視訊點播服務之資料儲存系統設計
The Design of Storage System for VOD Services

陳郁堂　杜睿哲
台灣科技大學電子系
ytchen@et.ntit.edu.tw

摘要

視訊點播服務系統的負載變動性相當大，尖峰時段的高負載可能造成系統資源不足。據研究顯示，負載常集中在少數熱門影片，本論文從儲存系統角度，提出整合磁碟機儲存系統與動態記憶體管理的架構，針對存取頻率較高的影片，以更有效率的資料存取方式，提供多重相位(Multiple Phases)播放服務，並在尖峰時段，並以動態資源使用方式來姐決尖峰時段資源不足的問題。

Abstract

Previous research has shown that the VOD services are scare in system resources during peak hours. This paper presents a storage system for supporting multiple access points to the same video object. The proposed storage architecture enables smooth integration of the disk storage system and dynamic buffer management within the video server. We also explore the feasibility of dynamic resource allocation to attack the shortage of resources at peak hours.

1. 緒論

近年來，有關視訊點播系統(Video on Demand, VOD)的研究，深受國內外學術與產業界重視。在市場上亦出現小型視訊點播(VOD)的產品。使用者透過高速網路，以交談的方式 (Interactive)觀賞影片。然而對於大型視訊點播系統經驗並不多。近來有關大型視訊點播服務研究顯示：

- 大型視訊點播服務並不適合採用真實性視訊點播服務(True VOD) [1]。近似性視訊點播(Near-VOD) [1]，將點播同一部影片的使用者集中，於固定的時段(time interval) ，以廣播(broadcast)方式進行，是較經濟的服務模式。然而使用者等待的時間增加，無法得到即時性的服務。

視訊點播服務在尖峰與離峰負載變動性相當大，特別在尖峰時段的高負載可能造成系統資源不足[1]。然而負載常集中在少數熱門影片。因此針對存取頻率較高的影片提供系統更有效的存取方式，不失為降低系統負載良策。

近年來關於視訊伺服器(Video Server)儲存系對同一部影片統的研究如汗牛充棟，然而關於以儲存系統提供使用者做多重相位(Multiple Phases)播放的研究並不多，僅 Rotem[3]提出以緩衝記憶體為主體的架構，然而視訊資料體積龐大，以緩衝記憶體建構，價格勢必昂貴，從經濟觀點，並不可行。本論文提出整合磁碟機儲存系統與動態記憶體管理的架構來解決這個問題。

我們將探討大型視訊伺服器中支援多重相位影片播放模組(稱之為複製型記憶模組)之設計，以降低磁碟機讀寫的次數，並改良近似性視訊點播中使用者的等待時間過長的缺點，除了提供使用者對同一部影片多重相位的播放服務以外，並以動態資源使用方式來姐決尖峰時段資源不足的問題。

在磁碟機儲存系統方面，為了降低搜尋時間(Seek Time)耗費與負荷不平衡對系統產生的負面影響，我們以使用者存取型態控制(Access Pattern Control)的方式，配合系統設計流程(System Design Procedure)步驟，在符合多媒體播放連續性限制(Continuous Constraint)的條件之下，利用交替式(Interleaved)的資料讀取來達到搜尋時間的疊覆(Overlapping)與負載平衡(Load Balancing)。

在記憶體管理方面，以提高資料重複使用率的機，來降低熱門影片的負載。我們利用預測的方式，加入使用者等待限度(Wait Tolerance)與離開機率(Turn Away Probability)的人為因素考量，發展動態記憶體視窗調整演算法，並配合實際的工作負荷(Workload)模型加以模擬，驗證該演算法在資源使用上的優越之處。

本文的章節編排如下：第二節說明儲存裝置的架構及其資料佈局，第三節提出使用者的存取型態控制演算法；第四節探討複製型記憶模組系統設計流程；第五節有關以動態調整記憶體視窗來改良系統，第六章結論。

2. 系統架構

複製型記憶模組的硬體架構如圖 1，由多磁碟機與共用記憶體緩衝區，並配合高速匯流排構成。圖 2 為複製型記憶模組系統整體架構。

2.1 視訊資料佈局

多重磁碟機儲存子系統，將磁碟機分成兩群，資料佈局方式如下(圖 3)：

- 視訊資料分割成固定大小的資料區塊(Striping Unit)以條列式(Striping)的方式輪流地(Round-Robin)放置在每一部磁碟機上。
- 為了在存取資料時，能降低讀寫頭的移動

