ON THE RATE CONSTRAINT OF TRANSMITTING MULTIPLE PRIORITY CLASSES WITH QOS

W. Kumwilaisak†, Q. Zhang‡, W. Zhu‡, C.-C. Jay Kuo†, and Ya-Qin Zhang‡

† Integrated Media Systems Center and Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089-2564
‡ Microsoft Research Asia, 3F 49 Zhichun Road, Haidian, Beijing, 100080, P.R. China
E-mail: kumwilai@sipi.usc.edu

ABSTRACT

The rate constraint of transmitting multiple priority classes over a time-varying service-rate channel is studied in this work. This constraint specifies the maximum data rate that can be transmitted reliably with QoS (quality of service) guarantee. In our framework, the time-varying service channel is modeled by an N-state discrete Markov process, where each Markov state is associated with a channel service rate, and the absolute priority scheduling is used to transport packets of different classes. The transmission rate constraint is derived based on effective bandwidth and capacity. To be more specific, given channel parameters and the maximum buffer size for each priority class, statistical QoS guarantees in terms of packet loss probabilities can be determined and translated to the transmission rate constraint. The derived result is verified by simulation in a time-varying wireless environment.

1. INTRODUCTION

Multiple QoS provisioning is one of the most important issues in the next generation IP-based Internet and wireless network [1]. Different from the best effort IP system, the multiple QoS provisioning system offers differentiated guarantees of loss and delay depending on the requirements of applications. For example, UMTS QoS specifications [1] have four classes of services: conversational, streaming, interactive, and background.

It is difficult to provide absolute QoS guarantees [2] over unreliable and time-varying links of IP-based channels. Under a dynamic environment, absolute QoS provisioning may not be able to maintain the QoS guarantee as committed, which may result in interrupted or even terminated transmission services. Thus, statistical QoS provisioning, where QoS is guaranteed within some probability, provides an attractive alternative in the next generation wired and wireless IP networks. One important problem in statistical QoS provisioning is the choice of the transmission rate constraint, which is the maximum data rate transmitted reliably with the statistical QoS guarantee. It serves as an effective tool to allocate the resource for data transmission with varying link quality. A methodology is presented in this work to compute the transmission rate constraint of multiple priority classes under time-varying service-rate channels.

This work extends results in [5], where a single service class and an infinite buffer size were considered in determining the effective capacity of a wireless channel, to multiple priority classes with a finite buffer size. Generally speaking, we use theory of effective bandwidth [3, 4] as well as effective capacity [5] in deriving the transmission rate constraint. Under the proposed framework, the time-varying service-rate channel is first modeled by an N-state discrete Markov model, where each Markov state is associated with a certain channel service rate, and the N-state Markov model is characterized by its transitional probabilities. Furthermore, the absolute priority scheduling is adopted to transport information packets of different classes. Then, given channel parameters and maximum buffer sizes for multiple priority classes, the statistical QoS guarantee in terms of packet loss probabilities for multiple classes can be determined and translated to the transmission rate constraint. The derived result is verified by simulation in a time-varying wireless environment.

The rest of the paper is organized as follows. Section 2 presents the N-state Markov model to characterize the time-varying service-rate channel and reviews concepts of effective bandwidth and effective capacity. In Section 3, the transmission rate constraint of the multiple priority classes is derived based on theory of effective bandwidth and effective capacity. Simulation results are shown in Section 4. Finally, concluding remarks are given in Section 5.

2. BACKGROUND

2.1. Markov Model for Time-Varying Service-Rate Channel

The time-varying service-rate channel is modeled by the first order N-state Markov model in this work. The channel state at time \( u \) is expressed as \( X_c(u) \), where \( X_c(u) \in \{0, 1, \ldots, N-1\} \), in this model. Each possible channel state \( X_c(u) = i \) represents a channel link condition characterized by channel service rate \( r_i \), which is the available transmission data rate.

