Delay Analysis for Wireless Networks with Single Hop Traffic and General Interference Constraints

Gagan Raj Gupta, Ness B. Shroff, *Fellow, IEEE*

***Abstract*—We consider a class of wireless networks with general interference constraints on the set of links that can be served simultaneously at any given time. We restrict the traffic to be single-hop, but allow for simultaneous transmissions as long as they satisfy the underlying interference constraints. We begin by proving a lower bound on the delay performance of any scheduling scheme for this system.**

We then analyze a large class of throughput optimal policies which have been studied extensively in the literature. The delay analysis of these systems has been limited to asymptotic behavior in the heavy traffic regime and order results. We obtain a tighter upper bound on the delay performance for these systems. We use the insights gained by the upper and lower bound analysis to develop an estimate for the expected delay of these networks operating under the well-known Maximum Weighted Matching (MWM) scheduling policy. We show via simulations that the MWM policy is often close to the lower bound, which means that it is not only throughput optimal, but also provides excellent delay performance.

***Index Terms*—Wireless Networks, Scheduling, Delay Analysis, Interference, Lyapunov.**

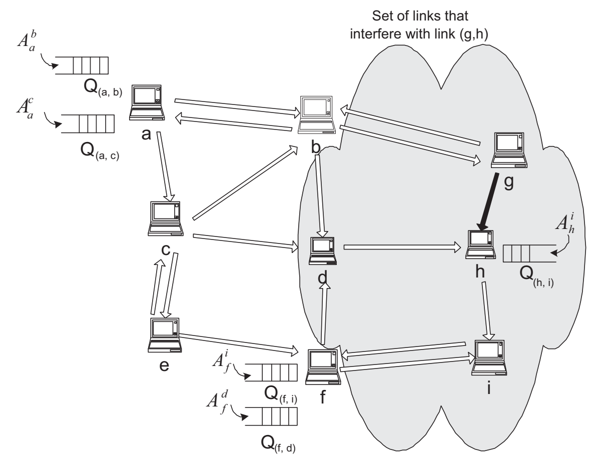

Fig. 1. Figure showing a wireless network with single-hop traffic. All packets A_s^d, transmitted on link (s,d) are exogenous and are queued ($Q_{s,d}$ denotes the queue length). All the links that interfere with link (g,h) are shown.

I. Introduction

In a wireless system, users compete for accessing a shared transmission medium. Since link transmissions cause mutual interference, the medium access layer (MAC) is needed to schedule the links carefully so that packets can be transmitted with minimal collisions. Many scheduling policies have been studied at the MAC layer with the objective of maximizing throughput. These schemes are often called throughput-optimal scheduling schemes. However, the delay analysis of these systems has largely been limited. Our focus in this paper is to analyze the expected delay for this system. To that end, we will derive upper and lower bounds on the expected delay, and also provide an accurate estimate of the expected delay for a well-known and extensively-studied (e.g., [1]–[4]) throughput-optimal scheme called the Maximum Weighted Matching (MWM).

To simplify the analysis we, in common with related work [3], [5], [6], restrict the traffic model to single-hop traffic. Under the single-hop traffic model, all packets transmitted on a link (s,d) are generated by an exogenous arrival process A_s^d at the source node s. As shown in Figure 1, the exogenous arrivals waiting to be transmitted at each link are queued in their respective queues. This approach has also been adopted in the literature while studying the throughput performance of scheduling policies for wireless networks. This allows us to study the effect of scheduling policy on the delay of the system, independent of routing. *We note that this model allows for simultaneous transmissions as long as they satisfy the underlying interference constraints. Such systems are more general than the cellular type systems where the system is divided into non-interfering cells. The results presented here work for any underlying model for interference constraints*.

The design of scheduling policies which stabilize the system even under single-hop traffic is a challenging task. Intuitively, the scheduler must schedule as many links as possible in every time slot. Such schedulers are called maximal schedulers (as opposed to maximum weighted schedulers that also take the queue length into account). However, even with maximal scheduling, some of the queue lengths may become unbounded. The reason is that if the scheduler does not use the queue length information, some of the queues may grow large, while others remain very small or become empty. This, in turn, does not allow the scheduler to schedule a large number of queues and leads to instability. Thus a throughput optimal policy like MWM, carefully uses the information of the queue lengths while scheduling the links.

The above behavior caused by throughput-efficient schedulers significantly complicates the delay analysis of these systems, because the service process of each link is governed not only by the interference constraints, but also by its queue length.

For example, in a wireless network operating under a throughput optimal policy, such as the MWM policy, the expected delay at a link may be large even if the arrival rate is small. This is because these policies try to schedule the longer queues in the system or in other words, they prevent

Gagan Raj Gupta is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA e-mail: grgupta@purdue.edu

Ness B. Shroff is with the Departments of ECE and CSE, The Ohio State University, Columbus, Ohio, USA e-mail: shroff@ecn.osu.edu

the queues from becoming very large. This can be thought of as a mechanism to balance the queue lengths in the system.

We now state our main contributions in this paper:

- Development of a fundamental lower bound on the expected queuing delay of a wireless network regardless of the scheduling policy used.
- Development of an upper bound on the expected delay of a throughput optimal scheduling policy, GMWM (a generalization of MWM), under a single-hop traffic model.
- Development of an estimate for the expected delay in a wireless network under a throughput-optimal policy, given the load and the interference constraints. Further, the estimate is shown to lie between the upper and lower bounds developed above. We show through simulations that for single-hop traffic and any given load within the capacity region, the estimate is accurate.

II. Related Work

Most of the analysis of scheduling policies for the wireless systems has been limited to stability results. A stable scheduling policy is guaranteed to keep the average queue lengths in the system finite, but the tightness of the upper bound on the average queue length is not known. One of the techniques used for deriving upper bounds on the average queue length for these systems is the method of Lyapunov drifts developed in [2], [5], [7], [8]. However, these results are order results and provide only a limited understanding of the delay of the system. For example, it has been shown in [5] that the maximal matching policies achieve $\mathcal{O}(1)$ delay for networks with single-hop independent Poisson traffic when the input load is in the reduced capacity region. However, for arbitrary networks, this region may be only a small fraction of the capacity region, C (see [9]). Informally, the (maximum) capacity region C is the set of mean flow rate vectors $(\lambda_1, ..., \lambda_N)$ such that there exists a scheduling rule making the queue length process stable.

Simulations have shown that two schemes that guarantee stability for the full capacity region can have very different delay characteristics. The results presented in [3] suggest that a policy that provides stability guarantees in the full capacity region may have worse delay characteristics than another policy which provides weaker guarantees. The comparison of an implementation of a throughput optimal algorithm (*Pick and Compare*) with sub-optimal algorithms like maximal matching is studied in [9]. It is shown that under *Pick and Compare* type scheduling algorithms, queues in the system grow very large and are hence such idealized algorithms are not realizable in practice.

Since throughput by itself does not seem to be a good metric to differentiate between scheduling algorithms, the development of analytical techniques to compare other metrics of performance such as delay is crucial. In [10], the authors observe that there is no theoretical result comparing the delay performance of a RANDOM scheduler to the MWM algorithm. The upper bound developed in this paper allows us to show that the expected delay performance of GMWM is no worse than the performance of any stationary randomized policy.

In [11]–[13], cellular systems are analyzed and large deviations results are obtained to calculate queue-overflow probability. The analysis is much harder for the wireless network considered here, due to the complex interactions of the arrival, service, and backlog process. Order-optimal results for the expected delay a wireless up-link down-link system are presented in [8]. The bounds presented here are sharper than the those obtained by [8] and are also order-optimal in the context of the system studied in that paper.

One of the results that has been shown about the MWM scheduling policy is that it is asymptotically optimal in the heavy traffic regime [14], [15] under the assumption of resource pooling. However, this result does not provide any estimate of the delay. It is also not known whether these policies continue to be optimal for an arbitrary load in the capacity region.

The lower bound presented in this paper uses the concept of *exclusive sets* (defined in Section III) to characterize constraints on the scheduling policy. We analyze a fictitious scheduling policy based on exclusive sets that is amenable to analysis and show that its expected delay is a lower bound on the performance of any other scheduling policy. The exclusive sets were also studied in [16] for the purpose of analyzing the impact of interference on the throughput capacity of a multi-hop wireless network. The authors proved that the polytope generated by these sets is an upper bound on the capacity region C and may be loose. We find that these exclusive set constraints are nonetheless very useful for delay analysis, since they also constitute some of the faces of the capacity polytope C. We observe in our simulations that for several representative topologies, the performance of MWM scheduling policy is close to the lower bound. The upper bound on the other hand captures all the interference constraints in the system and whenever the upper bound goes to infinity, the average delay of the system under the GMWM policy also becomes infinite.

Delay optimal schemes have been proposed in the literature [17] for wireless networks, which typically minimize an expected delay metric (assuming that the system behaves as M/M/1). We note that there is no reason to assume that M/M/1 approximation will be accurate because the service process could be very complex in this system, given that the interference constraints have to be met at every time-slot. Neither are we aware of any result which shows that a policy that minimizes the M/M/1 delay metric also minimizes the delay for the system. In fact, we expect that such an argument will likely not be true given the complexity involved in scheduling link transmissions in a wireless system. We provide a more accurate estimate of the expected delay for wireless networks, which could be used as a delay metric that would be useful in the development of such delay optimal schemes.

We begin with a brief description of the system model and notations. We then derive the lower bound and the upper bound on the expected delay in the system. We then propose a method to estimate the expected delay of the system. We study the accuracy of the results for several important classes of wireless networks through simulations.

