Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Performance modeling of a bottleneck node in an IEEE 802.11 ad-hoc network

J.L. van den Berg, M.R.H. Mandjes, F. Roijers

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

Performance modeling of a bottleneck node in an IEEE 802.11 ad-hoc network

ABSTRACT

This paper presents a performance analysis of wireless ad-hoc networks, with IEEE 802.11 as the underlying Wireless LAN technology. WLAN has, due to the fair radio resource sharing at the MAC-layer, the tendency to share the capacity equally amongst the active nodes, irrespective of their loads. An inherent drawback of this sharing policy is that a node that serves as a relay-node for multiple flows is likely to become a bottleneck. This paper proposes to model such a bottleneck by a fluid-flow model. Importantly, this is a model at the flow-level: flows arrive at the bottleneck node, and are served according to the sharing policy mentioned above. Assuming Poisson initiations of new flow transfers, we obtain insightful, robust, and explicit expressions for characteristics related to the overall flow transfer time, the buffer occupancy, and the packet delay at the bottleneck node. The analysis is enabled by a translation of the buffer dynamics at the bottleneck node in terms of an M/G/1 queueing model. We conclude the paper by an assessment of the impact of alternative sharing policies (which can be obtained by the IEEE 802.11E version), in order to improve the performance of the bottleneck.

Performance Modeling of a Bottleneck Node
in an IEEE 802.11 Ad-hoc Network

Hans van den Berg[1,2], Michel Mandjes[3,4], and Frank Roijers[1,3]

[1] TNO Information and Communication Technology, the Netherlands
[2] Department of Design and Analysis of Communication Systems, University of Twente, the Netherlands
[3] Centre for Mathematics and Computer Science, the Netherlands
[4] Korteweg-de Vries Institute, University of Amsterdam, the Netherlands

Abstract

This paper presents a performance analysis of wireless ad-hoc networks, with IEEE 802.11 as the underlying wireless LAN technology. WLAN has, due to the fair radio resource sharing at the MAC-layer, the tendency to share the capacity equally amongst the active nodes, irrespective of their loads. An inherent drawback of this sharing policy is that a node that serves as a relay-node for multiple flows is likely to become a *bottleneck*. This paper proposes to model such a bottleneck by a fluid-flow model. Importantly, this is a model at the *flow-level*: flows arrive at the bottleneck node, and are served according to the sharing policy mentioned above. Assuming Poisson initiations of new flow transfers, we obtain insightful, robust, and explicit expressions for characteristics related to the overall flow transfer time, the buffer occupancy, and the packet delay at the bottleneck node. The analysis is enabled by a translation of the buffer dynamics at the bottleneck node in terms of an M/G/1 queueing model. We conclude the paper by an assessment of the impact of alternative sharing policies (which can be obtained by the IEEE 802.11E version), in order to improve the performance of the bottleneck.

1 Introduction

Developments in wireless communication technology open up the possibility of wireless ad-hoc networks; these networks can be deployed instantly without a fixed infrastructure or pre-advanced configuration and multi-hop connectivity is one of the key-features. Currently, IEEE 802.11 Wireless LAN [15] is the most popular technology used for wireless ad-hoc networks.

In IEEE 802.11 based ad-hoc networks stations have to contend for access to the wireless medium according to the DISTRIBUTED COORDINATION FUNCTION (DCF). DCF is a distributed random medium access mechanism based on CSMA/CA — for a more elaborate description, see for instance [2, 15]. In the literature it is shown that in a single-cell setting DCF tends to share the wireless medium capacity *equally* amongst contending stations, see e.g. [2, 18]. Clearly, DCF is particularly appropriate in the context of ad-hoc networks as it operates in a fully distributed fashion.

DCF has, however, also significant drawbacks. Most notably, it facilitates only equal sharing, i.e., it is not capable of granting stations different shares of the available capacity. The lack of such differentiation options may cause nodes to become bottlenecks, as can be seen as follows. Stations that cannot communicate directly with each other use other stations to relay their traffic; in particular, nodes that have a central location in the ad-hoc network are likely to become such a relay node. Let n stations send traffic via the same relay node. Then, due to the sharing policy, the relay-node obtains just the same share of the medium capacity, viz. $1/(n+1)$, as each of the 'sending nodes'. In other words, as soon as $n > 1$, the node's input rate exceeds its output rate, and hence the excess traffic

accumulates in the node's buffer; only when $n = 0$ the queue drains. This entails that these relay stations become the bottleneck, and will strongly affect the performance of the flow transfers through the network.

In this paper we study the performance of wireless ad-hoc networks focusing on the Medium Access Control (MAC) layer. In particular, the goal of this paper is to investigate the flow-level performance of a multi-hop flow, e.g. the transfer time of an entire flow consisting of multiple packets. The attractive feature of this performance measure is that it directly relates to the perspective of the users: the quality perceived by a user is primarily based on the time required to transmit the entire flow, rather than the delays or loss probabilities of individual packets. Unfortunately, as will be clear from the next paragraph, the literature does not cover analytical models for flow transfer times of multi-hop flows. For that purpose we develop and analyze a stochastic model of a simple ad-hoc network scenario capturing the essential characteristics of the resource sharing enforced by the IEEE 802.11 MAC protocol of a 2-hop flow.

In contrast with the modeling approach presented in our paper, almost all earlier studies on the performance of IEEE 802.11 multi-hop ad-hoc networks available in the literature are based on simulation, see e.g. the survey paper [3] and [11, 13]. These studies usually capture many details of the ad-hoc network protocols, but, unfortunately, are mostly limited with respect to the considered traffic scenarios and often do not provide much insight into the essentials of the behavior of the system. In particular, in most existing studies it is assumed that there are a fixed number of persistent traffic streams, thus ignoring the dynamics and random nature of the traffic generated by the users in these networks. Moreover, to simulate the system details and to obtain sufficiently accurate performance estimates, time consuming simulation runs (taking many hours) are required even for relatively small ad-hoc network scenarios. Analytical studies on ad-hoc networks only consider single-cell flow transfers and do not consider multi-hop flows. Also, they only consider the overall throughput or the node throughput. These omissions of the existing literature will be illustrated in the literature overview below.

Literature

This overview covers analytical studies on ad-hoc networks. These studies can roughly be divided into three classes: packet-level in a single-cell, flow-level in a single-cell, and packet-level in ad-hoc networks. As will become clear, a flow-level analysis of ad-hoc networks was not covered yet.

The packet-level behavior of IEEE 802.11 WLAN in a single-cell has been investigated extensively. A detailed mathematical performance model of the DCF has been developed and analyzed by Bianchi [2] that was slightly improved by Wu *et al.* [26]. These papers assume a constant number of persistently contending stations and rely on a relatively simple Markov chain analysis, neglecting only minor dependencies among the behavior of different stations, and is used to obtain the saturation throughput of the cell. Comparison with simulation shows that the analytical results are in general remarkably accurate. Bianchi's model is also used to obtain packet contention delays [4, 8]. [4] present a different analytical model, in which the transmission and collision probabilities of Bianchi's model are used; the first two moments of the contention delay are derived.

