JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

2/2003

Internet, telecommunications and signal processing

Special issue edited by Tadeusz Antoni Wysocki

ISSN 1509-4553

Circulation: 300 copies

Sowa - Druk na życzenie, www.sowadruk.pl, tel. 022 431-81-40

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

Preface

In recent years, we have been witnessing an information revolution caused by explosion of the Internet and associated applications. This effect, firstly confined to computers connected to wired networks has been later magnified by an introduction of wireless data networks and incorporation of the Internet into the domain of mobile telephony. Because of the resulting volume of the Internet related traffic in all telecommunication areas, the research efforts in both telecommunications and the associated signal processing started to be more and more linked to the Internet. That connection is either in terms of an underlying technologies enabling performance improvements or development of new applications, or in the use of the Internet Protocol (IP) for better utilization of telecommunication resources.

The coupling of the Internet related research with the telecommunications and signal processing has influenced a group of researchers in Australia to organize the 1st Workshop on the Internet, Telecommunications and Signal Processing, held in Wollongong on 9–11 December 2002 (the 2nd Workshop on the Internet, Telecommunications and Signal Processing will be held in conjunction with the 7th International Symposium on DSP and Communication Systems in Coolangatta, Gold Coast, on 8–11 December 2003). This special issue on the Internet, Telecommunications and Signal Processing contains 12 selected and revised papers from that workshop.

The first paper *Applications of Hadamard matrices* by Evangelaras, Koukouvinos, and Seberry deals with the theory of Hadamard matrices, which play a very important role in several areas of modern telecommunications like error-control coding or code division multiple access (CDMA) systems. That paper is followed by two papers *Design of filterbank transceivers for dispersive channels with arbitrary-length impulse response* by Mertins, and *Blind frequency offset estimation for overlap PCC-OFDM systems in presence of phase noise* by Shentu and Armstrong, investigating very important issues of designing physical channels for the future wireless networks. The next two papers *LPAR: an adaptive routing strategy for MANETs* by Abolhasan, Wysocki and Dutkiewicz, and *Performance analysis of reactive shortest path and multi-path routing mechanism with load balance* by Pham and Perreau, are concerned with routing protocols for mobile ad hoc networks (MANETs).

The sixth paper of the issue *Linear quadratic power control for CDMA systems* by Anderson, Perreau, and White introduces a novel approach to solve one of the bottlenecks of CDMA systems, which is the power control. It is followed by Jayasuriya's paper *Queuing models for cellular networks with generalised Erlang service distributions*, which proposes a simplified model of cellular networks. This model can be used to calculate blocking probabilities for handover and new users.

The next group of three papers: *Adaptive handover control in IP-based mobility networks* by Park and Dadej, *A concept of Differentiated Services architecture supporting military oriented Quality of Service* by Kwiatkowski, and *Bandwidth broker extension for optimal resource management* by Sohail and Jha, deal with applications of IP, mobile IP, and quality of service (QoS) issues in the IP based networks.

The final two papers of the issue *Manipulation of compressed data using MPEG-7 low level audio descriptors* by Lukasiak, Stirling, Perrow and Harders, and *Fully spatial and SNR scalable, SPIHT-based image coding for transmission over heterogeneous networks* by Danyali and Mertins are concerned with applications of signal processing for multimedia communications over the Internet.

The guest editor would like to thank here all the authors for their contributions and the reviewers for their hard work in preparing the submissions, reviewing, and revising the papers on time.

Tadeusz Antoni Wysocki
Guest Editor

Invited paper

Applications of Hadamard matrices

Haralambos Evangelaras, Christos Koukouvinos, and Jennifer Seberry

Abstract — We present a number of applications of Hadamard matrices to signal processing, optical multiplexing, error correction coding, and design and analysis of statistics.

Keywords — Hadamard matrices, orthogonal sequences, CDMA spreading codes, Walsh functions, optical multiplexing.

1. Hadamard matrices and definitions

A square matrix with elements ± 1 and size h, whose distinct row vectors are orthogonal is an *Hadamard matrix* of order h. The smallest examples are

$$[\,1\,],\quad \begin{bmatrix} 1 & 1 \\ 1 & - \end{bmatrix},\quad \begin{bmatrix} - & 1 & 1 & 1 \\ 1 & - & 1 & 1 \\ 1 & 1 & - & 1 \\ 1 & 1 & 1 & - \end{bmatrix}$$

where we write $-$ for -1. These were first studied by J. J. Sylvester [17] who observed that if H is an Hadamard matrix then

$$\begin{bmatrix} H & H \\ H & -H \end{bmatrix}$$

is also an Hadamard matrix. Indeed, using the matrix of order 2 we have

Lemma 1 (Sylvester [17]): *There is an Hadamard matrix of order 2^t for all integers t.*

We call matrices of order 2^t constructed by Sylvester's construction *Sylvester-Hadamard matrices*. They are naturally associated with discrete orthogonal functions called *Walsh functions*. Using Sylvester's method the first few Hadamard matrices obtained are:

$$\begin{bmatrix} 1 & 1 \\ 1 & - \end{bmatrix} \quad \left[\begin{array}{cc|cc} 1 & 1 & 1 & 1 \\ 1 & - & 1 & - \\ \hline 1 & 1 & - & - \\ 1 & - & - & 1 \end{array}\right] \quad \left[\begin{array}{cccc|cccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & - & 1 & - & 1 & - & 1 & - \\ 1 & 1 & - & - & 1 & 1 & - & - \\ 1 & - & - & 1 & 1 & - & - & 1 \\ \hline 1 & 1 & 1 & 1 & - & - & - & - \\ 1 & - & 1 & - & - & 1 & - & 1 \\ 1 & 1 & - & - & - & - & 1 & 1 \\ 1 & - & - & 1 & - & 1 & 1 & - \end{array}\right]$$

For these matrices we count, row by row, the number of times the sign changes so $1--1$ changes sign twice. This gives:

for the matrix of order 2 : 0,1
for the matrix of order 4 : 0,3,1,2,
for the matrix of order 8 : 0,7,3,4,1,6,2,5

Indeed we shall see that the set of the number of sign changes in a Sylvester-Hadamard matrix of order n is $\{0,1,\ldots,n-1\}$ corresponding to the number of zero crossings of the Walsh functions.

In 1893 Jacques Hadamard [4] gave examples for a few small orders and conjectured they exist for every order divisible by 4. An example for order 12 is:

$$\begin{bmatrix}
1 & 1 & 1 & - & 1 & 1 & - & 1 & 1 & - & 1 & 1 \\
1 & 1 & 1 & 1 & - & 1 & 1 & - & 1 & 1 & - & 1 \\
1 & 1 & 1 & 1 & 1 & - & 1 & 1 & - & 1 & 1 & - \\
1 & - & - & 1 & 1 & 1 & - & 1 & 1 & 1 & - & - \\
- & 1 & - & 1 & 1 & 1 & 1 & - & 1 & - & 1 & - \\
- & - & 1 & 1 & 1 & 1 & 1 & 1 & - & - & - & 1 \\
1 & - & - & 1 & - & - & 1 & 1 & 1 & - & 1 & 1 \\
- & 1 & - & - & 1 & - & 1 & 1 & 1 & 1 & - & 1 \\
- & - & 1 & - & - & 1 & 1 & 1 & 1 & 1 & 1 & - \\
1 & - & - & - & 1 & 1 & 1 & - & - & 1 & 1 & 1 \\
- & 1 & - & 1 & - & 1 & - & 1 & - & 1 & 1 & 1 \\
- & - & 1 & 1 & 1 & - & - & - & 1 & 1 & 1 & 1
\end{bmatrix}$$

We now look at some basic properties of Hadamard matrices:

Lemma 2: *Let H be an Hadamard matrix of order h. Then:*

(i) $HH^T = hI_h$;

(ii) $|\det H| = h^{\frac{1}{2}h}$;

(iii) $HH^T = H^T H$;

(iv) Hadamard matrices may be changed into other Hadamard matrices by permuting rows and columns and by multiplying rows and columns by -1. We call matrices which can be obtained from one another by these methods H-equivalent (not all Hadamard matrices of the same order are H-equivalent);

(v) every Hadamard matrix is H-equivalent to an Hadamard matrix which has every element of its first row and column $+1$ – matrices of this latter form are called normalized;

(vi) if H is a normalized Hadamard matrix of order $4n$, then every row (column) except the first has $2n$ minus ones and $2n$ plus ones in each row (column), further n minus ones in any row (column) overlap with n minus ones in each other row (column);

(vii) the order of an Hadamard matrix is 1,2, or $4n$, n positive integer.

Definition 1: If $M = (m_{ij})$ is a $m \times p$ matrix and $N = (n_{ij})$ is an $n \times q$ matrix, then the *Kronecker product* $M \times N$ is the $mn \times pq$ matrix given by

$$M \times N = \begin{bmatrix} m_{11}N & m_{21}N & \cdots & m_{1p}N \\ m_{12}N & m_{22}N & \cdots & m_{2p}N \\ \vdots & & & \\ m_{m1}N & m_{m2}N & \cdots & m_{mp}N \end{bmatrix}$$

Example:

$$\text{Let } M = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \text{ and}$$

$$N = \begin{bmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix} \text{ then}$$

$$M \times N = \begin{bmatrix} N & N \\ N & -N \end{bmatrix} =$$

$$= \begin{bmatrix} -1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 \\ & & & & & & & \\ -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

Lemma 3 (Sylvester-Hadamard): *Let H_1 and H_2 be Hadamard matrices of orders h_1 and h_2. Then by the properties of Kronecker products $H = H_1 \times H_2$ is an Hadamard matrix of order $h_1 h_2$.*

2. Historical background

More than one hundred years ago, in 1893, Jacques Hadamard [4] found square matrices of orders 12 and 20, with entries ± 1, which had all their rows (and columns) orthogonal. These matrices, $X = (x_{ij})$, satisfied the equality of the following inequality

$$|det\ X|^2 \leq \prod_{i=1}^{n} \sum_{j=1}^{n} |x_{ij}|^2$$

and had maximal determinant. Hadamard actually asked the question of matrices with entries on the unit disc but his name has become associated with the real matrices.
Hadamard was not the first to study these matrices for J. J. Sylvester in 1867 in his seminal paper "Thoughts on inverse orthogonal matrices, simultaneous sign-successions and tesselated pavements in two or more colours with application to Newton's rule, ornamental tile work and the theory of numbers" [17] had found such matrices for all orders which are powers of two. Nevertheless, Hadamard showed matrices with elements ± 1 and maximal determinant could exist for all orders 1, 2, and $4t$ and so the Hadamard conjecture "that there exists an *Hadamard matrix*, or square matrix with every element ± 1 and all row (column) vectors orthogonal" came from here. This survey discusses some of the applications of hadamard matrices.

2.1. Hadamard codes

Definition 2: The rows of an Hadamard matrix H of order $4n$ give a $(4n,\ 8n,\ n-1)$ block error correction code as each of the rows has distance at least $2n$ from each of the other rows. The block code is:

$$\begin{bmatrix} H \\ -H \end{bmatrix}$$

In the 1960's the U.S. Jet Propulsion Laboratories (JPL) was working toward building the Mariner and Voyager space probes to visit Mars and the other planets of the solar system. Those of us who saw early black and white pictures of the back of the moon remember that whole lines were missing. The first black and white television pictures from the first landing on the moon were extremely poor quality. How many of us now take the glorious high quality colour pictures of Jupiter, Saturn, Uranus, Neptune and their moons for granted.
In brief, these high quality colour pictures are taken by using three black and white pictures taken in turn through red, green and blue filters. Each picture is then considered as a thousand by a thousand matrix of black and white pixels. Each picture is graded on a scale of, say, one to sixteen, according to its greyness. So white is one and black is sixteen. These grades are then used to choose a codeword in, say, an eight error correction code based on, say, the Hadamard matrix of order 32. The codeword is transmitted to Earth, error corrected, the three black and white pictures reconstructed and then a computer used to reconstruct the coloured pictures.
Hadamard matrices were used for these codewords for two reasons, first, error correction codes based on Hadamard matrices have maximal error correction capability for a given length of codeword and, second, the Hadamard matrices of powers of two are analogous to the Walsh functions, thus all the computer processing can be accomplished using additions (which are very fast and easy to implement in computer hardware) rather than multiplications (which are far slower).

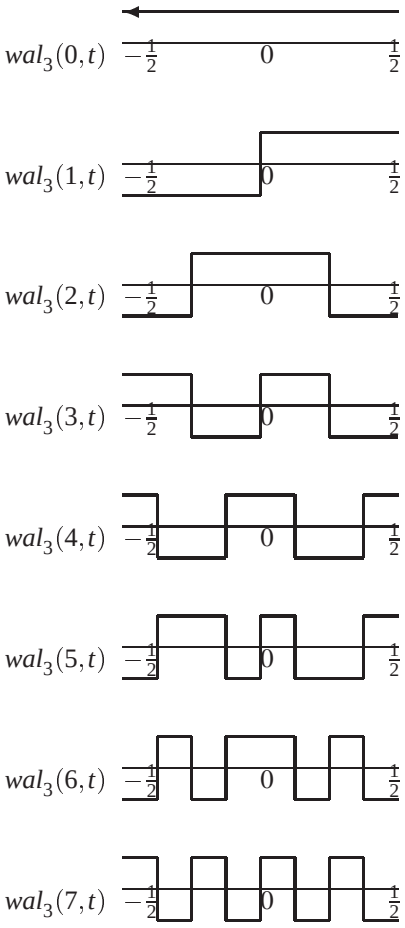
3. Walsh functions

Sylvester's original construction for Hadamard matrices is equivalent to finding Walsh functions which are the discrete analogue of Fourier series.

