AN OVERVIEW OF BEAMFORMING AND POWER ALLOCATION FOR MIMO RELAYS

Yingbo Hua

Department of Electrical Engineering, University of California, Riverside, CA 92521.
Email: yhua@ee.ucr.edu

ABSTRACT

This paper provides an overview of the design of transmit and receive beamformers and transmit power allocation for MIMO relays. Soft and hard methods for interference cancelation are presented, which play a critical role for all MIMO relays whether they are full-duplex or half-duplex, regenerative or non-regenerative, one-way relay or two-way relay. A perspective of MIMO relays in a network of many hops is illustrated. A distinction is made between feedback loop of energy and feedback loop of noise. Subspace computation and a generalized water-filling (GWF) algorithm are shown as important building blocks for the design.

1. INTRODUCTION

MIMO relays are an emerging technology important for both civilian and military communications. A MIMO relay has multiple transmit and receive antennas that provide highly valuable spatial diversity. Such spatial diversity can translate to increased spectral efficiency, increased reliability, reduced power consumption, reduced interference to the environment, and other benefits. MIMO relays are particularly important for environments with severe path loss and rich scattering due to woods, buildings, low elevation of antennas, etc. MIMO relays can be mounted on soldiers, vehicles or many other types of platforms for long distance multihop communication in cities, remote or isolated areas.

In this paper, we present a number of strategies for beamforming and power allocation for MIMO relays. A MIMO relay can be regenerative or non-regenerative, full-duplex or half-duplex, one-way or two-way. A regenerative relay decodes its received information and then re-encodes it for transmission, which can cause a long processing delay but stop the propagation of noise. A non-regenerative relay performs amplify-and-forward, which has a much shorter delay but at the expense of noise propagation. A full-duplex relay transmits and receives in the same time and same frequency, which is spectrally efficient but causes a problem of self-interference. A half-duplex relay transmits and receives separately in two orthogonal channels in either time or frequency, which is easy to implement but not spectrally the most efficient. A one-way relay relays information in a single direction in a single time and single frequency, which is good for most situations where traffics are not symmetric. A two-way relay relays information in two directions in a single time and single frequency, which can improve the spectral efficiency of a half-duplex relay.

There are eight possible strategies for classifying relay systems: regenerative vs non-regenerative, full-duplex vs half-duplex, and one-way vs two-way. For easy reference, we will use the following notations: *R–regenerative, N–non-regenerative; F–full-duplex, H–half-duplex; 1–one-way, 2– two-way*. For example, RF1 denotes regenerative, full-duplex and one-way relays. For each of the eight combinations, beamforming and power allocation are important issues, which we will discuss next in detail. Our goal of beamforming and power allocation is to reduce or eliminate interferences and increase the end-to-end data rate of relays subject to power constraints. Interference at a relay can be weak or strong, from another node or its own, in an open loop or closed loop. We will point out these properties and their implications. For all schemes shown in this paper, the knowledge of channel matrices but not signal waveforms is needed.

2. RH1 MIMO RELAYS

A RH1 MIMO relay system consists of three or more nodes cascaded with one another. These nodes can be indexed sequentially as 0, 1, 2, $\cdots$, L. The nodes 0 and L are respectively source and destination nodes, and all other nodes in between are half-duplex regenerative MIMO relays. For the purpose of multihop long distance communications, we can assume that the (average) channel gain from node $i \pm j$ to node i is negligible compared to that from node $i \pm 1$ to node i for all i and $j \geq 2$. In other words, we assume that the channel gain over a distance of two or more hops is negligible compared to that over a one-hop distance. This assumption is valid for strong path loss environment where the path loss exponent could range from 4 to 5 or even higher (possibly due to shadowing). The residue interference over a distance of two or more hops can be treated as noise. Then, node i can transmit to node $i+1$ at the same time and same frequency as node $i+j$ to node $i+j+1$ for $j \geq 3$, without mutual interference. For large L, the end-to-end (ETE) spectral efficiency of the relay system in terms of bits per channel use (i.e., bits per sec-

ond per Hertz) is given by $C_{RH1} = \frac{1}{3}\min_{i=0,1,\cdots,L-1} C_{i+1,i}$ where $C_{i+1,i}$ is the capacity from node i to node $i+1$ and $1/3$ is a penalty factor due to three time slots needed to isolate any three adjacent links. See Fig. 1. When $L = 2$, the penalty factor is $1/2$.

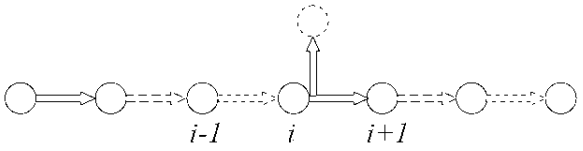

Fig. 1. *A RH1 (regenerative, half-duplex, one-way) relay system, and a scenario that a node outside the relay system may receive interference from relay i. The three line patterns represent three orthogonal time/frequency slots. This illustration also applies to NH1 (non-regenerative, half-duplex, one-way).*

The ETE delay of this relay system is proportional to $L-1$, i.e., $D_{RH1} = \alpha_{RH1}(L-1)$. For regenerative relays, the delay at each relay, α_{RH1}, is lower bounded by the time needed for digital decoding and re-encoding. At high SNR, robust decoding and re-encoding can be done symbol by symbol, and in this case α_{RH1} is in the order of a symbol interval. At low SNR, robust decoding and re-encoding need to be done across multiple symbols, and in this case α_{RH1} is increased accordingly.

