

RESEARCH

Open Access



Buffer-aided relaying improves both throughput and end-to-end delay

Javad Hajipour^{1*}, Rukhsana Ruby¹, Amr Mohamed² and Victor C. M. Leung¹

Abstract

Buffer-aided relaying has recently attracted a lot of attention due to the improvement in the system throughput. However, a side effect usually deemed is that buffering at relay nodes results in the increase of packet delays. In this paper, we study the effect of buffering at relays on the end-to-end delay of users' data, from the time they arrive at the source until delivery to the destination. We use simple discussions to provide an insight on the overall waiting time of the packets in the system, taking into account the queue dynamics both in the source and relay. We analyze the end-to-end delay in the relay networks with Bernoulli data arrivals and channel conditions and prove that the data packets experience lower average end-to-end delay in the buffer-aided relaying system compared with the conventional one. Moreover, using intuitive generalizations, we conclude that the use of buffers at relays improves not only throughput but ironically the average end-to-end packet delay. Through extensive simulations, we validate our analytical results for the system when the data arrival and channel condition processes follow Bernoulli distribution. Furthermore, via the simulations under the settings of practical systems, we confirm our intuition for the general scenarios.

Keywords: Wireless relay networks, Buffering capability, Throughput, Delay

1 Introduction

Wireless relays have received significant attention in the past decade because of their capability to enhance the capacity and coverage of wireless networks. The authors in [1] have investigated the bounds on the ergodic and outage capacities for wireless relay channels in Rayleigh fading environment. Similarly, Azarian et al. [2] have studied amplify-and-forward (AF) and decode-and-forward (DF) relay channels and proposed variants of these protocols for reaching the bounds on achievable diversity-multiplexing trade-offs. Employing wireless relays in contemporary cellular networks is also considered as a promising solution for meeting the growing demands of users in these systems, due to the cost-effective and fast deployment possibility of relay stations [3]. Therefore, there has been extensive research in this area to identify the challenges and address them accordingly [4–7]. In particular, Ng et al. [4] studied resource allocation in an orthogonal frequency division multiple access (OFDMA)-based system with AF relays and proposed

optimal subchannel and power allocation for maximizing the system goodput. In [5], the authors investigated joint relay selection and resource allocation taking link asymmetry and imperfect channel state information (CSI) into account. Zhang et al. [6, 7] studied resource allocation to provide quality of service (QoS) for the users with minimum rate or maximum packet delay requirements.

Usually in the literature in this area, it is assumed that relaying procedure is performed in two consecutive sub-slots of a transmission interval; i.e., in the first subslot, the base station (BS) transmits to the relay and in the second one, the relay forwards the received data to the mobile terminal. We refer to this method as “conventional relaying” in which the end-to-end transmission rate in each transmission interval is limited by the poorest link quality. Recently, it has been shown that using buffer in the relay node can improve the system throughput [8–11]. This is achieved due to the fact that the buffering capability allows the relay to store packets when its channel condition is bad and transmit when it is good. Motivated by this, several other works have studied buffer-aided relaying scheme in different areas [12–16]. While Krikidis et al. [12] have studied adaptive relay link selection in a single-source multi-relay system, the authors of [13] have

*Correspondence: hajipour@ece.ubc.ca

¹ECE Department, The University of British Columbia, Vancouver, Canada
Full list of author information is available at the end of the article

investigated that for a multi-source multi-relay scenario. On the other hand, Ahmed et al. [14] have discussed the advantages of buffer-aided relaying for the operation of nodes with energy harvesting capability. Moreover, Liu et al. and Darabi et al. [15, 16] have confirmed the advantage of using buffer-aided relays in two-way relaying and cognitive radio networks, respectively.

Any improvement in a system usually comes at a cost. In the case of buffer-aided relaying, the cost is usually deemed to be the increase in packet delays due to queueing in the relay. Consequently, the works in [8–11] have tried to investigate and discuss the trade-off between throughput and delay. This is however based on the assumption of infinitely backlogged buffers in the source (i.e., BS) and considering the queueing delay only at the relay buffer without taking into account the queue dynamics at the BS.

In this paper, we aim at filling the abovementioned gap by taking into account the queue dynamics both in the source node and the relay node. Whereas the existing literature [8–11] mostly considers packet delay as the time delay between packet transmission (departure) at the source node and reception (arrival) at the destination node, in this paper, we study end-to-end packet delay, i.e., the delay that data packets experience since their arrival at the source node until reception at the destination node. The difference between the end-to-end delay considered in this paper and the delay investigated in the abovementioned works is that the end-to-end delay includes both the queueing delay that data packets experience in the source node and the time interval between their transmission at the source and reception at the destination. Noting that the delay perceived by the end user is affected by the queueing at both the BS and the relay, in this paper, we investigate the effect of buffer-aided relaying on the end-to-end packet delay. For this, we first provide simple reasoning and discuss the cause of queue formation in a simple queueing system. Based on that, we provide an insight on the delay performance in the buffer-aided and conventional relaying systems. Then, we study the delay performance when data arrival and channel condition processes of the system follow Bernoulli distribution and derive closed form expressions for the average end-to-end packet delay. Using these, we prove that the buffer-aided relaying system incurs lower average end-to-end packet

delay compared with the conventional one. Finally, we discuss general scenarios and based on intuitive discussions, we conclude that buffering at relays improves the system throughput as well as the average end-to-end packet delay. Using extensive simulations, we verify our analysis and demonstrate the validity of the presented perspective. To the best of our knowledge, this is the first work that discusses the effect of buffering at relays on the overall waiting time in a relay-based network and provides the above conclusion and insight. We note that the discussions in this paper assume infinite buffer capacities in the BS and relay. However, the insights provided can be used in future works to study the scenarios with limited buffer capacities, where the buffer overflow events can also affect the end-to-end packet delays due to the need for repeated transmissions. In such scenarios, for reducing buffer overflow incidents, a buffer-aware source rate control mechanism can be exploited to adjust the traffic load in the network [17]. Also, buffer-aware resource allocation methods similar to [18] can be employed to efficiently serve the system queues. Then, considering the discussions presented in this paper as well as the probability of repeated transmissions, the end-to-end packet delay can be investigated for conventional and buffer-aided relaying systems with finite buffer capacities in the BS and relay.

The rest of this paper is organized as follows. Section 2 provides a background on the queueing delay based on a simple queueing system. In Section 3, through the mathematical analysis and generalized intuitions, we study the end-to-end delay performance of conventional and buffer-aided relaying systems. We validate our analytical study and provide the results for general scenarios through the extensive simulations in Section 4, and finally Section 5 concludes the paper.

2 Background

In this section, we study a simple queueing system and discuss the cause of packet delays to provide a basis for the next section, which studies the end-to-end packet delay in relaying networks.

Let us consider a single buffer, as shown in Fig. 1, which is fed by a deterministic data arrival process and served by a single server. We assume that time is divided into slots with equal lengths, indexed by $t \in \{1, 2, \dots\}$. The total number of data packets that arrive at the buffer is

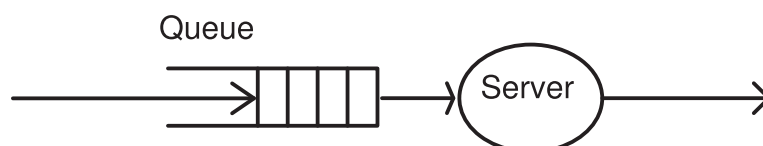


Fig. 1 Simple queueing system

N . Starting from $t = 1$, one packet arrives per time slot. Therefore, the last packet arrives at $t = N$. For simplicity, we assume that the arrivals occur at the beginning of time slots. The server might be active or inactive in each time slot. When it is active, it can serve only one packet per time slot, where the service implies delivering the packet successfully to the destination. When it is inactive, no packet is served.

We note that if the server is active in each time slot $t \in \{1, \dots, N\}$, each packet will be served immediately after its arrival. In this case, there is no queue formed in the buffer and consequently, each packet experiences an overall delay of one time slot, which is due to the time spent in the server. Accordingly, the packets will arrive in the destination at the beginning of time slots $t \in \{2, \dots, N + 1\}$. However, if the server is inactive in the first time slot, the first packet has to wait in the buffer until time slot 2, to get served. Then, in time slot 2, when the second packet arrives, the server is busy with serving the first packet. Therefore, the second packet also experiences one slot delay in the queue and one slot delay in the server. In a similar manner, all the following packets incur the same queueing and service delays. In other words, the delayed operation of the server causes the nonzero queueing delay for the first packet, which is transferred to the subsequent packets as well.

Based on the above discussion, if the server is inactive in time slot $x \in \{1, \dots, N\}$, it adds one slot to the queueing delay (and the overall waiting time) of every packet arrived in slot x or afterward. In general, the packet which arrived in time slot t will experience a queueing delay of n_t and will be delivered at time slot $t + n_t + 1$, where n_t indicates the number of slots before and including t in which the server was inactive. It is clear that the cause of queue formation in such systems is the interruption in the operation of the server, which is translated to queueing delays of the data packets.

3 Effect of buffer-aided relaying on the end-to-end packet delay

In this section, first we assume that data arrive in a deterministic manner and the availability of the channels follows Bernoulli distribution and provide an insight on the end-to-end delay performance for conventional and buffer-aided relaying systems. Then, we analytically derive the average end-to-end packet delay for

these systems, in the case that both the data arrival process and the availability of the channels follow Bernoulli distribution. Finally, we discuss general cases and present the intuitions about the end-to-end delay performance.

