



(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 I540436 B

(45)公告日：中華民國 105 (2016) 年 07 月 01 日

(21)申請案號：102103475

(22)申請日：中華民國 102 (2013) 年 01 月 30 日

(51)Int. Cl. : G06F12/16 (2006.01)

G06F12/08 (2006.01)

(30)優先權：2012/01/31 美國

13/362,099

(71)申請人：安華高科技通用 I P (新加坡) 公司 (新加坡) AVAGO TECHNOLOGIES
GENERAL IP (SINGAPORE) PTE. LTD. (SG)

新加坡

(72)發明人：艾許 馬克 ISH, MARK (US) ; 巴德帝尼 安納特 BADERDINNI, ANANT (IN) ;
斯莫登 加里 傑 SMERDON, GARY J. (US)

(74)代理人：陳長文

(56)參考文獻：

TW 200410216A

TW 200535615A

TW 200712919A

US 5586291

US 6279078B1

US 6772288B1

US 2004/0205092A1

審查人員：陳泰龍

申請專利範圍項數：20 項 圖式數：11 共 35 頁

(54)名稱

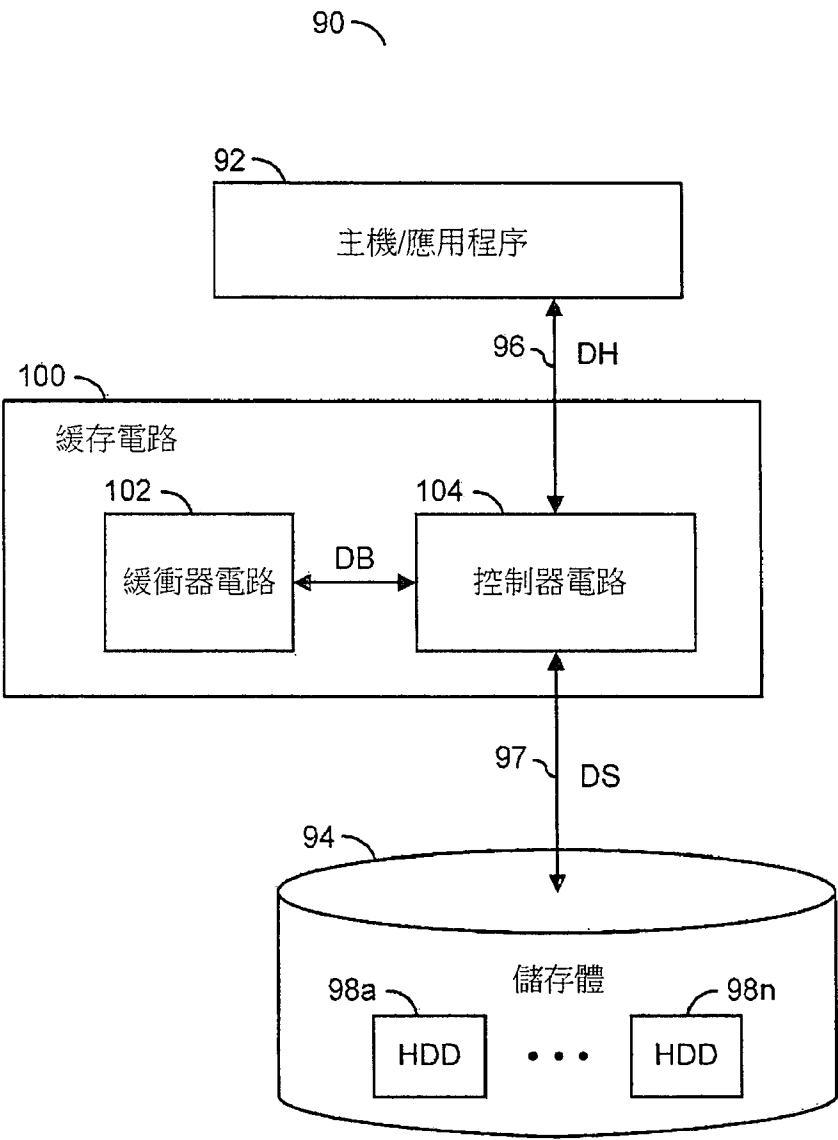
冗餘緩存數據的彈性緩存

ELASTIC CACHE OF REDUNDANT CACHE DATA

(57)摘要

本發明涉及冗餘緩存數據的彈性緩存，具體提供了一種用於冗餘緩存數據的彈性緩存的設備。該設備可具有多個緩衝器和電路。該電路可被配置為(i)從主機接收寫入請求以在儲存儲存體中儲存寫入數據，(ii)基於與寫入請求相關聯的冗餘組織，在緩衝器中分配多個區域，以及(iii)在多個區域中儲存寫入數據，其中，(a)多個區域中的每一個位於緩衝器中的不同一個內，以及(b)響應寫入請求，多個區域被動態鏈接在一起。

指定代表圖：



符號簡單說明：

90 . . . 系統

92 . . . 電路

94 . . . 電路

96 . . . 網路

97 . . . 網路

98a-98n . . . 電路

100 . . . 電路

102 . . . 電路

104 . . . 電路

DB、DH、DS . . . 雙向訊號

圖1

發明摘要

※ 申請案號：102103475

※ 申請日：102.1.30

※IPC 分類：

G06F 12/16

(2006.01)

12/08

(2006.01)

【發明名稱】(中文/英文)

冗餘緩存數據的彈性緩存

ELASTIC CACHE OF REDUNDANT CACHE DATA

【中文】

本發明涉及冗餘緩存數據的彈性緩存，具體提供了一種用於冗餘緩存數據的彈性緩存的設備。該設備可具有多個緩衝器和電路。該電路可被配置為 (i) 從主機接收寫入請求以在儲存儲存體中儲存寫入數據，(ii) 基於與寫入請求相關聯的冗餘組織，在緩衝器中分配多個區域，以及 (iii) 在多個區域中儲存寫入數據，其中，(a) 多個區域中的每一個位於緩衝器中的不同一個內，以及 (b) 響應寫入請求，多個區域被動態鏈接在一起。

【英文】

【代表圖】

【本案指定代表圖】：圖 1。

【本代表圖之符號簡單說明】：

90	系統
92	電路
94	電路
96	網路
97	網路
98a-98n	電路
100	電路
102	電路
104	電路
DB、DH、DS	雙向訊號

【本案若有化學式時，請揭示最能顯示發明特徵的化學式】：

發明專利說明書

(本說明書格式、順序，請勿任意更動)

【發明名稱】(中文/英文)

冗餘緩存數據的彈性緩存

ELASTIC CACHE OF REDUNDANT CACHE DATA

【技術領域】

本發明總體上涉及儲存裝置緩存，更具體地，涉及用於實施冗餘緩存數據的彈性緩存的方法和/或設備。

【先前技術】

在常規儲存系統中，整個系統和應用程序的性能通過緩存經常存取的數據來提高。在許多應用中，若緩存使用回寫模式，則工作負荷受益於此且性能得到提高。保持在回寫模式中的緩存數據有時通過提供冗餘來保護以防止數據丟失。冗餘通常利用緩存控制器中的兩個閃存模組來實施。兩個閃存模組利用鏡像永久配對，有效減少了將緩存數據保持在單個閃存模組中的有效空間。假定僅有某些主機數據受益於冗餘，則使閃存模組中有效空間被低效使用。

將期望實施冗餘緩存數據的彈性緩存。

【發明內容】

本發明涉及用於冗餘緩存數據的彈性緩存的設備。該裝置可具有多個緩衝器和電路。該電路可被配置為(i)從主機接收寫入請求以在儲存體中儲存寫入數據，(ii)基於與寫入請求相關聯的冗餘組織，在緩衝器中分配多個區域(延伸區，extent)，以及(iii)在多個區域中儲存寫入數據，其中，(a)多個區域中的每一個位於緩衝器中的不同一個內，以及(b)響應寫入請求，多個區域被動態鏈接在一起。

本發明的目標、特徵和優勢包括提供了一種用於實施冗餘緩存數據的彈性緩存的方法和/或設備，該冗餘緩存數據可以(i)提

供附加有效緩存空間，(ii) 與常規技術相比，提供了對給定數量的緩存介質的更好緩存覆蓋，(iii) 使用現有區域結構，(iv) 使用現有管理基礎設施，(v) 為緩存冗餘實施各種 RAID 級，(vi) 與常規技術相比，減少了對於給定數量的主機輸入/輸出活動的緩存寫入數量，(vii) 增強了對於給定主機輸入/輸出活動的閃存緩衝器的壽命，和/或 (viii) 實施固態驅動器作為用於緩存數據的大緩衝器。

【圖式簡單說明】

根據以下詳細描述和所附申請專利範圍以及附圖，本發明的這些和其他目標、特徵和優勢將變得顯而易見，其中：

圖 1 是系統的實例性實施的框圖；

圖 2 是根據本發明優選實施方式的系統中緩存電路的實例性實施的框圖；

圖 3 是區域的實例性結構的示意圖；

圖 4 是實例性緩存數據的示意圖；

圖 5 是區域的幾個實例性組的示意圖；

圖 6 是具有彈性鏡像緩存的實例性緩存組織的示意圖；

圖 7 是用於服務寫入請求的實例性方法的流程圖；

圖 8 是用於服務讀取請求的實例性方法的流程圖；

圖 9 是用於分配區域的實例性方法的流程圖；

圖 10 是用於去分配區域的實例性方法的流程圖；以及

圖 11 是示出在緩存電路中來自彈性鏡像髒寫 (elastically mirroring dirty writes) 的空間益處的圖表示意圖。

【實施方式】

本發明的某些實施方式可彈性使用有效緩存空間，以提供寫入緩存空間的部分的冗餘。彈性方式通常利用增加緩存空間整體有效部分來緩存讀取數據和/或未保護（例如，非冗餘）寫入數據的智慧管理技術來實施。此外，彈性方法可被模型化為示出其他

益處，諸如聚集（例如，某些數據可以每個區域為基礎在單獨物理裝置上鏡像，使得若儲存和緩存通過多個節點（服務器）共享，則被鏡像區域的每個部分可駐留在單獨節點上）以及策略選擇（例如，某些儲存磁碟使用受保護的（或冗餘的）緩存數據和基於選擇策略，某些不保護緩存數據）。

參照圖 1，示出了系統 90 的實例性實施的框圖。該系統（或設備）90 通常包括塊（或電路）92、塊（或電路）94、網路（或總線）96、網路（或總線）97 和塊（或電路）100。電路 94 通常包括多個塊（或電路）98a-98n。電路 100 通常包括塊（或電路）102 和塊（或電路）104。電路 92-104 可表示可被實施為硬體、軟體、硬體和軟體的結合或其他實施的模組和/或塊。

