Relay Load Balancing in Queued Cooperative Wireless Networks with Rateless Codes

Gaurav Bansal Vinod Sharma Neelesh B. Mehta Eitan Altman

***Abstract*—Relay selection combined with buffering of packets of relays can substantially increase the throughput of a cooperative network that uses rateless codes. However, buffering also increases the end-to-end delays due to the additional queuing delays at the relay nodes. In this paper we propose a novel method that exploits a unique property of rateless codes that enables a receiver to decode a packet from non-contiguous and unordered portions of the received signal. In it, each relay, depending on its queue length, ignores its received coded bits with a given probability. We show that this substantially reduces the end-to-end delays while retaining almost all of the throughput gain achieved by buffering. In effect, the method increases the odds that the packet is first decoded by a relay with a smaller queue. Thus, the queuing load is balanced across the relays and traded off with transmission times. We derive explicit necessary and sufficient conditions for the stability of this system when the various channels undergo fading. Despite encountering analytically intractable G/GI/1 queues in our system, we also gain insights about the method by analyzing a similar system with a simpler model for the relay-to-destination transmission times.**

I. Introduction

In a cooperative relay communication system, a source collaborates with a single relay to transmit to a destination. When multiple relays are present, the source selects the most suitable relay on the basis of instantaneous channel conditions [1], [2]. Thus, cooperation exploits the broadcast nature of the wireless channel and the spatial diversity from multiple relays [3]–[5].

Channel knowledge at the transmitting source and relay nodes is often required for cooperation. However, acquiring it for the purpose of selection and transmission incurs significant bandwidth and throughput inefficiencies [6], and complicates the system design. Recently, rateless codes have been shown to be well suited for cooperative decode-and-forward (DF) relay networks [7]–[10] as they effectively mitigate this problem – they do not require the transmitting nodes to have any instantaneous channel knowledge.

Unlike conventional codes, which generate a finite number of parity bits, rateless codes generate a potentially unbounded number of parity bits that are transmitted until one acknowledgment (ACK) is received from the receiver [11]–[16]. Rateless codes are also spectrally efficient because the realized rate, which equals the number of information bits divided by the total times taken to transmit, is close to the mutual information of the channel. Luby-Transform codes and Raptor codes are practical examples of such codes [11].

G. Bansal, V. Sharma, and N. B. Mehta are with the Dept. of Electrical Communication Eng. at the Indian Institute of Science (IISc), Bangalore, India. E. Altman is with INRIA, France.

This work was supported in part by the DRDO-IISc Program on Advanced Research in Mathematical Engineering.

When multiple relays are present, the source transmits rateless coded bits/symbols until the relay that is the first to successfully decode the packet transmitted by the source sends an ACK [7], [10]. This relay, in turn, transmits the packet using another rateless code until it receives an ACK from the destination. The other relays play no further role in the transmission of this packet. An important property of this protocol is that the relay that first decodes the source's packet is automatically the one with the best source-to-relay (SR) channel among all the relays. Thus, the system exploits the spatial diversity from multiple relays without the source requiring any instantaneous channel knowledge.

Combining relay selection with buffering of packets at the relays further increases the throughput of the system, as was shown in [17]. This is because the source can start transmitting its next packet earlier – and to the relay with the best channel condition – while the current packet is en route to the destination. However, this gain comes at the expense of a larger average end-to-end (E-E) delay in the network due to queuing delays that now occur at the relays. One reason for these delays is the mismatch in the transmission rates of the SR and relay-to-destination (RD) channels. For example, the transmission rate over an SR channel may be high, but it might be low over its corresponding RD channel.

In this paper, we propose a novel and practical method that substantially reduces the queuing delays with only a marginal reduction in overall throughput. Thus, the significantly higher throughput of queuing and cooperation can be enjoyed with a much smaller increase in average delay. *In our method, relays with longer queues reduce the rate at which they receive new packets by ignoring portions of the incoming rateless encoded signal.* We present simple and effective criteria for ignoring portions of the incoming signals, and evaluate them using both analysis and simulations. The method utilizes a unique property of rateless codes that allows the receiver to decode the packet so long as the total mutual information that it has accumulated from portions of the received signal exceeds the entropy of the source information. Notably, the coded bits do not need to be received consecutively [11]–[13].

Intuitively, the method increases the odds that a relay with a smaller queue length and a good – but perhaps not the best – channel from the source is the first to receive the packet in its queue and send back an ACK. The method, thus, balances the queuing load across the many relays in the system. Notably, this is achieved without the source knowing the CSI or even the queue lengths at the relays.

The paper is organized as follows. We describe the system

978-1-4244-6404-3/10/$26.00 ©2010 IEEE

model in Sec. II, and analyze it in Sec. III-B. Simulation results are presented in Sec. IV, and are followed by our conclusions in Sec. V. Mathematical details are relegated to the Appendix.

