A Generic Admission-Control Methodology for Packet Networks

Chi Harold Liu, *Member, IEEE*, Kin K. Leung, *Fellow, IEEE* and Athanasios Gkelias, *Member, IEEE*

***Abstract*—Admission control (AC) is highly important in packet networks with quality-of-service (QoS) requirements since the uncontrolled admission of new data connections can jeopardize the QoS of existing connections and degrade the overall network performance. It requires *a priory* knowledge of the network capacity, the estimation of which is intricate and complex due to the operational characteristics of various communication protocols, complex network topologies, and dynamic traffic behavior with QoS requirements. In this work, we propose a generic admission control (GAC) methodology where the network conditions between any given ingress-egress node pair are summarized into a single parameter, referred to as the "QoS index". It accounts for various traffic volumes and QoS requirements, like throughput, delay, and packet error rate. Using this QoS index we track the network performance by predicting the potential impact of a new connection admission on the index. The decision depends on whether the predicted value lies below 1. We discuss its wide applicability and feasibility to many types of packet networks, wired or wireless, independent of what communication protocols in use at various layers. We finally validate the proposed GAC methodology by extensive simulations, confirming a significant improvement in terms of network goodput and QoS outage probability compared to other existing statistics-based AC methodologies.**

***Index Terms*—Admission-control, packet networks, QoS.**

I. INTRODUCTION

WITH the rapid development of Internet applications and wireless devices, networking has lately experienced unprecedented advances that have been pushing high-speed wired networking into new domains, making mobile and wireless networking much more ubiquitous, and driving the needs for all optical, 3G wireless, and quality-of-service (QoS)-based packet networks [1]. Moreover, the increase in processing power and memory availability of current user devices such as PDAs, game consoles, i-pad and laptops, give rise to a new wave of bandwidth-hungry and delay-sensitive mobile services and applications that will push the QoS demands to their limits or beyond. To satisfy the QoS of all connections in a packet network, we have identified the following three challenges when designing an efficient network architecture or network protocols.

The first challenge is the multi-dimensional QoS requirements of the data traffic. The question of how to support the multimedia traffic, like web browsing, Voice over IP (VoIP), interactive video, to name a few, is always central for the design of efficient network architecture or protocols like routing and scheduling, which are always associated with different combinations of QoS metrics to be enforced by the network. These metrics include but not limited to throughput, packet delay, packet-error-rate (PER), etc. It is proven NP complete in finding an optimal route satisfying more than one metric simultaneously [2], and thus challenging to accurately quantify the overall QoS experience for the completed connections and for the newly arrived connections which will be admitted into the packet network in the immediate future.

The second challenge is that the limited amount of network resources cannot allow an arbitrary large number of connections with strict and multiple QoS requirements admitted into the network, which can easily jeopardize the QoS experience for all ongoing connections in network and make the overall network inefficient and fragile. Furthermore, in wireless networks, due to the co-channel interference and channel fluctuations, the improper admission of the new connections can highly affect the resource availability of adjacent transmissions.

Finally, due to the complexity of network protocols in use, the stochastic nature of traffic and vastly different QoS requirements for various connections, it is very difficult to estimate the remaining network capacity (or the amount of available network resources) that can be used to admit new connections. This difficulty extends beyond estimating the remaining capacity of a single wireless network [3], that for the support of multimedia traffic, connections with different QoS requirements will consume different amount of network resources, making the network capacity highly dynamic and the estimation of the remaining resources extremely difficult.

All these challenges have motivated us to devise a new generic admission-control (GAC) methodology to enforce the QoS control in packet networks where we make the following three contributions:

First, we propose to aggregate multiple QoS metrics of a connection into one single performance index, called the "QoS index" for that connection. Using this index, the overall QoS for the completed and ongoing connections is quantified by a single scaler, despite multi-dimensional QoS requirements

Manuscript received November 23, 2012; revised September 16, 2013; accepted November 18, 2013. The associate editor coordinating the review of this paper and approving it for publication was S. Bahk.

This research was sponsored in part by the National Natural Science Foundation of China (No. 61300179). This research was also sponsored in part by the US Army Research Laboratory and the UK Ministry of Defence, and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Army Research Laboratory, the U.S. Government, the UK Ministry of Defense, or the UK Government. The US and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

C. H. Liu is with the School of Software at Beijing Institute of Technology, China (e-mail: chiliu@bit.edu.cn).

K. K. Leung and A. Gkelias are with the Department of Electrical and Electronic Engineering, Imperial College, London, United Kingdom (e-mail: {kin.leung, a.gkelias}@imperial.ac.uk).

Digital Object Identifier 10.1109/TWC.2013.010214.121856

for the connections. The index can change over time as the network provides services to the connections, and ranges from 0 to infinity, whereas the index value between 0 and 1 represents satisfaction of all QoS requirements for the completed and ongoing connections.

Second, all communication paths between a given pair of ingress and egress nodes in a network are treated as a "black box" of network resources available for sharing among connections from the ingress and egress nodes. Without considering the detailed network operations and protocols, the "performance" of the subnetwork is characterized by a general function that maps the connection parameters (inputs) such as the number of ongoing connections, their performance requirements and traffic information from the ingress to the egress nodes into the QoS index (output), as introduced above. The key idea here is to map the multiple performance parameters associated with the connections into a single scaler quantity to reflect the degree of QoS satisfaction of the connections. Specifically, when the QoS index is less than 1, it is certain that all QoS requirements are met. As expected, the mapping function is general and time-dependent, and not even its form is known. To overcome such difficulty, we propose to approximate the function by a Taylor expansion for which the unknown coefficients for the first several low-order terms can be properly estimated by realtime measurements as the network serves the connections. As a result, the approximate function is used to predict impacts on the QoS index by a new connection, if admitted, and thereby forming the basis for a new, generic framework for connection admission control.

Finally, we show that the proposed admission methodology has wide applicability to any packet networks, wired and/or wireless. It is completely transparent to layers of protocols in use, including the physical (PHY), medium access control (MAC) and network layers. We also discuss various important feasibility issues like the impact of large connection throughput requirement, measurement delay, and measurement collection time.

The rest of this paper is organized as follows. After introducing the related work in Section II, the system model and QoS index is described in Section III. Followed by the mathematical representation of a packet network from ingress to egress nodes in Section IV, Section V presents the admission impacts on the QoS index. Section VI presents the proposed GAC methodology. Next, numerical results and detailed analysis are given in Section VII. Finally, Section VIII presents the discussions on the applicability and the feasibility issues, and a conclusion is drawn in Section IX.

II. Related Work

A great deal of research attention has been paid to the topic of admission control (AC) algorithm in a variety of packet networks [4], [5], [6], [7], [8], [9], due to the growing popularity of multimedia applications (such as voice, video, and broadband data) and the central role AC scheme plays in QoS provisioning in terms of the connection quality, blocking probabilities, packet delay, and throughput, etc.

More specifically, many AC algorithms have been proposed for mobile ad-hoc wireless networks [10] and wireless mesh networks (WMNs, [11]). For example, [12] proposes an algorithm to support rate and delay requirements, but it assumed no channel fading and co-channel interference among wireless links, and uses a tree-structure MAC scheduling [13]. A distributed AC algorithm is proposed in [14] for each node to estimate the bandwidth used by its neighbors. A set of papers in [15], [16], [17] study the design of optimal joint AC control and routing that can maximize the overall "revenue" while guaranteeing the QoS for multiple classes in mesh networks has not been addressed, where the AC problem is formulated as a semi-Markov decision process (SMDP) and solved by a linear programming (LP) based algorithm. Unfortunately, the notion of QoS is only denoted as the SNR constraint. In [18], an adaptive admission control (AAC) protocol estimates the resource availability in contention-based WLAN MAC layer to control QoS. However, neither provides a degree of transparency to lower protocol layers nor the guarantee of multi-dimensional QoS requirements. This work is followed by a similar approach in [19] and [20]. In [21], a joint centralized scheduling and time-slot-allocation-based AC algorithm is proposed for WiMAX networks, which allows to admit a connection if extra unused slots are sufficient to satisfy bandwidth requirement. The integrated framework of routing and AC for IEEE 802.16 distributed mesh networks is studied in [22]. It estimates available bandwidth in a token bucket to perform AC with the minimum time-slot requirement for each connection, and uses the shortest-widest efficient bandwidth metric for route discovery. [23] makes admission decision by estimating the achievable capacity between any pair of ingress and egress nodes with only packet loss constraint, assuming that traffic arrives according to Gaussian distribution. Finally, work [24] studies extensively on the integrated QoS routing protocol and the actual interface between the scheduling and routing schemes, to provide the optimal routes that guarantee multiple QoS constraints.

In a summary, the existing AC algorithms in the literature do assume certain operational characteristics of the underlying communication protocols in use and AC schemes are developed for specific network settings. Furthermore, none of these techniques properly estimate and quantify the impact of a connection admission in terms of QoS experience to existing connections, nor are they widely applicable to other packet network settings. These are the central issues to be addressed in this paper.

III. System Model

Consider a generic packet network that comprises a set of subnetworks. Each subnetwork can be an access network, backhaul network or backbone network. The packet network is composed of a finite number of nodes. Let $V_I = \{v_i | i = 1, 2, \ldots, N_I\}$ and $V_E = \{v_e | e = 1, 2, \ldots, N_E\}$ denote the sets of N_I ingress and N_E egress nodes, respectively, and let us focus on a specific pair of ingress and egress nodes. Data traffic generated in the application layer of a user device, reaches the ingress node, and intends to be transferred to the egress node as shown in Fig. 1. Without loss of generality, assume that each connection $q \in \mathcal{Q}$ (where $\mathcal{Q}$ denotes the set of connections currently being served from the specific pair of ingress to egress nodes) has a set of QoS performance

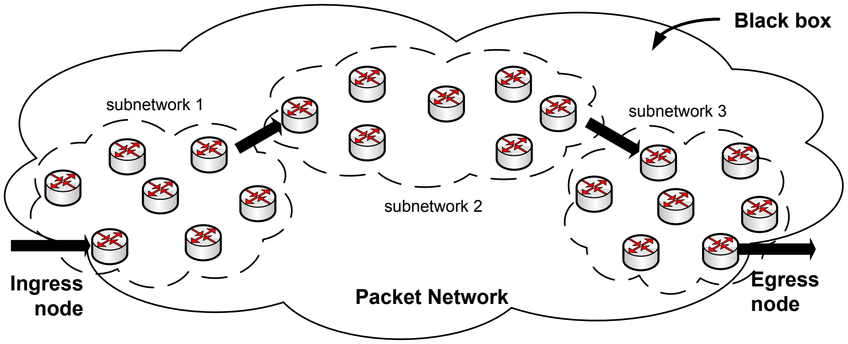


Fig. 1. An illustrative example of a packet network consisting of three separate subnetworks.

requirements. In practice, most considered QoS performance requirements generally fall into three categories, addictive, concave, and multiplicative constraints. Without loss of generality, here we pick one from each category, and they are the packet delay D (addictive constraint for multiple hops), throughput T (concave constraint as the bottleneck of the end-to-end route), PER E (multiplicative constraint for multiple hops). For connection q, we denote its required QoS value as packet delay D_q^r, throughput T_q^r, and PER E_q^r, where the superscript r denotes the *required* QoS performance value.

