的某些程式模組可由第3圖的上半部所示來說明。由於儲存伺服器10可依據Ceph解決方案來運作，相關程式模組可包含一物件儲存裝置（Object Storage Device，以下簡稱OSD）資料模組、一OSD

元資料模組、一OSD日誌模組、一鑰值（Key-Value，以下簡稱KV）資料及元資料模組、一KV日誌模組、一第四延伸檔案系統（Fourth Extended File System，以下簡稱EXT4）資料模組、一ÉXT4元資料模組、以及一EXT4日誌模組（在第3圖中分別標示為「OSD資料」、「OSD元資料」、「OSD日誌」、「KV資料及元資料」、「KV日誌」、「EXT4資料」、「EXT4元資料」、及「EXT4日誌」以求簡明），用來依據Ceph解決方案分別記錄（例如寫入）OSD資料、OSD元資料、OSD日誌、KV資料及元資料、KV日誌、EXT4資料、EXT4元資料、以及EXT4日誌，其中這些程式模組可屬於程式模組52P中之一或多個中介層的程式模組，但本發明不限於此。尤其，元資料模組諸如該OSD元資料模組、該KV資料與元資料模組、以及該EXT4元資料模組可分別產生或更新對應的元資料諸如該OSD元資料、該KV元資料以及該EXT4元資料，而日誌模組諸如該OSD日誌模組、該KV日誌模組、以及該EXT4日誌模組可分別產生或更新對應的日誌諸如該OSD日誌、該KV日誌、以及該EXT4日誌。當儲存伺服器10透過網路介面電路58自一用戶端裝置接收一寫入指令，運行著程式模組52P（例如程式模組52P中之該一或多個中介層上方的一上層（upper layer））的處理電路52可將該寫入指令轉譯為一轉譯後寫入指令諸如針對該物件命名空間的一物件寫入指令，以供指出一使用者寫入請求諸如寫入使用者資料的請求（在第3圖中標示為「使用者寫入」以求簡明），其中該物件寫入指令可依據該Ceph解決方案伴隨一物件名稱。該轉譯後寫入指令諸如該物件寫入指令可被傳送至該一或多個中介層中之特定的程式模組。

【0024】　由於該一或多個中介層中之全部的程式模組可依據該Ceph解決方案運作，嚴重的寫入放大會在基於管線的存取管理模組53的該寫入放大減少機制被暫時禁用的情況下發生。第3圖的上半部所示的多個箭號的至少一部分（例如一部分或全部），諸如指向這些程式模組的箭號，可指出可被基於管線的存取管理模組53的該寫入放大減少機制減少的相關寫入放大。例如，該架構的多個部

分的箭號的數量可約略地分別指出對應的寫入放大值。本發明的方法及相關設備能以極佳的方式極度地減少寫入放大（例如減少箭號數量及／或移除這些箭號），以最小化整體的寫入放大。

【0025】　如第3圖的下半部所示，該管線架構可包含一元資料模組諸如該OSD元資料模組、管線快取模組64中之一元快取（Meta-Cache）模組64M（在第3圖中標示為「元快取」以求簡明）、至少一KV模組（例如一或多個KV模組）諸如該KV資料及元資料模組以及該KV日誌模組、以及至少一EXT4模組（例如一或多個EXT4模組）諸如該EXT4日誌模組以及該EXT4元資料模組。這些模組可視為至少一管線（諸如一或多個管線）的多個管線模組，其中元快取模組64M可視為一硬體管線模組，而其它模組可視為軟體管線模組，但本發明不限於此。相較於第3圖上半部所示之架構，一管線模組諸如元快取模組64M可被插入至該管線中以處理該元資料。例如，該管線架構可利用這個管線模組來存取非揮發性雙列直插式記憶體模組60。只要在非揮發性雙列直插式記憶體模組60中提供給這個管線模組的儲存區的相關儲存空間是足夠的，在該管線架構中的通往後續模組（例如該KV模組以及該EXT4模組）的資訊流可變成不活躍的（inactive），如第3圖下半部所示以虛線的箭號來表示以便於理解。例如，運行著程式模組52P的處理電路52可控制儲存伺服器10提供複數個節點（例如四個節點），並且可利用上述至少一非揮發性雙列直插式記憶體模組中之一特定非揮發性雙列直插式記憶體模組諸如非揮發性雙列直插式記憶體模組60作為對應於這個節點的一專用的非揮發性雙列直插式記憶體模組，以針對複數個儲存裝置90中之對應於這個節點的一組儲存裝置（例如六個儲存裝置）的存取進行元資料與日誌管理；複數個儲存裝置90的每一者的儲存容量可為4兆位元組（Terabytes, TB）；以及當在非揮發性雙列直插式記憶體模組60中的該儲存區的儲存空間的大小可達到（例如大於或等於）32吉位元組（Gigabytes, GB），該管線架構可在不激活

（activate）通往後續模組的資訊流的情況下正確地運作；但本發明不限於此。在某些實施例中，基於管線的存取管理模組53的該寫入放大減少機制可在需要時激活通往後續模組的資訊流。

【0026】　依據某些實施例，運行著程式模組52P的處理電路52可將快取資訊（例如該元資料）儲存在非揮發性雙列直插式記憶體模組60中，並且在複數個預定條件的至少一預定條件（例如一或多個預定條件）被滿足時清理管線快取模組64，其中處理電路52可發送該清理指令至非揮發性雙列直插式記憶體模組60以將在非揮發性雙列直插式記憶體模組60中之管線快取模組64中之該快取資訊（例如在元快取模組64M中的該元資料）清理至儲存伺服器10中之該儲存裝置層（在第3圖中標示為「儲存」以求簡明），例如複數個儲存裝置90中之一或多個儲存裝置，諸如儲存裝置100（例如該固態硬碟），但本發明不限於此。該複數個預定條件可包含：儲存伺服器10為閒置；以及管線快取模組64具有高快取壓力（例如在管線快取模組64中的快取資訊的大小達到一預定值）。因此，運行著程式模組52P的處理電路52大部分的時間可繼續使用管線快取模組64中之元快取模組64M以快取並保留該儲存元資料（例如該OSD元資料的最新版本，諸如更新後的OSD元資料）。當儲存伺服器10為閒置且管線快取模組64具有高快取壓力，處理電路52可將於管線快取模組64中之元快取模組64M內的該儲存元資料清理至該儲存裝置層，例如，透過上述至少一KV模組（例如該KV資料及元資料模組以及該KV日誌模組）以及上述至少一EXT4模組（例如該EXT4日誌模組以及該EXT4元資料模組）以及相關清理路徑（例如第3圖的下半部所示之以虛線繪示的箭號）。

【0027】　依據某些實施例，發送該清理指令可透過發送一快取線寫回（Cache Line Write Back, CLWB）指令來實施，以及基於管線的存取管理模組53的該寫入放大減少機制可利用該快取線寫回指令觸發清理運作，例如，透過處理電路52

中之一處理器的一裸晶上（on-die）記憶體控制器，但本發明不限於此。依據某些實施例，一非同步動態隨機存取記憶體刷新（Asynchronous DRAM Refresh，簡稱「ADR」）控制方案可應用於處理電路52，以及管線快取模組64可被用來當作多個ADR保護緩衝器（ADR protected buffer）的其中一者。處理電路52（例如其內的該處理器／處理器核心）可清理該多個ADR保護緩衝器並且控制上述至少一非揮發性雙列直插式記憶體模組諸如非揮發性雙列直插式記憶體模組60的全部處於一自行刷新狀態，並且宣稱（assert）一旗標諸如一ADR完成旗標。因應宣稱該旗標，上述至少一非揮發性雙列直插式記憶體模組諸如非揮發性雙列直插式記憶體模組60可將揮發性記憶體62（例如該動態隨機存取記憶體）從該主裝置（例如處理電路52）隔離並且切換至該備用電源諸如該超級電容器以依據該超級電容器的電源運作，並且將於揮發性記憶體62內的資訊（例如該快取資訊）複製到非揮發性記憶體66以保存該資訊，再接著關閉該超級電容器。

【0028】　針對第3圖所示之基於管線的快取控制方案的某些實施細節可說明如下。依據某些實施例，該管線架構可利用上述至少一非揮發性雙列直插式記憶體模組諸如非揮發性雙列直插式記憶體模組60（例如管線快取模組64）作為如上所述之用來快取該儲存元資料（例如該OSD元資料）的媒介。類似地，該管線架構可利用上述至少一非揮發性雙列直插式記憶體模組（例如一或多個非揮發性雙列直插式記憶體模組）諸如非揮發性雙列直插式記憶體模組60作為用來快取該儲存日誌（例如該OSD日誌）的媒介。為便於理解，假設上述至少一非揮發性雙列直插式記憶體模組可包含一特定非揮發性雙列直插式記憶體模組諸如非揮發性雙列直插式記憶體模組60以供在一每節點一個非揮發性雙列直插式記憶體模組的組態（one NVDIMM per node configuration）中之該複數個節點之一特定節點之用。在此情況下，管線快取模組64可包含多個子模組諸如子快取模組64A及64B，以分別用來快取該儲存元資料（例如該OSD元資料）以及該

儲存日誌（例如該OSD日誌），其中子快取模組61A可代表元快取模組64M。例如，在上述至少一管線（例如該一或多個管線）諸如第3圖下半部所示管線之中，子快取模組64A可被配置在該元資料模組（例如該OSD元資料模組）與上述至少一KV模組（例如該KV資料及元資料模組以及該KV日誌模組）之間。又例如，在該管線架構的另一管線中，子快取模組64B可被配置在一日誌模組諸如該OSD日誌模組與該儲存裝置層之間。上述至少一管線以及上述另一管線中之兩者均可用來以寫入的方向（諸如自該上層透過該一或多個中介層至該儲存裝置層的方向）傳送資訊（例如OSD、KV及EXT4之各自的資料、各自的元資料、以及各自的日誌），以及運行著程式模組52P（例如對應於該方法之基於管線的存取管理模組53）的處理電路52可分別利用子快取模組64A及64B來快取該儲存元資料（例如該OSD元資料的最新版本諸如更新後的OSD元資料）以及該儲存日誌（例如該OSD日誌的最新版本諸如更新後的OSD日誌），以在大部分的時間消除在該管線架構的這些管線內的後續資訊流。因此，本發明的方法以及相關設備能大幅地減少整體寫入放大並且改善整體效能。

【0029】　依據某些實施例，上述至少一非揮發性雙列直插式記憶體模組可包含複數個非揮發性雙列直插式記憶體模組{60}諸如N_{NVDIMM}個非揮發性雙列直插式記憶體模組{60_1, ..., 60_N_{NVDIMM}}，其數量N_{NVDIMM}可為大於一的正整數。該N_{NVDIMM}個非揮發性雙列直插式記憶體模組{60_1, ..., 60_N_{NVDIMM}}可具有與非揮發性雙列直插式記憶體模組60相同的架構。例如，非揮發性雙列直插式記憶體模組60_1可包含一揮發性記憶體62_1以及一非揮發性記憶體66_1，而一管線快取模組64_1可被設置於揮發性記憶體62_1中；非揮發性雙列直插式記憶體模組60_2可包含一揮發性記憶體62_2以及一非揮發性記憶體66_2，而一管線快取模組64_2可被設置於揮發性記憶體62_2中；而其餘可依此類推。另外，主裝置50可如上所述，以該每節點一個非揮發性雙列直插式記憶體模組的組態來運

作，其中揮發性記憶體62_1中之管線快取模組64_1可扮演子快取模組64A的角色，而揮發性記憶體62_2中之管線快取模組64_2可扮演子快取模組64B的角色，但本發明不限於此。例如，上述至少一非揮發性雙列直插式記憶體模組可包含多個非揮發性雙列直插式記憶體模組諸如非揮發性雙列直插式記憶體模組60_1及60_2以供在一每節點多個非揮發性雙列直插式記憶體模組的組態（multi-NVDIMM per node configuration）中之該複數個節點之一特定節點之用。尤其，運行著程式模組52P（例如對應於該方法之基於管線的存取管理模組53）的處理電路52可利用管線快取模組64_1及64_2之各自的子模組（諸如子快取模組{64_1A, 64_1B}及{64_2A, 64_2B }）來分別快取該儲存元資料（例如該OSD元資料的最新版本諸如更新後的OSD元資料）以及該儲存日誌（例如該OSD日誌的最新版本諸如更新後的OSD日誌），以在大部分的時間消除在該管線架構的這些管線內的後續資訊流。在某些實施例中，運行著程式模組52P（例如對應於該方法之基於管線的存取管理模組53）的處理電路52可分別利用管線快取模組64_1及64_2來快取該儲存元資料（例如該OSD元資料的最新版本諸如更新後的OSD元資料）以及該儲存日誌（例如該OSD日誌的最新版本諸如更新後的OSD日誌），以在大部分的時間消除在該管線架構的這些管線內的後續資訊流。

