

주문형 비디오 시스템을 위한 예약 기반 동적 배칭 정책

이 경숙, 하 숙정, 배 인한
대구효성가톨릭대학교 전자정보공학부

A Reservation-based Dynamic Batching Policy for Video-on-Demand Systems

Kyung-Sook Lee, Sook-Jeong Ha, Ihn-Han Bae

Department of Computer Engineering, Catholic University of Taegu-Hyosung

요 약

주문형 비디오 시스템의 성능에서 중요한 문제는 클라이언트 요구를 만족시키기 위한 입출력 대역폭이다. 이러한 입출력 대역폭을 감소시키기 위해 하나의 비디오 스트림을 사용하여 다중 요청을 처리하는 많은 정책들; 배칭, 피기백 머징, 버퍼링이 연구되고 있다. 본 논문에서는 인기 있는 비디오들에 대한 서비스 요청들이 매 배칭 간격마다 스케줄될 수 있도록 인기 있는 비디오들을 위한 서버 스트림 용량을 동적으로 예약해두는 예약 기반 동적 배칭 정책을 제안하고, 그것의 성능을 시뮬레이션을 통하여 단순 배칭 정책과 비교하고 평가한다.

1. 서 론

정보통신 분야에서 최근의 기술적인 발전은 주문형 비디오(video-on-demand), 홈쇼핑(home-shopping)과 같은 여러 가지 주문형 멀티미디어 시스템을 가능하게 만들었다. 오늘날의 정보 시스템은 단순히 커다란 멀티미디어 객체를 저장하고 검색하는 것만이 아니라, 객체를 일정한 대역폭에서 계속적으로 제공하는 엄격한 실시간 요구사항을 만족시킬 수 있어야 한다. 멀티미디어 시스템은 교육용 어플리케이션, 오락 기술, 도서관 정보 시스템등에서 중요한 역할을 하고 있다. 이러한 시스템에서 가장 중요한 것은 다중 요청들을 처리하는 방법이다. 즉, 사용자들은 영화 등과 같은 객체를 요구하고, 적절한 지연시간내에 볼 것을 기대한다. 지연시간은 요구의 도착한 때부터 시스템이 디스크로부터 객체 읽기를 초기화하는 시간 간격

으로 정의한다. 데이터가 실제적으로 디스플레이 장치에 전달될 때까지의 추가적인 지연은 상대적으로 무시할 수 있다. 지연은 서비스 요구를 위한 불충분한 대역폭, 디스크로부터 읽은 내용을 스케줄링 하기 위한 불충분한 버퍼 공간, 불충분한 디스크 기억장치 등의 요인으로 발생한다. 이러한 지연 요소 중에서 입출력 대역폭(I/O bandwidth)은 매우 중요한 자원이다. 그러므로 기억장치 서버에서 공유될 수 있는 사용자 요구의 수를 증가시키기 위하여 입출력 대역폭을 줄이는 여러 접근 방법들이 다음과 같이 제시되고 있다[3].

- 배칭(batching) : 일정한 배칭 간격동안 도착하는 동일한 객체에 대한 요구들을 모아 단일 입출력 스트림으로 전체 그룹을 서비스하는 방법이다.
- 브리징(bridging) : 중앙 처리기의 메모리를 버퍼로서 이용하는 방법이다. 특정 비디오에 대해 고정된

숫자의 프레임이 버퍼링 된다면, 대응하는 시간 간격 안에 발생하는 그 비디오에 대한 어떤 요구도 디스크가 아닌 버퍼에서 읽혀올 수 있다. 브리징은 많은 양의 버퍼 공간을 요구하는 단점이 있다.

- 피기백(piggyback) : 동일한 객체에 대한 입출력 스트림이 하나로 머지될 때까지 진행중인 스트림의 디스플레이율을 조정하는 기법이다.

