UNIVERSITY OF CALIFORNIA
IRVINE

CENTER FOR PERVASIVE COMMUNICATIONS AND COMPUTING

GRADUATE FELLOWSHIP PROJECTS PROGRESS REPORTS SPRING 2008

PROJECTS

ALPHABETIZED ACCORDING TO STUDENT LASTNAME

STUDENT NAME	PROJECT TITLE	ADVISOR
LUAY AZZAM	DECODING COMPLEXITY OF QUASI-ORTHOGONAL SPACE-TIME CODES	ENDER AYANOGLU
ALIREZA S. BEHBAHANI	OPTIMIZING MIMO RELAY NETWORKS	AHMED ELTAWIL
VIVEK CADAMBE	OPTIMAL INTERFERENCE ALIGNMENT FOR WIRELESS NETWORKS	SYED A. JAFAR
FATEMEH FAZEL	PRACTICAL CODE DESIGN FOR RECONFIGURABLE MIMO SYSTEMS	HAMID JAFARKHANI
MINAS GJOKA	NOVEL USES AND MISUSES OF PEER-TO-PEER SYSTEMS	ATHINA MARKOPOULOU
AMIN JAHANIAN	AN ULTRAHIGH-SPEED ANALOG-TO-DIGITAL CONVERTER (ADC) IN 130NM CMOS	PAYAM HEYDARI
MAHYAR KARGAR	ADAPTIVE EQUALIZATION OF MULTIMODE FIBER CHANNELS IN 0.13µM CMOS	MICHAEL GREEN
AMIN KHAJEH DJAHROMI	CHALLENGES AND OPPORTUNITIES IN DESIGNING ULTRA EFFICIENT COMMUNICATION SYSTEMS IN HIGHLY SCALED TECHNOLOGIES	AHMED ELTAWIL
SEUNG EUN LEE	POWER-AWARE INTERCONNECTION NETWORK FOR WIRELESS MOBILE SYSTEMS	NADER BAGHERZADEH
MOHAMMAD A. MAKHZAN	FAULT TOLERANT VOLTAGE SCALABLE CACHES	FADI KURDAHI
NEGAR NEJATI	OPTIMAL TRANSMISSION OF PACKETIZED VIDEO OVER TANDEM CHANNEL	HAMID JAFARKHANI
HULYA SEFEROGLU	NETWORK CODING FOR PRACTICAL APPLICATIONS OVER WIRELESS MESH NETWORKS	ATHINA MARKOPOULOU
SRINIVASAN SURENDRAN	CAN INTERFERENCE ALIGNMENT DOUBLE THE CAPACITY OF AN INTERFERENCE NETWORK WITH CHANNEL UNCERTAINTY AND LIMITED POWER?	SYED A. JAFAR

An Efficient Tree Search for Reduced Complexity Sphere Decoding

Luay Azzam and Ender Ayanoglu

Center for Pervasive Communications and Computing
Department of Electrical Engineering and Computer Science
The Henry Samueli School of Engineering
University of California, Irvine
email: lazzam@uci.edu, ayanoglu@uci.edu

***Abstract*— The complexity of sphere decoding (SD) has been widely studied due to the importance of this algorithm in obtaining the optimal Maximum Likelihood (ML) performance with lower complexity. In this paper, we propose a proper tree search traversal technique that reduces the overall SD computational complexity without sacrificing the performance. We exploit the similarity among the complex symbols in a square QAM lattice representation and rewrite the squared norm ML metric in a simpler form allowing significant reduction of the number of operations required to decode the transmitted symbols. We also show that this approach achieves $> 45\%$ complexity gain for systems employing 4-QAM, and that this gain becomes bigger as the constellation size is larger.**

I. Introduction

The use of multiple antennas at both the transmitter and the receiver provides high capacity gains without increasing the bandwidth or transmitted power. Thus, multiple-input multiple-output (MIMO) systems have attracted much attention and serious research interests. Improving the performance is the main challenge of receiver design. Therefore, a number of decoding algorithms with different complexity-performance tradeoffs can be used. For instance, linear detection methods such as a zero-forcing (ZF) or minimum mean squared error (MMSE) have linear complexity but provide suboptimal performance. Ordered successive interference cancelation decoders, on the other hand, such as the vertical Bell Laboratories layered space-time (V-BLAST) algorithm can also be used which provide suboptimal performance with higher complexity compared to ZF and MMSE [1]. The optimum detection method is the Maximum Likelihood (ML) detection. However, in MIMO systems, ML algorithm has an exponential complexity with the constellation size and the number of antennas [2]. Therefore, sphere decoding (SD) [3] was proposed as an alternative for ML that provides optimal performance with reduced computational complexity [4], [5].

Different techniques have been proposed in the literature to reduce the complexity of SD. Among these, the increased radius search (IRS) [6] and the improved increasing radius search (IIRS) [7] are noteworthy. These two techniques attempt tackling this complexity by making a proper choice of the sphere radius. Other methods such as the application of the K-best lattice decoder [8] or a combination of the SD and the K-best decoder [9] were used where the complexity reduction came at the cost of a BER performance degradation.

In this paper, we improve the SD complexity efficiency by reducing the number of computations required to obtain the ML optimal solution. This complexity reduction is accomplished by exploiting the similarity between the complex symbols in a square QAM lattice representation. The proposed technique is generic and can be used with depth-first and breadth-first sphere decoders. Moreover, we show that the conventional ML metric can be rewritten in a simpler form which can be also used for reduced complexity ML decoding. In our previous work presented in [10], we proposed a new lattice representation that enables decoding the real and imaginary parts of each complex symbol independently. In this work, we use that same lattice representation and show that a complexity gain of at least 45% is obtained without sacrificing the performance.

The remainder of this paper is organized as follows: In Section II, a problem definition is introduced and a brief review of the conventional SD algorithm is presented. In Section III, we propose the new lattice representation and perform the mathematical derivations for complexity reduction. Performance and complexity comparisons for different number of antennas or modulation schemes are included in Section IV. Finally, we conclude the paper in Section V.

II. Conventional Sphere Decoder

Consider a MIMO system with N transmit and M receive antennas. The received signal at each instant of time is given by

$$y = Hs + v \tag{1}$$

where $y \in \mathbb{C}^M$, $H \in \mathbb{C}^{M \times N}$ is the channel matrix, $s \in \mathbb{C}^N$ is an N dimensional transmitted complex vector whose entries have real and imaginary parts that are integers, $v \in \mathbb{C}^M$ is the i.i.d complex additive white Gaussian noise (AWGN) vector with zero-mean and covariance matrix $\sigma^2 I$. Usually, the elements of the vector s are constrained to a finite set Ω where $\Omega \subset \mathbb{Z}^{2N}$, e.g., $\Omega = \{-3, -1, 1, 3\}^{2N}$ for 16-QAM where $\mathbb{Z}$ and $\mathbb{C}$ denote the sets of integers and complex numbers respectively.

Assuming H is known at the receiver, the ML detection is carried out as

$$\hat{s} = \arg \min_{s \in \Omega^N} ||y - Hs||^2. \quad (2)$$

Solving (2) is well known to be NP-hard since a full search over the entire lattice space is performed [11]. Consequently, SD was proposed where (2) is being solved by searching only those lattice points that lie inside a sphere of radius d centered around the received vector y.

A frequently used solution for the QAM-modulated complex signal model given in (1) is to decompose the N-dimensional complex-valued problem into a $2N$-dimensional real-valued problem, which then can be written as

$$\begin{bmatrix} \Re\{y\} \\ \Im\{y\} \end{bmatrix} = \begin{bmatrix} \Re\{H\} & -\Im\{H\} \\ \Im\{H\} & \Re\{H\} \end{bmatrix} \begin{bmatrix} \Re\{s\} \\ \Im\{s\} \end{bmatrix} + \begin{bmatrix} \Re\{v\} \\ \Im\{v\} \end{bmatrix}$$

where $\Re\{y\}$ and $\Im\{y\}$ denote the real and imaginary parts of y.

Assuming $N = M$ in the sequel, and introducing the QR decomposition of H, where R is an upper triangular matrix, and the matrix Q is unitary, (1) can be written as

$$\begin{aligned} y &= QRs + v \\ Q^H y &= Rs + Q^H v \\ \bar{y} &= Rs + \bar{v} \end{aligned} \quad (3)$$

where $\bar{v}$ and v have the same statistical properties since Q is unitary and so is Q^H. Then, SD solves

$$\hat{s} = \arg \min_{s \in \Omega^{2N}} ||\bar{y} - Rs||^2 < d^2 \quad (4)$$

The SD algorithm can be viewed as a pruning algorithm on a tree of depth $2N$, having branches that correspond to the elements drawn by the set Ω. This is shown in Figure 1 for a 16-QAM constellation. Note that pruned branches are shown as dashed.

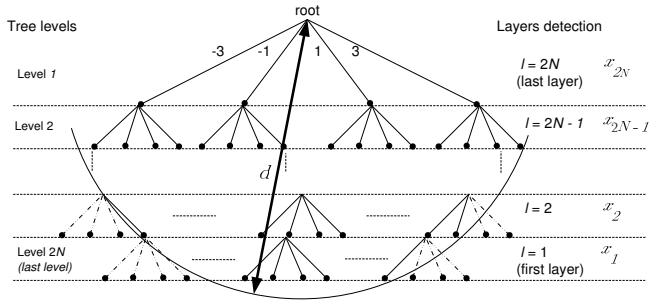


Fig. 1. Tree search example for a 16-QAM showing sphere radius, tree levels and detection layers.

Sphere decoding starts the search process from the root of the tree and works its way down along the branches until (4) is violated. This occurs when the total weight of a node exceeds the square of the sphere radius d^2. At this point, the corresponding branch is pruned and any path that passes through that node is declared as an improbable way to a candidate solution. Then, the algorithm backtracks and proceeds down a different branch. Whenever a valid lattice point at the bottom level of the tree is found within the sphere, the square of the sphere radius d^2 is set to the newly found point weight, thus reducing the search space for finding other candidate solutions. Finally, the path from the root to the leaf that is inside the sphere with the lowest weight is chosen to be the estimated solution $\hat{s}$.

III. Proposed Sphere Decoder

The complexity of SD is measured in terms of the number of operations required per visited node multiplied by the number of visited nodes through out the algorithm search [11]. This complexity can be reduced by either reducing the number of nodes to be visited or the number of operations to be carried out at each node or both. Making a good choice of the sphere radius to start the algorithm execution with, and thus reducing the number of visited nodes, has been widely studied in [6], [7] and the references therein. In this paper, we attempt to reduce the SD complexity by reducing the number of operations required at each node. We start by writing the node weight as [10]

$$\begin{aligned} w_l(x^{(l)}) &= w_p + w_{pw} \\ &= w_p + |\bar{y}_l - \sum_{k=l}^{2N} r_{l,k} x_k|^2 \end{aligned} \quad (5)$$

with $l = 2N, 2N-1, \ldots, 1$, w_p is the weight of the node's parent, w_{pw} is the partial weight of the node, and where $\{x_1, x_2, ..., x_N\}$, $\{x_{N+1}, x_{N+2}, ..., x_{2N}\}$ are the real and imaginary parts of $\{s_1, s_2, ..., s_N\}$ respectively.

This means that the total weight of a node is merely the weight of its parent summed with its partial weight represented by the second term in (5). This squared term can be simplified by exploiting the similarity between the symbols in a square QAM lattice structure. To make this clearer, let us consider a 16-QAM modulation scheme. Every node in the tree has 4 branches; each represents one of the real values given by $\Omega = \{-3, -1, 1, 3\}$. For a specific l, the computation of $|\bar{y}_l - \sum_{k=l}^{2N} r_{l,k} x_k|^2$ for $x_k = -3$ and $x_k = 3$ is very similar. Basically, all the terms are common except for the minus sign inside the summation. The same applies for $x_k = -1$ and $x_k = 1$. This similarity makes it possible to divide the set $\Omega = \{-3, -1, 1, 3\}$ into two smaller sets namely $\Omega_1 = \{-3, -1\}$ and $\Omega_2 = \{1, 3\}$, and enforce the algorithm to compute the weight of the nodes that correspond only to one set. These weights can then be reused to find the weight of the nodes in the other set. We provide the following example to explain this.

Example: Consider a MIMO system having $N = M = 2$ and employing 4-QAM modulation scheme. Then, SD constructs a tree with $2N = 4$ levels where the branches coming out from each node represent the real values in the set $\Omega = \{-1, 1\}$. This tree is shown in Figure 2. Now, using the real-valued lattice representation developed in [10], and applying the QR decomposition to the channel matrix, the

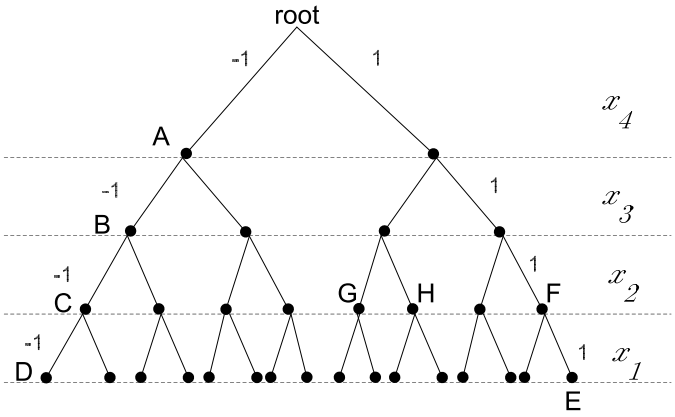


Fig. 2. Tree structure for 2×2 system employing 4-QAM.

input-output relation is given by

$$\begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \\ \bar{y}_4 \end{bmatrix} = \begin{bmatrix} r_{1,1} & 0 & r_{1,3} & r_{1,4} \\ 0 & r_{2,2} & r_{2,3} & r_{2,4} \\ 0 & 0 & r_{3,3} & 0 \\ 0 & 0 & 0 & r_{4,4} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} \bar{v}_1 \\ \bar{v}_2 \\ \bar{v}_3 \\ \bar{v}_4 \end{bmatrix}$$

Proposed SD starts by dividing the set Ω into two smaller sets $\Omega_1 = \{-1\}$ and $\Omega_2 = \{1\}$. Then using (5), it calculates the weights of nodes A, B, C, and D which correspond to $x_k \in \Omega_1$ for $k = 1, 2, 3, 4$. For e.g., D represents the solution $(x_1, x_2, x_3, x_4) = (-1, -1, -1, -1)$. Now these basic nodes act as other parents for the nodes that are on the same tree level. For example, the node E has now two parents which are the original parent F and the new parent D. For level k in the tree, we denote all branches in that level which share the same property with their new parents in having the same value of x_k, by $\tilde{x}_k$, and denote the others which have a different values than their parent nodes by $\check{x}_k$. For e.g., G represents $\tilde{x}_k = -1$, whereas H represents $\check{x}_k = 1$. By setting up the above definitions, (5) is no longer used and the weight of any node in the tree can be found using a less computational expression given by

$$\begin{aligned} w_l(x^{(l)}) =& w_p + w_{np} \\ &+ 4\left(\bar{y}_l - \sum_{\tilde{x}_k \in \Omega_1} r_{l,k}|\tilde{x}_k|\right)\left(\sum_{\check{x}_k \in \Omega_2} r_{l,k}|\check{x}_k|\right) \end{aligned} \quad (6)$$

where w_{np} is the weight of the new parent. Note that it is straight forward to obtain (6) by expanding the squared term in (5) and taking out the common terms between any node weight and its new parent weight.

Dealing only with the absolute values in the above expression reduces the number of required computations significantly. For 4-QAM, we have $|\tilde{x}_k| = |\check{x}_k|$, thus (6) can be rewritten as

$$\begin{aligned} w_l(x^{(l)}) =& w_p + w_{np} \\ &+ 4\left(\bar{y}_l - \sum_{\tilde{x}_k \in \Omega_1} r_{l,k}\right)\left(\sum_{\check{x}_k \in \Omega_2} r_{l,k}\right). \end{aligned} \quad (7)$$

Then the partial weight of any node, say E and H can be simply found as

$$w_E = w_D + 4\bar{y}_1(r_{1,1} + r_{1,3} + r_{1,4})$$

and,

$$w_H = w_C + 4(\bar{y}_2 - r_{2,3})(r_{2,2} + r_{2,4}).$$

Calculating the partial weight of node E using the conventional SD requires 25 real multiplications and 24 real additions, whereas these numbers reduce to 4 real multiplications and 3 real additions using the proposed technique. These savings even becomes bigger as the constellation size is larger. This is because the number of tree nodes increases exponentially with the constellation size. For e.g., the number of nodes for a 4×4 system and 16-QAM is 87380. This number increases to $19 \cdot 10^6$ for a 4×4 system and 64-QAM.

To this end, it is worth mentioning here that in the proposed technique, we basically rewrite the conventional ML metric in a simpler form. As a result, this new form can be used for conventional ML decoding as well. It is also important to emphasize the fact that the proposed SD provides the same performance as the conventional SD with much lower computational complexity. This complexity gain becomes bigger as the constellation size and the number of antennas are larger.

