I620072

(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 I620072 B**

(45)公告日： 中華民國 107 (2018) 年 04 月 01 日

(21)申請案號：105124918 (22)申請日： 中華民國 105 (2016) 年 08 月 05 日

(51)Int. Cl.： *G06F13/40 (2006.01)* *G06F13/42 (2006.01)*
G06F3/06 (2006.01)

(30)優先權：2016/03/07 美國 62/304,483
2016/06/28 美國 15/195,261

(71)申請人：廣達電腦股份有限公司 (中華民國) QUANTA COMPUTER INC. (TW)
桃園市龜山區文化二路 188 號

(72)發明人：施青志 SHIH, CHING CHIH (TW)

(74)代理人：洪澄文；顏錦順

(56)參考文獻：

TW	201443647A	EP	1705574A2
US	8,234,424B2	US	20150254088A1
US	2015/0039804A1		

審查人員：林文琦

申請專利範圍項數：10 項 圖式數：5 共 50 頁

(54)名稱

可擴充集中式非揮發性記憶體儲存盒、電腦實施方法以及非暫態電腦可讀取儲存裝置

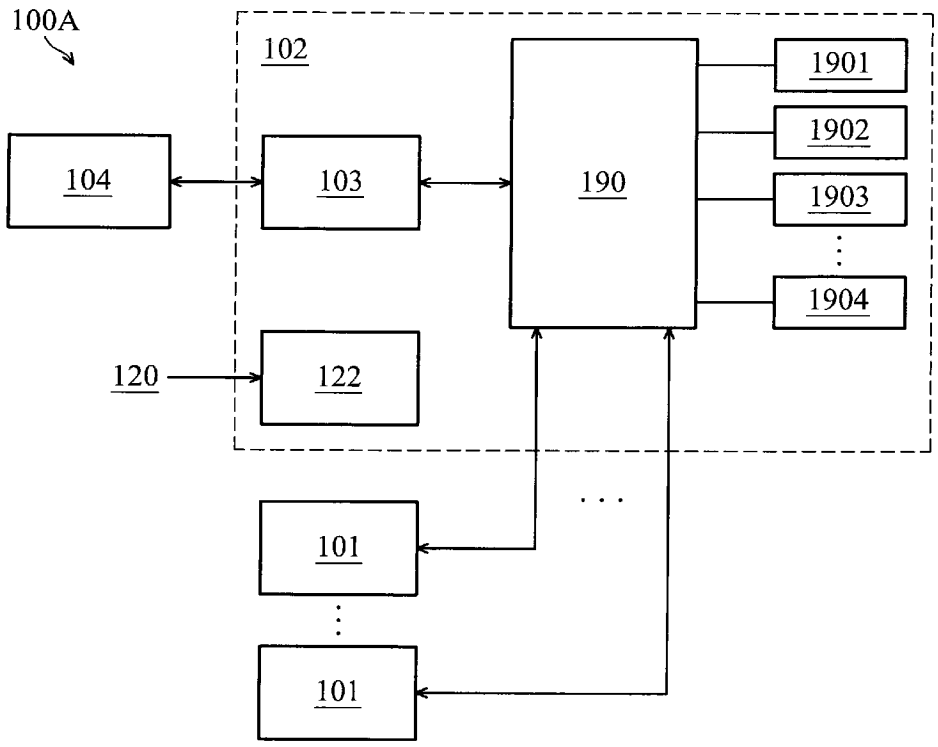
SCALABLE POOLED-NON-VOLATILE MEMORY EXPRESS STORAGE BOXES, COMPUTER-IMPLEMENTED METHODS AND NON-TRANSITORY COMPUTER-READABLE STORAGE MEDIUM

(57)摘要

本發明提供一種支援可共用於複數節點之間於可擴充集中式非揮發性記憶體(NVMe)儲存盒之叢集架構。可擴充集中式 NVMe 儲存盒包括 NVMe 硬碟、一個或者多個交換器以及一個或者多個交換器埠。可擴充集中式 NVMe 儲存盒中 NVMe 硬碟之數量可根據需要共用可擴充集中式 NVMe 儲存盒之節點之數量增加或者減少。

The present invention provides a cluster-architecture to support a scalable pooled-NVMe storage box that can be shared among a scalable number of nodes. The scalable pooled-NVMe storage box comprises NVMe drives, one or more switches and one or more switch ports. The number of NVMe drives in the scalable scalable-pooled-NVMe storage box can be scaled up or down based upon a number of nodes that need to share the scalable pooled-NVMe storage box.

指定代表圖：



第 1A 圖

符號簡單說明：

100A・・・示範系統

101・・・伺服器系統

102・・・儲存子系統

103・・・結構控制器

104・・・硬體編制管理器

120・・・電源供應器

122・・・電源供應單元

190・・・PCIe 交換器

1901~1904・・・儲存裝置

發明專利說明書

（本說明書格式、順序，請勿任何更動）

【發明名稱】 可擴充集中式非揮發性記憶體儲存盒、電腦實施方法以及非暫態電腦可讀取儲存裝置

SCALABLE POOLED-NON-VOLATILE MEMORY EXPRESS STORAGE BOXES, COMPUTER-IMPLEMENTED METHODS AND NON-TRANSITORY COMPUTER-READABLE STORAGE MEDIUM

【技術領域】

【0001】 本發明係有關於電信網路中之儲存系統。

【先前技術】

【0002】 SATA（Serial At Attachment）硬碟之容量以及進階主機控制器介面（Advanced Host Controller Interface, AHCIs）已經超越處理器之發展速度。同時，資料量的增加亦需要更有效率之儲存設計。為了滿足企業以及個人用戶之資料儲存以及處理速度之需求，係開發非揮發性記憶體（Non-Volatile Memory Express, 以下簡稱NVMe）以用於資料之各種應用上。一些NVMe固態硬碟可集中於NVMe儲存盒中以改善資料中心之效率（例如NVMe儲存盒之固態硬碟之間可共用電源以及冷卻單元）。

【0003】 然而，如何改善NVMe儲存盒之可擴充性仍為一項挑戰。

【發明內容】

【0004】 根據本發明各種實施例所述之系統以及方法係透過提供具有可共用於複數節點之間之可擴充集中式非揮發記憶體儲存盒之一叢集架構以解決前述機架系統中之問題。可擴充集中式NVMe儲存盒包括NVMe硬碟、一個或者多個交換器、以及一個或者多個交換器埠。可擴充集中式NVMe儲存盒中之NVMe硬碟之數量可根據共用可擴充集中式NVMe儲存盒之節點之數量增加或者減少。

【0005】 於傳統之儲存系統中，SATA硬碟之頻寬最大可達600Gbps。本發明係提供一種集中式NVMe儲存盒之可擴充叢集架構，可擴充叢集架構包括可透過PCIe（PCI Express）存取之NVMe硬碟（例如NVMe固態硬碟）。NVMe硬碟可提供大於1.5GBps之容量。舉例來說，第二代之PCIe之每個通道可提供約500MBps之頻寬以及第三代之PCIe之每個通道可提供約985MBps之頻寬。四個第二代PCIe之插槽可提供2GBps之頻寬以及四個第三代PCIe之插槽則可提供接近4GBps之頻寬。

【0006】 本發明一些實施例提供一種可擴充叢集架構，用以提供包括具有雙埠之NVMe硬碟之可擴充集中式NVMe儲存盒。可擴充集中式NVMe儲存盒中之NVMe硬碟之數量可根據共用可擴充集中式NVMe儲存盒之節點之數量增加或者減少。舉例來說，可擴充集中式NVMe儲存盒可包括一第一複數NVMe硬碟、一第一交換器、一第一交換器埠、以及一第一裝置埠。第一複數NVMe硬碟之每一者具有一第一埠以及一第二埠。第一複數NVMe硬碟可透過對應於NVMe硬碟之第一埠以及第一

交換器連接至第一複數節點。於一些實施例中，第一交換器埠係與第一裝置埠連接。第一複數NVMe硬碟亦可透過對應於NVMe硬碟之第二埠、第一裝置埠、第一交換器埠、以及第一交換器連接至第一複數節點。

【0007】 可擴充集中式NVMe儲存盒可透過利用一小參數加倍第一複數NVMe硬碟之數量以支援兩倍數量之第一複數節點（即第一複數節點以及第二複數節點）。舉例來說，透過更包括第二複數NVMe硬碟、第二交換器、第二交換器埠、以及第二裝置埠以擴充可擴充集中式NVMe儲存盒。第二複數NVMe硬碟之每一者具有一第一埠以及一第二埠。第二複數節點可透過第二交換器以及對應之NVMe硬碟之第一埠連接至第二複數NVMe硬碟。於一些實施例中，第二裝置埠亦可透過第二交換器埠連接至第二交換器。第二複數節點亦可透過第二交換器、第二交換器埠、第二裝置埠、以及對應之NVMe硬碟之第二埠連接至第二複數NVMe硬碟。可擴充集中式NVMe儲存盒中之一特定NVMe硬碟可被第一複數節點以及第二複數節點之任何節點存取。換言之，可擴充集中式NVMe儲存盒中之NVMe硬碟可共用於第一複數節點以及第二複數節點之間。

【0008】 於一些實施例中，可擴充集中式NVMe儲存盒中之裝置埠以及交換器埠可結合為單一介面埠以提供裝置埠以及交換器埠兩者之功能。舉例來說，一節點可透過特定NVMe硬碟以及單一介面埠之複數埠一者存取可擴充集中式NVMe儲存盒中特定之NVMe硬碟。於一些實施例中，可擴充集中式NVMe儲存盒中兩個或者多個NVMe硬碟可動態地配置至兩個或者多

個節點以產生一高可用度叢集系統（High-Availability cluster, HA cluster）。高可用度叢集系統可以最短停機時間提供伺服器程序。

【0009】 一些實施例提供一種可擴充叢集架構以支援包括NVMe硬碟之可擴充集中式NVMe儲存盒，NVMe硬碟係具有高頻寬單一埠。可擴充集中式NVMe儲存盒中NVMe硬碟之數量可根據需要共用可擴充集中式NVMe儲存盒之節點數量增加以及減少。可擴充集中式NVMe儲存盒可包括第一複數NVMe硬碟、第一交換器、以及第一介面埠。第一交換器係與第一介面埠連接。第一複數NVMe硬碟之每一者具有一單一埠。第一複數NVMe硬碟可透過對應之NVMe硬碟之單一埠以及第一交換器連接至第一複數節點。

【0010】 可擴充集中式NVMe儲存盒可藉由加倍具有單一埠之第一複數NVMe硬碟之數量以支援兩倍之第一複數節點（即第一複數節點以及第二複數節點）。舉例來說，透過更包括第二複數NVMe硬碟、第二交換器、以及第二介面埠以擴充可擴充集中式NVMe儲存盒。第二交換器係與第二介面埠連接。第二複數NVMe硬碟之每一者包括單一埠。第二複數節點係透過第二交換器以及對應之NVMe硬碟之單一埠連接至第二複數NVMe硬碟。第一介面埠係與第二介面埠連接。第一複數節點以及第二複數節點可存取可擴充集中式NVMe儲存盒中任何特定之硬碟。舉例來說，可透過第一交換器或者第二交換器、或者透過一第一連接路徑（第二交換器-第二介面埠-第一介面埠）或者第二連接路徑（第一交換器-第一介面埠-第二介

面埠）直接存取可擴充集中式NVMe儲存盒中之一特定NVMe硬碟。

【圖式簡單說明】

【0011】 為了描述前述之方法以及本發明之其它優點和特徵，有關前述之原理更具體之描述可藉由參照後附圖式中所示之具體實施例。必須理解的是，這些圖式僅用以描繪有關本發明之實施例，並非用以對本發明之範圍有所限制，本發明之原理係藉由圖式以及其附加特徵以及細節說明以及解釋，其中：

