♦ Layered Space-Time Architecture for Wireless Communication in a Fading Environment When Using Multi-Element Antennas

Gerard J. Foschini

This paper addresses digital communication in a Rayleigh fading environment when the channel characteristic is unknown at the transmitter but is known (tracked) at the receiver. Inventing a codec architecture that can realize a significant portion of the great capacity promised by information theory is essential to a standout long-term position in highly competitive arenas like fixed and indoor wireless. Use (n_T, n_R) to express the number of antenna elements at the transmitter and receiver. An (n, n) analysis shows that despite the n received waves interfering randomly, capacity grows linearly with n and is enormous. With n = 8 at 1% outage and 21-dB average SNR at each receiving element, 42 b/s/Hz is achieved. The capacity is more than 40 times that of a (1, 1) system at the same total radiated transmitter power and bandwidth. Moreover, in some applications, n could be much larger than 8. In striving for significant fractions of such huge capacities, the question arises: Can one construct an (n, n) system whose capacity scales linearly with n, using as building blocks n separately coded one-dimensional (1-D) subsystems of equal capacity? With the aim of leveraging the already highly developed 1-D codec technology, this paper reports just such an invention. In this new architecture, signals are layered in space and time as suggested by a tight capacity bound.

Introduction

This paper describes a new point-to-point communication architecture employing an equal number of antenna array elements at both the transmitter and receiver. The architecture is designed for a Rayleigh fading environment in circumstances in which the transmitter does not have knowledge of the channel characteristic. This new communication structure, termed the *layered space-time architecture*, targets application in future generations of fixed wireless systems, bringing high bit rates to the office and home. The architecture might also be used in future wireless local area network (LAN) applications for which it promises extraordinarily high bit rates.

The architecture is a method of presenting and processing higher dimensional signals with the aim of leveraging the already highly developed one-dimensional (1-D) codec technology. Note that in this context, "higher dimensional" refers to *space*. (Generally, a bandwidth-efficient 1-D code involves many dimensions over the temporal domain. 1-D refers to a *complex* alphabet which is, of course, 2-D in terms of reals.)

As the paper points out, the capacity that this architecture enables is enormous. At first, the number of bits per cycle might seem too great to be meaningful. The capacity is achieved, however, in terms of n equal lower component capacities, one for each antenna at the receiver (or transmitter). A form of the new architecture attains a capacity equal to a tight

Panel 1. Abbreviations, Acronyms, and Terms

AWGN—additive white Gaussian noise
BER—bit error rate
codec—coder/decoder
iid—independent identically distributed
LAN—local area network
MEA—multi-element array
MMSE—minimum mean square error
SNR—signal-to-noise ratio

lower bound on the capacity attainable using multi-element arrays (MEAs) with an equal number of elements at both ends of the link. The next section describes this lower bound on capacity. Subsequently, the layered space-time architecture is discussed.

Although additional background details are available,[1] this paper provides a self-contained description of the architecture. The perspective is one of a complex baseband view of signaling over a fixed linear matrix channel with additive white Gaussian noise (AWGN). Time proceeds in discrete steps, normalized so that t = 0, 1, 2, ... The following notation and basic assumptions should be reviewed:

- *Number of antennas*. The MEA at the transmitter has n_T. The MEA at the receiver has n_R. For convenience, the pair (n_T, n_R) denotes a communication system with n_T transmit elements and n_R receive elements. **Figure 1** illustrates the notation.
- *Transmitted signal s(t)*. This signal has fixed narrow bandwidth. The total power is constrained to P regardless of n_T (which is the dimension of s(t)). The bandwidth is narrow enough that the channel frequency characteristic can be treated as flat across the band.
- *Noise at receiver* $\nu(t)$. This is the complex n_R-dimensional AWGN. The components are statistically independent and of identical power N at each of the n_R antenna outputs.
- *Received signal r(t)*. At each point in time, this is an n_R-dimensional signal. There is one complex vector component per receive antenna. With each transmit antenna transmitting power P/n_T, P denotes the average power at the output of each receiving antenna, with "average" meaning spatial average.
- *Average signal-to-noise ratio (SNR) at each receive antenna*. This is $\rho = P/N$, independent of n_T.
- *Matrix channel impulse response g(t)*. This matrix has n_T columns and n_R rows. The notation h(t) is used for the normalized form of g(t). The normalization is such that each element of h(t) has a spatial average power loss of unity.

The basic vector equation describing the channel operating on the signal is

$$r = g * s + \nu, \qquad (1)$$

where "$*$" means convolution. These three vectors are complex n_R-dimensional vectors ($2n_R$ real dimensions). Because of the narrowband assumption, the channel Fourier transform G is treated as a matrix constant over the band of interest. Thus, g is written for the nonzero value of the channel impulse response, thereby suppressing the time dependence of g(t). The same is true for h and its Fourier transform H. Thus, in normalized form, (1) becomes

$$r = \sqrt{\rho / n_T} \cdot h \cdot s + \nu\,. \qquad (1a)$$

The random channel model we use is the Rayleigh channel model. Assume that the MEA elements at each end of the communication link are separated by about half a wavelength. At 5 GHz, for example, half a wavelength measures only about 3 cm, so many antenna elements are often possible. (Additionally, there are two states of polarization [see **Figure 2**]). With a half-wavelength separation, the Rayleigh model for the $n_R \times n_T$ matrix H representing the channel in the frequency domain is approximated by a matrix having the following independent identically distributed (iid), complex, zero-mean, unit-variance entries:

- $H_{ij} = \text{Normal}\left(0, 1/\sqrt{2}\right) + \sqrt{-1} \cdot \text{Normal}\left(0, 1/\sqrt{2}\right)$.
- $|H_{ij}|^2$ is a chi-squared variate with two degrees of freedom denoted by χ_2^2 but normalized so $E|H_{ij}|^2 = 1$.

Capacity

The viewpoint assumed in treating capacity is discussed next. We stress that capacity is a limit to error-free bit rate that is provided by information theory,

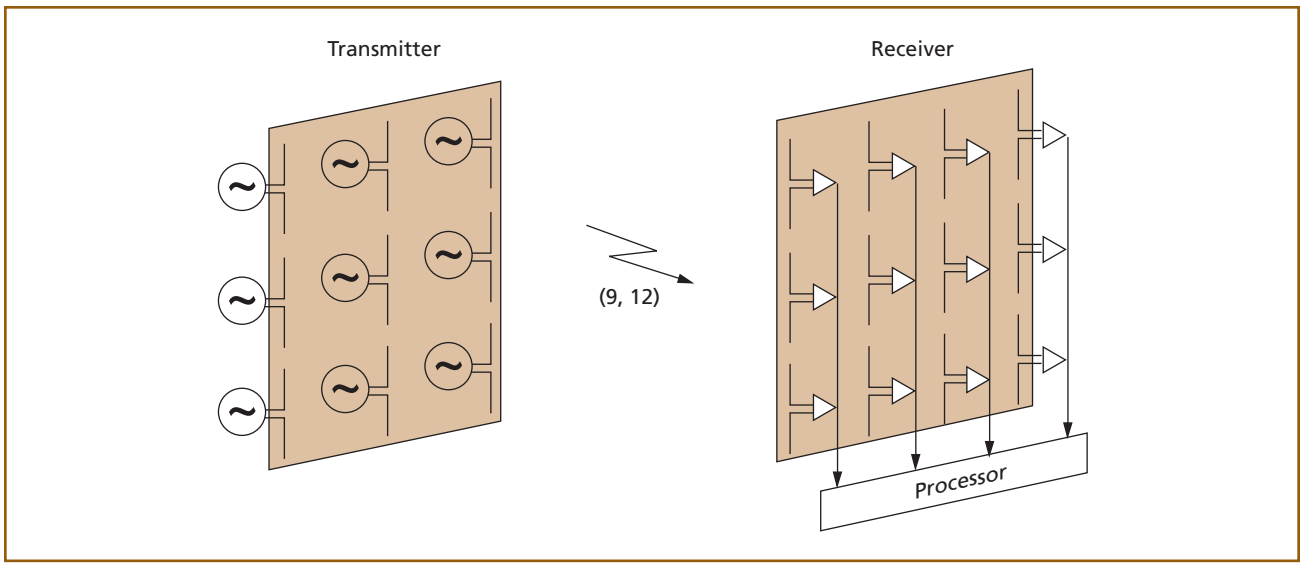

Figure 1.
Shorthand notation (n_T, n_R) for the number of transmit antennas n_T and receive antennas n_R (dipole antennas shown).

and this limit can only be approached in practice with the advance of technology: any working system can only achieve a bit rate (at some desired small bit error rate [BER]) that is only a fraction of capacity. In what follows, the term "capacity" will often be used as an indicator of some smaller deliverable bit rate.

Long-Burst Perspective

Communication in long bursts means bursts having many symbols—so many that an infinite time horizon information-theoretic description of communication portrays a meaningful idealization. Yet bursts are assumed to be of short enough duration that a channel is essentially unchanged during a burst. The channel is assumed to be unknown to the transmitter but learned (tracked) by the receiver. The channel might change considerably from one burst to the next.

By a channel being unknown to the transmitter, we mean that the realization of H during a burst is unknown. Actually, the average SNR value and even n_R might not be known to the transmitter. Nonetheless, for purposes of this discussion, these two parameters are considered to be known. The reason for this is that at the transmitter, one assumes that communication is taking place with a user for which *at least* a certain n_R and average SNR are available. These minimum values represent what the transmitter conservatively uses to determine a capacity value that is nearly always available.

In a given system, not all communication bursts are successful. As explained below, for some small percentage of instances of H, the transmitter's assumed capacity value may be too optimistic. In such cases, delivering the bit rate at the desired BER required of a successful burst may be impossible. When it is impossible, a channel outage is said to have occurred and the channel is considered to be in the OUT state.

Outage is dealt with probabilistically because H is random; thus, capacity is a random variable. The channel is random even though the base and user in an office LAN environment or the communication sites in fixed wireless applications may be "fixed." In actuality, the reason for this is that such sites are only nominally fixed because perturbations of the communicators and the communication medium are possible.