Concurrent connections share a limited amount of network resources, such as buffer, bandwidth, transmission power, time slot, etc. To ensure QoS for ongoing connections from the given ingress to egress nodes, we must consider all connections that share any resources along the paths between the two nodes. For this reason, all communication paths between the ingress and egress nodes are treated as a "black box" of network resources. Without considering the detailed network operations and protocols, the "performance" of the packet network is characterized by a general function that maps the input parameters such as the number of ongoing connections from the ingress to the egress nodes, their performance requirements and other traffic information into a single quantity referred to as the QoS index. As discussed below, the index can adequately reveal the degree of satisfaction of QoS requirements for all connections from the ingress to the egress nodes.

A. QoS Outage Ratio for Each Connection

The key for admission control is to identify how a new connection, if accepted, will experience and impact the QoS of existing connections from the ingress to the egress nodes. To this end, a unique QoS utility function based on the *QoS outage ratio*, denoted by R, is defined for each considered QoS parameter, i.e., R_q^D for packet delay, R_q^T for throughput, and R_q^E for PER, $\forall q \in \mathcal{Q}$. These ratios are defined as the ratio between the *attained* performance by measurements (denoted by a superscript a), and the *required* QoS value (denoted by a superscript r). That is,

$$R_q^D = \beta_D \frac{D_q^a}{D_q^r}, \quad R_q^T = \beta_T \frac{T_q^r}{T_q^a}, \quad R_q^E = \beta_E \frac{E_q^a}{E_q^r}, \quad \forall q \in \mathcal{Q}, \tag{1}$$

where $\{\beta_D, \beta_T, \beta_E\} \geq 1$ are a set of *error margins* introduced for delay, throughput, and PER, respectively, to safeguard against imperfect estimations and system fluctuations. Note that the QoS requirements for any given connection q are satisfied if the corresponding QoS outage ratios are less than 1. It is understood that the attained performance values (parameters) in (1) can be readily obtained by time-stamping packets associated with a connection and tracking at the egress node.

B. QoS Index for Each Connection

Due to the multi-dimensional nature of QoS requirements, despite the use a set of QoS outage ratios to denote the degree of satisfaction for each performance metric, it is still difficult to judge the *overall* QoS experience any connection has received.

Definition 1: **QoS Index for Each Connection:** given the multi-dimensional QoS requirements of each connection, user experience is satisfactory if and only if every QoS performance requirement is met; and QoS index for each connection $\mathrm{I}(q), \forall q \in \mathcal{Q}$ is defined as a scaler $\mathrm{I}(q) = g\left(R_q^D, R_q^T, R_q^E\right)$ to consider all QoS outage ratios associated with a connection.

Mathematically, many function definition $g(\cdot)$ can be applied here. However, for our purpose of aggregating the multi-dimensional QoS performance parameters into a single scaler quantity to facilitate efficient admission decisions, we specifically use $g(\cdot) \triangleq \max(\cdot)$ as an illustrative example as follows:

$$\begin{aligned} \mathrm{I}(q) &= g\left(R_q^D, R_q^T, R_q^E\right) \\ &\triangleq \max\left(R_q^D, R_q^T, R_q^E\right), \quad \forall q \in \mathcal{Q}. \end{aligned} \tag{2}$$

The reason why we choose $\max$ operator is as follows. Higher R indicates the higher degree of performance dissatisfaction with a particular QoS parameter (i.e., it captures the percentage of performance deviation from the required value). Then, if we take the maximum value of all considered QoS outage ratios, it clearly provides a maximum degree of performance dissatisfaction if all QoS parameters are considered simultaneously. Therefore, it follows immediately from the definition of QoS index for a connection that: for any connection $q \in \mathcal{Q}$, its multi-dimensional QoS requirements are simultaneously satisfied within the network if and only if $\mathrm{I}(q) \in [0, 1]$.

The main advantages of using this index are twofold. First, it combines different heterogeneous performance parameters and their requirements into one single quantity without considering the details of network operations and protocols in use. Second, although simple, it will be shown in the following that the index adequately reflects the degree of resource availability for maintaining the required QoS performance in realtime.

The egress node is assumed to constantly monitor traffic flows and measurements associated with each connection q. Therefore, it is able to compute the QoS outage ratios in (1), and the corresponding QoS index for each connection in (2).

Finally, it is worth to point out that different from existing approaches to compute the aggregated statistics (e.g., throughput) for *each link* in a network, which generally needs detailed lower layer information like topology, routing and resource allocation results (which can be quite dynamic and difficult to obtain in real-time), here we opt to only compute the "degree of QoS satisfaction" from the ingress to egress node, captured by the QoS index for each connection as above, irrespective what resource the connection is consuming (e.g., which route(s) the connection is taken, the time slot allocation etc.). Then, the collective effects of multiple connections on the considered network can also be sufficiently reflected by the operator g that takes into account all connections $\mathcal{Q}$ being served simultaneously.

IV. Mathematical Representation of Packet Network From Ingress To Egress Nodes

All resources in the network from the given ingress to the egress nodes are utilized and shared by connections between the two nodes (referred to as the main connections) and other connections originating and terminating elsewhere (referred to as the cross connections). At this point, let us assume that the number of the *cross connections* remains fixed, although their traffic and thus its resource utilization fluctuate stochastically. Clearly, as additional connections and traffic are admitted and served from the ingress to the egress, their *overall* QoS experience, as reflected by the individual QoS index for each connection, will start to deteriorate. For a network with a fixed amount of resources available, the QoS index is expected to depend on a set of "connection parameters", defined as follows.

Definition 2: **Connection Parameters:** For any given pair of ingress and egress nodes, they are associated with a set of *connection parameters* that characterize the traffic information and their performance requirements.
Examples of connection parameters are the total number of connections N being serviced from the ingress to egress nodes, the total throughput requirements of these connections T, the packet delay requirement D, and the PER requirement E.

From this perspective, it is appropriate to use a general mathematical function to relate the QoS index in terms of the connection parameters. Let such a function be denoted by f, which maps an M-dimensional vector $\underline{x} = (x_1, \ldots, x_i, \ldots, x_M) \in \mathbb{R}^M$ to the index $\mathrm{I} \in \mathbb{R}$ (defined as the QoS index for all completed connections under the same connection parameters, or simply "QoS index" in the next section), which reflects the degree of resource occupancy within the packet network. In other words, the mapping is

$$f : \mathbb{R}^M \to \mathbb{R}, \text{ or } \mathrm{I} = f(\underline{x}), \tag{3}$$

where there are M input variables (connection parameters) $\underline{x}$ that captures the system status.

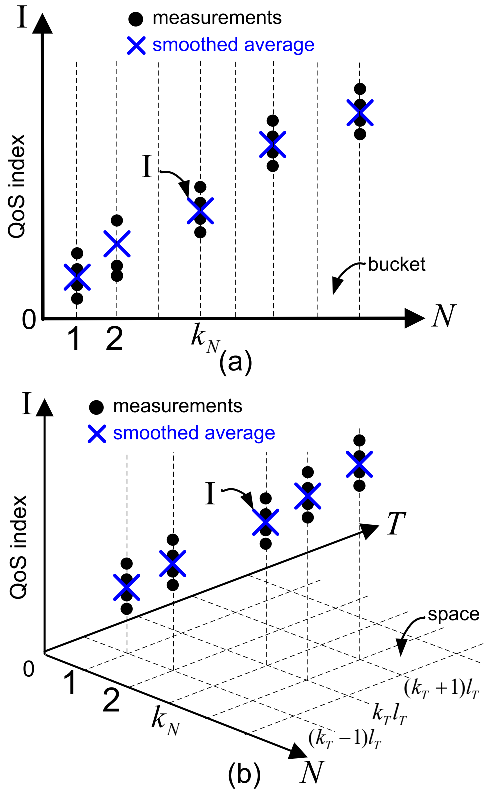


Fig. 2. An illustrative example for the egress node to collect and smooth measurements to produce a time average. (a) 1-D case, (b) 2-D case.

A. Measurements Collection and Smoothing

From the definition of connection parameters, some of them are discrete while others are continuous. For the latter, we discretize each connection parameter into "buckets" without border effects. Then x_i is rewritten as a joining set of buckets as $x_i \triangleq \left\{ \left[k_i l_i, (k_i+1) l_i \right) \right\}, \forall k_i \in \mathbb{Z}, i = 1, \ldots, M$, where k_i and l_i denote the bucket index of the i-th dimension, and the length of a bucket, respectively. We use the left-hand side value $k_i l_i$ to represent all values falling into the bucket, as the only discrete point to complete the discretization process. For example, considering aggregated throughput requirement T as the connection parameter, l_T can be $50 \sim 100$kbps for VoIP traffic, and $200 \sim 500$kbps for video connections. For VoIP, the discretized throughput dimension is thus represented by a set of points, *e.g.,* 0kbps, 50kbps, 100kbps, ...

We next illustrate the process of collecting measurements, and start from one-dimensional case. We use the discrete connection parameter, *i.e.,* the total number of connections being serviced N, for illustration purposes, as shown in Fig. 2(a). In the figure, we show all measured QoS index for each connection (black dots) according to their associated number of connections concurrently running in the network.

