Novel Resource Management Approach for End-to-End QoS Support in Wireless Mesh Networks

Ho Ting Cheng, Atef Abdrabou, and Weihua Zhuang

Department of Electrical and Computer Engineering, University of Waterloo, Canada

Email: {htcheng, alotfy, wzhuang}@bbcr.uwaterloo.ca

Abstract— **In this paper, we propose a novel end-to-end resource allocation approach for wireless mesh networks with quality-of-service (QoS) assurance. By introducing additional interference tolerability to each multimedia flow, our proposed approach is shown effective in fostering frequency reuse and increasing the number of multimedia flows supported in the system, outperforming its conventional resource allocation counterpart. Further, the proposed approach is of low complexity, leading to a preferred candidate for practical implementation.**

I. INTRODUCTION

Future wireless networks are expected to support a variety of applications, ranging from file sharing to voice communications to video streaming. To all intents and purposes, wireless mesh networking has emerged as a promising solution for future broadband wireless access [1,2], providing an easy and economical solution to Internet access. In specific, wireless mesh networks (WMNs) generally consist of gateways, mesh routers, and mesh clients, organized in a multi-tier hierarchical architecture [2]. Recently, the notion of this networking paradigm for suburban and rural residential areas has been attracting a plethora of attentions from academia and industry. In order for WMNs to be successful, efficient and effective radio resource management is indispensable.

In multi-hop wireless networks such as WMNs, end-to-end resource allocation is imperative to support multimedia services with diverse quality-of-service (QoS) demands [3]. In fact, a multimedia traffic flow initiated from a source node far away from its destination needs to forward its packets via multi-hop transmissions. For the sake of fine-grain QoS assurance, reserving sufficient network resources is vital at every link along the path of a multi-hop traffic flow, leading to the necessity of end-to-end resource reservation [4]. If powerful central controllers (such as base stations) are in place, resource allocation decisions for new and the existing multimedia flows in the system can be updated on the fly, where the notion of *frequency reuse* can be greatly fostered. Since austere suburban/rural environments discourage the setup of base stations [5], however, applying such a centralized approach directly to WMNs for suburban/rural residential areas can be inefficient or ineffective. Thus, distributed end-to-end resource allocation is imperative for WMNs with decentralized control.

In the literature, there exists a rich body of research work aiming at increasing throughput by admitting as many multimedia calls as possible in a distributed manner (e.g., [6,7]). With limited global information, conventional distributed approaches are to allocate a minimal amount of resources to new calls only, if admitted, without adjusting the resource allocation decisions previously made to the existing calls, striving for a balance between computational complexity and system performance [3]. Nevertheless, such a conventional methodology can discourage frequency reuse even in the presence of flexible medium access control (MAC)-layer resource allocation, thereby plausibly curbing system performance and resource utilization. Without extra interference tolerance, allocating only a minimum amount of resources to each incoming call can reduce its interference tolerability, thereby undermining the effectiveness of both frequency reuse and QoS support. Thus, a new resource management approach tailored for decentralized WMNs with QoS support and effective frequency reuse is needed.

On the other hand, recent studies show that the quantity of best-effort or background data traffic has been skyrocketing, constituting a substantial portion of the total traffic in WMNs [8]. In future broadband WMNs with abundant background data traffic, admitting more multimedia calls can, in fact, decrease the system throughput, for QoS provisioning and throughput maximization are conflicting performance measures [9]. As such, allocating resources strategically is imperative to not only increase the overall system throughput but also facilitate QoS support. In this paper, we devise a novel distributed resource management strategy for efficient WMNs with the consideration of end-to-end QoS assurance. In specific, we introduce the notion of rate cushions, whereby the interference tolerability of each multi-hop flow can be increased. With the proposed rate cushions in place, QoS support can be effectively facilitated while system performance can be greatly improved. Simulation results show that, by carefully adjusting the value of rate cushions, the proposed approach is effective in maximizing the number of QoS-sensitive calls to be supported in the system over its conventional counterpart without considering additional interference tolerability.

II. SYSTEM MODEL

We consider an orthogonal frequency division multiplexing (OFDM)-based synchronized WMN for a suburban or rural residential area, consisting of one wireline gateway attached to the Internet backbone and a number of mesh routers scattered around, rendering a multi-hop network (see Fig. 1). In specific, mesh routers mounted on the rooftops of the premises comprise a wireless mesh backbone, thereby providing an *all-wireless* environment. Mesh routers are assumed non-mobile and hence the channel gains can be estimated accurately. We further assume that each mesh node can operate in a full-duplex mode so that it can transmit and receive at the same time. To effectively facilitate bandwidth reservation and
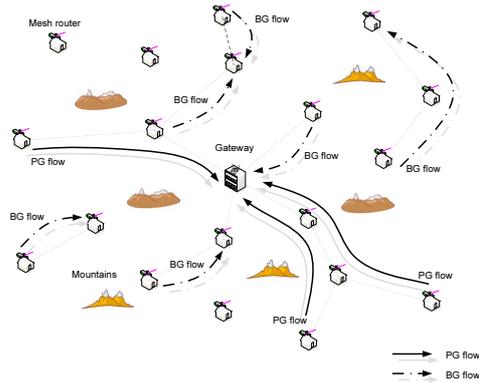
Fig. 1. An illustration of the WMN under consideration, with stationary mesh routers mounted on rooftops of a residential area and two classes of traffic flows (i.e., performance-guaranteed (PG) traffic and background (BG) traffic) traversing the WMN.