【0030】　不論主裝置50是否以該每節點一個非揮發性雙列直插式記憶體模組的組態或該每節點多個非揮發性雙列直插式記憶體模組的組態來運作，具有至少一快取模組（例如一或多個快取模組諸如管線快取模組64、子快取模組64A及64B、管線快取模組64_1及64_2等）的管線結構能急劇地緩解在儲存運作中的寫入放大，尤其，能改善在儲存裝置層中的底層的儲存裝置（例如固態硬碟）的壽命並且減少在該複數個儲存系統（例如該多個儲存伺服器{10}）之中及／或之間的資料流量。另外，備有上述至少一快取模組的該管線架構能減少因為該Ceph解決方案所帶來的副作用（諸如相較於非Ceph的架構更大的輸入輸出

（input/output, I/O）延遲、更大量的記憶體複製以及更大量的鎖定），尤其，能使儲存伺服器10近乎完全地利用在該儲存裝置層的該多個儲存裝置（例如固態硬碟）並且輕易地將資料直接寫入至該多個儲存裝置中。例如，該管線架構可配置為具有一組指定的緩衝器（例如子快取模組64A及64B、管線快取模組64_1及64_2等）以增加寫入流通量（write throughput）。又例如，該管線架構可包含元資料之一基於層級的（tier-based）的儲存架構以減少詢問資料的反應時間。於是，依據本發明來實施能達到改善整體效能的目標。

【0031】　依據某些實施例，具有上述至少一快取模組（例如一或多個快取模組諸如管線快取模組64、子快取模組64A及64B、管線快取模組64_1及64_2等）的該管線架構可自網路直接接收該使用者資料（例如透過該有線或無線網路連接自該一或多個用戶端裝置取得該OSD資料）至非揮發性雙列直插式記憶體模組60，並且可將對應於使用者寫入操作的元資料（例如該OSD元資料）直接產生至非揮發性雙列直插式記憶體模組60中，並可另寫入相關日誌（例如該OSD日誌），例如，藉由將（關於通訊的）交易（transaction）的標頭（header）編碼至非揮發性雙列直插式記憶體模組60中。由於上述運作的資料或資訊流會先進入非揮發性雙列直插式記憶體模組60，並且由於對應的元資料以及相關日誌會在非揮發性雙列直插式記憶體模組60中被產生、改變或更新，所以，自該一或多個中介層朝向該儲存裝置層的後續資訊流，連同對應於這些後續資訊流的寫入放大，能被大幅地減少（尤其，消除），且因此整體寫入放大能被大幅減少。於是，依據本發明來實施能達到改善整體效能的目標。

【0032】　依據某些實施例，本發明的方法以及相關設備（例如主裝置50，尤其，運行著基於管線的存取管理模組53的處理電路52）可僅利用該管線架構之上述至少一快取模組（例如一或多個快取模組諸如管線快取模組64、子快取模組64A及64B、管線快取模組64_1及64_2等），而非該儲存裝置層，來儲存該儲存

元資料（例如該OSD元資料）以及該儲存日誌（例如該OSD日誌）。在此情況下，自該一或多個中介層朝向該儲存裝置層的後續資訊流會被消除，且因此，將該儲存元資料（例如該OSD元資料）以及該儲存日誌（例如該OSD日誌）快取在該管線架構之上述至少一快取模組中可視為將該儲存元資料（例如該OSD元資料）以及該儲存日誌（例如該OSD日誌）儲存在該管線架構之上述至少一快取模組中。

【0033】　第4圖依據本發明一實施例繪示對應於第3圖所示之方法的效能增強。為便於理解，由在一用戶端裝置測量一普通Ceph伺服器的用戶端延遲（client latency）得到的一第一用戶端延遲圖可繪示如第4圖的左半部所示，而由在這個用戶端裝置測量儲存伺服器10的用戶端延遲得到的一第二用戶端延遲圖可繪示如第4圖的右半部所示，其中圖例諸如Max、Min及Avg可分別代表最大值、最小值以及平均值。例如，在該第一用戶端延遲圖中的量測結果（例如資料點）有往上方散布的趨勢且到達40毫秒（millisecond，以下簡稱「ms」）。另外，在該第二用戶端延遲圖中的量測結果（例如資料點）的大部分均維持在低處且互相靠近，這表示相對於時間的整體延遲可被大幅地減少。因此，備有上述至少一快取模組的管線架構確實能使該儲存系統（例如儲存伺服器10）加速。

【0034】　第5圖繪示用於每一寫入運作（例如因應該物件寫入指令的一物件寫入運作）的該OSD元資料，諸如將被儲存的相關資訊。例如，該資訊的至少一部分（例如一部分或全部）可被儲存為複數組的鑰（key, K）與值（value, V）。依據本實施例，本發明的方法以及相關設備可更新（尤其，重新寫入）物件資訊（information, Info）、物件快照（snapshot）資訊、放置群組（placement group, PG）資訊、以及放置群組登錄（log），例如，透過維護在非揮發性雙列直插式記憶體模組60中的該OSD元資料。當該OSD元資料被儲存為多個鑰值組（KV set），該放置群組資訊、該放置群組登錄、該物件資訊及該物件快照資訊之各自

的鑰可分別載有（carry）放置群組名稱加固定後置（PG name plus fixed postfix）、放置群組名稱加登錄版本（PG name plus log version）、物件名稱加固定後置（object name plus fixed postfix）以及物件名稱加固定後置（例如針對快照），而該放置群組資訊、該放置群組登錄、該物件資訊及該物件快照資訊之各自的值長度可分別是至少數百位元組、大約一百八十或更多位元組（以寫入一特定類型的資料物件為例）、數百至數千位元組、及至少數十位元組（以空資訊為例），但本發明不限於此。另外，一放置群組可包含多個物件。例如，該放置群組可被用來作為用於進行復原（recovery）的單元、用於移動物件的單元、及／或用於平衡在該儲存系統諸如儲存伺服器10中之複數個區塊裝置之間的物件數量（例如當一新區塊裝置諸如一新固態硬碟被加入該儲存系統）。由於每一寫入運作（例如該物件寫入運作）之針對該OSD元資料的資訊流先進入非揮發性雙列直插式記憶體模組60，且由於該物件資訊、該物件快照資訊、該放置群組資訊以及該放置群組登錄可在非揮發性雙列直插式記憶體模組60中被產生、更改、或更新，自該一或多個中介層朝向該儲存裝置層的後續資訊流，連同對應於這些後續資訊流的寫入放大，能被大幅地減少（尤其，消除），且因此整體寫入放大能被減少。於是，依據本發明來實施能達到改善整體效能的目標。

【0035】　第6圖繪示一原始OSD日誌控制方案的一系列運作。為便於理解，假設基於管線的存取管理模組53的該寫入放大減少機制可被暫時禁用，並且在此情況下對應於上述嚴重的寫入放大的某些原始行為可繪示如第6圖所示，但本發明不限於此。由於典型地需要與作業系統核心（Operating System kernel, OS kernel）互動，故整體效能在此情況下會衰退。例如，來自多個執行緒（thread）的複數個OSD交易可分別被接收，並且可分別被佇列（queue），其在第6圖中標示為「佇列交易」以求簡明。雖然可分別對這些交易進行編碼（在第6圖中標示為「編碼交易」以求簡明），在寫入佇列、日誌區塊裝置、及相關互動（例如一

體傳遞（all in one submit, AIO submit）以及得到事件）的運作中可能會發生各種問題，諸如鎖定（lock）、系統調用（system call）、記憶體複製、佇列／排程等問題。尤其，該原始OSD日誌控制方案的該系列運作可包含：透過該系統調用自使用者空間切換至核心空間（kernel space）以進行該使用者空間與該核心空間之間的傳輸；進行該記憶體複製；透過上述佇列／排程存取該日誌區塊裝置；以及對應於相反方向的某些運作。另外，基於該原始OSD日誌控制方案，這些交易可先被編碼並且收集在一起，接著發送至於處理電路52中的隨機存取記憶體，接著再以作業系統核心區塊輸入輸出來處理，並且透過直接記憶體存取寫入（Direct Memory Access write, DMA write）發送至該儲存裝置層（例如複數個儲存裝置90的其中一者，諸如一固態硬碟）。由於對應這些交易中之一特定交易的使用者資料可能需要被儲存至一外部緩衝器，且由於對應這個交易的OSD元資料可被表示為在處理電路52中之隨機存取記憶體當中之一複雜的樹狀資料，其具有大量的分布於其各種位置之鑰與值，故針對這個交易的處理會很複雜，且針對這些交易的全部的處理可對應於相當繁重的工作負荷。例如，整個程序可能被於該寫入佇列中所佇列的交易之大量的鎖定阻塞，且可能被該系統調用、該記憶體複製等運作延遲，尤其，可能需要等待該日誌區塊裝置的佇列／排程。於是，該儲存系統的整體反應時間可能會增加。

【0036】　第7圖依據本發明一實施例繪示在第3圖所示之方法的一專門OSD日誌控制方案中的複數個每執行緒日誌環（per thread journal ring）以及在該專門OSD日誌控制方案中的相關運作的一部分，而第8~10圖繪示在第7圖所示之專門OSD日誌控制方案中的相關運作的其它部分。基於該專門OSD日誌控制方案，本發明的方法及相關設備（例如主裝置50，尤其，運行著基於管線的存取管理模組53的處理電路52）可藉助於上述至少一非揮發性雙列直插式記憶體模組（例如非揮發性雙列直插式記憶體模組60或複數個非揮發性雙列直插式記憶體模組

{60}（在第7~10圖中標示為「NVDIMM」以求簡明））中之該複數個每執行緒日誌環（例如分別對應複數個執行緒的日誌環，諸如分別專用於這些執行緒的專屬日誌環）以及複數個固定大小緩衝器池（fixed size buffer pool）來進行OSD資訊管理，以避免第6圖所示之該原始OSD日誌控制方案的問題（例如鎖定、系統調用、記憶體複製、及佇列／排程等問題）。

【0037】　如第7圖所示，該設備可進行緩衝器的無鎖定分配（lock free allocation）以分配（allocate）對應該使用者資料的緩衝器，並且將該使用者資料直接接收至非揮發性雙列直插式記憶體模組60（而不是先接收至於處理電路52中的隨機存取記憶體），例如，自儲存伺服器10的外部透過一網路服務諸如傳輸控制通訊協定（Transmission Control Protocol，簡稱「TCP」）的網路服務進入非揮發性雙列直插式記憶體模組60中之分配到的緩衝器，其中來自發信者（Messenger）的使用者寫入請求可包含該使用者資料以及該使用者資料的一標頭。例如，基於管線的存取管理模組53的該寫入放大減少機制可準備並且管理上述至少一非揮發性雙列直插式記憶體模組（例如非揮發性雙列直插式記憶體模組60或複數個非揮發性雙列直插式記憶體模組{60}）中之該複數個固定大小緩衝器池，諸如分別對應一較小緩衝大小與一較大緩衝大小的一小緩衝器池與一大緩衝器池，尤其，準備並管理各種大小的緩衝器，諸如在該小緩衝器池內的多個小緩衝器以及在該大緩衝器池內的多個大緩衝器，以供在緩衝器分配的期間被選擇，但本發明不限於此。於是，主裝置50（例如運行著基於管線的存取管理模組53的處理電路52）能避免第6圖所示之該原始OSD日誌控制方案的鎖定問題。

【0038】　如第8圖所示，在建立對應於該使用者資料的交易（例如該OSD交易）時，該設備可將相關OSD元資料直接寫入至非揮發性雙列直插式記憶體模組60，而不是先寫入至在處理電路52中的隨機存取記憶體。例如，對應於該使用

者資料的交易（例如該OSD交易）可包含複數個運作（operation, OP）諸如五個運作：使用者資料（user data，在第8圖中標示為「UD」以求簡明）；實質上可等於該OSD元資料的三個包裝後（packed）鑰值組（例如三組包裝後鑰值），其中該三個包裝後鑰值組可包含一第一包裝後鑰值組{(K_1, V_1)}諸如多個第一鑰值組{($K_1(1)$, $V_1(1)$), ($K_1(2)$, $V_1(2)$), …}、一第二包裝後鑰值組{(K_2, V_2)諸如多個第二鑰值組{($K_2(1)$, $V_2(1)$), ($K_2(2)$, $V_2(2)$), …}以及一第三包裝後鑰值組{(K_3, V_3)}諸如多個第三鑰值組{($K_3(1)$, $V_3(1)$), ($K_3(2)$, $V_3(2)$), …}；以及一放置群組登錄。由於第二、第三、及第四運作諸如該三個包裝後鑰值組可視為該OSD元資料，該設備可建立對應於該使用者資料的OSD交易，利用該OSD交易的該三個包裝後鑰值組作為該OSD元資料，以及將該OSD交易的該三個包裝後鑰值組直接寫入至非揮發性雙列直插式記憶體模組60以作為該OSD元資料，不需要進行額外的處理來產生該OSD元資料。尤其，基於管線的存取管理模組53的該寫入放大減少機制可預先依據該標頭判斷該使用者資料、該三個包裝後鑰值組及該放置群組登錄之各自的大小，以判斷並指定該OSD元資料中之該三個包裝後鑰值組被寫入至非揮發性雙列直插式記憶體模組60之各自的儲存位置，其中預先依據該標投決定的這些儲存位置可視為給該OSD元資料的內容的預定儲存位置，並且該OSD元資料可被附加在非揮發性雙列直插式記憶體模組60中之該使用者資料（例如就在該使用者資料之後）。於是，當有需要時（例如該使用者資料被更改時），基於管線的存取管理模組53的該寫入放大減少機制可直接地更改或更新在非揮發性雙列直插式記憶體模組60中之該OSD元資料的任何部分而不需移動（或重新寫入）該OSD元資料的另一部分，其中第6圖所示之該原始OSD日誌控制方案的系統調用問題、記憶體複製問題等就不會發生。

