I474181

(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 I474181 B**

(45)公告日： 中華民國 104 (2015) 年 02 月 21 日

(21)申請案號：098139783 (22)申請日： 中華民國 98 (2009) 年 11 月 23 日

(51)Int. Cl.： ***G06F13/38 (2006.01)***

(30)優先權：2008/12/30 美國 12/346,251

(71)申請人：萬國商業機器公司 (美國) INTERNATIONAL BUSINESS MACHINES CORPORATION (US)
美國

(72)發明人：布朗亞倫C BROWN, AARON C. (US)；瑞可瑞南多J RACIO, RANATO J. (US)；費謬道格拉斯M FREIMUTH, DOUGLAS M. (US)；社柏史帝芬M THURBER, STEVEN M. (US)

(74)代理人：蔡坤財；李世章

(56)參考文獻：

TW M309135U TW M342008

TW 200817914A US 7293129B2

審查人員：栗永欣

申請專利範圍項數：18 項 圖式數：17 共 69 頁

(54)名稱

在多根ＰＣＩＥ環境中區別刀鋒型目的地及訊務類型

DIFFERENTIATING BLADE DESTINATION AND TRAFFIC TYPES IN A MULTI-ROOT PCIE ENVIRONMENT

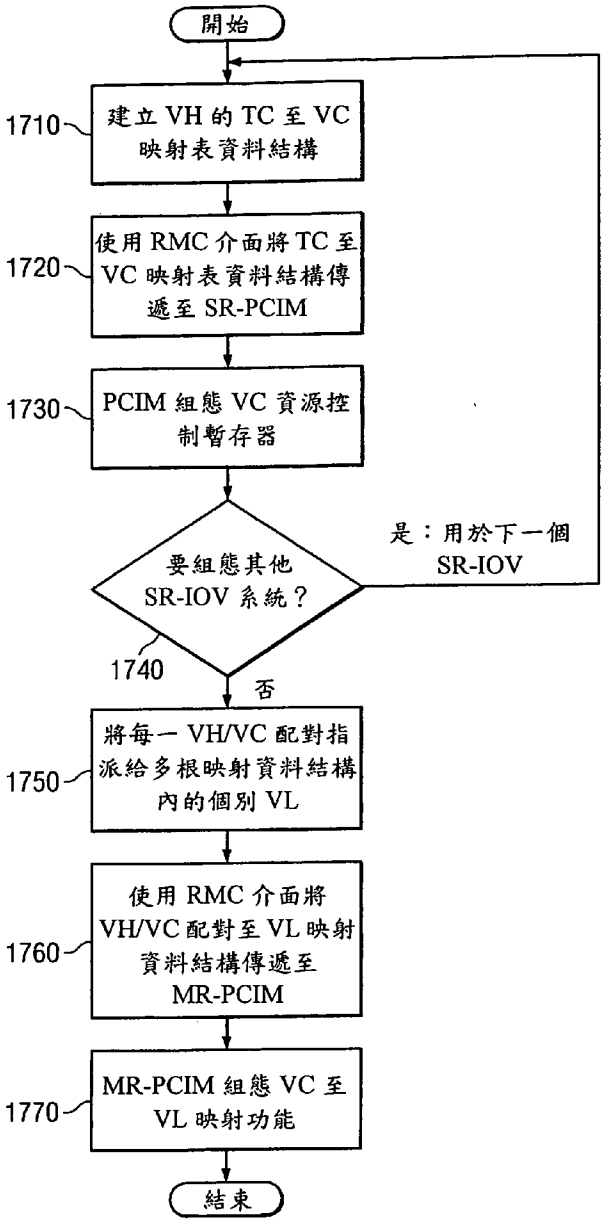
(57)摘要

本發明揭示一種區別多根 PCI Express 環境內每一主機系統刀鋒的訊務類型之機制。該機制產生一第一映射資料結構，對於該多根資料處理系統內每一單根虛擬階層，該第一映射資料結構將複數個訊務等級與複數個優先群組相關，並將該複數個訊務等級內的每一訊務等級映射至複數個虛擬通道內的一對應虛擬通道。再者，產生一第二映射資料結構，其映射該複數個虛擬通道內的每一虛擬通道至該多根資料處理系統內複數個虛擬連結內的每一對應主機系統刀鋒型虛擬連結。根據該第一映射資料結構以及第二映射資料結構，將一特定優先群組的訊務從一單根虛擬階層繞送至該複數個虛擬連結內一特定虛擬連結。

Mechanisms for differentiating traffic types per host system blade in a multi-root PCI Express environment are provided. The mechanisms generate a first mapping data structure that, for each single-root virtual hierarchy in the multi-root data processing system, associates a plurality of traffic classes with a plurality of priority groups and maps each traffic class in the plurality of traffic classes to a corresponding virtual channel in a plurality of virtual channels. Moreover, a second mapping data structure is generated that maps each virtual channel in the plurality of virtual channels to corresponding per host system blade virtual links in a plurality of virtual links of the multi-root data processing system. Traffic of a particular priority group is routed from a single-root virtual hierarchy to a particular virtual link in the plurality of the virtual links based on the first mapping data structure and second mapping data structure.

1710-1770・・・步驟

流程



第 17 圖

公告本

發明專利說明書

（本說明書格式、順序，請勿任意更動，※記號部分請勿填寫；惟已有申請案號者請填寫）

※ 申請案號：98139783

※ 申請日期：2009 年 11 月 23 日　　※IPC 分類：

G06F 13/38 (2006.01)

一、發明名稱：（中文/英文）

在多根 PCIE 環境中區別刀鋒型目的地及訊務類型

DIFFERENTIATING BLADE DESTINATION AND TRAFFIC TYPES IN A MULTI-ROOT PCIE ENVIRONMENT

二、中文發明摘要：

本發明揭示一種區別多根 PCI Express 環境內每一主機系統刀鋒的訊務類型之機制。該機制產生一第一映射資料結構，對於該多根資料處理系統內每一單根虛擬階層，該第一映射資料結構將複數個訊務等級與複數個優先群組相關，並將該複數個訊務等級內的每一訊務等級映射至複數個虛擬通道內的一對應虛擬通道。再者，產生一第二映射資料結構，其映射該複數個虛擬通道內的每一虛擬通道至該多根資料處理系統內複數個虛擬連結內的每一對應主機系統刀鋒型虛擬連結。根據該第一映射資料結構以及第二映射資料結構，將一特定優先群組的訊務從一單根虛擬階層繞送至該複數個虛擬連結內一特定虛擬連結。

三、英文發明摘要：

Mechanisms for differentiating traffic types per host system blade in a multi-root PCI Express environment are provided. The mechanisms generate a first mapping data structure that, for each single-root virtual hierarchy in the

multi-root data processing system, associates a plurality of traffic classes with a plurality of priority groups and maps each traffic class in the plurality of traffic classes to a corresponding virtual channel in a plurality of virtual channels. Moreover, a second mapping data structure is generated that maps each virtual channel in the plurality of virtual channels to corresponding per host system blade virtual links in a plurality of virtual links of the multi-root data processing system. Traffic of a particular priority group is routed from a single-root virtual hierarchy to a particular virtual link in the plurality of the virtual links based on the first mapping data structure and second mapping data structure.

四、指定代表圖：

(一)本案指定代表圖為：第（17）圖。

(二)本代表圖之元件符號簡單說明:

1710-1770　步驟流程

五、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

六、發明說明：

【發明所屬之技術領域】

本發明一般係關於改良式資料處理系統以及方法，尤其是本發明係關於根據訊務類型與伺服器刀鋒型目的地兩者區別通過多根 PCI Express 環境內根複合體的訊務類型之系統及方法。該區別將避免一個訊務等級阻擋其他訊務等級流通過多根系統。

【先前技術】

大多數現代計算裝置都會使用到輸入/輸出(input/output，I/O)配接器以及匯流排，這些都運用到1990 年代 Intel 所制定的周邊組件互連標準(Peripheral Component Interconnect standard)的某些版本。周邊組件互連(Peripheral Component Interconnect，PCI)標準指定讓周邊裝置附加至電腦主機板的電腦匯流排。PCI Express 或 PCIe 屬於 PCI 電腦匯流排的一種實施，其使用現有的 PCI 程式設計概念，但是使用完全不同並且更快速的序列實體層通訊協定的電腦匯流排。該實體層並非由可在許多裝置之間共享的雙向匯流排，而是確實連接至兩裝置的單一單向連結所構成。

【發明內容】

在一個說明的環境內，在多根資料處理系統內提供一種區別不同訊務類型與伺服器刀鋒型目的地之方法。該方法包含以下步驟：產生一第一映射資料結構，對於該多根資料處理系統內每一單根虛擬階層，該第一映射資料結構將複數個訊務等級與複數個優先群組相關，並將該複數個訊務等級內每一訊務等級映射至複數個虛擬通道內的一對應虛擬通道。該方法另包含以下步驟：產生一第二映射資料結構，其映射該複數個虛擬通道內的每一虛擬通道至該多根資料處理系統內複數個虛擬連結內的每一對應主機刀鋒型虛擬連結。再者，該方法包含以下步驟：根據該第一映射資料結構以及第二映射資料結構，將一特定優先群組的訊務從一單根虛擬階層繞送至該複數個虛擬連結內的一特定虛擬連結。

在其他說明的具體實施例內，一電腦程式產品包含：一電腦可使用或可讀取媒體，其中提供一電腦可讀取程式。該電腦可讀取程式當在一計算裝置上執行時，會導致該計算裝置執行上述有關所說明具體實施例方法的多個操作或操作組合。

仍舊在其他說明的具體實施例內，提供一系統/設備。該系統/設備可包含一或多個處理器以及耦合至該一或多個處理器的一記憶體。該記憶體可包含指令，當在該一或多個處理器上執行時，該指令會導致該一或多個處

理器執行上述有關該所說明具體實施例方法的多個操作和操作組合。

從下列本發明示範具體實施例的實施方式當中，精通此技術的人士將瞭解本發明的這些與其他特徵及優點。

【實施方式】

周邊組件互連特殊相關群組(peripheral component interconnect special interest group，PCI-SIG)已經發展出周邊組件互連(peripheral component interconnect，PCI)和 PCI Express (PCIe)規格，指定 PCI 和 PCIe 在資料處理系統內實施之方式。第 1 圖為說明合併根據 PCIe 規格的 PCI Express (PCIe)結構拓撲之系統範例圖。如第 1 圖內所示，系統 100 包含一主機處理器(CPU) 110 以及耦合至根複合體 130 的記憶體 120，其中該根複合體依序耦合至一或多個 PCIe 端點 140 (PCIe 規格內使用術語「端點」代表 PCIe 啟用的 I/O 配接器)、PCI Express 對 PCI 橋接器 150 以及一或多個互連交換器 160。根複合體 130 代表將 CPU/記憶體連接至 I/O 配接器的 I/O 階層根部。根複合體 130 包含一主機橋接器、沒有或多個根複合體整合端點、沒有或多個複合體事件收集器以及一或多個根連接埠。每一根連接埠都支援一個別 I/O 階層。I/O 階層可包含一根複合體 130、沒有或多個互連交換器 160

以及/或橋接器150(包含一交換器或PCIe結構)和一或多個端點，像是端點140、170和182-188。有關PCI和PCIe的更多資訊，請參閱可從PCI-SIG網站中取得的PCI和PCIe規格，網址為www.pcisig.com。

除了PCI和PCIe規格以外，PCI-SIG也定義輸入/輸出虛擬化(input/output virtualization，IOV)標準，用於定義如何設計可由許多邏輯分割(logical partition，LPAR)共享的I/O配接器(I/O adapter，IOA)。在單根系統上共享IOA稱為單根IO虛擬化(Single Root IO Virtualization，SR-IOV)，橫跨多根系統(例如刀鋒型系統)共享的IOA稱為多根IO虛擬化(Multi-root IO Virtualization，MR-IOV)。LPAR為將電腦處理器、記憶體和儲存裝置分成多個資源集合的分區，如此每一資源集合都可用自己的作業系統實體與應用程式獨立操作。可建立的邏輯分割數量取決於系統的處理器模型以及可用的資源，一般來說，分割用於不同目的，像是資料庫操作、主從式操作，以分開測試與生產環境等。每一分割都可與其他分割通訊，如同其他分割在分開的機器內。在支援LPAR的現代系統中，某些資源可在LPAR之間共享。如上面所提及，在PCI和PCIe規格中，一種可共享的資源為使用I/O虛擬化機制的I/O配接器。

雖然PCI-SIG提供標準來定義如何設計可由SR-IOV

或 MR-IOV 環境內許多 LPAR 共享的 IOA，此規格並未定義如何避免佇列前端阻擋在多根系統的個別虛擬階層內所有訊務中的不同訊務等級與不同主機刀鋒。PCIe 多根 I/O 虛擬化(Multi-root I/O Virtualization，MR-IOV)規格包含虛擬通道至虛擬連結映射的細節，來建立通過系統的多獨立資料流與資源。不過，MR-IOV 規格並不提供如何避免佇列前端阻擋不同訊務等級通過虛擬階層的虛擬連結之細節。MR-IOV 規格陳述「一般來說，MR-PCIM 事先並不知道 VH 內的軟體操作將使用哪個 VC ID 或如何配置。在 MR-PCIM 處理此確認或其中 MR-PCIM 可將所要配置傳遞至 VH 內所操作軟體之系統內，VC ID 可在 MR-PCIM 初始化時(即是 VH 內所操作軟體的實體化之前)映射至 VL。」如此，該規格並不提供如何避免佇列前端阻擋通過虛擬連結內之虛擬階層(virtual hierarchies，VH)的任何方法。該規格說明可幫助避免佇列前端阻擋的旁通佇列，但是最終這些旁通佇列將耗盡緩衝區資源。這導致佇列前端阻擋以及由於使用旁通佇列造成訊務佇列處理內的額外延遲。

所說明的具體實施例定義將使用一訊務等級通過虛擬階層到達虛擬通道的訊務區別成虛擬連結映射能力之機制，在此虛擬連結提供多重獨立邏輯資料流通過單實體多根(Multi-Root，MR) PCIe 連結之支援，並且在 MR 拓

撲中扮演與 PCIe 基本拓撲內之虛擬通道(VC)相同的角色。虛擬通道可在 PCI Express 單根階層內建立多重獨立資料流。虛擬連結可在 PCI Express 多根階層內建立多重獨立資料流。多根系統的每一虛擬階層都可指派單一虛擬通道給虛擬連結。多訊務類型共享單一虛擬連結會導致佇列前端阻擋，例如：儲存訊務會阻擋虛擬連結上的高效能計算(high performance computing，HPC)訊務，即是與超級電腦和叢集相關的訊務。進一步，儲存訊務可來自與 HPC 訊務不同的虛擬階層。如此具有較長傳輸時間的訊務會阻擋需要較低延遲的訊務。阻擋虛擬連結的較慢訊務可來自不同的虛擬階層，其導致一個系統的工作阻擋目標為其他系統的訊務。說明的具體實施例定義指派優先群組給訊務等級、虛擬通道以及虛擬連結的機制，以避免像是儲存訊務這類較慢訊務阻擋對於延遲比較敏感的訊務，像是 HPC 應用訊務。

精通此技術的人士將瞭解，本發明可具體實施為系統、方法或電腦程式產品。因此，本發明可為完整硬體具體實施例、完整軟體具體實施例(包含韌體、常駐軟體、微代碼等)或軟體與硬體態樣的組合具體實施例之形式，在此通稱為「電路」、「模組」或「系統」。更進一步，本發明可採用具有媒體內具體實施之電腦可使用程式碼的任何實質媒體內具體實施之電腦程式產品之形式。

任何一或多個電腦可使用或電腦可讀取媒體的組合都可利用。電腦可使用或電腦可讀取媒體可為例如但不受限於電、磁、光學、電磁、紅外線或半導體系統、設備、裝置或傳播媒體。電腦可讀取媒體的更多特定範例(非窮舉清單)包含下列：具有一或多條線的電連接、可攜式電腦磁碟片、硬碟、隨機存取記憶體(random access memory，RAM)、唯讀記憶體(read-only memory，ROM)、可抹除可程式唯讀記憶體(erasable programmable read-only memory(EPROM)或快閃記憶體)、光纖、可攜式光碟唯讀記憶體(compact disc read-only memory，CDROM)、光學儲存裝置、像是支援網際網路或企業內部網路的這些傳輸媒體或磁性儲存裝置。請注意，電腦可使用或電腦可讀取媒體可為上面列印程式的紙張或其他合適媒體，而該程式可透過例如紙張或其他媒體上的光學掃描以電子方式擷取，然後若有需要以合適的方式組譯、解釋、或處理，然後儲存在電腦記憶體內。在本文件的內容中，電腦可使用或電腦可讀取媒體可為可以包含、儲存、通訊、傳播或傳輸程式，來讓指令執行系統、設備或裝置使用或相連之任何媒體。在基頻或部分載波內，電腦可使用媒體可包含其內具體實施電腦可使用程式碼的傳播資料訊號。電腦可使用程式碼可使用任何適當的媒體來傳輸，該媒體包含但不受限於無線、有

