Collision Resolution in Packet Radio Networks Using Rotational Invariance Techniques

Ruifeng Zhang, *Member, IEEE*, Nicholas D. Sidiropoulos, *Senior Member, IEEE*, and Michail K. Tsatsanis, *Member, IEEE*

***Abstract*—This paper deals with the multiuser medium access problem in the packet radio environment. Under the framework of network diversity multiple access (NDMA), a recently proposed medium access method, a blind collision resolution scheme is proposed employing rotational invariance and factor analysis techniques. The proposed approach (dubbed B-NDMA for Blind NDMA) overcomes the difficulty of orthogonal identification codes required by the original protocol, thereby improving channel utilization and system capacity, while being insensitive to multipath effects and synchronization errors. Performance issues of the proposed technique are addressed both analytically and numerically.**

***Index Terms*—Access protocols, blind signal separation, matrix decomposition, packet radio, rotational invariance.**

I. Introduction

In a wireless cellular setup, the transmission medium is inherently broadcast, and provides no protection from interference among the wireless users. Hence the medium access problem, that is, the multiplexing of multiple wireless users in the same cell is far from trivial. Various solutions to this resource allocation problem have been proposed, depending on the traffic characteristics of the users.

A great majority of the existing wireless cellular infrastructure supports voice services or other constant-bit-rate (CBR) connections. Under these idealized traffic conditions, simple solutions like FDMA or TDMA schemes can isolate different transmissions and transform the problem into multiple single-user communication setups (e.g., [4]). However, even when only voice users are concerned, it appears that multi-user solutions (CDMA) coupled with advanced signal processing at the receiver can provide added flexibility and improved capacity, and are preferred for future wireless networks [18].

The medium access problem becomes substantially more challenging when data users or variable-bit-rate (VBR) sources are considered. With the advent of mobile computing, video conferencing, web browsing, etc., it is expected that VBR services will become more and more important in the future, and will require efficient medium access strategies. Unfortunately, FDMA and TDMA are known to be very inefficient for bursty sources [9]. While CDMA is somewhat better due to its relatively graceful degradation under increasingly heavier load, it still represents a fixed resource allocation scheme that is inefficient when dealing with VBR traffic.

Relatively simple medium access strategies for bursty users include random access protocols of the ALOHA type, e.g., [1], wherein collided packets are discarded and later retransmitted. Although better suited to handle VBR traffic than FDMA/TDMA, ALOHA schemes suffer from substantial throughput penalties and underutilization of the channel under relatively heavy loads. Improvements like carrier sensing (CSMA) and collision detection (CD) are helpful but cannot be reliably implemented in a wireless environment as they require a "listen-while-talk" feature which is not possible in wireless transmissions. Instead, data sensing (DSMA) is usually implemented, wherein the base station broadcasts the channel status on a separate down-link control channel [10].

Notice that in most random access protocols little signal processing is involved and no attempt is made to separate the colliding packets. In the communications and signal processing literature, there has recently been intense research activity in user separation in the context of CDMA multiuser detection [18], and also in Space-Division Multiple-Access (SDMA) wireless networks, cf. [8] and references therein. Either way, collision resolution through multiuser detection/spatial source separation requires *extra diversity*, either in the form of spreading codes [18] with which the individual symbols are spread, or powerful error control codes (in the spirit of GMAC [3]) with which the user packets are coded, or in the form of antenna arrays. Spreading is bandwidth-consuming, and so is GMAC-type coding, which also requires complex decoding. Antenna arrays can only help resolve up to a relatively small fixed number of colliding packets ($\leq$ number of elements), and require relatively expensive hardware. It therefore makes sense to investigate whether one can rely on network resources alone to carry out the collision resolution burden, possibly aided by, but not counting on other forms of diversity.

In [7], a novel random access protocol was proposed and analyzed. The protocol, named Network Diversity Multiple Access (NDMA), exploits network resources to provide diversity and allow the receiver to separate the colliding users. NDMA is a slotted random access protocol that builds on the following idea:

Paper approved by A. K. Singh, the Editor for Packet Radio Access of the IEEE Communications Society. Manuscript received September 23, 1999; revised June 14, 2001. The work of N. D. Sidiropoulos was supported by the National Science Foundation under NSF/Wireless CCR-0096164 (9979295) and NSF/CAREER under CCR-0096165 (9733540). The work of M. K. Tsatsanis was supported by the National Science Foundation under NSF/NCR-9706658, NSF/CAREER CCR-9733048 and NSF/Wireless CCR-9979288.

R. Zhang is with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19014 USA.

N. D. Sidiropoulos is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA.

M. K. Tsatsanis was with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030 USA. He is now with Voyan Technology, Santa Clara, CA 95054 USA.

Publisher Item Identifier S 0090-6778(02)00513-5.

0090–6778/02$17.00 © 2002 IEEE

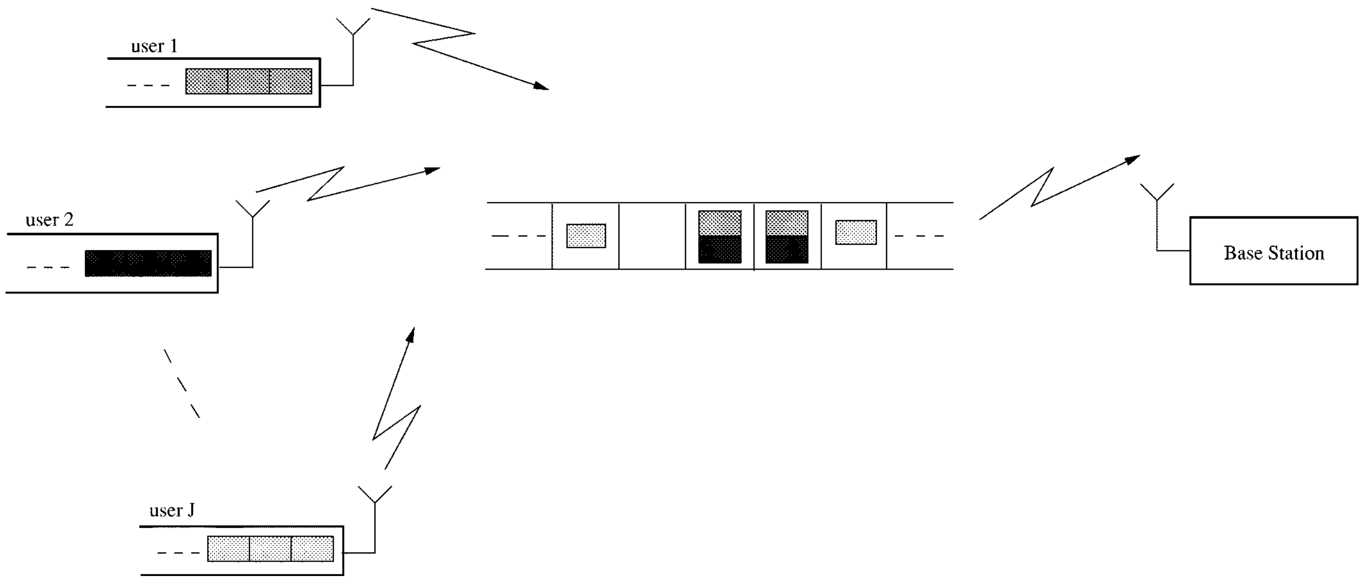


Fig. 1. Slotted random access cellular radio system.

if K users collide, the base station (BS) will instruct them to transmit a total of K times, resulting in K different versions of the original superposition of the K user signals. Then a source separation algorithm is invoked to resolve the collision. The method only requires K channel slots to transmit a total of K user packets and therefore has excellent throughput characteristics.