雙向訊號（例如，DH）可經由網路 96 在電路 92 與電路 100 之間交換。訊號 DH 可被用於在電路 92 與電路 100 之間傳送數據（例如，讀取數據和/或寫入數據）、請求和狀態訊息。雙向訊號（例如，DB）可在電路 102 與電路 104 之間交換。訊號 DB 可在電路 102 與電路 104 之間傳遞緩存數據和元數據。雙向訊號（例如，DS）可經由網路 97 在電路 94 與電路 100 之間交換。訊號 DS 可在電路 100 與電路 94 之間傳送數據和關聯命令（例如，讀取命令，寫入命令和配置命令）。

電路 92 可實施一個或多個主機電路（或服務器）和/或一個或多個軟體應用程序。電路 92 通常可操作以通過訊號 DH 和 DS 經由電路 100 和網路 96 從電路 94 讀取數據和/或向電路 94 寫入數據。電路 92 還可操作以通過訊號 DH 沿相應地址訊息（例如，邏輯塊地址（LBAs））向電路 100 提出請求（例如，讀取請求和寫入請求）。此外，電路 92 可操作以通過訊號 DH 向電路 100 發出命令（例如，預取、刷新、配置）。訊號 DH 還可由電路 100 控制以將來自電路 100 的狀態訊息（例如，緩存命中（hit），緩存錯失）回傳到電路 92。

電路 94 可實施一個或多個儲存體 (storage volume)。電路 94 響應寫入命令通常可操作以儲存從電路 100 經由訊號 DS 接收到的數據。電路 94 響應讀取命令還可操作以經由訊號 DS 向電路 100 提供數據。儲存體可被實施為邏輯體、虛擬體和/或物理體。電路 94 可以是儲存區域網路 (SAN)、網路附加儲存 (NAS) 和/或磁碟陣列子系統 (DAS) 架構的一部分。

網路 96 可實施一個或多個數位通訊網路和/或總線。網路 96 通常可操作以提供電路 92 與電路 100 之間的通訊。網路 96 的實施可包括但不限於網際網路、乙太網、光纖網路、無線網路、有線網路、射頻通訊網路和/或底板總線中的一個或多個。

網路 97 可實施一個或多個數位通訊網路和/或總線。網路 97 通常可操作以提供電路 94 與電路 100 之間的通訊。網路 97 的實施可包括但不限於網際網路、乙太網、光纖網路、無線網路、有線網路、射頻通訊網路和/或底板總線中的一個或多個。

每個電路 98a-98n 可實施大容量儲存驅動器。在某些實施方式中，每個電路 98a-98n 可實施硬碟機 (例如，HDD)。電路 98a-98n 通常可操作以經由訊號 DS 從電路 100 儲存數據和向電路 100 提供數據。電路 98a-98n 可被佈置和操作為一個或多個儲存體。在某些實施方式中，電路 98a-98n 可被組織和操作為一個或多個冗餘和獨立的磁碟 (例如，RAID) 配置。

RAID 組織可包括但不限於 RAID0、RAID1、RAID2、RAID3、RAID4、RAID5、RAID6、RAID7、RAID10、RAID53 和/或 RAID0+1 的組織。可實施其他 RAID 組織級別以滿足特定應用程序的標準。此外，多個不同 RAID 組織可由電路 98a-98n 同時表示。

電路 100 可實施緩存電路。電路 100 通常可操作以緩存在電路 92 與電路 94 之間被交換的數據。對於從電路 92 向電路 94 移動的寫入數據，緩存可實施回寫 (或後寫) 策略。髒寫入數據可利用在多個介質之間傳播的冗餘組織緩存到電路 100，冗餘組織和

傳播可被配置為使得在單個介質失效導致失效介質中部分或全部的寫入數據變為不可存取和/或丟失之後，保持對該寫入數據的存取。電路 100 可實施位於電路 92、電路 94 中或在中間點處的緩存。在某些實施方式中，電路 100 可被實施為一個或多個集成電路。

電路 102 可實施多個緩衝器電路。電路 102 可操作以儲存在電路 92 與電路 94 之間傳送的緩存（或緩衝）數據。電路 102 還可操作以儲存與緩存數據相關聯的元數據。緩存數據可在區域中佈置。每個區域可表示為特定數據文件或軟體程序保留的儲存空間的連續塊。元數據通常包括用於管理區域的訊息。在某些實施方式中，電路 102 可整個被實施在電路 100 的內部（如圖所示）、整個在電路 100 的外部、部分在電路 100 的內部和外部、或跨多個節點（或服務器）而聚集。

電路 104 可實施控制器裝置。電路 104 可操作以從電路 92 接收在電路 94 中儲存寫入數據的寫入請求，基於與寫入請求相關聯的冗餘組織在電路 102 的多個緩衝器中分配多個區域，並在多個區域中儲存寫入數據。多個區域中的每個區域通常位於不同緩衝器中。響應寫入請求，多個區域可被動態鏈接在一起。響應從緩衝器向電路 94 複製寫入數據，電路 104 還可操作以解鏈多個區域中的至少一個。在某些實施方式中，電路 104 可被實施為一個或多個集成電路。

參照圖 2，示出了根據本發明優選實施方式的電路 100 的實例性實施的框圖。電路 102 通常包括多個塊（或電路）106a-106n。電路 104 通常包括塊（或模組）108、塊（或模組）110 和塊（或模組）112。電路 106a-106n 可表示可被實施為硬體、軟體、硬體和軟體的結合或其他實施的模組和/或塊。模組 108-112 可被實施為可執行控制電路 104 的固件（或軟體或程序指令或者代碼）。

訊號 DH 可通過介面 114 由電路 104 發送和接收。訊號 DS 可通過介面 116 由電路 104 發送和接收。訊號 DB 可被實施為多個訊

號（例如，DBa-DBn）。每個訊號 DBa-DBn 通常建立電路 104 的介面 118a-118n 與電路 106a-106n 的相應介面 120a-120n 之間的通訊。

每個電路 106a-106n 可實施緩衝器電路。每個電路 106a-106n 可操作以儲存緩存數據和關聯元數據。在某些實施方式中，每個電路 106a-106n 可被實施為固態驅動器（例如，SSD）。固態驅動器的常見大小範圍可從 1TB（terabyte）到 2TB。在其他實施方式中，每個電路 106a-106n 可被實施為閃存驅動器。閃存的常見大小範圍可從 1GB（gigabyte）到 32GB。在又一實施方式中，每個電路 106a-106n 可被實施為雙倍數據速率（例如，DDR）儲存器電路。DDR 儲存器的常見大小範圍可從 1GB 到 64GB。可實施其他大小以滿足特定應用程序的標準。電路 102 中的電路 106a-106n 的數量可多於、等於或少於電路 94 中的電路 98a-98n 的數量。

模組 108 可實施區域控制器。模組 108 可操作以在電路 102 中分配和去分配用於服務從電路 92 接收到的請求（例如，讀取請求和/或寫入請求）的一個或多個區域。模組 108 還可操作以改變在與電路 102 中的每個區域相關聯的元數據中的參數來創建分組空間（例如，讀取空間和/或寫入空間）。分組可通過將區域鏈接在一起以形成可在冗餘保護組織中容納緩存數據的更大儲存器空間（或超級區域）來實現。將區域分配（或鏈接）到分組空間通常包括改變關聯區域中的每個參數以指向分組中的下一區域。指針可在從一個或多個初始區域到一個或多個最終區域的鏈路中繼續。最終區域的參數可指回初始區域以閉合鏈接。模組 108 還可改變元數據中的參數以將分組的空間向後分解成單獨區域和/或更小分組空間。在緩存數據被緩衝在非冗餘組織中的情況下，模組 108 可將一個或多個區域的參數改變為分配（空）值，以指示當前該區域被用於緩衝緩存數據。當區域不再用於緩衝緩存數據時，模組 108 可將關聯參數改變為未分配（或未使用）值。

模組 110 可實施緩存控制器。模組 110 可操作以執行標準緩存操作和標準緩存策略。緩存操作通常包括但不限於將數據寫入電路 102，從電路 102 讀取緩存（或緩衝）數據，將元數據寫入電路 102，從電路 102 讀取元數據，搜索緩存數據的標籤，當發現匹配標籤時指示緩存命中，當發現不匹配標籤時指示緩存錯失，跟蹤電路 102 中的髒寫入數據，將髒寫入數據刷新（例如，複製）到電路 94，跟蹤有效和無效緩存行，跟蹤有效和無效緩存字等。模組 110 還可操作以實施關於緩存數據的保留和去除的一個或多個緩存策略。緩存策略可包括但不限於用於髒寫入數據的回寫模式，用於某些寫入數據類型的連續寫入模式，某些寫入數據類型的冗餘（或保護）緩衝，其他寫入數據類型的非冗餘（或非保護）緩衝，用於髒寫入數據的一個或多個刷新策略，清除電路 102 內的空間的一個或多個去除策略等。

模組 112 可實施 RAID 控制器。模組 112 通常可操作以基於分配給數據的 RAID 策略在各種 RAID 配置中組織（或佈置或者格式化）緩衝在電路 102 和儲存在電路 94 中的數據。

當應保護的寫入數據從電路 92 接收時，模組 112 可將寫入數據重新安排到 RAID 組織間的給定組織中。重新安排的寫入數據可基於寫入數據的大小（例如，寫入數據跨過多少區域邊界）和對應於寫入請求的 RAID 組織（例如，RAID 0、RAID 3、RAID 5 等）隨後被儲存在兩個或更多個電路 106a-106n 中。可實施其他數據保護形式以滿足特定應用程序的標準。

當受保護的緩存數據從電路 102 向電路 92 提供時，模組 112 可操作以檢查數據錯誤以及如果可用則修正錯誤。若數據已被條形化（stripe），則模組 112 還可在將被請求數據提供給電路 92 之前，將柄（stipe）收集回相應的塊。

可在將數據保持在 RAID 組織中的同時執行 RAID 保護的寫入數據從電路 102 向電路 94 的傳送。例如，若數據在被鏡像 RAID 1

組織中，則數據可經由訊號 DB 和 DS 從兩個電路 106a-106n（例如，106a 和 106b）被傳送到相應的兩個電路 98a-98n（例如，98f 和 98g）。在另一實例中，若數據在利用帶有兩個條區域的奇偶區域的 RAID 3 組織中，則數據可經由網路 97 從三個電路 106a-106n（例如，106a、106b 和 106c）被傳送到相應的三個電路 98a-98n（例如，98d、98g 和 98k）。

某些 RAID 保護儲存數據從電路 94 向電路 102 的傳送（例如，讀取數據）可在從電路 94 到模組 112 的 RAID 組織和從模組 112 到電路 102 的非冗餘組織中執行。例如，儲存在 RAID 3 配置中的數據可從電路 94 被複製到模組 112 作為三個並行傳送（例如，兩個區域和奇偶區域）。模組 112 可與正常一樣執行錯誤檢查和錯誤糾正，並將被儲存數據的單個非冗餘副本傳送到電路 102。