Flow-level behavior in a single-cell is considered in [10, 18, 25]. The situation with non-persistent traffic sources is considered, i.e., the number of active stations varies dynamically in time according to the initiation and completion of file transfers at random time instants. These papers propose and analyze simplified analytical models yielding approximations for the expected flow (file) transfer time. In particular, in [18] the analysis is based on the modeling assumption that, from the flow-level point of view, the WLAN can be regarded as a Processor Sharing type of queueing system. The analyses in [10, 18, 25] ignore the effects of higher layer protocols, in particular TCP, on the traffic behavior. Several papers consider flow transfer times of TCP-flows over WLAN focusing on the impact of the interaction between TCP's feedback control loop and the DCF MAC protocol on TCP throughputs and

fairness, see e.g. [20, 21, 22].

Analytical models for packet-level performance in ad-hoc networks are presented in e.g. [5, 12, 14, 24]. These papers assume that stations have a limited interference range and after stations are distributed on a plain, the stations experience different channel-conditions as the number of neighbors and the distances to the neighbors varies per station. The performance measures considered in these papers are the overall aggregate throughput and the throughput per node including the impact of hidden stations; unfortunately they do not consider the performance of multi-hop flows.

Contribution

This paper analyzes the transfer time of a multi-hop flow in a bottleneck of an ad-hoc network. Importantly, the study focuses on the flow level: source nodes arrive according to a Poisson process and instantly initate a flow transfer to a destination via the bottleneck node. After a source node has transmitted its entire flow to the bottleneck node, it leaves the system (note that a part of its flow may still be in the buffer of the bottleneck node waiting for service). The model is of a fluid nature: the packet stream is approximated by a continuous stream. Emphasis is on flow-level performance measures, most notably the flow transfer time.

As remarked earlier, the bottleneck node's buffer only drains when there are no flows feeding into the node, due to the way DCF shares the medium capacity. It is exactly this property that makes it possible to recast the buffer dynamics in terms of the classical M/G/1 queueing model. This facilitates an explicit and insightful characterization of the mean overall flow transfer time. Interestingly, the resulting formula depends on the flow-size distribution only through its first two moments; in other words, there is a high degree of insensitivity. We also derive expressions for the mean buffer occupancy and the packet delay at the bottleneck node.

Finally, we discuss the possibility of improving the overall flow transfer time by alternative resource sharing between the stations (which can be obtained by the IEEE 802.11E version).

Organization

This paper is organized as follows. In Section 2 we introduce the ad-hoc network scenario considered in this paper and we develop an analytical performance model for this scenario. Section 3 contains the analysis of this model, whereas the corresponding numerics are presented in Section 4. Section 5 contains some considerations about the performance improvement that can be obtained by differentiation between source and bottleneck nodes (i.e., by using non-equal sharing of the capacity). Finally, concluding remarks and directions for further research are presented in Section 6.

2 System and model description

This section describes the ad-hoc network scenario and introduces the fluid-flow model that is used to analyse the ad-hoc network in the next section.

2.1 The ad-hoc network scenario

We consider a wireless ad-hoc network scenario in which a single node is used by many other nodes as a relay node for data transmission to receivers elsewhere in the network. This may, for example, occur when connectivity between two ad-hoc network clusters is (during some time) provided by only one node acting as a kind of 'bridge', see Figure 1. It is clear that in such a situation this node forms a potential bottleneck. The aim of the present paper is to investigate the performance of this potential bottleneck in terms of, e.g., the flow transfer times and the queueing delays that are encountered by the traffic generated by other network nodes. In particular, we are interested in the impact of the
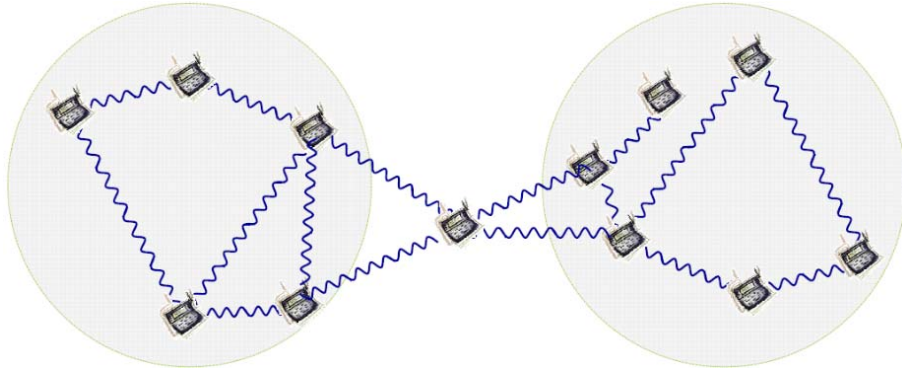
Figure 1: Bottleneck node connecting two ad-hoc network clusters.

interaction between the traffic generating (neighbor) nodes and the bottleneck node due to sharing common radio transmission resources.

For that purpose we focus on a simple, special case of the above scenario yet capturing the essentials of this interaction; a two-hop ad-hoc network consisting of a number of nodes (sources) that may initiate data transmissions, and a single relay node that forwards the traffic generated by the other nodes to a next-hop destination, see Figure 2. The source nodes and the relay node (bottleneck node) are all within each others transmission range. Data transmissions are controlled by the Distributed Coordination Function (DCF) of the IEEE 802.11B MAC protocol (including the RTS/CTS-access mechanism for collision avoidance), see [15]. Data flow transmissions by the sources are initiated at random time instants; the sizes of the flows transmitted by the sources are also random.

The following important observations with respect to the behavior of IEEE 802.11B ad-hoc network nodes sharing common radio transmission resources, obtained in other papers (see e.g. [2, 18]), motivate the fluid-flow modeling approach described below in detail:

- Assuming RTS/CTS-access the aggregate throughput C provided by the system is (almost) independent of the number of active nodes (i.e., nodes contending for the radio resources in order to send their data), cf. the model of Bianchi [2]. The aggregate throughput C can be modelled as a fixed rate which is a percentage of the data transmission rate; the actual percentage depends on the used physical layer technology (in case of IEEE 802.11B physical layer with a data transmission rate of 11 Mbit/s rate the aggregate throughput is about 60% of the data transmission rate).

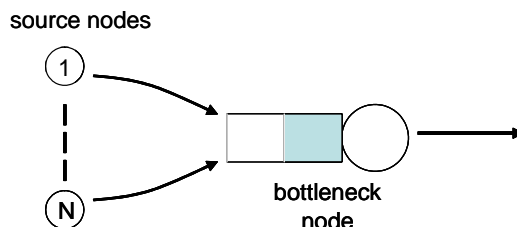- The aggregate system throughput is more or less equally shared by the active network nodes, cf. see [18].

Figure 2: Two-hop network with a single (bottleneck) node used as relay node by many sources.