For connection parameters of M-dimensions, we apply the same technique descretizing each dimension (of connection parameter) into buckets, and then by combining any bucket $\left[k_i l_i, (k_i+1) l_i \right)$ of dimension i together, one is able to construct an M-dimensional Euclidean space, denoted by

$\Omega \in \mathbb{R}^M$, composed of a set of $2M$ vertexes, as $V \triangleq \Big\{k_i l_i, (k_i + 1) l_i\Big\}, \forall i = 1, \ldots, M$. Similar to the 1-D case, we use the set of left-most side values $\Big\{k_i l_i\Big\}$ to represent all values fall into the space as the only set of discrete points (of M dimension).

For instance, considering the 2-D connection parameters $(N, T) \in \mathbb{R}^2$, each 2-D space (or *area* in this case) Ω is defined by a set of four vertexes: $V \triangleq \Big\{k_N, k_N + 1, k_T l_T, (k_T + 1) l_T\Big\}, \forall k_N, k_T \in \mathbb{Z}$. Clearly, V are the grid points. Fig. 2(b) shows an illustrative example to quantize the entire area formed by connection parameters N, T into grid points. In the figure, we also show the collected QoS index for each connection (black dots), according to their associated concurrent number of running connections and aggregated (and discretized) throughput requirement.

The idea of discretizing each dimension of the connection parameter is to simplify the aggregation complexity of all measured QoS index values for each connection into one single QoS index under the same connection parameters, or simply the *QoS index* used in the rest of the paper. It is defined as follows.

Definition 3: **QoS Index:** for a given pair of ingress and egress nodes, the QoS index is the smoothed value of all connections, if their associated connection parameters during service from the ingress to egress nodes are identical.

An example is to consider (N, T) as connection parameters, and QoS index values for each connection is smoothed if they are running with the same number of concurrent connections and total required (and discretized) throughput requirement. Note that the difference between Definition 1 and 3 is that the former characterizes the received QoS experience of *one single connection*, while the latter represents the *overall QoS experience* under the same set of connection parameters.

Consider that the egress node computes QoS indexes $\mathrm{I}(q), \forall q \in \mathcal{Q}$ when they complete, and their associated connection parameters may vary significantly. Therefore, based on the quantization on connection parameters we introduced earlier, the egress node can easily decide which quantized space the QoS index for each connection $\mathrm{I}(q)$ belongs to. Then, assuming the associated QoS index for that space is denoted as I, the egress node revises I by exponential smoothing as follows:

$$\mathrm{I} \leftarrow \gamma \mathrm{I}(q) + (1 - \gamma)\mathrm{I}. \tag{4}$$

$\gamma \in (0, 1)$ is the forgetting factor and initial value of I is set to 0. The 1-D and 2-D illustrative examples are shown in Fig. 2(a) and Fig. 2(b), respectively. Measurements (black dots) that fall into the same quantized space are smoothed by the sequence of their appearance into one scaler, the QoS index (the blue crosses) as in (4). Then, it follows immediately, that for all ongoing connections $q \in \mathcal{Q}$, their multi-dimensional QoS requirements are simultaneously satisfied within a subnetwork if and only if $\mathrm{I} \in [0, 1]$.

Finally, the egress node is expected to inform the ingress node of this index I when a connection q completes the service, which will be used for admission decisions of future connections initiated from the ingress node.

B. Approximating the Mapping f

Next, we demonstrate the steps of approximating the mapping f by measurements. It is worth noting that the mapping f is usually quite complicated, which generally cannot be expressed in a closed-form expression. However, we aim to approximate this mapping through realtime measurements. For a packet network from ingress to egress nodes, each accepted connection originating at the ingress and destinating at the egress node, perform the following steps:

Step-1 Connection Parameter Update: The ingress node updates the M-dimensional connection parameters $\underline{x}$, corresponding to the newly admitted connection q's connection requirements. For instance, if the k_N connections are being serviced, with total $k_T l_T$ aggregated throughput requirement, then a new connection admission (of throughput requirement T_q^r) would result in the change of $N = k_N + 1$, and total served throughput is increased by corresponding (and discretized) throughput requirement $T = (k_T + k_T^q) l_T$, where $k_T^q = \lfloor \frac{T_q^r}{l_T} \rfloor$.

Step-2 Measurement Collection: The egress node constantly collects per-packet-based statistics of all admitted connections, including packet arrival and departure time, connection lifetime, and number of packets in error. Then, it computes QoS outage ratios R_q^D, R_q^T, R_q^E in (1).

Step-3 QoS Index Computation: Upon connection completion, for each connection:

(a) The egress node computes the QoS index for connection q as in (2).

(b) Based on the corresponding quantized space, compute an updated value I by exponential smoothing in (4).

(c) The ingress node updates the inputs (N, T) again. Take (N, T) again for example, the connection completion would result in the change of $N = k_N - 1$, and total served throughput decreased by corresponding throughput requirement $T = (k_T - k_T^q) l_T$.

Step-4 Update f: The egress node passes the QoS index I to the ingress node (in practice, this can be done by piggybacking this piece of information on the control plane messages). The ingress node records the measurement (N, T, I).

The same steps apply for any dimension of the input connection parameters $\underline{x}$, and iteratively, if enough connections are monitored and measurements are received, the mapping f is represented by a set of statistic pairs $(\underline{x}, \mathrm{I})$.

Before we move to the mathematical modeling of connection admission, we discuss the fundamental reasons why using a generic function f to map network parameters associated with connections between a pair of ingress and egress nodes into the QoS index will be sufficient to predict QoS for the connections. It is particularly so because network resources between the ingress and egress nodes are shared by cross traffic (*i.e.,* connections not related to the ingress or egress nodes). Our implicit assumption is that the usage of network resources by the cross traffic is assumed to be "quasi-stationary" despite being random or stochastic. Under such a quasi-stationary assumption, resource sharing and usage by the cross traffic has already been captured and reflected by the performance measurements such as throughput, delay and PER for the connections between the ingress and egress nodes under consideration. As a result, we do not need to explicitly consider the cross traffic and its performance impacts on connections between the two nodes. Numerical results from our simulation platform have confirmed the validity of the proposed GAC methodology.

V. Impacts on the QoS Index by Connection Admission

This section aims to estimate the impacts of the new connection admission on the QoS index, which is the central part to perform the accurate admission control.

When a new connection is admitted by the ingress node, the connection parameters in the mapping function defined above will be changed. Accordingly, to determine whether the connection should be admitted or not while providing satisfactory QoS for all connections (including the new one) becomes a problem of predicting if the QoS index I increases beyond 1, the unsatisfactory level, when the connection parameters are changed due to the connection admission. Toward that end, let a new connection admission cause a change of input parameters by $\Delta\underline{x} = (\Delta x_1, \Delta x_2, ..., \Delta x_M)$. The predicted QoS index is now given by

$$\tilde{\mathrm{I}} = f(\underline{x} + \Delta\underline{x}). \tag{5}$$

Then, we can appropriately approximate this mapping by a Taylor expansion [25] with the first several order terms and the unknown coefficients for those terms can be estimated by realtime measurements. Specifically, we consider the expansion with the first (representing the long-term average) and second-order partial derivatives (representing the fast change, or the variance) as follows:

$$\tilde{\mathrm{I}} = f(\underline{x} + \Delta\underline{x}) \approx f(\underline{x}) + \sum_{i=1}^{M} \frac{\partial f}{\partial x_i}\Delta x_i + \frac{1}{2}\left(\sum_{i=1}^{M} \frac{\partial f^2}{\partial x_i^2}(\Delta x_i)^2 + \sum_{i=1}^{M}\sum_{j\neq i} \frac{\partial f^2}{\partial x_i \partial x_j}\Delta x_i \Delta x_j\right). \tag{6}$$

Due to the discretized nature of each dimension of the connection parameters, these partial derivatives could be approximated by using adjacent measurements on the shape of curve produced by the mapping f. Section V-A and V-B demonstrate two examples when considering the 1-D and 2-D inputs. Furthermore, in (6), $(\mathrm{I}, \Delta x)$ are known through measurements, and the only unknown variable, to be computed, is $\tilde{\mathrm{I}}$, reflecting the change of output by admitting the new connection.

A. One-Dimensional Input

We start by using only a scaler input, *i.e.,* $M = 1$, as an illustrative example to demonstrate the steps of estimating this impact. The total number of connections being serviced is considered, $\underline{x} \triangleq N \in \mathbb{Z}$, as shown in Fig. 3. It is worth noting that N cannot accurately capture of the overall operational status of the packet network, since connections demand different amount of network resources with different QoS requirements; and we use this as an illustrative example only, and will consider a more practical scenario later.

The mapping f becomes

$$\mathrm{I} = f(N). \tag{7}$$

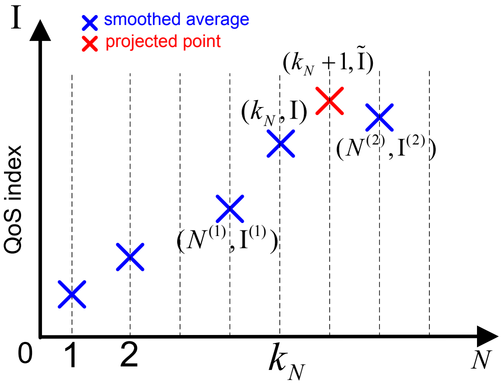


Fig. 3. An illustrative example for admission estimation on the mapping f, where scaler input N is considered, *i.e.,* $M = 1$.

