

一、本案已向

國家(地區)申請專利

申請日期

案號

主張專利法第二十四條第一項優先權

美國 US

2002/02/15

10/077,129

有

二、☐主張專利法第二十五條之一第一項優先權：

申請案號：

無

日期：

三、主張本案係符合專利法第二十條第一項☐第一款但書或☐第二款但書規定之期間

日期：

四、☐有關微生物已寄存於國外：

寄存國家：

寄存機構：

寄存日期：

寄存號碼：

無

☐有關微生物已寄存於國內(本局所指定之寄存機構)：

寄存機構：

寄存日期：

寄存號碼：

無

☐熟習該項技術者易於獲得,不須寄存。



五、發明說明 (1)

本專利申請案係與以下發明人同時申請之美國專利申請案相關：美國專利申請號10/077,371(代理人登錄號POU920020009US1)、美國專利申請號10/077,345(代理人登錄號POU920020010US1)、美國專利申請號10/077,320(代理人登錄號POU920020011US1)、美國專利申請號10/077,246(代理人登錄號POU920020012US1)、美國專利申請號10/077,201(代理人登錄號POU920020013US1)，所有申請案之申請日為2002年2月15日，且均讓與給國際商業機器股份有限公司(IBM)。

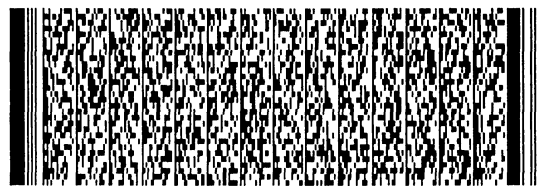
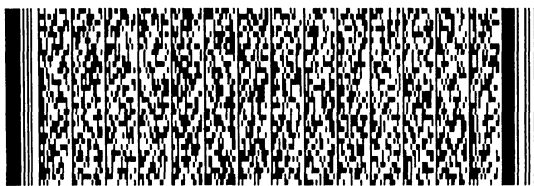
一、【發明所屬之技術領域】

本發明係關於資料儲存系統，尤其是關於儲存快照之資料儲存系統(即特定時點所儲存資料狀態之表徵)。

二、【先前技術】

許多資料儲存應用係將資料儲存在機電系統中，而這樣的應用是很容易於發生物理上失敗的，磁碟機就是這種儲存系統之例子。磁碟機使用具有一讀取/寫入頭十分接近地停留在轉動磁盤上，資料之儲存係藉由產生一磁性記錄於磁盤上。在磁盤表面上的污損經常會損壞磁性表面以及記錄，進而導致無法存取資料。其它資料儲存系統則會受到物理或電性損壞，進而可能遺失資料。

許多資料儲存系統係根據檔案隱喻(file metaphor)來組織被儲存的資料。在這些儲存系統中，相

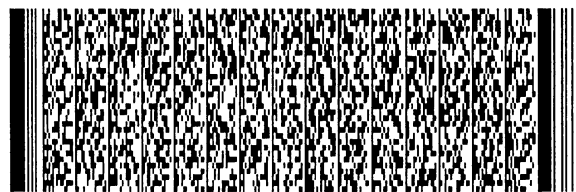
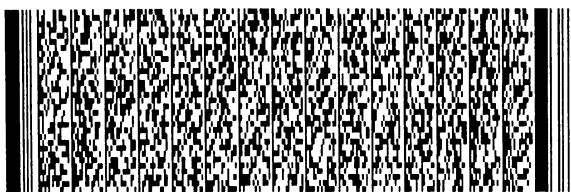


五、發明說明 (2)

關的資料被儲存在一檔案中，而資料儲存系統則會存放數個檔案，接著存放指向那些檔案之指標 (reference)，供存取在那些檔案中的資料。單一檔案可以存放在資料儲存裝置中之連續位置，或存放在不同位置。當一個大的資料檔案需要儲存，而資料儲存裝置已儲存許多檔案時，這個大的資料檔案必須被切割存放在未被佔用的空間，這種將資料存放在資料儲存裝置中之不同位置的情況經常會發生。若有額外的資料加入一個已經存在的檔案時，資料也經常存放在不同的位置。將資料儲存裝置上之儲存資料組合成檔案及檔案結構便稱為檔案系統。

資料儲存系統通常儲存目前存放在檔案系統中之資料的影像 (image) 或快照 (snapshot)。快照的資料內容為存放在現正開啟 (active) 的檔案系統中之資料，或者是在擷取快照當時之先前快照。快照的其中一個用途係用來將檔案系統的狀態存放在其他儲存系統上，例如其他磁碟機或磁帶儲存系統。另一個用途則用來重新產生已經被刪除的資料，即存取已經被刪除或更新的檔案之先前版本。

存放在檔案系統之檔案中之資料具有相關的超資料 (metadata)，以描述資料並且允許資料的存取。有些現有用來取得檔案系統快照的方法直到原始系統中的資料被修改，才實際上複製原始檔案系統中的資料至其快照。這樣的系統被稱為「寫入時複製 (copy-on-write)」系統，因為在這個系統中，在原始資料執行寫入時，才會被複製至快照資料。

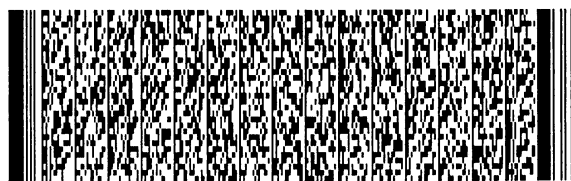
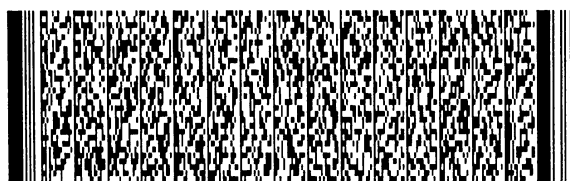


五、發明說明 (3)

現有的「寫入時複製」系統彼此間不同處在於如何及何時複製超資料。現有的快照系統會將某些或所有的超資料複製至快照資料檔案中，其中超資料用來描述在產生快照時，資料檔案所儲存的位置。這些現有的系統會產生快照資料集合，其包含指向原始檔案系統中之原始資料檔案的檔案指標，這會產生原始檔案系統中指向相同資料區塊之數個指標、原始檔案系統之超資料中之指標、以及每一快照資料集合中之指標。

圖4顯示一個示範性的檔案系統資料結構400，其包含二個資料節點(inode)，其中一個在現正開啟的檔案系統中，而另外一個則在快照資料集合中。每一資料節點均指向相同的資料區塊406。在原始檔案系統中，存在數個指向單一資料區塊之指標會影響原始檔案系統的需求。在使用快照之檔案系統中，每一快照存有一指向原始資料區塊之指標，必須維持每一指向該資料區塊之指標的表徵，以判斷資料區塊是處於使用中或空閒的狀態。

在不具有數個指標時，單一個位元即可以表示資料區塊是處於使用中或空閒的狀態。而具有數個指標時，則需要數個位元以追蹤數個指標，並且確保在宣告資料區塊為空閒前，沒有任何指向該資料區塊的指標存在。這個必須追蹤數個指標之需要將會使得檔案系統的操作複雜化，並限制快照的總數量，同時也複雜化這種快照系統之執行，或者使得這種快照系統變得無法執行，如果這種快照系統為不支援追蹤數個指向一資料區塊之指標的檔案系統。



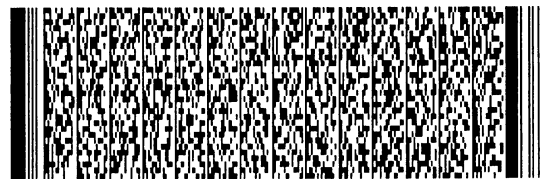
五、發明說明 (4)

因此，必須克服上述習知技術之問題，並且特別需要提供一個方法，更有效率地使用資料處理裝置中之系統核心記憶體，以支援對時間敏感之處理任務，例如一外部資料通訊處理。

三、【發明內容】

簡言之，本發明係關於一種系統、方法以及電腦可讀取媒體，供提供一檔案系統子集之快照。在本發明之一實施例中，此方法包含在一檔案系統中一組原始檔之第一快照之產生。此第一快照包含一資料節點(inode)，對應在此第一組檔案之每一原始檔。儲存在每一資料節點的是有關第一組檔案之一第一識別符及有關第一快照時間之一第二識別符。每隔一第一期間取得隨後第一組檔案之快照。接著，取得第二組原始檔之第二快照，此第二快照包含一資料節點，對應在此第二組檔案之每一原始檔。儲存在每一資料節點的是有關第二組檔案之一識別符及有關第二快照時間之一識別符。每隔一第二期間取得隨後第二組檔案之快照。此第一組檔案與此第二組檔案不相同，此第一期間與此第二期間不相等。

本發明之另一實施例為，接收重存一快照之一請求。此請求的快照被擷取，為在快照中之每一資料節點，決定在檔案系統中之對應資料節點是否可用。若可用，在快照中對應此資料節點之資料塊(data block)重存至檔案系統中之對應資料節點。否則，在快照中對應此資料節點之資



五、發明說明 (5)

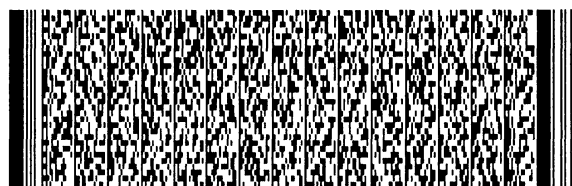
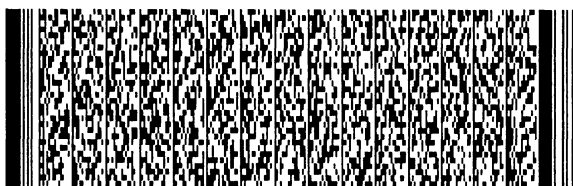
料塊重存至檔案系統中一未分配之資料節點。

以下將配合圖式，詳細描述本發明之較佳實施例，以突顯本發明之特徵和優點。

四、【實施方式】

為了克服習知技術之問題，本發明依據較佳實施例，提供一種系統與方法，用來有效地取得檔案系統資料之快照，並且將這些快照運用於不同的用途。本發明之實施例所產生之快照資料集合，係利用明示的邏輯指標 (inferred logical reference)，而非物理指標 (physical reference) 來指向原始資料區塊。

在以下之圖式中，相同的編號係表示相同的部分。圖1為本發明實施例之整體系統結構100之方塊圖。本發明實施例係操作於一用來存放一個或多個資料檔案之檔案系統102中，或與其連結操作。本發明之實施例取得並且保留一個或多個快照資料集合104，以下將詳細描述。具有這種檔案系統102之電腦或客戶資訊處理系統係經由網路106連接至其它電腦和資料處理系統。本發明實施例之一個應用為支援一資料儲存系統，以有效地處理備份資料。一個示範性的備份系統108係顯示於示範性整體系統結構100中，示範性備份系統108用來維持一備份，其包含檔案系統102中所有資料之副本。快照104的一個用途係為有效地通訊和儲存備份資料集合於遠端備份系統，如備份系統108。利用本發明實施例所取得並保留之快照資料，除了

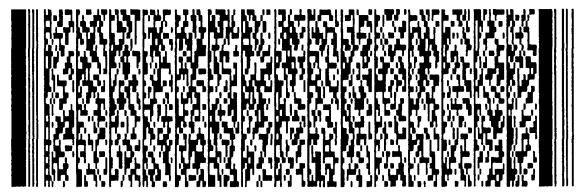
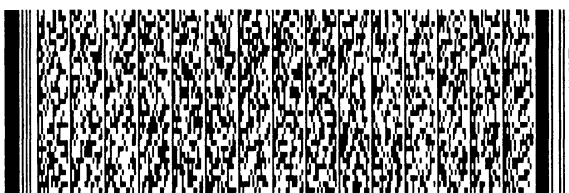


五、發明說明 (6)

執行資料備份之外，也可以使用於其他不同的用途，例如用來回復意外刪除的檔案，或者擷取有意或無意被重寫的資料。

系統結構100顯示有一網路106，用來連接檔案系統102與備份108。在本發明之一實施例中，網路106為一個線路交換網路(circuit-switched network)，像是公共服務電話網路(Public Service Telephone Network, PSTN)。在本發明之另一實施例中，網路106是一個封包交換網路(packet switched network)，封包交換網路為一個廣域網路(wide area network, WAN)，例如全球網路(global Internet)、私有WAN、區域網路(local area network, LAN)、電信網路(telecommunications network)或上述網路之任何組合。在本發明之另一實施例中，網路106是有線、無線、廣播或者點對點(point-to-point)。在本發明之另一實施例中，檔案系統102執行於與備份108之電腦系統相同的電腦系統上。

在本發明之一實施例中，檔案系統102和備份108之電腦系統為一個或多個個人電腦(如執行微軟視窗95/98/2000/ME/CE/NT/XP作業系統之IBM或相容的個人電腦工作站、執行Mac OS作業系統之麥金塔電腦或等同物)、個人數位助理(Personal Digital Assistant, PDA)、遊戲操作台(game console)或其他電腦處理裝置。在本發明另一實施例中，檔案系統102和備份108之電腦系統為伺服器，例如一個或多個執行像是SunOS或AIX作



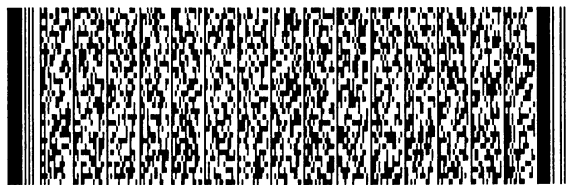
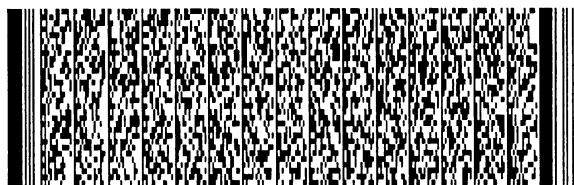
五、發明說明 (7)

業系統之電腦、執行SunOS作業系統或IBM RS/6000之SUN Ultra工作站、以及執行AIX作業系統之工作站和伺服器。

檔案系統和檔案結構之例子

本發明之實施例係操作於將資料存放在大容量儲存裝置 (mass storage devices) 之電腦系統，但不限於上述之大容量儲存裝置，也可以操作於具有不同特徵或結構之資料儲存裝置。

在說明書中，檔案系統(即圖1之檔案系統102)包含一電腦程式，其允許存放其他應用程式並且於儲存裝置上擷取資料，例如磁碟機、非揮發性記憶體(non-volatile memory)或其它資料儲存裝置。在實施例中之檔案係指一個任意大小且已命名的資料物件，檔案系統通常允許應用程式產生檔案，並且替那些檔案命名，進而在那些檔案中儲存資料、讀取資料、刪除和執行其他操作。檔案結構亦定義儲存裝置上之資料組織，除了檔案資料本身以外，檔案系統一般包含一個目錄(directory)，將檔案名稱映射(map)於資料檔案；一檔案超資料，包含檔案資料在儲存裝置上之位置之相關檔案資訊(即哪些裝置區塊存放有檔案資料)；一配置映射(allocation map)，用以記錄哪些裝置區塊正在使用中，以便存放超資料和檔案資料；以及一包含關於檔案結構整體資訊之超區塊(superblock)(如目錄位置、配置映射，以及其它超資



五、發明說明 (8)

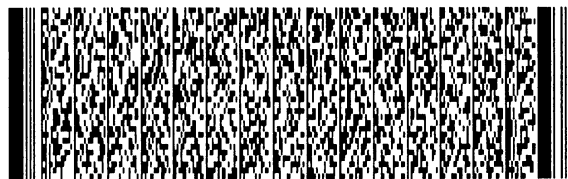
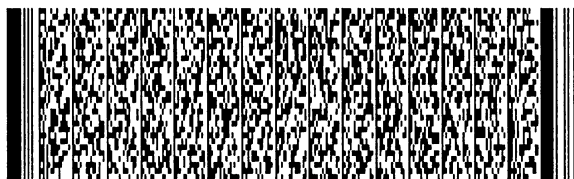
料結構)。

本發明之實施例使用一檔案系統，係根據下列超資料結構來執行檔案和目錄：

資料節點：超資料元件，包含檔案屬性(如所有人、存取許可、修改時間、檔案大小)，並且指明資料區塊(用於小檔案)或間接區塊(用於大檔案，其資料節點中之資料區塊的數量大於磁碟位址之數量)之物理磁碟位址。在本發明實施例的說明中，資料節點之匯集(collection)稱為「資料節點檔案」。實施例將資料節點檔案視為一般檔案(資料節點加上間接區塊)來儲存，但是其它實施例則利用不同的方式來表示資料節點之匯集，而包含在資料節點中某些或全部資訊的匯集稱為「資料節點資訊」。

間接區塊：超資料元件，包含資料區塊之磁碟位址或其他間接區塊(用於需要多個間接區塊層級之大檔案)。

目錄：超資料元件，包含將檔案名字映射至資料節點號碼之目錄項目。本發明之實施例使用相似於其它檔案資料之儲存方式，來存放資料區塊中之目錄項目。實施例之超資料結構用於存有相似於其它資料檔案所使用之目錄資訊之檔案(資料節點加上包含目錄區塊磁碟位址之間接區塊)。實施例之目錄可以包含指向其他目錄之項目，以產生一階層目錄結構(hierarchical directory structure)，由另一目錄之項目所指向之目錄係稱為子目錄。格式之細節以及目錄之其他超資料對於本發明來說並非特定的，而其他可選擇的實施例則使用其它目錄和超資



五、發明說明 (9)

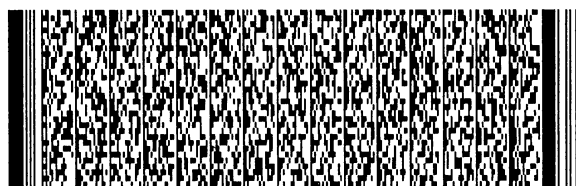
料儲存結構。

區塊配置映射(block allocation map)：一超資料結構，用來表示每一資料區塊是閒置的或是使用中(即每一資料區塊是否包含現有檔案之資料或超資料)。

超區塊：一超資料元件，包含整體性檔案系統配置資料(global file system configuration data)，特別是包含定位資料節點和配置映射所需的資訊。

資料區塊：檔案系統記憶體區域，包含未來會被用戶或其它應用擷取和使用之用戶資料。

圖2A為與資料儲存於一檔案系統相關之示範性資料結構集合200。在本發明實施例所操作之檔案系統中，每一檔案係配有一稱為資料節點202之結構。資料節點202包含檔案之超資料204，例如檔案之建立時間、所有人身分、指向資料區塊和與資料或資料檔案相關之其它資料的指標。每一資料節點202係由一資料節點號碼來單一地辨識，而資料節點中之資料供檔案系統來使用，以定位連結至那個資料節點之檔案。實施例中之資料節點並不存放用戶資料，也不存放用戶指派給檔案之名稱。一個現正開啟的檔案系統之資料節點一般存放在單一特定檔案，稱為資料節點檔案，通常用戶是看不見它的。在實施例中之檔案系統的資料節點具有固定大小，並且連續地存放在資料節點檔案中。藉由將資料節點數量乘上每一資料節點長度，來決定資料節點檔案之一偏移量(offset amount)，然後



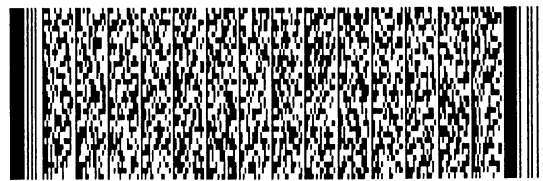
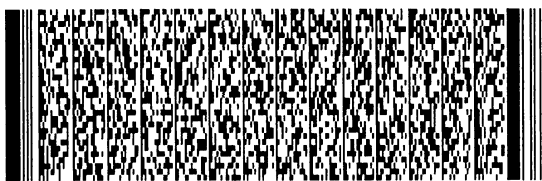
五、發明說明 (10)

從資料節點檔案中，由計算出的值來偏移資料結構，以存取其特定資料節點號碼之資料節點。

資料節點202進一步包含一磁碟位址206，其指向資料區塊208之指標。在實施例中，包含在資料節點202中並且指向資料區塊208之磁碟位址資料206包含那些資料區塊208之磁碟位址。磁碟位址資料206可以指出特定磁碟機之身分，以及包含資料區塊208之磁碟機中的一區段或區塊。

圖2B為第二示範性資料結構集合250，其說明在與本發明實施例一起使用之檔案系統中的資料檔案。資料結構之示範性集合200係用於小檔案，其中磁碟位址資料206可以存放在資料節點中。資料結構之第二示範性集合250係用於較大檔案，其中資料節點252之磁碟位址資料256存放有指向一個或多個間接區塊258之指標，每一間接區塊258包含一個或多個指標，例如間接區塊第一磁碟位址260和間接區塊第二磁碟位址262指向資料區塊，資料區塊可以是第一資料區塊264和第二資料區塊266，或者其他包含磁碟位址資訊之間接區塊。

在與本發明實施例一起使用之檔案系統中，由用戶指派給檔案之名稱係存放在一特定類型之檔案，稱為目錄。圖3顯示一個示範性檔案系統目錄結構。示範性目錄結構300包含二個目錄項目或稱為記錄(record)，像是第一目錄記錄304和第二目錄記錄310，實施例所使用之目錄可以包含大量的目錄項目。於此將描述與第一目錄記錄304相

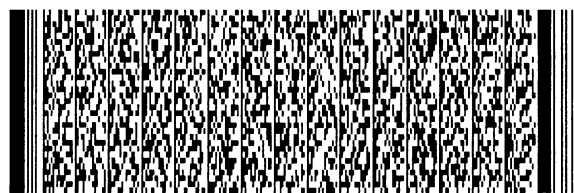
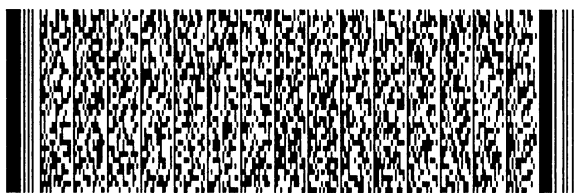


五、發明說明 (11)

關之結構，而其他目錄記錄(包含目錄記錄310)則相似於第一目錄記錄304。

第一目錄記錄304之資料包含一檔案名稱306和資料節點號碼308。目錄302係由檔案系統來管理，而記錄304將使用者所指派的檔案名稱306映射至資料節點號碼308。資料節點號碼308用來單一地辨識與此目錄項目304相關檔案的資料節點316，在實施例中，這些目錄項目有時候稱為連結(link)。這些實施例中之連結指向用戶檔案、其他目錄、或其它未在此說明之檔案系統對象。實施例之目錄施加一階層命名結構於檔案系統之檔案上，階層之開端係檔案系統之根目錄，在檔案系統中之檔案和目錄係可從根目錄或來自根目錄之子目錄來存取。檔案可以具有多於一個的目錄項目，以及多於一個的名稱，供用戶存取。

檔案系統也定義存放在一個或多個大容量儲存裝置之資料結構。許多檔案系統支援同時建立、修改和刪除多個檔案，這樣的檔案系統需要連結與寫入的一致性。如果每一檔案具有至少一個目錄項目，而每一目錄項目係連結至連結建立時所連結之檔案，檔案系統就會被視為具有連結一致性。此外，具有寫入一致性之檔案系統會保留單一檔案中或位於檔案之間已經安排好之寫入次序，例如在寫入至第一檔案，並且強迫第一檔案中被寫入之資料寫入至非揮發性儲存裝置後，才寫入至第二檔案之應用程式，只要有資料位於第一檔案中，就會要求資料必須出現在第二檔案中。維持此特性之檔案系統會被視為具有寫入一致性。



五、發明說明 (12)

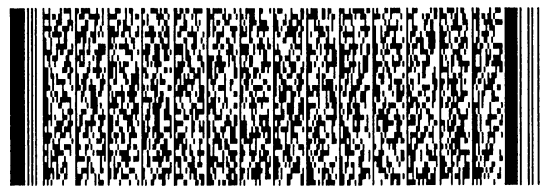
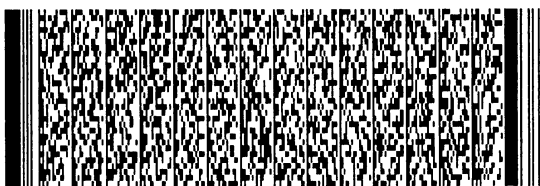
檔案系統快照

本發明之實施例會取得檔案系統之一個或多個快照，以產生一資料集合，用來在取得快照時，保存檔案系統中之資料狀態，但必須產生並取得檔案系統中所有檔案之快照，以維持一致的檔案系統圖像，同時在取得快照後修改前，有效率地複製檔案系統中之舊資料。

在一開始取得快照之後，實施例之快照會被保留。在取得快照之後，快照保留程序會增加對現正開啟檔案系統之修改。在取得快照之後，修改位於現正開啟檔案系統中之資料時，實施例之程序會在修改原始檔案中之資料前，將資料的原始版本(即快照產生時之資料版本)複製至快照資料集合。這個系統之操作會使得在修改現正開啟檔案系統中之檔案時，檔案快照內容所需的儲存空間會增加。

圖5顯示本發明實施例之整體控制流程500。控制流程500係從步驟502開始，並且直接進行至步驟504。在步驟504中，一快照初次被取得，以獲得檔案系統當時的狀況。一旦快照被取得，流程會進行至步驟510之快照保留程序，其中由於在現正開啟檔案系統中更新，被取得的快照也會被更新(步驟506)，或者會擷取來自快照之資料，以回應用戶的請求或存放在快照資料集合中之資料的程序(步驟508)。隨後，當最新的快照被取得並保留時，控制流程會回到步驟504。

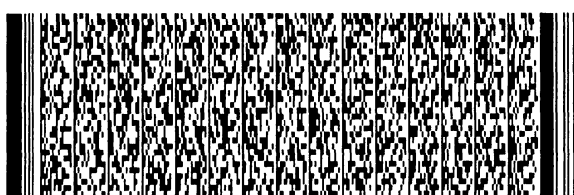
圖6A為本發明實施例之快照取得和控制流程600。控



五、發明說明 (13)

制流程600開始於步驟602，並且直接進行至步驟604，以暫停對現正開啟檔案系統的更新。在產生快照之前，藉由延遲未來對檔案系統中之檔案與目錄的更新，並等待即將發生之更新，實施例之流程會將檔案系統帶入一靜止狀態(quiet state)。在延遲更新檔案系統後，在步驟606中，流程會寫入快取資料更新(cached data update)至檔案系統，以確保所有資料存放在檔案系統上，且操作中之中斷(如電力中斷)不會過度地損壞檔案系統中之資料。在資料寫入檔案系統後，在步驟608中，藉由建立映像資料節點檔案之一資料節點，並將此快照之存在記錄於檔案系統超區塊中，來會產生實際快照。

在實施例中所產生之映像資料節點檔案係存在於原始檔案系統上，並且如同任何其他檔案一樣的產生和存取。實施例之稀疏(sparse)映像資料節點檔案藉由建立與快照相關之映像資料節點檔案之資料節點來產生的。在本發明之實施例中，建立快照並不會涉及用於映像資料節點檔案或快照資料集合之其他元件之資料區塊的配置。在取得映像資料節點檔案之快照時所產生之資料節點反映了映像資料節點檔案具有與現正開啟檔案系統之資料節點檔案之資料節點檔案相同之長度，但是包含在此資料節點中之磁碟位址全部等於無效值(NULL value)(即全部等於零)，無效值用以表示映像資料節點之資料區塊尚未被配置。這個特徵使得快照可以僅藉由建立快照資料檔案之資料節點來取得，而無須將資料區塊配置在檔案系統上。

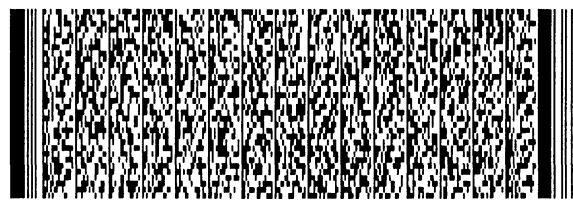
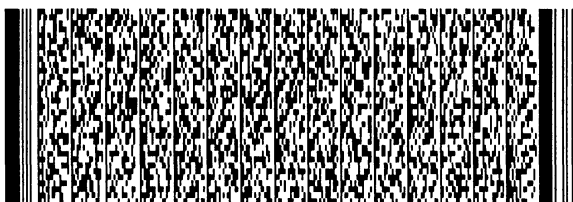


五、發明說明 (14)

稀疏資料節點檔案因此會產生於僅需要少量磁碟寫入操作之程序中，因為產生一個大的且完全稀疏的檔案資料節點僅需要寫入其資料節點（最近產生的稀疏資料節點檔案不包含任何資料，實際上僅是用於未來使用之「保留」空間）。實施例之稀疏映像資料節點檔案一開始包含所有無效值，在快照資料節點檔案中，一個特定資料節點資料之存在係藉由辨識快照資料節點檔案中之資料節點資料是否全部包含無效值來決定，如果快照資料集合之資料節點資料結構包含非無效值，則假設有效的資料節點資料存在於那個區塊中。映像資料節點檔案之資料節點位置記錄在超區塊中，以產生快照之記錄並且支援後續之處理。

藉由存放指向剛取得快照之映像資料節點檔案之資料節點位置來更新超區塊後，原始資料檔案系統中之一般檔案系統活動會重新開始（步驟610），被延遲之更新會被處理，而應用程式可以再次存取原始檔案系統資料。在本發明之實施例中，存取檔案系統被有效地攔阻一段短暫的時間，以完成上述任務。在完成產生快照之前，所有髒資料（dirty data）和超資料都經由快照所取得之相關檔案之磁碟快取記憶體之傳統程序，沖洗（flush）至原始檔案系統。這確保後續的故障（如電力損失）不會導致記錄在快照中之檔案系統的狀態（如下所述，一開始係儲存為原始資料檔案）突然發生改變。

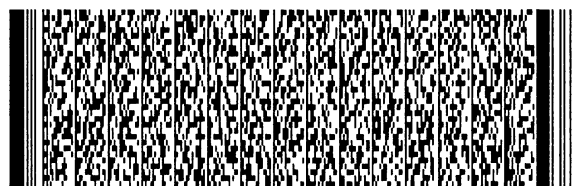
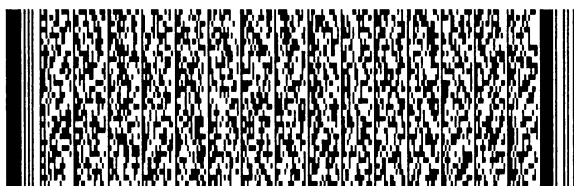
更新原始檔案系統可以區分為數種不同的類型。以下將描述實施例之流程，在快照已經取得並且存在之後，對



五、發明說明 (15)

檔案系統進行更新，此流程係用於執行在實施例中之不同更新類型。實施例執行以下所述之流程，在經由傳統流程對原始檔案系統進行更新之前，影響快照資料集合中之變化。圖7A為本發明實施例之快照更新程序操作處理和控制流程700。控制流程700係從步驟702開始，並且直接進行步驟704、706或708之其中一個步驟。在更新現正開啟檔案系統之檔案時，實施例之流程會進入快照更新程序700，而即將執行之流程藉由執行於現正開啟檔案系統之更新類型來決定。

在步驟704中，快照更新處理700在此判斷是否只有原始檔案之資料節點資訊將被更新。更新檔案系統中之檔案屬性改變傳統上僅藉由更新包含在檔案中之資料節點之相關超資料來執行(例如經由Unix指令chown或chmod來改變檔案歸屬或許可)。如果是產生快照以來資料節點的第一次改變，在更新原始檔案系統之資料節點前，本發明實施例首先會在步驟710中，將一原始資料節點內容之複本存入或複製到映像資料節點檔案中。這操作是利用來自正在更新之檔案之資料節點的資料來填滿映像資料節點檔案中一部份的稀疏區域。當原始資料節點存放在檔案系統所保留之資料節點檔案中時，因為實施例將快照資料節點存放於映像資料節點檔案中之相同偏移處，所以快照資料節點檔案中之資料節點複本具有與原始資料節點相同之有效資料節點號碼。來自原始資料節點之檔案屬性會如現狀地複製，但是存放於快照資料節點檔案之映像資料節點中之磁

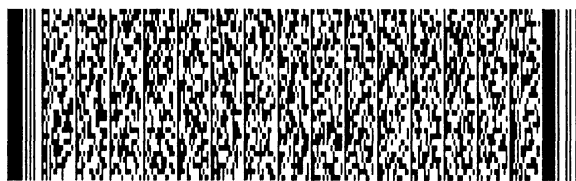


五、發明說明 (16)

碟位址則利用「相同(ditto)」值來替換，在此表示相同的磁碟位址。

實施例的相同值為一個保留值，其表示磁碟位址值不是一個真正的磁碟位址，而需要額外的處理來存取資料，以下將對此進行描述。在實施例中，磁碟位址之「相同」值係以"-1"來儲存。由於-1不是一個有效的磁碟區塊位址，此流程會發現這是一個相同值，需要特定的處理才可擷取用戶資料，以下將對此進行描述。一旦原始資料節點資料係複製至快照資料集合中之陰影快照資料節點檔案時，快照更新流程會在步驟716終止，在這個例子中，原始資料節點(現正開啟檔案系統中之資料節點)係正常地更新。

實施例使用特定處理，以覆寫或刪除檔案中之資料。本發明實施例刪除檔案中資料之流程包含刪除整體檔案，其相似於覆寫檔案之處理。以下將說明實施例用來覆寫資料檔案中之資料的流程，同時說明用來刪除檔案中之資料，而習知技藝者可以輕易了解，這兩個流程之間明顯的不同之處。用於檔案覆寫操作之流程包含覆寫一資料檔案全部或部份內容。在步驟708中，若更新原始資料檔案是重寫原始檔案中之用戶資料，快照更新流程會決定原始檔案中之資料即將被覆寫。在步驟712中，此流程會將來自原始資料檔案之資料移動或複製至快照資料集合。在步驟714中，利用指向已移動或複製至快照資料集合之資料區塊之資料節點的磁碟位址部份，快照資料節點檔案中之資



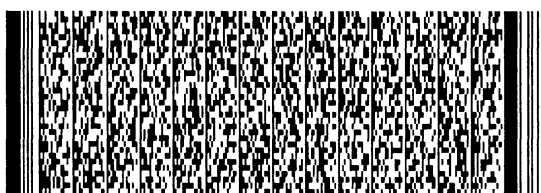
五、發明說明 (17)

