

Queue-Architecture and Stability Analysis in Cooperative Relay Networks

Jubin Jose and Sriram Vishwanath

Dept. of Electrical and Computer Engineering

University of Texas at Austin

Email: {jubin, sriram}@austin.utexas.edu

Abstract

An abstraction of the physical layer coding using bit pipes that are coupled through data-rates is insufficient to capture notions such as node cooperation in cooperative relay networks. Consequently, network-stability analyses based on such abstractions are valid for non-cooperative schemes alone and meaningless for cooperative schemes. Motivated from this, this paper develops a framework that brings the information-theoretic coding scheme together with network-stability analysis. This framework does not constrain the system to any particular achievable scheme, i.e., the relays can use any cooperative coding strategy of its choice, be it amplify/compress/quantize or any alter-and-forward scheme. The paper focuses on the scenario when coherence duration is of the same order of the packet/codeword duration, the channel distribution is unknown and the fading state is only known causally. The main contributions of this paper are two-fold: first, it develops a low-complexity queue-architecture to enable stable operation of cooperative relay networks, and, second, it establishes the throughput optimality of a simple network algorithm that utilizes this queue-architecture.

Index Terms

Cooperative relay networking, Network algorithm, Stability analysis

I. INTRODUCTION

Cooperative relaying is traditionally seen as a physical layer scheme for analyzing and designing wireless link layer protocols [1], with limited network-layer insights originating from such schemes. Indeed, the not-so-uncommon perception is: whatever be the physical layer transmission/coding scheme, the network can abstract it into a “rate region” and then determine

algorithms to stabilize queues, perform rate control and other tasks at the higher layers. From this perspective, it seems unimportant for researchers at either layer to learn much about the intricacies of the other.

There is a significant and growing body of work suggesting that such abstractions may not be accurate [2] and that physical layer parameters must be included into the analysis. A large class of this work is based on signal-to-noise ratio (SNR) or signal-to-interference-and-noise ratio (SINR) models for the physical medium. While S(I)NR is a worthwhile abstraction for physical-layer schemes that “treat interference as noise”, it is often overused and does not capture more involved physical layer transmission schemes [3]. From information theory, it is well known that “treating interference as noise” represents a very limited class of transmission schemes, and a much larger class of schemes exist that achieve significantly higher throughput. Therefore, a framework that brings the information-theoretic coding scheme together with network-stability analysis is needed, to bridge the gap caused by the “unconsummated union” [4]. In this paper, we explore building this bridge in the context of cooperative relay networks.

We emphasize that a natural separation between network stability and physical layer coding exists only for specific classes of networks (such as capacitated networks [5]) and not in general, and a joint framework is needed that can capture notions such as physical layer cooperation. In this paper, we focus on cooperative relay networks, where multiple reasons exist for combining network and physical layer aspects.

- First, the rate-maximizing physical-layer coding strategy automatically imposes scheduling restrictions on the relays/transmitters in the network. For coherent combination at the receivers to be at all possible, all nodes involved must transmit simultaneously in that block.
- Second, it is codebooks and functions of codebooks being received, stored and transmitted by nodes and not traditional data packets.
- Finally, the codebook chosen by the source(s) determines the rate of transmission, which may or may not be alterable at intermediate nodes (this is a key distinction between general information-theoretic coding theorems and say, packetized or linear network coded systems where rate can always be varied at every node). For example, if a relay were to use amplify-and-forward or compress-and-forward as its physical-layer strategies, it has no control over rate and has a real vector as its “packet”.

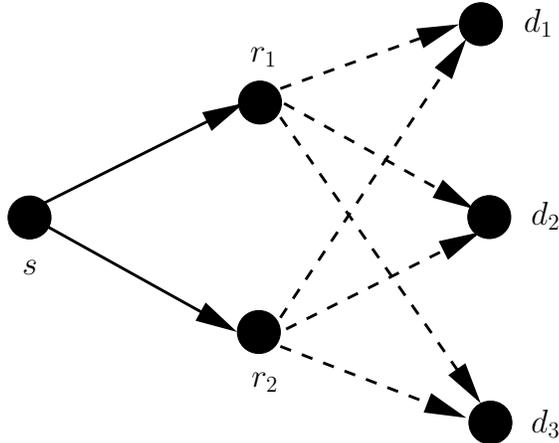


Fig. 1. Two-hop Cooperative Network

Given the need for a joint physical and network layer framework for cooperative networks, the rest of the paper is organized as follows: in the next section, we present a brief summary of cooperative relay networking from a physical layer perspective. In Section III, we present our main results in this paper. In Section IV, we describe our system model in the context of heterogeneous cellular networks. In Section V, we describe cooperative schemes for such networks in detail and present a queue-architecture that enables both efficient and optimal operation of the network. In Section VI, we present the main algorithm for operating such networks, and establish that this algorithm is throughput-optimal. We conclude with Section VII.