Suppose a new connection q with a set of performance requirements denoted as (D_q^r, T_q^r, E_q^r) arrives at the ingress node for admission, and the updated connection parameter N falls into the k_N-th bucket of the input dimension; in other words, $N = k_N$. As discussed earlier, the potential new connection admission with requirements (D_q^r, T_q^r, E_q^r) would result in the change of input parameter for the mapping function is $\Delta x = \Delta N = 1$. When this input change occurs, it changes the QoS index as

$$\tilde{\mathrm{I}} = f(N+1). \tag{8}$$

By the reference of Taylor expansion in (6), we rewrite (8) as

$$\tilde{\mathrm{I}} = \mathrm{I} + \left.\frac{df}{dN}\right|_{k_N} + \frac{1}{2}\left.\frac{d^2 f}{dN^2}\right|_{k_N}, \tag{9}$$

where first and second-order derivatives are taken at $N = k_N$, as shown in Fig. 3. Since N is discrete, we approximate the "derivatives" by the slopes of adjacent network measurements. For example, assuming that two most adjacent measurements $(N^{(1)}, \mathrm{I}^{(1)}), (N^{(2)}, \mathrm{I}^{(2)})$, with a superscript "(1)" and "(2)", around the current bucket k_N can be obtained. Without lost of generality, $N^{(1)} < k_N$ and $N^{(2)} > k_N$. Then, the first-order derivative is computed as the average of two adjacent slopes of measurements,

$$\left.\frac{df}{dN}\right|_{k_N}^{(1)} \approx \frac{\mathrm{I} - \mathrm{I}^{(1)}}{k_N - N^{(1)}}, \quad \left.\frac{df}{dN}\right|_{k_N}^{(2)} \approx \frac{\mathrm{I}^{(2)} - \mathrm{I}}{N^{(2)} - k_N}, \tag{10}$$

and,

$$\left.\frac{df}{dN}\right|_{k_N} = \frac{1}{2}\left(\left.\frac{df}{dN}\right|_{k_N}^{(1)} + \left.\frac{df}{dN}\right|_{k_N}^{(2)}\right). \tag{11}$$

The second-order partial derivative is computed as the change of the above two slopes:

$$\left.\frac{d^2 f}{dN^2}\right|_{k_N} \approx \frac{\left.\frac{df}{dN}\right|_{k_N}^{(2)} - \left.\frac{df}{dN}\right|_{k_N}^{(1)}}{N^{(2)} - 2k_N + N^{(1)}}. \tag{12}$$

As shown in Fig. 3, the potential connection admission causes the QoS index, the function value of f to change from I to $\tilde{\mathrm{I}}$, when the connection parameter changes from k_N to $k_N + 1$. Although the Taylor approximation by (11) and (12) is not perfect, the first two order terms are expected to provide good estimates of the performance index.

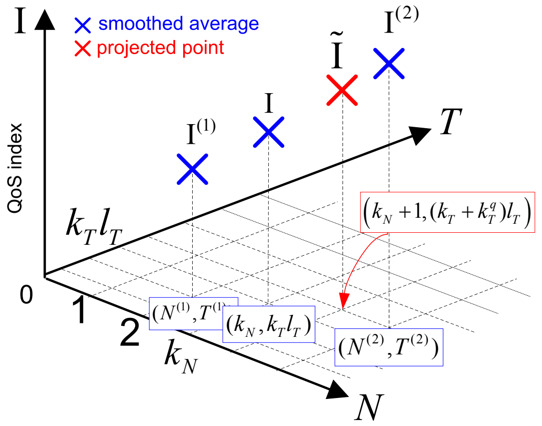


Fig. 4. An illustrative example for admission estimation on the mapping f, where two-dimensional inputs are considered, *i.e.*, $M = 2$.

B. Two-Dimensional Inputs

As discussed at the beginning of this section, a single scaler input variable, the number of currently served connections N, cannot accurately capture of the overall operational statistics of the packet network. Towards this end, we now move to a more realistic case where two-dimensional input variables are considered as the connection parameters: the number of ongoing connections $N = \sum_{q\in\mathcal{Q}} = k_N$ and the total amount of throughput requirement of all served connections $T = \lfloor \sum_{q\in\mathcal{Q}} T_q^r \rfloor = k_T l_T$. Therefore, we have $\underline{x} \triangleq (N, T) \in \mathbb{R}^2$, as shown in Fig. 4. Then, the mapping f becomes

$$\mathrm{I} = f\,(N, T)\,. \tag{13}$$

As discussed earlier, the potential new connection admission with requirements (D_q^r, T_q^r, E_q^r) would result in an input change of $\Delta\underline{x} = (\Delta N, \Delta T) = (1, T_q^r)$, into the packet network. Since throughput dimension is discretized, we write new demand as $k_T^q = \lfloor T_q^r / l_T \rfloor$, where k_T^q denotes the increased number of discretized bucket along that dimension, or $\Delta\underline{x} = (1, k_T^q l_T)$. As shown in Fig. 4, once this input change $\Delta\underline{x}$ is incurred, it will result in an output change to,

$$\tilde{\mathrm{I}} = f\Big(k_N + 1, (k_T + k_T^q) l_T\Big). \tag{14}$$

With the reference of Taylor expansion in (6), we rewrite (14) as,

$$\begin{aligned}\tilde{\mathrm{I}} = \mathrm{I} &+ \frac{\partial f}{\partial N} + k_T^q l_T \frac{\partial f}{\partial T} + \frac{1}{2}\frac{\partial^2 f}{\partial N^2} \\ &+ \frac{(k_T^q l_T)^2}{2}\frac{\partial^2 f}{\partial T^2} + k_T^q l_T \frac{\partial^2 f}{\partial N \partial T},\end{aligned} \tag{15}$$

where all first and second-order partial derivatives are taken at $(k_N, k_T l_T)$, if assuming that after admitting the new connection q, as shown in Fig. 4.

Similar to (11) and (12) as computing the derivatives for dimension of N, we show the first and second-order statistics for T, as

$$\left.\frac{\partial f}{\partial T}\right|_{k_T} = \frac{1}{2}\left(\frac{\mathrm{I} - \mathrm{I}^{(1)}}{k_T l_T - T^{(1)}} + \frac{\mathrm{I}^{(2)} - \mathrm{I}}{T^{(2)} - k_T l_T}\right), \tag{16}$$

and

$$\left.\frac{\partial^2 f}{\partial T^2}\right|_{k_T} \approx \frac{\left.\frac{\partial f}{\partial T}\right|_{k_T}^{(2)} - \left.\frac{\partial f}{\partial T}\right|_{k_T}^{(1)}}{T^{(2)} - 2k_T l_T + T^{(1)}}, \tag{17}$$

where $T^{(2)} > k_T l_T$ and $T^{(1)} < k_T l_T$ are the most adjacent right- and left-hand side measurements around $k_T l_T$ of the k_T-th bucket.

In order to obtain the cross second-order partial derivative $\partial^2 f / \partial N \partial T$, we use the limit definition of the first-order partial derivative:

$$\left.\frac{\partial T}{\partial N}\right|_{k_N} = \lim_{\Delta N \to 1} \frac{(k_T + k_T^q) l_T - k_T l_T}{\Delta N} = k_T^q l_T. \tag{18}$$

Therefore,

$$\left.\frac{\partial^2 f}{\partial N \partial T}\right|_{k_N, k_T} = \left.\frac{\partial^2 f}{\partial T^2}\right|_{k_T} \left.\frac{\partial T}{\partial N}\right|_{k_N} = k_T^q l_T \left.\frac{\partial^2 f}{\partial T^2}\right|_{k_T}. \tag{19}$$

Substituting (19) into (15) yields,

$$\begin{aligned}\tilde{\mathrm{I}} = \mathrm{I} + \left.\frac{\partial f}{\partial N}\right|_{k_N} &+ k_T^q l_T \left.\frac{\partial f}{\partial T}\right|_{k_T} + \frac{1}{2}\left.\frac{\partial^2 f}{\partial N^2}\right|_{k_N} \\ &+ \frac{3}{2}(k_T^q l_T)^2 \left.\frac{\partial^2 f}{\partial T^2}\right|_{k_T}.\end{aligned} \tag{20}$$

It is interesting to observe that the potential connection admission eventually updates the output of the mapping f from I to $\tilde{\mathrm{I}}$ (and this would be the admission impact we are aiming to estimate), when the connection parameters change from $(k_N, k_T l_T)$ to $\big(k_N + 1, (k_T + k_T^q) l_T\big)$.

For some network scenarios, if the shape of the curve produced by the mapping f is smooth enough around current connection parameters $(k_N, k_T l_T)$ so that the second-order derivatives are negligible and the first-order statistics are sufficient, we simplify (20) as:

$$\tilde{\mathrm{I}} = \mathrm{I} + \left.\frac{\partial f}{\partial N}\right|_{k_N} + k_T^q l_T \left.\frac{\partial f}{\partial T}\right|_{k_T}. \tag{21}$$

To conclude this section, it is worth noting that although only two-dimensional input variables are demonstrated to show the steps of estimating the impacts of admitting the new connection, it is easy to observe the applicability of the the impact estimation steps to M-dimensional inputs in general.

VI. Proposed GAC Algorithm

Our proposed GAC methodology for QoS control operates in an *on-demand* fashion where it is initialized when a new connection q arrives at an ingress node of the packet network with multiple QoS constraints, intended to communicate with the egress node (the connections may potentially go through several heterogeneous subnetworks). It is important to point out that in this way, our algorithm only considers the admission decision for one connection at a time, so that we leave the multiple connection admission with resource negotiations to the future work. Without loss of generality, we use $M = 2$, *i.e.*, the number of connections being serviced N, and corresponding total throughput T as an example. The following steps summarize and describe the algorithm.

Step-1 Measurement Collection and Smoothing: as shown in Section IV-A, the ingress node keeps track of pairs of measurements (N, T, I) to approximate the shape of curve produced by the mapping $f : \mathrm{I} = f(N, T)$. We re-iterate here that the mapping f is not modeled by any closed-form, but approximated through the realtime measurements.

Step-2 Derivative Derivation: upon receiving the measurements, the ingress node computes all partial derivatives, $\frac{\partial f}{\partial N}, \frac{\partial f}{\partial T}, \frac{\partial^2 f}{\partial N^2}, \frac{\partial^2 f}{\partial T^2}$, around the connection parameter $(k_N, k_T l_T)$ as in equations (11), (12), (16), and (17).

Step-3 Admission Decision: the impacts of admitting the new connection on the QoS index is characterized and approximated by Taylor expansion while taking both the first and second partial derivatives in (20), or only the first-order derivatives in (21), as inputs. Then, the new QoS index $\widetilde{\mathrm{I}}$ is calculated in a closed-form. Finally, we verify if there is enough network resources within the packet network available for the new connection, as:

$$\begin{cases} \text{Admission}, & \text{if } \widetilde{\mathrm{I}} \leq 1 \\ \text{Rejection}, & \text{otherwise}. \end{cases} \tag{22}$$

VII. Performance Evaluation

Although our proposed algorithm is applicable to many kinds of packet networks, in this section we use the OPNET [26] modeler to establish a wireless mesh network (WMN) simulation environment, as one of the representative packet network example for performance evaluation. It creates more complicated network environment if compared with wired network that has less interference than wireless networks. Eighteen wireless mesh routers are randomly and independently deployed on a two dimensional area, and variable number of client and server couples are connected through the WMN, as shown in Fig. 5.