The NDMA method in [17] introduced a signal processing viewpoint to the collision resolution problem. Its critical assumption is that the base station is able to reliably detect how many users have collided based on the first reception of the collision signal. Given this information, intelligent retransmission strategies could be initiated by the base station. The detection of the active users is based on training sequences assumed to be embedded in each user's packet. In order to make the detection problem tractable, those training or identification (ID) sequences are assumed orthogonal to each other.

The orthogonality assumption for the ID sequences is responsible for a major drawback of the method of [17]. The length of the user ID sequence is required to grow linearly in the number of users and not logarithmically, as is usually the case with addressing schemes. Therefore, for a large user population (active + inactive users), the induced overhead may be substantial. Furthermore, the orthogonality assumption is sensitive to multipath effects and lack of synchronization of the colliding users [20].

In order to overcome these difficulties, in this paper we explore the applicability of blind signal separation methods to the collision resolution problem. Our goal is to avoid the use of training ID sequences in the NDMA approach. We modify the protocol, and control each packet's initial phase in such a way that a structured mixing matrix is constructed through the retransmissions. The signature of each colliding user has a Vandermonde form and can be blindly estimated through rotational invariance techniques.

The proposed method results in substantial computational savings since the complexity of the method is proportional to the number of colliding users, which is typically a small number. In contrast, the method of [17] has a complexity that is proportional to the total number of users in the system, irrespective of how many of them have collided. Furthermore, the proposed method is insensitive to multipath effects and synchronization errors among the users.

The rest of this paper is organized as follows: a brief description of existing NDMA methods and associated drawbacks are presented in Section II. A new blind collision resolution method using rotational invariance and factor analysis techniques is discussed in Section III, while its performance is studied in Section IV. Section V presents simulation results that corroborate our performance analysis, and Section VI summarizes our findings.

II. Problem Statement

The NDMA approach is meant to be implemented in slotted cellular packet radio systems, as illustrated in Fig. 1. In each slot a user may send a fixed-length data packet composed of N QAM symbols, provided that it is allowed to transmit and its data buffer is nonempty. Since there is no coordination among the users, two or more users may collide in a channel slot. In traditional random access schemes the collided data are customarily discarded, and later retransmitted under the control of upper layer protocols. The channel slots in which collisions occur are therefore wasted. The novel idea of NDMA originated from the observation that the collided data do contain useful information; exploiting this information enables increased channel utilization.

Assume that the channel between every user and the BS is slowly (compared to the slot duration) frequency-flat fading, and that the users are perfectly synchronized with the slot timing. Then the baseband discrete-time symbol-rate signal received at the BS in a channel slot will be the superposition of signals from all users transmitting in the said slot. It can be written in vector form as

$$\mathbf{y}(l) = \sum_{i \in \mathcal{I}(l)} \alpha_i(l)\mathbf{s}_i(l) + \mathbf{v}(l), \tag{1}$$

where $\mathbf{s}_i(l)$ is an $N \times 1$ vector representing the packet sent by user i in slot l; $\mathbf{y}(l)$ and $\mathbf{v}(l)$ represent the received signal vector and noise vector in slot l, respectively; and $\alpha_i(l)$ is a random initial phase[1]. It is assumed that $\alpha_i(l)$ is constant within a slot and i.i.d. in i and l, $\mathbf{v}(l)$ is white Gaussian noise (WGN), and both are independent of the user signals $\mathbf{s}_i(l)$. The indices of all users colliding in slot l are collected in the set $\mathcal{I}(l)$.

Note that the signal model described in (1) constitutes a classic signal separation problem, where the colliding packets are the sources that need to be separated. When there are K users colliding, $\mathcal{I}(l) = \{i_1, \ldots, i_K\}$, the separation of the K users' packets can be accomplished if K branches of diversity of the mixture signal are available (using multiple antennas, for example). The novelty of the NDMA approach lies in that it builds the diversity through retransmissions dictated by the protocol, without requiring multiple antennas or excess bandwidth. The procedure involves the following operations:

If the BS finds K users colliding in a slot, it will instruct those colliding users to retransmit their packets $K-1$ more times in the next $K-1$ slots; meanwhile, other users are not allowed to transmit until this retransmission procedure is completed[2].

Finally, the BS receives a total of K different versions of the original superposition of the K colliding signals, which can be arranged in a matrix form as (2)

$$\begin{bmatrix} \mathbf{y}^T(l) \\ \vdots \\ \mathbf{y}^T(l+K-1) \end{bmatrix} = \begin{bmatrix} \alpha_{i_1}(l) & \cdots & \alpha_{i_K}(l) \\ \vdots & \ddots & \vdots \\ \alpha_{i_1}(l+K-1) & \cdots & \alpha_{i_K}(l+K-1) \end{bmatrix} \cdot \begin{bmatrix} \mathbf{s}_{i_1}^T(l) \\ \vdots \\ \mathbf{s}_{i_K}^T(l) \end{bmatrix} + \begin{bmatrix} \mathbf{v}^T(l) \\ \vdots \\ \mathbf{v}^T(l+K-1) \end{bmatrix}, \quad (2)$$

or equivalently,

$$\mathbf{Y}_{K\times N} = \mathbf{A}_{K\times K}\mathbf{S}_{K\times N} + \mathbf{V}_{K\times N} \quad (3)$$

with obvious definitions for the matrices in (3). Now, it becomes a relatively easy task to resolve the user packets from the collision. A simple linear solution, for example, is $\widehat{\mathbf{S}} = \mathbf{A}^{-1}\mathbf{Y}$, provided that $\mathbf{A}$ has full rank, and an estimate of it is available. The full-rank property of $\mathbf{A}$ is guaranteed w.p. 1 under the i.i.d. assumption for the coefficients $\alpha_i(l)$. More involved approaches (e.g., maximum likelihood and successive cancellation) are also applicable. Note that only K slots are required to resolve K colliding users. No slots are wasted.

The model of (1) can also incorporate the effects of multipath fading, if we consider $\mathbf{s}_i(l)$ as the colored version of the original user data. As long as the multipath channel is slowly varying (in particular, is time invariant for the duration of the collision resolution period), the colored user data are the same in each (re)transmission. After the separation of the colliding users, single user equalizers, either training-based or blind, can be employed for each user packet to extract the original user data. A detailed discussion on this topic can be found in [20].

The NDMA scheme requires that the BS be able to determine the collision multiplicity K. The original NDMA technique in [17] uses training sequences to solve this problems. Specifically, a unique ID sequence is assigned to each user, and is used as a prefix for all packets originating from this user. Then the BS can identify the active users by checking the ID's. In order to simplify the problem, the ID sequences are designed to be mutually orthogonal. Since the ID part of the user data obeys the same I/O relation as (1), a matched filterbank (one filter matched to each user ID) will be the optimal detector.

The orthogonal ID requirement simplifies detection, but also brings a major drawback to the NDMA method. The length of the ID sequence is required to grow linearly in the number of the users and not logarithmically, as is usually the case with addressing schemes. Therefore, for a large user population, the induced overhead may be substantial. Furthermore, the orthogonality assumption is sensitive to multipath effects and lack of synchronization of the colliding users. Special joint design of ID sequences and receive filters that can combat dispersive effects of the channel are possible [20], but at the cost of even longer IDs.

From the viewpoint of signal separation, the NDMA model has many analogies with the popular array signal processing model, where each (re)transmission acts like a "virtual sensor". We will explore this link in the sequel to develop a blind collision detection and resolution NDMA scheme, called B-NDMA for *Blind* NDMA[3], using rotational invariance and related factor analysis tools.