【0039】　如第9圖所示，該設備可以僅針對該交易（例如該OSD交易）之運作來編碼該交易（在第9圖中標示為「以僅針對交易之運作來編碼交易」以求簡明）

而不碰觸該使用者資料，這對中央處理單元的快取是友善的。由於該OSD元資料中之該使用者資料以及該三個包裝後鑰值組已被儲存於非揮發性雙列直插式記憶體模組60中之分配到的緩衝器（例如在該緩衝分配的期間自該複數個固定大小緩衝器池中選擇的緩衝器）中，該OSD交易之大部分的內容在非揮發性雙列直插式記憶體模組60中是可用的（available），並且可透過指向分配到的緩衝器的位址資訊（例如至少一位址）來存取，諸如透過指向分配到的緩衝器的起始處的一緩衝器位址、分別指向分配到的緩衝器中之該使用者資料及該OSD元資料的兩個子緩衝器位址（sub-buffer address）、分別指向分配到的緩衝器中之該使用者資料及該三個包裝後鑰值組（其被寫入作為該OSD元資料）的一組子緩衝器位址等來存取。例如，基於管線的存取管理模組53的該寫入放大減少機制可在一執行緒擁有環（thread owned ring）諸如該複數個每執行緒日誌環中之一者上工作，而不會有任何鎖定諸如該原始OSD日誌控制方案的鎖定（在第9圖中標示為「沒有鎖定」以求簡明），尤其，可針對該OSD交易的運作來編碼該OSD交易以在該執行緒擁有環中產生並記錄該OSD交易的一編碼結果，其中該OSD交易的該編碼結果可包含指向分配的緩衝器的該位址資訊（例如該緩衝器位址、該兩個子緩衝器位址、該組子緩衝器位址等中之任何位址），且可另包含該放置群組登錄，但本發明不限於此。於是，被編碼的交易（例如該OSD交易的在該執行緒擁有環中的該編碼結果）可等於被寫入的日誌（在第10圖中標示為「編碼後交易==寫入的日誌」以便於理解）。如第10圖所示，當該交易諸如該OSD交易已被完全編碼，該日誌諸如該OSD日誌已被完全寫入，故該交易被編碼的狀態可視為該日誌被寫入的狀態。另外，主裝置50（例如運行著基於管線的存取管理模組53的處理電路52）在該日誌被寫入後可即時地回覆該使用者（例如該使用者的用戶端裝置）。為簡明起見，本實施例中與前述實施例類似之內容在此不重複贅述。

【0040】　依據某些實施例，該設備（例如主裝置50，尤其，運行著基於管線的存取管理模組53的處理電路52）可依據基於預先剖析（parsing）該標投的結果的一計畫進行該專門OSD日誌控制方案的相關運作。該設備可剖析第7圖所示之標頭以預先判斷該OSD元資料中之各組資訊將被儲存於分配到的緩衝器中之各自的位置（例如在非揮發性雙列直插式記憶體模組60中的儲存位置）。依據預先判斷的這些位置，該設備可將該OSD元資料中之各組資訊於建立該交易時直接寫入非揮發性雙列直插式記憶體模組60。由於本發明的方法以及相關設備能以極佳的方式極度地盡力減少與該作業系統核心的互動，儲存伺服器10的整體反應時間能被大幅地減少，故整體的效能能被提升。為簡明起見，這些實施例中與前述實施例類似之內容在此不重複贅述。

【0041】　依據某些實施例，該複數個固定大小緩衝器池可包含分別對應一第一緩衝大小、一第二緩衝大小、以及一第三緩衝大小的一第一緩衝器池、一第二緩衝器池、以及一第三緩衝器池。例如，針對具有一第一資料大小（例如4千位元組（kilobytes，以下簡稱KB））的第一使用者資料（例如OSD資料#1），基於管線的存取管理模組53的該寫入放大減少機制可自該第一緩衝器池分配具有該第一緩衝大小（例如12 KB）的一第一緩衝器以供處理該第一使用者資料（例如該OSD資料#1）以及相關的儲存元資料與儲存日誌（例如其OSD元資料#1與OSD日誌#1）；針對具有一第二資料大小（例如12 KB）的第二使用者資料（例如OSD資料#2），基於管線的存取管理模組53的該寫入放大減少機制可自該第二緩衝器池分配具有該第二緩衝大小（例如24 KB）的一第二緩衝器以供處理該第二使用者資料（例如該OSD資料#2）以及相關的儲存元資料與儲存日誌（例如其OSD元資料#2與OSD日誌#2）；以及針對具有一第三資料大小（例如50 KB）的第三使用者資料（例如OSD資料#3），基於管線的存取管理模組53的該寫入放大減少機制可自該第三緩衝器池分配具有該第三緩衝大小（例如72 KB）的一第

三緩衝器以供處理該第三使用者資料（例如該OSD資料#3）以及相關的儲存元資料與儲存日誌（例如其OSD元資料#3與OSD日誌#3）；但本發明不限於此。為簡明起見，這些實施例中與前述實施例類似之內容在此不重複贅述。

【0042】　第11圖為依據本發明一實施例之第3圖所示之方法的工作流程。

【0043】　在步驟S11中，因應將該使用者資料（例如該OSD資料）寫入儲存伺服器10的請求，儲存伺服器10利用儲存伺服器10中之主裝置50將該使用者資料寫入儲存伺服器10的該儲存裝置層，並且開始以儲存伺服器10的該管線架構處理對應於寫入該使用者資料的該請求的物件寫入指令，其中儲存伺服器10包含主裝置50以及該儲存裝置層，以及該儲存裝置層包含耦接至主裝置50的至少一儲存裝置（例如複數個儲存裝置90中之該一或多個儲存裝置，諸如儲存裝置100）。

【0044】　在步驟S12中，在以該管線架構處理該物件寫入指令的期間，儲存伺服器10利用主裝置50將對應於該使用者資料的元資料（例如對應於該OSD資料的該OSD元資料）輸入該管線架構中之上述至少一管線，其中該管線架構穿過運行於主裝置50之處理電路52上的程式模組52P中之該一或多個中介層的程式模組，並且達到該儲存裝置層。

【0045】　在步驟S13中，在以該管線架構處理該物件寫入指令的期間，儲存伺服器10利用主裝置50以上述至少一管線的一第一快取模組（例如管線快取模組64中之元快取模組64M、子快取模組64A、管線快取模組64_1等）快取該元資料（例如該OSD元資料），以控制儲存伺服器10在不產生該元資料（例如該OSD元資料）的針對該一或多個中介層的程式模組之任何寫入放大的情況下完成該請求，其中該第一快取模組係在該儲存裝置層外部的一硬體管線模組。

【0046】　在步驟S14中，在以該管線架構處理該物件寫入指令的期間，儲存伺服器10利用主裝置50將對應於該使用者資料的一日誌（例如對應於該OSD資料

的該OSD日誌）輸入該管線架構中之另一管線（諸如上述者）。

【0047】　在步驟S15中，在以該管線架構處理該物件寫入指令的期間，儲存伺服器利用主裝置50以該另一管線的一第二快取模組（例如管線快取模組64中之另一快取模組、子快取模組64B、管線快取模組64_2等）快取該日誌（例如該OSD日誌），以控制儲存伺服器10在不產生該日誌（例如該OSD日誌）的針對該一或多個中介層的程式模組之任何寫入放大的情況下完成該請求，其中該第二快取模組係在該儲存裝置層外部的一硬體管線模組。

【0048】　依據本實施例，主裝置50可以上述至少一管線之該第一快取模組快取該元資料（例如該OSD資料），以避免上述至少一管線的複數個後續管線模組（例如第3圖所示之實施例中提及的後續模組，諸如該KV模組及該EXT4模組）產生該元資料的衍生資訊，尤其，避免該元資料的該衍生資訊被產生並寫入至該儲存裝置層，其中該一或多個中介層的程式模組包含上述至少一管線的該複數個後續管線模組。另外，該第一快取模組以及該第二快取模組可被設置於上述至少一非揮發性雙列直插式記憶體模組（例如非揮發性雙列直插式記憶體模組60或複數個非揮發性雙列直插式記憶體模組{60}）中，並且該第一快取模組與該第二快取模組中之任一快取模組可代表上述至少一非揮發性雙列直插式記憶體模組中之一特定非揮發性雙列直插式記憶體模組的揮發性記憶體中之一儲存區。例如，該第一快取模組以及該第二快取模組（例如子快取模組64A及64B）可代表非揮發性雙列直插式記憶體模組60的揮發性記憶體62中之不同的儲存區。又例如，該第一快取模組以及該第二快取模組（例如管線快取模組64_1及64_2）可分別代表非揮發性雙列直插式記憶體模組60_1及60_2之各自的揮發性記憶體62_1及62_2之各自的儲存區。另外，上述至少一非揮發性雙列直插式記憶體模組可提供該儲存裝置層外部的至少一硬體管線模組（例如一或多個硬體管線模組），諸如包含有該第一快取模組以及該第二快取模組的複數個硬體管線

模組。為簡明起見，本實施例中與前述實施例類似之內容在此不重複贅述。

【0049】　為便於理解，該方法可用第11圖所示之工作流程來說明，但本發明不限於此。依據某些實施例，一或多個步驟可於第11圖所示之工作流程中被新增、刪除、或修改。

【0050】　第12圖依據本發明一實施例繪示第1圖所示之管線快取模組64的某些子模組。如第12圖所示，管線快取模組64可包含子快取模組64A及64B，以分別快取該儲存元資料（例如該OSD元資料）以及該儲存日誌（例如該OSD日誌）。為簡明起見，本實施例中與前述實施例類似之內容在此不重複贅述。

【0051】　第13圖依據本發明一實施例繪示第1圖所示之主裝置10的某些非揮發性雙列直插式記憶體模組。如第13圖所示，主裝置10可包含該N_{NVDIMM}個非揮發性雙列直插式記憶體模組{60_1, 60_2, …, 60_N_{NVDIMM}}（在第13圖中分別標示為「NVDIMM」以求簡明）。例如，非揮發性雙列直插式記憶體模組60_1可包含揮發性記憶體62_1以及非揮發性記憶體66_1（在第13圖中分別標示為「VM」及「NVM」以求簡明），其可分別用來為處理電路52儲存（例如緩衝或快取）資訊以及為處理電路52保存在揮發性記憶體62_1中的該資訊，並且可耦接至處理電路52以供處理電路52存取；非揮發性雙列直插式記憶體模組60_2可包含揮發性記憶體62_2以及非揮發性記憶體66_2（在第13圖中分別標示為「VM」及「NVM」以求簡明），其可分別用來為處理電路52儲存（例如緩衝或快取）資訊以及為處理電路52保存在揮發性記憶體62_2中的該資訊，並且可耦接至處理電路52以供處理電路52存取；而其餘可依此類推。類似地，非揮發性雙列直插式記憶體模組60_N_{NVDIMM}可包含一揮發性記憶體62_N_{NVDIMM}以及一非揮發性記憶體66_N_{NVDIMM}（在第13圖中分別標示為「VM」及「NVM」以求簡明），其可分別用來為處理電路52儲存（例如緩衝或快取）資訊以及為處理電路52保存在揮發性記憶體62_N_{NVDIMM}中的該資訊，並且可耦接至處理電路52以供處理電路52存

取。另外，管線快取模組64_1、64_2、…及64_N_{NVDIMM}（在第13圖中分別標示為「管線快取」以求簡明）可分別被設置在揮發性記憶體62_1、62_2、…及62_N_{NVDIMM}中。為簡明起見，本實施例中與前述實施例類似之內容在此不重複贅述。

