

公告本

申請日期	90.8.21
案 號	90170537
類 別	G06F15/16

A4
C4

542964

(以上各欄由本局填註)

發 明 專 利 說 明 書		
一、發明 名稱	中 文	用於中心化窺探過濾器的方法和裝置
	英 文	METHOD AND APPARATUS FOR CENTRALIZED SNOOP FILTERING
二、發明 創作人	姓 名	1. 曼諾 凱爾 MANOJ KHARE 2. 法葉 A. 布利格 FAYE A. BRIGGS 3. 鄭凱 KAI CHENG 4. 莉莉 包 羅伊 LILY PAO LOOI
	國 籍	1. 印度 2. 美國 3. 美國 4. 美國
三、申請人	住、居所	1. 美國加州薩拉多加市奧瑞拉廣場12650號 2. 美國俄勒岡州波特蘭市西北法拉奇廣場1230號 3. 美國俄勒岡州波特蘭市西北特拉肯納路15885號 4. 美國俄勒岡州波特蘭市西北密查爾街15134號
	姓 名 (名 稱)	美商英特爾公司 INTEL CORPORATION
	國 籍	美國
	住、居所 (事務所)	美國加州聖塔卡拉瓦市米遜大學路2200號
	代 表 人 姓 名	F. 湯姆士·當烈二世 F. THOMAS DUNLAP, JR.

A6
B6

國(地區) 申請專利, 申請日期:

案號：

， ☐ 有 ☐ 無主張優先權

美國

2000年08月21日

09/643,382

☒ 有 ☐ 無 主張優先權

有關微生物已寄存於：

寄存日期：

，寄存號碼：

装

訂

線

五、發明說明(1)

發明領域

該發明關於電腦系統領域。尤其，該發明關於執行多節點電腦系統的窺探操作。

發明背景

快取器的一致性(coherency)是任何電腦系統的關鍵性要素。在網路伺服器電腦系統，維持快取器一致性是困難的工作是由於一個典型網路伺服器包含在多快取器外工作的多處理器。一些電腦系統建立於節似的架構其中一些處理器，它們的快取器記憶體，和系統記憶體區塊構成一節點。電腦系統可以擁有多節點。

在一個典型的先前系統，為了去維護快取器一致性，窺探過濾器包含在每一節點中。這個結構可以被解釋為分散式窺探架構。每個節點中的該窺探過濾器追蹤(所有同時快速緩衝儲存於任何節點處理器快取器的快取器列)的快取器一致性協定狀態資訊。一個較流行的快取器一致性協定是MESI(修改的、獨佔的、共享的或無效的)協定。

在分散式的窺探架構下，當一個節點希望去從記憶體去讀取一系列的資料，首先必須清楚是否有任何系統中的快取器具有那列資料的備份。該列資料的許多備份存於一些節點的快取器中是可能的並且備份可能快速緩衝儲存於任何狀態，只要MESI協定許可。為了決定哪一節點有該資料列的備份以及是在什麼狀態該列被快速緩衝儲存，節點希望讀取該資料列必須詢問其它所有點。因此，所有節點必須被詢問縱使其中一些節點可能沒有詢問(in question)快速緩

五、發明說明(2)

衝儲存該資料列。這個結果浪費排線頻寬並創造出非期望的系統額外負擔和增加延遲。

另一個先前的系統來維護快取器一致性引入了窺探目錄(directory)。在這個架構下，每個有系統記憶體的節點包含一個窺探目錄。該窺探目錄追蹤所有快取器動作資訊包含座落於那個節點的記憶體位置(involving memory locations located at that node)。每個窺探目錄從窺探目錄節點系統記憶體追蹤哪個節點有快速緩衝儲存該資料列並且追蹤被快速緩衝儲存該列在什麼狀態。如此，有了這個分散式目錄架構，當一個節點(提出請求的節點)希望從其他節點(終點節點)上讀取記憶體位置，提出請求的節點必須與查詢它的目錄的(who consults its directory)終點節點聯繫。接著終點節點送出窺探請求到任一節點，其中該任意節點有詢問(in question)快速緩衝儲存記憶列(memory line)的備份。然後各個不同的節點回應給終點節點而終點節點因此可以提供要求的資料給提出要求的節點。第一步驟聯繫終點節點然後終點節點聯繫其它節點來窺探資訊會導致非期待的系統額外負擔，尤其是在電腦系統中，其中該電腦系統的節點藉由交換裝置連結在一起。

