Noname manuscript No.
(will be inserted by the editor)

Queue-Proportional Rate Allocation with Per-Link Information in Multihop Wireless Networks

Bin Li · R. Srikant

the date of receipt and acceptance should be inserted later

Abstract The backpressure scheduling algorithm for multihop wireless networks is known to be throughput optimal, but it requires each node to maintain per-destination queues. Recently, a clever generalization of processor sharing has been proposed which is also throughput optimal, but which only uses per-link queues. Here we propose another algorithm called Queue Proportional Rate Allocation (QPRA) which also only uses per-link queues, and allocates service rates to links in proportion to their queue-lengths and employs the Serve-In-Random-Order (SIRO) queueing discipline within each link. Through fluid limit techniques and using a novel Lyapunov function, we show that the QPRA achieves the maximum throughput. We demonstrate an advantage of QPRA by showing that, for the so-called primary interference model, it is able to develop a low-complexity scheduling scheme which approximates QPRA and achieves a constant fraction of the maximum throughput region, independent of network size.

Keywords Resource Allocation · Low-complexity Algorithm Design · Multihop Networks · Maximum Throughput · Lyapunov Function · Fluid Limit Analysis

1 Introduction

We consider the resource allocation problem in multihop wireless networks with fixed routing, where each packet may traverse multiple links before departure. The work of Tassiulas and Ephremides (see [35]) developed a throughput-optimal

An earlier version of this paper has appeared in ACM SIGMETRICS 2015 [20].

Bin Li
Coordinated Science Lab
University of Illinois at Urbana-Champaign
lib@illinois.edu

R. Srikant
Coordinated Science Lab and Department of ECE
University of Illinois at Urbana-Champaign
rsrikant@illinois.edu

backpressure algorithm, which prioritizes activation of routes with the largest differential backlog awaiting service subject to network interference constraints. Here, the throughput-optimal strategy means that it can achieve any throughput subject to network stability that is achievable by any other scheduling strategy. A large body of work has extended throughput performance to other metrics, such as fairness (e.g., [10,28,22,34,19]), average energy consumption (e.g., [29,27,5]), Quality-of-Service (e.g., [13,14,37,18]) etc.; see [33] for an overview.

However, all the policies mentioned above make transmission decisions by maintaining per-destination queueing information at each node and frequently exchanging this information among neighboring nodes, which is usually a difficult task in large networks handling thousands of flows. Such a restriction has motivated recent research efforts to develop more practical throughput-optimal schedulers (e.g., [23,15]) with reduced queueing information. For example, the authors in [15] developed a scheduling algorithm with per-link queueing information in limited network setups, where routes do not form loops. But, it is likely to happen in practice that different routes do form a loop, in which case the results in [15] do not apply. Additionally, the arrival rates are assumed to be either known or measured in this line of work.

More recently, motivated by the research activities in bandwidth sharing networks (e.g., [25,2,26,32]), the author in [36] intelligently generalized the idea of processor-sharing in queueing networks, and proposed the throughput-optimal Proportional Scheduler that only utilizes per-link queueing information. Therefore, the Proportional Scheduler significantly simplifies the queueing structure compared to the well-known backpressure algorithm, since the number of traffic flows is generally orders of magnitude greater than the number of links in communication networks. However, the Proportional Scheduler requires the network to solve a concave optimization problem with the full knowledge of the capacity region and its low-complexity implementation still remains an open question.

In this work, we propose alternative throughput-optimal scheduling algorithm, called Queue-Proportional Rate Allocation (QPRA), which also makes scheduling decisions only based on per-link queueing information. The proposed QPRA algorithm is a natural generalization of the algorithm developed in [11] that has been shown to be throughput-optimal in single-hop networks, where packets immediately leave the network once they are served. In particular, the QPRA algorithm allocates the service rates to links in proportion to their queue-lengths as in [11] and additionally employs the Service-In-Random-Order (SIRO) queueing discipline within each link. However, stability in a single-hop network does not necessarily imply stability in a multihop network (e.g., [24,30,4]).

By using fluid limit techniques and a novel Lyapunov function, we are able to prove that the proposed QPRA algorithm achieves maximum throughput. Further, for the commonly-used primary interference model (see, for example, [12]), we develop a low-complexity scheduling scheme approximating the QPRA algorithm that not only uses per-link queueing information but also achieves a constant fraction of the maximum throughput region, independent of network size. To the best of our knowledge, this is the first work that addresses both computational and queueing complexities in multihop wireless networks. The following items list our main contributions along with references on where they appear in the text:

• In Section 3, we develop a scheduling algorithm that allocates service rates to links in proportion to their queue-lengths and employs the SIRO queueing

discipline within each link. Then, we show the stability of our proposed algorithm for any arrival rate vector within the maximum throughput region.

• To illustrate the advantage of our proposed algorithm, we develop an efficient low-complexity scheduling algorithm that mimics our proposed QPRA algorithm in Section 4, and show that it at least achieves a constant fraction of the maximum throughput region, independent of network size, in networks with primary interference.

A note on Notation: We use bold and script font to denote a vector and a set, respectively. Also, let $|\mathcal{A}|$ denote the cardinality of the set $\mathcal{A}$. We use $\text{Int}(\mathcal{A})$ to denote the set of interior points of the set $\mathcal{A}$. We use the convention that $0/0 = 0$.

2 System Model

We consider a wireless network represented by a graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$, where $\mathcal{N}$ is the set of nodes and $\mathcal{L}$ is the set of links. A node represents a wireless transmitter or receiver, while a link represents a pair of transmitter and receiver that are within the transmission range of each other. We assume that the system operates in slotted time. For ease of exposition, we assume that each link can at most transfer one packet in one time slot. We consider the *link-based conflict model*, where links conflicting with each other cannot be active at the same time. We call a set of links that can be active simultaneously as a *feasible schedule* and denote it as $\mathbf{R}(t) \triangleq (R_l(t))_{l\in\mathcal{L}}$, where $R_l(t) = 1$ if the link l is scheduled in time slot t and $R_l(t) = 0$, otherwise. We define the *capacity region* $\mathbf{\Lambda}$ as the convex hull of all feasible schedules.

We consider a *multihop traffic model*, where packets traverse multiple links within the network before they depart. We assume that each route r consists of $K^{(r)}$ consecutive links that do not form a cycle. That is, a route r packet served at link $l_k^{(r)}$ must next go to link $l_{k+1}^{(r)}$ until it is served at the last link $l_{K^{(r)}}^{(r)}$, where $k = 1, 2, \cdots, K^{(r)} - 1$. Let $\mathcal{R}$ be a set of routes through the network. We use the notation $l \in r$ to denote that link l is part of route r. We also say pair (l, r) if $l \in r$. Further, we denote $l_-^{(r)}$ as the previous (upstream) link of link l on route r.

We use $A_{l,r}(t)$ to denote the number of *exogenous* route r packets arriving at the ingress link $l = l_1^{(r)}$ in time slot t. Note that $A_{l,r}(t) \equiv 0$ if $l \neq l_1^{(r)}$. Let $A_{l,r}^{\Sigma}(t) \triangleq \sum_{\tau=0}^{t-1} A_{l,r}(\tau)$ denote the cumulative number of route r packets arriving at the ingress link l up to time slot $t-1$. The arrival processes are assumed to satisfy the Strong Law of Large Numbers (SLLN), i.e.,

$$\lambda_{l,r} = \lim_{t\to\infty} \frac{A_{l,r}^{\Sigma}(t)}{t}, \text{ with probability 1,} \tag{1}$$

where $\lambda_{l,r} \triangleq \lambda_r \mathbb{1}_{\{l=l_1^{(r)}\}}$, and $\lambda_r > 0$ is the mean arrival rate of route r packets. For simplicity, here we will assume that the arrival process for each route are i.i.d., and independent across routes[1].

[1] Since we use fluid limit techniques, this assumption can be relaxed in many different ways at the cost of additional notation. In particular, we only need a Markovian description of the queueing system for our results to hold.