For example, in time-varying service rate wireless channels, the channel service rate of Markov state \( i \) (in the unit of bps) can be computed as

\[
r_i = R \log_2 (1 + \gamma_i),
\]

where \( R \) is the available channel capacity and \( \gamma_i \) is the channel link condition. The time-varying service-rate channel is first modeled by an N-state discrete Markov model, where each Markov state is associated with a certain channel service rate, and the N-state Markov model is characterized by its transitional probabilities. Furthermore, the absolute priority scheduling is adopted to transport information packets of different classes. Then, given channel parameters and maximum buffer sizes for multiple priority classes, the statistical QoS guarantee in terms of packet loss probabilities for multiple classes can be determined and translated to the transmission rate constraint. The derived result is verified by simulation in a time-varying wireless environment.

Finally, concluding remarks are given in Section 5.
where $R$ is the bandwidth in Hz and $\gamma_i$ is the SNR value used to measure the wireless channel condition at Markov state $i$. The Markov channel model can be characterized by the transitional probability matrix

$$P_{\text{state}} = \begin{pmatrix}
    p_{11} & \ldots & p_{1N} \\
    \vdots & \ddots & \vdots \\
    p_{N1} & \ldots & p_{NN}
\end{pmatrix}$$

where $p_{ij} = P(X_c(u) = j|X_c(u-1) = i)$ is the transitional probability from state $i$ to state $j$.

### 2.2. Effective Bandwidth and Capacity

First, with $K$ priority classes, it is assumed that service class $i$ has a higher priority than service class $j$, if $i < j$. Service class $i$ provides a statistical QoS guarantee for transmitting data in terms of the packet loss probability. It can be derived from theory of large deviation as [5]

$$P(B_i(t) > B_{\max,i}) \approx \xi_i \cdot e^{-\theta_i_i B_{\max,i}},$$

(2)

where $B_i(t)$ is the buffer occupancy of service class $i$ at time $t$, $B_{\max,i}$ is the maximum buffer size of service class $i$, $\theta_i$ is the QoS exponent corresponding to the guaranteed packet loss probability provided by service class $i$, $\xi_i$ is the probability that the buffer of service class $i$ is not empty, and $\xi_i \cdot e^{-\theta_i_i B_{\max,i}}$ is the guaranteed packet loss probability of service class $i$.

![Diagram](image)

**Fig. 1.** A multiple QoS transmission system over a time-varying service-rate channel.

The accumulated data amount of the source stream generated for transmission under service class $i$ from time 0 to $t$ is a random variable of the form:

$$\alpha_i(t) = \int_0^t X_{c,i}(u)du,$$

where $X_{c,i}(u)$ is the source data rate generated for service class $i$ at time $u$ under state $X_{c,i}(u)$. The data of amount $\alpha_i(t)$ will be stored in the buffer of size $B_{\max,i}$ before transmission over service class $i$. Next, the time-varying service-rate channel effect for substream transmission should be considered. Let the random variable of data to be transmitted over service class $i$ from time 0 to $t$ be of the form

$$S_i(t) = \int_0^t X_{c,i}(u)du,$$

where $X_{c,i}(u)$ is the channel service rate of class $i$ at time $u$ with the random channel state $X_{c,i}(u)$.

The stochastic behavior of the source and channel random processes can be described using the concept of effective bandwidth [3, 4] and effective capacity [5]. Given the QoS exponent $\theta_i$, we have

$$e_{\alpha_i(t)}(\theta_i) = \frac{\Lambda_{\alpha_i(t)}(\theta_i)}{\theta_i},$$

$$\mu_{S_i}(\theta_i) = -\frac{\Lambda_{S_i}(\theta_i)}{\theta_i},$$

where $e_{\alpha_i(t)}(\theta_i)$ and $\mu_{S_i}(\theta_i)$ are the effective bandwidth and the effective capacity, respectively. Given the source characteristic, the effective bandwidth provides a guideline about how much bandwidth a service channel should allocate to a data substream to meet the statistical QoS requirement. In contrast, the effective capacity imposes the constraint on the amount of information substream to be transmitted over the time-varying service rate channel while maintaining the statistical QoS guarantee provided by service class [5] under given channel characteristics.

