A New Approach to Layered Space–Time Coding and Signal Processing

Hesham El Gamal, *Member, IEEE,* and A. Roger Hammons, Jr., *Member, IEEE*

***Abstract*—The information-theoretic capacity of multiple antenna systems was shown to be significantly higher than that of single antenna systems in Rayleigh-fading channels. In an attempt to realize this capacity, Foschini proposed the layered space–time architecture. This scheme was argued to asymptotically achieve a lower bound on the capacity. Another line of work has focused on the design of channel codes that exploit the spatial diversity provided by multiple transmit antennas [2], [3].**

In this paper, we take a fresh look at the problem of designing multiple-input–multiple-output (MIMO) wireless systems. First, we develop a generalized framework for the design of layered space–time systems. Then, we present a novel layered architecture that combines efficient algebraic code design with iterative signal processing techniques. This novel layered system is referred to as the threaded space–time (TST) architecture. The TST architecture provides more flexibility in the tradeoff between power efficiency, bandwidth efficiency, and receiver complexity. It also allows for exploiting the temporal diversity provided by time-varying fading channels. Simulation results are provided for the various techniques that demonstrate the superiority of the proposed TST architecture over both the diagonal layered space–time architecture in [1] and the recently proposed multilayering approach [4].

***Index Terms*—Array processing, fading channels, multiple transmit and receive antennas, multiuser detection, space–time coding.**

I. Introduction

RECENTLY, information-theoretic studies have shown that spatial diversity provided by multiple transmit and/or receive antennas allows for a significant increase in the capacity of coherent wireless communication systems operated in a flat Rayleigh-fading environment [5]–[7]. Following this discovery, two approaches for exploiting this spatial diversity have been proposed [2], [8], [3], [1]. In the first approach [2], [3], channel coding is performed across the spatial dimension as well as time to benefit from the spatial diversity provided by using multiple transmit antennas. Tarokh *et al.* coined the term "space–time coding" for this coding scheme. One potential drawback of this scheme is that the complexity of the maximum-likelihood (ML) decoder is exponential in the number of transmit antennas.

Manuscript received November 13, 1999; revised December 1, 2000. The material in this paper was presented in part at the Thirty-Seventh Annual Allerton Conference on Communication, Control, and Computing, University of Illinois at Urbana Champaign, Urbana, IL, September 22–24, 1999.

H. El Gamal was with the Advanced Development Group, Hughes Network Systems, Germantown, MD 20876 USA. He is now with the Electrical Engineering Department, Ohio State University, Columbus, OH 43210 USA (e-mail: helgamal@ee.eng.ohio-state.edu).

A. R. Hammons, Jr. was with the Advanced Development Group, Hughes Network Systems, Germantown, MD 20876 USA. He is now with Corvis Cooperation, Columbia, MD 21046 USA (e-mail: rhammons@corvis.com).

Communicated by M. L. Honig, Associate Editor for Communications.

Publisher Item Identifier S 0018-9448(01)06214-9.

The second approach, proposed by Foschini [1], relies upon suboptimal signal processing techniques at the receiver to achieve performance asymptotically close to the outage capacity with reasonable complexity. In this approach, no effort is made to optimize the channel coding scheme. The preferred approach in [1] is referred to as the Diagonal Bell Laboratories Layered Space–Time (D-BLAST) architecture. One of the contributions of this paper is a new scheme that combines ideas from these two approaches. Specifically, we present a new layered space–time transmission architecture—the threaded space–time (TST) architecture—that benefits from the advantages provided by efficient algebraic code design and advanced iterative signal processing [9]–[11].

Recently, Tarokh *et al.* proposed a new scheme for combined array processing and space–time coding [4] that likewise addresses some of the problems encountered with D-BLAST. This approach relies upon a zero-forcing group interference suppression technique and shows performance that is 6–9 dB from the outage capacity at 10% frame error rate [4]. The threaded architecture and signal processing proposed in this paper, however, close the gap to less than 3 dB from the outage capacity with the same frame length, error rate, and receiver complexity. It also provides greater flexibility in terms of the tradeoff between power efficiency, bandwidth efficiency, and receiver complexity.

The rest of this paper is organized as follows. The system description and a brief review of previous work on the design of space–time modems are presented in Section II. In Section III, we present a novel approach for the design of layered space–time systems. This approach combines iterative multiuser detection and decoding with algebraic space–time coding. Algebraic space–time code constructions for the new architecture are given in Section III-A1). In Section III-A2), the turbo processing principle is utilized to develop an iterative minimum mean-square error (MMSE) receiver. Comparisons of the various layered architectures in terms of efficiency and achievable diversity order are presented in Section IV, while simulation results are compared in Section V. Finally, Section VI presents our conclusions.

II. Overview of Space–Time Concepts

In this section, we lay out the basic concepts for space–time code design and signal processing. The key ideas involved in space–time coding for coherent channels [2], [8], [3], layered space–time processing [1], and a recently proposed hybrid multilayered approach [4] are briefly explained. This overview serves to establish our perspective and notation in the context of the prior body of work.

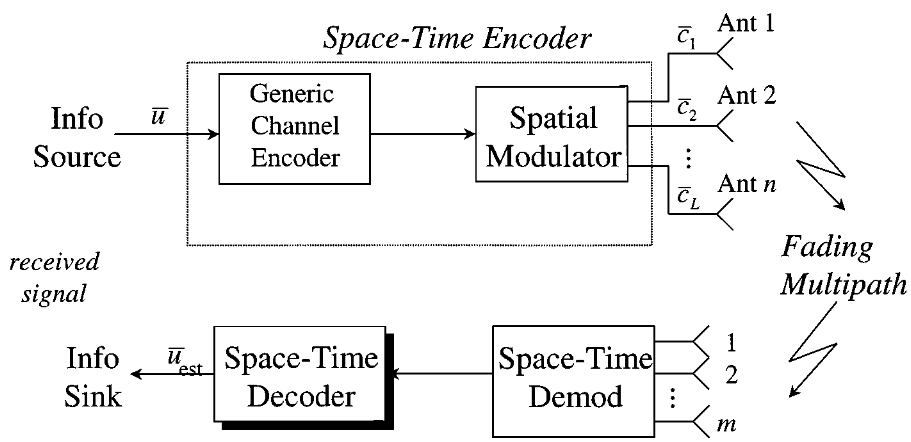

Fig. 1. Multiple antenna communication system.

A. Signal Model

We consider a multiple-antenna communication system with n transmit and m receive antennas as shown in Fig. 1. In this paper, we are interested in the scenario where the fading channel is frequency nonselective and channel state information is only available at the receiver [3], [2], [8]. In Fig. 1, the channel encoder accepts input from the information source and outputs a coded stream of higher redundancy suitable for error correction processing at the receiver. The encoded output stream is modulated and distributed among the n antennas. The transmissions from each of the n transmit antennas are simultaneous and synchronous. The signal received at each antenna is therefore a superposition of the n transmitted signals corrupted by additive white Gaussian noise (AWGN) and multiplicative fading. At the receiver end, the signal r_t^j received by antenna j at time t is given by

$$r_t^j = \sqrt{E_s} \sum_{i=1}^{n} \alpha_t^{(ij)} c_t^i + n_t^j \tag{1}$$

where E_s is the energy per transmitted symbol, $\alpha_t^{(ij)}$ is the complex path gain from transmit antenna i to receive antenna j at time t, c_t^i is the symbol transmitted from antenna i at time t, and n_t^j is the AWGN sample for receive antenna j at time t. The noise samples are independent samples of circularly symmetric zero-mean complex Gaussian random variables with variance $N_0/2$ per dimension. At each time t, the different path gains $\alpha_t^{(ij)}$ are assumed to be statistically independent. The fading model of primary interest is that of a block flat Rayleigh-fading process in which the codeword encompasses B fading blocks. The complex fading gains are constant over one fading block but are independent from block to block. The quasi-static fading model studied extensively in [8], [2], [3], [7] is a special case of the block fading model in which $B = 1$.

The received signal can be expressed in vector notation as

$$\underline{r}_t = S_t \underline{c}_t + \underline{n}_t \tag{2}$$

where $\underline{r}_t$ is the $m \times 1$ received vector at time t, S_t is the $m \times n$ complex channel matrix whose ith column corresponds to the path gains for the ith transmit antenna, $\underline{c}_t$ is the $n \times 1$ transmitted vector at time t, and $\underline{n}_t$ is the $m \times 1$ white Gaussian noise vector.

B. Space–Time Channel Codes

In the concept of a space–time code, the channel encoding, modulation, and distribution of symbols across antennas are intrinsically connected—i.e., a two-dimensional (2-D) coded modulation technique. Given a set $\mathcal{Y}$, the space of $1 \times d$ row vectors and the space of $d \times e$ matrices taking values in $\mathcal{Y}$ will be denoted by $\mathcal{Y}^d$ and $\mathcal{Y}^{d\times e}$, respectively. Then, a block code of length N over the discrete symbol alphabet $\mathcal{Y}$ is a subset C of the N-dimensional space $\mathcal{Y}^N$. Usually, the number of codewords in C is a power of the alphabet size, $|C| = |\mathcal{Y}|^k$, so that there is a one-to-one mapping, $\gamma: \mathcal{Y}^k \to C$, of information k-tuples onto codewords. The mapping γ is an encoder for C. In this paper, we will be primarily interested in the case in which C is a binary linear code—i.e., $\mathcal{Y}$ is the elementary binary field $\mathbb{F} = \mathrm{GF}(2)$ and C is linear.

The baseband modulation mapping $\mu: \mathcal{Y}^b \to \Omega$ assigns to each b-tuple of alphabet symbols a unique point in the discrete, complex-valued signaling constellation Ω, which is assumed not to contain the point zero. Conversely, the inverse map μ^{-1} provides a b-symbol labeling of the constellation points. By extension, $\mu(\underline{v})$ denotes the modulated version of the vector $\underline{v} \in \mathcal{Y}^N$. In this case, it is understood that N must be a multiple of b and that the blocking of symbols into b-tuples for the modulator is performed left to right.