A queue is maintained for each link l with $Q_l(t)$ denoting its queue-length at the beginning of time slot t. Let $X_{l,r}(t)$ be the number of route r packets at link l in time slot t. Thus,

$$Q_l(t) = \sum_{r:l\in r} X_{l,r}(t), \quad \forall t \geq 0. \tag{2}$$

We use $H_l(t)$ to denote the number of packets departing from link l in time slot t. Let $D_{l,r}(t)$ be the number of route r packets departing from link l in time slot t. Therefore, we have

$$H_l(t) = \sum_{r:l\in r} D_{l,r}(t), \quad D_{l,r}(t) \leq X_{l,r}(t), \quad \forall t \geq 0. \tag{3}$$

Based on the above setup, the queueing dynamics of a multihop network can be described as follows:

$$X_{l,r}(t+1) = X_{l,r}(t) + A_{l,r}(t) + D_{l_-^{(r)},r}(t) - D_{l,r}(t), \tag{4}$$

holds for any pair (l,r) and $t \geq 0$. Here, we note that $D_{l_-^{(r)},r}(t) = 0$ if $l = l_1^{(r)}$.

In this paper, we consider the policies under which the system evolves as a Markov Chain. We call system *stable* if the underlying Markov Chain is positive recurrent. We define the *maximum throughput region* $\overline{\mathbf{\Lambda}}$ as the set of arrival rate vectors $\boldsymbol{\lambda} \triangleq (\lambda_r)_{r\in\mathcal{R}}$ for which the network is stable under some policy. It has been shown in [35] that the maximum throughput region $\overline{\mathbf{\Lambda}}$ can be represented as $\overline{\mathbf{\Lambda}} \triangleq \{(\lambda_r)_{r\in\mathcal{R}} : (\sum_{r:l\in r}\lambda_r)_{l\in\mathcal{L}} \in \mathbf{\Lambda}\}$, where we recall that $\mathbf{\Lambda}$ denotes the capacity region. We call an algorithm *throughput-optimal* if it makes the system stable for any arrival rate vector $\boldsymbol{\lambda}$ that lies strictly within the maximum throughput region $\overline{\mathbf{\Lambda}}$, i.e., $\boldsymbol{\lambda} \in \mathrm{Int}(\overline{\mathbf{\Lambda}})$. An algorithm is said to achieve an *efficiency ratio* $\kappa \in (0,1]$ if it can stabilize the system for any $\boldsymbol{\lambda}$ strictly within a fraction κ of the maximum throughput region $\overline{\mathbf{\Lambda}}$, i.e., $\boldsymbol{\lambda} \in \mathrm{Int}(\kappa\overline{\mathbf{\Lambda}})$.

The aim of this work is to address the scheduler design with only per-link queueing information in multihop networks. We first present a centralized throughput-optimal algorithm that makes transmission decisions only based on the link queue-lengths. Then, we show that this algorithm can be easily modified for low-complexity implementations, which achieve a strictly positive efficiency ratio for certain interference models, independent of the network size.

3 Queue-Proportional Rate Allocation scheduler

In this section, we present a throughput-optimal scheduling algorithm that makes transmission decisions only based on per-link queueing information.

3.1 Algorithm description

Here we develop a scheduling algorithm that only utilizes per-link queueing information and achieves the maximum throughput region. Given the link queue-length vector $\mathbf{Q}(t) = (Q_l(t))_{l\in\mathcal{L}}$, the algorithm operates as follows:

Queue-Proportional Rate Allocation (QPRA):
(1) In each time slot t, first allocate the service rate $\sigma_l(\mathbf{Q}(t))$ for each link l, such that

$$\begin{aligned}&\sigma_l(\mathbf{Q}(t)) = 0, \text{whenever } Q_l(t) = 0,\\ &\frac{\sigma_l(\mathbf{Q}(t))}{Q_l(t)} = \frac{\sigma_{l'}(\mathbf{Q}(t))}{Q_{l'}(t)}, \text{ whenever } Q_l(t) > 0 \text{ and } Q_{l'}(t) > 0,\end{aligned} \tag{5}$$

and $\boldsymbol{\sigma}(\mathbf{Q}(t)) = (\sigma_l(\mathbf{Q}(t)))_{l\in\mathcal{L}}$ lies on the boundary of the capacity region $\boldsymbol{\Lambda}$. That is, the allocated service rate vector $\boldsymbol{\sigma}(\mathbf{Q}(t))$ is the longest vector within the capacity region $\boldsymbol{\Lambda}$ that is along the same direction of the link queue-length vector $\mathbf{Q}(t)$. Let $\mathbf{R}(t) = (R_l(t))_{l\in\mathcal{L}}$ be a random vector with the support on the set of all feasible schedules and mean $\boldsymbol{\sigma}(\mathbf{Q}(t))$, i.e., $\mathbb{E}\left[R_l(t)|\mathbf{Q}(t)\right] = \sigma_l(\mathbf{Q}(t)), \forall l \in \mathcal{L}$. Recall that link l is scheduled in time slot t if $R_l(t) = 1$.
(2) Then, serve each link l according to Serve-In-Random-Order (SIRO) queueing discipline, i.e., serve packets at each link l uniformly at random. This implies that the average departure rate of each route at link l is proportional to the number of packets of its route, i.e.,

$$\begin{aligned}&\mu_{l,r}(\mathbf{X}(t)) = 0, \text{ whenever } X_{l,r}(t) = 0,\\ &\frac{\mu_{l,r}(\mathbf{X}(t))}{X_{l,r}(t)} = \frac{\mu_{l,r'}(\mathbf{X}(t))}{X_{l,r'}(t)}, \text{ whenever } X_{l,r}(t) > 0 \text{ and } X_{l,r'}(t) > 0,\end{aligned} \tag{6}$$

where $\mu_{l,r}(\mathbf{X}(t)) \triangleq \mathbb{E}\left[D_{l,r}(t)\middle|\mathbf{X}(t)\right]$ denotes the average departure rate of route r packets at link l in time slot t.

Remark: If $Q_l(t) = 0$, then $R_l(t) = 0$ and hence $H_l(t) = 0$. If $Q_l(t) > 0$, then $R_l(t) = 1$ implies that $H_l(t) = 1$. Therefore, we also have $\mathbb{E}\left[H_l(t)|\mathbf{Q}(t)\right] = \sigma_l(\mathbf{Q}(t))$. Since $H_l(t) = \sum_{r:l\in r} D_{l,r}(t)$, we have $\sigma_l(\mathbf{Q}(t)) = \sum_{r:l\in r} \mu_{l,r}(\mathbf{X}(t))$. This combines the fact that $Q_l(t) = \sum_{r:l\in r} X_{l,r}(t)$ and (6), implying that

$$\frac{\mu_{l,r}(\mathbf{X}(t))}{X_{l,r}(t)} = \frac{\sigma_l(\mathbf{Q}(t))}{Q_l(t)}, \text{whenever } X_{l,r}(t) > 0. \tag{7}$$

This, together with (5), yields

$$\frac{\mu_{l,r}(\mathbf{X}(t))}{X_{l,r}(t)} = \frac{\mu_{l',r'}(\mathbf{X}(t))}{X_{l',r'}(t)}, \text{whenever } X_{l,r}(t) > 0 \text{ and } X_{l',r}(t) > 0, \tag{8}$$

where $l \neq l'$ or $r \neq r'$.

Unlike the well-known backpressure algorithm, the QPRA algorithm only requires the per-link queueing information to make transmission decisions and thus significantly simplifies the queueing structure of multihop networks. Here, we want to point out that the QPRA algorithm generalizes the algorithm proposed in [11] under the single-hop traffic model, i.e., packets immediately leave the network once they are served. In particular, the link rate allocation procedure of the QPRA algorithm was proposed in [11], where the authors showed that this link rate allocation scheme can achieve maximum throughput under the single-hop traffic model. To see this, let us consider a network with two links, where the capacity region $\boldsymbol{\Lambda}$ is

shown in Fig. 1. Suppose that the arrival rate vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ is strictly within the capacity region. If the queue-length vector is below the line $Q_1/\lambda_1 = Q_2/\lambda_2$, i.e., $Q_1/\lambda_1 > Q_2/\lambda_2$, then according to the QRPA algorithm, we can observe from Fig. 1 that the service rate of the first link is always greater than its arrival rate (i.e., $\sigma_1 > \lambda_1$) and thus the queue length of the first link tends to decrease. This suggests that the QPRA algorithm always tries to reduce $\max_{l\in\mathcal{L}} Q_l(t)/\lambda_l$ (link 1 achieves $\max_{l\in\mathcal{L}} Q_l(t)/\lambda_l$ in the current example) and hence keeps the mean queue-length finite for any arrival rate vector strictly within the capacity region.

