I521359

(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 I521359 B**

(45)公告日： 中華民國 105 (2016) 年 02 月 11 日

(21)申請案號：100143929 (22)申請日： 中華民國 100 (2011) 年 11 月 30 日

(51)Int. Cl.： *G06F15/167 (2006.01)* *G06F11/07 (2006.01)*

(30)優先權：2011/09/23 中國大陸 201110285802.8

(71)申請人：阿里巴巴集團控股有限公司 (開曼群島) ALIBABA GROUP HOLDING LIMITED (KY)

香港

(72)發明人：李智慧 (CN) ；何坤 (CN) ；余俊 (CN) ；周異 (CN)

(74)代理人：林志剛

(56)參考文獻：

TW	I298128	TW	I352477
TW	I356310	US	5754781
US	2005/0108593A1		

審查人員：何旭智

申請專利範圍項數：10 項 圖式數：3 共 29 頁

(54)名稱

分散式儲存系統管理裝置及方法

(57)摘要

本發明公開了一種分散式儲存系統管理裝置，應用於包括 N 個儲存伺服器的分散式儲存系統，該裝置將其中 M 個儲存伺服器分為 x 個對等序列並形成 y 個虛擬節點組，且每個虛擬節點組中包括 z 個彼此屬於不同對等序列的儲存伺服器，其餘 N-M 個儲存伺服器為臨時儲存伺服器。本發明還相應公開了一種分散式儲存系統管理方法。本發明實施例提供的分散式儲存系統管理裝置及方法，藉由對等序列及虛擬節點組的雙重功能劃分對儲存伺服器進行管理，能夠保證資料備份在虛擬節點組中分屬不同對等序列的儲存伺服器中，從而在某個對等序列中的儲存伺服器失效時可以繼續由該虛擬節點組中其他對等序列的儲存伺服器提供資料讀寫服務。

指定代表圖：

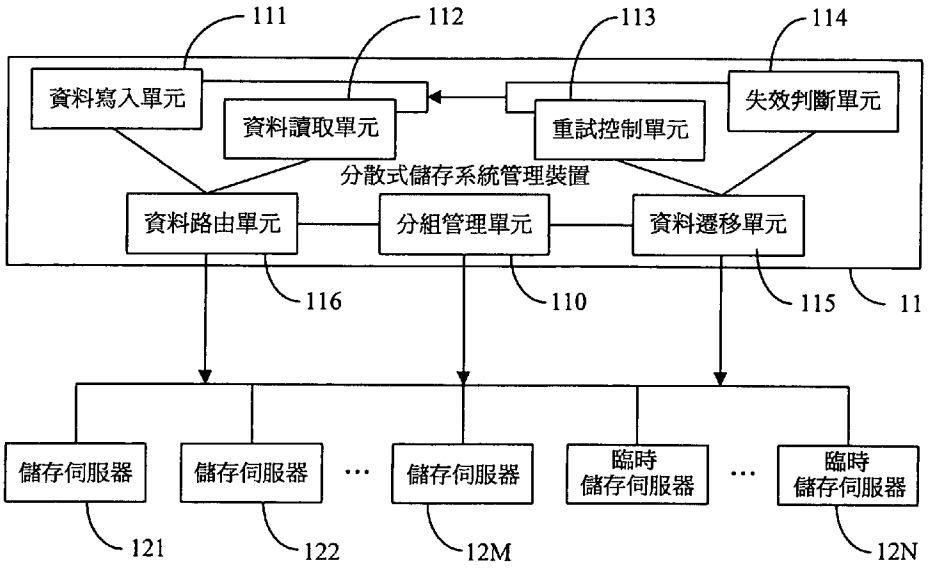

圖1

符號簡單說明：

11・・・分散式儲存系統管理裝置

12M・・・儲存伺服器

12N・・・儲存伺服器

110・・・分組管理單元

111・・・資料寫入單元

112・・・資料讀取單元

113・・・重試控制單元

114・・・失效判斷單元

115・・・資料遷移單元

116・・・資料路由單元

121・・・儲存伺服器

122・・・儲存伺服器

公告本

發明專利說明書

（本申請書格式、順序，請勿任意更動，※記號部分請勿填寫）

※申請案號：100143929

※申請日：100 年 11 月 30 日　　※IPC 分類：G06F 15/167 (2006.01)
G06F 11/07 (2006.01)

一、發明名稱：（中文／英文）

分散式儲存系統管理裝置及方法

二、中文發明摘要：

本發明公開了一種分散式儲存系統管理裝置，應用於包括 N 個儲存伺服器的分散式儲存系統，該裝置將其中 M 個儲存伺服器分爲 x 個對等序列並形成 y 個虛擬節點組，且每個虛擬節點組中包括 z 個彼此屬於不同對等序列的儲存伺服器，其餘 N-M 個儲存伺服器爲臨時儲存伺服器。本發明還相應公開了一種分散式儲存系統管理方法。本發明實施例提供的分散式儲存系統管理裝置及方法，藉由對等序列及虛擬節點組的雙重功能劃分對儲存伺服器進行管理，能夠保證資料備份在虛擬節點組中分屬不同對等序列的儲存伺服器中，從而在某個對等序列中的儲存伺服器失效時可以繼續由該虛擬節點組中其他對等序列的儲存伺服器提供資料讀寫服務。

三、英文發明摘要：

四、指定代表圖：

（一）本案指定代表圖為：第（1）圖。

（二）本代表圖之元件符號簡單說明：

11：分散式儲存系統管理裝置

12M：儲存伺服器

12N：儲存伺服器

110：分組管理單元

111：資料寫入單元

112：資料讀取單元

113：重試控制單元

114：失效判斷單元

115：資料遷移單元

116：資料路由單元

121：儲存伺服器

122：儲存伺服器

五、本案若有化學式時，請揭示最能顯示發明特徵的化學式：無

六、發明說明：

【發明所屬之技術領域】

本發明涉及分散式儲存技術領域，尤其涉及一種具有高可用儲存失效處理能力的分散式儲存系統管理裝置及方法。

【先前技術】

傳統的儲存系統採用集中的儲存伺服器存放所有資料，儲存伺服器成爲系統性能的瓶頸，也是可靠性和安全性的焦點，不能滿足大規模儲存應用的需要。分散式儲存系統採用可擴展的系統結構，藉由利用多台儲存伺服器分擔儲存負荷，並利用位置伺服器定位儲存資訊，不僅提高了系統的可靠性、可用性和存取效率，還使得後續的擴展更爲便利。

