Multihop Wireless Networks with Quality-of-Service: Fluid and Diffusion Approximations

Ashok Krishnan K. S. and Vinod Sharma, *Senior Member, IEEE*

***Abstract*—We consider a multihop wireless network. There are multiple source destination pairs. We propose a scheduling algorithm to provide end-to-end mean delay guarantees to the flows. We demonstrate that the algorithm is also throughput optimal. Under heavy traffic, we obtain a Brownian approximation of the network. Furthermore, we also show that the stationary distribution of the scaled process of the network converges to that of the Brownian limit, providing an approximation to the stationary distribution under heavy traffic. Finally simulations further verify our claims.**

***Index Terms*—Fluid Limit,Diffusion Approximation,Multihop Network,Convergence of Stationary Distributions**

I. Introduction & Literature Review

A Multihop wireless network consists of nodes communicating to each other over a wireless medium. The data generated at a source node, may have to travel across multiple nodes before reaching its destination. These data flows have different Quality-of-Service (QoS) requirements, depending on the application which generated them. These requirements include guarantees on the mean end-to-end delay, hard bounds on the delay deadline and stability. To design scheduling algorithms that meet these varying QoS demands, is a complex problem.

Flows in a network may arise from various applications, each requiring its specific form of QoS. The coming together of different kinds of data, both human and machine generated, sent over a common network, is central to the idea of the Internet of Things (IoT) [1]. Applications using Voice over IP (VoIP) are sensitive to delay variations in the network [2]. Remote healthcare systems, which involve collecting data from a patient in a remote location and transmitting it elsewhere to be analysed, require delay QoS guarantees to ensure timely interventions [3].

Stability is a minimum QoS requirement that many applications seek. Algorithms are said to be *throughput optimal* if they stabilize all the network queues for all arrival rates within the *capacity region* [4], [5]. The class of *backpressure* based *maxweight* algorithms is throughput optimal. In [6], a joint power control, scheduling and routing algorithm using backpressure is shown to be throughput optimal using Lyapunov drift based arguments. To improve delay performance, one may incorporate weights into the backpressure based optimization formulation [7]. In general, however, they may not yield good performance in terms of delay [8].

Indian Institute of Science,Bangalore

A popular approach in the control of networks is to use solutions from the theory of Markov Decision Processes (MDPs). In [9], the authors decompose an optimization to packet level to optimize delay. In [10], the algorithm optimizes the sum of the product of the rate and a value function. The algorithm is shown to be delay optimal in the asymptotic regime of slot length going to zero. An alternate approach, in the large queue length regime, is to transform delay requirements of flows to *effective bandwidth* and *effective delay*, given by large deviations theory. This transforms delay requirements into equivalent physical optimization problems. However, these approximations are easier for single hop, than multihop systems, because the coupling between queues is more complex [11]. A MAC layer algorithm to provide QoS to different users, using a priority scheme, is studied in [12]. Here, multiple users are connected to a single base station, and priority is provided based on channel and service status. The problem of minimizing power while providing mean and hard delay guarantees is studied in [13].

While the direct analysis of queueing networks can be a hard problem, one may look at various scaling regimes to obtain useful insights. In [14], [15], the authors used the technique of fluid limit analysis to study network performance and establish sufficient conditions for stability of the network. A sequence of scaled networks is studied, and its (fluid) limit yields insights about the performance of the original system. In [16], it is shown that the stability of the fluid limit implies the positive recurrence of the Markov chain corresponding to the original unscaled system. In [17], the authors provide ufficient conditions for obtaining bounds on the steady state moments of queue lengths in a multiclass queueing network. Optimizing the fluid equivalent of a cost function is studied in [18]. The technique of *discrete review* is used in [19]. Here, the network is viewed at certain review instants, and control decisions are taken till the next review instant using information from the current state. A distributed algorithm that provides mean delay and hard deadline guarantees in a multihop network, in a discrete review set-up, is studied in [20]. The resource allocation uses (approximate) draining time of the fluid network to compute its optimal policy. A throughput optimal algorithm that provides mean delay guarantees is presented in [21]. In [22] a throughput optimal, per-queue based scheduling algorithm is presented.

Diffusion approximations of networks is the study of queueing systems scaled to reveal asymptotics corresponding to the Functional Central Limit Theorem (FCLT) [23]. The networks

are scaled while simultaneously increasing the traffic intensity to the boundary of capacity. This is called the Heavy Traffic regime [24]. A reflected Brownian motion is obtained as a weak limit of a sequence of processes. This process provides approximations for different statistics, such as mean delay and queue length. Sufficient conditions for the existence of a diffusion limit for multiclass queueing networks are given in [25]. This assumes a *work conserving* service policy, i.e., the queues are never idle when a customer is present. In the case of the network studied in [26], work conservation holds only asymptotically. The techniques in [27] are used therein to demonstrate state space collapse. Approximations for queues in heavy traffic is given in [28].

The diffusion limit has a stationary distribution, which is easier to calculate than the stationary distribution of the actual system. This provides an approximation for various system statistics of interest, such as mean queue length and delay. Earlier papers on diffusion approximation did not provide convergence of stationary distributions. The first paper to do so seems to be [29] for general Jackson networks. They obtain convergence under the assumption that the inter-arrival and service times have exponential moments. In [30], convergence is shown under weaker assumptions, using techniques refined from [17]. The same problem is solved in [31] for multiclass queueing networks. In [32], the authors justify the heavy traffic diffusion approximation by showing convergence of moment generating functions (MGF) of the stationary distributions of diffusion scaled processes. using the *basic adjoint relationship* and bypassing the intermediate step of showing the existence of the diffusion process.

Our main contributions in this work are summarized below.

- We propose a new scheduling algorithm to guarantee end-to-end mean delay for different traffic flows, in a wireless multihop network. The delay guarantees are implemented by a dynamic weight function which incorporates feedback. Using fluid limit analysis of the system, we demonstrate that it is throughput optimal. This is a slight variation of our algorithm proposed in [21]. The main difference is that [21] assumes a discrete review set-up, and hence the optimization is performed only at certain review instants. In the current work we do not make such an assumption. The performance of the present algorithm is similar to that in [21] as seen from simulations, and has the same fluid limit. Works that provide for *targeted* end-to-end mean delay are not commonly available in the literature.
- We obtain a reflected Brownian motion (with drift) as the weak limit of the system under diffusion scaling in the heavy traffic regime, and show that the stationary distribution of our network converges to the stationary distribution of this Brownian motion. This allows us to approximate the stationary distribution of our network by that of the Brownian limit, which is explicitly available. For this we use properties of the fluid limit, unlike the techniques in [30] which seem difficult to use in our case. Convergence of stationary distributions to the diffusion limit for a complex system with fading and end-to-end QoS guarantees has not been demonstrated previously.
- In the literature, there have been works regarding throughput optimality of maxweight type algorithms in networks [4], [6]. The knowledge of fluid limits has enhanced our understanding of these algorithms, through works such as [26]. The study of diffusion approximations to obtain queue length approximations also has along history, while studies on converegence of stationary distribution are far more recent [30]. To the best of our knowledge, this is the first work to provide a throughput optimal algorithm with end-to-end mean delay guarantee, obtains the diffusion approximation and convergence of stationary distribution to that of the approximation is demonstrated. In [21], diffusion approximation, and hence the theoretical approximation of the performance of the algorithm was not obtained.

A. Organization of this paper

This paper is organized as follows. In Section II, we describe the system model, and the optimization used for resource allocation. In Section III, we obtain the fluid limit of the system, and demonstrate throughput optimality. In Section IV, we obtain the diffusion scaled limit of the system in the heavy traffic regime. In Section V, we show the convergence of the stationary distributions of the sequence of diffusion scaled systems. In Section VI, we provide the numerical results and simulations, followed by concluding remarks in Section VII.

B. Notational Convention

We denote the set of real numbers by $\mathbb{R}$, positive reals by $\mathbb{R}_+$ and integers by $\mathbb{Z}$. We use $\mathscr{D}[0,\infty)$ the set of all right continuous functions with left limits (RCLL) from $[0,\infty)$ to $\mathbb{R}$. We use $\xrightarrow{\mathscr{L}}$ to denote weak convergence. For a real vector x, $||x||$ denotes its Euclidean norm. For a real number x, we use $x^+ = \max(x,0)$, and $|x|$ is its modulus. For a set $A \subset \mathbb{R}$, A^+ denotes $A \cap \mathbb{R}_+$. The vector of variables of the form x_i^j over all $i \in \mathcal{I}$ and $j \in \mathcal{J}$ will be denoted by $(x_i^j)_{i\in\mathcal{I},j\in\mathcal{J}}$. If x is the vector $(x_i^j)_{i\in\mathcal{I},j\in\mathcal{J}}$, and we have a sequence x^n with scaling parameter n, the components of the scaled vectors will be represented as $x_i^{j,n}$. The inner product of two vectors x and y is $\langle x,y\rangle$. We use *s.t.* to denote *such that*. Union of sets is denoted by $\cup$.

II. System Model

We consider a multihop wireless network (see Fig. 1), represented by a graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ where $\mathcal{V} = \{1,2,..,N\}$ is the set of vertices and $\mathcal{E}$, the set of links (edges) on V. The links are directional, with (i,j) representing a link from node i to node j. The system evolves in slotted time, $t \in \{0,1,2,...\}$. Associated with each link is a time varying channel gain, $H_{ij}(t)$. The *channel vector* is $H(t)$, defined as the vector of all $H_{ij}(t)$. We will assume that the channel vector evolves i.i.d. across time, taking values from a finite set $\mathcal{H}$ with distribution $\gamma = (\gamma_1 \ ... \ \gamma_{|\mathcal{H}|})$, where γ_i is the probability that channel state takes the i-th value from the set $\mathcal{H}$. Associated with this graph is a set of *flows* $\mathcal{F}$. Each flow corresponds to a stream of packets being generated at a source node to be transmitted to a destination node. For simplicity, we assume that packets have a size of one bit. We also assume that the flows have

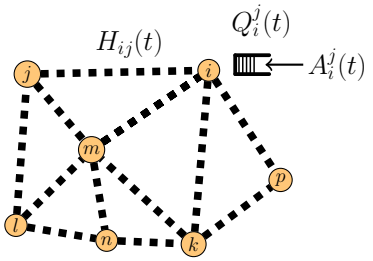

Fig. 1. A simplified depiction of a Wireless Multihop Network

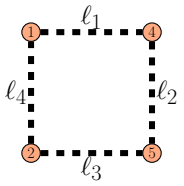

Fig. 2. Example network

fixed paths, from source to destination. For a flow $f \in \mathcal{F}$, let $src(f)$ denote its source node, and $des(f)$ its destination. For a flow $f \in \mathcal{F}$, let $A^{\text{des}(f)}_{\text{src}(f)}(t)$ denote the number of packets of flow f generated at the source node at time t. This is called an *arrival process*. Denote the vector of all such arrival processes by $A(t)$.

Depending on the physical constraints of the system, the wireless links will have interference constraints. These in turn determine $\mathbb{L}$, the set of *activation sets* of a network. An activation set is a set of links which can be ON simultaneously, i.e., they do not violate the system interference constraints. Consider the network in Fig 2, with four links. One can have different interference constraints on this system, depending on the communication protocol. An example is the case where all links use the same frequency band and are close to each other. Here, $\mathbb{L}$ consists of $\{\ell_1\}$, $\{\ell_2\}$, $\{\ell_3\}$ and $\{\ell_4\}$. In another scenario it may be that no two links that have a common node can be ON simulatneously. This is a half-duplex type constraint. Here, $\mathbb{L}$ contains $\{\ell_1, \ell_3\}$, $\{\ell_2, \ell_4\}$, $\{\ell_1\}$, $\{\ell_2\}$, $\{\ell_3\}$ and $\{\ell_4\}$. In this work, we assume that $\mathbb{L}$ is nonempty, and that each link of the network belongs to at least one activation set in $\mathbb{L}$. This ensures that all communication paths exist. We do not require any other assumption on the interference structure of $\mathbb{L}$.