MAC-layer packet scheduling, time is partitioned into frames, each of which is further divided into a number of DATA slots [10].

Concerning the traffic flows[1] traversing the WMN of interest, two classes of traffic are considered, namely 1) performance-guaranteed (PG) traffic and 2) background (BG) traffic. In particular, the PG traffic has a minimum rate requirement and an end-to-end delay bound for information delivery, whereas the BG traffic has no QoS requirement. We further consider that PG call arrivals are random, while BG traffic flows are already in the system and can be transmitted whenever there are available resources. In addition, the PG (BG) traffic flows are assigned higher (lower) priority. Here, we consider that PG traffic traverses only from mesh routers to the gateway, whereas BG traffic generated from mesh routers can be disseminated in a peer-to-peer fashion. Using OFDM, simultaneous transmissions over different subcarriers are allowed in the network. In this work, we mainly focus on call-level resource allocation in WMNs. Once a PG call is admitted, a flow path for this PG traffic source is determined by our proposed resource allocation approach (discussed in Section IV-B), and the allocated resources are reserved for its packet transmissions during a call duration. As such, resource allocation for PG traffic and/or BG traffic is performed when there is a PG call arrival or PG call departure. In specific, our proposed approach allocates resources only to a new PG call, if admitted, and BG traffic without updating the resource allocation solutions for the PG calls already in service. On the other hand, if a PG call departs, our approach re-allocates resources to BG traffic only, yet without changing the resource allocation solutions for the other ongoing PG calls.

III. QoS-Aware End-to-End Resource Allocation

In this section, we formulate the problem of end-to-end resource allocation in WMNs with QoS assurance. First, we describe the system constraints associated with our problem. The sum of the non-negative transmit power levels of the

[1]In this paper, the traffic flows refer to multi-hop flows traversing the WMN under consideration.

(allocated) subcarriers at each link's transmitter is bounded by a maximum power level. On the other hand, subcarriers should be reused as much as possible so as to increase system capacity. Whether or not a subcarrier can be reused depends upon the interference incurred. A subcarrier can be reused and allocated to a new call if the QoS support of the existing calls can be maintained with the increased interference level; otherwise, the subcarrier cannot be reused. Without superposition coding, information of two or more traffic flows traversing the same link cannot be transmitted over the same subcarrier simultaneously. Concerning the PG traffic, as mentioned in Section II, each PG flow has its rate requirement. Thus, to satisfy the rate demand, the achievable rate on each link of the path along which a traffic flow traverses is at least its required rate. Here, we employ the notion of *effective bandwidth* [11] to characterize the rate demand of each PG call. In addition, each PG traffic flow has a delay requirement for end-to-end information delivery. Since a PG flow generally has more than one possible path to the destination (i.e., the gateway), here we only consider the set of possible paths of a PG traffic flow which meet its end-to-end delay bound requirement. With the above constraints, we employ the well-known utility maximization framework to abstract the objective[2] [12].

Problem Formulation: Let I, M, N, and L denote the number of PG traffic flows to be admitted, the number of links, the number of subcarriers available, and the number of timeslots (i.e., DATA slots) in a frame, respectively. Since the PG traffic is assigned higher priority, we first address the end-to-end resource allocation problem for the PG traffic flows, and then further optimize the system performance with the remaining resources, if any, for the BG traffic. Notice that we perform resource allocation for the PG traffic only when there is a new PG call arrival. For $1 \le m \le M$, $1 \le n \le N$, and $1 \le l \le L$, consider the following end-to-end resource allocation optimization problem (ERAOP)

$$\max_{I, \mathbf{c}, \mathbf{p}, \mathbf{q}} \left\{ \sum_{i=1}^{I} U_i \left(f_i \left(\mathbf{R}^{(i)} \right) \right) \right\} \tag{1}$$

$$\text{subject to} \quad \sum_{i=1}^{I} \sum_{n=1}^{N} {p^{(i)}}^l_{m,n} \le P_m^{\max}, \forall m, l \tag{2}$$

$$\sum_{i=1}^{I} {c^{(i)}}^l_{m,n} \le 1, \forall m, n, l \tag{3}$$

$$f_i \left(\mathbf{R}^{(i)} \right) > f_i^d, \forall i \tag{4}$$

$$\sum_{j=1}^{h\left(q^{(i)}\right)} {D^{(i)}}_j \le D^{\max(i)}, \forall i, q^{(i)} \tag{5}$$

$${p^{(i)}}^l_{m,n} \ge 0, \forall m, n, l, i \tag{6}$$

$${c^{(i)}}^l_{m,n} \in \{0, 1\}, \forall m, n, l, i \tag{7}$$

where ${p^{(i)}}^l_{m,n}$ is the transmit power of the mth link's transmitter over the nth subcarrier *dedicated to the ith flow* on the lth timeslot, ${c^{(i)}}^l_{m,n}$ is an indicator of allocating the nth subcarrier to the ith flow traversing the mth link on the lth timeslot, $P_m^{\max}$ is the maximum power constraint of the

