I497316

(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 I497316 B**

(45)公告日： 中華民國 104 (2015) 年 08 月 21 日

(21)申請案號：099111941 (22)申請日： 中華民國 99 (2010) 年 04 月 16 日

(51)Int. Cl.： ***G06F15/17 (2006.01)***

(30)優先權：2009/09/01 美國 12/552,058

(71)申請人：ＬＳＩ公司 (美國) LSI CORPORATION (US)
美國

(72)發明人：威斯勒 羅絲Ｅ ZWISLER, ROSS E. (US)；薛費爾德 羅伯特Ｌ SHEFFIELD, ROBERT L. (US)；史布 安德魯Ｊ SPRY, ANDREW J. (US)；費丁 傑瑞德Ｊ FREDIN, GERALD J. (US)；吉布森 肯尼斯Ｊ GIBSON, KENNETH J. (US)

(74)代理人：陳長文

(56)參考文獻：

TW 200519597 TW 200636482

EP 1811378A2

審查人員：游象甫

申請專利範圍項數：20 項 圖式數：11 共 47 頁

(54)名稱

透過小型計算機系統介面輸入／輸出(SCSI I/O)指示提供多重路徑之方法、電腦可讀取媒體及系統

METHOD, COMPUTER-READABLE MEDIUM, AND SYSTEM FOR PROVIDING MULTI-PATHING VIA SMALL COMPUTER SYSTEM INTERFACE INPUT/OUTPUT (SCSI I/O) REFERRAL

(57)摘要

本發明係一種透過小型計算機系統介面(SCSI)輸入/輸出指示而在透過一網路來連通耦接之一起始端和一儲存叢集之間提供多重路徑之方法，該儲存叢集至少包含一第一目標端裝置及一第二目標端裝置。該方法包含於該第一目標端裝置處透過該網路接收來自該起始端之輸入/輸出(I/O)。該輸入/輸出包含一資料要求。該方法進一步包含當包含於該資料要求中之資料未儲存於該第一目標端裝置上而是儲存於該第二目標端裝置上時，將一小型計算機系統介面輸入/輸出指示列表傳送至該起始端。該指示列表包含用以分別辨識該第二目標端裝置之第一及第二埠之第一及第二埠識別符。該第一及第二埠識別符係小型計算機系統介面相對埠識別符。該目標端裝置之第一及第二埠被辨識為用以存取該資料要求中所要求資料之存取埠。

The present invention is a method for providing multi-pathing via Small Computer System Interface Input/Output (SCSI I/O) referral between an initiator and a storage cluster which are communicatively coupled via a network, the storage cluster including at least a first target device and a second target device. The method includes receiving an input/output (I/O) at the first target device from the initiator via the network. The I/O includes a data request. The method further includes transmitting a SCSI I/O referral list to the initiator when data included in the data request is not stored on the first target device, but is stored on the second target device. The referral list includes first and second port identifiers for identifying first and second ports of the second target device respectively. The first and second port identifiers are SCSI relative port identifiers. The first and second ports of the target device are identified as access ports for accessing the data requested in the data request.

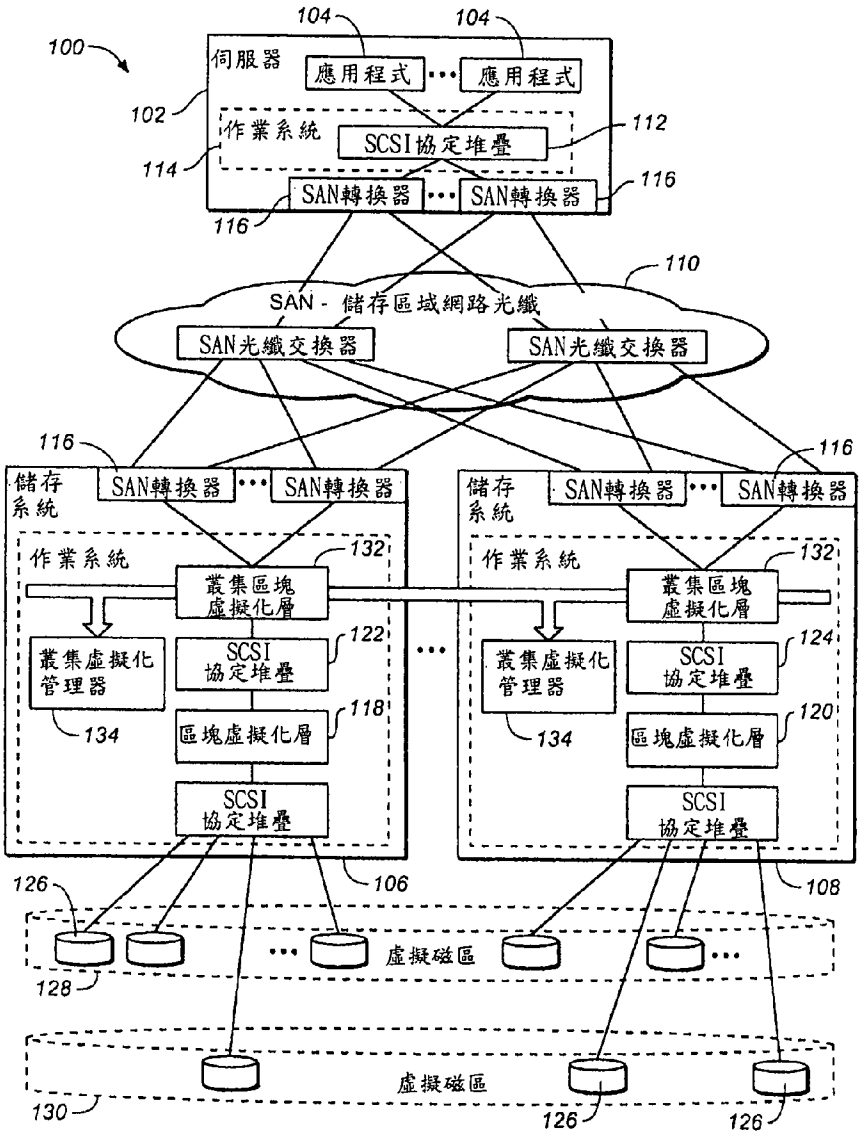

圖1

100・・・網路式儲存配置/系統/儲存基礎架構

102・・・應用系統/伺服器

104・・・應用程式

106/108・・・儲存系統

110・・・儲存區域網路

112・・・小型計算機系統介面協定堆疊

114・・・作業系統

116・・・儲存區域網路轉換器

118/120・・・本地區塊虛擬化層

122/124・・・內部區塊協定堆疊

126・・・實體儲存裝置

128/130・・・虛擬磁區

132・・・叢集區塊虛擬化層

134・・・叢集虛擬化管理器函式

發明專利說明書

公告本

（本說明書格式、順序，請勿任意更動，※記號部分請勿填寫）

※ 申請案號：099111941

※ 申請日：99年4月16日　　※IPC分類：G06F 15/17 (2006.01)

一、發明名稱：(中文/英文)

透過小型計算機系統介面輸入/輸出(SCSI I/O)指示提供多重路徑之方法、電腦可讀取媒體及系統

METHOD, COMPUTER-READABLE MEDIUM, AND SYSTEM FOR PROVIDING MULTI-PATHING VIA SMALL COMPUTER SYSTEM INTERFACE INPUT/OUTPUT (SCSI I/O) REFERRAL

二、中文發明摘要：

本發明係一種透過小型計算機系統介面(SCSI)輸入/輸出指示而在透過一網路來連通耦接之一起始端和一儲存叢集之間提供多重路徑之方法，該儲存叢集至少包含一第一目標端裝置及一第二目標端裝置。該方法包含於該第一目標端裝置處透過該網路接收來自該起始端之輸入/輸出(I/O)。該輸入/輸出包含一資料要求。該方法進一步包含當包含於該資料要求中之資料未儲存於該第一目標端裝置上而是儲存於該第二目標端裝置上時，將一小型計算機系統介面輸入/輸出指示列表傳送至該起始端。該指示列表包含用以分別辨識該第二目標端裝置之第一及第二埠之第一及第二埠識別符。該第一及第二埠識別符係小型計算機系統介面相對埠識別符。該目標端裝置之第一及第二埠被辨識為用以存取該資料要求中所要求資料之存取埠。

三、英文發明摘要：

The present invention is a method for providing multi-pathing via Small Computer System Interface Input/Output (SCSI I/O) referral between an initiator and a storage cluster which are communicatively coupled via a network, the storage cluster including at least a first target device and a second target device. The method includes receiving an input/output (I/O) at the first target device from the initiator via the network. The I/O includes a data request. The method further includes transmitting a SCSI I/O referral list to the initiator when data included in the data request is not stored on the first target device, but is stored on the second target device. The referral list includes first and second port identifiers for identifying first and second ports of the second target device respectively. The first and second port identifiers are SCSI relative port identifiers. The first and second ports of the target device are identified as access ports for accessing the data requested in the data request.

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:

100~網路式儲存配置/系統/儲存基礎架構

102~應用系統/伺服器

104~應用程式

106/108~儲存系統

110~儲存區域網路

112~小型計算機系統介面協定堆疊

114~作業系統

116~儲存區域網路轉換器

118/120~本地區塊虛擬化層

122/124~內部區塊協定堆疊

126~實體儲存裝置

128/130~虛擬磁區

132~叢集區塊虛擬化層

134~叢集虛擬化管理器函式

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】

本發明關於透過一區塊儲存協定所存取之網路型儲存領域，且更特別關於一種於一起始端系統及一區塊儲存叢集之間透過由多個傳輸協定所配置之一網路來提供具有小型計算機系統介面(SCSI)輸入/輸出指示之多重路徑之系統及方法。

【先前技術】

目前在一區塊儲存叢集及一起始端之間提供通訊之可行系統/方法也許無法提供想要的執行效率水準。

因此期待提供在一區塊儲存叢集及一起始端之間提供通訊之系統/方法，上述所提涉及問題目前可行之解決方案。

【發明內容】

有鑑於此，本發明一實施例係指向一種透過小型計算機系統介面(SCSI)輸入/輸出指示於透過一網路來連通耦接之一起始端和一儲存叢集之間提供多重路徑之方法，該儲存叢集至少包含一第一目標端裝置及一第二目標端裝置，該方法包含：於該第一目標端裝置處透過該網路接收來自該起始端之一輸入/輸出(I/O)，該輸入/輸出包含對一第一部分資料及對一第二部分資料之一要求；以及當不是該第二部分資料而是該第一部分資料被儲存於該第一目標端裝置上，且該第二部分資料被儲存於該第二目標端裝置上時，