4

2.2 Model description

The ad-hoc network scenario described above is modeled as a fluid-flow queueing system. We assume that source nodes arrive according a Poisson process with rate λ ('flow arrival rate') and instantly initiate a new flow transfer to a destination node via the bottleneck node. The active source nodes and the bottleneck node equally share the system capacity c; i.e., when n source nodes have a flow transfer in progress, any source transmits its traffic (fluid) into the buffer of the bottleneck node at rate $c/(n+1)$, while a rate $c/(n+1)$ is used by the bottleneck node to 'serve' the buffer (i.e., to forward the traffic stored in its buffer to the next node). Recall that for $n > 1$ the bottleneck node's input rate exceeds its output rate, and the excess traffic accumulates in the node's buffer; the queue only drains when $n = 0$. Once a source has completed its flow transmission (i.e., the source node has transmitted the flow's last traffic particle to the bottleneck node), the source node leaves the system (although the last part of the flow may still be at the buffer of the bottleneck node waiting for service). Flow sizes (in terms of the amount of traffic/fluid) are i.i.d. random variables (denoted by F) with finite mean f and second moment f_2. We define the load of the system by $\rho = \lambda f/c$. All flows have to be served twice, i.e., once by the source node and once by the bottleneck node, resulting in the overall stability condition $\rho < 1/2$.

Our main performance measures of interest are the steady-state buffer occupancy (delay) Q_{BUFFER} (D_{BUFFER}) at the bottleneck node and the overall flow transfer time total time D_{OVERALL}, i.e., the time needed to transmit a complete flow from source to destination. The overall flow transfer time is the sum of two other performance measures: (i) the time (D_{SOURCE}) it takes a source to completely transmit a particular flow to the bottleneck node, and (ii) the delay at the bottleneck node (D^*_{BUFFER}) of the last particle of fluid of the flow (here the asterisk denotes that the performance measure is considered for the last particle of a flow).

The analysis in the next section aims at the derivation of the mean values of these performance measures; insightful, explicit formulas are obtained, which can easily be evaluated in order to generate numerical results. Derivation of higher-order statistics of the performance measures appears to be also possible, but requires substantially more complicated mathematical analysis. In the complementary paper [19] we consider the same model and derive expressions for the Laplace Transforms of D_{BUFFER} and D_{OVERALL}, and characterize the tail probabilities of these performance measures.

3 Analysis

This section analyzes the model presented in Section 2.2. Exact expressions are derived for the performance measures flow transfer time and the buffer occupancy and approximations are presented for the buffer delay of the last particle and the overall flow transfer time.

3.1 Source behavior

The dynamics of the source nodes in our model can be described by a Processor Sharing (PS) queueing model with state dependent service rates $cn/(n+1)$, whenever n flows are in progress. This model is a special case of the so called Generalized Processor Sharing (GPS) model extensively studied by Cohen [7]. In the GPS model, whenever there are i customers present in the system, each customer receives service at a rate $r(i)$, where $r(\cdot)$ is some arbitrary positive function (under some weak assumptions). Cohen derives the following result for the joint stationary probability/density function of the number of customers N in the GPS system and their residual service requirements $T := (T(1), ..., T(N))$, cf. formula (7.19) in [7]:

$$\mathbb{P}(N = n, T = \tau) = \frac{\frac{(\lambda\beta)^n}{n!}\varphi(n)}{\sum_{k=0}^{\infty}\frac{(\lambda\beta)^k}{k!}\varphi(k)}\prod_{i=1}^{n}\frac{1 - B(\tau(i))}{\beta}, \quad n = 0, 1, ..., \quad \tau(i) \geq 0, \tag{1}$$

where $\varphi(0) := 1$ and $\varphi(n) := (\prod_{i=1}^{n} r(i))^{-1}$, for $n = 1, 2, ...$, and where $B(\cdot)$ denotes the customers' service requirement distribution, β is the mean service requirement and λ the customer arrival rate. This general result will be extensively exploited in the remainder of this paper.

In our model $r(i) = \text{C}/(i+1)$ and we obtain the following result for the distribution of the number of active source nodes in our model:

$$\pi_n := \mathbb{P}(N = n) = (n + 1)(1 - \rho)^2 \rho^n.$$

Note, cf. equation (1), that the distribution of N, and hence also its mean, is *insensitive* to the flow-size distribution apart from its mean. The mean number of source nodes that is simultaneously active is given by:

$$\mathbb{E}N = 2\frac{\rho}{1 - \rho}.$$

Flow transfer time D_{SOURCE}

From the above result, using Little's formula, we obtain the mean flow transfer time at the sources:

$$\mathbb{E}D_{\text{SOURCE}} = \frac{\mathbb{E}N}{\lambda} = 2\frac{f/\text{C}}{1 - \rho}. \tag{2}$$

$\mathbb{E}D_{\text{source}}$ is also *insensitive* to the flow-size distribution.

The *conditional* flow transfer time $\mathbb{E}D_{\text{source}}(x)$, the time required by a source node to transmit a flow of give size x, is linear in x, see Cohen [7]:

$$\mathbb{E}D_{\text{SOURCE}}(x) = \frac{x}{\text{C}\rho}\mathbb{E}N = \frac{2x/\text{C}}{1 - \rho}. \tag{3}$$

3.2 Buffer occupancy

The mean buffer occupancy is derived for two different epochs: an arbitrary epoch (which relates to the buffer occupancy at the moment that a new source node arrives and initiates a flow transfer) and the buffer occupancy that is seen by the last particle of a flow that arrives at the bottleneck node.

Buffer occupancy at an arbitrary epoch Q_{BUFFER}

In order to analyze the buffer occupancy we consider the amount of 'work' in the system (i.e., the service time required to serve the buffer occupancy related to the service rate C). The total amount of work in the system W_{TOTAL} consists of two parts: the amount of work W_{SOURCES} present at the source nodes (i.e., fluid that remains to be sent by the actively transmitting source nodes) and the amount of work W_{BUFFER} present in the buffer of the bottleneck node, so

$$\mathbb{E}W_{\text{TOTAL}} = \mathbb{E}W_{\text{SOURCES}} + \mathbb{E}W_{\text{BUFFER}}.$$

$\mathbb{E}W_{\text{TOTAL}}$ is obtained by considering the whole system (i.e., sources plus bottleneck node with buffer) as a single service center. Flows arrive at the service center according to a Poisson process with rate λ and have to be served essentially *twice* (i.e., one service corresponds to a source node transmitting its flow to the bottleneck node and the second service corresponds to the forwarding of the flow by the bottleneck node). Using that the overall system is 'work conserving' (i.e., the full capacity C is used whenever work is present in the system), it follows that the (steady-state) distribution of the total amount of work W_{TOTAL} in the overall system equals the distribution of the amount of work $W_{\text{M/G/1}}$ in a 'corresponding' M/G/1 queueing system with flow arrival rate λ, flow size $2F$ and service rate C. Hence, from the Pollaczek-Khintchine formula (see e.g. [17]) we get

$$\mathbb{E}W_{\text{TOTAL}} = \mathbb{E}W_{\text{M/G/1}} = \frac{2\lambda f_2}{1 - 2\rho}\frac{1}{\text{C}^2}.$$

The amount of work present at the sources W_{SOURCES} consists of the residuals of the flows that are still with the active sources, i.e., not yet transmitted to the bottleneck node.