參照圖 3，示出了區域 140 的實例性結構的示意圖。在電路 102 中的緩存數據和在電路 94 中的儲存數據可以區域（或窗口）的形式組織。每個區域通常表示數據連續塊的集合，該數據連續塊作為用於分配、去分配、跟蹤和管理數據的單個單元來處理。實例性區域 140 通常包括多個數據線 142a-142n。每個數據線 142a-142n 可包括多個數據塊 144a-144n。每個數據塊 144a-144n 通常包括多個字元、字或其他數據單元。此外，每個數據塊 144a-144n 可具有相應邏輯塊地址（例如，LBA A 至 LBA N）。可實施其他區域組織以滿足特定應用程序的標準。

參照圖 4，示出了實例性緩存數據的示意圖。該實例示出了用於冗餘的 RAID 1 組織。為簡化起見，僅示出了電路 106a-106n 中的兩個電路（例如，106a-106b）。每個電路 106a-106b 可被組織為一個或多個區域和相應元數據。某些區域（例如，寫入區域）可承載未交給電路 94 的髒寫入數據。其他區域（例如，讀取區域）可基於某些基本緩存標準承載來自電路 94 的緩存數據。在電路 106a 或 106b 失效的情況下，承載髒寫入數據的寫入區域可利用冗

餘保護來緩衝。承載緩存讀取的所有其他區域可在沒有冗餘保護的情況下緩衝，因為在電路 106a 或 106b 失效時丟失的緩存數據可在失效發生後從電路 94 向工作電路 106a 和 106b 重新填充。

參照圖 5，示出了在電路 102 內的幾個實例性區域組的示意圖。在所示實例中，電路 102 可包括電路 106a-106d。可實施更少或更多電路 106a-106n 以滿足特定應用程序的標準。如圖所示，多個區域組、多個組類型（例如，RAID 1、RAID 3、RAID 4、RAID 5、RAID 10、RAID 0+1 等）和/或多個未分組區域（例如，讀取和寫入區域）可同時存在於電路 102 中。

每個區域和相應元數據可利用現有方法（或技術）管理。具體地，每個區域可具有用於跟蹤和識別區域的唯一區域識別碼。對於每個主機輸入/輸出請求，LBA 可被映射到區域以定位到正確區域，從而服務主機請求。參數（或字段）可被添加到每個區域的元數據以跟蹤用於冗餘保護的另一區域識別碼。

當在單個區域內的寫入數據由電路 100 從電路 92 接收時，兩個或更多個區域可被分配給一組。相應參數可被更新以在環路中指向（或鏈接）彼此。將區域分配給組可被配置為使得單個組中的每個區域被分配在不同電路 106a-106n 中。

在寫入數據跨多個區域和/或橫跨區域邊界的某些實施方式中，兩個或更多個區域的集合可被處理為大區域。兩個或更多個大區域可被分配給一組，且相應參數可被更新以在環路中指向彼此。大區域中的每個集合可在不同電路 106a-106n 中分配。例如，RAID 1 組 150 通常包括電路 106b 中兩個區域的集合和電路 106c 中兩個區域的另一集合。電路 106b 中的兩個區域的每個參數可指向電路 106c 中的相應區域。類似地，電路 106c 中的兩個區域的每個參數可指向電路 106b 中的相應區域。

在寫入數據跨多個區域和/或橫跨區域邊界的其他實施方式中，鄰接區域可位於不同電路 106a-106n 中且被每個冗餘組織鏈

接。例如，在 RAID 3 組 152 中的數據可具有緩衝在電路 106b 中的初始區域，緩衝在電路 106c 中的下一區域和緩衝在電路 106d 中的奇偶區域。

基於緩存策略，某些寫入數據可被緩衝在非冗餘組織 154 內的電路 102 中。在由於電路 102 失效而導致數據丟失的情況下，這種未保護寫入數據可由電路 92 取代。未保護寫入數據還可具有適於潛在數據丟失的低重要性。

對於由於讀取失誤而執行的每個緩存填充操作，可根據要重新填充電路 102 中的數據的位置和數量，一個或多個區域的集合可被分配給一個或多個電路 106a-106n。被分配集合中的每個參數可被改變（或程序化）為指示相應區域被分配到非冗餘組織中的值（例如，空值或分配值）。

在區域組中的髒寫入數據被刷新到電路 94 之後，組中的區域可不被分配以獲得有效空間用於進一步分配。可通過更新相鄰區域的參數而不分配區域，以將未分配區域從組中解鏈。未分配區域可由用於保護寫入數據的冗餘組織來確定。例如，由於兩個被鏡像區域包括所有寫入數據，所以具有 RAID 1 組織的區域組可去分配被鏡像區域中的任一個。具有 RAID 3 組織的區域組可解鏈被分配以緩衝奇偶數據的區域。緩衝實際數據的剩餘區域可被保留在組中。

當電路 106a-106n 失效時，標準恢復技術可被用於服務來自倖存區域的主機請求。例如，在失效電路 106a-106n（例如，電路 106a）中丟失塊的具有 RAID 5 組織的區域組 156 可使用倖存塊和在倖存電路 106a-106n（例如，106b 和 106d）中的奇偶塊來重構丟失的塊。重構的塊可隨後用於滿足讀取/寫入請求。

參照圖 6，示出了帶有彈性鏡像緩存的實例性緩存組織 160 的示意圖。該實例通常示出了利用兩個電路 106a-106n（例如，106a 和 106b）的鏡像組織。區域 1 和區域 10 兩者可被分配給共用冗餘

寫入空間 162。該分配可通過改變以下兩者來實現，即 (i) 與指向區域 10 的識別碼的區域 1 相關聯的參數，以及 (ii) 與指向區域 1 的識別碼的區域 10 相關聯的參數。緩衝在電路 106a 中的區域 4 可通過將關聯參數改變為空值（或分配值）而被分配給讀取空間 164。緩衝在電路 106b 中的另一區域 8 可通過將關聯參數更新為空值（或分配值）而被分配給另一讀取空間 166。其他區域（例如，帶有附圖標記 168 的區域 12）可不駐留數據，且因此相應參數可被設置為未分配（或未使用）值。用於位於電路 106a 中的區域（例如，區域 1-6）的元數據可被緩衝在元數據空間 170 中。用於位於電路 106b 中的區域（例如，區域 7-12）的元數據可被緩衝在元數據空間 172 中。

參照圖 7，示出了用於服務寫入請求的實例性方法 180 的流程圖。方法（或過程）180 可在電路 100 中實施。方法 180 通常包括步驟（或狀態）182、步驟（或狀態）184、步驟（或狀態）186、步驟（或狀態）188、步驟（或狀態）190 以及步驟（或狀態）192。步驟 182-192 可表示可被實施為硬體、軟體、硬體和軟體的結合或其他實施的模組和/或塊。

在步驟 182 中，電路 104 可經由訊號 DH 從電路 92 接收寫入請求。對緩衝在電路 102 中的緩存數據的搜索可在步驟 184 中由電路 104（例如，模組 110）執行，以確定用於寫入數據的適當區域是否已存在於電路 102 中。若適當區域在電路 102 中不可用（例如，緩存錯失），則電路 104（例如，模組 108）可向電路 106a-106n 分配電路 102 中未使用的區域，以服務步驟 186 中的寫入請求。

一旦區域組可用於服務寫入請求，則模組 110 可在步驟 188 中從電路 94 複製對應於寫入區域的儲存數據。若從電路 94 接收到的儲存數據已不在受保護組織（或格式）中，則電路 104（例如，模組 112）可將數據安排到受保護組織中，作為步驟 188 的一部分。在步驟 188 中，數據可由模組 110 連續緩衝到合適的電路

106a-106n。在步驟 190 中，模組 110 和 112 可由受保護組織來組織從電路 92 接收到的寫入數據，並將該寫入數據儲存到電路 102 中。由於寫入數據尚未被複製回電路 94 中（例如，回寫策略），所以模組 110 可更新相應元數據以標記新寫入的數據為髒寫入數據。

若在電路 102 中發現寫入數據（例如，緩存命中），則模組 108 可將附加區域分配和鏈接到導致緩存命中的現有一個或多個區域。此後，在步驟 188 中，區域組可以與已發生緩存錯失時的相同方式從電路 94 填充。在步驟 190 中，新寫入數據可由模組 110 添加到電路 106a-106n。

參照圖 8，示出了用於服務讀取請求的實例性方法 200 的流程圖。方法（或過程）200 可在電路 100 中實施。方法 200 通常包括狀態（或步驟）202、狀態（或步驟）204、狀態（或步驟）206、狀態（或步驟）208 和狀態（或步驟）210。步驟 202-210 可表示可被實施為硬體、軟體、硬體和軟體的結合或其他實施的模組和/或塊。

在步驟 202 中，電路 104 可經由訊號 DH 從電路 92 接收讀取請求。在步驟 204 中，對緩衝在電路 102 中的緩存數據的搜索可由電路 104（例如，模組 110）來執行。該搜索可確定用於讀取數據的合適區域是否已存在於電路 102 中。若合適區域在電路 102 中不可用（例如，緩存錯失），則電路 104（例如，模組 108）可在電路 106a-106n 中分配一個或多個未使用區域，以服務步驟 206 中的讀取請求。

一旦所分配區域可用於服務讀取請求，則模組 110 可在步驟 208 中從電路 94 複製對應於讀取區域的儲存數據。在步驟 208 中，接收數據可由模組 110 連續緩衝到合適電路 106a-106n。在步驟 210 中，模組 110 可將被請求讀取數據的副本從該區域傳送到電路 92。在步驟 204 中，若在電路 102 中發現讀取數據（例如，緩存命中），

則方法 200 可繼續步驟 210，其中，電路 104 將被請求讀取數據傳送到電路 92。

參照圖 9，示出了用於分配區域的實例性方法 220 的流程圖。方法（或過程）220 可在電路 100 中實施。方法 220 通常包括步驟（或狀態）222、步驟（或狀態）224、步驟（或狀態）226、步驟（或狀態）228、步驟（或狀態）230、步驟（或狀態）232、步驟（或狀態）234 以及步驟（或狀態）236。步驟 222-236 可表示可被實施為硬體、軟體、硬體和軟體的結合或其他實施的模組和/或塊。