Each activation set can also be represented by an *activation vector*. This is a vector $[\iota_{ij}]_{(i,j)\in\mathcal{E}}$ of length $\mathcal{E}$, with each element corresponding one-to-one with a link. The value of that element is one if the link belongs to that activation set, and zero otherwise. Thus, one generates from $\mathbb{L}$, the set of all activation vectors, $\mathbb{L}_0$. A *schedule* $\varsigma = [\varsigma^f_{ij}]_{(i,j),f}$ is a vector whose components take value 0 or 1. If $\varsigma^f_{ij} = 1$, it means that bits of flow f are to be sent over link (i, j). Further, for any schedule ς, there must exist an activation vector $\iota \in \mathbb{L}_0$ such that, $\sum_f \varsigma^f_{ij} \leq \iota_{ij}$. This ensures that the schedule is *feasible*, i.e., there exists an activation set which enables this schedule. Let $\mathcal{S}$, the set of such *feasible* schedules that additionally satisfy $\sum_f \varsigma^f_{ij} \leq 1$ (only one flow to be sent over a link at a time). We assume there exists a rate function μ, that assigns channel rates to each link in every time slot. When the channel state is $h \in \mathcal{H}$ and the schedule $\varsigma \in \mathcal{S}$ is chosen, the rate function $\mu(H(t), \varsigma)$, assigns a rate $\mu_{ij}(H(t), \varsigma)$ to a link (i, j). Note that $\mu_{ij}(h, \varsigma) = \varphi_{ij}(h)$, where φ is some achievable rate function (for some fixed transmit power).

At each node, packets present are sorted into queues, with a first-in-first-out priority, for each flow. Let $Q^f_i(t)$ denote the number of packets of flow f queued at node i. The vector of all $Q^f_i(t)$, over all nodes and flows, is denoted by $Q(t)$.

At each time instant, a network controller must decide to transmit packets contained in the queues, across links, over the network. Any such control decision can be interpreted as deciding the values of a vector $S(t) = [S^f_{ij}(t)]_{(i,j),f}$. Here, $S^f_{ij}(t)$ denotes the number of bits of flow f transmitted from node i to node j at time t. Clearly, $S(t)$ must obey some natural constraints. Since we cannot transmit more packets from a queue than what we have, $\sum_j S^f_{ij}t \leq Q^f_i(t)$. Moreover, the number of bits we can send will be constrained by the maximum channel rates available to us. These rates will depend on the channel vector, as well as the interference contraints between different links. Given a control decision $S(t)$, the queues evolve as,

$$Q^f_i(t+1) = Q^f_i(t) + A^f_i(t) + \sum_k S^f_{ki}(t) - \sum_j S^f_{ij}(t), \quad (1)$$

where $i \neq des(f)$. Define,

$$R^f_i(t) = \sum_{k\neq i} S^f_{ki}(t), \;\; D^f_i(t) = \sum_{j\neq i} S^f_{ij}(t). \quad (2)$$

These denote the net inflow and outflow from a queue, by routing. Let $R(t)$ and $D(t)$ denote the vectors $[R^f_i(t)]_{i\in\mathcal{V}, f\in\mathcal{F}}$ and $[D^f_i(t)]_{i\in\mathcal{V}, f\in\mathcal{F}}$ respectively. Then we have the queueing equation in vector notation,

$$Q(t+1) = Q(t) + A(t) + R(t) - D(t). \quad (3)$$

If a rate μ_{ij} is available across a link (i, j), it can be allocated to one of the flows over that link. Such an *allocation* vector is represented by $\hat{\mu} = (\hat{\mu}^f_{ij})_{(i,j)\in\mathcal{E}, f\in\mathcal{F}}$, where the $\hat{\mu}^f_{ij}$ are non negative and satisfy $\sum_{f\in\mathcal{F}} \hat{\mu}^f_{ij} \leq \mu_{ij}$, $\forall (i, j) \in \mathcal{E}$. Let $\tilde{\mathcal{U}}(\mu(h, \varsigma))$ be the set of all allocations corresponding to a rate vector $\mu(h, \varsigma)$, and let $\mathcal{U}(h) = \cup_{\varsigma\in\mathcal{S}} \tilde{\mathcal{U}}(h, \varsigma)$. We define the optimal allocation at time t, $\hat{\mu}^*(t)$, as

$$\hat{\mu}^*(t) = \arg_{\hat{\mu}\in\mathcal{U}(H(t))} \max \sum_{(i,j)\in\mathcal{E}, f\in\mathcal{F}} \alpha(Q^f(t), \overline{Q}^f) Q^f_{ij}(t) \hat{\mu}^f_{ij}, \quad (4)$$

assuming $Q^f_{ij} > 0$ for at least one link flow pair $(i, j), f$. If all Q^f_{ij} are zero, we define the solution to be $\hat{\mu}^f_{ij}(t) = 0$, for all i, j, f. The optimal schedule $\varsigma^*(t)$ is that ς such that $\hat{\mu}^* \in \tilde{\mathcal{U}}(\mu(H(t), \varsigma))$, with ties broken arbitrarily.

In (4), 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, which is determined by the end-to-end mean delay requirement of flow f. We use,

$$\alpha(Q^f(t), \overline{Q}^f) = 1 + \frac{a_1}{1 + \exp(-a_2(Q^f(t) - \overline{Q}^f))}. \quad (5)$$

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 $\overline{Q}^f = \lambda^f \overline{\tau}^f$, where $\overline{\tau}^f$ 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.

We define, $\mathcal{M}_h = \{\mu(h, \varsigma) : \varsigma \in \mathcal{S}\}$. Let $\overline{\mathcal{M}_h}$ be the convex hull of $\mathcal{M}_h$, and $\mathcal{M} = \sum_{h \in \mathcal{H}} \gamma_h \overline{\mathcal{M}_h}$. The *capacity region*, which is the set of all arrival rates for which a stabilizing policy exists, is defined below.

Definition II.1. *The capacity region, Λ, is the set of all arrival rate vectors λ for which there exists a non-negative vector $\varpi = [\varpi_{ij}^f]_{(i,j)\in\mathcal{E}, f\in\mathcal{F}}$ which satisfies,*

$$\varpi_{ii}^f = 0, \ \forall i, f, \quad \varpi_{ij}^i = 0, \ \forall i, j, f, \quad (6)$$

$$\lambda_i^f \leq \sum_j \varpi_{ij}^f - \sum_k \varpi_{ki}^f, \ \forall i, f, \quad (7)$$

$$\sum_f \varpi_{ij}^f \leq m_{ij}, \ \textit{for some } m \in \mathcal{M}. \quad (8)$$

We define the *rate region* $\mathcal{W}$ as follows.

$$\mathcal{W} = \{\varpi : \exists m \in \mathcal{M} \ s.t. \ \sum_f \varpi_{ij}^f \leq m_{ij}, \ \forall i, j\}. \quad (9)$$

The rate region $\mathcal{W}$ is the set of all feasible rate vectors, i.e, all rates ϖ for which there exists a schedule that achieves it.

In the next section we show that the present algorithm is throughput optimal in the sense that if there is any other algorithm that will stabilise the network for given traffic and channel statistics, then this algorithm will. This is done by obtaining the fluid limit of the system.

III. Fluid Limit

For any process $L(t)$ evolving in discrete time, define its cumulative form, $\breve{L}(t) := \sum_{\tau=1}^{t} L(t)$. Thus, we obtain $\breve{A}(t)$, $\breve{S}(t)$, $\breve{R}(t)$ and $\breve{D}(t)$, as the time cumulated $A(t)$, $S(t)$, $R(t)$ and $D(t)$ respectively. Let $\breve{E}_h(t)$ denote the number of slots till time t that the channel state was $h \in \mathcal{H}$. The vector $[\breve{E}_h(t)]_{h\in\mathcal{H}}$ will be denoted by $\breve{E}(t)$. Let $\breve{G}_{\hat{\mu}}^{h\varsigma}(t)$ denote the cumulative number of slots till time t when channel state was h, the schedule chosen was ς and the allocation vector was $\hat{\mu}$. It will be assumed that the possible allocations $\hat{\mu}$ forms a finite set. It follows that,

$$\sum_{\hat{\mu},\varsigma} \breve{G}_{\hat{\mu}}^{h\varsigma}(t) = \breve{E}_h(t). \quad (10)$$

We have,

$$Q(t) = Q(0) + \breve{A}(t) + \breve{R}(t) - \breve{D}(t). \quad (11)$$

Define the system state to be $Y(t) = Q(t)$. From the queue evolution (3) and the allocation, it is clear that the system $Y(t)$ evolves as a discrete time countable state Markov chain. The associated norm is $||Y(t)|| = \sum_{i,f} Q_i^f$. Positive recurrence of this Markov chain would imply stability. We will show the positive recurrence of this Markov process via its fluid limit.

Define the process $Z(t) = (\breve{A}(t), \breve{E}(t), \breve{G}(t), \breve{D}(t), \breve{R}(t), \breve{S}(t), Q(t))$. Let $Z = \{Z(t), t \geq 0\}$ and $Y = \{Y(t), t \geq 0\}$. For the components of the process $Z(t)$, define the corresponding scaled (continuous time) processes indexed by n, for $t \geq 0$,

$$a^n(t) = \frac{\breve{A}(\lfloor nt \rfloor)}{n}, e^n(t) = \frac{\breve{E}(\lfloor nt \rfloor)}{n}, g^n(t) = \frac{\breve{G}(\lfloor nt \rfloor)}{n},$$
$$d^n(t) = \frac{\breve{D}(\lfloor nt \rfloor)}{n}, r^n(t) = \frac{\breve{R}(\lfloor nt \rfloor)}{n}, s^n(t) = \frac{\breve{S}(\lfloor nt \rfloor)}{n},$$
$$q^n(t) = \frac{Q(\lfloor nt \rfloor)}{n}.$$

Thus we obtain $z^n(t) = (a^n(t), e^n(t), g^n(t), d^n(t), r^n(t), s^n(t), q^n(t))$. Let z^n denote $\{z^n(t), t \geq 0\}$. Note that, $z^n = (a^n, e^n, g^n, d^n, r^n, s^n, q^n)$. The term fluid limit denotes the limits obtained as we scale $n \to \infty$ for this process.

We will use the following definition.

Definition III.1. *A sequence of functions ξ^n is said to converge uniformly on compact sets (u.o.c) if $\xi^n \to \xi$ uniformly on every compact subset of the domain.*

Now we show the existence of a fluid limit for the scaled sequence of processes $\{z^n, n \geq 0\}$.

Theorem III.1. *Consider a sequence of scaled systems $\{z^n, n \geq 0\}$ such that the initial condition $||Q(0)|| = n$ in the n-th system. Then, for almost every sample path ω, there exists a subsequence $n_k(\omega) \to \infty$ such that, along this subsequence, $z^n \to z$, where $z = (a, e, g, d, r, s, q)$. The component functions of z^n converge to the respective component functions of z u.o.c. as well. The limiting functions are Lipschitz continuous, and hence almost everywhere differentiable, and satisfy the following properties for all $t \geq 0$.*

$$a(t) = \lambda t, \quad e(t) = \gamma t, \quad (12)$$

$$r_i^f(t) = \sum_{k\neq i} s_{ki}^f(t), \quad d_i^f(t) = \sum_{j\neq i} s_{ij}^f(t), \quad (13)$$

$$q_i^f(t) = q_i^f(0) + a_i^f(t) + r_i^f(t) - d_i^f(t), \quad (14)$$

$$\dot{q}_i^f(t) = \lambda_i^f + \dot{r}_i^f(t) - \dot{d}_i^f(t), \quad (15)$$

$$\sum_{I,\hat{\mu}} g_{\hat{\mu}}^{hI}(t) = e_h(t), \quad ||q(0)|| \leq 1, \quad (16)$$

$$s_{ij}^f(t) = \int_0^t \dot{s}_{ij}^f(\tau) d\tau, \quad (17)$$

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

$$\sum_{i,j,f} \alpha(q^f(t)) q_{ij}^f(t) \dot{s}_{ij}^f(t) = \max_{\varpi\in\mathcal{W}} \sum_{i,j,f} \alpha(q^f(t)) q_{ij}^f(t) \varpi_{ij}^f, \quad (18)$$

where the dot indicates derivative, at regular t (the points where the function is differentiable).

The proof is provided in appendix A. Denote the vector of all $q_i^f(t)$ by $q(t)$.

The following theorem establishes the stability of the queues under the our policy.

Theorem III.2. *The control policy defined in (4), stabilizes the process $\{Q(t), t \geq 0\}$ for all arrivals in the interior of Λ.*

The proof uses the fluid limit of the queue process, and is provided in appendix B.