3.1 Relaying systems with deterministic data arrivals and Bernoulli channel conditions

Let us consider a relay network, with one source node, i.e., the BS, one relay node and one destination (or user) node, where the relay works based on the DF technique. It is assumed that there is no direct link between the BS and the user, and the transmissions are done only through the relay. There is only one channel in the system, which can be used for transmissions either from the BS to the relay or from the relay to the user. We use c_1 and c_2 to indicate the BS channel condition (for the link between the BS and relay) and relay channel condition (for the link between the relay and user), respectively. These variables can be either “Good” or “Bad”, meaning respectively that it is possible to transmit one or zero packet successfully on the corresponding channel. It is assumed that the channel conditions remain constant during each time slot but vary independently from one time slot to another. The probability of being “Good” is s_1 and s_2 for the BS and relay channel conditions, respectively. We assume that each time slot is further divided into two subslots, where the BS and relay can transmit a packet in the first and second subslots, respectively. The reason for considering subslots is stated later in Remark 1.

Figure 2a shows the queueing model for a conventional relaying system, where the relay does not have buffer and therefore, if it receives a packet in a subslot, it has to transmit it immediately in the next subslot. The server 1 and server 2 indicate the wireless channel from the BS to relay and from the relay to user, respectively. On the other hand, Fig. 2b indicates a relaying network, where the relay has a buffer which allows it to store the data packets and transmit whenever its channel is good. In both of the figures, the rectangle enclosed around the servers is to abstract the overall serving behavior of the system from the time that the BS starts to transmit a data packet until it is delivered to the user. Note that the works in [8–11] in fact study the delay by considering only the time a packet spends inside this rectangle and do not take into account the waiting time in the BS queue, as they assume infinitely backlogged

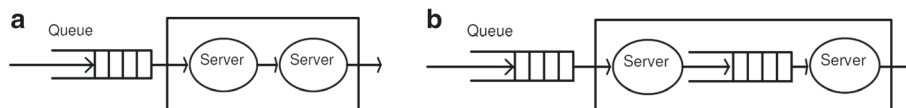


Fig. 2 Queueing model of (a) conventional relaying system and (b) buffer-aided relaying system

buffer for the BS. However, in practice, the packets arrive finitely at the BS and experience a queueing delay before the transmission from the BS to the relay. Therefore, the overall waiting time of a packet in the relay network, which we also refer to as end-to-end delay, includes both its waiting time in the BS queue and the time it spends inside the aforementioned rectangle (i.e., the time period between the transmission at the BS and successful reception at the destination). Note that the data arrivals at the BS might be due the packet generation in the BS itself, in which case the end-to-end packet delay might be referred to as “generation-to-delivery packet delay”; or it might be due to exogenous packet arrivals from an external network (e.g., the Internet), in which case the end-to-end packet delay might be referred to as “enter-to-exit packet delay” to specify the portion of the delay that a packet experiences since it enters the relay network until it exits it in the destination. In this paper, for simplicity, we will use the term “end-to-end” instead of “generation-to-delivery” or “enter-to-exit,” to specify the delay from the time that a packet arrives at the BS buffer until it is delivered to the destination.

In the following, we consider the data arrivals at the BS buffer as the deterministic process, with N packets, mentioned in the previous section. Taking the overall service behavior of the systems into account and based on the previous section, we discuss the overall waiting time of data packets in both the conventional and buffer-aided relaying systems.

Table 1 shows the different states for the joint conditions of the BS and relay channels, in which G and B indicate “Good” and “Bad” conditions, respectively. We assume that a central scheduler at the relay has the CSI in each time slot, using the pilot signals transmitted by the BS and destination through error-free control channels at the beginning of that time slot. Based on the CSI and the buffering capability of relay, the scheduler decides about the packet transmissions over the links and notifies the BS and the destination accordingly. These are explained in detail in the following.

In the case of conventional relaying (no buffer in the relay), only when $c_1c_2 = GG$, the scheduler notifies the BS to transmit a packet in the first subslot. Then, the relay forwards the received packet to the destination in the second subslot. In the other three cases, i.e., when one or both of c_1 and c_2 are “Bad,” the packets remain in the BS buffer

and are not transmitted. Therefore, conventional relaying serves the packets with the probability of $s = s_1s_2$ in each time slot. Consequently, based on the discussions in the previous section, the overall server in the system is inactive with the probability of

$$u_{nb} = P(GB) + P(BG) + P(BB) = 1 - s = 1 - s_1s_2 \quad (1)$$

where u_{nb} indicates the interruption probability for the overall server in the system without buffering at the relay. Considering this and the discussions in the previous section, in each time slot, the probability of “increase of one slot” in the overall waiting time of the packets present in that time slot or arrived after that is $u_{nb} = 1 - s_1s_2$. Here, the increase in the overall waiting time is due to the increase in the BS queueing delay of those packets.

Remark 1. Now, we explain the reason for considering subslots in each time slot. First note that the conventional relaying protocol stated above takes into account the CSI to decide about the packet transmission. This is reasonable as the CSI is assumed to be available at the scheduler, irrespective of exploiting buffer or not in the relay, and using it can prevent information loss in the case that successful delivery of packet is not possible (i.e., either one or both of the channel conditions are “Bad”). Second, since the channel conditions might vary from one time slot to another, the scheduler does not know what the CSI will be in the next time slot. Therefore, the CSI obtained in the beginning of a time slot can only be used to decide about the packet transmissions from the BS and relay during that time slot. Consequently, it is needed to have a subslot for the BS transmission and another one for the relay transmission, in conventional relaying. Note that alternatively, one might refer to subslots as slots, in which case the channel conditions remain constant over two time slots and the BS transmissions happen in the odd-numbered slots and the relay transmissions happen in even-numbered slots.

Now consider the system where the relay has a buffer, but as before, the BS can only transmit in the first subslot and the relay can transmit in the second subslot. We note that if the channel conditions are as BB in time slot x , similar to the system with conventional relaying, no transmission will be scheduled and therefore, there will be an increase of one slot in the overall waiting time of the packets present in time slot x or arriving afterward. However, for the channel conditions as GB and BG , the case is different. In order to clearly investigate these states, first we consider the following example:

- In time slot $t = 1$, the channel conditions are as GB . Therefore, in the first subslot, the BS transmission is

Table 1 Joint channel condition probabilities

c_1c_2	Probability
GG	s_1s_2
GB	$s_1(1 - s_2)$
BG	$(1 - s_1)s_2$
BB	$(1 - s_1)(1 - s_2)$

scheduled and packet 1 will be transmitted from the BS to relay; but due to the “Bad” channel condition of relay, it will not be transmitted to the user in the second subslot and will be stored in the relay buffer.

- In time slot $t = 2$, the channel conditions are as *BG*. Therefore, in the first subslot, there will not be any transmission from the BS to relay and the overall waiting time of the packets $2, \dots, N$ will be increased by one slot. However, due to good condition of the relay channel, packet 1 will be transmitted from the buffer of the relay to the user in the second subslot.

In the above example, it is observed that packet 1 is served by the relay in time slot $t = 2$ and therefore, it is delivered to the user in time slot $t = 3$. This has become possible due to the queueing of that packet in the relay buffer. Note that with conventional relaying, however, in the above example, packet 1 would remain in the BS queue in both time slots $t = 1$ and $t = 2$, and the overall waiting time would increase by two slots for all the packets. Based on the above discussion and considering the nonzero probability of having channel conditions as *GB* and *BG* in two consecutive time slots, it can be concluded that $u_b < u_{nb}$, where u_b is the interruption probability of the overall server in the buffer-aided relaying system. In other words, the buffering capability in relay reduces the interruption probability of the overall server and consequently, it reduces the overall waiting time for the data packets. This is achieved due to the fact that the queue size in the BS is reduced, and the data packets transferred to the relay buffer enable the efficient use of the relay channel.

3.2 Relaying systems with Bernoulli data arrivals and channel conditions

Now, we consider the relaying networks where both data arrivals and channel conditions follow Bernoulli distribution. We assume that in each time slot, the probability of one packet arrival at the BS buffer is a , and, as before,

the probability of “Good” channel condition for the BS and relay is equal to s_1 and s_2 , respectively. It is assumed that $a < s_1 s_2$ and therefore, the system queues are stable in the case of conventional and buffer-aided relaying [19, Chapter 2]. In the following, when we use subscripts b and nb for the variables, we refer to them in the case with buffering and without buffering at the relay, respectively.