線、光纖纜線、RF等。

執行本發明操作的電腦程式碼可用任何一或多種程式語言的組合來撰寫，包含像是 Java、Smalltalk、C++等物件導向程式語言，以及像是「C」程式語言或類似程式語言的傳統程序程式語言。程式碼可完全在使用者的電腦上、部分在使用者的電腦上、作為單機套裝軟體、部分在使用者的電腦上並且部分在遠端電腦上或完全在遠端電腦或伺服器上執行。在稍後的案例中，遠端電腦可透過任何網路，包含區域網路(local area network，LAN)或廣域網路(wide area network，WAN)，連接至使用者的電腦，或與外部電腦連線(例如透過網際網路系統服務商提供的網際網路)。

底下參考根據本發明說明具體實施例的方法、設備(系統)和電腦程式產品之流程圖以及/或方塊圖，來說明具體實施例。吾人將瞭解，電腦程式指令可實施流程圖及/或方塊圖的每一方塊以及流程圖及/或方塊圖內方塊的組合。這些電腦程式指令可提供給一般用途電腦、特殊用途電腦或其他可程式資料處理設備的處理器以產生一機器，如此透過電腦或其他可程式資料處裡設備的處理器所執行之指令，建立用於實施流程圖及/或方塊圖方塊內所指定功能/步驟之構件。

這些電腦程式指令也可儲存在電腦可讀取媒體內，指

引電腦或其他可程式資料處理設備以特定方式運作，如此儲存在電腦可讀取媒體內的指令產生製造物品，包含實施流程圖及/或方塊圖方塊內所指定功能/步驟之構件。

電腦程式指令也可載入電腦或其他可程式資料處理設備，導致在電腦或其他可程式設備上執行一連串操作步驟來產生電腦實施的處理，如此在電腦或其他可程式設備上執行的指令提供用於實施流程圖及/或方塊圖方塊內所指定功能/步驟之處理。

圖式內的流程圖和方塊圖說明根據本發明許多具體實施例的系統、方法和電腦程式產品可能實施之架構、功能和操作。如此，流程圖或方塊圖內每一方塊都可代表模組、區段或程式碼部分，這部分程式碼可包含一或多個可執行指令來實施特定邏輯功能。吾人也應該注意，在某些替代實施當中，方塊內提到的功能可以不依照圖式內順序來執行。例如：兩連續顯示的方塊實際上可同時執行，或可以顛倒順序執行，這取決於所牽涉到的功能。吾人也注意，使用執行特殊功能或步驟的特殊用途硬體式系統或特殊用途硬體與電腦指令的組合，實施方塊圖及/或流程圖的每一方塊以及方塊圖及/或流程圖內方塊的組合。

因為說明的具體實施例定義使用虛擬通道通過虛擬階層到達虛擬連結映射以區別訊務類型之機制，以便瞭解

所說明具體實施例的機制，所以最重要是先瞭解如何利用管理程序或其他虛擬化平台實施 I/O 虛擬化。吾人應該瞭解，雖然說明關於周邊組件互連 Express (Peripheral Component Interconnect Express，PCIe)配接器或端點的說明性具體實施例，但本發明並不受限於此。而是所說明具體實施例的機制可用支援 I/O 配接器內 I/O 虛擬化的任何 I/O 結構實施。再者吾人應該瞭解，雖然將用其中使用管理程序(hypervisor)的實施來描述所說明的具體實施例，但本發明並不受限於此。相較之下，在不悖離本發明精神與範疇之下可使用管理程序以外的其他種虛擬平台(用目前已知或稍後開發的軟體、硬體或軟硬體的任何組合來實施)。

第 2 圖為說明業界一般已知的系統虛擬化之範例圖。系統虛擬化為實體系統處理器、記憶體、I/O 配接器、儲存裝置以及其他資源的區別，如此每一資源集合都可用自己的系統映像實體與應用程式獨立操作。在這種系統虛擬化內，虛擬資源由實體資源構成並且作為具有相同外部介面與功能的實體資源代理來操作，例如記憶體、磁碟機以及具有架構介面/功能的其他硬體組件。系統虛擬化通常運用建立虛擬資源的虛擬層，並且將他們映射至實體資源，讓虛擬資源之間隔離。虛擬層通常作為軟體、韌體與硬體機制中之一個或組合。

如第2圖內所示，通常在虛擬化系統內，應用程式210與屬於軟體組件的系統映像(system image，SI) 220通訊，該應用程式像是一般或特殊作業系統，利用其指派特定虛擬與實體資源。系統映像220相關於虛擬系統230，該系統由執行例如虛擬化處理器、記憶體、I/O配接器、儲存裝置等的單一SI實體所需之實體或虛擬化資源所構成。

透過使用虛擬系統230，系統映像220藉由虛擬化層240存取實體系統資源250。虛擬化層240管理資源到SI的配置，並隔離指派給SI的資源以免遭其他SI存取。此配置與隔離通常根據虛擬化層240所執行的資源映射以及虛擬化層240所維護的一或多個資源映射資料結構來執行。

這種虛擬化可用於I/O操作與I/O資源的虛擬化。也就是關於I/O虛擬化(I/O virtualization，IOV)，超過一個SI可使用部分或完整實施為管理程序的虛擬化層240來共享單一實體I/O單元。管理程序可為軟體、韌體等，其利用介入例如SI的一或多個組態、I/O和記憶體操作以及完成直接記憶體存取(direct memory access，DMA)和中斷SI操作，用來支援IOV。

第3圖為說明使用虛擬化層將PCI根複合體的I/O虛擬化之第一方式範例圖。如第3圖內所示，可為晶片、

主機板、刀鋒等等當中一或多個處理器的主機處理器集合 310 可支援複數個系統映像 320-330，應用程式(未顯示)可透過此來存取系統資源，像是 PCIe 端點 370-390。系統映像透過虛擬化層 340、PCIe 根複合體 350、一或多個 PCIe 交換器 360 以及/或其他 PCIe 結構元件與虛擬化資源通訊。

運用第 3 圖內說明的方式，可部分或全部實施為管理程序或其他種虛擬化平台的虛擬化層 340 都牽涉在所有 I/O 交易內，並且執行所有 I/O 虛擬化功能。例如：虛擬化層 340 將來自許多 SI 中 I/O 佇列的 I/O 要求多工至 PCIe 端點 370-390 內的單一佇列，如此虛擬化層 340 作為 SI 320-330 與實體 PCIe 端點 370-390 之間的代理層。

第 4 圖為說明使用本質上共享的 PCI I/O 配接器，以將 PCI 根複合體的 I/O 虛擬化之第二方式範例圖。如第 4 圖內所示，可為晶片、主機板、刀鋒等等當中一或多個處理器的主機處理器集合 410 可支援複數個系統映像 420-430，應用程式(未顯示)可透過此來存取系統資源，像是 PCIe I/O 虛擬化(IOV)端點 470-490。系統映像 420-430 透過 PCIe 根複合體 440、一或多個 PCIe 交換器 460 以及/或其他 PCIe 結構元件與虛擬化資源通訊。

PCIe 根複合體 440 包含根複合體虛擬化啟用器(root complex virtualization enabler，RCVE) 442，其可包含一

或多個位址轉譯和保護表資料結構、中斷表資料結構等等，幫助使用 IOV 啟用端點 470-490 的 I/O 操作虛擬化。PCIe 根複合體 440 可使用位址轉譯與保護表資料結構執行例如虛擬化資源的虛擬與實際位址間之位址轉譯、根據虛擬資源對 SI 的映射來控制對虛擬資源的存取，以及其他虛擬化操作。這些根複合體中斷表資料結構可透過 PCIe 記憶體位址空間存取，並且用於例如將中斷映射至與 SI 相關的適當中斷處置器。

第 3 圖內所示的方式，第 4 圖的虛擬化結構內也提供虛擬化層 450。虛擬化層 450 與可耦合至 PCIe 交換器 460 的非 IOV 啟用 PCIe 端點搭配使用。也就是，可部分或全部實施為管理程序或其他虛擬化平台的虛擬化層 450 針對非本質上(即是端點內部)支援 I/O 虛擬化(IOV)的這些 PCIe 端點，用和上述關於第 3 圖的類似方式來運用 PCIe 端點。

針對 IOV 啟用的 PCIe 端點 470-490，虛擬化層 450 主要用於組態交易目的，不牽涉到記憶體位址空間操作，像是來自 SI 或直接記憶體存取(DMA)操作的記憶體映射輸入/輸出(memory mapped input/output，MMIO)操作從 PCIe 端點 470-490 開始。相較之下，直接執行 SI 420-430 與端點 470-490 之間的資料傳輸，虛擬化層 450 並不介入。SI 420-430 與端點 470-490 之間的直接 I/O 操作可由

RCVE 442 和 IOV 啟用的 PCIe 端點 470-490 內建的 I/O 虛擬化邏輯，例如實體與虛擬功能，來進行。執行直接 I/O 操作逐漸增加 I/O 操作所執行速度的能力，但是需要 PCIe 端點 470-490 支援 I/O 虛擬化。

第 5 圖為 PCIe I/O 虛擬化(IOV)啟用端點的範例圖。如第 5 圖內所示，PCIe IOV 端點 500 包含透過此與 PCIe 交換器等通訊的 PCIe 連接埠 510，並且可執行 PCIe 結構。內部繞送 520 提供通訊通道給組態管理功能 530 以及複數個虛擬功能(virtual function，VF) 540-560。組態管理功能 530 可為相對於虛擬功能 540-560 的實體功能(physical function，PF)。如 PCI 規格內所使用的術語，實體「功能」為單一組態空間所表示的邏輯集合。換言之，實體「功能」為可根據記憶體內功能的相關組態空間內所儲存之資料組態之電路邏輯，像是例如非可分離資源 570 內所提供者。

組態管理功能 530 可用來組態虛擬功能 540-560。在 I/O 虛擬啟用端點內，虛擬功能為共享一或多個實體端點資源(例如連結)的功能，並且 PCIe IOV 端點 500 的共享資源佇池 580 內可提供該功能，例如與其他功能一起。管理程序未在執行時間時介入，虛擬功能可直接成為 I/O 的佇槽，自系統映像的記憶體操作以及來自直接記憶體存取(DMA)的來源，以及對於系統映像(SI)的中斷操作。

PCIe配接器/端點可具有許多不同種組態，其關於PCIe配接器/端點所支援的「功能」。例如：端點可支援單一實體功能(PF)、多重獨立PF或甚至多重相依PF。在支援本質I/O虛擬化的端點內，端點所支援的每一PF都可相關於一或多個虛擬功能(VF)，功能本身可取決於與其他PF相關的VF。此後將用第6圖和第7圖說明實體與虛擬功能之間的範例關係。

第6圖為說明單根端點在無本質虛擬化時實體與虛擬功能之範例圖。術語「單根端點」代表相關於單根結點的單根複合體之端點，即是單主機系統。運用單根端點，端點可由相關於單根複合體的複數個系統映像(SI)所共享，但是不在相同或不同根節點上複數個根複合體之間共享。

如第6圖內所示，根節點600包含複數個系統映像610、612，其與PCIe端點670-690、I/O虛擬化中間體630(如先前所述)、PCIe根複合體640以及一或多個PCIe交換器650及/或其他PCIe結構元件。根節點600進一步包含一單根PCIe組態管理(single root PCIe configuration management，SR-PCIM)單元620。SR-PCIM單元620負責管理PCIe結構，其包含根複合體640、一或多個PCIe交換器650等以及端點670-690。SR-PCIM 620的管理責任包含決定哪個功能指派給哪個SI 610、

612 以及設定端點 670-690 的組態空間。SR-PCIM 620 可根據 SI 能力以及來自使用者，像是系統管理員，或負載平衡軟體的輸入組態許多端點 670-690 的功能，以及哪個資源已指派給哪個 SI 610、612。SI 的能力可包含許多因素，包含有多少位址空間可配置給端點 670-690、有多少中斷可用於指派給端點 670-690 等。

每一 PCIe 端點 670-690 可支援一或多個實體功能(PF)。一或多個 PF 可彼此獨立，或在某些方式內彼此相依。根據供應商定義的功能從屬，PF 可與其他 PF 相依，其中一個 PF 需要其他 PF 的操作或其他 PF 所產生的結果，例如為了正確操作。在說明的範例中，PCIe 端點 670 支援單 PF 並且 PCIe 端點 680 支援複數個不同類型 1 至 M 的獨立 PF，即是 PF_0 至 PF_N。類型係關於 PF 或 VF 的功能性，例如乙太網路功能以及光纖通道功能為兩種不同功能類型。端點 690 支援不同類型，其中二或多個 PF 相依的多個 PF。在說明的範例中，PF_0 視 PF_1 而定，或反之。

在第 6 圖顯示的範例中，系統映像(SI) 610 和 612 透過由 I/O 虛擬化中間體(I/O virtualization intermediary，IOVI) 630 讓虛擬化機制變成可用，來共享端點 670-690。如先前所說明，在這種配置內，IOVI 630 牽涉在 SI 610、612 與 PCIe 端點 670-690 之間的所有 PCIe

交易內。個別 PCIe 端點 670-690 本身並不需要支援虛擬化，因為處理虛擬化的負荷都完全加諸在 IOVI 630 上。結果，雖然這種配置內可使用虛擬化的已知機制，不過若 IOVI 630 並未牽涉到每一 I/O 操作，則 I/O 操作可執行的速率相較於潛在的 I/O 速率來說相對較慢。

第 7 圖為說明單根端點起用本質 I/O 虛擬化時實體與虛擬功能之範例圖。第 7 圖內顯示的配置類似於第 6 圖的配置，不過某些重要差異在於 PCIe 端點 770-790 原本(即是在端點本身內)就支援 I/O 虛擬化(IOV)。結果，可有效消除第 6 圖內的 I/O 虛擬化中間體 630，當然組態操作除外，有關 IOV 啟用的 PCIe 端點 770-790。不過，若配置內也使用非 IOV 啟用的 PCIe 端點(未顯示)，例如舊端點，則 I/O 虛擬化中間體可搭配第 7 圖內顯示的元件使用，來處理系統映像 710 與 712 之間這種非 IOV 啟用 PCIe 端點的共享。

如第 7 圖內所示，IOV 啟用 PCIe 端點 770-790 可支援一或多個獨立或相依實體功能(PF)，其輪流相關於一或多個獨立或相依虛擬功能(VF)。在此範疇內，SR-PCIM 720 使用 PF 管理 VF 集合，也用於管理端點功能，像是實體錯誤與事件。與 PF 相關的組態空間定義 VF 的能力，包含相關於 PF 的 VF 最大數量、PF 與 VF 與其他 PF 與 VF 的組合等等。

SI 使用 VF 存取 IOV 啟用 PCIe 端點 770-790 上的資源，例如記憶體空間、佇列、中斷等等。如此，產生不同 VF 給要共享特定 PF 的每一 SI 710、712。根據對應 PF 的組態空間內 SR-PCIM 720 所設定的 VF 數量，由端點 770-790 產生 VF。在此方式中，PF 虛擬化，如此可由複數個 SI 710、712 共享。

如第 7 圖內所示，VF 和 PF 視其他 VF 和 PF 而定。通常，若 PF 為相依的 PF，然後相關於 PF 的所有 VF 也將為相依。如此例如 PF_0 的 VF 可視對應的 PF_1 之 VF 而定。

如第 7 圖內顯示的配置，SI 710、712 可透過 PCI 根複合體 730 和 PCIe 交換器 740 直接與 IOV 啟用的 PCIe 端點 770-790 通訊，反之亦然，而不需要牽涉到 I/O 虛擬中間體。這種直接通訊由端點 770-790 內以及 SR-PCIM 720 內提供的 IOV 支援所進行，其設定端點 770-790 內的 PF 和 VF。

