



(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 I441476 B

(45)公告日：中華民國 103 (2014) 年 06 月 11 日

(21)申請案號：100115092

(22)申請日：中華民國 100 (2011) 年 04 月 29 日

(51)Int. Cl. : **H04L12/00 (2006.01)**

(30)優先權：2010/04/30 世界智慧財產權組織 PCT/US10/33258

(71)申請人：惠普研發公司(美國) HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P.
(US)

美國

(72)發明人：史蓮斯克 麥可 SCHLANSKER, MICHAEL (US)；托瑞西斯 珍 TOURRILHES,
JEAN (FR)；托納 尤須歐 TURNER, YOSHIO (US)

(74)代理人：惲軼群；陳文郎

(56)參考文獻：

US 2008/0285562A1

US 2009/0028172A1

US 2009/0292855A1

審查人員：陳哲賢

申請專利範圍項數：15 項 圖式數：7 共 0 頁

(54)名稱

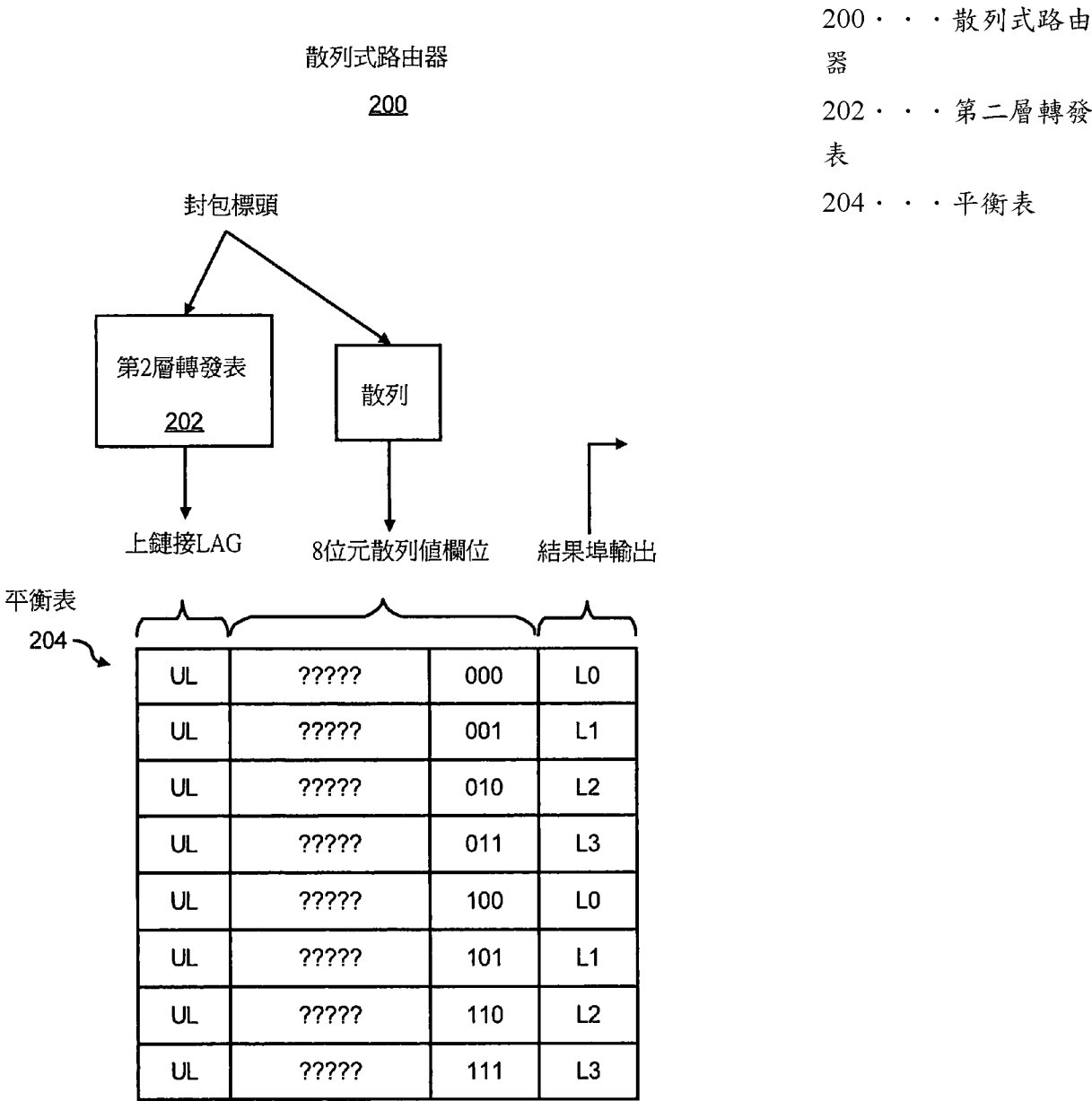
用以安排胖樹網路中之資料封包傳送路由的方法

METHOD FOR ROUTING DATA PACKETS IN A FAT TREE NETWORK

(57)摘要

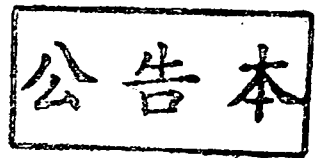
一種用於一胖樹網路中使用一平衡表來安排多個頂部交換器與多個葉片交換器間之封包傳送路由的方法中，可檢測至少一頂部交換器與至少一葉片交換器間之一失效鏈接。此外，該平衡表可根據該檢測之失效鏈接來修改，而該等封包可根據該已修改平衡表來於該胖樹網路中之該等多個頂部交換器與該等多個葉片交換器之間安排傳送路由。

In a method (400) for routing packets between a plurality of top switches (110a-110n) and a plurality of leaf switches (120a-120n) using a balancing table (204, 208, 210) in a fat tree network (100), a failed link between at least one top switch (110n) and at least one leaf switch (120n) is detected (402). In addition, the balancing table (204, 208, 210) is modified (406) based on the detected failed link, and the packets are routed (408) between the plurality of top switches (110a-110n) and the plurality of leaf switches (120a-120n) in the fat tree network (100) based on the modified balancing table (204, 208, 210).



第2A圖

200 . . . 散列式路由
器
202 . . . 第二層轉發
表
204 . . . 平衡表



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)

※申請案號：100115092

H04L 12/00 (2006.01)

※申請日：100.4.29

※IPC 分類：

一、發明名稱：(中文/英文)

用以安排胖樹網路中之資料封包傳送路由的方法
METHOD FOR ROUTING DATA PACKETS IN A FAT TREE NETWORK

二、中文發明摘要：

一種用於一胖樹網路中使用一平衡表來安排多個頂部交換器與多個葉片交換器間之封包傳送路由的方法中，可檢測至少一頂部交換器與至少一葉片交換器間之一失效鏈接。此外，該平衡表可根據該檢測之失效鏈接來修改，而該等封包可根據該已修改平衡表來於該胖樹網路中之該等多個頂部交換器與該等多個葉片交換器之間安排傳送路由。

三、英文發明摘要：

In a method (400) for routing packets between a plurality of top switches (110a-110n) and a plurality of leaf switches (120a-120n) using a balancing table (204, 208, 210) in a fat tree network (100), a failed link between at least one top switch (110n) and at least one leaf switch (120n) is detected (402). In addition, the balancing table (204, 208, 210) is modified (406) based on the detected failed link, and the packets are routed (408) between the plurality of top switches (110a-110n) and the plurality of leaf switches (120a-120n) in the fat tree network (100) based on the modified balancing table (204, 208, 210).