3.2.1 Buffer-aided relaying system

Based on [20, Section 7.5], Fig. 3 shows the Markov Chain model for the queue dynamics at the BS buffer for the buffer-aided relaying network, where each state represents the number of packets in the queue. Let p_n , $n \in \{0, 1, \dots\}$ denote the probability that in steady state, there are n packets in the BS queue. Note that due to equilibrium in the steady state, we have:

$$p_0 = [1 - a(1 - s_1)]p_0 + s_1(1 - a)p_1,$$

$$p_n = a(1 - s_1)p_{n-1} + [1 - \{a(1 - s_1) + s_1(1 - a)\}]p_n + s_1(1 - a)p_{n+1}, n = 1, 2, \dots$$

Based on the above equations, the probability of each state can be written as

$$p_n = \rho^n p_0, n = 0, 1, 2, \dots, \tag{2}$$

where

$$\rho = \frac{a(1 - s_1)}{s_1(1 - a)}. \tag{3}$$

Considering the fact that $\sum_{n=0}^{\infty} p_n = 1$, we have:

$$p_0 = 1 - \rho. \tag{4}$$

Therefore, the expected number of packets in the BS queue can be expressed by

$$E(Q_b^B) = \sum_{n=0}^{\infty} np_n = \frac{\rho}{1 - \rho} = \frac{a(1 - s_1)}{s_1 - a}. \tag{5}$$

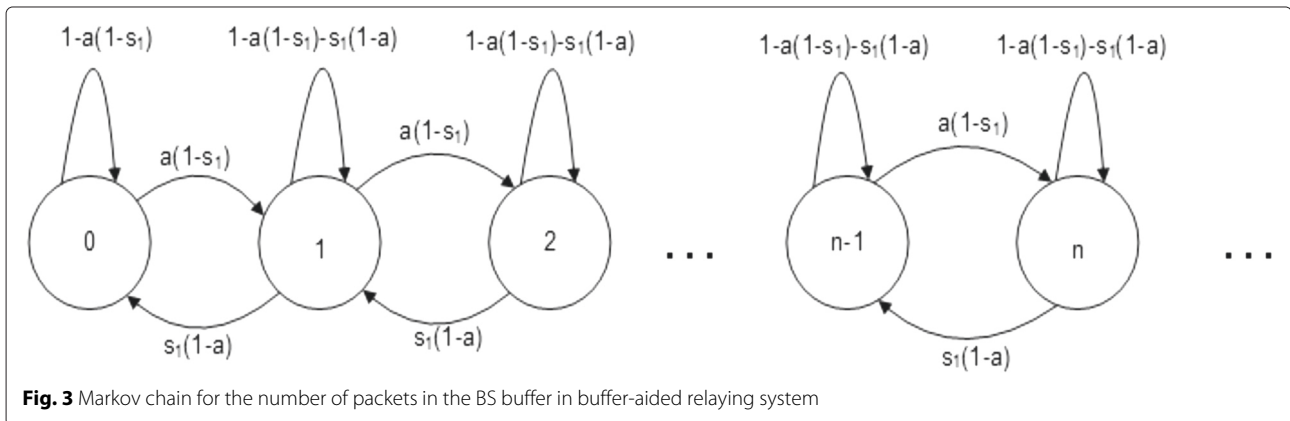


Fig. 3 Markov chain for the number of packets in the BS buffer in buffer-aided relaying system

Note that when a new packet arrives at the BS buffer, its expected delay until completion of its service by the BS can be split into two parts. The first part is the expected time that it has to wait until the packets already in the queue are served, i.e., $E(Q_b^B)E(T_b^B)$, where $E(T_b^B)$ is the expected delay imposed due the service of each packet when it is in the head of queue. The second part is the expected time since the packet itself gets to the head of the queue until its service is completed, which is denoted as $E(T_b^{B*})$. Therefore, the expected waiting time of a packet in the BS, $E(D_b^B)$, can be written as $E(D_b^B) = E(Q_b^B)E(T_b^B) + E(T_b^{B*})$. This is in fact the well known *mean value approach* which holds for queueing systems with memoryless data arrival processes [21, Section 4.3].

The interpretation of $E(T_b^B)$ is as follows. The delay caused due the service of a packet in the head of the queue is 1 slot with the probability of s_1 (this is in the case that the BS channel is good at the time that the packet gets to the head of queue). It is $(1 + 1)$ slots with the probability of $(1 - s_1)s_1$, $(2 + 1)$ slots with the probability of $(1 - s_1)^2s_1$, $(k + 1)$ slots with the probability of $(1 - s_1)^ks_1$, and so on. Therefore, the expected delay caused due to the service of a packet in the head of queue is given by

$$\begin{aligned}
 E(T_b^B) &= s_1 + (1 + 1)(1 - s_1)s_1 + (2 + 1)(1 - s_1)^2s_1 + \dots \\
 &= \sum_{k=0}^{\infty} (1 - s_1)^k s_1 (k + 1) \\
 &= s_1 \sum_{k=0}^{\infty} (1 - s_1)^k k + s_1 \sum_{k=0}^{\infty} (1 - s_1)^k \\
 &= s_1(1 - s_1) \left[\frac{d}{ds_1} \left(- \sum_{k=0}^{\infty} (1 - s_1)^k \right) \right] \\
 &\quad + s_1 \frac{1}{1 - (1 - s_1)} \\
 &= -s_1(1 - s_1) \frac{d}{ds_1} \frac{1}{s_1} + s_1 \frac{1}{s_1} \\
 &= \frac{1 - s_1}{s_1} + 1 \\
 &= \frac{1}{s_1}
 \end{aligned} \tag{6}$$

On the other hand, we can compute $E(T_b^{B*})$ as follows. Considering that the packet has reached the head of the queue, its delay until the departure from the BS is equal to 0.5 with the probability of s_1 , $(1 + 0.5)$ with the probability of $(1 - s_1)s_1$, $(2 + 0.5)$ with the probability of $(1 - s_1)^2s_1$, $(k + 0.5)$ with the probability of $(1 - s_1)^ks_1$, and so on. Hence, the expected waiting time of the packet after reaching the head of the queue is

$$\begin{aligned}
 E(T_b^{B*}) &= 0.5s_1 + (1 + 0.5)(1 - s_1)s_1 \\
 &\quad + (2 + 0.5)(1 - s_1)^2s_1 + \dots \\
 &= \sum_{k=0}^{\infty} (1 - s_1)^k s_1 (k + 0.5) \\
 &= s_1 \sum_{k=0}^{\infty} (1 - s_1)^k k + 0.5s_1 \sum_{k=0}^{\infty} (1 - s_1)^k \\
 &= \frac{1 - s_1}{s_1} + 0.5.
 \end{aligned} \tag{7}$$

Based on the above discussions, the expected total delay of a packet in the BS is equal to

$$\begin{aligned}
 E(D_b^B) &= E(Q_b^B)E(T_b^B) + E(T_b^{B*}) \\
 &= \frac{a(1 - s_1)}{s_1 - a} \frac{1}{s_1} + \frac{1 - s_1}{s_1} + 0.5 \\
 &= \frac{1}{s_1} \left[\frac{a(1 - s_1)}{s_1 - a} + 1 \right] - 0.5 \\
 &= \frac{1 - a}{s_1 - a} - 0.5.
 \end{aligned} \tag{8}$$

We note that in each time slot, either one or zero packet departs the BS. Therefore, the packet departures from the BS can be modeled as a Bernoulli process. Due to the stability of the queues, the data departure rate from the BS is equal to the data arrival rate in its buffer. Consequently, the probability that one packet departs the BS, or, equivalently, the probability that one packet arrives at the relay buffer, is equal to a . As a result, the average delay that a packet experiences in the relay can be computed in the similar manner as the average delay in the BS, which is expressed by

$$E(D_b^R) = \frac{1 - a}{s_2 - a} - 0.5. \tag{9}$$

Based on (8) and (9), the average waiting time of a packet in the buffer-aided relaying system is given by

$$E(D_b) = E(D_b^B) + E(D_b^R) = \frac{1 - a}{s_1 - a} + \frac{1 - a}{s_2 - a} - 1. \tag{10}$$

3.2.2 Conventional relaying system

Note that in the conventional relaying system, the BS can serve the packets in its buffer only when both its own channel and the relay channel are in good condition. Hence, the service probability for serving the BS buffer is s_1s_2 . Considering that, the average number of packets in the BS buffer can be obtained by replacing s_1 in (5) with s_1s_2 , i.e., $E(Q_{nb}^B) = \frac{a(1 - s_1s_2)}{s_1s_2 - a}$. Similarly, the average delay caused for a packet due to the service of each packet in front of it can be computed based on (6) and by using s_1s_2 instead of s_1 , i.e., $E(T_{nb}^B) = \frac{1}{s_1s_2}$. Also, the average delay that a packet experiences when it gets to the head

of the queue can be obtained, based on (7), as $E(T_{nb}^{B*}) = \frac{1-s_1s_2}{s_1s_2} + 0.5$. Therefore, we have:

$$E(D_{nb}^B) = E(Q_{nb}^B)E(T_{nb}^B) + E(T_{nb}^{B*}) = \frac{1-a}{s_1s_2-a} - 0.5. \quad (11)$$

On the other hand, when a packet arrives at the relay, it is immediately served without waiting in any buffer. Therefore, it only spends 0.5 of a slot in the relay, which is due to the service time in the relay. Consequently, we have:

$$E(D_{nb}^R) = 0.5. \quad (12)$$

Based on (11) and (12), the average waiting time of a packet in the conventional relaying system is given by

$$E(D_{nb}) = E(D_{nb}^B) + E(D_{nb}^R) = \frac{1-a}{s_1s_2-a}. \quad (13)$$

In order to compare the delay performance of the conventional and buffer-aided relaying systems, *Theorem 1* states and proves the main result of this subsection.

Theorem 1. Consider a relaying network where the data arrival process at the BS and the channel availability processes follow Bernoulli distribution and the packet arrival probability satisfies stability condition $a < s_1s_2$. Then, the average end-to-end packet delay in the buffer-aided relaying system is less than or equal to that in the conventional one. In other words, we have:

$$E(D_b) \leq E(D_{nb}), \quad (14)$$

where the equality holds only in the case that the channels are always in "Good" condition, i.e., $s_1 = s_2 = 1$.

