Analyzing Scheduling Performance for Real-time Traffic in Wireless Networks

Emily Lee and David Taubman
The University of New South Wales, Sydney, Australia

Abstract— **This paper is a sequel to an earlier study on scheduling real-time traffic in wireless TDMA channels. In particular, we develop mathematical analysis to model the system behavior of the equal-delay scheme and its extension of the multi-class scheme, both of which were proposed previously. The main usefulness of the proposed scheduling mechanisms is the improvement on the worst-case delay, and consequently the reduction in probability of delay violation or loss. We achieve the analytical solution by evaluating the corresponding cumulative distribution of delay. The derivation of the distribution function is based on a well-known method of solving a set of first-order differential equations and eliminating all unstable modes. However, the existence of channel variations generate unique difficulties which renders the original method unsuitable. To solve this problem, we devise novel transformation techniques and an algorithm to recursively estimate the state probabilities in steady-state, so the original method can be applied in an alternative manner. Furthermore, we develop a reduced state model for homogeneous systems to alleviate the requirement on computational complexity of the full model. On the other hand, an additional challenge to extend the mathematical model to support the multi-class system is to overcome the varying progression rates of individual transmissions. To tackle this problem, we model the system aggregate delay value and re-calculate the scaled statistics for each class of flows according to their pre-assigned weights. Finally, numerical results and computational complexity analysis are presented.**

I. INTRODUCTION

This paper is a continual study on the subject of scheduling real-time traffic in wireless networks from its preceding investigation [1]. There have been intensive research focus on developing effective scheduling schemes for wireless networks in recent years [6][7][8], satisfying a wide range of goals, such as minimizing delay, maximizing channel utilization, and maintaining fairness. On the other hand, exchanging multimedia contents across the wireless medium is an area with growing research interest. However, the unique service requirements imposed by real-time traffic are largely overlooked by most of the existing wireless schedulers, resulting in a big lag in achievements behind industrial expectations. The Equal-Delay scheme and its extension to support multiple traffic classes proposed in [1] aim to narrow this gap by paying specific attention to this important type of network traffic. In this paper, we propose mathematical models for evaluating the scheduling performance of the schemes, and demonstrate the analytical solutions and their computational complexities.

The main advantages of achieving analytical solution over simulation analysis are its inherent proven integrity and reliability of the obtained results, and the lesser requirement on computational power, allowing the tasks of admission control and network planning to be executed more efficiently and accurately. In the powerful paper published in 1982 [2], Anick et al. proposed a mathematical model to represent a data-handling system with multiple sources, and devised an elegant analytical solution for it in closed form. Although the system considered in [2] is primitive, the work stimulates the development of extended mathematical analysis for systems with other characteristics [3][4][5]. These include multiplexing more than one type of sources onto one transmission channel, and multiplexing one type of sources onto a network of channels.

Like the above antecedent references, the development of our analytical model is based on the framework established in [2]. However, the varying progression rates of the two existing time scales, namely, the source and channel times, create unique difficulty not presented or studied in any of the these works before. We develop two mechanisms in order to convert our system model to a form that can utilize the original approach. The first procedure is a transformation technique which combines the two times into a single one by formulating the relationship between them and expressing one in terms of the other, whereas the second approach is to obtain the steady-state distribution of the combined system, which is done via recursive estimation.

Although a closed-from solution is not achieved in our paper, the cumulative distribution of delay can be evaluated by numerical procedures. For systems containing only small numbers of individual source and channel states, the computational complexity is relatively low. However, the processing and memory requirements go up dramatically with increasing number of states. To lessen the complexity involved, we propose an improved model for homogeneous sources so that the number of resulting states, and hence the system size, are vastly reduced, which in turn enables a more efficient computation.

The equal-delay scheme has been extended to support flows with diverse delay requirements in [1], and the resulting multi-class scheme imposes different delay degradations to flows belonging to different classes. This unbalanced delay allocation poses an additional problem to the extension of the mathematical model from the equal-delay framework, since it introduces yet another irregularly-varying time scale to the multi-class system. Inspired by the original idea of transformation proposed for the equal-delay system, we merge the three different time scales into one that can be used for analytical derivation. Instead of evaluating individual delay statistics, the aggregate delay statistics are evaluated. Each individual class result can then be obtained by scaling the aggregate results with the associated class weight.

Brief introduction of the equal-delay and the multi-class scheme are given in Section II and IV, whereas the mathematical analysis of the two schemes are presented in Section III and V. Finally, numerical results are provided in Section VI.