四、指定代表圖：

(一)本案指定代表圖為：第 (2A) 圖。

(二)本代表圖之元件符號簡單說明：

200...散列式路由器

204...平衡表

202...第二層轉發表

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】

相關申請案

本申請案包含某些與2010年4月30日由麥可schlansker、珍tourrihles以及特納良夫(Yoshio turner)申請之美國專利申請案第12/770978號，且名為「使用VLAN來安排資料封包傳送路由的方法」共同的標的，其揭示內容完整地合併以供參考並以附接字母“A”來附接於本文中。

本發明係有關於用以安排胖樹網路中之資料封包傳送路由的方法。

【先前技術】

發明背景

乙太網路交換技術可針對用於資料通訊之大型資料中心與電腦來提供高頻寬、低潛伏期、以及可擴充性。已有許多方法用來開發一乙太網路中之多路徑。例如，該網路可使用第三層網際網路協定(IP)子網路或第二層虛擬區域網路(VLAN)來分區。雖然該等方法會限制純屬第二層網路之範疇與協助開發多路徑，但仍需要複雜且昂貴的人工管理。

發展來增強乙太網路之另一協定為根據胖樹之散列式安排傳送路由。胖樹是以層來建構以使固定數基之交換器可縮放到任意大小與對分頻寬的網路。當一訊息從一來源流至一目的時，需面對一序列多個上鏈接的選擇直到達到一根交換器為止。該等技術針對多重路徑的網路提供了負

載平衡，並保有隨插即用網路之簡單管理，其為透通橋接之學習功能的結果。然而，一或更多鏈接或交換器失效時散列式安排傳送路由會出現一關鍵問題，其會中斷正常的網路通訊。不像僅包括一單一頂部交換器之一規則樹，該胖樹結構包括大量上鏈接與頂部交換器。因此，該胖樹結構中之失效機率會增加。

【發明內容】

依據本發明之一實施例，係特地提出一種用於一胖樹網路中使用一平衡表來安排多個頂部交換器與多個葉片交換器間之封包傳送路由的方法，該方法包含由一處理器執行之下列步驟：檢測至少一頂部交換器與至少一葉片交換器間之一失效鏈接；根據該檢測之失效鏈接來修改該平衡表，其中該平衡表包括根據該等每一封包之散列值來指出一上鏈接輸出埠的一上鏈接虛擬埠、以及與用於安排該等封包傳送路由之該等多個頂部交換器其中之一相關聯的一封包目的位址，其中該等每一封包之散列值可使用該等每一封包之一封包標頭來計算；以及根據該已修改平衡表來安排該胖樹網路中之該等多個頂部交換器與該等多個葉片交換器間之該等封包的傳送路由。

圖式簡單說明

對業界熟於此技者而言，本發明之特徵可從下列說明並參照附圖而變得更加明顯，其中：

第1圖顯示一根據本發明之一示範實施例的一胖樹網路圖；

第2A圖顯示一根據本發明之一示範實施例，用於一胖樹網路中之散列式路由器；

第2B圖顯示一根據本發明之一示範實施例，用於容錯一故障葉片交換器中之一失效鏈接的已修改散列式路由器；

第2C圖顯示一根據本發明之一示範實施例，用於容錯一故障葉片交換器中之一失效鏈接的已修改散列式路由器；

第3圖繪示一根據本發明之一示範實施例，第1圖中描繪之該控制器的簡化方塊圖；

第4圖繪示一根據本發明之一示範實施例，用於安排一胖樹網路中之封包傳送路由的方法之流程圖；

第5圖顯示一根據本發明之一示範實施例，可用來作為一平台以實施或執行第4圖中描繪之該等一或更多程序的一計算裝置之方塊圖。

【實施方式】

較佳實施例之詳細說明

為了簡化與舉例解說，本發明主要參照其一示範實施例來說明。下列說明中，提出許多特定細節來提供對該等實施例之一全面了解。然而，很明顯地對業界熟於此技者而言本發明在不侷限於該等特定細節的情況下亦可加以實作。其他實例中，並不詳細說明著名的方法與結構以避免不必要地混淆該等實施例之說明。

本文揭示內容為一胖樹網路中用於安排一頂部交換器

與一葉片交換器間之封包傳送路由的一方法與控制器。更特別是，本文揭示之該控制器與方法可組配來於一胖樹網路中使用散列式安排傳送路由表，諸如一三位元內容定址記憶體(TCAM)，經由多個頂部交換器與多個葉片交換器間之多重上鏈接來安排封包之傳送路由。此外，如下文中更詳細說明，該等散列式安排傳送路由表可以各種不同方式來修改以適於該等頂部交換器與該等葉片交換器間之失效鏈接。此外，該等封包可根據該等已修改散列式安排傳送路由表，於該胖樹網路中之該等多個頂部交換器與該等多個葉片交換器之間安排傳送路由，以使該胖樹網路中之封包通訊中斷最小化。

透過本文揭示之該方法與控制器的實施態樣，可自動決定該等頂部交換器與該等葉片交換器間之一失效鏈接。本文揭示之該方法與控制器可使所有剩餘之操作鏈接與交換器能夠承載訊務而不需不必要地停用任何頂部或葉片交換器。就本身而言，本文揭示之該方法與控制器可使具有一失效鏈接或一失效交換器之一封包傳送路由能夠以一相當快且有效的方式來容錯。

如本文說明，於傳送與接收裝置間之一網路上，一封包可為一資料封包而一串流資料封包可承載傳輸資料串流之通訊流所需的資訊。封包串流的類型之範例為多玩家遊戲資料、串流視頻或音頻、或資料之大量轉移。該等來源與目的裝置可組配為經由一網路中之一傳送路由安排來傳送或接收封包，而透過該網路或透過另一網路，封包可經

過該傳送路由安排至該目的裝置。

於一胖樹網路中，一上鏈接之一端連接至一頂部交換器而該上鏈接之另一端連接至一葉片交換器。可有超過一個上鏈接連接至一頂部交換器與一葉片交換器。於一胖樹網路中，亦可有一或更多下鏈接連接至一葉片交換器。於一胖樹網路中之多層交換器可使用專門的記憶體架構，諸如一平衡表來配置記憶體表。一適當平衡之一特別範例為一三位元內容定址記憶體(TCAM)表，其為將該等頂部交換器間之訊務分佈的一散列式安排傳送路由表之一範例。

一散列功能可將訊息指派至隨機選擇之散列類別。針對每一來源位址“s”與每一目的位址“d”， $H(s, d)$ 表示從s傳送至d之每一訊息的散列類別。該散列可合併第二層與更高層的資訊。對稱散列具有 $H(s, d) = H(d, s)$ 。此可確保一訊息從一來源流至任何目的後該路徑可由一回應來回掃。

本文說明之方法並不侷限於僅應用於一胖樹網路中。該等方法亦可應用於其他有線或無線網路，諸如有線或無線電信網路、電腦網路、或其他類型的網路。

首先參照第1圖，其顯示根據一範例之一胖樹網路100的圖形。應清楚了解該胖樹網路100可包括額外的組件，而在不違背該胖樹網路之範疇的情況下，本文說明之某些該等組件可被移除與/或加入。就本身而言，該胖樹網路100可包括任何數量的頂部交換器、葉片交換器、上鏈接、下鏈接、來源裝置、以及網路裝置。此外，該胖樹網路100可連接至一接取網路或任何廣域網路(WAN)或其他網路。

該胖樹網路100包括多個頂部交換器110a-110n與多個葉片交換器120a-120n。該胖樹網路100亦包括多個上鏈接。該等多個上鏈接之每一個將該等頂部交換器110a-110n的其中之一連接至該等葉片交換器120a-120n的其中之一。於一實施例中，該等頂部交換器110a-110n的其中之一，例如，該頂部交換器110a，可針對將該等葉片交換器120a-120n互連之一跨距樹來使用。該跨距樹適合支援可附接於該等葉片交換器120a-120n之額外乙太網路交換器的加入與重新組配。該等剩餘頂部交換器並不參與該跨距樹協定但提供可由散列式安排傳送路由使用之額外冗餘路徑來傳輸封包。另一實施例中，若支援該跨距樹協定之該頂部交換器110a失效，則可決定從該等剩餘頂部交換器中的另一頂部交換器開始的一新跨距樹。

該胖樹網路100更包括多個下鏈接。如第1圖所示，該等多個下鏈接連接至該等葉片交換器120a-120n。此外，一控制器130組配來透過應用散列式安排傳送路由表來安排該等頂部交換器110a-110n與該等葉片交換器120a-120n間之封包傳送路由。如下文更詳細說明，該控制器130組配來檢測一失效鏈接、修改該散列式安排傳送路由表來適於該失效鏈接、以及根據該已修改散列式安排傳送路由表來安排該等頂部交換器110a-110n與該等葉片交換器120a-120n間之封包傳送路由。此外，該控制器130可包含一資料中心網路控制器或一更局部化網路控制器。

如第1圖所示，該最右邊頂部交換器110n與該最右邊葉

片交換器120n間之一鏈接已經以一“x”來表示為失效。例如，若該頂部交換器110n或該葉片交換器120n之一埠失效或者該頂部交換器110n或該葉片交換器120n完全失效時，則一鏈接可能失效。此外，或者，若連接該頂部交換器110n與該葉片交換器120n之埠的纜線損壞或斷線時，則一鏈接可能失效。