Proof. Please refer to the Appendix. \square

3.3 General relaying system

Now we consider a general scenario, where the data arrival and channel condition processes follow general distributions, and are stationary and ergodic. We assume that the data arrivals and transmission rates have finite mean and variance. We use $r_{br}(t)$, $r_{rd}(t)$, and $r_{bd}(t)$ to show the achievable transmission rate in time slot t between the BS and relay, the relay and destination, and the BS and destination, respectively. Without buffering, the BS needs to transmit to the relay in the first subslot, and then the relay has to forward it immediately in the next subslot. We know that in this case, the end-to-end achievable rate between the BS and the user is $r_{bd}(t) = \frac{1}{2} \min\{r_{br}(t), r_{rd}(t)\}$. Therefore, the scheduler in the relay notifies the BS in the beginning of each time slot to transmit with a rate that can be supported by both of the links to lead to a successful reception at the destination. Due to

this, the transmission rate in each slot is limited by the link with the worst channel condition in that time slot.

However, when the relay has a buffer, there is no necessity for the immediate forwarding of the data and the abovementioned limitation is relaxed; therefore, the BS has the opportunity for transmitting continuously to the relay when the channel condition from the BS to relay is good. Then, the relay can store them in the buffer to transmit when the channel from the relay to user is good. Because of this, the buffering makes it possible to improve the system throughput as shown in [8–11]. Improvement in the throughput is equivalent to the improvement in the average end-to-end service rate of the data arrived at the BS buffer. In other words, the increase in the system throughput means that more data is transferred from the BS to the user, or equivalently, the same data is transferred from the BS to the user in a less amount of time. Therefore, on average, packets experience lower end-to-end delay, i.e., the delay since their arrival at the BS until delivery to the destination.

Based on the above discussion, we make the conclusion as follows. Although buffer-aided relaying results in queueing delay in the relay, it also facilitates data transfer from the BS to the user and leads to a large reduction in the queueing delay at the BS. Therefore, the overall effect is the improvement of the average end-to-end packet delay. In summary, we state this as follows.

Proposition. Using buffer in the relay improves the system throughput, and therefore, it reduces the average end-to-end packet delay. \square

Remark 2. We note that the given proposition is about the *average* end-to-end packet delay. There might be some packets that experience larger end-to-end delays in buffer-aided relaying compared with the conventional relaying. However, reduction in the average end-to-end packet delay indicates that *most* of the packets experience less delay in the case of buffer-aided relaying compared with conventional relaying. This is confirmed in the next section, in Figs. 11 and 16. Moreover, we note that the above discussions do not explain anything about the maximum and minimum possible end-to-end packet delays in buffer-aided relays. In general, considering the queueing dynamics in both the BS and relay, the maximum possible end-to-end packet delay in both conventional and buffer-aided relaying is infinite, which is due to the infinite buffer size of the BS and relay. However, simulation results presented in the next section indicate that usually the maximum end-to-end packet delay is less in buffer-aided relaying compared with the conventional relaying. On the other hand, the minimum possible end-to-end packet delay in both of the relaying systems is one time slot, which happens when there is no queue neither in the

BS buffer nor in the relay buffer; in such a case, when a packet arrives at the BS, it can be immediately transmitted to the relay and then, the relay can immediately transmit it to the destination. This will take totally two subslots or equivalently one time slot.

In the previous subsection, even in the buffer-aided relaying system, we assumed that the BS and relay transmissions are a priori scheduled to be done in the first and second subslots. In general, when buffering is exploited in the relay, each subslot can be used dynamically for the BS transmission or relay transmission, if there are data in their buffers. In this regard, a dynamic scheduling policy is required to stabilize the system queues. Specifically, in each subslot, this policy should decide on allocating the channel to the BS or relay such that the system queues remain bounded. For this, the well-known Max-Weight (MW) algorithm can be used, which has the largest stability region [19, 22, 23]. MW aims at maximizing the weighted rates of the links, where the weight of a link is considered equal to the difference of the queue sizes at the two ends of the link. Note that due to the interdependence of the queue sizes and the scheduling decision in MW, it is highly intractable to derive the expressions for average queue sizes and delays under the MW policy. However, if the data arrival rate is inside the stability region (so it can be supported by the network capacity), it is guaranteed that scheduling the links by MW policy will result in bounded average queue sizes and delays [19, 22, 23]. MW is an attractive scheduling policy for stabilizing the queues in buffer-aided relay networks as it works by utilizing just the instantaneous queue and channel state information (QCSI) and does not require information about the probability distribution of packet arrival processes and channel states. Considering the abovementioned, we summarize the costs of buffer-aided relaying in the following remark.

Remark 3. Note that the costs for the improvements brought by buffer-aided relaying are the requirement for a memory to buffer data at the relay and the necessity for a scheduling algorithm to keep the queues stable.

Remark 4. It is worth noting that the proposition stated above can also be considered for the case of relay networks with more than two hops or more than one relays. For successful data transmission with conventional relaying in a multihop network, it is needed to have the channel states of all the hops from the source to destination favorable during the transmission interval. However, with buffer-aided relaying, it is possible to use the channels more opportunistically. This is studied in [24], where it is shown that the outage (or unsuccessful packet reception) is reduced in buffer-aided multihop networks, which is equivalent to improvement in throughput. Similarly, it is shown in [25] that in a network with multiple relays,

the system throughput is improved in buffer-aided relaying compared with the case without buffers at the relays. Therefore, based on those results and considering the queueing delays both in the source and the relay, we conclude that the packets that are successfully received at the destination experience lower average end-to-end delay in the aforementioned relaying systems when buffering used in relays compared with the case without buffers in relays. The analysis for deriving the exact expressions of average end-to-end packet delay in these scenarios needs more investigation, as the effect of the relay selection policy should also be taken into account, and is an interesting research topic for future works.

4 Numerical results

To verify the presented discussions, we have conducted extensive Matlab simulations over 10,000 time slots and more. We have investigated the cases that the data arrival and channel condition processes follow Bernoulli distribution, as well as general cases with the settings of a practical system. We present the simulation results in the following.

4.1 Bernoulli data arrivals and channel conditions

In order to validate the analysis provided in Subsection 3.2, in Figs. 4, 5, and 6, we present the average packet delay obtained from both the analytical expressions and the simulation. In each of these figures, we have fixed the values of s_1 and s_2 and have evaluated the effect of increase in a on the average end-to-end packet delay. In order to maintain the stability of the system queues with both conventional and buffer-aided relaying, for each figure, we have considered $a < s_1 s_2$. Figure 4 displays the case with high probability for the good channel conditions at the BS and relay, i.e., $s_1 = s_2 = 0.9$. It is clear that the analytical results are quite close to the simulation ones. Moreover, the results confirm that the buffer-aided relaying has lower packet delays compared with the conventional relaying. As expected, both of the systems incur larger delay as the packet arrival probability increases. However, the delay in the conventional relaying system increases faster comparing with that in the buffer-aided relaying system.

Furthermore, Figs. 5 and 6 show the results for the cases that either one or both of the channels have relatively lower probability of being in good condition. It is observed that in these cases, the conventional relaying results in significantly higher delays even at the lower data arrival rates. In particular, the performance difference of these relaying systems is larger in Fig. 5 compared with Fig. 4 and the largest in Fig. 6. This is because when the probability of good channel conditions is low, in the case of conventional relaying, the BS has to wait for a long time before having both the channels favorable for transmission. However,

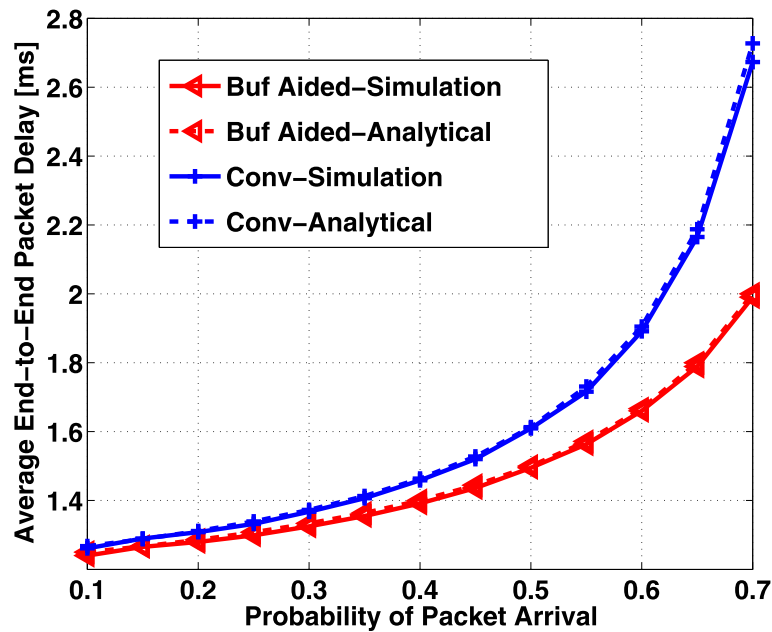


Fig. 4 Average end-to-end packet delay in the case of Bernoulli channel distribution with $s_1 = s_2 = 0.9$

in the case of buffer-aided relaying, the BS can transmit to the relay even when the relay channel is bad. Then, the relay can buffer the received data and transmit in its subslots whenever its channel is good.

