



(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 I745353 B

(45)公告日：中華民國 110 (2021) 年 11 月 11 日

(21)申請案號：106106752

(22)申請日：中華民國 106 (2017) 年 03 月 02 日

(51)Int. Cl. : G06F13/16 (2006.01)

G06F3/06 (2006.01)

(30)優先權：2016/04/13 美國

62/322,035

2016/07/13 美國

15/209,566

(71)申請人：南韓商三星電子股份有限公司 (南韓) SAMSUNG ELECTRONICS CO., LTD. (KR)  
南韓(72)發明人：葛拉普迪 文卡塔 巴努 普拉卡什 GOLLAPUDI, VENKATA BHANU PRAKASH  
(IN)；雅庫拉 維傑亞 JAKKULA, VIJAYA (IN)

(74)代理人：林孟閱；盧佩君；陳怡如

(56)參考文獻：

TW I537831B

US 6240335B1

US 2005/0273190A1

US 2008/0120489A1

審查人員：易昶霈

申請專利範圍項數：24 項 圖式數：9 共 46 頁

(54)名稱

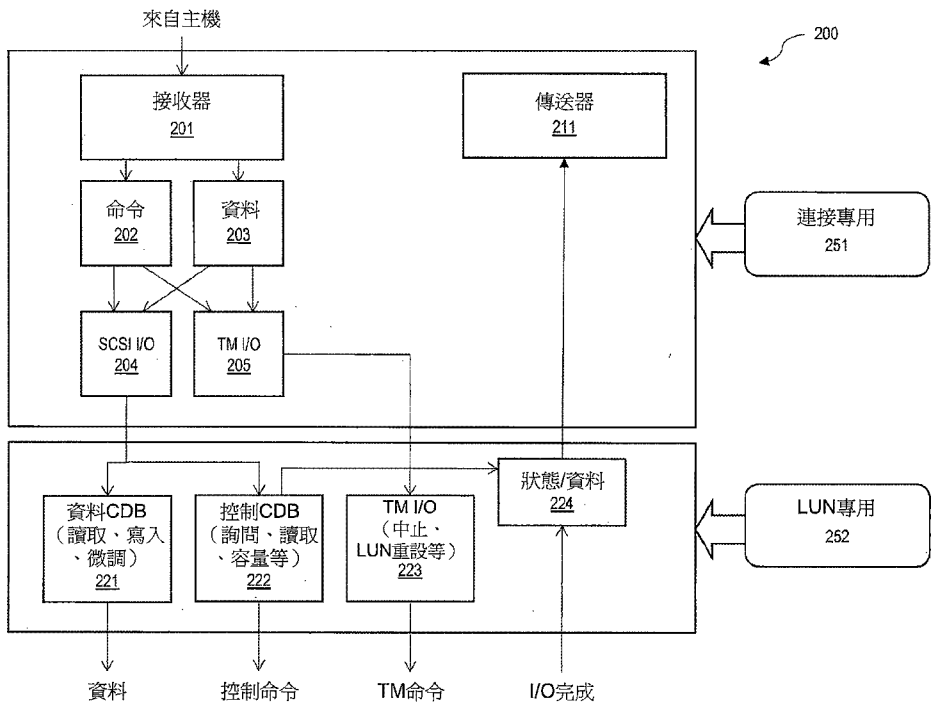
提供高效能無鎖可分級目標之方法與系統

(57)摘要

一種對儲存於資料儲存系統的目標中的資料進行存取的方法包括：在第一中央處理單元(CPU)群組上運行多個輸送執行緒，其中每一輸送執行緒包括命令接收器路徑及命令傳送器路徑；在第二CPU群組上運行多個資料路徑執行緒，其中每一資料路徑執行緒包括命令發出路徑及命令完成路徑；使用輸送執行緒的命令接收器路徑將輸入/輸出(I/O)命令投遞至 I/O 命令發出隊列，並使用資料路徑執行緒的命令發出路徑處理所述 I/O 命令；以及使用資料路徑執行緒的命令完成路徑將 I/O 完成通知投遞至 I/O 完成隊列，並使用輸送執行緒的命令傳送器路徑處理所述 I/O 出完成通知。

A method for accessing data stored in a target of a data storage system includes: running a plurality of transport threads on a first group of CPUs, wherein each of the plurality of transport threads comprises a command receiver path and a command transmitter path; running a plurality of data path threads on a second group of CPUs, wherein each of the plurality of data path threads comprises a command issue path and a command completion path; posting an I/O command to an I/O command issue queue using the command receiver path of a transport thread, and processing the I/O command using the command issue path of a data path thread; and posting an I/O completion notification to an I/O completion queue using the command completion path of the data path thread, and processing the I/O completion notification using the command transmitter path of the transport thread.

指定代表圖：



【圖2】

符號簡單說明：

- 200:資料路徑
- 201:接收器
- 202:網際網路小電腦系統介面命令
- 203:資料/相關聯資料
- 204:小電腦系統介面輸入/輸出模組
- 205:任務管理輸入/輸出模組
- 211:傳送器
- 221:資料命令描述符區塊
- 222:控制命令描述符區塊
- 223:任務管理輸入/輸出區塊
- 224:狀態/資料區塊
- 251:連接專用路徑
- 252:邏輯單元號碼專用路徑



I745353

## 【發明摘要】

【中文發明名稱】提供高效能無鎖可分級目標之方法與系統

【英文發明名稱】SYSTEM AND METHOD FOR HIGH  
PERFORMANCE LOCKLESS SCALABLE TARGET

【中文】一種對儲存於資料儲存系統的目標中的資料進行存取的方法包括：在第一中央處理單元（CPU）群組上運行多個輸送執行緒，其中每一輸送執行緒包括命令接收器路徑及命令傳送器路徑；在第二 CPU 群組上運行多個資料路徑執行緒，其中每一資料路徑執行緒包括命令發出路徑及命令完成路徑；使用輸送執行緒的命令接收器路徑將輸入/輸出（I/O）命令投遞至 I/O 命令發出隊列，並使用資料路徑執行緒的命令發出路徑處理所述 I/O 命令；以及使用資料路徑執行緒的命令完成路徑將 I/O 完成通知投遞至 I/O 完成隊列，並使用輸送執行緒的命令傳送器路徑處理所述 I/O 出完成通知。

【英文】A method for accessing data stored in a target of a data storage system includes: running a plurality of transport threads on a first group of CPUs, wherein each of the plurality of transport threads comprises a command receiver path and a command transmitter path; running a plurality of data path threads on a second

group of CPUs, wherein each of the plurality of data path threads comprises a command issue path and a command completion path; posting an I/O command to an I/O command issue queue using the command receiver path of a transport thread, and processing the I/O command using the command issue path of a data path thread; and posting an I/O completion notification to an I/O completion queue using the command completion path of the data path thread, and processing the I/O completion notification using the command transmitter path of the transport thread.

【指定代表圖】圖2。

【代表圖之符號簡單說明】

200：資料路徑

201：接收器

202：網際網路小電腦系統介面命令

203：資料/相關聯資料

204：小電腦系統介面輸入/輸出模組

205：任務管理輸入/輸出模組

211：傳送器

221：資料命令描述符區塊

222：控制命令描述符區塊

223：任務管理輸入/輸出區塊

224：狀態/資料區塊

106-05-22

251：連接專用路徑

252：邏輯單元號碼專用路徑

【特徵化學式】

無

## 【發明說明書】

【中文發明名稱】提供高效能無鎖可分級目標之方法與系統

【英文發明名稱】SYSTEM AND METHOD FOR HIGH

PERFORMANCE LOCKLESS SCALABLE TARGET

[相關申請案的交叉參考]

【0001】 本申請案主張於 2016 年 4 月 13 日提出申請且序列號為 62/322,035 的美國臨時專利申請案的權利及優先權，所述美國臨時專利申請案的揭露內容全文併入本案供參考。

【技術領域】

【0002】 本發明大體而言是有關於一種資料儲存系統，更具體而言是有關於一種提供高效能無鎖可分級目標的系統與方法。

【先前技術】

【0003】 在典型的儲存區域網路（storage area network，SAN）中，目標藉由光纖連接（Fiber connection）或開關網路（switching network）而將持久性資料儲存空間（例如，邏輯單元號碼（logical unit number，LUN）、命名空間）暴露給一或多個發起者（initiator）。發起者是指發起介面會話（例如，小電腦系統介面（small computer system interface，SCSI）會話）並發送命令（例如，小電腦系統介面命令）的端點。目標是指等待發起者的工作請求並執行輸入/輸出（input/output，I/O）操作的端點。在通用電腦架構的上下文中，發起者可被稱作客戶端，且目標可被稱作伺服器。單一目標可服務於多個發起者並為所述發起者提供一或多個邏輯單元號碼。

【0004】 一個目標與一或多個發起者可藉由相互認可的儲存區域網路協定進行通訊。儲存區域網路協定的實例包括但不限於光纖通道協定 (Fibre Channel Protocol, FCP)、並列小電腦系統介面 (parallel SCSI, pSCSI)、網際網路小電腦系統介面 (Internet small computer system interface, iSCSI)、超小電腦系統介面 (HyperSCSI)、光纖通道 (Fibre Channel)、先進技術附接 (Advanced Technology Attachment, ATA)、串列先進技術附接 (Serial ATA, SATA)、基於乙太網路的先進技術附接 (ATA over Ethernet, AoE)、無限帶頻 (InfiniBand) 及基於纖維的快速非揮發性記憶體 (Non-Volatile Memory Express, NVMe)。儲存區域網路協定使得發起者能夠向目標發送輸入/輸出命令。資料中心的資料儲存裝置可包括多個目標，所述多個目標可實體地及/或邏輯地分佈於儲存區域網路上。儲存區域網路協定將各所述資料儲存裝置合併成目標陣列、同時向主機提供各所述資料儲存裝置似乎在本地進行附接此一錯覺。

【0005】 儲存目標通常可藉由多個連接而將後端邏輯單元號碼 (backend LUN) 暴露給多個發起者。每一發起者可與目標建成一或多個連接並對所述目標中的一或多個邏輯單元號碼進行存取。從輸入/輸出 (I/O) 流角度來看，在資料路徑中建立的多個同步點可能會影響輸入/輸出效能，當多個會話並列地對多個邏輯單元號碼進行存取時尤其如此。

【0006】 前端輸入/輸出堆疊 (frontend I/O stack) 中的資料路徑可

被分成輸送協定層及小電腦系統介面核心層。在輸送協定層中進行的處理是基於每一連接的，而在小電腦系統介面核心層中進行的處理是基於每一邏輯單元號碼的。在輸送協定層中進行的處理包括在特定連接上發出輸入/輸出命令並在此特定連接上完成所述輸入/輸出命令。在小電腦系統介面核心層中進行的處理包括將輸入/輸出命令發出至特定邏輯單元號碼並完成被發出至所述特定邏輯單元號碼的輸入/輸出命令。若整個輸入/輸出路徑是在輸送協定層的執行緒上下文中執行，則在邏輯單元號碼層階處及連接層階處可能建立有各種同步點，進而影響總體輸入/輸出效能。此乃因輸送協定層處的輸入/輸出上下文是連接專用的（connection-specific），而小電腦系統介面核心層中的輸入/輸出上下文是邏輯單元號碼專用的（LUN-specific）。