一鏈接聚合群組(LAG)中之所有實體鏈接失效時，一鏈接亦可視為失效。如下所述，一LAG之每一實體鏈接將一不同的葉片交換器連接至附接於該LAG之一頂部交換器。第1圖中，不附接於該LAG之剩餘葉片交換器120a-120c視為遠端葉片交換器。直接附接於該失效鏈接之該頂部交換器110n視為一失效頂部交換器。除了該中斷鏈接外，該失效葉片交換器120n與該失效頂部交換器110n兩者可完全運作。

現參照第2A圖，其顯示根據一範例，用於一胖樹網路中之散列式路由器200。應了解第2A圖中描繪之該散列式路由器200係作為舉例解說而因此不應視為將本發明侷限於任何方面中。

一般而言，該控制器130(第1圖)可組配來使用該散列式路由器200來主動管理封包之傳送路由安排經過該胖樹網路100以平衡若干頂部交換器110a-110n間之負載。換言之，該控制器130可組配來接取本文中共同參照為散列式安排傳送路由表之一第二層轉發表202與一平衡表204，並以各種不同方法來安排封包之傳送路由經過該等頂部交換器

110a-110n以實質防止該等頂部交換器110a-110n之任何一個過載而因此降低該胖樹網路100之效能。此外，該控制器130可組配來修改下文說明之平衡表204以實質維持保持作用中之頂部交換器110a-110n間的負載平衡。

該平衡表204中之每一項目包含一散列值欄位、一轉發埠欄位、以及一實際埠欄位。若該散列值與轉發埠欄位與對應的金鑰欄位匹配，則該實際埠欄位被取代並決定該上鏈接。針對每一可能的散列位元組合(例如，針對一8位元散列值欄位會有256種組合)，其散列值欄位中有包含該常數的一項目。每一項目亦可針對該項目之散列值來提供規定該所欲上鏈接之一實際埠欄位。該控制器130可收集交換器統計資訊來獲得用以負載平衡之資料中心的測量。

根據第1圖所示之範例，該胖樹網路100不包含中斷鏈接時，在檢測到任何故障之前，該控制器130使用第2A圖所示之散列式路由器200來安排該等四個頂部交換器110a-110n之封包傳送路由。假設為一對稱的安排傳送路由，於是所有該等交換器110a-110n與120a-120n使用相同的第二層轉發表202與平衡表204。第2A圖中，每一交換器中之該第二層轉發表202包含乙太網路位址與該位址之所需轉發埠的一結合。針對上鏈接，該第二層轉發表202可根據一平衡表204構造之結果來規定該上鏈接虛擬埠以指出該等上鏈接其中之一將用來將資料傳送經過一頂部交換器。於一實施例中，該平衡表204之項目可被建立來一致地分散來自每一葉片交換器120a-120n並標示為L0、L1、L2、與L3

之四個上鏈接的訊務。於每一葉片交換器上具有相同標示之上鏈接會到達相同的頂部交換器。每一封包之散列值與具有特定散列值之直接封包與特定上鏈接匹配。例如，可計算8位元的散列結果。然而，該平衡表204之項目可忽略散列結果的5個位元並與具有二進制數值：000、001、010、011、100、101、110與111之散列結果的三個低階位元匹配。

於一實施例中，一匹配操作可選擇決定使用哪個頂部交換器來傳輸一封包之輸出埠。例如，八個散列類別可以確實兩個散列類別於每個上鏈接上映射的方式而於該等四個上鏈接上映射。該散列映射為對稱，而相同的散列式安排傳送路由表安裝於每個葉片交換器120a-120n中。

現特別參照第2B圖，其顯示一根據一範例，包含一中斷鏈接之局部修改的已修改散列式路由器220。於一實施例中，無訊務經過該失效上鏈接時，即使該等葉片交換器120a-120n與頂部交換器110a-110n的其中之一已被決定具有一失效鏈接後，所有葉片交換器120a-120n與所有頂部交換器110a-110n仍維持使用中。為了適於本實施例之失效鏈接，會計算具有局部修改之新平衡表208項目來重新安排該失效鏈接之封包訊務的傳送路由。例如，針對該失效葉片交換器，葉片交換器120n，無封包透過(第1圖)標示為L3之該上鏈接來傳送或接收。因為該失效葉片交換器之散列式安排傳送路由表直接控制透過該失效鏈接傳送之封包，所以只需要第2A圖所示之該平衡表208少許的局部修改來對該失效上鏈接容錯。

第2A圖所示之該平衡表204具有規定失效上鏈接L3來作為該所需上鏈接之兩個項目(散列低階位元“011”與“111”)。第2B圖所示之該平衡表208中，該等每一項目已被修改而所有具有低階散列“011”之訊務描繪為透過鏈接L0來傳送，而具有低階散列“111”之訊務描繪為透過鏈接L1來傳送。該兩改變可重新安排該失效上鏈接L3附近來自該失效葉片交換器120n之一上鏈接上傳送的所有封包之傳送路由。

現來到第2C圖，其顯示一根據另一範例，包含一中斷鏈接之一局部修改的已修改散列式路由器240。於一實施例中，為了繼續使用該失效頂部交換器，例如，(第1圖)連接至該失效上鏈接之頂部交換器110n，每一遠端葉片交換器可識別附接於該失效葉片交換器之下鏈接的裝置之乙太網路目的位址。因為傳送至該等位址之封包可達到該失效上鏈接之一終端，所以該等封包應可避開該失效頂部交換器。於一範例中，從一遠端葉片交換器，其他目的位址可傳送至包括該失效頂部交換器之任何頂部交換器，因此允許使用該失效頂部交換器來傳送封包至不附接於該失效葉片交換器之目的。

第2C圖繪示保持使用該失效頂部交換器與該故障葉片交換器之情形下，為了正確安排該失效上鏈接訊務之傳送路由，針對第二層轉發表202以及一已修改平衡表210兩者遠端葉片交換器所需之改變。第2C圖中，該第二層轉發表202會隨著列舉連接至該失效葉片交換器之裝置的目的位

址之一中斷鏈接表206而增大。於一實施例中，針對連接至該失效或中斷葉片交換器之每一裝置，需要一機構來建立每一遠端交換器中之該中斷鏈接表的一項目。該項目將該特定裝置位址(例如，Dx)與為人所知之該中斷交換器(或失效交換器)的虛擬埠之一虛擬埠中斷鏈接(“BL”)相關聯。因此，針對每一葉片交換器，若該封包目的位址不附接於該失效葉片交換器，則該正常學習程序會以上鏈接(“UL”)來將其上鏈接目的貼標籤。若該封包目的位址是針對附接於該失效目的交換器之一裝置，則使用該中斷鏈接表之特別程序以“BL”來將其上鏈接目的貼標籤。另一實施例中，使用一未修改的第二層轉發表來處理之該中斷鏈接表206可藉由固定該標準表中之該等特定位址以及將其上鏈接設定為“BL”來執行。

於一實施例中，亦可增大該平衡表210。針對以該“UL”上鏈接虛擬埠標籤來標示之任何封包而言，該增大之平衡表210包含該未修改之正常平衡表，諸如第2A圖所示之該平衡表204，其指出該等封包不附接於該失效葉片交換器。無任何訊務嘗試經過該失效上鏈接時，該等封包可被安排傳送路由至該等四個頂部交換器之任一個。此外，新產生之項目可使用該BL標籤來安排欲前往附接於該失效葉片交換器之目的的封包傳送路由。該等項目包含結果埠指定作業，諸如第2B圖所示之已修改平衡表210，其可避免使用該失效的頂部交換器，而針對具有該“BL”標籤之目的使用該交換器會造成該失效上鏈接的不合法使用。

現特別參照第3圖，其顯示一根據一範例，第1圖描繪之該控制器130的簡化方塊圖300。對業界熟於此技者而言很明顯地該控制器130之方塊圖300代表一歸納圖，並且在不違背該控制器130之範疇的情況下可加入其他組件或可移除、修改或重新安排現存組件。

如第3圖所示，該控制器130包括一上鏈接監控模組302、一交換器監控模組304、一散列式安排傳送路由表更新模組306、以及一安排傳送路由模組308。該等模組302-308可包含軟體模組、硬體模組、或軟體與硬體模組之一組合。因此，於一實施例中，該等模組302-308的其中之一或更多模組包含電路組件。另一實施例中，該等模組302-308的其中之一或更多模組包含由一處理器執行並儲存於一電腦可讀儲存媒體之軟體碼。