IV. Simulation Results

We have considered 2×2, 4×4 systems using 4-QAM, 16-QAM and 64-QAM modulation schemes. Since the multiplications are the most expensive operations in terms of machine cycles compared to additions, the complexity is measured in terms of the number of real multiplications required to decode the transmitted complex symbols. We use the real-valued lattice representation presented in [10] for the conventional and proposed SD. This means that the complexity gain shown in all the figures below is on top of the gain obtained in [10]. We denote the conventional SD by Conv and the proposed SD by PR. The radius d should be chosen properly so that it is not too small to result in an empty sphere and thus restarting the search, and at the same time, it should not be too large to increase the number of lattice points to be searched. We use the formula presented in [12] for the radius, which is $d^2 = 2\sigma^2 N$, where N is the problem dimension and σ^2 is the noise variance.

Figure 3 shows the complexity curves for both algorithms using 4-QAM. For 2×2 system, the complexity gain is 45%. This gain increases up to reach 60% for the 4×4 case.

Similarly, Figures 4 and 5 show complexity comparison using the same configuration with 16-QAM, and 64-QAM respectively. Again, the proposed SD achieves high complexity reduction compared to conventional SD. The gain ranges from

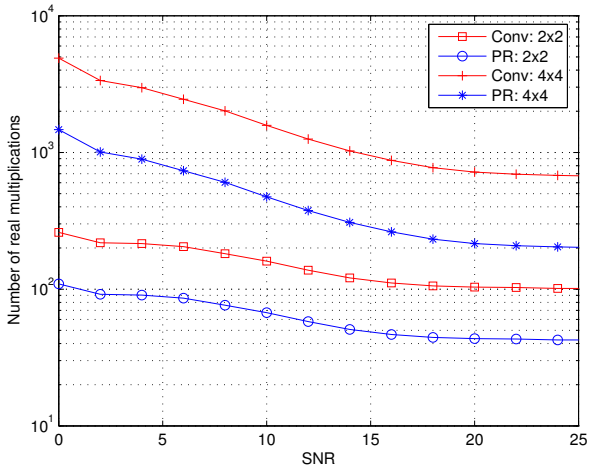


Fig. 3. # of real multiplications vs SNR for the proposed and conventional SD over a 2×2, and 4×4 MIMO flat fading channel using 4-QAM.

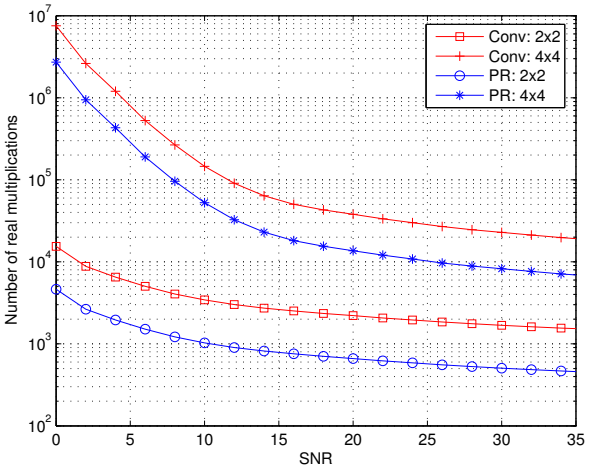


Fig. 5. # of real multiplications vs SNR for the proposed and conventional SD over a 2×2, and 4×4 MIMO flat fading channel using 64-QAM.

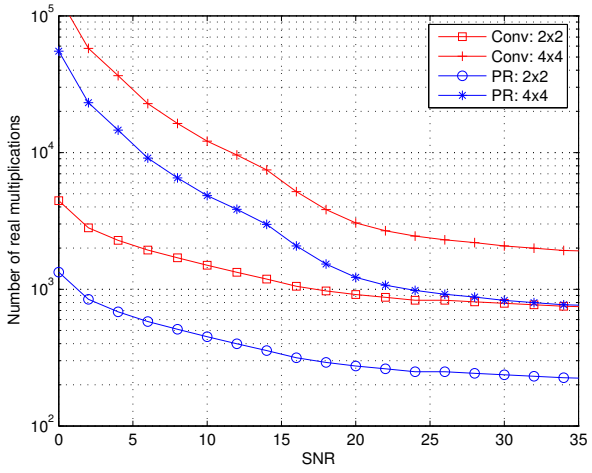


Fig. 4. # of real multiplications vs SNR for the proposed and conventional SD over a 2×2, and 4×4 MIMO flat fading channel using 16-QAM.

50% for 2×2 with 16-QAM, up to 65% for 4×4 and 64-QAM.

Finally, we should emphasize that these complexity gains are for free. This is because the proposed algorithm provides the same performance results as the conventional SD.

V. Conclusions

A simple and general tree search technique for sphere decoding is proposed in this paper. The performance of the proposed SD is the same as that of conventional SD. However, a significant complexity reduction compared to conventional SD is achieved. This complexity reduction is accomplished by exploiting the similarity between the complex symbols in a square QAM constellation, allowing for rewriting the conventional ML metric in a simpler form. This new form is also applicable to performing ML decoding with reduced computational complexity. The complexity gain ranges between 45% and 65% depending on the number of antennas and constellation size being used. Simulation results are provided for 2×2 and 4×4 systems employing 4-QAM, 16-QAM, and 64-QAM. We show that the gain achieved by using the proposed SD becomes bigger as the constellation size is larger.

References

[1] A. Burg, M. Borgmann, M. Wenk, M. Zellweger, W. Fichtner, and H. Bolcskei, "VLSI implementation of MIMO detection using the sphere decoding algorithm," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 1566–1577, July 2005.

[2] E. Zimmermann, W. Rave, and G. Fettweis, "On the Complexity of Sphere Decoding," in Proc. International Symp. on Wireless Pers. Multimedia Commun., Abano Terme, Italy, Sep. 2004.

[3] U. Fincke and M.Pohst, "Improved Methods for Calculating Vectors of Short Length in Lattice, Including a Complexity Analysis," *Mathematics of Computation*, vol. 44, pp. 463–471, April 1985.

[4] J. Jalden and B. Ottersten, "On the Complexity of Sphere Decoding in Digital Communications," *IEEE Trans. Signal Processing*, vol. 53, no. 4, pp. 1474–1484, April 2005.

[5] B. Hassibi and H. Vikalo, "On the Sphere-Decoding Algorithm I. Expected complexity," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2806–2818, August 2005.

[6] E. Viterbo and J. Boutros, "A Universal Lattice Code Decoder for Fading Channels," *IEEE Trans. Inform. Theory*, vol. 45, no. 5, pp. 1639–1642, July 1999.

[7] W. Zhao and G. Giannakis, "Sphere Decoding Algorithms With Improved Radius Search," *IEEE Trans. Commun.*, vol. 53, pp. 1104–1109, Jul. 2005.

[8] K. Wong, C. Tsui, R.-K. Cheng, and W. Mow, "A VLSI architecture of a K-best lattice decoding algorithm for MIMO channels," *in Proc. IEEE ISCAS02*, vol. 3, pp. 273–276, 2002.

[9] J. Tang, A. Tewfik, and K.K.Parhi, "Reduced Complexity Sphere Decoding and Application to interfering IEEE 802.15.3a piconets," *IEEE ICC*, vol. 5, pp. 2864–2868, June 2004.

[10] L. Azzam and E. Ayanoglu, "Reduced Complexity Sphere Decoding for Square QAM via a New Lattice Representation," *IEEE GLOBECOM*, pp. 4242–4246, Nov. 2007.

[11] B. Hassibi and H. Vikalo, "On The Expected Complexity of Integer Least-Squares Problems," *IEEE ICASSP*, pp. 1497–1500, 2002.

[12] G. Rekaya and J. Belfiore, "On the Complexity of ML Lattice Decoders for Decoding Linear Full-Rate Space-Time Codes," *in Proceedings 2003 IEEE International Symposium on Information Theory.*, July 2003.

Progress Report on "Relay Networks" Spring 2008

Alireza S. Behbahani, Advisor: Prof. Ahmed Eltawil
Electrical Engineering Department
University of California, Irvine, USA
Email: sshahanb@uci.edu

In Winter 2008 quarter, We designed and analyzed a training based linear mean square error (LMMSE) channel estimator for time division multiplex amplify-and-forward (AF) relay networks. We started with the scenario where relays have no knowledge of their backward and forward channels. In Spring 2008 quarter we continued that work and considered two more scenarios where relays know backward and forward channels perfectly, and relays estimate backward and forward channels. We could show that the scenario where each relay estimates its backward and forward channels is a general case for the other two scenarios.

We consider an AF relay network consisting of 1 transmit antenna at the source, 1 receive antenna at the destination, and K relays, each equipped with 1 antenna. We denote by h_{s_i} the channel between the source and relay i, backward channel, while h_{t_i} is the channel between the relay i and the destination, forward channel. We assume that the channels follow the block-fading law, where the channels are constant for some coherence interval T_c, which is measured in symbols, and after that they change to an independent value which hold for another interval T_c. we further assume that channel estimation and data transmission is to be done during the interval T_c. Also backward and forward channels are independent and Rayleigh flat fading distributed which are $h_{s_i},\ h_{t_i} \sim \mathcal{CN}(0, \sigma_h^2),\quad i = 1, \cdots, K$ and where for convenience we assumed that backward and forward channels have the same variance, σ_h^2.

In the first phase transmitter sends the signal block $\boldsymbol{s} = [s_1, \cdots, s_{T_c}]^T$ to the relays. The received signal at relay i can be modeled as

$$\boldsymbol{r}_i = h_{s_i}\boldsymbol{s} + \boldsymbol{v}_{s_i}, \tag{1}$$

where $\boldsymbol{r}_i = [r_{i_1}, \cdots, r_{i_{T_c}}]^T$ is the received signal and $\boldsymbol{v}_{s_i} = [v_{s_{i_1}}, \cdots, v_{s_{i_{T_c}}}]^T$ is the zero mean additive white complex Gaussian noise at the relay i with covariance matrix $\boldsymbol{R}_{\boldsymbol{v}_{s_i}} = \sigma_{v_s}^2 \boldsymbol{I}$. Also the transmitter has the total power $\mathsf{E}\boldsymbol{s}^*\boldsymbol{s} = T_c P_s$, where P_s is the average transmitting power of the source. In the second phase each relay, multiplies its received signal by a scalar coefficient β_i which is constant during coherence time T_c and sends it, on the same time slot, to the destination. The received signal at the destination can be expressed as

$$\boldsymbol{y} = \sum_{i=1}^{K} \beta_i \boldsymbol{r}_i h_{t_i} + \boldsymbol{v}_t, \tag{2}$$

where $\boldsymbol{v}_t$ is $T_c \times 1$ zero mean additive white complex Gaussian noise at the destination with covariance matrix $\boldsymbol{R}_{\boldsymbol{v}_t} = \sigma_{v_t}^2 \boldsymbol{I}$ and also independent of $\boldsymbol{v}_{s_i}$ for all i. By plugging (1) in (2) the received signal can be expressed as

$$\boldsymbol{y} = \underbrace{\sum_{i=1}^{K} \beta_i h_{s_i} h_{t_i}}_{h_{tot}} \boldsymbol{s} + \underbrace{\sum_{i=1}^{K} \beta_i h_{t_i} \boldsymbol{v}_{s_i} + \boldsymbol{v}_t}_{\boldsymbol{n}} = h_{tot}\ \boldsymbol{s} + \boldsymbol{n}, \tag{3}$$

where h_{tot} is the overall channel from the source to the destination and $\boldsymbol{n}$ is the overall noise at the destination which is zero mean and has the covariance matrix

$$\boldsymbol{R}_n = (\sigma_{v_s}^2 \sigma_h^2 \sum_{i=1}^{K} |\beta_i|^2 + \sigma_{v_t}^2)\boldsymbol{I} = \sigma_n^2 \boldsymbol{I}\,. \tag{4}$$

Since the receiver does not know h_{tot}, training-based schemes assign part of the transmitted signal $\boldsymbol{s}$ to be a known training signal from which the receiver can learn h_{tot}. Here we investigate three different scenarios and investigate the effect of relay functionality on the overall channel estimation.

A. *Relays estimate backward and forward channels*

In this scenario each relay needs to estimate its backward, h_{s_i}, and forward, h_{t_i} channels by using training symbols. Since we are assuming that the relay network is a time division multiplex system, the channel between relays and destination, h_{t_i}, are reciprocal. In order to estimate the forward channel, after the transmitter sends control signals to the destination, the transmission starts from the receiver side by sending some training symbols to the relays, such that relays can estimate their forward channel. After that the transmitter starts sending training symbols to the destination through relays. The backward channel can be estimated at each relay by the same training symbols which is sent to the destination. After training phase the transmitter starts sending data information. In this scenario the scaling factor at each relay is defined as

$$\beta_i = \sqrt{\frac{P_r}{\sigma_h^2 P_s + \sigma_{v_s}^2}}\ e^{-j[\angle \hat{h}_{s_i} + \angle \hat{h}_{t_i}]}\,, \tag{5}$$

where $\hat{h}_{s_i}$ and $\hat{h}_{t_i}$ are estimated backward and forward channels respectively, and P_r is the output power of each relay. Having a good estimation of backward and forward channels will result in constructive addition of signals at the destination and hence higher SNR. The expression $\sqrt{\frac{P_r}{P_s+\sigma_{v_s}^2}}$ adjusts the output power of relays from long term point of view. Also conservation of time yields

$$T_c = 2T_\tau + T_d \ . \tag{6}$$

B. *Relays know backward and forward channels perfectly*

In this scenario, each relay knows its backward and forward channels perfectly and there is no need for training symbols from the receiver side. Here the transmission starts by sending training symbols from the transmitter, and relays just scale and forward the training symbols to the destination. The scaling factor is given by

$$\beta_i = \sqrt{\frac{P_r}{\sigma_h^2 P_s + \sigma_{v_s}^2}}\ e^{-j[\angle h_{s_i} + \angle h_{t_i}]} \ , \tag{7}$$

and

$$T_c = T_\tau + T_d \ . \tag{8}$$

C. *Relays do not have knowledge of channels*

In this scenario, relays just scale and forward its received signal to the destination and they do not have any knowledge of the backward and forward channels. Here, as in B, the transmission starts by sending training symbols from the transmitter to the destination through relays and there is no need for training symbols from the receiver and also estimation at each relay. Therefore,

$$\beta_i = \sqrt{\frac{P_r}{\sigma_h^2 P_s + \sigma_{v_s}^2}} \ , \quad \text{and} \quad T_c = T_\tau + T_d \ . \tag{9}$$

In order to investigate the performance of our proposed channel estimation We define the signal to noise ratio as SNR $= P_s/\sigma_{v_s}^2$, where $\sigma_{v_s}^2 = \sigma_{v_t}^2$ and also $\sigma_h^2 = 1$. We further assume that the transmitter output power, P_s, and the relays average output power, P_r, are set to 10 dB.

Figure 1 shows channel estimation mean square error (MSE) for the case that relays do not have knowledge of channels for coherence interval of $T_c = 300$ and training interval of $T_\tau = 10$. It can be seen that increasing the number of relays increases the MSE which is consistent with mathematical analysis where it says that estimation error variance increases with K.

Figure 2 compares MSE channel estimation of different scenarios for the same setting as figure 1 except that here $K = 10$. It can be seen that the estimator performs the same for different schemes except at low SNR which by having knowledge of channels we can get around 0.5 dB gain.

We could show that the scenario where each relay estimates its backward and forward channels is a general case for the other two scenarios.

We are going to continue this work by finding a lower bound for the capacity considering the effect of training and estimation error.

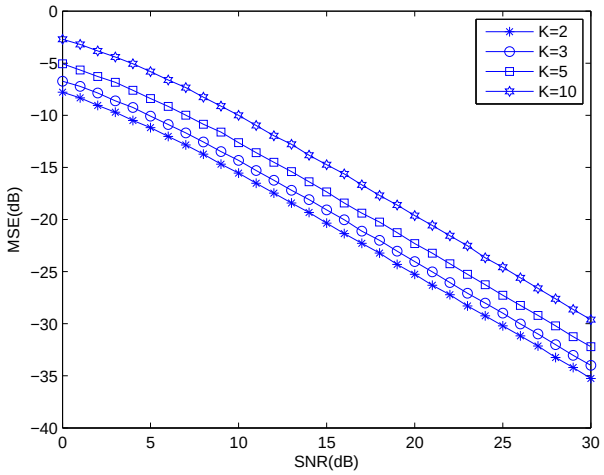


Fig. 1. MSE in dB versus SNR for $T_c = 300$, and $T_\tau = 10$ for different number of relays. Here relays do not have knowledge of their backward and forward channels.

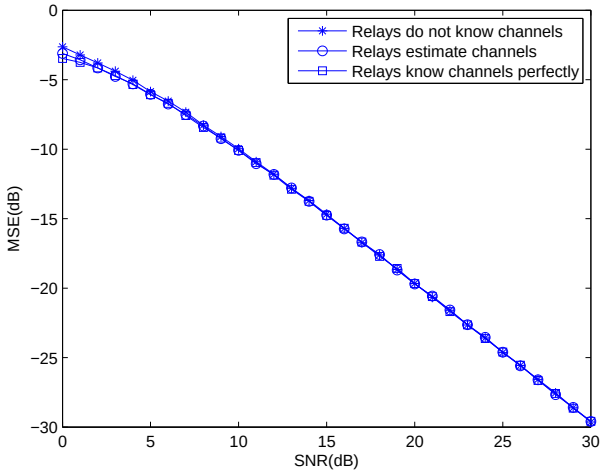


Fig. 2. Comparison of MSE for different scenarios with the same setting of figure 1 except that here $K = 10$.