Furthermore, we let $\Lambda_{\alpha_i(t)}(\theta_i)$ and $\Lambda_{S_i}(\theta_i)$ denote log-moment generating functions defined as [3, 4, 5]

$$\Lambda_{\alpha_i}(\theta_i) = \lim_{t \to \infty} \frac{\log E[e^{\theta_i \alpha_i(t)}]}{t},$$

$$\Lambda_{S_i}(\theta_i) = \lim_{t \to \infty} \frac{\log E[e^{-\theta_i S_i(t)}]}{t}.$$

To derive the transmission rate constraint of transmitting multiple priority classes, the theorem of the effective bandwidth and effective capacity [3] is utilized. It is summarized below.

**Theorem 1** Suppose the time-varying channel service rate is a stationary ergodic process satisfying the Gartner-Ellis condition with time-varying channel service rate $r_{X_{c,i}(u)}$ and $\theta^*_i$ is the QoS exponent provided by transmission service class, which corresponds to the effective capacity

$$\mu_{S_i}(\theta^*_i) = \frac{\Lambda_{S_i}(\theta^*_i)}{\theta^*_i},$$

(3)

then, we have

$$\mu(\theta_r) = \begin{cases}
    \mu_{S_i}(\theta^*_i), \\
    \mu_{S_i}(\theta^*_i) \frac{\theta_r}{\theta^*_i} - e_{S_i}(\theta_r - \theta^*_i), & 0 \leq \theta_r \leq \theta^*_i, \\
    0, & \theta_r > \theta^*_i,
\end{cases}$$

(4)

where $\theta_r > 0$ is the QoS exponent corresponding to the packet loss probability required by transmitting information, $S(t)$ is the random variable of transmitted data over service class from time 0 to time $t$, and $\mu(\theta_r)$ is the effective capacity with QoS exponent $\theta_r$. Note that

$$e_{S_i}(\theta_r - \theta^*_i) = \frac{\Lambda_{S_i}(\theta^*_i - \theta^*_i)}{\theta_r - \theta^*_i},$$

in above can be viewed as the effective bandwidth of $S(t)$ with the QoS exponent $\theta_r - \theta^*_i$. 

1 - 454
3. DERIVATION OF TRANSMISSION RATE CONSTRAINT

The transmission rate constraint of a priority transmission system under a time-varying service rate channel $r_{X_{s}(u)}$ is derived in this section. We start the derivation by first assuming that there are only 2 priority classes with QoS exponents $\theta_{1}$ and $\theta_{2}$ corresponding to their guaranteed packet loss probabilities. The generated data stream for transmitting over the first and second service classes from time 0 to $t$ are $\alpha_{1}(t)$ and $\alpha_{2}(t)$ and stored in different buffers of sizes $B_{\text{max}1}$ and $B_{\text{max}2}$, respectively. The statistical QoS guarantee of each service class is provided in form of the packet loss probability as shown in Eq. (2), which is computed based on the QoS exponent and the buffer size. With the absolute priority scheduling, the second service class has a lower priority than the first service class, and will be served only after all data in the buffer of the first service class is served.

For the first stream in the high priority buffer, it is easy to see that the rate constraint of substream 1 with QoS exponent $\theta_{1}$ transmitted over the service class 1 with QoS exponent $\theta_{1}$ and buffer size $B_{\text{max}1}$ can be shown as

$$\min(\mu_{1}(\theta_{1}) \cdot t, \alpha_{1}(t)) < r_{\text{channel}} \cdot t,$$

or

$$\min(\mu_{S_{1}(t)}(\theta_{1}, t, \alpha_{1}(t)) < r_{\text{channel}} \cdot t, 0 \leq \theta_{r,1} \leq \theta_{l,1}.$$

$$e_{S_{1}(t)}(\theta_{r,1} - \theta_{l,1}) \cdot t, \alpha_{1}(t)) < r_{\text{channel}} \cdot t, \quad \theta_{r,1} > \theta_{l,1},$$

where $\mu_{1}(\theta_{1})$ is the transmission rate constraint of substream 1 and $r_{\text{channel}}$ is the expected channel service rate. $S_{1}(t)$ is the random variable of information that can be transmitted over service class 1 under the time-varying service rate $r_{X_{s}(u)} = r_{X_{s}(u)}$ from time 0 to $t$.