II. The Equal-Delay Scheme

We consider a TDMA transmission system containing F mobile stations, labeled $f = 1, 2, \ldots, F$. Each station is equipped with its own sending buffer of infinite size, with a common base station acting as a central access controller through polling. In our scheme, the base station schedules individual traffic flows in cycles, and the objective of the scheduling scheme is to maintain the same cycle delay across all flows in the system. We achieve this by allowing every flow to transmit one cell of traffic, corresponding to the amount of source generation within the same time interval as any other flows.

As the quality of a channel degrades, various channel adaption procedures including the FEC can be carried out to maintain roughly the same perceived error probability. The resulting transmission speed of the link thus appears to vary over time. We assume that the time required to transmit a certain amount of source traffic in the j^{th} degraded channel is C_j times longer than in the highest quality channel, and consequently define C_j as the channel factor for state j. As a result of the variations in source activity level and channel quality, the n^{th} cycle duration, denoted by λ_n, also varies over time. Let us represent the average source generation rate for flow f from time $(k-1)\Lambda$ to $k\Lambda$ by $S_{f,k}$; the average channel factor for flow f during its transmission interval within the k^{th} cycle by $C_{f,k}$; and the nominal cycle duration by Λ. The relationship between the n^{th} cycle delay, $D_{n\Lambda}$, and its cycle duration may then be conveniently captured by

$$0 \leq D_{n\Lambda} = \sum_{k=0}^{n-1}(\lambda_k - \Lambda) = \sum_{k=0}^{n-1}\left(\sum_{f=1}^{F}\frac{S_{f,k}C_{f,k}}{L}\Lambda\right) + G_k - \Lambda \quad (1)$$

where L represents the highest link rate achievable by any channel, and G_k represents any portion of the k^{th} cycle which is not used for real-time traffic.

We consider the limiting behavior of equation (1) in order to derive analytical expressions for the cumulative delay probability distribution. As $\Lambda \to 0$ and $n \to \infty$, such that $n\Lambda \to \tau$, the rate of increase in D_τ, for positive D_τ, becomes

$$\frac{\partial D_\tau}{\partial \tau} = \sum_{f=1}^{F}\frac{S_f(\tau)C_f(t_\tau)}{L} - 1$$

III. Mathematical Analysis for the Equal-Delay Scheme

A. The Model

For a homogeneous system with F flows, let there be N_S source states and N_C channel states for each flow. We consider a composite process in which each of its states is composed of the source and channel states of all flows. The i^{th} composite state, m_i, is thus defined as

$$m_i = [(s_{1,i}, c_{1,i}), (s_{2,i}, c_{2,i}), \ldots, (s_{F,i}, c_{F,i})]$$

where s_f and c_f are the source and channel states for flow f, respectively.

In this paper we model the variations in both source generation and channel condition by Markov process. For Markovian individual source and channel processes, the composite process is a multi-dimensional Markov process. The idea is to express the system behavior within a very short time interval $[\tau, \tau + \Delta\tau]$. By passing $\Delta\tau \to 0$, all compound transitions are eliminated because of its Markovian nature. This simplification allows the steady-state distribution of the delay to be found by solving a simple first-order differential equation and by eliminating all unstable modes.

Let $P_{m_\nu}(\tau + \Delta\tau, D)$ be the joint probability that at source time $(\tau + \Delta\tau)$, the composite state is m_ν, and the cumulative delay is less than or equal to D. Logically both τ and D have to be non-negative. Furthermore, let us use $M(\tau)$ to represent the composite state at source time τ. $P_{m_\nu}(\tau + \Delta\tau, D)$ is equal to the sum of the product of all the possible composite states at τ and their corresponding conditional probabilities. That is,

$$\begin{aligned} &P_{m_\nu}(\tau + \Delta\tau, D) \\ &= \sum_{i=1}^{N_M}[P(M(\tau + \Delta\tau) = m_\nu | M(\tau) = m_i, D - D_i) \\ &P_{m_i}(\tau, D - D_i)] \end{aligned} \quad (2)$$

where D_i is the amount of delay incurred during the interval $[\tau, \tau + \Delta\tau]$. There are N_M composite states, where $N_M = (N_S N_C)^F$.

