LINEAR VERSUS CHANNEL CODING TRADE-OFFS IN FULL DIVERSITY FULL RATE MIMO SYSTEMS

Abdelkader Medles, Dirk T.M. Slock

Eurecom Institute, 2229 route des Crêtes, B.P. 193, 06904 Sophia Antipolis Cedex, FRANCE
Tel: +33 4 9300 2916/2606; Fax: +33 4 9300 2627
e-mail:{medles,slock}@eurecom.fr

ABSTRACT

The use of multiple transmit (TX) and receive (RX) antennas allows to transmit multiple signal streams in parallel and hence to increase communication capacity. We have previously introduced simple convolutive linear precoding schemes that spread transmitted symbols in time and space, involving spatial spreading, delay diversity and possibly temporal spreading. Such linear precoding allows to attain full diversity without loss in ergodic capacity. Linear precoding however cannot provide coding gain. Hence practical transmission systems have to involve channel coding. Threading is an example of a MIMO transmission system in which spatial diversity gets exploited via channel coding only. Practical symbol constellations however only allow the exploitation of a limited diversity order by the channel coding. Hence powerful yet simple MIMO TX schemes can be obtained by combining the coding gain and diversity exploitation of classical channel codes with linear precoding to exploit the remaining diversity degrees. A typical design would use channel coding to exploit temporal fading with linear precoding to exploit spatiofrequential fading.

1. INTRODUCTION

The $N_{tx} \times N_{rx}$ MIMO system is essentially described by

$$\boldsymbol{y}_k = \mathbf{H}(q)\, \boldsymbol{a}_k + \boldsymbol{v}_k = \mathbf{H}(q)\, \mathbf{T}(q)\, \boldsymbol{b}_k + \boldsymbol{v}_k \qquad (1)$$

where the white noise power spectral density matrix is $S_{\boldsymbol{v}\boldsymbol{v}}(z) = \sigma_v^2\, I$, and $q^{-1}\, \boldsymbol{b}_k = \boldsymbol{b}_{k-1}$. We consider the case of channel state information being absent at the transmitter (TX) and perfect at the receiver (RX). The linear precoding considered here (introduced in Allerton01 and further analyzed in [1]) consists of a modification of VBLAST, obtained by inserting a square matrix prefilter before inputting the vector signal $\boldsymbol{b}_k$ into the channel $\mathbf{H}(q)$. The $N_s = N_{tx}$ ("full rate") component signals of $\boldsymbol{b}_k$ are called streams or layers. The suggested prefilter is

$$\begin{array}{lcll} \mathbf{T}(z) & = & \mathbf{D}(z)\, \mathbf{Q}\,, & |\mathbf{Q}_{ij}| = \frac{1}{\sqrt{N_{tx}}} \\ \mathbf{D}(z) & = & diag\{1,\, z^{-1}, \ldots, z^{-(N_{tx}-1)}\}\,, & \mathbf{Q}^H \mathbf{Q} = I \end{array}$$

Note that for a channel with delay spread L, the prefilter can be immediately adapted by replacing the elementary delay z^{-1} by z^{-L}, though we mostly focus on the flat channel case, in which case symbol stream n ($b_{n\,,k}$) passes through the equivalent SIMO channel $\sum_{i=1}^{N_{tx}} z^{-(i-1)} \mathbf{H}_{:,i} \mathbf{Q}_{i,n}$ which now has memory due to the delay diversity introduced by $\mathbf{D}(z)$. It is important that the different columns $\mathbf{H}_{:,i}$ of the channel matrix get spread out in time to get full diversity (otherwise the streams just pass through a linear combination of the columns, as in VBLAST, which offers limited diversity). The delay diversity only becomes effective by the introduction of the spatial spreading matrix $\mathbf{Q}$, which has equal magnitude elements for uniform diversity spreading (a specific choice for $\mathbf{Q}$ exists for maximum coding gain in case of QAM symbols [1]). We can see that each symbol stream has the same Matched Filter Bound (MFB), which is proportional to the channel Frobenius norm, hence full diversity is exploited. Also, since the prefilter $\mathbf{T}(z)$ is paraunitary and leaves the white stream $\boldsymbol{b}_k$ white, no loss in ergodic capacity is incurred.