Draining time τ_{drain} is defined as the time by which the fluid queue $q(t)$ has norm zero. We have the following result, which relates the draining time to the time T obtained in the proof of Lemma III.2. This result will be used subsequently, in section V. (For proof see appendix C)

Lemma III.3. *If the fluid limit satisfies $||q(0)|| \leq \delta_1 < 1$, we have $\tau_{drain} \leq \frac{T}{1-\delta_1}$.*

The fluid limit gives insights into the stability properties of the system. However, it only proves the existence of a stationary distribution. To predict the behaviour of the system, one needs the stationary distribution, or some approximation to it. 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 provides an approximation of the stationary distribution of our system under heavy traffic, the scenario of most practical interest.

IV. Diffusion Scaling and Heavy Traffic Limit

Now we consider a new sequence of scaled systems, Z^n. The n-th process is the above system but with arrival rate vector λ^n and standard deviation σ^n. As $n \to \infty$, $\lambda^n \to \lambda^*$, and,

$$\lim_{n\to\infty} n\langle \psi, \lambda^n - \lambda^* \rangle = b^* \in \mathbb{R}, \tag{19}$$

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 *heavy traffic scaling*. The arrival rate increases towards the maximum rate that the system can support. The 'speed'at which this happens is controlled by (19). This is a technical necessity to ensure that the scaled queue length process converges in the limit to a well defined Brownian motion. Heavy traffic implies that the system is facing resource constraints, and hence, it is of great practical interest as well. We will also assume that λ^* falls in the relative interior (the *relative interior* of a set Ω is its interior within the affine hull of Ω [33]) of one of the faces of the boundary of Λ (this is the *resource pooling* condition). Define the diffusion scaling,

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

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,

$$W^n(t) = \langle \psi, Q^n(t) \rangle, \text{and } \hat{w}^n(t) = \frac{W(\lfloor n^2 t \rfloor)}{n}. \tag{20}$$

Denote $\hat{w}^n = \{\hat{w}^n(t), t \geq 0\}$. Define an *invariant point* to be a vector ϕ that satisfies, for some $k > 0$, $\alpha(\phi)\phi = k\psi$, where $\alpha(\phi)$ is the vector of all $\alpha(\phi_j)$, with α defined in (5). Assume that $\sigma^n \to \sigma$, as $n \to \infty$, and the arrival process $A^n(t)$ satisfies, for all i, f,

$$\lim_{x\to\infty} \sup_{n\geq 1} \mathbb{E}[(A_i^{f,n}(1))^2 \mathbf{1}_{\{A_i^{f,n}(1)\geq x\}}] = 0. \tag{21}$$

This is a sufficient condition for Donsker's Theorem to hold for the arrival process [23]. Under these assumptions, we have the following result, which characterizes the weak convergence of the diffusion scaled processes.

Theorem IV.1. *Consider $\{\hat{z}^n, n \in \mathcal{N}\}$, under heavy traffic scaling satisfying (19),and $\mathcal{N}$ a sequence of positive integers n increasing to infinity. Assume that the arrival process satisfies (21). Further, assume that, $\hat{q}^n(0) \xrightarrow{\mathcal{L}} c\phi$, 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 $\mathcal{D}[0,\infty)$. Further, $\{\hat{q}^n, n \in \mathcal{N}\}$ converges weakly to $\phi\hat{w}$.*

The proof of this Theorem proceeds in the following manner. The process $\hat{w}^n$ is decomposed into two parts. The first of these parts converges to a Brownian motion. The second converges to the *unique regulator* corresponding to the Brownian motion. Together, they add up and form a *reflected* Brownian motion. Define $\mathcal{W}_h = \{\varpi : \exists m \in \mathcal{M}_h \; s.t. \; \sum_f \varpi_{ij}^f \leq m_{ij}, \; \forall i, j\}$. For a vector ϖ, define the transformation ζ by,

$$\zeta_i^f(\varpi) = \sum_j \varpi_{ij}^f - \sum_k \varpi_{ki}^f. \tag{22}$$

Applied to a rate vector, this shows the net outflow by routing. Define the set $\zeta(\mathcal{W}_h) = \{\zeta(\varpi) : \varpi \in \mathcal{M}_h\}$. Let us denote the maximum allocation in the direction ψ, when the channel is in state h, by $\rho_h = \max_{\vartheta\in\zeta(\mathcal{W}_h)}\langle\psi, \vartheta\rangle$, for all $h \in \mathcal{H}$. Define the vectors, $\rho = [\rho_h]_{h\in\mathcal{H}}, \hat{\rho} = [(\rho_h)^2]_{h\in\mathcal{H}}$, and the random variables, $X_\mu(t) = \rho_{H(t)}$, $t \geq 1$. The random variables $\{X_\mu(t), t \geq 0\}$ are i.i.d, with mean and variance given by, $\hat{\nu} = \langle\rho, \gamma\rangle$, $\hat{\sigma}^2 = \mathbb{E}[(X_\mu(1) - \hat{\nu})^2] = \langle\hat{\rho}, \gamma\rangle - \hat{\nu}^2 \geq 0$.

Define $\check{X}(t) = \sum_{k=1}^t X_\mu(k)$. This is the cumulative maximum possible service along ψ. Write,

$$U(t) = W(0) + \langle\psi, \check{A}(t)\rangle - \check{X}(t), \tag{23}$$

$$V(t) = \check{X}(t) + \langle\psi, \check{R}(t)\rangle - \langle\psi, \check{D}(t)\rangle, \tag{24}$$

and, consequently, $W(t) = U(t) + V(t)$. Hence, $W^n(n^2t) = U^n(n^2t) + V^n(n^2t)$. Define, $\hat{u}^n(t) = \frac{U^n(\lfloor n^2 t\rfloor)}{n}$,$\hat{v}^n(t) = \frac{V^n(\lfloor n^2 t\rfloor)}{n}$. Thus we have, $\hat{w}^n(t) = \hat{u}^n(t) + \hat{v}^n(t)$. Let us denote $\hat{w}^n = \{\hat{w}^n(t), t \geq 0\}$, $\hat{u}^n = \{\hat{u}^n(t), t \geq 0\}$ and $\hat{v}^n = \{\hat{v}^n(t), t \geq 0\}$. Now we look at convergence of $\hat{u}^n$.

Lemma IV.2. *Assume the initial condition $\hat{w}^n(0)$ converges weakly to an invariant point, $\hat{w}(0)$, as $n \to \infty$ along $\mathcal{N}$, where $\alpha(\hat{w}(0))\hat{w}(0) = \psi$. Then, it follows that, $\hat{u}^n \xrightarrow{\mathcal{L}} \hat{u}$, in $\mathcal{D}[0,\infty)$ as $n \to \infty$ along $\mathcal{N}$, where $\hat{u} = (\hat{u}(t), t \geq 0)$ is a Brownian motion with drift, given by, $\hat{u}(t) = \hat{w}(0) + b^*t + \sigma\mathcal{B}(t)$, where*

$\mathcal{B}(t)$ is a standard Brownian motion, $\sigma^2 = \sum_{i,f} (\sigma_i^f)^2 + \hat{\sigma}^2$, and b^* is given by (19).

Proof. Note that $\hat{u}^n = \frac{U^n(n^2 t)}{n} = \hat{w}^n(0) + \langle \psi, \hat{a}^n(t) \rangle - \hat{x}^n(t)$, and hence,

$$\hat{u}^n(t) = \hat{w}^n(0) + \langle \psi, \hat{a}^n(t) - \lambda^n nt \rangle - (\hat{x}^n(t) - \hat{\nu} nt) + (\langle \psi, \lambda^n \rangle - \hat{\nu}) nt.$$

Since $\hat{\nu} = \langle \rho, \gamma \rangle$, we can see that,

$$\hat{\nu} = \sum_{h \in \mathcal{H}} \gamma_h \rho_h = \sum_{h \in \mathcal{H}} \gamma_h \max_{\vartheta \in \zeta(\mathcal{M}_h)} \langle \psi, \vartheta \rangle \tag{25}$$

$$= \max_{\tilde{\vartheta} \in \sum_h \gamma_h \zeta(\mathcal{M}_h)} \langle \psi, \tilde{\vartheta} \rangle = \langle \psi, \lambda^* \rangle, \tag{26}$$

where the last equality holds since λ^* is at the boundary of the capacity region and $\tilde{\vartheta} \in \sum_h \gamma_h \zeta(\mathcal{M}_h)$ represents service rate in the system,whose inner product with ψ is maximized when it is λ^*. From (19), $(\langle \psi, \lambda^n \rangle - \hat{\nu}) nt \to b^* t$. The convergence of $(\langle \psi, \hat{a}^n(t) - \lambda^n nt \rangle, t \geq 0)$ and $(\hat{x}^n(t) - \hat{\nu} nt, t \geq 0)$ to independent Brownian motions follows by Donsker's theorem [23]. □

This establishes the weak convergence of the processes $\{\hat{u}^n(t), t \geq 0\}$. Using Skorohod representation [34] , one can construct a probability space where we have $\mathcal{D}[0, \infty)$ valued processes $\hat{u}_S^n$ and $\hat{u}_S$, such that, almost surely, $\hat{u}_S^n \to \hat{u}_S$ *u.o.c.*, where $\hat{u}_S^n$ and $\hat{u}_S$ are identical in distribution to $\hat{u}^n$ and $\hat{u}$. Thus $\hat{u}_S$ is the Brownian motion given in (IV.2). We augment this probability space to include the other components of Z as well. On this probability space, we will have the functions $\hat{\nu}^n$ and $\hat{w}^n$ as before. In this augmented probability space, we will prove the convergence of $\hat{\nu}^n$. In particular, we will show that the limit of the processes $\{\nu^n(t), t \geq 0\}$ has a limit which satisfies certain conditions necessary for it to be the unique *regulator* corresponding to the Brownian motion $\hat{u}$. The relationship between a one dimensional Brownian motion and its regulator is given by the following result [35].

Lemma IV.3 (One dimensional Skorohod Problem)**.** *Let $\xi \in \mathcal{D}[0, \infty)$, such that ξ is continuous, and $\xi(0) \geq 0$. Then there exists a unique pair of non-negative functions ξ_1, ξ_2, both in $\mathcal{D}[0, \infty)$ with $\xi_2(t)$ non decreasing and continuous, $\xi_2(0) = 0$ such that $\xi_1(t) = \xi(t) + \xi_2(t)$ for all $t \geq 0$. For any $t \geq 0$, if $\xi_1(t) > 0$, then it is not a point of increase of $\xi_2(t)$. This pair is given by, $\xi_2(t) = \sup_{0 \leq \tau \leq t} (-\xi(\tau))^+$, $\xi_1(t) = \xi(t) + \xi_2(t)$, $t \geq 0$.*

If the process $\xi(t)$ is a sample path of a Brownian motion, $\xi_2(t)$ is called its regulator, and $\xi_1(t)$ is called the reflected (regulated) Brownian motion. It is clear that the proof of convergence of the processes $\{\hat{w}^n, t \geq 0\}$ to the reflected Brownian motion corresponding to the Brownian motion $\{\hat{u}(t), t \geq 0\}$ would involve showing the limit of the processes $\{\hat{\nu}^n, t \geq 0\}$ as ξ_2 satisfies properties required by Lemma IV.3. This is done in the following theorem (see appendix D).

Theorem IV.4. *For any subsequence $\mathcal{N}_1$ of $\mathcal{N}$ as given in Theorem IV.1, there is a further subsequence $\mathcal{N}_2$ along which the processes $\{\hat{\nu}^n, t \geq 0\}$ has a limit $\hat{\nu} = \{\hat{\nu}, \geq 0\}$, such that, 1) $\hat{\nu}(t)$ is continuous, 2) $\hat{\nu}(t)$ is finite for $t \in [0, \infty)$, 3) $\hat{\nu}(0) = 0$, and,*
4) if $\hat{w}(t) > 0$, then t is not a point of increase of $\hat{\nu}$.

Now we outline the proof of Theorem IV.1.

Proof of Theorem IV.1. From Lemma IV.2, using Skorohod representation, one can construct a probability space where we have $\mathcal{D}[0, \infty)$ valued processes $\hat{u}_S^n$ and $\hat{u}_S$, such that, almost surely, $\hat{u}_S^n \to \hat{u}_S$ *u.o.c.*, where $\hat{u}_S^n$ and $\hat{u}_S$ are identical in distribution to $\hat{u}^n$ and $\hat{u}$. Thus $\hat{u}_S$ is the Brownian motion given in (IV.2). We augment this probability space to include the other components of Z as well. On this probability space, we will have $\hat{\nu}^n$ and $\hat{w}^n$ as before.