Progress in Spring 2008 : On the Generalized Degrees of Freedom of Interference Networks with Feedback

Viveck R. Cadambe, Syed A. Jafar
Center For Pervasive Communications and Computing Electrical Engineering and Computer Science
University of California Irvine,
Irvine, California, 92697, USA
Email: vcadambe@uci.edu, syed@uci.edu

***Abstract*— In this report, we study the generalized degrees of freedom of the interference channel with feedback. In particular, we find out the exact characterization of the generalized degrees of freedom of the strong interference channel and bounds for the weak interference channel with feedback. Interestingly, we find that feedback can improve the generalized degrees of freedom of the interference channel, in both the strong and weak interference regimes.**

I. Introduction

In Winter 2008, we discovered that the techniques of feedback, relays, full duplex operation and noisy cooperation do not increase the degrees of freedom of fully connected wireless networks [1]. In other words, the degrees of freedom tool is too coarse to capture the benefits of feedback on wireless networks. In this work, we study the positive effects of feedback on the interference network using the *generalized* degrees of freedom tool. The generalized degrees of freedom of a network tool was used in [2] to identify the various operating regimes of an interference network. In this work, we discover that feedback can improve the generalized degrees of freedom of both strong and weak interference channels. The model and the main results are presented in the next section.

II. System Model

We only consider symmetric interference channel in this study (Fig. 1). The input-output relations of this channel are described as

$$Y_1(\tau) = \sqrt{\text{SNR}}\, H_{11}X_1(\tau) + \sqrt{\text{INR}}\, H_{12}X_2(\tau) + Z_1(\tau)$$

$$Y_2(\tau) = \sqrt{\text{INR}}\, H_{21}X_1(\tau) + \sqrt{\text{SNR}}\, H_{22}X_2(\tau) + Z_2(\tau)$$

where at the τ^{th} channel use, $X_i(\tau)$ represents the complex symbol transmitted by transmitter i. $Y_i(\tau)$ and $Z_i(\tau)$ respectively represent the received symbol and the zero-mean unit-variance AWGN symbol at receiver i. H_{ij} represents the phase of the channel gain from transmitter j to receiver i, i.e, it satisfies $||H_{ij}||^2 = 1$, For a code spanning T uses of the channel, the codeword transmitted by transmitter i satisfies an average power constraint that may be expressed as $\frac{1}{T}E\left[\sum_{\tau=1}^{T}||X_i(\tau)||^2\right] \leq 1, i = 1, 2$. This power constraint assumption ensures SNR represents the actual signal-to-noise ratio between transmitter i and receiver i.

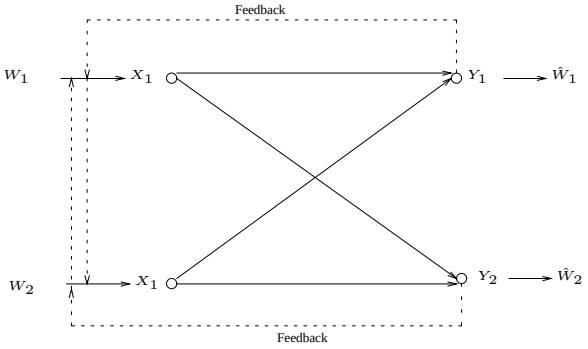


Fig. 1. The 2 user interference channel with feedback

A. Generalized Degrees of Freedom

Let

$$\alpha = \frac{\log(\text{INR})}{\log(\text{SNR})}$$

Equivalently, $\text{INR} = \text{SNR}^{\alpha}$. The *sum*-capacity of the interference channel is represented by $C_{\Sigma}(\alpha, \text{SNR})$ Then, the generalized degrees of freedom (GDOF) of the interference channel $d(\alpha)$ is defined as

$$d(\alpha) = \lim_{\text{SNR}\to\infty} \frac{C_{\Sigma}(\alpha, \text{SNR})}{\log(\text{SNR})}$$

III. Main Results

When there is no feedback, the various operating interference regimes in the *symmetric* interference channel identified in [2] as

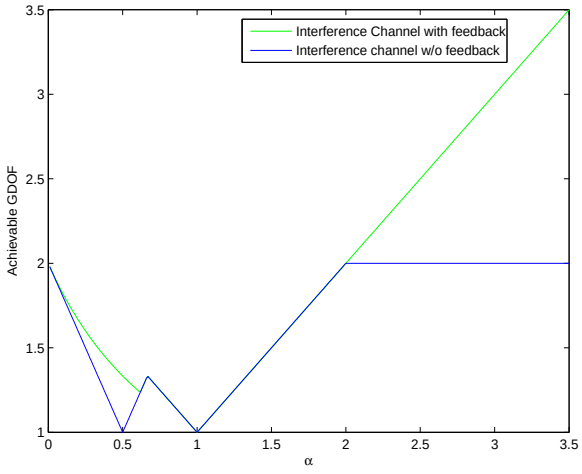


Fig. 2. Achievable GDOF of the interference channel - the effect of feedback

1) Very weak interference ($0 < \alpha \leq 2/3$) : $d(\alpha) = 2\max(\alpha, 1-\alpha)$
2) Moderately weak interference ($2/3 < \alpha \leq 1$) : $d(\alpha) = 2 - \alpha$
3) Moderately strong interference ($1 < \alpha \leq 2$) : $d(\alpha) = \alpha$
4) Very strong interference ($2 < \alpha$) : $d(\alpha) = 2$

We find the exact generalized degrees of freedom characterization of the interference channel with feedback for $\alpha > 2/3$ in the Theorem below

Theorem 1: For $\alpha \geq 2/3$, the generalized degrees of freedom of the interference channel with feedback can be characterized as

$$d(\alpha) = \max(2-\alpha, \alpha)$$

Theorem 2: For $\alpha < 2/3$, the generalized degrees of freedom of the interference channel with feedback can be bounded as

$$\max(\alpha, \frac{2}{1+\alpha}) < d(\alpha) < 2 - \alpha$$

We refer the reader to the extended paper [3] for the proofs. The results of Theorem 1 and Theorem 2 imply that for $\alpha > 2$ and for $\alpha \leq 2/3$, feedback improves the GDOF of the interference channel (See Figure 2).

IV. Conclusion

We found a tight characterization of the GDOF of the symmetric interference channel with feedback for $\alpha \geq 2/3$ and bounds for the GDOF for $\alpha \leq 2/3$. In the interference channel without feedback, the GDOF characterization leads to an approximation of its capacity within one bit. Pursuit of approximations of the interference channel with feedback within a constant number of bits is an interesting area of future work.

References

[1] V. R. Cadambe and S. A. Jafar, "Can feedback, cooperation, relays and full duplex operation increase the degrees of freedom of wireless networks ?," *Information Theory, 2008 IEEE International Symposium on*, July 2008.

[2] R. Etkin, D. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *arxiv.org*, Feb 2007. eprint - cs/0702045.

[3] V. R. Cadambe and S. A. Jafar, "On the generalized degrees of freedom of the interference channel with feedback," 2008. Under Preparation.

CPCC Fellowship Progress Report, Spring 2008:

Space-Time-State Block Coded MIMO Communication Systems Using Reconfigurable PIXEL Antennas: Simulation Results

Fatemeh Fazel, Advisor: Hamid Jafarkhani
Center for Pervasive Communications and Computing
EECS, University of California, Irvine
{fazel,hamidj}@uci.edu

I. Introduction

In this report, we consider a MIMO channel model using reconfigurable PIXEL antennas at the receiver. We take into account the practical issues pertaining to channel propagation. The goal is to study the performance of a practical reconfigurable MIMO system using the state-selection scheme proposed in [1].

II. PIXEL Antenna

We assume a system using reconfigurable PIXEL antennas over correlated channels. The PIXEL antenna is a multifunctional MEMS-reconfigurable radiator that works based on the principle that different radiation modes and operating frequencies can be excited by changing the dimensions (radius) of a circular patch and the relative location of the feed line within the patch. The PIXEL antenna can produce up to five distinct radiation patterns (states), which correspond to the modes $n = 1$, $n = 1$ with a rotation of $\phi_0 = 90^\circ$, $n = 2$, $n = 2$ with a rotation of $\phi_0 = 45^\circ$, and $n = 0$ of a circular patch antenna.

III. Simulation Setup

For this set of simulations, we use a RX-reconfigurable 2×1 system, with $M_T = 2$ transmit antennas, composed of two dipole antennas separated by a distance of 0.5λ, and $M_R = 1$ reconfigurable PIXEL antenna at the receiver. The radiation patterns of the PIXEL and the dipole antennas are computed using HFFS electromagnetic software. In this configuration $\Psi = Q = 5$. We investigate the performance of the abovementioned system on three distinct NLOS propagation scenarios, which are described through the characteristics of the receive power azimuth spectrum, as follows:

1) case 1: single cluster with mean angle $\phi_c = 0^\circ$ and variance $\sigma_\phi = 5^\circ$.
2) case 2: single cluster with mean angle $\phi_c = 0^\circ$ and variance $\sigma_\phi = 60^\circ$.
3) case 3: single cluster with mean angle $\phi_c = 0^\circ$ and variance $\sigma_\phi = 220^\circ$.

where ϕ_c denotes the mean angle of arrival and σ_ϕ denotes the variance. For the transmit power azimuth spectrum, we always assume a mean angle of $\phi_c = 0^\circ$ and a variance of $\sigma_\phi = 220^\circ$. Using the radiation patterns of the PIXEL antenna, one can calculate the channel correlation properties in the above propagation scenarios. Then, one can incorporate these correlation characteristics into the simulations, as noted by

$$\underline{\mathcal{H}} = \sqrt{g}\mathcal{R}_{\underline{\mathcal{H}}}^{1/2}\underline{\tilde{\mathcal{H}}} \tag{1}$$

where, $\underline{\tilde{\mathcal{H}}}$ is a zero-mean i.i.d complex vector with covariance $\mathcal{E}\{\underline{\tilde{\mathcal{H}}}\underline{\tilde{\mathcal{H}}}^H\} = I_{M_T M_R P}$, and having into consideration that $\mathcal{R}_{\underline{\mathcal{H}}} = \mathbf{R}^{\mathbf{R}} \otimes \mathbf{R}^{\mathbf{T}}$. For the PIXEL antenna, the intra and inter-state receive correlation matrices $\mathbf{R}^R_{q_1 q_2}$ for $q_1, q_2 \in \{1, \ldots, 5\}$ for the three abovementioned cases are given by Eqs. (2)-(4) at the top of next page. Notice that throughout these simulations, we have assumed that only the energy of the channel in case 1 is equal to $M_R M_T PQ$, thus $g_{case1} = \frac{M_R M_T PQ}{tr(\mathbf{R}^{\mathbf{R}}_{\mathbf{case1}} \otimes \mathbf{R}^{\mathbf{T}})}$, where $\mathbf{R}^{\mathbf{T}}$ is given by Eq. (5). We assign $g_{case3} = g_{case2} = g_{case1}$ in order to preserve different levels of received power not only among distinct radiation states but also among different simulation scenarios. Note that we use the Space-Time-State Block Code (STS-BC) for $\psi = 5$ which is given in [1]. The optimal rotation angles for the STS-BC-5, using BPSK, are derived by exhaustive search using a step size of $\pi/32$. The optimal values, which maximize the coding gain, are found to be $\theta_1 = 0.5890, \theta_2 = 1.1781, \theta_3 = 2.4544$, and $\theta_4 = 1.8653$.

IV. Simulation Results

Fig. 1 depicts the BER vs. SNR performance of the reconfigurable MIMO system for cases 1, 2 and 3 discussed above, employing both STS-BC-1 and STS-BC-5. Note that STS-BC-1 is equivalent to Alamouti structure. We perform state-selection at the receiver by selecting the state providing the largest SNR. Note that we use the notation (1,5) in the figure to refer to selecting 1 state out of a possible $Q = 5$ states. The maximum possible diversity gain offered by the system is determined by $\mathrm{rank}\{\mathcal{R}_{\underline{\mathcal{H}}}\} = \mathrm{rank}\{\mathbf{R}^{\mathbf{R}} \otimes \mathbf{R}^{\mathbf{T}}\}$. From Fig. 1, we notice that for STS-BC-1, the angular spread impacts the level of diversity obtained. As the angular spread

$$\mathbf{R}^{R}_{case1} = 10000 \begin{bmatrix} 0.4310 & -0.0475 - 0.0635i & 0.1225 + 0.4856i & 0.6488 - 0.0387i & -0.1486 - 0.1467i \\ -0.0475 + 0.0635i & 0.3427 & -0.1036 - 0.0105i & -0.1090 + 0.1466i & -0.1338 + 0.3518i \\ 0.1225 - 0.4856i & -0.1036 + 0.0105i & 0.6344 & 0.1785 - 0.7929i & -0.1875 + 0.1248i \\ 0.6488 + 0.0387i & -0.1090 - 0.1466i & 0.1785 + 0.7929i & 1.0673 & -0.1556 - 0.2694i \\ -0.1486 + 0.1467i & -0.1338 - 0.3518i & -0.1875 - 0.1248i & -0.1556 + 0.2694i & 0.5928 \end{bmatrix}, \quad (2)$$

$$\mathbf{R}^{R}_{case2} = 1000 \begin{bmatrix} 5.6877 & -1.3503 - 0.3958i & 0.2965 + 2.8659i & 2.8735 - 0.9275i & -1.1063 - 1.5173i \\ -1.3503 + 0.3958i & 3.5267 & -0.6809 - 0.1700i & -0.3993 + 0.9495i & -0.3994 + 2.9586i \\ 0.2965 - 2.8659i & -0.6809 + 0.1700i & 4.4298 & 0.3655 - 3.4931i & -0.5211 + 0.6223i \\ 2.8735 + 0.9275i & -0.3993 - 0.9495i & 0.3655 + 3.4931i & 6.0393 & -0.4637 - 1.5262i \\ -1.1063 + 1.5173i & -0.3994 - 2.9586i & -0.5211 - 0.6223i & -0.4637 + 1.5262i & 6.1639 \end{bmatrix}, \quad (3)$$

$$\mathbf{R}^{R}_{case3} = 1000 \begin{bmatrix} 4.0195 & -1.1617 - 0.2077i & 0.2182 + 0.5086i & 0.9848 - 0.6060i & -0.6712 - 0.8665i \\ -1.1617 + 0.2077i & 2.7355 & -0.1715 - 0.0221i & -0.1004 + 0.5175i & 0.4415 + 1.1542i \\ 0.2182 - 0.5086i & -0.1715 + 0.0221i & 2.8882 & 0.2675 - 0.8158i & 0.2820 + 0.2286i \\ 0.9848 + 0.6060i & -0.1004 - 0.5175i & 0.2675 + 0.8158i & 3.4968 & -0.0704 - 0.8536i \\ -0.6712 + 0.8665i & 0.4415 - 1.1542i & 0.2820 - 0.2286i & -0.0704 + 0.8536i & 4.3716 \end{bmatrix}. \quad (4)$$

$$\mathbf{R}^{T} = 1000 \begin{bmatrix} 7.6585 & -1.6245 + 0.0238i \\ -1.6245 - 0.0238i & 7.6194 \end{bmatrix}. \quad (5)$$

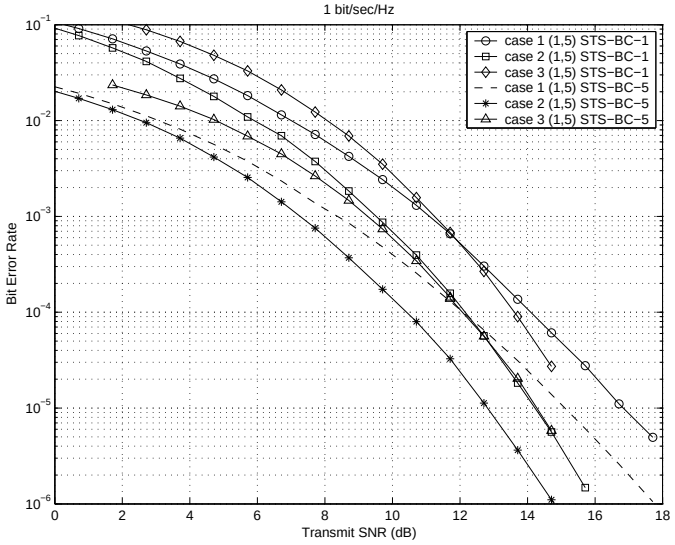


Fig. 1. BER vs. SNR for a reconfigurable MIMO system using PIXEL antennas at the receiver; $M_T = 2$ and $M_R = 1$; 1 bit/sec/Hz using BPSK.