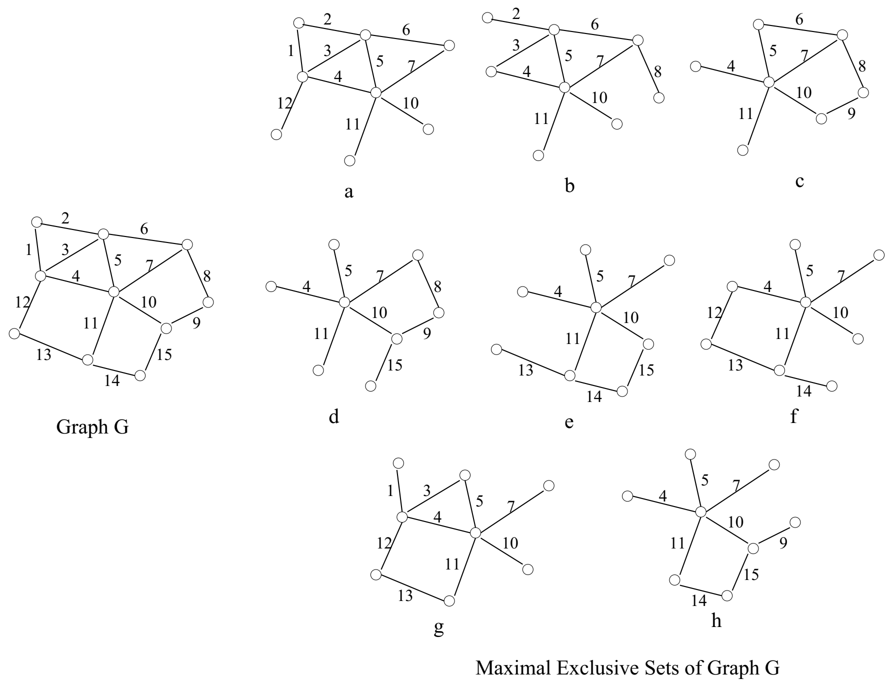

Fig. 2. Maximal exclusive Sets under 2-hop interference model

III. System Model

We consider a wireless network, G with N links denoted by set L. Each link l has its own exogenous arrival stream $\{A_l(t)\}_{t=1}^{\infty}$. Each arrival stream is *i.i.d.* in time. The distribution of the number of packets, $A_l(t)$, arriving to a link l in any given time slot t may be arbitrary but time invariant. Assume that the second moments, $\mathbf{E}[A_l^2]$, of the arrival processes are finite. Different input streams may be correlated with each other. Let $\mathbf{A}(t) = (A_1(t), \ldots, A_N(t))$ represent the vector of exogenous arrivals, where $A_l(t)$ is the number of packets that arrive to link l during time slot t (for $l \in 1, \ldots, N$). Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ represent the corresponding arrival rate vector.

The packets arriving at each link are queued. Let $Q_l(t)$ denote the queue length at link l. The queue length vector is denoted by $\mathbf{Q}(t) = (Q_l(t) : l = 1, 2, \ldots, N)$. A link can be activated in a time slot t only if the queue is non empty. We use the term activation (scheduling) of a link or a queue interchangeably in the paper. At most, one packet is served at a queue in a given time slot. After service, each packet leaves the system. There is a slotted structure in service. For each link l, the indicator function $I_l(t)$ indicates whether or not link l received service at time slot t. Note that

$$I_l(t) = \begin{cases} 1 & \text{if } Q_l(t) > 0 \text{ and } l \text{ is scheduled} \\ 0 & \text{otherwise} \end{cases} \tag{III.1}$$

The evolution of the queue is as follows,

$$Q_l(t+1) = Q_l(t) - I_l(t) + A_l(t), l = 1, .., N \tag{III.2}$$

The vector of the scheduled queues is denoted by $\mathbf{I}(t) = (I_n(t)) : n = 1, ..N$. Because of interference, there are constraints on the combination of links that can be activated simultaneously. We allow these constraints to be arbitrary. $\mathbf{I}(t)$ is a valid activation vector if it satisfies these constraints. Let S be the collection of all valid activation vectors. Let $\mathbf{I}^j$ be the j^{th} activation vector in S. At each time-slot an activation vector $\mathbf{I}(t)$ is scheduled. A scheduling policy decides which activation vector is used in every time slot.

For any given link l, we define an *exclusive set*, χ_l, as a set of links including l in which no more than one link can be scheduled at any given slot. In particular, we are interested in the *maximal* exclusive sets, i.e., sets in which no more links can be added without violating the above property. A link may be present in multiple exclusive sets.

In this paper, we will use exclusive sets to derive the fundamental lower bounds on the delay of the system. We will be interested in those exclusive sets χ_l, where the sum of arrival rates is large. We use λ_{χ_l} to denote the sum of arrival rates to the queues in the set χ_l.

$$\lambda_{\chi_l} = \sum_{i \in \chi_l} \lambda_i \tag{III.3}$$

Similarly, A_{χ_l} and Q_{χ_l} are used to denote the the sum of arrivals and the sum of queues in the set χ_l respectively,

$$A_{\chi_l}(t) = \sum_{i \in \chi_l} A_i(t) \tag{III.4}$$

$$Q_{\chi_l}(t) = \sum_{i \in \chi_l} Q_i(t) \tag{III.5}$$

Figure 2, shows all the maximal exclusive sets of a graph G under an example interference model called the 2-hop interference model. In a 2-hop interference model, any two active links in $\mathbf{I}$(t) are always separated by two or more hops in the underlying network graph. Let us consider subgraph *a* in Figure 2. Every link in the subgraph interferes with any other link because it is within two hop distance. Moreover, no more link from graph G can be added to this subgraph without violating the above property.

The 2-hop interference model is used again in our simulation studies since it has been often used to model the behavior of a large class of MAC protocols based on virtual carrier sensing using RTS/CTS messages, which includes the IEEE 802.11 protocol [18], [19].

Let $\|\mathbf{Y}\|$ denote the Euclidean norm of vector $\mathbf{Y}$. The system is considered to be stable [2] if $\lim_{t\to+\infty} \sup \mathbf{E}[\|\mathbf{Q}(t)\|] < \infty$. If the system is stable then the throughput is the same as the arrival rates. A throughput vector $\boldsymbol{\lambda}$ is admissible if there is some scheduling policy under which the system is stable when the arrival rate vector is $\boldsymbol{\lambda}$. Let us denote by Λ the closure of the convex hull of the set of activation vectors, $\mathbf{I}^j$ and by C the interior of the convex hull. Note that Λ is a closed convex set. It has been shown in [1] that if each arrival process is *i.i.d.* in time, and the first two moments of all the arrival streams $\{A_l(t)\}_{t=1}^{\infty}$ are finite, then $\boldsymbol{\lambda} \in C$ is a necessary condition for a stabilizing scheduling policy to exist. It is also shown that the MWM policy, that chooses the maximum weighted activation vector (matching), stabilizes the system for any arrival rate satisfying the preceding condition.

MWM Scheduling Policy

$$\mathbf{I}(t) = \operatorname*{argmax}_{\mathbf{I}^j \in S} \sum_{i=1}^{N} Q_i(t) I_i^j \tag{III.6}$$

where I_i^j is the i^{th} component of the j^{th} activation vector, $\mathbf{I}^j$, in set S.

Fig. 3. MWM Scheduling Policy

The definition of the capacity region of these systems is related to the existence of a scheduler that chooses to activate the queues by a stationary process. These results have been derived in [7].

Lemma 3.1: For any feasible input rate vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_N)$ which lies in the interior of the capacity region, C there exists a vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_N) \in C$ such that $\lambda_l < \mu_l$ for all queues $l \in L$. Also, there exists a stationary randomized scheduling policy which chooses activation vectors $\mathbf{I}^{\mathbf{R}}(t)$ such that $\mathbf{E}[I_l^R(t)] = \mu_l$ and hence stabilizes the system.

The exclusive sets define the constraints on the rate vector $\boldsymbol{\mu}$. We let μ_{χ_l} denote the sum of service rates of the queues in χ_l of a stationary randomized policy. *A given vector $\boldsymbol{\mu}$ is in the capacity region if μ_{χ_l} is less than one for all exclusive sets in the system.*

IV. Fundamental Lower Bounds on the System

In this section, we develop an algorithm to calculate a lower bound on the delay of the system, independent of the scheduling policy used. Recall the definition of the exclusive sets, χ_l of link l in the system. Only one of the queues in χ_l can be scheduled at any given time slot. The notion of exclusive sets is helpful for deriving fundamental lower bounds on the expected delay of the system.

Let us consider a fictitious scheduling policy Π_{lower} that guarantees to schedule one of the links in χ_l whenever there is at least one non-empty queue. Although Π_{lower} policy satisfies the interference constraints within χ_l, it ignores the interference of the scheduled link with other links in the network. We denote the sum of queue lengths in χ_l under the policy Π_{lower} as $\mathcal{Q}_{\chi_l}$.

$$\mathcal{Q}_{\chi_l}(t) = \sum_{i \in \chi_l} Q_i(t) \tag{IV.7}$$

Then, the queue evolution under Π_{lower} is given by the following Equation.

$$\mathcal{Q}_{\chi_l}(t+1) = (\mathcal{Q}_{\chi_l}(t) - \mathbf{1}_{\{\mathcal{Q}_{\chi_l}(t)>0\}} + A_{\chi_l}(t))^+ \tag{IV.8}$$

where $\mathbf{1}$ is the indicator function and A_{χ_l} is as defined in Equation (III.4).

We now compare the evolution of queues in χ_l under the Π_{lower} policy to an arbitrary scheduling policy. We assume that both the systems are driven by the same sequence of arrivals. In Lemma 4.1 we compare the sum of queue lengths Q_{χ_l} in χ_l with $\mathcal{Q}_{\chi_l}$ at a given time T. The periods of time in which at least one of the queues in χ_l is non-empty under the Π_{lower} policy are called *busy periods*.