We have also conducted simulations to investigate the average total queue sizes when the packet arrival

probabilities get close to the stability region boundaries, in the case of $s_1 = 0.5, s_2 = 0.5$. We note that based on [19, Chapter 2], the stability region boundary in conventional relaying is equal to $s_1 s_2 = 0.25$ whereas it is equal to $\min[s_1, s_2] = 0.5$ in buffer-aided relaying. Also, note that in conventional relaying, the total queue size is

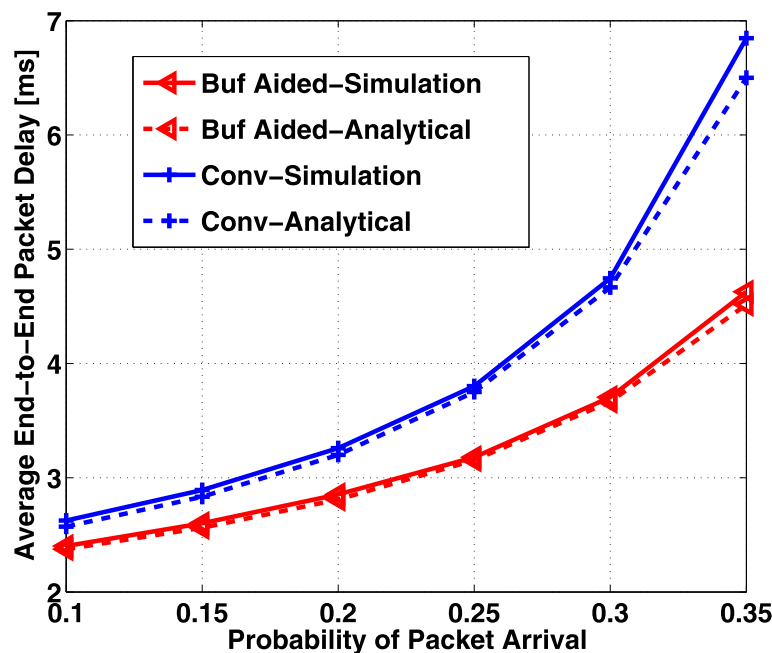


Fig. 5 Average end-to-end packet delay in the case of Bernoulli channel distribution with $s_1 = 0.5, s_2 = 0.9$

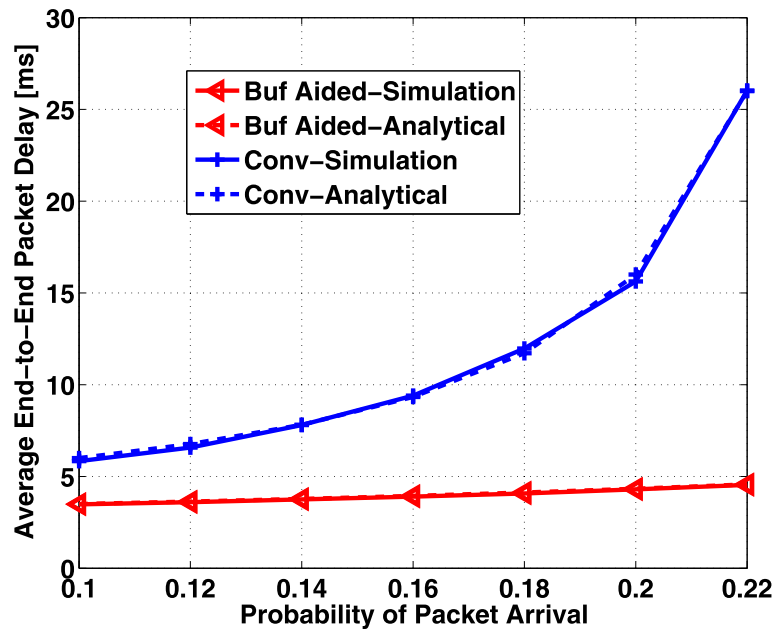


Fig. 6 Average end-to-end packet delay in the case of Bernoulli channel distribution with $s_1 = s_2 = 0.5$

the BS queue size whereas in buffer-aided relaying, the total queue size is the sum of the BS queue size and relay queue size. Therefore, based on (5) and the discussions in Subsection 3.2, the mathematical expression of average total queue size is $\frac{a(1-s_1s_2)}{s_1s_2-a}$ in conventional relaying and $\frac{a(1-s_1)}{s_1-a} + \frac{a(1-s_2)}{s_2-a}$ in buffer-aided relaying. The graphs for these equations as well as the results of simulations

are shown in Figs. 7 and 8. It is observed that the analytical results are close to the simulation ones. Furthermore, Fig. 7 shows that the average total queue size in conventional relaying system increases rapidly when the packet arrival probability gets close to 0.25 (the stability region boundary for conventional relaying). This is due to the fact that the probability of having both the channel conditions

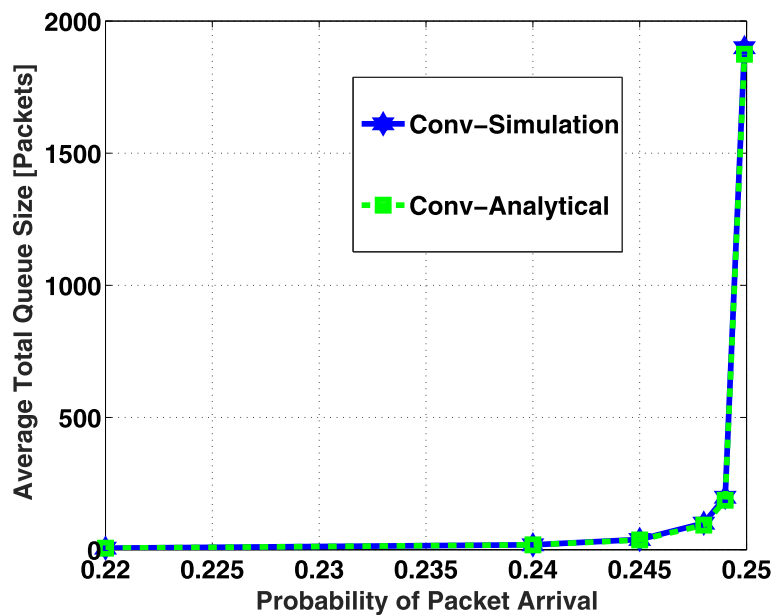


Fig. 7 Average total queue size in conventional relaying; the case of Bernoulli channel distribution with $s_1 = s_2 = 0.5$ and the packet arrival probability close to the stability region boundary

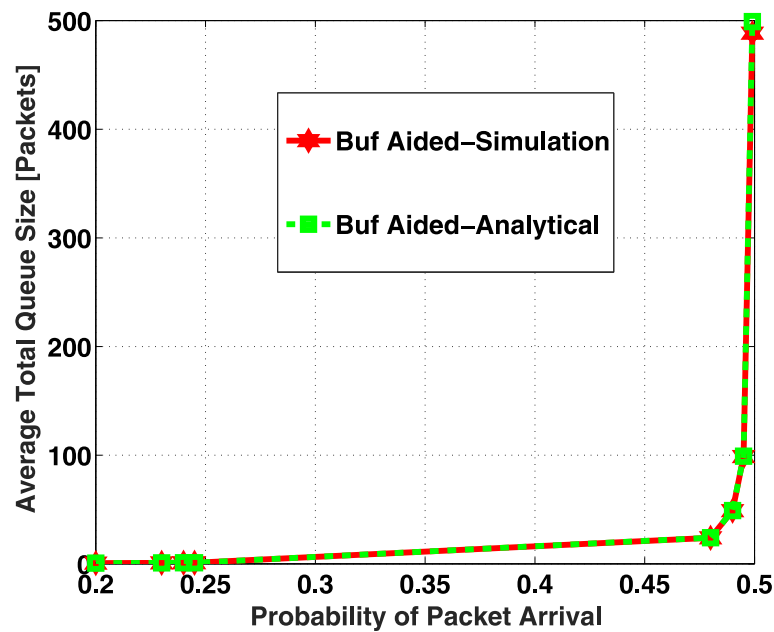


Fig. 8 Average total queue size in buffer-aided relaying; the case of Bernoulli channel distribution with $s_1 = s_2 = 0.5$ and the packet arrival probability close to the stability region boundary

“Good” (to be able to serve the packets in the BS queue in conventional relaying) is 0.25. When the data arrival probability gets close to this value, more packets have to wait in the queue until they get to the head of buffer and get the chance to be transmitted. On the other hand, it is observed in Fig. 8 that the similar effect happens in buffer-aided relaying in considerably larger packet arrival probability, i.e., 0.5 (the stability region boundary for buffer-aided relaying), and before that, the average total queue size is small. This means that in the arrival probabilities larger than 0.25 in conventional relaying, when a packet arrives at the BS, it is expected to encounter an infinite queue size (and end-to-end delay) in its path to the destination. However, even though in buffer-aided relaying, there are two buffers in the path of packets from the BS to the destination, it is expected that the packets will encounter a finite total queue size (and end-to-end delay) before reaching the destination, as long as their arrival probability is less than 0.5.