另一方面，在各種應用環境中，儲存系統中的資料都是寶貴的財富，各種儲存系統均會致力於保證所儲存的資料不因各種原因丟失。在分散式儲存系統中，個別儲存伺服器的宕機、停機維護或網路故障等問題都有可能導致資料的丟失，甚至可能會導致整個儲存系統的不可用，進而會影響到儲存系統所服務的應用系統的可用性。爲了避免這種狀況的發生，分散式儲存系統中通常採用以下兩種儲存失效處理機制，同時，這兩種機制也都各自存在一定的缺點。

第一種儲存失效處理機制中，是將兩個不同的物理儲

存伺服器配置爲主從伺服器，例如伺服器 A（主伺服器）和伺服器 B（從伺服器），正常情況下應用系統訪問伺服器 A 進行資料的讀取和寫入，並且寫入伺服器 A 的資料會同步至伺服器 B；一旦伺服器 A 發生宕機等故障，應用系統便切換至伺服器 B 進行資料讀寫；而在伺服器 A 恢復正常後，再將伺服器 B 的資料回遷到伺服器 A 上。該機制的缺點在於，首先，正常情況下，從主伺服器到從伺服器的資料同步也會存在延時，當主伺服器出現故障時，可能會導致小部分資料無法及時同步到從伺服器，從而出現資料丟失；其次，主伺服器從故障恢復正常後，需要將從伺服器的資料遷回主伺服器，由於此期間從伺服器的資料一直在不斷寫入，所以這個遷回過程會比較複雜；最後，一旦主伺服器、從伺服器同時都出現故障，便會導致儲存系統完全不可用。

另外一種新近出現的儲存失效處理機制中，應用系統在進行資料寫入時，是將同一個資料寫入到多個（例如 3 個）彼此不相關聯的伺服器上，只要成功寫入一定數量（例如 2 個）的伺服器即認定爲寫入操作完成；而應用系統在進行資料讀取時，便到上述多個伺服器上同時讀取，只要從一定數量（例如 2 個）的伺服器上讀取成功便認定爲讀取操作完成。由上述可以看出，該機制是藉由資料備份的數量來彌補可能的伺服器故障所造成的儲存失效，其缺點在於，應用系統的資料讀取和寫入操作都需要同時訪問多個伺服器，然而寫入操作並不能保證每次寫入的資料完

全不出差錯，而讀取操作便可能會讀取到不一致的資料，這樣應用系統在讀取資料時還需要進行資料校驗並修復不正確的資料，因此整個過程實現比較複雜且會導致系統性能變差。

【發明內容】

本發明的實施例旨在提供一種分散式儲存系統管理裝置及方法，以解決上述分散式系統中儲存失效處理機制所存在的問題。

爲實現上述目的，本發明的實施例提供了一種分散式儲存系統管理裝置，應用於包括 N 個儲存伺服器的分散式儲存系統，該裝置包括分組管理單元、資料寫入單元及資料讀取單元，其中，

所述分組管理單元用於將所述 N 個儲存伺服器中的 M 個儲存伺服器分爲 x 個對等序列並形成 y 個虛擬節點組，且每個虛擬節點組中包括 z 個彼此屬於不同對等序列的儲存伺服器，其餘 N-M 個儲存伺服器爲臨時儲存伺服器，上述 N、M、x、y 爲自然數常量且滿足：$N \geq 3$，$2 \leq M < N$，$x \geq 2$，$y \geq 1$，$x \cdot y \geq M$；z 爲自然數變數且滿足：$2 \leq z \leq x$；

所述資料寫入單元用於將資料寫入到選擇的一個虛擬節點組的每個儲存伺服器中，並在該虛擬節點組的部分儲存伺服器不可用時，將該資料寫入到該虛擬節點組剩餘可用的儲存伺服器以及所述臨時儲存伺服器中；

所述資料讀取單元用於從資料被寫入的虛擬節點組中任一可用的儲存伺服器處讀取該資料。

本發明的實施例還提供一種分散式儲存系統管理方法，應用於包括 N 個儲存伺服器的分散式儲存系統，並包括以下步驟：

S1. 將 M 個儲存伺服器分爲 x 個對等序列並形成 y 個虛擬節點組，且每個虛擬節點組中包括 z 個彼此屬於不同對等序列的儲存伺服器，其餘 N-M 個儲存伺服器爲臨時儲存伺服器，上述 N、M、x、y 爲自然數常量且滿足：$N \geq 3$，$2 \leq M < N$，$x \geq 2$，$y \geq 1$，$x \cdot y \geq M$；z 爲自然數變數且滿足：$2 \leq z \leq x$；

S21. 在執行資料寫入操作時將資料寫入到選擇的一個虛擬節點組的每個儲存伺服器中，並在該虛擬節點組的部分儲存伺服器不可用時，將該資料寫入到該虛擬節點組剩餘可用的儲存伺服器以及所述臨時儲存伺服器中；

S22. 在執行資料讀取操作時從資料被寫入的虛擬節點組中任一可用的儲存伺服器處讀取該資料。

由上述技術方案可知，本發明實施例提供的分散式儲存系統管理裝置及方法，藉由對等序列及虛擬節點組的雙重功能劃分對儲存伺服器進行管理，能夠保證資料備份在虛擬節點組中分屬不同對等序列的儲存伺服器中，從而在某個對等序列中的儲存伺服器失效時可以繼續由該虛擬節點組中其他對等序列的儲存伺服器提供資料讀寫服務。

【實施方式】

下面將詳細描述本發明的具體實施例。應當注意，這裏描述的實施例只用於舉例說明，並不用於限制本發明。

圖 1 爲本發明分散式儲存系統管理裝置實施例的結構示意圖，如圖所示，本實施例的分散式儲存系統管理裝置 11，用於對包括 N 個儲存伺服器 121~12N 的分散式儲存系統進行管理。分散式儲存系統管理裝置 11 進一步包括分組管理單元 110，其用於對上述分散式儲存系統中的 N 個儲存伺服器 121~12N 進行分組管理，具體包括：在儲存伺服器 121~12N 中，選取 M 個儲存伺服器劃分爲 x 個對等序列，每個對等序列中包含一定數量的儲存伺服器，在一個實施例中，每個對等序列包含的儲存伺服器的個數相同，當然每個對等序列中包含的伺服器的個數也可以不相同；接下來，繼續將上述所選取的 M 個儲存伺服器劃分爲 y 個虛擬節點組，且使得每個虛擬節點組中都包括 z 個彼此屬於不同對等序列的儲存伺服器；最後，將其餘 N-M 個儲存伺服器設置爲臨時儲存伺服器。上述各字母所表示的量中，N、M、x、y 爲自然數常量且其取值分別滿足：$N \geq 3$，$2 \leq M < N$，$x \geq 2$，$y \geq 1$，$x \cdot y \geq M$（“ ·爲乘法標記）；z 爲自然數變數且取值滿足：$2 \leq z \leq x$。在一個實施例中，z 可以是自然數常量，也即，使得每個虛擬節點組中儲存伺服器的個數相同，在這種情況下，爲了滿足每個虛擬節點組中各個儲存伺服器彼此屬於不同對等序列，實際上 x、y、M 的取值需要滿足 $M=x \cdot y$，而此時 $z=x$。

