

VOD 서비스의 스케줄링 기법에 관한 연구

이준화, 유재욱, 박태근, 박재현
명지대학교 전자공학과

A Study on the Scheduling Scheme for VOD Services

Jun-Hwa Lee, Jae-Wook Yoo, Tae-Keun Park, Jae-Hyun Park
Dept. of Electronic Engineering, Myongji Univ.

Abstract

Generally, in a VOD(Video On Demand) system, batching technique is often used to improve throughput. Since subscribers to VOD services tend to withdraw their requests as the latency becomes large, an optimal scheduling scheme should consider not only batching size but also withdrawal rate, service latency and fairness between different programs.

In this paper, we propose a new scheduling scheme, which shows improvement in fairness and reduction in withdrawal rate and service latency compared with other well known schemes. When there are 2 or more streams are available in the VOD server, proposed scheme apply different program selection policies to each streams. Using this approach effectiveness of stream usage can be improved.

1. 서론

흔히 주문형 비디오라 불리는 VOD는 가입자가 원하는 프로그램을 원하는 시간에 제공받는 서비스로 정보통신 기술의 발달로 상용화가 임박한 멀티미디어 기술의 총아라 하겠다[1]. 본격적인 VOD 서비스의 상용화를 이루기 위하여 선결되어야 할 과제들 중의 하나로 최적의 스케줄링 기법 개발을 들 수 있다. 여기서 최적이란 제한된 VOD 서버의 스트림 용량(Stream Capacity)으로 최대의 가입자를 수용함과 동시에 시청 포기율과 서비스 지연시간을 최소화하고 프로그램간의 공평성을 향상시키는 것을 의미한다.

VOD 서버는 잠재적으로 동시에 수천명의 서비스 요청을 수용할 수 있어야 한다. 서버의 제한된 스트림 용량으로 다수의 가입자에게 서비스를 제공하여 수용을 높이기 위한 방법으로 배칭 기법이 있다. 이것은 인접한 시간대에 발생하는 동일한 프로그램에 대한 요청들을 하나의 스트림으로 동시에 서비스하는 방법이다. 배칭 기법을 사용한 스케줄링에서는 필수적으로 서비스의 지연을 수반하게 되는데 이 지연시간이 일정 수준을 초과하면 신청한 프로그램의 시청을 포기하는 가입자가 발생하게 된다[2]. 따라서 VOD 서비스에서 최고의 수용과 최저의 시청 포기율을 보장하는 스케줄링 기법의 개발이 요구된다. 또한 가입자가 프로그램을 요청할 때 요청된 프로그램의 종류에 따라 서비스 지연시간이 달라지고 시청 가능성이 좌우되는 것은 바람직하지 못하다. 즉 가입자 입장에서 프로그램간의 공평성이 있어야 한다.

본 논문에서는 기존의 스케줄링 기법 중 하나인 BML 방법의 단점을 보완하여 포기율과 지연시간을 줄이고 공평성을 향상시킨 스케줄링 기법을 제안한다. 제안된 방법은 VOD 서버에 2개 이상의 스트림이 사용 가능한 경우 각각의 스트림에 대한 서비스 프로그램 선택을 서로 다른 기준으로 하여 효율을 향상시키는 방법이다.

2. VOD 시스템과 가입자 특성

2.1 VOD 시스템 구조

VOD 시스템의 기본 요소는 VOD 서버, 통신망, 그리고 서비스 가입자 세 가지이며, 가입자는 통신망을 통하여 VOD 서버에 원하는 프로그램을 요청하게 된다. 서비스 제공자 측의 시스템은 크게 VOD 서버와 메모리 버퍼 그리고 저장 매체로 구성된다. 메모리 버퍼는 가입자의 요구에 따른 대화형 기능을 제공키 위한 임시 데이터 저장용으로 사용되며 서버는 다시 CPU와 이를 제어하는 프로그램을 포함하고 있다.

2.2 스트림 스케줄링

특정 프로그램에 스트림을 할당하는 과정은 어떤 프로그램을 상영할 것인가와 언제 프로그램이 시작되는가의 두 가지 요소에 의해 결정된다[3].

요청된 서비스를 처리하는 순서는 공정성과 저장된 프로그램의 갯수에 영향을 받는다. 일반적으로 서비스의 공정성과 효율성은 상호 대립의 관계에 있고, 이것은 프로그램의 기호도(Popularity)에 기인한다. 즉, 효율성을 고려하면 인기 있는 프로그램 위주로 스케줄링이 되어야 하는데 이에 따라 공정성은 떨어지게 되는 것이다.

