

發明專利說明書

公告本

(本說明書格式、順序及粗體字，請勿任意更動，※記號部分請勿填寫)

※ 申請案號：95111529

※ 申請日期：95.3.31

※IPC 分類：H04L 12/28 H04L 12/56
G06F 13/38

一、發明名稱：(中文/英文)

於伺服器環境中用以網路卸載之主乙太網路配接器

HOST ETHERNET ADAPTER FOR NETWORKING OFFLOAD IN
SERVER ENVIRONMENT

二、申請人：(共 1 人)

姓名或名稱：(中文/英文)

美商萬國商業機器公司

INTERNATIONAL BUSINESS MACHINES CORPORATION

代表人：(中文/英文)

琳恩 D 安德森

ANDERSON, LYNNE D.

住居所或營業所地址：(中文/英文)

美國紐約州阿蒙市新果園路

NEW ORCHARD ROAD, ARMONK, NY 10504, U.S.A.

國籍：(中文/英文)

美國 U.S.A.

三、發明人：(共 11 人)

姓 名：(中文/英文)

1. 瑞維 庫瑪 阿瑞米黎
ARIMILLI, RAVI KUMAR
2. 克勞迪 巴索
BASSO, CLAUDE
3. 尚 路易斯 凱維納
CALVIGNAC, JEAN LOUIS
4. 張志任
CHANG, CHIH-JEN
5. 菲利普 戴蒙
DAMON, PHILIPPE
6. 羅諾 艾德華 傅斯
FUHS, RONALD EDWARD
7. 沙提雅 普拉卡希 夏瑪
SHARMA, SATYA PRAKASH
8. 納塔瑞詹 維顏納森
VAIDHYANATHAN, NATARAJAN
9. 法布瑞斯 尚 維普蘭肯
VERPLANKEN, FABRICE JEAN
10. 柯林 畢頓 維瑞里
VERRILLI, COLIN BEATON
11. 史考特 麥可 威倫伯格
WILLENBORG, SCOTT MICHAEL

國 籍：(中文/英文)

1. 美國 U.S.A.
2. 法國 FRANCE
3. 法國 FRANCE
4. 中華民國 ROC (TAIWAN)
5. 法國 FRANCE
6. 美國 U.S.A.
7. 美國 U.S.A.
8. 印度 INDIA
9. 法國 FRANCE
10. 美國 U.S.A.
11. 美國 U.S.A.

四、聲明事項：

☐ 主張專利法第二十二條第二項 ☐ 第一款或 ☐ 第二款規定之事實，其事實發生日期為： 年 月 日。

☒ 申請前已向下列國家（地區）申請專利：

【格式請依：受理國家（地區）、申請日、申請案號 順序註記】

☒ 有主張專利法第二十七條第一項國際優先權：

1. 美國；2005年04月01日；11/097,608

2.

☐ 無主張專利法第二十七條第一項國際優先權：

1.

2.

☐ 主張專利法第二十九條第一項國內優先權：

【格式請依：申請日、申請案號 順序註記】

☐ 主張專利法第三十條生物材料：

☐ 須寄存生物材料者：

國內生物材料 【格式請依：寄存機構、日期、號碼 順序註記】

國外生物材料 【格式請依：寄存國家、機構、日期、號碼 順序註記】

☐ 不須寄存生物材料者：

所屬技術領域中具有通常知識者易於獲得時，不須寄存。

五、中文發明摘要：

本發明揭示一種乙太網路配接器。該乙太網路配接器包含允許該配接器自一處理器接收封包及將封包傳輸至一處理器的複數個層。該複數個層包括一解多工機制以考慮該處理器之分割。一主乙太網路配接器(HEA)係一為乙太網路及TCP加速提供一新方法的整合乙太網路配接器。已將一組TCP/IP加速特徵引入一工具組方法中：當要求時及如要求，伺服器TCP/IP堆疊使用此等加速器。在伺服器與網路介面控制器之間之介面已藉由繞過PCI匯流排而簡化。HEA支援網路虛擬化。HEA可由提供必要隔離及保護之多個OS共用而不會影響其效能。

六、英文發明摘要：

七、指定代表圖：

(一)本案指定代表圖為：第 (2) 圖。

(二)本代表圖之元件符號簡單說明：

100	伺服器系統
102	處理器
104	記憶體
106	介面配接器晶片
108	介面
110	主乙太網路配接器

八、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

(無)

九、發明說明：

【發明所屬之技術領域】

本發明大體係關於一種伺服器環境，且更特定言之係關於用於該環境中之配接器。

【先前技術】

圖1說明一習知伺服器系統10。伺服器系統10包括一耦接至一主記憶體14之處理器12。處理器12經由其私有匯流排(GX)16耦接至包括網路介面系統18之系統。網路介面系統18又經由一PCI匯流排22或其類似物耦接至一配接器20。如熟知，PCI 22匯流排具有影響經由其流動之訊務量的有限頻寬。

網際網路及其應用已極大地增加伺服器必須滿足之用戶端請求的數目。每一用戶端請求產生網路及儲存I/O。另外，百億位元乙太網路及IP儲存的出現使得可能在單一骨幹基礎結構(乙太網路、TCP/IP)上合併資料中心通信。

然而，百億位元速度下之TCP/IP協定極大地消耗主流伺服器中之處理及記憶體頻寬，因此嚴重限制伺服器運行應用程式之能力。

在現今之伺服器網路介面控制器(NIC)中，諸如TCP及IP總和檢查碼、大量發送(或TCP分段卸載)之有限功能卸載受到支援。然而，此等功能在1 G時是足夠的，但卻不能解決更高速度如10 G及更高時的問題。

眾所周知，使用TCP卸載引擎以自伺服器完全卸載全部TCP/IP協定堆疊。然而，TOE之實施通常在硬體或在相對

複雜之微微處理器架構中之微微碼中進行。亦有除錯、問題確定及堆疊可維護性問題。另外，當使用微微碼時有可擴充性問題，因為微微引擎不跟隨主處理器路標。最後，卸載引擎通常引入新協定及API且因此要求在應用程式以及互用性問題的改變。

因此，需要一用以考慮克服以上識別之問題的乙太網路環境中之高頻寬資料的系統及方法。本發明處理該需要。

【發明內容】

本發明揭示一種乙太網路配接器。該乙太網路配接器包括用於允許該配接器自一處理器接收封包及將封包傳輸至一處理器之複數個層。該複數個層包括一解多工機制以考慮處理器的分割。

主乙太網路配接器(HEA)係一為乙太網路及TCP加速提供一種新方法之整合乙太網路配接器。一組TCP/IP加速特徵已在工具組方法中引入：當有要求時伺服器TCP/IP堆疊使用此等加速器。在伺服器與網路介面控制器之間之介面已藉由繞過PCI匯流排而得以簡化。

HEA支援網路虛擬化。HEA可由提供必要隔離及保護之多個OS共用而不會影響其效能。

【實施方式】

本發明大體係關於伺服器環境且更特定言之係關於用於該環境中之配接器。呈現以下描述以使得一般熟習此項技術者能夠製造及使用本發明且在專利應用及專利要求的內容中提供以下描述。熟習此項技術者易於瞭解本文中所描

述之較佳實施例、一般原理及特徵的各種修正。因此，本發明並不希望受限於所展示之實施例，而與本文描述之原理及特徵最廣泛地一致。

圖2為根據本發明之伺服器系統100的方塊圖。伺服器系統100包括在記憶體104與介面配接器晶片106之間耦接之處理器102。介面配接器晶片106包括至處理器102之私有(Gx)匯流排之介面108及主乙太網路配接器(HEA)110。HEA 110接收來自處理器102之訊號且將訊號傳輸至處理器102。

HEA 110為整合乙太網路配接器。在伺服器內之TCP/IP堆疊中提供一組加速器特徵。在處理器102與介面配接器晶片106之間之介面100藉由繞過PCI匯流排且提供啟用解多工及多佇列與封包標頭分離的介面技術而得以簡化。

HEA 110藉由直接連接至GX+匯流排及因此享有極大的頻寬(在866 MHz下為55.42 Gbps)以真正地支援兩個10 Gbps埠之全40 Gbps頻寬而達成未匹配效能水平。注意一64位元PCI-X 133 MHz匯流排受限於8.51 Mbps且需要至少一PCI Express x8匯流排以匹配兩個10 Gbps埠之產出。在GX匯流排上亦移除中間邏輯且因此改良轉移等待時間。

在該種情況下，提供乙太網路配接器以當考慮與遺留伺服器環境之相容性時考慮具有高速系統之改良的功能性。此改良之功能性的某些關鍵特徵在下文中描述。

加速功能

HEA 110支援先進加速特徵。一關鍵觀察為當前加速功

能在傳輸側(意即，自處理器傳輸封包)上充分執行而在接收側(意即，經由配接器接收封包)執行不充分。HEA 110藉由引入諸如封包解多工及多佇列及標頭分離之新特徵而定址此間隙。

所有HEA 110新特徵係可選擇的；若要求且當要求時由TCP/IP堆疊利用其。舉例而言，TCP/IP堆疊可使用HEA 110且利用HEA之諸如產出、低等待時間及虛擬化支援的其他特徵。

封包解多工及多佇列

多佇列及解多工為支援諸如虛擬化、每一連接佇列及OS旁路之功能的關鍵特徵。HEA解多工使用佇列對、完成佇列及事件佇列之概念。已添加增益以較佳地處理OS協定堆疊要求及短封包等待時間減少。

視系統要求及組態而定，HEA可解多工傳入之封包，其基於：

- 目的MAC位址(通常為每一分割區有一MAC位址及一預設佇列)。

- 用於建立連接之連接識別符(協定、源IP位址、目的IP位址、源埠、目的埠)。

- 用於TCP連接設置封包(SYN)之目的埠及視需要的目的IP位址。

封包標頭分離

HEA 110能夠自資料有效負載分離TCP/IP標頭。此特徵允許標頭導引至協定堆疊用於處理而不會污染由應用程式

發送的所接收緩衝區。

增強的特徵

在伺服器環境中，HEA 110提供許多增強的特徵。此等特徵中之某些在下文中列出。

1.多接收佇列：佇列對概念延伸以每一對支援一個以上接收佇列。此使得堆疊較佳地管理其緩衝池記憶體。舉例而言，可將一佇列指派至小封包、將一佇列指派至中型封包且將一佇列指派至大封包。HEA 110將根據所接收封包大小選擇特別的佇列。

2.低等待時間佇列：在傳輸側，描述符(WQE)可含有即時資料，在該種狀況下無間接，意即，不要求來自系統記憶體之額外DMA來發送資料。在接收側，低等待時間佇列不提供緩衝區而是接收即時封包資料。HEA 110直接寫入接收佇列。短封包利用此導致DMA運作急劇減少之特徵：每一封包有一單一DMA寫入，如與每一封包有一DMA讀取及一DMA寫入相對。

3.亦使用接收低等待時間佇列以支援封包標頭分離：當有效負載經DMA至特別接收佇列中指示之緩衝區中時，標頭寫入低等待時間佇列中。

總之，解多工、多佇列及封包標頭分離為虛擬化及提供低等待時間運作之基本構建組塊。此外，應注意此等特徵亦可用於改良傳統OS協定堆疊效能，例如，每一連接佇列考慮程式碼之移除且更重要的是減少經消耗以在記憶體中定位適當資訊之記憶體存取及相關聯之遲延/快取記憶體