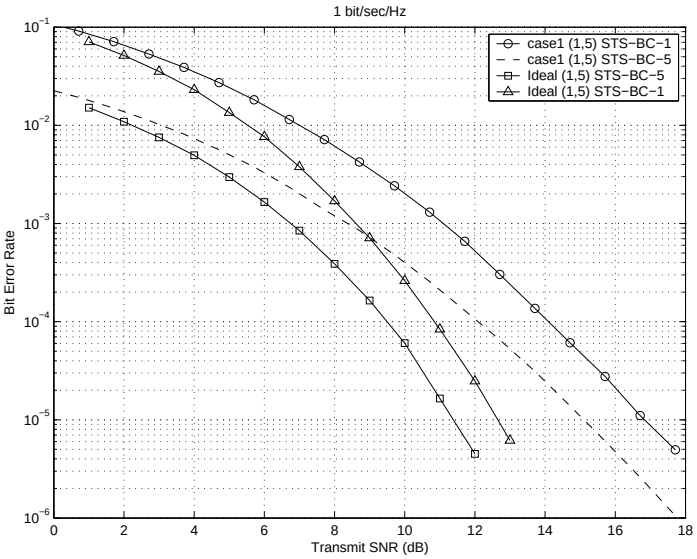


Fig. 2. BER vs. SNR for a reconfigurable MIMO system; $M_T = 2$ and $M_R = 1$; 1 bit/sec/Hz using BPSK.

increases, the diversity of the system increases as well. Note that case 1, which has the smallest angular spread, exhibits the lowest diversity among the 3 cases. Also, the amount of received power depends on both the angular spread and the shape of the radiation pattern, thus the curves corresponding to cases 1, 2 and 3 are shifted up or down with respect to each other. As expected, a similar pattern is observed using STS-BC-5, in terms of diversity and received power level. However, compared to STS-BC-1, larger coding gain is achieved. In Fig. 2 we compare the performance of case 1 to that of a scenario with ideal correlation characteristics. For both STS-BC-1 and STS-BC-5 at a BER of 10^{-5}, the performance degradation in case 1 due to correlation is about 4 dB.

REFERENCES

[1] F. Fazel, A. Grau, H. Jafarkhani and F. De Flaviis, "Space-Time-State Block Coded MIMO Communication Systems Using Reconfigurable Antennas", submitted to *IEEE Trans. on Wireless Commun.*

Characterization of Online Social Networks

Progress Report - Spring 2008

Minas Gjoka
Advisor: Athina Markopoulou
Center of Pervasive Communications and Computing
The Henry Samueli School of Engineering
University of California, Irvine
mgjoka@uci.edu

***Abstract*— In Spring Quarter 2008, I continued doing research in Online Social Networks (OSN). I extended our work on characterization of OSN applications which led to publication [1]. I also started a new research thread in Online Social Networks which involves crawling the social graph of a pure social network, Facebook, and analyzing its network properties.**

I. Characterization of OSN Applications

In the previous quarter, I described the project which involved the characterization of Facebook applications in terms of their popularity and adoption dynamics. In that work, we proposed a method to simulate the process with which users install applications so as to determine the user coverage from the popularity of applications, without detailed knowledge of how applications are distributed among users.

Since then, we realized that the work at [3] describes a non-linear preferential installation process of balls into bins, which is similar to our model. In the model of [3], a phase transition at $\rho = 1$ is observed with the system tending towards monopoly for values of the preferential exponent greater than one. Our model has a basic difference compared to that: we have many copies of a unique ball (many installations of a unique application) and a certain type of ball cannot be installed twice in a bin (a user cannot install an application twice). We are in the process of applying the continuum approach introduced at [4] to derive analytically the distribution of the number of application installs for users by calculating the time dependence of the degree d_i of a certain node i.

II. Characterization of the Facebook Social Graph

A. Motivation

The popularity of social networking has given rise to online social networks, such as Facebook and MySpace, with tens of millions of users. We believe that a detailed understanding of the network structure and connectivity of such networks will lead to better architecture design to support future social networking systems. In this project, we crawl Facebook and retrieve the whole social network of friends. To the best of our knowledge this is the largest ever study of OSNs. We use a distributed measurement platform to measure the whole graph as fast as possible.

B. Contributions

The contributions of this work include the following:

- **Novel measurement methodology.** We use Planetlabs as a distributed measurement platform with more than 800 nodes at our disposal to shorten the crawling time for such a huge network. PlanetLab is a worldwide-spread research network with the aim of fostering the development of new network services. The Planetlab machines crawl independently, with a coordination point (our machine at UCI) to avoid downloading duplicate lists of friend relationships. The latter feature is introduced to reduce the load at the Facebook servers. Another challenge that this distributed crawling framework faces is the privacy features that Facebook has developed to protect against such automated crawlers.
- **Measuring the Facebook Social Graph.** We plan to make available the partially crawled social graph after applying anonymization. The full graph is known to contain more than 70 million nodes and by approximation consists of more than 10 billion edges. Another fact that needs to be stressed here is that Facebook is a pure social network. We expect this dataset to be useful in many fields which include computer networks, sociology, complex networks, graph theory.
- **Characterizing the Facebook Social Graph.** We characterize the properties of the underlying graph

in regard to degree distribution, joint degree distribution, clustering coefficient, rich club connectivity, shortest path distribution and other interesting graph metrics that are used in the literature to compare graphs. This is one of the largest analysis on online social networks and it is performed on a large subset of the full network. We plan to capture a snapshot which contains more than 50% of the users and we only miss the users with the strictest privacy settings during the crawling.

C. Progress Report

We have developed the distributed crawling framework and have used it to crawl more than 4 million nodes. The measurement phase is in-progress and we expect to scale up to at least 50% of the total Facebook userbase. We are currently in the phase of developing a massive-scale graph framework on top of which we will build the algorithms needed to analyze/characterize the social graph. In the development of the graph framework we face the challenge of working with huge graphs which do not fit in memory (needs more than 64 GBytes of RAM). If we were to let the Operating System fetch in memory the necessary records and swap out when needed then the O/S will very fast reach a thrashing state with very slow progress observed afterwards. On the other hand, if we were to go to the other extreme and only access records from the hard disk we would observe again a large slowdown because of large access times in hard disks.

In addition to the characterization part of the social graph, we are looking into other uses of the crawled dataset such as modeling the diffusion process of information in such networks or modeling their evolution.

This is an ambitious project and we expect its proper completion to take 1-2 more quarters. Paper [2] is in preparation.

REFERENCES

[1] Minas Gjoka, Michael Sirivianos, Athina Markopoulou, Xiaowei Yang, "Poking Facebook: Characterization of OSN Applications", *to appear in Proc. of ACM SIGCOMM Workshop on Online Social Networks '08*

[2] Minas Gjoka, Pegah Sattari, Athina Markopoulou, "Measurements and Network Analysis of the Facebook Social Graph", *in preparation*

[3] E Drinea, A. Frieze, M. Mitzenmacher, "Balls and Bins Models with Feedback", *in Proc of SODA '02*.

[4] A. Barabasi, R. Albert, H. Jeong, "Mean-field theory for scale-free random networks", *Physica A: Statistical Mechanics and its Applications*, 1999

Progress Report on Multiple-Antenna Front-End Design Research: Spring 2008

Amin Jahanian Advisor: Payam Heydari
Nanoscale Communications Integrated Circuits Lab
Center for Pervasive Communications and Computing
Department of Electrical Engineering and Computer Science
The Henry Samueli School of Engineering
University of California, Irvine, CA 92697-2625
Email: jahanian@uci.edu

***Abstract* — In the spring quarter, we worked on incorporation of non-orthogonal codes on the previously designed multiple-antenna receiver architecture, and its effect on the overall bit-error-rate (BER) of the system. The use of non-orthogonal codes makes it possible for us to decrease the BW expansion inherent in the code-modulated path-sharing multiple antenna (CPMA) receiver architecture. The CPMA receiver is capable of accommodating multiple input multiple output (MIMO) including spatial multiplexing, spatial diversity, and beamforming. The use of a unique code-modulation scheme at the RF stages of the signal paths enables linear combination of all mutually orthogonal code-modulated received signals. The combined signal is then fed to a single RF/baseband/ADC chain, resulting in a significant reduction of power consumption and area, as well as mitigating the issue of LO routing/distribution. In the digital domain, all antenna signals are fully recovered.**

I. INTRODUCTION

Multi-antenna communications promises higher data rates using spatial multiplexing (SM), and increased range using spatial diversity (SD) and beamforming (BF). The use of multiple antennas in any MIMO RX may entail multiple RF chains, baseband blocks and ADCs [1], [2]. Consequently, there will be considerable increase in power consumption and chip area. In addition, having multiple receive chains results in a complicated LO routing and distribution task.

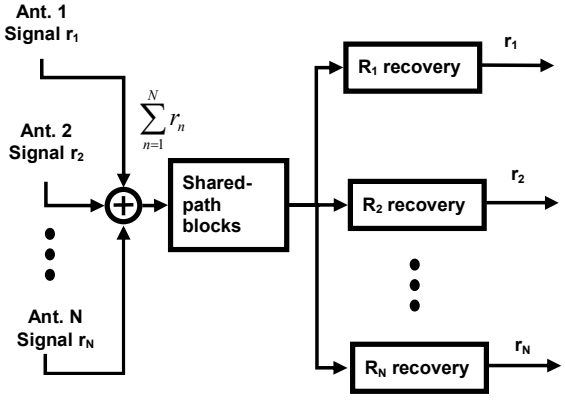


Fig. 1. Conceptual diagram of the proposed CPMA receiver

Previously in this project, we had presented a new **code-modulated path-sharing multi-antenna** (CPMA) RX front-end architecture (Fig. 2) that enables sharing of RF, baseband, and ADC blocks among multi-antenna signals. The underlying idea is to implement a code modulation system *within* the multi-antenna RX in order to distinguish antenna signals before combining them in the RF domain. More specifically, N antenna signals are modulated by N orthogonal code sequences. The mutual orthogonality of code-modulated signals allows signal combination in the RF domain, while promising full recovery of each signal in the baseband using digital matched filters (DMF). The recovered signal is then fed to the MIMO DSP for further processing. The proposed CPMA RX front-end is capable of accommodating any multi-antenna scheme, including SM, SD (including OSTBC, MRC, and BF). The advantages of this architecture include significant reduction of power consumption and chip area, and mitigation of coupling between antenna signals in the RX. Moreover, the single path alleviates the problem of LO distribution and routing in multi-antenna architectures.

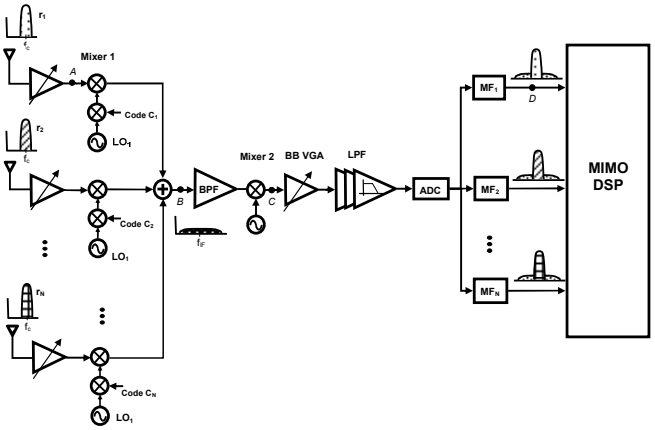


Fig. 2. Proposed CPMA receiver

II. EFFECT OF NON-ORTHOGONAL CODES ON BER

In spatial multiplexing schemes, the MIMO receiver suffers from inherent multi-stream interference (MSI). Consequently, MIMO detectors are needed to separate the multi-stream into its original transmitted streams. In the proposed CPMA receiver, the use of non-orthogonal codes adds non-zero code cross-correlation in addition to MSI. Thus, the MIMO detectors should be designed to detect the transmitted symbol in the presence of both of these phenomena. Even so, we can use the effective channel and noise matrices to build the expressions for a maximum

likelihood (ML) and minimum mean squared error (MMSE) detectors.

Simulation results of the BER performance of the CPMA receiver are provided under uniform-valued code cross-correlation ρ for spatial multiplexing and spatial diversity. In both experiments, we assume no CSI at the transmitter and perfect CSI at the receiver. The data is QPSK modulated and Monte Carlo simulations are executed.

In the first experiment, we implement uncoded spatial multiplexing and realize an MMSE detector for MSI separation under various ρ values. Fig. 3 shows the BER performance for a 4×4 system for ρ varying in 0.1 step sizes from 0 to 1. From Fig. 3, we see that for ρ<0.7, the performance degradation is quite negligible. This demonstrates the feasibility of using non-orthogonal codes for $N>G$ if ρ<0.7. Note that when ρ=1 the BER is 0.5 because each antenna signal is modulated with the same code and then added. This would not happen in practice because a different code is applied to each antenna signal.

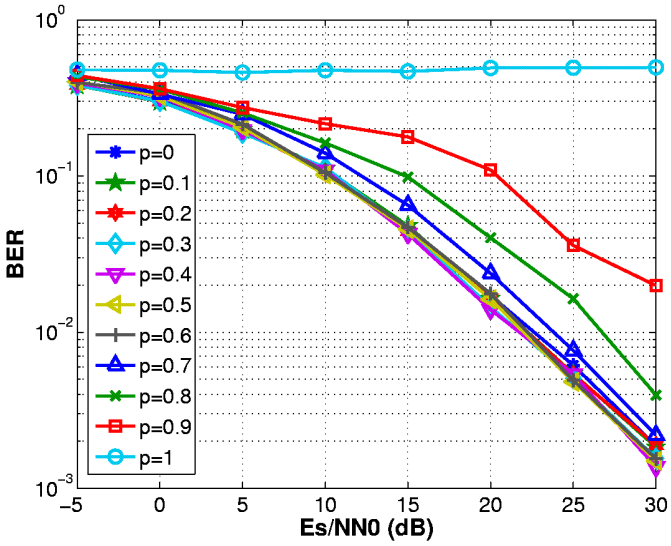


Fig. 3. BER vs. E_s/NN_0 vs. ρ for 4×4 MIMO spatial multiplexing

In the spatial diversity experiment, we use the ½ rate OSTBC for a 4×4 MIMO system. Since OSTBC simplifies the MIMO channel into equivalent SISO channels, it is reasonable to use ML detection for each symbol without imposing high complexity. Fig. 4 shows the BER performance for ρ varying in 0.1 step sizes from 0 to 1. Depending on the tolerable performance loss, an appropriate ρ can be determined.

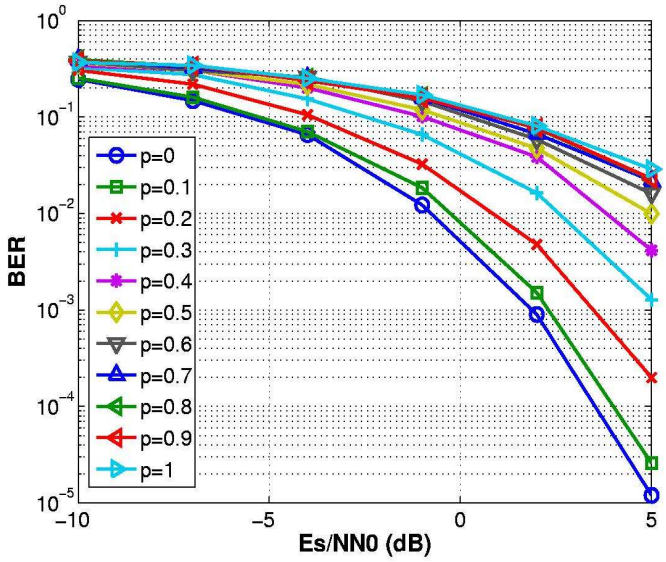


Fig. 4. BER vs. E_s/NN_0 vs. ρ for 4×4 spatial diversity

REFERENCES

[1] A. Behzad, *et al*, "A fully integrated MIMO multi-band direct-conversion CMOS transceiver for WLAN applications (802.11n)," *ISSCC Proc.*, pp. 560-561, Feb., 2007.

[2] Y. Palaskas, *et al*, "A 5-GHz 108-Mb/s 2x2 MIMO transceiver RFIC with fully integrated 20.5-dBm P_{1dB} power amplifiers in 90-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 41, no. 12, pp. 2746-2756, Dec. 2006.

UC Irvine Center For Pervasive Communications and Computing

Graduate Fellowship Progress Report Spring 2008

July 19, 2008

Project Name: Adaptive Equalization of Multimode Fiber Channels in $0.13\mu m$ CMOS
CPCC Affiliate Professor: Michael Green
Mailing Address: 544 Engineering Tower, Irvine CA 92697-2625
Phone No.: 949-824-1656 **E-mail:** mgreen@uci.edu
Student Fellowship Recipient: Mahyar Kargar

Introduction

With increasing data rates and link distance in fiber-optic systems, the transmission path becomes severely limited by fiber non-idealities, especially dispersion. Intersymbol interference (ISI) is a fundamental limiting factor in band-limited communication links. In particular, in multimode optical fiber, which are the dominant fiber type in local area network (LAN) links like 10Gb/s Ethernet,the ISI is mainly due to modal dispersion. Electronic dispersion compensation (EDC) is used to combat this ISI in short-distance fiber links. Use of adaptive equalizers as an EDC method is known to make the data communications over short ranges of the MMF possible. In this project a high-speed adaptive DFE is designed to combat the ISI caused by the band-limited MMF channel.

Summary of Accomplishments

Due to the change of the design kit all the previous work had to be redone in the new IBM cmrf8 $0.13\mu m$ CMOS kit. The design of feed-forward filter (including the DLL and gain control loop) and the adaptive feedback filter (including high-speed slicer and DFF's) has been completed in new $0.13\mu m$ CMOS. LMS algorithm is implemented to adaptively change the feedback filter coefficients using an analog integrator and a Gilbert multiplier.

The project is in the layout and integration mode now. The building blocks of the feed-forward and feedback filters have been laid out and the results have been verified with post-layout RC-extracted simulations.

The LMS circuit block with Gilbert multiplier and integrator has been laid out and verified.

A 10Gb/s binary Alexander phase detector for the CDR loop designed using high-speed DFFs and symmetric XOR circuits.

