視訊點播服務系統中影片配置與負載平衡
Video Placement and Load Balancing for VOD Services

陳郁堂
Yie-Tarng Chen
台灣科技大學電子系
ytchen@et.ntit.edu.tw

林明毅
Ming-Yil Lin
台北商專資管科
linmy@mail.ntcb.edu.tw

摘要

儲存系統是發展大型視訊點播系統(Video On Demand)的瓶頸，困難在於如何達成儲存群組間的負載平衡(Load Balancing)，條列式資料配置(Striping)與熱門影片複製(Replica)為目前常用方式，然而對大型視訊點播並不完全適用。本論文利用改良式 Webster's Monotone Divisor Method[2]決定熱門影片複製(Replica)份數，將較受歡迎的影片複製在不同儲存群組中。再提出負載轉移能力的概念，利用圖形理論(Graph Theory)的分析模式，提出Major Copy Round Robin(*MCRR*)的靜態影片配置方法，並以電腦模擬來進行驗證，結果證實 *MCRR* 靜態影片配置方法，在不同的用戶存取模式下，仍能達到儲存裝置間的負載平衡(Load Balancing)，對提高系統的效能(Performance)，有顯著的效果。

Abstract

Load balancing is an important design issue for large-scale Video-on-Demand (VOD) system. Striping and replica are general approaches to attack this problem. However, the variation of access patterns still incurs load imbalancing . In this paper, we present a video placement scheme for solving this problem.. The replica number of a video stream is decided by modified Webster's Monotone Divisor Method. Based on the concept of load-shifting, we proposed a static video placement scheme, Major Copy Round Robin (MCRR) The MCRR scheme is evaluated by the simulation study. The result reveals that MCRR has an obvious effect ion load balancing.

1 緒 論

近年來，由於高速網路的迅速發展，動態影像壓縮與大容量儲存技術的成熟，帶動多媒體應用的澎湃發展，其中以視訊點播的服務(Video on Demand)，備受矚目。用戶透過高速網路，可隨時從視訊伺服器(Video Server)接收到所希望觀賞的影片。

經過這幾年國內外通訊與電腦界的努力，小型視訊伺服器(支援 30- 40 video streams) 的技術日漸成熟，然而對於發展中、大型視訊點播(支援 100 video streams 以上)的經驗卻十分陌生。其中儲存系統是發展的主要瓶頸，困難在如何達成儲存裝置間的負載平衡(Load Balancing)。由於每部影片被用戶點播的機率不同，若影片配置不當或用戶的存取模式(access patterns)產生非預期性的改變，可能造成資料的讀取動作集中在少數的儲存裝置中，造成儲存系統的負載不平衡，使得後續用戶被拒絕(Reject)的機率大為提高，而降低系統的效能。因此，影片配置便成為設計視訊點播服務系統時的主要考量。

為了達成儲存系統的負載平衡，條列式資料配置(Striping)與熱門影片複製(Replica)為目前常用方式。條列式資料配置(Striping)，將固定大小的影片資料區塊以循序式(round-robin)平均分散在磁碟陣列，因此每部磁碟機的負載基本上是平衡，然而對大型視訊點播中，將影片直接分散於數十甚至上百個磁碟中，是不可行。熱門影片複製(Replica)，將較受歡迎的影片複製在不同儲存裝置中，以降低其對某部儲存裝置負載的影響程度。在系統資源固定的情況下，如何決定每部影片的複製份數及其配置的位置，是設計視訊儲存系統，重要的議題。

本論文的主題在探討在中型視訊點播環境下如何以影片配置(Video Placement)的方法，來達到儲存裝置間的負載平衡(Load Balancing)，以提高系統的效能。我們的儲存系統是許多儲存群組(storage unit)構成，每個儲存群組(storage unit)是個小型的磁碟陣列，資料以條列式(Striping)的方式，分散在磁碟機中。在這樣的儲存系統架構下，利用改良式 Webster's Monotone Divisor Method[2]決定熱門影片複製(Replica)份數，將較受歡迎的影片複製在不同儲存群組中。再提出負載轉移能力的概念，最佳的影片配置在使負載轉移能力得到極大值，配合圖形理論(Graph Theory)的分析，我們發展出靜態影片配置方法;Major Copy Round Robin(*MCRR*)。再根據影片點播頻率與負載平衡的原則，將影片資料一一的放置到儲存群組中，使得每部儲存群組除了負載能夠平衡之外，同時也具有很好的負載轉移能力，在面對不同的用戶存取模式時，儲存群組間負載平衡的情況仍有很好的表現，系統整體資源的使用率提高，系統的用

