연속 미디어 서버의 접속 빈도 수를 이용한 자원 예약 승인 제어 알고리즘

홍재인*, 최홍묵*, 박병수**, 최명렬*
* 한양대학교 전자전기제어계측과
** 상명대학교 컴퓨터정보통신공학부

A Resource reserved-base Admission Control Algorithm using access frequency on CM server

Jae-In Hong*, Hong-Mook Choi*, Byoung-Soo Park**, Myung-Ryul Choi*
*Dept. of EECI, Hanyang University
E-mail :{monadic1979, chmook, choimy}@asic.hanyang.ac.kr
**Dept. of Computer, Information and Telecommunication, Sangmyung University
E-mail : bpark@smuc.ac.kr

요 약

최근 컴퓨터와 통신망 기술의 급속한 발달에 따라 오디오 및 비디오 등의 연속미디어를 통한 정보 응용에 대한 연구와 연속 미디어를 다루는 연속 미디어 저장 서버에 관한 연구가 활발히 이루어지고 있다. 본 논문에서는 연속 미디어 서버의 설계와 관련한 승인 제어에 대하여, 기존의 결정적 승인 제어, 통계적 승인 제어, 자원 예약 승인 제어 알고리즘을 이용하여 새로운 방법으로 실제 서버의 자원 이용도와 접속 빈도 수에 따라 자원 양을 할당하고 자원의 할당 및 예약을 행하고 QoS를 조정하는 승인 제어 알고리즘을 제안하였다. 제안된 알고리즘은 향후 구체적 통계 모델을 통한 시뮬레이션 검증을 필요로 한다.

1. 서론

최근 컴퓨터와 통신망 기술이 급속도로 발달함에 따라, 사운드 및 동영상 등을 통해 정보를 제공하는 멀티미디어 정보 응용에 대한 연구가 활발히 진행되고 있다. 또한 이러한 멀티미디어의 핵심인 오디오나 비디오와 같은 연속 미디어(Continuous Media, CM)를 다루는 연속미디어 저장 서버에 대한 많은 연구가 이루어지고 있다.

연속미디어 저장 서버의 설계에서 고려해야할 사항으로는 1)디스크에 데이터를 배치하는 방법; 2)디스크 액세스 스케줄링; 3)서버의 버퍼관리; 4)승인 제어 등이 있다[1,2]. 이러한 사항들 중에서 승인 제어란 연속미디어 저장 서버에 새로운 사용자가 서비스를 요청하였을 때, 현재 서비스 중인 스트림에 영향을 주지 않고 서비스 할 수 있는지 없는 지를 결정하는 것이다. 기존의 승인 제어 알고리즘에는 시스템의 모든 실시간 성능 제약조건을 만족하는 범위 내에서 서비스를 허용하는 결정적 방법, 실시간 성능 제약 조건을 통계적으로 만족하는 범위 내에서 서비스를 허용하는 통계적 방법, 실시간 성능 제약 조건에 상관없이 서비스를 허용하는 방법 그리고, 남아있는 자원의 양에 따라 서비스의 허용을 결정하는 자원 예약 승인제어 알고리즘 등이 있다[3].

본 논문에서는 기존의 알고리즘들을 이용하여 새로운 서비스 요청의 승인을 위한 자원을 실제 서버의 접속 빈도 수(access frequency)에 따라 자원을 할당 및 예약하고, 서비스의 QoS 등급을 조정하는 새로운 승인 제어 알고리즘을 제안하였으며, 2장에서 승인 제어를 위한 조건과 기존의 승인 제어 알고리즘을 설명하고, 3장에서 제안한 승인 제어 알고리즘의 자원 할당량 계산방식, 접속 빈도 수를 이용한 자원 예약, QoS 등급 조정등을 설명하고 4장에서 결론을 맺는다.

2. 연속 미디어 서버의 구조와 승인 제어

연속 미디어 서버는 디지털화된 연속미디어 데이터를 디스크 배열 혹은 계층적 디스크 시스템에 저장하

며, 클라이언트는 실시간 재생을 위한 연속 미디어 스트림의 검색을 서버에 요청한다. 그림 1에서 이러한 서버의 구조를 나타낸다.

