I737395

(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 I737395 B**

(45)公告日： 中華民國 110 (2021) 年 08 月 21 日

(21)申請案號：109123042 (22)申請日： 中華民國 109 (2020) 年 07 月 08 日

(51)Int. Cl.： ***G06F9/312 (2006.01)*** ***G06F12/0877(2016.01)***

(30)優先權：2019/09/12 世界智慧財產權組織 PCT/CN2019/105753

(71)申請人：開曼群島商創新先進技術有限公司 (開曼群島) ADVANCED NEW TECHNOLOGIES CO., LTD. (KY)
開曼群島

(72)發明人：田世坤 (CN)

(74)代理人：林志剛

(56)參考文獻：

TW	201812674A	TW	201935299A
CN	109144414A	CN	109375868A
CN	109857337A	US	5926834
US	9946657B1	US	2017/0124077A1

審查人員：吳兆平

申請專利範圍項數：21 項 圖式數：15 共 114 頁

(54)名稱

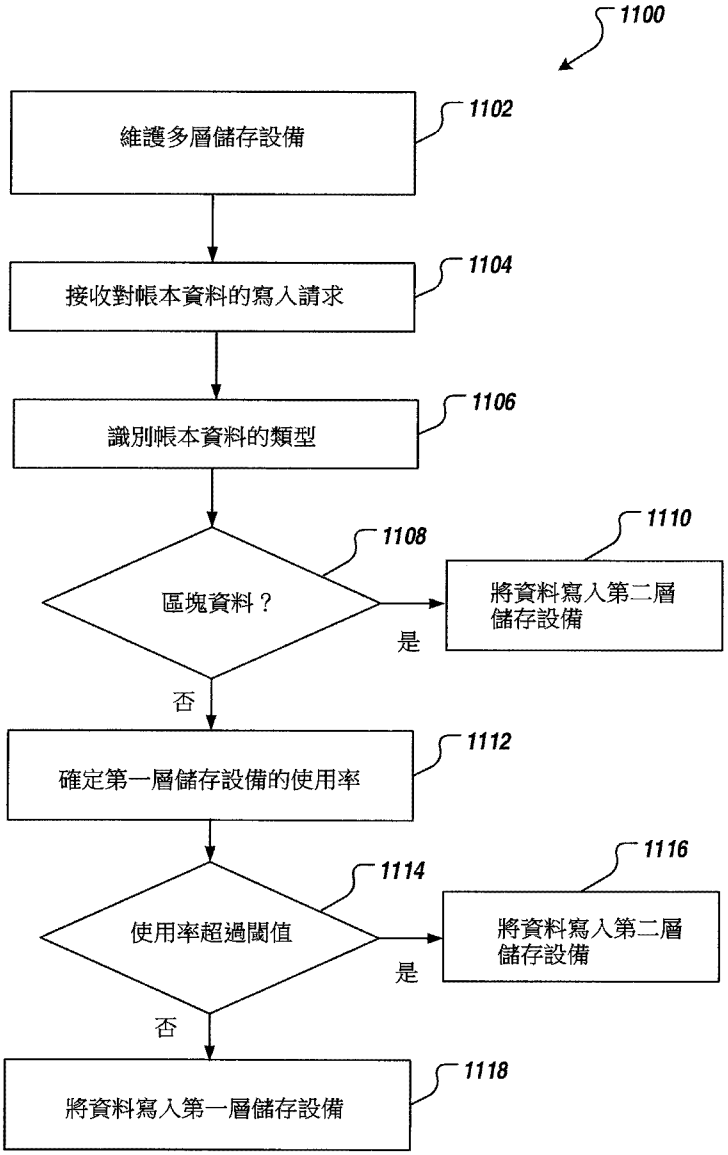
日誌結構儲存系統及方法

(57)摘要

本文揭露了用於資料處理的方法、系統和裝置，包括編碼在電腦儲存設備上的電腦程式。方法之一包括：儲存系統維護多個儲存設備，所述多個儲存設備至少包括第一層儲存設備和第二層儲存設備。所述儲存系統接收針對帳本資料的寫入請求，確定所述帳本資料的類型是否是區塊資料，響應於確定所述帳本資料的所述類型是區塊資料，將所述資料寫入所述第二層儲存設備。

Disclosed herein are methods, systems, and apparatus, including computer programs encoded on computer storage devices, for data processing. One of the methods includes maintaining, by a storage system, a plurality of storage devices that include at least a first tier storage device and a second tier storage device. The storage system receives a write request of a ledger data, determines whether a type of the ledger data is block data, and, in response to determining that the type of the ledger data is block data, writes the data into the second tier storage device.

指定代表圖：


1100
維護多層儲存設備
1102
接收對帳本資料的寫入請求
1104
識別帳本資料的類型
1106
區塊資料？
1108
是
將資料寫入第二層儲存設備
1110
否
確定第一層儲存設備的使用率
1112
使用率超過閾值
1114
是
將資料寫入第二層儲存設備
1116
否
將資料寫入第一層儲存設備
1118

【圖 11】

專利案號：109123042

公告本

民國 110 年 6 月 18 日修正

I737395

【發明摘要】

【中文發明名稱】

日誌結構儲存系統及方法

【英文發明名稱】

LOG-STRUCTURED STORAGE SYSTEMS AND METHOD

【中文】

本文揭露了用於資料處理的方法、系統和裝置，包括編碼在電腦儲存設備上的電腦程式。方法之一包括：儲存系統維護多個儲存設備，所述多個儲存設備至少包括第一層儲存設備和第二層儲存設備。所述儲存系統接收針對帳本資料的寫入請求，確定所述帳本資料的類型是否是區塊資料，響應於確定所述帳本資料的所述類型是區塊資料，將所述資料寫入所述第二層儲存設備。

【英文】

Disclosed herein are methods, systems, and apparatus, including computer programs encoded on computer storage devices, for data processing. One of the methods includes maintaining, by a storage system, a plurality of storage devices that include at least a first tier storage device and a second tier storage device. The storage system receives a write request of a ledger data, determines whether a type of the ledger data is block data, and, in response to determining that the type of the ledger data is block data, writes the data into the second tier storage device.

【指定代表圖】第(11)圖。
【代表圖之符號簡單說明】無

【特徵化學式】無

第 109123042 號　　民國 110 年 6 月 18 日修正

【發明說明書】

【中文發明名稱】

日誌結構儲存系統及方法

【英文發明名稱】

LOG-STRUCTURED STORAGE SYSTEMS AND METHOD

【技術領域】

本文涉及日誌結構儲存系統。

【先前技術】

分散式帳本系統(DLS)，也可稱為共識網路及/或區塊鏈網路，使得參與的實體能夠安全且不可變地儲存資料。在不引用任何特定用例的情況下，DLS通常被稱為區塊鏈網路。區塊鏈網路類型的示例可以包括公共區塊鏈網路、私有區塊鏈網路和聯盟區塊鏈網路。為選定的實體群組提供聯盟區塊鏈網路，所述實體控制共識處理，並且所述聯盟區塊鏈網路包括存取控制層。

通常，DLS的每個節點(例如，區塊鏈網路節點)儲存或具有區塊鏈網路資料的完整備份，使得每個節點可以是獨立的，並且每個節點處的本地資料可以被信任以提供服務。然而，這種儲存方案提出了苛刻的儲存要求，並增加了每個節點的儲存成本，尤其是隨著DLS達到規模。因此，期望用於提高效率並降低儲存系統成本的解決方案。

【發明內容】

本文描述了用於將資料儲存在例如分散式帳本系統(例如，區塊鏈網路)及/或基於區塊鏈的中心化帳本系統(例如，通用可審計帳本服務系統)中的日誌結構儲存系統的技術，所述分散式帳本系統及/或基於區塊鏈的中心化帳本系統採用區塊鏈的資料結構以利用儲存在區塊鏈上的資料的不可變性、可靠性以及可信性。

本文還提供了耦接到一個或多個處理器並且其上儲存有指令的一個或多個非暫態電腦可讀儲存媒介，當所述指令由所述一個或多個處理器執行時，所述指令將促使所述一個或多個處理器按照本文提供的方法的實施例執行操作。

本文還提供了用於實施本文提供的所述方法的系統。日誌結構儲存系統包括一個或多個處理器以及耦接到所述一個或多個處理器並且其上儲存有指令的電腦可讀儲存媒介，當所述指令由所述一個或多個處理器執行時，所述指令將促使所述一個或多個處理器按照本文提供的方法的實施例執行操作。

應瞭解，依據本文的方法可以包括本文描述的方面和特徵的任意組合。也就是說，根據本文的方法不限於本文具體描述的方面和特徵的組合，還包括所提供的方面和特徵的任意組合。

以下在附圖和描述中闡述了本文的一個或多個實施例

的細節。根據說明書和圖式以及請求項，本文的其他特徵和優點將顯現。

【圖式簡單說明】

[圖 1]是示出可用於執行本文實施例的環境的示例的圖。

[圖 2]是示出根據本文實施例的架構的示例的圖。

[圖 3]是示出根據本文實施例的基於區塊鏈的日誌結構儲存系統的示例的圖。

[圖 4]是示出根據本文實施例的分層儲存系統的示例的圖。

[圖 5]是示出根據本文實施例的用於執行日誌結構儲存系統的寫入操作的處理的示例的流程圖。

[圖 6]是示出根據本文實施例的用於產生與日誌結構儲存系統的寫入操作有關的索引的處理的示例的流程圖。

[圖 7]是示出根據本文實施例的用於執行日誌結構儲存系統的讀取操作的處理的示例的流程圖。

[圖 8]是示出根據本文實施例的用於改善日誌結構儲存系統的讀取操作的處理的示例的流程圖。

[圖 9]是示出根據本文實施例的用於管理儲存在日誌結構儲存系統中的資料日誌文件的處理的示例的流程圖。

[圖 10]是示出根據本文實施例的用於在日誌結構儲存系統中執行資料遷移的處理的示例的流程圖。

[圖 11]是示出根據本文實施例的用於在日誌結構儲存

系統中執行資料流控制的處理的示例的流程圖。

[圖 12]是示出可根據本文實施例執行的處理的示例的流程圖。

[圖 13]是示出可根據本文實施例執行的處理的示例的流程圖。

[圖 14]是示出可根據本文實施例執行的處理的示例的流程圖。

[圖 15]是示出根據本文實施例的裝置的模組的示例的圖。

在各個圖式中，相同的圖式標記和名稱表示相同的元件。

【實施方式】

本文描述了用於將資料儲存在例如分散式帳本系統(例如，區塊鏈網路)及/或基於區塊鏈的中心化帳本系統(例如，通用可審計帳本服務系統)中的日誌結構儲存系統的技術，所述分散式帳本系統及/或基於區塊鏈的中心化帳本系統採用區塊鏈的資料結構以利用儲存在區塊鏈上的資料的不可變性、可靠性以及可信性。在一些實施例中，分散式帳本系統和基於區塊鏈的中心化帳本系統可以統稱為基於區塊鏈的帳本系統。

在一些實施例中，基於區塊鏈的中心化帳本系統可以是基於中心化的帳本系統，其可以提供具有時間關鍵性審計(具有不可否認性和防竄改性)的、密碼學可驗證的與狀

態無關的資料帳本儲存。在一些實施例中，基於區塊鏈的中心化帳本系統可以基於雲端平臺提供帳本服務，該雲端平臺的特徵在於具有可信度和中立性的中心化背書。基於區塊鏈的中心化帳本系統可以提供高可靠性和高性能的可審計流水帳本服務，其組合了區塊鏈系統的高可信度以及中心化系統的高性能和低延遲，以處理具有審計要求、可追溯性和跟蹤的各種資料和日誌。

本文中描述的技術產生若干技術效果。在一些實施例中，所描述的技術可以被應用在各種應用和場景中，以提供有效的、可信的、可擴展的、有成本效益的和高性能的資料儲存和管理。所描述的技術可以為包括例如交易資料、區塊資料、狀態資料和索引資料的區塊鏈資料提供簡單且定義良好的應用程式介面(API)集。

所描述的技術提供了一種日誌結構儲存系統，該系統不僅提供I/O服務，還考慮了成本和個性化需求，以提供諸如分層、資料壓縮、共享儲存、抹除編碼和狀態快照的功能，尤其是在儲存在區塊鏈系統中的資料量達到規模之後。日誌結構儲存系統可以提供諸如日誌結構資料儲存以及異步及/或併發處理等特徵，以實施性能優化、高效處理、可信環境、通用性(例如，分散式帳本系統和基於區塊鏈的中心化帳本系統均可用)和改進的儲存方案。所描述的技術可以提供用於提供這樣的功能和特徵的總體框架或架構。

通常，在分散式帳本系統(例如，區塊鏈網路)中產生

及/或儲存的資料可以被稱為區塊鏈資料。區塊鏈資料可以包括或分類為交易資料、區塊資料、狀態資料和索引資料。在一些實施例中，在基於區塊鏈的中心化帳本系統(例如，通用可審計帳本服務系統)中產生及/或儲存的資料可以包括或分類為交易資料、區塊資料和索引資料。

在一些實施例中，可以以表示為<hash (value), value>的鍵值對(KVP)的形式來接收各種區塊鏈資料。該值可以是表示區塊、交易或區塊鏈網路狀態中一個或多個的實際資料。鍵可以是該值的雜湊值。

在一些實施例中，對於區塊資料，每個區塊可以包括區塊頭和區塊體。區塊頭可以包括特定區塊的身分資訊，並且區塊體可以包括用該區塊確認的交易。在一些實施例中，區塊資料是區塊鏈系統中的資料結構，並且通常具有以下一個或多個特性。例如，(1)在區塊鏈網路中達成共識後，儲存在區塊鏈網路中的每個節點中的區塊資料的內容在理論上是一致的。(2)區塊號密集地增加。(3)連續區塊之間存在雜湊糾纏(hash entanglement)。(4)區塊資料是僅追加(append-only)的。也就是說，一旦達成共識，歷史區塊資料將不會被修改。(5)區塊資料的存取頻率通常較低。區塊資料佔用的儲存空間通常很大。

在一些實施例中，狀態資料可以被組裝(assemble)為全域共享狀態(也稱為世界狀態)。世界狀態可以包括帳戶地址和帳戶狀態之間的映射。世界狀態可以儲存在諸如默克爾帕特麗夏樹(Merkle Patricia tree，MPT)的資料結構

中。在一些實施例中，例如，在智慧合約場景中，可以基於默克爾樹的內容來設計狀態資料。它是一個增量式的內容尋址資料集。狀態資料佔用的儲存空間通常很大。

在一些實施例中，狀態資料可以進一步分類為當前狀態和歷史狀態。在一些實施例中，當前狀態是與最新區塊相對應的狀態資料，並且是在區塊鏈網路上執行最新交易時的資料源。在一些實施例中，歷史狀態是儲存從創世區塊到最新區塊的所有狀態資料的內容尋址資料集。在一些實施例中，歷史狀態資料被儲存在歷史狀態樹中。歷史狀態樹可以將狀態資訊儲存為內容可尋址的表示為<hash (node value), node value>的鍵值對(KVP)。值或節點值可以是與區塊鏈節點關聯的帳戶的帳戶狀態，而鍵可以是相應帳戶狀態的雜湊值。在一些實施例中，當前狀態資料被儲存在當前狀態樹中。在一些實施例中，可以基於一個或多個位置相關識別符(ID)來對當前狀態樹進行位置尋址。例如，當前狀態樹可以將狀態資訊儲存為表示為<node ID, node value>的KVP，其中可以根據相應的節點ID來尋址節點值。

在一些實施例中，交易資料可以包括與一系列操作的輸入和輸出有關的資料。在一些實施例中，交易資料可以包括與有價物(例如，資產、產品、服務、貨幣)的交換有關的資料。

在一些實施例中，索引資料可以指示資料(例如，交易資料、區塊資料和狀態資料)與儲存該資料以用於尋址

或檢索該資料的資料日誌文件之間的映射對應關係。在一些實施例中，索引資料可以指示對應資料儲存在儲存系統中的實體位置。在一些實施例中，索引資料可以包括以下中的一個或多個：指示從區塊雜湊值到區塊號的對應關係的索引、指示從區塊雜湊值到儲存位置的對應關係的索引、指示從交易雜湊值到交易的對應關係的索引、或指示從收據雜湊值到收據的對應關係的索引。在一些實施例中，索引資料不包括區塊鏈資料的內容。

當越來越多的交易輸入到區塊鏈中時，區塊鏈資料(例如狀態資料和區塊資料)的大小會越來越大。在DLS中，DLS的每個節點儲存區塊鏈的整個副本，即使某些舊區塊資料或狀態資料不經常被訪問，也可能佔用大量儲存空間。

在一些實施例中，區塊鏈資料由日誌結構系統儲存在資料文件中，並且資料文件基於時間被連續地追加和劃分。在一些實施例中，可以不根據鍵的分類來重新排列資料(例如，資料不是按鍵值或其他度量排列的，使得熱資料和冷資料不被混合在多個資料日誌文件中)，從而大大降低了分層實施的技術挑戰。

在一些實施例中，日誌結構儲存系統使用兩種僅追加資料文件來儲存區塊鏈資料以提供資料持久性：資料日誌文件和索引日誌文件。例如，區塊資料、交易資料、狀態資料和附加自描述資料可以被儲存在資料日誌文件中，而指示交易資料、區塊資料和狀態資料的儲存位置的索引資

料(例如，資料日誌文件的識別符和偏移量)可以儲存在索引日誌文件中。

在區塊鏈資料中，交易資料和區塊資料可以是日誌結構友好的，可以包括僅追加資料，使得可以透過直接將這些資料添加或追加到相應的資料日誌文件中來將其寫入資料日誌文件中。在一些實施例中，交易資料和區塊資料的寫入不需要大量壓合(compaction)。例如，它可能需要相對少量的交易重現，並且可能不需要區塊回滾(rollback)。在一些實施例中，狀態資料可以是日誌結構友好的資料，使得歷史狀態資料可以在不需要壓合的情況下增加。

在一些實施例中，日誌結構儲存系統可以支援多級資料分層，並支援多種儲存設備，例如雲端磁碟、網路連接系統(NAS)和對象儲存服務(OSS)(低頻，存檔)。例如，日誌文件可以儲存在基於雲端的儲存系統、NAS或OSS設備、或自建的分散式儲存系統中。

在一些實施例中，不同類型的日誌文件可以具有不同的儲存策略。例如，可以將相對較長時間未被存取的資料日誌文件儲存在廉價且相對低速的例如NAS/OSS的儲存設備中，並可以使用壓縮和抹除編碼進行處理以用於儲存。作為另一示例，索引日誌文件可以儲存在諸如雲端磁碟的高速儲存設備上。

在一些實施例中，日誌結構儲存系統可以透過使用最近最少使用的(LRU)記憶體快取和磁碟快取來執行資料分層，以優化低速儲存設備的讀取性能。

在一些實施例中，日誌結構儲存系統可以提供管理多級的儲存設備池的分層池管理器。在一些實施例中，每個池支援集群中的多個磁碟或儲存設備。分層池管理器可以管理池的空間、壓力和運行狀況。在一些實施例中，日誌結構儲存系統可以提供遷移任務管理器，該遷移任務管理器管理不同級的儲存設備之間的資料的雙向遷移任務；管理結果回呼、統計資訊、遷移任務的生命週期等。在一些實施例中，日誌結構儲存系統可以提供支援可插拔策略的遷移調度器、管理資料遷移策略並提供資料創建/查詢/更新/刪除介面。

所揭露的日誌結構儲存系統採用了合併樹(LSM-Tree)架構的思想。在一些實施例中，日誌結構儲存系統可以包括多個日誌結構儲存實例(或流)，其中每個日誌結構儲存實例負責儲存和管理分散式帳本系統(例如，區塊鏈系統)或基於區塊鏈的中心化帳本系統的資料。在一些實施例中，日誌結構儲存系統可以將隨機寫入操作轉換為有序追加操作，以減輕由於大量隨機寫入操作引起的頻繁的“髒”頁刷新而導致的寫入放大問題。在一些實施例中，日誌結構儲存系統可以延遲高性能場景中的寫入刷新操作並減少同步操作的次數，以提高整個系統的效率和性能。

為了提供本文實施例的進一步的背景，並且如上所述，分散式帳本系統(DLS)，也可以被稱為共識網路(例如，由點對點節點組成)和區塊鏈網路，使得參與實體能夠安全地並且不可變地進行交易並儲存資料。儘管用語區

塊鏈通常與特定網路及/或用例相關聯，但是在不參考任何特定用例的情況下，本文中使用的區塊鏈通常是指DLS。

區塊鏈是以交易不可變的方式儲存交易的資料結構。因此，區塊鏈上記錄的交易是可靠且可信的。區塊鏈包括一個或多個區塊。鏈中的每個區塊透過包含鏈中其緊接在前的前一區塊的加密雜湊值(cryptographic hash)鏈接到該前一區塊。每個區塊還包括時間戳、自身的加密雜湊值以及一個或多個交易。已經被區塊鏈網路中的節點驗證的交易經雜湊處理並編碼入默克爾(Merkle)樹中。Merkle樹是一種資料結構，對樹的葉節點處的資料進行雜湊處理，並且在樹的每個分支中的所有雜湊值在該分支的根處級聯。沿著樹持續該處理一直到整個樹的根，在整個樹的根處儲存了代表樹中所有資料的雜湊值。透過確定雜湊值是否與該樹的結構一致而可快速驗證該雜湊值是否為儲存在該樹中的交易的雜湊值。

在一些實施例中，例如，在作為計算節點網路的區塊鏈網路中，可以以分散式或去中心化或至少部分去中心化的方式來實施區塊鏈以用於儲存交易。每個計算節點(又稱為區塊鏈網路節點)可以透過廣播交易、驗證交易和確認交易有效性等來管理、更新和維護一個或多個區塊鏈。如上所述，區塊鏈網路可作為公有區塊鏈網路、私有區塊鏈網路或聯盟區塊鏈網路被提供。在本文中參考聯盟區塊鏈網路進一步詳細描述本文實施例。然而，可以預期，本

文實施例可以在任何適當類型的區塊鏈網路中實現。

通常，聯盟區塊鏈網路在參與的實體中是私有的。在聯盟區塊鏈網路中，共識處理由授權的節點集控制，該節點集可以被稱為共識節點，一個或多個共識節點由相應實體(例如，金融機構、保險公司)操作。例如，由10個實體(例如，金融機構、保險公司)組成的聯盟可以操作聯盟區塊鏈網路，每個實體操作聯盟區塊鏈網路中的至少一個節點。

在一些示例中，在聯盟區塊鏈網路內，提供全域區塊鏈作為跨所有節點複製的區塊鏈。也就是說，對於全域區塊鏈，所有的共識節點處於完全共識狀態。為了達成共識(例如，同意將區塊添加到區塊鏈)，在聯盟區塊鏈網路內實施共識協定。例如，聯盟區塊鏈網路可以實施實用拜占庭容錯(PBFT)共識，下面將進一步詳細描述。

在一些實施例中，中心化帳本系統還可以採用區塊鏈的資料結構，以利用儲存在區塊鏈上的資料的不可變性、可靠性和可信性。在一些實施例中，例如中心化帳本系統可以被稱為基於區塊鏈的中心化帳本系統或通用可審計帳本服務系統。在一些實施例中，基於區塊鏈的中心化帳本系統可以包括中央可信機構，該中央可信機構提供儲存在區塊鏈資料結構的區塊中的透明、不可變且可密碼學可驗證的資料。所儲存的資料可以是日誌格式，例如不僅包括交易日誌，還包括其他交易資料和區塊資料。由於中央信任機構的存在，基於區塊鏈的中心化帳本系統無需執行共

識處理即可建立信任。在一些實施例中，與典型基於區塊鏈的分散式或去中心化帳本系統相比，基於區塊鏈的中心化帳本系統可以更高效。在一些實施例中，基於區塊鏈的中心化帳本系統可以提供具有增強的信任、效率和儲存性能的基於雲端的儲存服務。

