광대역 이동 액세스 시스템에서의 실시간 및 비실시간 통합 서비스 지원을 위한 적응적 임계값 기반 패킷 스케줄링 기법

준회원 구 진 모*, 정회원 김 성 경*, 준회원 김 태 완*, 정회원 김 재 훈**, 종신회원 강 충 구*

Adaptive Delay Threshold-based Priority Queueing Scheme for Packet Scheduling in Mobile Broadband Wireless Access System

Jin-mo Ku* *Associate Member*, Sung-kyung Kim* *Regular Member*, Tae-wan Kim* *Associate Member*, Jae-hoon Kim** *Regular Member*, Chung G. Kang* *Lifelong Member*

요 약

실시간과 비실시간 서비스 클래스의 각 QoS요구 사항을 만족하면서 전체 시스템 용량을 최대화하기 위하여 DTPQ스케줄링 알고리즘이 제안된 바 있다. 기존 DTPQ방식은 개별 서비스 클래스의 QoS를 만족하면서 전체 시스템의 평균 수율을 극대화하기 위해 실시간 서비스의 HOL 패킷의 지연 시간이 정해진 임계값을 넘는 경우에만 서비스하는 접근으로서, 이를 구현하기 위해서는 주어진 환경에서 적절한 지연 임계값을 설정할 수 있어야 한다. 본 논문에서는 실시간 사용자의 채널 상태와 지연 시간에 따라 지연 임계값을 적응적으로 적용하는 A-DTPQ 기반의 스케줄링 기법을 제안한다. 이는 기존의 DTPQ알고리즘과 달리 트래픽 부하 및 채널 특성 등에 따라 지연 임계값을 적응적으로 결정하는 방식이다. 시스템 레벨 시뮬레이션을 통해 제안 방식의 고정된 최적 임계값을 가지는 DTPQ방식보다 성능이 향상시킬 수 있음을 확인하였으며, 향후 연구에서는 다양한 서비스 클래스에 대해 이와 같은 개념을 일반화할 수 있을 것이다.

Key Words : Packet scheduling, DTPQ, Adaptive delay threshold, MBWA, RT&NRT integral service

ABSTRACT

The Delay Threshold-based Priority Queueing (DTPQ) scheme has been shown useful for scheduling both real-time (RT) and non-real-time (NRT) service traffic in mobile broadband wireless access (MBWA) systems. The overall system capacity can be maximized subject to their QoS requirement by the DTPQ scheme, which takes the urgency of the RT service into account only when their head-of-line (HOL) packet delays exceed a given delay threshold. In practice, the optimum delay threshold must be configured under the varying service scenarios and a corresponding traffic load, e.g., the number of RT and NRTusers in the system. In this paper, we propose an adaptive version of DTPQ scheme, which updates the delay threshold by taking the urgency and channel conditions of RT service users into account. By evaluating the proposed approach in an orthogonal frequency division multiple access/time division duplex (OFDMA/TDD)-based broadband mobile access system, it has been found that our adaptive scheme significantly improves the system capacity as compared to the existing DTPQ scheme with a fixed delay threshold.

※ 본 연구의 일부는 대학 ITRC 및 한국과학재단 목적기초연구(과제번호: R01-2005-000-10155-0) 지원으로 수행되었습니다.
* 고려대학교 정보통신대학 전파공학과 (ccgkang@korea.ac.kr), ** SK텔레콤 네트워크 연구원 (jayhoon@sktelecom.com)
논문번호 : KICS2006-11-503, 접수일자 : 2006년 11월 21일, 최종논문접수일자 : 2007년 3월 14일

Ⅰ. 서 론

WiBro 등과 같은 광대역 이동 인터넷 액세스 시스템은 OFDMA방식의 다중 접속을 통해 OFDM (orthogonal frequency division multiplexing)고유의 주파수 다이버시티와 더불어 다중 사용자 다이버시티에 의해 이동 환경에서 효율적인 패킷 데이터 서비스를 제공할 수 있다. 이러한 시스템에서 실시간과 비실시간 서비스 클래스의 각 QoS요구 사항을 만족하면서 전체 시스템 용량을 최대화하기 위하여 DTPQ스케줄링 알고리즘이 제안된 바 있다[1].

기존 DTPQ방식은 개별 서비스 클래스의 QoS를 만족하면서 전체 시스템의 평균 수율을 극대화하기 위해 실시간 서비스의 HOL 패킷의 지연 시간이 정해진 임계값을 넘는 경우에만 서비스하는 접근으로서, 이를 구현하기 위해서는 주어진 환경에서 적절한 지연 임계값을 설정할 수 있어야 한다. 본 논문에서는 실시간 사용자의 채널 상태와 지연 시간에 따라 지연 임계값을 적응적으로 적용하는A-DTPQ 기반의 스케줄링 기법을 제안한다. 이를 위해 허용 가능한 지연시간에 임박한 실시간 사용자에 대한 urgency metric을 정의하고, 이때 트래픽 부하와 각 사용자별 채널 상태를 반영함으로써 시스템의 수율 극대화하고자 한다. OFDMA/TDD 기반의 시스템 레벨 시뮬레이션을 통하여 각 서비스 클래스의 QoS 성능 분석을 수행하고, 이를 통해 제안된 기법과 기존 DTPQ 기반 기법간의 성능을 비교한다.

본 논문의 2장에서는 기존 DTPQ기반 스케줄링 알고리즘을 중심으로 본 연구의 배경을 살펴본다. 그리고 3장에서는 제안하는 기법을 설명하고, 4장에서 이에 대한 성능 분석 결과를 제시한다. 마지막으로 5장에서는 본 논문의 결론을 맺는다.