Ongoing Work

1. Verification of the feed-forward filter with post-layout simulations on extracted blocks after layout.
2. Verification of the feedback filter with post-layout simulations extracted blocks after layout.
3. Verification LMS circuit convergence properties using transistor level post-layout simulations.
4. Design, lay out and verification of a 10GHz LC VCO to be used in the CDR loop.
5. Top level simulation of the 10Gb/s CDR with extracted VCO verifying the locking behavior and performance.

Progress Report on Embedded SRAM Reliability considering the Temperature Effect: Spring 2008

Amin Khajeh Djahromi, Advisor: Professor Ahmed Eltawil
Center for Pervasive Communications and Computing
Department of Electrical Engineering and Computer Science
The Henry Samueli School of Engineering
University of California, Irvine
Irvine, California 92697-2625
Email: akhajehd@uci.edu

Abstract—**In older technologies (above 100nm), it was common practice to increase the supply voltage of memories as compared to logic to mask the effects of process variation as well as improve performance. For nano-scale CMOS circuits, this assumption is no longer valid. Overdriving the memories by increasing the supply voltage is functional only up a point after which an inflexion in the probability of errors occurs and the assumption breaks down. In fact, due to temperature increase, the probability of error increases at higher supply voltage. In this report we have developed a mathematical model to quantify the effect of the temperature on the embedded SRAM reliability.**

I. Introduction

The traditional statistical power consumption optimization approaches does not factor in second order effects such as temperature. According to ITRS 2007 "A key form of variability is due to thermal effects during operation; this variation is on the time scale of billions of clock cycles and can affect timing and noise phenomena" [2]. Elevated chip operating temperatures impose constraints on the circuit performance in terms of frequency of operation and power consumption, (specifically leakage) [3][4]. It has been shown that sub-threshold leakage increases exponentially with temperature while gate leakage is rather insensitive to temperature. BTBT leakage, on the other hand shows a weak linear dependence on temperature [10]. Furthermore, as leakage current increases, the die temperature increases, causing a further increase in leakage current. This indirectly sets a limit on the scaling of threshold voltage to avoid an unstable heat dissipation situation [5]. Modern techniques for temperature management include using temperature sensors to monitor the temperature gradients on a die, then applying throttling techniques that control the frequency of operation such that a target temperature is maintained across the die [6]. Other techniques include resource spilling to distribute heat density or cooling techniques both on die and on package.

II. Analytical Approach

In the following we try to find the minimum supply voltage required to achieve a specific probability of error given an expected temperature profile for the system. The variation in the threshold voltage will results in a Gaussian distribution for the access time/write time/storing node voltage which can be modeled as a Gaussian distribution with moments that are a function of supply voltage and temperature. In this case the access time, T_{Access} can be expressed as:

$$T_{Access} \sim \mathcal{N}(\mu_{Access}(V_{dd},T), \sigma_{Access}(V_{dd},T))$$

Where T is the temperature and V_{dd} is the supply voltage. For a given frequency, F, one can find the maximum allowed time for the read operation, T_{MAX}. Therefore, we can define the probability of error as:

$$P_e(V_{dd},T,T_{MAX}) = P[T_{Access} > T_{MAX}] = Q(\tau)|_{\tau=t}$$

where

$$t = \frac{T_{MAX} - \mu_{Access}(V_{dd},T)}{\sigma_{Access}(V_{dd},T)}$$

and $Q(\cdot)$ is the Gaussian Error Integral or Q −function and is given by:

$$Q(\tau) = \frac{1}{\sqrt{2\pi}} \int_{\tau}^{+\infty} e^{\left(-\frac{x^2}{2}\right)} dx.$$

For a given probability of error of P_{e0} and, t_0 can be calculated such that $Q(t_0) = P_{e0}$. Thus for P_{e0} and F, we will have:

$$t_0 = \frac{T_{MAX} - \mu_{Access}(V_{dd},T)}{\sigma_{Access}(V_{dd},T)}$$

where t_0 and T_{MAX} are constant. If we solve this equation for V_{dd}, we will have:

$$V_{dd} = G(T).$$

On the other hand, the choice of floor plan, frequency of operation and activity of the neighboring block result in the dependency of the temperature to V_{dd} which can be expressed as:

$$T = H(V_{dd})$$

therefore for a given probability of error and frequency, we can find the appropriate V_{dd}, which results the probability of error of P_{e0} for a given frequency while factoring in the effect of the temperature by solving:

$$V_{dd} - G\big(H(V_{dd})\big) = 0.$$

It is important to note that solving the above equation may results in more than one real solution for V_{dd} (as shown in figure 1) which in that case the optimum solution will be the minimum V_{dd}.

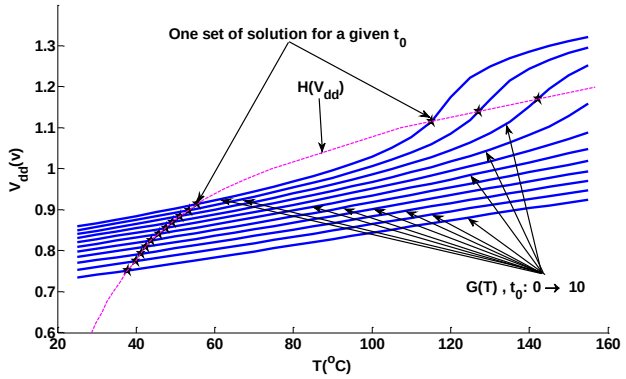


Figure1: analytical solution to the optimization problem

REFERENCES

[1] A. Gupta et al., "A System Level Leakage-Aware Floorplanner for SoCs,"Proc. of ASP-DAC, 2007.

[2] International Technology Roadmap for Semiconductors (ITRS), http://www.itrs.net/Links/2007ITRS/Home2007.htm

[3] Stephan Ohr, "Efforts heat up to remove processor hot spots," EE Times, February 2005.

[4] K. Banerjee, A. Sangiovanni-Vincentelli et al., "On Thermal Effects in Deep Sub-Micron VLSI Interconnects," Proceedings of Design Automation Conference, 1999

[5] Aseem Gupta, Nikil Dutt, Fadi Kurdahi, Kamal Khouri, Magdy Abadir,``STEFAL: A system level temperature- and floorplan- aware leakage power estimator for SoCs," International Conference of VLSI Design, 2007

[6] D. Brooks, M. Martonosi, "Dynamic Thermal Management for High-Performance Microprocessors," Proceedings of International Symposium on High Performance Computer Architecture, January 2001

[7] J. Donald et al.,"Techniques for Multicore Thermal Management: Classification and New Exploration,"Proceedings of International Symposium on Computer Architecture, 2006

[8] K. Skadron et al.," Control-Theoretic Techniques and Thermal-RC Modeling for Accurate and Localized Dynamic Thermal Management," Proceedings of International Symposium on High-Performance Computer Architecture, 2002.

[9] Predictive Technology Model (PTM): http://www.eas.asu.edu/~ptm/

[10] Roy, K.; Mukhopadhyay, S.; Mahmoodi-Meimand, H., "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proceedings of the IEEE* , vol.91, no.2, pp. 305-327, Feb 2003

A Generic Network Interface Architecture for Network-on-Chip: SPRING 2008

Seung Eun Lee, Advisor: Nader Bagherzadeh
Center for Pervasive Communications and Computing
University of California-Irvine
Email: seunglee@uci.edu

***Abstract*—In Spring 2008, we developed a generic architecture for network interface (NI) and associated wrappers for a networked processor array (NoC based multiprocessor SoC) in order to allow systematic design flow for accelerating the design cycle.**

I. Introduction

As a new SoC design paradigm, the Network-on-Chip (NoC) [1] has been proposed to support the integration of multiple IP cores on a single chip. In NoC, the reuse of IP cores in plug-and-play manner can be achieved by using a generic NI, reducing the design time of new systems. NI translates packet-based communication into a higher level protocol that is required by the IP cores by packetizing and depacketizing the requests and responses of the cores. Decoupling of computation from communication is a key ingredient in NoC design. This requires well defined NI that integrates IP cores to on-chip interconnection network to hide the implementation details of an interconnection.

In Spring 2008, we developed a generic architecture for NI and associated wrappers for a networked processor array (NoC based multiprocessor SoC) in order to integrate IP cores into on-chip interconnection networks efficiently. We split the design of a generic NI into master core interface and slave core interface. First, we implement an NI architecture for an embedded RISC core. Then, an application specific wrapper for a slave IP core is proposed based on the NI. In order to implement a wrapper, we start by choosing application-specific parameters and writing an allocation table for architecture description. The allocation table is used for the configuration of the modular wrapper and for the software adaptation.

II. Generic Network Interface

NI consists of a packetization unit (PU), a depacketization unit (DU) and PE interface. NI is located between a router and a PE, decoupling the communication and computation. It offers a memory-mapped view on all control registers in NI. With this interface model, a simple implementation can be accomplished. All of the register accesses are done by bus interface and BLOCK data transfer can be handled by the DMA controller. DMA controller manages BLOCK data transfer from/to the internal memory. In order to achieve high performance, all operations are completed in one cycle.

Fig. 1 shows the micro-architecture of a modular wrapper for a slave IP core interface. The input control signals

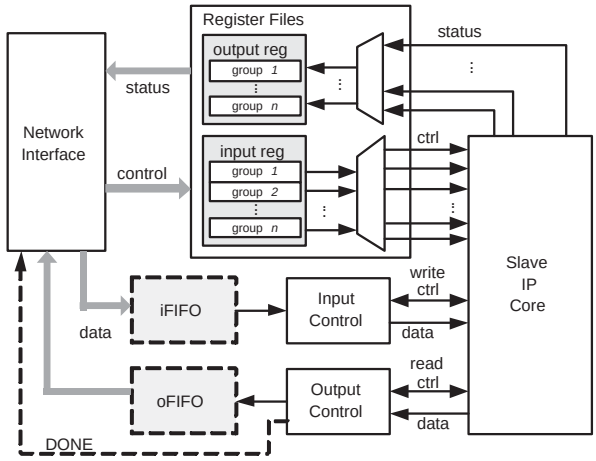


Fig. 1. Micro-architecture of a modular wrapper for a slave IP core

are grouped by their functionality and then assigned to the application specific registers in the wrapper. These registers are accessed by NI using SINGLE packet to initialize the control signals which are allocated to dedicated signals and fed to the slave IP core completing initialization. Status signals have specific functions. For instance, the *error* signal requires special services such as generating trap to another PE or stop the operation of the slave IP core. The *done* signal initiates communication to another PE to transmit the results of the slave IP. These status signals need dedicated logic for each signal. There are a set of status signals and associated control logic to generate the controller for status signals. Input data for a slave IP core is sent by other cores through network and NI translates the incoming packet for the slave IP core. In order to handle this data width mismatch, we adopted two operation modes for the data interface such as Unbuffered and Buffered modes. The generic NI with modular wrapper allows to accelerate the design cycle and a proposal of a systematic design flow for an application specific interface.

References

[1] S. E. Lee, J. H. Bahn, and N. Bagherzadeh, "Design of a feasible on-chip interconnection network for a chip multiprocessor (cmp)," in *SBAC-PAD'07: Proceedings of the 19th International Symposium on Computer Architecture and High Performance Computing*, 2007.

Architectural and Algorithm level Fault Tolerant Techniques for Low Power High Yield Multimedia Devices

Mohammad A Makhzan (Avesta Sasan), Student Member IEEE, Ahmed Eltawil, Member, IEEE, and Fadi J. Kurdahi, Fellow, IEEE

***Abstract* — This paper presents a novel architecture that allows a high level of fault tolerance in embedded memory devices for multimedia applications. The benefits are two fold by allowing systems to operate at a lower voltage thus saving power, while improving yield via error masking. The proposed architecture performs a remapping of defective parts of the memory while allowing single cycle access to the remapped portions. Furthermore, it provides run time control of the enforced protection policies leading to an expanded design space that trades off power, error tolerance and quality. Simulations indicate a reduction of up to 35% in encoder power for a 65 nm CMOS process.**

I. INTRODUCTION

JPEG2000 is an example of many multimedia applications that require significant processing power yet must be implemented in an increasingly power-thrifty and cost-conscious context in order to exist in handheld devices. The push to reduce power and cost have necessitated increasing levels of integration, leading to Systems-on-Chip (SoC) that integrate interface, processing and storage on a single silicon chip. These SoCs, tend to have significant amounts of memory needed to store input, output and intermediate forms of media (e.g. images, audio or video) during the encoding and/or decoding process. The increasingly demanding requirements on image and display resolution has resulted in the need for large memories to store such images, to the point where memory is already the dominant part in most advanced SoCs today-a trend that is expected to increase for the foreseeable future.

The increasing amount of on-chip memory, while beneficial in reducing cost, has a two-fold shortcoming: 1) Power consumption metrics become dominated by the metrics of the embedded memory, 2) Yield of the overall integrated circuit is negatively impacted since the defect density of memories is roughly twice as much as logic 18. The problem becomes compounded in advanced CMOS technologies where Random Dopant Fluctuation (RDF) 1,2,3 is starting to become the dominant factor causing intra-die variations. Intra-die variations shift the process parameters of different transistors in a die in different directions, which can result in significant mismatch between neighboring transistors leading to memory cell instability and failure 1234. These failures are manifested as either an increase in the cell access time or unstable read and write operations. In this paper, we will focus on parametric failures, which are failures caused by process variations that can be lumped into an effective change in the individual transistor threshold voltage. Since for a single memory cell the transistors are in close proximity, RDF will be the primary cause of mismatches between the transistors. To counter the effect of RDF induced faults designers typically raise the supply voltage of the circuit to assure the masking of these errors. This approach in turn creates an undesired relation between yield and power consumption, where power consumption (voltage supply) must be high to ensure a high yield. This statement counters the well known technique of supply scaling to reduce power consumption, where in CMOS, the exponential relation of the leakage power with the supply voltage 5 and the square relation of the dynamic power with Supply voltage5, are typically used as one of the most effective means of reducing power consumption.

An in-depth view of these two issues namely power and yield, indicates that they are really two facets of the same problem. If memory dominated systems can be designed to be fault tolerant, they will also be low-power by definition. While it is true that some applications cannot function unless the data processed is 100% correct (such as processor instruction code), there exists a broad family of applications that are inherently fault tolerant such as wireless applications and multimedia systems. The authors have shown in previous publications 1011 that maintaining 100% correctness in such embedded systems is not the optimum power saving approach. Resilience to hardware-induced faults makes power reduction techniques such as aggressive Vdd scaling feasible, while increasing the effective yield of the chip via masking errors. As features sizes shrink, the rate of operation condition induced errors becomes significantly higher 19 and as a result the cost of correcting them increases significantly.

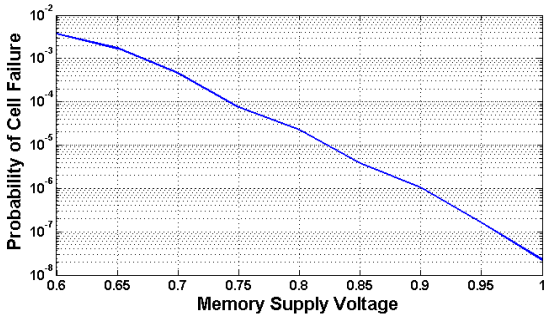


Figure 1: Probability of bit error VS V_{dd}

In this paper we intend to focus on embedded memories and identify means of detection and correction of operating condition faults (Vdd scaling). Figure 1 illustrates the relation between supply voltage and probability of bit error P(e) for an SRAM memory cell in 65 nm CMOS technology19. The probability of failure is the cumulative sum of access, read and write failures due to process variation. Obviously as the voltage is reduced, errors due to process variations start to increase in an exponential manner. These errors must be detected and corrected to minimize the impact on measurable quality metrics such as peak signal to noise ratio (PSNR).

The paper proposes a means of identifying and remapping faulty parts of the memory in such a way which affords selective protection to different parts of the memory. For example, most significant bits can be protected while least significant bits can be allowed to have errors. We show that such an approach leads to the possibility of reducing voltage to aggressive levels while still maintaining target metrics such as PSNR levels.

II. Prior Work

Fault tolerance in embedded multimedia system could be achieved in hardware or software. Depending on the nature of defects or faults one or the other will be better suited. However, since they operate in different domains inclusion of one doesn't mean exclusion of the other. In this work, we will consider approaches that deal with process variations at both software and hardware levels.

At the circuit and architecture level many techniques have been previously proposed to deal with SRAM faults, many of which could be used to deal with process variation. A standard technique, used by many manufacturers, is utilization of redundant rows and columns [24][25]. In this technique by remapping the rows and columns that contain defective cells to the redundant rows, the yield of production is improved. However, typically, the number of redundant rows available are limited to minimize area overhead. Moving to smaller geometries, the number of weak cells (due to process variation) grows exponentially [26][27], which renders static systems such as redundancy unusable. Another commonly used technique is the use of Error Correcting Codes (ECC) [28][29] to deal with the transient defects. ECC memories can handle dynamic faults albeit at a heavy cost in power consumption, area and complexity[30]

Works presented in [31], [32] and [33] accordingly introduce, padded caches, resizable caches and cached caches. These architecture level techniques are designed to leverage tolerability against process variation; however they rely on the existence of the lower level memories and the ability to re-fetch the data if lost. In addition these protocols are designed to guarantee 100% correctness for the residing data where as in multimedia applications, 100% correctness could be relaxed for better power savings with unnoticeable loss in picture quality.