배치 정책에서 매 배치 간격마다 대기 큐에 인기 있는 비디오에 대한 다수의 스트림 요청들이 존재할 것이고, 인기있는 비디오에 대한 요청들이 매 배치 간격마다 스케줄되어야 많은 비디오 스트림들이 배치되고 하나의 입출력 스트림으로 서비스되어 많은 비디오 스트림들이 절약된다. 따라서 비디오 스트림들이 많이 스케줄되어 스트림 이탈율도 감소할 것이다. 본 논문에서는 인기 있는 비디오들에 대한 스트림들이 매 배치 간격마다 스케줄될 수 있도록 인기 있는 비디오들을 위한 서버 스트림 용량을 동적으로 예약해두는 예약 기반 동적 배치 정책을 제안하고, 그것의 성능을 시뮬레이션을 통하여 단순 배치 정책과 비교하고 평가한다.

본 논문의 구성은 다음과 같다. 2장에서는 배치 정책에 대해서 살펴보고 3장에서는 예약 기반 동적 배치 정책을 제안하고, 4장에서는 본 논문에서 제안하는 예약 기반 동적 배치 정책의 성능을 시뮬레이션 통하여 평가하고, 마지막으로 5장에서 결론을 맺는다.

2. 배치

배치는 같은 객체에 대한 요구들을 모아 기억장치 서버에 대한 하나의 입출력 요구를 만드는 것이다. 배치 방법에서의 상반관계는 스트림 요구의 지연시간과 기억장치 서버에서 대응되는 입출력 요구의 감소이다. 배치 정책은 크기에 의한 배치와 시간에 의한 배치로 나눌 수 있다. 크기에 의한 배치는 각 객체 j 에 대해 미리 정해진 배치 윈도우 B_j 를 두고, 객체 j 에 대해 B_j 만큼의 크기가 누적되면 객체 j 에 대해 입출력 스트림을 초기화한다. 그러므로 객체 j 에 대해 감소되는 입출력 스트림의 크기는 $B_j - 1$ 이다. 따라서 $E[N_j] = B_j - 1$ 이다. 여기서 $E[N_j]$ 는 요청된 I/O 스트림의 개수에서 배치에 의한 예상 감소 스트림의 개수를 L_j 는 오브젝트 j 에 대한 각 요청에 의해 발생하는 대기 시간을 나타내는 랜덤 변수를 나타낸다.

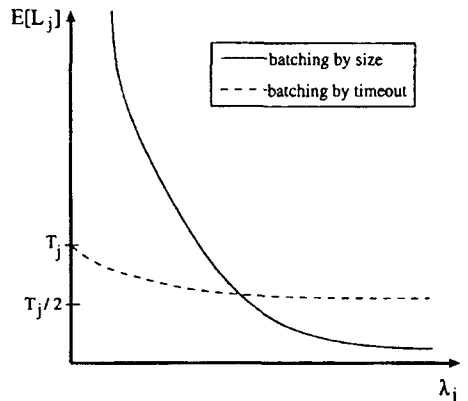
$$E[L_j] = \frac{1}{B_j} \sum_{i=1}^{B_j} \frac{B_j - i}{\lambda_j} = \frac{B_j - 1}{2\lambda_j}$$

비록 이 정책이 저장 서버상의 I/O 요구를 감소시키지만 특히 적절한 도착율 보다 낮은 요청들에서 긴 지연이 발생할 수 있다.

시간에 의한 배치는 시간단위로 배치 윈도우를 설정하는 것이다. 하나의 요구가 기억장치 서버에 도착하고 같은 객체 j 에 대한 요구가 존재하지 않을 때 타이머가 설정된다. 시스템은 타이머의 초기화 후 T_j 단위 시간내에 저장 서버에 I/O 요청을 제출한다. T_j 단위 시간 동안에 도착하는 같은 오브젝트에 대해 어떤 요청이 배치되고 타이머가 끝났을 때 서비스된다. N_j 를 배치으로 인하여 절약된 I/O 스트림들의 개수를 나타내는 랜덤 변수라 하고, p_j^k 를 시간 T_j 동안에 k 요청 도착 확률로 가정한다. 도착 프로세스가 포아송이므로 $E[N_j] = \lambda_j T_j$ 이다. 각 요청에 의해 발생하는 예상 대기 시간을 평가하기 위하여, 그 시스템을 상수 준비 시간(T_j)과 결정적 서비스 분포를 갖는 M/G/1 큐로 볼 수 있다. 이런 종류의 시스템에 대한 예상 지연 시간은 다음과 같다.