Formula (1) shows that the residual flow sizes at the sources are all identically distributed, mutually independent and independent of the total number of active sources. In particular, this distribution of the residual flow size (with mean $f_2/(2f)$) is the so-called excess distribution of the initial flow sizes well-known from renewal theory, see e.g. Ch. 1 of [23]. From these observations it follows that the mean total amount of fluid present at the sources is given by $\mathbb{E}N \cdot f_2/(2f)$. Hence, taking into account that the residuals of the flows at the sources have to be 'served' twice (i.e., to be transmitted by the sources and by the bottleneck node), we have:

$$\mathbb{E}W_{\mathrm{SOURCES}} = 2\mathbb{E}N \frac{f_2}{2f\mathrm{C}}.$$

Hence, the mean amount of work in the buffer at the bottleneck node is given by:

$$\mathbb{E}W_{\mathrm{BUFFER}} = \mathbb{E}W_{\mathrm{TOTAL}} - \mathbb{E}W_{\mathrm{SOURCES}} = \frac{2\lambda f_2}{1-2\rho}\frac{1}{\mathrm{C}^2} - 4\frac{\rho}{1-\rho}\frac{f_2}{2f}\frac{1}{\mathrm{C}} = \frac{2\rho^2 f_2}{f\mathrm{C}}\frac{1}{(1-2\rho)(1-\rho)}. \tag{4}$$

Finally, the expected buffer occupancy in terms of fluid at an arbitrary epoch is obtained from:

$$\mathbb{E}Q_{\mathrm{BUFFER}} = \mathrm{C} \cdot \mathbb{E}W_{\mathrm{BUFFER}}.$$

Buffer occupancy seen by the last particle of a flow Q^*_{BUFFER}

The buffer occupancy seen by the last particle of a particular flow is the buffer occupancy at arrival of the flow plus the buffer increase during its flow transfer time D_{SOURCE} (note that the buffer occupancy cannot decrease during the presence of a flow in our specific sharing policy). By the PASTA-property, the mean buffer occupancy at the arrival of the flow is equal to (the time-average mean) $\mathbb{E}Q_{\mathrm{BUFFER}}$.

The expected growth of the buffer $\mathbb{E}\Delta Q_{\mathrm{BUFFER}}$ during the flow transfer time can be derived easily as the sending rates of the source nodes and the bottleneck node are coupled. In particular, if a source requires time $D_{\mathrm{SOURCE}}(x)$ to transmit a flow of size x to the bottleneck node, the bottleneck node will have served also an amount x. This means that the aggregate input of all sources is $D_{\mathrm{SOURCE}}(x)\mathrm{C} - x$, resulting in a buffer growth of $D_{\mathrm{SOURCE}}(x)\mathrm{C} - 2x$, hence

$$\mathbb{E}\Delta Q_{\mathrm{BUFFER}}(x) = \mathbb{E}D_{\mathrm{SOURCE}}(x)\mathrm{C} - 2x \tag{5}$$

which is linear in x. Hence,

$$\mathbb{E}\Delta Q_{\mathrm{BUFFER}} = \mathbb{E}D_{\mathrm{SOURCE}}\mathrm{C} - 2f = \frac{2f\rho}{1-\rho},$$

and

$$\mathbb{E}Q^*_{\mathrm{BUFFER}} = \mathbb{E}Q_{\mathrm{BUFFER}} + \mathbb{E}\Delta Q_{\mathrm{BUFFER}} = \frac{2\rho^2 f_2/f}{(1-2\rho)(1-\rho)} + \frac{2f\rho}{1-\rho}. \tag{6}$$

Finally, the conditional buffer occupancy, also by the PASTA-property, is

$$\mathbb{E}Q^*_{\mathrm{BUFFER}}(x) = \mathbb{E}Q_{\mathrm{BUFFER}} + \mathbb{E}\Delta Q_{\mathrm{BUFFER}}(x). \tag{7}$$

3.3 Buffer delay

The mean buffer delay is derived for two different fluid particles: an arbitrary particle and the last particle of a flow that arrives at the bottleneck node.

Buffer delay of an arbitrary particle D_{BUFFER}

The mean delay of an arbitrary fluid particle in the buffer of the bottleneck node can be directly obtained from Little's formula:

$$\mathbb{E}D_{\text{BUFFER}} = \frac{\mathbb{E}Q_{\text{BUFFER}}}{\lambda f} = \frac{2\rho f_2/(f\text{C})}{(1 - 2\rho)(1 - \rho)}. \tag{8}$$

Buffer delay of the last particle of a flow D^*_{BUFFER}

The buffer delay of the last particle of a flow is equal to the time required by the bottleneck node to serve the buffer content Q^*_{BUFFER} present upon the last particle's arrival. For the analysis of this quantity it is useful to note that, from the point of view of the bottleneck node, the system behaves as a standard Processor Sharing (PS) queueing model; the bottleneck node receives capacity $\text{C}/(n+1)$ if there are n other active sources. So the buffer delay of the last particle corresponds to the time required to serve a given amount of fluid (i.e., the buffer content upon the last particle's arrival) in a PS queueing model.

For the moment we assume exponentially distributed flow sizes; the remark at the end of this subsection generalizes the result to non-exponentially distributed flow sizes. Coffman, Muntz, and Trotter [6] studied the so-called *response time* for jobs in an M/M/1-PS queue. They obtained an explicit expression ((33) of [6]) for the mean response time $\mathbb{E}X_n(\tau)$ of a tagged job requiring an amount τ of service, when there are n other active source nodes upon arrival of the tagged job. In terms of our system this expression reads:

$$\mathbb{E}X_n(\tau) = \tau + \frac{\rho\tau}{1 - \rho} + (n(1 - \rho) - \rho)\,(f/\text{C})\frac{1 - \exp(-(1 - \rho)\tau\text{C}/f)}{(1 - \rho)^2}. \tag{9}$$

Using this result we can derive an expression for the expected buffer delay of the last particle of a flow in the case that flow sizes are exponentially distributed. Let $w_n(\tau)$ denote the probability density function of the amount of work at the buffer at a departure epoch of a source node leaving n active source nodes behind. Taking into consideration that the distribution of the number of source nodes left behind at a source node departure epoch is distributed according to the stationary distribution ('departure theorem' of product form networks, see also [7]), we obtain

$$\mathbb{E}D^*_{\text{BUFFER}} = \sum_{n=0}^{\infty} \pi_n \int_0^{\infty} \mathbb{E}X_n(\tau)w_n(\tau)\mathrm{d}\tau.$$