啟動該第一部分資料至該起始端之傳送並將一小型計算機系統介面輸入/輸出指示列表傳送至該起始端，其中該指示列表包含用以辨識該第二目標端裝置之一第一埠之一第一埠識別符及用以辨識該第二目標端裝置之一第二埠之一第二埠識別符，該第一埠及該第二埠被辨識為用以存取該第二部分資料之存取埠，其中該第一埠識別符及該第二埠識別符係小型計算機系統介面相對埠識別符。

本發明另一實施例係指向一種具有電腦可執行指令以執行用以透過小型計算機系統介面(SCSI)輸入/輸出指示於透過一網路來連通耦接之一起始端和一儲存叢集之間提供多重路徑之方法之電腦可讀取媒體，該儲存叢集至少包含一第一目標端裝置及一第二目標端裝置，該方法包含：於該第一目標端裝置處透過該網路接收來自該起始端之一輸入/輸出(I/O)，該輸入/輸出包含對一第一部分資料及對一第二部分資料之一要求；以及當不是該第二部分資料而是該第一部分資料被儲存於該第一目標端裝置上，且該第二部分資料被儲存於該第二目標端裝置上時，啟動該第一部分資料至該起始端之傳送並將一小型計算機系統介面輸入/輸出指示列表傳送至該起始端，其中該指示列表包含用以辨識該第二目標端裝置之一第一埠之一第一埠識別符及用以辨識該第二目標端裝置之一第二埠之一第二埠識別符，該第一埠及該第二埠被辨識為用以存取該第二部分資料之存取埠，其中該第一埠識別符及該第二埠識別符係小型計算機系統介面相對埠識別符。

本發明一進一步實施例係指向一種具有電腦可執行指令以執行用以透過小型計算機系統介面(SCSI)輸入/輸出指示於透過一網路來連通耦接之一起始端和一儲存叢集之間提供多重路徑之方法之電腦可讀取媒體，該儲存叢集至少包含一第一目標端裝置及一第二目標端裝置，該方法包含：於該第一目標端裝置處透過該網路接收來自該起始端之一輸入/輸出(I/O)，該輸入/輸出包含一資料要求；當包含於該資料要求中之資料並未儲存於該第一目標端裝置上，但是該資料儲存於該第二目標端裝置上時，將一小型計算機系統介面輸入/輸出指示列表傳送至該起始端，其中該指示列表包含用以辨識該第二目標端裝置之一第一埠之一第一埠識別符及用以辨識該第二目標端裝置之一第二埠之一第二埠識別符，該第一埠及該第二埠被辨識為用以存取該資料之存取埠；於該第二目標端裝置處透過該網路接收來自該起始端之一指示輸入/輸出，該指示輸入/輸出係回應至該小型計算機系統介面輸入/輸出指示列表，且該指示輸入/輸出要求透過該起始端所示之該第一埠及該第二埠中一者進行資料存取；以及啟動該資料至該起始端之一傳送，其中該第一埠識別符及該第二埠識別符係小型計算機系統介面相對埠識別符。

要了解到前面大體性說明及後面詳細說明兩者只是示範和說明，並不是如申請專利範圍般地限制本發明。聚合並構成本說明一部分之該些附圖說明本發明各實施例並連同該大體性說明一起用於說明本發明原理。

【實施方式】

現在將詳細參考本發明所示較佳實施例，其範例係示於該些附圖中。

大體上參考至圖 1，根據本發明一示範實施例之一種可透過一區塊儲存協定進行存取之網路型儲存配置/系統被顯示。例如，該區塊儲存協定可為配置於例如光纖通道、網際網路式小型計算機系統介面、或串接式小型計算機系統介面(SAS)之網路功能媒體傳輸裝置上之一小型計算機系統介面(SCSI)協定。本發明可進一步被配置於例如小型計算機系統介面之遠端直接記憶體存取協定(SRP)之其它區塊儲存協定中。本發明系統/方法擴充該區塊儲存協定以允許資料分佈於構成一區塊儲存叢集之多個不連續協同儲存系統之共同邏輯區塊位址空間。

在所示實施例(示於圖 1)，該網路型儲存配置/系統/儲存基礎架構 100 包含一應用系統/伺服器 102。該應用系統 102 可執行一或更多應用程式 104。執行於該應用系統 102 上之應用程式 104 可經由/透過/使用一儲存區域網路(SAN)110 來存取一或更多儲存系統(106、108)上所儲存之儲存資源。應用程式可使用該伺服器(未損失一般性)之作業系統 114 一區塊儲存協定堆疊(例如一小型計算機系統介面(SCSI)協定堆疊)112 來存取該些儲存資源/網路儲存資源。該伺服器 102 之作業系統 114 可直接在伺服器硬體上執行或於一虛擬化環境之虛擬機器內執行而不會損失一般性。

在本發明目前實施例中，該伺服器 102 之小型計算機系統介面協定堆疊 112 可為該些應用程式 104 之本地(例如在該伺服器上)或遠端(例如在該網路上)儲存資源而以區塊儲存裝置/邏輯單元/小型計算機系統介面邏輯單元來展現。每一個邏輯單元/小型計算機系統介面邏輯單元可具有一獨一無二的邏輯區塊位址空間。該些遠端儲存資源/遠端儲存裝置(106、108)可由該伺服器 102 及/或儲存系統(106、108)之一或更多儲存區域網路轉換器 116 進行存取，其可執行一有區塊儲存協定映射於其上之網路媒體傳輸協定。例如，小型計算機系統介面協定可映射於各類型可配置網路傳輸裝置上。不會損及一般性地，該儲存區域網路轉換器 116 及其媒體傳輸協定層可為實體或虛擬網路轉換器。

在本發明示範實施例中，該儲存區域網路 110 可由允許埠位準定址(例如光纖通道、乙太網、InfiniBand、及串接式小型計算機系統介面(SAS))之任何網路媒體及傳輸協定來配置。媒體傳輸層協定可處理所有橫跨網路光纖 110 之點對點埠之協定封包路徑。不會損及一般性地，該網路 110 可被配置成單一光纖或多條冗餘光纖。網路 110 可為由單一網路媒體及傳輸協定所配置之單一網路或由多個網路媒體及傳輸協定結合所配置之多個網路。較佳地，應用系統/伺服器 102 上之埠可真地到達儲存系統(106、108)上之埠。

在本發明另外實施例中，儲存系統(106、108)可為網路連接型儲存裝置。例如，該些儲存系統(106、108)可為通用電腦、專屬儲存陣列、或可讓其本地裝置於該儲存區域網

路 110 上顯而易見之網路型磁碟機。該儲存系統之儲存資源可透過正在執行媒體傳輸協定層之儲存區域埠來存取。小型計算機系統介面層可使用該些儲存區域埠做為儲存埠來與該儲存網路通訊。每一個儲存系統 106、108 可包含一可施加資料保護或區塊抽象至其實體儲存裝置之本地區塊虛擬化層(118、120)。例如，磁碟陣列(RAID)類之資料保護可被使用於專屬網路儲存系統上。每一個儲存系統 106、108 可進一步包含一可存取該些真正附接之實體儲存裝置 126 之內部區塊協定堆疊(122、124)，其輸出至該網路 110。

應用伺服器/應用系統/伺服器 102 可用之儲存量可藉由將更多儲存裝置 126 添加至個別儲存系統(106、108)或將額外儲存系統(106、108)添加至該儲存區域網路而延伸。當額外儲存裝置 126 被添加至該些個別儲存系統(106、108)時，在該些儲存系統(106、108)中之本地區塊虛擬化層(118、120)可被使用以自多個實體磁碟(126)中產生更大的虛擬磁區儲存裝置(128、130)。這個雖然可保留虛擬磁區(128、130)之單邏輯區塊位址空間，但在某種觀點下個別儲存系統(106、108)上之實體附接點數量可能被用盡因而產生總容量擴展限制。當儲存系統被添加至該儲存區域網路時，應用裝置可用之總儲存量可能增加超過單一儲存系統之物理限制。然而，多個儲存系統(106、108)所提供之儲存可能需要聚合至一共同邏輯區塊位址空間中以供應用伺服器(102)使用。

一些技術可被利用以在多個網路連接型儲存系統(106、108)上之儲存資源 126 中產生單一名稱空間/共同邏

輯區塊位址空間。例如，該些技術可使用例如叢集檔案系統或物件儲存協定之不同儲存協定。區塊儲存叢集聚合可被添加至該儲存網路 110，以使區塊叢集聚合可由在複數個冗餘儲存區域網路光纖中每一個之叢集區塊虛擬化裝置來提供。該叢集區塊虛擬化裝置可位在一網路儲存系統及一應用系統間。該叢集區塊虛擬化裝置可輸入該些網路儲存系統/儲存系統所輸出之區塊儲存邏輯單元，並可藉由產生虛擬磁區來產生一額外區塊虛擬化層。該叢集區塊虛擬化裝置接著可輸出該些虛擬磁區以做為應用系統之邏輯單元。該應用系統並不會看見或存取該儲存系統所輸出之邏輯單元，而是只會看見該些虛擬磁區/叢集虛擬磁區。該叢集實體結構探索、虛擬化映射及管理可由一叢集虛擬化管理器來提供。該叢集虛擬化管理器可位在該儲存區域網路邊緣內或其上之任何地方的一對獨立冗餘裝置上。不會損及一般性地，該區塊儲存叢集聚合函式可被分佈遍及該些叢集區塊虛擬化裝置/區塊叢集虛擬化裝置各處。

替代性地，區塊儲存叢集聚合/區塊叢集聚合可被添加至應用系統 102(應用系統聚合)。例如：區塊叢集聚合可由所添加至該應用系統的區塊儲存協定堆疊之一額外抽象層來提供。許多選項可被配置以將該抽象層放置在該應用系統上。該區塊虛擬化層可遮蓋或隱藏儲存系統所輸出之邏輯單元，也可將虛擬磁區呈現於該區塊儲存協定堆疊中該區塊虛擬化層上方各層。不像在區塊儲存叢集聚合被添加至該儲存區域網路(網路聚合)時，在將區塊儲存叢集聚合添

加至應用系統時，儲存系統所輸出之邏輯單元係可存取應用系統/伺服器。該區塊虛擬化層可隱藏對正在應用系統/伺服器上執行之應用程式之邏輯單元的存取。像網路聚合般，在將區塊叢集聚合添加至應用系統時，可存在一叢集虛擬化管理器函式以探索該叢集中之儲存資源並將虛擬化映射分佈遍及應用伺服器各處。本管理方法之一變化例可在每一個伺服器內含有獨立叢集虛擬化架構，而可防止虛擬磁區被各應用伺服器所分享。替代性地，為了提供共享該些虛擬磁區，需要一叢集級虛擬化管理器。