### 【發明內容】

【0007】 根據一個實施例，一種方法包括：接收對儲存於資料儲存系統的目標中的資料進行存取的命令，其中所述資料儲存系統包括多個中央處理單元；在所述多個中央處理單元中的第一中央處理單元群組上運行多個輸送執行緒，其中所述多個輸送執行緒中的每一者包括命令接收器路徑及命令傳送器路徑；在所述多個中央處理單元中的第二中央處理單元群組上運行多個資料路徑執行緒，其中所述多個資料路徑執行緒中的每一者包括命令發出路徑及命令完成路徑；使用輸送執行緒的所述命令接收器路徑將輸入/輸出命令投遞至輸入/輸出命令發出隊列，並使用資料路徑執行

緒的所述命令發出路徑處理所述輸入/輸出命令；以及使用所述資料路徑執行緒的所述命令完成路徑將輸入/輸出完成通知投遞至輸入/輸出完成隊列，並使用所述輸送執行緒的所述命令傳送器路徑處理所述輸入/輸出完成通知。所述輸入/輸出命令發出隊列包括第一隊列陣列，且所述第一隊列陣列中的每一者對應於所述第一中央處理單元群組中的相應中央處理單元。所述輸入/輸出完成隊列包括第二隊列陣列，且所述第二隊列陣列中的每一者對應於所述第二中央處理單元群組中的相應中央處理單元。

【0008】 根據另一實施例，一種資料儲存系統包括：多個中央處理單元；多個目標；以及記憶體，儲存用於處理輸入/輸出流以及輸入/輸出命令發出隊列、及輸入/輸出完成隊列的一組軟體模組。所述一組軟體模組用以：接收對儲存於資料儲存系統的目標中的資料進行存取的輸入/輸出命令，其中所述資料儲存系統包括多個中央處理單元；在所述多個中央處理單元中的第一中央處理單元群組的第一中央處理單元上運行多個輸送執行緒，其中所述多個輸送執行緒中的每一者包括命令接收器路徑及命令傳送器路徑；以及在所述多個中央處理單元中的第二中央處理單元群組的第二中央處理單元上運行多個資料路徑執行緒，其中所述多個資料路徑執行緒中的每一者包括命令發出路徑及命令完成路徑。輸送執行緒的所述命令接收器路徑將輸入/輸出命令投遞至輸入/輸出命令發出隊列，且資料路徑執行緒的所述命令發出路徑處理所述輸入/輸出命令。所述資料路徑執行緒的所述命令完成路徑將輸入/

輸出完成通知傳遞至所述輸入/輸出完成隊列，且所述輸送執行緒的所述命令傳送器路徑處理所述輸入/輸出完成通知。所述輸入/輸出命令發出隊列包括第一隊列陣列，且所述第一隊列陣列中的每一者對應於所述第一中央處理單元群組中的相應中央處理單元。所述輸入/輸出完成隊列包括第二隊列陣列，且所述第二隊列陣列中的每一者對應於所述第二中央處理單元群組中的相應中央處理單元。

【0009】 現將參照附圖來更具體地闡述並在申請專利範圍中指出以上及其他較佳特徵（包括實施方案的各種新穎細節及事件組合）。應理解，本文所述特定系統及方法僅用於說明而非用於限制。如熟習此項技術者所應理解，在各種各樣的諸多實施例中均可採用本文所述原理及特徵，而此並不背離本發明的範圍。

#### 【圖式簡單說明】

【0010】 作為本說明書的一部分而包括在內的各個附圖說明當前較佳的實施例，且與以上所給出的大體說明及以下所給出的對較佳實施例的詳細說明一起用於闡釋及教示本文所述原理。

#### 【0011】

圖 1 說明根據一個實施例的執行緒模型中的單片資料路徑（monolithic data path）的示例圖。

圖 2 說明根據一個實施例的示例性執行緒模型的方塊圖。

圖 3 說明根據一個實施例的示例性劃分執行緒模型。

圖 4 說明根據一個實施例的示例性執行緒模型的示意圖。

圖 5 說明根據一個實施例的示例性單一生產者單一消費者（single-producer single-consumer，SPSC）無鎖隊列的方塊圖。

圖 6 說明根據一個實施例的示例性多生產者多消費者（multi-producer multi-consumer，MPMC）無鎖隊列的方塊圖。

圖 7 說明根據一個實施例的示例性執行緒模型的方塊圖。

圖 8 說明根據另一實施例的示例性多生產者多消費者無鎖隊列的方塊圖。

圖 9 說明根據又一實施例的示例性多生產者多消費者無鎖隊列的方塊圖。

**【0012】** 各個圖未必是按比例繪製，且出於說明目的，在所有圖中具有相似結構或功能的元件一般是由相同的參考編號表示。各個圖僅旨在方便闡述本文所述各種實施例。各個圖並不闡述本文所揭露教示內容的每一個態樣且並不限制申請專利範圍的範圍。

#### **【實施方式】**

**【0013】** 本文所揭露的特徵及教示內容中的每一者可單獨使用或結合其他特徵及教示內容來使用以提供高效能無鎖可分級目標。參照附圖更詳細地闡述代表性實例，所述代表性實例單獨地使用及組合地使用該些附加特徵及教示內容中的諸多特徵及教示內容。此詳細說明僅旨在教示熟習此項技術者關於實踐本教示內容的態樣的進一步細節而並非旨在限制申請專利範圍的範圍。因此，在本詳細說明中以上所揭露特徵的各種組合可能未必是在最廣泛意義上實踐本教示內容所必需的，而是相反僅是為了具體闡

述本教示內容的代表性實例而教示。

【0014】 在以下說明中，僅出於闡釋目的而闡述特定術語以提供對本發明的透徹理解。然而，對於熟習此項技術者而言將顯而易見，該些特定細節並非是實踐本發明的教示內容所必需的。

【0015】 本文中的詳細說明的某些部分是以演算法及對電腦記憶體內的資料位元進行的運算的符號表示法來展現。該些演算法描述及表示法被熟習資料處理技術的技術人員用於向其他熟習此項技術者有效地傳達其工作的實質。演算法在此處且一般而言均被視作能得到所期望結果的步驟的自治序列（self-consistent sequence）。所述步驟需要對物理量進行物理調處。通常（儘管未必必需），該些量會呈現能夠被儲存、傳遞、組合、比較、及以其他方式被調處的電性訊號或磁性訊號的形式。業已證明，主要出於通用的原因，將該些訊號稱作位元、值、元素、符號、字元、項、數字等有時是方便的。

【0016】 然而，應記住，該些用語中的所有者及所有相似用語皆欲與適宜的物理量相關聯且僅作為應用於該些量的便利標籤。除非藉由閱讀以下論述顯而易見地另有具體說明，否則應知在本說明通篇中，使用例如「處理（processing）」、「計算（computing）」、「運算（calculating）」、「判斷（determining）」、「顯示（displaying）」等用語進行的論述是指電腦系統或相似電子計算裝置的動作及處理，所述電腦系統或相似電子計算裝置調處在電腦系統的暫存器及記憶體內被表示為物理（電子）量的資料並將所述資料轉換成

在電腦系統記憶體或暫存器或者其他此類資訊儲存裝置、資訊傳輸裝置或資訊顯示裝置內被相似地表示為物理量的其他資料。

【0017】 本文中所展現的演算法並非固有地有關於任何特定電腦或其他設備。可根據本文中的教示內容將各種通用系統、電腦伺服器、或個人電腦與程式一起使用，或者構建更專門的設備來執行所需方法步驟可被證明是方便的。該些系統中的各種系統的所需結構將自以下說明顯而易見。應知，可使用各種程式化語言來實作本文所述發明的教示內容。

【0018】 此外，代表性實例及附屬請求項的各個特徵可以並未具體地及明確地枚舉的方式加以組合以提供本教示內容的額外的有用實施例。亦應明確注意，出於原始揭露內容的目的以及出於限制所主張主題的目的，所有值範圍或對實體的群組的指示均揭露每個可能的中間值或中間實體。亦應明確注意，圖中所示各組件的尺寸及形狀被設計成有助於理解如何實踐本教示內容，而並非旨在限制實例中所示尺寸及形狀。

【0019】 圖 1 說明根據一個實施例的單片執行緒模型中的資料路徑的示例圖。資料路徑包括由網際網路小電腦系統介面接收器路徑 101、小電腦系統介面命令發出路徑 102、小電腦系統介面命令完成路徑 103、及網際網路小電腦系統介面傳送器路徑 104 構成的單片序列。儲存系統可具有多個中央處理單元（central processing unit，CPU）（例如，N 個中央處理單元）且所述儲存系統中的每一中央處理單元負責在資料路徑的整個生命週期期間建立、處

理、管理、及完成相應資料路徑。在資料路徑 101 至 104 中，資料路徑 101 及 104 是建立於發起者與小電腦系統介面目標之間的輸送協定（即，網際網路小電腦系統介面）中且是基於每一連接的。舉例而言，登入/登出（login/logout）、緩衝器管理（buffer management）、及任務管理處置（task management handling）是基於每一連接而在會話/連接層階處執行。輸入/輸出命令是在會話/連接層階處進行排隊及跟蹤。在資料路徑 101 至 104 中，資料路徑 102 及 103 是在邏輯單元號碼層階處建立於小電腦系統介面目標與邏輯單元號碼之間。舉例而言，往來於目標儲存裝置的小電腦系統介面命令以及錯誤處置（error handling）是在邏輯單元號碼層階處進行跟蹤。根據圖 1 中所示單片執行緒模型，目標系統的可用中央處理單元資源被共享以運行輸送協定及小電腦系統介面核心層階協定二者。單片執行緒模型不會使快取局部性（cache locality）最大化。

【0020】本發明提供一種對輸送協定及小電腦系統介面核心協定進行劃分並在各單獨的執行緒上獨立地處置所述輸送協定及小電腦系統介面核心協定的新穎執行緒模型。本執行緒模型在輸送協定執行緒與小電腦系統介面核心執行緒之間採用無鎖隊列設計。無鎖隊列設計會藉由對輸送協定及小電腦系統介面核心協定進行劃分並針對輸送協定及小電腦系統介面核心協定獨立地運行執行緒而提高輸入/輸出效能。儲存系統的中央處理單元資源分佈於輸送協定層與小電腦系統介面核心層之間。輸送協定執行緒僅排程

於被指配成對在輸送協定層中運行的執行緒進行處置的那些中央處理單元上。小電腦系統介面核心執行緒則僅排程於被指配成對在小電腦系統介面核心層中運行的執行緒進行處置的中央處理單元上。輸送協定執行緒處理接收路徑及傳送路徑。小電腦系統介面核心執行緒處理對特定邏輯單元號碼的輸入/輸出請求及關於此特定邏輯單元號碼的輸入/輸出完成。