Example: Let H be a Sylvester-Hadamard matrix of order 8 and sequency order:

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{pmatrix}$$

The Walsh function generated by H is the following:



The points of intersections of Walsh functions are identical with that of trigonometrical functions. By mapping $w(i,t) = wal_n(i,t)$ into the interval $[-\frac{1}{2},0]$, then by mapping axial symmetrically into $[0,\frac{1}{2}]$, we get $w(2i,t)$ which is an even function. By operating similarly we get $w(2i-1,t)$, an odd function.

Just as any curve can be written as an infinite Fourier series

$$\sum_n a_n \sin nt + b_n \cos nt$$

the curve can be written in terms of Walsh functions

$$\sum_n a_n\, sal(i,t) + b_n\, cal(i,t) = \sum_n c_n\, wal(i,t).$$

The hardest curve to model with Fourier series is the step function $wal_2(0,t)$ and errors lead to the Gibbes phenomenom. Similarly, the hardest curve to model with Walsh functions is the basic $\sin 2\pi t$ or $\cos 2\pi t$ curve. Still, we see that we can transform from one to another.

Many problems require Fourier transforms to be taken, but Fourier transforms require many multiplications which are slow and expensive to execute. On the other hand, the fast Walsh-Hadamard transform uses only additions and subtractions (addition of the complement) and so is extensively used to transform power sequency spectrum density, band compression of television signals or facsimile signals or image processing.

4. Desired characteristics of CDMA spreading codes

Hadamard matrices have a significant role to play in the search for desirable CDMA spreading codes.

For bipolar spreading codes $\{s_n^{(i)}\}$ and $\{s_n^{(l)}\}$ of length N, the normalized discrete aperiodic correlation function is defined as [9]:

$$c_{i,l}(\tau) = \begin{cases} \frac{1}{N}\sum_{n=0}^{N-1-\tau} s_n^{(i)} s_{n+\tau}^{(l)}, & 0 \le \tau \le N-1 \\ \frac{1}{N}\sum_{n=0}^{N-1+\tau} s_{n-\tau}^{(i)} s_n^{(l)}, & 1-N \le \tau < 0 \\ 0, & |\tau| \ge N \end{cases}.$$

When $\{s_n^{(i)}\}$ equals $\{s_n^{(l)}\}$, the above equation defines the normalized discrete aperiodic auto-correlation function.

In order to evaluate the performance of a whole set of M spreading codes, the average mean sqaure value of cross-correlation for all codes in the set, denoted by R_{CC}, was introduced by Oppermann and Vucetic [12] as a measure of the set cross-correlation performance:

$$R_{CC} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{k=1 \\ k\neq i}}^{M} \sum_{\tau=1-N}^{N-1} |c_{i,k}(\tau)|^2 .$$

A similar measure, denoted by R_{AC} was introduced there for comparing the auto-correlation performance:

$$R_{AC} = \frac{1}{M} \sum_{i=1}^{M} \sum_{\substack{\tau=1-N \\ \tau\neq 0}}^{N-1} |c_{i,j}(\tau)|^2 .$$

The R_{AC} allows for comparison of the auto-correlation properties of the set of spreading codes on the same basis as their cross-correlation properties.

It is higly desirable to have both R_{CC} and R_{AC} as low as possible, as the higher value of R_{CC} results in stronger mult-access interferance (MAI), and an increase in the value of R_{AC} impedes code acquisition process. Unfortunately, decreasing the value of R_{CC} causes increase in the value of R_{AC}, and vice versa.

Both R_{CC} and R_{AC} are very useful for large code sets and large number of active users, when the constellation of interferers (i.e. relative delays among the active users and the spreading codes used) changes randomly for every transmitted information symbol. However, for a more static situation, when the constellation of interferers stays constant for the duration of many information symbols, it is also important to consider the worst-case scenarios. This can be accounted for by analyzing the maximum value of peaks in the aperiodic cross-correlation functions over the whole set of sequences and in the aperiodic auto-correlation function for $\tau \neq 0$. Hence, one needs to consider two additional measures to compare the spreading sequence sets.
Maximum value of the aperiodic cross-correlation functions C_{max}:

$$c_{max}(\tau) = \max_{\substack{i=1,\cdots,M \\ k=1,\cdots,M \\ i\neq k}} |c_{i,k}(\tau)|; \quad \tau=(-N+1),\cdots,(N-1).$$

Maximum value of the off-peak aperiodic auto-correlation functions A_{max}

$$a_{max}(\tau) = \max_{k=1,\cdots,M} |c_{k,k}(\tau)|;$$

$$A_{max} = \max_{\tau\neq 0}\{a_{max}(\tau)\}.$$

The known relationships between C_{max} and A_{max} are due to Welch [18] and Levenshtein [10].
The Welch bound states that for any set of M bipolar sequences of length N

$$max\{C_{max}, A_{max}\} \geq \sqrt{\frac{M-1}{2NM-M-1}}.$$

A tighter Levenshtein bound is expressed by:

$$max\{C_{max}, A_{max}\} \geq \sqrt{\frac{(2N^2+1)M-3N^2}{3N^2(MN-1)}}.$$

It must be noted here that both Welch and Levenshtein bounds are derived for sets of bipolar sequences where the condition of orthogonality for perfect synchronization is not imposed. Hence, one can expect that by introducing the orthogonality condition, the lower bound for the aperiodic cross-correlation and aperiodic out-of-phase auto-correlation magnitudes must be significantly lifted.

4.1. Constructions for Hadamard matrices for CDMA

There are many constructions for Hadamard matrices and recent work of Seberry, B. Wysocki and T. Wysocki [15] have found that different constructions give different auto-correlation and cross correlation coefficients when tested for CDMA coding.

5. Boolean functions

Hadamard matrices are intimately related with two families of symmetric balanced incomplete block designs. These families of designs are also connected with *boolean functions* used in the construction of *S*-boxes for cryptographic algorithms. The family SBIBD$(4t-1, 2t-1, t-1)$ is related to linear boolean functions and the SBIBD$(4s^2, 2s^2 \pm s, s^2 \pm s)$ to those functions which are "furthest" from linear functions the *bent functions*.

6. The existence and construction of a complete set of orthogonal $F(4t; 2t, 2t)$-squares

This material is from Walter T. Federer [2].

6.1. Introduction and definitions

Hedayat [5] and Hedayat and Seiden [6] have defined an F-square as follows:

Definition 3: Let $A = [a_{ij}]$ be an $n \times n$ matrix and let $\Sigma = (c_1, c_2, ..., c_m)$ be the ordered set of m distinct elements or symbols of A. In addition, suppose that for each $k = 1, 2, \cdots, m$, c_k appears exactly λ_k times $(\lambda_k \geq 1)$ in each row and in each column of A. Then A will be called a *frequency square* or, more concisely, an *F-square* on Σ of order n and frequency vector $(\lambda_1, \lambda_2, \cdots, \lambda_m)$ and will be denoted by $F(n; \lambda_1, \lambda_2, \cdots, \lambda_m)$. Note that $(\lambda_1 + \lambda_2 + \cdots + \lambda_m) = n$ and that when $\lambda_k = 1$ and $m = n$, a latin square results.
As with latin squares, one may consider orthogonality of a pair or a set of F-squares of the same order. The above cited authors give the following two definitions covering these cases:

Definition 4: Given an F-square $F_1(n; \lambda_1, \lambda_2, \cdots, \lambda_k)$ on a set $\Sigma = (a_1, a_2, \ldots, a_k)$ and an F-square $F_2(n; u_1, u_2, ..., u_t)$ on a set $\Omega = (b_1, b_2, ..., b_t)$, we say F_2 is an *orthogonal mate* for F_1 (and write $F_2 \perp F_1$) if upon supposition of F_2 on F_1, a_i appears $\lambda_i u_j$ times with b_j.

Definition 5: Let S_i be an n_i-set, $i = 1, 2, \cdots, t$, and let F_i be an F-square of order n on the set S_i with frequency vector $\lambda_i = (\lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{ih})$. Then $F_1, F_2, ..., F_t$ is a set of t *mutually (pairwise) orthogonal F-squares* if $F_i \perp F_j$, $i \neq j$; $i, j = 1, 2, \cdots, t$. If every $n_i = n$ and every $\lambda_{i\ell} = 1$, $\ell = 1, 2, \cdots, n$, a set of t mutually orthogonal latin squares results and is denoted as $OL(n, t)$.

If a complete set of orthogonal latin squares exists, then $t = n-1$ and the set is denoted as $OL(n, n-1)$. If a complete set of orthogonal F-squares of order n exists, the number will depend upon the values of the n_i in Definition 5. This leads to the following definition:

Definition 6: A complete set of orthogonal F-squares of order n is denoted as $CSOFS(\cdot,\cdot,\cdot)$, and the number of F-squares with i distinct elements is given by the terms in the summation $\sum_{i=2}^{n} N_i F(n; \lambda_1, \lambda_2, \cdots, \lambda_i)$ where $\sum h = 1^i \lambda_h = n$, $\sum i = 2^n N_i (i-1) = (n-1)^2$ and N_i is the number of the squares with i distinct elements.

The fact that $\sum i = 2^n N_i (i-1) = (n-1)^2$ in order to have a *CSOFS* follows directly from analysis of variance theory and from factorial theory in that the interaction of two n-level factors has $(n-1)^2$ degrees of freedom and from the fact that only interection degrees of freedom are available to construct F-squares. For each $F_i(n; \lambda_1, \lambda_2, \cdots, \lambda_i)$-square, there are $(i-1)$ degrees of freedom associated with the i distinct symbols of an F-square, there are N_i F-squares containing i symbols, and hence $(n-1)^2 = \sum_{i=2}^{n} N_i (i-1)$. Federer [2] showed that a *CSOFS* exists for $n = 4t$ and for $i = 2$ distinct symbols. The results have application in zero-one graph theory, in orthogonal arrays, in coding theory, and other areas. It illustrates now analysis of variance and factorial theory can be used to construct the CSOFS and thus provides a new tool for construction purposes.

6.2. Construction of a complete set of $F(4t; 2t, 2t)$-squares

The use of orthogonal contrasts in the analysis of variance for factorial experiments to construct latin squares was indicated by Federer et al [1], Mandeli [11] also used this procedure. It would appear that there is considerable potential in using the orthogonality of single degree of freedom contrasts from the interaction to construct F-squares and latin squares. The following theorem represents one such example:

Theorem 1: *There exists a complete set of* $(4t-1)^2$ *mutually orthogonal* $F(4t; 2t, 2t)$ *squares.*

Proof: A normalised Hadamard matrix is one in which there are all plus ones in the first row and in the first column. The remaining elements are plus and minus ones. Hadamard matrices of side $4t$ are known to exist for all $1 \le t \le 105$ and are presumed to exist for all $4t$. In the last $4t-1$ rows of a normalised $4t \times 4t$ Hadamard matrix, the number of plus ones is equal to the number of minus ones. The Kronecker product of two normalised Hadamard matrices, i.e. $H_{4t} \times H_{4t}$, is a normalised Hadamard matrix of side $16t^2$. Delete the first $4t$ rows of the resulting H_{16t^2} and delete the $4t+1$st, the $8t+1$st, ..., $16t^2-4t+1$st row of the H_{16t^2} matrix. $8t-1$ rows are thus deleted, leaving $(16t^2-8t+1) = (4t-1)^2$ rows having $2t$ plus ones and $2t$ minus ones. Let the plus one be symbol a_1 and the minus on be symbol a_2 in these remaining $(4t-1)^2$ rows. Thus, an $F(4t; 2t, 2t)$-square will be formed from each row resulting in $(4t-1)^2 F(4t; 2t, 2t)$-squares. The resulting F-squares will be mutually orthogonal from the properties of Hadamard matrices. Hence, the *CSOFS* is constructed in this manner. □

7. Hadamard matrices and optimal weighing designs

Suppose we are given p objects to be weighed in n weighings with a chemical balance (two–pan balance) having no bias. Let

$x_{ij} = 1$ if the jth object is placed in the left pan in the ith weighing,

$x_{ij} = -1$ if the jth object is placed in the right pan in the ith weighing,

Then the $n \times p$ matrix $X = (x_{ij})$ completely characterizes the weighing experiment.

Let us write $w_1, w_2, \cdots, w_p$ for the true weights of the p objects, and $y_1, y_2, \cdots, y_n$ for the results of n weighings (so that the readings indicate that the weight of the left pan exceeds that of the right pan by y_i in the weighing of i), denote the column vectors of w's and y's by W and Y respectively.

Then the readings can be represented by the linear model

$$Y = XW + e,$$

where e is the column vector of $e_1, e_2, \ldots, e_n$ and e_i is the error between observed and expected readings. We assume that e is a random vector distributed with mean zero and covariance matrix $\sigma^2 I$. This is a reasonable assumption in the case where the objects to be weighed have small mass compared to the mass of the balance.

We assume X to be a non-singular matrix. Then the best linear unbiased estimator of W is

$$\hat{W} = (X^T X)^{-1} X^T Y$$

with covariance of $\hat{W}$

$$Cov(\hat{W}) = \sigma^2 (X^T X)^{-1}.$$

Hotelling showed that for any weighing design the variance of $\hat{w}_i$ cannot be less than σ^2/n. Therefore, we shall call a weighing design X optimal if it estimates each of the weights with this minimum variance, σ^2/n. Kiefer [8] proved that an optimal weighing design in our sense is actually optimal with respect to a very general class of criteria. It can be shown that X is optimal if and only if $X^T X = nI$. This means that a chemical balance weighing design X is optimal if it is an $n \times p$ matrix of ± 1 whose columns are orthogonal, that is an *Hadamard matrix*.