在步驟 222 中，電路（例如，模組 108）可確定應被分配以服務請求的多個區域。區域的數量可基於被緩衝的數據量和用於數據的組織類型（例如，冗餘保護或非冗餘）。若在步驟 224 中確定組織類型為非冗餘，則方法 220 可繼續步驟 226。電路 104（例如，模組 108）可基於與從電路 92 接收到的讀取請求相關聯的非冗餘組織來分配一個或多個區域。在步驟 226 中，該分配通常包括改變從未分配值到分配值（或空值）的區域的相應參數。所產生的一個或多個區域（例如，圖 6 中區域 164）通常準備填充來自電路 94 的數據。

若在步驟 224 中確定組織類型為冗餘，則方法 220 可繼續步驟 228。在步驟 228-236 期間，多個區域可通過改變對應於多個區域中的每一個的相應參數以指向多個區域中的另一區域而動態鏈接在一起。在步驟 228 中，模組 108 可初始化計數器（例如， $N=1$ ）。在步驟 230 中，初始分配區域的相應參數可被更新（或改變）以指向下一被分配區域。計數器在步驟 232 中可由模組 108 更新。對更多區域的檢查可在步驟 234 中執行。若更多區域應被鏈接到組中，則模組 108 可更新下一區域的參數以在步驟 230 中指向其後區域。可繼續從步驟 230 到 234 的循環，直到所有區域已被添加到組中。在步驟 236 中，模組 108 可通過更新最後區域的參數

以指回初始區域來閉合鏈接。所產生的區域組（例如，圖 6 中的組 162）可準備填充數據。

參照圖 10，示出了用於去分配區域的實例性方法 240 的流程圖。方法（或過程）240 可在電路 100 中實施。方法 220 通常包括步驟（或狀態）242、步驟（或狀態）244、步驟（或狀態）246 以及步驟（或狀態）248。步驟 242-248 可表示可被實施為硬體、軟體、硬體和軟體的結合或其他實施的模組和/或塊。

在步驟 242 中，電路 104（例如，模組 110）可將髒寫入數據從電路 102 刷新到電路 94 中。在某些情況下，寫入數據可從電路 102 讀取到電路 104、被糾錯、被恢復到原始形式、被佈置到 RAID 配置中、以及從電路 104 被傳送到電路 94。在其他情況下，髒寫入數據可被複製為類似從電路 106a-106n 中的每一個被複製到相應電路 98a-98n。例如，當區域被組織為兩個或更多個數據塊和奇偶塊時，每個數據塊和奇偶塊可無變化地從適當電路 106a-106n 經由訊號 DB 和 DS 複製到相應電路 98a-98n。

在步驟 244 中，電路 104（例如，模組 108）可確定哪個區域可從剛刷新的區域組中被解鏈和不分配。例如，若寫入數據由鏡像（例如，RAID 1 組織）保護，則由於兩個區域包括所有寫入數據，所以任一區域可被解鏈和不分配。在另一實例中，若寫入數據以 RAID 4 格式被保護，則緩衝奇偶數據的區域可被解鏈和不分配。在步驟 246 中，模組 108 可通過改變相鄰區域中的相應參數以旁路所選擇區域來解鏈所選擇區域。在步驟 248 中，模組 108 可將所選擇區域的參數改變為未分配值。

參照圖 11，示出了說明來自電路 102 中的彈性鏡像髒寫的空間益處的圖表 260 的示意圖。沿垂直軸的空間益處指示可用於緩衝非鏡像數據的總區域的百分比。如圖表 260 所示，空間益處可增加為使更少的區域表示數據的寫入部分。當所有區域被鏡像以保護髒寫入數據時（例如，100%），可實現 0 百分比空間節省。在

區域的更低百分比被用於緩衝鏡像數據時，空間益處動態（或彈性）增加。當沒有區域被用於緩衝鏡像數據時（例如，0%），正常將被用於保持鏡像數據的電路 106a-106n 中的 50%可被替換以用於保持非髒寫入數據和/或讀取數據。

根據被緩衝到電路 100 中的數據的讀/寫混合，可通過區域的彈性分組和取消分組來釋放附加緩存空間。對於緩存介質（例如，電路 106a-106n）中的給定投入，自由緩存空間中的動態變化通常提供了更好的緩存覆蓋。因此，具有擴展頻繁存取數據的服務器（例如，電路 92）可利用附加有效緩存空間。

本發明的某些實施方式一般利用在緩存裝置中正常發現的大多數區域和元數據管理架構。用於彈性分配和不分配區域的附加參數可構建在現有架構的頂端。用於創建、修改和破壞區域組的技術可從鏡像保護（例如，RAID 1 級別）擴展到其他 RAID 級別。此外，對於給定量的主機輸入/輸出數據，由於冗餘可限制髒數據，所以該技術通常減少向電路 102 中的寫入的數量。在電路 106a-106n 實施有閃存時，對於給定主機輸入/輸出活動，寫入數量的減少可提高閃存介質的壽命。

由圖 1-圖 11 的示意圖執行的功能可利用根據本說明書的教導程序化的常規通用處理器、數位計算機、微處理器、微控制器、RISC（精簡指令集計算機）處理器、CISC（複雜指令集計算機）處理器、SIMD（單指令多數據）處理器、訊號處理器、中央處理單元（CPU）、算術邏輯單元（ALU）、視頻數位訊號處理器（VDSP）和/或類似計算機器中的一個或多個來實施，這對相關領域技術人員將是顯而易見的。熟練程序員可基於本公開的教導很容易地準備適當軟體、固件、代碼、例程、指令、操作碼、微代碼和/或程序模組，這對相關領域技術人員也將是顯而易見的。來自一個或幾個介質中的軟體通常由機器實施的一個或多個處理器執行。

本發明還可通過 ASIC（專用集成電路）、平臺 ASIC、FPGA

(現場可編程門陣列)、PLD(可編程邏輯器件)、CPLD(複雜可編程邏輯器件)、門海、RFIC(射頻集成電路)、ASSP(專用標準產品)、一個或多個單片集成電路、佈置成倒裝晶片模組和/或多晶片模組的一個或多個晶片或管芯的製備來實施，或者通過互聯常規部件電路的適當網路來實施，如本文所述，其修改對於本領域技術人員將是顯而易見的。

因此，本發明還可包括計算機產品，其可以是儲存介質或介質和/或傳輸介質或包括指令的介質，該指令可被用於對機器編程以執行根據本發明的一個或多個過程或方法。由機器執行包括在計算機產品中的指令以及外圍線路的操作，可將輸入數據轉換成儲存介質上的一個或多個文件和/或表示物理對象或物質的一個或多個輸出訊號，諸如音頻和/或視頻描述。儲存介質可包括但不限於任何類型的碟，包括軟碟、硬碟機、磁碟、光碟、CD-ROM、DVD 和磁光碟以及諸如 ROM(只讀儲存器)、RAM(隨機存取儲存器)、EPROM(可擦寫可編程 ROM)、EEPROM(電可擦寫可編程 ROM)、UVPROM(紫外可擦寫可編程 ROM)、閃存、磁卡、光卡和/或適於儲存電子指令的任何介質類型的電路。

本發明的元件可形成一個或多個裝置、單元、部件、系統、機器和/或設備的一部分或全部。裝置可包括但不限於服務器、工作站、儲存陣列控制器、儲存系統、個人電腦、膝上型電腦、筆記型電腦、掌上電腦、個人數位終端、便攜式電子裝置、電池供電裝置、機上盒、編碼器、解碼器、代碼轉換器、壓縮器、解壓縮器、預處理器、後處理器、發送器、接收器、收發器、密碼電路、蜂窩電話、數位相機、定位和/或導航系統、醫療設備、頭戴式顯示器、無線裝置、錄音、音頻儲存和/或音頻回放裝置、視頻記錄、視頻儲存和/或視頻回放裝置、遊戲平臺、外設和/或多晶片模組。相關領域技術人員將理解，本發明的元件可以其他器件類型來實施以滿足特定應用的標準。如本文所使用，術語“同時”

是指描述共享某些共用時間段的事件，但該術語並非限於在相同時間點開始、在相同時間點結束的事件或具有相同時間段的事件。

儘管已參照其優選實施方式示出並描述了本發明，但本領域技術人員將理解，在不偏離本發明範圍的情況下，可做出形式和細節上的各種變更。

【符號說明】

90	系統
92	電路
94	電路
96	網路
97	網路
98a-98n	電路
100	電路
102	電路
104	電路
106a-106n	電路
108	模組
110	模組
112	模組
114	介面
116	介面
118a-118n	介面
120a-120n	介面
140	區域
142a-142n	數據線
144a-144n	數據塊

150	RAID 1 組
152	RAID 3 組
154	非冗餘組織
156	區域組
160	緩存組織
162	共用冗餘寫入空間
164	讀取空間
166	另一讀取空間
168	附圖標記
170	元數據空間
172	元數據空間
180	方法
182-192	步驟
200	方法
202-210	步驟
220	方法
222-236	步驟
240	方法
242-248	步驟
260	圖表
DB、DH、DS	雙向訊號

申請專利範圍

1. 一種用於冗餘緩存數據(redundant cache data)的彈性緩存(elastic caching)的設備，其包括：多個緩衝器；以及電路，所述電路被配置為(i)從主機接收寫入請求以在儲存體(storage volume)中儲存寫入數據，(ii)基於所述寫入請求而自複數個冗餘組織選擇一特定冗餘組織，(iii)基於所述寫入數據之大小及所述特定冗餘組織兩者而在所述緩衝器中分配多個區域(extents)，以及(iv)在所述多個區域中儲存所述寫入數據，其中，(a)所述多個區域中的每一個各別區域被分配在所述緩衝器中的一不同緩衝器內，(b)響應所述寫入請求，所述多個區域被動態鏈接在一起，以及(c)所述兩個或多個鏈接橫跨所述緩衝器之間的一或多個邊界。
2. 根據申請專利範圍第1項所述的設備，其中，所述電路還被配置為(i)自所述緩衝器複製少於全部所述寫入數據之一部份至所述儲存體，以及(ii)響應所述寫入數據之所述部分之所述複製，解鏈所述多個區域中的至少一個及少於所述多個區域之數目。
3. 根據申請專利範圍第1項所述的設備，其中，所述多個區域通過改變與所述多個區域中的每一個相對應的相應參數來動態鏈接。
4. 根據申請專利範圍第3項所述的設備，其中，所述相應參數中的每一個被改變為指向分配至所述寫入數據之所述多個區域中的另一區域。
5. 根據申請專利範圍第1項所述的設備，其中，所述電路還被

配置為(i)基於與從所述主機接收到的另一請求而選擇非冗餘組織(non-redundant organization)，以及(ii)基於所述非冗餘組織來分配一個或多個區域。

