Energy Efficient Quality of Service Traffic Scheduler for MIMO Downlink SVD Channels

Dan J. Dechene, *Student Member, IEEE*, and Abdallah Shami, *Member, IEEE*

***Abstract*—In this paper we focus on minimizing the long-term average power consumption of a single transmitter providing Quality of Service (QoS) enabled traffic to a single receiver. Both the transmitting and receiving stations are equipped with multiple antennas. First, we present a general $\{K \times M\}$ system model where K is the number of independently buffered QoS streams and M is the number of parallel channels available through MIMO SVD eigenmode transmission. Through application of the constrained Markov decision process (MDP) framework combined with a novel MAC layer rate assignment scheme, a randomized per-buffer scheduling policy is obtained. The designed policy exploits queue state information to schedule traffic while meeting throughput, delay and loss constraints. Packets scheduled for transmission during each frame are mapped across the set eigenmode channels subject to available channel resources and the set of channel eigenvalues. Simulation results are provided for several scenarios. System drawbacks, limitations and extensions are also discussed.**

***Index Terms*—Quality of Service, Markov Decision Process, MIMO, Scheduling, Cross-Layer, QSI**

I. Introduction

RECENTLY, the delivery of multimedia services in the home has began to shift from traditional wireline to wireless technologies. Transmission of multimedia traffic however is more complex than that of generic Internet traffic due to stringent Quality of Service (QoS) requirements traffic. As such, it is even more difficult to provide such guarantees over the unreliable wireless medium.

A number of cross-layer protocols have surfaced in recent years, as a method of ensuring QoS over the wireless channel [1]–[8]. In [2] and [4] for example, the authors focus on constrained QoS where they show that knowledge of the instantaneous buffer occupancy, also known as queue state information (QSI), combined with knowledge of the wireless channel can guarantee MAC layer throughput, delay and improve energy performance via dynamic scheduling techniques. These works however are mainly focused on a single queue which is serviced over a single channel which we herein denote these as $\{1 \times 1\}$ systems. Most modern wireless communications systems however, are comprised of multiple input queues with various QoS requirements, and may in turn be transmitted over multiple channel systems (such as those channels provided by Multiple Antenna or MIMO systems). More generally, we define a $\{K \times M\}$ system as a transmission system with K queues as inputs to the system and M is the number of channels available for transmission. While several recent works [3], [7]–[9] have looked at scheduling techniques for supporting QoS in $\{K \times M\}$ systems, to the best of our knowledge there exists no prior works which exploits full QSI in these systems while meeting hard, heterogeneous QoS constraints. In Lau and Chen's work [8] for example, a framework for a delay-optimal power and resource allocation is proposed for such a heterogeneous traffic system, however the weights employed for delay in the optimization framework do not impose hard guarantees on heterogenous average delay and losses which is required for QoS stringent traffic streams.

In this work, we design a cross-layer scheduler in the presence of full QSI for a generic $\{K \times M\}$ system to target specific average delay and packet loss rates. Through application of a novel MAC layer rate assignment scheme, we design a scheduler that is able to exploit full QSI to meet QoS requirements while reducing long-term average power consumption and reduced complexity compared to the full-scale optimization problem.

The remainder of this paper is organized as follows. In Section II we describe both the MAC and PHY layer models used in this work. Later, in Section III we describe our general $\{K \times M\}$ scheduler design. In Section IV we describe how to formulate the scheduler design in general optimization frameworks. Section V provides detailed simulation results and Section VI proposes extensions for time-varying channels. Finally Section VII draws conclusions on this work.

II. System Model

The system model used in the work is a general $\{K \times M\}$ downlink model where K is used to denote the number of independent MAC layer queues as an input to the system and M is used to denote the number of PHY layer channels available for transmission. The overall system model is shown in Fig. 1 where the MAC and PHY layer subcomponents are discussed below.

A. MAC Layer Model

The media access control (MAC) layer model used in this work is as follows. Consider the downlink system shown in Fig. 1. Traffic is received from upper layers and classified into K traffic streams. A single traffic stream has an associated set of QoS parameters $\{D_i, L_i, \overline{\lambda}_i, B_i, \delta_i\}$ which denotes the maximum tolerable average delay, packet length, average arrival rate, buffer size and maximum tolerable packet loss rate respectively for that stream. Each stream may represent a broad service class (such as voice over IP or video) or a

Manuscript received December 17, 2009; revised June 15, 2010; accepted September 26, 2010. The associate editor coordinating the review of this paper and approving it for publication was E. Hossain.

The authors are with the Department of Electrical & Computer Engineering, The University of Western Ontario, London, Ontario, Canada N6A 5B9 (e-mail: ddechene@ieee.org, ashami2@uwo.ca).

Digital Object Identifier 10.1109/TWC.2010.100510.091821

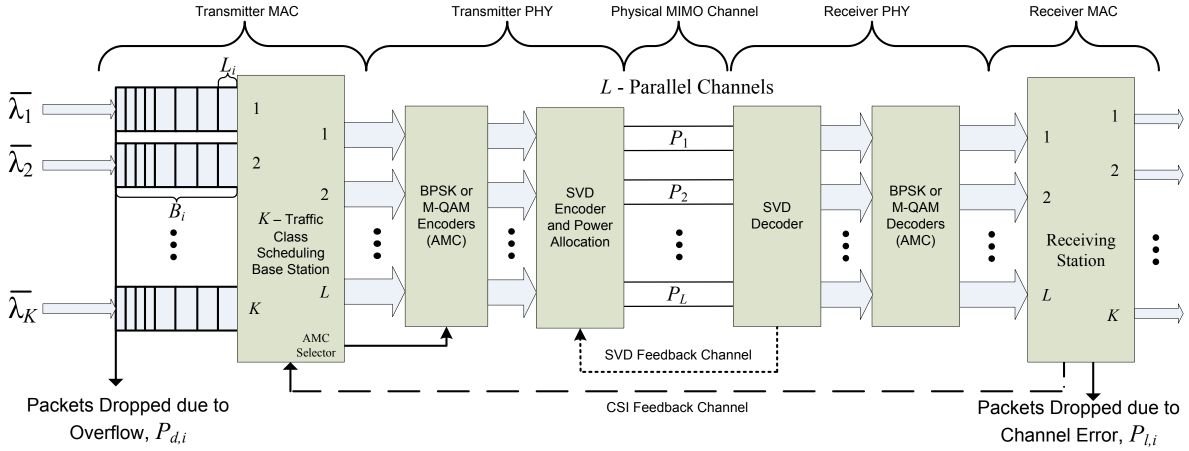


Fig. 1. Multi-queue cross-layer model.

particular application-layer stream. Each incoming stream is stored in a finite-length first-in, first-out (FIFO) buffer where incoming packets are dropped when the buffer is full.

The target loss (δ_i) can be further broken down into packets dropped at the source (due to buffer overflow, $P_{d,i}$) and packets dropped at the destination (due to channel errors, $P_{l,i}$). The total probability of an erroneous packet for queue i is $\delta_i = 1-(1-P_{d,i})(1-P_{l,i})$. The total average tolerable loss rate is known for each particular traffic class. In this work, we also assume that the target total probability of packet failure on the channel is also known.

The time horizon is divided into fixed scheduling intervals denoted as frames. Each frame has a duration of T_f seconds. Packets arriving during the frame are assumed to be enqueued at the end of the frame. The time duration lying in $[nT_f,(n+1)T_f)$ is denoted as time frame n.

The K buffers are statistically multiplexed into a QoS-aware K-queue scheduler which makes scheduling decisions based on the CSI feedback from the subscribing station, the number of packets in each of the K MAC buffers, and their parameter set. While the theoretical number of queues (K) with varying QoS constraints is large, practical implementations employ a finite number of classes [10] as most multimedia services can be categorized into one of several QoS classes.

B. PHY Layer Model

The PHY layer transmits packets scheduled for transmission during each frame. Packets are separated into $M = \min(M_T, M_R)$ parallel streams and encoded with BPSK or M-QAM, where the constellation is decided based on channel conditions and MAC rate demands. The M streams are then reconstructed at the receiver and forwarded to the receiver MAC. The M streams are provided by MIMO singular value decomposition (SVD)[1] achieved by assuming full channel knowledge is available at the transmitter error-free. The state of the M channels is characterized by their ordered eigenvalues where $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_M^2 \geq 0$. For simplicity at this stage, we assume these eigenvalues are known, and do not change in time.

In general, the maximum number of parallel channels is equal to the minimum number of antennas at either the transmitting or receiving station. Recent measurement campaigns conducted in urban environments however suggest [11], [12] that there are only a finite number of resolvable non-zero eigenvalues, which in general, can be less than the number of antennas.