污染。

為更詳細地描述HEA 110之特徵，現結合附圖參考以下描述。

圖3為根據本發明之HEA 110的簡單方塊圖。正所示，HEA 110具有一個三層架構。第一層包含一媒體存取控制器(MAC)及串列化/去串列化(串化器/解串化器)層202，層202提供來自乙太網路上之其他設備的複數個介面且將複數個介面提供至乙太網路上之其他設備。在層202中，相同晶片I/O用以提供複數個介面。舉例而言，在一較佳實施例中，利用相同晶片I/O以提供一個百億位元介面或十億位元介面。

第二層包含封包加速及虛擬化層204。層204準備接收封包及解多工用於啟用虛擬化之封包的流動。層204啟用基於封包之伺服器運作系統的虛擬化或分割。層204亦提供封包標頭分離以啟用零複本運作且因此提供改良之等待時間。同樣因為層204經由主介面層206直接與私有匯流排(Gx)互相作用，所以提供低等待時間、高頻寬連接。

第三層包含主介面層206。主介面層206提供至Gx或處理器之私有匯流排的介面且與介面204通信。層206給每一佇列對(QP)提供多接收子佇列以啟用TCP堆疊之有效緩衝區管理。主機層206提供用於給定資料封包流的內容管理。

為更詳細地描述HEA 100之層202、204及206中之每一者之特徵，現結合附圖參考以下論述。

MAC及串化器/解串化器層202

圖 4 為具有 MAC 及串化器/解串化器層 202 之更詳細視圖之 HEA 100 的方塊圖。正如此實施例中所示，有一個百億位元 MAC 302 及四個十億位元 MAC 304a 及 304b。MAC 302、304 及 304b 包括用於對準及編碼所接收封包之類比編碼單元 308a、308b 及 308c。MAC 302、304a 及 304b 耦接至高速串聯器/去串列化(HSS)306。高速串化器/解串化器 306 能夠接收來自一個百億位元源或四個十億位元源之資料。

接收乙太網路功能(RxNet)概述

此部分通過層 202 內之接收乙太網路功能展示高水準結構及流程。如將在下文中更詳細解釋之 Rx 加速器單元 400 為封包加速及虛擬化層 204 的部分。

圖 5 展示用於 RxNet 之一的成分及資料流。資料到達介面 302 且由高速串化器/解串化器 304、類比編碼單元 308a 及 308b 與 MAC 處理，該 MAC 在 64 位元(10 G)或 32 位元(1 G)平行資料匯流排中在此實施例中裝配及對準封包資料。亦產生指示訊框之開始及結束以及其他封包資訊的控制訊號。資料及控制通過 RxAccel 單元 400，該 RxAccel 單元 400 在準備用層 206(圖 2)之接收封包處理器(RPP)進行處理時執行剖析、過濾、總和檢查碼及查找功能。在此實施例中，時脈轉換為 4.6 ns 時脈且當資料進入 RxAccel 單元 400 中時資料寬度轉換為 128b。

當資料經由 RxAccel 單元 400 流至主機層 206 內之資料緩衝區時，RxAccel 單元 400 探聽控制及資料且開始其處理。資料流在 RxAccel 單元 400 中受到延遲以使得 RxAccel 單元

400之結果與封包末端同步。此時，RxAccel單元400之結果連同一些來自於MAC 302之原始控制資訊傳遞至一指令佇列。此控制資訊連同資料儲存於緩衝區中。

若RxAccel單元400不具有快取之查找入口，則其需要經由GX匯流排介面進入主記憶體。GX匯流排在4.6 ns下運作。主機層206可非同步地讀取來自RxAccel單元400之佇列對解析資訊。

傳輸乙太網路功能(TxEnet)概述

此部分透過乙太網路及加速功能提供傳輸結構及流程的概述。如將在下文中更詳細解釋之Tx加速器單元500為封包加速及虛擬化層204之部分。

圖6展示一TxEnet之成分及資料流。封包資料及控制自HEA 110之TxAccel 500成分到達。Tx加速器(TxAccel)單元500解譯控制資訊且修正在流過單元500之封包之標頭中的欄位。其基於控制資訊或在封包標頭中發現之資訊做出回繞對埠之決定。其亦產生用於TxMAC 302及304之適當控制。資料流在TxAccel單元500中遲延以使得TxAccel單元500在流入MAC 302及304之前可更新封包標頭。在出口處，資料寬度自128位元轉換為64位元(10 G)或32位元(1 G)。資料及控制通過在TxAccel單元500中之時脈轉換功能，以進入MAC 302及304之不同時脈領域。MAC 302及304、類比轉換器508a及508b與高速串化器/解串化器306格式化用於乙太網路介面之封包。

封包加速及虛擬化層204

圖 7 為具有封包加速及虛擬化層 204 之更詳細視圖之 HEA 110 的方塊圖。HEA 層 204 包含前述接收 (RxAccel) 加速單元 400 及傳輸加速 (TxAccel) 單元 500。RxAccel 單元 400 包含接收骨幹 (RBB) 402、剖析過濾總和檢查碼單元 (PFC) 404、查找引擎 (LUE) 406 及 MIB 資料庫 408。TxAccel 單元 500 包含傳輸骨幹 502、查找檢查 504 及 MIB 引擎 506。Rx 加速單元 400 及 Tx 加速單元 500 之運作將在下文中更詳細地描述。

接收加速 (RxAccel) 單元 400

此部分通過 RxAccel 單元 400 描述高水平結構。圖 8 為 RxAccel 單元 400 之更詳細視圖。RxAccel 單元 400 包括接收骨幹 (RBB) 402、剖析、過濾及總和檢查碼單元 (PFC) 404、局部查找單元 (LLU) 406、遠端查找單元 (RLU) 408 及 MIB 資料庫 410。

資料自未改變之接收 MAC 流過 RxAccel 單元 400。RBB 402 管理資料的流動且負責時脈及資料匯流排寬度轉換功能。自接收 MAC 接收之控制及資料由 PFC 404 使用以執行加速功能且做出丟棄決定。PFC 404 將自包括 5-元組鍵之訊框提取之控制及資料傳遞至 LLU 406，以解決 RBB 402 的佇列對數目 (QPN)。LLU 406 或者立即發現 QPN 或者配置快取記憶體入口以保留槽。若當前鍵不在快取記憶體中，則 LLU 406 在主記憶體中搜索鍵。PFC 404 介面連接至 MIB 資料庫 410 以儲存封包統計。

Tx 加速 500

此部分通過傳輸加速單元 500 (TxAccel) 描述高水平結構

及流程。

圖 9 展示 TxAccel 單元 500 由兩個傳輸骨幹 (XBB) 502a 及 502b、兩個傳輸總和檢查碼單元 (XCS) 504a 及 504b、兩個傳輸 MIB 506a 及 506b、一回繞單元 (WRP) 508 及一暫停單元 (PAU) 邏輯 510 組成。資料自 ENop 流過 TxAccel 且經修正以調整 IP 及 TCP 總和檢查碼欄位。XBB 502a 及 502b 管理資料流動且負責時脈及資料匯流排寬度轉換功能。自 ENop 接收之控制及資料由 XCS 504a 及 504b 使用以執行總和檢查碼功能。在封包由 MAC 層 202 傳輸 (或丟棄) 後，傳輸狀態返回至用於計數之 TxAccel。XBB 502 將資訊轉換至 TxAccel 單元 500 之時脈領域。狀態資訊藉由 XCS 504 與自封包獲得之原始資訊合併且傳遞至 MIB 計數器邏輯 506a 及 506b。MIB 邏輯 506a 及 506b 在 MIB 陣列中更新適當計數器。回繞單元 (WRP) 508 負責轉移至 XCS 504a 及 504b 已決定回繞的接收側封包。暫停單元 (PAU) 510 基於接收緩衝區之佔用而命令 MAC 傳輸暫停訊框。

主介面層 206

圖 10 為具有主介面層 206 之更詳細視圖之 HEA 110 的方塊圖。主介面層 206 包括用於接收來自層 204 之封包且將封包提供至層 204 的輸入及輸出緩衝區 602 及 604。層 206 包括用於適當處理輸入緩衝區中之封包的接收封包處理器 (RPP) 606。內容管理機制 908 在啟用 TCP 堆疊之有效緩衝區管理之前給每一佇列提供多個子佇列。

解多工功能

層 204 之 Rx 單元 400 結合主介面層 206 之成分解多工封包以確保其提供至處理器之適當部分。因此，接收封包必須經解多工以確保其流向伺服器的適當部分。

為描述此解多工功能之細節，現結合圖 8 及圖 9 參考以下內容。

在 HEA 配接器上之解多工實施

在接收封包處理器 (RPP) 606 對接收封包起作用之前，必須擷取佇列對內容。QP 連接管理器使用 QP 號擷取佇列對內容。因為 QP 號不在 TCP/IP 封包中傳輸，則該號必須用其他方式確定。有兩種一般種類之 QP，每一連接 QP 及預設 QP。

每一連接佇列對 (QP)

希望每一連接 QP 用於長久連接，其中不期望有 IP 封包之斷片且期望有低等待時間。要求應用程式支援由 HEA 110 提供之使用者間隔佇列機制。在此實施例中，邏輯埠必須使用目的 MAC 位址而首先被發現。存在三種類型之查找用於每一連接 QP：

1. 用於特定目的 IP 位址及目的 TCP 埠之新 TCP 連接。若封包為 TCP SYN 封包，則基於 TCP/IP (DA、DP、邏輯埠) 執行查找。

2. 僅用於特定目的 TCP 埠之新 TCP 連接 (不顧 DA)。若封包為 TCP SYN 封包，則基於 TCP/IP (DP、邏輯埠) 執行查找。

3. 現有 TCP/UDP 連接。若封包為非斷片單播 TCP 或 UDP

封包，則基於TCP/IP 5-元組加邏輯埠執行查找。

預設佇列對

若未發現用於封包之每一連接QP，或若每一連接查找未經啟用用於MAC位址或若封包為再循環多播/廣播封包，則使用預設QP。通常，預設QP由核心網路堆疊處理。此類型之預設QP 存在於HEA 110中：

1. 每一邏輯埠之預設OS佇列。邏輯埠對應於具有其自己的預設佇列的邏輯乙太網路介面。每一邏輯埠在邏輯轉換器上具有一分離埠。

基於MAC位址執行查找。

對預設OS佇列之直接索引(邏輯埠號)具備再循環(回繞)多播/廣播封包。

2. 多播(MC)或廣播(BC)佇列。

若封包為與MAC查找表中之MAC位址之一不匹配的多播或廣播封包，則使用一組態值。

3. 超預設單播(UC)佇列。

若UC封包與組態之MAC位址之一不匹配，則可使用預設UC QPN。

此機制考慮在每一連接之佇列與每一邏輯埠之佇列(OS佇列)的兩個末端之間的靈活性。兩種模型均可與具有其自己之佇列的一些連接及與預設邏輯埠佇列排隊的一些連接一起運作。

連接查找由RxAccel單元400執行。存在一個該單元用於每一埠組。在RxAccel單元400內，每一成分執行處理之一

部分。PFC 404自封包標頭提取需要之欄位且基於目的MAC位址確定邏輯埠數目。局部查找單元(LLU)406及遠端查找單元(RLU)408接著負責分解QP號。LLU 406試圖僅使用局部資源(快取記憶體及暫存器)來發現QPN。

LLU 406之目的為試圖確定與所接收封包相關聯之QP號。VLIM及RPP 606需要QP號。若可能(意即，沒有進入系統記憶體)，則局部執行此任務。

QP號可用若干方式之一局部地發現：

- 在TS快取記憶體中查找
- 預設分割區QP
- 預設UC QP

若局部未發現匹配，則進行初步檢查以察看入口是否可能存在於系統記憶體中。若存在，則調用RLU 408以執行搜索。若RLU 408繁忙，則可形成一要求佇列，該佇列在RLU 408空閒時提供給RLU 408。