Lemma 4.1: For any exclusive set χ_l in the system, the sum of queue lengths Q_{χ_l} in χ_l, under any scheduling policy is no smaller than those under Π_{lower} policy at all times, T, i.e. $Q_{\chi_l}(T) \geq \mathcal{Q}_{\chi_l}(T)$.

Proof: Depending on whether T lies in the busy period of the system under the Π_{lower} policy or not, the following two cases arise.

Case 1: $\mathcal{Q}_{\chi_l}(T) = 0$

Since $Q_{\chi_l}(T)$ is always non-negative, the result holds trivially.

Case 2: $\mathcal{Q}_{\chi_l}(T) > 0$.

Let T_o be the time that initiated the current busy period, i.e. $T_o < T$. Then the queue length can obtained by summing Equation (IV.8), is as follows:

$$\mathcal{Q}_{\chi_l}(T) = \sum_{t=T_o-1}^{T-1} A_{\chi_l}(t) - \sum_{t=T_o-1}^{T-1} \mathbf{1}_{\{\mathcal{Q}_{\chi_l}(t)>0\}} \tag{IV.9}$$

Since the system is in the middle of a busy period, $\mathbf{1}_{\{\mathcal{Q}_{\chi_l}(t)>0\}} = 1$ for all $T_o \leq t \leq T$, and the above equation reduces to

$$\mathcal{Q}_{\chi_l}(T) = \sum_{t=T_o-1}^{T-1} A_{\chi_l}(t) - (T - T_o) \tag{IV.10}$$

Now we consider the evolution of the queues in χ_l under an arbitrary scheduling policy. By the definition of χ_l , not more than one of the queues in χ_l can be scheduled at any given time-slot, i.e.,

$$\sum_{i \in \chi_l} I_i(t) = I_{\chi_l}(t) \leq 1 \tag{IV.11}$$

The evolution of the queues in χ_l is given by the following equation.

$$Q_{\chi_l}(t+1) = Q_{\chi_l}(t) - I_{\chi_l}(t) + A_{\chi_l}(t) \tag{IV.12}$$

In particular,

$$Q_{\chi_l}(T_o) = Q_{\chi_l}(T_o-1) - I_{\chi_l}(T_o-1) + A_{\chi_l}(T_o-1) \tag{IV.13}$$

This system (under the arbitrary scheduling policy) may or may not be in the middle of a busy period at $T_o - 1$. If it is in the middle of a busy period, $Q_{\chi_l}(T_o - 1) \geq 1$ and thus,

$$(Q_{\chi_l}(T_o - 1) - I_{\chi_l}(T_o - 1)) \geq 0. \tag{IV.14}$$

If the system is not in the middle of a busy period, then

$$I_{\chi_l}(T_o - 1) = 0 \tag{IV.15}$$

since an empty queue cannot be scheduled at any time slot (see Equation (III.1)).

Combining Equations (IV.14) and (IV.15), we obtain the following.

$$Q_{\chi_l}(T_o) \geq A_{\chi_l}(T_o - 1) \tag{IV.16}$$

By summing Equation (IV.12) to obtain $Q_{\chi_l}(T)$, and simplifying using Equations (IV.16) and (IV.11), we obtain the desired result.

$$\begin{aligned} Q_{\chi_l}(T) &= Q_{\chi_l}(T_o) + \sum_{t=T_o}^{T-1} A_{\chi_l}(t) - \sum_{t=T_o}^{T-1} I_{\chi_l}(t) \\ &\geq A_{\chi_l}(T_o - 1) + \sum_{t=T_o}^{T-1} A_{\chi_l}(t) - \sum_{t=T_o}^{T-1} I_{\chi_l}(t) \\ &\geq \sum_{t=T_o-1}^{T-1} A_{\chi_l}(t) - \sum_{t=T_o}^{T-1} 1 \\ &\geq \boldsymbol{\mathcal{Q}}_{\chi_l}(T) \end{aligned} \tag{IV.17}$$

■

Using the above lemma, we derive the following lower bound on the queues in χ_l.

Theorem 4.1: For any exclusive set χ_l in the system, the expected value of the sum of queue lengths in χ_l under any scheduling policy is lower bounded by the following.

$$\mathbf{E}[Q_{\chi_l}] \geq \sum_{i \in \chi_l} \frac{\lambda_i + \mathbf{E}[A_i(\sum_{j \in \chi_l} A_j)] - 2\lambda_i \lambda_{\chi_l}}{2(1 - \lambda_{\chi_l})} = LB_{\chi_l}$$

Proof: Lemma 4.1 shows that at all times, T, $Q_{\chi_l}(T) \geq \boldsymbol{\mathcal{Q}}_{\chi_l}(T)$. It follows then, that the expected value of the sum of queue lengths in χ_l under any other scheduling policy Π will be lower bounded by the expected value of sum of queue lengths in χ_l under Π_{lower}. Then

$$\mathbf{E}[Q_{\chi_l}] \geq \mathbf{E}[\boldsymbol{\mathcal{Q}}_{\chi_l}] \tag{IV.18}$$

The analysis of the exclusive set under the Π_{lower} policy reduces to that of single server queue being fed by multiple arrival streams, i.e. A_{χ_l}. Since the arrival streams are assumed to independent over time, the expected value of $\boldsymbol{\mathcal{Q}}_{\chi_l}$ under the Π_{lower} policy can be derived using the standard GI/D/1 analysis and is given by.

$$\mathbf{E}[\boldsymbol{\mathcal{Q}}_{\chi_l}] = \frac{\lambda_{\chi_l} + \mathbf{E}[(\sum_{i \in \chi_l} A_i)^2] - 2(\lambda_{\chi_l})^2}{2(1 - \lambda_{\chi_l})} \tag{IV.19}$$

It follows that:

$$\begin{aligned} &\mathbf{E}[Q_{\chi_l}] \geq \frac{\sum_{i \in \chi_l} \lambda_i + \mathbf{E}[(\sum_{i \in \chi_l} A_i)^2] - 2(\sum_{i \in \chi_l} \lambda_i)^2}{2(1 - \lambda_{\chi_l})} \\ \implies &\mathbf{E}[Q_{\chi_l}] \geq \sum_{i \in \chi_l} \frac{\lambda_i + \mathbf{E}[A_i(\sum_{j \in \chi_l} A_j)] - 2\lambda_i(\sum_{j \in \chi_l} \lambda_j)}{2(1 - \lambda_{\chi_l})} \\ \implies &\mathbf{E}[Q_{\chi_l}] \geq \sum_{i \in \chi_l} \frac{\lambda_i + \mathbf{E}[A_i(\sum_{j \in \chi_l} A_j)] - 2\lambda_i \lambda_{\chi_l}}{2(1 - \lambda_{\chi_l})} \end{aligned}$$

■

We use LB_{χ_l} to denote the lower bound derived above on the set χ_l. We now develop a greedy algorithm (see Algorithm 1) to compute a lower bound on the sum of expected queue lengths on the entire system. At every iteration of the "repeat-until" loop, an exclusive set with the highest value of LB_{χ_l} is computed among the links in set X. Note that this set is a maximal exclusive set in X and may not be maximal in the original set of links L. For any link l, we use $\widetilde{\chi_l}$ to denote the set of links it was grouped with by the greedy algorithm. Note that $l \in \widetilde{\chi_l}$.

Assume that the Π_{lower} policy schedules one link in every exclusive set $\widetilde{\chi_l}$, computed by Algorithm 1, whenever there is a non-empty queue in the corresponding set. Since $\widetilde{\chi_l}$ is an exclusive set, a lower bound on the sum of its queues can be obtained by applying Theorem 4.1. The value of the lower bound is incremented and the links in the chosen exclusive set are removed from further consideration. This process is repeated until every link in the system has been used. Since each link appears in exactly one exclusive set, the system-wide lower bound on the expected queue length can be obtained as the sum of the contribution of each link towards the lower bound given by Corollary 4.1.

Algorithm 1 Computing the Lower Bound

```
1: X ← {1, 2 . . . N}
2: BOUND ← 0
3: repeat
4:     Find an exclusive set χ̃ ⊂ X which maximizes LB_χ̃
5:     BOUND ← BOUND + LB_χ̃
6:     X ← X \ χ̃
7: until X = φ
8: return BOUND
```

Corollary 4.1: The sum of expected value of the queue length satisfies:

$$\sum_{i=1}^{N} \mathbf{E}[Q_i] \geq \sum_{i=1}^{N} \frac{\lambda_i + \mathbf{E}[A_i(\sum_{j \in \widetilde{\chi_i}} A_j)] - 2\lambda_i \lambda_{\widetilde{\chi_i}}}{2(1 - \lambda_{\widetilde{\chi_i}})} \tag{IV.20}$$

The total expected network delay, $\bar{D}$, satisfies:

$$\bar{D} = \frac{\sum_{i=1}^{N} \mathbf{E}[Q_i]}{\sum_{i=1}^{N} \lambda_i} \geq \sum_{i=1}^{N} \frac{\lambda_i + \mathbf{E}[A_i(\sum_{j \in \widetilde{\chi_i}} A_j)] - 2\lambda_i \lambda_{\widetilde{\chi_i}}}{2(\sum_{j=1}^{N} \lambda_j)(1 - \lambda_{\widetilde{\chi_i}})} \tag{IV.21}$$

Note, that the above result only requires each arrival processes to be independent over time. In the case where all the arrival stream are also independent of each other, we obtain the following result.