圖1是示出了可用於執行本文實施例的環境100的示例的圖。在一些示例中，環境100使得實體能夠參與聯盟區塊鏈網路102。環境100包括計算設備106、108和網路110。在一些示例中，網路110包括區域網路(LAN)、廣域網路(WAN)、網際網路或其組合，並且連接網站、用戶設備(例如，計算設備)和後端系統。在一些示例中，可以透過有線及/或無線通訊鏈路來存取網路110。在一些示例中，網路110使得能夠與聯盟區塊鏈網路102通訊或在聯盟區塊鏈網路102內部通訊。通常，網路110表示一個或多個通訊網路。在一些情況下，計算設備106、108可以是雲端計算系統(未示出)的節點，或者每個計算設備106、108可以是單獨的雲端計算系統，其包括透過網路互連並且用作分散式處理系統的多個電腦。

在所描繪的示例中，計算設備106、108可以各自包括能夠作為節點參與至聯盟區塊鏈網路102中的任何適當的計算系統。計算設備的示例包括(但不限於)伺服器、臺式電腦、筆記本電腦、平板電腦和智慧手機。在一些示例中，計算設備106、108承載用於與聯盟區塊鏈網路102交互的一個或多個由電腦實施的服務。例如，計算設備106

可以承載第一實體(例如，用戶A)的由電腦實施的、例如交易管理系統的服務，例如，第一實體使用該交易管理系統管理其與一個或多個其他實體(例如，其他用戶)的交易。計算設備108可以承載第二實體(例如，用戶B)的由電腦實施的、例如交易管理系統的服務，例如，第二實體使用該交易管理系統管理其與一個或多個其他實體(例如，其他用戶)的交易。在圖1的示例中，聯盟區塊鏈網路102被表示為節點的點對點網路(Peer-to-Peer network)，並且計算設備106、108分別提供參與聯盟區塊鏈網路102的第一實體和第二實體的節點。

圖2是示出根據本文實施例的架構200的圖。示例性概念架構200包括分別對應於參與者A、參與者B和參與者C的參與者系統202、204、206。每個參與者(例如，用戶、企業)參與到作為點對點網路提供的區塊鏈網路212中，該點對點網路包括多個節點214，至少一些節點將資訊不可變地記錄在區塊鏈216中。如圖中進一步詳述，儘管在區塊鏈網路212中示意性地描述了單個區塊鏈216，但是在區塊鏈網路212上提供並維護了區塊鏈216的多個副本。

在所描繪的示例中，每個參與者系統202、204、206分別由參與者A、參與者B和參與者C提供或代表參與者A、參與者B和參與者C，並且在區塊鏈網路中作為各自的節點214發揮作用。如這裡所使用的，節點通常是指連接到區塊鏈網路212且使相應的參與者能夠參與到區塊鏈網路中的個體系統(例如，電腦、伺服器)。在圖2的示例

中，參與者對應於每個節點214。然而，可以預期，一個參與者可以操作區塊鏈網路212內的多個節點214，及/或多個參與者可以共享一個節點214。在一些示例中，參與者系統202、204、206使用協定(例如，超文字傳輸協定安全(HTTPS))及/或使用遠端程序呼叫(RPC)與區塊鏈網路212通訊或透過區塊鏈網路212進行通訊。

節點214可以在區塊鏈網路212內具有不同的參與程度。例如，一些節點214可以參與共識處理(例如，作為將區塊添加到區塊鏈216的監視節點)，而其他節點214不參與此共識處理。作為另一示例，一些節點214儲存區塊鏈216的完整的副本，而其他節點214僅儲存區塊鏈216的一部分的副本。例如，資料存取特權可以限制相應的參與者在其相應系統內儲存的區塊鏈資料。在圖2的示例中，參與者系統202、204和206儲存區塊鏈216的相應的完整副本216'、216''和216'''。

區塊鏈(例如，圖2的區塊鏈216)由一系列區塊組成，每個區塊儲存資料。資料的示例包括表示兩個或更多參與者之間的交易的交易資料。儘管本文透過非限制性示例使用了“交易”，但是可以預期，任何適當的資料可以儲存在區塊鏈中(例如，文檔、圖像、視訊、音訊)。交易的示例可以包括(但不限於)有價物(例如，資產、產品、服務、貨幣)的交換。交易資料不可竄改地儲存在區塊鏈中。也就是說，交易資料不能改變。

在將交易資料儲存在區塊中之前，對交易資料進行雜

湊處理。雜湊處理是將交易資料(作為字符串資料提供)轉換為固定長度雜湊值(也作為字符串資料提供)的過程。不可能對雜湊值進行去雜湊處理(un-hash)以獲取交易資料。雜湊處理可確保即使交易資料輕微改變也會導致完全不同的雜湊值。此外，如上所述，雜湊值具有固定長度。也就是說，無論交易資料的大小如何，雜湊值的長度都是固定的。雜湊處理包括透過雜湊函數處理交易資料以產生雜湊值。雜湊函數的示例包括(但不限於)輸出256位元雜湊值的安全雜湊算法(SHA)-256。

多個交易的交易資料被雜湊處理並儲存在區塊中。例如，提供兩個交易的雜湊值，並對它們本身進行雜湊處理以提供另一個雜湊值。重複此過程，直到針對所有要儲存在區塊中的交易提供單個雜湊值為止。該雜湊值被稱為Merkle根雜湊值，並儲存在區塊的頭中。任何交易中的更改都會導致其雜湊值發生變化，並最終導致Merkle根雜湊值發生變化。

透過共識協定將區塊添加到區塊鏈。區塊鏈網路中的多個節點參與共識協定，並競相將區塊添加到區塊鏈中。這種節點稱為共識節點。上面介紹的PBFT用作共識協定的非限制性示例。共識節點執行共識協定以將交易添加到區塊鏈，並更新區塊鏈網路的整體狀態。

更詳細地，共識節點產生區塊頭，對區塊中的所有交易進行雜湊處理，並將所得的雜湊值成對地組合以產生進一步的雜湊值，直到為區塊中的所有交易提供單個雜湊值

(Merkle根雜湊值)。將此雜湊值添加到區塊頭中。共識節點還確定區塊鏈中最新區塊(即，添加到區塊鏈中的最後一個區塊)的雜湊值。共識節點還向區塊頭添加隨機數(nonce)和時間戳。

通常，PBFT提供容忍拜占庭故障(例如，故障節點、惡意節點)的實用拜占庭狀態機複製。這透過在PBFT中假設將發生故障(例如，假設存在獨立節點故障及/或由共識節點發送的操縱訊息)而實現。在PBFT中，以包括主共識節點和備共識節點的順序提供共識節點。主共識節點被週期性地改變。透過由區塊鏈網路內的所有共識節點對區塊鏈網路的全域狀態達成一致，將交易添加到區塊鏈中。在該處理中，訊息在共識節點之間傳輸，並且每個共識節點證明訊息是從指定的對等節點(peer node)接收的，並驗證在傳輸期間訊息未被修改。

在PBFT中，共識協定是在所有共識節點以相同的狀態開始的情況下分多個階段提供的。首先，客戶端向主共識節點發送調用服務操作(例如，在區塊鏈網路內執行交易)的請求。響應於接收到請求，主共識節點將請求組播到備共識節點。備共識節點執行請求，並且各自向客戶端發送回覆。客戶端等待直到接收到閾值數量的回覆。在一些示例中，客戶端等待直到接收到f+1個回覆，其中f是區塊鏈網路內可以容忍的錯誤共識節點的最大數量。最終結果是，足夠數量的共識節點就將記錄添加到區塊鏈的順序達成一致，並且該記錄或被接受或被拒絕。

在一些區塊鏈網路中，用密碼學來維護交易的隱私。例如，如果兩個節點想要保持交易隱私，以使得區塊鏈網路中的其他節點不能看出交易的細節，則這兩個節點可以對交易資料進行加密處理。加密處理的示例包括但不限於對稱加密和非對稱加密。對稱加密是指使用單個密鑰既進行加密(從明文產生密文)又進行解密(從密文產生明文)的加密過程。在對稱加密中，同一密鑰可用於多個節點，因此每個節點都可以對交易資料進行加密/解密。

非對稱加密使用密鑰對，每個密鑰對包括私鑰和公鑰，私鑰僅對於相應節點是已知的，而公鑰對於區塊鏈網路中的任何或所有其他節點是已知的。節點可以使用另一個節點的公鑰來加密資料，並且該加密的資料可以使用其他節點的私鑰被解密。例如，再次參考圖2，參與者A可以使用參與者B的公鑰來加密資料，並將加密資料發送給參與者B。參與者B可以使用其私鑰來解密該加密資料(密文)並提取原始資料(明文)。使用節點的公鑰加密的訊息只能使用該節點的私鑰解密。

非對稱加密用於提供數位簽名，這使得交易中的參與者能夠確認交易中的其他參與者以及交易的有效性。例如，節點可以對訊息進行數位簽名，而另一個節點可以根據參與者A的該數位簽名來確認該訊息是由該節點發送的。數位簽名也可以用於確保訊息在傳輸過程中不被竄改。例如，再次參考圖2，參與者A將向參與者B發送訊息。參與者A產生該訊息的雜湊值，然後使用其私鑰加密

該雜湊值以提供作為加密雜湊值的數位簽名。參與者A將該數位簽名附加到該訊息上，並將該具有數位簽名的訊息發送給參與者B。參與者B使用參與者A的公鑰解密該數位簽名，並提取雜湊值。參與者B對該訊息進行雜湊處理並比較雜湊值。如果雜湊值相同，參與者B可以確認該訊息確實來自參與者A，且未被竄改。

圖3是示出根據本文實施例的日誌結構儲存系統的示例的圖。日誌結構儲存系統300可以儲存將資料儲存在一個或多個區塊鏈上的分散式帳本系統(例如，區塊鏈網路)及/或基於區塊鏈的中心化帳本系統(例如，通用可審計帳本服務系統)(統稱為基於區塊鏈的帳本系統)的資料。

在一些實施例中，日誌結構儲存系統300可以由區塊鏈網路的每個共識節點或基於區塊鏈的中心化帳本系統的中央節點來實施。在一些實施例中，日誌結構儲存系統300可以連接到由基於區塊鏈的帳本系統的客戶端節點構建的分散式儲存系統340。如圖所示，日誌結構儲存系統300包括前端輸入/輸出(I/O)子系統310、多層儲存子系統320和後端資料管理子系統330。在一些實施例中，前端I/O子系統310可以執行寫入操作以將資料寫入到儲存在多層儲存子系統320中的資料文件(例如，資料日誌文件和索引日誌文件)中，並執行讀取操作，以存取來自儲存在多層儲存子系統320中的資料文件的資料。在一些實施例中，後端資料管理子系統330可以根據不同的需求對資料文件中的資料進行處理、重組和管理，以提高整個系統的

效率和性能。

前端I/O子系統310可以包括任何合適的計算元件(例如，處理器、記憶體315等中的一個或多個)以執行本文所述的方法。在一些實施例中，前端I/O子系統310可以對多種資料元素執行包括各種讀寫操作(例如，插入、更新、刪除、查詢等)的前端I/O操作。

在一些實施例中，前端I/O子系統310處理的所有資料元素(例如，交易資料、區塊資料和狀態資料)可以以日誌文件格式被儲存，無論該日誌文件是從寫入操作產生的還是從後端資料管理子系統330的例如儲存分層、壓合、資料壓縮、抹除編碼等的操作產生的文件。

在一些實施例中，前端I/O子系統310處理的資料可以被儲存在以下兩種日誌文件中：(1)儲存諸如區塊鏈資料(例如，交易資料、區塊資料、狀態資料)和自描述元資料的實質資料之資料日誌文件(例如，資料日誌文件390、362、364、366、372、374和376)；以及(2)儲存指示資料的實體位置的索引資訊(例如，資料日誌文件的識別符和偏移量)的索引日誌文件(例如，索引日誌文件380)。在一些實施例中，資料日誌文件不儲存索引資訊，而索引資訊由單獨的索引日誌文件維護。

在一些實施例中，前端I/O子系統310可以被配置為執行寫入操作以將區塊鏈資料寫入到資料日誌文件390中。在一些實施例中，區塊鏈資料可以包括由區塊鏈網路或分散式帳本系統產生的區塊資料、交易資料或狀態資料。在

一些實施例中，區塊鏈資料可以包括由基於區塊鏈的中心化帳本系統產生的區塊資料和交易資料。在一些實施例中，寫入資料日誌文件390的資料可以包括描述資料區塊的元資料，諸如交易雜湊值和序列值、區塊雜湊值和區塊號、快照版本號、循環冗餘校驗(CRC)碼、加密資訊等。在一些實施例中，資料日誌文件390可以是僅追加文件。

在一些實施例中，前端I/O子系統310可以被配置為產生指示相應資料儲存在日誌結構儲存系統300中(例如，多層儲存子系統320中的資料日誌文件中)的實體位置的索引。在一些實施例中，索引可以被儲存在索引日誌文件380中。在一些實施例中，資料日誌文件和索引日誌文件可以被儲存在多層儲存子系統320中。在一些實施例中，索引可以被儲存在索引日誌文件380中，該索引日誌文件380被儲存在多層儲存子系統320的儲存設備中存取速度最高的一個儲存設備中。

在一些實施例中，可以基於資料寫入或追加操作來持續更新資料日誌文件。在一些實施例中，資料日誌文件可具有可配置的最大長度，例如在512MB和2GB之間。在一些實施例中，如果確定資料日誌文件已經達到最大長度或大小，則可以將資料日誌文件密封或設置為唯讀，並且可以為新寫入操作分配新資料日誌文件。

在一些實施例中，前端I/O子系統310可執行寫入操作，包括對儲存在日誌結構儲存系統300中的資料的修改。在一些實施例中，前端I/O子系統310透過以日誌格式

將資料添加或追加到資料日誌文件中，來處理對資料的修改，從而不覆蓋原始資料。在一些實施例中，資料日誌文件可以形成可用於崩潰恢復的先寫日誌(WAL)層。

在一些實施例中，前端I/O子系統310將索引資訊儲存在記憶體315中，該索引資訊指示資料(例如，交易資料、區塊資料和狀態資料)與儲存該資料的資料日誌文件之間的映射對應關係，以尋址或檢索資料。在一些實施例中，可以使用日誌結構合併(LSM)方法來組織記憶體中的索引資料。在一些實施例中，新寫入的資料的索引可以被儲存在記憶體315中，並在記憶體使用率(memory usage)超過預定閾值時被刷新(flush)到索引日誌文件380中。這樣，舊資料的索引可以被儲存在磁碟儲存設備或硬碟驅動器儲存設備中的索引日誌文件380中，從而釋放出空間以在記憶體315中快取新的熱點資料的索引。

在一些實施例中，索引資料可以包括以下中的一個或多個：指示從區塊雜湊值到區塊號的對應關係的索引、指示從區塊雜湊值到儲存位置的對應關係的索引、指示從交易雜湊值到交易的對應關係的索引、或指示從收據雜湊值到收據的對應關係的索引。在一些實施例中，用於基於區塊鏈的中心化帳本系統的索引資料可以包括指示從序列到交易儲存位置的對應關係的索引及/或指示從時序到交易雜湊值的對應關係的索引。

在一些實施例中，前端I/O子系統310可以包括儲存在記憶體315中的多個記憶體內索引映射。在一些實施例

中，記憶體內索引映射可以被認為是用於維護前端I/O子系統310的記憶體中的索引資料的任何合適的組件、單元、模組或資料結構(例如，表或結構)。記憶體內索引映射可以是前端I/O子系統310的關鍵組件，其確定前端I/O子系統310和整個日誌結構儲存系統300的可擴展性和性能。在一些實施例中，因為區塊鏈資料的時間敏感性強，並且最近寫入的交易資料和區塊資料被再次存取的機會相對較高，所以日誌結構儲存系統300可以將熱資料的索引儲存在記憶體315中的索引映射中，以改善整個日誌結構儲存系統300的性能。

在一些實施例中，記憶體內索引映射可以維護指示從交易雜湊值到序列值的映射的索引及/或指示從區塊雜湊值和區塊號到資料的實體位置的映射的索引中的一個或多個。在一些實施例中，前端I/O子系統310將記憶體315中的索引映射的檢查點定期持久化到索引日誌文件中。例如，前端I/O子系統310可以週期性地或在某個時間點捕獲記憶體315中的索引資料的快照，並將快照儲存在多層儲存子系統320中的索引日誌文件380中。這可以創建一個時間點，在日誌結構儲存系統300意外關閉或崩潰之後的恢復過程中，日誌結構儲存系統300可以在該時間點應用包含在索引日誌文件380中的更改。在一些實施例中，前端I/O子系統310可以透過查詢記憶體內索引映射並確定所請求資料的當前位置來讀取資料(例如，交易資料、區塊資料和狀態資料)。

在一些實施例中，可以在創建索引日誌文件時將記憶體內索引映射的完整檢查點寫入索引日誌文件。在一些實施例中，可以透過批量處理寫入操作的索引來更新索引日誌文件。在一些實施例中，批大小可以是動態可配置的，例如數千個交易寫入操作或幾兆位元組(MB)的寫入操作。在一些實施例中，當已經針對特定批次數的寫入操作更新了索引日誌文件時，可以將索引日誌文件密封或設置為唯讀，並且可以創建新索引日誌文件以寫入新資料。

在一些實施例中，為了從異常崩潰中恢復，前端I/O子系統310可以將索引日誌文件(例如，索引日誌文件380)加載到記憶體315中，並且掃描資料日誌文件390的頁底部以確保資料日誌文件390和索引日誌文件380的一致性。在一些實施例中，索引日誌文件可能落後於資料日誌文件幾個批次，因此恢復時間可能佔用有限的I/O資源和時間。

在一些實施例中，可以將新寫入的交易資料和區塊資料的索引添加到索引映射和索引日誌文件中，但是除了在重放攻擊和區塊回滾場景中之外，可以不修改現有交易資料和區塊資料的索引。在一些實施例中，為了實現讀寫操作的高併發性，可以將記憶體內索引映射分為唯讀基本索引映射316和讀寫增量索引映射312。在一些實施例中，基本索引映射316可以儲存冷資料的索引，而增量索引映射312可以儲存新寫入的資料的索引。在一些實施例中，雜湊值索引可以儲存在雜湊值表中，而序列索引可以儲存在B樹(B-tree)中。

在一些實施例中，在前端I/O子系統310的寫入操作期間，可以首先將資料的索引資訊更新到增量索引映射312。在讀取操作期間，前端I/O子系統310可以首先在增量索引映射312中搜索所請求的資料。如果在增量索引映射312中未找到所請求的資料，則前端I/O子系統310可以隨後搜索基本索引映射316。

在一些實施例中，前端I/O子系統310可以定期將索引資料從記憶體315刷新到索引日誌文件380。在一些實施例中，索引刷新的基本過程可以包括以下操作：(1)組合增量索引映射312和基本索引映射316；(2)對基本索引映射316進行持久化處理(例如，將基本索引映射儲存到索引日誌文件中)；(3)從記憶體315釋放基本索引映射316的部分或全部；(4)透過讀取請求的索引資料將索引資料交換到記憶體315。

在一些實施例中，前端I/O子系統310可以將記憶體315中的增量索引映射312轉換為不可變索引映射314，然後將它們刷新到索引日誌文件380，並創建新增量索引映射以接收根據新請求產生的索引。這樣，可以減少增量索引映射的儲存佔用，以改善日誌結構儲存系統300的性能。

在一些實施例中，為了減少對前端I/O的影響，記憶體中的索引映射可以在後端異步地合併。在一些實施例中，可以透過以下兩個條件中的至少一個來觸發合併處理：(1)增量索引映射的大小超過預定閾值；及(2)新快照

被創建。在一些實施例中，前端I/O子系統310可以產生合併索引映射以包括要被刷新到索引日誌文件380中的不可變索引映射314。在一些實施例中，前端I/O子系統310可以將合併索引映射與當前基本索引映射316組合以產生新基本索引映射。

在一些實施例中，在操作期間，前端I/O子系統310可以與多個基本索引映射和索引日誌文件一起運行。在一些實施例中，當在某些場景下需要壓合時，可以透過將所有基本索引映射和增量索引映射組合成一個基本索引映射來定期執行次要壓合(minor compaction)和主要壓合(major compaction)。主要壓合主要合併和管理索引，這可用於例如快照、垃圾回收加載和索引文件管理的場景。

在一些實施例中，可以透過合併基本索引映射和增量索引映射並產生新基本索引映射，並且將其儲存到新索引日誌文件中來執行主要壓合。在一些實施例中，可以透過組合若干索引日誌文件並產生新索引日誌文件來執行次要壓合，這可以減少索引日誌文件的數量。在一些實施例中，如果當前索引日誌文件的大小達到預定閾值，則可以將當前索引日誌文件設置為密封或不可變狀態並關閉，並且可以為新索引資料創建新索引日誌文件。

在一些實施例中，在讀取操作期間，如果在記憶體內索引映射中的搜索失敗，則可能需要兩個或多個I/O操作，這可能給日誌結構儲存系統300帶來負擔。在一些實施例中，前端I/O子系統310可以提供具有記憶體快取313

和區塊快取317的多級快取機制(例如，使用快閃記憶體媒介(例如，SSD雲端磁碟))。

在一些情況下，日誌結構儲存系統300可以接收較大的讀取請求，使得日誌結構儲存系統300需要存取多個資料日誌文件從而為客戶端獲取完整的請求資料。然而，存取多個資料日誌文件可能會導致不小的開銷。在一些實施例中，後端資料管理子系統330可以執行壓合操作以級聯邏輯上相鄰的資料區塊從而減少碎片。在一些實施例中，壓合操作可能具有開銷，並且可以在資料碎片嚴重時執行。

在一些實施例中，多層儲存子系統320可以包括多層儲存設備。儲存設備可以包括儲存媒介以及相應的軟體及/或硬體介面。在一些實施例中，多層儲存設備可以包括具有不同性能特性的多個儲存設備。例如，多層儲存設備可以包括雲端磁碟、網路附加儲存(NAS)設備和對象儲存服務(OSS)設備。在一些實施例中，多層儲存設備是按根據一個或多個性能特性的分層結構分層的。在一些實施例中，一個或多個性能特性可以包括存取速度、存取頻寬或存取延遲。例如，多層儲存設備可以包括具有第一性能特性(例如，存取速度)的第一層儲存設備和具有低於第一性能特性的第二性能特性(例如，相對於第一層儲存設備而言相對較低的存取速度)的第二層儲存設備等。如圖3所示，多層儲存子系統320的示例可以包括第一層儲存設備350、第二層儲存設備360和第三層儲存設備370，該第一

層儲存設備350包括雲端磁碟或基於雲端的儲存設備(例如，固態驅動器(SSD)雲端磁碟、嵌入式SSD(ESSD)雲端磁碟)，該第二層儲存設備360包括NAS設備，該第三層儲存設備370包括OSS設備。

在一些實施例中，多層儲存設備可以儲存不同類型的資料。在一些實施例中，可以基於例如資料被產生或接收的時間或資料被存取的頻率將資料分類為熱資料355、暖資料365和冷資料375。例如，最新交易的資料可以是熱資料；昨天的交易資料可以是暖資料，而一周前進行的歷史交易資料可以是冷資料。又例如，區塊鏈中最近產生的10個區塊中的資料可以是熱資料；最近產生的11～20個區塊中的資料可以是暖資料，其他較早的區塊中的資料可以是冷資料。然而，在一些實施例中，區塊鏈的創世區塊可以被視為熱資料，因為其頻繁被存取。

在一些實施例中，多層儲存子系統320可以將熱資料355、暖資料365和冷資料375分別儲存到多層儲存設備中。例如，第一層儲存設備350可以儲存熱資料355；第二層儲存設備360可以儲存暖資料365；第三層儲存設備370可以儲存冷資料375。在一些實施例中，一層儲存設備可以例如基於儲存空間和成本來儲存一種或多種資料。例如，第一層儲存設備350可以儲存熱資料355和一些暖資料365，而第二層儲存設備360可以儲存其餘的暖資料365和一些冷資料375。

在一些實施例中，儲存設備的每一層可以儲存資料日

誌文件，該資料日誌文件包括由基於區塊鏈的帳本系統(例如，分散式帳本系統及/或基於區塊鏈的中心化帳本系統)產生的區塊鏈資料。例如，第一層儲存設備350可以儲存包括由基於區塊鏈的帳本網路產生的第一區塊鏈資料的第一資料日誌文件390，並且第二層儲存設備360可以儲存包括由基於區塊鏈的帳本系統產生的第二區塊鏈資料的第二資料日誌文件362，等等。