8. Hadamard matrices and optical multiplexing

The connection between Hadamard designs and multiplexing optics is now straightforward. In the optical case

the unknowns w_i represent intensities of individual spatial and/or spectral elements in a beam of radiation. In contrast to scanning instruments which measure the intensities one at a time, the multiplexing optical system measures (i.e. weighs) several intensities (or w_i's) simultaneously. The y_i's now represent the readings of the detector (instead of the reading of the balance). Finally, the weighing design itself, X, is represented by a mask. More precisely, one row of X, which specifies which objects are present in a single weighing, corresponds to the row of transmitting, absorbing or reflecting elements. We usually refer to such a row as a mask configuration.

The two types of weighing designs – chemical and spring balance designs – are realized by masks which contain either transmitting, absorbing and reflecting elements (for the chemical balance design) or simply open and close slots (for the spring balance design). Note that the former case requires two detectors, whereas in the latter case the reference detector can be omitted. In Hadamard transform spectrometry the separated light is sent to a mask. Various parts of the mask will be clear, allowing the light to pass through, reflective (sending light to a secondary detector), or opaque. Let us represent clear, reflective and opaque by $1,\ -1$ respectively. Then the configuration of the mask is represented by a sequence of elements $1,\ -1$.

Suppose k measurements are to be made, and suppose it is convenient to measure the intensity of light at n points of the spectrum. Then the experiment will involve k masks, which can be thought of as $n \times k$ matrix of entries 1, and -1. The efficiency of the experiment is the same as the efficiency of the matrix as a weighing design. The best systems of mask are thus derived from Hadamard matrices.

9. Screening properties of Hadamard matrices

An array on two symbols with N rows and k columns is a (N, k, p) screening design if for each choice of p columns, each of the 2^p row vectors appears at least once. Screening designs are useful for situations where a large number of factors (q) is examined but only few (k) of these are expected to be important.

Screening designs that arise from Hadamard matrices have traditionally been used for identifying main effects only, because of their complex aliasing structures. Without loss of generality we can insist that the first column of any Hadamard matrix contain only 1's. Then, by removing this column we obtain a $(N, N-1, p)$ screening design, with $p \geq 2$. Some screening designs of this form were introduced by Plackett and Burman [13] and they are termed as *Plackett-Burman* designs. These designs can be generated from the first row, that consists of $N-1$ elements, by cyclic arrangement. The second row is generated by removing all the entries of the first row one position to the right and placing the last element in the first position. The third row is generated from the second row with the same procedure, and the process stops when $N-1$ rows are generated. A row of -1's is then added as the last row, completing the design with N runs and $N-1$ columns. By adding a column of all 1's in a Plackett and Burman design with N runs we obtain a Hadamard matrix of order N. In fact, Plackett and Burman constructed Hadamard matrices of order N, for all $N \leq 100$ except $N = 92$ which was later given by Baumert, Colomb and Hall in 1962. For more details see [16]. As an example, the first rows that generate the Plackett and Burman designs with $N = 8,\ 12,\ 16,\ 20$ and 24 runs are given below.

```
 8  + + + − + − −
12  + + − + + + − − − + −
16  + + + + − + − + + − − + − − −
20  + + − − + + + + − + − + − − − − + + −
24  + + + + + − + − + + − − + + − − + − + − − − −
```

After the identification of the active factors, the original design is then projected into k dimensions for further analysis, that is, we select the columns that correspond to the active factors to form a new design with N runs and k columns which is called a *projection*.

Since the choice of k columns varies with the outcome of the analysis, it is desired to study the properties of all projection designs that may arise.

Projection designs that arise from Hadamard matrices are either regular or non-regular factorial plans. Regular fractional factorial designs, have simple aliasing structures and usually arise from Hadamard matrices of orders $N = 2^p$; non-regular fractional factorial designs have complex aliasing structures.

The aliasing structure of regular factorial designs can easily be computed. On the other hand, the alias structure of non-regular designs cannot easily be computed. For more details on fractional factorial designs and screening experiments we refer the interested reader to Wu and Hamada [20].

10. Supersaturated designs

Supersaturated designs are useful in situations in which the number of active factors is very small compared to the total number of factors being considered.

The use of Hadamard matrices to construct supersaturated designs that can examine $k = N-2$ factors in $n = N/2$ runs, where N is the order of the normalized Hadamard matrix used. The first column of all 1's is not taken into consideration since it is fully aliased with the mean. Then, we choose a *branching column* out of the remaining $N-1$ columns and we split the N runs into two groups. Group I contains all the runs with the sign $+1$ in the branching column and Group II contains the remaining runs. Then by deleting the branching column either from Group I or Group II causes the remaining $N-2$ columns to form a super saturated design to examine $k = N-2$ factors in $N/2$ runs.

11. Edge designs

These designs allow a model-independent estimate of the set of relevant variables, thus providing more robustness than traditional designs. They use a construction known as skew-Hadamard matrices.

If the first row and first column of C is removed, a $(N-1)\times(N-1)$ matrix S is obtained to be used in the form

$$X = \begin{pmatrix} \mathbf{1} & S+I_{N-1} \\ \mathbf{1} & S-I_{N-1} \end{pmatrix}$$

in order to obtain the resulting edge design, where $\mathbf{1}^T = (1,1,\ldots,1)$ is a $1\times(N-1)$ vector with all entries equal to 1.

12. Idle column method

This uses Hadamard matrices of order $N = 8t$ to construct multi-level idle column arrays.

Let $H_{N/2}$ be a normalized Hadamard matrix of order $N/2$. We can denote this matrix by $H_{N/2} = (\mathbf{1}, \mathbf{C}_1, \ldots, \mathbf{C}_{N/2-1})$. Then it is well known that

$$\begin{aligned} H_N &= \begin{pmatrix} H_{N/2} & H_{N/2} \\ H_{N/2} & -H_{N/2} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{1}\ \mathbf{C}_1 \cdots \mathbf{C}_{N/2-1} & \mathbf{1} & \mathbf{C}_1 \cdots & \mathbf{C}_{N/2-1} \\ \mathbf{1}\ \mathbf{C}_1 \cdots \mathbf{C}_{N/2-1} & -\mathbf{1} & -\mathbf{C}_1 \cdots & -\mathbf{C}_{N/2-1} \end{pmatrix} \end{aligned}$$

is a Hadamard matrix of order N.

Remove the first column of H_N and by treating the column $(\mathbf{1}^T, -\mathbf{1}^T)^T$ as the idle column, the product of columns $(\mathbf{C}_i^T, \mathbf{C}_i^T)^T$ and $(\mathbf{C}_i^T, -\mathbf{C}_i^T)^T$ for $1 \le i \le \frac{N}{2} - 1$, equals to the idle column. Then, for the level combinations of the two columns in a pair, the recoding scheme

$$\begin{aligned} (-1,-1) &\longrightarrow -1 \\ (-1,1) &\longrightarrow 0 \\ (1,-1) &\longrightarrow 1 \\ (1,1) &\longrightarrow 0 \end{aligned}$$

is used to construct a three level column.

References

[1] W. T. Federer, A. Hedayat, E. T. Parker, B. L. Raktoe, E. Seiden, and R. J. Turyn, "Some techniques for constructing mutually orthogonal latin squares", MRC Technical Summary Report no. 1030, Mathematics Research Centre University of Wisconsin, June 1971. (A preliminary version of this report appeared in the proceedings of the Fifteenth Conference on the Design of Experiments in Army Research Development and Testing, ARO-D Report 70-2, July 1970, The Office of Chief of Research and Development, Durham, North Carolina).

[2] W. T. Federer, "On the existence and construction of a complete set of orthogonal $F(4t;2t,2t)$-squares", Paper no. BU-564-M in the Biometrics Unit Mimeo Series, Department of Plant Breeding and Biometry, Cornell University, Ithica, New York, 14853, 1975.

[3] A. V. Geramita and J. Seberry, *Orthogonal Designs: Quadratic Forms and Hadamard Matrices*. New York-Basel: Marcel Dekker, 1979.

[4] J. Hadamard, "Resolution d'une question relative aux determinants", *Bull. Sci. Math.*, vol. 17, pp. 240–246, 1993.

[5] A. Hedayat, "On the theory of the existence, non-existence and the construction of mutually orthogonal F-squares and latin squares". Ph.D. dissertation, Cornell University, June 1969.

[6] A. Hedayat and E. Seiden, "F-square and orthogonal F-square design: A generalization of Latin square and orthogonal Latin squares design", *Ann. Math. Stat.*, vol. 41, pp. 2035–2044, 1970.

[7] C. Koukouvinos and J. Seberry, "New weighing matrices and orthogonal designs constructed using two sequences with zero autocorrelation function – a review", *J. Stat. Plan. Infer.*, vol. 81, pp. 153–182, 1999.

[8] J. Kiefer, "Construction and optimality of generalized Youden designs", in: *Statistical Design and Linear Models*, J. N. Srivastava, Ed. Amsterdam: North-Holland, 1975, pp. 333–353.

[9] A. W. Lam and S. Tantaratana, "Theory and applications of spread-spectrum systems", IEEE/EAB Self-Study Course, IEEE Inc., Piscataway, 1994.

[10] V. I. Levenshtein, "A new lower bound on aperiodic crosscorrelation of binary codes", in *4th Int. Symp. Commun. Theory & Appl., ISCTA'97*, 1997, pp. 147–149.

[11] J. P. Mandeli, "Complete sets of orthogonal F-squares". M.Sc. thesis, Cornell University, Aug. 1975.

[12] I. Oppermann and B. S. Vucetic, "Complex spreading sequences with a wide range of correlation properties", *IEEE Trans. Commun.*, vol. 45, pp. 365–375, 1997.

[13] R. L. Plackett and J. P. Burman, "The design of optimum multifactorial experiments", *Biometrika*, vol. 33, pp. 305–325, 1946.

[14] J. Seberry and R. Craigen, "Orthogonal designs", in *Handbook of Combinatorial Designs,* C. J. Colbourn and J. H. Dinitz, Eds. CRC Press, 1996, pp. 400–406.

[15] J. Seberry, B. J. Wysocki, and T. A. Wysocki, "Golay sequences for DS CDMA applications", in *Sixth Int. Symp. DSP Commun. Syst. DSPCS'02*, Manly, TITR, Wollongong, Jan. 2002, pp. 103–108.

[16] J. Seberry and M. Yamada, "Hadamard matrices, sequences and designs", in *Contemporary Design Theory – a Collection of Surveys*, D. J. Stinson and J. Dinitz, Eds. Wiley, 1992, pp. 431–560.

[17] J. J. Sylvester, "Thoughts on inverse orthogonal matrices, simultaneous sign successions, and tesselated pavements in two or more colours, with applications to Newton's rule, ornamental tile-work, and the theory of numbers", *Phil. Mag.*, vol. 34, pp. 461–475, 1967.

[18] L. R. Welch, "Lower bounds on the maximum cross-correlation of signals", *IEEE Trans. Inform. Theory*, vol. 20, pp. 397–399, 1974.

[19] J. Seberry Wallis, "Part IV of combinatorics: Room squares, sum-free sets and Hadamard matrices", *Lecture Notes in Mathematics*, W. D. Wallis, A. Penfold Street, and J. Seberry Wallis, Eds. Berlin-Heidelberg-New York: Springer, 1972, vol. 292.

[20] C. F. J. Wu and M. Hamada, *Experiments, Planning, Analysis, and Parameter Design Optimization*. New York: Wiley, 2000.

Haralambos Evangelaras received the B.Sc. degree in mathematics from the University of Athens, Athens, Greece, in 1998. He is currently a Ph.D. student at the National Technical University of Athens in the area of statistics.

Department of Mathematics
National Technical University of Athens
Zografou 15773
Athens, Greece

Christos Koukouvinos received the B.Sc. degree in mathematics and the Ph.D. in statistics, both from the University

of Thessaloniki, Thessaloniki, Greece, in 1983 and 1988, respectively. In 1996, he received the Hall medal (a research award) from the ICA. He is a fellow of the ICA and since 2001 an elected member of the Council of the ICA. He is in the Editorial Board for the *Australasian Journal of Combinatorics*, and the *Journal of Modern Applied Statistical Methods*. He is currently a Professor at the National Technical University of Athens. His research interests include statistics, combinatorics, matrix theory, coding theory and computational mathematics.

Department of Mathematics
National Technical University of Athens
Zografou 15773, Athens, Greece

Jennifer Seberry received her B.Sc.(Hons), M.Sc. in mathematics from the University of NSW and her M.Sc. and Ph.D. in mathematics from La Trobe University, Victoria, Australia. For the past twent years she has also been working in cryptography and computer security. In 1987 she founded the Centre for Computer Security Research. She is presently Professor of Computer Science at the University of Wollongong. She has over 300 publication and has successfully supervised 22 Ph.D. theses.

School of IT and Computer Science
University of Wollongong
Wollongong, NSW, 2522
Australia

Paper

Design of filterbank transceivers for dispersive channels with arbitrary-length impulse response

Alfred Mertins

Abstract — This paper addresses the joint design of transmitter and receiver for multichannel data transmission over dispersive channels. The transmitter is assumed to consist of FIR filters and the channel impulse response is allowed to have an arbitrary length. The design criterion is the maximization of the information rate between transmitter input and receiver output under the constraint of a fixed transmit power. A link to minimum mean squared error designs for a similar setting is established. The proposed algorithm allows a straightforward transmitter design and generally yields a near-optimum solution for the transmit filters. Under certain conditions, the exact solution for the globally optimal transmitter is obtained.

Keywords — joint transmitter/receiver design, filterbanks, information rate maximization, dispersive channels.

1. Introduction

The joint design of transmitter and receiver for data transmission over dispersive channels has attracted numerous researchers, as it has the potential to yield very high throughput without the need of costly algorithms on the receiver side, such as maximum likelihood sequence estimation with the Viterbi algorithm.