【0021】 本系統及方法利用多個邏輯單元號碼及其連接來提供高分級性 (scalability)。本系統及方法進一步將輸送協定執行緒與邏輯單元號碼執行緒分開以使快取局部性最大化。此外，無鎖隊列設計會在經由多個連接來存取多個邏輯單元號碼時消除鎖競爭 (lock contention)。儘管針對網際網路小電腦系統介面目標來闡述本發明，然而應理解，本發明可應用於任何儲存區域網路協定 (例如，光纖通道協定、並列小電腦系統介面、網際網路小電腦系統介面、超小電腦系統介面、光纖通道、先進技術附接、串列先進技術附接、基於乙太網路的先進技術附接、無限帶頻、及基於纖維的快速非揮發性記憶體) 以高效地使用系統資源及提供高效能目標輸入/輸出效能，進而避免同步瓶頸 (synchronization bottleneck)。舉例而言，在基於纖維的快速非揮發性記憶體協定中，與小電腦系統介面核心層等效的核心層可被稱作快速非揮發性記憶體排隊介面及命令集合。

【0022】 圖 2 說明根據一個實施例的示例性執行緒模型中的資料路徑的方塊圖。資料路徑 200 可被劃分成連接專用路徑 251 及邏

輯單元號碼專用路徑 252。連接專用路徑 251 包括接收器 201 及傳送器 211。發起者（客戶端）的接收器 201 自主機接收指令，產生網際網路小電腦系統介面命令 202 及相關聯資料 203，建立與目標的連接，並藉由小電腦系統介面輸入/輸出模組 204 及任務管理（task management，TM）輸入/輸出模組 205 而向所述目標發送網際網路小電腦系統介面命令 202 及資料 203。

【0023】邏輯單元號碼專用路徑 252 包括用於儲存與小電腦系統介面命令（例如，讀取、寫入、及微調（trim））相關聯的資料的資料命令描述符區塊（command descriptor block，CDB）221、及用於儲存小電腦系統介面命令（例如，詢問、讀取、及容量）的控制命令描述符區塊 222。網際網路小電腦系統介面命令及所述資料分別儲存於資料命令描述符區塊 221 及控制命令描述符區塊 222 中。邏輯單元號碼專用路徑 252 更包括用於儲存小電腦系統介面管理命令（例如，中止（abort）、邏輯單元號碼重設）的任務管理輸入/輸出區塊 223、以及狀態/資料區塊 224。自主機接收到的任務管理命令可儲存於任務管理輸入/輸出區塊 223 中。所述目標的狀態及自所述目標接收到的相關聯資料儲存於狀態/資料區塊 224 中。控制命令描述符區塊 222 可直接對狀態/資料區塊 224 進行存取以更新特定控制命令的狀態及資料。所述目標的狀態/資料資訊（例如，輸入/輸出完成）被發送回連接專用路徑 251 的傳送器 211。

【0024】圖 3 說明根據一個實施例的示例性劃分執行緒模型。目

標系統可包括一定數目的中央處理單元，所述一定數目的中央處理單元是基於工作負載的類型而分佈以在網際網路小電腦系統介面輸送協定與小電腦系統介面命令之間進行處置。連接專用執行緒（即，輸送協定中的網際網路小電腦系統介面）包括位於發起者與小電腦系統介面目標之間的網際網路小電腦系統介面接收器路徑 301 及網際網路小電腦系統介面傳送器路徑 304。連接專用執行緒是基於每一連接的。小電腦系統介面命令發出路徑 302 及小電腦系統介面命令完成路徑 303 是建立於邏輯單元號碼層階處。在目標系統中的所述一定數目的可用中央處理單元中，M 個中央處理單元被指配用於建立資料路徑、在發起者與小電腦系統介面目標之間輸送小電腦系統介面命令、及在輸送協定層處完成所建立資料路徑，而 N 個中央處理單元則被指配用於在小電腦系統介面目標與邏輯單元號碼之間處理、管理、及處置小電腦系統介面命令。

【0025】 發起者與小電腦系統介面目標之間的網際網路小電腦系統介面協定運行針對每一連接建立的連接專用執行緒且包括由網際網路小電腦系統介面接收器路徑 301 及網際網路小電腦系統介面傳送器路徑 304 構成的序列。連接專用執行緒被指配給所述 M 個中央處理單元中的一或多者。網際網路小電腦系統介面接收器路徑 301 對包括小電腦系統介面命令發出路徑 302（即，小電腦系統介面接收器）的小電腦系統介面輸入/輸出請求進行排隊。在完成小電腦系統介面命令之後，小電腦系統介面目標對小電腦系統

介面命令完成路徑 303（即，小電腦系統介面傳送器）進行排隊。基於每一邏輯單元號碼而建立的小電腦系統介面命令發出路徑 302 及小電腦系統介面命令完成路徑 303 被指配給所述 N 個中央處理單元。在完成小電腦系統介面命令之後，小電腦系統介面目標在先前建立的位於發起者與小電腦系統介面目標之間的輸送連接上將輸入/輸出完成（例如，網際網路小電腦系統介面傳送器路徑 304）排隊至所述發起者。最後，由在發起者與小電腦系統介面目標之間建立所述連接專用執行緒的中央處理單元來處置輸入/輸出完成。

【0026】圖 4 說明根據一個實施例的示例性執行緒模型的示意圖。本執行緒模型提供資料路徑，所述資料路徑包括網際網路小電腦系統介面接收器路徑 401、網際網路小電腦系統介面傳送器路徑 404、小電腦系統介面命令發出路徑 402、及小電腦系統介面命令完成路徑 403。在輸入/輸出命令發出路徑上，網際網路小電腦系統介面接收器路徑 401 使用輸入/輸出命令發出隊列 410 將輸入/輸出命令投遞至小電腦系統介面命令發出路徑 402。在輸入/輸出命令返回路徑上，小電腦系統介面命令完成路徑 403 使用輸入/輸出完成隊列 411 將輸入/輸出完成投遞至網際網路小電腦系統介面傳送器路徑 404。根據一個實施例，輸入/輸出命令發出隊列 410 及輸入/輸出完成隊列 411 均為多生產者及多消費者（MPMC）無鎖隊列，在所述無鎖隊列中，來自多個連接的輸入/輸出命令可到達單一邏輯單元號碼且來自單一連接的輸入/輸出命令可到達多個

邏輯單元號碼。

【0027】 根據一個實施例，在隊列的上下文中，生產者是指發起者，且消費者是指目標。在某些實施例中，生產者可指目標，且消費者可指發起者。舉例而言，網際網路小電腦系統介面接收器路徑 401 及網際網路小電腦系統介面傳送器路徑 404 可由生產者擁有，且小電腦系統介面命令發出路徑 402 及小電腦系統介面命令完成路徑 403 可由消費者擁有。在另一實例中，小電腦系統介面命令發出路徑 402 及小電腦系統介面命令完成路徑 403 可由生產者擁有，且網際網路小電腦系統介面接收器路徑 401 及網際網路小電腦系統介面傳送器路徑 404 可由消費者擁有。

【0028】 在多生產者/多消費者（MPMC）隊列中，生產者任務及消費者任務均在多個執行緒上運行。舉例而言，生產者任務在  $n$  個執行緒上運行，且消費者任務在  $m$  個執行緒上運行。可存在與特定生產者執行緒綁定的多個生產者、及與特定消費者執行緒綁定的多個消費者。在輸入/輸出發出路徑中，輸送協定層產生髮至小電腦系統介面核心層的工作請求。在此種情形中，輸送協定層為生產者，且小電腦系統介面核心層為消費者。另一方面，在輸入/輸出完成路徑中，小電腦系統介面核心層產生髮至輸送協定層的工作請求。在此種情形中，小電腦系統介面核心層為生產者，且輸送協定層為消費者。由針對每一連接的輸送協定層產生的工作請求可去往多個邏輯單元號碼。相似地，由針對每一邏輯單元號碼的小電腦系統介面核心層產生的工作請求可去往多個連接。

輸送協定層與小電腦系統介面核心層之間的通訊具有通常需要鎖的同步點。根據一個實施例，本系統使得輸送協定層及小電腦系統介面核心層能夠以無鎖方式被存取。

【0029】 圖 5 說明根據一個實施例的示例性單一生產者單一消費者（SPSC）無鎖隊列的方塊圖。無鎖隊列中的每一節點包括資料容器（data container）及指針（pointer），且一系列節點被鏈結於鏈結列表中。根據一個實施例，圖 4 所示輸入/輸出命令發出隊列 410 及輸入/輸出完成隊列 411 中的每一者可具有圖 5 中所示無鎖隊列。本無鎖隊列具有首部節點 501、尾部節點 503。若所述隊列中僅存在單一節點，則首部節點 501 與尾部節點 503 為同一節點。當前節點 502 是指列表中的起點，消費者可自所述起點開始列表遍歷（list traversal）並消費所述節點。在某些實施例中，當前節點 502 可被稱作消費者起始部（consumer start）或消費者首部（consumer head）（與由生產者擁有的首部節點 501 相對）。

【0030】 圖 5 中所示鏈結列表中的每一節點具有資料容器 511 及指向下一節點的指針 512。生產者創建新節點（例如，與新輸入/輸出命令對應的節點），將新創建的節點鏈結至隊列中的尾部節點 503，並將尾部節點 503 的指針更新成指向所述新節點。藉此，新節點可被添加至現有隊列。相似地，生產者可騰空自首部節點 501 至當前節點 502 的各所消費節點。生產者可在系統資源可用時與由消費者進行的節點消費過程獨立地騰空所消費節點。在此種意義上，騰空由生產者消費的節點的過程被稱作怠惰刪除（lazy

delete)。自首部節點 501 至當前節點 502 之前的一個節點的各節點是由生產者擁有，而自當前節點 502 至尾部節點 503 的各節點則是由消費者擁有。消費者遍歷自當前節點 502 至尾部節點 503 的列表，消費當前節點 502 處的項，並將當前指針更新至下一節點。若當前指針指向尾部節點 503，則消費者不會消費節點。

【0031】 本無鎖隊列提供容納用於管理無鎖列表的控制資訊的控制結構。生產者擁有第一指針及最末指針。最末指針是當生產者將新節點添加至無鎖列表時進行更新。第一指針是當生產者刪除所消費節點時進行更新。當消費者遍歷自當前節點 502 至尾部節點 503 的列表時，當前指針被所述消費者更新。由於無鎖隊列的當前指針的控制資訊及所有權能利用所述無鎖隊列的控制結構在生產者與消費者之間無縫互換，因此本無鎖隊列能消除對在資料路徑中存在同步點及鎖的需求並提高輸入/輸出效能，當多個會話並列地對多個邏輯單元號碼進行存取時尤其如此。

【0032】 圖 6 說明根據一個實施例的示例性多生產者多消費者（MPMC）無鎖隊列的方塊圖。多生產者多消費者無鎖隊列具有由  $N$  個首部節點（Head 1 至 Head  $N$ ）構成的控制陣列 601，其中  $N$  對應於供生產者在其上排程運行的執行緒的數目。在生產者在所述  $N$  個執行緒上排程運行的同時，消費者在  $M$  個執行緒上排程運行。生產者的數目與執行緒的數目  $N$  可不同，乃因單一生產者可在多個執行緒上運行。相似地，消費者的數目與執行緒的數目  $M$  可不同，乃因單一消費者可在多個執行緒上運行並消費由多個

生產者生產的節點。在控制陣列 601 上，生產者可生產節點，且消費者可消費所述節點。與所述  $N$  個執行緒中的每一者對應的控制結構維持單一生產者單一消費者無鎖隊列。消費者維持所述  $N$  個執行緒的控制陣列 601。