II. System Model

Shown in Fig. 1 is a two-hop network in which the source, S, has a continuous stream of packets to transmit to the destination, D, via M decode-and-forward relays, $R_1, \ldots, R_M$. Each packet has a payload of B bits. The signal transmitted by a source is received by multiple relays due to the broadcast nature of the wireless channel.

The source as well as the relays use different rateless codes to transmit their packets. They transmit their signals over non-interfering orthogonal channels, each of bandwidth W Hz. This enables the various nodes to operate in a decentralized manner independently of each other. So long as a relay receives a sufficient number of coded bits, it can successfully decode the packet [13]. The decoded packet is then queued in the relay's buffer for transmission in a first come first serve fashion to the destination [17].

The time taken by a receiver to successfully decode a packet from the transmitter is $B_{\text{nrm}}/\log_e(1+\gamma)$, where $B_{\text{nrm}} = \frac{2B(1+\delta)}{W}\log_e 2$, δ is the inefficiency of the code, and γ is the signal to noise ratio (SNR) of the channel between the transmitter and receiver that includes fading [13]. We shall refer to $\frac{W}{2}\log_2(1+\gamma)$ as the transmission rate of the channel. The wireless channels between different nodes are assumed to be independent, frequency-flat, block-fading channels. The channel is assumed to be constant over the duration of transmission of the packet; it changes to an independent value thereafter.[1]

The source transmits until it receives an ACK back from *any one* of the relays. The ACKs are assumed to be error-free and are sent on a low bandwidth control channel that incurs negligible delays. Since only one bit needs to be communicated in an ACK, this can be easily ensured. The other relays thereafter play no further role in transmitting this packet.[2] The source then starts transmitting the next packet. As mentioned, the source and relays need not know any of the SR channel states and RD channel states, respectively.

To keep the analysis tractable, we assume infinitely large buffers at each relay; in practice, this is easily approximated by having sufficiently large buffer sizes. The relays are full-duplex capable. This capability is practically implemented using a duplexer and is extensively used today, for example, in most cell phones. All that is required is that the carrier frequencies used by the source and relays are spaced sufficiently apart.

A. Notation

Let t_n be the time when the nth packet of the source is decoded successfully by one of the relays, and enters that relay's queue. Then, $T_n = t_n - t_{n-1}$ is the time taken by the source to transmit the nth packet. Let $q_n(i)$ be the queue length, including the packet in service, at R_i's queue, $\mathcal{Q}_i$, at the time t_n^+. Let $\mathbf{q}_n = (q_n(1), \ldots, q_n(M))$. Let $r_n(i)$ and $r'_n(i)$ be the transmission rates of the S to R_i and R_i to D channels, respectively, during transmission of their nth packets. $\mathbf{E}[A|B]$ and $P[A|B]$ denote the expectation and probability of A conditioned on B. Let $\gamma(i)$ and $\lambda(i)$ denote the power gains of the channels from S to R_i and R_i to D, respectively. Let $\overline{\gamma}(i) = \mathbf{E}[\gamma(i)]$ and $\overline{\lambda}(i) = \mathbf{E}[\lambda(i)]$. Henceforth, we interchangeably use the terms 'packet transmission time' and 'service time'.

[1]For rateless codes, this assumption is valid in low and medium mobility scenarios of current high data rate wireless systems. It is commonly used in cooperative rateless code literature to ensure analytical tractability [7], [10].

[2]This is achieved by making the relays overhear the ACK. Alternately, the source itself can inform the relays about it when it transmits the next packet.

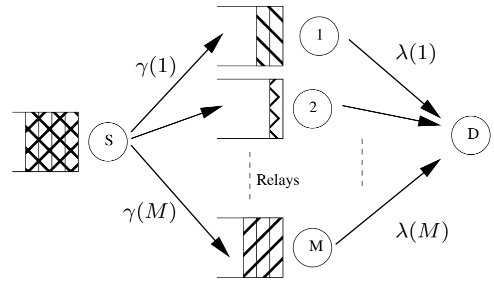

Fig. 1. Schematic of a two-hop queued cooperative relay network

B. Load Balancing Method

In the load balancing method, each relay autonomously reduces the rate at which it accumulates mutual information from the source from $r_n(i)$ *to* $(1-p)r_n(i)$ *when* $q_{n-1}(i) \geq q_{th}$*. Since the source transmits using rateless codes, this is achieved naturally when the relay ignores, with probability* p*, the signal its receives over the duration of a coded bit.* Here, p and q_{th} are system parameters that will be fine-tuned.

The key idea behind doing this is that it increases the odds that the packet is decoded first by a relay with a smaller queue, even if it has a marginally weaker SR channel. This packet then suffers from a smaller queuing delay. The method, thus, exploits the properties of rateless codes to trade off between queuing delays and transmission times. As we shall see, a larger q_{th} increases the throughput but also the delays. Where as, a larger p reduces delays but also the throughput.