下面結合陣列的概念可以更好地理解本實施例中分組管理單元 110 以對等序列和虛擬節點組來對 M 個儲存伺服器進行雙重劃分的過程。例如，假設陣列 $a_{[x][y]}$ 如下：

$$a_{[x][y]}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1y} \\ a_{21} & \cdot & \cdot & \cdot \\ \vdots & \cdot & \cdot & \cdot \\ a_{x1} & \cdot & \cdot & a_{xy} \end{bmatrix}$$

結合以上陣列 $a_{[x][y]}$，可以將上述儲存伺服器的劃分過程看作將 M 個儲存伺服器填充至陣列 $a_{[x][y]}$ 中作爲陣列元素的過程。在 $M= x \cdot y$ 的情況下，陣列中的每個元素剛好與一個儲存伺服器對應，而每個虛擬節點組中儲存伺服器的個數 z 即爲常量且 $z=x$；而在 $M < x \cdot y$ 的情況下，在將 M 個儲存伺服器填充至陣列後，陣列中還會存在一定數量的空白元素，這時每個虛擬節點組中儲存伺服器的個數 z（對應於每一列中非空白元素的個數）就可能是不固定的，但在滿足 $x \geq 2$，$y \geq 1$，$2 \leq z \leq x$ 等條件下，這一情況應是允許的。

在一個實施例中，可以使得 x、y、M 的取值滿足 $M= x \cdot y$，即，使得每個對等序列（也即每個虛擬節點組）中儲存伺服器的個數均相等，這樣能夠得到更好的負載均衡效果，後續的失效恢復也更爲便利。爲使描述更爲簡便，下文中除非有特別說明，均是基於這種每個對等序列中儲存伺服器個數相同的情形。在這種情形下，以上文所示的陣列 $a_{[x][y]}$ 爲例，M 個儲存伺服器被劃分爲 x 個對等序列，例如，第 1 個和第 x 個對等序列所包含的儲存伺服器對

應的陣列元素分別是[a_{11}、a_{12}、...、a_{1y}]和[a_{x1}、a_{x2}、...、a_{xy}]；同時，這 M 個儲存伺服器又形成 y 個虛擬節點組，例如，第 1 個和第 y 個虛擬節點組所包含的儲存伺服器對應的陣列元素分別是[a_{11}、a_{21}、...、a_{x1}]和[a_{1y}、a_{2y}、...、a_{xy}]。

進一步如圖 1 所示，本實施例的分散式儲存系統管理裝置 11 還包括資料寫入單元 111 及資料讀取單元 112。其中，資料寫入單元 111 用於將資料寫入到任意選擇或按預設規則選擇的一個虛擬節點組的每個儲存伺服器中，並在該虛擬節點組的部分儲存伺服器不可用時，將該資料寫入到該虛擬節點組裏除去不可用儲存伺服器之外剩餘所有可用的儲存伺服器以及臨時儲存伺服器中；資料讀取單元 112 用於從資料被寫入的虛擬節點組中任一可用的儲存伺服器處讀取該資料。由上述可知，基於對等序列和虛擬節點組的資料讀寫操作，每個資料寫入時先選擇要寫入的虛擬節點組，再將該資料寫入所選擇的虛擬節點組中的每個儲存伺服器，這時實際上每個對等序列中都有儲存伺服器被寫入該同一資料，最終，如果不考慮故障狀況，在每個對等序列之間所有儲存伺服器儲存的資料應該是對等的（相同的），這也是所稱“對等序列”的含義所在；而在虛擬節點組中有儲存伺服器出現故障等不可用的情況時，本應寫入該不可用儲存伺服器的資料則轉而寫入到選擇的一個臨時儲存伺服器中（至於同時應寫入虛擬節點組中剩餘可用儲存伺服器的資料則不受影響繼續正常寫入），以便

後續儲存伺服器恢復可用後進行資料的遷回。另一方面，在進行資料的讀取操作時，只需從該資料被寫入的虛擬節點組中任意選取一個可用儲存伺服器進行讀取即可。在一個實施例中，爲了實現負載均衡，資料寫入的虛擬節點組可以是隨機進行選擇，後文實施例中還將介紹一種依據特定演算法進行隨機選擇的情況。

繼續如圖 1 所示，本實施例的分散式儲存系統管理裝置 11 還包括資料遷移單元 115、重試控制單元 113、失效判斷單元 114 及資料路由單元 116，下文將分別予以描述。

在一個實施例中，資料遷移單元 115 用於在不可用的儲存伺服器恢復可用時，將對應的臨時儲存伺服器中儲存的資料遷回該恢復可用的儲存伺服器；並用於在不可用的儲存伺服器無法恢復可用時，將該儲存伺服器所在的虛擬節點組裏可用的儲存伺服器中儲存的資料遷移至選擇的一個臨時儲存伺服器，並以該臨時儲存伺服器替換不可用的儲存伺服器。從此處描述可以看出，臨時儲存伺服器的設置有兩方面的功能。第一個方面，臨時儲存伺服器用於對儲存伺服器不可用期間寫入的資料進行臨時儲存，以便在儲存伺服器恢復可用時進行資料遷回；第二個方面，臨時儲存伺服器還作爲儲存伺服器的替換，用於在後者不可恢復時及時從相應虛擬節點組中其他對等序列的儲存伺服器遷移全部資料，進而替換不可恢復的儲存伺服器，這裏的“替換”是指從功能（相應的資料讀寫）到角色（相應所

屬的虛擬節點組和對等序列）的徹底替換。需要說明的是，上述臨時儲存伺服器兩個方面的功能是從整體上而言，但如果從個體考慮，爲了保證兩個方面的功能互不干擾，同時也考慮到這兩個方面功能所適用的資料儲存結構有所區別，較佳地，每個臨時儲存伺服器應當被設置爲始終只承擔單獨一個方面的功能；例如，在一個實施例中，所有臨時儲存伺服器在初始設置時就被劃分爲臨時伺服器和備份伺服器，其中臨時伺服器只承擔上述第一個方面的功能，而備份伺服器則只承擔上述第二個方面的功能。