以上所述僅為本發明之較佳實施例，凡依本發明申請專利範圍所做之均等變化與修飾，皆應屬本發明之涵蓋範圍。

【符號說明】

【0052】

10・・・儲存伺服器

50・・・主裝置

52・・・處理電路

52P・・・程式模組

53・・・基於管線的存取管理模組

54・・・儲存介面電路

56・・・儲存裝置

58・・・網路介面電路

60、

60_1、60_2、…、60_N_{NVDIMM}・・・非揮發性雙列直插式記憶體模組

62、

62_1、62_2、…、62_N_{NVDIMM}・・・揮發性記憶體

64、

64_1、64_2、…、64_N_{NVDIMM}・・・管線快取模組

64_A、64_B・・・子快取模組

66、

66_1、66_2、…、66_N_{NVDIMM}・・・非揮發性記憶體

90・・・儲存裝置

100・・・儲存裝置

110・・・記憶體控制器

112・・・微處理器

112M・・・唯讀記憶體

112C・・・程式碼

114・・・控制邏輯電路

116・・・隨機存取記憶體

118・・・傳輸介面電路

120・・・非揮發性記憶體

122-1、122-2、…、122-N・・・非揮發性記憶體元件

S11、S12、S13、S14、S15・・・步驟

【發明申請專利範圍】

【第1項】 一種用來在一儲存伺服器中進行基於管線的（pipeline-based）存取管理的方法，該方法係應用於該儲存伺服器，該方法包含：

因應將使用者資料寫入該儲存伺服器的一請求，利用該儲存伺服器中之一主裝置將該使用者資料寫入該儲存伺服器之一儲存裝置層並且以該儲存伺服器的一管線（pipeline）架構開始處理對應於寫入該使用者資料的該請求的一物件寫入指令，其中該儲存伺服器包含該主裝置以及該儲存裝置層，該儲存裝置層包含耦接至該主裝置的至少一儲存裝置，該主裝置係用來控制該儲存伺服器的運作，以及所述至少一儲存裝置係用來為該儲存伺服器儲存資訊；

在以該管線架構處理該物件寫入指令的期間，利用該主裝置將對應於該使用者資料的元資料（metadata）輸入該管線架構中之至少一管線，其中該管線架構穿過運行於該主裝置之一處理電路上的複數個程式模組中之一或多個中介層（intermediate layer）的程式模組，並且達到該儲存裝置層；以及

在以該管線架構處理該物件寫入指令的期間，利用該主裝置以所述至少一管線的一第一快取（cache）模組快取該元資料，以控制該儲存伺服器在不產生該元資料的針對該一或多個中介層的程式模組之寫入放大的情況下完成該請求，其中該第一快取模組係在該儲存裝置層外部的一硬體管線模組。

【第2項】 如申請專利範圍第1項所述之方法，其中利用該主裝置以所述至少一管線的該第一快取模組快取該元資料的步驟另包含：

利用該主裝置以所述至少一管線的該第一快取模組快取該元資料，以避免所

述至少一管線的複數個後續管線模組產生該元資料的衍生資訊，其中該一或多個中介層的程式模組包含所述至少一管線的該複數個後續管線模組。

【第3項】 如申請專利範圍第1項所述之方法，其中利用該主裝置以所述至少一管線的該第一快取模組快取該元資料的步驟另包含：

利用該主裝置以所述至少一管線的該第一快取模組快取該元資料，以避免該元資料的衍生資訊被產生以及被寫入該儲存裝置層。

【第4項】 如申請專利範圍第1項所述之方法，其中該主裝置包含至少一非揮發性雙列直插式記憶體模組（non-volatile dual in-line memory module, NVDIMM），以及該第一快取模組位於所述至少一非揮發性雙列直插式記憶體模組中。

【第5項】 如申請專利範圍第4項所述之方法，其中所述至少一非揮發性雙列直插式記憶體模組中之任一非揮發性雙列直插式記憶體模組包含一揮發性（volatile）記憶體以及一非揮發性（non-volatile, NV）記憶體以分別用來為該處理電路儲存資訊以及為該處理電路保存在該揮發性記憶體中的該資訊；以及該第一快取模組代表該揮發性記憶體中之一儲存區。

【第6項】 如申請專利範圍第1項所述之方法，另包含：

在以該管線架構處理該物件寫入指令的期間，利用該主裝置將對應於該使用者資料的一日誌（journal）輸入該管線架構中之另一管線；以及

在以該管線架構處理該物件寫入指令的期間，利用該主裝置以該另一管線的

一第二快取模組快取該日誌，以控制該儲存伺服器在不產生該日誌的針對該一或多個中介層的程式模組之寫入放大的情況下完成該請求，其中該第二快取模組係在該儲存裝置層外部的一硬體管線模組。

【第7項】 如申請專利範圍第6項所述之方法，其中該主裝置包含至少一非揮發性雙列直插式記憶體模組（non-volatile dual in-line memory module, NVDIMM），以及該第一快取模組以及該第二快取模組位於所述至少一非揮發性雙列直插式記憶體模組中。

【第8項】 如申請專利範圍第7項所述之方法，其中所述至少一非揮發性雙列直插式記憶體模組中之任一非揮發性雙列直插式記憶體模組包含一揮發性（volatile）記憶體以及一非揮發性（non-volatile, NV）記憶體以分別用來為該處理電路儲存資訊以及為該處理電路保存在該揮發性記憶體中的該資訊；以及該第一快取模組與該第二快取模組中之任一快取模組代表該揮發性記憶體中之一儲存區。

【第9項】 如申請專利範圍第1項所述之方法，其中該管線架構自運行於該主裝置之該處理電路上的該複數個程式模組中之一上層起始，穿過該一或多個中介層的程式模組，並且達到該儲存裝置層；以及該方法另包含：

利用至少一非揮發性雙列直插式記憶體模組（non-volatile dual in-line memory module, NVDIMM）來提供該儲存裝置層外部的至少一硬體管線模組，其中所述至少一硬體管線模組包含該第一快取模組。

【第10項】如申請專利範圍第9項所述之方法，其中所述至少一硬體管線模組包

含複數個硬體管線模組；以及該方法另包含：

在以該管線架構處理該物件寫入指令的期間，利用該主裝置將對應於該使用者資料的一日誌（journal）輸入該管線架構中之另一管線；以及

在以該管線架構處理該物件寫入指令的期間，利用該主裝置以該另一管線的一第二快取模組快取該日誌，以控制該儲存伺服器在不產生該日誌的針對該一或多個中介層的程式模組之寫入放大的情況下完成該請求，其中該複數個硬體管線模組包含該第一快取模組以及該第二快取模組。

【第11項】一種主裝置，包含：

一處理電路，用來控制該主裝置在一儲存伺服器中進行基於管線的（pipeline-based）存取管理，其中該儲存伺服器包含該主裝置以及一儲存裝置層，該儲存裝置層包含耦接至該主裝置的至少一儲存裝置，該主裝置係用來控制該儲存伺服器的運作，以及所述至少一儲存裝置係用來為該儲存伺服器儲存資訊，其中：

因應將使用者資料寫入該儲存伺服器的一請求，該儲存伺服器中之該主裝置將該使用者資料寫入該儲存伺服器之該儲存裝置層並且以該儲存伺服器的一管線（pipeline）架構開始處理對應於寫入該使用者資料的該請求的一物件寫入指令；

在以該管線架構處理該物件寫入指令的期間，該主裝置將對應於該使用者資料的元資料（metadata）輸入該管線架構中之至少一管線，其中該管線架構穿過運行於該主裝置之該處理電路上的複數個程式模組中之一或多個中介層（intermediate layer）的程式模組，並且達到該儲存裝置層；以及

在以該管線架構處理該物件寫入指令的期間，該主裝置以所述至少一管

線的一第一快取（cache）模組快取該元資料，以控制該儲存伺服器在不產生該元資料的針對該一或多個中介層的程式模組之寫入放大的情況下完成該請求，其中該第一快取模組係在該儲存裝置層外部的一硬體管線模組。

【第12項】如申請專利範圍第11項所述之主裝置，其中該主裝置以所述至少一管線的該第一快取模組快取該元資料，以避免所述至少一管線的複數個後續管線模組產生該元資料的衍生資訊，其中該一或多個中介層的程式模組包含所述至少一管線的該複數個後續管線模組。

【第13項】如申請專利範圍第11項所述之主裝置，其中該主裝置以所述至少一管線的該第一快取模組快取該元資料，以避免該元資料的衍生資訊被產生以及被寫入該儲存裝置層。

【第14項】　如申請專利範圍第11項所述之主裝置，另包含：

至少一非揮發性雙列直插式記憶體模組（non-volatile dual in-line memory module, NVDIMM），其中該第一快取模組位於所述至少一非揮發性雙列直插式記憶體模組中。

【第15項】如申請專利範圍第14項所述之主裝置，其中所述至少一非揮發性雙列直插式記憶體模組中之任一非揮發性雙列直插式記憶體模組包含：

一揮發性（volatile）記憶體，用來為該處理電路儲存資訊；以及

一非揮發性（non-volatile, NV）記憶體，用來為該處理電路保存在該揮發性記憶體中的該資訊；

其中該第一快取模組代表該揮發性記憶體中之一儲存區。

【第16項】如申請專利範圍第11項所述之主裝置，其中：

在以該管線架構處理該物件寫入指令的期間，該主裝置將對應於該使用者資料的一日誌（journal）輸入該管線架構中之另一管線；以及

在以該管線架構處理該物件寫入指令的期間，該主裝置以該另一管線的一第二快取模組快取該日誌，以控制該儲存伺服器在不產生該日誌的針對該一或多個中介層的程式模組之寫入放大的情況下完成該請求，其中該第二快取模組係在該儲存裝置層外部的一硬體管線模組。

【第17項】如申請專利範圍第16項所述之主裝置，另包含：

至少一非揮發性雙列直插式記憶體模組（non-volatile dual in-line memory module, NVDIMM），其中該第一快取模組以及該第二快取模組位於所述至少一非揮發性雙列直插式記憶體模組中。

【第18項】如申請專利範圍第17項所述之主裝置，其中所述至少一非揮發性雙列直插式記憶體模組中之任一非揮發性雙列直插式記憶體模組包含：

一揮發性（volatile）記憶體，用來為該處理電路儲存資訊；以及

一非揮發性（non-volatile, NV）記憶體，用來為該處理電路保存在該揮發性記憶體中的該資訊；

其中該第一快取模組與該第二快取模組中之任一快取模組代表該揮發性記憶體中之一儲存區。

【第19項】如申請專利範圍第11項所述之主裝置，其中該管線架構自運行於該

主裝置之該處理電路上的該複數個程式模組中之一上層起始，穿過該一或多個中介層的程式模組，並且達到該儲存裝置層；以及該主裝置另包含：

至少一非揮發性雙列直插式記憶體模組（non-volatile dual in-line memory module, NVDIMM），用來提供該儲存裝置層外部的至少一硬體管線模組，其中所述至少一硬體管線模組包含該第一快取模組。

【第20項】如申請專利範圍第19項所述之主裝置，其中所述至少一硬體管線模組包含複數個硬體管線模組，其中：

在以該管線架構處理該物件寫入指令的期間，該主裝置將對應於該使用者資料的一日誌（journal）輸入該管線架構中之另一管線；以及

在以該管線架構處理該物件寫入指令的期間，該主裝置以該另一管線的一第二快取模組快取該日誌，以控制該儲存伺服器在不產生該日誌的針對該一或多個中介層的程式模組之寫入放大的情況下完成該請求，其中該複數個硬體管線模組包含該第一快取模組以及該第二快取模組。