Note that the source does not need to know the queue lengths at the relays, and continues to transmit as before. Only a relay needs to know its own queue length. Thus, no extra signalling between the various nodes is required.

C. Discussion: Alternate Models

We comment further on two modeling assumptions. (i) In this paper, we focus on DF relays since they have been studied for rateless code based cooperation in several papers [7]–[10]. (ii) We assign orthogonal channels to the source and relays since our goal is to demonstrate the inherent capability of a rateless code based cooperation system to balance load across its relays. In general, ideas related to rateless codes for multiple-input-multiple-output channels [16] can be used to further improve the spectral efficiency of the system.

III. Analysis

Our goal is to analyze the (i) *average throughput*, ψ, which is the rate at which D successfully receives packets, and (ii) stationary *mean E-E delay* of a packet from S to D, which captures the time taken by S to transmit the packet and the delay faced by the packet in a relay's queue. In order to do so, we will first develop criteria for determining when a relay's queue is stable. (We define stability below.) This is important because the queing delay and, thus, the E-E delay are unbounded if one or more relays have unstable queues. To conserve space, we focus on the scenario where the various SR channels are statistically identical, and so are the RD channels.

Given the packet-wise block fading assumption, it follows that $\{r_n(i)\}$ and $\{r'_n(i)\}$ are independent and identically distributed (i.i.d.) sequences. The sequence $\{\mathbf{q}_n\}$ is a regenerative sequence [18]. Its regeneration epochs are the times t_n at which the nth packet joins one of the queues, such that all the queues are empty just before t_n. A regeneration length is denoted by τ.[3]

A. General Conditions for Stability

A system is said to be stable if none of its queue lengths tends to ∞. If $\mathbf{E}[\tau] < \infty$ then $\{\mathbf{q}_n\}$ is stable and has a unique stationary distribution π, which equals $\lim_{n\to\infty} P[\mathbf{q}_n = \mathbf{q}] = \pi(\mathbf{q})$ for any initial distribution on $\mathbf{q}_n$ [18]. If $\mathbf{E}[\tau] = \infty$, then the system is unstable and $\lim_{n\to\infty} P\left[\sum_i \mathbf{q}_n(i) > x\right] = 1$, for any x. The following theorem, which holds for any Relay i, provides explicit necessary and sufficient conditions for stability. Let $r_n^{\max} = \max_i r_n(i)$.

Theorem *1:* The system is stable if and only if

$$M\mathbf{E}\left[\frac{1}{r_n^{\max}}\right]\frac{1}{(1-p)} > \mathbf{E}\left[\frac{1}{r'_n(i)}\right]. \qquad (1)$$

Proof: The proof is relegated to the Appendix. ∎

$\mathbf{E}\left[\frac{1}{r_n^{\max}}\right]$ and $\mathbf{E}\left[\frac{1}{r'_n(i)}\right]$ can be computed given the fading distribution. Closed-form expressions for Rayleigh fading are derived in [17], and are not repeated here to conserve space.

Comment: In practical systems, the source times out if it does not receive an acknowledgment within a time t_{out}. The stability condition in (1) and the analysis in Sec. III-B can be suitably modified to handle time-out. We do not show these modifications to conserve space.

B. Analysis of Throughput and E-E Delay

On account of the load balancing method, the rate at which packets arrive at a relay's queue depends on the queue length itself. Therefore, our system is a G/GI/1 queue, which is more general that a GI/GI/1 queue [18], and is well known to be analytically intractable. Consequently, general expressions for the throughput and stationary E-E delays are not known.

In order to understand the system dynamics better, this intractability motivates us to come up with a simpler phyiscal model that is relevant and yet enables an analysis. This is achieved by assuming that the service times of the R_i to D channel are i.i.d. exponential random variables (RVs), with mean rate μ. Note that the SR channels still undergo block fading. Henceforth, we shall refer to this model as the *SR Fading Channel & RD Exponential Service Time* model, and the general one as the *SR and RD Fading Channels* model. As we shall see in Sec. IV, the quantitative trade-offs in this simplified model and the general model are similar. Thus, the simplified model is useful for understanding the role of p and q_{th} and fine-tuning their values.

The queue evolution at Relay i is

$$q_n(i) = (q_{n-1}(i) - X_n(i))^+ + Y_n(i), \qquad (2)$$

where $Y_n(i) \leq 1$ is the number of arrivals to $\mathcal{Q}_i$ at time t_n and $X_n(i)$ is the *potential service completions* at $\mathcal{Q}_i$ during time T_n.[4] Since $\{X_n(i)\}$ are i.i.d. and the RD transmission times are exponential RVs, $\{\mathbf{q}_n\}$ is a discrete time countable state Markov chain with transition matrix $\mathbf{P}$. Furthermore, the Markov chain is aperiodic since the probability of packet arrival at each queue is non-zero.