Ⅱ. 문제 개요 및 접근 방법

2.1 기존 방식 : DTPQ 기법

실시간 서비스에서는 허용 가능한 최대 지연 시간 $W_{\max}$을 초과하는 경우 패킷 손실이 발생하며, 해당 QoS 요구사항은 패킷 손실률에 의해 규정될 수 있다. 따라서, 실시간 서비스 사용자의 HOL패킷 지연 시간이 $W_{\max}$에 근접해짐에 따라 실시간 서비스 사용자의 우선권이 증가해야 한다. 그러나 만약 우선권 할당이 너무 일찍 이루어지는 경우에는 실시간 서비스 클래스의 QoS요구 사항이 불필요할 만큼 과도하게 만족되어 결과적으로 비실시간 서비스 클래스의 성능이 저하될 수 있다. DTPQ 기반의 스케줄링 알고리즘은 실시간 서비스의 HOL 패킷에 대해 지연 임계값(delay threshold)을 둠으로써 우선적으로 할당되어야 하는 서비스 클래스를 결정하는 방식이다.

그림 1에서 보는 바와 같이 지연 임계값은 $kW_{\max}$로 주어지게 되며, 여기서 k는 실시간 서비스 클래스의 우선권을 결정하는 제어 파라미터이다. 지연 임계값 $kW_{\max}$는 $W_{\max}$보다 작아야 하므로 $0 \le k \le 1$이다. 즉, 실시간 서비스 클래스 사용자의 HOL 패킷 지연 시간이 $kW_{\max}$를 넘지 않는 한 비실시간 서비스 클래스 사용자들이 우선적으로 서비스되며, 그렇지 않은 경우에는 실시간 서비스 클래스 사용자들이 우선적으로 서비스된다.

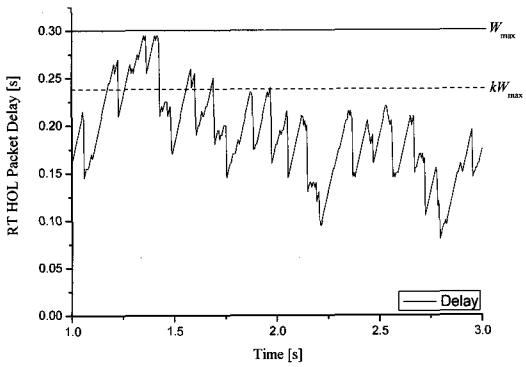


그림 1. 실시간 서비스의 HOL 패킷 지연시간 : 예시

한편 각 서비스 클래스에 대하여 opportunistic scheduling을 적용하기 위하여 적절한 priority metric이 정해진다. 즉, 같은 서비스 클래스 내의 모든 사용자들에 대해서는 주어진 priority metric을 기반으로 우선권이 결정된다. 본 논문에서는 실시간 서비스 클래스와 비실시간 서비스 클래스를 위하여 Proportional Fair (PF) 스케줄링 알고리즘을 고려한다[2]. PF 알고리즘의 경우에는 일반적으로 공정성을 고려한 best effort 방식의 비실시간 서비스에 적합하나, DTPQ 기법의 priority queueing에 의해 클래스간에 차별화된QoS 지원이 가능하므로 동일한 클래스 내의 사용자가 공정성을 유지하면서 지연시간을 우선적으로 만족할 수 있다.

2.2 QoS 및 성능 척도

실시간 서비스 클래스의 성능은 패킷 손실률로서 나타낼 수 있으며, 이는 전체 전송된 패킷과 폐기된 패킷의 비율로서 정의된다. 실시간 서비스의 경우에

는 버퍼에서 대기하는 패킷이 시스템에서 정의된 최대 허용 가능한 지연시간 $W_{\max}$를 초과하는 경우 해당 패킷은 폐기되며, 이와 같이 목표 성능을 만족하지 못했을 때 해당 패킷을 폐기하는 형태의 품질을 Hard QoS라고 한다. U_{rt}와 U_{nrt}를 각각 실시간 서비스와 비실시간 서비스 클래스에 속한 사용자의 집합이라고 하자. 이때, 프레임 t까지 사용자 i가 전송한 패킷의 총 수를 $J_i(t)$라 하고, 사용자 i의 j 번째 패킷의 지연시간을 $W_i^j(t)$라고 하자. 이때, 임의의 프레임 t에서 사용자 i의 패킷 손실률은 다음과 정의할 수 있다.

$$L_i(t)=\frac{\sum_{j=1}^{J_i(t)} U(W_i^j(t)-W_{\max})}{J_i(t)}, i\in U_{rt} \quad (1)$$

여기서, $U(x)$는 unit step function을 나타낸다. 사용자별 최대 허용 가능한 패킷 손실률을 $L_{\max}$라 할 때, 프레임 t에서 시스템의 실시간 서비스 불능 확률은 전체 실시간 서비스 사용자와 $L_{\max}$를 넘는 사용자수의 비로서 정의하며, 다음과 같이 나타낼 수 있다.

$$P_{out}^{(rt)}(t)=\frac{1}{N_{rt}}\sum_{i=1}^{N_{rt}} U(L_i(t)-L_{\max}) \quad (2)$$

여기서, N_{rt}는 시스템 내의 유효 실시간 서비스 사용자 수를 나타낸다($N_{rt}=|U_{rt}|$). 실시간 서비스 클래스에 대한 시스템 평균 서비스 불능 확률은 다음과 같다.

$$\overline{P}_{out}^{(rt)}=\lim_{t\to\infty}\frac{1}{t}\int_0^t P_{out}^{(rt)}(\tau)d\tau \quad (3)$$

본 논문에서는 비실시간 서비스 클래스에 대해서는 최소 요구 데이터 전송률 $R_{\min}$이 주어진다고 가정한다. 하지만 실시간 서비스에서 고려되는 hard QoS와 달리 $R_{\min}$을 만족하지 못하더라도 해당 패킷을 폐기하지 않으며, 이와 같은 품질 특성을 soft QoS라고 한다. 프레임 t에서 사용자 i의 데이터 전송률을 $R_i(t)$라 하면 OFDMA 시스템의 경우 부채널(subchannel) 집합이 프레임 단위로 사용자들 사이에 공유되므로 만약 사용자 i의 현재 전송률과 부채널의 수를 각각 $r_i(t)$와 $s_i(t)$라 했을 때, $R_i(t)$는 다음과 같다.