在一些實施例中，相比於儲存在儲存設備的相對較高層中的資料日誌文件中的區塊鏈資料，儲存在儲存設備的相對較低層中的資料日誌文件中的區塊鏈資料可以更早地被寫入。例如，相比於儲存在第一層儲存設備350上的第一資料日誌文件390中的第一區塊鏈資料，儲存在第二層儲存設備360上的第二資料日誌文件362中的第二區塊鏈資料可以更早地被寫入。

在一些實施例中，第一層儲存設備350可以進一步儲存一個或多個索引日誌文件380，該索引日誌文件380包括索引資料，該索引資料用於指示多層儲存設備350、360和370儲存的資料日誌文件390、362、364、366、372、374和376中資料的實體儲存位置。例如，如圖3所示，第一層儲存設備350可以儲存索引日誌文件380，該索引日誌文件380包括的索引資料指示第一層儲存設備350儲存的資料日誌文件390中區塊鏈資料的實體儲存位置、第二層儲存設備360儲存的資料日誌文件362、364和366中區塊鏈資料的實體儲存位置、以及第三層儲存設備370儲存的資料日誌

文件372、374和376中區塊鏈資料的實體儲存位置。

在一些實施例中，一個或多個索引日誌文件可以被儲存在第二層儲存設備360及/或第三層儲存設備370。

在一些實施例中，儲存在多層儲存子系統320上的索引日誌文件和資料日誌文件是僅追加日誌文件。在一些實施例中，儲存在資料日誌文件中的區塊鏈資料可以包括區塊資料、交易資料和歷史狀態資料。

在一些實施例中，較高層的儲存設備可以儲存包括從較低層的儲存設備遷移來的區塊鏈資料的資料日誌文件。例如，第一層儲存設備可以儲存包括這樣的區塊鏈資料的資料日誌文件：該區塊鏈資料比第二層儲存設備的資料日誌文件中的區塊鏈資料和從第二層儲存設備遷移來的區塊鏈資料更頻繁地被存取。

在一些實施例中，儲存系統300可以進一步包括分散式儲存系統340，該分散式儲存系統340包括諸如非揮發性記憶體快速(non-volatile memory express，NVME)、SSD、硬碟驅動器(HDD)和疊片式磁記錄(SMR)的儲存媒介。在一些實施例中，分散式儲存系統340可以由基於區塊鏈的帳本系統的客戶端節點產生或擴展，以具有更好的可用性、分區容忍度、靈活性和成本。例如，分散式儲存系統340可以透過添加更多伺服器或儲存節點並由此線性地增加容量和性能來允許擴展。它可以使用價格便宜的標準伺服器、驅動器和網路。在一些實施例中，分散式儲存系統340可以提高標準伺服器的利用率，因此導致更少的功

耗、更好的冷卻效率、更好的空間利用率、更少的維護成本等。

前端I/O子系統310可以對區塊鏈資料執行寫入操作並產生索引日誌文件380以及儲存在多層儲存子系統320上的資料日誌文件390、362、364、366、372、374和376。隨著時間的流逝，儲存在多層儲存子系統320上的資料可能會累積和聚集，並且可能會降低日誌結構儲存系統300的性能。後端資料管理子系統330可以根據不同的需求來處理和重組資料，例如，從而提高日誌結構儲存系統300的性能並降低其成本。在一些實施例中，後端資料管理子系統330可以獨立於前端I/O子系統310來管理資料。例如，後端資料管理子系統330可以在後端對密封索引日誌文件或唯讀索引日誌文件以及資料日誌文件執行諸如分層、壓縮、抹除編碼、狀態快照、壓合和驗證的資料管理操作。在一些實施例中，後端資料管理子系統330可以實施流控制以最小化對前端I/O子系統310的前端I/O處理的影響。

在一些實施例中，後端資料管理子系統330的任務可以包括對儲存的資料進行重寫和與該重寫的資料相對應的索引的替換。在一些實施例中，後端資料管理子系統330可以在後端上自動確定是否需要重寫資料日誌文件。在一些實施例中，後端資料管理子系統330可以基於諸如分層、壓縮和抹除編碼的配置來確定重寫的放置。在一些實施例中，後端資料管理子系統330可以從一個或多個源資料日誌文件讀取資料，並將資料重寫到目標資料日誌文

件。在一些實施例中，當重寫完成時，後端資料管理子系統330可以將目標資料日誌文件設置為密封或不可變狀態，並產生對應的目標索引日誌文件。在一些實施例中，目標索引日誌文件可以包括可被安全刪除的資料日誌文件的列表，以及目標索引日誌文件所引用的資料日誌文件。在一些實施例中，後端資料管理子系統330不回收仍然可以由前端I/O子系統310的實時實例使用的舊資料日誌文件。

在一些實施例中，後端資料管理子系統330可以處理根據前端I/O子系統310的I/O操作產生的唯讀索引日誌文件和對應的唯讀資料日誌文件。在一些實施例中，後端資料管理子系統330可以分析索引日誌文件並確定例如資料的熱、暖或冷的水平、資料量、垃圾率及/或碎片量。在一些實施例中，基於垃圾率、磁碟使用率及/或系統請求，後端資料管理子系統330可以執行以下一項或多項任務：

(1)資料分層。例如，當儲存媒介使用率接近安全上限時，可能需要將資料遷移到下一層或更低層儲存設備的儲存媒介中。

(2)資料壓縮。例如，當儲存媒介使用率接近安全上限時，可能需要壓縮資料文件。

(3)抹除編碼(EC)。例如，當儲存媒介使用率接近安全上限時，可能需要透過抹除編碼來釋放儲存空間。

(4)狀態快照。例如，當存在狀態修改(例如，在資料

刪除後回收儲存空間)時，可以對區塊鏈狀態執行快照。

(5)資料壓合。例如，如果資料日誌文件中的垃圾或碎片增長到一定大小從而明顯影響日誌結構儲存系統300的性能，則可能需要清理垃圾或碎片。

(6)驗證。例如，可以定期或按需在儲存媒介上對資料執行循環冗餘校驗(CRC)。

資料分層：

在一些實施例中，對於需要相對較高性能的寫入請求，可以將寫入請求寫入到多個不同儲存設備中的更快的儲存設備(例如，SSD雲端磁碟、ESSD雲端磁碟、NVME等)。對於需要較低性能以換取較低開銷的寫入請求，可以將寫入請求寫入儲存設備媒介(例如，NAS等)。在一些實施例中，後端資料管理子系統330可以使用一組混合的慢速和快速儲存設備來進行資料分層和資料遷移。例如，區塊鏈網路產生的新區塊資料的存取頻率可能比舊區塊資料相對更高，並且新區塊資料可以儲存在更快的儲存設備中。在一些實施例中，可以將存取頻率最高的新區塊資料的一部分儲存在記憶體快取(例如，記憶體快取313)及/或高速的磁碟快取(例如，區塊快取317)中。

在一些實施例中，分散式帳本系統和基於區塊鏈的中心化帳本系統均具有強的冷熱特性，這使其適於分層儲存。例如，諸如多層儲存子系統320的分層儲存系統可用於包括以下一項或多項特徵：(1)具有較小儲存空間的快速儲存媒介和具有大儲存空間的慢速儲存媒介的組合提高

了空間使用率而不犧牲性能；(2)支援冷遷移(例如，冷資料自動從快速媒介遷移到慢速媒介)和預熱(例如，資料從慢速媒介遷移到快速媒介)；(3)可擴展性，以在規模增加時減少維護成本；(4)支援基於用戶需求的靈活配置；(5)支援多媒體儲存池；或(6)快速遷移到新儲存媒介。

圖4是示出根據本文實施例的分層儲存系統400的示例的圖。在一些實施例中，分層儲存系統可以例如基於儲存設備的存取速度來包括多個級或多個層的儲存設備。例如，參考圖4，用於分層的多個儲存設備可以分為四個層或級，包括熱、暖、冷和用於基於日誌文件的熱和冷特性來儲存日誌文件的存檔。例如，分層儲存系統400的儲存設備可以分為用於分別儲存熱日誌文件410、暖日誌文件412、冷日誌文件414和存檔文件416的四個層或級。

在一些實施例中，儲存設備的每個層或級可以被視為虛擬池，並且每個池可以支援多個實體或虛擬文件系統(也稱為儲存設備)。例如，分層儲存系統400可以包括第一級池402、第二級池404、第三級池406以及第四級池408。在一些實施例中，池中支援的文件系統可以包括以下短期文件系統中的一個或多個：雲端磁碟(例如，安裝在ext4/xfs文件系統上的虛擬機(VM)的區塊設備)；NAS(例如，具有posix介面的nfs文件系統)； OSS低頻(適用於像是虛擬文件系統、軟體開發套件(SDK)系統、代表性狀態傳輸(REST)介面等格式)；和OSS存檔(適用於像是虛擬文件系統、SDK系統、REST介面等格式)。

例如，如圖4所示，第一級池402可以包括雲端儲存系統(例如，多層儲存子系統320)中儲存熱日誌文件410的ESSD和SSD設備。第二級池404可以包括雲端儲存系統中儲存暖日誌文件412的NAS設備和雲端磁碟。第三級池406可包括雲端儲存系統中儲存冷日誌文件414的OSS低頻設備。第四級池408可以包括雲端儲存系統中儲存存檔文件416的OSS存檔設備。

在一些實施例中，文件系統還可以包括長期文件系統，諸如自建分散式系統(例如，由基於區塊鏈的帳本系統的客戶端節點構建的分散式儲存系統340)。例如，第一級池402還可以包括由區塊鏈網路的客戶端節點產生的分散式儲存系統的儲存熱日誌文件410的NVME設備(例如，作為分散式儲存系統340的一部分)。第二級池404還可包括分散式儲存系統的儲存熱日誌文件412的SSD設備。第三級池406還可包括分散式儲存系統的儲存冷日誌文件414的HDD設備。第四級池408還可以包括分散式儲存系統的儲存存檔文件416的SMR設備。在一些實施例中，可以為所有文件系統提供與整個日誌結構儲存系統300的統一介面。

在一些實施例中，分層儲存系統400可以包括一個或多個子系統或組件，諸如(1)分層池管理器418、(2)遷移任務管理器420、(3)用於管理資料分層的遷移調度器422、或(4)服務品質(QoS)管理器423。在一些實施例中，每個管理器可以包括任何合適的計算元件(例如，處理器、記憶

體等中的一個或多個)以執行本文所述的功能。例如，這些管理器可以管理不同性能和成本的多個儲存設備之間的資料流，例如，透過利用不同儲存設備之間的性能和成本差異來提高整個日誌結構儲存系統的性能和效率。

在一些實施例中，分層池管理器418可以被配置為管理儲存設備的每個層。在一些實施例中，分層池管理器418可以執行以下功能中的一項或多項：管理多層儲存設備的儲存空間和壓力；提供指定層的文件創建、刪除和統計分析功能(例如，根據系統請求選擇儲存設備以創建資料日誌文件)；維護層文件映射表，該表指示資料文件的對應關係，它們在儲存設備的相應層中的儲存位置以及資料文件的熱度或冷度等。

在一些實施例中，遷移任務管理器420可以管理不同儲存設備之間的雙向資料遷移任務、管理任務生命週期、回呼結果、執行統計分析、執行遷移任務等等。

在一些實施例中，遷移調度器422可以支援可插拔遷移策略、管理資料遷移策略並提供資料創建/查詢/更新/刪除介面。在一些實施例中，遷移調度器422可以對遷移任務執行調度管理，以實現對遷移任務的有效流控制。在一些實施例中，遷移調度器422可以在後端對資料日誌文件進行打分或以其他方式分配相應得分，並且根據得分排名和遷移策略來產生資料日誌文件的遷移任務。在一些實施例中，可以根據評分公式對資料日誌文件進行打分，該評分公式考慮了儲存設備的層、存取頻率、原始資料創建時

間、遷移成本及/或其他因素。在一些實施例中，遷移調度器422可以與分層儲存系統400的其他子系統或組件一起工作以快速驗證不同的遷移策略。

在一些實施例中，可以根據預定的資料遷移策略自動執行資料遷移。例如，可以根據預定的評分方案對高速儲存設備中的不同資料進行打分，並基於不同資料的相應得分在後端將其遷移到低速設備中，以釋放快取空間。在一些實施例中，在某些應用中，低速設備中的一些資料可以被確定為熱資料。可以將熱資料首先保存在磁碟快取中，如果資料日誌文件的得分滿足要求，則可以將其遷移到高速設備。在一些實施例中，在資料文件從源儲存設備遷移到目標儲存設備之後，原始資料可以在源儲存設備中被刪除或可以不被刪除。例如，如果目標儲存設備是頂層儲存設備，則不需要刪除磁碟快取中的資料日誌文件，而是可以自動將其替換為其他資料。

在一些實施例中，QoS管理器423可以被配置為管理分層儲存系統400的資料流或其他性能度量以改善QoS。例如，在一些情況下，對高速儲存設備突發的I/O寫入可能導致較高層中的高速儲存設備被大量佔用或使用。在一些實施例中，QoS管理器423可以以高使用率(例如85%或另一閾值)來控制進入儲存池的資料流，以避免儲存池被過快填滿。流控制可以防止分層儲存系統400的性能下降，並可以釋放儲存空間以進行資料遷移。為了提高資料遷移的效率，同時減少對前端I/O操作(例如，透過前端I/O

子系統310)的影響，可以在後端(例如，透過後端資料管理子系統330)執行流控制資料遷移。在一些實施例中，遷移速度可以與儲存設備的使用率正相關。例如，如果儲存設備的使用率低，則可以減少流控制，以避免對前端I/O產生過多影響。如果儲存設備的使用率高，則可以取消流控制以加速資料遷移。

在一些情況下，高速儲存設備的使用率可能已滿，並且前端I/O操作可能受到嚴重限制。在一些實施例中，可以將資料直接寫入較低層儲存設備，而無需將資料從較高層儲存設備遷移到較低層儲存設備。例如，如果圖3中的第一層儲存設備350已滿或達到使用率閾值，則可以將資料直接寫入第二層儲存設備360。在一些實施例中，像區塊鏈區塊資料這樣的大(例如，具有大於閾值的大小)的資料可以被直接寫入低層儲存設備中的資料日誌文件中，以節省由於資料遷移引起的成本。

在一些實施例中，為了進一步減少由於資料遷移引起的網路資源、硬碟吞吐量、儲存空間和其他資源的消耗，並且為了減小對前端I/O操作的影響，當資料被遷移到低層儲存設備時可以透過預設執行壓縮和抹除編碼。

與高速儲存設備相比，低速或存檔儲存設備的性能相對較差。通常，大部分資料最終會寫入儲存設備的低速層。根據資料的熱特性和冷特性在高速儲存設備上快取熱資料並將其遷移到高速儲存設備有助於提高讀取性能。在一些實施例中，可以實施以下兩種或更多種快取以提高讀

取性能：(1)記憶體快取(例如，最近最少使用(LRU)快取424)；及(2)快速磁碟快取(例如，高速儲存設備上的最不常用(LFU)磁碟快取426)。在一些實施例中，可以例如以數百MB到幾GB來動態地配置記憶體快取424的總大小。類似地，可以例如以1GB到數十GB動態地配置快速磁碟快取的總大小。

在一些實施例中，可以將已經頻繁被存取的一些歷史資料，例如區塊鏈的創世區塊，放置在快速儲存設備的LFU快取中。

資料壓縮：

對於分散式帳本系統和基於區塊鏈的中心化帳本系統，對資料區塊的壓縮可以有效降低成本並提高日誌結構儲存系統的性能。日誌結構由於其固有的特性和特徵，可以有助於日誌結構儲存系統中的壓縮。

在一些實施例中，寫入前端上的資料可以不被壓縮，並且可以例如透過將資料追加在資料日誌文件中而直接寫入到高速儲存設備(例如，SSD雲端磁碟)。在一些實施例中，當資料日誌文件達到一定大小時，可以將其設置為不可變的。在一些實施例中，後端資料管理子系統330可以在後端上壓縮原始資料日誌文件，並將原始資料日誌文件替換為壓縮後的資料文件。這樣，由於壓縮操作是在後端執行的，因此可以減少或最小化壓縮操作對前端I/O操作的影響。

通常，在確定資料日誌文件的壓縮大小或容量時，可

能需要考慮壓縮和讀取放大的有效性並達到平衡。例如，在一些情況下，如果資料日誌文件的壓縮大小或容量太小(例如，小於4 KB)，則由於壓縮而節省的空間可能會受到限制，並且壓縮性能可能不是最佳的。另一方面，如果資料日誌文件的壓縮大小或容量太大，則讀取放大也可能變得更大(例如，要讀取交易條目，則包括交易條目的整個壓縮資料日誌文件需要首先解壓縮)。在一些實施例中，資料日誌文件的壓縮大小可以被設置為16KB-128KB。

在一些實施例中，壓縮資料日誌文件可以包括多個記錄，其中每個記錄可以包括壓縮頭和壓縮資料體。在一些實施例中，壓縮資料的元資料可以包括版本資訊、壓縮算法、長度和CRC等。

對於加密的資料，加密本身的隨機性會使資料壓縮的性能不理想。因此，在一些實施例中，對於需要加密的資料(例如在可信執行環境(TEE)中)，可以在加密之前或解密之後對資料進行壓縮。

在一些實施例中，對於壓縮日誌文件，基本索引映射可以對壓縮日誌文件的實體資料大小進行編碼、修改並記錄相應的索引，並且記錄日誌文件的文件ID、日誌文件的偏移量和日誌文件的壓縮資料大小。

抹除編碼

在一些實施例中，後端資料管理子系統330可以對資料日誌文件中的資料執行抹除編碼。例如，後端資料管理子系統330可以在後端使用抹除編碼將進入的資料寫入資

料日誌文件。

對於分散式帳本系統，為了在區塊鏈網路的共識節點之間實現拜占庭容錯日誌文件層，可以執行抹除編碼以減少儲存在分散式帳本系統的每個共識節點上的冷資料量。例如，對於4個共識節點，可以在執行抹除編碼之前寫入4個資料副本。在執行抹除編碼(例如，8+3抹除編碼方案)之後，這4個節點可以儲存少於2個資料副本(例如，1.375個資料副本)。

對於基於區塊鏈的中心化帳本系統，中心化結構不需要由於多個節點的備份而導致的資料冗餘。在一些實施例中，可以在基於區塊鏈的中心化帳本系統中的分層儲存系統中執行抹除編碼，以減少資料冗餘，例如，在頂層儲存設備或分散式儲存系統中的資料備份中。

資料壓合

在一些實施例中，交易資料、區塊資料和歷史狀態資料是僅追加的，並且不能被刪除或覆蓋，因此不對這些資料執行壓合。在一些實施例中，可以使用資料壓合來處理當前狀態資料。資料壓合通常包括垃圾收集和資料碎片整理。

在一些實施例中，後端資料管理子系統330可以根據資料日誌文件各自的垃圾率對它們進行排序，並且以從高垃圾率到低垃圾率的降序排列它們。在一些實施例中，後端資料管理子系統330可以重寫具有相對較高的垃圾率的資料日誌文件。例如，後端資料管理子系統330可以重寫

垃圾率超過預定閾值的資料日誌文件。在一些實施例中，資料日誌文件創建得越早，資料日誌文件中的資料就越有可能被覆蓋，這意味著較早的資料日誌文件的垃圾率可能高於新資料日誌文件的垃圾率。

在一些實施例中，後端資料管理子系統330可以實施垃圾回收機制，該垃圾回收機制可以為每次重寫設置最大資料量。在一些實施例中，例如可以由前端I/O子系統310的多個實時實例流並行地執行多個回收程式，以提高垃圾收集的整體效率。

在一些實施例中，前端I/O子系統310的實時實例流可以獲得垃圾率，並將獲得的垃圾率報告給後端資料管理子系統330，後端資料管理子系統330可以確定適當或最佳的流來重寫資料。

碎片整理通常是以下處理：定位儲存在儲存設備上的不連續資料碎片，然後重新排列這些碎片並將其還原為更少的碎片或整個文件。碎片整理可以減少資料存取時間，並可以更有效率地使用儲存設備。在一些實施例中，後端資料管理子系統330可以週期性地、不時地或根據請求執行碎片整理。

在一些實施例中，對於某些類型的資料，例如世界狀態資料或狀態對象資料，這些資料的鍵具有一定水平的雜湊特性。如果鍵具有前綴(例如，不同的狀態對象具有不同的前綴)，則透過將資料放入同一文件或相鄰文件中來對此類資料執行壓合可以提高讀取性能。

狀態快照

狀態快照可以捕獲系統(例如，分散式帳本系統)在特定時間點的狀態。在一些實施例中，後端資料管理子系統330可以執行狀態快照操作以產生和儲存日誌結構儲存系統300的狀態資料。在一些實施例中，狀態資料可以包括歷史狀態資料和當前狀態資料。歷史狀態資料可以包括分散式帳本系統的歷史狀態以用於回溯，而當前狀態資料可以包括分散式帳本系統的最新狀態資料。隨著時間的流逝，歷史狀態資料的大小可能會變大，並佔用大量儲存空間。在一些實施例中，為了改善歷史資料回溯和儲存空間使用率，後端資料管理子系統330可以對當前狀態資料執行快照操作。

日誌結構儲存系統300的日誌結構設計可以有助於快照操作並提高日誌結構儲存系統300的性能和效率。在一些實施例中，可以基於寫時重定向(ROW)方法來執行快照操作，該方法為與快照相對應的資料集提供了高效索引。

在一些實施例中，日誌結構儲存系統300的快照功能可以支援快閃記憶體創建(例如，秒級)和回滾，這可能對前端I/O操作僅具有有限或最小的影響。在一些實施例中，後端資料管理子系統330可以創建資料日誌文件和索引日誌文件的硬鏈接以避免資料複製。

在一些實施例中，當來自寫入請求的資料被儲存到資料日誌文件時，可以產生記錄以包括指示快照版本的快照識別符(ID)。在一些實施例中，後端資料管理子系統330

可以響應於接收到狀態快照請求而執行以下操作中的一個或多個：

(1)寫入與快照創建相對應的操作日誌(op log)；

(2)將快照版本增加1；

(3)將所有新快照寫入請求寫入新記憶體內增量索引映射中(重定向)；

(4)與舊快照關聯的所有寫入請求完成後，將索引刷新到當前索引日誌文件，對所有索引文件執行壓合，將索引文件合併為單個索引日誌文件，並將合併的單個索引日誌文件設置為密封狀態(資料日誌文件在壓合處理期間也被密封)；

(5)基於新快照版本創建新索引文件；以及

(6)創建與快照相對應的目錄，並為與快照相關聯的資料日誌文件和索引日誌文件創建到該目錄的硬鏈接。

在一些實施例中，後端資料管理子系統330可以在後端上執行壓合以恢復被刪除的快照。

在一些實施例中，如果需要快照上傳，則可以維護快照的資料結構(例如具有1位元的點陣圖表示資料範圍)。例如，在創建快照時創建的索引日誌文件中，可以將與快照相對應的點陣圖設置為全0。在接收到寫入請求後，點陣圖可以更新為1，指示該快照版本中的資料被修改。

在一些實施例中，快照版本號可以對應於索引日誌文件，該快照版本號指示與索引日誌文件中的所有索引相對應的寫入請求。

驗證

在一些實施例中，後端資料管理子系統330可以對記錄在日誌文件中的資料執行CRC校驗。在一些實施例中，後端資料管理子系統330可以週期性地、不時地或根據請求執行CRC校驗。

在一些實施例中，當將由後端資料管理子系統330產生的索引文件導入到前端I/O子系統310的實時實例流中時，實時實例流的記憶體內索引映射可以比後端資料管理子系統330產生的索引文件更新，並且可以包括新資料區塊和舊資料區塊的位置資訊。在一些實施例中，實時實例流可以遍歷記憶體內基本映射，替換相應的索引條目，然後產生沒有引用舊資料日誌文件的新索引日誌文件。然後，實時實例流可以安全地刪除舊資料日誌文件和索引日誌文件。