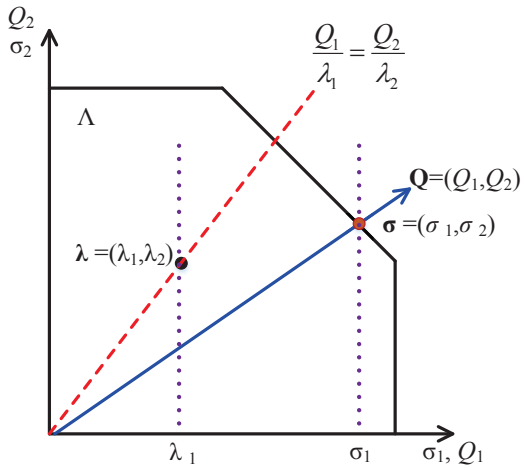

Fig. 1: An example illustrating the throughput optimality of the QPRA algorithm in networks with single-hop traffic.

Similarly, in multihop networks, the QPRA algorithm tends to reduce the maximum of $X_{l,r}/\lambda_r$ over all pairs (l, r). In particular, the following lemma shows that if $\theta(\sum_{r:l\in r}\lambda_r)_{l\in\mathcal{L}} \in \boldsymbol{\Lambda}$ for some $\theta > 0$, and the pair (l^*, r^*) achieves the maximum of $X_{l,r}/\lambda_r$ over all pairs (l, r), then the departure rate of route r^* at link l^* is at least $\theta\lambda_{r^*}$ under the QPRA algorithm.

Lemma 1 *Assume* $\theta(\sum_{r:l\in r}\lambda_r)_{l\in\mathcal{L}} \in \boldsymbol{\Lambda}$ *for some* $\theta > 0$. *If* $\mathbf{X} \neq 0$ *and* $(l^*, r^*) \in \arg\max_{(l,r)} X_{l,r}/\lambda_r$, *then* $\mu_{l^*,r^*}(\mathbf{X}) \geq \theta\lambda_{r^*}$.

Proof The proof generalizes [21, Lemma 5] to the multihop setup. Since $\mathbf{X} \neq 0$, we have $X_{l^*,r^*} > 0$. Assume $\mu_{l^*,r^*}(\mathbf{X}) < \theta\lambda_{r^*}$. Then, for any route r within the link l ($r \neq r^*$ or $l \neq l^*$), there are two cases:

(i) If $X_{l,r} = 0$, then $\mu_{l,r}(\mathbf{X}) = 0$.

(ii) If $X_{l,r} > 0$, then we have

$$\mu_{l,r}(\mathbf{X}) \overset{(a)}{=} \frac{X_{l,r}}{X_{l^*,r^*}}\mu_{l^*,r^*}(\mathbf{X}) = \frac{X_{l,r}/\lambda_r}{X_{l^*,r^*}/\lambda_{r^*}}\frac{\lambda_r}{\lambda_{r^*}}\mu_{l^*,r^*}(\mathbf{X}) \overset{(b)}{\leq} \lambda_r\frac{\mu_{l^*,r^*}(\mathbf{X})}{\lambda_{r^*}} \overset{(c)}{<} \theta\lambda_r, \tag{9}$$

where the step (a) follows from equations (8); (b) is true since $\frac{X_{l^*,r^*}}{\lambda_{r^*}} = \max_{l,r} \frac{X_{l,r}}{\lambda_r}$; (c) follows from the assumption.

Combining (i) and (ii), we have $\mu_{l,r}(\mathbf{X}) < \theta\lambda_r$ for any pair (l,r). Therefore, we have $\sigma_l(\mathbf{Q}) = \sum_{r:l\in r} \mu_{l,r}(\mathbf{X}) < \theta \sum_{r:l\in r} \lambda_r$ for each link $l \in \mathcal{L}$. This contradicts the fact that the allocated service rate vector $(\sigma_l(\mathbf{Q}))_{l\in\mathcal{L}}$ lies on the boundary of the capacity region $\mathbf{\Lambda}$, since $\theta(\sum_{r:l\in r} \lambda_r)_{l\in\mathcal{L}} \in \mathbf{\Lambda}$. Hence, we have the desired result.

However, Lemma 1 does not necessarily imply the stability in a multihop system. Indeed, as we mentioned in the Introduction, the stability in a single-hop system does not necessarily imply stability in a multihop system (e.g., [24,30,4]). Nevertheless, we are able to show that the QPRA algorithm achieves the maximum throughput under the multihop traffic model through fluid limit techniques. Next, we introduce the fluid limit and fluid model equations under the QPRA algorithm.

3.2 Fluid Model

In this subsection, we present the fluid limit and fluid model equations associated with the QPRA algorithm.

Proposition 1 *Under the QPRA algorithm, with probability* 1*, for any positive sequence $w_n \to \infty$, there exists a subsequence w_{n_j} with $w_{n_j} \to \infty$ such that the following convergence holds uniformly over compact intervals of time t:*

$$\frac{1}{w_{n_j}} A^{\Sigma}_{l,r}(w_{n_j}t) \to \lambda_{l,r}t, \quad \forall(l,r), \tag{10}$$

$$\frac{1}{w_{n_j}} H^{\Sigma}_{l}(w_{n_j}t) \to \phi^{\Sigma}_{l}(t), \quad \forall l \in \mathcal{L}, \tag{11}$$

$$\frac{1}{w_{n_j}} D^{\Sigma}_{l,r}(w_{n_j}t) \to \mu^{\Sigma}_{l,r}(t), \quad \forall(l,r), \tag{12}$$

$$\frac{1}{w_{n_j}} X_{l,r}(w_{n_j}t) \to x_{l,r}(t), \quad \forall(l,r), \tag{13}$$

$$\frac{1}{w_{n_j}} Q_{l}(w_{n_j}t) \to q_{l}(t), \quad \forall l \in \mathcal{L}, \tag{14}$$

where the limiting functions $\phi^{\Sigma}_{l}(t), \mu^{\Sigma}_{l,r}(t), x_{l,r}(t), q_l(t)$ are Lipschitz-continuous in $[0,\infty)$, which implies that these limiting functions are differentiable for almost all t. Let $\mathcal{T}$ be the set of time instants where these functions are differentiable. Then, the following equations hold for all $t \in \mathcal{T}$:

$$\frac{d}{dt}\phi^{\Sigma}_{l}(t) = \sigma_l(\mathbf{q}(t)), \quad \forall l \in \mathcal{L}, \quad \textit{whenever } \mathbf{q}(t) \neq 0, \tag{15}$$

$$\frac{d}{dt}\mu^{\Sigma}_{l,r}(t) = \mu_{l,r}(\mathbf{x}(t)), \quad \forall(l,r), \quad \textit{whenever } \mathbf{x}(t) \neq 0, \tag{16}$$

$$\phi^{\Sigma}_{l}(t) = \sum_{r:l\in r} \mu^{\Sigma}_{l,r}(t), \quad \forall(l,r), \tag{17}$$

$$q_l(t) = \sum_{r:l\in r} x_{l,r}(t), \quad \forall(l,r), \tag{18}$$

$$\frac{d}{dt}x_{l,r}(t) = \lambda_{l,r} + \frac{d}{dt}\mu^{\Sigma}_{l^{(r)}_{-},r}(t) - \frac{d}{dt}\mu^{\Sigma}_{l,r}(t), \quad \forall(l,r), \tag{19}$$

where $\frac{d}{dt}\mu^{\Sigma}_{l^{(r)}_{-},r}(t) = 0$ *if* $l = l^{(r)}_1$. *Here* $\boldsymbol{\sigma}(\mathbf{q}(t)) = (\sigma_l(\mathbf{q}(t)))_{l\in\mathcal{L}}$ *and* $\boldsymbol{\mu}(\mathbf{x}(t)) = (\mu_l(\mathbf{x}(t)))_{l\in\mathcal{L},r\in\mathcal{R}}$ *are defined in the QPRA algorithm.*

The proof of Proposition 1 is somewhat standard and technical using the techniques developed in [8,9,21], and is available in Appendix A for completeness. Proposition 1 indicates that the stochastic queueing networks and its fluid model are strongly coupled. In particular, the queue-length dynamics (18) and (19) are equivalent to (2) and (4), respectively. Moreover, equations (15) and (16) follow from the QPRA algorithm. By [8, Theorem 4.2], the stability of the fluid model implies that the original stochastic system is positive recurrent, where the fluid model is *stable* if there exists a finite time $T \geq 0$ such that for every fluid limit model $\mathbf{x} \triangleq (x_{l,r}(t))_{r\in\mathcal{R},l\in\mathcal{L}}$ with $\|\mathbf{x}(0)\| = 1$, $\mathbf{x}(t) = 0$ for all $t \geq T$. Therefore, it is sufficient to show that the fluid model is stable in the rest of the paper.