The process of shaping the transmit signal and/or introducing redundancy based on the knowledge of the channel is also known as precoding. Salz [1] provided a first solution to the joint transmit/receive filter design problem, but it required the filters to have support within the first Nyquist zone $[-1/2T, 1/2T]$. Yang and Roy proposed an algorithm for the design of precoders that use excess bandwidth to introduce redundancy [2]. Their method required an iteration to find the optimum solution. Xia studied the existence of redundant precoders that allow a perfect inversion of FIR channels with FIR receivers [3]. The effects of noise were not considered in [3].

Direct solutions to the joint design problem for the case of block transforms with a sufficiently long guard interval to avoid interblock interference (IBI) were provided in [4–6]. The optimality criteria considered in [4] were the zero forcing (ZF) and minimum mean squared error (MMSE) criteria. In [5] and [6] the maximization of mutual information between transmitter and receiver was studied, using results derived in [7]. A drawback of the block transforms of [4–6] is that the length of the guard interval needs to be at least equal to the channel order. This is the same problem as with the well-known DMT and OFDM techniques [8, 9]. To cope with longer channel impulse responses one can increase the length of the guard interval, but this will decrease the efficiency, as less data symbols can be transmitted. Increasing both the length of the guard interval and the number of subchannels allows one to maintain a desired bandwidth efficiency, but this strategy also has its limits, because the delay between transmitter and receiver may become unacceptably high.

Li and Ding provided a direct solution to the problem of minimizing the mean squared error (MSE) between transmitter input and receiver output under the power constraint for arbitrary channel lengths with overlapping blocks [10]. However, their solution generally yields IIR transmit filters, which restricts the practical use of their exact solution. An FIR approximation of the technique in [10] was provided in [11]. Finally, transmitter design methods for the case where decision feedback receivers are employed have been proposed in [7, 12, 13].

This paper addresses the design of FIR precoders for the case where the channel impulse response has arbitrary length. Note that this configuration is of a significant interest for practical applications, because real-world channel impulse responses may become extremely long and the use of sufficiently long guard intervals, as required for DMT, OFDM, or the methods in [4–6], may be prohibitive due to delay constraints. During transmitter optimization an approximation is used that allows us to simplify the objective function and obtain a straightforward solution. For $L \leq N - M$, where L is the channel order, M is the number of subchannels, and N is the upsampling factor in the transmitter, the algorithm yields the exact optimum solutions of [5, 6], and for $L > N - M$ it leads to near optimum solutions.

The paper is organized as follows. Section 2 describes the input-output relationships of the considered transmit/receive system. Section 3 then addresses the maximization of the information rate through the choice of optimal transmit and receive filters. Also a link to MMSE designs for similar settings is established. Section 4 demonstrates the properties of the proposed algorithm in several examples, and finally Section 5 gives some conclusions.

Notation. Vectors and matrices are printed in boldface. The superscripts $\{\cdot\}^T$, $\{\cdot\}^H$, $\{\cdot\}^+$ denote transposition,

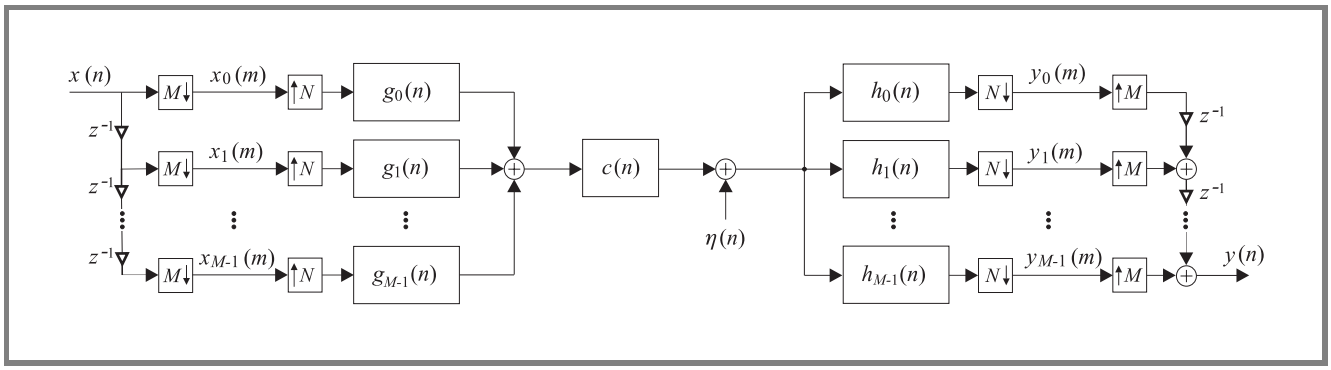


Fig. 1. Redundant precoder.

Hermitian transposition, and the pseudoinverse, respectively. The determinant and trace of a matrix are denoted as $|\cdot|$ and $\mathrm{tr}\{\cdot\}$, respectively. $E\{\cdot\}$ is the expectation operation.

2. System description

A block diagram of the considered system is depicted in Fig. 1. The input stream $x(m)$ is split into M parallel streams which are then upsampled by a factor of $N \geq M$ and fed into the M transmit filters with impulse responses $g_k(n)$, $k = 0, 1, \ldots, M-1$. The channel is described by its impulse response $c(n)$ and an additive, data independent, zero-mean, stationary, Gaussian noise process $\eta(n)$. The receive signal is filtered with the analysis filters $h_k(n)$, $k = 0, 1, \ldots, M-1$ and subsampled by N to yield the parallel output data $y_k(m)$. Finally, a parallel-to-serial conversion yields the output sequence $y(n)$.

For further analysis it is advantageous to decompose the filters into their polyphase components and to describe the system as a multiple-input multiple-output (MIMO) system as depicted in Fig. 2. The input vector at time m is given by $\boldsymbol{x}(m) = [x_0(m), x_1(m), \ldots, x_{M-1}(m)]^T$ with $x_k(m) = x(mM-k)$. Accordingly, the output process $\boldsymbol{y}(m)$ is defined as $\boldsymbol{y}(m) = [y_0(m), \ldots, y_{M-1}(m)]^T$. The transmit filter bank can be described via its $N \times M$ polyphase matrix [14]

$$\boldsymbol{G}(z) = \begin{bmatrix} G_{00}(z) & \cdots & G_{M-1,0}(z) \\ \vdots & & \vdots \\ G_{0,N-1}(z) & \cdots & G_{M-1,N-1}(z) \end{bmatrix} \tag{1}$$

where $G_{k,\ell}(z)$ is the ℓth polyphase component of the kth transmit filter, given by

$$G_{k,\ell}(z) = \sum_n g_k(nN+\ell)\, z^{-n}. \tag{2}$$

Alternatively, $\boldsymbol{G}(z)$ may be expressed as $\boldsymbol{G}(z) = \sum_n \boldsymbol{G}_n z^{-n}$ with $[\boldsymbol{G}_n]_{\ell,k} = g_k(nN+\ell)$ where $[\boldsymbol{G}_n]_{\ell,k}$ denotes the element of $[\boldsymbol{G}_n]$ at position ℓ, k.

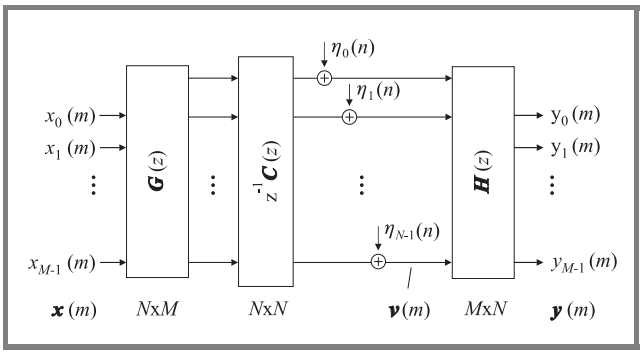


Fig. 2. Redundant precoder in polyphase (MIMO) representation.

The polyphase matrix of the receiver filter bank is given by

$$\begin{aligned} \boldsymbol{H}(z) &= \sum_n \boldsymbol{H}_n z^{-n} = \\ &= \begin{bmatrix} H'_{00}(z) & \cdots & H'_{0,N-1}(z) \\ \vdots & & \vdots \\ H'_{M-1,0}(z) & \cdots & H'_{M-1,N-1}(z) \end{bmatrix} \end{aligned} \tag{3}$$

with

$$\begin{aligned} H'_{k,\ell}(z) &= \sum_n h_k(nN+N-1-\ell)\, z^{-n}, \\ [\boldsymbol{H}_n]_{k,\ell} &= h_k(nN+N-1-\ell). \end{aligned} \tag{4}$$

The channel can be described via the pseudo-circulant $N \times N$ matrix

$$\boldsymbol{C}(z) = \begin{bmatrix} C_0(z) & z^{-1}C_{N-1}(z) & \cdots & z^{-1}C_1(z) \\ C_1(z) & C_0(z) & \cdots & z^{-1}C_2(z) \\ \vdots & & \ddots & \vdots \\ C_{N-1}(z) & C_{N-2}(z) & \cdots & C_0(z) \end{bmatrix} \tag{5}$$

with $C_\ell(z) = \sum_n c(nN+\ell)\, z^{-n}$. Alternatively, $\boldsymbol{C}(z)$ can be written as a polynomial of matrices:

$$\boldsymbol{C}(z) = \sum\nolimits_k z^{-k}\, \boldsymbol{C}_k. \tag{6}$$

The often desired (zero forcing) property

$$\boldsymbol{y}(n) = \boldsymbol{x}(n-n_0) \tag{7}$$

is obtained in the noise free case if $\boldsymbol{H}(z)$ and $\boldsymbol{G}(z)$ are chosen such that the perfect reconstruction (PR) condition

$$\boldsymbol{H}(z)\ \boldsymbol{C}(z)\ \boldsymbol{G}(z) = z^{-n_0+1}\boldsymbol{I}_{M\times M} \tag{8}$$

holds. Conditions to satisfy (7) for a given channel $c(n)$ are for example discussed in [3, 4].

3. Maximizing information rate

In this section, we address the problem of maximizing the information rate through the choice of the transmit and receive filters. We will first consider a straightforward matrix model, similar to block transforms, and will show for this model that the mutual information can be expressed via the error covariance matrix of MMSE receive filters. Using this fact, an algorithm for determining optimal FIR transmit filters is presented.

3.1. A general expression for mutual information

The mutual information between a block of input symbols, $\boldsymbol{x}$, and a block of output symbols, $\boldsymbol{y}$, of a transceiver is defined as $I_0(\boldsymbol{x};\boldsymbol{y}) = H(\boldsymbol{x}) - H(\boldsymbol{x}|\boldsymbol{y})$ where $H(\boldsymbol{x})$ is the entropy of $\boldsymbol{x}$ and $H(\boldsymbol{x}|\boldsymbol{y})$ is the conditional entropy of $\boldsymbol{x}$ given $\boldsymbol{y}$ [15]. We define a normalized mutual information as

$$I(\boldsymbol{x};\boldsymbol{y}) = \frac{1}{N}[H(\boldsymbol{x}) - H(\boldsymbol{x}|\boldsymbol{y})] \tag{9}$$

where N is the upsampling factor in Fig. 1. The length of $\boldsymbol{x}$ is M with $M \le N$, and the length of $\boldsymbol{y}$ will be defined as needed. It is known that $I(\boldsymbol{x};\boldsymbol{y})$ becomes maximal if $\boldsymbol{x}$ is Gaussian [15], and therefore we will assume Gaussian processes henceforth. For this case it was shown in [7] that

$$I(\boldsymbol{x};\boldsymbol{y}) = \frac{1}{N}\log_2\left(\frac{|\boldsymbol{R}_{xx}|}{|\boldsymbol{R}^{\perp}_{x|y}|}\right) \tag{10}$$

with

$$\boldsymbol{R}^{\perp}_{x|y} = \boldsymbol{R}_{xx} - \boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx} \tag{11}$$

and $\boldsymbol{R}_{xx} = E\{\boldsymbol{x}\boldsymbol{x}^H\}$, $\boldsymbol{R}_{xy} = \boldsymbol{R}_{yx}^H = E\{\boldsymbol{x}\boldsymbol{y}^H\}$, $\boldsymbol{R}_{yy} = E\{\boldsymbol{y}\boldsymbol{y}^H\}$. We now consider the model

$$\boldsymbol{y} = \boldsymbol{H}[\boldsymbol{C}\boldsymbol{G}\boldsymbol{x} + \boldsymbol{n}] \tag{12}$$

where the matrices $\boldsymbol{G},\boldsymbol{C},\boldsymbol{H}$ describe the transmitter, channel, and receiver, respectively, and vector $\boldsymbol{n}$ describes additive noise. At this point, no assumptions are made about the size of vectors and matrices in (12) and the type of noise. With (12) one obtains for $\boldsymbol{R}^{\perp}_{x|y}$

$$\boldsymbol{R}^{\perp}_{x|y} = \boldsymbol{R}_{xx} - \boldsymbol{R}_{xx}\boldsymbol{G}^H\boldsymbol{C}^H\boldsymbol{H}^H\Big[\boldsymbol{H}(\boldsymbol{C}\boldsymbol{G}\boldsymbol{R}_{xx}\boldsymbol{G}^H\boldsymbol{C}^H + \boldsymbol{R}_{nn})\boldsymbol{H}^H\Big]^{-1}\boldsymbol{H}\boldsymbol{C}\boldsymbol{G}\boldsymbol{R}_{xx}, \tag{13}$$

with $\boldsymbol{R}_{nn} = E\{\boldsymbol{n}\boldsymbol{n}^H\}$. By using the pseudoinverse of $\boldsymbol{R}^{\perp}_{x|y}$ given by

$$(\boldsymbol{R}^{\perp}_{x|y})^+ = \boldsymbol{R}^+_{xx} + \boldsymbol{G}^H\boldsymbol{C}^H\boldsymbol{H}^H[\boldsymbol{H}\boldsymbol{R}_{nn}\boldsymbol{H}^H]^{-1}\boldsymbol{H}\boldsymbol{C}\boldsymbol{G} \tag{14}$$

the quantity $I(\boldsymbol{x};\boldsymbol{y})$ can be alternatively expressed as

$$I(\boldsymbol{x};\boldsymbol{y}) = \frac{1}{N}\log_2\left|\boldsymbol{R}_{xx}\,(\boldsymbol{R}^{\perp}_{x|y})^+\right|. \tag{15}$$

Note that the expression (14) for $(\boldsymbol{R}^{\perp}_{x|y})^+$ includes the shaping of the transmit signal with matrix $\boldsymbol{G}$ and the influence of the receive filters in matrix $\boldsymbol{H}$. A similar expression for mutual information has been derived in [7], but for the simpler model $\boldsymbol{y} = \boldsymbol{C}\boldsymbol{x} + \boldsymbol{n}$ with $\boldsymbol{n}$ being white noise. Using the results of [7] and a model similar to (12), but without possible interblock interference, a related expression has also been obtained in [5].