The developed wireless mesh router models representing the functionalities of different layers are shown in Fig. 6(a). In PHY layer, original 14 stage pipeline model is enhanced with the Rayleigh fading channel (of Jake's model [27]) while the required PER is derived based on SINR curves for the used adaptive modulation and coding scheme (see Fig. 7). Furthermore, in order to reduce the interference to adjacent concurrent transmissions, increase the frequency reuse and channel capacity, the nodes are equipped with directional antennas. Furthermore, the integrated QoS scheduling and routing protocol (IQoSR, [24]) is used in the network and MAC layers to provide sub-optimal solutions for QoS. Specifically, IQoSR scheme selects a sub-optimal route satisfying multiple QoS requirements (throughput, delay and PER) simultaneous overcoming the NP hard problem in [2], and it employs a distributed proportional-fair (PF) scheduling algorithm to support the long-term throughput constraint (specified by the routing layer). TCP retransmission is also employed in each node where the packets are retransmitted from the ingress node if its received PER is higher than a QoS threshold. We refer our solution as "IQoSR+GAC".

The client and server models that are generating and receiving traffic (as ingress and egress nodes) are shown in Fig. 6(b) that have exactly the protocol stack of the real environment, with all the standard TCP/IP protocol layers. The network configuration parameters in our simulation environment are summarized in Table I.

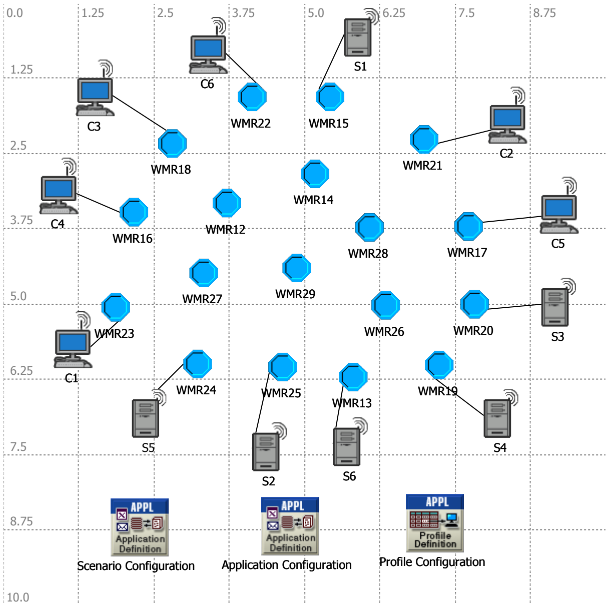
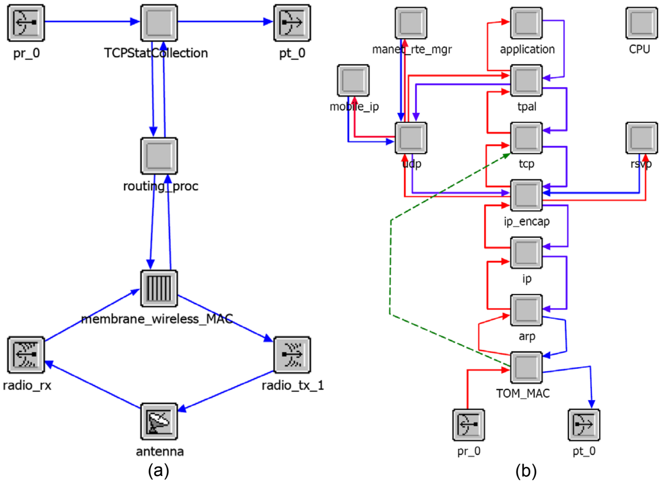


Fig. 5. Example of the standard scenario used for the simulation campaign. Eighteen wireless mesh routers consist of the backhaul network, where six client/server pairs connect to.



(a)

(b)

Fig. 6. (a) Protocol layer models for wireless mesh routers, (b) The client and server models.

Furthermore, some other processes have been added as seen in Fig. 5. The "Scenario Configuration" configures the global parameters of the simulation and provides perfect TDMA synchronization among all wireless mesh routers. Here we use TDMA, because it is widely adopted in most WMN MAC protocols [28], however it is worth noting that our proposed GAC does not rely on TDMA itself, nor any inputs from the used MAC protocol. Two processes "Application Configuration" and "Profile Configuration" defines, instead, application profiles like HTTP, FTP, VoIP. Both nodes allow us to select and configure multiple application models (for the selected application in our model, HTTP, VoIP and FTP, Table II, III and IV indicate the most important parameters) and application usage patterns, like how often the application is used, the usage during each session, the number of users and the usage fluctuations etc. Finally, new connection is chosen from the three considered type of network traffic (HTTP, FTP and VoIP), and associated with three QoS requirements, namely: packet delay, throughput, and PER (see Table II, III

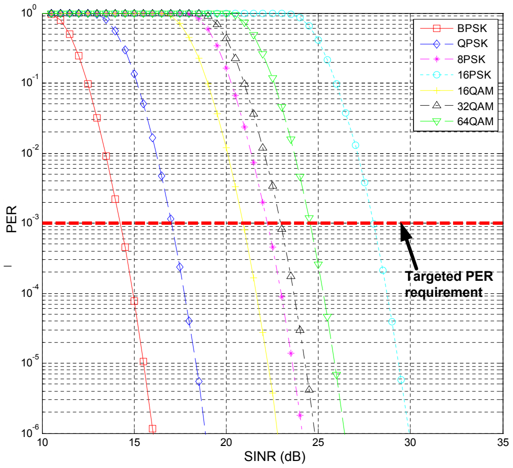


Fig. 7. The used mapping from the received SINR (dB) to PER with different combinations of modulation and coding schemes.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Channel Model	Rayleigh fading model
Path Loss Coefficient	3.5
Directional Antenna Pattern	Side lobe: -25dB, Main lobe: 30°
Adaptive Modulation and Coding	BPSK, QPSK, 16QAM, 64QAM
Doppler Frequency	25Hz
System Bandwidth	10MHz (3.5GHz band)
Slot Duration	400μs
Slots per TDMA Frame	8
Frame Duration	3.2ms
MAC Packet Length	512 bytes
Number of Mesh Routers	5-35, Typical number 15
Network Area	3 mile × 3 mile square
Transmission Range	1 mile
Traffic Patterns	FTP and VoIP
Queue Length	Infinite
Error margins $\beta_D, \beta_T, \beta_E$	1.2
HTTP, VoIP and FTP occupancy	20%, 40%, 40% (if not specified)

and IV); and a total of 10,000 connections are simulated. Here we use HTTP traffic to simulate and investigate the increasingly popular web browsing behavior on the Internet. It is worth noting that although some web pages can be really big, we specify its size between 300-800KB (as of typical for www.google.com, 399KB) to investigate the impact of "small" traffic on the proposed algorithm, if compared with FTP and VoIP connections (see Section VII-C). The servers in the scenario periodically send a feedback packet containing the measurements like (N, T, I) to ingress nodes to aid the admission decision making. The ingress node maintains a sliding window (of fixed size 5s, and its effect will be analyzed in Section VIII-B2) collecting those measurements sent by the egress node, and we assume this feedback delay is only incurred by packet scheduling along the path (and its effect will be analyzed in Section VIII-B3).

The applicability of GAC to other wireless and wired IP networks are discussed in Section VIII.

A. An Example: A Six-Node WMN

TABLE II
FTP PROFILE

Parameter	Value
File Size	Constant 1 MBytes, *i.e.,* ≈ 2,000 packets
Type of Service	Best Effort
Delay Requirement	Not Req. (set as a sufficiently large number)
Throughput Requirement	100kbps-2Mbps, uniformly chosen
PER Requirement	0

TABLE III
VOIP PROFILE

Parameter	Value
Encoder scheme	G.729A
Voice frame per packet	1
Type of Service	Interactive Voice
Duration	10s-50s, uniformly chosen
Delay Requirement	100ms-300ms, uniformly chosen
Throughput Requirement	17kbps-106kbps, uniformly chosen
PER Requirement	0.01

We first assess the proposed GAC methodology in a six-node WMN in OPNET as an illustrative example in Fig. 8, the parameter settings are the same as Table I. Node 1 serves as the ingress node (client) to generate connections (inter-arrived VoIP and FTP traffic) and node 4 serves as the egress node (*i.e.,* the Internet gateway as the intended receiver, or server to be consistent). Node connectivity is also shown in the figure where nodes 2, 3, 5, and 6 may serve multiple cross-traffic of different QoS requirements due to the decision of underlying IQoSR protocol, and there exist co-channel interference between cross connections; and thus this network setting will confirm the effectiveness of our proposed algorithm on handling cross-traffic. The number of ongoing connections N from node 1 to 4 and their total required throughput T are considered as connection parameters, and bucket length $l_N = 1, l_T = 17$kbps. Table V summarizes the simulation results to estimate impacts of a new connection on the QoS index, *i.e.,* to use only the first-order partial derivatives as in (21), and both the first and second-order partial derivatives as in (20). We compare their performance in terms of:

1) maximum supported throughput $T_{\max}$: defined as the maximum amount of throughput that the gateway can support, beyond which all ongoing connections' QoS cannot be guaranteed;
2) maximum number of accepted connections $N_{\max}$: corresponding to the maximum amount of supported throughput, its associated number of accepted connections;
3) achieved QoS outage: defined as the percentage of total accepted connections that fail to provide the required QoS parameters;
4) blocking probabilities: defined as the percentage of connections refused to be admitted to the network);
5) error bar (defined as the 95% confidence interval) of the first two metrics.

We next depict the attained QoS index and network capacity prediction process in Fig. 9. Here we only consider the 1-D input as the aggregated throughput requirement when varying different bucket lengths $l_T = \{1, 2\}$Mbps. We observe that the finer discretization (smaller l_T) will provide more accuracy of constructing the mapping f and as a result, more throughputs are allowed in the network. Also, the measurement points

TABLE IV
WEB BROWSING (HTTP) PROFILE

Parameter	**Value**
Size of Web page	300KB-800KB, uniformly chosen (HTTP GET, without caching)
Delay Requirement	Not Req.
Throughput Requirement	Not Req.
PER Requirement	Not Req.

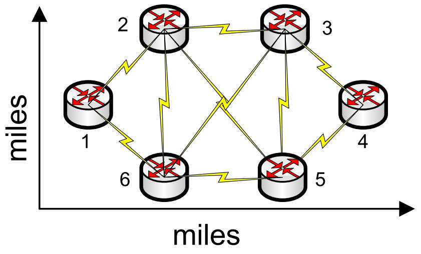


Fig. 8. A six-node WMN setting, where node 1 serves as the ingress node to generate connections with multiple QoS requirements and node 4 serves as the egress node.

around the smoothed value exhibit fewer variations in term of the received QoS index value when $l_T = 1$Mbps.