圖式簡述

本發明將從下列具體實施例中詳細的說明和伴隨的圖式更加了解，然而這些說明和圖式不應該被拿來限制發明於所敘述的特定具體實施例，而只是拿來說明和了解而已。

圖1是一個系統的具體實施例的方塊圖，該系統係包含幾

五、發明說明(3)

個藉由一個交換裝置連結的節點，其中該交換裝置包含一個中央窺探過濾器。

圖2是一個系統的具體實施例之方塊圖，該系統係包含幾個藉由兩個交換裝置連結的節點，其中該兩個交換裝置每個包含一個中央窺探過濾器。

圖3是一個方法的具體實施例之流程圖，該方法是為了使用在交換裝置中的中央窺探過濾器而執行一個窺探週期。

圖4是一個方法的具體實施例之流程圖，該方法是為了執行一讀取處理，該讀取處理包含一個使用在交換裝置中的中央窺探過濾器之窺探週期。

詳細說明

一個使用中央窺探過濾器之電腦系統的具體實施例，該中央窺探過濾器包含數個經由交換裝置連結在一起的節點。每個節點可能包含數個處理器和快取器和系統記憶體。從一個節點到另一節點的所有流量皆透過交換裝置發生。該交換裝置包含一個窺探過濾器，該窺探過濾器針對所有在電腦系統中的快取器追蹤快取器列的連貫性資訊，該窺探過濾器有至少足夠的入口(entries)來追蹤所有系統節點中所有快取器的所有入口的標記位址(tag address)和狀態資訊。除了標記位址和狀態資訊外，該窺探過濾器儲存資訊指出哪些節點具有每個快取器列的備份。該窺探過濾器部分適用(serves in part to)於避免窺探動作(transactions)執行於沒有包含對象主體快取器列的一個備份之節點中藉此減少系統額外負擔處理，減少跨系統互連匯流排的流量，

五、發明說明(4)

和減少執行窺探作業所需的時間。

一個中央窺探器的具體實施例使用一種不同的MESI協定。在這個具體實施例中的窺探過濾器追蹤快取器列是否是無效的、共享的或修改/獨佔的。該完整的MESI協定支援在不同節點上的快取器中。其它協定可以使用，只要在所有快取器代理人程式(agents)上的動作(transactions)是一致的。

因為中央窺探過濾器包含所有在快取器代理人程式上的快取器，所以快取器代理人程式不應該具有沒有出現在窺探過濾器中之快取器列的備份。如果一列被從窺探過濾器中除去，該列也必須從包含該列備份的所有節點之快取器代理人程式中除去。

該窺探過濾器利用存在向量(vector)追蹤那些節點，具有著一個列(line)的備份。該存在向量是一位元欄位，其中每個位元對應到一個特定節點。

不同的具體實施例可能在每一個節點上不是使用向量組(vector field)而是使用不同的窺探過濾器。窺探過濾器位於交換裝置上並且因此集中放置。可以實施該窺探過濾器或多個窺探過濾器使得特定窺探過濾器入口適用於一個節點，其它入口適用於其它節點，如此推演下去。

圖1是一個系統的區塊圖，該系統包含節點A、節點B和節點C。節點A可以包含多個標為105到L的處理器。節點B可以包含多個標為115到M之處理器。節點C可以包含多個標為125到N之處理器。節點A也可以包含一節點控制電路

五、發明說明(5)

102和一個系統記憶體區塊104。節點B也包含節點控制電路112和一個系統記憶體區塊114。另外節點C包含一節點控制電路122和一個系統記憶體區塊124。節點A、B和C的處理器包含快取器。