【0033】 根據一個實施例，多生產者多消費者無鎖隊列使用執行緒辨識符（**identifier**，**ID**）提供對執行緒的排他性存取。舉例而言，在給定執行緒上運行的生產者可能需要向對應隊列中生產新節點。當添加所述新節點時，生產者可藉由將隊列的執行緒辨識符（**ID**）索引至控制陣列 601 中來對所述隊列進行排他性存取。由於每一生產者需要獲得對隊列的排他性存取來生產新節點，因此在多個生產者之間不存在競爭（**contention**）。在給定執行緒上運行的消費者可消費屬於多個生產者的多個節點。相似地，由於每一消費者對控制陣列 601 中的隊列進行排他性存取，因此多個消費者之間不存在競爭。

【0034】 圖 7 說明根據一個實施例的示例性執行緒模型的方塊圖。本執行緒模型包括多個網際網路小電腦系統介面連接執行緒 701 及多個小電腦系統介面邏輯單元號碼執行緒 751。出於闡釋目的，網際網路小電腦系統介面連接執行緒 701 在三個中央處理單元上運行，且小電腦系統介面邏輯單元號碼執行緒 751 在兩個中央處理單元上運行。網際網路小電腦系統介面連接執行緒 701 中的每一者包括與運行小電腦系統介面邏輯單元號碼執行緒 751 的中央處理單元中的每一者對應的兩個節點（即，**Head 1a** 及 **Head**

2a)。小電腦系統介面邏輯單元號碼執行緒 751 中的每一者包括與運行網際網路小電腦系統介面連接執行緒 701 的中央處理單元中的每一者對應的三個節點（即，Head 1b、Head 2b、及 Head 3b）。儘管本實例包括被指配成分別運行三個網際網路小電腦系統介面連接執行緒 701 中的每一者的三個中央處理單元及被指配成分別運行小電腦系統介面邏輯單元號碼執行緒 751 中的每一者的兩個中央處理單元，然而應理解，本執行緒模型可應用於任何數目的網際網路小電腦系統介面連接執行緒及小電腦系統介面邏輯單元號碼執行緒，而此並不偏離本發明的範圍。每一網際網路小電腦系統介面連接執行緒（CT1、CT2、CT3）被指配給每一中央處理單元（CPU1、CPU2、及 CPU3）。相似地，每一小電腦系統介面邏輯單元號碼執行緒（LT1 及 LT2）被指配給每一中央處理單元（CPU4 及 CPU5）。所述 6 個網際網路小電腦系統介面連接 C1...C6 可由網際網路小電腦系統介面執行緒 CT1、CT2、及 CT3 提供服務。所述 4 個邏輯單元號碼 L1...L4 可由小電腦系統介面邏輯單元號碼執行緒 LT1 及 LT2 提供服務。

【0035】 生產者連接執行緒（例如，網際網路小電腦系統介面連接執行緒 701）直接向藉由相應中央處理單元辨識符（CPU-ID）進行索引的針對每一邏輯單元號碼的隊列（per-LUN queue）生產輸入/輸出命令（供處理的）。小電腦系統介面邏輯單元號碼執行緒 751 根據由藉由相應中央處理單元辨識符進行索引的網際網路小電腦系統介面連接執行緒 701 所產生的輸入/輸出命令而直接向針

對每一連接的隊列 (per-connection queue) 生產輸入/輸出命令 (已完成的)。消費者連接執行緒 (例如, 小電腦系統介面邏輯單元號碼執行緒) 消費屬於此連接的輸入/輸出命令。小電腦系統介面邏輯單元號碼執行緒 751 消費屬於所述邏輯單元號碼的輸入/輸出命令。網際網路小電腦系統介面連接執行緒 701 (生產者) 可直接向各別小電腦系統介面邏輯單元號碼執行緒 751 (消費者) 發出輸入/輸出命令, 且小電腦系統介面邏輯單元號碼執行緒 751 (生產者) 可直接向各別網際網路小電腦系統介面連接執行緒 701 (消費者) 發出輸入/輸出完成。每一輸入/輸出命令可藉由對於運行特定執行緒的處理器而言唯一的中央處理單元辨識符來辨識, 藉此消除對在獨立運行的執行緒中存在同步點或鎖的需求。

【0036】 圖 8 說明根據另一實施例的示例性多生產者多消費者無鎖隊列的方塊圖。在輸入/輸出發出路徑中, 生產者可為輸送協定層, 且消費者可為小電腦系統介面核心層。輸送協定執行緒 (例如, 圖 3 所示網際網路小電腦系統介面接收器路徑 301) 對  $N$  個中央處理單元的邏輯單元號碼陣列 801 進行存取, 其中  $N$  為對輸送協定執行緒指配的中央處理單元的數目。陣列索引是指用於在上面運行輸送協定執行緒的中央處理單元的數目 (1、2、...、 $N$ )。所述陣列索引中的每一者具有無鎖單一鏈結列表, 在所述無鎖單一鏈結列表上, 輸送協定執行緒產生對被發出至特定邏輯單元號碼的輸入/輸出命令的工作請求。小電腦系統介面核心執行緒 (例如, 圖 3 所示小電腦系統介面命令發出路徑 302) 是在處理工作表

項的對應小電腦系統介面核心中央處理單元上執行，所述工作表項由輸送協定執行緒進行排隊。小電腦系統介面核心執行緒更新在對應小電腦系統介面核心中央處理單元上運行的執行緒中的當前指針。小電腦系統介面模組（邏輯單元號碼）擁有如圖 8 中所述的資料結構且提供與網際網路小電腦系統介面模組（輸送協定層）的介面以生產及投遞輸入/輸出請求。

【0037】 圖 9 說明根據又一實施例的示例性多生產者多消費者無鎖隊列的方塊圖。在輸入/輸出完成路徑中，生產者可為小電腦系統介面核心層，且消費者可為輸送協定層。邏輯單元號碼執行緒（例如，圖 3 所示小電腦系統介面命令完成路徑 303）對  $M$  個中央處理單元的針對每一連接的陣列（per-connection array）901 進行存取，其中  $M$  為對小電腦系統介面核心執行緒指配的中央處理單元的數目。陣列索引是指在上面運行小電腦系統介面核心執行緒的中央處理單元的數目（1、2、...、 $M$ ）。所述陣列索引中的每一者具有無鎖單一鏈結列表，小電腦系統介面核心執行緒依據所述無鎖單一鏈結列表而產生對針對特定連接的已完成輸入/輸出命令的工作請求。在對應中央處理單元上執行的輸送協定執行緒（例如，圖 3 所示網際網路小電腦系統介面傳送器路徑 304）處理由所有小電腦系統介面核心執行緒進行排隊的輸入/輸出命令。網際網路小電腦系統介面執行緒更新在對應輸送協定中央處理單元上運行的網際網路小電腦系統介面執行緒中的當前指針。網際網路小電腦系統介面模組（網際網路小電腦系統介面連接）擁有如圖 9

中所述的資料結構且提供與小電腦系統介面模組（小電腦系統介面核心層）的介面以產生及投遞輸入/輸出完成。

**【0038】** 本系統及方法能將輸送處理與資料（或核心）處理分開。本系統及方法實作多生產者-多消費者無鎖設計，所述無鎖設計解決了輸送處理與資料處理所固有的同步問題。本系統及方法提供在輸送處理與資料處理之間共享的資源（例如，中央處理單元），以適應對輸送層及核心層中的每一者的資源的不同的且變化的需求。輸入/輸出寫入（I/O write，IOW）及各中央處理單元資源的比率可基於系統配置而變化。

**【0039】** 根據一個實施例，一種方法包括：接收對儲存於資料儲存系統的目標中的資料進行存取的命令，其中所述資料儲存系統包括多個中央處理單元；在所述多個中央處理單元中的第一中央處理單元群組上運行多個輸送執行緒，其中所述多個輸送執行緒中的每一者包括命令接收器路徑及命令傳送器路徑；在所述多個中央處理單元中的第二中央處理單元群組上運行多個資料路徑執行緒，其中所述多個資料路徑執行緒中的每一者包括命令發出路徑及命令完成路徑；使用輸送執行緒的所述命令接收器路徑將輸入/輸出命令投遞至輸入/輸出命令發出隊列，並使用資料路徑執行緒的所述命令發出路徑處理所述輸入/輸出命令；以及使用所述資料路徑執行緒的所述命令完成路徑將輸入/輸出完成通知投遞至輸入/輸出完成隊列，並使用所述輸送執行緒的所述命令傳送器路徑處理所述輸入/輸出完成通知。所述輸入/輸出命令發出隊列包括第

一隊列陣列，且所述第一隊列陣列中的每一者對應於所述第一中央處理單元群組中的相應中央處理單元。所述輸入/輸出完成隊列包括第二隊列陣列，且所述第二隊列陣列中的每一者對應於所述第二中央處理單元群組中的相應中央處理單元。

【0040】 所述輸送執行緒可包括接收器及傳送器，且所述接收器可向所述資料路徑發送輸入/輸出命令。

【0041】 所述資料路徑執行緒可將所述輸入/輸出命令發送至所述目標並自所述目標接收狀態及/或資料，並且將所述狀態及/或資料傳送至所述輸送執行緒的所述傳送器。

【0042】 所述第一隊列陣列可包括與被指配給所述資料路徑執行緒的所述第二中央處理單元群組對應的第一數目的節點。

【0043】 所述第一數目的節點可包括首部節點、尾部節點、及指向所述第一隊列陣列中的隊列的當前節點，且自所述當前節點至所述尾部節點的各節點可由消費者擁有，且所述第一隊列陣列中的其餘節點可由生產者擁有。

【0044】 所述生產者可為發起者且所述消費者可為所述目標。

【0045】 所述消費者可使用對於所述資料路徑執行緒中的每一者而言唯一的執行緒辨識符來獲得對所述隊列的排他性存取。

【0046】 所述第二隊列陣列可包括與被指配給所述輸送執行緒的所述第一中央處理單元群組對應的第二數目的節點。

【0047】 所述第二數目的節點可包括首部節點、尾部節點、及指向所述第二隊列陣列中的隊列的當前節點，且自所述當前節點至

所述尾部節點的各節點可由消費者擁有，且所述第二隊列陣列中的其餘節點可由生產者擁有。

【0048】 所述生產者可為所述目標且所述消費者可為所述發起者。

【0049】 所述消費者可使用對於所述傳送執行緒中的每一者而言唯一的執行緒辨識符來獲得對所述隊列的排他性存取。

【0050】 所述輸入/輸出命令發出隊列及所述輸入/輸出完成隊列中的每一者均可為多生產者多消費者（MPMC）無鎖隊列。來自一或多個傳送執行緒的工作請求能夠到達特定資料路徑執行緒，且來自特定傳送執行緒的工作請求能夠到達一或多個資料路徑執行緒。相似地，來自一或多個資料路徑執行緒的工作請求能夠到達特定傳送執行緒且來自特定資料路徑執行緒的工作請求能夠到達一或多個傳送執行緒。