3.2. Incorporating the filterbank model

Now let the model (12) describe the filterbank transceiver of Section 2 with $\boldsymbol{x} := \boldsymbol{x}(m)$ and $\boldsymbol{y} := \boldsymbol{y}(m-n_0)$. The columns of matrix $\boldsymbol{G}$ are the transmit filter impulse responses, and the channel matrix $\boldsymbol{C}$ has the structure

$$\boldsymbol{C} = \begin{bmatrix} c(0) & 0 & 0 & 0 & \dots & 0 \\ c(1) & c(0) & 0 & 0 & \dots & 0 \\ c(2) & c(1) & c(0) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \end{bmatrix} \tag{16}$$

The size of $\boldsymbol{C}$ depends on the lengths of the transmit filters and the channel. $\boldsymbol{C}$ may even be of infinite dimension, and similarly, the vector $\boldsymbol{v} = \boldsymbol{C}\boldsymbol{G}\boldsymbol{x} + \boldsymbol{n}$ observed at the channel output may be of infinite length. However, both $\boldsymbol{x}$ and $\boldsymbol{y}$ are of length M. The noise process $\boldsymbol{n}$ contains the additive channel noise and the IBI from other data blocks.

In the following we show that the optimal receive matrix $\boldsymbol{H}$ has the structure

$$\boldsymbol{H} = \boldsymbol{X}\boldsymbol{G}^H\boldsymbol{C}^H\boldsymbol{R}_{nn}^{-1} \tag{17}$$

with an arbitrary, full-rank $M \times M$ matrix $\boldsymbol{X}$. Depending on $\boldsymbol{X}$ one obtains, for example, the ZF or MMSE receive

filters. Inserting (17) into (14) and rearranging the obtained expression yields

$$(\boldsymbol{R}_{x|y}^{\perp})^{+} = \boldsymbol{R}_{xx}^{+} + \boldsymbol{G}^H \boldsymbol{C}^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{C}\boldsymbol{G}. \tag{18}$$

Note that (18) is independent of $\boldsymbol{X}$. Obviously, $(\boldsymbol{R}_{x|y}^{\perp})^{+}$ according to (18) is the same as the matrix $(\boldsymbol{R}_{x|v}^{\perp})^{+}$, which relates to the conditional entropy $H(\boldsymbol{x}|\boldsymbol{v})$ based on the observation $\boldsymbol{v}$. Because of $H(\boldsymbol{x}|\boldsymbol{y}) \leq H(\boldsymbol{x}|\boldsymbol{v})$, we can conclude that any matrix $\boldsymbol{H}$ of the form (17) maximizes the mutual information. Thus, due to the structure of $\boldsymbol{H}$ in (17) this means that the optimal receive filters are "matched filters", given by the term $\boldsymbol{G}^H \boldsymbol{C}^H \boldsymbol{R}_{nn}^{-1}$, followed by an arbitrary, full-rank matrix operation $\boldsymbol{X}$. Through the choice of $\boldsymbol{X}$ one can obtain, for example, the optimal zero forcing and MMSE solutions.

Interestingly, the matrix $(\boldsymbol{R}_{x|y}^{\perp})$ is the same as the error correlation matrix

$$\boldsymbol{R}_{x|y}^{\perp} := \boldsymbol{R}_{ee} = E\left\{(\boldsymbol{y}-\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})^H\right\}$$

for the case of linear MMSE estimation of $\boldsymbol{x}$ from the noisy observation $\boldsymbol{v}$.[1] This observation has also been made in [7]. For the filterbank transceivers considered in this papers it means that we can concentrate on minimizing the determinant of the error correlation matrix in the presence of an MMSE receive filterbank. To simplify the notation we assume white channel noise with variance σ_η^2 and white data $x(n)$ with variance σ_x^2. The incorporation of nonwhite data and noise processes is straightforward.

For further derivations, the expression (18) for $(\boldsymbol{R}_{x|y}^{\perp})^{+}$ is not very convenient, as it contains the inverse correlation matrix of the noise which is comprised of channel noise and IBI. Knowing that we need the error correlation matrix of MMSE estimation we can alternatively use the expression obtained in [11] for MMSE precoders:

$$\boldsymbol{R}_{ee} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \sigma_x^2 \left[\boldsymbol{I}_{M\times M} + \right.$$
$$\left. + \frac{\sigma_x^2}{\sigma_\eta^2} \boldsymbol{G}^H(e^{j\omega}) \left[\sum\nolimits_k \boldsymbol{\mathcal{R}}_{cc}(k) e^{-j\omega k}\right] \boldsymbol{G}(e^{j\omega}) \right]^{-1} d\omega \tag{19}$$

where

$$\boldsymbol{\mathcal{R}}_{cc}(k) = \sum\nolimits_\ell \boldsymbol{C}_\ell^H \boldsymbol{C}_{\ell+k}. \tag{20}$$

3.3. Using FIR transmit filters

To minimize the transmitter complexity and system delay, we assume transmit filters of length N where N is the upsampling factor in Fig. 1. For this filter length we have $\boldsymbol{G}(z) = \boldsymbol{G}_0$ and obtain

$$\boldsymbol{R}_{ee} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \sigma_x^2 \left[\boldsymbol{I}_{M\times M} + \right.$$
$$\left. + \frac{\sigma_x^2}{\sigma_\eta^2} \boldsymbol{G}_0^H \left[\sum\nolimits_k \boldsymbol{\mathcal{R}}_{cc}(k) e^{-j\omega k}\right] \boldsymbol{G}_0 \right]^{-1} d\omega. \tag{21}$$

The next step is to approximate (21) by a simpler expression. Because the summation terms for $k \neq 0$ in (21) relate to IBI we choose $\boldsymbol{G}_0$ from a subspace such that the terms $\boldsymbol{G}_0^H \boldsymbol{\mathcal{R}}_{cc}(k) \boldsymbol{G}_0$ for $k \neq 0$ become so small that they can be neglected in (21). To determine a suitable subspace for the choice of $\boldsymbol{G}_0$ we employ an iterative procedure based on the singular value decomposition (svd). We do not explicitly formulate a basis for the required subspace, and rather consider a projection $\boldsymbol{P}$ that projects onto the required subspace.

The algorithm is as follows:

Step 1: Let $\boldsymbol{P} = \boldsymbol{I}_{N\times N}$.

Step 2: Compute the svd's

$$\boldsymbol{A}_k \boldsymbol{\Sigma}_k \boldsymbol{B}_k^H = \boldsymbol{P}^H \boldsymbol{\mathcal{R}}_{cc}(k) \boldsymbol{P}$$

for all $k \neq 0$ for which $\boldsymbol{\mathcal{R}}_{cc}(k) \neq \boldsymbol{0}$.

Step 3: Determine the largest singular value for $k \neq 0$ and denote it as σ_{max}. Assuming that σ_{max} is contained in matrix $\boldsymbol{\Sigma}_K$ denote the corresponding column of $\boldsymbol{A}_K$ as $\boldsymbol{a}$.

Step 4: If $\mathrm{rank}(\boldsymbol{P}) > M$ and $\sigma_{max} > 0$ set

$$\boldsymbol{P} := [\boldsymbol{I}_{N\times N} - \boldsymbol{a}\boldsymbol{a}^H]\boldsymbol{P}$$

and go back to Step 2. Otherwise, end the algorithm.

When incorporating the projection matrix $\boldsymbol{P}$, the error correlation matrix can be approximated by

$$\tilde{\boldsymbol{R}}_{ee} = \sigma_x^2 \left[\boldsymbol{I}_{M\times M} + \frac{\sigma_x^2}{\sigma_\eta^2} \boldsymbol{G}_0^H \boldsymbol{P}^H \boldsymbol{\mathcal{R}}_{cc}(0) \boldsymbol{P} \boldsymbol{G}_0\right]^{-1}, \tag{22}$$

and the normalized mutual information can thus be approximated as

$$\bar{I}(\boldsymbol{x};\boldsymbol{y}) = \frac{1}{N}\log_2(|\boldsymbol{M}|) \tag{23}$$

with

$$\boldsymbol{M} = \left[\boldsymbol{I}_{M\times M} + \frac{\sigma_x^2}{\sigma_\eta^2} \boldsymbol{G}_0^H \boldsymbol{P}^H \boldsymbol{\mathcal{R}}_{cc}(0) \boldsymbol{P} \boldsymbol{G}_0\right]. \tag{24}$$

According to Hadamard's inequality [15], $\boldsymbol{M}$ must be diagonal in order to maximize $|\boldsymbol{M}|$ under the transmit power constraint

$$\sigma_x^2 \,\mathrm{tr}\left\{\boldsymbol{G}_0 \boldsymbol{G}_0^H\right\} = N P_0. \tag{25}$$

[1] Introductions to linear estimation theory can be found in [16].

This means that the columns of $\boldsymbol{G}_0$ have to be scaled eigenvectors of $\boldsymbol{P}^H\mathfrak{R}_{cc}(0)\boldsymbol{P}$. We now consider the eigendecompositions

$$\boldsymbol{P}^H\mathfrak{R}_{cc}(0)\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^H \tag{26}$$

and

$$\boldsymbol{G}_0\boldsymbol{G}_0^H = \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^H \tag{27}$$

with

$$\boldsymbol{\Lambda} = \mathrm{diag}\left[\lambda_1, \ldots, \lambda_N\right] \tag{28}$$

and

$$\boldsymbol{Q} = \mathrm{diag}\left[q_1, \ldots, \sigma_N\right] \tag{29}$$

where the eigenvalues λ_i are assumed to be sorted such that $\lambda_i \geq \lambda_i + 1$. Note that some of the eigenvalues λ_i may be zero and that only the first M values $q_1, \ldots, q_M$ are nonzero. Using (26) and (27) the mutual information $\bar{I}(\boldsymbol{x};\boldsymbol{y})$ according to (23) and (24) can be rewritten as

$$\bar{I}(\boldsymbol{x};\boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{M}\log_2(1+\frac{\sigma_x^2}{\sigma_\eta^2}\lambda_i q_i). \tag{30}$$

A standard Lagrange optimization, similar to [5, 7], yields

$$q_i = \max(c - \frac{\sigma_\eta^2}{\sigma_x^2\lambda_i}, 0) \tag{31}$$

where c is to be determined from the power constraint (25). As one can see in (31), the optimal values q_i obey the waterpouring distribution. Assuming that M is chosen such that q_i, $i = 1, \ldots, M$ are nonzero, the transmit filters finally become

$$\boldsymbol{G}_0 = \bar{\boldsymbol{U}}\,\mathrm{diag}\left[\sqrt{q_1}, \ldots, \sqrt{q_M}\right] \tag{32}$$

where $\bar{\boldsymbol{U}}$ contains the M eigenvectors that belong to the largest eigenvalues $\lambda_1, \ldots, \lambda_M$. A comparison with the solution in [11] shows that maximizing the information rate and minimizing the overall MSE leads to the same transmit filters, but with different power loading factors q_i, $i = 1, \ldots, M$. Moreover, it is straightforward to show that if the channel order L is smaller or equal to $N - M$ we have $\bar{I}(\boldsymbol{x};\boldsymbol{y}) = I(\boldsymbol{x};\boldsymbol{y})$, and the proposed algorithm yields the solutions of [5, 6].

4. A design example

We demonstrate the performance of the precoder design algorithm using a simple example where significant IBI between adjacent data blocks occurs. The chosen parameters are $L = 6$, $N = 16$, $M = 14$, and the E_b/N_0 ratio at the receiver input is set to 20 dB. The channel impulse response is $c(n) = [1, 1, 1, 1, 1, 1, 1]$. Note that all channel zeros lie on the unit circle of the z-plane. The frequency response of the channel is depicted in Fig. 3, together with the transmit power spectra for the following two precoder design methods: (i) the MMSE precoder of [11] and (ii) the precoder maximizing information rate proposed in this paper. The comparison between the two power spectra shows that the MMSE precoder tends to spend power in frequency bands where the channel gain is low, whereas the precoder maximizing information rate reduces the transmit power for such frequencies.