6. 根據申請專利範圍第 1 項所述的設備，其中，(i)所述電路還被配置為基於與從所述主機接收到的另一寫入請求相關聯的所述冗餘組織之另一冗餘組織，分配另一多個區域，以及(ii)所述另一冗餘組織不同於所述特定冗餘組織。
7. 根據申請專利範圍第 1 項所述的設備，其中，所述特定冗餘組織包括以在所述儲存體中之獨立盤的冗餘陣列使用的組織。
8. 根據申請專利範圍第 1 項所述的設備，其中，所述設備被實施為一個或多個集成電路。
9. 根據申請專利範圍第 1 項所述的設備，其中，所述電路還被配置為(i)自所述主機接收讀取請求，以及(ii)響應所述讀取請求之緩存錯失(cache miss)而分配一或多個未使用區域。
10. 根據申請專利範圍第 9 項所述的設備，其中，所述未使用區域係分配在非冗餘組織中。
11. 一種用於冗餘緩存數據的彈性緩存的方法，包括以下步驟：
(A)從主機接收對設備的寫入請求以在儲存體中儲存寫入數據，(B)基於所述寫入請求而自複數個冗餘組織選擇一特定冗餘組織，(C)基於所述寫入數據之大小及所述特定冗餘組織兩者而在所述緩衝器中分配多個區域；以及(D)在所述多個區域中儲存所述寫入數據，其中，(i)所述多個區域中的每一個各別區域位於所述緩衝器中的一不同緩衝器內，(ii)響