Now, since the composite process is a continuous-time Markov Chain, the conditional probabilities are all independent of the delay, and are equivalent to the corresponding Markov transition probabilities. As $\Delta\tau \to 0$, only single source / channel transitions originated from one flow remain significant, the remaining compound events all have zero probability of occurring. Composite states m_i and m_j are called neighboring states in the multi-dimensional Markov Chain if they are differed by exactly one individual source or channel state. Let us denote the transition rate matrix of this composite Markov process by $\mathbf{Q}$, and the transition rate from m_i to neighboring state m_j by $Q_{m_i \to m_j}$, equation (2) then becomes

$$\begin{aligned} P_{m_\nu}(\tau + \Delta\tau, D) = &\sum_{i=1, i\neq\nu}^{N_M}[Q_{m_i \to m_\nu}\Delta\tau P_{m_i}(\tau, D - D_i)] \\ &+ (1 - \sum_{i=1, i\neq\nu}^{N_M} Q_{m_i \to m_\nu}\Delta\tau)P_{m_\nu}(\tau, D - D_\nu) \\ &+ O(\Delta\tau^2) \end{aligned} \quad (3)$$

where $lim_{\Delta\tau \to 0}\frac{O(\Delta\tau^2)}{\Delta\tau} = 0$.

Since the system is described in source-time progression, but the underlying channel processes change states according to the real-time axis, $Q_{m_i \to m_j}$ must include a stretching / shrinking factor of r_{m_i} in such cases. Its value equals the total flow rate of the composite state m_i. That is, with $S_{s_{f,i}}$ and $C_{c_{f,i}}$ representing the source rate and the channel factor of source state $s_{f,i}$ and channel state $c_{f,i}$, respectively,

$$r_{m_i} = \sum_{f=1}^{F} \frac{S_{s_{f,i}} C_{c_{f,i}}}{L}$$

and L is defined as the constant link rate, as before. This represents the situations where the probability of a channel state change is higher within a source time interval of $\Delta\tau$ when the system is running slow (stretching) than when the system is running at nominal speed, or when it is catching up (shrinking). Similarly, the probability of a channel state change is lower when the system is catching up than when it is running at nominal speed, or slowing down. Let the transition rate of the underlying source state change be $\alpha_{m_i \to m_j}$ and channel state change be $\beta_{m_i \to m_j}$. $Q_{m_i \to m_j}$ can be formally expressed as

$$Q_{m_i \to m_j} = \begin{cases} \alpha_{m_i \to m_j} & \text{for source state change} \\ r_{m_i}\beta_{m_i \to m_j} & \text{for channel state change} \\ 0 & \text{otherwise} \end{cases}$$

Now passing (3) to the limit $\Delta\tau \to 0$, we get

$$\begin{aligned} &\frac{\partial P_{m_\nu}(\tau, D)}{\partial \tau} + (r_{m_\nu} - 1)\frac{\partial P_{m_\nu}(\tau, D)}{\partial D} \\ &= \sum_{i=1, i\neq\nu}^{N_M} [Q_{m_i \to m_\nu} P_{m_i}(\tau, D)] \\ &- (\sum_{i=1, i\neq\nu}^{N_M} Q_{m_i \to m_\nu}) P_{m_\nu}(\tau, D) \end{aligned} \quad (4)$$

If existing, the steady-state probability of $P_{m_\nu}(\tau, D)$, denoted by $\widetilde{P}_{m_\nu}(D)$, can be found by setting the rate of change of $P_{m_\nu}(\tau, D)$, i.e. $\frac{\delta P_{m_\nu}(\tau,D)}{\delta\tau}$, to 0. Therefore in steady-state, equation (4) becomes

$$\begin{aligned} (r_{m_\nu} - 1)\frac{d\widetilde{P}_{m_\nu}(D)}{dD} = &\sum_{i=1, i\neq\nu}^{N_M} [Q_{m_i \to m_\nu} \widetilde{P}_{m_i}(D)] \\ &- (\sum_{i=1, i\neq\nu}^{N_M} Q_{m_i \to m_\nu}) \widetilde{P}_{m_\nu}(D) \end{aligned} \quad (5)$$

Defining $\mathbf{D}$ as a $N_M \times N_M$ diagonal matrix with non-zero entries $\{r_{m_1} - 1, r_{m_2} - 1, \ldots, r_{m_{N_M}} - 1\}$, and $\mathbf{M}$ also a $N_M \times N_M$ matrix, whose entries are

$$\mathbf{M}_{j,i} = \begin{cases} Q_{m_i \to m_j} & \text{if} i \neq j, \\ -\sum_{k=1, k\neq i}^{N_M} Q_{m_k \to m_j} & \text{otherwise} \end{cases}$$

we can write down an equivalent matrix expression for equation (5) as

$$\mathbf{D}\frac{d\widetilde{\mathbf{P}}(D)}{dD} = \mathbf{M}\widetilde{\mathbf{P}}(D) \quad (6)$$