Each independent channel is subject to noise. In the presence of additive white Gaussian noise (AWGN), the bit error rate of an uncoded M-ary signal is approximately [13]

$$Pb(\gamma_j, M_j) \approx 0.15 \exp\left(\frac{-1.55\gamma_j}{M_j - 1}\right), j = 1, 2, \dots, M \quad (1)$$

where M_j is the size of the constellation set used in channel j, γ_j is the per symbol SNR given as $\gamma_j = P_j \gamma_0 \lambda_j^2$, P_j is the power allocated to channel j and γ_0 is the reference SNR level.

C. System Operation

From frame $n-1$ to frame n, the evolution of each buffer i follows

$$u_i(n) = \min\{B_i, \max\{0, u_i(n-1) - c_i(n)\} + A_i(n)\} \quad (2)$$

where $u_i(n)$ describes the buffer occupancy (number of buffer spaces in use) at time frame n, B_i denotes the maximum buffer occupancy, $c_i(n)$ denotes the number of transmitted packets during the frame n (*i.e.,* the transmission action taken by queue i) and $A_i(n)$ denotes the number of arrivals in the queue. Here, $c_i(n), u_i(n), A_i(n) \geq 0, \forall n$ and $c_i(n), u_i(n), A_i(n) \in \mathbb{I}$ where $\mathbb{I}$ is the set of all Integers. The number of arrivals during frame n to a given queue (or $A_i(n)$) is a Poisson process with an average arrival rate of $\bar{\lambda}_i$ and a constant packet length of L_i. Further to this, packet arrivals are assumed to be independent of the current queue occupancy, service process and arrivals to other queues. For a Poisson process, the probability of k packets arriving to queue i during a frame of duration T_f is well-known to be

$$Pr[A_i(T_f) = k] = \begin{cases} \frac{(\bar{\lambda}_i T_f)^k \exp(-\bar{\lambda}_i T_f)}{k!}, & \text{if } k \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

[1] The framework can easily be extended to any multi-channel system with known subchannel error performance.

Packets are serviced in FIFO discipline over the previously described MIMO physical layer. In each individual parallel channel, bits are encoded from a finite adaptive modulation and coding (AMC) alphabet $\mathcal{M}$ which determines the number of bits that can be encoded onto a single symbol. The selection of the set is described in later sections. Denoting k_j as the spectral efficiency in bits/symbol for choosing a constellation of size $M_j \in \mathcal{M}$ for the j^{th} channel and T_s as the symbol duration, the maximum number of bits that can be transmitted through the j^{th} channel over a duration of T_f seconds is $\frac{k_j T_f}{T_s}$ with a bit error rate given in (1).

III. $\{K \times M\}$ Scheduler Design

The proposed scheduler utilizes queue state information (QSI) to design a scheduling policy $\mathbf{\Omega}$. A given scheduling policy $\mathbf{\Omega}$ describes the channel, power and rate assignment for all time frames n. The benefit of utilizing QSI has been well demonstrated by [1] and [2] in the case of a $\{1 \times 1\}$ system, however extensions for general $\{K \times M\}$ systems are non-trivial. Now, consider the following practical observation.

Examining only the buffer states for a set of K queues, the global set of states to describe the joint occupancy level across all queues is a K-dimensional set spanning the possible occupancy levels for the model discussed above and can be expressed as $\mathcal{U} := \mathcal{U}_1 \times \cdots \times \mathcal{U}_K$ where $\mathcal{U}_i = \{0, 1, \ldots, B_i\}$ is the state set for any buffer i and where 0 denotes an empty buffer.

It was previously noted that arrivals to each queue are independent, however the service process couples the queues. In practice however, the number of possible service rates is comprised of a finite subset of the possible buffer states, and in our work physically represents the number of packets taken from the queue during a given frame. We previously denoted this quantity as $c_i(n)$ (*i.e.,* the transmission action taken by queue i during frame n. Now suppose, we have a set of rates which we denote $\mathcal{C}_i$ as the set of possible MAC rates (measured in packets) that can be serviced during each frame for queue i (*i.e.,* all possible values that can be taken on by $c_i(n),\ \forall n$). This set is independent of the current frame n and we assume that $\mathcal{C}_i$ is chosen such that the maximum simultaneous packet transmission rate is achievable (*i.e.,* admission control is performed in advance to ensure the maximum transmission rate of all streams is less than the maximum rate on the channel). The overall MAC rate state-space of transmission actions is by extension simply $\mathcal{C} := \mathcal{C}_1 \times \cdots \times \mathcal{C}_K$. Assuming that $|\mathcal{C}_i| << |\mathcal{U}_i|$ which we argue occurs in practice, then by trivial extension $|\mathcal{C}| << |\mathcal{U}|$, where $|\cdot|$ denotes the size of a set. The preceding implies that any channel mapping and power control scheme need only consider $|\mathcal{C}|$ possible MAC layer rates rather than $|\mathcal{U}|$ possible states as a method of reducing the system complexity.

Using the above arguments, given a predetermined number of MAC service rates $\mathcal{C}_i$ for each queue i, the design of a $\{K \times M\}$ scheduler can be constructed as two components:

A: A mechanism to determine how to map a set of packets during each frame across all parallel channels while performing rate and power adaptation to minimize power usage and maintain channel error rate requirements. This is computed for each $c \in \mathcal{C}$, where $\mathcal{C}$ is the MAC layer rate state-space, and

B: A mechanism to select the appropriate MAC layer transmission rate from each queue during each time frame to minimize total average transmission power found for $c \in \mathcal{C}$ in addition to ensuring QoS constraints are met by exploiting QSI.

The above problem segmentation allows the power, rate and channel allocation problems to be solved with reduced complexity by considering only a subset of information in each stage. As a result of the segmentation, components comprising the total average packet losses (*i.e.,* dropping probability and channel error probability) are constrained individually at each stage.

A. Channel Mapping and Power Control

The first component of the scheduling mechanism maps a set of packets during a frame over the set of parallel channels. This procedure is performed for each possible MAC layer rate combination (*i.e.,* each $c \in \mathcal{C}$). The resulting outputs of this component are:

- A bit loading map $X_{j,i}(c), \forall j, i$ which denotes the number of bits per queue i that are mapped to channel j during the time frame,
- Constellation selection for each channel ($k_j(c), \forall j$), and
- Power assignment for each channel ($P_j(c), \forall j$)

where $c \in \mathcal{C}$. Relevant complexity issues will be addressed in a later Section. Further, we have the following constraints applied in this component.

1) Packet Loss Due to Channel Error Constraint: Firstly, the average packet error rate experienced by a packet from stream i in state c given as

$$PER_i(c) = 1 - \prod_{j=1}^{M} (1 - P_{b,j}(c))^{x_{j,i}(c) L_i}, \forall i \tag{4}$$

where $P_{b,j}(c)$ is the bit error rate for channel j with power $P_j(c)$ assigned, spectral efficiency of $k_j(c)$ during state c given from (1) as $P_{b,j}(c) = Pb(P_j(c)\gamma_0\lambda_j^2, 2^{k_j(c)})$ and $x_{j,i}(c)$ is the percentage of bits per packet transmitted in the channel j from queue i such that $\sum_{j=1}^{M} x_{j,i}(c) = 1$.

Next, we assume the average packet error rate is targeted at each instant of time (*i.e.,* $PER_i(c) = \mathbb{E}_c[PER_i(c)]$ where $\mathbb{E}_c[\cdot]$ is the expectation over $\mathcal{C}$). The constraint on channel losses is then

$$1 - \prod_{j=1}^{M} (1 - P_{b,j}(c))^{x_{j,i}(c) L_i} \le P_{l,i}, \quad \forall c, i \tag{5}$$

where $P_{l,i}$ is the target channel loss rate and is a portion of the total loss rate.

To further simplify (5), it can be approximated as[2]

$$1 - \left(1 - \sum_{j=1}^{M} x_{j,i}(c) P_{b,j}(c)\right)^{L_i} \le P_{l,i}, \forall i \tag{6}$$

[2] See Proof Online: http://www.dechene.ca/TWC/perproof.pdf

Defining $X_{j,i}(c)$ as the number of bits mapped to channel j from queue i over a frame of duration T_f. Relating this to the quantity $x_{j,i}(c)$ discussed above we have

$$X_{j,i}(c) = x_{j,i}(c)c_i(c)L_i = x_{j,i}(c)R_i(c) \tag{7}$$

where $c_i(c)$ is defined as before in number of packets taken from queue i in state c and $R_i(c)$ is this quantity measured in bits. From (6), the final expression for the packet loss constraint is

$$P_{l,i} \geq 1 - \left(1 - \sum_{j=1}^{M} \frac{X_{j,i}(c)P_{b,j}(c)}{R_i(c)}\right)^{L_i}, \forall i \tag{8}$$

Which can be constrained in terms of the bit error rates such that

$$\sum_{j=1}^{M} \frac{X_{j,i}(c)P_{b,j}(c)}{R_i(c)} \leq BER_i, \forall i, \tag{9}$$

where $BER_i = 1 - (1 - P_{l,i})^{\frac{1}{L_i}}$

2) Rate Selection Constraints: We also select constellation schemes for each channel such that the requested MAC layer rate requirement is met. The total MAC layer rate requirement is given as $\sum_{i=1}^{K} R_i(c)$. Therefore we note that

$$\sum_{j=1}^{M} \frac{k_j(c)T_f}{T_s} \geq \sum_{i=1}^{K} R_i(c) \tag{10}$$

is a necessary condition. We also note that the set $\{k_j\}_{j=1}^{M}$ that satisfies the above such that $\sum_{j=1}^{M} k_j$ should be minimized to achieve the minimum power usage.