料節點會被更新，以包含原始檔案資料節點。

如果原始資料區塊之操作是覆寫一整個資料區塊，則在更新前，本發明有些實施例不會將資料複製至一個新的磁碟區塊，而藉由將原始資料區塊之磁碟位址儲存至快照資料集合之間接區塊，同時利用最新配置之資料區塊位址來取代原始檔案之間接區塊中之位址，將舊的磁碟區塊移動至快照資料集合。當原始檔案不再需要這些原始資料時，上述動作會將與原始資料複製有關之磁碟I/O由原始資料區塊儲存至快照資料集合。

在檔案已經在快照中被取得後，在現正開啟檔案系統中之現正開啟檔案中之更新資料係根據所需的更新類型，藉由實施例之數個演算法其中一個來執行。如果更新是在產生快照之後，對原始檔案系統中，給定檔案之特定資料區塊的第一次更新，則原始資料區塊之複本首先會複製至快照資料集合。如果資料節點和原始檔案中被選擇之間接區塊尚未藉由先前操作被複製至快照資料集合中，則那些超資料元件首先會被複製至快照資料集合。如果資料節點先前尚未被複製，資料節點則被複製至映像資料節點檔案中，如同上述之屬性改變。

在快照資料集合之映像資料節點檔案中，藉由判斷資料節點記錄是否包含無效值，資料節點資料之存在係決定於實施例中。如果檔案夠大且具有間接區塊，指向更新中資料區塊之間接區塊也會被複製至快照資料集合。實施例會藉由配置一個新的間接區塊於快照資料集合中，來複製

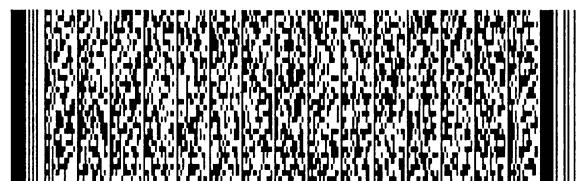
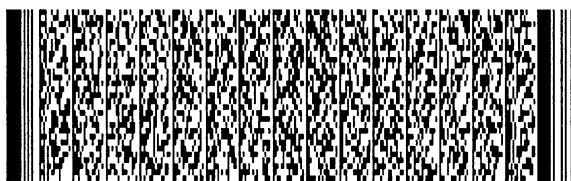


五、發明說明 (18)

這些超資料元件，利用「相同」磁碟位址來填滿間接區塊，並且將新(快照)間接區塊之位址存放至快照資料節點(進而取代當資料節點複製至映像資料節點檔案時所存放之「相同」磁碟位址)。如果檔案是具有數個層級之間接區塊，此流程會重覆執行於資料節點導向更新中的每一間接區塊。如果在先前的更新期間，部分或所有受到影響之間接區塊已經複製至快照資料集合，那麼這些間接區塊則不需要在快照資料集合中被更新。

在更新資料節點和任何與更新資料區塊有關之間接區塊後，藉由配置一個新的磁碟區塊、將來自舊區塊之資料複製至新區塊、以及將新區塊之磁碟位址存放至快照資料集合之被複製間接區塊(因而在間接區塊從原始檔案複製至快照資料集合時，取代存放在間接區塊之「相同」磁碟位址)，欲更新資料區塊係被複製至快照資料集合。在資料區塊複製至快照資料集合之後，原始檔案中之資料區塊係正常地被更新。

其它類型之資料檔案更新係附加資料至現有檔案。在步驟706中，快照更新處理控制流程700會判斷更新原始資料檔案是否包含一檔案附加操作，其會使得資料附加至原始資料檔案。本發明實施例與附加資料至檔案有關之處理與更新資料區塊之處理相似：在步驟710中，原始檔案之資料節點首先被複製至快照資料集合之映像資料節點檔案，以存放舊檔案大小以及資料修改時間之複本。然而，在附加資料至檔案的情況下，現有資料檔案中之任何資料

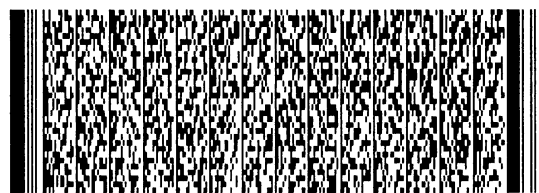


五、發明說明 (19)

均不會被覆寫，因此不需要複製任何資料或間接區塊至快照資料集合。

另一種類型之資料檔案更新為削減一資料檔案，通常發生在執行清除或刪除檔案之過程中。關於更新這些檔案之流程亦相似於更新整個資料區塊時之流程。這些程序包含：首先將即將被清除或刪除之資料複製至快照資料集合，然後在需要的時候，在更新資料區塊之流程下複製資料節點和被選擇之間接區塊。在資料區塊、資料節點和任何間接區塊被複製至快照資料集合之後，原始檔案之操作則可以正常地進行。

本發明有些實施例可執行清除或刪除整個資料區塊之最佳化操作，這些實施例執行相似於部分實施例覆寫整個資料區塊時所使用之最佳化流程。這些實施例藉由將即將被清除或刪除之資料區塊從原始資料檔案移動至快照資料集合，並且利用一個無效的磁碟位址來取代原始檔案超資料，以執行清除或刪除整個資料區塊之操作。藉由辨識間接區塊之整體內容是否指向即將被刪除或清除之資料之部分或全部範圍，這些實施例會進一步最佳化清除或刪除資料區塊之流程。再藉由複製間接區塊之磁碟位址至資料節點，這些實施例會移動原始檔案之整個間接區塊至快照資料集合中之資料節點或更高層級之間接區塊。這些實施例亦可僅藉由將磁碟位址從原始資料節點複製至快照資料集合中之快照資料節點檔案的映像資料節點，來執行削減檔案長度至零，或者刪除從取得快照以來尚未被更新之檔案



五、發明說明 (20)

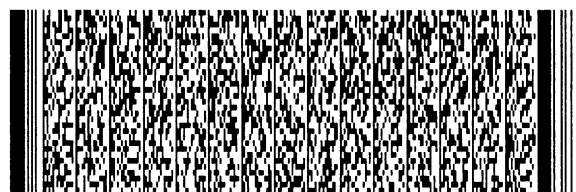
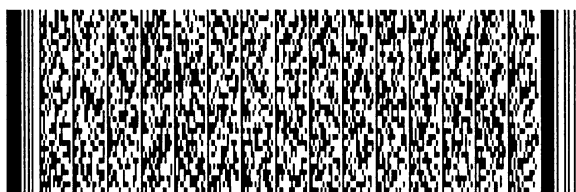
的特例。

實施例執行目錄更新之流程相似於更新資料區塊之流程。如果需要的話，目錄資訊首先會被複製至快照資料集合，然後資料節點和被選擇之間接區塊會如同上述之更新資料區塊之流程被複製。

本發明實施例之快照流程支援其它檔案系統管理操作，包含成長、萎縮，或重新組織檔案系統。這些操作涉及將現有資料與超資料移動至不同的磁碟位置，舉例來說，如可從IBM取得之通用平行檔案系統(General Parallel File System, GPFS)橫跨多個磁碟，將條紋(stripe)加至資料和超資料，並且支援從現有檔案系統移除一個或多個磁碟之操作，或者橫跨所有可利用之磁碟，以增加磁碟同時重新分配現有資料之操作，這些均涉及將資料和/或超資料從一磁碟移動至另一磁碟之例子。這些操作可以獨立於原始和快照資料上，因為它們不會改變檔案的內容。本發明實施例藉由先處理原始檔案，再處理所有快照資料集合，來執行這些操作。在這些操作之流程中，一個「相同」磁碟位址係被當作一個無效位址，即如果發現「相同」位址，無須進行任何動作。

讀取在快照中之資料

本發明之實施例允許用戶或其它程序讀取和修改已經在快照中被取得之資料。在快照中讀取和修改資料可以直接使用快照中存在之資料，而無須將快照之資料回復至現



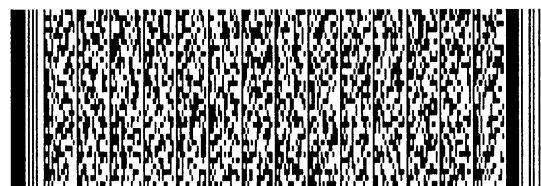
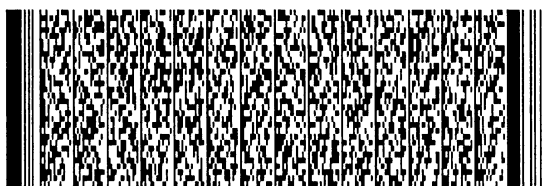
五、發明說明 (21)

正開啟檔案系統。

圖8A為本發明實施例之起始快照資料集合結構800之示意圖。在這個例子中，快照資料集合802是空的且代表一快照，其中從此快照產生以來，尚未更新原始檔案系統。此流程係用於從快照資料集合恢復資料，以下將會更詳細地描述。首先會檢查快照之資料節點檔案，以判斷有效的資料節點資料是否存在於正被恢復之資料檔案之快照資料集合802中。快照資料集合802中不存在資料節點資料表示實施例之一種流程，此流程從取得快照以來，所要求的資料尚未被修改，而"在"快照中之被要求資料位於現正開啟檔案系統之原始資料檔案，這會導致這些實施例所執行之流程將一個明示指標812指向現正開啟檔案系統之資料節點804，以存取所要求之資料區塊。

當本發明實施例試圖存取位於被取得快照中之資料區塊810，但快照資料集合不包含關於被要求資料檔案之資料節點資料時，實施例之流程會藉由使用現正開啟檔案系統之資料節點804中之磁碟位址808來存取資料區塊。如果沒有有效的資料節點資料位於快照資料集合802中，實施例之流程會同樣地藉由現正開啟檔案系統之資料節點804中之超資料806的明示指標，來存取檔案之超資料。

圖8B顯示本發明實施例之更新屬性資料結構820。在更新屬性資料結構820中，映像資料節點822存放在快照資料集合802中，其代表被取得且存入快照資料集合802之資料檔案的資料節點，且從快照被取得以來，屬性(即超資



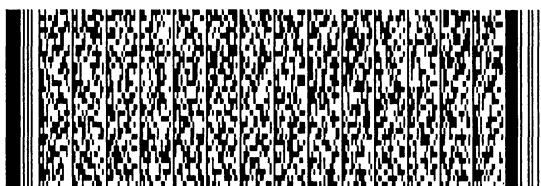
五、發明說明 (22)

料824)已經被改變。在更新現正開啟檔案系統中之資料節點超資料前,上述流程已經將資料節點資料從現正開啟檔案系統複製至快照資料集合802。這種結構使得包含超資料之映像資料節點822出現在相關的檔案中,當快照被取得時,此檔案亦隨之產生。

如上所述,映像資料節點之磁碟位址826包含「相同」值,在實施例中為"-1"值,以表示這是一個映像資料節點,而實際資料區塊810之真實磁碟位址係存放在現正開啟檔案系統之資料節點804中。存放在磁碟位址826之相同值表示一個指向實施例流程之明示指標828,此實施例表示擷取資料區塊810之實際磁碟位址的流程,利用現正開啟檔案系統之資料節點804中之磁碟位址808,其仍舊是在此情況下之原始檔案的一部份。

圖8C顯示本發明實施例之更新資料區塊儲存資料結構830。更新資料區塊儲存資料結構830代表從快照被取得以來,已經更新之資料區塊810。如上所述,實施例之流程在快照被取得之後更新資料區塊之前,將現正開啟檔案系統中之資料區塊810移動或複製至快照資料集合802。更新資料區塊儲存資料結構830說明了快照資料集合802中之映像資料節點822不只包含有效的超資料824,還包含一個有效的磁碟位址826。磁碟位址826表示資料區塊810之儲存位置,其為快照資料集合802之一部分。

圖8D為本發明實施例之間接區塊儲存資料結構840。快照資料集合802在這種情況下包含映像資料節點822,其

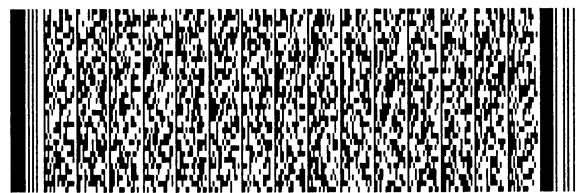
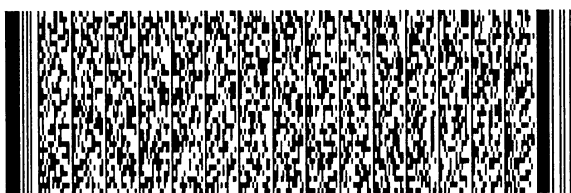


五、發明說明 (23)

具有最初被取得資料檔案之超資料824，以及一有效的磁碟位址826，其中磁碟位址826存放間接區塊842之位置，係與一個大的資料檔案相關。這個例子之間接區塊842包含二個磁碟位址：第一磁碟位址844和第二磁碟位址846，分別指向第一資料區塊848和第二資料區塊850，其為包含在快照資料集合802中之檔案的一部分。在這個例子，二個說明用的資料區塊已經移動至快照資料集合802中。

如果給定檔案之資料節點並非位於快照資料集合中，則表示從快照被取得以來，檔案均未改變，而操作會藉由存取儲存在原始檔案中之資料來讀取目前的資料檔案。如果被要求檔案之資料節點係位於快照資料集合中，此操作將如同一個使用快照資料集合之資料節點的傳統讀取操作來進行，不同之處在於當在快照資料集合中發現「相同」磁碟位址時，相對應的資料區塊(具有與存放在資料節點中相同邏輯區塊號碼之區塊)係從原始檔案中讀取(如下所述，如果快照不是多個快照之快照系列之部分)。

圖9為實施例之取得單一快照之資料檔案讀取流程900。控制流程900係從步驟902開始，並且直接進行至步驟904。在步驟904中，根據從快照資料集合讀取一資料檔案之請求，資料檔案讀取處理控制流程900一開始會存取快照資料集合中之映像資料節點，為了回應存取快照資料集合中之資料節點資料，三個結果其中一個可能會發生在這個實施例中，而特定之流程會執行這三種情況之每一個。以下將描述這三種情況之每一個流程。



五、發明說明 (24)

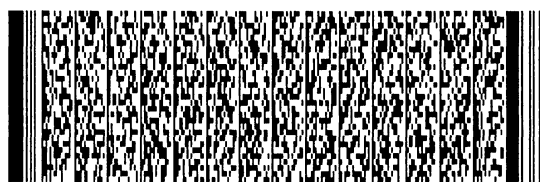
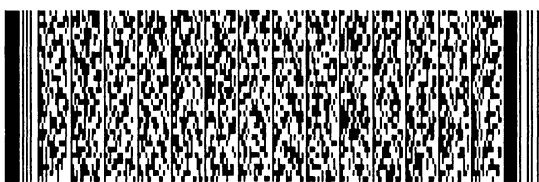
快照資料集合一開始是空白的且未包含任何資料。在由快照取得之原始檔案被修改至快照取得(snapshot capture)之後，資料節點被複製至快照資料集合。

在步驟906中，如果從快照被取得以來，原始資料檔案尚未被修改，此流程會決定快照資料集合不包含有效的資料節點資料。在這決定之後，在步驟912中，此流程會擷取來自現正開啟檔案系統所請求之檔案的資料節點，因為從快照被取得以來，原始資料檔案尚未被修改，而在現正開啟檔案系統上之檔案狀態係相同於當快照被取得時之狀態。在步驟914中，藉由使用存放在現正開啟檔案系統之資料節點中之資料區塊位址，資料檔案會被存取。

在步驟908中，除了磁碟位址部份包含「相同」位址指令之外，如果此流程判斷快照資料集合中之映像資料節點包含有效的資料節點資料，那麼從快照被複製以來，在檔案中之用戶資料則尚未被修改。如果需要的話，檔案超資料會從映像資料節點被擷取，並且在步驟912中，從現正開啟檔案系統中之原始檔案資料節點，擷取資料節點之資料區塊位址。一旦資料區塊位址被擷取，此流程會從原始檔案擷取被請求的資料（步驟914）。

在步驟910中，如果確定映像資料節點包含有效的資料節點資料以及有效的資料區塊位址，此流程會從資料區塊位址所指向之資料區塊擷取被請求的資料（步驟914）。

快照目錄中之資料查詢係藉由相似的處理來操作。首

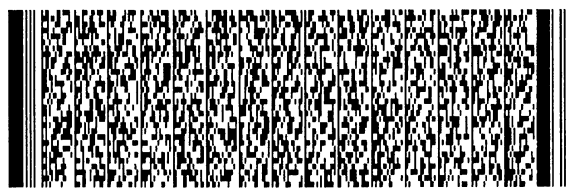
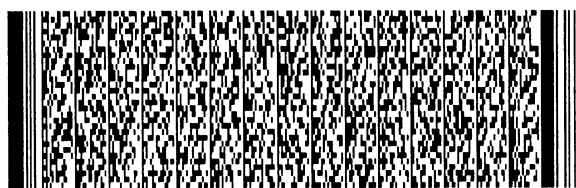


五、發明說明 (25)

先藉由檢查資料之快照目錄，以及隨後每當發現「相同」位址時，或當目錄資料節點落入映像資料節點檔案中之稀疏區域時，則從原始目錄讀取資料，查詢資料係執行於不支援如下所述之快照系列 (snapshot chain) 的實施例中。

支援多個快照

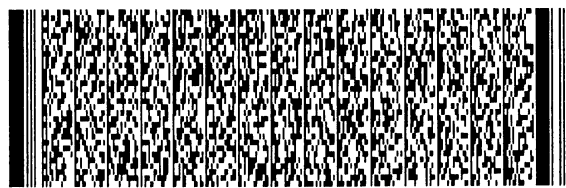
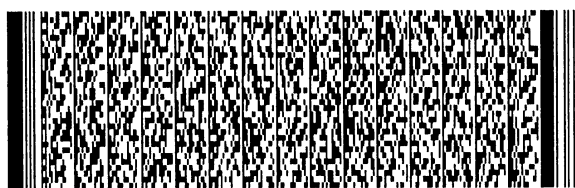
本發明實施例也支援任何數量的多個快照。多個快照用來取得現正開啟檔案系統在不同時間之狀態。圖6B為示範性快照時間軸線 (timeline) 620。示範性快照時間軸線620 以時間的角度說明有 n 個快照被取得，第一快照 S_1 是這個例子中之起始快照，後續的快照則按照時間順序而被取得，其中最近的快照係以快照 S_n 來表示。中間的快照 S_i 在快照 S_n 之前與快照 S_1 之後被取得的，快照 S_{i-1} 緊接在快照 S_i 之前被取得之快照，並且從快照 S_i 以來下一個最舊的檔案系統快照。快照 S_{i+1} 是緊接在快照 S_i 之後被取得之快照，並且在快照 S_i 之後下一個最近的快照檔案。為了產生一個新快照，一個新的稀疏的映像資料節點檔案會被產生，同時記錄在檔案系統之超區塊 (superblock) 中。當藉由根據上述之流程將資料節點、間接區塊和資料區塊複製至最近快照，以呈現多個快照時，實施例流程會對原始檔案系統進行更新。藉由快照寫入超區塊之順序，及/或快照資料集合之產生時間，實施例可判斷最近快照和在不同快照中相對之取得時間。



五、發明說明 (26)

當出現多個快照時，在先前時間檔案系統取得快照狀態會由用戶或其它應用程式來讀取。圖11說明一示範性快照系列1100，其顯示包含特定資料區塊之連續快照中之資料。示範性快照系列1100說明產生在本發明實施例之明示指標 (inferred reference)，用以讀取作為快照Si-1 1102之一部份資料集合之資料區塊。示範性快照系列說明快照Si-1 1102、Si 1106和Si+1 1112係顯示於示範性快照時間軸線620中。圖10說明當出現多個快照時，快照擷取程序之操作和控制流程1000。控制流程1000係從步驟1002開始，並且直接進行至步驟1004。

快照擷取程序1000開始於步驟1004，其存取與包含在快照Si-1 1102中之欲取得的資料檔案有關之映像資料節點。在這個例子中，此流程會發現一個不具有有效資料節點資料之第一映像資料節點1104，這表示在快照Si 1106被取得的時候，並未對資料檔案進行任何的更新。在第一映像資料節點1104中不具有有效資料節點資料表示在流程中，將會檢查下一個最近快照，以獲得資料節點資料。第一映像資料節點1104不具有有效資料節點資料也會使得沒有任何磁碟位址包含在資料節點中，此流程會繼續進行並且在步驟1006中判斷是否有一個磁碟位址在映像資料節點中。在這個例子中，第一映像資料節點1104不包含有效的磁碟位址或其他資料，因此藉由步驟1012判斷是否有較新的快照，此流程會繼續進行。藉由檢查檔案系統之超區塊，實施例會判斷一個較新的快照是否存在。在這個例子中，在

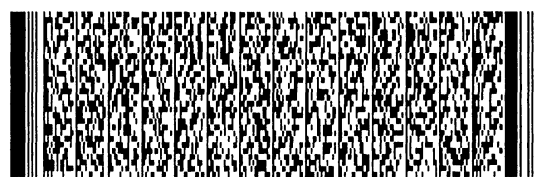
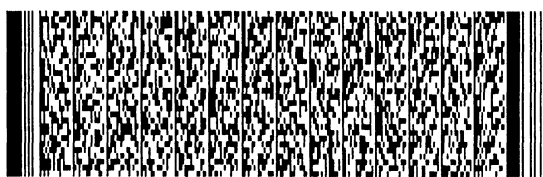


五、發明說明 (27)

步驟1010中，如果有較新的快照存在，此流程會更新快照以進行檢查。在這個例子中，將被檢查之快照會被更新為快照Si 1106。

在步驟1004中，藉由檢查第二映像資料節點1108，此流程會繼續進行，其中第二映像資料節點1108是快照Si 1106中之映像資料節點，快照Si 1106係對應於本發明實施例之快照Si-1中，第一映像資料節點1104之相同檔案。在這個例子中，快照Si 1106包含具有有效資料節點資料之第二映像資料節點1108，但是「相同」位址存放在第二映像資料節點1108之磁碟位址資料1110，第二映像資料節點1108之有效資料節點資料表示在取得快照Si+1 1112之前，檔案之超資料已經被修改，而第二映像資料節點1108包含當取得快照Si 1106時所存在之超資料(包含存在於及時快照Si-1 1102時之狀態)。

快照Si之檔案超資料以及先前修改資料檔案後所取得之先前快照(在這個例子還包含快照Si-1 1102)，係從第二映像資料節點1108讀取的。在這個例子中，第二映像資料節點1108之磁碟位址資料1110包含一個相同值，進而表示後續的快照會被檢查，以用於快照資料之資料區塊位址。在步驟1006中，藉由判斷未包含任何(有效的)磁碟位址，使得此流程繼續進行。而在步驟1012中，藉由判斷是否有較新的快照，使得此流程繼續進行。在這個例子中，有一個較新的快照，即快照Si+1 1112，因此在步驟1010中，被設定為下一個將被檢查之快照。

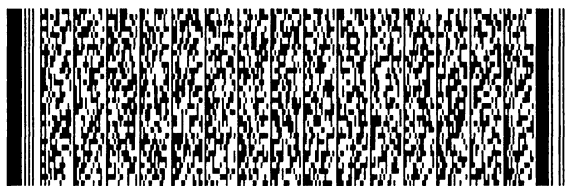
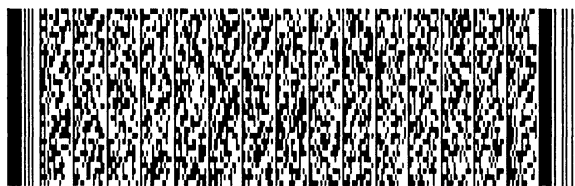


五、發明說明 (28)

下一個較新的快照(即快照Si+1 1112)包含一第三映像資料節點1114，位於快照資料集合中對應於被請求檔案之資料節點，用以指向資料檔案。第三映像資料節點1114在這個例子中，具有一個有效磁碟位址位於磁碟位址資料1116中，存放在第三映像資料節點1114之磁碟位址資料1116之有效磁碟位址表示在取得快照Si+1 1112之後，資料區塊已經被更新，而資料區塊之先前狀態係存放在資料區塊1118之快照資料集合1112中。在步驟1006中，判斷這個有效磁碟位址的存在，而在步驟1008中，資料區塊1118被擷取，以支援在這個例子中，在快照Si-1期間，對應擷取檔案狀態之快照資料。

在這個例子中，透過反覆檢查以追蹤映像資料節點系列，且藉由存取和判斷後續的快照，直到被取得的資料在快照資料集合中被發現，以擷取存在於取得快照Si-1 1102期間之被請求的資料檔案中之資料。為了簡化說明，這例子具有額外的資料存放在後續的快照中，本發明之實施例會利用不包含資料但包含表示在後續快照中找到實際資料之相同位址的中間快照(intervening snapshot)來操作。換句話說，在下一個較新快照資料也包含「相同」位址之情況下，流程會重覆，且「相同系列(ditto chain)」會被跟隨著，直到發現一個非相同的磁碟位址。

快照擷取程序1000會繼續進行，直到步驟1012中，判斷沒有任何比剛剛被檢查的快照更新的快照。如果所有較

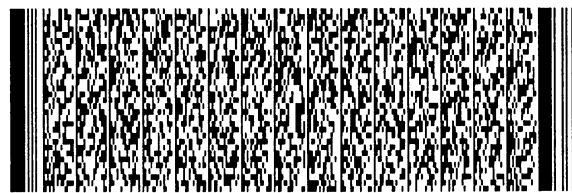
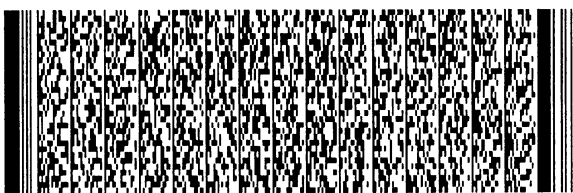


五、發明說明 (29)

新的快照都包含「相同」位址，那麼在步驟1012中，會判斷缺乏一個較新的快照，而在步驟1014中，會從現正開啟檔案系統上之原始檔案擷取資料或目錄區塊。相似地，當在較舊的快照中存取檔案時，如果沒有任何快照包含資料節點複本，操作會從最舊但較新包含有效資料節點複本之快照，或者從原始檔案擷取相對應的資料節點。此操作會發現一個稀疏區域存在於映像資料節點檔案中，藉由在資料節點記錄中之所有無效資料位元組之出現，可於實施例中判斷。

在一系列多個快照中之一個快照可以被刪除或用於回復資料至檔案系統。在實施例中，經由使用本說明書中其他部分所描述之流程，從非最近取得的快照之快照資料集合，回復資料至檔案系統可以解出該快照之暗示指標

(implied reference)。本發明刪除位於一系列快照中之一個快照資料集合之實施例會先判斷是否有一個快照位於在快照被刪除之前所取得的快照序列中，如果存在先前或母快照 (parent snapshot)，那麼流程會判斷下一個最舊的快照，或是任何母快照是否包含暗示指標，用來指向存放即將刪除之快照中之資料。若資料存放在即將刪除之快照中(即不是經由即將被刪除之快照中之相同位址來指向)，且如果下一個最舊快照包含指向該快照之暗示指標，則刪除快照之流程會將存放在該快照中之資料複製至下一個最舊快照，一旦這資料已經複製或移動至下一個最舊快照，快照就可以被刪除。



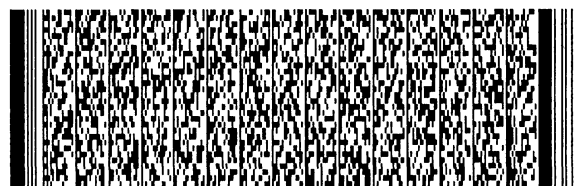
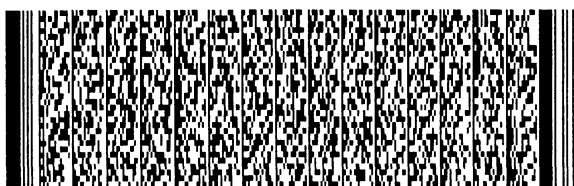
五、發明說明 (30)

支援多個可寫入快照

上述快照流程會產生「唯讀」快照。唯讀快照是在快照產生時取得檔案系統狀態之快照資料集合，但是不支援用戶或應用程式直接寫入，或更新存放在快照中之資料。本發明有些實施例支援寫入與更新先前建立的快照中被取得之資料，寫入資料至先前建立之快照具有使用快照建立時存在之檔案系統的效果。

特定之流程在具有一個以上的快照時，係藉由本發明之實施例來執行，以修改或更新快照。在具有一系列快照600(分別為 $S_1, \dots, S_{i-1}, S_i, S_{i+1}, \dots, S_n$)之檔案系統之例子中，快照係利用遞減的年齡來排序，即 S_1 是最舊的快照，而 S_n 是最近的快照。圖7B顯示利用本發明實施例來執行之快照更新處理控制流程720。

控制流程720係從步驟722開始，並且直接進行至步驟724。實施例之更新或修改現有快照(快照 S_i)中之檔案的流程會在步驟724藉由解出即將被修改，並且快照 S_i 中之暗示指標所指向之資料和/或超資料來開始，也就是說，快照即將被更新。解出資料區塊位址係根據執行於實施例中之快照擷取程序1000。其利用相同值所表示之暗示指標中之資料區塊位址來解出，包含在快照 S_i 中之相同值以及任何如上所述之後續快照之解出。快照 S_i 和任何後續快照之相同值利用後續快照中指向暗示資料之指標來解出，直到從後續快照或從原始檔案系統取得實際資料值。在解出



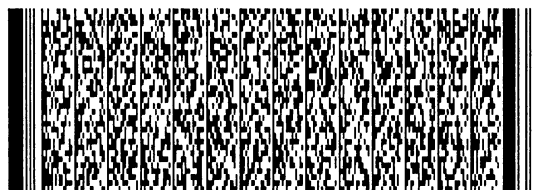
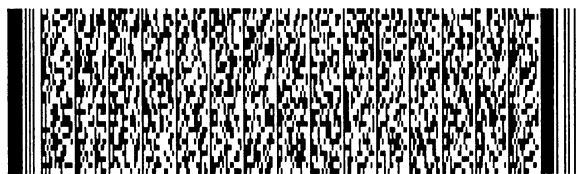
五、發明說明 (31)

Si 之暗示指標後，包含即將被修改之資料的資料區塊被複製至目前的快照Si中（步驟726）。

一旦擷取用於快照Si之資料，下一個最舊快照（在這個例子中為快照Si-1）係在步驟726中被檢查，以判斷下一個較舊快照Si-1是否包含指向位於快照Si中之資料之暗示指標，其中，快照Si對應於即將被目前的操作更新之資料。在步驟728中，如果快照Si-1包含一個指向位於快照Si中之資料之暗示指標，即如果快照Si-1包含一個對應於或定址即將被目前操作修改之資料的相同值，由於上述所執行之流程，暗示指標係藉由判斷資料出現於快照Si來解出。在步驟730中，一旦暗示指標或相同位址被解出，此流程會將目前操作所修改之資料複製到下一個最舊快照，即快照Si-1。

如上所述，為了達到更新包含在快照Si中之資料的目的，下一個最舊快照（即快照Si-1）等同於檔案系統之唯讀快照，其以快照Si表示。在更新快照Si之情況下，當快照Si中之資料被更新或改變時，相對於上述從原始檔案系統取得資料之情形，快照Si-1則用來從快照Si中「取得」資料。然而，與原始檔案系統不同的是，快照Si可以包含「相同」值。在讀取快照Si中之資料時，所發現的「相同」值藉由跟隨包含隨後所取得之快照中之相同系列（ditto chain）來解出，以擷取即將從較新快照或從原始檔案系統複製至快照Si-1之資料/超資料。

在步驟732中，一旦即將被修改之資料被複製至與快



五、發明說明 (32)

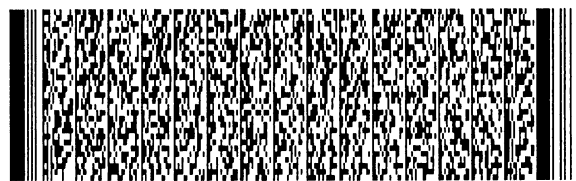
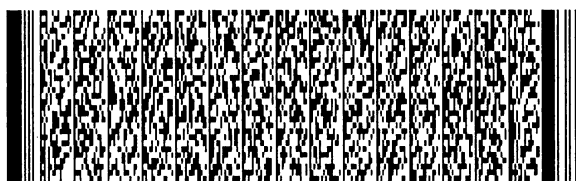
照Si有關之快照資料集合中，而欲取得之適當資料包含於快照Si-1時，快照Si中之資料會被修改，如同由用戶或實施例中之其他流程所決定的。

取得一快照之快照

本發明有些實施例可執行取得現有快照之快照的流程。如果先前快照係可寫入的並且在取得後係可以修改的，則會特別需要這樣的流程。可寫入快照之示範性使用乃用於新軟體的測試，新軟體的測試通常藉由使用相同資料執行軟體數次來完成的。在藉由寫入至原始快照資料來開始多次執行之前，可利用取得原始快照資料之新快照來執行多次相同輸入資料的軟體。

產生一快照之快照之流程包含與產生原始檔案系統之快照相似之步驟，這些步驟包含1)提昇所有資料更新活動至靜止狀態；2)產生一新映像資料節點檔案；以及3)記錄超區塊中之新快照之存在(包含新映像資料節點檔案之位置)。然而，在產生一快照之快照時，存放在超區塊中之資訊亦包含一個「原型(ancestor)」區域，用以表示該快照是原始檔案系統之快照或是一快照之快照。在一快照之快照之情況下，原型區域表示該快照所依據之較早快照。在檔案系統之快照的情況下，原型區域表示快照被取得之檔案系統。

支援取得快照之快照的實施例將「先前」快照(由記號parent(S)來表示)定義為最舊的快照，但比快照S還

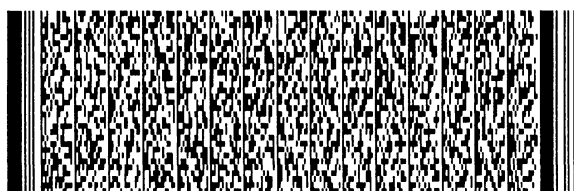


五、發明說明 (33)