Let $\Omega^* = \Omega \cup \{0\}$ denote the expanded constellation. Then, the spatial modulator is a mapping $\boldsymbol{f}: \mathcal{Y}^N \to (\Omega^*)^{n\times\ell}$ that sends the vector $\underline{v}$ to an $n \times \ell$ complex-valued matrix $\boldsymbol{c} = \boldsymbol{f}(\underline{v})$, whose nonzero entries are a rearrangement of the entries of $\mu(\underline{v})$. Specifically, $\boldsymbol{c}$ is the baseband version of the codeword $\underline{v}$ as transmitted across the channel. Thus, in the notation of (1), the matrix $\boldsymbol{c}$ has (i, t)th entry equal to c_t^i. Note that, in this formulation, it is expressly allowed that a complex zero (i.e., no transmission) be assigned to a given antenna at a given signaling interval; thus, $N/b \le n\ell$. This provision is intended to simplify the generalized layering framework outlined in Section III. We will refer to n and ℓ, respectively, as the spatial span and temporal span of $\boldsymbol{f}$.

Finally, for convenience, let $\hat{\boldsymbol{c}} = \mu^{-1}(\boldsymbol{c})$ denote the $n \times b\ell$ matrix in which each constellation point is replaced by its b-symbol label and any zero entry is replaced by a b-tuple of special blank symbols.

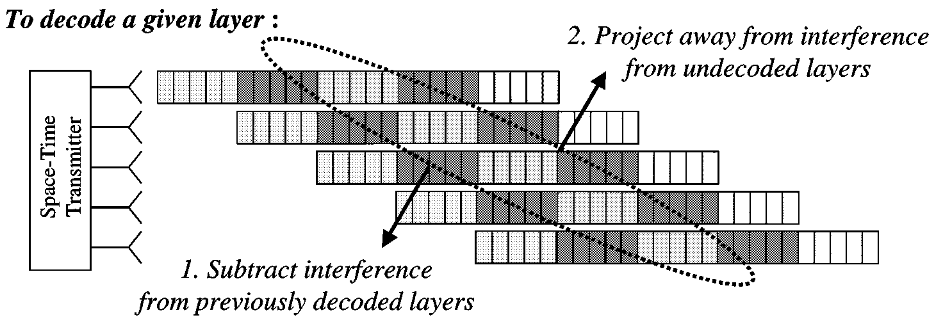

Fig. 2. Layering and signal processing for D-BLAST.

Definition 1: A space–time code $\mathcal{C}$ consists of an underlying channel code C together with the spatial modulator function $\boldsymbol{f}$.

The fundamental performance parameters [8], [2] for space–time codes are the following: 1) diversity advantage, which describes the exponential decrease of decoded error rate versus signal-to-noise ratio (SNR) (asymptotic slope of the performance curve on a log–log scale); and 2) coding advantage which does not affect the asymptotic slope but results in a shift in the performance curve. The diversity advantage is the more critical of the two performance metrics as it determines the asymptotic slope of the performance curve. Ideally, the coding advantage should be optimized after the diversity advantage is maximized [2], [8], [3].

For quasi-static fading channels, it has been shown [2], [8] that the spatial diversity advantage of the code, assuming ML decoding, is the product of the number of receive antennas and the minimum rank among the set of complex-valued matrices associated with differences between baseband-modulated codewords. It is clear that full spatial diversity nm will be achieved if and only if all the difference matrices have full rank.

In [3], we developed an algebraic framework for systematic design of binary phase-shift keying (BPSK) and quaternary phase-shift keying (QPSK) space–time codes that achieve full spatial diversity. This framework will be utilized in Section III-A1) to design algebraic space–time codes for the layered scenario.

C. Layered and Multilayered Space–Time Architectures

In the layered space–time architecture, the channel encoder of Fig. 1 is composite and the multiple, independent coded streams are distributed throughout the transmission resource array in so-called layers. The primary design objective is to design the layering architecture and associated signal processing so that the receiver can efficiently separate the individual layers from one another and can decode each of the layers effectively. Foschini [1] discusses different layering schemes for the proposed BLAST architecture. In the simplest variation, the code words are transmitted in horizontal layers (H-BLAST). The preferred scheme, however, involves the transmission of code words in diagonal layers (D-BLAST).

The BLAST receiver uses a multiuser detection strategy based on a combination of interference *cancellation* and *suppression*. In D-BLAST, each diagonal layer constitutes a complete codeword, so decoding is performed layer-by-layer. Consider the codeword matrix shown in Fig. 2, the entries below the first diagonal layer are zeros. To decode the first diagonal, the receiver generates a soft-decision statistic for each entry in that diagonal. In doing so, the interference from the upper diagonals is *suppressed* by projecting the received signal onto the null space of the *upper* interference. The soft statistics are then used by the corresponding channel decoder to decode this diagonal. The decoder output is then fed back to *cancel* the first diagonal contribution in the interference while decoding the next diagonal. The receiver then proceeds to decode the next diagonal in the same manner. It is worth noting that the zero-forcing suppression strategy requires that $m \geq n$; however, this requirement can be relaxed by using MMSE filtering instead of the zero-forcing strategy.

The multilayered space–time architecture, as introduced by Tarokh *et al.* [4], is a hybrid approach involving use of both space–time channel codes and layered processing. In this scheme, the input stream is divided, for example, into n/n' substreams. The different substreams are encoded using n'-level diversity component space–time trellis codes $\mathcal{C}_1, \ldots, \mathcal{C}_{n/n'}$. Each component code is then transmitted from n' antennas (horizontal n'-layering). At the receiver, each component code is decoded separately while suppressing signals from other component codes. The group interference suppression strategy [4] is based on the zero-forcing principle and requires that $m \geq n - n' + 1$. In quasi-static fading channel, the spatial diversity gain achieved by $\mathcal{C}_1$ is $n' \times (m - n + n')$. The decoded output from $\mathcal{C}_1$ is subtracted from signals at different receive antennas. This gives a communication system with $n - n'$ transmit and m receive antennas. Hence, assuming correct decoding of $\mathcal{C}_1$, the space–time code $\mathcal{C}_2$ affords a diversity gain of $n' \times (m - n + 2n')$, and so on. Using the fact that the diversity gain increases with each decoding stage, unequal power levels are allocated to the different component codes. Because all the space–time codes proposed in [2] were two-level diversity codes, except for the delay diversity, the design examples in [4] were limited to $n' = 2$.

III. Generalized Space–Time Layering

The different layering and multilayering approaches available in the literature were partly inspired by the signal processing techniques employed at the receiver. For example, in the D-BLAST approach, each layer is constrained to occupy a diagonal in the 2-D transmission resources array. It is easy to see that this constraint is imposed by the interference cancellation/suppression technique proposed in [1]. In this paper, we follow a different path. First, we generalize the notion of space–time layering independent of the signal processing employed at the receiver. Based on this generalized notion, we recognize a certain type of layering—TST layering—that efficiently exploits the diversity available in the system. Then, we consider the design of algebraic space–time codes and iterative signal processing techniques that optimize the performance of TST systems.

In our framework, a *layer* is defined as a section of the transmission resources array having the property that each symbol interval within the section is allocated to at most one antenna. This property ensures that all spatial interference experienced by the layer comes from outside the layer.

Formally, a layer in an $n \times \ell$ transmission resource array may be identified by an indexing set $L \subset I_n \times I_\ell$ having the property that the tth-symbol interval on antenna a belongs to the layer if and only if $(a, t) \in L$. Then, our formal notion of a layer requires that, if $(a, t) \in L$ and $(a', t') \in L$, then either $t \neq t'$ or $a = a'$ (i.e., that a is a function of t). The pair of spatial and temporal spans of a layer is defined as

$$([\max(a) - \min(a)], [\max(t) - \min(t)])$$

where $(a, t) \in L$. This pair represents the ability of the layer to exploit the available spatial and temporal diversity, and hence, it is desirable to develop a layering approach in which all layers have full spatial and temporal spans (n, ℓ).

Consider a composite channel encoder γ consisting of n constituent encoders $\gamma_1, \gamma_2, \ldots, \gamma_n$ operating on independent information streams. Let $\gamma_i\colon \mathcal{Y}^{k_i} \to \mathcal{Y}^{N_i}$, so that

$$k = k_1 + k_2 + \cdots + k_n$$

and

$$N = N_1 + N_2 + \cdots + N_n.$$

Then, there is a partitioning $\underline{u} = \underline{u}_1 \,|\, \underline{u}_2 \,|\, \cdots \,|\, \underline{u}_n$ of the composite information vector $\underline{u} \in \mathcal{Y}^k$ into a set of disjoint component vectors $\underline{u}_i$, of length k_i, and a corresponding partitioning

$$\gamma(\underline{u}) = \gamma_1(\underline{u}_1) \,|\, \gamma_2(\underline{u}_2) \,|\, \cdots \,|\, \gamma_n(\underline{u}_n)$$

of the composite codeword $\gamma(\underline{u})$ into a set of constituent codewords $\gamma_i(\underline{u}_i)$, of length N_i. In the generalized layering architecture approach, the space–time transmitter assigns each of the constituent codewords $\gamma_i(\underline{u}_i)$ to one of the set of n disjoint layers. For simplicity, we consider the case in which the constituent codes are all of the same rate and have the same codeword length: $N_i = N/n$ and $k_i = k/n$ for all i.

There is a corresponding decomposition of the spatial modulating function that is induced by the layering. Let $\boldsymbol{f}_i\colon \mathcal{Y}^{N/n} \to (\Omega^*)^{n \times \ell}$ denote the *component spatial modulating function*, associated with layer L_i, which agrees with the composite spatial modulator $\boldsymbol{f}$ regarding the modulation and formatting of the layer elements but which sets all off-layer elements to complex zero. Then

$$\boldsymbol{f}(\gamma(\underline{u})) = \boldsymbol{f}_1(\gamma_1(\underline{u}_1)) + \boldsymbol{f}_2(\gamma_2(\underline{u}_2) + \cdots + \boldsymbol{f}_n(\gamma_n(\underline{u}_n)).$$

It is straightforward to see that the layered architectures proposed by Foschini in [1] are special cases of this generalized layering. For example, in the D-BLAST architecture [1], the output of each encoder is distributed among the n antennas along the diagonal layers such that

$$L_i = \{(\lfloor t/w \rfloor_n - i + 2, t)\colon (i-1)w \le t < \ell - (n-i)w\} \tag{3}$$

where $w = N/(n^2 b)$ is the width of the diagonal, $\ell = (2n-1)w$ is the temporal span, and $\lfloor \cdot \rfloor_n$ denotes the function returning the integer part of a real-valued input reduced modulo n.