在本發明(如圖 1 所示)所示實施例中，區塊儲存叢集聚合可被添加至該些儲存系統(106、108) (儲存系統聚合)。該區塊叢集聚合可由添加至一或二儲存系統(106、108)之驅塊協定堆疊(122、124)之叢集區塊虛擬化層 132 所提供。該叢集區塊虛擬化層 132 可將本地及遠端儲存系統上之儲存裝置 126 結合成虛擬磁區(128、130)。在該叢集內之每一個儲存系統(106、108)上之儲存裝置 126 係可由其餘儲存系統(例如，儲存系統 106 之儲存裝置可被儲存系統 108 看到，且儲存系統 108 之儲存裝置可被儲存系統 106 看到)之一或更多偵測到/看到中一者，以讓該叢集區塊虛擬化層 132 可產生虛擬磁區(128、130)。在一些儲存系統聚合配置中，只有虛擬磁區(128、130)係由該叢集區塊虛擬化層 132 經該儲存區域網路 110 輸出至應用系統 102。在一些網路型儲存配置中，到達一儲存系統(106 或 108)之一輸入/輸出(I/O)要求，其需要一或更多不同儲存系統上之資料，可被傳送至正確

儲存系統以滿足該輸入/輸出要求。一些技術可被配置以執行例如代理伺服器輸入/輸出及命令傳送之輸入/輸出重新導向。如上述其它區塊儲存叢集技術，在儲存系統聚合中需要一獨立叢集虛擬化管理器函式 134 存在於該儲存基礎架構之儲存系統(106、108)中至少一者。不會損及一般性地，該叢集虛擬化管理器函式 134 可被分佈遍及該叢集內之儲存系統(106、108)各處，藉此提供一低成本、低侵入性配置給該儲存管理函式。

上述區塊儲存叢集產生方法/配置提供一些有用的特徵。例如，區塊儲存磁區可遍佈於多個儲存系統(106、108)。同時，應用系統 102 可存取該叢集內任何儲存系統上之資料。再進一步，該些虛擬磁區(128、130)可提供所有儲存節點/儲存系統(106、108)一共同區塊位址空間。然而，上面方法/配置(網路聚合、儲存系統聚合、應用系統聚合)中之每一個具有不同的缺點。

該網路 110(網路聚合)內之儲存聚合缺點在該網路 110 內可需額外特殊用途元件。該些特殊用途元件可增加該網路光纖 110 之成本而迫使單一儲存系統往高成本之多系統儲存叢集移動。進一步，若儲存聚合裝置未納入為了極小化整體成本所產生之網路光纖內，該網路也許需要重新佈線以聚合該些聚合裝置及叢集虛擬化管理器。再進一步，自單一儲存系統移至一儲存叢集可需要重新架構所有應用系統以使用虛擬磁區來取代該些儲存系統之原始磁區。

該應用伺服器 102(應用系統聚合)上之儲存聚合缺點在

於可能需要添加額外元件至該伺服器區塊儲存堆疊 112。也許需要該些元件來遮蓋所有非虛擬邏輯單元以避開正在該系統上執行之應用程式 104 之存取。若存取該叢集之所有作業系統 114 未完成遮蓋，則資料錯誤或遺失可能會因為對該些非虛擬邏輯單元之不規則性存取而發生。在該應用系統中之叢集區塊虛擬化層也是需要以提供該些應用程式區塊虛擬化。每一個作業系統可能需要獨一無二的叢集區塊虛擬化元件。這些獨一無二的叢集區塊虛擬化元件可能被迫在該系統的儲存堆疊內使用無授證介面來達成它們的功能，其可創造維護及測試所需架構之擴張。進一步，該叢集虛擬化管理器仍需要一獨立於該些應用系統之外部系統。若該叢集虛擬化管理器被置放在一應用系統上，它可能消耗應用程式可能使用之資源，同時，該叢集虛擬化管理器也可能需要與該基礎架構內之所有其它應用系統進行通訊。無關於儲存虛擬化管理器位置，需要一獨立協定來散佈並更新該叢集虛擬化管理器所維護及該應用系統內之叢集區塊虛擬化層所使用之區塊儲存地圖。

儲存系統(106、108)(儲存系統聚合)內之儲存聚合可消除額外網路元件成本。儲存系統聚合可進一步消除對該儲存堆疊 112 內之額外元件之應用伺服器 102 之影響，也可消除正在該應用伺服器 102 上執行之儲存虛擬化管理器之影響。儲存系統聚合可允許該區塊儲存叢集/區塊儲存串所需之所有元件位在該些儲存系統(106、108)上。儲存系統聚合可能需要輸入/輸出要求在被傳送至錯誤的儲存系統時重

新導向。如上所述，代理伺服器輸入/輸出及/或命令傳送可被使用於重新導向，然而兩者皆有其缺點。當代理伺服器輸入/輸出被使用時，這個對透過接收該原始錯誤導向要求之儲存系統來安排資料傳送路徑可能會添加一額外儲存及傳送延遲。傳遍一相連的私有儲存叢集之命令可能增加額外成本至該儲存叢集且可能限制該叢集之最大尺寸。

本發明藉由提供一小型計算機系統介面指示技術/方法來搭配例如圖 1 所示配置/系統 100 般之網路型儲存配置/系統使用而克服上述區塊儲存叢集技術弱點。本發明技術/方法被設計成可產生區塊儲存叢集而不需要該應用系統區塊儲存堆疊 112 內之非標準元件或該儲存網路 110 內之額外特殊用途叢集裝置。

大體上參考至圖 3，根據本發明一示範實施例之一種透過一網路型儲存配置(例如一在一起始端系統/起始端及一叢集式儲存陣列/區塊儲存叢集式陣列間之通訊方法)進行資料傳送之方法被顯示。例如，該方法可如下所述地(並如圖 2 至圖 3 所示)使用一儲存協定命令及回應序列(例如一小型計算機系統介面命令/回應遠端程序呼叫模式)來配置用於區塊儲存叢集之技術。在本發明一目前實施例中，該方法 300 包含該叢集式儲存陣列之複數個儲存系統內含之第一儲存系統接收一命令之步驟 302。例如該區塊儲存叢集可包含二或更多儲存系統(106、108)，每一個連通性地耦接/包含實體儲存裝置 126。進一步，該命令可由一起始端/起始端系統/主機/伺服器 102 透過一儲存區域網路 110 傳送至

該第一儲存系統 106(例如一目標端系統/目標端)。在示範實施例中，該命令可為例如一資料要求(例如讀取要求)之輸入/輸出要求。在進一步實施例中，該目標端可為該叢集式陣列中之任何儲存系統，且可使用該叢集/叢集式儲存陣列中之任何預期目標端儲存系統上之任何埠(例如如圖 2 所示之主要埠)來傳送該命令。又進一步，該命令可為一小型計算機系統介面命令，該起始端/起始端系統 102 可為一小型計算機系統介面起始端，且該目標端(例如第一儲存系統 106)可為一小型計算機系統介面目標端。

在額外實施例中，在該儲存區域網路 110/網路傳輸裝置上傳送時，該命令可在一已建立之起始端及目標端夥伴關係(例如一起始端-目標端連結)間傳輸。在小型計算機系統介面協定中，介於該起始端及目標間之起始端-目標端連結可被建立於一在該起始端之小型計算機系統介面埠(例如該伺服器/應用系統 102之小型計算機系統介面埠)及一在該目標端之小型計算機系統介面埠(例如該第一儲存系統 106之小型計算機系統介面埠)之間。一具有多個埠之目標端可提供每一個埠一獨一無二之小型計算機系統介面埠識別符。一具有多個儲存系統(例如該區塊儲存叢集之目標端)之區塊儲存叢集可提供一獨一無二之埠識別符給在該叢集內之所有儲存系統上之每一個埠。該些小型計算機系統介面埠識別符可為該小型計算機系統介面架構模型規格書中所定義之小型計算機系統介面相對埠識別符。在另一實施例中，每一個小型計算機系統介面命令可藉由它在該磁區

的邏輯區塊位址空間內之起始位址及長度來辨識要傳送之資料。

在示範實施例中，該方法 300 可進一步包含將該資料要求所要求之儲存於該第一儲存系統上之資料透過該儲存區域網路傳送至該起始端系統之步驟 304。在本發明目前實施例中，儲存/存在於接收該命令(例如該目標端儲存系統)之儲存系統上之任何部分要求資料可被移送/傳送至該起始端。例如，資料可藉由一連串小型計算機系統介面資料傳送步驟，透過上述/相同起始端-目標端連結(例如儲存於該第一資料儲存系統 106 上之資料可被傳送至該應用系統/起始端系統 102)而移動於該目標端 106 及該起始端 102 之間。在本發明目前實施例中，資料可視該特定小型計算機系統介面命令所要以任一方向或雙向方式流通於該起使端及目標端之間。

在本發明進一步實施例中，該方法 300 可進一步包含步驟 306 為在該資料要求所要求之部分資料雖未儲存/未存在於該第一儲存系統但儲存/存在於該儲存叢集/叢集式儲存陣列之複數個儲存系統內含之第二儲存系統上時，將一指示回應自該第一儲存系統傳送至該起始端系統。在示範實施例中，該指示回應可提供不是所有原始資料要求所要求之資料已被傳送之指示給該起始端，該指示回應可提供將該起始端系統指向該第二儲存系統之資訊，及/或該指示回應可指示/提供該叢集之一或更多其它儲存系統(例如該第二儲存系統 108)儲存著該資料中一部分/其餘部分之指標

給該起始端系統。例如，該指示回應可包含指向該要求資料之其餘部分(例如步驟 302 所接收之原始資料要求所要求資料之其餘部分)所在/所儲存叢集中之一或更多其它儲存系統/叢集節點(例如該第二儲存系統 108)之指示列表。

如上所述，在該起始端必須取得資料以滿足該原始資料要求所在之每一個額外叢集節點/儲存系統具有一指示。在本發明目前實施例中，該指示列表中之每一個指示可包含提供屬於該起始端之每一個儲存系統/節點下列資訊(如圖 4 所示)：一埠識別符(例如與內含該原始資料要求所要求資料之其餘部分中至少其中一些之叢集節點/儲存系統上之一埠相關)；一位移量(例如在它的相關資料儲存系統/儲存節點上之第一位元組資料之邏輯區塊位址)；及一資料長度(例如該指示要傳送之資料量)。該埠識別符可如該小型計算機系統介面架構模型規格書中所定義地遵守小型計算機系統介面相對埠識別符規定。完成一指示所需之其它資訊(例如磁區、邏輯單元、及目標端)係可由該小型計算機系統介面指示所產生之命令內容中取得。

在本發明示範實施例中，該方法 300 可進一步包含該第二儲存系統接收一第二命令之步驟 308。例如，回應於接收該指示列表，該起始端 102 可傳送該第二命令(例如，透過該儲存區域網路)至叢集內之其它儲存系統中一者，其中在該叢集係該指示列表中被辨識為儲存該資料其餘部分中至少一部分。例如，該起始端 102 可傳送該第二命令(例如，其係依據該指示回應)至一該指示列表所辨識之埠，該埠與