Since by design $k_j \in \mathcal{M}$ is a discrete set and (1) is monotonic in k_j, one can easily show that any set $\{k_j\} \in \mathcal{M}$ that minimizes transmission power must satisfy

$$\sum_{j=1}^{M} k_j(c) = \left\lceil \frac{T_s}{T_f} \sum_{i=1}^{K} R_i(c) \right\rceil \tag{11}$$

where in this case, $\lceil \cdot \rceil$ denotes rounding up to the nearest valid sum of k_j values. The set of all valid AMC mode combinations that satisfies the above for each $c \in \mathcal{C}$ is denoted $\mathcal{K}_{min}(c)$.

While (11) is true if valid spectral efficiencies in each channel increment by 1, the values that can be taken on by k_j fall within the set of allowable AMC modes and in this work is further restricted to $\{0, 1, 2, 4, 6\}$ representing no transmission, BPSK, QPSK and 16/64-QAM respectively. Therefore a marginal increase in transmission rate may result in an minimum increase of spectral efficiency of 2 in one channel. To integrate this phenomenon, the search set of valid AMC modes is extended such that

$$\mathcal{K}(c) = \begin{cases} \mathcal{K}_{min}(c) \bigcup \mathcal{K}_{min+1}(c), & \sum_{i=1}^{K} R_i > 0 \\ 0 & otherwise \end{cases} \tag{12}$$

where $\mathcal{K}_{min+1}(c)$ is the set of AMC modes satisfying $\left\lceil \frac{T_s}{T_f} \sum_{i=1}^{K} R_i(c) \right\rceil + 1$.

3) Channel Mapping Constraints: Based on the system design, we also have the following constraints on the mapping coefficients $X_{j,i}(c)$:

$$\sum_{j=1}^{M} X_{j,i}(c) = R_i(c), i = 1, 2, \ldots K \tag{13}$$

$$\sum_{i=1}^{K} X_{j,i}(c) \leq \frac{k_j(c)T_f}{T_s}, j = 1, 2, \ldots, M \tag{14}$$

4) Power Control: Minimization of the average applied power is the objective function. Average power is computed as the sum of power allocated in each subchannel multiplied by the number of symbols transmitted over that subchannel or simply

$$\sum_{j=1}^{M} \frac{P_j(c)}{k_j(c)} \sum_{i=1}^{K} X_{j,i}(c) \tag{15}$$

The above set of constraints, which contains both Integer and discrete variables can be formulated as a mixed-Integer non-linear programming (MINLP) problem. The details of the MINLP formulation and elements are discussed in Section IV-A.

B. Locally Optimized MAC Rate Selection

The second component of the $\{K \times M\}$ scheduler design is to select a MAC layer transmission rate (or $c_i \in \mathcal{C}_i$) to determine the number of packets to transmit during frame n from queue i. This decision is based on both the QoS constraints and the power allocation computed in (15).

Each queue is characterized by its current state $u_i \in \mathcal{U}_i$ denoting the current occupancy level. During anytime n, c_i packets may be taken from the queue where $\mathcal{C}_i$ is the set of all transmission actions (packets that can be transmitted) during a given frame. The scheduling policy $\mathbf{\Omega}$ defines the set probabilities of choosing c_i when the current queue state is u_i for each queue i. From (2) it can be seen that the queue occupancy during any frame n depends only on the occupancy during frame $n-1$ and arrivals during that frame. As such, the above can be solved as constrained Markov decision process (CMDP) [4] to obtain a scheduling policy $\mathbf{\Omega}$. Let $\theta_i(c_i, u_i|\mathbf{\Omega})$ be a steady-state distribution function that exists for a particular policy $\mathbf{\Omega}$ which denotes the probability of being in state u_i and transmitting c_i packets during frame n. The scheduling policy $\mathbf{\Omega}$ is obtained through application of the constraints on average delay and MAC layer throughput given as follows.

1) Throughput Constraint: The dropping probability is related to the MAC throughput by

$$\overline{\chi}_i = \lambda_i(1 - P_{d,i})T_f \tag{16}$$

Therefore we do not constrain the dropping probability directly, but rather constrain the minimum MAC layer throughput. The throughput at each state u_i is dependant on both the queue state and the action taken (*i.e.,* u_i and c_i). For a given set $\{c_i, u_i\}$ during frame n, the throughput is

$$\chi_{i:n}(c_i, u_i) = \min(c_i, u_i) \tag{17}$$

2) Delay Constraint: From Little's Theorem, the average queueing delay constraint is

$$D_i \geq \mathcal{D}_i = \frac{\bar{q}_i}{\lambda_{q,i} T_f} \tag{18}$$

where $\bar{q}_i$ is the average queue size and $\lambda_{q,i}$ is the average enqueued arrival rate for queue i. By design we can express $\bar{q}_i$ using the steady-state distribution $\theta_i(c_i, u_i|\mathbf{\Omega})$ as:

$$\bar{q}_i = \sum_{u_i \in \mathcal{U}_i} u_i \sum_{c_i \in \mathcal{C}_i} \theta_i(c_i, u_i|\mathbf{\Omega}) \tag{19}$$

and since $\lambda_{q,i}$ is also equal to the average service rate in steady-state, it can be expressed as

$$\lambda_{q,i} = \sum_{u_i \in \mathcal{U}_i} \sum_{c_i \in \mathcal{C}_i} \min(c_i, u_i)\theta_i(c_i, u_i|\mathbf{\Omega}) \tag{20}$$

3) Transition Probabilities: The transition probabilities denote the probability of transitioning from one queue state to another. By design this is based on the arrival process, the given state, the next state and the transmission action taken. With all quantities defined as before, we denote $p^{c_i}_{u_i;u_i'}$ as the probability of transitioning from state u_i to u_i' given action c_i is taken. From (2) and (3) this is given as

$$p^{c_i}_{u_i;u_i'} = \begin{cases} P(A_i(T_f) = u_i' - [u_i - \min(u_i, c_i)]), u_i' < B_i \\ \displaystyle\sum_{j=B_i-[u_i-\min(u_i,c_i)]}^{\infty} P(A_i(T_f) = j), \qquad\qquad u_i' = B_i \end{cases} \tag{21}$$

By design, the steady-state distribution $\theta_i(c_i, u_i|\mathbf{\Omega})$ must also satisfy the following balance property

$$\sum_{u_i' \in \mathcal{U}_i} \sum_{c_i' \in \mathcal{C}_i} \theta(c_i', u_i'|\mathbf{\Omega}) p^{c_i'}_{u_i';u_i} = \sum_{c_i \in \mathcal{C}_i} \theta(c_i, u_i|\mathbf{\Omega}), \forall u_i \tag{22}$$

C. Per Queue Objective Function

The computed power allocation found in the first component is used to derive the objective function for the local MAC layer rate assignment. First, the average marginal cost for taking an action c_i in queue 1 can be given as

$$\Upsilon_{1,x} = \sum_{c_2 \in \mathcal{C}_2} \cdots \sum_{c_K \in \mathcal{C}_K} P(x, c_2, \ldots, c_K) \cdot \pi_2(c_2|\mathbf{\Omega}) \times \ldots \times \pi_K(c_K|\mathbf{\Omega}) \tag{23}$$

where there are $i-1$ summations. Similar expressions can be found for all actions $c_i \in \mathcal{C}_i$ and found for all queues $k = 1, \ldots, K$ and where

$$\pi_i(x|\mathbf{\Omega}) = \sum_{u_i \in \mathcal{U}_i} \theta(x, u_i|\mathbf{\Omega}), x \in \mathcal{C}_i \tag{24}$$

The above steady-state action probabilities are coupled through the policy $\mathbf{\Omega}$. The value $P(c_1, c_2, \ldots, c_K)$ is the total power associated with taking actions c_1 through c_K in each queue (or one for each state $c \in \mathcal{C}$) found as the solution to (15). Here we need to highlight that the above expression contains the steady-state probability of choosing an action in each queue. The result of which implies that it is not possible to directly decouple and consider each queue independently. We can however consider the following special cases.