第1A圖係顯示根據本發明一些實施例所述之包括一儲存子系統以及一伺服器系統之示範系統之示意方塊圖。

第1B~1F圖係顯示根據本發明一些實施例所述之示範集中式NVMe儲存盒之示意方塊圖。

第2圖係顯示根據本發明一些實施例所述之支援可擴充集中式NVMe儲存盒之示範方法。

第3圖係顯示根據本發明各個實施例所述之示範計算裝置。

第4、5圖係顯示根據本發明各個實施例所述之示範系統。

【實施方式】

【0012】 本發明各個實施例係提供支援可擴充集中式NVMe儲存盒之叢集架構，其中可擴充集中式NVMe儲存盒可共用於可擴充節點之間。可擴充集中式NVMe儲存盒包括NVMe硬碟、至少一交換器以及至少一交換器埠。可擴充集中式NVMe儲存盒中NVMe硬碟之數量可根據需要共用可擴充集中式

Windows、Window Phone、以及IBM z/OS，但並不以此為限。

【0027】 根據示範系統100A~100F所期望之實施，係可使用各種網路以及訊息協定，包括智慧平台管理介面（Intelligent Platform Management Interface, IPMI）或者Redfish等，但並不以此為限。其它合適之協定亦可用於實施本發明。本領域技術人員皆可理解，第1A~1F圖所示之示範系統100A~100F僅用以作為說明之目的。因此，在仍提供本發明各種實施例所述之網路平台配置之情況下，可對本發明之網路系統進行各種適當之改變。

【0028】 第1A~1F圖所示之示範配置中，儲存子系統102亦可包括用以與特定無線頻道之計算範圍中之一個或者多個電子裝置進行通訊之一個或者多個無線元件。無線通道可為任何使裝置可進行無線通訊之合適通道，例如藍芽、蜂巢式系統、NFC或者Wi-Fi通道。必須理解的是，裝置可具有一個或著多個本領域公知的傳統有線通訊連接。各種其它元件和/或組合亦可包括於本發明實施例之範圍內。

【0029】 前述之內容主要用以說明本發明之原理以及各種實施例。一旦完全理解本發明之內容，各種改變以及修正為顯而易見的。

【0030】 第2圖係顯示根據本發明一些實施例所述之支援可擴充集中式NVMe儲存盒之示範方法200。於此實施例中，可擴充集中式NVMe儲存盒包括第一複數NVMe硬碟。如示範系統100A~100F所示，第一複數NVMe硬碟之每一者包括一第一埠以及一第二埠。必須理解的是，示範方法200僅用以作為說明

之目的，以及根據本發明之技術所實施之其它方法中可包括以類似或者替代之順序執行或者平行處理之額外的、更少的、或者替代的步驟。

【0031】 示範方法200起始於步驟202，將第一複數NVMe硬碟之第一埠連接至可擴充集中式NVMe儲存盒之第一交換器。於步驟204，將第一複數NVMe硬碟之第二埠透過可擴充集中式NVMe儲存盒之第一裝置埠以及一第一交換器埠連接至第一交換器。於步驟206，第一複數NVMe硬碟可透過第一交換器連接至第一複數節點。如第1B~1F圖所示，第一複數節點之間透過對應之NVMe硬碟之第一埠或者第二埠共用第一複數NVMe硬碟之每一者。於一些實施例中，第一裝置埠以及第一交換器埠可結合為第一介面埠以提供第一裝置埠以及第一交換器埠兩者之功能。

【0032】 於步驟208，根據需要共用可擴充集中式NVMe儲存盒之節點之數量判斷是否增加可擴充集中式NVMe儲存盒之NVMe硬碟之數量。於步驟210，若判斷不需要增加可擴充集中式NVMe儲存盒之NVMe硬碟之數量，則示範方法200結束。

【0033】 於步驟212，若判斷需要增加可擴充集中式NVMe儲存盒之NVMe硬碟之數量，則示範方法200更將第二複數NVMe硬碟之第一埠連接至第二交換器。於步驟214，將第二複數NVMe硬碟之第二埠透過第二裝置埠以及第二交換器埠連接至第二交換器。第二複數NVMe硬碟之每一者包括一第一埠以及一第二埠。於步驟216，第二複數NVMe硬碟可透過第二交換器連接至第二複數節點。第二複數節點之間可透過對應之

NVMe硬碟之第一埠或者第二埠共用第二複數NVMe硬碟之每一者。

【0034】 於步驟218，第一裝置埠係與第二交換器埠連接。如第1E~1F圖所示，於步驟220，第一交換器埠係與第二裝置埠連接。可擴充集中式NVMe儲存盒包括擴充之NVMe硬碟，擴充之NVMe硬碟包括第一複數NVMe硬碟以及第二複數NVMe硬碟。可擴充集中式NVMe儲存盒之NVMe硬碟之每一者可被共用於第一複數節點以及第二複數節點之間。可擴充集中式NVMe儲存盒之一特定NVMe硬碟可直接透過第一交換器或者第二交換器、或者透過一連接路徑被存取。連接路徑可包括第一交換器-第一交換器埠-第一裝置埠、第二交換器-第二交換器埠-第一裝置埠、第一交換器-第一交換器埠-第二裝置埠、或者第二交換器-第二交換器埠-第二裝置埠，但並不以此為限。於一些實施例中，連接路徑亦可包括第一交換器-第一交換器埠-第二裝置埠-第二交換器埠-第一裝置埠、或者第二交換器-第二交換器埠-第一裝置埠-第一交換器埠-第二裝置埠。

【0035】 於一些實施例中，可擴充集中式NVMe儲存盒之NVMe硬碟之每一者僅包括一高頻寬單一埠。可擴充集中式NVMe儲存核可包括一第一複數NVMe硬碟、一第一交換器、以及一第一介面埠。第一交換器係與第一介面埠連接。第一複數NVMe硬碟可透過對應之NVMe硬碟之單一埠以及一第一交換器連接至一第一複數節點。可擴充集中式NVMe儲存盒中之NVMe硬碟之數量可根據需要共用可擴充集中式NVMe儲存盒之節點之數量增加或者減少。

【0036】 雲端計算可提供於一個或者多個網路中，以供電腦伺服器使用共用之資源。雲端運算通常可包括基於網路之運算，即可用之資源集合係透過網路（例如雲端）將運算資源動態分配並依照需求分配給客戶或者使用者之電腦或者其它裝置。舉例來說，雲端運算資源可包括任何類型之資源，例如運算、儲存、以及網路裝置、虛擬主機（VM）等。例如資源可包括服務裝置（防火牆、深度封包檢測、流量監控、負載平衡器等）、運算/處理裝置（伺服器、CPU、記憶體、暴力處理能力（brute force processing capability））、儲存裝置（例如網路附加儲存（network attached storages）、儲存區域網路裝置）等。除此之外，上述之資源可用以支援虛擬網路、虛擬機器、資料庫、應用程式等。

【0037】 雲端運算資源可包括"私有雲"、"公共雲"、和/或"混合雲"。 "混合雲"可為由兩個或者多個互操作或者透過技術聯合之雲所組成之雲端基礎架構。於本質上來說，混合雲為私有雲以及公共雲之交互運作，而私有雲係以一安全且可擴充之方式加入公共雲，並利用公共雲之資源。雲端運算資源亦可透過一覆蓋網路中之虛擬網路（例如VXLAN）進行分配。

【0038】 以下將對第3~5圖之示範系統以及網路進行簡述。各個實施例將描述各種變化。接著參閱第3圖。

【0039】 第3圖係顯示適合執行本發明之技術之示範運算裝置300。運算裝置300包括一主CPU 362、介面368、以及一匯流排315（例如PCI匯流排）。當啟動於合適之軟體或者韌體之控制之下時，CPU 362係負責執行封包管理、錯誤偵測、和/

或路由功能，例如佈線錯誤偵測功能。CPU 362較佳地為於作業系統以及任何其它合適之軟體之控制下完成上述功能。CPU 362可包括一個或者多個處理器363，例如微處理器之Motorola family或者微處理器之MIPS family。於另一實施例中，處理器363係為專門設計之硬體，用以控制運算裝置300之操作。於一特定之實施例中，記憶體361（例如非揮發性記憶體和/或唯讀記憶體）亦構成CPU 362之部分。然而，記憶體可透過許多不同之方式與系統耦接。

【0040】 介面368一般係作為介面卡（有時被稱為"網路卡（line card）"）。一般而言，介面卡368控制透過網路發送以及接收之資料封包，以及有時支援其它與運算裝置300一同使用之周邊裝置。介面之間可使用乙太網路介面、訊框中繼介面（frame relay interface）、電纜介面、DSL介面、信號環（token ring）介面等。除此之外，介面之間亦可使用各種超高速介面，例如快速信號環介面、無線介面、乙太網路介面、十億位元乙太網路（Gigabit Ethernet）介面、ATM介面、高速串列（High Speed Serial Interface, HSSI）介面、POS介面、光纖分散式資料介面（Fiber Distributed Data Interface, FDDI）等。一般而言，這些介面可包括可與合適之媒介進行通訊之埠。於一些實施例中，介面亦可包括獨立的處理器以及揮發性隨機存取記憶體。獨立的處理器可控制上述之通訊密集型任務作為封包交換、媒介控制以及管理。藉由提供分開之處理器以處理通訊密集之任務，上述介面允許主微處理器362有效率地執行路由運算、網路診斷、安全功能等。

【0041】 儘管第3圖所示之系統為本發明之一特定運算裝置，但並不表示其為本發明唯一之網路裝置架構。舉例來說，亦常使用具有單一處理器之架構以處理通訊以及路由運算等。除此之外，其它類型之介面以及媒介亦可與路由一同使用。

【0042】 無論網路裝置如何配置，其可採用一個或者多個記憶體或者記憶體模組（包括記憶體361），記憶體或者記憶體模組係用以儲存通用網路操作以及漫遊（roaming）、路徑優化以及路由功能之機制之程式指令。舉例來說，程式指令可控制作業系統之操作和/或一個或者多個應用程式。一個或者多個記憶體亦可用以儲存查找表，例如移動綁定（mobility binding）、註冊、以及連結表（association table）等。

【0043】 第4、5圖係顯示根據本發明一些實施例所述之示範系統。對本領域技術人員而言，於實施本發明之技術時，更適當之實施例為顯而易見的。本領域技術人員亦可理解其它可能之系統為可行的。

【0044】 第4圖係顯示系統匯流排運算系統架構400，其中系統之元件係彼此係使用匯流排402進行電性通訊。示範系統400包括處理單元（CPU或者處理器）430以及將包括系統記憶體404（例如唯讀記憶體406以及隨機存取記憶體408）之各個系統元件耦接至處理器430之系統匯流排402。系統400可包括直接與處理器430連接、靠近處理器430或者整合為處理器430之一部分之快速記憶體之快取。系統400可自記憶體404和/或儲存裝置412複製資料至快取428以供處理器430快速存取。透過這種方法快取可避免處理器430因等待資料所造成之延遲以

提升性能。上述之模組或者其它模組可控制或者用以控制處理器430執行各種動作。其它系統記憶體404亦可作為選擇。記憶體404可包括各種具有不同性能特徵之不同類型之記憶體。處理器430可包括任何通用處理器，以及用以控制處理器430以及將軟體指令結合至實際的處理器設中之專門處理器之硬體模組或者軟體模組（例如儲存於儲存裝置412中之模組1 414、模組2 416、以及模組3 418）。處理器430實質上可為完全獨立之運算系統，包括多個核心或者處理器、匯流排、記憶體控制器、快取等。多核心處理器可為對稱或者不對稱的。

【0045】 為了致能使用者與運算裝置400之間之互動，輸入裝置400可代表任何數量之輸入機制，例如語音麥克風、產生手勢或者圖形輸入之處控感應螢幕、鍵盤、滑鼠、運動輸入、語音等。輸出裝置422亦可為一個或者多個本領域人員已知之輸出機制。於一些實施例中，多模組系統可致能使用者提供多種類型之輸入以與系統400進行溝通。通訊介面424可用以操作以及管理使用者之輸入以及系統之輸出。在此並無限制操作於任何特定之硬體配置上，因此當開發出改進之硬體或者韌體配置時，可輕易地取代本發明之基本特徵。