節點A、B和C每個都連結到一交換裝置130上。該交換裝置130包含一個中央窺探過濾器132。一輸入/輸出集線器140亦連接至該交換裝置130。輸入/輸出集線器140也可以包含一個快取器。該窺探過濾器132包含足夠入口來追蹤目前所有系統節點的所有快取器的所有入口的標記、狀態和資訊，該等系統節點包含該處理器快取器和該輸入/輸出集線器140快取器。

現在討論一個向沒有快速緩衝儲存於任何系統快取器之中的記憶列執行資料讀取動作的例子。對這個例子而言，節點A是提出請求的節點(該結點發送讀取動作)，節點C是終點節點(該節點包含讀取請求所定址(addressed)的系統記憶體部分)。節點A發出一個讀取請求到交換裝置130。該窺探過濾器132接著執行一個查詢(lookup)操作來決定被讀取請求定址的記憶列是否已經被快速緩衝儲存於任何節點A、B或C，或在輸入/輸出集線器140中。當該窺探過濾器132正在執行查詢(lookup)操作時，該交換裝置130發送出一個推測的(speculative)記憶體讀取到節點C。該交換裝置130包含記憶體解碼電路，其中該記憶體解碼電路使得交換裝置130可以決定哪一個節點正被該讀取請求所定址。對這個例子而言，該窺探過濾器查詢(lookup)指出由該讀取請求

五、發明說明(6)

所定址的記憶列沒有快速緩衝儲存於任何系統快取器中。該交換裝置130傳送一個窺探回應給節點A並且也送一個記憶體讀取確認訊息給節點C。回應記憶體讀取確認訊息，節點C傳送所請求的資料到交換裝置130。該交換裝置130然後傳送該請求的資料到節點A。因為節點A是唯一快速緩衝儲存主(subject)快取器列的節點，該窺探過濾器132將它的用於此快取器列之入口標記為獨佔(exclusive)並且設定存在向量以指出節點A有著該快取器列的一個備份。

現在將討論一個寫入到由兩個節點共享之快取器列的記憶體寫入動作之例子，對這個例子而言，節點A是提出請求的節點並且也是終點節點。換句話說，節點A正在寫入一系列資料到它自己的記憶體。節點A藉由發送出一個無效的周期到該交換裝置130而開始。該窺探過濾器132接著執行一個查詢操作來決定那些節點有著快取器列的備份並且也決定在何種狀態快取器列被快速緩衝儲存。該查詢操作在這個例子中指出快取器列是“共享的”並且節點B和C都有備份。該交換裝置130接著發送出多個無效的循環到節點B和C。節點B和C以窺探回應訊號來回應，接著該交換裝置130發送一個窺探回應給節點A來完成整個動作。該窺探過濾器132現在顯示，節點A有著該快取器列的唯一備份並且該列是在“獨佔／修改”的狀態。

現在將討論一個針對一個“修改的”快取器列之讀取週期(cycle)的例子，對這個例子而言，節點A是發出請求的

五、發明說明 (7)

節點，節點 B 是已修改過所關心的快取器列之備份的節點，並且節點 C 是終點節點(該節點的系統記憶體區塊包含對應到快取器列的記憶體位置)。節點 A 藉由發送一個取請求到交換裝置 130 開始動作。當交換裝置 130 發出一個推測性(speculative)的記憶體讀取到節點 C 時，該窺探過濾器 132 執行一個查行操作。窺探過濾器查詢的結果指出節點 B 有一個快取器列修改過的備份。該交換裝置 130 送出一個窺探請求到節點 B，節點 B 以伴隨有該修改資料的窺探回應來回應。該交換裝置 130 接著傳送該窺探回應和該資料到節點 A，並且同時傳送該資料到節點 C，使得系統記憶體可以被更新。節點 C 接著發信一個完成信號到交換裝置 130，並且該完成信號從該交換裝置 130 轉送到節點 A 來完成該動作(transaction)。

有可能在一個系統中使用多個窺探過濾器，以提供充分的窺探過濾器處理效能(throughput)和能力以符合系統可擴展性(scalability)需求。在這種情況下，不同的窺探過濾器持續追蹤快取器列的互斥集合。所有時候一個快取器列都被唯一的一個窺探過濾器所追蹤。