4.2 General scenario

Note that the mathematical analysis presented in Subsection 3.2 and the numerical results shown in Subsection 4.1 are for Bernoulli data arrivals and channel conditions and provide an insight on the effect of using a buffer in relay on the end-to-end packet delay. In order to verify the discussions presented in Subsection 3.3 for general data arrival and channel condition processes, we consider a scenario

with more realistic settings. For this scenario, the simulation parameters are shown in Table 2. It is assumed that the channel fading is flat over the system bandwidth and constant during each time slot; however, it can vary from one slot to another. For the link between the relay and user, Rayleigh channel model is used, and for the link from the BS to relay, Rician channel model is used with κ factor equal to 6 dB [26]. In the case of conventional relaying, the transmissions at the BS and relay are done in consecutive subslots. For buffer-aided relaying, we have used MW

Table 2 Simulation parameters

Parameter name	Setting
Cell radius	1000 m
Min UE-BS distance	50 m
BS antenna height	15 m
Relay antenna height	10 m
User antenna height	1.5 m
Relay distance from BS	1/2 cell radius
Pathloss model	From [27]
Channel bandwidth	180 KHz
Time slot duration	1 ms
Noise power spectral density	-174 dBm/Hz
Traffic model	Poisson
Packet size	1 Kbits

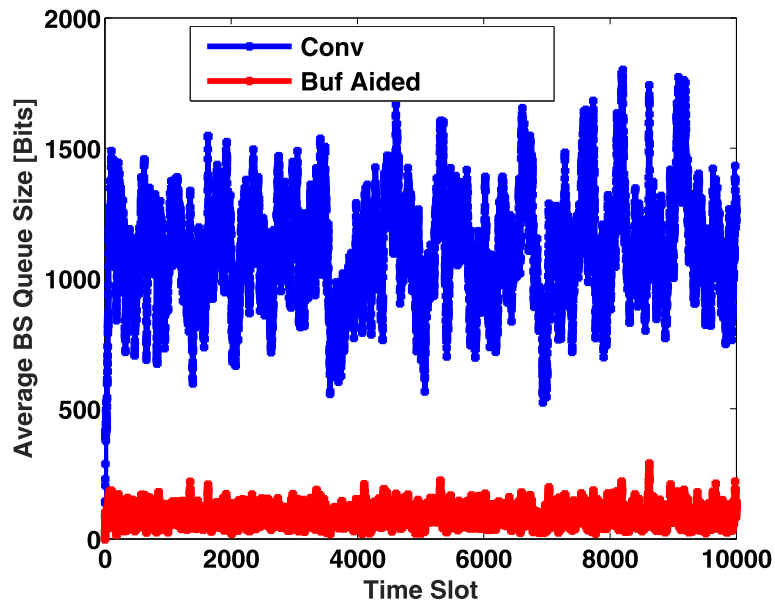


Fig. 9 Average BS queue size over time at the arrival rate of 50 packets/second

policy [19, 22, 23] to decide in an adaptive way, about the transmission in each subslot, either from the BS or from the relay buffer. The simulations were conducted for 100 independent realizations of channel condition and data arrival processes, each over 10,000 time slots.

Figures 9 and 10 show the BS and relay average queue sizes over time, respectively, at the arrival rate of 50 packets/second. The average queue size is obtained by

taking the average of queue sizes over 100 simulations. It is observed that with buffer-aided relaying, although data are queued in the relay, the average BS queue size in each time slot is reduced significantly. This results in lower average end-to-end packet delays in buffer-aided relaying compared with the conventional relaying, as shown in Fig. 11. In particular, in this scenario, the average end-to-end packet delays are 11 ms and 30 ms in buffer-aided

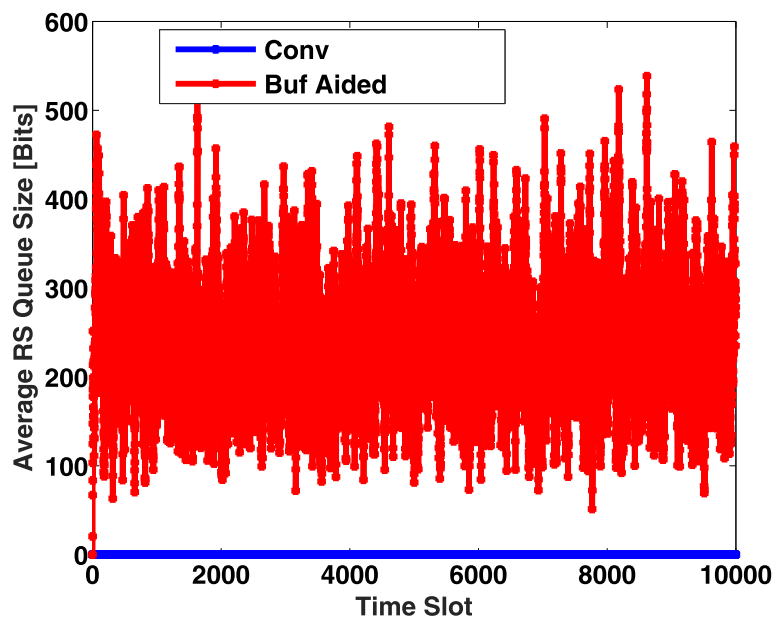
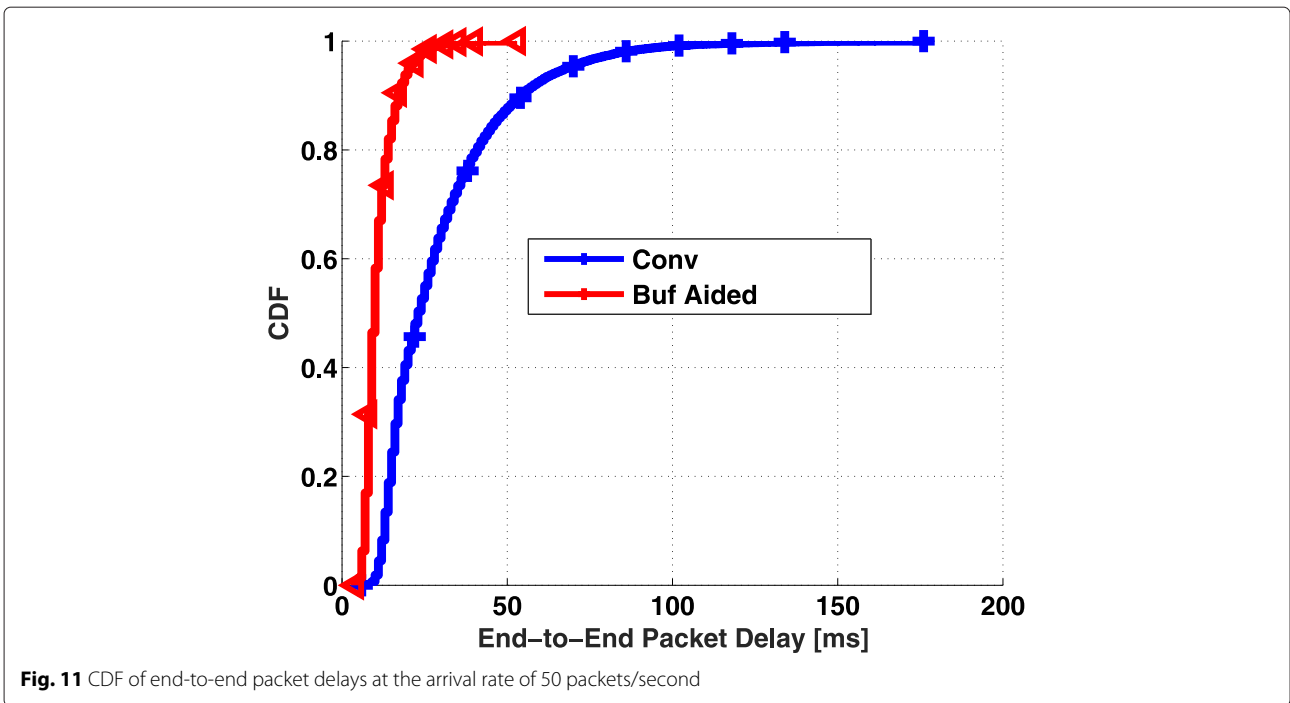