2. LINEAR PRECODING + CHANNEL CODING

Linear Precoding was introduced to exploit the transmit diversity, leading to a maximum diversity gain (exponent of the error probability: $N_{tx} N_{rx}$ (flat channel case)). This gain corresponds to the slope of the error probability vs SNR curve (in logarithmic scale), but to achieve this regime (fastly decaying error probability (Pe)) we need an SNR such that $\rho = \frac{\sigma_b^2}{\sigma_v^2} = \frac{SNR}{N_{tx}} >> 4\left(\frac{1}{coding gain}\right)^{\frac{1}{N_{tx}}} = \frac{2(4M^2-1)N_{tx}}{3}$. This SNR range is out of scope for practical systems. For lower SNR values, it will be important to improve the position of the Pe curve by increasing the coding gain via channel coding, see fig. 1.

For channel decoding we can use an iterative decoder that combines a SISO decoder with a MIMO linear filter and Interference Canceller(IC), this is represented in fig. 2. This decoder structure was first used for CDMA reception [2], and was then proposed for the MIMO reception [3],[4],[5], it is the analog to the turbo detection when the the mapping, Linear Precoding and the channel, resp. the channel coding, are seen as Inner coding, resp. Outer coding. This structure of the decoder is shown to give a good performances for small size constellations and exploits the diversity when a LMMSE front-end equalizer is used ([4]). In the following we will give a short overview of the iterative channel decoding with interference cancellation, and for simplicity we will denote the overall channel $\mathbf{H}\,\mathbf{T}(q)$ by $\mathbf{G}(q)$, in the sense that the Linear Precoder transforms the flat channel in a frequency-selective channel in order to exploit the spatial multi TX antenna diversity.

Eurécom's research is partially supported by its industrial partners: Ascom, Swisscom, Thales Communications, ST Microelectronics, CEGETEL, France Télécom, Bouygues Telecom, Hitachi Europe Ltd. and Texas Instruments. The work reported herein was also partially supported by the French RNRT project ERMITAGES.

2.1. Encoding

Fig. 1 shows the encoding operation. The channel coder output is followed by the interleaver, the output is then mapped into symbols before the Serial to Parallel (S/P) conversion. The symbol vector $\boldsymbol{b}_k$ is then filtered by $\mathbf{T}(z)$. If the iterative decoder would succeed to cancel all the interference (genie aided decoder), then each symbol would be interfered only by noise. Performance would then be the matched filter bound, which corresponds to full diversity exploitation. The channel coder and interleaver are then only used to lower the error probability (coding gain). Now, by considering the overall channel and the channel code as the two constituents of a serial turbo code, then lowering of the error probability can be obtained by increasing the minimum distance. Therefore a good choice for the interleaver for large frame size is to choose a random interleaver, without specific structure.

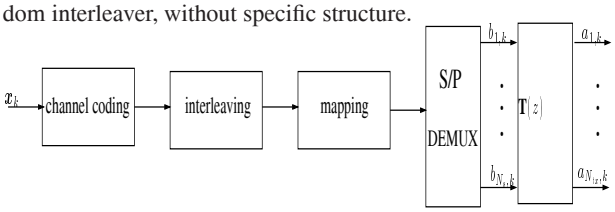

Fig. 1. Encoder for Space-Time Spreading.

2.2. Iterative decoding

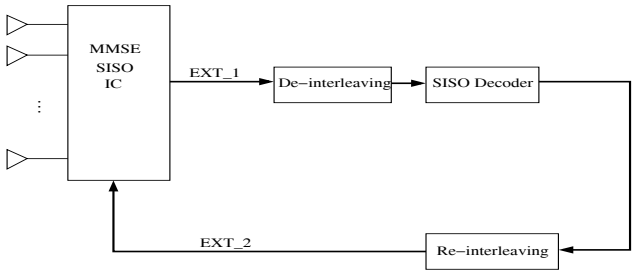

Fig. 2. Iterative decoder with interference cancellation.