在一些實施例中，在日誌結構儲存框架(例如，日誌結構儲存系統300)中，流可以被用於作為處理I/O請求的處理引擎、組件、單元或模組來操作。每個流可以透過不同的配置適應不同的業務場景。在一些實施例中，流可以由與軟體耦接的一個或多個處理器來實施以執行諸如管理資料日誌文件、索引日誌文件、清單文件、請求佇列等的操作。在一些實施例中，實時流可以指代處理前端I/O子系統310的前端I/O操作的實時實例。在一些實施例中，可以存在由後端資料管理子系統330在後端管理由實時流寫入的資料的對應dredger流。

在一些實施例中，流可以包括管理介面，該管理介面允許針對不同類型的資料的例如快照、統計和故障恢復的不同操作的不同配置。例如，用於處理區塊資料、狀態資料和交易資料的流可以根據區塊資料、狀態資料和交易資料的各自的特性採用不同的配置。例如，可以將與區塊相對應的流配置為具有分層及/或壓縮功能，但不具有壓合、快照或表格功能。

在一些實施例中，可以透過對應自定義的或以其他方式配置的流來處理不同類型的資料。例如，可以透過與區塊相對應的流來處理寫入區塊的請求。

在一些實施例中，可以將多個流組合成束(bundle)以提供適合於分散式帳本系統及/或基於區塊鏈的中心化帳本系統的特定應用的靈活的實施方式。所描述的技術可以支援分散式帳本系統(例如，區塊鏈網路)及/或基於區塊鏈的中心化帳本系統中的服務。在一些實施例中，兩種系統可以具有根據兩種日誌結構儲存系統300的需要自定義的或以其他方式配置的不同的流。例如，分散式帳本系統可以具有四種資料：交易共識日誌、區塊、狀態和索引。因此，可以將四種流配置為分別處理四種資料。基於區塊鏈的中心化帳本系統可以具有三種資料：交易、區塊和索引，而沒有狀態(或複雜合約狀態)。因此，可以將三種流配置為分別處理三種資料。

在一些實施例中，每種流可以被分別配置為處理不同類型的資料。例如，區塊、交易共識日誌、索引不需要快

照。因此，用於處理區塊、交易共識日誌和索引的流不需要配置快照功能。另一方面，狀態資料流可以配置有快照功能。作為另一示例，索引資料相對較小，但是它需要良好的性能，並且不需要分層分級儲存。長期操作和大量區塊資料可能需要分層分級儲存、共享儲存和抹除編碼。

在一些實施例中，分散式帳本系統和基於區塊鏈的中心化帳本系統可以對流具有不同的要求，以執行諸如分層、壓縮、抹除編碼、狀態快照、壓合和資料驗證等操作。

表1提供了不同場景下的配置示例。如圖所示，“兩者”表示可以針對分散式帳本系統和基於區塊鏈的中心化帳本系統都執行對特定類型資料的操作。“DLS”表示可以只針對分散式帳本系統執行對特定類型資料的操作。“都不”表示可以針對分散式帳本系統和基於區塊鏈的中心化帳本系統都不執行對特定類型資料的操作。

表格1

項目	交易	區塊	當前狀態	歷史狀態	額外	共識日誌
分層	兩者	兩者	DLS	DLS	都不	都不
壓縮	兩者	兩者	DLS	DLS	都不	都不
抹除編碼	DLS	DLS	DLS	DLS	都不	都不
快照	都不	都不	DLS	都不	都不	都不
壓合	都不	都不	DLS	都不	都不	DLS
驗證	兩者	兩者	兩者	兩者	兩者	都不

例如，如表1所示，可以針對分散式帳本系統和基於

區塊鏈的中心化帳本系統執行對交易資料及/或區塊資料的分層操作。可以只針對分散式帳本系統執行對當前狀態資料及/或歷史狀態的分層操作。可以針對分散式帳本系統和基於區塊鏈的中心化帳本系統都不執行對交易資料的快照操作。

在一些實施例中，日誌結構儲存系統採用基於每個執行緒一個循環一個佇列和併發的多執行緒全異步機制，從而提供了高效的異步模式和便捷的併發同步程式化模式。在一些實施例中，不同的流可以並行處理不同類型的資料。例如，為區塊資料配置的流可以將區塊資料寫入被分配以儲存區塊資料的資料日誌文件中，而為交易資料配置的流可以從包括交易資料的資料日誌文件中讀取特定請求交易資料。

圖5是示出根據本文實施例的用於執行日誌結構儲存系統的寫入操作的處理500的流程圖。在一些實施例中，處理500的一些或全部操作可以是由前端I/O子系統(例如，圖3的前端I/O子系統310)執行的寫入程式的示例。為了方便，處理500將被描述為由圖3的前端I/O子系統310系統執行。然而，處理500可以由位於一個或多個位置的一個或多個電腦的系統執行，並且根據本文被適當地程式化。例如，適當程式化的資料處理和儲存系統(例如，圖3的日誌結構儲存系統300)可以執行處理500。

在502，在資料儲存系統(例如，日誌結構儲存系統300)中維護資料日誌文件(例如，資料日誌文件390、

362、364、366、372、374或376)。在一些實施例中，資料日誌文件可以儲存包括交易資料、區塊資料、狀態資料和自描述元資料的資料。例如，資料日誌文件可以儲存由區塊鏈網路產生的區塊鏈資料，包括區塊資料、交易資料及/或狀態資料。在一些實施例中，資料日誌文件中的元資料可以包括描述資料區塊的元資料，諸如交易雜湊值和序列值、區塊雜湊值和區塊號、快照版本號、循環冗餘校驗(CRC)碼、加密資訊等。在一些實施例中，一個資料日誌文件儲存單一類型的區塊鏈資料，因此多種區塊鏈資料不混合在單個資料文件中。例如，資料儲存系統可以維護用於交易資料的資料日誌文件、用於區塊資料的資料日誌文件和用於狀態資料的資料日誌文件中的一個或多個。在一些實施例中，資料日誌文件可以是僅追加文件。在一些實施例中，資料日誌文件不儲存索引資訊。在一些實施例中，資料日誌文件可以被儲存在多層儲存子系統(例如，多層儲存子系統320)中。

在504，資料儲存系統的前端I/O子系統(例如，前端I/O子系統310)接收用於將資料寫入資料儲存系統的寫入請求。在一些實施例中，前端I/O子系統310可以處理寫入操作，包括對儲存在日誌結構儲存系統300上的資料的修改。在一些實施例中，對資料的修改由前端I/O子系統310處理，從而不覆蓋原始資料。取而代之的，可以透過以日誌形式向資料日誌文件添加或追加資料來處理修改。

在506，前端I/O子系統310將資料追加到資料日誌文

件。在一些實施例中，可以基於資料寫入或追加操作來持續更新資料日誌文件。在一些實施例中，資料日誌文件可以具有在512MB和2GB之間的可配置的最大長度，或者取決於儲存系統的需求或應用的其他大小。

在508，前端I/O子系統310確定是否滿足用於產生新資料日誌文件的條件。在一些實施例中，前端I/O子系統310可以確定資料日誌文件是否已經達到預定的最大長度或大小。如果確定資料日誌文件已經達到預定的最大長度或大小，則前端I/O子系統310可以確定滿足了用於產生新資料日誌文件的條件。如果確定滿足了用於產生新資料日誌文件的條件，則處理進行到步驟510。如果確定沒有滿足用於產生新資料日誌文件的條件，則處理返回到步驟504。

在510，如果確定滿足了用於產生新資料日誌文件的條件，則前端I/O子系統310密封資料日誌文件。在一些實施例中，如果確定滿足了用於產生新資料日誌文件的條件(例如，資料日誌文件已達到最大長度或大小)，則前端I/O子系統310可以將資料日誌文件設置為密封的、不可變的或唯讀狀態。

在512，前端I/O子系統310產生新資料日誌文件。在一些實施例中，新資料日誌文件也可以是僅追加的，並儲存在多層儲存子系統320中。

在一些實施例中，前端I/O子系統310可以確定在寫入請求中被請求寫入的資料的類型(例如，交易資料、區塊

資料、狀態資料)。響應於該確定，前端I/O子系統310將資料追加到與資料的類型相對應的資料日誌文件中。在一些實施例中，前端I/O子系統310可以使用與資料類型相對應的相應處理引擎來執行處理500中的一些或全部。

例如，響應於確定資料是交易資料，前端I/O子系統310使用針對處理交易資料指定的處理引擎，以將該資料追加到用於交易資料的資料日誌文件中。在一些實施例中，響應於確定資料是區塊資料，前端I/O子系統310使用針對處理區塊資料指定的處理引擎，以將該資料追加到用於區塊資料的資料日誌文件中。在一些實施例中，響應於確定資料是狀態資料，前端I/O子系統310使用針對處理狀態資料指定的處理引擎，以將該資料追加到用於狀態資料的資料日誌文件中。

圖6是示出根據本文實施例的用於產生與日誌結構儲存系統的寫入操作有關的索引的處理600的流程圖。在一些實施例中，處理600的一些或全部操作可以是由前端I/O子系統(例如，圖3的前端I/O子系統310)執行的寫入程式的示例。為了方便，處理600將被描述為由圖3的前端I/O子系統310執行。然而，處理600可以由位於一個或多個位置的一個或多個電腦的系統執行，並且根據本文被適當地程式化。例如，適當程式化的資料處理和儲存系統(例如，圖3的日誌結構儲存系統300)可以執行處理600。

在602，成功將資料寫入儲存系統(例如，日誌結構儲存系統300)。在一些實施例中，資料儲存系統的前端I/O

子系統(例如，前端I/O子系統310)可以將資料區塊寫入儲存在資料儲存系統的多層儲存子系統(例如，多層儲存子系統320)的資料日誌文件中。

在604，前端I/O子系統310產生指示資料在日誌結構儲存系統300中的實體儲存位置的索引。在一些實施例中，索引資料可以包括：指示從區塊雜湊值到區塊號的對應關係的索引、指示從區塊雜湊值到儲存位置的對應關係的索引、指示從交易雜湊值到交易的對應關係的索引、和指示從收據雜湊值到收據的對應關係的索引。在一些實施例中，用於基於區塊鏈的中心化帳本系統的索引資料可以包括指示從序列到交易儲存位置的對應關係的索引或指示從時序到交易雜湊值的對應關係的索引。

在606，前端I/O子系統將索引保存到前端I/O子系統310的記憶體(例如，記憶體315)中的增量索引映射(例如，增量索引映射312)中。在一些實施例中，前端I/O子系統310可以包括儲存在記憶體315中的多個記憶體內索引映射。在一些實施例中，記憶體內索引映射可以分為唯讀基本索引映射316和讀寫增量索引映射312。在一些實施例中，基本索引映射316可以儲存冷資料(例如，舊資料及/或較不頻繁被存取的資料)的索引，而增量索引映射312可以儲存新寫入的資料的索引。

在608，前端I/O子系統310確定是否發生觸發事件。觸發事件可以包括導致密封當前增量索引映射並產生新增量索引映射的一個或多個事件。觸發事件可包括，例如，

當前增量索引映射的大小達到閾值、記憶體315的儲存使用率達到閾值、或指定的時間到來(例如，日誌結構儲存系統300可以定期密封增量索引映射)。如果確定觸發事件發生，則處理進行到步驟610。如果確定觸發事件未發生，則處理返回到步驟602。

在610，如果確定觸發事件發生，則前端I/O子系統310將增量索引映射312設置為不可變的。在一些實施例中，前端I/O子系統可以將記憶體315中的增量索引映射312設置為不可變索引映射314，將它們刷新到索引日誌文件(例如，索引日誌文件380)，並創建新增量索引映射312以接收根據新寫入請求產生的索引。

在612，在儲存系統300中維護索引日誌文件380。在一些實施例中，可以將新寫入的交易資料和區塊資料的索引添加到索引映射312和316以及索引日誌文件390，但是可以不修改現有交易資料和區塊資料的索引。在一些實施例中，索引日誌文件390可以與資料日誌文件一起儲存在多層儲存子系統320中。

在614，前端I/O子系統310例如將增量索引映射312刷新到索引日誌文件380中，以釋放增量索引映射312所使用的記憶體。在一些實施例中，前端I/O子系統310可以創建新增量索引映射312以接收根據新請求產生的索引。在一些實施例中，前端I/O子系統310可以組合增量索引映射312和基本索引映射316，並產生新基本索引映射316，並將產生的基本索引映射316刷新到索引日誌文件380。

在一些實施例中，在616，前端I/O子系統310將熱資料的索引保存在記憶體快取(例如，記憶體快取313)中。例如，如果將特定資料確定為具有頻繁被存取可能性的熱資料，則可以將資料的索引保存到記憶體快取中以提高讀取速度。

在618，前端I/O子系統310確定是否滿足用於產生新的索引日誌文件380的條件。在一些實施例中，用於產生新索引日誌文件的條件可以包括索引日誌文件380的最大長度或大小。在一些實施例中，用於產生新索引日誌文件380的條件可以包括由前端I/O子系統執行的寫入操作的批次數。例如，在一些實施例中，可以透過批量處理寫入操作的索引來更新索引日誌文件380。在一些實施例中，當已經針對特定批次數的寫入操作更新了索引日誌文件380時，可以將索引日誌文件380密封或設置為唯讀，並且可以創建新的索引日誌文件380以寫入新資料。如果確定滿足了用於產生新的索引日誌文件380的條件，則處理進行到步驟620。

在620，如果確定滿足了用於產生新的索引日誌文件380的條件，則前端I/O子系統310密封索引日誌文件380。例如，當索引日誌文件380已經達到最大長度或大小，或者已經針對特定批次數的寫入操作進行了更新時，可以將索引日誌文件380密封或設置為唯讀。

在622，前端I/O子系統310在密封舊索引日誌文件380之後，產生新的索引日誌文件380，以儲存後續的索引資

料。

圖7是示出了根據本文實施例的用於執行日誌結構儲存系統的讀取操作的處理700的流程圖。在一些實施例中，處理700的一些或全部操作可以是由前端I/O子系統(例如，圖3的前端I/O子系統310)執行的讀取程序的示例。為了方便，處理700將被描述為由前端I/O子系統310執行。然而，處理700可以由位於一個或多個位置的一個或多個電腦的系統執行，並且根據本文被適當地程式化。例如，適當程式化的資料處理和儲存系統(例如，圖3的日誌結構儲存系統300)可以執行處理700。

在702處，儲存系統(例如，日誌結構儲存系統300)的前端I/O子系統(例如，前端I/O子系統310)接收用於從儲存系統讀取資料的讀取請求。

在704，前端I/O子系統310在前端I/O子系統310的記憶體(例如，記憶體315)中的增量索引映射(例如，增量索引映射312)中搜索與資料相對應的索引。在一些實施例中，與資料相對應的索引可以包括資料的實體位置資訊。在一些實施例中，前端I/O子系統310的記憶體315可以儲存包括唯讀基本索引映射316和讀寫增量索引映射312的多個記憶體內索引映射。

在706，前端I/O子系統310確定是否在增量索引映射312中找到了與資料相對應的索引。如果在增量索引映射312中找到了與資料相對應的索引，則處理進行到步驟708，其中，前端I/O子系統310可以基於由索引指示的實

體位置來定位資料。如果在增量索引映射312中未找到與資料相對應的索引，則處理進行到步驟710。

在710，如果確定在增量索引映射312中未找到與資料相對應的索引，則前端I/O子系統310在記憶體315中的基本索引映射316中搜索與資料相對應的索引。

在712，前端I/O子系統310確定是否在基本索引映射316中找到了與資料相對應的索引。如果確定在基本索引映射316中找到了與資料相對應的索引，則處理進行到步驟714，其中，前端I/O子系統310可以基於由索引指示的實體位置資訊來定位資料。如果在基本索引映射316中未找到與資料相對應的索引，則處理進行到步驟716。

在716，如果確定在基本索引映射316中未找到與資料相對應的索引，則前端I/O子系統310在磁碟儲存設備中的索引日誌文件(例如，索引日誌文件380)中搜索與資料相對應的索引。例如，前端I/O子系統310可以在儲存在儲存系統300的多層儲存子系統(例如，多層儲存子系統320)中的索引日誌文件380中搜索與資料相對應的索引。

在一些實施例中，前端I/O子系統310可以確定在讀取請求中被請求讀取的資料的類型(例如，交易資料、區塊資料、狀態資料)。響應於該確定，前端I/O子系統310可以使用與資料類型相對應的相應處理引擎來執行處理700中的一些或全部。

圖8是示出根據本文實施例的用於改善日誌結構儲存系統的讀取操作的處理800的流程圖。在一些實施例中，

處理800的一些或全部操作可以是由日誌結構儲存系統(例如，圖3的日誌結構儲存系統300)執行的I/O操作的示例。為了方便，處理800將被描述為由日誌結構儲存系統300執行。然而，處理800可以由位於一個或多個位置的一個或多個電腦的系統執行，並且根據本文被適當地程式化。例如，適當程式化的資料處理和儲存系統可以執行處理800。

在802，儲存系統(例如，日誌結構儲存系統300或分層儲存系統400)維護多層儲存設備(例如，多層儲存子系統320的儲存設備350、360和370)和一層或多層快取(例如，記憶體快取313和區塊快取317)。在一些實施例中，多層儲存設備是按基於一個或多個性能特性(例如，存取速度、存取頻寬或存取延遲)的分層結構分層的。例如，多層儲存設備可以包括具有不同存取速度並儲存具有不同特性的資料的多個儲存設備。例如，第一層儲存設備可以儲存比第二層儲存設備中儲存的區塊鏈資料更頻繁被存取的區塊鏈資料。

在804，例如前端I/O子系統(例如，前端I/O子系統310)或後端資料管理系統(例如，儲存系統300的後端資料管理子系統330)確定儲存在較低層儲存設備(例如，儲存設備350、360和370)中的資料日誌文件(例如，資料日誌文件362、364、366、372、374、376、390)中的資料對象是活躍資料對象。在一些實施例中，資料對象可以包括交易資料、區塊資料和狀態資料。在一些實施例中，例如，

如果資料對象最近已經被存取一定次數(例如，在預定時間窗內被存取一定次數)或者如果資料對象已經被識別為具有特定優先級，則可以基於一個或多個活躍性或熱度策略將資料對象確定為活躍資料對象。

在806，將資料對象寫入快取(例如，記憶體快取313和區塊快取317)。例如，前端I/O子系統310可以將資料對象寫入高速儲存媒介的記憶體快取313或磁碟區塊快取317中。

在808，產生指示資料對象在快取中的實體儲存位置的索引。在一些實施例中，可以使用LSM方法來組織記憶體315中的索引資料。

在810，可以將索引保存到記憶體315中的增量索引映射(例如，增量索引映射312)。在一些實施例中，記憶體315可以維護多個記憶體內索引映射，包括唯讀基本索引映射316和讀寫增量索引映射312。在一些實施例中，增量索引映射312可以被配置為儲存比儲存在基本索引映射316中的索引更頻繁被存取及/或更新的資料的索引。

在812，前端I/O子系統310接收針對資料對象的讀取請求。

在814，前端I/O子系統310在記憶體315中的增量索引映射312中搜索與資料對象相對應的索引。在一些實施例中，前端I/O子系統310可以首先搜索增量索引映射312。如果在增量索引映射312中未找到索引，則前端I/O子系統310可以隨後在基本索引映射316中搜索與資料相對應的索

引。

在816，前端I/O子系統310從快取返回資料對象，與需要從下一快取級、主記憶體或多層儲存子系統320的儲存設備的較低層獲取所請求的資料對象的情況相比，這可以提供對所請求的資料對象的更快存取。例如，如果前端I/O子系統310成功地識別出與增量索引映射312或基本索引映射316中的資料相對應的索引，則前端I/O子系統310可以使用該索引來識別資料在快取中的實體位置，並從快取中檢索資料。

在一些實施例中，前端I/O子系統310可以確定在讀取請求中被請求讀取的資料的類型(例如，交易資料、區塊資料、狀態資料)。響應於該確定，前端I/O子系統310可以使用與資料類型相對應的相應處理引擎來執行處理800中的一些或全部。

圖9是示出根據本文實施例的用於管理儲存在日誌結構儲存系統中的資料日誌文件的處理900的流程圖。在一些實施例中，處理900的一些或全部操作可以是由日誌結構儲存系統的後端資料管理系統(例如，圖3的日誌結構儲存系統300的後端資料管理子系統330)執行的重寫放置程序的示例。為了方便，處理900將被描述為由後端資料管理子系統330執行。然而，處理900可以由位於一個或多個位置的一個或多個電腦的系統執行，並且根據本文被適當地程式化。例如，適當程式化的資料處理和儲存系統(例如，圖3的日誌結構儲存系統300)可以執行處理900。

在902，後端資料管理系統(例如，後端資料管理子系統330)從索引日誌文件(例如，索引日誌文件380)確定儲存在儲存設備(例如，儲存設備350、360和370)中的資料日誌文件(例如，資料日誌文件390、362、364、366、372、374和376)的資訊。在一些實施例中，儲存設備中的資料日誌文件的資訊可以包括資料日誌文件的活躍性(例如，存取頻率)、大小、垃圾率或碎片水平中的一個或多個。

在904，後端資料管理系統330確定儲存設備的資訊。在一些實施例中，儲存設備的資訊可以包括儲存設備的使用率、垃圾率、碎片水平或輸入/輸出(I/O)請求中的一個或多個。

在906，後端資料管理系統330確定資料日誌文件是否需要重寫放置。在一些實施例中，後端資料管理子系統330可以基於儲存在儲存設備中的資料日誌文件的資訊及/或儲存設備的資訊來確定重寫放置。在一些實施例中，重寫放置可以包括分層、壓縮、抹除編碼、狀態快照、壓合或驗證中的至少一個。如果確定資料日誌文件需要重寫放置，則處理進行到步驟908。如果確定資料日誌文件不需要重寫放置，則處理返回到步驟902。

在908，後端資料管理系統330從源位置讀取資料日誌文件，並且如果確定資料日誌文件需要重寫放置，則將資料日誌文件重寫到目標位置。

在910，後端資料管理系統330將資料日誌文件密封在目標位置中。例如，後端資料管理系統330可以在重寫放

置完成之後將資料日誌文件設置為密封狀態或唯讀。

在912，後端資料管理系統330產生與目標位置中的資料日誌文件相對應的目標索引日誌文件。在一些實施例中，目標索引日誌文件可以包括可被安全刪除的資料日誌文件的列表，及/或目標索引日誌文件所引用的資料日誌文件的列表。

在914，後端資料管理系統330密封目標索引日誌文件。例如，後端資料管理系統330可以將目標索引日誌文件設置為不可變的或唯讀的。

在916，將目標索引日誌文件導入到記憶體中的可讀索引映射中。例如，可以將目標索引日誌文件導入到增量索引映射或基本索引映射，使得可以尋址或讀取目標位置中的資料日誌文件。

圖10是示出根據本文實施例的用於在日誌結構儲存系統中執行資料遷移的處理1000的流程圖。在一些實施例中，處理1000的一些或全部操作可以是由日誌結構儲存系統的後端資料管理系統(例如，圖3的日誌結構儲存系統300的後端資料管理子系統330)執行的分層/遷移程序的示例。為了方便，處理1000將被描述為由後端資料管理子系統330執行。然而，處理1000可以由位於一個或多個位置的一個或多個電腦的系統執行，並且根據本文被適當地程式化。例如，適當程式化的資料處理和儲存系統(例如，圖3的日誌結構儲存系統300)可以執行處理1000。

在1002，後端資料管理系統(例如，後端資料管理子

系統330)識別資料日誌文件(例如，資料日誌文件390、362、364、366、372、374和376)的一個或多個特性和儲存設備(例如，儲存設備350、360和370)的一個或多個特性。在一些實施例中，資料日誌文件的一個或多個特性可以包括資料日誌文件的資料類型(例如，區塊資料、狀態資料和交易資料)、創建時間、資料大小、活躍性、垃圾率或碎片水平等。在一些實施例中，儲存設備的一個或多個特性可以包括儲存設備的存取速度、存取頻寬、存取延遲、使用率、垃圾率、碎片水平或輸入/輸出(I/O)請求。