【第21項】如申請專利範圍第11項所述之主裝置，另包含：

一機殼（case），用來設置該主裝置的多個元件以及所述至少一儲存裝置，其中該主裝置的該多個元件包含該處理電路。

【第22項】一種儲存伺服器，包含：

一主裝置，用來控制該儲存伺服器的運作，該主裝置包含：

一處理電路，用來控制該主裝置在該儲存伺服器中進行基於管線的（pipeline-based）存取管理；以及

一儲存裝置層，包含：

至少一儲存裝置，耦接至該主裝置，用來為該儲存伺服器儲存資訊；

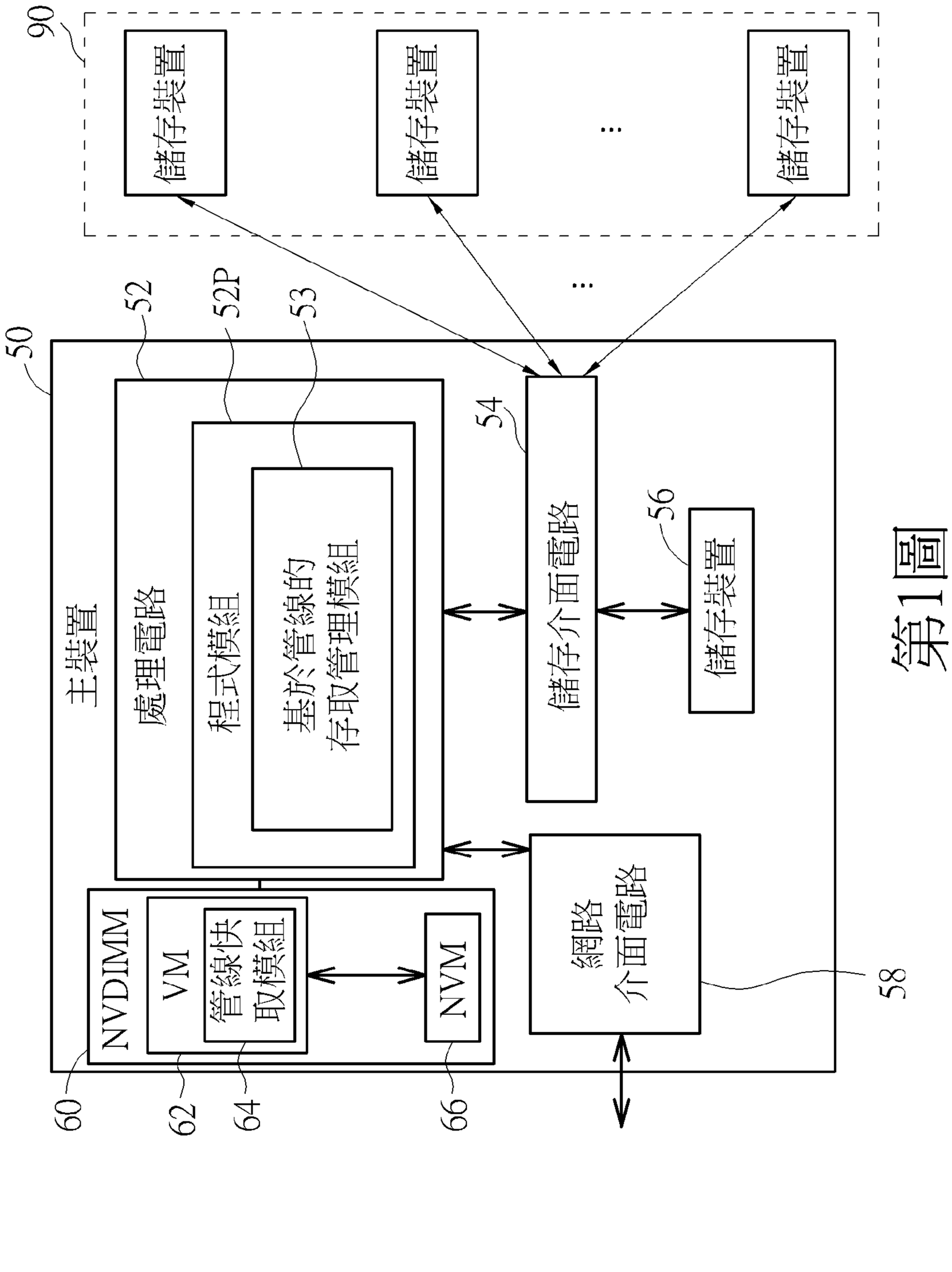
其中：

因應將使用者資料寫入該儲存伺服器的一請求，該儲存伺服器中之該主裝置將該使用者資料寫入該儲存伺服器之該儲存裝置層並且以該儲存伺服器的一管線（pipeline）架構開始處理對應於寫入該使用者資料的該請求的一物件寫入指令；

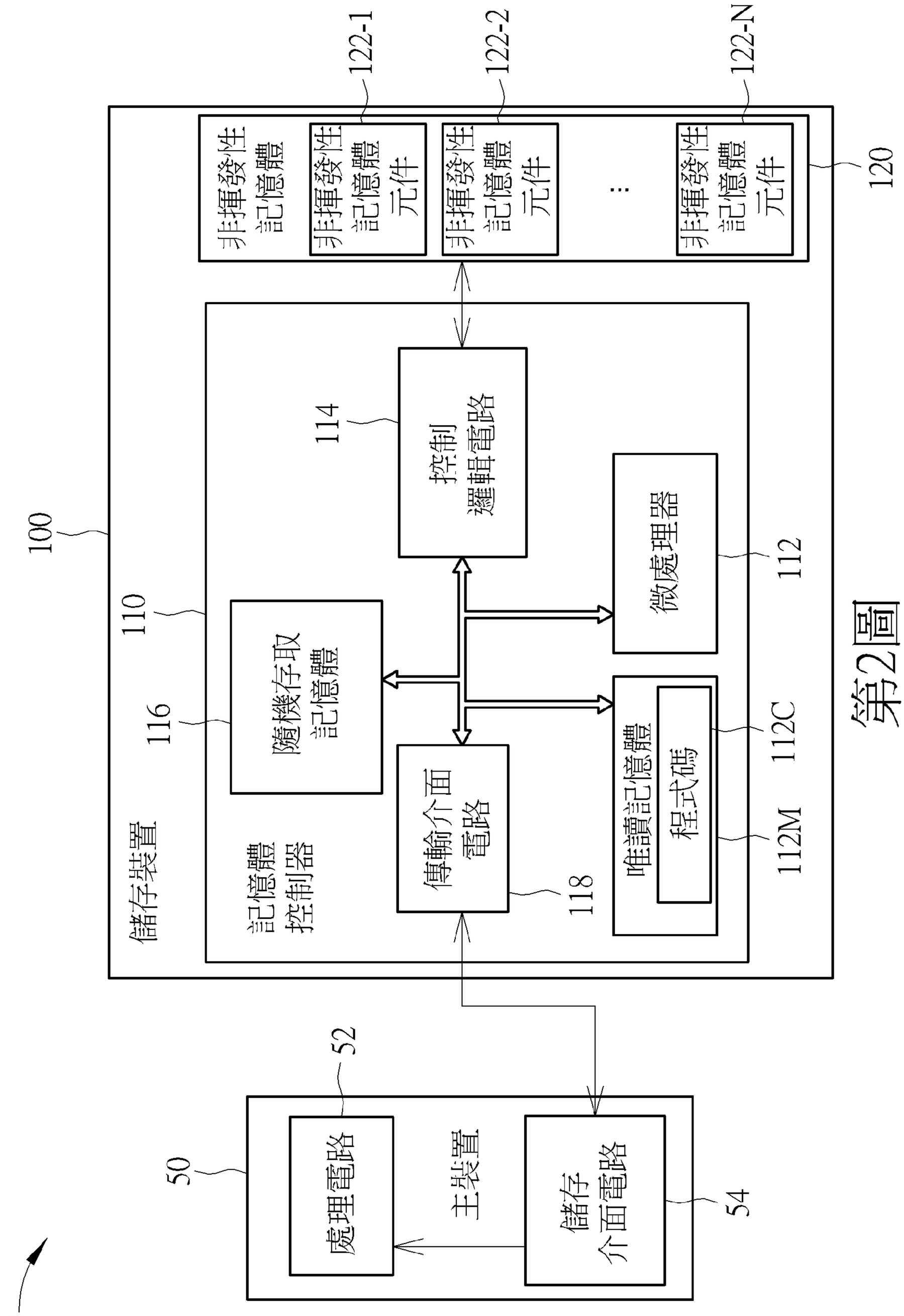
在以該管線架構處理該物件寫入指令的期間，該主裝置將對應於該使用者資料的元資料（metadata）輸入該管線架構中之至少一管線，其中該管線架構穿過運行於該主裝置之該處理電路上的複數個程式模組中之一或多個中介層（intermediate layer）的程式模組，並且達到該儲存裝置層；以及

在以該管線架構處理該物件寫入指令的期間，該主裝置以所述至少一管線的一第一快取（cache）模組快取該元資料，以控制該儲存伺服器在不產生該元資料的針對該一或多個中介層的程式模組寫入放大的情況下完成該請求，其中該第一快取模組係在該儲存裝置層外部的一硬體管線模組。

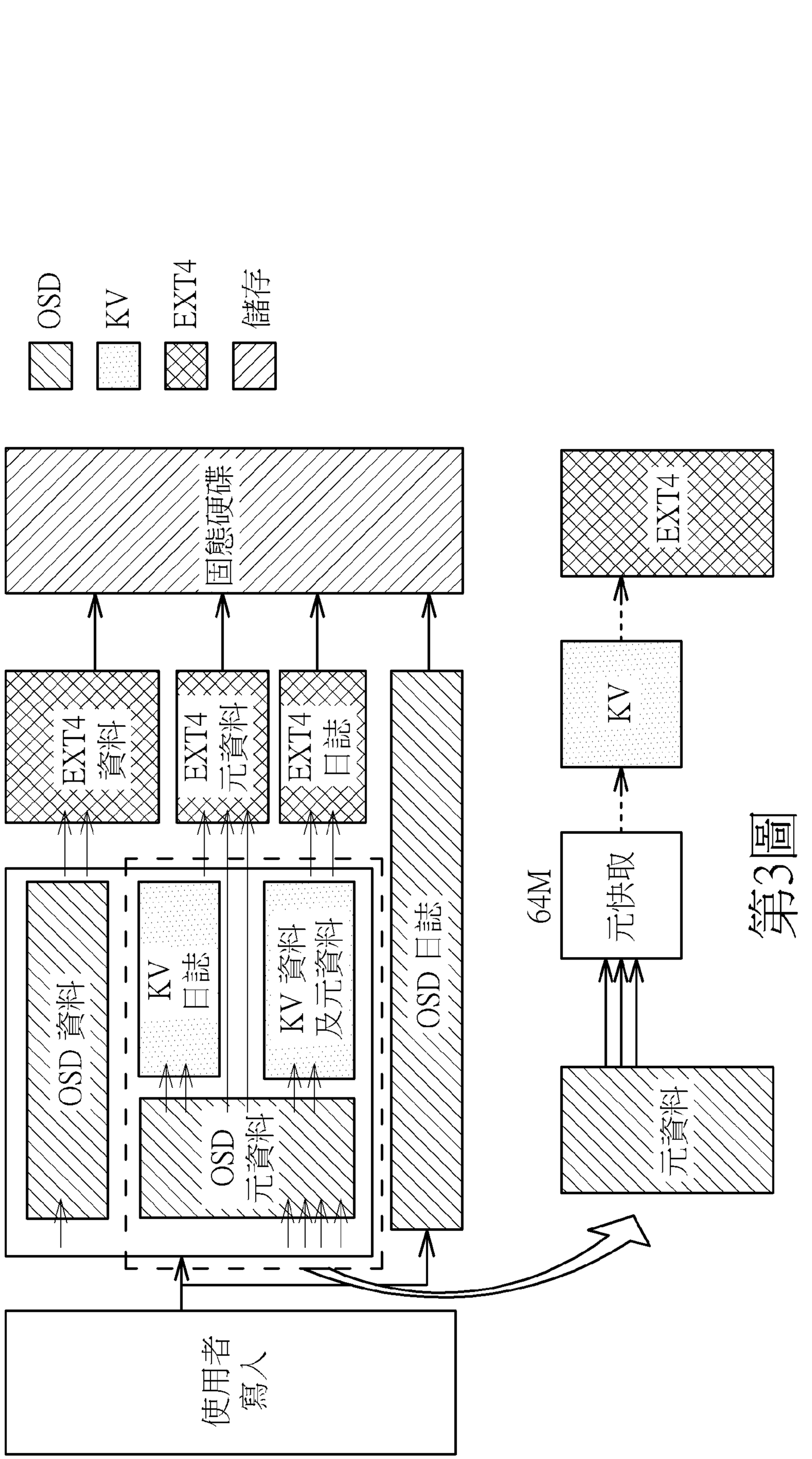
【發明圖式】


90
儲存裝置
儲存裝置
...
儲存裝置
...
50
52
52P
53
54
主裝置
處理電路
程式模組
基於管線的
存取管理模組
儲存介面電路
56
儲存裝置
NVDIMM
VM
管線快
取模組
NVM
網路
介面電路
58
60
62
64
66
10

第1圖


10
50
處理電路
主裝置
儲存
介面電路
54
52
儲存裝置
100
110
記憶體
控制器
116
隨機存取
記憶體
118
傳輸介面
電路
114
控制
邏輯電路
112
微處理器
112C
唯讀記憶體
程式碼
112M
120
非揮發性
記憶體
元件
122-1
122-2
122-N

第2圖


OSD
KV
EXT4
儲存
固態硬碟
EXT4 資料
EXT4 元資料
EXT4 日誌
OSD 資料
KV 日誌
KV 資料及元資料
OSD 元資料
OSD 日誌
使用者寫入
EXT4
KV
64M
元快取
元資料