연속 미디어 저장 서버에서 서비스의 질을 보장하는 일반적인 방법으로 미리 시스템의 용량을 예약하는 방법이 있다. 이때 예약해야 할 자원으로는 CPU의 일정부분, 메모리 버퍼, 디스크 전송량, 네트워크 대역폭등이 있다[4]. 새로운 사용자가 서버에 접속하여 음악을 듣거나 영화를 보기 원할 때, 서버 혹은 네트워크가 사용자를 위한 충분한 시스템 용량을 갖고 있지 않다면, 현재 서버 사용자의 서비스 질을 저하시키지 않기 위해서 새로운 사용자의 요구를 거절한다. 따라서 멀티미디어 서버에는 더 많은 사용 요청을 처리하기 위한 승인 제어 알고리즘이 요구된다.

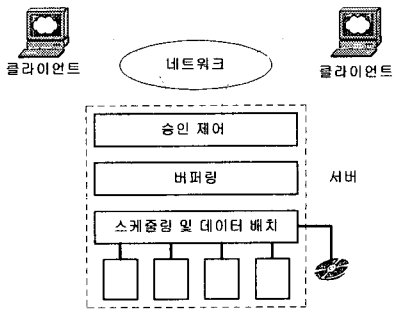

그림 1. 연속미디어 서버의 구조

아래의 그림 2는 승인 제어 알고리즘의 구성을 나타낸다. 새로 요청되는 서비스에 대한 승인 제어는 승인 제어 모듈과 QoS 처리 모듈로 구성되는 QoS 관리기에 의해 이루어진다. 이때 승인 제어기는 시스템의 I/O 대역폭과 사용 가능한 버퍼 용량을 테스트한다. 이를 통하여 서비스 요청이 일어났을 때, 서비스를 승인할지 혹은 승인하지 않을지를 결정한다.

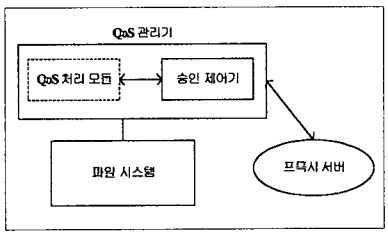

그림 2. 승인 제어의 구성

QoS 처리 모듈은 클라이언트의 요청에 의해 제공되는 입력 매개 변수 등에 따라 데이터 처리율의 관리를 담당한다. 서비스의 질을 결정하는 요소에는 탐색 시간과 회전 지연 시간이 있는데, 이러한 오버헤드로 인해 재생 시간이 길어질 경우 이 모듈은 데이터 처리율을 유지하기 위해서 몇몇 프레임을 삭제할 수 있으며, 동적인 QoS조정과 관리 등을 제어한다[5,6].

기존의 승인 제어 알고리즘에서 승인은 주어진 스트림의 실시간 성능 요구사항을 기반으로 모든 실시간 제약 조건을 만족할 때 이루어질 수 있다. 서버는 모든 실시간 제약 조건을 만족하기 위해서 탐색 시간과 회전 지연 시간에 대한 최악의 시나리오를 결정하며 이를 통해 서비스 질이 결정된다. 이 때, 최악의 시나리오를 고려하는 것이 결정적 승인 제어, 확률 값을 고려하는 것이 통계적 승인 제어 방식이다.

2.1 결정적 승인 제어

결정적 승인 제어는 각 스트림의 자원을 최악의 시나리오를 고려하여 계산하는 방식이다. 또한 새로운 스트림이 추가될 때, 현재 서비스 중인 스트림의 버퍼 사용량이 부족하지 않도록 버퍼링을 보장해야한다. 이때 버퍼 공간은 모든 스트림의 최대 사용 시간을 가정하여, 남아있는 시간에 따라 동적으로 할당되도록 한다. 따라서 승인된 서비스 사용자에게 최상의 서비스를 제공할 수 있다. 그러나 서비스의 승인에 많은 제약을 갖기 때문에 서비스 승인 비율이 낮고, 자원을 효율적으로 활용하지 못하는 단점을 갖는다.