在1004，後端資料管理系統330基於所述特性確定資料日誌文件的遷移度量。在一些實施例中，後端資料管理系統330可以將得分分配給資料日誌文件，並根據得分排名和預定的遷移策略來產生遷移任務。在一些實施例中，可以根據評分公式對資料日誌文件進行打分或分配得分，該評分公式考慮了媒介水平、存取頻率、原始資料創建時間和遷移成本等。

在1006，後端資料管理系統330確定是否遷移資料日誌文件。例如，可以根據預定的評分方案對資料日誌文件進行評分。如果資料日誌文件的得分超過預定閾值，則後端資料管理系統330可以確定需要遷移資料日誌文件。如果確定需要遷移資料日誌文件，則處理進行到步驟1008。如果確定不需要遷移資料日誌文件，則處理返回到步驟1002。

在1008，如果確定需要遷移資料日誌文件，則後端資

料管理系統330將資料日誌文件從源位置遷移到目標儲存設備。在一些實施例中，可以根據預定的評分方案對高速儲存設備中的資料日誌文件進行打分，並基於得分將其遷移到低速儲存設備(例如，在對資料日誌文件的得分進行排序或排名之後)，以釋放儲存空間。在一些實施例中，低速儲存設備中儲存的資料日誌文件中的熱資料可以首先保存在磁碟快取中，如果資料日誌文件的得分達到預定閾值，則可以將其遷移到高速儲存設備中。

圖11是示出根據本文實施例的用於在日誌結構儲存系統中執行資料流控制的處理1100的流程圖。在一些實施例中，處理1100的一些或全部操作可以是由日誌結構儲存系統(例如，圖3的日誌結構儲存系統300)執行的流控制/優化程序的示例。為了方便，處理1100將被描述為由日誌結構儲存系統執行。然而，處理1100可以由位於一個或多個位置的一個或多個電腦的系統執行，並且根據本文被適當地程式化。例如，適當程式化的資料處理和儲存系統可以執行處理1100。

在1102，儲存系統(例如，日誌結構儲存系統300)維護多層儲存設備(例如，儲存設備350、360和370)。在一些實施例中，多層儲存設備是按基於一個或多個性能特性(例如，存取速度、存取頻寬或存取延遲)的分層結構來分層的。例如，多層儲存設備可以包括具有不同存取速度的多個儲存設備，並且可以儲存具有對應特性的資料(例如，第一層儲存設備可以儲存比儲存在第二層儲存設備中

的區塊鏈資料更頻繁被存取的區塊鏈資料)。

在一些實施例中，日誌結構儲存系統300可以將流控制策略分配給多層儲存設備。例如，日誌結構儲存系統300可以基於第一層儲存設備和第二層儲存設備的一個或多個特性(例如，存取速度、存取頻寬、存取延遲、使用率、垃圾率、碎片水平)，將第一流控制策略分配給第一層儲存設備，將第二流控制策略分配給第二層儲存設備。在一些實施例中，第一流控制策略可以包括以下一項或多項：將資料寫入第一層儲存設備的第一速度，或用於調整將資料寫入第一層儲存設備的第一速度的一個或多個第一閾值，並且第二流控制策略可以包括以下一項或多項：將資料寫入第一層儲存設備的第二速度，或用於調整將資料寫入第二層儲存設備的第二速度的一個或多個第二閾值。

在1104，日誌結構儲存系統300接收針對帳本資料的寫入請求。在一些實施例中，帳本資料可以包括例如交易資料、區塊資料和狀態資料等的區塊鏈資料。

在1106，日誌結構儲存系統300識別帳本資料的類型。例如，後端資料管理子系統330可以確定帳本資料是交易資料、區塊資料還是狀態資料。在1108，日誌結構儲存系統300確定帳本資料是否是區塊資料。在一些實施例中，區塊資料具有比其他類型的區塊鏈資料(例如，交易資料、狀態資料或索引資料)更大的大小，並且可能對日誌結構儲存系統300的I/O操作的吞吐量具有更大的影響。如果確定帳本資料是區塊資料，則處理進行到步驟1110，

其中後端資料管理子系統330將資料直接寫入第二層儲存設備(例如，儲存設備360)，並跳過第一層儲存設備，例如，以節省之後執行遷移的成本。在一些實施例中，第二層儲存設備位於比第一層儲存設備低的層。例如，第二層儲存設備的存取速度可以比第一層儲存設備的存取速度更低。在一些實施例中，第二層儲存設備的成本可以比第一層儲存設備的成本更低。在一些實施例中，第二層儲存設備的儲存空間可以比第一層儲存設備的儲存空間更大。如果確定帳本資料不是區塊資料，則處理進行到步驟1112。

在1112，如果確定帳本資料不是區塊資料，則日誌結構儲存系統300確定第一層儲存設備的使用率。

在1114，日誌結構儲存系統300確定使用率是否達到或超過預定閾值。在一些實施例中，預定閾值用於確定第一層儲存設備是否基本上已滿。例如，如果確定使用率達到或超過閾值(例如，85%)，則日誌結構儲存系統300可以確定第一層儲存設備基本上已滿。如果確定使用率達到或超過預定閾值，則處理進行到步驟1116，在步驟1116，將資料寫入第二層儲存設備。如果確定使用率低於預定閾值，則處理進行到步驟1118。

在1118，日誌結構儲存系統300在確定使用率低於預定閾值的情況下，將資料寫入第一層儲存設備。

在一些實施例中，日誌結構儲存系統300可以基於第一層儲存設備的使用率來調整將資料寫入第一層儲存設備的速度。例如，如果確定第一層儲存設備的使用率達到或

超過第一預定閾值(例如65%)，則日誌結構儲存系統300可以降低將資料寫入第一層儲存設備的速度。在一些實施例中，日誌結構儲存系統300可以基於第一層儲存設備的使用率來降低將資料寫入第一層儲存設備的速度。在一些實施例中，日誌結構儲存系統300可以隨著第一層儲存設備的使用率的增加來持續降低將資料寫入第一層儲存設備的速度。例如，當第一層儲存設備的使用率是第一值(例如，70%)時，日誌結構儲存系統300可以將資料寫入第一層儲存設備中的速度降低到第一速率(例如，500MB/s)，以及當第一層儲存設備的使用率為比第一值大的第二值(例如，75%)時，將資料寫入第一層儲存設備的速度降低到低於第一速率的第二速率(例如400MB/s)。

在一些實施例中，如果確定第一層儲存設備的使用率低於第二預定閾值(例如，35%)，則日誌結構儲存系統300可以提高將資料寫入第一層儲存設備的速度。在一些實施例中，日誌結構儲存系統300可以基於第一層儲存設備的使用率來提高將資料寫入第一層儲存設備的速度。在一些實施例中，日誌結構儲存系統300可以隨著第一層儲存設備的使用率的降低來持續提高將資料寫入第一層儲存設備的速度。例如，當第一層儲存設備的使用率是第三值(例如，30%)時，日誌結構儲存系統300可以將資料寫入第一層儲存設備中的速度提高到第三速率(例如，550MB/s)，以及當第一層儲存設備的使用率為比第三值小的第四值(例如，20%)時，將資料寫入第一層儲存設備的速度提高

到高於第三速率的第四速率(例如600MB/s)。

圖12是示出可根據本文實施例執行的處理1200的流程圖。為了方便，處理1200將被描述為由圖3的日誌結構儲存系統300執行。然而，處理1200可以由位於一個或多個位置的一個或多個電腦的系統執行，並且根據本文被適當地程式化。例如，適當程式化的資料處理和儲存系統可以執行處理1200。

在1202，儲存系統(例如，日誌結構儲存系統300)接收多個處理引擎的配置。在一些實施例中，所述配置可以根據例如表1的分散式帳本系統的多種資料類型中的每種資料類型的特性，配置用於處理這種資料類型的相應的處理引擎類型。在一些實施例中，儲存系統300可以包括被指定用於處理區塊資料的處理引擎類型；被指定用於處理交易資料的處理引擎類型；被指定用於處理狀態資料的處理引擎類型；被指定用於處理索引資料的處理引擎類型。

在一些實施例中，狀態資料可以包括當前狀態資料和歷史狀態資料，並且儲存系統300可以包括被指定用於處理當前狀態資料的處理引擎類型和被指定用於處理歷史狀態資料的處理引擎類型。

在1204，儲存系統300接收針對分散式帳本系統的資料的處理請求。在一些實施例中，分散式帳本系統的資料類型可以包括區塊資料、交易資料、狀態資料和索引資料。

在一些實施例中，儲存系統300可以接收針對分散式

帳本系統的資料的I/O請求。在一些實施例中，被指定用於處理分散式帳本系統的一種資料類型的相應的處理引擎類型可以包括被指定用於對分散式帳本系統的一種資料類型執行讀取或寫入操作的相應的I/O處理引擎類型。

在一些實施例中，儲存系統300可以接收針對分散式帳本系統的資料的資料管理請求。在一些實施例中，被指定用於處理一種資料類型的相應的處理引擎類型可以包括被指定用於對儲存系統中的一種資料類型執行資料管理操作的相應的資料管理處理引擎類型。在一些實施例中，管理操作包括分層、壓合、壓縮、抹除編碼或快照中的一個或多個。

在1206，儲存系統300確定資料是分散式帳本系統的多種資料類型中的一種資料類型。在一些實施例中，所述一種資料類型可以是區塊資料或交易資料。在一些實施例中，所述一種資料類型可以是狀態資料。

在1208，儲存系統300應用被指定用於處理所述一種資料類型的處理引擎類型。在一些實施例中，被指定用於處理所述一種資料類型的所述處理引擎類型可以被配置為具有包括對儲存系統300中的區塊資料或交易資料進行分層、抹除編碼和壓縮的功能。在一些實施例中，被指定用於處理所述一種資料類型的處理引擎類型可以被配置為具有包括對儲存系統300中的狀態資料進行快照和壓縮的功能。

圖13是示出可根據本文實施例執行的處理1300的流程

圖。為了方便，處理1300將被描述為由圖3的日誌結構儲存系統300執行。然而，處理1300可以由位於一個或多個位置的一個或多個電腦的系統執行，並且根據本文被適當地程式化。例如，適當程式化的資料處理和儲存系統可以執行處理1300。

在1302，儲存系統(例如，日誌結構儲存系統300)接收多個處理引擎的配置。在一些實施例中，所述配置可以根據例如表1的基於區塊鏈的中心化帳本系統的多種資料類型中的每種資料類型的特性，配置用於處理這種資料類型的相應的處理引擎類型。

在一些實施例中，儲存系統300可以包括被指定用於處理區塊資料的處理引擎類型；被指定用於處理交易資料的處理引擎類型；被指定用於處理索引資料的處理引擎類型。

在1304，儲存系統300接收針對基於區塊鏈的中心化帳本系統的資料的處理請求。在一些實施例中，基於區塊鏈的中心化帳本系統的資料類型可以包括區塊資料、交易資料和索引資料。

在一些實施例中，儲存系統300可以接收針對基於區塊鏈的中心化帳本系統的資料的I/O請求。在一些實施例中，被指定用於處理基於區塊鏈的中心化帳本系統的一種資料類型的相應的處理引擎類型可以包括被指定用於對基於區塊鏈的中心化帳本系統的所述一種資料類型執行讀取或寫入操作的相應的I/O處理引擎類型，例如，根據處理

500、600、700、1100和1400的部分或全部操作。

在一些實施例中，儲存系統300可以接收針對基於區塊鏈的中心化帳本系統的資料的資料管理請求。在一些實施例中，被指定用於處理所述一種資料類型的相應的處理引擎類型可以包括被指定用於對儲存系統中的所述一種資料類型執行資料管理操作的相應的資料管理處理引擎類型。在一些實施例中，管理操作可以包括分層、壓合、壓縮、抹除編碼或快照中的一個或多個。

在1306，儲存系統300確定所述資料是基於區塊鏈的中心化帳本系統的資料類型中的一種資料類型。在一些實施例中，所述一種資料類型可以是區塊資料或交易資料。

在1308，儲存系統300根據一種資料類型的特性應用被指定用於處理所述一種資料類型的處理引擎類型。在一些實施例中，被指定用於處理所述一種資料類型的處理引擎類型可以被配置為具有包括對儲存系統中的區塊資料或交易資料進行分層、抹除編碼和壓縮的功能。在一些實施例中，儲存系統300根據處理800、900、1000和1400的一些或全部操作來應用被指定用於處理所述一種資料類型的處理引擎類型。

圖14是示出可根據本文實施例執行的處理1400的流程圖。為了方便，處理1400將被描述為由圖3的日誌結構儲存系統300執行。然而，處理1400可以由位於一個或多個位置的一個或多個電腦的系統執行，並且根據本文被適當地程式化。例如，適當程式化的資料處理和儲存系統可以

執行處理1400。

在1402，儲存系統(例如，日誌結構儲存系統300)接收用於在儲存系統中儲存相應的多個區塊鏈資料的多個寫入請求。在一些實施例中，多個區塊鏈資料中的每個可以包括區塊、交易或區塊鏈網路的狀態中的一個或多個的值以及與該值相對應的鍵。在一些實施例中，鍵可以包括與該值相對應的雜湊值。

在1404，儲存系統300根據多個區塊鏈資料的時間順序將多個區塊鏈資料追加到資料日誌文件(例如，資料日誌文件390、362、364、366、372、374和376)。例如，之後接收到的區塊鏈資料將被追加到已儲存在資料日誌文件中的先前接收到的資料中。在一些實施例中，資料日誌文件可以是僅追加文件。在一些實施例中，資料日誌文件可以儲存在日誌結構儲存系統300的包括多層儲存設備的多層儲存子系統(例如，多層儲存子系統320)中的第一層儲存設備(例如，儲存設備350)中，並且第一層儲存設備的存取速度在多層儲存設備中最高。

在1406，日誌結構儲存系統300被限制根據任何其他度量，例如根據多個區塊鏈資料中的值的對應鍵(例如，在KVP中)對資料日誌文件中的多個區塊鏈資料進行排序。在一些實施例中，不同於現有儲存系統將根據多個區塊鏈資料中的值的對應鍵對資料日誌文件中的多個區塊鏈資料進行重新排列，日誌結構儲存系統300的資料日誌文件中的多個區塊鏈資料是根據日誌結構儲存系統300產生

或接收到所述多個區塊鏈資料的時間來排列的。在1408，日誌結構儲存系統300例如根據處理600的相應操作產生指示資料日誌文件中的多個區塊鏈資料的對應實體儲存位置的索引。

在1410，日誌結構儲存系統300例如根據處理600的相應操作將索引寫入第一層儲存設備中。

在1412，日誌結構儲存系統300例如根據處理1000的相應操作，確定多個區塊鏈資料的對應遷移優先級、得分或度量。在一些實施例中，日誌結構儲存系統300根據多個區塊鏈資料的時間順序確定對應的遷移優先級。在一些實施例中，較舊的區塊鏈資料的遷移優先級可以高於較新區塊鏈資料的遷移優先級。

在1414，日誌結構儲存系統300根據對應的遷移優先級將儲存在第一層儲存設備中的多個區塊鏈資料遷移到第二層儲存設備(例如，儲存設備360)中。在一些實施例中，第二層儲存設備的存取速度可以比第一層儲存設備的存取速度更低。

圖15描繪了根據本文的實施例的裝置1500的模組的示例。裝置1500可以是儲存系統(例如，圖3的日誌結構儲存系統300)的實施例的示例。裝置1500可以對應於上述實施例，裝置1500包括：維護模組1502，用於維護多個儲存設備，所述多個儲存設備至少包括第一層儲存設備和第二層儲存設備；接收模組1504，用於接收針對帳本資料的寫入請求；確定模組1506，用於確定帳本資料的類型是否是區

塊資料；以及寫入模組1508，用於響應於確定帳本資料的類型是區塊資料，將資料寫入第二層儲存設備。

在可選實施例中，第一層儲存設備的性能特性比第二層儲存設備的性能特性更好，其中，所述性能特性包括存取速度、存取頻寬或存取延遲中的一個或多個。

在可選實施例中，裝置1500還包括：第一確定子模組，用於響應於確定帳本資料的類型不是區塊資料，確定第一層儲存設備的使用率；第二確定子模組，用於確定第一儲存設備的使用率是否達到或超過預定閾值；以及寫入子模組，用於響應於確定第一儲存設備的使用率達到或超過預定閾值，將資料寫入第二層儲存設備。

在可選實施例中，裝置1500還包括寫入子模組，用於響應於確定第一儲存設備的使用率低於預定閾值，將資料寫入第一層儲存設備。

在可選實施例中，裝置1500還包括：確定子模組，用於確定第一層儲存設備的使用率；以及降低子模組，用於響應於確定第一層儲存設備的使用率達到或超過第一預定閾值，基於第一層儲存設備的使用率降低將資料寫入第一層儲存設備的速度。

在可選實施例中，裝置1500還包括：第一降低子模組，用於當第一層儲存設備的使用率是第一值時，將資料寫入第一層儲存設備的速度降低到第一速率；第二降低子模組，用於當第一層儲存設備的使用率是第二值時，將資料寫入第一層儲存設備的速度降低到第二速率。第二速率

低於第一速率，並且第二值大於第一值。

在可選實施例中，裝置1500還包括：提高模組，用於響應於確定第一儲存設備的使用率低於第二預定閾值，基於第一層儲存設備的使用率提高將資料寫入第一層儲存設備的速度。第二預定閾值低於第一預定閾值。

在可選實施例中，裝置1500還包括：提高子模組，用於當第一層儲存設備的使用率是第三值時，將資料寫入第一層儲存設備的速度提高到第三速率；第二提高子模組，用於當第一層儲存設備的使用率是第四值時，將資料寫入第一層儲存設備的速度提高到第四速率。第四速率高於第三速率，並且第四值小於第三值。

在先前實施中示出的系統、裝置、模組或單元可以透過使用電腦晶片或實體來實施，或者可以透過使用具有特定功能的產品來實施。典型的實施設備是電腦，電腦可以是個人電腦、膝上型電腦、蜂窩電話、相機電話、智慧電話、個人數位助理、媒體播放器、導航設備、電子郵件接收和發送設備、遊戲控制台、平板電腦、可穿戴設備或這些設備的任意組合。

對於裝置中各個單元的功能和角色的實施過程，可以參考前一方法中相應步驟的實施過程。為簡單起見，這裡省略了細節。

由於裝置實施例基本上與方法實施例相對應，因此對於相關部分，可以參考方法實施例中的相關描述。前述裝置實施例僅僅是示例。被描述為單獨部分的模組可以或可

以不是實體上分離的，並且顯示為模組的部分可以是或可以不是實體模組，可以位於一個位置，或者可以分佈在多個網路模組上。可以基於實際需求來選擇一些或所有模組，以實現本文方案的目標。本技術領域具有通常知識者無需付出創造性努力就能理解和實施本申請的實施例。

再次參見圖15，它可以被解釋為示出了資料處理和儲存裝置的內部功能模組和結構。資料處理和儲存裝置可以是日誌結構儲存系統(例如，圖3的日誌結構儲存系統300)的示例。本質上，執行主體實質上可以是電子設備，該電子設備包括以下：一個或多個處理器；一個或多個電腦可讀記憶體，其被配置為儲存一個或多個處理器的可執行指令。在一些實施例中，一個或多個電腦可讀記憶體耦接到一個或多個處理器並且具有儲存在其上的程式指令，所述指令可由一個或多個處理器執行以執行本文描述的算法、方法、功能、處理、流程和程序。

所描述的主題的實施例可單獨地或組合地包括一個或多個特徵。例如，在第一實施例中，一種方法包括：儲存系統維護多個儲存設備，所述多個儲存設備至少包括第一層儲存設備和第二層儲存設備；接收針對帳本資料的寫入請求；確定所述帳本資料的類型是否是區塊資料；以及響應於確定所述帳本資料的所述類型是區塊資料，將所述資料寫入所述第二層儲存設備。

前述和其它描述的實施例可以各自可選地包括一個或多個以下特徵：

第一特徵，可與以下任意特徵組合，指定第一層儲存設備的性能特性比第二層儲存設備的性能特性更好，其中，所述性能特性包括存取速度、存取頻寬或存取延遲中的一個或多個。

第二特徵，可與任意先前或以下特徵組合，指定響應於確定帳本資料的類型不是區塊資料，確定第一層儲存設備的使用率；確定第一儲存設備的使用率是否達到或超過預定閾值；以及響應於確定第一儲存設備的使用率達到或超過預定閾值，將資料寫入第二層儲存設備。

第三特徵，可與任意先前或以下特徵組合，指定所述方法還包括：響應於確定所述第一儲存設備的使用率低於所述預定閾值，將所述資料寫入所述第一層儲存設備。

第四特徵，可與任意先前或以下特徵組合，指定所述方法還包括：確定第一層儲存設備的使用率；以及響應於確定第一儲存設備的使用率達到或超過第一預定閾值，基於第一層儲存設備的使用率降低將資料寫入第一層儲存設備的速度。

第五特徵，可與任意先前或以下特徵組合，指定基於第一層儲存設備的使用率降低將資料寫入第一層儲存設備的速度包括：當第一層儲存設備的使用率是第一值時，將資料寫入第一層儲存設備的速度降低到第一速率；當第一層儲存設備的使用率是第二值時，將資料寫入第一層儲存設備的速度降低到第二速率；第二速率低於第一速率，並且第二值大於第一值。

第六特徵，可與任意先前或以下特徵組合，指定所述方法還包括：響應於確定第一層儲存設備的使用率低於第二預定閾值，基於第一層儲存設備的使用率提高將資料寫入第一層儲存設備的速度，其中，第二預定閾值低於第一預定閾值。

第七特徵，可與任意先前特徵組合，指定基於第一層儲存設備的使用率提高將資料寫入第一層儲存設備的速度包括：當第一層儲存設備的使用率是第三值時，將資料寫入第一層儲存設備的速度提高到第三速率；當第一層儲存設備的使用率是第四值時，將資料寫入第一層儲存設備的速度提高到第四速率；第四速率高於第三速率，並且第四值小於第三值。

本文中描述的主題、動作和操作的實施例可以在數位電子電路中、有形體現的電腦軟體或韌體中、包括本文中揭露的結構及其結構等同物的電腦硬體中，或者它們中的一個或多個的組合中實施。本文中描述的主題的實施例可以實施為一個或多個電腦程式，例如，一個或多個電腦程式指令模組，編碼在電腦程式載體上，用於由資料處理裝置執行或控制資料處理的操作。例如，電腦程式載體可以包括具有一個或多個電腦可讀儲存媒介，其上編碼或儲存有指令。載體可以是有形的非暫態電腦可讀媒介，例如磁碟、磁光碟或光碟、固態驅動器、隨機存取記憶體(RAM)、唯讀記憶體(ROM)或其他媒體類型。可選地或附加地，載體可以是人工產生的傳播信號，例如，機器產生

的電、光或電磁信號，其被產生以編碼用於傳輸到合適的接收器裝置以供資料處理裝置執行的資訊。電腦儲存媒介可以是或可以部分是機器可讀儲存設備、機器可讀儲存基板、隨機或序列存取記憶體設備或它們中的一個或多個的組合。電腦儲存媒介不是傳播信號。

電腦程式也可以被稱為或描述為程式、軟體、軟體應用、app、模組、軟體模組、引擎、腳本或代碼，可以以任何形式的程式化語言編寫，包括編譯或演繹性語言、說明或程式性語言；它可以配置為任何形式，包括作為獨立程式，或者作為模組、組件、引擎、子程式或適合在計算環境中執行的其他單元，該環境可包括由資料通訊網路互連的在一個或多個位置一台或多台電腦。

電腦程式可以但非必須對應於文件系統中的文件。電腦程式可以儲存在：保存其他程式或資料的文件的一部分中，例如，儲存在標記語言文檔中的一個或多個腳本；專用於所討論的程式的單個文件；或者多個協調文件，例如，儲存一個或多個模組、子程式或代碼部分的多個文件。

舉例來說，用於執行電腦程式的處理器包括通用和專用微處理器，以及任何類型的數位電腦的任何一個或多個處理器。通常，處理器將接收用於執行的電腦程式的指令、以及接收來自耦接到處理器的非暫態電腦可讀媒介的資料。