The beamforming and power allocation problem for the RH1 MIMO relay system can be done by treating each link separately. Let the channel from node i to node $i+1$ be modeled as

$$\mathbf{y}_{i+1} = \mathbf{H}_{i+1,i}\mathbf{x}_i + \mathbf{w}_{i+1} \tag{1}$$

where $\mathbf{x}_i$ is the output from node i, $\mathbf{y}_{i+1}$ the input to node $i+1$, $\mathbf{w}_{i+1}$ the noise-plus-interference received by node $i+1$, and $\mathbf{H}_{i+1,i}$ the channel matrix from node i to node $i+1$. This channel model holds for any narrow enough subband (i.e., the bandwidth of the subband is significantly smaller than the inverse of the channel delay spread). With known $\mathbf{H}_{i+1,i}$, we can compute the singular value decomposition (SVD):

$$\mathbf{H}_{i+1,i} = \mathbf{U}_{i+1,i}\mathbf{\Sigma}_{i+1,i}\mathbf{V}_{i+1,i}^H \tag{2}$$

with $\mathbf{\Sigma}_{i+1,i} = diag(\sigma_{i+1,i,1}, \sigma_{i+1,i,2}, \cdots) \in R^{r_{i+1,i}\times r_{i+1,i}}$, $\sigma_{i+1,i,l} \geq \sigma_{i+1,i,l+1}$ and $r_{i+1,i} = rank(\mathbf{H}_{i+1,i})$.

Assuming Gaussian $\mathbf{w}_{i+1}$ with $E\{\mathbf{w}_{i+1}\mathbf{w}_{i+1}^H\} = \mathbf{I}$, the optimal transmit beamforming (or precoding) matrix at node i is given by $\mathbf{Q}_i^{(t)} = \mathbf{V}_{i+1,i}$, and the optimal power allocation (diagonal) matrix $\mathbf{D}_i$ at the node i is given by the *water-filling* algorithm (e.g., see [1]): $(\mathbf{D}_i)_{l,l} = (v - 1/\sigma_{i+1,i,l}^2)^+$ where $(x)^+ = \max(0, x)$, v is chosen to meet the total power constraint $Tr(\mathbf{D}_i) = P_i$. Let m_i be the number of nonzero diagonal elements of $\mathbf{D}_i$, $\mathbf{D}_{i,m_i}$ be the $m_i \times m_i$ upper left submatrix of $\mathbf{D}_i$, and $\mathbf{Q}_{i,m_i}^{(t)}$ the left most m_i columns of $\mathbf{Q}_i^{(t)}$. Then, the optimal $\mathbf{x}_i$ to be transmitted from node i is $\mathbf{x}_i = \mathbf{Q}_{i,m_i}^{(t)}\mathbf{D}_{i,m_i}^{1/2}\mathbf{s}_i$ where $\mathbf{s}_i$ is a $m_i \times 1$ vector of independent symbols each with zero-mean and unit-variance. If $E\{\mathbf{w}_{i+1}\mathbf{w}_{i+1}^H\} = \mathbf{W}_{i+1}$ which is known, the above discussion still holds if we replace $\mathbf{y}_{i+1}$ by $\mathbf{W}_{i+1}^{-1/2}\mathbf{y}_{i+1}$ and $\mathbf{H}_{i+1,i}$ by $\mathbf{W}_{i+1}^{-1/2}\mathbf{H}_{i+1,i}$.

If the node i is not allowed to interfere with any of its neighboring nodes (not belonging to the relay system) and the channel matrix $\tilde{\mathbf{H}}_{k,i}$ from the node i to each of these neighboring nodes ($k = 1, \cdots, K$) is known, then the transmit beamforming matrix $\mathbf{Q}_i^{(t)}$ at node i must satisfy

$$\tilde{\mathbf{H}}_i\mathbf{Q}_i^{(t)} = 0 \text{ with } \tilde{\mathbf{H}}_i = \begin{bmatrix} \tilde{\mathbf{H}}_{1,i} \\ \cdots \\ \tilde{\mathbf{H}}_{K,i} \end{bmatrix} \tag{3}$$

The case of $K = 1$ is illustrated in Fig. 1. If the number of transmit antennas at node i, denoted by $n_i^{(t)}$, is larger than $r_i \doteq rank(\tilde{\mathbf{H}}_i)$, then we can write $\mathbf{Q}_i^{(t)} = \mathbf{Q}_i^{(t,1)}\mathbf{Q}_i^{(t,2)}$ where $\mathbf{Q}_i^{(t,1)} \in C^{n_i^{(t)}\times(n_i^{(t)}-r_i)}$ is orthonormal and satisfying $span(\mathbf{Q}_i^{(t,1)}) = null(\tilde{\mathbf{H}}_i)$. The second layer beamformer $\mathbf{Q}_i^{(t,2)}$, and the corresponding optimal power allocation matrix, $\mathbf{D}_i$, can be found in the same way as shown previously but with $\mathbf{H}_{i+1,i}$ replaced by $\mathbf{H}_{i+1,i}\mathbf{Q}_i^{(t,1)}$. The transmitted signal from node i is then $\mathbf{x}_i = \mathbf{Q}_i^{(t,1)}\mathbf{Q}_{i,m_i}^{(t,2)}\mathbf{D}_{i,m_i}^{1/2}\mathbf{s}_i$.

If the node i is allowed to cause a small (rather than zero) amount of interference to its neighboring nodes, then we have a much different problem of beamforming and power allocation. In this case, we should impose soft constraints on interference such as

$$Tr(\tilde{\mathbf{H}}_{k,i}\mathbf{Q}_i^{(t)}\mathbf{D}_i\mathbf{Q}_i^{(t)H}\tilde{\mathbf{H}}_{k,i}^H) \leq \epsilon_{k,i} \tag{4}$$

with $k = 1, \cdots, K$, and $\epsilon_{k,i}$ is the upper limit of the allowed average power of the interference from node i to node k. In addition to the total power constraint: $Tr(\mathbf{D}_i) \leq P_i$, we have $K+1$ inequality constraints to determine the optimal transmit beamforming matrix and the optimal power allocation matrix for node i. This is not a conventional problem. However, as shown in [2], the maximization of the capacity from node i to node $i+1$ subject to these constraints, with respect to $\mathbf{Q}_i^{(t)}$ and $\mathbf{D}_i$, is a convex optimization problem and can be solved by a *generalized water-filling* (GWF) algorithm. The soft constraint (4) is more general and allows a higher capacity than the hard constraint (3).