2.2 통계적 승인 제어

통계적 승인 제어는 결정적 승인 제어 방식과 달리, 각 스트림의 자원을 확률 값을 고려한다. 따라서 새로운 사용자의 서비스 승인을 위해서 서버의 디스크에 접근하는 시간의 통계적 변화를 이용한 시스템의 통계적 모델을 고려한다. 통계적 모델은 대개 임의 분포에 따른 회전 지연시간의 확률을 바탕으로 계산된다. 이러한 확률 값은 최악의 시나리오보다 작기 때문에, 결정적 승인 제어와 달리 어느 정도의 결함을 허용하지만 서비스 승인 비율을 2배 이상 증가시킬 수 있으며, 자원을 더욱 효율적으로 이용할 수 있다.

2.3 자원 예약 승인 제어

자원 예약 승인 제어는 자원에 대한 제약 조건들을 초기화하고 혼잡을 체크하는 비트와 자원의 경계 값을 조정 가능한 값으로 설정한다[6,7]. 새로운 서비스 요청이 들어왔을 때 혼잡 비트가 '0'이 되면 서버는 기존의 서비스 질의 저하 없이 새로운 서비스를 승인할 수 있으며, 서비스 승인으로 인해 자원 혼잡이 생기면 다른 서비스의 질을 저하시킴으로써 보상한다.

결정적 승인 제어와 통계적 승인 제어의 경우에는 요청 승인의 많은 제약으로 인해 서비스 승인율이 낮고, 상대적으로 자원을 효율적으로 사용하지 못한다. 그에 비해 자원 예약 방식은 기존 서비스의 질에 영향을 줄 가능성이 있지만, 더 많은 요청을 승인할 수 있으며, 더욱 효율적으로 자원을 이용할 수 있다.

3. 새로운 승인 제어 알고리즘

3.1 접속 빈도 수에 따른 자원 할당

실제 연속 미디어 어플리케이션에서 절반 이상의 서비스 요청은 최근 영화 혹은 인기 있는 영화 등에 집중된다[8]. 새로운 승인 제어 알고리즘은 일정 기간 동안의 서버의 접속 빈도 수에 따라서 승인 제어에 요구되는 자원의 양을 계산하고, 계산된 값을 바탕으로 승인제어를 실시한다. 객체에 대한 접속 빈도 수에 대한 자료는 실제 영화 대여 웹사이트의 자료를 참고로 하여 구할 수 있고, Zipf 분포와 같은 통계 모델을 이용한다.

다음 <표 1>은 비디오 대여 정보를 알려주는 웹사이트 www.imdb.com를 참조하여 2002년도 9월 첫 번째 주에 대한 영화의 실제 접속 빈도 수를 백분율로 나타낸 모델이다.

<표 1> 영화 대여 통계 (2002. 09. 01)

object ID (Oi)	Access Freq. (Fi)	object ID (Oi)	Access Freq (Fi)	object ID (Oi)	Access Freq (Fi)
O_0	8.31	O_1	7.15	O_2	6.55
O_3	6.44	O_4	5.13	O_5	5.12
O_6	4.57	O_7	4.32	O_8	3.77
O_9	3.65	O_{10}	3.60	O_{11}	3.49
O_{12}	3.43	O_{13}	2.82	O_{14}	2.47
O_{15}	2.04	O_{16}	1.99	O_{17}	1.72
O_{18}	1.53	O_{19}	1.50	O_{20}	1.49
O_{21}	1.49	O_{22}	1.48	O_{23}	1.48
O_{24}	1.33	O_{25}	1.27	O_{26}	0.89
O_{27}	0.89	O_{28}	0.84	O_{29}	0.77
O_{30}	0.75	O_{31}	0.73	O_{32}	0.66
O_{33}	0.65	O_{34}	0.49	O_{35}	0.49
O_{36}	0.47	O_{37}	0.45	O_{38}	0.44
O_{39}	0.43	O_{40}	0.43	O_{41}	0.38
O_{42}	0.33	O_{43}	0.31	O_{44}	0.27
O_{45}	0.25	O_{46}	0.25	O_{47}	0.24
O_{48}	0.23	O_{49}	0.22		