用語“資料處理裝置”包括用於處理資料的所有類型

的裝置、設備和機器，包括例如可程式化處理器、電腦或多個處理器或電腦。資料處理設備可以包括例如FPGA(現場可程式化閘陣列)，ASIC(專用積體電路)或GPU(圖形處理單元)的專用邏輯電路。除了硬體，該裝置還可以包括為電腦程式創建執行環境的代碼，例如，構成處理器韌體、協定棧、資料庫管理系統、操作系統、或者它們一個或多個的組合的代碼。

本文中描述的處理和邏輯流程可以由一台或多台電腦或處理器執行一個或多個電腦程式進行，以透過對輸入資料進行運算並產生輸出來執行操作。過程和邏輯流程也可以由例如FPGA、ASIC或GPU等的專用邏輯電路或專用邏輯電路與一個或多個程式化電腦的組合來執行。

適合於執行電腦程式的電腦可以基於通用及/或專用微處理器，或任何其他種類的中央處理單元。通常，中央處理單元將從唯讀記憶體及/或隨機存取記憶體接收指令和資料。電腦的元件可包括用於執行指令的中央處理單元和用於儲存指令和資料的一個或多個儲存設備。中央處理單元和記憶體可以補充有專用邏輯電路或整合在專用邏輯電路中。

通常，電腦還將包括或可操作地耦接至一個或多個大容量儲存設備，以從一個或多個儲存設備接收資料或將資料傳輸到一個或多個大容量儲存設備。儲存設備可以是例如磁碟、磁光碟或光碟、固態驅動器或任何其他類型的非暫態電腦可讀媒介。然而，電腦不需要具有這樣的設備。

因此，電腦可以耦接到本地及/或遠端的例如一個或多個記憶體的一個或多個儲存設備。例如，電腦可以包括作為電腦的整合組件的一個或多個區域記憶體，或者電腦可以耦接到雲網路中的一個或多個遠端記憶體。此外，電腦可以嵌入在另一個設備中，例如行動電話，個人數位助理(PDA)，行動音訊或視訊播放器，遊戲控制台，全球定位系統(GPS)接收器或例如通用序列匯流排(USB)快閃記憶體驅動器的便攜式儲存設備，僅舉幾例。

組件可以透過諸如直接地連接、或透過一個或多個中間組件彼此電學連接或光學連接可通訊地連接而彼此“耦接”。如果其中一個組件被整合到另一個組件中，組件也可以彼此“耦接”。例如，整合到處理器(例如，L2快取組件)中的儲存組件“耦接到”處理器。

為了提供與用戶的交互，本文中所描述的主題的實施例可以在電腦上實施或配置為與該電腦通訊，該電腦具有：顯示設備，例如，LCD(液晶顯示器)監視器，用於向用戶顯示資訊；以及輸入設備，用戶可以透過該輸入設備向該電腦提供輸入，例如鍵盤和例如滑鼠、軌跡球或觸摸板等的指針設備。其他類型的設備也可用於提供與用戶的交互；例如，提供給用戶的反饋可以是任何形式的感覺反饋，例如視覺反饋、聽覺反饋或觸覺反饋；並且可以接收來自用戶的任何形式的輸入，包括聲音、語音或觸覺輸入。此外，電腦可以透過向用戶使用的設備發送文檔和從用戶使用的設備接收文檔來與用戶交互；例如，透過向用

戶設備上的web瀏覽器發送web頁面以響應從web瀏覽器收到的請求，或者透過與例如智慧電話或電子平板電腦等的用戶設備上運行的應用程式(app)進行交互。此外，電腦可以透過向個人設備(例如，運行訊息應用的智慧手機)輪流發送文本訊息或其他形式的訊息來並接收來自用戶的響應訊息來與用戶交互。

本文使用與系統，裝置和電腦程式組件有關的用語“配置為”。對於被配置為執行特定操作或動作的一個或多個電腦的系統，意味著系統已經在其上安裝了在運行中促使該系統執行所述操作或動作的軟體、韌體、硬體或它們的組合。對於被配置為執行特定操作或動作的一個或多個電腦程式，意味著一個或多個程式包括當被資料處理裝置執行時促使該裝置執行所述操作或動作的指令。對於被配置為執行特定操作或動作的專用邏輯電路，意味著該電路具有執行所述操作或動作的電子邏輯。

儘管本文包含許多具體實施細節，但這些不應被解釋為由請求項本身限定的對要求保護的範圍的限制，而是作為對特定實施例的具體特徵的描述。在本文單獨實施例的上下文中描述的某些特徵也可以在單個實施例中組合實現。相反，在單個實施例的上下文中描述的各種特徵也可以單獨地或以任何合適的子組合在多個實施例中實現。此外，儘管上面的特徵可以描述為以某些組合起作用並且甚至最初如此要求保護，但是在一些情況下，可以從要求保護的組合中刪除來自該組合的一個或多個特徵，並且要求

保護可以指向子組合或子組合的變體。

類似地，雖然以特定順序在圖式中描繪了操作並且在請求項中敘述了操作，但是這不應該被理解為：為了達到期望的結果，要求以所示的特定順序或依次執行這些操作，或者要求執行所有示出的操作。在某些情況下，多任務和並行處理可能是有利的。此外，上述實施例中的各種系統模組和組件的劃分不應被理解為所有實施例中都要求如此劃分，而應當理解，所描述的程式組件和系統通常可以一起整合在單個軟體產品中或打包成多個軟體產品。

已經描述了主題的特定實施例。其他實施例在以下請求項的範圍內。例如，請求項中記載的動作可以以不同的順序執行並且仍然實施期望的結果。作為一個示例，附圖中描繪的過程無需要求所示的特定順序或次序來實現期望的結果。在一些情況下，多任務和並行處理可能是有利的。

【符號說明】

100:環境

102:聯盟區塊鏈網路

106:計算設備

108:計算設備

110:網路

200:架構

202:參與者系統

204:參與者系統

206:參與者系統

212:區塊鏈網路

214:節點

216:區塊鏈

216':副本

216'':副本

216''':副本

300:日誌結構儲存系統

310:前端I/O子系統

312:增量索引映射

313:記憶體快取

314:不可變索引映射

315:記憶體

316:基本索引映射

317:區塊快取

320:儲存子系統

330:後端資料管理子系統

340:分散式儲存系統

350:儲存設備

355:熱資料

360:儲存設備

362:資料日誌文件

364:資料日誌文件

365:暖資料

366:資料日誌文件

370:儲存設備

372:資料日誌文件

374:資料日誌文件

375:冷資料

376:資料日誌文件

380:索引日誌文件

390:資料日誌文件

400:分層儲存系統

402:第一級池

404:第二級池

406:第三級池

408:第四級池

410:熱日誌文件

412:暖日誌文件

414:冷日誌文件

416:存檔文件

418:分層池管理器

420:遷移任務管理器

422:遷移調度器

423:服務品質管理器

424:記憶體快取

426:磁碟快取

500:處理

502:步驟

504:步驟

506:步驟

508:步驟

510:步驟

512:步驟

600:處理

602:步驟

604:步驟

606:步驟

608:步驟

610:步驟

612:步驟

614:步驟

616:步驟

618:步驟

620:步驟

622:步驟

700:處理

702:步驟

704:步驟

706:步驟

708:步驟

710:步驟

712:步驟

714:步驟

716:步驟

800:處理

802:步驟

804:步驟

806:步驟

808:步驟

810:步驟

812:步驟

814:步驟

816:步驟

900:處理

902:步驟

904:步驟

906:步驟

908:步驟

910:步驟

912:步驟

914:步驟

916:步驟

1000:處理

1002:步驟

1004:步 驟

1006:步 驟

1008:步 驟

1100:處 理

1102:步 驟

1104:步 驟

1106:步 驟

1108:步 驟

1110:步 驟

1112:步 驟

1114:步 驟

1116:步 驟

1118:步 驟

1200:處 理

1202:步 驟

1204:步 驟

1206:步 驟

1208:步 驟

1300:處 理

1302:步 驟

1304:步 驟

1306:步 驟

1308:步 驟

1400:處 理

1402:步驟

1404:步驟

1406:步驟

1408:步驟

1410:步驟

1412:步驟

1414:步驟

1500:裝置

1502:維護模組

1504:接收模組

1506:確定模組

1508:寫入模組

【發明申請專利範圍】

【請求項1】一種電腦實施的方法，包括：

儲存系統維護多個儲存設備，所述多個儲存設備至少包括第一層儲存設備和第二層儲存設備；

接收第一帳本資料的第一寫入請求，其中所述第一帳本資料包括下述一者：區塊鏈交易資料、區塊鏈區塊資料、區塊鏈索引資料、或由區塊鏈網路所產生之區塊鏈狀態資料；

確定所述第一帳本資料的類型是否是區塊鏈區塊資料，其中所述區塊鏈區塊資料包括區塊鏈中至少一區塊，以及所述該至少一區塊包括一或多區塊鏈交易以及所述區塊鏈中先前區塊之雜湊值；

響應於確定所述第一帳本資料的類型是區塊鏈區塊資料，將所述第一帳本資料寫入所述第二層儲存設備，其中所述第二層儲存設備具有比所述第一層儲存設備更大的儲存空間；

接收由所述區塊鏈網路產生的第二帳本資料的第二寫入請求；

確定所述第二帳本資料的類型不是區塊鏈區塊資料，其中所述第二帳本資料之類型是區塊鏈交易資料、區塊鏈索引資料、或區塊鏈狀態資料；

響應於確定所述第二帳本資料的類型不是區塊鏈區塊資料，確定所述第一層儲存設備之使用率；

確定所述第一層儲存設備之所述使用率是否低於預定

閾值；以及

響應於確定所述第一層儲存設備之所述使用率是低於所述預定閾值，將所述第二帳本資料寫入所述第一層儲存設備。

【請求項2】如請求項1所述的電腦實施的方法，其中，所述第一層儲存設備的性能特性比所述第二層儲存設備的性能特性更好，所述性能特性包括存取速度、存取頻寬或存取延遲中的一個或多個。

【請求項3】如請求項1所述的電腦實施的方法，還包括：

接收由所述區塊鏈網路產生的第三帳本資料的第三寫入請求；

確定所述第三帳本資料的類型不是區塊鏈區塊資料；

響應於確定所述第三帳本資料的類型不是區塊鏈區塊資料，確定所述第一層儲存設備之所述使用率；

確定所述第一層儲存設備之所述使用率是否達到或超過所述預定閾值；以及

響應於確定所述第一層儲存設備之所述使用率是達到或超過所述預定閾值，將所述第三帳本資料寫入所述第二層儲存設備。

【請求項4】如請求項1所述的電腦實施的方法，還包括：

確定所述第一層儲存設備之所述使用率；以及

響應於確定所述第一層儲存設備的所述使用率達到或

超過第一預定閾值，基於所述第一層儲存設備的所述使用率降低將資料寫入所述第一層儲存設備的速度。

【請求項5】如請求項4所述的電腦實施的方法，其中，基於所述第一層儲存設備的所述使用率降低將資料寫入所述第一層儲存設備的速度包括：

當所述第一層儲存設備的所述使用率是第一值時，使將資料寫入所述第一層儲存設備的所述速度降低到第一速率；以及

當所述第一層儲存設備的所述使用率是第二值時，使將資料寫入所述第一層儲存設備的所述速度降低到第二速率；

其中，所述第二速率低於所述第一速率，並且所述第二值大於所述第一值。

【請求項6】如請求項4所述的電腦實施的方法，還包括：

響應於確定所述第一層儲存設備的所述使用率低於第二預定閾值，基於所述第一層儲存設備的所述使用率提高將資料寫入所述第一層儲存設備的速度，其中，所述第二預定閾值低於所述第一預定閾值。

【請求項7】如請求項6所述的電腦實施的方法，其中，基於所述第一層儲存設備的所述使用率提高將資料寫入所述第一層儲存設備的速度包括：

當所述第一層儲存設備的所述使用率是第三值時，使將資料寫入所述第一層儲存設備的所述速度提高到第三速

率；以及

當所述第一層儲存設備的所述使用率是第四值時，使將資料寫入所述第一層儲存設備的所述速度提高到第四速率；

其中，所述第四速率高於所述第三速率，並且所述第四值小於所述第三值。

【請求項8】一種非暫態電腦可讀儲存媒體，其上儲存有一或多指令，所述指令可由儲存系統執行以執行以下操作，包括：

維護多個儲存設備，所述多個儲存設備至少包括第一層儲存設備和第二層儲存設備；

接收第一帳本資料的第一寫入請求，其中所述第一帳本資料包括下述一者：區塊鏈交易資料、區塊鏈區塊資料、區塊鏈索引資料、或由區塊鏈網路所產生之區塊鏈狀態資料；

確定所述第一帳本資料的類型是否是區塊鏈區塊資料，其中所述區塊鏈區塊資料包括區塊鏈中至少一區塊，以及所述該至少一區塊包括一或多區塊鏈交易以及所述區塊鏈中先前區塊之雜湊值；

響應於確定所述第一帳本資料的類型是區塊鏈區塊資料，將所述第一帳本資料寫入所述第二層儲存設備，其中所述第二層儲存設備具有比所述第一層儲存設備更大的儲存空間；

接收由所述區塊鏈網路產生的第二帳本資料的第二寫

入請求；

確定所述第二帳本資料的類型不是區塊鏈區塊資料，其中所述第二帳本資料之類型是區塊鏈交易資料、區塊鏈索引資料、或區塊鏈狀態資料；以及

響應於確定所述第二帳本資料的類型不是區塊鏈區塊資料，確定所述第一層儲存設備之使用率；

確定所述第一層儲存設備之所述使用率是否低於預定閾值；以及

響應於確定所述第一層儲存設備之所述使用率是低於所述預定閾值，將所述第二帳本資料寫入所述第一層儲存設備。

【請求項9】如請求項8所述的非暫態電腦可讀儲存媒體，其中，所述第一層儲存設備的性能特性比所述第二層儲存設備的性能特性更好，所述性能特性包括存取速度、存取頻寬或存取延遲中的一個或多個。

【請求項10】如請求項8所述的非暫態電腦可讀儲存媒體，所述操作還包括：

接收由所述區塊鏈網路產生的第三帳本資料的第三寫入請求；

確定所述第三帳本資料的類型不是區塊鏈區塊資料；

響應於確定所述第三帳本資料的類型不是區塊鏈區塊資料，確定所述第一層儲存設備之所述使用率；

確定所述第一層儲存設備之所述使用率是否達到或超過所述預定閾值；以及

響應於確定所述第一層儲存設備之所述使用率是達到或超過所述預定閾值，將所述第三帳本資料寫入所述第二層儲存設備。

【請求項11】如請求項8所述的非暫態電腦可讀儲存媒體，所述操作還包括：

確定所述第一層儲存設備之所述使用率；以及

響應於確定所述第一層儲存設備的所述使用率達到或超過第一預定閾值，基於所述第一層儲存設備的所述使用率降低將資料寫入所述第一層儲存設備的速度。

【請求項12】如請求項11所述的非暫態電腦可讀儲存媒體，其中，基於所述第一層儲存設備的所述使用率降低將資料寫入所述第一層儲存設備的速度包括：

當所述第一層儲存設備的所述使用率是第一值時，使將資料寫入所述第一層儲存設備的所述速度降低到第一速率；以及

當所述第一層儲存設備的所述使用率是第二值時，使將資料寫入所述第一層儲存設備的所述速度降低到第二速率；

其中，所述第二速率低於所述第一速率，並且所述第二值大於所述第一值。

【請求項13】如請求項11所述的非暫態電腦可讀儲存媒體，所述操作還包括：

響應於確定所述第一層儲存設備的所述使用率低於第二預定閾值，基於所述第一層儲存設備的所述使用率提高

將資料寫入所述第一層儲存設備的速度，其中，所述第二預定閾值低於所述第一預定閾值。

【請求項14】如請求項13所述的非暫態電腦可讀儲存媒體，其中，基於所述第一層儲存設備的所述使用率提高將資料寫入所述第一層儲存設備的速度包括：

當所述第一層儲存設備的所述使用率是第三值時，使將資料寫入所述第一層儲存設備的所述速度提高到第三速率；以及

當所述第一層儲存設備的所述使用率是第四值時，使將資料寫入所述第一層儲存設備的所述速度提高到第四速率；

其中，所述第四速率高於所述第三速率，並且所述第四值小於所述第三值。

【請求項15】一種儲存系統，包括：

一或多處理器；以及

一或多電腦可讀記憶體，所述一或多電腦可讀記憶體與所述一或多處理器耦接，且所述一或多電腦可讀記憶體其上儲存有指令，當所述指令由所述一或多處理器執行時致使執行操作，所述操作包括:

所述儲存系統維護多個儲存設備，所述多個儲存設備至少包括第一層儲存設備和第二層儲存設備；

接收第一帳本資料的第一寫入請求，其中所述第一帳本資料包括下述一者：區塊鏈交易資料、區塊鏈區塊資料、區塊鏈索引資料、或由區塊鏈網路所產生之區塊鏈狀

態資料；

確定所述第一帳本資料的類型是否是區塊鏈區塊資料，其中所述區塊鏈區塊資料包括區塊鏈中至少一區塊，以及所述該至少一區塊包括一或多區塊鏈交易以及所述區塊鏈中先前區塊之雜湊值；

響應於確定所述第一帳本資料的類型是區塊鏈區塊資料，將所述第一帳本資料寫入所述第二層儲存設備，其中所述第二層儲存設備具有比所述第一層儲存設備更大的儲存空間；

接收由所述區塊鏈網路產生的第二帳本資料的第二寫入請求；

確定所述第二帳本資料的類型不是區塊鏈區塊資料，其中所述第二帳本資料之類型是區塊鏈交易資料、區塊鏈索引資料、或區塊鏈狀態資料；以及

響應於確定所述第二帳本資料的類型不是區塊鏈區塊資料，確定所述第一層儲存設備之使用率；

確定所述第一層儲存設備之所述使用率是否低於預定閾值；以及

響應於確定所述第一層儲存設備之所述使用率是低於所述預定閾值，將所述第二帳本資料寫入所述第一層儲存設備。

【請求項16】如請求項15所述的儲存系統，其中，所述第一層儲存設備的性能特性比所述第二層儲存設備的性能特性更好，所述性能特性包括存取速度、存取頻寬或存

取延遲中的一個或多個。

【請求項17】如請求項15所述的儲存系統，所述操作還包括：

接收由所述區塊鏈網路產生的第三帳本資料的第三寫入請求；

確定所述第三帳本資料的類型不是區塊鏈區塊資料；

響應於確定所述第三帳本資料的類型不是區塊鏈區塊資料，確定所述第一層儲存設備之所述使用率；

確定所述第一層儲存設備之所述使用率是否達到或超過所述預定閾值；以及

響應於確定所述第一層儲存設備之所述使用率是達到或超過所述預定閾值，將所述第三帳本資料寫入所述第二層儲存設備。

【請求項18】如請求項15所述的儲存系統，所述操作還包括：

確定所述第一層儲存設備之所述使用率；以及

響應於確定所述第一層儲存設備的所述使用率達到或超過第一預定閾值，基於所述第一層儲存設備的所述使用率降低將資料寫入所述第一層儲存設備的速度。

【請求項19】如請求項18所述的儲存系統，其中，基於所述第一層儲存設備的所述使用率降低將資料寫入所述第一層儲存設備的速度包括：

當所述第一層儲存設備的所述使用率是第一值時，使將資料寫入所述第一層儲存設備的所述速度降低到第一速

率；以及

當所述第一層儲存設備的所述使用率是第二值時，使將資料寫入所述第一層儲存設備的所述速度降低到第二速率；

其中，所述第二速率低於所述第一速率，並且所述第二值大於所述第一值。

【請求項20】如請求項18所述的儲存系統，所述操作還包括：

響應於確定所述第一層儲存設備的所述使用率低於第二預定閾值，基於所述第一層儲存設備的所述使用率提高將資料寫入所述第一層儲存設備的速度，其中，所述第二預定閾值低於所述第一預定閾值。

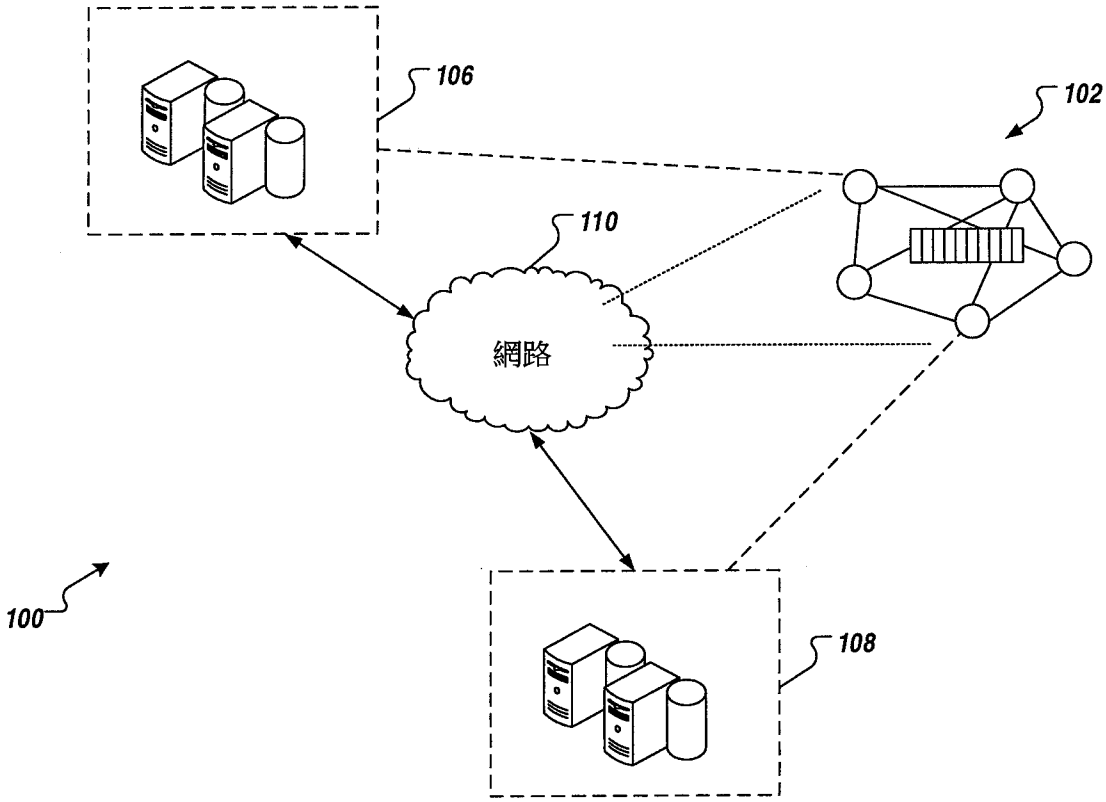
【請求項21】如請求項20所述的儲存系統，其中，基於所述第一層儲存設備的所述使用率提高將資料寫入所述第一層儲存設備的速度包括：