For the low priority substream, the existence of substream 1 affects the transmission rate constraint of substream 2 due to the absolute priority scheduling algorithm. The derivation of the transmission rate constraint of substream 2 can be simply viewed as trying to transmit substream 2 alone with time-varying channel service rate

$$r_{X_{s,2}(u)} = r_{X_{s}(u)} - \min(\mu_{1}(\theta_{1}), \frac{\alpha_{1}(t)}{t}),$$

where $r_{X_{s,2}(u)}$ is the time-varying channel service rate, which is seen by substream 2 with the existence of substream 1. Suppose that substream 2 has its own QoS exponent requirement equaling $\theta_{r,2}$. Hence, from Theorem 1, the transmission rate constraint of substream 2 can be computed based on $r_{X_{s,2}(u)}$ as

$$\mu_{2}(\theta_{r,2}) = \begin{cases} \mu_{S_{2}(t)}(\theta_{r,2}), & 0 \leq \theta_{r,2} \leq \theta_{l,2} \\ \mu_{S_{2}(t)}(\theta_{r,2}) + e_{S_{2}(t)}(\theta_{r,2} - \theta_{l,2}) & \theta_{r,2} > \theta_{l,2} \end{cases}$$

(8)

where $e_{S_{2}(t)}(\theta_{r,2} - \theta_{l,2})$ is the effective bandwidth of $S_{2}(t)$ with QoS exponent provisioning $\theta_{r,2} - \theta_{l,2}$. $S_{2}(t)$ is the random variable of information that can be transmitted over service class 2 under the time-varying service rate $r_{X_{s,2}(u)}$ from time 0 to $t$.

Together with Eq. (5), the transmission rate constraint on both substreams 1 and 2 can be expressed as

$$\min(\mu_{1}(\theta_{r,1}) \cdot t, \alpha_{1}(t)) < r_{\text{channel}} \cdot t,$$

and

$$\min(\mu_{1}(\theta_{r,1}) \cdot t, \alpha_{1}(t)) + \min(\mu_{2}(\theta_{r,2}) \cdot t, \alpha_{2}(t)) < r_{\text{channel}} \cdot t.$$

Eqs. (9) and (10) show that the transmission rates of substreams 1 and 2 are limited by $\mu_{1}(\theta_{r,1})$ and $\mu_{2}(\theta_{r,2})$, respectively. Moreover, the summation of the constraint on the rate of both substreams 1 and 2 should not exceed the expected channel service rate $r_{\text{channel}}$. Therefore, when the substream demands to send more data than the transmission rate constraint, in which the service class can allow with the statistical QoS guarantee, the rate shaper algorithm has to be applied to shape the information rate to meet with the transmission rate constraints.

The procedure for deriving of the transmission rate constraint for 2 data substreams can be easily extended to $K$ substreams via

$$\sum_{i=1}^{K} \min(\mu_{i}(\theta_{r,i}) \cdot t, \alpha_{i}(t)) < r_{\text{channel}} \cdot t, \quad k = 1, 2, \ldots, K,$$

(11)

where $\mu_{i}(\theta_{r,i})$ is the transmission rate constraint of substream $i$ with QoS exponent $\theta_{r,i}$ computed by assuming that the channel service rate seen by substream $i$ can be written as

$$r_{X_{s,i}(u)} = r_{X_{s}(u)} - \sum_{j=1}^{i-1} \min(\mu_{j}(\theta_{r,j}), \frac{\alpha_{j}(t)}{t})$$

(12)

where $\alpha_{i}(t)$ is the data generation by the source of class $i$.