圖 2 是一個系統的方塊圖，該系統包含節點 D、節點 E 和節點 F，並且也包含兩個窺探過濾器 232 和 234。節點 D 可以包含多個標為 205 到 P 之處理器。節點 E 可以包含多個標為 215 到 Q 之處理器。節點 F 可以包含多個標為 225 到 R 之處理器。節點 D 也包含一個節點控制電路 202 和一個系統記憶體區塊 204。節點 E 也包含一個節點控制電路 212 和一個系統

五、發明說明(8)

記憶體區塊214。另外節點F包含一個節點控制電路222和一個系統記憶體區塊224。該節點D、E、和F的處理器包含快取器。

節點D、E和F每一個由交換裝置230所連結並且也連到交換裝置240。該交換裝置230包含該中央窺探過濾器232。一個輸入/輸出集線器250也和交換裝置240一樣連結到交換裝置230。該輸入/輸出集線器250也可以包含一個快取器。該交換裝置240包含該中央窺探過濾器242。一個輸入/輸出集線器260也連接到交換裝置240並且也連到交換裝置230。該輸入/輸出集線器260也可以包含一個快取器。

該窺探過濾器232包含足夠的入口來追蹤針對所有系統節點中的所有快取器的所有一半的入口之標籤、狀態和現在的資訊，該等系統節點包含該處理器快取器和輸入/輸出集線器快取器。該窺探過濾器242包含針對該系統快取器的其餘部分之入口。在一個具體實施例中，該節點D、E和F在一個交替(alternating)的快取器列基礎上，交替存取交換裝置230和240。

圖3是一個方法的具體實施例的流程圖。該方法係關於執行一個窺探週期(cycle)，其中該窺探週期係使用位於交換裝置中的中央窺探過濾器。在區塊310中，一個讀取請求從第一個節點發送出一個交換裝置。該交換裝置包含一個窺探過濾器。在方塊320中，執行一個窺探過濾器查詢來決定是否位於第二節點的一個快取器有著對應於該讀取請求的快取器列之備份。

五、發明說明(9)

圖4是一個方法的具體實施例的流程圖。該方法執行一個包含使用位於一個交換裝置中的中央窺探裝置的一個窺探週期之讀取動作。在方塊410中，一個讀取請求從第一個節點發送到一個交換裝置，該交換裝置包含一個窺探過濾器。該讀取請求定址(addressed)於位於第二節點中的記憶體位置。在方塊420中執行，一個窺探過濾器查詢，同時在方塊430中，一個推測的記憶體讀取提取從該交換裝置發送到第二個節點。方塊440指出如果該窺探過濾器查詢決定，由讀取請求所載明的記憶體列被快速緩衝儲存於一個或多個該節點中，然後一個快取器列演算法在方塊450中執行。這種形式的演算法的例子在上面被討論並與圖1有關。方塊440也指出如果該窺探過濾器查詢決定，由讀取請求所載明的該記憶體列不藉由任何快取器代理者快速緩衝儲存，則在方塊460中一個窺探回應從該交換裝置傳送到第一個節點。同時，在方塊470一個記憶體讀取確認訊息從該交換裝置傳送到第二節點。最後，在方塊480該提取的資料藉由交換裝置從第二節點傳送到第一節點。

在前面的說明中，本發明已經參考特定範例性的具體實施例說明。然而很明顯地，可生不背離如附加的申請專利範圍所述之本發明的較廣的精神和範圍，而對本發明進行不同的修改和改變。因此，本說明書和圖式可視為範例而非限制。

涉及到說明中的“一具體實施例”、“一個具體實施例”、“一些具體實施例”或“其它具體實施例”意思是，參考各具

五、發明說明 (10)

體實施例所述之一個特別的特徵、結構或特性是包含在本發明的至少一些具體實施例但未必是所有具體實施例。“一具體實施例”、“一個具體實施例”或“一些具體實施例”的出現未必指的是相同的具體實施例。

四、中文發明摘要(發明之名稱：用於中心化窺探過濾器的方法和裝置)