For indoor LANs—even for a user at a desk—some motion in and around the workspace is likely. Not only people but various (especially metal) objects could be moving in the propagation path. For predominately outdoor fixed wireless links, weather-related motion of antenna structures occurs, as well as significant channel changes due to, say, vehicles and foliage. Half-wavelength movements can be important. Thus, assuming that the channel is fixed during a burst, the channel may vary from burst to burst,

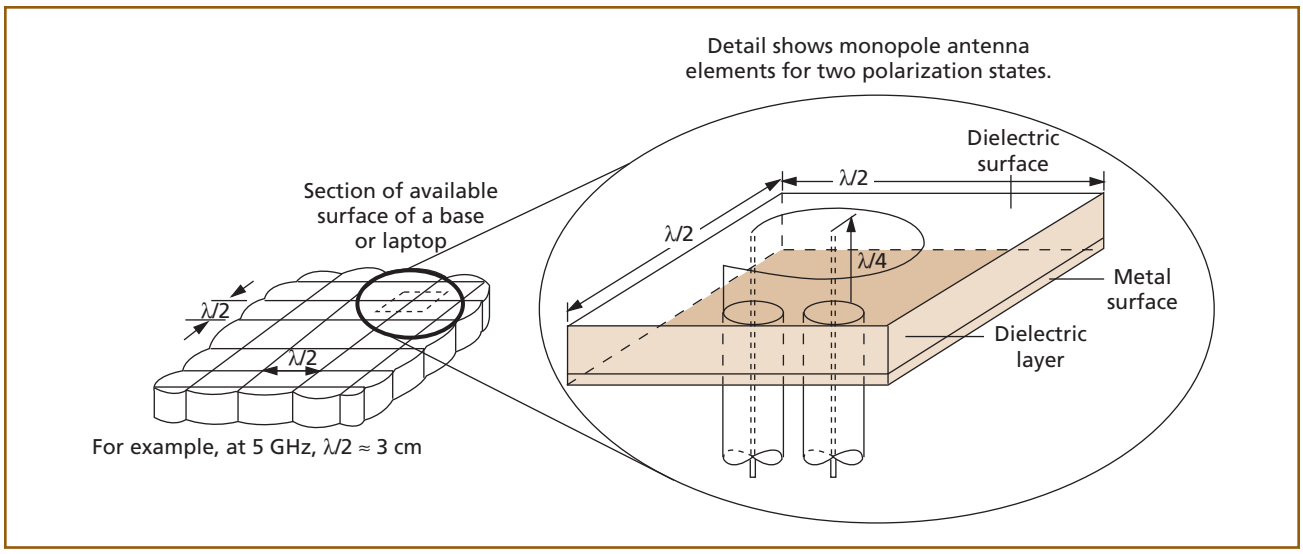

Figure 2.
Section of casing paved with half-wavelength lattice.

and one might be interested in the capacity that can be attained in nearly all transmissions (for example, 95 to 99% or even higher).

Complementary capacity distributions, discussed below, focus on the high-probability tail. (In special applications like very large file transfers, however, maximum attainable throughput over long time durations may be a preferable figure of merit.) A companion article[1] mentions that based on the results of research,[2,3] the transmitter can use a single code even though the specific value of the H matrix is unknown.

The distribution of capacity is derived from an ensemble of statistically independent Gaussian $n_R \times n_T$ H matrices (the aforementioned Rayleigh model). In this paper, the system is considered to be either OUT or NOT OUT for each realization of H. As mentioned earlier, the OUT state corresponds to the event that a prespecified capacity level (for example, X) cannot be met. For instance, given a 1% outage level, one would say a certain capacity can be assured at that level.

By employing MEAs, capacity tail probabilities can be significantly improved. The subsection "Opportunity for Enormous Bit Rates" below discusses the great capacity available and how the tail probability improves with larger and larger n.

(For cases in which the time constant of channel change is very large, extra receive antenna elements may be needed to ensure that outage is minimal (see the end of the "Conclusion" section). In very severe situations, provision for movement of the receive antenna may be desirable to avoid the risk of being stuck with an undesirable H for excessive time. Deployment of a relay site is yet another alternative. Fading correlation time and its incorporation into more refined performance criteria is an interesting subject for future investigation in measurement and analytical studies.)

Key Capacity Expressions

A generalized capacity formula and a capacity lower-bound formula are referred to below.[1] The generalized formula is derived from other basic formulas.[4-6] Additionally, the capacity formula for optimum ratio combining is needed.

The generalized capacity formula for the general (n_T, n_R) case is

$$C = \log_2 \det\left[I_{n_R} + (\rho / n_T)\cdot HH^{\dagger}\right] \text{ b/s/Hz}. \quad (2)$$

In this equation, "det" means determinant, I_{n_R} is the $n_R \times n_R$ identity matrix and "†" means transpose conjugate.

The capacity lower bound for the (n, n) case in terms of n independent chi-squared variates is

$$C > \sum_{k=1}^{n} \log_2\left[1 + (\rho / n)\cdot \chi^2_{2k}\right] \text{ b/s/Hz}. \quad (3)$$

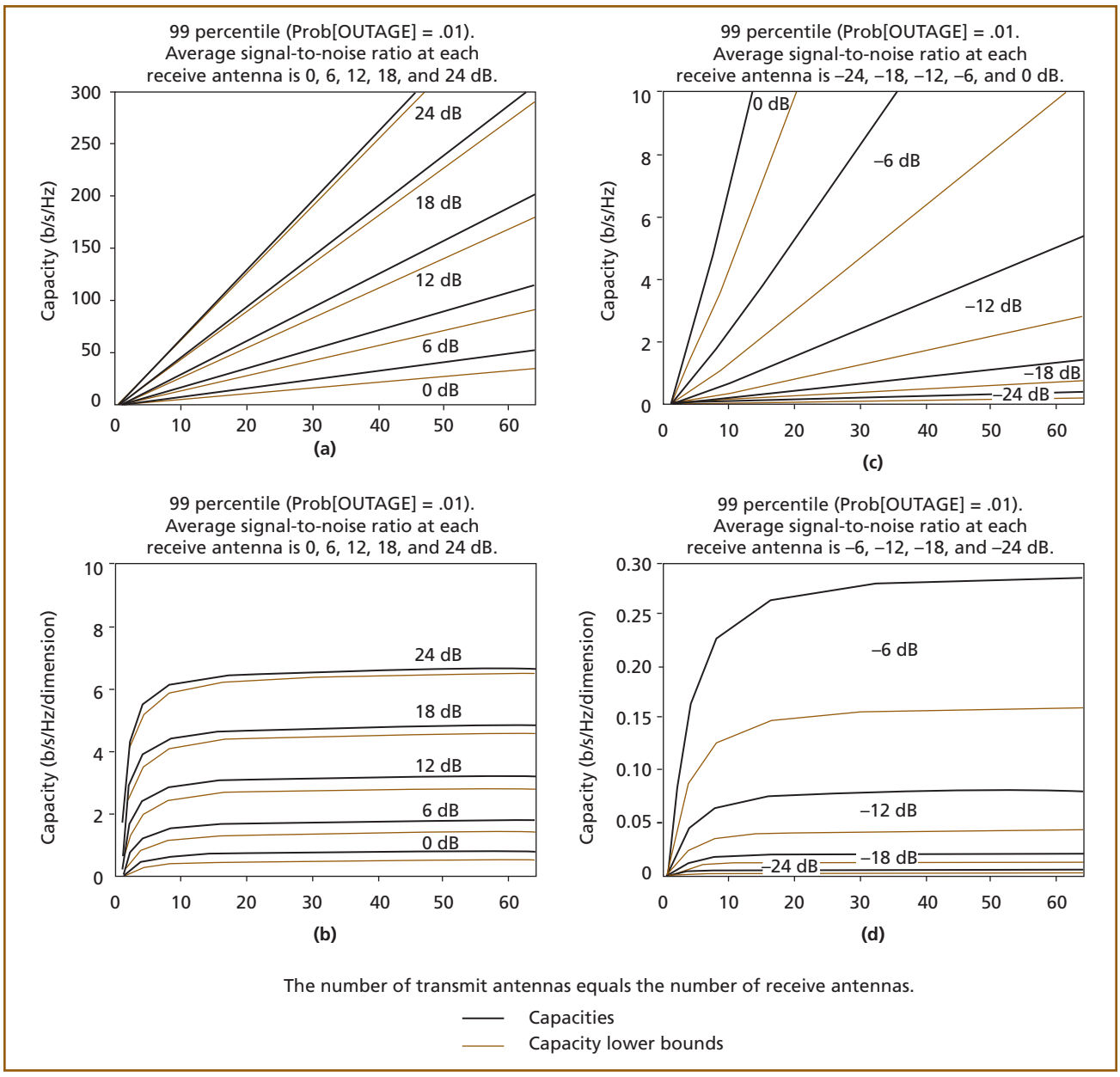

Figure 3(a).
Capacity in b/s/Hz versus the number of antenna elements at each site.
Figure 3(b).
Capacity in b/s/Hz/dimension versus the number of antenna elements at each site.
Figure 3(c).
Capacity in b/s/Hz versus the number of antenna elements at each site.
Figure 3(d).
Capacity in b/s/Hz/dimension versus the number of antenna elements at each site.