Given the transmit beamformer $\mathbf{Q}_i^{(t)}$ and the transmit power allocation $\mathbf{D}_i$ at relay i, the input to relay $i+1$ is $\mathbf{y}_{i+1} = \mathbf{H}_{i+1,i}\mathbf{Q}_{i,m_i}^{(t)}\mathbf{D}_{i,m_i}^{1/2}\mathbf{s}_i + \mathbf{w}_{i+1}$ and the minimum mean square error (MMSE) estimate $\hat{\mathbf{s}}_i$ of $\mathbf{s}_i$ is standard, i.e., $\hat{\mathbf{s}}_i = \mathbf{G}_{i+1}\mathbf{y}_{i+1}$ where $\mathbf{G}_{i+1} = \bar{\mathbf{H}}_{i+1,i}^H(\bar{\mathbf{H}}_{i+1,i}\bar{\mathbf{H}}_{i+1,i}^H + \mathbf{I})^{-1} = (\bar{\mathbf{H}}_{i+1,i}^H\bar{\mathbf{H}}_{i+1,i} + \mathbf{I})^{-1}\bar{\mathbf{H}}_{i+1,i}^H$ and $\bar{\mathbf{H}}_{i+1,i} = \mathbf{H}_{i+1,i}\cdot$

$\mathbf{Q}_{i,m_i}^{(t)}\mathbf{D}_{i,m_i}^{1/2}$. The optimal receive beamformer at relay $i+1$ is embedded in $\mathbf{G}_{i+1}$. In the absence of any interference constraint on the transmit beamformer, we have $\mathbf{G}_{i+1} = (\mathbf{D}_{i,m_i}\mathbf{\Sigma}_{i+1,i,m_i}^2 + \mathbf{I})^{-1}\mathbf{D}_{i,m_i}^{1/2}\mathbf{\Sigma}_{i+1,i,m_i}\mathbf{U}_{i+1,i,m_i}^H$ where $\mathbf{U}_{i+1,i}^H$ can be viewed as the optimal receive beamformer since all other matrices are diagonal.

If node $i+1$, in its receive mode, is strongly interfered by sources of known subspace, then we can use $\hat{\mathbf{s}}_i = \mathbf{G}_{i+1}\mathbf{Q}_{i+1}^{(r,1)}\mathbf{y}_{i+1}$ where $\mathbf{Q}_{i+1}^{(r,1)}$ is orthonormal and chosen to be orthogonal to the interference subspace in $\mathbf{y}_{i+1}$. The optimal design of $\mathbf{Q}_i^{(t)}$, $\mathbf{D}_i$ and $\mathbf{G}_{i+1}$ follows the previous procedure but with $\mathbf{H}_{i+1,i}$ replaced by $\mathbf{Q}_{i+1}^{(r,1)}\mathbf{H}_{i+1,i}$.

In the rest of this paper, we will assume (unless mentioned otherwise for self-interference) that all channel matrices $\mathbf{H}_{i+1,i}$ are effective channel matrices after the hard interference constraints are already implemented at the transmitting and receiving nodes and the residue noise-plus-interference at receiving nodes is already whitened.

3. NH1 MIMO RELAYS

We now consider a NH1 MIMO relay system, which is similar to the RH1 MIMO relay system discussed earlier, except that all relays are non-regenerative. See Fig. 1. A non-regenerative half-duplex relay does not perform digital decoding and re-encoding, and its half-duplex mode can be realized by using two frequency channels (instead of two time slots). Analog circuitry can be designed to perform RF-to-baseband conversion, baseband algorithmic operations (addition, subtraction, multiplication and division), baseband-to-RF conversion, and amplification. Hence, the delay at each relay, α_{NH1}, is much smaller than a symbol interval. The ETE delay of the system is $D_{NH1} = \alpha_{NH1}(L-1)$ and clearly $D_{NH1} << D_{RH1}$.

Because of the non-regenerative nature, the noise received at each relay is propagated to its next node. Non-regenerative relays are suitable only for high SNR scenarios. The number of non-regenerative hops should be limited to prevent the accumulated noise from becoming too large. Assuming a large L, the ETE spectral efficiency of the NH1 relay system in terms of bits/s/Hz is $C_{NH1} = \frac{1}{3}\tilde{C}_{L,0}$. Here, $\tilde{C}_{L,0}$ is the capacity of the analog channel from node 0 to node L without counting the three different frequency channels needed for isolating any three adjacent links. When $L = 2$, the penalty factor is $1/2$. If the number of antennas at each node is large and all elements in all channel matrices of adjacent links are i.i.d. Gaussian, then the distribution of the singular values of each channel matrix "hardens" to a quarter circle law, and in this case we expect $C_{NH1} < C_{RH1}$.

The beamforming and power allocation problem for the NH1 MIMO relay system can be treated as follows. Relay i can use a receive beamforming matrix $\mathbf{Q}_i^{(r)}$, a transmit beamforming matrix $\mathbf{Q}_i^{(t)}$ and a diagonal power (amplification) allocation matrix $\mathbf{D}_i$. The product of the three matrices, $\mathbf{F}_i = \mathbf{Q}_i^{(t)}\mathbf{D}_i^{1/2}\mathbf{Q}_i^{(r)}$, governs the baseband input-output relationship of (non-regenerative) relay i, i.e., $\mathbf{x}_i = \mathbf{F}_i\mathbf{y}_i$. Both $\mathbf{Q}_i^{(r)}$ and $\mathbf{Q}_i^{(t)}$ are orthonormal and of equal rank in general. The source node uses a transmit beamforming matrix $\mathbf{Q}_0^{(t)}$ and a diagonal power allocation matrix $\mathbf{D}_0$. The destination node uses a receive (front-end) beamforming matrix $\mathbf{Q}_L^{(r)}$. The choices of these beamforming matrices and power allocation matrices affect significantly the capacity $\tilde{C}_{L,0}$ of the system. However, if all noises at the relays are white (or whitened by using known covariances) and if for each i the channel matrix $\mathbf{H}_{i+1,i}$ (see (2)) from node i to node $i+1$ is known (for example, to these two nodes), then the optimal choice of $\mathbf{Q}_i^{(t)}$ and $\mathbf{Q}_{i+1}^{(r)}$, can be shown as in [3] to be

$$\mathbf{Q}_i^{(t)} = \mathbf{V}_{i+1,i} \text{ and } \mathbf{Q}_{i+1}^{(r)} = \mathbf{U}_{i+1,i}^H \tag{5}$$