Using Theorem IV.4 with IV.3, we can see that $\hat{\nu}$ is the unique regulator corresponding to $\hat{u}$. Consequently, the process $\hat{w}$ converges to a reflected Brownian motion.

The weak convergence of $\{\hat{q}^n, n \in \mathcal{N}\}$ to $\phi \hat{w}$ will follow if $\hat{q}^n$ converges to $\phi \hat{w}$ u.o.c.. From (60), it follows that for any $t \geq 0$ and $\epsilon > 0$, there exists $\delta > 0$ *s.t.*,

$$\limsup_{n \to \infty} \sup_{\tau \in [t-\delta, t+\delta]_+} ||\hat{q}^n(\tau) - \phi \hat{w}^n(\tau)|| < \epsilon. \tag{27}$$

Let $\mathbb{C} \subset \mathbb{R}_+$ be a compact set. Let ϵ be fixed. Then, for every $t \in \mathbb{C}$, there exists a δ_t such that (27) holds. Consider all sets of the form $(t - \frac{\delta}{2}, t + \frac{\delta}{2})$. These form an open cover for $\mathbb{C}$. Since the set is compact, there exists a finite subcover [36]. Therefore, there exists some finite number K such that, we have numbers $t_1, \ldots, t_K$ all from $\mathbb{C}$, such that, $\mathbb{C} \subset \cup_{i=1}^K \left(t_i - \frac{\delta_{t_i}}{2}, t_i + \frac{\delta_{t_i}}{2}\right)$. This with (27) yields the result. □

Now we demonstrate that the stationary distributions of the scaled systems converge to the stationary distribution of the Brownian motion.

V. Convergence of Stationary Distributions

We have the following result.

Theorem V.1. *As $n \to \infty$, $\hat{q}^n(\infty) \xrightarrow{\mathcal{L}} \phi \hat{w}(\infty)$, where the time argument being infinity denotes the respective stationary distributions.*

To prove this, we define a new set of fluid limit processes, $\bar{z}^{n,r}(t) = \frac{Z^n(\lfloor rt \rfloor)}{r}$. Let $\bar{z}^{n,r} = (\bar{a}^{n,r}, \bar{e}^{n,r}, \bar{g}^{n,r}, \bar{d}^{n,r}, \bar{r}^{n,r}, \bar{s}^{n,r}, \bar{q}^{n,r})$, denote the process $(\bar{z}^{n,r}(t), t \geq 0)$, and $\bar{z}^n$ the fluid limit process obtained, for each n, by taking the limit $r \to \infty$. For each Z^n, let π_n denote the stationary distribution of the corresponding network. These exist because for each n, the system Q^n is stable. The draining time for the n-th fluid system is denoted by τ_{drain}^n. From Lemma III.3, we see that τ_{drain}^n is inversely proportional to the distance from the boundary of the capacity region Λ. It is also easy to see that, due to (19), the distance to the boundary of the capacity region, which is the plane whose normal vector is ψ, decreases as $\frac{1}{n}$. Hence,

$$\tau_{drain}^n \leq nT_1, \tag{28}$$

for some finite T_1, assuming that the initial fluid level is unity.

Now, we state a sufficient condition for the sequence $\{\pi_n, n \geq 0\}$ to be tight. Note that by writing $\hat{q}_x^n(\cdot)$ we indicate that the initial condition of the queue is x.

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

$$\mathbb{E}[\sup_{0\leq k\leq t} |\check{A}_i^{f,n}(k) - \bar{a}_i^{f,n}(k)|^2] \leq Bt, \tag{29}$$

$$\mathbb{E}[\sup_{0\leq k\leq t} |\check{R}_i^{f,n}(k) - \bar{r}_i^{f,n}(k)|^2] \leq Bt, \tag{30}$$

$$\mathbb{E}[\sup_{0\leq k\leq t} |\check{D}_i^{f,n}(k) - \bar{d}_i^{f,n}(k)|^2] \leq Bt. \tag{31}$$

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

$$\lim_{||x||\to\infty} \sup_n \frac{1}{||x||^2}\mathbb{E}|\hat{q}_x^n(t||x||)|^2 = 0. \tag{32}$$

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

The result is an adaptation of the techniques in [30] to our case. We give an outline below.

Proof. From (32), it follows that there exists M, $0 < M < \infty$, such that, with $\mathbb{D} = \{x : ||x|| < M\}$, for all $x \notin \mathbb{D}$,

$$\sup_n \mathbb{E}||\hat{q}_x^n(T||x||)|^2 \leq \frac{||x||^2}{2}. \tag{33}$$

Define $\delta = TM$ and $\tau^n(\delta) = \inf\{t \geq \delta : ||\hat{q}_x^n(t)|| \leq M\}$. Define a sequence of stopping times, $\mathcal{T}_0 = 0$, $\mathcal{T}_m = \mathcal{T}_{m-1} + T\max(||\hat{q}_x^n(\mathcal{T}_{m-1})||, M)$. Define, $m_n^* = \min\{m \geq 1 : ||\hat{q}_x^n(\mathcal{T}_m)|| \leq M\}$. and, $\hat{V}_n(x) = \mathbb{E}[\int_0^{\tau^n(\delta)}(1 + ||\hat{q}_x^n(t)||)dt]$. It follows that,

$$\hat{V}_n(x) \leq \mathbb{E}[\int_0^{\mathcal{T}_{m_n^*}} (1 + ||\hat{q}_x^n(t)||)dt] \tag{34}$$

$$= \sum_{k=0}^{\infty} \mathbb{E}[\int_{\mathcal{T}_k}^{\mathcal{T}_{k+1}} (1 + ||\hat{q}_x^n(t)||)dt\mathbf{1}_{\{k<m_n^*\}}]. \tag{35}$$

Define the filtration $\mathcal{F}_t$ as the sigma algebra generated by $\{\hat{q}_x^n(s) < 0 \leq s \leq t\}$. It can be shown (see Appendix F) that there exists a finite non negative constant c_0 such that, for all n, k, x,

$$\mathbb{E}[\int_{\mathcal{T}_k}^{\mathcal{T}_{k+1}} (1 + ||\hat{q}_x^n(t)||)dt|\mathcal{F}_{\mathcal{T}_k}]\mathbf{1}_{\{k<m_n^*\}} \tag{36}$$

$$\leq c_0(1 + ||\hat{q}_x^n(\mathcal{T}_k)||^2)\mathbf{1}_{\{k<m_n^*\}}. \tag{37}$$

Using this, one obtains the estimate,

$$\sup_n \hat{V}_n(x) \leq c_0 \sup_n \mathbb{E}[\sum_{k=0}^{m_n^*-1} (1 + ||\hat{q}_x^n(\mathcal{T}_k)||^2)]. \tag{38}$$

Observe that the Markov chain $\{\hat{q}_x^n(\mathcal{T}_m), m \geq 1\}$ has the single step transition kernel $P_n(x, A) = \hat{P}_n^{T\max(||x||,M)}(x, A)$, where $\hat{P}_n^t$ was the transition kernel of $\hat{q}^n$. Using (32) and (33), we have, for some $B \in (0, \infty)$,

$$sup_n \int_x P_n(x, dy)||y||^2 \leq ||x||^2 - \frac{||x||^2}{2} + B\mathbf{1}_{[1,M]}(||x||). \tag{39}$$

Using this in Lemma E.1, and using (38), we see that, for all x, $\sup_n \int_0^{\tau^n(\delta)}(1+||\hat{q}_x^n(t)||)dt \leq c(1+||x||^2)$. It can be shown [30] that there exists a positive $\kappa < \infty$ such that, for all t, x, n,

$$\frac{\mathbb{E}[\hat{V}_n(\hat{q}_x^n(t))]}{t} + \frac{\int_0^t \mathbb{E}(1 + ||\hat{q}_x^n(s)||)ds}{t} \leq \frac{\hat{V}_n(x)}{t} + \kappa. \tag{40}$$

Define the functions,

$$V_n^k(x) = \min(\hat{V}_n(x), k),\ \Gamma_n^k(x) = \frac{1}{t}(V_n^k(x) - \mathbb{E}[V_n^k(\hat{q}_x^n(t))]),$$

$$\Gamma_n(x) = \frac{1}{t}(\hat{V}_n(x) - \mathbb{E}[\hat{V}_n(\hat{q}_x^n(t))]).$$

Now, $\Gamma_n^k(x) \to \Gamma_n(x)$ as $k \to \infty$, by the monotone convergence theorem. Also, since π_n is the invariant measure of the n-th system, we have, $\int_x \Gamma_n^k(x)\pi_n(dx) = 0$. By Fatou's Lemma,

$$\int_x \Gamma_n(x)\pi_n(dx) \leq \liminf_{k\to\infty} \int_x \Gamma_n^k(x)\pi_n(dx) = 0. \tag{41}$$

If $\hat{V}_n(x) \leq k$, from (40), we know that, $\Gamma_n^k(x) \geq -\kappa$. If $\hat{V}_n(x) > k$, we have, $\Gamma_n^k(x) \geq 0$. Hence, $\Gamma_n^k(x) \geq -\kappa$ for all x. From (40), we can see that, $\Gamma_n(x) \geq \frac{\int_0^t \mathbb{E}(1+||\hat{q}_x^n(s)||)ds}{t} - \kappa$. Thus we obtain the bound,

$$\int_x \Gamma_n(x)\pi_n(dx) \geq \frac{\int_0^t \int_x \mathbb{E}(1 + ||\hat{q}_x^n(s)||)\pi_n(dx)ds}{t} - \kappa.$$

Combining with (41), and noting that the systems are assumed to be stationary, we obtain, $\int_x \mathbb{E}(1 + ||\hat{q}_x^n(t)||)\pi_n(dx) \leq \kappa$. Since π_n is the invariant measure for the n-th system, this is equivalent to,

$$\int_x (1 + ||x||)\pi_n(dx) \leq \kappa. \tag{42}$$

Let ϵ be fixed. Let $\mathbb{M} = \{x : ||x|| \leq M\}$, for some $M > \frac{\kappa}{\epsilon} - 1$. Then, $\int_{x\notin\mathbb{M}}(1 + ||x||)\pi_n(dx) \geq (1 + M)\pi_n(\mathbb{M}^c)$. Using (42), we have that, $\pi_n(\mathbb{M}^c) \leq \frac{\kappa}{1+M} < \epsilon$, by our choice of M. Since this is true for all n, it implies that the sequence of probability measures $\{\pi_n, n \geq 1\}$ is tight. □

Lemma V.3. *In our system model, conditions (29)-(31) hold. Further, there exists T such that (32) holds. Consequently, the sequence $\{\pi_n\}$ is tight.*