It is interesting to observe from Table V that the second-order statistics enhance the accuracy of the QoS index predictions for AC over the method using the first-order statistics. As a result, the estimations using the second-order terms improve the volume of maximum supported throughput and connection number by 25% and decrease the prediction error, QoS outage, and blocking probabilities significantly. This is because the higher order statistics show finer horizon of the shape of the curve produced by mapping f, especially when the packet network operates around its capacity where one single connection admission with large throughput requirement may jeopardize all existing connections' QoS. Thus, second-order statistics aid to admit the most *appropriate* connection (in term of throughput requirement) with the knowledge of the satisfactory QoS index $\mathrm{I} \in [0, 1]$, while maintaining QoS satisfactions to all ongoing connections. To this end, we can further conclude that the second-order statistics is enough and the third-order terms are not needed.

B. The Overall Network Performance

The proposed algorithm, referred as "IQoSR+GAC", is compared with the existing method "IQoSR" that does not include any AC scheme, and the statistical connection admission control (SCAC) algorithm in [23], referred as "IQoSR+SCAC". We choose "SCAC" because it is the most recent research that also makes uses of collected statistics to perform AC by estimating the achievable capacity as the amount of bandwidth that can probabilistically guarantee PLR to be smaller than a threshold, and uses the central limit theorem to approximate the capacity in a closed-form by a Gaussian process. As before, we choose N, T as connection parameters. Our algorithm is also compared with the conventional protocol layer 2 and 3 techniques: the round-robin scheduler (RR, [29]) and AODV routing protocol [30], to be referred as "RR+AODV" scheme below.

TABLE V
EFFECTS OF USING DIFFERENT COMBINATIONS OF PARTIAL DERIVATIVES FOR ADMISSION ESTIMATION (WHEN $l_N = 1, l_T = 1$MBPS)

	First-order stat.	**First and second-order stat.**
$T_{\max}$ (Error Bar)	12($\pm$1.3)Mbps	15($\pm$0.8)Mbps
$N_{\max}$ (Error Bar)	15($\pm$2)	19($\pm$2)
QoS Outage	$\approx 5\%$	$\approx 2\%$
Blocking Probability	$\approx 12\%$	$\approx 8\%$

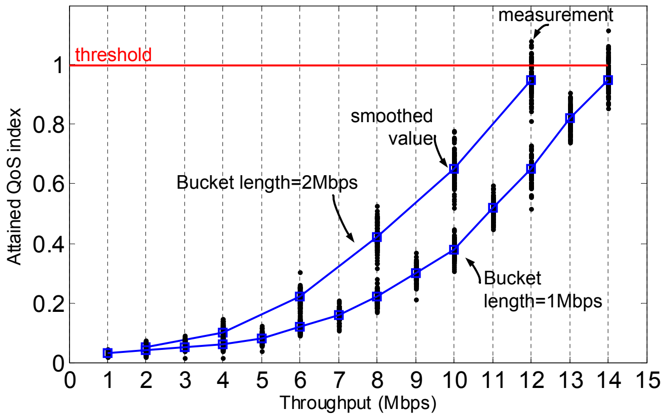


Fig. 9. The attained QoS index and network capacity prediction process.

As shown a complete simulation topology in Fig. 5. The overall performance for various schemes is shown in terms of (a) the gateway goodput (defined as the amount of traffic flowing through the gateway without QoS failures) in Fig. 10, and (b) the average QoS outage probability of completed connections in Fig. 11 (defined as the probability of a connection's QoS requirements to fail during their lifetime, or the condition $\mathrm{I}(q) \leq 1, \forall q \in \mathcal{Q}$ at any time is not satisfied at all.).

Fig. 10 shows that "IQoSR+GAC" outperforms all other schemes in terms of the overall gateway goodput even under heavy load conditions (*i.e.,* small connection inter-arrival time). It is observed that 1.6, 2.8 and 4.5 times of goodput improvements are achieved when connection inter-arrival time is 5ms, and "IQoSR+GAC" is compared with "IQoSR+SCAC", "IQoSR", and "RR+AODV" scheme, respectively. Such huge improvements are made possible by the proposed technique because it adequately tracks the overall QoS index as connections are admitted and served by the network and the QoS estimations by Taylor approximation with the use of the first two-order statistics are sufficiently accurate. We also find that under high traffic-load conditions, the arrival process could be more accurately assumed to be Gaussian, which helps the "SCAC" scheme achieve a relatively high goodput. However, when the traffic load is relatively low, the Gaussian approximation is no longer accurate and the "SCAC" scheme makes wrong admission decisions, which turn into less goodput and higher QoS outage when compared with our proposed technique. Furthermore, the closed-form derivation for achievable capacity of "SCAC" scheme does not consider throughput and delay requirements. It is also interesting to observe that the gateway goodput saturates when the traffic load becomes excessive. Finally, when we increase the node density from 15 to 25 nodes per squared miles, similar results have been obtained where the gateway goodput decreases by

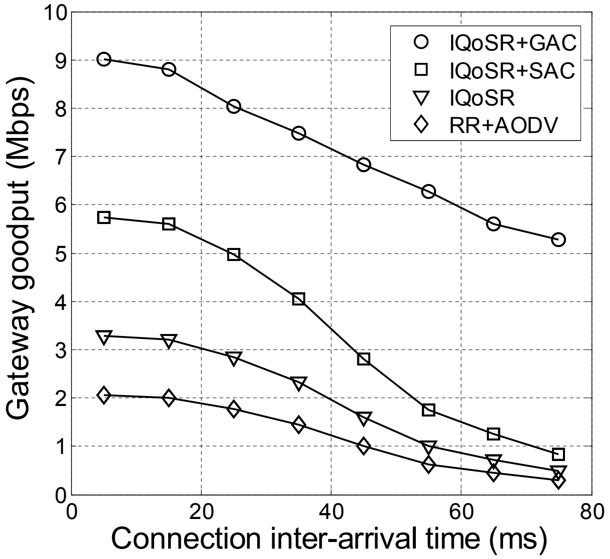


Fig. 10. Simulation result of the overall gateway goodput with respect to (w.r.t.) the different new connection inter-arrival time.

10%. This is not only because of the keen competition for network resources by ongoing connections, but also much more co-channel interference from the adjacent transmissions.

Fig. 11 illustrates the QoS outage probability. It can be seen that under the highest traffic condition considered in the figure, the proposed GAC scheme can guarantee 95% of the QoS satisfaction for the underlying applications, as compared to only 81% if no AC is used, 83% if "SCAC" scheme is used, and 58% if "RR+AODV" is employed. For low traffic load, proposed scheme can achieve almost 0% QoS outage as compared to at least 7% of all other algorithms. The similar results have been obtained when 25 nodes are randomly deployed, and for all schemes the achieved QoS outage is around 5% higher than the 15-node case, since limited amount of network resources are shared by more traffic and thus higher probability of QoS failures. Benchmark protocol "RR+AODV" does not consider QoS at all, and thus connections may be concentrated on some nodes that leads to congestion and QoS outage. "IQoSR" performs better than "RR+AODV" because it uses QoS-aware scheduling and routing protocols, however it has no AC policy enforced for new connections. "SAC" is also a measurement based approach but assuming traffic is Gaussian which can be more accurate under heavy load conditions. In contrast, the improvement of GAC over all others is made possible because the impact of the newly admitted connections on the QoS experience of the existing and new connections can accurately be reflected by the predicted QoS index that integrate multiple QoS parameters, including the attained and required values. GAC properly rejects connections if its admission will jeopardize the existing connections' QoS experience which guarantees the received low QoS outage probability.

C. Impact of HTTP Traffic

We next investigate the impact of HTTP traffic on the admission control process and capacity estimation accuracy.

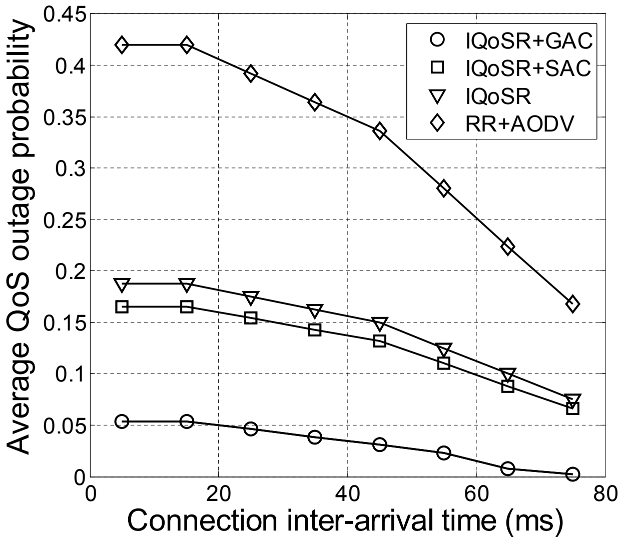


Fig. 11. Simulation result of the average QoS outage probability w.r.t. the different new connection inter-arrival time.

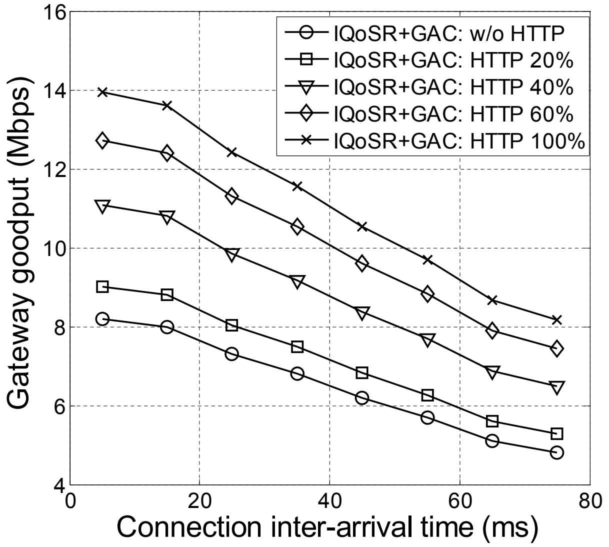


Fig. 12. Simulation result of the average QoS outage probability w.r.t. the different new connection inter-arrival time.

