



(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 I746846 B

(45)公告日：中華民國 110 (2021) 年 11 月 21 日

(21)申請案號：107115599

(22)申請日：中華民國 107 (2018) 年 05 月 08 日

(51)Int. Cl. : G06F9/50 (2006.01)

G06F11/30 (2006.01)

G06F12/00 (2006.01)

(30)優先權：2017/07/19 美國

62/534,647

2018/02/02 美國

15/887,980

(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：楊征宇 YANG, ZHENGYU (CN)；艾文斯 湯瑪斯 大衛 EVANS, THOMAS DAVID
(US)；安德魯斯 艾倫 ANDREWS, ALLEN (US)；邁爾斯 克雷 MAYERS, CLAY
(US)；伯爾特 湯瑪斯羅里 BOLT, THOMAS RORY (US)

(74)代理人：林孟閱；盧佩君；陳怡如

(56)參考文獻：

US 9703664B1

US 9710178B2

US 2016/0306554A1

US 2016/0378614A1

US 2017/0147237A1

審查人員：詹効儒

申請專利範圍項數：項 圖式數： 共頁

(54)名稱

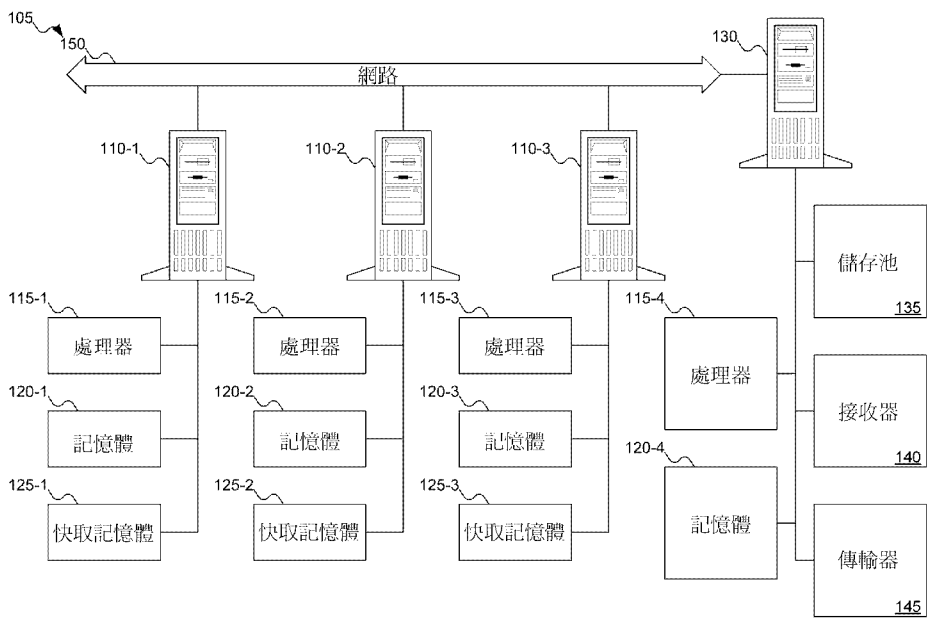
資料儲存管理系統、方法及物件

(57)摘要

本發明提供一種資料儲存管理系統、資料儲存管理方法及資料儲存管理物件。本發明公開一種用於執行自動分層的系統。系統可以包含提供多個資源並組織成儲存層的多個儲存裝置。儲存裝置能夠儲存虛擬機器的資料。接收器能夠接收虛擬機器的輸入/輸出命令和性能資料。傳輸器能夠傳輸對於輸入/輸出命令的回應。自動分層控制器能夠選擇儲存層以儲存虛擬機器的資料並且能夠回應於性能資料而在儲存層之間遷移資料。將虛擬機器的性能改變和在儲存層之間遷移資料的遷移成本作為考慮因素，儲存層的選擇能夠優化所有虛擬機器在所有儲存層中的性能。

The invention provides a data storage managing system, a data storage managing method and a data storage managing article. The system may include a plurality of storage devices offering a plurality of resources and organized into storage tiers. The storage devices may store data for virtual machines. A receiver may receive I/O commands and performance data for the virtual machines. A transmitter may transmit responses to the I/O commands. An auto-tiering controller may select storage tiers to store the data for the virtual machines and may migrate data between storage tiers responsive to the performance data. The selection of the storage tiers may optimize the performance of all virtual machines across all storage tiers, factoring the change in performance of the virtual machines and a migration cost to migrate data between storage tiers.

指定代表圖：



【圖1】

符號簡單說明：

- 105:系統
- 110-1、110-2、110-3:
虛擬機器伺服器
- 115-1、115-2、
115-3、115-4:處理器
- 120-1、120-2、
120-3、120-4:記憶體
- 125-1、125-2、125-3:
快取記憶體
- 130:資料中心伺服器
- 135:儲存池
- 140:接收器
- 145:傳輸器
- 150:網路



公告本

107-06-21

I746846

【發明摘要】

【中文發明名稱】資料儲存管理系統、方法及物件

【英文發明名稱】DATA STORAGE MANAGING SYSTEM,

METHOD AND ARTICLE

【中文】本發明提供一種資料儲存管理系統、資料儲存管理方法及資料儲存管理物件。本發明公開一種用於執行自動分層的系統。系統可以包含提供多個資源並組織成儲存層的多個儲存裝置。儲存裝置能夠儲存虛擬機器的資料。接收器能夠接收虛擬機器的輸入/輸出命令和性能資料。傳輸器能夠傳輸對於輸入/輸出命令的回應。自動分層控制器能夠選擇儲存層以儲存虛擬機器的資料並且能夠回應於性能資料而在儲存層之間遷移資料。將虛擬機器的性能改變和在儲存層之間遷移資料的遷移成本作為考慮因素，儲存層的選擇能夠優化所有虛擬機器在所有儲存層中的性能。

【英文】The invention provides a data storage managing system, a data storage managing method and a data storage managing article. The system may include a plurality of storage devices offering a plurality of resources and organized into storage tiers. The storage devices may store data for virtual machines. A receiver may receive I/O commands and performance data for the virtual machines. A transmitter may transmit responses to the I/O

commands. An auto-tiering controller may select storage tiers to store the data for the virtual machines and may migrate data between storage tiers responsive to the performance data. The selection of the storage tiers may optimize the performance of all virtual machines across all storage tiers, factoring the change in performance of the virtual machines and a migration cost to migrate data between storage tiers.

【指定代表圖】圖1。

【代表圖之符號簡單說明】

105：系統

110-1、110-2、110-3：虛擬機器伺服器

115-1、115-2、115-3、115-4：處理器

120-1、120-2、120-3、120-4：記憶體

125-1、125-2、125-3：快取記憶體

130：資料中心伺服器

135：儲存池

140：接收器

145：傳輸器

150：網路

【特徵化學式】

無

【發明說明書】

【中文發明名稱】 資料儲存管理系統、方法及物件

【英文發明名稱】 DATA STORAGE MANAGING SYSTEM,
METHOD AND ARTICLE

【技術領域】

【0001】 本發明是有關於一種資料儲存方法，且特別是有關於一種將虛擬機器資料儲存在哪個儲存層中的自動管理。

【先前技術】

【0002】 傳統的資料儲存中心依賴於硬碟驅動器和其它早期形式的資料儲存裝置。由於硬碟驅動器和其它早期形式的資料儲存裝置具有相對較高的延遲，所以虛擬機器快取在本地儲存裝置中頻繁存取的資料。通常，快取資料的相關益處為資料存取時間增大十倍或大於十倍。

【0003】 但是隨著快閃記憶體儲存裝置變得更常用且更負擔得起，這種在資料儲存中心與快取記憶體之間的存取時間上的差異減小。這種快取的益處減少意味著資料頻繁地專門儲存在資料中心且根本不快取。當結合將資料的冗餘複本（快取記憶體和資料中心兩者中的）儲存在快閃記憶體儲存裝置中是昂貴的事實時，以最高效的方式將資料儲存在資料儲存中心中變得更重要。

【0004】 另外，不同於傳統的儲存裝置，例如硬碟驅動器，快閃

記憶體儲存裝置可以具有不同的專業化。由於傳統的儲存裝置並不考慮裝置專業化，所以使用傳統技術進行資料儲存的任何層選擇是次優的。

【0005】 仍需要一種使用快閃記憶體儲存裝置更優化地將資料儲存在資料儲存中心中的方式，所述快閃記憶體儲存裝置同時提供令人滿意的存取時間且利用任何由不同儲存層提供的專業化。

【發明內容】

【0006】 本發明概念的實施例包含一種系統，其包括：提供多個資源的多個儲存裝置，所述多個儲存裝置組織成多個儲存層並且儲存第一虛擬機器的第一資料和第二虛擬機器的第二資料；接收器，接收來自所述第一虛擬機器的第一輸入/輸出（I/O）命令、來自所述第二虛擬機器的第二 I/O 命令、模擬所述第一虛擬機器在所述多個儲存層中的性能的第一性能資料，以及模擬所述多個儲存層中的所述第二虛擬機器的性能的第二性能資料；傳輸器，將對於所述第一 I/O 命令的第一回應傳輸到所述第一虛擬機器並且將對於所述第二 I/O 命令的第二回應傳輸到所述第二虛擬機器；以及自動分層控制器，選擇第一儲存層以儲存所述第一虛擬機器的所述第一資料、選擇第二儲存層以儲存所述第二虛擬機器的所述第二資料，且回應於所述第一性能資料和所述第二性能資料移轉所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個，其

中所述自動分層控制器可操作以選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料並且選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料來優化所有虛擬機器在所述多個儲存層中的性能，其中所述自動分層控制器可操作以將由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所引起的性能改變和遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個的遷移成本作為考慮因素。

【0007】 本發明概念的實施例包含一種方法，其包括：接收模擬第一虛擬機器在多個儲存層中的性能的第一性能資料、接收模擬第二虛擬機器在所述多個儲存層中的性能的第二性能資料、使用所述第一性能資料和所述第二性能資料生成使所述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯的資源矩陣、使用所述資源矩陣選擇所述多個儲存層中的第一儲存層以儲存所述第一虛擬機器的第一資料、使用所述資源矩陣選擇所述多個儲存層中的第二儲存層以儲存所述第二虛擬機器的第二資料，以及遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個；其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料使所有虛擬機

器在所述多個儲存層中的性能優化，其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料將由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所引起的性能改變作為考慮因素，以及其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料將遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個的遷移成本作為考慮因素。

【圖式簡單說明】

【0008】

圖 1 示出根據本發明概念的實施例的可以託管虛擬機器並且使用經裝備以執行自動分層的儲存池的連網機器。

圖 2 示出圖 1 的機器的額外細節。

圖 3 示出根據本發明概念的實施例的在自動分層的情況下與圖 1 的儲存池進行交互的圖 1 的一個伺服器的額外細節。

圖 4 示出由不同儲存裝置提供的不同可用資源。

圖 5 示出將從一個儲存裝置遷移到另一個儲存裝置的圖 3 的虛擬機器。

圖 6A 到圖 6B 示出在不同時間點處虛擬機器到儲存層的最優

分配。

圖 7 示出增加延遲並且確定圖 3 的虛擬機器的性能資料的圖 3 的 I/O 濾波器。

圖 8 示出圖 3 的自動分層控制器的細節。

圖 9 示出圖 7 的性能資料轉換成使虛擬機器、儲存層和所述儲存層的資源相關聯的矩陣。

圖 10 示出圖 7 的性能資料的線性回歸。

圖 11 示出關於圖 3 的儲存層的專業化資訊。

圖 12 示出待用於將圖 3 的虛擬機器遷移到圖 3 的不同儲存層中的得分矩陣的計算。

圖 13 示出監測期與遷移期之間的關係。

圖 14 示出根據本發明概念的實施例的用於圖 3 的 I/O 濾波器將延遲增加到輸入/輸出 (I/O) 命令以產生圖 7 的性能資料的實例過程的流程圖。

圖 15A 到圖 15B 示出根據本發明概念的實施例的用於圖 3 的自動分層控制器管理虛擬機器在儲存層中的遷移的實例過程的流程圖。

圖 16 示出根據本發明概念的實施例的用於圖 3 的自動分層控制器生成圖 9 的資源矩陣的實例過程的流程圖。

【實施方式】

【0009】 現在將詳細參考本發明概念的實施例，在附圖中圖示所

述實施例的實例。在以下詳細描述中，闡述眾多具體細節以實現對本發明概念的透徹理解。然而，應理解，本領域技術人員可以在沒有這些具體細節的情況下實踐本發明概念。在其它情況下，未詳細描述眾所周知的方法、過程、元件、電路和網路，以便不會不必要地混淆實施例的各方面。

【0010】 應理解，雖然術語第一、第二等可在本文中用以描述各種元件，但這些元件不應受這些術語限制。這些術語僅用於將一個元件與另一元件區分開來。舉例來說，在不脫離本發明概念的範圍內，第一模組可以被稱為第二模組，並且類似地，第二模組可以被稱為第一模組。

【0011】 本文中用於本發明概念的描述中的術語是僅出於描述具體實施例的目的且並不希望限制本發明概念。如本發明概念和所附申請專利範圍的描述中所使用，單數形式“一(a/an)”和“所述”意在也包含複數形式，除非上下文另外明確指示。還將理解，如本文中所使用的術語“和/或”指代且涵蓋相關聯的所列項目中的一或多者的任何及所有可能的組合。將進一步理解，當用於本說明書中時，術語“包括(comprises)”和/或“包括(comprising)”指定存在所陳述的特徵、整數、步驟、操作、元件和/或元件，但不排除存在或添加一個或大於一個其它特徵、整數、步驟、操作、元件、元件和/或其群組。圖式中的元件和特徵未必是按比例繪製。

【0012】 全快閃記憶體儲存系統由於固態驅動器(SSD)價格持

續降低且 SSD 容量持續增大的事實將用於近期的企業資料中心。因此，硬碟驅動器（HDD）的優勢減弱，將被具有較大電容量的低端 SSD，例如多層單元（MLC）SSD 替換。因此，有成本效益的設計將引入多層 SSD，其中每一層可能具有獨特的特殊性，例如高速、大容量、擅長讀取輸入/輸出（I/O）命令、長壽命等。

【0013】 在另一方面，典型的企業資料中心使用範例使用管理程式-虛擬機器（VM）結構。VM 的 I/O 工作負載模式可以不同，且其升級到更高層服務所發現的性能改進並不相同。由此推動，稱作“自動分層（自動分層）”的整體解決方案可以操控全快閃多層資料中心中的 VM 配置和遷移以最佳的利用儲存資源、優化性能，以及減小遷移開銷。自動分層也可以考慮歷史性能和預測性能兩種因素。

【0014】 VM 運行客戶機作業系統（OS），例如 Linux 和 Windows，以及應用程式。VM 彼此隔離。雲服務供應商向使用者“出租”其 VM。VM 基於使用者應用可以具有不同的工作負載特徵，且因此其對儲存裝置速度可以具有不同程度的靈敏度。

【0015】 VM 管理程式軟體，例如 Xen、KVM 和 VMware，可以託管單個伺服器內的多個虛擬機器（VM）。VM 管理程式軟體負責調度、資源管理、系統軟體應用程式設計介面（API）以及硬體虛擬化。本發明概念的實施例可以包含裝設在 VM 管理程式層級處的自動分層幕後程式，所述幕後程式的工作是與自動分層 I/O 濾波器和自動分層控制器連通。

【0016】 所有 VM 伺服器可以連接到集中式儲存池以共用後端全快閃多層 SSD 磁片，所述磁片可以包含例如非揮發性記憶體 Express (NVMe) SSD、3D XPoint 非揮發性記憶體(NVM)SSD，和 MLC/三階儲存單元 (TLC)/四階儲存單元 (QLC) SSD。每一個層可以具有不同的特殊性，例如快速、較大儲存容量等。

【0017】 自動分層可以包含三個元件：自動分層 I/O 濾波器、自動分層幕後程式和自動分層控制器。

【0018】 自動分層 I/O 濾波器驅動器可以附接到將管理的每一個 VM (即，虛擬機器磁片 (VMDK) 檔)。自動分層 I/O 濾波器驅動器可以收集 I/O 相關的統計且對每一 VMDK 運行特定的延遲測試。可以一些樣本週期速率收集資料且結果可以傳輸到負責收集來自所有 VMDK 的所有資料的主機系統上的自動分層幕後程式。

【0019】 自動分層幕後程式可以在所有物理機器的 VM 管理程式上運行。自動分層幕後程式跟蹤 VM 伺服器的工作負載變化 (I/O 存取模式變化)，且從自動分層 I/O 濾波器收集增加延遲測試的結果以傳送到自動分層控制器。

【0020】 自動分層控制器可以在專用伺服器或儲存系統中的嵌入式系統上運行。若 VM 遷移到另一層以及其對應的遷移開銷，則自動分層控制器可以作出決策以基於預測的 VM 性能觸發遷移。

【0021】 從每一個 VM 的視角來看，某一 VM 從一個層遷移到另一層的原因是 VM 可以在遷移之後更好的運行 (較小的平均 I/O 延遲、每秒更高的 I/O 操作 (IOPS)、更高的輸送量等)。但同時，

存在相應的遷移成本，這是因為將 VM 從一個層遷移到另一層相對昂貴且不可忽略。

【0022】 從全域優化的視角來看，由於資源限制和其相應的服務水準協定 (SLA)，難以滿足所有 VM 要求同時遷移到其最喜歡層。全域優化應考慮 VM 隨時間的變化以及先前對遷移效應。舉例來說，當前最好地分配解決方案由於在執行時間期間改變 VM 行為可以導致在未來不良的情況。因此，全域優化應考慮潛在的益處和代償、遷移開銷、每一層上的 VM 的歷史和預測性能、SLA (服務水準協定) 等。基於這種資訊，自動分層可以試圖使用方程式 1 到方程式 6 來實現多層全快閃記憶體儲存系統的最優解決方案。

【0023】 表 1. 符號。

符號	含義
v	VM 標識， $v \in [1, v_{max}]$ ，其中 v_{max} 是最後一個 VM。
t	層標識， $t \in [1, t_{max}]$ ，其中 t_{max} 是最後一個層。
$t_{v,\tau}$	VM v 在時期 τ 期間的託管層。
k	不同的資源類型， $t \in [1, k_{max}]$ ，例如 IOPS、輸送量、儲存大小等。
τ	時期標識，其中 $\tau \in [0, -\infty)$ 。
α_k	資源 k 的權重。
β	遷移成本的權重。
$r(v, t_{v,\tau}, \tau, k)$	在層 $t_{v,\tau}$ 上運行的 VM v 的 k 資源使用的預測類型。
$m(v, t_{v,\tau-1}, t_{v,\tau})$	VM v 在時期 τ 期間從層 $t_{v,\tau-1}$ 到層 $t_{v,\tau}$ 的遷移成本。
k_{stg}	儲存大小。
$Pr(v, t, \tau), Pw(v, t, \tau)$	在時期 τ 期間 VM v 在層 t 上的讀取或寫入命令的輸送量。
$Pr(\mathcal{A}, t, \tau), Pw(\mathcal{A}, t, \tau)$	在時期 τ 期間層 t 的所有剩餘可用讀取或寫入輸送量。

$W_{v,\tau}$	VM v 的權重
Γ_k	可以使用的每種類型資源的上限（以百分比計）。
$R(t_{v,\tau}, k)$	在層 $t_{v,\tau}$ 上的第 k 種類型資源的總容量。

最小化：

$\sum_{\forall v, \forall t, \forall \tau} w_{v,\tau} \cdot [\sum_{\forall k} \alpha_k \cdot r(v, t_{v,\tau}, \tau, k) - \beta \cdot m(v, t_{v,\tau-1}, t_{v,\tau})]$	(1)
---	-----

進行：

$r(v, t_{v,\tau}, \tau, k) = r_{prd}(v, t_{v,\tau-1}, t_{v,\tau}, \tau - 1, k) \geq 0,$	$\forall v, \forall t, \forall \tau$	(2)
$r(v, t_{v,\tau}, \tau_1, k_s) \equiv r(v, t_{v,\tau}, \tau_2, k_s) \geq 0,$	$\forall \tau_1, \tau_2 \in [0, -\infty)$	(3)
$m(v, t_{v,\tau-1}, t_{v,\tau}) = \frac{r(v, t_{v,\tau-1}, \tau - 1, k_s)}{\min(Pr(A, t_{v,\tau-1}, \tau - 1) + Pr(v, t_{v,\tau-1}, \tau - 1), Pw(A, t_{v,\tau-1}, \tau - 1))}$		(4)
$t_{v,\tau} \neq \emptyset,$	$\forall v, \forall \tau$	(5)
$\sum_{\forall v} r(v, t_{v,\tau}, \tau, k) \leq \Gamma_k \cdot R(t_{v,\tau}, k),$	$\forall t, \forall k, \forall \tau$	(6)