逐漸增加 SI 與端點之間的直接通訊速度，其上可執行複數個 SI 710、712 與共享 IOV 啟用的 PCIe 端點 770-790 之間的 I/O 操作。不過，為了進行這種效能強化，PCIe 端點 770-790 必須利用提供 SR-PCIM 720 內的機制以及用於產生與管理虛擬功能(VF)的端點 770-790 之實體功能(PF)，來支援 I/O 虛擬化。

上述 PCIe 階層受限為單根階層。換言之，PCIe 端點

只能由與單 PCI 根複合體 730 相關的單根節點 700 上之 SI 710、712 共享。上述機制並不支援多根複合體共享 PCIe 端點。如此，多根節點無法提供 PCIe 端點資源的共享存取。因為每一根節點需要個別端點集合，所以限制了運用這種配置的系統之擴充性。

此處所說明的具體實施例可運用多根 I/O 虛擬化，其中多 PCI 根複合體可共享存取相同的 IOV 啟用 PCIe 端點集合。結果，與每一這種 PCI 根複合體相關的系統映像每一都可共享相同的 IOV 啟用 PCIe 端點資源集合的存取，但是針對每一根節點上每一 SI 安置虛擬化保護。如此，利用提供允許增加根節點以及對應 PCI 根複合體的機制，其可共享相同現有 IOV 啟用 PCIe 端點集合，將擴充性最大化。

第 8 圖為說明根據一個說明具體實施例的多根虛擬化 I/O 拓撲之範例圖。如第 8 圖內所示，提供複數個根節點 810 和 820，其中每一根節點具有單根 PCI 組態管理員 (SR-PCIM) 812、822、一或多個系統映像 (SI) 814、816、824 和 826 以及 PCI 根複合體 818 和 828。這些例如刀鋒型伺服器內刀鋒的根節點 810 和 820 可耦合至 PCIe 交換器結構的一或多個多根覺醒 (multi-root aware，MRA) PCIe 交換器 840，其中該結構可包含一或多個這種 MRA PCIe 交換器 840 及/或其他 PCIe 結構元件。MRA 交換器

840 與第 7 圖內的非 MRA 交換器 740 不同，因為 MRA 交換器 840 具有額外根節點的連接，並且內含維持這些不同根節點的位址空間分隔與分開所需的機制。

除了這些根節點 810 和 820 以外，提供第三根節點 830，其包含多根 PCI 組態管理員(MR-PCIM) 832 和對應的 PCI 根複合器 834。MR-PCIM 832 負責發現並組態第 8 圖內所示多根(MR)拓撲內的虛擬階層，此後將有更詳細討論。如此 MR-PCIM 832 就多根節點的多根複合體，組態端點的實體與虛擬功能。SR-PCIM 812 和 822 組態其相關單根複合體的實體與虛擬功能。換言之，MR-PCIM 將 MR 拓撲看待成一個整體，而 SR-PCIM 只看見 MR 拓撲內自己擁有的虛擬階層，此後將會更詳細說明。

如第 8 圖內所示，IOV 啟用 PCIe 端點 850 和 860 支援一或多個虛擬端點(VE) 852、854、862 和 864。VE 為指派給根複合體的實體與虛擬功能集合，如此例如在 IOV 啟用 PCIe 端點 850 和 860 上提供個別 VE 852 和 862 給根節點 810 的 PCI 根複合體 818。類似地，在 IOV 啟用 PCIe 端點 850 和 860 上提供個別 VE 854 和 864 給根節點 820 的 PCI 根複合體 828。

每一 VE 都指派給虛擬階層(VH)，該階層具有作為 VH 根部的單根複合體並且 VE 作為階層內的終止節點。VH 為完整功能的 PCIe 階層，其已經指派給根複合體或

SR-PCIM。吾人應該注意，VE內所有實體功能(PF)和虛擬功能(VF)都已經指派給相同VH。

每一IOV啟用PCIe端點850和860都支援基本功能(base function，BF)859和869。BF 859、869為MR-PCIM 832用來管理對應端點850、860的VE之實體功能。例如：BF 859、869負責指派功能給對應端點850、860的VE。MR-PCIM 832使用BF組態空間內允許指派VH編號給端點850、860內每一PF的欄位，將功能指派給VE。在說明的具體實施例內，每一端點只有一個BF，不過本發明並不受限於此。

如第8圖內所示，每一VE 852、854、862和864都支援自己的實體與虛擬功能集合。如先前所說明，這種函數集合可包含獨立實體功能、相依實體功能以及其相關的獨立/相依虛擬功能。如第8圖內所示，VE 852支援具有其相關虛擬功能(VF)的單一實體功能(PF_0)。VE 854同樣支援具有其相關虛擬功能(VF)的單一實體功能(PF_0)。VE 862支援複數個獨立實體功能(PF_0-PF_N)以及其相關虛擬功能(VF)。不過，VE 864支援複數個獨立實體功能(PF_0-PF_N)。

若及只有若VE指派至SI已經存取的VH，VE 852、854、862或864可直接與根節點810和820的SI 814、816、824和826通訊，反之亦然。端點850和860本身

必須支援單根 I/O 虛擬化，像是先前所說明，以及多根 I/O 虛擬化，如關於本說明具體實施例所說明。此需求係根據拓撲支援多根複合體但是每一個別根節點只看見其相關單根式虛擬階層之事實。

第 9 圖為說明根據一個說明具體實施例從根節點的根複合體觀點來看，多根虛擬化 I/O 拓撲的虛擬階層圖之範例圖。如第 9 圖內所示，雖然多根(MR)拓撲可能如第 8 圖內所示，不過每一個別根節點的每一根複合體只看見其 MR 拓撲的部分。如此例如相關於根節點 810 的 PCI 根複合體 818 瞭解其主機處理器集合、自己的系統映像(SI) 814、816、MRA 交換器 840 以及自己的虛擬端點(VE) 852 和 862。在此虛擬階層內有完整 PCIe 功能性，不過 PCI 根複合體 818 不瞭解不屬於其虛擬階層部分的 VE、根複合體、系統映像等。

第 10 圖為其中根據一個說明具體實施例運用 IOV 啟用端點或配接器的系統結構之範例圖。第 10 圖內所示的機制可與第 4 圖內說明的機制結合實施。例如：可提供可為單根(SR)或多根(MR) PCIM 之第 10 圖內顯示的 PCI 管理員(PCI manager，PCIM) 1003，來與第 4 圖內系統映像 1 420 相關，而提供第 10 圖內的用戶端分割 1004 來與第 4 圖內系統映像 2 430 相關。類似地，第 10 圖的 I/O 結構 1011 可包含第 4 圖內的 PCIe 交換器 460，IOV

端點 1014 可類似於第 4 圖內任一 PCIe 端點 470-490 並且端點 1015 和 1016 可為 IOV 啟用端點或非 IOV 啟用端點，像是第 3 圖內的端點 370-390。

如第 10 圖內所示，系統 1001 包含可為任何資料處理裝置的主機系統 1026，例如伺服器、用戶端計算裝置等、I/O 結構 1011 (例如 PCIe 結構)，其可包含一或多個通訊連結及一或多個交換器，以及一或多個 I/O 端點 1014-1016，其在一個說明的具體實施例內可為 PCIe I/O 端點，I/O 端點 1014 為 I/O 虛擬化(IOV)啟用端點，而其他端點 1015-1016 為 IOV 啟用或非 IOV 啟用端點。主機系統 1026 包含平台硬體 1010，其為資料處理裝置的硬體、管理程序 1025、邏輯分割(LPARS) 1003 和 1004 以及對應的分割韌體(partition firmware，PFW) 1023 和 1024。雖然此處就使用管理程序 1025 來描述說明的具體實施例，吾人應該瞭解在不悖離先前提及本發明精神與範疇之下，可運用其他種虛擬化平台。

在一個說明的具體實施例內，管理程序 1025 可為在平台硬體 1010 上執行的程式碼，並且屬於平台韌體的一部分。類似地，分割韌體(PFW) 1023-1024 也可為平台韌體的一部分，但是因為考慮到其邏輯上屬於 LPAR 內所執行 LAPR 程式碼的一部分，所以顯示成相關於 LPAR 1003 和 1004。

LPAR 1003 和 1004 具有已配置的資源以及在 LPAR 內執行的作業系統映像或實體。此外，LPAR 1003 和 1004 可執行 LPAR 內其他應用程式、軟體、程式碼等。例如：對於所說明具體實施例的一個 LPAR，例如 LPAR 1003，特別重要，執行讓 LPAR 1003 操作成為單根通訊結構管理員，例如 SR-PCIM，或成為多根通訊結構管理員，例如 MR-PCIM 1003 (此後簡稱為「PCIM」)，的程式碼。其他 LPAR 1004 可操作成為用戶端分割。雖然第 10 圖內只顯示一個 PCIM 1003 和一個用戶端分割 1004，吾人應該瞭解，在不悖離所說明具體實施例的精神與範疇之下，主機系統 1026 內可提供超過一個 PCIM 1003 和用戶端分割 1004。

管理程序 1025 已經存取至 IOV 端點 1014 的組態空間 1019、1021 以及 I/O 結構 1011 組態空間 1017。此處所用的術語「組態空間」代表來自記憶體映射中 I/O (memory mapped I/O，MMIO)位址空間之散開位址空間，為 I/O 配接器上的記憶體，由主機作業系統映射用於主機作業系統的可定址性，其可配置用於儲存用於系統 1001 的特定組件之組態資料。進一步，PCIM 的作業系統 1031 和裝置驅動器 1005 在指派給 PCIM 1003 時存取至實體功能(PF) 1012 的組態空間 1019，以及存取至屬於已指派給 PCIM 1003 中 PF 的虛擬功能(VF)之組態空間

1021。

管理應用程式 1040 位於硬體管理控制台(Hardware Management Console，HMC) 1009 上，該控制台位於主機系統 1026 上或個別資料處理裝置之內(如所示)，HMC 1009 本身通訊通過遠端管理指令(Remote Management Command，RMC)介面 1002 到 PCIM 1003 和用戶端分割 1004，並且通過相同介面 1020 到管理程序 1025。管理應用程式 1040 (此後將 HMC 1009 的集合通稱為 HMC 1009)作為指揮者來控制存取系統 1001 內許多組件的功能，並且提供使用者介面 1042 讓人檢視系統組態，並且輸入要將哪些資源指派給哪些 LPAR 1003-1004 之資訊。管理應用程式 1040 可提供許多不同功能，這些功能可由使用者引動，此後將會更詳細說明。另外，不用使用者就可自動引動這些功能來回應觸發這些功能啟動的事件或輸入。

如此後將根據所說明具體實施例來說明，某些這些功能包含映射資料結構之產生或建立，來根據優先群組從訊務等級映射至虛擬通道以及從虛擬通道映射至虛擬連結。再者，如稍後所討論，這些功能可進一步包含組態這些映射的通訊結構管理員功能，如此可適當繞送許多訊務等級和優先群組的資料至適當虛擬連結，避免佇列前端阻擋。

如上述，虛擬通道可在 PCI Express 單根階層內建立多重獨立資料流，而虛擬連結可在 PCI Express 多根階層內建立多重獨立資料流。多根系統的每一虛擬階層都可指派單一虛擬通道給虛擬連結。不過，多訊務類型共享單一虛擬連結會導致佇列前端阻擋，如此具有較長傳輸時間的訊務會阻擋需要較低延遲的訊務。說明的具體實施例定義指派優先群組給訊務等級、虛擬通道以及虛擬連結的機制，以避免較慢訊務阻擋對於延遲比較敏感的訊務。

第 11 圖為說明建立阻擋通過 MR-IOV 系統的佇列前端(head of line，HOL)之虛擬連結範例圖。MR-IOV 規格可讓虛擬通道(VC)映射至虛擬連結(VL)，如第 11 圖內所說明。例如，如第 11 圖內所示，計算系統 1105 的第一計算裝置 1110，在說明範例中為刀鋒型伺服器 1105 內的刀鋒 1110，相關於例如圖式內顯示為 VH1 的第一虛擬階層(VH)的根節點。第二計算裝置 1120 相關於第二虛擬階層(VH2)的根節點，第三計算裝置 1130 相關於第三虛擬階層(VH3)的根節點。不同虛擬階層可相關於不同虛擬通道(VC0-VC1)。每一虛擬連結(VL)，即是 VL0 和 VL1，可通訊資料通過與許多虛擬階層(VH)相關的這些 VL。

吾人可瞭解，在未定義每一 VL 的優先群組之下，VH VC 配對的任意放置會導致第 11 圖內所示的組態。也就

是，不同虛擬階層的計算裝置 1110-1130 可傳輸資料，即是訊務，其繞送通過虛擬連結(VL)，該連結提供 MR PCIe 交換器與計算裝置 1110-1130 之間的邏輯連線。每一 VL，即是 VL0 和 VL1，可通訊資料通過與許多虛擬階層(VH)相關的這些 VL。結果如方塊 1150 內所示，來自 VH1 的資料通過 VC0，其可為需要更多處理器週期來處理或在處理時效性上具有比其他資料慢的依存性之較慢訊務，可繞送通過相同虛擬連結，例如 VL0，而較快訊務需要相對較少數處理器週期來處理，例如通過 VH3 VC1 的資料訊務。也就是，例如 HPC 訊務可根據短延遲或回應(即是較快訊務)產生小封包，而儲存的訊務可為無法容許損耗行為的較大區塊資料(即是較慢訊務)。每一訊務等級的簽署完全取決於應用環境，例如：HPC 資料可為較大區塊，而儲存訊務可為較小區塊。

較慢訊務和較快訊務可繞送過相同虛擬連結的事實會導致在虛擬連結(VL)的緩衝區內，之前的較慢訊務 VH1 VC0 佇列前端(HOL)阻擋較快訊務 VH3 VC1。此外，一個系統或計算裝置，例如相關於第一虛擬階層 VH1 的刀鋒 1110，本質上會阻擋其他系統或計算裝置的訊務，例如相關於第三虛擬階層 VH3 的刀鋒 130。若發生 HOL 阻擋時可發出旁通佇列，不過旁通佇列資源受到限制並且會耗盡，最終導致 HOL 阻擋。旁通佇列也可導致實施的

系統顯著效能負擔。

第 12 圖為一個說明具體實施例的範例圖，其中訊務等級映射在每一主機刀鋒的虛擬連結上，以避免佇列前端阻擋通過 MR-IOV 系統。優先群組相關於虛擬通道可利用在自己的優先群組內放入較快訊務，並且將這些優先群組相關於虛擬連結(VL)，以避免佇列前端阻擋。這避免一個優先群組內較慢訊務阻擋其他優先群組內更快訊務。此外，當具有不同效能目標時，避免一個主機系統或虛擬階層(VH)阻擋其他主機系統或虛擬階層(VH)。如此允許說明的具體實施例根據其優先並且也根據訊務要前往的系統刀鋒，來區別訊務。在說明的具體實施例內，若緩衝區資源擁擠，則此擁擠只會影響進入特定系統刀鋒的特定訊務優先。例如：若相關於 VH1/VC3 組合的虛擬連結(VL)變得擁擠，則只有到系統刀鋒 1 (VH1)的 HPC 訊務(VC3)變得擁擠。所提到一個重要的說明具體實施例之一個態樣為所說明具體實施例提供實施加權訊務的能力。例如：若有一個比較重要或系統內廣泛使用的訊務優先或一個特定系統刀鋒，可配置更多緩衝區資源給此訊務優先或系統刀鋒，確定進入特定刀鋒的特定訊務優先群組不會輕易造成擁擠。

運用所說明具體實施例的機制，計算裝置 1210-1230 內的 PCI 管理員 1202、1204 和 1206，可構成對應虛擬

階層的根節點，可例如依序為刀鋒伺服器系統的系統刀鋒，使用映射表程式設計，用於根據所傳輸訊務的優先，將訊務等級映射至虛擬通道，最終映射到虛擬連結。如此不同的虛擬階層可傳輸資料，即是訊務，其繞送通過許多虛擬通道(VC0-VC3)。每一 VH 的 VC 都可相關於特定優先群組，如稍後更詳細討論。VH 和 VC 的組合可輪流映射至特定虛擬連結(VL)，例如 VL1-VL12。