[2]The objective function can be generalized to optimize system throughput, fairness, or a tradeoff among several system performance metrics (e.g., a tradeoff between throughput and fairness [9]).

m^{th} link's transmitter, $f_i\left(\mathbf{R}^{(i)}\right)$ is the achievable end-to-end rate of the ith flow, f_i^d is the minimum rate demand (i.e., effective bandwidth) of the ith flow, $D^{\max(i)}$ is the end-to-end delay bound of the ith flow, $D^{(i)}{}_j$ is the average delay that a packet of the ith flow experiences on the jth hop of the path $q^{(i)}$ with $q^{(i)}$ being an element of the set $\mathcal{Q}^{(i)}$ consisting of all the possible multi-hop paths for the ith flow, i.e., $q^{(i)} \in \mathcal{Q}^{(i)}$, $h\left(q^{(i)}\right)$ is the hop count of the ith flow over the path $q^{(i)}$, $U_i\left(f_i\left(\mathbf{R}^{(i)}\right)\right)$ is the utility function of the ith PG flow, $\mathbf{c} = \left[c^{(i)l}{}_{m,n}\right]_{M \times N \times L \times I}$, $\mathbf{p} = \left[p^{(i)l}{}_{m,n}\right]_{M \times N \times L \times I}$, and $\mathbf{q} = [q^{(i)}]_{I \times 1}$. The achievable end-to-end rate of the ith flow, $f_i\left(\mathbf{R}^{(i)}\right)$, is given by $f_i\left(\mathbf{R}^{(i)}\right) = \min_{m \in q(i)}\left\{R^{(i)}{}_m\right\}$, where $R^{(i)}{}_m$ is the achievable rate of the ith flow on the mth link over the path $q^{(i)}$, given by $R^{(i)}{}_m = \sum_{l=1}^{L}\sum_{n=1}^{N} c^{(i)l}{}_{m,n} r^{(i)l}{}_{m,n}$. Here, $r^{(i)l}{}_{m,n}$ is given by $r^{(i)l}{}_{m,n} = \ln\left(1 + \gamma^{(i)l}{}_{m,n}\right)$, where $\gamma^{(i)l}{}_{m,n} = \frac{\varphi G_{mm,n} p^{(i)l}{}_{m,n}}{\sigma \sum_{\forall u,k \backslash (i,m)} c^{(u)l}{}_{k,n} G_{mk,n} p^{(u)l}{}_{k,n} + \eta}$ with $G_{mk,n}$ being the channel gain from the kth link's transmitter to the mth link's receiver over the nth subcarrier, φ the bit-error-rate measure, σ the cross-correlation factor between any two signals, i.e., $\sigma \in [0, 1]$, and η the noise power. For simplicity, we assume that $D^{(i)}{}_j, \forall i, j$, is equal to the duration of a transmission frame in this work. The utility function of the ith (PG) flow, $U_i\left(f_i\left(\mathbf{R}^{(i)}\right)\right)$, is defined as

$$U_i\left(f_i\left(\mathbf{R}^{(i)}\right)\right)$$
$$= \begin{cases} f_i^d + \lambda \Delta_i, & f_i\left(\mathbf{R}^{(i)}\right) - f_i^d > \Delta_i \\ f_i^d + \lambda\left(f_i\left(\mathbf{R}^{(i)}\right) - f_i^d - \varepsilon\right), & \varepsilon < f_i\left(\mathbf{R}^{(i)}\right) - f_i^d \leq \Delta_i \\ \frac{f_i^d}{\log\left(1+f_i^d\right)} \log\left(1 + f_i(\mathbf{R}^{(i)})\right), & f_i\left(\mathbf{R}^{(i)}\right) - f_i^d \leq 0 \end{cases} \quad (8)$$

where $\lambda > 0$, $0 < \varepsilon \ll 1$, and Δ_i (≥ 0) is referred to as a *rate cushion* of the ith flow. Here, we view a utility function as a perceived satisfaction level or *satiation* [12]. Generally speaking, when $f_i\left(\mathbf{R}^{(i)}\right) < f_i^d$, the perceived satisfaction level of the ith PG flow is a concave increasing function of $f_i\left(\mathbf{R}^{(i)}\right)$. When $f_i\left(\mathbf{R}^{(i)}\right) > f_i^d + \Delta_i$, the rate demand of the ith PG flow is met, and hence allocating more resources does not further increase its perceived satisfaction level (i.e., satiated [12]). For $\varepsilon < f_i\left(\mathbf{R}^{(i)}\right) - f_i^d \leq \Delta_i$, the perceived satisfaction level of the ith PG flow increases as its interference tolerability increases against an increased interference level. For the sake of mathematical convenience, we assume that for $0 < f_i\left(\mathbf{R}^{(i)}\right) - f_i^d \leq \varepsilon$, $\frac{\partial U_i\left(f_i\left(\mathbf{R}^{(i)}\right)\right)}{\partial p^{(u)l}{}_{m,n}} \to -\infty$, for $u \neq i$. In other words, if its rate demand is only minimally satisfied, a flow cannot tolerate any additional interference, or its QoS support is void otherwise. Here, our objective function is to maximize the sum of all the utility functions. Notice that, in the ERAOP, $I, \mathbf{c}, \mathbf{v}$, and $\mathbf{q}$ are the optimization variables, while M, N, and L are the system parameters. By reducing the well-known NP-complete *number partitioning problem* [13] to the ERAOP, it can be proved that the ERAOP is NP-hard.