In this section we propose an iterative decoding strategy for a general block fading channel. We consider an iterative decoding scheme with Interference Cancellation(IC), see fig. 2. The first block of the scheme contains the IC operation followed by a MIMO linear equalizer, a symbol to bit demapping and de-interleaving. The second block of the decoder is the maximum-a-posteriori(MAP) soft-input soft-output(sIsO, to distinguish from SISO) decoder of the channel code (for instance, we use a convolutional code and the corresponding BCJR SISO decoder [6]) followed by the interleaver and the bit to symbol mapping. These two blocks exchange information in the form of log-likelihood ratios (LLRs) during iterations, the overall decoder can be seen as an application of the belief propagation principle, known also as the sum-product algorithm [2],[3]. We assume that the residual interference plus noise at the output of the equalizer follows a Gaussian distribution. This is cleary an approximation, however it tends to be valid for large systems (large N_{tx} and/or delay spread), see [2] for the case of CDMA.

As linear equalizer we use the Ubiased MMSE (UMMSE) design, where

$$\mathbf{f}_n^{(i)}(z) = \frac{1}{\frac{1}{2\pi j}\oint \frac{dz}{z}\rho \mathbf{G}_{:,n}^{\dagger}(z)\mathbf{R}^{(i)}(z)^{-1}\mathbf{G}_{:,n}(z)}\mathbf{G}_{:,n}^{\dagger}(z)\mathbf{R}^{(i)}(z)^{-1}$$

is the equalizer filter for $\mathbf{T}(z)$ input (stream) n of the MIMO system at iteration i, $\mathbf{R}^{(i)}(z) = \sigma_v^2 I + \sum_{n=1}^{N_{tx}} \tilde{\sigma}_{b_n}^{2\,(i-1)} \mathbf{G}_{:,n}(z)\mathbf{G}_{:,n}^{\dagger}(z)$ is the spectrum of the noise plus residual interference and $\tilde{\sigma}_{b_n}^{2\,(i)} = \mathrm{E}\,|\tilde{b}_{n,k}^{(i)}|^2 = \mathrm{E}\,|b_{n,k} - \hat{b}_{n,k}^{(i)}|^2$ is the variance of the residual interference at the input n (of $\mathbf{T}(z)$). $\hat{b}_{n,k}^{(i)} = E(b_{n,k}\,|\,\mathrm{EXT}_2^{(i)})$ is the MMSE estimate of $b_{n,k}$ based on the information contained in $\mathrm{EXT}_2^{(i)}$. For the residual interference spectrum we assume that the residual interference $\tilde{b}_{n,k}$ is temporally and spatially white and decorrelated from the noise. This approximation is again valid for large systems (and hence works better when linear precoding is used). Finally, the equalizer output on interference cancelled received signal at iteration (i) for stream n at time k is:

$$s_{n,k}^{(i)} = \mathbf{f}_n^{(i)}(q)\ \left(\boldsymbol{y}_k - \mathbf{G}(q)\hat{\boldsymbol{b}}_k^{(i-1)}\right) + \hat{b}_{n,k}^{(i-1)} \qquad (2)$$

Due to the unbiasedness of $\mathbf{f}_n^{(i)}$, $\hat{b}_{n,k}^{(i-1)}$ does not appear in $s_{n,k}^{(i)}$. Let's denote the bit-to-symbol mapping by $\mu : \mathbb{F}_2^P \to \mathcal{A}$, where $\mathbb{F}$ is the binary alphabet and $P = \log_2(|\mathcal{A}|)$ is the number of coded (and interleaved) bits per symbol: $b_{n,k} = \mu(x_{n,k}^1, \ldots, x_{n,k}^P)$. The extrinsic information of the p-th bit of the binary mapping of the k-th symbol of stream n at the output of the IC in the (i)-th iteration is:

$$\mathrm{EXT}_2^{(i)}(x_{n,k}^p) = \log\frac{p(x_{n,k}^p=1\,|\,s_{n,k}^{(i)},\mathbf{G})}{p(x_{n,k}^p=0\,|\,s_{n,k}^{(i)},\mathbf{G})} = \log\frac{\sum_{b_{n,k}\in\mathcal{A}\,|\,x_{n,k}^p=1} p(s_{n,k}^{(i)}|b_{n,k},\mathbf{G})\ \exp\left(\sum_{p'=1,\neq p}^{P}\mathrm{EXT}_1^{(i-1)}(x_{n,k}^{p'})\right)}{\sum_{b_{n,k}\in\mathcal{A}\,|\,x_{n,k}^p=0} p(s_{n,k}^{(i)}|b_{n,k},\mathbf{G})\ \exp\left(\sum_{p'=1,\neq p}^{P}\mathrm{EXT}_1^{(i-1)}(x_{n,k}^{p'})\right)}$$