Unfortunately, an exact expression for $w_n(\tau)$ is not available. As an approximation we assume

$$w_n(\tau) \approx w(\tau), \tag{10}$$

where $w(\tau)$ denotes the unconditional probability density function of the amount of work at the buffer at a source node departure epoch. Approximation (10) is expected to work well as the dependency between the number of active source n and the buffer occupancy is not expected to be strong; the buffer occupancy cannot decrease if source nodes are active which means that leaving behind a higher number of active source nodes does not necessarily imply that the buffer occupancy is higher than when a lower number of active source nodes are left behind. Hence,

$$\mathbb{E}D^*_{\text{BUFFER}} \approx \sum_{n=0}^{\infty} \pi_n \int_0^{\infty} \mathbb{E}X_n(\tau)w(\tau)\mathrm{d}\tau.$$

Unfortunately, an expression for $w(\tau)$ is also not available. However, an insightful observation is that if $\mathbb{E}X_n(\tau)$ would be linear in τ (possibly with a non-zero intercept), then $\int \mathbb{E}X_n(\tau)w_n(\tau)\mathrm{d}\tau =$

$\mathbb{E}X_n(\mathbb{E}W^*_{\text{BUFFER}})$ where $W^*_{\text{BUFFER}} = Q^*_{\text{BUFFER}}/\text{C}$. From expression (9) it is seen that only the last part of $\mathbb{E}X_n(\tau)$ does not depend on τ linearly, but for large τ the last part is (almost) constant. This observation justifies our assumption that (9) *is* linear in τ, which entails:

$$\int_0^\infty \mathbb{E}X_n(\tau)w(\tau)\mathrm{d}\tau \approx \mathbb{E}X_n(\mathbb{E}W^*_{\text{BUFFER}}). \tag{11}$$

Combining approximations (10) and (11) yields the following approximate expression for the mean delay of the last particle of a flow in the buffer of the bottleneck node:

$$\mathbb{E}D^*_{\text{BUFFER}} \approx \sum_{n=0}^\infty \pi_n \mathbb{E}X_n(\mathbb{E}W^*_{\text{BUFFER}}). \tag{12}$$

Remark on generally distributed flow sizes. Above we considered exponentially distributed flow sizes (note that the assumption of exponentially distributed flow sizes is only used in expression (9)). Now we consider the buffer delay of generally distributed flow sizes. Asare and Foster [1] derived an expression for the expected conditional response time $\mathbb{E}X_n(\tau)$ of an M/G/1-PS queue, which can be considered as the analogue of (9) obtained in [6] for the special case of exponentially distributed flow sizes. The result of Asare and Foster for general flow sizes, however, requires the distribution of the workload in the system, and an explicit formula is not available.

As an approximation we propose to use (6) also for the expected conditional buffer delay for non-exponentially distributed flow sizes. Then, the approximation for the buffer delay of the last particle is approximated by using (6) in (9), which gives approximation (12).

Now, we will argue that using (9) for the buffer delay is expected to work well also for generally distributed flow sizes. An important observation is that rate at which the buffer is drained only depends on the number of active source nodes; recall from Section 3.1 that the steady-state distribution of the number of active source nodes is insensitive to the flow-size distribution apart from its mean. In particular if the buffer occupancy is large, it is expected that the behavior experienced by the last particle will resemble the steady-state behavior of the source nodes. Also for a very small buffer occupancy (smaller than the residuals at the source nodes) the influence of the flow-size distribution is small, as the number of sources will most likely remain constant. For an intermediate buffer occupancy, the influence of the flow-size distribution will be larger; the source nodes do not immediately behave like the steady-state behavior as the expected sizes of the residuals of the flows at the source nodes strongly depend on the flow-size distribution.

The flow-size distribution affects approximation (12) via both (6) and (9). However, as was argued in the previous paragraph, the impact on (9) is only minor. On the contrary, the flow-size distribution has a strong impact on (6); this is captured through the second moment of the flow size.

3.4 Overall flow transfer time D_{OVERALL}

The overall flow transfer time D_{OVERALL} of a flow is the sum of its flow transfer time D_{SOURCE} and the buffer delay of its last particle D^*_{BUFFER}. Hence

$$\mathbb{E}D_{\text{OVERALL}} = \mathbb{E}D_{\text{SOURCE}} + \mathbb{E}D^*_{\text{BUFFER}},$$

where $\mathbb{E}D_{\text{SOURCE}}$ is given by expression (2) and $\mathbb{E}D^*_{\text{BUFFER}}$ is appromximated by expression (12). Remark that $\mathbb{E}D_{\text{OVERALL}}$ is linear in f.

Remark on the overall flow transfer time conditional on the flow size. In the previous section the expectation of the overall flow transfer time was derived for an arbitrary flow. These results can easily be extended to the situation where a flow has a given size x:

$$\mathbb{E}D_{\text{OVERALL}}(x) = \mathbb{E}D_{\text{SOURCE}}(x) + \mathbb{E}D^*_{\text{BUFFER}}(x). \tag{13}$$

The conditional flow transfer time $\mathbb{E}D_{\text{SOURCE}}(x)$ is already presented by (3). The expected conditional buffer delay $\mathbb{E}D^*_{\text{BUFFER}}(x)$ can be obtained also by approximation (12) of the previous section, but now the buffer occupancy is given by $\mathbb{E}W^*_{\text{BUFFER}}(x)$, the expected amount of work in the buffer upon arrival of the last particle of the particular flow with given size x. Note that $\mathbb{E}W^*_{\text{BUFFER}}(x) = \mathbb{E}Q^*_{\text{BUFFER}}(x)/\text{C}$ and $\mathbb{E}Q^*_{\text{BUFFER}}(x)$ is given in (7). Then the conditional buffer delay is obtained by

$$\mathbb{E}D^*_{\text{BUFFER}}(x) \approx \sum_{n=0}^{\infty} \pi_n \mathbb{E}X_n(\mathbb{E}W^*_{\text{BUFFER}}(x)),$$

which completes (13).

4 Numerical results

This section presents numerical results of the exact and approximate results of the previous section. The analytical approximations are validated by simulations of the model as it is described in Section 2.2. The simulation tool was built in Delphi and sufficient replications have been simulated in order to obtain small confidence intervals. The simulations are only performed using $\text{C} = 1$ and $f = 1$ as the performance measures are linear in f, so that the performance measures for different flow sizes f can be directly obtained from $f = 1$. For flow-size distributions we used Deterministic, Erlang-4, Exponential and Hyper-Exponential (with balanced means (see e.g. [23]) and $\text{COV}= 4$) distributions and the graphs refer to these distributions by their COVs which are $0, 0.25, 1,$ and 4 respectively. The load ρ is varied between 0.05 and 0.45 (by varying the arrival rate λ) to observe the system under different load settings.

Figure 3 presents the mean flow transfer time $\mathbb{E}D_{\text{SOURCE}}$ as a function of the offered traffic load ρ; as noticed earlier $\mathbb{E}D_{\text{SOURCE}}$ is insensitive to the flow-size distribution apart from its mean. Observe that the variation in the flow transfer time is small; the stability condition of the source nodes is $\rho < 1$ (while the stability condition for the *total* system is $\rho < 0.5$) and, as a consequence, the source nodes experience a lightly loaded system resulting in a small number of active source nodes and short flow transfer times.