Proposition 1: The ERAOP is an NP-hard problem.

Proof: Omitted due to space limitation. ∎

After the network resources are allocated to the PG traffic flows, the allocation of remaining resources is performed for the BG traffic, thereby further improving the system

throughput. Let J ($\gg 1$) denote the number of BG traffic flows. By viewing BG traffic as *elastic* traffic [12], we set $U_j\left(f_j\left(\mathbf{R}^{(j)}\right)\right) = f_j\left(\mathbf{R}^{(j)}\right)$, where $U_j\left(f_j\left(\mathbf{R}^{(j)}\right)\right)$ is the utility function of the jth (BG) flow. Consider the following system throughput optimization problem (STOP) for the WMN

$$\max_{\check{\mathbf{c}}, \check{\mathbf{p}}} \left\{ \sum_{i=1}^{\check{I}} U_i\left(f_i\left(\mathbf{R}^{(i)}\right)\right) + \sum_{j=\check{I}+1}^{\check{I}+J} f_j\left(\mathbf{R}^{(j)}\right) \right\} \quad (9)$$

subject to

$$\sum_{j=\check{I}+1}^{\check{I}+J}\sum_{n=1}^{N} p^{(j)l}{}_{m,n} \leq P_m^{\max} - \sum_{i=1}^{\check{I}}\sum_{n=1}^{N} p^{*(i)l}{}_{m,n}, \forall m, l \quad (10)$$

$$\sum_{j=\check{I}+1}^{\check{I}+J} c^{(j)l}{}_{m,n} \leq 1 - \sum_{i=1}^{\check{I}} c^{*(i)l}{}_{m,n}, \forall m, n, l \quad (11)$$

$$p^{(j)l}{}_{m,n} \geq 0, \forall m, n, l, j \quad (12)$$

$$c^{(j)l}{}_{m,n} \in \{0, 1\}, \forall m, n, l, j \quad (13)$$

$$f_i\left(\mathbf{R}^{(i)}\right) > f_i^d, \forall i \quad (14)$$

where $\check{\mathbf{c}} = \left[c^{(j)l}{}_{m,n}\right]_{M \times N \times L \times J}$ and $\check{\mathbf{p}} = \left[p^{(j)l}{}_{m,n}\right]_{M \times N \times L \times J}$ are the optimization variables, while $\check{I}, p^{*(i)l}{}_{m,n}$, and $c^{*(i)l}{}_{m,n}$ are the system parameters. Note that $p^{*(i)l}{}_{m,n}$ and $c^{*(i)l}{}_{m,n}$ are the optimal solutions obtained from the ERAOP with $I = \check{I}$, where $0 \leq \check{I} \leq I^*$, with I^* being the maximal value of I. Concerning the STOP, it can be observed that the problems of CAC optimization (i.e., maximizing the number of PG calls admitted to the system and providing QoS support) and system throughput maximization are strongly coupled. As such, there is always a tradeoff between CAC optimization and throughput maximization, proved in Proposition 2.

Proposition 2: With a large value of J, the system throughput (i.e., $\sum_i U_i\left(f_i\left(\mathbf{R}^{(i)}\right)\right) + \sum_j f_j\left(\mathbf{R}^{(j)}\right)$) is a decreasing function of $\check{I}$.

Proof: Let $\mathcal{J}$ be a set of active BG traffic flows (i.e., the BG traffic flows with $f_j(\mathbf{R}^{(j)}) > 0$). Denote $\mathcal{J}_1$ be the set of active BG traffic flows for the STOP with $\check{I}_1$ and $\mathcal{J}_2$ the set of active BG traffic flows for the STOP with $\check{I}_2$, where $0 \leq \check{I}_1 < \check{I}_2 \leq I^*$. Let $\check{\mathbf{c}}_k$ and $\check{\mathbf{p}}_k$ be the optimal solutions obtained from the STOP with $\check{I}_k$, where $k = 1, 2$. When $\check{I}$ increases (decreases), the feasible region for $\check{\mathbf{c}}$ and $\check{\mathbf{p}}$ in the STOP shrinks (expands). For $\check{I}_1 < \check{I}_2$, the feasible region for $\check{\mathbf{c}}$ and $\check{\mathbf{p}}$ in the STOP with $\check{I}_2$ is only a proper subset of that with $\check{I}_1$. Thus,

$$\sum_{i=1}^{\check{I}_1} U_i\left(f_i\left(\mathbf{R}^{(i)}\right)\right) + \sum_{j \in \mathcal{J}_1} f_j\left(\mathbf{R}^{(j)}\right)_{|\check{\mathbf{c}}_1, \check{\mathbf{p}}_1} \geq$$
$$\sum_{i=1}^{\check{I}_2} U_i\left(f_i\left(\mathbf{R}^{(i)}\right)\right) + \sum_{j \in \mathcal{J}_2} f_j\left(\mathbf{R}^{(j)}\right)_{|\check{\mathbf{c}}_2, \check{\mathbf{p}}_2}. \quad (15)$$