LLU 406與提供QP號及/或用於臨時排隊之佇列索引的RBB 402通信。若在快取記憶體中無適當入口可用，則LLU 406向RBB 402指示搜索繁忙。封包必須在此種狀況下丟棄。

當佇列索引解析受到請求且得到解決時，LLU 406將QPN提供至主機層406。RLU 408試圖使用系統記憶體表來發現QPN。

LLU 406利用局部64入口快取記憶體，以發現TCP/UDP封包之QPN。若在快取記憶體中發現入口，則無需調用

RLU 408。若在快取記憶體中未發現入口，則在快取記憶體中進行初步檢查以察看入口是否可能在連接表中。當有少量組態佇列時，快取記憶體用於消除對主記憶體之不必要存取。

若調用RLU 408，則RLU 408使用一6-元組雜湊(包括邏輯埠號)以自記憶體獲取一128字組直接表(DT)入口。此DT入口含有高達八個6-元組模式及相關聯之QPN。若發現匹配，則不再要求進一步的動作。

當RLU 408必須受到調用時，QPN在工作中得不到確定，因為封包置放入輸入緩衝區中。實際上，QPN在若干封包後得到確定。由於此原因，RxAccel單元400或者提供QPN或者提供對用於封包排隊之主機層206的佇列索引。若提供QPN，則主機層206(卸載器)可用RPP直接將封包排隊用於工作。若提供佇列索引，則主機層206必須保持此封包以等待QPN之解析。QPN總是由調度RPP之時間確定。

虛擬化

因為高速資料路徑可能由多個分割區共用，且因為高速乙太網路效能在此伺服器上係關鍵的，所以HEA之以下行為係至關重要的：

- 提供在多個分割區之間共用之配接器
- 考慮自然效能，意即，"如同專用配接器"
- 考慮自然增值特徵，意即，"如同專用配接器"(大量發送，每一連接佇列，.....)

- 考慮在分割區之間之隔離

- 提供分割區連通性

分割區必須能夠透明地通信，意即，使用相同方式而不管其配置於相同實體伺服器上還是位於由實際乙太網路連接之不同實體伺服器上。

現今，乙太網路虛擬化藉由在擁有配接器之伺服器分割區中轉換或投送而得到支援，此額外跳躍產生效能瓶頸(資料複本、三個驅動器驅動，.....)。HEA 110經設計以在使用分割區與配接器之間提供直接資料及控制路徑(無額外跳躍)。換言之，HEA提供具有其各自"虛擬"配接器及"邏輯"埠之每一分割區。如同HCA，所有HEA資源及功能可在每一分割區受到配置/啟用，精確相同機制用於提供中間分割區保護及隔離。

資料路徑

如圖 11 所示，關於資料路徑，HEA 110提供每一實體埠 902及904之邏輯層2轉換器906及908，以將多播處理及分割區提供至分割區910a-910c通信。在HEA內實施此支援保持整個系統解決方案簡單(詳言之，軟體之透明性)且提供高效能。所有HEA硬體加速及保護可用於分割區以劃分通信。

為支援以上流程，所考慮之習知方式為形象描述邏輯層2轉換器902及904，與給定實體埠相關聯之所有邏輯埠以及實體埠本身附著至邏輯層2轉換器902及904。問題是此邏輯轉換器是如何以及在何處實施的，此邏輯轉換器是如

何以及在何處替代自在韌體/軟體中之完全模擬至在HEA硬體中之完全實施的跨度的。每一實體埠有一邏輯層2轉換器；此等邏輯轉換器未連接在一起。

系統組態

具有遺留OS TCP/IP堆疊之虛擬化HEA

圖 12 展示與遺留 OS TCP/IP堆疊 1102、1104及1106一起使用之HEA。

- 未改變應用程式 1108a-1108c
- 未改變TCP/IP堆疊 1102、1104及1106
- 需要支援HEA 110之設備驅動器 1107a、1107b、1107c

TCP/IP堆疊(OS)可視需要增強以利用諸如用於短封包或每一TCP連接之封包解多工之低等待時間佇列的特徵。如所示，封包解多工基於每一分割區之MAC位址及QPN而執行。

具有遺留OS堆疊及使用者空間TCP/IP之虛擬化HEA

圖 13 展示用於某些分割區支援使用者空間TCP堆疊 1220以及遺留OS堆疊 108a及1208b之系統中的HEA 110：

-需要支援使用者空間TCP之應用程式以使用通訊端擴充功能API

- 其他分割區可使用規則OS TCP/IP堆疊

-在分割區中支援已使用空間TCP之某些應用程式亦可使用規則TCP/IP堆疊(預設路徑)。使用者空間TCP 1220由HEA 119基於用戶識別(Cid)資訊及用於用戶之QPN而解多工。

邏輯轉換器在配接器中受到完全支援。對於多播/廣播封包複製，為使HEA硬體複雜性最小，HEA依賴於硬體實體、多播管理器。HEA為多播管理器提供幫助以將封包複本輸送至目的分割區。

外部單播訊務

傳輸單播訊務通過配置至分割區之QP來處理。其可為每一連接之專用佇列對或每一邏輯埠之單一佇列對或兩者。在發送佇列中之合理排程由HEA提供。視系統組態而定，QP存取可頒予應用程式(使用者空間)或僅頒予OS堆疊(有特權)。

所接收單播訊務如下解多工：

- 在目的MAC位址上執行查找以發現目的邏輯埠
- 若每一連接佇列經啟用以用於目的邏輯埠，則在連接ID上執行第二查找以發現與連接相關聯之QP
- 若每一連接佇列未受到啟用或者連接QP未對於一特定連接而建立，則傳入訊息排隊等待與目的邏輯埠相關聯的"預設"QP。

劃分單播訊務之分割區

圖14為說明包括每一連接佇列之所有HEA加速特徵受到支援的方塊圖。將完全透明性提供至分割區之設備驅動器。

分割區堆疊或者使用每一連接QP或者預設QP以傳輸封包。因為封包由HEA傳輸側處理，所以HEA偵測到目的MAC位址為與界定於相同實體埠上之邏輯埠相關聯的

MAC位址(換言之，目的MAC位址識別屬於與傳輸邏輯鏈接相同之層2邏輯轉換器的接收邏輯鏈接)。因此，HEA回繞封包。HEA接收側接著處理封包，如同其自實體鏈接收一般且因此使用精確相同的加速特徵。

在IP狀況下，IP堆疊可使用規則機制以發現位於相同IP子網上之目的分割區之目的MAC位址。此分割區可配置於相同伺服器上或並非配置於相同伺服器上，此對於堆疊及設備驅動器均係透明的。

劃分多播/廣播訊務之外部及分割區

HEA不將複製多播及廣播封包提供至所關注分割區。實情為，HEA將所有接收MC/BC封包轉發至由多播管理器功能所擁有的QP。此功能複製所需要之封包且使用HEA傳輸能力以將複本分配至所關注分割區。

接收

圖15說明輸入多播傳輸。所接收多播及廣播封包首先通過HEA過濾功能。若未丟棄，則封包導引至多播管理器1500所擁有之QP。封包轉移至系統記憶體且多播管理器1500得到啟動。多播管理器1500確定哪一個邏輯埠應接收一封包複本(多播過濾)且處理封包複製。多播管理器可使用HEA 110以促進將封包重新分配至接受分割區1502a-1502c。

如此做，多播管理器參考所接收封包將n號接受者-描述符(WQF)排入其發送佇列中。注意封包必須完整無缺地傳送至其接受者，詳言之，用各接受者之單播位址替代多播

目的MAC位址是不可接受的。因此，封包描述符必須含有資訊以使得HEA可將封包導引至其適當地目的地。此資訊可為接受者之預設QP或者為其邏輯埠ID或其MAC位址。一旦封包經選擇而發送，歸因於含於QP及WQE中之資訊，HEA傳輸側確定封包需要經回繞而發送。連同資料，確定接受者QP之資訊轉移至HEA接收側。HEA接收側使用此資訊以將對於接受者QP之封包入隊。

傳輸

圖16說明輸出多播傳輸。廣播/多播封包係使用正常程序由原始分割區傳輸。因為封包由HEA傳輸側處理，所以HEA偵測到目的MAC位址受到廣播或多播且偵測到"擠出"選項未在WQE中設置。因此，HEA回繞封包。HEA接收側接著處理上述封包。多播管理器處理上述封包，具有以下添加：

- 必須確保發送者自接受者列表中移除，其確實將封包之源MAC位址用作過濾器。

- VLAN過濾可在封包複製處理期間執行。封包將僅發送至VLAN之部件。

- 一旦內部複製已發生，則封包必須發送出實體埠外。其藉由啟用其QP之擠出功能及在WQE中設置"擠出"位元而如此做。當設置此位元時，HEA在實體鏈接上直接發送出封包。

在多分割區環境中之多播過濾

在接收側，HEA提供多播過濾。如其他"現用"配接器之

HEA基於目的MAC位址之散列值及每一實體埠之在過濾表中之查找提供最佳的效應過濾。此功能的目的是限制多播訊務，但"最終"過濾留給堆疊。在多分割區狀況下，來自所有有關分割區之過濾要求應由多播管理器合併，接著在HEA中組態。

當處理封包分配至所關注分割區時，多播管理器可接著對每一分割區進行多播過濾。

封包標頭分離

HEA 110能夠自資料有效負載分離TCP/IP標頭。此特徵啟用零複本運作且因此改良等待時間。

當在QP內容中組態時，由HEA 110執行封包標頭分離。當組態時，乙太網路/IP/TCP或乙太網路/IP/UDP標頭自封包本體分離且置放於不同記憶體中。通常，TCP/IP堆疊處理標頭且應用程式處理本體。在硬體中之分離允許將使用者資料對準於使用者緩衝區中從而避免複本。

在層204內之PFC 404(圖8)將總標頭長度(8位元)傳遞至指示乙太網路、IP及TCP/UDP標頭之字組數目的主介面206之RPP 606(圖10)。當不執行標頭分裂時，標頭長度設為0。

必須為兩或兩個以上接收查詢(RQ)組態QP。

若封包為TCP或UDP(標頭長度不為零)，則RPP 606將標頭置放入RQ1 WQE。RPP 606接著為封包之資料部分(RQ2或RQ3)選擇一適當RQ。RQ2或RQ3 WQE中之描述符用於置放剩餘資料。RPP 606指示應當產生具有完整資訊之

CQE。設置標頭分裂旗標。在CQE之相關器領域中的相關器自所使用之RQ2或RQ3 WQE複製。置放於第一WQE中之標頭字組的計數亦放在CQE中。

若在RQ1 WQE中標頭大於可用空間，則WQE填充有盡可能多的資料且在CQE中設置標頭過長旗標。標頭之餘項與資料置放於RQ2/RQ3 WQE中。

當標頭分裂模式設置為ALL(所有)且標頭分裂執行(標頭長度非零)時，無封包本體置放在RQ1 WQE中。QP可視需要組態以將短封包整個置放入RQ1 WQE(標頭分裂模式 = ML)中。若如此組態，若封包長度小於RQ2臨限值，則僅使用RQ1 WQE且不執行標頭分離。注意本體在RQ1與RQ2/RQ3之間從未分裂。

若封包為IP斷片或不為TCP或UDP(標頭長度為零)且封包過長而不適合RQ1 WQE，則整個封包使用RQ2或RQ3 WQE置放。標頭計數設置為零。關閉標頭分裂旗標。RQ1 WQE未消耗(除非競爭資訊置放於RQ1 WQE中)。