【0051】 根據另一實施例，一種資料儲存系統包括：多個中央處理單元；多個目標；以及記憶體，儲存用於處理輸入/輸出流以及輸入/輸出命令發出隊列、及輸入/輸出完成隊列的一組軟體模組。所述一組軟體模組用以：接收對儲存於資料儲存系統的目標中的資料進行存取的輸入/輸出命令，所述資料儲存系統包括多個中央處理單元；在所述多個中央處理單元中的第一中央處理單元群組的第一中央處理單元上運行多個傳送執行緒，其中所述多個傳送執行緒中的每一者包括命令接收器路徑及命令傳送器路徑；以及在所述多個中央處理單元中的第二中央處理單元群組的第二中央

處理單元上運行多個資料路徑執行緒，其中所述多個資料路徑執行緒中的每一者包括命令發出路徑及命令完成路徑。輸送執行緒的所述命令接收器路徑將輸入/輸出命令投遞至所述輸入/輸出命令發出隊列，且資料路徑執行緒的所述命令發出路徑處理所述輸入/輸出命令。所述資料路徑執行緒的所述命令完成路徑將輸入/輸出完成通知投遞至所述輸入/輸出完成隊列，且所述輸送執行緒的所述命令傳送器路徑處理所述輸入/輸出完成通知。所述輸入/輸出命令發出隊列包括第一隊列陣列，且所述第一隊列陣列中的每一者對應於所述第一中央處理單元群組中的相應中央處理單元。所述輸入/輸出完成隊列包括第二隊列陣列，且所述第二隊列陣列中的每一者對應於所述第二中央處理單元群組中的相應中央處理單元。

**【0052】** 所述輸送執行緒可包括接收器及傳送器，且其中所述接收器可向所述資料路徑發送輸入/輸出命令。

**【0053】** 所述資料路徑執行緒可將所述輸入/輸出命令發送至所述目標並自所述目標接收狀態及/或資料，並且將所述狀態及/或資料傳送至所述輸送執行緒的所述傳送器。

**【0054】** 所述第一隊列陣列可包括與被指配給所述資料路徑執行緒的所述第二中央處理單元群組對應的第一數目的節點。

**【0055】** 所述第一數目的節點可包括首部節點、尾部節點、及指向所述第一隊列陣列中的隊列的當前節點，且自所述當前節點至所述尾部節點的各節點可由消費者擁有，且所述第一隊列陣列中

的其餘節點可由生產者擁有。

【0056】 所述生產者可為發起者且所述消費者可為所述目標。

【0057】 所述消費者可使用對於所述資料路徑執行緒中的每一者而言唯一的執行緒辨識符來獲得對所述隊列的排他性存取。

【0058】 所述第二隊列陣列可包括與被指配給所述輸送執行緒的所述第一中央處理單元群組對應的第二數目的節點。

【0059】 所述第二數目的節點可包括首部節點、尾部節點、及指向所述第二隊列陣列中的隊列的當前節點，且自所述當前節點至所述尾部節點的各節點可由消費者擁有，且所述第二隊列陣列中的其餘節點可由生產者擁有。

【0060】 所述生產者可為所述目標且所述消費者可為所述發起者。

【0061】 所述消費者可使用對於所述傳送執行緒中的每一者而言唯一的執行緒辨識符來獲得對所述隊列的排他性存取。

【0062】 所述輸入/輸出命令發出隊列及所述輸入/輸出完成隊列中的每一者均可為多生產者多消費者（MPMC）無鎖隊列。來自一或多個輸送執行緒的工作請求能夠到達特定資料路徑執行緒，且來自特定輸送執行緒的工作請求能夠到達一或多個資料路徑執行緒。相似地，來自一或多個資料路徑執行緒的工作請求能夠到達特定輸送執行緒且來自特定資料路徑執行緒的工作請求能夠到達一或多個輸送執行緒。

【0063】 上文已闡述以上示例性實施例來說明實作一種用於提供

高效能無鎖可分級目標的系統與方法的各種實施例。此項技術中具有通常知識者將意識到，存在對所揭露示例性實施例的各種潤飾及相對於所揭露示例性實施例的各種不同之處。在以下申請專利範圍中闡述旨在落於本發明範圍內的主題。

**【符號說明】**

**【0064】**

101：網際網路小電腦系統介面接收器路徑/資料路徑

102：小電腦系統介面命令發出路徑/資料路徑

103：小電腦系統介面命令完成路徑/資料路徑

104：網際網路小電腦系統介面傳送器路徑/資料路徑

200：資料路徑

201：接收器

202：網際網路小電腦系統介面命令

203：資料/相關聯資料

204：小電腦系統介面輸入/輸出模組

205：任務管理輸入/輸出模組

211：傳送器

221：資料命令描述符區塊

222：控制命令描述符區塊

223：任務管理輸入/輸出區塊

224：狀態/資料區塊

251：連接專用路徑

252：邏輯單元號碼專用路徑

301、401：網際網路小電腦系統介面接收器路徑

302、402：小電腦系統介面命令發出路徑

303、403：小電腦系統介面命令完成路徑

304、404：網際網路小電腦系統介面傳送器路徑

410：輸入/輸出命令發出隊列

411：輸入/輸出完成隊列

501：首部節點

502：當前節點

503：尾部節點

511：資料容器

512：指針

601：控制陣列

701：網際網路小電腦系統介面連接執行緒

751：小電腦系統介面邏輯單元號碼執行緒

801：邏輯單元號碼陣列

901：針對每一連接的陣列

C1、C2、C3、C4、C5、C6：網際網路小電腦系統介面連接

CPU1、CPU2、CPU3、CPU4、CPU5：中央處理單元

CT1、CT2、CT3：網際網路小電腦系統介面連接執行緒

Head 1、Head 2、Head M-1、Head M、Head N-1、Head N：

首部節點

Head 1a、Head 1b、Head 2a、Head 2b、Head 3b：節點

L1、L2、L3、L4：邏輯單元號碼

LT1、LT2：小電腦系統介面邏輯單元號碼執行緒

## 【發明申請專利範圍】

【第1項】 一種提供高效能無鎖可分級目標的方法，包括：

接收對儲存於資料儲存系統的目標中的資料進行存取的命令，其中所述資料儲存系統包括多個中央處理單元；

在所述多個中央處理單元中的第一中央處理單元群組上運行多個輸送執行緒，其中所述多個輸送執行緒中的每一者包括命令接收器路徑及命令傳送器路徑；

在所述多個中央處理單元中的第二中央處理單元群組上運行多個資料路徑執行緒，其中所述多個資料路徑執行緒中的每一者包括命令發出路徑及命令完成路徑；

使用所述輸送執行緒的所述命令接收器路徑將輸入/輸出命令投遞至輸入/輸出命令發出隊列，並使用所述資料路徑執行緒的所述命令發出路徑處理所述輸入/輸出命令；以及

使用所述資料路徑執行緒的所述命令完成路徑將輸入/輸出完成通知投遞至輸入/輸出完成隊列，並使用所述輸送執行緒的所述命令傳送器路徑處理所述輸入/輸出完成通知，

其中所述輸入/輸出命令發出隊列包括第一隊列陣列，且所述第一隊列陣列中的每一者對應於所述第一中央處理單元群組中的相應中央處理單元，且

其中所述輸入/輸出完成隊列包括第二隊列陣列，且所述第二隊列陣列中的每一者對應於所述第二中央處理單元群組中的相應中央處理單元。

【第2項】 如申請專利範圍第1項所述的方法，其中所述輸送執行緒包括接收器及傳送器，且其中所述接收器向所述資料路徑執行緒發送所述輸入/輸出命令。

【第3項】 如申請專利範圍第2項所述的方法，其中所述資料路徑執行緒將所述輸入/輸出命令發送至所述目標並自所述目標接收狀態及/或資料，並且將所述狀態及/或資料傳送至所述輸送執行緒的所述傳送器。

【第4項】 如申請專利範圍第1項所述的方法，其中所述第一隊列陣列包括與被指配給所述資料路徑執行緒的所述第二中央處理單元群組對應的第一數目的節點。

【第5項】 如申請專利範圍第4項所述的方法，其中所述第一數目的節點包括首部節點、尾部節點、及指向所述第一隊列陣列中的隊列的當前節點，且其中自所述當前節點至所述尾部節點的節點是由消費者擁有，且所述第一隊列陣列中的其餘節點是由生產者擁有。

【第6項】 如申請專利範圍第5項所述的方法，其中所述生產者是發起者且所述消費者是所述目標。

【第7項】 如申請專利範圍第6項所述的方法，其中所述消費者使用對於所述資料路徑執行緒中的每一者而言唯一的執行緒辨識符來獲得對所述隊列的排他性存取。

【第8項】 如申請專利範圍第1項所述的方法，其中所述第二隊列陣列包括與被指配給所述輸送執行緒的所述第一中央處理單元群

組對應的第二數目的節點。

【第9項】 如申請專利範圍第8項所述的方法，其中所述第二數目的節點包括首部節點、尾部節點、及指向所述第二隊列陣列中的隊列的當前節點，且其中自所述當前節點至所述尾部節點的節點是由消費者擁有，且所述第二隊列陣列中的其餘節點是由生產者擁有。

【第10項】 如申請專利範圍第9項所述的方法，其中所述生產者是所述目標且所述消費者是發起者。

【第11項】 如申請專利範圍第10項所述的方法，其中所述消費者使用對於所述輸送執行緒中的每一者而言唯一的執行緒辨識符來獲得對所述隊列的排他性存取。

【第12項】 如申請專利範圍第1項所述的方法，其中所述輸入/輸出命令發出隊列及所述輸入/輸出完成隊列中的每一者均是多生產者多消費者（MPMC）無鎖隊列，其中來自一或多個所述輸送執行緒的工作請求能夠到達特定資料路徑執行緒，且來自特定輸送執行緒的工作請求能夠到達一或多個所述資料路徑執行緒，且其中來自所述一或多個資料路徑執行緒的工作請求能夠到達所述特定輸送執行緒且來自所述特定資料路徑執行緒的工作請求能夠到達所述一或多個輸送執行緒。

【第13項】 一種資料儲存系統，包括：

多個中央處理單元；

多個目標；以及

記憶體，儲存用於處理輸入/輸出流以及輸入/輸出命令發出  
隊列、及輸入/輸出完成隊列的一組軟體模組，

其中所述一組軟體模組用以：

接收對儲存於資料儲存系統的目標中的資料進行存取的  
輸入/輸出命令；

在所述多個中央處理單元中的第一中央處理單元群組的  
第一中央處理單元上運行多個輸送執行緒，其中所述多個輸送  
執行緒中的每一者包括命令接收器路徑及命令傳送器路徑；以  
及

在所述多個中央處理單元中的第二中央處理單元群組的  
第二中央處理單元上運行多個資料路徑執行緒，其中所述多個  
資料路徑執行緒中的每一者包括命令發出路徑及命令完成路  
徑；

其中所述輸送執行緒的所述命令接收器路徑將所述輸入/輸  
出命令投遞至所述輸入/輸出命令發出隊列，且所述資料路徑執行  
緒的所述命令發出路徑處理所述輸入/輸出命令，

其中所述資料路徑執行緒的所述命令完成路徑將輸入/輸出  
完成通知投遞至所述輸入/輸出完成隊列，且所述輸送執行緒的所  
述命令傳送器路徑處理所述輸入/輸出完成通知，