新，並且具有與S相同之原型。在沒有具備相同原型之較新快照之情況下， $\text{parent}(S)$ 係S的原型。此關係會將所有的快照組織成一棵樹，而原始檔案系統位於樹之根部。例如，考慮取得現正開啟檔案系統之第一快照(快照S1)，有時候，在第一快照後有一個後續的快照(快照S2)從現正開啟檔案系統取得，這會產生S1和S2之快照系列，第三快照(快照S3)然後從快照S2取得。在這個例子中之快照樹具有二分支，從快照S2所取得之快照資料集合分裂開來的，而快照S1形成一個分支，快照S3形成另一個分支。一旦這兩個分支形成時，在每一分支中之資料集合的改變(如果它們是可寫入的)不會影響到其他快照分支中之資料，即 $\text{parent}(S1)=S2$ 、 $\text{parent}(S2)=$ 現正開啟檔案系統、而 $\text{parent}(S3)=S2$ 。在這個例子中，作為二個分支之結果，更新快照S2需要將快照S2中被修改之資料複製至以快照S2為母快照的兩快照，即快照S1和S3。

一般處理演算法係使用於讀取和更新快照資料集合之本發明實施例。此方法中之一個元件會讀取在快照S時所發現之「相同」值，其值藉由從快照 $\text{parent}(S)$ 讀取相對應資料節點或資料區塊來解出。如果快照 $\text{parent}(S)$ 亦包含一個「相同」值，資料則藉由讀取快照 $\text{parent}(\text{parent}(S))$ 來解出。這流程會重覆執行，直到相同值藉由發現實際資料或資料節點值來解出。

在更新快照S中之檔案前，支援取得快照之快照並且寫入至快照之本發明實施例也會將目前的資料和超資料內



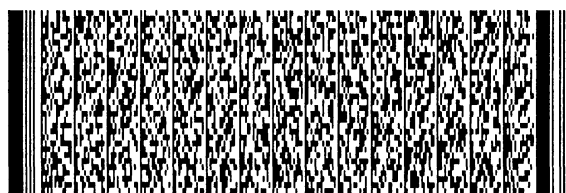
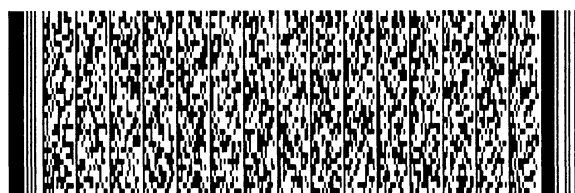
五、發明說明 (34)

容從快照S複製至S的所有子快照S'，其中 $\text{parent}(S') = S$ 。快照S之子快照就是原型為S之快照，且其中沒有比S更新的快照。為了準備更新快照Si中之資料，在更新快照Si中被修改的資料/超資料會從 $\text{parent}(Si)$ 複製至Si。

熱待命檔案系統(hot standby file system)

本發明之實施例使用快照以維持一熱待命檔案系統。熱待命檔案系統是一個分開的檔案系統，通常用來維持一個不同的資料儲存裝置，其保留存放在主要檔案系統之資料複本，並且與主要檔案系統一起使用，以確保整個系統的高度可用性。一旦由於硬體、通訊或其它的問題，使得主要檔案系統變得無法使用時，熱待命檔案系統可以負起主要檔案系統的功能，使得系統繼續操作。熱待命檔案系統在依靠一連續可用檔案系統(continually available file system)中特別好用。傳統系統係使用磁碟鏡像技術(disk mirroring)來實現熱待命檔案系統，在執行傳統磁碟鏡像技術時，鏡像(mirror)是一個第二磁碟，不斷地更新寫在主磁碟上之資料。傳統方法需要一個在主要(來源)檔案系統和待命(目標)檔案系統間連續且高速的連接。

另一方面，在實施例中之檔案系統快照為檔案系統內容在單一時點之邏輯複本。例如，作為檔案系統之先前狀態記錄，一個快照允許用戶擷取一個意外被刪除之檔案的舊複本，也就是還原刪除(undelete)。快照之另一個常見

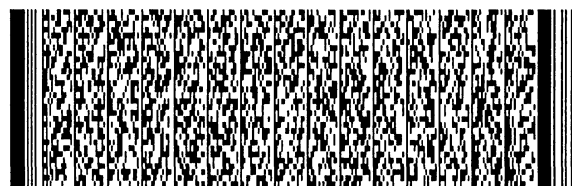
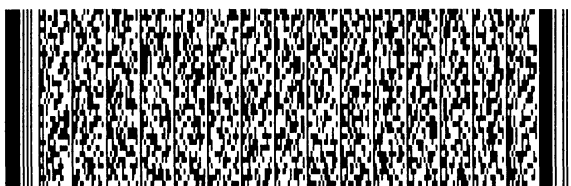


五、發明說明 (35)

用途是用來在備份期間，當允許對檔案系統進行持續的讀取/寫入之存取時，將檔案系統備份至磁帶(tape)。在某些實施例中，產生一個快照時不需要製作所有檔案系統資料之物理複本，相反的，複製檔案資料會延遲，直到下次資料被修改(稱為「寫入時複製」程序)。因為減少產生快照所需的時間，且減少保留快照所需的磁碟空間，使得快照更有效率。

本發明之實施例使用快照以產生並且維持熱待命檔案系統，使得原始和鏡像檔案系統之間不再需要高速的連續連接。這些實施例之鏡像檔案系統還可以作為數個主要檔案系統之鏡像，並作為其它資料用戶之主要檔案系統。

圖15A為本發明實施例之主要與待命檔案系統1500之整體系統結構方塊圖。這個示範性系統結構具有一個主要檔案系統，以檔案系統A 1502來表示；一個待命檔案系統，以檔案系統B 1504來表示；以及一網路106，以提供在這些檔案系統間之通訊。其他實施例維持主要和待命檔案系統於單一個處理器中，因而不需要網路106。在這個例子中之檔案系統A 1502具有二個快照資料集合：第一快照資料集合1506和第二快照資料集合1508。這兩個快照資料集合在不同時間取得檔案系統A 1502的狀態。檔案系統A 1502藉由將快照資料集合(例如第一快照資料集合1506和第二快照1508)與檔案系統B 1504聯繫來操作，檔案系統B 1504存放來自檔案系統A 1502之快照資料集合之複本，即存放第一快照資料集合複本1510和第二快照資料集



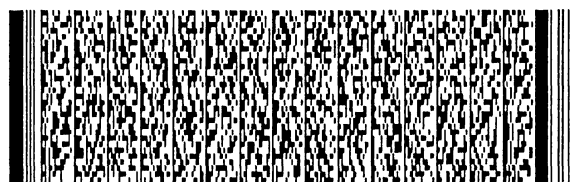
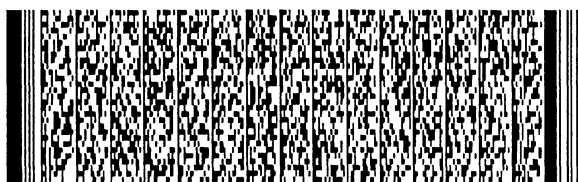
五、發明說明 (36)

合複本1512，以支援待命資料儲存操作。

藉由先產生一個原始(來源)檔案系統之快照，並且將其快照整個資料集合轉移至第二檔案系統，本發明這些實施例會產生一個熱待命檔案系統，以產生原始檔案系統之相同複本(即鏡像檔案系統)。

藉由產生原始檔案系統之新快照，同時判斷這些新的最近取得或產生的快照間之改變，以及由已經轉移至鏡像檔案系統之原始檔案系統中先前快照所取得之狀態的改變，這些實施例會定期地產生最新的待命或鏡像檔案系統。原始檔案系統會產生一組改變，其連接和應用於待命檔案系統，使得待命檔案系統處於原始檔案系統上所取得之新快照之狀態。原始檔案系統快照和由這些檔案系統所產生之一組改變包含標記(tag)，以藉由辨識產生順序或這些改變之應用順序，來確保鏡像檔案系統之完整性。在這個說明書中，「回復」(restore)為藉由處理表示來自其它檔案系統之整個快照之資料集合，檔案系統已經處於其它檔案系統之狀態。「應用」(apply)為藉由處理產生於其它檔案系統上之二個快照間之一組改變，檔案系統已經更新為其它檔案系統之較新狀態。

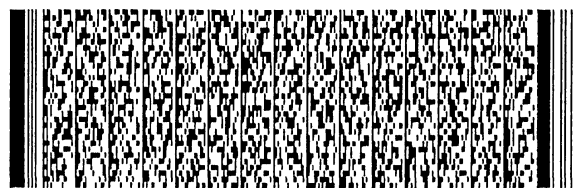
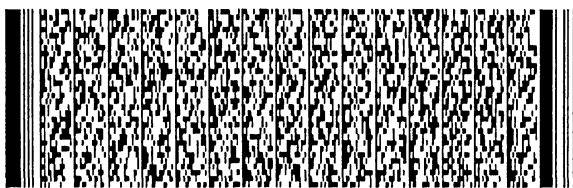
利用保留快照標記，實施例可以使得維持待命檔案系統更為容易，其中快照標記係用來單一地辨識兩個記錄一檔案系統在不同時間之狀態的不同快照，並且辨識產生在二個快照間之改變。快照標記在用來將鏡像檔案系統從唯讀檔案系統切換至現正開啟讀取/寫入檔案系統時，藉由



五、發明說明 (37)

確保故障使得原始檔案系統無法運作前最新快照已應用，來協調同步鏡像檔案系統和現正開啟檔案系統間之資料。一旦最初的鏡像檔案系統變成客戶端處理器所使用之現正開啟檔案系統(即新的原始檔案系統)，快照會從新的原始檔案系統取得，而快照標記則用來回復先前的原始檔案系統(現在變成鏡像)，將原始檔案系統維持為新的待命或鏡像檔案系統。

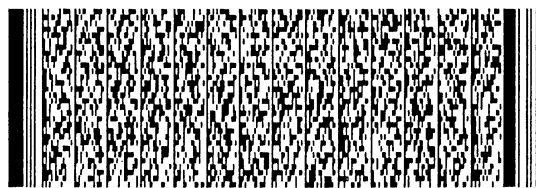
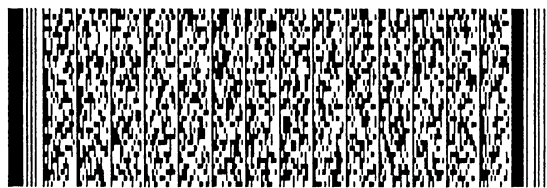
本發明之實施例使用快照標記來辨識每一快照及檔案系統，其中快照係從該檔案系統中取得。此處所使用之快照標記記號包含(A:S1)的格式，其中第一元素(在這個例子中為"A")用來辨識檔案系統，而第二元素(在這個例子中為"S1")為該快照之快照辨識碼(snapshot identifier)，使得此處所述之熱待命系統中之不同檔案系統可以取得不同時間之快照，並且只使用那些快照之一個子集合，即可讓那些檔案系統間之資料同步。實施例之檔案系統產生擷取自檔案系統之快照間之一組改變，這些改變包含一對標記，用以辨識在改變其間被決定之快照。如快照標記對(A:S2, A:S3)包含在一組改變中，那些改變被產生以作為檔案系統A上取得之快照S2與快照S3間的改變，這組改變只可以成功地應用於已經被回復至快照S2之狀態的檔案系統，例如，如果檔案系統B接收到這個快照，而來自檔案系統A之快照S2尚未被回復至檔案系統B，或改變尚未被應用至檔案系統B，將使得檔案系統B具有快照之狀態(A:S2)，而應用具有快照標記對(A:S2, A:



五、發明說明 (38)

S3) 之改變組是不適當的。檔案系統會排除接收到之一組改變，並且不具有開始於一快照標記之一快照對，其中快照標記對應於最近被回復或應用至檔案系統之快照。示範性系統辨識最後被應用或回復之快照，以及來自其他檔案系統中該組改變之請求，其中該組改變係對應於從最後被應用或回復之快照以來所產生之改變。

快照標記係存放在快照中及每一檔案系統中。存放在檔案系統中之快照標記會存放在檔案系統之超區塊，並且用來辨識被回復之最新快照，以建立一個基本檔案系統，而應用於基本檔案系統之最新快照的快照標記也是存放在檔案系統之超區塊中。在新的快照或一組改變應用於檔案系統之前，在檔案系統中之快照標記被視為一個新接收到的快照或一組改變之快照標記。只有在一個快照或一組改變具有對應於最近被使用於檔案系統上之基本快照的基本快照標記，才會應用於檔案系統。一旦來自來源檔案系統之快照被應用於一鏡像檔案系統，另一個快照係從鏡像檔案系統被取得，使其同步於原始檔案系統。實施例之檔案系統會將最後被回復或應用之資料的快照標記存放在檔案系統之超區塊。快照標記辨識來源檔案系統及被複製至此檔案系統之遠端系統中最後快照之快照辨識碼。這資料的一個使用例子用於當檔案系統接收到時，一系列快照更新被遺失或錯誤的情況下。在檔案系統無法正確地接收到一組或多組改變的情況下，會判斷最後被正確應用之一組改變，並詢問遠端檔案系統自從對應於該組已經適當地被回



五、發明說明 (39)

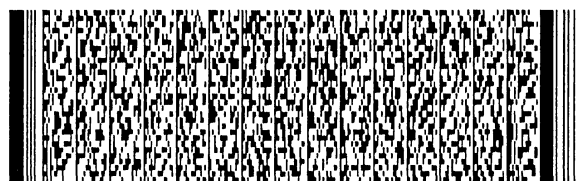
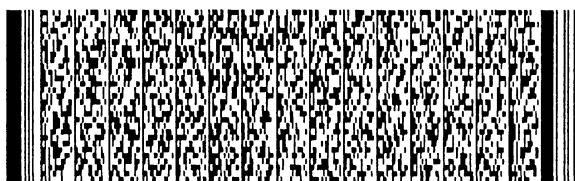
復或應用之資料的快照以來，發生於檔案系統之該組改變。

在利用原始檔案系統之所有快照來更新先前之鏡像檔案系統，且先前鏡像檔案系統被轉換成現正開啟檔案系統之後，新的快照會從目前現正開啟檔案系統(本來是影像檔案系統)取得。此時，現正開啟和鏡像角色互換，而目前之現正開啟檔案系統的快照會應用在已經變成新的鏡像檔案系統之原始檔案系統。

熱待命檔案系統之切割(cutting over to the hot standby file system)

在發生網路問題或來源系統本身其他故障，使得客戶(如應用伺服器)無法存取原始(來源)檔案系統之情況下，本發明實施例會切換以存取鏡像(熱待命檔案系統)，使得客戶可以快速地重新開始他們的工作，而無須等待網路與來源檔案的修復。這樣的切換執行於這些實施例中，通常會由系統管理員來執行以下步驟。

系統管理員執行FileImageRestore指令，以利用在目標檔案系統上之最新快照(即在FileImageMirror指令之前增加或底線執行之後所產生之快照)。FileImageRestore指令將快照應用於特定的圖像(image)，在這裡是鏡像檔案系統圖像。FileImageMirror指令在執行時，會產生原始檔案系統之複本以及相關的快照資料，以取得原始檔案系統之狀態，此程序確保現正開啟檔案系統為「乾淨

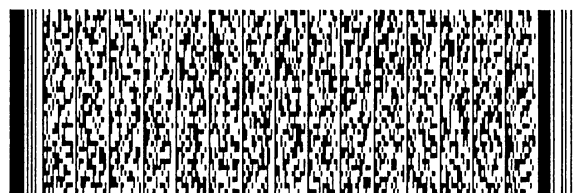
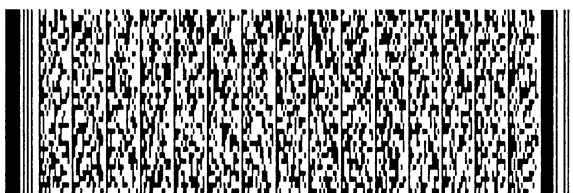


五、發明說明 (40)

的」。系統管理員然後會輸出現正開啟之目標檔案系統給用戶，然後再將客戶的伺服器配置從原始檔案系統切換至待命檔案系統。

圖15B、15C和15D為本發明實施例之待命檔案系統更新處理之處理流程之三個部分。圖15B顯示示範性待命檔案系統更新處理之第一部分的控制流程1520。控制流程1520從步驟1522開始，並且直接進行至步驟1524。在步驟1524中，在準備操作時，程序會藉由初始化檔案系統A和檔案系統B來開始。用來初始化檔案系統之流程係取決於電腦作業系統，及其它影響特定實施例之檔案系統之操作因素。例如，檔案系統係在Unix作業系統下被初始化時，會根據檔案系統之配置來使用「mount」指令和其他指令。某些實施例之檔案系統的初始化也會要求準備資料儲存硬體。在步驟1526中，檔案系統開始操作後，檔案系統A會取得快照0，並且將快照資料集合連接至檔案系統B。在步驟1528中，當檔案系統B接收到快照0資料集合時，會回復快照資料集合，以作為一個待命檔案系統。檔案系統B然後會將快照標記(A:S0)存放至其超區塊，以辨識對應於快照0之快照辨識碼的快照，其中來自檔案系統A之快照0是最後一個被回復或應用至檔案系統B之資料集合。

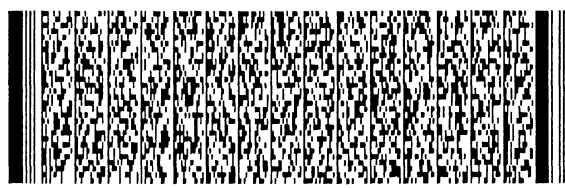
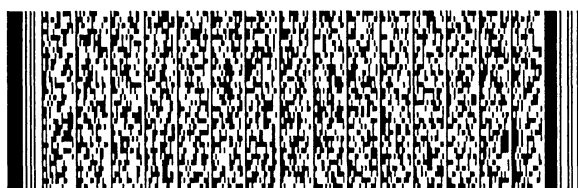
回復之後，檔案系統B即包含在快照0被取得時出現在檔案系統A之資料複本。在步驟1530中，檔案系統B在回復後取得快照0，在檔案系統B上再一次取得快照0具有在檔案系統B上建立一個乾淨的快照資料集合之效果，因此允



五、發明說明 (41)

許更有效率地存取檔案系統B上之快照。必須注意的是，這個說明使用相同的快照辨識碼(如快照1)，來用於在兩個檔案系統上之相同資料，這是為了清楚地說明範例，此外，使用於二個檔案系統上之快照辨識碼不必具有相同資料之相同值。如這裡所述，用以辨識檔案系統和快照辨識碼之快照標記係存放在檔案系統上，用來使得這些實施例中之二個檔案系統的資料能夠同步。

當這些檔案系統繼續操作時，在步驟1532中，檔案系統A會取得快照1。在取得快照1之後，在步驟1534中，實施例之流程會判斷介於快照0和快照1之間，發生在檔案系統中之資料的改變。檔案系統A然後會產生包含這些改變之資料集合，並且將資料集合連接至檔案系統B。這組改變包含快照標記對(A:S0, A:S1)，以辨識包含在檔案系統A上，發生於取得快照0和快照1間之改變。當接收到快照0和快照1之間的一組改變後，檔案系統B會確認包含在該組改變之快照標記是否適當地對應於已經回復或應用至檔案系統B之資料，檔案系統B之超區塊會辨識出檔案系統B之狀態係最近回復或應用至快照標記(A:S0)。在檔案系統B確認最後被回復或應用的資料匹配於接收到該組改變前之資料後，在步驟1536中，檔案系統B會將那些改變應用至在檔案系統B上之資料，使得快照1被取得時，更新包含在檔案系統B中之資料，以確保與檔案系統A上之資料相同。然後檔案系統B會將快照標記(A:S1)存放至其超區塊，來辨識對應於快照1之快照辨識碼的快照，其中來自

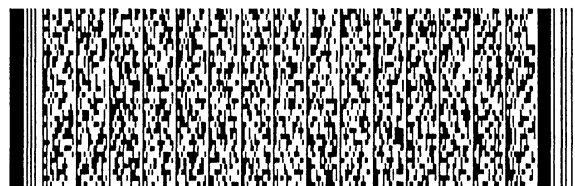
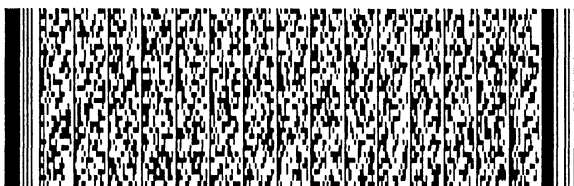


五、發明說明 (42)

檔案系統A之快照1為最後一個被回復或被應用至檔案系統B之資料集合。在步驟1536中，檔案系統B會取得快照1之複本。在檔案系統正常操作期間，這些步驟會重覆執行。

圖15C為示範性待命檔案系統更新處理之第二部分之控制流程1540。控制流程1540係步驟1542開始，並且直接進行至步驟1544。在步驟1544中，這個操作部分由於檔案系統之機械或其他問題，使得檔案系統A無法運作。在步驟1546中，接管程序 (takeover sequence) 會開始判斷檔案系統A已經無法運作。在步驟1548中，檔案系統B會回復接收到的最後快照，這個回復會使得檔案系統B現在是乾淨的，表示檔案系統B係處於前後一致的狀態。這會讓下列情況的發生，當一組改變藉由將檔案系統B回復至其最後知道的一致狀態，而移轉或應用至檔案系統B時，檔案系統A將無法運作。

在檔案系統B已經回復從檔案系統A接收到之最近快照後，在步驟1550中，檔案系統B會接管而成為主要檔案系統，這會使得檔案系統B成為客戶所使用之主要儲存裝置。當檔案系統B係作為主要檔案系統來操作時，在步驟1552中，檔案系統B會取得快照2，以獲得檔案系統B之狀態。在這個例子中，檔案系統B會繼續操作並且取得快照3 (步驟1554)。接著例子中之檔案系統A會再度變得可使用 (available)，而檔案系統A會回復檔案系統A之最後快照，在這個例子中是快照1 (步驟1556)，其中快照1係已連接至檔案系統B。在檔案系統A上之檔案系統回復之

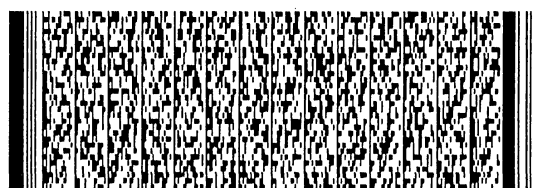
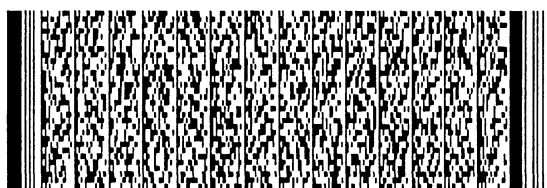


五、發明說明 (43)

後，檔案系統A則處於一個已知狀態，並且在步驟1558中被初始化，成為待命檔案系統來運作。

在步驟1560中，檔案系統A被初始化，並且變成待命檔案系統之後，檔案系統B會從檔案系統A接收到最近快照間產生一組改變(在這個例子中為快照1)，並且將該組改變連接至檔案系統A，這組改變包含快照標記對(A：S1，B：S3)。在步驟1562中，檔案系統A會從檔案系統B接收所產生的該組改變，同時將那些改變應用於存放在檔案系統A上之資料，以建立檔案系統B之資料複本。在應用這些改變之後，檔案系統A會取得檔案系統上之資料快照(在這個例子中為快照3)。如果在這個例子中之檔案系統A之先前快照不存在於檔案系統A上，檔案系統B之整個備份資料集合會在檔案系統B中產生，並且連接至檔案系統A，進而回復檔案系統A。

系統管理員現在可以重新建立檔案系統A，以作為主要檔案系統，並且建立檔案系統B為備份系統。圖15D顯示在本發明實施例中執行恢復之恢復處理流程圖1590。在步驟1568中，這個流程會從待命檔案系統更新處理1540繼續執行，並且進行至步驟1570。在步驟1570中，檔案系統B會取得快照4，以獲得目前的主檔案系統(即檔案系統B)之狀態。檔案系統B則會產生一組改變，介於這快照、快照4，以及連接至檔案系統A之最後快照。在步驟1574中，這組改變會被檔案系統A接收，並且應用至存放在檔案系統A中之資料。在這組改變被應用之後，檔案系統A會取得快



五、發明說明 (44)

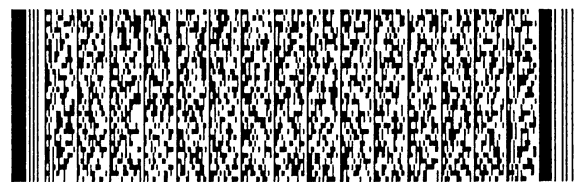
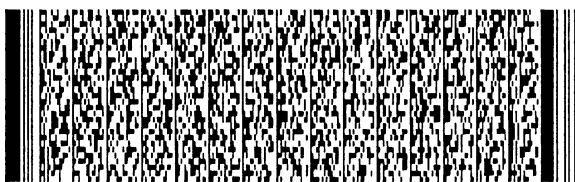
照4，以在檔案系統A中保存一個乾淨的快照。在步驟1578中，恢復程序然後會開始使得檔案系統A再度變成主檔案系統。在步驟1580中，檔案系統B會回復為待命檔案系統，而在步驟1582中，檔案系統A會回復成主檔案系統。在步驟1584中，流程會終止。

可以明顯看出的是，上述執行熱待命檔案系統之快照資料集合也可以用於其他目的。除了用於熱待命檔案系統之快照外，實施例之流程也可以在不同的時間取得額外的快照，而不將那些快照用於維持熱待命檔案系統，例如，其他快照可以用來製造檔案系統之磁帶備份。

動態快照連結

實施例之操作產生原始檔案系統之數個快照，每一快照反映檔案系統在取得快照時之狀態。本發明之不同實施例使用不同的技術來組織和辨識這些多個快照，藉由使用指派給原始檔案系統中之檔案的相同檔案名稱，但是增加有些額外的附註(context)以區別原始檔案與存放在快照中的版本，有些實施例可以辨識快照資料集合中之檔案。不同的實施例會以數種方式呈現這個額外的附註給用戶，例如定位在原始檔案系統中，但是卻位於一個對應於快照辨識碼之特殊目錄下之快照。

其它實施例藉由一個不同的檔案系統或經由相同檔案系統之不同的安裝點(mount point)來存取快照資料集合，以呈現附註給用戶。在內部中，在這些實施例之檔案

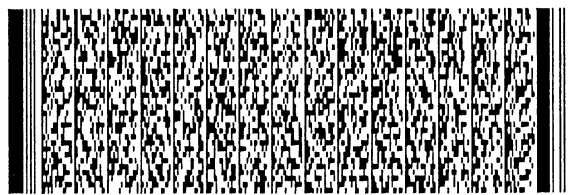
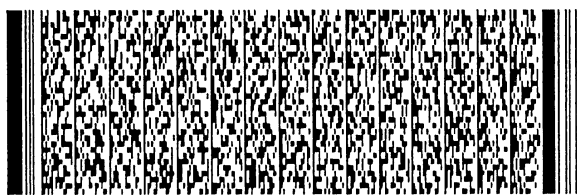


五、發明說明 (45)

系統會藉由某種方式編碼額外的附註，例如在資料節點號碼中排列在前面 (high-order) 之位元，或者編碼在單一檔案系統號碼中之附註，來區別快照資料集合中之檔案與現正開啟檔案系統中之原始檔案。用來單一地辨識快照的編碼資訊在這裡稱為快照辨識碼。

本發明實施例會使用動態快照連結來存取快照，並且協助辨識所需的快照資料集合和相關的快照辨識碼。這些實施例會使用二種類型之動態快照連結其中一個或兩個，其中動態快照連結包含：「快照最新連結」，為一個檔案之先前快照版本的單一連結；以及「快照連結目錄」，為一個包含連結項目之目錄，其中連結項目係用於一個或多個可用快照中之檔案。這些實施例之部分會結合這兩個變形 (variant)，因此快照連結目錄也包含快照最新連結，用來輕易地辨識目標之最新版本。

本發明其他可選擇之實施例使用動態快照鏈來指向快照，此快照是藉由不同的技術來取得和維持的，包含不同於本說明書中其他部分所描述之快照結構。例如，使用快照資料節點檔案之快照可以使用動態快照連結，以有效率地指向數個使用這種結構之快照資料集合，其中快照資料節點檔案包含一組完整資料節點，而每一資料節點係指向現正開啟檔案系統或後續快照中之資料區塊，包含使用指標計數器 (reference counter) 之系統，其指標計數器用來維持快照資料節點數量之計數，以指向一個特定的資料區塊，確保在刪除資料區塊之前，沒有任何資料節點指



五、發明說明 (46)

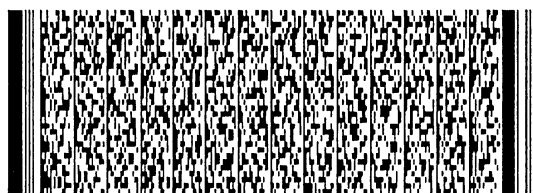
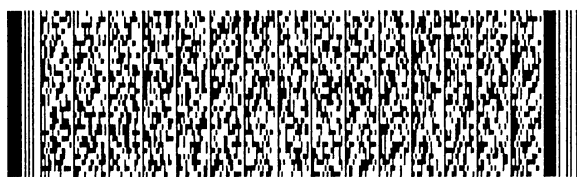
向該資料區塊。

快照最新連結

如同一個傳統檔案系統連結，本發明實施例所使用之快照最新連結為一個檔案系統目錄中之項目，其中檔案系統目錄將連結名稱映射至檔案系統中之一特定檔案(或目錄)，而傳統連結係靜態地將名稱映射至特定檔案的資料節點號碼以及特定的快照辨識碼。相較於傳統連結，快照最新連結將連結名稱映射至一個特定檔案之資料節點號碼，但是並不將快照辨識碼束縛在該連結，直到系統執行存取與那個快照最近連結相關之檔案的操作。存取檔案之操作會啟動那個連結，其中該檔案為快照最新連結之目標，藉由特定快照資料集合中之目標檔案(如資料節點檔案)，進而觸發執行實施例之連結的解析

(resolution)，這會導致快照最新連結在讀取目錄時，會動態地指向目標檔案之最近快照版本。

實施例中快照最新連結以相似於產生傳統目錄連結的方式來產生的，也就是指定兩個目標檔案(或目錄)以用於連結以及即將被指派至此連結之名稱，產生之附註中之連結係用於已經存在之檔案或目錄。這些實施例之檔案系統會維持連結之計數，其中連結係用於每一檔案和目錄。在產生一個新連結時，目標檔案之連結計數會增加，當刪除連結時，連結計數則會減少。連結計數可確保檔案或目錄之所有連結被刪除前，不會刪除目標檔案或目錄。維持一

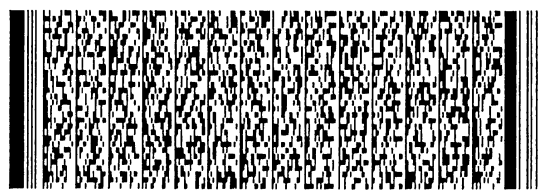
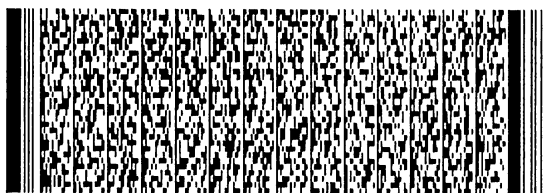


五、發明說明 (47)

個快照最新連結之實施例會確保由快照最新連結所指向之檔案亦指向原始檔案系統之最近快照版本，當產生快照最新連結時，如果沒有可用的快照，這些實施例會延遲連結的產生，直到讀取快照目錄。

圖12A為示範性快照最新連結資料結構圖1200。示範性快照最新連結資料結構1200顯示三個連續快照：快照 $i-1$ 1202、快照 i 1212和快照 $i+1$ 1222。這三個快照在不同的時間被取得，且每一個在某個時間點都是最新的快照，即在時間 $i-1$ ，快照 $i-1$ 1202是最新的快照；在時間 i ，快照 i 1212是最新的快照；而在時間 $i+1$ ，快照 $i+1$ 1222是最新的快照。特定檔案之示範性快照最新連結(動態連結1240)係指向最新快照，快照最新連結顯示在時間 $i-1$ ，當快照 $i-1$ 1202變成最新快照時，快照最新連結(即動態連結1240)會指向快照 $i-1$ 1202。接著在時間 i ，將快照 i 1212變成最新快照，而快照最新連結1240會被更新，以指向快照 i 1212。在時間 $i+1$ ，當快照 $i+1$ 1222成為最新快照時，快照最新連結1240會被指派至快照 $i+1$ 。在這個例子中，與快照最新連結1240有關之檔案的所有快照包含相同磁碟位址，以用於與檔案有關之資料節點，因此在這三個快照中，所有映像資料節點表示一個指向目前檔案系統資料節點之指標，其中檔案資料係存放在檔案系統中。

圖12B為本發明實施例之快照最新連結資料擷取處理控制流程1250。控制流程1250從步驟1252開始，並且直接進行到步驟1254。快照最新連結資料擷取處理控制流程



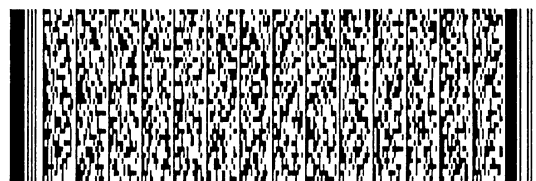
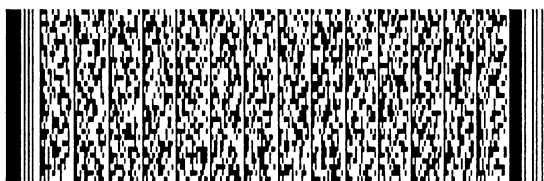
五、發明說明 (48)

1250 說明如何從一個被請求之檔案的最新快照擷取資料之程序。在步驟1252中，程序開始於啟動快照最新連結1240，其指向包含與快照最新連結1240有關之檔案的最近可用快照。檔案系統會在步驟1256中擷取最近快照之快照辨識碼，然後在步驟1258中擷取最近快照資料集合。接著，程序會進行至步驟1260，以判斷先前被擷取之快照資料集合中的資料節點是否包含有效資料節點資料。如果被擷取之映像資料節點確定包含有效資料，那麼程序會繼續進行至步驟1262，判斷快照資料節點中之磁碟位址是否包含非相同值(即有效值)。如果確定資料節點具有一個有效磁碟位址，程序會繼續進行，以讀取包含在快照資料節點中之磁碟位址所指向之資料區塊。

如果在步驟1260中，確定擷取之映像資料節點不包含有效資料節點資料，或者在步驟1262中，確定映像資料節點不包含一個有效磁碟位址，那麼在步驟1266中，程序會存取目前的檔案系統之資料節點，來判斷儲存檔案之磁碟位址。然後程序會在步驟1268中從目前檔案讀取資料。