該控制器130亦可與一資料儲存器320通訊，其可包含該控制器130可從其擷取資料之任何相當適合的媒體。雖然該資料儲存器320已描繪為形成與該控制器130分開的一組件，但應了解在不違背該控制器130之範疇的情況下該資料儲存器320可與該控制器130整合。

參照第4圖更詳細說明可執行該控制器130之該等模組302-308的各種不同方法，其描繪根據一範例，於一胖樹網路中使用散列式安排傳送路由表來安排多個頂部交換器與多個葉片交換器間之封包傳送路由的一方法400之流程圖。對業界熟於此技者而言很明顯地該方法400代表一歸納圖，並且在不違背該方法400之範疇的情況下可加入其他步

驟或者移除、修改或重新安排現存步驟。

該方法400之說明可特別參照第1圖描繪之該控制器130以及特別參照包含於該控制器130中之組件來完成。然而，應了解在不違背該方法400之範疇的情況下，該方法400可於與該控制器130不同之一裝置中執行。

步驟402中，可由例如，該上鏈接監控模組302來檢測至少一頂部交換器110a-110n與至少一葉片交換器120a-120n間之一失效鏈接。於一實施例中，於步驟402，該上鏈接監控模組302決定一LAG，其為與一對交換器互連之一組實體鏈接。該實施例中，一LAG中之每一實體鏈接將每一葉片交換器連接至一頂部交換器，而該LAG中之每一實體鏈接附接於相同的頂部交換器。如第1圖所示，於一LAG中聚合並作為一單一虛擬鏈接之四個實體鏈接附接於該頂部交換器110a。該等四個不同的實體鏈接之每一個將一不同的葉片交換器120a-120n連接至該頂部交換器110a。一LAG中之任何實體鏈接持續操作時，該LAG可操作，而一LAG中之所有實體鏈接已失效時，該LAG視為已失效。參照第1圖，其描繪之該等上鏈接為LAG而標記“L0”、等等參照為對應該LAG之一虛擬鏈接。每一虛擬鏈接附接於一交換器以及可識別附接於該LAG中之實體鏈接的該組實體埠之一對應虛擬埠。該等下鏈接亦可視為LAG。

第2A圖、第2B圖、與第2C圖中所示之該轉發表202以及該等平衡表204、208、210中的項目皆可參照為對應一LAG之一端點的一虛擬埠。該等平衡表204、208、210決定

於哪個LAG上傳送資料。該下鏈接聚合層之後決定於該LAG中之哪個實體鏈接上傳送資料。

根據一範例，於步驟402，例如，由該上鏈接監控模組302來完成有關該LAG是否包括一失效上鏈接的決定。若該LAG包括失效上鏈接，則如下所述，一或更多封包可經由該LAG中未失效之上鏈接來於該頂部交換器與不同的葉片交換器之間安排傳送路由。

步驟404中，例如，由該交換器監控模組304來完成有關哪個頂部交換器110a-110n與/或葉片交換器120a-120n附接於該失效上鏈接的決定。

步驟406中，例如，由該散列式安排傳送路由表更新模組306根據該檢測之失效鏈接來修改該等平衡表208、210。如上所述，該等平衡表208、210包括根據該等每一封包之散列值來指出一上鏈接輸出埠的一上鏈接虛擬埠。該等每一封包之散列值可使用該等每一封包之一封包標頭來計算。藉由一範例，該等每一封包可根據該等每一封包之不同散列值來指派至該等多個上鏈接之每一個。

於一實施例中，於步驟406，執行該失效葉片交換器中之該平衡表208的一局部修改，如第2B圖所示。另一實施例中，執行所有遠端葉片交換器之該轉發表202與該平衡表210兩者的遠端修改，如第2C圖所示。如上述參照第2C圖，該遠端修改可包括將該失效鏈接項目206加入一遠端轉發表202，其規定該特定目的位址是傳送至該中斷鏈接(BL)(該中斷交換器背後之目的)虛擬埠。亦可修改該遠端平

衡表210使得傳送至該BL虛擬埠的項目之後於該遠端平衡表中匹配並於未失效之頂部交換器間分散。

根據一實施例，於步驟404，其完成有關該等多個葉片交換器120a-120n的哪一個附接於該失效鏈接的決定。該實施例之一範例中，該平衡表208可藉由修改與附接於該失效鏈接之該決定葉片交換器的該平衡表208中之失效鏈接相關聯的上鏈接輸出埠來修改，其中與該失效鏈接相關聯之上鏈接輸出埠可被修改來指出與與該失效鏈接相關聯之上鏈接輸出埠不同的一上鏈接輸出埠。該實施例之另一範例中，其建構不附接於該失效鏈接之一遠端葉片交換器的一轉發表202，而來自與附接於該失效鏈接之一遠端葉片交換器相關聯的一裝置之一封包的一封包目的位址可被決定，該平衡表208可藉由下列步驟來修改：修改來自與附接於該失效鏈接之葉片交換器相關聯的裝置之封包的上鏈接虛擬埠與上鏈接輸出埠，其中該修改並不安排封包從該裝置傳送路由至附接於該失效鏈接之該頂部交換器，以及將該遠端葉片交換器之轉發表202與具有來自與附接於該失效鏈接之葉片交換器相關聯的裝置之封包的已修改上鏈接虛擬埠以及已修改上鏈接輸出埠的平衡表208整合。

根據另一實施例，於步驟404，其完成有關該等多個頂部交換器的哪一個附接於該失效鏈接的決定。此實施例中，決定附接於該失效鏈接之至少一個頂部交換器停用，而因此停用附接於該決定之至少一個頂部交換器的鏈接。此外，可修改該平衡表208來避免使用該至少一個停用頂部

交換器之鏈接。此實施例中，其完成有關該停用頂部交換器是否為一跨距樹頂部交換器的決定。為用以響應該至少一個停用頂部交換器為一跨距樹頂部交換器，該胖樹網路中會選擇一新的跨距樹頂部交換器，而使用該新的跨距樹頂部交換器來運作一跨距樹協定以產生一新的跨距樹。

於上述實施例中，該等散列式安排傳送路由表(轉發表202、平衡表208、210)的其中之一或兩者可於步驟406中藉由修改該等散列式安排傳送路由表的其中之一或兩者中與該失效上鏈接相關聯的上鏈接輸出埠來修改。根據一範例，如上述參照第2B圖與第2C圖，與該失效上鏈接相關聯之上鏈接輸出埠可被修改來指出與與該失效上鏈接相關聯之上鏈接輸出埠不同的一上鏈接輸出埠。如其顯示與上述討論，如第2B圖所示，例如，該平衡表208中可修改該上鏈接虛擬埠與該上鏈接輸出埠，使得該封包不會從該裝置安排傳送路由至附接於該失效上鏈接之頂部交換器。另一範例中，於步驟406，如第2C圖所示，所有該等遠端葉片交換器之該轉發表202與該平衡表210可被修改，使得來自與附接於該失效上鏈接之失效葉片交換器相關聯的裝置之封包的上鏈接虛擬埠與已修改上鏈接輸出埠可整合。

步驟408中，一或更多封包可例如，由該安排傳送路由模組308，根據該已修改散列式安排傳送路由表來於該胖樹網路100中之該等多個頂部交換器110a-110n與該等多個葉片交換器120a-120n間安排傳送路由。更特別是，例如，如上述若該LAG包括該失效上鏈接，則一或更多封包可經由

如上述該LAG中未失效之上鏈接於該頂部交換器與不同的葉片交換器之間安排傳送路由。

圖形中提出之某些或所有操作可包含在任何所欲之電腦可讀儲存媒體中之一公用程式、程式、或子程式。此外，該等操作可由以作用中與不作用中之各種不同型式存在的電腦程式來具體化。例如，其可以由原始碼、物件碼、可執行碼或其他型式的程式指令組成之軟體程式存在。上述物件之任一個可於包括儲存裝置之一電腦可讀儲存媒體中具體化。

作示範的電腦可讀儲存裝置包括習知電腦系統隨機存取記憶體(RAM)、唯讀記憶體(ROM)、可抹除可程式化唯讀記憶體(EPROM)、電子可抹除可程式化唯讀記憶體(EEPROM)、以及磁碟或光碟或磁帶。上述具體範例包括一光碟唯讀記憶體(CD ROM)中之程式分佈或經由網際網路下載。因此應了解能夠執行上述功能之任何電子裝置皆可執行以上列舉之功能。

