公告本

申請日期	88, 12, 23
案號	88122711
類別	G06F 15/163; G06F 12/08

（以上各欄由本局填註）

A4
C4

449701

發明 ~~新型~~ 專利說明書

一、發明 ~~新型~~ 名稱	中文	混合非單一記憶體架構/簡單唯快取記憶體架構之系統及方法
	英文	HYBRID NUMA/S-COMA SYSTEM AND METHOD
二、發明 ~~創作~~ 人	姓名	狄恩 A. 利伯堤
	國籍	美國
	住、居所	美國紐約州帕倫維爾市HCR路1號312號郵政信箱
三、申請人	姓名（名稱）	美商萬國商業機器公司
	國籍	美國
	住、居所（事務所）	美國紐約州阿蒙市新果園路
	代表人姓名	傑拉德 羅森賓

449701

（由本局填寫）

承辦人代碼：	
大　　類：	
ＩＰＣ分類：	

A6
B6

本案已向：

國（地區）	申請專利，申請日期：	案號：	，□有 □無主張優先權
美國	1999 年 01 月 27 日	09/238,203	☑有 □無主張優先權

有關微生物已寄存於：　　，寄存日期：　　，寄存號碼：

（請先閱讀背面之注意事項再填寫本頁各欄）

裝

訂

線

五、發明說明（1）

技術領域

本發明說明分散式共享記憶體系統和快取記憶體的範疇；且更詳盡地說，本發明係說明一種混合的配置，其中一第一種記憶體形態(簡單CMOA)建立在上面，並和另一種記憶體形態(NUMA)整合在一起。

定義

於本說明書中使用下面的名詞：

全域記憶體：意指藉不同節點上的處理而可定址的記憶體物件。以一種類似UNIX系統V的方法產生和接附，並接附至每一個欲定址該全域記憶體物件的處理的該有效位址空間中。

DSM：分散式共享記憶體。提供該共享記憶體功能的配置的一種等級(即使該實體記憶體分散在該系統的節點之間)。

S-COMA：簡單唯快取記憶體配置。一種DSM設計，其中每一個節點將其區域記憶體的一部分保留、作爲全域記憶體的一個快取記憶體之用。經由S-COMA軟體和硬體的一個組合管理該快取記憶體。處理藉處理特殊虛擬位址參考該資料，節點記憶體硬體藉區域實際位址參考該資料，且S-COMA硬體在節點之間傳遞全域位址。該S-COMA次系統管理區域實際位址和全域位址之間的翻譯。

NUMA：非單一記憶體存取。一種DSM設計，其中一個系統中該等n個節點的每一個節點保留該系統其1/n的實際記憶體(和實際位址空間)。處理藉該虛擬位址參考資料，

五、發明說明(2)

且節點記憶體硬體藉該實際位址參考資料。該NUMA基礎在節點之間傳遞實際位址。

UMA：單一記憶體存取。一種共享記憶體結構，由是任何處理器可在同樣的(單一的)時間參考任何記憶體位置。

邊界功能(BF)：一種層次功能或邏輯功能，在一個節點的邊界上執行一組行動。於本發明中，該BF對"藉該DSM次系統進入或離開一個節點"的位址執行位址翻譯。

客戶：一個參考(快取)資料的節點，但並非該資料的HOME。

本地(HOME)：一個節點，爲該資料的所有者或該目錄的所有者，管理該資料的相干性。

等待時間：該與一個特殊行動或操作相關聯的延遲，像是從記憶體中擷取資料。

窺察邏輯："監督(窺察)一條線或一個匯流排"的邏輯、尋找特殊位址、標籤或其它的關鍵資訊。

網路邏輯：界面至一個網路或通訊結構中的邏輯。

實際位址空間：由位址翻譯產生之該等實際位址的範圍。該實際記憶體的該等位址。

區域實際位址：引用至一個區域節點上的一個實際位址。

全域實際位址：引用至所有節點上的一個實際位址。

實體位址：一個實際位址。該實體記憶體的位址。

輸入位址：提供該位址作爲"輸入至一個元件中"。

關聯位址：在一個由位址對組成的資料結構中，該位址

五、發明說明(3)

對的該第二個位址，該第一個位址爲該輸入位址。

發明背景

共享記憶體多處理器系統允許多個處理器的每一個處理器藉讀取和寫入(載入和儲存)操作參考該系統中的任何儲存體位置。該等處理器或程式無法知曉該共享記憶體的基本結構，目前只能顧及該共享記憶體的性能。

可由多個處理器更新一個單一的記憶體位置。該結果爲一個單一的更新序列，且所有的處理器以該相同的順序察看該記憶體位置的該等更新。該特性以“相干性”著稱聞名。在一個相干的系統上，沒有任何一個處理器可察看一個不同的更新順序(除了另一個處理器之外)。

快取記憶體相干、共享記憶體多處理器系統提供快取記憶體給該記憶體結構，以改良記憶體存取的性能(減少等待時間)。因將該等快取記憶體保持相干，故就一個指定的記憶體位置而言，可維持其一個單一更新序列的特徵(爲該系統中所有處理器察看的)。

於該專利中所討論的該等系統配置均爲快取記憶體相干、共享記憶體多處理器系統。將於下說明該等系統三種特殊的變化，即UMA、NUMA和S-COMA。

“UMA”意指單一記憶體存取，且說明一種系統配置，其中一個電腦系統中的多個處理器共享一個實際位址空間，且任何處理器到任何記憶體位置的該記憶體等待時間係相同的或爲單一的。即一個指定的處理器可在單一的時間參考任何記憶體位置。現代大部分對稱性的多處理器(SMP)

五、發明說明（ 4 ）

爲UMA系統。圖1說明一個典型的UMA系統10的配置。若干處理器12連接至一個公用的系統匯流排14上，並連接至一個記憶體16上。因任何處理器12到記憶體16中任何位置的該路徑均爲相同的(即橫過該系統匯流排)，故任何處理器到任何記憶體位置的該等待時間係相同的。

圖1亦說明快取記憶體18。必須有一個快取記憶體相干協定管理快取記憶體18，並確保依順序地更新一個單一的記憶體位置，以便所有的處理器察看該相同的更新序列。於UMA系統中，如該描述的UMA系統，經常藉著使每一個快取記憶體控制器在該系統匯流排上“窺察”以完成此。其包括觀察該匯流排上所有的異動，且當該匯流排上的一個操作意指“該窺察器快取記憶體中正在保留的一個記憶體位置”時、則採取行動(即參與該相干協定)。

該結構形態的利益爲簡化平行的程式設計，因爲處理可對資料置放較不敏感；即可以一個特殊的時間量存取資料，而不管用以保留該資料的該記憶體位置。

該結構形態的缺點爲UMA系統無法適當地衡量。當系統設計的越來越大時(具有越來越多的處理器和記憶體)，則維持記憶體存取時間的該單一性即變得越來越困難、越來越昂貴。此外，要求快取記憶體控制器窺察的設計需要一個公用的通訊媒體(如一個公用的系統匯流排)作爲資料位址。然而，該系統匯流排爲一種串列式的資源，當該系統匯流排上置放越多的處理器和越多的記憶體操作時，該系統匯流排將變成超載。當該系統匯流排達到飽和時，則增

五、發明說明（ 5 ）

加越多和高速的處理器並不能改良系統的性能。

一種更進一步的系統變化爲“NUMA”，其意指非單一記憶體存取，且說明一種系統配置，其中一個電腦系統中的多個處理器共享一個實際位址空間，且記憶體等待時間隨著該正在存取的記憶體位置而變。即某些記憶體位置比較靠近某些處理器，而和其它處理器較遠。和一個UMA系統不同，在NUMA中，不可在同樣的時間存取一個指定處理器中所有的記憶體位置；即某些記憶體位置花費較長的存取時間（較其它記憶體位置長），因此記憶體的存取時間爲非單一的。

如圖2中所示，一個NUMA系統履行分散式共享記憶體：即該系統的總記憶體爲節點22中該等記憶體M_1、M_2、M_3的總和。存在一個單一的實際位址空間，爲系統20中該等所有節點22所共享，且於圖2中，每一個節點包含該三分之一的系統記憶體。每一個節點22包括一個UMA系統10。若干節點的每一個節點經由一個網路界面(NI) 26連接至一個公用的通訊結構或網路24上。

一個節點中的一個處理器可經由一個載入或儲存指令存取另一個節點中的一個記憶體位置。該NUMA記憶體控制器(NMC) 28功能負責“擷取該區域節點系統匯流排上的該記憶體請求”和“將該記憶體請求傳遞給包含該目標記憶體位置的該節點(即該本地節點)”。因“一個處理器到一個遠端記憶體位置的該路徑”比“該相同的處理器到一個區域記憶體位置的該路徑”遠，故該等記憶體的存取時間爲非單

五、發明說明（ 6 ）

一的。

就如該UMA系統一樣，藉相同的協定將快取記憶體保持相干。所有節點上的所有處理器將串列式地察看一個單一記憶體位置的更新。然而，和UMA系統不同，NUMA系統典型地無法將記憶體操作傳播給所有的節點，以使所有的快取記憶體控制器能夠窺察記憶體操作。反之，該本地節點NMC負責“將相干請求傳遞給該等對相干請求感興趣的遠端節點”。於典型的NUMA履行中，每一個NMC對其節點中該全部的記憶體維持一個目錄。該目錄追蹤區域記憶體的每一條快取記憶體線，以保持該條快取記憶體線的狀態、並使知道哪些其它的節點正在快取該條快取記憶體線。例如，節點1中的NMC 28追蹤M_1中該全部的記憶體。如在節點1上發生一個記憶體操作，則節點1上的NMC 28諮詢它的目錄，且可將該請求傳遞至所有使該目標線被快取的節點。將於美國專利5,710,907、定名爲“用以在該等快取記憶體模式之間選取之混合HUMA COMA快取記憶體系統和方法”中，詳細地說明在一個NUMA系統中、遠端記憶體存取之一個範例的資料流程，藉此以提及的方式將其完全併入本文中。

該一種配置的利益爲較易於建立一種擊越一個UMA系統該等限制的系統。其主要的原因爲所有的快取記憶體控制器不須窺察一個單一、公用的通訊結構。反之，該等快取記憶體控制器僅窺察一個區域的結構－僅當該等快取記憶體控制器影響該節點時、才察看記憶體操作。

五、發明說明（ 7 ）

一個NUMA系統的缺點爲性能敏感程式將視資料置放在記憶體中的何處而不同地執行。此對平行的程式而言係特別重要，其可在該程式執行的許多條串列之間共享資料。

一個使該分散記憶體其記憶體等待時間加劇增加的第二個問題爲該等NUMA系統其有限的快取記憶體大小。某些NUMA系統無法提供比該等SMP更高的快取能力（因該等SMP而建立NUMA系統），於該事例中，該增加的記憶體等待時間減少了該等硬體快取記憶體的利益。選擇性地，可提供另一種硬體快取等級，可能在該NMC中。然而，此傾向於專屬硬體，其意指僅當大量存取遠端記憶體時、才會成爲一項實質上的費用。

如一種更進一步的系統變化"S-COMA"，意指簡單唯快取記憶體配置（即該一個改變的唯快取記憶體配置(COMA)），並說明一種分散式共享記憶體配置，其中一個電腦系統中的多個處理器可透通地存取該複合物中任何的記憶體，且記憶體等待時間隨著該目前正在存取的記憶體位置而變。然而，和NUMA系統不同，在S-COMA中節點維持獨立的實際位址空間。利用每一個節點該區域實際記憶體的一部分作爲一個快取記憶體，配置成頁次、並由系統軟體作成某種大小。於上面編入的美國專利5,710,907中詳細說明該S-COMA操作的特質。

該一種配置的一項利益爲易於建立一種在攀越上比UMA或NUMA更佳的系統。該主要的原因爲每一個節點僅管理其區域的實際記憶體空間，減少了系統的複雜性和節點之

五、發明說明（ 8 ）

間的干擾。同時，就如NUMA一樣，所有的快取記憶體控制器不須窺察一個單一、公用的通訊結構。反之，該等快取記憶體控制器僅窺察一個區域的結構－僅當該等快取記憶體控制器影響該節點時、才察看記憶體操作。

此外，S-COMA藉提供極大的主記憶體快取、以比NUMA提供更佳的平均等待時間給許多程式。因提供極大的快取記憶體，故可顯著地減少快取記憶體未得到的總數，且因此顯著地減少遠端記憶體的存取，藉以改良程式的性能。此外，S-COMA藉減少記憶體管理資料結構的競爭狀況、以比NUMA提供更佳的可縮放性和節點隔離特徵。其亦經由一個翻譯功能過濾位址以限制遠端節點的直接記憶體存取。

參考圖3，就S-COMA配置產生一個全域記憶體物件，並配置一個全域位址(GA)給該全域記憶體物件，且指定一個單一的節點作爲任何特殊資料頁的本地節點30。藉指定一個虛擬位址(VA)給一個全域物件、以將該全域物件接附在每一個感興趣的處理位址空間上。隨後，利用頁次表(PT)將該VA翻譯成一個區域實際位址(RA)。

每一個節點30、32維持一個S-COMA快取記憶體34；即由該S-COMA次系統維持主記憶體36中的一個快取記憶體。當參考該全域資料區域時，則配置該S-COMA快取記憶體中的一個磁格，且藉著將資料置放在本地節點的S-COMA快取記憶體34中，使該資料成爲常駐在本地節點30中的記憶體。提供本地30上的該S-COMA裝置，使本地S-

五、發明說明(9)

COMA快取記憶體線34的位址(RA)和該目標線的全域位址(GA)結合。提供客户節點32上的該S-COMA裝置，使客户S-COMA快取記憶體線34的位址(RA')和該目標線的全域位址結合。

當一個客户32企圖參考"不在該區域L2快取記憶體中"的全域資料時，則檢視該客户S-COMA快取記憶體34。如該資料在那裡爲有效的，則從區域記憶體(RA')中擷取該資料，並完成該請求。如該資料未出現或在客户S-COMA快取記憶體34中不是一個有效的狀態，則該客户S-COMA dir.' 38和該本地S-COMA dir. 38相通訊，以擷取該資料的一個有效複印。

在該節點間通訊的每一個節點上，該S-COMA裝置執行一個邊界功能，將該S-COMA快取記憶體磁格該等相關的區域實際位址(RA)、(RA')翻譯成一個全域位址(GA)。注意，每一個節點30、32可對該目標線的S-COMA快取記憶體磁格利用一個不同的實際位址，但所有的節點均使用該相同的全域位址、用以識別一個特殊的全域線。以此方式，在節點之間維持獨立性。

一個S-COMA快取系統承受了"在頁次增量中配置快取記憶體磁格"的缺點，雖然根據一條快取記憶體線執行相干性。如一個處理利用該S-COMA快取記憶體中配置的每一個頁次中該記憶體的一個大的百分比，則S-COMA可藉提供許多快取記憶體容量、以勝過NUMA。然而，如僅利用每一個配置頁次其相當小的記憶體百分比時，則S-COMA

五、發明說明(10)