1) *Single Queue:* For the single queue case, the cost function in (23) reduces to the total power required and can be solved as an linear programming (LP) problem.
2) *Two Queues:* For the two queue scenario, the cost function model can be considered a Quadratic Programming (QP) problem where the number of degrees of freedom is twice that of the single queue problem or $|\mathcal{C}_1 \times \mathcal{U}_1| + |\mathcal{C}_2 \times \mathcal{U}_2|$, rather than $|\mathcal{C}_1 \times \mathcal{U}_1 \times \mathcal{C}_2 \times \mathcal{U}_2|$.
3) *General Number of Queues:* In general, the problem can be solved using iterative methods. This process is as follows. Firstly, given the power allocation values, we iteratively solve the C-MDP problem (as an LP problem) for each queue and update the corresponding cost function until the steady state distribution $\theta_i(c_i, u_i|\mathbf{\Omega})$ in each queue converges. Convergence details are discussed in Section V-C.

IV. Forming Programming Elements

Both the channel/power allocation and the local MAC rate assignment mechanisms are formulated as optimization problems. The channel and power allocation scheme can be formulated as a generic MINLP problem, while the local MAC rate assignment can be formulated as a general LP (or QP) problem.

A. Formation of MINLP Problem

A general solution to a NLP problem is non-trivial, this is further complicated by introduction of discrete or Integer constraints on several variables. To relax these discrete constraints we perform the following:

1) Consider $X_{j,i}(c)$ as a continuous variable as we note in practice rounding $X_{j,i}$ to the nearest Integer affects only a single bit of information. Since during any frame where there is an active the transmission the number of transmitted bits is in general much larger than 1, rounding does not dramatically affect the result, and
2) We formulate a general NLP problem for each subset satisfying (11) and choose the allocation strategy achieving the lowest power consumption.

Based on the above, we formulate a general NLP problem such that we solve $\arg\min_{\mathbf{x}} f(\mathbf{x})$ subject to $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, $\mathbf{A}_{eq}\mathbf{x}_{eq} = \mathbf{b}_{eq}$, $\mathbf{x} \geq 0$ and $\mathbf{c}(\mathbf{x}) \leq 0$ where $\mathbf{A}$ and $\mathbf{A}_{eq}$ are matrices, $\mathbf{b}$ and $\mathbf{b}_{eq}$ are vectors, $\mathbf{c}(\mathbf{x})$ is a vector of nonlinear functions evaluated at $\mathbf{x}$ and $f(\mathbf{x})$ is a scalar nonlinear function evaluated at $\mathbf{x}$. The above is evaluated at each state $c \in \mathcal{C}$ and over the space $\mathcal{K}(c)$ which is the space containing each combination $\{k_j\}_{j=1}^{M}$ that meets the rate selection restrictions above described in (11).

The derivations of the NLP elements is given below. The vector $\mathbf{x}$ is a $(M + MK) \times 1$ vector with elements $P_j, j = 1, 2, \ldots, M$ and $X_{j,i}, j = 1, 2, \ldots, M, i = 1, 2, \ldots, K$ given as

$$\mathbf{x} = [P_1, \ \ldots, \ P_M, \ X_{1,1}, \ \ldots, \ X_{1,K}, \ \ldots, \ X_{M,K}]^T \tag{25}$$

TABLE I
FREQUENTLY USED NOTATION

Quantity	Symbol	Quantity	Symbol
Number of Traffic Streams	K	Subchannel SNR	γ_j
Number of Parallel Channels	M	BER of Channel j	$P_{b,j}$
Average Delay Constraint	$\mathcal{D}_i$	Scheduling Policy	$\mathbf{\Omega}$
Packet Size in Bytes	L_i	Buffer state-space	$\mathcal{U}_i$
Average Arrival Rate	$\bar{\lambda}_i$	MAC Rate state-space	$\mathcal{C}_i$
Buffer Size	B_i	Joint MAC Rate state-space	$\mathcal{C}$
Total Average Loss Constraint	δ_i	Transition probability	$p^{c_i}_{u_i,u_i'}$
Packet Dropping Probability	$P_{d,i}$	Fraction of bits allocated to channel j from stream i	$x_{j,i}$
Probability of Channel Packet Loss	$P_{l,i}$	Number of bits allocated to channel j from stream i	$X_{j,i}$
Frame Duration	T_f	Stream Rate of channel i	R_i
Symbol Duration	T_s	Throughput of stream i	$\bar{\chi}_i$
Frame Number	n	Per queue cost function	$\Upsilon_{i,x}$
Subchannel Eigenvalue	λ_j^2	Steady-State policy distribution	$\theta_i(c_i, u_i\|\mathbf{\Omega})$
Reference SNR	γ_0	Steady-state action probability	$\pi_i(x\|\mathbf{\Omega})$
Set of Valid AMC Modes	$\mathcal{M}$	Number of Arrivals during frame n	$A_i(n)$
Spectral efficiency in channel j	k_j	Buffer Occupancy during frame n	$u_i(n)$
M-ary Mode	M_j	Packet Service rate of queue i during frame n	$c_i(n)$

1) Objective Function: The objective function from $f(\mathbf{x})$ given in (15) is formulated as

$$f(\mathbf{x}) = \sum_{j=1}^{M} \frac{\mathbf{x}(\mathcal{P}_j)}{k_j} \sum_{i \in \mathcal{I}'_j} \mathbf{x}(i) \tag{26}$$

where $\mathcal{I}'{}_j$ and $\mathcal{P}_j$ are the sets containing location indices of $X_{j,i}, \forall i$ and P_j respectively in $\mathbf{x}$.

2) Equality Constraints: The K equality constraints from (13) are given in the $K \times (M + MK)$ matrix $\mathbf{A}_{eq}$ with entries

$$A_{eq:i,k} = \begin{cases} 1, k \in \mathcal{I}_i \\ 0, otherwise \end{cases} \tag{27}$$

where $\mathcal{I}_i$ is the set containing location indices of $X_{j,i}, \forall j$ in $\mathbf{x}$. The coefficient vector $\mathbf{b}_{eq}$ is given as

$$\mathbf{b}_{eq} = [R_1,\ R_2, \ldots,\ R_K]^T \tag{28}$$

3) Inequality Constraints: The M equality constraints from (14) are defined in the $M \times (M + MK)$ matrix $\mathbf{A}$ with entries

$$A_{j,k} = \begin{cases} 1, k \in \mathcal{I}'{}_j \\ 0, otherwise \end{cases} \tag{29}$$

The coefficient vector $\mathbf{b}$ is given as

$$\mathbf{b} = \frac{T_f}{T_s}[k_1,\ k_2, \ldots,\ k_M]^T \tag{30}$$

4) Non-Linear Inequality Constraint: The K non-linear inequality constraints are given as a $K \times 1$ vector of functions of $\mathbf{x}$. For simplicity of notation with further define $\mathcal{I}_{j,i}$ as the indices of $X_{j,i}, \forall i, j$. Using the bit error rate expression in (1), we then have

$$\mathbf{c}(\mathbf{x}) = [c_1(\mathbf{x}),\ \ldots,\ c_K(\mathbf{x})]^T \tag{31}$$

where

$$c_i(\mathbf{x}) = \frac{\displaystyle\sum_{j=1}^{M} 0.15 \exp\left(\frac{-1.55\mathbf{x}(\mathcal{P}_j)\gamma_0\lambda_j^2}{(2^{k_j}-1)}\right)\mathbf{x}(\mathcal{I}_{j,i})}{R_i} - BER_i$$

The above framework can now be computed using general NLP methods such as those provided by *fmincon* included in the MATLAB optimization toolbox.

B. Forming LP Problem

As in [4], the constrained MDP problem formulated for the local MAC rate selection can be solved using Linear Programming (LP) techniques for each queue. LP techniques efficiently solve convex optimization problems of the form $\underset{\mathbf{x}}{\arg\min}\, \mathbf{c}^T\mathbf{x}$, subject to $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, $\mathbf{A}_{eq}\mathbf{x} = \mathbf{b}_{eq}$, $\mathbf{x} \geq 0$ where $\mathbf{A}$ and $\mathbf{A}_{eq}$ are matrices and $\mathbf{x}, \mathbf{b}, \mathbf{b}_{eq}$ and $\mathbf{c}$ are column vectors. The vector $\mathbf{x}$ is the solution to the optimization problem. In our problem, the elements are given as

$$\mathbf{x} = [\boldsymbol{\theta}_i(\mathcal{C}_i, 0|\mathbf{\Omega}), \ldots, \boldsymbol{\theta}_i(\mathcal{C}_i, B_i|\mathbf{\Omega})]^T \tag{32}$$

with each $\boldsymbol{\theta}_i(\mathcal{C}_i, u_i|\mathbf{\Omega})$ being a row vector with entries for each $c_i \in \mathcal{C}_i$.