Proposition 4.1: When the arrival streams are independent, the expected value of the sum of queue lengths in the system under any scheduling policy satisfies:

$$\sum_{i=1}^{N} \mathbf{E}[Q_i] \geq \sum_{i=1}^{N} \frac{\lambda_i + \mathbf{Var}[A_i] - \lambda_i \lambda_{\widetilde{\chi_i}}}{2(1-\lambda_{\widetilde{\chi_i}})} \tag{IV.22}$$

The total expected delay in the network, $\bar{D}$, satisfies:

$$\bar{D} \geq \sum_{i=1}^{N} \frac{\lambda_i + \mathbf{Var}[A_i] - \lambda_i \lambda_{\widetilde{\chi_i}}}{2(\sum_{j=1}^{N} \lambda_j)(1-\lambda_{\widetilde{\chi_i}})} \tag{IV.23}$$

A. Discussion

The lower bound is achieved by a fictitious scheduling policy, Π_{lower}, which schedules one link in every exclusive set $\widetilde{\chi}_l$, computed by the algorithm, whenever there is a non-empty queue in the corresponding set. This policy may violate the interference constraints, because the set of scheduled queues may not be a valid activation vector. This is because the links in two exclusive sets may interfere with each other. In other words, we have relaxed the constraints in the queuing system to obtain this bound. Therefore, in general, it is not possible to design a scheduling policy that achieves the lower bound. However, we observe through simulation studies that for several different values of the input load, the performance of the MWM policy is indeed quite close to this bound.

Since the exclusive sets do not completely characterize the capacity region of the network, it may also be expected that if the input load is close to a boundary of the capacity region C, which is different from the boundaries generated by the exclusive sets, the lower bound may perform poorly. Thus, in certain cases, the delay of the system under MWM policy may be close to infinity while the lower bound is much smaller. This motivates the development of an upper bound for the system, which is tight in the sense that whenever the upper bound goes to infinity, the delay of the system under a throughput optimal policy also becomes infinite.

V. Development of an Upper Bound

In this section, we analyze a class of Generalized Maximum Weighted Matching (GMWM($\mathbf{w}$)) policies, parametrized by weights w_i which is described in Figure 4. The MWM policy is a special case, where all the weights w_i are unity. We prove that GMWM achieves 100% throughput for every choice of w, s.t. $\forall i, w_i > 0$, using the Foster-Lyapunov drift criteria for countable Markov chains. The following well known theorem provides Foster's criteria for Positive Recurrent and Ergodic Markov chains [2], [20]

GMWM Scheduling Policy

$$\mathbf{I}(t) = \underset{\mathbf{I}^j \in S}{\operatorname{argmax}} \sum_{i=1}^{N} (w_i Q_i) I_i^j \tag{V.24}$$

where I_i^j is the i^{th} component of the j^{th} activation vector, $\mathbf{I}^j$, in set S and $w_i > 0$ are fixed constants.

Fig. 4. GMWM Scheduling Policy

Theorem 5.1: A countable Markov chain is positive recurrent and ergodic if and only if there exists a positive function $V > 0$ and a finite set of states $\mathcal{E}_o$, such that the following hold:

- Bounded drift from the finite set $\mathcal{E}_o$:
  $\forall\, \mathbf{Q}(t) \in \mathcal{E}_o, \quad \Delta(\mathbf{Q}(t)) < \infty$
- Negative drift from the complement:
  $\forall\, \mathbf{Q}(t) \notin \mathcal{E}_o,\ \exists \epsilon > 0\ s.t., \quad \Delta(\mathbf{Q}(t)) < -\epsilon$
  where

$$\Delta(\mathbf{Q}(t)) \equiv \mathbf{E}[V(\mathbf{Q}(t+1) - V(\mathbf{Q}(t))|\mathbf{Q}(t)]. \tag{V.25}$$

We first design an appropriate Lyapunov function for the system.

$$V(\mathbf{Q}(t)) = \frac{1}{2}\sum_{i=1}^{N} w_i Q_i^2(t) \tag{V.26}$$

Note that if all the weights w_i are chosen to be 1, this is exactly the quadratic Lyapunov function used in [1]. Before we move on to prove the throughput optimality of GMWM, we state a couple of useful definitions.

Definition 5.1: $B(t) = \frac{1}{2}\sum_{i=1}^{N} w_i (A_i(t) - I_i(t))^2$

Since the second moments of the arrival processes are bounded, it follows that $\mathbf{E}[B(t)|\mathbf{Q}(t)]$ is bounded from above by a positive constant c.

Definition 5.2: We define $E_o := \{0, 1, 2, ..\frac{c}{\epsilon w_{min}}\}^N$ to be a finite set of states as required by the Foster's criteria, where w_{min} is the minimum of the weights among w_i and $\epsilon > 0$.

Theorem 5.2: For any input load $\boldsymbol{\lambda} \in C$, the GMWM scheduling algorithm ensures that the resulting DTMC is positive recurrent and ergodic.

Proof: See Appendix A. ■

We now analyze GMWM and derive upper bounds using the following lemma from Lyapunov drift theory [7], [8].

Lemma 5.1: Let $V(\mathbf{Q})$ be a non-negative function of the queue vector and the drift $\Delta(\mathbf{Q}(t))$ be as defined above. Let $P(t)$ be a non-negative process and let $\epsilon > 0$ such that for all time t and all possible $\mathbf{Q}(t)$,
$\Delta(\mathbf{Q}(t)) \leq \mathbf{E}[P(t) - \epsilon h(t)|\mathbf{Q}(t)]$ where h(t) represents a non-negative process that might depend on the queue state. Then the following holds:

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbf{E}[h(\tau)] \leq \limsup_{t\to\infty} \frac{1}{t}\sum_{\tau=0}^{t-1} \frac{\mathbf{E}[P(\tau)]}{\epsilon} \tag{V.27}$$

We are now ready to state our main result that bounds the sum of the expected queue lengths and the expected delay in the system.

Theorem 5.3: Given any input load vector $\boldsymbol{\lambda} \in C$ and any vector $\boldsymbol{\mu} \in C: \quad \forall i, \quad \mu_i > \lambda_i$, the following bound on the expectation of the sum of lengths of queues holds true in a system operating under the GMWM policy where the weights w_i are chosen as $w_i = \frac{1}{(\mu_i - \lambda_i)}$:

$$\sum_{i=1}^{N} \mathbf{E}[Q_i] \leq \sum_{i=1}^{N} \frac{(\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2)}{2(\mu_i - \lambda_i)} \tag{V.28}$$

The total expected network delay, $\bar{D}$, satisfies:

$$\bar{D} \leq \sum_{i=1}^{N} \frac{(\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2)}{2(\sum_{i=1}^{N} \lambda i)(\mu_i - \lambda_i)} \tag{V.29}$$

Proof: See Appendix B. ■

A. Discussion

We have been able to obtain an upper bound that is explicit in the statistics of the arrival process. Note that the upper bound also decouples the contribution of each link towards the total network delay. It is interesting to note that the correlations between the arrival streams do not affect the upper bound. We have analyzed the system when each arrival process is *i.i.d.* in time. The above analysis can be extended to the case when each arrival process $\{A_l(t)\}_{t=1}^{\infty}$, is modulated by a discrete-time, stationary, ergodic Markov chain using the techniques developed in [5].

The upper bound derived in [21] (Theorem 2) for the same system has ϵ in the denominator for each of the N terms in the sum in Eq. (V.28) (with the same numerator), where $\epsilon = \min_i(\mu_i - \lambda_i)$. Hence, the upper bound obtained here is numerically smaller than the state-of-the-art. This has been achieved by choosing the weights w_i, such that the second term on the right hand side of Eq. (B.48) in the Appendix B is equal to the negative of the sum of queues in the system. Thus, the contribution of each queue towards the drift is equal to its queue length, i.e., balanced, resulting in a tighter lower bound.

The above analysis naturally leads us to the question of which $\boldsymbol{\mu} > \boldsymbol{\lambda}$ should be selected in the capacity region C such that the upper bound is minimized. Intuitively this means that the distance between the load vector and the service process should be as large as possible. This can be formulated as an optimization problem to compute the value of $\boldsymbol{\mu}$ that minimizes the upper bound.

Upper Bounding Expected Delay

$$\text{Minimize} \sum_{i=1}^{N} \frac{(\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2)}{2(\mu_i - \lambda_i)}$$
$$\text{subject to } \boldsymbol{\mu} \in C$$

Fig. 5. Optimization Problem for Minimizing the Upper Bound

The optimization problem in Figure 5 is convex because the objective function is convex and the capacity region is also convex, being a convex hull of the activation vectors. The formulation of the problem is very similar to the network utility maximization using convex optimization techniques (see [22]–[24]). Using Lagrangian techniques, the dual, $U(\mathbf{a})$, of the above problem can be decomposed into the following two sub-problems. $\mathbf{a}$ is the set of prices.

$$U(\mathbf{a}) = X_i(\mathbf{a}) + Y(\mathbf{a}) \tag{V.30}$$

where

$$\begin{aligned} X_i(\mathbf{a}) &= \max_{\mu_i > \lambda_i} \left\{ -\frac{(\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2)}{2(\mu_i - \lambda_i)} - a_i \mu_i \right\} \\ &= \min_{\mu_i > \lambda_i} \left\{ \frac{(\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2)}{2(\mu_i - \lambda_i)} + a_i \mu_i \right\} \end{aligned} \tag{V.31}$$

and

$$Y(\mathbf{a}) = \underset{\mathbf{I}^j \in S}{\operatorname{argmax}} \sum_{i=1}^{N} a_i I_i^j \tag{V.32}$$

Algorithm 2 Computing the Optimal Value of $\boldsymbol{\mu}^{opt}$

1: $n \leftarrow 1$
2: Initialize the prices $\mathbf{a}^n$
3: **repeat**
4: $\quad \mu_i^{(n)} \leftarrow \lambda_i + \sqrt{\frac{\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2}{2a_i^{(n)}}}$
5: $\quad Y_i^{(n)} \leftarrow I_i^j$ where $I^j = \operatorname{argmax}_{I^j \in S} \sum_{i=1}^{N} a_i^{(n)} I_i^j$
6: $\quad a_i^{(n+1)} \leftarrow a_i^{(n)} + h^{(n)}(\mu_i^{(n)} - V_i^{(n)})$
7: $\quad n \leftarrow n + 1$
8: **until** $\boldsymbol{\mu}$ converges
9: **return** $BOUND$

The dual problem can be solved using an iterative sub-gradient method shown in Algorithm 2. The dual prices a_i are updated in each iteration. It has been shown in the literature [22]–[24] that if the sequence of values of $\{h\}$ are chosen such that $\lim_{n\to\infty} h^{(n)} \to 0$ and $\sum_{n=0}^{\infty} h^{(n)} = \infty$, then the values of $\mu_i^{(n)}$ converge to the optimal value μ_i^{opt}, which minimizes the upper bound on the expected queue lengths in the system.