在一個實施例中，重試控制單元 113 用於控制資料寫入單元 111 及資料讀取單元 112 在執行對應的資料寫入或讀取操作失敗時按第一預定次數重試該資料寫入或讀取操作。失效判斷單元 114 用於在重試控制單元 113 控制的重試達到第一預定次數時判斷對應的儲存伺服器爲不可用，並將該判斷結果通知資料寫入單元 111 及資料讀取單元 112；以及用於在儲存伺服器被判斷爲不可用後，利用重試控制單元 113 按第二預定次數檢測該儲存伺服器的狀態，在檢測爲可用時判斷該儲存伺服器恢復可用或者在檢測爲不可用並達到第二預定次數時判斷該儲存伺服器無法恢復可用，並將該判斷結果通知資料遷移單元 115。由上述可以看出，本發明的分散式儲存系統管理裝置實施例提供了暫態失效、臨時失效和永久失效三種情況下的處理機制。這裏，暫態失效是指由於網路瞬斷或其他原因導致的應用伺服器（應用程式）在極短時間內不能連接儲存伺服器

的狀況；爲此，本實施例藉由設置重試控制單元 113 來控制資料寫入單元 111 或資料讀取單元 112 重試一定次數的讀寫請求後即可恢復訪問，並且爲了不影響應用伺服器（應用程式）的性能及使用體驗，這裏的重試次數通常設置較小，而且每兩次重試之間可以不設置時間間隔。進一步，臨時失效是指由於升級或者維護等原因導致儲存伺服器臨時不可用，但經過適當時間後，儲存伺服器便能恢復提供服務；而永久失效是指由於硬碟損壞等極端原因導致的資料丟失而無法恢復的狀況；爲此，本實施例設置了失效判斷單元 114 的判斷和通知機制。首先，在重試控制單元 113 控制的重試次數達到第一預定次數後，失效判斷單元 114 便可以判定對應的儲存伺服器失效（不可用），並將判定結果通知資料寫入單元 111 及資料讀取單元 112，此後便如前文所述，資料寫入操作會在失效儲存伺服器所對應虛擬節點組剩餘所有可用的儲存伺服器以及臨時儲存伺服器中進行，資料讀取則僅在失效儲存伺服器所對應虛擬節點組中任一可用儲存伺服器中進行；之後，失效判斷單元 114 利用重試控制單元 113 按第二預定次數重複檢測該儲存伺服器的狀態，在檢測爲可用時（表明該儲存伺服器經過“臨時失效”後恢復可用）便可以通知資料遷移單元 115 開始進行從臨時儲存伺服器到恢復可用儲存伺服器的資料遷回操作，在失效判斷單元 114 檢測到對應儲存伺服器爲不可用並達到第二預定次數時（表明該儲存伺服器爲“永久失效”），便可以通知資料遷移單元 115 開始進行

從失效儲存伺服器所在虛擬組中的可用儲存伺服器到臨時儲存伺服器的全部資料複製操作，最終該臨時儲存伺服器將取代永久失效的儲存伺服器。

需要說明的是，在一個實施例中，上述重試控制單元 113 按第二預定次數重複檢測該儲存伺服器的狀態時，具體仍然可以像之前暫態失效的重試時那樣，藉由控制資料寫入單元 111 或資料讀取單元 112 不斷重試資料的寫入或讀取操作，並以重試是否成功來判斷儲存伺服器的狀態是否恢復可用；在一個實施例中，之前暫態失效時按第一預定次數進行讀寫重試的資料可以就是原本要進行讀寫但未讀寫成功的正常資料，但之後按第二預定次數進行讀寫重試的資料可以是預先專門設置的測試資料，這是爲了避免因正常資料可能較大而影響後一重試過程的執行效率。另外，爲了使“永久失效”的判定維持嚴格標準，與前述第一預定次數相比，這裏的第二預定次數通常設置較大，而且每兩次重試之間可以設置一定的時間間隔；例如，可以在一天內按數秒的間隔不斷重複檢測，此時最終的第二預定次數可能會達到幾萬甚至幾十萬次。另外，由上述可知，臨時失效和永久失效通常伴隨著一些人爲可以控制（例如升級、維護）或可以檢測（例如硬碟故障）的因素，因此，在一個實施例中，失效判斷單元 114 也可以借助人爲的手段來判定臨時失效的恢復和永久失效，例如可以藉由人工手動輸入在失效判斷單元 114 管理的儲存伺服器狀態表中將某個儲存伺服器的狀態修改爲“臨時失效”、“可

用”、“永久失效”等等。

在一個實施例中，資料路由單元 116 用於確定資料寫入單元 111 寫入資料以及資料讀取單元讀取資料時所選擇的虛擬節點組；在一個實施例中，爲了提高路由效率，資料路由單元 116 可以按照前文所述陣列 $a_{[x][y]}$的方式對儲存伺服器加以管理。進一步，資料路由單元 116 還包括路由選擇子單元 1161，路由選擇子單元 1161 用於按以下過程進行虛擬節點組的選擇：將上述 y 虛擬節點組分別賦以 0、1、...、y-1 的編號進行管理；將待寫入的資料的哈希值對 y 取模，得到[0,...,y-1]範圍內的一個值，選擇與該值對應編號的虛擬節點組作爲資料寫入單元寫入該資料的虛擬節點組；同理，將待讀取的資料的哈希值對 y 取模，得到[0,...,y-1]範圍內的一個值，選擇與該值對應編號的虛擬節點組作爲資料讀取單元讀取該資料的虛擬節點組。這裏，基於 kv 儲存的原理，由於哈希值的唯一性，同一條資料在進行寫入時由資料路由單元 116 所選擇的虛擬節點組，與該條資料在進行讀取時由資料路由單元 116 所選擇的虛擬節點組，會絕對保持一致，由此便能夠保證被隨機寫入到某個虛擬節點組的資料能夠被準確地在同一虛擬節點組中讀取到。進一步，藉由上述過程可知，每條資料會隨機分配到一個虛擬節點組並被寫入該虛擬節點組裏屬於每個對等序列的儲存伺服器中，由此可以看出，從虛擬節點組的劃分角度，是保證了分散式儲存系統的負載均衡；從對等序列的劃分角度，是保證了資料在所有對等序列

中的備份。另外需要說明的是，上述根據取模演算法選取虛擬節點組的過程只是示例，在不考慮儲存的其他特性（例如資料分類、用戶分組等等）時可以使用，而一般儲存實現會使用更複雜的演算法，儘管如此，該細節並不影響本發明的理解和實施，此處不再加以贅述。

在一個實施例中，上述分散式儲存系統管理裝置 11 中的資料寫入單元 111、資料讀取單元 112 及資料路由單元 116，可以設置在應用伺服器上作爲儲存系統的用戶端使用；而分組管理單元 110、資料遷移單元 115、重試控制單元 113 及失效判斷單元 114 則可以單獨設置在一個區別於儲存伺服器和應用伺服器的管理伺服器中，作爲儲存伺服器與應用伺服器之間的仲裁仲介發揮作用。可以看出，這種具體位置上的設置並不影響上述分散式儲存系統管理裝置 11 中各單元的功能劃分，因此並不影響本發明的理解和實施，此處不再加以贅述，圖中也未例示。