Theorem 1 provides the necessary and sufficient conditions required for $\{\mathbf{q}_n\}$ to be ergodic and have a unique stationary distribution, π. We now derive expressions for the throughput and mean E-E delay as a function of π. For this, we first explicitly charaterize $\mathbf{P}$, from which π is easily computed by solving $\pi\mathbf{P} = \pi$, such that $\sum_{\mathbf{m}} \pi(\mathbf{m}) = 1$, where $\mathbf{m} = (m_1, \ldots, m_M)$ and $m_i \geq 0$, for all $1 \leq i \leq M$.

1) Transition probabilites: Let $P(\mathbf{m}'|\mathbf{m})$ denote the transition probability from $\mathbf{q}_{n-1} = \mathbf{m}$ to $\mathbf{q}_n = \mathbf{m}'$, where $\mathbf{m}' = (m'_1, \ldots, m'_M)$. Then,

$$P(\mathbf{m}'|\mathbf{m}) = \sum_{i=1}^{M} \int_0^{\infty} P[R_i \text{ selected}, T_n = x|\mathbf{q}_{n-1} = \mathbf{m}] \times P[\mathbf{q}_n = \mathbf{m}'|\mathbf{q}_{n-1} = \mathbf{m}, R_i \text{ selected}, T_n = x]\, dx. \qquad (3)$$

Since the various RD channels are independent, we have

$$P[\mathbf{q}_n = \mathbf{m}'|\mathbf{q}_{n-1} = \mathbf{m}, R_i \text{ selected}, T_n = x] = \prod_{j=1}^{M} P\left[q_n(j) = m'_j|q_{n-1}(j) = m_j, R_i \text{ selected}, T_n = x\right].$$

Since the RD service times are exponential RVs, for $j \neq i$:

$$P\left[q_n(j) = m'_j|q_{n-1}(j) = m_j, R_i \text{ selected}, T_n = x\right] = \begin{cases} 0, & m'_j > m_j \\ 1, & m'_j = m_j = 0 \\ e^{-\mu x}\frac{(\mu x)^{m_j - m'_j}}{(m_j - m'_j)!}, & m_j \geq m'_j \geq 1 \\ 1 - \sum_{d=0}^{m_j - 1} e^{-\mu x}\frac{(\mu x)^d}{d!}, & m'_j = 0, m_j \geq 1 \end{cases}. \qquad (4)$$

[3]For simplicity, we will also assume that τ is aperiodic [18]. A sufficient condition for this is that the service times on the R_i to D link have a non-zero probability of being less than T_n. As we shall see, this condition will always be satisfied when the queues are stable.

[4]By potential service completions, we mean the number of service completions that would occur if there were an infinite number of packets in the queue. This ensures that $X_n(i)$ is independent of $q_n(i)$.

For $j = i$, we get

$$P\left[q_n(i) = m_i' | q_{n-1}(i) = m_i, R_i \text{ selected}, T_n = x\right]$$
$$= \begin{cases} 0, & m_i' = 0 \text{ or } m_i' > m_i + 1 \\ 1, & m_i' = 1, m_i = 0 \\ e^{-\mu x} \frac{(\mu x)^{m_i - m_i' + 1}}{(m_i - m_i' + 1)!}, & m_i \geq m_i' - 1 \geq 1 \\ 1 - \sum_{d=0}^{m_i - 1} e^{-\mu x} \frac{(\mu x)^d}{d!}, & m_i' = 1, m_i \geq 1 \end{cases}. \quad (5)$$

The term $P\left[R_i \text{ selected}, T_n = x | \mathbf{q}_{n-1} = \mathbf{m}\right]$ in (3), depends on the fading distribution. For Rayleigh fading on the SR channels, an analysis similar to that in [17] yields

$$\begin{aligned} &P\left[R_i \text{ selected}, T_n = x | \mathbf{q}_{n-1} = \mathbf{m}\right] \\ &\quad = \frac{\tilde{B}_{\text{nrm},i}}{\overline{\gamma}(i) x^2} \exp\left(\frac{\tilde{B}_{\text{nrm},i}}{x} + \frac{1 - e^{\tilde{B}_{\text{nrm},i}/x}}{\overline{\gamma}(i)} \right) \\ &\quad \times \sum_{k=1}^{2^{M-1}} (-1)^{|\omega_k|} \exp\left(\sum_{j \in \omega_k} \frac{1 - e^{\tilde{B}_{\text{nrm},j}/x}}{\overline{\gamma}(j)} \right), \end{aligned} \quad (6)$$

where $\tilde{B}_{\text{nrm},j} = B_{\text{nrm}}/(1-p)$, if $q_{n-1}(j) \geq q_{\text{th}}$, and $\tilde{B}_{\text{nrm},j} = B_{\text{nrm}}$, if $0 \leq q_{n-1}(j) < q_{\text{th}}$, ω_k is the kth subset of $\{1, \ldots, M\} \setminus \{i\}$, and $|\omega_k|$ is the number of elements in ω_k. Intuitively, the scaling of B_{nrm} by a factor $1-p$ occurs because ignoring each received symbol with probability $1-p$ increases the average time required to receive the packet by a factor $1/(1-p)$. The single integral in (3) is then numerically computed.