where $\widetilde{\mathbf{P}}(D)$ is a $N_M \times 1$ column vector whose i^{th} entry is $\widetilde{P}_{m_i}(D)$. The general solution to the differential equation (6) which gives rise to a stable system is thus

$$\widetilde{\mathbf{P}}(D) = \sum_{Re_{\lambda_i} \leq 0} u_i \mathbf{V}_i e^{\lambda_i D} \quad (7)$$

where λ_i and $\mathbf{V}_i$ are the i^{th} eigenvalue and eigenvector of $\mathbf{D}^{-1}\mathbf{M}$, and u_i's are the set of coefficients to be determined by the boundary conditions. For $\mathbf{D}$ to be invertible, the diagonal entries must be non-zero, which implies $r_{m_i} \neq 1, \forall i$. On the other hand, by restricting to only non-positive eigenvalues (i.e. $Re_{\lambda_i} \leq 0$), the unstable modes of the system are eliminated. Otherwise $|\widetilde{\mathbf{P}}(D)|$ will grow without bound as $D \to \infty$.

With the inclusion of the ∞ boundary condition, equation (7) then becomes

$$\widetilde{\mathbf{P}}(D) = \widetilde{\mathbf{P}}(\infty) + \sum_{Re_{\lambda_i} < 0} u_i \mathbf{V}_i e^{\lambda_i D} \quad (8)$$

Further, the zero boundary condition gives the following set of linear equations

$$\widetilde{P}_{m_i}(0) = 0, \text{ for } i = 1, 2, \ldots, N_M \text{ such that } r_{m_i} > 1 \quad (9)$$

which reads that in steady-state, the probability of no delay is zero when the aggregate rate exceeds the nominal link rate. Therefore the remaining N_- constant coefficients of u_i can be solved, since the number of composite states with effective rate $r_{m_i} > L$ equals N_-.

The characteristics of the $2F$-dimensional Markov process, and hence the procedure for obtaining the steady-state distribution, are noticeably different from our antecedent references using method proposed in [2]. Firstly, the inclusion of the stretching / shrinking factor in each transition rate makes the property of it being only dependent on the individual state change, invalid. As a result, the steady-state distribution of the composite Markov process cannot be expressed in a simple product form, that is,

$$\widetilde{P}_{m_i}(\infty) \neq \prod_{f=1}^{F} \widetilde{P}_{s_{f,i}}(\infty)\widetilde{P}_{c_{f,i}}(\infty), \text{ for } i = 1, 2, \ldots, N_M$$

where $\widetilde{P}_{s_{f,i}}(\infty)$ and $\widetilde{P}_{c_{f,i}}(\infty)$ are the steady-state probabilities of flow f's individual source and channel being in state $s_{f,i}$ and $c_{f,i}$, respectively. Furthermore, the inclusion of the stretching / shrinking factor in the composite Markov transition rate matrix is valid only when the delay is strictly positive. On the other hand, when the delay equals zero, they should not be included at all. This is because when the system is not running behind, the real time would not run slower than the source time, even when there is less to transmit than what the link can handle. In other words, the scheduling system is governed by two separate composite Markov processes. When the delay is positive, it follows the process which includes both stretching and shrinking factors, which is the same as the one that governs the dynamics of the system in equation (3). When the delay reaches zero, on the other hand, it is governed by the Markov process that excludes the stretching / shrinking factors. It is thus apparent that obtaining the real steady-state distribution under such a complicated system dynamics is not an easy task. In the following section we describe a method to recursively estimate this distribution.

B. Estimation of the Steady-State Probabilities

We start with an initial estimate of $\widetilde{\mathbf{P}}(\infty)$, denoted by $\widetilde{\mathbf{P}}(\infty)^0$, and the value of which is obtained by solving a set of N_M linear equations from the transition rate matrix $\mathbf{Q_{delay}} = \mathbf{Q}$. The first estimate of $\widetilde{\mathbf{P}}(D)$, denoted by $\widetilde{\mathbf{P}}(D)^1$, can be obtained by evaluating (8) and substitute (9) in the solution. In the following updating procedures, the knowledge of $\widetilde{\mathbf{P}}(0)^1$ is required. Now the first estimate of the portion of time spent in any particular composite state, m_i, in steady-state with zero delay, let us represent it by $p^1_{m_i}$, can be obtained by evaluating