The graphs of Figure 4 present the mean buffer occupancy at the bottleneck node for various flow-size distributions. The left graph shows the results for arbitrary fluid particles, whereas the right graph relates to the last particle of a flow. In both graphs it can be seen that the flow-size distribution has a strong impact on the buffer occupancy. Further, it can be observed that the buffer occupancy seen by the last particle is only slightly higher than the buffer occupancy upon flow arrival; Figure 3
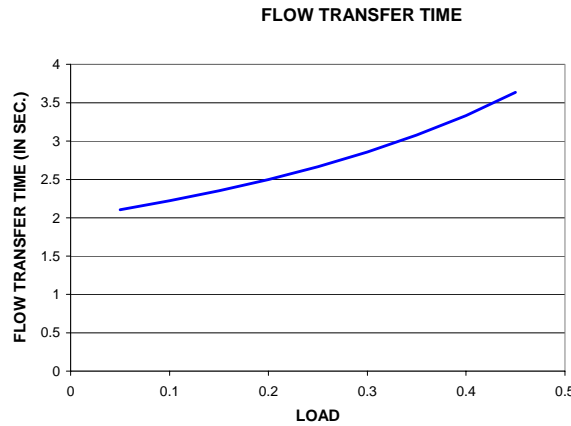

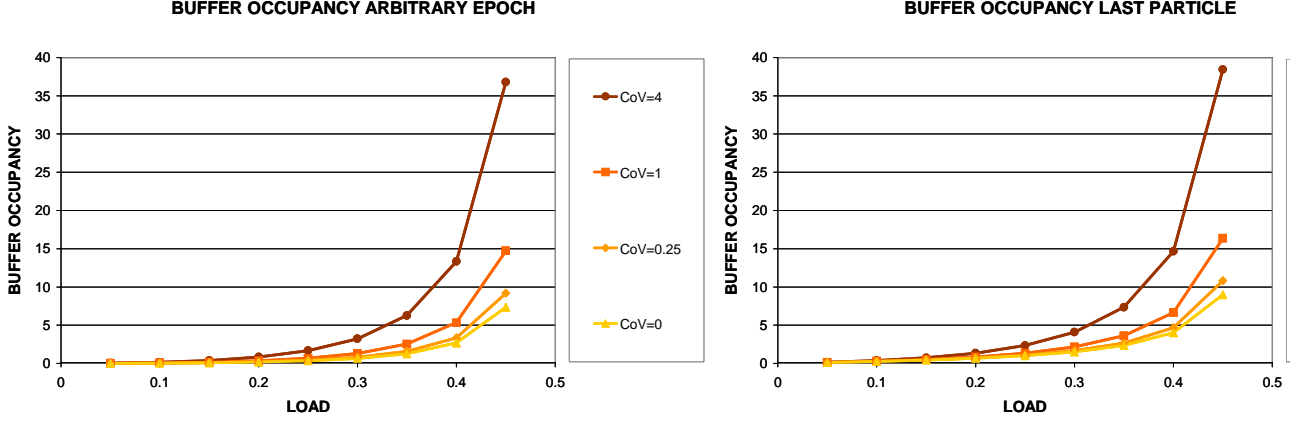Figure 3: Mean flow transfer time ($\mathbb{E}D_{\text{SOURCE}}$).

Figure 4: Mean buffer occupancy. Left: mean buffer occupancy in steady-state ($\mathbb{E}Q_{\text{BUFFER}}$). Right: mean buffer occupancy seen by the last particle ($\mathbb{E}Q^*_{\text{BUFFER}}$).

already illustrated that the flow transfer time is relatively short and the number of active sources is also low, resulting in a minor increase of the buffer during the flow transfer time.

Figure 5 presents the results for the mean buffer delay for both an arbitrary particle (left) and the last particle of a flow (right). The buffer delay of the last particle in the right graph is an approximation, but it captures the behavior very well, not only for exponential flow sizes (for which the approximation is originally designed), but also for other flow-size distributions (cf. remark in Section 3.2). An interesting observation is that buffer delay corresponding to an arbitrary particle is higher than the buffer delay of the last particle. This effect can be explained by the waiting time paradox (see e.g. [9]) as with high probability an arbitrary particle belongs to a large flow. Belonging to a large flow has two negative effects: first, before the particle enters the buffer the source node has been transmitting for a long period, so the buffer occupancy will be high, and second, when the particle has entered the buffer, the source node will remain active (in the system) for a long period resulting in a low rate for the bottleneck node.

Figure 6 presents approximation and simulation results for the mean overall flow transfer time. It can be seen that the mean overall transfer time is largely determined by the buffer delay. There is also a small error which is obviously due to the error in the approximation of the buffer delay of the last particle. Figure 7 presents the mean overall flow transfer time conditional on the flow size for a
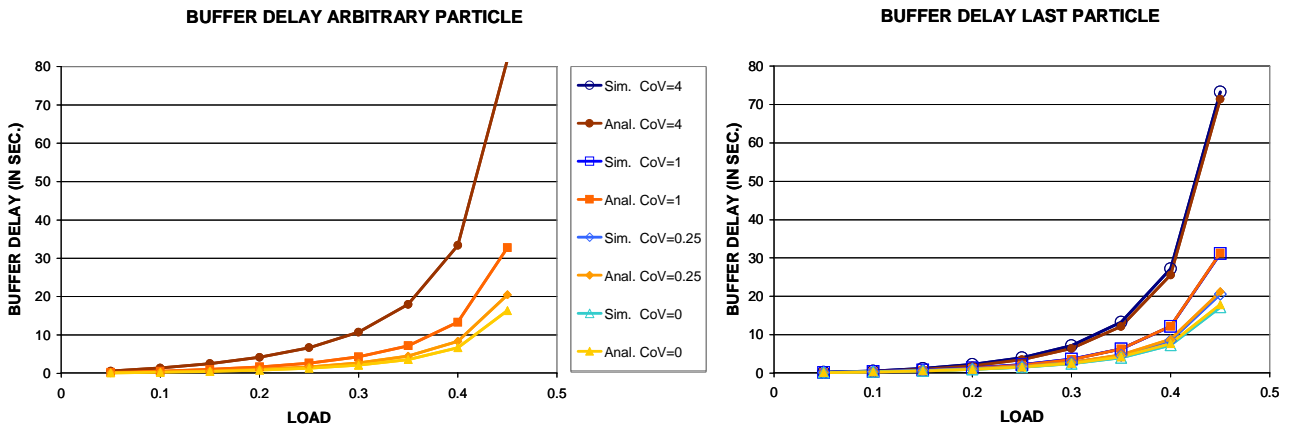