其中所述輸入/輸出命令發出隊列包括第一隊列陣列，且所述  
第一隊列陣列中的每一者對應於所述第一中央處理單元群組中的  
相應中央處理單元，且

其中所述輸入/輸出完成隊列包括第二隊列陣列，且所述第二隊列陣列中的每一者對應於所述第二中央處理單元群組中的相應中央處理單元。

【第14項】 如申請專利範圍第13項所述的資料儲存系統，其中所述輸送執行緒包括接收器及傳送器，且其中所述接收器向所述資料路徑執行緒發送所述輸入/輸出命令。

【第15項】 如申請專利範圍第14項所述的資料儲存系統，其中所述資料路徑執行緒將所述輸入/輸出命令發送至所述目標並自所述目標接收狀態及/或資料，並且將所述狀態及/或資料傳送至所述輸送執行緒的所述傳送器。

【第16項】 如申請專利範圍第13項所述的資料儲存系統，其中所述第一隊列陣列包括與被指配給所述資料路徑執行緒的所述第二中央處理單元群組對應的第一數目的節點。

【第17項】 如申請專利範圍第16項所述的資料儲存系統，其中所述第一數目的節點包括首部節點、尾部節點、及指向所述第一隊列陣列中的隊列的當前節點，且其中自所述當前節點至所述尾部節點的節點是由消費者擁有，且所述第一隊列陣列中的其餘節點是由生產者擁有。

【第18項】 如申請專利範圍第17項所述的資料儲存系統，其中所述生產者是發起者且所述消費者是所述目標。

【第19項】 如申請專利範圍第18項所述的資料儲存系統，其中所述消費者使用對於所述資料路徑執行緒中的每一者而言唯一的執

行緒辨識符來獲得對所述隊列的排他性存取。

【第20項】 如申請專利範圍第13項所述的資料儲存系統，其中所述第二隊列陣列包括與被指配給所述輸送執行緒的所述第一中央處理單元群組對應的第二數目的節點。

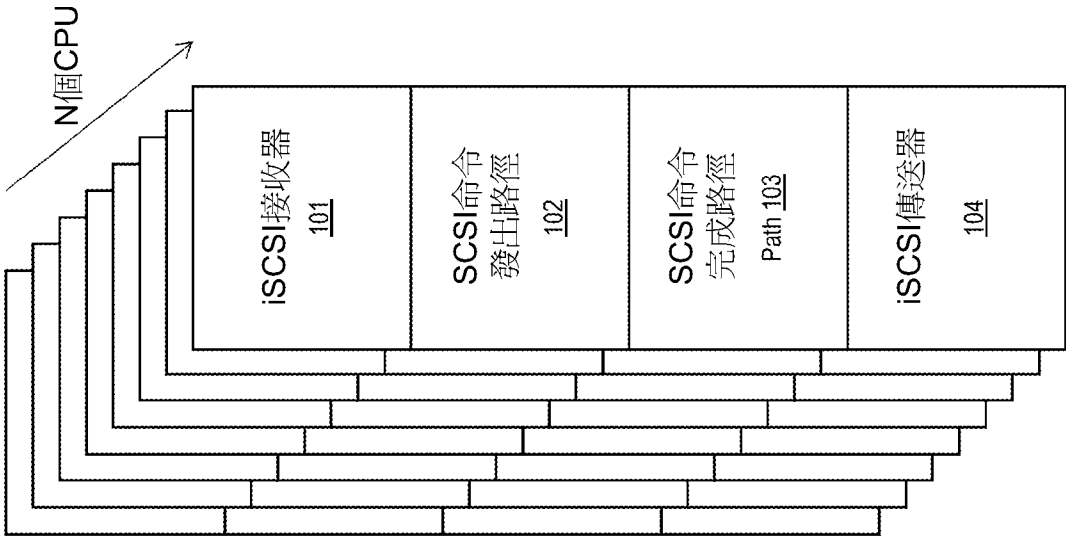
【第21項】 如申請專利範圍第20項所述的資料儲存系統，其中所述第二數目的節點包括首部節點、尾部節點、及指向所述第二隊列陣列中的隊列的當前節點，且其中自所述當前節點至所述尾部節點的節點是由消費者擁有，且所述第二隊列陣列中的其餘節點是由生產者擁有。

【第22項】 如申請專利範圍第21項所述的資料儲存系統，其中所述生產者是所述目標且所述消費者是發起者。

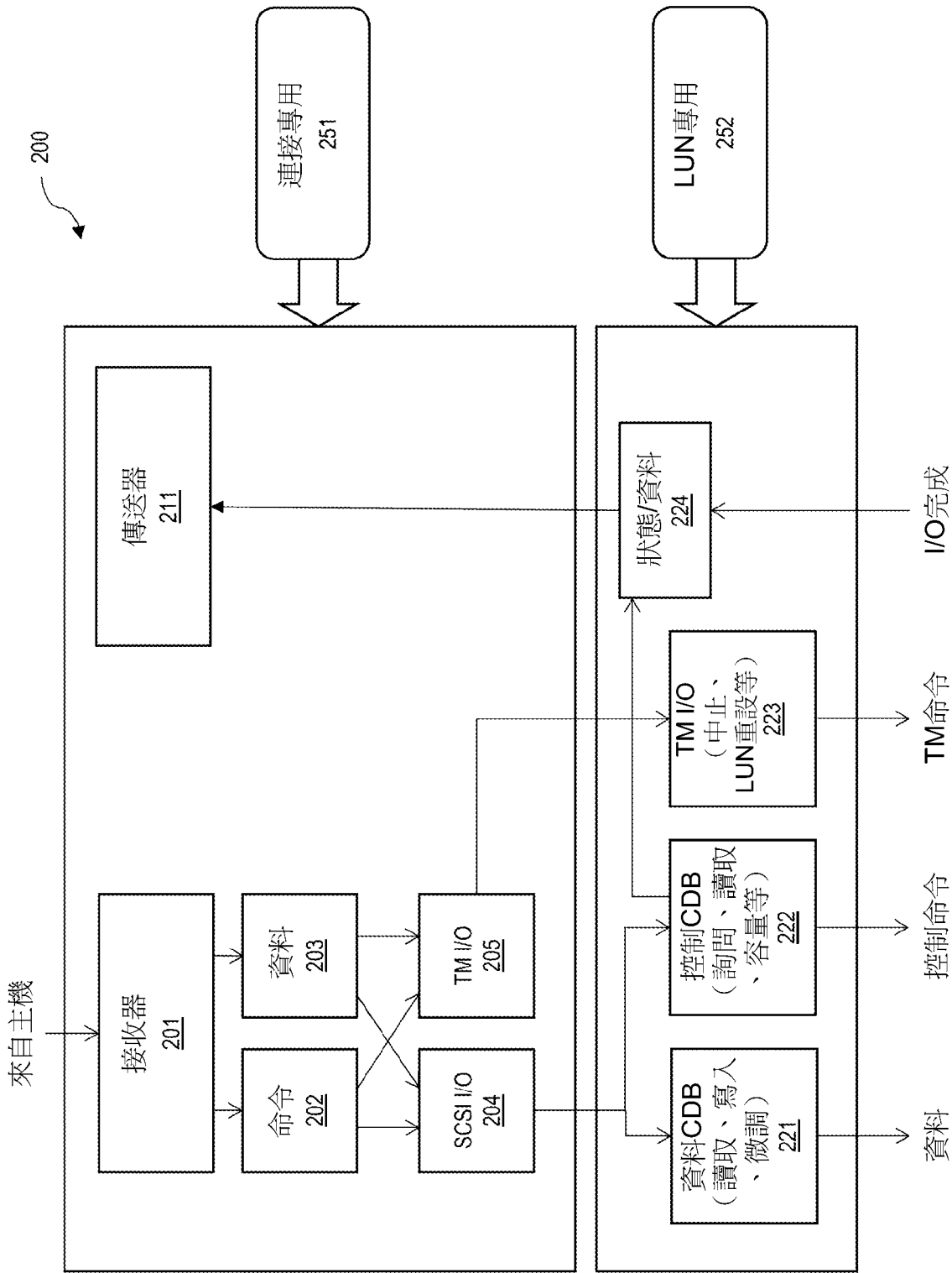
【第23項】 如申請專利範圍第22項所述的資料儲存系統，其中所述消費者使用對於所述輸送執行緒中的每一者而言唯一的執行緒辨識符來獲得對所述隊列的排他性存取。

【第24項】 如申請專利範圍第13項所述的資料儲存系統，其中所述輸入/輸出命令發出隊列及所述輸入/輸出完成隊列中的每一者均是多生產者多消費者（MPMC）無鎖隊列，其中來自一或多個所述輸送執行緒的工作請求能夠到達特定資料路徑執行緒，且來自特定輸送執行緒的工作請求能夠到達一或多個所述資料路徑執行緒，且其中來自所述一或多個資料路徑執行緒的工作請求能夠到達所述特定輸送執行緒且來自所述特定資料路徑執行緒的工作請求能夠到達一或多個輸送執行緒。