該第二系統有關。在進一步實施例中，該區塊儲存協定起始器 102 可使用該指示列表所示埠(例如，第二埠)來傳送各個命令至持有該原始要求所要求資料之叢集內之所有其它儲存系統。

在本發明進一步實施例中，該方法 300 進一步包含透過該儲存區域網路 310 將該要求資料中之儲存部分自該第二儲存系統傳送至該起始端系統之步驟 310。例如，如上所述之起始端 102 可使用該指示列表所示埠(例如，第二埠)來傳送命令至持有該原始要求所要求資料之叢集內之所有其它儲存系統(例如，儲存系統 108)，該些儲存系統將其本地資料連同用以指出已傳送其本地資料之一狀態資料一起回送至該起始端。在完成所有資料傳送以回應依據該些指示所發送之命令後，該區塊儲存協定可返回到它的呼叫程式以完成該操作。

在該原始資料要求(屬於步驟 302)所要求資料全部被該第一儲存系統所儲存並傳送之本發明替代性實施例中，該方法 300 可進一步包含將一傳送完成回應自該第一儲存系統傳送至該起始端系統之步驟 312，該傳送完成回應指示著該資料要求所要求資料全部被傳送。在進一步實施例中，當對應一命令之全部資料已被傳送時、或若不是該命令傳送就是該資料傳送發生一錯誤狀況時，該小型計算機系統介面目標端可藉由使包含一命令狀態之小型計算機系統介面回應回到/傳送至該起始端(方法步驟 314)而完成該操作。

為支援本發明，可能需要該儲存陣列叢集技術來提供

一些屬性。例如，該區塊儲存協定目標端可能需要被遍佈於該叢集中之所有儲存系統(106、108)。進一步，可能需要該叢集中之所有儲存系統上之所有埠具有獨一無二之埠識別符。該些埠識別符可能需要是該小型計算機系統介面架構模型規格書所定義之小型計算機系統介面相對埠識別符。再進一步，用於一虛擬磁區之邏輯區塊位址空間可被需要以讓存在於該虛擬磁區之所有儲存系統共用。此外，需要位在所有儲存系統(106、108)上之叢集區塊虛擬化函式(134)以決定該叢集內那個儲存系統可持有虛擬磁區(128、130)內那些位址範圍之資料。

如上所述，本發明方法可被配置於區塊儲存叢集內以提供儲存系統(106、108)上之區塊虛擬化。在示範實施例中，本發明系統/方法並非利用命令傳送或代理伺服器輸入/輸出，而是藉由利用內含一小型計算機系統介面檢查條件之狀態資訊和在小型計算機系統介面感測資料內之指示列表來完成它的本地資料傳送而配置指示著資料存在於其它叢集節點上之叢集區塊虛擬化(132、134)。

在進一步實施例中，該小型計算機系統介面起始端 102 可被架構以偵測一新的檢查條件、發送用於每一個指示之新小型計算機系統介面命令、及在完成所有指示時進行追蹤。該起始端 102 可進一步被架構以透過遍及多個起始端-目標端連結之指示來累積所取得之資料。

本發明系統/方法透過目前區塊儲存叢集方法提供一些優勢。第一，在該儲存基礎架構中不需額外硬體來支援叢

集進行。若是存在有主機代理者及儲存系統之區塊虛擬化例子，不必添加硬體至該儲存區域網路 110。進一步，目標端及邏輯單元之探索方式對於一起始端 102 而言並未改變，在該儲存叢集之所有節點上可看見該目標端，且該儲存叢集之所有節點被架構以決定那些邏輯單元可用至該目標端。再進一步，不必隱藏來自起始端之非虛擬磁區，且只有虛擬磁區自該叢集內之儲存系統中輸出。此外，該起始端不必保留關於該叢集內之資料分佈之資訊。本發明起始端/主機 102 被架構以決定資料存在於該叢集內的那裡。資料可由該叢集之任何節點上之任何埠所要求。該指示將該起始端指向持有該資料之叢集節點。進一步，儲存系統(106、108)上之資料可能被移動而沒有通知起始端，因為若是一起始端嘗試透過一錯誤叢集節點上之埠來存取資料，則該起始端只是簡單地重新導向(藉由該指示)內含該資料之叢集節點上之埠。再進一步，相對於在該儲存區域網路 110 中受限於所添加至該儲存區域網路之硬體容量之儲存虛擬化，本發明方法可施用任意數量之儲存裝置。此外，本發明方法可被施用至在一叢集節點上具有多個埠之儲存叢集。若是可透過多於一個路徑來存取資料，則該指示只需包含單一埠以經其存取該資料。將本發明配置於一標準儲存協定中之優勢在於沒有獨一無二之軟體需安裝於該些起始端系統上之區塊儲存協定中。

上述用以提供小型計算機系統介面輸入/輸出指示之系統/方法讓起始端可存取散佈於複數個目標端裝置之邏輯磁

碟區(LUN)上之資料。該些目標端裝置可以是磁片、儲存陣列、磁帶庫、或任何其它類型之儲存裝置。在本發明一進一步示範實施例中，我們可提供一系統/方法以讓起始端可透過多於一個目標端埠來存取一虛擬磁區中之一部分(例如，該部分係一資料段，該部分可由該目標端上可用之實體碟片、虛擬碟片、或任何其它資料段所組成)。例如，若存取至一資料段之目標端裝置具有多於一個目標端埠連接至該儲存區域網路(SAN)的情況可能會發生。在這類例子中，本發明系統/方法可被提供以讓目標端裝置(其可利用上述小型計算機系統介面輸入/輸出指示之方法)可通知起始端關於可用於那個資料段之多重路徑中每一者。

大體上參考圖 5，根據本發明一示範實施例所架構之一種用以透過小型計算機系統介面輸入/輸出指示來提供多重路徑之系統(例如拓樸)被顯示。該系統 500 可包含一起始端 502(例如一應用伺服器)。該起始端 502 可被架構以透過一儲存區域網路 504 連通耦接複數個目標端/目標端裝置/儲存裝置(506、508、510、512)。在本發明目前實施例中，該系統 500 可進一步包含複數個資料段(514、516、518、520)，其構成一虛擬磁區 522/為該虛擬磁區一部分/包含於該虛擬磁區內。該起始端 502 可被架構以透過該些目標端裝置之一或更多埠(圖 5 所示之埠 0 至埠 7)來存取一虛擬磁區 522 中之一部分(例如，該部分係該些資料段 514、516、518、520 中之一或更多)。例如，運用上述小型計算機系統介面輸入/輸出指示之方法之起始端 502 可具有多重路徑至構成

該虛擬磁區 522 之資料段中之每一個。

在本發明一示範實施例中，為通知該起始端 502 可透過多個目標端埠取得一資料段(514、516、518 或 520)，該些目標端裝置(506、508、510、512)可傳回多個小型計算機系統介面輸入/輸出指示以列出不同埠並聯結該些不同埠/指示著該些不同埠係聯結著同一資料段。該些資料段於該小型計算機系統介面指示列表中可因它們的資料位移及資料長度值而被獨一無二地辨識。例如該起始端 502 不是透過要求資料之目標端 506 之埠 0 就是埠 1 來接觸/傳送一輸入/輸出至該些目標端中其中之一(例如，目標端 506)，該資料係散佈於該些資料段(514、516、518、520)並可透過多個埠進行存取之。例如，該輸入/輸出可要求儲存/散佈於資料段 514、資料段 516、資料段 518、及資料段 520 中每一個之資料。目標端 506 接著可針對本地持有資料(例如，在資料段 514 上之資料)啟動資料傳送。進一步，目標端 506 接著可回送一內含下列指示之小型計算機系統介面輸入/輸出指示列表：

埠 2 識別符，資料段 516 資料位移量，資料段 516 資料長度

埠 3 識別符，資料段 516 資料位移量，資料段 516 資料長度

埠 4 識別符，資料段 518 資料位移量，資料段 518 資料長度

埠 5 識別符，資料段 518 資料位移量，資料段 518 資

料長度

埠 6 識別符，資料段 520 資料位移量，資料段 520 資料長度

埠 7 識別符，資料段 520 資料位移量，資料段 520 資料長度

在本發明目前實施例中，回應於接收該輸入/輸出指示列表，該起始端 502 接著可被架構以發送指示輸入/輸出至該小型計算機系統介面輸入/輸出指示列表所辨識資料段中每一個。例如：該起始端 502 可發送輸入/輸出指示至資料段 516、518、520 以取出該起始端 502 所送之原始輸入/輸出中所要求資料之剩餘資料。在本發明示範實施例中，該起始端 502 可選擇性決定透過那個埠來存取每一個資料段。因為上述本發明輸入/輸出指示多重路徑表列方法，該起始端 502 可依據該輸入/輸出指示列表來決定用以存取同一資料段之可行替代路線/多重路徑。例如依據上面小型計算機系統介面輸入/輸出指示列表，該起始端 502 不是選擇埠 2 就是選擇埠 3 以透過一指示輸入/輸出來存取資料段 516。進一步，若一給予指示輸入/輸出遇到問題，該起始端 502 可選擇在另一埠上再試著發送以存取相同資料段。例如若該起始端 502 透過埠 2 發送一指示輸入/輸出以存取/要求存取資料段 516 且該指示輸入/輸出遇到問題，該起始端可以埠 3 替代(例如，該起始端可在至一資料段之給予路徑變得不通順時切換至一替代路徑)來發送該指示輸入/輸出而

再試著發送該指示輸入/輸出。圖 6 顯示在該原始輸入/輸出被送至目標端 514 之埠 0 時由上示拓樸中一目標端所提供之示範小型計算機系統介面輸入/輸出指示列表，該原始輸入/輸出大小為四百(400)個區塊，且該些資料段(514、516、518、520)中之每一個持有該原始輸入/輸出要求所要求資料中之一百(100)個區塊。

上述輸入/輸出指示多重路徑方法/功能讓起始端可隨時運用通達一資料段之新/替代路徑。該功能可能效用為例如透過循環排序取得系統 500 之負載平衡。依據起始端行為，上述輸入/輸出指示多重路徑方法/功能讓目標端可依據小型計算機系統介面輸入/輸出指示列表之指示順序分配負載。這類順序可透過循環式排序、目前負載分佈等決定。

在本發明替代性實施例中，該小型計算機系統介面輸入/輸出指示列表並非如上所述配置隱含性網路分組而是明確將多個埠分至相同資料段。該明確分組可藉由添加額外結構至該小型計算機系統介面輸入/輸出指示列表而得。大體上參考至圖 7，一替代性示範小型計算機系統介面輸入/輸出指示列表(針對圖 6 所示彼等者)可由根據本發明一進一步示範實施例之上示拓樸中之目標端提供。該埠識別符遵守該小型計算機系統介面架構模型規格書所定義之小型計算機系統介面相對埠識別符之定義。在資料段可透過多個埠定期存取之架構中，圖 7 所示之明確埠分組/替代性小型計算機系統介面輸入/輸出指示列表可消除/排除每一個可用埠必須重複/重列該資料段位移量及資料段長度之需要