在說明的具體實施例內，VH 和 VC 的每一組合都可映射至個別虛擬連結。如此相同 VH 和 VC 可映射至不同訊務等級，使得不同訊務等級的訊務可透過與相同虛擬階層相關的相同虛擬通道來傳輸/接收，但是這些訊務等級必須具有相關的相同優先群組。不過，VH 和 VC 的每一組合都映射至個別虛擬連結(VL)，使得相關於相同優先群組的訊務根據訊務所相關的虛擬階層組合或系統刀鋒，以及訊務的特定訊優先(相關於虛擬通道(VC))，傳輸通過不同虛擬連結。

如第 12 圖內所示，每一虛擬連結(VL)都有自己的 VH 和 VC 相關組合。因為 VC 對應至特定優先群組，如此每一 VL 都相關於來自特定 VH 的不同訊務優先。因為 VH 相關於特定主機系統或系統刀鋒、每一 VL 都相關於特定主機系統或系統刀鋒，並且特定訊務優先相關於該主機系統或系統刀鋒。例如：虛擬連結 VL1 相關於 VH1

和 VC0，其對應至刀鋒 1210 以及「管理」優先群組。虛擬連結 VL2 相關於 VH1 和 VC0，其對應至刀鋒 1220 以及「管理」優先群組。類似地，虛擬連結 VL6 相關於 VH3 和 VC1，其對應至刀鋒 1230 以及「SAN」優先群組。虛擬連結 VL10 相關於 VH1 和 VC3，其對應至刀鋒 1210 以及「HPC」優先群組。其他虛擬連結 VL 同樣具有不同 VH 和 VC 組合，其對應至不同刀鋒 1210-1230 組合和優先群組。

來自許多計算裝置或系統 1210-1230 的訊務可根據訊務的訊務等級(TC)(用於單根 I/O 虛擬化)或虛擬通道(VC)(用於多根 I/O 虛擬化)，映射及繞送至許多虛擬連結 VL1 至 VL12。如此，根據所傳輸訊務的訊務等級，PCI 管理員 1202、1204 和 1206 可將來自計算裝置 1210 的訊務繞送至多根 PCIe 交換器 1240 的任何虛擬連結 VL1-VL12。訊務要繞送的特定虛擬連結 VL1-VL12 根據傳送資料或訊務以及訊務的訊務等級(TC)之刀鋒 1210-1230 之 VH 映射來決定。TC 映射至刀鋒 1210-1230 的 VH 之特定 VC，以及用於決定哪個 VL 要繞送訊務的 VC 和 VH 之組合。

如此例如運用單根 I/O 虛擬化，來自刀鋒 1210，即是 VH1，具有相關於「HPC」優先群組的訊務等級(TC)之訊務將由 PCI 管理員 1202 映射並繞送至虛擬通道 VC3。在

說明的範例中，VH1 和 VC3 的組合對應至虛擬連結 VL10。類似地，來自計算裝置或刀鋒 1220-1230 並具有相關於相同「HPC」優先群組的訊務等級之訊務由個別 PCI 管理員 1204 和 1206 分別映射至虛擬連結 VL11 和 VL12。同樣地，來自每一這些計算裝置或刀鋒 1210-1230 並具有映射至其他優先群組的訊務等級之訊務由 PCI 管理員 1202、1204 和 1206 映射並繞送至適當虛擬連結 VL1-VL9。

如第 12 圖內所示，PCI 管理員(PCIM) 1202-1206 包含映射模組 1250-1254 和映射表資料庫 1260-1264。映射模組 1250-1254 根據虛擬階層(VH)和訊務等級(TC)將通過 PCIM 1202-1206 的訊務映射至虛擬通道(VC)，然後將 VC 和 VH 結合映射至特定虛擬連結(VL)。映射表資料庫 1260-1264 儲存映射表資料，其指定如何從訊務等級和虛擬階層映射至虛擬通道並且最終至優先群組，反之亦然，從虛擬通道映射至虛擬連結，反之亦然等。如之後所討論，儲存在映射表資料庫 1260-1264 內的映射表資料可由硬體管理控制台(HMC)，如第 10 圖內的 HMC 1009，提供給 PCI 管理員 1202-1206。

當決定例如如何將來自計算裝置或刀鋒 1210-1230 的訊務映射並繞送至虛擬連結 VL4 至 VL7 時，映射表資料庫 1260-1264 內的映射表用來作為 PCIM 1202-1206 的映

射模組 1250-1254 所執行的映射查找操作之基礎，反之亦然。如此例如當資料從計算裝置或刀鋒 1210 上執行的處理流進 PCIM 1202 時，PCIM 1202 的映射模組 1250 決定與資料相關的虛擬階層(VH)和訊務等級(TC)。此決定可例如根據端點與交換器內的組態資料來進行，如第 15 圖和第 16 圖內所示。與組態資料相關的 PCI Express 封包具有可貼附於資料的標頭來實施，否則使用第 14 圖內所示的映射資訊，如稍後所述。此資訊使用映射表資料庫 1260 內的映射表，以決定對應於 VH 和 TC 的虛擬通道(VC)。然後使用 VC 和 VH 選擇與 VH 和 VC 組合相關的虛擬連結(VL)，資料透過此連結傳輸至 MR PCIe 交換器 1240。然後 PCIM 1250 將對應至計算裝置或刀鋒 1210 訊務的資料，透過選取的虛擬連結(VL)繞送至 MR PCIe 交換器 1240 的適當連接埠。

如此，VC 定義優先群組並且該優先群組用於管理應用層上，如此管理員可適當指派訊務等級(TC)給 VC 以及 VH/VC 組合給 VL。通常無法只使用 VC 至 VL 映射，因為 MR-PCIM 事先並不知道 VH 內的軟體操作將使用哪個 VC ID 或如何配置。在不使用管理應用程式將 VC 映射至優先群組並且機制根據該映射組態 PCI Express 端點時，如所說明具體實施例所提供，當相同 VL 上混合優先群組時會導致佇列前端阻擋(PCI Express 標準所允

許)，如上面關於第 11 圖的討論。

在一個說明具體實施例內，映射表可由特權使用者、自動機制等透過硬體管理控制台(HMC)，如第 10 圖內的 HMC 1009，程式設計進入映射表資料庫 1260-1264。第 13 圖為說明硬體管理控制台(HMC)映射單階層，例如 VH1 1320，單根 I/O 虛擬化(SR-IOV)訊務等級(TC) 1330 至虛擬通道(VC) 1310 映射，並且接著至優先群組 1340 的範例圖。優先群組 1340 跨單根與多根環境。如第 13 圖內所示，一個優先群組 1340 內所有 TC 1330 都指派給相同虛擬通道 VC 1310。如此例如針對優先群組「HPC」，訊務等級 TC7 和 TC3 指派給相同虛擬通道 VC3，針對優先群組「LAN」，訊務等級 TC6 和 TC2 指派給虛擬通道 VC2，以此類推。

如此在單一虛擬階層(VH)內，相同虛擬通道(VC)對應至相同優先群組。不過，不同訊務等級(TC)可相關於相同虛擬通道(VC)。多重虛擬階層(VH)可結合成多根 I/O 虛擬化(MR-IOV)啟用架構，如前所述。如此在 MR-IOV 環境內，會有多個第 13 圖內所示表格的副本，或這些表格可結合在單一映射表內。不過，針對每一虛擬階層，不同 TC 可相關於像是第一 VH 內的相同 VC，TC7 對應至 VC 1 並且在第二 VH 內，TC7 對應至 VC4。結果由於不同映射至不同虛擬通道，所以相同訊務等級，例如

TC7，可對應至不同虛擬階層(VH)內不同優先群組。

第 14 圖為說明根據一個說明具體實施例的硬體維護控制台將到虛擬連結的虛擬通道映射至每一主機系統刀鋒的優先群組之範例圖。此圖式顯示單根環境 VH1，如第 13 圖內所示，VH2 和 VH3 的 VC 如何映射至多根環境內的虛擬連結。再者，此圖式顯示 HMC 如何進行第 12 圖內所示多根環境的映射。此映射避免佇列前端阻擋指派給每一主機刀鋒的不同優先群組。基本上第 14 圖用於顯示若單獨管理但是有共同的 VH 和 VC 組合映射至個別 VL 時，每一 VH，例如每一實體伺服器、刀鋒等，具有自己的 TC 至 VC 映射。然而，在此有不同的映射用於每一 VH，根據特定 VH 和 VC 適當映射，如第 14 圖內所示，至個別 VL 是相當重要的，如此避免一個系統或刀鋒的佇列前端阻擋其他系統或刀鋒。

如第 14 圖內所示，在說明的範例中，每一 VH 內的 VC3 都相關於 HPC 優先群組，不過根據找尋哪個 VH，該 VH 的 HPC 優先訊務相關於不同 VL。如此例如在第 14 圖的映射資料結構內，VH1 的 VC3 訊務相關於 VL10，而 VH2 的 VC3 訊務相關於 VL11。吾人應該瞭解，VC 和 VH 的每一組合都具有個別 VL，如此不同優先的訊務在每一 VH 的不同虛擬連結上傳輸。基本上，根據訊務的系統/刀鋒和優先特定組合，例如訊務的 VC，不同優

先和不同系統或刀鋒的訊務繞送通過多根系統。這種繞送可由於映射結構和所說明具體實施例所提供機制來達成。藉由將來自不同 VH 的不同優先訊務放置在個別 VL 上，可避免不同系統上不同優先群組之間的佇列前端阻擋。

如此運用所說明具體實施例的機制，特定 VH 的 TC 映射至 VC，並且 VH/VC 的組合映射至個別 VL。如此，使用所說明具體實施例映射表的映射操作以及所說明具體實施例的機制，避免佇列前端阻擋不同優先群組和不同系統/刀鋒。因為不同優先群組和不同系統/刀鋒的所有訊務都相關於不同虛擬連結，所以避免佇列前端阻擋。結果，較低優先訊務、較高優先訊務或甚至不同系統/刀鋒的相同優先訊務都不會導致目前系統/刀鋒上訊務效能降級。

第 15 圖為說明個別 SR-IOV 系統的 TC/VC 映射表資料結構之範例圖。此 TC/VC 映射表可儲存在例如虛擬通道擴充能力結構的 VC 資源控制暫存器內，這說明於 PCI-SIG 網站內可取得的 PCI Express Base Specification，1.1 版第 428 頁內。VC 資源控制配置 VC 資源。TC/VC 映射表資料結構說明用於參考功能(虛擬或實體功能)或端點的 TC 至 VC 映射。端點可具有多個功能並且每一功能都有自己的 TC/VC 映射。實體功能指派

至單一 VH。單一功能裝置表示該端點具有一個 TC/VC 映射。

每一 SR-IOV 系統代表一個虛擬階層 VH，像是第 12 圖內所說明，並且在第 14 圖的映射表資料結構內顯示為個別輸入。在第 15 圖說明的範例中，TC/VC 映射表資料結構已經傳播資料值以代表上面第 13 圖內所說明的 TC/VC 映射。例如在第 15 圖內，VC ID 值「000」對應至 VC0 並且相關值「00010001」對應至 TC 的 4 和 0，如第 13 圖的表內所示。啟用位元讓 VC 相關於 VC ID 來使用 (有關啟用 VC 的進一步資訊，請參閱 PCI Express Base Specification，1.1 版第 428 頁)。第 13 圖和第 14 圖的實施造成在 PCI Express 端點上設定組態空間。

第 16 圖為說明具有來自第 14 圖的虛擬通道至虛擬連結映射之 MR-IOV 裝置功能表範例圖。使用 HMC 建立優先群組映射，然後傳送至 MR-PCIM。MR-PCIM 根據說明具體實施例組態具有來自第 14 圖的虛擬通道至虛擬連結映射之裝置功能表。第 16 圖表示所說明具體實施例的 VC 至 VL 映射儲存在 PCI Express 組態空間的裝置功能表內之方式，可由根據所說明具體實施例機制之 HMC 所產生。

在第 16 圖內，第一欄位表示指向 32 位元值的位元欄位，像是第 15 圖內的位元 0 至 31。如此例如位元 2:0

指向3位元值，這對應至VC0 VL映射，如第三欄位內所示。PCI SIG多根I/O虛擬化1.0規格內保留位元3 1610，但是在說明具體實施例的範圍內，此位元用來增加可定址虛擬連結的空間。也就是，使用VL提供與VH、VC配對的一對一映射，並且使用3位元指定VC0 VL映射，只有支援8 VH、VC配對。運用所說明具體實施例的機制，使用相鄰保留位元以擴充指定VC VL映射用的位元。如此例如位元3:0，即是新增位元3 1610，可用來代表VC VL映射，藉此提供可支援至少12 VH、VC配對的4位元值。再者吾人應該注意，第16圖內的所有VCx VL映射具有相鄰保留位元1620-1670，可用來增加VCx VL映射至4位元。

第二欄位為十六進位值，代表特定VC至VL映射，例如「0001」對應至VC0。位元4為啟用VC0 VL映射的啟用位元。在此提供類似位元值給VH的每一VC1 VL、VC2 VL和VC3 VL映射。啟用位元的設定指示是否運用對應的VC至VL映射。第16圖內顯示的表格資料結構可與上面提到的其他表格資料結構搭配使用，以幫助PCI管理員、交換器等等進行的映射查找操作，以決定繞送來自特定VC上特定主機系統/刀鋒(例如VH)的資料流之虛擬連結。

吾人應該瞭解，第13圖至第16圖內表示的複雜映射

只用於三個 VH，且當在多根系統內運用更多多重 VH 時，例如高達 256 個 VH，就會變得更複雜。在此情況下，需要額外 VCx VL 映射位元。

第 17 圖為描述根據一個說明具體實施例用於執行優先群組映射和繞送的範例操作之流程圖。如第 17 圖內所示，操作由特權使用者，像是系統管理員，或自動機制開始，建立訊務等級(TC)至虛擬通道(VC)映射表資料結構，其使用硬體管理控制台(HMC)將訊務等級(TC)指派至單一虛擬階層(VH)內的虛擬通道(VC)，例如像是之前關於第 13 圖的說明(步驟 1710)。HMC 使用例如第 10 圖內的遠端管理指令介面 1002，將 TC 至 VC 映射表資料結構傳遞至 SR-PCIM (步驟 1720)。SR-PCIM 使用例如第 15 圖內所示的 PCI Express 組態位元欄，組態具有 TC 至 VC 映射的 VC 資源控制暫存器(步驟 1730)。根據 SR-PCIM 如何程式設定 VC，VC 根據限制排程、循環法(round robin)或加權循環法來仲裁。利用判斷是否還有額外 SR-IOV 系統要組態(步驟 1740)，針對每一 SR-IOV 系統重複上面關於步驟 1710-1730 描述的操作，若有，則針對下一個 SR-IOV 系統返回步驟 1710，其中每一 SR-IOV 系統都代表 MR-IOV 系統內的虛擬階層(VH)。

一旦用上述方式定義每一 SR-IOV 系統的 TC 至 VC 映射表資料結構，例如系統管理員這類特權使用者或自動

機制使用 HMC 將每一 VH/VC 配對指派給 VH/VC 配對至 VL 映射表資料結構內個別虛擬連結(VL) (步驟 1750)。這導致每一主機刀鋒的優先群組，因為如第 14 圖內所示 VL 相關於每一主機刀鋒的優先群組。HMC 使用例如第 10 圖內的遠端管理指令介面 1002，將定義優先群組映射的此 VH/VC 配對至 VL 映射表資料結構傳遞至 MR-PCIM (步驟 1760)。MR-PCIM 使用例如第 16 圖內所示的 PCI Express 組態位元，組態來自 MR-IOV 裝置功能表映射的功能 VC 至 VL 映射(步驟 1770)。然後操作結束。

一旦已經根據所說明具體實施例的機制設定映射，則根據優先群組運用這些映射將特定訊務等級的資料繞送至適當虛擬連結。例如在一個說明的具體實施例內，第 10 圖內的 HMC 1009 具有所說明具體實施例的每一主機刀鋒之優先群組映射，如第 13 圖和第 14 圖內所示的映射。然後 RMC 介面 1002 可例如用來將資料傳輸至 PCIM 1003。HMC 1009 可為單一管理介面，透過 RMC 介面 1002 可傳輸資料至 MR-PCIM 或每一個別 SR-PCIM。然後 SR-PCIM 或 MR-PCIM 可將組態資料寫入端點及/或交換器內的 PCI Express 組態空間之中。

吾人應該瞭解，功能 VC 至 VL 映射依照 VH 來組態。不過為了簡化範例，每一 VH 的位元欄都相同。根據 MR-PCIM 如何程式設計 VL，VL 可根據限制優先法或包