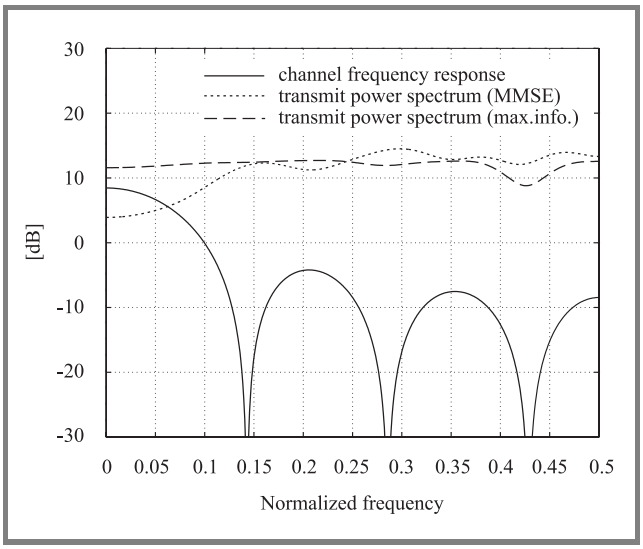


Fig. 3. Channel frequency response and transmit power spectra.

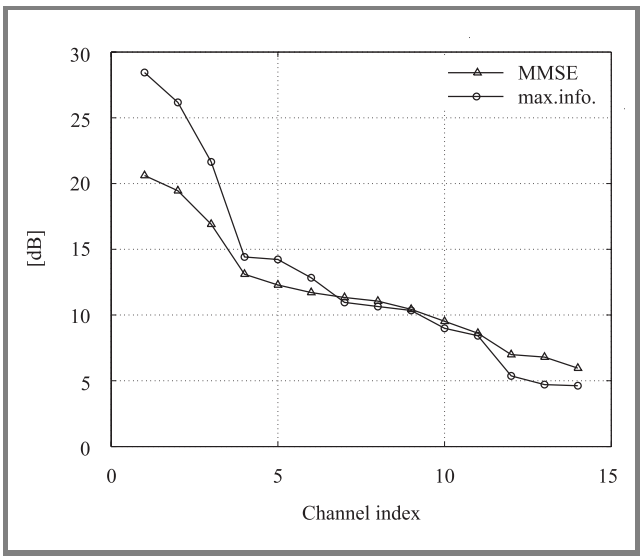


Fig. 4. Signal to noise ratios in subchannels at the receiver output.

Figure 4 shows the obtained SNR's at the receiver output for the two design methods. One can see that maximizing the information rate yields several subchannels with very good SNR and a few with poor SNR. The MMSE design, on the other hand, tries to uphold all SNR's in order to minimize

the MSE. The obtained normalized information rates are 3.49 bit/symbol for the MMSE design and 3.84 bit/symbol when maximizing $\overline{I}(\boldsymbol{x},\boldsymbol{y})$.
When reducing the number of subchannels to $M = 10$, all IBI vanishes, and the design method becomes equivalent to the ones in [5, 6]. However, the maximum normalized mutual information is only 3.47 bit/symbol for this case, which shows that allowing IBI has the potential to improve performance compared to block transmission.

5. Conclusions

A method for the joint design of transmitter and receiver for data transmission over dispersive channels has been presented. The proposed method maximizes the information rate and can treat the practically important case where the transmitter is FIR and the channel has arbitrary length. This allows for low latency transmission over dispersive channels. Design examples have confirmed the effectiveness of the design method.

References

[1] J. Salz, "Digital transmission over cross-coupled linear channels," *AT&T Tech. J.*, pp. 1147–1159, July-Aug. 1985.

[2] J. Yang and S. Roy, "On joint transmitter and receiver optimization for multiple-input–multiple-output (MIMO) transmission systems," *IEEE Trans. Signal Proc.*, vol. 42, no. 12, pp. 3221–3231, 1994.

[3] X.-G. Xia, "New precoding for intersymbol interference cancellation using nonmaximally decimated multirate filterbanks with ideal FIR equalizers," *IEEE Trans. Signal Proc.*, vol. 45, no. 10, pp. 2431–2440, 1997.

[4] A. Scaglione, G. B. Giannakis, and S. Barbarossa, "Redundant filterbank precoders and equalizers. Part I: Unification and optimal designs," *IEEE Trans. Signal Proc.*, vol. 47, no. 7, pp. 1988–2006, 1999.

[5] A. Scaglione, S. Barbarossa, and G. B. Giannakis, "Filterbank transceivers optimizing information rate in block transmissions over dispersive channels," *IEEE Trans. Inform. Theory*, vol. 45, no. 3, pp. 1019–1032, 1999.

[6] N. Al-Dhahir, "Transmitter optimization for noisy isi channels in the presence of crosstalk," *IEEE Trans. Signal Proc.*, vol. 48, no. 3, pp. 907–911, 2000.

[7] N. Al-Dhahir and J. M. Cioffi, "Block transmission over dispersive channels: transmit filter optimization and realization, and MMSE-DFE receiver performance," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 137–160, 1996.

[8] I. Kalet, "The multitone channel," *IEEE Trans. Commun.*, vol. 37, no. 2, pp. 119–124, 1989.

[9] T. de Couasnon, R. Monnier, and J. B. Rault, "OFDM for digital TV broadcasting," *Signal Proc.*, vol. 39, no. 1–2, pp. 1–32, Sept. 1994.

[10] T. Li and Z. Ding, "Joint transmitter-receiver optimization for partial response channels based on nonmaximally decimated filterbank precoding technique," *IEEE Trans. Signal Proc.*, vol. 47, no. 9, pp. 2407–2414, 1999.

[11] A. Mertins, "Design of redundant FIR precoders for arbitrary channel lengths using an MMSE criterion," in *Proc. Int. Conf. Commun. (ICC'02)*, New York, USA, April-May 2002, vol. 1, pp. 212–216.

[12] J. Yang and S. Roy, "Joint transmitter-receiver optimization for multi-input multi-output systems with decision feedback," *IEEE Trans. Inform. Theory*, vol. 40, no. 5, pp. 1334–1347, Sept. 1994.

[13] A. Stamoulis, W. Tang, and G. B. Giannakis, "Information rate maximizing FIR transceivers: filterbank precoders and decision-feedback equalizers for block transmissions over dispersive channels," in *Proc. Glob. Telecommun. Conf. (GLOBECOM'99)*, Rio de Janeiro, Brazil, Dec. 1999, vol. 4, pp. 2142–2146.

[14] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[16] J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications, and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

Alfred Mertins received his Diplomingenieur degree from the University of Paderborn, Germany, in 1984, the Dr.-Ing. degree in electrical engineering and the Dr.-Ing. habil. degree in telecommunications from the Hamburg University of Technology, Germany, in 1991 and 1994, respectively. From 1986 to 1991 he was with the Hamburg University of Technology, Germany, from 1991 to 1995 with the Microelectronics Applications Center Hamburg, Germany, from 1996 to 1997 with the University of Kiel, Germany, from 1997 to 1998 with the University of Western Australia, and from 1998 to 2003 with the University of Wollongong, Australia. In April 2003, he joined the University of Oldenburg, Germany, where he is a Professor in the School of Mathematics and Natural Sciences. His research interests include speech, audio, image and video processing, wavelets and filter banks, and digital communications.
e-mail: alfred.mertins@uni-oldenburg.de
School of Mathematics and Natural Sciences
University of Oldenburg
D-26111 Oldenburg, Germany

Paper

Blind frequency offset estimation for overlap PCC-OFDM systems in presence of phase noise

Jinwen Shentu and Jean Armstrong

Abstract — This paper presents a technique for frequency offset estimation for polynomial cancellation coded orthogonal frequency division multiplexing with symbols overlapped in the time domain (overlap PCC-OFDM) in the presence of phase noise. The frequency offset estimator is designed based on the subcarrier pair imbalance (SPI) caused by frequency offset. The estimation is performed in the frequency domain at the output of the receiver discrete Fourier transform (DFT). No training symbols or pilot tones are required. Simulations show that this estimator is an approximately linear function of frequency offset. Phase noise does not significantly affect the variance performance of the estimator.

Keywords — OFDM, frequency offset, estimation, phase noise, overlap PCC-OFDM.

1. Introduction

Orthogonal frequency division multiplexing (OFDM) techniques have been widely used as the modulation method in many high-speed data transmission systems [1, 2]. One of the major disadvantages is its high sensitivity to frequency offset caused by the frequency difference between the oscillators in the transmitter and the receiver. A frequency offset will introduce a common phase error (CPE) in all subcarriers in a symbol, will cause noise-like intercarrier interference (ICI) crossing all subcarriers in a symbol and reduce the amplitude of the wanted signal. Since phase noise will cause CPE and ICI, OFDM is also very sensitive to phase noise [3].

Polynomial cancellation coded OFDM is an ICI cancellation scheme which was originally designed to reduce the ICI caused by frequency offset rather than phase noise [4–6]. A recent study shows that PCC-OFDM can also significantly reduce the ICI caused by phase noise [7]. Furthermore, it is shown in [8] that in PCC-OFDM systems where the symbols are entirely coded in PCC, blind frequency offset estimation can be achieved by exploiting the subcarrier pair imbalance caused by frequency offset.

In PCC-OFDM, each data value to be transmitted is mapped onto a group of subcarriers. The number of subcarriers in a group depends on the requirement for the system performance [6]. In this paper, the case where each data value to be transmitted is mapped onto a pair of subcarriers is considered. PCC-OFDM has many advantages including insensitivity to frequency offset and Doppler spread, much faster power spectrum roll-off and much lower out-of-band power than conventional OFDM. Despite its advantages, PCC-OFDM is not bandwidth efficient in its simplest form. One way to overcome this drawback is to overlap PCC-OFDM symbols in the time domain [9, 10]. Figure 1 shows how PCC-OFDM symbols are overlapped with an overlapping offset of $T/2$, where T is the symbol period. In this case, the first half of the current PCC-OFDM symbol is overlapped with the second half of the previous PCC-OFDM symbol. The second half of the current symbol is overlapped with the first half of the next PCC-OFDM symbol. Using the overlapping technique, the same data rate as conventional OFDM can be achieved while the benefits of PCC-OFDM are retained. After the overlapping, at any time instant an overlap PCC-OFDM symbol contains overlapping components from adjacent PCC-OFDM symbols. In this paper we propose a frequency offset estimator for overlap PCC-OFDM, which has the same form of the estimator as that we proposed for PCC-OFDM [8].

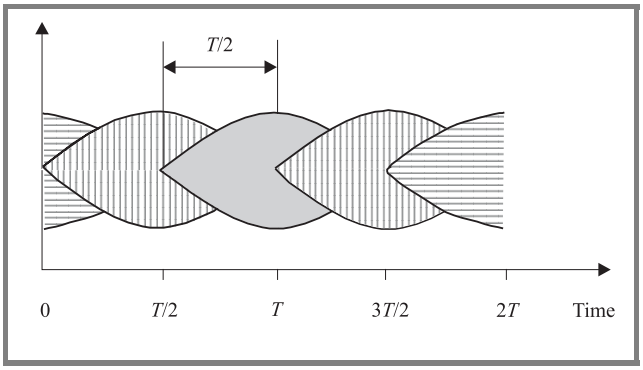


Fig. 1. PCC-OFDM symbols overlapped in the time domain.

This paper is organized as follows. Sections 2 and 3 are the background for overlap PCC-OFDM techniques. They are essential for understanding the frequency offset estimation in overlap PCC-OFDM. In Section 2, the basic concepts for overlap PCC-OFDM are introduced and an overlap OFDM system is described. In Sections 3 and 4, the SPI which is the basis of the technique of frequency offset estimation is investigated, and the expression for a demodulated overlap PCC-OFDM subcarrier is derived. In Section 5, the frequency offset estimator for overlap PCC-OFDM is

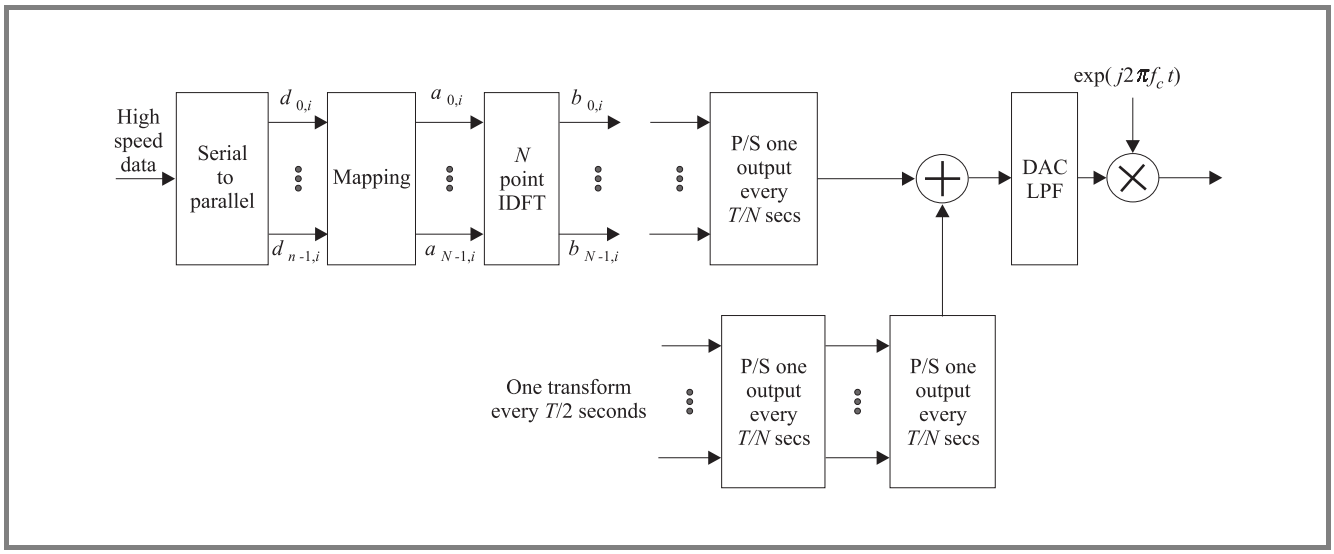


Fig. 2. Transmitter for an overlap PCC-OFDM system.