藉配置大量的快取記憶體磁格(無效地使用)消耗記憶體。

如一個更進一步的變化，一種混合快取記憶體配置、連同一個多處理器電腦系統的一個快取記憶體相干協定(於上面編入的美國許可專利號碼5,710,907、定名爲"用以在該等快取記憶體模式之間選取之混合HUMA COMA快取記憶體系統和方法"中說明之)。於該混合系統的一個履行中，每一個次系統包括至少一個處理器，一個頁次導向的COMA快取記憶體及一個線導向的混合NUMA/COMA快取記憶體。每一個次系統能夠在COMA模式或NUMA模式下獨立地儲存資料。當在COMA模式下快取記憶體時，一個次系統配置一頁記憶體空間，並接著將該配置頁次內的該資料儲存在其COMA快取記憶體中。依據該履行，當在COMA模式下快取記憶體時，該次系統亦可將該相同的資料儲存在其混合的快取記憶體中、作爲高速存取。反之，當在NUMA模式下快取記憶體時，該次系統將該資料(典型地爲一條資料線)儲存在其混合的快取記憶體中。

該上面概述的混合系統有一個缺點爲"該系統依賴一個S-COMA相干裝置"，其中該S-COMA裝置係和一個NUMA相干裝置無關、但卻和該NUMA相干裝置同等的。使用兩種邏輯上完全的裝置履行該等其中所述的混合觀念。此外，本地節點和客戶節點兩者必須同時維持資料的S-COMA快取記憶體，並在全域位址和實際位址之間翻譯。

縱使具有該等上面概述的系統變化，但利用一第一種記憶體形態和一第二種記憶體形態更進一步地增強系統記憶

五、發明說明(11)

體配置(特別是增強一種混合配置)係較合適的。

發明說明

簡短地說，就一個觀點而言、本發明包括一種混合非單一記憶體配置/簡單唯快取記憶體配置(NUMA/S-COMA)的記憶體系統，該記憶體系統在與一具有多個相互耦合的節點之電腦系統聯合時係有效益的，其中資料係儲存在該混合的NUMA/S-COMA記憶體系統中、作爲多個頁次，且每一個頁次包括至少一條資料線。該混合NUMA/S-COMA記憶體系統包括多個NUMA記憶體，該等多個NUMA記憶體配置成儲存該至少一條的資料線。該等多個NUMA記憶體的每一個NUMA記憶體常駐在該電腦系統該等多個節點之一個不同的節點上。該等多個NUMA記憶體包括一個NUMA相干次系統，用以整合該等NUMA記憶體之間的資料轉移。該混合NUMA/S-COMA記憶體次系統更進一步包括NUMA記憶體中至少一個S-COMA快取記憶體，配置成儲存該等多個頁次的至少一個頁次。每一個S-COMA快取記憶體常駐在該電腦系統該等多個節點之一個不同的節點上。該至少一個S-COMA快取記憶體使用該NUMA相干次系統，以於該電腦系統該等多個節點的另一個節點中接收或轉移資料。

就另一個觀點而言，本發明包括一種"用以在一具有多個相互耦合的節點之電腦系統的一個客户節點和一個本地節點之間通訊"之方法，其中該電腦系統使用一種混合非單一記憶體配置/簡單唯快取記憶體配置(NUMA/S-COMA)

五、發明說明（ 12 ）

記憶體系統，該記憶體系統具有多個NUMA記憶體，該等多個NUMA記憶體配置成儲存資料，每一個NUMA記憶體常駐在該電腦系統一個不同的節點上，且至少一個S-COMA快取記憶體配置成儲存資料，每一個S-COMA快取記憶體常駐在該電腦系統一個不同的節點上，該客户節點包括該至少一個S-COMA快取記憶體的一個S-COMA快取記憶體。該通訊方法包括：在該電腦系統該等多個節點的該客户節點上產生一具有一實際記憶體位址的資料請求；判定該實際位址是否包括該客户節點上的一個區域實際位址；當該實際位址包括一個區域實際位址時，判定該區域實際位址是否需要一個邊界功能翻譯、以將該區域實際位址翻譯成一個本地節點實際位址；及當請求該邊界功能翻譯時，則將該區域實際位址翻譯成該本地節點實際位址，其中該本地節點實際位址包括一個網路位址，其中該客户節點使用該網路位址、以要求存取該本地節點實際位址上的資料。

就一更進一步的觀點而言，本發明包括一種“用以配置一個混合非單一記憶體配置/簡單唯快取記憶體配置(NUMA/S-COMA)記憶體系統”之方法。該方法包括：提供一具有多個相互耦合的節點之電腦系統；以一個非單一記憶體存取(NUMA)配置配置該電腦系統，該NUMA配置包括一個NUMA相干次系統；及在該電腦系統該等多個節點之至少一個節點上建立一個簡單唯快取記憶體配置(S-COMA)的快取記憶體，建立該S-COMA快取記憶體包括：

五、發明說明（ 13）

利用該NUMA相干次系統(無需使用一個S-COMA相干次系統)對資料參考擷取、資料移動及相干管理配置該S-COMA快取記憶體。

又就一更進一步的觀點而言，本發明包括一個製造物件，該製造物件包括至少一個電腦可用的媒體，且該電腦可讀取程式碼裝置編入在該製造物件中、導致一具有多個相互耦合的節點之電腦系統的一個客户節點和一個本地節點之間的資料通訊，其中該電腦系統使用一種混合非單一記憶體配置/簡單唯快取記憶體配置(NUMA/S-COMA)記憶體系統，該記憶體系統具有多個NUMA記憶體，該等多個NUMA記憶體配置成儲存資料，每一個NUMA記憶體常駐在該電腦系統一個不同的節點上，且至少一個S-COMA快取記憶體配置成儲存資料，每一個S-COMA快取記憶體常駐在該電腦系統一個不同的節點上，該客户節點包括該至少一個S-COMA快取記憶體的一個S-COMA快取記憶體。該製造物件中的該電腦可讀取程式碼裝置包括：電腦可讀取程式碼裝置，用以使一個電腦實現“在該等多個節點的該客户節點上產生一具有一個實際記憶體位址的資料請求”；電腦可讀取程式碼裝置，用以使一個電腦實現“判定該實際位址是否包括該客户節點上的一個區域實際位址”；電腦可讀取程式碼裝置，當該實際位址包括一個區域實際位址時，用以使一個電腦實現“判定是否需要一個邊界功能翻譯”，該邊界功能翻譯將該區域實際位址翻譯成一個本地節點實際位址；及電腦可讀取程式碼裝置，當

五、發明說明（ 14 ）

需要該邊界功能翻譯時，用以使一個電腦實現“將該區域實際位址翻譯成該本地節點實際位址”，其中該本地節點實際位址包括一個網路位址，其中該客户節點使用該網路位址、以要求存取該本地節點實際位址上的資料。

如此處所述，一個混合NUMA/S-COMA系統有極多的優點。就如呈遞之一個眞實合併的系統而言，可避免需要“各別的S-COMA相干和通訊裝置”和“各別的NUMA相干和通訊裝置”。如此處所述，使用該NUMA導引器移動資料和維持節點間的相干性(縱使隨後整合該S-COMA功能)。就此處所提之一個結合的系統而言，可利用該大量、有彈性的主記憶體快取的S-COMA，其提供額外的快取容量(較一個純粹NUMA的履行上而言)給每一個節點，而無需專屬的快取記憶體。根據本發明所履行的該混合系統應在一個S-COMA的履行上達到極高的性能，爲獲得該最佳化NUMA裝置的手段，並避免該本地節點上全域位址和實際位址之間的一個翻譯。

圖式簡單說明

當考量下面本發明某些較可取體細的詳述、連同該等伴隨的圖式時，將迅速地察知本發明上面所述及其它的目地、優點和特徵，其中：

圖1，爲一個典型的單一記憶體存取(UMA)配置之一個體系的圖示；

圖2，爲一個非單一記憶體存取(NUMA)配置之一個體系的圖示；

五、發明說明（ 15 ）

圖3，描述一個一般的簡單唯快取記憶體配置(S-COMA)；

圖4，爲根據本發明所履行的一個混合NUMA/S-COMA系統之一個高階說明；

圖5，爲根據本發明、在一個混合NUMA/S-COMA系統內的一個客户節點上所履行的記憶體請求邏輯的一個體系之流程圖；

圖6，爲根據本發明、在一個混合NUMA/S-COMA系統內的一個本地節點上所履行的記憶體請求邏輯之流程圖；

圖7，爲根據本發明、在一個混合NUMA/S-COMA系統中該要求節點上的入站記憶體請求處理邏輯的一個體系之流程圖；

圖8，爲根據本發明、在一個混合NUMA/S-COMA系統中一個本地節點上的混合記憶體出站處理邏輯的一個體系之流程圖；及

圖9，爲根據本發明、在一個混合NUMA/S-COMA系統的一個請求節點上的混合記憶體入站處理邏輯的一個體系之流程圖。

實行本發明的最佳模式

本發明提供一種利用該NUMA相干裝置用作資料參考擷取、資料移動及相干管理，用以在一個非單一記憶體存取(NUMA)基礎的上面建立一個簡單唯快取記憶體配置(S-COMA)系統。經由該S-COMA結構，利用每一個節點上一部分的主記憶體作爲全域記憶體的一個資料快取記憶體。

五、發明說明（ 16 ）

由使用該等S-COMA快取記憶體的該區域節點管理該等每一個S-COMA快取記憶體。

藉由本發明，則每一個從一個本地節點中快取資料的節點均能夠快取該正規NUMA層次中的資料，且可視需要快取S-COMA快取記憶體(該等S-COMA快取記憶體間係相互無關的，且該本地節點無需特別知曉該等S-COMA快取記憶體的快取決策)中的該資料。

“利用該NUMA家區域實際位址作爲S-COMA快取記憶體的該全域位址”，且“將該家區域實際位址翻譯成一個客户區域實際位址，用以在該客户節點上執行該相干協定”以完成此。同樣地，就從該客户移動到該家的相干訊息而言，在將該客户區域實際位址傳輸至連接該等節點的該網路中之前，先將該客户區域實際位址翻譯成一個全域位址(爲一個家區域實際位址的該形式)。

於該標準的S-COMA履行中，所有的節點對參考全域資料維持S-COMA快取記憶體，即使該資料爲本地節點。當一個本地節點將資料引入其區域記憶體中給客户節點使用時，該本地節點必須將該資料保持在其自己的S-COMA快取記憶體中。爲使該S-COMA裝置獲得控制“該資料的參考”、以執行該相干管理，此係必須的。

藉由本發明，則僅利用S-COMA保持在別處成家的快取記憶體線。因由該家NUMA裝置執行所有的相干管理，故不須在該等本地節點的該等S-COMA快取記憶體中快取線；即正規的(非S-COMA的)記憶體保持區域線，且該

五、發明說明(17)

NUMA目錄將追蹤在該節點成家的線的使用。當一個相干行動爲必要的時，則利用該等標準NUMA裝置將相干訊息傳遞給其它的節點，並傳遞該線的家實際位址。

以該同樣的方式利用該家實際位址作爲一個S-COMA系統上的一個全域位址。S-COMA客户節點將於正在該節點中傳進和傳出的該等位址上執行一個邊界功能翻譯。此類似於以一個標準S-COMA履行而發生的該翻譯(在全域位址和區域實際位址之間翻譯)。然而，於該混合履行中，該客户直接在"該本地節點使用的一個家實際位址"、和"代表該客户區域實際記憶體中該S-COMA快取記憶體的一個客户實際位址"之間翻譯。

圖4描述本發明一個體系，例證說明此處所建議之一個混合NUMA/S-COMA環境該等不同位址型態之間的關係。該環境包括一個本地節點40、和客户節點42，其中本地節點40和客户節點42分別包括實際位址(RA_H)和(RA_L)。就如圖3一樣，客户節點42維持一個與本地節點40無關的區域實際位址空間46中的一個S-COMA快取記憶體44。然而和圖3的該體系不同，本地節點40和客户節點42之間的通訊並非以一個正規的全域位址發生的，而是以本地節點40利用的該實際位址(RA_H)或(RA_{home})發生的。客户節點42將具邊界功能49之NUMA導引器中該接收的實際位址翻譯成一個相對應的區域實際位址(RA_L)或(RA_{local})，並指向客户S-COMA快取記憶體44。

本發明的一個觀點爲將根據一個系統寬而使用的該

五、發明說明（ 18）

NUMA位址翻譯成一個區域位址，其中該區域位址僅引用至每一個與該家通訊的節點上的該S-COMA快取記憶體上。

一個全域記憶體系統(如NUMA)利用分散在一個系統中多個節點之間的一個位址空間。該位址本身指定由哪一個節點提供該向後記憶體給該正在被定址的資料。例如，在一個非常簡單的系統中，一個位址的該高階位元組可指定該節點號碼，且該位址的剩餘部分可指定該節點內該記憶體的位置。

當一個節點從另一個節點中接收一個家位址作爲輸入時，可將該位址翻譯成某一個關聯的位址，即一個直接相對應至該家位址的區域位址、以作更進一步的處理。它的好處爲該區域節點可管理其自己的位址、而與該剩餘的系統無關，就如S-COMA一樣。在一個多重作業系統的環境中此係特別重要，其中每一個節點執行一個獨立的作業系統事例。注意，亦可以節點的子集合取代該名詞每一個節點。爲了簡化，今後均使用該名詞每一個節點，且可察知每一個節點意指在一個單一作業系統事例的控制下所操作的一個節點或一個節點的集合。

此處，該位址翻譯裝置意指爲該邊界功能。在每一個客戶節點上，該邊界功能維持一個目錄用以追蹤記憶體位址(其中當該節點正在傳遞或接收該等記憶體位址時、該等記憶體位址需要翻譯)。可由該窺察邏輯(用作區域記憶體參考)和該網路邏輯(用作遠端節點中的參考)存取該目

五、發明說明（ 19 ）

錄。每一個目錄條目包含該區域實際位址和該興趣線其相對應的本地節點位址。如該條目中的一個旗標指示需要該翻譯，則以該目錄條目中的該關聯位址取代該輸入位址。

經由該S-COMA結構，利用主記憶體的一部分作爲全域記憶體的一個資料快取記憶體。由利用該等S-COMA快取記憶體的該區域節點管理該等每一個S-COMA快取記憶體。

將利用該NUMA裝置提供資料參考擷取、資料移動和相干裝置。此在資料參考擷取、資料移動和相干性上避免需要一個各別的S-COMA裝置。將於下更進一步詳列本發明的該等觀點。

1. 從一個NUMA系統（如圖2中所描述的）開始。於該NUMA系統中，該系統的實際位址空間分散在該等節點之間，且爲共享的（即任何節點可參考任何記憶體位置）。一包含一個用以維持每一個記憶體位置相干性的NUMA系統，維持一個客户表列，該客户表列快取一個特殊的記憶體位置、並將相干請求的路線定在該主記憶體位置或該客户記憶體位置上。

2. 相干目錄：每一個NMC保留一個表，該表給予每一條在該節點上成家的記憶體線一個條目。每一個條目中均有一個客户節點表列，其中在該等客户節點上正在快取該條線。以該實際記憶體位址索引該表。

五、發明說明（ 20 ）

3. 系統軟體在每一個節點上配置一部分的主記憶體，作爲一個S-COMA快取記憶體。

4. 邊界功能翻譯表：每一個NMC保留一個表，該表給予每一個正作爲S-COMA快取記憶體之用的記憶體頁次一個條目。每一個條目包含該快取資料的該本地節點實際位址(RA_{home})。以該區域實際記憶體位址(RA_{local})索引該表。亦可利用該家實際記憶體位址作爲輸入、並藉導出該區域實際記憶體位址發現一個條目(雖然此可能係一種效率較低的查尋)。