國科會計劃: NSC-86-2213-E011-061

戶拒絕率(User Reject Ratio)也因此而降低了。。

相關研究

視訊點播環境下影片配置的研究相當多[1-6]，其中IBM提出 MMPacking[1]，先將影片按其點播機率的遞增順序排列，然後將影片依照 Round Robin 的方式，循序地放置到每部儲存伺服器中，同時累計此儲存伺服器所儲存之影片的點播機率總和，此時若儲存伺服器的機率累計值大於某一數值(假設系統共有 N 個儲存伺服器，則此數值等於 $1/N$)，即決定將此影片繼續複製到下一個儲存伺服器中，重複此方式直到所有的影片均放置完畢為止。此種方法雖然可使每部儲存伺服器的機率累計值相等，但在不同的用戶存取模式情況下，其負載平衡的差異性極大。

Philip S. Yu 與 Hadas Shachnai 提出 *CLLF* [3]法，則是先根據預測的影片點播機率與總儲存空間的關係，利用 Webster's Monotone Divisor Method[2]計算出每部影片的最佳複製份數，然後使用圖形理論的觀念，優先將多重複製的影片放置到儲存群組，以建構出 Clique Trees 的圖形，剩餘的單一複製影片再以最少負載優先(Least Loaded First ; *LLF*)的方式將其放置完畢。

一般對於影片點播機率的預測，經常採用 Zipf Distributions 方法來預測影片點播機率的分佈情況。在 Pure Zipf Distribution 方法中，只使用影片總數 V 當作唯一的參考依據，如公式(1)，其中 π_i 表示影片 i 被要求點播的機率。

$$\pi_i = \frac{1}{i}\left[\sum_{i=1}^{V}\frac{1}{i}\right]^{-1} \qquad (1)$$

2 靜態影片配置方法

本論文中儲存系統是許多儲存群組(storage unit)構成，每個儲存群組(storage unit)採小型磁碟陣列的架構(例如以 8 部磁碟機組成一個儲存群組)，其中資料的佈局是以條列式(Striping)的方式，分散在磁碟機中，因此在儲存群組內，每部磁碟機的負載是平衡。當新用戶提出點播要求時，系統是以 Least Load First-(*LLF*)的方式進行，即從有此用戶點播之影片且有足夠的輸入/輸出頻寬的儲存裝置中，挑選儲存負載最少。

靜態影片配置方法決定每部影片所需的複製份數及其放置位置，以便於負載平均分配到每部儲存群組，以提高儲存系統效率，增加能同時能提供服務的人數。由於用戶提出服務要求的存取模式(access pattern)並非固定，因此我們除了考慮每個儲存群組所儲存之影片的點播機率總和外，儲存群組間的負載轉移能力也是考量的重要因素。

儲存群組的負載能轉移至別處，表示其他的儲存群組中至少需儲存有一相同的影片方能達成。一個好的儲存系統應具備最佳的負載平衡與負載轉移能力，才能在不同的用戶存取模式下，提供最多的同時性視訊資料流。對於儲存系統負載轉移能力的分析，我們以圖形理論的方式來說明。首先定義使用的參數：

V：表示所有影片的總數。

V_i：表示某部影片 i。

π_i：表示影片 V_i 的點播機率故 $\sum_{i=1}^{V}\pi_i = 1$。

r_i：表示影片 V_i 的複製份數。

β_i：表示影片 V_i 複製的每個影片資料的平均點播機率故 $\beta_i = \pi_i / r_i$。

D：表示所有儲存群組的總數。

d_j：表示儲存群組 j。

c_j：表示儲存群組 d_j 所能儲存的影片資料總數。

α_j：表示儲存群組 d_j 所儲存之影片資料的點播機率總和。

L_j：表示儲存群組 d_j 的輸入/輸出頻寬最大可提供的同時性視訊資料流總數。

C：表示全部儲存群組所能儲存的影片資料總數，故 $C = \sum_{j=1}^{D} c_j = \sum_{i=1}^{V} r_i$。

為了方便說明起見，在此我們假設影片資料的大小均相同，所需的頻寬亦相等，每部儲存群組的特性(包括儲存空間，輸入/輸出頻寬)均一致的情況。

任何影片配置的結果，均可將其轉換成相對應的圖形 G(如圖 2)，其定義如下：圖形 G 中的每個節點(Node)代表某部儲存群組，每個邊(Edge)則代表其相鄰的兩個節點中至少儲存有一相同的影片資料，此邊所代表的意義為：相鄰的兩個節點彼此具有負載轉移的能力，至於負載的轉移量，與節點中所配置的影片資料及當時用戶提出服務要求的存取模式有關。