The optimality of this beamforming structure holds for a class of Schur-convex or Schur-concave objectives, which include the maximization of $\tilde{C}_{L,0}$, subject to the transmit power constraints at all nodes: $E(\|\mathbf{x}_i\|^2) \le P_i$, $i = 1, \cdots, L-1$. The optimal power allocation matrices $\mathbf{D}_i$ for all i can be found by a cyclic water-filling algorithm [3]. In other words, for each $i = 0, \cdots, L-1$, we fix $\mathbf{D}_j$, $j \neq i$, at their previous values, and determine $\mathbf{D}_i$ using the water-filling algorithm. We repeat the process until convergence. For $L = 2$, the optimality of (5) and the cyclic water-filling algorithm were first shown in [5], which uses a key result from [6], and later generalized in [4]. This result is further generalized for any integer L in [3]. Upon convergence of the cyclic water-filling algorithm, the number of nonzero diagonal entries in each of $\mathbf{D}_i$ can be denoted by m, which is independent of i. Then, the optimal symbol vector to be transmitted from node 0 is $\mathbf{x}_0 = \mathbf{V}_{1,0,m}\mathbf{D}_{0,m}^{1/2}\mathbf{s}_0$ where $\mathbf{s}_0$ is a $m \times m$ symbol vector. Recall that the subscript $_m$ denotes the upper left $m \times m$ submatrix of a diagonal matrix or the left most m columns of an orthonormal matrix. The optimal output from relay i is $\mathbf{x}_i = \mathbf{V}_{i+1,i,m}\mathbf{D}_{i,m}^{1/2}\mathbf{U}_{i,i-1,m}^H\mathbf{y}_i$. The receiver at node L follows from the MMSE estimation of $\mathbf{s}_0$.

If there is any hard interference constraint like (3) for transmit beamformers, the above optimal solution for beamforming and power allocation can be modified easily. But if a soft interference constraint like (4) is required, there is currently no known optimal solution for the NH1 relay system.

4. RF1 MIMO RELAYS

In a RF1 MIMO relay system, each relay is full-duplex which transmits and receives at the same time and same frequency. However, since each relay is regenerative, the symbols a relay transmits at one time and one frequency are different from those received by the relay at the same time and same frequency. Furthermore, there is no *feedback loop of energy*

within a regenerative relay because of the digital decoding and re-encoding operations. However, there is potentially a *feedback loop of noise* within each relay because its transmitted signal is an interference to its received signal. The self-interference is generally of very high power, much higher (e.g., 100 dB) than that of a signal received from another node. Therefore, any method for self-interference cancelation should aim to cancel the interference completely, or otherwise the power of the residue interference in practice could be still significant compared to the power of the received signal.

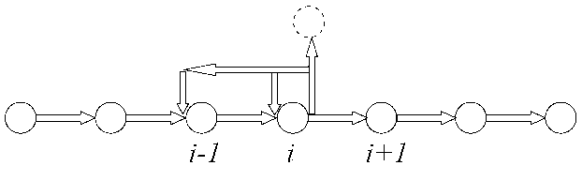

Fig. 2. *A RF1 (regenerative, full-duplex, one-way) relay system where interferences from relay i are highlighted. This illustration also applies to NF1 (non-regenerative, full-duplex, one-way).*

The purpose of using full-duplex relays is to improve spectral efficiency. For a linear network of relays, as described previously, we would ideally expect all relays to receive and transmit at the same time and same frequency so that the ETE capacity in bits/s/Hz is simply $C_{RF1} = \min_{i=0,\cdots,L-1} C_{i+1,i}$ where there is no penalty factor (such as 1/3 or 1/2) at all. But the signal transmitted from node i not only interferes its own input signal but also the input signal at node $i-1$ which is the adjacent upper-stream node of node i.

We assume that node i uses an $n_i^{(r)}$-antenna array for receive and an $n_i^{(t)}$-antenna array for transmit. The separation of receive antennas from transmit antennas for a full-duplex relay should make the circuit design easy especially when considering the large power ratio between the output and input signals. The antennas could stick up high in air, and in this case the self-interference channel matrix $\mathbf{H}_{i,i} \in C^{n_i^{(r)} \times n_i^{(t)}}$ from transmit antennas to receive antennas for node i should have a low rank (e.g., rank one) property which is useful for interference cancelation. But the antennas could also be embedded inside a more complex medium (such as hidden inside a vehicle), and in this case $\mathbf{H}_{i,i}$ may not have a low rank property. For a general treatment, we will not rely on any low rank assumption of $\mathbf{H}_{i,i}$. However, we have $r_{i,i} \doteq rank(\mathbf{H}_{i,i}) \leq \min(n_i^{(r)}, n_i^{(t)})$.