$$R_i(t)=\frac{1}{T}\sum_{t=t_0}^{t_0+T} r_i(t)\frac{s_i(t)}{S_d}, i\in U_{nrt} \quad (4)$$

여기서, S_d는 하향 링크 부프레임에서 사용 가능한 부채널의 총 개수이고, T는 프레임 단위로 주어지는 시간 윈도우의 크기이다. 따라서 평균 데이터 전송률은 $\overline{R}_i=\lim_{T\to\infty} R_i(t)$ 이다. Hard QoS에서와 유사한 방법으로 soft QoS요구 사항도 평균 서비스 불능 확률로 표현할 수 있으며, 다음과 같이 주어진다.

$$\overline{P}_{out}^{(nrt)}=\frac{1}{N_{nrt}}\sum_{i=1}^{N_{nrt}} U(R_{\min}-\overline{R}_i) \quad (5)$$

여기서, N_{nrt}는 시스템내의 유효 비실시간 사용자수를 나타낸다($N_{nrt}=|U_{nrt}|$).

2.3 문제 개요

본 논문에서 제안하는 방식의 스케줄링 알고리즘의 목적은 최소 서비스 손실 성능을 만족시키는 사용자, 즉 $\overline{P}_{out}^{(rt)}\le\overline{P}_{\max}^{(rt)}$과 $\overline{P}_{out}^{(nrt)}\le\overline{P}_{\max}^{(nrt)}$을 만족하는 전체 사용자 수를 최대화하는 것이다. 즉, 지연 임계값을 고정하지 않고 트래픽 부하와 개별 사용자의 채널 상황에 따라 급박하지 않은 실시간 서비스 사용자의 전송을 최대한 지연함으로써 비실시간 서비스 클래스의 서비스 불능 확률을 최소화할 수 있다. 프레임 t에서의 지연 임계값을 k_t라 하고 이때 실시간과 비실시간 서비스 클래스의 사용자 수를 각각 $N_{RT}(k_t)$와 $N_{NRT}(k_t)$로 나타낸다. 전체 서비스 매출(revenue)을 최대화하기 위해 사용자별 평균 revenue (average revenue per user: ARPU)는 서비스 클래스에 따라 다르게 가중치를 주어야 한다. 본 논문에서 실시간과 비실시간 사용자의 weighting factor를 각각 w_{rt}와 w_{nrt}라 하면 전체 서비스 revenue는 다음과 같이 주어진다.

$$f(k_t)=w_{rt}N_{RT}(k_t)+w_{nrt}N_{NRT}(k_t) \quad (6)$$

따라서, 실시간과 비실시간 서비스 클래스의 서비스 불능 확률을 만족하면서 동시에 목적 함수 (6)식을 최대화하는 제어 파라미터 k_t가 정해져야 한다. 즉, 매 프레임마다 다음의 최적화 문제를 풀어야 한다.

$$\max_{0\le k_i\le 1}\{w_{rt}N_{RT}(k_i)+w_{nrt}N_{NRT}(k_i)\} \quad (7)$$

$$\text{subject to } \begin{cases} P_{out}^{(rt)}(k_i)\le P_{\max}^{(rt)}, \forall i\in U_{rt} \\ P_{out}^{(nrt)}(k_i)\le P_{\max}^{(nrt)}, \forall i\in U_{nrt} \end{cases}$$

Ⅲ. Adaptive DTPQ 기반 패킷 스케줄링 알고리즘

[1]에서 제안된 기존 DTPQ기법은 실시간과 비실시간 사용자의 비율에 따라 최적의 지연 임계값을 다르게 설정하여야 한다. 즉, (7)식에서 보는 바와 같이 최적의 임계값 k_i는 매 프레임마다 사용자 수와 채널 상태, 그리고 QoS 요구사항 등에 따라 결정되어야 하나, (7)식의 최적화를 실시간으로 수행하는 것은 매우 어렵다.

본 논문에서는 시스템의 상태에 따라 DTPQ 기법의 지연 임계값을 적응적으로 변경할 수 있는 방안을 고려한다. 즉, 이것은 기존 DTPQ기법의 적응적인 형태이다. 이전 프레임에서 설정된 지연 임계값은 사용자 i의 HOL패킷 지연 시간인 $W_i^{hol}(t)$와 현재 전송률 $r_i(t)$를 이용하여 현재 프레임 t에서 재설정된다. 즉, 실시간 서비스 패킷의 지연 시간과 현재 채널 상태에 따라 임계값을 증가 또는 감소시켜야 한다. 실시간 사용자들의 평균적인 채널 상태가 나쁘거나 서비스 불능의 위험에 처할 경우에는 비실시간 사용자보다 더 급박하므로 임계값을 감소시켜야 한다. 한편, 실시간 사용자의 전체 잔여 수명이 짧아지면 지연 임계값을 낮추어야 하며, 이는 모든 실시간 사용자의 평균 지연시간에 의해 반영될 수 있다. 따라서, 각 프레임에서 시스템 전체의 긴급성을 지연 임계값에 반영하기 위한 긴급도(urgency metric)을 다음과 같이 정의한다.

$$c(t)=\frac{\sum_{i=1}^{Y_{rt}(t)}\frac{W_i^{hol}(t)}{W_{i,\max}}}{\frac{1}{Y_{rt}(t)}\sum_{i=1}^{Y_{rt}(t)}r_i(t)} \quad (8)$$