III. B-NDMA

A. Retransmission Protocol for B-NDMA

In the model of (3), the data matrix $\mathbf{Y}$ can be viewed as the response of an array of K virtual antennas to the K colliding data packets, where each row of $\mathbf{Y}$ corresponds to one antenna output. The mixing matrix $\mathbf{A}$ is like the steering matrix of the array. For a uniform linear array (ULA) the steering matrix exhibits special (Vandermonde) structure, which facilitates signal separation. In order to force the mixing matrix $\mathbf{A}$ to have the desired Vandermonde structure, we can modify the NDMA retransmission protocol and control each packet's initial phase $\alpha_i(l)$, instead of leaving it random. Specifically:

Upon collision and retransmission request by the BS, every user who transmitted in the previous slot will transmit again for as many times as instructed by the BS, but each time multiplying the whole packet by z_i^r, where i is a user index, z_i is a complex

[1]The random phase changes from slot to slot due to an arbitrary initial transmitter phase at the beginning of each transmitted packet.

[2]The required central control from the BS can be implemented through a reverse control channel (for details, see [17]).

[3]The *blind* here means that there is no training needed for collision detection and resolution purposes. Training may still exist in the user packets, e.g., to allow training-based single-user channel equalization.

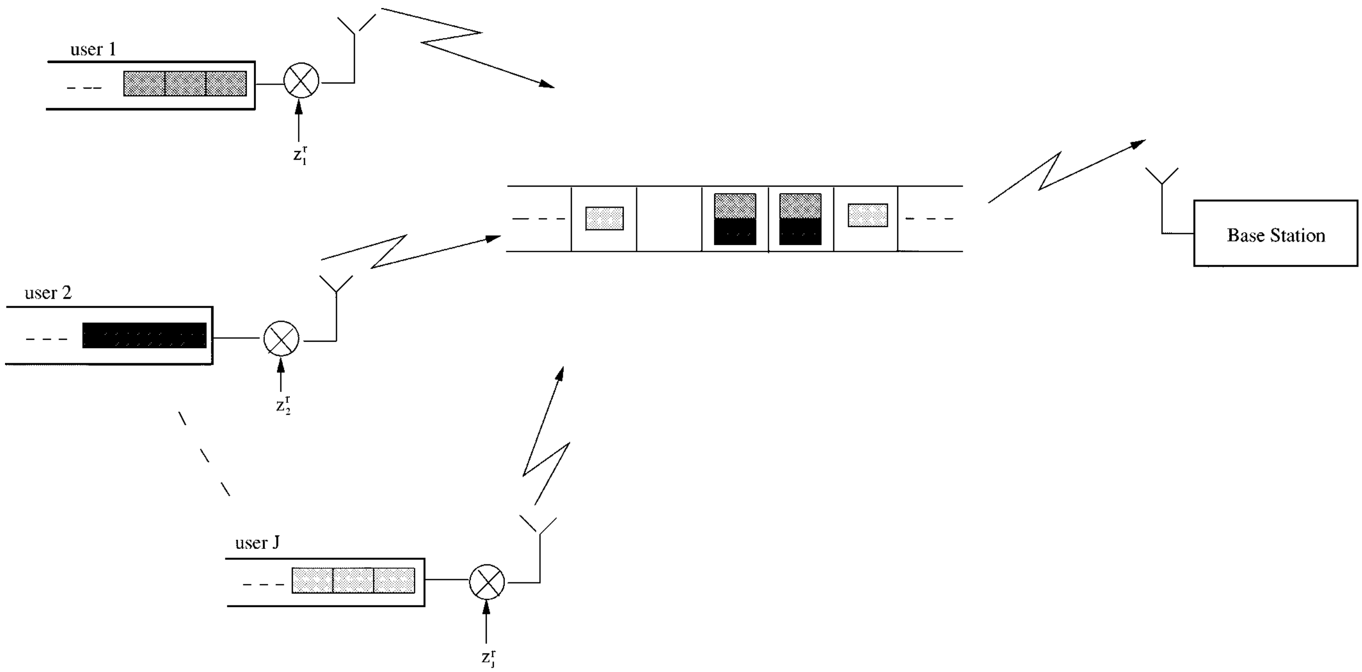


Fig. 2. B-NDMA retransmission scheme.

number[4] *and r stands for the rth retransmission of the given packet.*

This setup is illustrated in Fig. 2 (cf. Fig. 1 for comparison).

Now suppose that a total of R transmissions take place for a K-fold collision (R is not necessarily equal to K). Similar to (3), we have the received signal matrix

$$\mathbf{Y}_{R\times N} = \mathbf{Z}_{R\times K}\mathbf{S}_{K\times N} + \mathbf{V}_{R\times N} \tag{4}$$

where $\mathbf{Y}$, $\mathbf{S}$ and $\mathbf{V}$ are received signal matrix, user signal matrix and observation noise matrix, respectively; and $\mathbf{Z}$ is a Vandermonde matrix[5]

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ z_{i_1} & z_{i_2} & \dots & z_{i_K} \\ \vdots & \vdots & \ddots & \vdots \\ z_{i_1}^{R-1} & z_{i_2}^{R-1} & \dots & z_{i_K}^{R-1} \end{bmatrix}$$

where $i_1, \dots, i_K$ are the indices of the colliding users. We assume the channels to be slowly fading compared to the collision resolution period. Then frequency-flat fading scales the rows of $\mathbf{S}$ (multiplying each with a complex constant), while frequency-selective fading colors the rows of $\mathbf{S}$ (convolving each with a FIR channel). Since our focus here is on user separation and not on the equalization of each individual transmission, we absorb the channel effects in the rows of $\mathbf{S}$ in the model of (4). Hence, in the sequel, the rows of $\mathbf{S}$ will represent scaled or possibly colored versions of the transmitted packets. Except for the possibility of a deep fade, this has no effect on the ensuing discussion.

[4]The z_i's can be pre-assigned by the BS, or drawn at random on-line by each user, in a user-wise independent fashion. For large user populations the choice is immaterial regarding collision resolution *perse*, but random assignment requires less system setup/administration overhead.

[5]We view the initial transmission as the zeroth retransmission, corresponding to initial phase $z_i^0 = 1$.

An important point regarding (4) is that the identity of the colliding users is unknown, hence the resulting steering matrix $\mathbf{Z}$ is unknown, even if the z_i's are pre-assigned. Therefore, blind techniques are needed for determining the number of retransmissions required (equivalent to determining the collision multiplicity K), and resolving the user data $\mathbf{S}$. In Section IV, we first develop our blind collision resolution method, assuming K is known; we then return to the issue of determining K in Section III-D.

B. An ESPRIT-Type Method for Blind Collision Resolution

Consider the noiseless case of (4)

$$\mathbf{X}_{R\times N} = \mathbf{Z}_{R\times K}\mathbf{S}_{K\times N} \tag{5}$$

where $\mathbf{X}$ is the noiseless data. Suppose for the moment that the number of colliding users K is known, and the BS requests more retransmissions than the collision multiplicity, i.e., $R > K$. Define $(R-1)\times N$ matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, consisting of the first (respectively last) $R-1$ rows of $\mathbf{X}$. Due to the Vandermonde structure of $\mathbf{Z}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ admit the following factorization:

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{A}_{(R-1)\times K}\mathbf{S}_{K\times N} \\ \mathbf{X}_2 &= \mathbf{A}_{(R-1)\times K}\mathbf{D}_{K\times K}\mathbf{S}_{K\times N} \end{aligned} \tag{6}$$

where $\mathbf{A}$ is a matrix consisting of the first $R-1$ rows of $\mathbf{Z}$, (different from $\mathbf{A}$ in (3)), and $\mathbf{D} = \mathrm{diag}([z_{i_1}, \dots, z_{i_K}])$. Under certain conditions, this model is unique, i.e., $\mathbf{A}$, $\mathbf{D}$, and $\mathbf{S}$ can be found from $\mathbf{X}_1$, $\mathbf{X}_2$. The idea is similar in spirit to data-domain ESPRIT [11] (but also cf. [12]). In particular, note that $\mathbf{A}$ is square/tall and full column rank (true w.p. 1. under our retransmission scheme and the Vandermonde structure), $\mathbf{S}$ is fat and full rank (persistence of excitation—note that the packet length N is much larger than the number of colliding users), and also that $\mathbf{D}$ contains distinct nonzero elements (true w.p. 1 under the assignment scheme for the z_i's).