【0046】 儲存裝置412係為非揮發性記憶體，並可為硬碟或者為可儲存供電腦存取之資料之其它類型電腦可讀取媒介，例如磁帶、快閃記憶卡、固態儲存裝置、數位視頻光碟（Digital Video Disc, DVD）、卡帶、隨機存取記憶體408、唯讀記憶體406、以及前述裝置之結合等。

【0047】 儲存裝置412可包括用以控制處理器430之軟體模

組414、416、418，亦可為其它硬體或者軟體模組。儲存裝置412可連接至系統匯流排402。於一方面，執行特定功能之硬體模組可包括儲存於與必要硬體元件連接之電腦可讀取媒介中之軟體元件，例如處理器430、匯流排402、顯示器436等，以進行對應之功能。

【0048】 控制器410可為位於系統400上之專門微處理器或者處理器，例如主控晶片（baseboard management controller, BMC）。於一些實施例中，控制器410可為智慧平台管理介面之一部分。除此之外，於一些實施例中，控制器410可嵌於系統400之主機板或者主電路板上。控制器410可管理介於系統管理軟體以及硬體平台之間之介面。控制器亦可與各個系統裝置以及元件（內部和/或外部）進行通訊，例如控制器或者周邊元件，以下將提出更詳細之描述。

【0049】 控制器410可對通知、警示、和/或事件產生特定之回應，並與遠端裝置或者元件（例如電子郵件訊息、網路訊息等）進行通訊，產生指令或者命令以執行自動硬體修復程序等。系統管理員亦可與控制器410遠端通訊以啟動或者進行特定硬體修復程序或者操作，以下將提出更詳細之描述。

【0050】 系統400上不同類型之感應器（例如感應器426）可將參數回報給控制器410，例如冷卻風扇速度、電源狀態、作業系統狀態、硬體狀態等。控制器410亦可包括用以管理以及保存控制器410所接收之事件、警示、以及通知之系統事件日誌控制器和/或儲存器。舉例來說，控制器410或者系統事件日誌控制器可自一個或者多個裝置以及元件接收警示或者通

知，並將警示或者通知保存於系統事件日誌儲存元件中。

【0051】 快閃記憶體432可為系統400用以儲存和/或作為資料傳輸之電子非揮發電腦儲存媒介或者晶片。快閃記憶體432可電性抹除和/或編程。舉例來說，快閃記憶體432可包括可抹除可編程唯讀記憶體（EPROM）、電性可抹除可編程唯讀記憶體（EEPROM）、ROM、NVRAM、或者互補金屬氧化物半導體（CMOS）。快閃記憶體432可儲存當系統400開機時系統400所執行之韌體434，以及韌體434之一組特定配置。快閃記憶體432亦可儲存韌體434所使用之配置。

【0052】 韌體434可包括基本輸入/輸出系統或者其替代或者類似之韌體，例如EFI（Extensible Firmware Interface）或者UEFI（Unified Extensible Firmware Interface）。韌體434可於系統400啟動時作為順序程序載入以及執行。韌體434可根據一組配置設定辨識、初始化、以及測試系統400中之硬體。韌體434可於系統400上執行自測試，例如開機自我檢測（Power-on Self-Test, POST）。自測試可測試各個硬體元件之功能，例如硬碟、光學讀取裝置、冷卻裝置、記憶體模組、擴充卡等。韌體434可於記憶體404、唯讀記憶體406、隨機存取記憶體408、和/或儲存裝置412中定址以及定位一區域以儲存作業系統。韌體434可載入啟動載入器和/或作業系統，並將系統400之控制權交給作業系統。

【0053】 系統400之韌體434可包括定義韌體434如何控制系統400中各個硬體元件之韌體配置。韌體配置可決定系統400中各個硬體元件之啟動順序。韌體434可提供設定各種不同參

數之介面（例如UEFI），其中參數之設定可與韌體預設配置中之參數不同。舉例來說，使用者（例如系統管理員）可使用韌體434以指定時脈以及匯流排速度、定義哪些周邊裝置連接至系統400、設定系統健康（例如風扇速度以及CPU溫度限制）之監控、和/或提供各種其它影響系統400之整體效能以及電源使用之參數。

【0054】 儘管本發明顯示韌體434係儲存於快閃記憶體432中，但本領域技術人員將容易理解韌體434可儲存於其它記憶體元件（例如記憶體404或者唯獨記憶體406中）。然而，本發明所示儲存於快閃記憶體432中之韌體434僅作為說明目的之非限制實施例。

【0055】 系統400可包括一個或者多個感應器426。舉例來說，一個或者多個感應器426可包括一個或著多個溫度感測器、熱感測器（thermal sensor）、氧感測器、化學感測器、噪音感測器、熱感測器（heat sensor）、電流感測器、電壓檢測器、氣流感測器、紅外線溫度計、熱流感測器、溫度計、高溫計等。一個或者多個感測器426可透過匯流排402與處理器、快取428、快閃記憶體432、通訊介面424、記憶體404、唯讀記憶體406、隨機存取記憶體408、控制器410、以及儲存裝置412進行通訊。一個或者多個感測器426亦可透過不同之方法（例如內部整合電路（inter-integrated circuit, I2C）、通用型輸出（general purpose output, GPO）、以及類似之介面）與系統中之其它元件進行通訊。

【0056】 第5圖係顯示具有可用以執行前述之方法或者操

作、並產生以及顯示圖形化使用者介面之晶片組架構之示範電腦系統500。電腦系統500可包括用以執行本發明技術特徵之電腦硬體、軟體、以及韌體。系統500可包括處理器510，代表任何數量之可運作用以執行運算之軟體、韌體、以及硬體之物理和/或邏輯上之不同資源。處理器510可與控制處理器510之輸入以及輸出之晶片502進行通訊。於此實施例中，晶片502輸出資訊至輸出514（例如顯示器），並可讀取以及寫入資訊至儲存裝置516，例如磁性媒介以及固態媒介。晶片502亦可自隨機存取記憶體518讀取資料以及寫入資料至隨機存取記憶體518。連接各種使用者介面元件506之橋接器（bridge）504可供用以與晶片502橋接。上述之使用者介面元件506可包括鍵盤、麥克風、觸控偵測以及處理電路、指向裝置（例如滑鼠）等。一般而言，系統500之輸入可源自任何來源、機器所產生之輸入和/或使用者所產生之輸入。

【0057】 晶片502亦可與一個或者多個具有不同物理介面之通訊介面508連接。上述之通訊介面可包括有線以及無線區域網路、寬頻無線網路、以及個人區域網路之介面。用以產生、顯示、以及使用在此所述之圖形化使用者介面之方法之一些應用包括藉由處理器510接收透過物理介面傳輸或者機器本身所產生之資料集，其中處理器510係用以分析儲存於儲存裝置516或者518中之資料。除此之外，機器可透過使用者介面元件506自使用者接收輸入，並執行合適之功能，例如透過使用處理器510分析上述輸入以執行瀏覽功能。

【0058】 除此之外，晶片502亦可與電腦系統500於開機時

所執行之韌體512進行通訊。韌體502可根據一組韌體配置辨識、初始化、以及測試電腦系統500中之硬體。韌體512可於系統500上執行自測試，例如開機自我檢測（Power-on Self-Test, POST）。自測試可測試各個硬體元件502~518之功能。韌體512可於記憶體518中定址以及定位一區域以儲存作業系統。韌體512可載入啟動載入器和/或作業系統，並將系統500之控制權交給作業系統。於一些實施例中，韌體512可與硬體元件502~510以及514~518進行通訊。在此韌體512可透過晶片502和/或透過一個或著多個其它元件與硬體元件502~510以及514~518進行通訊。於一些實施例中，韌體512可直接與硬體元件502~510以及514~518進行通訊。

【0059】 可以理解的是，示範系統400以及500可具有多於一個之處理器（例如430、510）或者為透過網路連接在一起之電腦裝置之群組或者叢集之部分，以提供更強大之處理能力。

【0060】 為了清楚說明，本發明一些實施例係顯示包括具有裝置、裝置元件、透過軟體或者軟硬體之結合實現方法中之步驟或者程序之功能區塊。

【0061】 於一些實施例中，電腦可讀取儲存裝置、媒介、以及記憶體可包括含有位元流之纜線或者無線訊號等。然而，當提到非暫態電腦可讀取儲存媒介時，係明確地排除例如能量、載波信號、電磁波、以及信號本身。

【0062】 根據前述實施例所述之方法可藉由使用儲存於或者透過其它方法自電腦可讀取媒介中存取之電腦可執行指令實現。舉例來說，上述之指令可包括致使或者以其它方式配置

通用電腦、專門電腦、或者專門處理裝置執行特定功能或者複數功能之指令或者資料。部分之電腦資源可透過網路存取。電腦可執行指令可為二進制編碼、中間格式指令（例如組合語言、韌體、或者原始程式碼）。可用以儲存指令、所使用之資訊、和/或前述實施例之方法執行期間所建立之資訊之電腦可讀取介質之範例可包括磁片或者光碟片、快閃記憶體、提供非揮發性記憶體之通用序列匯流排裝置、網路儲存裝置等。

【0063】 用以實施這些方法之裝置可包括硬體、韌體和/或軟體，並可帶有任何複數形式參數。帶有任何複數形式參數之典型範例包括筆記型電腦、智慧型手機、小型個人電腦、個人數位助理、機架式裝置、獨立設備等。在此所述之功能亦可實現於周邊裝置或者外接卡中。上述之功能亦可透過其它示例實現於不同晶片之間之電路板或者單一裝置中所執行之不同程序。

【0064】 指令、傳輸上述指令之媒介、執行上述指令之運算資源、以及其它支援上述運算資源之架構係為提供在此所述之功能之手段。

【0065】 本發明之各個方面係提供支援伺服器系統中可擴充集中式非揮發性記憶體儲存盒之系統以及方法。儘管前面所述之具體實施例已顯示如何將可選操作以不同之指令實現，但其它實施例係可將可選操作結合至不同指令中。為了清楚說明，本發明一些實施例可表示為包括具有裝置、裝置元件、透過軟體或者軟硬體之結合實現方法中之步驟或者程序之功能區塊。

【0066】 各個實施例更可於各種操作環境中實現，於一些例子中可包括一個或者多個伺服器電腦、用戶電腦或者可運作任何數量之應用程式之運算裝置。用戶或者客戶端裝置可包括任何數量之通用個人電腦，例如運作標準作業系統之桌上型電腦或者筆記型電腦，以及運作手機軟體以及可支援一些網路以及訊息協定之移動、無線以及手持裝置。上述之系統亦可包括一些運作任何類型之商用作業系統以及用於其它已知應用（例如開發以及資料庫管理）之工作站。上述裝置亦可包括其它電子裝置，例如虛擬輸出端、瘦客戶端（thin client）、遊戲系統以及其它可透過網路進行通訊之裝置。

【0067】 本發明一些實施例或者部分可實現於硬體中，本發明所述之方法可以下列技術之任一者或者其結合實現：具有用以對資料訊號執行邏輯功能之邏輯閘之離散邏輯電路、具有合適之組合邏輯閘之特殊應用積體電路（application specific integrated circuit, ASIC）、可編程硬體（例如可程式閘陣列（programmable gate array, PGA）、現場可程式閘陣列（field programmable gate array, FPGA））等。

【0068】 大多數之實施例係使用對本領域技術人員為熟知的至少一網路以支援通訊。舉例來說，網路可為區域網路、廣域網路、虛擬私有網路、網際網路、企業內部網路（intranet）、商際網路（extranet）、公共電話交換網（public switched telephone network）、紅外線網路、無線網路以及上述網路之結合。