國科會計劃: NSC-85-2213-E011-040

(磁碟機搜尋),視訊資料區塊在同一部磁碟機以連續放置。

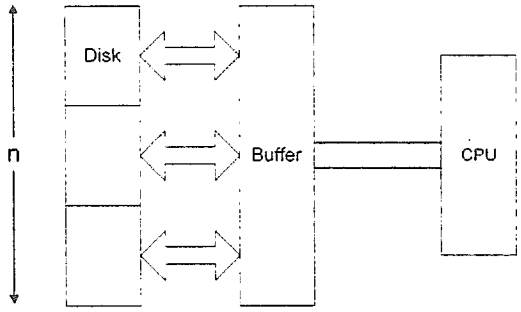

圖 1 複製型記憶模組系統硬體架構示意圖

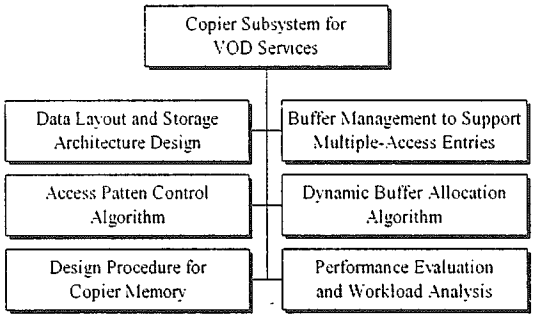

圖 2 系統整體架構

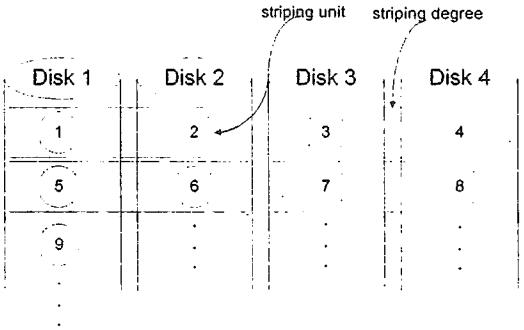

圖3 資料在磁碟機上的放置方式

B	buffer size measured in bytes	T_{wl}	rotation latency
b	max. bytes in a block	T_{ws}	maximal disk seek time
b_r	MPEG blocks refreshed per cycle	S_i	system idle time
d	the display rate in block/sec	W_t	user wait tolerance
R_d	the display rate in Mbps	C_t	service cycle time
r	the disk transfer rate inMB/sec	N_c	the number of users be served
λ	the mean request arrival rate	P_s	the memory partition size
γ	the mean request wait tolerance	w	register window size
S	latency and seek time in ms	D	degree of striping

表 3-1 使用參數與其代表的意義

2.2 暫存視窗

複製型記憶模組以記憶體作為緩衝區,利用視訊伺服器儲存資料的唯讀(Read Only)特性與區域性(Locality),以重複利用先前讀進來的資料。對同時段內先後點播相同影片的觀眾提供不同時段的切入點的服務,可達到降低磁碟機讀寫的次數,提升系統的效率。

記憶體分割為數個暫存視窗,每個暫存視窗資料更新與的步驟如下[3],(如圖 4.所示):

1. 開始時,由系統先讀進資料充滿整個記憶暫存視窗,並開始服務第一位到來的使用者。
2. 隨後在相鄰近的時間內到來點播相同影片的觀眾,從記憶體中讀取既有的資料。
3. 當第一位使用者消耗完記憶體資料前 $y = d(s + \frac{b}{r})$ 時,關閉暫存視窗,暫停使用者進入,進行資料更新的動作。
4. 從資料開始更新到更新完畢這一時段,記憶體中的資料流亦同時消耗資料,故實際上讀進的資料量只有 $b_r - d*s$,其中 b_r 表示每一回讀進的資料量,$d*s$ 表示磁碟機搜尋時記憶體資料的消耗量。