【0024】 方程式 1 表示目標函數。主要想法是使整個性能增益減遷移成本：“性能增益-性能損失” 最大化。內部“求和”運算子進行每一個 VM 的所有資源類型（例如 IOPS、輸送量和儲存大小等）的使用率的加權和，假設將 v 從層 tτ-1 遷移到層 tτ。注意這裡的遷移並不將 VM 從一個主機伺服器遷移到另一個主機伺服器：僅後端 VMDK 檔從一個 SSD 層遷移到另一個 SSD 層。因此，不需要考慮非磁片 I/O 相關的資源，例如 CPU、記憶體等。外部“求和”運算子進一步反覆運算所有可能的遷移情況。注意權重參數 Wv,τ 可以反映每一個 VM 的 SLA。

【0025】 方程式 2 表示可以用於預測性能增益的預測模型函數。

方程式 3 確保儲存大小（即，VMDK 大小）在每一個遷移之前和之後將不會改變。方程式 4 估計時間遷移開銷。詳細地說，時間遷移開銷是待遷移的 VM 大小除以通過在對應層上讀取和寫入資料所施加的瓶頸。詳細地說， $Pr(\Lambda, t_{v,\tau-1}, \tau-1)$ 表示可用的剩餘速度，且由於 VM 在遷移期間可以暫停，所以這個 VM 所使用的輸送量也可用（即， $Pr(\Lambda, t_{v,\tau-1}, \tau-1)$ 可以加回）。 $Pw(\Lambda, t_{v,\tau-1}, \tau-1)$ 表示呈寫入速度的遷移。方程式 5 反映需要保證 VM 被一個且僅一個磁片層代管。最後，方程式 6 表示資源限制，其中 Γ_k 是可以使用的每種資源類型的預設上限（以百分比計）。

【0026】 由於以下挑戰，很難在執行時間期間對所有時期進行全域優化：

1) 系統幾乎沒有關於未來工作負載模式的資訊，所以很難在執行時間期間對所有 τ 時段進行全域優化。

2) 當執行全域優化時，對於每一個時期的決策並非無關，這意味著傳統的次優類動力程式設計技術可能不能用於通過優化每一個時期的目標函數值“性能增益-性能損失”來解決問題。

3) 根據性能預測模型的複雜度（所述性能預測模型預測在不遷移的情況下在其它層上的 VM 性能），優化問題可以容易地變成 NP 難題。

【0027】 因此，為了在可接受的時間內解決這個問題，一些限制需要放寬。

【0028】 代替試圖對所有時間執行全域優化，可以尋求僅對每一

個時期優化的解決方案（即，執行時間貪婪演算法）。

【0029】 隨後的問題類似於加權背包問題。另外，由於每一個層的性能“專業”（例如高 IOPS、高輸送量、大空間、較小的寫入放大因數（WAF）函數、大量過度配置和較大程式設計/擦除（P/E）週期等）事先已知，所以可以計算關於每一個 VM 的以專業資源維度（例如層 1 的 IOPS 和輸送量）為重點的估計性能的“得分”，並且基於性能專業得分排序作出遷移決策，以及對遷移開銷的相應估計。

【0030】 可以安排兩個不同的時期：監測期，表示評估性能的頻率，且回歸性能估計模型，期可設定為任何所需的間隔：例如 15 分鐘；和遷移期，表示可以執行從一個層到另一個層的 VM 遷移的頻率，其可設定為任何所需的間隔：例如 1 天或 1 周。監測期和遷移期不一定固定：其可以使用能夠適應性地調節其在執行時間期間的大小的函數來確定。遷移期也不需要是監測期的整數倍數。其視窗大小越小，監測並測量儲存系統且遷移 VM 就越頻繁。系統管理員可以通過在使用之前進行靈敏度分析來平衡精確度與遷移成本之間的折衷。

【0031】 為了在不遷移的情況下估計 VM 在其它層上的性能，可以人工地將額外延遲增加到每一個 VM 的 I/O 命令且通過調用自動分層 I/O 濾波器 API（例如，用於 I/O 過濾的 VMware vSphere API）測量對總 I/O 延遲產生的效果來“模擬”VM 在另一層上的性能。

【0032】 可以使用線形曲線模擬 VM 在不同層中的性能改變。平均 I/O 延遲對比所增加的延遲的曲線可以回歸成線形方程式。線條的斜率可能取決於當時正在運行的應用程式是同步或非同步。同步應用程式和非同步應用程式在其如何操作上不同，其反映了線條的斜率。具體地說，當同步應用程式傳送 I/O 命令時，同步應用程式等待 I/O 命令完成，這會在通信正在進行時阻斷應用程式的進度，並使系統資源閒置。相比之下，非同步應用程式可以傳送 I/O 命令並且可以在傳輸已結束之前執行其它處理。這些事實意味著非同步應用程式傾向於受 I/O 命令延遲的較小影響，但改進出現的代價是：非同步應用程式並不知道回應將何時傳輸，且必須能夠在任何時間接收所述回應。

【0033】 VM 對層速度的靈敏度不同。具有同步應用程式的 VM 比具有非同步應用程式的 VM 對層速度變化更具有靈敏性。主要原因是非同步應用程式通常具有不太頻繁的 I/O 命令待刷新，且對 I/O 時序具有更小的依賴。

【0034】 由這些觀測結果推動，“層速度靈敏度校正會話”(TSSCS)可以用於通過在不真正遷移的情況下遷移 VM 到更高(或更低)速度層來預測所述 VM 可以增益(或損失)多少。詳細地說，TSSCS 具有以下特性：

1) 羽量級：在自動分層 I/O 濾波器上運行，TSSCS 將一些延遲增加到每一個 VM 的 I/O 命令中，且花費極短的時間以便不影響當前的託管工作負載。

2) 多個樣本/延遲/TSSCS：為改進在每一個層下的每一個 VM 性能的模擬精確度，可以平均在每一個層的相同增加延遲下所獲得的結果。

3) 多個延遲/TSSCS：為優化回歸，可以增加多個不同的延遲。此外，所增加的延遲並不必須受限於當前儲存池中的可用層。實際上，額外的延遲資料點可以用於提高回歸精確度。

4) 在執行時間期間的多個 TSSCS：TSSCS 可以被週期性觸發以更新曲線和回歸函數。

【0035】 基於相應的性能，性能增益與層速度之間的相關性可以通過線性回歸來確定，所述相關性反映了每一個 VM 對磁片速度的靈敏度提高或降低。曲線有可能回歸成多項式或非線性函數，但改變回歸並不影響自動分層如何操作。給定當時的每一個 VM 的增加延遲曲線，可以計算每一個 VM 在每一個層的 IOPS、輸送量、儲存大小等。注意到雖然 IOPS、輸送量和儲存大小為焦點，但其它類型的資源也可考慮。

【0036】 表 2 概述這個分段中所使用的符號，且表 3 進一步示出使用在三個不同層下的三個 VM 的實例的估計平均 I/O 延遲的結果。

【0037】 表 2. 符號。

符號	含義
L_t	層 t 的初始平均 I/O 延遲（無增加延遲）
b_x, m_x	用於 VM x 的經回歸層速度敏感度相關性模型的兩個參數，即 $y = mx - b$ ，其中 y 為平均 I/O 延遲且 x 為增加的延遲。

S_v	VM v 的平均 I/O 大小。
$S_v, VM[v].size$	VM v 的儲存大小。
$maxIOPS[t], maxMBPS[t], maxSize[t]$	層 t 的預設可用資源上限，其中我們設定所述上限為每一種類型資源的總容量的 90%。

【0038】 表 3. 在不遷移的情況下估計平均 I/O 延遲的實例

	層 1	層 2	層 3
V1 平均 I/O 延遲 (μs)	b_1	$m_1(L_2 - L_1) - b_1$	$m_1(L_3 - L_1) - b_1$
V2 平均 I/O 延遲 (μs)	$m_2(L_1 - L_2) - b_2$	b_2	$m_2(L_3 - L_2) - b_2$
V3 平均 I/O 延遲 (μs)	$m_3(L_1 - L_3) - b_3$	$m_3(L_2 - L_3) - b_3$	b_3

【0039】 層速度靈敏度校正會話（TSSCS）可以反覆運算儲存系統中的每一個 VM 和每一個層。所述層速度靈敏度校正會話可以針對每一個增加延遲計算樣本的 I/O 延遲的係數變數（即，CV）。隨後可以針對每一個增加延遲計算 I/O 延遲的平均值。隨後可以求所有增加延遲的 CV 值的平均值，且基於結果，自動分層可以界定用於產生每一個 VM 的線性回歸的增加延遲的置信度，以及因此隨後自身線性回歸中的置信度。CV 越低，這種 VM 具有的置信度就越高。若 CV 大於 1，則置信度可以設定為可以根據系統管理員的要求來調節的下限：例如 0.05。最後，TSSCS 可以調用回歸函數來計算相應的 m 值和 b 值。注意到雖然本發明概念的實施例可以使用在返回元組中具有僅兩個參數的線性回歸模型，但系統管理員可以基於其選擇的回歸模型添加更多專案到返回元組中。

【0040】 自動分層具有三種二維矩陣來記錄三種類型的資源，例如 IOPS（I/O 命令/秒）、輸送量（數百萬位元組/秒或 MBPS）和

儲存大小（位元組）。可以視需要添加用於儲存系統中的任何其它類型的資源的額外矩陣。與 IOPS 和輸送量相比，用於儲存大小的矩陣相對地簡單通過調用管理程式 API 測量每一個 VM 佔用的儲存大小來得到。

【0041】 自動分層可以使用基於經回歸 M 值和 B 值的線形估計來估計 IOPS。在得到用於計算不同層上的每一個 VM 的相應 IOPS 和輸送量的估計平均 I/O 延遲結果後，可以使用以下方程式來計算 IOPS 和輸送量：

$$IOPS = \frac{\text{平均IO大小}(v) \times 10^6}{\text{平均IO延遲}(v, t)}$$

(7)

$$\text{吞吐量} = \text{平均IO大小} \times IOPS$$

(8)

【0042】 由於不同類型的資源具有不同的單位，所以難以將其一起放置在單個多維矩陣中。因此，每一種資源可以轉換成每一個層的總的可用資源容量的百分比。這種歸一化的實例示出在表 4 中。

【0043】 表 4. 3 個 VM 在層 1 上運行時的歸一化性能矩陣的實例

	IOPS	輸送量	儲存
VM1	$\frac{10^6}{b_1 \times \max IOPS[t]}$	$\frac{s_1 \times 10^6}{b_1 \times \max MBPS[t]}$	$\frac{s_1}{\max Size[t]}$
VM2	$\frac{10^6}{[m_2(L_1 - L_2) + b_2] \times \max IOPS[t]}$	$\frac{s_2 \times 10^6}{[m_2(L_1 - L_2) + b_2] \times \max MBPS[t]}$	$\frac{s_2}{\max Size[t]}$
VM3	$\frac{10^6}{[m_3(L_1 - L_3) + b_3] \times \max IOPS[t]}$	$\frac{s_3 \times 10^6}{[m_3(L_1 - L_3) + b_3] \times \max MBPS[t]}$	$\frac{s_3}{\max Size[t]}$

【0044】 計算出的得分應反映以下事實：

1) 層和 VM 兩者的特徵：如果在每一個層上運行，則得分應反映每一個層的特殊性和每一個 VM 的工作負載特徵。因此，單獨地計算每一個層上的每一個 VM 的得分。

2) SLA 權重：VM 由於其具有不同的 SLA 權重而不相同，如方程式 1 中所示。

3) 估計的置信度：在性能矩陣中計算的係數變數可以用於反映估計的置信度。

4) 歷史成本和遷移成本：卷積老化因數可以用於將歷史得分包含到得分計算值中。遷移成本也可以被估計且包含在得分計算值中。

【0045】 基於使用者配置或預設配置，由每一個層提供的資源為事先已知的。自動分層可以對給定層的多種資源進行優化。為反映特殊性，可以使用層特殊性矩陣。實例層特殊性矩陣示出於表 5 中。在表 5 中，針對特殊性配置使用所有間隔尺寸或不使用間隔尺寸。具有 1 的儲存單元意謂當前層在相應類型的資源中具有特殊性。否則，所述層不針對所述類型的資源進行優化。但在本發明概念的其它實施例中，這個矩陣中的值可不一定為 0 和 1。如果系統管理員想要更細微性的間隔尺寸來控制特殊性，則其可以設定任何權值代替 0 和 1。系統管理員也可以通過添加更多類型的資源（即，更多行）來延伸這個矩陣。

【0046】 表 5.層特殊性矩陣的實例。

	IOPS	輸送量 (MBPS)	儲存大小 (位元組)
層 1	1	1	0
層 2	0	1	0
層 3	0	0	1

【0047】 在通過預設最昂貴到最不昂貴層的次序來分選所述層時是有幫助的，且這種次序被視為在遷移決策制定時段期間的優先次序。注意到系統管理員也可以基於其對每一個層的偏好來改變這種次序。

【0048】 下一個問題是“如何在反映每一個層的特殊性時反映每一個 VM 在每一個層上的性能？”。由於術語“性能”考慮多個維度或類型的資源（例如 IOPS、輸送量和儲存大小），所以加權總和用於反映每一個層的一種或多種特殊性。可以使用“正交匹配”，其是對“特殊性”矩陣和 VM 容量利用率矩陣的每一層每一 VM 的相乘操作，如方程式 9 中所示：

$$\begin{bmatrix} VM容量率矩阵[t][v].IOPS \\ VM容量率矩阵[t][v].MBPS \\ VM容量率矩阵[t][v].大小 \end{bmatrix} \times [特殊性[t].IOPS, 特殊性[t].MBPS, 特殊性[t].大小]$$

(9)

【0049】 另外，自動分層還允許系統管理員為每種類型的資源配置不同的權值。

【0050】 計算層 1 的容量得分的實例示出於表 5 中。第一行(“容

量率”) 中的(IOPS%，MBPS%，大小%)元組是對這個層上的每種類型的資源的估計 VM 利用率。在這個實例中，這些值中無一者大於 1，所以全部三個 VM 都通過限制測試。最後，假設特殊性矩陣如表 5 中所示。

【0051】 表 6. 3 個 VM 在層 1 上運行時的容量率、容量限制和容量得分值的實例。

	VM1	VM2	VM3
容量率	(0.5, 0.6, 0.3)	(0.2, 0.3, 0.3)	(0.1, 0.5, 0.4)
容量限制	通過	通過	通過
當前容量得分	$(0.5 + 0.6)/2 = 0.55$	$(0.2 + 0.3)/2 = 0.25$	$(0.1 + 0.5)/2 = 0.30$

【0052】 最終得分是歷史得分、當前得分和遷移成本的卷積總和，如方程式 10 中所示：

$\text{總得分} = \text{老化因數} \times \text{歷史得分} + \text{當前容量得分} - \text{遷移權重} \times \text{遷移成本}$	(10)
--	------

【0053】 這個最終得分具有三種組分：

- 1) 歷史得分：由於僅基於最新時期作出決策，為了避免一些 VM 在兩個層之間頻繁地遷移且切換，自動分層考慮歷史得分，以預設老化因數來衰減舊得分。最初，當前未分配到特定層的任一 VM 將接收零得分。
- 2) 當前容量得分：當前容量得分參考方程式（9）和表 6 如上文所描述計算。又，當前未分配到特定層的任一 VM 將接收零得分。

3) 遷移成本：自動分層通過選擇遷出和遷入可用輸送量的較小值來計算遷移速度瓶頸。自動分層隨著遷移時間成本傳回 VM 大小除以遷移速度的結果。

【0054】 為計算得分，針對每一個層，自動分層可以反覆運算其託管 VM 並檢查當前 VM 是否超過當前層 t 的任何類型資源的上限。若 VM 超過任一資源容量上限，則相應的 VM 容量利用率可以設定為 0，且層 t 上的這個 VM_v 的得分設定為 -1。否則，將資源使用率轉換成容量百分比。隨後如上文計算總得分。注意到自動分層亦允許系統管理員通過調節資源的權重值來控制每種類型資源的權重。

【0055】 自動分層在“長期視圖”中進行遷移。也就是說，自動分層遵循相對較低的頻率來遷移，這是因為遷移操作相對昂貴。

【0056】 自動分層可以檢查遷移操作是否需要且在必要時可以觸發相應的遷移。自動分層可以反覆運算每一個層，優先地以最昂貴到最便宜層的次序。注意到系統管理員可以基於其對每一個層待偏好來改變這種次序。隨後自動分層按降冪分選層得分。對於這個分選清單中的每一個 VM，若 VM 在以下情況下，則自動分層可以跳過所述 VM：

- 1) 當前反覆運算 VM 在當時已經被分配（遷移），或
- 2) 由於這個 VM 超過了這個層的容量限制，所以所述層上的所述 VM 沒有得分，或
- 3) 在分配期間層 t 上沒有充足的資源容量用於這個 VM_v。

【0057】 否則，自動分層可以將 VM 從最多得分到最少得分逐一地分配到這個層。注意到若 VM 已經在其目標層上，則這種分配不會遷移所述 VM。一旦 VM 被分配另一層後，則所述 VM 可以標記用於遷移，其包括從源儲存層複製資料（VMDK 檔）到目標儲存層。

【0058】 傳統資料中心使用 HDD 儲存資料。HDD 在大多數意義上通常與彼此相當類似，主要不同在於其每千百萬位元組的成本，其讀取/寫入速度和其整體容量。正如可預期的那樣，容量越大或讀取/寫入速度越快，HDD 就越昂貴。但即使考慮到其差異，HDD 也具有平均\$0.25/千百萬位元組的成本，且具有平均 10 毫秒（ms）的讀取/寫入速度。但甚至 10 ms 在電腦操作中也可以是較長的延遲，通常使得一些資料被局部快取以提高存取時間。

【0059】 在另一方面，SSD 可以提供不同的專業化。一些 SSD 可以具有較低的讀取/寫入延遲，其它 SSD 可以具有高頻寬，又其它 SSD 可以具有大容量，且又其它 SSD 可以具有較小的寫入放大因數。在本文中，術語“資源”用於描述單個儲存裝置可提供到更大程度或更小程度的這些專業化。其它 SSD 仍可以提供這些資源的組合。這些額外專業化為應用程式提供可以利用這些資源的益處，但使確定哪個 SSD 對於特定應用是最好的方法複雜化。

【0060】 與 HDD 相比，SSD 通常更昂貴，其中視格式而定，每千百萬位元組的價格在\$0.28 到\$4.68 的範圍內變化。但 SSD 通常還極具回應性：根據將使用的 SSD 格式，讀取/寫入時間可以在

1-2 ms 到微秒 (μs) 或甚至納秒 (ns) 的範圍內變化。

【0061】 雖然 SSD 通過 HDD 提供的附加資源是如何最好地將應用程式分配給 SSD 的重要因素，但存取時間是對應用程式可如何分配給 SSD 的影響最大。傳統快取演算法假設儲存層之間的性能得到相當大的改進：通常根據本地快取層相對於 HDD 儲存，預期性能得到 10 倍改進。此外，鑒於 SSD 比 HDD 儲存資料每千百萬位元組更昂貴，快取資料將導致資料以一定的成本儲存兩次。因此，在使用除 HDD 以外的儲存裝置時，傳統的分層和快取演算法產生次優結果。

【0062】 圖 1 示出根據本發明概念的實施例的可以託管虛擬機器並且使用經裝備以執行自動分層的儲存池的連網機器。在圖 1 中，示出系統 105。系統 105 可以包含虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3。雖然圖 1 示出了三種伺服器，但本發明概念的實施例可以支援任何數量的虛擬機器伺服器。

【0063】 虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 中的每一個可以分別包含處理器 115-1、處理器 115-2 和處理器 115-3，且分別包含記憶體 120-1、記憶體 120-2 和記憶體 120-3。在其它使用中，處理器 115-1、處理器 115-2 和處理器 115-3 可以執行應用程式，例如虛擬機器在虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 上運行所使用的應用程式，且所述應用程式可以儲存在記憶體 120-1、記憶