當所述第一層儲存設備的所述使用率是第三值時，使將資料寫入所述第一層儲存設備的所述速度提高到第三速率；以及

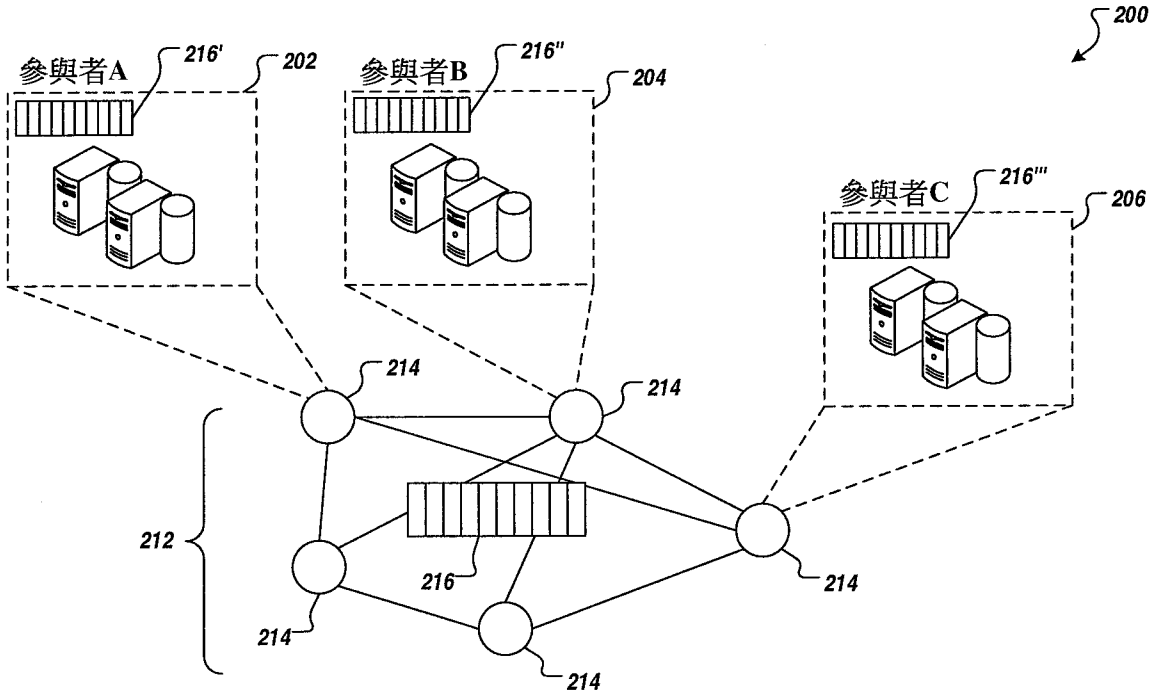
當所述第一層儲存設備的所述使用率是第四值時，使將資料寫入所述第一層儲存設備的所述速度提高到第四速率；

其中，所述第四速率高於所述第三速率，並且所述第四值小於所述第三值。

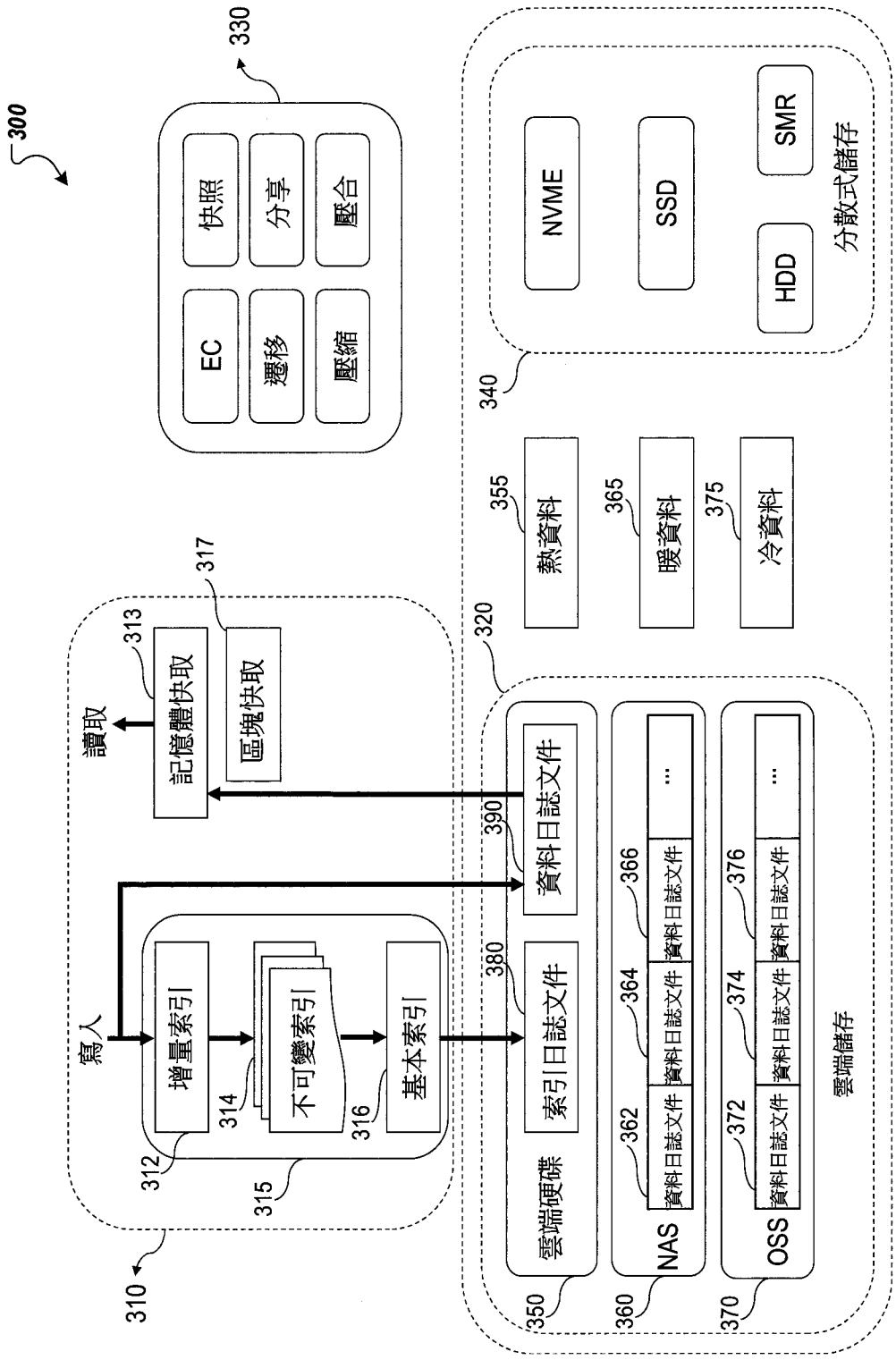
【發明圖式】


106
102
110
網路
100
108

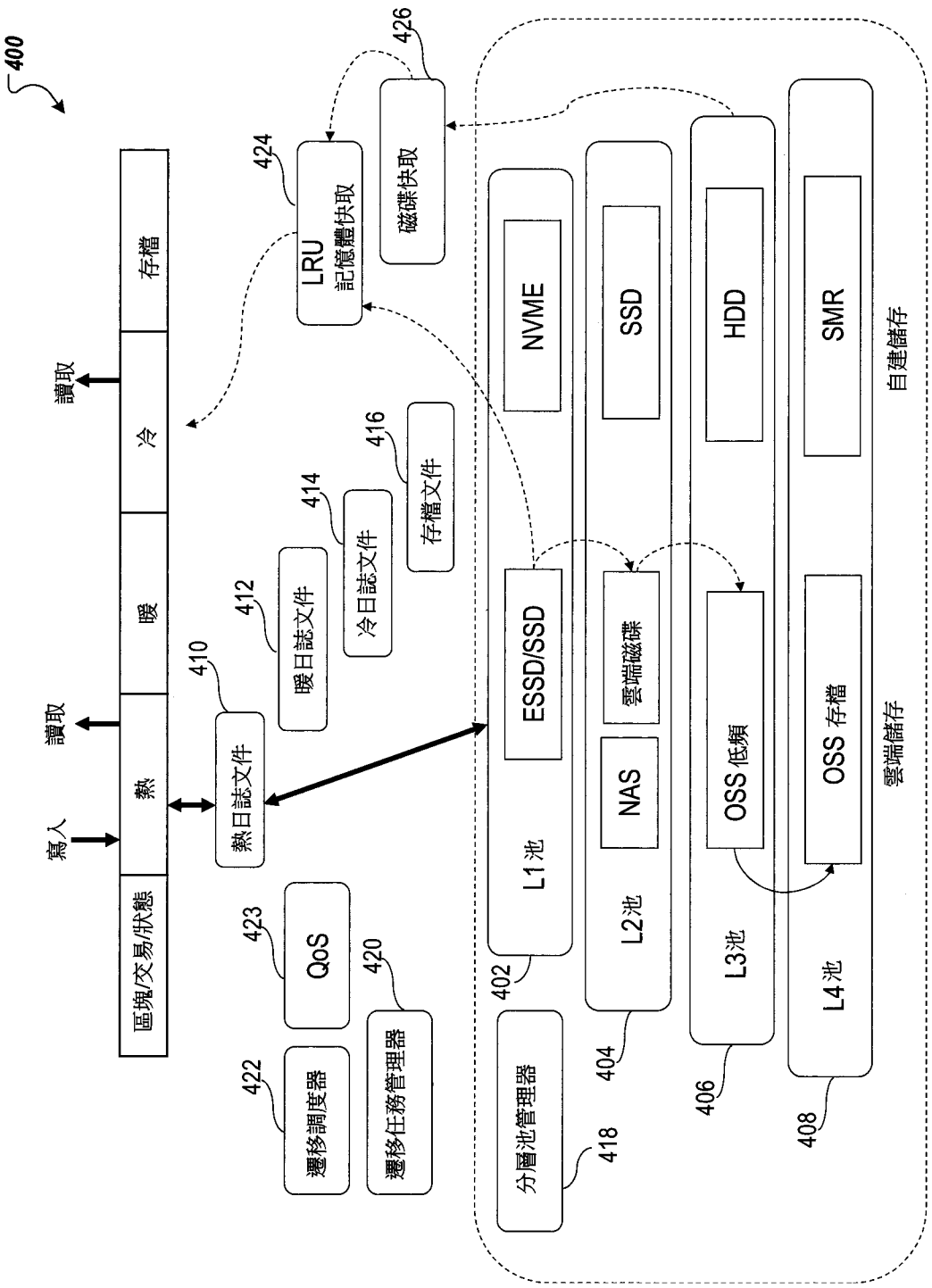
【圖 1】


200
參與者A
216'
202
參與者B
216"
204
參與者C
216'''
206
214
214
212
216
214
214
214

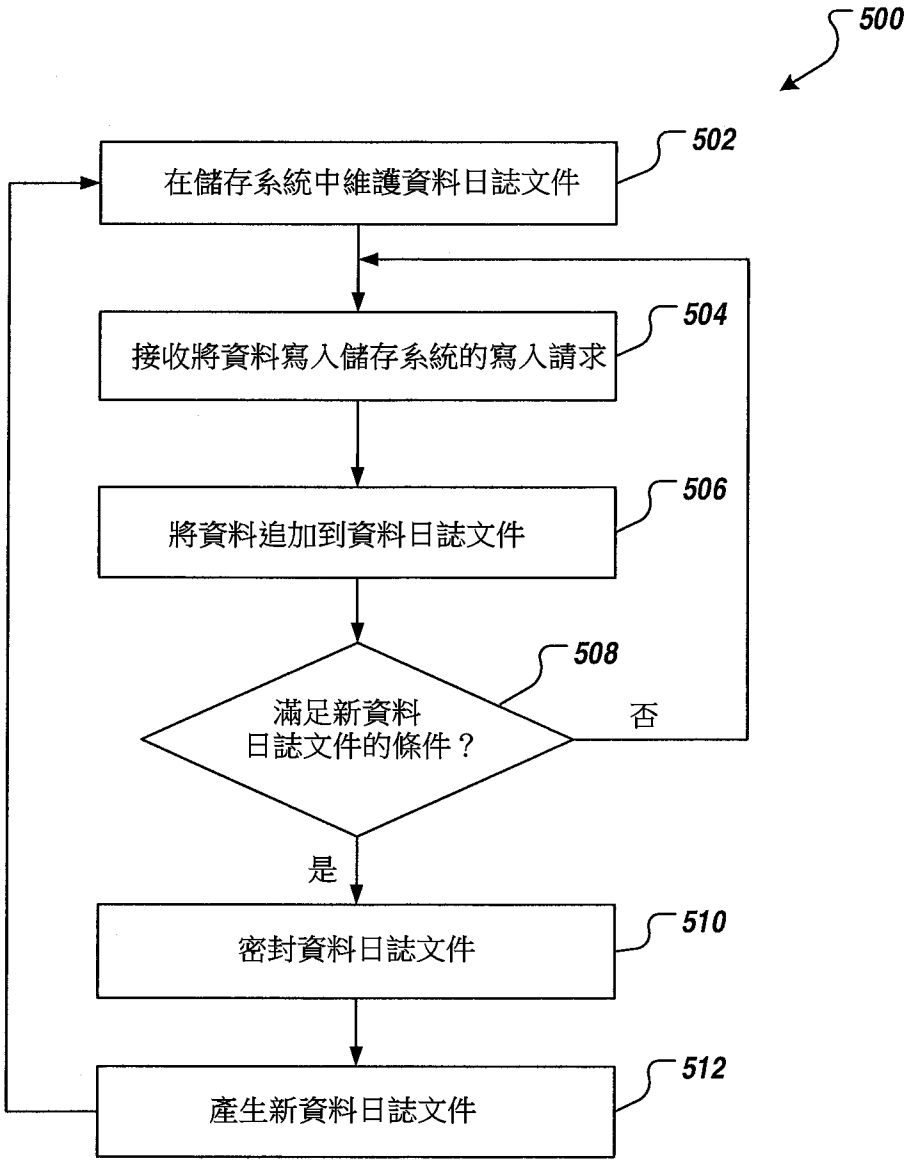
【圖 2】


300
330
EC
遷移
壓縮
快照
分享
壓合
317
313
讀取
記憶體快取
區塊快取
寫入
增量索引
314
不可變索引
316
基本索引
312
315
310
320
340
NVME
SSD
HDD
SMR
分散式儲存
355
熱資料
365
暖資料
375
冷資料
390
資料日誌文件
380
索引日誌文件
雲端硬碟
350
362
資料日誌文件
364
資料日誌文件
366
資料日誌文件
...
NAS
360
372
資料日誌文件
374
資料日誌文件
376
資料日誌文件
...
OSS
370
雲端儲存

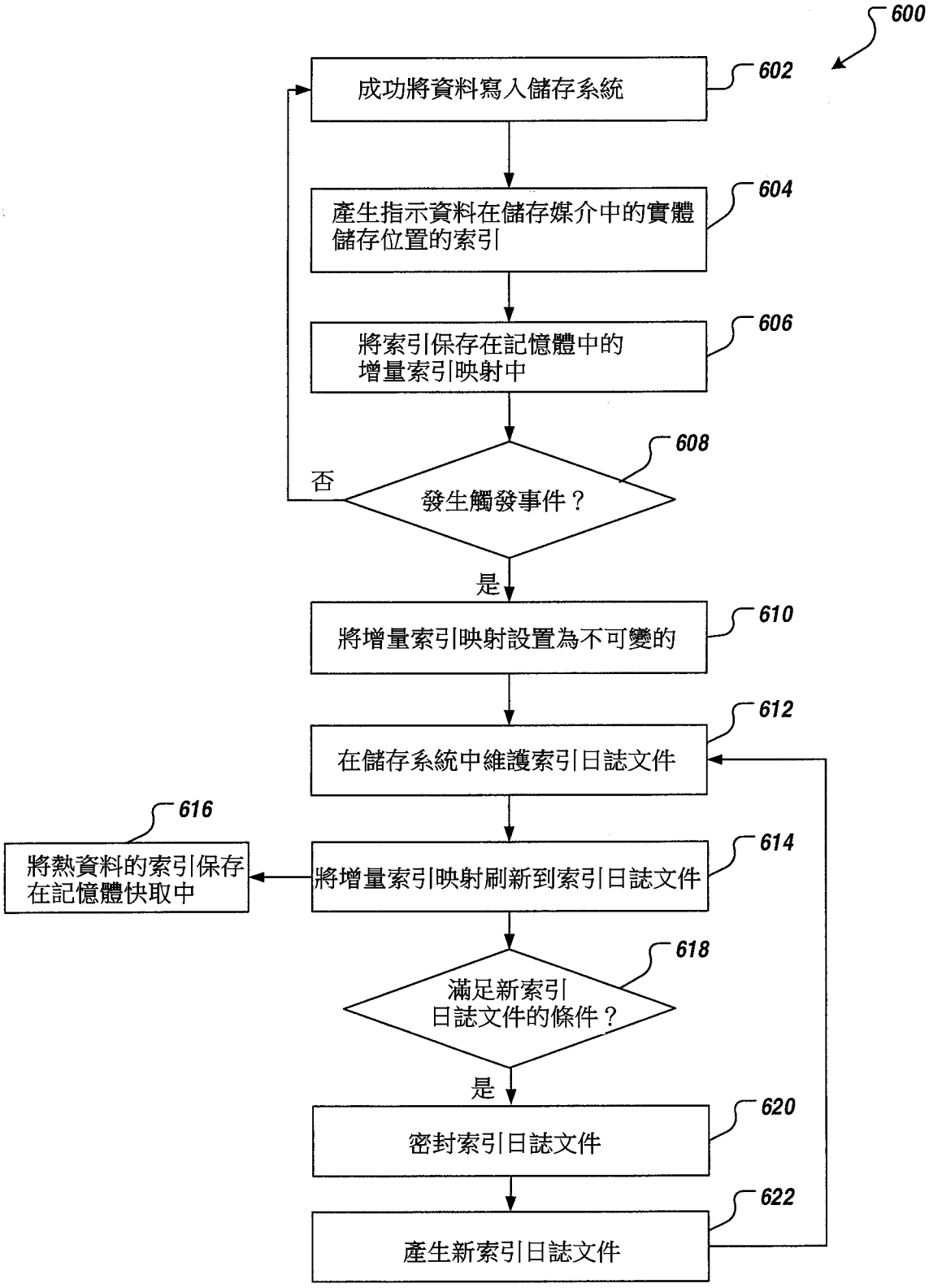
【圖 3】


400
寫入
讀取
讀取
區塊/交易/狀態
熱
暖
冷
存檔
熱日誌文件
410
暖日誌文件
412
冷日誌文件
414
存檔文件
416
LRU
記憶體快取
424
磁碟快取
426
QoS
423
遷移調度器
422
遷移任務管理器
420
分層池管理器
418
L1池
ESSD/SSD
NVME
402
L2池
NAS
雲端磁碟
SSD
404
L3池
OSS 低頻
HDD
406
L4池
OSS 存檔
SMR
408
雲端儲存
自建儲存

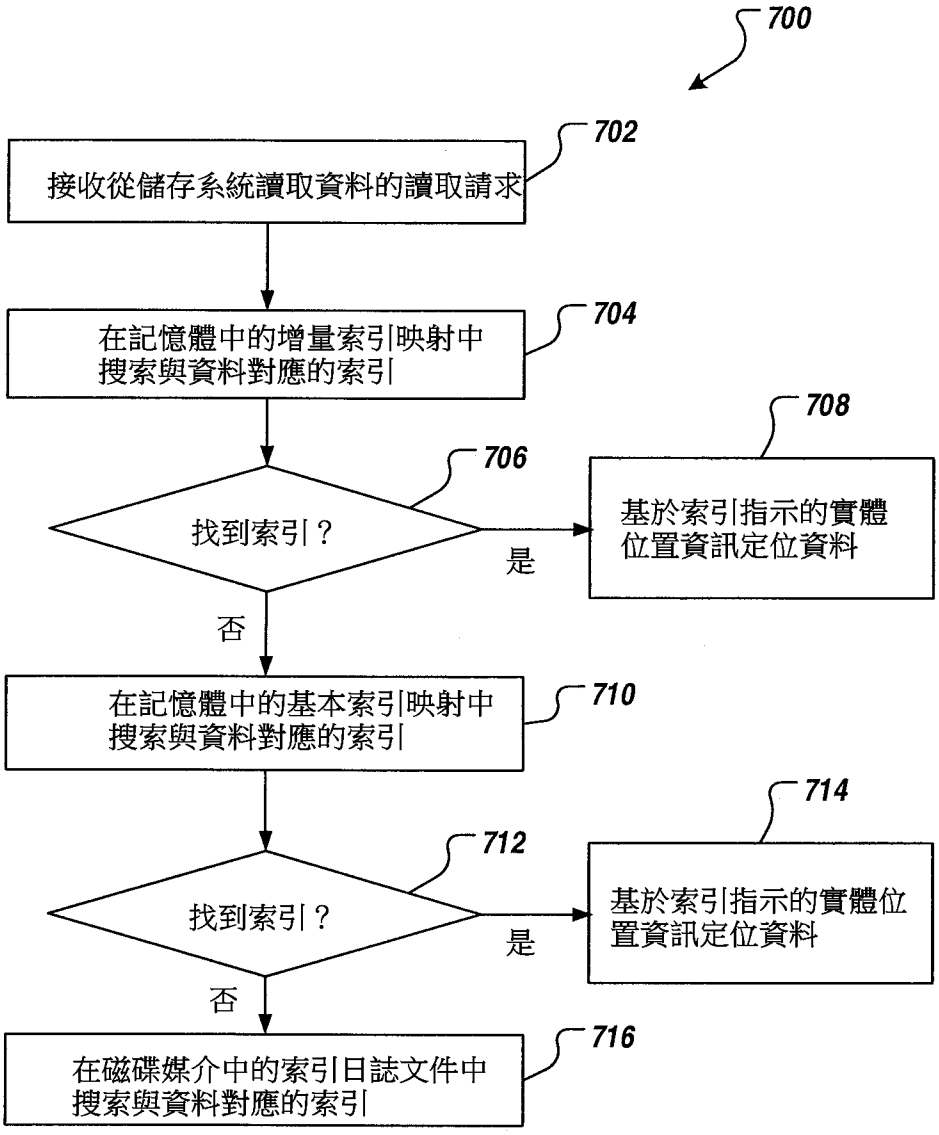
【圖 4】


500
502
在儲存系統中維護資料日誌文件
504
接收將資料寫入儲存系統的寫入請求
506
將資料追加到資料日誌文件
508
滿足新資料
日誌文件的條件？
否
是
510
密封資料日誌文件
512
產生新資料日誌文件

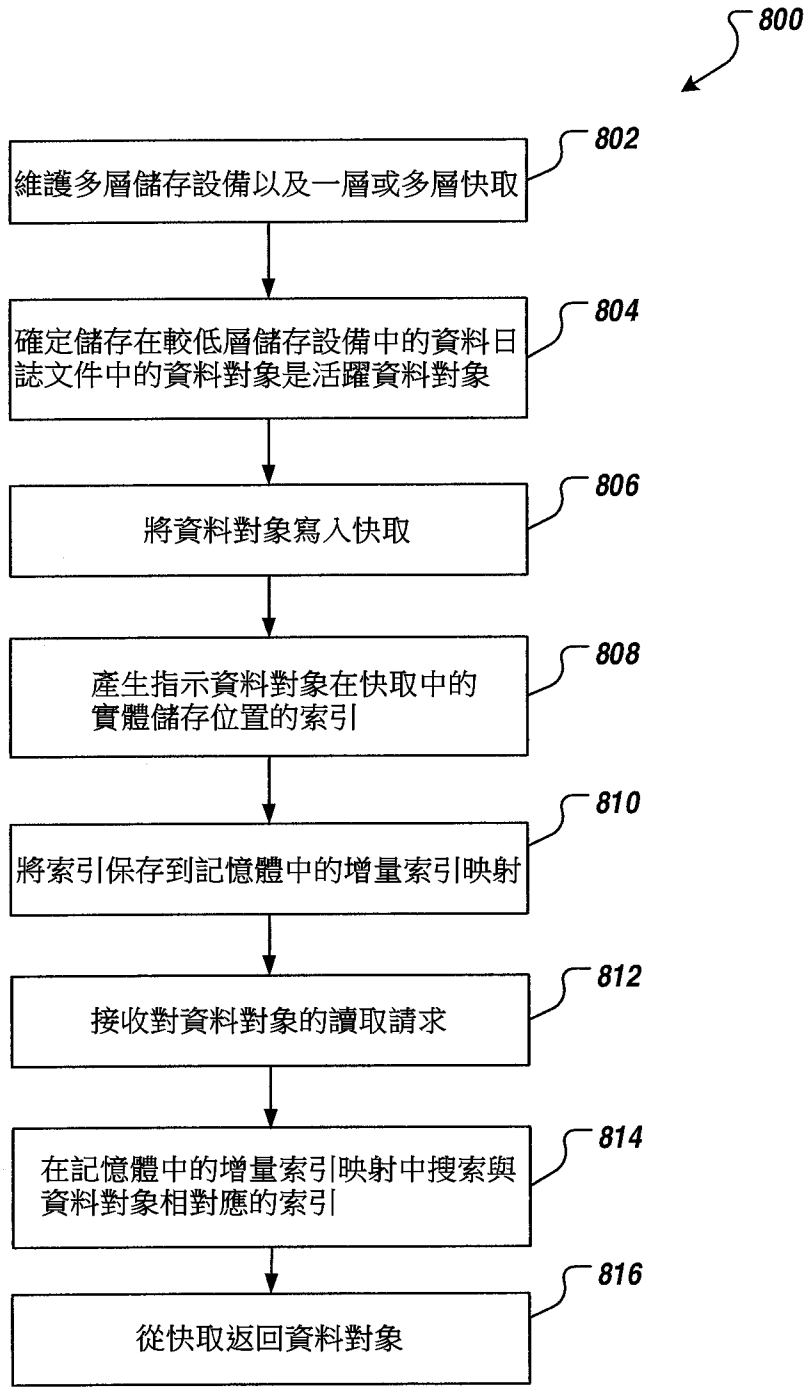
【圖 5】


600
成功將資料寫入儲存系統
602
產生指示資料在儲存媒介中的實體儲存位置的索引
604
將索引保存在記憶體中的增量索引映射中
606
發生觸發事件？
608
否
是
將增量索引映射設置為不可變的
610
在儲存系統中維護索引日誌文件
612
將增量索引映射刷新到索引日誌文件
614
將熱資料的索引保存在記憶體快取中
616
滿足新索引日誌文件的條件？
618
是
密封索引日誌文件
620
產生新索引日誌文件
622