因此，HEA 110能夠自資料有效負載分離TCP/IP標頭。此特徵允許標頭導引至協定堆疊用於處理而不會污染由應用程式發送之所接收緩衝區且因此減少用於某些異動的等待時間週期。

概述

因此，根據本發明之主乙太網路配接器(HEA)藉由直接連接至處理器之私有匯流排且因此具有足夠頻寬(例如在866 MHz下為55.42 Gbps)以支援兩個10 Gbps埠的全40

Gbps頻寬而達成未匹配效能水平。藉由在處理器之私有匯流排上具有配接器，亦移除中間邏輯且因此改良轉移等待時間。因此，一網路介面控制器(NIC)可利用HEA 110而提供，該網路介面控制器與習知NIC相比考慮更高速度、更低等待時間及更簡單的邏輯。

雖然本發明已根據所展示之實施例進行描述，但是一般熟習此項技術者將易於瞭解實施例可有一些變化且彼等變化在本發明之精神及範疇內。因此，一般熟習此項技術者可進行許多修正而不偏離附屬申請專利範圍的精神及範疇。

【圖式簡單說明】

圖1說明一習知伺服器系統。

圖2為根據本發明之伺服器系統的方塊圖。

圖3為根據本發明之HEA的簡單方塊圖。

圖4為具有MAC及串化器/解串化器層之更詳細視圖之HEA的方塊圖。

圖5展示RxNet之一的成分及資料流。

圖6展示一TxEnet的成分及資料流。

圖7為具有封包加速及虛擬化層之更詳細視圖之HEA的方塊圖。

圖8為RxAccel單元之更詳細視圖。

圖9展示RxAccel單元由兩個傳輸骨幹(XBB)、兩個傳輸總和檢查碼單元、兩個傳輸MIB、一回繞單元及一暫停單元組成。

圖 10 為具有主介面層之更詳細視圖之 HEA 110 的方塊圖。

圖 11 說明給每一實體埠提供邏輯層 2 轉換器的 HEA。

圖 12 展示與遺留 OS TCP/IP 堆疊一起使用的 HEA。

圖 13 展示用於某些分割區支援使用者空間 TCP 堆疊之系統中的 HEA。

圖 14 說明支援包括每一連接佇列之加速特徵的所有 HEA。

圖 15 說明輸入多播傳輸。

圖 16 說明輸出多播傳輸。

【主要元件符號說明】

10	伺服器系統
12	處理器
14	主記憶體
16	匯流排
18	網路介面系統
20	配接器
22	PCI 匯流排
100	伺服器系統
102	處理器
104	記憶體
106	介面配接器晶片
108	介面
110	主乙太網路配接器

202	MAC及串化器/解串化器層
204	封包加速及虛擬化層
206	主介面層
302	百億位元MAC
304a	十億位元MAC
304b	十億位元MAC
306	高速串化器/解串化器
308a	類比編碼單元
308b	類比編碼單元
308c	類比編碼單元
400	Rx加速器單元
402	接收骨幹
404	剖析過濾總和檢查碼單元
406	查找引擎
408	遠端管理單元
410	MIB資料庫
500	Tx加速器單元
502	傳輸骨幹
502a	傳輸骨幹
502b	傳輸骨幹
504	查找檢查
504a	傳輸總和檢查碼單元
504b	傳輸總和檢查碼單元
506	MIB引擎

506a	傳 輸 MIB
506b	傳 輸 MIB
508	回 繞 單 元
510	暫 停 單 元 邏 輯
602	輸 入 緩 衝 區
604	輸 出 緩 衝 區
606	接 收 封 包 處 理 器
902	實 體 埠 / 邏 輯 層 2 轉 換 器
904	實 體 埠 / 邏 輯 層 2 轉 換 器
906	邏 輯 層 2 轉 換 器
908	邏 輯 層 2 轉 換 器
910a	分 割 區
910b	分 割 區
910c	分 割 區
1102	OS TCP/IP 堆 疊
1104	OS TCP/IP 堆 疊
1106	OS TCP/IP 堆 疊
1107a	設 備 驅 動 器
1107b	設 備 驅 動 器
1107c	設 備 驅 動 器
1108a	應 用 程 式
1108b	應 用 程 式
1108c	應 用 程 式
1208a	遺 留 OS 堆 疊

1208b	遺留 OS 堆疊
1402a	分割區
1402b	分割區
1402c	分割區
1450	伺服器
1452	伺服器
1454	實際轉換器
1500	多播管理器
1502a	分割區
1502b	分割區
1502c	分割區
1550	伺服器
1552	伺服器
1554	實際轉換器
1602a	分割區
1602b	分割區
1602c	分割區
1650	伺服器
1652	伺服器
1654	實際轉換器

十、申請專利範圍：

1. 一種乙太網路配接器，其包含：

一架構，其允許該配接器自一處理器接收封包及將封包傳輸至該處理器；該架構包括一解多工機制以考慮該處理器的分割，及複數個層，其包含：

一媒體存取控制器及序列化/去序列化(MAC及串化器/解串化器)層；

一封包加速及虛擬化層，其用於自該MAC及串化器/解串化器層接收封包及將封包提供至該MAC及串化器/解串化器層；及

一主介面層，其用於與該封包加速器及虛擬化層通信且用於與該處理器之一私有匯流排介面連接。

2. 如請求項1之乙太網路配接器，其中該解多工機制考慮一至在該處理器內同時利用之複數個作業系統的直接路徑。

3. 如請求項1之乙太網路配接器，其中該解多工機制考慮一至在該處理器內同時利用之複數個應用程式的直接路徑。

4. 如請求項1之乙太網路配接器，其中該解多工機制考慮一至在該處理器內同時利用之複數個作業系統之直接路徑，且考慮在該處理器內同時利用的複數個應用程式。

5. 如請求項1之乙太網路配接器，其中一邏輯轉換器係用於該配接器之每一實體埠。

6. 如請求項5之乙太網路配接器，其中一邏輯轉換器提供

複數個邏輯埠，其中該複數個邏輯埠中之每一者支援該處理器之一分割。

7. 如請求項1之乙太網路配接器，其中該MAC及串化器/解串化器層進一步包含一用於接收封包之接收乙太網路功能；及一用於傳輸封包之傳輸乙太網路功能。
8. 如請求項1之乙太網路配接器，其中當適當時，一封包標頭自資料有效負載分離以減少用於某些異動之等待時間週期。

9. 一種網路介面控制器(NIC)，其包含：

一介面配接器，其耦接至一處理器之一私有匯流排；及

一乙太網路配接器，其具有允許該配接器自一處理器接收封包及將封包傳輸至該處理器之一架構；該架構包括一解多工機制以考慮用於該處理器的分割，及複數個層，其包含：

一媒體存取控制器及串列化/去串列化(MAC及串化器/解串化器)層；

一封包加速及虛擬化層，其用於自該MAC及串化器/解串化器層接收封包及將封包提供至該MAC及串化器/解串化器層；及

一主介面層，其用於與該封包加速器及虛擬化層通信且用於與該處理器之一私有匯流排介面連接。

10. 如請求項9之NIC，其中該解多工機制考慮一在該處理器內同時利用之複數個作業系統的直接路徑。

11. 如請求項9之NIC，其中該解多工機制考慮一至在該處理器內同時利用之複數個應用程式的直接路徑。
12. 如請求項9之NIC，其中該解多工機制考慮一至在該處理器內同時利用之複數個作業系統之直接路徑，且考慮在該處理器內同時利用的複數個應用程式。
13. 如請求項9之NIC，其中一邏輯轉換器係用於該配接器之每一實體埠。
14. 如請求項13之NIC，其中一邏輯轉換器提供複數個邏輯埠，其中該複數個邏輯埠中之每一者支援該處理器之一分割。
15. 如請求項9之NIC，其中該MAC及串化器/解串化器層進一步包含一用於接收封包之接收乙太網路功能；及一用於傳輸封包之傳輸乙太網路功能。
16. 如請求項9之NIC，其中當適當時，一封包標頭自資料有效負載分離以減少用於某些異動之等待時間週期。
17. 一種伺服器系統，其包含：
 - 一伺服器，該伺服器包括一處理器及一耦接至該處理器之記憶體；及
 - 一網路介面控制器(NIC)，其經由該處理器之一私有匯流排耦接至該處理器；該NIC進一步包括一經由一私有匯流排介面耦接至該私有匯流排的乙太網路配接器，
 - 該乙太網路配接器包含允許該配接器自該處理器接收封包及將封包傳輸至該處理器之一架構；該架構包括一解多工機制以考慮用於該處理器的分割，及複數個層，

其包含：

一媒體存取控制器及序列化/去序列化(MAC及串化器/解串化器)層；

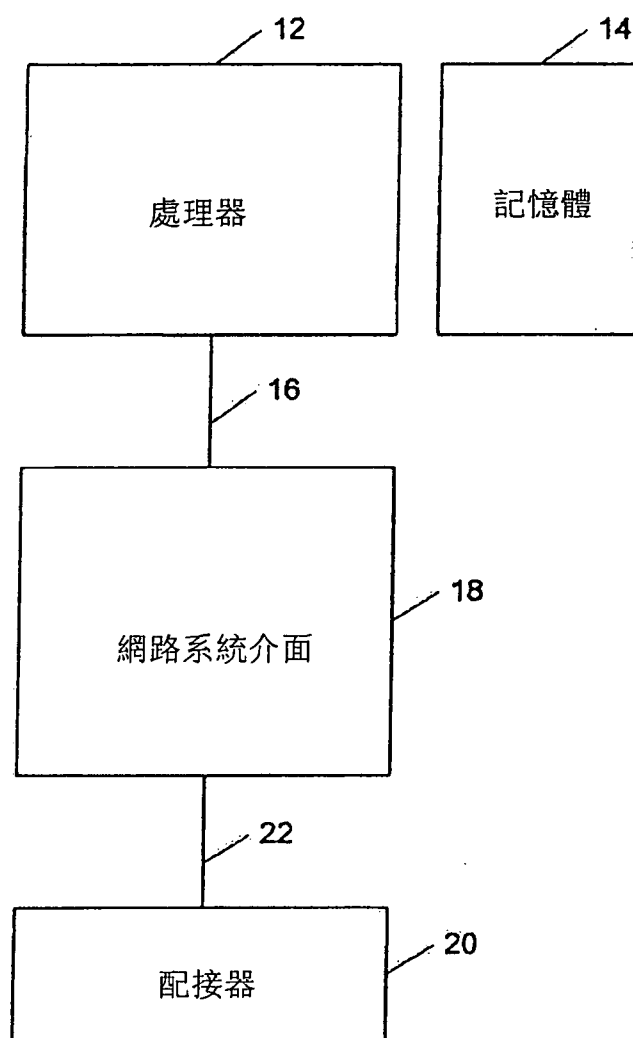
一封包加速及虛擬化層，其用於自該MAC及串化器/解串化器層接收封包及將封包提供至該MAC及串化器/解串化器層；及