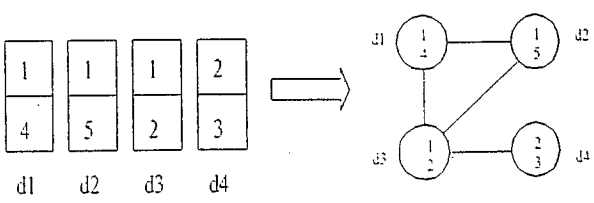

圖 1 影片配置結果及其對應的圖形 G

為衡量儲存系統負載轉移能力的優劣，定義負載轉移力 S_L 如下：

$$S_L = \sum_{j=1}^{D} \alpha_j \cdot deg(d_j) \qquad (3)$$

即每個節點中所儲存之影片資料的點播機率總和乘以此節點的度(Degree ：$deg(d_j)$表示連接至此節點的邊的總數)，此數據愈大，代表儲存系統的負載轉移能力愈好。

由圖形的性質得知，當圖形 G 為完全圖(Complete Graph)時，每個節點的$deg(d_j)$有最大值，即

$deg(d_1) = deg(d_2) = ... = deg(d_D) = D-1$，公式(3)可改寫成

$$S_L = (D-1)\sum_{j=1}^{D} \alpha_j \qquad (4)$$

公式(4)中 $\sum_{j=1}^{D}\alpha_j = 1$(因影片的點播機率總和等於1)，由此可知$0 \le S_L \le D-1$。

圖形 G要成為完全圖，表示每個節點中至少均需儲存有一相同的影片資料才能達成，而依據負載平衡的原則，則希望每部儲存群組所儲存之影片資料的點播機率總和能夠相等：即$\alpha_1 = \alpha_2 = .. = \alpha_D$。

根據先前的分析結果，每部影片之複製份數決定的方式，採取先將最受歡迎的影片放置在所有儲存群組中(也就是讓$r_1 = D$)，讓圖形 G 成為一完全圖，使得每部儲存群組均有很好的負載轉移能力，再將剩餘的影片與儲存空間，根據 Webster's Monotone Divisor Method 的計算方式，決定出每部影片的最佳複製份數，使每個影片資料所分配到的點播機率較為平均。我們假設儲存系統的全部儲存空間$C > V-1+D$，其執行步驟如下所述：

影片複製份數計算方式

(0) 讓最受歡迎的影片其複製份數$r_1 = D$，其餘的影片其複製份數暫定$r_i = 1, i = 2,3,.....,V$。

(1) 讓除數規範(Divisor Criterion)$d(r_i) = r_i + (1/2)$。

(2) 計算每部影片的$\pi_i / d(r_i)$，並從複製份數$r_i < D$的影片中找到$\pi_i / d(r_i)$值最大的影片為V_k。

(3) 讓$r_k = r_k + 1$，如果$\sum_{i=1}^{V} r_i = C$，已完成影片複製份數計算；停止；否則回到步驟(1)。

在此稱最受歡迎的影片V_1為 Major Copy Video ，$r_i > 1$的影片稱為 Multiple-copy Video ，$r_i = 1$的影片稱為 Single-copy Video。

當每部影片的複製份數決定後，將影片配置到儲存群組中。由於用戶要求點播的影片可能為 Single-copy Video 或 Multiple-copy Video ，對儲存系統負載平衡的影響程度不同。當用戶點 Single-copy Video 時，視訊伺服器無法根據目前儲存系統資源的使用情況做出適當的調配；相反的，若為 Multiple-copy Video 時，則視訊伺服器會挑選負載最少且儲存有此一影片資料的儲存群組來提供此服務，以平衡儲存系統的負載。