【0069】 根據前述實施例所述之方法可藉由使用儲存於或

者透過其它方法自電腦可讀取媒介中存取之電腦可執行指令實現。舉例來說，上述之指令可包括致使或者以其它方式配置通用電腦、專門電腦、或者專門處理裝置執行特定功能或者複數功能之指令或者資料。部分之電腦資源可透過網路存取。電腦可執行指令可為二進制編碼、中間格式指令（例如組合語言、韌體、或者原始程式碼）。可用以儲存指令、所使用之資訊、和/或前述實施例之方法執行期間所建立之資訊之電腦可讀取介質之範例可包括磁片或者光碟片、快閃記憶體、提供非揮發性記憶體之通用序列匯流排裝置、網路儲存裝置等。

【0070】 用以實施這些方法之裝置可包括硬體、韌體和/或軟體，並可帶有任何複數形式參數。帶有任何複數形式參數之典型範例包括伺服器電腦、筆記型電腦、智慧型手機、小型個人電腦、個人數位助理等。在此所述之功能亦可實現於周邊裝置或者外接卡中。上述之功能亦可透過其它示例實現於不同晶片之間之電路板或者單一裝置中所執行之不同程序。

【0071】 在利用網頁伺服器的實施例中，網頁伺服器可執行任何種類之伺服器或中層（mid-tier）應用程式，包括超文件傳送協定（HTTP）伺服器、檔案傳送協定（FTP）伺服器、共同閘道介面（CGI）伺服器、資料伺服器、JAVA伺服器、及商業應用伺服器。這些伺服器可用以執行回應來自用戶裝置之要求之程式或腳本（script），例如藉由執行一或者多個網頁應用程式，網頁應用程式可利用任何程式語言（例如：Java®、C、C#或者C++）、或者任何腳本語言（例如Perl、Python、TCL）以及其組合之腳本或程式撰寫實施。伺服器亦可包括資料庫伺

服器，並不限於來自開放市場之商用可用軟體。

【0072】 伺服器系統可包括前述之各種資料儲存以及其它記憶體以及儲存媒介。上述伺服器系統可註冊於各種位址，例如一儲存多媒體本地連結（和/或註冊）至一或多個電腦或從透過網路從任何或所有電腦遠端連結。於一組特別之實施例中，資訊可註冊於本領域具有通常知識者所熟知的儲存區域網路（SAN）。同樣地，用以執行對電腦、伺服器或其他網路裝置有貢獻功能之任意有需要之資料夾可被本地和/或遠端儲存。其中系統係包括複數個電腦化裝置，每個裝置包括可透過一匯流排電性耦合之多個硬體元件。舉例來說，這些硬體元件至少包括一中央處理單元、一輸入裝置（例如滑鼠、鍵盤、控制器、觸控感應顯示元件或輔助鍵盤）以及至少一輸出裝置（例如顯示器裝置、印表機或喇叭）。上述系統亦可包括一或多個儲存裝置，例如光碟裝置、光學儲存裝置、固態儲存裝置（例如隨機存取記憶體或唯讀記憶體）以及可移除式多媒體裝置、記憶卡、快閃記憶卡等。

【0073】 上述裝置亦可包括一電腦可讀取儲存媒介閱讀器、通訊裝置（例如數據機、網路卡（有線或無線）、紅外線運算裝置）以及以上所述之工作記憶體裝置（working memory）。電腦可讀取儲存媒介讀取器可連接至或者用以接收自電腦可讀取儲存媒介，電腦可讀取儲存媒介代表遠端、本地、混合和/或可移除式儲存裝置，用以暫時性及/或更永久地包含、儲存、傳送、以及取回電腦可讀取資訊之儲存媒介。系統和多種裝置可典型地將包括若干個至少位於一工作記憶體

裝置之軟體應用程式、模組、服務或其他元件，包括一作業系統以及應用程式（例如用戶端應用程式或者網頁瀏覽器）。必須理解的是，亦可根據前述之範例作各種變化。舉例來說，亦可使用客製化硬體和/或特殊元件可實施於硬體、軟體（包括可攜式軟體，例如applets）或者兩者之上。除此之外，連結至其它計算裝置的連結像是網路輸入輸出裝置可被採用。

【0074】 包含程式碼、或者部分程式碼之儲存媒介以及電腦可讀取媒介可包括任何習知技術之適當多媒體，包括儲存式媒介以及運算媒介，包含揮發性以及非揮發性、可移除和不可移除媒介，以便以任何方法或技術實現用以傳輸資訊（例如電腦可讀取指令、資料結構、程式模組或其它資料），包括隨機存取記憶體、唯讀記憶體、可抹除可編程唯讀記憶體、電子可抹除可編程唯讀記憶體、快閃記憶體、或其他記憶體技術、光碟唯讀記憶體（CD-ROM）、DVD、或其它光學儲存裝置、磁卡、磁帶磁片除儲存裝置或其他磁儲存裝置或者任何其它可用以儲存所需資訊以及系統裝置可存取接收之媒介。本領域技術人員可根據本發明提供之方法與技術將本發明描述之功能以各種不同方法作實現。

【0075】 儘管以上已揭露本發明較佳之實施例，但其並非用以限定本發明，任何熟習此技藝者，在不脫離本發明之精神以及範圍內，當可作些許之更動以及潤飾，因此本發明之保護範圍當視後附之申請專利範圍所界定者為準。

【符號說明】

【0076】

100A~100D~ 示範系統

101~ 伺服器系統

1011~1018~ 節點

102~ 儲存子系統

1021~ 裝置埠

1022~ 交換器埠

1023~ 交換器埠

1024~ 裝置埠

103~ 結構控制器

104~ 硬體編制管理器

109~ 交換器

120~ 電源供應器

122~ 電源供應單元

190~192~ PCIe交換器

1901~1904~ 儲存裝置

300~ 運算裝置

315~ 匯流排

361~ 記憶體

362~ CPU

363~ 處理器

368~ 介面

400~ 系統匯流排運算系統架構

402~ 匯流排

404~ 系統記憶體

406~ 唯讀記憶體

408~ 隨機存取記憶體

410~ 控制器

412~ 儲存裝置

414~ 模組1

416~ 模組2

418~ 模組3

420~ 輸入裝置

422~ 輸出裝置

424~ 通訊介面

426~ 感應器

428~ 快取

430~ 處理器

432~ 快閃記憶體

434~ 韌體

436~ 顯示器

500~ 電腦系統

502~ 晶片

504~ 橋接器

506~ 使用者介面元件

508~ 通訊介面

510~ 處理器

512~ 韌體

514~ 輸出

516~ 儲存裝置

518~ 隨機存取記憶體

公告本

發明摘要

※ 申請案號：105124918
※ 申請日：105/08/05 ※IPC 分類：G06F 13/40 (2006.01)
G06F 13/42 (2006.01)
G06F 3/06 (2006.01)

【發明名稱】可擴充集中式非揮發性記憶體儲存盒、電腦實施方法以及非暫態電腦可讀取儲存裝置
SCALABLE POOLED-NON-VOLATILE MEMORY EXPRESS STORAGE BOXES, COMPUTER-IMPLEMENTED METHODS AND NON-TRANSITORY COMPUTER-READABLE STORAGE MEDIUM

【中文】

本發明提供一種支援可共用於複數節點之間於可擴充集中式非揮發性記憶體（NVMe）儲存盒之叢集架構。可擴充集中式NVMe儲存盒包括NVMe硬碟、一個或者多個交換器以及一個或者多個交換器埠。可擴充集中式NVMe儲存盒中NVMe硬碟之數量可根據需要共用可擴充集中式NVMe儲存盒之節點之數量增加或者減少。

【英文】

The present invention provides a cluster-architecture to support a scalable pooled-NVMe storage box that can be shared among a scalable number of nodes. The scalable pooled-NVMe storage box comprises NVMe drives, one or more switches and one or more switch ports. The number of NVMe drives in the

scalable scalable-pooled-NVMe storage box can be scaled up or down based upon a number of nodes that need to share the scalable pooled-NVMe storage box.

申請專利範圍

1. 一種可擴充集中式非揮發性記憶體儲存盒，包括：

一第一PCIe交換器；

一第一交換器埠，連接至上述第一PCIe交換器；

一第一裝置埠；以及

一第一複數非揮發性記憶體硬碟，上述第一複數非揮發性記憶體硬碟之每一者包括一第一埠以及一第二埠；

其中，上述第一複數非揮發性記憶體硬碟之上述第一埠係透過上述第一PCIe交換器連接至一第一複數節點；

其中，上述第一複數非揮發性記憶體硬碟之上述第二埠係連接至上述第一裝置埠。

2. 如申請專利範圍第1項所述之可擴充集中式非揮發性記憶體儲存盒，更包括：

一第二PCIe交換器；

一第二交換器埠，連接至上述第二PCIe交換器；

一第二裝置埠；以及

一第二複數非揮發性記憶體硬碟，上述第二複數非揮發性記憶體硬碟之每一者包括一第一埠以及一第二埠；

其中，上述第二複數非揮發性記憶體硬碟之上述第一埠係透過上述第二PCIe交換器連接至一第二複數節點；

其中，上述第二複數非揮發性記憶體硬碟之上述第二埠係連接至上述第二裝置埠；

其中，上述第一裝置埠係連接至上述第二交換器埠；

其中，上述第一交換器埠係連接至上述第二裝置埠；

其中，上述第一複數節點之間係透過上述第一PCIe交換器共用上述第一複數非揮發性記憶體硬碟之每一者，以及上述第二複數節點之間係透過一第一連接路徑共用上述第一複數非揮發性記憶體硬碟之每一者，上述第一連接路徑為：第二PCIe交換器-第二交換器埠-第一裝置埠；和/或

其中，上述第二複數節點係透過上述第二PCIe交換器共用上述第二複數非揮發性記憶體硬碟之每一者，以及上述第一複數節點係透過一第二連接路徑共用上述第二複數非揮發性記憶體硬碟之每一者，上述第二連接路徑為：第一PCIe交換器-第一交換器埠-第二裝置埠。

3. 如申請專利範圍第1項所述之可擴充集中式非揮發性記憶體儲存盒，其中上述第一交換器埠係與上述第一裝置埠連接。

4. 如申請專利範圍第1項所述之可擴充集中式非揮發性記憶體儲存盒，更包括：

一第二PCIe交換器；

一第二介面埠，連接至上述第二PCIe交換器；以及

一第二複數非揮發性記憶體硬碟，上述第二複數非揮發性記憶體硬碟之每一者包括一第一埠以及一第二埠；

其中，上述第一交換器埠以及上述第一裝置埠係結合為一第一介面埠；

其中，上述第一複數非揮發性記憶體硬碟之一特定非揮發性記憶體硬碟係透過上述特定非揮發性記憶體硬碟之上述第一埠以及上述第一PCIe交換器或者透過上述特定非揮發性記

憶體硬碟之上述第二埠、上述第一介面埠以及上述第一PCIe交換器連接至上述第一複數非揮發性記憶體硬碟；

其中，上述第二複數非揮發性記憶體硬碟之上述第一埠係透過上述第二PCIe交換器連接至一第二複數節點；

其中，上述第二複數非揮發性記憶體硬碟之上述第二埠係連接至上述第二介面埠；

其中，上述第二介面埠係連接至上述第一介面埠；

其中，上述第二複數節點之間係透過上述第二PCIe交換器共用上述第二複數非揮發性記憶體硬碟之每一者；以及

其中，上述第一複數節點之間係透過一連接路徑共用上述第二複數非揮發性記憶體硬碟之每一者，上述連接路徑為：第一PCIe交換器-第一介面埠-第二介面埠。

5. 如申請專利範圍第1項所述之可擴充集中式非揮發性記憶體儲存盒，其中兩個或者多個上述第一非揮發性記憶體硬碟係與兩個或者多個上述第一複數節點集群以產生一高可用性叢集。