應所述寫入請求，所述多個區域被動態鏈接在一起，以及(iii)所述兩個或多個鏈接橫跨所述緩衝器之間的一或多個邊界。

12. 根據申請專利範圍第 11 項所述的方法，還包括以下步驟：

自所述緩衝器複製少於全部所述寫入數據之一部份至所述儲存體；以及

響應所述寫入數據之所述部分之所述複製，解鏈所述多個區域中的至少一個及少於所述多個區域之數目。

13. 根據申請專利範圍第 11 項所述的方法，其中，所述多個區域通過改變與所述多個區域中的每一個相對應的相應參數來動態鏈接。

14. 根據申請專利範圍第 13 項所述的方法，其中，所述相應參數中的每一個被改變為指向分配至所述寫入數據之所述多個區域中的另一區域。

15. 根據申請專利範圍第 11 項所述的方法，還包括以下步驟：

基於與從所述主機接收到的另一請求而選擇非冗餘組織，以及

基於所述非冗餘組織來分配一個或多個區域。

16. 根據申請專利範圍第 11 項所述的方法，還包括以下步驟：

基於與從所述主機接收到的另一寫入請求相關聯的在所述儲存體中之另一冗餘組織，分配另一多個區域，其中，所述另一冗餘組織不同於所述特定冗餘組織。

17. 根據申請專利範圍第 11 項所述的方法，其中，所述特定冗餘組織包括以在所述儲存體中之獨立盤的冗餘陣列使用的

組織。

18. 根據申請專利範圍第 11 項所述的方法，還包括以下步驟：

自所述主機接收讀取請求；及

響應於所述讀取請求之緩存錯失而分配一或多個未使用區域。

19. 根據申請專利範圍第 18 項所述的方法，其中，所述未使用區域係分配在非冗餘組織中。

20. 根據申請專利範圍第 11 項所述的方法，其中，所述設備包括緩存電路。

圖式

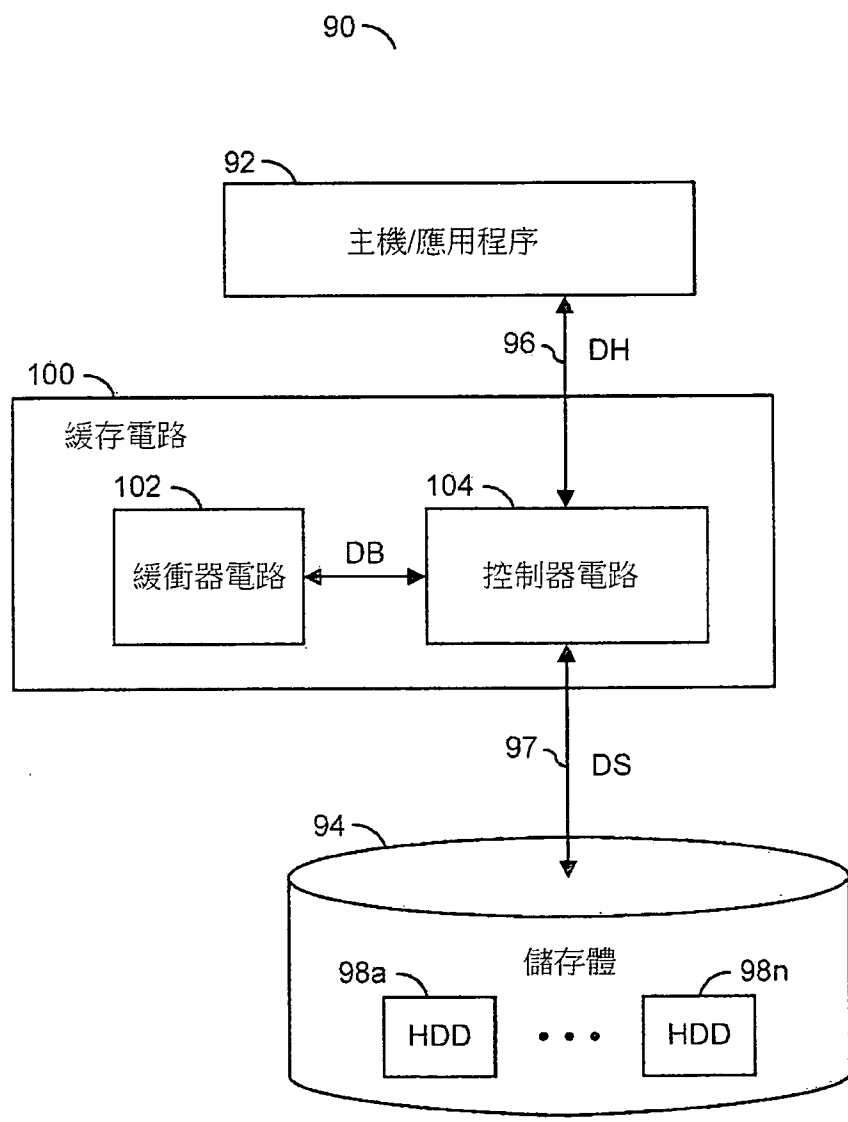


圖1

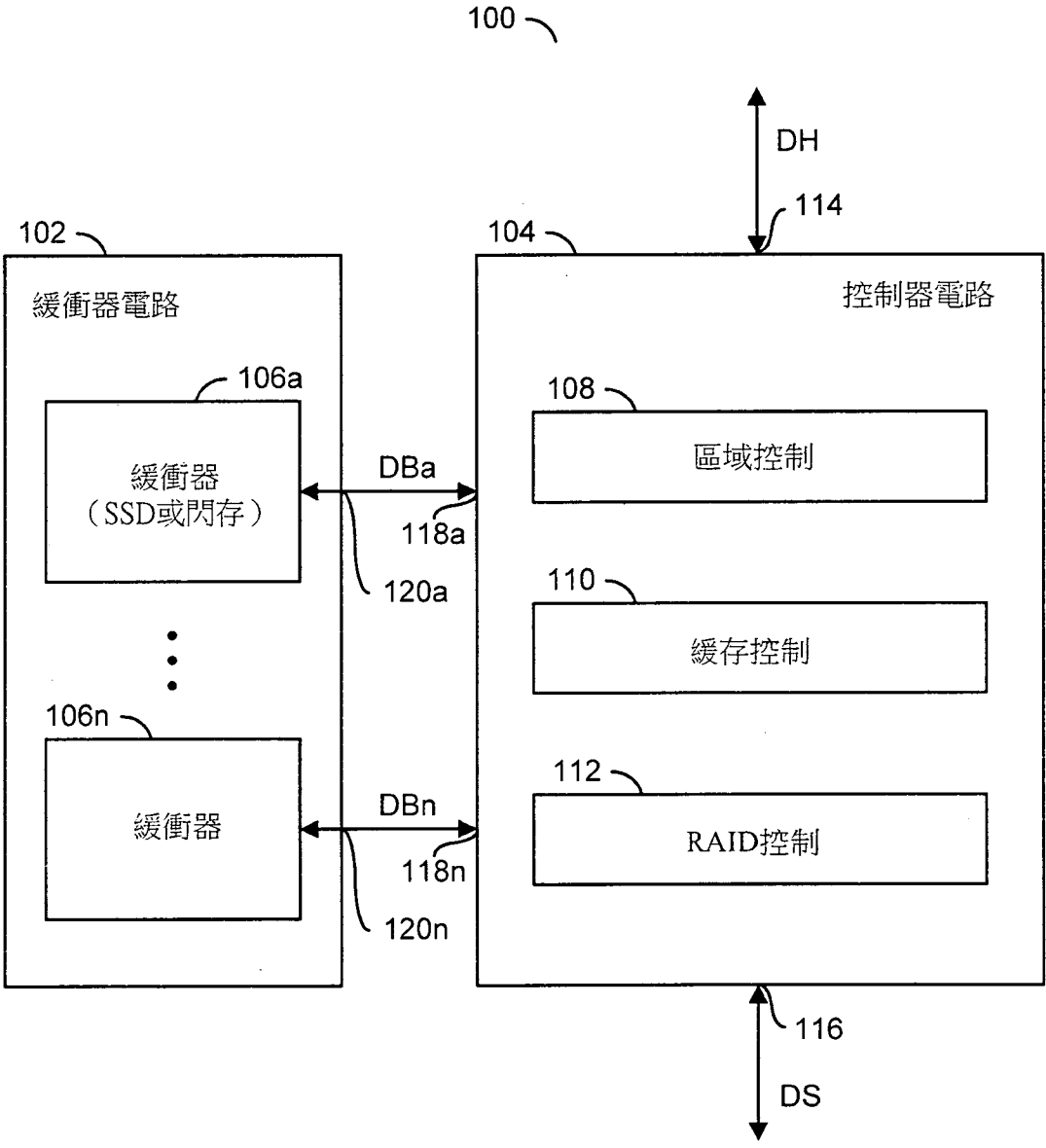


圖2