4. EXPERIMENTAL RESULTS

We conducted experiments to study the derived transmission rate constraint of multiple priority classes in this section. In particular, we consider a time-varying service-rate channel modeled by the 4-state Markov process representing a wireless link with the SNR value and the normalized Doppler frequency equal to 16 dB and $10^{-2}$, respectively [6].

If the time-varying service rate channel $X_{s}(u)$ can be well modeled by the Markov process with $N$ states, the closed form of effective capacity and effective bandwidth for transmission data rate $S(t)$ could be written as [3, 4]

$$\mu_{S(t)}(\theta_{l}) = \frac{\ln(\Omega(e^{-\theta_{l}A}P_{\text{state}}))}{-\theta_{l}},$$

and

$$e_{S(t)}(\theta_{l}) = \frac{\ln(\Omega(e^{\theta_{l}A}P_{\text{state}}))}{\theta_{l}},$$

where $\mu_{S(t)}(\theta_{l})$ and $e_{S(t)}(\theta_{l})$ are the effective capacity and the effective bandwidth of $S(t)$, respectively, $P_{\text{state}}$ is the transitional probability matrix of the discrete Markov model, and $\Omega(U)$ is the spectral radius of matrix $U$. 

1 - 455
Let $\vec{r} = \{ r_1, \ldots, r_N \}$ be the vector of service rates at various discrete Markov states, and $\Lambda = \text{diag}(\vec{r})$ be the diagonal matrix with its diagonal components corresponding to those in $\vec{r}$. Furthermore, the QoS exponent of the guaranteed packet loss probability can be expressed as

$$\theta_i = -\frac{\ln(P(B > B_{\text{max}}))}{B_{\text{max}}}$$

where $\xi$ is the probability that the buffer is empty.

First, consider 2 service classes with absolute priority scheduling for packet transmission. The first class has a higher priority than the second class. The packet size for transmission is set to be 200 bytes. The expected service rate of the wireless channel is 380 kbps. As seen in Fig. 2, the transmission rate constraint of the first service class (i.e., the high priority class) computed from the closed form of effective bandwidth and effective capacity and those obtained from the simulation are close over a wide range of the packet loss probability guarantee. The lower the packet loss probability requirement, the less reliable the transmitted data rate. Simulation results given in Fig.2 also demonstrate the buffer size effect on the transmission rate constraint. The larger the buffer size, the more data rate we can transmit under the same packet loss probability requirement.

![Fig. 2. The transmission rate constraint of a wireless channel of service class 1 computed from discrete Markov channel model corresponding to the normalized Doppler frequency = $10^{-2}$ and average power = 16 dB. The buffer sizes are set to 250 and 500 packets.](image)

Based on the transmission rate constraint derived in Section 3, Fig. 3 shows the transmission rate constraint of service class 2 (i.e., the low priority class) over a wide range of guaranteed packet loss probability requirement. As seen from simulation results, the transmission rate constraint of service class 2 with a buffer size equal 250 packets is dependent on how much service class 1 occupies the wireless link. The transmission rate constraint of service class 2 with a lower QoS requirement of service class 1 (guaranteed packet loss probability = $10^{-2}$ with transmission rate constraint = 109 kbps) can provide a higher transmission rate than that with a higher QoS requirement of service class 1 (guaranteed packet loss probability = $10^{-4}$ with transmission rate constraint = 54.5 kbps).

![Fig. 3. The transmission rate constraint of a wireless channel of service class 2 with a buffer size equal to 250 packets based on absolute priority scheduling when the packet loss rate requirement of class 1 is equal to $10^{-2}$ and $10^{-4}$.](image)

5. CONCLUSION

The transmission rate constraint of multiple priority classes under the time-varying service rate channel was studied in this paper. Theory of effective bandwidth and effective capacity was used to derive the transmission rate constraint. It was demonstrated by theoretical derivation and simulation results that the reliable transmitting data rates over multiple priority classes depends on the statistical QoS provisions. Given the same statistical QoS provision and buffer size, the high priority class can transmit more data rate than the lower priority class. The transmission rate can be increased by enlarging the buffer size.

6. REFERENCES