一個電腦系統使用中央窺探過濾器的具體實施例包含幾個藉由交換裝置連結在一起的節點。每個節點可能包含一些處理器和快取器以及一塊系統記憶體。所有從一個節點到另一節點的流量透過該交換裝置發生。該交換裝置包含一個窺探過濾器，該窺探過濾器追蹤所有電腦系統中快取器中的快取器列相干(coherency)資訊。該窺探過濾器有足夠的入口來追蹤所有系統節點快取器入口的標籤和狀態資訊。除了標籤和狀態資訊外，該窺探過濾器儲存指示哪一些節點有快取器列的備份的資訊。該窺探過濾器部分適用(serves in part to)於保持窺探動作(transactions)在沒有包含主(subject)快取器列的節點中被執行，因此減少系統額外負擔處理，減少跨系統互連匯流排的流量，以及減少執行窺探動作所需的時間。

英文發明摘要(發明之名稱：METHOD AND APPARATUS FOR
CENTRALIZED SNOOP FILTERING)

An example embodiment of a computer system utilizing a central snoop filter includes several nodes coupled together via a switching device. Each of the nodes may include several processors and caches as well as a block of system memory. All traffic from one node to another takes place through the switching device. The switching device includes a snoop filter that tracks cache line coherency information for all caches in the computer system. The snoop filter has enough entries to track the tags and state information for all entries in all caches in all of the system's nodes. In addition to the tag and state information, the snoop filter stores information indicating which of the nodes has a copy of each cache line. The snoop filter serves in part to keep snoop transactions from being performed at nodes that do not contain a copy of the subject cache line, thereby reducing system overhead, reducing traffic across the system interconnect busses, and reducing the amount of time required to perform snoop transactions.