圖 2 爲本發明分散式儲存系統管理方法的實施例一流程圖，本實施例的分散式儲存系統管理方法應用於包括 N 個儲存伺服器的分散式儲存系統，如圖 2 所示，其包括以下步驟：

S201、將 M 個儲存伺服器分爲 x 個對等序列並形成 y 個虛擬節點組，且每個虛擬節點組中包括 z 個彼此屬於不同對等序列的儲存伺服器，其餘 N-M 個儲存伺服器爲臨時儲存伺服器；

具體而言，本步驟首先是在 N 個儲存伺服器中，選

取 M 個儲存伺服器劃分爲 x 個對等序列，每個對等序列中包含一定數量的儲存伺服器。在一個實施例中，每個對等序列包含的儲存伺服器的個數相同。接下來，繼續將上述所選取的 M 個儲存伺服器劃分爲 y 個虛擬節點組，且使得每個虛擬節點組中都包括 z 個彼此屬於不同對等序列的儲存伺服器。最後，將其餘 N-M 個儲存伺服器設置爲臨時儲存伺服器。上述各字母所表示的量中，N、M、x、y 爲自然數常量且其取值分別滿足：$N \geq 3$，$2 \leq M < N$，$x \geq 2$，$y \geq 1$，$y \cdot x \geq M$（“ ·爲乘法標記）；z 爲自然數變數且取值滿足：$2 \leq z \leq x$。在一個實施例中，z 可以是自然數常量，也即，使得每個虛擬節點組中儲存伺服器的個數相同，在這種情況下，爲了滿足每個虛擬節點組中各個儲存伺服器彼此屬於不同對等序列，實際上 x、y、M 的取值需要滿足 $M = x \cdot y$，而此時 $z = x$。

S202、在執行資料寫入操作時將資料寫入到選擇的一個虛擬節點組的每個儲存伺服器中，並在該虛擬節點組的部分儲存伺服器不可用時，將該資料寫入到該虛擬節點組剩餘所有可用的儲存伺服器以及選擇的一個臨時儲存伺服器中；

S203、在執行資料讀取操作時從資料被寫入的虛擬節點組中任一可用的儲存伺服器處讀取該資料。

從步驟 S202~ S203 可以看出，對於在對等序列和虛擬節點組雙重功能劃分基礎上的資料讀寫操作，每個資料寫入時先選擇要寫入的虛擬節點組，再將該資料寫入所選

擇的虛擬節點組中的每個儲存伺服器，這時實際上每個對等序列中都有儲存伺服器被寫入該資料；而在虛擬節點組中有儲存伺服器出現故障等不可用的情況時，本應寫入該不可用儲存伺服器的資料則轉而寫入到選擇的一個臨時儲存伺服器中（至於同時應寫入虛擬節點組中剩餘可用儲存伺服器的資料則不受影響繼續正常寫入），以便後續儲存伺服器恢復可用後進行資料的遷回。另一方面，在進行資料的讀取操作時，只需從該資料被寫入的虛擬節點組中任意選取一個可用儲存伺服器進行讀取即可。在一個實施例中，爲了實現負載均衡，資料寫入的虛擬節點組可以是隨機進行選擇，後文實施例中還將介紹一種依據特定演算法進行隨機選擇的情況。

圖 3 爲本發明分散式儲存系統管理方法實施例二的流程圖，本實施例中將描述暫態失效的處理機制，以及臨時失效和永久失效的判斷機制和處理機制，如圖所示，本實施例的方法包括以下步驟：

S301、在執行資料讀寫操作失敗時重試該資料讀寫操作，重試成功則進行下一條資料讀寫操作，失敗則繼續步驟 S302；

S302、判斷步驟 S301 重試是否達到第一預定次數，如果是則繼續步驟 S303，否則轉回步驟 S301 繼續重試；

步驟 S301~S302 可以看作是暫態失效的處理機制，這時通常是由於網路瞬斷或其他原因導致的應用伺服器（應用程式）在極短時間內不能連接儲存伺服器，爲此，本實

施例的方法藉由重試一定次數的讀寫請求後即可恢復訪問，並且爲了不影響應用伺服器（應用程式）的性能及使用體驗，這裏重試的第一預定次數通常設置較小，而且每兩次重試之間通常不設置時間間隔。

S303、判斷對應的儲存伺服器爲不可用，並繼續步驟S304；

在一個實施例中，本步驟的判斷結果可以通知給相關的資料讀寫操作單元，此後便如前文所述，資料寫入操作會在失效儲存伺服器所對應虛擬節點組剩餘所有可用的儲存伺服器以及臨時儲存伺服器中進行，資料讀取則僅在失效儲存伺服器所對應虛擬節點組中任一可用的儲存伺服器中進行。

S304、重複檢測該儲存伺服器的狀態，檢測爲可用時繼續步驟S305，否則轉入步驟S306；

S305、將對應的臨時儲存伺服器中儲存的資料遷回該恢復可用的儲存伺服器；

經過步驟S304，重複檢測到失效儲存伺服器爲可用時，可以認爲是表明相應儲存伺服器經過“臨時失效”後恢復可用，此時通常是伺服器的升級和維護完成等情況，接下來便可以開始進行從臨時儲存伺服器到恢復可用儲存伺服器的資料遷回操作。由於臨時儲存伺服器中僅儲存失效儲存伺服器失效期間所寫入的資料，因此上述資料遷回操作會十分簡便、迅速。

S306、判斷步驟S304的重複檢測是否達到第二預定

次數，如果是則繼續步驟 S307，否則回到步驟 S304 繼續重複檢測。

S307、將該儲存伺服器所在的虛擬節點組裏可用的儲存伺服器中儲存的資料遷移至選擇的一個臨時儲存伺服器，並以該臨時儲存伺服器替換不可用的儲存伺服器；

經過步驟 S304 和 S305，在重複檢測到失效儲存伺服器爲不可用且達到第二預定次數時，可以認爲是表明該儲存伺服器爲“永久失效”，此時通常是發生了硬碟損壞等極端情況，接下來便可以開始進行從失效儲存伺服器所在虛擬組中的可用儲存伺服器到臨時儲存伺服器的全部資料複製操作，並且最終以該臨時儲存伺服器取代永久失效的儲存伺服器。在一個實施例中，爲了使“永久失效”的判定維持嚴格標準，與前述的第一預定次數相比，這裏的第二預定次數通常設置較大，而且每兩次重試之間可以設置一定的時間間隔，例如，可以在一天內按數秒的間隔不斷重複檢測，此時最終的第二預定次數可能會達到幾萬甚至幾十萬次。另外，在一個實施例中，也可以借助人爲介入的手段分別在步驟 S304、S306 中實現臨時失效恢復和永久失效的判定。