快照連結目錄

快照連結目錄是一個包含指向存放在快照資料集合中之檔案及/或目錄連結之子目錄。快照連結目錄可以存在於原始檔案系統中或一個不同的檔案系統。這些實施例將項目存放在原始檔案(或目錄)之每一快照版本中的快照連結目錄。快照連結目錄的每一項目將單一快照名稱映射至

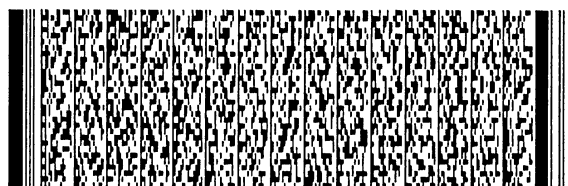
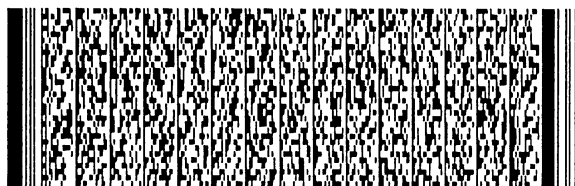


五、發明說明 (49)

存放在快照中之資料檔案版本。當新的快照被產生或舊的快照被刪除時，快照連結目錄中之項目亦會改變，使得快照連結目錄可維持原始檔案之可用快照版本的完整名單。

這些實施例之快照連結目錄藉由指定連結之目標檔案(或目錄)，以及指派至快照連結目錄之名稱來產生的。連結之目標檔案(存放在或由快照資料集合所指向之檔案)之連結計數會增加以反映定位檔案之新連結。一旦快照連結目錄被產生，母目錄會維持子目錄之項目，其中子目錄為新的快照連結目錄。母目錄中指向快照連結目錄之項目是靜態的，而本發明之這些實施例指向快照連結目錄之連結被當作一個常規目錄項目來儲存。然而，這些實施例並不儲存靜態項目於快照連結目錄中，而是每當有存取請求進入目錄時，例如當快照連結目錄被讀取時，就會產生快照連結目錄項目。這些實施例會產生一個項目於原始檔案之每一可用快照版本的快照連結目錄中，在快照連結目錄中之項目會將快照名稱映射至快照資料檔案資料節點之資料節點和快照辨識碼。如果沒有可用的快照，快照連結目錄不會產生任何用於快照之項目。在實施例中，讀取不具有快照之快照連結目錄會產生本身(". ")和其母(".. ")標準目錄項目。

在傳統檔案系統中之資料節點號碼係用來在給定時間內單一地辨識一個檔案。然而，在檔案系統的操作中，刪除檔案以及產生新檔案都是非常常見的，在這樣的操作期間，被刪除的檔案之資料節點有時候會被重新配置至一個



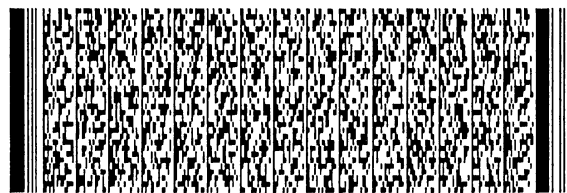
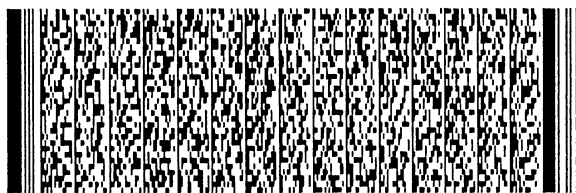
五、發明說明 (50)

新檔案，因而重複利用那個資料節點號碼。為了區別這兩個無關的檔案，檔案系統一般會保留每一資料節點之「產生號碼」，每當資料節點被重複利用時，其產生號碼就會增加。由於在檔案存活期間產生號碼會維持不變，因此實施例的檔案系統利用它來判斷存放原始檔案版本之可用快照的名單。使用快照連結目錄之本發明實施例只會產生項目於具有相同資料節點號碼與相同產生號碼之檔案的快照連結目錄中，當快照連結目錄產生時，包含特定檔案版本之最舊可用快照會被決定，而對應於那個檔案之值會被存放在用於後續流程之快取記憶體中。存放在快取記憶體中之值允許快照名單之產生，而無須一個後續的詢問來決定哪個快照包含目標檔案的版本。

本發明有些實施例所使用之動態快照連結是提供用戶一個裝置之有效方式，其中該裝置係用來定位檔案之所有快照版本或定位最近快照版本。相較於使用傳統檔案系統，使用動態快照連結提供一個好處，由於傳統檔案系統的連結是靜態的，因此當產生新快照和刪除舊快照時，這些連結會變得過時和無效。

寫入時移動資料取得流程 (Move-On-Write Data Capture Processing)

本發明有些實施例使用寫入時複製 (copy-on-write) 技術，以更有效率地配置與維持快照資料集合。因為未被修改的資料並不需要被複製，所以這個技術只會

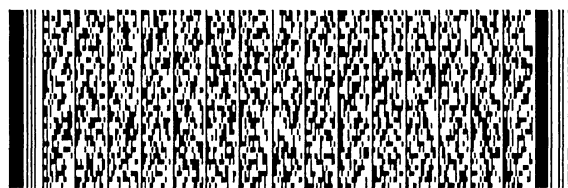
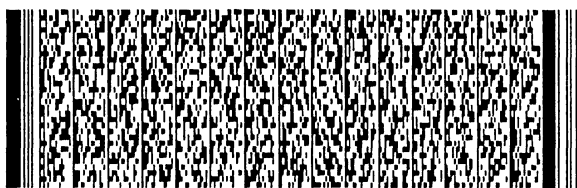


五、發明說明 (51)

將修改過的檔案系統資料複製至快照檔案。使用寫入時複製更新技術會使得「包含」資料區塊之快照資料集合會被存放在一快照資料集合中及仍舊是現正開啟檔案系統之一部分的資料區塊中。當那些區塊中之資料被修改時，在快照資料集合中之資料區塊就會被複製或移動。如果原始檔案資料從產生快照以來未曾被修改過，本發明之實施例亦具有快照資料集合用來指向位於原始檔案中之資料。

圖13A為本發明之實施例在寫入時複製處理控制流程1300。控制流程1300係從步驟1302開始，並且直接進行至步驟1304。在步驟1304中，這個實施例所使用之寫入時複製操作之流程藉由配置即將成為快照資料集合之一部分新的資料區塊來開始。在步驟1306中，流程會將資料從原始檔案之資料區塊複製至記憶緩衝器。在步驟1308中，記憶緩衝器中之資料複本會被複製至先前配置之新資料區塊。在步驟1310中，此流程會修改記憶緩衝器中之資料。而在步驟1312中，被修改過的資料會寫入至原始資料檔案。本發明其他可選擇之實施例會將新資料緩衝器指派給原始資料檔案，並將未被改變的原始資料區塊移動至快照資料集合。指派原始資料區塊係藉由更新資料節點中之區塊指標或原始檔案之間接區塊來進行。

如果整個資料區塊即將被修改，則會使用本發明其它實施例所使用之第二方法。第二方法包含將舊區塊直接移動至快照，而無須將內容讀取至電腦的記憶體，也無須將資料寫入至一個新位置。在移動舊資料區塊之後，一個新



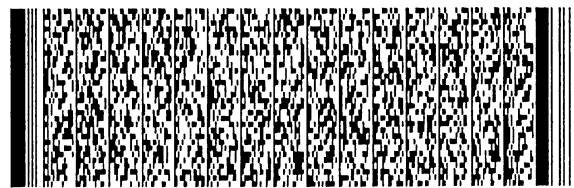
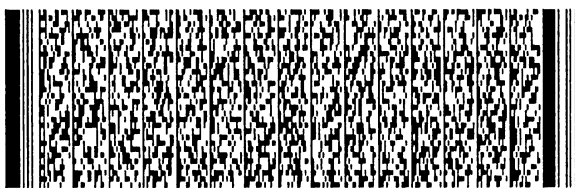
五、發明說明 (52)

的未初始化的區塊會被指派給修改過的原始檔案。這個方法係稱為「寫入時移動」操作，並且實質上會快於寫入時複製，而且使用較少電腦資源。

相反於寫入時複製操作，寫入時移動操作僅僅配置一個新區塊給檔案，同時更新資料節點中之區塊指標或原始檔案之間接區塊。舊區塊被移動至最新快照，以取代存放在那裡的相同位址。既然檔案的任何較早快照先前指向原始區塊並且仍然是如此，因此它們可以維持不變。寫入時移動避免讀取原始區塊的內容，並且實質上會快於寫入時複製之操作。

如果流程決定整個資料儲存區塊即將被修改，則會使用寫入時移動之方法。這寫入時移動方法開始於首次修改原始檔案中之資料區塊，以保存快照資料集合之原始資料。

圖13B為本發明實施例在寫入時移動處理控制流程1320。控制流程1320從步驟1322開始，並且直接進行至步驟1324。在步驟1324中，流程藉由將原始資料區塊移至快照資料集合來開始，其中該原始資料區塊為一開始包含在被修改之來源檔案中的資料區塊，這個實施例會藉由更新存放在快照資料集合中之資料節點中之磁碟位址指標值，來完成此移動。在步驟1325中，在磁碟區塊被移至快照資料集合之後，流程會藉由配置一個新資料區塊給來源檔案來繼續進行。然後，流程會繼續進行至步驟1325，配置檔案系統上之空間給新資料區塊，而在步驟1326中，配置一



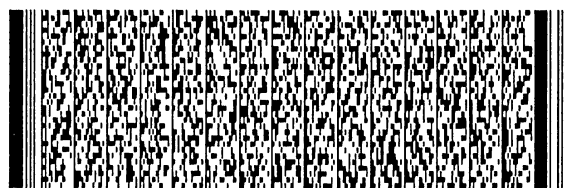
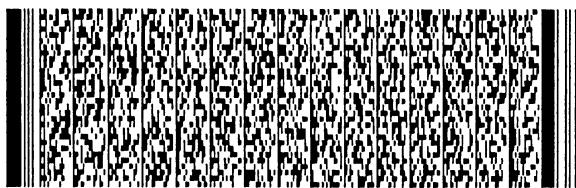
五、發明說明 (53)

記憶緩衝器給新資料區塊，其中被更新的資料會暫時存放在新資料區塊。實施例中之記憶緩衝器一開始為零，或者被標記 (mark) 起來，以表示記憶區塊中沒有任何資料是有效的。在步驟1327中，來源檔案資料會被複製至記憶緩衝器。在步驟1328中，當其位於記憶緩衝器中時，存放在記憶緩衝器之來源檔案資料會被修改。在步驟1329中，修改過的資料會被複製至最新配置的資料區塊。

如果原始資料檔案是以片斷的方式被寫入，即一次一個記錄，執行上述之寫入時移動操作之本發明有些實施例版本就不會使用這個更有效率的技術。在片斷式寫入資料之情況下，對原始檔案的第一次寫入只會修改區塊的一個部份。由於檔案系統流程無法判斷將來是否會對相同區塊進行另外的寫入操作，因此，這些實施例中之檔案系統會保守地運作，並且利用較為昂貴的寫入時複製方法，來複製原始資料至快照。

本發明有些實施例使用寫入時複製/寫入時移動方法之變形，在此稱為延遲性寫入時複製操作 (deferred copy-on-write operation)。當檔案是以片斷的方式寫入時，這個變形會延遲寫入時複製操作，延遲性寫入時複製操作通常會使得檔案系統使用更快的寫入時移動操作，而非一個或多個寫入時複製操作。

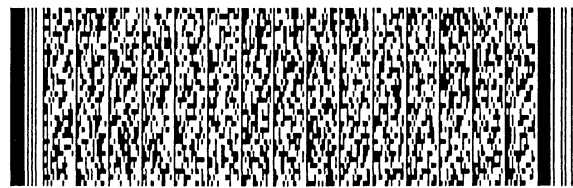
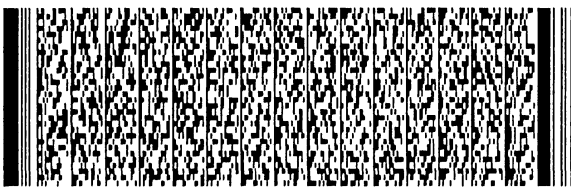
這些實施例之延遲性寫入時複製操作係運作於原始資料檔案區塊之修改。在實施例中，當首次使用延遲性寫入時複製操作來修改原始資料檔案區塊時，原始資料檔案區



五、發明說明 (54)

塊會立即移至快照資料集合，而一個新的尚未初始化之區塊會被指派給原始檔案。然後檔案系統會記錄在一變數中，其由實施例之檔案系統來維持，而用於修改資料區塊的寫入時複製操作已經被延遲。接著檔案系統程序將新資料寫入至新區塊之一部份，由於新資料即將存放於該新資料區塊中，因而使得有效資料存放在那個新的但是尚未初始化的區塊的部份。如果藉由將有效資料寫入至整個區塊，而後續寫入至新區塊使得那個資料區塊變成有效，那麼整個區塊將會包含有效資料，並且不需要任何寫入時複製操作。在這些情況下，延遲性寫入時複製操作可以有效地消除一次或多次的重複執行較為昂貴的寫入時複製操作。

執行延遲性寫入時複製操作之實施例在延遲性寫入時複製操作之間會用不同的方式表示一個特定區塊。本發明有些執行延遲性寫入時複製操作之實施例會在區塊的磁碟位址中保留一個位元，用來表示該區塊是否位於延遲性寫入時複製操作之間。如果那個位元是開啟的，則表示區塊位於延遲性寫入時複製操作中，而從那個區塊讀取資料可能會另外需要包含將資料與快照資料合併之處理。除了表示每一個區塊是否因延遲性寫入時複製操作而不完整之外，本發明之實施例也會為每一檔案保留一個變數，用來表示在該檔案中之任何資料區塊是否處於利用延遲性寫入時複製操作來更新之程序。這個變數可表示缺乏一個完整的來源檔案資料區塊，即來源檔案中至少有一個資料區塊

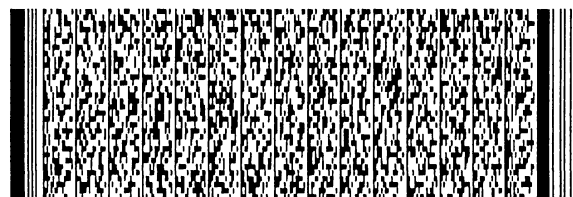
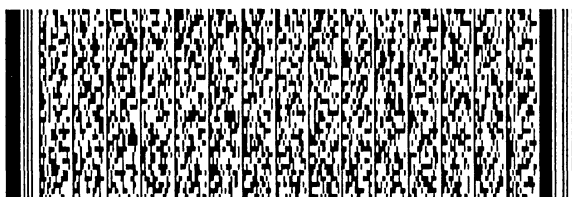


五、發明說明 (55)

不是完整的。這個變數可藉由避免一個區塊一個區塊的去確定每一個資料區塊是否完整且有效的，使得資料存取更有效率。

執行檔案系統之延遲性寫入時複製操作之實施例會記錄新區塊的哪個部份已經被寫入而是有效的。在有些實施例中，記錄已經被寫入之部份是藉由追蹤區塊中特定位元組的改變來執行的。然而，其它實施例之設計係利用啟發式觀察(heuristic observation)，來觀察由於檔案連續地被寫入，因此如果區塊第一位元組已被修改，則可判斷延遲性寫入時複製，故這些實施例之流程只存放偏移量(offset)在最後被修改的位元組。藉由觀察到檔案通常為整體地被寫入，有些實施例可以實現額外的處理效率，因此只要那個區塊係暫時存放在揮發性記憶體中，檔案系統便可藉由維持包含在區塊中之有效資訊的範圍來操作。如果在將區塊寫入磁碟以作為磁碟快取操作之部分前，區塊尚未被完全地重寫，這些實施例也會導致傳統寫入時複製操作，或將來自新緩衝器之資料合併至預先存在的資料區塊中。本發明之其它實施例會將資料之有效範圍存放在區塊中，以作為檔案資料節點之一部分，來支援寫入時複製之延遲，直到區塊已經實際地被讀取。清楚可見的是，一連串的執行是可能的。

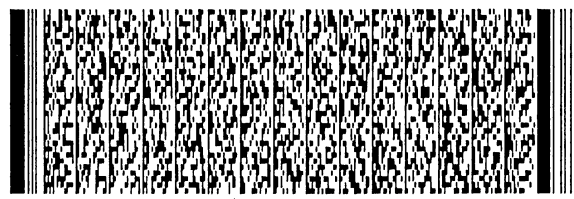
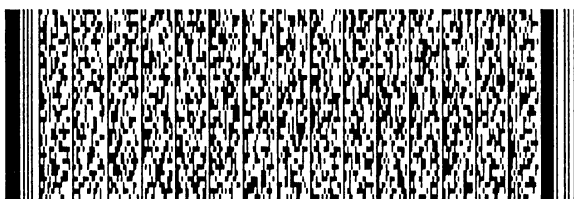
如果在區塊包含整個有效的資料之前，檔案系統需讀取正在延遲性寫入時複製操作之區塊時，檔案系統會偵測區塊正處在延遲性寫入時複製操作之間，為了回應偵測出



五、發明說明 (56)

區塊不是完全有效的，在這些實施例中之檔案系統會讀取儲存在原始區塊中之資料，並且與被寫入至新區塊之新資料合併，以達到最新且有效資料的讀取請求。當讀取處於延遲性寫入時複製操作間之資料區塊時，這種情況會造成較不具效率的操作，但是這樣的情況在系統操作中是罕見的，有利的是，大多數的檔案都是連續且完整地寫入。同樣可以觀察到的是，寫入請求和讀取請求同時發生是罕見的，因此，延遲性寫入時複製操作之技術大致可以提昇維持一檔案系統快照之整體效率。

執行延遲性寫入時複製操作之實施例的檔案系統會在完成區塊寫入之前，以不同的方式來讀取正在延遲性寫入時複製操作間之區塊之請求。如果資料係完全位於區塊之有效部份，即新資料已經被寫入至新檔案資料區塊之其中一部份，則請求讀取資料之處理(從新資料區塊擷取資料)是利用特定的流程來進行。如果讀取資料之請求要求存取橫跨區塊無效範圍(新資料沒有被寫入的範圍)之資料，或者請求讀取完全位於無效範圍內之資料，讀取操作會從原始資料區塊獲得部分或所有資料，其中原始資料區塊係存放或移至快照中。雖然檔案系統無須使用寫入時複製便可滿足讀取請求，但是寫入時複製的成本在於讀取原始資料至記憶體中，由於必須支付這個成本以滿足讀取請求，因此檔案系統可以執行具有相對較少額外工作的寫入時複製。一旦完成寫入時複製，或者當整個區塊已經重寫，有效範圍的資訊會被排除，而延遲性寫入時複製會取消，使



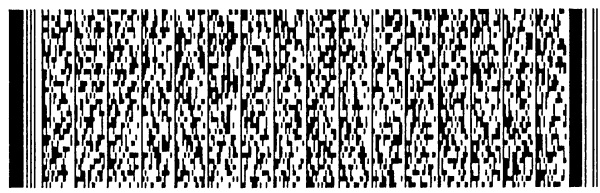
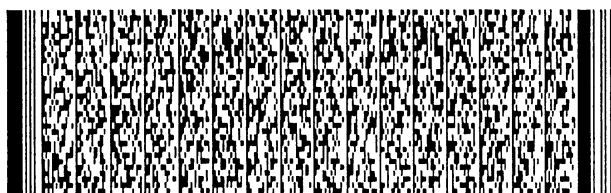
五、發明說明 (57)

得後續的讀取可以繼續進行，而不會受到干預。

藉由部份地存放完整區塊於磁碟上來執行延遲性寫入時複製操作之實施例中，允許原始檔案之部分資料只存在於快照中，即不同於上述之部分實施例，只有當區塊暫時存放在系統記憶體中時，才會延遲寫入時複製的處理。這個情況會導致如果原始區塊係移至快照資料集中（如上所述），現在作為原始檔案的部分之新資料區塊只具有最新寫入的資料，而舊資料只存在於先前已經移至快照的資料區塊中。這樣的實施例在刪除快照之前（並且潛在地排除存放在快照中之區塊）會對所有區塊執行寫入時複製，而其區塊在刪除快照之前，仍然被標記為要求延遲性寫入時複製操作。當如上所述之實施例在執行延遲性寫入時複製至磁碟時，其在沖洗磁碟存取記憶緩衝器至磁碟前所執行之寫入時複製不會受到這個情況的支配，因此在那些實施例中之快照可以被刪除，而無須檢查在現正開啟檔案系統中之檔案。

利用一般存取模式來存取檔案，在許多情況下可以減少寫入時複製操作之處理資源成本，這實質上減少了維持檔案系統之快照一致性的經常性花費（overhead）。

圖13C為本發明之一實施例在已經發生寫入時移動後，執行讀取程序之操作和控制流程1350之流程圖。控制流程1350係從步驟1352開始，並且直接進行至步驟1354。在步驟1354中，流程會藉由接收讀取來源檔案之請求而開始，其中該來源檔案目前正處於寫入時移動之程序中。此

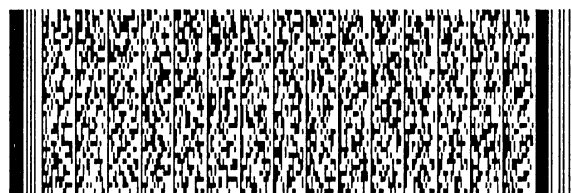
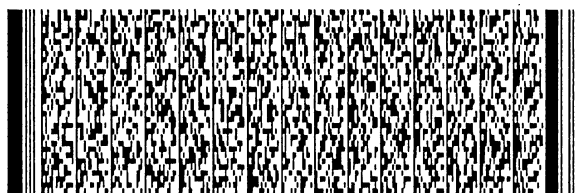


五、發明說明 (58)

示範性流程會在步驟1356中，從來源檔案讀取資料區塊至第一記憶緩衝器。接著進行至步驟1358，根據資料區塊中之更新活動，判斷讀取請求所要求之所有資料是否皆由指向來源檔案資料區塊有效部份之儲存指標所提供。如果所請求的資料係位於來源檔案之資料區塊的有效部份，程序則會在步驟1364中終止，而第一緩衝器中之資料則會被使用。否則，程序會進行至步驟1360，將對應於被請求資料之資料從快照中之資料區塊讀取至第二緩衝器。在步驟1362中，流程會從第二記憶緩衝器中之資料將其複製至第一記憶緩衝器之無效部份。其它可選擇的實施例並不使用第二記憶緩衝器，且將無效資料直接從快照資料集合複製至第一記憶緩衝器。

快照序列 (Snapshot Sequence)

本發明實施例取得快照之檔案系統通常由許多用戶來共同分享，這些檔案系統包含數個具有不同儲存需求的檔案，因此檔案系統管理員通常會在不同時間且/或間隔時段製作檔案的備份。為了支援在不同時間且/或間隔時段執行資料備份，本發明有些實施例會選擇性地產生僅包含在檔案系統中某些檔案之快照，例如有些用戶要求更為頻繁的備份。由於其它類型的檔案是臨時檔案，因此會被選出而不包含在備份中，故也不包含在快照中。因為資料庫應用程式具有其他裝置可以用來回復資料，所以資料庫檔案有時候會被選出而從備份中排除，此外，檔案系統管理



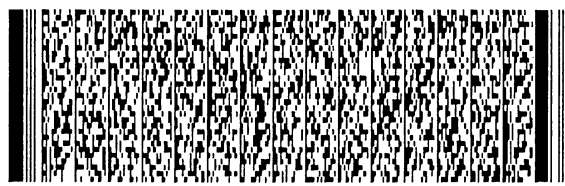
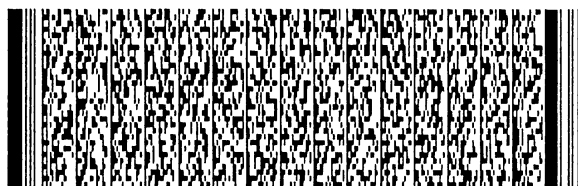
五、發明說明 (59)

員也不希望發生將這種資料庫檔案存放在快照中的成本。取得多個快照序列(每一取得只是檔案系統的一部分)允許不同的用戶可以具有不同的快照政策，或者可以完全從快照機制中排除。

本發明之實施例允許產生和取得僅是檔案系統之一部分的快照。這些實施例支援在檔案系統中切割檔案，因此只有在給定分割區中的檔案和目錄會包含在一特定快照中，這些快照(只包含在檔案系統中之一個檔案子集合)在此稱為快照序列或SnapSequence。在本發明的實施例中，檔案系統可以具有任何數量的SnapSequence，但是一個檔案只能屬於一個SnapSequence，或者不屬於任何SnapSequence，即從未被包含在快照中。執行快照序列之實施例中，快照序列可以被安排，使得每一SnapSequence可以產生、刪除或回復其快照，而與其它序列無關。在這些實施例中之SnapSequence還可以階層性地被安排，因此母序列中的快照操作可以操作於其子序列。

圖14A為本發明實施例之快照序列階層維恩圖(Venn diagram) 1400。維恩圖1400說明一個檔案系統102具有二個最高層級快照序列：SnapSequence 1 1404和SnapSequence 4 1410。二個子快照序列SnapSequence 2 1406和SnapSequence 3 1408則是完全包含在SnapSequence 1 1404中。為了確保檔案系統圖像之一致性，這些實施例之階層是嚴格的，且不可以相互重疊。

支援快照序列(SnapSequence)的本發明實施例係利用

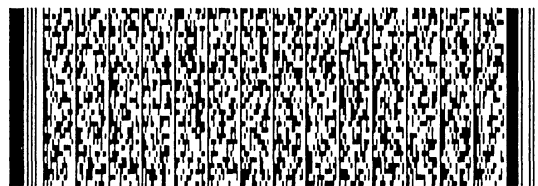
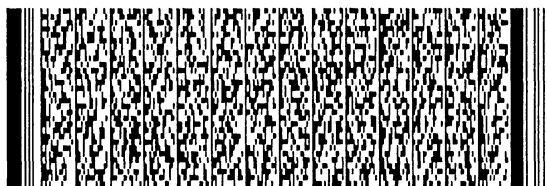


五、發明說明 (60)

單一的快照序列辨識碼來辨識每一個SnapSequence，在此稱為SnapSequenceId。SnapSequenceId用來辨識每一個快照相對於其它快照序列中的快照被取得之時間。每一SnapSequence具有自己的MaxSnapId值，其等同於已經產生於特定序列中之最大快照辨識碼值（即該序列中最近產生之快照之快照辨識碼）。MaxSnapId值分別存放在本發明之實施例中，以在刪除最近快照的情況下，保存最近快照辨識碼值。如果快照序列的最近快照被刪除，在那個序列中之下一個快照會被指派至MaxSnapId值之後的下一個快照辨識碼，以保存快照辨識碼編號，而不會重複使用被刪除快照的快照辨識碼。每一快照序列也包含該快照序列可用快照之名單。支援SnapSequence之實施例的檔案系統會將SnapSequence之SnapSequenceId和檔案快照辨識碼（於此稱為FileSnapId）之資料領域存放在原始檔案之資料節點中。

本發明之實施例會經由一個明確指令或從母目錄繼承SnapSequenceId，來指派一個檔案至一快照。當該檔案被指派至一快照序列，使得該檔案被包含在序列中之所有後續快照時，檔案之FileSnapId會被初始化至MaxSnapId。這些實施例之檔案系統會將一個檔案的FileSnapId與快照序列之MaxSnapId相比較，其中當檔案被修改，以將檔案的原始內容保存在序列之快照中時，該檔案會被指派至快照序列。

在本發明之實施例中，一個新快照序列係經由一個明

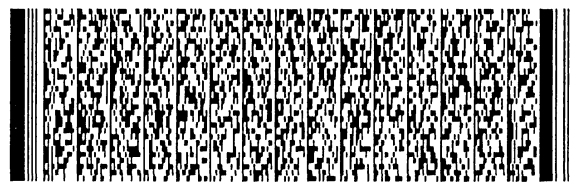
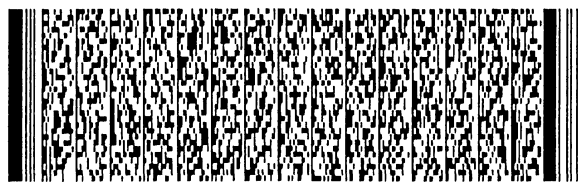


五、發明說明 (61)

確指令來產生。當建立一個新快照序列時，實施例之檔案系統會將新快照序列之存在記錄於檔案系統的超區塊中，並指派SnapSequenceId至新快照序列，同時初始化其MaxSnapId值。如果新快照序列是另一個快照序列的子序列，其母序列也會被指定出來。若無母序列被指定至快照序列，則此快照序列會與其他快照序列分開且無關。實施例之檔案系統會將快照序列間的關係記錄在超區塊中，因此在對快照序列的後續操作可以包含所有的子快照序列。在超區塊中的快照序列的連結使得子快照序列的部分檔案可以包含在一個或多個母快照序列中，故不須保留檔案資料在子和母快照序列中之複本。

例如，母快照序列和子快照序列為具有三個用戶：Able、Baker和Charlie的檔案系統，每一個用戶具有自己的快照序列。檔案系統管理員會定義一個標記為「用戶」之母快照序列，並且指派Able和Baker之快照序列作為用戶快照序列的子序列。在這個例子中用戶序列之快照包含Able和Baker之快照序列，但是不包含Charlie之快照序列。

包含在一快照序列中之資料檔案具有快照序列之SnapSequenceId和FileSnapId值，其中該快照序列係存放在資料檔案之資料節點中。圖14B為本發明之實施例中，用於母快照序列和子快照序列中之檔案資料節點資料結構1420之方塊圖。這個例子之快照資料集合1422包含二個快照序列SnapSequence 1 1430和SnapSequence 2 1444。如



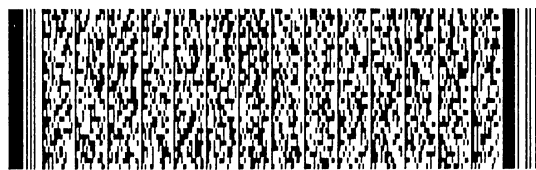
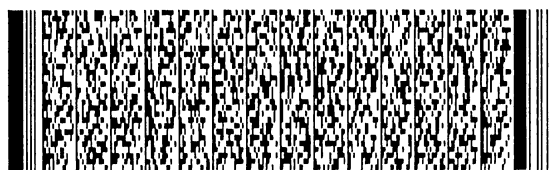
五、發明說明 (62)

同在檔案系統之超區塊中之快照序列資料中所描述的，這個例子之SnapSequence 2 1444為SnapSequence 1 1430的子序列。

在這個例子中，SnapSequence 1 1430包含二個資料檔案，其具有inode 1 1432和inode 2 1438。資料檔案之每一資料節點包含SnapSequenceId和FileSnapId之項目，例如inode 1 1432具有第一SnapSequenceId 1434和第一FileSnapId 1436，inode 2 1438也具有第二SnapSequenceId 1440和第二FileSnapId 1442。在第一快照序列中之每一檔案具有等於1的SnapSequenceId值，以表示它們所屬於的快照序列。子快照序列SnapSequence 2 1444在這個例子中具有與第三資料節點 1446相關之單一檔案，其中第三資料節點 1446具有第三SnapSequenceId 1448和第三FileSnapId 1450。這個例子中之SnapSequenceId 1448係被設定為2，以表示這個檔案屬於第二快照序列SnapSequence 2 1444。第二快照序列SnapSequence 2 1444也包含一個母連結1452，其指向該快照序列之母快照序列。

快照序列中之快照建立

在快照序列中產生一個快照基本上係使用相同於產生一標準快照之處理，序列之MaxSnapId會增加，並且被指派至新快照，而新快照之資料節點檔案會被產生，並且該快照會被加入至超區塊中之序列可用快照名單中。



五、發明說明 (63)

至於具有子序列之快照序列，檔案系統會產生一個新的快照，以用於使用相同程序之每一子序列。子快照連接至母快照，以確保整個集合為一個單一單元來回復或刪除。

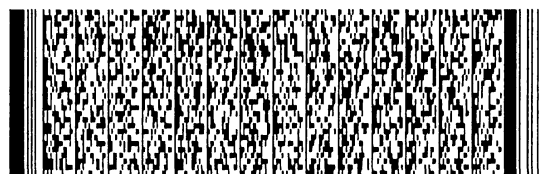
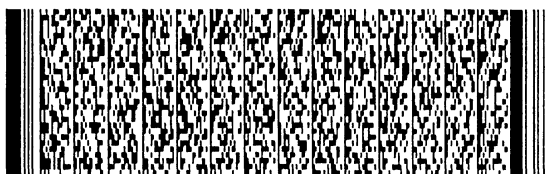
快照序列中之快照刪除

從一快照序列中刪除快照也會從每個子序列中刪除相對應的快照。當子快照序列中之快照被刪除時，本發明之實施例會刪除母快照序列中之相對應的快照，以維持一致的檔案系統。

快照序列中回復一快照

藉由取消配置 (de-allocating) 原始檔案系統中所有被改變的資料節點和資料區塊，並且回復存放在快照序列中被保存的值，本發之實施例可以將一快照序列回復至檔案系統，具有子序列之快照序列也必須從每個子序列中回復相對應的快照序列。

快照序列可能不包含出現在目前檔案系統之所有檔案，正被修改之檔案(或目錄)的原始內容被保存在對應於該檔案快照中之資料節點的快照，因此，快照回復操作會將資料節點返回至其保存狀態。在快照產生後刪除保存檔案的情況下，資料節點不再處於使用中的狀態，或被指派至不同的檔案或目錄，因此，檔案系統無法自動回復被保存的快照資料。由於檔案系統中之所有檔案並不必然會

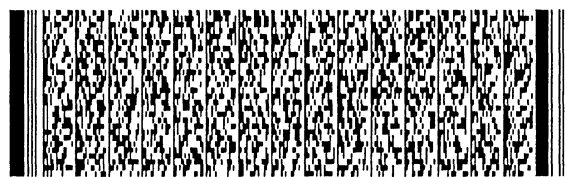
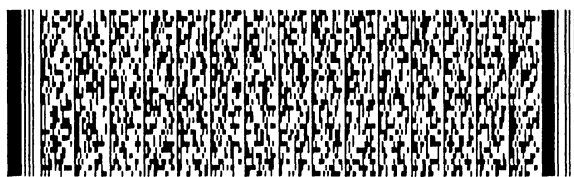


五、發明說明 (64)