Notice that some nonstandard notations have been used—for example, χ^2_{2k} to denote *directly* a chi-squared variate with 2k degrees of freedom. Because the entries of H are zero mean unit variance *complex* Gaussians, the mean of this variate is k. As discussed later in an asymptotic sense, the bound in (3) for large ρ and n is quite tight. While (3) was initially derived elsewhere,[1] the subsection "A (6, 6) Example of Processing at the Receiver" below includes a rederivation of (3) that is constructive. That is, the right-hand

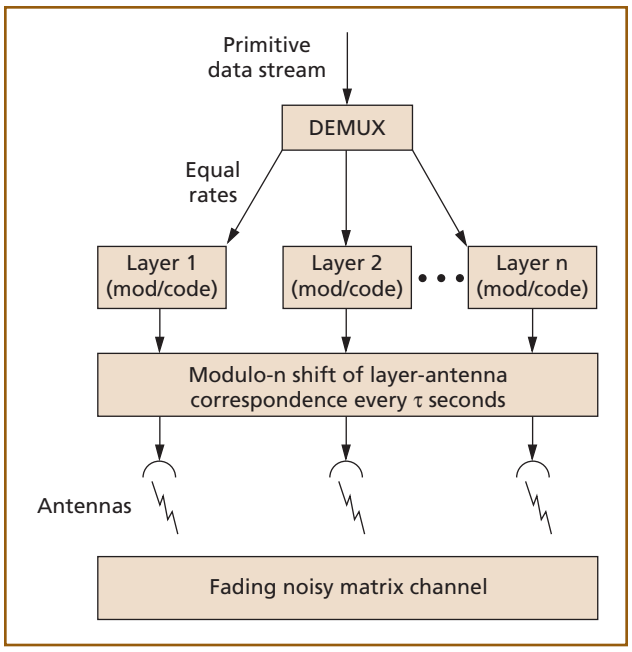

Figure 4.
Transmission process using space-time layering.

side will be associated with the capacity of a limiting form of a communication architecture using 1-D codecs.

A special application of (2)—important to this discussion—is the case of optimum combining, which corresponds to (1, n) receive diversity. OC(n) is written to convey a system employing optimum combining with n-fold receive diversity.

The capacity formula for optimum ratio combining or receive diversity ($n_T = 1$, $n_R = n$) is

$$C = \log_2\left[1 + \rho \cdot \chi^2_{2n}\right] \text{ b/s/Hz} . \qquad (4)$$

The lower-bound (3) suggests that in some sense, one might be able to embed n OC(k) systems (with k = 1, 2, ... n) in an (n, n) system. The argument of the logarithm in (3) suggests that each of the n single-transmit antenna systems would have transmit power ρ/n so that each receive array element has an average SNR of ρ/n. As explained below, such embedding is indeed possible.

Opportunity for Enormous Bit Rates

Before demonstrating how to do the embedding, the great capacity at stake is worth reviewing. **Figure 3(a)** shows the capacity 99 percentile (1% outage) for SNRs of 0 dB to 24 dB in steps of 6 dB. As noted from the ordinate, which ranges to 300 b/s/Hz, the capacities are enormous. For example, at a 12-dB SNR and even for modest numbers of antenna elements like eight or 12, significant capacity is available—that is, about 21 and 32 b/s/Hz, respectively. Even at a 0-dB SNR, very significant capacity exists. About 25 b/s/Hz is available for, say, n = 32.

From the standpoint of signal constellations, one must be concerned with per-dimension capacities. Figure 3(b) shows the same capacity results as Figure 3(a) but they are expressed in terms of the b/s/Hz/signal dimension. In preparing Figure 3(b), the cases n = 1, 2, 4, 8, 16, 32, and 64 were actually computed and the remaining cases interpolated. Note that even when the overall capacities are great, the per-dimension capacities can be reasonable. Figures 3(a) and 3(b) depict the capacity (bold curves), as well as the capacity lower bound (light curves). The capacity lower bound is quite tight at the higher ρ values.

As the antennas increase in number, saturation of the lower bound on the per-dimension capacity becomes evident as Figure 3(b) indicates. This asymptotic behavior can be explained by looking at the right-hand side of the inequality in (3). For the right-hand side divided by n, the large n asymptote is[1]

$$\int_0^1 \log_2\left(1+\rho x\right) dx = \left(1+\rho^{-1}\right)\cdot\log_2\left(1+\rho\right) - \log_2 e. \qquad (5)$$

Note that in the limit of large ρ, the dominant term in the last expression is $\log_2(\rho/e)$. For example, as Figure 3(b) shows, for an SNR of 24 dB and n = 64, an asymptotic value of about 6.5 b/s/Hz/dimension and indeed $\log_2(10^{2.4}) \approx 6.5$. The "Asymptotic Optimality of the Layered Architecture" subsection later on concludes that not just the lower bound but also the capacity per dimension or C/n converges to $\log_2(\rho/e)$ as ρ and n increase without bound.

Figures 3(c) and 3(d) correspond to Figures 3(a) and 3(b) but only for SNRs that are negative when expressed in dB. Note that even at negative SNRs, interesting capacity levels are possible. The tightness of the lower bound deteriorates more and more, however, as ρ is lowered. The "Related Options" section later in the paper further discusses these aspects.

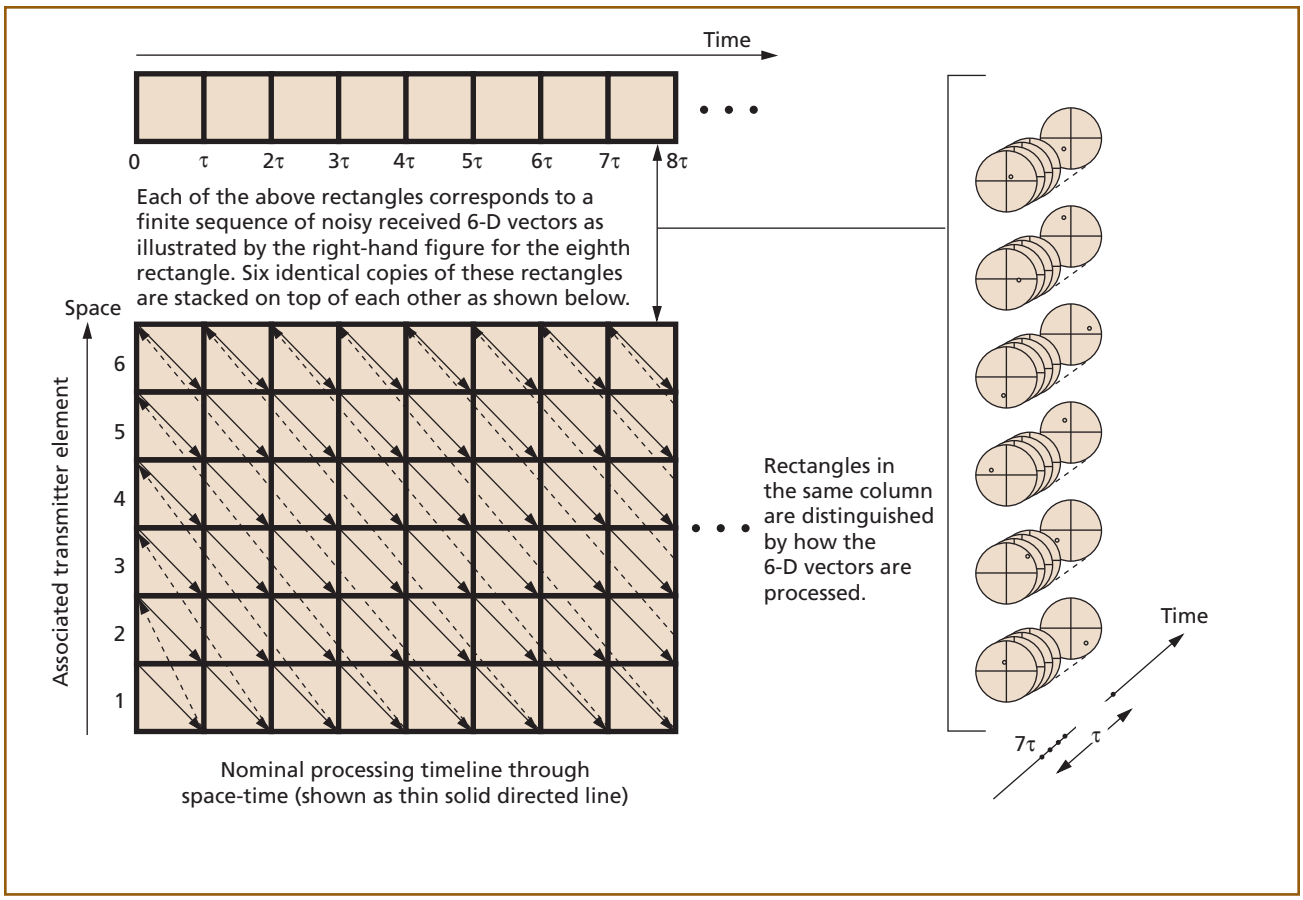

Figure 5.
Flow of nominal processing time for a received signal.

Layered Space-Time Architecture

This section provides a high-level description of a form of the new architecture having an equal number of antenna elements at both ends of the link. Its capacity is associated with the lower bound (3) on the capacity achievable with MEAs for the higher SNR values indicated in Figure 3(a). A simple description of the architecture is provided following some brief mathematical background information, which is needed to establish that the architecture indeed offers tremendous capacity. The previously mentioned "Related Options" section provides some advantageous variations on the architectural theme.

Mathematical Background

The following linear algebra helps clarify the architecture. Let $H_{\cdot j}$ with $1 \le j \le n$ denote the n columns of the H matrix ordered left to right so that $H = (H_{\cdot 1}, H_{\cdot 2}, \ldots H_{\cdot n})$. For each k such that $1 \le k \le n + 1$, let $\mathbf{H}_{\cdot[k,n]}$ denote the vector space spanned by the column vectors $H_{\cdot j}$ satisfying $k \le j \le n$. Because no such column vectors exist when $k = n + 1$, the space $\mathbf{H}_{\cdot[n+1,n]}$ is simply the null space. Note that the joint density of the entries of H is a spherically symmetric (complex) n^2-dimensional Gaussian. This makes it possible to state that, with probability one, $\mathbf{H}_{\cdot[k,n]}$ is of dimension $n - k + 1$. Furthermore, with probability one, the space of vectors perpendicular to $\mathbf{H}_{\cdot[k,n]}$ denoted as $\mathbf{H}^{\perp}_{\cdot[k,n]}$ is $k - 1$ dimensional. For $j = 1, 2, \ldots n$, η_j is defined as the projection of $H_{\cdot j}$ into the subspace $\mathbf{H}^{\perp}_{\cdot[j+1,n]}$.

