

#### 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：

美國；西元 2003 年 7 月 14 日；10/619,649

☐ 無主張專利法第二十七條第一項國際優先權：

☐ 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。

## 九、發明說明：

### 【發明所屬之技術領域】

本發明係關於儲存系統。本發明尤其是關於一種規劃一儲存系統之方法，此儲存系統包含有複數個儲存單元之陣列，供增加儲存系統可容忍的儲存單元失效數，而不會遺失儲存於系統中的資料。

### 【先前技術】

以下定義在此使用並且提供說明而非限制之目的：

一“元件(element)”，係位於一儲存單元的一資料塊。

一“基準陣列(base array)”，係用於一 ECC 且包含一陣列單元的一組元件。

一“陣列(array)”，係容納一或多個陣列的一組儲存單元。

一“條帶(stripe)”，係一陣列內的一基準陣列。

$n$  係基準陣列中的資料單元數。

$r$  係基準陣列中的冗餘單元數。

$m$  係陣列中的儲存單元數。

$d$  係基準陣列的最小漢明距離。

$D$  係陣列的最小漢明距離。

大型儲存系統通常包含數個獨立儲存單元之陣列。每一陣列一般藉由一擦去(Erasure)(或錯誤)校正碼(ECC)，於例如一鏡像規劃或一 RAID5(個別磁碟冗餘陣列第五階)規劃，來避免某些儲存系統失效(亦稱擦去)。ECC 碼提供對每一陣列為本區的冗餘儲存單元，並且藉由處理對陣列的子組為本區的單元失效，來增加一儲存系統的可靠性。

相較於元件可靠性的進步，以硬磁碟(HCC)為基準的儲存系統之儲存容量增加更快。因此，最小漢明距離  $d=2$  方案，如 RAID5 及鏡像技術，不再提供系統階層足夠可靠性的適當保護。另外的設計，如 RAID6(雙同位)於距離  $d=3$ ，雙鏡像於距離  $d=3$ ，以及 RAID51 於距離  $d=4$ ，已提出解決系統可靠性的不足之方案。在儲存系統中，提供備用單元(spare unit)來減少系統修復時間，並增加維修間距是常見的方法。然而，增加備用單元，則增加了系統的成本，以及減少了系統的效能。

加強系統可靠性的其他方法有包含較高位同位碼的使用。舉例來說，J.S. Plank, "針對失誤容限於 RAID 般的系統之 Reed-Solomon 編碼的教學，"軟體-練習&經驗，27(9)，1997 年九月，p.p.995-1012，揭露一 Reed-Solomon 碼的範例。

再者，E.J. Schwabe et al., "計算於磁碟陣列大約平衡的同位-取消叢集的佈局，"ACM 0-89791-813-4/96/95 1996，揭露針對改良的性能，冗餘資訊之有效位置的資料佈局。

P. Chen et al., "RAID：高性能，可靠次要儲存"，ACM 計算問卷，26 期，1994 年六月，p.p. 145-185，提供一冗餘磁碟陣列的概述。M. Holland et al., "於冗餘磁碟陣列之持續作業的同位取消叢集"，第五屆程式語言及作業系統之架構支援的國際會議之紀錄 (ASPLOS-V)，p.p. 23-25，1992 年十月，揭露 RAID 系統的取消叢集之同位。G.A.

Alvarez et al., “容忍多個失效於 RAID 架構”，ACM 0-89791-901-7/97/0006 1997 描述一般使用 8 位元有限欄位的多數同位陣列之屬性及建設。

M.M. Blaum et al.的美國專利號 5,579,475，標題為”使用簡單非回覆性對角及行同位，於一 DASD 陣列，來編碼及重建至多兩個不可使用的 DASDs 的方法及手段”，揭露一距離  $d=3$  陣列的執行。N.K. Ouchi,”兩階 DASD 失效修復方法，”IBM 技術揭露告示欄，36:03 期，1993 年三月，揭露從具有失效的距離  $d=3$  陣列，重建資料所需的作業。

然而，某些陣列設計，如產品碼(包含 RAID51)，不禁某些儲存單元失效型樣。這些陣列表現的像他們擁有本區冗餘一般。

需要的是一種藉由使陣列中的本區冗餘，可供陣列系統之全面性使用，以改良一儲存系統之可靠性的技術。再者，需要的是改良一儲存系統之可靠性的之技術，此儲存系統對儲存單元失效型樣是敏感的。再者，需要的是一種允許儲存系統的維修，較一般習知系統可有更長延遲的技術。

### 【發明內容】

本發明提供一種藉由使一陣列中的本區冗餘，可供陣列系統全面性使用，以改良一儲存系統之可靠性的技術。再者，本發明提供一種儲存系統之可靠性的改良之技術，此儲存系統對儲存單元失效型樣是敏感的。再者，較一般

習知系統，本發明提供一種允許儲存系統的維修延遲一段長時間的技術。

藉由增加具有複數個陣列之儲存系統的失效容限的方法，本發明之優點可被提供，其中每一陣列具有複數個儲存單元。儲存系統之陣列包含基於一擦去或錯誤校正碼，如一同位碼、一 Winograd 碼、一對稱碼、一 Reed-Solomon 碼、一 EVENODD 碼或一 EVENODD 碼之衍生，之冗餘。一儲存系統之失效容限係以系統之最小漢明距離  $D$  來給定。系統之最小距離，因此，為相對條帶之所有最小漢明距離之最小值，即， $D=\min(d_i)$ 。當捐贈者陣列之最小距離與接收者陣列之一最小距離間的差值，為大於或等於 2 時，捐贈者陣列從複數個陣列中選出。選取於捐贈者陣列之捐贈者儲存單元，係基於對捐贈者陣列有一最小性能衝擊。接收者儲存單元從接收者陣列被選出。接著，至少一部分的遺失資訊會從接收者陣列被重建至捐贈者陣列中的選取的儲存單元。接收者資訊基於接收者陣列之一改良性能來選取。選取的儲存單元對捐贈者陣列表明，在遺失資訊重建至選取的儲存單元前，其已經被捐贈。較佳地，在捐贈者陣列從複數個陣列被選出前，接收者陣列之最小漢明距離為  $d \geq 2$ 。當一備用儲存單元變得可使用時，備用儲存單元以習知方法被指派給一選取的陣列。

在從接收者陣列重建至少一部份的接收者資訊至選取的儲存單元步驟之期間，當儲存單元於捐贈者陣列失效時，重建步驟被結束，並且選取一第二捐贈者陣列。從接收者陣列的至少一部份遺失接收者資訊被重建至第二捐