而促使效率增加。

在本發明進一步實施例中，讓目標端裝置可指定主要及替代路徑會是有利的，儘管也允許本發明包含進一步額外特性。大體上參考至圖 8，一替代性示範小型計算機系統介面輸入/輸出指示列表可由上示拓樸中之目標端提供之，該列表如上所述根據本發明一進一步示範實施例來配置主要及替代路徑之指定。該埠識別符可遵守該小型計算機系統介面架構模型規格書所定義之小型計算機系統介面相對埠識別符之定義。

大體上參考至圖 9，根據本發明一示範實施例之一種透過小型計算機系統介面輸入/輸出(SCSI I/O)指示於透過一網路來連通耦接之一起始端和一儲存叢集之間提供多重路徑之方法 900 被顯示。在示範實施例中，該方法 900 可透過上述系統 500 來實現。在本發明目前實施例中，該儲存叢集至少包含一第一目標端裝置及一第二目標端裝置。例如，該第一目標端裝置及/或該第二目標端裝置可為磁片、儲存陣列、磁帶庫、及/或儲存裝置。在本發明示範實施例中，該方法 900 包含該第一目標端裝置透過該網路接收一來自該起始端之輸入/輸出(I/O)之步驟 902。例如，該輸入/輸出可包含對一第一部分資料及對一第二部分資料之要求(例如，讀取要求)。在本發明目前實施例中，該第一部分資料係位在一第一資料段上，且該第二部分資料係位在一第二資料段上。進一步，該第一部分資料及該第二部分資料係包含一虛擬磁區內。

在進一步實施例中，當不是該第二部分資料而是該第一部分資料被儲存於該第一目標端裝置上，且該第二部分資料被儲存於該第二目標端裝置上時，該方法 900 可進一步包含：啟動該第一部分資料至該起始端之傳送之步驟 904；及將一小型計算機系統介面輸入/輸出指示列表傳送至該起始端之步驟 906。在示範實施例中，該指示列表可包含用以辨識該第二目標端裝置之第一埠之第一埠識別符及用以辨識該第二目標端裝置之第二埠之第二埠識別符。該第一埠及該第二埠於該指示列表中可被辨識為用以存取該第二部分資料之存取埠(例如，該起始端可透過該第二目標端裝置之存取埠來存取該第二部分資料)。在額外實施例中，該指示列表可藉由該第二資料段之資料位移值及/或該第二資料段之資料長度值來辨識該第二資料段。進一步，該指示列表可聯結該第一埠識別符及該第二埠識別符與該第二資料段、該第二資料段之資料位移值、及/或該第二資料段之資料長度。該第一埠識別符及該第二埠識別符遵守該小型計算機系統介面架構模型規格書所定義之小型計算機系統介面相對埠識別符定義。

在額外實施例中，該方法 900 可進一步包含該第二目標端裝置透過該網路接收來自該起始端之一指示輸入/輸出之步驟 908。例如，該指示輸入/輸出係回應於該小型計算機系統介面輸入/輸出指示列表。同時，該指示輸入/輸出可要求透過該起始端所示之第一埠及第二埠中一者來存取該第二部分資料。在又一實施例中，該方法 900 更可包含啟

動該第二部分資料至該起始端之傳送之步驟 910。

大體上參考至圖 10，根據本發明一示範實施例之一種透過小型計算機系統介面輸入/輸出(SCSI I/O)指示於透過一網路來連通耦接之一起始端和一儲存叢集之間提供多重路徑之方法 1000 被顯示。在示範實施例中，該方法 1000 可透過上述系統 500 來實現。該儲存叢集至少包含一第一目標端裝置及一第二目標端裝置。該方法 1000 可包含該第一目標端裝置透過該網路(例如，該輸入/輸出包含一資料要求)接收來自該起始端之輸入/輸出(I/O)之步驟 1002。該方法 1000 可進一步包含當包含於該資料要求中之資料並未儲存於該第一目標端裝置上，但是該資料儲存於該第二目標端裝置上時，將一小型計算機系統介面輸入/輸出指示列表傳送至該起始端之步驟 1004。該指示列表包含用以辨識該第二目標端裝置之第一埠之第一埠識別符及用以辨識該第二目標端裝置之第二埠之第二埠識別符，該第一埠及該第二埠被辨識為用以存取該資料之存取埠。該第一埠識別符及該第二埠識別符遵守該小型計算機系統介面架構模型規格書所定義之小型計算機系統介面相對埠識別符定義。該方法 1000 可進一步包含該第二目標端裝置透過該網路接收一來自該起始端之指示輸入/輸出之步驟 1006。該指示輸入/輸出可回應於該小型計算機系統介面輸入/輸出指示列表，且該指示輸入/輸出可要求透過該起始端所示之第一埠或第二埠來存取資料。該方法 1000 可進一步包含啟動該資料至該起始端之傳送之步驟 1008。

大體上參考圖 11，根據本發明一示範實施例之一種用以透過過小型計算機系統介面輸入/輸出指示來提供多重路徑所架構之系統(例如拓樸)被顯示。該系統 1100 可包含一起始端 1102(例如一應用伺服器)。該起始端 1102 可被架構以透過儲存區域網路(1104、1106)連通耦接複數個目標端/目標端裝置/儲存裝置(1108、1110)。如圖 11 所示，儲存區域網路 1104(光纖通道)及 1106(網際網路式小型計算機系統介面)可由不同媒體傳輸層協定來實現。在本發明目前實施例中，該系統 1100 可進一步包含複數個資料段(1112、1114)，其構成一虛擬磁區 1122/為該虛擬磁區一部分/包含於該虛擬磁區內。該些目標端裝置、資料段、及虛擬磁區可全部包括於一儲存叢集 1116。該起始端 1102 可被架構以透過該些目標端裝置之一或更多埠(圖 11 所示埠 1 至埠 4)來存取一虛擬磁區 1122 中一部分，該部分係該些資料段(1112、1114)中之一或更多。例如，運用上述小型計算機系統介面輸入/輸出指示方法之起始端 1102 可具有多重路徑至構成該虛擬磁區 1122 之資料段中之每一個。

在本發明一示範實施例中，為了通知該起始端 1102 可透過多個目標端埠取得一資料段(1112、1114)，該些目標端裝置(1108、1110)可傳回多個小型計算機系統介面輸入/輸出指示以列出不同埠並聯結該些不同埠/指示著該些不同埠係聯結著同一資料段。該些資料段於該小型計算機系統介面指示列表中可因它們的邏輯區塊位址(LBA)及資料長度值而被獨一無二地辨識。例如，該起始端 1102 不是透過要

求資料之目標端 1110 之埠 3 就是埠 4 來接觸/傳送一輸入/輸出至該些目標端中其中之一(例如，目標端 1110)，該資料係散佈於該些資料段(1112、1114)並可透過多個埠進行存取。例如，該輸入/輸出可要求儲存/散佈於資料段 1112 及資料段 1114 中每一個之資料：

埠 4，邏輯區塊位址 0，資料長度 200

目標端 1110 接著可針對本地持有資料(如資料段 1114 之資料)啟動資料傳送。此外，所提供資料段 1112 及 1114 長度係各 100 個區塊，目標端 1110 接著可回送含一資料項之下列指示之小型計算機系統介面輸入/輸出指示列表：

邏輯區塊位址 0，資料長度 100，埠 1 相對埠識別符，埠 2 相對埠識別符

在本發明目前實施例中，回應於接收該輸入/輸出指示列表，該起始端 1102 接著可被架構以發送指示輸入/輸出至該小型計算機系統介面輸入/輸出指示列表所辨識資料段中之每一個。例如，該起始端 1102 可發送輸入/輸出指示至資料段 1112 及 114 以取出該起始端 1102 所傳送之原始輸入/輸出中所要求資料之剩餘資料。在本發明示範實施例中，該起始端 1102 可選擇性決定透過那個埠來存取每一個資料段。因為上述本發明輸入/輸出指示多重路徑表列方法，該起始端 1102 可依據該輸入/輸出指示列表來決定用以存取同一資料段之可行替代路線/多重路徑。例如，依據上面小

型計算機系統介面輸入/輸出指示列表，該起始端 1102 不是選擇埠 1 就是埠 2 以透過一指示輸入/輸出來存取資料段 1112。進一步，若給予指示輸入/輸出遇到問題，該起始端 1102可選擇在另一埠上再試著發送該給予指示輸入/輸出以存取相同資料段。例如，若該起始端 1102 透過埠 2 發送一指示輸入/輸出以存取/要求存取資料段 1112 且該指示輸入/輸出遇到問題，該起始端可以埠 1 替代(例如，該起始端可在至一資料段之給予路徑變得不通順時切換至一替代路徑)來發送該指示輸入/輸出而再試著發送該指示輸入/輸出。

上述輸入/輸出指示方法/功能讓多媒體傳輸協定可同時存在於一小型計算機系統介面指示儲存系統。該方法/功能讓一小型計算機系統介面指示儲存系統之埠識別符係一固定尺寸。此外，用於不同媒體傳輸協定之小型計算機系統介面埠之埠識別符可同時存在於本方法/功能所發送之指示/指示列表中。進一步，發送至一媒體傳輸協定埠之一輸入/輸出可被重新導向至一第二媒體傳輸協定之埠。例如，指向一串接式小型計算機系統介面埠之一輸入/輸出可被重新導向至一光纖通道埠。

注意，根據本發明之前述實施例可使用根據本說明書教示程式化之傳統通用數位電腦來便利地配置之，對那些熟知電腦技術之人士會是顯而易見的。適當的軟體程式碼可由熟知程式之程式設計師依據本揭示之教導來輕易地備製之，對那些熟知軟體技術之人士會是顯而易見的。

要解本發明可以一軟體套件形式來便利地實現。這類

軟體套件可以是一電腦程式產品，其運用包含用以程式化一電腦來執行本發明所示功能及方法而儲存之電腦程式碼之電腦可讀取儲存媒體。該電腦可讀取媒體包含任何類型之傳統軟碟片、光學碟片、唯讀式光碟片、磁性碟片、硬碟機、磁性-光學碟片、唯讀記憶體、隨機存取記憶體、可拭可程式唯讀記憶體、電性可拭可程式唯讀記憶體、磁性或光學卡片、或用以儲存電子指令之任何其它合適媒體。

了解到所示方法中之特定次序或層級係為示範方法之範例。依據設計偏好了解到本方法步驟之特定次序或層級可被重新安排而維持在本發明範圍內。所附方法項申請專利範圍以一實例次序呈現各步驟構件，並不必然意謂著受限於所示特定次序或層級。