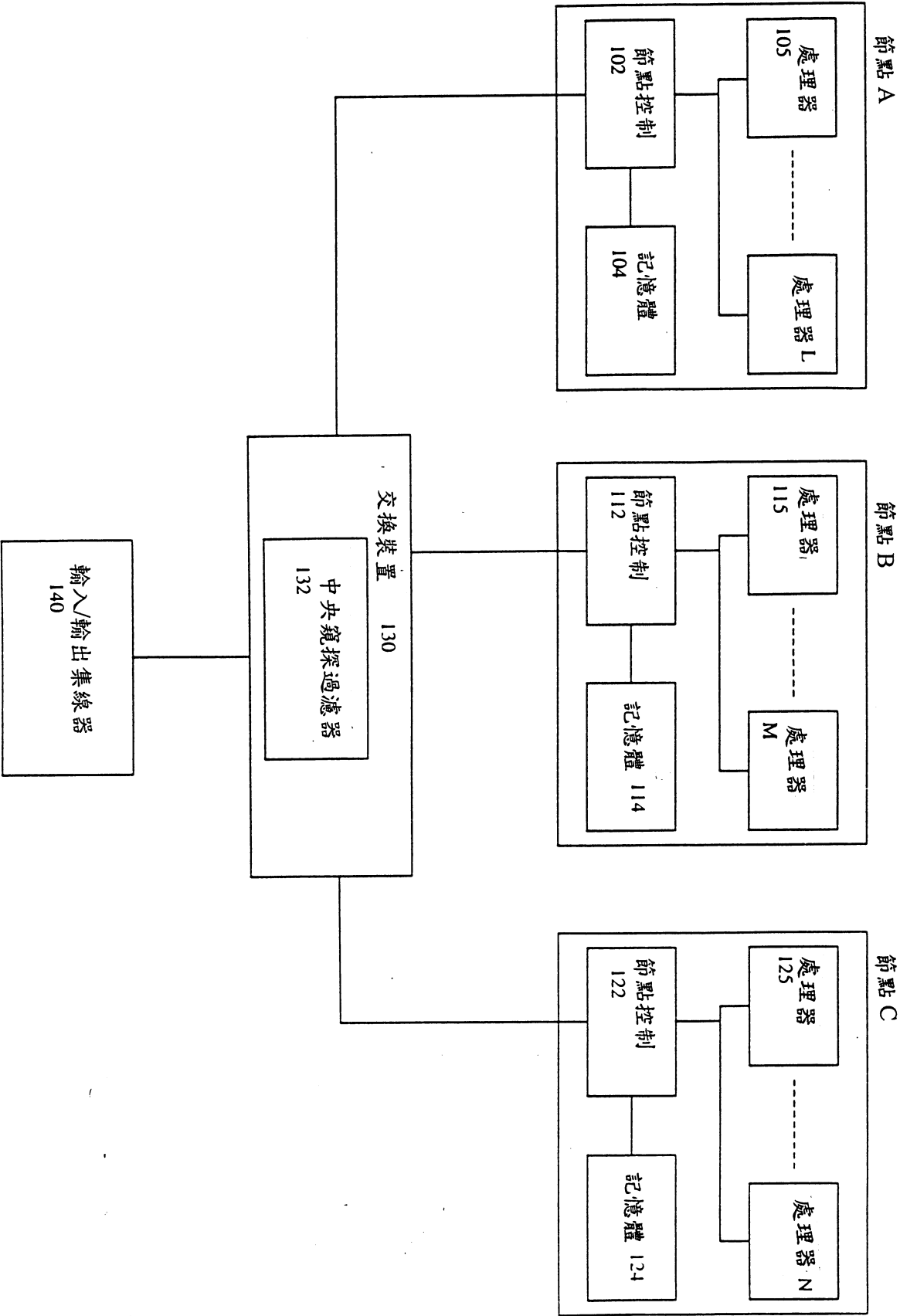


圖 1

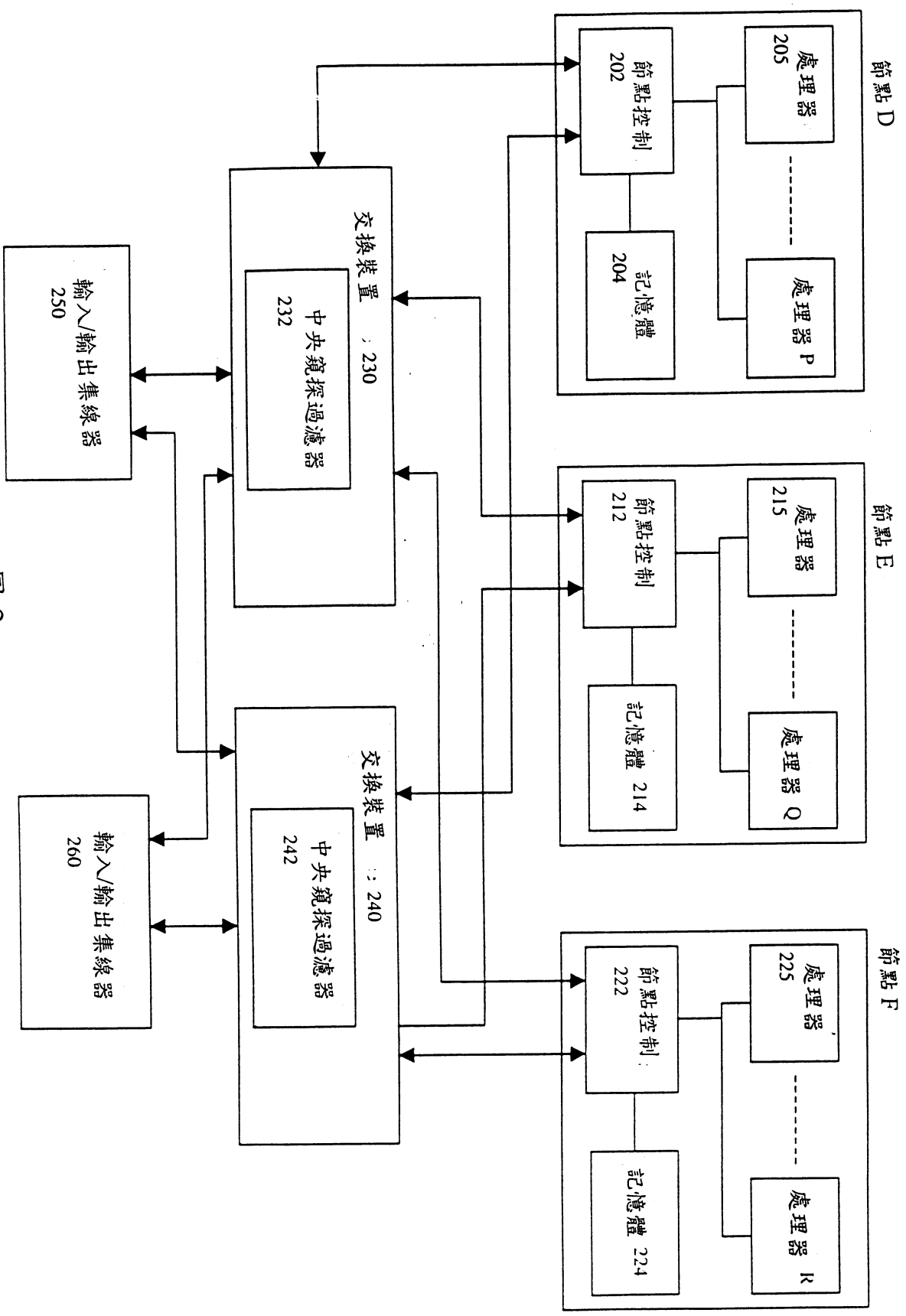


圖 2

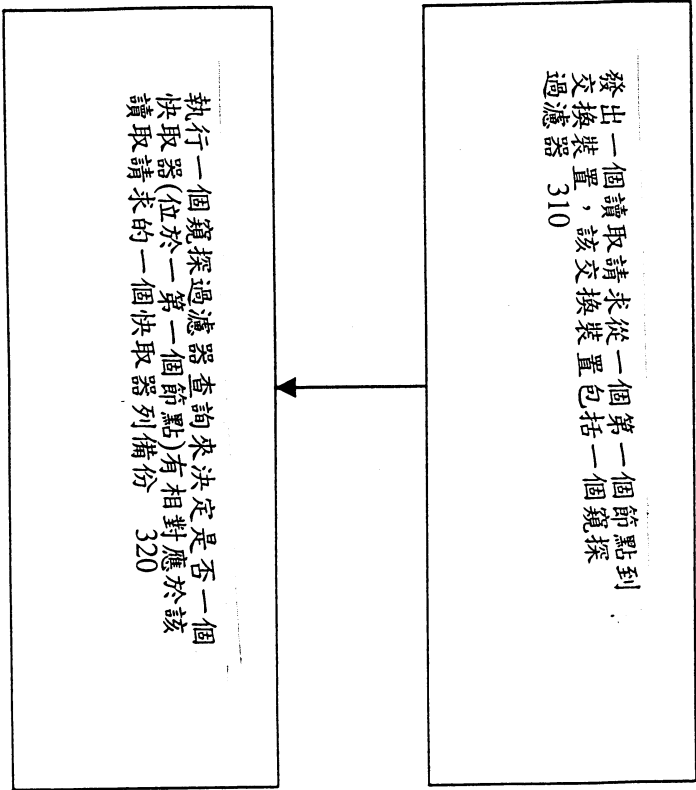


圖 3

發出一個讀取請求(從一個第一個節點到一個交換裝置包含一個窺探過濾器)其中該讀取請求地址於在第二個節點的記憶體位置 410

執行一個窺探過濾器查詢 420

發出一個預測性的記憶體讀取裝置 430
從第二個節點

列快速緩衝儲存 440
是
執行快速器列演算法 450

傳送一個窺探回應從第一個交換裝置 460
該節點

傳送一個記憶體讀取確認訊息從第二節點 470
裝置到該節點

傳送取得資料從該節點第二節透過該交換裝置 480

圖 4

六、申請專利範圍

1. 一種用於中心化窺探過濾器之裝置，包含：
複數個節點介面埠，提供複數個節點的通聯；和
一個窺探過濾器，用來維持該等多個節點的快取器一致性(coherency)資訊。
2. 如申請專利範圍第1項之裝置，其中該等複數個節點的每個節點係包含一個快取器，並且其中該窺探過濾器另外包含至少多個和位於多節點中所有快取器所包含的入口的數量相同之入口。
3. 如申請專利範圍第2項之裝置，其中每個窺探過濾器入口包含一個位址標記。
4. 如申請專利範圍第3項之裝置，其中每個窺探過濾器入口包含一個欄位(field)，指示一個相對應之快取器列的快取器連貫性協定狀態。
5. 如申請專利範圍第4項之裝置，其中每個窺探過濾器入口包含一個欄位(field)指出該多節點哪一個有相對應快取器列的備份在它的快取器中。