由上述技術方案可知，本發明實施例提供的分散式儲存系統管理裝置及方法，藉由對等序列及虛擬節點組的雙重功能劃分對儲存伺服器進行管理，在正常情況下，資料可以寫入在虛擬節點組裏分屬於每個對等序列的儲存伺服器中，而讀取時只需在該虛擬節點組裏任一可用的儲存伺

服器上進行；在某個儲存伺服器失效期間，資料的讀取仍然只需在對應虛擬節點組裏任一可用的儲存伺服器上進行，而寫入的資料會同時備份到該虛擬節點組裏所有可用的儲存伺服器以及臨時儲存伺服器上；當儲存伺服器從臨時失效狀態中恢復時，便可以從前述臨時儲存伺服器上遷回失效期間所寫入的資料；當儲存伺服器因永久失效而無法恢復時，還可以將對應虛擬節點組裏任一可用的儲存伺服器上的全部資料複製到臨時儲存伺服器上，再以該臨時儲存伺服器取代失效的儲存伺服器。

雖然已參照幾個典型實施例描述了本發明，但應當理解，所用的術語是說明和示例性、而非限制性的術語。由於本發明能夠以多種形式具體實施而不脫離發明的精神或實質，所以應當理解，上述實施例不限於任何前述的細節，而應在申請專利範圍第所限定的精神和範圍內廣泛地解釋，因此落入申請專利範圍第或其等效範圍內的全部變化和修改都應爲申請專利範圍第所涵蓋。

【圖式簡單說明】

圖 1 爲本發明分散式儲存系統管理裝置實施例的結構示意圖；

圖 2 爲本發明分散式儲存系統管理方法實施例一的流程圖；

圖 3 爲本發明分散式儲存系統管理方法實施例二的流程圖。

【主要元件符號說明】

11：分散式儲存系統管理裝置

12M：儲存伺服器

12N：儲存伺服器

110：分組管理單元

111：資料寫入單元

112：資料讀取單元

113：重試控制單元

114：失效判斷單元

115：資料遷移單元

116：資料路由單元

121：儲存伺服器

122：儲存伺服器

七、申請專利範圍：

1.一種分散式儲存系統管理裝置，應用於包括 N 個儲存伺服器的分散式儲存系統，該裝置包括分組管理單元、資料寫入單元及資料讀取單元，其中，

該分組管理單元用於將該 N 個儲存伺服器中的 M 個儲存伺服器分爲 x 個對等序列並形成 y 個虛擬節點組，且每個虛擬節點組中包括 z 個彼此屬於不同對等序列的儲存伺服器，其餘 N-M 個儲存伺服器爲臨時儲存伺服器，上述 N、M、x、y 爲自然數常數且滿足：$N \geq 3$，$2 \leq M < N$，$x \geq 2$，$y \geq 1$，$x \cdot y \geq M$；z 爲自然數變數且滿足：$2 \leq z \leq x$；

該資料寫入單元用於將資料寫入到選擇的一個虛擬節點組的每個儲存伺服器中，並在該虛擬節點組的部分儲存伺服器不可用時，將該資料寫入到該虛擬節點組剩餘可用的儲存伺服器以及該臨時儲存伺服器中；

該資料讀取單元用於從資料被寫入的虛擬節點組中任一可用的儲存伺服器處讀取該資料。

2.如申請專利範圍第 1 項之分散式儲存系統管理裝置，其中，還包括資料遷移單元，該資料遷移單元用於在不可用的儲存伺服器恢復可用時，將對應的臨時儲存伺服器中儲存的資料遷回該恢復可用的儲存伺服器；並用於在不可用的儲存伺服器無法恢復可用時，將該儲存伺服器所在的虛擬節點組裡可用的儲存伺服器中儲存的資料遷移至選擇的一個臨時儲存伺服器，並以該臨時儲存伺服器替換該

不可用的儲存伺服器。

3.如申請專利範圍第 2 項之分散式儲存系統管理裝置，其中，還包括重試控制單元及失效判斷單元，

該重試控制單元用於控制該資料寫入單元及該資料讀取單元在執行對應的資料寫入或讀取操作失敗時按第一預定次數重試該資料寫入或讀取操作；

該失效判斷單元用於在該重試控制單元控制的重試達到該第一預定次數時判斷對應的儲存伺服器爲不可用，並將該判斷結果通知該資料寫入單元及該資料讀取單元；以及用於在儲存伺服器被判斷爲不可用後，利用該重試控制單元按第二預定次數重複檢測該儲存伺服器的狀態，在檢測爲可用時判斷該儲存伺服器恢復可用或者在檢測爲不可用達到該第二預定次數時判斷該儲存伺服器無法恢復可用，並將該判斷結果通知該資料遷移單元。

4.如申請專利範圍第 1 項之分散式儲存系統管理裝置，其中，還包括資料路由單元，該資料路由單元用於確定該資料寫入單元寫入資料以及該資料讀取單元讀取資料時所選擇的虛擬節點組。

5.如申請專利範圍第 4 項之分散式儲存系統管理裝置，其中，該資料路由單元還包括路由選擇子單元，該路由選擇子單元用於分別賦以 0、1、...、y-1 的編號管理該 y 個虛擬節點組，並根據將待寫入的資料的雜湊值對 y 取模得到的值選擇與該值對應編號的虛擬節點組作爲該資料寫入單元寫入該資料的虛擬節點組，以及根據將待讀取的資

料的雜湊值對 y 取模得到的值選擇與該值對應編號的虛擬節點組作爲該資料讀取單元讀取該資料的虛擬節點組。

6.如申請專利範圍第 1 項之分散式儲存系統管理裝置，其中，x、y、M 的取值滿足 M= x·y。

7.一種分散式儲存系統管理方法，應用於包括 N 個儲存伺服器的分散式儲存系統，其中，該方法包括以下步驟：

S1. 將 M 個儲存伺服器分爲 x 個對等序列並形成 y 個虛擬節點組，且每個虛擬節點組中包括 z 個彼此屬於不同對等序列的儲存伺服器，其餘 N-M 個儲存伺服器爲臨時儲存伺服器，上述 N、M、x、y 爲自然數常數且滿足：$N \geq 3$，$2 \leq M < N$，$x \geq 2$，$y \geq 1$，$x \cdot y \geq M$；z 爲自然數變數且滿足：$2 \leq z \leq x$；