相信本揭示及其伴隨許多優勢能經由前述說明了解。也相信該些元件在形式、建構及安排的各種改變可被進行而不偏離所示主題或不犧牲其所有重要優勢係明顯。所述形式僅是說明且下列申請專利範圍企圖納入這類改變。

【圖式簡單說明】

那些熟知此項技術之人士可藉由參考該些附圖而對本發明許多優勢有較佳了解，其中：

圖 1 係根據本發明一示範實施例可透過一區塊儲存協定進行存取之網路型儲存配置/系統之一方塊圖。

圖 2 係根據本發明系統/方法之示範實施例所配置具有指示之小型計算機系統介面命令/回應遠端程序呼叫之一方

塊示意圖。

圖 3 係根據本發明一示範實施例說明一種用於一起始端系統及一叢集型儲存陣列間之通訊方法之流程圖。

圖 4 係根據本發明一示範實施例之一種小型計算機系統介面指示格式圖。

圖 5 係根據本發明一進一步示範實施例具有多重路徑至一虛擬磁區各部分之網路型儲存配置之一方塊圖拓樸。

圖 6 係根據本發明一進一步示範實施例具有多重路徑之小型計算機系統介面指示的一格式圖。

圖 7 係根據本發明一替代性示範實施例具有多重路徑之小型計算機系統介面指示的一格式圖。

圖 8 係根據本發明一進一步示範性實施例具有優先權化多重路徑之小型計算機系統介面指示的一格式圖。

圖 9 係根據本發明一示範實施例說明透過小型計算機系統介面(SCSI)輸入/輸出指示於透過一網路來連通耦接之起始端和儲存叢集之間提供多重路徑之方法流程圖。

圖 10 係根據本發明一替代示範實施例說明透過小型計算機系統介面(SCSI)輸入/輸出指示於透過一網路來連通耦接之起始端和儲存叢集之間提供多重路徑之方法流程圖。

圖 11 係根據本發明一進一步示範實施例具有自多個傳輸協定至虛擬磁區各部分之多重路徑之網路型儲存配置之一方塊圖拓樸。

【主要元件符號說明】

100~網路式儲存配置/系統/儲存基礎架構

102~應用系統/伺服器

104~應用程式

106/108~儲存系統

110~儲存區域網路

112~小型計算機系統介面協定堆疊

114~作業系統

116~儲存區域網路轉換器

118/120~本地區塊虛擬化層

122/124~內部區塊協定堆疊

126~實體儲存裝置

128/130~虛擬磁區

132~叢集區塊虛擬化層

134~叢集虛擬化管理器函式

500~系統

502~起始端

504~儲存區域網路

506/508/510/512~目標端/目標端裝置/儲存裝置

514/516/518/520~資料段

522~虛擬磁區

1100~系統

1102~起始端

1104/1106~儲存區域網路

1108/1110~目標端/目標端裝置/儲存裝置

1112/1114~資料段

1116~儲存叢集

1122~虛擬磁區

七、申請專利範圍：

1.一種透過小型計算機系統介面輸入/輸出(SCSI I/O)指示於透過一網路來連通耦接之一起始端和一儲存叢集之間提供多重路徑之方法，該儲存叢集至少包含一第一目標端裝置及一第二目標端裝置，該方法包括：

在該起始端決定資料先前係儲存於該儲存叢集之該第一目標端裝置後，透過該網路將來自該起始端之一輸入/輸出(I/O)要求指向該第一目標端裝置；以及

當不是該資料之一第二部分而是該資料之一第一部分目前被儲存於該第一目標端裝置上、且該資料之該第二部分目前被儲存於該第二目標端裝置上時，藉由該儲存叢集啟動該資料之該第一部分至該起始端之一傳送並將一小型計算機系統介面輸入/輸出指示列表傳送至該起始端，

其中該指示列表包含用以辨識該第二目標端裝置之一第一埠之一第一埠識別符、及用以辨識該第二目標端裝置之一第二埠之一第二埠識別符，該第一埠及該第二埠被辨識為使該起始端存取該資料之該第二部分之存取埠。

2.如申請專利範圍第1項之方法，進一步包括：

該第二目標端裝置透過該網路接收來自該起始端之一指示輸入/輸出，該指示輸入/輸出係回應於該小型計算機系統介面輸入/輸出指示列表，該指示輸入/輸出要求透過下列中一者來存取該資料之該第二部分：該起始端所示之第一埠及第二埠。

3.如申請專利範圍第2項之方法，進一步包括：

啟動該資料之該第二部分至該起始端之一傳送。

4.如申請專利範圍第3項之方法，其中該資料之該第一部分係位在一第一資料段上，且該資料之該第二部分係位在一第二資料段上。

5.如申請專利範圍第4項之方法，其中該第一資料段及該第二資料段係包含於一虛擬磁區內。

6.如申請專利範圍第5項之方法，其中該第一目標端裝置及該第二目標端裝置係下列中一者：碟片、儲存陣列、磁帶庫、及儲存裝置。

7.如申請專利範圍第5項之方法，其中該指示列表藉由該第二資料段之一資料位移值來辨識該第二資料段。

8.如申請專利範圍第7項之方法，其中該指示列表藉由該第二資料段之一資料長度值來辨識該第二資料段。

9.如申請專利範圍第8項之方法，其中該指示列表聯結該第一埠識別符及該第二埠識別符與該第二資料段、該第二資料段之資料位移值、和該第二資料段之資料長度值。

10.一種具有電腦可執行指令以執行用以透過小型計算機系統介面輸入/輸出(SCSI I/O)指示於透過一網路來連通耦接之一起始端和一儲存叢集之間提供多重路徑之方法之電腦可讀取媒體，該儲存叢集至少包含一第一目標端裝置及一第二目標端裝置，該方法包括：

在該起始端決定資料先前係儲存於該儲存叢集之該第一目標端裝置後，透過該網路將來自該起始端之一輸入/輸出(I/O)要求指向該第一目標端裝置；以及

當不是該資料之一第二部分而是該資料之一第一部分目前被儲存於該第一目標端裝置上、且該資料之該第二部分目前被儲存於該第二目標端裝置上時，藉由該儲存叢集啟動該資料之該第一部分至該起始端之一傳送並將一小型計算機系統介面輸入/輸出指示列表傳送至該起始端，

其中該指示列表包含用以辨識該第二目標端裝置之一第一埠之一第一埠識別符、及用以辨識該第二目標端裝置之一第二埠之一第二埠識別符，該第一埠及該第二埠被辨識為使該起始端存取該資料之該第二部分之存取埠。

11.如申請專利範圍第 10 項之電腦可讀取媒體，該方法進一步包括：

該第二目標端裝置透過該網路接收來自該起始端之一指示輸入/輸出，該指示輸入/輸出係回應於該小型計算機系統介面輸入/輸出指示列表，該指示輸入/輸出要求透過下列中一者來存取該資料之該第二部分：該起始端所示之第一埠及第二埠。

12.如申請專利範圍第 11 項之電腦可讀取媒體，該方法進一步包括：

啟動該資料之該第二部分至該起始端之一傳送。

13.如申請專利範圍第 12 項之電腦可讀取媒體，其中該資料之該第一部分係位在一第一資料段上，且該資料之該第二部分係位在一第二資料段上。

14.如申請專利範圍第 13 項之電腦可讀取媒體，其中該第一資料段及該第二資料段係包含於一虛擬磁區內。

15.如申請專利範圍第14項之電腦可讀取媒體，其中該第一目標端裝置及該第二目標端裝置係下列中一者：碟片、儲存陣列、磁帶庫、及儲存裝置。

16.如申請專利範圍第14項之電腦可讀取媒體，其中該指示列表藉由該第二資料段之一資料位移值來辨識該第二資料段。

17.如申請專利範圍第16項之電腦可讀取媒體，其中該指示列表藉由該第二資料段之一資料長度值來辨識該第二資料段。

18.如申請專利範圍第17項之電腦可讀取媒體，其中該指示列表聯結該第一埠識別符及該第二埠識別符與該第二資料段、該第二資料段之資料位移值、和該第二資料段之資料長度值。

19.一種透過小型計算機系統介面輸入/輸出(SCSI I/O)指示於透過一網路來連通耦接之一起始端和一儲存叢集之間提供多重路徑之系統，該儲存叢集至少包含一第一目標端裝置及一第二目標端裝置，該系統包括：

用於在該起始端決定資料先前係儲存於該儲存叢集之該第一目標端裝置後透過該網路將來自該起始端之一輸入/輸出(I/O)要求指向該第一目標端裝置的構件；

當不是該資料之一第二部分而是該資料之一第一部分目前被儲存於該第一目標端裝置上、且資料之該該第二部分目前被儲存於該第二目標端裝置上時，藉由該儲存叢集啟動該資料之該第一部分至該起始端之一傳送及將一小型

計算機系統介面輸入/輸出指示列表傳送至該起始端，

其中該指示列表包含用以辨識該第二目標端裝置之一第一埠之一第一埠識別符、及用以辨識該第二目標端裝置之一第二埠之一第二埠識別符，該第一埠及該第二埠被辨識為使該起始端存取該資料之該第二部分之存取埠。

20.一種具有電腦可執行指令以執行用以透過小型計算機系統介面輸入/輸出(SCSI I/O)指示於透過一網路來連通耦接之一起始端和一儲存叢集之間提供多重路徑之方法之電腦可讀取媒體，該儲存叢集至少包含一第一目標端裝置及一第二目標端裝置，該方法包括：

在該起始端決定資料先前係儲存於該儲存叢集之該第一目標端裝置後，透過該網路將來自該起始端之一輸入/輸出(I/O)要求指向該第一目標端裝置；

當包含於該資料要求中之該資料目前並未儲存於該第一目標端裝置上、但是該資料目前儲存於該第二目標端裝置上時，藉由該儲存叢集將一小型計算機系統介面輸入/輸出指示列表傳送至該起始端，

其中該指示列表包含用以辨識該第二目標端裝置之一第一埠之一第一埠識別符、及用以辨識該第二目標端裝置之一第二埠之一第二埠識別符，該第一埠及該第二埠被辨識為使該起始端存取該資料之存取埠；