II. BACKGROUND: COOPERATIVE RELAY NETWORKS

Cooperative relay networks have been researched extensively since the “MIMO effect” was established. Until recently, it was considered hard if not impractical for nodes to coordinate transmissions to enable cooperative relaying. However, emerging heterogeneous cellular networks are increasingly moving in the direction of standardizing and evaluating schemes with node cooperation [6], [7]. As cell sizes decrease, an increase in cell edges and interference requires node cooperation to increase throughput, and cooperative relaying is an important step in making this happen.

Figure 1 shows the most basic configuration that incorporates cooperative relaying in heterogeneous cellular networks. To motivate this setting, we take the example of a macrocellular

network. Here, the source node s corresponds to the macro-cell base-station, the relay nodes r_1 and r_2 correspond to pico-cell base-stations and the destination nodes d_1 , d_2 and d_3 correspond to mobiles. We focus on the downlink scenario where the source s has independent messages/bits for the mobiles. The relays' role is to help the source in transmitting these messages. Further, we assume a half-duplex cooperative constraint so that either the first-hop or the second-hop links can be activated at any given time, with no direct-links from the source to the destinations. A more general and detailed system model for such cooperative relay networks is provided in Section IV.

Even for such simple networks with two relays and one destination and fixed channels, information-theoretic capacity is not yet known. However, there has been significant progress in developing cooperative communication schemes for such systems by using coherence and physical-layer coordination among nodes. There are multiple strategies studied in literature that enable this coordination, referred to as *forwarding* schemes. One such scheme of interest is the so-called decode-and-forward scheme that requires relays to decode messages. In contrast to traditional networks, the relays decode common messages, that are then transmitted cooperatively. However, the relays still have decoded messages or packets as in traditional networks. In [8], the authors develop a throughput-optimal network algorithm that can handle common messages. In [9], the authors consider more general network configurations, but the applicability is still limited to decode-and-forward schemes with fixed channels. In essence, all of these apply only in packet-in-packet-out networks. Complimentary to this is the work on optimal resource allocation for non-cooperative wireless networks [10]–[12] (and references therein).

In our effort, we do not want to constrain the system to a packet-in-packet-out framework. We desire that the relays use *any* information-theoretic cooperative coding strategy of its choice, be it amplify/compress/quantize or any alter-and-forward scheme. This couples coding, resource allocation and stability into one joint problem, and the analyses in [8], [10], [12] and the vast literature on non-cooperative networks do not apply. Even the analyses in [8], [9] for decode-and-forward cooperative networks do not apply. This motivates the need for a new framework and stability analysis.

Before proceeding to describe our results, a note to state the obvious: if the channel state is fixed and thus its capacity is precomputed, a simple static split scheme will ensure stable operation while maximizing the information theoretic rate (region) for the network. The chal-

lenge, of course, is when the fading state distribution and input arrival rates are unknown, and the fading state can only be observed causally. For example, consider a fading channel with block fading of T symbols each. When T is much smaller compared to the packet duration (or equivalently the channel-coding duration), queueing/buffering of packets at relays is not required as the first-hop and second-hop can be operated sequentially without reducing data-rates. When T is comparable to (or larger than) the packet duration, queueing of packets at relays can provide significant gains in terms of data-rates. Furthermore, when T is roughly the same as the packet duration, queueing at relays is inevitable as the source does not know the fading state of the second-hop while encoding the packet. In this paper, we focus on the second scenario when T is larger than the packet/codeword duration. Given that the channel distribution is unknown and the fading state is only known causally, we ask the question: Is it possible to stabilize the network while operating it close to the boundary of its information-theoretic rate region?

III. MAIN RESULTS

The answer to the preceding question in Section II is “yes”, which is proved for a simpler network with two relays and one destination in [13]. In this setting, for cooperative schemes such as amplify/quantize-and-forward and partial-decode-and-forward, the relays receive and transmit real-valued “packets”. In order to accomplish this in [13], we introduce a new “state-based” virtual-queue-architecture for these real-valued “packets”, and develop a throughput-optimal network algorithm that does not require the knowledge of the fading distribution. Each “state” corresponds to a vector comprised of the *entire* channel-state of each link in the network. This approach, although analytically very helpful, suffers from a major issue that makes it practically uninteresting - requiring that a virtual-queue be maintained for each channel-state at each node in the network leads to an explosion of queues, even for simple network configurations. Moreover, the approach in [13] is particular to a single destination setting. In this paper, we develop a simpler queue-architecture to enable stable operation of cooperative relay networks. Further, we generalize it to any forwarding scheme with multiple destinations.

The virtual-queue-architecture we introduce in this paper is primarily *encoding-based*. This architecture is motivated by the manner in which adaptive modulation and coding is currently implemented in practice. In systems today, the source node implements a limited number of encoding schemes (encoding functions and rate-vectors). Each encoding scheme is designed so

that it can be successfully employed for a particular subset of states. Even though encoding schemes belong to a finite (and usually small) set, the mapping functions at the relays and the decoding functions at the destinations are usually state-dependent. A queue-architecture that keeps virtual queues at the relays for each state corresponding to the first-hop and each encoding scheme is sufficient. This considerably reduces the number of virtual queues that must be maintained while still remaining a “sufficient statistic”, i.e., these encoding-based queues are a sufficiently rich representation for us to develop throughput optimal algorithms using them. Using this new and somewhat intuitive virtual-queue-architecture, we develop a network algorithm that has the following properties.