含循環法和加權佇列的其他排程法來仲裁。

如此所說明具體實施例提供機制來區別多根環境內的訊務類型，如此較高優先執行緒不會因為佇列前端阻擋而遭到較低優先執行緒阻擋。所說明具體實施例的機制提供一種系統及方法，來指派優先群組給虛擬連結，以便提供訊務區別並避免像是儲存這類較慢訊務阻擋對於延遲比較敏感的訊務，像是 HPC 應用。所說明具體實施例定義使用訊務等級將通過虛擬階層的訊務類型區別至虛擬通道、至虛擬連結映射能力的機制。多根系統的每一虛擬階層都可指派單一虛擬通道給虛擬連結。多訊務類型共享單一虛擬連結會導致佇列前端阻擋，例如虛擬連結上儲存訊務會阻擋 HPC 訊務。進一步，該儲存訊務可來自與該 HPC 訊務不同的虛擬階層。如此傳輸時間比較久的訊務會阻擋需要較少延遲的訊務，並且較慢訊務阻擋來自不同虛擬階層的虛擬連結。說明的具體實施例定義一種系統及方法，來指派優先群組給訊務等級、虛擬通道以及虛擬連結，以避免像是儲存的較慢訊務阻擋對於延遲比較敏感的訊務，像是 HPC 應用。

如上述，吾人應該瞭解，所說明具體實施例的態樣可採用整個硬體具體實施例、整個軟體具體實施例或包含硬體與軟體元件兩者的具體實施例之形式。在一個範例具體實施例內，所說明具體實施例的機制實施在軟體或

程式碼內，這包含但不受限於韌體、常駐軟體、微代碼等等。

適合儲存以及/或執行程式碼的資料處理系統將包含至少一個直接或透過系統匯流排間接耦合至記憶體元件的處理器。該記憶體元件可包含實際執行程式碼期間運用的本機記憶體、大量儲存體以及提供至少某些程式碼暫存的快取記憶體，以便減少執行期間必須從大量儲存體擷取的時間碼次數。

輸入/輸出或I/O裝置(包含但不受限於鍵盤、顯示器、指標裝置等等)可直接或透過中間I/O控制器耦合至系統。網路配接器也可耦合至系統，讓該資料處理系統變成耦合至其他資料處理系統，或透過中間私用或公用網路耦合至遠端印表機或儲存裝置。數據機、纜線數據機以及乙太網路卡只是一些目前可用的網路配接器種類。

本發明的描述已經為了說明與描述而呈現，並非要將本發明窮盡或受限在所揭示形式中。精通此技術的人士將瞭解許多修改與變化，具體實施例經過選擇與說明來最佳闡述本發明原理及實際應用，並且以許多具體實施例讓其他精通此技術的人士對本發明有最佳瞭解，這些具體實施例可具有多種修正來適用於所考慮到的特定使用。

【圖式簡單說明】

閱讀本發明並結合附圖，利用參考下列實施方式就可對於本發明本身、較佳使用模式、進一步目的及其優點有最佳瞭解，其中：

第 1 圖為說明合併業界一般已知的 PCIe 結構拓撲的系統之範例圖；

第 2 圖為說明業界一般已知的系統虛擬化之範例圖；

第 3 圖為說明使用虛擬化層將 PCI 根複合體的 I/O 虛擬化之第一方式範例圖；

第 4 圖為說明使用本質上共享的 PCI I/O 配接器，將 PCI 根複合體的 I/O 虛擬化之第二方式範例圖；

第 5 圖為 PCIe I/O 虛擬化啟用端點的範例圖；

第 6 圖為說明單根端點在無本質虛擬化時實體與虛擬功能之範例圖；

第 7 圖為說明單根端點起用本質 I/O 虛擬化時實體與虛擬功能之範例圖；

第 8 圖為說明根據一個說明具體實施例的多根虛擬化 I/O 拓撲之範例圖；

第 9 圖為說明根據一個說明具體實施例從根節點之 SR-PCIM 的觀點來看，多根虛擬化 I/O 拓撲的虛擬階層圖之範例圖；

第 10 圖為說明根據一個說明具體實施例的硬體維護控制台(hardware maintenance console，HMC)和 PCIM 之範例圖；

第 11 圖為說明建立阻擋通過 MR-IOV 系統的佇列前端之虛擬連結範例圖；

第 12 圖為說明根據一個說明具體實施例使用映射在每一主機系統刀鋒型虛擬連結上的訊務等級，避免佇列前端阻擋通過 MR-IOV 系統之範例圖；

第 13 圖為說明根據一個說明具體實施例的硬體維護控制台將單階層 SR-IOV 訊務等級(traffic class，TC)映射至虛擬通道(virtual channel，VC)之範例圖；

第 14 圖為說明根據一個說明具體實施例的硬體維護控制台將到虛擬連結的虛擬通道(VC)映射至每一主機刀鋒型的優先群組之範例圖；

第 15 圖為說明根據一個說明具體實施例在用於個別 SR-IOV 系統的 VC 資源控制暫存器內之 TV/VC 映射範例圖；

第 16 圖為說明根據一個說明具體實施例具有來自第 14 圖的虛擬通道至虛擬連結映射之 MR-IOV 裝置功能表範例圖；以及

第 17 圖為描述根據一個說明具體實施例用於執行優先群組映射以及每一主機系統刀鋒與繞送的範例操作之

流程圖。

【主要元件符號說明】

100 系統
110 主機處理器
120 記憶體
130 根複合體
140 PCIe 端點
150 PCI Express 對 PCI 橋接器
160 互連交換器
170 端點
182 端點
184 端點
186 端點
188 端點
210 應用程式
220 系統映像
230 虛擬系統
240 虛擬化層
250 實體系統資源
310 主機處理器集合
320 系統映像
330 系統映像
340 虛擬化層
350 PCIe 根複合體
360 PCIe 交換器
370 PCIe 端點
380 PCIe 端點
390 PCIe 端點
410 主機處理器集合
420 系統映像
430 系統映像
440 PCIe 根複合體
442 根複合體虛擬化啟用器
450 虛擬化層
460 PCIe 交換器
470 PCIe I/O 虛擬化(IOV)端點
480 PCIe I/O 虛擬化(IOV)端點
490 PCIe I/O 虛擬化(IOV)

端點

500 PCIe IOV 端點

510 PCIe 連接埠

520 內部繞送

530 組態管理功能

540 虛擬功能

550 虛擬功能

560 虛擬功能

570 非可分離資源

580 共享資源佇池

600 根節點

610 系統映像

612 系統映像

620 單根 PCIe 組態管理單元

630 I/O 虛擬化中間體

640 PCIe 根複合體

650 PCIe 交換器

670 PCIe 端點

680 PCIe 端點

690 PCIe 端點

700 單根節點

710 系統映像

712 系統映像

720 SR- PCIM

730 單 PCI 根複合體

740 PCIe 交換器

770 PCIe 端點

780 PCIe 端點

790 PCIe 端點

810 根節點

812 SR- PCIM

814 系統映像

816 系統映像

818 PCI 根複合體

820 根節點

822 SR- PCIM

824 系統映像

826 系統映像

828 PCI 根複合體

830 第三根節點

832 多根 PCI 組態管理員

834 PCI 根複合體

840 多根覺醒 PCIe 交換器

850 IOV 啟用 PCIe 端點

852 虛擬端點

854 虛擬端點

859 基本功能

860 IOV 啟用 PCIe 端點
862 虛擬端點
864 虛擬端點
869 基本功能
1001 系統
1002 遠端管理指令介面
1003 PCI 管理員
1004 用戶端分割
1005 裝置驅動器
1006 裝置驅動器
1009 硬體管理控制台
1010 平台硬體
1011 I/O 結構
1012 實體功能
1014 IOV 端點
1014 I/O 端點
1015 I/O 端點
1016 I/O 端點
1017 組態空間
1019 組態空間
1020 介面
1021 組態空間
1023 分割韌體
1024 分割韌體
1025 管理程序
1026 主機系統
1031 作業系統
1032 作業系統
1040 管理應用程式
1042 使用者介面
1105 計算系統
1105 刀鋒型伺服器
1110 第一計算裝置
1120 第二計算裝置
1130 第三計算裝置
1140 多根 PCIe 交換器
1145 I/O 配接器
1150 方塊
1202 PCI 管理員
1204 PCI 管理員
1206 PCI 管理員
1210 刀鋒
1210 計算裝置
1220 計算裝置
1220 刀鋒
1230 計算裝置
1230 刀鋒
1245 I/O 配接器

1250 映射模組

1252 映射模組

1254 映射模組

1260 映射表資料庫

1262 映射表資料庫

1264 映射表資料庫

1240 MR PCIe 交換器

1310 虛擬通道

1320 VH1

1330 單根 I/O 虛擬化(SR-IOV)訊務等級

1340 優先群組

1410 VC

1420 VH

1430 VL

1440 優先群組

1610 位元 3

1620 保留位元

1630 保留位元

1640 保留位元

1650 保留位元

1660 保留位元

1670 保留位元

七、申請專利範圍：

1. 一種在資料處理系統內區別不同訊務類型之方法，包含以下步驟：

產生一第一映射資料結構，對於該資料處理系統之複數個單根虛擬階層內每一單根虛擬階層，該第一映射資料結構將複數個訊務等級內每一訊務等級與複數個優先群組內之一對應優先群組及複數個虛擬通道內之一對應虛擬通道相關；

產生一第二映射資料結構，其將該複數個虛擬通道內每一虛擬通道映射至該資料處理系統內複數個虛擬連結中之一對應虛擬連結；以及

根據該第一映射資料結構以及第二映射資料結構，將訊務從一單根虛擬階層的一特定優先群組繞送(route)至該複數個虛擬連結內一特定虛擬連結，其中每一虛擬階層與虛擬通道的組合都映射至該複數個虛擬連結內之一不同的虛擬連結，其中根據該第一映射資料結構以及第二映射資料結構，將訊務從該單根虛擬階層繞送至該複數個虛擬連結內該特定虛擬連結之步驟包含以下步驟：

決定與該訊務相關的一訊務等級；

在該第一映射資料結構內執行一查找操作，以識別對應至該訊務之該特定優先群組的一虛擬通道，該

訊務之該特定優先群組對應至該訊務之該單根虛擬階層和該訊務等級之一組合；

在該第二映射資料結構內執行一查找操作，以識別與該訊務之該虛擬通道和單根虛擬階層相關的一虛擬連結；以及

將該訊務繞送至該已識別虛擬連結。

2. 如申請專利範圍第1項所述之方法，其中該第一映射資料結構與第二映射資料結構由該資料處理系統的一硬體管理控制台所產生。

3. 如申請專利範圍第2項所述之方法，其中該硬體管理控制台根據該第一映射資料結構與第二映射資料結構組態一主機系統的一或多個通訊結構管理員。

4. 如申請專利範圍第1項所述之方法，其中該資料處理系統為具有複數個單根虛擬階層的一多根資料處理系統。

5. 如申請專利範圍第4項所述之方法，其中該第一映射資料結構與第二映射資料結構根據該單根虛擬階層的一識別碼組合與由該識別碼識別的該單根虛擬階層相關之複數個虛擬通道內一虛擬通道，將資料從該複數個單根虛擬階層內每一單根虛擬階層繞送至該複數個虛擬連結內的一不同虛擬連結。

6. 如申請專利範圍第1項所述之方法，其中針對一單根

虛擬階層，在該第一映射資料結構內，一相同虛擬通道對應至一相同優先群組，並且其中該相同虛擬通道具有相關於該相同虛擬通道的一或多個訊務等級。

7. 如申請專利範圍第 6 項所述之方法，其中該單根虛擬階層相關於一刀鋒型伺服器的一主機系統刀鋒。

8. 如申請專利範圍第 1 項所述之方法，其中由一刀鋒型伺服器內一主機系統刀鋒的一周邊組件互連(PCI) Express 管理組件實施該方法。

9. 如申請專利範圍第 8 項所述之方法，其中該 PCIExpress 管理組件為一單根 PCI Express 管理組件或一多根 PCI Express 管理組件中之一個。

10. 一種包含一電腦可讀取媒體的電腦程式產品，在該媒體上記錄一電腦可讀取程式，其中該電腦可讀取程式在一資料處理系統上執行時會導致該資料處理系統：

產生一第一映射資料結構，對於在該資料處理系統之複數個單根虛擬階層內每一單根虛擬階層，該第一映射資料結構將複數個訊務等級內每一訊務等級與複數個優先群組內之一對應優先群組及複數個虛擬通道內一對應虛擬通道相關；

產生一第二映射資料結構，其將該複數個虛擬通道內每一虛擬通道映射至該資料處理系統內複數個虛擬連結中之一對應虛擬連結；以及

根據該第一映射資料結構以及第二映射資料結構，將訊務從一單根虛擬階層的一特定優先群組繞送至該複數個虛擬連結內一特定虛擬連結，其中每一虛擬階層與虛擬通道的組合都映射至該複數個虛擬連結內一不同的虛擬連結，其中根據該第一映射資料結構以及第二映射資料結構，將訊務從該單根虛擬階層繞送至該複數個虛擬連結內該特定虛擬連結之步驟包含以下步驟：

決定與該訊務相關的一訊務等級；

在該第一映射資料結構內執行一查找操作，以識別對應至該訊務之該特定優先群組的一虛擬通道，該訊務之該特定優先群組對應至該訊務之該單根虛擬階層和該訊務等級之一組合；

在該第二映射資料結構內執行一查找操作，以識別與該訊務之該虛擬通道和單根虛擬階層相關的一虛擬連結；以及

將該訊務繞送至該已識別虛擬連結。

11. 如申請專利範圍第 10 項所述之電腦程式產品，其中該第一映射資料結構與第二映射資料結構由該資料處理系統的一硬體管理控制台所產生。

12. 如申請專利範圍第 11 項所述之電腦程式產品，其中該硬體管理控制台根據該第一映射資料結構與第二

映射資料結構組態一主機系統的一或多個通訊結構管理員。

13. 如申請專利範圍第 10 項所述之電腦程式產品，其中該資料處理系統為具有複數個單根虛擬階層的一多根資料處理系統。

14. 如申請專利範圍第 13 項所述之電腦程式產品，其中該第一映射資料結構與第二映射資料結構根據該單根虛擬階層的一識別碼組合與由該識別碼識別的該單根虛擬階層相關之複數個虛擬通道內一虛擬通道，將資料從該複數個單根虛擬階層內每一單根虛擬階層繞送至該複數個虛擬連結內之一不同虛擬連結。

15. 如申請專利範圍第 10 項所述之電腦程式產品，其中針對一單根虛擬階層，在該第一映射資料結構內，一相同虛擬通道對應至一相同優先群組，並且其中該相同虛擬通道具有相關於該相同虛擬通道的一或多個訊務等級。

16. 如申請專利範圍第 15 項所述之電腦程式產品，其中該單根虛擬階層相關於一刀鋒型伺服器的一主機系統刀鋒。

17. 如申請專利範圍第 10 項所述之電腦程式產品，其中由一刀鋒型伺服器內一主機系統刀鋒的一周邊組件互連(PCI) Express 管理組件執行該電腦可讀取程式。

18. 一種資料處理系統，包含：

一硬體管理控制台；以及

一主機系統，其耦合至該硬體管理控制台，其中該硬體管理控制台：

產生一第一映射資料結構，對於該資料處理系統之複數個單根虛擬階層內每一單根虛擬階層，該第一映射資料結構將複數個訊務等級內每一訊務等級與複數個優先群組內之一對應優先群組及複數個虛擬通道內一對應虛擬通道相關；

產生一第二映射資料結構，其將該複數個虛擬通道內每一虛擬通道映射至該資料處理系統內複數個虛擬連結中之一對應虛擬連結；以及

根據該第一映射資料結構以及第二映射資料結構，將訊務從一單根虛擬階層的一特定優先群組繞送至該複數個虛擬連結內之一特定虛擬連結，其中每一虛擬階層與虛擬通道的組合都映射至該複數個虛擬連結內之一不同的虛擬連結，其中根據該第一映射資料結構以及第二映射資料結構，將訊務從該單根虛擬階層繞送至該複數個虛擬連結內該特定虛擬連結之步驟包含以下步驟：