Therefore, the system throughput does not increase with the value of $\check{I}$. On the other hand, for $\check{I}_1 < \check{I}_2$, consider

$$\sum_{i=1}^{\check{I}_2} U_i\left(f_i\left(\mathbf{R}^{(i)}\right)\right)_{|\check{\mathbf{c}}_2, \check{\mathbf{p}}_2} - \sum_{i=1}^{\check{I}_1} U_i\left(f_i\left(\mathbf{R}^{(i)}\right)\right)_{|\check{\mathbf{c}}_1, \check{\mathbf{p}}_1} = \sum_{i=\check{I}_1+1}^{\check{I}_2} f_i^d + \epsilon \quad (16)$$

where $\epsilon > 0$. As we consider a large J in our system model, with an enlarged feasible region, it is always possible to

achieve the following inequality:

$$\sum_{j \in \mathcal{J}_1} f_j\left(\mathbf{R}^{(j)}\right) - \sum_{j \in \mathcal{J}_2} f_j\left(\mathbf{R}^{(j)}\right) > \sum_{i=\check{I}_1+1}^{\check{I}_2} f_i^d + \epsilon. \quad (17)$$

As a result, the system throughput is a decreasing function of $\check{\imath}$. ∎

In a nutshell, maximizing the number of admitted PG flows and maximizing system throughput are two conflicting performance measures. Since the STOP is a rather standard optimization problem, many existing approaches suggested in the literature can be applied to solve the STOP. In this paper, our focus is to devise an efficient and effective QoS-assured resource allocation strategy to solve the ERAOP.

IV. PROPOSED APPROACH

A. KKT Interpretations

Generally, solving the NP-hard ERAOP requires exponential time complexity [12], and all the multimedia flows have to be considered at once. Without centralized control and global information, distributed resource management is highly preferred, whereby resources are allocated to each multimedia flow in a decentralized fashion on a link-by-link basis. On the other hand, some mesh links are more congested than the others in multi-hop WMNs. With the gateway as the destination of all PG traffic flows, the closer a mesh link to the gateway, the more likely it experiences traffic congestion. As such, we consider that links closer to the gateway have higher priority, whereas others farther away from the gateway have lower priority. Thus, for any multi-hop traffic flow, distributed resource allocation is performed in the ascending order with respect to the distance between a link and the gateway. Regarding traffic arrivals, calls generally come to the system one by one, and call admission should be performed in sequence. However, the need of considering each traffic flow individually can complicate our resource allocation problem. The complication stems from the fact that the rate demands and flow paths of future calls are unknown and, therefore, a frequency reuse pattern cannot be determined *a priori*. As mentioned in Section I, due to distributed control, resource allocation solutions for new calls and ongoing calls can barely be updated on the fly simultaneously. To design a low-complexity yet effective resource allocation approach to solve the ERAOP, we investigate the relationships among different optimization variables by means of the interpretations of KKT conditions [12]. By interpreting the KKT conditions with respect to $p^{(i)l}_{m,n}$, we propose a novel decentralized end-to-end resource allocation strategy, referred to as *bottleneck link-avoidance resource allocation*, taking the network characteristics, frequency reuse, and QoS assurance into account, to be discussed in Section IV-B. On one hand, we minimize the number of subcarriers used to satisfy the QoS requirements (i.e., end-to-end delay and rate requirements) of each incoming call. On the other hand, we allocate *extra* subcarriers to establish a rate cushion for each admitted call so as to increase its interference tolerability, defined in Section III. In essence, the introduced rate cushion protects the QoS of the ongoing traffic flows, facilitates subcarrier reuse for future incoming calls, and potentially increases the system capacity.

B. Bottleneck Link-Avoidance Resource Allocation

Here, the bottleneck link of a multi-hop traffic flow is defined as the link over which the flow traverses that gives the lowest achievable transmission rate. To identify the bottleneck link of the ith flow along the path $q^{(i)}$, choose b^* such that $b^* = \arg\min_{m \in q^{(i)}} \left\{ R^{(i)}_m \right\}$. Let $T^{(ui)l}_{m,n} = -\frac{\partial U'_u\left(f_u(\mathbf{R}^{(u)})\right)}{\partial p^{(i)l}_{m,n}}$, where $T^{(ui)l}_{m,n}$ can be interpreted as the marginal decrease in the utility obtained by the uth flow per unit increase in the transmit power of the mth link on which the ith flow traverses over the nth subcarrier on the lth timeslot. Thus, $T^{(ui)l}_{m,n}$ is always non-negative. $T^{(ui)l}_{m,n}$ can also be viewed as a *price* paid by the ith flow traversing the mth link for generating interference to the uth flow over the nth subcarrier on the lth timeslot. Consider the following payoff function of the ith flow traversing the mth link over the nth subcarrier on the lth timeslot

$$S^{(i)l}_{m,n}\left(c^{(i)l}_{m,n}, p^{(i)l}_{m,n}\right) = c^{(i)l}_{m,n} r^{(i)l}_{m,n} - p^{(i)l}_{m,n} \sum_{u \neq i} T^{(ui)l}_{m,n}. \quad (18)$$