여기서, $Y_{rt}(t)=\sum_{i\in U_{rt}}U(W_i^{hol}(t))$로서, 이는 시간 t의 프레임에서 존재하는 실시간 사용자의 총 수를 나타낸다. 이와 같인 정의된 긴급도 값은 HOL패킷을 가진 실시간 사용자들에 대해 지연시간 $W_i^{hol}(t)$와 $W_{i,\max}$의 비를 이용하여 상대적인 긴급성을 반영하고, 또한 이들에 대한 평균 전송률을 통해 시스템 전체의 긴급성을 반영할 수 있도록 한 것이다. 이와 같이 정의된 긴급도에 의해 매 프레임 단위로 전체 시스템 관점에서 실시간 서비스의 긴급성이 정량화되고, 이 값에 따라 지연 임계값을 동적으로 갱신하고자 한다. 즉, $c(t)<c(t-1)$인 경우는 실시간 서비스의 긴급성이 낮아졌다고 간주하고 지연 임계값을 증가시키며, 그 반대인 경우에는 감소를 시킨다. 이와 같은 증감 과정을 통해 매 프레임에서 지연 임계값을 다음과 같이 갱신할 수 있다.

$$k_i=\begin{cases} k_{i-1}+\beta\Delta, & c(t)<c(t-1)\text{or } Y_{rt}(t)=0 \\ k_{i-1}-\gamma\Delta, & otherwise \end{cases} \quad (9)$$

여기서 β와 γ는 매 프레임 별로 설정되는 지연 임계값의 증감을 위한 상수로서, 다음과 같이 실시간 사용자와 비실시간 사용자의 비율에 의해 결정되도록 한다.

$$\beta=\frac{N_{nrt}}{N_{nrt}+N_{rt}} \quad (10)$$

$$\gamma=\frac{N_{rt}}{N_{nrt}+N_{rt}} \quad (11)$$

이는 실시간 사용자의 수가 비실시간 사용자의 수보다 많은 경우 지연 임계값을 더 크게 감소 시키고, 반대의 경우 지연 임계값을 더 크게 증가 시킴으로써 실시간 사용자와 비실시간 사용자의 비에 따라 증감의 단위에 가중치 역할을 하게 된다. 이는 긴급도 $c(t)$가 실시간 트래픽의 증감에만 반응하고 실시간 트래픽의 양에 반영할 수 있는 요소가 없기 때문에 이를 보완하기 위한 것이다. 한편, (9)식에서 Δ는 매 프레임 별로 설정되는 지연 임계값 증감분의 기본 단위로서, 다음과 같이 전체 허용 가능 지연시간 중에서 지연 임계값 증감이 적용된 시간의 비율로 설정한다.

$$\Delta=\frac{Frame\,Time\times N_{frame}}{W_{i,\max}} \quad (12)$$

여기서, $Frame\,Time$은 프레임 구간 길이를 나타내고, N_{frame}은 지연 임계값 증감 발생 횟수를 프레임 단위로 나타낸 것이다. $\Delta<1$이며, $W_{i,\max}$ 에 비해 Δ가 너무 작으면 지연 임계값의 증감이 시스템의 긴급성을 따라 가지 못하고, 너무 크면 시스템의 긴급성에 대한 과도한 반응으로 적응 기법의 성능을

얻을 수 없다. 한편, 초기 임계값은 $k_0 = \frac{N_{nrt}}{N_{rt}+N_{nrt}}$ 로 설정한다.

이와 같이 식(9)에 의해 지연 임계값 k_t 이 정해지면, 지연 시간이 $k_t W_{i,\max}$ 을 넘는 사용자들은 다른 모든 사용자들보다 먼저 서비스를 받게 된다. 즉, 가장 먼저 서비스를 받게 되는 사용자들의 집합을 다음과 같이 나타낼 수 있다.

$$S_{rt}(k_t) = \{i \mid W_i^{hol}(t) > k_t W_{i,\max}\} \subset U_{rt} \qquad (13)$$

즉, $S_{rt}(k_t)$에 속하지 않는 사용자들은 서비스를 지연시켜 비실시간 사용자를 서비스 함으로써 비실시간 사용자의 서비스 불능 확률 목표를 유지하면서 전체 시스템의 수율을 극대화할 수 있다.

Ⅳ. 성능 분석

4.1 시스템 모델

본 논문에서 광대역 이동인터넷 액세스 망의 한 예로서 국내에서 상용화를 예정하고 있는 WiBro규격을 고려한 시스템 레벨 시뮬레이션을 통해 제안 방식의 성능을 평가하고자 하며, 이에 따라 대역폭은 9MHz이고 768개의 유효 부반송파를 갖는 OFDMA/TDD시스템을 고려한다[3]. 각 OFDM심볼의 길이는 115.2μs이다. 그리고 한 프레임의 길이는 5ms이며, 하향 링크와 상향 링크의 심볼 비율은 24:12인 비대칭적인 프레임 구조를 갖는다. 따라서 한 프레임 동안 하향 링크부채널의 개수는 384개 (768개 부반송파* 1개 부채널/48개 부반송파/심볼 * 24개 심볼)이다. 본 분석에서는 AMC모드에 따른 사용자의 요구 SNR값을 모든 프레임마다 계산하여 반영하였다. 채널 모델은 ITU-R M.1225의 Pedestrian-A와 Vehicle-A를 사용하였으며, 각 채널 모델별 MCS레벨은 표 1과 같이 가정하였다. 본 시뮬레이션에서 사용된 시스템 파라미터를 요약하면 표 2와 같다.

표 1. AMC에 따른 MCS Table

MCS		Required C/I (dB)	
		Ped-A: 3km/h	Veh-A: 30km/h
QPSK	1/12	-2.45	-2.82
QPSK	1/6	-0.14	-0.43
QPSK	1/3	2.61	2.55
QPSK	1/2	5.36	5.34
QPSK	2/3	9.2	8.72
16-QAM	1/2	10.64	10.58
16-QAM	2/3	14.65	14.15
64-QAM	2/3	18.97	19.17
64-QAM	5/6	24.75	24.50

표 2. 시스템 파라미터

Parameter	Value	Note
Cell layout	Hexagonal cell/3 tiers	
Number of sectors	3 sectors	
Mobile speed	3km/h, 30 km/h	
Cell radius	1 km	
Antenna patterns	$A(\theta) = -\min\left[12\left(\frac{\theta}{\theta_{3dB}}\right)^2, A_m\right]$ $-180 \le \theta \le 180$	θ_{3dB}= 70 A_m= 20dB
BS Tx. Power	20 W	
Path loss	$L = 128.1 + 37.6\log_{10} d$	d: distance to BS
Standard deviation for log-normal shadowing	σ_s^2 = 8dB	
Correlation distance	50 m	