贈者陣列之選取的儲存元件。第二捐贈者陣列的選取經重新估計前述的狀態來進行。

本發明亦提供一種增加儲存單元之陣列的失效容限的方法，此儲存單元不禁失效的特定型樣。根據本發明，在儲存單元失效後，接收者儲存單元從儲存單元之陣列選出。捐贈者儲存單元從儲存單元之陣列選出，使得在重建作業後，陣列的失效容限被增加。最後，至少一部份遺失接收者資訊被重建於捐贈者儲存單元中。當一備用單元變成可使用時，備用單元以習知方法被指派給陣列。

再者，本發明提供具有複數個陣列及一系統陣列控制器的一種資料儲存系統。每一陣列具有複數個儲存單元，並且包含有基於一擦去或錯誤校正碼，如一同位碼、一Winograd碼、一對稱碼、一Reed-Solomon碼、一EVENODD碼或一EVENODD碼之衍生，的冗餘。系統陣列控制器與每一陣列耦合，並且偵測位於複數個陣列之第一陣列中的一儲存單元之一失效。當第二陣列的最小距離與第一陣列的最小距離間之差值，係大於或等於2時，系統控制器接著在複數個陣列之第二陣列中選取一儲存單元。從第一陣列的至少一部份遺失資訊，被重建至第二陣列的選取之儲存單元。每一儲存單元可為一HD、一揮發性隨機存取記憶體、一非揮發性隨機存取記憶體、一光學儲存裝置或一磁帶儲存裝置。

本發明亦提供具有複數個儲存單元之一陣列以及一陣列控制器的一種資料儲存系統。陣列包含有基於一擦去或

錯誤校正碼，如一同位碼、一 Winograd 碼、一對稱碼、一 Reed-Solomon 碼、一 EVENODD 碼或一 EVENODD 碼之衍生，的冗餘。陣列亦不禁選取的失效型樣，且/或一不相同失效機率。陣列控制器與陣列耦合，並且偵測在陣列中的一第一儲存單元之失效。陣列控制器接著選取一第二儲存單元於陣列中，使得陣列的一失效容限在一重建作業後被增加，並且重建至少一部份資訊從第一儲存單元到第二儲存單元。第二儲存單元係基於一陣列的失效型樣來選取，且/或基於一預定的目標型樣。在陣列控制器選取第二儲存單元前，陣列的最小漢明距離為  $d \geq 2$ ，且在從第一儲存單元重建至少一部份資訊至第二儲存單元完成後被增加。每一儲存單元可為一 HD、一揮發性隨機存取記憶體、一非揮發性隨機存取記憶體、一光學儲存裝置或一磁帶儲存裝置。

### 【實施方式】

本發明改良一儲存系統之可靠性，並且比一可相比而沒有同位交換的儲存系統，對儲存系統維修提供相當長時間的延遲。因此，本發明提供一顯著的可靠性改良，超越 RAID 系統所提供之可靠度。與對待多數陣列儲存系統之每一陣列為一獨立個體的 RAID 系統相反，本發明全面地耦合多數陣列儲存系統之獨立陣列，允許一陣列之冗餘被另一陣列來使用。此程序在此意指一自主同位交換 (APX)，或為一 APX 作業。

根據本發明，一 APX 作業允許藉由陣列的一系統，使在一陣列之本區冗餘被全面性使用，因此當儲存單元數增

加，增加系統可靠性。APX 亦減少或消除備用儲存單元的需求。

為說明本發明之優點，使用一特定之範例，考慮一 48 單元之儲存系統，此儲存系統包含有六個儲存單元之八個陣列，且沒有備用儲存單元。再者，此範例儲存系統使用具有距離  $d=4$  的一對稱碼。對稱碼具有相同數目的資料單元及冗餘單元，以及有由一半單位結合的遺失之恢復能力。藉由 APX，在至多九個儲存單元失效的情況，儲存系統之距離保持在  $D=2$ 。假設 APX 作業可以其作業期間，以少於兩個失效而完成，十個失效時，儲存系統達到  $D=1$ 。相反地，使用十六個儲存單元之三個陣列的一 RAID6 系統，以兩個儲存單元失效即可達到  $D=1$ 。更相反地，使用十六個儲存單元之三個陣列的一 RAID51 系統，以三個儲存單元失效即可達到  $D=1$ 。

再者，利用 APX 的一儲存系統，在失效堆積時可安全地降低性能，允許系統的維修的延遲伴隨著重要的節省成本。因此，與可相比之不利用 APX 的儲存系統之年保固成本相比，利用 APX 的一儲存系統之年保固成本要少的許多。對一習知的系統，在備用儲存單元池被耗盡時，服務通常被請求。當使用 APX，可在此範例系統至多九個單元失效後，服務才被請求。因此，與一沒有 APX 的系統相比較下，使用 APX 時，系統在一較長間段可維持一預定距離。

APX 允許一組陣列內的陣列交換冗餘，而克服集中在

一組陣列的單一陣列中之失效的顯現。舉例說明，若一第一陣列具有一最小漢明距離，比一第二陣列之最小漢明距離少 2 或更多，則第二陣列可捐贈一儲存單元給第一陣列。之後，第一陣列的失效容限會被增加，以及第二陣列的失效容限會被減少，但不低過第一陣列的層級。因此，系統的最小距離會被增加，而增加系統的失效容限。

圖 1a 顯示一範例儲存系統，一般以 100 表示，包含與一共同陣列控制器 101 連結之兩儲存陣列 102 及 103。儲存陣列 102 及 103 包含多數個儲存單元 104，並且與陣列控制器 101 經由介面 105 通訊。陣列控制器 101 與其他控制器以及主機系統藉由介面 106 通訊。此一規劃允許一陣列控制器與多數個儲存陣列通訊。

圖 1b 顯示一範例儲存系統，一般以 150 表示，包含兩儲存陣列 153 及 154，係分別與不同的陣列控制器 152 及 151 連結。儲存陣列 153 藉由介面 157 與陣列控制器 152 通訊，以及儲存陣列 154 藉由介面 156 與陣列控制器 151 通訊。陣列控制器 151 及 152 藉由介面 158 及 159，分別與其他陣列控制器以及儲存系統通訊。亦如圖 1b 所顯示，一通訊連結 160 係允許陣列控制器 151 及 152 相互通訊。