- 1) It does not require the knowledge of the fading distribution.
- 2) It does not require the knowledge of the arrival rates.
- 3) It keeps all the queues stable for any arrival rate-vector within the throughput region, i.e., it is throughput-optimal.

Note that limiting ourselves to a small set of possible encoding schemes and rates inherently reduces the network’s information-theoretic rate region. The more fine-grained the encoding schemes and resulting queue-architecture, the smaller the loss in rate region. However, note that the encoding-based queue-architecture itself does not introduce any sub-optimality.

In summary, we introduce and study a new encoding-based queue-architecture, which is inspired by an adaptive coded modulation system analyzed and implemented at the physical layer in systems today. However, in today’s systems, there is limited interaction, if any, between network-layer algorithms and adaptive coding/modulation, and we argue that coupling them together can be very useful in both the analysis and design of cooperative relay networks. Indeed, we show that such a queuing architecture can result in throughput optimal algorithms, and the network can achieve its information-theoretic rate region corresponding to its choice of encoding/decoding strategies while maintaining stability.

IV. SYSTEM MODEL

We consider discrete-time two-hop cooperative networks that include the network shown in Figure 1. We allow for arbitrary number of relays and destinations, i.e, the network consists of a source node denoted by s , N relay nodes denoted by r_1, r_2, \dots, r_N , and K destination nodes denoted by d_1, d_2, \dots, d_K . The source has independent messages for all the destinations.

The relays aid in transmitting these messages to their respective destinations. Throughout this paper, “first-hop” refers to the links from the source to the relays, and “second-hop” refers to the links from the relays to the destinations. At any given time, half-duplex and cooperative-communication constraints require that either the first-hop or the second-hop can be activated and not both. The presence of direct links from source to destinations will not invalidate the analysis presented in this paper, but would render it considerably harder. For simplicity, we assume that they are absent and thus concentrate on equal-path length networks.

The channel model does not directly impact the queue-architecture, and thus the network algorithm and stability analysis presented in this paper. The channel is state dependent, and the joint-state distribution be unknown. A particular channel model of interest is a linear interaction model with additive white Gaussian noise (AWGN). In the context of an AWGN channel, an example of state is a multiplicative fading parameter. We focus on a framework with i.i.d. block-fading model with a block-length of T symbols in the remainder of this paper. The channels remain constant for the duration of one block, and then change to a new (independent) realization from an underlying distribution from block to block. Let $t \in \mathbb{Z}_+$ denote the channel fading blocks, and let \mathcal{F} denote the fading state-space, which is assumed to be discrete. In block t , $\mathbf{f}_1[t] \in \mathcal{F}^N$ denotes the fading realization for the first-hop and $\mathbf{f}_2[t] \in \mathcal{F}^{NK}$ denotes the fading realization for the second-hop. The combined fading-state is denoted by $\mathbf{f}[t] = (\mathbf{f}_1[t], \mathbf{f}_2[t])$. The corresponding random vectors are denoted by $\mathbf{F}_1[t]$, $\mathbf{F}_2[t]$ and $\mathbf{F}[t]$. Note that $\mathbf{F}[t]$ is i.i.d. over time, but can be spatially correlated. Let the probability that $\mathbf{F}[t]$ takes value \mathbf{f} be $\pi_{\mathbf{f}}$. This is the underlying probability distribution that is unknown to the central controller.

Next, we explain the time-scales in which network and channel parameters evolve in our system. The coherence time T is assumed to be comparable to the channel-coding length in symbols. For the ease of presentation, the “packet” (which is either the channel codeword or any real-vector representing the actual data packet) length is assumed to be equal to the coherence time T . It is straightforward to extend the analysis when the “packet” length is a sub-multiple of the coherence time T . Each “packet” is transmitted on the first-hop and the second-hop exactly once. These transmissions need not happen in consecutive time-blocks, i.e., these “packets” can be buffered at the relays. The coding performed at the source, the mappings performed at the relays, and the decoding at the destinations can be arbitrary, i.e., this includes any and all schemes that are information-theoretically capacity-optimal or, if capacity is unknown, then the best known

coding scheme. Further, we assume that the instantaneous fading-state is causally known globally to the central controller. In other words, prior to transmission, the central controller is aware of the entire network channel state for that particular time-block.

A. Notation

Vectors are denoted by bold letters. For vectors, equality and inequality operators are defined component-wise. $\mathbf{a} \cdot \mathbf{b}$ denotes the dot product of \mathbf{a} and \mathbf{b} . $|\cdot|$ denotes the cardinality of a set. $\mathbf{1}_{\{E\}}$ denotes the indicator function of event E . $(a)^+$ denotes $\max(a, 0)$. $\mathbb{E}[\cdot]$ denotes the expectation operator.