3.3 Stability of the Fluid Model

In this subsection, we will show that the fluid model is stable under the QPRA algorithm for any arrival rate vector within the maximum throughput region $\overline{\boldsymbol{\Lambda}}$. We first state the following two simple lemmas that will be used for the stability proof.

Lemma 2 *If* $f(x) = \max_{i=1,2,...,K} f_i(x)$ *and* $f_i(x), \forall i$, *are locally Lipschitz continuous*[2], *then, we have*

$$\frac{D^+}{dx^+}f(x) \leq \max_{i\in\mathcal{K}}\left\{\frac{D^+}{dx^+}f_i(x)\right\}, \tag{20}$$

where $\mathcal{K} \triangleq \{i|f_i(x) = f(x)\}$, *and* $\frac{D^+}{dx^+}f(x)$ *is defined as*

$$\frac{D^+}{dx^+}f(x) \triangleq \limsup_{u\downarrow 0}\frac{f(x+u)-f(x)}{u}.$$

This lemma relaxes the assumption of differentiability of $f_i(x), \forall i = 1, 2, ..., K$ in [3, Proposition 2.3.2] to the case when these functions are locally Lipschitiz continuous. The proof is quite similar to that in [3, Proposition 2.3.2] and is available in Appendix B for completeness.

Lemma 3 *Let* $g : [0,\infty) \to [0,\infty)$ *be a locally Lipschitz continuous function.*

(i) Assume that $g(0) = 0$ *and* $\frac{D^+}{dt^+}g(t) \leq 0$ *whenever* $g(t) > 0$. *Then,* $g(t) = 0$ *for all* $t \geq 0$*;*

(ii) Assume that $g(0) > 0$ *and* $\frac{D^+}{dt^+}g(t) \leq -\gamma$ *for some* $\gamma > 0$ *whenever* $g(t) > 0$. *Then, there exists a* $T \geq 0$ *such that* $g(t) = 0$ *for all* $t \geq T$.

The proof of Lemma 3 is available in Appendix C.

We are ready to prove that the fluid model is weakly stable for any arrival rate vector within the maximum throughput region $\overline{\boldsymbol{\Lambda}}$.

[2] Locally Lipschitz continuity guarantees the existences of $\frac{D^+}{dx^+}f(x)$ and $\frac{D^+}{dx^+}f_i(x)$, $i = 1, 2, ..., K$ (see, for example, [7]).

Proposition 2 *The QPRA algorithm achieves throughput optimality, i.e., it stabilizes the system for any arrival rate vector that is in the interior of the maximum throughput region* $\overline{\mathbf{\Lambda}}$.

Proof Given any arrival rate vector $\boldsymbol{\lambda} \in \mathrm{Int}(\overline{\mathbf{\Lambda}})$, there always exists a real number $\epsilon \in (0,1)$ such that

$$(1+\epsilon)\left(\sum_{r:l\in r}\lambda_r\right)_{l\in\mathcal{L}} \in \mathbf{\Lambda}. \tag{21}$$

Consider the Lyapunov function

$$V(\mathbf{x}(t)) = \max_{(l,r)} \frac{a^{N_r-n_l}}{\lambda_r} x_{l,r}(t), \tag{22}$$

where $a>1$ is some parameter, N_r is the number of links belonging to route r, and n_l is the index of link l at route r, $\forall n_l = 1,2,\ldots,N_r$.

We would like to show that there exists a constant $\delta>0$ such that $\frac{D^+}{dt^+}V(\mathbf{x}(t)) \le -\delta$ whenever $V(\mathbf{x}(t))>0$, which implies the desired result according to Lemma 3 and [8, Theorem 4.2].

According to Lemma 2, we have

$$\frac{D^+}{dt^+}V(\mathbf{x}(t)) \le \max_{(\bar{l},\bar{r})\in\overline{\mathcal{K}}} \frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\frac{d}{dt}x_{\bar{l},\bar{r}}(t),$$

where

$$\overline{\mathcal{K}} \triangleq \left\{(\bar{l},\bar{r}) : V(\mathbf{x}(t)) = \frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}} x_{\bar{l},\bar{r}}(t)\right\}.$$

In the rest of the proof, we will omit the time index t for brevity. We consider the case when $V(\mathbf{x})>0$, i.e., $x_{\bar{l},\bar{r}}>0$. Let $(l^*,r^*)\in \arg\max_{(l,r)} x_{l,r}/\lambda_r$. Then, we have

$$\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}x_{\bar{l},\bar{r}} \overset{(a)}{\ge} \frac{a^{N_{r^*}-n_{l^*}}}{\lambda_{r^*}}x_{l^*,r^*} \overset{(b)}{\ge} \frac{x_{l^*,r^*}}{\lambda_{r^*}}, \tag{23}$$

where step (a) follows the definition of $(\bar{l},\bar{r})$ and (b) uses the fact that $a>1$ and the fact that $1\le n_{l^*}\le N_{r^*}$.

Hence, under QPRA policy, we have

$$\begin{aligned}
\mu_{\bar{l},\bar{r}} &= \frac{x_{\bar{l},\bar{r}}}{x_{l^*,r^*}}\mu_{l^*,r^*} \\
&= \frac{x_{\bar{l},\bar{r}}/\lambda_{\bar{r}}}{x_{l^*,r^*}/\lambda_{r^*}}\frac{\lambda_{\bar{r}}}{\lambda_{r^*}}\mu_{l^*,r^*} \\
&\overset{(a)}{\ge} \frac{1}{a^{N_{\bar{r}}-n_{\bar{l}}}}\frac{\lambda_{\bar{r}}}{\lambda_{r^*}}\mu_{l^*,r^*} \\
&\overset{(b)}{\ge} \frac{1}{a^{N_{\bar{r}}-n_{\bar{l}}}}\lambda_{\bar{r}}(1+\epsilon),
\end{aligned} \tag{24}$$

where step (a) uses inequality (23); (b) follows form Lemma 1.

(i) If $\bar{l}$ is the ingress link for route $\bar{r}$ (i.e., $n_{\bar{l}} = 1$), then

$$\begin{aligned}\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\frac{d}{dt}x_{\bar{l},\bar{r}} =& \frac{a^{N_{\bar{r}}-1}}{\lambda_{\bar{r}}}\left(\lambda_{\bar{r}} - \mu_{\bar{l},\bar{r}}\right)\\ \overset{(a)}{\leq}& a^{N_{\bar{r}}-1}\left(1 - \frac{1}{a^{N_{\bar{r}}-1}}(1+\epsilon)\right)\\ =& a^{N_{\bar{r}}-1} - (1+\epsilon)\\ \overset{(b)}{\leq}& a^{N_{\max}-1} - (1+\epsilon),\end{aligned}$$

where step (a) uses inequality (24); (b) is true for $N_{\max} \triangleq \max_{r\in\mathcal{R}} N_r \geq 1$.

(ii) If $\bar{l}$ is not the ingress link for route $\bar{r}$ (i.e., $n_{\bar{l}} \geq 2$), then

$$\begin{aligned}\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\frac{d}{dt}x_{\bar{l},\bar{r}} =& \frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\left(\mu_{\bar{l}-,\bar{r}} - \mu_{\bar{l},\bar{r}}\right)\\ \overset{(a)}{=}& \frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\left(\frac{x_{\bar{l}-,\bar{r}}}{x_{\bar{l},\bar{r}}} - 1\right)\mu_{\bar{l},\bar{r}}\\ \overset{(b)}{\leq}& \frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\left(\frac{1}{a} - 1\right)\mu_{\bar{l},\bar{r}}\\ \overset{(c)}{\leq}& \left(\frac{1}{a} - 1\right)(1+\epsilon)\,,\end{aligned}$$

where step (a) follows from the definition of QPRA policy; (b) uses the definition of $(\bar{l},\bar{r})$ (i.e., $a^{N_{\bar{r}}-n_{\bar{l}}}x_{\bar{l},\bar{r}}/\lambda_{\bar{r}} \geq a^{N_{\bar{r}}-(n_{\bar{l}}-1)}x_{\bar{l}-,\bar{r}}/\lambda_{\bar{r}}$); (c) utilizes inequality (24) and the fact that $a > 1$.