2.3 서비스 가입자 특성

- 프로그램 선택 : VOD 서비스 가입자는 자신이 원하는 프로그램을 선택하게 되는데 프로그램에 따라 서비스 요청이 많은 것도 있고 적은 것도 있다. 즉 각 프로그램은 그 요청 빈도에 따라 자주 선택되는 인기 프로그램과 자주 선택되지 않는 비인기 프로그램으로 구분 될 수 있다.
- 시청 포기 특성 : 시스템의 상태에 따라 다르지만 프로그램 요청이 즉시 서비스 되지 않을 수도 있다. 가입자는 일반적으로 적은 지연은 수용할 수 있지만 지연 시간이 길어지면 시청을 포기하게 될 것이다. 즉 서비스를 받기 위해 기다리는 시간이 기다릴 수 있는 시간보다 커지게 되면 시청 포기가 발생한다. 스케줄링 기법에서 가입자의 포기를 유발시키는 지연시간의 임계치가 중요한 역할을 하게 되며 또한 이에 따른 포기율은 성능 측정에 주요 요소가 된다.

2.4 스케줄링 기법의 성능 평가

일반적으로 VOD 시스템에서 스케줄링 기법의 평가를 위하여 사용하는 성능 측정 요소로 다음과 같은 것

이 있다.

- 포기율 : 요청된 서비스의 총 수 대 시청을 포기한 수의 비를 말한다. 포기율을 최소화함으로써 VOD 시스템의 수율을 최대화 할 수 있다.
- 평균 지연 시간 : 가입자의 서비스 요청으로부터 실제 요청된 프로그램의 상영까지 소요된 시간의 전체적인 평균값을 의미한다. 일반적으로는 서비스를 받은 요청의 지연시간만을 고려하나 본 논문에서는 포기된 요청이 서버의 큐(Queue)에서 소요한 시간도 고려하여 보다 의미 있는 측정요소가 되게 하였다.
- 공정성 : 각 프로그램간의 포기율에 대한 편차로 계산되며 스케줄링 기법이 프로그램을 공정하게 선택하는가에 대한 지표가 된다. 실제적으로는 아래식으로 표현된 불공평성을 측정하며 이 값이 적을수록 공평한 스케줄링이 된다.

$$\sqrt{\frac{\sum_{i=1}^M (\bar{d} - d_i)^2}{M}}$$

M = 프로그램 총 갯수
 \bar{d} = 평균포기율
 d_i = i 번째 프로그램의 포기율

3. 스케줄링 기법

멀티캐스트(Multicast) 기능을 사용하여 동시간 또는 근접한 시간대에 요구되어진 요청들을 하나의 스트림으로 서비스하는 기술을 배칭이라 한다[4,5]. VOD 서버당 수용 가능한 가입자의 수는 서비스 이용료에 직접적인 영향을 미치며 스케줄링 방법에 따라 크게 변화한다. 최적의 스케줄링은 배칭을 최대화 하여 스트림의 효율을 증대 시키는 동시에 서비스 지연시간과 서비스 신청 후 지연에 따른 포기율을 최소화하고 가입자들에게 최대한의 공정성을 보장하여 주는 것이다.

3.1 기존의 스케줄링 기법

기존에 제안된 스케줄링 기법으로는 다음과 같은 것들이 있다.

- FCFS (First-Come-First-Service) : 사용 가능한 스트림이 있으면 언제라도 즉시 요청순서에 따라 서비스를 시작하는 방법으로 프로그램의 종류에 따른 기호도에 좌우되지 않아 공정성이 뛰어나고 서비스 지연시간이 짧은 장점이 있는 방법이나 스트림 이용 효율성이 낮아서 포기율이 높은 것이 단점이다.
- MFQL (Maximum Factored Queue Length) : VOD 시스템에 저장된 프로그램 별로 요청된 서비스를 저장하는

독립적인 큐를 두어 스트림 할당시 가장 큰 FQL을 가진 비디오를 스케줄하는 방법이다[6,7]. 서비스 지연시간이 짧은 장점이 있으나 포기율이 크고 공평성이 떨어지는 단점이 있다

- BML(Batch Maxiimal Loss): 스트림 이용 효율성을 높이기 위하여 일정기간(Tolerance Interval)을 기다린 요청을 보유한 프로그램 중에서 다음 스트림이 서비스 완료 될때까지 포기가 가장 많이 기대되는 프로그램에 스트림을 할당하는 방법이다[8,9]. 낮은 포기율이 보장되나 일정기간 지연되는 동안에 여분의 스트림을 효율적으로 사용하지 못하는 단점이 있다.