4.2 트래픽 모델

본 분석에서 실시간 서비스는 비디오 스트리밍을 고려했으며, 이에 대한 트래픽 모델은 [4]를 따른다. 하나의 비디오 스트리밍 세션 동안 프레임 전송률에 따라 일정한 간격으로 프레임이 도착한다. 본 논문에서는 10fps (frame per second)를 고려하고 있으므로 비디오 프레임의 도착 간격은 100ms이다. 그리고 각 프레임은 고정된 수의 슬라이스(slice)로 구성되며, 각 슬라이스는 하나의 패킷을 발생시키고 패킷 도착 간격과 각 패킷의 크기는 truncated Pareto분포를 따른다고 가정한다. 표 3에서 비디오 스트리밍 트래픽 모델에서 사용되는 각종 파라미터를 나타낸 것이다. 한편, 비실시간 서비스로서 FTP 트래픽을 고려하며, 이는 일정한 간격으로 일정한 길이의 패킷이 발생하는 것으로 가정한다. 이에 따른 트래픽 모델의 구체적인 파라미터는 표 4를 따른다.

표 3. 트래픽 모델 파라미터: Video Streaming

Category	Distribution	Parameters
Slice size	Truncated Pareto	$\vartheta = 1.2$, $\mu_{\min} = 20$ byte, $\mu_{\max} = 125$ byte
Slice inter-arrival time	Truncated Pareto	$\vartheta = 1.2$, $\mu_{\min} = 2.5$ ms, $\mu_{\max} = 12.5$ ms

표 4. 트래픽 모델 파라미터: FTP

Information types	Distribution	Parameters
Packet call size	Deterministic	400 bytes
Inter-arrival time between packet calls	Deterministic	50 ms

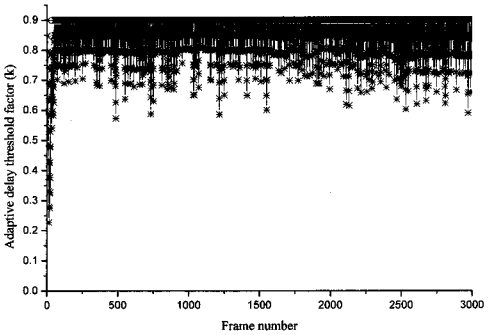
4.3 시뮬레이션 시나리오

다중 셀 환경에서 제안된 스케줄링 알고리즘에 대한 성능을 분석하기 위해 시스템 레벨 시뮬레이션을 수행한다. 각 사이트는 이상적인 육각형 셀로 모델링하며, 각 셀은 3개의 섹터로 분할한다. 기준 셀을 중심으로 모두 3개 층(tier)까지 고려하며, 이에 따라 총 19개의 셀을 동시에 고려한다. 이때 셀 경계 점 영향을 제거하고 실제 상황과 유사한 간섭 수준을 반영하기 위해 셀들이 3차원 공간에서 서로 연결된 wrap-around 구조를 적용한다[5]. COST-231 경로 손실 모델과 log-normal shadow fading 모델을 적용하였으며, 단말과 기지국간의 거리에 따른 상관 관계를 반영하기 위해 상관 거리(correlation distance)를 고려하였다.

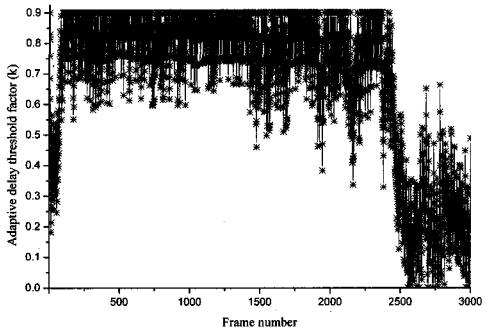
실시간 서비스 클래스의 최대 허용 가능한 지연 시간은 200msec이며, 비실시간 서비스 클래스의 최소 요구 데이터 전송률을 64kbps로 설정하였다 ($R_{\min}$= 64kbps). 그리고 서비스 클래스별 최대 허용 가능한 시스템 평균 서비스 불능 확률을 $P_{\max}^{(rt)} = P_{\max}^{(nrt)} = 0.1$로 하였다. 각 시뮬레이션은 3,000개 프레임에 대해서 수행하며, 최종 결과는 모든 셀에 대해 평균을 취한 것이다.

4.4 시뮬레이션 결과 및 검토

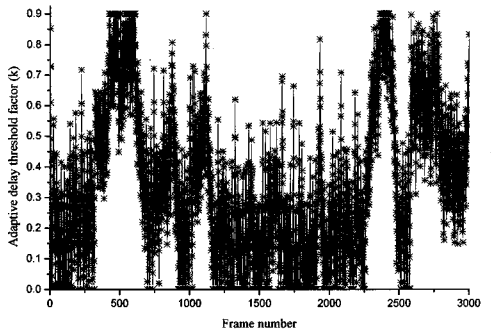
본 시뮬레이션에서는 실시간 서비스 사용자의 수를 고정하고 비실시간 서비스 사용자의 수를 증가해 가면서 각 서비스 클래스별 시스템 평균 서비스 불능 확률과 시스템 평균 데이터 수율을 살펴본다. 또한, 비실시간 서비스 사용자의 수를 고정하고 실시간 서비스 사용자의 수를 증가해 가면서 각 서비스 클래스별 시스템 평균 서비스 불능 성능을 고찰