6. 一種支援可擴充集中式非揮發性記憶體儲存盒之電腦實施方法，其中上述可擴充集中式非揮發性記憶體儲存盒包括一第一PCIe交換器、連接至上述第一PCIe交換器之一第一交換器埠、一第一裝置埠以及一第一複數非揮發性記憶體硬碟，上述第一複數非揮發性記憶體硬碟之每一者係包括一第一埠以及一第二埠，步驟包括：

致使上述第一複數非揮發性記憶硬碟之上述第一埠連接至上述第一PCIe交換器；

致使上述第一複數非揮發性記憶體硬碟之上述第二埠連接至上述第一裝置埠；以及

致使上述第一複數非揮發性記憶體硬碟透過上述第一PCIe交換器連接至一第一複數節點。

7. 如申請專利範圍第6項所述之電腦實施方法，更包括：

判斷上述可擴充集中式非揮發性記憶體儲存盒是否必須支援一第二複數節點，其中上述可擴充集中式非揮發性記憶體儲存盒更包括一第二PCIe交換器、連接至上述第二PCIe交換器之一第二交換器埠、一第二裝置埠、以及一第二複數非揮發性記憶體硬碟，上述第二複數非揮發性記憶體硬碟之每一者係包括一第一埠以及一第二埠；

致使上述第二複數非揮發性記憶體硬碟之上述第一埠連接至上述第二PCIe交換器；

致使上述第二複數非揮發性記憶體硬碟之上述第二埠連接至上述第二裝置埠；

致使上述第一裝置埠連接至上述第二交換器埠；以及

致使上述第一交換器埠連接至上述第二裝置埠；

其中，上述第一複數節點之間係透過上述第一PCIe交換器共用上述第一複數非揮發性記憶體硬碟之每一者，以及其中上述第二複數節點之間係透過一第一連接路徑共用上述第一複數非揮發性記憶體硬碟之每一者，上述第一連接路徑為：第二PCIe交換器-第二交換器埠-第一裝置埠；和/或

其中，上述第二複數節點係透過上述第二PCIe交換器共用上述第二複數非揮發性記憶體硬碟之每一者，以及其中上述第

一複數節點係透過一第二連接路徑共用上述第二複數非揮發性記憶體硬碟之每一者，上述第二連接路徑為：第一PCIe交換器-第一交換器埠-第二裝置埠。

8. 如申請專利範圍第6項所述之電腦實施方法，其中上述第一交換器埠係與上述第一裝置埠連接。

9. 如申請專利範圍第6項所述之電腦實施方法，更包括：

判斷上述可擴充集中式非揮發性記憶體儲存盒是否必須支援一第二複數節點，其中上述可擴充集中式非揮發性記憶體儲存盒更包括一第二PCIe交換器、連接至上述第二PCIe交換器之一第二介面埠、以及一第二複數非揮發性記憶體硬碟，上述第二複數非揮發性記憶體硬碟之每一者包括一第一埠以及一第二埠；

致使上述第二複數非揮發性記憶體硬碟之上述第一埠連接至上述第二PCIe交換器；

致使上述第二複數非揮發性記憶體硬碟之上述第二埠連接至上述第二介面埠；以及

致使上述第一介面埠連接至上述第二介面埠；

其中，上述第一交換器埠以及上述第一裝置埠係結合為一第一介面埠；

其中，上述第一複數非揮發性記憶體硬碟之一特定非揮發性記憶體硬碟係透過上述特定非揮發性記憶體硬碟之上述第一埠以及上述第一PCIe交換器或者透過上述特定非揮發性記憶體硬碟之上述第二埠、上述第一介面埠以及上述第一PCIe交換器連接至上述第一複數非揮發性記憶體硬碟；

其中，上述第二複數節點之間係透過上述第二PCIe交換器共用上述第二複數非揮發性記憶體硬碟之每一者，以及其中上述第一複數節點之間係透過一連接路徑共用上述第二複數非揮發性記憶體硬碟之每一者，上述連接路徑為：第一PCIe交換器-第一介面埠-第二介面埠。

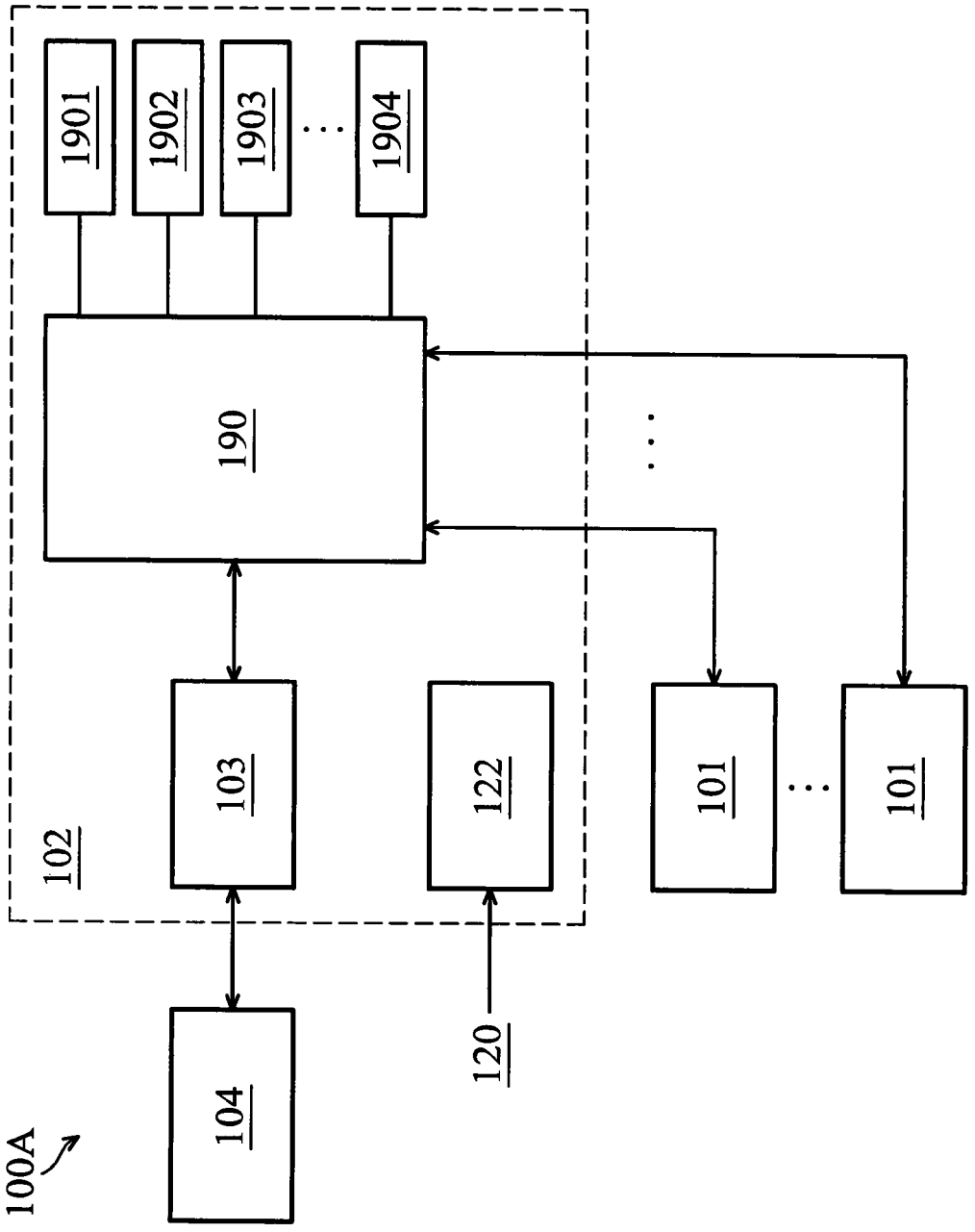
10. 一種非暫態電腦可讀取儲存裝置，儲存有支援一可擴充集中式非揮發性記憶體儲存盒之指令，其中上述可擴充集中式非揮發性記憶體儲存盒包括一第一PCIe交換器、連接至上述第一PCIe交換器之一第一交換器埠、一第一裝置埠以及一第一複數非揮發性記憶體硬碟，上述第一複數非揮發性記憶體硬碟之每一者係包括一第一埠以及一第二埠，當上述指令透過一計算機系統之至少一處理器執行時，致使上述計算機系統執行：