3.2 제안된 스케줄링 기법

본 논문에서는 기존의 BML 스케줄링 기법을 기본으로 스트림 할당에 유연성을 갖는 새로운 방법을 제시하여 포기율이나 공평성면에서 보다 우수한 스케줄링 기법을 제안한다. 제안된 기법에 대한 순서도는 그림 1과 같다.

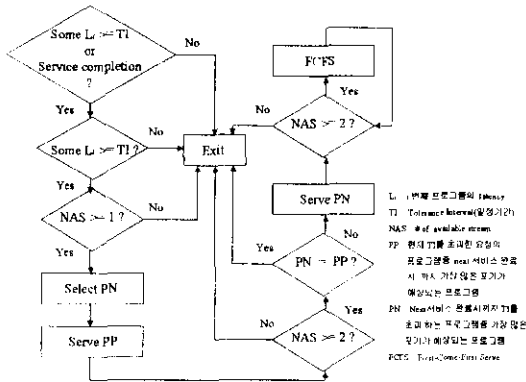


그림 1. 제안된 스케줄링 기법의 순서도

일정 기간을 기다린 요청을 보유한 프로그램 중에서 다음 스트림 할당시까지 가장 많이 포기가 예상되는 프로그램에 대하여 스트림을 할당하는 BML 방법은 강제적인 지연으로 인한 여분의 스트림 발생이 가능한데 이것을 효과적으로 이용하는 방법이 제안된 아이디어이다. 즉, 잔여 스트림이 발생시 이것을 일정기간이 지연되지 않은 프로그램에도 포기 기대치가 높은 경우 또는 가장 먼저 요청된 프로그램에 할당을 한다. 즉 스트림 할당에 유연성을 두어 스트림 이용 효율 및 공평성을 도모하는 것이다. 스트림 할당은 잔여 스트림

이 2개일때까지만 동작하는데 그 이유는 잔여 스트림 1개는 스트림 할당 중간에 발생할 수 있는 포기를 방지하는데 사용하기 위함이다.

4. 시뮬레이션

4.1 시뮬레이션 모델

프로그램 요청은 exponential 간격 분포를 갖는 Poisson process로 발생시킨다. 각 요청자의 프로그램 선택은 Zipf's 분포[10]를 따르며 아래의 식으로 나타내어진다.

$$z(i) = \frac{c}{(i^{1+\theta})}$$

여기서 $z(i)$ 는 i 번째로 인기있는 프로그램이 선택될 확률을 나타내며 c 는 normalizing 상수, $\theta=0.271$ 을 사용하였다.

각 프로그램의 상영 시간은 105~135분의 uniform 분포로, 프로그램 수는 100 개이고 스트림 수는 1000 개로 가정하였다. 가입자가 기다릴 수 있는 시간은 평균 10분 분산 2.5분을 갖는 normal 분포를 따르며 배정을 위한 일정 기간은 10분을 사용하였다.

본 시뮬레이션에서는 일시정지나 재시작과 같은 VCR 기능은 스케줄링 기법의 성능평가에 직접적인 영향을 주지 않아 생략하였으나 실제적인 VOD 서비스에서 간과 될 수 없는 주요 기능으로 연구되어야 할 과제이다.

4.2 성능 평가

그림 2는 VOD 서비스에서 분당 서비스 요청 수에 따른 포기율의 변화 및 스케줄링 기법간의 성능 차이를 보여준다. 제안된 기법이 가장 낮은 포기율을 보여주고 있다.

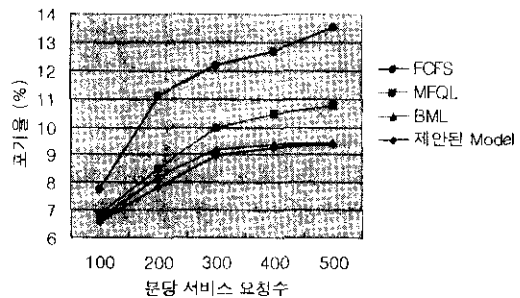


그림 2. 분당 요청 수에 따른 포기율

그림 3 은 분당 서비스 요청 수에 따른 평균지연시간을 보여준다. 본 논문에서는 포기된 요청이 VOD 서버의 큐에서 소비한 시간도 평균 지연시간에 포함시켜 보다 의미있는 측정 요소가 되게 하였다. 제안된 기법의 평균지연시간은 MFQL 보다는 길지만 다른 기법보다는 단축되었음을 보여준다.