決定與該訊務相關的一訊務等級；

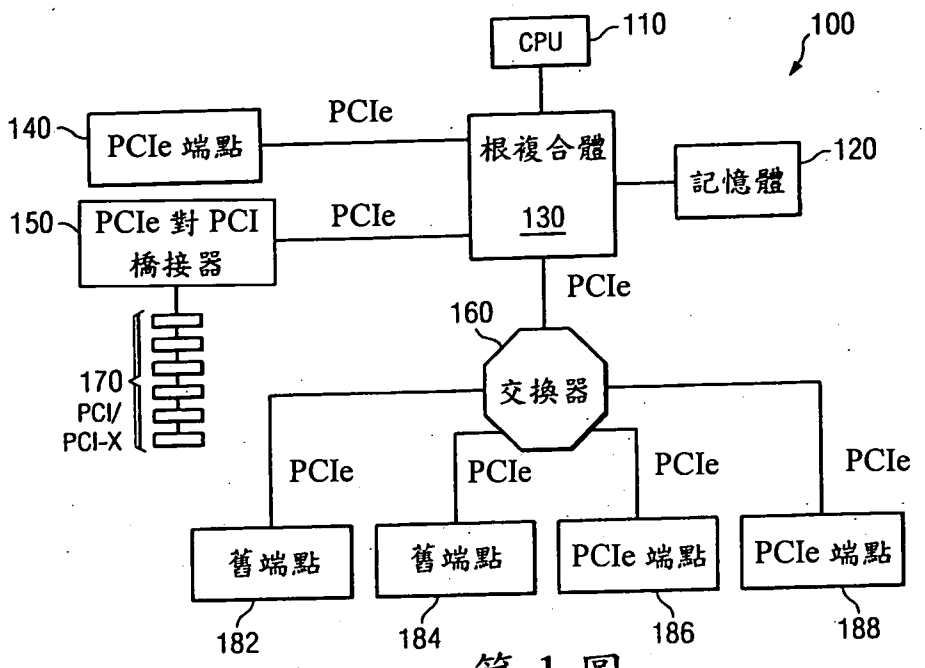
在該第一映射資料結構內執行一查找操作，以識

別對應至該訊務之該特定優先群組的一虛擬通道，該訊務之該特定優先群組對應至該訊務之該單根虛擬階層和該訊務等級之一組合；

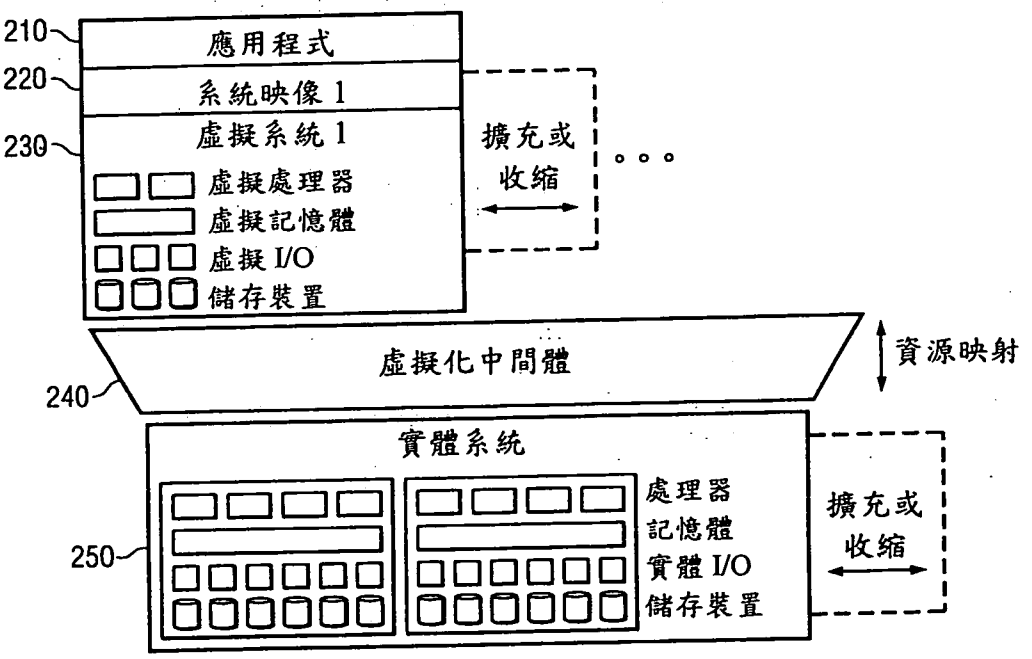
在該第二映射資料結構內執行一查找操作，以識別與該訊務之該虛擬通道和單根虛擬階層相關的一虛擬連結；以及

將該訊務繞送至該已識別虛擬連結。

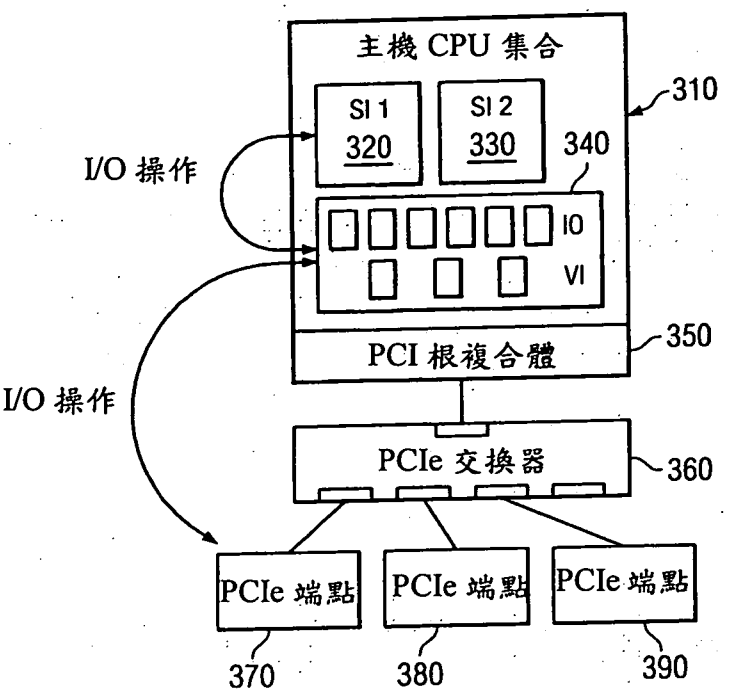
八、圖式：


CPU
110
100
140
PCIe 端點
PCIe
根複合體
130
記憶體
120
150
PCIe 對 PCI
橋接器
PCIe
PCIe
160
交換器
170
PCI/
PCI-X
PCIe
PCIe
PCIe
PCIe
舊端點
舊端點
PCIe 端點
PCIe 端點
182
184
186
188

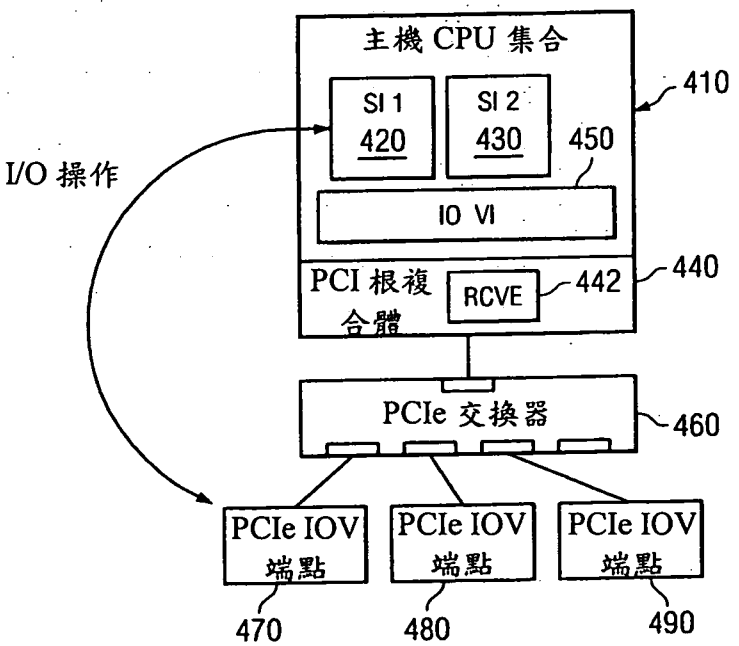
第 1 圖


210
應用程式
220
系統映像 1
230
虛擬系統 1
虛擬處理器
虛擬記憶體
虛擬 I/O
儲存裝置
擴充或
收縮
虛擬化中間體
資源映射
240
實體系統
處理器
記憶體
實體 I/O
儲存裝置
擴充或
收縮
250

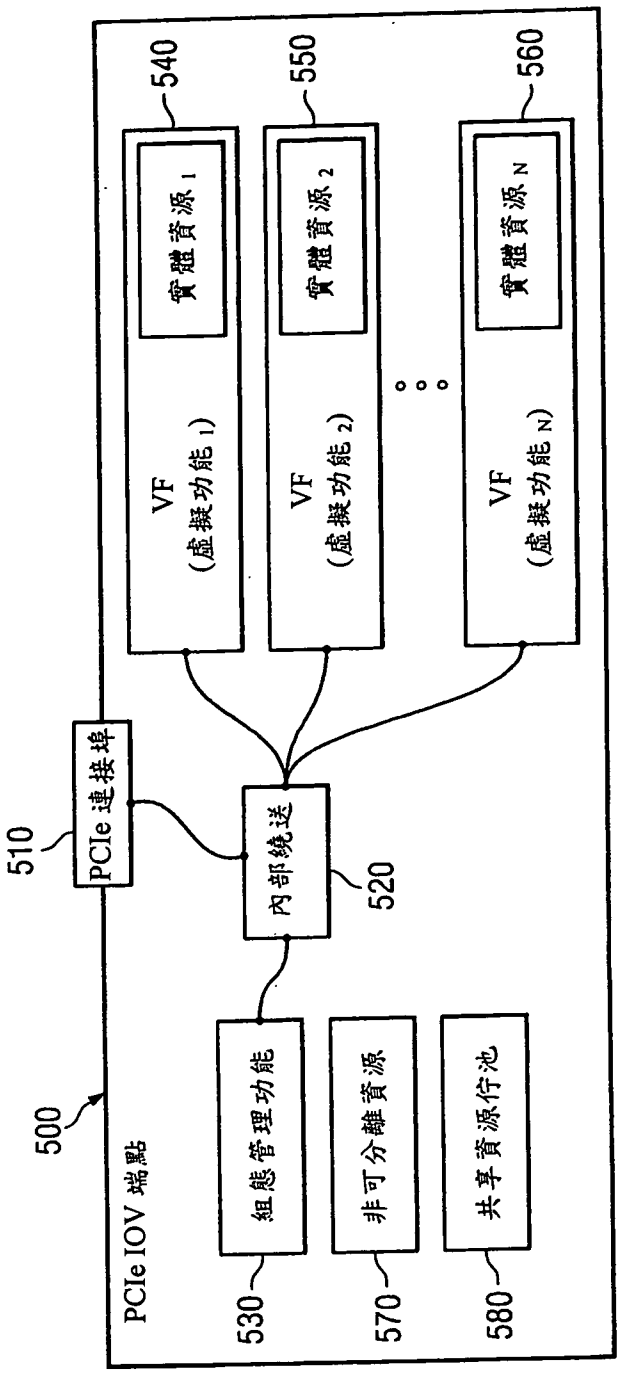
第 2 圖


主機 CPU 集合
SI 1
320
SI 2
330
340
310
I/O 操作
IO
VI
PCI 根複合體
350
I/O 操作
PCIe 交換器
360
PCIe 端點
PCIe 端點
PCIe 端點
370
380
390

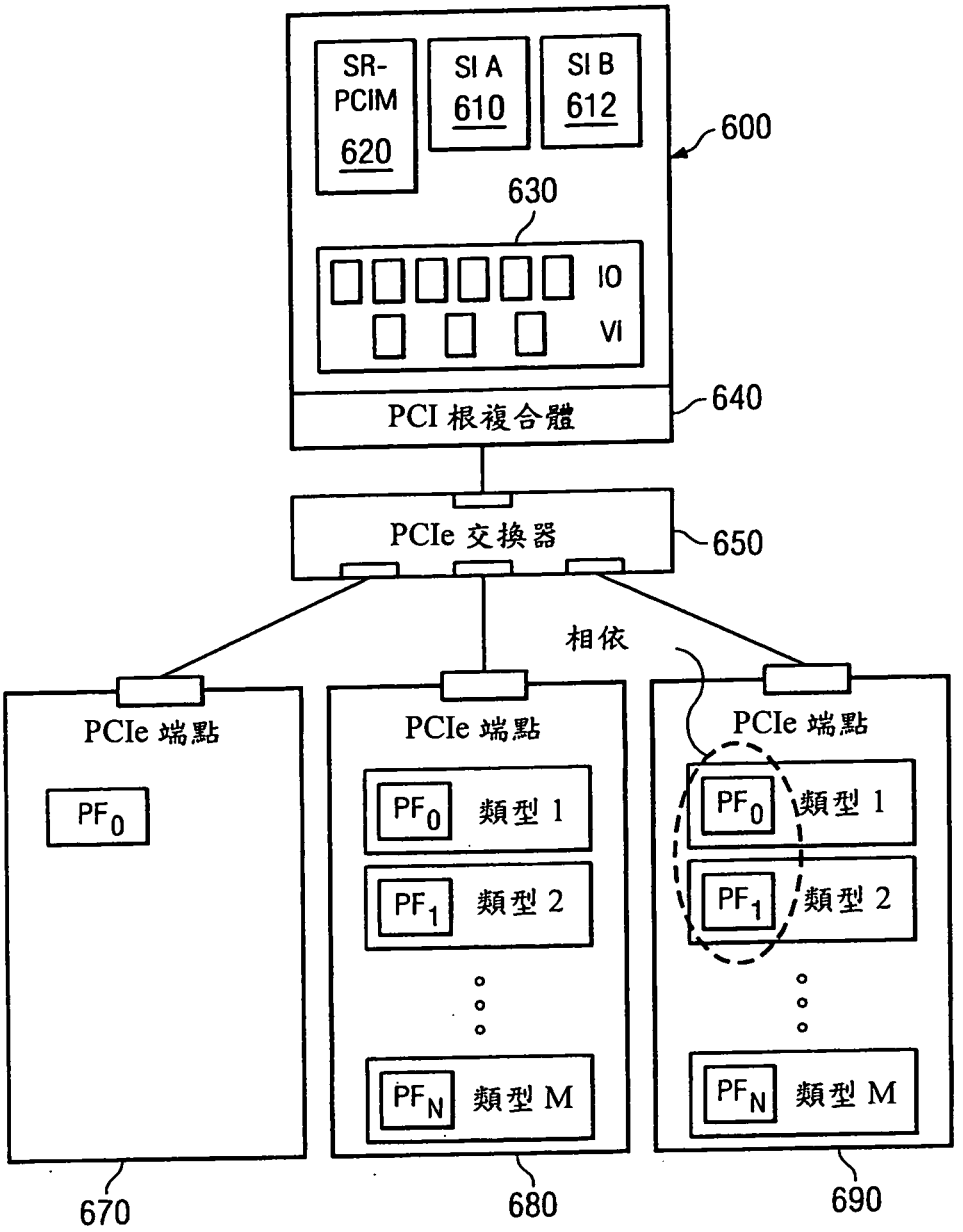
第 3 圖


主機 CPU 集合
SI 1
420
SI 2
430
410
450
I/O 操作
IO VI
PCI 根複合體
RCVE
442
440
PCIe 交換器
460
PCIe IOV 端點
PCIe IOV 端點
PCIe IOV 端點
470
480
490