The GMWM schemes in which the weights w_i satisfy $\forall i$, $w_i(\mu_i^{opt} - \lambda_i) = 1$ achieve the optimal delay bound and will be referred to as GMWM^{opt} for the rest of the paper. We now show that the delay performance of GMWM^{opt} is no worse than any other stationary randomized policy.

B. Comparison with a Stationary Randomized Policy

We analyze the delay of the wireless network when operated with a stationary randomized scheduler, Π_R. As noted before, in Lemma 3.1, for each link l in the system a service rate of $\mu_l > \lambda_l$ is guaranteed. The service process can be analyzed as follows. The scheduler Π_R is unaware of the backlog and chooses to schedule link l independent of whether the queue is empty or not. In every slot, if the link is scheduled, exactly one packet is served, otherwise the packets in the queue wait for the next available slot.

We define the following for the system.

- $q_l(t)$: Length of the queue l at the beginning of time slot t.

- $A_l(t)$: Number of arrivals at link l during the time slot t.
- $R_l(t)$: Random variable that is 1 if link l is scheduled and is 0 otherwise.
- $\bar{d}$: Average delay in the system.

The system evolves as follows

$$q_l(t+1) = q_l(t) + A_l(t) - R_l(t)\mathbf{1}_{\{q_l(t)>0\}} \tag{V.33}$$

The following is a standard result for GI/D/1 system with Bernoulli service process [8], i.e.,

$$\mathbf{E}[q_l] = \frac{\lambda_l + \mathbf{Var}[A_l] - \lambda_l^2}{2(\mu_l - \lambda_l)} \tag{V.34}$$

Under the stationary randomized policy the behavior of each queue in the system is independent of other queues. Using the fact that the expectation of the sum of independent random variables equals the sum of their expectation, the following lemma follows:

Lemma 5.2: The sum of expected queue lengths of the queues in a discrete-time system constrained queueing system with arrival process A_l (rate λ_l) and service rate μ_l, at link l, operating under a stationary randomized scheduling policy is given by:

$$\sum_{l=1}^{N} \mathbf{E}[q_l] = \sum_{l=1}^{N} \frac{\lambda_l + \mathbf{Var}[A_l] - \lambda_l^2}{2(\mu_l - \lambda_l)}$$

Proof: The proof follows by using Lemma V.34 and using the fact that the service process is Bernoulli with probability μ_i at the queue i independent of other queues in the system. ■

Theorem 5.4: Given any admissible arrival process $\{A_l(t)\}_{t=1}^{\infty}$ (with mean λ_l), the sum of expected queue lengths Q_l under the GMWM^{opt} policy is no worse than the sum of expected queue lengths q_l of any other stabilizing stationary randomized policy. In other words,

$$\sum_{l=1}^{N} \mathbf{E}[Q_l] \leq \sum_{l=1}^{N} \mathbf{E}[q_l]$$

It follows then, that the average delay $\bar{D}$ under GMWM^{opt} is no worse than the average delay $\bar{d}$ under any other stabilizing stationary randomized policy.

$$\bar{D} \leq \bar{d}$$

Proof: Among the class GMWM policies, the upper bound is minimum for the GMWM^{opt}. The result follows by comparing the bound established in Theorem 5.3 for the GMWM policy with weights $w_i = \frac{1}{\mu_i - \lambda_i}$ and expected value result for the stationary randomized policy in Lemma 5.2. ■

It is known that in the heavy traffic limit, the scheme GMWM is asymptotically optimal [14]. However, the result obtained here is true for all load vectors $\boldsymbol{\lambda} \in C$.

VI. Estimating the delay

We noted towards the end of Section IV that the lower bound may not be achieved by any policy because it may not be possible to schedule a link in every exclusive set due to the interference constraints. Therefore, we attempt to develop an accurate estimate for the delay performance in this section.

The lower bound analysis suggests that those exclusive sets that have a large λ_{χ_l}, must have longer queues lengths because the sum of the expected queue lengths in the exclusive set is proportional to $\frac{1}{1-\lambda_{\chi_l}}$. However, since a scheduling policy like MWM also balances the queue lengths in the system, the effect of congestion in a particular exclusive set is distributed over the whole system. Hence, instead of estimating the queue length at each link, we estimate the contribution of each link towards the aggregate expected queue length.

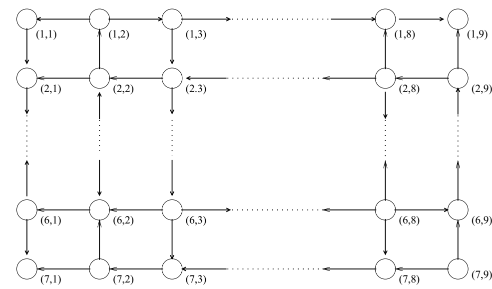

Fig. 6. Grid Topology (Independent Traffic)

The upper bound analysis indicates that the expected aggregate queue length in the system can be expressed a sum of the individual contributions of each link. It also suggests that the contribution of each link is inversely proportional to the congestion, $(\mu_l - \lambda_l)$, at the link l. A similar feature is also noted in the lower bound where the congestion is equal to $(1 - \lambda_{\widetilde{\chi}_l})$, where $\widetilde{\chi}_l$ are the sets computed by Algorithm 1 in Section IV. However, since the sets $\widetilde{\chi}_l$ used to compute the lower bound are not maximal, they do not accurately represent the effect of congestion and multiplexing in the system. Hence, we consider the sets $\widehat{\chi}_l$ (defined below).

We define $\widehat{\chi}_l$ as the exclusive set that has the largest sum of arrival rates, $\lambda_{\chi_l} = \sum_{i \in \chi_l} \lambda_i$ among all exclusive sets containing l. In the case where all the arrival streams are mutually independent, we propose to estimate the total expected delay in the network by the following equation.

$$\sum_{i=1}^{N} \mathbf{E}[Q_i] \approx \sum_{i=1}^{N} \frac{\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2}{2(1 - \lambda_{\widehat{\chi}_i})} \tag{VI.35}$$

The total expected delay in the network, $\bar{D}$ can be estimated as follows:

$$\bar{D} \approx \sum_{i=1}^{N} \frac{\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2}{2(\sum_{j=1}^{N} \lambda_j)(1 - \lambda_{\widehat{\chi}_i})} \tag{VI.36}$$

We call the r.h.s. of the above Equation (VI.35) as the $Estimate(G, \boldsymbol{\lambda})$. Similarly, we call the r.h.s. of Equation (IV.22) as the $LowerBound(G, \boldsymbol{\lambda})$ and the r.h.s. of Equation (V.28) as the $UpperBound(G, \boldsymbol{\lambda})$ respectively. We now show that when the arrival streams are independent, indeed the estimate lies between the upper and lower bounds.

Theorem 6.1: $UpperBound(G, \boldsymbol{\lambda}) \geq Estimate(G, \boldsymbol{\lambda}) \geq LowerBound(G, \boldsymbol{\lambda})$

Proof: The bounds and the estimates have been expressed as a sum of N terms. We first show that each term in the upper bound is no smaller than the corresponding term in the estimate.

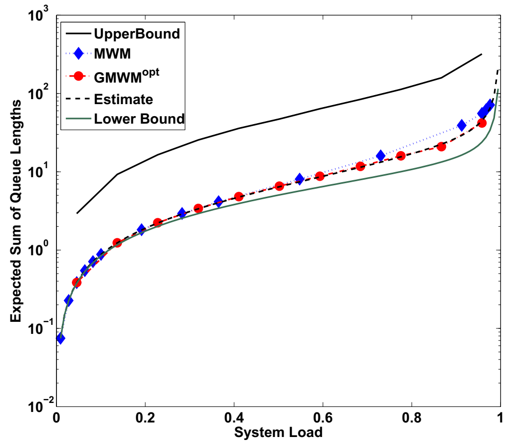

Fig. 7. Expected Queue lengths for Independent Traffic (Grid)

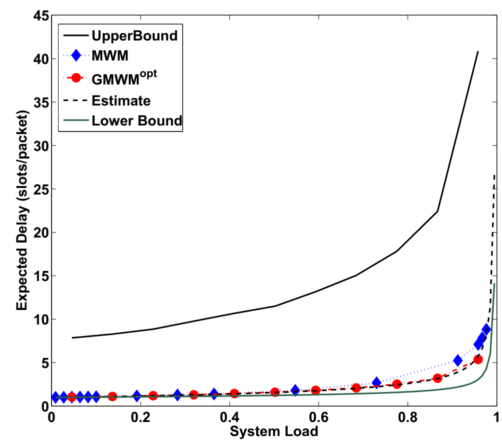

Fig. 8. Expected Delay for Independent Traffic (Grid)

Part 1: Consider link i in the system. As explained in Section III, for any exclusive set χ_i and any $\boldsymbol{\mu} > \boldsymbol{\lambda} : \boldsymbol{\mu} \in C$,

$$\mu_{\chi_i} = \sum_{j \in \chi_i} \mu_j \leq 1$$

$$\implies 1 - \lambda_i - \sum_{j \in \chi_i, j \neq i} \lambda_j \geq \mu_i - \lambda_i + \sum_{j \in \chi_i, j \neq i} (\mu_j - \lambda_j)$$

$$\implies 1 - \lambda_{\chi_i} \geq \mu_i - \lambda_i$$

since, each $\mu_j > \lambda_j$. In particular, we have

$$1 - \lambda_{\widehat{\chi_i}} \geq \mu_i - \lambda_i \quad \text{(VI.37)}$$

Since both sides in Equation (VI.37) are positive, we have the following result,

$$\frac{\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2}{2(1 - \lambda_{\widehat{\chi_i}})} \leq \frac{\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2}{2(\mu_i - \lambda_i)} \quad \text{(VI.38)}$$

Now, we show that each term in the Estimate is no smaller than the corresponding term in the lower bound.