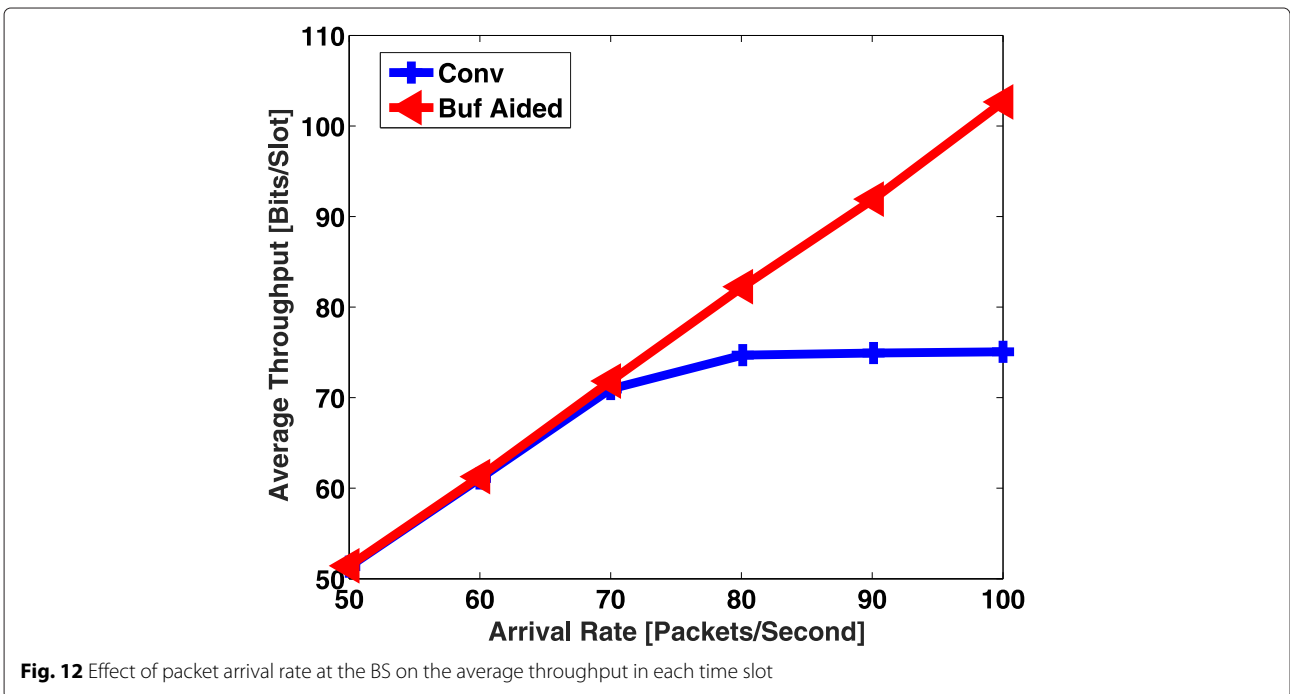
Fig. 10 Average relay queue size over time at the arrival rate of 50 packets/second



and conventional relaying, respectively. Note that Fig. 11 indicates that in general, the average end-to-end packet delay is less in buffer-aided relaying. In other words, even though some packets might experience larger overall waiting time compared with the conventional relaying, most of the packets experience lower delay since their arrival

at the BS until delivery to the destination. Moreover, it is observed that the maximum end-to-end packet delay is less in the case of buffer-aided relaying.

Next, we investigate the effect of increase in the packet arrival rate on the throughput and delay performance. It is observed in Fig. 12 that the conventional relaying is



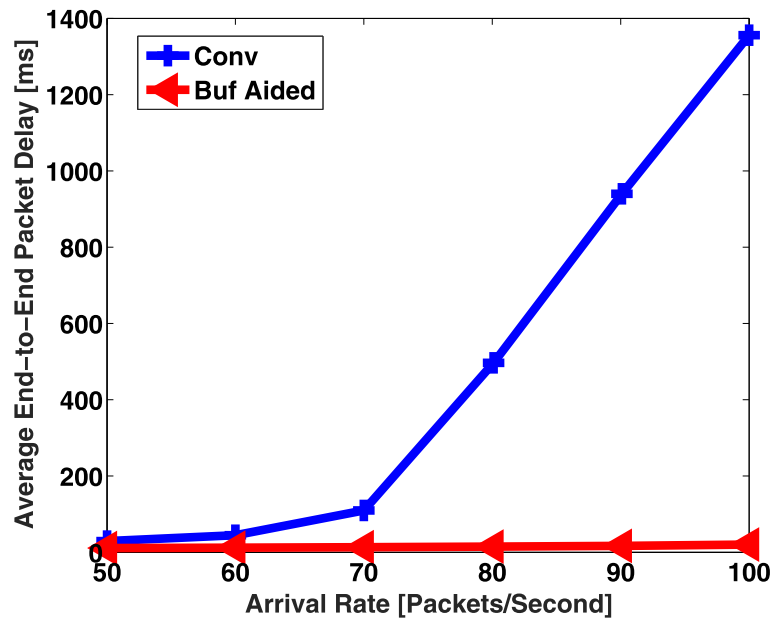


Fig. 13 Effect of packet arrival rate at the BS on the average end-to-end packet delay

able to support data arrival rates up to 60 packets/second, in which range it results in the average throughput equal to data arrival rate at the BS. However after that, due to low capacity, it starts to get saturated. This leads to queue instability and large end-to-end delays for packets, as shown in Fig. 13. In contrast, the buffer-aided relaying is able to provide the average throughput equal to the data

arrival rate, in all the packet arrival rates, and therefore leads to very low end-to-end packet delays.

In order to have a complete picture, we also present the system performance in the arrival rate of 100 packets/second, in Figs. 14 and 15. Figure 14 shows that in conventional relaying, the average BS queue size grows unbounded; this is due to the low capacity of relaying channel which

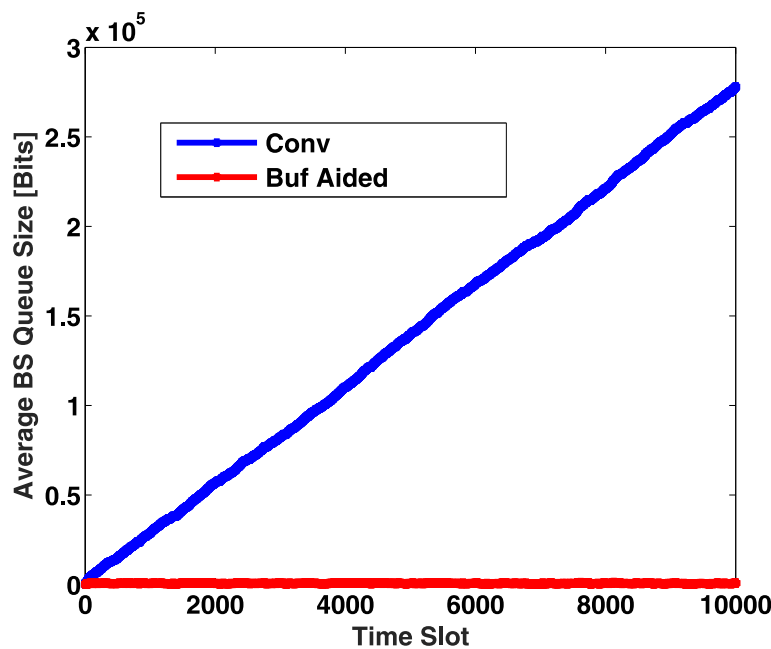


Fig. 14 Average BS queue size over time at the arrival rate of 100 packets/second

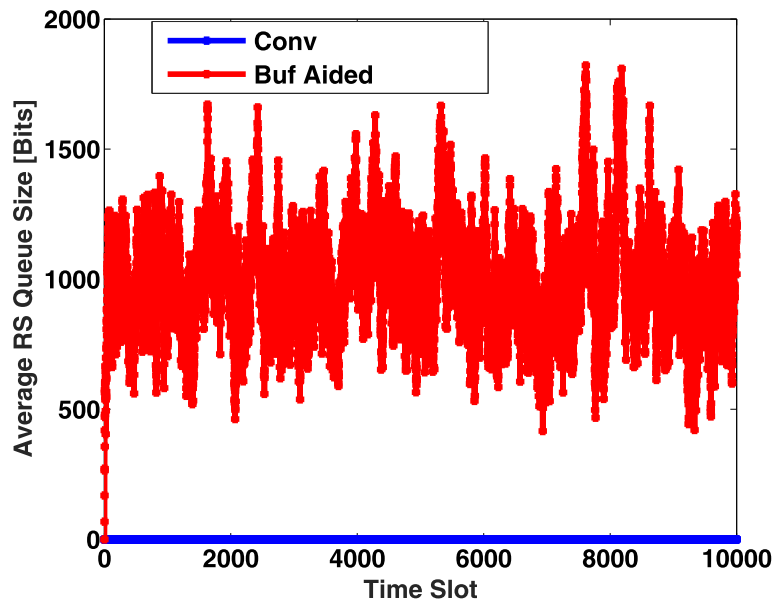


Fig. 15 Average relay queue size over time at the arrival rate of 100 packets/second

is unable to serve all the arrived data. This leads to large end-to-end packet delays as depicted in Fig. 16. On the other hand, as shown in Fig. 15, buffer-aided relaying leads to queueing in the relay buffer, which helps to utilize the channel variations efficiently. It allows to transfer the data from the BS buffer to relay buffer and from relay buffer to user, when the corresponding channels have

good conditions, and therefore leads to low end-to-end packet delays. In particular, in this scenario, the average end-to-end packet delays are 20 ms and 1355 ms, respectively, in buffer-aided and conventional relaying.

The above results confirm that using buffer in relay improves the throughput as well as the average end-to-end packet delay in the system.