5. 該NUMA記憶體次系統將該區域S-COMA快取記憶體其範圍中所有的位址視爲在別處成家的位址。即該區域NMC未對該區域S-COMA快取記憶體其範圍中的位址啓始相干行動。反之，該區域NMC執行一個邊界功能翻譯，並將該記憶體請求傳遞給相對應至該邊界功能翻譯表中該目標區域實際位址的該本地節點實際位址(RA_{home})的家中。

6. 每一個NMC僅對該節點中出站或入站的操作執行一個BF翻譯表查尋，並繼之僅對該節點不是其家的該等位址執行一個BF翻譯表查尋。此係以該邊界功能著稱。

7. 以該表中該關聯的數值資料位址替代該出站資料中的該資料位址。

當該客户節點經由其區域實際位址參考其S-COMA快取記憶體中的該資料時，且當該區域相干目錄中的狀態指示

五、發明說明（ 21 ）

爲無效的時，則由該正規的NUMA裝置停止該區域參考，該BF將該區域實際位址翻譯成該本地節點實際位址，並由該正規的NUMA裝置將該資料請求傳遞給該本地節點。注意，該NUMA次系統能夠定訂該線請求的路線，因該網路位址爲一個實際的NUMA位址。

未執行任何邊界功能的該本地節點，接收該請求、並以標準的NUMA邏輯、遵照該請求行事。以一個回應的訊息將該請求的資料傳遞回給該客户節點。

該客户節點接收該請求資料的回應，並在該網路位址（即該本地節點實際位址）上執行一個BF查尋。在該BF翻譯表中找出該RA_{home}，該BF以該表中相對應的區域實際位址替代該位址。該記憶體控制器在該相干目錄中將該條線標誌爲有效的，並將該資料回應傳遞給該客户節點該請求的處理器。此完成該請求。

當該客户經由其區域實際位址參考其S-COMA快取記憶體中的該資料時，且當該狀態爲有效的時，則該條線在其S-COMA快取記憶體中，且該條線未要求更進一步的行動。不需任何BF翻譯，且僅由該正規的NUMA裝置在區域記憶體等待時間時從區域記憶體中將該資料送回。

當該本地節點傳出一個相干請求給該等客户節點時，該本地節點係利用其自己的NUMA位址（其代替一個純粹S-COMA系統中的該全域位址）執行此。每一個客户節點接收該請求，且該BF將該進來的位址（該本地節點實際位址）翻譯成該區域位址（如該BF翻譯表中所找到的）。接著，

五、發明說明(22)

利用該區域位址(RA_{local})局部地處理該請求。

圖5-9描述根據本發明、在一個混合NUMA/S-COMA系統內履行之邏輯流程的一個體系。假設該討論係遵循圖4的該系統，其中一個請求處理器包括該客户節點、及一可能包括如上所述之一個NUMA記憶體控制器(NMC)的記憶體管理單元。此外，該實體位置(PA)相當於上面所述的一個實際位址(RA)。

從圖5開始，由一個客户節點啓始該邏輯流程(100)，或該邏輯流程要求處理器提出一條資料線的一個虛擬位址(VA)給其區域記憶體管理單元(MMU)(110)。該區域MMU將該VA轉換成一個實體位址(PA)，且該邏輯判定在該其中一個請求處理器硬體快取記憶體中是匹有該合適資料線的一個有效複印(120)。如此處所用，該等級2快取記憶體意指一個正規或標準的快取記憶體，而非一個S-COMA快取記憶體，將於下更進一步說明之。如該資料在該區域的標準快取記憶體中，則該L2快取記憶體將該資料線提供給請求處理器(130)，且該資料擷取處理完成(140)。

如該實體資料未在該其中一個請求器硬體快取記憶體中，則該L2快取記憶體將該實體位址(PA)提供給請求次系統的該NUMA記憶體控制器(NMC)(150)，並在該NMC上查詢該實體位址對該請求的處理器而言是否爲一個區域實體位址(160)。如是，則使用圖8的該混合記憶體出站處理(170)，如下更進一步說明之。基本上，圖8的邏輯判定該實體位址是否爲該請求處理器上一個區域S-COMA快取記

五、發明說明（ 23 ）

憶體的一部分。如該實體位址未包括一個區域實體位址，則該資料線係在一個遠端的實體位址上，且該邏輯改變成一個純粹的NUMA處理範例，其中該請求的NMC將一個資料請求傳遞給該訊息請求中具有該資料位址的該本地次系統(180)。

從該請求節點至圖6、200的該本地節點處理之該等邏輯轉換(190)，其中該家次系統接收該請求訊息、並更新其相干目錄，以將一個回應反映給該讀取請求，即該資料線的新狀態(210)。該資料線的新狀態代表“哪一個處理節點具有該資料線的一個有效複印”、“哪一個爲傳統的NUMA處理、以追蹤該等正在快取該資料線的客户節點”、及“該資料線的狀態(即有效或無效的)”。接著，該本地次系統判定其在本地記憶體中是否具有該資料的一個有效複印(220)，且如是，則該本地次系統使用該傳統的NUMA處理從快取客户次系統中撤回該資料線的快取複印(230)。一旦該本地次系統在記憶體中具有該資料線的一個有效複印時，即將該資料提供給該請求次系統(240)，且該等邏輯轉換至該請求次系統中、以作更進一步的處理，如圖7、250中所示。

一旦從該本地次系統中接收到該回返的資料時，該請求次系統上的處理即從300開始，判定該客户節點上接收的該實體位址是否爲一個區域實體位址(310)。如否，則根據本發明執行圖9的該混合記憶體入站處理(320)。

如圖9中所示，入站處理從500開始“在該請求次系統

五、發明說明（ 24 ）

上、正從該本地次系統中接收該請求的資料(510)”。接著，該請求的NMC利用該接收的資料位址作爲一個網路位址(如此處上面所述)、對該回應訊息中的該資料位址執行一個邊界功能目錄查尋(520)。如在該邊界功能翻譯表中找到該網路位址(NA)(530)，則將該資料位址從該網路位址翻譯成該邊界功能翻譯表中一個相對應的區域實際位址(RA_{local})，以由該請求次系統使用之(540)。如在該翻譯表中未找到該網路位址、或已依照該邊界功能執行與翻譯相關的部分，則處理折返至圖7的該邏輯流程(550)。

繼續圖7的該請求次系統處理，如該實體位址爲一個區域實體位址、或在圖9的該混合記憶體入站處理之後，則該請求的NMC將該資料回應傳遞給請求處理器(330)，且視需要將該資料線儲存在該L2快取記憶體中(即正規的硬體快取記憶體中)、以作爲隨後之用(340)。根據本發明的該體系而完成處理(350)。

折返至圖5，如該實體位址爲一個區域位址，則形成查詢(160)，並執行圖8的該混合記憶體出站處理。該處理從該請求次系統開始(400)“該請求的NMC存取該相干目錄、以判定該資料線在該請求器區域記憶體中是否爲一個有效的狀態(410)”。如該回應爲“是”，則該區域記憶體將該資料線提供給該請求處理器，不管該記憶體是否爲S-COMA快取記憶體(430)，藉以完成該擷取的處理(440)。

如在區域記憶體中沒有該資料的有效複印，則邏輯判定是否需要一個邊界功能(450)，即藉判定該請求的位址是否

五、發明說明(25)

在該S-COMA快取記憶體的範圍中。如否，則其不是一種S-COMA狀態，且已將該資料提至該系統的另一個節點中。如該選取的位址爲一個S-COMA快取記憶體的一部分，則必須將該位址翻譯成該次系統能夠瞭解的一個本地位址。如是，該請求的邊界功能存取該邊界功能翻譯表，並將該請求中的該區域實際記憶體位址翻譯成一個網路位址(460)，並可接如上所述著將該網路位址傳遞至該本地次系統(470)，藉以完成圖8的該混合記憶體出站處理(480)。

概述之，對那些熟知該技藝的人來說，將從上面的討論中察知此處所述之一種混合NUMA/S-COMA系統具有極多的優點。又該呈遞之系統包括一個眞實合併的系統，其中藉使用該NUMA相干和通訊裝置免除需要"一個各別的S-COMA相干和通訊裝置"。如此處所述，使用該NUMA導引器移動資料和維持節點間的相干性(縱使隨後整合該S-COMA功能)。就該所提之結合的系統而言，可利用該大量、有彈性的主記憶體快取的S-COMA，其提供額外的快取容量(較一個純粹NUMA的履行上而言)給每一個節點，而無需該專屬的快取記憶體。有利地，藉由該提出的履行避免一個本地節點上一個全域位址和實際位址之間的翻譯。

可將本發明包括在一個具有例如電腦可用的媒體之製造物件中(例如一個或多個電腦程式產品)。該媒體編入在該製造物件中，例如電腦可讀取程式碼裝置，用以提供和助益本發明的該等能力。可包括該等製造物件作爲該電腦系

五、發明說明（ 26 ）

統的一部分或各別地售賣。

此外，可提供至少一個機器可讀取的程式儲存體裝置，實實在在地編入該機器可執行之至少一個程式的指令），以執行本發明的該等能力。

藉由範例提供此處所描述的該等流程圖示。在未偏離本發明的精髓下，可對此處所述的該等圖示或該等步驟（或操作）作一些變化。例如，於某些事例中，可以不同的順序執行該等步驟，或可增加、刪除或修正某些步驟。該等所有的變化均考慮成包括本發明的一部分（如該附加申請專利範圍中所詳述的）。

當已根據某些與本發明相關之較可取體系於此處詳述本發明之後，對那些熟知此技藝的人來說，可在其中實現許多修正和變更。如是，該附加的申請專利範圍意欲涵蓋該等所有落在本發明該眞實精髓和範疇內的修正和變更。

四、中文發明摘要（發明之名稱：混合非單一記憶體架構/簡單唯快取記憶體架構之系統及方法）

本發明揭示一種混合非單一記憶體配置/簡單唯快取記憶體配置(NUMA/S-COMA)之記憶體系統及方法，該記憶體系統和方法在與一具有多個相互耦合的節點之電腦系統聯合時係非常有用。該等多個節點包括NUMA記憶體，配置以儲存資料線。該等NUMA記憶體包括一個NUMA相干次系統，用以整合該等節點之間的資料轉移。在該電腦系統的至少一個節點上提供至少一個S-COMA快取記憶體。將該至少一個S-COMA快取記憶體配置成使用該NUMA相干次系統在該電腦系統該等多個節點的另一個節點中傳遞或接收資料通訊。利用一個本地節點的實際位址作爲該網路位址以存取儲存在該系統另一個節點上的資料。利用一個邊

英文發明摘要（發明之名稱：HYBRID NUMA/S-COMA SYSTEM AND METHOD）

A hybrid non-uniform-memory-architecture/simple-cache-only-memory-architecture (NUMA/S-COMA) memory system and method are described useful in association with a computer system having a plurality of nodes coupled to each other. The plurality of nodes include NUMA memory which are configured to store data lines. The NUMA memories include a NUMA coherence subsystem for coordinating transfer of data between the nodes. At least one S-COMA cache is provided on at least one node of the computer system. The at least one S-COMA cache is configured to employ the NUMA coherence subsystem in sending data communication to or receiving data communication from another node of the plurality of nodes of the computer system. Data stored at another node of the system is accessed using a home node real address as the network address. The home node real address is translated

四、中文發明摘要（發明之名稱：　　）

界功能翻譯表將該本地節點的實際位址翻譯成該客户節點上的一個區域實際位址。該S-COMA快取記憶體使用該NUMA相干次系統以提供資料參考擷取、資料移動和相干機構，藉以避免需要一個各別的S-COMA相干裝置完成該等功能。

英文發明摘要（發明之名稱：　　）

into a local real address at the client node using a boundary function translation table. The NUMA coherence subsystem is employed by the S-COMA cache to provide data reference capture, data movement and coherence mechanisms, thereby avoiding the need for a separate S-COMA coherence mechanism to accomplish these functions.

六、申請專利範圍

1. 一種混合非單一記憶體配置/簡單唯快取記憶體配置(NUMA/S-COMA)之記憶體系統，該記憶體系統在與一具有多個相互耦合的節點之電腦系統聯合時係非常有用，且其中資料係儲存在該混合NUMA/S-COMA記憶體系統中並分爲多個頁次，該每一個頁次包括至少一條資料線，該混合NUMA/S-COMA記憶體系統包括：

多個NUMA記憶體，配置成儲存該至少一條資料線，該等多個NUMA記憶體的每一個NUMA記憶體常駐在該電腦系統該等多個節點之一個不同的節點上，且其中該等多個NUMA記憶體包括一個NUMA相干次系統，用以整合該等NUMA記憶體之間的資料轉移；及

至少一個S-COMA快取記憶體，配置成儲存該等多個頁次的至少一個頁次，該至少一個S-COMA快取記憶體的每一個S-COMA快取記憶體常駐在該電腦系統該等多個節點之一個不同的節點上，且其中該至少一個S-COMA快取記憶體使用該NUMA相干次系統，以向該電腦系統該等多個節點的另一個節點傳遞資料通訊或自該另一個節點接收資料通訊。

2. 如申請專利範圍第1項之混合NUMA/S-COMA記憶體系統，其中該至少一個S-COMA快取記憶體包括多個S-COMA快取記憶體，且其中該NUMA相干次系統包括該等多個節點中一個客戶節點上的一個NUMA相干協定，以從該電腦系統該等多個節點之一個本地節點中快取該其中一個S-COMA快取記憶體中的資料，使用該NUMA

六、申請專利範圍

相干協定以在該客户節點和該本地節點之間傳遞S-COMA相干訊息。

3. 如申請專利範圍第2項之混合NUMA/S-COMA記憶體系統，其中在該客户節點上具有該相干協定的該NUMA相干次系統包括該客户節點上的一個邊界功能翻譯表，當在該客户節點的該一個S-COMA快取記憶體和該本地節點之間傳遞一個訊息時、用以在一個本地節點實際位址和一個客户節點實際位址之間轉換，其中該本地節點實際位址包括一個該NUMA通訊次系統所使用的網路位址，該網路位址在該本地節點和該客户節點的該S-COMA快取記憶體之間轉移資料。

4. 如申請專利範圍第3項之混合NUMA/S-COMA記憶體系統，更進一步包括一連接該等多個節點的通訊網路，且其中該NUMA相干次系統包括使用該邊界功能翻譯表以將一個客户節點實際位址翻譯成一個本地節點實際位址之裝置，其中該本地節點實際位址包括一個網路位址，當從該客户節點通訊一個訊息至該本地節點中時，使用該網路位址在互連於多個節點之該通訊網路上作傳輸。

5. 如申請專利範圍第3項之混合NUMA/S-COMA記憶體系統，更進一步包括一連接該等多個節點的通訊網路，且其中該NUMA相干次系統包括用以一旦接收到該客户節點上的一個網路位址時即檢視該邊界功能翻譯表及用以當該網路位址與從一個本地節點到該客户節點該一個S-COMA快取記憶體中之一個訊息相關聯時，利用該邊界

六、申請專利範圍

功能翻譯表將該網路位址轉換成一個客户節點實際位址之裝置，該網路位址包括一個本地節點實際位址。

6. 如申請專利範圍第1項之混合NUMA/S-COMA記憶體系統，其中該至少一個S-COMA快取記憶體包括在該等多個節點其至少一個客户節點上實施的一個S-COMA快取記憶體，以在該電腦系統中保持至少一個在別處的本地快取記憶體頁次。