We can then select $a > 1$ satisfying $a^{N_{\max}-1} < (1+\epsilon)$ such that $\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\frac{d}{dt}x_{\bar{l},\bar{r}}$ has a negative drift in both cases, which implies that $\frac{D^+}{dt^+}V(\mathbf{x}(t)) < -\delta$ for some $\delta > 0$.

Even though the QPRA algorithm is throughput-optimal, it requires the knowledge of the full capacity region, which is generally unavailable in practice. Motivated by [21,16] and the QPRA Algorithm, we next propose an efficient low-complexity scheduling scheme that makes transmission decisions only based on per-link queueing information in multihop wireless networks.

4 Low-Complexity Implementation

In this section, we propose a low-complexity scheduling algorithm with only per-link queueing information.

We consider the primary interference model, where each link l interferes with all of its one-hop neighboring links. The capacity region $\mathbf{\Lambda}$ (see [12]) under the primary interference model is bounded by $\frac{2}{3}\mathbf{\Gamma} \subseteq \mathbf{\Lambda} \subseteq \mathbf{\Gamma}$, where

$$\mathbf{\Gamma} \triangleq \left\{\boldsymbol{\eta} = (\eta_l)_{l\in\mathcal{L}} \middle| \sum_{l\in\mathcal{E}(i)} \eta_l \leq 1, \forall i \in \mathcal{N}\right\}, \tag{25}$$

and $\mathcal{E}(i)$ denotes the set of links that connect to the neighbors of node i.

Motivated by the QPRA algorithm and [21], for any queue-length vector $\mathbf{Q} \neq 0$, let $\boldsymbol{\sigma} = (\sigma_l)_{l\in\mathcal{L}}$ be the longest vector in $\mathbf{\Gamma}$ such that $\sigma_l = \gamma Q_l, \forall l \in \mathcal{L}$, for some $\gamma \geq 0$. Therefore, it can be easily calculated that

$$\sigma_l = \frac{Q_l}{\max_{i\in\mathcal{N}} \sum_{l'\in\mathcal{E}(i)} Q_{l'}}, \quad \forall l \in \mathcal{L}. \tag{26}$$

Based on (26), we are able to develop a low-complexity scheduling algorithm with only per-link queueing information. We divide each time slot into a control slot and a data slot, where we further subdivide the control slot into M mini-slots[3]. The main purpose of the control slot is to determine the transmission schedule used for data transmission in the data slot.

Low-Complexity QPRA (LC-QPRA):
(1) In each mini-slot of the time slot t, each link l attempts transmission with probability $\sigma_l(t)(\sqrt{M}-1)/(2M)$, where $\sigma_l(t)$ is defined as follows:

$$\sigma_l(t) = \begin{cases} 0, & \text{if } Q_l(t) = 0; \\ \dfrac{Q_l(t)}{\max_{i\in\mathcal{N}} \sum_{k\in\mathcal{E}(i)} Q_k(t)}, & \text{if } Q_l(t) > 0. \end{cases} \tag{27}$$

(2) If the link l attempts transmission in the current mini-slot and does not overhear its neighbors' transmissions, then it will start transmission in the data slot. Otherwise, it repeats steps (1) and (2) until the end of the control slot.
(3) In the data slot, serve each link l according to the SIRO queueing discipline as in the QPRA algorithm.

Remarks: (1) Here we need to know $\max_{i\in\mathcal{N}} \sum_{k\in\mathcal{E}(i)} Q_k(t)$ in each time slot. However, since the throughput performance is insensitive to the accuracy of the queue-lengths (e.g., [38,17,31]), we may just need to update the attempt probability over a relatively long period so that all links have sufficient time to learn this global information.
(2) The proposed LC-QPRA algorithm can be easily extended to the two-hop interference model (see [21,16]), where each link l interferes with all of its two-hop neighboring links.

The next lemma provides the successful transmission probability of each link l in each time slot under the LC-QPRA algorithm.

Lemma 4 *Under the LC-QPRA algorithm, the successful transmission probability of link l in time slot t is at least $\sigma(t)\left(1/2 - 1/\sqrt{M}\right)$.*

The proof is available in [16, Proposition 3].

Proposition 3 *The LC-QPRA Algorithm can achieve a constant efficient ratio of at least ρ under the primary interference model, i.e., it stabilizes the system for any arrival rate vector that is strictly within the region $\rho^2\overline{\Lambda}$, where $\rho \triangleq 1/2 - 1/\sqrt{M}$.*

[3] In IEEE 802.11b standard, the total number of mini-slots M ranges between 32 and 1024, where each mini-slot lasts $20\mu s$.

Proof Given any arrival rate vector $\boldsymbol{\lambda} \in \text{Int}(\rho\overline{\boldsymbol{\Lambda}})$, there always exists a real number $\epsilon \in (0, 1/2)$ such that

$$\frac{1+\epsilon}{\rho^2}\left(\sum_{r:l\in r}\lambda_r\right)_{l\in\mathcal{L}} \in \boldsymbol{\Lambda}.$$

The fluid model under the LC-QPRA algorithm is the same as that in Proposition 1 except the departure rate vector $\boldsymbol{\mu}(t) = (\mu_{l,r}(t))_{r\in\mathcal{R},l\in\mathcal{L}}$, which satisfies

$$\mu_{l,r}(t) = \frac{x_{l,r}(t)}{q_l(t)}\nu_l(t), \tag{28}$$

$$\sigma_l(t) \geq \nu_l(t) \geq \rho\sigma_l(t), \tag{29}$$

where $\nu_l(t)$ is the service rate for link l at time t and $\boldsymbol{\sigma}(t) = (\sigma_l(t))_{l\in\mathcal{L}}$ lies on the boundary of the region $\boldsymbol{\Gamma}$ satisfying

$$\sigma_l(t) = 0, \text{ whenever } q_l(t) = 0; \tag{30}$$

$$\sigma_l(t) = \frac{q_l(t)}{\max_{i\in\mathcal{N}}\sum_{l'\in\mathcal{E}(i)} q_{l'}(t)}, \text{ whenever } q_l(t) > 0, \tag{31}$$

Here we note that equation (29) follows from Lemma 4, and (28), (30) and (31) are from the LC-QPRA algorithm.

We use the same Lyapunov function $V(\mathbf{x}(t))$ as in the proof of Proposition 2 except selecting $a > 1/\rho$, i.e.,

$$V(\mathbf{x}(t)) = \max_{(l,r)} \frac{a^{N_r - n_l}}{\lambda_r} x_{l,r}(t), \tag{32}$$

where $a > 1/\rho$ is some parameter, N_r is the number of links belonging to route r, and n_l is the index of link l at route r, $\forall n_l = 1, 2, \ldots, N_r$. We would like to show that there exists a constant $\delta > 0$ such that $\frac{D^+}{dt^+}V(\mathbf{x}(t)) \leq -\delta$ whenever $V(\mathbf{x}(t)) > 0$, which implies the desired result according to Lemma 3 and [8, Theorem 4.2].

According to Lemma 2, we have

$$\frac{D^+}{dt^+}V(\mathbf{x}(t)) \leq \max_{(\bar{l},\bar{r})\in\overline{\mathcal{K}}} \frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\frac{d}{dt}x_{\bar{l},\bar{r}}(t),$$

where

$$\overline{\mathcal{K}} \triangleq \left\{(\bar{l},\bar{r}) : V(\mathbf{x}(t)) = \frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}x_{\bar{l},\bar{r}}(t)\right\}.$$

In the rest of the proof, we will omit the time index t for brevity. We consider the case when $V(\mathbf{x}) > 0$, i.e., $x_{\bar{l},\bar{r}} > 0$. Let $(l^*, r^*) \in \arg\max_{(l,r)} x_{l,r}/\lambda_r$. Then, we have

$$\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}x_{\bar{l},\bar{r}} \overset{(a)}{\geq} \frac{a^{N_{r^*}-n_{l^*}}}{\lambda_{r^*}}x_{l^*,r^*} \overset{(b)}{\geq} \frac{x_{l^*,r^*}}{\lambda_{r^*}}, \tag{33}$$

where step (a) follows from the definition of $(\bar{l},\bar{r})$ and (b) uses the fact that $a > 1/\rho > 1$ and the fact that $1 \leq n_{l^*} \leq N_{r^*}$.