2) System Throughput: For a stable system with stationary distribution π, the throughput in packets/sec equals

$$\psi = \left(\sum_{\mathbf{m}} \pi(\mathbf{m}) \mathbf{E}\left[T_n | \mathbf{q}_{n-1} = \mathbf{m}\right] \right)^{-1}. \quad (7)$$

This follows from $\psi = 1/\mathbf{E}\left[T_n\right]$, and using the law of total expectation to simplify $\mathbf{E}\left[T_n\right]$. From (6), for Rayleigh fading on the SR channels, we get

$$\begin{aligned} &\mathbf{E}\left[T_n | \mathbf{q}_{n-1} = \mathbf{m}\right] \\ &\quad = \sum_{i=1}^{M} \int_0^\infty \frac{\tilde{B}_{\text{nrm},i}}{\overline{\gamma}(i) x} \exp\left(\frac{\tilde{B}_{\text{nrm},i}}{x} + \frac{1 - e^{\tilde{B}_{\text{nrm},i}/x}}{\overline{\gamma}(i)} \right) \\ &\quad \times \sum_{k=1}^{2^{M-1}} (-1)^{|\omega_k|} \exp\left(\sum_{j \in \omega_k} \frac{1 - e^{\tilde{B}_{\text{nrm},j}/x}}{\overline{\gamma}(j)} \right) dx, \end{aligned} \quad (8)$$

where $\tilde{B}_{\text{nrm},j}$, $1 \leq j \leq M$, and ω_k are as defined before.

In the unstable regime, $\psi = \sum_{i=1}^{M} \frac{1}{\mu}$. This is because, with probability 1, every relay always has at least packet in its unstable queue; R_i then transmits packets to D at rate $1/\mu$.

3) Mean End-to-End Delay: As mentioned, the mean E-E delay is bounded only when the system is stable and has a stationary distribution π. It is calculated as follows.

From symmetry, the mean waiting time of a packet, denoted by $\mathbf{E}\left[W_i\right]$, that includes the time taken by R_i to transmit the packet equals $\mathbf{E}\left[q_n(i)\right]/\mu$, where $\mathbf{E}\left[q_n(i)\right]$ is given by

$$\mathbf{E}\left[q_n(i)\right] = \sum_{i=1}^{M} \sum_{\mathbf{m}, \mathbf{m}'} m_i' P\left[\mathbf{q}_n = \mathbf{m}', R_i \text{ selected} | \mathbf{q}_{n-1} = \mathbf{m}\right] \pi(\mathbf{m}). \quad (9)$$

Here,

$$\begin{aligned} &P\left[\mathbf{q}_n = \mathbf{m}', R_i \text{ selected} | \mathbf{q}_{n-1} = \mathbf{m}\right] \\ &\quad = \int_0^\infty P\left[\mathbf{q}_n = \mathbf{m}', T_n = x, R_i \text{ selected} | \mathbf{q_{n-1}} = \mathbf{m}\right] dx, \\ &\quad = \int_0^\infty P\left[\mathbf{q}_n = \mathbf{m}' | T_n = x, R_i \text{ selected}, \mathbf{q_{n-1}} = \mathbf{m}\right] \\ &\qquad \times P\left[T_n = x, R_i \text{ selected}, \mathbf{q_{n-1}} = \mathbf{m}\right] dx. \end{aligned} \quad (10)$$

Both the terms in the integrand in (10) have already been computed in Sec. III-B1. Therefore, the mean E-E delay can be computed.

C. A Useful Simplifying Approximation

The analysis of the SR fading channel and RD exponential service time model above simplifies when we consider an approximate system in which the queue evolution of different relays is decoupled. This is done by focussing on a specific relay (say Relay 1) that ignores its received coded bit/symbol with probability p when $q_n(1) \geq q_{\text{th}}$, and *assuming that all the other $M-1$ relays never ignore their received signals. This decouples the queues of the various relays because, now, the arrival rate to a queue no longer depends on the state of the other queues.* This is advantageous because only a one-dimensional Markov chain whose state is $q_n(1)$ needs to be handled. This is unlike Sec. III-B1, which dealt with an M-dimensional chain whose state was $\mathbf{q}_n = (q_n(1), \ldots, q_n(M))$.

