Effective Capacity-Based Quality of Service Measures for Wireless Networks

Dapeng Wu* Rohit Negi†

Abstract

An important objective of next-generation wireless networks is to provide quality of service (QoS) guarantees. This requires a simple and efficient wireless channel model that can easily translate into connection-level QoS measures such as data rate, delay and delay-violation probability. To achieve this, in [8], we developed a link-layer channel model termed *effective capacity*, for the setting of a single hop, constant-bit-rate arrivals, fluid traffic, and wireless channels with negligible propagation delay. In this paper, we apply the effective capacity technique to deriving QoS measures for more general situations, namely, 1) networks with multiple wireless links, 2) variable-bit-rate sources, 3) packetized traffic, and 4) wireless channels with non-negligible propagation delay.

Key Words: Wireless channel model, QoS, delay, effective capacity, large deviations theory.

*Please direct all correspondence to Dapeng Wu, University of Florida, Dept. of Electrical & Computer Engineering, P.O.Box 116130, Gainesville, FL 32611, USA. Tel. (352) 392-4954, Fax (352) 392-0044, Email: wu@ece.ufl.edu. URL: http://www.wu.ece.ufl.edu.

†Carnegie Mellon University, Dept. of Electrical & Computer Engineering, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA. Tel. (412) 268-6264, Fax (412) 268-2860, Email: negi@ece.cmu.edu. URL: http://www.ece.cmu.edu/~negi.

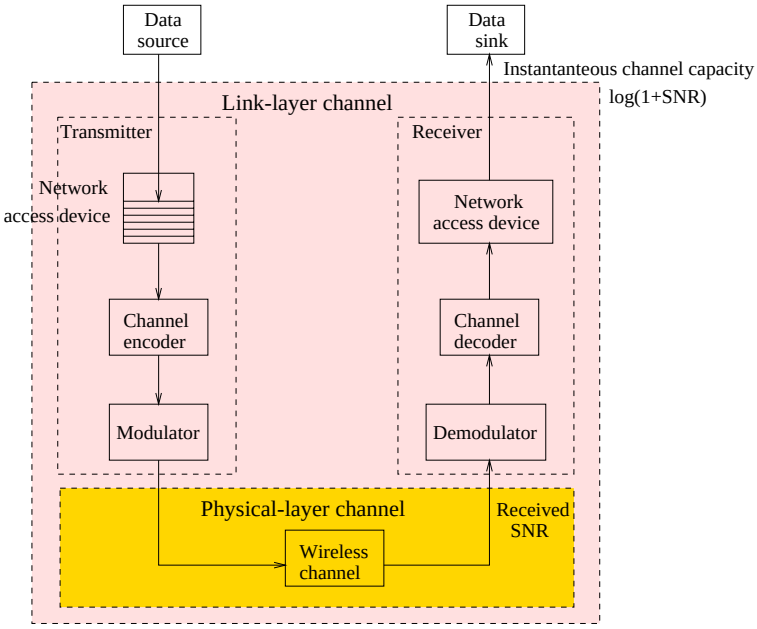

Figure 1: A wireless communication system.

1 Introduction

Providing QoS guarantees is crucial in the development of next-generation packet-based wireless communication networks [4]. To support QoS guarantees, QoS provisioning mechanisms are required. A major problem in designing QoS provisioning mechanisms is the *high complexity* in characterizing the relation between the control parameters of QoS provisioning mechanisms, and the calculated QoS measures, based on existing channel models, *i.e.*, physical-layer channel models (see Fig. 1). This is because the physical-layer channel models (*e.g.*, Rayleigh fading model with a specified Doppler spectrum) do not explicitly characterize a wireless channel in terms of the link-level QoS metrics specified by users, such as data rate, delay and delay-violation probability. To use the physical-layer channel models for QoS support, we first need to estimate the parameters for the channel model, and then extract the link-level QoS metrics from the model. This two-step approach is obviously complex, and may lead to inaccuracies due to possible approximations in extracting QoS metrics from the models.

Recognizing that the limitation of physical-layer channel models in QoS support, is the

difficulty in analyzing queues using them, in [8], we proposed moving the channel model up the protocol stack, from the physical-layer to the link-layer. We call the resulting model an *effective capacity* (EC) channel model [8], because it captures a generalized link-level capacity notion of the fading channel. Figure 1 illustrates the difference between the conventional physical-layer channel and the link-layer channel. In [8], we presented the EC channel model under the setting of a single hop, constant-bit-rate arrivals, fluid traffic, and wireless channels with negligible propagation delay; in this paper, we use the effective capacity technique to derive QoS measures for more general situations, namely, 1) networks with multiple wireless links, 2) variable-bit-rate sources, 3) packetized traffic, and 4) wireless channels with non-negligible propagation delay.

The remainder of this paper is organized as follows. In Section 2, we present preliminary results to familiarize the reader with the effective capacity technique. Sections 3 to 6 present effective capacity-based QoS measures for networks with multiple wireless links, variable-bit-rate sources, packetized traffic, and wireless channels with non-negligible propagation delay, respectively. Section 7 concludes the paper.

2 Preliminaries

We first formally define statistical QoS, which characterizes the requirement of a user. First, consider a single-hop system, where the user is allotted a single time varying channel. Assume that the user source has a fixed rate r_s and a specified delay bound D_{max}, and requires that the delay-bound violation probability is not greater than a certain value ε, that is,

$$Pr\{D(\infty) > D_{max}\} \leq \varepsilon, \tag{1}$$

where $D(\infty)$ is the steady-state delay experienced by a flow, and $Pr\{D(\infty) > D_{max}\}$ is the probability of $D(\infty)$ exceeding a delay bound D_{max}. Then, we say that the user is specified by the (statistical) QoS triplet $\{r_s, D_{max}, \varepsilon\}$. Even for this simple case, it is not immediately obvious as to which QoS triplets are feasible, for the given channel, since a rather complex queueing system (with an arbitrary channel capacity process) will need to be analyzed. The key contribution of [8] was to introduce a concept of statistical delay-constrained capacity

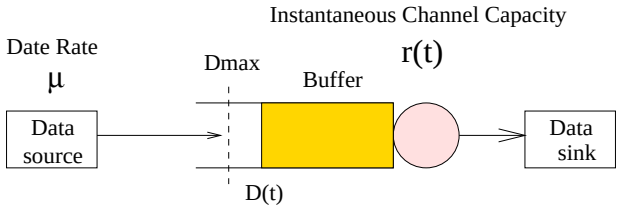

Figure 2: A queueing system model.

termed *effective capacity*, which allows us to obtain a simple and efficient test, to check the feasibility of QoS triplets for a single time-varying channel. That paper did not deal with general situations, *e.g.*, networks with multiple wireless links and multi-hops, variable-bit-rate sources, packetized traffic, and wireless channels with non-negligible propagation delay, which we consider in this paper.