At the software level many techniques have been developed to deal with defective pixels which could be easily adapted to handle process variation. Process variation due to its random and uniform distribution will result in salt and pepper noise in the images if they are saved in the memory using any simple and non-compressed data format. Many techniques [34][35][36] have previously addressed salt and pepper noise. In some multimedia application however the data is saved in compressed or transformed format in the memory. One such application is JPEG2000 where pictures after transformation to wavelets are saved in the memory. Any memory defect would then manifest itself as a faulty wavelet coefficient.

The work in [37] introduces a software level approach to deal with coefficient defects. This technique is a blind technique applied at the decoder site where by repetitive iteration, the decoder is capable of identifying faulty pixels and fixing a majority of these faulty locations. This technique increases the power consumption of the decoder but allows the use of very low power low cost encoders, since the 100% correctness criteria can be relaxed at the encoder site (which is typically a mobile device with a stringent power budget).

Our proposed approach allows the designer to tradeoff several key aspects of the design such as quality versus power consumption by offering the ability to selectively protect different parts of the memory. This approach can be used in conjunction with other software approaches as previously discussed. The remapping circuit presented in the proposed architecture is novel and is used to increase redundancy coverage even at very high rates of defects.

III. Architecture / Circuit level Masking technique

This technique relies on the inherent fault tolerant property in multimedia application. Using 8 bits for each color value 256 different color levels in each channel is realized. The naked human eye is able to recognize about 25-35 different gray levels. This implies that small changes in the color values of the picture most likely are unrecognizable for the human eye. This creates a design space where least significant bits (LSB) can afford to have more faults (or less protection) than most significant bits (MSB) without a significant drop in PSNR. Furthermore, in a lossy transformation these changes in the LSBs are further masked by the lossy nature of the transformation rendering the PSNR drop negligible. Having this in mind, in the following section, we introduce an architecture that guaranties the correctness in MSBs while defects in LSBs are allowed. As we will see the system is adaptive in a sense that 1) it tolerates different defect rates in different sections (banks or sub-banks) in the memory, 2) the policy of fault coverage could be changed dynamically based on decisions by the system user or designer on how much power savings versus PSNR drop are acceptable in the system.

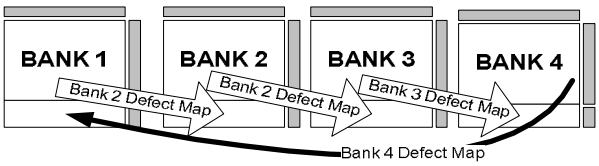


Figure 2: Suggested Defect Map organization in the new architecture

A. SRAM with Variable Remapping Size (VRS_SRAM)

The memory rows are divided into smaller groups called words. Each word contains 2^n bytes with n varying from 1 to 4. A word is considered defective if one or more error is detected in one of the MSBs of each byte in the word. The decision of how many bits are considered MSBs is a run or design time decision. The proposed technique relies on redirecting one single word rather than an entire row when a defective word is countered. This technique is applied to banked SRAMs such as the architecture depicted in figure 2, where banks are organized in a circular chain with each bank providing redundancy for the previous bank in the chain. Reading a block of data which contains defective words includes reading the redundancy section of the next bank, accessing the requested block from the current bank and finally replacing the defective words with the words read from the redundancy section of the next bank. The motivation behind this is three fold.

1- The redundancy section of the next bank is used because the accessed bank and the redundancy section of its next bank could be accessed at the same time to avoid multi cycle access for data to one bank. In this structure, a separate decoder is needed for the redundancy section of each bank removing the complexity of traditional row decoders and thus providing remapping capabilities and effectively reducing the area of row decoders.
2- To reduce power consumption via power gating of the unused sections in each bank.
3- Providing the ability to determine the protection policy (number of protected MSBs) based on desired operating voltage.

Since fault tolerance is performed at the word level, knowledge of the location of defective bits should be provided apriori and is stored in a defect map.

B. SRAM Defect Map (SDM)

A dynamic defect map is required due to the dynamic nature of the faults. As described in section I, a majority of the parametric faults are dependent on the operation condition (voltage, temperature etc.), furthermore, as shown in figure one, the errors are exponential in nature with lower supply voltage. These realities render static remapping (one time remapping by burning a fuse and redirecting data to a redundancy section) rather useless. An added disadvantage is that designing a decoder to provide remapping for a large number of redundant rows (for example 16 rows or larger for each 128 rows) will result in complex logic with a large area and delay overhead.

To construct this defect map we assign one bit per word to indicate the status of that word (healthy or defective). Due to the small size of the defect map it is safe to assume that it can be supplied from a stable higher supply and protected by regular built in self repair techniques (BISR). Note that at higher voltages, parametric errors are greatly reduced.

Depending on how it is updated, the defect map may be used for storing temperature induced defects, process variation defects and manufacturing defects. Since we need to operate at different voltages and protection policies, each new configuration will require a refresh of the SDM contents. This could be accomplished by creating a finite set of configurations and preloading the defect map at manufacture time or alternatively, the defect map can be updated with each new configuration via a built in self test (BIST) run. Populating SDM at manufacture time is an expensive solution since testing for multiple error vectors will significantly add to the final chip cost. This analysis favors the boot/run time defect detection for populating of SDM by using the already existing memory BIST infrastructure.

C. Access properties of the SDM

In Multimedia application, access to memory is sequential in nature (there is no branching and access to the image data is serial). This promotes the use of a serial pipeline access of the defect map to always pre-fetch the defect information before accessing the SRAM. In order to further reduce power consumption of SDM access to the SDM is serialized by a small "cache" register file referred to as Defect Map Buffer (DMB). DMB is an array of flip flops equal to the size of each row in the SDM and a tag section. Each time the SDM is accessed, the DMB tag is updated with the TAG of the accessed address shifted to the right by $\log_2[Size(DMB)]$ and buffer data section stores the defect map in the accessed row that could be fetched in one access to SDM. Since the image data is accessed serially, then after one access to the SDM, the next few accesses for defective data could be satisfied solely by the DMB without a need for accessing the SDM. This approach reduces the number of accesses to SDM and in turn reduces power consumption since the SDM is comparatively a much larger and more power consuming structure than the DMB register.

D. Power gating opportunities

When accessing a block of data in the SRAM that contains one or more defective words, some of the words (healthy) are read from the accessed bank, and other (remapped words) are read from the redundancy section of the next bank. In such cases we do not need to read out all the words in the SRAM. To save power/energy we added a gating feature to the column decoder and the sense amplifiers of each bank. Gating is performed prior to column activation (pre-gating). Using the SDM information, gating avoids reading the defective words in the main bank and instead the words are read from the next bank. Gating is also added to the redundancy section of each bank since we rarely access all the words simultaneously and instead need only one or a few words in

each row. To apply this gating technique to the redundancy sections we introduced some compaction and shifting logic. The compaction logic receives the SDM information (defect information) of the accessed row in the accessed bank and compacts all the 1's (faults) in the array to the right. For example, if the defective array contains {0,1,0,0,0,1,0,1} for a row with 8 words, the compaction logic will produce {1,1,1,0,0,0,0,0}. This indicates that only three locations should be accessed in the redundancy section. The remaining columns could be gated leaded to lower power consumption. Furthermore, the compaction allows tight packing of the faulty words in the redundancy sections.

E. Addressing the memory

In order to simplify the remapping logic we assumed that all the words with their defect map on one row of the DMB could be only redirected to one row in the redundancy sections. Since each row in the redundancy section has a limited number of words, an upper bound on the number of defects that could exist in each m (m being the size of the SDM row) consecutive words in the memory with their defect map placed on the same SDM row is introduced. To facilitate book keeping, an extension is added to each SDM row to identify the number of errors that have occurred in the memory up to the current row. This extension will be referred to as Error Counting Extension (ECE). Each SDM row starts with an ECE section of 'E' bits with $E = W \times R$, W being the number of words per block, and R being number of redundant rows per bank. The value stored in the ECE section is used for row selection in redundancy section of the next bank in which the remapped words exists. The ECE calculation for each row is performed based on the algorithm given in the flowchart of Figure 3. The flowchart march trough the memory locations and update the ECE of each row with the number of defects in the memory prior to that row. Essentially ECE is used as an offset for remapping to the redundancy section of next bank. However, in reality, the calculation of the ECE offsets is further complicated by the following issues:

Since ECE count is used for remapping to the redundancy section of the next bank, the defective words in the redundancy blocks of the next bank should also be considered in generating the ECE value. This case is illustrated in Figure 4.a.

Due to compaction, the data that is supposed to be available in one access may not fit in one redundant row (in case of multi word faults). To allow one cycle access, all defective words related to one fetch group should be stored in the same redundant row. This restriction requires special provisioning to assure correct addressing. This case is illustrated in Figure 4.b.

To ease the address calculation overhead we assume that there are no more that m defective words in each DMB row (where m is the number of words in each row). While the last rule results in sub-optimal compaction, it makes the address calculation of the redundant bank trivial therefore improving both timing and power consumption.

Finally, it is important to note that the ECE provides the error count up to its associated row in memory. To identify the exact offset we introduce two more variables. 1- The number of defects within a specific fetch group which we will identify as the Fetch Error Count (FEC) register, and 2- Number of words used for previous defective words in the SDM row associated to the fetch group prior to current fetch group. This register is called Fetch Error Counting Extension (FECE). FECE is obtained by addition of the ECE of the current row to the number of defects in the current SDM row prior to the fetch group and then taking the least R bits with R being the number of words in each redundancy row.

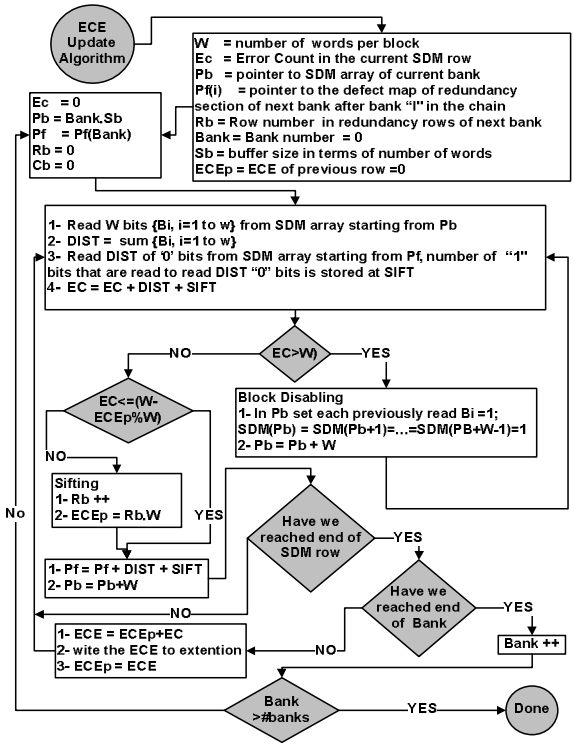


Figure 3: ECE update algorithm

After compaction the defect map is right shifted by value indicated by register FECE. Register FECE holds an offset indicating the number of words in the current redundant row that are used for the previous defects in the accessed bank. For example, if the FECE stores a (0011) it means that the first 3 word in the current redundancy word are already in use (or defective). If the output of the compaction is (11100000) after shifting we will have (00011100) at the output of the shifter. At this point any column that receives a 0 input from the shifter is gated and others are read. If the delay of the compaction and shifting logic exceeds the delay of the previous bank row decoder the generated data could be used for post-gating to only sense and output the bit-lines with valid remapped data. Since gating consumes some dynamic energy, if the number of defects is not very high the overhead of the gating could be more than the final savings therefore gating could be optional and applied only in lower voltages where the defect rate is exponentially higher.

Calculation of the ECE and FECE could be done either dynamically every time a buffer miss occur or once when ECE

values are calculated and saved in SDM. The first case is achieved by adding a tree of small 2bit, 3 bit and 4 bit adders. Although the number of those small adders is large, the switching activity (due to small number of defective bits) is very low and the adder tree doesn't consume much power. In the second case the FECE and FEC calculation is performed along with ECE calculation (with minor change to the ECE counting algorithm in Figure 3) and the calculated values are saved in the SDM. Every time the SDM is accessed for a new defect map, the auxiliary data (FEC and FECE for each Fetch group) is also obtained and saved in an Auxiliary-Buffer (AXB) along with DMB. AXB and DMB are both selected using the same selection logic.

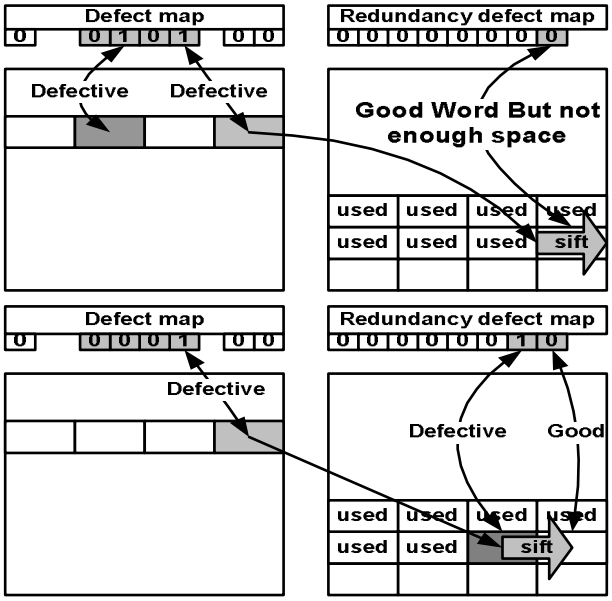


Figure 4: (top): Sifting mechanism a-sifting due to lack of space in the current row. (Bottom): Sifting due to defective location in current redundant row

F. *Demonstration of an Access Scenario*

In Figure 5 an access scenario where defect information resides in DMB and AXB is demonstrated. In this illustration we assumed 16 bit address for accessing the SRAM. On each row of the SRAM there are 8 words with each word being 32 bits. When the access is initiated the tag of the accessed address is checked against the tag of the DMB using the 7MSBs, if tag mismatches the SDM is accessed for defect map other wise, the defect map is obtained from DMB. Then the SRAM row/column decoder is activated. In parallel, the DMB, FEC and FECE are queried using the 8th through 11th MSBs of the address. If the defect map flags the Fetch Group (Block) as defective, the first two bits of the ECE are used to activate the redundancy row of the next bank (assuming a redundancy size of 4 rows per bank). In addition the FEC and FECE are used to activate the columns that contain the remapped words. Finally the sensed word from the redundant section of the next bank is inserted in place of the gated word from the main bank to produce the fetch block that is sent to the Fetch Unit (FU). Careful timing analysis using 45nm and 65nm library information indicates that the combining logic has more than enough time (decoding + MEM unit access) to prepare the routing scheme based on the defect map information.

G. *Fault policies*

Using a dynamic configuration as described, one could define different policies for dealing with defective locations. Forcing a protection policy reduces to modifying the BIST engine algorithm to make different decisions. For example, if a 2 MSB protection policy is to be enforced the BIST engine only need to check the correctness of the first two MSB and based on that update the SDM. Another policy could be area protection rather than bit protection (or bit protection in a certain area). For example, a defect in the low frequency sections of the wavelet in the JPEG encoding process is far more damaging than a defect in high frequency area. A possible design might allocate different number of MSB protection for each location. The important fact is that, based on how these priorities are defined, different defect maps are generated and designs with different fault tolerance can be generated. Compared to traditional static redundancy, this proposed technique allows:

1- **Larger Redundancy budget:** new design tolerate far more defective words because with each defective word only that word is remapped and not the entire row.
2- **Flexibility:** it allows enforcing different protection and remapping policies.
3- **Reconfiguration:** it allows discovery and coverage of new errors as they occur with each new configuration.
4- **Voltage Scalability:** it allows aggressive voltage scaling by trading off PSNR versus power.

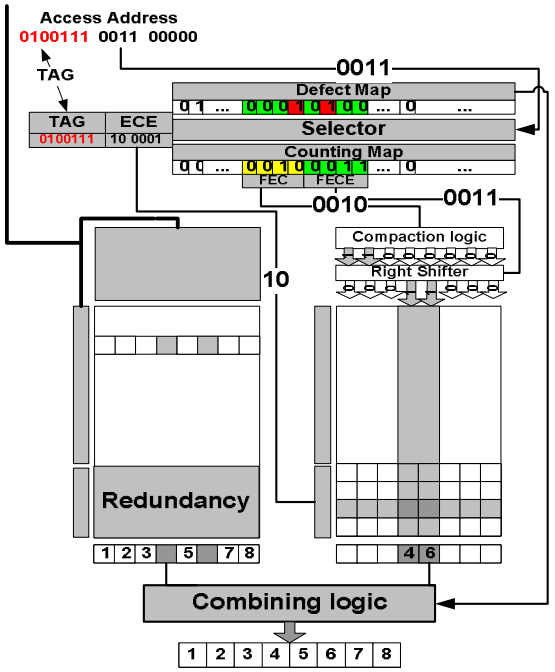


Figure 5: illustration of remapping mechanism

IV. Simulation Platform and Results

A. Simulation setup

To analyze how using a defective memory affects the quality of the JPEG image, we identified when and how the image data is saved in memory. Using JasPer[23] implementation of JPEG2000, we corrupted the memory image data at those points. Based on currently developed SoC architectures of JPEG2000 6789the image data is once saved at the initiation of the encoding process and then after the Discrete Wavelet Transformation "DWT" step (and before quantization). At the initiation of the encoding process, data is in a simple color format, such as RGB or YCrCb. Storing the image data in a defective memory as presented in Figure 6.c will result in changes in the RGB pixel values. This will result in salt and pepper like noise. Salt and pepper noise is well understood in the image processing fields and a number of filters, (e.g. median) have been devised as solutions to nearly eliminate such noise 13. Applying area or MSB protection reduces and nearly eliminates the salt and pepper noise, since MSB errors are not allowed to occur. After the DWT step, as shown in Figure 6.c data is saved in memory in the form of a wavelet, with each coefficient in the wavelet being 16 bits long. Each defective bit corrupts the coefficient that contains that bit. Corruption of one coefficient will result in corruption of all color values in the sampling window of that coefficient.