致使上述第一複數非揮發性記憶體硬碟之上述第一埠連接至上述第一PCIe交換器；

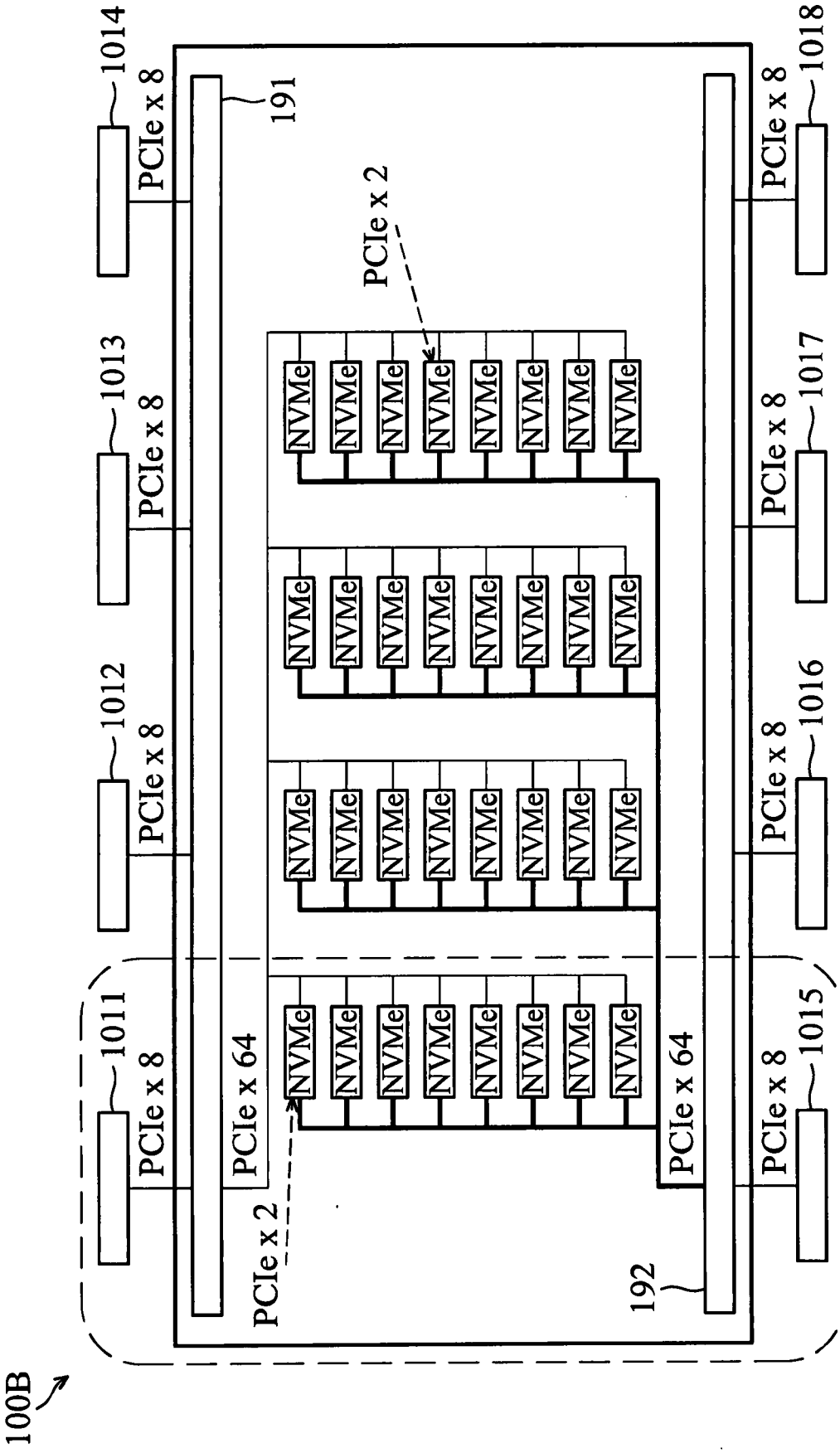
致使上述第一複數非揮發性記憶體硬碟之上述第二埠連接至上述第一裝置埠；以及

致使上述第一複數非揮發性記憶體硬碟透過上述第一PCIe交換器連接至一第一複數節點。

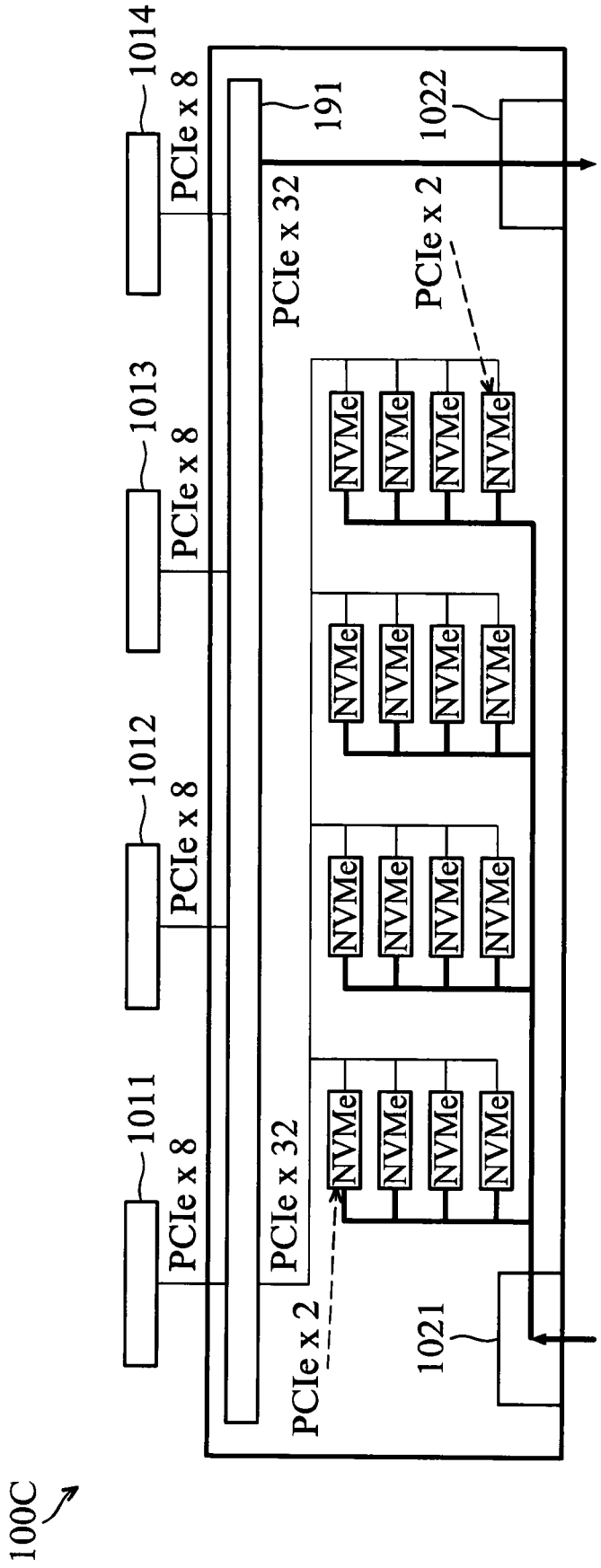
圖式



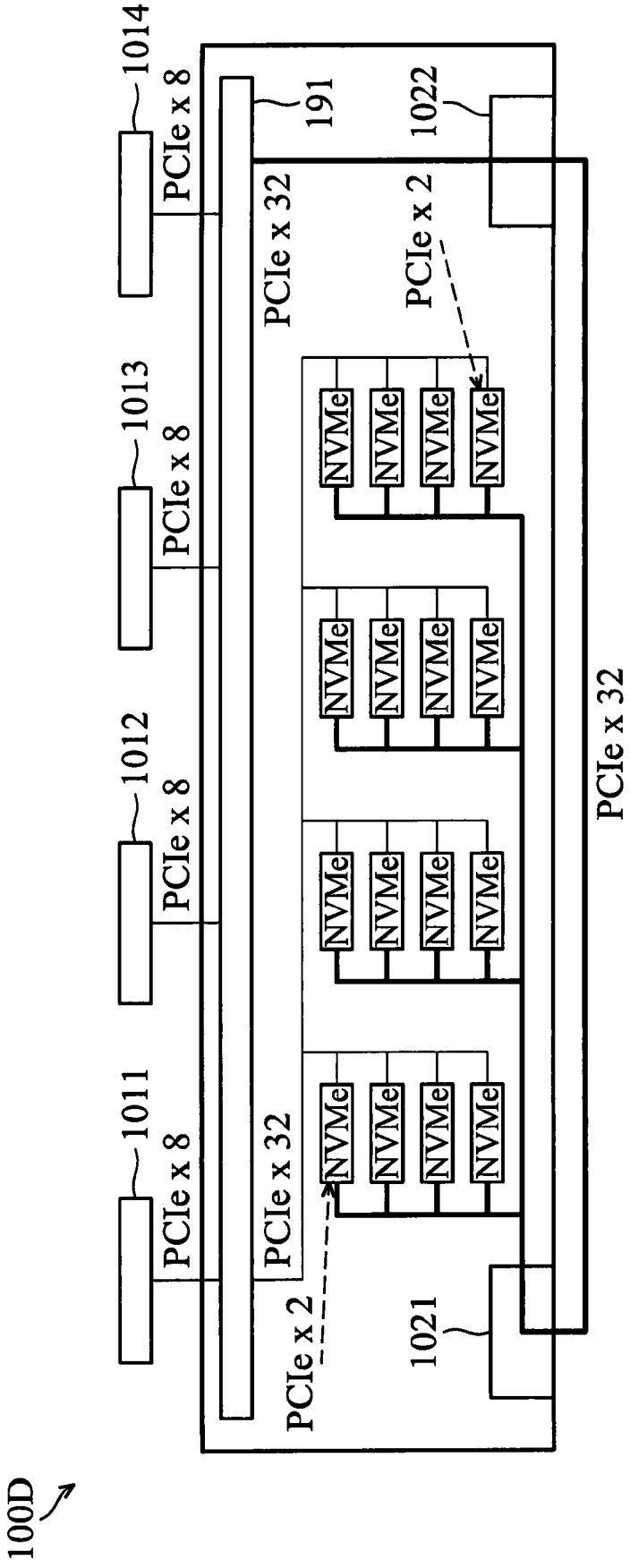
第 1A 圖


100B
1011
1012
1013
1014
1015
1016
1017
1018
PCIe x 8
PCIe x 64
PCIe x 2
191
192
NVMe

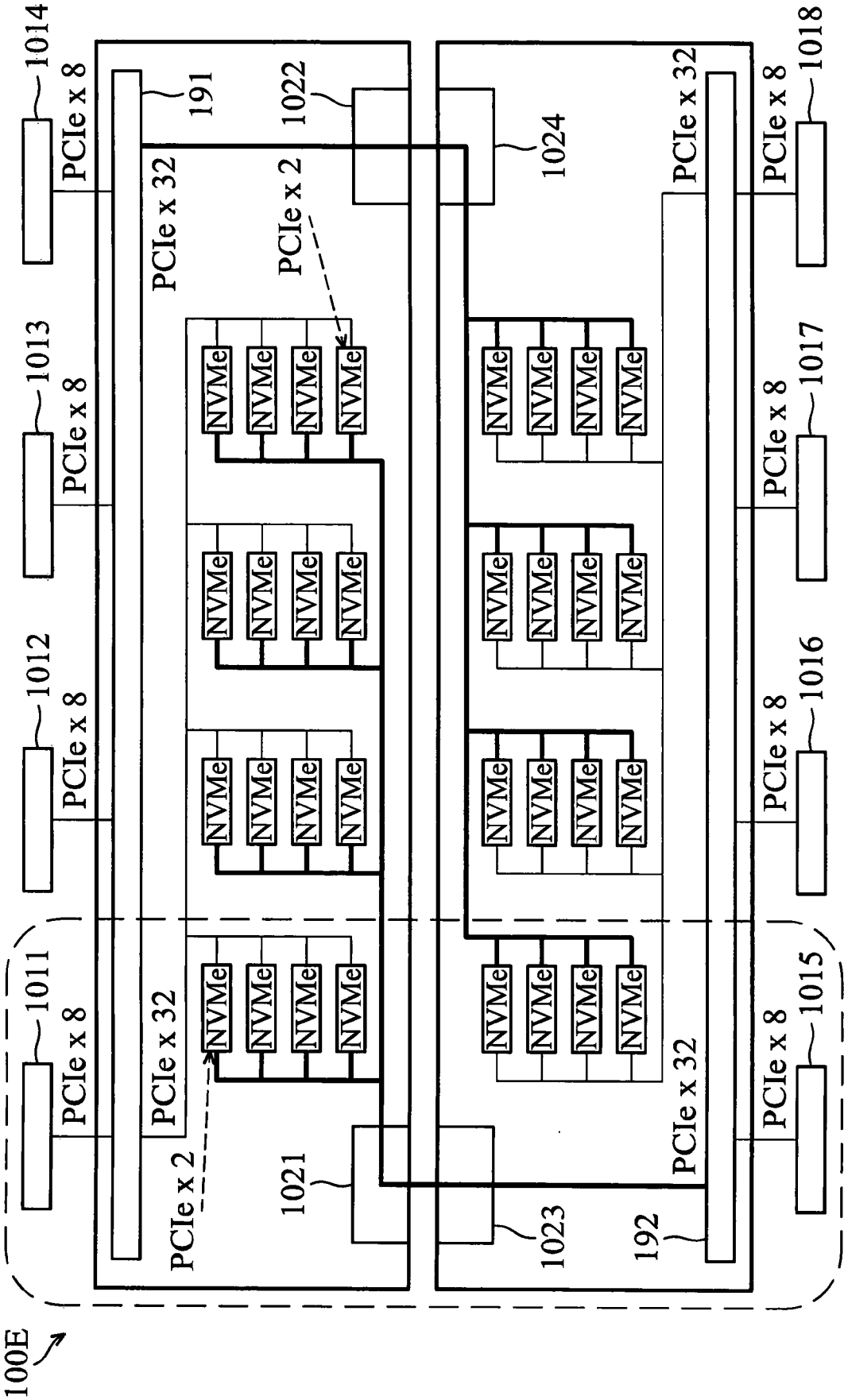
第 1B 圖


100C
1011
1012
1013
1014
PCIe x 8
PCIe x 8
PCIe x 8
PCIe x 8
PCIe x 32
PCIe x 32
PCIe x 2
PCIe x 2
191
1021
1022
NVMe