S21. 在執行資料寫入操作時將資料寫入到選擇的一個虛擬節點組的每個儲存伺服器中，並在該虛擬節點組的部分儲存伺服器不可用時，將該資料寫入到該虛擬節點組剩餘可用的儲存伺服器以及該臨時儲存伺服器中；

S22. 在執行資料讀取操作時從資料被寫入的虛擬節點組中任一可用的儲存伺服器處讀取該資料。

8.如申請專利範圍第 7 項之分散式儲存系統管理方法，其中，還包括以下步驟：

S31. 在不可用的儲存伺服器恢復可用時，將對應的臨時儲存伺服器中儲存的資料遷回該恢復可用的儲存伺服器；

S32. 在不可用的儲存伺服器無法恢復可用時，將該儲存伺服器所在的虛擬節點組裏可用的儲存伺服器中儲存的資料遷移至選擇的一個臨時儲存伺服器，並以該臨時儲存伺服器替換該不可用的儲存伺服器。

9.如申請專利範圍第 8 項之分散式儲存系統管理方法，其中，還包括以下步驟：

S41. 在執行資料寫入或讀取操作失敗時重試該資料寫入或讀取操作；

S42. 在步驟 S41 重試達到第一預定次數時判斷對應的儲存伺服器爲不可用；

S43. 在儲存伺服器被判斷爲不可用後，重複檢測該儲存伺服器的狀態，在檢測爲可用時判斷該儲存伺服器恢復可用並轉步驟 S31，在檢測爲不可用且達到第二預定次數時判斷該儲存伺服器無法恢復可用並轉步驟 S32。

10.如申請專利範圍第 7 項之分散式儲存系統管理方法，其中，還包括以下步驟：

分別賦以 0、1、...、y-1 的編號管理該 y 個虛擬節點組，並在步驟 S21 執行資料寫入操作時，根據將待寫入的資料的雜湊值對 y 取模得到的值選擇與該值對應編號的虛擬節點組作爲寫入該資料的虛擬節點組；以及在步驟 S22 執行資料讀取操作時，根據將待讀取的資料的雜湊值對 y 取模得到的值選擇與該值對應編號的虛擬節點組作爲讀取該資料的虛擬節點組。

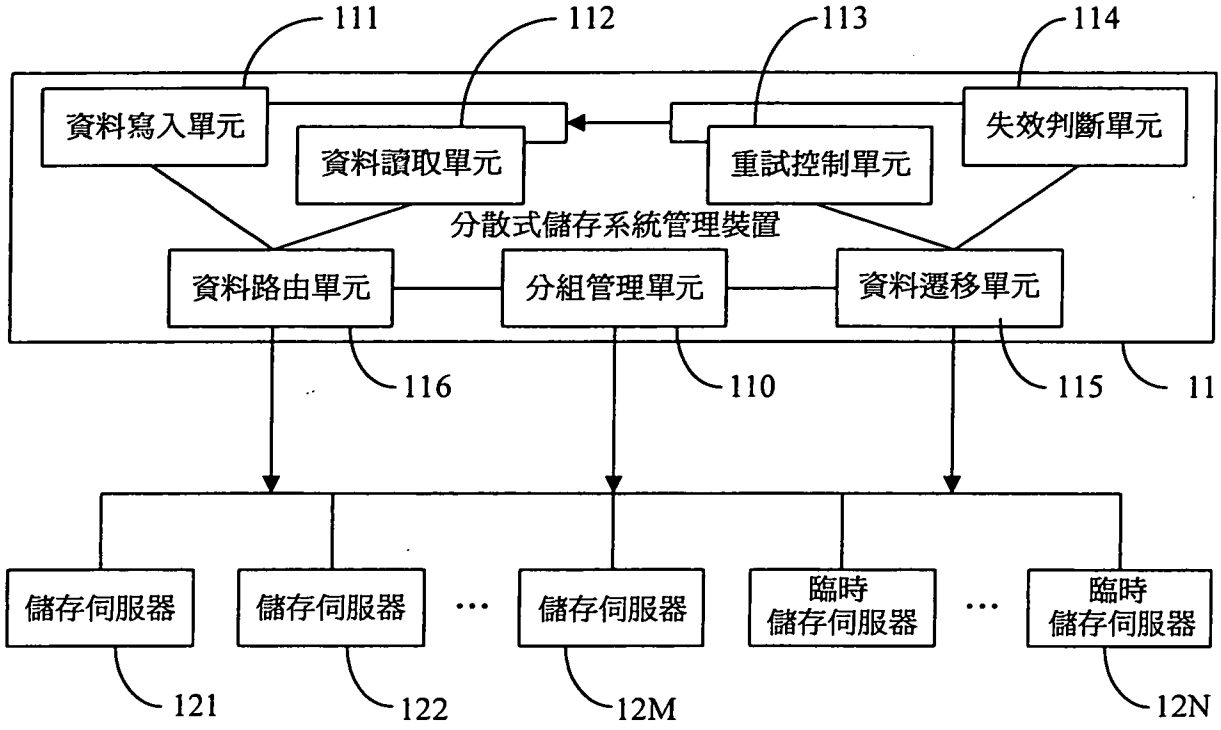
111
112
113
114
資料寫入單元
資料讀取單元
重試控制單元
失效判斷單元
分散式儲存系統管理裝置
資料路由單元
分組管理單元
資料遷移單元
116
110
115
11
儲存伺服器
儲存伺服器
儲存伺服器
臨時
儲存伺服器
臨時
儲存伺服器
121
122
12M
12N