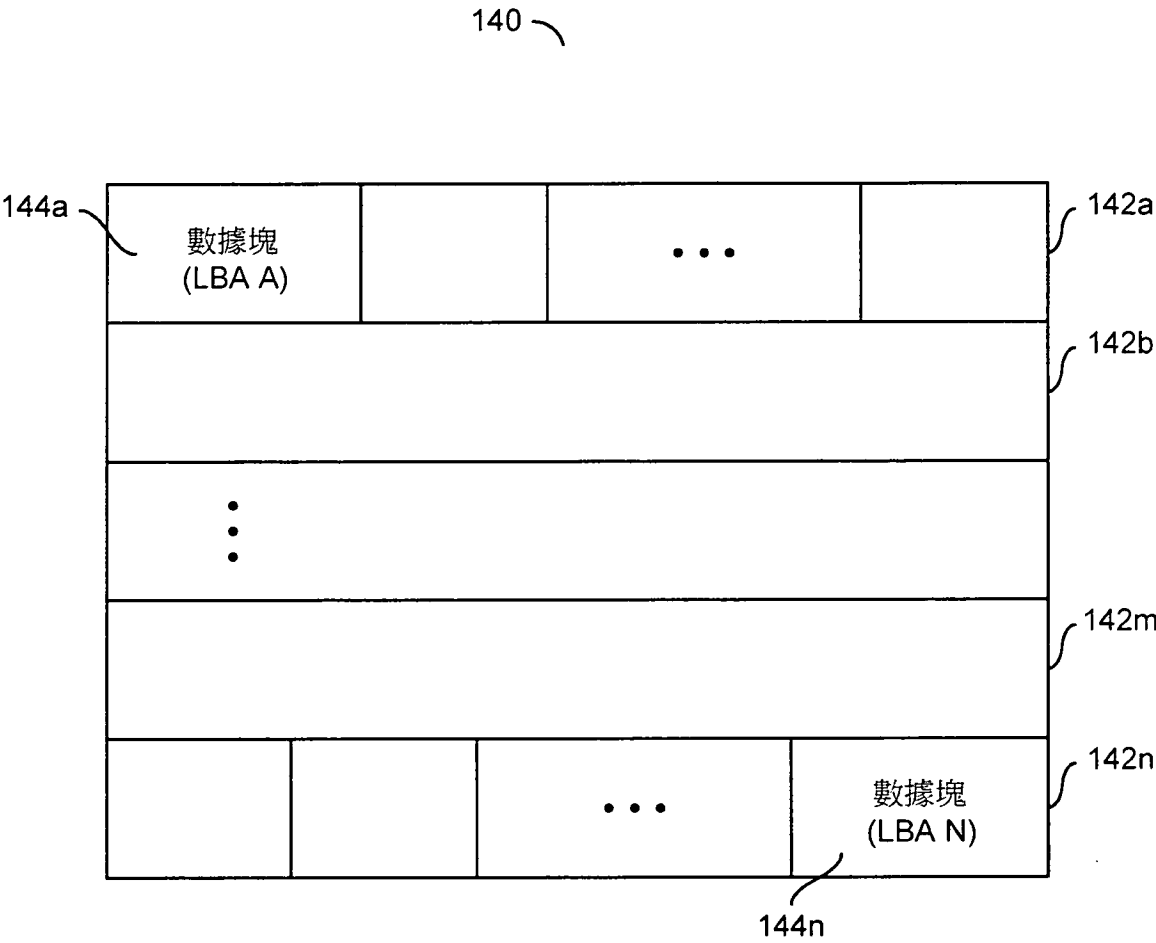


圖3

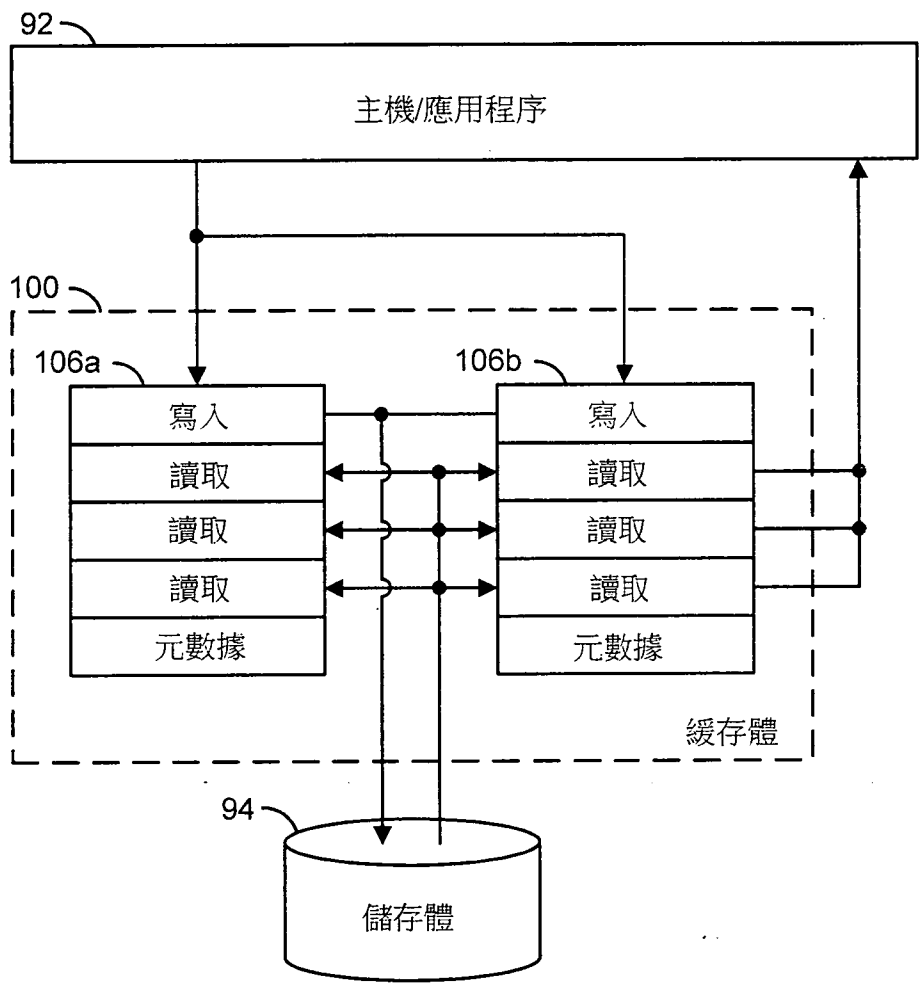


圖4

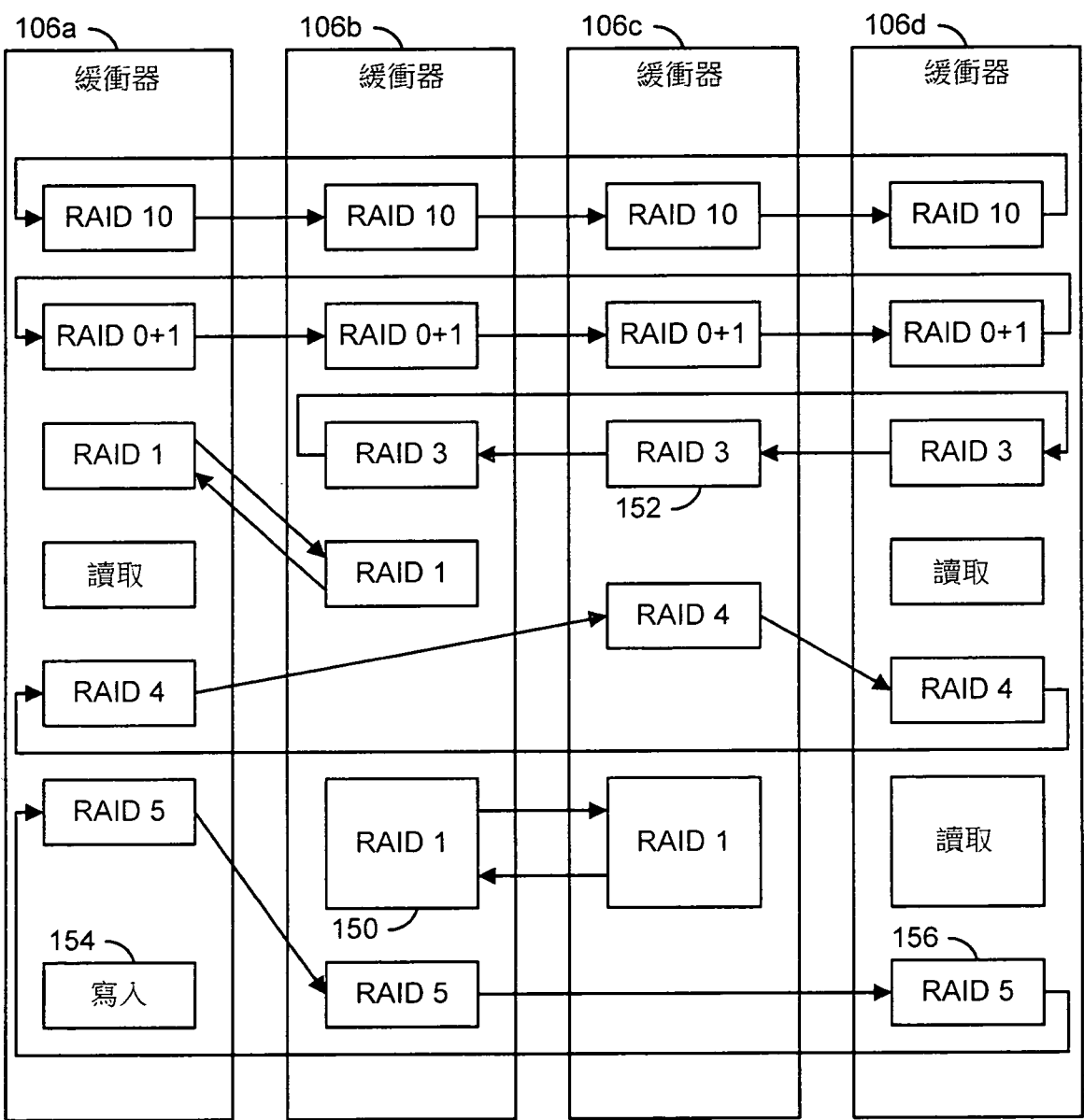


圖5

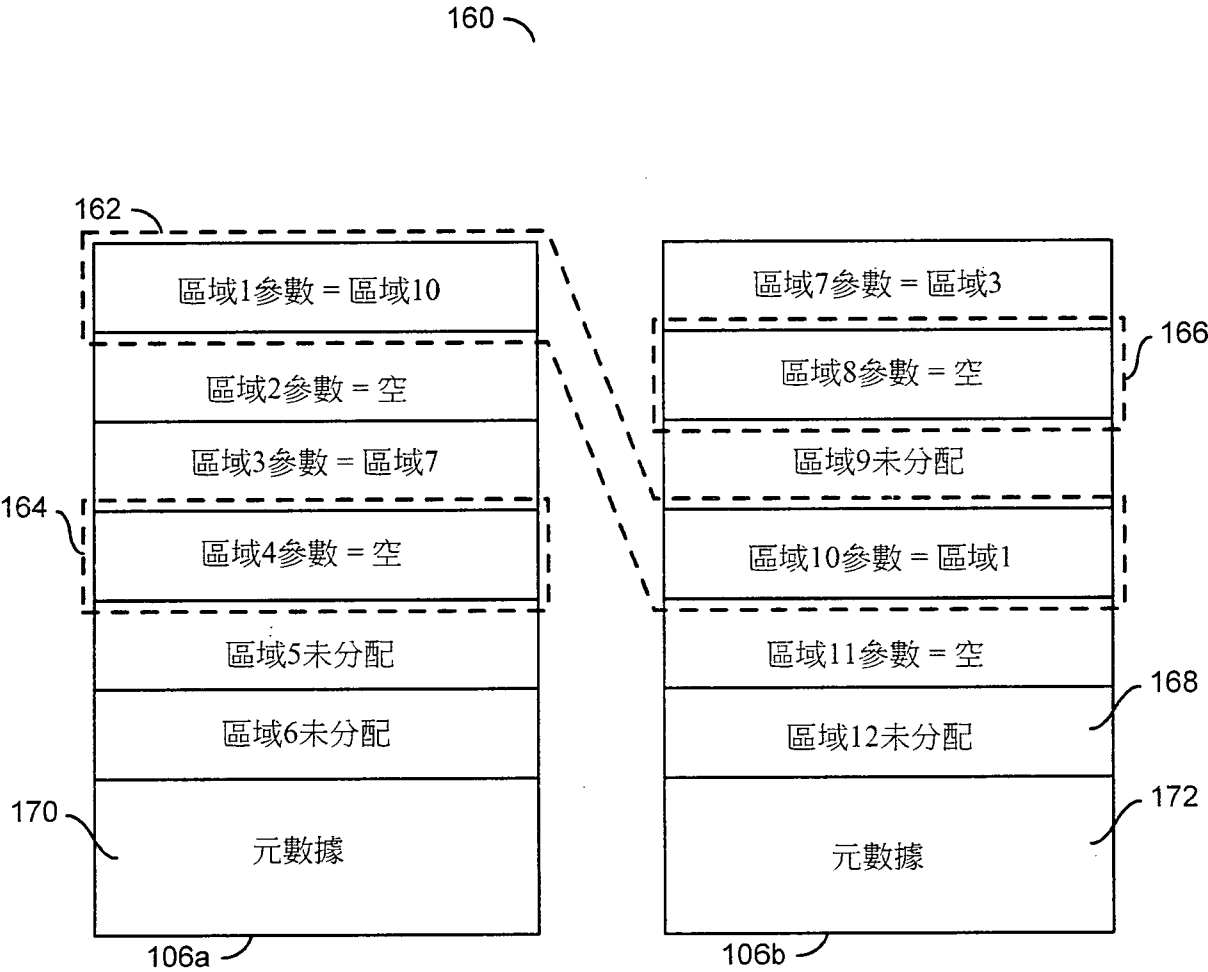


圖6

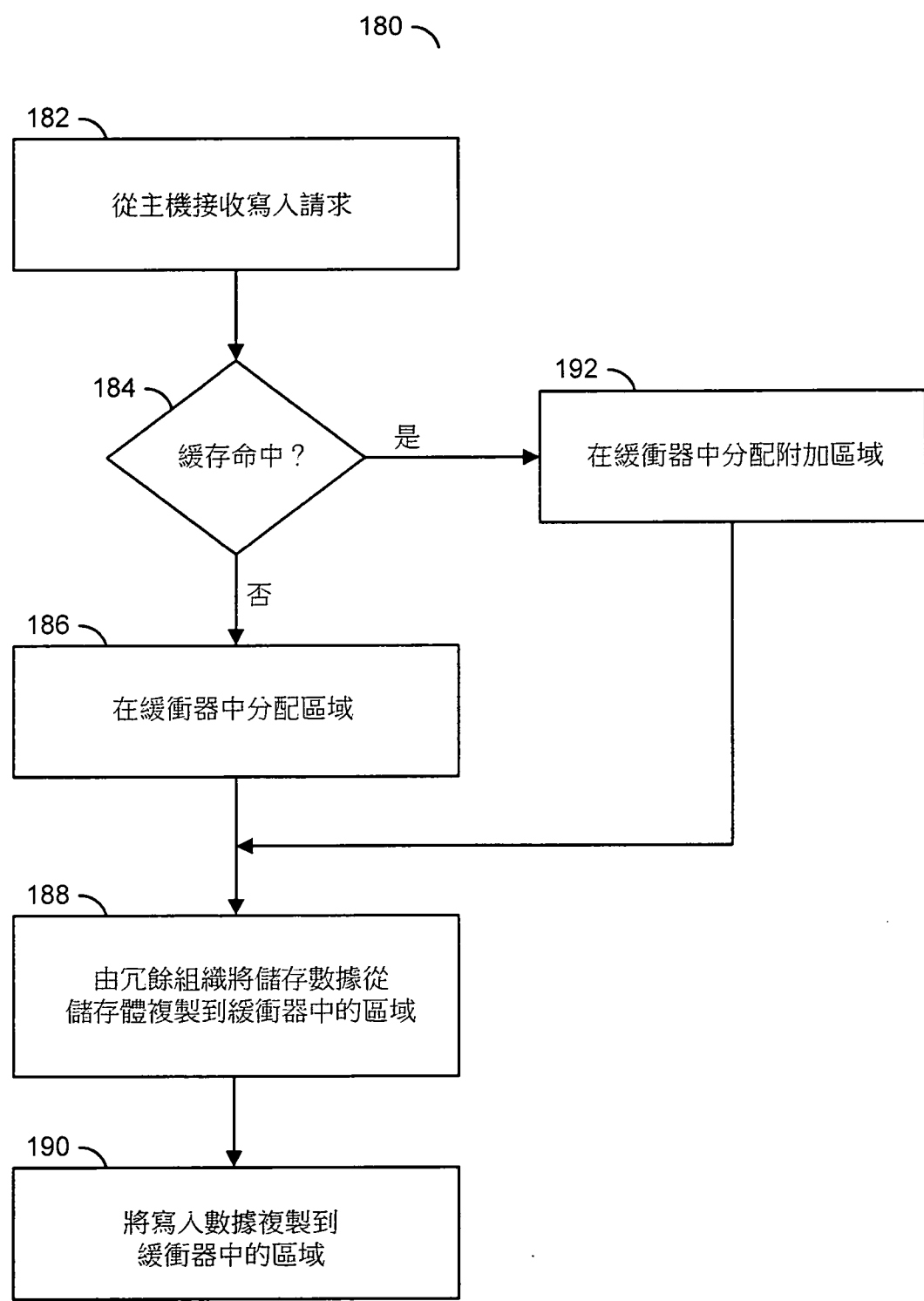


圖7

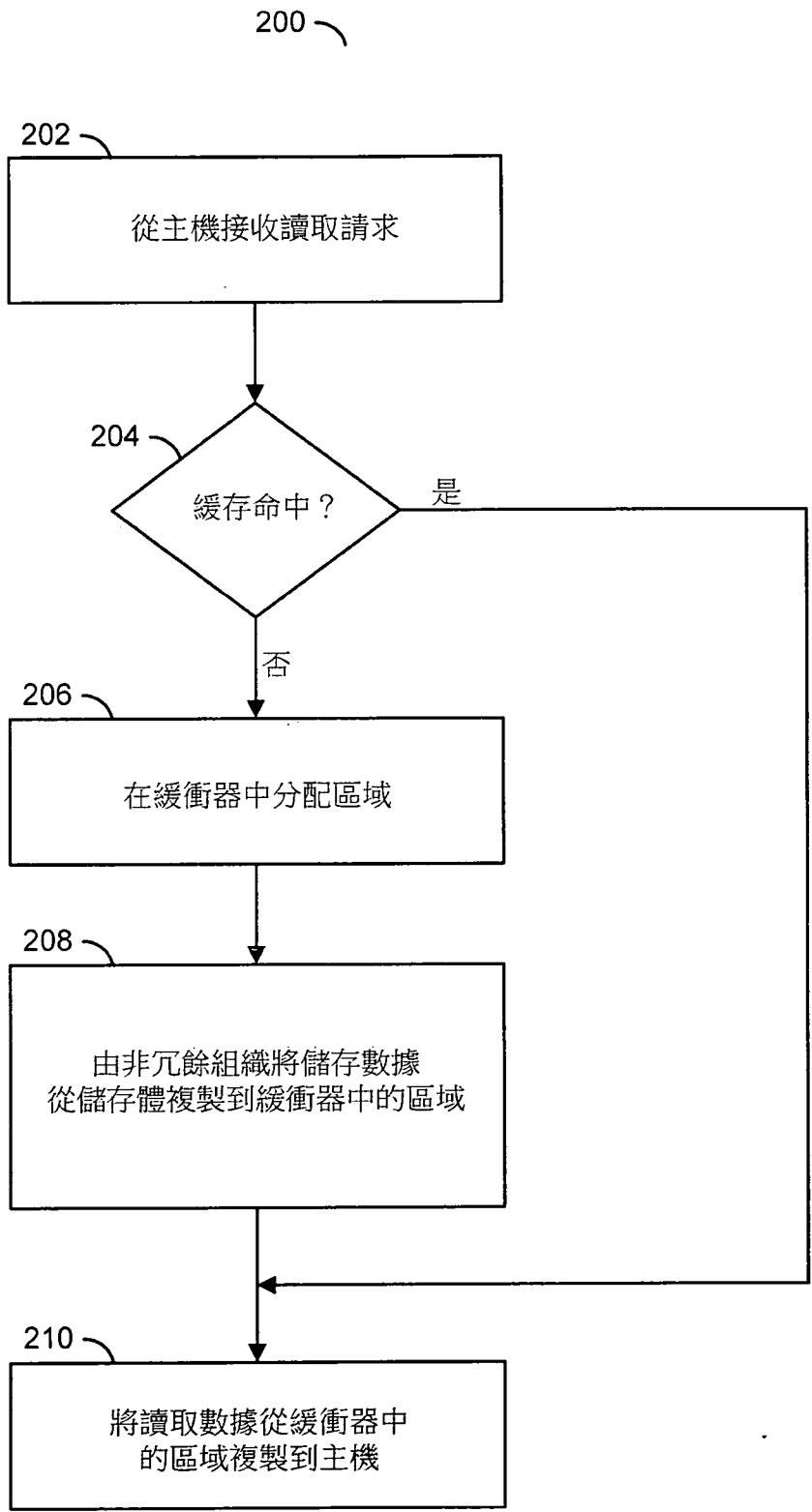


圖8

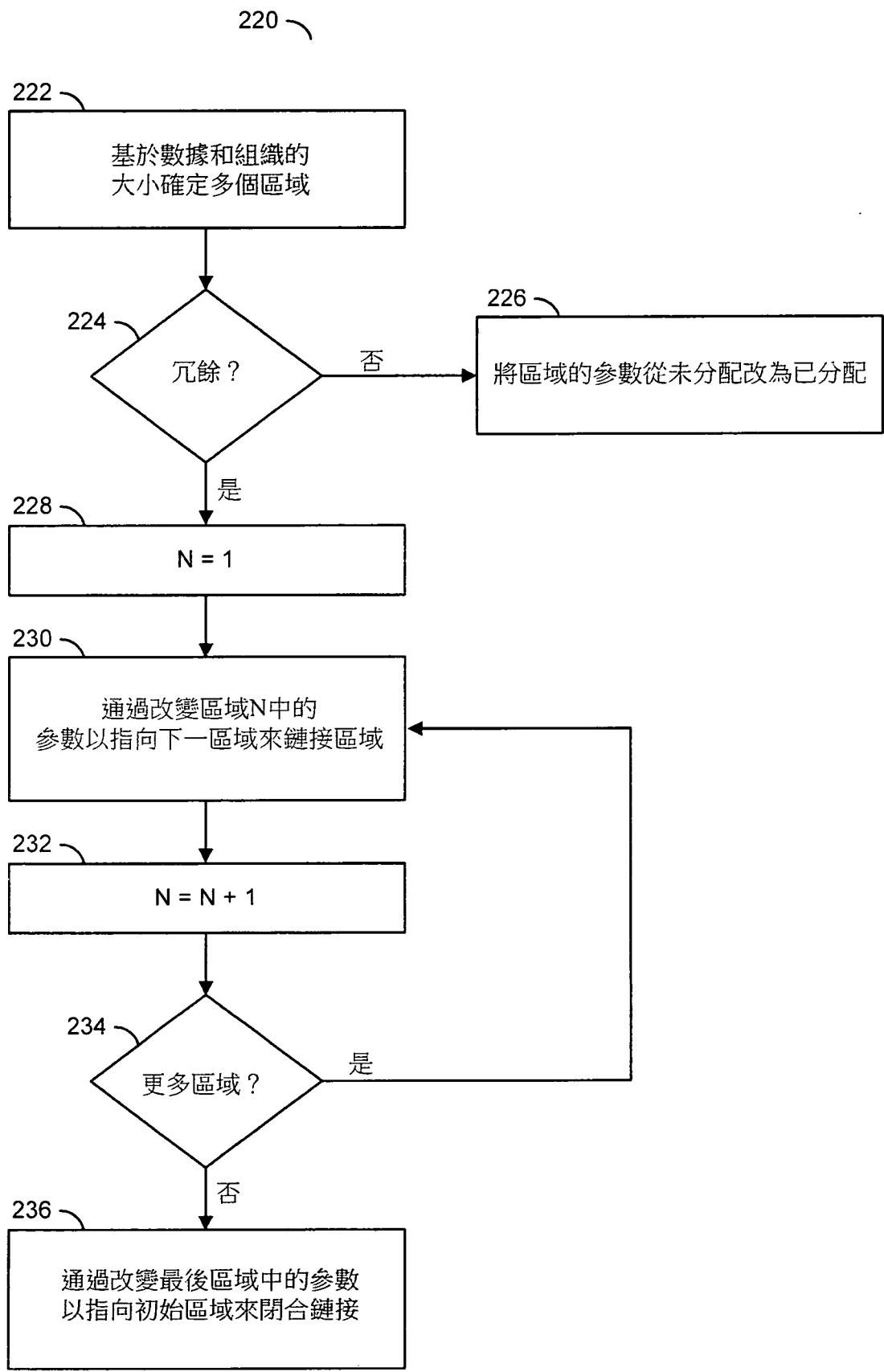


圖9

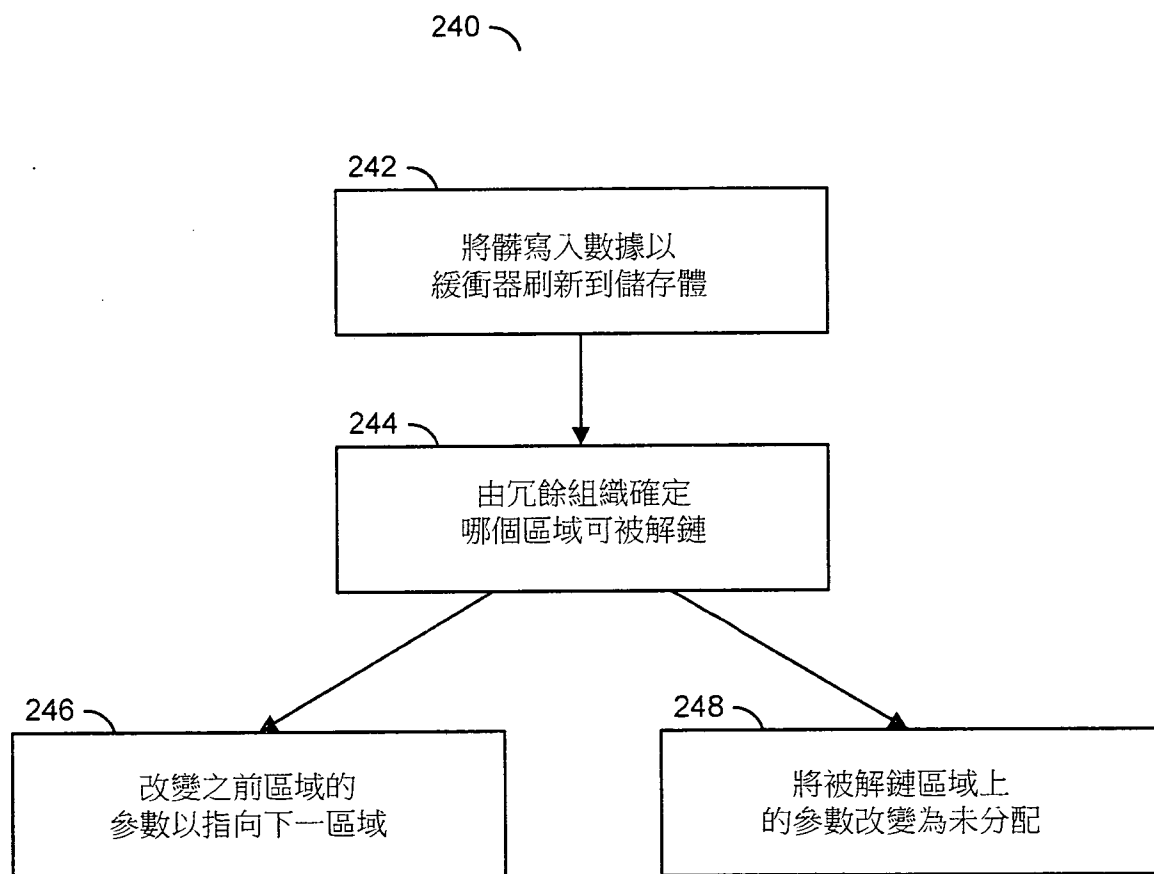


圖 10

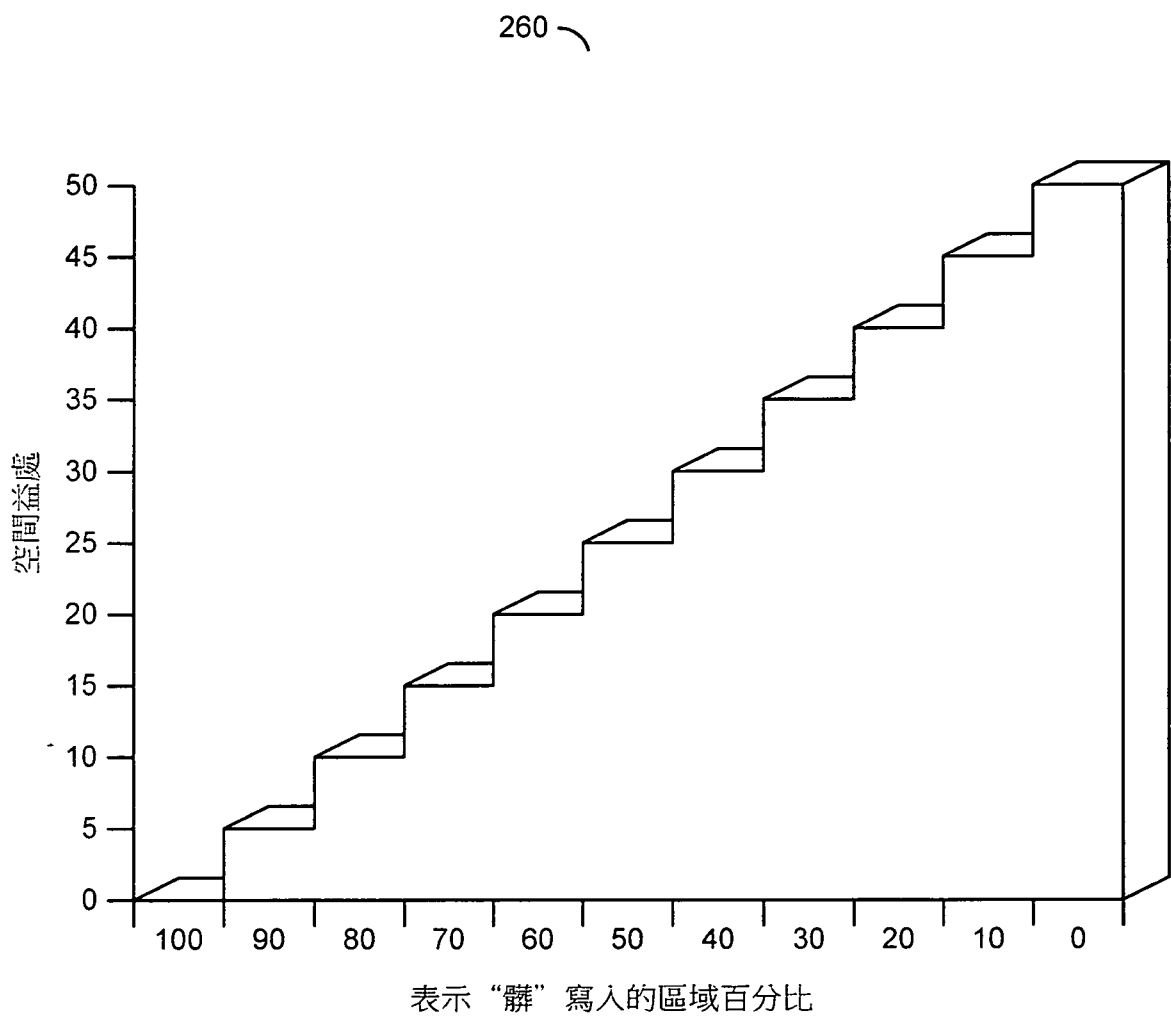


圖11