第5圖繪示一根據一實施例，組配來實施或執行第4圖描繪之該等程序的其中之一或更多程序的一計算裝置500之方塊圖。應了解該計算裝置500之圖形為一歸納圖，並且在不違背該計算裝置500之範疇的情況下，該計算裝置500可包括額外的組件以及上述某些組件可被移除及/或修改。

該計算裝置500包括一處理器502，其實施或執行第4圖描繪之該等程序的其中之一或更多程序中說明的某些或所有步驟。例如，該處理器502組配來檢測至少一頂部交換器

與至少一葉片交換器間之一失效鏈接，以根據該檢測之失效鏈接來修改該(等)散列式安排傳送路由表，以及根據該(等)散列式安排傳送路由表來安排該胖樹網路中之該等多個頂部交換器與該等多個葉片交換器間之封包傳送路由。來自該處理器502之命令與資料可於一通訊匯流排504上通訊。該計算裝置500亦包括一主要記憶體506、諸如該處理器502之程式碼於運行時間期間執行之一隨機存取記憶體(RAM)，以及一次要記憶體508。該次要記憶體508包括，例如，一或更多硬碟機510及/或一可移動式儲存器512、表示一軟碟機、一磁帶機、一光碟機、等等，其中可儲存第4圖描繪之該等程序的其中之一或更多程序的程式碼複本。

該可移動式儲存器512以一著名方式從一可移動式儲存單元514讀取及/或寫入一可移動式儲存單元514。使用者輸入與輸出裝置可包括一鍵盤516、一滑鼠518、以及一顯示器520。一顯示配接器522可與該通訊匯流排504及該顯示器520介接，並可從該處理器502接收顯示資料並將該顯示資料轉換為該顯示器520之顯示命令。此外，該(等)處理器502可透過一網路配接器524而於一網路，例如，網際網路、區域網路(LAN)、等等上通訊。

對業界熟於此技者而言，很明顯地該計算裝置500可加入或替代其他著名的電子組件。亦應很明顯地第5圖描繪之該等組件的其中之一或更多組件可任選(例如，使用者輸入裝置、次要記憶體、等等)。

雖然已特別說明直接揭露內容之全面性，但本發明之

代表實施例在大範圍的應用上亦相當實用，而上述說明並不意欲且不應詮釋為限制，但可提供來作為本發明觀點之舉例說明。

本文已說明及繪示的範例為本發明之實施例以及其某些變化型態。本文使用之用語、說明與圖形可僅藉由舉例解說來提出並且不表示為限制。業界熟於此技者可體認在本發明之精神與範疇中可有許多變化型態，其中本發明意欲由下列申請專利範圍及其等效項目來加以定義，除非於文中其他地方指出，否則所有用語皆以其最廣大適合的觀點來表示。

【圖式簡單說明】

第1圖顯示一根據本發明之一示範實施例的一胖樹網路圖；

第2A圖顯示一根據本發明之一示範實施例，用於一胖樹網路中之散列式路由器；

第2B圖顯示一根據本發明之一示範實施例，用於容錯一故障葉片交換器中之一失效鏈接的已修改散列式路由器；

第2C圖顯示一根據本發明之一示範實施例，用於容錯一故障葉片交換器中之一失效鏈接的已修改散列式路由器；

第3圖繪示一根據本發明之一示範實施例，第1圖中描繪之該控制器的簡化方塊圖；

第4圖繪示一根據本發明之一示範實施例，用於安排一

胖樹網路中之封包傳送路由的方法之流程圖；

第5圖顯示一根據本發明之一示範實施例，可用來作為一平台以實施或執行第4圖中描繪之該等一或更多程序的一計算裝置之方塊圖。

【主要元件符號說明】

100...胖樹網路	400...方法
110a-110n...頂部交換器	402、404、406、408...步驟
120a-120n...葉片交換器	500...計算裝置
130...控制器	502...處理器
200、220...散列式路由器	504...通訊匯流排
202...第二層轉發表	506...主要記憶體
204、208、210...平衡表	508...次要記憶體
206...中斷鏈接表	510...硬碟機
300...方塊圖	512...可移動式儲存器
302...上鏈接監控模組	514...可移動式儲存單元
304...交換器監控模組	516...鍵盤
306...散列式安排傳送路由表	518...滑鼠
更新模組	520...顯示器
308...安排傳送路由模組	522...顯示配接器
320...資料儲存器	524...網路配接器

七、申請專利範圍：

1. 一種用於一胖樹網路中使用一平衡表來安排多個頂部交換器與多個葉片交換器間之封包傳送路由的方法，該方法包含由一處理器執行之下列步驟：

檢測至少一頂部交換器與至少一葉片交換器間之一失效鏈接；

根據該檢測之失效鏈接來修改該平衡表，其中該平衡表包括根據該等封包各者之散列值來指出一上鏈接輸出埠的一上鏈接虛擬埠、以及與該等多個頂部交換器其中之一相關聯的用以安排該等封包傳送路由之一封包目的位址，其中該等封包各者之散列值係使用該等封包各者之一封包標頭來計算；以及

根據該已修改平衡表來安排該胖樹網路中之該等多個頂部交換器與該等多個葉片交換器間之封包傳送路由。

2. 如申請專利範圍第1項之方法，更包含下列步驟：

決定該等多個葉片交換器之哪個附接於該失效鏈接；而

其中修改該平衡表更包含修改附接於該失效鏈接之該決定葉片交換器的平衡表中與該失效鏈接相關聯的上鏈接輸出埠，其中與該失效鏈接相關聯之該上鏈接輸出埠被修改來指出與關聯於該失效鏈接之該上鏈接輸出埠不同的一上鏈接輸出埠。

3. 如申請專利範圍第1項之方法，更包含下列步驟：

決定該等多個葉片交換器之哪個附接於該失效鏈接；

建構一遠端葉片交換器之一平衡表，其中該遠端葉片交換器不附接於該失效鏈接；

決定來自與附接於該失效上鏈接之葉片交換器相關聯的一裝置之封包的一封包目的位址；而

其中修改該平衡表更包含：修改針對來自與附接於該失效鏈接之葉片交換器相關聯的裝置之封包的上鏈接虛擬埠與上鏈接輸出埠，其中該修改不會將來自該裝置的封包安排傳送路由至附接於該失效上鏈接之頂部交換器，以及，整合該遠端葉片交換器之平衡表與該具有針對來自與附接於該失效鏈接之葉片交換器相關聯的裝置之封包已修改之上鏈接虛擬埠與已修改之上鏈接輸出埠的平衡表。

4. 如申請專利範圍第3項之方法，更包含下列步驟：

修改該遠端葉片交換器之一轉發表以包括規定目的位址是要傳送至一中斷鏈接虛擬埠之失效鏈接項目。

5. 如申請專利範圍第1、2、3、4項之任一項的方法，更包含下列步驟：

根據該等封包各者之散列值來將該等封包各者指派至該等多個上鏈接之每一個。

6. 如申請專利範圍第1、2、3、4項之任一項的方法，更包含下列步驟：

決定該等多個頂部交換器之哪個附接於該失效鏈

接；以及

停用該決定之附接於該失效鏈接之至少一個頂部交換器，以藉此停用附接於該決定之至少一個頂部交換器的鏈接；以及

修改該平衡表以避免使用該至少一個停用頂部交換器之鏈接。

7. 如申請專利範圍第6項之方法，更包含下列步驟：

決定該停用之至少一個頂部交換器是否為一跨距樹頂部交換器；

響應於該停用之至少一個頂部交換器是一跨距樹頂部交換器而選擇一新的頂部交換器來操作以作為該跨距樹頂部交換器；以及

使用該新的跨距樹頂部交換器來執行一跨距樹協定以產生一新的跨距樹。

8. 如申請專利範圍第1、2、3、4項之任一項的方法，更包含下列步驟：

安排該等封包之傳送路由對稱地通過該胖樹網路。

9. 一種用於一胖樹網路中使用一平衡表來安排多個頂部交換器與多個葉片交換器間之封包傳送路由的電腦執行控制器，該電腦執行控制器包含：

一或更多模組，其組配來檢測至少一頂部交換器與至少一葉片交換器間之一失效鏈接；根據該檢測之失效鏈接來修改該平衡表，其中該平衡表包括根據該等封包各者之散列值來指出一上鏈接輸出埠的一上鏈接虛擬

埠、以及與該等多個頂部交換器其中之一相關聯的用於安排該等封包傳送路由之一封包目的位址，其中該等封包各者之散列值係使用該等封包各者之一封包標頭來計算；以及根據該已修改平衡表來安排該胖樹網路中之該等多個頂部交換器與該等多個葉片交換器間之該等封包的傳送路由。

