Diffusion Approximation Analysis of Multihop Wireless Networks: Quality-of-Service and Convergence of Stationary Distribution

K. S. Ashok Krishnan and Vinod Sharma

Abstract Consider a multihop wireless network, with multiple source–destination pairs. We obtain a channel scheduling policy which can guarantee end-to-end mean delay for different traffic streams. We show the stability of the network for this policy by convergence to a fluid limit. It is intractable to obtain the stationary distribution of this network. Thus, we also provide a diffusion approximation for this scheme under heavy traffic. We further show that the stationary distribution of the scaled process of the network converges to that of the Brownian limit. This theoretically justifies the performance of the system. We verify the theoretical properties by means of simulations.

Keywords Multihop wireless network · Quality-of-service · Diffusion approximation

1 Introduction and Literature Review

A multihop wireless network is constituted by nodes communicating over a wireless channel. Some of the nodes, called source nodes, have data to be sent to other nodes, called receivers. In general, the data will have to be transmitted across multiple hops. The data, originating from different applications, may have different quality-of-service (QoS) requirements, such as delay or bandwidth constraints. Therefore, we need to design routing and link scheduling algorithms that can meet all these requirements.

Network performance has been studied using various mathematical techniques. Stability of flows in a network is a minimum QoS requirement. Algorithms based

K. S. Ashok Krishnan (✉) · V. Sharma
Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India
e-mail: ashokk@iisc.ac.in; vinod@iisc.ac.in

V. C. Joshua et al. (eds.), *Applied Probability and Stochastic Processes*, Infosys Science Foundation Series, https://doi.org/10.1007/978-981-15-5951-8_2

on backpressure [7] are throughput optimal, which means that they stabilize the network if it is possible by any other policy. Another approach is to use the framework of Markov decision processes [17]. The problem of minimizing power while simultaneously providing mean and hard delay guarantees is studied in [12]. However, knowledge of system statistics is required, and the scheme is not throughput optimal. In [13], an algorithm using per hop queue length information is presented, along with a low complexity approximation that stabilizes a fraction of the capacity region. In [18], the problem of routing and scheduling transient flows in a multihop network is studied. They also provide schemes for optimal routing.

The analysis of fluid scaling of networks was pioneered in works such as [16] and [4], where it was demonstrated that stability of the fluid limit of the network implies the stability of the network. Further, one may obtain bounds on moments of asymptotic values of the queues using these techniques [5]. A comprehensive treatment of work in this direction is provided in [14]. A delay-based scheduling scheme is proposed in [9], where the analysis of stability uses fluid limits.

Diffusion approximation of networks [20] has been used to study the behaviour of the system under a scaling corresponding to the functional central limit theorem [1]. The weak limit of the diffusion scaled systems under heavy traffic is generally a reflected Brownian motion [8], which under certain assumptions on the scaling rate has a limiting stationary distribution. This distribution may be used as a proxy for the actual distribution of the system state. The diffusion approximation of the MaxWeight algorithm is studied in [19], using properties of certain fluid scaled paths to obtain properties of the diffusion scaled paths, as in [2]. Of these, [19] deals with a discrete time switch under the MaxWeight policy.

To obtain the behaviour of the network under stationarity, one also needs to show the convergence of the stationary distribution of the network to that of the limiting network. Sufficient conditions for these have only recently been studied, in [6] and [3], in the case of Jackson networks. An important requirement for the exchange of limits in [3] to hold is the Lipschitz continuity of an underlying Skorokhod map, which may not always hold in general. A recent concise survey of diffusion approximations and convergence of stationary distributions is given in [15].

Our main contributions in this work are summarized below.

- We propose a new link scheduling algorithm to guarantee end-to-end mean delay for different traffic flows. This algorithm is close to the one proposed in [10] and has the same fluid limit. Hence, it is also throughput optimal.
- We obtain a reflected Brownian motion (with drift) as the weak limit of the system under diffusion scaling. This Brownian motion exhibits state space collapse.
- We also show that the stationary distribution of our network converges to the stationary distribution of the limiting Brownian network. This allows us to approximate the stationary distribution of our network by that of the limiting network which is explicitly available. While diffusion approximations have been used to traditionally study networks, the proof of convergence of stationary distributions is still not known in many systems. Our work proves this in a

controlled multihop wireless system with a general scheduling policy with QoS provisions. However, our proof does not require Lipschitz continuity of the Skorokhod map, unlike [3].