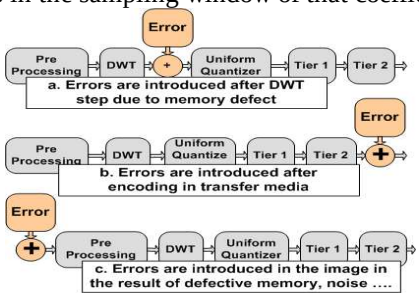


Figure 6: a. Memory error injection after DWT step
b. Errors due to noise in transfer median
c. Defect introduced to the original image (black and white noise and etc.)

Table 1: Encoder simulation settings

Parameter	Value
Offsets	Zeros
Tile size	512x512
Decomposition level	2
Precinct size	32x32
Block Size	32x32
Wavelet filters	(9,7) lossy
Layers	1
Marker segment	QOD
Image	Lena
Memory Size	786432 Byte
Number of Bank	4
# of Redundant Rows per bank	8

Errors due to RDF are uniform and randomly distributed so in the error injection step we randomly and uniformly corrupt the coefficient values by producing stuck at zero/one errors. In addition to error injection we need a comparison metric to compare the quality of decoded images with the original image. To achieve this, we have used the Peak Signal to Noise Ratio (PSNR) which is defined as follows:

$$MSE = \frac{\sum_{M,N}(I_1[m,n] - I_2[m,n])^2}{M \times N} \quad \textbf{(1)}$$

$$PSNR = 10 \times \log_{10}\left(\frac{R^2}{MSE}\right) \quad \textbf{(2)}$$

In equation (1), M and N are the number of rows and columns in the input images respectively, and R in the PSNR equation is the maximum pixel value in the image.

The simulation setting of the JPEG2000 encoder used for the simulation is illustrated in Table 1

B. Simulation results

Figure 7 illustrates the relation between voltage and Probability of failure for different protection policies. For this simulation, the word size was assumed to be four bytes and the probability of error versus voltage is derived from figure 1. The policies that are defined include protection of the first MSB to protection of the first 5 MSBs. These curves illustrate the achievable expanded design space. For example, a designer with the knowledge that a specific probability of failure is tolerable can choose the optimum operating voltage and protection scheme. To further elaborate on his point, assume a probability of failure of up to 2×10^{-2} is tolerable, and then a protection policy requiring the protection of only one MSB can operate at voltages down to 0.68v while 3 MSBs will require operation at 0.75v.

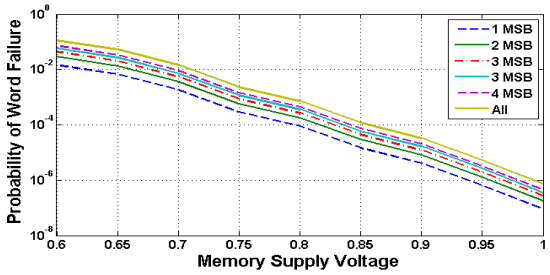
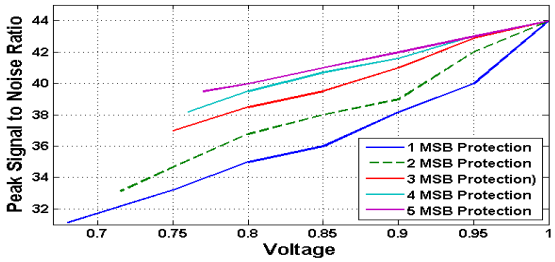


Figure 7: Probability of defective word in different protection policies versus voltage (only defects in protected bits is makes a word defect)

On the other hand, based on which protection scheme is used different pictures qualities result. It is important to correlate, voltage reduction, protection scheme and quality (reduction in PSNR).



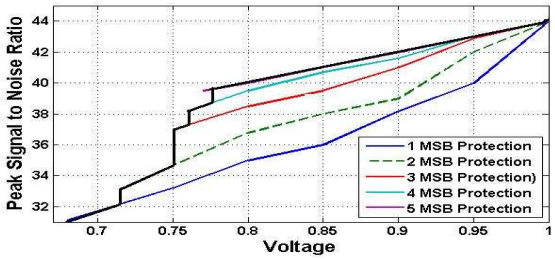


Figure 8:
a. Top: PSNR curve for different protection policies versus voltage
b. Bottom: Optimal operating point for the available redundancy budget.

Figure 8.a depicts these results (performed using the Lena image), where PSNR values are plotted for each protection scheme and voltage level averaged over 10000 random error injections for each voltage level. PSNR curves for some protection policies at lower voltages do not exist because as we explained previously each protection scheme passes the saturation point of the remapping architecture at a different point. Since always the best quality is desirable for each voltage, the protection policy that provides the highest PSNR at that voltage is the optimal choice. Figure 8.b illustrates the optimal operating point. For example, in figure 8.b at any voltage above 0.77V the 5 MSB protection provide the best quality but as the voltage drop to (0.77-0.76)V range the 5MSB protection is not possible due to saturation of available redundancy and a lower protection plan (4MSB in this case) should be used.

C. Power savings

The power savings associated with Vdd reduction are illustrated in Figure 9 which shows that a reduction of Vdd from 1.0v to 0.7v will result in more that 50% savings in dynamic power and 65% savings in leakage power. These savings are associated with the memory section of the encoder. In the following section, we will discuss the power reduction impact on the entire system (of which the memory is a subsystem) using two commercially available JPEG2000 encoders as case studies.

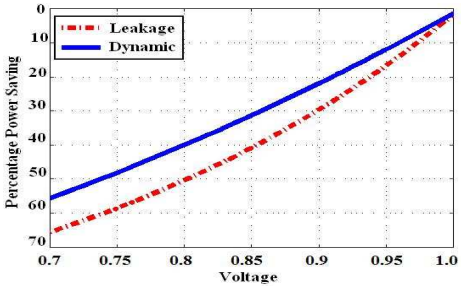


Figure 9: Power reduction with Vdd Scaling for 65 nm technology

V. POWER CONSUMPTION GAIN

We use two existing implementations of JPEG2000 encoders to validate our claims of power reduction. The first one is an Intellectual Property block (IP) from BARCO-SILEX that is described in 15. The second system is the ADV202 JPEG2000 codec chip from Analog Devices 16. The BARCO-SILEX IP implements the JPEG2000 encoding but requires additional tile memory of 512Kb (64KB) and another 135Kb of additional data storage. In our approach, the tile memory is error resilient and can be supplied at aggressively lower Vdd than the rest of the blocks. We used both TSMC 65nm LP process and G process standard cell libraries to estimate the power consumption of the core, Plogic. Additionally, we used CACTI [22] to estimate the tile memory's dynamic power consumption at the system's required operating frequency and 0.9v supply, Pmemd. Unfortunately, CACTI's estimates of leakage are not reliable at technologies below 90nm so we estimated the leakage at 180nm and scaled it based on TI's predictions by 10x 17 to obtain the memory leakage power, Pmeml. Figure 10 shows how Pmemd and Pmeml scale with Vdd. Since the power savings come only from scaling to the tile memory, the power savings percentage ratio, r is obtained by:

$$r = 100x\frac{P_{memd}(V_{ddl}) + P_{meml}(V_{ddl})}{P_{memd}(V_{dd0}) + P_{meml}(V_{dd0}) + P_{\log ic}(V_{dd0})} \quad \textbf{(3)}$$

Where Vdd0 is the nominal supply voltage and Vddl is the low supply voltage applied to the tile memories. As we can see the total savings, r can reach about 17% (25%) when the memory supply is at 0.7V and 25% (35%) when it is lowered to 0.6V in the LP process(G process) . The second system we considered is a complete JPEG 2000 codec SoC. The ADV 202 chip from analog devices includes all the memories, interfaces and control processor needed for encoding and decoding JPEG 2000 images. The tile memory is about 512KB. The chip was manufactured in 180nm technology so we scaled the logic power to 65nm. Similarly, we also used CACTI to estimate the memory's dynamic power and TI's leakage forecast to estimate its leakage power. Using the same method and Equation (3), the estimated savings are computed and shown in Figure 11. The savings in total system power are at 10% (16%) when the memory's Vdd is scaled to 0.7V and 15% (23%) when scaled to 0.6V in the LP (G) process.

VI. CONCLUSION

The paper proposes a means of identifying and remapping faulty parts of embedded memory in multimedia devices (specifically, JPEG 2000 encoder as a case study). The remapping allows the designer or user the flexibility to selectively choose different protection policies. Each protection policy in turn allows different levels of voltage scaling leading to aggressive power savings and improving overall yield by detection and correction of faults as they occur. We show that such an approach can lead to power savings of up to 35 % for commercial JPEG 2000 encoders while still maintaining target metrics such as PSNR levels.

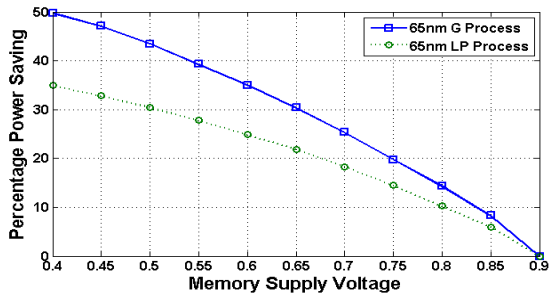


Figure 10: Estimated power savings percentage, r, for the JPEG 2000 IP in 15.

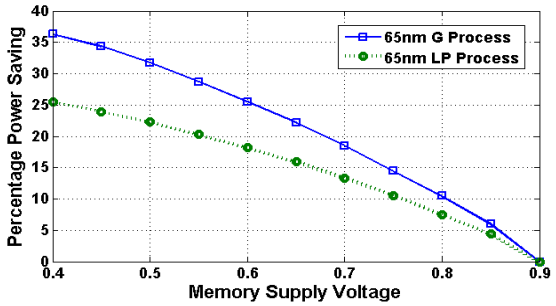


Figure 11: Estimated power savings percentage, r, for [17].

REFERENCES

1. S. R. Nassif, "Modeling and analysis of manufacturing variations," in Proceedings of Custom Integrated Circuit Conf., pp. 223–228, San Diego, CA, 2001,
2. C. Visweswariah, "Death, taxes and failing chips," in Proceedings of Design Automation Conf., pp. 343–347, Anaheim, CA, 2003.
3. S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variation and impact on circuits and microarchitecture," in Proceedings of Design Automation Conf., pp. 338–342, Anaheim, CA, 2003.
4. S. Mukhopadhyay, H. Mahmoodi, K. Roy, "Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in Nanoscaled CMOS" in Proceeding of IEEE Transactions on CAD of Integrated Circuits and Systems, Vol. 24, No. 12, Dec. 2005
5. J. Rabaey, A. Chandrakasan, B. Nilolic "Digital Integrated Circuit A design perspective" second edition, Pearson Education Inc, 2003.
6. C. Zhang, Y. Long and F. Kurdahi, "A Scalable Embedded JPEG2000 Architecture", in Embedded Computer Systems: Architecture, Modeling and Simulation, Lecture Notes in Computer Science LNCS 3553, Springer. (Hamalainen et. al eds.) 2005.
7. Y. Inoue, Y. Kajikawa, Y. Nomura, "Novel Motion-JPEG2000 video transmission system over CDMA environment" in Proceedings of the IEEE International Symposium on Communications and Information Technology, Page(s):265 - 268 vol.1 26-29 Oct. 2004
8. JPEG2000 image coding system, ISO/IEC International Standard 15444-1. ITU Recommendation T.800, 2000.
9. JPEG2000 Encoder Core. Cast Inc., Oct. 2002.
10. A. Agarwal, B.C. Paul, H. Mahmoodi, A. Datta, K. Roy, "A process-tolerant cache architecture for improved yield in nanoscale technologies," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Volume 13, Issue 1, Page(s):27 – 38, Jan. 2005.
11. M.A. Breuer, S.K. Gupta, T.M. Mak," Defect and error tolerance in the presence of massive numbers of defects," IEEE Design & Test of Computers, Volume 21, Issue 3, Page(s):216 – 227, May-June 2004.
12. S. Fossel, G. Fottinger, J. Mohr, "Motion JPEG2000 for high quality video systems," IEEE Transactions on Consumer Electronics, Volume 49, Issue 4, Page(s):787 – 791, Nov. 2003.
13. R. H. Chan, C. Wa Ho, and M. Nikolova "Salt-and-Pepper Noise Removal by Median-Type Noise Detectors and Detail-Preserving Regularization" in Proceeding of IEEE Transactions on Image Processing , Vol. 14, No. 10, Oct. 2005.
14. L. Bar, A. Brook , N. Sochen, N. Kiryati "Deblurring of Color Images Corrupted by Impulsive Noise" IEEE Transactions on Image Processing, This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.
15. http://www.barco.com/subcontracting/Downloads/IPProducts/BA112JPEG2000EFactSheet.pdf
16. http://www.analog.com/en/prod/0,,ADV202,00.html
17. http://www.cie-dfw.org/2006Convention/CIE-2006-microelctronic_Bill_Krenik.pdf
18. S. Shoukourian, V. Vardanian, Y. Zorian," SoC yield optimization via an embedded-memory test and repair infrastructure," IEEE Design & Test of Computers, Volume 21, Issue 3, Page(s):200 – 207, May-June 2004.
19. Amin Khajeh Djahromi, Ahmed M. Eltawil, Fadi Kurdahi and Rouwaida Kanj, UC Irvine and IBM "Cross Layer Error Exploitation for Aggressive Voltage Scaling" ISQED 2007.
20. S. Foessel, "Motion JPEG 2000 and digital cinema," presented at the SPIE Conf. Visual Communication and Image Processing, Lugano, Switzerland, 2003.
21. F. Dufaux and T. Ebrahimi, "Motion JPEG 2000 for wireless applications," presented at the SPIE Conf. Visual Communication and Image Processing, Lugano, Switzerland, 2003.
22. http://www.hpl.hp.com/personal/Norman_Jouppi/cacti4.html
23. http://www.ece.uvic.ca/~mdadams/jasper/
24. S. E. Schuster, "Multiple word/bit line redundancy for semiconductor memories," IEEE J. Solid-State Circuits, vol. SC-13, no. 5, pp. 698–703, Oct. 1978.
25. M. Horiguchi, "Redundancy techniques for high-density DRAMS," in Proc. 2nd Annu. IEEE Int. Conf. Innovative Systems in Silicon, Oct. 1997, pp. 22–29.
26. M. A. Makhzan (avesta sasan) , A. Eltawil, F. Kurdahi "Limits of Voltage Scaling for Caches Utilizing Fault Tolerant Techniques" Proc ICCD 2005
27. A. Argawal, B. C. Paul, S Mukhopadhyay, K. Roy "Process Variation in Embedded Memories: Failure Analysis and Variation Aware Architecture." IEEE Journal of Solid State Cuircuits, VOL. 40, NO. 9, SEPTEMBER 2005
28. H. L. Kalter et al., "A 50-ns 16-Mb DRAM with a 10 ns data rate and on-chip ECC," IEEE J. Solid-State Circuits, vol. 25, no. 5, pp. 1118–1128, Oct. 1990.
29. D. Weiss, J. J. Wuu, and V. Chin, "The on-chip 3-MB subarray-based third-level cache on an itanium microprocessor," IEEE J. Solid-StateCircuits, vol. 37, no. 11, pp. 1523–1529, Oct. 1990.
30. G. Sohi, ''Cache Memory Organization to Enhance the Yield of High Performance VLSI Processors'', IEEE Trans. Comp, vol.38(4), , pp.484-492, April 1989.
31. P. P. Shirvani and E. J. McCluskey, ''PADded Cache: A New Fault-Tolerance Technique for Cache Memories'', In Proc. Of 17th IEEE VLSI Test Symposium, pp.440-445, April 1999.
32. S. Mukhopadhyay, H. Mahmoodi, K. Roy "Modeling of Failiure Probability and Statistical Design of SRAM Array for Yield Enhancement in NanoScaled CMOS" CADICS Vol.24 NO. 12, DEC 2005
33. M. A. Makhzan (avesta sasan) , A. Eltawil, F. Kurdahi "Limits of Voltage Scaling for Caches Utilizing Fault Tolerant Techniques" Proc ICCD 2005
34. R. H. Chan, C. Wa Ho, and M. Nikolova "Salt-and-Pepper Noise Removal by Median-Type Noise Detectors and Detail-Preserving Regularization" in Proceeding of IEEE Transactions on Image Processing , Vol. 14, No. 10, Oct. 2005.
35. K. Nallaperumal, P. Kumar " An efficient Swithing Median Filter for Salt & Pepper Impulse Noise Reduction" DIM 2006 pages 161-166
36. K. Chinnasan, Y. Rangssanseri, O. Thitimajshima "Removing salt-and pepper noise in text/graphics images". APCCAS 1998 1998 pages 459-462
37. M. Makhzan, A. K. Djahromi, A. Eltawil, F. Kurdahi "A Low Power JPEG2000 Encoder with Iterative and Fault Tolerant Error Concealment" Trans on TVLSI 2008

Optimal Rate Allocation for Video Transmission over a Wireless Channel

Negar Nejati, Advisor: Hamid Jafarkhani
Center for Pervasive Communications and Computing
Department of Electrical Engineering and Computer Science
The Henry Samueli School of Engineering
University of California, Irvine
Email: nnejati@uci.edu

***Abstract*—Wireless channels introduce both packet erasures and bit errors which cause degradation in the quality of the transmitted video over such a channel. Previously we introduced an analytical expression for the expected distortion of a single layer encoded video bit-stream and based on the expected distortion model we proposed a distortion-optimal Unequal Error Protection (UEP) technique to transmit such a video bit-stream over wireless tandem channel. In Spring 2008, we worked on this distortion model to make it more accurate.**

I. Introduction

Temporally correlated tandem loss patterns of wireless channels appear in the form of bit errors related to fading and packet erasures related to network layer buffering. Most of the literature works consider just the effect of packet erasures on the quality of the transmitted video bitstream. Previously we introduced an accurate analytical model to capture the effects of both bit errors and packet erasures for a progressive enhancement layer of a compressed video bitstream [1]. In that case we assumed that the base layer is strongly protected by channel codes and the video codec was MoMuSys [2] [3] implementing MPEG4 standard.Then we introduced a simple analytical model for the base layer and based on that model we proposed a distortion-optimal technique to allocate the budget between different frames to minimize distortion. Our comparison results show that our transmission method can improve the quality of the received video over some previously proposed methods [4]. Fig. 1 shows these comparison results for different parity rates.