圖 1a 及 1b 所顯示的陣列控制器可被設計做硬體或軟體控制器。控制器一詞將會在此用來指任一上述的規劃。

圖 2 顯示兩個陣列 201 及 202 的一範例組，供說明本發明。陣列 201 包含有儲存單元 1A-1F，以及陣列 202 包

此捐贈。在重建期間，已被寫入接收者陣列資料的資料之分段屬於接收者。直到重建完成，捐贈者陣列可被允許寫至捐贈的單元，以在剩餘分段中維持最新的資料。在一儲存單元於一捐贈者陣列在同位交換期間失效的情況下，結束捐贈作業較佳，並且初始一新捐贈作業與一不同的捐贈者陣列。捐贈者陣列可接著重建至先前捐贈的分段，來增加陣列的距離。舉例來說，若失效發生於捐贈作業的最初部分，則轉回捐贈及初始一新捐贈者的方法，應以減少的距離減少系統執行的時間。

一旦一備用儲存單元變成可使用，例如經維修，可被指派來替換任一失效的單元。資訊可以已知的方法來重建。指派一接收者陣列之失效的單元至捐贈者陣列，可促進此作業，因為這顯示一失效的單元屬於哪一陣列。

圖 4 顯示根據本發明一 APX 作業後的圖 2 之陣列。重建的資料於圖 4 被劃線。當仍有兩個失效的儲存單元於系統中時，僅有一失效的儲存單元於每一陣列。針對圖 4 的系統規劃，每一陣列現在有距離  $d=3$ ，且可再容許兩個失效，而不會造成資料遺失。系統的總距離現為  $D=3$ 。

使用 APX，一儲存系統可比原先的系統容忍更多失效。在圖 2-4 的範例中，不使用 APX，第一次的系統失效會是在四個單元失效。相反地，當使用 APX 時，第一次的系統失效會是在六個單元失效。此改良尚未藉由請求額外冗餘或備用來達成，但可藉由調整儲存系統內冗餘的全面性指派，來符合所觀察到的儲存單元失效之狀態。APX

所提供的改良隨著系統內的陣列數增加而增加。

在具有相同距離陣列的儲存系統中，捐贈者陣列必有至少距離  $d=3$ ，使得 APX 作業得以執行。即，一捐贈者陣列一定有至少比接收者陣列的距離大 2 的距離。通常，當一接收者陣列在距離等於或大於 2 時，一 APX 作業較佳地被執行，以防止在一 APX 作業期間，硬體錯誤或更多的儲存單元失效。

許多習知系統，如 RAID5，使用散佈式同位(亦稱做取消叢集同位)來分散叢集型樣。在此系統中，每一儲存單元具有部分指派至每一單元類型(如資料 1、資料 2、冗餘 1 等)。APX 亦可應用於使用散佈式同位的系統。在此一案例中，一捐贈者儲存系統的選取較不重要，因為冗餘是分散在所有單元中。系統可選取簡單地重建任一接收者的失效儲存單元。

當捐贈者陣列有多數個選取時，一捐贈者陣列的選取標準可基於例如利用性、裝置的年份，以及先前錯誤歷史來考慮。

圖 2-4 所示的說明範例執行與陣列做同位交換，其中冗餘單元數資料單元數相同。當陣列中的儲存單元數大於基準陣列中的儲存單元數時，一 APX 作業可與一遮擋作業(dodging operation)一同執行，如專利申請案號(律師登錄號 ARC920030014US1)所揭露，加註於此供參考，更增加系統的可靠性。

有些陣列設計是對單元失效型樣敏感的。在此陣列中，捐贈者單元以及接收者單元皆可來自相同陣列。一捐贈者單元可以陣列規劃為基準來選取。圖 5 顯示包含有 30 儲存單元之一產品碼陣列 500，其中行 501-505 形成 RAID6 陣列( $d=3$ )，以及列 511-516 形成 RAID5 陣列( $d=2$ )。在陣列 500 中， $D=6$  伴隨 14 個冗餘磁碟。然而，僅有六個失效的某些位置排列會造成陣列失效，如圖 6 所示，其中儲存單元 1J、1K、1M、3J、3K 及 3M 已失效。失效的型樣可被辨識為在一第一行中的三個失效，與在一第二行中的三個失效匹配。使用 APX 可增加系統的失效容限。系統會以此方法選取捐贈者單元以及接收者單元，而避免這些型樣及返回陣列至  $D=3$  的狀態。

針對圖 5 的範例，當其排列如圖 7 所示時，系統可維持  $D=3$  與九個失效。在此，陣列有九個失效單元，1N、2N、3N、4N、5H、5J、5K、5L 及 5M。這些系統可選取具有高失效容限之一的目標型樣。當失效發生時，捐贈者單元從目標型樣被選取，並指派給非此目標型樣一部份的失效的接收者。針對圖 5 的範例，一 APX 作業可被執行許多次，使得陣列在達到  $D=2$  前，可容忍至少十個失效。藉由使用 APX，有效距離因此從  $D=6$  被增加到  $D=12$ 。APX 允許維修被延遲，直到 30 個單元中 10 個失效，陣列維持至少  $D=2$ 。沒有 APX，維持  $D=2$  可請求以少至四個單元失效的維修。

“失效”及“型樣”等詞在此被用來指自陣列中邏輯位置

之資訊的擦去，而非實體位置。此非 MDS 陣列可經如說明的產品碼、低密度同位碼、非均勻圖像碼或任一具有不禁特定型樣的碼來形成。

APX 可使用於簡單地增加一儲存系統之最小距離之外。許多其他因素可被包含來決定是否執行 APX，以及選取捐贈者及接收者。舉例說明，當元件為不相同時，其個別失效機率，導致資料遺失的失效結合，以及系統性能的效力皆可被考慮。在此案例中，系統的最小距離可在 APX 後保持不變。

APX 可與其他具有最小距離  $d \geq 3$  的陣列類型一同使用。再者，一較小的陣列大小使 APX 被更有效使用，以及使包含有小型陣列的大型系統達到高失效容限。當一儲存系統具有一備用池，最好在執行一 APX 作業前，執行重建於此備用池。

APX 亦可被執行於一儲存單元中的資料之子組中。舉例來說，在某些規劃重建次數會被減少。考慮圖 4 的範例。代替重建單元 1C 的內容至單元 2F，可能較有利的是重建一半的單元 1C 至一半的單元 2F，以及另一半的單元 1E 至一半的單元 2E。此總結果皆為陣例在  $D=3$ ，但重建數可能會減少，因為相同的資料數在被重建，但兩捐贈磁碟在被使用。其他結合亦明顯的可行。