一主介面層，其用於與該封包加速器及虛擬化層通信且用於與該處理器之一私有匯流排介面連接。

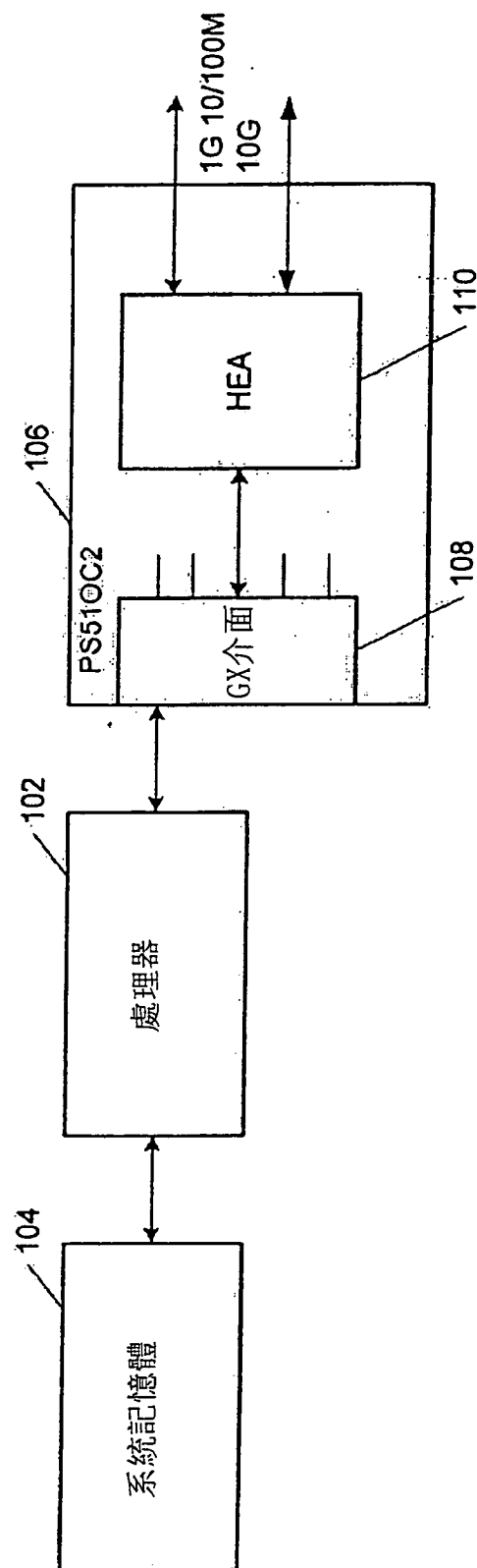
18. 如請求項17之伺服器系統，其中該解多工機制考慮一在該處理器內同時利用之複數個作業系統的直接路徑。
19. 如請求項17之伺服器系統，其中該解多工機制考慮一在該處理器內同時利用之複數個應用程式的直接路徑。
20. 如請求項17之伺服器系統，其中該解多工機制考慮一在該處理器內同時利用之複數個作業系統之直接路徑，且考慮在該處理器內同時利用的複數個應用程式。
21. 如請求項17之伺服器系統，其中一邏輯轉換器係用於該配接器之每一實體埠。
22. 如請求項21之伺服器系統，其中一邏輯轉換器提供複數個邏輯埠，其中該複數個邏輯埠中之每一者支援該處理器之一分割。
23. 如請求項17之伺服器系統，其中該MAC及串化器/解串化器層進一步包含一用於接收封包之接收乙太網路功能；及一用於傳輸封包之傳輸乙太網路功能。
24. 如請求項17之伺服器系統，其中當適當時，一封包標頭