整個靜態影片配置方式稱為 Major Copy Round Robin(*MCRR*)分成兩個階段進行：

1. 先將 Multiple-copy Videos 以 Round Robin 的方式循序地放置到每部儲存群組中，此時並不考慮儲存群組所儲存之影片資料的點播機率總和α_j。

2. 再將 Single-copy Videos 以最少負載優先(*LLF*)的方式，放置到儲存群組剩餘的空間中，在此階段計算每部儲存群組所儲存之影片資料的點播機率總和α_j，如此做法的主要考量因素為：每部 Single -copy Video 的影片資料，一天中佔用系統整體資源的比例約等於其點播機率π_i；但每部 Multiple-copy Video 的每個影片資料，一天中佔用系統整體資源的比例並非等於其平均的點播機率β_i，而是會因儲存系統的負載平衡狀況而有所不同。詳細的執行步驟如下敘述：

(一) Multiple-copy Video 的配置方式

(1) 將每部 Multiple-copy Video 放入影片佇列(Video Queue)中，影片佇列的排序方式先依影片份數r_i做遞減排序，對份數相同影片再依其平均點播機率β_i做遞增順序排列。

(2) 將每部儲存群組按其編號d_j依序地放入裝置佇列(Device Queue)中，裝置佇列的存取方式為先進先出(First In First Out ；*FIFO*)。

(3) 從影片佇列前端取出一部影片 v_i。

(4) 從裝置佇列前端取出一部儲存裝置 d_j。

(5) 將影片 v_i 存入儲存群組 d_j，讓影片份數 $r_i = r_i - 1$，儲存群組 d_j 尚能儲存的影片總數 $c_j = c_j - 1$，如果 $c_j = 0$，則將儲存群組 d_j 移除，否則將其放回裝置佇列的最末端；如果 $r_i > 0$，則回到步驟(4)，否則將影片 v_i 移除

(6) 如果影片佇列中尚有影片存在，則回到步驟(3)。如果影片佇列為空集合，則完成 Multiple-copy Video 的配置。

(二) Single-copy Video 的配置方式

(1) 將每部 Single-copy Video 放入影片佇列(Video Queue)中，影片佇列依影片點播機率 π_i 遞增排列。

(2) 將尚有剩餘空間的儲存群組放入裝置佇列(Device Queue)中，裝置佇列依其播機率總和 α_j 的做遞增順序排列。

(3) 從影片佇列前端取出一部影片 v_i。

(4) 從裝置佇列前端取出一部儲存群組 d_j。

(5) 影片 v_i 存入儲存群組 d_j 之後將其移除，讓 $c_j = c_j - 1$，$\alpha_j = \alpha_j + \pi_i$，如果 $c_j = 0$，則將儲存群組 d_j 移除，否則依 α_j 值將其放回裝置佇列中適當位置。

(6) 如果影片佇列中尚有影片存在，則回到步驟(3)。如果影片佇列為空集合，則完成 Single-copy Video 的配置。

4 模擬結果

這節中，主要在描述模擬方式的架構及模擬的結果。我們是以 Simscript V2.5 for Windows NT/95 作為主要的模擬程式語言。首先說明系統中使用到的參數，假設系統中總共有 200 部影片 V=200，影片的長度均為 1.5 小時，每部磁碟機的容量(Capacity)可儲存 3 部影片及輸入/輸出頻寬最多可提供 10 個同時性視訊資料流，儲存群組是以 8 部磁碟機的架構組合而成的，所以每部儲存群組的容量共可儲存 24 部影片 c_j=24 及輸入輸出頻寬最多可提供 80 個同時性視訊資料流 L_j=80。我們模擬了二種系統架構，也就是假設系統中，總共有 9 部儲存群組 D=9 與 12 部儲存群組 D=12 的環境。在模擬靜態影片配置方法時並與文獻[3]所提的 *CLLF* 演算法互相比較。我們是以下列二點作為衡量影片配置演算法優劣的依據：

1. 儲存群組間負載失衡(Load Unbalancing)的分佈情況。
2. 用戶拒絕率(User Reject Ratio)。

當有新用戶要求服務或已上線的用戶觀看影片結束時，儲存群組的負載才會有變動。此時，重新計算儲存群組的負載；即其使用率 $U(d_j) = x_j / L_i$，x_j 代表負載變動後的儲存群組 d_j 所提供的同時性視訊資料流總數。衡量儲存群組間負載平衡的方式以每隔 1 小時計算每部儲存群組的平均使用率，然後以最高使用率減去最低使用率 $(MAX(U(d_i)) - MIN(U(d_j)))$ 的差值，觀察儲存系統在 24 小時中的每個時段儲存群組間負載差異的分佈情況。