Next, we briefly explain the concept of effective capacity, and refer the reader to [8] for details.

Let $r(t)$ be the instantaneous channel capacity at time t. Assume that, the asymptotic log-moment generation function of $r(t)$

$$\Lambda(-u) = \lim_{t\to\infty} \frac{1}{t} \log E[e^{-u\int_0^t r(\tau)\mathrm{d}\tau}] \tag{2}$$

exists for all $u \geq 0$. Then, the *effective capacity function* of $r(t)$ is defined as

$$\alpha(u) = \frac{-\Lambda(-u)}{u} \qquad , \ \forall\ u > 0. \tag{3}$$

That is,

$$\alpha(u) = -\lim_{t\to\infty} \frac{1}{ut} \log E[e^{-u\int_0^t r(\tau)\mathrm{d}\tau}], \ \forall\ u > 0. \tag{4}$$

Consider a queue of infinite buffer size supplied by a data source of *constant* data rate μ (see Fig. 2). It can be shown [8] that if $\alpha(u)$ indeed exists (*e.g.*, for ergodic, stationary, Markovian $r(t)$), then the probability of $D(\infty)$ exceeding a delay bound D_{max} satisfies

$$Pr\{D(\infty) > D_{max}\} \approx e^{-\theta(\mu)D_{max}}, \tag{5}$$

where the function $\theta(\mu)$ of source rate μ depends only on the channel capacity process $r(t)$. $\theta(\mu)$ can be considered as a "channel model" that models the channel at the link layer (in contrast to "physical layer" models specified by Markov processes, or Doppler spectra). The approximation (5) is accurate for large D_{max}.

In terms of the effective capacity function (4) defined earlier, the *QoS exponent function* $\theta(\mu)$ can be written as [8]

$$\theta(\mu) = \mu\alpha^{-1}(\mu) \tag{6}$$

where $\alpha^{-1}(\cdot)$ is the inverse function of $\alpha(u)$. Once $\theta(\mu)$ has been measured for a given channel, it can be used to check the feasibility of QoS triplets. Specifically, a QoS triplet $\{r_s, D_{max}, \varepsilon\}$ is feasible if $\theta(r_s) \geq \rho$, where $\rho \doteq -\log\varepsilon/D_{max}$. Thus, we can use the effective capacity model $\alpha(u)$ (or equivalently, the function $\theta(\mu)$ via (6)) to relate the channel capacity process $r(t)$ to statistical QoS. Since our effective capacity method predicts an exponential dependence (5) between ε and D_{max}, we can henceforth consider the QoS *pair* $\{r_s, \rho\}$ to be equivalent to the QoS triplet $\{r_s, D_{max}, \varepsilon\}$, with the understanding that $\rho = -\log\varepsilon/D_{max}$.

In the following sections, we extend the effective capacity technique to more general situations. The following property is needed in the propositions in the rest of this paper.

Property 1 *(i) The asymptotic log-moment generation function $\Lambda(u)$ defined in (2) is finite for all $u \in \mathbb{R}$. (ii) $\Lambda(u)$ is differentiable for all $u \in \mathbb{R}$.*

3 QoS Measures for Wireless Networks

In this section, we consider two basic network structures for wireless networks: one with only tandem wireless links (see Figure 3) and the other with only parallel wireless links (see Figure 4). In the following, Propositions 1 and 2 give QoS measures for these two network structures, respectively.

Denote $r_k(t)$ $(k = 1, \cdots, K)$ the instantaneous capacity of channel k at time t. For a

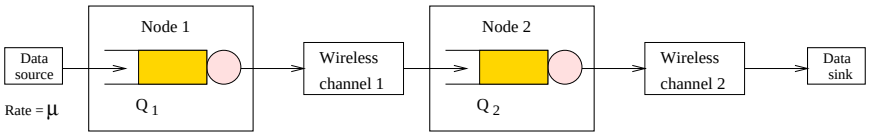

Figure 3: A network with tandem wireless links.

network with K tandem links, define the service $\tilde{S}(t_0, t)$, for $t \geq 0$ and any $t_0 \in [0, t]$, by

$$\tilde{S}(t_0, t) = \inf_{t_0 \leq t_1 \leq \cdots \leq t_{K-1} \leq t_K = t} \left\{ \sum_{k=1}^{K} \int_{t_{k-1}}^{t_k} r_k(\tau) \mathrm{d}\tau \right\}, \tag{7}$$

and the asymptotic log-moment generating function

$$\Lambda_{tandem}(-u) = \lim_{t \to \infty} \frac{1}{t} \log E[e^{-u\tilde{S}(0,t)}] \tag{8}$$

where $\tilde{S}(0, t)$ is defined by (7); also define the effective capacity of channel k by

$$\alpha_k(u) = -\lim_{t \to \infty} \frac{1}{ut} \log E[e^{-u \int_0^t r_k(\tau) \mathrm{d}\tau}], \ \forall \ u > 0. \tag{9}$$