Notice that, for some i, $T^{(ui)l}_{m,n}$ in (18) changes only when there is a call arrival or call departure. Essentially, each incoming flow is to maximize the difference between its throughput obtained over the mth link minus its total price paid to the other ongoing PG flows in service due to the induced interference. As mentioned in Section III, if the rate requirement of a flow, say the uth flow, is only minimally met, then $T^{(ui)l}_{m,n} \to \infty, \exists i$. As such, reusing the subcarriers already allocated to the uth flow to some other traffic flows is prohibited for the sake of QoS assurance. In other words, the gist of our proposed approach is to avoid generating too much interference to the bottleneck links of other ongoing multi-hop flows already in the system and to ensure their end-to-end QoS support. We also notice that, in (18), power allocation and subcarrier allocation should be jointly taken into consideration. For simplicity, however, we employ uniform power allocation and focus on subcarrier allocation, i.e., $p^{(i)l}_{m,n} = P^{\max}_m/N$. Thus, the decentralized subcarrier allocation condition can be deduced as follows. For the ith flow traversing the mth link on the lth timeslot, choose n^* such that

$$n^* = \arg\max_n \left\{ S^{(i)l}_{m,n}\left(c^{(i)l}_{m,n}, P^{\max}_m/N\right) \right\} \quad (19)$$

and set $c^{(i)l}_{m,n^*} = 1$. Since the gateway is the default destination node for the PG traffic, we consider that the price information (i.e., $T^{(ui)l}_{m,n}$) for each admitted PG call is stored at the gateway. Such information is to be provided to new PG call arrivals to determine their resource allocation solutions (i.e., $\mathbf{c}$ and $\mathbf{q}$) in a distributed fashion. Notice that employing the price information can also help facilitate the admission of new PG calls. In essence, a call admission routine is triggered whenever a new PG call arrives [4]. Each incoming call is admitted if a feasible solution of the ERAOP is found or rejected otherwise. In other words, the criterion of call

admission is contingent upon whether or not new PG calls can acquire a feasible resource allocation solution without voiding the QoS support of other existing PG calls. With such a CAC mechanism in place, the QoS requirements of all the admitted PG calls can be effectively guaranteed. Concerning path selection, in general, a shorter-path PG flow is likely to consume a smaller number of subcarriers to satisfy the QoS demands than a longer-path PG flow. Thus, our proposed approach is to start with the shortest path and search for the path such that the number of subcarriers employed is minimal. With the help of the payoff function given in (19), our proposed resource allocation approach for each call arrival is described as follows.

Step 1: Given the delay bound of an incoming PG call (referred to as the ith flow), $D^{\max(i)}$, the source node identifies a set of possible paths. All the possible paths are set to be unselected. Set $s = 1$.

Step 2: The unselected shortest possible path of the ith flow (in terms of the number of hops between the source node and the gateway) is chosen, denoted by $q(s)$. If no paths can be selected, go to Step 6.

Step 3: Resource allocation is performed in the ascending order with respect to the distance between a link on the path $q(s)$ and the gateway. For the link of interest, referred to as the mth link, the source node selects the best subcarriers according to our subcarrier selection criterion given in (19) until the rate condition for the mth link, $R^{(i)}_m \geq f^d_i + \Delta_i$, is satisfied.

Step 4: Let $N^{(i)}_{q(s)}$ denote the number of subcarriers used for the possible path $q(s)$, where $N^{(i)}_{q(s)} = \sum_{m \in q(s)} \sum_n \sum_l c^{(i)l}_{m,n}$ [3]. If the link rate conditions for all the links along the path $q(s)$ can be met, all the selected subcarriers are recorded by the ith flow. However, if any of the aforesaid link rate conditions cannot be met, remove the chosen path, and set $N^{(i)}_{q(s)} = 0$. Go back to Step 2.

Step 5: If $N^{(i)}_{q(s)} \geq N^{(i)}_{q(s-1)}$, the algorithm continues; otherwise, set $s = s + 1$, and go back to Step 2.

Step 6: If $N^{(i)}_{q(1)} = 0$, the incoming call is rejected; otherwise, the incoming call is accepted, and the source node chooses the path $q(s-1)$ for the ith flow for information delivery.

Our proposed resource allocation algorithm tries to find the path for each incoming call in which the number of subcarriers employed is minimal. The algorithm can be performed in a decentralized manner. Based on the price information of other existing PG flows, an incoming PG call can make its own resource allocation decision individually without any extra message exchange with other PG calls in service. As discussed in Section V-B, with the introduction of rate cushions, frequency reuse and hence system performance can be improved.