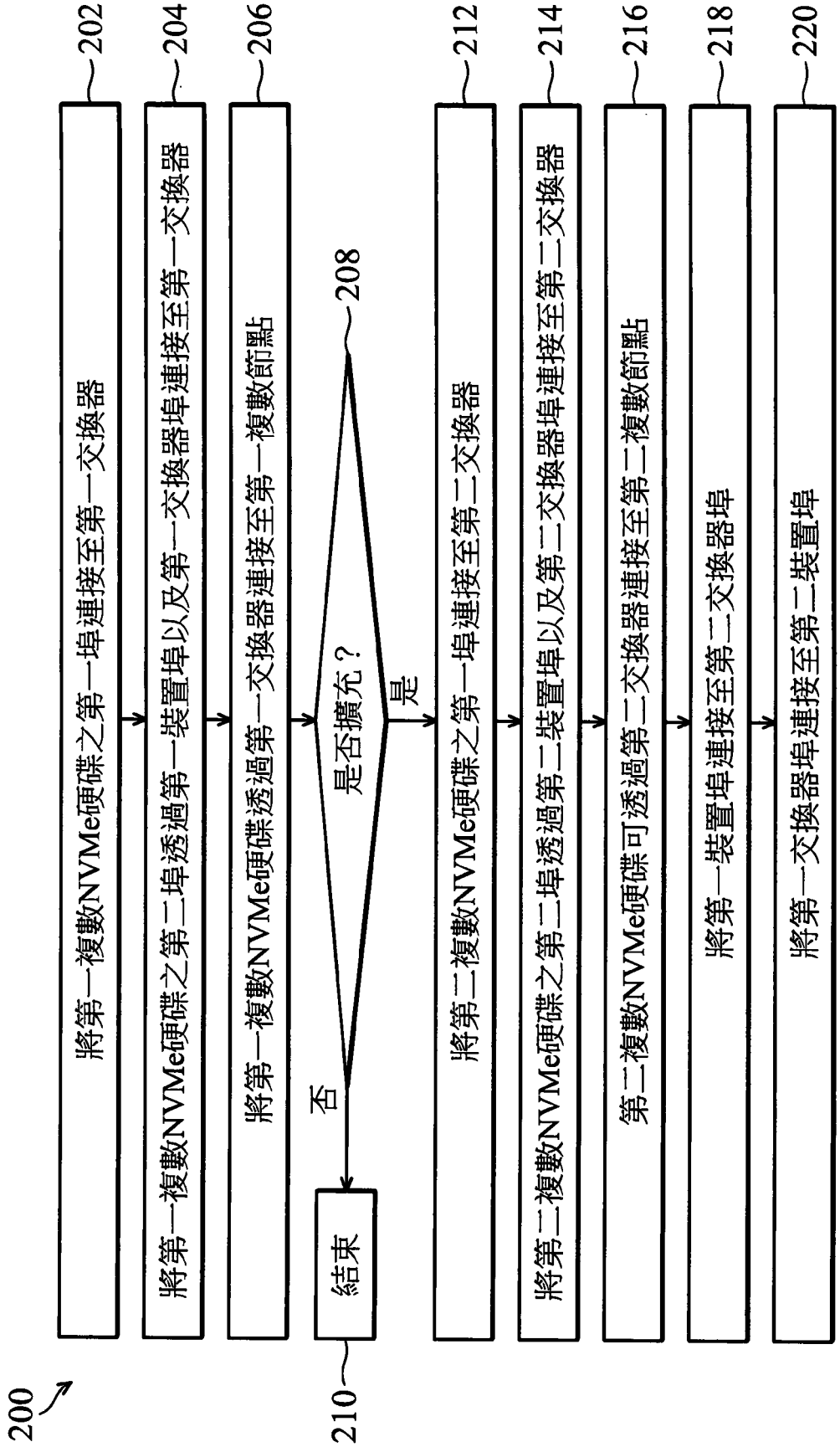
第 1C 圖


100D
1011
1012
1013
1014
PCIe x 8
PCIe x 8
PCIe x 8
PCIe x 8
PCIe x 32
PCIe x 2
191
1022
1021
PCIe x 32
PCIe x 2
PCIe x 32
NVMe

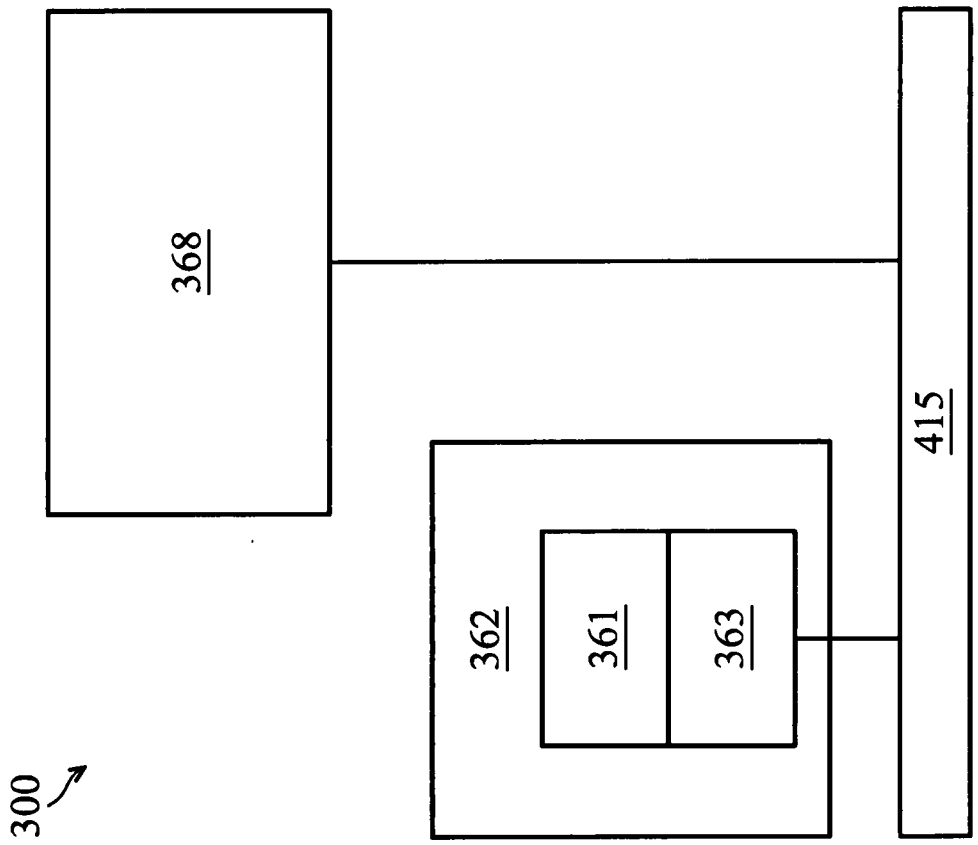
第 1D 圖


100E
1011
1012
1013
1014
1015
1016
1017
1018
1021
1022
1023
1024
191
192
PCIe x 8
PCIe x 32
PCIe x 2
NVMe

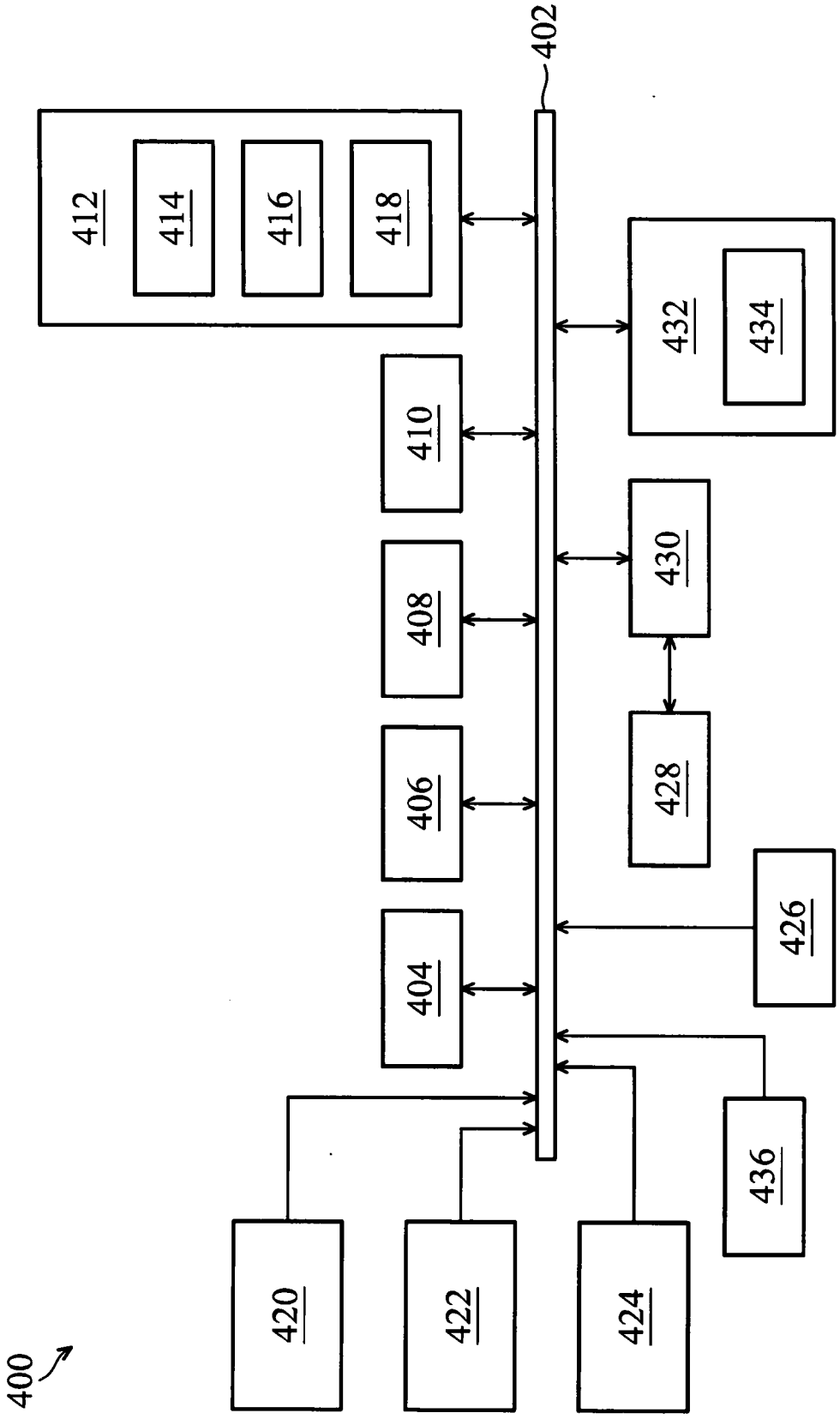
第 1E 圖


200
將第一複數NVMe硬碟之第一埠連接至第一交換器
202
將第一複數NVMe硬碟透過第二埠透過第一裝置埠以及第一交換器埠連接至第一交換器
204
將第一複數NVMe硬碟透過第一交換器連接至第一複數節點
206
是否擴充？
208
否
是
結束
210
將第二複數NVMe硬碟之第一埠連接至第二交換器
212
將第二複數NVMe硬碟之第二埠透過第二裝置埠以及第二交換器埠連接至第二交換器
214
第二複數NVMe硬碟可透過第二交換器連接至第二複數節點
216
將第一裝置埠連接至第二交換器埠
218
將第一交換器埠連接至第二裝置埠
220