圖1

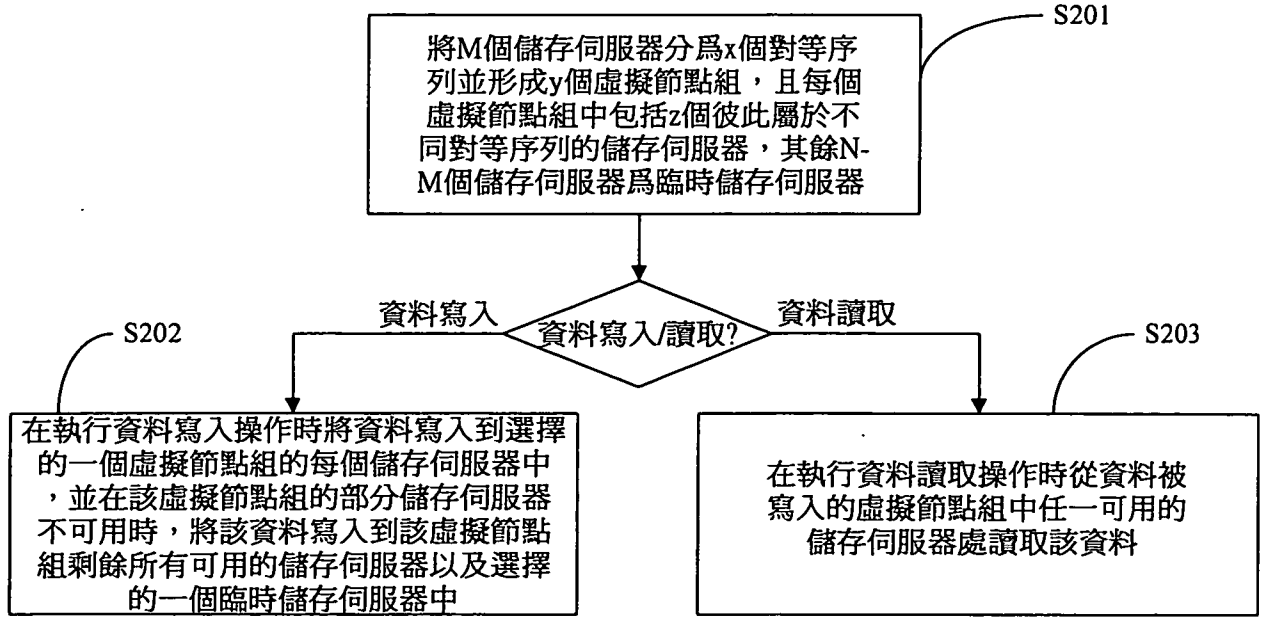
將M個儲存伺服器分爲x個對等序列並形成y個虛擬節點組，且每個虛擬節點組中包括z個彼此屬於不同對等序列的儲存伺服器，其餘N-M個儲存伺服器爲臨時儲存伺服器
S201
資料寫入/讀取?
資料寫入
資料讀取
S202
S203
在執行資料寫入操作時將資料寫入到選擇的一個虛擬節點組的每個儲存伺服器中，並在該虛擬節點組的部分儲存伺服器不可用時，將該資料寫入到該虛擬節點組剩餘所有可用的儲存伺服器以及選擇的一個臨時儲存伺服器中
在執行資料讀取操作時從資料被寫入的虛擬節點組中任一可用的儲存伺服器處讀取該資料

圖2

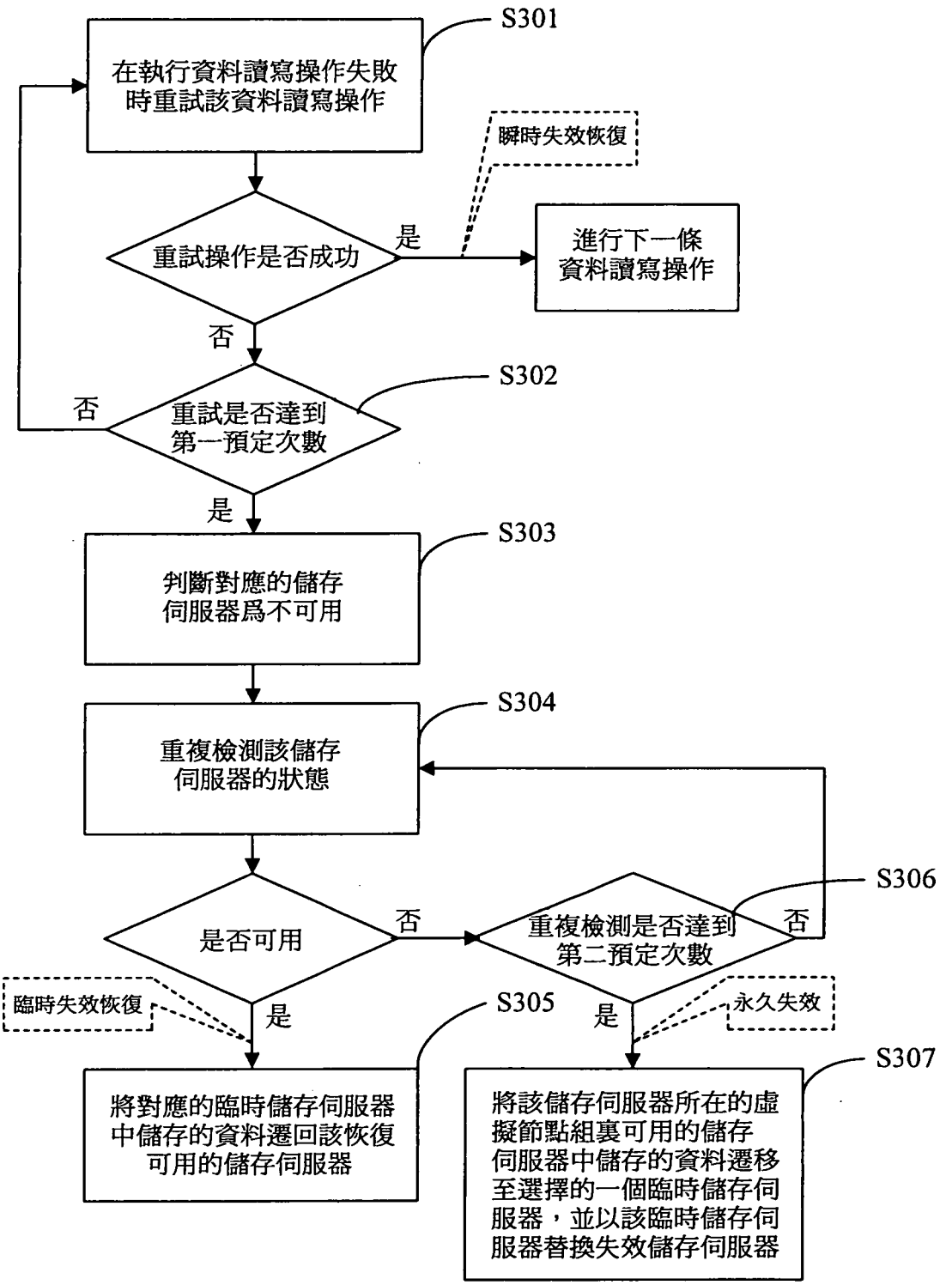
S301
在執行資料讀寫操作失敗時重試該資料讀寫操作
瞬時失效恢復
重試操作是否成功
是
進行下一條資料讀寫操作
否
S302
否
重試是否達到第一預定次數
是
S303
判斷對應的儲存伺服器爲不可用
S304
重複檢測該儲存伺服器的狀態
是否可用
否
S306
重複檢測是否達到第二預定次數
否
臨時失效恢復
是
S305
是
永久失效
S307
將對應的臨時儲存伺服器中儲存的資料遷回該恢復可用的儲存伺服器
將該儲存伺服器所在的虛擬節點組裏可用的儲存伺服器中儲存的資料遷移至選擇的一個臨時儲存伺服器，並以該臨時儲存伺服器替換失效儲存伺服器

圖3