7. 如申請專利範圍第6項之混合NUMA/S-COMA記憶體系統，其中該NUMA相干次系統包括標準的NUMA相干裝置，該至少一個S-COMA快取記憶體使用該等標準的NUMA相干裝置、以將S-COMA相干訊息傳遞給該等多個節點的其它節點。

8. 如申請專利範圍第6項之混合NUMA/S-COMA記憶體系統，其中該客户節點包括用以將資料儲存在客户節點上的該S-COMA快取記憶體中、而無需通知該儲存決策的一個本地節點。

9. 如申請專利範圍第8項之混合NUMA/S-COMA記憶體系統，其中該客户節點和該本地節點具有獨立的作業系統。

10. 如申請專利範圍第6項之混合NUMA/S-COMA記憶體系統，其中該電腦系統的該等多個節點包括多個客户節點，每一個客户節點具有一個S-COMA快取記憶體、且包括用以管理該S-COMA快取記憶體之裝置，該用以管理之裝置包括用以對該S-COMA快取記憶體其相干管理

六、申請專利範圍

使用該NUMA相干次系統之裝置。

11. 一種用以在一具有多個相互耦合的節點之電腦系統的一個本地節點和一個客户節點之間通訊資料之方法，其中該電腦系統使用一種混合非單一記憶體配置/簡單唯快取記憶體配置(NUMA/S-COMA)記憶體系統，該記憶體系統具有多個配置以儲存資料之NUMA記憶體，每一個NUMA記憶體常駐在該電腦系統不同的一個節點上，且至少一個S-COMA快取記憶體配置成儲存資料，每一個S-COMA快取記憶體常駐在該電腦系統不同的一個節點上，該客户節點包括該至少一個S-COMA快取記憶體的一個S-COMA快取記憶體，該通訊方法包括下列之步驟：

(i) 在該電腦系統該等多個節點的該客户節點上產生一個具有一實際記憶體位址的資料請求；

(ii) 判定該實際位址是否包括該客户節點上的一個區域實際位址；

(iii) 當該實際位址包括一個區域實際位址時，判定該區域實際位址是否需要一個邊界功能翻譯、以將該區域實際位址翻譯成一個本地節點實際位址；及

(iv) 當需要該邊界功能翻譯時，則將該區域實際位址翻譯成該本地節點實際位址，其中該本地節點實際位址包括一個網路位址，其中該客户節點使用該網路位址以要求存取該本地節點實際位址上的資料。

12. 如申請專利範圍第11項之方法，其中該區域實際位址包

六、申請專利範圍

括一個S-COMA快取記憶體位址，且該方法更進一步包括利用該等多個NUMA記憶體的一個NUMA相干次系統將一個請求訊息從該客户節點傳遞至該本地節點之步驟，該請求訊息包括該網路位址，該網路位址包括該請求資料的本地節點實際位址。

13. 如申請專利範圍第12項之方法，更進一步包括從該本地節點中接收該客户節點上的一個回應訊息之步驟，該回應訊息包括一個實際位址，且其中該方法更進一步包括決定該回應訊息的該實際位址是否包括該客户節點的一個區域實際位址。

14. 如申請專利範圍第13項之方法，其中當該實際位址與該客户節點上的一個區域實際位址不同時，則該方法包括在該客户節點上對該回應訊息中的該實際位址執行一個邊界功能目錄查尋，且如在該邊界功能目錄查尋中找到該實際位址時，則利用該客户節點的該邊界功能翻譯表將該網路位址中的該實際位址翻譯成一個相對應的區域實際位址。

15. 如申請專利範圍第11項之方法，其中該客户節點包括一個NUMA記憶體控制器，且其中該方法更進一步包括請求該客户節點的該NUMA記憶體控制器將該資料請求從該客户節點傳遞至使用(包括該本地節點實際位址)之該網路位址該本地節點。

16. 如申請專利範圍第11項之方法，其中當該實際位址包括不同於該判定(iii)中的一個區域實際位址時，則該方法更

六、申請專利範圍

進一步包括請求該客户節點上的一個NUMA記憶體控制器將使用該實際位址作爲一個本地節點實際位址的該資料請求傳遞給該本地節點。

17. 一種用以配置一個混合非單一記憶體配置/簡單唯快取記憶體配置(NUMA/S-COMA)記憶體系統之方法，包括：

提供一具有多個相互耦合的節點之電腦系統；

以一個非單一記憶體存取(NUMA)配置配置該電腦系統，該NUMA配置包括一個NUMA相干次系統；及

在該電腦系統該等多個節點之至少一個節點上建立一個簡單唯快取記憶體配置(S-COMA)的快取記憶體，建立該S-COMA快取記憶體包括：配置該S-COMA快取記憶體以無需使用一個S-COMA相干次系統並利用該NUMA相干次系統對資料參考擷取、資料移動及相干管理。

18. 一種製造物件，包括：

至少一個電腦可使用的媒體，且電腦可讀取程式碼裝置編入在該製造物件中、用以致使一具有多個相互耦合的節點之電腦系統的一個本地節點和一個客户節點之間的資料通訊，其中該電腦系統使用一種混合非單一記憶體配置/簡單唯快取記憶體配置(NUMA/S-COMA)記憶體系統，該記憶體系統具有多個配置以儲存資料之NUMA記憶體，每一個NUMA記憶體常駐在該電腦系統不同的一個節點上，且至少一個S-COMA快取記憶體配置成儲存資料，每一個S-COMA快取記憶體常駐在該電腦系統

六、申請專利範圍

不同的一個節點上，該客户節點包括該至少一個S-COMA快取記憶體的一個S-COMA快取記憶體，該製造物件中的該電腦可讀取程式碼裝置包括：

(i) 電腦可讀取程式碼裝置，用以使一個電腦實現在該等多個節點的該客户節點上產生一具有一個實際記憶體位址的資料請求；

(ii) 電腦可讀取程式碼裝置，用以使一個電腦實現判定該實際位址是否包括該客户節點上的一個區域實際位址；

(iii) 電腦可讀取程式碼裝置，當該實際位址包括一個區域實際位址時，用以使一個電腦實現判定是否需要一個邊界功能翻譯，該邊界功能翻譯將該區域實際位址翻譯成一個本地節點實際位址；及

(iv) 電腦可讀取程式碼裝置，當需要該邊界功能翻譯時，用以使一個電腦實現將該區域實際位址翻譯成該本地節點實際位址，其中該本地節點實際位址包括一個網路位址，其中該客户節點使用該網路位址、以要求存取該本地節點實際位址上的資料。

19. 如申請專利範圍第18項之製造物件，其中該區域實際位址包括一個S-COMA快取記憶體位址，且該製造物件更進一步包括電腦可讀取程式碼裝置，利用該等多個NUMA記憶體的一個NUMA相干次系統、用以使一個電腦實現將一個請求訊息從該客户節點傳遞至該本地節點，該請求訊息包括該網路位址，該網路位址包括該請

六、申請專利範圍

求資料的本地節點實際位址。

20. 如申請專利範圍第19項之製造物件，更進一步包括電腦可讀取程式碼裝置，用以使一個電腦實現從該本地節點中接收該客户節點上的一個回應訊息，該回應訊息包括一個實際位址，且其中該製造物件更進一步包括電腦可讀取程式碼裝置，用以使一個電腦實現判定該回應訊息的該實際位址是否包括該客户節點的一個區域實際位址。

21. 如申請專利範圍第20項之製造物件，其中當該實際位址與該客户節點上的一個區域實際位址不同時，則該製造物件包括電腦可讀取程式碼裝置，用以使一個電腦實現在該客户節點上對該回應訊息中的該實際位址執行一個邊界功能目錄查尋，且如在該邊界功能目錄查尋中找到該實際位址時，則該製造物件更進一步包括電腦可讀取程式碼裝置，用以使一個電腦實現利用該客户節點的該邊界功能翻譯表、將該網路位址中的該實際位址翻譯成一個相對應的區域實際位址。

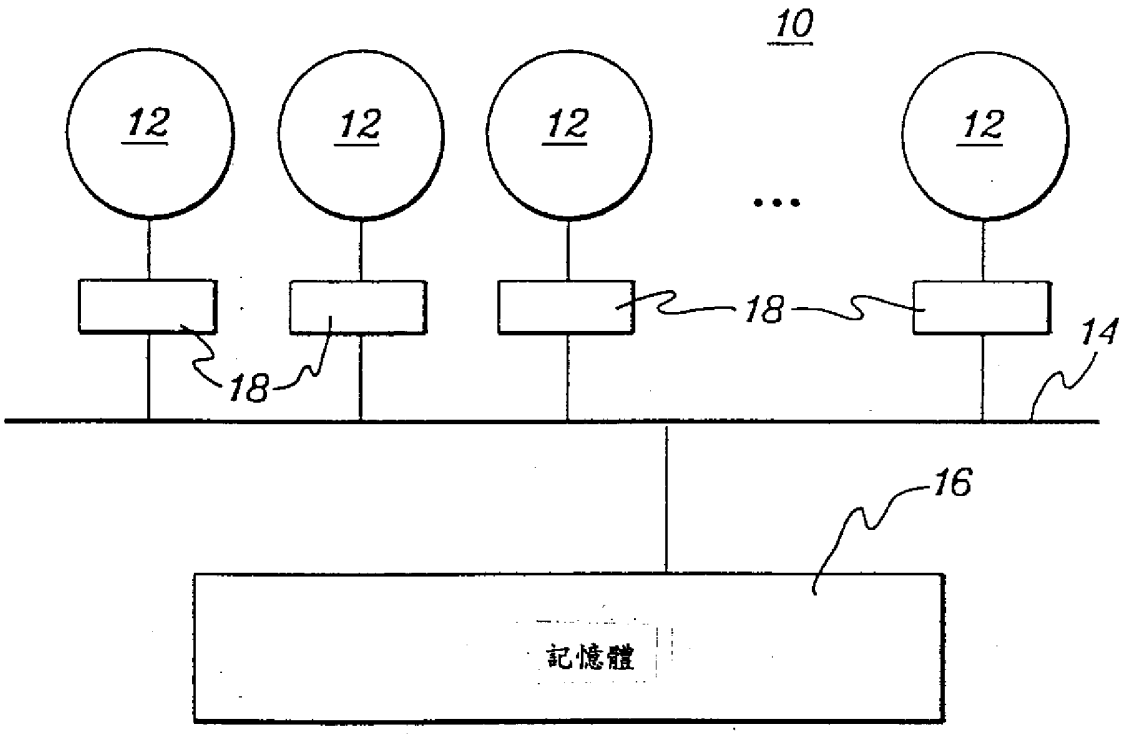

圖 1

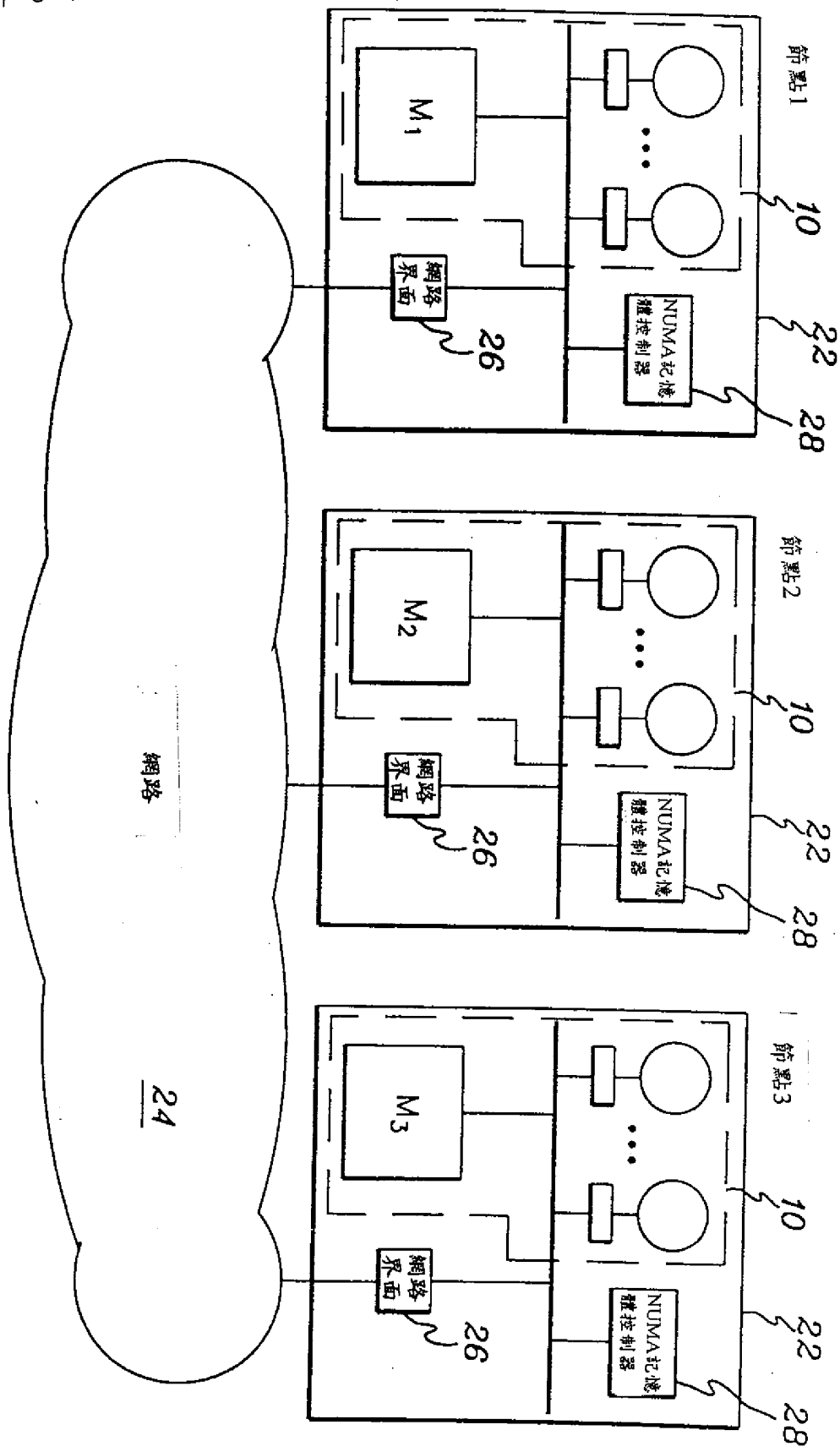
節點1
M_1
10
22
28
網路界面
26
NUMA記憶體控制器
節點2
M_2
10
22
28
網路界面
26
NUMA記憶體控制器
節點3
M_3
10
22
28
網路界面
26
NUMA記憶體控制器
網路
24