第3圖

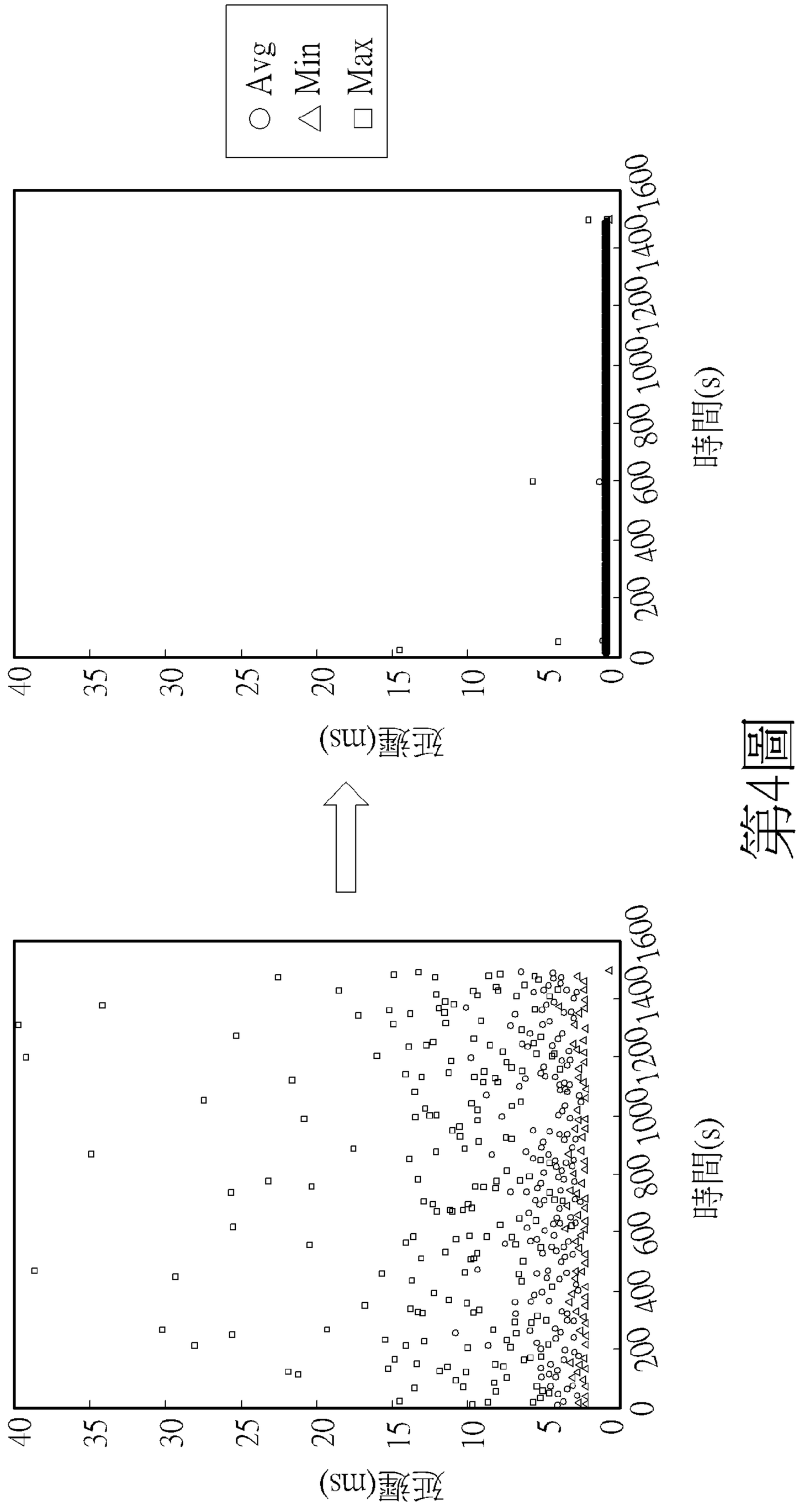
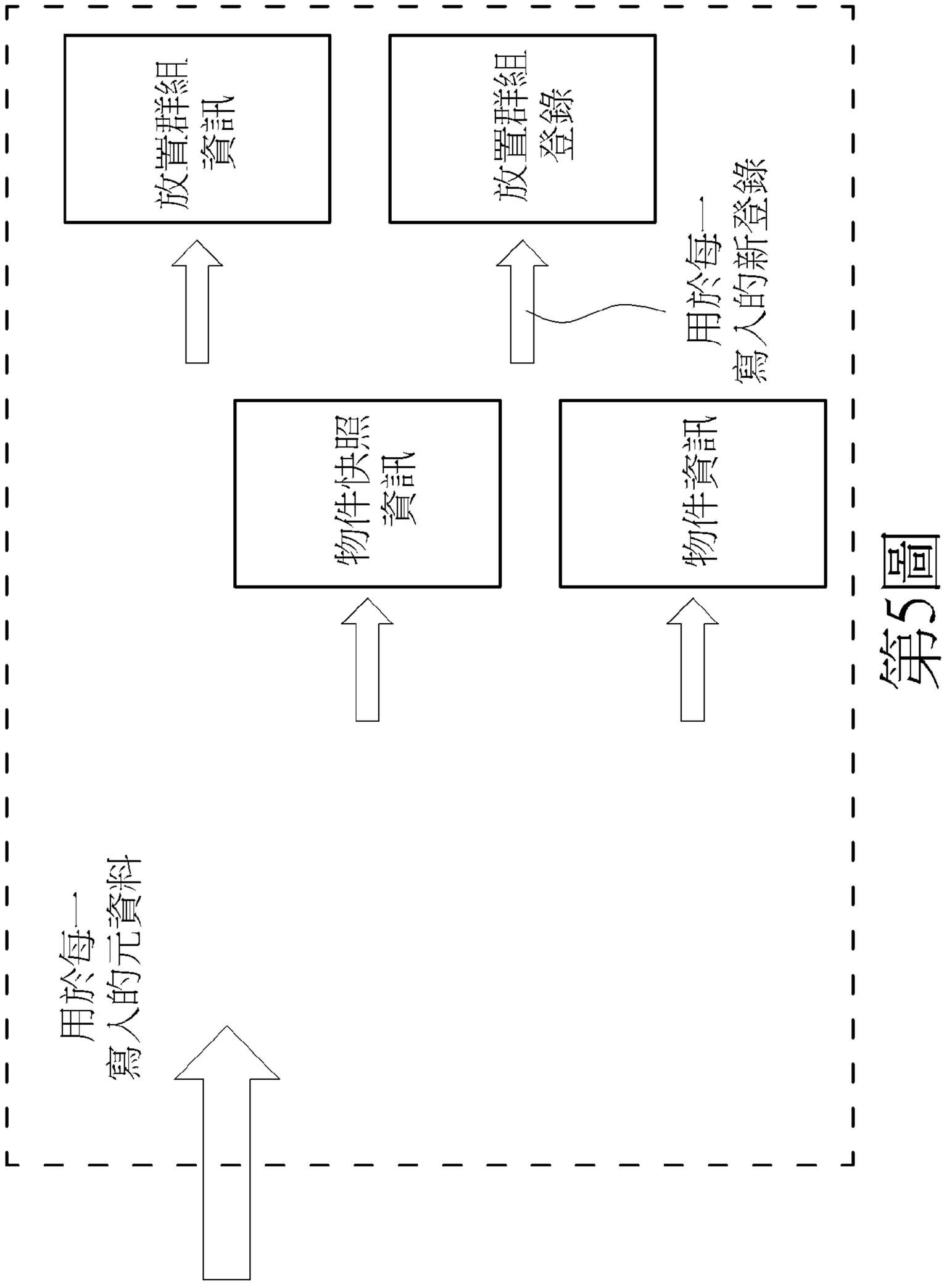
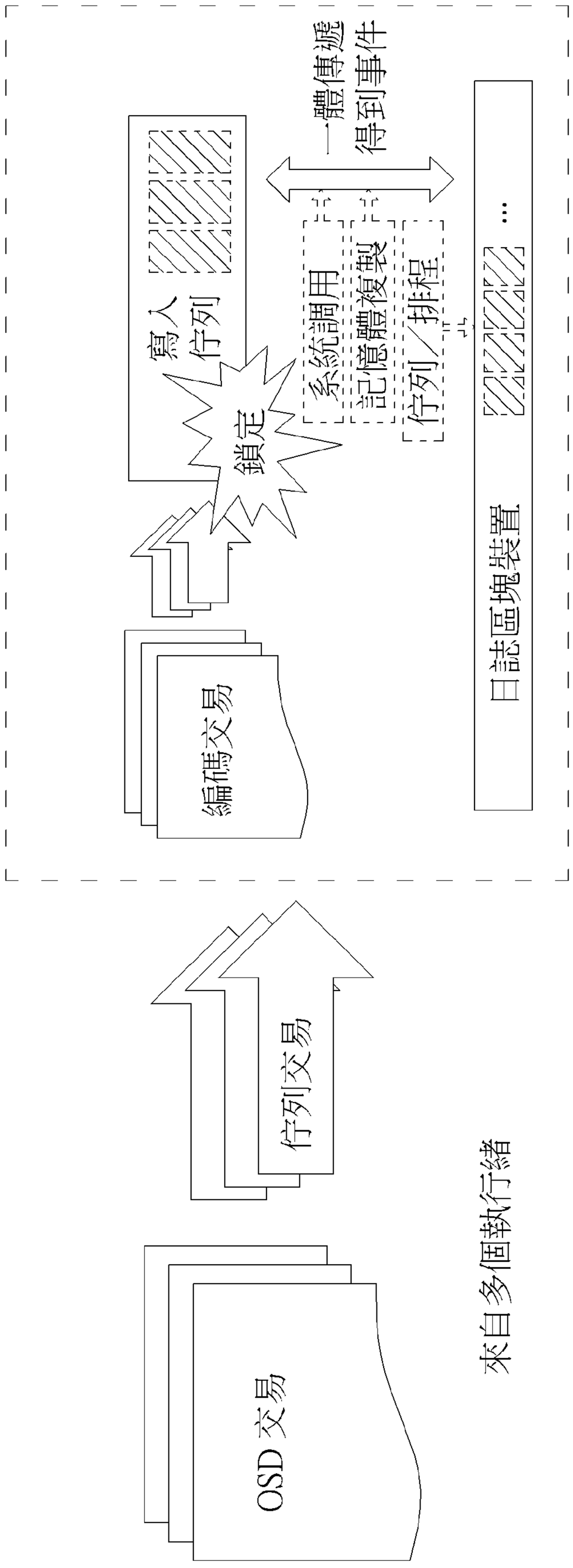

○ Avg
△ Min
□ Max
延遲(ms)
時間(s)
0
5
10
15
20
25
30
35
40
0
200
400
600
800
1000
1200
1400
1600

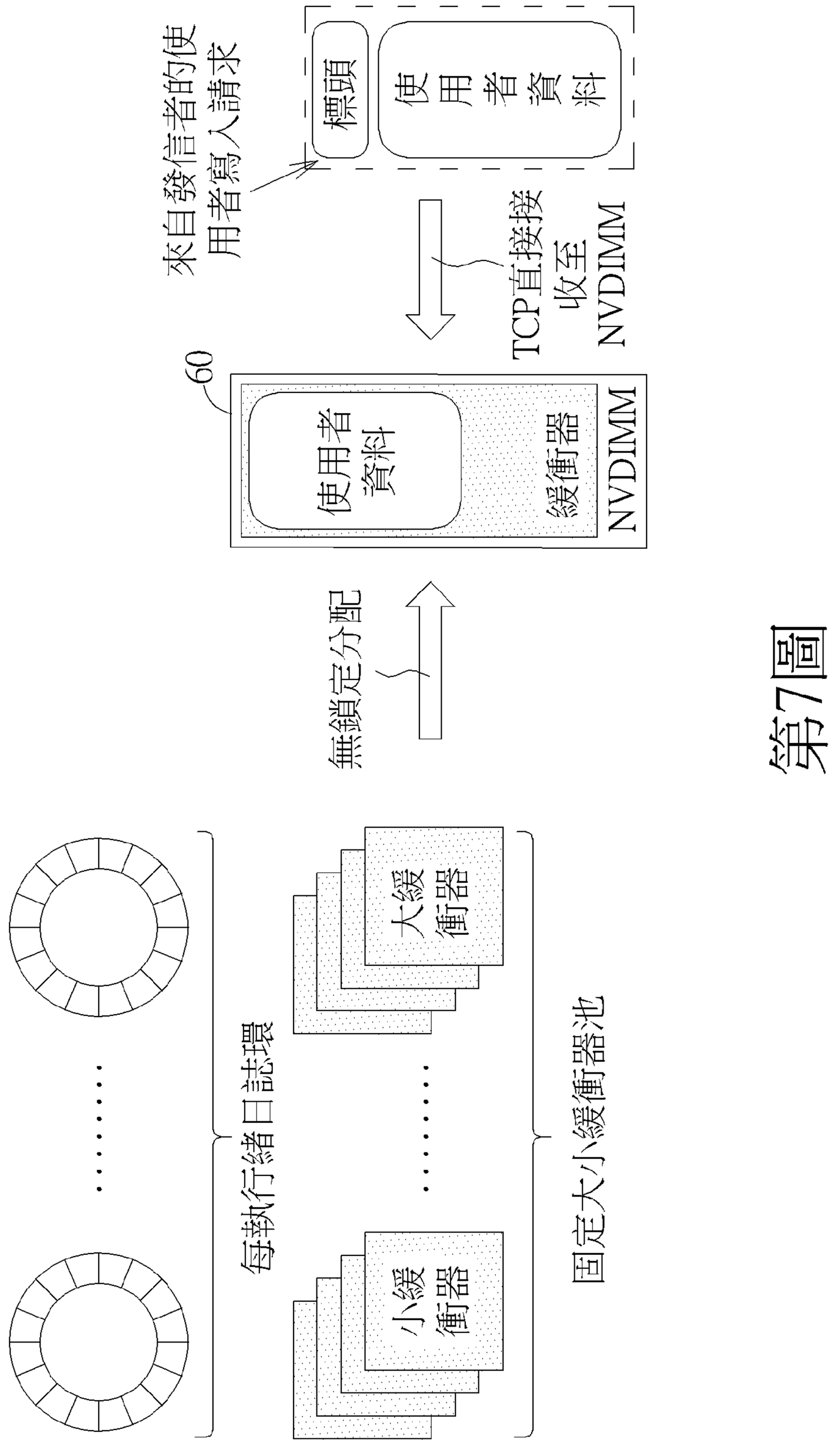
圖4



第5圖


寫入佇列
鎖定
系統調用
記憶體複製
佇列／排程
日誌區塊裝置
...
一體傳遞得到事件
編碼交易
佇列交易
OSD 交易
來自多個執行緒

第6圖


來自發信者的使
用者寫入請求
標頭
使用者資料
TCP直接接
收至
NVDIMM
60
使用者
資料
緩衝器
NVDIMM
無鎖定分配
每執行緒日誌環
大緩
衝器
小緩
衝器
固定大小緩衝器池