Under the LC-QPRA algorithm, we have

$$\begin{aligned}\mu_{\bar{l},\bar{r}} \overset{(a)}{=} \frac{x_{\bar{l},\bar{r}}}{q_{\bar{l}}}\nu_{\bar{l}} \overset{(b)}{\geq} \rho\frac{x_{\bar{l},\bar{r}}}{q_{\bar{l}}}\sigma_{\bar{l}} \overset{(c)}{=} &\rho x_{\bar{l},\bar{r}}\frac{\sigma_{l^*}}{q_{l^*}} \\ \overset{(d)}{\geq} &\rho x_{\bar{l},\bar{r}}\frac{\nu_{l^*}}{q_{l^*}} \\ \overset{(e)}{=} &\rho x_{\bar{l},\bar{r}}\frac{\mu_{l^*,r^*}}{x_{l^*,r*}} \\ = &\rho\frac{x_{\bar{l},\bar{r}}/\lambda_{\bar{r}}}{x_{l^*,r^*}/\lambda_{r^*}}\frac{\lambda_{\bar{r}}}{\lambda_{r^*}}\mu_{l^*,r^*} \\ \overset{(f)}{\geq} &\rho\frac{1}{a^{N_{\bar{r}}-n_{\bar{l}}}}\frac{\lambda_{\bar{r}}}{\lambda_{r^*}}(1+\epsilon)\frac{1}{\rho^2}\lambda_{r^*} \\ = &\frac{1}{a^{N_{\bar{r}}-n_{\bar{l}}}}(1+\epsilon)\frac{1}{\rho}\lambda_{\bar{r}}, \end{aligned} \tag{34}$$

where the steps (a) and (e) use equation (28); (b) and (d) follow from inequality (29); (c) utilizes equation (31); (f) utilizes inequality (33) and Lemma 1.

(i) If $\bar{l}$ is the ingress link for route $\bar{r}$ (i.e., $n_{\bar{l}} = 1$), then

$$\begin{aligned}\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\frac{d}{dt}x_{\bar{l},\bar{r}} = &\frac{a^{N_{\bar{r}}-1}}{\lambda_{\bar{r}}}\left(\lambda_{\bar{r}} - \mu_{\bar{l},\bar{r}}\right) \\ \overset{(a)}{\leq} &a^{N_{\bar{r}}-1}\left(1 - \frac{1}{a^{N_{\bar{r}}-1}}\frac{1}{\rho}(1+\epsilon)\right) \\ = &a^{N_{\bar{r}}-1} - \frac{1}{\rho}(1+\epsilon) \\ \overset{(b)}{\leq} &a^{N_{\max}-1} - \frac{1}{\rho}(1+\epsilon), \end{aligned}$$

where the step (a) uses inequality (34); (b) is true for $N_{\max} \triangleq \max_{r\in\mathcal{R}} N_r \geq 1$.

(ii) If $\bar{l}$ is not the ingress link for route $\bar{r}$ (i.e., $n_{\bar{l}} \geq 2$), then

$$\begin{aligned}\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\frac{d}{dt}x_{\bar{l},\bar{r}} = &\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\left(\mu_{\bar{l}-,\bar{r}} - \mu_{\bar{l},\bar{r}}\right) \\ \overset{(a)}{=} &\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\left(\frac{x_{\bar{l}-,\bar{r}}}{q_{\bar{l}-}}\nu_{\bar{l}-} - \frac{x_{\bar{l},\bar{r}}}{q_{\bar{l}}}\nu_{\bar{l}}\right) \\ \overset{(b)}{\leq} &\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\left(\frac{x_{\bar{l}-,\bar{r}}}{q_{\bar{l}-}}\sigma_{\bar{l}-} - \frac{x_{\bar{l},\bar{r}}}{q_{\bar{l}}}\rho\sigma_{\bar{l}}\right) \\ = &\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\left(x_{\bar{l}-,\bar{r}} - \rho x_{\bar{l},\bar{r}}\right)\frac{\sigma_{\bar{l}}}{q_{\bar{l}}} \\ = &\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\left(\frac{x_{\bar{l}-,\bar{r}}}{x_{\bar{l},\bar{r}}} - \rho\right)\mu_{\bar{l},\bar{r}} \\ \overset{(c)}{\leq} &\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\left(\frac{1}{a} - \rho\right)\mu_{\bar{l},\bar{r}} \\ \overset{(d)}{\leq} &\left(\frac{1}{a\rho} - 1\right)(1+\epsilon), \end{aligned}$$

where the step (a) follows from the definition of QPRA policy; (b) uses inequality (29); (c) uses the definition of $(\bar{l},\bar{r})$ (i.e., $a^{N_{\bar{r}}-n_{\bar{l}}}x_{\bar{l},\bar{r}}/\lambda_{\bar{r}} \geq a^{N_{\bar{r}}-(n_{\bar{l}}-1)}x_{\bar{l}-,\bar{r}}/\lambda_{\bar{r}}$); (d) utilizes inequality (34) and the fact that $a > 1$.

We can then select $a > 1/\rho$ satisfying $a^{N_{\max}-1} < (1+\epsilon)/\rho$ such that $\frac{a^{N_{\bar{r}}-n_{\bar{l}}}}{\lambda_{\bar{r}}}\frac{d}{dt}x_{\bar{l},\bar{r}}$ has a negative drift in both cases, which implies that $\frac{D^+}{dt^+}V(\mathbf{x}(t)) < -\delta$ for some $\delta > 0$.

5 Conclusion

In this work, we first developed a scheduling algorithm that makes transmission decisions only based on the per-link queueing information, which significantly reduces the queueing complexity compared to the well-known backpressure algorithm. By introducing a novel Lyapunov function, we are able to establish the throughput optimality of our proposed algorithm through fluid limit arguments. We further proposed a low-complexity scheduling algorithm that approximates our proposed algorithm and showed that it achieves a constant fraction of the maximum throughput region, independent of network size.

Acknowledgement

We thank Prof. Jim Dai for providing us the key idea behind the proof of Proposition 2. In particular, he showed that our algorithm is stable for the special case of a network with one flow over two links. The Lyapunov function in (22) is a generalization of the one he used for that network.

References

1. M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting. Scheduling in a queuing system with asynchronously varying service rates. *Probability in the Engineering and Informational Sciences*, 18(02):191–217, 2004.
2. T. Bonald and A. Proutiere. Insensitive bandwidth sharing in data networks. *Queueing Systems*, 44:69–100, 2003.
3. J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2000.
4. M. Bramson. Stability of queueing networks. *Probability Surveys*, 5:169–345, 2008.
5. M. Chiang. Balancing transport and physical layers in wireless multihop networks: Jointly optimal congestion control and power control. *IEEE Journal on Selected Areas in Communications, special issue on Nonlinear Optimization of Communication Systems*, 23(1):104 – 116, January 2005.
6. Y. S. Chow. On a strong law of large numbers for martingales. *The Annals of Mathematical Statistics*, 38:610, 1967.
7. F. Clarke. *Optimization and Nonsmooth Analysis (Classics in Applied Mathematics)*. Society for Industrial Mathematics, 1987.
8. J. G. Dai. On positive harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability*, 5:49–77, 1995.
9. J.G. Dai and B. Prabhakar. The throughput of data switches with and without speedup. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Tel Aviv, Israel, March 2000.
10. A. Eryilmaz and R. Srikant. Joint congestion control, routing and MAC for stability and fairness in wireless networks. *IEEE Journal on Selected Areas in Communications, special issue on Nonlinear Optimization of Communication Systems*, 14:1514–1524, August 2006.