A. TST Layering

In this section, we present a new space–time layering design that efficiently exploits the diversity available in the multiple-input–multiple-output (MIMO) channel. In the proposed approach, the encoding, interleaving, and distribution of each layer's symbols among different antennas are optimized to maximize spatial and temporal diversity for a given transmission rate, assuming no interference from the other layers. Meanwhile, interleaving is also optimized to maximize the efficiency of the iterative signal processing techniques necessary to suppress other layers' interference as described in Section III-A2). It is worth noting that the threaded approach is applicable to arbitrary constellations with binary (or nonbinary) codes.

As in the generalized layering architecture, the transmitter has available a disjoint set of layers $\mathcal{L} = \{L_1, L_2, \ldots, L_n\}$ and transmits the composite codeword

$$\gamma(\underline{u}) = \gamma_1(\underline{u}_1) \,|\, \gamma_2(\underline{u}_2) \,|\, \cdots \,|\, \gamma_n(\underline{u}_n)$$

by sending $\gamma_i(\underline{u}_i)$ in layer L_i.

The layer set $\mathcal{L}$ is designed so that each layer is active during all of the available symbol transmission intervals and, over time, uses each of the n antennas equally often. Thus, during each symbol transmission interval, each layer transmits a symbol using a different antenna; and, in terms of antenna usage, all of the layers are equivalent. Unlike the layered architectures of [1], the new design approach treats the coded transmission in each layer as a *bona fide* space–time code, constructions for which are given in the next section. Looking at the space–time coding performed on a single layer in isolation, one notes that the main limitation of this construction is the reduction in throughput resulting from the silence periods imposed on the different antennas. But, in the overall transmission scheme, the silent periods on antennas not used by a given layer are filled with the transmissions from the other component space–time codes. Iterative signal processing at the receiver, necessary to remove or suppress spatial interference among the layers, is discussed in Section III-A2). One innovation of the new architecture is that, under the assumption of error-free interference cancellation, the component space–time codes can be designed to achieve full spatial diversity without degradation in overall system throughput.

The new space–time architecture is not a multilayered approach since the transmit positions occupied by the modulated code symbols for a particular codeword constitute a single layer. Yet, the new architecture is not a layered architecture in the same sense as the BLAST architecture, since the layering is more general, well-suited for iterative multiuser techniques, and the channel coding design in the new approach is 2-D based on

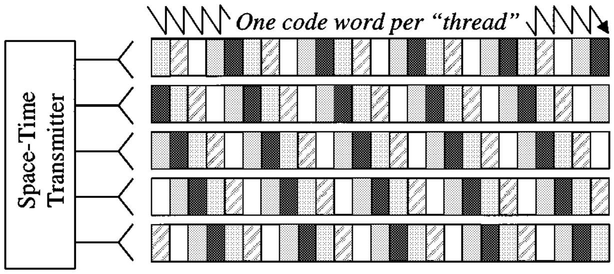

Fig. 3. A simple example for threaded layering (each shade represents a thread).

space–time coding principles designed to exploit both the spatial and temporal diversity. To distinguish this new approach, we refer to it as the *threaded space-time* (TST) architecture and each layer in the new architecture is referred to as a *thread*. A thread can be defined as a layer with full spatial span "n" and full temporal span "ℓ." The simplest example of threaded layering is the set $\mathcal{L}$ shown in Fig. 3 in which

$$L_i = \{(\lfloor t+i-1 \rfloor_n + 1, t): 0 \le t < \ell\}. \tag{4}$$

1) Design of TST Codes: Now, we look at the design of the component space–time codes used in the threaded architecture. The design of these codes follows the algebraic approach introduced in [3]. The layering provided by the threaded architecture allows the algebraic formulation to be extended to arbitrary signaling constellations. Importantly, the requirement for independent interleaving in the iterative multiuser receiver, discussed in the following section, is easily accommodated in these code designs. Our results are first developed for quasi-static fading channels, then we outline the extension to time-varying block fading channels.

Consider a single threaded layer L_i and the corresponding component space–time code $\mathcal{C}_i$ associated with encoder γ_i. The spatially modulated codewords of $\mathcal{C}_i$ are the $n \times \ell$ complex matrices $\boldsymbol{f}_i(\gamma_i(\underline{u}_i))$. To simplify notation, we will drop the indexes, letting $L = L_i$, $\mathcal{C} = \mathcal{C}_i$, and $g = \gamma_i$. We will let $\boldsymbol{f}_L$ denote the component spatial modulator function associated with layer L. Unsubscripted vectors such as $\underline{x}$ or $\underline{y}$ will be used to refer to the information stream.

For the design of the space–time code $\mathcal{C}$ associated with thread L, we have the following stacking construction using binary matrices for the quasi-static fading channel.

Theorem 2 (Threaded Stacking Construction): Let L be a threaded layer of spatial span n. Given binary matrices $\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_n$ of dimension $k \times b\ell/n$, let C be the binary code of dimension k consisting of all codewords of the form $g(\underline{x}) = \underline{x}\boldsymbol{M}_1 \,|\, \underline{x}\boldsymbol{M}_2 \,|\, \cdots \,|\, \underline{x}\boldsymbol{M}_n$, where $\underline{x}$ denotes an arbitrary k-tuple of information bits. Let $\boldsymbol{f}_L$ denote the spatial modulator having the property that $\mu(\underline{x}\boldsymbol{M}_i)$ is transmitted in the ℓ/n symbol intervals of L that are assigned to antenna i.

Then, as the space–time code in a communication system with n transmit antennas and m receive antennas, the space–time code $\mathcal{C}$ consisting of C and $\boldsymbol{f}_L$ achieves spatial diversity dm in a quasi-static fading channel if and only if d is the largest integer such that $\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_n$ have the property that

$$\forall\, a_1, a_2, \ldots, a_n \in \mathbb{F},\ a_1 + a_2 + \cdots + a_n = n - d + 1:$$
$$\boldsymbol{M} = [a_1\boldsymbol{M}_1 a_2\boldsymbol{M}_2 \cdots a_n\boldsymbol{M}_n] \text{ is of rank } k \text{ over the binary field.}$$

Proof: Due to the lack of spatial interference within a layer, the baseband rank criterion [2], [8] is straightforward to apply. In particular, note that the baseband difference $\boldsymbol{f}_L(g(\underline{x})) - \boldsymbol{f}_L(g(\underline{y}))$ has rank d if and only if it has precisely d nonzero rows.

Now suppose that, for some $a_1, a_2, \ldots, a_n \in \mathbb{F}$ satisfying

$$a_1 + a_2 + \cdots + a_n = n - d + 1$$

we have that

$$\boldsymbol{M} = [a_1\boldsymbol{M}_1 a_2\boldsymbol{M}_2 \cdots a_n\boldsymbol{M}_n]$$

is singular. Then, there exist $\underline{x}, \underline{y} \in \mathbb{F}^k$, $\underline{x} \neq \underline{y}$, such that $\underline{x}\boldsymbol{M} = \underline{y}\boldsymbol{M}$. In this case, $\boldsymbol{f}_L(g(\underline{x})) - \boldsymbol{f}_L(g(\underline{y}))$ has an all-zero row for every nonzero coefficient a_i. Since there are $n-d+1$ nonzero coefficients, $\boldsymbol{f}_L(g(\underline{x})) - \boldsymbol{f}_L(g(\underline{y}))$ has rank less than d. Thus, $\mathcal{C}$ does not achieve dm-level diversity.

Conversely, suppose $\mathcal{C}$ does not achieve dm-level diversity. Then, there exist $\underline{x}, \underline{y} \in \mathbb{F}^k$, $\underline{x} \neq \underline{y}$, such that the baseband difference $\boldsymbol{f}_L(g(\underline{x})) - \boldsymbol{f}_L(g(\underline{y}))$ has rank less than d. It must, therefore, have at least $n-d+1$ all-zero rows. Let I denote a set of indices for $n-d+1$ such rows, and set $a_i = 1$ for $i \in I$ and $a_i = 0$ otherwise. Then, the matrix $\boldsymbol{M} = [a_1\boldsymbol{M}_1 a_2\boldsymbol{M}_2 \cdots a_n\boldsymbol{M}_n]$ is singular since $\underline{x}\boldsymbol{M} = \underline{y}\boldsymbol{M}$. □

Corollary 3: Full spatial diversity nm is achieved if and only if $\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_n$ are of rank k over the binary field.

A space–time code that achieves dm-level spatial diversity in a communication system with n transmit and m receive antennas over the quasi-static fading channel is called a d-space–time code.

Corollary 4: The maximum transmission rate for a communication system using the threaded layering architecture with n transmit antennas, a signaling constellation of size 2^b, and component codes achieving d-level transmit spatial diversity is $b(n-d+1)$ bits per second per hertz.

Proof: By Theorem 2, in order for the code to achieve d-level spatial diversity, the number of columns in $\boldsymbol{M}_j$ must satisfy $b\ell/n \ge k/(n-d+1)$. Then the code rate for C is $k/(b\ell) \le (n-d+1)/n$. Therefore, the maximum transmission rate of each thread is $br \le b(n-d+1)/n$ bits per signaling interval. Then, the total transmission rate of the n threads is $b(n-d+1)$. A different proof can be obtained using the argument in [12] on the maximum lossless compression transmission rate. □

The following result is straightforward but quite important for the design of space-time threaded codes that allow for maximizing the efficiency of the iterative multiuser detector as discussed in the next section.