(a) $N_{rt} = 10$



(b) $N_{rt} = 15$



(c) $N_{rt} = 20$

그림 2. 실시간 사용자 수에 따른 적응 지연 임계값의 변화

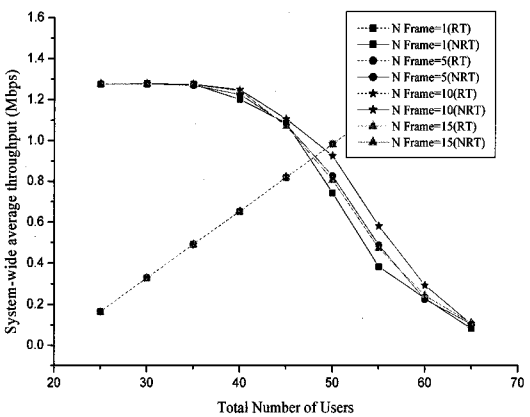


그림 3. N_{frame}에 따른 평균 수율 성능의 비교: $N_{nrt} = 20$

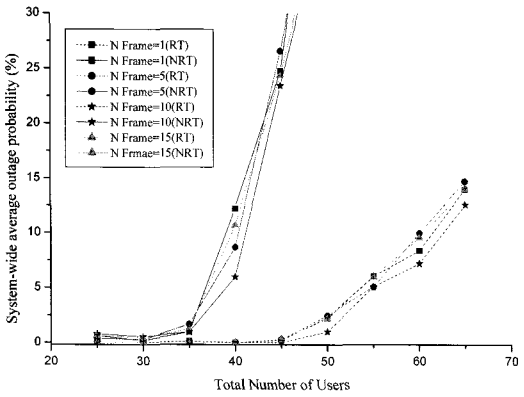


그림 4. N_{frame}에 따른 평균 서비스 불능 확률의 비교: $N_{nrt}=20$

한다.

그림 2에서는 실시간 사용자의 수에 따른 adaptive delay threshold factor (k) 값의 변화를 예시한 것이다. 예상한 바와 같이 실시간 사용자의 수가 증가할수록 k값은 상대적으로 낮은 값으로 분포되는 것을 확인할 수 있다. 그러나, 실제로는 트래픽의 분포와 사용자 채널의 상태에 따라 능동적으로 적응하고 있음을 보여준다.

그림 3과 4에서는 비실시간 사용자의 수를 $N_{nrt}=20$으로 고정하고 실시간 사용자의 수를 증가시키면서, $W_{\max}=0.2$ second의 지연 요구사항을 갖는 트래픽 모델에서 N_{frame}의 값에 따른 평균 수율 성능과 서비스 불능 확률을 비교하였다. 실시간 사용자 수가 증가하면서 동시에 수용할 수 있는 비실시간 사용자의 수율이 감소하는 것을 볼 수 있다. 이때 N_{frame}값이 증가하면서 비실시간 사용자의 수율이 증가하는 것을 볼 수가 있다. 반면, 실시간 사용자의 경우에는 N_{frame}값의 영향을 받지 않는 것을 알 수 있다. 그런데 N_{frame}이 과도하게 큰 경우, 즉 $N_{frame}=15$에서는 오히려 비실시간 성능의 향상이 떨어지게 된다. 이는 N_{frame}의 크기가 크면 그에 따른 적응 지연 임계값 k의 변화의 폭이 커지게 되어 과도한 변동으로 인해 실시간에는 영향이 없으나 비실시간의 성능 향상에 영향을 미치기 때문이다. 본 시뮬레이션을 통해 $0.1 \le \Delta \le 0.2$의 범위가 적절하며, 이 때 $W_{\max}$의 10% ~ 15%에서 가장 높은 성능을 낼 수 있음을 알 수 있다.

다음으로 그림 5와 6에서는 $N_{rt}=20$일 때 허용 가능한 지연시간을 각각 $W_{\max}=0.2$sec과 $W_{\max}=5$sec로 설정하고, 비실시간 사용자의 수를 변화하면서 Priority Queuing 방식, 지연 임계값에 따른 DTPQ방식, 그리고 제안한 A-DTPQ방식에 대해서 실시간 및 비실시간 사용자의 평균 데이터 수율을 나타낸다. 여기서 보는 바와 같이 Priority Queuing 방식보다는 지연 임계값에 따른 DTPQ 방식이 상대적으로 더 높은 수율 성능을 보임을 알 수 있고, 고정된 지연 임계값을 갖는 DTPQ 방식보다는 실시간 사용자의 긴급성과 채널상태를 반영하여 지연 임계값을 적응적으로 갱신하는 A-DTPQ방식을 통해 수율 성능이 향상될 수 있음을 알 수 있다. 이는 제안하는 방식이 실시간 사용자의 급박함과 채널 상태를 고려하여 지연 임계값을 적응적으로 설정한다는 것을 알 수 있으며 유휴 자원의 효율성이 높다는 것을 알 수 있다. 사용자의 수가 충분히 클 때, $W_{\max}=0.2$sec의 경우에는 A-DTPQ 방식에 의해 기존의 DTPQ 방식보다 0.24Mbps, 약 13%의 수율 향상을 볼 수 있다. 한편, $W_{\max}=5$sec의 경우에는 0.5Mbps, 약 25%의 수율 향상이 있음을 나타낸다. 일반적으로 $W_{\max}$가 클수록 A-DTPQ 방식의 성능이 더 향상되는 것을 볼 수 있다. 이는