於該第二目標端裝置處透過該網路接收來自該起始端之一指示輸入/輸出，該指示輸入/輸出係回應於該小型計算機系統介面輸入/輸出指示列表，且該指示輸入/輸出要求透

過下列一者來存取資料：該起始端所示之第一埠及第二埠；以及

啟動該資料至該起始端之一傳送。

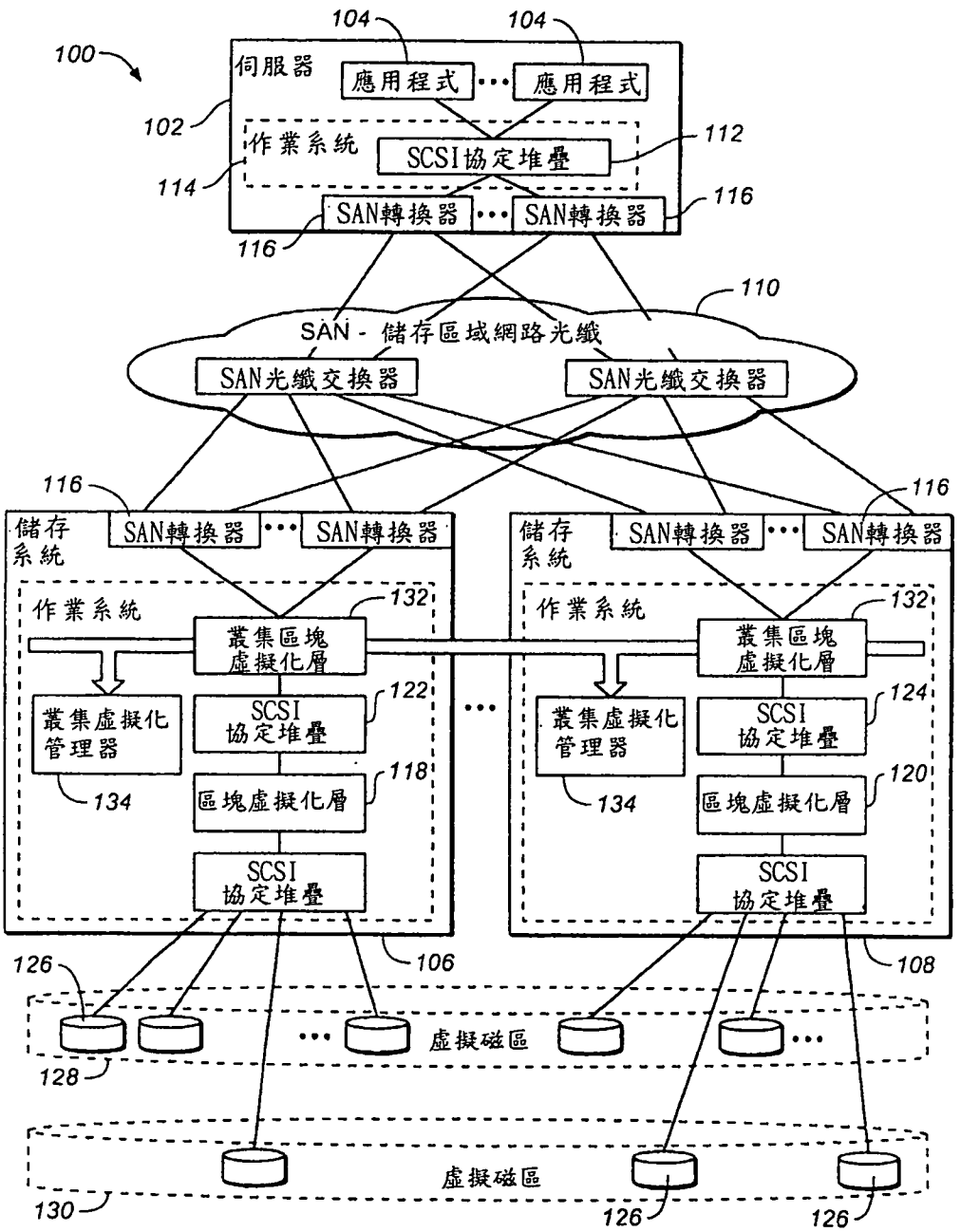
100
伺服器
104
104
應用程式
應用程式
102
作業系統
SCSI協定堆疊
112
114
SAN轉換器
SAN轉換器
116
116
110
SAN - 儲存區域網路光纖
SAN光纖交換器
SAN光纖交換器
116
116
儲存系統
SAN轉換器
SAN轉換器
儲存系統
SAN轉換器
SAN轉換器
作業系統
132
叢集區塊虛擬化層
叢集虛擬化管理器
134
SCSI協定堆疊
122
區塊虛擬化層
118
SCSI協定堆疊
作業系統
132
叢集區塊虛擬化層
叢集虛擬化管理器
134
SCSI協定堆疊
124
區塊虛擬化層
120
SCSI協定堆疊
106
108
126
虛擬磁區
128
虛擬磁區
130
126
126

圖1

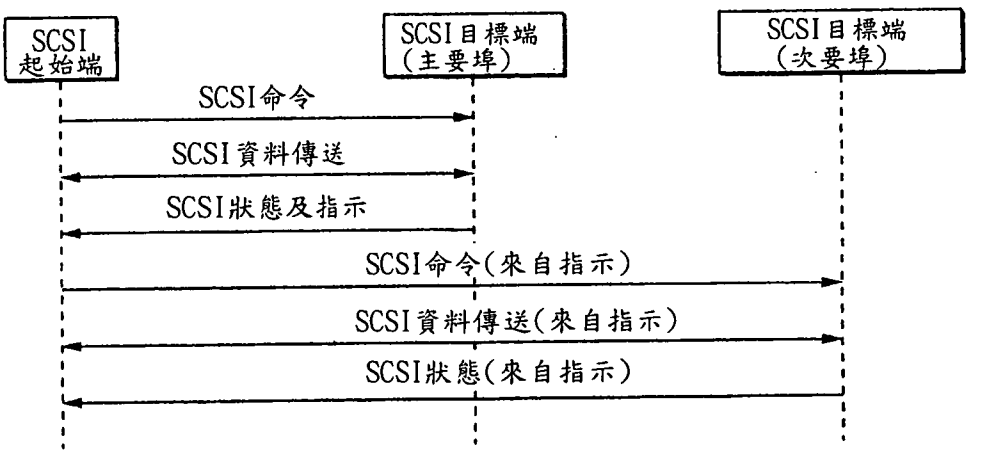
SCSI
起始端
SCSI目標端
(主要埠)
SCSI目標端
(次要埠)
SCSI命令
SCSI資料傳送
SCSI狀態及指示
SCSI命令(來自指示)
SCSI資料傳送(來自指示)
SCSI狀態(來自指示)

圖2

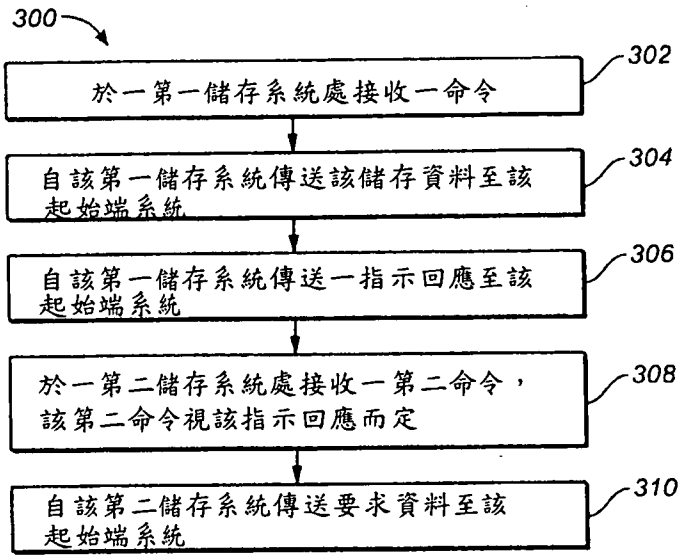
300
於一第一儲存系統處接收一命令
302
自該第一儲存系統傳送該儲存資料至該起始端系統
304
自該第一儲存系統傳送一指示回應至該起始端系統
306
於一第二儲存系統處接收一第二命令，該第二命令視該指示回應而定
308
自該第二儲存系統傳送要求資料至該起始端系統
310

圖3

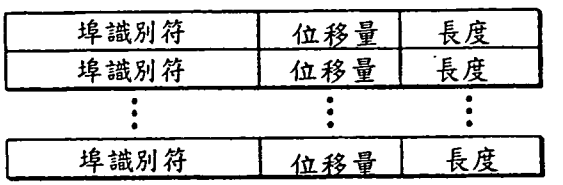
埠識別符
位移量
長度
埠識別符
位移量
長度
埠識別符
位移量
長度

圖4

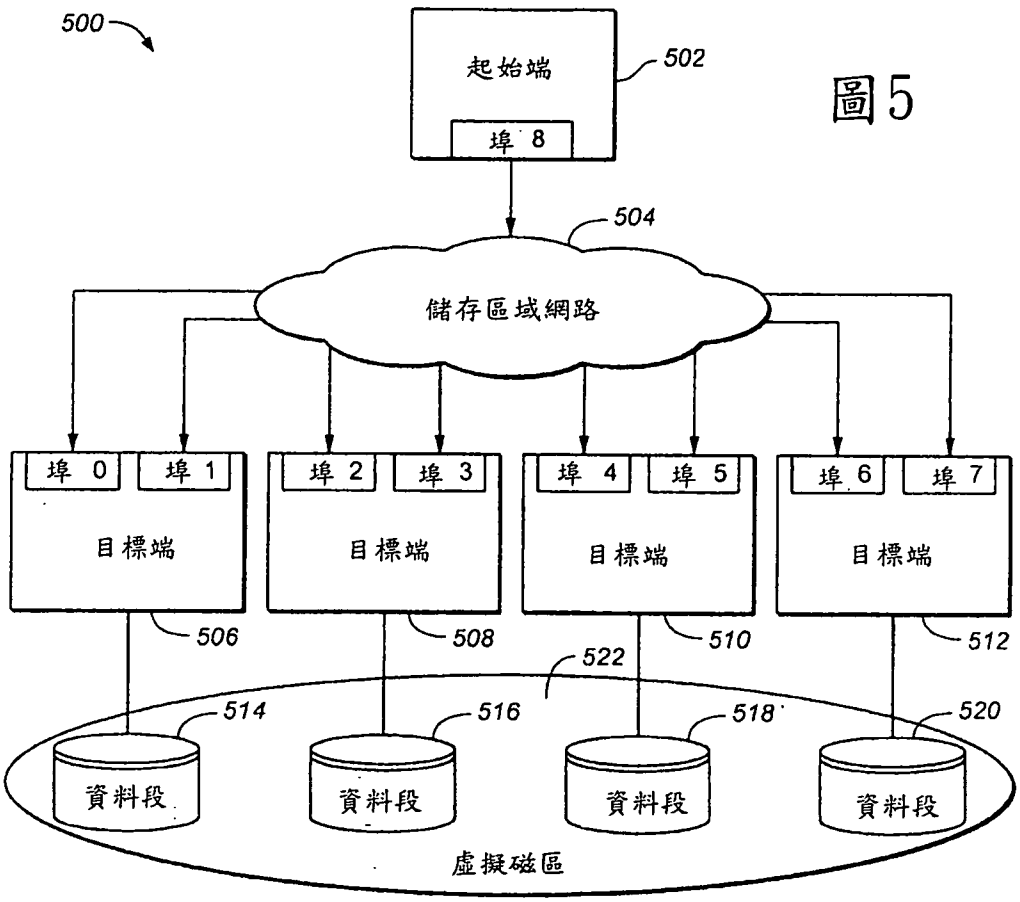
500
起始端
502
埠 8
504
儲存區域網路
埠 0
埠 1
埠 2
埠 3
埠 4
埠 5
埠 6
埠 7
目標端
目標端
目標端
目標端
506
508
510
512
522
514
516
518
520
資料段
資料段
資料段
資料段
虛擬磁區