Theorem 5: Let $\mathcal{C}$ be a d-space–time code consisting of the binary code C whose codewords are of the form

$$g(\underline{x}) = \underline{x}\boldsymbol{M}_1 \,|\, \underline{x}\boldsymbol{M}_2 \,|\, \cdots \,|\, \underline{x}\boldsymbol{M}_n$$

where $\underline{x}$ denotes an arbitrary k-tuple of information bits, and the spatial modulator $\boldsymbol{f}_L$ in which $\underline{x}\boldsymbol{M}_i$ is assigned to antenna i along threaded layer L. Given the linear vector-space transformations $T_1, \ldots, T_n$: $\mathbb{F}^{b\ell/n} \to \mathbb{F}^p$, we construct a new space–time code by assigning $T_i(\underline{x}\boldsymbol{M}_i)$ to antenna i along threaded layer L. Then, the new space–time code achieves the same spatial diversity order dm if $T_1, \ldots, T_n$ are nonsingular.

In particular, we can take the linear transformation T_i of the previous theorem to be an arbitrary permutation π_i. Then, the interleaved space–time code resulting from assigning $\pi_i(\underline{x}\boldsymbol{M}_i)$ to antenna i along threaded layer L achieves the same level of spatial diversity as the noninterleaved space–time code $\mathcal{C}$.

We now look at the special case of designing space–time trellis codes for the threaded architecture. The main advantage of such codes is the availability of computationally efficient, soft-input/soft-output (SISO) decoding algorithms. The natural space–time codes [3] associated with binary, rate $1/n$, convolutional codes with periodic bit interleaving are attractive candidates for the threaded space–time architecture as they can be easily formatted to satisfy the threaded stacking construction. Each output arm from the encoder is transmitted from a separate antenna. There is no restriction on the interleaving employed by each antenna (i.e., different interleaving can be used by the different antennas without violating the threaded stacking condition). As discussed earlier, this feature allows for the design of efficient iterative multiuser receivers. These convolutional codes were considered for a similar application (the block-erasure channel) in [12].

The prior literature on space–time trellis codes treats only the case in which the underlying code has rate $1/n$ matched to the number of transmit antennas. In our development of threaded space–time code design, we consider the more general case in which the convolutional code has rate greater than $1/n$. The treatment includes the case of rate k/n convolutional codes constructed by puncturing an underlying rate $1/n$ convolutional code.

Let C be a binary convolutional code of rate k/n. The encoder processes k binary input sequences $x_1(t), x_2(t), \ldots, x_k(t)$ and produces n coded output sequences $y_1(t), y_2(t), \ldots, y_n(t)$ which are multiplexed together to form the output codeword.

For quasi-static fading channels, the input and output sequences of interest are of fixed finite length; in the more general case, however, the sequences are semi-infinite indexed by $t = 0, 1, 2, \ldots$. We let $\mathbb{F}^\infty$ denote the space of all such binary sequences. A sequence $\{x(t)\}_{t=0}^{\infty} \in \mathbb{F}^\infty$ is often represented by the formal series

$$X(D) = x(0) + x(1)D + x(2)D^2 + \cdots.$$

We refer to $\{x(t)\} \leftrightarrow X(D)$ as a D-transform pair. The space $\mathbb{F}[[D]]$ of all formal series is an integral domain whose invertible elements are those that are not multiples of D.

The action of the binary convolutional encoder is linear and is characterized by the so-called impulse responses $g_{i,j}(t) \leftrightarrow G_{i,j}(D)$ associating output $y_j(t)$ with input $x_i(t)$. Thus, the encoder action is summarized by the matrix equation

$$\boldsymbol{Y}(D) = \boldsymbol{X}(D)\boldsymbol{G}(D)$$

where

$$\begin{aligned}\boldsymbol{Y}(D) &= [\,Y_1(D) \quad Y_2(D) \quad \cdots \quad Y_n(D)\,]\\ \boldsymbol{X}(D) &= [\,X_1(D) \quad X_2(D) \quad \cdots \quad X_k(D)\,]\end{aligned}$$

and

$$\boldsymbol{G}(D) = \begin{bmatrix} G_{1,1}(D) & G_{1,2}(D) & \cdots & G_{1,n}(D) \\ G_{2,1}(D) & G_{2,2}(D) & \cdots & G_{2,n}(D) \\ \vdots & \vdots & \ddots & \vdots \\ G_{k,1}(D) & G_{k,2}(D) & \cdots & G_{k,n}(D) \end{bmatrix}.$$

We consider the natural space–time formatting of C in which the output sequence corresponding to $Y_j(D)$ is assigned to the jth transmit antenna and wish to characterize the spatial diversity that can be achieved by this scheme. Our algebraic analysis technique considers the rank of matrices formed by concatenating the column vectors

$$\boldsymbol{F}_\ell(D) = \begin{bmatrix} G_{1,\ell}(D) \\ G_{2,\ell}(D) \\ \vdots \\ G_{k,\ell}(D) \end{bmatrix}.$$

Specifically, for $a_1, a_2, \ldots, a_n \in \mathbb{F}$, let

$$\boldsymbol{F}(a_1, a_2, \ldots, a_n) = [\,a_1\boldsymbol{F}_1 \quad a_2\boldsymbol{F}_2 \quad \cdots \quad a_n\boldsymbol{F}_n\,].$$

Then we have the following theorem relating the spatial diversity of the space–time code $\mathcal{C}$ in the quasi-static fading channel to the rank of these matrices over $\mathbb{F}[[D]]$.

Theorem 6: Let $\mathcal{C}$ denote the threaded space–time code consisting of the binary convolutional code C, whose $k \times n$ transfer function matrix is

$$\boldsymbol{G}(D) = [\boldsymbol{F}_1(D) \;\; \boldsymbol{F}_2(D) \;\; \cdots \;\; \boldsymbol{F}_n(D)]$$

and the spatial modulator $\boldsymbol{f}_L$ in which the output

$$Y_j(D) = \boldsymbol{X}(D) \cdot \boldsymbol{F}_j(D)$$

is assigned to antenna j along threaded layer L. Let v be the smallest integer having the property that, whenever $a_1 + a_2 + \cdots + a_n = v$, the $k \times n$ matrix $\boldsymbol{F}(a_1, a_2, \ldots, a_n)$ has full rank k over $\mathbb{F}[[D]]$. Then the space–time code $\mathcal{C}$ achieves d-level spatial transmit diversity over the quasi-static fading channel where $d = n - v + 1$ and $v \geq k$.

Proof: All of the codewords of $\mathcal{C}$ are of the form $\boldsymbol{Y}^{\mathrm{T}}(D) = \boldsymbol{G}^{\mathrm{T}}(D)\boldsymbol{X}^{\mathrm{T}}(D)$. Under the stipulated conditions of the theorem and following the argument of Theorem 2 (threaded stacking construction), only the all-zero codeword has v or more all-zero rows, so the spatial transmit diversity of $\mathcal{C}$ is at least $n - v + 1$. On the other hand, since v is the smallest integer having the stated property, there is some information sequence $\boldsymbol{X}(D)$ resulting in a codeword with $v - 1$ all-zero rows. Hence, the spatial transmit diversity of $\mathcal{C}$ is precisely $n-v+1$. □

Rate $1/n'$ convolutional codes with $n' < n$ can also be put into this framework. Let C be a binary convolutional code with transfer function matrix

$$\boldsymbol{G}(D) = [\, G_1(D) \quad G_2(D) \quad \cdots \quad G_{n'}(D)\,].$$

The coded bits are to be distributed among n transmit antennas. For simplicity, we consider the case in which $s = n/n'$ is an integer and the coded bits are assigned to the antennas periodically. Thus, for each of the coded bit streams $Y_i(D) \leftrightarrow \{y_i(t)\}$, the subsequence $y_i(0), y_i(s), y_i(2s), \ldots$ is assigned to antenna i; the subsequence $y_i(1), y_i(s+1), y_i(2s+1), \ldots$ is assigned to antenna $i+n'$; and so on. Alternate assignments such as symbol-based demultiplexing would also be possible and can be analyzed using the same framework.

In general, we partition the series $X(D)$ corresponding to $\{x(t)\}$ into its modulo s components $X_j(D)$ corresponding to the subsequences

$$\{x(st+j)\}_{t=0}^{\infty} \qquad (j = 0, 1, 2, \ldots, s-1).$$

Then

$$X(D) = X_0(D^s) + D \cdot X_1(D^s) + \cdots + D^{s-1} \cdot X_{s-1}(D^s).$$

Similarly, we partition $G_i(D)$ into components $G_{i,j}(D)$ and $Y_i(D)$ into components $Y_{i,j}(D)$. The space–time code $\mathcal{C}$ under consideration therefore consists of the binary code C together with a spatial modulator function in which $Y_{i,j}(D)$ is assigned to antenna $n'j+i$.

By multiplying the expansions for $X(D)$ and $G_i(D)$ and collecting terms, one may show that the coded bit stream assigned to antenna $n'j+i$ is given by

$$Y_{i,j}(D) = \sum_{k=0}^{s-1} X_k(D) F_{n'j+i,k}(D)$$

where

$$F_{n'j+i,k}(D) = G_{i,j-k}(D) + D \cdot G_{i,j-k+s}(D).$$

In matrix form, we have

$$Y_{i,j}(D) = \boldsymbol{X}(D)\boldsymbol{F}_{n'j+i}(D)$$

which is the dot product of row vector

$$\boldsymbol{X}(D) = [\, X_0(D) \quad X_1(D) \quad \cdots \quad X_{s-1}(D)\,]$$

and column vector

$$\boldsymbol{F}_{n'j+i}(D) = \begin{bmatrix} F_{n'j+i,0}(D) \\ F_{n'j+i,1}(D) \\ \vdots \\ F_{n'j+i,s-1}(D) \end{bmatrix}.$$

The theorem now applies directly. The spatial transmit diversity achieved by $\mathcal{C}$ is given by $d = n - v + 1$, where v is the smallest integer having the property that, whenever $a_0 + a_1 + \cdots + a_{n-1} = v$, the $s \times n$ matrix $\boldsymbol{F}(a_0, a_1, \ldots, a_{n-1})$ has full rank s. In particular, we note that the best possible spatial transmit diversity is $d = n - s + 1$. When $n' = n$, we have $s = 1$ so that full spatial transmit diversity $d = n$ is possible as expected.