The rest of the paper is organized as follows. In Sect. 2, we describe the system model and formulate the control policy used in the network. In Sect. 3, we describe the two scaling regimes in which we study the network and prove the existence of the Brownian limit. In Sect. 4, we show that the stationary distribution of the limit of the scaled process is the stationary distribution of the limiting Brownian process. In Sect. 5, we provide simulation results, followed by the conclusions in Sect. 6.

2 System Model and Control Policy

We consider a multihop wireless network (Fig. 1). The network is a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{1, 2, \ldots, N\}$ being the set of nodes and $\mathcal{E}$ being the set of links on $\mathcal{V}$. The system evolves in discrete time denoted by $t \in \{0, 1, 2, \ldots\}$. The links are directed, with link (i, j) from node i to node j having a time varying channel gain $H_{ij}(t)$ at time t. Denote the channel gain vector at time t by $H(t)$, evolving as independent and identically distributed (i.i.d.) process across slots with distribution γ over a finite set $\mathcal{H}$. Let $E_h(t)$ denote the cumulative number of slots till time t when the channel state was $h \in \mathcal{H}$. The vector of all $E_h(t)$ is denoted by $E(t)$.

At a node i, $A_i^f(t)$ denotes the cumulative (in time) process of exogenous arrival of packets destined to node f. The packets arrive as an i.i.d sequence across slots, with mean arrival rate λ_i^f and variance σ_i^f. Let λ denote the vector of all λ_i^f. All traffic in the network with the same destination f is called *flow* f; the set of all flows is denoted by $\mathcal{F}$. Each flow has a fixed route to follow to its destination. At each node there are queues, with $Q_i^f(t)$ denoting the queue length at node i corresponding to flow $f \in \mathcal{F}$ at time t. For a queue Q_i^f with $i \neq f$, we have the queue evolution given by,

$$Q_i^f(t) = Q_i^f(0) + A_i^f(t) + R_i^f(t) - D_i^f(t), \tag{1}$$

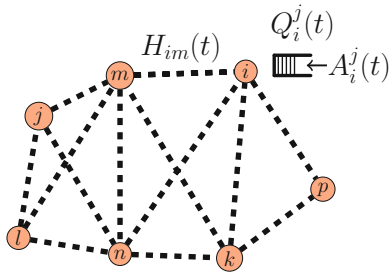

Fig. 1 A simplified depiction of a wireless multihop network

where $R_i^f(t)$ is the cumulative arrival of packets by routing (i.e., arrivals from other nodes), and $D_i^f(t)$ is the cumulative departure of packets. Let $S_{ij}^f(t)$ be the cumulative number of packets of flow f transmitted over link (i, j). We write

$$R_i^f(t) = \sum_{k \neq i} S_{ki}^f(t), \text{ and } D_i^f(t) = \sum_{j \neq i} S_{ij}^f(t). \tag{2}$$

We assume that the links are sorted into M *interference sets* $I_1, I_2, \ldots, I_M$. At any time, only one link from an interference set can be active. A link may belong to multiple interference sets. We also assume that each node transmits at unit power. Then, the rate of transmission between node i and node j is given by an achievable rate function, which depends on $H(t)$ and the schedule at time t.

The vector of queues at time t is denoted by $Q(t)$. Similarly, we have the vectors $A(t)$, $R(t)$, $D(t)$ and $S(t)$. Consider a vector $\varrho = [\varrho_{ij}^f]_{(i,j)\in\mathcal{E}, f\in\mathcal{F}}$. Define $\mathcal{S}$ to be the set of all ϱ that satisfies,

1. $\varrho_{ij}^f \in \{0, 1\}\ \forall i, j, f$,
2. $\sum_f \varrho_{ij}^f \leq 1,\ \forall (i, j) \in \mathcal{E}$,
3. $\sum_{(i,j)\in I_m} \sum_f \varrho_{ij}^f \leq 1,\ m = 1, \ldots, M$.