Part 2: Consider link i in the system. By definition of $\widehat{\chi_i}$, $\lambda_{\widehat{\chi_i}}$ is no smaller than $\lambda_{\widetilde{\chi_i}}$ for the sets $\widetilde{\chi_i}$, computed by the Algorithm 1 in Section IV of the paper. Also, $\lambda_{\widetilde{\chi_i}}$ is no smaller than λ_i, i.e.,

$$\lambda_{\widehat{\chi_i}} \geq \lambda_{\widetilde{\chi_i}} \geq \lambda_i \quad \text{(VI.39)}$$

It follows that $(1 - \lambda_{\widetilde{\chi_i}}) \geq (1 - \lambda_{\widehat{\chi_i}})$ and

$$(\lambda_i + \mathbf{Var}[A_i] - \lambda_i \lambda_{\widetilde{\chi_i}}) \leq (\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2) \quad \text{(VI.40)}$$

Using the above two inequalities, we get the desired result,

$$\frac{\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2}{2(1 - \lambda_{\widehat{\chi_i}})} \geq \frac{\lambda_i + \mathbf{Var}[A_i] - \lambda_i \lambda_{\widetilde{\chi_i}}}{2(1 - \lambda_{\widetilde{\chi_i}})} \quad \text{(VI.41)}$$

■

VII. Simulation Results

We present the simulation results for two types of network topologies, grid and random quasi unit disk graphs [25]. In each case, the lower bound is computed using Algorithm 1. The upper bound on the performance of GMWM policy is computed using Algorithm 2 and the corresponding weights are used by the $GMWM^{opt}$ policy. We also simulate MWM policy to provide comparison with the $GMWM^{opt}$ policy. We study the accuracy of the estimate for this class of throughput optimal policies when the arrival streams are mutually independent. We use CPLEX [26] to solve the combinatorial problems of computing the maximum weight scheduling problems at every iteration. The simulations are run until the half-width of the 95% confidence interval is within 2.5% of the mean. All simulation experiments have been conducted under the 2-hop interference model explained in Section III.

A. Grid topology

We simulate two cases, one with with mutually independent arrival streams and another with correlated arrival streams.

1) Independent Arrival Streams: For this simulation, the network is a 7x9 grid with 63 nodes and 110 links as shown in Figure 6. The direction of data transfer among a pair of neighboring nodes is chosen randomly. The arrival process at each link is Poisson with rate parameter λ chosen independently, randomly between 0 and 1 packets per slot. This arrival vector may even be outside the capacity region of the network. Once a random base-line load is chosen, we use a scaling factor to study the delay performance for different values of the (normalized) load in the network. The maximum value of the load that is supported by the system is determined from the simulations. Since MWM is throughput optimal, the point where the system becomes unstable must be outside the capacity region. The input load is then normalized with value 1 corresponding to the point on the boundary of the capacity region. It appears from our simulations that a randomly selected load, when scaled appropriately, usually hits the boundary generated by the exclusive set constraints.

Figure 7 shows the increase in the sum of expected queue lengths in the system as the load is scaled. The queue length increases almost like a quadratic function at low to medium loads. At high loads however, the denominator term $(1-\lambda_{\widetilde{\chi_l}})$, grows very fast. We observe that both the $GMWM^{opt}$ and

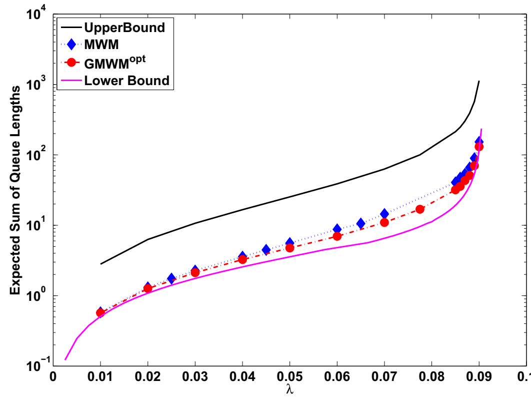

Fig. 9. Expected Queue lengths for Correlated Traffic (Grid)

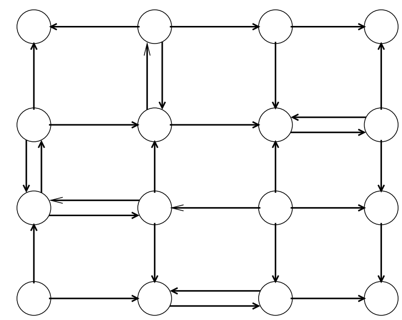

Fig. 10. Grid Topology (Correlated Traffic)

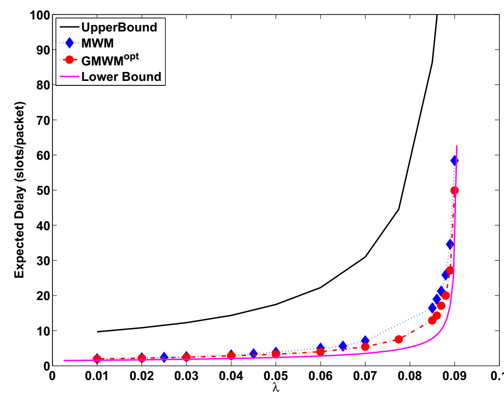

Fig. 11. Expected Delay for Correlated Traffic (Grid)

MWM policies perform close to the lower bound. The estimate closely matches the queue lengths of both MWM and $GMWM^{opt}$ policies, however it is more accurate for the $GMWM^{opt}$ policy. The upper bound, although tight in an order sense, is almost always a constant multiple of the average queue length in the system. It seems that for each link l, the term $(1-\lambda_{\widetilde{\chi_l}})$ in the estimate is a constant multiple of $(\mu_l-\lambda_l)$, selected by the $GMWM^{opt}$ policy. This suggests that under the MWM type scheduling policies, the system behaves as if all the queues in the exclusive set χ_l have been multiplexed into a single queue.

The delay in the system increases rather slowly when the system load is in the low to medium range. However, as expected, the increase is sharp as the load approaches the capacity region boundary. It seems that the lower bound analysis was rather optimistic for heavy loads because it assumed that all the exclusive sets generated by the Algorithm 1 can be scheduled at the same time if they have non zero queue lengths. At low and medium loads, since many of the exclusive sets are likely to have small queue lengths, the lower bound appears to be tight. The fact that even for an optimistic lower bound, the MWM and GMWM perform so close to the lower bound indicates that they are nearly optimal.

2) Correlated arrival streams: We simulated a 4×4 grid with 29 links with link directions as shown in Figure 10. The arrival process at each link is Poisson with the same rate parameter λ. All the flows originating from the same node have exactly the same arrivals, i.e. they are perfectly correlated. The upper bound and the lower bound analysis is general enough to correlations in the arrival process and the results are shown below.

Figure 9 shows the increase in the sum of expected queue lengths in the system as the value of λ is increased. We observe that the delay performance of the $GMWM^{opt}$ policy is better than that of the MWM on account of a better choice of weights which increase the chances of scheduling the more congested links in the network. Figure 11 shows that the lower bound is quite close to the performance of the $GMWM^{opt}$ even when there are correlations among the arrival streams.

B. Random Quasi Unit Disk Topology

We generate a random quasi unit disk graph shown in Figure 12 with 40 nodes and 92 links. We allow a neighboring pair of nodes to transfer data in both directions (for the sake of simplicity, the links in the figure are shown as undirected links). The arrival rate λ_l at each link l is chosen randomly between 0.1 and 1 packet per slot. Let Geometric(p) denote a sample from the geometric distribution with parameter p. The arrival process at each link l, is chosen as follows:

$$A_l(t) = \begin{cases} \text{Geometric}(\frac{2}{2+\lambda_l}) & \text{with probability } \frac{16}{16+\lambda_l} \\ \text{Geometric}(\frac{1}{9+\lambda_l}) & \text{with probability } \frac{\lambda_l}{16+\lambda_l} \end{cases}$$

The first two moments of A_l are λ_l and $9\lambda_l + 2\lambda_l^2$ respectively. This load is scaled in a manner similar to the previous case, to study the performance of the system at different loads. The results are practically similar to the previous case. We note additionally that the estimates and lower bounds capture the variance in the arrival process quite accurately.