Proof. Since $\{\check{A}_i^{f,n}(t) - a_i^{f,n}(t), t \geq 0\}$ is a martingale, using Doob's inequality [37] we get,

$$\mathbb{E}[\sup_{0\leq k\leq t} |\check{A}_i^{f,n}(s) - \bar{a}_i^{f,n}(s)|^2] \leq B_1'\mathbb{E}|\check{A}_i^{f,n}(t) - \bar{a}_i^{f,n}(t)|^2$$

$$\leq B_1' t\mathbb{E}|\check{A}_i^{f,n}(1) - \bar{a}_i^{f,n}(1)|^2 = B_1 t,$$

where the second inequality follows from the i.i.d nature of the arrival process [38]. Hence, (29) holds. The bounds for $\check{R}$ and $\check{D}$ would hold if a corresponding bound holds for the $\check{S}_{ij}^f$ processes, which depend on both the queue state at time t, and the channel state at time t. Let C be the set of possible values $S(t)$ can take. Since $\mathcal{H}$ is finite (and consequently, C), there are only a finite set of mappings from $\mathcal{H}$ to C. This set of mappings will be denoted by $\{\mathbb{F}_1, \ldots, \mathbb{F}_{K_1}\}$. Each $S(Q(t), H(t))$ will take the value of one of these functions. It is easy to see that the state space of queues can be partitioned

as, $\mathcal{Q} = \cup_{m=1,\ldots,K_1} \mathcal{Q}_m$, where, if $Q(t) \in \mathcal{Q}_m$, we have $S(Q(t), H(t)) = \mathbb{F}_m(H(t))$, and the $\mathcal{Q}_m$ are disjoint. Now we can write,

$$\check{S}_{ij}^f(t) = \sum_{t'=1}^{t} \sum_{m=1}^{K_1} \mathbb{F}_m(H(t)) \mathbf{1}_{\{Q(t)=m\}}, \tag{43}$$

where $\mathbf{1}$ is the indicator function. Rewrite this as, $\check{S}_{ij}^f(t) = \sum_{m=1}^{K_1} \sum_{k \in \hat{T}_m(t)} \mathbb{F}_m(H(k))$, where $\hat{T}_m(t)$ is the set of time slots till t when the queue state was in $\mathcal{Q}_m$. Since the system is stationary, we can also obtain $s_{ij}^f(t) = \mathbb{E}[\check{S}_{ij}^f(t)]$. Thus, we may write, with $\bar{\mathbb{F}}_m = \mathbb{E}[\mathbb{F}_m(H(1))]$,

$$|\check{S}_{ij}^f(t) - s_{ij}^f(t)|^2 \leq B_2' \sum_{m=1}^{K_1} || \sum_{k \in \hat{T}_m(t)} \mathbb{F}_m(H(k)) - \bar{\mathbb{F}}_m||^2,$$

where B_2' depends only on K_1. For any m, along $k \in \hat{T}_m(t)$, $\mathbb{F}_m(H(k))$ is an i.i.d sequence. Therefore, proceeding similar to what was done for A, we now obtain,

$$\mathbb{E}[\sup_{0 \leq k \leq t} |\hat{S}_{ij}^f(k) - s_{ij}^f(k)|^2] \leq B_2 \mathbb{E}[\sum_m |\hat{T}_m(t)|] = B_2 t,$$

where the equality follows, since $\sum_m |\hat{T}_m(t)| = t$. Hence the bounds hold for $\hat{R}$ and $\hat{D}$ as well. Hence (29)-(31) hold, choosing $B = \max\{B_1, B_2\}$.

To show (32), observe that, since $n\hat{q}_i^{f,n}(t) = Q_i^{f,n}(0) + \check{A}_i^{f,n}(n^2t) + \check{R}_i^{f,n}(n^2t) - \check{D}_i^{f,n}(n^2t)$. Subtract the fluid queue $q_i^{f,n}(t')$ at time $t' = n^2t$ and use triangle inequality to get,

$$\begin{aligned} |Q_i^{f,n}(n^2t) - \bar{q}_i^{f,n}(n^2t)|^2 \leq C(|Q_i^{f,n}(0) - \bar{q}_i^{f,n}(0)|^2 \\ +|\check{A}_i^{f,n}(n^2t) - \bar{a}_i^{f,n}(n^2t)|^2 + |\check{R}_i^{f,n}(n^2t) - \bar{r}_i^{f,n}(n^2t)|^2 \\ +|\check{D}_i^{f,n}(n^2t) - \bar{d}_i^{f,n}(n^2t)|^2). \end{aligned}$$

Choosing $Q_i^{f,n}(0) = \bar{q}_i^{f,n}(0)$, we obtain, using (29)-(31), that $\mathbb{E}|Q_i^{f,n}(n^2t) - \bar{q}_i^{f,n}(n^2t)|^2 \leq C_2 n^2 t$, and hence it follows for the vector process Q, with a higher constant C_2', $\mathbb{E}||Q^n(n^2t) - \bar{q}^n(n^2t)||^2 \leq C_2' n^2 t$. From (28), since the draining time of the fluid system $\bar{q}^n$ with initial condition equal to one, $\tau_{drain}^n \leq nT_1$, the fluid system with initial condition x, will be zero at any time greater than $\tau_{drain}^n ||x||$. Setting $t \geq T_1||x||$, and dividing by n^2, we get,

$$\mathbb{E}|\hat{q}_x^n(t||x||)|^2 \leq C_2' t||x||. \tag{44}$$

Since the bound is uniform over n, dividing by $||x||^2$ and taking $||x|| \to \infty$ gives the result. □

With this result, we are ready to prove Theorem V.1.

Proof of Theorem V.1. Since the π_n are tight, any subsequence of π_n has a convergent subsequence. Let such a limit point be π^*. On the convergent subsequence, assume that the initial conditions $\hat{Z}^n(0)$ are distributed as π_n. Since the systems $\hat{Z}^n$ converge to a reflected Brownian motion (RBM), the initial condition of the RBM $\hat{w}$ will have distribution π^*. Also, we have shown that finite dimensional distributions of $\hat{z}^n$ also converge to that of $\hat{w}$. In particular, $\hat{z}^n(t)$ weakly converges to $\hat{w}(t)$ for any $t \geq 0$. But the distribution of $\hat{z}^n(t)$ is π_n. Thus distribution of $\hat{w}(t)$ is π^* for each t. Hence π^* is the stationary distribution of $\hat{w}$. □

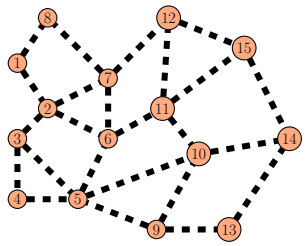

Fig. 3. Example Network for simulation

TABLE I
GETTING TARGETED END-TO-END MEAN DELAY FOR FLOW F2

Target Mean Delay $\overline{\tau}^{F2}$	Achieved Mean Delay for F2	Delay for F1	Delay for F3	Delay for F4	Delay for F5
∞	524.9	56.1	101.8	485.2	593.8
450	438.7	76.3	147.7	799.8	998.9
400	389.7	76.9	154.5	857.8	1066.0
350	348.5	72.2	141.1	724.1	905.0
300	301.9	75.2	148.1	797.6	1001.6
250	259.3	75.3	149.9	820.1	1024.0
220	237.6	74.4	149.0	805.9	1001.0

The Brownian motion $\hat{w}$ obtained as the limit of $\hat{w}^n$ is a unidimensional reflected Brownian motion, having drift $b^* < 0$. If $\hat{w}(\infty)$ has the stationary distribution of $\hat{w}$, from [35],

$$\mathbb{P}[\hat{w}(\infty) < y] = 1 - \exp(2b^* y/\sigma^2). \tag{45}$$

VI. Numerical Simulations

For simulations, we consider a 15 node network, with connectivity as depicted in Fig 3. In each slot, each link samples a channel state from the set $\{0, 1, 2, 3\}$ uniformly, independent of other links. On this network we have five flows, with the following paths F1: $2 \to 1 \to 8 \to 7 \to 12 \to 15$, F2: $12 \to 11 \to 10 \to 9 \to 5 \to 4$, F3: $2 \to 6 \to 11 \to 15 \to 14 \to 13$, F4: $7 \to 6 \to 5 \to 9 \to 13 \to 14$ and F5: $8 \to 1 \to 2 \to 3 \to 4 \to 5 \to 9$. For each flow, the arrival process at the source is Poisson distributed, with mean λ. The parameters a_1 and a_2 in (5) are both set equal to one. The set of activation vectors is chosen as a subset of the set of all activation vectors that satisfy the half-duplex constraint.

First, we demonstrate how we can deliver targeted end-to-end mean delay using the parameter $\overline{Q}^f$ of the function α in (4). Recall that for a flow with arrival rate λ^f, to obtain a target end-to-end mean delay of $\overline{\tau}^f$, we need to set $\overline{Q}^f = \lambda^f \overline{\tau}^f$. Thus, we set a target mean delay for flow F2. The results are provided in Table I. Let $\overline{\tau}^{F1}, \ldots, \overline{\tau}^{F5}$ denote the target mean delays of the five flows. We see the effect of varying the parameter $\overline{\tau}^{F2}$ alone, while $\overline{\tau}^f$ for the other flows is set to ∞ (this is equivalent to setting $\alpha = 1$). The arrival rate λ is fixed to be 0.37. The results are displayed in Table I. The first row displays the mean delays of five flows, with no targets (all $\overline{\tau}^f = \infty$). Without targets, F2 has a mean delay of 524.9 slots. We use $\overline{\tau}^{F2}$ to take the mean delay below this value. From 524.9, the mean delay can be brought down to around half that value, 237.6, by fixing a target of 220. Indeed, the delivered mean delay for this flow is close to the target, for

TABLE II
GETTING TARGETED END-TO-END MEAN DELAYS FOR FLOWS F4 AND F5

Target delay ($\overline{\tau}^{F4}$)	Achieved delay for F4	Target delay ($\overline{\tau}^{F5}$)	Achieved delay for F5
400	400.1	500	495.3
370	368.1	450	449.5
350	356.9	400	408.2
370	368.1	370	389.3
320	331.0	380	390.9
300	313.0	350	364.8

TABLE III
APPROXIMATION OF QUEUES. THE MEAN QUEUE LENGTH OF THE FLOW F2 IS COMPARED WITH THE UPPER AND LOWER BOUNDS.

Arrival Rate λ	Lower Bound	Mean Queue Length	Upper Bound
0.30	25.7	28.5	51.5
0.31	28.9	35.1	57.9
0.32	33.1	43.9	66.2
0.33	38.7	54.8	77.3
0.34	46.4	73.0	92.9
0.35	58.1	95.7	116.2
0.36	77.5	134.4	155.0
0.37	116.3	199.7	232.7
0.38	232.9	268.3	465.8

values from 520 till about 300. The delay cannot be reduced much below 237, owing to paucity of network resources. Note how the other flows lose out, since the resources are used up for providing priority service to F2. Correspondingly their delays blow up.

We can make multiple flows meet their delay targets simultaneously. Keeping the system parameters same as before, we now attempt to control the mean end-to-end delays corresponding to flows F4 and F5, simultaneously. These results are provided in Table II. We show the targeted mean delays and the delays obtained for these two flows. The delays of the other flows blow up as in the previous table. We have not included them in this table for ease of understanding. Observe that from the initial values of 485.2 and 593.8, we are able to bring down the delays to as low as 313 and 364, simultaneously. As in the previous table, the achieved mean delay tracks the targeted values very well. The amount of reduction we can bring to each flow is less than what we can do if we were controlling a single flow. This is because the algorithm is throughput optimal, and will not sacrifice stability. Consequently, the amount of 'tweaking' we can do is limited by the additional resources available to the system. These extra resources will get divided between the two flows, instead of one in the previous table.

Now we move on to the approximation of queue lengths by the limiting distribution obtained from the Brownian motion. From the diffusion approximation and (45), 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 time n, we may further also approximately write $b^* = n||\lambda - \lambda^*||$. Hence, we have the approximation,

$$Q(\infty) \approxeq \phi\frac{\sigma^2}{2||\lambda - \lambda^*||}. \tag{46}$$

Instead of approximating ϕ, which may not be straightforward, we observe that, $\frac{\psi}{2} \leq \phi \leq \psi$. From this we obtain upper and lower bounds for the queue length. We present the approximation values near the point where the arrival rates of all flows are equal to 0.39 (this is at the boundary of capacity). The value of σ^2 is $5\lambda + \hat{\sigma}^2$. We can approximate $\hat{\sigma}^2$ by 67.5. In Table III, we show the lower and upper bounds obtained from the approximation, for the flow F2. The component of ψ in the direction of F2 is approximately 0.3. The bounds track the queue length quite well, for a large range of values starting away from the boundary of the capacity region.

VII. 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 limit of the stationary distributions 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.

APPENDIX A
PROOF OF THEOREM III.1

We will use the following two results, the first from [15] and the second from [39].

Lemma A.1. *Let $\xi_n : [0,\infty) \to \mathbb{R}$ be a sequence of monotonically increasing functions. Let $\xi_n(x) \to \xi(x)$ for all rational x. If $\xi(x)$ is continuous, the convergence is u.o.c..*

Lemma A.2 (Helly's Selection Theorem)**.** *Let ξ_n be a sequence of monotonically increasing functions on $\mathbb{R}$, such that $0 \leq \xi_n(x) \leq B < \infty$, for all x and n. Then, there is a function ξ and a subsequence $\{n_k\}$ such that $\xi(x) = \lim_{n_k\to\infty}\xi_{n_k}(x)$.*

The Strong Law of Large Numbers (SLLN) in conjunction with Lemma A.1 gives (12). The family $\{\frac{1}{n}\check{S}_{ij}^f(nt)\}$ consists of monotone increasing functions, bounded by $\mu_{max}t$. Using Helly's selection theorem (Lemma A.2), one can obtain a convergent subsequence with limit s. Along this subsequence, $r_n \to r$ and $d_n \to d$ satisfying (13), due to (2).