Such a vector is called a *schedule*. Clearly, any $\varrho \in \mathcal{S}$ has elements ϱ_{ij}^f, which, if one, indicates that flow f is to be sent over link (i, j). The constraints listed above represent the fact that no two flows can be transmitted simultaneously on a link at any time. Furthermore, no two links in an interference set can transmit at the same time. For any schedule ϱ and channel state h, we assume there exists a channel rate function $\mu = [\mu_{ij}^f]_{(i,j)\in\mathcal{E}, f\in\mathcal{F}}$, where,

$$\mu_{ij}^f = \mathbb{F}(h, \varrho), \tag{3}$$

where $\mathbb{F}$ is some achievable rate function.

We want to develop scheduling policies such that the different flows obtain their end-to-end mean delay deadline guarantees. Define $Q_{ij}^f = \max(Q_i^f - Q_j^f, 0)$, $Q^f(t) = \sum_i Q_i^f(t)$. Let $\mathcal{M}(t) = \{\mathbb{F}(H(t), \varrho) : \varrho \in \mathcal{S}\}$ be the set of feasible rates at time t. Our network control policy is as follows. At each t, given the region of feasible rates $\mathcal{M}(t)$, we obtain the optimal allocation μ^*,

$$\mu^* = \arg_{\mu\in\mathcal{M}(t)} \max \sum_{i,j,f} \alpha(Q^f(t), \overline{Q}^f) Q_{ij}^f(t) \mu_{ij}^f, \tag{4}$$

assuming $Q_{ij}^f > 0$ for at least one link flow pair (i, j), f. If all Q_{ij}^f are zero, we define the solution to be $\mu^* = 0$. We optimize a weighted sum of rates, with more weight given to flows with larger backlogs, with α capturing the delay requirement

of the flow. The weights α are functions of $Q^f(t)$, and $\overline{Q}^f$ denotes a desired value for the queue length of flow f. We use,

$$\alpha(x, \overline{x}) = 1 + \frac{a_1}{1 + \exp(-a_2(x - \overline{x}))}. \tag{5}$$

Thus, flows requiring a lower mean delay would have a higher weight compared to flows needing a higher mean delay. Flows whose mean delay requirements are not met should get priority over the other flows. The $\overline{Q}^f$ are chosen, using Little's law, as $\overline{Q}^f = \lambda^f \overline{D}$, where $\overline{D}$ is the target end-to-end mean delay and λ^f is the arrival rate of flow f. Note that we will often use $\alpha(x)$ instead of $\alpha(x, \bar{x})$ for simplicity of notation.

Let $G^{hI}_{ijf}(t)$ be the number of slots till time t, in which channel state was h, the schedule was I and flow f was scheduled over (i, j). Denote the vector of all $G^{hI}_{ijf}(t)$ by $G(t)$. Define the process,

$$Z = (A, E, G, D, R, S, Q), \tag{6}$$

where we have $A = (A(t), t \geq 0)$ (and likewise for the other processes). This process describes the evolution of the system. The state of the system at time t is $Q(t)$, which takes values in a state space $\mathcal{Q}$. Define the *capacity region* as follows.

Definition 1 The capacity region Λ of the network is the set of all λ for which a stabilizing policy exists.

We denote the set of real numbers by $\mathbb{R}$, and the set of integers by $\mathbb{Z}$. We use $\mathscr{C}[0, \infty)$ to denote the set of all continuous functions from $[0, \infty)$ to $\mathbb{R}$, and $\mathscr{D}[0, \infty)$ the set of all right continuous functions with left limits (RCLL) from $[0, \infty)$ to $\mathbb{R}$. We use $\Longrightarrow$ to denote weak convergence. For a vector x, $|x|$ denotes its norm (modulus). The vector of variables of the form x_i^j over all i and j will be denoted by $(x_i^j)_{i,j}$. For any two vectors x and y, we denote their inner product by $\langle x, y \rangle$. For a vector $x = (x_1, \ldots, x_n)$ and scalar t, xt will be the product $(x_1 t, \ldots, x_n t)$. We will also need the following definition.