As explained next, with probability one, each η_j is essentially a complex j-dimensional vector with iid N(0,1) components (η_n is just $H_{\cdot n}$). Strictly speaking, an η_j is n dimensional. When viewing η_j using an orthonormal basis—with the first basis vectors being those spanning $\mathbf{H}^{\perp}_{\cdot[j+1,n]}$ and the remaining vectors

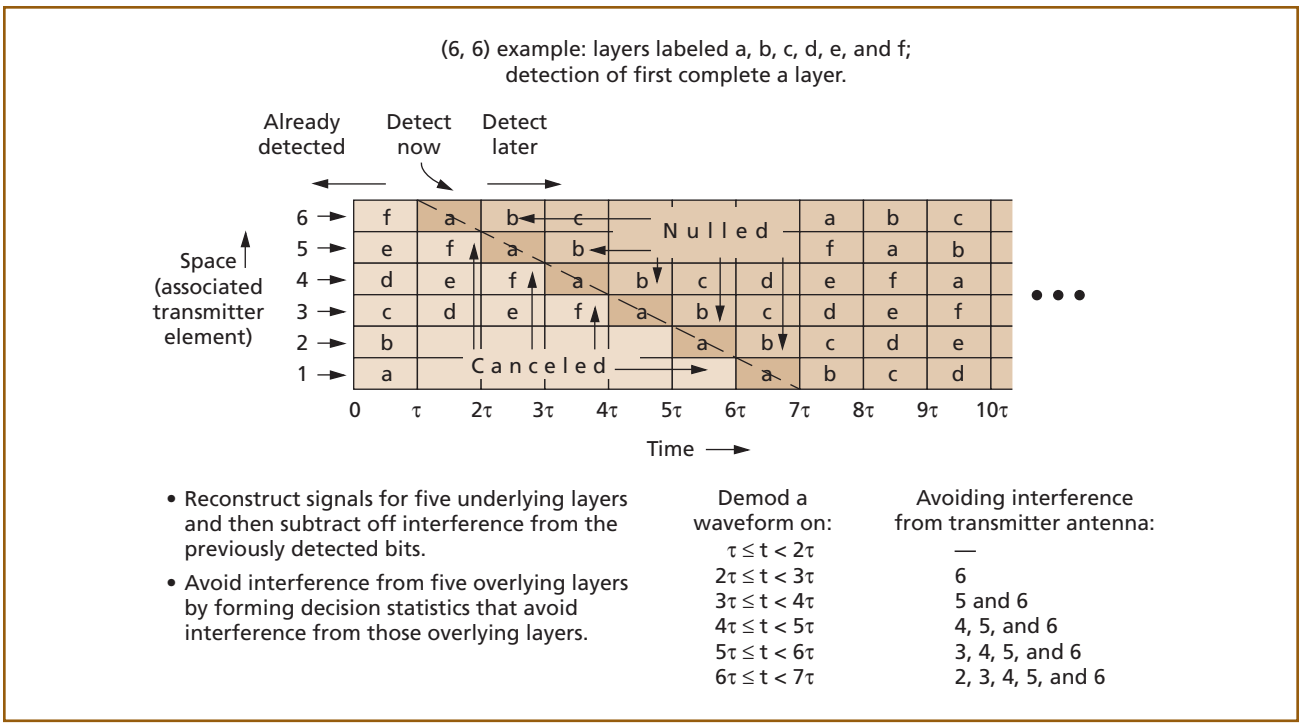

Figure 6.
Space-time layering in reception.

being those spanning $\mathbf{H}_{[j+1,n]}$ —the first j components of η_j are iid complex Gaussians while the remaining components are all zero. Looking at these projections in the order η_n, η_{n-1}, ... η_1 shows that the totality of the $n \times (n+1)/2$ nonzero components are all iid standard complex Gaussians. Consequently, the ordered sequence of squared lengths are statistically independent chi-squared variates having 2n, 2(n – 1), ... 2 degrees of freedom, respectively. With the choice of normalization presented in this paper, the mean of the squared length of η_j is j.

Transmission

In a spectrally economical system, the layered space-time architecture described here would be employed in conjunction with an efficient 1-D code. The form of the code employed in a specific instance of the architecture is not within the scope of this paper. For expositional simplicity, however, it is best to begin the description by considering some nonspecific block code rather than a convolutional code implementation.

Figure 4 illustrates the transmission process. A primitive data stream is demultiplexed into n data streams of equal rate. Each data stream is encoded in some unspecified way except to say that the encoders can proceed without sharing any information with each other. Rather than committing each of the n-encoded streams to an antenna, the bit-stream/antenna association is periodically cycled. The dwelling time on any association is τ seconds so that a full cycle takes $n \times \tau$ seconds. The n-encoded substreams, then, share a balanced presence over all n paths to the receiver. Therefore, none of the individual substreams is hostage to the worst of the n paths.

With communication structured in this balanced way, each subchannel has the same capacity. This setup serves to "uniformize" the multiplexing/demultiplexing and coding/decoding processes—that is, all n constituents are rendered virtually identical in structure. Of course, because the balance makes it possible to use the same constellation for each subchannel, the lowest maximum number of constellation points per subchannel is obtained. Each channel is essentially the same regarding the opportunity for coding.

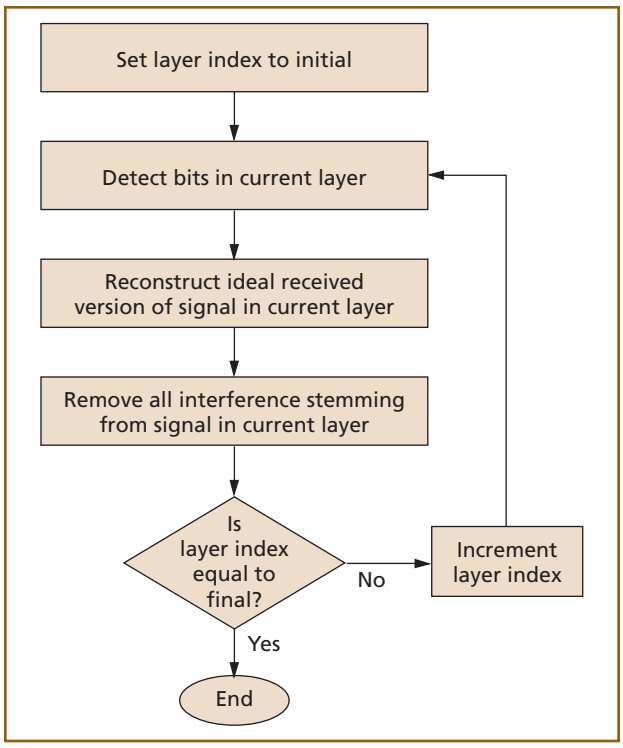

Figure 7.
Temporal view of the processing of successive space-time layers.

In the next subsection, it will be seen that within each substream, "good" symbols (those with a high SNR) can compensate for "bad" symbols (those with a low SNR) through coding. The subsection will help clarify that in a certain sense, the capacity obtained is the sum given by the right-hand side of (3).

As pointed out in the "Related Options" section below, advantages can be achieved by allowing an optional second stage of multiplexing/demultiplexing. For now, however, this discussion assumes only one stage of multiplexing/demultiplexing as indicated in Figure 4.

A (6, 6) Example of Processing at the Receiver

In describing processing at the receiver, a (6, 6) example is used. The extension to arbitrary (n, n) is immediate. A training phase (not described here) is assumed to be already completed. During this start-up phase, known signals were transmitted and processed at the receiver to expedite the H matrix becoming accurately known to the receiver. The transmitter, however, does not know the channel.

The top of **Figure 5** shows the first eight of a finite sequence of rectangles. Each rectangle in this linear sequence symbolizes a sequence of τ 6–D received vectors. The right side of the figure illustrates the succession of τ 6–D vectors (with complex components) corresponding to the eighth rectangle. These are the τ vectors arriving on the time interval $[7\tau, 8\tau)$. The heavy dots in the planes (circular sections shown) represent the complex received signal components that can be seen for the first and last of this sequence of τ vectors. Each of these vectors includes noise plus n-interfering transmitted signals from the transmit antennas.

For clarification, visualize constructing a stack of six identical copies of this aforementioned sequence of rectangles, one atop the other as shown in the figure. This stack is a visual aid for explaining how the received signals are preprocessed. A spatial element is associated with each rectangle on this "wall" of rectangles—specifically, a transmitter antenna—depending on the ordinate value. These elements are numbered 1, 2, ... 6 as are the ordinate values to which they correspond. The resulting rectangular partition of space-time must be understood figuratively as both space and time are discrete. The duration τ can span any number of time units and, as previously mentioned, each space element is associated with the single transmitter element as indicated on the left of the stack.

Note that for the same rectangle base interval of duration τ, the very same τ-consecutive vector signals with complex components are associated with each of the six vertically stacked rectangles. The six rectangles having the same τ-duration base interval will be distinguished by the preprocessing applied to the vector signals. The transmitter antenna associations were made because a rectangle at ordinate i will play a key role in the process of extracting the signal radiated by the i^{th} transmitter. As explained later, besides relating to the nature of information to be extracted, the ordinate also determines the way in which interference must be handled in the course of extracting information.