圖 2

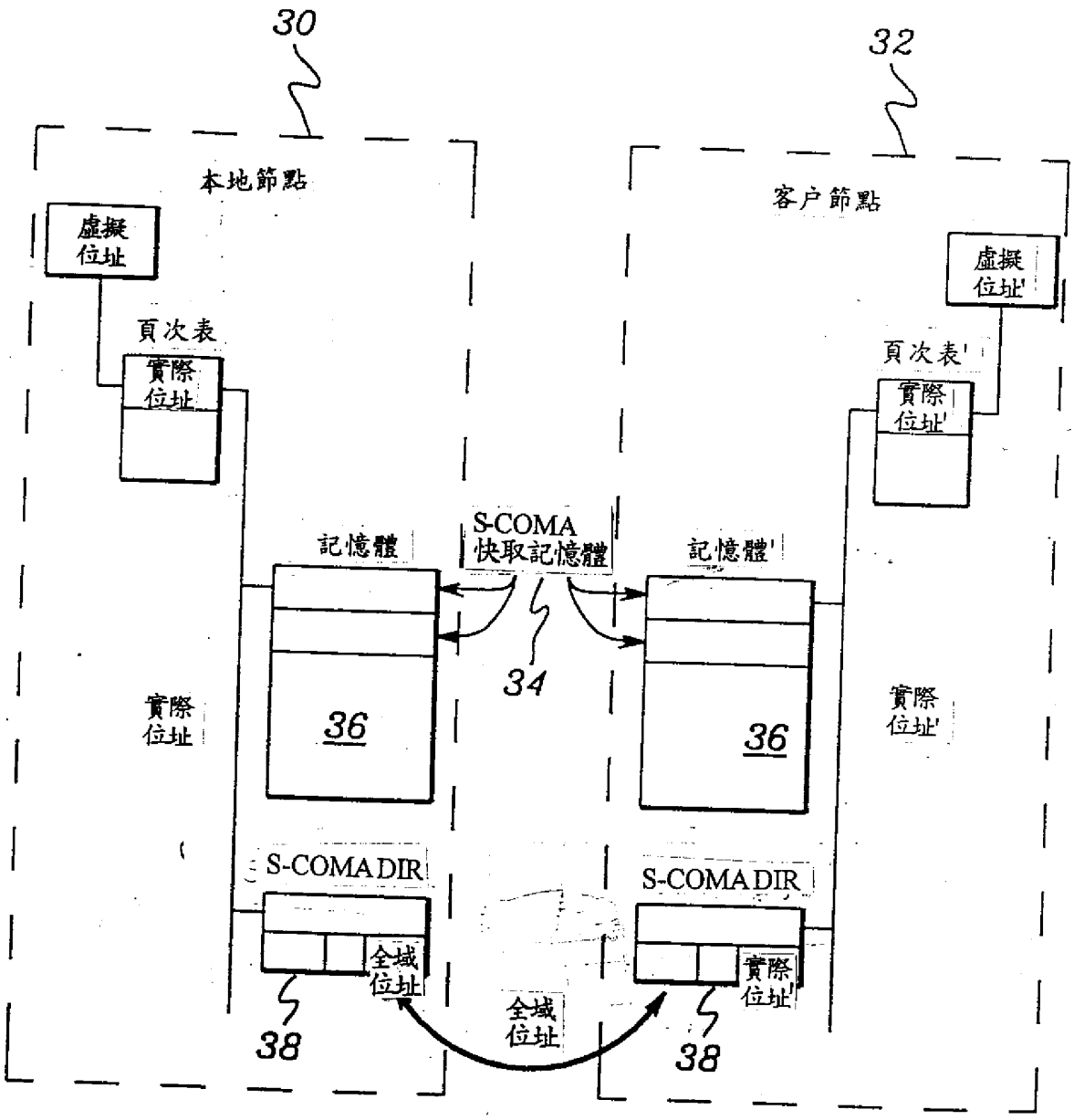
30
32
本地節點
客戶節點
虛擬位址
頁次表
實際位址
虛擬位址'
頁次表'
實際位址'
記憶體
S-COMA快取記憶體
記憶體'
34
36
36
實際位址
實際位址'
S-COMA DIR
S-COMA DIR
全域位址
實際位址'
全域位址
38
38

圖 3

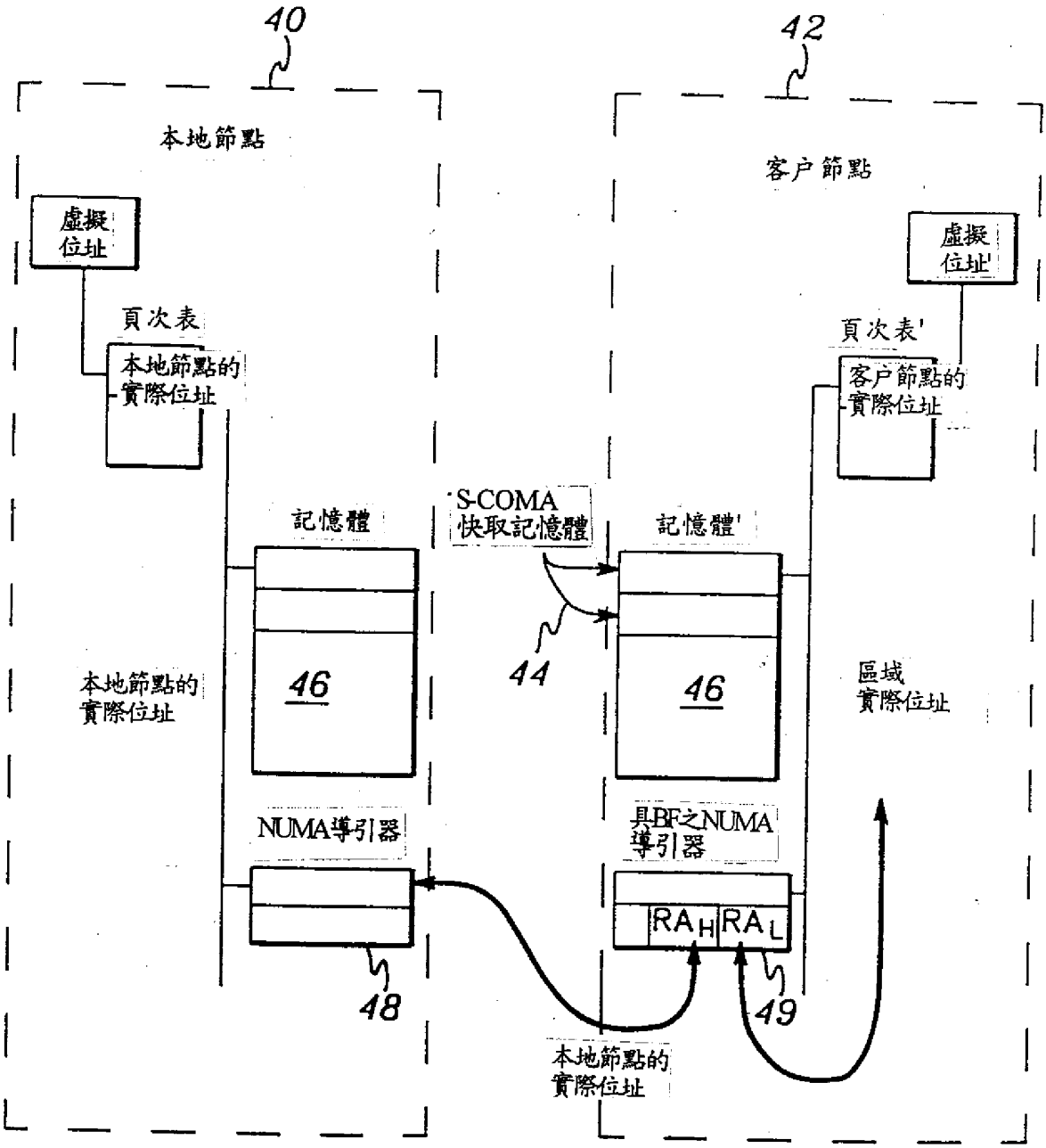
40
42
本地節點
客户節點
虛擬位址
虛擬位址'
頁次表
頁次表'
本地節點的實際位址
客户節點的實際位址
記憶體
S-COMA快取記憶體
記憶體'
46
44
46
本地節點的實際位址
區域實際位址
NUMA導引器
具BF之NUMA導引器
RA H
RA L
48
49
本地節點的實際位址

圖 4

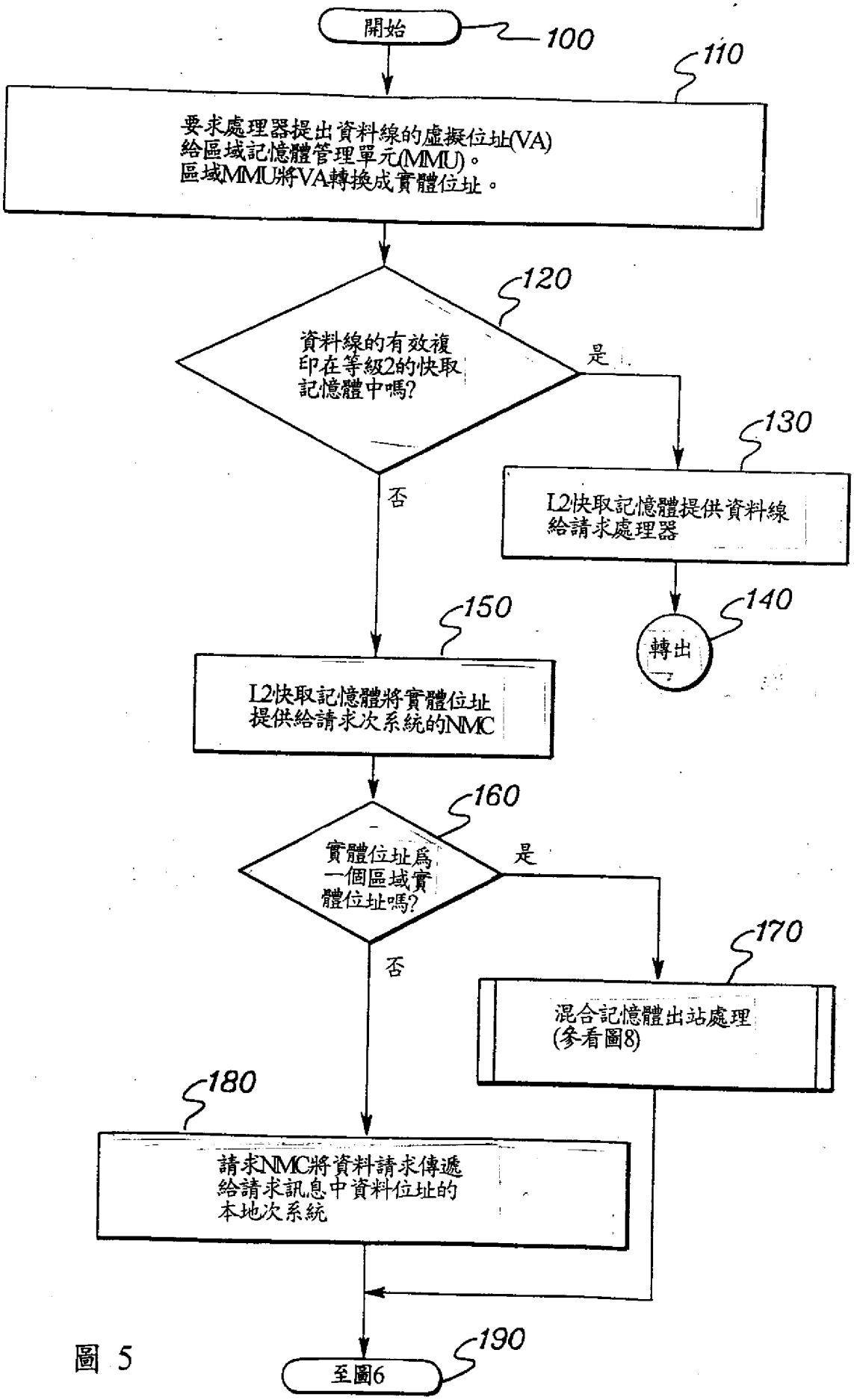
開始
100
110
要求處理器提出資料線的虛擬位址(VA)
給區域記憶體管理單元(MMU)。
區域MMU將VA轉換成實體位址。
120
資料線的有效複
印在等級2的快取
記憶體中嗎?
是
否
130
L2快取記憶體提供資料線
給請求處理器
140
轉出
150
L2快取記憶體將實體位址
提供給請求次系統的NMC
160
實體位址爲
一個區域實
體位址嗎?
是
否
170
混合記憶體出站處理
(參看圖8)
180
請求NMC將資料請求傳遞
給請求訊息中資料位址的
本地次系統
190
至圖6

圖 5

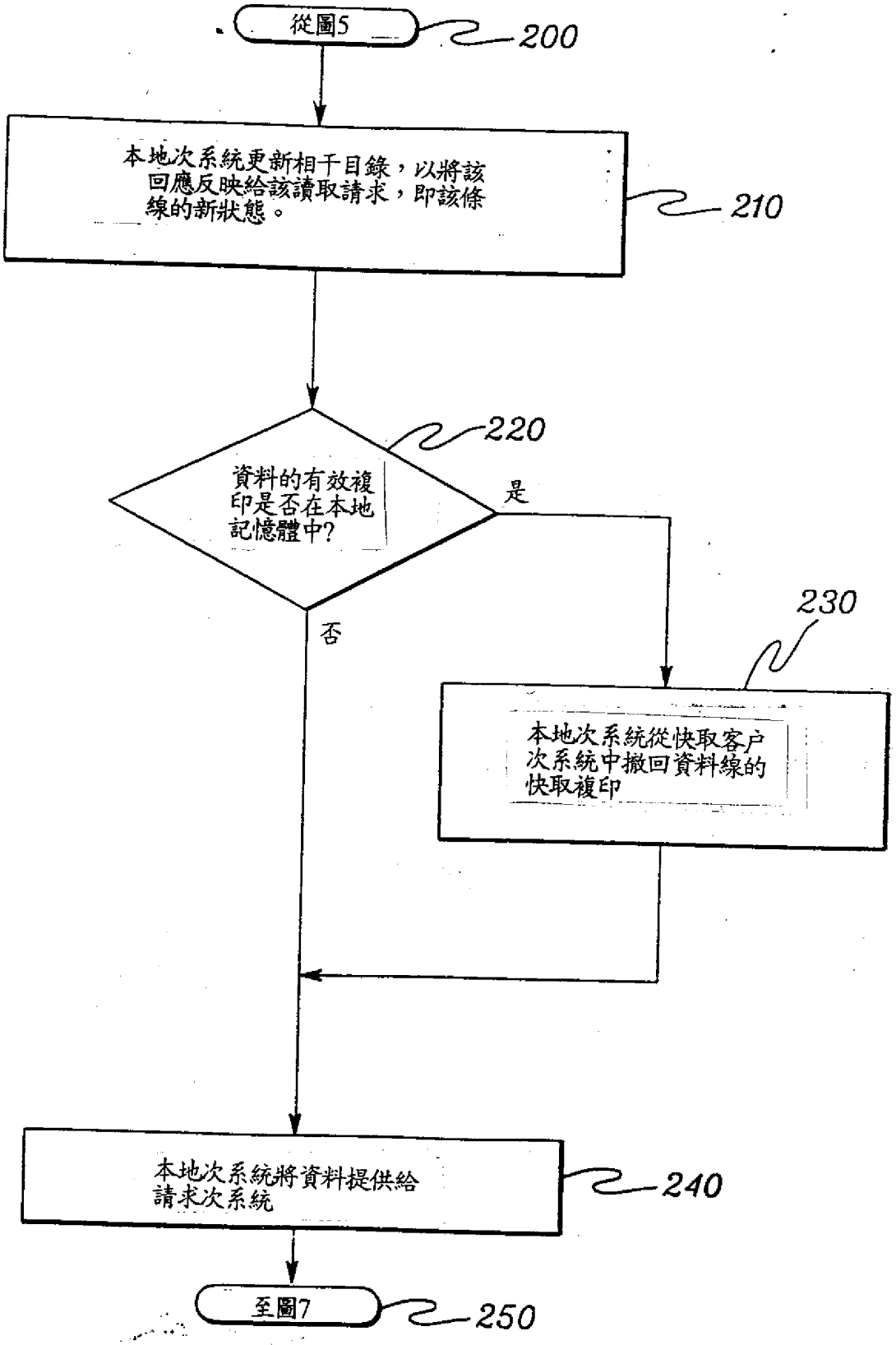
從圖5
200
本地次系統更新相干目錄，以將該回應反映給該讀取請求，即該條線的新狀態。
210
220
資料的有效複印是否在本地記憶體中?
是
否
230
本地次系統從快取客户次系統中撤回資料線的快取複印
本地次系統將資料提供給請求次系統
240
至圖7
250