Introduce a singular value decomposition of the stacked data matrix

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{AD} \end{bmatrix} \mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H \tag{7}$$

where $\mathbf{U}$ is $2(R-1) \times K$ holding the left singular vectors corresponding to the K nonzero singular values; $\boldsymbol{\Sigma}$ is $K \times K$ diagonal holding the said singular values; $\mathbf{V}$ is $N \times K$, holding the corresponding right singular vectors, and $(\cdot)^H$ stands for Hermitian (complex conjugate) transpose. $\mathbf{S}$ being full row rank assures that $\mathrm{span}(\mathbf{U}) = \mathrm{span}\left(\begin{bmatrix} \mathbf{A} \\ \mathbf{AD} \end{bmatrix}\right)$; hence there exists a nonsingular matrix $\mathbf{P}$ such that

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{AD} \end{bmatrix} \mathbf{P}. \tag{8}$$

Next, construct the auto- and cross-product matrices

$$\begin{aligned} \mathbf{R}_0 &= \mathbf{U}_1^H \mathbf{U}_1 = \mathbf{P}^H \mathbf{A}^H \mathbf{A} \mathbf{P} = \mathbf{QP}, \\ \mathbf{R}_1 &= \mathbf{U}_1^H \mathbf{U}_2 = \mathbf{P}^H \mathbf{A}^H \mathbf{A} \mathbf{D} \mathbf{P} = \mathbf{QDP}. \end{aligned} \tag{9}$$

Note that all matrices in (9) are square, full-rank. It then follows that

$$\left(\mathbf{R}_0^{-1}\mathbf{R}_1\right)\mathbf{P}^{-1} = \mathbf{P}^{-1}\mathbf{D} \tag{10}$$

i.e., a standard eigenvalue problem. $\mathbf{P}^{-1}$ can therefore be determined, up to permutation and scaling of columns. After that, $\mathbf{A}$ can be recovered from (8) as $\mathbf{A} = \mathbf{U}_1\mathbf{P}^{-1}$, and finally $\mathbf{S}$ can be recovered from (6) as[6] $\mathbf{S} = \mathbf{A}^\dagger \mathbf{X}_1$, both under the same permutation and scaling of columns, which carries over from the solution of the eigenvalue problem in (10).

The ESPRIT algorithm only needs the number but not the identities of the colliding users. Usually there are address strings embedded in the data packets. After recovering the user data packets, the users' identities can be deduced from the address strings. It follows that data-embedded ID's need only be as long as $\log J$, where J is the user population; no orthogonality requirements (wasteful in terms of bandwidth, especially if J is large) are imposed on the ID strings. In case the complex exponential generators z_i are pre-assigned (by the BS), the users' identities can also be deduced from the Vandermonde matrix $\mathbf{A}$.

If noise is present in the data ($\mathbf{Y} = \mathbf{X} + \mathbf{V}$ instead of $\mathbf{X}$ is available), then TLS-ESPRIT can be used [11]. Alternatively, an alternating least squares (ALS) refinement algorithm can be employed. This is explained next.

C. PARAFAC and Trilinear ALS Refinement

Consider the model in (6), and let $x_{r,n,i}$ stand for the (r,n)-element of $\mathbf{X}_i$, $i = 1, 2$. Then

$$x_{r,n,i} = \sum_{k=1}^{K} a_{r,k} d_{i,k} s_{k,n} \tag{11}$$

for all $r = 1, \ldots, R-1$, $n = 1, \ldots, N$, $i = 1, 2$, where $d_{1,k} = 1, \forall k$, and otherwise obvious notation. Equation (11) is a K-component trilinear decomposition of the three-way $(R-1) \times N \times 2$ array with typical element $x_{r,n,i}$. PARAllel FACtor (PARAFAC) analysis is a common name for the trilinear decomposition. The identifiability of the model in (6) (and therefore (5) as well) is actually a corollary of uniqueness of low-rank decomposition of three-way arrays [6], [7]; see also [2], [13]–[15] for a compact generalization to the complex case.

Now consider the noisy case, and assume that initial estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{S}}$ are available from eigenanalysis. One possible matrix view of the data in (11) is

$$\begin{aligned} \mathbf{X}_{N(R-1)\times 2} &= \begin{bmatrix} x_{1,1,1} \ldots x_{1,N,1} \ldots x_{R-1,1,1} \ldots x_{R-1,N,1} \\ x_{1,1,2} \ldots x_{1,N,2} \ldots x_{R-1,1,2} \ldots x_{R-1,N,2} \end{bmatrix}^T \\ &= \left(\mathbf{A} \odot \mathbf{S}^T\right) \boldsymbol{\Delta}^T \end{aligned} \tag{12}$$

where $\boldsymbol{\Delta}$ is a $2 \times K$ matrix holding the $d_{i,k}$'s, and $\odot$ stands for the Khatri-Rao (column-wise Kronecker) product. It follows that for noisy data $\tilde{x}_{r,n,i}$ giving rise to $\tilde{\mathbf{X}}_{N(R-1)\times 2}$, the conditional least squares estimate of $\boldsymbol{\Delta}$ given $\hat{\mathbf{A}}$ and $\hat{\mathbf{S}}$ is given by $\hat{\boldsymbol{\Delta}}^T = \left(\hat{\mathbf{A}} \odot \hat{\mathbf{S}}^T\right)^\dagger \tilde{\mathbf{X}}_{N(R-1)\times 2}$. Resorting to the symmetry of the model in (11), similar LS updates can be worked out for $\mathbf{A}$ and $\mathbf{S}$. The resulting round-robin updating algorithm is known in the PARAFAC literature as PARAFAC-ALS[7] or Trilinear ALS (TALS) and it is monotonically convergent (details can be found, e.g., in [14]). Due to the data smoothing step in going from (5) to (6), the noise in $\tilde{x}_{r,n,i}$ is actually correlated, hence weighted LS is more appropriate. Weighting can be incorporated in TALS; we skip the details for brevity, noting that the weighting gain is small at relatively high SNR.

D. Rank Test Method for Blind Collision Detection

We now return to the problem of detecting the collision multiplicity, K. This problem is crucial for the BS to properly schedule the retransmissions to guarantee the resolvability of the colliding users' packets. In addition the BS has to determine K from the received mixture signals only, i.e., blindly. This can be accomplished by examining the rank of the received data matrix.