Processing time is distinct from signal reception time. Figure 5 illustrates the flow of time in processing the received signal. As processing time passes, processing proceeds top to bottom along a succession of con-

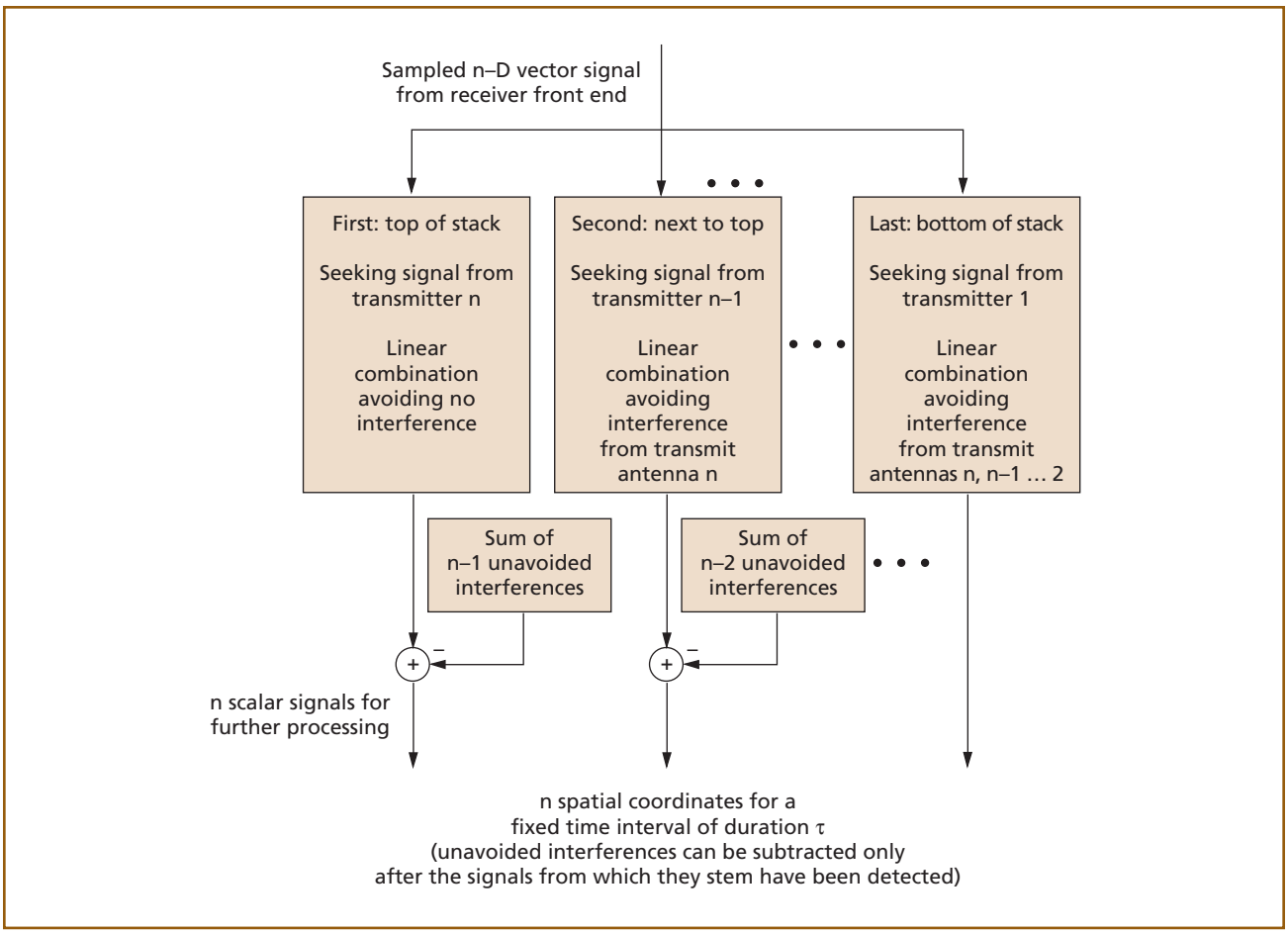

Figure 8.
Spatial view of receiver processing.

secutive space-time layers (diagonals), moving left to right as indicated by the thin solid directed line (really an ordered sequence of directed lines). The time flow in Figure 5 is only nominal for two reasons. First, with block coding where we assume that a layer is synonymous with a code block, no time arrow need be associated with the processing of the symbols of a block. Second, for convolutional coding, which has a definite time direction, a significant modification of the obvious time progression within each full layer might have an advantage as explained later.

The central theme of the architecture is *interference avoidance*, and this discussion assumes that interfering signals will be nulled out. (As discussed later, instead of nulling, SNR could be maximized. In such a case, "noise" means including not just AWGN but all interferers not yet subtracted out.) Fewer interferers must be nulled at the higher stack levels. The interferences that need not be nulled are those that will be subtracted out. Of course, when nulling interferers, any possible enhancement of the noise caused by the interference nulling process must be carefully assessed. As explained later in reference to the mathematical setup that has been carefully tailored for capacity analysis, the noise assessment will be easy to do.

Figure 6 illustrates additional details of the steps required for proceeding along iterated diagonal layers. For expositional convenience, a repetitive *abcdef* labeling on the stack is included. Detection of the first complete diagonal *a* layer through which is drawn a dashed diagonal line is described. Other layers, including boundary layers, are handled similarly. Boundary layers are those layers involved with where a burst starts or ends (those having fewer than six rectangles). The first complete *a* layer comprises six parts, $a_{j\tau}(t)$ (j = 1, 2, ... 6), in which the subscript indicates at

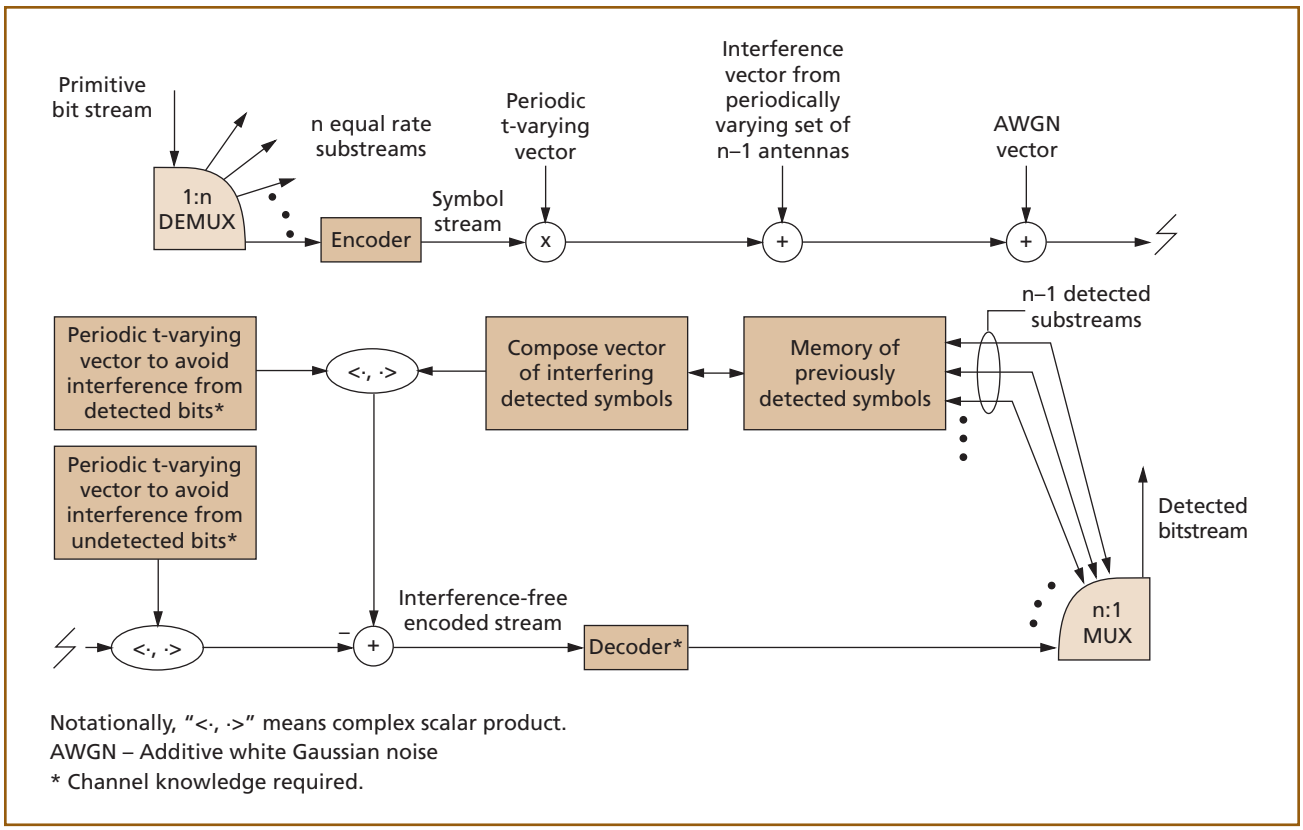

Figure 9.
System diagram of the processing involved at the receiver (discrete time baseband perspective).

what point in time the part that lasts for τ time units begins. All layers relatively disposed to be located partially underneath this *a* layer are assumed already successfully detected while all layers disposed to be partially above the *a* layer are yet to be detected. The capacity associated with this case will also be found, and then the capacity associated with the (n, n) case will be apparent. (With block coding, a full layer could correspond to exactly one block, although as pointed out later, associating more than one block within a layer can sometimes be advantageous).

Next, before computing capacity, a connection is made to the earlier "Mathematical Background" subsection by pointing out the relevance of projecting a received signal vector into $\mathbf{H}^{\perp}_{\cdot[k+1,n]}$. Assume that one received signal vector within a rectangle of ordinate k is being preprocessed. The purpose of the preprocessing stage is to help later determine the signal sent from antenna k. The aim of the preprocessing is to yield a vector free of interference from all signals that were simultaneously transmitted from antennas other than the k^{th}. The interference stemming from signals that were simultaneously transmitted from antennas 1, 2, ..., k–1 are inconsequential because these signals are assumed to have been already perfectly detected and subtracted out. What *is* necessary is to null out the interference from yet undetected signals—namely, those simultaneously transmitted from antennas k+1, k+2, ... n. This is exactly what is accomplished by projecting a received signal vector into $\mathbf{H}^{\perp}_{\cdot[k+1,n]}$ because that space is the maximal subspace orthogonal to the subspace spanned by the signals received from transmitters k+1, ... n.