所以只要使用者在資料更新之前進入系統,落在暫存視窗允許的範圍之內,都可以由記憶體讀取資料而不須對磁碟機直接存取,充分利用記憶體並有效地改善系統效率。

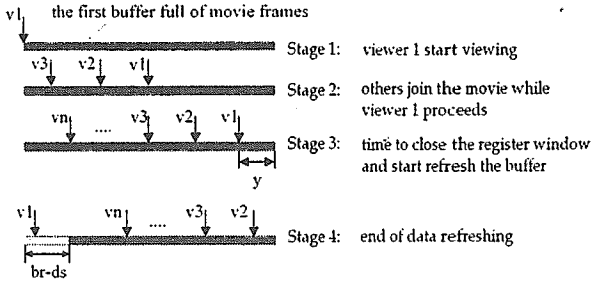

圖 4 記憶體資料更新過程

3. 存取型態控制與演算法

系統對資料存取型態(Access Pattern)會隨著時間和使用者而改變。在瞬間,如果大多數的請求都落在同一部或同一群磁碟機的情況,將造成系統負載不平衡(Load Unbalancing),使該群磁碟機成為輸出/輸入的瓶頸,而降低系統效率。

我們將多重磁碟機(multiple disks)分成兩群,配合交替式的資料讀取(interleaved retrieval);當一個資料流在傳送資料時,另一個資料流可以搜尋下一個將讀取的資料區塊,資料傳送時間可以疊覆搜尋時間,提昇系統的效率。當第 n 個資料流讀取磁碟機群 1 時,如果新加入的第 $n+1$ 個資料流也要讀取磁碟機群 1 的資料,則其讀取動作將被不被允許,必須等到下一個服務週期,第 n 個資料流讀取磁碟機群 2 時,新進來的第 $n+1$ 個資料流,才可以開始讀取磁碟機群 1 的資料。利用上述存取型態控制可達到磁碟系統負載平衡,

以下是我們針對存取型態控制所提出的演算法。

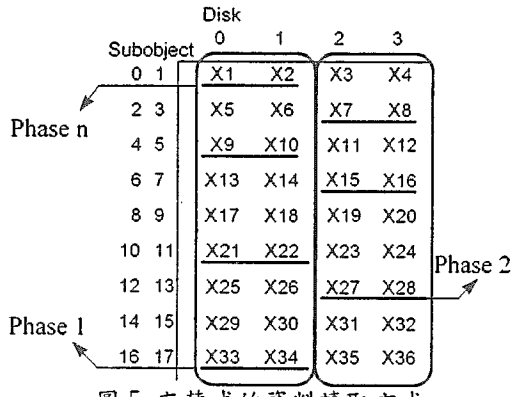

圖 5 交替式的資料讀取方式

步驟一： 對於系統裡的請求，產生形如(disk ID, process ID, block number)的請求序列(Request Sequence)，並以先到先服務(FCFS)的方式將請求放入請求佇列(Request Queue)，如圖 6 之①。

步驟二： 以旗號(Flag)做為請求進入系統的存取控制，並將之設為 1，所有系統裡的請求形成一個服務週期(Service Cycle)，依照 disk ID 分成兩群進入磁碟機系統，也是同樣以先到先服務的方式由磁碟機讀取所需要的資料區塊，如圖 6 之②。

步驟三： 假如這時有新的請求到來，若旗號為 1，則將新到來的請求放到存取型態控制佇列(Access Pattern Control Queue)並且等候。

步驟四： 磁碟機儲存系統服務完上一個週期的所有請求以後，把旗號設為 0，將所有請求的 block number 加 1，同時改變 disk ID，產生下一個週期的請求序列，如圖 6 之③。

步驟五： 檢查存取型態控制佇列，看是否有新的請求到來，假如有新的請求，而且讀取 block number 所在的 disk ID 不等於請求佇列裡頭最後一個請求的 disk ID，則將新的請求加入請求佇列。

步驟六： 檢查請求佇列是否為空，如果不是，重複步驟一到步驟六，直到所有請求離開系統。

4. 設計流程

由於磁碟機搜尋延遲(seek time)而降低傳輸頻寬，利用交替式的資料讀取方式疊覆資料搜尋時間技巧來解決這個問題，搜尋時間疊覆限制條件(Seek Time Overlapping Constraint)為資料區塊搜尋所需的時間(seek time)，小於資料的傳送時間，這樣系統才可以在替一個資料流傳送資料同時，搜尋下一個資料流所要讀的資料區塊。