10. 如申請專利範圍第9項之電腦執行控制器，其中該等一或更多模組更組配來決定該等多個葉片交換器之哪個附接於該失效鏈接，並修改附接於該失效上鏈接之一葉片交換器的平衡表中與該失效上鏈接相關聯的上鏈接輸出埠，其中與該失效上鏈接相關聯之該上鏈接輸出埠被修改來指出與關聯於失效上鏈接之該上鏈接輸出埠不同的一上鏈接輸出埠。
11. 如申請專利範圍第9項之電腦執行控制器，其中該等一或更多模組組配來決定該等多個葉片交換器之哪個附接於該失效鏈接；建構不附接於該失效鏈接之一遠端葉片交換器的一平衡表；決定來自與附接於該失效上鏈接之葉片交換器相關聯的一裝置之封包的一封包目的位址；修改針對來自與附接於該平衡表中之該失效鏈接的葉片交換器相關聯之裝置的封包之上鏈接虛擬埠與上鏈接輸出埠，使得該封包不會從該裝置安排傳送路由至附接於該失效上鏈接之頂部交換器；以及，整合該遠端葉片交換器之平衡表與該具有針對來自與附接於該失效鏈接之葉片交換器相關聯的裝置之封包已修改之上

鏈接虛擬埠與已修改之上鏈接輸出埠的平衡表。

12. 如申請專利範圍第9項之電腦執行控制器，其中該等一或更多模組更組配來修改該遠端葉片交換器之一轉發表以包括規定目的位址是要傳送至一中斷鏈接虛擬埠之失效鏈接項目。
13. 如申請專利範圍第9、10、11、12項之任一項的電腦執行控制器，其中該等一或更多模組更組配來決定該等多個頂部交換器之哪個附接於該失效鏈接、停用該決定之附接於該失效鏈接之頂部交換器並因此停用附接於該決定之頂部交換器的鏈接、以及修改該平衡表以避免使用該停用頂部交換器之鏈接。
14. 如申請專利範圍第9、10、11、12項之任一項的電腦執行控制器，其中該等一或更多模組更組配來決定該停用之頂部交換器是否為一跨距樹頂部交換器、響應於該停用之頂部交換器是一跨距樹頂部交換器而選擇一新的頂部交換器來操作以作為該跨距樹頂部交換器、以及使用該新的跨距樹頂部交換器來執行一跨距樹協定以產生一新的跨距樹。
15. 一種內嵌一或更多電腦程式之電腦可讀儲存媒體，該等一或更多電腦程式執行一種用於一胖樹網路中使用一平衡表來安排多個頂部交換器與多個葉片交換器間之封包傳送路由的方法，該等一或更多電腦程式包含一組指令以執行下列步驟：

檢測至少一頂部交換器與至少一葉片交換器間之

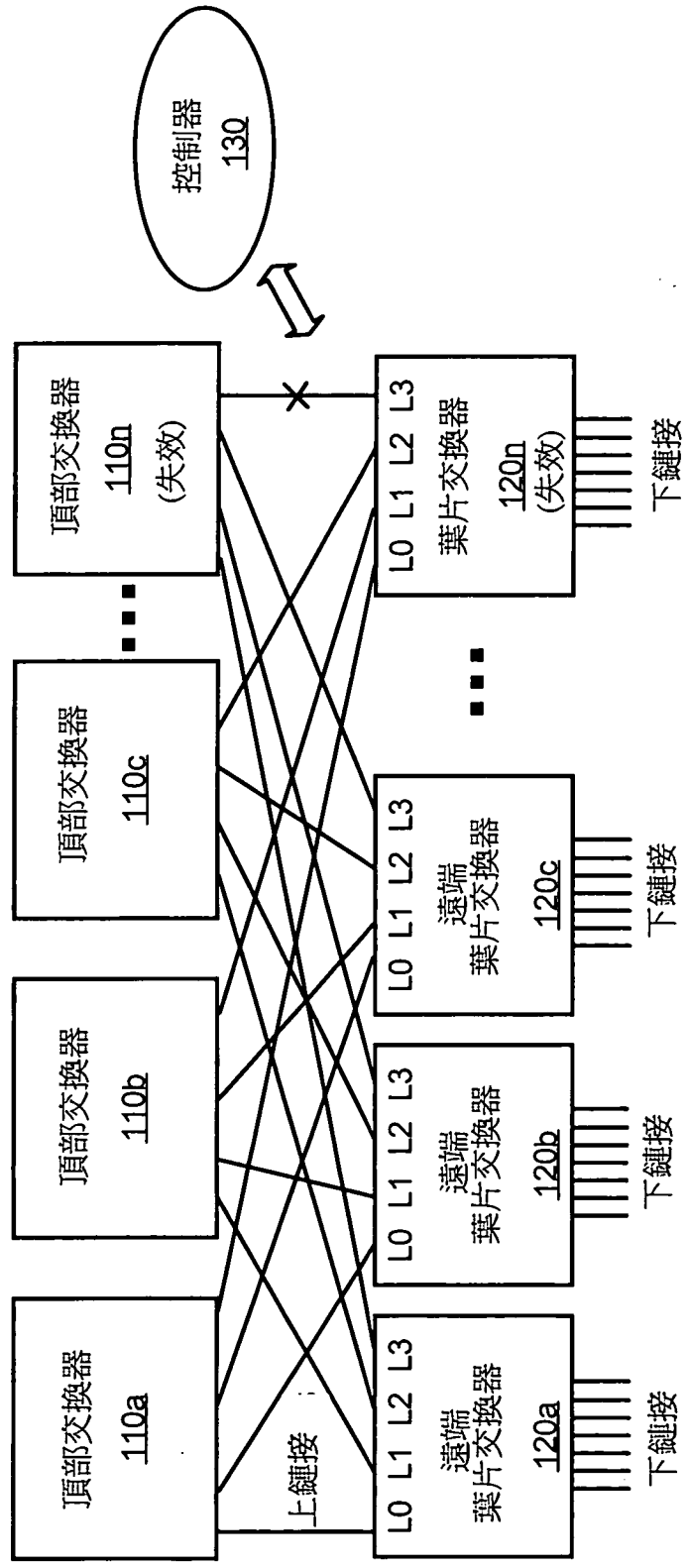
一 失效鏈接；

根據該檢測之失效鏈接來修改該平衡表，其中該平衡表包括根據該等封包各者之散列值來指出一上鏈接輸出埠的一上鏈接虛擬埠、以及與該等多個頂部交換器其中之一相關聯的用以安排該等封包傳送路由之一封包目的位址，其中該等封包各者之散列值係使用該等封包各者之一封包標頭來計算；以及