Example: Consider the four-state convolutional code with optimal $d_{\text{free}} = 5$ and generators $G_0(D) = 1 + D^2$ and $G_1(D) = 1 + D + D^2$. In the case of two transmit antennas, it is clear that the natural threaded space–time code achieves $d = 2$ level diversity.

In the case of four transmit antennas, we note that the rate–$1/2$ code can be written as a rate–$2/4$ convolutional code with generator matrix

$$\boldsymbol{G}(D) = \begin{bmatrix} 1+D & 0 & 1+D & 1 \\ 0 & 1+D & D & 1+D \end{bmatrix}.$$

By inspection, every pair of columns is linearly independent over $\mathbb{F}[[D]]$. Hence, the natural periodic distribution of the code across four transmit antennas produces a threaded space–time code achieving the maximum $d = 3$ transmit spatial diversity.

For six transmit antennas, we express the code as a rate–$3/6$ code with generator matrix

$$\boldsymbol{G}(D) = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 1 \\ D & 1 & 0 & D & 1 & 1 \\ 0 & D & 1 & D & D & 1 \end{bmatrix}.$$

Every set of three columns in the generator matrix has full rank over $\mathbb{F}[[D]]$, so the natural space-time code achieves maximum $d = 4$ transmit diversity.

Thus far, we have considered the design of threaded space–time codes that exploit the spatial diversity over quasi-static fading channels. However, one of the advantages of the threaded architecture is its ability to jointly exploit the spatial diversity provided by the multiple transmit and receive antennas, and the temporal diversity provided by the time variations in the block fading channel. In fact, the results obtained for threaded space–time code design for the quasi-static fading channel can be easily extended to the more general block fading channel.

In the absence of interference from other threads, the quasi-static fading channel under consideration may be viewed as a block fading channel with receive diversity, where each fading block is represented by a different antenna. For the threaded architecture with n transmit antennas and a quasi-static fading channel, there are n independent and noninterfering fading links per codeword that can be exploited for transmit diversity by proper code design. In the case of the block fading channel, there is a total of nB such links, where B is the number of independent fading blocks per codeword per antenna. Thus, the problem of block fading code design for the threaded architecture is addressed by simply replacing the parameter n by nB.

For example, the following "multistacking construction" is a direct generalization of Theorem 2 to the case of a block fading channel.

Theorem 7 (Threaded Multistacking Construction): Let L be a threaded layer of spatial span n. Given binary matrices

$$\boldsymbol{M}_{1,1}, \boldsymbol{M}_{2,1}, \ldots, \boldsymbol{M}_{n,1}, \ldots, \boldsymbol{M}_{1,B}, \boldsymbol{M}_{2,B}, \ldots, \boldsymbol{M}_{n,B}$$

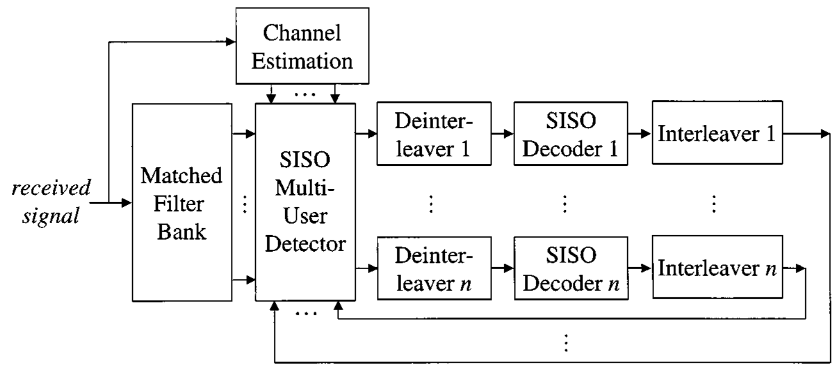

Fig. 4. Iterative multiuser detector for space–time signals.

of dimension $k \times b\ell/nB$, let C be the binary code of dimension k consisting of all codewords of the form

$$g(\underline{x}) = \underline{x}\boldsymbol{M}_{1,1} \,|\, \underline{x}\boldsymbol{M}_{2,1} \,|\, \cdots \,|\, \underline{x}\boldsymbol{M}_{n,1} \,|\, \cdots \,|\, \underline{x}\boldsymbol{M}_{1,B} \,|\, \underline{x}\boldsymbol{M}_{2,B} \,|\, \cdots \,|\, \underline{x}\boldsymbol{M}_{n,B}$$

where $\underline{x}$ denotes an arbitrary k-tuple of information bits, and B is the number of independent fading blocks spanning one codeword. Let $\boldsymbol{f}_L$ denote the spatial modulator having the property that $\mu(\underline{x}\boldsymbol{M}_{j,\nu})$ is transmitted in the symbol intervals of L that are assigned to antenna j in the fading block ν.

Then, as the space–time code in a communication system with n transmit antennas and m receive antennas, the space–time code $\mathcal{C}$ consisting of C and $\boldsymbol{f}_L$ achieves diversity dm in a B-block fading channel if and only if d is the largest integer such that $\boldsymbol{M}_{1,1}, \boldsymbol{M}_{2,1}, \ldots, \boldsymbol{M}_{n,B}$ have the property that

$$\forall a_{1,1}, a_{2,1}, \ldots, a_{n,B} \in \mathbb{F},\ a_{1,1} + a_{2,1} + \cdots + a_{n,B} = nB - d + 1:$$
$$\boldsymbol{M} = [a_{1,1}\boldsymbol{M}_{1,1} a_{2,1}\boldsymbol{M}_{2,1} \cdots a_{n,B}\boldsymbol{M}_{n,B}]$$

is of rank k over the binary field.

Proof: This result is immediate from the equivalent quasi-static model with nB transmit antennas. □

2) Iterative Signal Processing for TST Layering: In the previous section, we have considered the problem of designing TST systems assuming that a genie were to cancel the other layers' interference at the receiver. Ultimately, the performance of threaded systems will hinge upon the efficiency of the signal processing at the receiver in separating the signals from different threads. The problem of space–time signal processing can be formulated as a joint multiuser detection and decoding problem. Hence, the turbo processing principle [13] can be efficiently used to develop a set of iterative multiuser detection algorithms that allow tradeoffs between performance and complexity. A block diagram of the iterative receiver is shown in Fig. 4. In this block diagram, a SISO multiuser detector module provides soft-decision estimates of the n streams of data. The detected streams are decoded by the separate SISO channel decoders associated with the component channel codes. After each decoding iteration, the soft outputs from the channel decoders are used to refine the processing performed by the SISO multiuser detector. Note that, in the iterative receiver, each of the streams is independently interleaved to facilitate convergence. This key feature of the receiver is instrumental in ensuring good convergence characteristics for the iterative algorithm [10], [11]. We explicitly allowed for this random interleaving option in our algebraic code constructions in the previous section.

The complexity of the SISO multiuser detector constitutes a major part of the overall complexity of the iterative receiver. Three SISO multiuser detection algorithms that provide a tradeoff between performance and complexity have been proposed in the literature. The first is based on the maximum *a posteriori* (MAP) probability rule [9], [14]; the second is based on the MMSE criterion [10], [11]; and the third is the soft interference canceler which can be viewed a suboptimal approximation of the iterative MMSE receiver [15]. In this paper, we will focus our attention on the iterative MMSE receiver because it provides an efficient tradeoff between performance and complexity among the three iterative approaches [11], [10].

The iterative MMSE receiver is adapted from the first author's work [10] on iterative MMSE multiuser detectors for code-division multiple access (CDMA) systems. (Such receivers for CDMA applications were also investigated independently by Wang and Poor [11].) For simplicity of presentation, binary channel codes and BPSK modulation are assumed.

In this scheme, the soft outputs are used after each iteration to update the *a priori* probabilities of the transmitted symbols. These updated probabilities are then used to calculate the conditional MMSE filter feed-forward and feedback weights. The feedback connection represents the subtractive interference cancellation part of the receiver, while the feed-forward weights serve to suppress any residual interference. Let $y^{(i)}$ be the estimate of the ith-antenna symbol at time t given by (the subscript t will be omitted for convenience)

$$y^{(i)} = \underline{w}_f^{(i)^T} \underline{r} + w_b^{(i)} \tag{5}$$

where $\underline{w}_f^{(i)}$ is the $m \times 1$ optimized feed-forward coefficients vector and $w_b^{(i)}$ is a single coefficient that represents the soft cancellation part. The coefficients $\underline{w}_f^{(i)}$, $w_b^{(i)}$ are obtained through minimizing the conditional mean-square value of the error between the data symbol and its estimate. Now, let $\underline{S}^{(i)}$ be the $m \times 1$ complex channel vector of the ith transmit antenna; $S^{(n\backslash i)}$ be the $m \times (n-1)$ matrix composed of the complex channel vectors of the other $n-1$ transmit antennas; and $\underline{c}^{(n\backslash i)}$

be the $(n-1)\times 1$ transmitted data vector form the other $n-1$ transmit antennas. Assuming statistically independent *a priori* information and using standard minimization techniques, it is easily shown that the conditional MMSE solutions for $\underline{w}_f^{(i)}$, and $w_b^{(i)}$ are given by [10]

$$\underline{w}_f^{(i)^T} = \underline{S}^{(i)^H}\left(A+B+R_n-FF^H\right)^{-1} \quad (6)$$

$$w_b^{(i)} = -\underline{w}_f^{(i)^T}F, \quad (7)$$

where

$$A = \underline{S}^{(i)}\underline{S}^{(i)^H} \quad (8)$$

$$B = S^{(n\backslash i)}\Big[I_{(n-1)\times(n-1)} - \mathrm{Diag}\left(\underline{\tilde{c}}^{(n\backslash i)}\underline{\tilde{c}}^{(n\backslash i)^T}\right) + \underline{\tilde{c}}^{(n\backslash i)}\underline{\tilde{c}}^{(n\backslash i)^T}\Big]S^{(n\backslash i)^H} \quad (9)$$

$$F = S^{(n\backslash i)}\underline{\tilde{c}}^{(n\backslash i)} \quad (10)$$

$$R_n = N_0 I_{m\times m} \quad (11)$$

$I_{m\times m}$ is the identity matrix of order m, and $\underline{\tilde{c}}^{(n\backslash i)}$ is the $(n-1)\times 1$ vector of the conditional expected values of the transmitted symbols from the other $n-1$ antennas. The *a priori* probabilities used to evaluate these expected values are obtained from the previous decoding iteration soft outputs through the component-wise relation

$$P\left(c_t^j = 1\right) = \frac{e^{\lambda_t^j}}{1+e^{\lambda_t^j}} \quad (12)$$

where λ_t^j is the extrinsic information corresponding to the symbol transmitted from the jth antenna at time t [16]. Note that in the first iteration, one takes

$$P(c_t^j = 1) = P(c_t^j = -1) = 1/2.$$

3) Performance Bound: In this section, we investigate the spatial diversity advantage achieved by the threaded architecture over the quasi-static fading channel when the iterative MMSE algorithm is used.