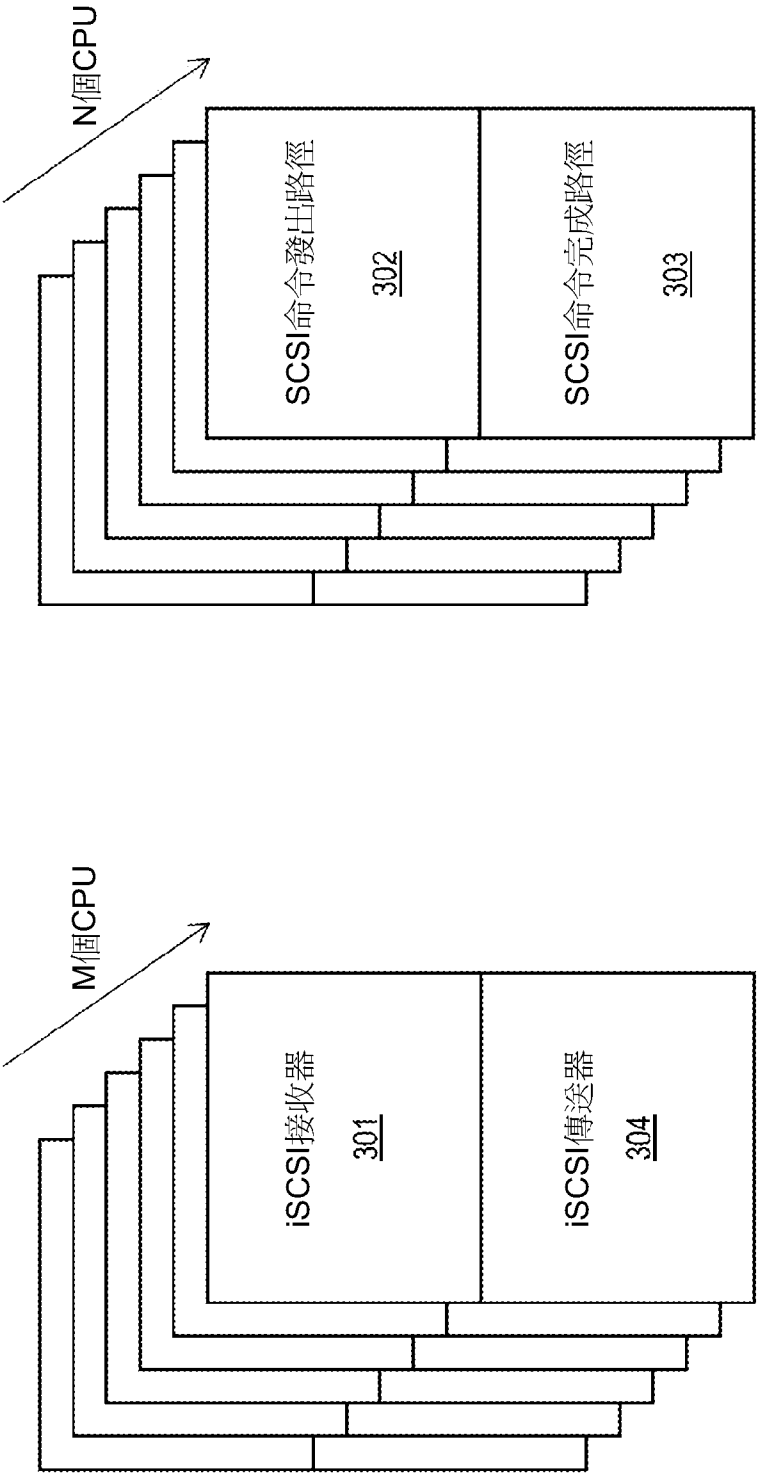
【發明圖式】



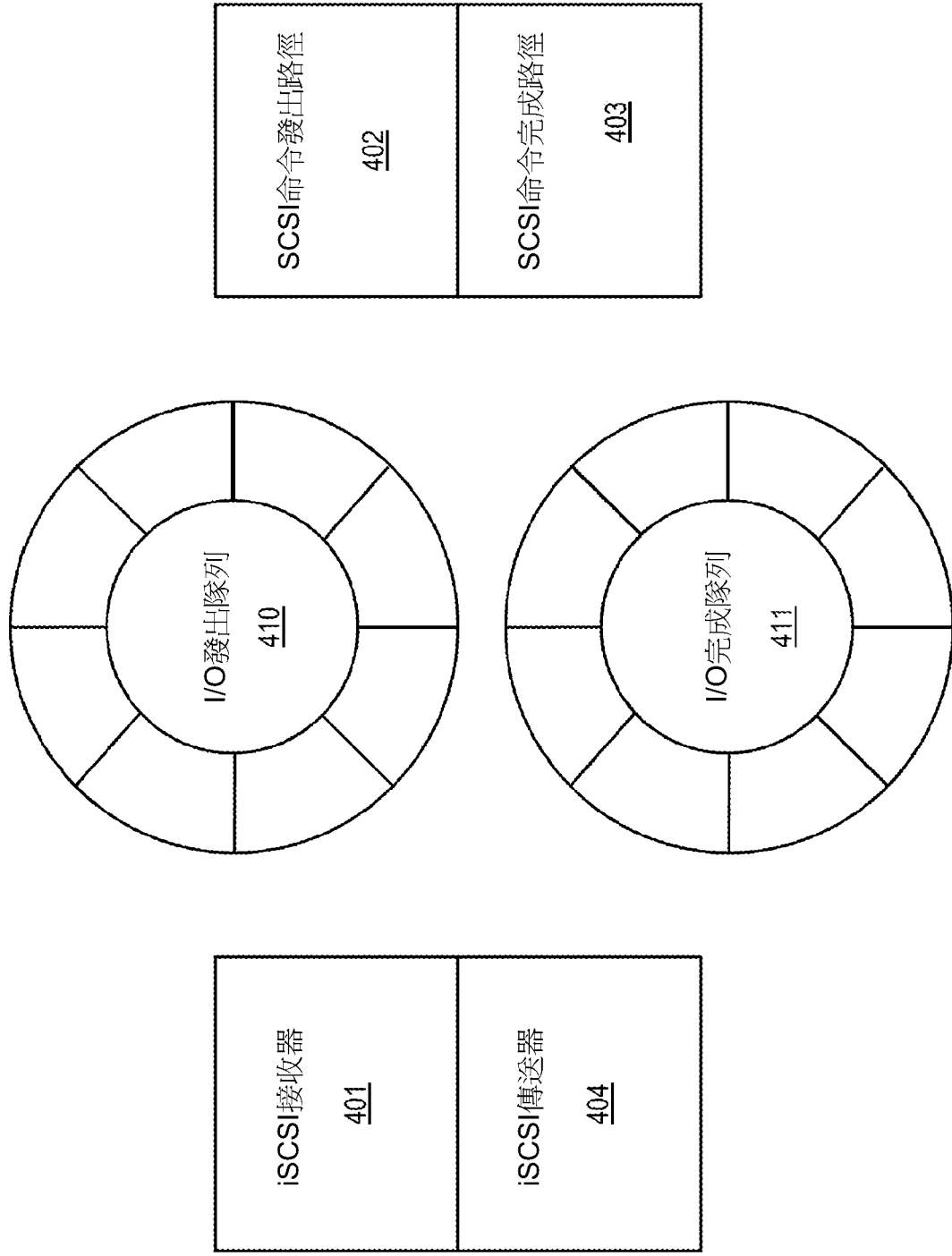
【圖1】



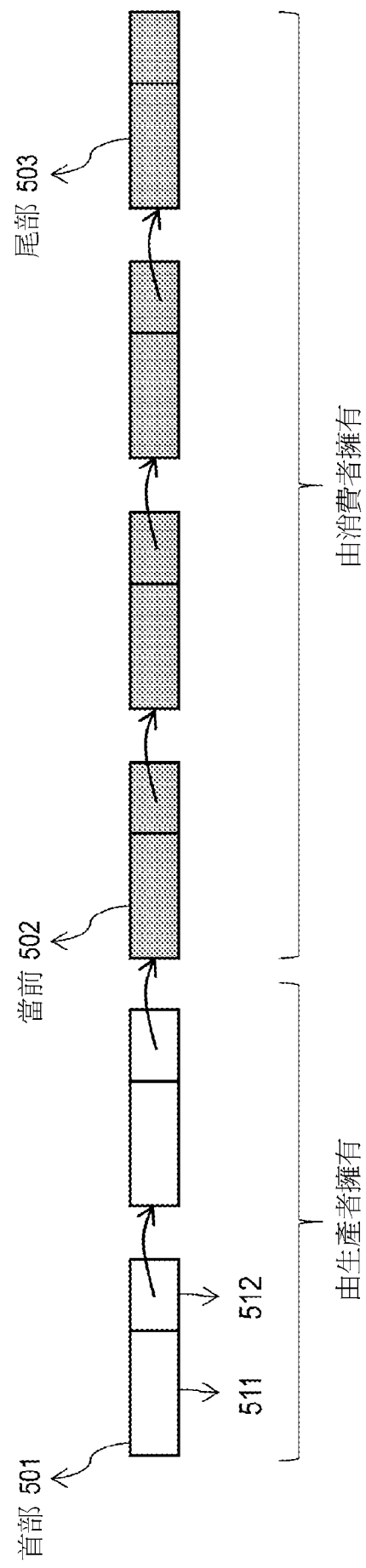
【圖2】



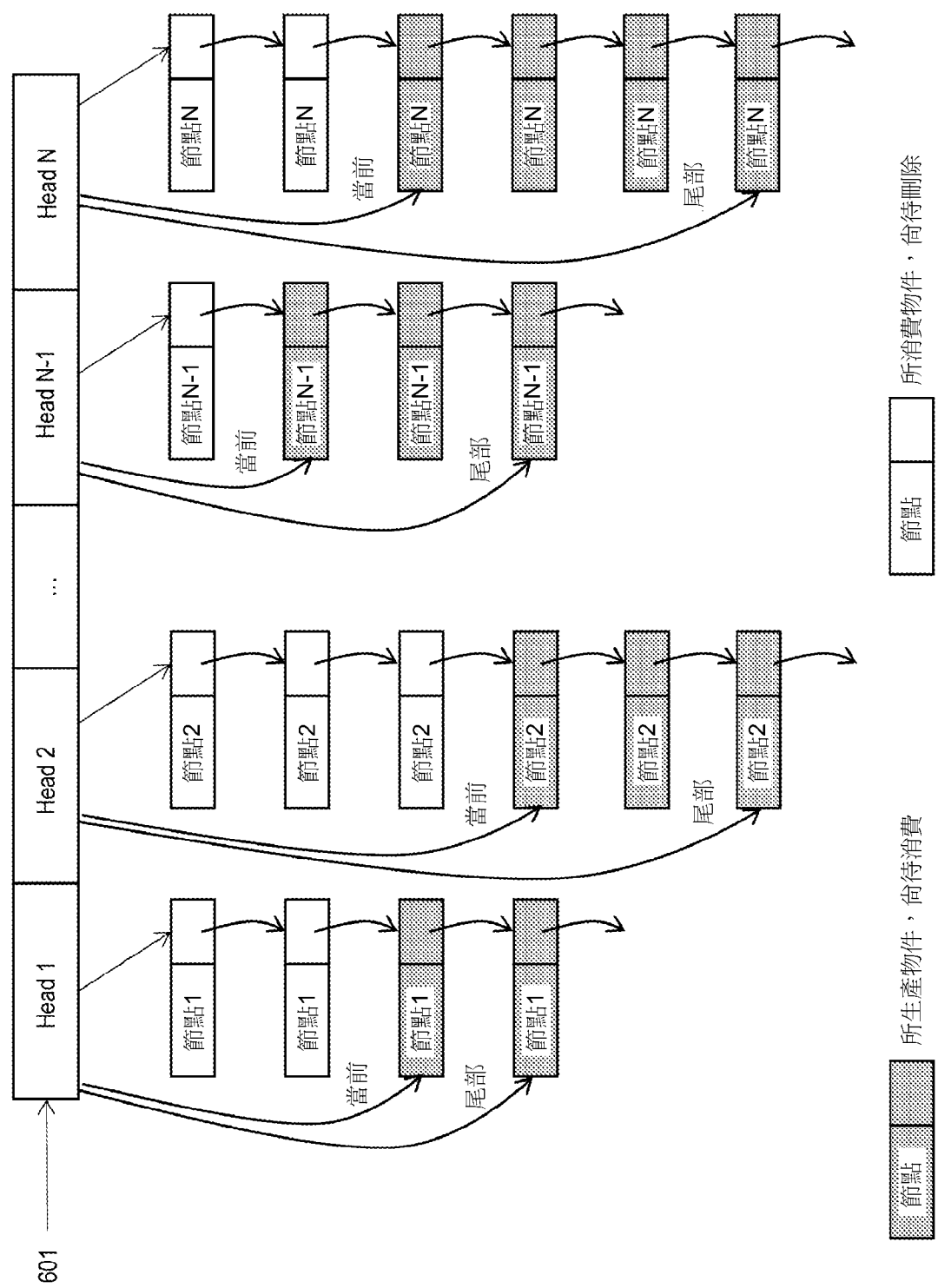
【圖3】