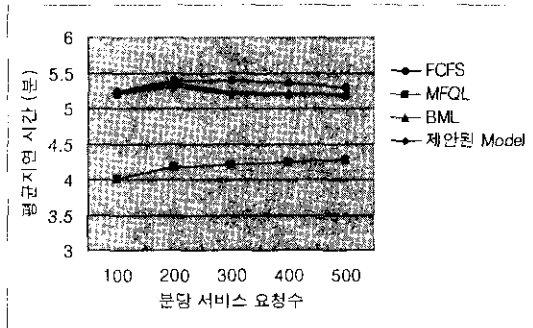


그림 3. 분당 요청 수에 따른 평균지연시간

그림 4 는 분당 서비스 요청 수에 따른 불공평성을 보여준다. 제안된 기법이 FCFS 를 제외하면 가장 공평한 성능을 보여주고 있다.

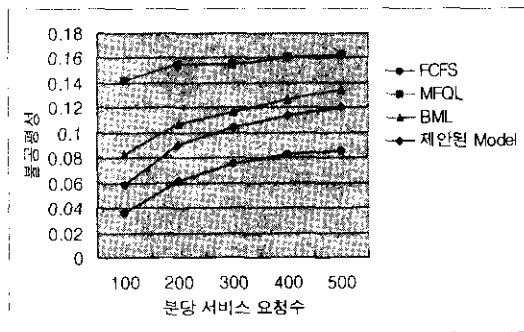


그림 4. 분당 요청 수에 따른 불공평성

5. 결론

VOD 서비스에서의 수출항상 및 공평성과 짧은 지연시간을 보장하기 위한 새로운 스케줄링 기법이 제안되었다. 제안된 방법은 기존 BML 방법에서 발생하는 여분의 스트림을 효율적으로 운용하는데 그 초점을 맞추고 있으며 시뮬레이션을 통하여 그 우수성이 다른 스케줄링 기법과의 비교를 통하여 입증되었다.

본문에서 언급되었듯이 VOD 서비스의 상용화를 위하여 최적의 스케줄링 기법 뿐만 아니라 VCR 기능의 효과적인 구현도 계속 연구 개발되어야 할 과제라 생각된다.

[참고문헌]

- [1] Jani Huoponen and Thorsten Wagner, "Video on Demand : A Survey" Telecommunication Networks Project 1, 1996.
- [2] Dan, A., Sitaram, D., and Shahabuddin, P., "Scheduling Policies for an On-Demand Video Server with Batching", Proceedings of the 2nd ACM Multimedia Conference, San Francisco, CA, pp. 25-32 1994
- [3] Dan, A., Sitaram, D., and William, H., "Long Term Resource Allocation in Video Delivery Systems", Report RC 20294, IBM T J Watson Research Center, Yorktown Heights, NY, 1996.
- [4] Anderson, D.P., "Metascheduling for Continuous Media.", ACM transactions on Computer Systems, Vol.11, No.3, pp.226-252, 1993.
- [5] A.Dan, D.Sitaram, and P.Sitaram, and D.Towsley, "Channel allocation under batching and ver control in movie-on-demand servers," Tech. Rep. RC19588, IBM Research Division, T.J. Watson Research Center, 1994.
- [6] Charu C.Aggarwal, Joel L.Wolf, and Philip, S.Yu, "The Maximum Factor Queue Length Batching Scheme for Video-on-Demand systems", Report RC 20621, IBM T J Watson Research Center, Yorktown Heights, NY, 1996.
- [7] Aggarwal, Wolf, Yu "5631694 : Maximum factor selection policy for batching VOD rquests", IBM Corporation, Armonk, NY, 1997.
- [8] Shachnai, H., and Yu, P., "The Role of Wait Tolerance in Effective Batching : A Paradigm for Multimedia Scheduling Schemes", Research Report RC 20038, IBM T J Watson Research Center, Yorktown Heights, NY, 1995.
- [9] Yu, Philip S. O., "5561456 : Return based scheduling to support video-on-demand applications", International Business machines Corporation, Armonk, NY, 1996.
- [10] Christer Samueleson, "Relating Turing's Formula and Zipf's Law", University Des Sarlandes, FR 8.7 Computerlinguistik Postfach 1150, D-66041.