This discussion has stressed that on each of the six time intervals for the *a* layer, a different number of interferers to be nulled must be addressed. For each of the six intervals in turn, the capacity of a corresponding hypothetical system is expressed in which the additive interference situation holds for all time. For the first time interval, the five layers below have

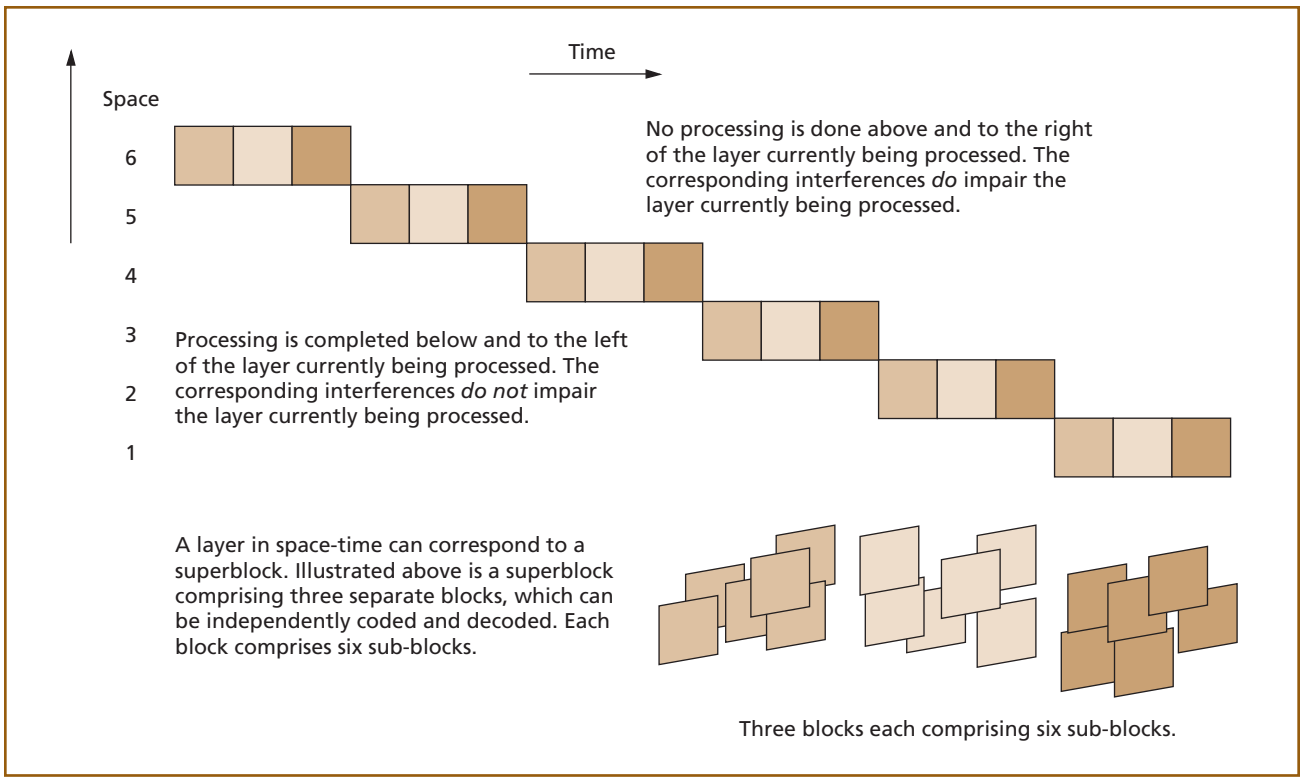

Figure 10.
Blocks and sub-blocks in a space-time layer show how parallel processing can be used to advantage.

already been detected and all interference from the signal components transmitted from antennas labeled 1 to 5 has been subtracted out. Therefore, no interferers are present. Consequently, for the first time interval, a sixfold receive diversity effectively exists. Under such nonexistence of interference, the capacity would be

$$C = \log_2\left[1 + (\rho / 6)\cdot \chi_{12}^{\,2}\right] \quad \text{b/s/Hz.}$$

One interferer is present during the next interval; the other four have been subtracted out. For a system in which this level of interference prevailed forever, the capacity would be

$$C = \log_2\left[1 + (\rho / 6)\cdot \chi_{10}^{\,2}\right] \quad \text{b/s/Hz.}$$

The process of nulling one interferer is what caused the reduction of the chi-squared subscript (giving $\chi_{10}^{\,2}$ in place of $\chi_{12}^{\,2}$). This process is repeated until finally encountering the sixth interval. In this case, all five signals from the other antennas interfere. Therefore, they must be nulled out so that the corresponding capacity would be

$$C = \log_2\left[1 + (\rho / 6)\cdot \chi_2^{\,2}\right] \quad \text{b/s/Hz.}$$

Because each signal radiated by the six transmitting elements multiply a different $H_{\cdot j}$, the six $\chi_{\cdot}^{2}$ variates are statistically independent of each other for the reason given in the previous section. Similar to what was reported elsewhere,[1] for a system cycling among these six conditions with an equal amount of time τ spent on each, the capacity would be

$$C = (1/6)\cdot \sum_{k=1}^{6} \log_2\left[1 + (\rho / 6)\cdot \chi_{2k}^{2}\right] \quad \text{b/s/Hz.}$$

Assume that six such systems are running in parallel with the same realization of χ_{2k}^{2} (k = 1, 2, ... 6) occurring in each one. The capacity would then be six times that given by the previous sixfold sum, or

$$C = \sum_{k=1}^{6} \log_2\left[1 + (\rho / 6)\cdot \chi_{2k}^{2}\right] \quad \text{b/s/Hz.}$$

In the limit of infinitely many symbols in a layer and because every sixth layer is an *a* layer, this last expression gives the capacity of the layered architecture for the (6, 6) case. Obviously then, in the large

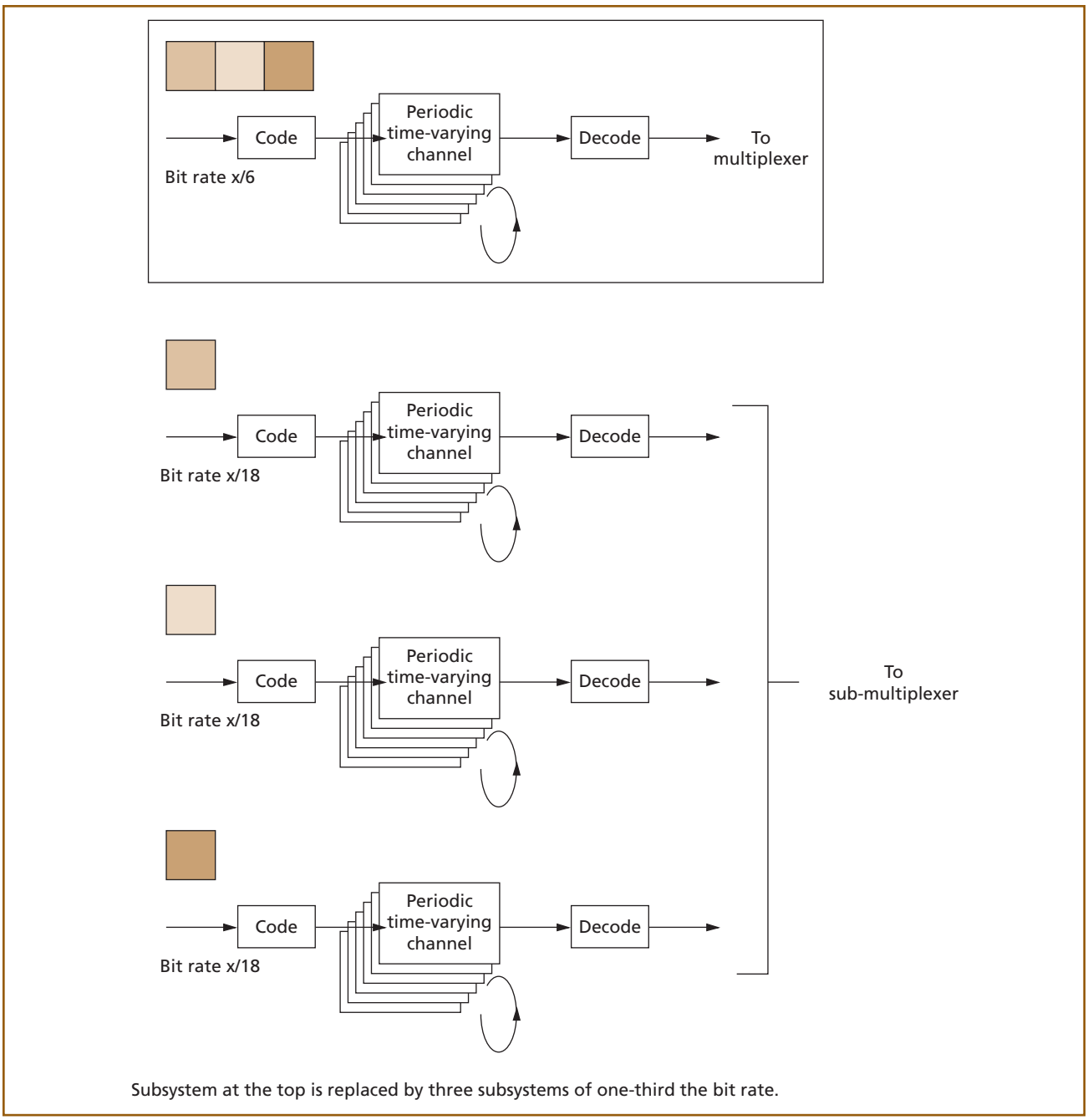

Figure 11.
In parallel receiver processing, three time-varying channels run in parallel.

number of symbols limit, the capacity for an (n, n) system is given by

$$C = \sum_{k=1}^{n} \log_2 \left[1 + (\rho / n) \cdot \chi^2_{2k} \right] \text{ b/s/Hz} \quad . \qquad (6)$$

Figure 7 provides a high-level temporal view of the major steps in the iterative detection of the n layers. In providing a spatial view, **Figure 8** highlights how interferences are handled differently at the distinct vertical levels for the same received vectors.

Figure 9 is a system diagram of the processing involved. The way past and future decisions are han-

dled is reminiscent of zero forcing decision feedback.[7] Factors corresponding to catastrophic error propagation in decision feedback systems are discussed next.

Robustness

In case the layered architecture described earlier seems fragile, an explanation of why it can be quite robust is included. At first, the architecture might *seem* fragile. After all, the successful detection of each layer relies on the successful detection of the underlying layers. Thus, any failure in any layer but the last will likely cause the detection of all subsequent layers to fail. A quantitative discussion is included in this subsection to illustrate that fragility generally is not a significant problem, especially when huge capacity is available. As depicted in Figure 3(a), a huge capacity can be a very reasonable assumption.