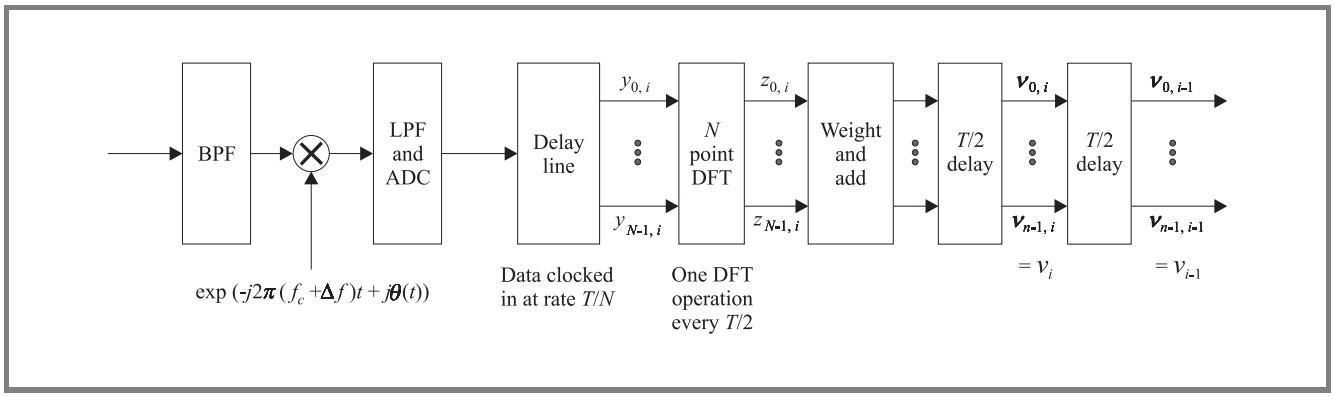


Fig. 3. Receiver for an overlap PCC-OFDM system.

introduced. In Sections 6 and 7, the effects of phase noise on the performance of the PCC-OFDM and the estimator have been evaluated. In Section 8, numerical simulation results are presented and the performance of the estimator is evaluated. Conclusions are drawn in Section 9.

2. An overlap PCC-OFDM system

We will now look into an overlap PCC-OFDM system. For the convenience of description, we define a symbol before overlapping as a PCC-OFDM symbol and define a symbol containing overlapping components as an overlap PCC-OFDM symbol.

Figure 2 shows the block diagram of an overlap PCC-OFDM transmitter. $d_{0,i}, \ldots, d_{n-1,i}$ are n data values in the ith data block to be transmitted. They are mapped onto N subcarriers $a_{0,i}, \ldots, a_{N-1,i}$, where N is a power of 2. In the case of data being mapped onto pairs of subcarriers, we have $n = N/2$ and $a_{2M+1,i} = -a_{2M,i}$. The output vectors of the inverse discrete Fourier transform (IDFT) are then overlapped in the way described in Fig. 1. After the parallel to serial conversion (P/S), the digital to analog conversion (DAC) and low pass filtering (LPF), the overlap PCC-OFDM symbol is up converted into the radio frequency (RF) signal.

Figure 3 shows the block diagram of an overlap PCC-OFDM receiver. The received signal is fed into a band pass filter (BPF) and down converted into a baseband signal. The frequency offset Δf is introduced. For convenience, the normalized frequency offset ΔfT is represented by ε. After the LPF and analog to digital conversion (ADC), a sequence of samples of the baseband signal is obtained. The data samples are clocked into the delay line at rate T/N and converted into parallel data blocks, each block has N data values. One DFT operation is performed on each block for every $T/2$ second.

The output of the DFT is then weighted and added to obtain the estimates of the transmitted data. The technique for

data recovery from a set of overlapped symbols has been presented in the reference [9]. A two-dimensional minimum mean square error (MMSE) equalizer can effectively recover the overlapped symbols. The equalization of the overlap PCC-OFDM systems is beyond the range of this paper, more detailed discussions can be found in the reference [9].

3. Subcarrier pair imbalance in PCC-OFDM in absence of phase noise

In overlap PCC-OFDM, each demodulated subcarrier contains overlapping components from adjacent PCC-OFDM symbols [11]. The SPI is defined as the amplitude or power difference between two subcarriers in a demodulated subcarrier pair. That is

$$F(\Delta fT) = \left|z_{2M+1,i}\right|^2 - \left|z_{2M,i}\right|^2. \tag{1}$$

Each value of imbalance is a combination of the imbalance of the PCC-OFDM subcarrier pair and the imbalance of the overlapping components in the demodulated subcarrier pair.

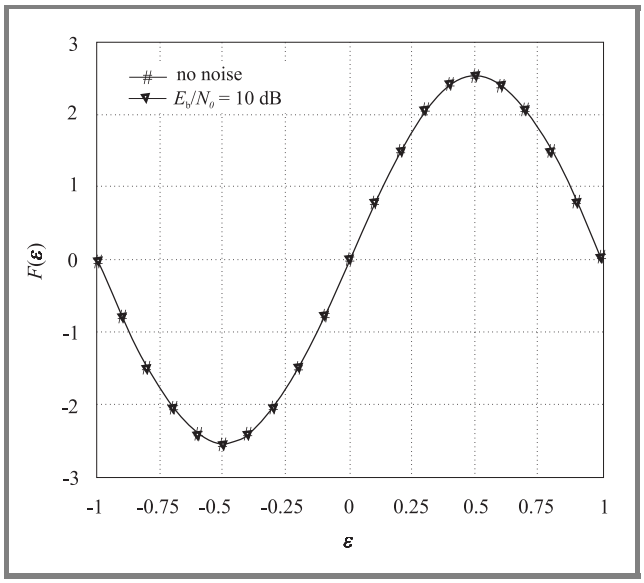


Fig. 4. Average SPI as a function of frequency offset for $N = 128$, 300 symbols.

In the absence of frequency offset, two demodulated subcarriers in an overlap PCC-OFDM subcarrier pair are balanced in a particular way. The SPI depends not only on the frequency offset and transmitted data values in the current PCC-OFDM symbol, but also on the overlapping components from the preceding and following PCC-OFDM symbols. The data dependency in overlap PCC-OFDM can be removed by averaging over a number of subcarrier pairs. The average SPI depends on the frequency offset, therefore, we can use the average imbalance for frequency offset estimation.

Figure 4 shows the average SPI as a function of frequency offset for $N = 128$. The number of simulated symbols was 300 and all subcarrier pairs were used. The figure shows some interesting features:

- Crossing zero at zero frequency offset.
- For $|\varepsilon| < 0.5$, the SPI increases monotonically as the frequency offset increases.
- Additive white Gaussian noise (AWGN) does not affect the zero crossing position.
- The frequency estimation range can be extended to one subcarrier spacing.

4. Expression for an overlap PCC-OFDM subcarrier

To investigate the relationship between the SPI and the frequency offset more clearly, the expression of an OFDM signal must be derived. The *k*th sample in the *i*th time domain symbol in PCC-OFDM is given by [12]

$$b_{k,i} = \frac{1}{N}\sum_{l=0}^{N-1} a_{l,i} \exp\left(\frac{j2\pi kl}{N}\right), \tag{2}$$

where $a_{l,i}$ is the signal in the *l*th subcarrier in the *i*th OFDM symbol. The overlapping components in the current symbol are introduced from the second half of the $(i-1)$th PCC-OFDM symbol and the first half of the $(i+1)$th PCC-OFDM symbol. The *k*th overlapping component is given by [11]

$$b'_{k,i} = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{l=0}^{N-1}(-1)^l a_{l,i-1}\exp\left(\frac{j2\pi kl}{N}\right), & \text{for } 0 \le k \le N/2-1, \\ \dfrac{1}{N}\displaystyle\sum_{l=0}^{N-1}(-1)^l a_{l,i+1}\exp\left(\frac{j2\pi kl}{N}\right), & \text{for } N/2 \le k \le N-1. \end{cases} \tag{3}$$

At the receiver, the *k*th value of the *i*th input data block to the DFT is given by

$$y_{k,i} = \exp\left(\frac{j2\pi k\varepsilon}{N}\right)\left(b'_{k,i} + b_{k,i}\right) + w_{k,i}, \tag{4}$$

where $w_{k,i}$ is the channel AWGN. The mth subcarrier in the ith demodulated symbol is then given by

$$z_{m,i} = \sum_{k=0}^{N-1} y_{k,i} \exp\left(\frac{-j2\pi km}{N}\right) =$$

$$= \sum_{k=0}^{N/2-1} y_{k,i} \exp\left(\frac{-j2\pi km}{N}\right) + \sum_{k=N/2}^{N-1} y_{k,i} \exp\left(\frac{-j2\pi km}{N}\right). \quad (5)$$

The $2M$th demodulated subcarrier is given by

$$z_{2M,i} = \sum_{L=0}^{N/2-1} (CF_{2(L-M)} + CF_{2(L-M)+1}) d_{L,i-1} + \\ + \sum_{L=0}^{N-1} (CS_{2(L-M)} + CS_{2(L-M)+1}) d_{L,i+1} + \\ + \sum_{L=0}^{N-1} (c_{2(L-M)} - c_{2(L-M)+1}) d_{L,i} + W_{2M,i}, \quad (6)$$

where $W_{2M,i}$ is the DFT of $w_{k,i}$, c_{l-m} are complex coefficients given by [4]

$$c_{l-m} = \frac{\sin\big(\pi(l-m+\varepsilon)\big)}{N \sin\big(\pi(l-m+\varepsilon)/N\big)} \times \exp\big(j\pi(N-1)(l-m+\varepsilon)/N\big). \quad (7)$$

It is shown in [13], $W_{2M,i}$ is AWGN. The coefficients CF_{l-m} and CS_{l-m} are given by

$$CF_{l-m} = \frac{1}{N} \sum_{k=0}^{N/2-1} \exp\left(\frac{j2\pi k(l-m+\varepsilon)}{N}\right), \quad (8)$$

$$CS_{l-m} = \frac{1}{N} \sum_{k=N/2}^{N-1} \exp\left(\frac{j2\pi k(l-m+\varepsilon)}{N}\right). \quad (9)$$

Similarly, the $(2M+1)$th demodulated subcarrier can be obtained.

5. Minimum mean square error frequency offset estimator for overlap PCC-OFDM

It is shown in Appendix A that Eq. (1) for overlap PCC-OFDM for 4 quadrature amplitude modulation (4QAM) can be written as

$$F(\varepsilon) = K \sin(\pi\varepsilon) + e_M, \quad (10)$$

where K is a constant given by [8]

$$K = \cos\left(\frac{\pi}{N}\right) \prod_{k=1}^{\log_2(N)-1} \cos\left(\frac{2^{k-1}\pi}{N}\right) \quad (11)$$

e_M is the error term. For small frequency offset, the error term can be approximated as additive noise with zero mean and finite variance. Applying MMSE techniques for Eq. (10) [14], we can obtain the frequency offset estimator

$$\hat{\varepsilon} = \frac{1}{\pi} \sin^{-1}\left(\frac{1}{KM_l} \sum_{t=1}^{M_l} F_t(\varepsilon)\right), \quad (12)$$

where M_l is number of subcarrier pairs used for the frequency offset estimation.

6. Phase noise in PCC-OFDM

The effects of phase noise on the performance of PCC-OFDM have been investigated in [15]. For convenience of discussion, the main results are outlined as follows. Let

$$\psi_p = \frac{1}{N} \sum_{k=0}^{N-1} \exp\left(\frac{-j2\pi kp}{N}\right) \exp\big(j\theta(k)\big), \quad (13)$$

where $p = -(N-1), \ldots, 0, \ldots, (N-1)$, $\theta(k)$ is the kth sample of the phase noise. Thus the quantity ψ_p is the DFT of $\exp\big(j\theta(k)\big)$, evaluated at the frequency p/N. Variable ψ_p depends on the frequency spectrum of the phase noise. If $\theta(k)$ is a constant, then $\psi_p = 0$, for $p \neq 0$, and there is no ICI. Furthermore, for $\theta(k) = 0$, we obtain $\psi_p = 1$, and the demodulated subcarrier is equal to its original data to be transmitted. For a small phase noise $\theta(k)$, using the approximation $\exp\big(j\theta(k)\big) \approx 1 + j\theta(k)$, Eq. (13) can be written as

$$\psi_p \approx \frac{1}{N} \sum_{k=0}^{N-1} \exp\left(\frac{-j2\pi kp}{N}\right) \big(1 + j\theta(k)\big). \quad (14)$$

For $p \neq 0$, Eq. (14) is given by

$$\psi_p \approx \frac{j}{N} \sum_{k=0}^{N-1} \exp\left(\frac{-j2\pi kp}{N}\right) \theta(k). \quad (15)$$

Quantity ψ_p is presented in terms of the DFT of the phase noise, which is the spectrum of the phase noise $\theta(k)$. For $p = 0$, we get

$$\psi_p \approx 1 + \frac{j}{N} \sum_{k=0}^{N-1} \theta(k). \quad (16)$$

The estimate of the Mth data value in the ith data block to be transmitted is given by

$$v_{M,i} = d_{M,i} + \frac{1}{2} d_{M,i} \big[-\psi_{-1} + 2(\psi_0 - 1) - \psi_1 \big] + \\ + \frac{1}{2} \sum_{\substack{L=0 \\ L \neq M}}^{N/2-1} d_{L,i} \big(-\psi_{2(M-L)+1} + 2\psi_{2(M-L)} - \psi_{2(M-L)-1} \big) \\ + \big(W_{2M,i} - W_{2M+1,i}\big)/2, \quad (17)$$

where the first term at the right hand is the wanted signal, the second term is the CPE, the third term is the ICI and the last one is the weighted AWGN. In OFDM, the CPE and ICI depend on the individual frequency spectrum of the phase noise [3, 16]. In PCC-OFDM, the CPE and ICI depend on the combinations of phase noise spectra rather than individual spectra. This makes it possible to reduce the effects of phase noise by using a phase lock loop (PLL) that can fit the overall phase noise spectrum to a particular pattern. For the special case where the individual spectrum ψ_p has a linear relationship with frequency, then the ICI caused by the phase noise can be completely cancelled.