Thus, even though the lower bound in not guaranteed to be tight in every case, it nonetheless provides a useful estimate of the delay. Notice that the upper bound is finite for any $lambda \in C$. Also note that the delay of any scheduling policy must be infinite if the load is outside the capacity region. Therefore, we can conclude that as the upper bound goes to infinity, the delay of any throughput optimal policy must also become infinite. Further, from our simulations, it appears that the upper bound is a constant multiple of the delay of the

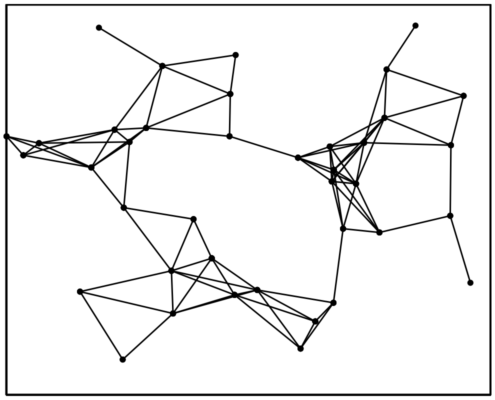

Fig. 12. Quasi Unit Disk Topology

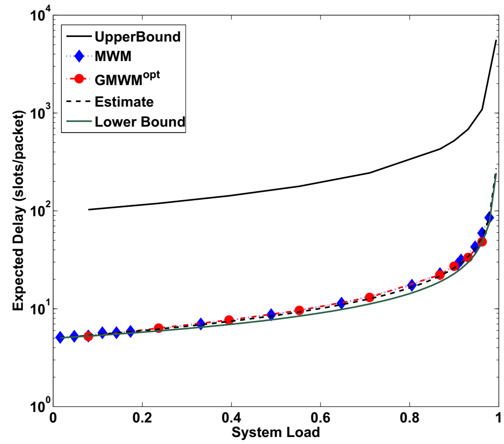

Fig. 14. Expected Delay for Quasi Unit Disk Topology

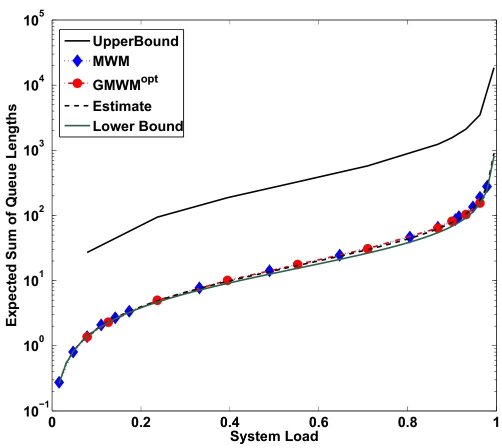

Fig. 13. Expected Queue lengths for Quasi Unit Disk Topology

MWM/GMWM policy.

VIII. Conclusion

We have established a fundamental lower bound on the performance of a wireless system with single-hop traffic and general interference constraints. This result can be used to study the relative performance of any scheduling policy. We observed through simulations that the performance of the throughput optimal policies such as the MWM policy is very close to the lower bound. It is interesting to note that the MWM type of policies, which were designed primarily for achieving maximum throughput, indeed also have good delay performance. This can be attributed to two reasons. Firstly, MWM schedules a maximal set of links in the system. Secondly, it performs load balancing in the system.

We have analyzed the impact of GMWM type of scheduling policies on the expected queue lengths and expected delay in the system. The GMWM^{opt} policy analyzed in the paper, uses the information of the arrival rates to the links to achieve load balancing by assigning higher weights w_i to more congested links. Thus, it improves the delay performance. We have shown that for any given $\lambda \in C$, the performance of GMWM^{opt} is no worse than any stationary randomized scheduling policy. It is interesting to note that the MWM policy achieves load balancing without explicit knowledge of the arrival statistics, simply by using the information of the backlogs and thus achieves a delay performance comparable to that of the GMWM^{opt} policy.

Note that our approach is orthogonal to that taken by [27] where functions of the type Q_i^{α}, $\alpha > 0$ were used to compute the weight of the matching. This was explored further in [28] where it was suggested that a smaller value of α may decrease the idling in the system, leading to smaller delays. In our approach, the knowledge of the arrival rates at different links in the system is used to compute the weight, w_i corresponding to each link i. In the GMWM policy w_i is a fixed constant that serves to increase the chances of scheduling a more congested link as compared to a less congested one, even when its instantaneous queue length is small.

Finally, we have developed an accurate estimate of the performance of MWM type scheduling schemes. This result can be used to study the relative performance of other scheduling policies for wireless networks. The proposed delay estimate can also be used as a more accurate metric for the development of the scheme studied in [17]. We have developed bounds and estimates for the expected value of the sum of all queue lengths in the system. Since the policies like MWM, balance queue lengths in the system, the above analysis can be used to estimate the individual queue lengths in the system. Thus, if the total expected queue length in the network is small, we can expect the average queue length at an individual link to be also small.

Since the complexity of implementing MWM/GMWM is high, the design of distributed algorithms based on these properties is an important avenue for future investigation. The study of throughput and stability of MWM has resulted in numerous interesting works on the development of far simpler practically implementable throughput-efficient schedulers. Similarly, we expect that this study of the delay characteristics of MWM will also result in simpler and more delay efficient schedulers.

As future work, we would like to analyze the delay of a wireless network with multi-hop traffic.

Appendix A
Proof of Theorem 5.2

We begin with the calculation of the drift for any state **Q**(t).

$$\begin{aligned}
&\Delta(\mathbf{Q}(t))\\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathbf{E}[(Q_i(t+1)-Q_i(t))(Q_i(t+1)+Q_i(t))|\mathbf{Q}(t)]\\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathbf{E}[(A_i(t)-I_i(t))(2Q_i(t)+A_i(t)-I_i(t))|\mathbf{Q}(t)]\\
&= \sum_{i=1}^{N} w_i \mathbf{E}[(A_i(t)-I_i(t))(Q_i(t))|\mathbf{Q}(t)]\\
&+ \frac{1}{2}\sum_{i=1}^{N} w_i \mathbf{E}[(A_i(t)-I_i(t))^2|\mathbf{Q}(t)]
\end{aligned} \tag{A.42}$$

We now invoke the assumption that the arrivals are *i.i.d.* over the time slots and hence have expected values that are independent of the current queue states. Also, since $\boldsymbol{\lambda} \in C$,

$$\lambda_i = \sum_{j=1}^{|S|} \alpha_j I_i^j \text{ such that } \sum_{j=1}^{|S|} \alpha_j < 1$$

Therefore we have

$$\begin{aligned}
&\sum_{i=1}^{N} w_i \mathbf{E}[(A_i(t)-I_i(t))(Q_i(t))|\mathbf{Q}(t)]\\
&= \sum_{i=1}^{N} w_i \lambda_i Q_i(t) - \sum_{i=1}^{N} w_i \mathbf{E}[I_i(t)Q_i(t)|\mathbf{Q}(t)] \qquad \text{(A.43)}\\
&= \sum_{i=1}^{N} w_i \sum_{j=1}^{|S|} \alpha_j I_i^j Q_i(t) - \sum_{i=1}^{N} w_i I_i(t) Q_i(t)
\end{aligned}$$

Since $\mathbf{I}(t)$ is the optimal activation vector chosen according to the GMWM rule,

$$\forall j, \ \sum_{i=1}^{N} w_i I_i(t) Q_i(t) \geq \sum_{i=1}^{N} w_i I_i^j Q_i(t)$$

Hence,

$$\begin{aligned}
&\sum_{i=1}^{N} w_i \mathbf{E}[(A_i(t)-I_i(t))(Q_i(t))|\mathbf{Q}(t)]\\
&\leq -(1-\sum_{j=1}^{|S|}\alpha_j)\sum_{i=1}^{N} w_i I_i(t) Q_i(t) \qquad \text{(A.44)}\\
&< -\epsilon \sum_{i=1}^{N} w_i I_i(t) Q_i(t), \ \epsilon > 0
\end{aligned}$$

Using Equations (A.42) and (A.44) and Definition 5.1, we have

$$\Delta(\mathbf{Q}(t)) < -\epsilon \sum_{i=1}^{N} w_i I_i(t) Q_i(t) + \mathbf{E}[B(t)|\mathbf{Q}(t)]$$

Then for $\mathbf{Q}(t) \in E_o$ the drift is bounded by c (defined in Section V).

For $\mathbf{Q}(t) \notin E_o$, $\epsilon \sum_{i=1}^{N} w_i I_i(t) Q_i(t) > c$ and hence $\Delta(\mathbf{Q}(t)) < -\eta, \eta > 0$. Hence by the Foster-Lyapunov criteria in Theorem 5.1, the DTMC **Q**(t) is positive recurrent and ergodic.