Definition 2 A sequence of functions ζ_n is said to converge uniformly on compact sets (u.o.c) to ζ if $\zeta_n \to \zeta$ uniformly on every compact subset of the domain.

3 Fluid and Diffusion Limits

Now we describe the behaviour of Z under two scaling regimes, *fluid* and *diffusion*.

3.1 Fluid Scaling

For the process Z, define the scaled continuous time process,

$$z^n(t) = \frac{Z(\lfloor nt \rfloor)}{n}, \tag{7}$$

where $\lfloor \cdot \rfloor$ represents the floor function. This is called the fluid scaled process. Note that the time argument t on the left side is continuous, while that on the right is discrete. Whether a time argument is discrete or continuous will be generally clear from the context. Let z^n denote the process $(z^n(t), t \geq 0)$. We have

$$z^n = (a^n, e^n, g^n, d^n, r^n, s^n, q^n), \tag{8}$$

with the scaling in (7) being applied to each component of Z. Note that $a^n = (a_i^{f,n})_{i,f}$, and a similar notational convention holds for all the constituent functions of z. The limit of z^n, as $n \to \infty$, offers insight into the behaviour of the system under the scheduling policy in (4). The following result can be shown for our policy.

Lemma 1 *The algorithm described by the slot-wise optimization in (4) stabilizes the system for all arrival rate vectors λ in the interior of Λ. Here, stability implies that the Markov chain $Q(t)$ is positive recurrent.*

The proof of this lemma proceeds on the lines of the proof of Theorems 1 and 3 in [10]. It can be shown that there exists a subsequential limit $z = (a, e, g, d, r, s, q)$ for the family $\{z^n, n \geq 0\}$. This z is called the *fluid limit*, and the convergence of the processes is u.o.c. The limiting functions are also Lipschitz continuous, and hence almost everywhere differentiable. The points t at which it is differentiable are called regular points. In addition, the limiting functions satisfy the following properties (see [11]):

$$a(t) = \lambda t, \quad e(t) = \gamma t, \tag{9}$$

$$r_i^f(t) = \sum_j s_{ji}^f(t), \quad d_i^f(t) = \sum_j s_{ij}^f(t), \tag{10}$$

$$q(t) = q(0) + a(t) + r(t) - d(t), \tag{11}$$

$$\dot{q}(t) = \lambda + \dot{r}(t) - \dot{d}(t), \tag{12}$$

$$\sum_I g_{ijf}^{hI}(t) = e_h(t), \quad s_{ij}^f(t) = \int_0^t \dot{s}_{ij}^f(\tau) d\tau, \tag{13}$$

and $\dot{s}(t)$ satisfies

$$\sum_{i,j,f} \alpha(q^f(t)) q_{ij}^f(t) \dot{s}_{ij}^f(t) = \max_{\bar{\mu}} \sum_{i,j,f} \alpha(q^f(t)) q_{ij}^f(t) \bar{\mu}_{ij}^f, \tag{14}$$

where the dot indicates derivative, at regular t and $\bar{\mu} = \sum_h \gamma_h \mu(h, \mathbb{S})$, where $\mu(h, \mathbb{S})$ is an achievable rate when channel is in state h and schedule is $\mathbb{S}$.

Using the Lyapunov function,

$$\mathcal{L}_1(q(t)) = -\int_t^\infty \exp(t-\tau) \sum_{i,f} \alpha(q^f(\tau)) q_i^f(\tau) \dot{q}_i^f(\tau) d\tau,$$

we can establish that the fluid system is stable, and consequently, so is the stochastic system. We can also show that the draining time of the system, which is the time for all the fluid queues to go to zero, is of the form $\frac{T|x|}{\epsilon}$, where T is a finite quantity,$|x|$ is the initial norm of the fluid queues and ϵ denotes the distance of λ to the boundary of Λ.