The transition probabilities, which determine the steady state probability $\pi_1^{\text{aprx}}(m)$ of $\mathcal{Q}_1$, now simplify as shown below. Let $P_1(m'|m)$ denote the transition probability from $q_{n-1}(1) = m$ to $q_n(1) = m'$. Then,

$$\begin{aligned} &P_1(m'|m) \\ &\quad = \sum_{i=1}^{M} \int_0^\infty P\left[q_n(1) = m' | q_{n-1} = m, R_i \text{ selected}, T_n = x\right] \\ &\qquad \times P\left[R_i \text{ selected}, T_n = x | q_{n-1} = m\right] dx. \end{aligned} \quad (11)$$

When $i \neq 1$, the first term in the integrand is the same as (4) (with $j = 1$). When $i = 1$, the first term is the same as (5) (with $j = 1$). The second term is the same as (6), except that $\tilde{B}_{\text{nrm},j} = B_{\text{nrm}}$, for all $j \neq 1$. Having obtained the steady state probability, the E-E delay and the throughput (using the approximation $\pi(\mathbf{m}) \approx \prod_i \pi_i^{\text{aprx}}(m_i)$ in (7)) can be evaluated.

IV. Simulation Results

We now plot the throughput and mean E-E delay as a function of the probability that a relay ignores a received bit, p, and the queue length threshold, q_{th}. We use $W = 2$ MHz, payload $B = 4096$ bits, $\delta = 0$ (ideal rateless code), and

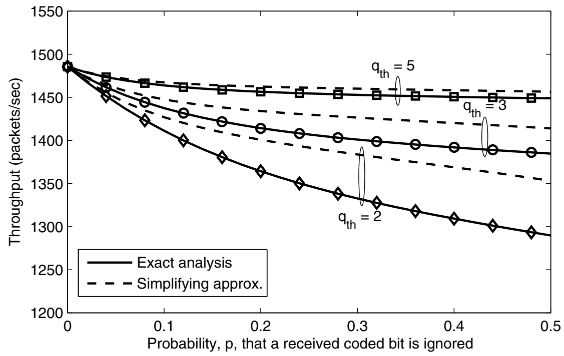

Fig. 2. Throughput as a function of p and q_{th} for the SR Fading Channel & RD Exponential Service Time model.

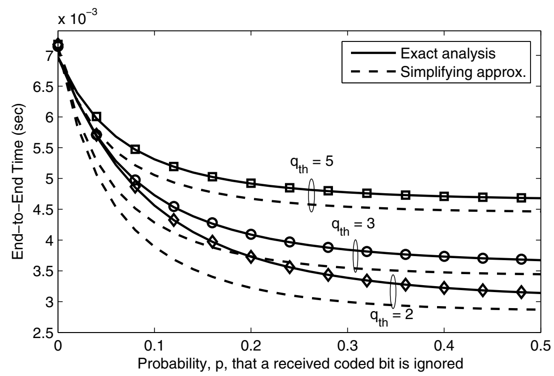

Fig. 3. End-To-End Delay as a function of p and q_{th} for the SR Fading Channel & RD Exponential Service Time model

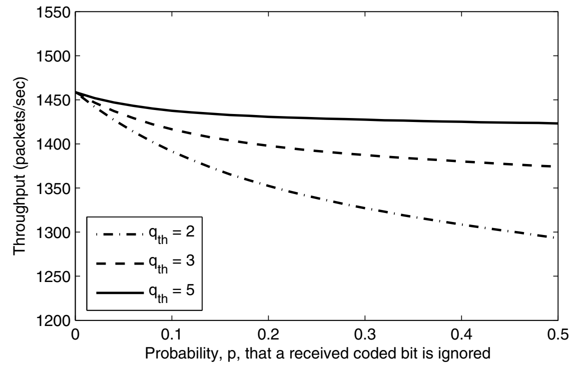

Fig. 4. Throughput as a function of p and q_{th} when the SR and RD channels undergo Rayleigh fading.

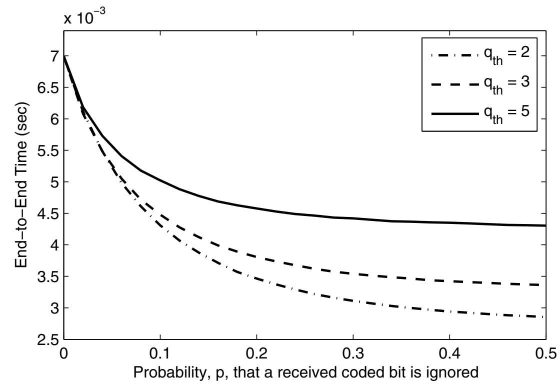

Fig. 5. E-E delay as a function of p and q_{th} when the SR and RD channels undergo Rayleigh fading.

$M = 3$ relays. We first study the behavior of the analytically tractable SR Fading Channel & RD Exponential Service Time model. Thereafter, we present simulation results for the model where the SR and RD channels both undergo Rayleigh fading. To enable a quantitative comparison, we set $\overline{\gamma}(i) = 17$ dB and the mean RD packet transmission time as $\mu = 1.6$ ms in both cases.