$$E[L_j] = \frac{T_j(2 + \lambda_j T_j)}{2(1 + \lambda_j T_j)}$$

$E[N_j] = \lambda_j T_j$ 이므로 적당히 높은 요청 도착율에서 엄청난 저장 서버상의 I/O 요구를 감소시킬 수 있다. 이것은 각 오브젝트 j 에 대한 각 요청들이 배치으로 인한 지연이 T_j 단위 시간 보다 크지 않기 때문에 주문형 영화 응용을 위한 적당한 정책이다. 그림 1은 2가지 배치 정책에 대한 예상 지연 시간을 보여준다. 시간에 의한 배치 정책은 FCFS 정책, MQL 정책, FCFS_n 정책 등이 있다[4].

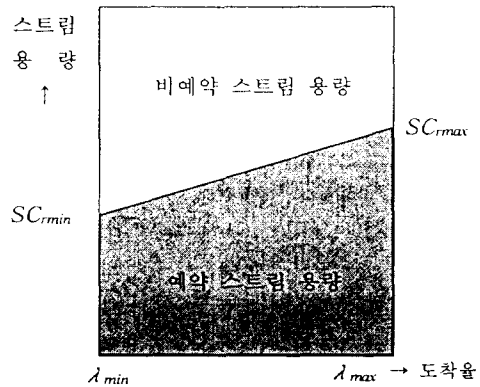


(그림 1) 배치 정책에 대한 예상 대기 시간

3. 예약 기반 동적 배칭

본 논문에서 제안하는 예약 기반 동적 배칭 정책은 기존의 단순 배칭에 인기있는 비디오를 위한 입출력 스트림 예약이라는 개념을 도입한다. 단순 배칭에서는 비디오 스트림들에게 도착순으로 입출력 스트림을 할당해 줌으로써 도착율이 어느 정도 이상 높아지면 도착한 비디오 스트림들은 무조건 기다려야 한다. 이때 인기있는 비디오를 위한 가용 스트림이 있으면 인기있는 비디오에 대한 같은 스트림들은 배치되어 많은 비디오 스트림들이 절약되어 보다 많은 비디오 스트림들을 스케줄할 수 있을 것이다. 그러나 가용 스트림이 없으면 가용 스트림이 생길 때 까지 기다려야 하고 결국 비디오 서비스 대기 시간이 길어지고 사용자 서비스 이탈율이 증가하여 고품질의 주문형 비디오 서비스를 제공할 수가 없게 된다. 따라서 본 논문에서는 단순 배칭에서의 이러한 문제점을 해결하기 위하여 인기 있는 비디오에 대한 스트림들이 매 배칭 간격마다 스케줄될 수 있도록 입출력 스트림(비디오 서버 스트림 용량)을 미리 예약해 둔다. 본 논문의 예약 기반 동적 배칭 정책의 경우, 비디오 서버의 부하에 따라 입출력 스트림이 예약되는 인기있는 영화의 개수가 가변적으로 변한다. 즉, 비디오 서버가 저부하일 경우에 대부분의 비디오 스트림들이 스케줄되므로 인기있는 비디오들을 위해 비디오 서버 스트림 용량을 예약하지 않거나 아주 인기 있는 비디오들에 대해서만 제한적으로 비디오 서버 스트림 용량(SC_{min})을 예약한다. 반면에 비디오 서버가 과부하일 경우에 많은 인기있는 비디오 스트림을 포함한 대부분의 비디오 스트림들이 블록되므로 인기있는 비디오 스트림이 스케줄될 경우 다수의 같은 비디오 스트림들이 배치되어 많은 입출력 스트림들이 절약될 뿐 아니라 많은 비디오 스트림들이 스케줄될 것이다. 따라서 비디오 서버가 과부하일 경우에 적당한 개수의 인기있는 비디오가 스케줄될 수 있을 정도의 비디오 서버 스트림 용량(SC_{max})을 예약한다. 특히 비디오 서버 과부하 상황에서 다수의 인기 있는 비디오를 항상 스케줄할 수 있을 만큼의 비디오 서버 스트림 용량을 예약하면, 비인기 비디오를 위해 할당되는 스트림 용량이 줄어들고 결국 비인기 비디오의 서비스 대기 시간이 길어져 이탈될 수밖에 없다. 따라서 비디오 서버의 부하에 따라 스트림 용량이 예약되는 인기있는 비디오의 개수와 그에 따른 예약되는 서버 스트림 용량을

동적으로 조정한다.