Studying the fluid limit gives us insights into the stability properties of the system. However, it only proves the existence of a stationary distribution. In order to predict the behaviour of the system, one needs the stationary distribution, or some approximation to the same. However, explicitly computing the stationary distribution for our system is not feasible. Thus, we define the heavy traffic regime, and the associated diffusion scaling, below. We will also show that the stationary distribution of our system process converges to that of the limiting Brownian network. This will provide us an approximation of the stationary distribution under heavy traffic, the scenario of most practical interest.

3.2 Diffusion Scaling

Consider a sequence of systems, Z^n. Each system differs from the other in its arrival rate, λ^n. The λ^n are chosen such that, as $n \to \infty$, $\lambda^n \to \lambda^*$, and,

$$\lim_{n\to\infty} n\langle \psi, \lambda^n - \lambda^* \rangle = b^* \in \mathbb{R}, \tag{15}$$

where λ^* is a point on the boundary of Λ, and ψ denotes the outer normal vector to Λ at the point λ^*. This is known as heavy traffic scaling. We will also assume that λ^* falls in the relative interior of one of the faces of the boundary of Λ. For this sequence of systems, we define the diffusion scaling, given by,

$$\hat{z}^n(t) = \frac{Z^n(\lfloor n^2 t \rfloor)}{n}. \tag{16}$$

Let $\hat{z}^n$ denote the process $(\hat{z}^n(t), t \geq 0)$. As before, we have

$$\hat{z}^n = (\hat{a}^n, \hat{e}^n, \hat{g}^n, \hat{d}^n, \hat{r}^n, \hat{s}^n, \hat{q}^n).$$

Define the system workload $W^n(t)$ in the direction ψ as,

$$W^n(t) = \langle \psi, Q^n(t) \rangle, \tag{17}$$

and,

$$\hat{w}^n(t) = \frac{W(\lfloor n^2 t \rfloor)}{n}.$$

Denote $\hat{w}^n = (\hat{w}^n(t), t \geq 0)$. Define an invariant point to be a vector ϕ that satisfies, for some $k > 0$,

$$\alpha(\phi)\phi = k\psi, \tag{18}$$

where $\alpha(\phi)$ is the vector of all $\alpha(\phi_j)$, with α defined in (5), and the product of the vectors is element-wise. Then, we have the following result, which characterizes the weak convergence of the diffusion scaled processes.

Theorem 1 *Consider $\{\hat{z}^n, n \in \mathcal{N}\}$, under heavy traffic scaling satisfying (15),and $\mathcal{N}$ a sequence of positive integers n increasing to infinity. Assume that the fluid scaled $z = (a, e, g, d, r, s, q)$ has components $a = (a_i^f)_{i,f}$ and $e = (e_h)_{h \in \mathcal{H}}$ that satisfy, with probability one, as $m \to \infty$, for any $T > 0$, for all $i, j, f, c \in \mathcal{H}$,*

$$\max_{0 \leq \ell \leq mT} \sup_{0 \leq \epsilon \leq 1} |a_i^f(m, \ell + \epsilon) - a_i^f(m, \ell) - \lambda_i^f \epsilon| \to 0, \tag{19}$$

$$\max_{0 \leq \ell \leq mT} \sup_{0 \leq \epsilon \leq 1} |e_c(m, \ell + \epsilon) - e_c(m, \ell) - \gamma_c \epsilon| \to 0. \tag{20}$$

Further, assume that,

$$\hat{q}^n(0) \implies c\phi, \tag{21}$$

where c is a non-negative real number. Then, the sequence $\{\hat{w}^n, n \in \mathcal{N}\}$ converges weakly to a reflected Brownian motion $\hat{w}$ as $n \to \infty$, in $\mathscr{D}[0, \infty)$. Further, $\{\hat{q}^n, n \in \mathcal{N}\}$ converges weakly to $\phi\hat{w}$.

The existence of the Brownian limit is demonstrated as follows. We write the scaled workload $\hat{w}^n$ as the sum of two terms, one of which converges to a Brownian motion, and the second as its corresponding *regulating* process. Together, they act as a reflected Brownian motion. The detailed proof is available in [11].