自資料有效負載分離以減少用於某些異動之等待時間週期。

十一、圖式：

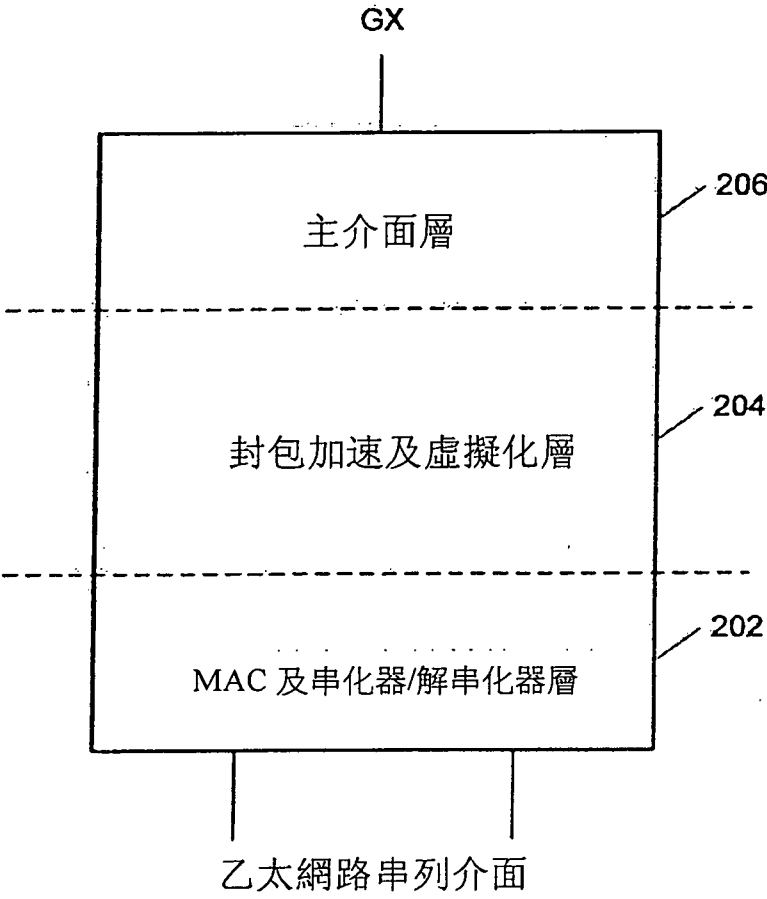


10
圖 1

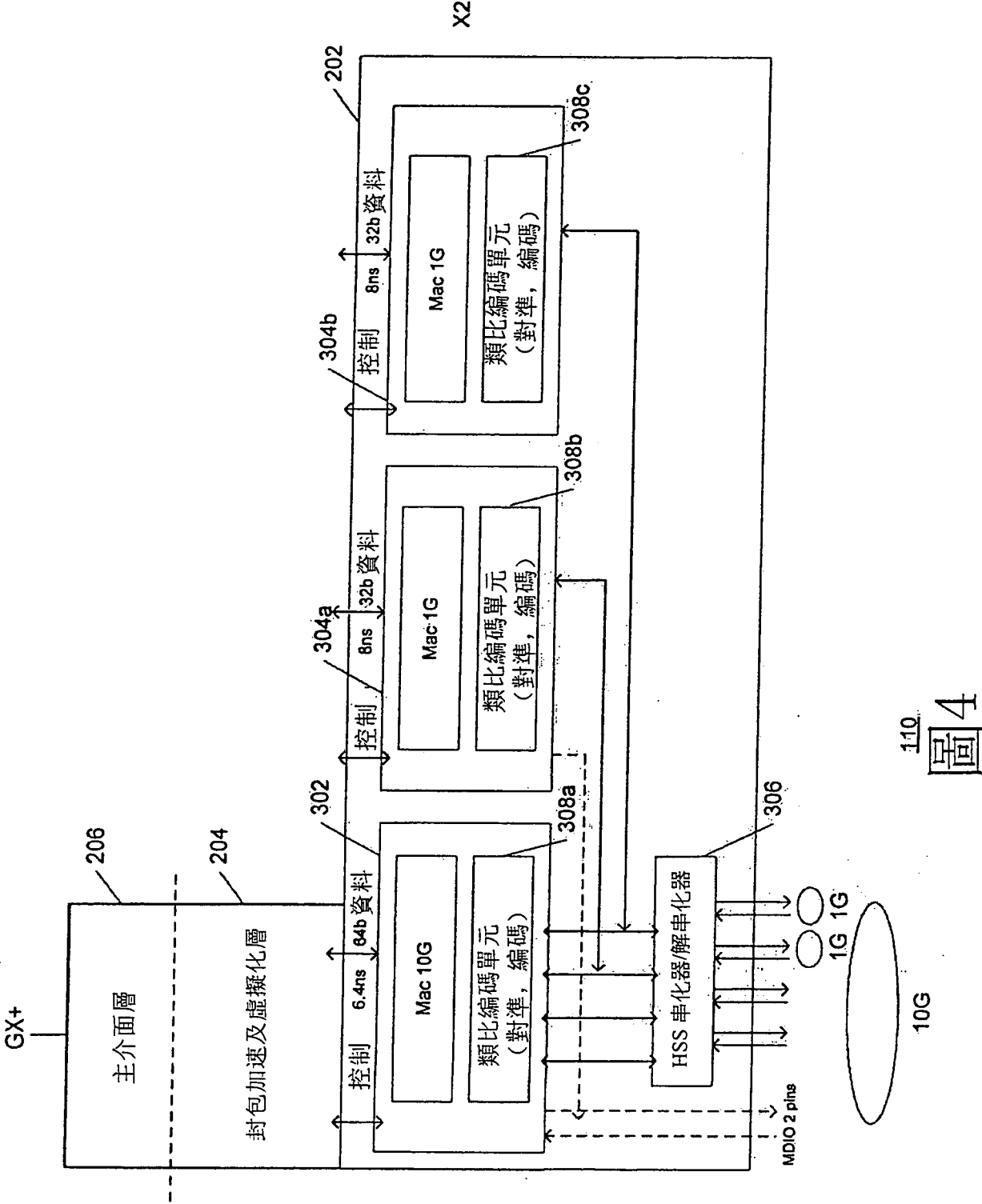


100

圖2



110
圖3



110 圖4

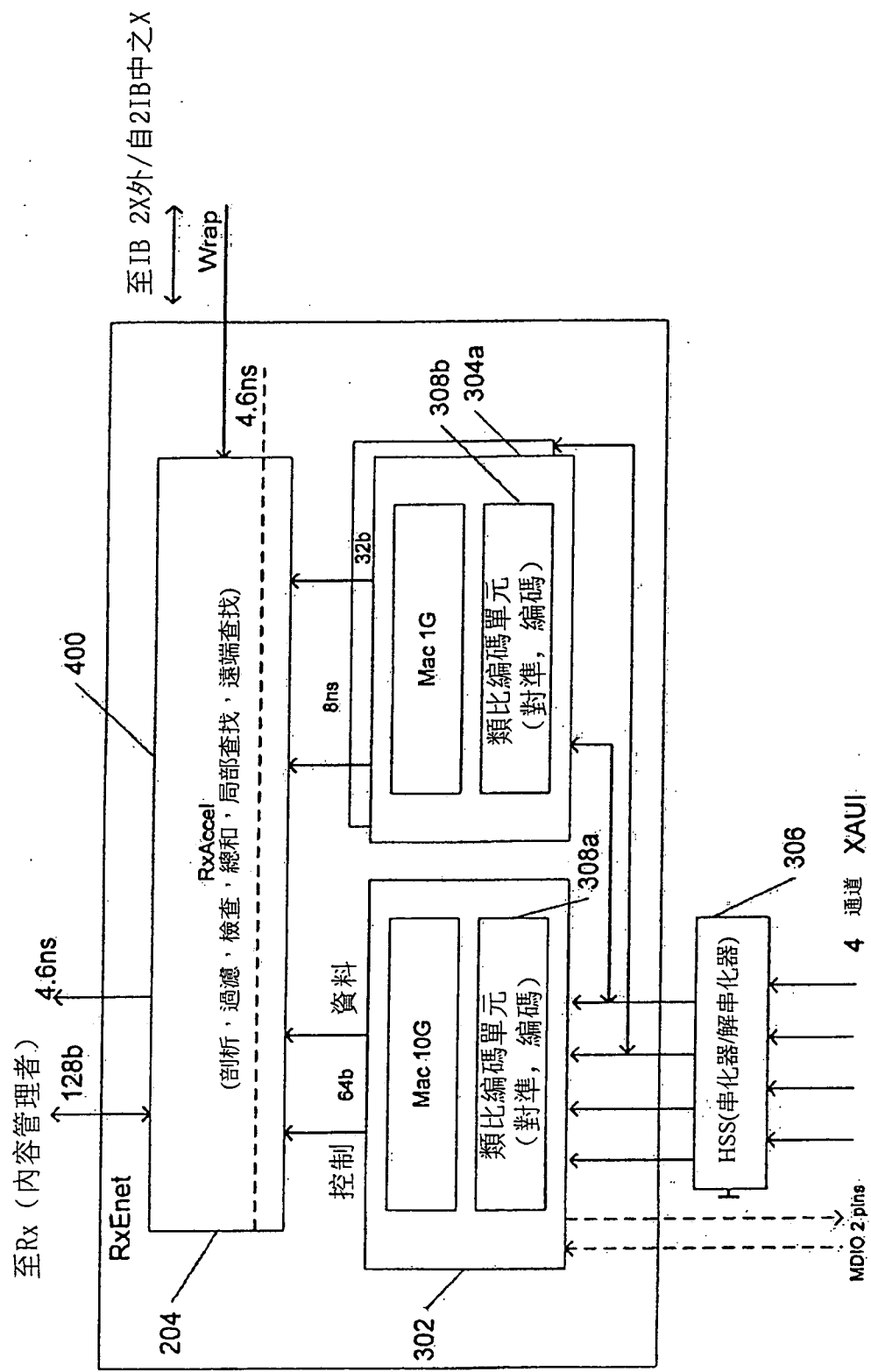


圖5

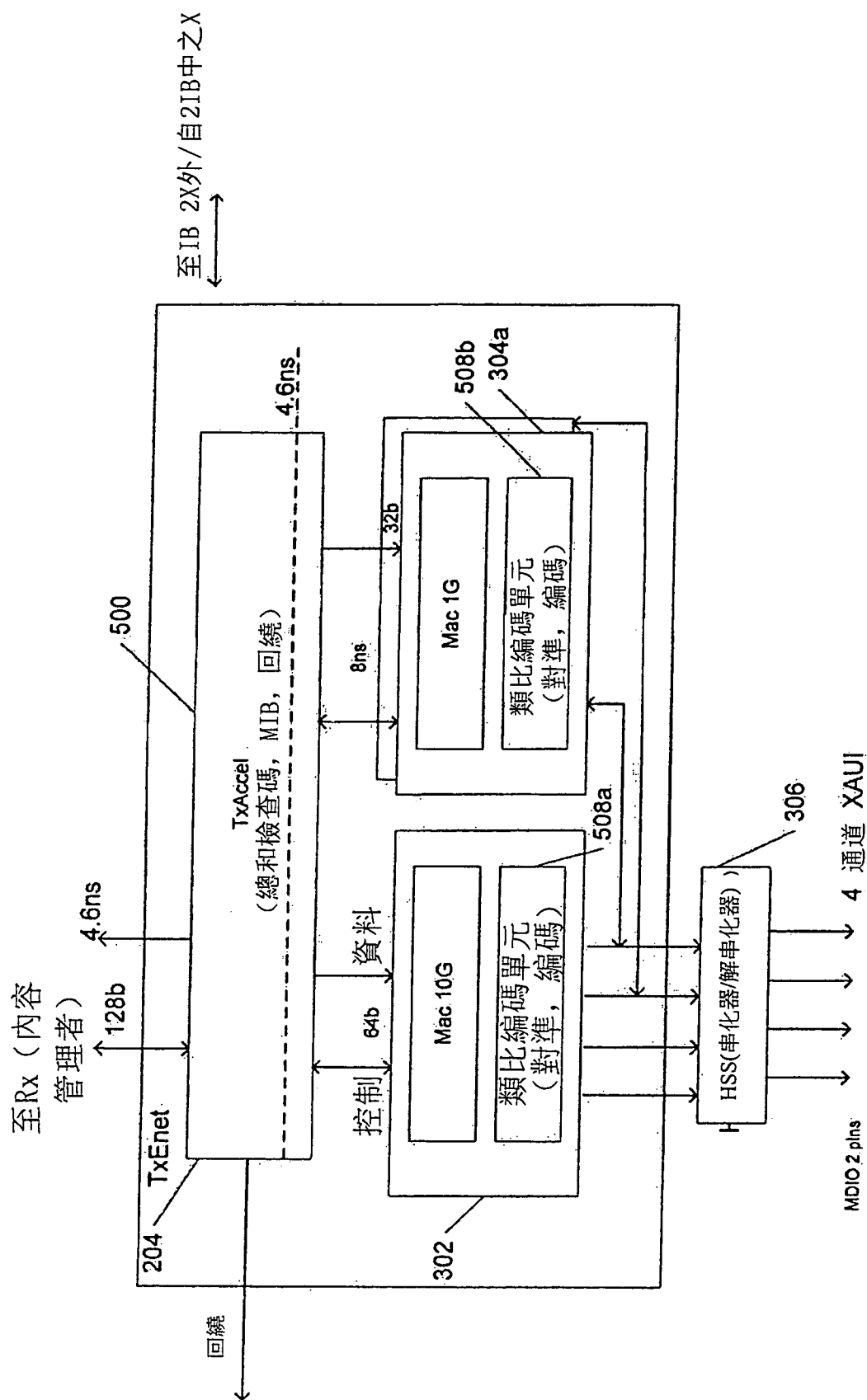
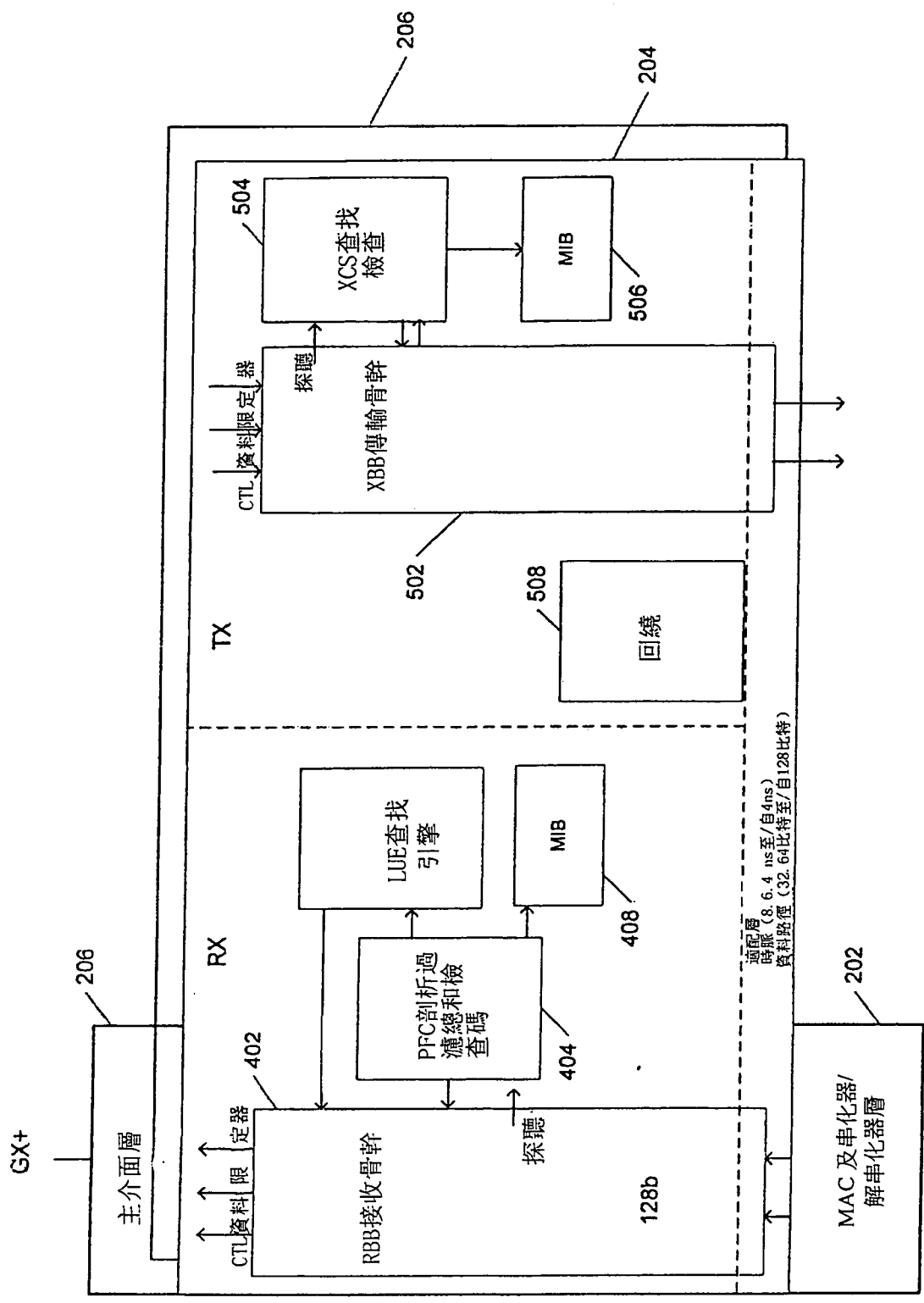
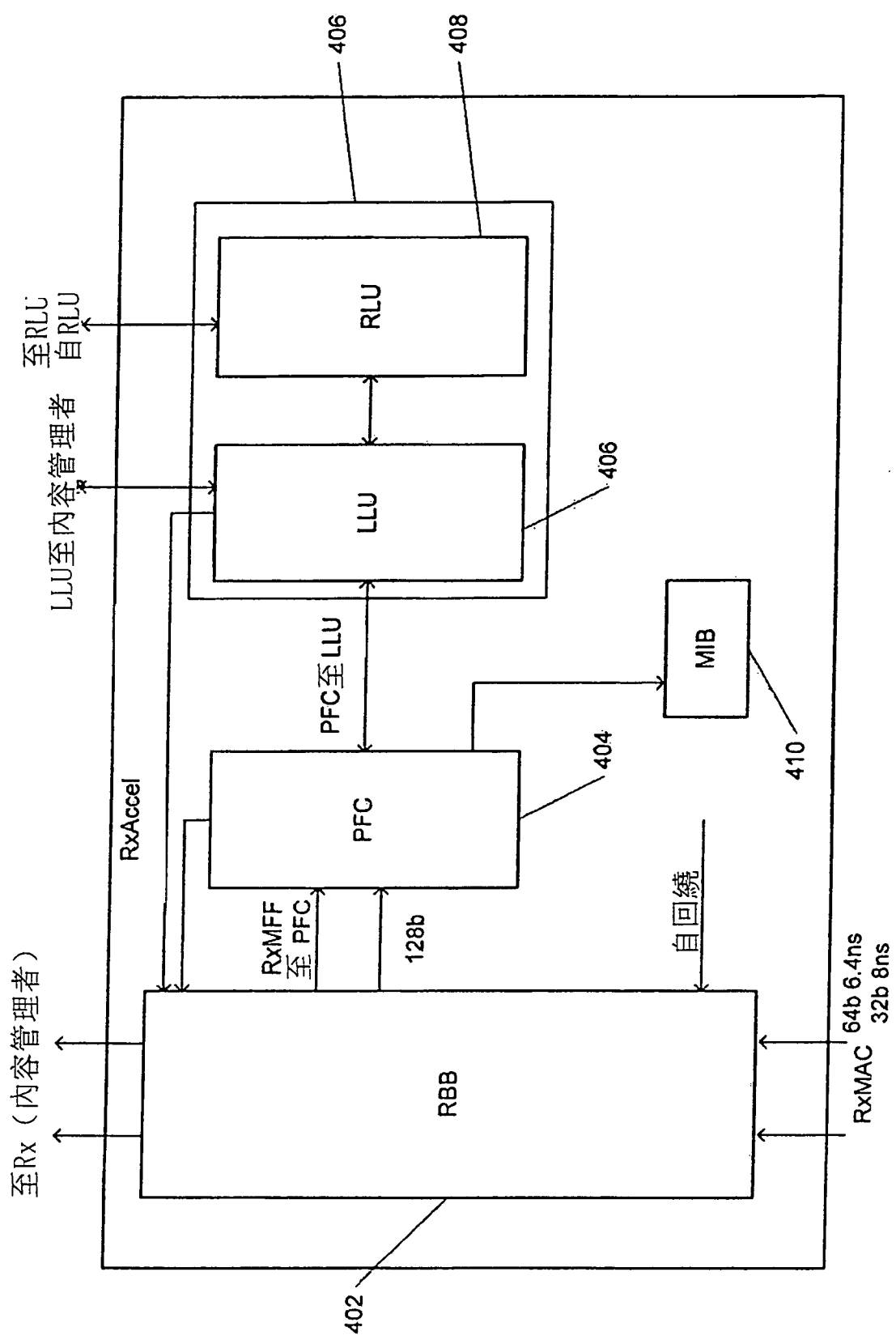