6. 如申請專利範圍第5項之裝置，該等多節點介面埠的一個接收來自該等多節點和窺探過濾器中的一個的窺探請求，決定(該多節點的其它哪一個有相對應於該窺探請求快取器列的備份)並且決定(哪一個快取器連貫性協定狀態被指定給該快取器列)。
7. 一種用於中心化窺探過濾器之系統，係包含：
一個交換裝置，其包括一個窺探過濾器；
一個第一個節點，其連到該交換裝置；和

92年5月21日修正
補充A8
B8
C8
D8

六、申請專利範圍

- 一個第二個節點，其連到該交換裝置。
8. 如申請專利範圍第7項之系統，其中該第一個節點係包括一個處理器和一個快取器。
9. 如申請專利範圍第8項之系統，其中該第二個節點係包括一個處理器、一個記憶裝置和一個快取器。
10. 如申請專利範圍第9項之系統，其中該窺探過濾器係包括和該第一個節點和該第二個節點中包括的入口至少相同數量的入口。
11. 如申請專利範圍第10項之系統，其中每個窺探過濾器入口係包括一個位址標記。
12. 如申請專利範圍第11項之系統，其中每個窺探過濾器入口係包括一個欄位，指出一個相對應的快取器列的一個快取器連貫性協定狀態。
13. 如申請專利範圍第12項之系統，其中每個窺探過濾器入口係包括一個欄位，指出該第一個節點和第二個節點哪一個有一個相對應快取器列的一個備份在它的快取器中。
14. 如申請專利範圍第13項之系統，該交換裝置接收一個來自於該第一個節點和該窺探過濾器的窺探請求，以決定該第一個節點和該第二個節點哪一個有相對應於該窺探請求之快取器列的備份並且決定哪一個快取器連貫性協定狀態被指定給該快取器列。