(그림 2) 서비스 도착율에 따른 비디오 서버 스트림 예약 용량

본 논문에서 제안하는 예약 기반 동적 배칭 알고리즘의 구조는 그림 3과 같다.

```

Algorithm Reservation-based Dynamic Batching
Begin
  Reserve video server stream capacity for popular videos
  according to arrival rate of video stream;
  Case arrival of a video stream:
    Put the arrival video stream in waiting queues;
  Case the batching interval of the front requesting stream
  in waiting queue is expired:
    If the requesting video is in the n-hottest video
      If reserved stream capacity is not full
        Batch the same video streams in waiting queue
        into a single I/O stream;
        Schedule the video stream with a stream in
        reserved stream capacity;
      Else
        Block the video stream;
    Else
      If unreserved stream capacity is not full
        Batch the same video streams in waiting queue
        into a single I/O stream;
        Schedule the video stream with a stream in
        unreserved stream capacity;
      Else
        Block the video stream;
  End
    
```

(그림 3) 예약 기반 동적 배칭 알고리즘

4. 시뮬레이션 및 성능 평가

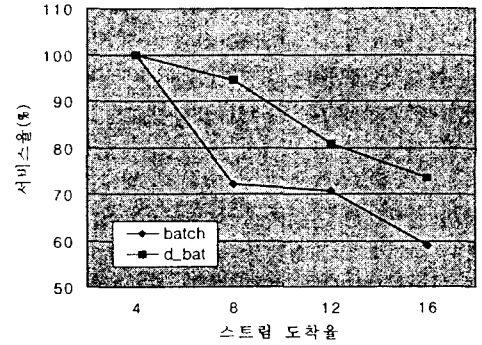
본 논문에서는 시뮬레이션을 통하여 배칭 정책과 예약 기반 동적 배칭 정책의 성능을 비교·평가하였

다. 시뮬레이션 파라미터는 표 1과 같다. 여기서 영화 i의 인기도는 Zipf의 법칙을 따른다고 가정한다. 실험 결과는 시뮬레이션의 공정성을 기하기 위하여 서비스를 시작한 후 100분과 200분 사이의 100분간 서비스된 비디오 스트림들로 정책들의 성능을 측정하였다. 여기서 평가된 성능 평가 항목들은 스트림 도착율에 따른 서비스율, 이탈 백분율, 평균 대기 시간 그리고 프레임 절약 백분율 등이다.

< 표1 > 시뮬레이션 파라미터

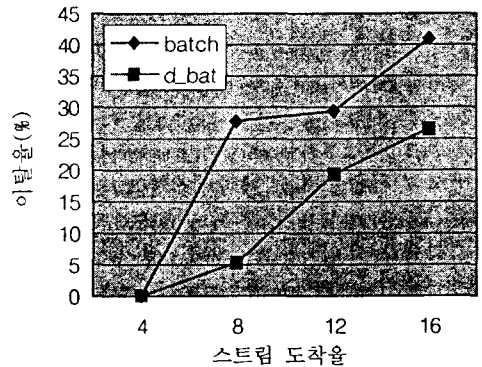
파라미터	값
비디오의 개수	60개
서버 스트림 용량	400
예약 서버 스트림 용량	75~125
예약 인기 비디오의 개수	3~6
배치 간격	3분
도착율/분	4, 8, 12, 16(포아송)
영화 길이	100
이탈시간	3분~5분(랜덤)