$$p^1_{m_i} = \frac{\widetilde{P}_{m_i}(0)^1}{\widetilde{P}_{m_i}(\infty)^0}$$

and the portion of time with positive delay, denoted by $q^1_{m_i}$, is thus

$$q^1_{m_i} = 1 - p^1_{m_i}$$

Essentially the updating process of $\widetilde{\mathbf{P}}(\infty)$, hence $\widetilde{\mathbf{P}}(D)$, is performed recursively by re-calculating the transition rate matrix of the combined Markov process, denoted by $\mathbf{A}^k$, where k represents the recursion step, and re-evaluating (8) and (9). This abstract random process governs the behavior of the system as a whole, under both circumstances of with and without delay. The k^{th} step update of each transition rate in $\mathbf{A}^k$ is given by

$$A^k_{m_i \to m_j} = p^k_{m_i} Q_{delay,m_i \to m_j} + q^k_{m_i} Q_{nodelay,m_i \to m_j} \quad (10)$$

where $Q_{delay,m_i \to m_j}$ is the transition rate from m_i to m_j according to transition rate matrix $\mathbf{Q_{delay}}$. Similarly $Q_{nodelay,m_i \to m_j}$ is the transition rate from m_i to m_j according to $\mathbf{Q_{nodelay}}$, which corresponds to the composite Markov process without including the stretching / shrinking factors. Given $A^k_{m_i \to m_j}$, the next estimate of $\widetilde{\mathbf{P}}(\infty)^k$, that is, $\widetilde{\mathbf{P}}(\infty)^{k+1}$, can be evaluated using the above mentioned method. Convergence usually occurs within just 5 iterations, and in each iteration we only need to solve a set of N_M linear equations, therefore the computational requirement for this updating process is modest.

C. A Reduced Model for Homogeneous Systems

A major problem associated with the mathematical model is its high computational complexity. Even for a modest case of 4 flows with 2 source and channel states each, a total number of 256 composite states would result. This means that in order to solve equation (7) numerically, the eigenvalues / eigenvectors of a 256×256 matrix must be found, which is a non-trivial task for modern computers equipped with consumer-grade power.

We investigate a reduced model for the case of homogeneous flows, in which the number of composite states can be vastly reduced. This improvement is achieved based on an important observation that it is unnecessary to maintain the identities of flows being in certain combinational source and channel states. For instance, the transition rate from the composite state $[(a_1, b_1), (a_2, b_2)]$ to $[(\mu, b_1), (a_2, b_2)]$ is identical to the transition rate from $[(a_2, b_2), (a_1, b_1)]$ to $[(a_2, b_2), (\mu, b_1)]$, which is equal to the single Markov transition rate from source state a_1 to μ. The same also applies to single channel transitions.

As a result, all states that maintain identities to individual flows can be combined to ones that only remember the aggregate number of flows being in certain combinations of source and channel states. We call such a state a reduced composite state. Furthermore, the reduced system also satisfies the Markov properties. Let us label the combinational states (s^1, c^1) the first, (s^1, c^2) the second, $\ldots$, (s^1, c^{N_C}) the N_C^{th}, (s^2, c^1) the $(N_C + 1)^{\text{th}}$, $\ldots$, and finally (s^{N_S}, c^{N_C}) the $(N_S N_C)^{\text{th}}$, and represent the number of flows being in combinational state k by η_k. The i^{th} reduced composite state, denoted by m'_i, can be expressed as

$$m'_i = [\eta_{1,i}, \eta_{2,i}, \ldots, \eta_{N_T,i}]$$

where $N_T = N_S N_C$. Since only the non-zero values of $\eta_{k,i}$ are significant, removing all zero components in the definition would result in exactly the same set of states.

The procedures for finding the cumulative delay probabilities are the same as in the previous case. On the other hand, the elements of the composite transition rate matrix $\mathbf{Q}$ and the value of r_{m_i} for each state are different. The aggregate flow rate for m'_i is modified to

$$r_{m'_i} = \frac{\eta_{1,i} S_{s^1} C_{c^1} + \eta_{2,i} S_{s^1} C_{c^2} + \cdots + \eta_{N_T,i} S_{s^{N_S}} C_{c^{N_C}}}{L}$$

Additionally, assuming that the only difference in the number of flows belonging to neighboring reduced composite states i and j is the k^{th} combinational state, that is,

$$|\eta_{k,i} - \eta_{k,j}| = 1$$

and use $\eta_{m'_i \to m'_j}$ to represent this value of $\eta_{k,i}$. The new transition rates can then be modified to

$$Q_{m'_i \to m'_j} = \begin{cases} \eta_{m'_i \to m'_j} \alpha_{m'_i \to m'_j} & \text{for source state change} \\ r_{m'_i} \eta_{m'_i \to m'_j} \beta_{m'_i \to m'_j} & \text{for channel state change} \\ 0 & \text{otherwise} \end{cases}$$

and the inclusion of the stretching / shrinking factor remains unchanged.