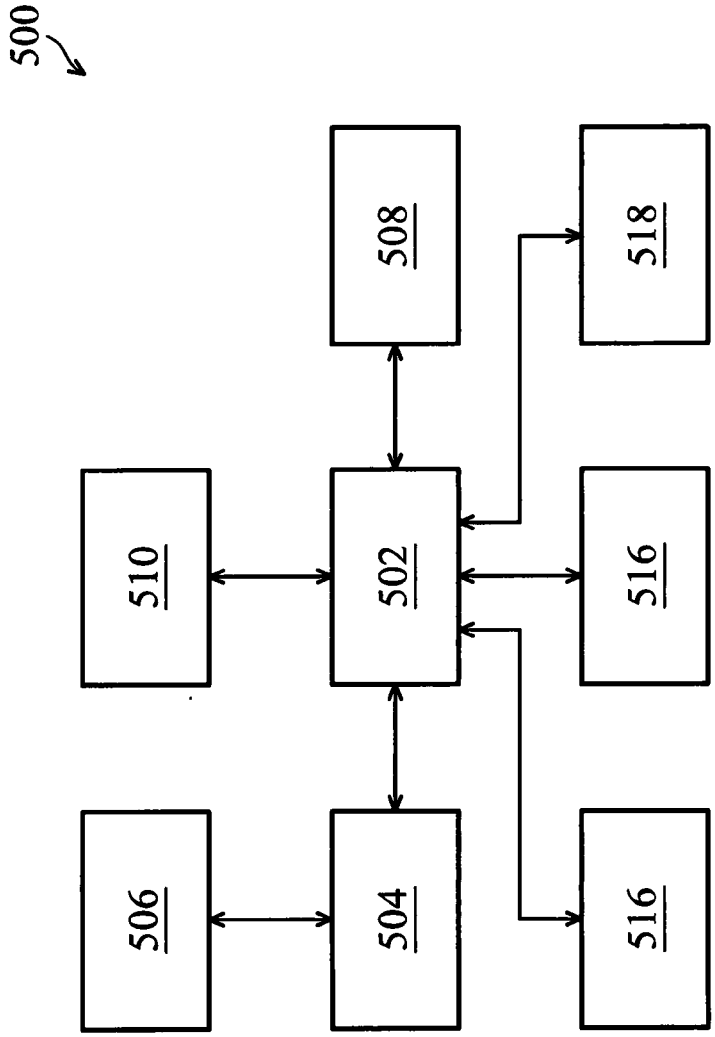
第 2 圖



第 3 圖


400
402
404
406
408
410
412
414
416
418
420
422
424
426
428
430
432
434
436

第 4 圖



第 5 圖

【代表圖】

【本案指定代表圖】：第（　1A　）圖。

【本代表圖之符號簡單說明】：

100A~ 示範系統

101~ 伺服器系統

102~ 儲存子系統

103~ 結構控制器

104~ 硬體編制管理器

120~ 電源供應器

122~ 電源供應單元

190~ PCIe交換器

1901~1904~ 儲存裝置

【本案若有化學式時，請揭示最能顯示發明特徵的化學式】：

無。

NVMe儲存盒之節點數量增加或者減少。

【0013】 第1A圖係顯示根據本發明一些實施例所述之包括一儲存子系統102（例如可擴充集中式NVMe儲存盒）以及複數節點（例如伺服器系統101）之示範系統100A之示意方塊圖。於此實施例中，儲存子系統102包括一個或者多個接收電源供應單元（PSU）122，電源供應單元122係自電源供應器120接收交流電源，並提供電源至儲存子系統102、至少一交換器（例如PCIe交換器190）、複數儲存NVMe硬碟（例如1901、1902、1903以及1904）、以及結構控制器（fabric controller）103。PCIe為用於顯示卡或者其它外接卡之底層資料傳輸層，以及傳輸介面（例如Thunderbolt傳輸介面）。儲存裝置可包括至少一NMVe硬碟（例如固態硬碟（solid state drive, SSD））。交換器之每一者係用以管理儲存子系統102之一個或者多個儲存裝置（例如接收指令並將指令路由至對應之儲存裝置）以及透過網路、管理模組、以及儲存子系統102中之其它交換器與遠端裝置進行通訊。指令可包括讀取或者寫入指令、資訊請求、或者管理指令（例如分區指令（zoning command））。指令可為文本之格式、或者PCIe介面。於此實施例中，交換器190係用以管理儲存裝置1901、1902、1903以及1904。

【0014】 結構控制器103可透過至少一交換器提供路由通訊之邏輯、指令和/或配置以將複數儲存NVMe硬碟連接至伺服器系統101。至少一交換器中之路由可透過結構控制器103配置。於一些實施例中，結構控制器103可於作業系統中執行，包括Andriod、BSD（Berkeley Software Distribution）、iOS

（iPhone OS）、Linux、OS X、Unix-like Real-time Operating System （例如QNX）、Microsoft Windows、Window Phone、以及IBM z/OS，但並不以此為限。

【0015】 儲存子系統102以及結構控制器103可透過網路介面（例如串列介面）與硬體編制管理器（hardware compose manager）104進行通訊。硬體編制管理器104可保存伺服器系統101以及一個或者多個特定資料中心和/或網路之資訊以及資料，例如硬體以及配置之細節。舉例來說，硬體編制管理器104可保存指示哪個NVMe硬碟係通信耦接至哪個伺服器系統101之資料。硬體編制管理器104亦可保存指示可供通信耦接至伺服器系統101之NVMe硬碟之資料。

【0016】 除此之外，硬體編制管理器104可儲存安裝、移除、和/或修復事件以及程序。舉例來說，硬體編制管理器104可保存有關任何自伺服器系統101新增或者移除之裝置、伺服器系統101所經歷之任何硬體故障、伺服器系統101所執行過之任何修復程序、伺服器系統101和/或複數儲存NVMe硬碟所經歷之任何硬體狀況、與伺服器系統101以及複數NVMe硬體相關之硬體狀態資訊、性能統計數據、配置資料、連結或者路由資訊等之資訊以及統計數據。

【0017】 於此實施例中，結構控制器103和/或至少一交換器（例如交換器190）亦可於硬體編制管理器104以及儲存子系統102之間提供命令列介面（command-line interface, CLI）。硬體編制管理器104或者遠端使用者可透過命令列介面或者網路介面輸入指令。命令列介面包括DCL（digital command language,

DCL)、各種Unix系統之命令直譯器(Unix shell)、CP/M(control program for microcomputers)、command.com、cmd.exe、以及RSTS(resource time sharing system)命令列介面。遠端裝置或者遠端使用者可登入儲存子系統102之命令列介面，並透過命令列介面使用應用層協定(application layer protocol)以及與儲存子系統102之複數儲存裝置(例如1901、1902、1903以及1904)相關之複數埠之輸入區域ID。

【0018】 於一些實施例中，兩個或者多個伺服器101以及儲存子系統102之交換器以及儲存裝置係叢集為一高可用度叢集系統，高可用度叢集系統係於一儲存裝置故障之情況下仍可繼續提供服務。在沒有高可用度叢集系統之情況下，若伺服器於執行一特定應用程式時崩潰，該應用程式將無法繼續執行直到崩潰之伺服器被修復為止。高可用度叢集系統可偵測伺服器上之硬體或者軟體故障，並可於無需管理員介入之情況下馬上於高可用度叢集系統中之其它伺服器上重啟應用程式。於一些實施例中，儲存裝置(例如NVMe硬碟)可為雙埠裝置。儲存裝置之每一埠可分配至高可用度叢集系統中之不同伺服器101。

【0019】 第1B~1F圖係顯示根據本發明一些實施例所述之包括至少一集中式NVMe儲存盒之示範系統之示意方塊圖。餘第1B圖中，示範系統100B包括複數節點(節點1011、1012、1013、1014、1015、1016、1017以及1018)，以及包括PCIe交換器(即191以及192)以及複數儲存裝置(即32個NVMe硬碟)之一集中式NVMe儲存盒。於此實施例中，複數儲存裝置之每一者具有兩個埠，一第一埠以及一第二埠。複數儲存裝置係透

過複數儲存裝置之第一埠以及PCIe交換器191連接至節點1011、1012、1013以及1014。複數儲存裝置亦透過複數儲存裝置之第二埠以及PCIe交換器192連接至節點1015、1016、1017以及1018。複數儲存裝置可被節點1011、1012、1013、1014、1015、1016、1017以及1018之任一者存取。

【0020】 第1C圖係顯示包括第一複數節點（即節點1011、1012、1013、以及1014）以及一可擴充集中式NVMe儲存盒之示範系統100C。可擴充集中式NVMe儲存盒包括一PCIe交換器（即191）以及一第一複數儲存盒（即16個NVMe硬碟）。於此實施例中，第一複數儲存盒之每一者具有兩個埠，一第一埠以及一第二埠。第一複數儲存裝置係透過第一複數儲存裝置之第一埠以及PCIe交換器191連接至節點1011、1012、1013以及1014。複數儲存裝置亦透過複數儲存裝置之第二埠連接至裝置埠1021。可擴充集中式NVMe儲存盒之NVMe硬碟之數量可根據需要共用可擴充集中式NVMe儲存盒之節點之數量增加或者減少。

【0021】 舉例來說，可擴充集中式NVMe儲存盒可擴充為更包括一第二複數儲存裝置、第二交換器、第二裝置埠以及第二交換器埠以提供第二複數節點（未顯示）。第一裝置埠可連接至第二交換器埠，而第一交換器埠係連接至第二裝置埠。第一複數節點以及第二複數節點之每一者可存取可擴充集中式NVMe儲存盒中之任何儲存裝置，包括第一複數儲存裝置以及第二複數儲存裝置。

【0022】 第1D圖係顯示包括第一複數節點（即節點1011、

1012、1013、以及1014）以及可擴充集中式NVMe儲存盒之示範系統100D。示範系統100D之可擴充集中式NVMe儲存盒係為示範系統100C之另一種架構。於此實施例，一第一裝置埠1021係與一第一交換器埠1022連接。可擴充集中式NVMe儲存盒之第一複數儲存裝置係透過第一複數儲存裝置之第一埠以及PCIe交換器191連接至節點1011、1012、1013以及1014。複數儲存裝置亦透過複數儲存裝置之第二埠、第一裝置埠、第一交換器埠以及PCIe交換器191連接至節點1011、1012、1013以及1014。於一些實施例中，第一裝置埠1021以及第一交換器埠1022可結合為單一介面埠以提供第一裝置埠1021以及第一交換器埠1022兩者之功能。

【0023】 第1E圖係顯示根據本發明一些實施例所述之擴充自示範系統100C之示範系統100E。於此實施例中，可擴充集中式NVMe儲存盒包括PCIe交換器（即191以及192）、交換器埠（即1022以及1023）、裝置埠（即1021以及1024）、以及儲存裝置（即32個NVMe硬碟）。裝置埠1021係連接至交換器埠1023，而裝置埠1024係連接至交換器埠1022。節點1011、1012、1013、1014、1015、1016、1017以及1018可直接透過PCIe交換器191或者PCIe交換器192、或者透過第一連接路徑（PCIe交換器191-交換器埠1022-裝置埠1024）、或者透過第二連接路徑（PCIe交換器192-交換器埠1023-裝置埠1021）存取可擴充集中式NVMe儲存盒中之任一者。

【0024】 第1F圖係顯示根據本發明一些實施例所述之包括示範系統100E之示範機架系統100F。於此實施例中，機架系統

100F之可擴充集中式NVMe儲存盒可根據需要共用可擴充集中式NVMe儲存盒之節點之數量擴充或者縮減。可擴充集中式NVMe儲存盒可包括交換器埠1022、裝置埠1021、以及第一複數NVMe硬碟（即16個NVMe硬碟）。可擴充集中式NVMe儲存盒亦可擴充至更包括一交換器埠1023、一裝置埠1024以及第二複數NVMe硬碟（即16個NVMe硬碟）。交換器埠1022係連接至裝置埠1024，而交換器埠1023係連接至裝置埠1021。可擴充集中式NVMe儲存盒之任一者（即32個NVMe硬碟之一者）可被分配給可擴充集中式NVMe儲存盒之節點存取。

【0025】 於一些實施例中，可擴充集中式NVMe儲存盒之兩個或者多個NVMe硬碟可分配給兩個或者多個節點以形成高可用度叢集系統。儘管高可用度叢集系統之一節點或者元件故障，高可用度叢集系統仍可繼續提供服務。高可用度叢集系統可偵測警示訊號或者硬體/軟體故障，並可馬上於其它沒有被任何管理干預之節點上重啟被影響或者可能被影響之應用程式。

【0026】 儘管第1A~1F圖僅顯示示範系統100A~100F之特定元件，但示範系統100A~100F亦可包括可處理或者儲存資料、或是接收或者傳輸訊號之各種類型之電子或者計算機元件。除此之外，示範系統100A~100F中之電子或者計算機元件可用以執行各種類性之應用程式和/或可使用各種類型之作業系統。這些作業系統可包括Andriod、BSD（Berkeley Software Distribution）、iOS（iPhone OS）、Linux、OS X、Unix-like Real-time Operating System （例如QNX）、Microsoft