對於系統的負載(Workload)方面，由實際情形與參考相關的論文研究，我們得知在一天內用戶提出服務要求的頻率，並不是均勻的分佈，會隨著時間的不同而有所改變，但每天的分佈情況卻很相似，如圖 2，在晚上 8 點至 10 點的尖峰時段有最多的用戶要觀看節目，而最少用戶使用的時間則是分佈在早上 5 點至 7 點的離峰時段。因此用戶提出服務要求的產生方式，是根據波瓦松過程(Poisson Process)配合每個時段用戶的平均到達率 λ(Mean Arrival Rate)，以 $1/\lambda$ 為均值的指數分佈(Exponential Distribution)方式，來決定下一個用戶要求到來的

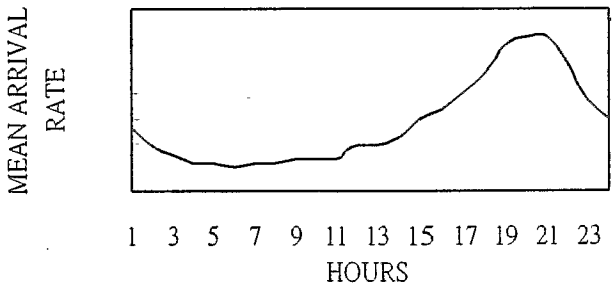

圖 2 系統負載(Workload)的分佈情形

時間。

在影片點播機率的預測方面，採用之前介紹的 Zipf Like Distribution 方法，選定 θ=0.271 來決定每部影片一天中被要求點播的機率為何。在模擬程式中不同的用戶存取模式產生的方式，是依據每部影片一天中被點播的次數而非其機率；且我們欲了解系統在不是很精確的預測結果情況下，儲存系統負載平衡的結果為何，因此在計算每部影片一天中被點播的次數時，刻意地將總使用人數提高之後，再根據其機

率計算其被點播的次數。

為了觀察 *MCRR* 與 *CLLF* 在不同的用戶存取模式時，對儲存系統負載平衡產生的影響，我們設計了三種用戶存取模式。在圖例中假設影片 V_1 與 V_2 經由預測的結果，其點播機率的比例為 $\pi_1/\pi_2 = 1.5$；且圖例中影片點播機率的分佈與系統負載的分佈無關，說明如下：

- 一致性(Uniform)的用戶存取模式：此種存取模式代表在任何時段中，影片 V_1 與 V_2 被要求點播的機率比例均不變的情況。
- 一般性(Normal)的用戶存取模式：此種存取模式，在某些時段中，影片 V_1 與 V_2 被要求點播的機率，會與原先預測的略有不同；π_1 與 π_2 的比例有時會小於 1.5，有時則會大於 1.5，但無論如何 π_1 始終會大於 π_2。
- 隨機性(Random)的用戶存取模式：此種存取模式，在某些時段中，影片 V_1 與 V_2 被要求點播的機率，會與原先預測的略有不同；有時 π_1 會大於 π_2，有時 π_1 則會小於 π_2。

以上三種的用戶存取模式，在模擬一天結束之後，每部影片被要求點播的機率會與原先預測的結果相似(會有些許的誤差存在)。而在模擬一般性與隨機性的用戶存取模式時，模擬程式每隔 1 小時便會隨機選定一介於 0.2 至 0.5 的 θ 值，然後以 Zipf Like Distribution 的方式來決定某個時段的影片點播機率。

模擬結果

以下是模擬得到的結果，圖表中各項數據的取得，是系統執行 100 天之後的平均值。在擁有 9 部儲存群組的系統中，共有可儲存 16 部影片的多餘空間；讓我們可複製一些較受歡迎的影片到其他的儲存群組中，靜態影片配置及動態影片調整方法的模擬結果如下：

靜態影片配置方法模擬結果：

擁有 9 部儲存群組的系統：在此系統架構下，使用的系統負載模式如圖 2 所示，在晚上 8 點至 10 點的尖峰時段，每分鐘平均約有 9 個人次的用戶提出服務的要求，依據此種系統負載模擬的結果，一天結束時約有 5400 人次的用戶提出服務的要求，因此每部影片一天中被要求點播的次數，是以 5600 人次乘以其點播機率計算得來的。