In order for the received signal at node i to be not affected at all by self-interference, we can design the transmit beamformer $\mathbf{Q}_i^{(t)}$ at relay i using two layers, i.e., $\mathbf{Q}_i^{(t)} = \mathbf{Q}_i^{(t,1)}\mathbf{Q}_i^{(t,2)}$ where $\mathbf{Q}_i^{(t,1)}$ satisfies

$$\mathbf{H}_{i,i}\mathbf{Q}_i^{(t,1)} = 0 \tag{6}$$

which is similar to (3). In fact, once we have the QR decomposition: $\mathbf{H}_{i,i}^H = \mathbf{Q}_{i,i}\mathbf{R}_{i,i}$, we can construct an $n_i^{(t)} \times (n_i^{(t)} - r_{i,i})$ orthonormal matrix $\mathbf{Q}_i^{(t,1)}$ from the orthogonal complement of $range(\mathbf{Q}_{i,i})$. Naturally, $\mathbf{Q}_{i,i}^H\mathbf{Q}_i^{(t,1)} = 0$ implies $\mathbf{H}_{i,i}\mathbf{Q}_i^{(t,1)} = 0$. In general, $\mathbf{Q}_{i,i}$ and the rank $r_{i,i}$ can be estimated by using the SVD or the rank-revealing QR of $\mathbf{H}_{i,i}^H$. To ensure that at least one stream of data can be transmitted from node i, we must have $n_i^{(t)} \geq r_{i,i} + 1$.

If $\mathbf{H}_{i,i}\mathbf{Q}_i^{(t,1)} = 0$ does not hold exactly, the residue self-interference at node i can be processed by its receive beamformer $\mathbf{Q}_i^{(r)}$. We can design $\mathbf{Q}_i^{(r)}$ with two layers, i.e., $\mathbf{Q}_i^{(r)} = \mathbf{Q}_i^{(r,2)}\mathbf{Q}_i^{(r,1)}$ where $\mathbf{Q}_i^{(r,1)}$ satisfies

$$\mathbf{Q}_i^{(r,1)}\mathbf{H}_{i,i}\mathbf{Q}_i^{(t,1)} = 0 \tag{7}$$

which is a generalization of (6). Here, $\mathbf{Q}_i^{(r,1)}$ cancels out the residue self-interference left out by $\mathbf{Q}_i^{(t,1)}$. The solution space of (7) is large in general because for any $\mathbf{Q}_i^{(t,1)}$ such that $\mathbf{H}_{i,i}\mathbf{Q}_i^{(t,1)}$ has a rank less than $n_i^{(r)}$, there is a $\mathbf{Q}_i^{(r,1)}$ satisfying (7). But a meaningful $\mathbf{Q}_i^{(t,1)}$ should at least make $rank(\mathbf{H}_{i,i}\mathbf{Q}_i^{(t,1)})$ less than $rank(\mathbf{H}_{i,i})$. Furthermore, in practice, we may need to ensure that the power of the residue interference $\mathbf{y}_i^{(i)} \doteq \mathbf{H}_{i,i}\mathbf{Q}_i^{(t,1)}\mathbf{Q}_i^{(t,2)}\mathbf{D}_i^{1/2}\mathbf{s}_i$ is not so large to saturate the input circuitry of node i.

A 2×2 special case of (7) was shown in [7] for a full duplex repeater. Using self-interference cancelation for full-duplex MIMO relay was also addressed recently in [8], [9], [10] and [11].

Let the SVD of $\mathbf{H}_{i,i} \in C^{n_i^{(r)} \times n_i^{(t)}}$ be written as two parts:

$$\mathbf{H}_{i,i} = \mathbf{U}_{i,i}^{(1)}\mathbf{\Sigma}_{i,i}^{(1)}\mathbf{V}_{i,i}^{(1)H} + \mathbf{U}_{i,i}^{(2)}\mathbf{\Sigma}_{i,i}^{(2)}\mathbf{V}_{i,i}^{(2)H} \tag{8}$$

where the first term is orthogonal to the second term. Then, a solution to (7) is $\mathbf{Q}_i^{(t,1)} = \mathbf{V}_{i,i}^{(1)\perp}$ and $\mathbf{Q}_i^{(r,1)} = \mathbf{U}_{i,i}^{(2)\perp H}$. Here, $\perp$ means that $range(\mathbf{X}^{\perp})$ is the orthogonal complement of $range(\mathbf{X})$. Given $r_{i,i} \doteq rank(\mathbf{H}_{i,i})$, the total number of the possible partitions like (8) is $\sum_{i=0}^{r_{i,i}} \binom{r_{i,i}}{i} = 2^{r_{i,i}}$. If we choose that all diagonal entries of $\mathbf{\Sigma}_{i,i}^{(1)} \in R^{\hat{r}_{i,i} \times \hat{r}_{i,i}}$ are larger than those of $\mathbf{\Sigma}_{i,i}^{(2)} \in R^{(r_{i,i}-\hat{r}_{i,i}) \times (r_{i,i}-\hat{r}_{i,i})}$ so that $\mathbf{Q}_i^{(t,1)}$ cancels the dominant interference while $\mathbf{Q}_i^{(r,1)}$ cancels the residual interference, then the total number of possible choices for $\mathbf{Q}_i^{(t,1)} \in C^{n_i^{(t)} \times (n_i^{(t)} - \hat{r}_{i,i})}$ and $\mathbf{Q}_i^{(r,1)} \in C^{(n_i^{(r)} - r_{i,i} + \hat{r}_{i,i}) \times n_i^{(r)}}$ is $r_{i,i}$. If we want the number of the input streams, $n_i^{(r)} - r_{i,i} + \hat{r}_{i,i}$, to be the same as that of the output streams, $n_i^{(t)} - \hat{r}_{i,i}$, at relay i, we should choose $\hat{r}_{i,i} = (n_i^{(t)} - n_i^{(r)} + r_{i,i})/2$.

As an alternative to (6) and (7), we can allow a small amount of residue interference from the transmit beamformer $\mathbf{Q}_i^{(t)}$ and design the entire receiver $\mathbf{G}_i$ at relay i based on MMSE. Note that for any given time slot (for either symbol or packet) the information transmitted by regenerative relay

$i-1$ can be treated as independent of the information transmitted by regenerative relay i. Hence, the residue interference $\mathbf{y}_i^{(i)}$ is independent of the signal transmitted from node $i-1$. Also, the covariance matrix of $\mathbf{y}_i^{(i)}$ is known to node i. The tolerance of a small interference leakage at relay i allows a larger number of data streams received by and/or transmitted from node i. In general, the degrees of freedom in the receive beamformer is the best exploited via MMSE estimation, and zero-forcing does not utilize the freedom optimally.