體 120-2 和記憶體 120-3 中。處理器 115-1、處理器 115-2 和處理器 115-3 可以是任何品種的處理器：例如 Intel Xeon、Celeron、Itanium 或 Atom 處理器、AMD Opteron 處理器、ARM 處理器等。雖然圖 1 示出了虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 中的每一個中的單一處理器，但虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 可以各自包含任何數量的處理器，其中的每一個可以是單核處理器或多核處理器，且可以任何所需的組合混合。記憶體 120-1、記憶體 120-2 和記憶體 120-3 各自可以是任何品種的記憶體，例如快閃記憶體、動態隨機記憶體（DRAM）、靜態隨機存取記憶體（SRAM）、永久性隨機存取記憶體、鐵電隨機存取記憶體（FRAM），或非揮發性隨機存取記憶體（NVRAM），例如磁阻隨機存取記憶體（MRAM）等。記憶體 120-1、記憶體 120-2 和記憶體 120-3 各自可以是不同記憶體類型的任何所需組合。儲器 120-1、記憶體 120-2 和記憶體 120-3 可以由記憶體控制器（在圖 1 中未示出）控制，所述記憶體控制器也是虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 的一部分。

【0064】 虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 各自還可以包含本地儲存裝置。舉例來說，虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 可以分別包含快取記憶體 125-1、快取記憶體 125-2 和快取記憶體 125-3。例如快取記憶體 125-1、快取記憶體 125-2 和快取記憶體

125-3 的本地儲存裝置可以採用任何所需形式，例如硬碟驅動器（HDD）或固態驅動器（SSD），且可以使用任何所需的連接器，例如周邊元件連接（PCI）、PCI Express（PCIe）、串列 AT 配件（SATA）、非揮發性記憶體 Express（NVMe）、網狀架構的 NVMe（NVMeoF）、M.2，或任何其它所需連接器。在本發明概念的一些實施例中，虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 可以忽略快取記憶體 125-1、快取記憶體 125-2 和快取記憶體 125-3，僅依賴資料中心中的儲存。在包含快取記憶體 125-1、快取記憶體 125-2 和快取記憶體 125-3 的本發明概念的實施例中，本發明概念的實施例可以與公開於 2018 年 2 月 7 日提交的美國專利申請案第 15/891350 號中的本發明概念的實施例組合，所述美國專利申請案要求 2017 年 11 月 29 日提交的美國專利申請案第 62/592,355 號的優先權，其兩者均出於所有目的以引用的方式併入本文中。

【0065】 系統 105 還可以包含資料中心伺服器 130。類似於虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3，資料中心伺服器 130 可以包含處理器 115-4 和記憶體 120-4。但資料中心伺服器 130 還可以包含儲存池 135、接收器 140 和傳輸器 145。儲存池 135 可以是經組織成儲存層的儲存裝置池。在本發明概念的一些實施例中，儲存池 135 可以包含僅 SSD 或其它快閃記憶體裝置；在本發明概念的其它實施例中，儲存池 135 可以包含使用其它儲存格式的儲存裝置。接收器 140 可以接收輸

入/輸出 (I/O) 命令和來自虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 的其它資料，且傳輸器 145 可以將資料傳輸到虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3。

【0066】 虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3，與資料中心伺服器 130 經示出已經由網路 150 連接。使用網路 150，虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 可以將 I/O 命令傳送到資料中心伺服器 130，且可以自其接收回應。虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 與資料中心伺服器 130 可以全部是局域網路 (Local Area Network; LAN) 的一部分，或可以包含其它網路類型，例如廣域網路 (Wide Area Network; WAN) 或例如網際網路的全球網路。網路 150 還可以是多個這類網路的組合，呈任何所需組合的形式。

【0067】 儘管圖 1 描繪虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 以及資料中心伺服器 130 作為伺服器 (且可以各自為獨立式或機架式伺服器)，但本發明概念的實施例可以包含各自具有任何所需類型而不受限的虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 以及資料中心伺服器 130。舉例來說，圖 1 所示的伺服器可以各自被桌上型或膝上型電腦或可以得益於本發明概念的實施例的任何其它裝置替換。虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器

伺服器 110-3 以及資料中心伺服器 130 各自還可包含特定得可攜式計算裝置、平板電腦、智慧手機和其它計算裝置。出於描述的目的，虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 被描述為託管虛擬機器，但本發明概念的實施例可以包含在虛擬機器伺服器 110-1、虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 上運行的不是虛擬機器的機器。

【0068】 圖 2 示出圖 1 的伺服器 110-1、伺服器 110-2、伺服器 110-3 和伺服器 130 的額外細節。在圖 2 中，通常地伺服器 110（其可以表示圖 1 的伺服器 110-1、伺服器 110-2 或伺服器 110-3 中的任一個）和伺服器 130 包含一個或多個處理器 115，所述處理器可以包含記憶體控制器 205 和時鐘 210，所述時鐘可以用於協調伺服器 110 和伺服器 130 的元件的操作。處理器 115 還可以耦合到記憶體 120，舉例來說，所述記憶體可以包含隨機存取記憶體（random access memory，RAM）、唯讀記憶體（read-only memory，ROM）或其它狀態保持媒體。在伺服器 130 中，處理器 115 還可以耦合到儲存裝置 135（伺服器 110 還可以包含局部附接的儲存裝置，但這類裝置不是所需的）。在伺服器 110 和伺服器 130 中，處理器 115 還可以耦合到網路連接器 215，所述網路連接器可以是例如乙太網連接器或無線連接器。處理器 115 還可以連接到匯流排 220，在其它元件中，使用者介面 225 和可以使用輸入/輸出引擎 230 管理的輸入/輸出介面埠可附接到所述匯流排。

【0069】 圖 3 示出根據本發明概念的實施例的圖 1 的虛擬機器伺

服器 110 與圖 1 的儲存池 135 在自動分層的情況下進行交互的附加細節。在圖 3 中，虛擬機器伺服器 110 示出託管三個虛擬機器 305-1、虛擬機器 305-2 和虛擬機器 305-3，但本發明概念的實施例可以包含由虛擬機器伺服器 110 託管的任何數量的虛擬機器。雖然僅虛擬機器 305-3 在下文更詳細地描述，但本發明概念的實施例包含其它虛擬機器，例如類似操作的虛擬機器 305-1 和虛擬機器 305-2。虛擬機器 305-3 示出為包含在作業系統 315 上運行的應用程式 310。管理程式 320 可以管理在虛擬機器伺服器 110 上運行的操作虛擬機器 305-1、操作虛擬機器 305-2 和操作虛擬機器 305-3。

【0070】 I/O 濾波器 325 可以充當傳遞濾波器，接收來自應用程式 310 的命令並將其傳送到管理程式 320 用於執行。I/O 濾波器 325 可以是任何所需達 I/O 濾波器，例如用於 I/O 過濾掉 VMware vSphere API。但除其正常操作以外，I/O 濾波器 325 可以將延遲增加到一些來自虛擬機器 305-3 的 I/O 命令中，所述虛擬機器可以提供關於 VM 在由其它儲存層支援時如何估計其性能的資訊。I/O 濾波器 325 的操作參考圖 7 進一步論述。

【0071】 自動分層幕後程式 330 可以接收來自 I/O 濾波器 325 的性能資料並將其傳送到自動分層控制器 335。自動分層幕後程式 330 可以是管理程式 320 的一部分，或其可以是獨立元件。另外，雖然自動分層幕後程式 330 可僅是簡單對傳遞裝置，接收來自 I/O 濾波器 325 的資料並將其傳送到自動分層控制器 335，但自動分層

幕後程式 330 可以包含稍後歸因於自動分層控制器 335 的功能中的一些，例如對關於虛擬機器 305-1、虛擬機器 305-2 和虛擬機器 305-3 的性能資料執行線性回歸的操作。而在本發明概念的一些實施例中，自動分層幕後程式 330 可以完全被省略，其中 I/O 濾波器 325 直接與自動分層控制器 335 連通而不是准許自動分層幕後程式 330 收集並傳送性能資料。

【0072】 自動分層控制器 335 負責取得從自動分層幕後程式 330 接收的性能資料並使用所述資料以估計虛擬機器 305-3 將在每一個儲存層上如何執行，之後自動分層控制器 335 可以按需要將資料從虛擬機器 305-3 遷移到另一儲存層。由自動分層控制器 335 管理的儲存池 135 可以包含任何數量的儲存層：圖 3 示出三個層 340-1、層 340-2 和層 340-3。在每一個層內，可存在任何數量的儲存裝置 345。

【0073】 如上文所提及，每一個儲存層可以提供利用由所述儲存層提供的特定資源的專業化。舉例來說，儲存層 340-1 可以提供快速的讀取/寫入時間，而儲存層 340-3 可以提供較小的寫入放大因數。由於儲存層 340-1、儲存層 340-2 和儲存層 340-3 提供專業化，所以所述儲存層內的儲存裝置 345 應全部提供專業化。因此，在本發明概念的一些實施例中，給定儲存層內的儲存裝置 345 相對於一個或多個特殊性而言是相對或完全均質的：在一些情況下，在給定儲存層內的儲存裝置 345 和所有儲存裝置 345 應是相同模式的儲存裝置。

【0074】 在本發明概念的一些實施例中，自動分層控制器 335 實施為圖 1 的資料中心伺服器 130 的一部分。在本發明概念的其它實施例中，自動分層控制器可以實施為可能在虛擬機器伺服器 110-1 虛擬機器伺服器 110-2 和虛擬機器伺服器 110-3 中的一個上運行的虛擬機器。

【0075】 圖 4 示出由不同儲存裝置提供的不同可用資源。在圖 4 中，示出資源 405。資源 405 可以包含輸送量 410，所述輸送量表示可以在給定時間量例如一秒中執行的 I/O 命令的數量（通常描述為 I/O 操作/秒，或 IOPS）；頻寬 415，所述頻寬反映在給定時間單位內可以傳輸多少資料（通常以數百萬位元組/秒，或 MBPS 描述）；容量 420，所述容量反映可以儲存多少資料；以及寫入放大因數 425，所述寫入放大因數反映由於垃圾收集操作而寫入多少額外資料。資源 405 可以容易擴展以包含由圖 1 的儲存池 135 提供的任何其它所需資源。

【0076】 圖 5 示出將從一個儲存裝置遷移到另一個儲存裝置的圖 3 的虛擬機器的檔。在圖 5 中，虛擬機器檔 505 儲存虛擬機器的資料。虛擬機器檔 505 當前駐存在儲存層 340-3 上，而第二虛擬機器的虛擬機器檔 510 目前駐存在儲存層 340-2 上。如箭頭 515 所示，虛擬機器檔 505 可以遷移到儲存層 340-2 或儲存層 340-1，根據何種遷移，圖 3 的自動分層控制器 335 將會引起圖 1 的系統 105 整體的最好提高。注意到圖 3 的自動分層控制器 335 可以決定不遷移虛擬機器到另一儲存層，如虛擬機器檔 510 在儲存層 340-2

上保持不遷移的事實所示。

【0077】 雖然最好存在一個如何將虛擬機器分配到不變（至少，只要虛擬機器集群不改變）的儲存層的最優解決方案，但實際上虛擬機器可能經歷工作負載改變，且可以與其它 VM（其需求也可以隨時間改變）競爭資源。工作負載和競爭的這些變化可以影響哪個儲存層對於特定虛擬機器來說是“最好的”。舉例來說，在一個時間點，一個虛擬機器可以非常忙且需要來自圖 1 的儲存池 135 的快速回應時間，但在另一時間點，同一虛擬機器可以暫停工作且不需要高速儲存層。因此，可能不存在一直為所有虛擬機器最好工作的單一全域優化。

【0078】 因此，代替尋求單一的全域優化，優選的是尋求一種在一段時間內工作的優化，之後可能優選的是不同的優化。舉例來說，圖 6A 到圖 6B 示出在不同時間點虛擬機器到儲存層的最優分配。舉例來說，在圖 6A 中，在時間 T0 處，虛擬機器檔 505 最優在儲存層 340-3 中且虛擬機器檔 510 最優在儲存層 340-1 中。相比之下，在圖 6B 中，在時間 T1 處，虛擬機器檔 505 和虛擬機器檔 510 兩個最優在儲存層 340-1 中。

【0079】 圖 7 示出增加延遲並且確定圖 3 的虛擬機器的性能資料的圖 3 的 I/O 濾波器 325。在圖 7 中，I/O 濾波器 325 可以接收來自圖 3 的虛擬機器 305-3 的應用程式 310 的 I/O 命令 705。I/O 濾波器 325 隨後可以將延遲 710 增加到 I/O 命令 705 中。這種增加延遲可以模仿更慢執行儲存的層，所述更慢執行儲存的層針對虛

擬機器產生不同的平均 I/O 延遲。

【0080】 使用所增加的延遲 710，I/O 濾波器 325 可以提供 I/O 命令 705 到資料中心伺服器 130。資料中心伺服器 130 隨後可以處理 I/O 命令 705 並返回回應 715。I/O 濾波器 325 可以返回回應 715 到應用程式 310 並且可以使用所增加的延遲 710 測量圖 3 的虛擬機器 305-3 的平均 I/O 延遲，且產生性能資料 720。圖 3 的自動分層控制器 335 隨後可以使用性能資料 720 來確定是否存在將虛擬機器遷移到另一儲存層（更快或更慢，這是因為遷移虛擬機器到更慢的儲存層可以為另一個具有性能補償改進的虛擬機器釋放空間）的優勢。

【0081】 雖然圖 7 示出使用所增加的延遲 710 以產生性能資料 720 的 I/O 濾波器 325，但所增加的延遲 710 具有其自身的開銷。視需要，所增加的延遲 710 可以用輕量路徑替換，條件是可以產生性能資料 720。舉例來說，I/O 濾波器 325 可以跟蹤未完成 I/O 命令的數量/虛擬機器，且使用所述資訊以提出在未完成的 I/O 命令數量仍不斷地增高或保持遞增時將虛擬機器遷移到另一儲存層。

【0082】 圖 8 示出圖 3 的自動分層控制器 335 的細節。在圖 8 中，自動分層控制器 335 可以包含線性回歸模組 805、資源矩陣生成器 810、度量轉換器 815、專業化資訊 820、層選擇模組 825 和遷移模組 830。線性回歸模組 805 可以採用圖 7 性能資料 720 並對所述資料執行線性回歸，產生估計虛擬機器對於給定儲存層的延遲的

平均 I/O 延遲的線形方程式。如下文參考圖 10 所描述，給定增加的延遲，可以確定給定虛擬機器的平均 I/O 延遲，並將這些資料點繪製在曲線上。在本發明概念的一些實施例中，這些資料點是近似地線形，且線性回歸可以產生近似給定的增加延遲的平均 I/O 延遲的方程式。隨後，已知每一個儲存層的延遲，可以針對每一個儲存層估計給定虛擬機器的平均 I/O 延遲。

【0083】 資源矩陣生成器 810 可以採用可以由使每一個虛擬機器與每一個儲存層和由每一個儲存層提供的資源相關聯的線性回歸產生的估計。舉例來說，可以根據虛擬機器的平均 I/O 大小和虛擬機器的平均 I/O 延遲來計算輸送量，且可以根據 VM 的 I/O 命令的平均 I/O 大小和 VM 發出的 IOPS 計算頻寬。資源矩陣生成器 810 可以使用度量轉換器 815 將一個度量的性能資料（例如輸送量）轉換成另一度量的性能資料（例如頻寬）。

【0084】 專業化資訊 820 可以指示每一個儲存層提供什麼資源。專業化資訊進一步參考下圖 11 來論述。

【0085】 層選擇模組 825 可以採用專業化資訊 820 和由資源矩陣生成器 810 產生的資源矩陣，並且選擇虛擬機器的儲存層。隨後，如果虛擬機器將從一個儲存層遷移到另一個儲存層，則遷移模組 830 可以執行所述遷移。

【0086】 圖 9 示出將圖 7 的性能資料 720 轉換成使虛擬機器、儲存層和所述儲存層的資源相關聯的矩陣。在圖 9 中，性能資料 720 表示為在使用給定儲存層時每一個虛擬機器的平均 I/O 延遲的矩

陣。因此舉例來說，估計虛擬機器 1 在使用儲存層 1 時具有 800 μ s 的平均延遲，在使用儲存層 2 時具有 500 μ s 的平均延遲，且在使用儲存層 3 時具有 300 μ s 的平均延遲，其中虛擬機器 2 和虛擬機器 3 的性能資料提供類似的資訊。圖 8 中所示出的數值僅為示範性的，且不必反映任何虛擬機器在儲存層中的真實性能值，這是因為實際性能很大程度上取決於特定虛擬機器將運行的應用程式和其它虛擬機器從同一儲存池存取資料的影響兩者。

【0087】 注意到性能資料 720 中的一些值不是估計值而是實際值。具體地說，每一個虛擬機器將其資料儲存在圖 1 的儲存池 135 中的一些儲存層上。因此，對於實際儲存虛擬機器的資料的儲存層來說，性能資料 720 中的平均 I/O 延遲表示所述虛擬機器在所述儲存層上的實際平均 I/O 延遲，而不是估計值。

【0088】 給定每一個虛擬機器在每一個儲存層上的估計平均延遲，可以產生使虛擬機器、儲存層和資源相關聯的多維矩陣。這種多維矩陣在圖 9 中示出為資源矩陣 905。由於很難在紙上示出多維矩陣，所以圖 9 將資源矩陣 905 示出為若干單獨的二維矩陣 910、二維矩陣 915 和二維矩陣 920，其中每一個二維矩陣儲存單個儲存層的資料，但這僅出於觀測目的。

【0089】 在儲存層 1 的二維矩陣 910 中，根據性能資料 720 的平均 I/O 延遲可以使用公式 $\frac{10^6}{AverageIOLatency * maxIOPS[t]}$ 轉換成輸送量：也就是說， 10^6 （1,000,000）除以所討論的使用所述儲存層的

虛擬機器的平均 I/O 延遲與儲存層可以支援的最大 IOPS（這個值可以由系統管理員設定為小於儲存層所提供的真正最大值的一些值：例如，真正最大值的 90%）的乘積。可以使用值 106，這是因為延遲通常以微秒來表達（一微秒 = $\frac{1}{1,000,000}$ ）。因此，除非將比如 12 μs 的延遲轉換成 0.000012 秒，否則平均 I/O 延遲的精確測量應反映方程式中出現的單位變化。輸送量可以通過用輸送量乘以由虛擬機器發出的 I/O 命令的平均寫入大小來轉換成頻寬。並且容量可以根據虛擬機器檔的儲存大小除以由儲存層提供的最大儲存來計算（另外，系統管理員可以將這個值設定成小於真正最大值的一些值）。可以使用圖 8 的度量轉換器 815 計算從性能資料 720 到儲存於資源矩陣 905 中的值的這些轉換。

【0090】 圖 10 示出圖 7 的性能資料 720 的線性回歸。在圖 10 中，示出兩種不同虛擬機器的性能資料。對於一個虛擬機器，曲線上的資料點可以通過線性回歸 1005 來粗略估計；對於另一個虛擬機器，曲線上的資料點可以通過線性回歸 1010 來粗略估計。雖然圖 10 示出僅兩個虛擬機器的資料點，但本發明概念第實施例可以支援線形回歸可以計算的任何數量的虛擬機器。

【0091】 每一個數據點使特定的增加延遲與虛擬機器的平均 I/O 延遲相關聯。因此，舉例來說，當沒有增加的延遲時，第一虛擬機器經歷 3000 μs 的平均 I/O 延遲，而第二虛擬機器經歷近似 700 μs 的平均 I/O 延遲。

【0092】 增加的延遲尤其在確定虛擬機器是否會使用較慢儲存層（即，具有較低性能水準的儲存層）來充分操作時起作用。畢竟，增加延遲的量可以正好是當前儲存層與目標儲存層之間的差。但使用增加的延遲不可能模擬較快儲存層：畢竟，不可能增加負值延遲，且儲存層不可能比其實際能夠回應更快速地回應。解決方案如線性回歸 1010 所示。使用關於虛擬機器如何回應較慢儲存層的資訊，有可能使用線性回歸模型估計更快儲存層的性能。在技術上，與“內插法”相比較，估計所提供資料範圍之外的值稱作“外推法”，所述內插法是對資料範圍內的值進行估計。但當使用例如線性回歸 1010 的線性回歸時，計算對於內插法和外推法是相同的。因此，使用線性回歸 1010，有可能估計虛擬機器在使用更快儲存層時將會如何回應。