IV. Multiple-Class Scheme

In order to accommodate multiple delay requirements, different numbers of cells can be transmitted by flows belonging to different classes in the multi-class scheme. Our multiple-class scheduling objective is to maintain the relative flow delays in proportion to a pre-defined set of weights, ρ_f. That is, we would like to achieve

$$\mathcal{D}_{f,n\Lambda} = \rho_f \mathcal{D}_{n\Lambda} \quad (11)$$

where $\mathcal{D}_{f,n\Lambda}$ is the delay for flow f at the beginning of the n^{th} cycle, $\mathcal{D}_{n\Lambda}$ is the aggregate delay of all $\mathcal{D}_{f,n\Lambda}$'s, with $\mathcal{D}_{n\Lambda} = \sum_{f=1}^{F} \mathcal{D}_{f,n\Lambda}$, and $\sum_{f=1}^{F} \rho_f = 1$.

To simplify matters, we insist that the total number of cells transmitted during any cycle is exactly F. That is,

$$\sum_{f=1}^{F} Z_{f,k} = F, \forall k \quad (12)$$

Let us denote the number of cells flow f is allowed to transmit during cycle n by $Z_{f,n}$. The total number of cells transmitted by flow f prior to cycle n is thus $\sum_{k=0}^{n-1} Z_{f,k}$, and its delay is

$$0 \leq \mathcal{D}_{f,n\Lambda} = F \left[\sum_{k=0}^{n-1} \left(\sum_{f=1}^{F} \frac{Z_{f,k} S_{f,k} C_{f,k}}{L} \Lambda \right) + G_k - \Lambda \right] \tag{13}$$

To distribute delay amongst the various flows, the scheduler first calculates the total number of outstanding un-transmitted cells amongst all flows, at the beginning of cycle n. We refer to this as the scheduler's *backlog*, B_n, where

$$B_n = \frac{\sum_{f=1}^{F} \mathcal{D}_{f,n\Lambda}}{\Lambda} = \frac{F}{\Lambda}(t_n - n\Lambda)$$

The scheduler selects the $Z_{f,n}$ values using a control loop which aims to distribute the backlog amongst the various flows according to equation (11). Assuming that the total backlog to neither grow nor shrink during the next cycle time, the cell allocation for each flow f can be calculated as

$$Z_{f,n} = 1 - \left(\rho_f B_n - \frac{\mathcal{D}_{f,n\Lambda}}{\Lambda} \right)$$

making appropriate adjustments as required to ensure that $Z_{f,n} \geq 0$ and $\sum_{f=1}^{F} Z_{f,n} = F$.

V. Mathematical Analysis for the Multi-Class Scheme

To derive the analytical expressions for the cumulative delay probability distribution, we again apply the liquid flow analysis. To find the statistics for each individual flow delay $\mathcal{D}_f$, the idea is to model the aggregate delay value $\mathcal{D}$. Then, assuming an ideal control loop, the delays of the various flows can be found according to equation (11).

Since flows are allowed to transmit different unit of cells in each cycle, as opposed to one cell per cycle in the equal-delay case, the individual source time τ_f of each flow no longer advances at the same pace as the system cycle time, $\mathcal{T}$. This condition requires the transition rate between neighboring composite states whose underlying state change is a source change to include a stretching / shrinking factor as well. The values of these factors, on the other hand, are different to those for channel changes. This is because the three time scales, namely, the source time, cycle time, and real time, all accelerate / de-accelerate differently from one another in the multiple-class system. Let the transition rates for the multiple-class system be represented by $\Psi_{m_i \to m_j}, \forall i, j \in [1, N_M], i \neq j$, the values can be expressed as

$$\Psi_{m_i \to m_j} = \begin{cases} Z_f(m_i)\alpha_{m_i \to m_j} & \text{for source state change} \\ x_{m_i}\beta_{m_i \to m_j} & \text{for channel state change} \\ 0 & \text{otherwise} \end{cases}$$

where

$$x_{m_i} = \sum_{f=1}^{F} \frac{Z_f(m_i) S_{s_{f,i}} C_{c_{f,i}}}{L}$$

and $Z_f(m_i)$ denotes the unit of cells allocated to flow f when the composite state is m_i. Since the flow delays maintain the weight proportion at all time, we can obtain the condition