To achieve this, we change the relevant amount of HTTP traffic occupying the entire traffic input (w.r.t. VoIP and FTP connections), from 0% gradually increased to 100%. As shown in Fig. 12, since our specified HTTP traffic load is relatively small compared with average size of FTP and VoIP connections, increasing the HTTP traffic will help boost up the overall gateway goodput. This is because that with the help of relatively small new arrival connections (as of specified by our HTTP profile, 500-800KB), we have a finer view of the mapping f (consistent with smaller bucket length as in Fig. 9. Therefore, more accurate prediction is achieved. This effect becomes weak with the increase of HTTP occupancy, because the network capacity gradually saturates and GAC starts to block new connections. As the overall trend, it is also consistent with all previous analysis that when increasing the traffic load (smaller connection inter-arrival time, the gateway goodput decreases).

VIII. Applicability and Feasibility of the Proposed Methodology

This section investigates and discusses about the applicability and feasibility issues of the proposed GAC methodology for QoS control.

A. The Applicability and Scalability

In reality, we keep track of all possible ingress and egress nodes in a network even some of them may be removed/added. This dynamics will not have any impact on our the proposed admission estimation procedure, because in that case the underlying routing protocol will forward the traffic to other ingress/egress node (in case of removal) that we are also monitoring, or we will start a new mapping function (in case of addition) that is identical to our normal process proposed in Section VI.

The proposed GAC methodology can be generally applied to a wide variety of packet networks, wired and/or wireless. It is so because the main essence of this methodology is to use a generic mapping f to represent sharing of network resources between the ingress and the egress. By doing this, without considering the heterogeneous operational and protocol features of each network component, we are able to estimate the potential impact of the new connection admission on the QoS experience of the existing and new connections. For example, MAC layer protocols/standards like IEEE 802.11 and IEEE 802.16 are fully compatible with our proposed GAC algorithm, because the inputs and output of GAC does not rely on any operational detail of lower layer protocols in use, but only the *measurement values* need to be exported (e.g., attained packet delay, however how packets are forwarded, as of scheduling, is not part of our proposal). Although different MAC protocols produce different interference levels, our AC algorithm runs in the application layer and the effects of these interferences will be well captured and reflected in the attained measurement values, and then QoS index. This transparency guarantees the seamless integration. For a network connected by multiple heterogeneous like a heterogeneous WiFi-WiMAX network [31], our algorithm simply take the source terminal as the ingress node, and the destination as the egress node, and between the ingress/egress node pair is a two-hop link (operating different MAC/PHY layer protocols).

Furthermore, this methodology is scalable in terms of computational complexity, since neither the number of nodes in the network, nor their generated/forwarded type of network traffic does not have any impact on the algorithm at all. To ensure QoS for the admitted connections between the ingress and egress nodes, no intermediate node other than the two is involved in measurement collection and performance estimations. Therefore, our approach can also cope with the possibility that some intermediate nodes appear or removed from the network.

Measurement and computation overhead can be further reduced by sampling the performance of a small subset of packets for each connection. Finally, when the proposed method is applied to wired networks, smooth channel fluctuation with no interference can be expected and we may only need the first-order statistics for the Taylor approximation, since the second-order statistics are used to track the fast change, especially applicable to wireless networks.

Since our proposed approach by essence is measurement based, it is possible that function f can be mapped wrongly, if the admitted connection's QoS requirements are more than what the network can support, but the mapping f may wrongly indicate that it can be admitted. Then, since network resources are shared (detailed resource allocation depends on the used scheduling/routing algorithms), the admitted new connection may potentially impact both the attained performance of existing and its own connections, which in turn will be reflected by the QoS index I (in this case it will exceed the threshold 1). After, it is the responsibility of underlying resource allocation scheme (which is not part of the AC) to optimize the limited amount of resources according to their Grade-of-Service, and no more connections can be admitted, until some resources are released so that the reflected QoS index falls below 1.

B. Feasibility of the Proposed Methodology

1) Impact of Connections with Large Throughput Requirements: As mentioned earlier, the Taylor expansion is used as a tool to estimate the potential admission impact. However, the approach is appropriate provided that sufficient measurement statistics have been collected within a relatively close region of the current operating point. When incoming connections are associated with large throughput requirements compared with the total network capacity, our proposed algorithm may not perform satisfactorily due to the discontinuity of the mapping f that characterizes the sharing of network resources. This will impact the accuracy of the partial-derivative calculations and ultimately the predicted QoS index $\widetilde{\mathrm{I}}$. The effect is depicted in Fig. 13 where only FTP traffic is simulated with different throughput requirements, and connection inter-arrival time is set as 35ms. We observe that larger the throughput requirement T_q^r is, the more severe impact on error accumulation and amplification for the estimation would be. Especially when $T_q^r > 2.5$Mbps per connection, erroneous admission of one single large connection may jeopardize QoS of all existing connections. On the other hand, negative effects of erroneous admission are not significant if small amount of throughput requirement (relative to the network capacity) is considered. For these reasons, the QoS outage probability increases significantly w.r.t. T_r^q for all node densities considered in Fig. 13. As a separate point, as the node density increases, more co-channel interference is generated, thus causing the QoS outage probability to increase.

2) Statistics Collection Time: This corresponds to the time window that performance measurements and statistics are periodically collected and aggregated at the ingress node, to estimate the coefficients for the Taylor expansion of mapping f. If the time window is too long, collected data may not represent the most recent network status, given the highly dynamic traffic. On the other hand, a relatively short collection time may lead to a lack of statistical significance of the measurements and thus inaccurate predictions. The effect of these can be seen from Fig. 14 where for a fixed connection inter-arrival time (*i.e.,* a given traffic loading), there is an optimal statistics collection time to achieve the lowest

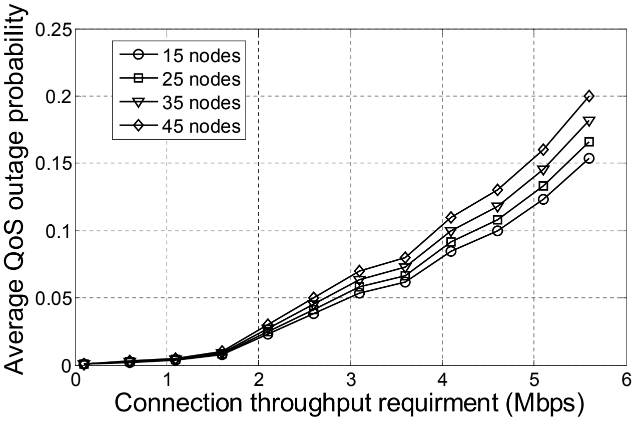


Fig. 13. The impact of different throughput requirements on the estimation of the new QoS index if the new connection is admitted. The figure is plotted with different number of nodes in a fixed size area.

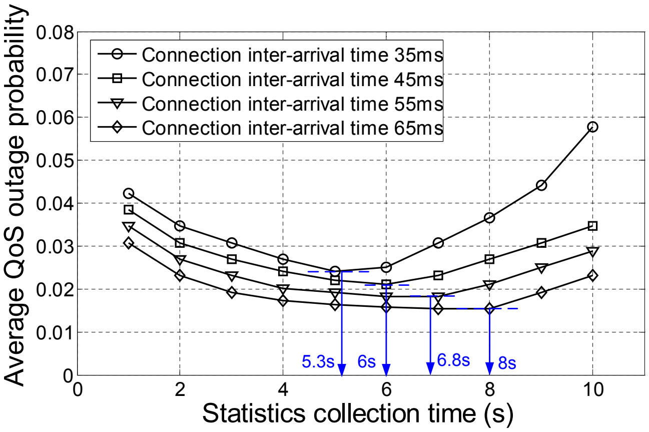


Fig. 14. The impact of the statistics collection time on the estimation of the new QoS index if the new connection is admitted. The figure is plotted with different connection inter-arrival time when $N = 15$ nodes are deployed.

QoS outage probability. For example, 5.3s (or equivalently 137 connections) when inter-arrival time is 35ms. Therefore, these simulation results can be used to identify the optimal collection time for the best system performance.

3) Statistics Feedback Delay: Performance measurements and associated computation are expected to carry out by the ingress node. Then, there will be delay in collecting the measurements and sending relevant data from the egress to ingress nodes, which is referred to as the feedback delay in the following. Fig. 15 shows its impacts on the QoS outage probability when connection inter-arrival time is set as 35ms, where the delay is simulated by holding the feedback packet at the egress node to the corresponding amount of time before sending out. As intuitively expected, long feedback delay makes the measurements and admission data obsolete, which in turns causes improper admission decisions at the ingress node. Results in Fig. 15 show a significant degradation of QoS when the feedback delay is bigger than 0.8s, relatively 16% of the optimal statistics collection time (*i.e.,* 5s) in Fig. 14. This is because the higher statistics feedback delay would eventually leads to the outdated measurements in the admission-control decision. Therefore, it is important to transfer the admission related data and information with the highest priority between the ingress and egress nodes.

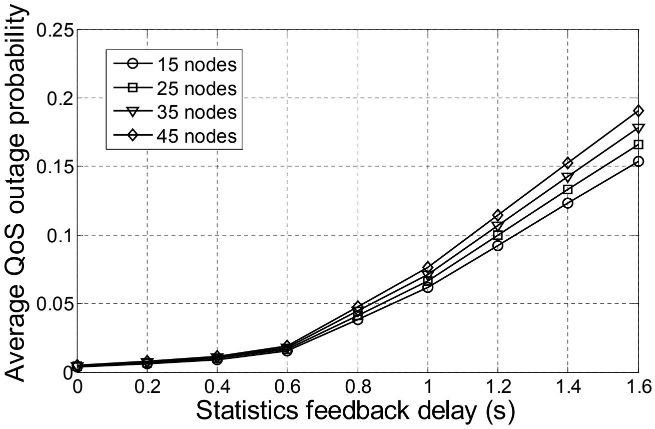


Fig. 15. The impact of the statistics feedback delay on the estimation of the new QoS index if the new connection is admitted. The figure is plotted with different number of nodes in a fixed size area.