Since the rates are bounded (owing to bounded channel gains and fixed power), it follows that $\check{S}_{ij}^f(t) \leq \mu_{max}t$. Therefore, for $0 \leq t_1 \leq t_2$, we have $\frac{\check{S}_{ij}^f(nt_2)}{n} - \frac{\check{S}_{ij}^f(nt_1)}{n} \leq \mu_{max}(t_2 - t_1)$. Along the subsequence along which $s^n \to s$, we have, $s_{ij}^f(t_2) - s_{ij}^f(t_1) \leq \mu_{max}(t_2 - t_1)$. It follows that s_{ij}^f is Lipschitz continuous, and hence so is s, and consequently r and d are Lipschitz as well. Hence, from Lemma A.1, we obtain u.o.c.

convergence for s^n, r^n and d^n along the chosen subsequence. Since s is Lipschitz and hence differentiable $a.e.$, (17) follows.

From (11), we can see that, $Q(nt) = Q(0) + \check{A}(nt) + \check{R}(nt) - \check{D}(nt)$. Dividing by n on both sides and taking $n \to \infty$ along the subsequence yields $q^n \to q$, with $q(t)$ satisfying (14). Since a, r and d are Lipschitz, q will also be Lipschitz, making it differentiable $a.e.$. At points where it is differentiable, we obtain (15) by differentiating (14).

The functions $\{\frac{1}{n}\check{G}^{hI}_{\mu}(nt)\}$ are also a monotone family, bounded uniformly on each compact interval. Hence, we can apply Helly's selection theorem to obtain a subsequence along which $g^n \to g$. Since $\frac{1}{n}(\check{G}^{hI}_{\hat{\mu}}(nt_2) - \check{G}^{hI}_{\hat{\mu}}(nt_1)) \leq t_2 - t_1$, for $t_2 > t_1$, g is Lipschitz; along this new subsequence $g^n \to g$ u.o.c.

To show (18), observe that, $\check{S}^f_{ij}(t) = \sum_{h,\varsigma,\hat{\mu}} \check{G}^{hI}_{\hat{\mu}}(t)\hat{\mu}^f_{ij}(h,\varsigma)$. Hence, we have, $\check{S}^f_{ij}(nt_2) - \check{S}^f_{ij}(nt_1) = \sum_{h,\varsigma,\hat{\mu}}(\check{G}^{h\varsigma}_{\hat{\mu}}(nt_2) - \check{G}^{h\varsigma}_{ijf}(nt_1))\hat{\mu}^f_{ij}(h,\varsigma)$. Multiplying LHS and RHS by $\alpha(\frac{Q^f(nt_1)}{n})\frac{Q^f_{ij}(nt_1)}{n}\frac{1}{n}$, summing over i, j, f, and taking $n \to \infty$, the LHS becomes

$$\sum_{i,j,f} \alpha(q^f(t_1))q^f_{ij}(t_1)[s^f_{ij}(t_2) - s^f_{ij}(t_1)], \tag{47}$$

where $q^f_{ij}(t) = \max(q^f_i(t) - q^f_j(t), 0)$ and $q^f(t_1) = \lim_{n\to\infty}\frac{Q^f(nt_1)}{n} = \sum_i q^f_i(t)$. The RHS becomes, $\sum_{i,j,f}\alpha(\frac{Q^f(nt_1)}{n})\frac{Q^f_{ij}(nt_1)}{n}\sum_{h,\varsigma,\hat{\mu}}(\frac{\check{G}^{h\varsigma}_{\hat{\mu}}(nt_2)}{n} - \frac{\check{G}^{h\varsigma}_{\hat{\mu}}(nt_1)}{n})\hat{\mu}^f_{ij}(h,\varsigma)$. The allocation satisfies, at every t where channel state is h,

$$\sum_{i,j,f}\alpha(\frac{Q^f(nt)}{n})\frac{Q^f_{ij}(nt)}{n}\hat{\mu}^f_{ij} \tag{48}$$

$$= \max_{\hat{\mu}\in\mathcal{U}(h)}\sum_{i,j,f}\alpha(\frac{Q^f(nt)}{n})\frac{Q^f_{ij}(nt)}{n}\hat{\mu}^f_{ij}. \tag{49}$$

Going along the subsequence along which $q^n \to q$, we obtain,

$$\sum_{i,j,f}\alpha(q^f(t))q^f_{ij}(t)\hat{\mu}^f_{ij}(h,\varsigma) = \max_{\hat{\mu}\in\mathcal{U}(h)}\sum_{i,j,f}\alpha(q^f(t))q^f_{ij}\hat{\mu}^f_{ij}. \tag{50}$$

Along the same subsequence, using (50), (16) and (12), this becomes

$$(t_2 - t_1)\sum_h \gamma_h \max_{\hat{\mu}\in\mathcal{U}(h)}\sum_{\hat{i},\hat{j},\hat{f}}\alpha(q^{\hat{f}}(t_1))q^{\hat{f}}_{\hat{i}\hat{j}}(t_1)\hat{\mu}^{\hat{f}}_{\hat{i}\hat{j}}. \tag{51}$$

Dividing (47) and the above by $t_2 - t_1$, equating, and taking $t_2 \to t_1$,

$$\sum_{i,j,f}\alpha(q^f(t_1))q^f_{ij}(t_1)\dot{s}^f_{ij}(t_1) \tag{52}$$

$$= \sum_h \gamma_h \max_{\hat{\mu}\in\mathcal{U}(h)}\sum_{i,\hat{j},\hat{f}}\alpha(q^{\hat{f}}(t_1))q^{\hat{f}}_{\hat{i}\hat{j}}(t_1)\hat{\mu}^{\hat{f}}_{\hat{i}\hat{j}}(h,\varsigma). \tag{53}$$

Since $\mathcal{W}$ is the convex hull of all points of the form $\sum_h \gamma_h\hat{\mu}(h)$ where $\hat{\mu}(h) \in \mathcal{U}(h)$, we obtain (18). The first part of (16) follows by applying the fluid scaling to (10). Since $||Q(0)|| = n$ for the n-th system, the second part of (16) follows.

Appendix B
Proof of Theorem III.2

We will be using the following result.

Theorem B.1. *(Theorem 4 of [16]) Let Y be a Markov Process with norm $||Y(.)||$. If there exist $\alpha > 0$ and a time $T > 0$ such that for a scaled sequence of processes $\{Y^n, n = 0, 1, 2, ..\}$, we have $\lim_{n\to\infty}\sup\mathbb{E}[||Y(n,T)||] \leq 1 - \alpha$, then the process Y is stable (positive recurrent).*

Pick an arrival rate $\lambda = \{\lambda^f_i\} \in int(\Lambda)$. Consider the Lyapunov function,

$$\mathcal{L}_1(q(t)) = -\int_t^{\infty}\exp(t-\tau)\sum_{i,f}\alpha(q^f(\tau))q^f_i(\tau)\dot{q}^f_i(\tau)d\tau, \tag{54}$$

where the dot indicates the derivative. This is a continuous function of $q(t)$, with $L(0) = 0$. Taking the derivaative of (14), we have, $\dot{\mathcal{L}}_1(q(t)) = \sum_{i,f}\alpha(q^f)q^f_i(\lambda^f_i + \sum_m \dot{s}^f_{mi}(t) - \sum_n \dot{s}^f_{in}(t))$. Since λ is in the interior of Λ, there exists a non negative vector $[\varpi^f_{ij}]_{(i,j)\in\mathcal{E},f\in\mathcal{F}}$ $s.t.$ $\lambda^f_i + \epsilon < \sum_j \varpi^f_{ij} - \sum_k \varpi^f_{ki}$ $\forall i, f$, and there exists $m \in \mathcal{M}$ such that $\sum_f \varpi^f_{ij} \leq m_{ij}$. Hence,

$$\dot{\mathcal{L}}_1(q(t)) < -\epsilon\sum_{i,f}\alpha(q^f)q^f_i + \sum_{i,f}\alpha(q^f)q^f_i(\sum_n \varpi^f_{in} - \sum_m \varpi^f_{mi} + \sum_m \dot{s}^f_{mi}(t) - \sum_n \dot{s}^f_{in}(t)).$$

Observing that $\sum_{i,f}\alpha(q^f)q^f_i(\sum_n \varpi^f_{in} - \sum_m \varpi^f_{mi}) = \sum_{i,j,f}\alpha(q^f)\varpi^f_{ij}(q^f_i - q^f_j)$, and that a similar equation holds for ϖ replaced by $\dot{s}$, it follows that if we show

$$\sum_{i,j,f}\alpha(q^f)\varpi^f_{ij}(q^f_i - q^f_j) \leq \sum_{i,j,f}\alpha(q^f)\dot{s}^f_{ij}(q^f_i - q^f_j), \tag{55}$$

it will imply $\dot{\mathcal{L}}_1(q(t)) < 0$. We have $\sum_{i,j,f}\alpha(q^f)\varpi^f_{ij}(q^f_i - q^f_j) \leq \sum_{i,j,f}\alpha(q^f)\varpi^f_{ij}q^f_{ij}$

$\leq \sum_{i,j,f}\alpha(q^f)\dot{s}^f_{ij}q^f_{ij}$, where the first inequality follows from the fact that $q^f_{ij} = (q^f_i - q^f_j)^+$, and the second from (18).

Now, if we show that $\dot{s}^f_{ij} = 0$ whenever $q^f_{ij} = 0$, (55) will follow. To see this, assume that at some t, $\dot{s}^f_{ij} = \delta^1 > 0$ and $q^f_{ij} = 0$. This would mean that for large enough n, there is a time s sufficiently close to t such that, for $\delta = \frac{\delta^1}{2}$, $S^f_{ij}(nt) - S^f_{ij}(ns) > n\delta(t-s)$. This implies that at a time $t_1 \in (s,t)$ with $Q^f_i(nt_1) - Q^f_j(nt_1) \leq 0$ the queue Q^f_i was served. This means that the optimization resulted in a positive μ^f_{ij}. This cannot happen when all Q^f_{ij} are zero, since in that state, by definition, all μ^f_{ij} are set to zero. Hence there exists k, l, m such that $Q^m_{kl} > 0$. If μ^f_{ij} is added to μ^m_{kl}, the value of the summand in (4) would only increase, thus contradicting its optimality. It follows that $\dot{s}^f_{ij} = 0$ whenever $q^f_{ij} = 0$, and hence, (55) is true.

Thus, $\dot{\mathcal{L}}_1(q(t)) < -\epsilon\sum_{i,f}\alpha(q^f)q^f_i$, and hence, $\mathcal{L}_1(q(t)) > 0$ whenever $q(t) \neq 0$. Fix $\delta_1 < 1$. There exists $T \leq T_1 = \frac{\mathcal{L}_1(q(0))}{\epsilon\,\delta_1} + \delta_1$ such that $\sum_{i,f} q^f_i \leq \delta_1$. To see

this, assume otherwise, that $\sum_{i,f} q_i^f(t) > \delta_1$ for $t \in [0, T_1]$. Now,

$$\mathcal{L}_1(q(t)) = \mathcal{L}_1(q(0)) + \int_0^t \dot{\mathcal{L}}_1(q(\tau))d\tau.$$

Since q is Lipschitz, $\dot{q}$ will be bounded. It is easy to see that $\mathcal{L}(q(0))$ is finite. Since $w(q^f) \geq 1$, $\mathcal{L}_1(q(t)) \leq \mathcal{L}_1(q(0)) - \epsilon\delta_1 t$, for $t \in [0, T_1]$, and by choosing $t = T_1$, we obtain $\mathcal{L}_1(q(T_1)) < 0$, which is a contradiction. Hence, $\sum_{i,f} q_i^f(T) \leq \delta_1$. Since the fluid queue follows the trajectory defined by equations (12)-(18), it follows that, almost surely, $\lim_{n\to\infty} \sup ||Q^n(T)|| = \sum_{i,j,f} q(T) \leq \delta_1 < 1$. From the definition of Q, we have that $||Q^n(T)|| \leq [1 + \sum_{i,f} \breve{A}_i^{f,n}(T) + T\sum_{i,j,f} \mu_{max}]$. Since $\mathbb{E}[\sum_{i,f} \breve{A}_i^{f,n}(T)] = T(\sum_{i,f} \lambda_i^f) < \infty$, we can use the Dominated Convergence Theorem [37] to see that Theorem B.1 holds for Q with $\alpha = 1 - \delta_1$. The result follows.