圖5

埠：2 位移量：100 長度：100	埠：3 位移量：100 長度：100	埠：4 位移量：200 長度：100	埠：5 位移量：200 長度：100	埠：6 位移量：300 長度：100	埠：7 位移量：300 長度：100

圖6

位移量：100 長度：100		位移量：200 長度：100		位移量：300 長度：100	
埠：2	埠：3	埠：4	埠：5	埠：6	埠：7

圖7

位移量：100 長度：100		位移量：200 長度：100		位移量：300 長度：100	
主要埠：2	替代埠：3	主要埠：4	替代埠：5	主要埠：6	替代埠：7

圖8

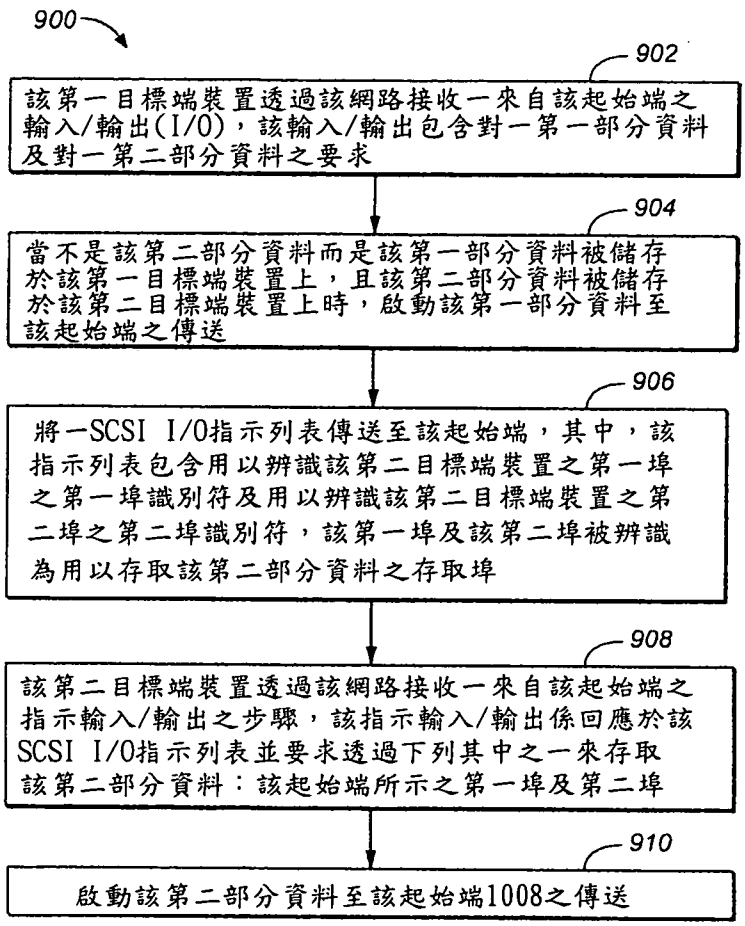
900
902
該第一目標端裝置透過該網路接收一來自該起始端之輸入/輸出(I/O)，該輸入/輸出包含對一第一部分資料及對一第二部分資料之要求
904
當不是該第二部分資料而是該第一部分資料被儲存於該第一目標端裝置上，且該第二部分資料被儲存於該第二目標端裝置上時，啟動該第一部分資料至該起始端之傳送
906
將一SCSI I/O指示列表傳送至該起始端，其中，該指示列表包含用以辨識該第二目標端裝置之第一埠之第一埠識別符及用以辨識該第二目標端裝置之第二埠之第二埠識別符，該第一埠及該第二埠被辨識為用以存取該第二部分資料之存取埠
908
該第二目標端裝置透過該網路接收一來自該起始端之指示輸入/輸出之步驟，該指示輸入/輸出係回應於該SCSI I/O指示列表並要求透過下列其中之一來存取該第二部分資料：該起始端所示之第一埠及第二埠
910
啟動該第二部分資料至該起始端1008之傳送

圖9

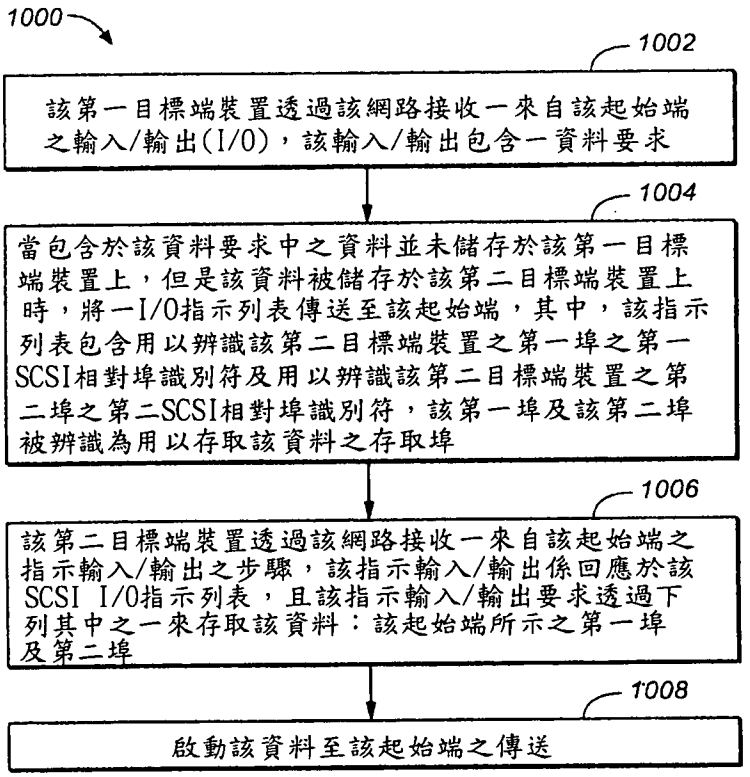
1000
1002
該第一目標端裝置透過該網路接收一來自該起始端之輸入/輸出(I/O)，該輸入/輸出包含一資料要求
1004
當包含於該資料要求中之資料並未儲存於該第一目標端裝置上，但是該資料被儲存於該第二目標端裝置上時，將一I/O指示列表傳送至該起始端，其中，該指示列表包含用以辨識該第二目標端裝置之第一埠之第一SCSI相對埠識別符及用以辨識該第二目標端裝置之第二埠之第二SCSI相對埠識別符，該第一埠及該第二埠被辨識為用以存取該資料之存取埠
1006
該第二目標端裝置透過該網路接收一來自該起始端之指示輸入/輸出之步驟，該指示輸入/輸出係回應於該SCSI I/O指示列表，且該指示輸入/輸出要求透過下列其中之一來存取該資料：該起始端所示之第一埠及第二埠
1008
啟動該資料至該起始端之傳送

圖10

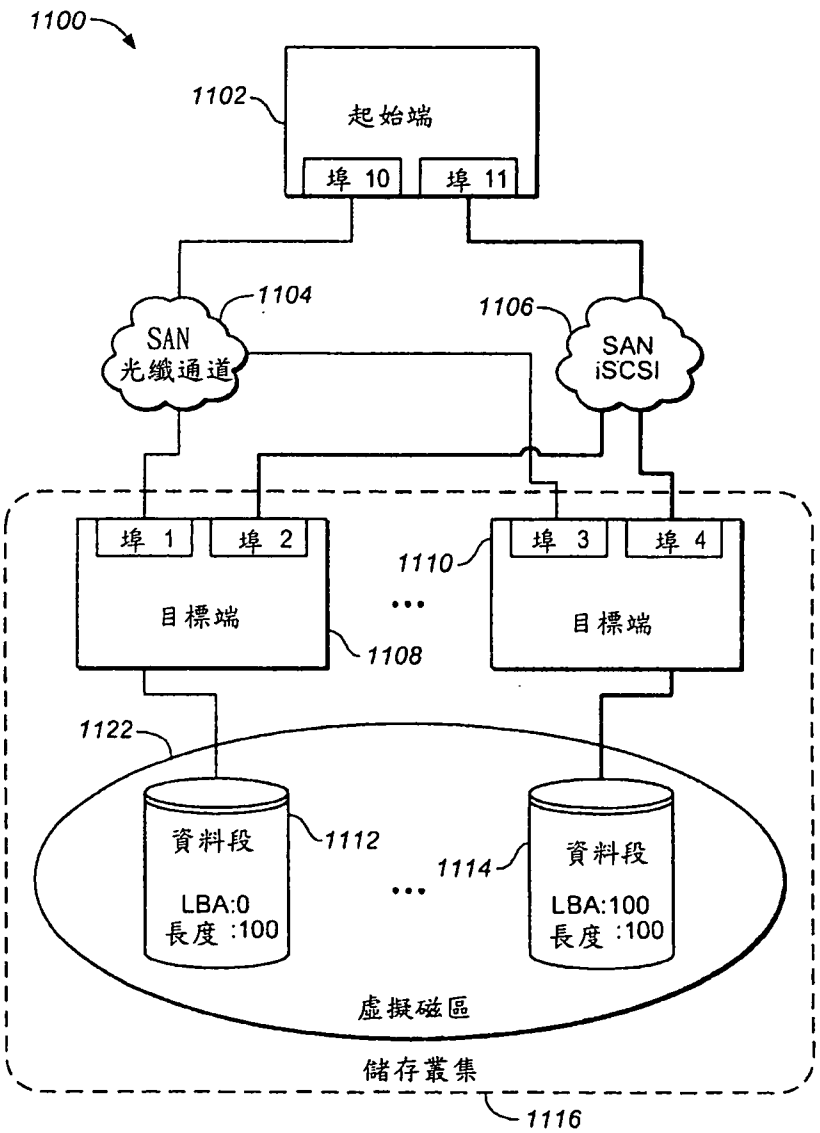
1100
1102
起始端
埠 10
埠 11
1104
SAN
光纖通道
1106
SAN
iSCSI
埠 1
埠 2
埠 3
埠 4
1110
目標端
目標端
1108
1122
資料段
LBA:0
長度 :100
1112
1114
資料段
LBA:100
長度 :100
虛擬磁區
儲存叢集
1116

圖11