圖 6

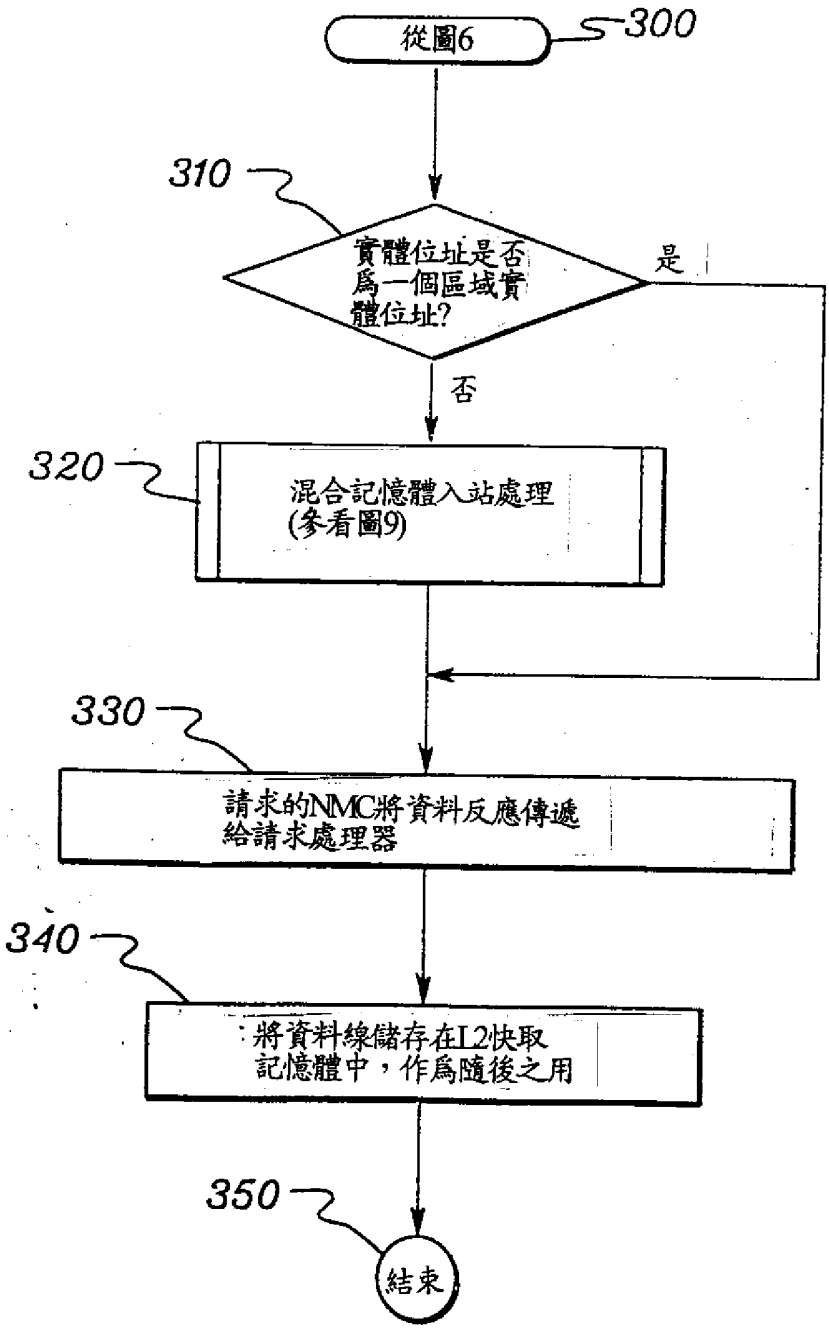
從圖6
300
310
實體位址是否爲一個區域實體位址?
是
否
320
混合記憶體入站處理(參看圖9)
330
請求的NMC將資料反應傳遞給請求處理器
340
將資料線儲存在L2快取記憶體中，作爲隨後之用
350
結束

圖 7

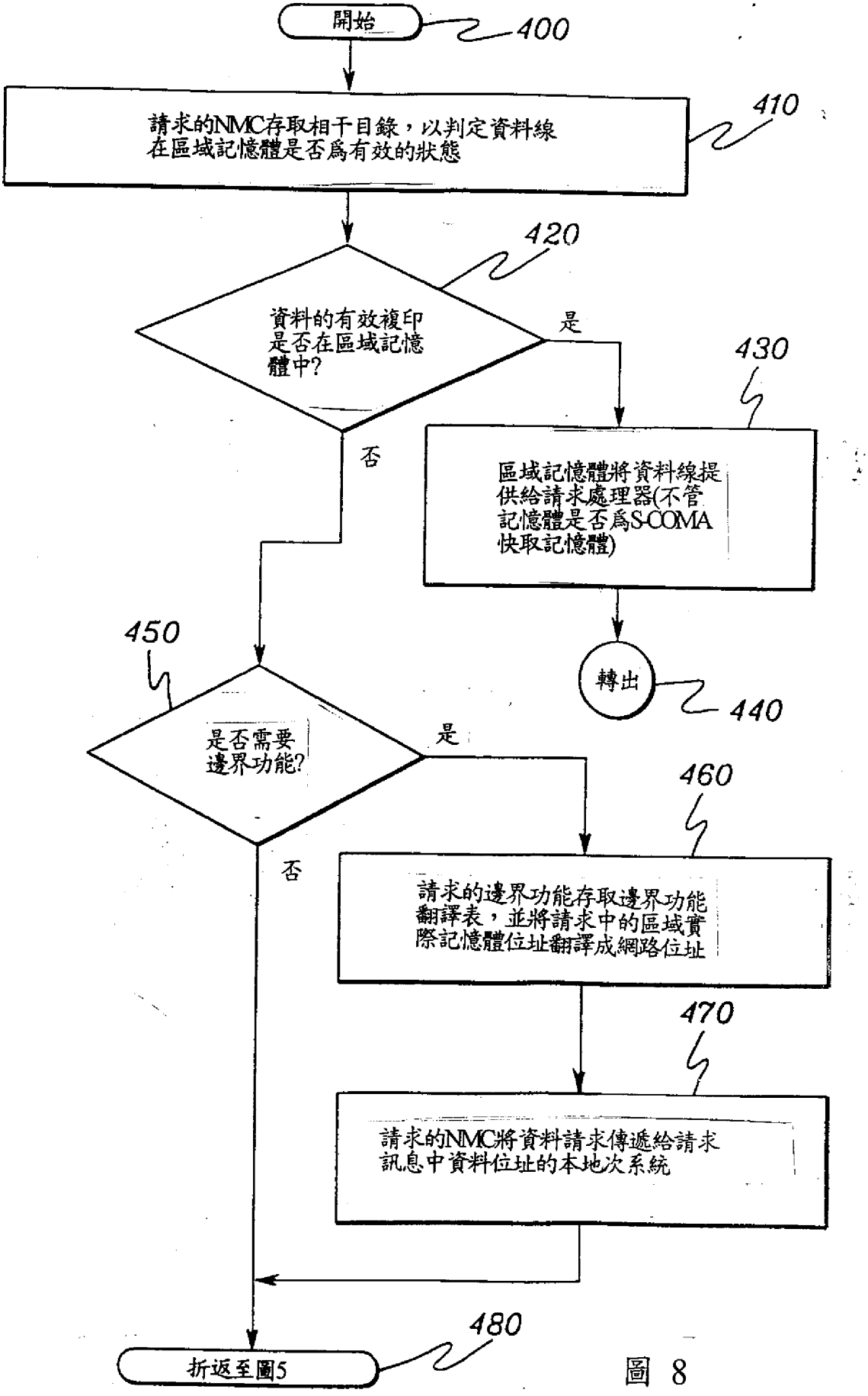
開始
400
請求的NMC存取相干目錄，以判定資料線在區域記憶體是否爲有效的狀態
410
420
資料的有效複印是否在區域記憶體中?
是
否
430
區域記憶體將資料線提供給請求處理器(不管記憶體是否爲S-COMA快取記憶體)
轉出
440
450
是否需要邊界功能?
是
否
460
請求的邊界功能存取邊界功能翻譯表，並將請求中的區域實際記憶體位址翻譯成網路位址
470
請求的NMC將資料請求傳遞給請求訊息中資料位址的本地次系統
480
折返至圖5

圖 8

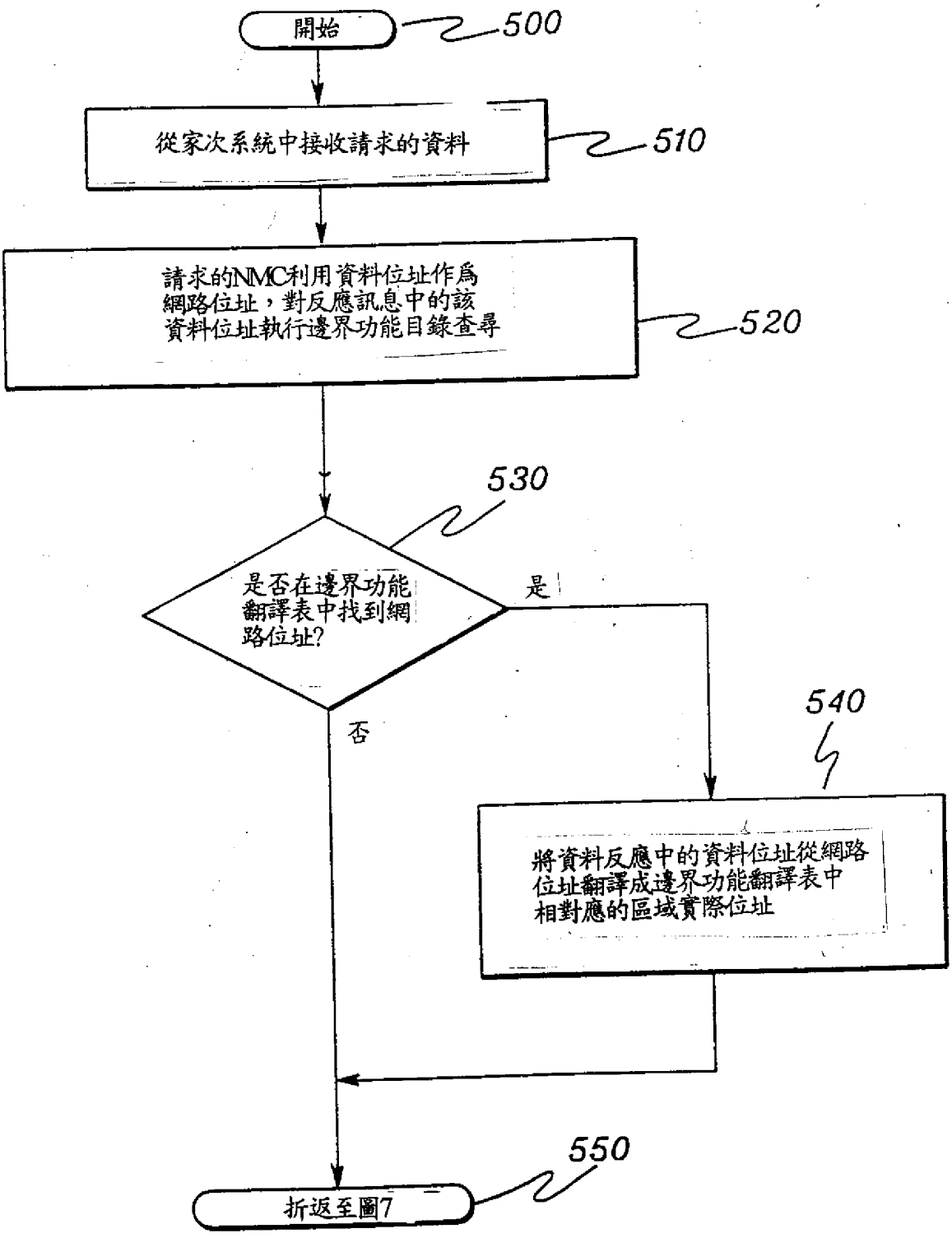
開始
500
從家次系統中接收請求的資料
510
請求的NMC利用資料位址作爲網路位址，對反應訊息中的該資料位址執行邊界功能目錄查尋
520
530
是否在邊界功能翻譯表中找到網路位址?
是
否
540
將資料反應中的資料位址從網路位址翻譯成邊界功能翻譯表中相對應的區域實際位址
550
折返至圖7