雖然本發明於此以從 HDD 儲存單元所形成之儲存單元作為描述，然而本發明係可應用於其他記憶體裝置之陣

列所形成的儲存系統，例如隨機存取記憶體(RAM)儲存裝置(皆揮發及非揮發)、光學儲存裝置以及磁帶儲存裝置。再者，適合用於虛擬的儲存系統，如從網路連結儲存器的陣列。更可應用於任一冗餘系統，其中有些狀態資訊係將一冗餘元件與特定元件子組關聯，以及狀態資訊可使用一捐贈作業被轉移。

雖然本發明以包含現在較佳的實施例來實施本發明的具體範例來描述，熟知技藝者皆知上述的系統及技術有許多改變及變更係界於附加之請求項內，並不偏離本發明之精神及範圍。

#### 【圖式簡單說明】

本發明係藉由範例來描述並不被伴隨之圖式限制。圖式中類似參考述表示相似元件，其中：

圖 1a 顯示一儲存系統之一般規劃有複數個陣列連結至一共同儲存控制器；

圖 1b 顯示一儲存系統之一般規劃有複數個陣列連結至一分離的儲存控制器；

圖 2 顯示兩陣列之一範例組來說明本發明；

圖 3 顯示在陣列之一中的兩個磁碟失效後的圖 2 的陣列來說明本發明；

圖 4 顯示根據本發明在一 APX 作業後的圖 2 的陣列；

圖 5 顯示一不禁儲存單元失效之型樣的一範例陣列；

圖 6 顯示導致資料遺失之圖 5 的陣列的六個儲存單元失效之一範例型樣；

圖 7 顯示具有距離  $d=3$  的圖 5 的甚列之九個儲存單元失效

之一範例目的型樣。

## 【主要元件符號說明】

100、150 儲存系統

101、102、103、151、152 陣列控制器

104 儲存單元

105、106、156、157、158、159 介面

153、154 儲存陣列

160 通訊連結

201、202 陣列

500 產品碼陣列

501-505 行

511-516 列

## 五、中文發明摘要：

藉由使在選取的陣列中之本區冗餘可於儲存系統中全面性使用(globally)，具有複數個陣列的儲存系統之錯誤容限(error tolerance)可被增加。為達成此增加的錯誤容限，當捐贈者陣列(donor array)之最小距離與接收者陣列(recipient array)之最小距離間的差值，係大於或等於 2 時，捐贈者陣列從複數個陣列被選取。捐贈者儲存單元在捐贈者陣列中被選取，以及接著從接收者陣列重建接收者資訊至選取的儲存單元。選取的儲存單元對捐贈者陣列表明，在遺失資訊重建至選取的儲存單元前，其已經被捐贈。較佳地，在捐贈者陣列從複數個陣列被選取前，接收者陣列之最小漢明距離(Hamming distance)為  $d \geq 2$ 。

## 六、英文發明摘要：

Error tolerance is increased for a storage system having a plurality of arrays by making local redundancy in a selected array globally available throughout the storage system. To achieve the increased error tolerance, a donor array is selected from the plurality of arrays when the difference between a minimum distance of the donor array and a minimum distance of a recipient array is greater or equal to 2. A donor storage unit is selected in the donor array and recipient information is then rebuilt from the recipient array on the selected storage unit. The selected storage unit is indicated to the donor array as having been donated before the lost information is rebuilt on the selected storage unit. Preferably, the minimum Hamming distance of the recipient array is  $d \geq 2$  before the donor array is selected from the plurality of

array.