V. ACHIEVABLE RATES & QUEUE-ARCHITECTURE

The notion of a “packet” here is different from traditional networks where a packet is decoded at all intermediate relays, and is usually meant for one destination. In this paper, the term “packet” refers to the set of coded symbols transmitted/received in the network. Note that each of the relays receives a different noisy version of the transmitted vector (transmitted “packet”), which is subsequently mapped to a transmit vector (“packet”) at each relay. Again, the destinations receive a noisy version of a linear combination of relays’ transmit “packets”. In this paper, we refer to the physical-layer signalling vectors as *packets* at each node in the network. We choose to use this language as the entire network layer analysis is based on understanding the dynamics of these transmit vectors as they traverse the system. Consider a packet that is transmitted from the source to the K destinations. Let this packet be transmitted on the first-hop during block t_1 , and be transmitted on the second-hop during block t_2 . Then, $\mathbf{g} = (\mathbf{f}_1[t_1], \mathbf{f}_2[t_2])$ is said to be the “state” seen by this packet. Note that this notion of state is different from physical channel fading state, but is it of equal importance in our analysis.

A packet transmitted by the source is received by all the destinations in two hops, but the amount of information each destination receives varies depending on the encoding rates. Given a state seen by the packet, the set of encoding rates that can be supported is known as the rate region for the given state. An extremely challenging problem even in the single destination setting is to find the set of all achievable rates, or the capacity region for the given state. Even though the capacity region is unknown in most cases, there are many efficient cooperative communication schemes that have been developed. Therefore, the main aims of this paper are: (i) to develop a

queue-architecture that can support existing (and future) cooperative schemes, and (ii) to develop a throughput-optimal network algorithm using this queue-architecture.

The queue-architecture developed in [13] for single-destination setting keeps “virtual” queues at relays for every state. Suppose that each rate-region can be quantized such that the convex-hull of the set of quantized rate-vectors is “nearly” same as the rate-region itself. Further, let us assume that the rate corresponding to each destination have to be quantized to L levels. Now, a direct extension of the state-based virtual-queue-architecture would require “virtual” queues at relays for each state and each quantized rate-vector, which results in $L^K |\mathcal{F}|^{K(N+1)}$ “virtual” queues. This scales exponentially in the number of destinations K . Clearly, such a queue-architecture is not scalable in practice, and will face implementation issues.

In order to design a low-complexity queue-architecture, we exploit the fact that practical systems implement limited number of encoding schemes, as in the case of adaptive modulation and coding. For example, the source might choose to encode only two destinations at a time using superposition encoding. In this case, the total number of encoding schemes would be $K(K-1)L^2$. In another example, the source might choose to encode at limited boundary rate-vectors again with superposition encoding. Let \mathcal{M} denote the set of encoding schemes, and \mathbf{r}_m denote the rate-vector corresponding to each encoding scheme $m \in \mathcal{M}$. Given that $|\mathcal{M}| \ll L^K |\mathcal{F}|^{KN}$, a queue-architecture needs to support these limited choices. While a queue-architecture can take advantage of this, it needs to allow for arbitrary mapping at the relays and decoding at the destinations. These are usually state-dependent, for example, an amplify-and-forward mapping is state-dependent.

Before describing our queue-architecture, we characterize the throughput region of the two-hop cooperative network. For this, we assume the knowledge of the fading distribution. Define $\mathcal{I} = \{(m, \mathbf{g}) | m \in \mathcal{M} \text{ can be supported by state } \mathbf{g} \in \mathcal{F}^{(N+1)K}\}$, which represents whether an encoding scheme is supported by a state or not¹. Now, let $\mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2)$ be any fading-state where \mathbf{f}_1 is the fading-state of first-hop and \mathbf{f}_2 is the fading-state of second-hop. Similarly, let $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2)$ by any state. We define $\hat{\mathcal{F}} = \mathcal{F}^{(N+1)K}$, $\mathcal{I}_1 = \{(\mathbf{f}, \mathbf{g}) | \mathbf{g}_1 = \mathbf{f}_1\}$, and $\mathcal{I}_2 = \{(\mathbf{f}, \mathbf{g}) | \mathbf{g}_2 = \mathbf{f}_2\}$. With the above definitions, the throughput region of the network is characterized in the following

¹We do not explicitly deal with packet error rate, as it is assumed that the achievable rate-vector is defined appropriately with required packet error rate.

lemma.