包含在快照序列中，在從快照序列回復一檔案系統之後，支援快照序列之實施例會執行額外的流程，以維持檔案的一致性。

執行快照序列的本發明實施例會移動 (traverse) 正被回復之快照，以判斷哪些儲存在快照中之資料節點屬於正被回復之快照序列。如果目前檔案系統中之資料節點屬於相同序列，或屬於正被回復快照序列階層中之另一個序列，同時如果那個資料節點不是正被現正開啟檔案系統中之其它檔案使用中，檔案系統會回復檔案至快照中之相同的資料節點。然而，若現正開啟檔案系統中之資料節點未配置在正被回復快照序列中之檔案的情況下，檔案系統會重新定位 (relocate) 被回復的檔案，從快照擷取至一個在現正開啟檔案系統中未被配置之資料節點。這些實施例之檔案系統最好能保留重新定位和原始資料節點號碼映射至被重新配置之資料節點號碼之檔案名單。當所有檔案從快照中被回復時，檔案系統會移動現正開啟檔案系統之資料節點檔案，以定位所有屬於被回復序列的目錄。每一個被回復的目錄會被掃描，以定位在回復時被重新定位之檔案的目錄項目，而目錄項目會被更新，以連接至在回復時被重新定位的檔案中被重新定位之資料節點。

圖14C為本發明實施例之快照序列回復處理之控制流程1460，其係用於回復來自快照之資料。控制流程1460係從步驟1462開始，並且直接進行至步驟1464。在步驟1464中，示範性快照序列回復流程藉由存取快照序列之第一資

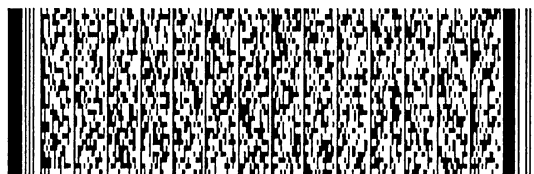
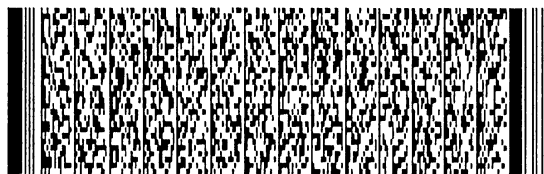


五、發明說明 (65)

料節點來開始。在步驟1466中，流程會判斷現正開啟檔案系統中相對應資料節點是否與正被回復之SnapSequence之檔案有關、是否位於正被回復之SnapSequence之任何子快照序列、或者是否現正開啟檔案系統中相對應資料節點未被指派。如果判斷為"是"，即現正開啟檔案系統中相對應資料節點為正被回復之快照序列的一部分，或是正被回復之快照序列中之子快照序列之一部份，或者未被指派，在步驟1468中，快照資料會被回復至使用該檔案相對應資料節點之現正開啟檔案系統。在步驟1470中，如果正被回復之檔案資料節點被指派，流程會回復快照資料至現正開啟檔案系統中未被配置之資料節點。

在回復上述資料節點至現正開啟檔案系統中相對應資料節點或未被指派的資料節點後，在步驟1472中，流程會判斷該資料節點是否為快照序列中之最後資料節點。如果該資料節點不是快照序列中之最後資料節點，在步驟1474中，此流程會藉由存取快照序列中之下一個資料節點來會繼續執行。下一個資料節點之處理會如上所述地繼續，直到在步驟1472中，判斷最後資料節點即將被處理。

在處理正被回復快照序列之最後資料節點後，在步驟1476中，流程會更新回復時被重新定位至未被配置資料節點之現正開啟檔案系統中的目錄項目，因其原始資料節點已經被現正開啟檔案系統重新配置。這更新涉及資料節點之重新定位，而其資料節點從初次擷取期間所使用之未配置資料節點更新至快照序列被取得時所使用之資料節點。



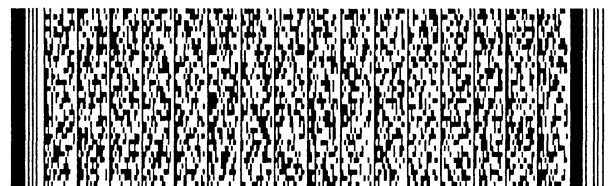
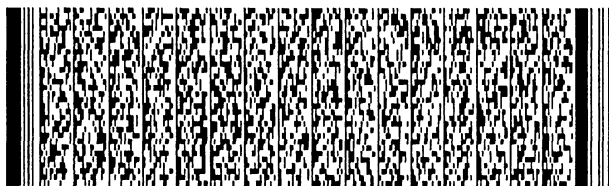
五、發明說明 (66)

在完成上述快照回復處理之後，實施例之檔案系統便恢復一致的連接。在回復期間被重新定位在現正開啟檔案系統上之檔案使得在被回復的目錄中之連結也被回復，以正確地連結現正開啟檔案系統中被重新定位之檔案。這個流程會維持相同快照序列中檔案間的寫入一致性。

本發明有些實施例會操作於分散式計算結構 (distributed computing architecture)。圖16為分散式計算環境1600之結構方塊圖，其包含經由複數個網路轉接器 (adapter) 1604 相互連接之複數個節點1602。每一節點1602是一台獨立的電腦，具有自己的作業系統圖像1608、記憶體1610、以及位於系統記憶體匯流排1618之處理器1606。一系統輸入/輸出匯流排1616係耦接至I/O轉接器1612和網路轉接器1604。每一網路轉接器係經由一網路開關1620相互連接在一起。必須注意的是，網路可以是一個光纖網路 (fiber optic network)，其具有一個相對應的光纖網路開關。

在一個例子中，分散式計算環境1600可以包含N個節點1602，其具有一個或多個處理器1606。在一個例子中，每一處理節點可以執行AIX之RISC/6000電腦，即UNIX作業系統之IBM版本，但處理節點並非需執行AIX作業系統之IBMRISC/6000電腦，有些或所有處理節點1602可以包含不同類型之電腦及/或不同的作業系統1608。所有這些變形可以被視為本發明所主張之一部份。

本發明可以實現於硬體、軟體、或硬體和軟體的組

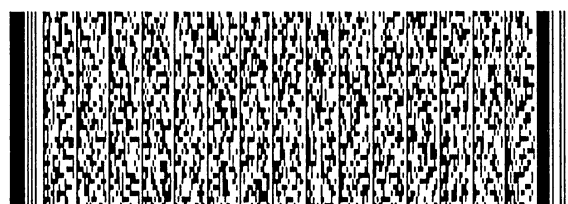
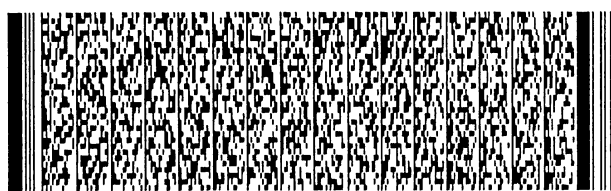


五、發明說明 (67)

合。依據本發明較佳實施例之系統，電腦系統可以集中模式 (centralized fashion) 或者以分散模式來實現，其中分散模式係指不同的元件分散於數個相互連接之電腦系統，任何一種電腦系統或其它適用於執行上述方法之裝置都是適合的。典型硬體和軟體之組合可以是一個一般用途電腦系統，其具有一個電腦程式，可用來在載入和執行時控制電腦系統，故可以執行於此所述之方法。

本發明之實施例也可以置入在一個電腦程式產品中，其包含所有可以執行上述方法之特徵，而當將此電腦程式安裝於電腦系統中時，可以用來執行這些方法。在本說明書中之電腦程式裝置或電腦程式係指一組指令，可以任何語言、代碼或記號來表示，這些指令使得一個系統具有資訊處理的能力，以直接執行一個特定的功能，或者在下列之一個或二個操作執行之後才執行：a) 轉換成另外一種語言、編碼或記號；以及b) 以不同的物質形式再生產。

電腦系統可以包含特別是一個或多個電腦，以及至少一個電腦可讀取媒體，以允許電腦系統讀取資料、指令、訊息或訊息封包，和其他來自電腦可讀取媒體之電腦可讀取資訊。電腦可讀取媒體可以包含非揮發性記憶體，例如ROM、快閃記憶體、磁碟機記憶體、CD-ROM、以及其它永久儲存裝置。此外，一電腦可讀取媒體亦可以包含如揮發性儲存裝置、RAM、緩衝器、快取記憶體、以及網路電路等。此外，電腦可讀取媒體可以包含短暫狀態媒體 (transitory state medium) 中之電腦可讀取資訊，像



五、發明說明 (68)

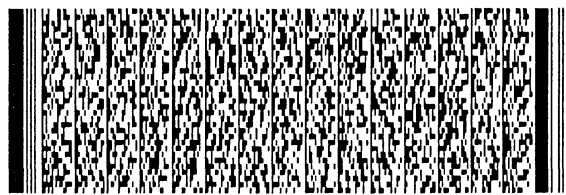
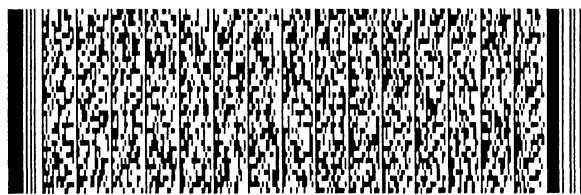
是一網路連結並且/或者一個網路介面，其包含一有線網路或一无線網路，可允許電腦系統讀取電腦可讀取資訊。

圖17為電腦系統1700之一個例子。電腦系統1700包含一個或多個處理器，例如處理器1704。處理器1704係連接至一通訊公用設施(communication infrastructure)

1702，如通訊匯流排、交越條(cross-over bar)、或網路，各種不同的軟體實施例係利用這個示範性電腦系統來描述，在閱讀這些描述之後，相關技術領域之習知技術者可以清楚了解如何利用其它電腦系統並且/或者電腦結構來實施本發明。

電腦系統1700可以包含一顯示介面1708，其係從通訊公用設施1702或從未顯示之框架緩衝器(frame buffer)來發送圖像、文字，和其它資料，以顯示在顯示單元1710上。電腦系統1700也可以包含一主記憶體1706，最好是隨機存取記憶體(RAM)，並且可以包含一第二記憶體1712。

第二記憶體1712可以包含像是硬碟機1714及/或可移除儲存裝置1716，可移除儲存裝置1716代表軟式磁碟機、磁帶機、光碟機等等。可移除儲存裝置1716係利用習知技術者所熟知之方式，從可移除儲存單元1718讀取資料，並且/或者寫入資料至可移除儲存單元1718。可移除儲存單元1718代表軟碟、磁帶、光碟等等，其可以利用可移除儲存裝置1716來讀取或寫入。可以了解的是，可移除儲存單元1718包含一電腦可使用儲存媒體，用來存放電腦軟體及/或資料。



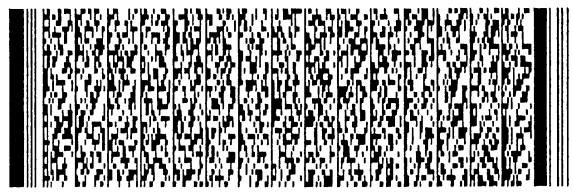
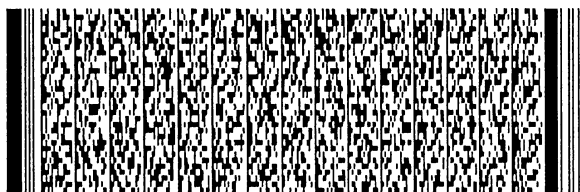
五、發明說明 (69)

在其他可選擇的實施例中，第二記憶體1712可以包含其它相似的裝置，使得電腦程式或其它指令可以載入電腦系統1700中，這樣的裝置可以包含像是可移除儲存單元1722和介面1720。這例子亦可以包含一個程式匣

(program cartridge) 和一匣介面 (cartridge interface) (存在於視訊遊戲裝置中 (video game devices))、一可移除記憶晶片 (例如EPROM或PROM)、相關的插槽、以及其它可移除儲存單元1722和介面1720，可使得軟體和資料從可移除儲存單元1722傳送至電腦系統1700。

電腦系統1700可以包含一通訊介面1724，使得軟體和資料可以在於電腦系統1700和外部裝置間轉移。通訊介面1724之例子可以包含數據機、網路介面 (例如乙太網路卡)、通訊埠、PCMCIA插槽和卡片等等。經由通訊介面1724所移轉之軟體和資料信號形式可以是電子信號、電磁信號、光學信號、或其他可以經由通訊介面1724接收之信號，這些信號可以經由通訊路徑 (即通道) 1726提供至通訊介面1724。通道1726可以運載信號，並且可以利用電線或同軸電纜、光纖、電話線、行動電話連結、RF連結、及/或其它通訊通道來實施。

在說明書中，名詞「電腦程式媒體」、「電腦可使用媒體」、以及「電腦可讀取媒體」一般都用來指稱媒體，例如主記憶體1706和第二記憶體1712、可移除儲存裝置1716、安裝在硬碟機1714之硬碟、以及信號。這些電腦程

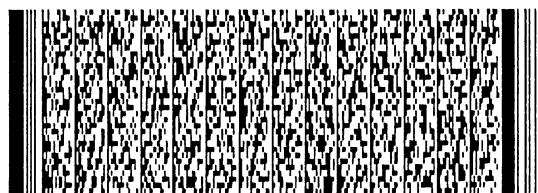
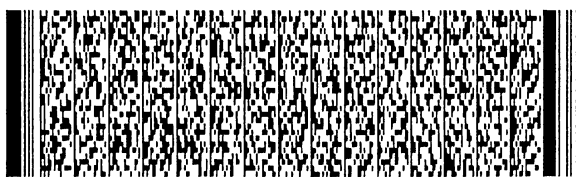


五、發明說明 (70)

式產品係用來提供軟體至電腦系統1700之裝置。電腦可讀取媒體使得電腦系統1700可以讀取資料、指令、訊息或訊息封包，和其它來自電腦可讀取媒體之電腦可讀取資訊。例如，電腦可讀取媒體可以包含非揮發性記憶體、軟磁碟、ROM、快閃記憶體、磁碟機記憶體、CD-ROM、以及其他永久儲存裝置。它可以用來將資訊、資料和電腦指令傳遞於電腦系統1700間。此外，電腦可讀取媒體可以包含短暫狀態媒體 (transitory state medium) 中之電腦可讀取資訊，例如一網路連結並且/或者一個網路介面，其包含一有線網路或一無線網路，以允許電腦系統可以讀取電腦可讀取資訊。

電腦程式(也稱為電腦控制邏輯)係存放在主記憶體1706及/或第二記憶體1712，可以經由通訊介面1724來接收。當執行這樣的電腦程式時，可以使得電腦系統1700執行上述之本發明特徵。特別是當執行電腦程式時，可以使得處理器1704執行電腦系統1700之特徵。因此，這樣的電腦程式代表電腦系統1700之控制器。

雖然本發明已參照較佳具體例及舉例性附圖敘述，惟其不應被認為限制者。熟悉此技藝者在不離開本發明之範圍內，當可對其形態及特殊具體例之內容作各種修改、省略及變化。



圖式簡單說明

五、【圖式簡單說明】

在說明書的結論部分，本發明之主題會在申請專利範圍中被特別指出及分別主張。參照實施方式的敘述及所附圖式，便可適當地了解本發明之特徵與優點。此外，標號最左邊的數字係表示該標號首次出現的圖式。

圖1為本發明實施例之整體系統結構方塊圖；

圖2A為本發明實施例之資料節點與相關資料區塊之方塊圖；

圖2B為本發明實施例之資料節點與相關間接區塊與資料區塊之方塊圖；

圖3為本發明實施例之目錄與相關結構方塊圖；

圖4為具有相同磁碟位址之一個以上資料節點之方塊圖；

圖5為本發明實施例之整體操作和控制流程之流程圖；

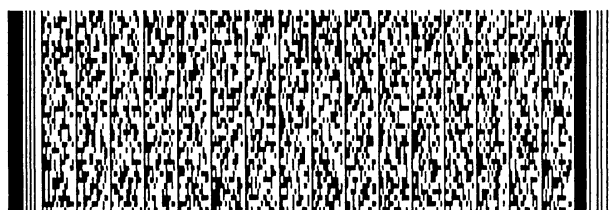
圖6A為本發明實施例之快照產生程序的操作和控制流程之流程圖；

圖6B為本發明實施例之多個快照產生順序之時間軸線；

圖7A為本發明實施例之快照更新程序操作和控制流程之流程圖；

圖7B為本發明實施例在有數個快照可以利用時之快照更新程序操作和控制流程之流程圖；

圖8A為本發明實施例在快照產生時之結構方塊圖；



圖式簡單說明

圖8B為本發明實施例之映像資料節點在修改其對應來源檔案屬性時之結構方塊圖；

圖8C為本發明實施例之映像資料節點在覆寫其對應來源檔案時之結構方塊圖；

圖8D為本發明實施例之映像資料節點在覆寫其對應具有相關間接區塊和資料區塊之來源檔案時的結構方塊圖；

圖9為本發明實施例之快照擷取程序之操作和控制流程之流程圖；

圖10為本發明實施例在有數個快照可以利用時快照擷取程序操作和控制流程之流程圖；

圖11為本發明實施例在有數個快照可以利用時之快照擷取方塊圖；

圖12A為本發明實施例在有數個快照可以利用時之最近映像資料節點動態連結方塊圖；

圖12B為本發明實施例使用動態連結之快照擷取程序之操作和控制流程之流程圖；

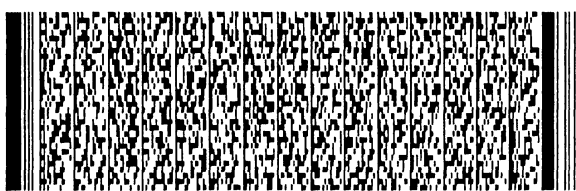
圖13A為寫入時複製程序之操作和控制流程之流程圖；

圖13B為本發明實施例在寫入時移動程序之操作和控制流程之流程圖；

圖13C為本發明實施例在寫入時移動發生後之讀取程序的操作和控制流程之流程圖；

圖14A為本發明實施例之數個快照序列之維恩圖；

圖14B為本發明實施例與不同快照序列相關之快照資



圖式簡單說明

料集合方塊圖；

圖14C為本發明實施例在有數個快照序列可以利用時之回復程序的操作和控制流程之流程圖；

圖15A為本發明實施例中具有待命檔案系統之主要檔案系統的整體系統結構方塊圖；

圖15B為本發明實施例之待命檔案系統更新程序之操作和控制流程之流程圖；

圖15C為本發明實施例之待命檔案系統激化程序之操作和控制流程之第一片段流程圖；

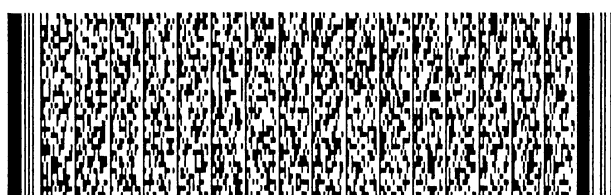
圖15D為本發明實施例之待命檔案系統激化程序之操作和控制流程之第二片段流程圖；

圖16為用來實施本發明示範性多處理器電腦系統之方塊圖；

圖17為用來實施本發明之示範性電腦系統之方塊圖。

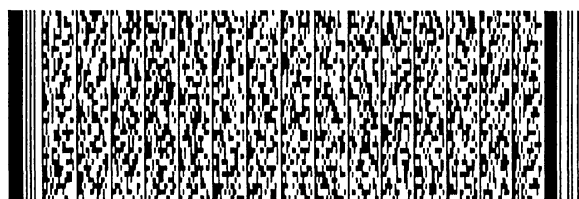
圖式元件符號說明

100	整體系統結構	102	檔案系統
104	快照資料集合	106	網路
108	備份系統	200	資料結構集合
202	資料節點	204	超資料
206	磁碟位址資料	208	資料區塊
250	資料結構集合	252	資料節點
254	超資料	256	磁碟位址資料
258	間接區塊	260	第一磁碟位址



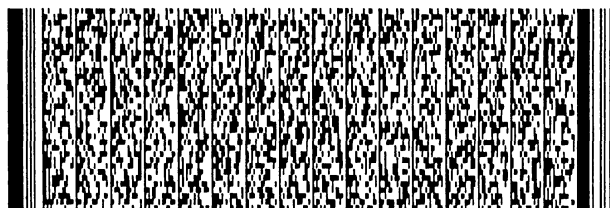
圖式簡單說明

262	第二磁碟位址	264	第一資料區塊
266	第二資料區塊	300	目錄結構
302	目錄	304	第一目錄記錄
306	檔案名稱	308	資料節點號碼
310	第二目錄記錄	312	檔案名稱
314	資料節點號碼	316	資料節點
318	資料節點	400	檔案系統資料結構
402	資料節點	404	磁碟位址
406	資料區塊	408	快照資料節點
410	磁碟位址		
800	起始快照資料集合結構		
802	快照資料集合	804	資料節點
806	超資料	808	磁碟位址
810	資料區塊	812	明示指標
820	更新屬性資料結構	822	映像節點
824	超資料	826	磁碟位址
828	明示指標		
830	更新資料區塊儲存資料結構		
840	間接區塊儲存資料結構		
842	超資料	844	磁碟位址
846	磁碟位址	848	資料區塊
850	資料區塊	1100	示範性快照系列
1102、1106、1112 快照			
1104、1108、1114 映像資料節點			



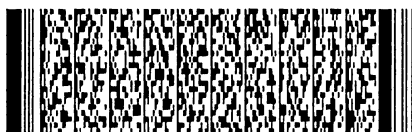
圖式簡單說明

1110	相同磁碟位址	1116	磁碟位址
1118	資料區塊		
1202、1212、1222	快照		
1204、1214、1224	映像資料節點		
1206、1216、1226	相同磁碟位址		
1228	資料節點	1230	磁碟位址
1232	資料區塊	1240	動態鏈結
1404、1406、1408、1410	快照序列		
1422	快照資料集合	1430	快照序列
1432、1438	資料節點	1444	快照序列
1446	資料節點	1452	鏈結
1502	主要檔案系統	1504	備用檔案系統
1506、1508、1510、1512	快照		
1600	分散式計算環境		
1602	資料節點	1604	網路轉接器
1606	處理器	1608	作業系統圖像
1610	記憶體	1612	I/O 轉接器
1614	通道控制器		
1616	系統輸入/輸出匯流排		
1618	系統記憶體匯流排		
1620	網路開關	1700	電腦系統
1702	通訊公用設施	1704	處理器
1706	主記憶體	1708	顯示介面
1710	顯示裝置	1712	第二記憶體



圖式簡單說明

1714	硬碟機	1716	可移除儲存裝置
1718	可移除儲存單元	1720	介面
1722	可移除儲存單元	1724	通訊介面
1726	通訊路徑		



四、中文發明摘要 (發明名稱：提供檔案系統子集之快照)

一種系統、方法以及電腦可讀取媒體，供提供一檔案系統子集之快照。在一檔案系統中產生一組原始檔之第一快照。此第一快照包含一資料節點(inode)，對應在此第一組檔案之每一原始檔。儲存在每一資料節點的是有關第一組檔案之一第一識別符及有關第一快照時間之一第二識別符。接著，取得第二組原始檔之第二快照，此第二快照包含一資料節點，對應在此第二組檔案之每一原始檔。儲存在每一資料節點的是第一識別符及一第二識別符。為第一組檔案及第二組檔案分別每隔第一期間及第二期間取得隨後之快照。

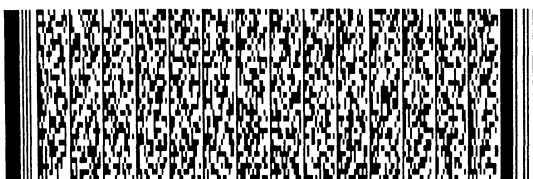
伍、(一)、本案代表圖為：圖1

(二)、本案代表圖之元件代表符號簡單說明：

100 整體系統結構	102 檔案系統
104 快照資料集合	106 網路
108 備份系統	

六、英文發明摘要 (發明名稱：PROVIDING A SNAPSHOT OF A SUBSET OF A FILE SYSTEM)

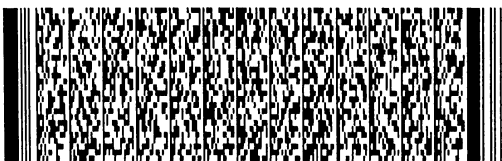
A system, method and computer readable medium for providing a snapshot of subset of a file system. A first snapshot of a first set of source files in a file system is generated. The first snapshot includes an inode corresponding to each source file in the first set of files. Stored in each inode is a first identifier associated with the first set of files and a second identifier



四、中文發明摘要 (發明名稱：提供檔案系統子集之快照)

六、英文發明摘要 (發明名稱：PROVIDING A SNAPSHOT OF A SUBSET OF A FILE SYSTEM)

associated with the time of the first snapshot. Next, a second snapshot of a second set of source files is taken. The second snapshot includes an inode corresponding to each source file in the second set of files. Stored in each inode are a first identifier and a second identifier. Subsequent snapshots are taken every first period and every second period for the first set of files

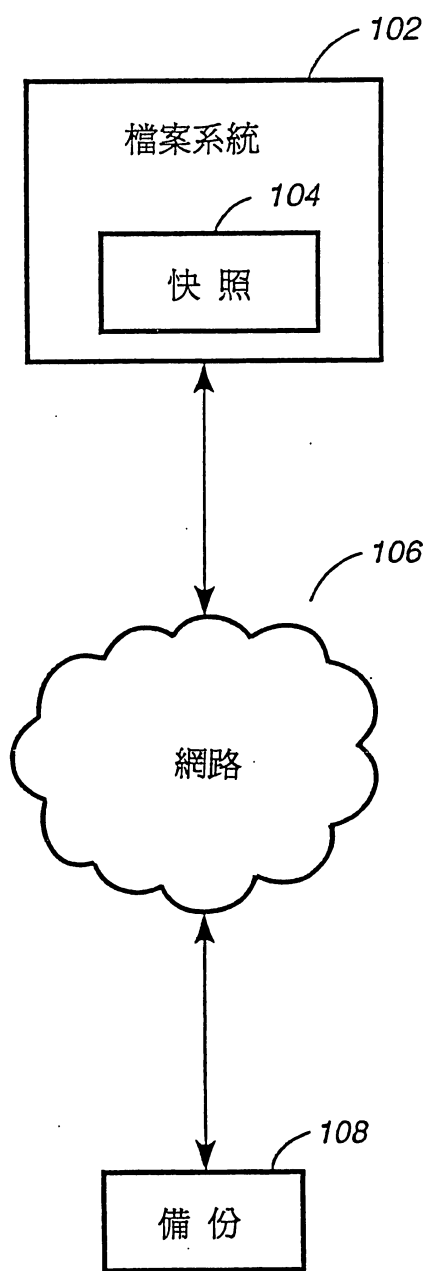


四、中文發明摘要 (發明名稱：提供檔案系統子集之快照)

六、英文發明摘要 (發明名稱：PROVIDING A SNAPSHOT OF A SUBSET OF A FILE SYSTEM)

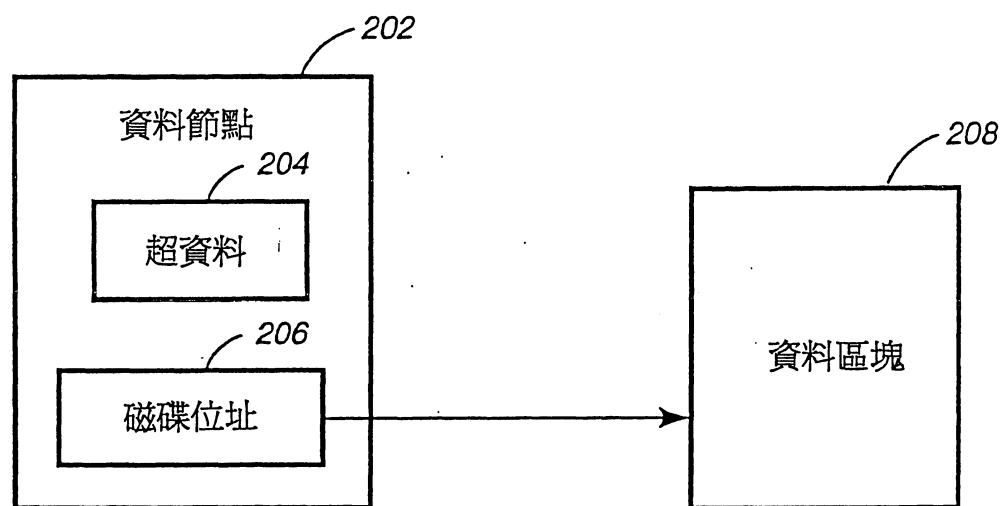
and the second set of files, respectively.