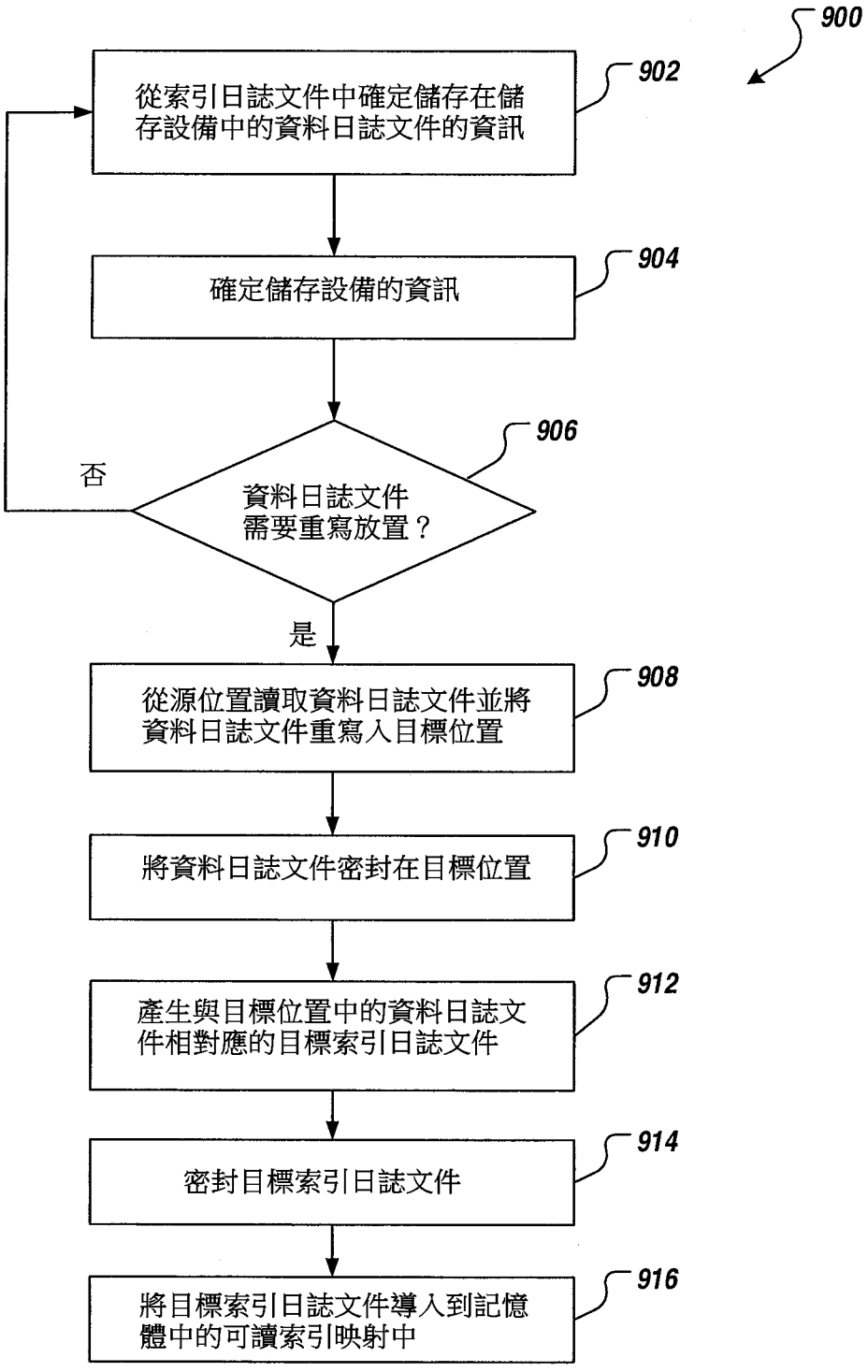
【圖 6】


700
702
接收從儲存系統讀取資料的讀取請求
704
在記憶體中的增量索引映射中搜索與資料對應的索引
706
找到索引？
是
708
基於索引指示的實體位置資訊定位資料
否
710
在記憶體中的基本索引映射中搜索與資料對應的索引
712
找到索引？
是
714
基於索引指示的實體位置資訊定位資料
否
716
在磁碟媒介中的索引日誌文件中搜索與資料對應的索引

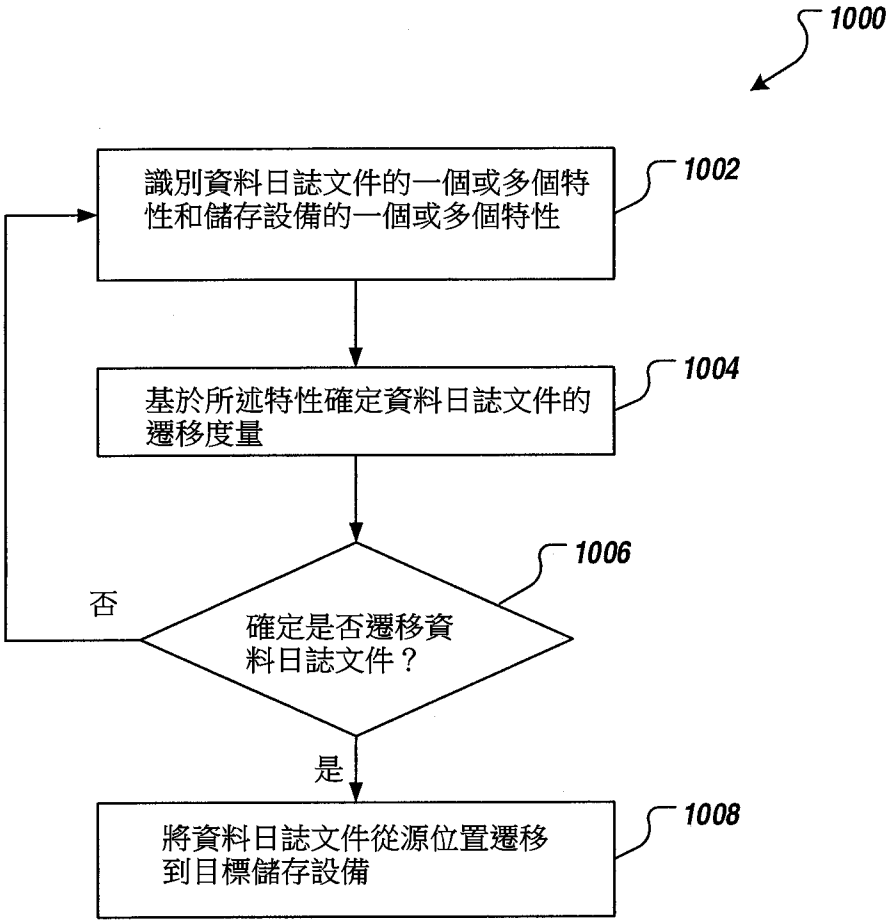
【圖 7】


800
802
維護多層儲存設備以及一層或多層快取
804
確定儲存在較低層儲存設備中的資料日誌文件中的資料對象是活躍資料對象
806
將資料對象寫入快取
808
產生指示資料對象在快取中的實體儲存位置的索引
810
將索引保存到記憶體中的增量索引映射
812
接收對資料對象的讀取請求
814
在記憶體中的增量索引映射中搜索與資料對象相對應的索引
816
從快取返回資料對象

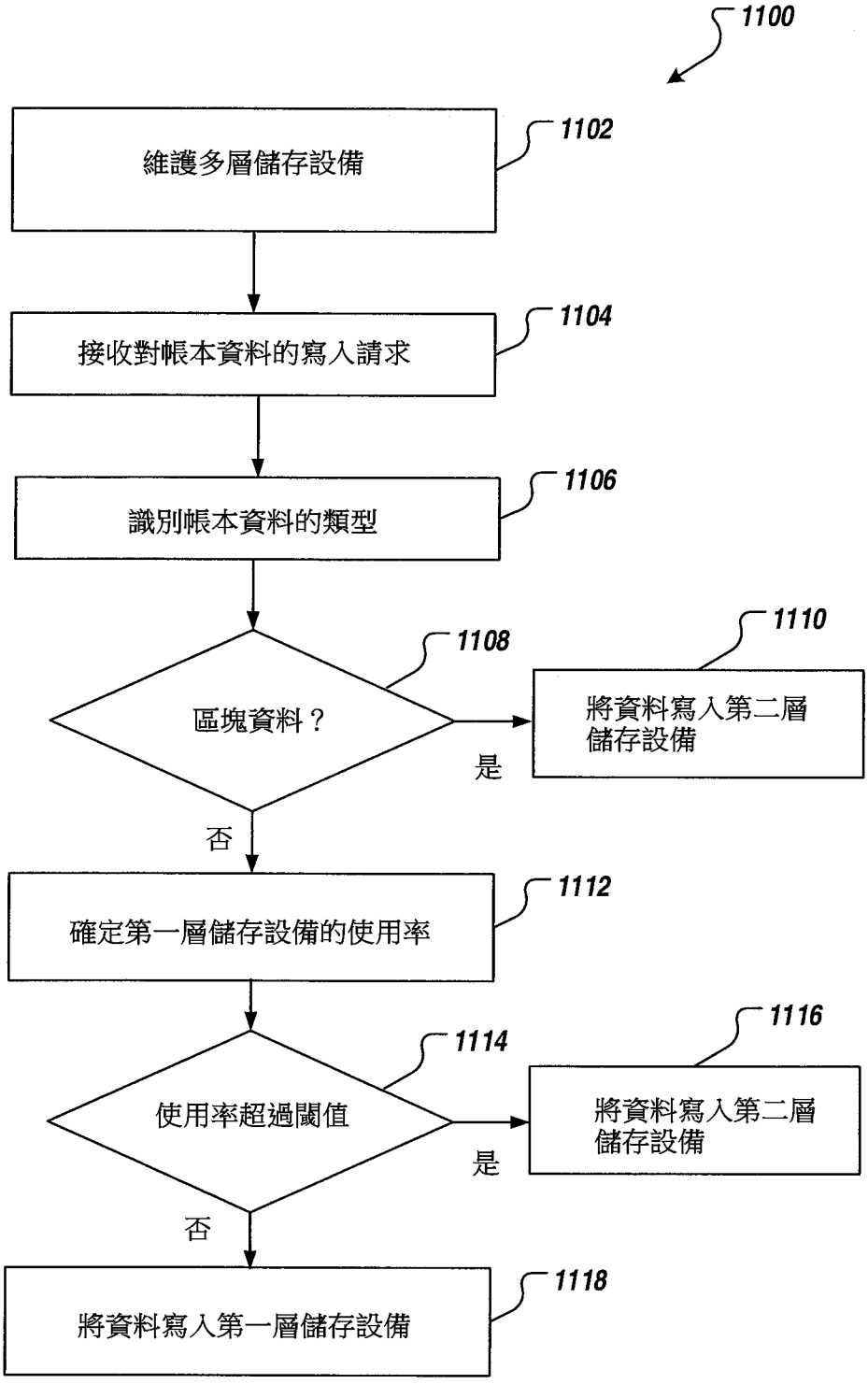
【圖 8】


900
從索引日誌文件中確定儲存在儲存設備中的資料日誌文件的資訊
902
確定儲存設備的資訊
904
資料日誌文件需要重寫放置？
906
否
是
從源位置讀取資料日誌文件並將資料日誌文件重寫入目標位置
908
將資料日誌文件密封在目標位置
910
產生與目標位置中的資料日誌文件相對應的目標索引日誌文件
912
密封目標索引日誌文件
914
將目標索引日誌文件導入到記憶體中的可讀索引映射中
916

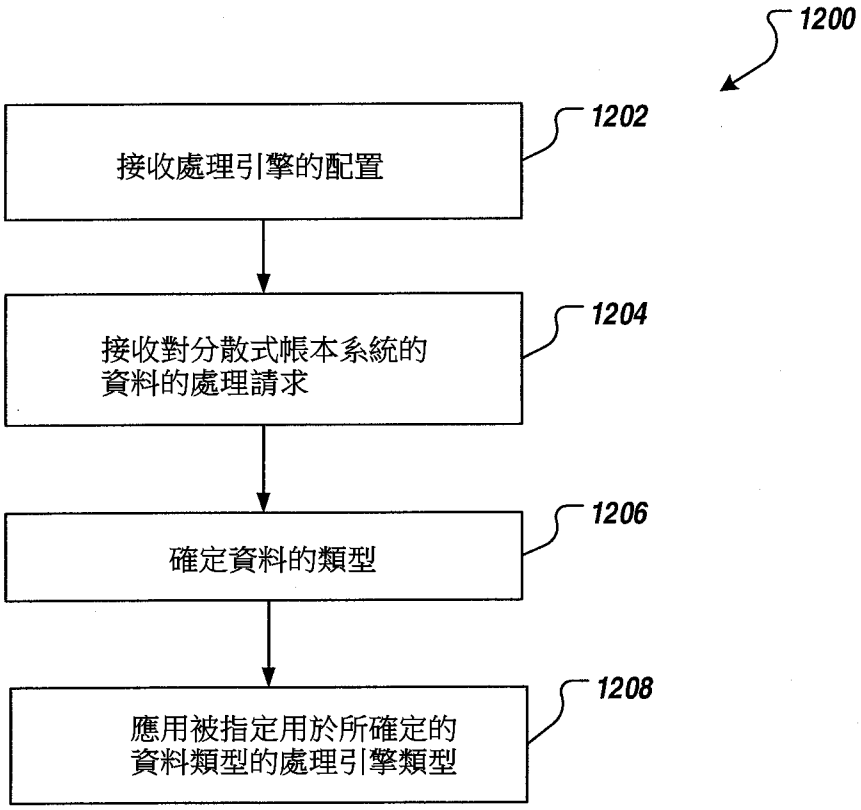
【圖 9】


1000
識別資料日誌文件的一個或多個特性和儲存設備的一個或多個特性
1002
基於所述特性確定資料日誌文件的遷移度量
1004
確定是否遷移資料日誌文件？
1006
否
是
將資料日誌文件從源位置遷移到目標儲存設備
1008

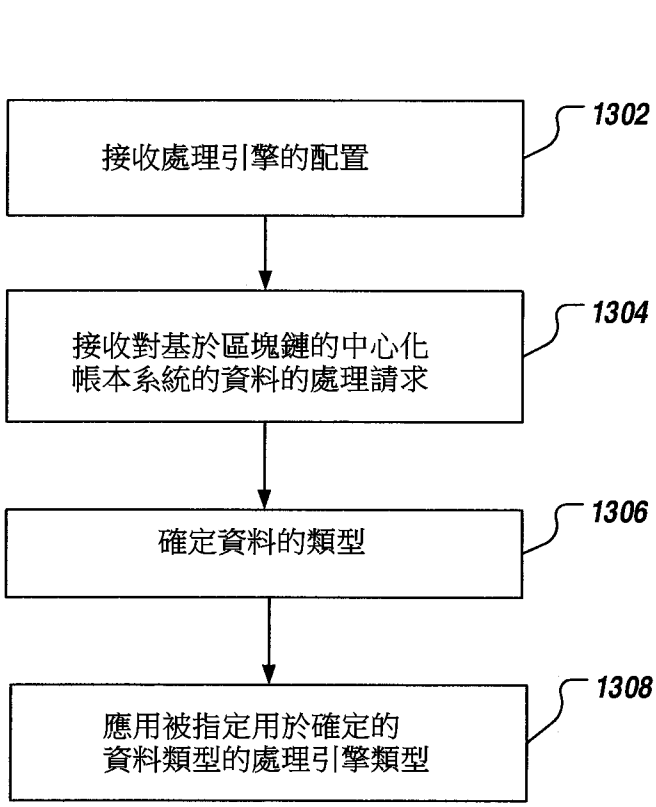
【圖 10】


1100
維護多層儲存設備
1102
接收對帳本資料的寫入請求
1104
識別帳本資料的類型
1106
區塊資料？
1108
是
將資料寫入第二層儲存設備
1110
否
確定第一層儲存設備的使用率
1112
使用率超過閾值
1114
是
將資料寫入第二層儲存設備
1116
否
將資料寫入第一層儲存設備
1118

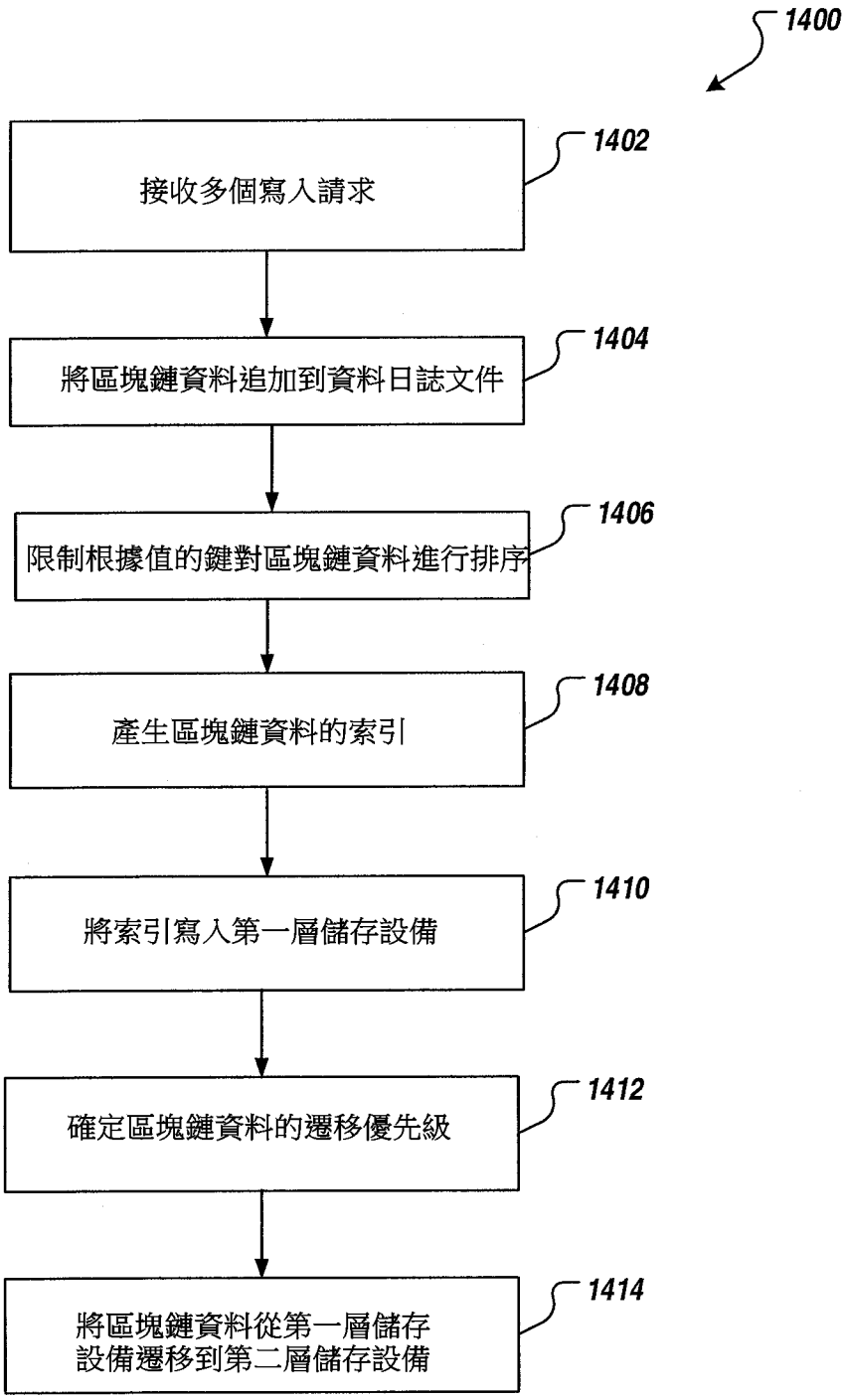
【圖 11】


1200
1202
接收處理引擎的配置
1204
接收對分散式帳本系統的
資料的處理請求
1206
確定資料的類型
1208
應用被指定用於所確定的
資料類型的處理引擎類型

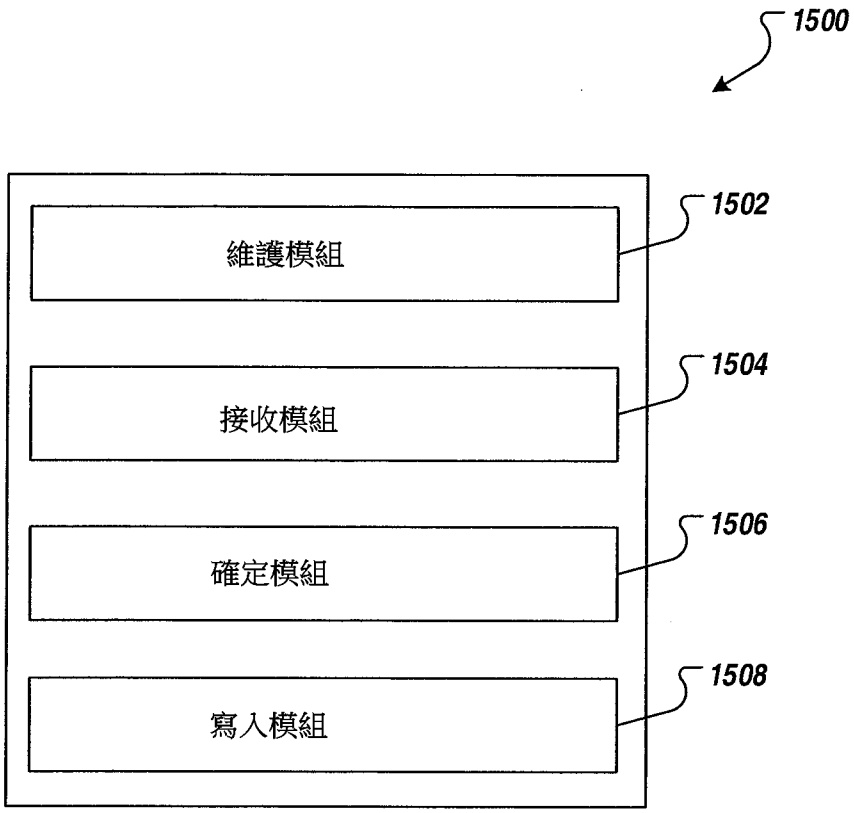
【圖 12】


1300
接收處理引擎的配置
1302
接收對基於區塊鏈的中心化帳本系統的資料的處理請求
1304
確定資料的類型
1306
應用被指定用於確定的資料類型的處理引擎類型
1308

【圖 13】


1400
接收多個寫入請求
1402
將區塊鏈資料追加到資料日誌文件
1404
限制根據值的鍵對區塊鏈資料進行排序
1406
產生區塊鏈資料的索引
1408
將索引寫入第一層儲存設備
1410
確定區塊鏈資料的遷移優先級
1412
將區塊鏈資料從第一層儲存設備遷移到第二層儲存設備
1414

【圖 14】



【圖 15】