Proposition 1 *Assume that the log-moment generating function $\Lambda_{tandem}(u)$ defined by (8) satisfies Property 1. Given the effective capacity functions $\{\alpha_k(u), k = 1, \cdots, K\}$ of K tandem links and an external arrival process with constant rate μ, the end-to-end delay $D(\infty)$ experienced by the traffic traversing the K tandem links satisfies*

$$\limsup_{D_{max} \to \infty} \frac{1}{D_{max}} \log Pr\{D(\infty) > D_{max}\} \leq -\theta, \qquad \text{if } \alpha(\theta/\mu) > \mu, \tag{10}$$

and

$$\lim_{D_{max} \to \infty} \frac{1}{D_{max}} \log Pr\{D(\infty) > D_{max}\} = -\theta^*, \qquad \text{where } \alpha(\theta^*/\mu) = \mu, \tag{11}$$

where $\alpha(u) = -\Lambda_{tandem}(-u)/u$. Moreover, the effective capacity $\alpha(u)$ satisfies

$$\alpha(u) \leq \min_k \alpha_k(u). \tag{12}$$

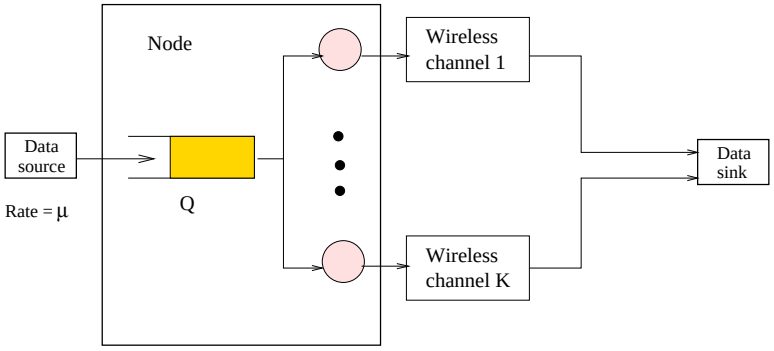

Figure 4: A network with parallel wireless links.

For a proof of Proposition 1, see the Appendix. Note that the capacity processes of the tandem channels are not required to be independent in Proposition 1.

Proposition 2 *Assume that the log-moment generating function $\Lambda_k(u)$ of each channel k in the network satisfies Property 1. Given the effective capacity functions $\{\alpha_k(u), k = 1, \cdots, K\}$ of K independent parallel links and an external arrival process with constant rate μ, the end-to-end delay $D(\infty)$ experienced by the traffic traversing the K parallel links satisfies*

$$\limsup_{D_{max}\to\infty} \frac{1}{D_{max}} \log Pr\{D(\infty) > D_{max}\} \leq -\theta, \qquad \text{if } \alpha(\theta/\mu) > \mu, \tag{13}$$

and

$$\lim_{D_{max}\to\infty} \frac{1}{D_{max}} \log Pr\{D(\infty) > D_{max}\} = -\theta^*, \qquad \text{where } \alpha(\theta^*/\mu) = \mu, \tag{14}$$

where $\alpha(u) = \sum_{k=1}^{K} \alpha_k(u)$.

For a proof of Proposition 2, see the Appendix.

Propositions 1 and 2 suggest the following approximation

$$Pr\{D(\infty) > D_{max}\} \approx e^{-\theta^* \times D_{max}}, \tag{15}$$

for large D_{max}. In addition, $\alpha(u)$ specified in Propositions 1 and 2 can be regarded as the effective capacity of the equivalent channel of the network, which consists of tandem links

only or independent parallel links only. In Sections 4 to 6, we will use $\alpha(u)$ to characterize the equivalent channel of the network; and we will use (14) only since (14) is tighter than (13).

4 QoS Measures for Variable-Bit-Rate Sources

In this section, we develop QoS measures for the case where the sources generate traffic at variable bit-rates (VBR). We consider two classes of VBR sources: leaky-bucket constrained arrival [2][9, page 15] and exponential process with its effective bandwidth function known [1][9, page 16]. Propositions 3 and 4 provide QoS measures for these two classes of VBR sources, respectively.

Proposition 3 *Assume that a wireless network consists of tandem links only or independent parallel links only; the effective capacity function of the equivalent channel of the wireless network is characterized by $\alpha(u)$; and the log-moment generating function $\Lambda_k(u)$ of each channel k in the network satisfies Property 1. Given an external arrival process constrained by a leaky bucket with bucket size $\sigma^{(s)}$ and token generating rate $\lambda_s^{(s)}$, the end-to-end delay $D(\infty)$ experienced by the traffic traversing the network satisfies*

$$\lim_{D_{max}\to\infty} \frac{1}{D_{max} - \sigma^{(s)}/\lambda_s^{(s)}} \log Pr\{D(\infty) > D_{max}\} = -\theta^*, \text{ where } \alpha(\theta^*/\lambda_s^{(s)}) = \lambda_s^{(s)}. \quad (16)$$

For a proof of Proposition 3, see the Appendix. Eq. (16) suggests the following approximation

$$Pr\{D(\infty) > D_{max}\} \approx e^{-\theta^* \times (D_{max} - \sigma^{(s)}/\lambda_s^{(s)})}, \quad (17)$$

for large D_{max}.

For convenience, we replicate the definition of the effective bandwidth [1] here. Consider an arrival process $\{A(t),\, t \geq 0\}$ where $A(t)$ represents the amount of source data (in bits) over the time interval $[0,\, t)$. Assume that the asymptotic log-moment generating function

of a stationary process $A(t)$, defined as

$$\Lambda(u) = \lim_{t\to\infty} \frac{1}{t} \log E[e^{uA(t)}], \tag{18}$$

exists for all $u \geq 0$. Then, the *effective bandwidth function* of $A(t)$ is defined as

$$\alpha^{(s)}(u) = \frac{\Lambda(u)}{u} \qquad , \ \forall \ u > 0. \tag{19}$$

Proposition 4 *Assume that a wireless network consists of tandem links only or independent parallel links only; the effective capacity function of the equivalent channel of the wireless network is characterized by $\alpha(u)$; an external arrival process is characterized by its effective bandwidth function $\alpha^{(s)}(u)$; and the log-moment generating function $\Lambda_k(u)$ of each channel k in the network and the log-moment generating function $\Lambda^{(s)}(u)$ of the external arrival process satisfy Property 1. Denote u^* the unique solution of the following equation*

$$\alpha^{(s)}(u) = \alpha(u). \tag{20}$$

The end-to-end delay $D(\infty)$ experienced by the traffic traversing the network satisfies

$$\lim_{D_{max}\to\infty} \frac{1}{D_{max}} \log Pr\{D(\infty) > D_{max}\} = -\theta^*, \ \textit{where } \theta^* = u^* \times \alpha^{(s)}(u^*). \tag{21}$$

For a proof of Proposition 4, see the Appendix. Note that a single-link network is a special case in Propositions 3 and 4.