Lemma 1: A rate-vector $\hat{\mathbf{r}}$ is in the throughput region denoted by \mathcal{T} only if there exists $a_{\mathbf{f}}^{m,\mathbf{g}} \geq 0$ and $b_{\mathbf{f}}^{m,\mathbf{g}} \geq 0$ for all $m \in \mathcal{M}$, $\mathbf{g} \in \hat{\mathcal{F}}$ and $\mathbf{f} \in \hat{\mathcal{F}}$ such that

$$\hat{\mathbf{r}} = \sum_{m,\mathbf{g},\mathbf{f}} \left(\pi_{\mathbf{f}} a_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{r}_m \mathbf{1}_{\{(\mathbf{f},\mathbf{g}) \in \mathcal{I}_1\}} \mathbf{1}_{\{(m,\mathbf{g}) \in \mathcal{I}\}} \right), \quad (1)$$

$$\sum_{\mathbf{f} \in \hat{\mathcal{F}}} \pi_{\mathbf{f}} a_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{1}_{\{(\mathbf{f},\mathbf{g}) \in \mathcal{I}_1\}} = \sum_{\mathbf{f} \in \hat{\mathcal{F}}} \pi_{\mathbf{f}} b_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{1}_{\{(\mathbf{f},\mathbf{g}) \in \mathcal{I}_2\}}, \forall (m, \mathbf{g}) \in \mathcal{I}, \quad (2)$$

$$\sum_{m,\mathbf{g}} a_{\mathbf{f}}^{m,\mathbf{g}} + b_{\mathbf{f}}^{m,\mathbf{g}} \leq 1, \forall \mathbf{f}. \quad (3)$$

Proof: Let $a_{\mathbf{f}}^{m,\mathbf{g}}$ be the fraction of time for which packets corresponding to encoding scheme m and state \mathbf{g} is transmitted from the source to the relays when the system is in fading state \mathbf{f} . Similarly, let $b_{\mathbf{f}}^{m,\mathbf{g}}$ be the fraction of time for which these packets are transmitted from the relays to the destinations. (1) is flow conservation constraint for the source, and (2) is the flow conservation constraint for each encoding scheme and state. (3) is the time conservation constraint for each fading-state. A central controller with the knowledge of the fading distribution can achieve these rates using static time-division. ■

An immediate corollary of this lemma is the following.

Corollary 2: The throughput region \mathcal{T} is convex.

Encoding-based Queue-architecture: At the source node s , there are K queues consisting of bits (or data) corresponding to the K destinations. We denote the queue at the source corresponding to k -th destination by Q_s^k with queue-length $Q_s^k[t]$ during block t . There is an exogenous i.i.d. arrival process $A^k[t]$ of data-bits into Q_s^k with mean rate $\lambda_k T$ bits/block and bounded variance. The vector of arrival rates λ_k is denoted by $\boldsymbol{\lambda}$. At each relay (say n), we keep virtual queues corresponding to each encoding scheme m and each fading state for the first-hop \mathbf{g}_1 denoted Q_n^{m,\mathbf{g}_1} with queue-length $Q_n^{m,\mathbf{g}_1}[t]$ during block t . This queue consists of real-valued packets encoded at rate \mathbf{r}_m . Since we keep virtual queues for each fading state corresponding to the first-hop, the mapping function performed at the relays can be a function of the fading state. Similarly, the decoding function can be a function of the fading state. With this queue-architecture, the number of virtual queues at each relay is $|\mathcal{M}| |\mathcal{F}|^N$. This is considerably less compared to the number of virtual queues required in the state-based approach, and thus provides a low-complexity queue-architecture. Note that the gain is high in the setting when the

number of destinations are large and number of relays are small, which is the case in cellular systems.

The queue dynamics is as follows: During block t , if the fading state for the first-hop is \mathbf{g}_1 and if the central controller decides that the source should transmit a packet using encoding scheme m , then the following queues get updated:

$$Q_s^k[t+1] = (Q_s^k[t] + A^k[t] - r_m^k T)^+, \forall k, \quad (4)$$

$$Q_n^{m, \mathbf{g}_1}[t+1] = Q_n^{m, \mathbf{g}_1}[t] + T, \forall n. \quad (5)$$

During block t , if the fading state for the second-hop is \mathbf{g}_2 , then the central controller can decide to transmit packets from queues $Q_n^{m, \mathbf{g}_1}, \forall n$ for some given m and \mathbf{g}_1 only if $(m, \mathbf{g}_1, \mathbf{g}_2) \in \mathcal{I}$. This ensures that the packet is received successfully at all the destinations. In this case, the following queues get updated:

$$Q_s^k[t+1] = Q_s^k[t] + A^k[t], \forall k, \quad (6)$$

$$Q_n^{m, \mathbf{g}_1}[t+1] = (Q_n^{m, \mathbf{g}_1}[t] - T)^+, \forall n. \quad (7)$$

Next, we address the question of designing a central controller that does not have the knowledge of the arrival rates or the fading state distribution.

VI. THROUGHPUT-OPTIMAL NETWORK ALGORITHM

In this section, we show that a throughput-optimal central controller can be designed without the knowledge of the arrival rates or the fading state distribution. Since cooperative schemes require strong node coordination, the centralized nature of the algorithm does not create additional system requirements. The following algorithm is motivated from back-pressure based Max-Weight algorithms for non-cooperative networks.