Let us consider the noiseless model in (5), reproduced here for convenience,

$$\mathbf{X}_{R\times N} = \mathbf{Z}_{R\times K}\mathbf{S}_{K\times N}. \tag{13}$$

Usually the packet length N is much larger than K and R, so $\mathbf{S}$ is fat and full rank, $\mathrm{rank}(\mathbf{S}) = \mathrm{K}$. The Vandermonde structure of $\mathbf{Z}$ and associated signature assignment scheme assure that $\mathrm{rank}(\mathbf{Z}) = \min(\mathrm{R}, \mathrm{K})$. Therefore, when $R \le K$, $\mathbf{X}$ is full rank, $\mathrm{rank}(\mathbf{X}) = \mathrm{R}$; while when $R > K$, $\mathbf{X}$ will lose rank, $\mathrm{rank}(\mathbf{X}) = \mathrm{K} < \min(\mathrm{R}, \mathrm{N})$. This property leads to the following collision detection and retransmission scheduling method:

After each retransmission, say the $(r-1)$th retransmission, the BS computes the rank of the data matrix $\mathbf{X}(r)$ (r indicating that $\mathbf{X}(r)$ has dimension $r \times N$). If $\mathbf{X}(r)$ is full rank, another retransmission is requested. The procedure continues until $\mathbf{X}(r)$ *loses rank.*

[6] $(\cdot)^\dagger$ stands for the matrix pseudo-inverse.

[7] Richard Harshman was the first to use ALS to fit the PARAFAC model in the early 1970's.

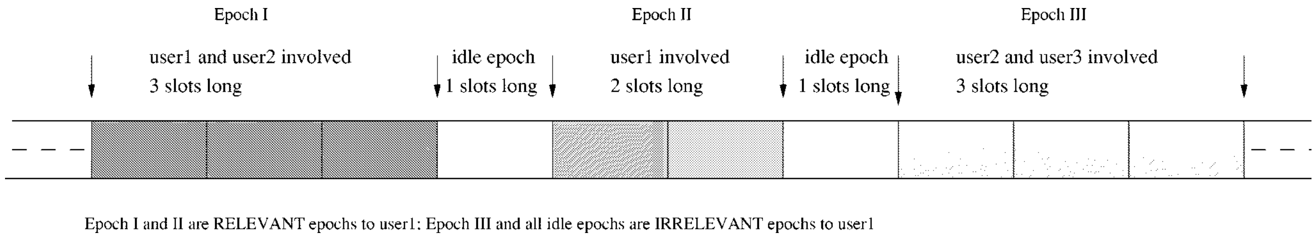


Fig. 3. Epoch flow in the channel.

When the retransmission procedure stops, $r-1=K$, so the number of retransmissions (not including the initial transmission) is just equal to the number of colliding users. Since the total number of transmissions is $r=K+1>K$, the collision resolution algorithm can proceed.

With noisy observations $\mathbf{Y}(r)$, we may examine the minimum eigenvalue of the correlation matrix $\mathbf{Y}(r)\mathbf{Y}^H(r)$ to determine the rank of $\mathbf{Y}(r)$, and thus the collision multiplicity K. Denote the minimum eigenvalue of $\mathbf{Y}(r)\mathbf{Y}^H(r)$ with $\lambda_{\min}(r)$. Comparing $\lambda_{\min}(r)$ to a tolerance threshold τ yields the following rank detector:

$$\begin{aligned} \lambda_{\min}(r) \geq \tau &\rightarrow H_0{:}\mathbf{Y}(r) \text{ is full rank} \\ \lambda_{\min}(r) < \tau &\rightarrow H_1{:}\mathbf{Y}(r) \text{ is rank-deficient.} \end{aligned} \tag{14}$$

Though without noise $\mathbf{Y}(r)$ first loses rank when $r=K+1$, $r-1$ is not necessarily a good estimate of K given a "singular" $\mathbf{Y}(r)$, when noise is present. In the array signal processing literature, information-theoretic criteria have been used to determine the number of signals impinging on a sensor array [19]. These techniques can also be applied here to estimate K from $\mathbf{Y}(r)$ in case of $r>K$, since the two problems share the same structure.

The retransmission scheduling, collision detection, and collision resolution methods developed in this section constitute a complete blind random access solution that spans the two lowest network layers: medium access and the physical layer. The acronym B-NDMA will henceforth stand for the above three components, with the understanding that collision detection is performed via the rank test method, while collision resolution is achieved via eigenanalysis followed by TALS refinement, although several other possibilities exist. B-NDMA performance is next.

IV. Performance of B-NDMA

Throughput and delay are the two most important performance metrics for random access techniques. Three factors come into play in the study of B-NDMA throughput and delay performance. The users' data traffic is of course the primary one. The performance of the collision detector is also very important. When the BS underestimates the number of colliding users and requests insufficient retransmissions, the resolvability condition will be violated and the collision resolution algorithm will fail, hence throughput will suffer. On the other hand, if the BS overestimates the number of colliding users, then extra channel slots will be wasted in unnecessary retransmissions, hence delay performance will suffer. The last factor is the residual bit-error-rate (BER) of the collision resolution algorithm, which may be nonzero even if collision detection is perfect. Residual bit errors beyond the correction capability of the error correcting code (ECC) used to protect the packet's payload translate to effective packet loss.

In order to avoid being overwhelmed by the numerous variations of collision detection and resolution algorithms and ECC schemes, and to make the analysis mathematically tractable and conceptually clear, we will mainly focus on the network aspect (the dependence of throughput and delay on the traffic load), and deal with the ideal case of infinite SNR wherein the BS makes no errors in collision detection and resolution. The results of the infinite SNR case may be viewed as an upper bound on B-NDMA system performance. A discussion of finite SNR effects will be given at the end of this section. The analysis will be corroborated by Monte Carlo simulation for a range of reasonable SNR's.

A. Collision Resolution Epoch

We first introduce the concept of the collision resolution epoch, which is the basis of our analysis. In the B-NDMA scheme, a group of consecutive channel slots are dedicated in transmitting several colliding packets. Hence these slots can be viewed as constituting a collision resolution period, which we call *epoch*. Fig. 3 illustrates the epoch flow in the channel. The epoch length (i.e., the number of slots in the epoch) is a random variable determined by the number of active users at the beginning of the epoch. In the ideal case (infinite SNR), K active users will result in an epoch of length $K+1$ to carry their payloads. Therefore, the probability mass function of the epoch length h is

$$P(\mathsf{h}=m)=\binom{J}{m-1}(1-p_e)^{m-1}p_e^{J-m+1}, m=1,\ldots,J+1 \tag{15}$$

where p_e is the probability that a user's packet buffer is empty and hence this user is inactive at the beginning of the epoch. p_e is a traffic-dependent parameter.

From the viewpoint of a specific user, say user i, two kinds of epochs can be distinguished, according to whether or not user i is active in this epoch. If user i sends packets during an epoch, this epoch is a *relevant epoch* to user i. Otherwise, if user i does not send anything during an epoch, this epoch is an *irrelevant epoch* to user i. Using h_R and h_I to denote the length of the relevant and irrelevant epoch, respectively, we have the mass functions

$$\begin{aligned} p(\mathsf{h}_R=m) =& p(\mathsf{h}=m|\text{user } i \text{ active}) \\ =& \binom{J-1}{m-2}(1-p_e)^{m-2}p_e^{J-m+1}, \\ m =& 2,\ldots,J+1, \end{aligned} \tag{16}$$

$$\begin{aligned} p(\mathsf{h}_I = m) =& p(\mathsf{h} = m|\text{user } i \text{ inactive}) \\ =& \binom{J-1}{m-1}(1-p_e)^{m-1}p_e^{J-m}, \\ m =& 1, \ldots, J. \end{aligned} \tag{17}$$

B. Queueing Analysis of the User's Packet Buffer

In this section we analyze the queue structure of the packet buffer of a typical user in the B-NDMA system. The analysis will produce an expression for p_e, which is the probability of a user's packet buffer being empty at the beginning of an epoch.

We assume that the user's packet buffer is an infinite queue, in which the packets are waiting to be served by the channel. We focus on the number of packets in one user's buffer at the beginning instance of the epoch, and specify it as a state variable. Denoting the state of the epoch l with q_l, we are seeking the mass function $p(q_l = k)$ (or equivalently, its generating function). As a special case $p_e = \lim_{l\to\infty} p(q_l = 0)$.