197/12/01

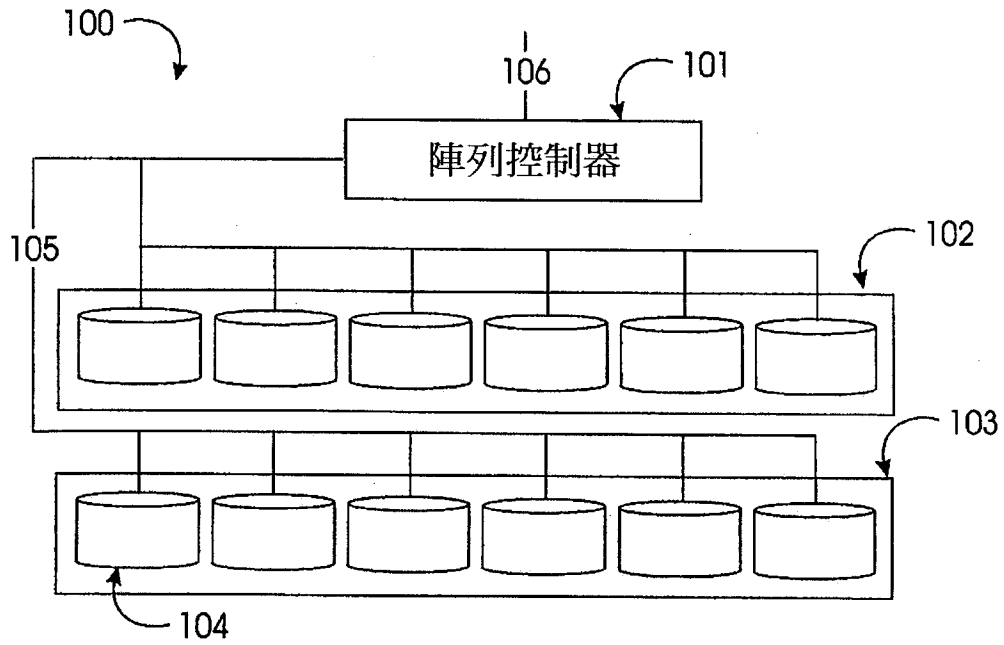


圖 1a

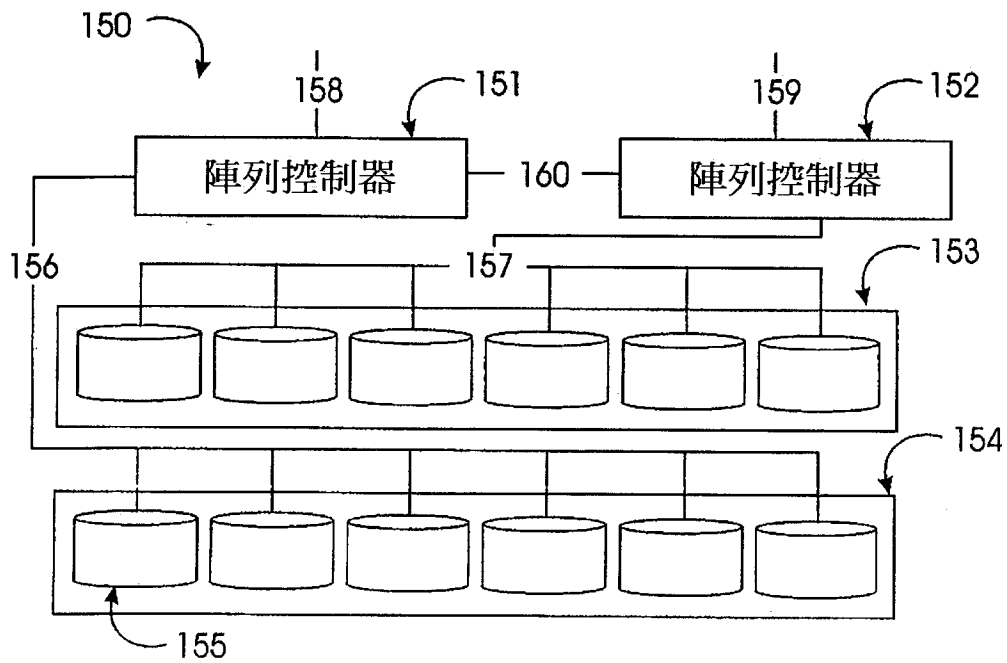


圖 1b

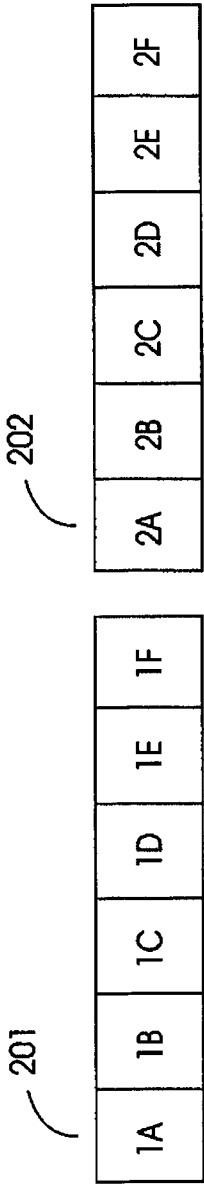


圖 2

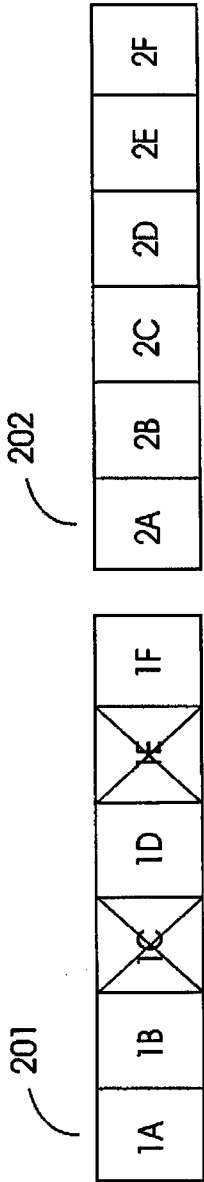


圖 3

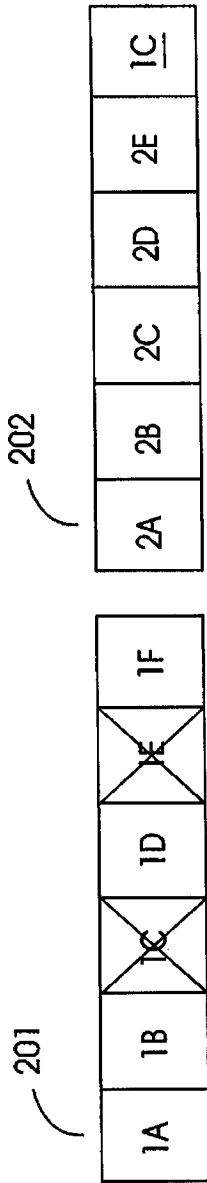


圖 4

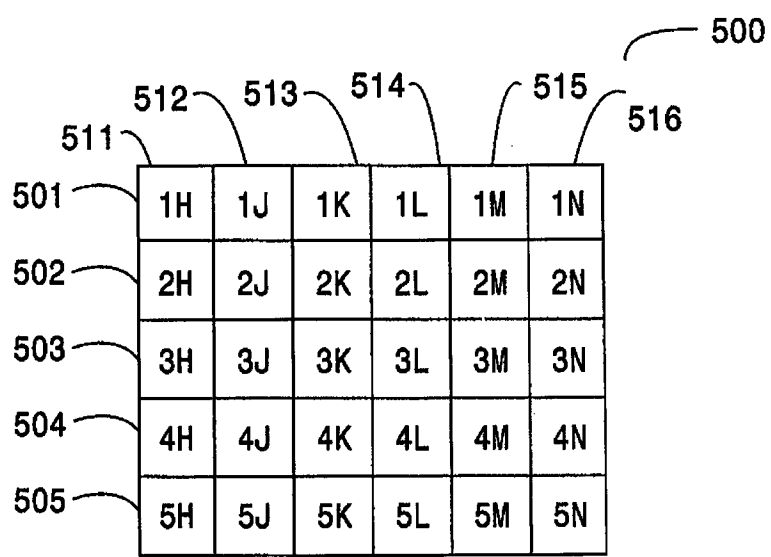


圖 5

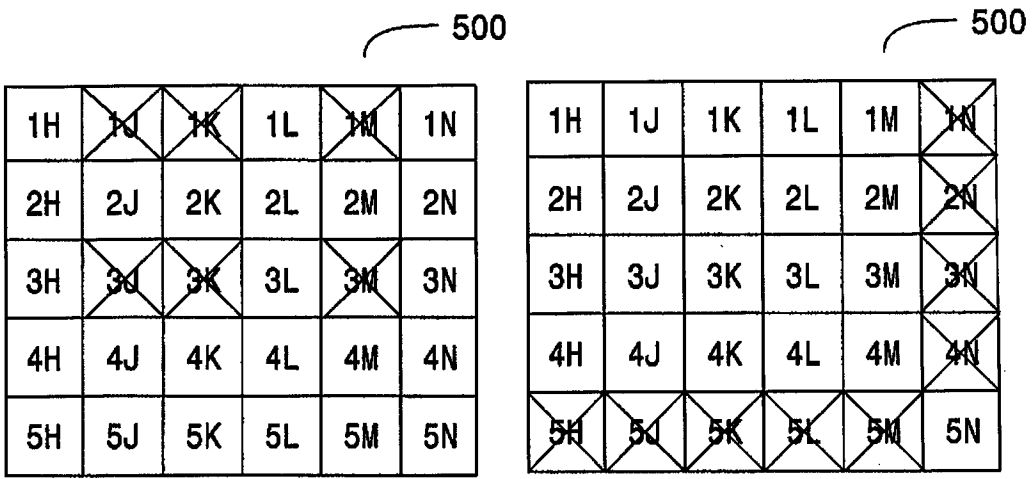


圖 6

圖 7

**七、指定代表圖：**

(一)本案指定代表圖為：圖 1。

(二)本代表圖之元件符號簡單說明：

100 儲存系統

101、102、103 陣列控制器

104 儲存單元

105、106 介面

**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**

# 發明專利說明書

(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：93120449

※申請日期：93年7月8日

※IPC分類：G06F11/07  
G11C5/00  
(2000.01)

## 一、發明名稱：(中文/英文)

增加一儲存系統的失效容限之方法及系統

METHOD AND SYSTEM FOR INCREASING FAILURE TOLERANCE  
OF A STORAGE SYSTEM

## 二、申請人：(共1人)

姓名或名稱：(中文/英文)

萬國商業機器公司

INTERNATIONAL BUSINESS MACHINES CORPORATION

代表人：(中文/英文) 傑羅 羅森梭 / ROSENTHAL, GERALD

住居所或營業所地址：(中文/英文)

美國紐約州 10504 亞芒克市新奧爾察德路

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2人)

姓名：(中文/英文)

斯帝芬 羅伯特 黑茲勒

HETZLER, STEVEN ROBERT

國籍：(中文/英文) 美國/U.S.A.

姓名：(中文/英文)

丹尼爾 菲力 斯密士

SMITH, DANIEL FELIX

國籍：(中文/英文) 英國/GB

含有儲存單元 2A-2F。每一陣列的儲存單元 A、B 及 C 為資料儲存單元，以及每一陣列的儲存單元 D、E 及 F 為有一 MDS 碼的冗餘儲存單元。因此，陣列 201 及 202 皆具有一最小漢明距離  $d=4$ 。圖 2 所示之規劃意指為一對稱碼，其中資料儲存單元數相等於冗餘儲存單元數。計算冗餘，使任三個單元失效可被對稱碼校正。擦去或錯誤校正碼(ECC)，如同位碼、Winograd 碼、對稱碼、Reed-Solomon 碼、EVENODD 碼及 EVENODD 碼之衍生，可被使用來產生冗餘。在此所使用，"距離"一詞是指最小漢明距離。