根據該已修改平衡表來安排該胖樹網路中之該等多個頂部交換器與該等多個葉片交換器間之該等封包的傳送路由。

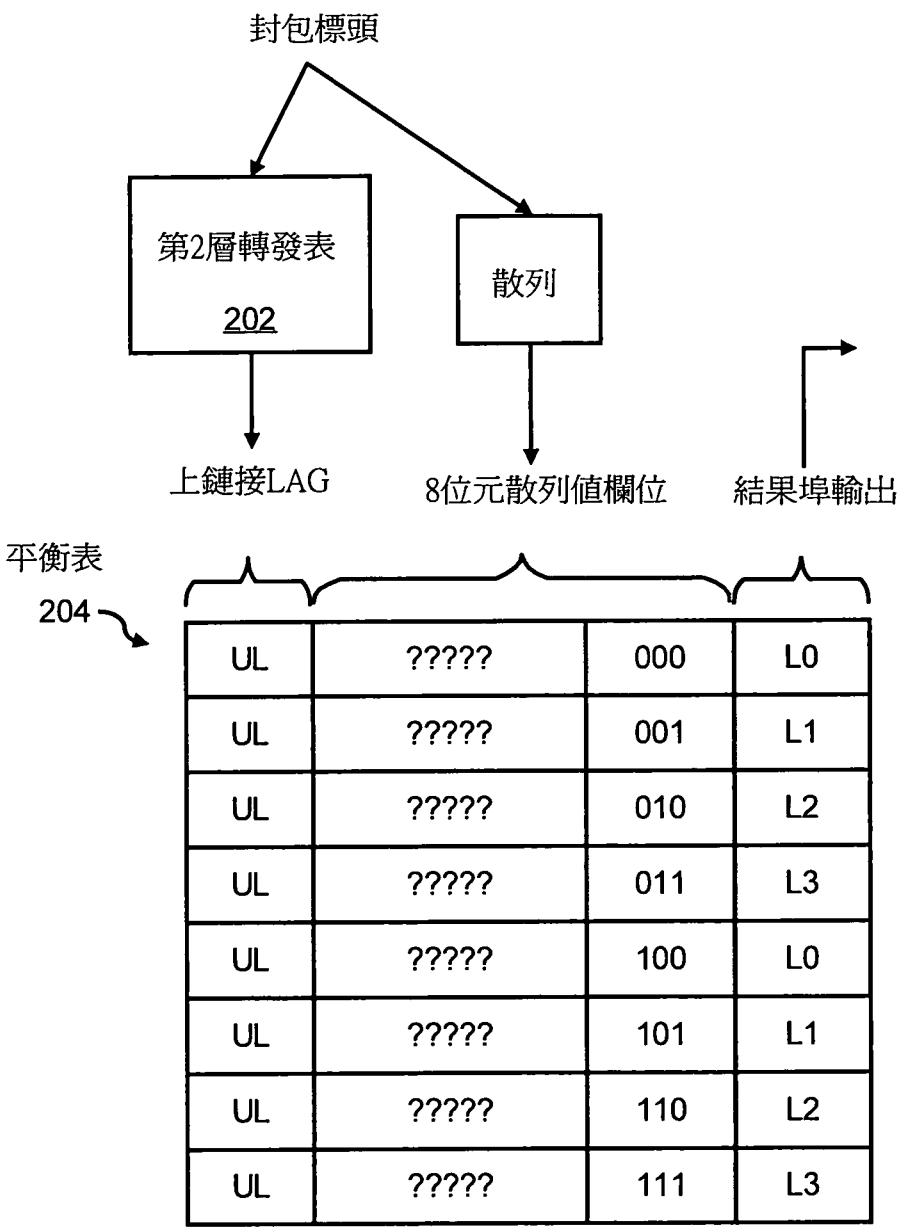
胖樹網路

100



第1圖

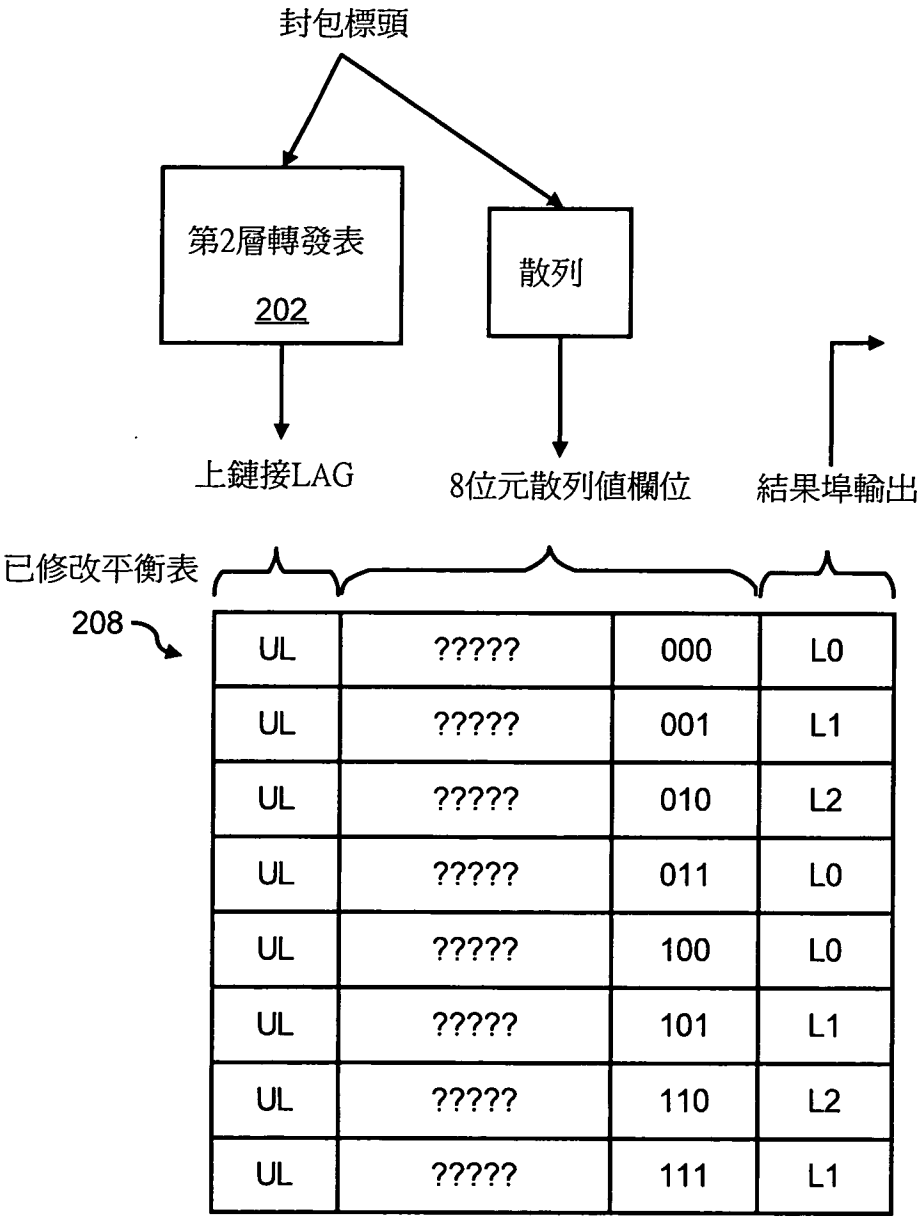
散列式路由器
200



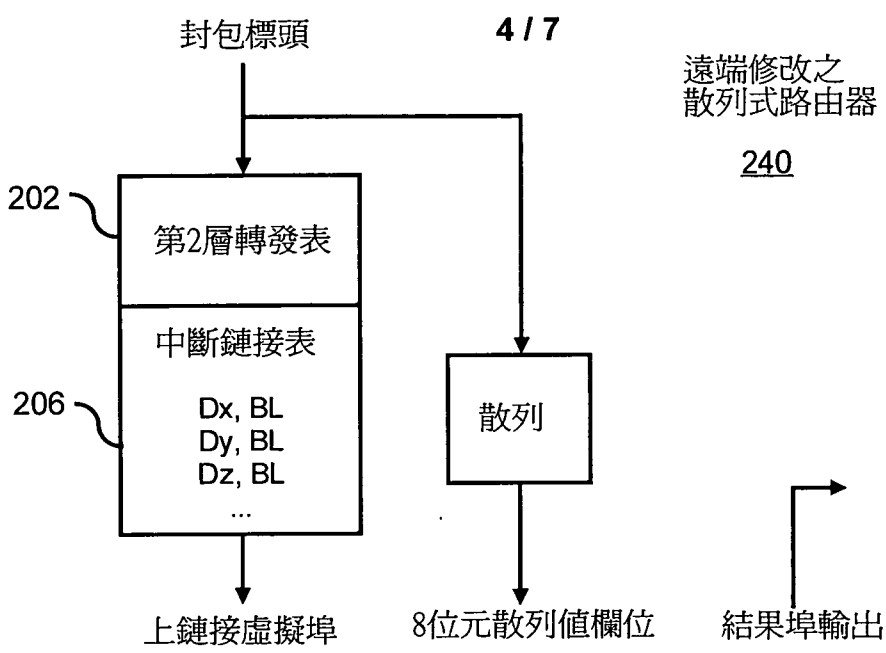
第2A圖

局部修改之散列式路由器

220



第2B圖

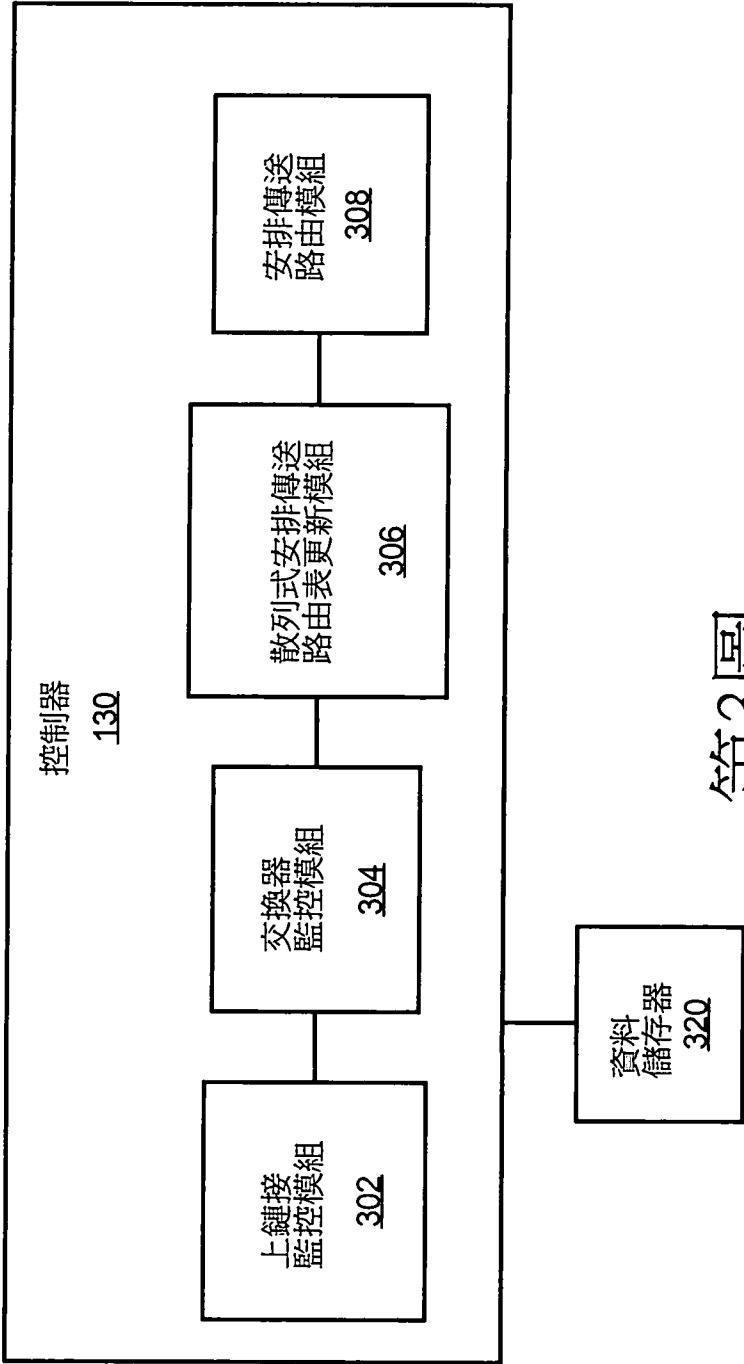


已修改平衡表 210

UL	?????	000	L0
UL	?????	001	L1