The sequence $q_l, l = 1, \ldots$ constitutes a Markov chain with the state transition relation

$$q_{l+1} = \begin{cases} q_l - 1 + r_l & q_l > 0 \\ r_l & q_l = 0 \end{cases} \tag{18}$$

where r_l is the number of new packets that come into the user's buffer during epoch l. Denote the steady state generating function of q_l as $Q(z) = \lim_{l\to\infty} E[z^{q_l}]$. We have the following result:

Proposition 1: If the buffer of every user in the system is driven by a user-wise independent Poisson packet arrival process of rate λ packets/slot, then the number of packets in a user's buffer at the beginning of an epoch has steady state generating function

$$Q(z) = p_e \frac{\left(z - e^{\lambda z - \lambda}\right)\left[(1-p_e)z + p_e\right]^{J-1}}{1 - \left[(1-p_e)z + p_e\right]^{J-1}} \tag{19}$$

□

The proof of this proposition can be found in Appendix.

To get p_e we may evaluate $Q(z)$ in (19) at $z = 1$ and (applying L'Hopital's rule) obtain

$$1 = p_e \frac{1 + \lambda H_I'(1) - \lambda H_R'(1)}{1 - \lambda H_R'(1)}. \tag{20}$$

Then using (16) and (17), and after straightforward derivations, we obtain

$$p_e = 1 - \frac{\lambda}{1 - \lambda J}. \tag{21}$$

C. Maximum Throughput

Since the collision detection and resolution procedure is perfect under the working assumption of infinite SNR, no packets will be lost in transmission. Throughput is therefore limited by traffic load, and maximum throughput (capacity) can be determined by driving the system with the maximum sustainable traffic load. This can be achieved by setting $p_e = 0$ in (20). This yields $\lambda_{\max} = 1/(J+1)$, and maximum throughput $J\lambda_{\max} = J/(J+1)$—which is monotonically increasing in the user population J. Note that the original training-based NDMA approach can only sustain maximum throughput of $N/(N+J)$ (again under infinite SNR), due to the fact that a J-user NDMA system needs orthogonal ID sequences of length $\geq J$. If the length of the payload of the data packet is N, the net utilization of the channel bandwidth is $N/(N+J)$—which is monotonically decreasing in J for a fixed N. For large user populations, B-NDMA exhibits better capacity, irrespective of N.

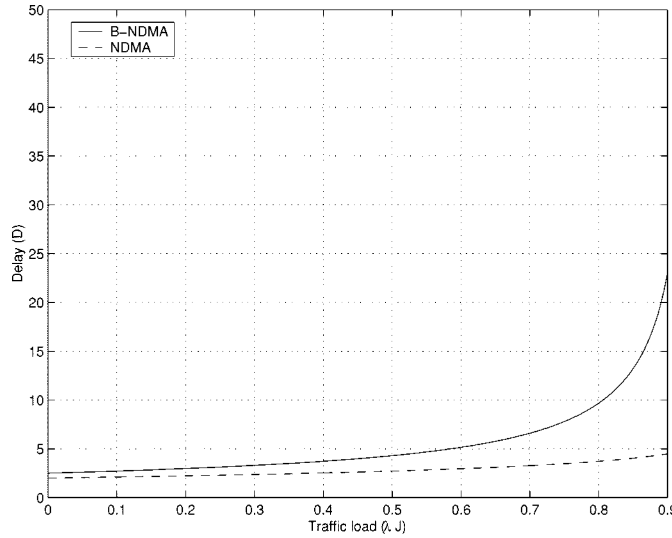


Fig. 4. Delay comparison of B-NDMA and original NDMA schemes ($J = 16$).

D. Delay

It is interesting to note that a user's buffer can be modeled as an M/G/1 queue with vacation[8], in which the relevant epoch and irrelevant epoch play the roles of the service time and vacation time, respectively. The average system delay (including waiting time in the buffer and transmission time in the channel) for a data packet through an M/G/1 queue with vacation is given by [5]

$$D = E[\mathsf{h}_R] + \frac{\lambda E[\mathsf{h}_R^2]}{2(1 - \lambda E[\mathsf{h}_R])} + \frac{E[\mathsf{h}_I^2]}{2E[\mathsf{h}_I]} \tag{22}$$

where the moments of h_R and h_I in (22) can be computed from their mass functions in (16) and (17).

B-NDMA guarantees blind resolvability at the cost of one extra retransmission relative to training-based NDMA. On the other hand, B-NDMA does not require orthogonal ID's (training), thereby effectively reducing packet length for a given payload. The trade-off improves in favor of B-NDMA as the size of the population increases. Even for small populations and even discounting the training issue, the extra latency is small: Fig. 4 is a comparison between B-NDMA and NDMA

[8]The M/G/1 queue model for the B-NDMA system is only an approximation, because the M/G/1 queue model requires that the service time and vacation time are independent, whereas in B-NDMA both are related to traffic load. The approximation is valid if the user population is large—compare the plot of (22) in Fig. 4 with the experimental result in Fig. 8 for $J = 16$ users.

delay performance for a population of $J = 16$ users. The curves depict mean delay versus traffic load λJ for the two systems, without taking training into account. B-NDMA does not incur a significant delay penalty relative to training-based NDMA for up to moderate traffic loads.

E. Some Considerations on the Finite SNR Scenario

For finite SNR, the analysis becomes more involved. It is no longer true that k active users always cause $k+1$ transmissions. A more complete description of the collision detector is needed. Such a description is the probability that m channels slots are dedicated to resolve a k-fold collision, denoted as $G(m|k)$. $G(m|k)$ is spread as a result of finite SNR, and this causes collision multiplicity under- and over-estimation errors. Under-estimation invariably leads to collision resolution failure, and hence the loss of all data packets involved, resulting in a throughput penalty; while over-estimation wastes channel slots, resulting in a delay penalty.

For the detector given in Section III-D,

$$G(m|k) = \mathrm{P}\left[\lambda_{\min}(1) \geq \tau, \ldots, \lambda_{\min}(m-1) \geq \tau, \lambda_{\min}(m) < \tau | k\right]. \quad (23)$$

It is clear from (23) that the selection of τ affects over/under-estimation and hence the throughput/delay performance tradeoff. Further analytical insights are possible, but the analysis is lengthy and perhaps of limited use. Section V presents finite SNR simulation results that corroborate our noiseless analysis, demonstrating that the method is actually quite robust in that it exhibits close to ideal performance for down to moderate SNR.

V. SIMULATIONS

We tested the B-NDMA protocol on a simulated slotted cellular packet radio system. Each user's buffer was fed by a Poisson process of rate λ packets/slot. The uncoded data packet was 274 bits long (equal to the length of the uncoded CDPD packet). A Cyclic Redundancy Code (CRC) was employed for error protection, and the resolved packets were CRCed to determine successful transmission. The channel was modeled as an additive white Gaussian noise channel. Throughout, the rank test method was used for collision detection, and eigenanalysis followed by TALS refinement was used for collision resolution. The collision detection threshold τ was set at two times the noise power level. We simulated two user population cases: $J = 8$ and 16, each under three SNR scenarios: 10, 15, and 20 dB. The number of channel slots simulated was varied as a function of traffic load, to assure sufficient statistical averaging.

Figs. 5–8 depict B-NDMA throughput and delay versus traffic load λJ. For down to moderate SNR (10 dB or above), B-NDMA exhibits an excellent throughput performance, very close to the ideal noiseless case (a straight line), while delay is low and relatively insensitive with respect to SNR. These results confirm our theoretical analysis.