第7圖

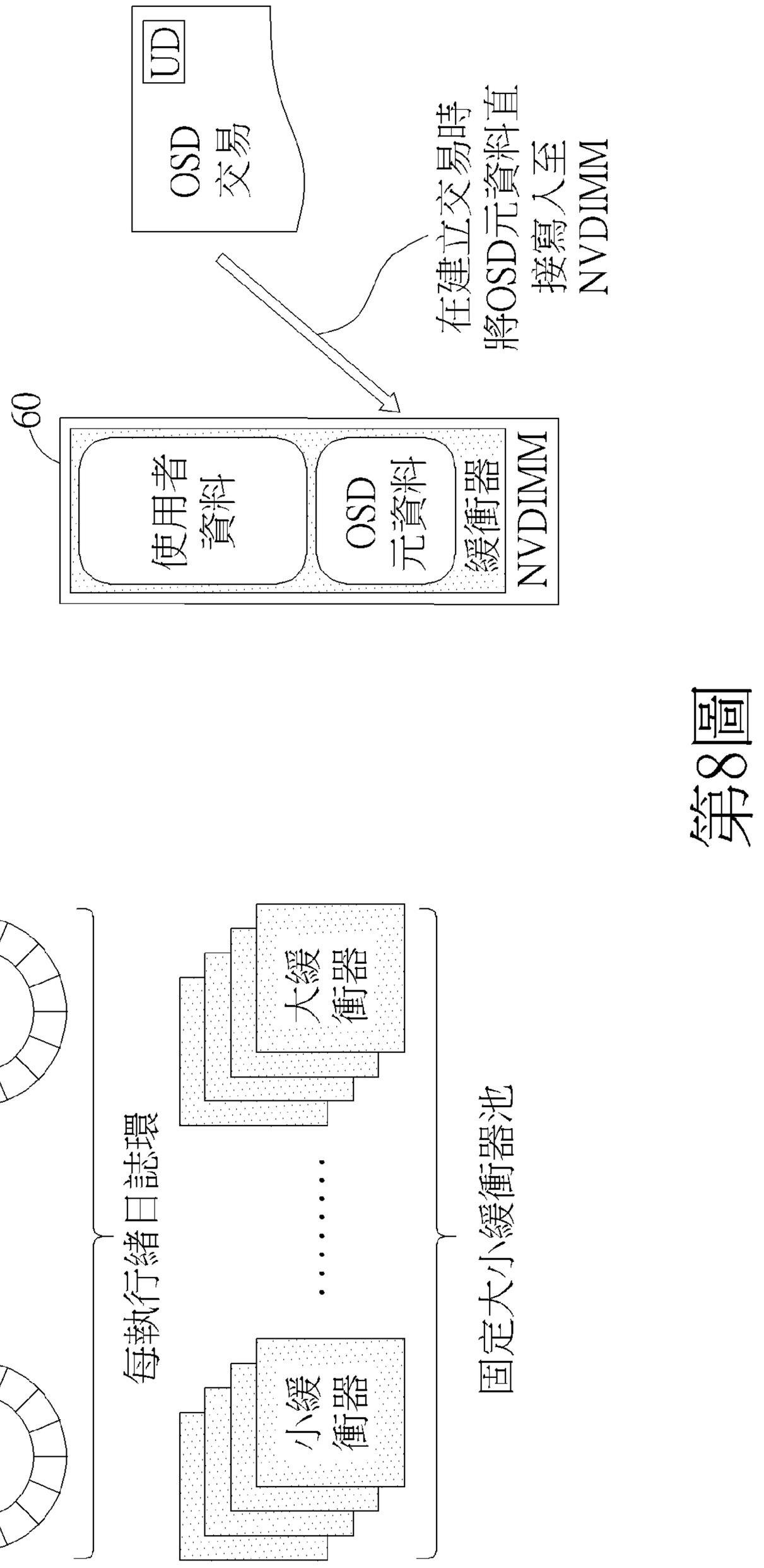
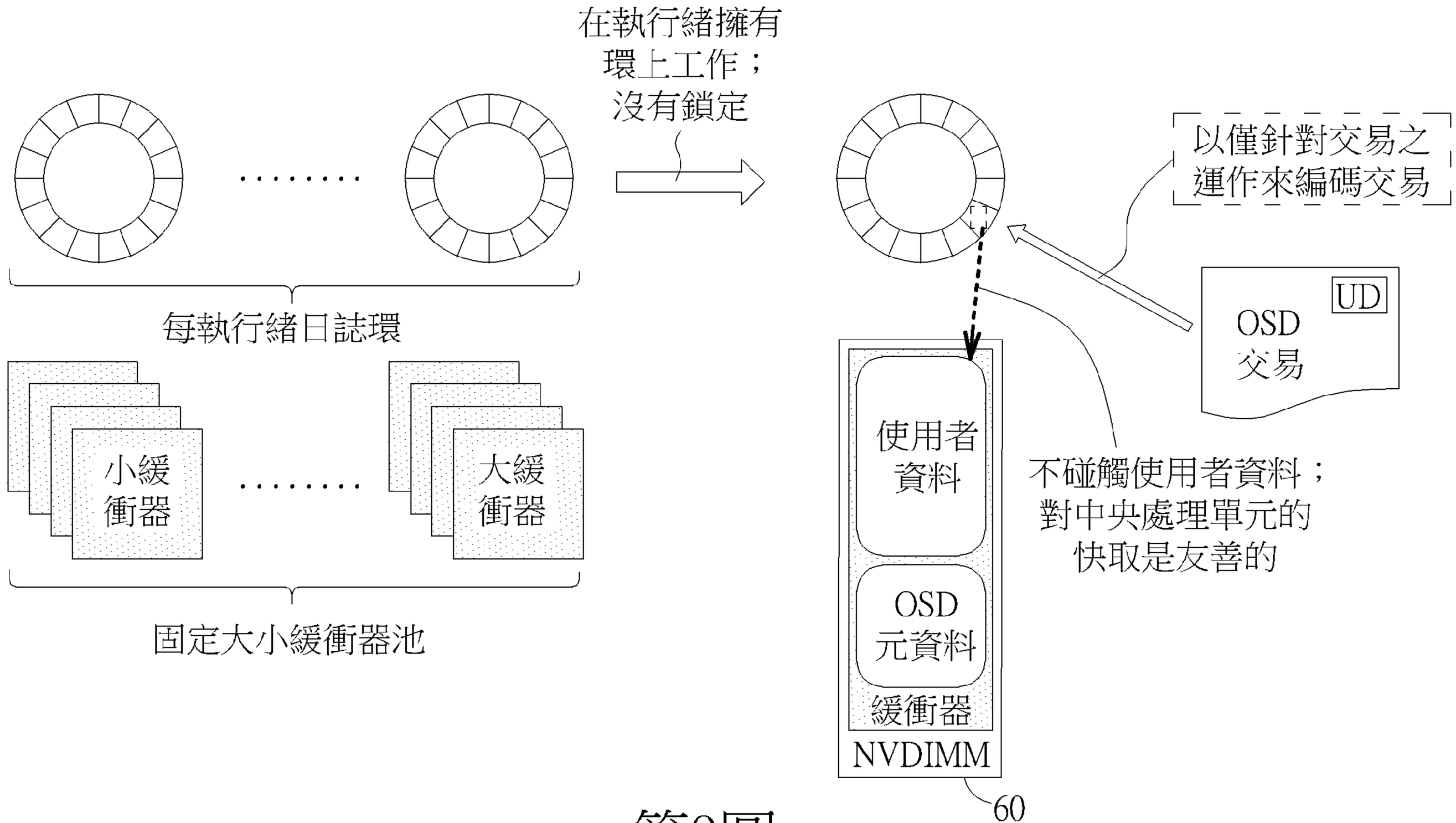
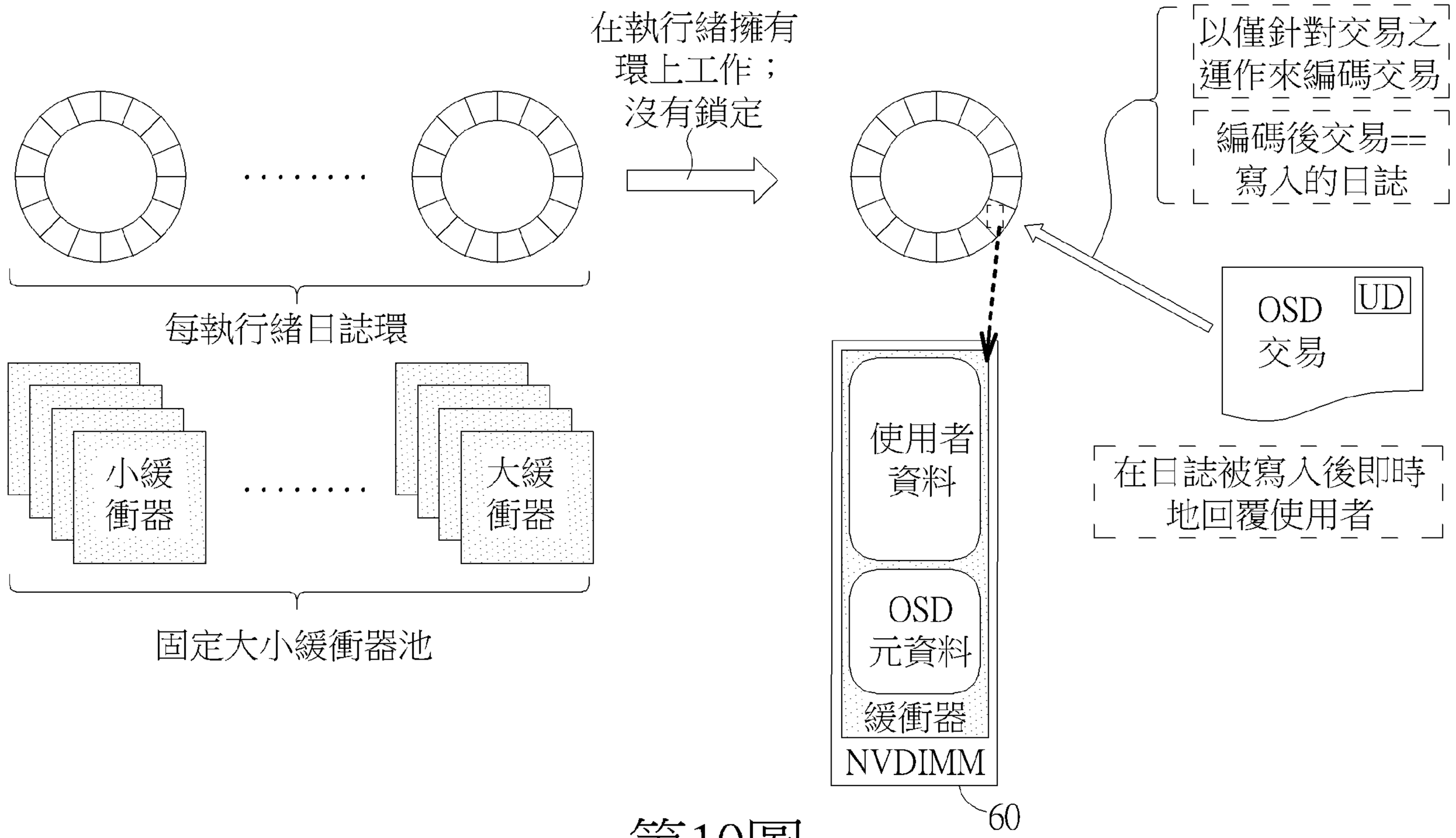

UD
OSD
交易
在建立交易時
將OSD元資料直
接寫入至
NVDIMM
60
使用者
資料
OSD
元資料
緩衝器
NVDIMM
每執行結束日誌環
大緩
衝器
小緩
衝器
固定大小緩衝器池