11. A. Eryilmaz, R. Srikant, and J. R. Perkins. Throughput-optimal scheduling for broadcast channels. In *Proceedings of ITCom (Modeling and Design of Wireless Networks)*, Denver, CO, August 2001.
12. B. Hajek and G. Sasaki. Link scheduling in polynomial time. *IEEE/ACM Transactions on Information Theory*, 34(5):910–917, September 1988.
13. I. H. Hou, V. Borkar, and P. R. Kumar. A theory of QoS for wireless. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Rio de Janeiro, Brazil, April 2009.
14. J.J. Jaramillo and R. Srikant. Optimal scheduling for fair resource allocation in Ad hoc networks with elastic and inelastic traffic. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, San Diego, CA, March 2010.
15. B. Ji, C. Joo, and N. B. Shroff. Throughput-optimal scheduling in multihop wireless networks without per-flow information. *IEEE/ACM Transactions on Networking*, 21(2):634–647, April 2013.
16. C. Joo and N. B. Shroff. Performance of random access scheduling schemes in multi-hop wireless networks. *IEEE/ACM Transactions on Networking*, 17(5):1481–1493, October 2009.
17. K. Kar, X. Luo, and S. Sarkar. Throughput-optimal scheduling in multichannel access point networks under infrequent channel measurements. *IEEE Transactions on Wireless Communications*, 7:2619 – 2629, 2008.
18. B. Li, R. Li, and A. Eryilmaz. Throughput-optimal scheduling design with regular service guarantees in wireless networks. *To appear in IEEE/ACM Transactions on Networking.*
19. B. Li, R. Li, and A. Eryilmaz. On the optimal convergence speed of wireless scheduling for fair resource allocation. *IEEE/ACM Transactions on Networking*, 23(2):631–643, 2015.
20. B. Li and R. Srikant. Queue-proportional rate allocation with per-link information in multihop networks. In *Proc. ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, Portland, Oregon, USA, June 2015.
21. X. Lin and S. B. Rasool. Constant-time distributed scheduling policies for ad hoc wireless networks. *IEEE Transactions on Automatic Control*, 54(2):231–242, February 2009.
22. X. Lin and N. Shroff. The impact of imperfect scheduling on cross-layer rate control in multihop wireless networks. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Miami, FL, March 2005.
23. S. Liu, E. Ekici, and L. Ying. Scheduling in multihop wireless networks without back-pressure. In *Proceedings of the Allerton Conference on Communications, Control and Computing*, Monticello, IL, October 2010.
24. S. H. Lu and P. R. Kumar. Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control*, 36(12):1406–1416, December 1991.
25. L. Massoulie and J. Roberts. Bandwidth sharing: Objectives and algorithms. *IEEE/ACM Transactions on Networking*, 10:320–328, June 2002.
26. J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8(5):556–567, October 2000.
27. M. J. Neely. Energy optimal control for time varying wireless networks. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Miami, FL, March 2005.
28. M.J. Neely, E. Modiano, and C. Li. Fairness and optimal stochastic control for heterogeneous networks. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Miami, FL, March 2005.
29. M.J. Neely, E. Modiano, and C.E. Rohrs. Dynamic power allocation and routing for time varying wireless networks. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, San Francisco, CA, April 2003.
30. A. Rybko and A. Stolyar. Ergodicity of stochastic processes describing the operation of open queueing networks. *Problemy Peredachi Informatsii*, 28(3):3–26, 1992.
31. D. Shah and J. Shin. Randomized scheduling algorithm for queueing networks. *The Annals of Applied Probability*, 22(1):128–171, 2012.
32. D. Shah, N. Walton, and Y. Zhong. Optimal queue-size scaling in switched networks. In *Proc. ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, London, United Kingdom, June 2012.
33. R. Srikant and L. Ying. *Communication Networks: An Optimization, Control, and Stochastic Networks Perspective.* Cambridge University Press, 2014.
34. A. Stolyar. Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm. *Queueing Systems*, 50(4):401–457, August 2005.

35. L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 36:1936–1948, December 1992.
36. N. Walton. Concave switching in single and multihop networks. In *Proc. ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, Austin, Texas, June 2014.
37. H. Wu, X. Lin, X. Liu, and Y. Zhang. Application-level scheduling with deadline constraints. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Toronto, Canada, April 2014.
38. L. Ying, R. Srikant, A. Eryilmaz, and G. E. Dullerud. Distributed fair resource allocation in cellular networks in the presence of heterogeneous delays. *IEEE Transactions on Automatic Control*, 52:129–134, January 2007.

A Proof of Proposition 1

For any integer $t \geq 1$, let $H_l^{\Sigma}(t) \triangleq \sum_{\tau=0}^{t-1} H_l(\tau)$ be the cumulative number of packets departing from link l up to time slot $t-1$. Let $D_{l,r}^{\Sigma}(t) \triangleq \sum_{\tau=0}^{t-1} D_{l,r}(\tau)$ denote the cumulative number of route r packets departing from link l up to time slot $t-1$. Further, let $H_l^{\Sigma}(0) = D_{l,r}^{\Sigma}(0) = 0$. Then, the evolution of the queue length can be rewritten as

$$X_{l,r}(t) = X_{l,r}(0) + A_{l,r}^{\Sigma}(t) + D_{l_-^{(r)},r}^{\Sigma}(t) - D_{l,r}^{\Sigma}(t), \tag{35}$$

for all pairs (l, r). Here we note that $D_{l_-^{(r)},r}^{\Sigma}(t) = 0$ if $l = l_1^{(r)}$.

For the purposes of our analysis, we interpolate the values of $A_{l,r}^{\Sigma}(t)$, $H_l^{\Sigma}(t)$ and $D_{l,r}^{\Sigma}(t)$ to all real number $t \geq 0$ by linear interpolation between $\lfloor t \rfloor$ and $\lfloor t \rfloor + 1$, where $\lfloor t \rfloor$ denotes the largest integer no greater than t. Then, we have the following lemma.

Proof of Lemma ??: Equation (10) follows from the Functional Strong Law of Large Numbers if $l = l_1^{(r)}$. If $l \neq l_1^{(r)}$, then $A_{l,r}^{\Sigma}(w_{n_j} t) = \lambda_{l,r} = 0$ and thus equation (10) always holds.

Note that for any $0 \leq t_1 \leq t_2$, we have

$$0 \leq \frac{1}{w_n} H_l^{\Sigma}(w_n t_2) - \frac{1}{w_n} H_l^{\Sigma}(w_n t_1) \leq t_2 - t_1, \tag{36}$$

where we use the fact that each link l can at most transfer one packet in one time slot. Thus, the sequence of functions $\{\frac{1}{w_n} H_l^{\Sigma}(w_n t)\}$ is uniformly equicontinuous, and since $H_l^{\Sigma}(0) = 0$, the sequence is also uniformly bounded. Similarly, the sequence $\{\frac{1}{w_n} D_{l,r}^{\Sigma}(w_n t)\}$ is uniformly bounded and uniformly equicontinuous. Consequently, according to the Arzela-Ascoli Theorem, there must exist a subsequence of $\{w_n\}_{n \geq 1}$ for which both (11) and (12) hold. Since

$$\frac{1}{w_n} X_{l,r}(w_n t) = \frac{1}{w_n} A_{l,r}^{\Sigma}(w_n t) - \frac{1}{w_n} D_{l_-^{(r)},r}^{\Sigma}(w_n t) + \frac{1}{w_n} D_{l,r}^{\Sigma}(w_n t),$$
$$\frac{1}{w_n} Q_l^{\Sigma}(w_n t) = \sum_{r: l \in r} \frac{1}{w_n} X_{l,r}^{\Sigma}(w_n t),$$

we have (13) and (14) by taking limits as $w_{n_j} \to \infty$.

Since the functions $H_l^{\Sigma}(t)$, $D_{l,r}^{\Sigma}(t)$, $X_{l,r}(t)$, $Q_l(t)$ are Lipschitz continuous, the Lipschitz continuity of $\phi_l^{\Sigma}(t)$, $\mu_{l,r}^{\Sigma}(t)$, $x_{l,r}(t)$, $q_l(t)$ also follows. Hence, these limiting functions are differentiable for almost all t. In the rest of proof, we consider all $t \in \mathcal{T}$, where $\mathcal{T}$ is the set of time instants where the limiting functions are differentiable.