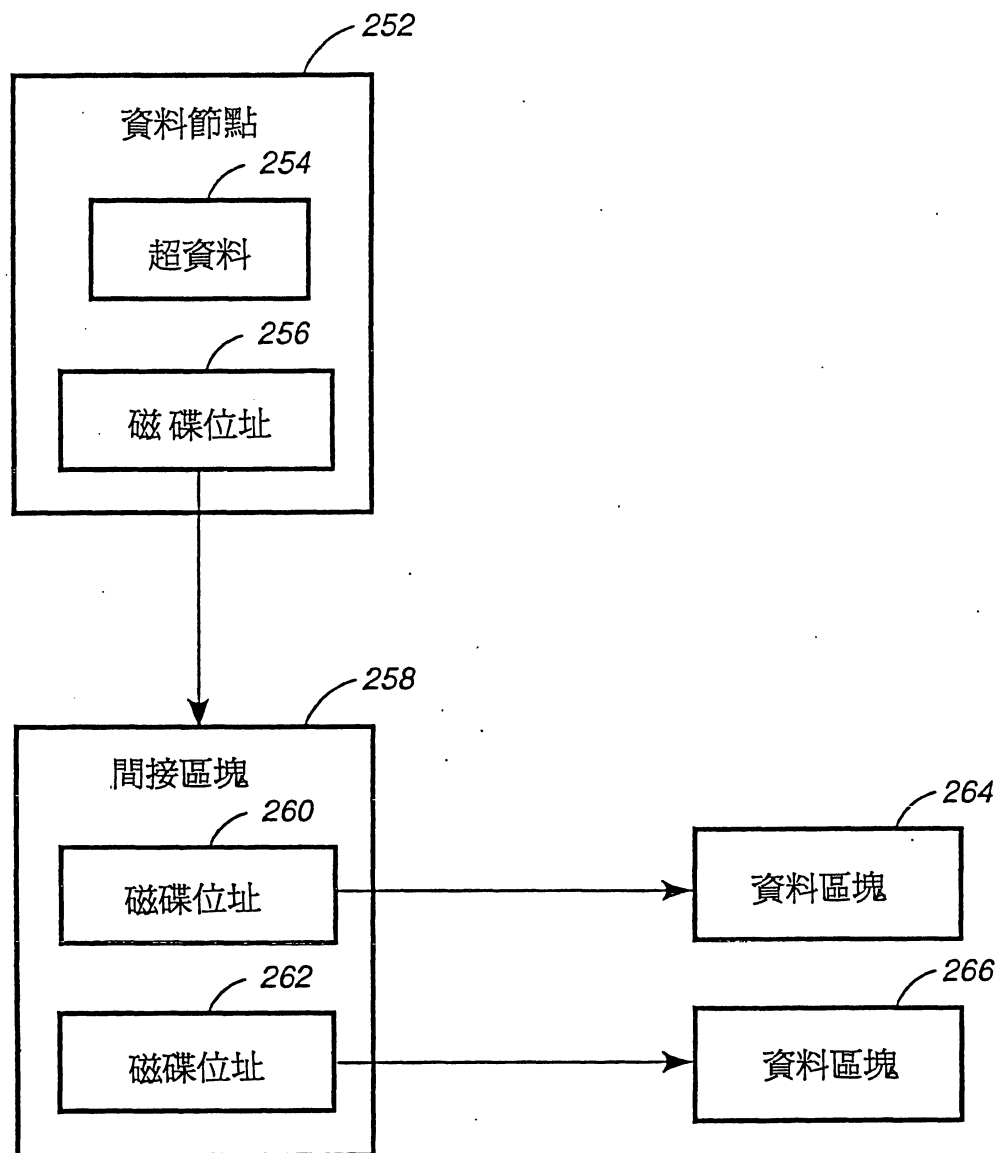
100

圖 1



200

圖 2A



250

圖 2B

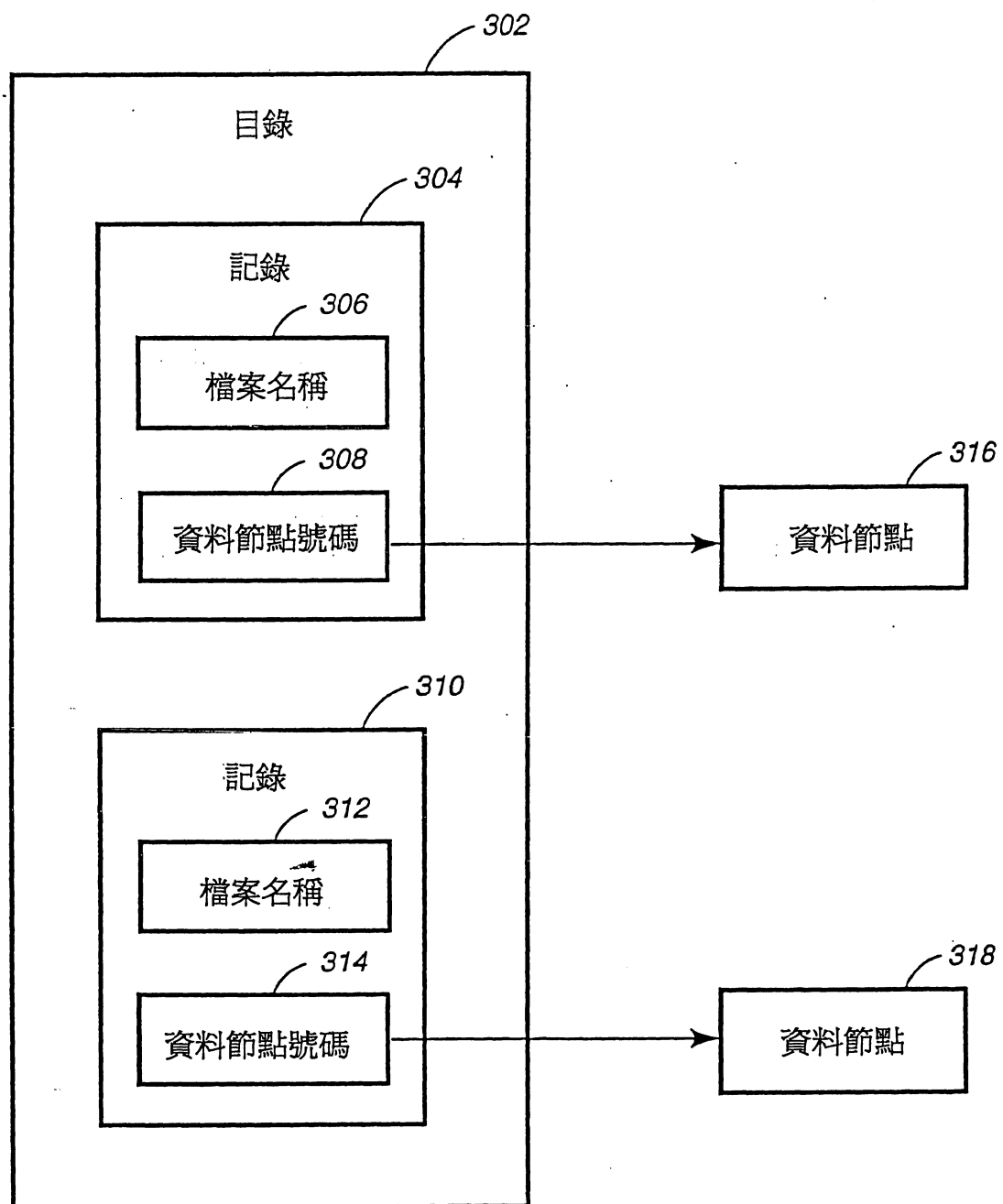
300

圖 3

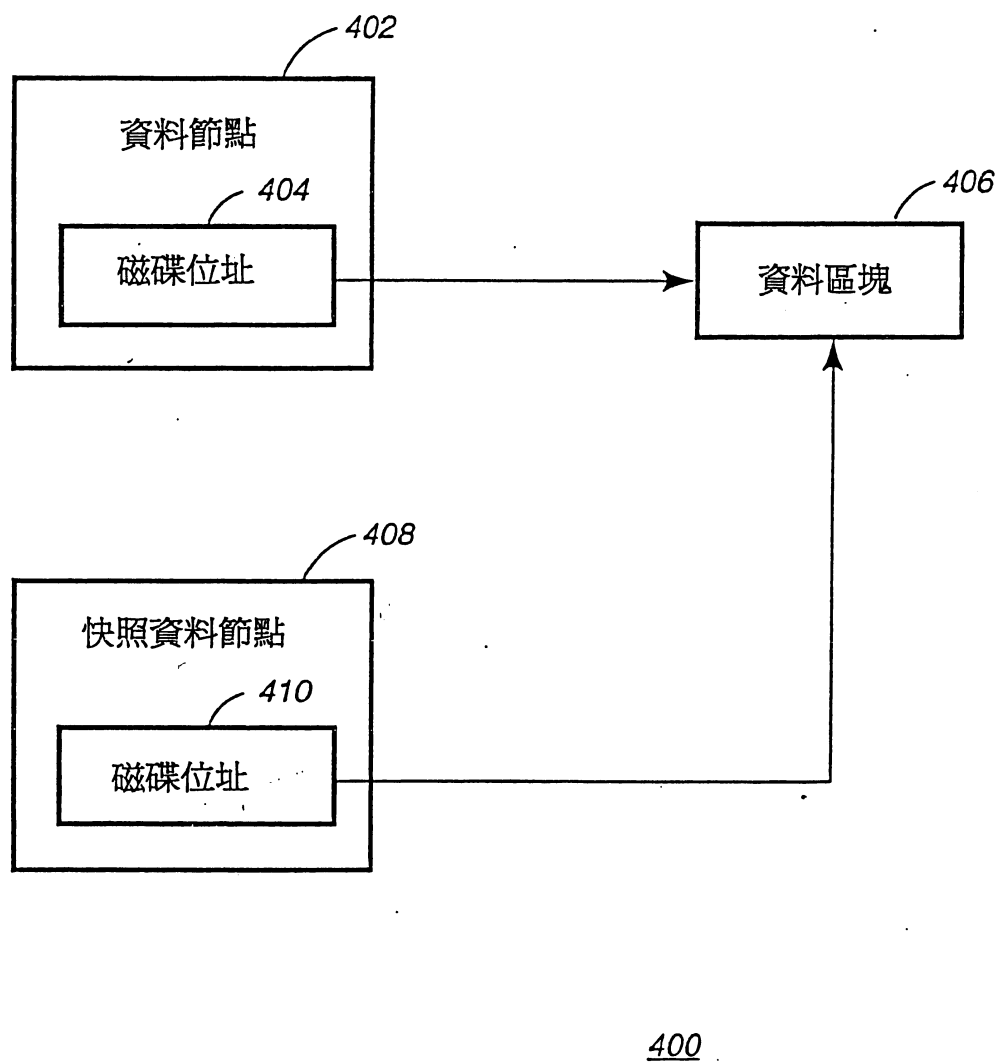


圖 4

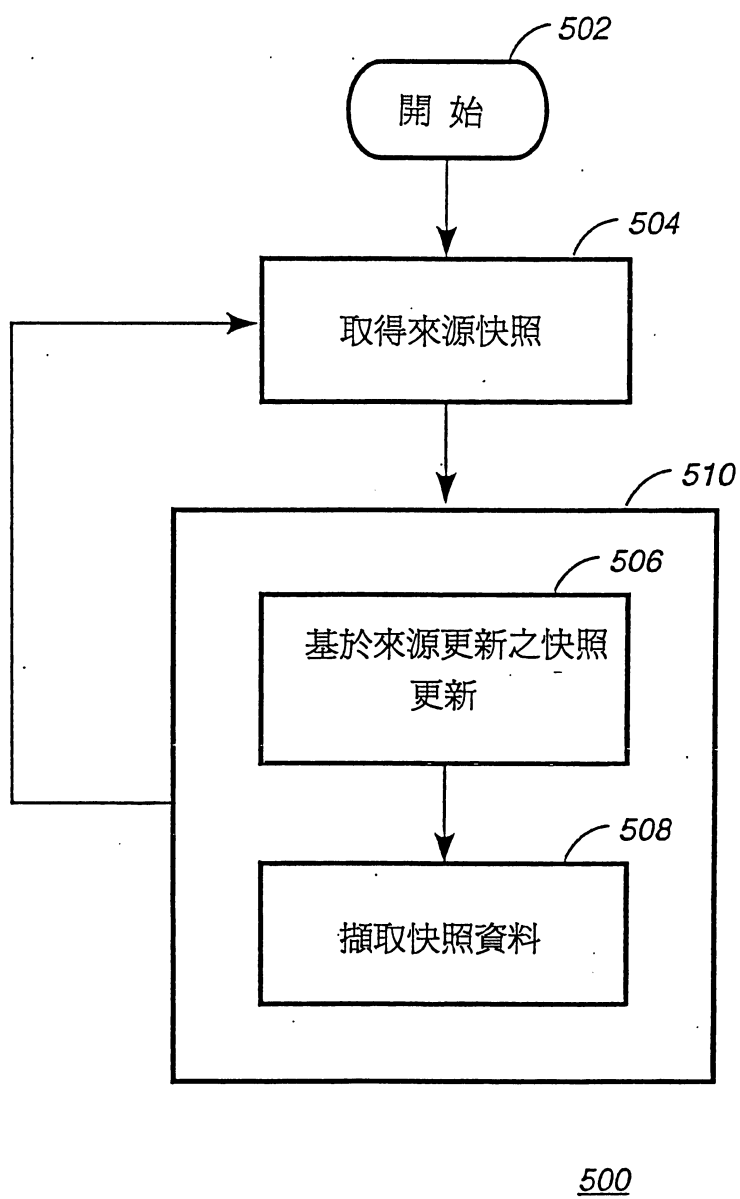
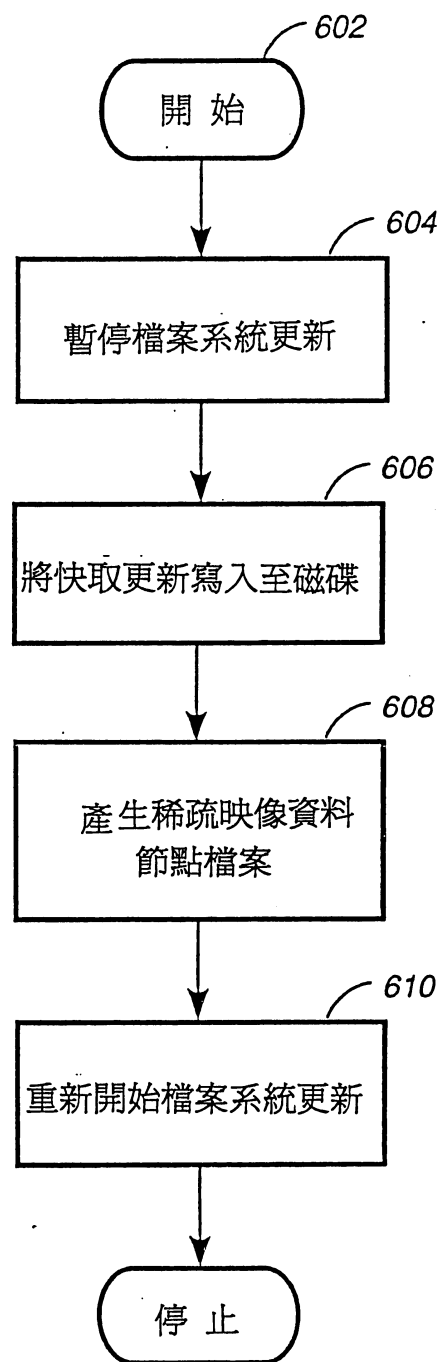
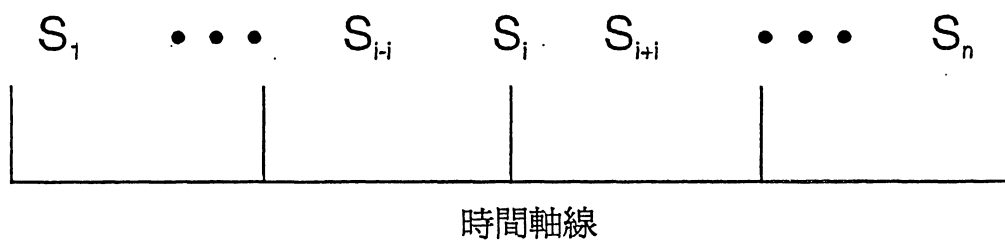


圖 5



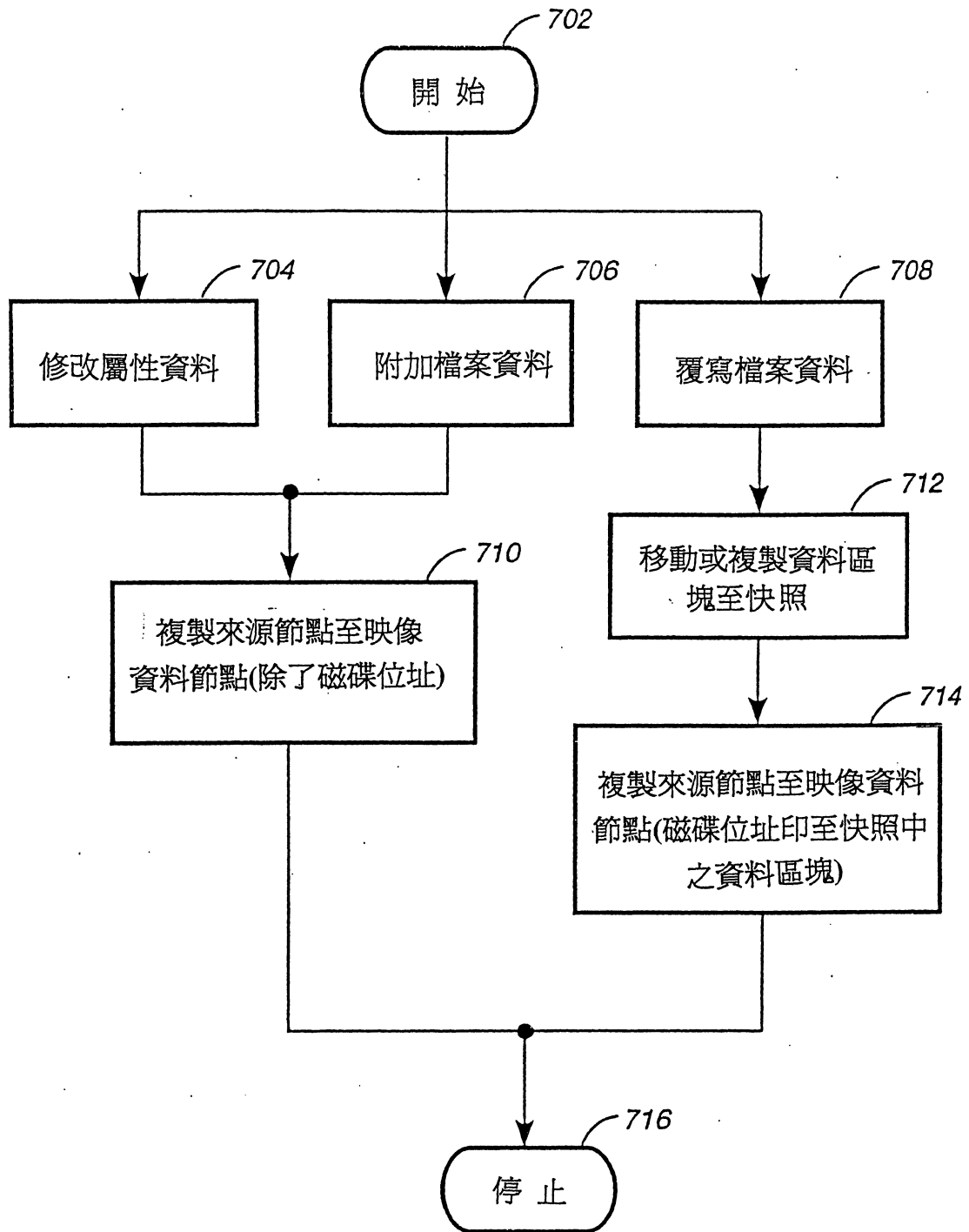
600

圖 6A



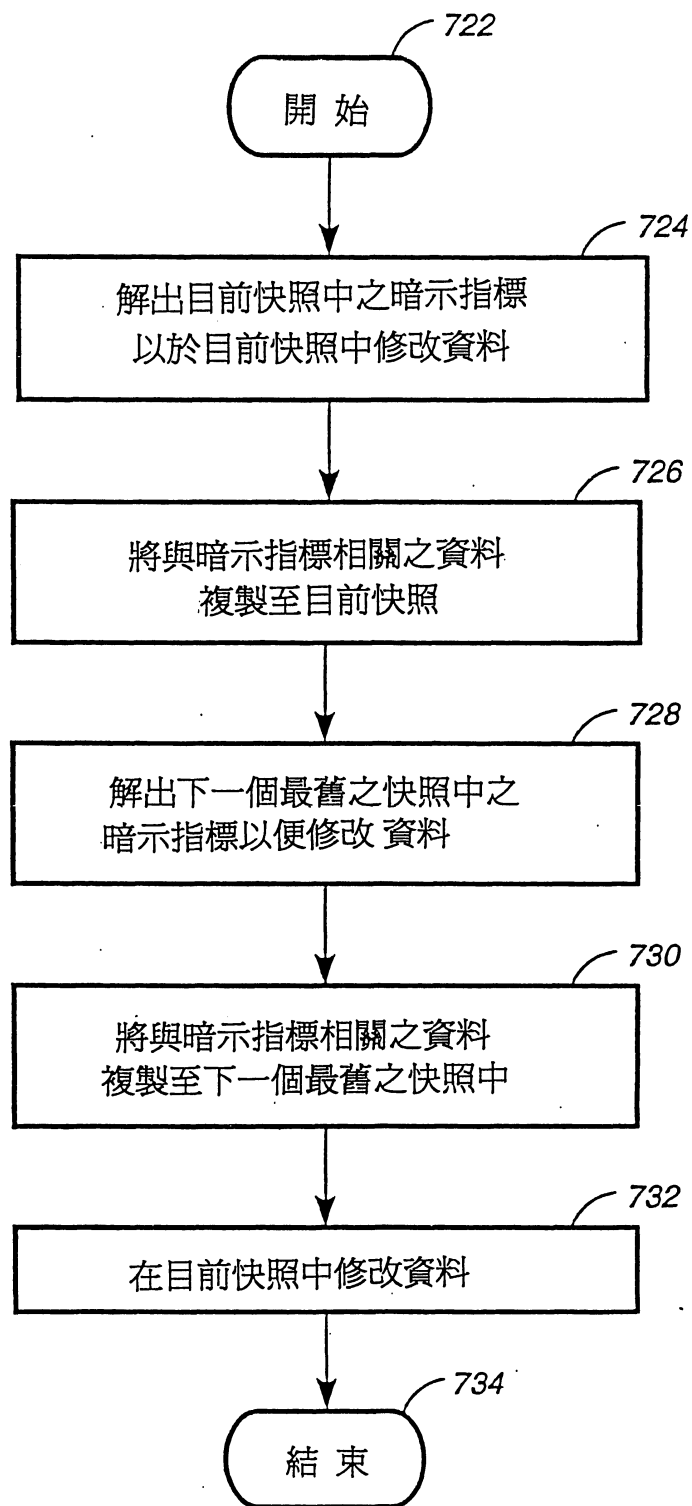
620

圖 6B



700

圖 7A



720

圖 7B

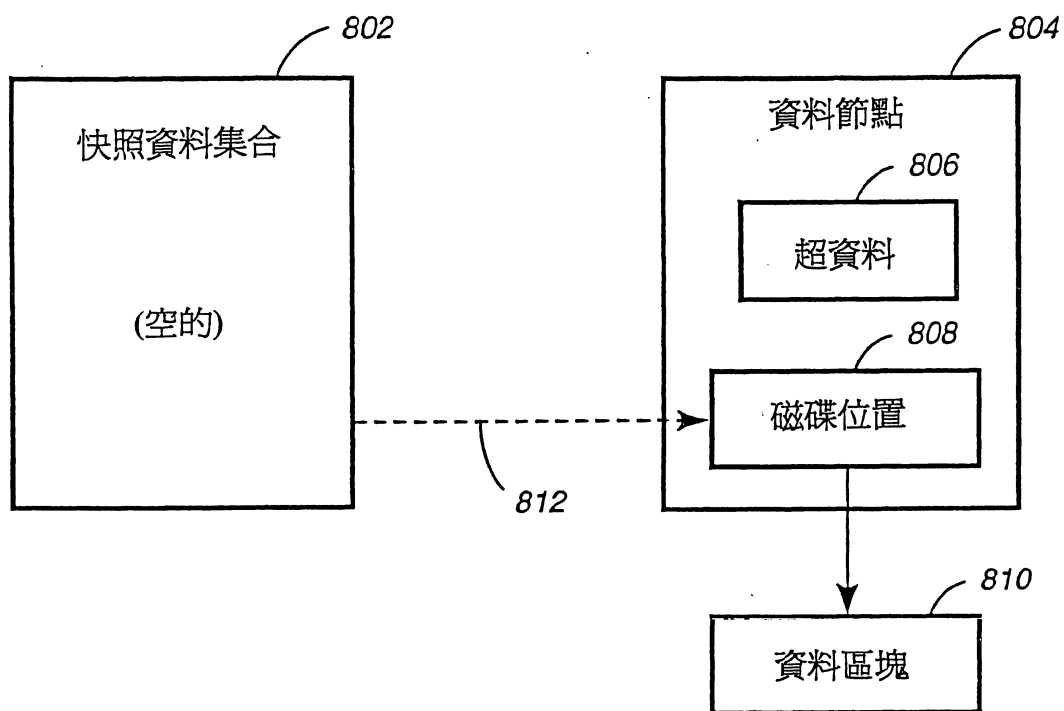
800

圖 8A

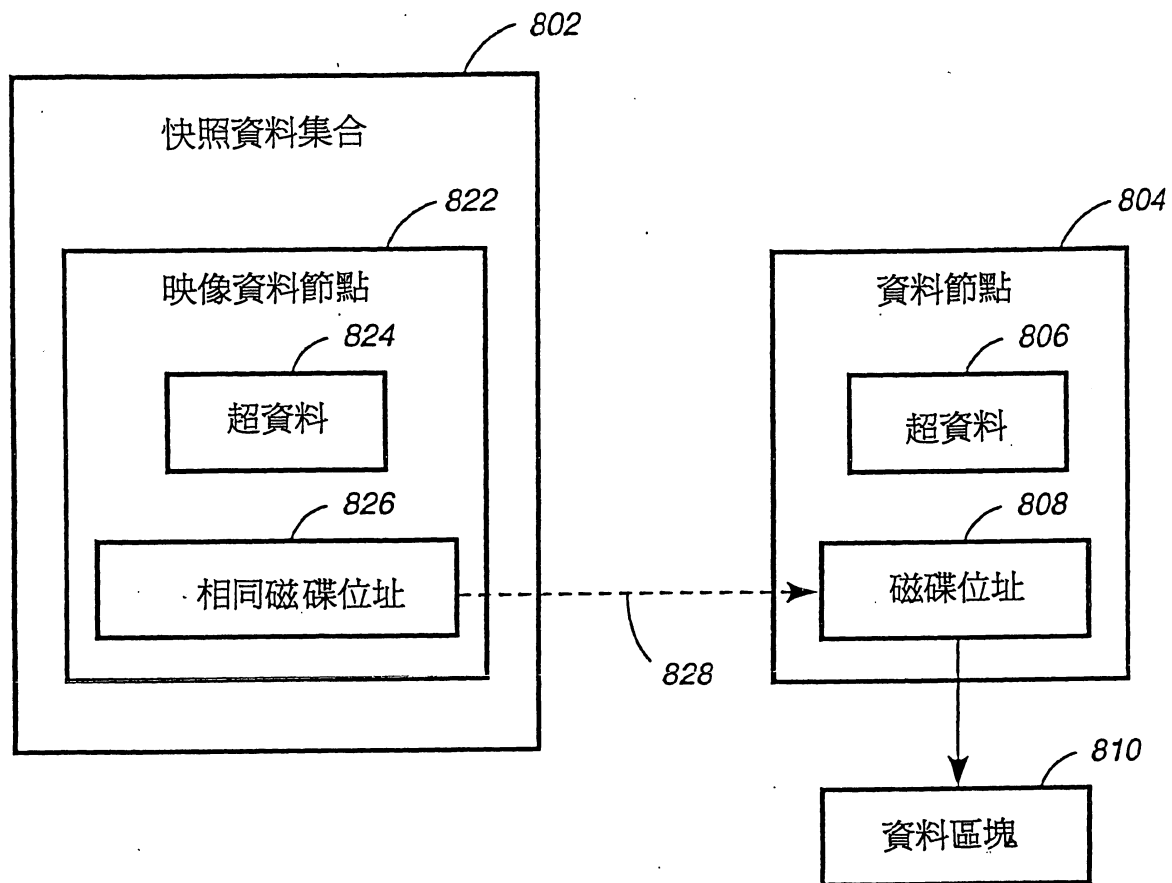
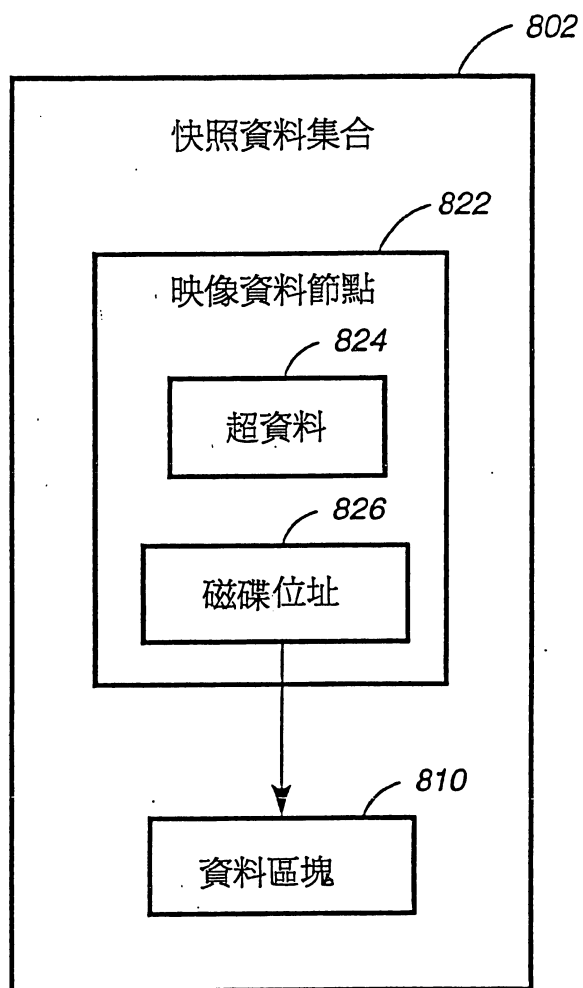
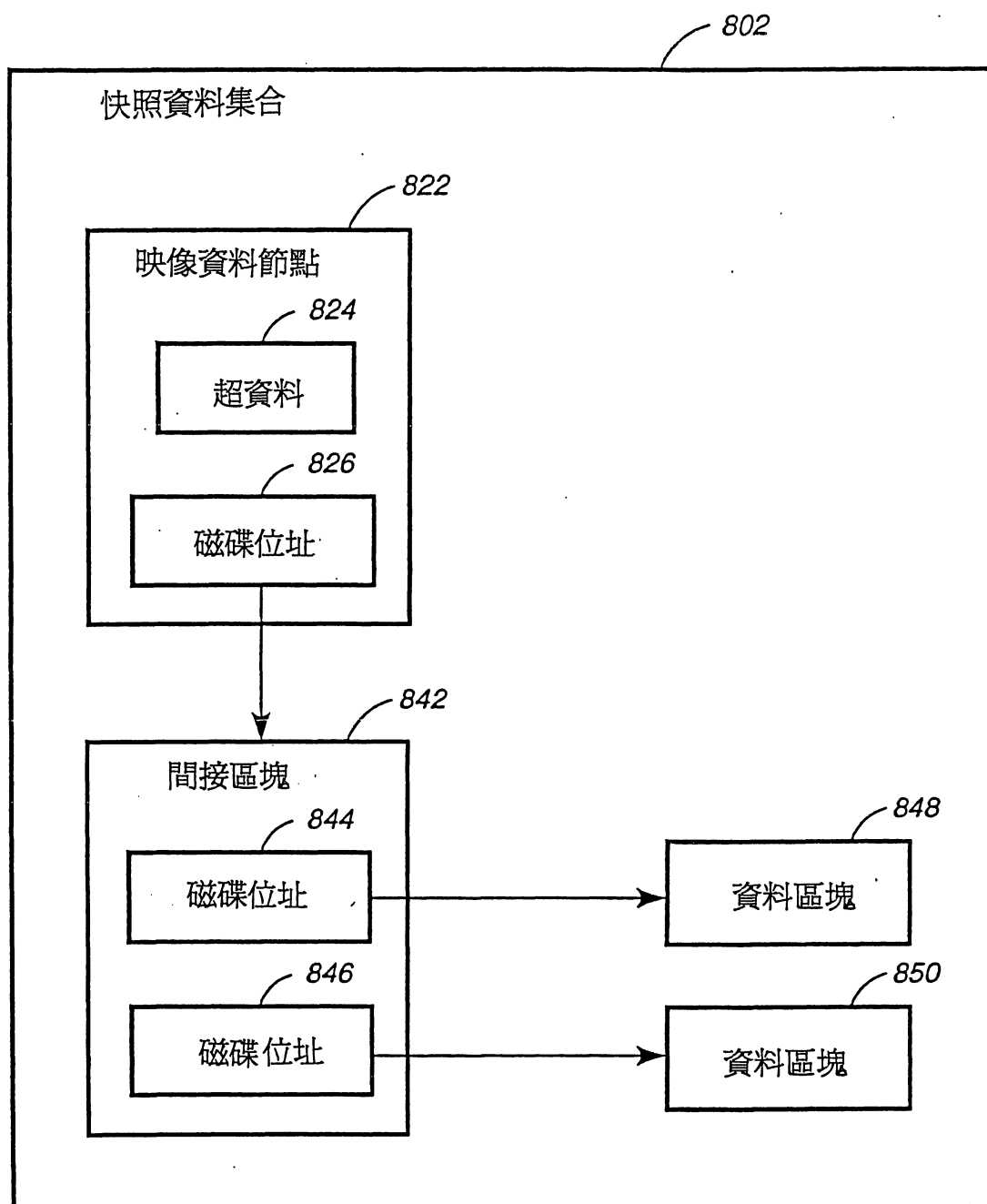
820