圖 9

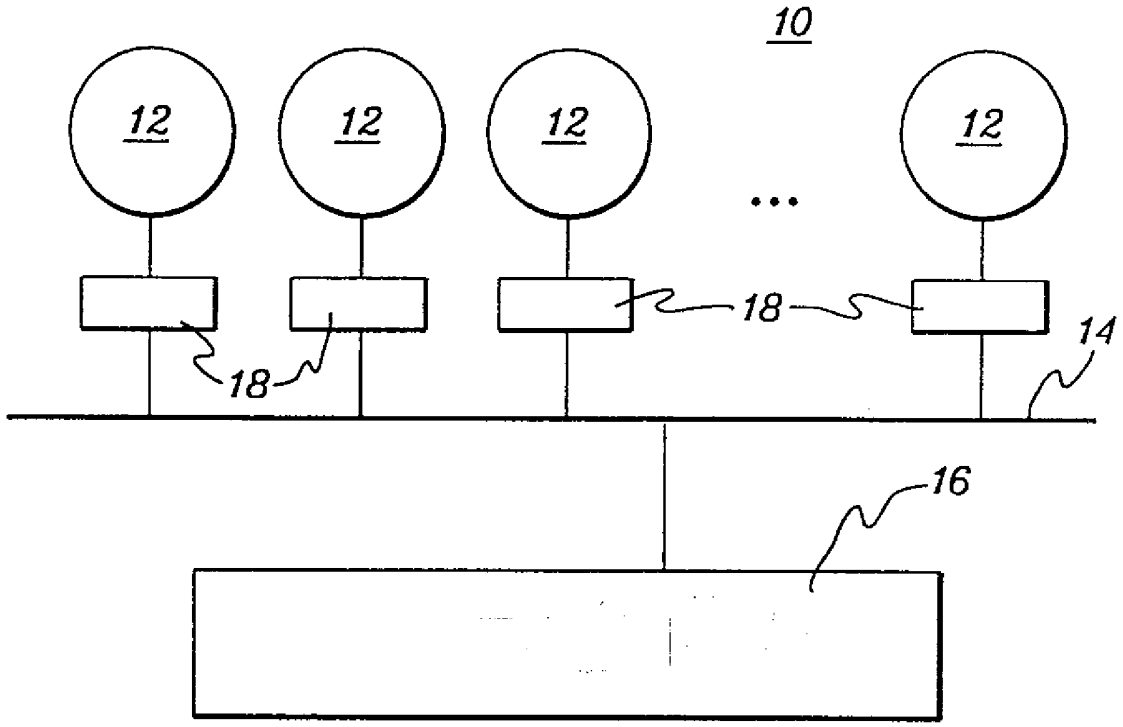

圖 1

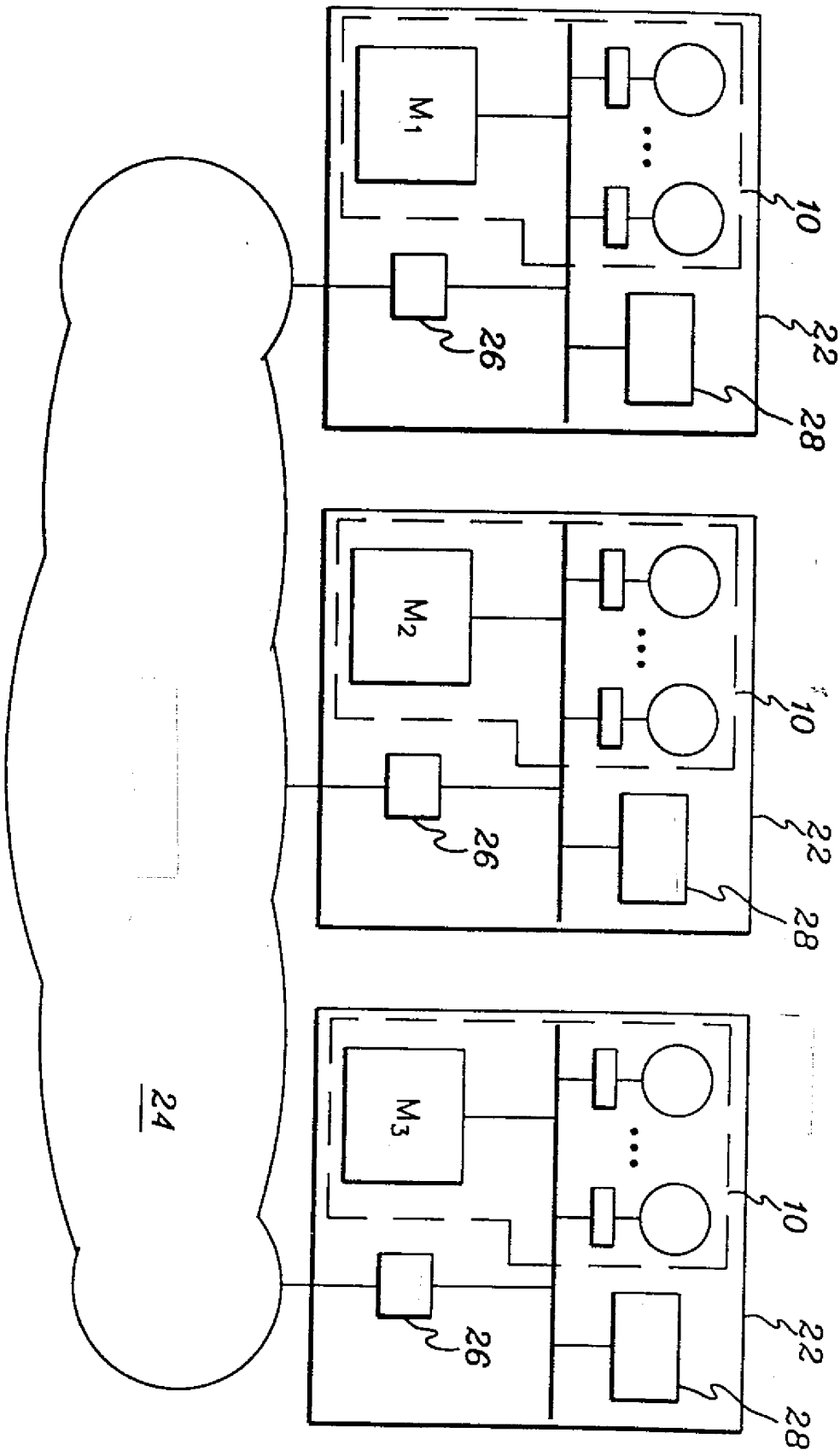

圖 2

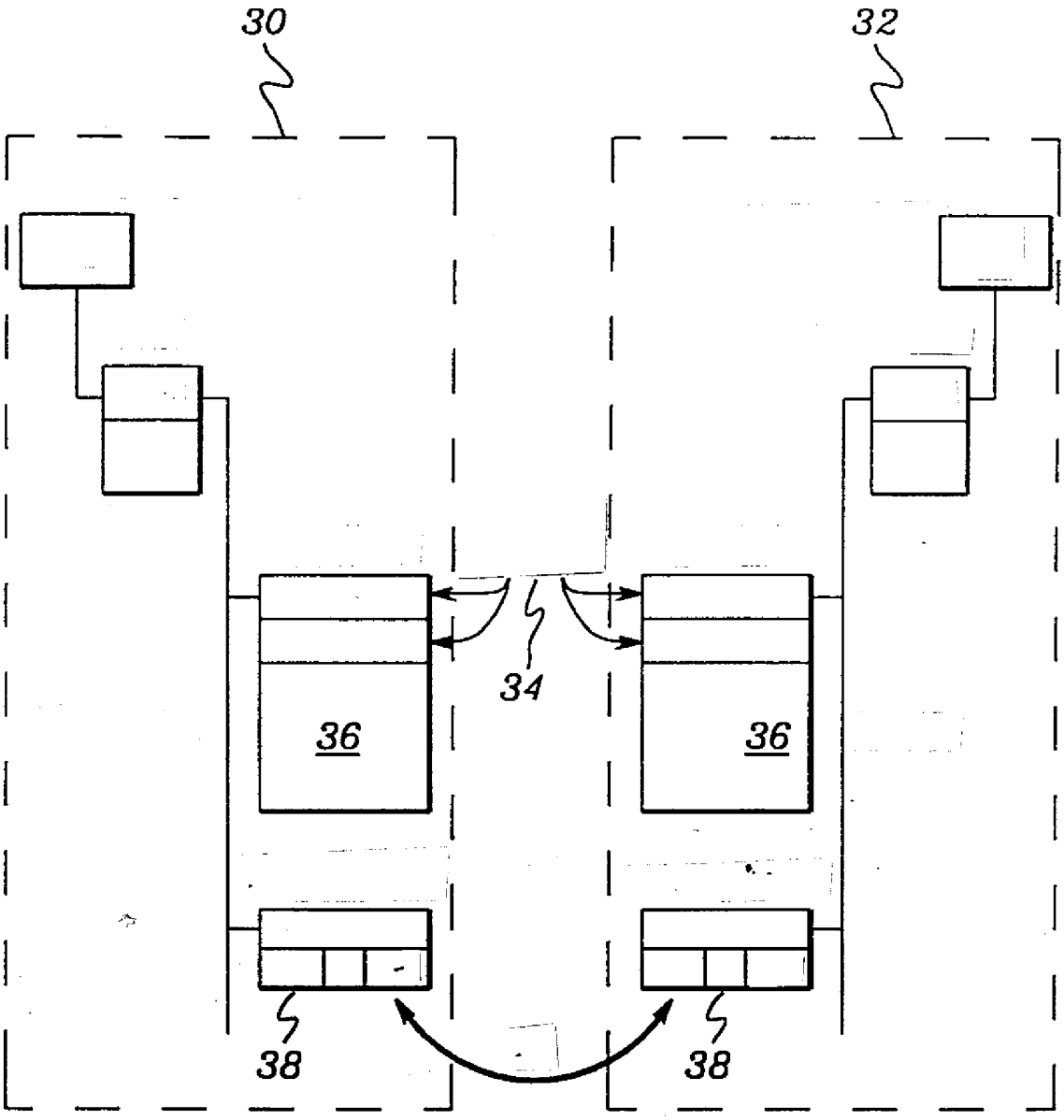

圖 3

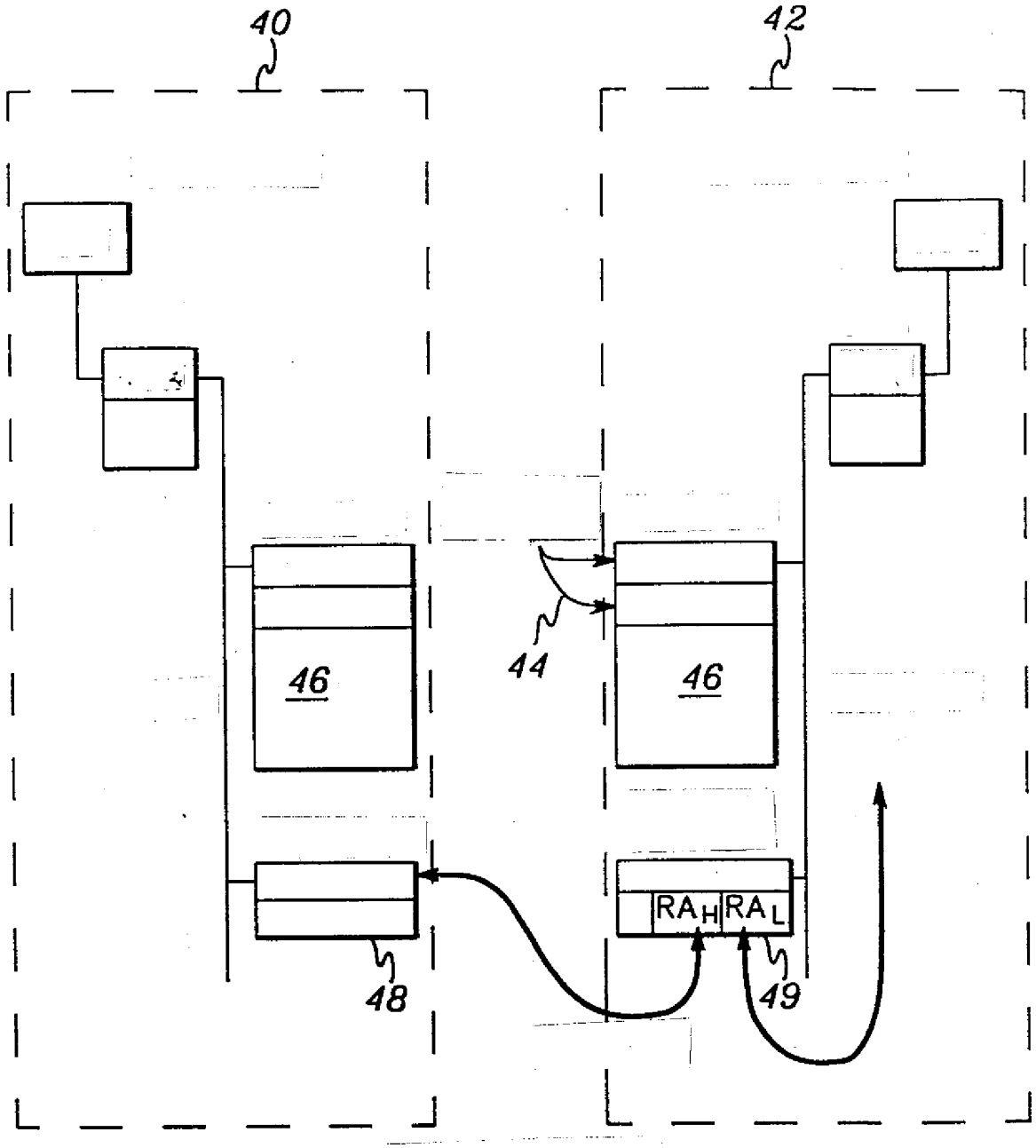

圖 4