---

[3]For mathematical convenience, we set $N^{(i)}_{q(0)} \rightarrow \infty$. Note that $c^{(i)l}_{m,n} = 1$ means that the nth subcarrier over the mth link on the lth timeslot is selected and recorded by the ith flow, or $c^{(i)l}_{m,n} = 0$ otherwise.

C. Complexity

Since the admission process of each PG call is done individually in sequence, the time complexity of the subcarrier allocation for each link of a flow is on the order of $O(NLk)$, where k is a constant. Let $\bar{H} = \max_i \left\{ D^{\max(i)}/L \right\}$ and $\bar{\mathcal{Q}} = \max_i \left\{ |\mathcal{Q}^{(i)}| \right\}$ [4]. The time complexity of the proposed bottleneck link-avoidance resource allocation algorithm given in Section IV-B is on the order of $O(I\bar{\mathcal{Q}}\bar{H}NLk)$. Compared to an exhaustive search, our proposed approach is of low complexity, a desired feature for practical implementation.

V. PERFORMANCE EVALUATION

A. Simulation Environment

We consider an OFDM-based WMN with 16 nodes randomly located in a 500m x 500m coverage area. The maximum transmission rate of each subcarrier is 100kb/s. Other system parameters for performance evaluation are chosen as follows: $\varphi = 1$, $\sigma = 1$, $\eta \sim N(0, 10^{-12}\text{W})$, $P^{\max}_m = 1\text{W}$, $L = 4$, $f^d_i = f^d$, $\Delta_i = \Delta$, and the duration of each DATA slot is 5ms. We adopt the channel model suggested in [14]. Concerning the traffic models, PG packets are generated according to a two-state ON-OFF model. In the ON state, a fixed-size packet arrives in every 20ms with the link rate demand 32kb/s, whereas in the OFF state, no packet is generated. The both durations of an ON period and an OFF period follow an exponential distribution, where the mean ON period and the mean OFF period are 1.2s and 1s, respectively. The end-to-end delay bound of PG traffic is set to be 100ms, and the desired delay violation probability of PG packets is less than 1%. Applying the results obtained in [11], the effective bandwidth (i.e., rate demand) of each incoming PG call is 18.55kb/s (i.e., $f^d = 18.55$kb/s). On the other hand, the BG traffic does not have any QoS requirements, and can be transmitted whenever there are available resources. In our simulations, we consider that there is one saturated BG traffic source residing at every mesh router, while PG traffic calls arrive randomly to the system. Here, to solve the STOP, we employ the approach given in [15]. We perform the simulations for 2,000 runs and average the results, where each simulation run sustains 2,000 frames.

B. Simulation Results

We first evaluate the impact of Δ on the system performance of our proposed bottleneck link-avoidance resource allocation in terms of the maximum number of PG flows that can be supported and frequency reuse ratio. Fig. 2 shows the maximum number of PG flows supported versus Δ with different values of N. The standard deviations of the results are also plotted for reference. Notice that the results at the $\Delta/f^d = 0$ refer to the system performance of a conventional approach without considering extra interference tolerance. As expected, for the same value of N, the maximum number of PG flows that can be supported first goes up and then drops as the value of Δ/f^d increases. With the introduced rate cushions, the existing PG flows already in the system can tolerate additional interference when a new PG flow attempts

---

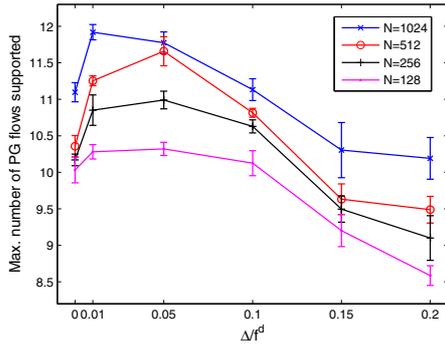[4]Note that both $\bar{\mathcal{Q}}$ and $\bar{H}$ are functions of the WMN topology and the number of links in the network.

Fig. 2. Maximum number of PG flows that can be supported in the system vs. the value of Δ/f^d.

Fig. 3. Frequency reuse ratio vs. the value of Δ/f^d.

TABLE I

RELATIONSHIP BETWEEN THE NUMBER OF ADMITTED PG FLOWS AND SYSTEM THROUGHPUT (WHERE $N = 512$ AND $\Delta/f^d = 0.05$)

Number of admitted PG calls	0	2	4	6	8	10
System throughput (in Mb/s)	19.71	17.60	16.74	16.34	15.93	15.78