5 QoS Measures for Packetized Traffic

In previous sections, we assumed fluid traffic. In this section, we extend the QoS measures obtained previously for the fluid model to the case with packetized traffic. This is important since in practical situations, the packet size is not negligible (not infinitesimal as in fluid model).

We assume the propagation delay of a wireless link is negligible, and the service at a network node is *non-cut-through*, *i.e.*, no packet is eligible for service until its last bit has arrived. We also assume a wireless network consists of tandem links only or parallel links only. For a network with tandem links only, the number of hops in the network is determined by the number of tandem links in the network; for a network with parallel links only, the number of hops in the network is one. We consider two cases: 1) a constant-bit-rate source with constant packet size, and 2) a variable-bit-rate source with variable packet size. Propositions 5 and 6 give QoS measures for these two cases, respectively.

Proposition 5 *Assume that a wireless network consists of tandem links only or independent parallel links only; the effective capacity function of the equivalent channel of the wireless network is characterized by $\alpha(u)$; the log-moment generating function $\Lambda_k(u)$ of each channel k in the network satisfies Property 1; and the network consists of N hops. Given an external arrival process with constant bit rate μ and constant packet size L_c, the end-to-end delay $D(\infty)$ experienced by the traffic traversing the network satisfies*

$$\lim_{D_{max}\to\infty} \frac{1}{D_{max} - N \times L_c/\mu} \log Pr\{D(\infty) > D_{max}\} = -\theta^*, \text{ where } \alpha(\theta^*/\mu) = \mu. \tag{22}$$

For a proof of Proposition 5, see the Appendix. Eq. (22) suggests the following approximation

$$Pr\{D(\infty) > D_{max}\} \approx e^{-\theta^*(D_{max} - N \times L_c/\mu)}, \tag{23}$$

for large D_{max}.

Proposition 6 *Assume that a wireless network consists of tandem links only or independent parallel links only; the effective capacity function of the equivalent channel of the wireless network is characterized by $\alpha(u)$; the log-moment generating function $\Lambda_k(u)$ of each channel k in the network satisfies Property 1; and the network consists of N hops. Given a traffic flow having maximum packet size L_{max} and constrained by a leaky bucket with bucket size $\sigma^{(s)}$ and token generating rate $\lambda_s^{(s)}$, the end-to-end delay $D(\infty)$ experienced by the traffic traversing the network satisfies*

$$\lim_{D_{max}\to\infty} \frac{1}{D_{max} - N \times L_{max}/\lambda_s^{(s)} - \sigma^{(s)}/\lambda_s^{(s)}} \log Pr\{D(\infty) > D_{max}\} = -\theta^*, \tag{24}$$

where $\alpha(\theta^*/\lambda_s^{(s)}) = \lambda_s^{(s)}$.

For a proof of Proposition 6, see the Appendix. Eq. (22) suggests the following approximation

$$Pr\{D(\infty) > D_{max}\} \approx e^{-\theta^*(D_{max} - N \times L_{max}/\lambda_s^{(s)} - \sigma^{(s)}/\lambda_s^{(s)})}, \tag{25}$$

for large D_{max}. Note that a single-link network is a special case in Propositions 5 and 6.

6 QoS Measures for Wireless Channels with Non-negligible Propagation Delay

In previous sections, we assumed the propagation delay of a wireless link is negligible. In this section, we extend the QoS measures obtained previously to the situation where the propagation delay of a wireless link is not negligible. We consider two cases: 1) a fluid source with a constant rate, and 2) a variable-bit-rate source with variable packet size. Propositions 7 and 8 give QoS measures for these two cases, respectively.

Proposition 7 *Assume that a wireless network consists of tandem links only or independent parallel links only; the effective capacity function of the equivalent channel of the wireless network is characterized by* $\alpha(u)$*; the log-moment generating function* $\Lambda_k(u)$ *of each channel* k *in the network satisfies Property 1; the network consists of* N *hops; and the* i*-th hop* $(i = 1, \cdots, N)$ *incurs a constant propagation delay* d_i*. Given a fluid traffic flow with constant rate* μ*, the end-to-end delay* $D(\infty)$ *experienced by the traffic traversing the network satisfies*

$$\lim_{D_{max} \to \infty} \frac{1}{D_{max} - \sum_{i=1}^{N} d_i} \log Pr\{D(\infty) > D_{max}\} = -\theta^*, \ \textit{where } \alpha(\theta^*/\mu) = \mu. \tag{26}$$

For a proof of Proposition 7, see the Appendix. Eq. (26) suggests the following approximation

$$Pr\{D(\infty) > D_{max}\} \approx e^{-\theta^*(D_{max} - \sum_{i=1}^{N} d_i)}, \tag{27}$$

for large D_{max}.

Proposition 8 *Assume that a wireless network consists of tandem links only or independent parallel links only; the effective capacity function of the equivalent channel of the wireless network is characterized by $\alpha(u)$; the log-moment generating function $\Lambda_k(u)$ of each channel k in the network satisfies Property 1; the network consists of N hops; and the i-th hop $(i = 1, \cdots, N)$ incurs a constant propagation delay d_i. Given a traffic flow having maximum packet size L_{max} and constrained by a leaky bucket with bucket size $\sigma^{(s)}$ and token generating rate $\lambda_s^{(s)}$, the end-to-end delay $D(\infty)$ experienced by the traffic traversing the network satisfies*

$$\lim_{D_{max}\to\infty} \frac{1}{D_{max} - N \times L_{max}/\lambda_s^{(s)} - \sigma^{(s)}/\lambda_s^{(s)} - \sum_{i=1}^{N} d_i} \log Pr\{D(\infty) > D_{max}\} = -\theta^*, \quad (28)$$

where $\alpha(\theta^*/\lambda_s^{(s)}) = \lambda_s^{(s)}$.