그림 4는 비디오 스트림 도착율에 따른 배치 정책별 비디오 스트림 서비스율을 보여준다. 여기서 비디오 스트림 서비스율은 전체 도착한 비디오 스트림 개수에 대한 스케줄된 비디오 스트림 개수의 백분율로 계산되었다. 여기서 비디오 서버가 저부하인 경우(도착율/분=4)에 단순 배치(batch)과 동적 배치(d_bat) 모두 도착한 비디오 스트림을 전부 스케줄 하였으나 두 정책 모두 비디오 서버의 부하가 높아질수록 비디오 스트림 서비스율이 떨어짐을 알 수 있다. 그러나 비디오 서버 과부하 상황에서도 본 논문에서 제안하는 예약 기반 동적 배치가 단순 배치 보다 10~20% 이상의 높은 비디오 스트림 서비스율을 보이고 있다. 이것은 단순 배치에서는 비디오 서버 과부하 상황에서 모든 비디오 스트림들이 블록되는데 반해 예약 기반 동적 배치에서는 인기 있는 비디오를 위한 비디오 서버 스트림 용량을 예약함으로써 과부하 상황에서도 많은 인기 있는 비디오 스트림들이 배치되어 하나의 입출력 스트림으로 스케줄되는 결과이다. 따라서 같은 서버 스트림 용량을 가지고 예약 기반 동적 배치가 단순 배치 보다 많은 고객들에게 주문형 비디오 서비스를 제공할 수 있음을 알 수 있다.



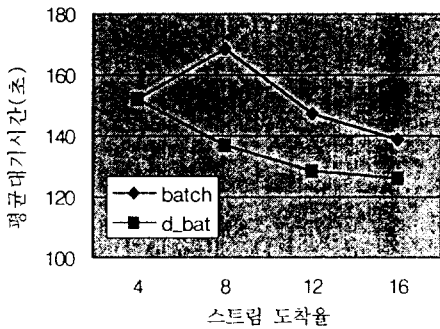
(그림 4) 스트림 도착율에 따른 서비스율

그림 5는 비디오 스트림 도착율에 따른 배치 정책별 서비스 이탈율을 보여준다. 서비스 이탈율은 그림 4의 서비스율과 상반관계가 있으므로 스트림 도착율이 높아질수록 이탈율은 높아짐을 알 수 있다. 여기서도 비디오 서버가 저부하인 경우에 단순 배치와 동적 배치 모두 서비스를 이탈하는 비디오 스트림은 전혀 없으나 비디오 서버 과부하 상황에서는 동적 배치가 단순 배치에 비해 10~20% 낮은 이탈율을 보이고 있다. 이것은 비디오 서버 과부하 상황에서 단순 배치는 모든 비디오 스트림들을 블록하는데 반해 예약 기반 동적 배치는 인기 없는 적은 개수의 비디오 스트림들은 블록되나 인기 있는 많은 개수의 비디오 스트림들은 여전히 배치되고 스케줄되는 결과이다. 따라서 본 논문의 예약 기반 동적 배치가 단순 배치에 비해 더 낮은 고객 이탈율을 보임으로써 더 나은 QoS를 제공할 수 있음을 알 수 있다.



(그림 5) 스트림 도착율에 따른 이탈 백분율

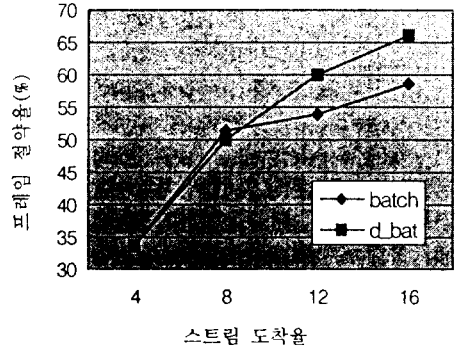
그림 6은 비디오 스트림 도착율에 따른 배치 정책 별 평균 대기시간을 보여준다. 평균 대기시간은 서비스된 영화의 대기시간의 합을 총 서비스된 영화의 개수로 나누어 계산하였다. 여기서는 비디오 스트림 도착율이 높아질수록 서비스된 비디오 스트림의 평균 대기시간은 오히려 줄어든다. 이것은 비디오 서버 과부하 상황에서는 지금 스케줄되는 비디오 스트림과 같은 비디오 스트림들이 배치 인터벌내에 많이 존재하여 짧은 대기시간을 갖는 많은 비디오 스트림들이 배치되는 결과이다. 그리고 동적 배치가 단순 배치에 비해 10초 이상 더 낮은 평균 대기시간을 보여주고 있다. 이것은 비디오 서버 과부하 상황에서 단순 배치는 모든 비디오 스트림 요청이 블록되는데 반해 동적 배치는 인기가 없는 비디오 스트림 요청들은 블록되나 인기 있는 비디오 스트림들을 여전히 배치되고 스케줄될 뿐 아니라 인기 있는 비디오의 더 많은 스트림들이 배치되는 결과이다.