Figure 5: Mean buffer delay. Left: arbitrary particle ($\mathbb{E}D_{\text{BUFFER}}$). Right: last particle ($\mathbb{E}D^*_{\text{BUFFER}}$).
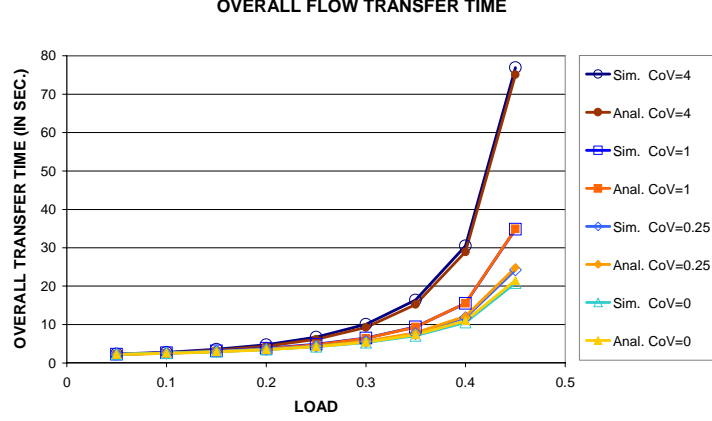
Figure 6: Mean overall flow transfer time ($\mathbb{E}D_{\text{OVERALL}}$).

load of 0.35, for different flow-size distributions (deterministic flow size (CoV=0) is only a single mark in the graph at $f = 1$). For Erlang-4 and exponentially distributed flow sizes the approximation is accurate as the simulations and analytical results coincide; for hyper-exponential it is seen that analysis underestimates the mean overall flow transfer time, but it is interesting to see that the absolute error remains constant for increasing flow sizes.

5   Model extension: impact of resource allocation strategy

In our IEEE 802.11B ad-hoc network (where all nodes can hear each other) the system capacity is equally shared amongst the active source nodes and the bottleneck node. However, the bottleneck node has to serve the traffic of all source nodes and if more than one source nodes are active, the buffer occupancy at the bottleneck node increases; the buffer content only decreases if there are no source nodes active. In this system the sources nodes can transmit their flows to the bottleneck node fast (cf. Figure 3), however, this is at the cost of large buffer delay (cf. right graph of Figure 5).

It is interesting to consider the possibility of assigning a large share of the system capacity to the bottleneck node (meaning a smaller share for each of the active source nodes); this will reduce the buffer delay of the last particle at the cost of larger flow transfer times. Obviously there is a trade-off between the flow transfer time and the buffer delay, where the main objective is to reduce the overall
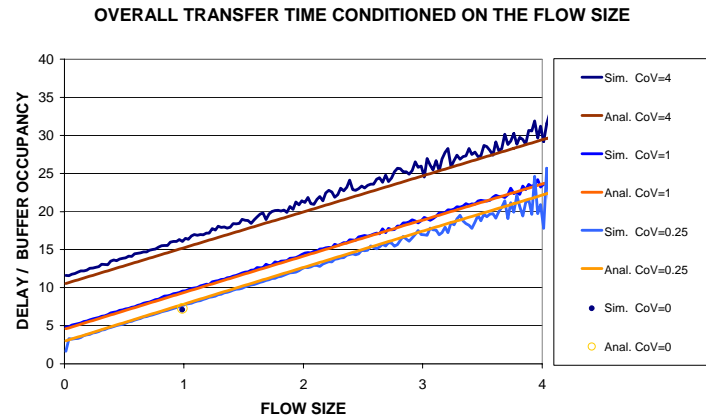
Figure 7: Mean overall flow transfer time conditioned on the flow size ($\mathbb{E}D_{\text{OVERALL}}(x)$) for $\rho = 0.35$.
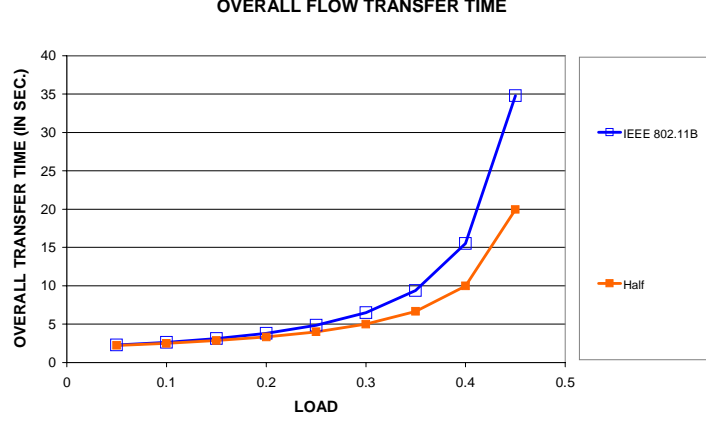
Figure 8: Mean overall transfer time ($\mathbb{E}D_{\text{OVERALL}}$) with a prioritized bottleneck node.

flow transfer time. With the recently standardized IEEE 802.11E EDCA, a QoS-aware version of the DCF, it is possible to 'assign' more weight to the bottleneck node by appropriately setting EDCA-parameters, e.g. CW_{min}, AIFS or the TXOP-limit.

In order to get some feeling for the impact of such alternative sharing policies on the network performance, we will compare the situation of equal sharing of the transmission resources among all individual active nodes (as described and analyzed above) with the (extreme) situation in which the bottleneck node is strongly favored and receives exactly as many transmission capacity as *all* active source nodes together. We will denote this latter resource allocation by 'half'.

It is clear that in the 'half' variant of our model, there is no queueing at the bottleneck node and the (overall) flow transfer time is completely determined by the behavior of the source nodes. The source nodes behave as a PROCESSOR SHARING model with service capacity $\text{C}/2$, flow arrival rate λ and mean flow size f. Hence, the mean flow transfer delay is given by (independent of the flow-size distribution):

$$\mathbb{E}D_{\text{OVERALL}}^{(\text{half})} = \frac{2f/\text{C}}{1 - 2\rho}.$$

Under the original 'IEEE 802.11B' resource allocation we have, see above,

$$\mathbb{E}D_{\text{OVERALL}} = \mathbb{E}D_{\text{SOURCE}} + \mathbb{E}D_{\text{BUFFER}}^* = 2\frac{f/\text{C}}{1 - \rho} + \frac{2\rho f_2/(f\text{C})}{(1 - 2\rho)(1 - \rho)}.$$

These results illustrate the considerable reduction of the overall flow transfer delay that can be achieved by modification of the resource allocation strategy realized by the MAC-protocol. It is expected that resource allocation strategies 'in between' the two strategies considered above will yield mean overall flow transfer times which are in between $\mathbb{E}D_{\text{OVERALL}}$ and $\mathbb{E}D_{\text{OVERALL}}^{(\text{half})}$. Figure 8 presents a comparison of $\mathbb{E}D_{\text{OVERALL}}$ (bottleneck node) and $\mathbb{E}D_{\text{OVERALL}}^{(\text{half})}$ (PROCESSOR SHARING) and it is seen that the scheme $\mathbb{E}D_{\text{OVERALL}}^{(\text{half})}$ performs significantly better for high loads.

Remark. It is expected that, if the resource allocation at the MAC-layer is according to 'IEEE 802.11B' and, in addition, the traffic rates are end-to-end controlled by TCP, the actually resulting allocation of radio transmission resources to these TCP flows would be according to the 'half' variant described above. The TCP-control loop will achieve that the source nodes can only transmit at a rate at which the bottleneck node can serve the traffic, see e.g. [21].