Appendix C
Proof of Lemma III.3

For any positive δ, as $n \to \infty$ along $\mathcal{N}_1$,

$$q(t) = \lim_{n\to\infty} \frac{Q(\lfloor \delta n t \rfloor)}{\delta n} = \frac{1}{\delta} q(\delta t). \tag{56}$$

Hence, a fluid limit path $q(t)$ is equivalent to a fluid path $\frac{q(\delta t)}{\delta}$. Define a fluid path $q^{'}(t) = q(t+T)$ for $t \geq 0$. This is a fluid path with initial condition, $|q^{'}(0)| \leq \delta_1$. From (56), $q^{'}(t) = q(t+T) = \frac{1}{\delta_1^{-1}} q(\delta_1^{-1}(t+T))$. If T_1 is the time for the path $q^{'}(t)$ to reach the level $(\delta_1)^2$, $|q(\delta_1^{-1}(T_1+T))| = \delta_1$. However, $|q(t)|$ reaches δ_1 in time T. Hence $|q(\delta_1^{-1}t)|$ reaches δ_1 in time $t = \delta_1 T$. Hence, $T_1 \leq \delta_1 T$. Thus we bound the time to reach δ_1, $(\delta_2)^2$, and so on by T_1, T_2, etc., where $T_n \leq (\delta_1)^n T$. Hence, the time for the queue to reach zero is bounded by, $T + \delta_1 T + (\delta_1)^2 T + \cdots = \frac{T}{1-\delta_1}$.

Appendix D
Properties of $\hat{\nu}$

We will show that, along the subsequence $\mathcal{N}_2$, we have a limit $\hat{\nu}$ of $\hat{\nu}^n$, for which the properties 1-4 of Theorem IV.4 hold. To study diffusion properties on an interval $[t_n, t_n + \delta]$ for $\delta > 0$, we look at fluid paths on the time $[nt_n, nt_n + n\delta]$. We consider the following family of fluid paths, started at a time T apart from each other. For a time evolving process $f(t)$, define the operator $\Theta(\tau)$ as the shift, corresponding to the process started at time τ. Consider the fluid scaled process z^n. Consider a shifted form of these processes, $\tilde{z}^{m,l} = \Theta(mt_m + Tl)z^m$, where $\Theta(x)f$ denotes the function f started at x. Define the family of processes, $\mathcal{Z} = \{\tilde{z}^{m,l(m)}, m \in \mathcal{N}_3\}$, where the index set $\mathcal{N}_3$ has the property that as $m \to \infty$ along $\mathcal{N}_3$, $t_m \to t$.

If $t_m \to t$, and $l(m) \in [0, 2\delta m/T - 1]$, a time $s \in [0, T]$ for the path $\tilde{z}(m, l(m))$, for m large enough, corresponds to a time, $s^{'} = t_m + l(m)T/m + s/m \in [t - 3\delta, t + 3\delta]^+$. We have the following results regarding the behaviour of the fluid sample paths. Here we follow [26].

Lemma D.1. *Consider the family $\mathcal{Z}$ with an associated sequence t_m, constants T and δ, both positive. Assume that $||\tilde{q}^{m,l(m)}|| \in [c_1, c_2]$, with $0 \leq c_1 \leq c_2 < \infty$, and $l(m) \in [0, 2\delta m/T - 1] \cap \mathbb{Z}$. Then, there is a subsequence m_k along which, $\tilde{z}^{m,l(m)} \to z$, u.o.c, with $||q(0)|| \in [c_1, c_2]$.*

The Lyapunov function $\mathcal{L}_1(q(t))$ defined in the proof of Lemma III.2. This function is non negative, finite and its time derivative is negative. If, along $q(t)$, if $\lim_{t\to\infty} \mathcal{L}_1(q(t)) = 0$, define $\mathcal{L}_3 = \mathcal{L}_1$. Else, if $\lim_{t\to\infty} \mathcal{L}_1(q(t)) = \mathcal{L}_* > 0$, define $\mathcal{L}_3(q(t)) = \frac{\mathcal{L}_1(q(t))}{\mathcal{L}_*} - 1$. Clearly, $\mathcal{L}_3(q(t))$ decreases to zero along any fluid path. We have the following result, with β a universal constant.

Lemma D.2. *Under our scheduling policy, assume that there is a subsequence such that, along this, $\hat{\nu}^n \to \hat{\nu}$. Suppose further that along this subsequence, we have $s_m \to s \geq 0$, $\hat{w}^m(s_m) \to K > 0$, $\limsup_{m\to\infty} ||\hat{q}^m(s_m)|| < K_1 K$, for some fixed $K_1 > 1$. Let $\delta > 0$ be chosen such that, $\epsilon = O_{\hat{u}}([s-3\delta, s+3\delta]^+) < 0.5K$, where $O_{\hat{u}}[a, b] = \sup_{x,y\in[a,b]} |u(x) - u(y)|$. Let $K_2 = \beta^2 K_1 K + 2\epsilon$. Then, for any $\epsilon_2 > 0$ sufficiently small, there exists a time T such that, for m sufficiently large, we have,*

$$K - 2\epsilon < \tilde{w}^{m,0}(u) < K_2, \text{ for } u \in [0, T], \tag{57}$$

$$(K - 2\epsilon)/\beta < ||\tilde{q}^{m,0}(u)|| < 2\beta K_2. \tag{58}$$

For $l \in [1, 2\delta r T^{-1} - 1] \cap \mathbb{Z}$, we have,

$$\mathcal{L}_3(\tilde{q}^{m.l}(0)) < 2\epsilon_2, \mathcal{L}_3(\tilde{q}^{m.l}(T)) < 2\epsilon_2, \mathcal{L}_3(\tilde{q}^{m.l}(u)) < 3\epsilon_2, \tag{59}$$

$$\text{for } u \in [0, T], \tag{60}$$

$$\tilde{v}^{m,l}(u) = \tilde{v}^{m,l}(u) - \tilde{v}^{m,l}(0) = 0, \text{ for } u \in [0, T], \tag{61}$$

$$K - 2\epsilon < \tilde{w}^{m,l}(u) < K_2, \text{ for } u \in [0, T], \tag{62}$$

$$(K - 2\epsilon)/\beta < ||\tilde{q}^{m,l}(u)|| < 2\beta K_2. \tag{63}$$

Proof of Lemma D.2. Since $\mathcal{L}_3$ is decreasing to zero, there exists T such that, $\mathcal{L}_3(t) \leq \epsilon_2$, $\forall t \geq T$. Consider the case $l = 0$. Observe that, for m large, we have $\limsup_{m\to\infty} \sup_{u\in[0,T]} ||\tilde{q}^{m,0}(u)|| < \beta \limsup_{m\to\infty} ||\tilde{q}^{m,0}(u)||$. This is true because, if it were not, using Lemma D.1, we could have a sequence of $\tilde{z}^{m,0}$ which converge to a fluid limit z with $||q(u)|| \geq \beta|q(0)|$ for some u. However, this is not possible since $\sup_{t\geq 0} ||q(t)|| < \beta||q(0)||$. This implies that,

$$\limsup_{m\to\infty} \sup_{u\in[0,T]} ||\tilde{q}^{m,0}(u)|| < \beta \limsup_{m\to\infty} ||\tilde{q}^{m,0}(u)|| < \beta K_1 K,$$

and $\limsup_{m\to\infty} \sup_{u\in[0,T]} \tilde{w}^{m,0}(u) < \beta^2 K_1 K$. By the non decreasing property of w, $\liminf_{m\to\infty} \inf_{u\in[0,T]} \tilde{w}^{m,0}(u) \geq K$. Choosing T large enough, we can have $\mathcal{L}_3(\tilde{q}^{m,0}(T)) < 2\epsilon_2$. Since $\tilde{q}^{m,0}(T) = \tilde{q}^{m,1}(0)$, it also follows that, $\mathcal{L}_3(\tilde{q}^{m,1}(0)) < 2\epsilon_2$. Consider the following properties, for $l \in [1, 2\delta m/T - 1]$.

$$\mathcal{L}_3(\tilde{q}^{m.l}(0)) < 2\epsilon_2, \ \mathcal{L}_3(\tilde{q}^{m.l}(T)) < 2\epsilon_2, \ \mathcal{L}_3(\tilde{q}^{m.l}(u)) < 3\epsilon_2, \tag{64}$$

$$\text{for } u \in [0, T], \tag{65}$$

$$\tilde{v}^{m,l}(u) = \tilde{v}^{m,l}(u) - \tilde{v}^{m,l}(0) = 0, \text{ for } u \in [0, T], \tag{66}$$

$$K - 2\epsilon < \tilde{w}^{m,l}(u) < K_2, \text{ for } u \in [0, T], \tag{67}$$

$$(K - 2\epsilon)/\beta < ||\tilde{q}^{m,l}(u)|| < 2\beta K_2. \tag{68}$$

We will show these hold, by induction on l. Assume the properties hold for all $l < l_1$, but at least one of the above properties is violated for $l = l_1$. Since the properties hold up to $l = l_1 - 1$, we have that $\mathcal{L}_3(\tilde{q}^{m,l_1}(0)) = \mathcal{L}_3(\tilde{q}^{m,l_1-1}(T)) < 2\epsilon_2$. Since w is non decreasing, we have, $\tilde{w}^{m,l_1}(0) > K - 2\epsilon$. From the relation between $||q||$ and w it follows that, $||\tilde{q}^{m,l_1}(0)|| \in \left[\frac{K-2\epsilon}{\beta}, 2\beta K_1\right]$. Thus, for a choice of T appropriately large, we will have,

$$\mathcal{L}_3(\tilde{q}^{m.l_1}(0)) < 2\epsilon_2, \mathcal{L}_3(\tilde{q}^{m.l_1}(T)) < 2\epsilon_2, \tag{69}$$

$$\mathcal{L}_3(\tilde{q}^{m.l_1}(u)) < 3\epsilon_2, \text{ for } u \in [0,T]. \tag{70}$$

To show the non-increasing property of $\tilde{v}$ as in (66), observe that the queue length and workload are strictly positive as shown above. Since we had $v^{m,l}(t) = x^{m,l}(t) - \langle \psi, d^{m,l}(t) - r^{m,l}(t)\rangle$, and since our optimization is such that we choose the allocation vector μ^* such that, $\mu^* = \arg_\mu \max \sum_{i,j,f} \alpha(q_i^f) q_{ij}^f \mu_{ij}^f$, $= \arg_\mu \max \sum_{i,j,f} \alpha(q_i^f)(q_i^f - q_j^f)\mu_{ij}^f$. The second equation holds because the allocation vector $\dot{s}_{ij}^f(t)$ is zero when $q_i^f - q_j^f \leq 0$. This optimization may be rewritten as a function of new variables $\tilde{\mu}$, where $\tilde{\mu}_i^f = \sum_j \mu_{ij}^f - \sum_k \mu_{ki}^f$. We have $\tilde{\mu}^*$ given by $\tilde{\mu}^* = \arg_{\tilde{\mu}} \max \sum_{i,j,f} \alpha(q_i^f) q_i^f \tilde{\mu}_i^f$. Since (70) holds, it will be that (choosing ϵ_2 small enough), this is exactly the result of the optimization, $\tilde{\mu}^* = \arg_{\tilde{\mu}} \max \sum_{i,j,f} \psi_i^f \tilde{\mu}_i^f$, since the function $\mathcal{L}_3$ indicates how close we are to the collapse vector ψ. From the definition of X, it follows that the scaled $\tilde{x}$ attains the value given above, and hence $\tilde{v}$ does not increase in the interval.

Since $\tilde{v}$ remains at zero, we can see that any increase in $\tilde{w}$ is an increase in $\tilde{u}$, and hence, $\tilde{w}^{m,l_1}(u) = \tilde{w}^{m,0}(T) + \tilde{u}^m(t_m + l_1 T/m + u/m) - \tilde{u}^m(t_m + T/m)$. Since the oscillation of $\hat{u}$ is bounded and since $\tilde{u}^m \to \hat{u}$, the bounds (in (68) also follow for l_1. Hence, we have inductively shown that the properties (60)-(63) hold. □

To obtain the properties of $\hat{v}$, we will require the following result, which is easy to obtain.