위와 같은 객체들이 연속 미디어 서버에 VBR 미디어 스트림으로 저장되어 있다고 가정하면, 예측되는 자원 요구량은 일정 시간 동안의 자원을 샘플링하는 방식으로 산출할 수 있다. 예측되는 자원 요구량은 크게 두 가지 경우로 고려할 수 있는데, 첫 번째 경우는 최대 자원 요구량을 산출하는 것으로 식(1)에 나타내었고, 두 번째 경우는 평균적인 자원의 요구량을 산출하는 것으로 식(2)에 나타내었다.

일정한 시간 [0, T]를 기준으로 하고 시간 단위를 t_s로 할 때, 임의의 t 시간에 예상되는 자원 요구량은 다음과 같다.

$$ER_{max}(t) = max\{A(t)\} \qquad (1)$$

$$ER_{avg}(t) = \frac{\int_T A(t)dt}{t_s} \qquad (2)$$

여기서 $ER(t)$: 예상되는 자원 할당량
$A(t)$: 실제 자원 사용량
$p \in [pt_s, (p+1)t_s]$
$\forall p \in \{0,1, \ldots n-1\}$, $n = P/t_s$

그리고 <표 1>에서와 나타낸 것과 같이 각 객체의 접속 빈도 수에 따라 자원을 예약하기 위해서 다음과 같은 divisior 알고리즘을 사용한다.

STEP 0 : Let $Rj = Lj$ for j = 0, 1, 2, ..., n-1 and
minimize = Min[Sj], *maximize* = Max[Sj],
(Lj : lower bound of resource)
STEP 1 : Compute $[Fj/d(Rj)]$ for all j
Find index j' having Max $[Fj/d(Rj)]$
STEP 2 : Let $rem = S - \Sigma\ (Sj \times Rj)$
if $rem < minimize$, then output R and stop
if $rem \geq Sj$, then $Rj' = Rj' + 1$
Else find index j'' which has Max$[Fj/d(Rj)]$
among those satisfying $Sj \leq rem$ and
let $Rj'' = Rj'' + 1$;

3.2 실제 자원할당을 이용한 승인 제어 알고리즘

제안한 승인 제어 알고리즘은 일정 기간동안의 실제 자원 이용도를 바탕으로 향후 요구될 자원의 양을 계산하고, 접속 빈도 수를 이용하여 서버의 객체별로 요구되는 자원을 계산하여 서비스의 QoS를 조정하는 방식이다. 서버에 새로운 서비스 요청이 들어왔을 때 이를 승인하고 서비스하는 방법은 크게 서버에 혼잡(congest)이 발생한 경우와 발생하지 않은 경우로 나누어진다. 이 때 혼잡이라는 것은 새로운 서비스에 할당해 줄 수 있는, 서버의 자원의 양이 새로운 서비스에 의해 요구된 자원의 양보다 작은 경우를 의미한다.

기존의 방식은 위와 같은 경우 새로운 서비스의 요청을 거절하게 되지만, 제안된 승인 제어 알고리즘에서는 승인율을 높이기 위해서 QoS 조정을 통해 서비스의 질을 저하하여 서비스하도록 하였다.

먼저 혼잡이 일어나지 않은 경우에는 최대 자원 요

구 량을 기준으로 자원을 할당하여 QoS의 등급을 guaranteed service가 되도록 한다. 다음으로 혼잡이 일어나는 경우에는 QoS 레벨을 best-effort로 저하시켜서 서비스하도록 하였다. 혼잡이 일어나 QoS레벨이 저하된 이후에도, 만일 할당되는 자원의 양이 객체에 요구되는 평균 자원 양보다 적은 경우에는 서비스에 지연 혹은 분열 현상(hiccup)이 발생할 수 있으므로 요청을 거절하도록 하였다.