【0093】 雖然有可能使用僅對應於儲存層之間的實際延遲差的資料點，但更多數據點是可用的，由線性回歸模組 805 提供的估計值有可能更好。因此，可以使用圖 3 的 I/O 濾波器 325 增加各種不同的延遲以視需要提供為多個數據點。

【0094】 當然，如果虛擬機器的每一個 I/O 命令使用增加的延遲而減緩，那麼虛擬機器將不接收經分配儲存層的真正益處。因此，所增加的延遲可以僅用於由虛擬機器發出的 I/O 命令子集。如果選擇以增加延遲的 I/O 命令是隨機選擇的，那麼圖 7 的性能資料 720 應提供有意義的線形回歸 1005 和線形回歸 1010。已知的統計技術可以用於確定資料點的變數係數（和置信度）。如果資料點中

的置信度足夠高，那麼可以信任線形回歸 1005 和線形回歸 1010 為精確估計值。可以指定任何所需的置信度：例如，90%或 95%。參考下圖 13 進一步論述如何選擇 I/O 命令用於增加延遲。

【0095】 注意到線形回歸 1005 和線形回歸 1010 完全非常不同的資料集：線性回歸 1005 包含極其陡峭的斜率，而線性回歸 1010 包含極其淺的斜率。並非所有虛擬機器以相同的方式回應其 I/O 命令的延遲變化：一些虛擬機器比其它更敏感。舉例來說，產生得到線性回歸 1005 的資料點的虛擬機器可以運行同步應用程式，其中延遲的較小變化可以具有對應用程式性能的顯著影響。在另一方面，產生得到線性回歸 1010 的資料點的虛擬機器可以運行非同步應用程式，其中延遲的變化具有較小影響。或者，線性回歸 1005 可以用於需要高輸送量的虛擬機器，而線性回歸 1010 可以用於需要低輸送量的虛擬機器。以類似的方式，虛擬機器的靈敏度可以影響其線性回歸：更敏感的虛擬機器將具有更陡的斜率，且不太敏感的虛擬機器將具有更淺的斜率。針對靈敏度的類似概念是虛擬機器是緊密耦合或不是緊密耦合（這裡的“緊密耦合”是指虛擬機器對儲存層的回應時間有多敏感或依賴，類似於同步應用程式與非同步應用程式之間的比較）：虛擬機器越緊密耦合，其線性回歸的斜率將越可能更陡。一般來說，具有陡峭斜率的線形回歸，如線性回歸 1005，表示在較快儲存層中執行更好的虛擬機器，而具有淺斜率的線形回歸，如線性回歸 1010，表示在較慢儲存層中充分執行的虛擬機器。

【0096】 圖 11 示出關於圖 3 的儲存層 340-1、儲存層 340-2 和儲存層 340-3 的圖 8 的專業化資訊 820。在圖 11 中，專業化資訊 820 示出為矩陣。每一個儲存層與每一個資源的交叉點可以指定反映儲存層是否提供所述資源（和/或通過多少所述資源）的值。在本發明概念的一些實施例中，如 11 圖中所示，專業化資訊 830 可以使用二進位資訊（0 與 1，是與否，或任何其它對值），其中一個值表明儲存層提供所述資源並且另一個值表明儲存層不提供所述資源因此舉例來說，專業化資訊 820 可以讀取為表明儲存層 1 提供高輸送量和高頻寬，儲存層 2 提供高頻寬，且儲存層 3 提供高容量。

【0097】 在本發明概念的其它實施例中，專業化資訊 820 中的值不必受限於二進位選項。舉例來說，所述值可以在 0 到 1 的範圍內，其中 0 表明儲存層不提供所述資源，1 表明儲存層完整提供所述資源，且兩者之間的值可以表示部分支援。因此，比如 0.75 的值可以表明儲存層在一定程度上提供資源，但不與“最好”儲存層一樣好，而比如 0.25 的值可以表明儲存層提供比“最差”儲存層略好的資源（其中“最好”和“最差”為相對術語）。

【0098】 圖 12 示出計算待用於將圖 3 的虛擬機器 305-1、虛擬機器 305-2 和虛擬機器 305-3 遷移到圖 3 的不同儲存層中的得分矩陣。通過專業化資訊 820 與資源矩陣 905 相乘，可以生成反映每一個層支持每一個虛擬機器的程度的得分矩陣 1205。舉例來說，值被提供給虛擬機器 1，表明儲存層 1 是 55% 匹配，儲存層 2 是

25%匹配，且儲存層 3 是 30%匹配。（注意到所述值並不總計為 100%，這是因為每一個值表示每一個儲存層將支援虛擬機器的程度，而不是特定虛擬機器應儲存於特定儲存層中多少。）可以類似地解釋另一虛擬機器的項目。

【0099】 一旦產生得分矩陣 1205 後，圖 8 的層選擇 825 可以為每一個虛擬機器選擇適當的層。這種選擇可以通過計算層和虛擬機器的每一個組合的以下值來完成：最終得分 = 老化因數 × 歷史最終得分 + 當前性能增益 - 當前性能代償。老化因數可以用於使虛擬機器和儲存層的組合的最終得分變舊的因數（以防止早期選擇過度偏置結果）。當前性能增益可以是根據虛擬機器和層的組合的圖 12 的得分矩陣 1205 繪製的值。當前性能代償表示從一個儲存層遷移資料到另一個儲存層的成本。

【0100】 應注意，由於虛擬機器可以駐存於除含有圖 1 的儲存池 135 的伺服器以外的不同伺服器上，當前性能代償僅反映移動資料本身的成本，且不需要將對虛擬機器操作的任何影響作為考慮因素。換句話說，由於虛擬機器駐存於圖 1 的虛擬機器伺服器 110-1 上，所以託管虛擬機器的伺服器不需要有處理能力來執行資料移轉：所有處理是在圖 1 的資料中心伺服器 130 上完成的。當前性能代償可以通過用需要遷移的虛擬機器檔大小除以源儲存層的更小讀取輸送率和目標儲存層的寫入輸送率來計算。換句話說，比較從源儲存層讀取資料所需的時間和將所述資料寫入目標儲存層所需的時間，且選擇更大值作為執行遷移所需的時間，這是因為

無論哪個較慢（讀取源儲存層或寫入目標儲存層）都表示資料移轉中的瓶頸。

【0101】 給定虛擬機器和儲存層的每一個組合的最終得分，層選擇模組 825 可以容易地為每一個虛擬機器選擇儲存層以通過確認虛擬機器到儲存層的分配來全域優化系統 105 的整體性能，所述分配提供對圖 1 的系統 105 的性能的整體最大改進，從而將性能改進和性能降低（通過將虛擬機器遷移到不太強大的儲存層）以及遷移成本作為考慮因素。

【0102】 如上文提示，針對所有時間執行單一全域優化假設每一個虛擬機器的工作負載不改變，且虛擬機器集群自身不改變（沒有添加新的虛擬機器且不移除現有的虛擬機器）。實際上，這些情況是不真實的。因此，如果圖 1 的系統 105 僅優化一次，隨時間推移，虛擬機器到儲存層的分配可能不再是最優的。這種問題的解決方案是以規律間隔執行優化，而不是僅一次。但在連續基礎上執行優化不是高效的：例如，虛擬機器 I/O 命令將必須在連續基礎上經歷增加的延遲，意味著無虛擬機器會接收其儲存層的全部益處。因此，並非執行連續優化，可以在定期基礎上執行優化。為此，引入遷移期和監測期的概念。

【0103】 監測期表示圖 3 的 I/O 濾波器 325 可以將延遲增加到一些 I/O 命令中以確定虛擬機器現在可以在不同儲存層中如何執行的時間點。當監測期開始時，圖 3 的 I/O 濾波器 325 可以將延遲增加到所選擇數量的 I/O 命令中。這些數量的 I/O 命令可以預設為

固定數量，或其可以設定為在先前監測期期間發出的 I/O 命令數量的百分比（比如，1%），或這兩個值（比如，這兩個值中的更大值）。監測期可以以規律間隔，例如每 15 分鐘、或每小時或每天出現，但本發明概念的實施例可以支援不規律的監測期。應注意不存在預測在監測期開始時虛擬機器的實際工作負載將是什麼的簡單方式。因此，觸發圖 3 的 I/O 濾波器 325 以在監測期開始處將延遲增加到 I/O 命令中有可能將選擇隨機的 I/O 命令集且提供圖 7 的有意義的性能資料 720。一旦已產生虛擬機器在給定監測期中的圖 7 的性能資料 720，圖 3 的自動分層幕後程式 330 可以將圖 7 的性能資料 720 傳送到圖 3 的自動分層控制器 335。替代地，圖 3 的自動分層幕後程式 330 可以等待直到遷移期開始並在當時傳送圖 7 的性能資料 720。

【0104】 遷移期表示圖 3 的自動分層控制器 335 可以確定虛擬機器是否應從一個儲存層遷移到另一個儲存層的時間點。如同監測期，遷移期可設定以規律或不規律時間間隔，例如每天或每週出現。當監測期出現時，圖 3 的自動分層控制器 335 可以分析圖 7 的性能資料 720，產生圖 9 的資源矩陣 905，計算最終得分，選擇用於每一個虛擬機器的儲存層，以及必要時遷移虛擬機器資料。

【0105】 圖 13 示出監測期與遷移期之間的關係。在圖 13 中，示出遷移期 1305 和監測期 1310。如圖 13 中可見，遷移期 1305 通常比監測期 1310 更長，但本發明概念的實施例可以具有與監測期 1310 一樣才或更短的遷移期 1305。另外，遷移期 1305 和監測期

1310 不需要對準，但其可以對準。

【0106】 圖 14 示出根據本發明概念的實施例的用於圖 3 的 I/O 濾波器 325 將延遲增加到圖 7 的輸入/輸出 (I/O) 命令 705 以產生圖 7 的性能資料 3615 的實例過程的流程圖。在圖 14 中，在塊 1405 處，圖 3 的 I/O 濾波器 325 可以確定圖 13 的新監測期 1310 是否已開始。如果未開始，那麼圖 3 的 I/O 濾波器 325 可以等待。否則，在塊 1410 處，圖 3 的 I/O 濾波器 325 可以接收來自圖 3 的應用程式 310 的圖 7 的一些 I/O 命令 705。圖 3 的 I/O 濾波器 325 以這種方式選擇多少命令來處理可以根據系統組態改變。在塊 1415 處，圖 3 的 I/O 濾波器 325 可以將圖 7 的延遲 710 增加到圖 7 的 I/O 命令 705 中。在塊 1420 處，圖 3 的 I/O 濾波器 325 可以將命令傳送到圖 1 的資料中心伺服器 130。在塊 1425 處，圖 3 的 I/O 濾波器 325 可以接收來自圖 1 的資料中心伺服器 130 的圖 7 的回應 715。在塊 1418 處，圖 3 的 I/O 濾波器 325 可以將圖 7 的回應 715 傳遞到圖 3 的應用程式 310。最後，在塊 1430 處，圖 3 的 I/O 濾波器 325 可以使用虛擬機器的平均 I/O 延遲計算圖 7 的性能資料 720。

【0107】 圖 15A 到圖 15B 示出根據本發明概念的實施例的用於圖 3 的自動分層控制器 335 管理虛擬機器在儲存層中的遷移的實例過程的流程圖。在圖 15A 中，在塊 1505 處，圖 3 的自動分層控制器 335 可以確定圖 13 的新監測期 1310 是否已開始。如果未開始，那麼圖 3 的自動分層控制器 335 可以等待。否則，在塊 1510 處，

圖 3 的自動分層控制器 335 可以接收來自圖 3 的自動分層幕後程式 330 的圖 7 的性能資料 720。在塊 1515 處，圖 3 的自動分層控制器 335 可以確定圖 13 的新遷移期 1305 是否已開始。如果未開始，那麼圖 3 的自動分層控制器 335 可以等待。

【0108】 否則，在塊 1520 處，圖 8 的資源矩陣生成器 810 可以產生圖 9 的資源矩陣 905(在其它元件中使用圖 8 的線性回歸模組 805 和圖 8 的度量轉換器 815)。在塊 1525 處，層選擇模組 825 可以使用資源矩陣 905 將虛擬機器分配到儲存層。在塊 1530 處，圖 3 的自動分層控制器 335 可以檢查以查看虛擬機器中的任一個是否從一個儲存層遷移到另一個儲存層。如果是，則在塊 1535 處，遷移模組 830 可以將圖 5 的虛擬機器檔 505 從一個儲存層遷移到另一個儲存層。

【0109】 圖 16 示出根據本發明概念的實施例的用於圖 3 的自動分層控制器 335 以產生圖 9 的資源矩陣 905 的實例過程的流程圖。在圖 16 中，在塊 1605 處，圖 8 的線性回歸模組 805 可以對圖 7 的性能資料 720 執行線性回歸以產生估計虛擬機器將在每一個儲存層中執行的程度的方程式。在塊 1610 處，圖 3 的自動分層控制器 335 可以使用圖 10 的線形回歸 1005 和線形回歸 1010 以估計每一個虛擬機器將在每一個儲存層中執行的程度。在塊 1615 處，資源矩陣生成器 810 可以根據每一個資源生成虛擬機器在每一個儲存層中的性能度量。如塊 1620 中所示，這可以包含使用度量轉換器 815 將一個度量(例如輸送量)轉換成另一個度量(例如頻寬)。

塊 1620 可以省略，如由虛線 1625 所示。

【0110】 如從以上論述中可見，本發明概念的實施例提供比傳統分層和快取解決方案顯著的優勢。第一，本發明概念的實施例支援考慮儲存層的多個資源。另外，由於資源數量可以改變，所以資源數量可以設定為 1，從而提供與傳統 HDD 儲存池一起使用的舊版相容性。第二，本發明概念的實施例對快閃記憶體儲存系統的較快性能做出解釋，忽略傳統分層和快取解決方案。第三，由於現有虛擬機器集群的改變和虛擬機工作負載（例如工作負載峰）出現，本發明概念的實施例支持尋找全域優化並隨時間推移調節所述全域優化。這與傳統解決方案相當，其通常為個人虛擬機器尋找本地解決方案而不是整體上優化系統。

【0111】 在圖 14 到圖 16 中，示出本發明概念的一些實施例。但本領域技術人員將認識到，通過改變塊的次序、通過省略塊，或通過包含未在附圖中示出的鏈路，本發明概念的其它實施例也是可能的。無論是否明確地描述，流程圖的所有此類變化都被視為本發明概念的實施例。

【0112】 以下論述預期提供其中可以實施本發明概念的某些方面的一個或多個合適機器的簡要一般描述。一個或多個機器可以至少部分地通過從例如鍵盤、滑鼠等傳統輸入裝置的輸入，以及通過從另一機器接收的指令、與虛擬實境（virtual reality, VR）環境的交互、生物計量回饋或其它輸入信號控制。本文所使用的術語“機器”意欲廣泛地涵蓋單個機器、虛擬機器，或以通信方式

耦合的機器、虛擬機器或一起操作的裝置的系統。示範性機器包含計算裝置，例如個人電腦、工作站、伺服器、可攜式電腦、掌上型裝置、電話、平板電腦等，以及交通裝置，例如私人或公共交通工具，例如汽車、火車、計程車等。

【0113】 一個或多個機器可以包含嵌入式控制器，例如，可程式設計或不可程式設計邏輯裝置或陣列、專用集成模組 (Application Specific Integrated Modules, ASIC)、嵌入式電腦、智慧卡等。一個或多個機器可以利用例如通過網路介面、數據機或其它通信耦合與一個或多個遠端機器的一個或多個連接。機器可以借助於物理和/或邏輯網路，例如企業內部網、網際網路、局域網、廣域網路等互連。本領域技術人員將理解，網路通信可以利用各種有線和/或無線短程或遠端載波和協議，包含射頻 (radio frequency, RF)、衛星、微波、電氣電子工程師學會 (Institute of Electrical and Electronics Engineer, IEEE) 802.11、Bluetooth®、光學、紅外線、電纜、鐳射等。

【0114】 可以參考或結合包含功能、過程、資料結構、應用程式等的相關聯資料描述本發明概念的實施例，所述相關聯資料在通過機器訪問時引起機器執行任務或定義抽象資料類型或低級硬體環境。相關聯資料可以儲存在例如揮發性和/或非揮發性記憶體，例如 RAM、ROM 等中，或其它儲存裝置以及其相關聯儲存媒體中，包含硬碟驅動器、軟碟、光學儲存裝置、磁帶、快閃記憶體、儲存棒、數位視訊光碟、生物儲存裝置等。相關聯資料可以採用

包、串列資料、並行資料、傳播信號等形式在包含物理和/或邏輯網路的傳輸環境上傳遞，並且可以用壓縮或加密格式使用。相關聯資料可以用於分散式環境中，並且本地和/或遠端地儲存用於機器訪問。

【0115】 本發明概念的實施例可以包含有形的非暫時性機器可讀媒體，包括可由一個或多個處理器執行的指令，所述指令包括用於執行如本文所描述的發明性概念的元件的指令。

【0116】 上文所描述的方法的各種操作可以通過能夠執行所述操作的任何合適構件（例如各種硬體和/或軟體元件、模組及/或一個或多個模組）來執行。軟體可以包含用於實施邏輯功能的可執行指令的有序列表，且可以任何供指令執行系統、設備或裝置使用或與指令執行系統、設備或裝置結合的“處理器可讀媒體”實施，所述指令執行系統、設備或裝置如單核處理器或多核處理器或含處理器的系統。

【0117】 結合本文中所公開的實施例而描述的方法或演算法和函數的步驟可直接體現在硬體、由處理器執行的軟體模組或所述兩者的組合中。如果實施於軟體中，則可以將功能作為一個或多個指令或代碼而儲存在有形的非暫時電腦可讀媒體上或經由電腦可讀媒體進行傳輸。軟體模組可駐留在隨機存取記憶體（RAM）、快閃記憶體、唯讀記憶體（ROM）、電可程式設計 ROM（EPROM）、電可擦除可程式設計 ROM（EEPROM）、暫存器、硬碟、可裝卸磁片、CD ROM 或所屬領域中已知的任何其它形式的儲存媒體中。

【0118】 在參考所說明實施例描述和說明本發明概念的原理之後，將認識到，所說明實施例的佈置和細節可以在不脫離此原理的情況下進行修改，並且可以按任何所需方式組合。而且，儘管前述論述集中於具體實施例，但預期其它配置。具體地說，即使本文使用例如“根據本發明概念的實施例”等的表達，但這些詞語一般表示參考實施例可能性，且並不意圖將本發明概念限制到具體實施例配置。如本文所使用，這些術語可以參考組合到其它實施例中的相同或不同實施例。

【0119】 前述說明性實施例不應被解釋為限制其本發明概念。雖然已經描述幾個實施例，但本領域技術人員將容易瞭解，在實質上不脫離本公開的新穎教示和優點的情況下，對這些實施例的許多修改是可能的。因此，所有此類修改意欲包含在如所附權利要求中定義的本發明概念的範圍內。

【0120】 本發明概念的實施例可以擴展到以下陳述，而不具有限制性：

【0121】 陳述 1. 本發明概念的實施例包含一種系統，所述系統包括：

提供多個資源的多個儲存裝置，所述多個儲存裝置組織成多個儲存層並且儲存第一虛擬機器的第一資料和第二虛擬機器的第二資料；

接收器，接收來自所述第一虛擬機器的第一輸入/輸出(I/O)命令、來自所述第二虛擬機器的第二 I/O 命令、模擬所述第一虛

擬機器在所述多個儲存層中的性能的第一性能資料，以及模擬所述第二虛擬機器在所述多個儲存層中的性能的第二性能資料；

傳輸器，將對於所述第一 I/O 命令的第一回應傳輸到所述第一虛擬機器並且將對於所述第二 I/O 命令的第二回應傳輸到所述第二虛擬機器；以及

自動分層控制器，選擇第一儲存層以儲存所述第一虛擬機器的所述第一資料、選擇第二儲存層以儲存所述第二虛擬機器的所述第二資料，並且回應於所述第一性能資料和所述第二性能資料來遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個，