where $p(s_{n,k}^{(i)}|b_{n,k},\mathbf{G})$ is evaluated by assuming that $\bar{v}_{n,k}^{(i)} = s_{n,k}^{(i)} - b_{n,k}$ is an AWGN. After de-interleaving, the EXT_1 information sequence is used as a priori LLR input to the MAP decoder of the channel code which is a convolutional code in our case. Using the forward-backward BCJR algorithm, the a posteriori LLR is calculated and the extrinsic information is defined as $\mathrm{EXT}_2^{(i)} = \mathrm{MAP}(\mathrm{EXT}_1^{(i)}) - \mathrm{EXT}_1^{(i)}$. Experimentally we observed that the number of iterations needed for the convergence of this algorithm is small, typically 3 or 4 iterations.

Remarks:

- For a flat channel with $N_s = N_{tx}$ we can show by induction that: $\tilde{\sigma}_{b_n}^{2\,(i)} = \tilde{\sigma}_b^{2\,(i)},\ n = 1, \ldots N_{tx}$, and that $\mathbf{f}_n^{(i)}(z) = \frac{N_{tx}/\rho}{\mathrm{tr}\{\mathbf{H}^H\mathbf{R}\mathbf{H}\}}\mathbf{T}_{:,n}^{\dagger}(z)\mathbf{H}^H\mathbf{R},\ \mathbf{R} = \left(\sigma_v^2 I + \tilde{\sigma}_b^{2\,(i-1)}\mathbf{H}\mathbf{H}^H\right)^{-1}$. This simplifies the equalization complexity: the joint equalizer for all streams consists of a channel equalizer followed by a precoder matched filter (= precoder inverse since $\mathbf{T}(z)$ is paraunitary). Even if these results cannot be generalized to the frequency selective channel case, experiments show that performance does not degrade using the suboptimal approach: average the $\tilde{\sigma}_{b_n}^{2\,(i)}$'s over streams, use a frequency-selective channel equalizer followed by the precoder matched filter.
- The stream equalizer can be designed using criteria other than UMMSE, typically Zero-Forcing (ZF) or channel Matched Filter (MF). These two alternative designs lead to performance loss. Especially the MF design leads to an error floor and therefore doesn't exploit the diversity.

2.3. Complexity Comparison with Threading

Whereas the proposed Space-Time Spreading (STS) strategy differs from VBALST by the insertion of the time-invariant precoder filter $\mathbf{T}(z)$, in Threading [5] $\mathbf{T}(z)$ is replaced by a periodically

time-varying cyclic shift matrix $\mathbf{Z}^{k \mod N_{tx}}$ where $\mathbf{Z}$ is the elementary circulant shift matrix. Comparing STS to Threading, in the encoding part, the precoding with $\mathbf{T}(z)$ leads to an additional complexity $\mathcal{O}(\log_2(N_{tx}))$ per symbol period (when we use the special structure of $\mathbf{Q}$ for $N_{tx} = 2^{n_t}$). This is a negligible increase compared to the remaining operations (channel coding and pulse shaping). At the receiver side the additional complexity comes from the inverse operation of the precoding, the matched filter $\mathbf{T}^{\dagger}(z)$, with same negligible complexity increase.

3. MULTI-BLOCK TIME DIVERSITY

In the usual SISO fading channel problem, time diversity of the channel resulting for the variation from block to block is used to improve performances. We can exploit block fading also for the MIMO channel. Below, we discuss how to differentially exploit the diversity sources in STS.

3.1. Combining Linear Precoding and Channel Coding to Exploit Time Diversity

We consider a block-fading environment with F i.i.d. blocks. In the STS approach, the spatiofrequential diversity is exploited in each block by the linear precoding. The problem of additionally exploiting temporal diversity is then reduced to the SISO channel fading problem. If we denote by d_F the diversity exploited in this latter problem, the overall diversity exploited is then $d = d_F\, N_{tx}\, N_{rx}\, (\times L)$. To exploit temporal diversity, we need to first use a block interleaver on the ensemble of fading blocks (see fig. 3), and then we apply a random interleaver within each block. Using a genie aided reasoning, the temporal diversity that can be exploited is limited by the fundamental Singleton Bound (SB) [7]:

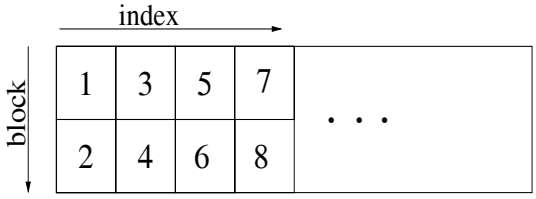

Fig. 3. Block interleaver for F=2.

$$d_F \leq 1 + \lfloor F_d(1 - \frac{r}{|\mathcal{A}|})\rfloor \leq F_d \qquad (3)$$

where r is the rate of the channel code, $\lfloor . \rfloor$ denotes the flooring operation, and F_d ($= F$ here) is the number of diversity branches. Table 1 gives the SB and the temporal diversity exploited by a set of channel convolutional codes with $r = \frac{1}{2}$ and for different symbol constellations and different number F of blocks. In the case of Threading no linear precoding is used to help the channel code exploit all the diversity sources. The number of diversity branches is here $F_d = FN_{tx}$([3]). By applying the same reasoning as before, the (source) diversity (d_s) exploited by the channel code is bounded by:

$$d_s \leq 1 + \lfloor FN_{tx}(1 - \frac{r}{|\mathcal{A}|})\rfloor \leq FN_{tx}\ . \qquad (4)$$

The overall exploited diversity is then $N_{rx}\, d_s$. From the table we can see that for $r = \frac{1}{2}$ and using convolutional code the effectively exploited diversity degree d_s is far from the available one (FN_{tx}). This even holds for SB for large number of diversity branches ($F_d \geq 8$). The only way to exploit higher diversity is by lowering the rate and increasing the constellation size. This leads to high decoding complexity and low performance (in comparison with the case of linear precoding).

d_F or d_s		BPSK			QPSK		
States	Generators	F_d=2	4	8	2	4	8
Singleton Bound		2	3	5	2	4	7
4	(5,7)	2	3	4	2	3	3
8	(15,17)	2	3	4	2	3	4
16	(23,35)	2	3	5	2	3	4
32	(53,75)	2	3	5	2	3	4
64	(133,171)	2	3	5	2	3	5

Table 1. Block diversity for some popular rate 1/2 binary convolutional codes mapped onto BPSK and QPSK (with Gray labeling). Code generators are expressed in octal notation.

Remarks:

- Using the SB we can interpret why the proposed linear precoding achieves full diversity for a single block MIMO channel. In fact, prefiltering the QAM constellation increase the constellation size at the channel input from $|\mathcal{A}|$ to $|\mathcal{A}|^{N_{tx}}$. Therefore the SB becomes: $1 + \lfloor N_{tx}(1 - \frac{1}{|\mathcal{A}|^{N_{tx}}})\rfloor = 1 + \lfloor N_{tx} - 2^{n_t - P\, 2^{n_t}}\rfloor$ where $n_t = \log_2(N_{tx}) \geq 1$ and $P = \log_2(|\mathcal{A}|) \geq 1$. These two last conditions imply that $0 \leq 2^{n_t - P\, 2^{n_t}} \leq 1$ and finally that the SB equals N_{tx}.
- Two other recent approaches are the Complex Field Coding approach of Giannakis *etal.* and the Universal Coding approach of El Gamal *etal*. These approaches (similar to earlier work by Belfiore *etal.*) correspond to linear dispersive block codes (LDBC) with block length equal to N_{tx}. As a result, each transmitted symbol sees a different SINR and symbol-independent equalization or residual interference variance in a turbo detection approach do not apply. That's why these authors consider other MLSE detection approximations in the form of sphere decoding. On the other hand, STS can also be seen as an instance of LDBC in which the z transform of the basis matrices corresponds to $z^{-k}\mathbf{T}_{:,n}(z)$ for all possible delays k within a (arbitrarily long) block. It can be shown that the coding gain increases with block length. Also, STS immediately applies in the case of delay spread $L > 1$ (frequency fading), whereas the other approaches (and also Threading) only apply for $L = 1$.