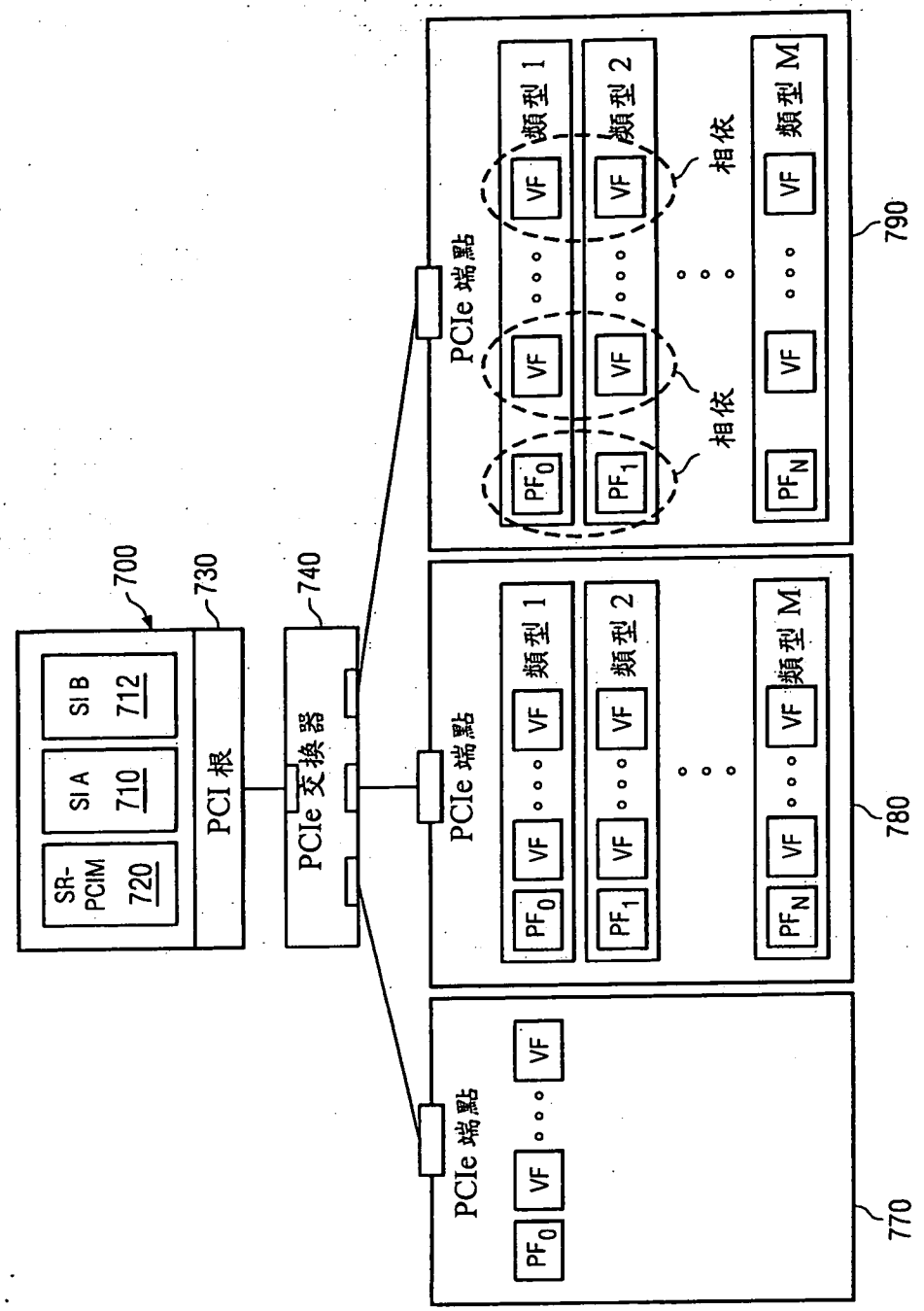
第 4 圖


PCIe IOV 端點
500
PCIe 連接埠
510
內部鏡送
520
VF
(虛擬功能 1)
實體資源 1
540
VF
(虛擬功能 2)
實體資源 2
550
VF
(虛擬功能 N)
實體資源 N
560
組態管理功能
530
非可分離資源
570
共享資源佇池
580

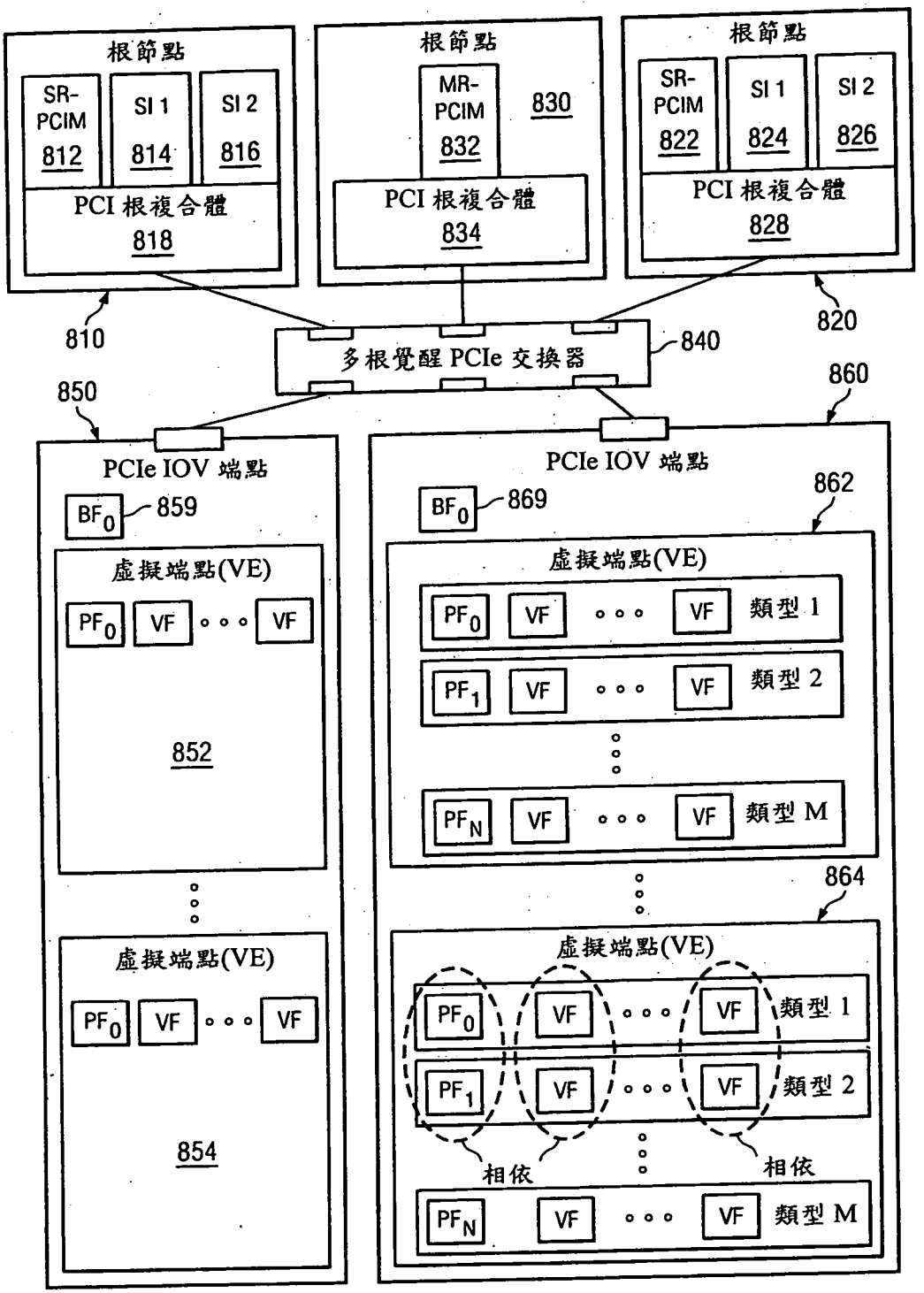
第 5 圖


SR-PCIM 620
SI A 610
SI B 612
600
630
IO
VI
PCI 根複合體
640
PCIe 交換器
650
相依
PCIe 端點
PF0
PCIe 端點
PF0 類型 1
PF1 類型 2
PFN 類型 M
PCIe 端點
PF0 類型 1
PF1 類型 2
PFN 類型 M
670
680
690

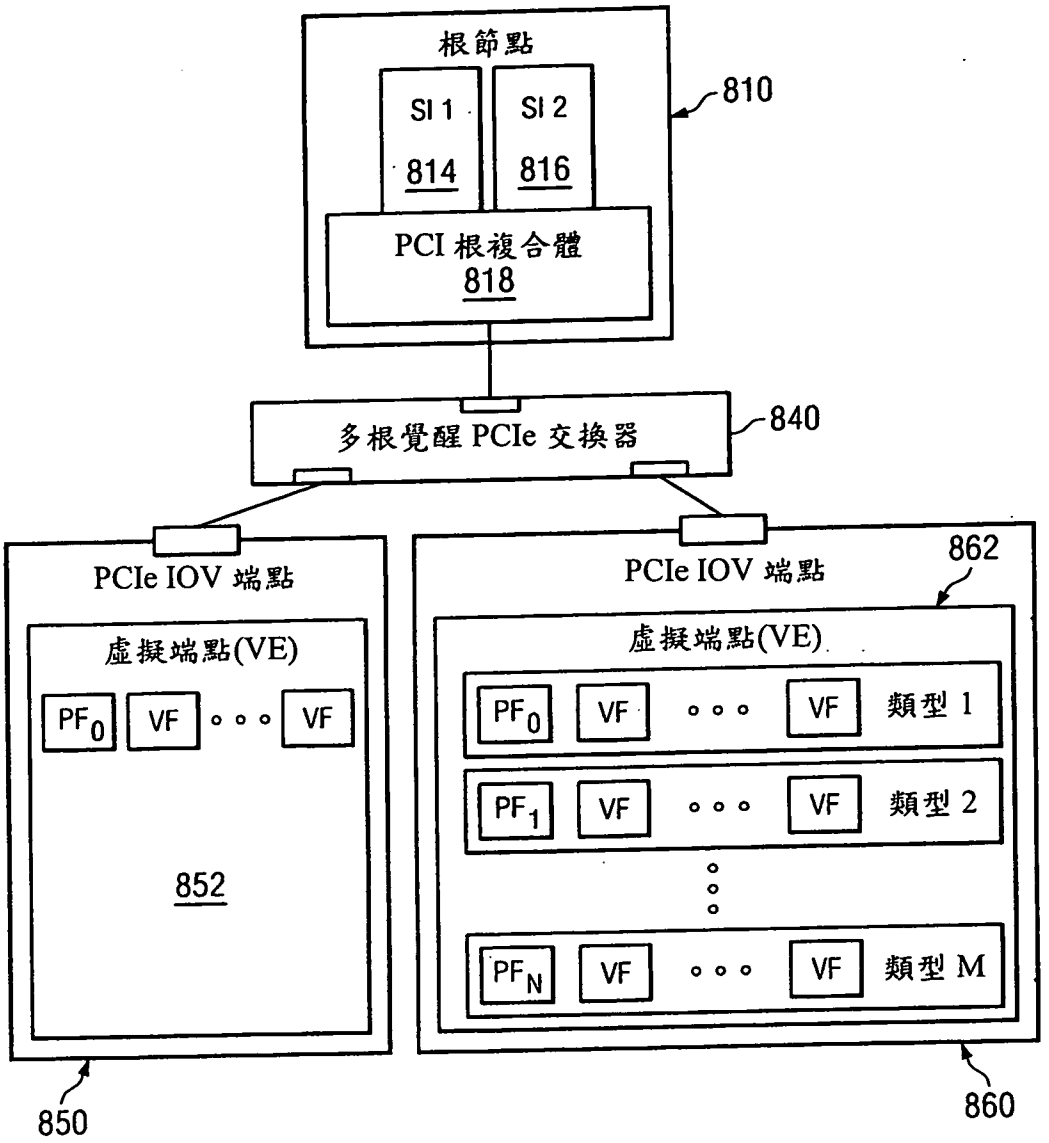
第 6 圖


700
730
740
SR-PCIM 720
SI A 710
SI B 712
PCI 根
PCIe 交換器
PCIe 端點
PF0
VF
類型 1
類型 2
類型 M
PF1
PFN
相依
770
780
790

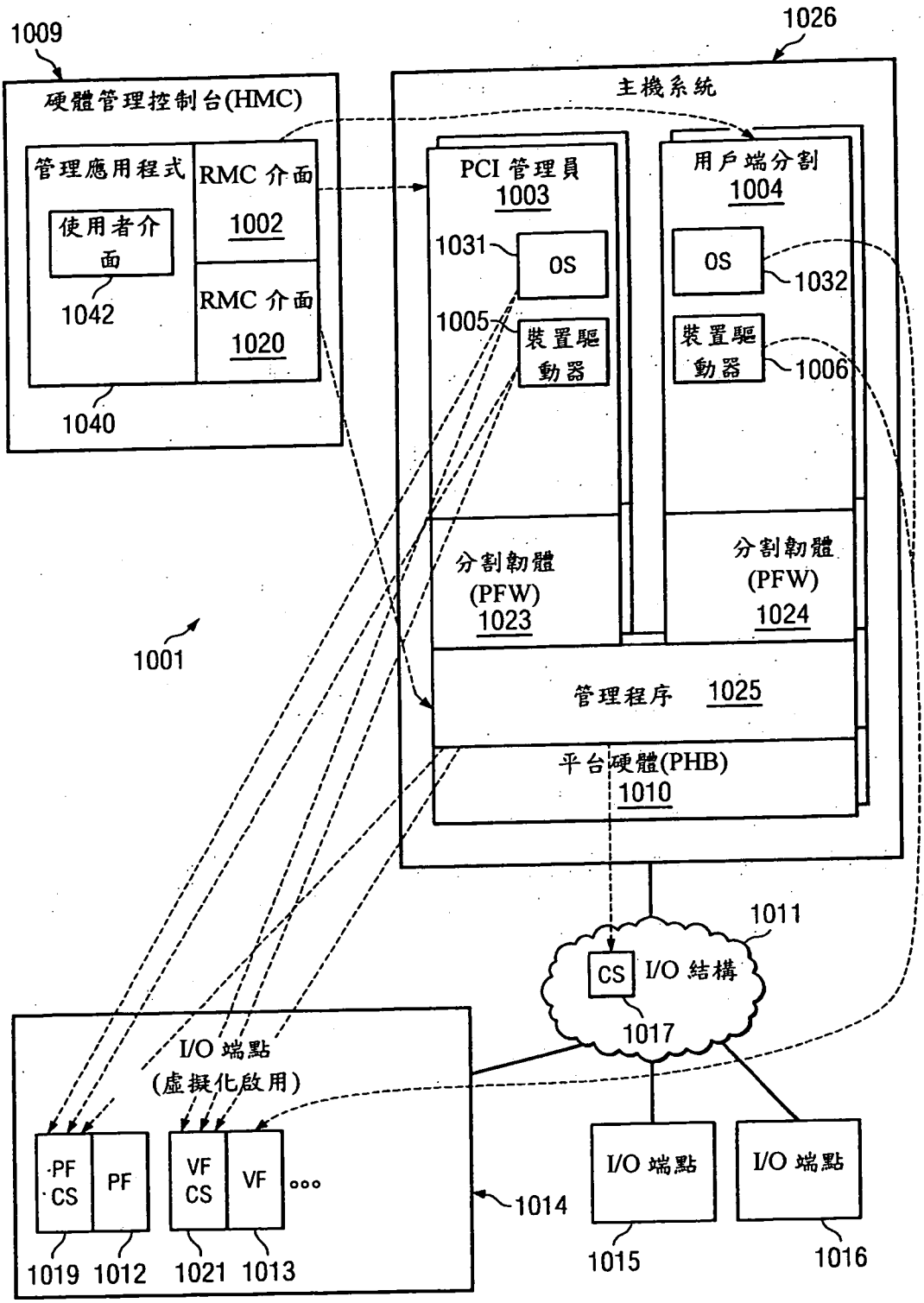
第 7 圖


根節點
SR-PCIM 812
SI 1 814
SI 2 816
PCI 根複合體 818
根節點
MR-PCIM 832
830
PCI 根複合體 834
根節點
SR-PCIM 822
SI 1 824
SI 2 826
PCI 根複合體 828
810
820
多根覺醒 PCIe 交換器
840
850
860
PCIe IOV 端點
BF0
859
虛擬端點(VE)
PF0 VF ... VF
852
虛擬端點(VE)
PF0 VF ... VF
854
PCIe IOV 端點
BF0
869
862
虛擬端點(VE)
PF0 VF ... VF 類型 1
PF1 VF ... VF 類型 2
PFN VF ... VF 類型 M
864
虛擬端點(VE)
PF0 VF ... VF 類型 1
PF1 VF ... VF 類型 2
相依
相依
PFN VF ... VF 類型 M