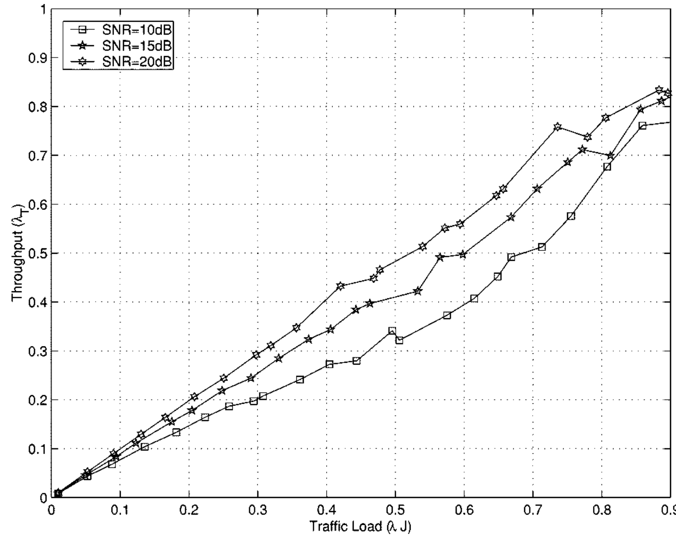


Fig. 5. B-NDMA throughput versus traffic load ($J = 8$).

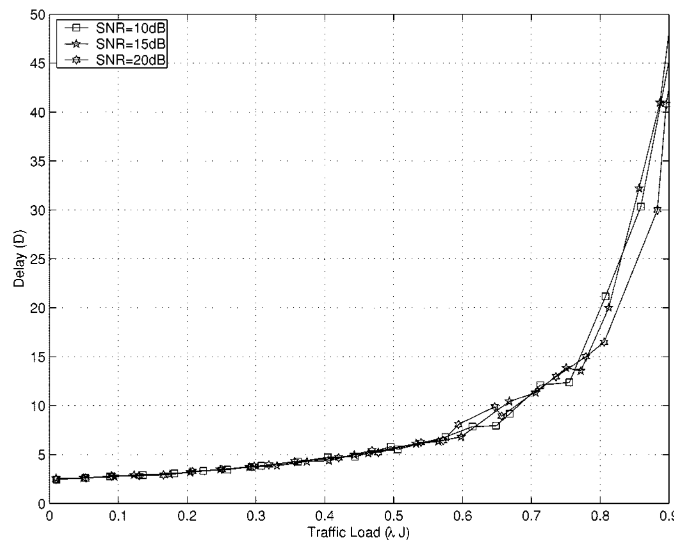


Fig. 6. B-NDMA delay versus traffic load ($J = 8$).

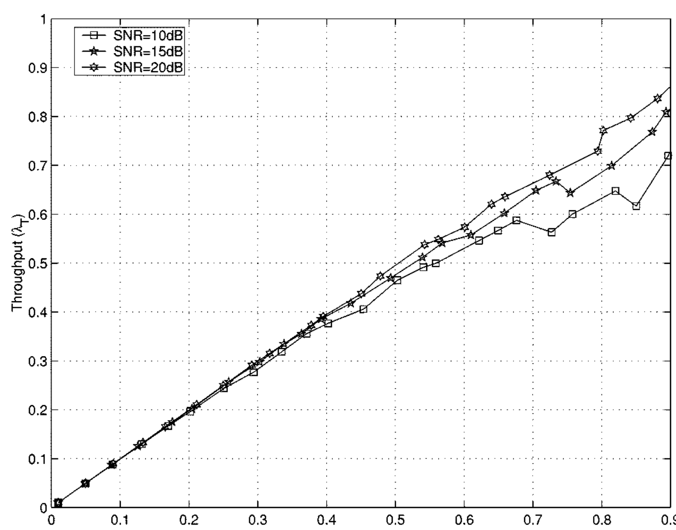


Fig. 7. B-NDMA throughput versus traffic load ($J = 16$).

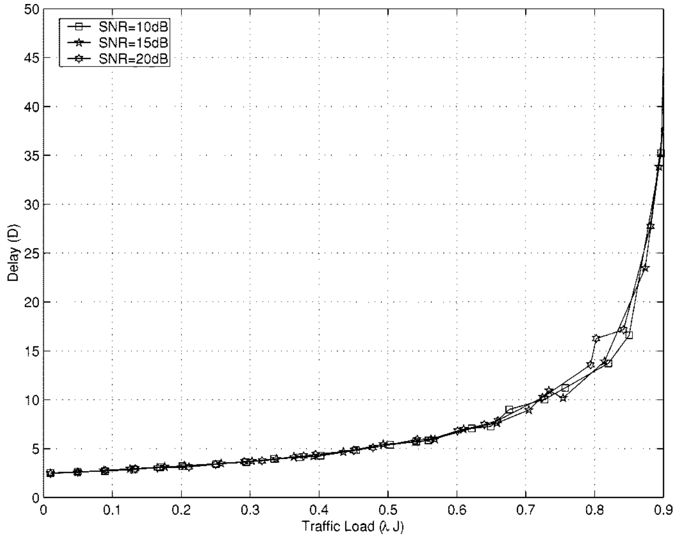


Fig. 8. B-NDMA delay versus traffic load $(J = 16)$.

VI. Conclusions

B-NDMA is a significant improvement of the original training-based NDMA scheme. Taking advantage of advanced factor analysis techniques, B-NDMA provides a blind deterministic collision resolution method that guarantees recovery of collided packets without requiring bandwidth-consuming training sequences, while incurring only a modest throughput-delay penalty (blind tradeoff). The method works with multipath channels as well since we do not require the user symbol streams to be i.i.d. Under the assumption that each user's channel is time-invariant for the duration of the collision resolution epoch, the method is applicable for colored symbol streams as well. Upon recovery of the packets, single-user equalization techniques can be used to equalize the individual user's packets using training based or blind techniques.

Appendix

The steady state generating function of the state q_l is defined as $Q_l(z) = E[z^{q_l}]$. From (18)

$$\begin{aligned} Q_{l+1}(z) =& (1 - P[q_l = 0])\, E\left[z^{q_l - 1 + r_l} | q_l > 0\right] \\ &+ P[q_l = 0] E\left[z^{r_l} | q_l = 0\right] \\ =& z^{-1}(1 - P[q_l = 0]) E\left[z^{q_l} | q_l > 0\right] E[z^{r_l} | q_l > 0] \\ &+ P[q_l = 0] E[z^{r_l} | q_l = 0]. \end{aligned} \tag{A.1}$$

It is easy to derive that

$$E\left[z^{q_l} | q_l > 0\right] = \frac{Q_l(z) - P[q_l = 0]}{1 - P[q_l = 0]}. \tag{A.2}$$

As for $E\left[z^{r_l} | q_l = 0\right]$ and $E\left[z^{r_l} | q_l > 0\right]$, we should note that when $q_l = 0$ epoch l will be an irrelevant epoch to the said user and when $q_l > 0$ epoch l will be a relevant epoch. So r_l when $q_l = 0$ and $q_l > 0$ are the number of new packets arriving during an irrelevant epoch and relevant epoch, respectively. Consequently, for Poisson arrivals of rate λ we have

$$\begin{aligned} E\left[z^{r_l} | q_l = 0\right] =& \sum_{k=0}^{\infty} \left[\sum_{m=1}^{J} e^{m\lambda} \frac{(m\lambda)^k}{k!} p(h_I = m) \right] z^k \\ =& H_I\left(e^{\lambda z - \lambda}\right), \end{aligned} \tag{A.3}$$

$$\begin{aligned} E\left[z^{r_l} | q_l > 0\right] =& \sum_{k=0}^{\infty} \left[\sum_{m=1}^{J} e^{m\lambda} \frac{(m\lambda)^k}{k!} p(h_R = m) \right] z^k \\ =& H_R(e^{\lambda z - \lambda}), \end{aligned} \tag{A.4}$$

where $H_R(z)$ and $H_I(z)$ are the generating function of the length of the relevant epoch and irrelevant epoch, respectively, defined as $H_R(z) = E\left[z^{h_R}\right]$ and $H_I(z) = E\left[z^{h_I}\right]$. It is straightforward from (16) and (17) to obtain that $H_R(z) = z^2[(1 - p_e)z + p_e]^{J-1}$ and $H_I(z) = z[(1 - p_e)z + p_e]^{J-1}$.