(그림 6) 스트림 도착율에 따른 평균 대기 시간

그림 7은 비디오 스트림 도착율에 따른 배치 정책 별 프레임 절약율을 보여준다. 프레임 절약율은 절약된 프레임의 합을 전체 처리된 프레임의 합으로 나눈 백분율이다. 이 때 전체 처리된 프레임의 합은 (서비스된 비디오의 개수 × 한 비디오의 프레임 개수)로 한 비디오의 프레임 개수는 $(100 \times 60 \times 30) = 18,000$ 프레임이다. 여기서 비디오 스트림 도착율이 높아질수록 배치와 동적 배치 모두 프레임 절약율이 높아짐을 알 수 있다. 그리고 비디오 서버 과부하 상황에서 예약 기반 동적 배치가 단순 배치에 비해 5% 이상의 높은 프레임 절약율을 보이고 있다. 이것은 비디오 서버 과부하 상황에서 단순 배치는 배치할 수 있는 많은 인기 있는 비디오 스트림들이 있으나 가용 서버 스트

림이 없어 블록되므로 비디오 프레임을 절약할 기회를 놓치게 된다. 그러나 본 논문의 예약 기반 동적 배치에서는 비디오 서버 과부하 상황에서도 인기 있는 비디오 스트림들은 여전히 배치되고 스케줄되므로 많은 비디오 프레임들이 절약되는 결과이다.



(그림 7) 스트림 도착율에 따른 절약 백분율

5. 결론

본 논문에서는 인기 있는 비디오 스트림을 위하여 비디오 서버 스트림 용량을 예약해두는 예약 기반 동적 배치 정책을 제안하고, 시뮬레이션을 통하여 단순 배치 정책과 본 논문의 예약 기반 동적 배치 정책의 성능을 비교·평가하였다. 여기서는 스트림 도착율에 따른 서비스율, 이탈 백분율, 평균 대기시간, 프레임 절약 백분율로 각 정책의 성능을 평가하였다. 그 결과 모든 성능 평가 항목에서 예약 기반 동적 배치 정책이 단순 배치 정책 보다 우수함을 알 수 있었다. 앞으로의 연구 내용은 다양한 시뮬레이션 환경 하에서의 다른 입출력 대역폭 감소 정책들간의 성능 평가와 그 정책들간의 QoS에 관한 평가 등이다.

참고문헌

- [1] C. Aggarwal, J. Wolf and P. Yu "On Optimal Piggyback Merging Policies for Video-on-Demand Systems" Technical Report, IBM RC 20337, Feb. 1996
- [2] C. Aggarwal, J. Wolf, and P. Yu "On Optimal Batching Policies for Video-On-Demand Servers" IEEE Multimedia Computing and Systems Conference, Hiroshima, Japan, 1996

- [3] L. Golubchik, J. Lui, and R. Muntz "Reducing I/O Demand in Video-On-Demand Storage Servers" ACM Sigmetrics Conference, pp. 25-36, May 1995.
- [4] A. Dan, D. Sitaram, and P. Shahabuddin "Scheduling Policies for an On-Demand Video server with Batching" Proceedings of the 2nd ACM Multimedia Conference, pp.25-32, 1994.
- [5] H. Shachnai, P. Yu "An Analytical Study of Multimedia Batching Schemes" Technical Report, IBM RC 20662 Dec. 1996
- [6] C. Aggarwal, J. Wolf, and P. Yu "The Maximum Factor Queue Length Batching Scheme for Video-on-Demand Systems" Technical Report, IBM RC 20621 Nov. 1996
- [7] 김 윤정, 하 숙정, 배 인한, "주문형 비디오 서버를 위한 예약 기반 Piggyback Merging 정책", 한국정보과학회 학술발표논문집, 제 24권 2호, pp.115~118, 1997.
- [8] 이 경숙, 오 선진, 배 인한, "주문형 비디오 시스템에서 Batching과 Piggyback Merging의 성능 비교", 한국정보과학회 영남지회 학술발표논문집, 제5권 1호, pp. 203~208, 1997.