其中所述自動分層控制器可操作以選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料並且選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料來優化所有虛擬機器在所述多個儲存層中的性能，

其中所述自動分層控制器可操作以將由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所引起的性能改變和遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個的遷移成本作為考慮因素。

【0122】 陳述 2. 本發明概念的實施例包含根據陳述 1 所述的系

統，其中：

所述系統進一步包括處理器；以及

所述自動分層控制器包含在所述處理器上運行的軟體。

【0123】 陳述 3. 本發明概念的實施例包含根據陳述 1 所述的系統，其中所述自動分層控制器包含：

層選擇模組，回應於所述第一性能資料和所述第二性能資料選擇所述第一虛擬機器的所述第一儲存層和所述第二虛擬機器的所述第二儲存層；以及

遷移模組，遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個。

【0124】 陳述 4. 本發明概念的實施例包含根據陳述 1 所述的系統，其中所述自動分層控制器包含使用所述第一性能資料和所述第二性能資料生成資源矩陣的資源矩陣生成器，所述資源矩陣使所述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯。

【0125】 陳述 5. 本發明概念的實施例包含根據陳述 4 所述的系統，其中所述自動分層控制器可操作以回應於所述資源矩陣來選擇第一儲存層以儲存所述第一虛擬機器的所述第一資料並且選擇第二儲存層以儲存所述第二虛擬機器的所述第二資料。

【0126】 陳述 6. 本發明概念的實施例包含根據陳述 5 所述的系統，其中所述自動分層控制器可操作以回應於所述多個儲存層中

的每一個的所述資源矩陣和專業化資訊來選擇第一儲存層意儲存所述第一虛擬機器的所述第一資料並且選擇第二儲存層以儲存所述第二虛擬機器的所述第二資料。

【0127】 陳述 7. 本發明概念的實施例包含根據陳述 6 所述的系統，其中所述多個儲存層中的每一個的所述專業化資訊確認所述多個儲存層中的每一個是否支持所述多個資源中的每一個。

【0128】 陳述 8. 本發明概念的實施例包含根據陳述 7 所述的系統，其中所述多個資源包含輸送量、頻寬、容量和寫入放大。

【0129】 陳述 9. 本發明概念的實施例包含根據陳述 4 所述的系統，其中所述資源矩陣生成器包含根據所述多個資源中的第二資源將所述第一虛擬機器在所述多個儲存層中的每一個中的第一性能度量轉換成所述第一虛擬機器在所述多個儲存層中的每一個中的第二性能度量的度量轉換器。

【0130】 陳述 10. 本發明概念的實施例包含根據陳述 9 所述的系統，其中所述第一性能度量包含所述第一虛擬機器的第一平均 I/O 延遲資料。

【0131】 陳述 11. 本發明概念的實施例包含根據陳述 1 所述的系統，其中：

所述自動分層控制器包含對所述第一性能資料執行第一線性回歸並且對所述第二性能資料執行第二線性回歸的線性回歸模組；以及

所述自動分層控制器可操作以回應於所述第一線性回歸和

所述第二線性回歸選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料且選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料。

【0132】 陳述 12. 本發明概念的實施例包含根據陳述 11 所述的系統，其中所述自動分層控制器可操作以回應於所述第一線性回歸和所述第二線性回歸，使用所述第一虛擬機器和所述第二虛擬機器的估計平均 I/O 延遲來選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料且選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料。

【0133】 陳述 13. 本發明概念的實施例包含根據陳述 1 所述的系統，其中：

所述第一性能資料包含使用第一資源的第一度量性能資料；以及

所述自動分層控制器包含將使用所述第一資源的所述第一度量性能資料轉換成使用第二資源的第二度量性能資料的度量轉換器。

【0134】 陳述 14. 本發明概念的實施例包含根據陳述 1 所述的系統，其中所述自動分層控制器可操作以選擇第一儲存層來儲存所述第一虛擬機器的所述第一資料、選擇第二儲存層來儲存所述第二虛擬機器的所述第二資料、並且每個遷移期一次回應於所述第一性能資料和所述第二性能資料移轉所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到

所述第二儲存層中的至少一個。

【0135】 陳述 15. 本發明概念的實施例包含根據陳述 14 所述的系統，其中：

所述第一性能資料包含來自所述遷移期內的多個監測期的第一性能資料；以及

所述第二性能資料包含來自所述遷移期內的所述多個監測期的第二性能資料。

【0136】 陳述 16. 本發明概念的實施例包含根據陳述 1 所述的系統，所述系統進一步包括：

第一虛擬機器伺服器，儲存所述第一虛擬機器；以及

第二虛擬機器伺服器，儲存所述第二虛擬機器。

【0137】 陳述 17. 本發明概念的實施例包含根據陳述 16 所述的系統，其中所述第二虛擬機器伺服器是所述第一虛擬機器伺服器。

【0138】 陳述 18. 本發明概念的實施例包含根據陳述 16 所述的系統，其中所述第一虛擬機器伺服器包含產生模擬所述第一虛擬機器在所述多個儲存層中的性能的所述第一性能資料的 I/O 濾波器。

【0139】 陳述 19. 本發明概念的實施例包含根據陳述 18 所述的系統，其中所述 I/O 濾波器可操作以將延遲增加到來自所述第一虛擬機器的多個 I/O 命令、包含所述第一 I/O 命令的多個 I/O 命令中，並且回應於所述增加的延遲確定所述第一虛擬機器的平均 I/O 延遲。

【0140】 陳述 20. 本發明概念的實施例包含根據陳述 19 所述的系統，其中：

所述 I/O 濾波器進一步可操作以將多個延遲增加到來自所述第一虛擬機器的多個 I/O 命令中，並且回應於所述增加的延遲確定所述第一虛擬機器的多個平均 I/O 延遲；以及

所述自動分層控制器包含對所述第一虛擬機器的所述多個平均 I/O 延遲執行第一線性回歸的線性回歸模組。

【0141】 陳述 21. 本發明概念的實施例包含根據陳述 18 所述的系統，其中所述第一虛擬機器伺服器進一步包含從所述 I/O 濾波器收集多個虛擬機器的性能資料並且將所彙集的性能資料傳輸到所述自動分層控制器的自動分層幕後程式。

【0142】 陳述 22. 本發明概念的實施例包含根據陳述 16 所述的系統，其中：

所述第一虛擬機器伺服器不包含儲存所述第一虛擬機器的第一資料中的任一種的第一本地快取記憶體；以及

所述第二虛擬機器伺服器不包含儲存所述第二虛擬機器的第二資料中的任一種的第二本地快取記憶體。

【0143】 陳述 23. 本發明概念的實施例包含根據陳述 1 所述的系統，其中由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所造成的性能改變包含所述第一虛擬機器的性能改進。

【0144】 陳述 24. 本發明概念的實施例包含根據陳述 1 所述的系統，其中由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所造成的性能改變包含所述第一虛擬機器的性能下降。

【0145】 陳述 25. 本發明概念的實施例包含根據陳述 1 所述的系統，其中遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個的遷移成本包含遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所需要的遷移時間。

【0146】 陳述 26. 本發明概念的實施例包含根據陳述 1 所述的系統，其中所述第二儲存層是所述第一儲存層。

【0147】 陳述 27. 本發明概念的實施例包含根據陳述 1 所述的系統，其中所述多個儲存層中的一個中的多個儲存裝置是均質的。

【0148】 陳述 28. 本發明概念的實施例包含一種方法，所述方法包括：

接收模擬第一虛擬機器在多個儲存層中的性能的第一性能資料；

接收模擬第二虛擬機器在所述多個儲存層中的性能的第二性能資料；

使用所述第一性能資料和所述第二性能資料生成使所述第

一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯的資源矩陣；

使用所述資源矩陣選擇所述多個儲存層中的第一儲存層以儲存所述第一虛擬機器的第一資料；

使用所述資源矩陣選擇所述多個儲存層中的第二儲存層以儲存所述第二虛擬機器的第二資料；以及

遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個，

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料使所有虛擬機器在所述多個儲存層中的性能優化，

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料將由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所引起的性能改變作為考慮因素，以及

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料將遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個的遷移成本作為考慮因素。

【0149】 陳述 29. 本發明概念的實施例包含根據陳述 28 所述的方法，其中所述第二儲存層是所述第一儲存層。

【0150】 陳述 30. 本發明概念的實施例包含根據陳述 28 所述的方法，其中接收第一性能資料、接收第二性能資料、生成資源矩陣、選擇第一儲存層和選擇第二儲存層是每遷移期執行一次。

【0151】 陳述 31. 本發明概念的實施例包含根據陳述 30 所述的方法，其中接收第一性能資料和接收第二性能資料是每監測期執行一次。

【0152】 陳述 32. 本發明概念的實施例包含根據陳述 31 所述的方法，其中所述遷移期包含多個監測期。

【0153】 陳述 33. 本發明概念的實施例包含根據陳述 28 所述的方法，其中：

使用所述資源矩陣選擇所述多個儲存層中的第一儲存層以儲存所述第一虛擬機器的第一資料包含使用所述資源矩陣和所述多個儲存層中的每一個的專業化資訊選擇所述多個儲存層中的所述第一儲存層以儲存所述第一虛擬機器的第一資料；以及

使用所述資源矩陣選擇所述多個儲存層中的第二儲存層以儲存所述第二虛擬機器的第二資料包含使用所述資源矩陣和所述多個儲存層中的每一個的所述專業化資訊選擇所述多個儲存層中的所述第二儲存層以儲存所述第二虛擬機器的第二資料。

【0154】 陳述 34. 本發明概念的實施例包含根據陳述 33 所述的方法，其中所述多個儲存層中的每一個的所述專業化資訊確認所

述多個儲存層中的每一個是否支持所述多個資源中的每一個。

【0155】 陳述 35. 本發明概念的實施例包含根據陳述 34 所述的方法，其中所述多個資源包含輸送量、頻寬、容量和寫入放大。

【0156】 陳述 36. 本發明概念的實施例包含根據陳述 34 所述的方法，其中模擬所述第一虛擬機器在所述多個儲存層中的性能的所述第一性能資料包含使所述第一虛擬機器的平均輸入/輸出 (I/O) 延遲與增加的延遲相關聯的多個數據點。

【0157】 陳述 37. 本發明概念的實施例包含根據陳述 36 所述的方法，其中使用所述第一性能資料和所述第二性能資料生成使所述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯的資源矩陣包含對使所述第一虛擬機器的平均 I/O 延遲與增加的延遲相關聯的所述多個數據點執行線性回歸。

【0158】 陳述 38. 本發明概念的實施例包含根據陳述 37 所述的方法，其中使用所述第一性能資料和所述第二性能資料生成使所述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯的資源矩陣進一步包含使用所述線性回歸估計所述第一虛擬機器在所述多個儲存層中的每一個中的性能。

【0159】 陳述 39. 本發明概念的實施例包含根據陳述 37 所述的方法，其中使用所述第一性能資料和所述第二性能資料生成使所述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一

個和所述多個儲存層中的每一個的多個資源相關聯的資源矩陣進一步包含根據所述多個資源中的每一個確定所述第一虛擬機器在所述多個儲存層中的每一個中的性能度量。

【0160】 陳述 40. 本發明概念的實施例包含根據陳述 39 所述的方法，其中根據所述多個資源中的每一個確定所述第一虛擬機器在所述多個儲存層中的每一個中的性能度量包含將根據所述多個資源中的第一資源的所述第一虛擬機器在所述多個儲存層中的每一個中的第一性能度量轉換成根據所述多個資源中的第二資源的所述第一虛擬機器在所述多個儲存層中的每一個中的第二性能度量。

【0161】 陳述 41. 本發明概念的實施例包含根據陳述 28 所述的方法，其中接收模擬第一虛擬機器在多個儲存層中的性能的第一性能資料包含：

在 I/O 濾波器處接收來自所述虛擬機器的多個第一 I/O 命令；

將第一延遲增加到所述多個第一 I/O 命令中；以及

針對所述多個第一 I/O 命令計算所述虛擬機器的第一平均 I/O 延遲。

【0162】 陳述 42. 本發明概念的實施例包含根據陳述 41 所述的方法，其中：

將第一延遲增加到所述多個第一 I/O 命令中包含將多個延遲增加到所述多個第一 I/O 命令中；

針對所述多個第一 I/O 命令計算所述虛擬機器的第一平均 I/O 延遲包含針對所述多個第一 I/O 命令計算所述虛擬機器的多個平均 I/O 延遲；以及

所述方法進一步包括對所述多個平均 I/O 延遲執行線性回歸。

【0163】 陳述 43. 本發明概念的實施例包含根據陳述 28 所述的方法，其中由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所造成的性能改變包含所述第一虛擬機器的性能改進。

【0164】 陳述 44. 本發明概念的實施例包含根據陳述 28 所述的方法，其中由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所造成的性能改變包含所述第一虛擬機器的性能下降。

【0165】 陳述 45. 本發明概念的實施例包含根據陳述 28 所述的方法，其中遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個的遷移成本包含遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所需要的遷移時間。

【0166】 陳述 46. 本發明概念的實施例包含一種物件，包括非暫

時性儲存媒體，所述非暫時性儲存媒體具有儲存於其上的在由機器執行時引起以下操作的指令：

接收模擬第一虛擬機器在多個儲存層中的性能的第一性能資料；

接收模擬第二虛擬機器在所述多個儲存層中的性能的第二性能資料；

使用所述第一性能資料和所述第二性能資料生成使所述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯的資源矩陣；

使用所述資源矩陣選擇所述多個儲存層中的第一儲存層以儲存所述第一虛擬機器的第一資料；

使用所述資源矩陣選擇所述多個儲存層中的第二儲存層以儲存所述第二虛擬機器的第二資料；以及

遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個，

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料使所有虛擬機器在所述多個儲存層中的性能優化，

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料將由遷移所述第一虛擬機器的所述第一資料到所述第一

儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所引起的性能改變作為考慮因素，以及

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料將遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個的遷移成本作為考慮因素。

【0167】 陳述 47. 本發明概念的實施例包含根據陳述 46 所述的物件，其中所述第二儲存層是所述第一儲存層。

【0168】 陳述 48. 本發明概念的實施例包含根據陳述 46 所述的物件，其中接收第一性能資料、接收第二性能資料、生成資源矩陣、選擇第一儲存層和選擇第二儲存層是每遷移期執行一次。

【0169】 陳述 49. 本發明概念的實施例包含根據陳述 48 所述的物件，其中接收第一性能資料和接收第二性能資料是每監測期執行一次。

【0170】 陳述 50. 本發明概念的實施例包含根據陳述 49 所述的物件，其中所述遷移期包含多個監測期。

【0171】 陳述 51. 本發明概念的實施例包含根據陳述 46 所述的物件，其中：

使用所述資源矩陣選擇所述多個儲存層中的第一儲存層以儲存所述第一虛擬機器的第一資料包含使用所述資源矩陣和所述多個儲存層中的每一個的專業化資訊選擇所述多個儲存層中的所

述第一儲存層以儲存所述第一虛擬機器的第一資料；以及

使用所述資源矩陣選擇所述多個儲存層中的第二儲存層以儲存所述第二虛擬機器的第二資料包含使用所述資源矩陣和所述多個儲存層中的每一個的所述專業化資訊選擇所述多個儲存層中的所述第二儲存層以儲存所述第二虛擬機器的第二資料。

【0172】 陳述 52. 本發明概念的實施例包含根據陳述 51 所述的物件，其中所述多個儲存層中的每一個的所述專業化資訊確認所述多個儲存層中的每一個是否支持所述多個資源中的每一個。

【0173】 陳述 53. 本發明概念的實施例包含根據陳述 52 所述的物件，其中所述多個資源包含輸送量、頻寬、容量和寫入放大。

【0174】 陳述 54. 本發明概念的實施例包含根據陳述 52 所述的物件，其中模擬所述第一虛擬機器在所述多個儲存層中的性能的所述第一性能資料包含使所述第一虛擬機器的平均輸入/輸出 (I/O) 延遲與增加的延遲相關聯的多個數據點。

【0175】 陳述 55. 本發明概念的實施例包含根據陳述 54 所述的物件，其中使用所述第一性能資料和所述第二性能資料生成使所述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯的資源矩陣包含對使所述第一虛擬機器的平均 I/O 延遲與增加的延遲相關聯的所述多個數據點執行線性回歸。

【0176】 陳述 56. 本發明概念的實施例包含根據陳述 55 所述的物件，其中使用所述第一性能資料和所述第二性能資料生成使所

述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯的資源矩陣進一步包含使用所述線性回歸估計所述第一虛擬機器在所述多個儲存層中的每一個中的性能。

【0177】 陳述 57. 本發明概念的實施例包含根據陳述 55 所述的物件，其中使用所述第一性能資料和所述第二性能資料生成使所述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯的資源矩陣進一步包含根據所述多個資源中的每一個確定所述第一虛擬機器在所述多個儲存層中的每一個中的性能度量。

【0178】 陳述 58. 本發明概念的實施例包含根據陳述 57 所述的物件，其中根據所述多個資源中的每一個確定所述第一虛擬機器在所述多個儲存層中的每一個中的性能度量包含將根據所述多個資源中的第一資源的所述第一虛擬機器在所述多個儲存層中的每一個中的第一性能度量轉換成根據所述多個資源中的第二資源的所述第一虛擬機器在所述多個儲存層中的每一個中的第二性能度量。

【0179】 陳述 59. 本發明概念的實施例包含根據陳述 46 所述的物件，其中接收模擬第一虛擬機器在多個儲存層中的性能的第一性能資料包含：

在 I/O 濾波器處接收來自所述虛擬機器的第一 I/O 命令；

將第一延遲增加到所述第一 I/O 命令中；以及

針對所述第一 I/O 命令計算所述虛擬機器的第一平均 I/O 延遲。

【0180】 陳述 60. 本發明概念的實施例包含根據陳述 59 所述的物件，其中：

將第一延遲增加到所述多個第一 I/O 命令中包含將多個延遲增加到所述多個第一 I/O 命令中；

針對所述多個第一 I/O 命令計算所述虛擬機器的第一平均 I/O 延遲包含針對所述多個第一 I/O 命令計算所述虛擬機器的多個平均 I/O 延遲；以及

所述非暫時性儲存媒體具有儲存於其上的當由機器執行時引起執行對所述多個平均 I/O 延遲的線性回歸的其它指令。

【0181】 陳述 61. 本發明概念的實施例包含根據陳述 46 所述的物件，其中由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所造成的性能改變包含所述第一虛擬機器的性能改進。

【0182】 陳述 62. 本發明概念的實施例包含根據陳述 46 所述的物件，其中由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所造成的性能改變包含所述第一虛擬機器的性能下降。

【0183】 陳述 63. 本發明概念的實施例包含根據陳述 46 所述的

物件，其中遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個的遷移成本包含遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所需要的遷移時間。

【0184】 因此，考慮到對本文所描述的實施例的各種置換，此詳細描述和隨附材料預期僅為說明性的，並且不應被視為限制本發明概念的範圍。因此，本發明概念所要求保護的是落入所附申請專利範圍以及其等效物的範圍和精神內的所有此類修改。

【符號說明】

【0185】

105：系統

110：虛擬機器伺服器

110-1、110-2、110-3：虛擬機器伺服器

115：處理器

115-1、115-2、115-3、115-4：處理器

120：記憶體

120-1、120-2、120-3、120-4：記憶體

125-1、125-2、125-3：快取記憶體

130：資料中心伺服器

135：儲存池

140：接收器

145：傳輸器

150：網路

205：記憶體控制器

210：時鐘

215：網路連接器

220：匯流排

225：使用者介面

230：輸入/輸出引擎

305-1、305-2、305-3：虛擬機器

310：應用程式

315：作業系統

320：管理程式

325：I/O 濾波器

330：自動分層幕後程式

335：自動分層控制器

340-1、340-2、340-3：儲存層

345：儲存裝置

405：資源

410：輸送量

415：頻寬

420：容量

425：寫入放大因數

505：虛擬機器文件

510：虛擬機器文件

515：箭頭

705：I/O 命令

710：延遲

715：回應

720：性能資料

805：線性回歸模組

810：資源矩陣生成器

815：度量轉換器

820：專業化資訊

825：層選擇模組

830：遷移模組

905：資源矩陣

910：二維矩陣

915：二維矩陣

920：二維矩陣

1005：線性回歸

1010：線性回歸

1205：得分矩陣

1305：遷移期

1310：監測期

1405、1410、1415、1418、1420、1425、1430、1505、1510、
1515、1520、1525、1530、1535、1605、1610、1615、1620：塊