$$T_{ws} + T_{wl} \le \frac{b * b_r}{r} \qquad (1)$$

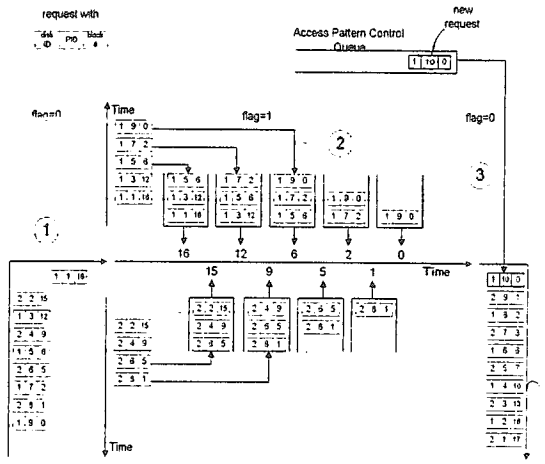

圖 6 存取型態控制

在式(1)中，不等號的左邊 T_{ws} 和 T_{wl} 分別代表磁碟機最大的搜尋時間及旋轉延遲，右邊則是將大小為 b 的資料區塊個數 b_r 傳入記憶體所需的時間。

在設計複製型記憶模組，必須考慮視訊資訊播放的連續性，即視訊資訊塊(video data block) 從記憶體消耗的時間，必須大於視訊資訊資訊塊從磁碟機讀出並傳到記憶體的時間。同時必須考慮記憶體與磁碟機儲存系統上的整合。傳統在磁碟機上的多媒體檔案系統，多採循序輪流(Round-Robin)的方式來服務系統中的使用者，必須在每個使用者緩衝區的資料消耗完之前，讀進新的資料，這樣才不會造成影片播放的中斷。

在一個週期裡服務 n 個使用者，其播放連續性的限制可以式(2)表示，s 為系統的搜尋時間，

$$S_i = \frac{b_r}{d} - n\left(s + \frac{b * b_r}{r * D}\right) \qquad (2)$$

S_i 代表第 i 個週期系統閒置的時間，$\frac{b_r}{d}$ 是在一個週期裡讀進的資料消耗完畢的時間，$\frac{b * b_r}{r * D}$ 為將大小為 b 的影片資料區塊 b_r 分由 D 個磁碟機傳送的時間，

我們會發現磁碟機的搜尋時間延遲在整個服務的時間裡佔了極大的比例，因而降低了系統的效能；但是如果採用多重磁碟機，並配合存取形態控制以及搜尋時間的疊覆，可以把搜尋時間多餘負擔對系統的影響降到最少，如式(3)所示，每一次讀進來的資料是 b_r，那麼系統閒置(Idle)的時間便可以用讀進資料播放的時間 $\frac{b_r}{d}$，減去搜尋時間 s，再減去服務 n 個使用者所需的時間。

$$S_i = \frac{b_r}{d} - s - n*\frac{b_r*b}{r*D} \quad (3)$$

從式(3)可以看出，當 $S_i = 0$ 時，即磁碟機一直都處於忙碌的狀態，系統能夠得到最大的效率，同時我們也可藉此算出每個資料流所需讀進最少的資料量，即 b_r，如式(4)所示

$$\text{Let } S_i = 0 \Rightarrow b_r = \frac{r*d*s*D}{r*D - n*b*d} \quad (4)$$

我們可以看出它除了與使用者人數、搜尋時間、傳輸速率和播放速率有關以外，還和資料在磁碟機上的條列式排列程度(Striping Degree) D 有關，根據這個參數可以決定磁碟機的個數，而系統所能夠服務最多的使用者即為總記憶體 B 除以每一個資料流所佔的空間大小，式(5)、(6)。

$$N_c = B/(b*b_r) \quad (5)$$

$$N_c = \frac{B*D*r}{D*b*d*r*s + B*b*d} \quad (6)$$

從以上分析得到的結果，我們可以由資料流對記憶體的需求，建構出資料區塊在磁碟機上所佔的大小與系統服務的使用者人數上限，並在決定系統預計支援的使用者人數之後，進而決定出條列程度 D，如式(8)

$$\text{Let } n = \frac{B*D*r}{D*b*d*r*s + B*b*d} \quad (7)$$

$$D = \frac{n*B*b*d}{B*r - n*b*r*d*s} \quad (8)$$

由於我們使用存取型態的控制，採取交替式的讀取方式，所以我們所需要磁碟機的個數至少要為兩倍的條列程度，亦即 $2*D$，每個資料流所佔分割為總記憶體大小除以使用者的人數，如式(9)，另外，暫存視窗的大小則如式(10)所示，等於整個分割的大小減掉資料流每一回更新資料時所讀進的資料，再減去花費在傳進一個資料區塊的傳輸時間所造成的資料播放消耗。

$$P_s = \frac{B}{n} \quad (9)$$

$$W = \frac{P_s}{b} - b_r - \frac{d*b}{r} \quad (10)$$

歸納複製型記憶模組設計流程如下：

步驟一： 以系統預計支援的人數 N_c，計算條列式程度 D 及決定儲存系統所需的磁碟機個數。 $D = \frac{n*B*b*d}{B*r - n*b*r*d*s}$