As mentioned earlier, node $i-1$ is also interfered by the transmitted signal from node i. To handle this interference, we need to design the second-layer transmit beamformer for node i, denoted by $\mathbf{Q}_i^{(t,2)}$. The power of this interference is generally much smaller than that of the self-interference within a single node. While the hard interference cancelation is still possible with a high cost of the freedom in $\mathbf{Q}_i^{(t,2)}$, a better strategy should be based on a soft interference constraint like (4). Once the first-layer transmit beamformer $\mathbf{Q}_i^{(t,1)}$ is determined as discussed before, the second-layer transmit beamformer $\mathbf{Q}_i^{(t,2)}$ along with the power allocation matrix $\mathbf{D}_i$ can be found by following the GWF algorithm [2] as discussed in Section 2. To determine $\mathbf{Q}_i^{(t,2)}$ and $\mathbf{D}_i$, we need to use $\mathbf{H}_{i+1,i}\mathbf{Q}_i^{(t,1)}$ as the effective channel matrix from node i to node $i+1$ and $\mathbf{H}_{i-1,i}\mathbf{Q}_i^{(t,1)}$ as the effective interference channel matrix from node i to node $i-1$.

The ETE delay of the RF1 system is the same as that of the RH1 system although the spectral efficiency of the former is higher.

5. NF1 MIMO RELAYS

The illustration in Fig. 2 also applies to a NF1 MIMO relay system where each relay is now non-regenerative. As in Section 3, the matrix $\mathbf{F}_i = \mathbf{Q}_i^{(t)}\mathbf{D}_i^{1/2}\mathbf{Q}_i^{(r)}$ governs the input and output of the non-regenerative MIMO relay i, i.e., $\mathbf{x}_i = \mathbf{F}_i\mathbf{y}_i$. But unlike the NH1 system, we now have full-duplexity and the signal $\mathbf{x}_i$ transmitted from relay i can be strongly present in the signal $\mathbf{y}_i$ received by relay i and also in the signal $\mathbf{y}_{i-1}$ received by relay $i-1$. To cancel the self-interference within relay i, we can let $\mathbf{F}_i = \mathbf{Q}_i^{(t,1)}\mathbf{F}_i^{(2)}\mathbf{Q}_i^{(r,1)}$ where $\mathbf{Q}_i^{(t,1)}$ and $\mathbf{Q}_i^{(r,1)}$ satisfy (7) as discussed in Section 4. To cancel the interference from node i to node $i-1$, we can have an additional beamformer $\mathbf{Q}_i^{(t,2)}$ satisfying $\mathbf{H}_{i-1,i}\mathbf{Q}_i^{(t,1)}\mathbf{Q}_i^{(t,2)} = 0$. The overall relay matrix at relay i is now $\mathbf{F}_i = \mathbf{Q}_i^{(t,1)}\mathbf{Q}_i^{(t,2)}\mathbf{F}_i^{(3)}\mathbf{Q}_i^{(r,1)}$.

The optimal design of $\mathbf{F}_i^{(3)}$ for $i = 1, \cdots, L-1$, along with $\mathbf{Q}_0^{(t)}$, $\mathbf{D}_0$ and $\mathbf{Q}_{L+1}^{(r)}$ (see Section 3), can be done similarly as in [3]. To be more specific, we can write $\mathbf{F}_i^{(3)} = \mathbf{Q}_i^{(t,3)}\mathbf{D}_i^{1/2}\mathbf{Q}_i^{(r,2)}$ for $i = 1, \cdots, L-1$. And define $\mathbf{Q}_0^{(t,1)} = \mathbf{Q}_0^{(t,2)} = \mathbf{I}$, $\mathbf{Q}_{L+1}^{(r,1)} = \mathbf{I}$ and $\mathbf{H}_{i+1,i}^{(3)} = \mathbf{Q}_{i+1}^{(r,1)}\mathbf{H}_{i+1,i}\mathbf{Q}_i^{(t,1)}\mathbf{Q}_i^{(t,2)}$. Similar to (2), we can write the SVD: $\mathbf{H}_{i+1,i}^{(3)} = \mathbf{U}_{i+1,i}^{(3)}\mathbf{\Sigma}_{i+1,i}^{(3)}\mathbf{V}_{i+1,i}^{(3)H}$. Then, it can be shown [3] that the following is optimal: $\mathbf{Q}_i^{(r,2)} = \mathbf{U}_{i,i-1}^{(3)H}$ and $\mathbf{Q}_i^{(t,3)} = \mathbf{V}_{i+1,i}^{(3)}$. The optimal design of $\mathbf{D}_i$ follows the same cyclic water-filling.

The ETE spectral efficiency of the NF1 system does not have the penalty factor (such as $1/3$ or $1/2$) as for the NH1 system. The ETE delay of the NF1 system is the same as that of the NH1 system.

If $\mathbf{Q}_i^{(t,1)}$ and $\mathbf{Q}_i^{(r,1)}$ do not meet (7) exactly, there will be a feedback loop of energy which can cause instability of the non-regenerative relay. The overall open-loop transfer matrix of relay i (with respect to the signal to be amplified by $\mathbf{D}_i^{(1/2)}$) is

$$\mathbf{T}_i^{(open)} = \mathbf{Q}_i^{(r,2)}\mathbf{Q}_i^{(r,1)}\mathbf{H}_{i,i}\mathbf{Q}_i^{(t,1)}\mathbf{Q}_i^{(t,2)}\mathbf{Q}_i^{(t,3)}\mathbf{D}_i^{(1/2)} \quad (9)$$

To ensure stability within relay i, all singular values of $\mathbf{T}_i^{(open)}$ must be less than one. Furthermore, they should be much less than one in general, or otherwise the input-output of the non-regenerative relay i is no longer governed by $\mathbf{x}_i = \mathbf{F}_i\mathbf{y}_i$ but rather by $\mathbf{x}_i = (\mathbf{I} - \mathbf{F}_i\mathbf{H}_{i,i})^{-1}\mathbf{F}_i\mathbf{y}_i$. It is important to stress that a small (non-zero) singular value of $\mathbf{Q}_i^{(r,1)}\mathbf{H}_{i,i}\mathbf{Q}_i^{(t,1)}$ could cause a large singular value of $\mathbf{T}_i^{(open)}$ because of the largeness of $\mathbf{D}_i^{(1/2)}$. This problem does not exist for the RF1 system because the input energy of a regenerative relay does not affect its output energy.