圖 8B



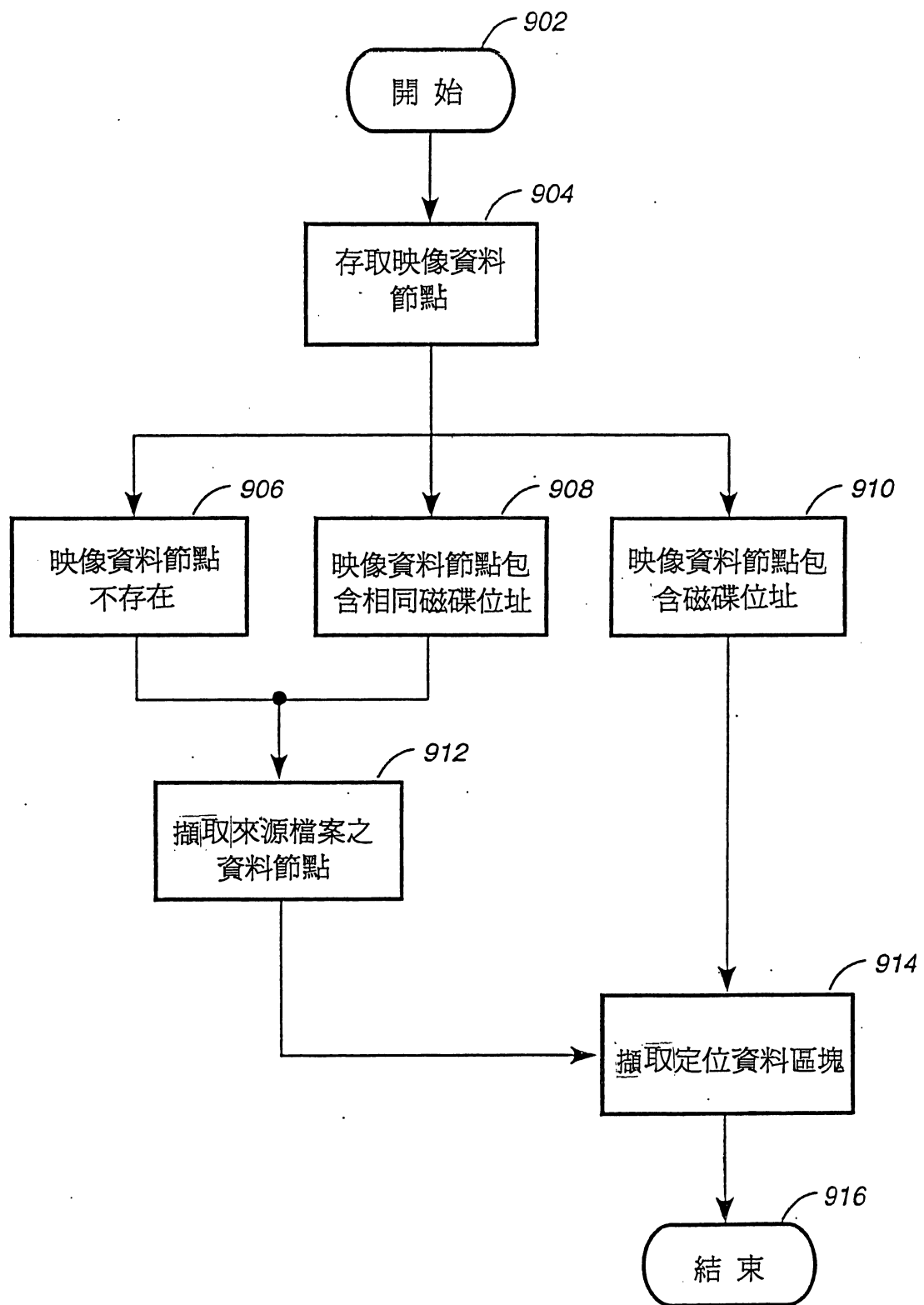
830

圖 8C

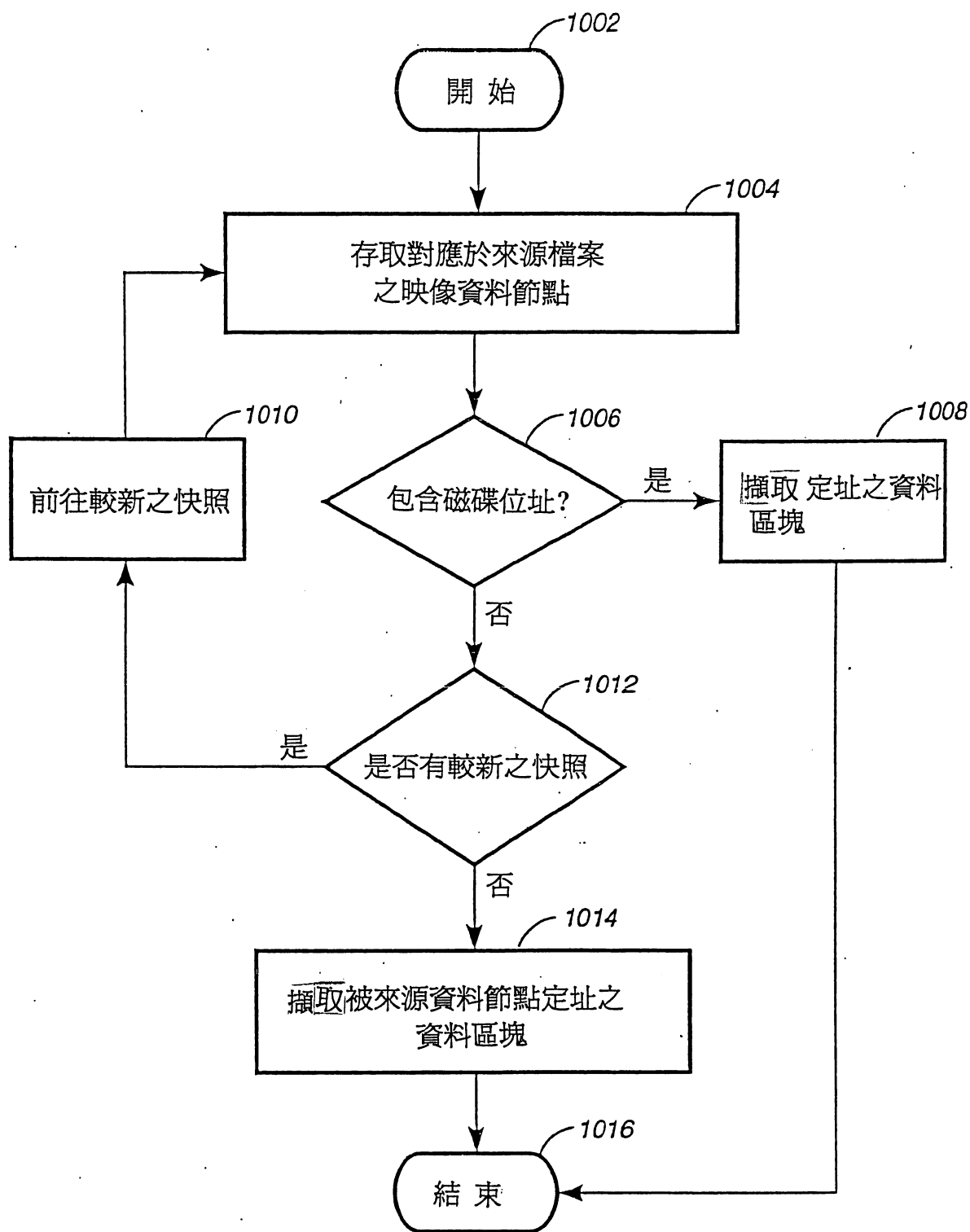


840

圖 8D



900
圖 9



1000

圖 10

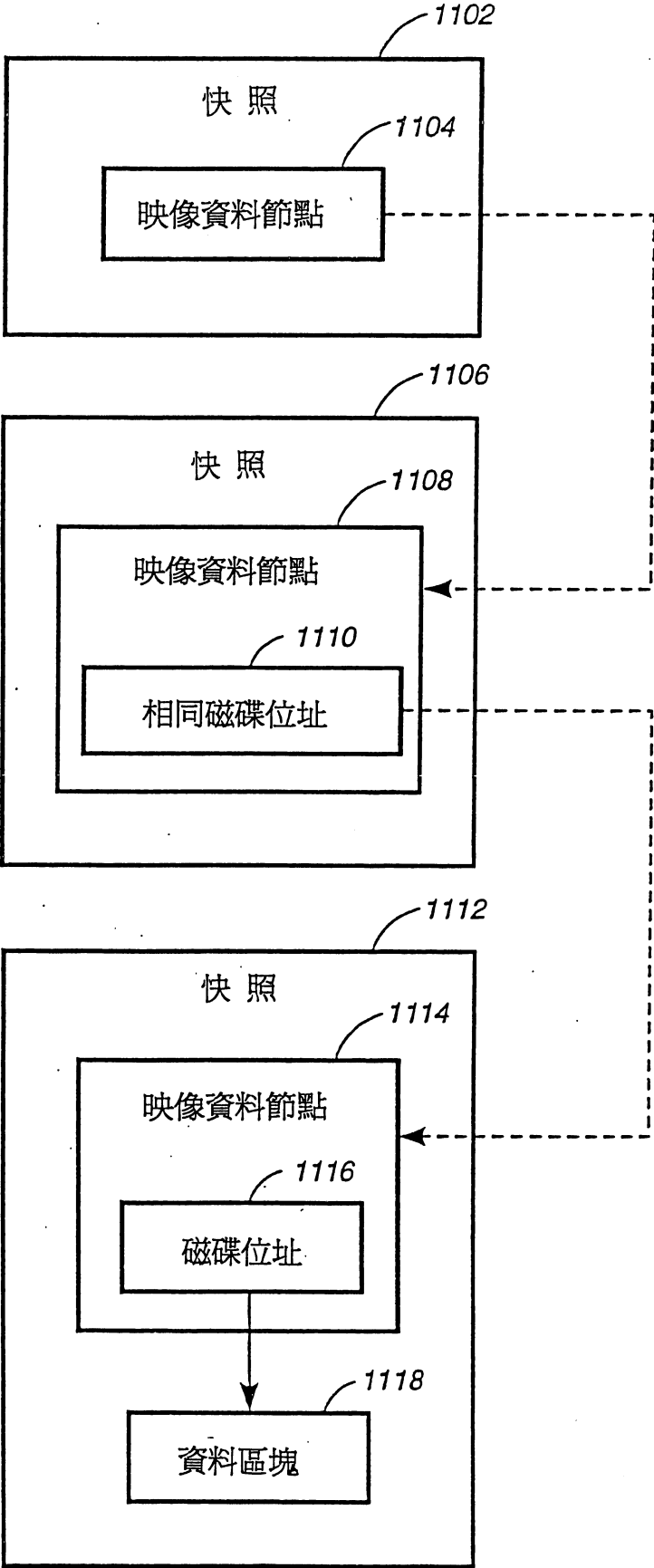


圖 11

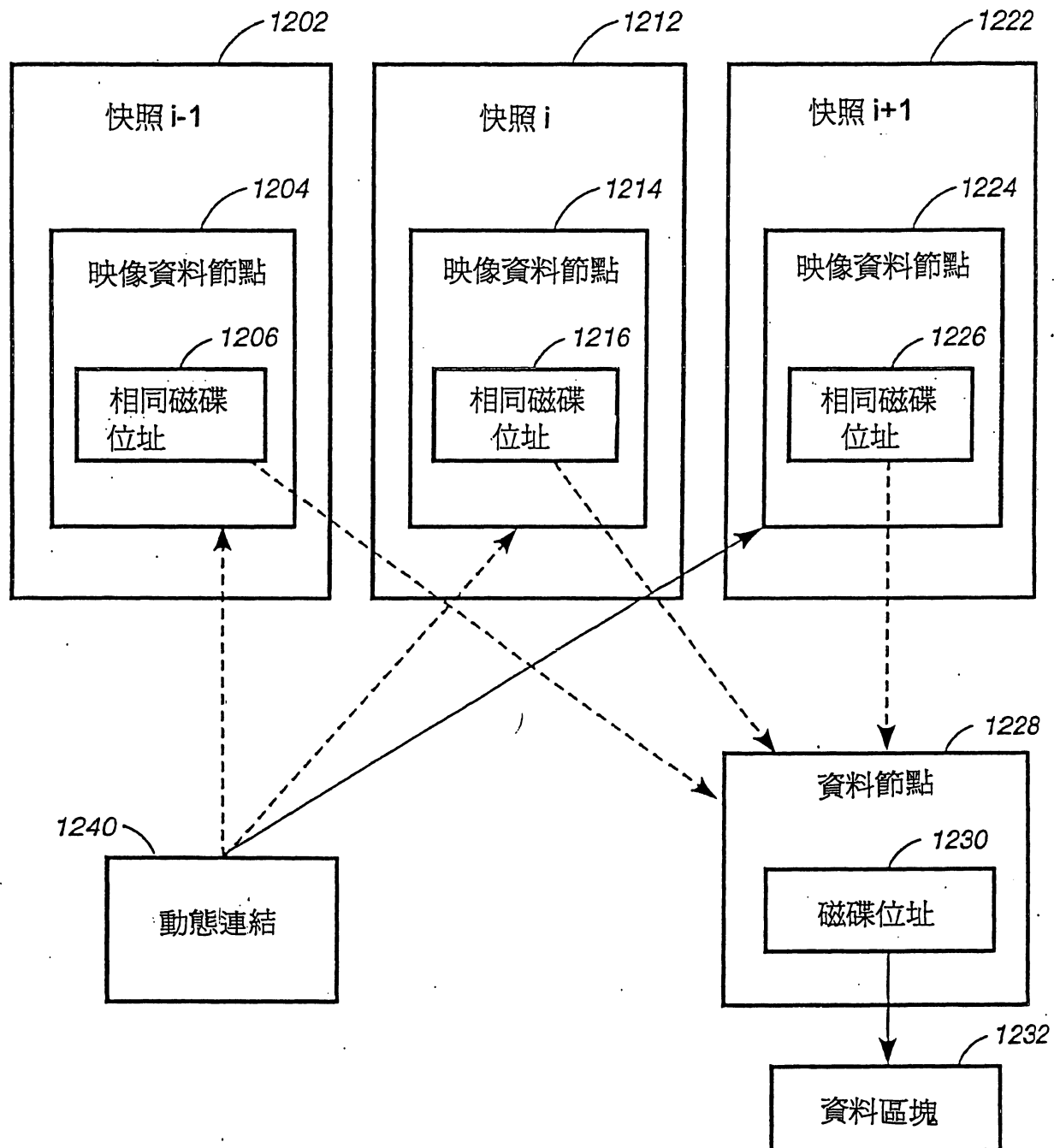


圖 12A

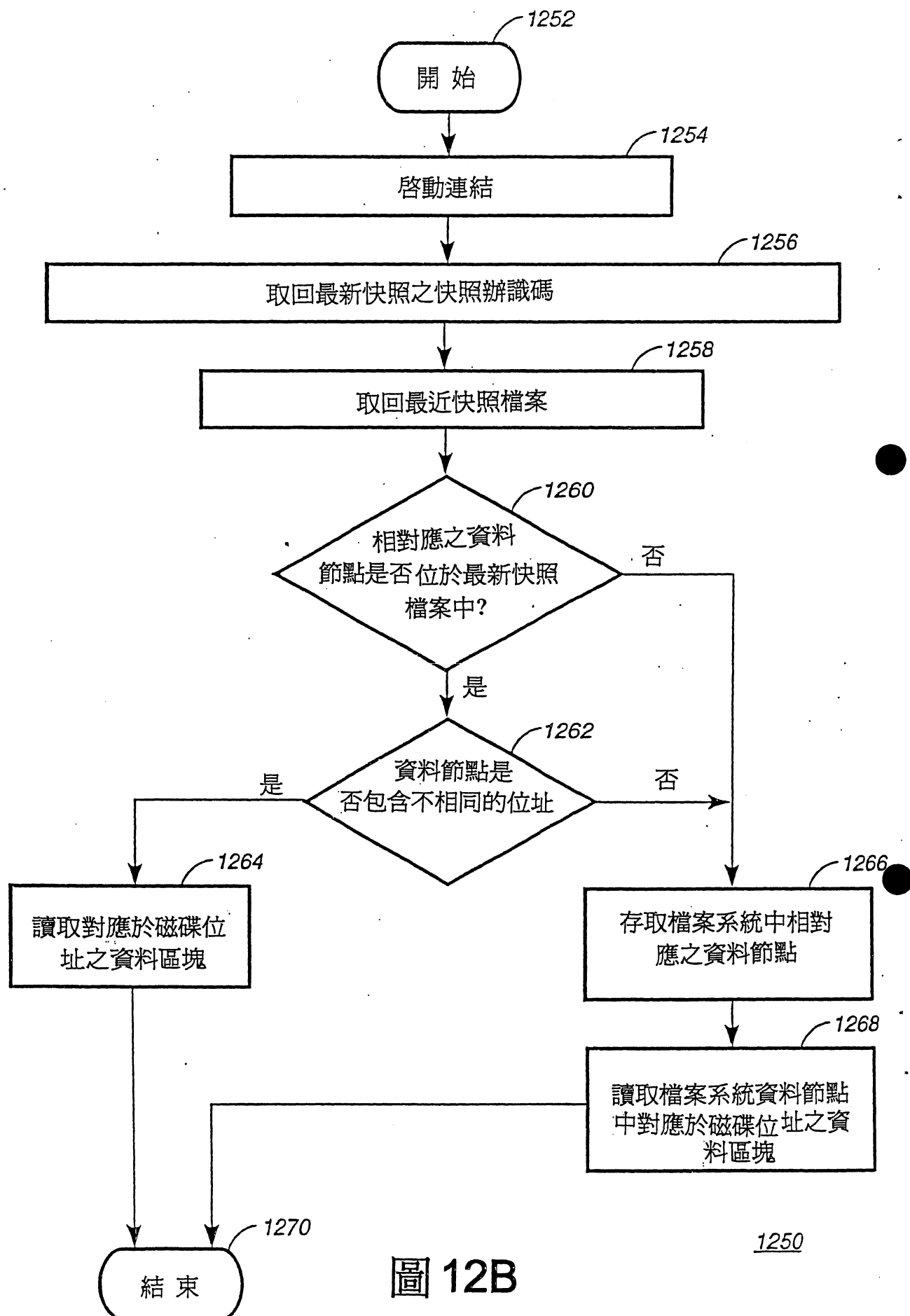
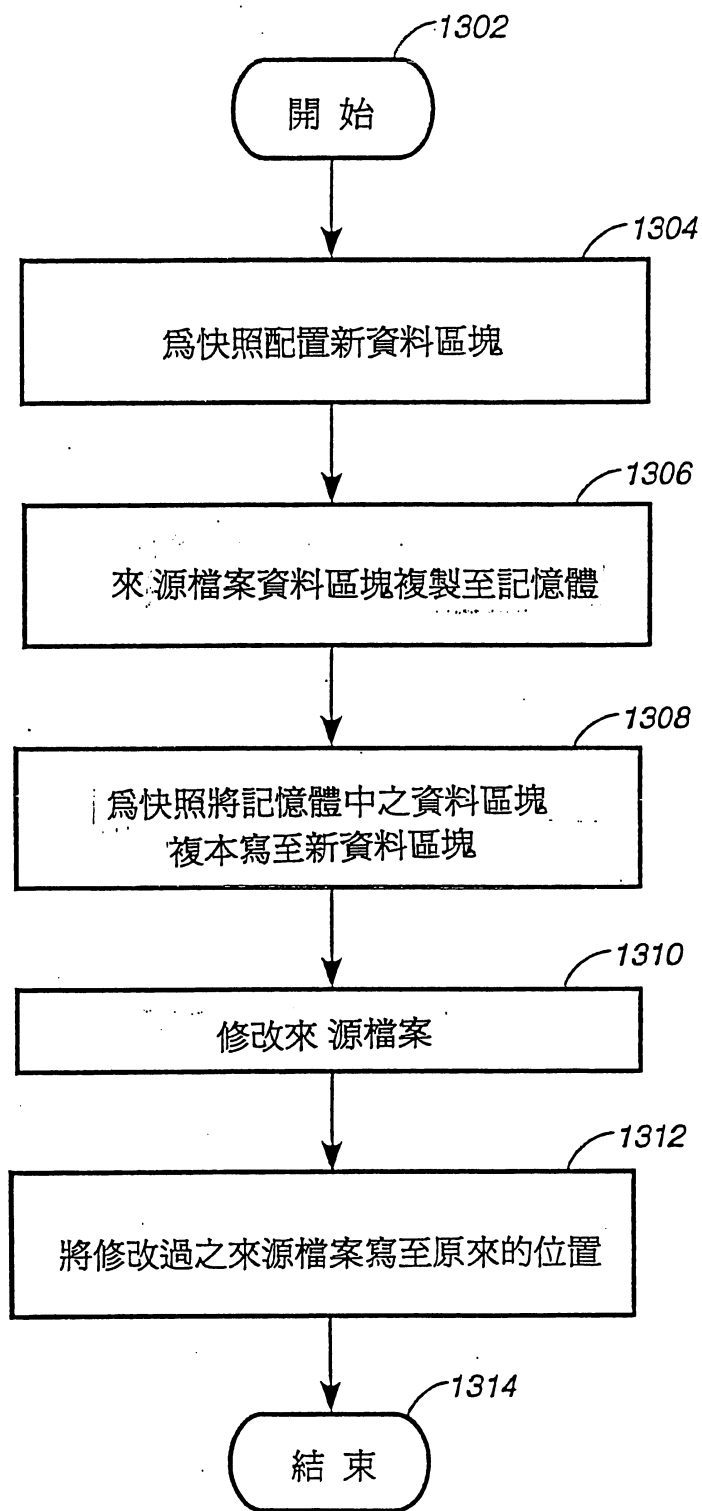
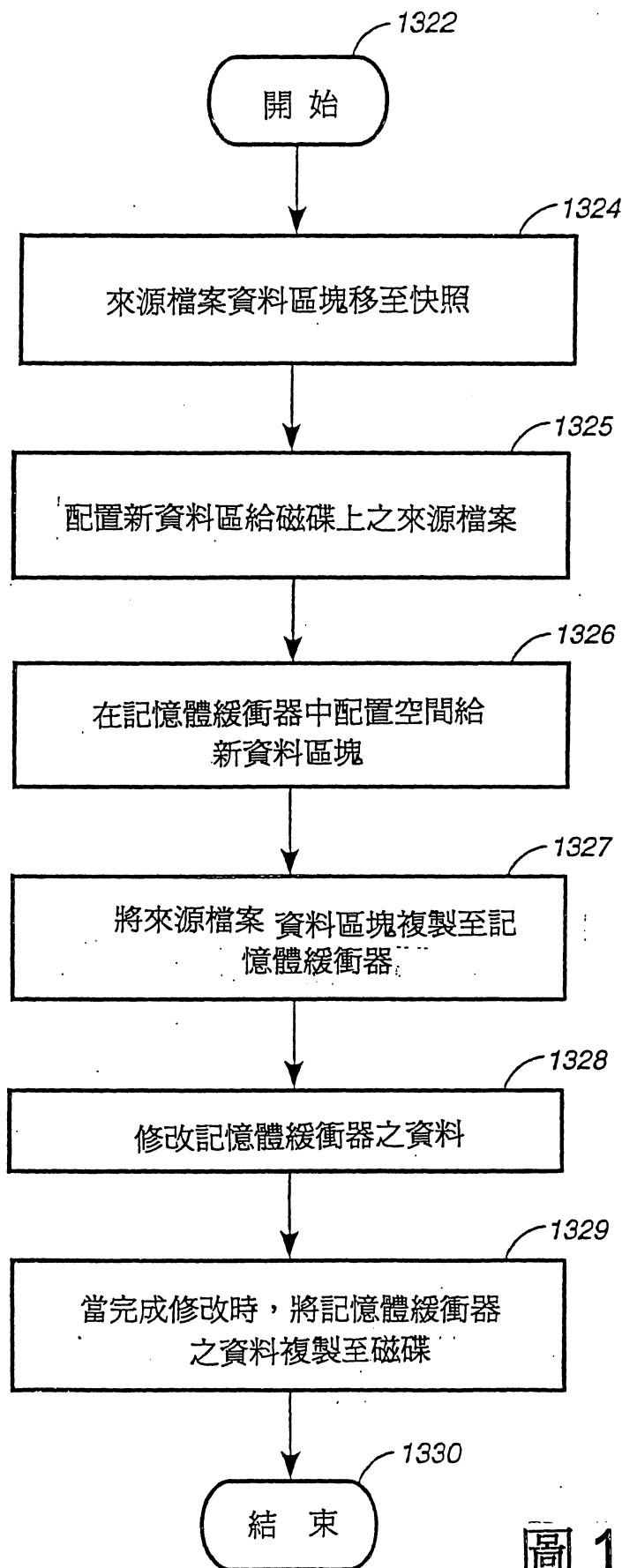


圖 12B



1300

圖 13A



1320

圖 13B

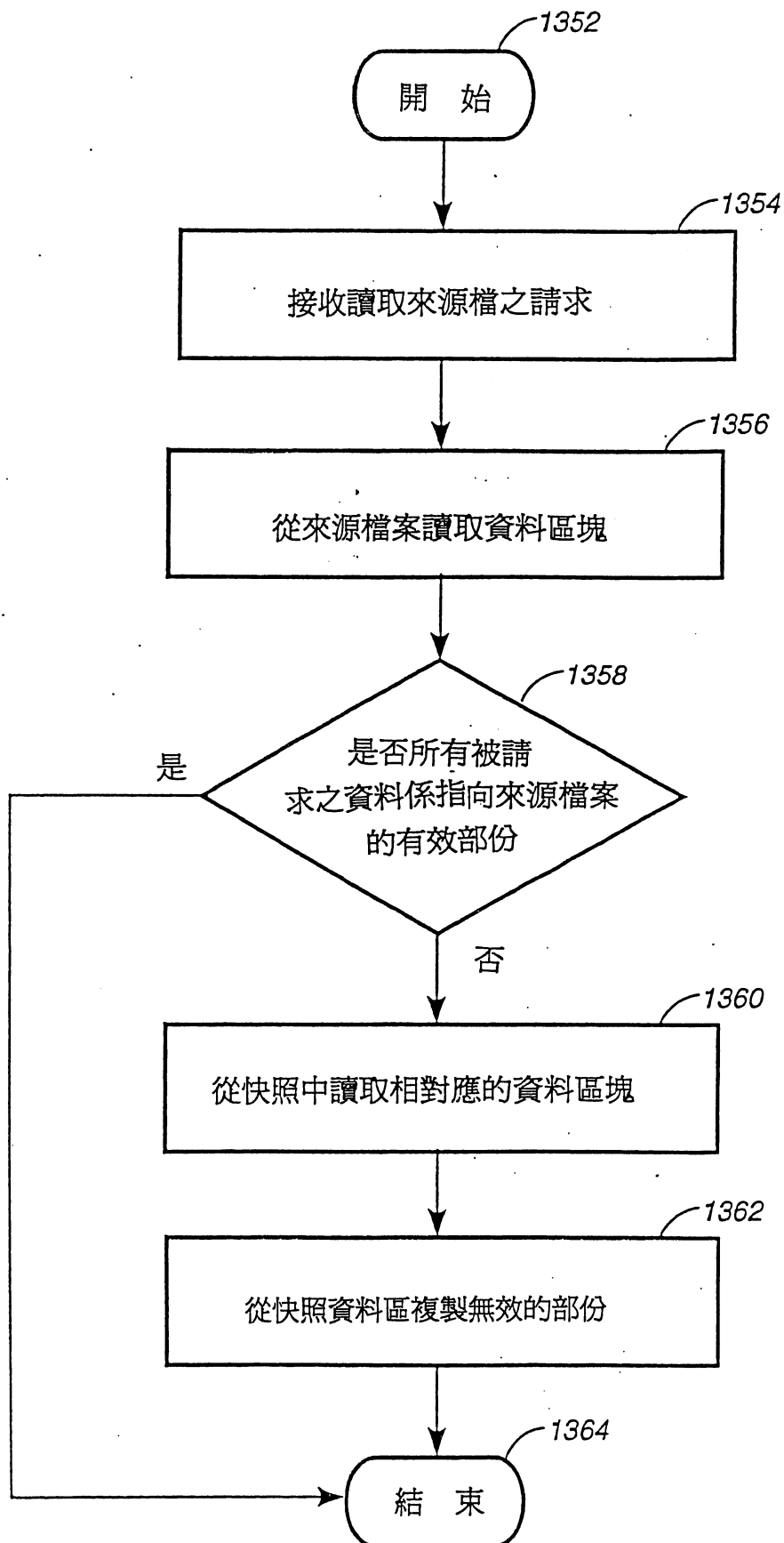
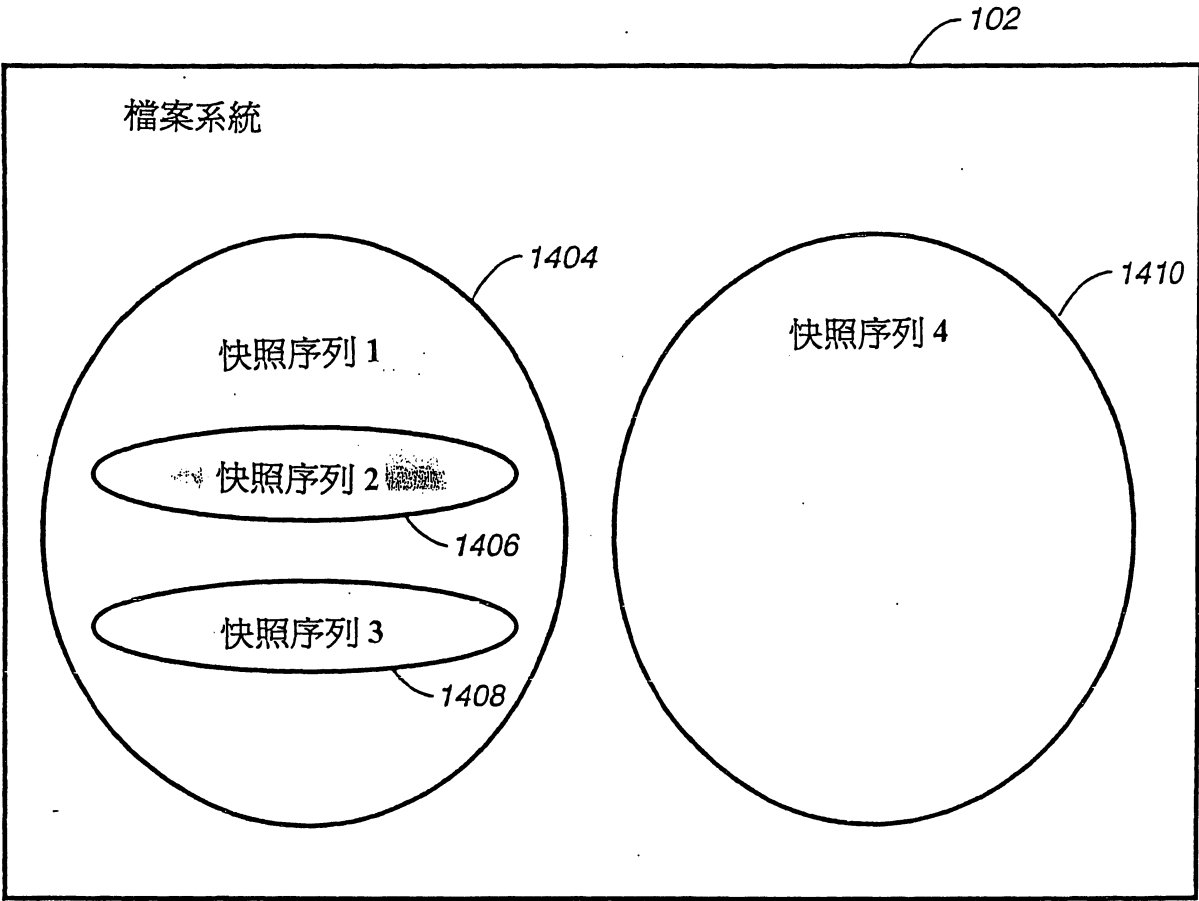
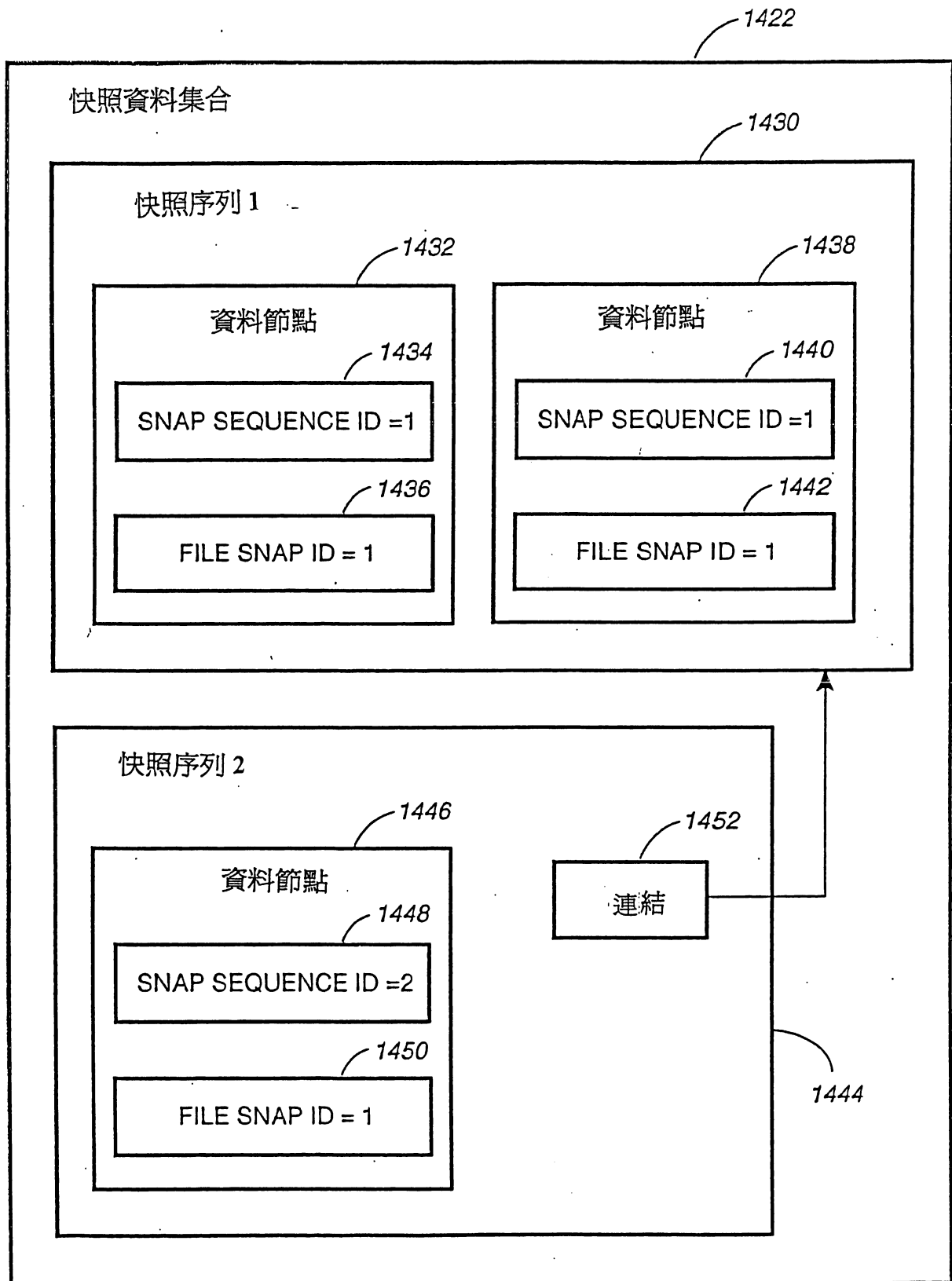