In practical implementations, the huge capacity available can be invested in selecting a code that provides the required bit rate with very substantial error protection. Let ERROR denote the event that a packet (= long burst) contains at least one error for whatever reason. Decomposing the ERROR event into two disjoint events gives

$$\mathrm{ERROR} = \mathrm{ERROR}_{\mathrm{nonsupp}} \cup \mathrm{ERROR}_{\mathrm{supp}} .$$

$\mathrm{ERROR}_{\mathrm{nonsupp}}$ denotes the event that channel realization simply does not support the required BER even if receiver processing could be enhanced magically by a genie removing interference entirely from all underlying layers.

$\mathrm{ERROR}_{\mathrm{supp}}$ denotes the remaining ERROR events. Assume that the required outage is 1%, packet size (payload) is 10,000 bits, and a BER of 10^{-7} is required. The extra capacity can be used instead to provide a BER at least one order of magnitude lower. Because $10^{4} \times 10^{-8} = 10^{-4}$, roughly one packet in 10^{4} contains an error. Inflate the bit-error occurrences by labeling all bits in such a packet "in error." Such a drastic inflation in the accounting of errors is a harmless exaggeration. The reason for this is those packets containing errors can be ascribed to outage because they carry insignificant probability compared to Probability[$\mathrm{ERROR}_{\mathrm{nonsupp}}$]. In effect, the huge capacity available allows the luxury of taking the perspective that ERROR = OUT. When the system is not out, essentially error-free transmission is provided.

Despite the robustness just described, providing error-free transmission over a burst for an extremely high fraction of the bursts can erode the bit rate so that codes must be carefully selected for any application.

Related Options

This section discusses some modifications of the communication architecture previously described. Suggestions are provided as to what might be gained or lost by these changes. Some of these items are preliminary ideas that are included as possibilities for future research.

Asymptotic Optimality of the Layered Architecture

The layered space-time approach to communication was based on the premise that the channel was not known at the transmitter. Suppose instead that the channel *is* known at the transmitter and that this knowledge is used to transmit n noninterfering signal components of equal power. Typically, this is done by using the eigenmodes of $HH^{\dagger}$ to derive what amount to n uncoupled subchannels. Other research has been published on how these eigenmodes arise as the natural modes to drive when the channel is known to the transmitter.[8]

For the large n and large ρ asymptote, the capacity benefit of communicating in the way just described is compared below with using the layered space-time architecture. It will be seen later in this paper that in an asymptotic sense, the per-dimension capacity is not improved by knowing the channel at the transmitter.

Under the assumption of a Rayleigh fading channel,[9] research has shown that as n increases, the density of the eigenvalues of $HH^{\dagger}$ approaches

$$\frac{1}{n\pi}\sqrt{\frac{n}{\lambda}-\frac{1}{4}}\,d\lambda \text{ on } 0 \le \lambda \le 4n \quad (0 \text{ elsewhere}).$$

For the purpose of computing the per-dimension capacity, convenience motivates renormalization in the following capacity-invariant ways:

- Channel $H/n^{1/2}$ in place of H,
- Transmit signal power per dimension ρ instead of ρ/n, and
- Noise power per dimension unity.

Previously, an $n^{-1/2}$ multiplier was attached to each scalar transmit signal component to keep the transmitter's total radiated power constant at $\mathcal{P}$ independent of n. This $n^{-1/2}$ multiplier has been moved off the transmit signal components and onto the H_{ij}s. This action fixes the limiting set of eigenvalue support of $HH^{\dagger}/n$ to [0, 4].

Consider that for each n, the matrix representation for each random channel stems from an infinite random matrix with indices ij ($i \geq 1$, $j \geq 1$), where for each n the infinite matrix is projected into its northwest $n \times n$ corner submatrix to obtain the random $n \times n$ channel matrix. For any fixed x on [0, 4] and a corresponding integer $\eta(n)$ on [0, n], it can be written that, with probability one,

$$\begin{aligned} &\operatorname{Lim}_{n\to\infty} \left[\text{Number of Channel Eigenvalues} \leq \eta\right] \\ &\sim \frac{n}{\pi} \cdot \int_0^{\eta} \sqrt{\frac{1}{\xi} - \frac{1}{4}}\, d\xi. \end{aligned} \tag{7}$$

Obviously, the per-dimension capacity C_{ch_knwn}/n is now expressed by

$$\begin{aligned} &\frac{2}{\pi} \int_0^1 \log_2 (1+4\rho x) \cdot \sqrt{\frac{1-x}{x}}\, dx \\ &= \log_2 (1+4\rho) \\ &- \frac{8\rho}{\pi \ln 2} \cdot \int_0^1 \frac{\sqrt{x(1-x)} + \sin^{-1}\sqrt{x}}{1+4\rho x}\, dx. \end{aligned} \tag{8}$$

The right-hand side of (8) follows from integration by parts. In the limit of large ρ, the last integral simplifies, enabling one to conclude that

$$\begin{aligned} C_{ch\_knwn} / n &\to \log_2 (\rho / e) \text{ b/s/Hz/signal} \\ &\text{dimension} \quad (n, \rho \text{ both large}). \end{aligned} \tag{9a}$$

This is the same asymptotic behavior as that for the layered space-time architecture for which the assumption was that the transmitter *does not know* the channel. In the large-ρ large-n asymptote, capacity per dimension is not lost by lack of channel knowledge. Furthermore, in light of (9a), one can now conclude that the equation's right-hand side also expresses the limiting behavior of C/n for the capacity C given by (2).

The large ρ asymptote is anticipated to be of interest in some applications. However, the following is worth mentioning: One can derive that the advantage of knowing the channel for large n but vanishingly small ρ is a factor of two in capacity

$$\begin{aligned} &C_{ch\_knwn} / n \\ &\to \rho / \ln 2 \text{ b/s/Hz/signal dimension} \\ &= 2C / n \ (\text{as } n \to \infty, \rho \to 0). \end{aligned} \tag{9b}$$

The next subsection provides additional information about the small-ρ realm.

The following question, related to the subject of this subsection, is left for possible future research: Does the lack of channel knowledge significantly diminish the per-dimension capacity if x is allowed dependence in the power distribution at the transmitter (optimal ρ(x) in place of constant ρ)? So far, the constraint of equal power out of all n_T transmit elements has been tacitly imposed. It would be worthwhile to explore the aforementioned question while relaxing this restriction even when H is unknown to the transmitter.

Despite the distributional spherical symmetry of the elements of H, the receiver can break from symmetry in the reception process in an H-independent manner that is known to the transmitter. Indeed, the layered space-time reception process involves a symmetry breaking. The example associated with Figure 6 shows that the signal from transmit antenna six is attended to first, then the signal from five, and so on. The capacity advantage that comes from using information on reception asymmetry to distribute power judiciously among the transmit antennas is an area currently being researched.

Slowing Processing With Parallelism

Parallel processing can be used to advantage as shown in **Figure 10**. The example involves slowing processors by a factor of three. (In theory, one could slow processing by any factor). To do so, each of the six streams is further demultiplexed at the transmitter into three demultiplexed substreams. At the receiver and for each of the six streams, three separate processors operate in parallel on the three distinctly encoded substreams. For block codes, each of the three sublayers could constitute separate blocks. In effect, **Figure 11**

stresses that three time-varying channels run in parallel. Even a sublayer could be decomposed into blocks for the purpose of block coding, especially if n is large.

Maximizing SNR Instead of Nulling

Assume that interference from underlying layers is subtracted out. In the previous "Layered Space-Time Architecture" section, the linear combinations formed to avoid remaining interferences were those combinations that null out all the remaining interferers. In the (6, 6) example, zero to five such interferers existed depending on the transmit antenna. The following option can surpass the capacity denoted by the right-hand side of (3): In processing the signal component radiated from a specified transmit array element, proceed as before except choose the linear combination that gives the maximum SNR instead of nulling. Note that here, the meaning of "noise" includes all non-canceled interference along with thermal noise.

The maximum SNR alternative to nulling is reminiscent of minimum mean square error (MMSE) decision feedback[10-13] as an alternative to the zero forcing[7] approach. For the maximum SNR method, capacity could be assessed assuming the code is so advanced that it produces essentially white Gaussian interference signals. From the curves (essentially lines) in Figure 3(a), the capacity improvement over nulling offered by maximizing SNR is marginal for the higher SNR values. Not much improvement can be expected for a ρ of more than about 12 dB. This conclusion is reached even though the curves for maximum SNR are not depicted in Figure 3(a); such curves would occur between the corresponding bold and light lines shown.

Given the marginality and the Gauss-like requirement on the interference, nulling could be preferred over maximizing SNR. For the lower SNR values in Figure 3(a), however—and especially for all the low SNR values of Figure 3(c)—the lower bound has lost its tightness. Thus, maximizing SNR looms as an alternative.

Does the maximum SNR method perform well when ρ is small? While it is beyond the scope of this paper to report performance curves, the maximum SNR method *does* perform well in an important limiting sense. This conclusion is reached after reviewing (2) in the context of the form of the expressions for the n-maximum SNRs. One can establish that for fixed n—as ρ tends toward zero—the capacity constrained to use the maximum SNRs divided by the true capacity converges to one. The next paragraph delineates the two steps required to show this effect.

First, consider each one of the n-maximum SNRs, taking care to note that the so-called noise power in the denominators involves thermal noise plus interference from yet unsubtracted signal components. When expanding each of the n SNRs in a power series in ρ, the interference terms in the denominator do not register in the terms of first-order importance. Second, derive the linear-ρ term in the argument of the logarithm in (2). Some of these terms clearly can be neglected. From this derivation, one learns that the capacity using the maximum SNR on each subchannel is precisely the same as the capacity given by (2) when contributions of only first-order importance are retained. This derivation for the small-ρ realm results in a capacity tending toward the total capacity of n parallel systems with n-fold optimal combining. Namely, as ρ tends to zero,

$$\begin{aligned} C &\to \sum_{k=1}^{n} \log_2 \left[1+(\rho/n)\cdot H_{\cdot k} \bullet H_{\cdot k}\right] \text{ b/s/Hz} \\ &= \sum_{k=1}^{n} \log_2 \left[1+(\rho/n)\cdot \chi^2_{2n,k}\right] \text{ b/s/Hz}. \end{aligned} \tag{10}$$

The dot between vectors is the complex n-D scalar product. As before, the extra k subscript on $\chi^2_{2n,k}$ is employed to index over independent chi-squared variates. Because the limit of small ρ has been carefully analyzed insofar as retaining terms that are linear in ρ, one can write

$$C \to (\ln 2)^{-1} \ (\rho/n) \ \text{trace}\left[HH^{\dagger}\right] \text{ b/s/Hz} \quad (\text{as } \rho \to 0) \ . \tag{11}$$

"Trace" in this context means the sum of the diagonal elements.