II. Distortion Model

To model distortion of a video sequence, first we model distortion of a Group of Pictures (GOP). Each GOP starts with an I frame which is intra-coded. Therefore the I frame will protect its GOP from the error propagation caused by error in previous frames. We assume that in the case of packet loss, receiver conceals an entire video packet by copying the shape and texture of the corresponding macroblocks in the previous VOP. The expected distortion of GOP can be represented by using table I. In table I each row indicates the pattern of the received frames in the GOP, and also the distortion of GOP associated with this pattern, after using the error concealment technique. These distortions are calculated based on the sequence. $P(1,j)$ indicates the probability of receiving frame j without error and it is calculated as follows:

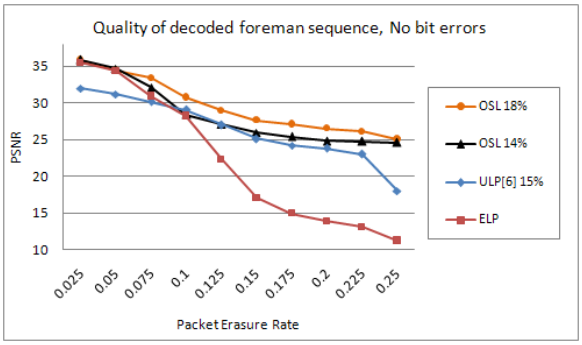


Fig. 1. Comparison results of OSL, ULP and ELP schemes for packet erasure channel.

TABLE I
Distortion of GOP based on the pattern of received frames of GOP

Frame 0	Frame 1	...	Frame n-1	Distortion
$P(0,0)$	$P(0,1)$	...	$P(0,n-1)$	d_0
$P(0,0)$	$P(0,1)$	...	$P(1,n-1)$	d_1
			...	
$P(1,0)$	$P(1,1)$	...	$P(1,n-1)$	d_{2^n-1}

$$P(1,j) = \prod_{k=0}^{N_j-1} P_{pck}(1,k) \tag{1}$$

where N_j defines the number of packets for transmitting frame j and $P_{pck}(1,k)$ represents the probability of receiving the kth packet of frame j free of error.

$$P_{pck}(1,k) = (1-P_{es})^L \tag{2}$$

L is the number of symbols in each packet. We assume each byte is a symbol and P_{es} is the probability of symbol error which is calculated using the channel model. $P(0,j) = 1 - P(1,j)$ is the probability of loss for frame j, respectively. $P^{i,j}$ defines the probability of frame j at row i in the table which is either $P(0,j)$ or $P(1,j)$ depending on the loss pattern of row i in the probability table of that GOP. Assuming n is the size of GOP, the total number of rows is 2^n. Using the introduced

table, the expected distortion of GOP k can be defined as follows:

$$\epsilon_d(k) = \sum_{i=0}^{2^n-1} \prod_{j=0}^{n-1} P^{i,j} d_i \quad (3)$$

Figure 2 shows the results of analytical model compared to the experimental distortion of the received video sequence in different channel conditions.

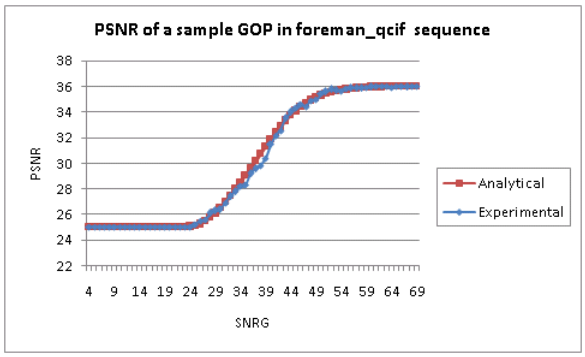


Fig. 2. A comparison of analytical and experimental distortion results of a sample GOP of Foreman sequence.

The distortion of a video sequence is related to the distortion of its GOPs. As it was mentioned before, I frames can stop error propagation but if there is any error in the I frame, both propagated errors from previous GOPs and low quality of the concealed I frame will effect other frames of that GOP. Capturing these propagation effects in modeling the distortion of a video sequence makes the model very complicated specially for optimization process. To model the video distortion and still keep the simplicity, we have assumed that I frames can be transmitted using stronger error correcting codes. It meas that if most intra-coded frames are received the effect of error propagation gets less and less. In such a case we can model the distortion of the whole sequence as the summation of distortion of all GOPs. We used different levels of error protection for the I frames of the sequence and it shows that as the error protection code gets stronger, the difference between the analytical model of the sequence gets closer to the experimental results. This means that if we could use enough budget to protect the I frames, the analytical model of the sequence distortion could be simplified as the summation of the distortion of GOPs.

III. Future Work

Next we want to used the new model to solve the transmission budget allocation optimally among all the sequence frames for the base layer bitstream and then combine this work with the previously proposed technique for the progressive enhancement layer.

References

[1] N. Nejati, H. Yousefi'zadeh, and H. Jafarkhani, "Wireless video transmission: A distortion-optimal approach," *DCC*, Mar 2008.

[2] "General project information, the momusys web site," *available at http://www.tnt.uni-honnover.de/project/eu/momusys/.*

[3] "Final report on standardisation activities, MoMuSys Public Deliverable," *available at http://www.tnt.uni-honnover.de/project/eu/momusys/doc/mom-docs.html.*

[4] X. Y. et al., "Unequal loss protection for robust transmission of motion compensated video over the internet," *Signal Process. Image Communications*, March 2003.

UC IRVINE CENTER FOR PERVASIVE COMMUNICATIONS AND COMPUTING
Graduate Fellowship - Progress Report for Spring 2008

Ph.D. Student: Hulya Seferoglu
CPCC Affiliate Professor: Athina Markopoulou
Project: Rate-Allocation in Wireless Networks with Network Coding

Overview: During spring 2008, we continued our research on *rate allocation for network coding over wireless*. Specifically, we have developed an optimization model, proposed a distributed solution and conducted some simulations.

Rate-Allocation for Wireless Networks with Network Coding: Resource allocation is an important class of problem for both wired and wireless networks; the goal is to use network resources efficiently and also provide fair resource allocation among multiple users. Most of the work in this area follows the seminal work of Kelly et al. [1]. Recently, there has been increased interest in resource allocation problems in wireless networks [2]. This problem is already challenging due to broadcast nature of wireless and of the presence of multiple-hops. It becomes even more challenging when network coding is employed. In this project, we study rate allocation in wireless networks with network coding.

Motivation:
Let's discuss the basic example of Fig. 1 that demonstrates the main idea. Nodes A and B communicate over a relay R in a wireless network where A and B do not hear each other. Let us suppose that node A transmits at rate r_1 and node B transmits at rate r_2, and assume that node R does a basic network coding operation (XORing of packets from node A and node B) and broadcasts, as in [3]. If $r_1 = r_2$, R will combine two flows from A and B and broadcast; therefore, the relay transmits at rate $r_3 = r_1 = r_2$. When $r_1 > r_2$, the relay node will transmit at a rate $r_3 = r_1$, because it should serve all packets from node A with rate r_1 and XOR the packets from node B with these packets. The reverse is also true. We can conclude that the relay node should transmit at a rate $r_3 = \max(r_1, r_2)$. This example shows that when network coding is used, an intermediate node (R in this example) may transmit at lower rates (i.e. $\max(r_1,r_2)$) than the sum of all rates in the neighborhood (r_1+r_2). This observation affects the capacity region and the rate allocation problem should be formulated considering this fact.

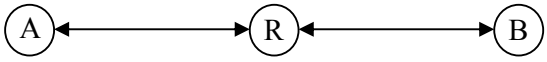


Figure 1: Cross topology with two nodes talking to each other

Optimization Model:
Consider that each source in a network transmits at a rate σ^t and has associated utility function $U(\sigma^t)$. Our goal is to optimize the aggregated utility function subject to capacity constraints. The rate allocation problem can be formulated as follows:

$$\max_{\sigma^t} \quad U(\sigma^t)$$

$$s.t \quad \sum_j x_{ij}^t - \sum_j x_{ji}^t = \sigma^t I_{\{i=f\}}, \quad \forall i,t$$

$$\sum_{t \in \phi_{J_n^{(i)}}, j \in J_n^{(i)}} x_{ij}^{tn} \le z_{iJ_n^{(i)}}, \quad \forall i,n$$

$$\sum_{\left\{n | t \in \phi_{J_n^i}\right\}} x_{ij}^{tn} = x_{ij}^t, \quad \forall i,j,t$$

where source t is transmitted from node i to node j with rate $x_{i,j}{}^{t}$. The first constraint expresses the flow conservation. The second constraint is due to the broadcast advantage provided by network coding. We assume that at each node i, there are n different network coding opportunities. The destination node set and flow set corresponding to network code n at node i are defined as $J_n^{(i)}$ and $\sigma_{J_n^{(i)}}$ respectively. The third constraint is the flow conservation for different network codes where x_{ij}^{tn} is a data rate assigned to flow t from node i to node j for network code n.

Solution:
When we solve the problem using Lagrangian optimization and do algebraic manipulations, the solution consists of two parts;

Rate Control:

$$\max_{\sigma^t} \quad (U(\sigma^t) - q_f^t \sigma^t)$$

Scheduling:

$$\max \quad \sum_{i,j,t} x_{ij}^t \left(q_i^t - q_j^t\right)$$

$$s.t. \quad \sum_{t \in \phi_{J_n^{(i)}}, j \in J_n^{(i)}} x_{ij}^{tn} \le z_{iJ_n^{(i)}}, \quad \forall i, n$$

$$\sum_{\left\{n | t \in \phi_{J_n^i}\right\}} x_{ij}^{tn} = x_{ij}^t, \quad \forall i, j, t$$

where q_i^t is a Lagrange multiplier and representative of the queue size at node i of flow t.

In the spring quarter, we reviewed the extensive literature in the areas of cross-layer optimization in wireless networks and network coding (especially within the linear programming framework). We then formulated and solved the problem as outlined above and conducted numerical simulations to confirm the optimality and stability of the solution.

Future Directions:
We are currently developing efficient sub-optimal algorithms. We will also conduct simulations over more realistic environments, study the interaction of the proposed algorithms with current protocols and compare it to current rate control implementation of TCP. We are currently working towards a paper submission for Infocom'08.

References:
[1] F. Kelly, A. Maulloo, D. Tan, "Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stability," *Journal of the Operations Research Society,* vol. 49 (3), pp. 237 – 252, 1998.
[2] X. Lin, N. B. Shroff, R. Srikant, "A Tutorial on Cross Layer Optimization in Wireless Networks," *IEEE Journal on Selected Areas in Communications,* vol. 24 (8), pp. 1452 – 1463, 2006.
[3] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Medard, and J. Crowcroft, "XORs in the air: Practical network coding," *in Proc. of ACM SIGCOMM*, vol. 36(4), pp. 243-254, Pisa, Italy, Sept. 2006.

Progress Report Spring 2008: Interference Alignment and Degrees of Freedom of Interference Channels under Channel Estimation Error

S. Surendran
Under the supervision of Prof. Syed A. Jafar
Center for Pervasive Communications and Computing
Department of Electrical Engineering and Computer Science
University of California, Irvine
Irvine, California, 92697, USA
Email: ssurendr@uci.edu

***Abstract*— Interference alignment is a powerful technique to achieve significant degrees of freedom for the interference channel. However, perfect channel state information (CSI) is assumed. Our objective is to characterize the effect of channel estimation error on interference alignment for interference channels. In particular, we wish to obtain scaling laws that govern the relationship between achievable degrees of freedom and channel estimation error. An appropriate channel estimation scheme for our problem is the maximum likelihood (ML) channel estimation procedure. An error model based on the ML scheme is used in conjunction with interference alignment.**

I. 3 USER MIMO INTERFERENCE CHANNEL

Consider a 3 user ($K = 3$) MIMO interference channel with M antennas at each node. For this channel $3M/2$ degrees of freedom can be obtained with constant channel matrices with $M > 1$ antennas at each node [1]. Each user has $M/2$ degrees of freedom yielding a total of $3M/2$ degrees of freedom for the network.

To illustrate the effect of channel estimation error on interference alignment, consider a 3 user interference channel with $M = 2$ antennas. Interference alignment is performed using $\hat{H}_{ML}^{[ij]}$ from transmitter j ($j = 1, 2, 3$) to receiver i ($i = 1, 2, 3$). Note that $\hat{H}_{ML}^{[ij]} = H^{[ij]} + \Delta H^{[ij]}$; now let us consider the received signal at receiver 1 ($i = 1$):

$$\begin{aligned} Y^{[1]} = {} & H^{[11]}V^{[1]}X^{[1]} + H^{[12]}V^{[2]}X^{[2]} + H^{[13]}V^{[3]}X^{[3]} \\ & + \Delta H^{[11]}V^{[1]}X^{[1]} + \Delta H^{[12]}V^{[2]}X^{[2]} \\ & + \Delta H^{[13]}V^{[3]}X^{[3]} + Z^{[1]} \end{aligned} \tag{1}$$

where $V^{[j]}$ and $X^{[j]}$ represent the beamforming and input vectors, respectively, corresponding to transmitter j, and $Z^{[1]}$ is the additive white noise at receiver 1.

Once the set of appropriate alignment vectors U^1 has been chosen, U^1 is orthogonal to both $H^{[12]}V^{[2]}$ and $H^{[13]}V^{[3]}$, and interference terms due to transmitters 2 and 3 are canceled. Thus we have:

$$\begin{aligned} y_1 = {} & < Y^{[1]}, U^1 > \\ = {} & < H^{[11]}V^{[1]}X^{[1]}, U^1 > + < H^{[12]}V^{[2]}X^{[2]}, U^1 > \\ & + < H^{[13]}V^{[3]}X^{[3]}, U^1 > + < \Delta H^{[11]}V^{[1]}X^{[1]}, U^1 > \\ & + < \Delta H^{[12]}V^{[2]}X^{[2]}, U^1 > + < \Delta H^{[13]}V^{[3]}X^{[3]}, U^1 > \\ & + Z^{[1]} \end{aligned} \tag{2}$$

Now U^1 is orthogonal to both $H^{[12]}V^{[2]}$ and $H^{[13]}V^{[3]}$, and $< ., . >$ represents inner-product, we have:

$$\begin{aligned} y_1 = {} & < Y^{[1]}, U^1 > \\ = {} & < H^{[11]}V^{[1]} + \Delta H^{[11]}V^{[1]}, U^1 > X^{[1]} \\ & + < \Delta H^{[12]}V^{[2]}, U^1 > X^{[2]} + < \Delta H^{[13]}V^{[3]}, U^1 > X^{[3]} \end{aligned} \tag{3}$$

Note that if we were to have perfect channel estimation, that is, $\Delta H^{[ij]} = 0$, we have would have full degrees of freedom. So the effect of white noise matrices is to introduce interference, and we would want the estimation error to scale as the inverse of the signal-to-noise ratio (SNR).

II. SIGNAL-TO-NOISE RATIO RELATIONSHIPS

Let

$$\alpha = H^{[11]}V^{[1]} + \Delta H^{[11]}V^{[1]}. \tag{4}$$

$$\beta = \Delta H^{[12]}V^{[2]}. \tag{5}$$

$$\gamma = H^{[13]}V^{[3]}. \tag{6}$$

Then the signal-to-noise ratio at receiver 1 is

$$SNR_1 = \alpha^2 P/(\beta^2 P + \gamma^2 P + 1) \tag{7}$$

where P is the transmit power/user and the additive white noise at receiver 1 is assumed to have unit variance.

REFERENCES

[1] V. R. Cadambe and S. A. Jafar, "Interference Alignment and the Degrees of Freedom for the K User Interference Channel," *arXiv: 0707.0323, preprint*.