Lemma D.3. *Let $z^n = (a^n, e^n, g^n, d^n, r^n, s^n, q^n)$ be the fluid scaled process, with components $a^n = (a_i^{f,n})_{i,f}$ and $e^n = (e_h^n)_{h\in\mathcal{H}}$. Let $\mathcal{N}_1$ be an arbitrary subsequence of $\mathcal{N}$. Then, there exists a further subsequence $\mathcal{N}_2$ of $\mathcal{N}_1$, such that almost surely, as $n \to \infty$ along $\mathcal{N}_2$, the fluid scaled process satisfies, for any $T > 0$, for all $i, j, f, c \in \mathcal{H}$,*

$$\max_{0\leq \ell \leq nT} \sup_{0\leq \epsilon \leq 1} |a_i^{f,n}(\ell+\epsilon) - a_i^{f,n}(\ell) - \lambda_i^f \epsilon| \to 0,$$

$$\max_{0\leq \ell \leq nT} \sup_{0\leq \epsilon \leq 1} |e_c^n(\ell+\epsilon) - e_c^n(\ell) - \gamma_c \epsilon| \to 0.$$

First we show that $\hat{v}(t)$ is finite for all $t \in [0,\infty)$. Suppose this is not true. Then we will have $t^0 = \inf\{t \geq 0 : \hat{v}(t) = \infty\}$. Fix $\delta > 0$, and $\epsilon = O_{\hat{u}}[t - 4\delta, t + 4\delta]_+$. Choose $\Delta \in (0, \min(t,\delta))$ and $C > \hat{w}(t-\Delta) + 2\epsilon$. Define the sequence, $t_n = \min\{s \geq t - \Delta : \hat{w}(s) \geq C\}$. Since $\hat{v}$ is RCLL, and since $\hat{v}(t) = \infty$, it follows that, $\limsup_n t_n \leq t$. Also, $\limsup_n \hat{w}^n(t-\Delta) < C$. Now, in a small interval, the process $\hat{w}$ will not have jumps, since,

$$\hat{w}^n(t) - \hat{w}^n(t-) \leq \langle \psi, \hat{a}^n(t) - \hat{a}^n(t-)\rangle + \langle \psi, \hat{r}^n(t) - \hat{r}^n(t-)\rangle, \tag{71}$$

and because the process R is bounded by the i.i.d channel process H. Using Lemma D.3, the above quantity goes to zero. Hence it will follow that, as $n \to \infty$, $\hat{w}^n(t_n) \to C$. Choose a further subsequence along which, $t_n \to t' \in [t-\Delta, t]$. Along this, applying Lemma D.2, we see that $\hat{v}$ is finite on the interval $[0, t' + \delta]$. Thus we have a contradiction, and hence $\hat{v}$ is finite. Note that a similar construction can be done for $t = 0$ as well. The proof for continuity can also be done similarly, by finding point t which is a point of discontinuity. Choosing a suitable time before t, one can construct a sequence as before, which converges to a value C. Again, we will use Lemma D.2 to claim a contradiction. A similar proof holds for other properties of $\hat{v}$ as well.

Appendix E
Comparison Theorem from [40]

Lemma E.1. *For a Markov chain $\{X_k, k \geq 1\}$ with transition kernel P. Suppose there exist non negative functions $\Phi_1(x)$, $\Phi_2(x)$ and $\Phi_3(x)$ that satisfy, for all x, $\int_x P(x,dy)\Phi_1(y) \leq \Phi_1(x) - \Phi_2(x) + \Phi_3(x)$, then, for any stopping time $\mathcal{T}$, $\mathbb{E}_x[\sum_{k=0}^{\mathcal{T}-1} \Phi_2(X_k)] \leq \Phi_1(x) + \mathbb{E}_x[\sum_{k=0}^{\mathcal{T}-1} \Phi_3(X_k)]$.*

Appendix F
Proof of (36)

Due to the strong Markov property, it suffices to show that,

$$\mathbb{E}\int_0^{\mathcal{T}_1} (1 + ||\hat{q}_x^n(s)||ds) \leq c_0(1 + ||x||^2). \tag{72}$$

Observe that, $Q^n(n^2t) = x + q^n(n^2t) + A^n(n^2t) - a^n(n^2t) + R^n(n^2t) - r^n(n^2t) - D^n(n^2t) + d^n(n^2t)$, where, $q^n(t) = a^n(t) + r^n(t) - d^n(t)$, is the fluid limit corresponding to the n-th system. Thus, one obtains the inequality,

$$\mathbb{E}[\sup_{0\leq t\leq \mathcal{T}} ||Q^n(n^2t)||] \leq ||x|| + \mathbb{E}[\sup_{0\leq t\leq \mathcal{T}} ||q^n(n^2t)||] \tag{73}$$

$$+\mathbb{E}[\sup_{0\leq t\leq \mathcal{T}} ||A^n(n^2t) - a^n(n^2t)||] \tag{74}$$

$$+ \mathbb{E}[\sup_{0\leq t\leq \mathcal{T}} ||R^n(n^2t) - r^n(n^2t)||] \tag{75}$$

$$+ \mathbb{E}[\sup_{0\leq t\leq \mathcal{T}} ||D^n(n^2t) - d^n(n^2t)||]. \tag{76}$$

Since $\sup_{0\leq t\leq \mathcal{T}} q^n(n^2t) \leq \sup_t q^n(t)$, and the queue is non zero only till the draining time (given by (28)), and since the total input rate to a queue is bounded by the sum of all mean arrival rates and mean channel gains, it follows that there exists a constant c independent of t and n, such that, $\sup_t ||q^n(t)|| \leq ||x|| + nT_1$, where T_1 is from (28). For the process A (and similarly for R and D), we can see that, $\mathbb{E}[\sup_{0\leq t\leq \mathcal{T}} ||A^n(n^2t) - a^n(n^2t)||]$ is bounded by $\sqrt{\mathbb{E}[\sup_{0\leq t\leq \mathcal{T}} ||A^n(n^2t) - a^n(n^2t)||^2]}$, using Jensen's inequality. Combining these bounds, we see that, for some constant c_1 $\mathbb{E}[\sup_{0\leq t\leq \mathcal{T}} \hat{q}_x^n(t)] \leq c_1(1 + ||x|| + \mathcal{T})$. By definition, $\mathcal{T}_1 \leq c_2(1 + ||x||)$. Using these facts in (73), we obtain (72).

REFERENCES

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[2] A. P. Markopoulou, F. A. Tobagi, and M. J. Karam, "Assessment of voip quality over internet backbones," in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1. IEEE, 2002, pp. 150–159.

[3] T. Shah, A. Yavari, K. Mitra, S. Saguna, P. P. Jayaraman, F. Rabhi, and R. Ranjan, "Remote health care cyber-physical system: quality of service (qos) challenges and opportunities," *IET Cyber-Physical Systems: Theory & Applications*, vol. 1, no. 1, pp. 40–48, 2016.

[4] 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*, vol. 37, no. 12, pp. 1936–1948, 1992.

[5] L. Jiang and J. Walrand, "A distributed csma algorithm for throughput and utility maximization in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 18, no. 3, pp. 960–972, 2009.

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

[7] A. Krishnan and V. Sharma, "A distributed algorithm for quality-of-service provisioning in multihop networks," in *Communications (NCC), 2017 Twenty-third National Conference on*. IEEE, 2017, pp. 1–6.

[8] E. Stai, S. Papavassiliou, and J. S. Baras, "Performance-aware cross-layer design in wireless multihop networks via a weighted backpressure approach," *IEEE/ACM Transactions on Networking*, vol. 24, no. 1, pp. 245–258, 2016.

[9] R. Singh and P. Kumar, "Throughput optimal decentralized scheduling of multi-hop networks with end-to-end deadline constraints: Unreliable links," *IEEE Transactions on Automatic Control*, 2018.

[10] Y. Cui, E. M. Yeh, and R. Liu, "Enhancing the delay performance of dynamic backpressure algorithms," *IEEE/ACM Transactions on Networking (TON)*, vol. 24, no. 2, pp. 954–967, 2016.

[11] Y. Cui, V. K. Lau, R. Wang, H. Huang, and S. Zhang, "A survey on delay-aware resource control for wireless systems—large deviation theory, stochastic lyapunov drift, and distributed stochastic learning," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1677–1701, 2012.

[12] Q. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with qos support in wireless networks," *IEEE Transactions on vehicular Technology*, vol. 55, no. 3, pp. 839–847, 2006.

[13] S. V. Kumar and V. Sharma, "Joint routing, scheduling and power control providing hard deadline in wireless multihop networks," in *2017 Information Theory and Applications Workshop (ITA)*, San Diego, Feb. 12-17, 2017.

[14] A. N. Rybko and A. L. Stolyar, "Ergodicity of stochastic processes describing the operation of open queueing networks," *Problemy Peredachi Informatsii*, vol. 28, no. 3, pp. 3–26, 1992.

[15] J. G. Dai, "On positive harris recurrence of multiclass queueing networks: a unified approach via fluid limit models," *The Annals of Applied Probability*, pp. 49–77, 1995.

[16] 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*, vol. 18, no. 2, pp. 191–217, 2004.

[17] J. G. Dai and S. P. Meyn, "Stability and convergence of moments for multiclass queueing networks via fluid limit models," *IEEE Transactions on Automatic Control*, vol. 40, no. 11, pp. 1889–1904, 1995.

[18] S. Meyn, "Dynamic safety-stocks for asymptotic optimality in stochastic networks," *Queueing Systems*, vol. 50, no. 2, pp. 255–297, 2005.

[19] C. Maglaras, "Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality," *Annals of Applied Probability*, pp. 897–929, 2000.

[20] A. Krishnan and V. Sharma, "A distributed scheduling algorithm to provide quality-of-service in multihop wireless networks," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.

[21] ——, "Distributed control and quality-of-service in multihop wireless networks," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–7.

[22] B. Ji, C. Joo, and N. Shroff, "Throughput-optimal scheduling in multihop wireless networks without per-flow information," *IEEE/ACM Transactions On Networking*, vol. 21, no. 2, pp. 634–647, 2013.

[23] P. Billingsley, *Convergence of probability measures*. John Wiley & Sons, 1968.

[24] D. L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic. i," *Advances in Applied Probability*, vol. 2, no. 1, pp. 150–177, 1970.

[25] R. J. Williams, "Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse," *Queueing systems*, vol. 30, no. 1-2, pp. 27–88, 1998.

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

[27] M. Bramson, "State space collapse with application to heavy traffic limits for multiclass queueing networks," *Queueing Systems*, vol. 30, no. 1-2, pp. 89–140, 1998.

[28] A. Eryilmaz and R. Srikant, "Asymptotically tight steady-state queue length bounds implied by drift conditions," *Queueing Systems*, vol. 72, no. 3-4, pp. 311–359, 2012.

[29] D. Gamarnik and A. Zeevi, "Validity of heavy traffic steady-state approximations in generalized jackson networks," *The Annals of Applied Probability*, vol. 16, no. 1, pp. 56–90, 2006.

[30] A. Budhiraja and C. Lee, "Stationary distribution convergence for generalized jackson networks in heavy traffic," *Mathematics of Operations Research*, vol. 34, no. 1, pp. 45–56, 2009.

[31] H.-Q. Ye and D. D. Yao, "Diffusion limit of fair resource control-stationarity and interchange of limits," *Mathematics of Operations Research*, vol. 41, no. 4, pp. 1161–1207, 2016.

[32] A. Braverman, J. Dai, and M. Miyazawa, "Heavy traffic approximation for the stationary distribution of a generalized jackson network: The bar approach," *Stochastic Systems*, vol. 7, no. 1, pp. 143–196, 2017.

[33] R. T. Rockafellar, *Convex analysis*. Princeton university press, 1970.

[34] W. Whitt, *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer Science & Business Media, 2002.

[35] J. M. Harrison and M. I. Reiman, "Reflected brownian motion on an orthant," *The Annals of Probability*, pp. 302–308, 1981.

[36] W. Rudin, *Principles of mathematical analysis*. McGraw-hill New York, 1964.

[37] K. B. Athreya and S. N. Lahiri, *Measure theory and probability theory*. Springer Science & Business Media, 2006.

[38] A. Gut, *Stopped random walks*. Springer, 2009.

[39] I. P. Natanson, *Theory of functions of a real variable*. Frederick Ungar Publishing Co, New York, 1964.

[40] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.