Appendix B
Proof of Theorem 5.3

We use Equation (A.43) from the proof of Theorem 5.2 to arrive at the following:

$$\Delta(\mathbf{Q}(t)) = \mathbf{E}[B(t)|\mathbf{Q}(t)] + \sum_{i=1}^{N} w_i \lambda_i Q_i(t) - \sum_{i=1}^{N} w_i \mathbf{E}[I_i(t)Q_i(t)|\mathbf{Q}(t)]$$

Note that $\mathbf{I}(t)$ is the activation vector chosen by the GMWM scheme at time-slot t. For any other activation vector $\mathbf{I}^* \in S$, the following holds true:

$$\sum_{i=1}^{N} w_i \mathbf{E}[I_i^*(t)Q_i(t)|\mathbf{Q}(t)] \leq \sum_{i=1}^{N} w_i \mathbf{E}[I_i(t)Q_i(t)|\mathbf{Q}(t)] \tag{B.45}$$

Hence,

$$\begin{aligned}
\Delta(\mathbf{Q}(t)) \leq \mathbf{E}[B(t)|\mathbf{Q}(t)] + \sum_{i=1}^{N} w_i \lambda_i Q_i(t) -\\
\sum_{i=1}^{N} w_i \mathbf{E}[I_i^*(t)Q_i(t)|\mathbf{Q}(t)]
\end{aligned}$$

Now, we use Lemma 3.1 which shows the existence of a stationary randomized policy Π_R with rates greater than $\boldsymbol{\lambda}$. Suppose the activation vector picked by Π_R at time t is $\mathbf{I^R}(t)$. We define another scheduling policy $\mathbf{I}^*$ which schedules at time t, all the queues scheduled by $\mathbf{I^R}(t)$ except for those whose queues are empty. We define $\mathbf{I}^*$ as follows:

$$I_i^*(t) = \begin{cases} I_i^R(t) & \text{if } Q_i(t) > 0 \\ 0 & \text{if } Q_i(t) = 0 \end{cases}$$

It follows that

$$\begin{aligned}
&\mathbf{E}[I_i^*(t)Q_i(t)|\mathbf{Q}(t)] = \mathbf{E}[I_i^R(t)Q_i(t)|\mathbf{Q}(t)],\\
&\sum_{i=1}^{N} w_i \mathbf{E}[I_i^*(t)Q_i(t)|\mathbf{Q}(t)] = \sum_{i=1}^{N} w_i \mathbf{E}[I_i^R(t)Q_i(t)|\mathbf{Q}(t)]
\end{aligned} \tag{B.46}$$

Therefore,

$$\begin{aligned}
\Delta(\mathbf{Q}(t)) \leq \mathbf{E}[B(t)|\mathbf{Q}(t)] + \sum_{i=1}^{N} w_i \lambda_i Q_i(t)\\
- \sum_{i=1}^{N} w_i \mathbf{E}[I_i^R(t)Q_i(t)|\mathbf{Q}(t)]
\end{aligned} \tag{B.47}$$

But, I_i^R is a stationary randomized policy and we have

$$\begin{aligned}
&\mathbf{E}[I_i^R] = \mu_i, \mu_i \geq \lambda_i\\
&\mathbf{E}[I_i^R(t)Q_i(t)|\mathbf{Q}(t)] = \mu_i Q_i(t)
\end{aligned}$$

Hence,

$$\Delta(\mathbf{Q}(t)) \leq \mathbf{E}[B(t)|\mathbf{Q}(t)] + \sum_{i=1}^{N} w_i(\lambda_i - \mu_i)Q_i(t) \quad \text{(B.48)}$$

Plugging the value of the weights, $w_i = \frac{1}{(\mu_i - \lambda_i)}$ in Equation (B.48), we have

$$\Delta(\mathbf{Q}(t)) \leq \mathbf{E}[B(t)|\mathbf{Q}(t)] - \sum_{i=1}^{N} Q_i(t)$$

and thus by the application of Lyapunov drift Lemma 5.1 we have:

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbf{E}[\sum_{i=1}^{N} Q_i(\tau)] \leq \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbf{E}[B(\tau)] \quad \text{(B.49)}$$

Let us now compute $\mathbf{E}[B(t)]$.

$$B(t) = \frac{1}{2} \sum_{i=1}^{N} w_i(A_i(t) - I_i(t))^2$$

The queueing system is stable under the GMWM policy and since $I_i(t)$ takes value either 0 or 1, it follows that

$\limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbf{E}[I_i^2(\tau)] = \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbf{E}[I_i(\tau)] = \lambda_i$

Also, $\mathbf{E}[A_i(t)] = \lambda_i$ and

$\mathbf{E}[A_i(t)I_i(t)] = \mathbf{E}[A_i(t)]\mathbf{E}[I_i(t)] = \lambda_i^2$

Finally, we arrive at the following:

$\limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbf{E}[B(\tau)] = \frac{1}{2}[\sum_{i=1}^{N} w_i(\lambda_i + \mathbf{E}[A_i^2] - 2\lambda_i^2)]$.

We have already established the ergodicity of the queue length process and we conclude that the steady state queue occupancies can be upper bounded by

$$\begin{aligned}\sum_{i=1}^{N} \mathbf{E}[Q_i] \leq& \frac{1}{2}[\sum_{i=1}^{N} w_i(\lambda_i + \mathbf{E}[A_i^2] - 2\lambda_i^2)] \\ =& \sum_{i=1}^{N} \frac{(\lambda_i + \mathbf{E}[A_i^2] - 2\lambda_i^2)}{2(\mu_i - \lambda_i)} \\ =& \sum_{i=1}^{N} \frac{(\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2)}{2(\mu_i - \lambda_i)}\end{aligned}$$

The upper bound for average network delay follows by the application of Little's law.

$$\bar{D} \leq \sum_{i=1}^{N} \frac{(\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2)}{2(\sum_{i=1}^{N} \lambda i)(\mu_i - \lambda_i)}$$

References

[1] T. Leandros and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Aut. Contr.37*, vol. 37, no. 12, pp. 1936–1948, 1992.

[2] E. Leonardi, M. Mellia, F. Neri, and M. A. Marsan, "On the stability of input-queued switches with speed-up," *IEEE/ACM Transactions on Networking*, vol. 9, Feburary 2001.

[3] A. Eryilmaz, R. Srikant, and J. R. Perkins, "Stable scheduling policies for fading wireless channels," *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 411–424, 2005.

[4] M. Andrews and L. Zhang, "Scheduling algorithms for multi-carrier wireless data systems.," in *MOBICOM*, pp. 3–14, ACM, 2007.

[5] M. J. Neely, "Delay analysis for maximal scheduling in wireless networks with bursty traffic," in *IEEE INFOCOM*, 2008.

[6] A. Gupta, X. Lin, and R. Srikant, "Low-complexity distributed scheduling algorithms for wireless networks." IEEE INFOCOM, 2007.

[7] L. Georgiadis, M. J. Neely, and L. Tassiulas, *Resource Allocation and Cross-Layer Control in Wireless Networks, Foundations and Trends in Networking*, vol. 1. Now Publishers, 2006.

[8] M. J. Neely, "Order optimal delay for opportunistic scheduling in multi-user wireless uplinks and downlinks," in *44th Annual Allerton Conference on Communication, Control, and Computing*, September 2006.

[9] G. Sharma, C. Joo, and N. B. Shroff, "Distributed scheduling schemes for throughput guarantees in wireless networks," in *the 44th Annual Allerton Conference on Communications, Control, and Computing*, Sep 2006.

[10] D. Shah and M. Kopikare, "Delay bounds for the approximate maximum weight matching algorithm for input queued switches," in *INFOCOM*, 2002.

[11] L. Ying, R. Srikant, A. Eryilmaz, and G. E. Dullerud, "A large deviations analysis of scheduling in wireless networks," *IEEE Transactions on Information Theory*, vol. 52, pp. 5088–5098, 2006.

[12] X. Lin, "On characterizing the delay performance of wireless scheduling algorithms," in *44th Annual Allerton Conference on Communication, Control and Computing*, September 2006.

[13] V. J. Venkataramanan and X. Lin, "Structural properties of ldp for queue-length based wireless scheduling algorithms," in *45th Annual Allerton Conference on Communication, Control and Computing*, September 2007.

[14] A. L. Stolyar, "Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic," *Annals of Applied Probability*, vol. Vol.14, no. No.1, pp. pp.1–53., 2004.

[15] S. Shakkottai, R. Srikant, and A. Stolyar, "Pathwise optimality of the exponential scheduling rule for wireless channels," *Advances in Applied Probability*, December 2004.

[16] K. Jain, J. Padhye, V. N. Padmanabhan, and L. Qiu, "Impact of interference on multi-hop wireless network performance," in *MOBICOM*, (New York, NY, USA), pp. 66–80, ACM, 2003.

[17] Y. Xi and E. M. Yeh, "Optimal capacity allocation, routing, and congestion control in wireless networks," in *Proceedings of the 2006 International Symposium on Information Theory*, July 2006.

[18] H. Balakrishnan, C. Barrett, V. Kumar, M. Marathe, and S. Thite, "The distance-2 matching problem and its relationship to the maclayer capacity of ad hoc networks," *IEEE Journal on Selected Area in Communications*, vol. 22, 2004.

[19] G. Sharma, R. R. Mazumdar, and N. B. Shroff, "On the complexity of scheduling in wireless networks," in *MOBICOM*, (New York, NY, USA), pp. 227–238, ACM Press, 2006.

[20] G. Fayolle, V. A. Malyshev, and M. V. Menshikov, *Topics in the Constructive Theory of Countable Markov Chains*. Cambridge University Press, 1995.

[21] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time varying wireless networks," *IEEE/ACM Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 89–103, 2005.

[22] S. H. Low and D. E. Lapsley, "Optimization flow control: basic algorithm and convergence," *IEEE/ACM Trans. Netw.*, vol. 7, no. 6, pp. 861–874, 1999.

[23] X. Wang and K. Kar, "Cross-layer rate control for end-to-end proportional fairness in wireless networks with random access," in *MobiHoc '05: Proceedings of the 6th ACM international symposium on Mobile ad hoc networking and computing*, (New York, NY, USA), pp. 157–168, ACM, 2005.

[24] X. Lin and N. B. Shroff, "The impact of imperfect scheduling on cross-layer congestion control in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 14, pp. 302–315, April 2006.

[25] F. Kuhn, R. Wattenhofer, and A. Zollinger, "Ad-hoc networks beyond unit disk graphs," in *In Proceedings on the 1st ACM Joint Workshop on Foundations of Mobile Computing (DIALM-POMC), San Diego, California, USA.*, Sepetember 2003.

[26] I. Ilog, "Solver cplex," 2007. http://www.ilog.com/products/cplex/.

[27] I. Keslassy and N. McKeown, "Analysis of scheduling algorithms that provide 100% throughput in input-queued switches," in *Proceedings of the 39th Annual Allerton Conference on Communication, Control, and Computing*, October 2001.

[28] D. Shah and D. Wischik, "Optimal scheduling algorithms for input-queued switches," in *INFOCOM*, 2006.