1) Objective Function: The objective function is of the form $\mathbf{c}^T\mathbf{x}$. The vector $\mathbf{c}$ is comprised of the total power cost for taking an action. Each entry of $\mathbf{c}$ corresponds to the entry in $\mathbf{x}$ with the value of entries in $\mathbf{c}$ given by Υ_{i,c_i} in (23).

$$\mathbf{c} = [\underbrace{\Upsilon_{i,1}, \ldots, \Upsilon_{i,|\mathcal{C}_i|}}_{1}, \overbrace{\ldots, \ldots, \ldots}^{2..B_i}, \underbrace{\Upsilon_{i,1}, \ldots, \Upsilon_{i,|\mathcal{C}_i|}}_{B_i+1}] \tag{33}$$

2) Equality Constraints: The equality constraints are comprised of the balance equations and the causality constraint (total probability space) given in (21) and (22) respectively. In matrix form, the balance equations can be expressed as $\mathbf{P} \times \mathbf{x} = \mathbf{\Phi}_0 \times \mathbf{x}$ where $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}^{\mathcal{C}_i}_{0;0} & \cdots & \cdots & \mathbf{p}^{\mathcal{C}_i}_{B_i;0} \\ \vdots & \mathbf{p}^{\mathcal{C}_i}_{1;1} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{p}^{\mathcal{C}_i}_{0;B_i} & \cdots & \cdots & \mathbf{p}^{\mathcal{C}_i}_{B_i;B_i} \end{bmatrix} \tag{34}$$

with $\mathbf{p}^{\mathcal{C}_i}_{q;q'}$ as a $1 \times |\mathcal{C}_i|$ row vector with entries

$$\mathbf{p}^{\mathcal{C}_i}_{q;q'} = [p^1_{q;q'}, \ldots, p^{|\mathcal{C}_i|}_{q;q'}] \tag{35}$$

and the quantity $\mathbf{\Phi}_0$ is given as the B_i+1 row matrix

$$\mathbf{\Phi}_0 = \begin{bmatrix} \mathbf{1}_{1\times|\mathcal{C}_i|} & 0 & \cdots & 0 \\ 0 & \mathbf{1}_{1\times|\mathcal{C}_i|} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}_{1\times|\mathcal{C}_i|} \end{bmatrix} \tag{36}$$

Combining the above with the causality constraint on the total probability space we have our overall equality constraints given as

$$\mathbf{A}_{eq} = \begin{bmatrix} \mathbf{P} - \mathbf{\Phi}_0 \\ \mathbf{1}_{1\times(|\mathcal{C}_i|(B_i+1))} \end{bmatrix} \qquad \mathbf{b}_{eq} = [\mathbf{0}_{1\times(B_i+1)} \quad 1]^T \tag{37}$$

3) Inequality Constraints: The inequality constraints are used to describe the throughput and delay constraints and are given by (16) and (18) respectively. These constraints are given in two parts as

$$\mathbf{A} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} \tag{38}$$

where $\mathbf{w}_1$ is given as

$$\mathbf{w}_1 = -[\boldsymbol{\chi}_{i:n}(\mathcal{C}_i, 0), \ldots, \boldsymbol{\chi}_{i:n}(\mathcal{C}_i, B_i)] \tag{39}$$

where $\boldsymbol{\chi}_{i:n}(\mathcal{C}_i, u_i)$ is a row vector with entries $\chi_{i:n}(c_i, u_i)$ for all $c_i \in \mathcal{C}_i$ and $\mathbf{z}_1$ is given as

$$\mathbf{z}_1 = -\bar{\lambda}_i(1 - P_{d,i})T_f \tag{40}$$

Further, using (18)-(20), and after some trivial manipulation we can obtain

$$\mathbf{w}_2 = \mathbf{Q} \times \mathbf{\Phi}_0 - \mathcal{D}_i\mathbf{U} \qquad \mathbf{z}_2 = 0 \tag{41}$$

where $\mathbf{Q} = [0, 1, \ldots, B_i]$ and $\mathbf{U}$ is given in (42).

The above framework can be computed using general LP methods such as those provided by *linprog* included in the MATLAB optimization toolbox for both the single queue and iterative methods. Extensions are trivial for the special case of two queues using QP methods.

The LP problem above yields the steady-state distribution $\theta_i(c_i, u_i|\mathbf{\Omega}), c_i \in \mathcal{C}_i, u_i \in \mathcal{U}_i, i \in \{1, 2, \ldots, K\}$. The physical meaning of this solution is that in a given queue i while the buffer is in a given state u_i, the scheduler selects c_i packets from the queue for transmission with probability given by

$$\Pr[c_i|u_i, i, \mathbf{\Omega}] = \frac{\theta_i(c_i, u_i|\mathbf{\Omega})}{\sum\limits_{c_i' \in \mathcal{C}_i} \theta_i(c_i', u_i|\mathbf{\Omega})} \tag{43}$$

C. Scheduler Implementation

The above optimization problem is solved offline, in advance with proper channel measurements. The advantage of this method is that all quantities can be stored in a lookup table (LUT). The LUT stores information on the power, AMC mode and bit allocation for each state $c \in \mathcal{C}$. The offline scheduler works as follows. At the beginning of each frame n, each queue has u_i packets waiting for transmission. All queues then choose actions c_i with probability given in (43). Given a joint action $c = \{c_i, \forall i\}$, the scheduler selects the stored AMC and power modes for each channel and allocates bits from all queues as found in the stored bit allocation.

TABLE II
SIMULATION PARAMETERS

Parameter	Value
Number of Antennas (M)	4
Number of Queues (K)	1
Spectral Efficiencies	$\{0,\ 1,\ 2,\ 4,\ 6\}$
Length of Packet (bits)	200
Arrival Rate (Packets/frame)	1
Queue Size (Packets)	25
Average Packet Delay (Frames)	5
Total Loss Rate (δ, % of Packets)	1%
Target Channel Loss Rate	$\delta/2$
Frame Duration (T_f)	1
Symbol Duration (T_s)	0.01
MAC Rates (Packets per Frame)	$\{0, 1, 2, 3, 4, 5\}$
MIMO Channel Eigenvalues	$[2,\ 1.5,\ 0.6,\ 0.4]$
Reference SNR (γ_0)	20dB

TABLE III
SIMULATION PARAMETERS - 2 QUEUES

Parameter	Value
Length of Packet (bits)	[200 250]
Arrival Rate (Packets/frame)	[1 1]
Queue Size (Packets)	[25 25]
Average Packet Delay (Frames)	[4 5]
Total Loss Rate (δ, % of Packets)	[1 1]%
Target Channel Loss Rate	$\delta/2$

All quantities in the LUT can be accurately stored as 64 bit double. The space required to store each state is

$$Size_c = 64(KM + 2M) \quad \text{bits/state} \tag{44}$$

as we require storage of KM bit allocations, M power levels and M AMC mode selections. Further, we note the system has $|\mathcal{C}|$ states, therefore the total size of the LUT in bits is given as

$$Size_{LUT} = |\mathcal{C}|Size_c = 64|\mathcal{C}|(KM + 2M) \quad \text{bits} \tag{45}$$

where the above describes the relation of number of channels, the number of queues and the possible MAC rates to the storage size of the LUT.

V. SIMULATION RESULTS

We provide simulation results for the three MAC layer rate assignment approaches. Firstly, results are provided for the special case of a single queue using the LP approach, followed by application of the QP approach for the case of two queues. Finally we validate the iterative approach by comparing the accuracy of the two queue system with QP approach. Convergence details of the iterative method are also

$$\mathbf{U} \;=\; [\min(1,0), \min(2,0), \ldots, \min(|\mathcal{C}_i|,0), \min(1,1), \ldots, \min(|\mathcal{C}_i|, B_i)] = -\mathbf{w}_1 \tag{42}$$

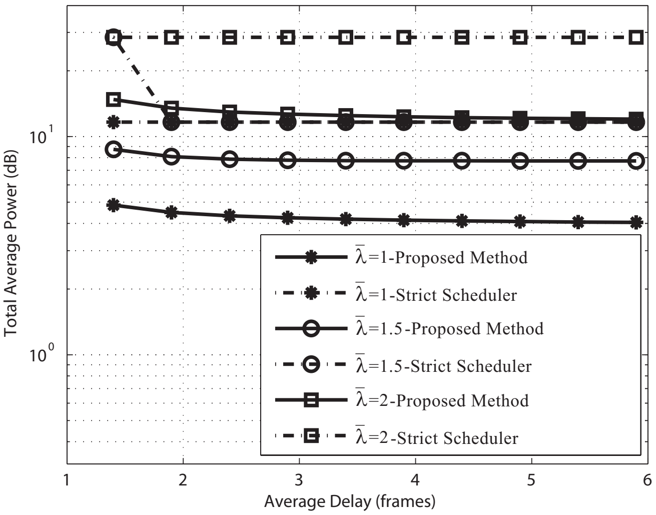


Fig. 2. Total average power vs. arrival rate - one queue.