6 Concluding remarks and directions for further research

We have developed a mathematical model describing the behavior of a bottleneck node in a wireless ad-hoc network. This model provides useful insights into the (overall) flow transfer time and the buffer delay occurring at the bottleneck node, in particular how it depends on the various system and traffic parameters. The modeling approach presented in this paper may enable refinements and extensions to more general network settings; this is a subject for further research.

Topics for further research also include:

- extending the model to multi-hop flows where not all the nodes are within each others sensing range. This allows for spatial reuse of the medium capacity, but it also introduces hidden and exposed nodes. Both effects strongly impact the medium capacity and the resource allocation to the nodes; multi-hop flows also may have to pass multiple bottlenecks. It is interesting to investigate the impact of these effects on the queueing and flow transfer times.

- alternative service disciplines at the bottleneck node. In the above analysis of the overall flow transfer time, it is assumed that the packet scheduling at the bottleneck node is FIRST COME FIRST SERVE. Alternative service disciplines, e.g. round robin (with respect to packets of different flows), may yield considerably smaller mean overall flow transfer times.

- investigating the influence of higher-layer protocols, such as TCP, on the flow transfer time (cf. the discussion in Section 5).

- investigating the possible performance gain that can be obtained by IEEE 802.11E EDCA (cf. the discussion in Section 5).

Acknowledgements

References

[1] B.K. ASARE and F.G. FOSTER. Conditional response times in the M/G/1 Processor-sharing system. *Journal of Applied Probability*, 20: 910–915, 1983.

[2] G. BIANCHI. Performance analysis of the IEEE 802.11 Distributed Coordination Function. *IEEE Journal on Selected Areas in Communications*, 18: 535–547, 2000.

[3] T. BHEEMARJUNA REDDY, I. KARTHIGEYAN, B.S. MANOJ, and C. SIVA RAM MURTHY. Quality of service provisioning in ad-hoc wireless networks: a survey of issues and solutions. *Ad Hoc Networks*, 4: 83–124, 2006.

[4] M.M. CARVALHO and J.J. GARCIA-LUNA-ACEVES. Delay analysis of IEEE 802.11 single-hop networks. *Proceedings of the 11th IEEE International Conference on Network Protocols*, Atlanta, USA, 2003.

[5] M.M. CARVALHO and J.J. GARCIA-LUNA-ACEVES. A scalable model for channel access protocols in mult-hop ad hoc networks. *Proceedings of MobiHoc'04*, Philadelphia, USA, 2004.

[6] E.G. COFFMAN JR., R.R. MUNTZ, and H. TROTTER. Waiting time distributions for Processor-sharing systems. *Journal of the Association for Computing Machinery*, 17: 123–130, 1970.

[7] J.W. COHEN. The multiple phase service network with GENERALIZED PROCESSOR SHARING. *Acta informatica*, 12: 245–284, 1979.

[8] P.E. ENGELSTAD and O.N. ØSTERBØ. Non-saturation and saturation analysis of IEEE 802.11e EDCA with starvation prediction. *Proceedings of MSWiM '05,* Montreal, Canada, 2005.

[9] W. FELLER. An introduction to probability theory and its applications, vol. II, 2nd edition. John Wiley & Sons, 1971.

[10] C.H. FOH and M. ZUKERMAN. Performance analysis of the IEEE 802.11 MAC protocol. *Proceedings of European Wireless '02*, Florence, Italy, 2002.

[11] Z. FU, P. ZERFOS, H. LUO, S. LU, L. ZHANG, and M. GERLA. The impact of multihop wireless channel on TCP throughput and loss. *Proceedings of INFOCOM '03*, San Francisco, USA, 2003.

[12] J. HE, D. KALESHI, A. MUNRO, Y. WANG, A. DOUFEXI, J. MCGEEHAN, and Z. FAN. Performance investigation of IEEE 802.11 MAC multihop Wireless Networks. *Proceedings of MSWiM '05,* Montreal, Canada, 2005.

[13] J. HE and H.K. PUNG. Fairness of medium access control for multi-hop ad hoc networks. Computer Networks, 48: 867–890, 2005.

[14] J. HE and H.K. PUNG. Performance modelling and evaluation of IEEE 802.11 DISTRIBUTED COORDINATION FUNCTION in multihop wireless networks. *To appear in Computer Communications.*

[15] IEEE P802.11B/D7.0, Supplement: higher speed physical layer extension in the 2.4 GHz band, 1999.

[16] IEEE P802.11E/D11.0, Draft supplement to standard for telecommunications and information exchange between systems - LAN/MAN specific requirements - part 11: wireless MEDIUM ACCESS CONTROL (MAC) and PHYSICAL layer (PHY) specifications: enhancements for QUALITY OF SERVICE (QOS), October 2004.

[17] L. KLEINROCK. Queueing Systems, Vol. 1. John Wiley & Sons, 1976.

[18] R. LITJENS, F. ROIJERS, J.L. VAN DEN BERG, R.J. BOUCHERIE, and M.J. FLEUREN. Analysis of flow transfer times in IEEE 802.11 wireless LANs. *Annals of Telecommunications*, 59: 1407–1432, 2004.

[19] M.R.H. MANDJES and F. ROIJERS. A fluid system with coupled input and output and its application to bottlenecks in ad hoc networks. In preparation. Available as CWI-report at http://www.cwi.nl/reports.

[20] D. MIORANDI, A.A. KHERANI AND E. ALTMAN. A queueing model for HTTP traffic over IEEE 802.11 WLANs. *Proceedings of ITC Specialist Seminar on Performance Evaluation of Wireless and Mobile Systems*, Antwerp, Belgium, 2004.

[21] F. ROIJERS, J.L. VAN DEN BERG, and X. FAN. Analytical modelling of TCP file transfer times over IEEE 802.11 wireless LANs. *Proceedings of ITC-19*, Beijing, China, 2005.

[22] T. SAKURAI and S. HANLY. Modelling TCP flows over an 802.11 wireless LAN. *Proceedings of European Wireless Conference*, Nicosia, Cyprus, 2005.

[23] H.C. TIJMS. Stochastic models: an algorithmic approach. John Wiley & Sons, 1994.

[24] Y. Wang and J.J. Garcia-Luna-Aceves. Performance of collision avoidance protocols in single-channel ad hoc networks. *Proceedings of the 10th IEEE International Conference on Network Protocols*, Paris, France, 2002.

[25] E.M.M. Winands, T.J.J. Denteneer, J.A.C. Resing, and R. Rietman. A finite-source feedback queueing network as a model of the ieee 802.11 distributed coordination function. *Proceedings of European Wireless'04*, Barcalona, Spain , 2004.

[26] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma. Performance of reliable transport protocol over ieee 802.11 wireless lan: analysis and enhancement. *Proceedings of IEEE INFOCOM '02*, New York, USA, 2002.