Step 0 : When EVENT occured,
  if ($R_{avail}(i) > Q_i$) then
    Admit the application
    Service_level : guaranteed service
    R_{alloc} <= Q_i;
    R_{used} <= R_{used} + Q_i;
    R_{avail} = R_{total} - R_{used};
    $flag_{R\_degraded}$ =0;
  else
    Reject the application

Step 1 : When EVENT occured
  if (! congested) then
    return to step 0;
  else
    if(R_i < Q_i) then
      Admit at a lower quality
      R_{alloc} <= R_i;
      R_{used} <= R_{used} + R_{alloc};
      $flag_{R\_degraded}$ =1;
    else
      Admit at a high quality
      R_{alloc} <= Q_i;
      R_{used} <= R_{used} + Q_i;

Step 3 : When $flag_{R\_degraded}$ =1
  if($R_i \geq ER_{avg}$)
    Service_level = best_effort;
  else
    reject the request;

여기서 R_{total} : 전체 자원의 양
R_{avail} : 사용할 수 있는 자원의 양
R_{used} : 사용한 자원의 양
R_{alloc} : 할당된 자원의 양
Q_i : 요청된 서비스의 자원 양
R_i : 요청된 객체의 자원 예약 할당 양
$flag_{R\_degraded}$: 서비스 저하를 나타내는 플래그

4. 결론 및 향후 연구 방향

본 논문에서는 연속 미디어 어플리케이션에서 사용자의 서비스 요청이 들어왔을 때 그에 대한 승인을 실제 자원 사용 정도에 의하여 결정하고, 시스템의 접속 빈도 수에 따라 자원을 예약해 두고 이를 바탕으로 QoS 등급을 조정하도록 하는 승인 제어 알고리즘을 제안하였다. 제안한 알고리즘은 기존의 알고리즘과 달리 부하의 불균형을 고려하여 자원을 예약함으로써 접속 빈도가 높은 객체에 대한 자원 예약 할당 정도를 높여서 부하 불균형을 고려하였으며, 예약된 자원의 양에 따라서 QoS를 조정하도록 하였다.

향후 시간에 따라 변하는 접속 빈도 수를 고려하여 자원 할당 및 예약을 동적으로 재구성할 수 있도록 할 예정이며, 접속 빈도 수에 대한 통계적 모델을 바탕으로 한 시뮬레이션을 통해 검증을 행할 예정이다.

[참고문헌]

[1] Dinkar Sitaram and Asit Dan "Multimedia Server ", Morgan Kaufmann Publishers, San Francisco.

[2] Tat Seng Chua, Jiandong Li, Beng Chin Ooi, Li an, Lee Tan. "Disk Striping Strategies for Larg e Video-on-Demand Servers", ACM Multimedia 96, Boston MA USA

[3] D. James Gemmell, Harrick M, Vin, Dilip D. Ka ndlur, P. Venkat Rangan, Lawrence A. Rowe, "Multimedia Storage Servers : A Tutorial", IEE E Computer, pp. 40-49, May 1995

[4] Andrew S. Tanenbaum, "Modern Operating Syst ems second edition" PRANTICE HALL Upper Saddle river, New Jersey 07458

[5] Xiaoye Jiang and Prasant Mohapatra. "An Aggr essive Admission Control Scheme Multimedia S ervers", Multimedia Computing and Systems pp. 620-621, IEEE international conference on 1997

[6] Wonjun Lee, Jaideep Srivastava, "Reserve-based Admission Control and Bandwidth Scheduling Strategy for Continuous Media Servers", Compu ter communication and Networ ks, pp. 232-237 IEEE, Ninth International conference on 2000

[7] Sabata, B, Chatterjee, S, Davis, M, Sydir, J.J, L awrence, T.F. "Taxonomy for QoS specification s.", Object-Oriented Real-Time Dependable Sys tems, pp.100-107 IEEE International Workshop on 1997.

[8] Seon Ho Kim. "Replication Techniques to Min-imize the Startup Latency of Continuous Media Servers." To appear in SCI 2001. July 2001.