磁碟機的個數為 $2*D$

步驟二： 計算更新資料時所需最少的資料區塊個數 b_r 以及資料量的大小。由式(4) $b_r = \frac{r*d*s*D}{r*D - n*b*d}$

資料量更新的大小為 $b*b_r$

步驟三： 由上述結果分析資料流所占記憶體分割 P_s 的大小。

$$P_S = \frac{B}{n}$$

步驟四： 計算儲存資料所需的緩衝區大小 W。

$$W = \frac{P_s}{b} - b_r - \frac{d*b}{r}$$

5. 動態記憶體管理

在近似性的視訊點播(Near VOD)，系統在固定的時段(Time Interval) 以整批(Batch)的方式對觀賞相同影片的觀眾提供服務，這種做法能有效地提升系統效率，但對部份使用者，可能等待超過容忍限度(Wait Tolerance)之後離開系統，我們處理的方式是在於利用複製型記憶模組，延長資料在記憶體中的時間，提供即時性不同相位存取點的服務，並滿足耐性較差的使用者。

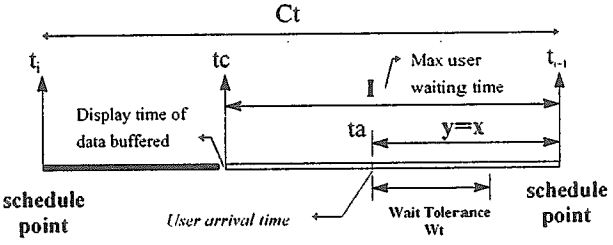

圖 7 使用者離開的機率

在一大型且點播頻繁的視訊點播服務系統之中，系統負載超過允入控制(Admission Control)的臨界值，系統必須限制使用者進入，如圖 7，假設該使用者進入系統的時間為 t_i，t_{i+1} 則代表下一位允許進入系統的使用者，而每一位使用者的到來是獨立而不相干的 Poisson Process，到來的時間間隔(Inter Arrival Time)是以 $\frac{1}{\lambda}$ 為均值的指數分佈，而使用者的等待限度亦是以 $\frac{1}{\gamma}$ 為均值的指數分佈函數，若 t_c 代表儲存在記憶體中的資料所提供的播放時間，在 $t_{i+1} - t_c = I$ 這一段時間區段之內，也就是使用者

到下一個處理時間點的最大等待時間，P 代表等待限度太短造成使用者流失的機率，換句話說，當使用者在 t_d 的時間到來，而等待限度 W_t 小於等於 y 的時候，使用者會離開系統，所以我們可以推得使用者離開的機率，如式(11)：

$$\begin{aligned}&P\{W_t \le y\}\\&= \int_{x=0}^{x=I} P\{W_t \le y | y = x\} * P\{y = x\} dx\\&= \int_{x=0}^{x=I}\left(\int_{t=0}^{t=x} \gamma * e^{-\gamma t} * dt\right) * \frac{1}{I} * dx\\&= \int_{x=0}^{x=I}\left(1 - e^{-\gamma x}\right) * \frac{1}{I} * dx = 1 + \frac{e^{-\gamma I}}{\gamma I} - \frac{1}{\gamma I} \quad (11)\end{aligned}$$

，γ 和 Waiting Time 成反比，當 γ 愈大而 Waiting Time 愈小的時候，使用者流失的機率幾乎等於 1 ，

我們計算使用者進入系統之後，因為不足以等待到下一個處理時間點而離開系統的人數，亦即在算出 I 時間區段內使用者流失的期望值，若 λ 代表使用者到來的速率，則使用者離開系統之期望值則為 I 時段內所有到來的人數乘以流失的機率 P ，如式(12)，

$$E = \lambda * I * P \quad (12)$$

我們定義使用者流失率(User Lost Ratio) φ ，式(13)，代表該段時間之內流失的人數與到來總人數之比，其中 E 為式(12)所求的期望值，C_t 則是使用者因為允入控制的關係而必須等待下一次進入系統的時間，我們可以由式(14)算出使用者可能等待的最長時間 I ，

$$\varphi = \frac{E}{\lambda * C_t} = \frac{I * P}{C_t} \quad (13)$$

$$I = \frac{\varphi * C_t}{P} \quad (14)$$

$$t_c - t_i = \frac{W}{R_d} = C_t - I \quad (15)$$

從式(15)計算可能需要的暫存視窗 W 的大小，得式(16)

$$\begin{aligned}W &= \left(C_t - I\right) * R_d \quad (16)\\&= C_t * \left(1 - \frac{\varphi}{P}\right) * R_d\end{aligned}$$