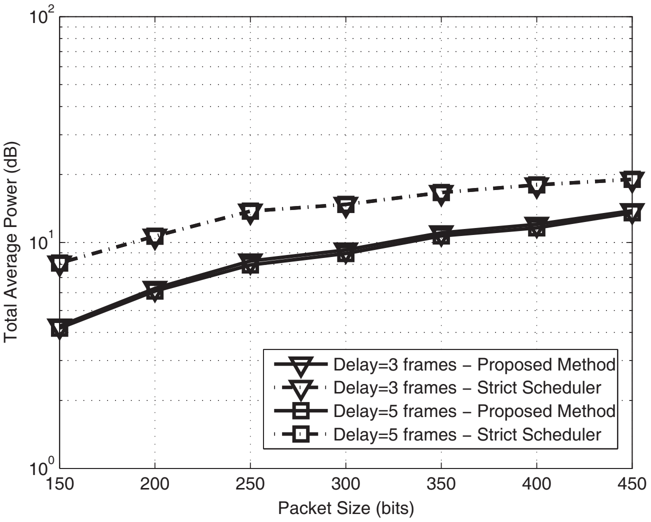


Fig. 3. Total average power vs. packet size - one queue.

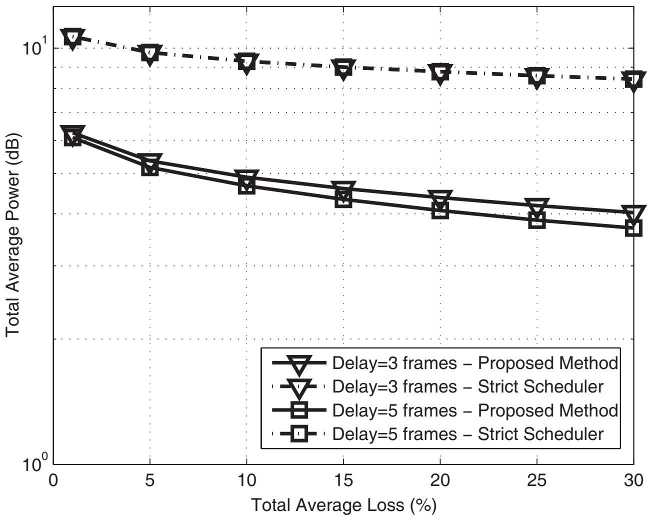


Fig. 4. Total average power vs. loss rate - one queue.

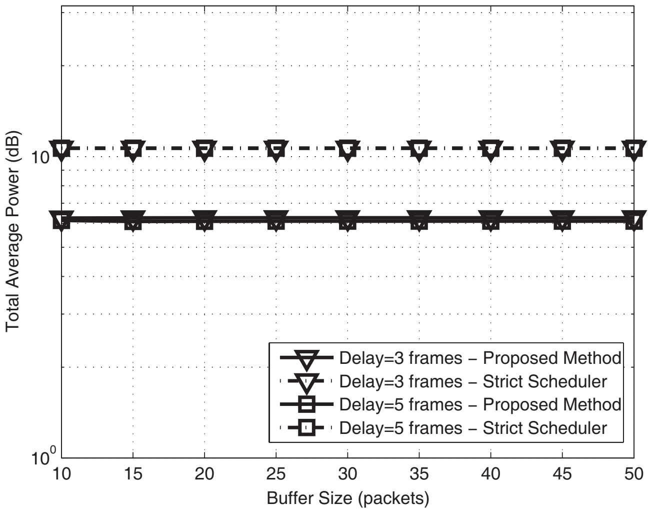


Fig. 5. Total average power vs. buffer size - one queue.

provided. Further, we also provide some metrics for the system complexity. Universal and single queue simulation parameters are shown in Table II while parameters for the two queue system are given in Table III.

The performance of the proposed scheme is compared with a strict scheduler. The strict scheduler implementation employs the same physical layer as the proposed scheme, however the transmission rate (packets per frame) is constant. This transmission rate is chosen from the set of allowable transmission rates such that it is the minimum rate that ensures QoS constraints (dropping probability and delay) are guaranteed and is found through a M/D/1/N model [14].

A. *Single Queue Results*

Figures 2-5 show the results for total average power for the single queue scenario for varying of major parameters. In all cases the proposed scheduler requires lower average power for transmission. For the case of increasing arrival rate or packet size we note an increase in the power selection which is expected while in the case of delay, increasing delay constraints result in a reduction of power. Once the delay constraint exceeds a certain threshold, power is no longer reduced as the scheduler is no longer dominated by the delay constraint. In the case of increasing loss rate we see a steady decline in power. This is due to the large amount of power required to maintain the stringent loss requirements (by maintaining a low channel error rate and lower number of packets dropped in the buffer). Finally, power selection is *u*-shaped when the buffer size is varied. This variation is negligibly small relative to the difference between the proposed method and the strict scheduler (shown in Fig. 5). The explanation for this phenomenon is as follows. In the smaller buffer region, power selection is dominated by trying to maintain acceptable loss in the buffer while in the larger buffer region, the power selection hits a plateau due to the limitation in the maximum tolerable delay is no longer a dominant factor.

B. *Two Queue Results*

In Figs. 6-9 the results for the two queue scenario is shown. As expected, the trends for the two queue scenario follow

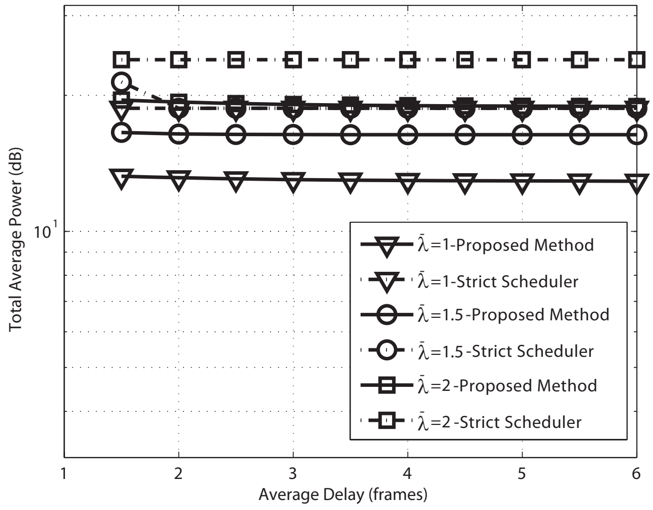


Fig. 6. Total average power vs. arrival rate - two queues.

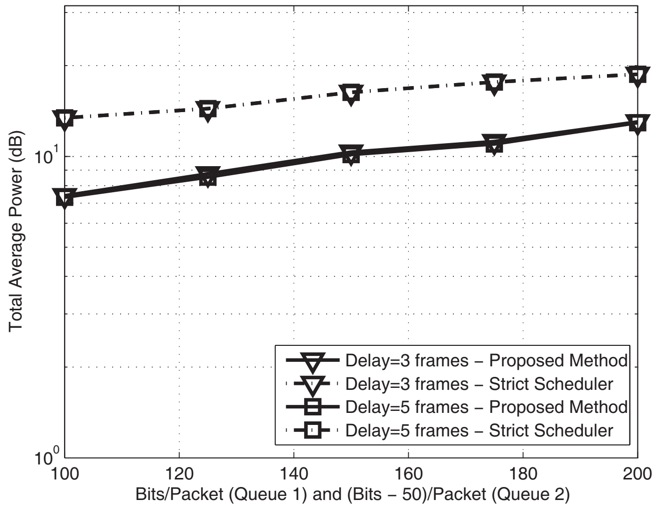


Fig. 7. Total average power vs. packet size - two queues.

those for the case of a single queue with higher transmission power. This is due to the increase in the total required physical layer transmission rate. Further, for the case of two queues, we utilize both the iterative and QP methods to validate its convergence. Our proposed iterative method converges to the same solution as the QP approach. Further, the storage requirements of the LUT in this case is only $36,864$ bits (or 4.5 kilobytes).

C. Iterative Convergence

For 3 or more queues, it is necessary to utilize the iterative LP method where each queue is solved as a LP problem and updates the cost function in (23). This method is guaranteed to find a local minimum due to the monotonicity of (23) in c_i irrespective of the values of $\pi_{i'}(c_{i'}|\mathbf{\Omega})$ for $\forall i', i' \neq i$ and due to the convexity of the LP problem within each queue.