圖 13C



1400

圖 14A



1420

圖 14B

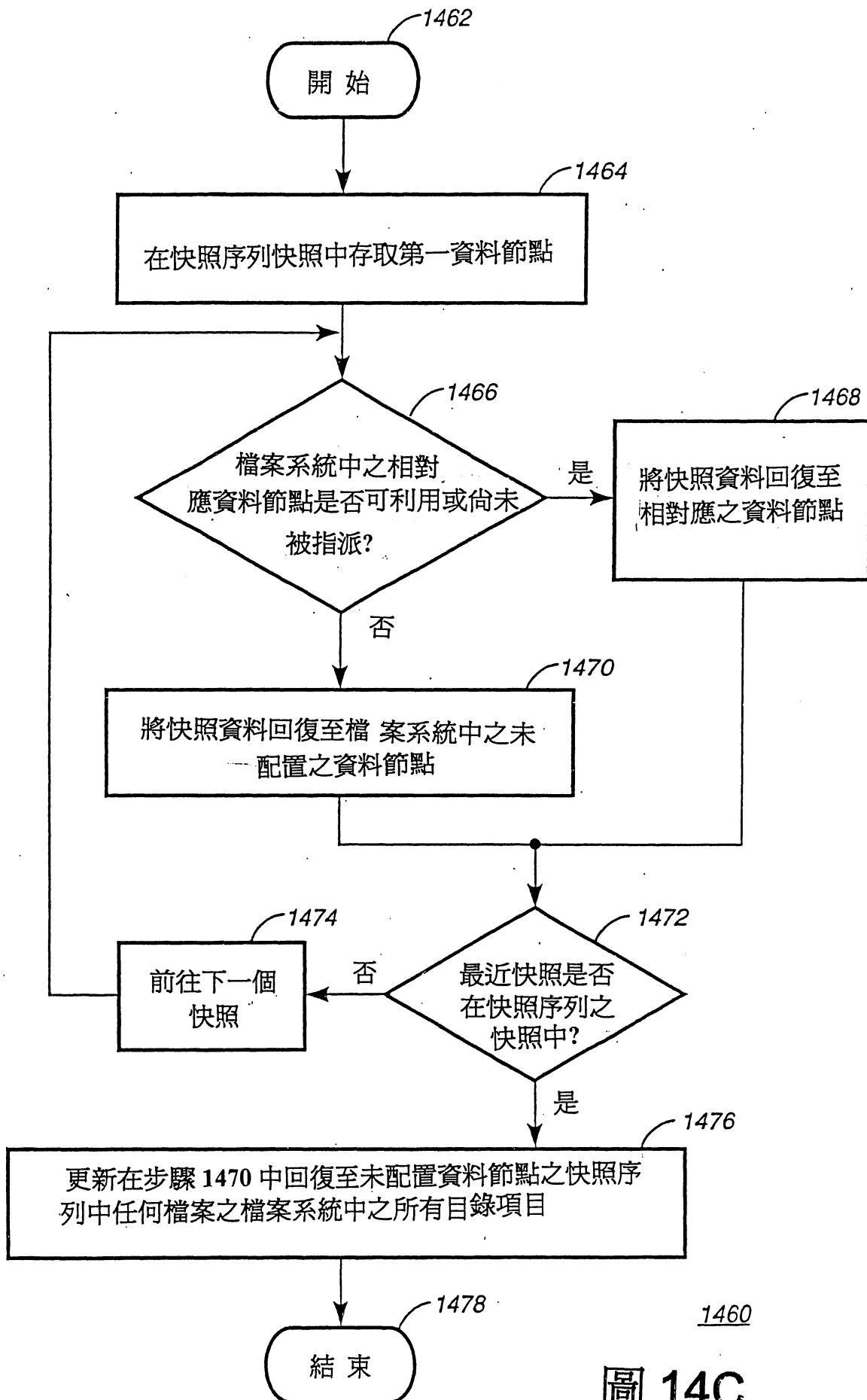


圖 14C

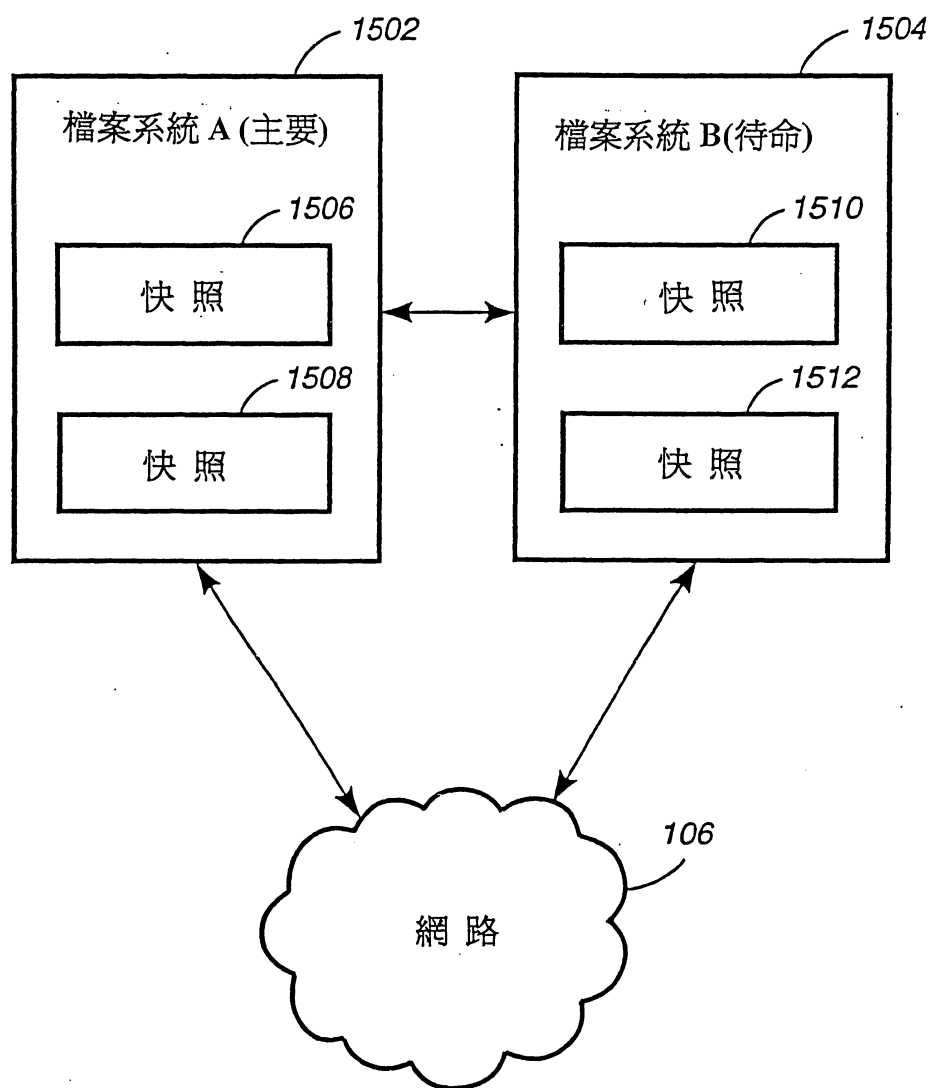


圖 15A

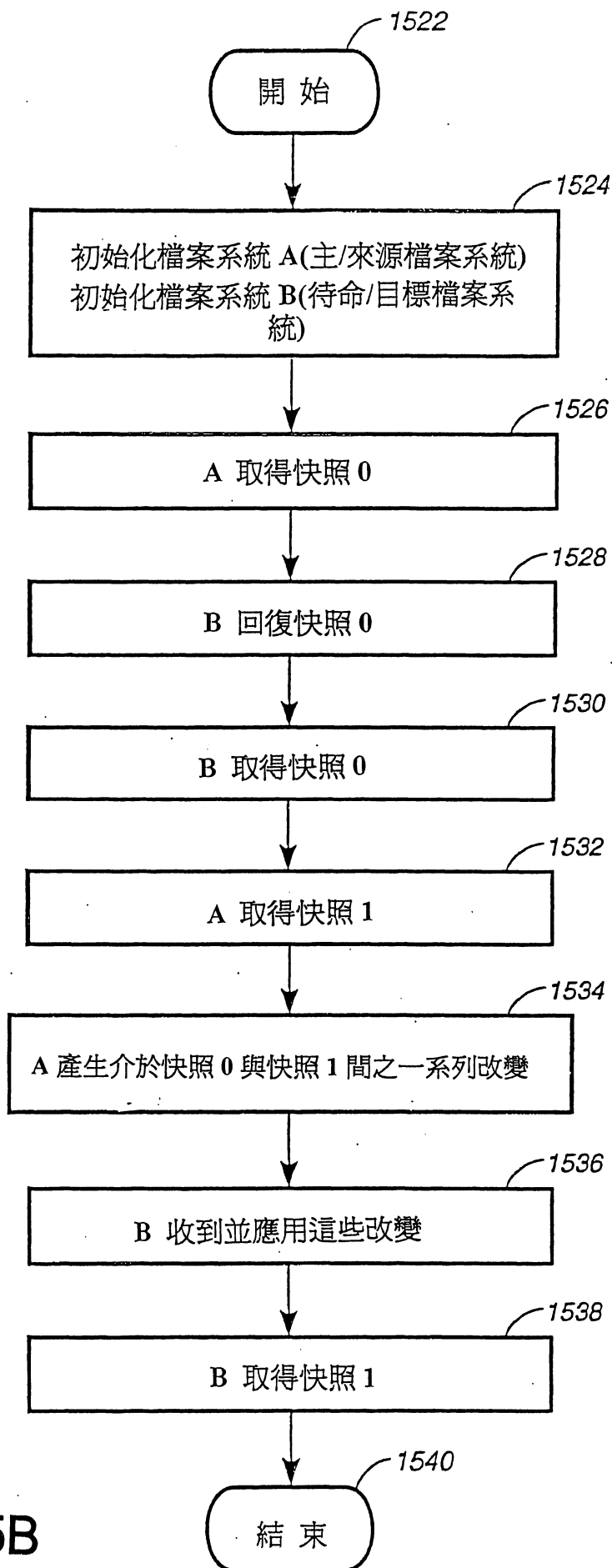
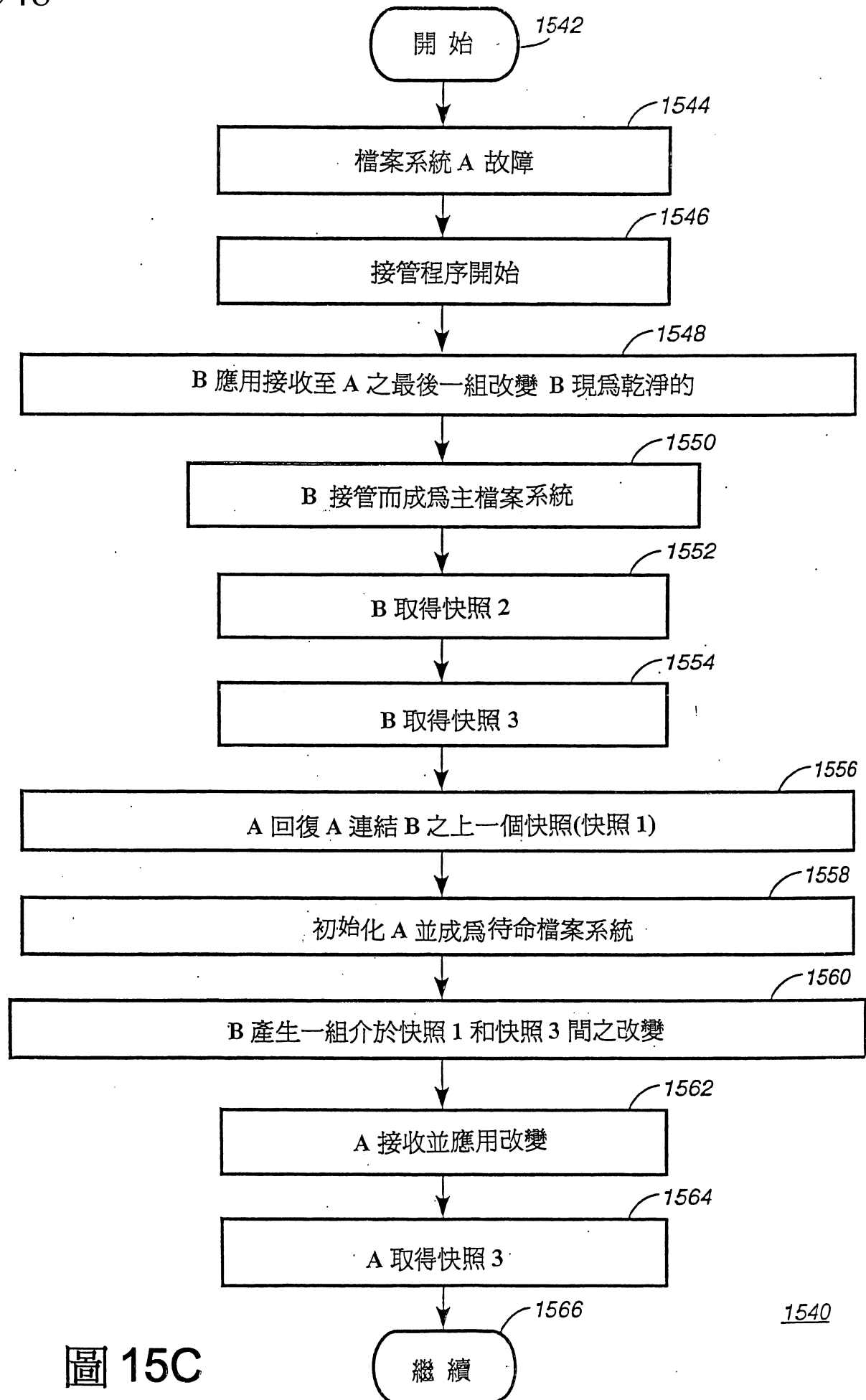
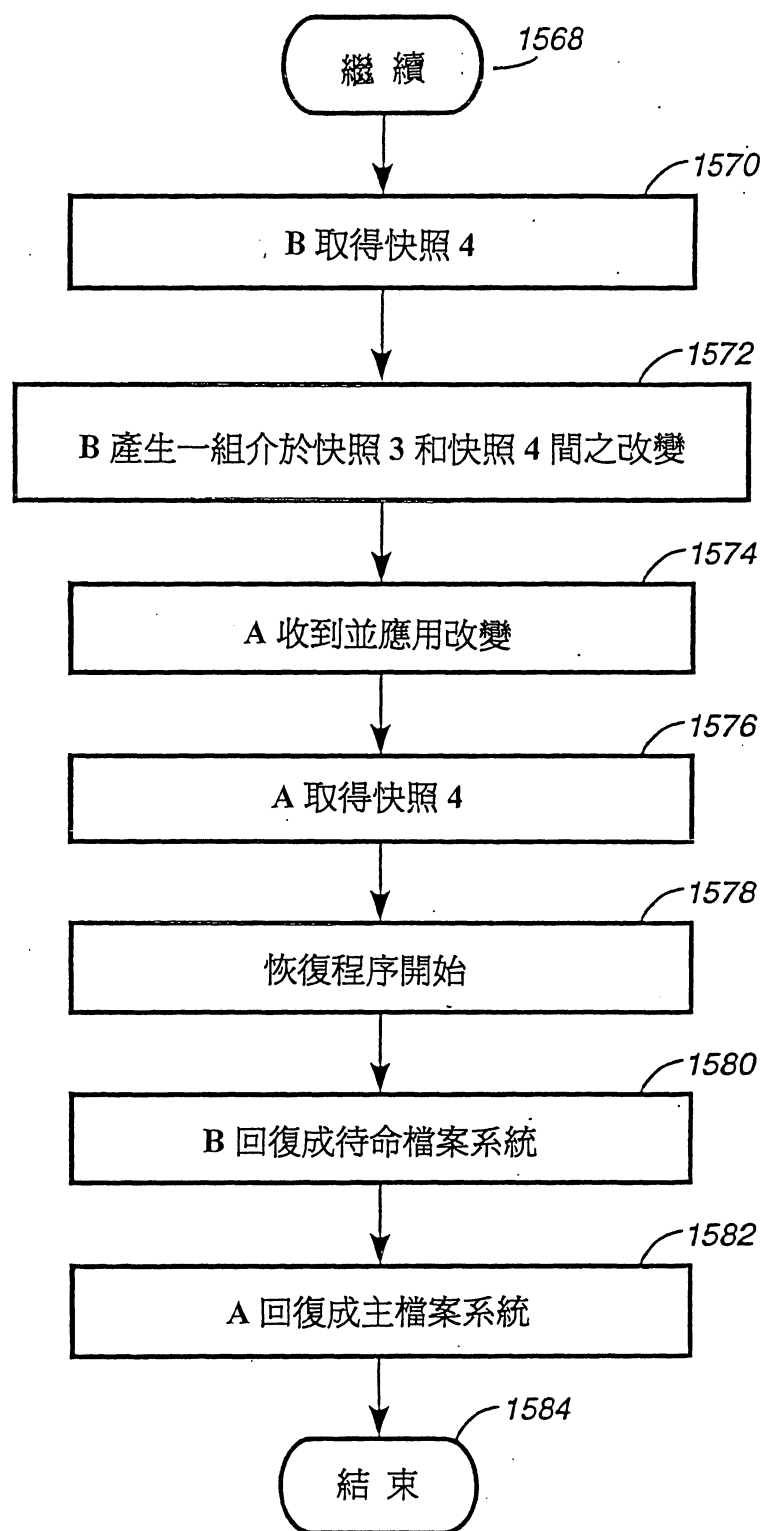


圖 15B





1590

圖 15D

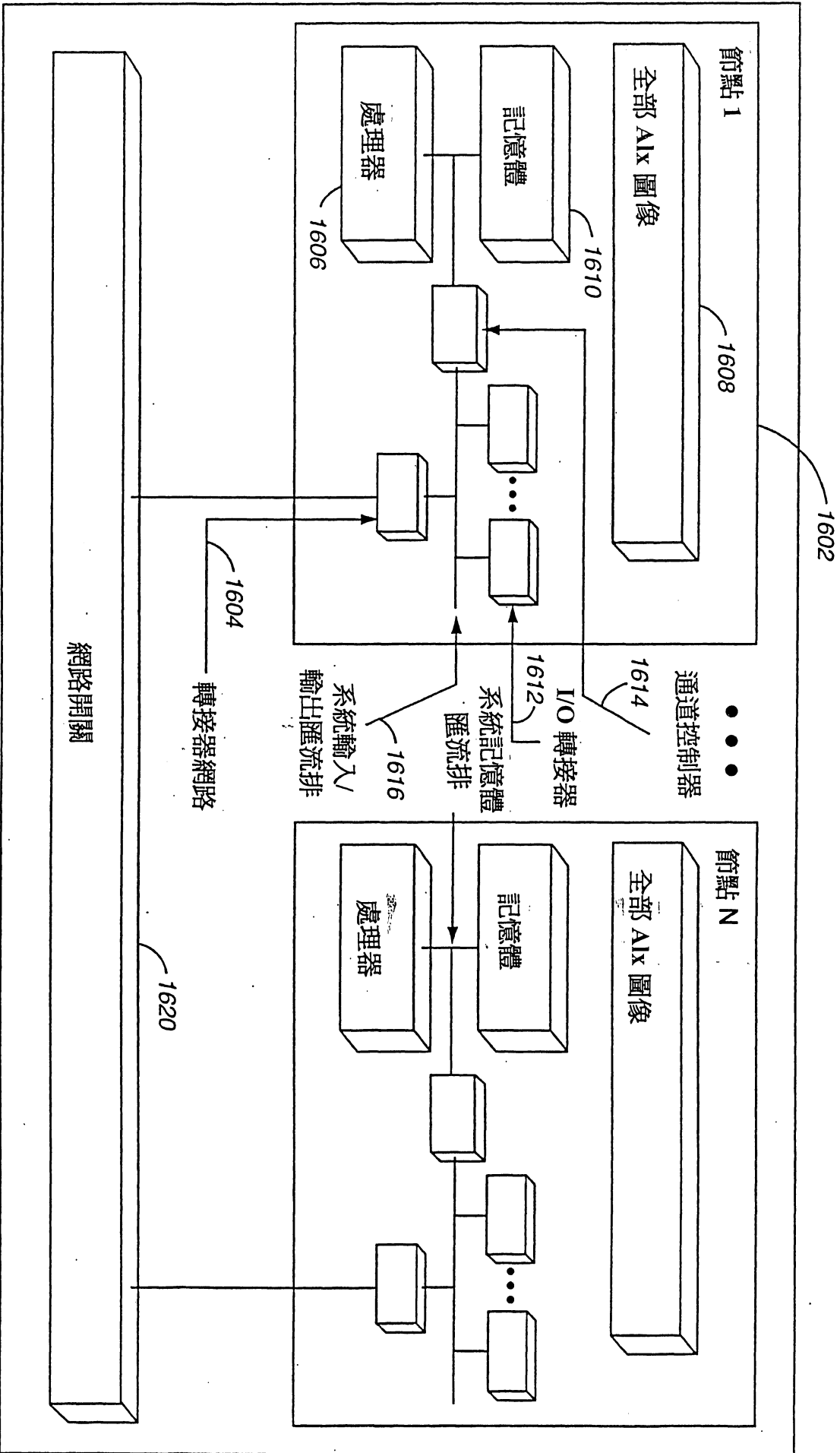


圖 16

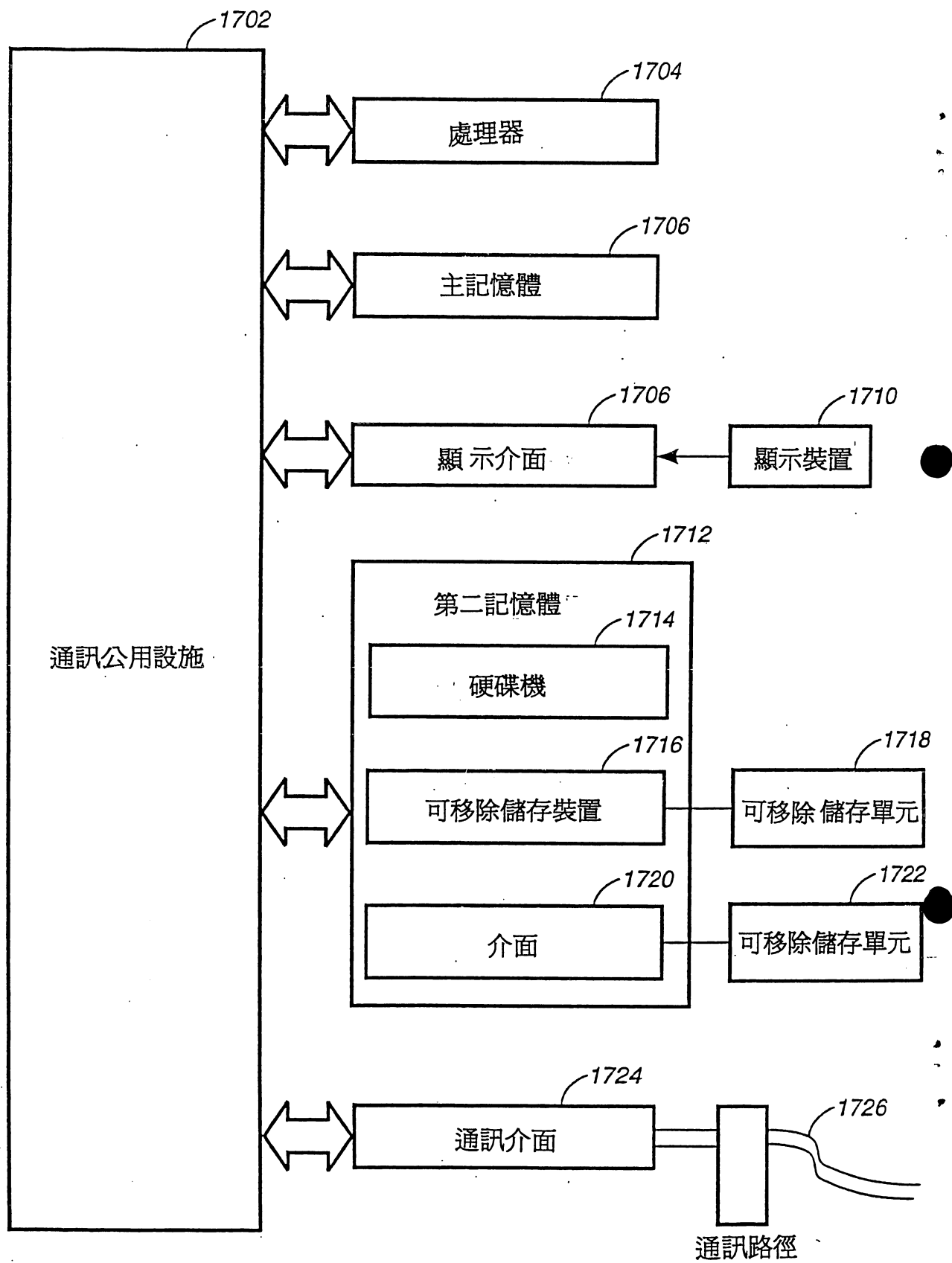


圖 17

申請日期：92.2.6

IPC分類

申請案號：92102430

G06F12/16

(以上各欄由本局填註)

發明專利說明書

一、 發明名稱	中文	提供檔案系統子集之快照
	英文	PROVIDING A SNAPSHOT OF A SUBSET OF A FILE SYSTEM
二、 發明人 (共2人)	姓名 (中文)	1. 梭頓, 韋恩 A. 2. 舒馬克, 法蘭克 B.
	姓名 (英文)	1. SAWDON, Wayne A. 2. SCHMUCK, Frank B.
	國籍 (中英文)	1. 美國 US 2. 美國 US
	住居所 (中文)	1. 美國加州95119聖荷西市聖安撒莫路6362號 2. 美國加州95008坎貝爾市聯盟大道406-A號
	住居所 (英文)	1. 6362 San Anselmo Way, San Jose, California 95119 USA 2. 406-A Union Avenue, Campbell, California 95008 USA
三、 申請人 (共1人)	名稱或姓名 (中文)	1. 萬國商業機器公司
	名稱或姓名 (英文)	1. INTERNATIONAL BUSINESS MACHINES CORPORATION
	國籍 (中英文)	1. 美國 US
	住居所 (營業所) (中文)	1. 美國紐約州10504亞芒克市新奧爾察德路 (本地址與前向貴局申請者相同)
	住居所 (營業所) (英文)	1. New Orchard Road, Armonk, NY 10504, USA
	代表人 (中文)	1. 傑羅 羅森梭
	代表人 (英文)	1. Gerald Rosenthal



六、申請專利範圍

1. 一種產生快照的方法，包含：

供在一檔案系統中之一第一複數個原始檔 (source file)，產生一第一快照，其中該第一快照包含一資料節點 (inode)，對應每一該第一複數個原始檔；

在每一對應每一該第一複數個原始檔之該資料節點 (inode)，儲存有關該第一複數個原始檔之一識別符 (identifier)；以及

在每一對應每一該第一複數個原始檔之該資料節點 (inode)，儲存有關該第一快照之時間。

2. 如申請專利範圍第 1 項所述之方法，進一步包含：

每隔一第一期間執行該產生步驟，該第一儲存步驟及該第二儲存步驟。

3. 如申請專利範圍第 2 項所述之方法，進一步包含：

供在一檔案系統中之一第二複數個原始檔 (source file)，產生一第二快照，其中該第二快照包含一資料節點 (inode)，對應每一該第二複數個原始檔；

在每一對應每一該第二複數個原始檔之該資料節點 (inode)，儲存有關該第二複數個原始檔之一識別符 (identifier)；以及

在每一對應每一該第二複數個原始檔之該資料節點 (inode)，儲存有關該第二快照之時間。

4. 如申請專利範圍第 3 項所述之方法，進一步包含：

每隔一第二期間執行該第二產生步驟，該第三儲存步驟及該第四儲存步驟。



六、申請專利範圍

5.如申請專利第4項所述之方法，其中至少一些該第一複數個原始檔及該第二複數個原始檔不相同。

6.如申請專利第4項所述之方法，其中該第一期間及該第二期間不相等。

7.如申請專利第4項所述之方法，進一步包含：

接收一請求，以重存一目的快照；

為每一在該目的快照中之資料節點，決定在該檔案系統中之一對應資料節點是否可用；

其中若在該檔案系統中之該對應資料節點可用，重存該目的快照之該資料節點至在該檔案系統中之該對應資料節點；以及

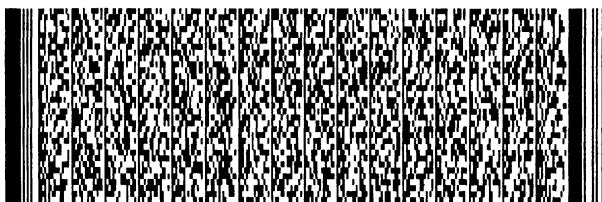
其中若在該檔案系統中之該對應資料節點不可用，重存該目的快照之該資料節點至在該檔案系統中之一未分配之資料節點。

8.如申請專利範圍第7項所述之方法，其中一可用資料節點包含任何一個：

一未分配資料節點；以及

一資料節點，對應一原始檔，該原始檔為該相同複數個檔案之一，該相同複數個檔案係包含於已被重存之該快照中。

9.如申請專利範圍第7項所述之方法，其中該第二重存步驟包含，重存在該目的快照之該資料節點至該檔案系統中之一未分配資料節點，及供該檔案系統中之該對應資料節點更新一目錄入口，以參考之前之該未分配節點。



六、申請專利範圍

10.一種重存一快照之方法，其中複數個快照可用，包含：

接收一請求，以重存一目的快照；

為每一在該目的快照中之資料節點，決定在該檔案系統中之一對應資料節點是否可用；

其中若在該檔案系統中之該對應資料節點可用，重存該目的快照之該資料節點至在該檔案系統中之該對應資料節點；以及

其中若在該檔案系統中之該對應資料節點不可用，重存該目的快照之該資料節點至在該檔案系統中之一未分配之資料節點。

11.一種電腦可讀媒體，包含供產生快照之電腦指令，該電腦指令包含指令供：

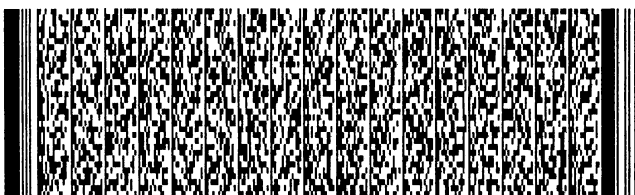
供在一檔案系統中之一第一複數個原始檔 (source file)，產生一第一快照，其中該第一快照包含一資料節點 (inode)，對應每一該第一複數個原始檔；

在每一對應每一該第一複數個原始檔之該資料節點 (inode)，儲存有關該第一複數個原始檔之一識別符 (identifier)；以及

在每一對應每一該第一複數個原始檔之該資料節點 (inode)，儲存有關該第一快照之時間。

12.如申請專利範圍第 11 項所述之電腦可讀媒體，該電腦指令進一步包含指令供：

每隔一第一期間執行供產生之指令，供儲存之第一指



六、申請專利範圍

令及供儲存之第二指令。

13.如申請專利範圍第12項所述之電腦可讀媒體，該電腦指令進一步包含指令供：

供在一檔案系統中之一第二複數個原始檔(source file)，產生一第二快照，其中該第二快照包含一資料節點(inode)，對應每一該第二複數個原始檔；

在每一對應每一該第二複數個原始檔之該資料節點(inode)，儲存有關該第二複數個原始檔之一識別符(identifier)；以及

在每一對應每一該第二複數個原始檔之該資料節點(inode)，儲存有關該第二快照之時間。

14.如申請專利範圍第13項所述之電腦可讀媒體，該電腦指令進一步包含指令供：

每隔一第二期間執行供產生之指令，供儲存之第三指令及供儲存之第四指令。

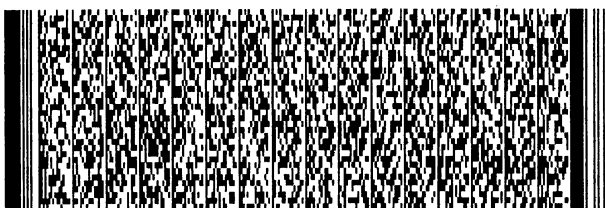
15.如申請專利第14項所述之電腦可讀媒體，其中至少一些該第一複數個原始檔及該第二複數個原始檔不相同。

16.如申請專利第14項所述之電腦可讀媒體，其中該第一期間及該第二期間不相等。

17.如申請專利第14項所述之電腦可讀媒體，該電腦指令進一步包含指令供：

接收一請求，以重存一目的快照；

為每一在該目的快照中之資料節點，決定在該檔案系統中之一對應資料節點是否可用；



六、申請專利範圍

其中若在該檔案系統中之該對應資料節點可用，重存該目的快照之該資料節點至在該檔案系統中之該對應資料節點；以及

其中若在該檔案系統中之該對應資料節點不可用，重存該目的快照之該資料節點至在該檔案系統中之一未分配之資料節點。

18.如申請專利範圍第17項所述之電腦可讀媒體，其中一可用資料節點包含任何一個：

一未分配資料節點；以及

一資料節點，對應一原始檔，該原始檔為該相同複數個檔案之一，該相同複數個檔案係包含於已被重存之該快照中。

19.如申請專利範圍第17項所述之電腦可讀媒體，其中供重存之該第二指令包含，重存在該目的快照之該資料節點至該檔案系統中之一未分配資料節點，及供該檔案系統中之該對應資料節點更新一目錄入口，以參考之前之該未分配節點。

20.一種電腦可讀媒體，包含電腦指令供重存一快照，其中複數個快照可用，該電腦指令包含指令供：

接收一請求，以重存一目的快照；

為每一在該目的快照中之資料節點，決定在該檔案系統中之一對應資料節點是否可用；

其中若在該檔案系統中之該對應資料節點可用，重存該目的快照之該資料節點至在該檔案系統中之該對應資料



六、申請專利範圍

節點；以及

其中若在該檔案系統中之該對應資料節點不可用，重存該目的快照之該資料節點至在該檔案系統中之一未分配之資料節點。

21.一種產生快照的系統，包含：

一裝置供在一檔案系統中之一第一複數個原始檔 (source file)，產生一第一快照，其中該第一快照包含一資料節點 (inode)，對應每一該第一複數個原始檔；

一裝置供在每一對應每一該第一複數個原始檔之該資料節點 (inode)，儲存有關該第一複數個原始檔之一識別符 (identifier)；以及

一裝置供在每一對應每一該第一複數個原始檔之該資料節點 (inode)，儲存有關該第一快照之時間。

22.如申請專利範圍第21項所述之系統，進一步包含：

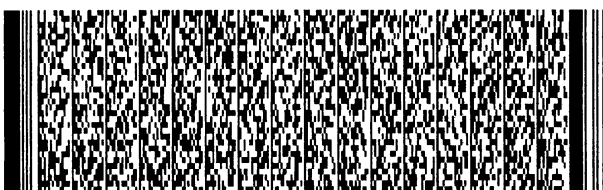
一裝置供每隔一第一期間執行供產生之該裝置，供儲存之該第一裝置及供儲存之該第二裝置。

23.如申請專利範圍第22項所述之系統，進一步包含：

一裝置供在一檔案系統中之一第二複數個原始檔 (source file)，產生一第二快照，其中該第二快照包含一資料節點 (inode)，對應每一該第二複數個原始檔；

一裝置供在每一對應每一該第二複數個原始檔之該資料節點 (inode)，儲存有關該第二複數個原始檔之一識別符 (identifier)；以及

在每一對應每一該第二複數個原始檔之該資料節點



六、申請專利範圍

(inode)，儲存有關該第二快照之時間。

24.如申請專利範圍第23項所述之系統，進一步包含：

一裝置供每隔一第二期間執行供產生之該第二裝置，供儲存之該第三裝置及供儲存之該第四裝置。

25.如申請專利第24項所述之系統，其中至少一些該第一複數個原始檔及該第二複數個原始檔不相同。

26.如申請專利第24項所述之系統，其中該第一期間及該第二期間不相等。

27.如申請專利第24項所述之系統，進一步包含：

一裝置供接收一請求，以重存一目的快照；

一裝置供為每一在該目的快照中之資料節點，決定在該檔案系統中之一對應資料節點是否可用；

一裝置供重存該目的快照之該資料節點至在該檔案系統中之該對應資料節點；以及

一裝置供重存該目的快照之該資料節點至在該檔案系統中之一未分配之資料節點。

28.如申請專利範圍第27項所述之系統，其中一可用資料節點包含任何一個：

一未分配資料節點；以及

一資料節點，對應一原始檔，該原始檔為該相同複數個檔案之一，該相同複數個檔案係包含於已被重存之該快照中。

29.如申請專利範圍第27項所述之系統，其中供重存之該第二裝置包含，一裝置供重存在該目的快照之該資料節點



六、申請專利範圍

至該檔案系統中之一未分配資料節點，及供該檔案系統中之該對應資料節點更新一目錄入口，以參考之前之該未分配節點。

30.一種重存一快照之系統，其中複數個快照可用，包含：

一裝置供接收一請求，以重存一目的快照；

一裝置供為每一在該目的快照中之資料節點，決定在該檔案系統中之一對應資料節點是否可用；

一裝置供重存該目的快照之該資料節點至在該檔案系統中之該對應資料節點；以及

一裝置供重存該目的快照之該資料節點至在該檔案系統中之一未分配之資料節點。

31.一種產生快照的系統，包含：

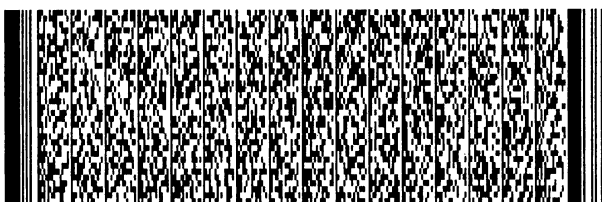
一第一快照，供在一檔案系統中之一第一複數個原始檔 (source file)，其中該第一快照包含一資料節點 (inode)，對應每一該第一複數個原始檔；

一識別符，有關該第一複數個原始檔，該識別符儲存在每一對應每一該第一複數個原始檔之資料節點 (inode)；以及

一識別符，有關該第一快照之時間，該識別符儲存在每一對應每一該第一複數個原始檔之資料節點。

32.如申請專利範圍第31項所述之系統，進一步包含：

一第二快照，供在一檔案系統中之一第二複數個原始檔 (source file)，其中該第二快照包含一資料節點



六、申請專利範圍

(inode)，對應每一該第二複數個原始檔；

一識別符，有關該第二複數個原始檔，該識別符儲存在每一對應每一該第二複數個原始檔之資料節點

(inode)；以及

一識別符，有關該第二快照之時間，該識別符儲存在每一對應每一該第二複數個原始檔之資料節點。

33.如申請專利範圍第32項所述之系統，其中至少一些該第一複數個原始檔及該第二複數個原始檔不相同。