Having established the existence of a limiting Brownian motion, we proceed to demonstrate that the stationary distributions of the scaled systems converge to the stationary distribution of the Brownian motion, in the next section.

4 Convergence of Stationary Distributions

In order to establish the convergence of stationary distributions, we use the following result, which is a consequence of Theorems 3.2, 3.3 and 3.4 of [3].

Lemma 2 *Assume that, for all nodes i, j, flows f, for any $n \geq 1$, $t \geq 0$, we have, for some $B < \infty$,*

$$\mathbb{E}\left[\sup_{0\leq k\leq t}\left|A_i^{f,n}(k) - \bar{a}_i^{f,n}(k)\right|^2\right] \leq Bt, \tag{22}$$

$$\mathbb{E}\left[\sup_{0\leq k\leq t}\left|R_i^{f,n}(k) - \bar{r}_i^{f,n}(k)\right|^2\right] \leq Bt, \tag{23}$$

$$\mathbb{E}\left[\sup_{0\leq k\leq t}\left|D_i^{f,n}(k) - \bar{d}_i^{f,n}(k)\right|^2\right] \leq Bt. \tag{24}$$

Further, assume that there exists T such that for all $t \geq T$, we have,

$$\lim_{|x|\to\infty}\sup_n \frac{1}{|x|^2}\mathbb{E}|\hat{q}_x(n, t|x|)|^2 = 0. \tag{25}$$

Then the sequence of distributions $\{\pi_n\}$ is tight.

It can be shown that the above conditions are satisfied in our case, as stated below.

Lemma 3 *In our system model, conditions (22)–(24) hold. Further, there exists T such that (25) holds. Consequently, the sequence $\{\pi_n\}$ is tight.*

Proof See [11]. □

As a consequence of the above two lemmas, we have the following result.

Theorem 2 *As $n \to \infty$,*

$$\hat{q}^n(\infty) \implies \phi\hat{w}(\infty), \text{ as } n \to \infty, \tag{26}$$

where the time argument being infinity denotes the respective stationary distributions.

Proof See [11]. □

The Brownian motion $\hat{w}$ obtained as the limit of $\hat{w}^n$ is a unidimensional reflected Brownian motion, having drift $b^* < 0$. The distribution of $\hat{w}(\infty)$ is given by [8],

$$\mathbb{P}[\hat{w}(\infty) < y] = 1 - \exp(2b^* y/\sigma^2). \tag{27}$$

This therefore becomes an approximation for the queue length distribution of the system under heavy traffic.

5 Numerical Simulations

For simulations, we consider two topologies. In both cases, the slot-wise allocation is done by performing the optimization (4). We compute the solution numerically by means of an exhaustive search. Since the search space of the optimization increases exponentially in the number of channel states, we limit the channel states to take values over a finite set of small size.

Example 1 We consider a star network topology (Fig. 2). There are two Poisson distributed arrival processes, one arriving at node 1, with node 4 as its destination. The other arrives at node 2, with node 5 as destination. Two links which share a common node interfere with each other. Thus, there is one interference set, which contains all the links. Consequently, only one link can be active at a time. We assume that the channels are independent and identically distributed, with the distribution being uniform over the values {0, 1, 2, 3}. The arrival vector $(\lambda_1, \lambda_2) = (\lambda, \lambda)$, i.e., increasing along the line of unit slope. Under heavy traffic, it is easy to see that, given the interference constraints, it is optimal to schedule the link with the highest channel gain. From simulations, the maximum arrival rate that can be supported by scheduling the link with the highest channel gain yields $\lambda^* = (0.65, 0.65)$. From the diffusion approximation and (27), we can see that the mean of the Brownian motion corresponding to the queue can be approximated by the vector $\phi\frac{\sigma^2}{2b^*}$. The Brownian motion is a limit of the scaled process of the form $\frac{Q(n^2t)}{n}$. For a large n, we may approximately write, $Q(n^2t) \approxeq n\phi\frac{\sigma^2}{2b^*}$. If we run the simulations for a