Back-pressure-based Algorithm: In every block, the central controller makes decisions based on the current fading state of the system and the current queue-lengths. Let the fading-state during block t be $\mathbf{f}[t] = (\mathbf{f}_1, \mathbf{f}_2)$. The network algorithm run by the controller is as follows:

- 1) It computes

$$A = \max_m \sum_k \left(Q_s^k[t] - r_m^k \sum_{n=1}^N Q_n^{m, \mathbf{f}_1}[t] \right) r_m^k$$

and an optimal parameter m^* for this problem.

2) It computes

$$B = \max_{m, \mathbf{g}_1} (\mathbf{r}_m \cdot \mathbf{1})^2 \sum_{n=1}^N Q_n^{m, \mathbf{g}_1}[t],$$

$$\text{s.t. } (m, (\mathbf{g}_1, \mathbf{f}_2)) \in \mathcal{I},$$

and a set of optimal parameters \hat{m} and $\hat{\mathbf{g}}_1$ for this problem.

- 3) If $A \geq B$, then the central controller decides to transmit a packet from the source to the relays using encoding scheme m^* .
- 4) Otherwise, the central controller decides to transmit a packet from queues $Q_n^{\hat{m}, \hat{\mathbf{g}}_1}, \forall n$, i.e., from the relays to the destinations.

The controller repeats steps 1 – 4 in every block.

The following theorem provides a strong theoretical guarantee on the throughput performance of this algorithm.

Theorem 3: The above algorithm stochastically stabilizes all the queues for any λ if there exists $\epsilon > 0$ such that $\lambda + \epsilon \mathbf{1}$ is within the throughput region given in Lemma 1, i.e., the underlying network Markov chain is positive recurrent. In simple terms, the algorithm is throughput-optimal.

Before proceeding to the proof of this theorem, we state the following lemma that is used in the proof.

Lemma 4: Suppose that there exists $\epsilon > 0$ such that $\lambda + \epsilon \mathbf{1}$ is within the throughput region. Then, there exists $a_{\mathbf{f}}^{m, \mathbf{g}} \geq 0$, $b_{\mathbf{f}}^{m, \mathbf{g}} \geq 0$ and $\delta > 0$ such that the following set of conditions are satisfied:

$$\lambda_k - \sum_{m, \mathbf{g}, \mathbf{f}} (\pi_{\mathbf{f}} r_m^k a_{\mathbf{f}}^{m, \mathbf{g}}) \leq -\delta, \forall k,$$

$$\sum_{\mathbf{f}} \pi_{\mathbf{f}} (a_{\mathbf{f}}^{m, \mathbf{g}} - b_{\mathbf{f}}^{m, \mathbf{g}}) \leq -\delta, \forall m, \mathbf{g},$$

$$\sum_{m, \mathbf{g}} a_{\mathbf{f}}^{m, \mathbf{g}} + b_{\mathbf{f}}^{m, \mathbf{g}} \leq 1, \forall \mathbf{f},$$

$$a_{\mathbf{f}}^{m, \mathbf{g}} = 0, \forall (\mathbf{f}, \mathbf{g}) \notin \mathcal{I}_1, \forall (m, \mathbf{g}) \notin \mathcal{I},$$

$$b_{\mathbf{f}}^{m, \mathbf{g}} = 0, \forall (\mathbf{f}, \mathbf{g}) \notin \mathcal{I}_2, \forall (m, \mathbf{g}) \notin \mathcal{I}.$$

Proof: The proof of this lemma is fairly straightforward, and is omitted for brevity. ■

A. Proof of Theorem 3

Since the queues form a Markov chain, we use Foster-Lyapunov theorem in order to prove the stability [14], [15]. Without loss of generality, we assume that $\mathbf{r}_m \neq \mathbf{0}, \forall m$. Otherwise, those queues at the relays can be removed without affecting the throughput region and the stability of the system. Now, consider the Lyapunov function

$$V(\mathbf{Q}[t]) = \sum_k (Q_s^k[t])^2 + \sum_{n=1}^N \sum_{m, \mathbf{g}_1} (\mathbf{r}_m \cdot \mathbf{1} Q_n^{m, \mathbf{g}_1}[t])^2,$$

where $\mathbf{Q}[t]$ denotes the vector of all queue lengths.

Next, we consider an optimization problem that captures the algorithm given in this section. Consider a fading-state \mathbf{f} and the following discrete optimization problem:

$$\begin{aligned} \max_{\alpha_{\mathbf{f}}^{m, \mathbf{g}}, \beta_{\mathbf{f}}^{m, \mathbf{g}}} \quad & \sum_{m, \mathbf{g}, k} \left[\left(Q_s^k[t] - r_m^k \sum_{n=1}^N Q_n^{m, \mathbf{g}_1}[t] \right) r_m^k \alpha_{\mathbf{f}}^{m, \mathbf{g}} \right] \\ & + \sum_{m, \mathbf{g}} \left[(\mathbf{r}_m \cdot \mathbf{1})^2 \left(\sum_{n=1}^N Q_n^{m, \mathbf{g}_1}[t] \right) \beta_{\mathbf{f}}^{m, \mathbf{g}} \right], \quad (8) \\ \text{s.t.} \quad & \sum_{m, \mathbf{g}} (\alpha_{\mathbf{f}}^{m, \mathbf{g}} + \beta_{\mathbf{f}}^{m, \mathbf{g}}) \leq 1, \\ & \alpha_{\mathbf{f}}^{m, \mathbf{g}} = 0, \forall (\mathbf{f}, \mathbf{g}) \notin \mathcal{I}_1, \\ & \beta_{\mathbf{f}}^{m, \mathbf{g}} = 0, \forall (\mathbf{f}, \mathbf{g}) \notin \mathcal{I}_2, \forall (m, \mathbf{g}) \notin \mathcal{I}, \\ & \alpha_{\mathbf{f}}^{m, \mathbf{g}}, \beta_{\mathbf{f}}^{m, \mathbf{g}} \in \{0, 1\}, \forall m, \mathbf{g}. \end{aligned}$$

It is fairly straightforward to check that the algorithm given in this section results from this optimization problem. We remark that this optimization has many redundant variables that are introduced for the purpose of the proof.

Let an optimal assignment to the optimization problem in (8) be $\hat{\alpha}_{\mathbf{f}}^{m, \mathbf{g}}, \hat{\beta}_{\mathbf{f}}^{m, \mathbf{g}}$. Now, from (4), (6), (5) and (7), we can bound queue-lengths during block $t + 1$ as follows:

$$\begin{aligned} (Q_s^k[t + 1])^2 &= \left(Q_s^k[t] + A^k[t] - \left(\sum_{m, \mathbf{g}} r_m^k T \hat{\alpha}_{\mathbf{f}}^{m, \mathbf{g}} \right) \right)^2 \\ &\leq (Q_s^k[t])^2 + (A^k[t])^2 + \left(\sum_{m, \mathbf{g}} r_m^k T \hat{\alpha}_{\mathbf{f}}^{m, \mathbf{g}} \right)^2 \\ &\quad - 2Q_s^k[t] \left(\sum_{m, \mathbf{g}} r_m^k T \hat{\alpha}_{\mathbf{f}}^{m, \mathbf{g}} - A^k[t] \right), \forall k, \end{aligned}$$

$$\begin{aligned}
(\mathbf{r}_m \cdot \mathbf{1} Q_n^{m, \mathbf{g}_1}[t+1])^2 &\leq \left(\mathbf{r}_m \cdot \mathbf{1} Q_n^{m, \mathbf{g}_1}[t] + \mathbf{r}_m \cdot \mathbf{1} T \sum_{\mathbf{g}_2} (\hat{\alpha}_f^{m, \mathbf{g}} - \hat{\beta}_f^{m, \mathbf{g}}) \right)^2 \\
&= (\mathbf{r}_m \cdot \mathbf{1} Q_n^{m, \mathbf{g}_1}[t])^2 + \left(\mathbf{r}_m \cdot \mathbf{1} T \sum_{\mathbf{g}_2} (\hat{\alpha}_f^{m, \mathbf{g}} - \hat{\beta}_f^{m, \mathbf{g}}) \right)^2 \\
&\quad - 2(\mathbf{r}_m \cdot \mathbf{1})^2 Q_n^{m, \mathbf{g}_1}[t] T \sum_{\mathbf{g}_2} (\hat{\alpha}_f^{m, \mathbf{g}} - \hat{\beta}_f^{m, \mathbf{g}}), \forall m, \mathbf{g}_1.
\end{aligned}$$

Applying the law of iterated expectations, we obtain

$$\begin{aligned}
\mathbf{E}[V(\mathbf{Q}[t+1]) - V(\mathbf{Q}[t]) | \mathbf{Q}[t]] - M &\leq \sum_{\mathbf{f}} \pi_{\mathbf{f}} \left[- \sum_k 2Q_s^k[t] \left(\sum_{m, \mathbf{g}} r_m^k T \hat{\alpha}_f^{m, \mathbf{g}} - \lambda_k T \right) - \right. \\
&\quad \left. \sum_{m, \mathbf{g}_1, n} \left(2(\mathbf{r}_m \cdot \mathbf{1})^2 Q_n^{m, \mathbf{g}_1}[t] T \sum_{\mathbf{g}_2} (\hat{\alpha}_f^{m, \mathbf{g}} - \hat{\beta}_f^{m, \mathbf{g}}) \right) \right] \\
&= 2T \left[\sum_k Q_s^k[t] \left(\lambda_k - \sum_{m, \mathbf{g}, \mathbf{f}} (\pi_{\mathbf{f}} r_m^k \hat{\alpha}_f^{m, \mathbf{g}}) \right) + \right. \\
&\quad \left. \sum_{m, \mathbf{g}, n} (\mathbf{r}_m \cdot \mathbf{1})^2 Q_n^{m, \mathbf{g}_1}[t] \left(\sum_{\mathbf{f}} \pi_{\mathbf{f}} \right) \right]. \quad (9)
\end{aligned}$$

where M is a finite positive value, as the variance associated with the arrival processes are bounded and the throughput region is compact.