Proposition 8: Let $\mathcal{C}$ be a d-diversity code used in each thread in a setting with n transmit and m receive antennas in quasi-static fading channels, then the zero-forcing receiver achieves spatial diversity $d' = d\cdot(m-n+1)$.

Proof: To detect the signal transmitted from the ith antenna, the zero-forcing receiver projects the received signal on the null space of $S^{(n\backslash i)}$. Let $\mathcal{N}_i$ be the null space of $S^{(n\backslash i)}$, and $\mathcal{V}_i$ be an $(m-n+1)\times m$ matrix whose rows are orthonormal vectors of $\mathcal{N}_i$. Then the $(m-n+1)\times 1$ output vector corresponding to $c_t^{(i)}$ is computed as

$$\underline{y}_t^{(i)} = \mathcal{V}_i\underline{r}_t = \underline{\tilde{S}}^{(i)}c_t^{(i)} + \underline{\tilde{n}}_t^{(i)}. \quad (13)$$

The elements of $\underline{\tilde{S}}^{(i)}$, $\underline{\tilde{n}}_t^{(i)}$ are Gaussian random variables with $E\{\underline{\tilde{n}}_t^{(i)}\underline{\tilde{n}}_t^{(k)H}\} = 0$. Note that, in general, $E\{\underline{\tilde{S}}^{(i)}\underline{\tilde{S}}^{(k)H}\} \neq 0$. Hence, at the output of the zero-forcing filter, the channel is equivalent to an interference-free correlated block fading channel with n blocks and $m-n+1$ receive antennas. Since the different equivalent Gaussian fading gains are linearly independent, the equivalent channel correlation matrix is of full rank [17]. Thus, by the argument in [2], the diversity order is $d\cdot(m-n+1)$. □

Let $\mathrm{SIR}_j^{(i)}$ denote the signal-to-interference-plus-noise ratio (SIR) for a symbol transmitted from the ith antenna after the jth iteration of the iterative MMSE algorithm. Then, conditioning on the set of path gains, we have

$$\mathrm{SIR}_j^{(i)} = \frac{E_s\left\|\underline{w}_j^T\underline{S}^{(i)}\right\|^2}{N_0\|\underline{w}_j\|^2 + \sum_{k\neq i}E_s\left\|\underline{w}_j^T\underline{S}^{(k)}\right\|^2\left(1-|\tilde{c}^{(k)}|^2\right)} \quad (14)$$

where $\underline{w}_j$ is the vector of feed-forward filter coefficients used in the jth iteration.

Proposition 9: Let $\mathcal{C}$ be a d-diversity code used in each thread in a setting with n transmit and m receive antennas. The SIR at the output of the iterative MMSE detector after j iterations is at least as large as the SIR after one iteration. Furthermore, the output SIR is at least as large as that produced by the zero-forcing detector.

Proof: If $\mathrm{SNR}_{\mathrm{zf}}$ denotes the SIR at the output of the zero-forcing detector, then it follows from the definition of the MMSE receiver that $\mathrm{SNR}_1^{(i)} \geq \mathrm{SNR}_{\mathrm{zf}}$. Also, from the definition of the MMSE filter, it follows that

$$\begin{aligned}\mathrm{SIR}_j^{(i)} &\geq \frac{E_s\left\|\underline{w}_1^T\underline{S}^{(i)}\right\|^2}{N_0\|\underline{w}_1\|^2 + \sum_{k\neq i}E_s\left\|\underline{w}_1^T\underline{S}^{(k)}\right\|^2\left(1-|\tilde{c}^{(k)}|^2\right)}\\ &\geq \frac{E_s\left\|\underline{w}_1^T\underline{S}^{(i)}\right\|^2}{N_0\|\underline{w}_1\|^2 + \sum_{k\neq i}E_s\left\|\underline{w}_1^T\underline{S}^{(k)}\right\|^2}\\ &= \mathrm{SIR}_1^{(i)}\end{aligned}$$

as was to be shown. □

In [18], Poor and Verdú have shown that the output of the MMSE receiver in AWGN channels can be tightly approximated by a Gaussian random variable. In the space–time code setting, the channel is AWGN when conditioned on the path gains. Thus, the two propositions imply that the diversity advantage achieved by the iterative MMSE receiver for the threaded architecture is approximately lower-bounded by the performance achieved by the zero-forcing receiver. Consequently, in a threaded architecture using d-space–time codes, the iterative MMSE receiver should achieve diversity d' satisfying

$$d\cdot(m-n+1) \leq d' \leq dm. \quad (15)$$

We note that this lower bound justifies our approach to code design for the threaded architecture. In particular, the design criteria developed in Theorems 2 and 7 for optimizing the channel coding for each thread in the absence of interference also serves to maximize a lower bound on the diversity advantage when the iterative MMSE detector is used to mitigate the interference from other threads. The simulation results of Section V suggest that the lower bound is, in fact, a pessimistic estimate of the performance of the threaded architecture with iterative MMSE multiuser detection.

TABLE I
COMPARISON OF DIFFERENT LAYERED ARCHITECTURES

Architecture	Efficiency	Quasi-Static Fading Diversity	B-Block Fading Diversity
H-Layering	nrb	$1 : m$ Layer-to-layer	$\lfloor B(1-r)\rfloor + 1 : m\,(\lfloor B(1-r)\rfloor + 1)$ Layer-to-layer
D-Layering	$nrb - \frac{n(n-1)rb}{\ell}$	$1 : m$ Symbol-to-symbol	$1 : m$ Symbol-to-symbol
H-Multi-Layering	nrb	$\lfloor \frac{1}{r}\rfloor(m-n+\lfloor\frac{1}{r}\rfloor){:}\lfloor\frac{1}{r}\rfloor m$ Layer-to-layer	?
Threaded	nrb	$(m-n+1)\,(\lfloor n(1-r)\rfloor+1) \le$ $d \le m\,(\lfloor n(1-r)\rfloor + 1)$	$(m-n+1)\,(\lfloor nB(1-r)\rfloor+1) \le$ $d \le m\,(\lfloor nB(1-r)\rfloor + 1)$

IV. SYSTEM COMPARISONS

A high-level comparison of the various architectures is shown in Table I. As shown in the table, all of the transmission formats achieve comparable efficiency. Here, efficiency refers to the number of information symbols per vector channel use. For example, in the horizontal layering scheme, there are n layers each containing a codeword of length ℓb and rate r. Thus, successful use of all transmission resources provides a total of $n \cdot (r\ell b)$ information symbols. Normalizing by the total number of symbol transmission intervals ℓ gives an efficiency of nrb information symbols per transmitted symbol interval. For the diagonal-layering approach, the efficiency is somewhat less since the diagonal layers cannot utilize a portion of the transmission resources (the result in the table assumes the width of the diagonal $w = 1$).

We also report the diversity orders achieved by the various architectures in both quasi-static and block fading channels. In the different approaches, the channel-coding schemes are assumed to achieve the maximum possible diversity level for rate r codes. Since no attempt was made in [1] to optimize the coding for the diagonal layering architecture, the results reported in the table are on a per-symbol basis. The diversity order achieved by the previous layered and multilayered architectures is variable. For these approaches, Table I shows the range of values (*minimum* : *maximum*) and notes whether the variation is from layer to layer or from symbol to symbol. In the case of the proposed threaded architecture, the diversity order is not variable, but the exact value is difficult to determine. In this case, the upper and lower bounds from (15) are used in Table I. For the block fading channel, the parameter B denotes the number of fading blocks per codeword.

The threaded layering is similar to H-BLAST in that each transmitted symbol in a thread is subject to interference from $n-1$ other layers, but better spatial diversity is achieved through more efficient transmit diversity and multiuser detection signal processing. The threaded layering is similar to D-BLAST in that all of the transmit antennas are used equally by each component coded transmission, but it more fully exploits the available temporal diversity since temporal interleaving is allowed across each transmit antenna. Furthermore, unlike D-BLAST, the threaded layering with space–time code design and iterative multiuser detection algorithms provide uniform spatial diversity from symbol to symbol. Finally, unlike the horizontal multilayering approach with group interference suppression, the threaded architecture provides uniform performance from one component space–time code to the next; and each component space–time code can, under the ideal interference cancellation assumption, achieve the maximum possible spatial and temporal diversity.

V. PERFORMANCE COMPARISONS

In this section, the different schemes are compared via simulation. In the study, we have used convolutional codes, the main advantage of which is the availability of computationally efficient SISO decoders. Periodic bit demultiplexers are used to distribute the encoder outputs across the different antennas. In addition, inner random interleavers are used to aid the convergence of the iterative MMSE receiver as discussed in Section III-A2). The error statistics are obtained by averaging the frame error rates of all the component codes. The channel decoder is based on the soft-output Viterbi algorithm (SOVA). Unless otherwise stated, the channel is assumed to follow the quasi-static fading model. The number of iterations for the iterative MMSE receiver is four. The code rate of the component codes is $1/2$.