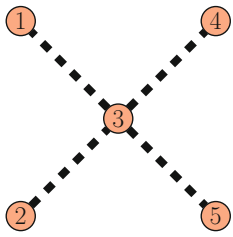

Fig. 2 Example 1: the network

Table 1 Approximation of queues

Arrival rate λ	Mean queue length	Approximation
0.64	233	232
0.641	263	258
0.642	319	290
0.643	367	332
0.644	381	387
0.645	479	465
0.646	517	581
0.647	568	775

The mean queue length of the flow $1 \to 3 \to 5$ corresponding to various arrival rates is displayed, along with the numerical approximation

Table 2 Mean queue length target and obtained, for both flows

λ	Mean queue length asked	Queue length obtained
0.63	(250, 100)	(213, 98)
0.64	(250, 100)	(264, 110)
0.641	(250, 100)	(292, 120)

time n, we may further also approximately write $b^* = n|\lambda - \lambda^*|$. Hence, we have the approximation,

$$Q(\infty) \approxeq \phi \frac{\sigma^2}{2|\lambda - \lambda^*|}. \tag{28}$$

We will be looking at the total queue length of the flow $1 \to 3 \to 4$. The value of σ^2 is $2\lambda + \hat{\sigma}^2$. The vector ϕ is approximately $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. (The value of $\bar{Q}$ for both queues is set at 100.) We take $\hat{\sigma}^2 \approxeq 8$. The values of the total queue length of the flow $1 \to 3 \to 5$ are listed in Table 1 (owing to symmetry both queue lengths are same), for simulation runs of length 10^5, averaged over 20 simulations. It can be seen that the approximations follow the queue length closely.

In order to demonstrate that the algorithm can satisfy different QoS requirements, we simulate the network at three points in the interior of the capacity region. The mean queue length asked from the flows is 250 and 100, respectively. We also pick a_2 in the expression of α for the second flow to be 4, since it requires a tighter constraint to be met. In Table 2, the first column gives the arrival rate, the second shows the target queue length for the two flows and the final column shows the queue length obtained. We see that the end-to-end mean queue length requirement is met for both the flows till rate 0.64. The capacity boundary is at 0.65. Thus, our algorithm can provide QoS under heavy traffic as well.

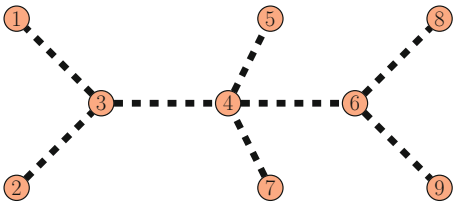

Fig. 3 Example 2: the network

Example 2 Consider the network in Fig. 3. The arrival process, channel state distribution and interference constraints are the same as in Example 1. There are three flows, $1 \rightarrow 3 \rightarrow 4 \rightarrow 6 \rightarrow 8$, $2 \rightarrow 3 \rightarrow 4 \rightarrow 5$ and $7 \rightarrow 4 \rightarrow 6 \rightarrow 9$. They will be called flow 8, flow 5 and flow 9. From simulations, the boundary of the capacity region, $\lambda^* \approx (0.59, 0.59, 0.01)$. We take arrival rates close to this point and show the values of total queue length of flow 8 obtained by simulations and the numerical approximations (using (28)), in Table 3. For calculating the approximation, we use $\hat{\sigma}^2 \approx 9$. In this case also, the approximations track the queue lengths well. Just as in the previous case, we provide an example to show how the queue length values meet targets, in Table 4. These are simulated at the arrival rate (0.55.0.55, 0.01), which is in the interior of the capacity region. In the weight function α, we use $a_1 = 5, a_2 = 1$ to give weights to flows. Since flows 8 and 5 are competing for network resources, delays of both cannot be reduced simultaneously. This is also clear from the simulations.