to reuse some of the subcarriers. As a result, increasing the value of Δ/f^d can foster more frequency reuse, whereby more PG flows can be supported. At a large Δ/f^d, however, the system performance of our approach becomes worse than that of a conventional approach. The rationale for the performance degradation is that too many subcarriers are allocated to the existing PG flows, resulting in a poor resource allocation solution and hence a weaker system performance. On the other hand, it is clear that the maximum number of PG flows that can be supported increases with the number of subcarriers, N. The more the available subcarriers in the system, the more the PG flows can be supported for the same value of Δ/f^d. Another observation is that the relationship between the maximum number of PG flows supported and Δ changes with different values of N. For example, the maximum number of PG flows supported is attained at the $\Delta/f^d = 0.05$ for $N = 256$, whereas for $N = 1024$, the maximum number of PG flows supported at the $\Delta/f^d = 0.01$. Fig. 3 depicts the average frequency reuse ratio versus Δ/f^d. As seen, our proposed approach with $\Delta/f^d > 0$ attains a higher frequency reuse ratio than a conventional approach, thanks to the introduced rate cushions. One interesting phenomenon is that, from $\Delta/f^d = 0.05$ onward, the frequency reuse ratio remains roughly the same (see Fig. 3), but the maximum number of PG calls that can be supported plummets (see Fig. 2). That decline is based on the fact that a lot more than enough subcarriers are allocated to admitted PG flows, thereby causing a waste of network resources and hence reducing the number of PG calls supported in the system. Thus, the value of Δ should be chosen carefully in order to achieve an improved system performance; however, how to obtain the optimal Δ is left for further work. We then evaluate the impact of the number of admitted PG flows on the system throughput performance. For $N = 512$ and $\Delta/f^d = 0.05$, the relationship between system throughput and the number of admitted PG flows is given in Table I. We observe that the system throughput decreases with the number of admitted PG flows. The decline of system throughput is ascribed to the bandwidth reservation for PG traffic flows. As such, our results reinforce the idea that there is always a natural tradeoff between throughput optimization and CAC optimization (i.e., Proposition 2), yet finding a desired balance between them is
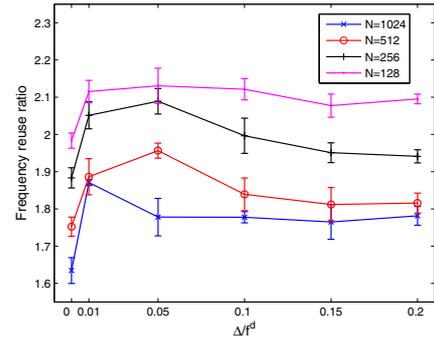
of great importance.

VI. CONCLUSIONS

In this paper, a novel QoS-aware end-to-end resource allocation approach is proposed, tailored for WMNs supporting heterogeneous traffic. Thanks to the introduced rate cushions, the proposed bottleneck link-avoidance resource allocation approach is demonstrated promising in terms of frequency reuse ratio and the maximum number of PG flows supported. Simulation results show that, by strategically adjusting the value of rate cushions, our approach can support more PG flows, outperforming its counterpart without considering interference tolerability. Our results also confirm that there is a natural tradeoff between throughput optimization and CAC optimization. Further, our approach is simple and can be performed in a decentralized fashion, leading to a viable candidate for practical implementation.

REFERENCES

[1] I. Akyildiz, X. Wang, and W. Wang, "Wireless mesh networks: a survey," *Computer Networks*, vol. 47, no. 4, pp. 445–487, Mar. 2005.

[2] H. T. Cheng, H. Jiang, and W. Zhuang, "Distributed medium access control for wireless mesh networks," *Wireless Communications and Mobile Computing*, vol. 6, no. 6, pp. 845–864, Sept. 2006.

[3] D. Camara and F. Filali, "Scheduling and call admission control: a WiMax mesh networks view," *Guide to Wireless Mesh Networks*, pp. 449–469, Feb. 2009.

[4] A. Abdrabou and W. Zhuang, "Statistical QoS routing for IEEE 802.11 multihop ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1542–1552, Mar. 2009.

[5] H. T. Cheng and W. Zhuang, "Pareto optimal resource management for wireless mesh networks with QoS assurance: joint node clustering and subcarrier allocation," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1573–1583, Mar. 2009.

[6] A. Abdrabou and W. Zhuang, "Stochastic delay guarantees and statistical call admission control for IEEE 802.11 single-hop ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3972–3981, Oct. 2008.

[7] W. Song, Y. Cheng, and W. Zhuang, "Improving voice and data services in cellular/WLAN integrated network by admission control," *IEEE Trans. Wireless Commun.*, vol. 6, no. 11, pp. 4025–4037, Nov. 2007.

[8] M. Afanasyev, T. Chen, G. M. Voelker, and A. C. Snoeren, "Analysis of a mixed-use urban WiFi network: when metropolitan becomes neapolitan," in *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement (IMC)*, 2008, pp. 85–98.

[9] H. T. Cheng and W. Zhuang, "An optimization framework for balancing throughput and fairness in wireless networks with QoS support," *IEEE Trans. Wireless Commun.*, vol. 7, no. 2, pp. 584–593, Feb. 2008.

[10] ——, "Novel packet-level resource allocation with effective QoS provisioning for wireless mesh networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 694–700, Feb. 2009.

[11] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, July 2003.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[13] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., 1979.

[14] IEEE 802.16 Broadband Wireless Access Working Group, "Channel models for fixed wireless applications," 2003. [Online]. Available: http://www.ieee802.org/16

[15] H. T. Cheng and W. Zhuang, "Joint power-frequency-time resource allocation in clustered wireless mesh networks," *IEEE Network*, vol. 22, no. 1, pp. 45–51, Jan.-Feb. 2008.