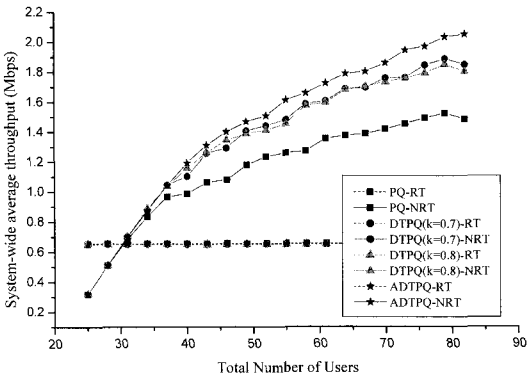


그림 5. 평균 수율 성능 비교: $N_{rt}=20$, $W_{\max}=0.2$sec

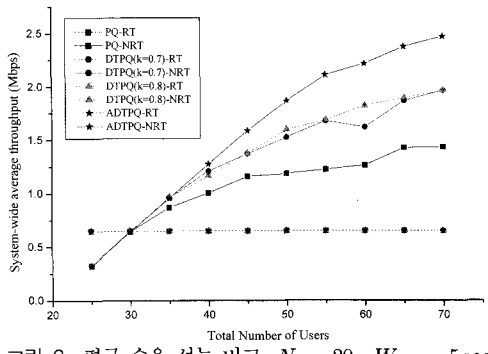


그림 6. 평균 수율 성능 비교: $N_{rt}=20$, $W_{\max}=5$sec

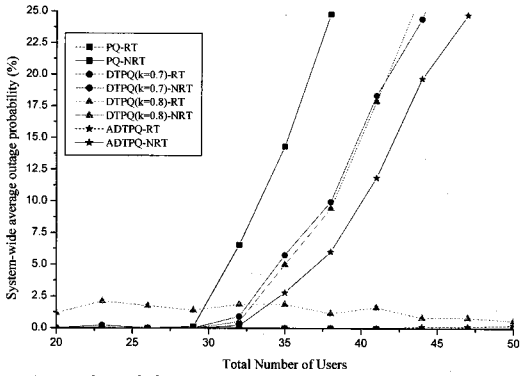


그림 7. 평균 서비스 불능 확률: $N_{rt}=15$, $W_{\max}=0.2\text{sec}$

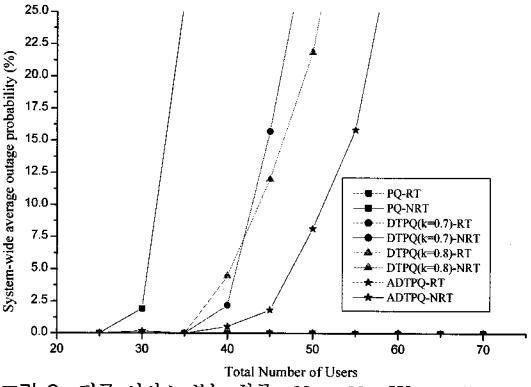


그림 8. 평균 서비스 불능 확률: $N_{rt}=20$, $W_{\max}=5\text{sec}$

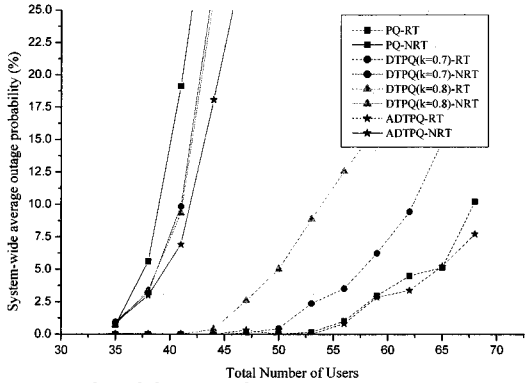


그림 9. 평균 서비스 불능 확률: $N_{nrt}=30$, $W_{\max}=0.2\text{sec}$

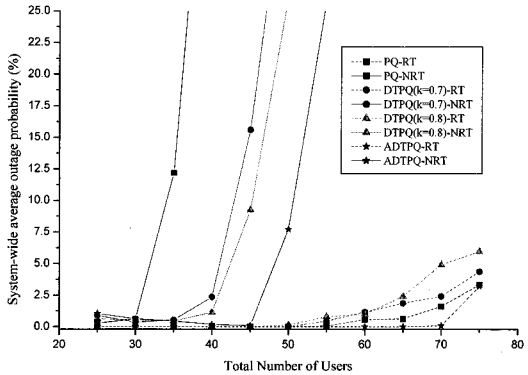


그림 10. 평균 서비스 불능 확률:$N_{nrt}=20$, $W_{\max}=5\text{sec}$

$W_{\max}$가 클수록 실시간 트래픽을 지연할 수 있는 여유시간이 늘어나 비실시간 트래픽을 보다 기회적으로 수용할 수 있기 때문이며, 임계값을 적응적으로 적용할 경우 이와 같은 효과는 더욱 더 강력하게 나타난다.

그림 7 및 8에서는 실시간 서비스 사용자를 고정하고 비실시간 사용자의 수를 변화하면서 각 방식에 대한 서비스 불능 확률을 보여준다. 수율 성능과 마찬가지로 서비스 불능 확률에 대한 성능에서도 A-DTPQ, DTPQ, Priority Queuing방식의 순으로 성능이 높음을 알 수 있다. 또한 DTPQ방식에서 k = 0.8로 고정했을 경우 실시간 사용자의 서비스 불능확률이 약 2%정도 발생하며, 이는 실시간 사용자의 성능을 적절히 희생하여 비실시간 사용자의 용량을 증대시키는 것을 알 수 있다. 또한, DTPQ 방식에서는 실시간과 비실시간 사용자의 구성 비율이 달라짐에 따라 QoS요구 사항을 만족하기 위한 최적의 임계값이 다르게 설정되어야 함을 알 수 있다. 하지만 A-DTPQ방식의 경우 사용자 수에 관계없이 항상 DTPQ 방식보다 나은 성능을 보여준다. 최대 허용 가능한 시스템 평균 서비스 불능 확률을 $P_{\max}^{(rt)}=P_{\max}^{(nrt)}=0.1$로 설정했을 때, 주어진 실시간 서비스 사용자에 대하여 목표 성능을 만족하면서 수용할 수 있는 최대 비실시간 사용자의 수를 비교할 수 있다. 그림 7에서 보는 바와 같이 $N_{rt}=15$및 $W_{\max}=0.2\text{sec}$의 경우에 수락 가능한 비실시간 사용자의 용량은 약13%의 증가를 보여준다. 한편, 그림 8에서 보는 바와 같이, $N_{rt}=20$ 및 $W_{\max}=5\text{sec}$의 경우에는 비실시간 사용자의 용량이 22명에서 31명으로 증가하며, 이는 약 40% 의 용량 증가에 해당한다.