【圖4】



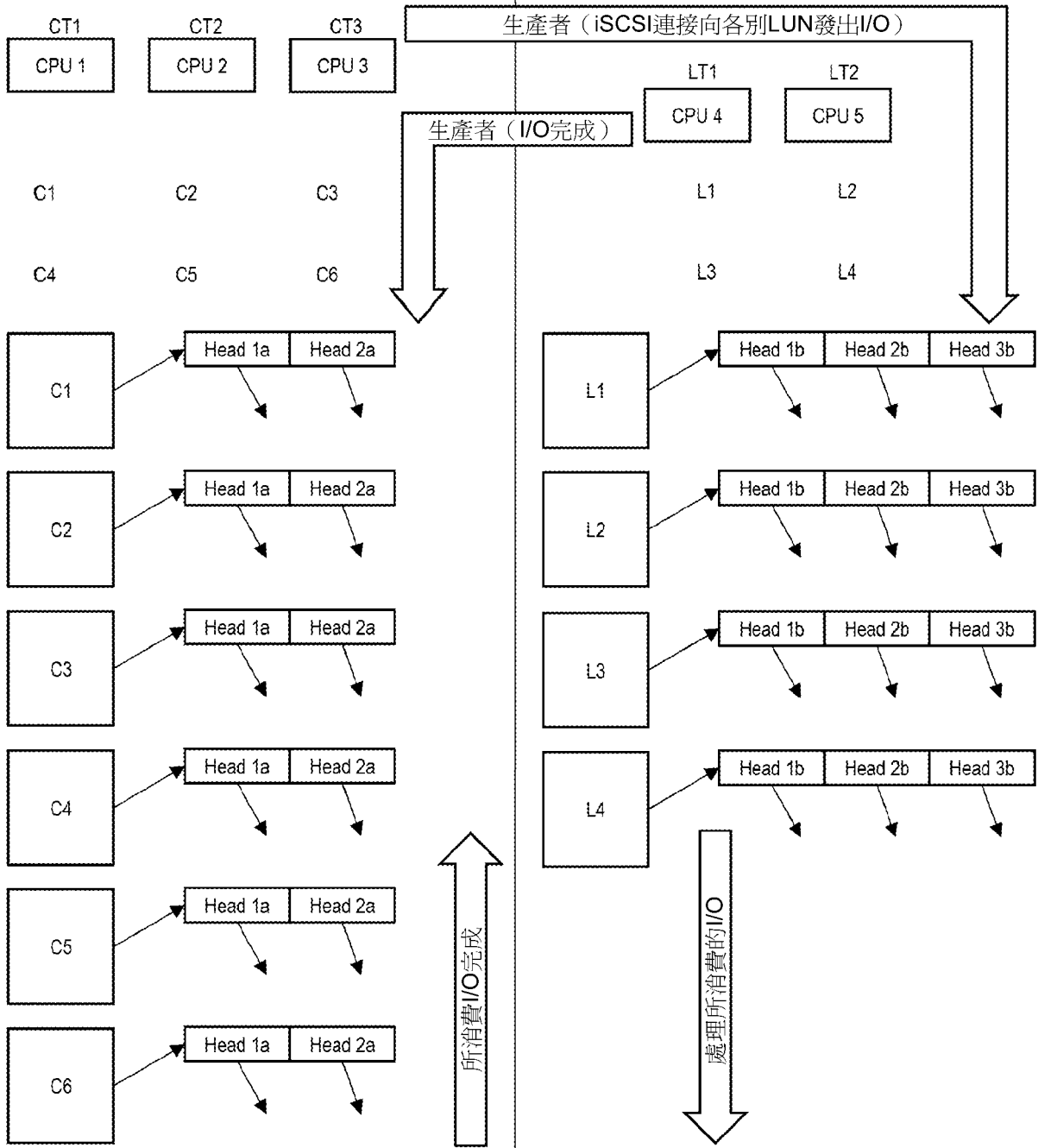
【圖5】



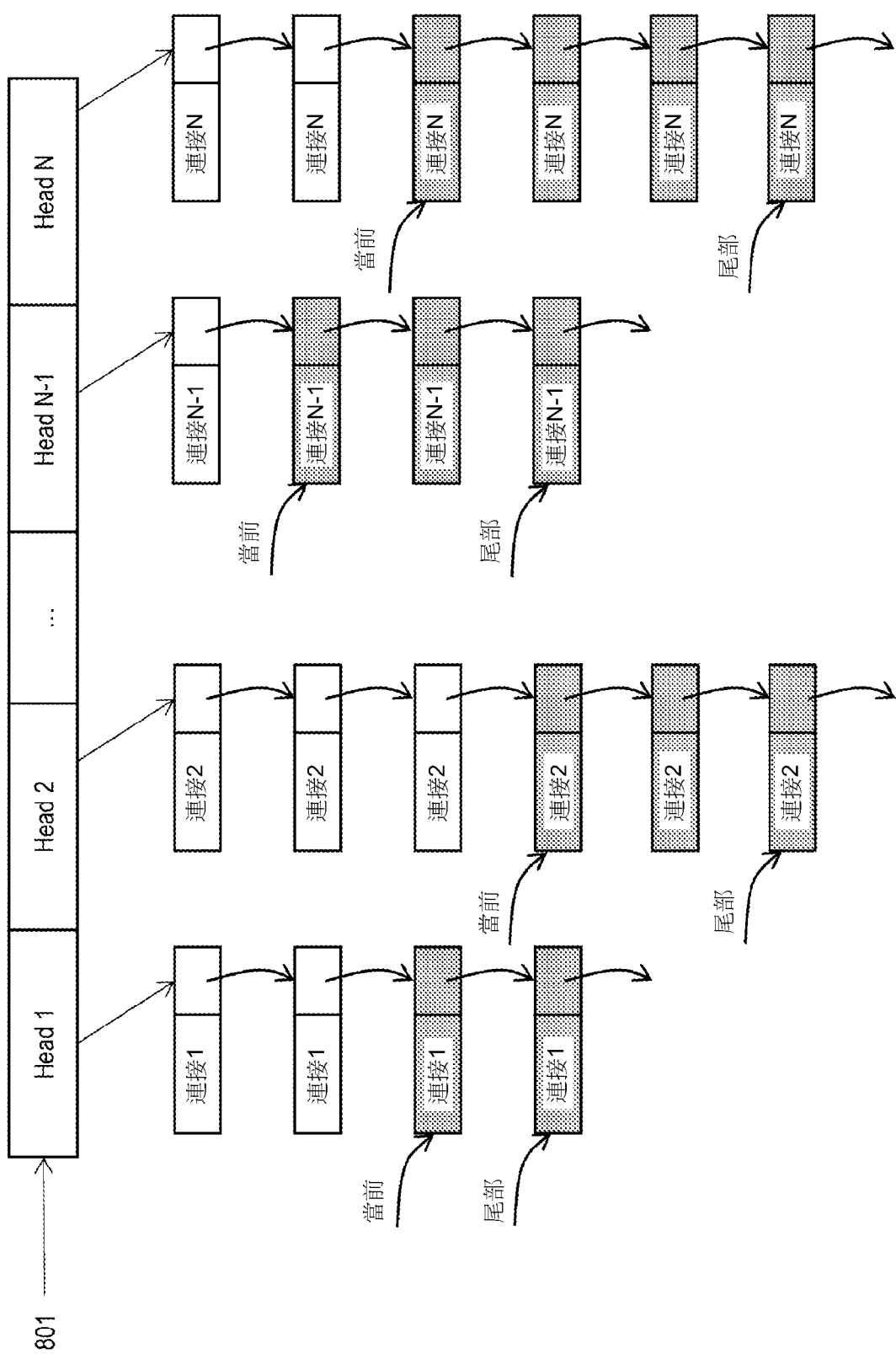
【圖6】

iSCSI連接執行緒 701

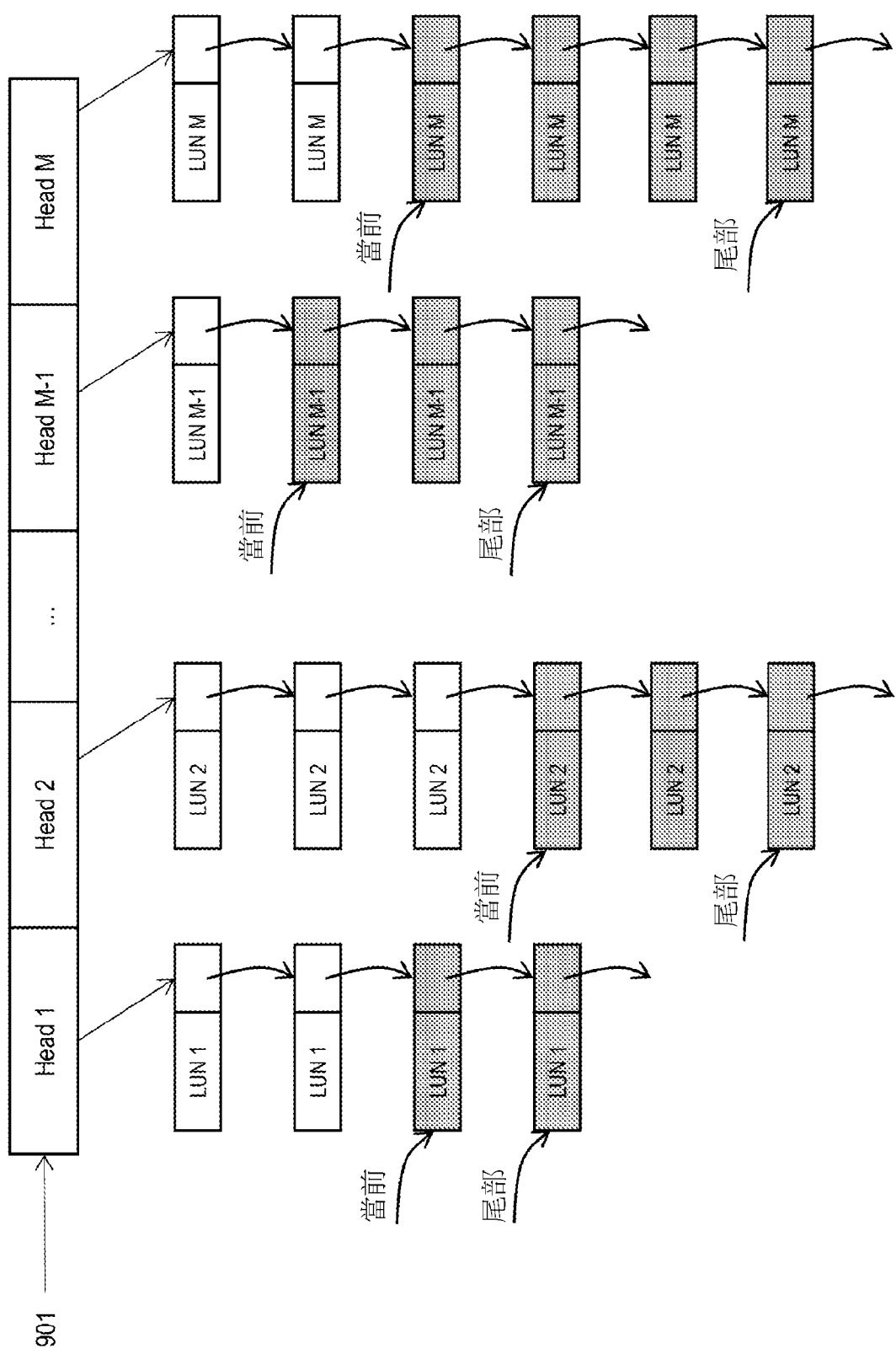
iSCSI連接執行緒 751



【圖7】



【圖8】



【圖9】