6. TWO-WAY MIMO RELAYS

In this section, we discuss two-way MIMO relays: the NH2, RH2, RF2 and NF2 systems. We first discuss the NH2 MIMO relay system as shown in Fig. 3. Here, nodes 1 and 3 exchange information with help from node 2 – the relay. The relay is half-duplex, which receives a combined signal $\mathbf{y}_2$ from nodes 1 and 3 in frequency band 1, and broadcasts a signal $\mathbf{x}_2 = \mathbf{F}_2\mathbf{y}_2$ back to nodes 1 and 2 in frequency band 2. The matrix $\mathbf{F}_2 \in C^{n_2^{(t)} \times n_2^{(r)}}$ has an SVD $\mathbf{F}_2 = \mathbf{Q}_2^{(t)}\mathbf{D}_2^{1/2}\mathbf{Q}_2^{(r)}$ where $\mathbf{Q}_2^{(t)}$ is the transmit beamformer, $\mathbf{Q}_2^{(r)}$ the receive beamformer and $\mathbf{D}_2$ the (diagonal) power allocation matrix.

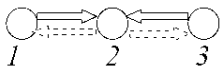

Fig. 3. *A NH2 (non-regenerative, half-duplex, two-way) relay system. This illustration also applies to a RH2 (regenerative, half-duplex, two-way) system.*

In frequency band 1, the signals transmitted from nodes 1 and 3 can be expressed as $\mathbf{x}_i = \mathbf{Q}_i^{(t)}\mathbf{D}_i^{1/2}\mathbf{s}_i$ for $i = 1, 3$, and the signal received at node 2 is $\mathbf{y}_2 = \sum_{i=1,3}\mathbf{H}_{2,i}\mathbf{x}_i + \mathbf{w}_2$. In frequency band 2, the signals received at nodes 1 and 3 are

$\mathbf{y}_i = \mathbf{H}_{i,2}\mathbf{x}_2 + \mathbf{w}_i = \mathbf{H}_{i,2}\mathbf{F}_2\mathbf{y}_2 + \mathbf{w}_i$, $i = 1, 3$. Here, $\mathbf{w}_i$, i=1,2,3, are noise. Since $\mathbf{x}_i$, $i = 1, 3$, are known to nodes 1 and 3, respectively, with some additional knowledge of channel matrices, the self-interference term $\mathbf{H}_{i,2}\mathbf{F}_2\mathbf{H}_{2,i}\mathbf{x}_i$ can be subtracted from $\mathbf{y}_i$ at node i. This self-interference subtraction is practically feasible. (It is very different from the subtraction of $\mathbf{H}_{i,i}\mathbf{x}_i$ from $\mathbf{y}_i$ at a full-duplex relay where $\mathbf{y}_i$ is the input, $\mathbf{x}_i$ the output, and the power of interference is much higher than that of desired signal.) Therefore, after the interference subtractions, we can replace $\mathbf{y}_1$ and $\mathbf{y}_3$ by

$$\mathbf{y}_1' = \mathbf{H}_{1,2}\mathbf{F}_2\mathbf{H}_{2,3}\mathbf{x}_3 + \mathbf{H}_{1,2}\mathbf{F}_2\mathbf{w}_2 + \mathbf{w}_1 \quad (10)$$

$$\mathbf{y}_3' = \mathbf{H}_{3,2}\mathbf{F}_2\mathbf{H}_{2,1}\mathbf{x}_1 + \mathbf{H}_{3,2}\mathbf{F}_2\mathbf{w}_2 + \mathbf{w}_3 \quad (11)$$

Since $\mathbf{y}_1'$ and $\mathbf{y}_3'$ are correlated with each other due to the noise $\mathbf{w}_2$, the optimal design of $\mathbf{Q}_i^{(t)}$ and $\mathbf{D}_i$, $i = 1, 3$, must be done jointly even if $\mathbf{F}_2$ is fixed. However, as shown in [12], for any fixed $\mathbf{F}_2$, the optimal design of $\mathbf{Q}_i^{(t)}$ and $\mathbf{D}_i$, $i = 1, 3$, to maximize the sum rate received at nodes 1 and 3 subject to the power constraint $Tr(E\{\mathbf{x}_i\mathbf{x}_i^H\}) \leq P_i, i = 1, 3$, can be done by following the GWF algorithm [2]. Note that $E\{\mathbf{x}_i\mathbf{x}_i^H\} = \mathbf{Q}_i^{(t)}\mathbf{D}_i\mathbf{Q}_i^{(t)H}$, $i = 1, 3$.