1. 一致性用戶存取模式的結果：此情況，一天中總共產生的用戶要求次數為 5387 次，以 *CLLF* 方法造成用戶被系統拒絕服務的次數為 300 次，其用戶拒絕率為 0.0557；而以 *MCRR* 方法造成的用戶被系統拒絕服務的次數為 107 次，其用戶拒絕率為 0.0199，如圖 4，負載差異分佈的情形，如圖 3。

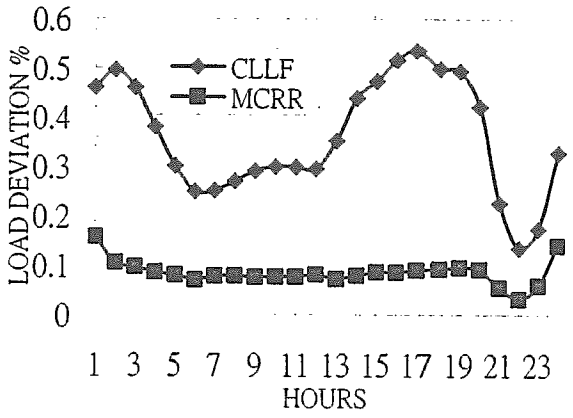

圖 3 不同方法之負載差異分佈情形

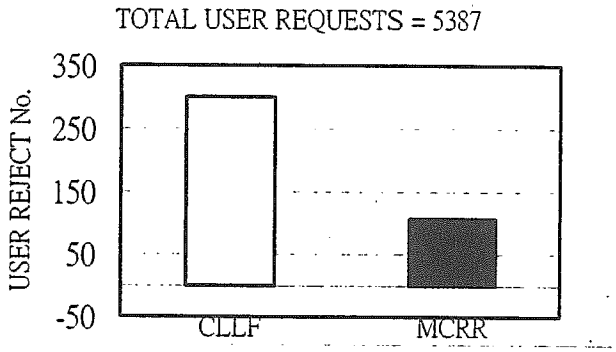

圖 4 不同方法之用戶拒絕次數比較

2. 一般性用戶存取模式的結果：此情況，一天中總共產生的用戶要求次數為 5386 次，以 *CLLF* 方法造成用戶被系統拒絕服務的次數為 334 次，其用戶拒絕率為 0.062；而以 *MCRR* 方法造成的用戶被系統拒絕服務的次數為 115 次，其用戶拒絕率為 0.0214，如圖 6，負載差異分佈的情形，如圖 5。

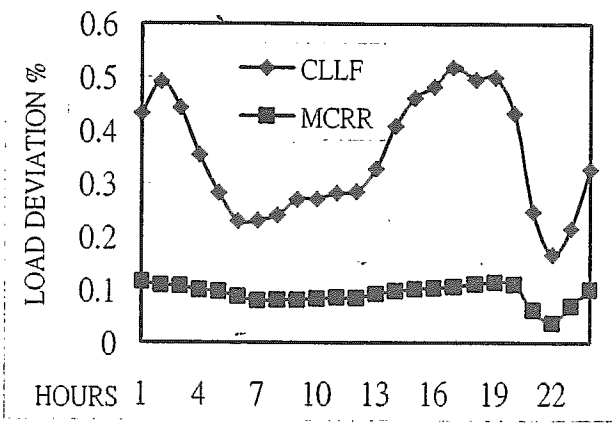

圖 5 不同方法之負載差異分佈情形

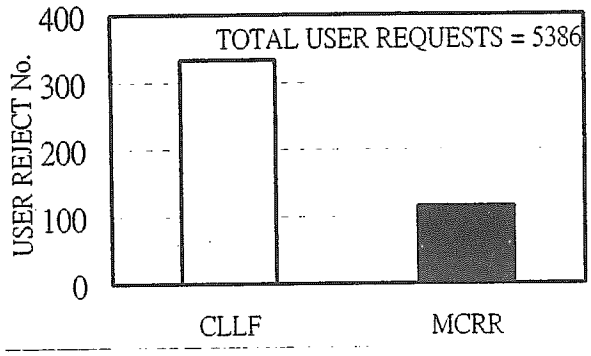

圖 6 不同方法之用戶拒絕次數比較

3. 機性用戶存取模式的結果：此情況，一天中總共產生的用戶要求次數為 5390 次，以 *CLLF* 方法造成用戶被系統拒絕服務的次數為 780 次，其用戶拒絕率為 0.1447；而以 *MCRR* 方法造成的用戶被系統拒絕服務的次數為 178 次，其用戶拒絕率為 0.033，如圖 8，負載差異分佈的情形，如圖 7。