Fig. 5 compares the performance of the iterative MMSE receiver with horizontal layering versus interference-free performance. In the case of the iterative MMSE, there are four transmit and four receive antennas, and the bandwidth efficiency is $\eta = 2$ b/s/Hz (i.e., BPSK modulation). The frame length corresponds to 100 transmissions. For the interference-free reference, there are four receive antennas but only one transmit antenna. The bandwidth efficiency in this case is $\eta = 0.5$ b/s/Hz. In general, the relation between the energy per bit to noise ratio E_b/N_0 and the total transmitted SNR is

$$\text{SNR} = \eta \frac{E_b}{N_0}. \tag{16}$$

The same SNR per transmit antenna is used in both scenarios, and the SNR reported in the figure is the total SNR for the four transmit antennas in the iterative MMSE case. The interference-free scenario represents a lower bound on the performance

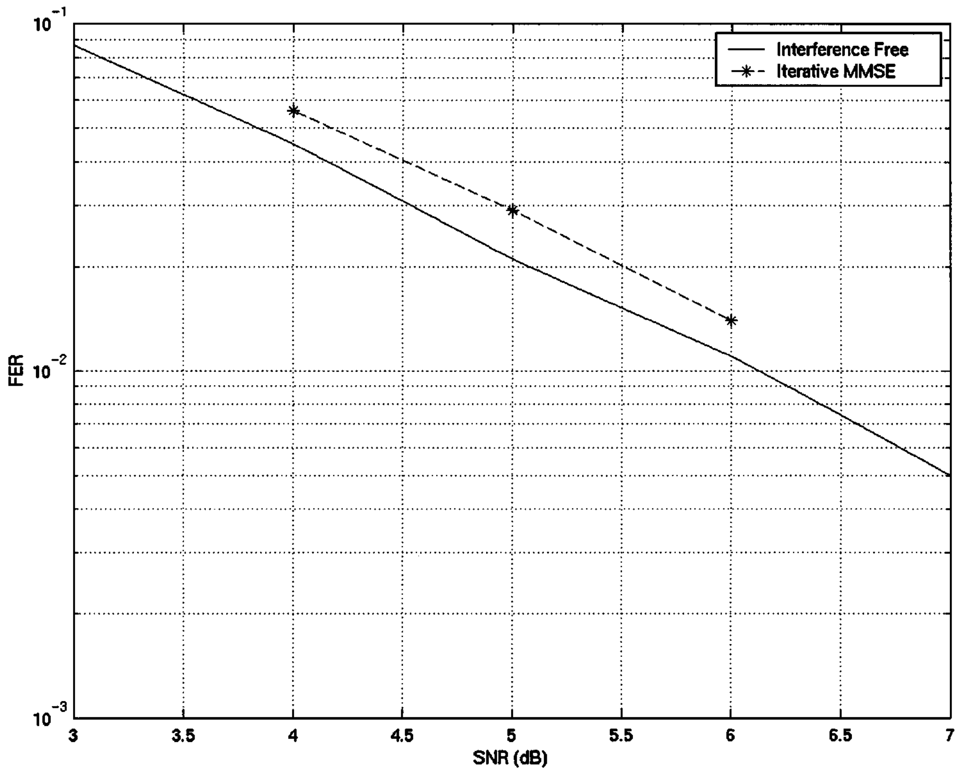

Fig. 5. Performance of the iterative MMSE receiver.

achieved by the optimum receiver. It is shown that the iterative MMSE receiver performs within a fraction of a decibel from the lower bound.

Now, we are interested in comparing the performance of the TST architecture presented in Section III with the D-BLAST architecture, and with the multilayering architecture proposed by Tarokh, Naguib, Seshadri, and Calderbank (TNSC) in [4].

Fig. 6 compares the TST architecture with a lower bound on the frame error rate achieved by D-BLAST in quasi-static fading channels. This lower bound assumes error-free decision feedback. In practice, the performance of D-BLAST is expected to be close to the lower bound at high SNRs (where the bound is tight) but much worse than the bound at lower SNRs (where the bound is loose). The same four-state convolutional code with generator polynomials $(5_8, 7_8)$ is used for both schemes. The iterative MMSE receiver is shown to provide a 3-dB gain over the D-BLAST lower bound under these conditions. Since the same code was used in both approaches, we can attribute the performance gain to the superiority of the iterative MMSE receiver over the signal processing algorithm used in the D-BLAST.

To further highlight the advantages of the threaded architecture, we report in Fig. 7 the same performance comparison for a block fading channel with three independent blocks per codeword. Due to the diagonal restriction imposed on each layer, the performance of the D-BLAST in this scenario is the same as that in the quasi-static fading channel. On the other hand, it is shown that the performance of the threaded architecture is improved by about 1 dB at 1% frame error rate without any additional complexity. This improvement is due to the increased diversity advantage achieved by efficient code design that exploits the additional temporal diversity.

Figs. 8 and 9 compare the performance of the TST and TNSC architectures for the cases of four transmit/four receive and eight transmit/eight receive antennas, respectively. QPSK modulation with Gray mapping is used to map the binary input at each antenna to a complex constellation. Hence, the spectral efficiency is 4 and 8 b/s/Hz, respectively. The frame length corresponds to 130 transmissions. The results of the TNSC scheme are obtained from [4, Figs. 4 and 6], respectively. The same four-state encoders are used for the TST architecture as in the previous case. Therefore, the overall complexity of the TST receiver including the iterations and soft-output decoding is in the same order as the 32-state decoders used in the TNSC [4]. From the figures, the significant gain provided by the TST over the TNSC scheme is clear. Indeed, the TST approach shows a gain of 4–8 dB over the TNSC scheme. The TST results are within 2–3 dB of the outage capacity. The gain in diversity advantage achieved by the TST architecture can be seen in the steeper asymptotic slope of the performance curve. It is also shown that the gain provided by the TST increases with the number of antennas. This can be attributed to the better exploitation of the diversity in the TST.

Finally, we note that by replacing the four-state code with a more powerful 64-state code we can close the gap between the TST frame error rate performance and the 10% outage capacity to less than a fraction of a decibel with the same system parameters.

VI. Conclusion

In this paper, we took a fresh look at the design problem for multiple-antenna systems operating over the fading channel. The problem was addressed from both a signal processing and a space–time coding perspective. From the space–time coding

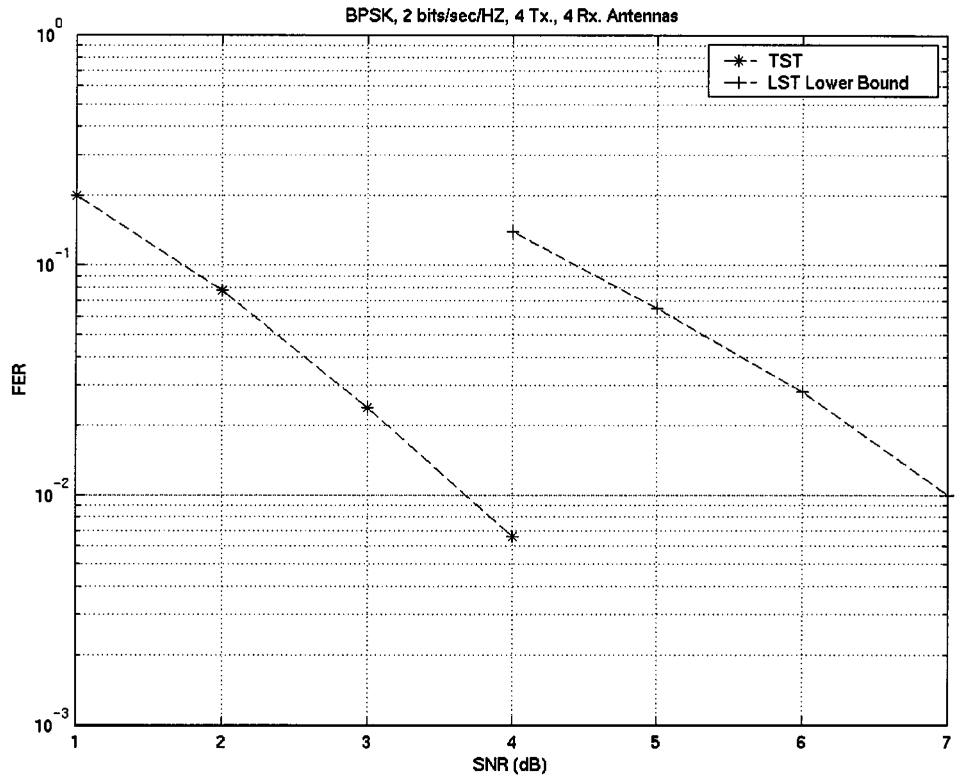

Fig. 6. Performance of the TST and BLAST architectures in quasi-static fading channels.

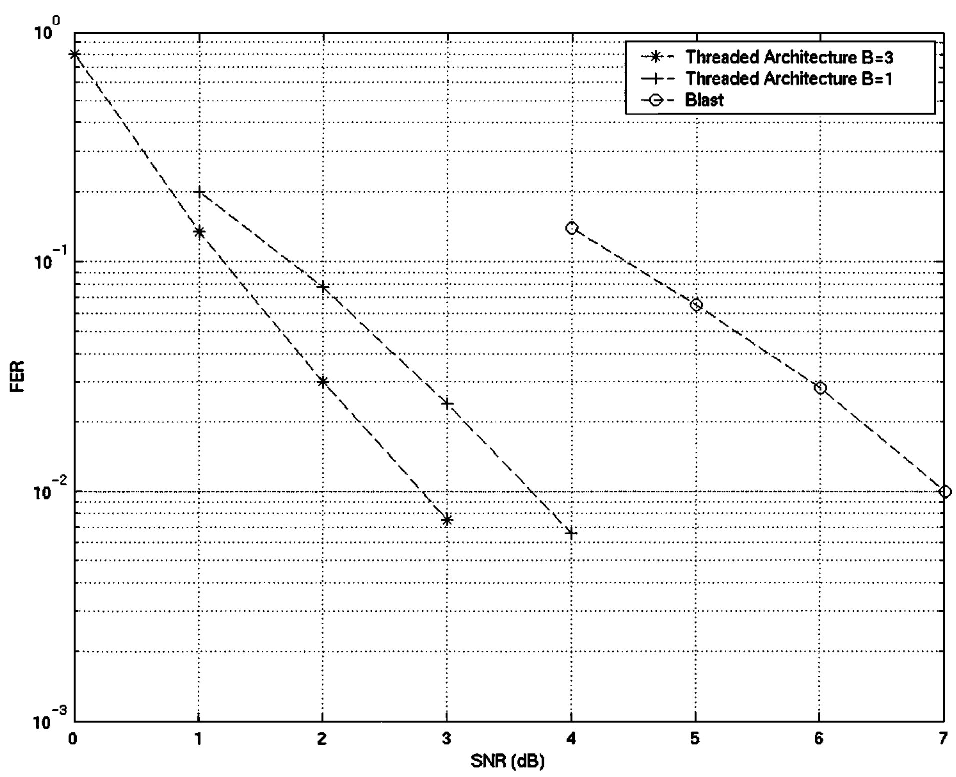

Fig. 7. Performance of the TST and BLAST architectures in block fading channels.