A. SR Fading Channel & RD Exponential Service Time Model

Figure 2 plots the throughput as a function of p for different values of q_{th}. The simulations results, which are shown using markers ($\diamond$, etc.), match the analysis curves (–) well. To determine π, the Markov state space was truncated, as is typically done. The throughput decreases as either p increases or q_{th} decreases since the source transmission time increases. This occurs because, for larger values of q_{th}, the odds that the number of packets in the queue exceeds q_{th} decreases and so does the probability that a relay accumulates mutual information at a reduced rate.

The corresponding E-E delay is shown in Figure 3. The simulation results, which are shown using markers ($\diamond$, etc.), again match the analysis curves (–) well. As p increases, the E-E delay decreases because the queue lengths and the queuing delays decrease. As expected, the smaller the value of q_{th}, the larger the decrease.

The two figures together show that the proposed load balancing method reduces the E-E delay substantially with only a marginal decrease in throughput. For example, at $p = 0.28$ and $q_{\text{th}} = 2$, the E-E delay decreases by 52%, while the throughput decreases by 10% compared to the case without load balancing ($p = 0$). The reduction in throughput is, in fact, a minor issue since introducing queuing at the relays itself improves throughput by 100% or more [17]. Also shown in both figures are results from the simplifying approximation of Sec. III-C, which are plotted using dashed lines (- -). It can be seen that it is reasonably accurate despite its considerable simplicity.

B. SR Fading and RD Fading Channels Model

We now study, using simulations, the more realistic scenario in which the RD links also undergo Rayleigh fading. We set $\overline{\lambda}(i) = 12.5$ dB so that the mean RD transmission time equals 1.6 ms, which is the same as for the RD exponential service time model studied in Section IV-A. The relay times out after 10 ms. (The effect of time-out on the mean SR and RD transmission times and stability is negligible since the probability that a packet is dropped by a relay is only 1.8%; the probability that the source times out is 0%.) As before, Figures 4 and 5 plot the throughput and E-E delay, respectively, as a function of p for different q_{th}.

The throughput and E-E delay marginally decrease compared to the SR Fading Channel & RD Exponential Service Time model studied in Sec. IV-A. Most notably, the behavior with respect to p and q_{th} is quantitatively very similar. Now, a 54.8% decrease in the E-E delay can be achieved with just a 8.7% reduction in throughput at $p = 0.28$ and $q_{\text{th}} = 2$. Thus, the analytically tractable model of Sec. III-B can be used to determined the optimal p and q_{th}.

V. Conclusions

We proposed a relay load balancing method in which a relay, depending on its queue state, autonomously reduces the rate at which it receives packets. It does this by randomly ignoring a fraction of the coded bits/symbols that arrive at its receiver. Doing so increases the odds that the packet will enter the queue of a relay with a smaller queue. The method exploits the fact that a rateless code receiver can decode the packet without having access to some portions of the transmitted signal, so long as it has accumulated sufficient mutual information about the packet. Over a wide range of operating points, the system fully reaped the benefits of spatial diversity provided by multiple relays; its throughput decreased marginally while the end-to-end delays decreased substantially. While the ideas in this paper consider a basic two-hop network, they are useful for multihop networks as well. Another interesting problem is deriving tight upper and lower bounds for the general model in which source-to-relay and relay-to-destination channels undergo Rayleigh fading.

Appendix

A. Proof of Theorem 1

First consider the case when the system is unstable. Then, $\lim_{n\to\infty} P\left[\sum_i q_n(i) \geq M(q_{\text{th}}+1)\right] = 1$. Thus, $\lim_{n\to\infty} P[q_n(i) \geq q_{\text{th}}+1, \text{ for some } i] = 1$. By symmetry,

$$\lim_{n\to\infty} P\left[\bigcap_i \{q_n(i) \geq q_{\text{th}}+1\}\right] \geq 1 - \lim_{n\to\infty} \sum_i P\left[q_n(i) \leq q_{\text{th}}\right]. \tag{12}$$

This implies that for any given $\epsilon > 0$, for all large n,

$$P\left[\bigcap_i \{q_n(i) \geq q_{\text{th}}+1\}\right] \geq 1 - \epsilon. \tag{13}$$

Thus, the mean service completion time of the source for all large n exceeds $(1-\epsilon)\mathbf{E}\left[\frac{B}{r_n^{\max}(1-p)}\right] + \epsilon\mathbf{E}\left[\frac{B}{r_n^{\max}}\right]$, where $r_n^{\max} = \max_i r_n(i)$.[5]

Also, from (12), the probability that any of the queues is empty is less than ϵ. Hence, by symmetry, the mean inter-arrival time of $\mathcal{Q}_i$, for any i, is greater than $M(1-\epsilon)\mathbf{E}\left[\frac{B}{r_n^{\max}(1-p)}\right] + \epsilon M\mathbf{E}\left[\frac{B}{r_n^{\max}}\right]$ and the mean inter-departure time is arbitrarily close to $\mathbf{E}\left[\frac{B}{r'_n(i)}\right]$. Thus, $\mathcal{Q}_i$ is unstable only if $\frac{M}{1-p}\mathbf{E}\left[\frac{B}{r_n^{\max}}\right] \leq \mathbf{E}\left[\frac{B}{r'_n(i)}\right]$, which occurs only if (1) is violated. This proves that (1) is a sufficient condition.