For a proof of Proposition 8, see the Appendix. Eq. (28) suggests the following approximation

$$Pr\{D(\infty) > D_{max}\} \approx e^{-\theta^*(D_{max} - N\times L_{max}/\lambda_s^{(s)} - \sigma^{(s)}/\lambda_s^{(s)} - \sum_{i=1}^{N} d_i)}, \quad (29)$$

for large D_{max}.

7 Concluding Remarks

The design of QoS provisioning mechanisms in wireless networks calls for a simple and effective wireless channel model. In [8], we proposed and developed such a simple and effective channel model, called effective capacity, for the setting of a single hop, constant-bit-rate arrivals, fluid traffic, and wireless channels with negligible propagation delay. In this paper, we employed the effective capacity technique to derive QoS measures for more general situations, *i.e.*, networks with multiple wireless links, variable-bit-rate sources, packetized traffic, and wireless channels with non-negligible propagation delay.

In our future work, the QoS measures developed in this paper will be used to design efficient mechanisms to provide end-to-end QoS guarantees in a multihop wireless network. This will involve developing algorithms for QoS routing, resource reservation, admission control and scheduling.

Acknowledgment

This work was supported by the National Science Foundation under the grant ANI-0111818.

Appendix

Proof of Proposition 1

Denote $Q_k(t)$ the queue length at time t at node k ($k = 1, \cdots, K$), $Q(t)$ the end-to-end queue length at time t, $Q(\infty)$ the steady state of the end-to-end queue length, $A(t_0, t)$ the amount of arrival to node 1 (see Figure 3) over the time interval $[t_0, t]$. Define $\tilde{S}_k(t_0, t) = \int_{t_0}^{t} r_k(\tau)\mathrm{d}\tau$, which is the service provided by channel k over the time interval $[t_0, t]$.

We first prove an upper bound. It can be proved [10, page 81] that

$$Q(t) = \sum_{k=1}^{K} Q_k(t) = \sup_{0 \le t_0 \le t} \Big\{ A(t_0, t) - \tilde{S}(t_0, t) \Big\} \tag{30}$$

where $\tilde{S}(t_0, t)$ is defined by (7). Without loss of generality, we consider the discrete time case only, *i.e.*, $t \in \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers. From (30) and Loynes Theorem [6], we obtain

$$Q(\infty) = \sup_{t \in \mathbb{N}} \Big\{ A(0, t) - \tilde{S}(0, t) \Big\} = \sup_{t \in \mathbb{N}} \Big\{ \mu t - \tilde{S}(0, t) \Big\} \tag{31}$$

Then, we have

$$\begin{aligned} Pr\left\{Q(\infty) > q\right\} &= Pr\left\{\sup_{t\in\mathbb{N}}\left\{\mu t - \tilde{S}(0,t)\right\} > q\right\} && (32)\\ &\overset{(a)}{\leq} Pr\left\{\bigcup_{t\in\mathbb{N}}\left\{\mu t - \tilde{S}(0,t) > q\right\}\right\} && (33)\\ &\overset{(b)}{\leq} \sum_{t\in\mathbb{N}} Pr\left\{\mu t - \tilde{S}(0,t) > q\right\} && (34)\\ &\overset{(c)}{\leq} \sum_{t\in\mathbb{N}} e^{-uq} E[e^{u(\mu t - \tilde{S}(0,t))}] && (35) \end{aligned}$$

where (a) since the event $\left\{\sup_{t\in\mathbb{N}}\left\{\mu t - \tilde{S}(0,t)\right\} > q\right\} \subset \bigcup_{t\in\mathbb{N}}\left\{\mu t - \tilde{S}(0,t) > q\right\}$, (b) is due to the union bound, and (c) from the Chernoff bound. Since $\alpha(u) = -\Lambda_{tandem}(-u)/u$, we have

$$\alpha(u) = -\lim_{t\to\infty}\frac{1}{ut}\log E[e^{-u\tilde{S}(0,t)}],\ \forall\ u > 0, \tag{36}$$

Hence, for any $\epsilon > 0$, there exists a number $\tilde{t} > 0$ such that for $t \geq \tilde{t}$, we have

$$E[e^{-u\tilde{S}(0,t)}] \leq e^{u(-\alpha(u)+\epsilon)t}, \qquad \forall u > 0. \tag{37}$$

If $\mu + \epsilon < \alpha(u)$, we have

$$\begin{aligned} \sum_{t\in\mathbb{N}} e^{-uq} E[e^{u(\mu t - \tilde{S}(0,t))}] &\overset{(a)}{\leq} \sum_{t\geq\tilde{t}} e^{-uq} e^{u(\mu-\alpha(u)+\epsilon)t} + \sum_{t=1}^{\tilde{t}-1} e^{-uq} E[e^{u(\mu t - \tilde{S}(0,t))}] && (38)\\ &\overset{(b)}{\leq} e^{-uq} \times \left(\frac{e^{u(\mu-\alpha(u)+\epsilon)\tilde{t}}}{1 - e^{u(\mu-\alpha(u)+\epsilon)}} + \sum_{t=1}^{\tilde{t}-1} E[e^{u(\mu t - \tilde{S}(0,t))}]\right) && (39) \end{aligned}$$

where (a) from (37), and (b) from geometric sum. From (35) and (39), we have

$$Pr\left\{Q(\infty) > q\right\} \leq \gamma \times e^{-uq}, \qquad \text{if } \mu + \epsilon < \alpha(u), \tag{40}$$

where γ is a constant independent of q. Hence, we obtain

$$\limsup_{q\to\infty} \frac{1}{q} \log Pr\left\{Q(\infty) > q\right\} \quad \leq \quad -u, \qquad \text{if } \mu + \epsilon < \alpha(u). \tag{41}$$

Letting $\epsilon \to 0$, we have

$$\limsup_{q\to\infty} \frac{1}{q} \log Pr\left\{Q(\infty) > q\right\} \quad \leq \quad -u, \qquad \text{if } \mu < \alpha(u). \tag{42}$$

Since $Q(t) = \mu \times D(t),\ \forall t \geq 0$, (42) results in

$$\limsup_{D_{max}\to\infty} \frac{1}{D_{max}} \log Pr\{D(\infty) > D_{max}\} \leq -\mu \times u, \qquad \text{if } \mu < \alpha(u). \tag{43}$$

Let $\theta = \mu \times u$. Then (43) becomes (10).