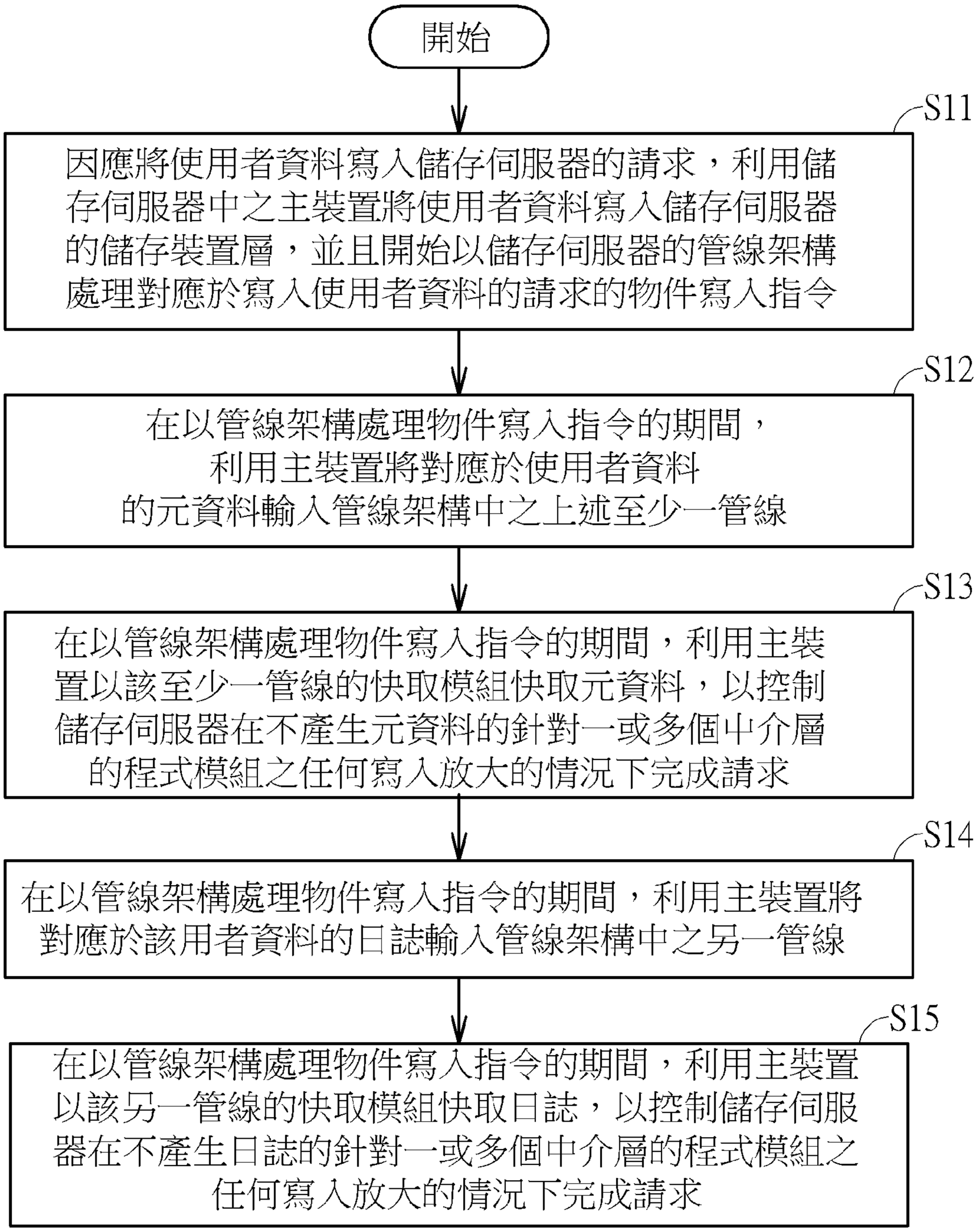
圖8


在執行緒擁有
環上工作；
沒有鎖定
每執行緒日誌環
小緩
衝器
大緩
衝器
固定大小緩衝器池
以僅針對交易之
運作來編碼交易
UD
OSD
交易
使用者
資料
OSD
元資料
緩衝器
NVDIMM
60
不碰觸使用者資料；
對中央處理單元的
快取是友善的

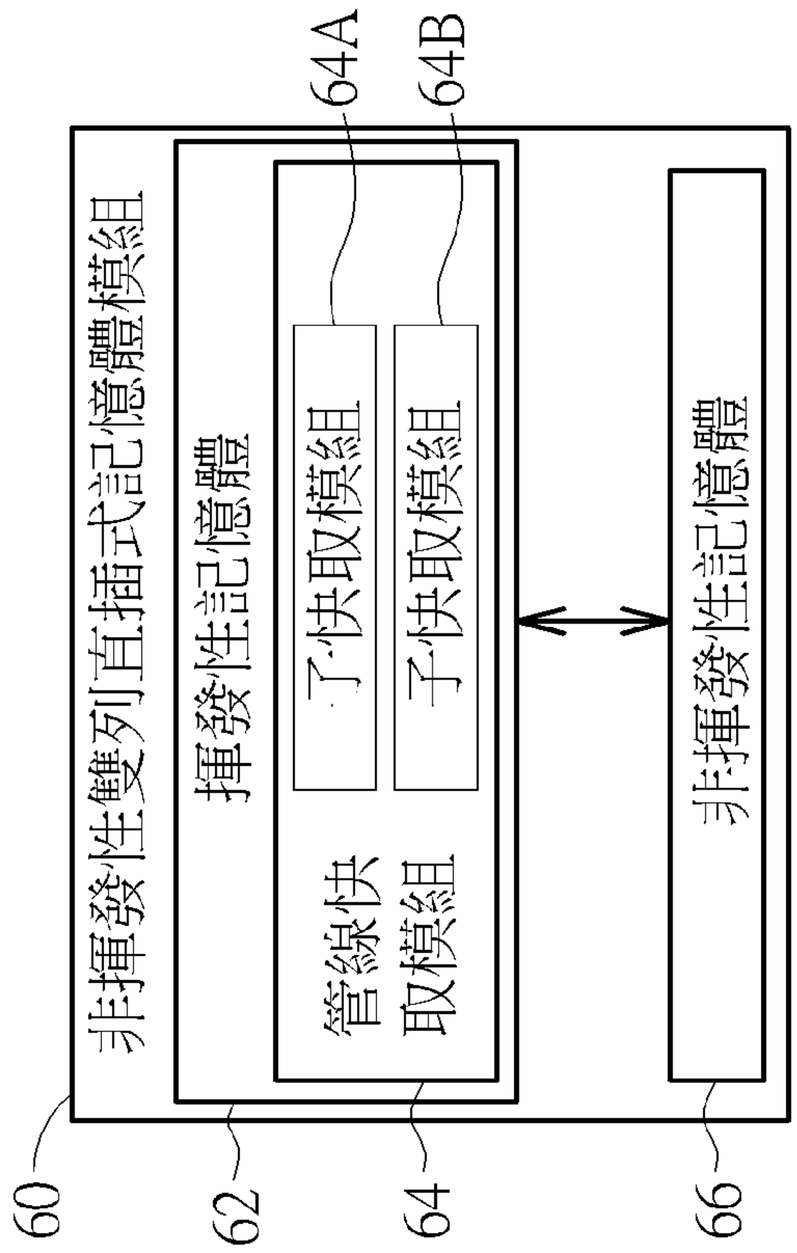
第9圖


在執行緒擁有
環上工作；
沒有鎖定
以僅針對交易之
運作來編碼交易
編碼後交易==
寫入的日誌
每執行緒日誌環
OSD
交易
UD
在日誌被寫入後即時
地回覆使用者
小緩
衝器
大緩
衝器
固定大小緩衝器池
使用者
資料
OSD
元資料
緩衝器
NVDIMM
60

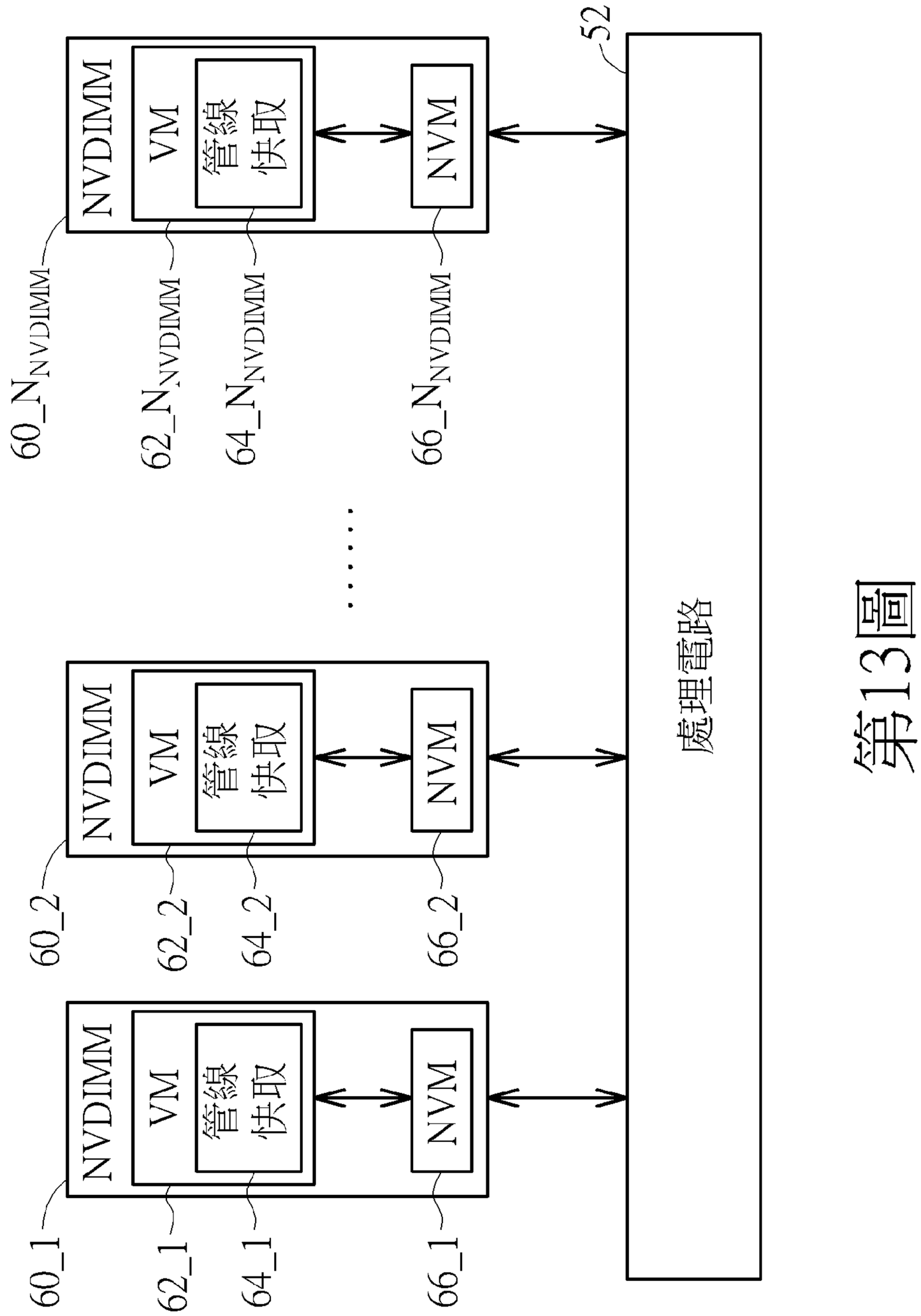
第10圖


開始
S11
因應將使用者資料寫入儲存伺服器的請求，利用儲存伺服器中之主裝置將使用者資料寫入儲存伺服器的儲存裝置層，並且開始以儲存伺服器的管線架構處理對應於寫入使用者資料的請求的物件寫入指令
S12
在以管線架構處理物件寫入指令的期間，利用主裝置將對應於使用者資料的元資料輸入管線架構中之上述至少一管線
S13
在以管線架構處理物件寫入指令的期間，利用主裝置以該至少一管線的快取模組快取元資料，以控制儲存伺服器在不產生元資料的針對一或多個中介層的程式模組之任何寫入放大的情況下完成請求
S14
在以管線架構處理物件寫入指令的期間，利用主裝置將對應於該用者資料的日誌輸入管線架構中之另一管線
S15
在以管線架構處理物件寫入指令的期間，利用主裝置以該另一管線的快取模組快取日誌，以控制儲存伺服器在不產生日誌的針對一或多個中介層的程式模組之任何寫入放大的情況下完成請求

第11圖


非揮發性雙列直插式記憶體模組
揮發性記憶體
64A
64B
子快取模組
子快取模組
管線快取模組
非揮發性記憶體
60
62
64
66

第12圖


NVDIMM
VM
管線供取
NVM
60_N_NVDIMM
62_N_NVDIMM
64_N_NVDIMM
66_N_NVDIMM
60_2
62_2
64_2
66_2
60_1
62_1
64_1
66_1
處理電路

第13圖