We further demonstrate the global convergence using Monte Carlo simulations for both 2 and 3 queue scenarios over 10000 random initial solutions. For the two queue scenario, results are compared with the QP result, and with the 3 queue scenario results are measured as relative error (deviation from minimal solution). The average relative error versus iteration number in this case is shown in Fig. 10. As shown, the iterative solution has a maximal deviation of less than 10^{-7} over 10000 random realizations after just 3 iterations strongly suggesting that the iterative method converges to the global minimum.

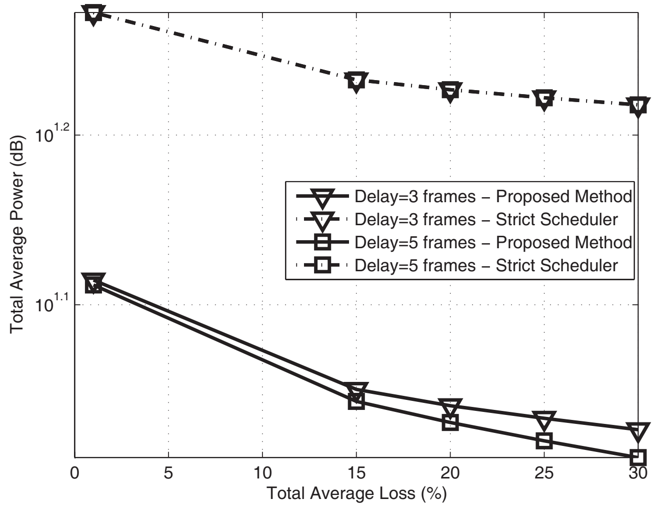


Fig. 8. Total average power vs. target loss - two queues.

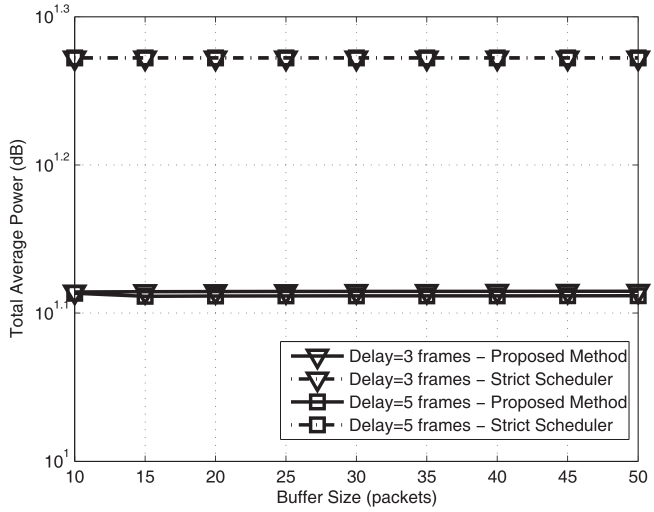


Fig. 9. Total average power vs. buffer size - two queues.

D. Loss Tradeoff

As previously discussed, we briefly overview the tradeoff between buffer and channel losses on the energy efficiency of the proposed design. Figure 11 shows the total average power versus the percentage of the total loss rate for both the buffer and channel losses. This is given for a several configurations of total loss rates, average delay constraints and buffer sizes. Overall, the trends demonstrate that is more efficient to incur a larger percentage of losses in the channel, particularly when the buffer size is large. However for smaller buffer sizes (*i.e.* delay is the dominant factor), it is beneficial to target a tradeoff between types of losses.

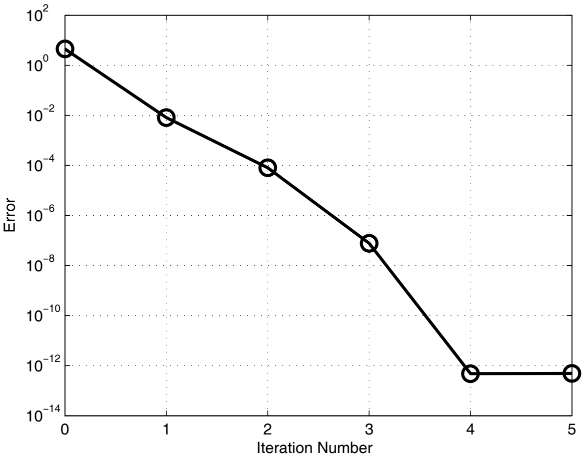


Fig. 10. Relative error vs. iteration number.

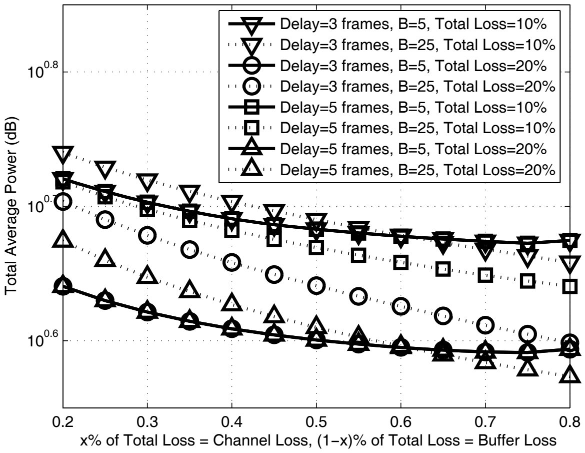


Fig. 11. Effect of varying loss rate component.

E. Complexity

System complexity can looked at from both the first and second stage optimizations. Due to the non-generalities of nonlinear programming solutions, in general it is not possible to model the complexity of the proposed NLP problem directly. We herein address this in two parts:

1) Demonstrate that the complexity of the sub optimization problem is less complex than the full-scale optimization problem, and
2) Demonstrate that the number of sub problems is less than the number of sub problems of the full-scale optimization technique.

For point 1, the full-scale optimization problem requires exploiting the joint queue state-space ($\mathcal{U}$) combined with the constraints and MAC rate state-space ($\mathcal{C}$). As the first stage of our optimization formulation does not rely on the $u \in \mathcal{U}$ states, the resulting computational complexity is not affected by the joint queue state-space size. As such, the complexity of the first-stage of our optimization is less complex than considering the full queue state-space.

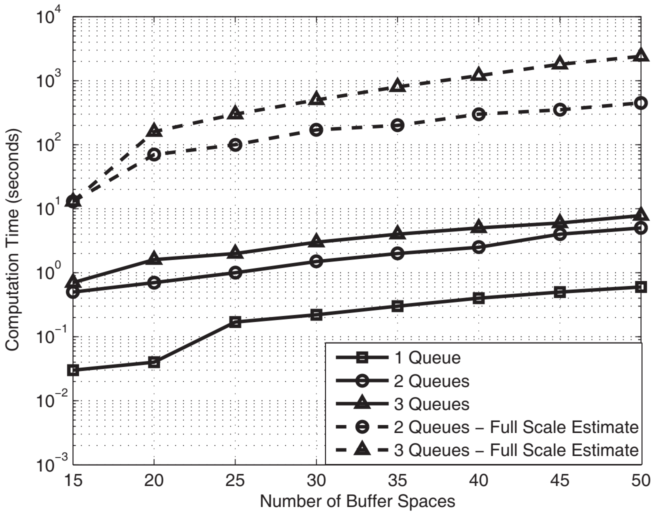


Fig. 12. Computation time.

The system complexity from the second stage (point 2), incurs a large complexity reduction. As we note from the previous section, the convergence of the series of LP problems occurs within several iterations. The dimension of each LP problem is $\Phi_i = |\mathcal{C}_i|(B_i + 1)$. Assuming computation time is polynomial in Φ_i, the computation time of the above is

$$T_{iter} \propto \kappa \sum_{i=1}^{K} \Phi_i \tag{46}$$

where κ denotes the number of iterations and $\propto$ means proportional to. Further, the computation time of the full-scale problem is

$$T_{full} \propto \prod_{i=1}^{K} \Phi_i \tag{47}$$

Comparing (46) and (47) its clear to see that for small κ, the computation time of (47) is sufficiently larger (even for $K = 2$).

In Fig. 12 we further study the complexity by showing the computation time measured in seconds of the MAC rate assignment component for a one, two and three queue configuration averaged over 50 realizations using the iterative method. In the graph, we also provide an estimate of the computational time of both two and three queue systems assuming a full-scale optimization scheme (i.e., $\mathcal{C} \times \mathcal{U}$ as defined before) and under the assumption that full-scale optimization scales in polynomial time. We note that there is a large reduction in terms of computation time using the proposed iterative method when examining the MAC rate assignment.