Table 3 Entries of the form (a, b) indicate delay target a, delay achieved b

Arrival rate λ	Mean queue length	Approximation
0.5	21	26
0.54	52	47
0.56	99	79
0.57	119	144
0.58	253	239
0.582	331	299
0.584	403	399
0.585	457	479

Table 4 Entries of the form (a, b) indicate delay target a, delay achieved b

Mean delay (slots) for each flow		
Flow 8	Flow 5	Flow 9
(50, 52)	(100, 112)	9
(40, 46)	(100, 114)	9
(100, 139)	(50, 53)	21

Arrival rate is (0.55.0.55, 0.01)

6 Conclusion

We have presented an algorithm for scheduling in multihop wireless networks that guarantees end-to-end mean delays of the packets transmitted in the network. The algorithm is throughput optimal. Using diffusion scaling, we obtain the Brownian approximation of the algorithm. We also prove theoretically that the stationary distribution of the limiting Brownian motion is the distribution of a sequence of scaled systems, and is consequently a good approximation for the stationary distribution of the original system. Using these relations, we obtain an approximation for queue lengths, and demonstrate via simulations that these are accurate.

References

1. Billingsley, P.: Convergence of Probability Measures. Wiley, London (1968)
2. Bramson, M.: State space collapse with application to heavy traffic limits for multiclass queueing networks. Queueing Syst. **30**(1–2), 89–140 (1998)
3. Budhiraja, A., Lee, C.: Stationary distribution convergence for generalized Jackson networks in heavy traffic. Math. Oper. Res. **34**(1), 45–56 (2009)
4. Dai, J.G.: On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. Ann. Appl. Probab. **5**, 49–77 (1995)
5. Dai, J.G., Meyn, S.P.: Stability and convergence of moments for multiclass queueing networks via fluid limit models. IEEE Trans. Autom. Control **40**(11), 1889–1904 (1995)
6. Gamarnik, D., Zeevi, A., et al.: Validity of heavy traffic steady-state approximations in generalized Jackson networks. Ann. Appl. Probab. **16**(1), 56–90 (2006)
7. Georgiadis, L., Neely, M.J., Tassiulas, L., et al.: Resource allocation and cross-layer control in wireless networks. Found. Trends® Netw. **1**(1), 1–144 (2006)
8. Harrison, J.: Brownian Motion and Stochastic Flow Systems. Wiley, London (1985)
9. Ji, B., Joo, C., Shroff, N.B.: Delay-based back-pressure scheduling in multihop wireless networks. IEEE/ACM Trans. Netw. **21**(5), 1539–1552 (2013)
10. Krishnan, A., Sharma, V.: Distributed control and quality-of-service in multihop wireless networks. In: 2018 IEEE International Conference on Communications (ICC), pp. 1–7. IEEE, Piscataway (2018)
11. Krishnan, A., Sharma, V.: Quality-of-service in multihop wireless networks: diffusion approximation (2018). arXiv:1810.12209
12. Kumar, S.V., Sharma, V.: Joint routing, scheduling and power control providing hard deadline in wireless multihop networks. In: 2017 Information Theory and Applications Workshop (ITA). San Diego (2017)
13. Li, B., Srikant, R.: Queue-proportional rate allocation with per-link information in multihop wireless networks. Queueing Syst. **83**(3–4), 329–359 (2016)
14. Meyn, S.: Control Techniques for Complex Networks. Cambridge University Press, Cambridge (2008)
15. Miyazawa, M.: Diffusion approximation for stationary analysis of queues and their networks: a review. J. Oper. Res. Soc. Jpn. **58**(1), 104–148 (2015)
16. Rybko, A.N., Stolyar, A.L.: Ergodicity of stochastic processes describing the operation of open queueing networks. Problemy Peredachi Informatsii **28**(3), 3–26 (1992)
17. Singh, R., Kumar, P.: Throughput optimal decentralized scheduling of multi-hop networks with end-to-end deadline constraints: unreliable links (2016). arXiv:1606.01608

18. Siram, V., Varma, K., et al.: Routing and scheduling transient flows for QoS in multi-hop wireless networks. In: 2018 International Conference on Signal Processing and Communications (SPCOM). Bangalore (2018)
19. Stolyar, A.L., et al.: Maxweight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. Ann. Appl. Probab. **14**(1), 1–53 (2004)
20. Williams, R.J.: Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. Queueing Syst. **30**(1–2), 27–88 (1998)