Next we prove a lower bound. Let $q = \beta t$ where $\beta > 0$. Then, we have

$$\begin{aligned}
\liminf_{q\to\infty} \frac{1}{q} \log Pr\left\{Q(\infty) > q\right\} &= \liminf_{t\to\infty} \frac{1}{\beta t} \log Pr\left\{Q(\infty) > \beta t\right\} && (44)\\
&= \liminf_{t\to\infty} \frac{1}{\beta t} \log Pr\left\{\sup_{t\in\mathbb{N}} \left\{\mu t - \tilde{S}(0,t)\right\} > \beta t\right\} && (45)\\
&\geq \liminf_{t\to\infty} \frac{1}{\beta t} \log Pr\left\{\mu t - \tilde{S}(0,t) > \beta t\right\} && (46)\\
&= \liminf_{t\to\infty} \frac{1}{\beta t} \log Pr\left\{\frac{-\tilde{S}(0,t)}{t} > \beta - \mu\right\} && (47)\\
&\overset{(a)}{\geq} -\frac{1}{\beta} \inf_{x>\beta-\mu} \Lambda^*(x) && (48)
\end{aligned}$$

where (a) from Gärtner-Ellis Theorem [1] since $\Lambda_{tandem}(u)$ satisfies Property 1, and the Legendre-Fenchel transform $\Lambda^*(x)$ of $\Lambda_{tandem}(-u)$ is defined by

$$\Lambda^*(x) = \sup_{u\in\mathbb{R}} \{u \times x - \Lambda_{tandem}(-u)\}. \tag{49}$$

Since (48) holds for any $\beta > 0$, we have

$$\liminf_{q\to\infty} \frac{1}{q} \log Pr\{Q(\infty) > q\} \geq \sup_{\beta>0} -\frac{1}{\beta} \inf_{x>\beta-\mu} \Lambda^*(x) \tag{50}$$

$$= -\inf_{y>-\mu} \frac{\Lambda^*(y)}{y+\mu} \tag{51}$$

It can be proved [3] that

$$\inf_{y>-\mu} \frac{\Lambda^*(y)}{y+\mu} = u^*, \qquad \text{where } \Lambda_{tandem}(-u^*) = -\mu \times u^*. \tag{52}$$

From the definition of effective capacity $\alpha(u)$ in (36), $\Lambda_{tandem}(-u^*) = -\mu \times u^*$ implies $\alpha(u^*) = \mu$. Then, applying (52) to (51) leads to

$$\liminf_{q\to\infty} \frac{1}{q} \log Pr\{Q(\infty) > q\} \geq u^*, \qquad \text{where } \alpha(u^*) = \mu. \tag{53}$$

Due to the continuity of $\alpha(u)$, we have

$$\lim_{u\to u^*} \alpha(u) = \mu \tag{54}$$

Hence, letting $u \to u^*$, (42) and (53) result in

$$u^* \leq \liminf_{q\to\infty} \frac{1}{q} \log Pr\{Q(\infty) > q\} \leq \limsup_{q\to\infty} \frac{1}{q} \log Pr\{Q(\infty) > q\} \leq u^* \tag{55}$$

Hence, we have

$$\lim_{q\to\infty} \frac{1}{q} \log Pr\{Q(\infty) > q\} = u^*, \qquad \text{where } \alpha(u^*) = \mu. \tag{56}$$

Since $Q(t) = \mu \times D(t),\ \forall t \geq 0$, (56) results in

$$\lim_{D_{max}\to\infty} \frac{1}{D_{max}} \log Pr\{D(\infty) > D_{max}\} = -\mu \times u^*, \qquad \text{where } \alpha(u^*) = \mu. \tag{57}$$

Let $\theta^* = \mu \times u^*$. Then (57) becomes (11).

From (7), it is obvious that

$$\tilde{S}(t_0,t) \leq \min_k \tilde{S}_k(t_0,t). \tag{58}$$

Then we have for $u > 0$,

$$\begin{aligned}
\alpha(u) &= -\lim_{t\to\infty} \frac{1}{ut} \log E[e^{-u\tilde{S}(0,t)}] && (59)\\
&\overset{(a)}{\leq} -\lim_{t\to\infty} \frac{1}{ut} \log E[e^{-u \min_k \tilde{S}_k(0,t)}] && (60)\\
&\leq \min_k -\lim_{t\to\infty} \frac{1}{ut} \log E[e^{-u\tilde{S}_k(0,t)}] && (61)\\
&= \min_k \alpha_k(u) && (62)
\end{aligned}$$

where (a) from (58). This completes the proof. ■

Proof of Proposition 2

Denote $r_k(t)$ $(k = 1, \cdots, K)$ channel capacity of link k at time t. From Figure 4, it is clear that the network has only one queue and multiple servers, each of which corresponds to a wireless link. Since the total instantaneous channel capacity $r(t) = \sum_{k=1}^{K} r_k(t)$, the effective capacity function for the aggregate parallel links is

$$\begin{aligned}
\alpha(u) &\overset{(a)}{=} -\lim_{t\to\infty} \frac{1}{ut} \log E[e^{-u\int_0^t r(\tau)\mathrm{d}\tau}]\\
&= -\lim_{t\to\infty} \frac{1}{ut} \log E[e^{-u\int_0^t \sum_{k=1}^{K} r_k(\tau)\mathrm{d}\tau}]\\
&\overset{(b)}{=} -\lim_{t\to\infty} \frac{1}{ut} \sum_{k=1}^{K} \log E[e^{-u\int_0^t r_k(\tau)\mathrm{d}\tau}]\\
&\overset{(c)}{=} \sum_{k=1}^{K} \alpha_k(u) && (63)
\end{aligned}$$

where (a) from (4), (b) since $\{r_k(t), k = 1, \cdots, K\}$ are independent, and (c) from (9).