UL	?????	010	L2
UL	?????	011	L3
UL	?????	100	L0
UL	?????	101	L1
UL	?????	110	L2
UL	?????	111	L3
BL	?????	000	L0
BL	?????	001	L1
BL	?????	010	L2
BL	?????	011	L0
BL	?????	100	L0
BL	?????	101	L1
BL	?????	110	L2
BL	?????	111	L1

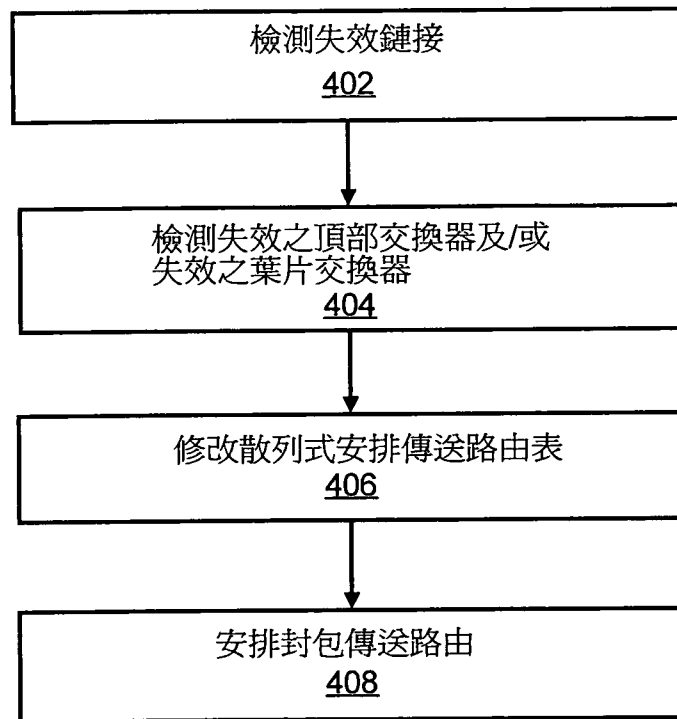
第2C圖

300

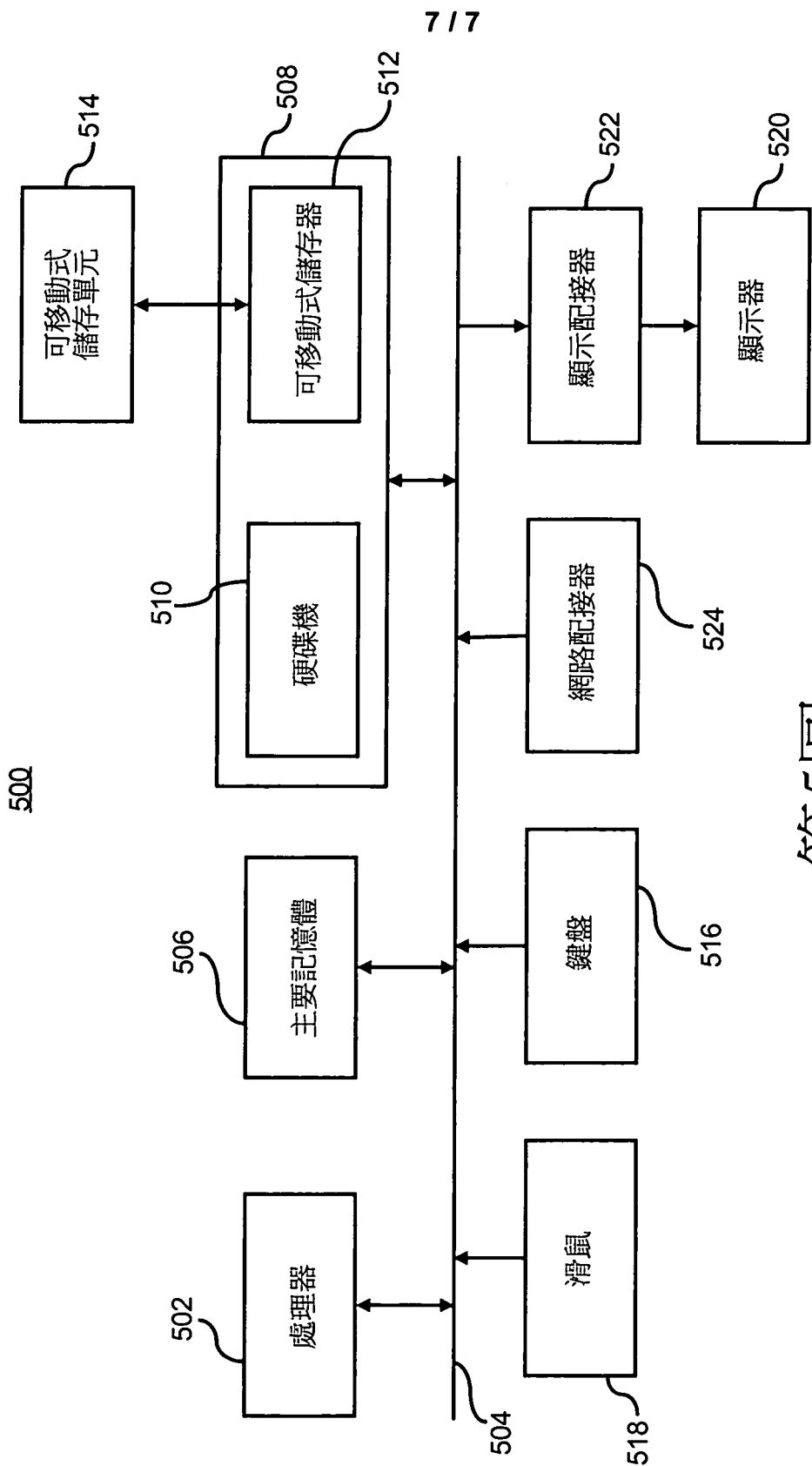


第3圖

400



第4圖



第5圖