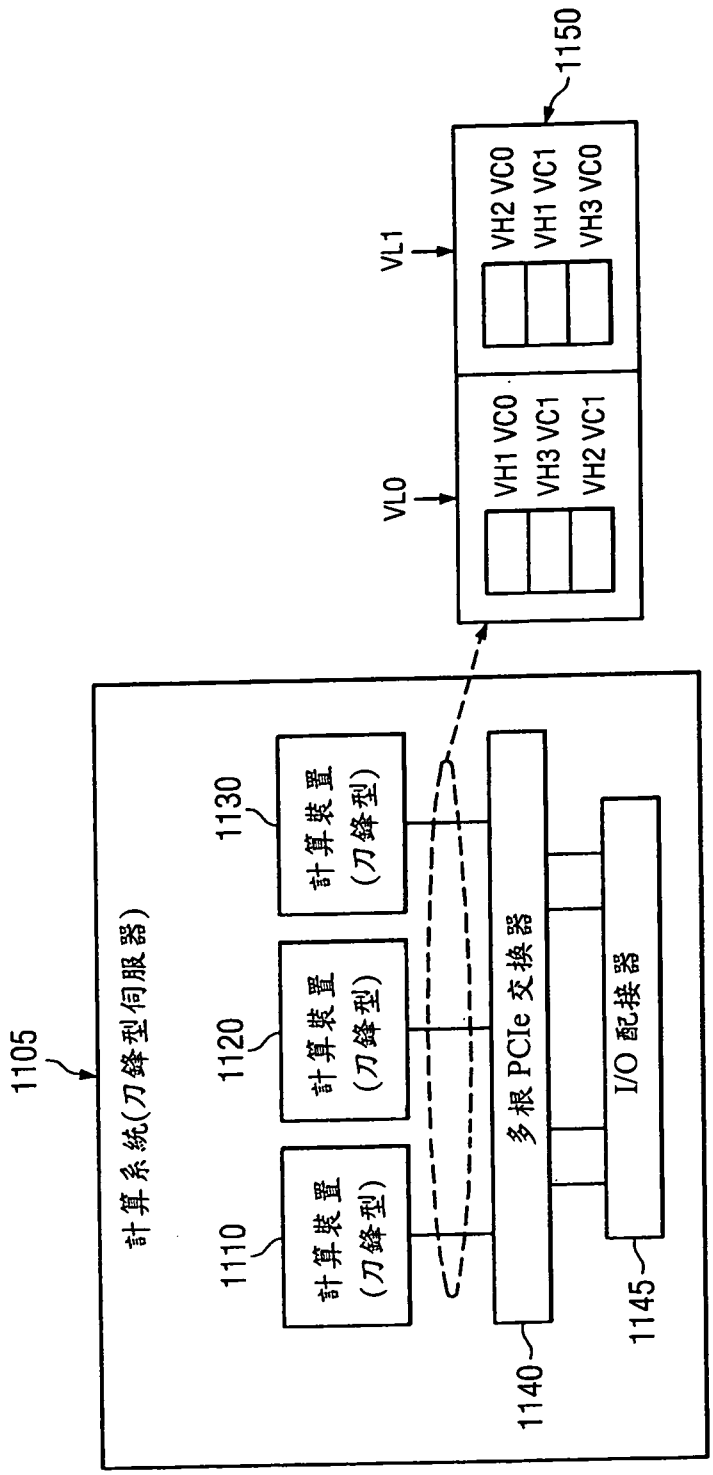
第 8 圖


根節點
SI 1
814
SI 2
816
PCI 根複合體
818
810
多根覺醒 PCIe 交換器
840
PCIe IOV 端點
虛擬端點(VE)
PF0
VF
VF
852
850
PCIe IOV 端點
862
虛擬端點(VE)
PF0
VF
VF
類型 1
PF1
VF
VF
類型 2
PFN
VF
VF
類型 M
860

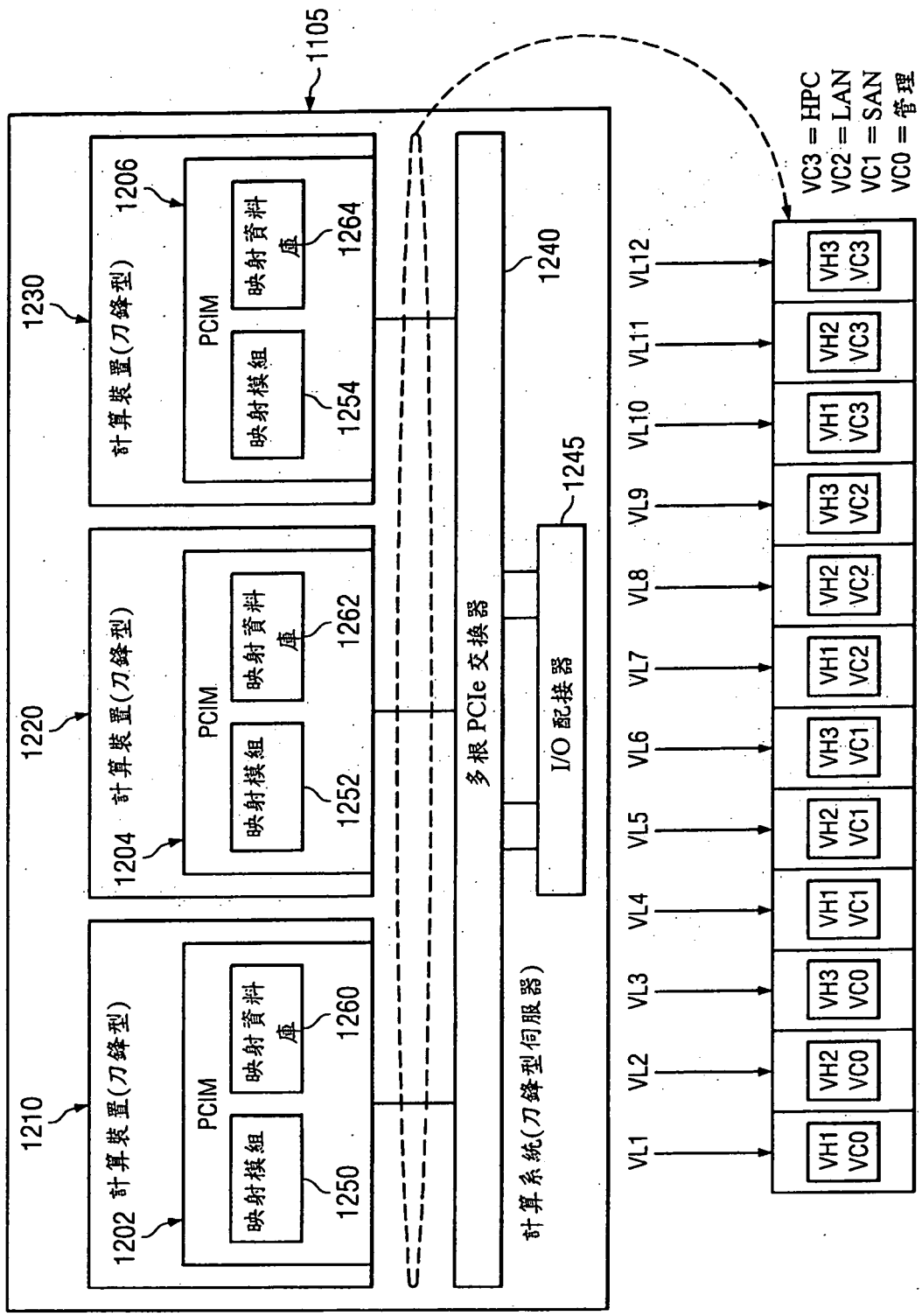
第 9 圖


1009
硬體管理控制台(HMC)
管理應用程式
使用者介面
1042
RMC 介面
1002
RMC 介面
1020
1040
1026
主機系統
PCI 管理員
1003
1031
OS
1005
裝置驅動器
用戶端分割
1004
OS
1032
裝置驅動器
1006
分割韌體
(PFW)
1023
分割韌體
(PFW)
1024
1001
管理程序 1025
平台硬體(PHB)
1010
1011
CS
I/O 結構
1017
I/O 端點
(虛擬化啟用)
PF
CS
PF
VF
CS
VF
ooo
1014
I/O 端點
I/O 端點
1019
1012
1021
1013
1015
1016

第 10 圖


1105
計算系統(刀鋒型伺服器)
1110
計算裝置
(刀鋒型)
1120
計算裝置
(刀鋒型)
1130
計算裝置
(刀鋒型)
多根 PCIe 交換器
1140
I/O 配接器
1145
VL0
VH1 VC0
VH3 VC1
VH2 VC1
VL1
VH2 VC0
VH1 VC1
VH3 VC0
1150

第 11 圖


1105
1210
計算裝置(刀鋒型)
1202
PCIM
映射模組
1250
映射資料庫
1260
1220
計算裝置(刀鋒型)
1204
PCIM
映射模組
1252
映射資料庫
1262
1230
計算裝置(刀鋒型)
1206
PCIM
映射模組
1254
映射資料庫
1264
多根 PCIe 交換器
1240
I/O 配接器
1245
計算系統(刀鋒型伺服器)
VL1
VL2
VL3
VL4
VL5
VL6
VL7
VL8
VL9
VL10
VL11
VL12
VH1 VC0
VH2 VC0
VH3 VC0
VH1 VC1
VH2 VC1
VH3 VC1
VH1 VC2
VH2 VC2
VH3 VC2
VH1 VC3
VH2 VC3
VH3 VC3
VC3 = HPC
VC2 = LAN
VC1 = SAN
VC0 = 管理

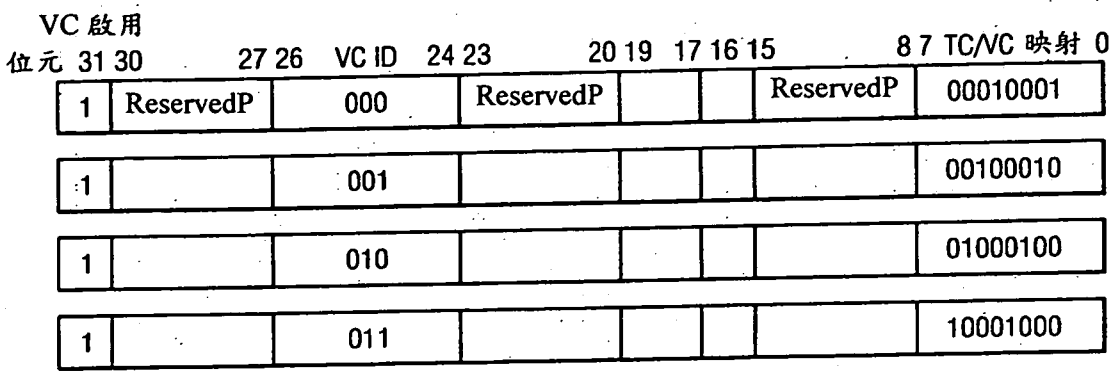
第 12 圖

1310 VC	1320 VH	1330 TC	1340 優先群組
VC3	VH1	TC7	HPC
VC2	VH1	TC6	LAN
VC1	VH1	TC5	SAN
VC0	VH1	TC4	管理
VC3	VH1	TC3	HPC
VC2	VH1	TC2	LAN
VC1	VH1	TC1	SAN
VC0	VH1	TC0	管理

第 13 圖

1410 VC	1420 VH	1430 VL	1440 優先群組
VC3	VH1	VL10	HPC
VC2	VH1	VL7	LAN
VC1	VH1	VL4	SAN
VC0	VH1	VL1	管理
VC3	VH2	VL11	HPC
VC2	VH2	VL8	LAN
VC1	VH2	VL5	SAN
VC0	VH2	VL2	管理
VC3	VH3	VL12	HPC
VC2	VH3	VL9	LAN
VC1	VH3	VL6	SAN
VC0	VH3	VL3	管理

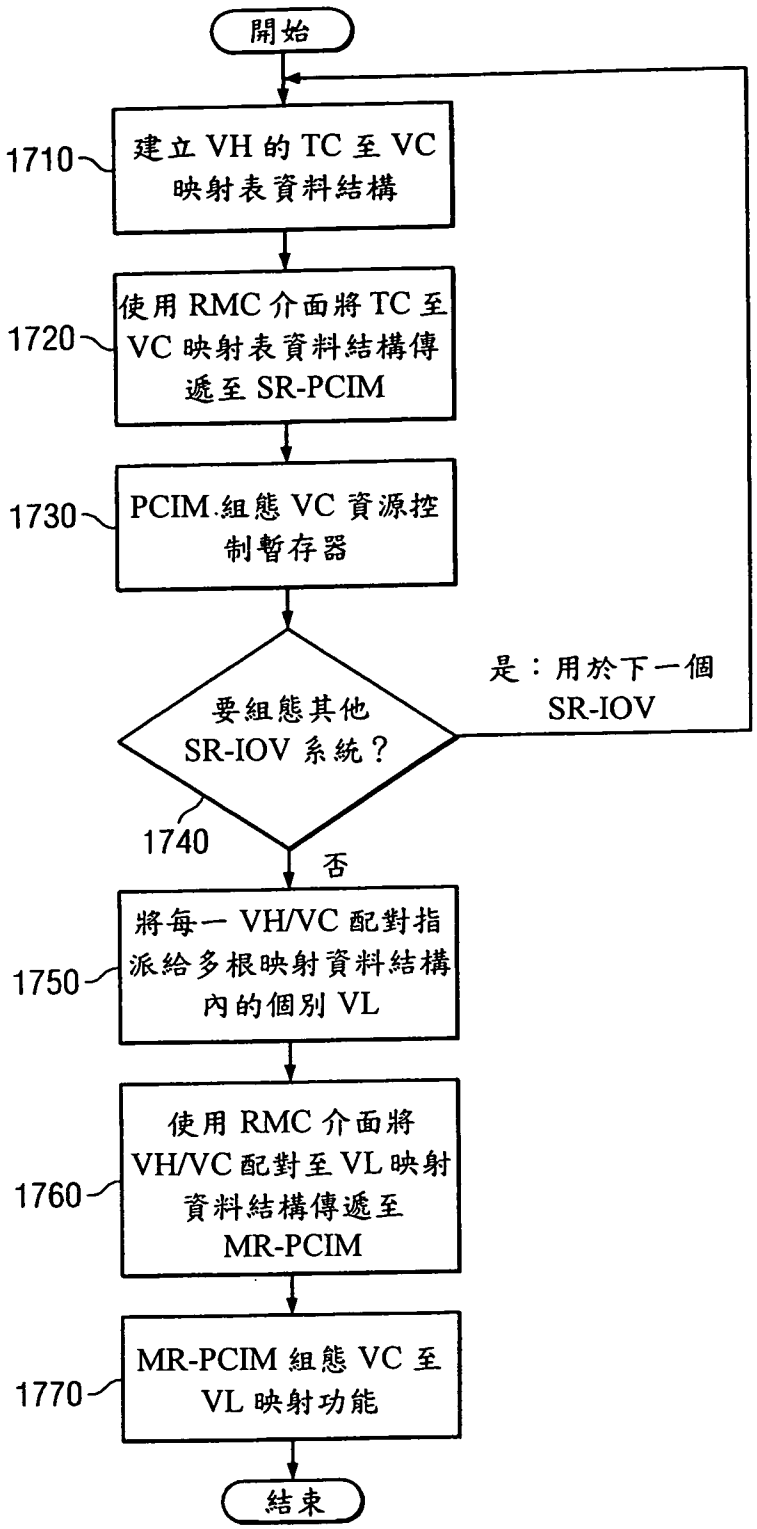
第 14 圖



第 15 圖

	2:0	100	VC0 VL 映射
1610	3		ReservedP
	4	1	VC0 VL 映射啟用
1620	7:5		ReservedP
	10:8	101	VC1 VL 映射
1630	11		ReservedP
	12	1	VC1 VL 映射啟用
1640	15:13		ReservedP
	18:16	110	VC2 VL 映射
1650	19		ReservedP
	20	1	VC2 VL 映射啟用
1660	23:21		ReservedP
	26:24	111	VC3 VL 映射
1670	27		ReservedP
	28	1	VC3 VL 映射啟用

第 16 圖


開始
1710
建立 VH 的 TC 至 VC 映射表資料結構
1720
使用 RMC 介面將 TC 至 VC 映射表資料結構傳遞至 SR-PCIM
1730
PCIM 組態 VC 資源控制暫存器
要組態其他 SR-IOV 系統？
是：用於下一個 SR-IOV
1740
否
1750
將每一 VH/VC 配對指派給多根映射資料結構內的個別 VL
1760
使用 RMC 介面將 VH/VC 配對至 VL 映射資料結構傳遞至 MR-PCIM
1770
MR-PCIM 組態 VC 至 VL 映射功能
結束

第 17 圖