Next we show that if the system is stable then (1) holds. Consider a centralized multiserver queueing system where all the packets leaving the source enter the same queue and the departure process from the source is the same as the original process. Also, each R_i to D link serves a separate packet in the queue, i.e., if there are $l < M$ packets in the queue then exactly l RD links will be busy. The total queue length in this system stochastically lower bounds the total number of packets in all the relay queues in our system. For the stability of the centralized system – and thus our sytsem – its mean inter-arrival time must exceed the mean service time of packets in the queue. A necessary condition for this is (1).

[5] Although the condition $\lim_{n\to\infty} P\left[\sum_i \mathbf{q}_n(i) > x\right] = 1$, for any x is asserted in probability, it often happens almost surely (a.s.). This occurs, for example, in the case of a Markov chain when the chain is transient. When this is so, (13) also holds a.s., and the arguments that follow in this proof go through directly. On the other hand, when the condition above holds only in probability, then (13) and the other arguments in this proof can be asserted for any finite sequence $\mathbf{q}_n, \mathbf{q}_{n+1}, \ldots, \mathbf{q}_{n+m}$. Consequently, the arguments that follow still go through with arbitrarily small inaccuracy.

References

[1] Z. Zhou, S. Zhou, J.-H. Cui, and S. Cui, "Energy-efficient cooperative communication based on power control and selective single-relay in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 3066–3078, Aug. 2008.

[2] D. S. Michalopoulos and G. K. Karagiannidis, "Performance analysis of single relay selection in Rayleigh fading," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 3718–3724, Oct. 2008.

[3] A. Stefanov and E. Erkip, "Cooperative coding for wireless networks," *IEEE Trans. Commun.*, vol. 52, pp. 1470–1476, Sept. 2004.

[4] I. Maric and R. D. Yates, "Cooperative multihop broadcast for wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 22, pp. 1080–1088, Aug. 2004.

[5] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE J. Sel. Areas Commun.*, vol. 24, pp. 659–672, Mar. 2006.

[6] R. Madan, N. B. Mehta, A. F. Molisch, and J. Zhang, "Energy-efficient cooperative relaying over fading channels with simple relay selection," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 3013–3025, Aug. 2008.

[7] A. F. Molisch, N. B. Mehta, J. Yedidia, and J. Zhang, "Performance of fountain codes in collaborative relay networks," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 4108–4119, Nov. 2007.

[8] J. Castura and Y. Mao, "Rateless coding for wireless relay channels," in *Proc. IEEE Int. Symp. Inf. Theory*, pp. 810–814, 2005.

[9] A. Eckford, J. Chu, and R. Adve, "Low-complexity cooperative coding for sensor networks using rateless and LDGM codes," in *Proc. ICC*, pp. 1537–1542, Jun. 2006.

[10] R. Nikjah and N. C. Beaulieu, "Achievable rates and fairness in rateless coded decode-and-forward half-duplex and full-duplex opportunistic relaying," in *Proc. ICC*, pp. 3701–3707, 2008.

[11] M. Mitzenmacher, "Digital fountains: A survey and look forward," in *Proc. IEEE Inf. Theory Workshop*, pp. 271–276, Oct. 2004.

[12] U. Erez, M. D. Trott, and G. W. Wornell, "Rateless coding for Gaussian channels," *Submitted to the IEEE Trans. Inf. Theory*, 2007.

[13] S. C. Draper, B. Frey, and F. R. Kschischang, "Rateless coding for non-ergodic channels with decoder channel state information," *Submitted to IEEE Trans. Inf. Theory*, 2008.

[14] A. Shokrollahi, "Raptor codes," in *Proc. IEEE Int. Symp. Inf. Theory*, p. 36, 2004.

[15] R. Palanki and J. Yedidia, "Rateless codes on noisy channels," in *Proc. IEEE Int. Symp. Inf. Theory*, p. 37, 2004.

[16] Y. Fan, L. Lai, E. Erkip, and H. V. Poor, "Rateless coding for MIMO block fading channels," in *Proc. IEEE Int. Symp. Inf. Theory*, pp. 2252–2256, 2008.

[17] N. B. Mehta, V. Sharma, and G. Bansal, "Queued cooperative wireless networks with rateless codes," *Proc. Globecom*, 2008.

[18] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Prentice Hall, 1989.