그림 9와 그림 10에서는 비실시간 사용자의 수를 고정하고 실시간 서비스 사용자의 수를 증가하면서 각 서비스 클래스 별 시스템 평균 서비스 불능 확률을 보여 주고 있다. 여기서 실시간 서비스에 대한 서비스 불능 확률을 살펴 보면 A-DTPQ 방식이 실시간 서비스를 항상 우선시 하는 Priority Queuing 방식과 유사한 수준의 성능을 유지하면서 비실시간 서비스의 불능 성능을 향상 시키는 것을 알 수 있다.

그림 9에서는 $N_{nrt}=30$으로 고정하고 $W_{\max}=0.2$sec에 대해서 실시간 사용자에 대한 허용 가능한 시스템의 평균 서비스 불능 확률 $P_{\max}^{(rt)}=0.1$을 만족하는 사용자의 수가 A-DTPQ방식에서는 40명이며, DTPQ 방식의 경우에는 k = 0.8일 때와 k= 0.7일 때 각각 24명 및 33명 가량이 된다. 이는 DTPQ 방식에 비해 각각 67%와 약21%의 용량 증가에 해당한다. 한편, 그림 10에서는 $N_{nrt}=20$으로 고정하고 $W_{\max}=5$sec에 대해서 평균 서비스 불능 확률을 보여준다. 이때 $P_{\max}^{(rt)}=0.1$을 기준으로 A-DTPQ방식은 DTPQ 방식보다 약 40%의 용량 증대를 가져오는 것을 확인할 수 있다.

V. 결 론

본 논문에서는 실시간과 비실시간 서비스를 통합 지원할 때 실시간 서비스 사용자를 우선적으로 고려하는 Priority Queueing 기반 스케줄링 알고리즘의 한계를 극복하고 광대역 이동 접속 시스템의 특성을 이용하는 기존 DTPQ기법의 성능을 극대화할 수 있는 접근 방법을 제시하였다. 이는 기존의 DTPQ알고리즘과 달리 트래픽 부하 및 채널 특성 등에 따라 지연 임계값을 적응적으로 결정하는 방식이다. 시스템 레벨 시뮬레이션을 통해 제안 방식의 고정된 최적 임계값을 가지는 DTPQ방식보다 성능이 향상시킬 수 있음을 확인하였으며, 향후 연구에서는 다양한 서비스 클래스에 대해 이와 같은 개념을 일반화할 수 있을 것이다.

참 고 문 헌

[1] D.H. Kim and C.G.Kang, "Delay Threshold-based Priority Queueing Packet Scheduling for Integrated Services in Mobile Broadband Wireless Access System," *Lecture Note in Computer Science, LNCS* 3726, September 2005.

[2] A. Jalali, R. Padovani and R. Pankaj, "Data Throughput of CDMA HDR a High Efficiency-High Data Rate Personal Communication Wireless System," *Proceedings of IEEE VTC 2000,* pp. 1854-1858, Vol. 3, 2000.

[3] 한국정보기술협회, "2.3GHz 휴대인터넷 표준 - 매체 접근 제어 계층" *정보통신단체표준 PG302*, Jun, 7, 2004.

[4] 3GPP2, "cdma2000 Evaluation Methodology," 2005

[5] Y.-B. Lin and V. W. Mak, "Eliminating the Boundary Effect of Large-Scale Personal Communication Service Network Simulation," *ACM Transaction on Modeling and Computer Simulation,* pp.165~190, Vol. 4, No. 2, April 1994.

구 진 모 (Jin-mo Ku) 준회원

2004년 2월 : 고려대학교 전자공학과 졸업

2006년 2월 : 고려대학교 전파공학과 석사

2006년 2월~현재 : LG전자 통신연구소 연구원

<관심분야> 광대역 무선전송 기술 및 매체접근 제어 프로토콜 설계/구현

김 성 경 (Sung-kyung Kim) 정회원

1999년 2월 : 고려대학교 전자공학과 졸업

2001년 2월 : 고려대학교 전파공학과 석사

2001년 3월~현재 : 한국전자통신연구원 연구원

<관심분야> 광대역 무선 통신 시스템 성능 분석, 매체접근 제어 프로토콜 설계/구현, 4세대 이동통신

김 태 완 (Tae-wan Kim) 준회원

1998년 2월 : 광운대학교 전자공학과 졸업

1998년 3월~현재 : 삼성탈레스 기술연구소 책임연구원

2006년 2월~현재 : 고려대학교 전파공학과 석사과정

<관심분야> 매체접근 제어 프로토콜 설계/구현, WiBro, MMR

김 재 훈 (Jae-hoon Kim) 정회원

1996년 2월 : KAIST 산업경영학과 졸업
1998년 2월 : KAIST 테크노경영/정보통신 석사
2003년 8월 : KAIST 테크노경영/정보통신 박사
2003년 2월~2005년 6월 : 삼성전자 네트워크 사업부 책임연구원
2005년 6월~현재 : SK 텔레콤 전략기술 부문 Manager
<관심분야> CDMA/WCDMA/WiBro 셀룰러 네트워크, 무선 서비스 Application

강 충 구 (Chung G. Kang) 종신회원

1987년 6월 : Univ. of California (San Diego), 전자공학과 학사
1989년 6월 : Univ. of California (Irvine), 전자 및 컴퓨터 공학과 석사
1993년 3월 : Univ. of California (Irvine), 전자 및 컴퓨터 공학과 박사
1992년 7월~1993년 6월 : Aerospace Corp. 연구원
1993년 4월~1994년 2월 : Rockwell International 연구원
1994년 3월~2006년 2월 : 고려대학교 전파통신공학과 조교수/부교수/정교수
2006년 3월~현재 : 고려대학교 전기전자전파공학부 정교수
<관심분야> 광대역 무선 전송 기술 및 매체접근제어 프로토콜 설계/구현, 무선 네트워크 모델링 및 성능분석, 4세대 이동통신