VI. Extension to Time Varying Channels

The power, rate and channel allocation scheme described in the first stage of Section III can be further extended to the case where MIMO eigenvalues evolve in time. In order to design such a system, knowledge about the evolution of the channel eigenvalues must be known. Here[3] we consider a

[3] The described framework can be applied to any channel model with known eigenvalue distribution.

sparse MIMO channel model well-described in [15] with time-varying quantities of interest derived in [16]. The unordered eigenvalue distribution in this case is simply

$$f_{\Lambda_j}(\lambda_j) = \frac{1}{M_R M_T \eta_j} \exp\left(\frac{-\lambda_j}{M_R M_T \eta_j}\right), j = 1, \ldots, L \tag{48}$$

where L is the number of independent scatters and η_j is the relative power of the j^{th} scatter such that $\sum_{j=1}^{L} \eta_j = 1$. The number of channels (using MIMO SVD eigenbeamforming) in this case is equal to L (*i.e.,* $M = L$).

Each eigenvalue can be partitioned into a finite number of states. Partitioning methods are well discussed in the literature [2], [17], however for the sake of brevity we describe the equal partitioning method where bounds for each eigenvalue are simply given as

$$\varphi_{j,k} = -M_R M_T \eta_j \ln\left(1 - \frac{k-1}{|\mathcal{J}_j|}\right), k = 2, \ldots, |\mathcal{J}_j| \tag{49}$$

with $\varphi_{j,1} = 0$, $\varphi_{j,|\mathcal{J}_j|+1} = \infty$, and $|\mathcal{J}_j|$ equals the number of partitions in channel j. The state-space for eigenvalue j is $\mathcal{J}_j$, and the overall channel space is the L-dimensional space given as $\mathcal{J} := \mathcal{J}_1 \times \cdots \times \mathcal{J}_L$.

Modifications to above optimization formulation in Section IV-A proceeds as follows. First, we define a new space $\mathcal{S}$ such that $\mathcal{S} := \mathcal{C} \times \mathcal{J}$. The optimization procedure is performed for each $s \in \mathcal{S}$ and for each valid AMC mode combination in $\mathcal{K}(c)$.

Next, to compensate for the increased optimization space $\mathcal{J}$, we can reduce the solving complexity to an LP problem from an MINLP problem (which has tractable complexity) by making the following two adjustments

1) The target BER for a particular class of traffic is the same in all subchannels, and
2) Allowing power to be configurable per traffic stream in each channel in a frame.

The LP problem is of the form $\arg\min_{\mathbf{x}} \mathbf{c}^T\mathbf{x}$, subject to $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, $\mathbf{A}_{eq}\mathbf{x} = \mathbf{b}_{eq}$, $\mathbf{x} \geq 0$ where

$$\mathbf{x} = [X_{1,1}, \ \ldots, \ X_{1,K}, \ \ldots, \ X_{M,K}]^T \tag{50}$$

and we have $\mathbf{A}_{eq}$ and $\mathbf{A}$ given as $K \times MK$ and $M \times MK$ matrices respectively with entries as described by (27) and (29). Both $\mathbf{b}_{eq}$ and $\mathbf{b}$ are given as in (28) and (30).

Finally, the cost function $\mathbf{c}$ is given as the average power required to maintain a particular BER in a given channel for a particular stream as

$$\mathbf{c} = \left[\frac{P_{1,1}}{k_1}, \ \ldots, \ \frac{P_{1,K}}{k_1}, \ \ldots, \ \frac{P_{M,K}}{k_M}\right]^T \tag{51}$$

where $P_{j,i}$ is the minimum average power required to maintain a given average bit error rate and is given as the solution to

$$BER_i = \frac{\int_{\varphi_{j,k}}^{\varphi_{j,k+1}} Pb(P_{j,i} \cdot r, 2^{k_j}) f_{\Lambda_j}(r) dr}{\int_{\varphi_{j,k}}^{\varphi_{j,k+1}} f_{\Lambda_j}(r) dr} \tag{52}$$

when channel j is in state $k \in \mathcal{J}_j$ and where $Pb(\cdots)$ is given in (1). Due to the monotonicity of $Pb(\gamma, k_j)$ in γ, the above can be solved efficiently using simple bisection techniques for $P_{j,i}$ for each $s \in \mathcal{S}$ and possible AMC mode.

VII. Conclusion

In this work, a general $\{K \times M\}$ model is presented with a corresponding scheduler to minimize average transmission power by utilizing both channel and queue knowledge. Through a novel MAC rate assignment scheme, it is possible to decouple the problem and utilize queue state information in all queues to derive a QoS-aware scheduling policy over both static and time-varying MIMO SVD channels. The proposed dynamic scheduler is shown to outperform static scheduling and is able to meet hard QoS constraints. Future work will focus on further reducing the system complexity beyond that discussed here, simplifying the complexities of the non-linear optimization problem by employing packet-based rate assignment. In this way, the non-linear constraints can be lifted and our ideology can be applied to more general multi-channel systems.

References

[1] X. Bai and A. Shami, "Two dimensional cross-layer optimization for packet transmission oer fading channel," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3813–3822, Oct. 2008.

[2] Q. Liu, S. Zhou, and G. Giannakis, "Queuing with adaptive modulation and coding over wireless links: cross-lyer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142–1153, May 2005.

[3] X. Zhu, J. Huo, X. Xu, C. Xu, and W. Ding, "QoS-guaranteed scheduling and resource allocation algorithm for IEEE 802.16 OFDMA system," in *Proc. IEEE ICC'08*, May 2008, pp. 3463–3468.

[4] X. Bai, A. Shami, and S. Primak, "Optimal power control over fading channel with cross-layer performance constraint," in *Proc. IEEE ICC'08*, May 2008, pp. 2265–2269.

[5] C. Wang, P.-C. Lin, and T. Lin, "A cross-layer adaptation scheme for improving IEEE 802.11e QoS by learning," *IEEE Trans. Neural Networks*, vol. 17, no. 6, pp. 1661–1665, Nov. 2006.

[6] F. Meshkati, H. Poor, S. Schwartz, and R. Balan, "Energy-efficient resource allocation in wireless networks with quality-of-service constraints," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3406–3414, Nov. 2009.

[7] R. Louie, M. Mckay, and I. Collings, "Maximum sum-rate of MIMO multiuser scheduling with linear receivers," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3500–3510, Nov. 2009.

[8] V. Lau and Y. Chen, "Delay-optimal power and precoder adaptation for multi-stream MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 3104–3111, 2009.

[9] V. Lau and Y. Cui, "Delay-optimal power and subcarrier allocation for OFDMA systems via stochastic approximation," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 227–233, Jan. 2010.

[10] IEEE Standard 802.11e-2005: Wireless LAN MAC and PHY Specifications, Amendment 8: MAC QoS Enhancements, 2005.

[11] R. Bultitude, G. Brussaard, M. Herben, and T. Willink, "Radio channel modelling for terrestrial vehicular mobile applications," in *Proc. Millenium Conf. on Antennas and Propogation*, 2000, pp. 1–5.

[12] A. Burr, "Capacity bounds and estimates for the finite scatterers MIMO wireless channel," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 812–818, May 2003.

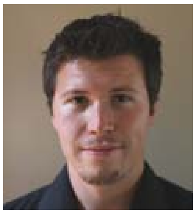
[13] S. Chung and A. Goldsmith, "Degrees of freedom in adaptive modulation: a unified view," *IEEE Trans. Commun.*, vol. 49, no. 9, pp. 1561–1571, 2001.

[14] O. Brun and J. Garcia, "Analytical solution of finite capacity M/D/1 queues," *J. Applied Probability*, vol. 37, no. 4, pp. 1092–1098, 2000.

[15] S. Primak and E. Sejdic, "Application of multitaper analysis to wireless communications problems," in *Proc. ISABEL'08*, Oct. 2008.

[16] D. J. Dechene, S. L. Primak, and A. Shami, "On the first and second order statistics of sparse MIMO channels," in *Proc. 25th Queens Biennial Symposium on Communications*.

[17] Q. Zhang and S. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, 1999.



Dan J. Dechene (M'03) received the B.Eng. Degree in Electrical Engineering from Lakehead University, Thunder Bay, ON, Canada in 2004 and his M.E.Sc. Degree in Electrical and Computer Engineering from The University of Western Ontario, London, ON, Canada in 2008. Currently, he is currently pursuing his Ph.D. Degree in Communication Systems Engineering at The University of Western Ontario. Dans research interests include energy efficient resource allocation schemes, heterogeneous quality of service (QoS) guarantees and multiple antenna (MIMO) wireless systems. He currently holds an NSERC PGS-D3 scholarship.

Abdallah Shami (M'03, SM'09) received the B.E. Degree in Electrical and Computer Engineering from the Lebanese University, Beirut, Lebanon in 1997, and the Ph.D. Degree in Electrical Engineering from the Graduate School and University Center, City University of New York, New York, NY in September 2002. In September 2002, he joined the Department of Electrical Engineering at Lakehead University, Thunder Bay, ON, Canada as an Assistant Professor. Since July 2004, he has been with The University of Western Ontario, London, ON, Canada where he is currently an Associate Professor in the Department of Electrical and Computer Engineering. His current research interests are in the area of wireless/optical networking.