IX. Conclusion

In this paper, a GAC methodology for packet networks, wired and/or wireless, has been proposed. The methodology consists of several components. First, the concept of QoS index is introduced to integrate the multi-dimensional QoS requirements to indicate the degree of QoS satisfaction for all connections between a given pair of ingress and egress nodes in the network. Second, a generic mathematical function is used to represent sharing of network resources along all communication paths between the ingress and egress nodes, thus enabling us to ignore the details of heterogeneous network operations and protocols in use. Since the mapping function is generally unknown, to make the proposed method practical, the function is approximated by a Taylor expansion with the first few order terms. The associated coefficients in the Taylor approximation can be estimated and computed by realtime performance measurements at the ingress and egress nodes. Third, an admission-control algorithm has been proposed to decide whether or not to admit a new connection by estimating the potential impact of the new connection on the QoS index by the Taylor approximation. By extensive simulations using WMNs as example, we have validated with the proposed methodology for connection admission. Furthermore, our new method performs better when compared with other statistics-based algorithm and conventional algorithms. By using the simulation platform, we have also investigated a number of implementation and feasibility issues of the proposed method. Some of these issues represent open problems that deserve to be studied further. Nevertheless, results reported here reveal that the proposed framework for connection admission control represents a new and efficient approach to the problem that can be further extended and refined for potential use in practical networks in the future.

References

[1] M. El-Sayed and J. Jaffe, “A view of telecommunications network evolution,” *IEEE Commun. Mag.*, vol. 40, no. 12, pp. 74–81, Dec. 2002.

[2] Z. Wang and J. Crowcroft, “Quality-of-service routing for supporting multimedia applications,” *IEEE J. Sel. Areas Commun.*, vol. 14, no. 7, pp. 1228–1234, Sept. 1996.

[3] P. Gupta and P. R. Kumar, “The capacity of wireless networks,” *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2003.

[4] M. H. Ahmed, "Call admission control in wireless networks: a comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 7, no. 1, pp. 49–68, Qtr. 2005.

[5] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 14, no. 4, pp. 711–717, May 1996.

[6] A. Djouama, M. Abdennebi, L. Mokdad, and S. Tohme, "Lifetime aware admission control for infrastructure-less wireless networks," in *Proc. 2009 IEEE Symp. Comput. Commun.*, pp. 67–72.

[7] F. Didi, M. Feham, H. Labiod, and G. Pujolle, "Dynamic admission control algorithm for WLANs 802.11," in *Proc. 2008 Int'l Conf. Inf. Commun. Technol.: From Theory Applications*, pp. 1–6.

[8] O. Baldo, "A cross-layer distributed call admission control," in *Proc. 2009 IEEE WiMob*, pp. 441–446.

[9] E. Stevens-Navarro, A. H. Mohsenian-Rad, and V. Wong, "Connection admission control for multiservice integrated cellular/WLAN system," *IEEE Trans. Veh. Technol.*, vol. 57, no. 6, pp. 3789–3800, Nov. 2008.

[10] J. Ratica and L. Dobos, "Mobile ad-hoc networks connection admission control protocols overview," in *Proc. 2007 Int'l Conf. Radioelektronika*, pp. 1–4.

[11] S. Zhang, F. R. Yu, and V. Leung, "Joint connection admission control and routing in IEEE 802.16-based mesh networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1370–1379, Apr. 2010.

[12] L. Seungjoon, G. Narlikar, M. Pal, G. Wilfong, and L. Zhang, "Admission control for multihop wireless backhaul networks with QoS support," in *Proc. 2006 IEEE WCNC*, vol. 1, pp. 92–97.

[13] G. Narlikar, G. Wilfong, and L. Zhang, "Designing multihop wireless backhaul networks with delay guarantees," in *Proc. 2006 IEEE INFOCOM*, pp. 1–12.

[14] A. Herms, S. Ivanov, and G. Lukas, "Precise admission control for bandwidth reservation in wireless mesh networks," in *Proc. 2007 IEEE MASS*, pp. 1–3.

[15] F. Yu, V. Krishnamurthy, and V. C. M. Leung, "Cross-layer optimal connection admission control for variable bit rate multimedia traffic in packet wireless cdma networks," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 542–555, Feb. 2006.

[16] ——, "Cross-layer optimal connection admission control for variable bit rate multimedia traffic in packet wireless cdma networks," in *Proc. 2004 IEEE GLOBECOM*, vol. 5, pp. 3347–3351.

[17] S. Zhang, F. R. Yu, and V. Leung, "Joint connection admission control and routing in IEEE 802.16-based mesh networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1370–1379, Apr. 2010.

[18] R. de Renesse, V. Friderikos, and H. Aghvami, "Cross-layer cooperation for accurate admission control decisions in mobile ad hoc networks," *IET Commun.*, vol. 1, no. 4, pp. 577–586, Aug. 2007.

[19] Q. Shen, X. Fang, P. Li, and Y. Fang, "Admission control for providing QoS in wireless mesh networks," in *Proc. 2008 IEEE ICC*, pp. 2910–2914.

[20] S.-L. Su, Y.-W. Su, and J.-Y. Jung, "A novel QoS admission control for ad hoc networks," in *Proc. 2007 IEEE WCNC*, pp. 4193–4197.

[21] D. Ghosh, A. Gupta, and P. Mohapatra, "Admission control and interference-aware scheduling in multi-hop WiMAX networks," in *Proc. 2007 IEEE MASS*, pp. 1–9.

[22] T.-C. Tsai and C.-Y. Wang, "Routing and admission control in IEEE 802.16 distributed mesh networks," in *Proc. 2007 IFIP Int'l Conf. Wireless Optical Commun. Netw.*, pp. 1–5.

[23] H. Zhu, V. O. K. Li, Z. Ma, and M. Zhao, "Statistical connection admission control framework based on achievable capacity estimation," in *Proc. 2006 IEEE ICC*, vol. 2, pp. 748–753.

[24] C. H. Liu, A. Gkelias, and K. K. Leung, "A cross-layer framework of QoS routing and distributed scheduling for mesh networks," in *Proc. 2008 IEEE VTC – Spring*, pp. 2193–2197.

[25] T. Apostol, *Calculus*. John Wiley & Sons, Inc., 1967.

[26] OPNET Inc. Available: http://www.opnet.com.

[27] B. Sklar, "Rayleigh fading channels in mobile digital communication systems—I: characterization," *IEEE Commun. Mag.*, vol. 35, no. 7, pp. 90–100, July 1997.

[28] I. F. Akyildiz and X. Wang, "A survey on wireless mesh networks," *IEEE Commun. Mag.*, vol. 43, no. 9, pp. S23–S30, Sept. 2005.

[29] X. Yuan and Z. Duan, "FRR: a proportional and worst-case fair round robin scheduler," in *Proc. 2005 IEEE INFOCOM*, vol. 2, pp. 831–842.

[30] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing," in *Proc. 1999 IEEE WMCSA*, pp. 90–100.

[31] R. Fantacci and D. Tarchi, "Bridging solutions for a heterogeneous wimax-wifi scenario," *J. Commun. Netw.*, vol. 8, no. 4, pp. 369–377, 2006.

Chi Harold Liu (S'05-M'10) is an Associate Professor and Department Head of Software Service Engineering at Beijing Institute of Technology, China. He holds a Ph.D. degree from Imperial College, U.K., and a B.Eng. degree from Tsinghua University, China. Before moving to academia, he joined IBM Research - China as a staff researcher and project manager, after working as a Postdoctoral researcher at Deutsche Telekom Laboratories, Germany, and a visiting scholar at IBM T.J. Watson Research Center, USA. His current research interests include the Internet-of-Things (IoT), big data analytics, mobile computing, and wireless ad-hoc, sensor and mesh networks. He receives the Distinguished Young Scholar Award in 2013, IBM First Plateau Invention Achievement Award in 2012, IBM First Patent Application Award in 2011, and interviewed by EEWeb.com as the Featured Engineer in 2011. He has published more than 30 prestigious conference and journal papers, and owned a few EU/US/China patents. He serves as the editor for KSII Trans. on Internet and Information Systems, and the book editor for four books published by Taylor & Francis Group, USA. He also has served as the General Chair of IEEE SECON'13 workshop on IoT Networking and Control, IEEE WCNC'12 workshop on IoT Enabling Technologies, and ACM UbiComp'11 Workshop on Networking and Object Memories for IoT. He is a member of IEEE and ACM.

Kin K. Leung (F'01) received his B.S. degree from the Chinese University of Hong Kong in 1980, and his M.S. and Ph.D. degrees from University of California, Los Angeles, in 1982 and 1985, respectively.

He joined AT&T Bell Labs in New Jersey in 1986 and worked at its successor companies, AT&T Labs and Bell Labs of Lucent Technologies, until 2004. Since then, he has been the Tanaka Chair Professor in the Electrical and Electronic Engineering (EEE), and Computing Departments at Imperial College in London. He serves as the Head of Communications and Signal Processing Group in the EEE Department at Imperial. His research interests focus on networking, protocols, optimization and modeling issues of wireless broadband, sensor and ad-hoc networks. He also works on multi-antenna and cross-layer designs for the physical layer of these networks.

He received the Distinguished Member of Technical Staff Award from AT&T Bell Labs in 1994, and was a co-recipient of the 1997 Lanchester Prize Honorable Mention Award. He was elected as an IEEE Fellow in 2001. He received the Royal Society Wolfson Research Merits Award from 2004 to 2009 and became a member of Academia Europaea in 2012. He has actively served on many conference committees. He serves as a member (2009-11) and the chairman (2012-13) of the IEEE Fellow Evaluation Committee for Communications Society. He was a guest editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC), *IEEE Wireless Communications* and the MONET journal, and as an editor for the JSAC: Wireless Series, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE TRANSACTIONS ON COMMUNICATIONS. Currently, he is an editor for the *ACM Computing Survey* and *International Journal on Sensor Networks*.

Athanasios Gkelias (S'05-M'08) received his Ph.D. and M.Sc from King's College London, UK in 2005 and 2001 respectively and his Diploma in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece, in 2000. Currently, he is a post-doctoral researcher at Imperial College London. In the past he served as the project manager of the University Defense Research Centre (UDRC) in Signal Processing at Imperial College, sponsored by the UK Ministry of Defence (MoD). He also participated in several ICT funded projects such as Mobile-VCE, IBM-ITA, MEMBRANE, e-SENSE and MIND. In the summer of 2008 he was at the Bell-Labs Research Centre, Alcatel-Lucent, UK, working as a visiting researcher on wireless mesh networks. He has published more than 40 peer-reviewed journal and conference papers and has been TPC member and in the organizing committee of various international conferences.