Let $a_f^{m, \mathbf{g}}, b_f^{m, \mathbf{g}}$ be the values given by Lemma 4. Now, substituting values $a_f^{m, \mathbf{g}}$ instead of $\hat{\alpha}_f^{m, \mathbf{g}}$ and $b_f^{m, \mathbf{g}}$ instead of $\hat{\beta}_f^{m, \mathbf{g}}$ in right hand side of (9) increases its value. This is due to the following reason. First, consider the linear program (LP) obtained by relaxing the integer constraints of the optimization problem (8) and introducing non-negativity constraints. This relaxation is tight as LPs have at least one optimal solution which is a boundary point. Next, the possible values for $a_f^{m, \mathbf{g}}, b_f^{m, \mathbf{g}}$ is a subset of the feasible set for the LP. Therefore, by substituting results from Lemma 4 in (9), we have

$$\begin{aligned}
\mathbf{E}[V(\mathbf{Q}[t+1]) - V(\mathbf{Q}[t]) | \mathbf{Q}[t]] - M &\leq 2T \left[\sum_k Q_s^k[t] \left(\lambda_k - \sum_{m, \mathbf{g}, \mathbf{f}} (\pi_{\mathbf{f}} r_m^k a_f^{m, \mathbf{g}}) \right) + \right. \\
&\quad \left. \sum_{m, \mathbf{g}, n} (\mathbf{r}_m \cdot \mathbf{1})^2 Q_n^{m, \mathbf{g}_1}[t] \left(\sum_{\mathbf{f}} \pi_{\mathbf{f}} (a_f^{m, \mathbf{g}} - b_f^{m, \mathbf{g}}) \right) \right] \\
&\leq -2T\delta \left[\sum_k Q_s^k[t] + \sum_{m, \mathbf{g}, n} (\mathbf{r}_m \cdot \mathbf{1})^2 Q_n^{m, \mathbf{g}_1}[t] \right]. \quad (10)
\end{aligned}$$

Now, from (10), it is fairly straightforward to see that there is strict negative drift except on a compact subset of the set of queue-lengths. This completes the proof. ■

VII. CONCLUSION

In this paper, we develop encoding-based queue architecture for cooperative relay networks. Cooperative relay networks are fundamentally different from traditional capacitated and non-cooperative wireless networks as they require physical layer coordination. This physical layer coordination cannot be abstracted out at the network layer in terms of bits-in-bits-out models, and thus a stability analysis that incorporates both the physical layer encoding and the network layer dynamics is needed, as performed in this paper. The encoding-based queue architecture is a succinct representation needed for generating network stabilizing algorithms. Using this queue-architecture, we show that throughput-optimal network algorithms can be developed even when the fade-distribution and input queue distributions are unknown.

REFERENCES

- [1] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity part I and part II," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927–48, Nov. 2003.
- [2] L. Dong, "Cross-layer design for cooperative wireless networks," Ph.D. dissertation, Drexel University, 2008.
- [3] A. Ephremides, "The audacity of throughput – a trilogy of rates," Plenary Lecture, ISIT, 2010.
- [4] A. Ephremides and B. Hajek, "Information theory and communication networks: An unconsummated union," *IEEE Trans. Inform. Theory*, pp. 2416–2434, 1998.
- [5] S. Bodas, J. Grubb, S. Sridharan, T. Ho, and S. Vishwanath, "Network with costs: Timing and flow decomposition," in *Proc. of WNC3*, Apr. 2007.
- [6] "3GPP long term evolution (LTE) coordinated multipoint transmission/reception (CoMP)," details at <http://www.3gpp.org/>.
- [7] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, "Coordinated multipoint transmission/reception techniques for LTE-advanced [Coordinated and Distributed MIMO]," *IEEE Wireless Communications*, vol. 17, no. 3, pp. 26–34, Jun. 2010.
- [8] E. Yeh and R. Berry, "Throughput optimal control of cooperative relaying networks," *IEEE Trans. Inform. Theory*, vol. 53, pp. 3827–3832, Oct. 2007.
- [9] —, "Throughput optimal control of wireless networks with two-hop cooperative relaying," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Nice, France, Jun. 2007.
- [10] L. Tassiulas and A. Ephremides, "Stability properties of constrained queuing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.
- [11] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 8, pp. 1452–1463, July 2006. [Online]. Available: <http://dx.doi.org/10.1109/JSAC.2006.879351>

- [12] L. Georgiadis, M. J. Neely, and L. Tassiulas, *Resource Allocation and Cross-Layer Control in Wireless Networks*. Now Publishers, 2006.
- [13] J. Jose, L. Ying, and S. Vishwanath, "On the stability region of amplify-and-forward cooperative relay networks," in *IEEE Info. Theory Workshop (ITW)*, Aug. 2009.
- [14] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.
- [15] S. Asmussen, *Applied Probability and Queues*. New York: Springer-Verlag, 2003.