$$(1 - \rho_{\mathsf{f}})\Delta\mathcal{D}_{\mathsf{f}} = \rho_{\mathsf{f}} \sum_{f=1, f \neq \mathsf{f}}^{F} \Delta\mathcal{D}_f, \forall \mathsf{f} \in [1, F] \tag{14}$$

where $\Delta\mathcal{D}_f$ denotes the change in flow f's delay within a short interval of cycle time, $\Delta\mathcal{T}$. Furthermore, the value of each Z_f is independent to the current individual and aggregate delay and thus depends entirely on the composite state the system is currently in. We can construct a set of F linear equations to solve the F unknowns of $Z_1, Z_2, \ldots, Z_F$ by expressing $\Delta\mathcal{D}_f$ in terms of the change in flow f's source time, $\Delta\tau_{f,\mathcal{T}}$, and the change in real time, $\Delta t_{\mathcal{T}}$, within the same interval of $\Delta\mathcal{T}$, that is,

$$\Delta\mathcal{D}_f = \Delta t_{\mathcal{T}} - \Delta\tau_{f,\mathcal{T}} = \Delta t_{\mathcal{T}} - Z_f(m_i)\Delta\mathcal{T}$$

and including conditions (12) and (14).

Whenever $\mathcal{D} > 0$, as $\Lambda \to 0$ and $n \to \infty$, so that $n\Lambda \to \mathcal{T}$, we get, analogous to equation (4),

$$\frac{\partial P_{m_\nu}(\mathcal{T}, \mathcal{D})}{\partial \mathcal{T}} + F \times (x_{m_i} - 1) \frac{\partial P_{m_\nu}(\mathcal{T}, \mathcal{D})}{\partial \mathcal{D}} = \sum_{i=1, i \neq \nu}^{N_M} [\Psi_{m_i \to m_\nu} P_{m_i}(\mathcal{T}, \mathcal{D})] - (\sum_{i=1, i \neq \nu}^{N_M} \Psi_{m_i \to m_\nu}) P_{m_\nu}(\mathcal{T}, \mathcal{D})$$

This can be solved by following the same procedures as provided in the equal-delay section.

VI. Numerical Results

In this section, the numerical results are presented. In particular, the multiplexing power and the implementation issues of the equal-delay scheme are investigated. We omit the multi-class results since they extend naturally from the equal-delay case. We consider a case where the source and channel characteristics of each flow are captured by two separate Markov processes, each of which contains two states. The generation rate of the lower source state is 200 Kbps, with a mean sojourn time of 10 seconds, while the generation rate of the higher source state is 500 Kbps, with a mean sojourn time of 5 seconds. Each of the two channel states has a mean sojourn time of 20 seconds and the channel factors are 1 and 3 for the lower and higher channel states, respectively. The transition rate matrices for the individual source, α, and channel, β process are given by

$$\alpha = \begin{bmatrix} 0.6 & 0.2 \\ 0.4 & 0.8 \end{bmatrix} \text{ and } \beta = \begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix}$$

Fig. 1 shows the probability of delay exceeding D, $P(d > D)$, as a function of D for the cases of 2 to 6 flows. It is equivalent to find $1 - P(d \leq D)$, where $P(d \leq D)$ is the cumulative probability of delay. This value can be obtained from equation (7), with the insertion of the boundary conditions described in (9), and the inclusion of the steady-state probabilities that can recursively be re-estimated using the updating process outlined in Section III. The general trend of each curve is the decrease of delay violation probability

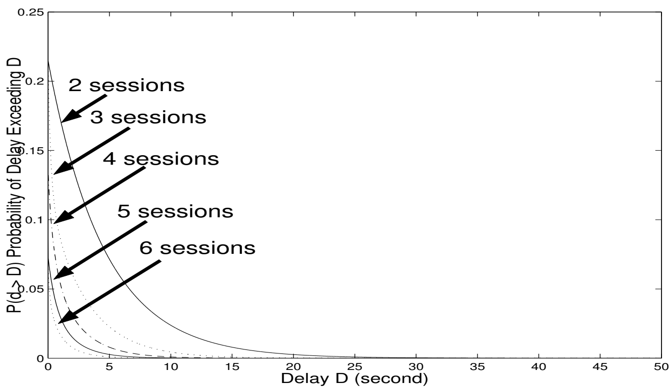

Fig. 1. Numerical results on the delay violation probability as a function of delay threshold, for 2 to 6 flows.