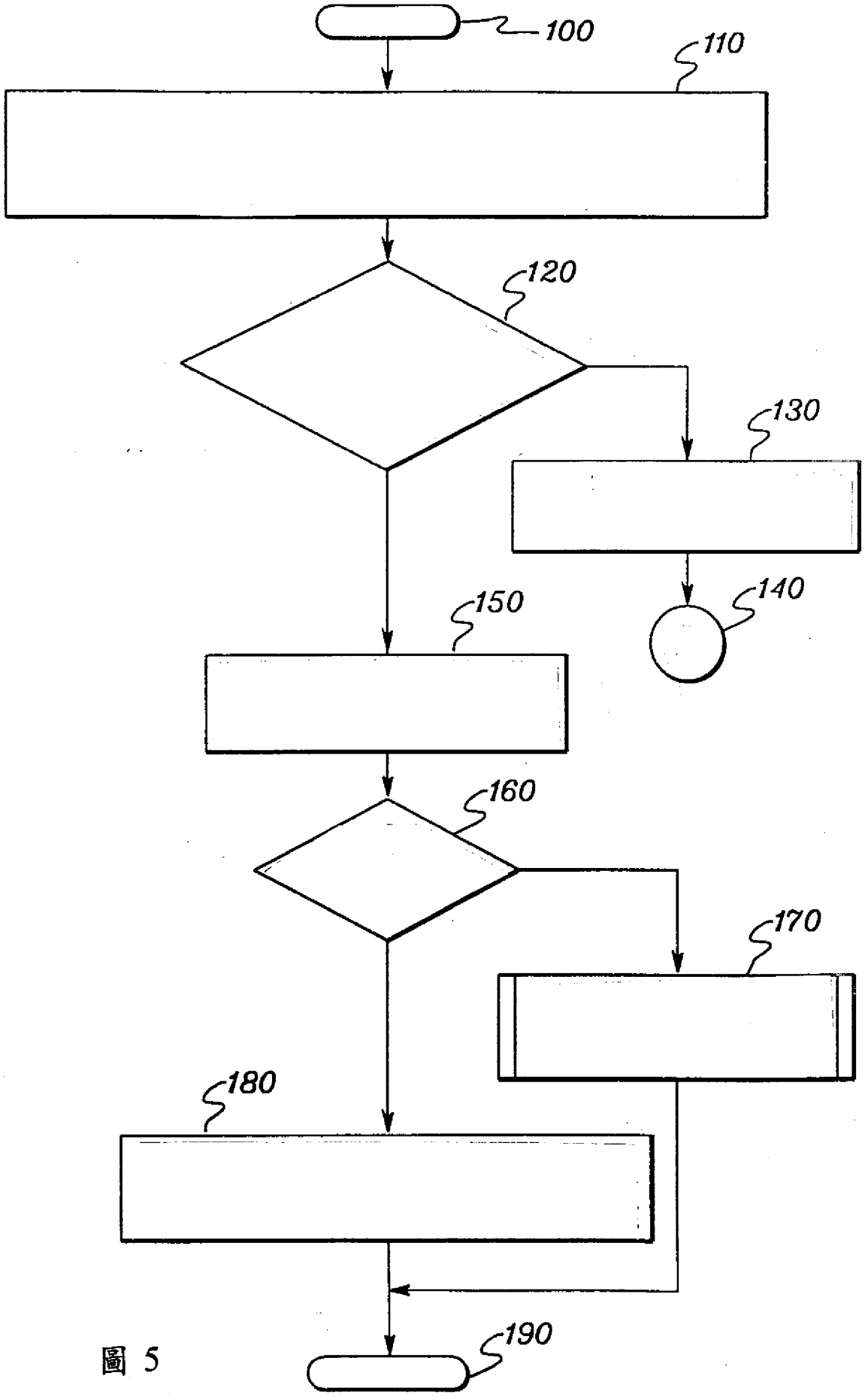

圖 5

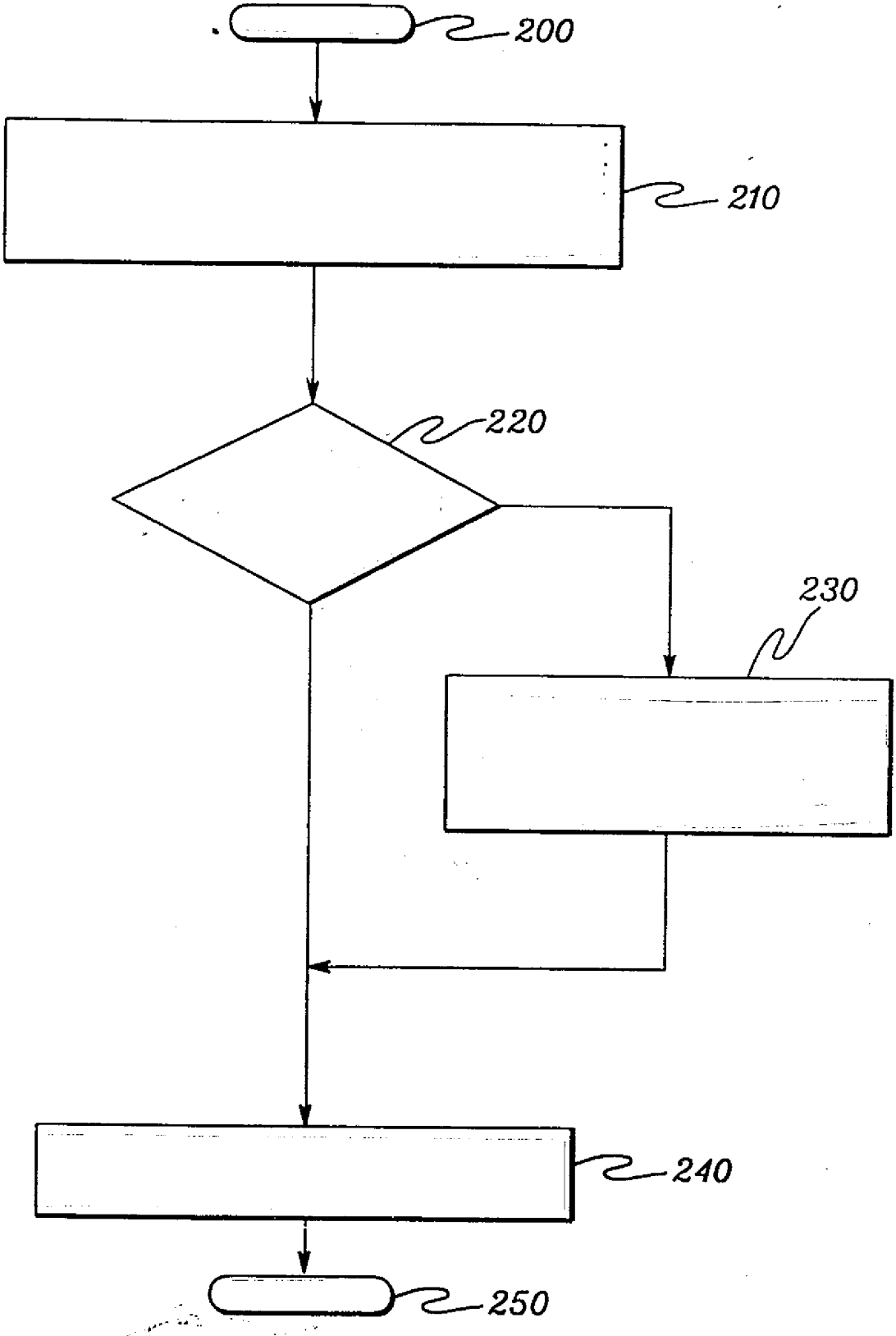

圖 6

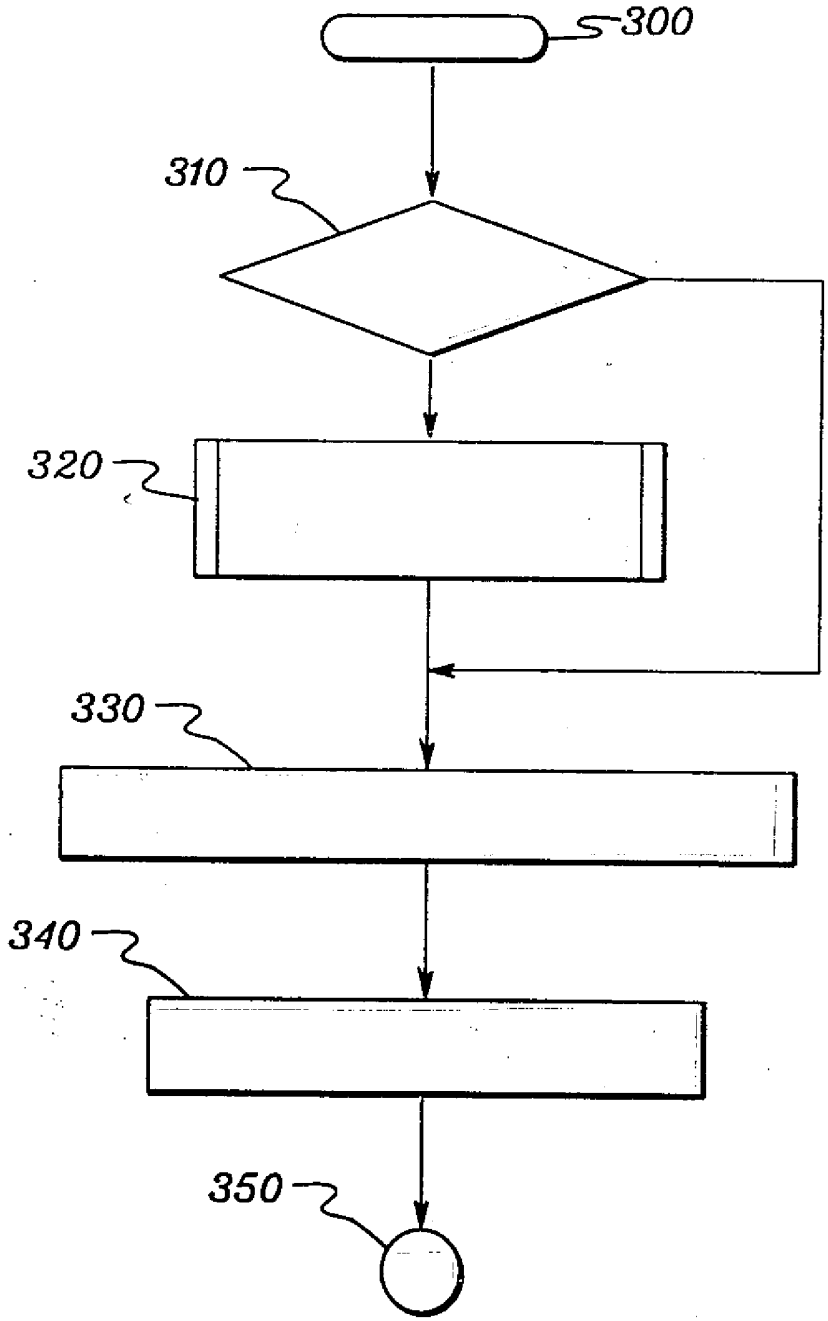

圖 7

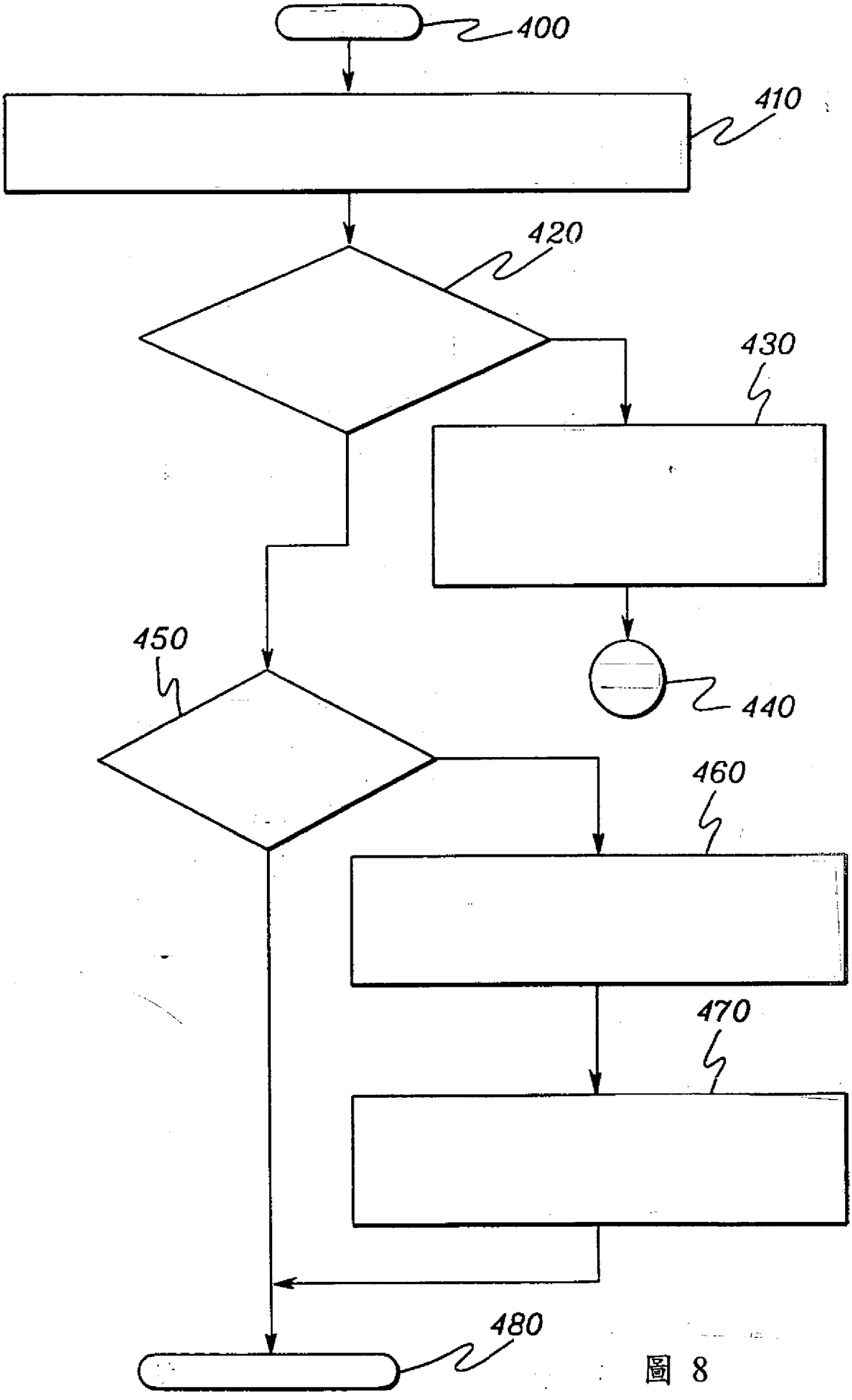

圖 8

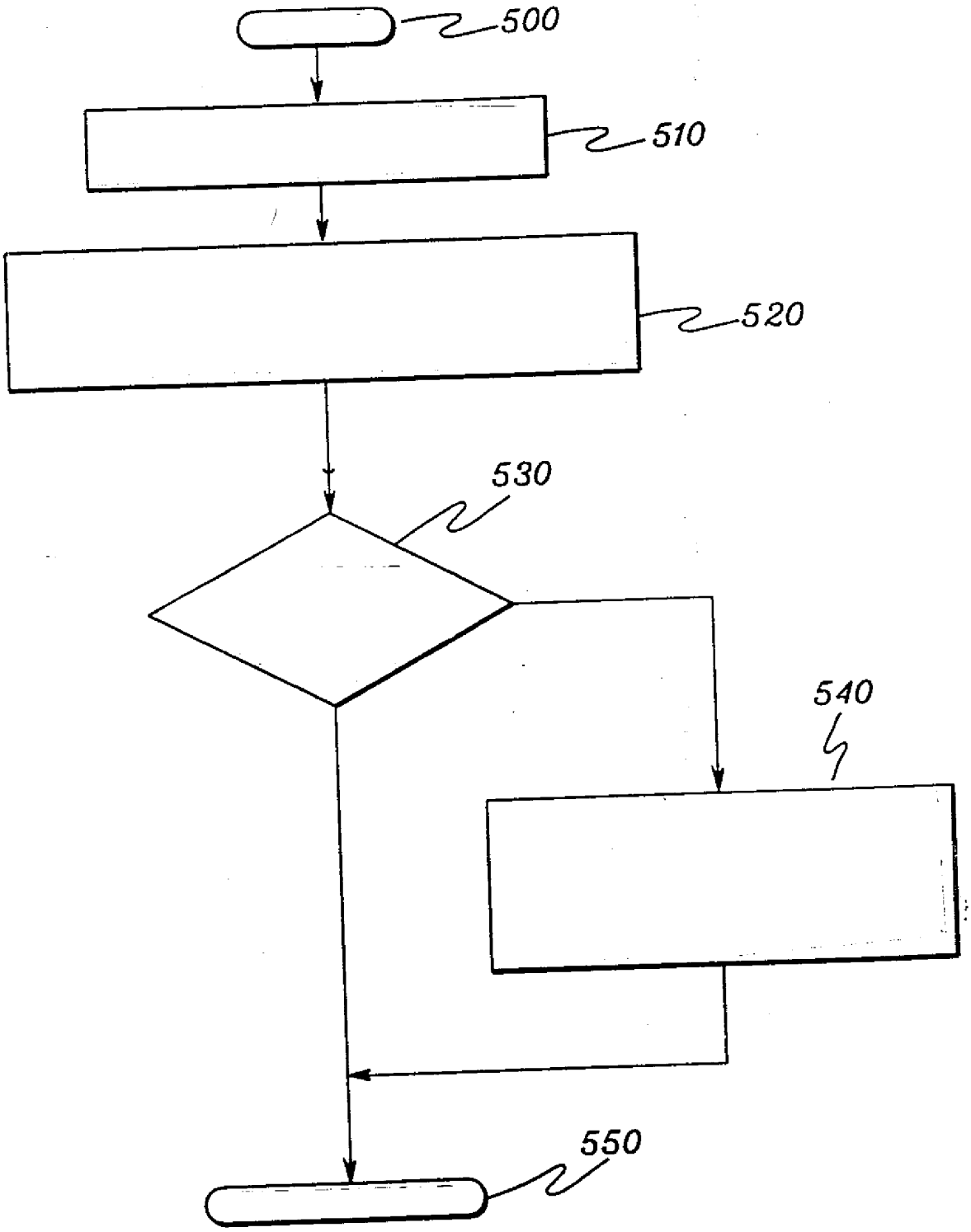

圖 9