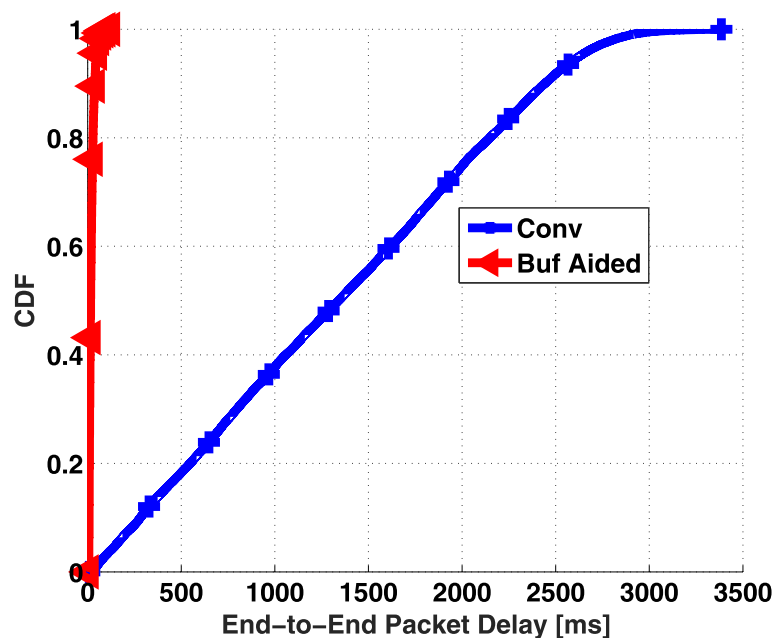


Fig. 16 CDF of end-to-end packet delays at the arrival rate of 100 packets/second

5 Conclusions

In this paper, we have studied the effect of buffering at the relay on the end-to-end delay performance. Through the discussions about queueing delay, we have explained the cause of delay in a simple queueing system. Based on that, we have provided an insight on the overall delay in the conventional and buffer-aided relaying networks. Moreover, for the case of Bernoulli data arrivals and channel conditions, we have proved analytically that the average packet delay is lower in buffer-aided relaying system compared with the conventional one. Finally, based on intuitive reasoning for general scenarios, we have concluded that employing buffer in the relay improves both the system's throughput and average end-to-end packet delay. Using numerical results, we have verified our analysis and discussions, and shown that using buffer in the relay leads to higher system throughput and lower average end-to-end packet delay.

Appendix

In order to prove that the buffer-aided relaying system incurs equal or lower delay compared with the conventional one, it is required to prove $E(D_{nb}) - E(D_b) \geq 0$. To show this, note that

$$E(D_{nb}) - E(D_b) = \frac{1-a}{s_1 s_2 - a} - \frac{1-a}{s_1 - a} - \frac{1-a}{s_2 - a} + 1 \quad (15)$$

By adding and subtracting the term $\frac{1-a}{s_1-a} \frac{1-a}{s_2-a}$ and rearranging the equations, we have

$$\begin{aligned} E(D_{nb}) - E(D_b) &= \frac{1-a}{s_1 - a} \frac{1-a}{s_2 - a} - \frac{1-a}{s_1 - a} - \frac{1-a}{s_2 - a} \\ &\quad + 1 + \frac{1-a}{s_1 s_2 - a} - \frac{1-a}{s_1 - a} \frac{1-a}{s_2 - a} \\ &= \left(\frac{1-a}{s_1 - a} - 1 \right) \left(\frac{1-a}{s_2 - a} - 1 \right) \\ &\quad + \frac{1-a}{s_1 s_2 - a} - \frac{1-a}{s_1 - a} \frac{1-a}{s_2 - a}. \quad (16) \end{aligned}$$

Note that the packet arrival probability is nonzero and the stability condition holds, i.e., $0 < a < s_1 s_2$. Since $s_i \leq 1$, $i = 1, 2$, we have $0 < a < s_i$, $i = 1, 2$, and $\frac{1-a}{s_i-a} \geq 1$, $i = 1, 2$. Therefore, the first term in the right hand side of (16) is non-negative. Hence, it suffices to show

$$\frac{1-a}{s_1 s_2 - a} \geq \frac{1-a}{s_1 - a} \frac{1-a}{s_2 - a}. \quad (17)$$

By canceling $1-a$ and cross-multiplying in (17), we obtain

$$(s_1 - a)(s_2 - a) \geq (1-a)(s_1 s_2 - a). \quad (18)$$

After multiplying both sides out and canceling the common terms of (18), we have

$$a(1-s_1)(1-s_2) \geq 0, \quad (19)$$

which is always true since $s_1 \leq 1$ and $s_2 \leq 1$.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by the Canadian Natural Sciences and Engineering Research Council through grants RGPIN-2014-06119 and RGPAS-462031-2014, and the National Natural Science Foundation of China through Grant No. 61271182. The work of Amr Mohamed was supported by NPRP 5-782-2-322 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Author details

¹ECE Department, The University of British Columbia, Vancouver, Canada. ²CSE Department, Qatar University, Doha, Qatar.

Received: 29 May 2015 Accepted: 12 November 2015

Published online: 15 December 2015

References

1. A Host-Madsen, J Zhang, Capacity bounds and power allocation for wireless relay channels. *IEEE Trans. Inf. Theory*. **51**(6), 2020–2040 (2005)
2. K Azarian, H El Gamal, P Schniter, On the achievable diversity-multiplexing tradeoff in half-duplex cooperative channels. *IEEE Trans. Inf. Theory*. **51**(12), 4152–4172 (2005)
3. C Hoymann, W Chen, J Montojo, A Golitschek, C Koutsimanis, X Shen, Relaying operation in 3GPP LTE: challenges and solutions. *IEEE Commun. Mag.* **50**(2), 156–162 (2012)
4. D Ng, R Schober, Cross-layer scheduling for OFDMA amplify-and-forward relay networks. *IEEE Trans. Veh. Technol.* **59**(3), 1443–1458 (2010)
5. Z Chang, T Ristaniemi, Z Niu, Radio resource allocation for collaborative OFDMA relay networks with imperfect channel state information. *IEEE Trans. Wirel. Commun.* **13**(5), 2824–2835 (2014)
6. D Zhang, Y Wang, J Lu, Qos aware relay selection and subcarrier allocation in cooperative OFDMA systems. *IEEE Commun. Lett.* **14**, 294–296 (2010)
7. D Zhang, X Tao, J Lu, M Wang, Dynamic resource allocation for real-time services in cooperative OFDMA systems. *IEEE Commun. Lett.* **15**, 497–499 (2011)
8. B Xia, Y Fan, J Thompson, H Poor, Buffering in a three-node relay network. *IEEE Trans. Wirel. Commun.* **7**, 4492–4496 (2008)
9. N Mehta, V Sharma, G Bansal, Performance analysis of a cooperative system with rateless codes and buffered relays. *IEEE Trans. Wirel. Commun.* **10**, 2816–2840 (2011)
10. N Zlatanov, R Schober, Buffer-aided relaying with adaptive link selection-fixed and mixed rate transmission. *IEEE Trans. Inf. Theory*. **59**, 2816–2840 (2013)
11. N Zlatanov, A Ikhlef, T Islam, R Schober, Buffer-aided cooperative communications: opportunities and challenges. *IEEE Commun. Mag.* **52**, 146–153 (2014)
12. I Krikidis, T Charalambous, J Thompson, Buffer-aided relay selection for cooperative diversity systems without delay constraints. *IEEE Trans. Wirel. Commun.* **11**(5), 1957–1967 (2012)
13. T Islam, A Ikhlef, R Schober, V Bhargava, in *Proc. IEEE Global Telecommun. Conf. Multisource buffer-aided relay networks: Adaptive rate transmission*, (2013), pp. 3577–3582
14. I Ahmed, A Ikhlef, R Schober, R Mallik, Power allocation for conventional and buffer-aided link adaptive relaying systems with energy harvesting nodes. *IEEE Trans. Wirel. Commun.* **13**(3), 1182–1195 (2014)
15. H Liu, P Popovski, E de Carvalho, Y Zhao, Sum-rate optimization in a two-way relay network with buffering. *IEEE Commun. Lett.* **17**(1), 95–98 (2013)
16. M Darabi, V Jamali, B Maham, R Schober, Adaptive link selection for cognitive buffer-aided relay networks. *IEEE Commun. Lett.* **19**(4), 693–696 (2015)

17. J Yang, Y Ran, S Chen, W Li, L Hanzo, Online source rate control for adaptive video streaming over HSPA and LTE-style variable bitrate downlink channels. To appear in *IEEE Trans. Veh. Technol.*, 1–1 (2015)
18. R Zhu, J Yang, Buffer-aware adaptive resource allocation scheme in LTE transmission systems. *EURASIP J. Wirel. Commun. Netw.*, 1–16 (2015)
19. M Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. (Morgan & Claypool, San Rafael, 2010)
20. F Gebali, *Analysis of Computer and Communication Networks*. (Springer, New York, 2008)
21. I Adan, J Resing, Queueing Systems. Online-Available: <http://www.win.tue.nl/~iadan/queueing.pdf>, Eindhoven University Netherlands (2015)
22. M Neely, E Modiano, C Rohrs, Dynamic power allocation and routing for time varying wireless networks. *IEEE J. Sel. Areas Commun., Special Issue on Wireless Ad-hoc Networks*. **23**(1), 89–103 (2005)
23. L Georgiadis, M Neely, L Tassiulas, Resource allocation and cross-layer control in wireless networks. *Foundation and Trends in Networking*. **1**(1), 1–144 (2006)
24. C Dong, L Yang, L Hanzo, Performance analysis of multihop-diversity-aided multihop links. *IEEE Trans. Veh. Technol.* **61**(6), 2504–2516 (2012)
25. A Ikhlef, DS Michalopoulos, R Schober, Max-max relay selection for relays with buffers. *IEEE Trans. Wirel. Commun.* **11**(3), 1124–1135 (2012)
26. M Jeruchim, P Balaban, K Shanmugan, *Simulation of Communication Systems: Modeling, Methodology and Techniques*, 2nd edn. (Kluwer Academic, Dordrecht, 2000)
27. Tech. Rep 3GPP TR 25.996 V7.0.0 (2007-06), Spatial channel model for multiple input multiple output (MIMO) simulations. available at: <http://www.3gpp.org/DynaReport/25996.htm>

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Immediate publication on acceptance
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ springeropen.com