Next, we prove equation (15). Since $\sigma_l(\mathbf{q}(t))$ is continuous with respect to $\mathbf{q}(t)$ when $\mathbf{q}(t) \neq 0$. Therefore, for any $\epsilon > 0$, there exists a $u > 0$ such that for all $t' \in [t, t+u]$, we have

$$\left|\sigma_l(\mathbf{q}(t')) - \sigma_l(\mathbf{q}(t))\right| \leq \epsilon. \tag{37}$$

Since $\frac{1}{w_{n_j}}\mathbf{Q}(\lfloor w_{n_j}t'\rfloor) \to \mathbf{q}(t')$ uniformly over compact intervals of time, and $\sigma_l(a\mathbf{Q}) = \sigma_l(\mathbf{Q})$ for any $a > 0$, we have $\sigma_l(\mathbf{Q}(\lfloor w_{n_j}t'\rfloor)) \to \sigma_l(\mathbf{q}(t'))$ with probability 1. Thus, there exists an integer $J > 0$ such that for all $j > J$ and $t' \in [t, t+u]$,

$$\sigma_l(\mathbf{q}(t')) - \epsilon \le \sigma_l(\mathbf{Q}(\lfloor w_{n_j}t'\rfloor)) \le \sigma_l(\mathbf{q}(t')) + \epsilon. \tag{38}$$

Combining with (37), we have

$$\sigma_l(\mathbf{q}(t)) - 2\epsilon \le \sigma_l(\mathbf{Q}(\lfloor w_{n_j}t'\rfloor)) \le \sigma_l(\mathbf{q}(t)) + 2\epsilon. \tag{39}$$

By the definition of the limit in (11), for any $t' \in [t, t+u]$, we have

$$\phi_l^{\Sigma}(t') - \phi_l^{\Sigma}(t) = \lim_{j\to\infty} \frac{1}{w_{n_j}} \sum_{k=\lfloor w_{n_j}t\rfloor}^{\lfloor w_{n_j}t'\rfloor - 1} H_l(k). \tag{40}$$

Define the filtration $\mathcal{F}_k, k = 1, 2, \cdots$, where $\mathcal{F}_k$ is the $\sigma-$algebra generated by the random variables $A_{l,r}^{\Sigma}(\lfloor w_{n_j}t\rfloor + k')$, $D_{l,r}^{\Sigma}(\lfloor w_{n_j}t\rfloor + k')$, $X_{l,r}(\lfloor w_{n_j}t\rfloor + k')$ for all pairs (l, r) and for $k' = 0, 1, 2, \cdots, k-1$. Let

$$Y_k = H_l(\lfloor w_{n_j}t\rfloor + k) - \mathbb{E}\left[H_l(\lfloor w_{n_j}t\rfloor + k)\middle|\mathbf{Q}(\lfloor w_{n_j}t\rfloor + k)\right].$$

Therefore, $\sum_{m=0}^{k-1} Y_m, k = 1, 2, \cdots$, is a martingale with respect to the filtration $\mathcal{F}_k, k = 1, 2, \cdots$. Further, $\mathbb{E}\left[Y_k^2\right]$ is always bounded for all k. Hence, using a strong law of large numbers for martingales [6], we have

$$\lim_{k\to\infty} \frac{1}{k} \sum_{m=0}^{k-1} Y_m = 0, \text{ with probability 1,} \tag{41}$$

which implies that

$$\begin{aligned}
\phi_l^{\Sigma}(t') - \phi_l^{\Sigma}(t) &= \lim_{j\to\infty} \frac{1}{w_{n_j}} \sum_{k=\lfloor w_{n_j}t\rfloor}^{\lfloor w_{n_j}t'\rfloor - 1} H_l(k) \\
&= \lim_{j\to\infty} \frac{1}{w_{n_j}} \sum_{k=\lfloor w_{n_j}t\rfloor}^{\lfloor w_{n_j}t'\rfloor - 1} \mathbb{E}\left[H_l(k)|\mathbf{Q}(k)\right] \\
&= \lim_{j\to\infty} \frac{1}{w_{n_j}} \sum_{k=\lfloor w_{n_j}t\rfloor}^{\lfloor w_{n_j}t'\rfloor - 1} \sigma_l\left(\mathbf{Q}(k)\right).
\end{aligned} \tag{42}$$

By using (39), we have

$$(t'-t)(\sigma_l(\mathbf{q}(t)) - 2\epsilon) \le \phi_l^{\Sigma}(t') - \phi_l^{\Sigma}(t) \le (t'-t)(\sigma_l(\mathbf{q}(t)) + 2\epsilon),$$

for $t' \in [t, t+u]$. Since we assume that $\sigma_l^{\Sigma}(t)$ is differentiable at t (i.e., $t \in \mathcal{T}$), we have

$$\sigma_l(\mathbf{q}(t)) - 2\epsilon \le \frac{d}{dt}\sigma_l^{\Sigma}(t) \le \sigma_l(\mathbf{q}(t)) + 2\epsilon. \tag{43}$$

Finally, since this is true for any $\epsilon > 0$, equation (15) follows. The proof of equation (16) follows the similar technique and is omitted here for brevity.

Equations (17) and (18) follow from the equation $H_l^{\Sigma}(t) = \sum_{r:l\in r} D_{l,r}^{\Sigma}(t)$ and $Q_l(t) = \sum_{r:l\in r} X_{l,r}(t)$ by taking limits as $w_{n_j} \to \infty$, respectively. Finally, by using the queue-evolution equation (35) and taking limits as $w_{n_j} \to \infty$, we have (19).

B Proof of Lemma 1

We prove it by contradiction. Assume

$$\frac{D^+}{dx^+}f(x) > \max_{i\in\mathcal{K}}\left\{\frac{D^+}{dx^+}f_i(x)\right\}. \tag{44}$$

Then, for a sufficient small $\epsilon > 0$, there exist a decreasing sequence $\{u_k, k = 1, 2, ...\}$ with $\lim_{k\to\infty} u_k = 0$ such that

$$\frac{f(x+u_k)-f(x)}{u_k} \geq \max_{i\in\mathcal{K}}\left\{\frac{D^+}{dx^+}f_i(x)\right\} + \epsilon, \forall k = 1, 2, \cdots.$$

Note that $f(x) = f_i(x), \forall i \in \mathcal{K}$. Since there are a finite number of locally Lipschitz continuous functions $f_i(x), i = 1, 2, \cdots, K$, there must exist a $j \in \mathcal{K}$ and a decreasing subsequence $\{u_{t_k}, k = 1, 2, \cdots\}$ of $\{u_k, k = 1, 2, \cdots\}$ such that $f_j(x + u_{t_k}) = f(x + u_{t_k}) = \max_{i=1,2,\cdots,K} f_i(x + u_{t_k}), \forall k = 1, 2, \cdots$, which implies that

$$\frac{f_j(x+u_{t_k})-f_j(x)}{u_{t_k}} \geq \max_{i\in\mathcal{K}}\left\{\frac{D^+}{dx^+}f_i(x)\right\} + \epsilon, \forall k = 1, 2, \cdots.$$

Therefore, we obtain the contradiction

$$\frac{D^+}{dx^+}f_j(x) \geq \max_{i\in\mathcal{K}}\left\{\frac{D^+}{dx^+}f_i(x)\right\} + \epsilon. \tag{45}$$

Hence, we have the desired result.

C Proof of Lemma 2

(i) Assume that there exist a $t_1 > 0$ and a $\zeta > 0$ such that $g(t_1) = \zeta$. Since $g(0) = 0$, according to the continuity property of the function $g(\cdot)$, there exists a $t_2 \in (0, t_1)$ such that $g(t_2) = \zeta/2$ and $g(t) \geq \zeta/2 > 0$ for any $t \in (t_2, t_1]$. Since $g(t) > 0$ for any $t \in [t_2, t_1]$ and $\frac{D^+}{dt^+}g(t) \leq 0$ whenever $g(t) > 0$, we have $g(t_1) \leq g(t_2)$, which contradicts that $g(t_1) = \zeta > g(t_2) = \zeta/2$. Therefore, $g(t) = 0$ for all $t \geq 0$.

(ii) Since $\frac{D^+}{dt^+}g(t) \leq -\gamma$ whenever $g(t) > 0$, the function $g(\cdot)$ will first hit zero at time $T = g(0)/\gamma$. Then, by using the technique in (i), we can show that $g(t) = 0$ for all $t \geq T$.