4. PERFORMANCE ANALYSIS

We compare the performance of STS and Threading via simulation. We use for both a rate 1/2, (5,7) four states convolutional code, convolutional to take advantage of the availability of computationally efficient sIsO decoders (BCJR). Performances are evaluated in terms of frame error rate (FER) as a function of E_b/N_0 ($SNR = R\, E_b/N_0, R = rN_{tx}\log_2|\mathcal{A}|, \rho = SNR/N_{tx}$). We run simulations for an input frame of 512(1024) information bits for $N_{tx} < (=)8$. We fix the number of decoding iterations to 5. We use QPSK with Gray labeling.

Comparison of Threading and STS In fig. 4, for $F = 1, 2, 4$ blocks, we see that STS (solid lines) succeeds in exploiting more diversity than Threading (dash-dot) except for $F = 1$ block. E.g. for $F = 2$, the asymptotic slope for ML decoding would be proportional, for Threading, to $d_s N_{rx} = 3 \times 2 = 6$ and for STS to

$d_F N_{tx} N_{rx} = 3 \times 2 \times 2 = 12$. In fig. 5, the slopes roughly double when delay spread L doubles. In fig. 6, when the number of antennas double, STS and Threading differentiate even for $F = 1$ block and the slopes again increase when the number of antennas further double in fig. 7. The increase in number of antennas (N_{rx}) also leads to an array gain and hence a translation of the curves to the left. In fig. 8, we consider the case $2 = N_{rx} < N_{tx} = 4$. For STS, we vary the number of streams N_s by varying the number of inputs to $\mathbf{T}(z)$. With $N_s = 2$, STS achieves the same diversity in a 2×4 MIMO system with $F = 1$ as in a square 2×2 system with $F = 2$: we observe equivalence between N_{tx}/N_{rx} and F. We also consider the Space Time Orthogonal Design (STOD) of Tarokh which leads to the leftmost curve, but at rate 0.75b/s/Hz. We see that at 2 b/s/Hz (the two middle curves), STS (solid) with $N_s = 2$ ("half rate": $\frac{N_s}{N_{tx}} = \frac{1}{2}$) and QPSK performs much better than Threading (dash-dot) with $N_s = 4$ ("full rate") and BPSK.

5. REFERENCES

[1] A. Medles and D.T.M. Slock. Multistream space-time coding by spatial spreading, scrambling and delay diversity. In *In Proc. ICASSP*, Orlando, FL, May 2002.

[2] J. Boutros and G. Caire. Iterative joint decoding: Unified frame work and asymptotic analysis. *IEEE Trans. on Inform. Theory*, 48(7):1772–1793, July 2002.

[3] G. Caire and A. Guillen i Fabregas. Design of space-time bit-interleaved coded modulation for block fading channels with iterative decoding. In *In Proc. 37th Conf. on Information Sciences and Systems (CISS)*, Baltimore, USA, March 2003.

[4] G. Caire and A. Guillen i Fabregas. Analysis and design of natural and threaded space-time codes with iterative decoding. In *In Proc. 36th Asilomar Conf. on Signals, Systems & Computers*, Pacific Grove, CA, Nov. 2002.

[5] H. EL Gamal and R. Hammons. A new approach to layered space-time coding and signal processing. *IEEE Trans. Info. Theory*, 47(6):2321–2334, Sept. 2001.

[6] C. Berrou, A. Glavieux, and P. Thitimajshima. Near shannon limit error-correcting coding and decoding: Turbo-codes. In *In Proc. ICC 1993*, Geneva, Switzerland, May 1993.

[7] R. Knopp and P.A. Humblet. On coding for block fading channels. *IEEE Trans. Info. Theory*, 46(1):189 –205, Jan. 2000.

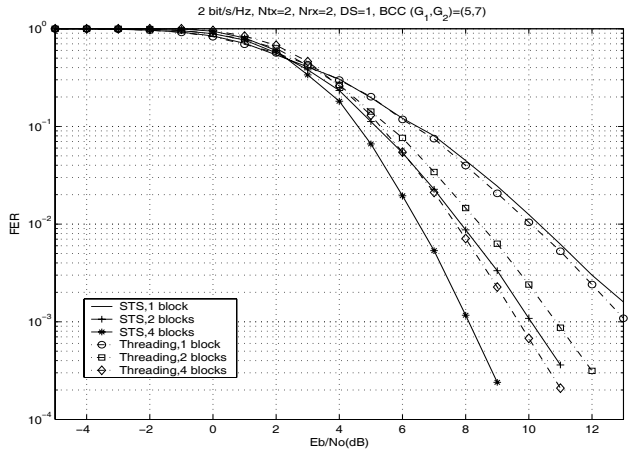

Fig. 4. STS/Threading for $(N_{tx}, N_{rx}) = (2, 2)$, $L = 1$, $F = 1, 2, 4$.