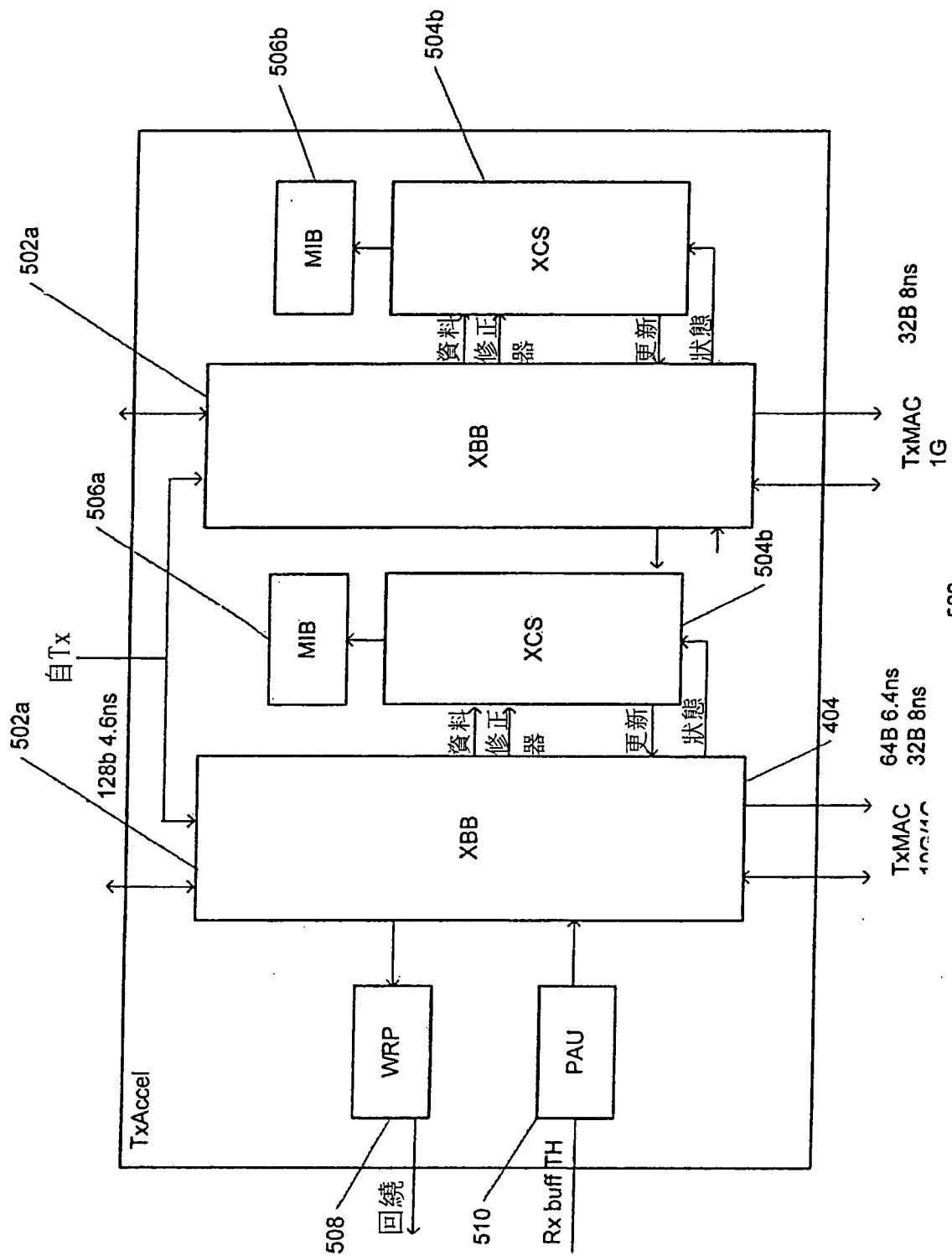
圖6



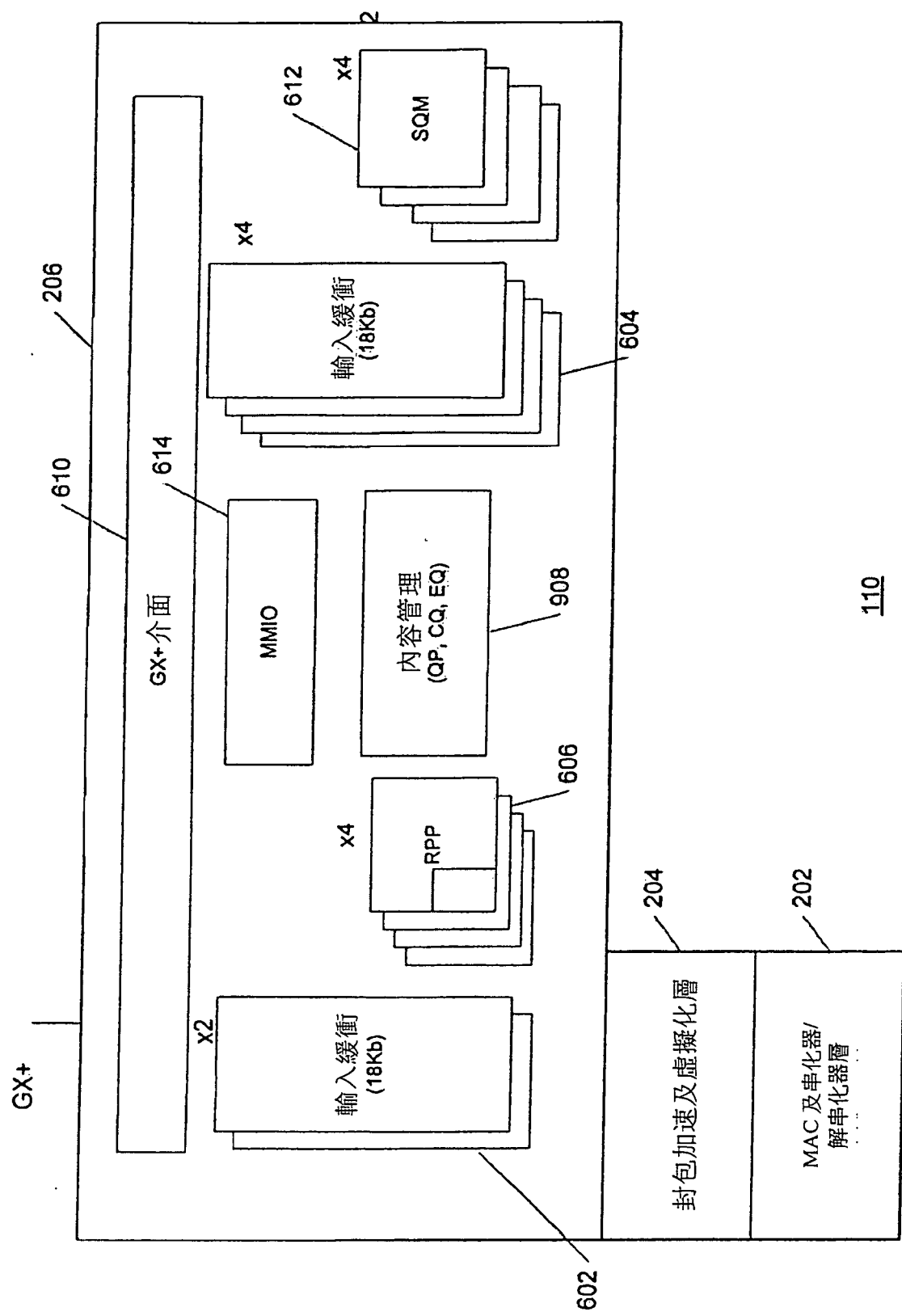
110 圖7



400 圖8



500 圖9



110
圖10

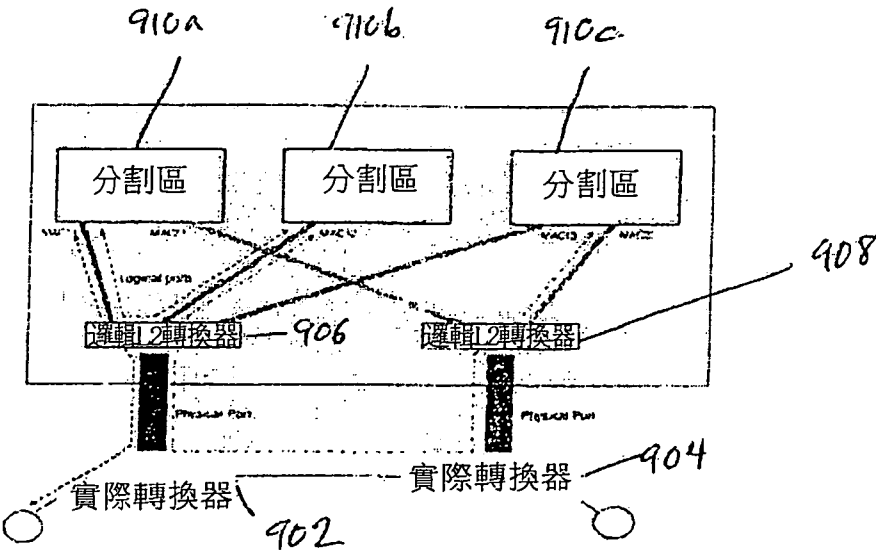


圖11

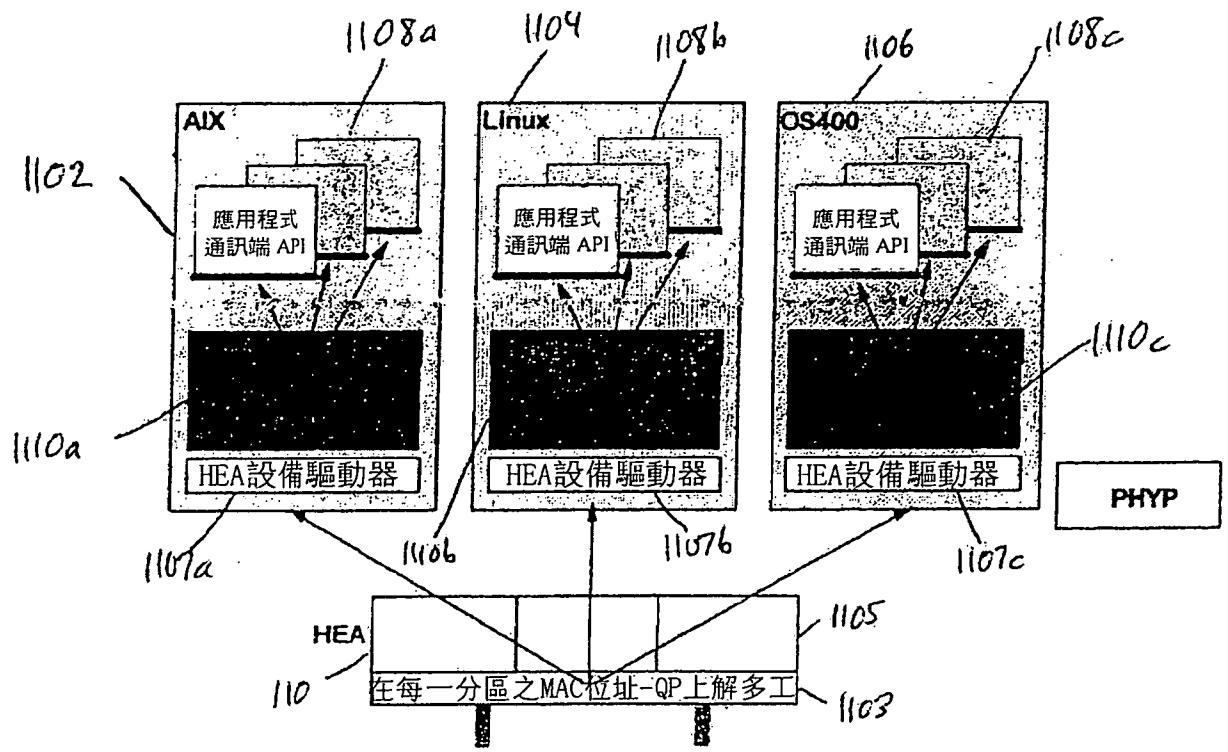