圖 3 顯示在陣列 201 之儲存單元 1C 及 1E 失效後之圖 2 的陣列。在儲存單元 1C 及 1E 失效後，陣列 201 具有距離  $d=2$ ，而陣列 202 仍有距離  $d=4$ 。陣列 201 僅可再容許一失效，而無資料遺失事件的可能。整體系統距離為陣列 201 之距離，即距離  $D=2$ 。藉由使用 APX 重新平衡冗餘，系統的整體失效容限可增加，這是因為系統的距離從  $D=2$  被增加到  $D=3$ 。

藉由捐贈包含於陣列 202(即捐贈者陣列)內的一儲存系統，可達成重新平衡冗餘，接著提供捐贈的儲存單元至陣列 201(即接收者陣列)宛如捐贈的儲存系統是一備用儲存單元。捐贈者陣列必須知道捐贈的儲存單元不再是捐贈者陣列的一部份，以預防捐贈者陣列讀取或寫入資料至捐贈的單元。從接收者陣列指派失效的儲存單元之一至捐贈者陣列，使得兩陣列皆維持一不變的儲存單元數，以及維持失效的單元的認知是有益的。沒有資訊可被捐贈者陣列寫至捐贈的單元。系統可選取捐贈哪一儲存單元。選取一

捐贈者單元之主要標準，係基於選取對捐贈者陣列的可靠性有最小衝擊之捐贈者單元。一次要標準，係基於選取對性能的儲存單元有最小衝擊，例如有最複雜冗餘計算的單元。系統可選取從哪一失效單元之資料來重建至捐贈者的單元。選取資料被重建的主要標準，係基於提供最能增加可靠性之資料組。一次要標準，係選取在重建執行後提供最佳陣列性能資料組。在上述圖 3 的範例中，藉由重建 1C 單元的資料組，接收者陣列 201 將會有最佳性能，這是因為 1C 單元為一資料單元。類似地，藉由捐贈者儲存冗餘資料的一單元，如單元 2F，捐贈者陣列 202 將會有最少性能衝擊。在兩者情況中，在 APX 作業後，讀取指令會因此被處理，而不需要從儲存陣列 201 及 202 的冗餘單元重建資料。

從一捐贈者陣列捐贈一儲存單元至一接收者陣列，需要儲存系統有從一陣列指派儲存單元至另一陣列的能力。當捐贈者及接收者陣列連結至一共同陣列控制器 101 時，如圖 1a 所示，則控制器 101 可內部地執行此作業。當捐贈者及接收者陣列連結至一分離陣列控制器 151 及 152 時，如圖 1b 所示，則控制器 151 及 152 可交換資訊。舉例說明，控制器可經由通訊連結 160，如一簡易磁碟串 (JBOD) 陣列規劃的方法，接觸個別磁碟。另外，控制器可從牽涉的儲存單元之位置，來交換有關要被讀取及寫入的資料之資訊。

經由人工指示，捐贈已失效於捐贈者陣列之捐贈的單元是可能達成的。然而，較有益的是以分段的方式來執行

## 十、申請專利範圍：

1. 一種增加一儲存系統的失效容限之方法，該儲存系統具有複數個陣列，每一陣列具有複數個儲存單元，該方法包含以下步驟：

選取一接收者陣列(recipient array)；

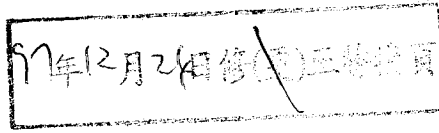
從該複數個陣列選取一捐贈者陣列(donor array)，係當該捐贈者陣列之一最小距離，與該接收者陣列之一最小距離間的一差值，係大於或等於 2 時；