1625：虛線

3615：性能資料

T0、T1：時間

【發明申請專利範圍】

【第 1 項】一種資料儲存管理系統，包括：

提供多個資源的多個儲存裝置，所述多個儲存裝置被組織成多個儲存層並且儲存第一虛擬機器的第一資料和第二虛擬機器的第二資料；

接收器，接收來自所述第一虛擬機器的第一輸入/輸出命令、來自所述第二虛擬機器的第二輸入/輸出命令、模擬所述第一虛擬機器在所述多個儲存層中的性能的第一性能資料，以及模擬所述第二虛擬機器在所述多個儲存層中的性能的第二性能資料；

傳輸器，將對於所述第一輸入/輸出命令的第一回應傳輸到所述第一虛擬機器，並且將對於所述第二輸入/輸出命令的第二回應傳輸到所述第二虛擬機器；以及

自動分層控制器，選擇第一儲存層以儲存所述第一虛擬機器的所述第一資料，選擇第二儲存層以儲存所述第二虛擬機器的所述第二資料，並且回應於所述第一性能資料和所述第二性能資料來遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個，

其中所述自動分層控制器能夠操作以選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料並且選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料，從而優化所有虛擬機器在所述多個儲存層中的性能，

其中所述自動分層控制器能夠操作以將由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所引起的性能改變和遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個的遷移成本作為考慮因素。

【第 2 項】如申請專利範圍第 1 項所述的資料儲存管理系統，其中所述自動分層控制器包含：

層選擇模組，以至少一處理器實施，回應於所述第一性能資料和所述第二性能資料選擇所述第一虛擬機器的所述第一儲存層和所述第二虛擬機器的所述第二儲存層；以及

遷移模組，遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個。

【第 3 項】如申請專利範圍第 1 項所述的資料儲存管理系統，其中所述自動分層控制器包含使用所述第一性能資料和所述第二性能資料生成資源矩陣的資源矩陣生成器，以至少一處理器實施，所述資源矩陣使所述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯。

【第 4 項】如申請專利範圍第 3 項所述的資料儲存管理系統，其中所述資源矩陣生成器包含度量轉換器，以至少一處理器實施，其

根據所述多個資源的第二資源將所述第一虛擬機器在所述多個儲存層中的每一個中的第一性能度量轉換成所述第一虛擬機器在所述多個儲存層中的每一個中的第二性能度量。

【第 5 項】如申請專利範圍第 4 項所述的資料儲存管理系統，其中所述第一性能度量包含所述第一虛擬機器的第一平均輸入/輸出延遲資料。

【第 6 項】如申請專利範圍第 1 項所述的資料儲存管理系統，其中：

所述自動分層控制器包含線性回歸模組，其對所述第一性能資料執行第一線性回歸並且對所述第二性能資料執行第二線性回歸；以及

所述自動分層控制器能夠操作以回應於所述第一線性回歸和所述第二線性回歸而選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料且選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料。

【第 7 項】如申請專利範圍第 6 項所述的資料儲存管理系統，其中所述自動分層控制器能夠操作以回應於所述第一線性回歸和所述第二線性回歸，使用所述第一虛擬機器和所述第二虛擬機器的估計平均輸入/輸出延遲來選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料並且選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料。

【第 8 項】如申請專利範圍第 1 項所述的資料儲存管理系統，其中所述自動分層控制器能夠操作以選擇第一儲存層來儲存所述第一

虛擬機器的所述第一資料，選擇第二儲存層來儲存所述第二虛擬機器的所述第二資料，並且每個遷移期一次回應於所述第一性能資料和所述第二性能資料移轉所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個。

【第 9 項】如申請專利範圍第 8 項所述的資料儲存管理系統，其中：

所述第一性能資料包含來自所述遷移期內的多個監測期的第一性能資料；以及

所述第二性能資料包含來自所述遷移期內的所述多個監測期的第二性能資料。

【第 10 項】如申請專利範圍第 1 項所述的資料儲存管理系統，進一步包括：

第一虛擬機器伺服器，儲存所述第一虛擬機器；以及

第二虛擬機器伺服器，儲存所述第二虛擬機器。

【第 11 項】如申請專利範圍第 10 項所述的資料儲存管理系統，其中所述第一虛擬機器伺服器包含生成所述第一性能資料的輸入/輸出濾波器，所述第一性能資料模擬所述第一虛擬機器在所述多個儲存層中的性能。

【第 12 項】如申請專利範圍第 11 項所述的資料儲存管理系統，其中所述輸入/輸出濾波器能夠操作以將延遲增加到來自所述第一虛擬機器的多個輸入/輸出命令中，所述多個輸入/輸出命令包含所述第一輸入/輸出命令，並且回應於所增加的延遲確定所述第一虛擬

機器的平均輸入/輸出延遲。

【第 13 項】一種資料儲存管理方法，包括：

接收模擬第一虛擬機器在多個儲存層中的性能的第一性能資料；

接收模擬第二虛擬機器在所述多個儲存層中的性能的第二性能資料；

使用所述第一性能資料和所述第二性能資料生成資源矩陣，所述資源矩陣使所述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯；

使用所述資源矩陣選擇所述多個儲存層中的第一儲存層以儲存所述第一虛擬機器的第一資料；

使用所述資源矩陣選擇所述多個儲存層中的第二儲存層以儲存所述第二虛擬機器的第二資料；以及

遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個，

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料使所有虛擬機器在所述多個儲存層中的性能優化，

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述

第二資料將由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所引起的性能改變作為考慮因素，以及

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料將遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個的遷移成本作為考慮因素。

【第 14 項】如申請專利範圍第 13 項所述的資料儲存管理方法，其中：

使用所述資源矩陣選擇所述多個儲存層中的第一儲存層以儲存所述第一虛擬機器的第一資料包含使用所述資源矩陣和所述多個儲存層中的每一個的專業化資訊選擇所述多個儲存層中的所述第一儲存層以儲存所述第一虛擬機器的第一資料；以及

使用所述資源矩陣選擇所述多個儲存層中的第二儲存層以儲存所述第二虛擬機器的第二資料包含使用所述資源矩陣和所述多個儲存層中的每一個的所述專業化資訊選擇所述多個儲存層中的所述第二儲存層以儲存所述第二虛擬機器的第二資料。

【第 15 項】如申請專利範圍第 14 項所述的資料儲存管理方法，其中所述多個儲存層中的每一個的所述專業化資訊確認所述多個儲存層中的每一個是否支持所述多個資源中的每一個。

【第 16 項】如申請專利範圍第 15 項所述的資料儲存管理方法，其

中模擬所述第一虛擬機器在所述多個儲存層中的性能的所述第一性能資料包含使所述第一虛擬機器的平均輸入/輸出延遲與所增加的延遲相關聯的多個數據點。

【第 17 項】如申請專利範圍第 13 項所述的資料儲存管理方法，其中接收模擬第一虛擬機器在多個儲存層中的性能的第一性能資料包含：

在輸入/輸出濾波器處接收來自所述虛擬機器的多個第一輸入/輸出命令；

將第一延遲增加到所述多個第一輸入/輸出命令中的每一個中；以及

針對所述多個第一輸入/輸出命令計算所述虛擬機器的第一平均輸入/輸出延遲。

【第 18 項】如申請專利範圍第 13 項所述的資料儲存管理方法，其中遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個的遷移成本包含遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所需要的遷移時間。

【第 19 項】一種資料儲存管理物件，包括非暫時性儲存媒體，所述非暫時性儲存媒體上儲存有指令，所述指令由機器執行時引起以下操作：

接收模擬第一虛擬機器在多個儲存層中的性能的第一性能

資料；

接收模擬第二虛擬機器在所述多個儲存層中的性能的第二性能資料；

使用所述第一性能資料和所述第二性能資料生成資源矩陣，所述資源矩陣使所述第一虛擬機器和所述第二虛擬機器與所述多個儲存層中的每一個和所述多個儲存層中的每一個的多個資源相關聯；

使用所述資源矩陣選擇所述多個儲存層中的第一儲存層以儲存所述第一虛擬機器的第一資料；

使用所述資源矩陣選擇所述多個儲存層中的第二儲存層以儲存所述第二虛擬機器的第二資料；以及

遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的至少一個，

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料使所有虛擬機器在所述多個儲存層中的性能優化，

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述第二資料將由遷移所述第一虛擬機器的所述第一資料到所述第一儲存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的所述至少一個所引起的性能改變作為考慮因素，以及

其中選擇所述第一儲存層以儲存所述第一虛擬機器的所述
第一資料和選擇所述第二儲存層以儲存所述第二虛擬機器的所述
第二資料將遷移所述第一虛擬機器的所述第一資料到所述第一儲
存層或所述第二虛擬機器的所述第二資料到所述第二儲存層中的
至少一個的遷移成本作為考慮因素。

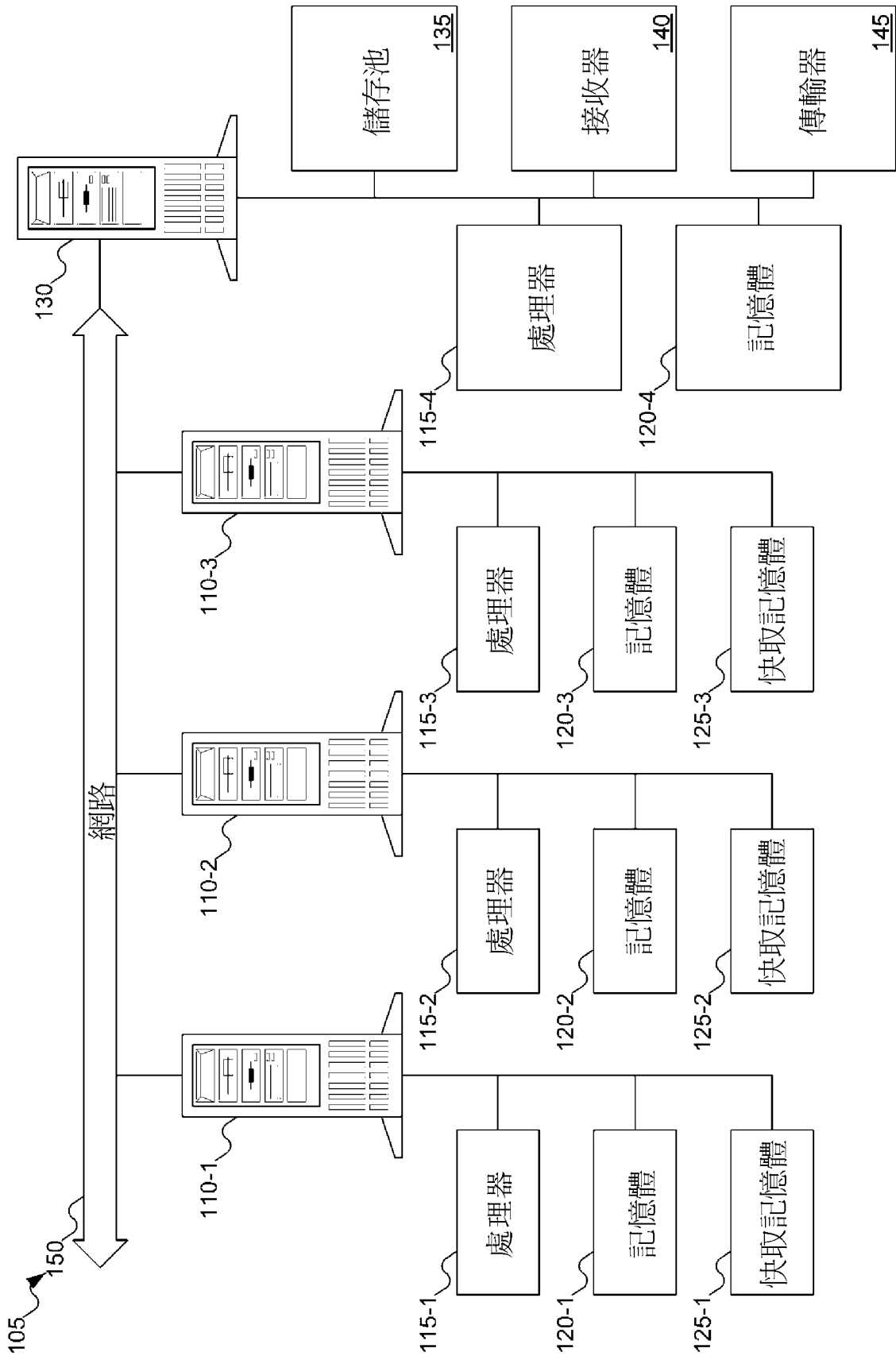
【第 20 項】如申請專利範圍第 19 項所述的資料儲存管理物件，其
中接收模擬第一虛擬機器在多個儲存層中的性能的第一性能資料
包含：