Given the effective capacity $\alpha(u)$, we can prove (13) and (14) with the same technique used in proving (10) and (11). This completes the proof. ■

Proof of Proposition 3

For the traffic of constant rate $\lambda_s^{(s)}$, denote $\tilde{Q}(\infty)$ the steady state of the end-to-end queue length and $\tilde{D}(\infty)$ the end-to-end delay. Using the result in [5, page 30]), we can show

$$D(\infty) - \tilde{D}(\infty) \leq \sigma^{(s)}/\lambda_s^{(s)}, \tag{64}$$

Note that $D(\infty)$ is the end-to-end delay experienced by the traffic constrained by a leaky bucket with bucket size $\sigma^{(s)}$ and token generating rate $\lambda_s^{(s)}$. Hence, we have

$$D(\infty) \leq \tilde{D}(\infty) + \sigma^{(s)}/\lambda_s^{(s)}, \tag{65}$$

From (40) and $\tilde{Q}(\infty) = \lambda_s^{(s)} \times \tilde{D}(\infty)$, we have

$$Pr\left\{\tilde{D}(\infty) > D_{max}\right\} \leq \gamma \times e^{-u \times \lambda_s^{(s)} \times D_{max}}, \qquad \text{if } \alpha(u) > \lambda_s^{(s)}, \tag{66}$$

where γ is a constant independent of D_{max}. Then, we have

$$\begin{aligned} Pr\left\{D(\infty) > D_{max}\right\} &\overset{(a)}{=} Pr\left\{\tilde{D}(\infty) > D_{max} - \sigma^{(s)}/\lambda_s^{(s)}\right\} \\ &\overset{(b)}{\leq} \gamma \times e^{-u \times \lambda_s^{(s)} \times (D_{max} - \sigma^{(s)}/\lambda_s^{(s)})} \end{aligned} \tag{67}$$

where (a) from (65), and (b) from (66). Hence, we have

$$\limsup_{D_{max} \to \infty} \frac{1}{D_{max} - \sigma^{(s)}/\lambda_s^{(s)}} \log Pr\{D(\infty) > D_{max}\} \leq -\theta, \qquad \text{if } \alpha(\theta/\lambda_s^{(s)}) > \lambda_s^{(s)}. \tag{68}$$

Similar to the proof of Proposition 1, we can obtain a lower bound

$$\liminf_{D_{max} \to \infty} \frac{1}{D_{max} - \sigma^{(s)}/\lambda_s^{(s)}} \log Pr\{D(\infty) > D_{max}\} \geq -\theta^*, \qquad \text{where } \alpha(\theta^*/\lambda_s^{(s)}) = \lambda_s^{(s)}. \tag{69}$$

Combining (68) and (69), we obtain (16). This completes the proof. ■

Proof of Proposition 4

The proof is similar to that of Proposition 1.

Denote $Q(t)$ the end-to-end queue length at time t, $Q(\infty)$ the steady state of the end-to-end queue length, $A(t_0,t)$ the amount of external arrival to the network over the time interval $[t_0,t]$. From (30), we know

$$Q(t) = \sup_{0\le t_0\le t}\left\{A(t_0,t) - \tilde{S}(t_0,t)\right\} \tag{70}$$

where $\tilde{S}(t_0,t)$ is defined by (7) for the tandem links, and is defined by $\tilde{S}(t_0,t) = \int_0^t \sum_{k=1}^K r_k(\tau)\mathrm{d}\tau$ for independent parallel links.

We first prove an upper bound. Without loss of generality, we consider the discrete time case only, *i.e.*, $t \in \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers. From (70) and Loynes Theorem [6], we obtain

$$Q(\infty) = \sup_{t\in\mathbb{N}}\left\{A(0,t) - \tilde{S}(0,t)\right\} \tag{71}$$

Then, we have

$$\begin{aligned}
Pr\left\{Q(\infty) > q\right\} &= Pr\left\{\sup_{t\in\mathbb{N}}\left\{A(0,t) - \tilde{S}(0,t)\right\} > q\right\} &(72)\\
&\le Pr\left\{\bigcup_{t\in\mathbb{N}}\left\{A(0,t) - \tilde{S}(0,t) > q\right\}\right\} &(73)\\
&\le \sum_{t\in\mathbb{N}} Pr\left\{A(0,t) - \tilde{S}(0,t) > q\right\} &(74)\\
&\le \sum_{t\in\mathbb{N}} e^{-uq}E[e^{u(A(0,t)-\tilde{S}(0,t))}] &(75)
\end{aligned}$$

From the definition of effective capacity in (36), for any $\epsilon/2 > 0$, there exists a number $\tilde{t} > 0$ such that for $t \ge \tilde{t}$, we have

$$E[e^{-u\tilde{S}(0,t)}] \le e^{u(-\alpha(u)+\epsilon/2)t}, \qquad \forall u > 0. \tag{76}$$

Similarly, from the definition of effective bandwidth in (19), for any $\epsilon/2 > 0$, there exists a number $\tilde{t} > 0$ such that for $t \geq \tilde{t}$, we have

$$E[e^{uA(0,t)}] \leq e^{u(\alpha^{(s)}(u)+\epsilon/2)t}, \qquad \forall u > 0. \tag{77}$$

Without loss of generality, here we choose the same $\tilde{t}$ for both (76) and (77), since we can always choose the maximum of the two to make (76) and (77) hold. Then, if $\alpha^{(s)}(u) + \epsilon < \alpha(u)$, we have

$$\sum_{t\in\mathbb{N}} e^{-uq} E[e^{u(A(0,t)-\tilde{S}(0,t))}] \overset{(a)}{\leq} \sum_{t\geq\tilde{t}} e^{-uq} e^{u(\alpha^{(s)}(u)-\alpha(u)+\epsilon)t} + \sum_{t=1}^{\tilde{t}-1} e^{-uq} E[e^{u(A(0,t)-\tilde{S}(0,t))}] \tag{78}$$

$$\leq e^{-uq} \times \left(\frac{e^{u(\alpha^{(s)}(u)-\alpha(u)+\epsilon)\tilde{t}}}{1 - e^{u(\alpha^{(s)}(u)-\alpha(u)+\epsilon)}} + \sum_{t=1}^{\tilde{t}-1} E[e^{u(A(0,t)-\tilde{S}(0,t))}] \right) \tag{79}$$

where (a) from (76) and (77). From (75) and (79), we have

$$\limsup_{q\to\infty} \frac{1}{q} \log Pr\{Q(\infty) > q\} \leq -u, \text{ if } \alpha^{(s)}(u) + \epsilon < \alpha(u). \tag{80}$$