選取於該捐贈者陣列之一捐贈者儲存單元；以及

從該接收者陣列重建至少一部份的遺失接收者資訊至該捐贈者陣列中該選取的儲存單元。

2. 如請求項 1 所述之方法，其中在從該複數個陣列選取該捐贈者陣列之該步驟前，該接收者陣列之該最小漢明距離為  $d \geq 2$ 。
3. 如請求項 1 所述之方法，進一步包含對該捐贈者陣列表明，於重建該遺失資訊至該選取的儲存元件的該步驟前，該選取的儲存已被捐贈之一步驟。
4. 如請求項 1 所述之方法，其中該儲存單元係硬碟機。
5. 如請求項 1 所述之方法，其中該儲存單元係隨機存取記憶儲存裝置。
6. 如請求項 1 所述之方法，進一步包含從該接收者陣列中選取一接收者儲存單元的一步驟。
7. 如請求項 6 所述之方法，其中該接收者儲存單元之至少一部份的該遺失內容，重建至該捐贈者儲存單元中。
8. 如請求項 1 所述之方法，其中該捐贈者儲存單元進一步，係基於對該捐贈者陣列有一最小性能衝擊而被選取。

9. 如請求項 1 所述之方法，更包含基於該接收者陣列之一改良性能，而選取該接收者資訊之一步驟。
10. 如請求項 1 所述之方法，其中該儲存系統之該陣列包含基於一擦去或錯誤校正碼之冗餘。
11. 如請求項 10 所述之方法，其中該擦去或錯誤校正碼係一同位碼。
12. 如請求項 10 所述之方法，其中該擦去或錯誤校正碼係一 Winograd 碼。
13. 如請求項 10 所述之方法，其中該擦去或錯誤校正碼係一對稱碼。
14. 如請求項 10 所述之方法，其中該擦去或錯誤校正碼係一 Reed-Solomon 碼。
15. 如請求項 10 所述之方法，其中該擦去或錯誤校正碼係一 EVENODD 碼。
16. 如請求項 10 所述之方法，其中該擦去或錯誤校正碼係一 EVENODD 碼之一衍生。
17. 如請求項 10 所述之方法，其中該儲存系統之該陣列，包含基於複數個擦去或錯誤校正碼之一乘積的冗餘。
18. 如請求項 17 所述之方法，其中至少一該擦去或錯誤校正碼係一同位碼。
19. 如請求項 17 所述之方法，其中至少一該擦去或錯誤校正碼係一 Winograd 碼。
20. 如請求項 17 所述之方法，其中至少一該擦去或錯誤校正碼係一對稱碼。
21. 如請求項 17 所述之方法，其中至少一該擦去或錯誤校正碼係一 Reed-Solomon 碼。
22. 如請求項 17 所述之方法，其中至少一該擦去或錯誤校正



碼係一 EVENODD 碼。

23.如請求項 17 所述之方法，其中至少一該擦去或錯誤校正碼係一 EVENODD 碼之一衍生。

24.如請求項 1 所述之方法，其中在從該接收者陣列重建至少一部份之接收者資訊至該選取的儲存單元的該步驟期間，當該捐贈者陣列之一儲存單元失效時，該方法更包含以下步驟：

結束從該接收者陣列重建至少一部份之接收者資訊至該選取的元件的該步驟；

從該複數個陣列中選取一第二捐贈者陣列，係當該第二捐贈者陣列之一最小距離與該第二接收者陣列之一最小距離間的一差值，係大於或等於 2 時；

選取該第二捐贈者陣列之一捐贈者儲存元件；以及從該接收者陣列重建至少一部份之遺失接收者資訊至該第二捐贈者陣列之該選取的儲存元件。

25.如請求項 1 所述之方法，其中當一備用的儲存單元變成可使用時，該方法更包含指派該備用的儲存單元至一選取的陣列之一步驟。

26.一種增加儲存單元之一陣列的失效容限之方法，該儲存單元不禁失效的特定型樣，該方法包含以下步驟：

在一儲存單元失效後，從儲存單元之該陣列選取一接收者儲存單元；

從儲存單元之該陣列選取一捐贈者儲存單元，使得在一重建作業後，該陣列之一失效容限被增加；以及

重建至少一部份之遺失接收者資訊至該捐贈者儲存單元中。

27.如請求項 26 所述之方法，其中在選取該捐贈者儲存單

元之該步驟前，該陣列之該最小漢明距離為  $d \geq 2$ 。

28. 如請求項 26 所述之方法，其中在該重建之步驟後，該陣列之該最小漢明距離被增加。
29. 如請求項 26 所述之方法，其中該接收者儲存單元係基於該陣列之一失效型樣而被選取。
30. 如請求項 26 所述之方法，其中該捐贈者儲存單元係基於一預定目標型樣而被選取。
31. 如請求項 26 所述之方法，其中該儲存單元之該陣列，包含基於一擦去或錯誤校正碼的冗餘。
32. 如請求項 31 所述之方法，其中該擦去或錯誤校正碼係一同位碼。
33. 如請求項 31 所述之方法，其中該擦去或錯誤校正碼係一 Winograd 碼。
34. 如請求項 31 所述之方法，其中該擦去或錯誤校正碼係一對稱碼。
35. 如請求項 31 所述之方法，其中該擦去或錯誤校正碼係一 Reed-Solomon 碼。
36. 如請求項 31 所述之方法，其中該擦去或錯誤校正碼係一 EVENODD 碼。
37. 如請求項 31 所述之方法，其中該擦去或錯誤校正碼係一 EVENODD 碼之一衍生。
38. 如請求項 26 所述之方法，其中該陣列包含基於複數個擦去或錯誤校正碼之一乘積的冗餘。
39. 如請求項 38 所述之方法，其中至少一該擦去或錯誤校正碼係一同位碼。
40. 如請求項 38 所述之方法，其中至少一該擦去或錯誤校正碼係一 Winograd 碼。