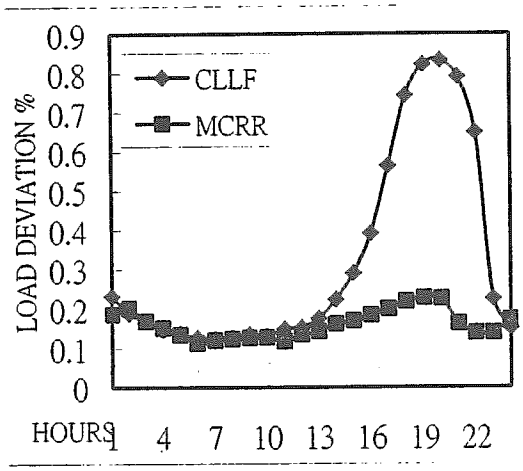

圖 7 不同方法之負載差異分佈情形

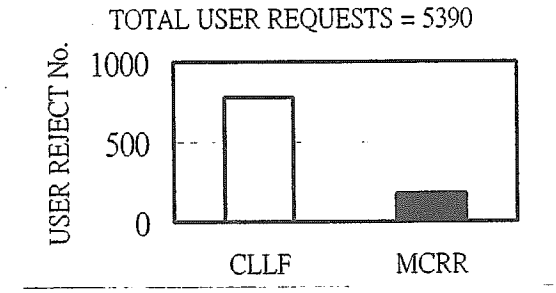

圖 8 不同方法之用戶拒絕次數比較

5 結 論

本論文針對視訊點播服務系統之儲存系統，提出了 *MCRR* 靜態影片配置方法配合動態影片調整演算法，來解決儲存系統負載平衡的問題。影響負載平衡的主要因素可區分為：影片的配置方式與用戶的存取模式，我們根據上述的影響因素，以圖形理論的分析模式，得到 *MCRR* 靜態影片配置方法。其先將最受歡迎的影片複製到每部儲存群組中，使每部儲存群組均有很好的負載轉移能力，其餘的影片與儲存空間，再依據負載平衡的原則決定出每部影片的最佳複製份數。而在放置影片資料時，為了避免 One Copy Video 的放置位置受到已配置在儲存群組中之 Multiple Copy Videos 的影響，我們採取只計算儲存群組所儲存之 One Copy Videos 的點播機率總和的方式，將 One Copy Videos 一一的放置到儲存群組中。

由模擬的結果得知，*MCRR* 的靜態影片配置方法配合動態影片調整演算法，使儲存系統在不同的用戶存取模式情況下，仍能將負載平均的分散到每部儲存群組中，對於提高系統效能與降低用戶拒絕率方面，有很明顯的效果。

參考文獻

[1] N. Serpanos, L. Georgiadis and T. Bouloutas, "MMPacking: A Load and Storage Balancing Algorithm for Distributed Multimedia Servers", Technical Report RC 20410, IBM Research Division, T.J. Watson Research Center, March 1996

[2] T. Ibarkai and N. Katoh, "Resource Allocation Problems - Algorithmic Approaches", The MIT Press, 1998

[3] Joel L. Wolf, Philip S. Yu and Hadas Shachnai, "DASD Dancing: A Disk Load Balancing Optimization Scheme for Video-on-Demand Computer Systems", Technical Report RC 19987, IBM Research Division, T.J. Watson Research Center, March 1995

[4] Robert Flynn and William Tetzlaff, "Disk Striping and Block Replication Algorithms for Video File Servers", Technical Report RC 20328, IBM Research Division, T.J. Watson Research Center, 1996

[5] Asit Dan and Dinkar Sitaram, "An Online Video Placement Policy based on Bandwidth to Space Ratio (BSR)", IBM Research Division, T.J. Watson Research Center

[6] Chatschik C. Bisdikian and Bai ju V. Patel, "Issues on Movie Allocation in Distributed Video-on-Demand Systems", IBM Research Division, T.J. Watson Research Center