圖12

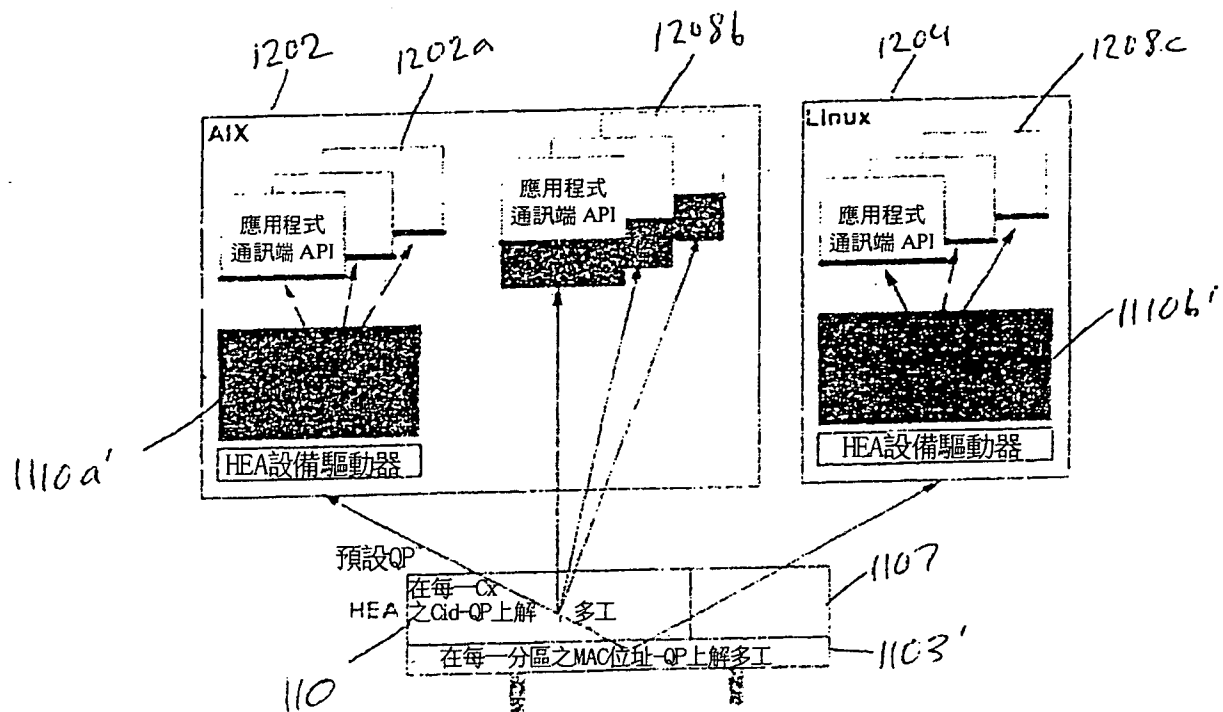


圖13

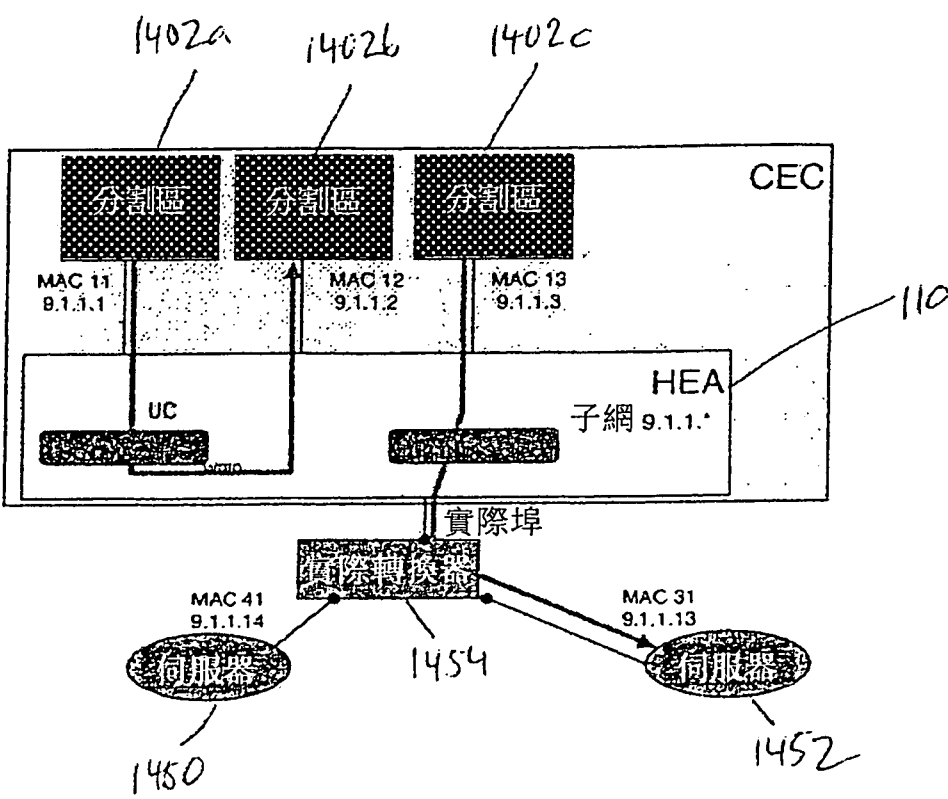


圖 14

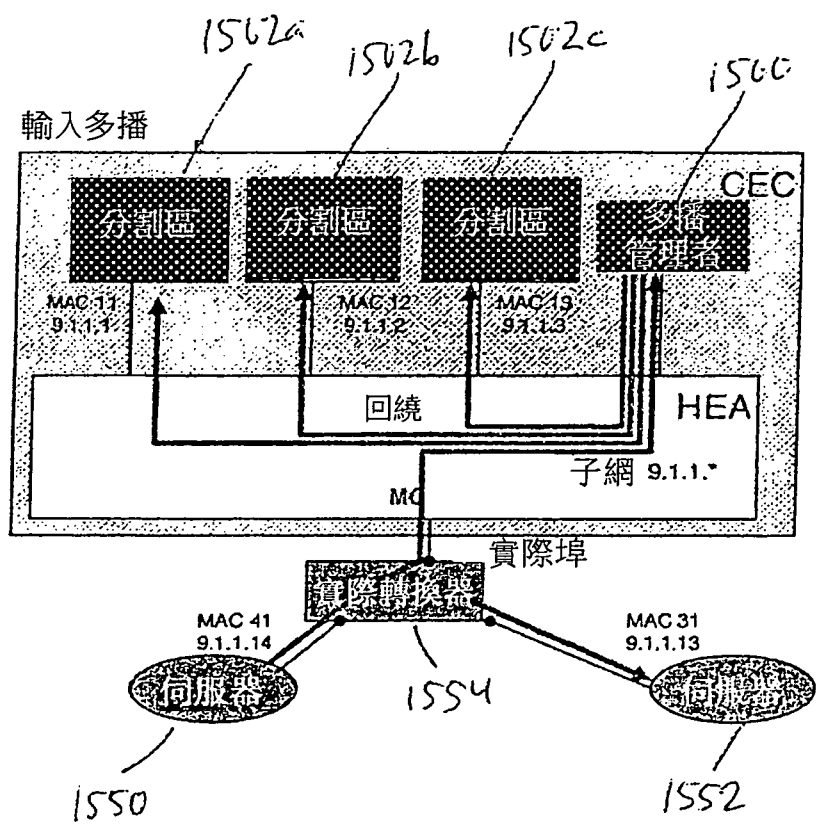


圖15

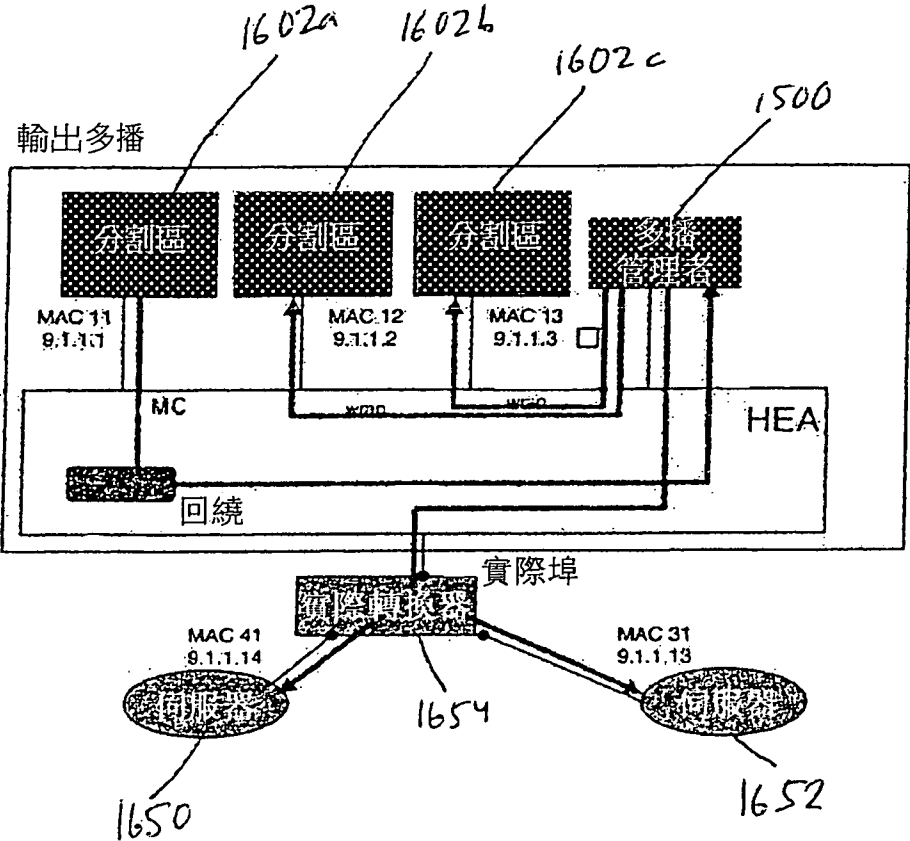


圖16