Substituting (A.2), (A.3) and (A.4) and the expressions for $H_R(z)$ and $H_I(z)$ into (A.1), and taking the limit of both sides as $l \to \infty$, we obtain the steady state generating function (19).

References

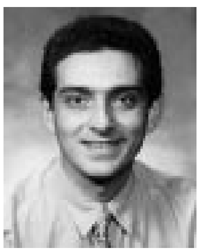
[1] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1992.
[2] R. Bro, N. D. Sidiropoulos, and G. B. Giannakis, “Optimal joint azimuth-elevation and signal-array response estimation using parallel factor analysis,” in *Proc. 32nd Asilomar Conf. on Signals, Systems, and Computers*, Monterey, CA, Nov. 1998.
[3] T. Cover, “Some advances in broadcast channels,” in *Advances in Communications Systems*. New York: Academic, 1975, pp. 229–260.
[4] D. J. Goodman, *Wireless Personal Communications Systems*. Reading, MA: Addison Wesley Longman, 1997.
[5] L. Kleinrock, *Queueing Systems*. New York: Wiley, 1975, vol. I.
[6] J. B. Kruskal, “Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics,” *Linear Algebra and Its Applicat.*, vol. 18, pp. 95–138, 1977.
[7] ——, “Rank decomposition, and uniqueness for 3-way and N-way arrays,” in *Multiway Data Analysis*, R. Coppi and S. Bolasco, Eds. Amsterdam, The Netherlands: North-Holland, 1988, pp. 7–18.
[8] B. Ottersten and A. L. Swindlehurst, “Antenna arrays for wireless networks,” *IEEE Signal Processing Mag.*, pp. 25–27, Mar. 1999.
[9] K. Pahlavan and A. H. Levesque, *Wireless Information Networks*. New York: Wiley, 1995.
[10] T. S. Rappaport, *Wireless Communications, Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 1996.
[11] R. Roy and T. Kailath, “ESPRIT—estimation of signal parameters via rotational invariance techniques,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 984–995, July 1989.
[12] P. H. Schonemann, “An algebraic solution for a class of subjective metrics models,” *Psychometrika*, vol. 37, p. 441, 1972.
[13] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro, “Deterministic waveform-preserving blind separation of DS-CDMA signals using an antenna array,” in *Proc. IEEE Signal Processing Workshop on Statist. Signal and Array*, Portland, OR, Sept. 14–16, 1998, pp. 304–307.
[14] ——, “Blind PARAFAC receivers for DS-CDMA systems,” *IEEE Trans. Signal Processing*, vol. 48, pp. 810–823, Mar. 2000.
[15] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis, “Parallel factor analysis in sensor array processing,” *IEEE Trans. Signal Processing*, vol. 48, pp. 2377–2388, Aug. 2000.
[16] P. Stoica and A. Nehorai, “MUSIC, maximum likelihood, and Cramér-Rao bound,” *IEEE Trans. Acoust.,Speech, Signal Processing*, vol. 37, pp. 720–741, May 1989.
[17] M. K. Tsatsanis, R. Zhang, and S. Banerjee, “Network assisted diversity multiple access for wireless networks,” *IEEE Trans. Signal Processing*, vol. 48, pp. 702–711, Mar. 2000.
[18] S. Verdu, *Multiuser Detection*. Cambridge, U.K: Cambridge Univ. Press, 1998.

[19] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoustic, Speech, Signal Processing*, vol. ASSP-33, pp. 384–392, Feb. 1985.

[20] R. Zhang and M. Tsatsanis, "Network assisted diversity multiple access in dispersive channels," in *Proc. 1999 2nd IEEE Workshop on Signal Processing Advances in Wireless Communications*, Annapolis, MD, May 9–12, 1998, pp. 263–287.

Ruifeng Zhang (S'97–M'00) received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 1993, the M.E. degree from Beijing Institute of Technology, Beijing, China, in 1996, and the Ph.D degree from Stevens Institute of Technology, Hoboken, NJ, in 2000, all in electrical engineering.

He joined the faculty of Department of Electrical and Computer Engineering, Drexel University, Philadelphia, Pensylvenia in 2000, where he is currently an Assistant Professor. His research interests are in the area of signal processing for communications and networking, with focus on multiple access, multiuser detection, and signal processing methods for networking.



Nicholas D. Sidiropoulos (M'92–SM'99) received the Diploma in electrical engineering from the Aristotelian University of Thessaloniki, Greece, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland at College Park (UMCP), in 1988, 1990, and 1992, respectively.

From 1988 to 1992 he was a Fulbright Fellow and a Research Assistant at the Institute for Systems Research (ISR) of the University of Maryland. From September 1992 to June 1994 he served his militaryservice as a Lecturer in the Hellenic Air Force Academy. From October 1993 to June 1994 he also was a member of the technical staff, Systems Integration Division, G-Systems Ltd., Athens, Greece. He has held Postdoctoral (1994–1995) and Research Scientist (1996–1997) positions at ISR-UMCP, before joining the Department of Electrical Engineering at the University of Virginia in July 1997 as an Assistant Professor. He is currently an Associate Professor in the Department of Electrical and Computer Engineering at the University of Minnesota, Minneapolis. His current research interests are primarily in multi-way analysis and its applications in signal processing for communications & networking.

Dr. Sidiropoulos is a member of the Signal Processing for Communications Technical Committee (SPCOM-TC) of the IEEE SP Society, and currently serves as Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING, and IEEE SIGNAL PROCESSING LETTERS. He received the NSF/CAREER award (Signal Processing Systems Program) in June 1998.

Michail K. Tsatsanis (M'93) received his diploma degree in electrical engineering from the National Technical University of Athens, Greece, in 1987, and the M.Sc. and Ph.D degrees in electrical engineering from the University of Virginia in 1990 and 1993, respectively.

He currently holds the position of Chief Scientist at Voyan Technology, Santa Clara, CA, which he joined in 2000. Before that, he has held various academic positions as an assistant and associate professor in the Electrical and Computer Engineering Department at the Stevens Institute of Technology, Hoboken, NJ, (1995–2000) and as a research associate in the Electrical Engineering Department at the University of Virginia. His general research interests lie in the areas of statistical signal and array processing, with applications to communications and networking. His current interests focus on signal processing techniques for wireless and wireline communications including blind equalization, multiuser detection, fading channel estimation and tracking, and signal processing methods for networking problems.

Dr. Tsatsanis is a member of the IEEE technical committee on SPCOM. He has served as a member of the organizing committee for the 1996 IEEE Signal Processing Workshop on SSAP and as the technical co-chair of the organizing committee for the 1999 IEEE Workshop on Signal Processing Advances in Wireless Communications. he received the 1998 NSF CAREER award and the 1999 IEEE Signal Processing Society Young Author Best Paper Award. He has also served the IEEE in the capacity of the associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE COMMUNICATIONS LETTERS.