15. 一種用於中心化窺探過濾器之方法，係包含：
- 從一個第一個節點發出一個讀取請求到一交換裝置，

六、申請專利範圍

該交換裝置包括一個窺探過濾器；

執行一個窺探過濾器查詢，來決定是否一個位於一個第二個節點的快取器有一個相對應於該讀取請求快取器列的備份。

16.如申請專利範圍第15項之方法，另外包含緊接著該窺探過濾器查詢，從該交換裝置發出一個窺探回應到該第一個節點。

17.如申請專利範圍第16項之方法，另外包含藉由從一個位於該第二個節點之記憶體提取資料而執行該讀取請求，以回應該讀取請求，該資料提取由該交換裝置執行。

18.一種用於中心化窺探過濾器之方法，係包含：

從一個第一個節點發出一個讀取請求到一交換裝置，該交換裝置包括一個窺探過濾器，該第一個節點是該多節點中的一個，該讀取請求定址(addressed)至位於第二個節點的記憶體位置，該第二個節點也是該多節點中的一個，該多節點中每一個包括一個快取器；

執行一個窺探過濾器查詢而從該交換裝置發出一個推測的(speculative)記憶體讀取提取到該第二個節點；

從該交換裝置傳送一個窺探回應到該第一個節點，而同時，如果該窺探過濾查詢指出該無節點有一個相對應於該讀取請求所指明的記憶體位址的記憶體快取器列，一個記憶體讀取確認訊息被傳送到該第二個節點；

傳送資料，以回應從該第二個節點到該交換裝置的該記憶體讀取確認訊息，以及

六、申請專利範圍

裝置傳送該資料從該交換到該第一個節點。

19. 如申請專利範圍第18項之方法，另外包含在窺探過濾器內指出該傳送資料已被該第一個節點在一個獨佔狀態下快速緩衝儲存。