Letting $\epsilon \to 0$, we have

$$\limsup_{q\to\infty} \frac{1}{q} \log Pr\{Q(\infty) > q\} \leq -u, \text{ if } \alpha^{(s)}(u) < \alpha(u). \tag{81}$$

Similar to the proof of Proposition 1, we can obtain a lower bound

$$\liminf_{q\to\infty} \frac{1}{q} \log Pr\{Q(\infty) > q\} \geq -u^*, \text{ where } \alpha^{(s)}(u^*) = \alpha(u^*). \tag{82}$$

Combining (81) and (82), we have

$$\lim_{q\to\infty} \frac{1}{q} \log Pr\{Q(\infty) > q\} = -u^*, \text{ where } \alpha^{(s)}(u^*) = \alpha(u^*). \tag{83}$$

Since $Q(\infty) = \alpha^{(s)}(u^*) \times D(\infty)$, (83) results in (21). This completes the proof. ■

Proof of Proposition 5

For the packetized traffic, denote $Q_k(t)$ the queue length at time t at node k ($k = 1, \cdots, N$), $Q(t)$ the end-to-end queue length at time t, and $Q(\infty)$ the steady state of the end-to-end queue length. Correspondingly, for the 'fluid' traffic of constant arrival rate μ, denote $\tilde{Q}_k(t)$ the queue length at time t at node k, $\tilde{Q}(t)$ the end-to-end queue length at time t, $\tilde{Q}(\infty)$ the steady state of the end-to-end queue length, and $\tilde{D}(\infty)$ the end-to-end delay.

For each node k, we have the sample path relation as below [7]

$$Q_k(t) - \tilde{Q}_k(t) \leq L_c, \qquad \forall t \geq 0. \tag{84}$$

Summing up over k, we obtain

$$\sum_{k=1}^{N} [Q_k(t) - \tilde{Q}_k(t)] = Q(t) - \tilde{Q}(t) \leq N \times L_c, \qquad \forall t \geq 0. \tag{85}$$

Hence, for the steady state, we have

$$Q(\infty) - \tilde{Q}(\infty) \leq N \times L_c. \tag{86}$$

Since $Q(\infty) = \mu \times D(\infty)$ and $\tilde{Q}(\infty) = \mu \times \tilde{D}(\infty)$, we have

$$D(\infty) - \tilde{D}(\infty) \leq N \times L_c/\mu. \tag{87}$$

Note that $D(\infty)$ is the end-to-end delay experienced by the packetized traffic with constant bit rate μ and constant packet size L_c. Then, we can prove (22) in the same way as we prove (16) in Proposition 3. ■

Proof of Proposition 6

Denote $\tilde{D}(\infty)$ the end-to-end delay experienced by the 'fluid' traffic with constant arrival rate $\lambda_s^{(s)}$. Using the sample path relation in [5, page 35]), we obtain

$$D(\infty) - \tilde{D}(\infty) \leq N \times L_{max}/\lambda_s^{(s)} + \sigma^{(s)}/\lambda_s^{(s)}, \tag{88}$$

Note that $D(\infty)$ is the end-to-end delay experienced by the packetized traffic having maximum packet size L_{max} and constrained by a leaky bucket with bucket size $\sigma^{(s)}$ and token generating rate $\lambda_s^{(s)}$. Then, we can prove (24) in the same way as we prove (16) in Proposition 3. ■

Proof of Proposition 7

Denote $\tilde{D}(\infty)$ the end-to-end delay experienced by the fluid traffic with constant arrival rate μ and without propagation delay. Using the sample path relation between the two cases (with/without propagation delay), it is easy to show

$$D(\infty) - \tilde{D}(\infty) \leq \sum_{i=1}^{N} d_i, \tag{89}$$

Then, we can prove (26) in the same way as we prove (16) in Proposition 3. ■

Proof of Proposition 8

Denote $\tilde{D}(\infty)$ the end-to-end delay experienced by the 'fluid' traffic with constant arrival rate $\lambda_s^{(s)}$ and without propagation delay. Using the sample path relation in [5, page 35]), we obtain

$$D(\infty) - \tilde{D}(\infty) \leq N \times L_{max}/\lambda_s^{(s)} + \sigma^{(s)}/\lambda_s^{(s)} + \sum_{i=1}^{N} d_i, \tag{90}$$

Then, we can prove (28) in the same way as we prove (16) in Proposition 3. ■

References

[1] C.-S. Chang, "Performance guarantees in communication networks," Springer, 2000.

[2] R. L. Cruz, "A calculus for network delay, Part I: network elements in isolation," *IEEE Trans. on Information Theory,* vol. 37, no. 1, pp. 114–131, Jan. 1991.

[3] G. de Veciana and J. Walrand, "Effective bandwidths: call admission, traffic policing and filtering for ATM networks," *Queuing Systems,* vol. 20, pp. 37–59, 1995.

[4] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications,* Wiley, 2000.

[5] J.-Y. Le Boudec and P. Thiran, "Network calculus: a theory of deterministic queueing systems for the Internet," Springer, 2001.

[6] R. M. Loynes, "The stability of a queue with non-independent inter-arrivals and service times," *Proc. Camb. Phil. Soc.,* vol. 58, pp. 497–520, 1962.

[7] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single node case," *IEEE/ACM Trans. on Networking,* vol. 1, no. 3, pp. 344–357, June 1993.

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

[9] D. Wu, "Providing quality of service guarantees in wireless networks," *Ph.D. Dissertation,* Dept. of Electrical & Computer Engineering, Carnegie Mellon University, Aug. 2003. Available at `http://www.wu.ece.ufl.edu/mypapers/Thesis.pdf`.

[10] Z.-L. Zhang, "End-to-end support for statistical quality-of-service guarantees in multimedia networks," *Ph.D. Dissertation,* Department of Computer Science, University of Massachusetts, Feb. 1997.