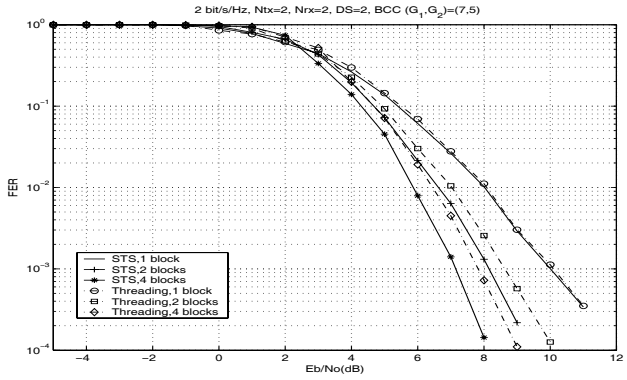

Fig. 5. STS/Threading for $(N_{tx}, N_{rx}) = (2, 2)$, $L = 2$, $F = 1, 2, 4$.

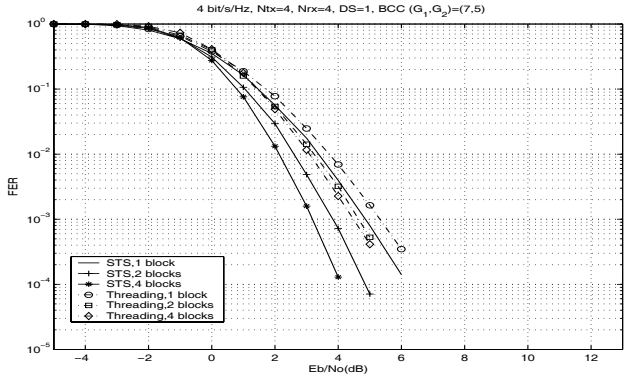

Fig. 6. STS/Threading for $(N_{tx}, N_{rx}) = (4, 4)$, $L = 1$, $F = 1, 2, 4$.

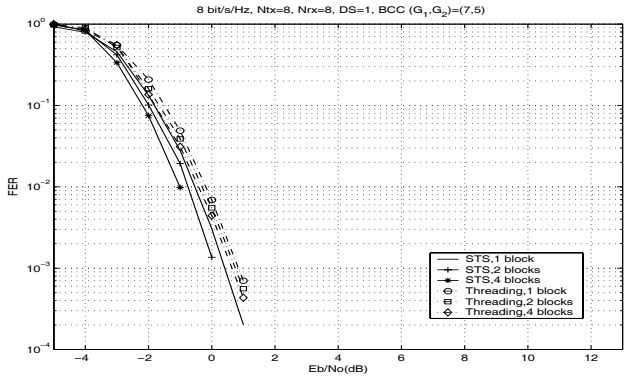

Fig. 7. STS/Threading for $(N_{tx}, N_{rx}) = (8, 8)$, $L = 1$, $F = 1, 2, 4$.

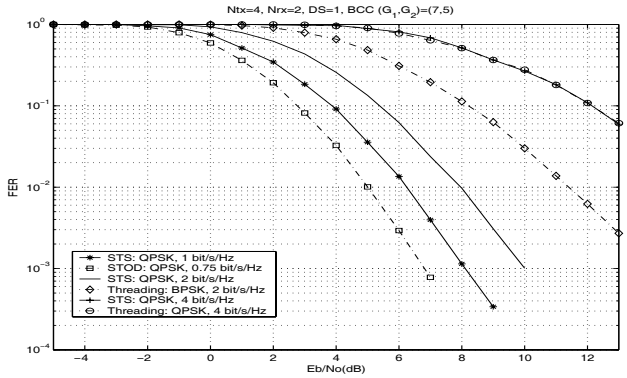

Fig. 8. STS ($N_s = 1, 2, 4$) vs Threading (QPSK, BPSK) and STOD for $(N_{tx}, N_{rx}) = (4, 2)$, $L = 1$ and $F = 1$.