- 41.如請求項 38 所述之方法，其中至少一該擦去或錯誤校正碼係一對稱碼。
- 42.如請求項 38 所述之方法，其中至少一該擦去或錯誤校正碼係一 Reed-Solomon 碼。
- 43.如請求項 38 所述之方法，其中至少一該擦去或錯誤校正碼係一 EVENODD 碼。
- 44.如請求項 38 所述之方法，其中至少一該擦去或錯誤校正碼係一 EVENODD 碼之一衍生。
- 45.一種資料儲存系統，包含：  
    複數個陣列，每一陣列具有複數個儲存單元；以及  
    與每一陣列耦合之一系統陣列控制器，該系統陣列控制器偵測該複數個陣列之一第一陣列中的一儲存單元之一失效，選取該複數個陣列之一第二陣列之一儲存單元，係當該第二陣列之一最小距離與該第一陣列之一最小距離間的一差值，係大於或等於 2 時，以及從該第一陣列重建至少一部份的資訊至該第二陣列之該選取的儲存單元中。
- 46.如請求項 45 所述之資料儲存系統，其中當有一備用儲存單元變成可使用的，該備用儲存單元係被指派至該第二陣列。
- 47.如請求項 45 所述之資料儲存系統，其中至少一陣列具有一不相同的失效機率。
- 48.如請求項 45 所述之資料儲存系統，其中該系統陣列控制器包含複數個陣列控制器，一陣列控制器與該複數個陣列之至少一陣列耦合，每個陣列控制器偵測一儲存單元之一失效，該儲存單元於與該陣列控制器相關之每一陣列中，該複數個陣列控制器之一第一陣列控制器選取一儲存單元，該儲存單元於與該第一陣列控制器相關的

每一陣列中，該複數個陣列控制器選取一儲存單元，該儲存單元於與該第一陣列控制器相關之一陣列中，此係當該第二捐贈者陣列之一最小距離與該第二接收者陣列之一最小距離間的一差值，係大於或等於 2 時，以及該第一及第二陣列控制器從與該第二陣列控制器相關的該陣列，重建至少一部分之遺失資訊至該選取的儲存單元，該選取的儲存單元於與該第一陣列控制器相關之該陣列中。

49. 如請求項 48 所述之資料儲存系統，其中當有一備用儲存單元變成可使用的，該備用儲存單元係被指派至該選取的儲存單元之該陣列。
50. 如請求項 48 所述之資料儲存系統，其中至少一陣列具有一不相同的失效機率。
51. 如請求項 45 所述之資料儲存系統，其中該資料儲存系統之該陣列，包含基於一擦去或錯誤校正碼的冗餘。
52. 如請求項 51 所述之資料儲存系統，其中該擦去或錯誤校正碼係一同位碼。
53. 如請求項 51 所述之資料儲存系統，其中該擦去或錯誤校正碼係一 Winograd 碼。
54. 如請求項 51 所述之資料儲存系統，其中該擦去或錯誤校正碼係一對稱碼。
55. 如請求項 51 所述之資料儲存系統，其中該擦去或錯誤校正碼係一 Reed-Solomon 碼。
56. 如請求項 51 所述之資料儲存系統，其中該擦去或錯誤校正碼係一 EVENODD 碼。
57. 如請求項 51 所述之資料儲存系統，其中該擦去或錯誤校正碼係一 EVENODD 碼之一衍生。

- 58.如請求項 45 所述之資料儲存系統，其中該儲存系統之該陣列，包含基於複數個擦去或錯誤校正碼之一乘積的冗餘。
- 59.如請求項 58 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一同位碼。
- 60.如請求項 58 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一 Winograd 碼。
- 61.如請求項 58 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一對稱碼。
- 62.如請求項 58 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一 Reed-Solomon 碼。
- 63.如請求項 58 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一 EVENODD 碼。
- 64.如請求項 58 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一 EVENODD 碼之一衍生。
- 65.如請求項 45 所述之資料儲存系統，其中每一儲存單元係一硬碟機、一揮發性隨機存取記憶裝置、一非揮發性隨機存取記憶裝置、一光學儲存裝置以及一磁帶儲存裝置之一。
- 66.一種資料儲存系統，包含：
- 具有複數個儲存單元之一陣列，該陣列不禁失效的特定型樣；以及
- 與該陣列耦合之一陣列控制器，該陣列控制器偵測該陣列之一第一儲存單元的一失效，選取該陣列之一第二儲存單元，使得在一重建作業後，該陣列之一失效容限被增加，以及從該第一儲存單元重建至少一部份的資訊至該第二儲存單元中。

- 67.如請求項 66 所述之資料儲存系統，其中在該陣列控制器選取該第二儲存單元前，該陣列之該最小漢明距離為  $d \geq 2$ 。
- 68.如請求項 66 所述之資料儲存系統，其中在從該第一儲存單元重建至少一部份的資訊至該第二儲存單元後，該陣列之該最小漢明距離被增加。
- 69.如請求項 66 所述之資料儲存系統，其中該第二儲存單元係基於該陣列之一失效型樣而被選取。
- 70.如請求項 66 所述之資料儲存系統，其中該第二儲存單元係基於一預定目標型樣而被選取。
- 71.如請求項 66 所述之資料儲存系統，其中該陣列包含基於一擦去或錯誤校正碼的冗餘。
- 72.如請求項 71 所述之資料儲存系統，其中該擦去或錯誤校正碼係一同位碼。
- 73.如請求項 71 所述之資料儲存系統，其中該擦去或錯誤校正碼係一 Winograd 碼。
- 74.如請求項 71 所述之資料儲存系統，其中該擦去或錯誤校正碼係一對稱碼。
- 75.如請求項 71 所述之資料儲存系統其中該擦去或錯誤校正碼係一 Reed-Solomon 碼。
- 76.如請求項 71 所述之資料儲存系統，其中該擦去或錯誤校正碼係一 EVENODD 碼。
- 77.如請求項 71 所述之資料儲存系統，其中該擦去或錯誤校正碼係一 EVENODD 碼之一衍生。
- 78.如請求項 66 所述之資料儲存系統，其中該陣列具有一不相同的失效機率。
- 79.如請求項 66 所述之資料儲存系統，其中每一儲存單元

係一硬碟機、一揮發性隨機存取記憶裝置、一非揮發性隨機存取記憶裝置、一光學儲存裝置以及一磁帶儲存裝置之一。

- 80.如請求項 66 所述之資料儲存系統，其中該陣列包含基於複數個擦去或錯誤校正碼的一乘績之冗餘。
- 81.如請求項 80 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一同位碼。
- 82.如請求項 80 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一 Winograd 碼。
- 83.如請求項 80 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一同位碼。
- 84.如請求項 80 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一對稱碼。
- 85.如請求項 80 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一 Reed-Solomon 碼。
- 86.如請求項 80 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一 EVENODD 碼。
- 87.如請求項 80 所述之資料儲存系統，其中至少一該擦去或錯誤校正碼係一 EVENODD 碼之一衍生。