我們可以看出暫存視窗 W 的大小與可能流失的機率 P 和使用者流失率 φ 之間的關係，若希望沒有使用者的流失，即 $\varphi = 0$ ，

得 $$W = C_t * R_d \quad (17)$$

式(17)是在兩次相鄰的對磁碟機的實際讀取之中，若不想造成使用者因等待太久而離開系統，或是使用者等待限度太短而需要提供即時性服務所需的記憶體大小。

動態視窗調整演算法

對一個視訊伺服器而言，它的負荷隨著時間而有不同的分佈[1]，我們希望分配給每一個視訊資料流的視窗大小也能夠隨著負荷的不同而改變，在使用者稀疏，配置較小的緩衝區給資料流，而在使用者密集到來時，能以較大的緩衝區空間滿足較多的使用者，並根據使用者點播統計的時間分佈圖，預測下一階段可能的負荷分佈，

以下就是我們所提出動態調整記憶體視窗的演算法：

步驟一：　首先，記錄從上一次對磁碟機的實際存取到這一次的實際存取時間 C_t ，減去原來暫存視窗 W_i 所涵蓋的時間以後，得到如圖 7 中的時間區段 I ，由式(11)計算使用者可能流失的機率 P ，用以預測到下一次的實際存取使用者可能的動態。

步驟二：　由流失機率 P 與使用者上一次流失的比率 φ 利用式(16)

$$\Delta W = C_t * \left(1 - \frac{\varphi}{P}\right) * R_d$$

計算視窗的調整大小。，

步驟三：　根據使用者到來密集的程度以及到來的數目，比較前後兩次使用者流失的情況，得到 Diff-ratio，如果這一次到來的速率大於等於上一次使用者到來的速率，即 Diff-ratio = 1 ，則將原來的視窗大小加上 ΔW ，得 $W_{i+1} = W_i + \Delta W$ ，做為新的資料流所需的暫存視窗大小，反之則減去 ΔW ，得 $W_{i+1} = W_i - \Delta W$ 。

步驟四：　隨著系統負荷的輕重，動態調整分配視窗大小，重複循環步驟一到步驟四，直到所有的使用者離開系統。

6.結論

本論文中提出了複製型記憶模組系統的架構，希望藉著磁碟機儲存系統的改良與緩衝區資料的重複用，提供使用者在同一時段，欣賞不同片段的影片播放，同時以動態的資源運用方式，克服伺服器於尖峰時段，

因請求密集到來所產生的資源不足現象。
在磁碟機儲存系統方面，我們發展存取形態控制的演算法，將進入系統的請求平均分散到磁碟機，藉由交替式的資料讀取，來達成儲存系統的負荷平衡以及搜尋時間的疊覆，滿足多媒體物件的播放條件，有效提昇儲存系統的性能。
在記憶體管理方面，解決系統在尖峰時段所面臨播放的瓶頸，是發展複製型記憶模組架構的出發點，我們利用影像的唯讀與資料的重複使用特性，配合使用者的等待限度(Wait Tolerance)及離開機率(Turn Away Probability)的分析，我們提供了理論上的模型，並且以動態記憶體視窗調整的方式，節省視訊資料流所耗費的系統資源與降低使用者流失的比率。

7.參考文獻

[1] Thomas D. C. Little and Dinesh Venkatesh, "Prospects for Interactive Video-on-Demand," IEEE Multimedia, Fall 1994, pp. 14-24.

[2] W. Sincoskie, "System Architecture for Large Scale Video on Demand", Computer Networks ISDN System," Vol. 22, 1991, pp. 155-162

[3] D.Rotem, and J.L.Zhao, "Buffer Management for Video Database Systems," Proc. Intl. Conf. On Data Eng., Taipei, Taiwan, March 1995, pp. 439-448.