在輸入/輸出濾波器處接收來自所述虛擬機器的第一輸入/
輸出命令；

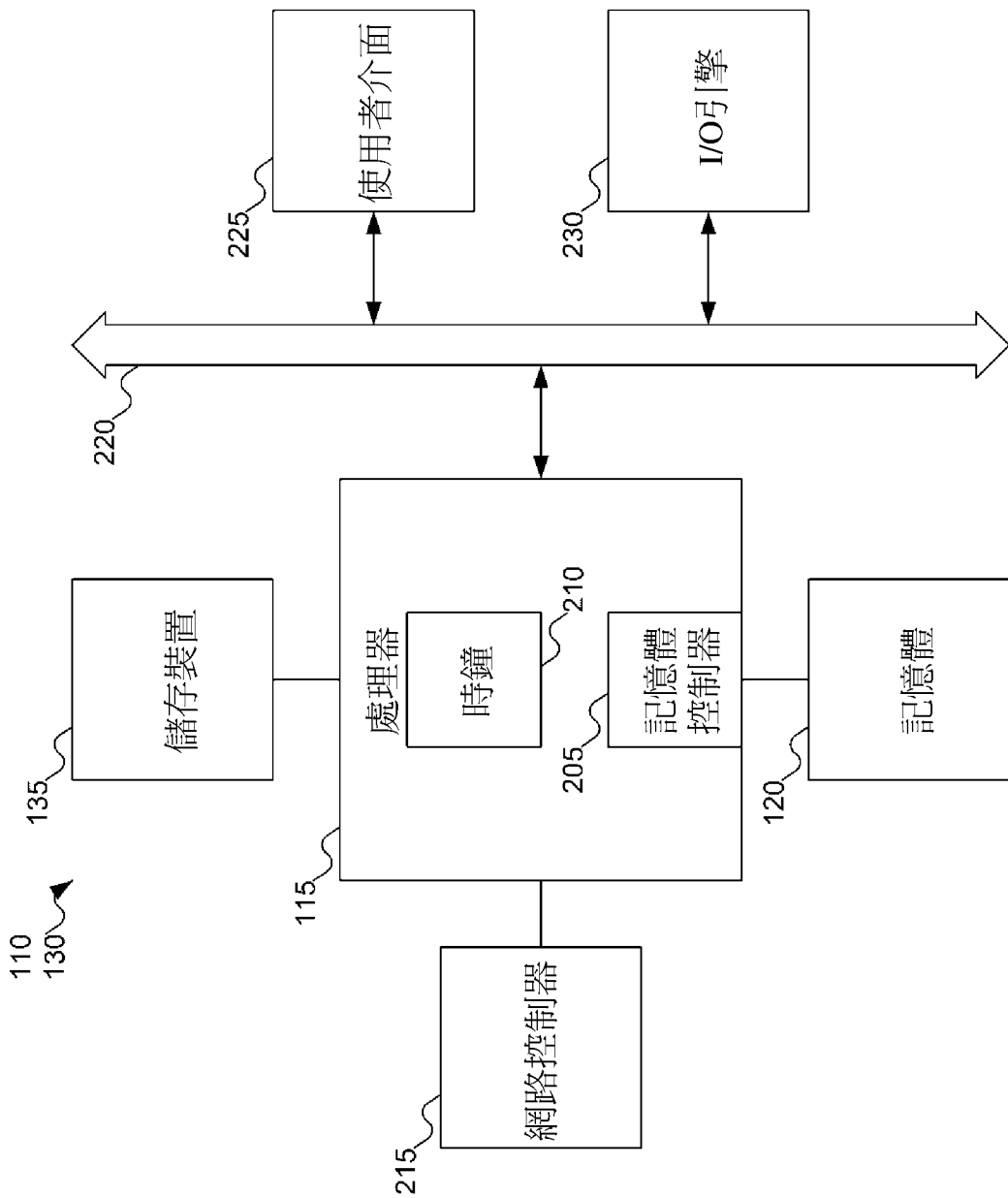
將第一延遲增加到所述第一輸入/輸出命令中；以及

針對所述第一輸入/輸出命令計算所述虛擬機器的第一平
均輸入/輸出延遲。

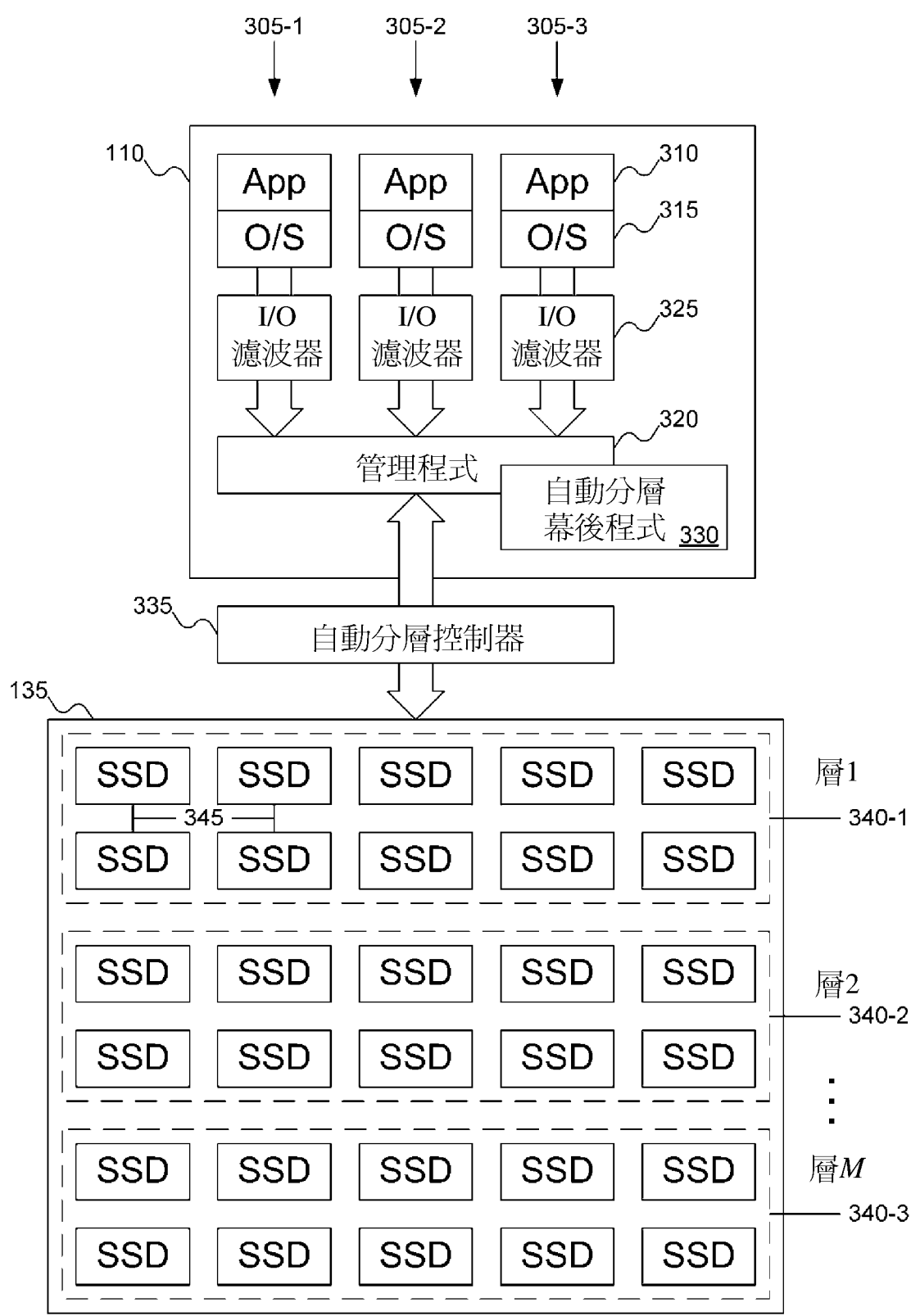
【發明圖式】



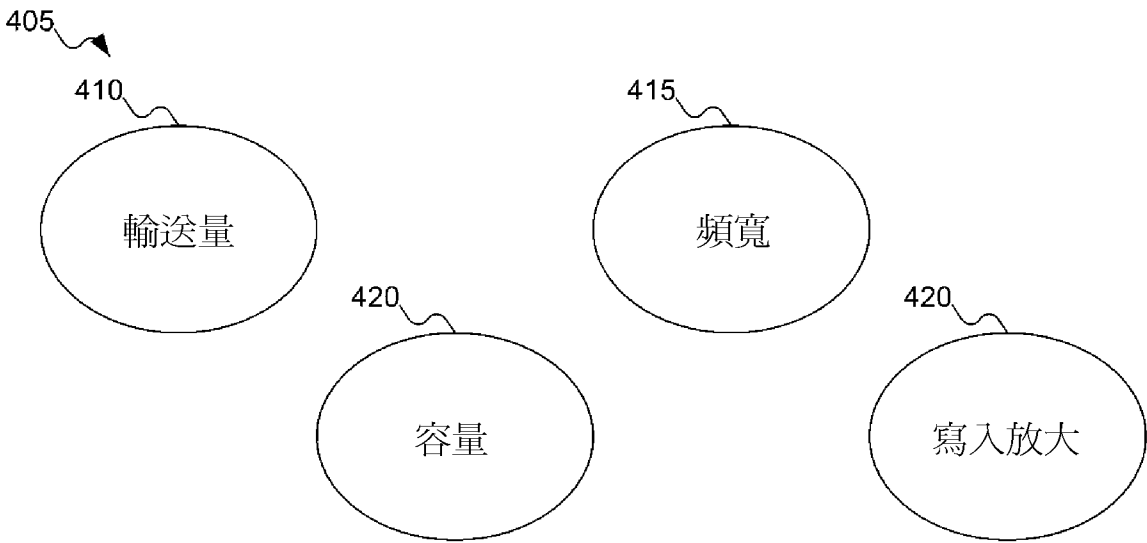
【圖1】



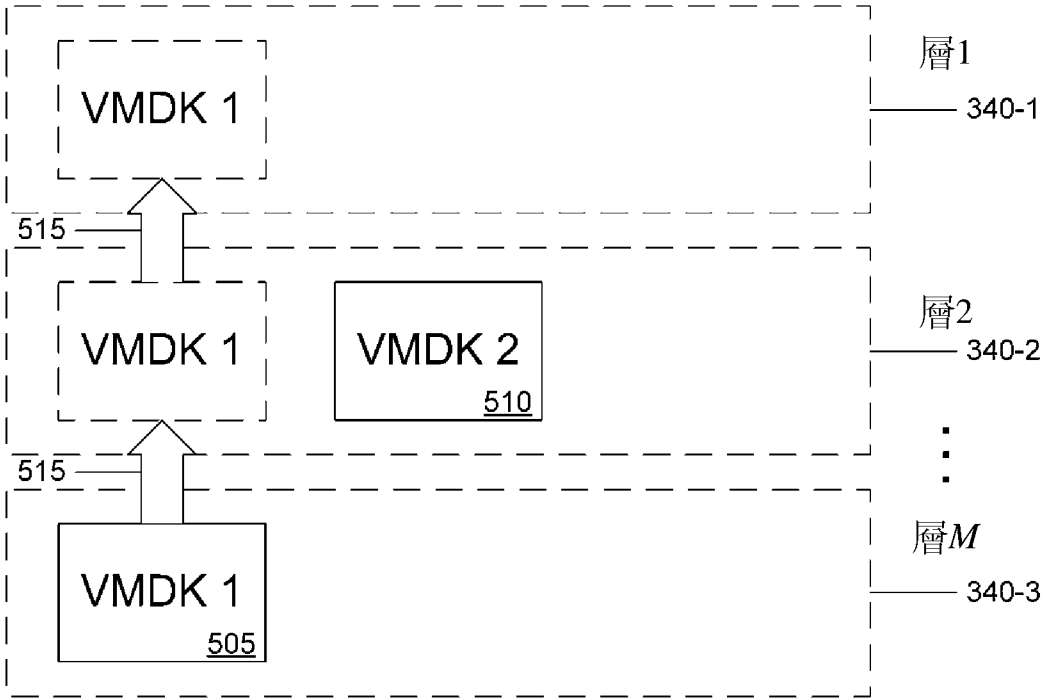
【圖2】



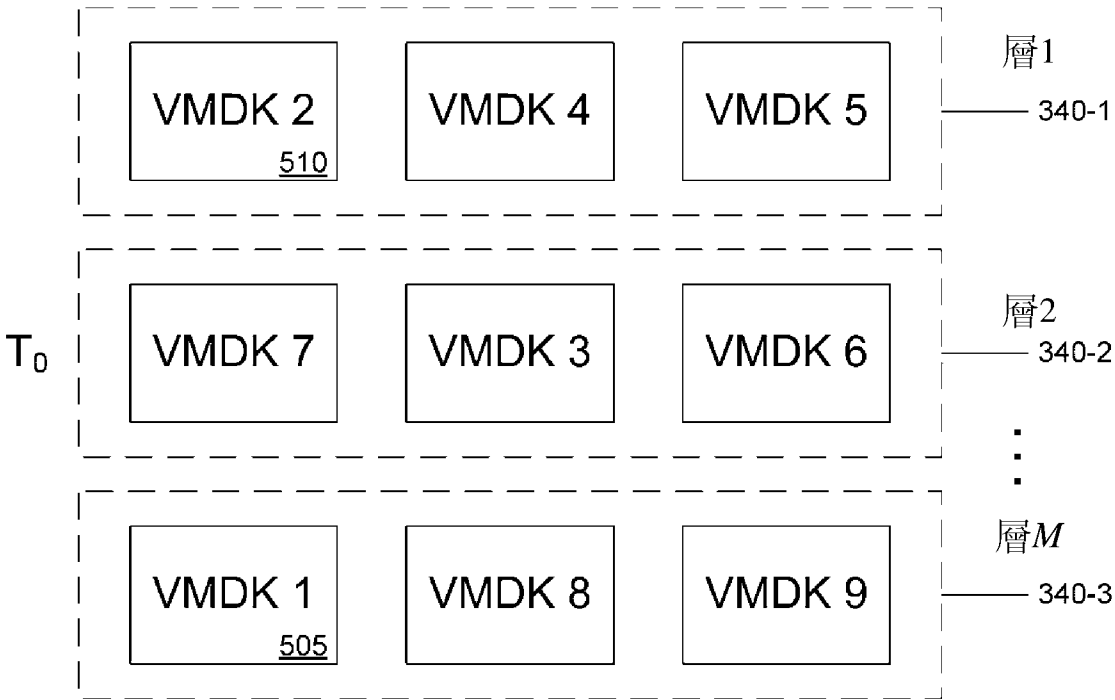
【圖3】



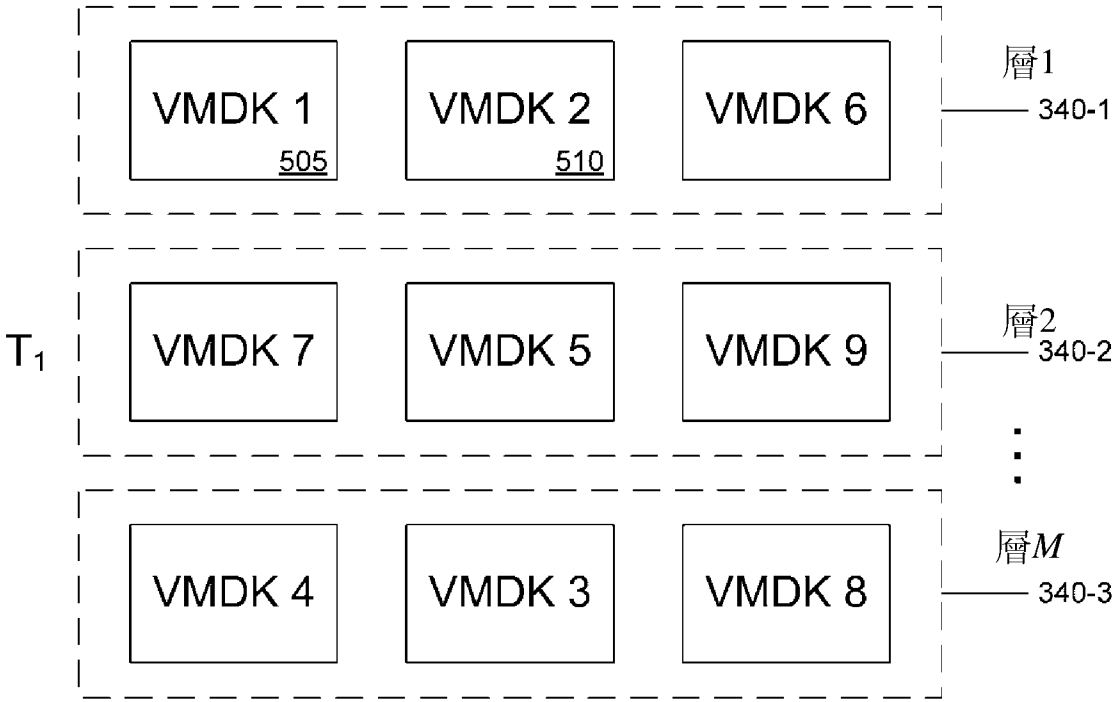
【圖4】



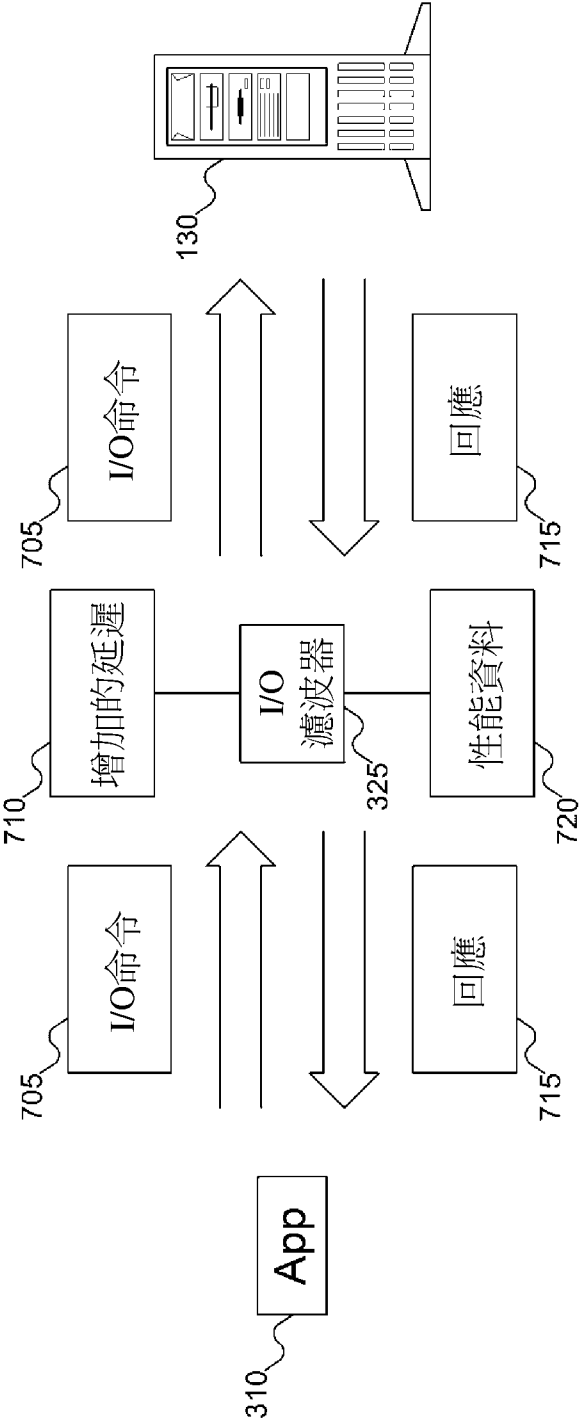
【圖5】



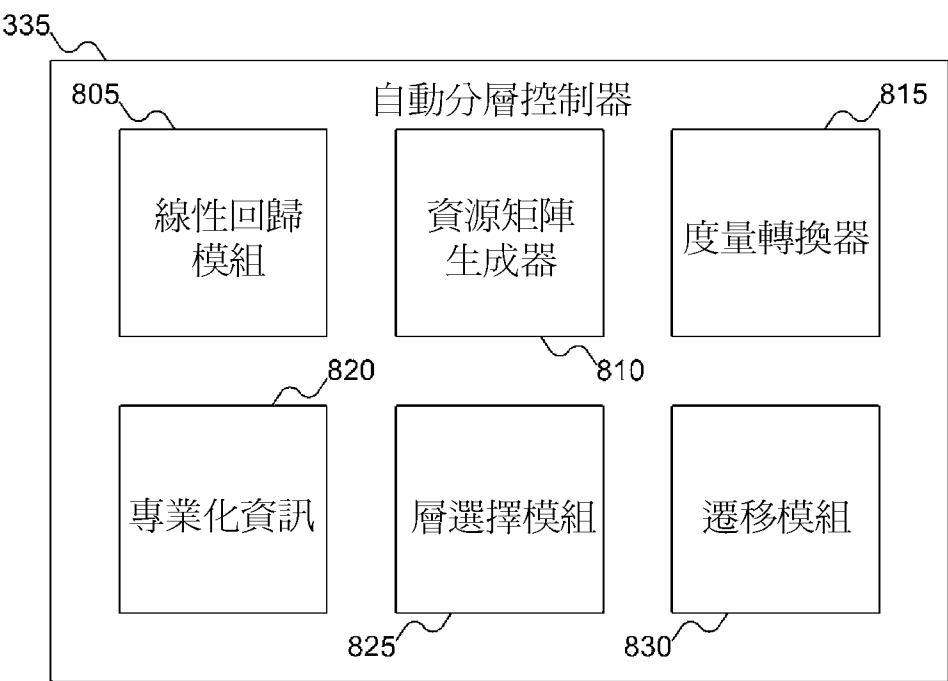
【圖6A】