perspective, we presented a new generic approach, the TST architecture, that allows for exploiting the spatial and temporal diversity available in the system. From the signal processing side, we proposed to utilize the turbo processing principle to develop iterative algorithms for joint decoding and detection which offer several advantages over previously proposed techniques [1], [4]. Simulation results were provided for the iterative MMSE receiver establishing its ability to approach the in-

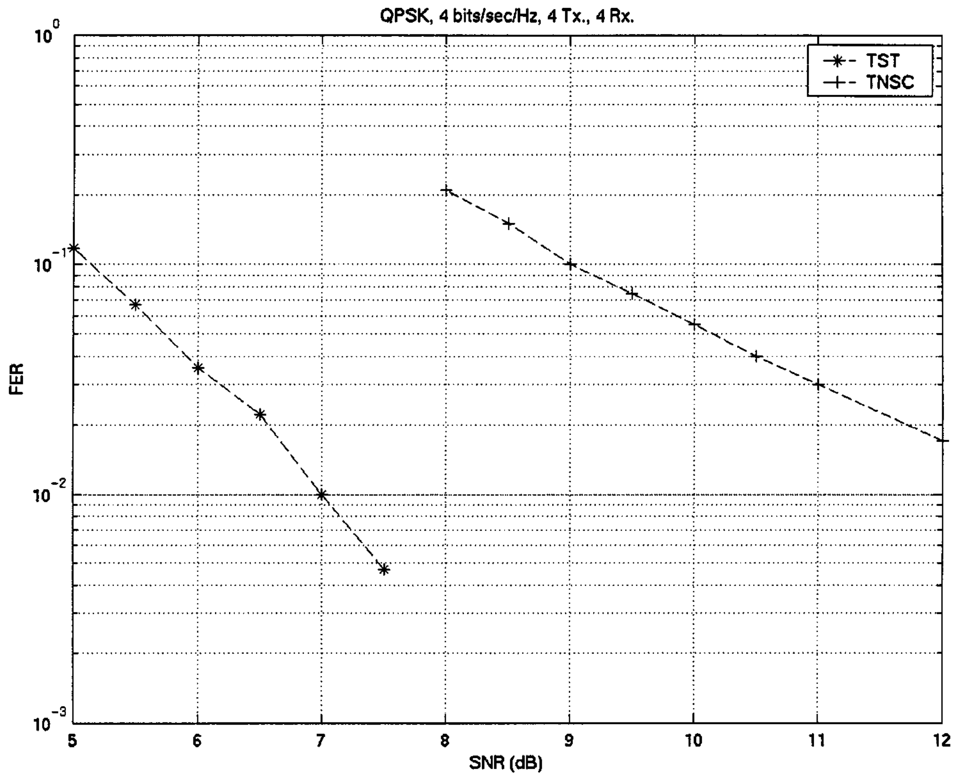

Fig. 8. Performance of the TST and TNSC architectures.

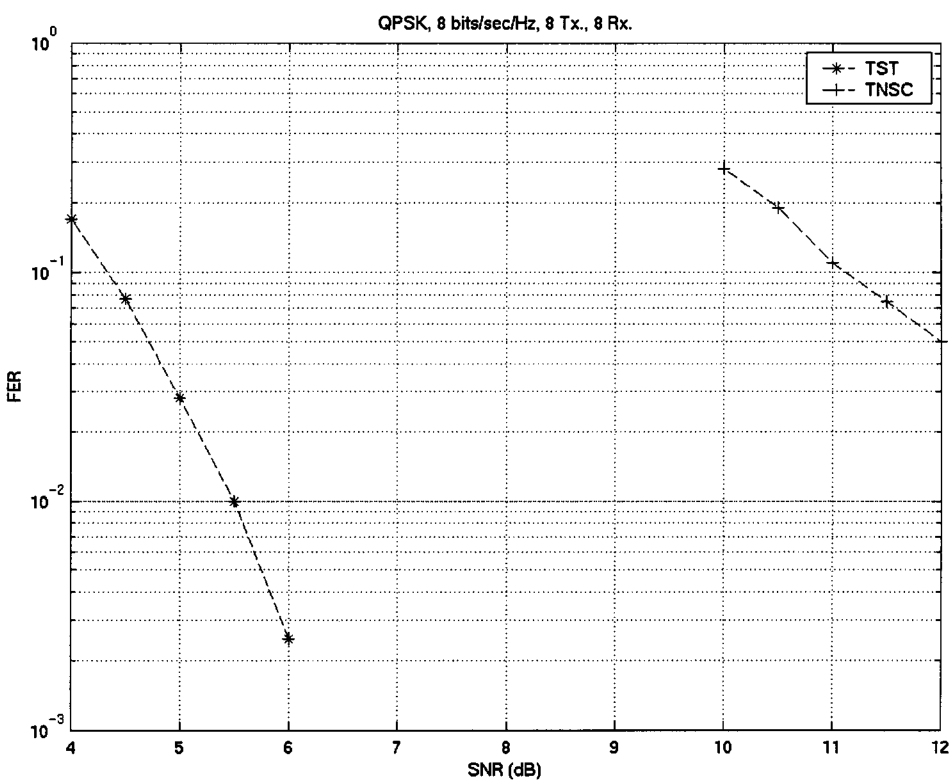

Fig. 9. Performance of the TST and TNSC architectures.

terference-free performance lower bound within a fraction of a decibel. The threaded architecture with efficient code design and iterative signal processing was shown, through simulation, to achieve significant gains over the D-BLAST and the combined array processing and space–time coding recently proposed by Tarokh *et al.* [4].

As a final remark, we note that, in the absence of interference from other threads, the fading channel is equivalent to the block fading channel with receive diversity, where the number of independent blocks in the equivalent model is equal to the product of the number of transmit antennas and the number of fading blocks. The algebraic framework that we developed for TST code design is, therefore, also useful in the study of code design for block fading channels and is applicable to both block and trellis-based codes. Conversely, optimization of the TST channel coding and interleaving schemes would also benefit from prior work on code design for such channels (see, for example, Lapidoth [12] or Wesel and Cioffi [19]).

REFERENCES

[1] G. J. Foschini, "Layered space-time architecture for wireless communication in fading environments when using multiple antennas," *Bell Labs. Tech. J.*, vol. 2, Autumn 1996.

[2] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Inform. Theory*, vol. 44, pp. 774–765, Mar. 1998.

[3] A. R. Hammons, Jr. and H. El Gamal, "On the theory of space–time codes for PSK modulation," *IEEE Trans. Inform. Theory*, vol. 46, pp. 524–542, Mar. 2000.

[4] V. Tarokh, A. Naguib, N. Seshadri, and A. R. Calderbank, "Combined array processing and space-time coding," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1121–1128, May 1999.

[5] E. Teletar, "Capacity of multi-antenna Gaussian channels," AT&T-Bell Labs, Tech. Rep., June 1995.

[6] G. J. Foschini and M. Gans, "On the limits of wireless communication in a fading environment when using multiple antennas," *Wireless Personal Commun.*, vol. 6, pp. 311–335, Mar. 1998.

[7] T. Marzetta and B. Hochwald, "Capacity of a mobile multiple antenna communication link in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 139–158, Jan. 1999.

[8] J.-C. Guey, M. R. Bell, M. P. Fitz, and W.-Y. Kuo, "Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels," in *Proc. IEEE Vehicular Technology Conf.*, Atlanta, GA, 1996, pp. 136–140.

[9] M. C. Reed, C. B. Schlegel, P. D. Alexander, and J. A. Asenstorfer, "Iterative multiuser detection for DS-CDMA with FEC," in *Proc. Int. Symp. Turbo Codes and Related Topics*, Brest, France, Sept. 1997, pp. 162–165.

[10] H. El Gamal and E. Geraniotis, "Iterative multiuser detection for coded CDMA signals in AWGN and fading channels," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 30–41, Jan. 2000. Special issue on Spread Spectrum for Global Communications.

[11] X. Wang and H. V. Poor, "Iterative (turbo) soft interference cancellation and decoding for coded CDMA," *IEEE Trans. Commun.*, vol. 47, pp. 1046–1061, July 1999.

[12] A. Lapidoth, "The performance of convolutional codes on the block erasure channel using various finite interleaving techniques," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1459–1473, Sept. 1997.

[13] J. Hagenauer, "The turbo principle: Tutorial introduction and state of the art," in *Proc. Int. Symp. Turbo Codes and Related Topics*, Brest, France, Sept. 1997, pp. 1–9.

[14] M. Moher, "An iterative multiuser decoder for near-capacity communications," *IEEE Trans. Commun.*, vol. 46, pp. 870–880, July 1998.

[15] P. D. Alexander, A. J. Grant, and M. C. Reed, "Iterative detection in code division multiple access with error control coding," *Europ. Trans. Telecommun.*, vol. 9, pp. 419–425, Sept./Oct. 1998.

[16] S. Benedetto, G. Montorsi, D. Divsalar, and F. Pollara, "A soft-input soft-output maximum *a posteriori* module to decode parallel and serially concatenated codes," *Telecom. and Data Acquisition Progr. Reps., Jet Propulsion Lab.*, no. 42-127, Nov. 1996.

[17] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1984.

[18] H. V. Poor and S. Verdú, "Probability of error in MMSE multiuser detection," *IEEE Trans. Inform. Theory*, vol. 43, pp. 858–871, May 1997.

[19] R. D. Wesel and J. M. Cioffi, "Joint trellis code and interleaver design," in *Proc. IEEE GLOBECOM'97*, Nov. 1997.