For fixed $\mathbf{Q}_i^{(t)}$ and $\mathbf{D}_i$, $i = 1, 3$, the problem of finding $\mathbf{F}_2$ to maximize the sum rate subject to the power constraint $Tr(E\{\mathbf{x}_2\mathbf{x}_2^H\}) \leq P_2$ is a non-convex problem. Only locally optimal solutions for $\mathbf{F}_2$ can be found. In [12], an iterative LMMSE algorithm and a hybrid gradient algorithm for computing $\mathbf{F}_2$ are presented and compared. In particular, it can be shown that if we have the QR decompositions $[\mathbf{H}_{2,3}, \mathbf{H}_{2,1}] = \mathbf{Q}_a\mathbf{R}_a$ and $[\mathbf{H}_{1,2}^H, \mathbf{H}_{3,2}^H] = \mathbf{Q}_b\mathbf{R}_b$, then the optimal $\mathbf{F}_2$ can be expressed as $\mathbf{F}_2 = \mathbf{Q}_b\mathbf{A}_2\mathbf{Q}_a^H$ where $\mathbf{F}_2 \in C^{n_2^{(t)} \times n_2^{(r)}}$, $\mathbf{Q}_a \in C^{n_2^{(r)} \times r_a}$, $\mathbf{Q}_b \in C^{n_2^{(t)} \times r_b}$, $r_a = rank(\mathbf{H}_{2,3}, \mathbf{H}_{2,1})$ and $r_b = rank(\mathbf{H}_{1,2}^H, \mathbf{H}_{3,2}^H)$. If $r_b \times r_a < n_2^{(t)} \times n_2^{(r)}$, the above result reduces the search space. The results in [12] extends those in [13], [14] and [15] from single antenna at nodes 1 and 3 to multiple antennas.

Although each of the three nodes in Fig. 3 can be generalized to a cluster of sub-nodes, the two-hop nature of the NH2 system can not be generalized to more than two hops. For a network of many hops, the NH2 system does not seem a good relay strategy.

Also illustrated in Fig. 3 is a RH2 system where node 2 is regenerative. For the RH2 system, we need two time slots (instead of two frequency bands) for nodes 1 and 3 to exchange information. In the first time slot, it is like a conventional multiple access (MAC) for which successive interference cancelation (SIC) can be used at node 2. In the second time slot, it is like a conventional broadcast (BC) scheme for which (vector) dirty paper coding (DPC) can be used at node 2. For SIC and DPC, see [1] and the references therein. The spectral efficiency of the system is $C_{RH2} = \frac{1}{2}\min(C_{MAC}, C_{BC})$. Because of the regenerative nature of each node, the RH2 system can be applied to the case of many hops.

If the MAC and BC of the RH2 system are implemented within a single time and single frequency, we have a RF2 system. With multiple antennas at node 2, it is feasible to construct the first layer beamformers $\mathbf{Q}_2^{(t,1)}$ and $\mathbf{Q}_2^{(r,1)}$ at node 2 such that the self-interference at node 2 is canceled. See (7). The design of the second layer beamforming and power allocation can be done as if the system is RH2.

The same beamformers $\mathbf{Q}_2^{(t,1)}$ and $\mathbf{Q}_2^{(r,1)}$ also apply to the NF2 system in a similar way. And the design procedure of the remaining beamforming and power allocation at all three nodes becomes the same as for NH2.

7. REFERENCES

[1] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, 2005.

[2] Y. Yu and Y. Hua, "Power allocation for a MIMO relay system with multiple-antenna users", *IEEE Transactions on Signal Processing*, Vol. 58, No. 5, pp. 2823–2835, May 2010.

[3] Y. Rong and Y. Hua, "Optimality of diagonalization of multi-hop MIMO relays," *IEEE Transactions on Wireless Communications*, Vol. 8, No. 12, pp. 6068–6077, Dec. 2009.

[4] Y. Rong, X. Tang and Y. Hua, "A unified framework for optimizing linear non-regenerative multicarrier MIMO relay communication systems," *IEEE Transactions on Signal Processing*, Vol. 57, No. 12, pp. 4837–4852, Dec 2009.

[5] Z. Fang, Y. Hua, and J. Koshy, "Joint source and relay optimization for a non-regenerative MIMO relay," *Proc. of IEEE Workshop Sensor Array Multi-Channel Processing*, Waltham, MA, Jul. 2006.

[6] X. Tang and Y. Hua, "Optimal design of non-regenerative MIMO wireless relays", *IEEE Transactions on Wireless Communications*, Vol. 6, No. 4, pp. 1398-1407, April 2007.

[7] H. Ju, E. Oh, D. Hong, "Improving efficieny of resource usage in two-hop full duplex relay systems based on resource sharing and interference cancellation," *IEEE Transactions on Wireless Communications*, Vol. 8, No. 8, pp. 3933–3938, Aug. 2009.

[8] D. W. Bliss, P. A. Parker, A. R. Margetts, "Simultaneous transmission and reception for improved wireless network performance," *IEEE Workshop on Statistical Signal Processing*, pp. 478–482, Aug. 2007.

[9] K. M. Nasr, J. P. Cosmas, M. Bard, and J. Gledhill, "Performance of an echo canceller and channel estimator for on-channel repeaters in DVB-T/H networks", *IEEE Transactions on Broadcasting*, Vol. 53, No. 3, pp. 609–618, Sept 2007.

[10] B. Chun, E.-R. Jeong, J. Joung, Y. Oh and Y. H. Lee, "Pre-nulling for self-interference supression in full-duplex relays," *Annual Summit of Asia-Pacific Signal and Information Processing Association*, 2009.

[11] T. Riihonen, S. Werner, and R. Wichman, "Spatial loop interference suppression in full-duplex MIMO relays," *Asilomar Conference on Signals, Systems and Computers*, Monterey, CA, Nov. 2009.

[12] S. Xu and Y. Hua, "Source-relay optimization for a two-way MIMO relay system," *Proc. of IEEE ICASSP*, Dallas, TX, March 2010.

[13] B. Rankov and A. Wittneben, "Spectral efficient protocols for half-duplex fading relay channels," *IEEE Journal on Selected Area In Communications*, vol. 25, no. 2, Feb 2007.

[14] N. Lee, H. J. Yang, and J. Chun, "Achievable sum-rate maximizing af relay beamforming scheme in two-way relay channels," *IEEE International Conference on Communications Workshops*, May 2008.

[15] R. Zhang, Y.-C. Liang, C. C. Chai, and S. Cui, "Optimal beamforming for two-way multi-antenna relay channel with analogue network coding," *IEEE J. Sel. Areas Commun.*, Vol. 27, No. 5, pp. 699-712, June 2009.