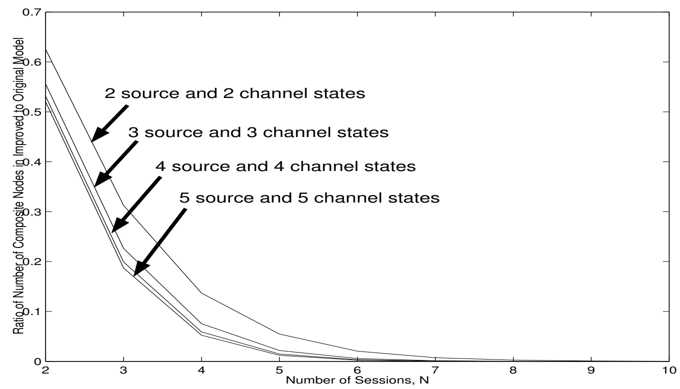

Fig. 2. The ratio of the number of states resulted from the original model to the reduced model for multiplexing systems with 2 to 5 initial source and channel states each.

as delay threshold increases. This is because the cumulative probability of delay approaches 1 at the tail as delay threshold approaches infinity. The figure clearly demonstrates the statistical multiplexing power of the equal-delay scheme, as the delay violation probability for all delay threshold values decrease substantially as the number of multiplexing flows increases.

In calculating the cumulative probability of delay, obtaining the eigenvalues and eigenvectors numerically for the ODE given in equation (6) is a mathematically expensive task. The cost of the operation shoots up dramatically and its precision drops with increasing matrix dimensions. Fortunately, for homogeneous systems, the reduced model provides an alleviation to the burden created by the expensive mathematically procedures. We investigate the savings in the resulting total number of composite states for different number of simple source generation and channel condition states and multiplexing flows. Fig.2 shows the ratio, $\frac{N_{M'}}{N_M}$, of the number of states for the reduced model, $N_{M'}$ to our original model, N_M, as a function of the number of multiplexing flows, F. It can be observed from the figure that $\frac{N_{M'}}{N_M}$ drops rapidly as the number of multiplexing flows increases, indicating a very wide improvement at that region. In addition, greater savings can also be observed for higher number of initial source generation and channel condition states.

On the other hand, the time consumed in finding the composite states and the transitional probability matrix according to the reduced model would become significantly larger as either the number of initial source and channel states or the multiplexing flows increases. This is because the relationship between these quantities vary as the quantities themselves vary, resulting in no definite procedure in determining the composite states. We record the time taken to carry out these computations for each of the cases and observe the longest time to be roughly half an hour on a Pentium IV 1.4GHz PC with 512MB of memory.

VII. Conclusion

In this paper, we have proposed mathematical models for the equal-delay scheme and the multiple-class scheme, and have devised analytical solution for each of the two cases. In particular, we solve the problem caused by the existence of multiple time-scales variation by transforming all of them into a single system time scale, and recursively estimating the steady-state probability distribution for the resulting transformed system. Furthermore, we have shown that an improved model can be applied to homogeneous systems, so that the computational complexity for obtaining the system eigenvalues and eigenvectors can be reduced. Finally, we have presented numerical results to confirm and demonstrate the strengths of the scheduling schemes.

References

[1] E. Lee and D. Taubman "Efficient scheduling schemes for real-time traffic in wireless networks," *Proceedings of IEEE Globecom'05*, 2005.

[2] D. Anick, D. Mitra and M. M. Sondhi "Stochastic theory of a data-handling system with multiple sources," *The Bell System Technical Journal*, Vol. 61, No. 8, October 1982.

[3] J. Daigle and J. Langford "Models for analysis of packet voice communications systems," *IEEE Journal on Selected Areas in Communications*, Vol. SAC-4, No. 6, pp. 847-855, Sep. 1986.

[4] R. Gibbens and P. Hunt "Effective bandwidths for the multi-type UAS channel," *Queueing Systems*, Vol. 9, pp. 17-28, 1991.

[5] C. Coutras, S. Gupta and N. B. Shroff, "Scheduling of real-time traffic in IEEE 802.11 wireless LANs," *Wireless Networks*, Vol. 6, Issue 6, pp. 457-466, Dec. 2000.

[6] Y. Liu and E. Knightly, "Opportunistic fair scheduling over multiple wireless channels," in *Proceedings of IEEE INFOCOM'03*, 2003.

[7] L. Wang, Y. K. Kwok, W. C. Lau and V. Lau, "Channel capacity fair queueing in wireless networks: Issues and a new algorithm," in *Proceedings of IEEE ICC'02*, 2002.

[8] V. Tsibonis, L. Georgiadis and L. Tassiulas, "Exploiting wireless channel state information for throughput maximization," in *Proceedings of IEEE INFOCOM'03*, 2003.