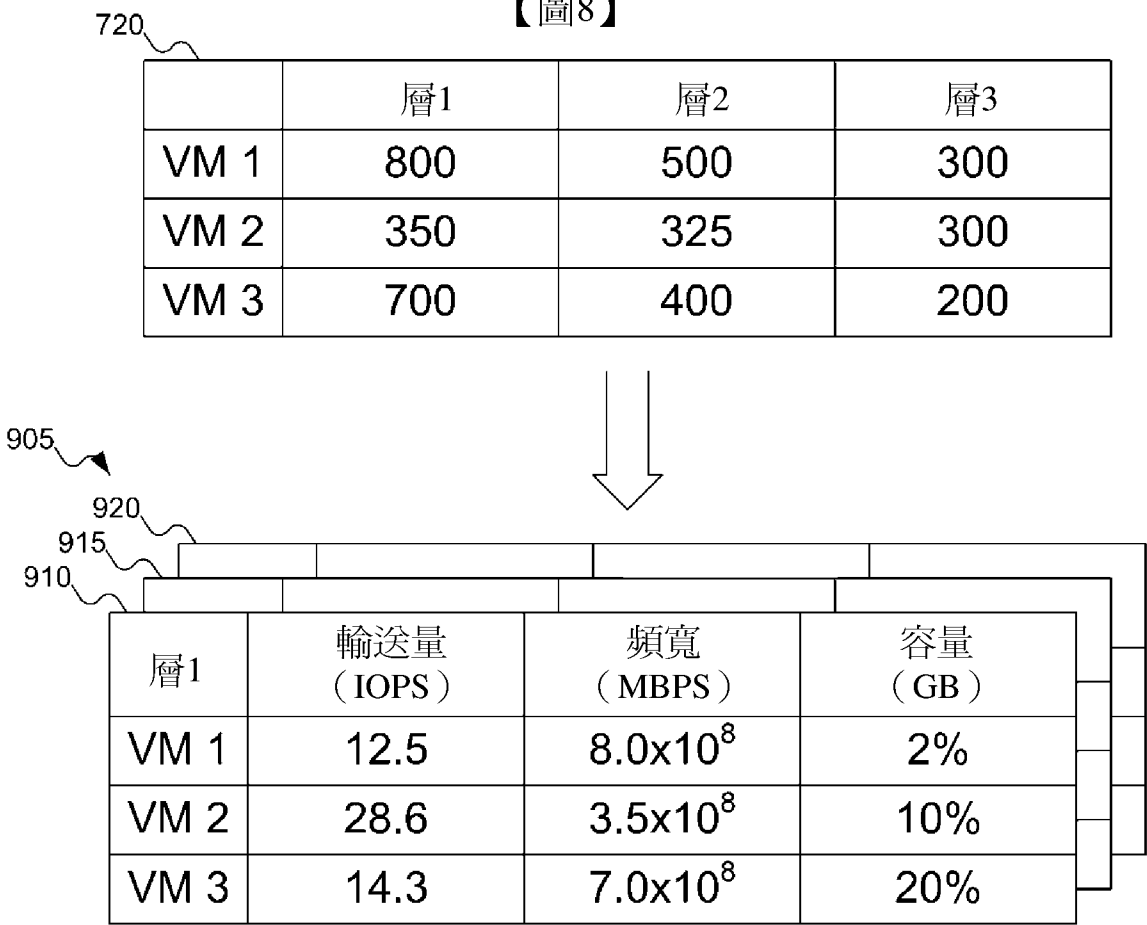
【圖6B】



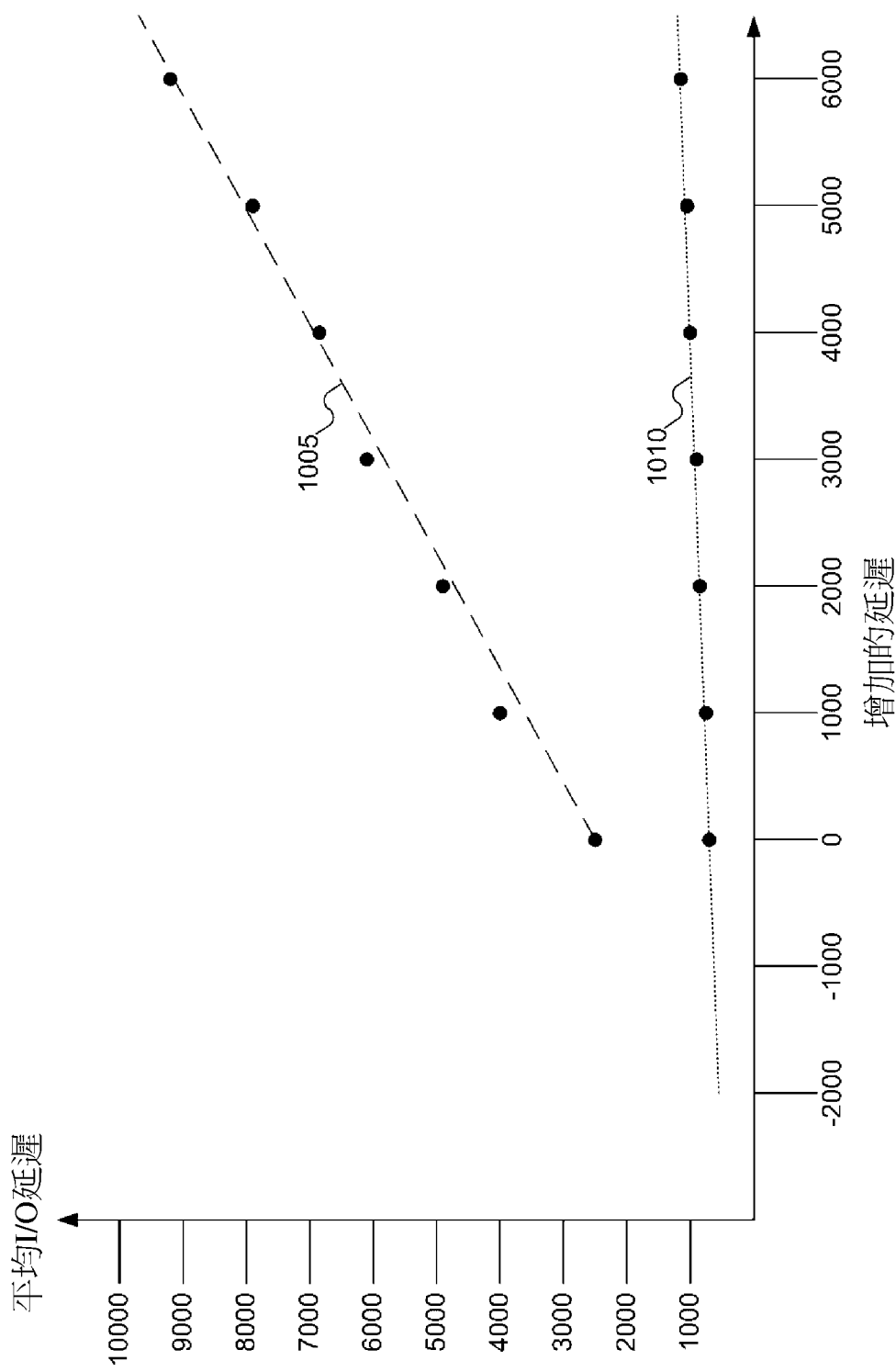
【圖7】



【圖8】



【圖9】

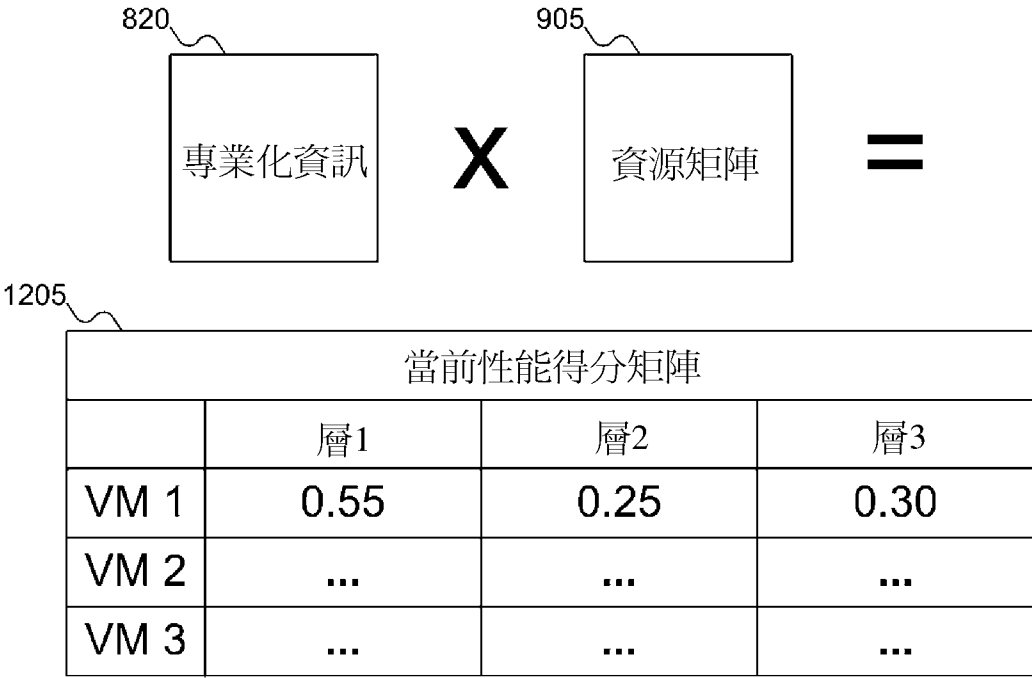


【圖10】

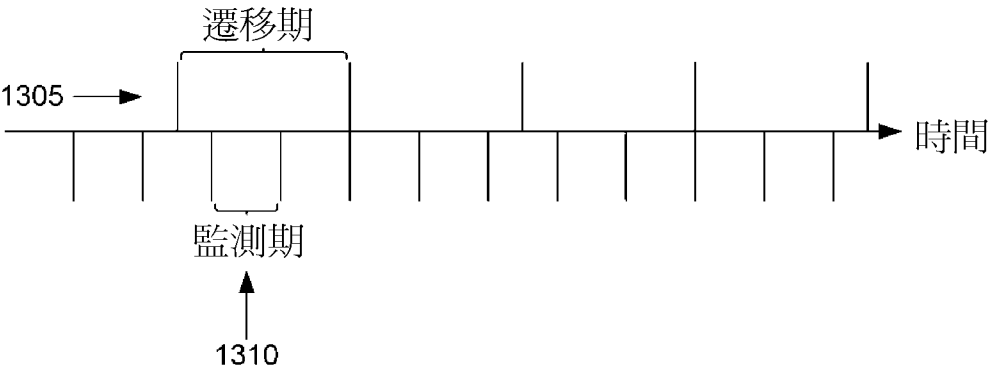
820

專業化資訊			
	輸送量 (IOPS)	頻寬 (MBPS)	儲存大小 (GB)
層1	1	1	0
層2	0	1	0
層3	0	0	1

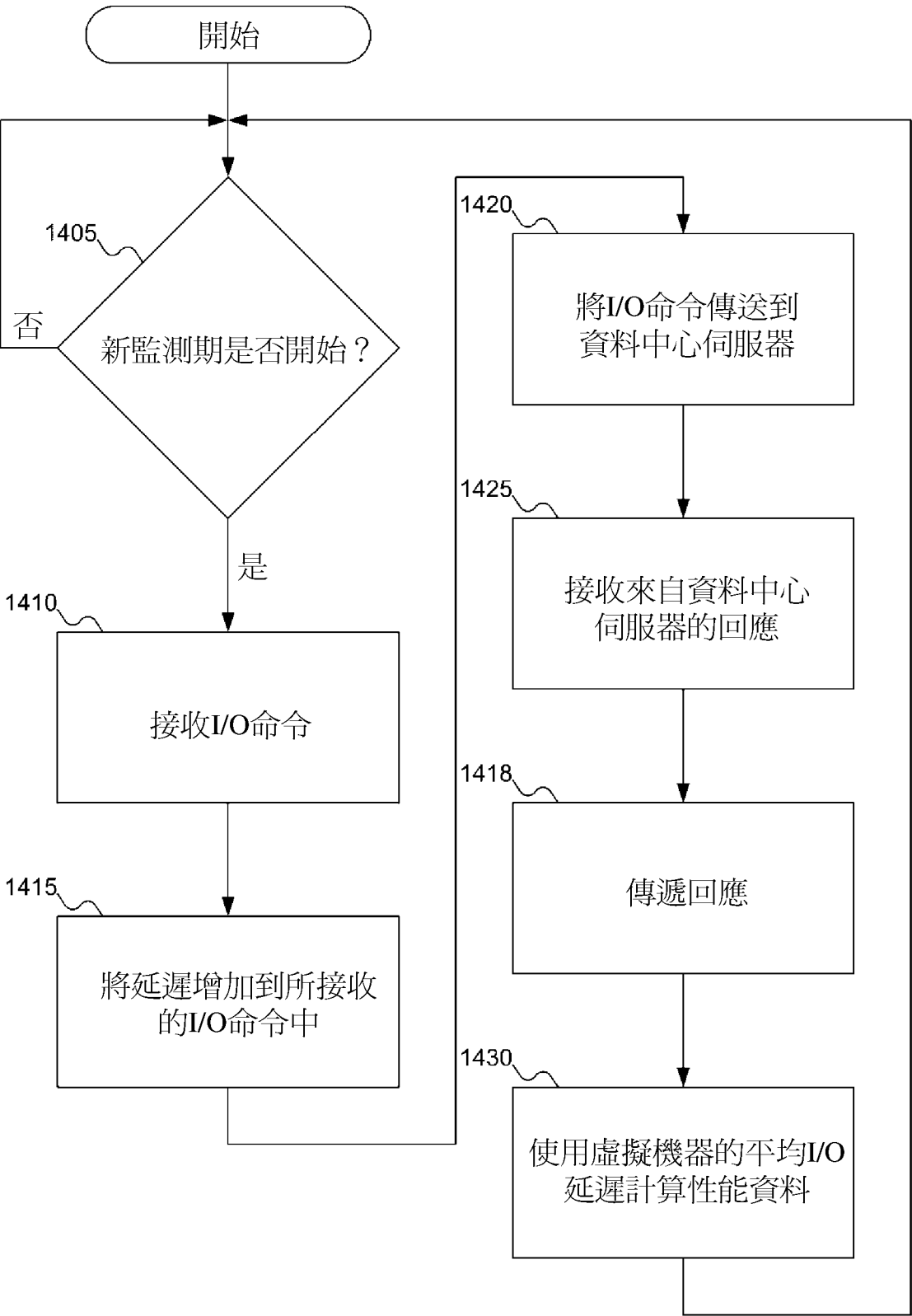
【圖11】



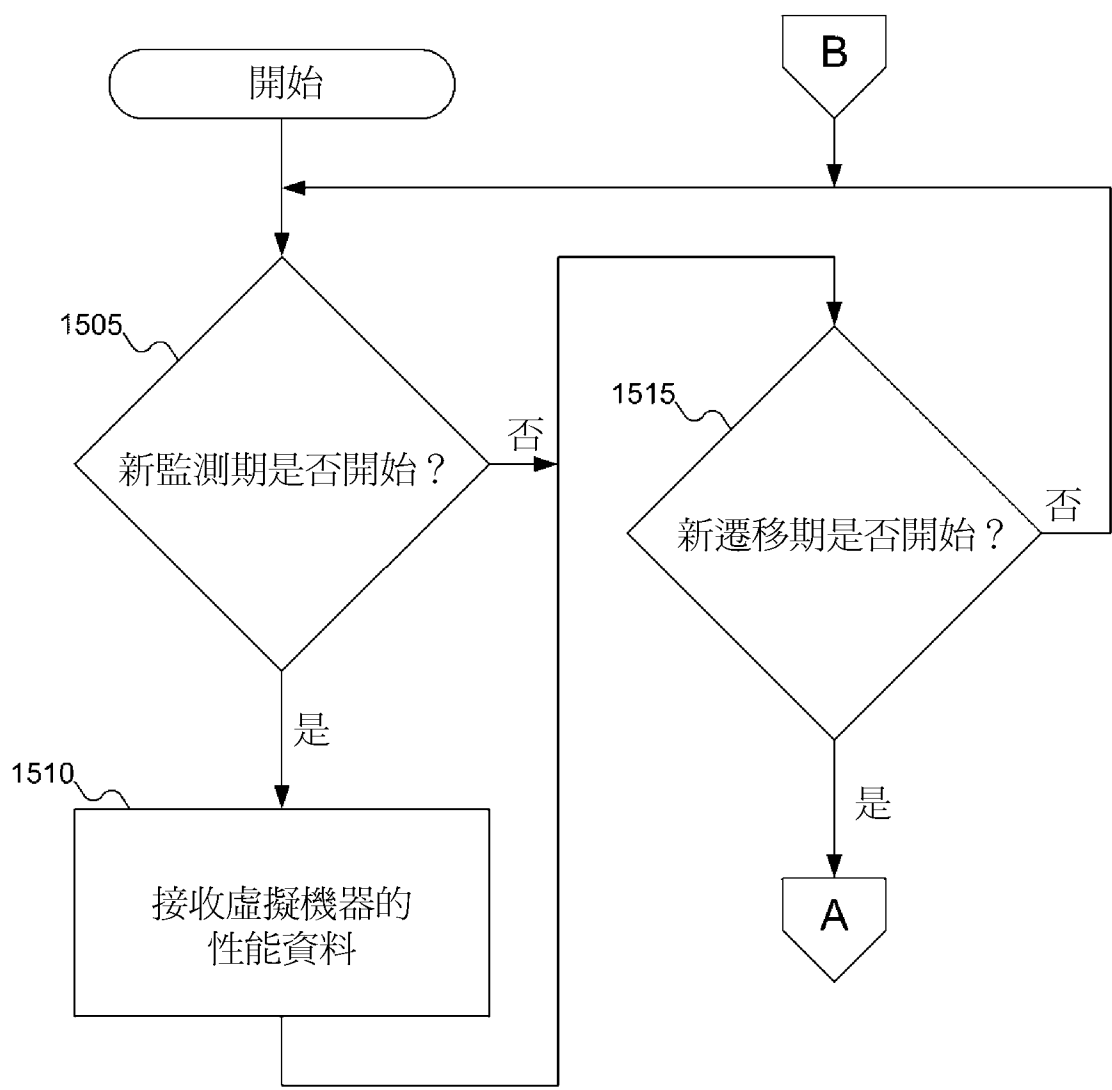
【圖12】



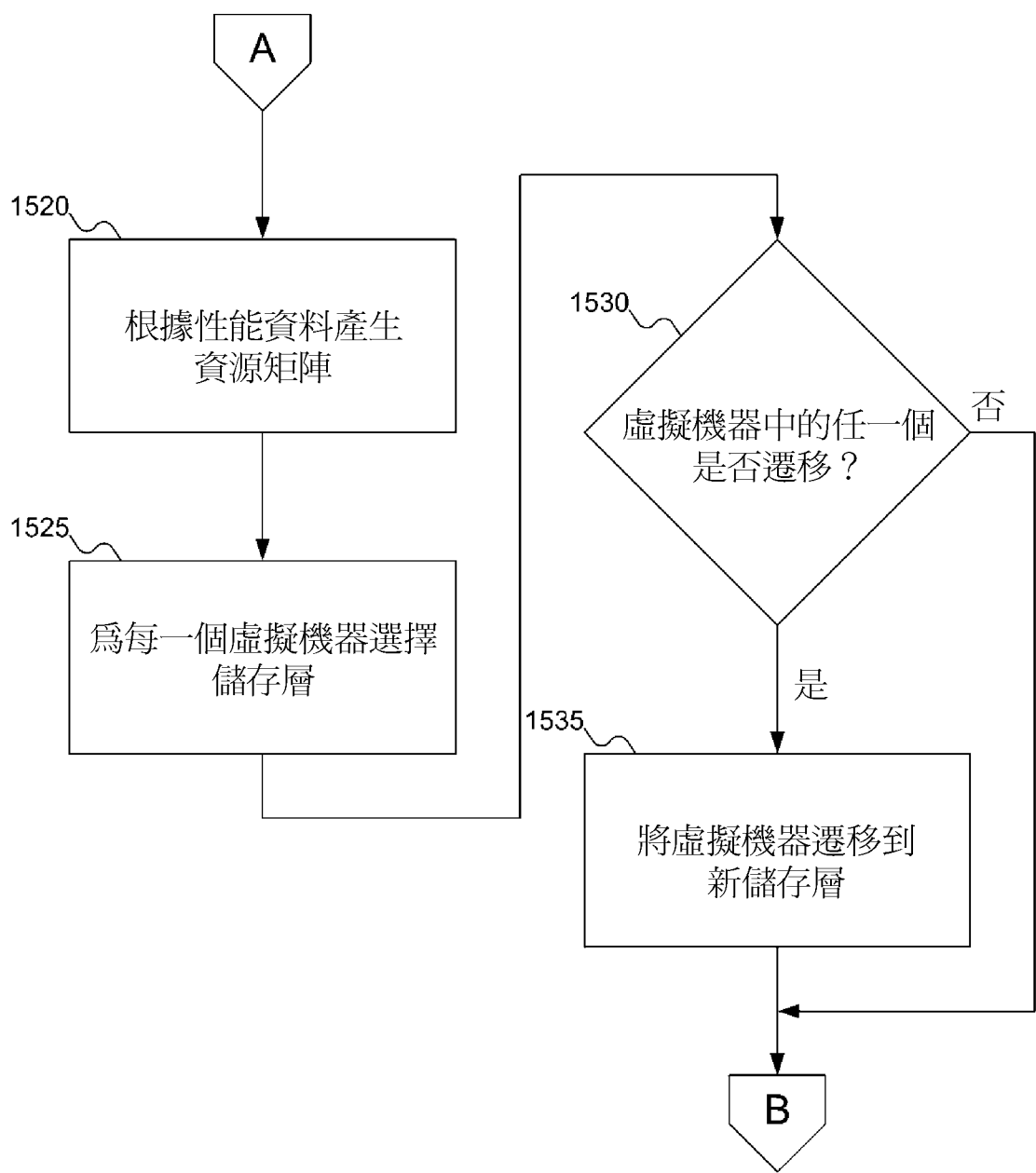
【圖13】



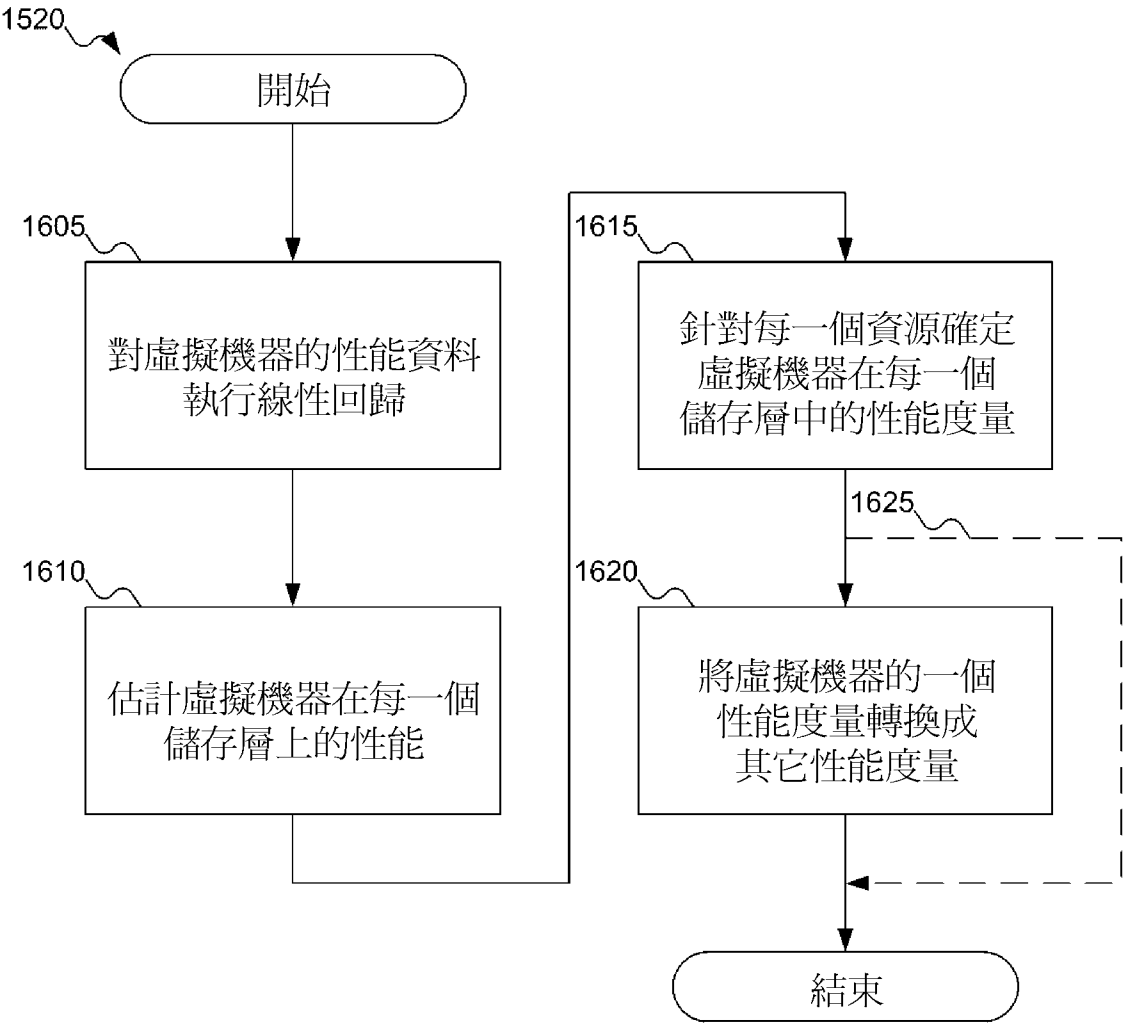
【圖14】



【圖15A】



【圖15B】



【圖16】