Coding

We previously stressed block coding for expositional simplicity. In practice, convolutional codes or a form of trellis-coded modulation for more bandwidth efficiency might have a role. (The block-convolutional distinction is blurred if a block code is

formed by blocking off a convolutional code.)

Comments about parallelism are equally true for both convolutional and block codes. With convolutional codes, adjoining layers can be decoded simultaneously at times as long as a decision depth requirement is met. The depth constraint stipulates that the detection processes of the different processors be staggered in such a way to ensure that each layer is decoded after the interferences below them have been detected and subtracted off. Such subtraction is essential for approaching the high capacities expressed by (2) and (3).

A preliminary idea related to convolutional coding is mentioned for addressing the nonstandard context in which the AWGN variance is a periodically changing value. For illustrative purposes, assume the existence of five symbols per sub-block and elaborate the sequence of transmit antenna elements used for any of the three sub-blocks over time. The $6 \times 5 = 30$ consecutive time intervals result in 666665555544444333332222211111. A 6 symbol (meaning a symbol transmitted from antenna 6) tends to need the least error protection (no interferences). A 5 symbol tends to need more protection (one interference)—and so on down to a 1 symbol, which tends to need the most protection (five interferences). "Tend" is used because noise, interference level, and channel realization are all random variables.

In decoding convolutional codes, one could pair protector and protected symbols in a more sensible way if doing so significantly expedites bit decisions. For example, 61616161612525252525434343434343 would be an improvement. The encoding (decoding) to accomplish this involves nothing more than a straightforward permutation (inverse permutation) in the encoding (decoding) process. Careful study is needed to quantify the resulting benefit of this idea for promoting timeliness in making bit decisions.

The actual choice of codes remains an important open issue, one that is best addressed in the context of specific applications, However, it is worth mentioning that a transformation of the architecture that renders the coding context much more standard does exist—albeit at a price of about one-half the available capacity.

In explaining the transformation, n is assumed to be even; n odd is a trivial extension. At the transmitter, the primitive bit stream is demultiplexed to n/2 streams rather than n streams. However, n transmit (and receive) antennas are still used. During each interval of duration τ, each of the n/2 demultiplexed signals is now associated with a distinct pair of transmit antennas. The same coded and modulated signal is transmitted out of the array elements in each pair but at different times. The pairing is as follows: the best with the worst symbols, the next to best with the next to worst, and so on. (To be precise, one should not simply say "best" and "worst." "Tending to be the best" and "tending to be the worst" are more definitive because of the *random* power-transfer characteristics.)

The motivation for pairing transmit antennas is to have each demultiplexed signal component possess the same optimum combining diversity level. Say, for explanatory purposes, n = 6. As far as time of transmission is concerned, refer also to the (6, 6) example in Figure 6. A signal transmitted out of antenna 6 during $[\tau, 2\tau)$ is the same as that transmitted out of antenna 1 during $[6\tau, 7\tau)$. Antenna pairs 5 and 2 and 4 and 3 are configured the same way, the latter pair being associated with the contiguous time intervals, the union of which is $[3\tau, 5\tau)$.

Evidently, with this pairing, n/2 optimal combiners exist in effect for arbitrary n, each of which has n+1-fold diversity. Therefore, subject to the constraint to communicate in this way, the formula for capacity differs from (6). Specifically, instead of chi-squared variates having the arithmetically progressing indices 2, 4, ..., 2n, all n/2 indices are 2n+2 so that the capacity is given by

$$C = \sum_{k=1}^{n/2} \log_2 \left[1 + (\rho / n) \cdot \chi^2_{2(n+1),k} \right] \text{ b/s/Hz} \quad . \quad (12)$$

The subscript k indexes statistically independent χ^2_{2n+2} variates. Thus, a good part of the cyclic volatility of the received signal SNR has been removed. Clearly, as n increases, the arguments of the n/2 base two logarithms in (12) converge to $1+\rho$. The price in lost capacity for the large n asymptote is easy to see. That is, the capacity now increases linearly with only

an n/2 slope rather than an n slope. Considerable capacity, however, is still possible.

No Cycling

As Figure 4 illustrates, cycling the substream to antenna association was required at the transmitter. Is this cycling really necessary? A straightforward but tedious asymptotic argument shows that, in the limit of large n, the receive diversity compensates for any inferior H_{ij}s. Consequently, the asymptotic linear capacity growth with n also occurs even without cycling.

Discussion and Conclusion

With growing multiuser applications, efficient use of spectral resources is especially important to avoid highly contentious channel demand in a limited frequency band. Bit-rate delivery issues can be difficult to decide solely from a fundamental standpoint. Indeed, determination of implementation complexity can be influenced by the legacy of past technology choices limiting what is readily available for the short term. Nonetheless, the results discussed in this paper inform the evolution toward meeting future demand for greater bit rates.

When the transmit volume is sufficient to allow driving transmit antenna elements separated by one-half a wavelength, the results presented in this paper suggest considering doing so. When the receive volume (also assumed to be amply sized) is radiated by waves involving distinct spatial degrees of freedom—all in the same frequency band—receive antenna elements with half-wavelength spacing can serve to capture that energy. Thus, transmit/receive volume can be used to improve capacity dramatically over that of systems in which the spatial dimension is not exploited.

The layered space-time architecture is designed largely to undo the coupling between distinct spatial modes, yielding a system in which capacity increases linearly with n for both fixed bandwidth and fixed total radiated power. This n-D architecture can be viewed in terms of n 1-D architectures of equal capacities. In future theoretical studies, a comparison of extreme approaches would be informative—for example, a narrowband system using MEAs at *both* the transmitter and receiver with wideband alternatives to using MEAs to meet required capacity demands. Understanding the relative merits of extreme approaches could help clarify how the spatial and frequency domains should be combined to provide channels in various applications.

When adding more receiver elements than n to an (n, n) system, the excess $n_R - n$ can be used to improve performance simply by adding twice the excess to the degrees of freedom of the chi-squared variates appearing on the right-hand side of (6). If other users of the same frequency spectrum are identified, the excess receiver elements are effective for reducing co-channel interference. The results of research have been published concerning handling co-channel interferers in a Rayleigh fading environment.[14]

Acknowledgments

Figures 1 and 2 were largely the creation of M. J. Gans, who also advised the author on antenna theory. Valuable discussions with I. Bar-David, J. Mazo, A. Saleh, J. Salz, and L-F. Wei are also gratefully acknowledged.

References

1. G. J. Foschini and M. J. Gans, "On Limits of Wireless Communication in a Fading Environment When Using Multiple Antennas," *Wireless Personal Communications*, accepted for publication.
2. E. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.
3. S. Z. Stambler, "Shannon's Theorems for a Complete Class of Discrete Channels Whose State Is Known at the Output," *Problems of Information Transmission*, No. 11, Plenum Press, New York, November 1976, pp. 263-270.
4. S. Kullback, *Information Theory and Statistics*, John Wiley and Sons, New York, 1959.
5. D. B. Osteyee and I. J. Good, *Information Weight of Evidence: the Singularity Between Probability Measures and Signal Detection*, Springer-Verlag, New York, 1970.
6. M. S. Pinsker, *Information and Information Stability of Random Processes*, Holden Bay, San Francisco, 1964, Chapter 10.
7. R. Price, "Nonlinearly Feedback-Equalized PAM vs. Capacity for Noisy Filter Channels," *Proceedings of the IEEE International Conference on Communications*, June 1972, IEEE Publishing

Company, New York, Sessions 22-12 to 22-17.
8. G. J. Foschini and R. K. Mueller, "The Capacity of Linear Channels with Additive Gaussian Noise," *Bell System Technical Journal*, January 1970, pp. 81-94.
9. V. Grenander and J. W. Silverstein, "Spectral Analysis of Networks with Random Topologies", *SIAM Journal of Applied Mathematics*, Vol. 32, No. 2, March 1997, pp. 499-519.
10. C. A. Belfiore and J. H. Park Jr., "Decision Feedback Equalization," *Proceedings of the IEEE*, 67(8):1143-1156, August 1979.
11. J. Salz, "Optimum Mean-Square Decision Feedback Equalization," *Bell System Technical Journal*, October 1973, pp. 1341-1372.
12. D. D. Falconer and G. J. Foschini, "Theory of Minimum Mean-Square-Error QAM Systems Employing Decision Feedback Equalization," *Bell System Technical Journal*, December 1973, pp. 1821-1849.
13. R. D. Gitlin, J. F. Hayes, and S. Weinstein, *Data Communication Principles*, Plenum Press, New York, 1992, Chapters 5 and 7.
14. R. D. Gitlin, J. Salz, and J. H. Winters, "The Impact of Antenna Diversity on the Capacity of Wireless Communication Systems," *IEEE Transactions on Communications*, Vol. 42, No. 4, April 1994, pp. 1740-1751.

(Manuscript approved October 1996)

GERARD J. FOSCHINI is a distinguished member of technical staff in the Wireless Communications Research Department at Bell Labs in Holmdel, New Jersey. He is conducting communication and information theory investigations for wireless communications at both the point-to-point and systems levels. Mr. Foschini is an IEEE Fellow and has taught at three New Jersey universities. He holds a B.S.E.E. from the New Jersey Institute of Technology in Newark, an M.E.E. from New York University, and a Ph.D. in mathematics from the Stevens Institute of Technology in Hoboken, New Jersey. ◆