The CPE term is given by

$$Y_{CM,i} = \frac{1}{2} d_{M,i}\left[-\psi_{-1} + 2(\psi_0 - 1) - \psi_1\right] = j d_{M,i}\Theta_0 \quad (18)$$

where Θ_0 is a constant,

$$\Theta_0 = \frac{2}{N}\sum_{k=0}^{N-1} \sin^2\left(\frac{\pi k}{N}\right)\theta(k). \quad (19)$$

This result indicates that all PCC-OFDM subcarriers experience a CPE. This rotation can be detected and therefore compensated using techniques provided in the literature [17]. One simple way to do this is to insert pilot tones in a symbol and estimate the rotation angle. Once the CPE is corrected in the pilot tones, the remaining phase noise in other subcarriers can be corrected.

Similarly, the ICI term in Eq. (17) can be presented by

$$Y_{IM,i} = \frac{1}{2}\sum_{\substack{L=0 \\ L\neq M}}^{N/2-1} d_{L,i}\left(-\psi_{2(M-L)+1} + \right.$$

$$\left. + 2\psi_{2(M-L)} - \psi_{2(M-L)-1}\right) = j\sum_{\substack{L=0 \\ L\neq M}}^{N/2-1} d_{L,i}\Theta_{L-M}, \quad (20)$$

where

$$\Theta_{L-M} = \frac{2}{N}\sum_{k=0}^{N-1} \sin^2\left(\frac{\pi k}{N}\right)\exp\left(\frac{j4\pi k(L-M)}{N}\right)\theta(k). \quad (21)$$

We will now investigate the case where phase noise is white Gaussian. In this case, no correlation between the samples of phase noise is assumed, and the possible difference between two consecutive samples is highest. The variance of the CPE can be calculated by

$$\mathrm{var}\left[\Theta_0\right] = \frac{4}{N^2}\sum_{k=0}^{N-1} \sin^4\left(\frac{\pi k}{N}\right) E\left[\left|\theta(k)\right|^2\right] = \frac{3\sigma_\theta^2}{2N}, \quad (22)$$

where σ_θ^2 is the variance of the phase noise, $\sigma_\theta^2 = E\left[\left|\theta(k)\right|^2\right]$. Note the summation in Eq. (22) is a constant, $\sum_{k=0}^{N-1} \sin^4(\pi k/N) = 3N/8$. The variance of the CPE is proportional to the variance of the phase noise. Thus the variance of the CPE is then given by

$$\mathrm{var}\left[Y_{CM,i}\right] = \frac{3\sigma_\theta^2\sigma_s^2}{2N} = \frac{3\sigma_\theta^2}{2N}. \quad (23)$$

Note we have used $\sigma_s^2 = 1$ for 4QAM. The variance of the angle in ICI is given by

$$\mathrm{var}\left[\Theta_{L-M}\right] = E\left[\Theta_{L-M}\Theta_{L-M}^*\right] =$$

$$= \frac{4}{N^2}\sum_{k=0}^{N-1} \sin^4\left(\frac{\pi k}{N}\right) E\left[\left|\theta(k)\right|^2\right] = \frac{3\sigma_\theta^2}{2N}. \quad (24)$$

Similarly, we can calculate the variance of the ICI from Eq. (17)

$$\mathrm{var}\left[Y_{IM,i}\right] = \left(\frac{N}{2}+1\right)\frac{3}{2N}\sigma_s^2\sigma_\theta^2 = \left(\frac{3}{4}+\frac{3}{2N}\right)\sigma_\theta^2. \quad (25)$$

The result indicates that the variance of ICI caused by phase noise has been significantly reduced in PCC-OFDM [3]. For a sufficiently large N, the value of ICI in PCC-OFDM is almost one quarter less than that in OFDM.

7. Effects of phase noise on SPI

The frequency offset estimator is derived for overlap PCC-OFDM systems without phase noise. When the phase noise is taken into account, the performance of the estimator is affected by the phase noise since the phase noise also contributes to the SPI. Figure 5 shows the SPI in the pres-

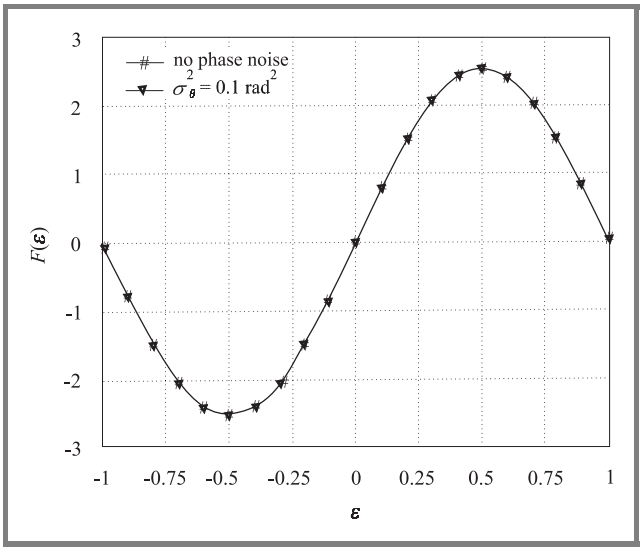


Fig. 5. Average SPI as a function of frequency offset in presence of phase noise for $N = 128$, 300 symbols.

ence of phase noise for $N = 128$ averaging over 300 symbols. It is shown that the phase noise does not significantly contribute to the SPI. In other words, the frequency offset estimator can be used for frequency offset estimation even in the presence of phase noise. Phase noise will increase the variance of estimate of frequency offset. This issue will be discussed in the next section.

8. Simulation results

To evaluate the performance of the MMSE estimator, simulations were performed. In the following simulations, 4QAM is used for the modulation scheme with $M_l = 512$ and $N = 128$. Figure 6 shows the estimator as a function of the frequency offset for $E_b/N_0 = 10$ dB. It is evident that the estimator has an approximately linear relationship with frequency offset.

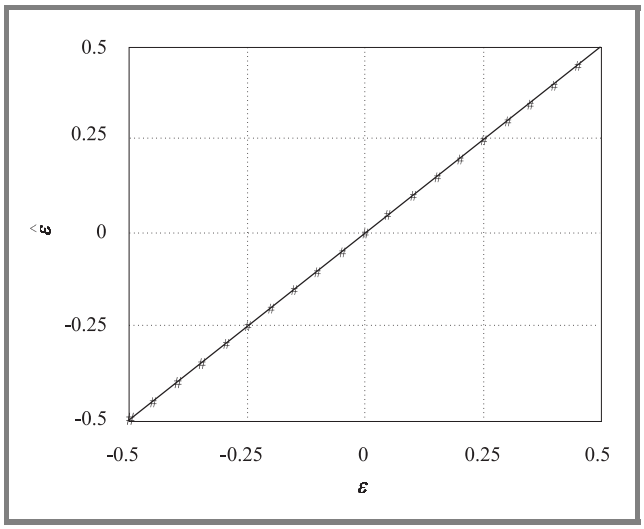


Fig. 6. $F(\varepsilon)$ as a function of frequency offset for $M_l = 512$ and $N = 128$.

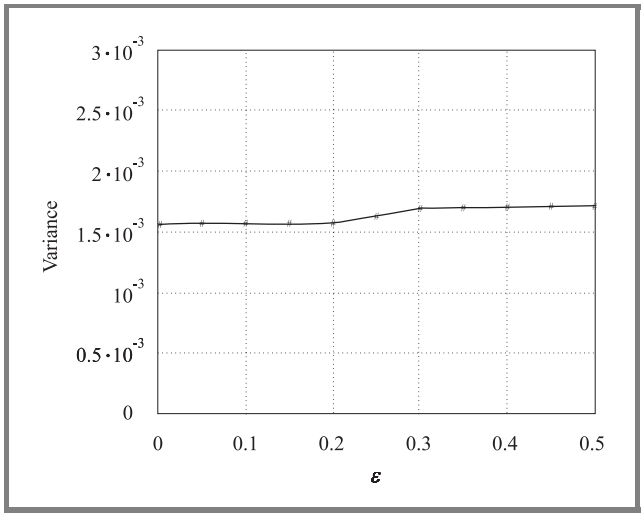


Fig. 7. Variance as a function of frequency offset for $M_l = 512$ and $N = 128$.

Figure 7 shows the variance of the frequency offset estimator as a function of the frequency offset for an ideal channel. The variance does not change significantly as frequency offset increases. This means that the overlapping components in overlap PCC-OFDM are the dominant factor for the variance. Lower variance can be obtained by increasing M_l or using a two-dimensional MMSE equalizer before the frequency offset estimation.

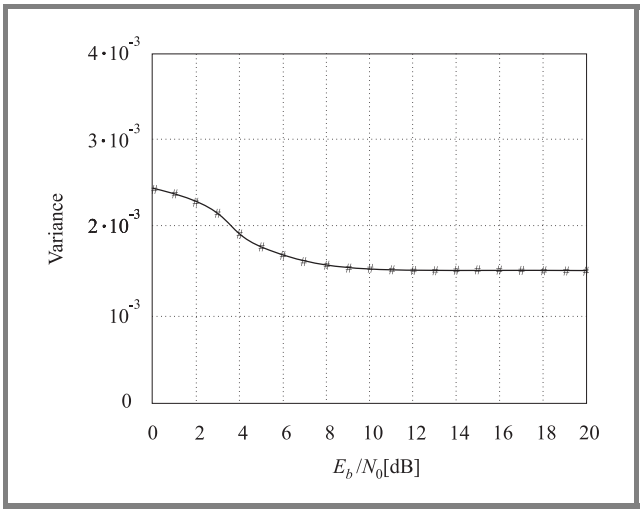


Fig. 8. Variance as a function of E_b/N_0 for $M_l = 512$ and $N = 128$.

Figure 8 shows the variance of frequency offset estimator as a function of E_b/N_0. The frequency offset simulated was zero. The variance does not significantly change as the channel noise increases. This is because the variance is dominated by the overlapping components rather than channel noise.

9. Conclusions

An MMSE frequency offset estimator for overlap PCC-OFDM systems has been presented. A frequency offset estimate is obtained by using the average SPI at the output of the receiver DFT. No training symbols or pilot tones are required. The estimator has an approximately linear relationship with frequency offset. The effects of phase noise on the performance of PCC-OFDM have been theoretically discussed. Particularly, the effect of phase noise on the SPI has been analyzed and simulated. Simulations show that the phase noise does not affect variance performance of the MMSE frequency offset estimator.

Appendix A

From Eq. (6), the power of the $2M$th subcarrier is given by

$$\begin{aligned}\left|z_{2M,i}\right|^2 = &\sum_{L=0}^{N/2-1}\left|\left(CF_{2(L-M)}+CF_{2(L-M)+1}\right)\right|^2\left|d_{L,i-1}\right|^2+\\ &+\sum_{L=0}^{N/2-1}\left|\left(CS_{2(L-M)}+CS_{2(L-M)+1}\right)\right|^2\left|d_{L,i+1}\right|^2+\\ &+\sum_{L=0}^{N/2-1}\left|\left(c_{2(L-M)}-c_{2(L-M)+1}\right)\right|^2\left|d_{L,i}\right|^2+e_{2M,i}.\end{aligned} \tag{A.1}$$

Equation (A.1) indicates that the power of each demodulated subcarrier can be represented in terms of the individual power components from the preceding, current and following PCC-OFDM symbols. Note for 4QAM with each subcarrier normalized to unity, $\left|d_{L,i}\right|^2=1$. The error term of Eq. (A.1) is then given by

$$\begin{aligned}e_{2M,i} = &\sum_{\substack{L=0\\L\neq K}}^{N/2-1}\sum_{K=0}^{N/2-1}\left(c_{2(L-M)}-c_{2(L-M)+1}\right)\left(c_{2(K-M)}-c_{2(K-M)+1}\right)^* d_{L,i}d^*_{K,i}+\\ &+\sum_{\substack{L=0\\L\neq K}}^{N/2-1}\sum_{K=0}^{N/2-1}\left(CF_{2(L-M)}+CF_{2(L-M)+1}\right)\left(CF_{2(K-M)}+CF_{2(K-M)+1}\right)^* d_{L,i-1}d^*_{K,i-1}+\\ &+\sum_{\substack{L=0\\L\neq K}}^{N/2-1}\sum_{K=0}^{N/2-1}\left(CS_{2(L-M)}+CS_{2(L-M)+1}\right)\left(CS_{2(K-M)}+CS_{2(K-M)+1}\right)^* d_{L,i+1}d^*_{K,i+1}+\\ &+\sum_{L=0}^{N/2-1}\sum_{K=0}^{N/2-1}\left(c_{2(L-M)}-c_{2(L-M)+1}\right)\left(CS_{2(K-M)}+CS_{2(K-M)+1}\right)^* d_{L,i}d^*_{K,i+1}+\\ &+\sum_{L=0}^{N/2-1}\sum_{K=0}^{N/2-1}\left(CF_{2(L-M)}+CF_{2(L-M)+1}\right)\left(c_{2(K-M)}+c_{2(K-M)+1}\right)^* d_{L,i-1}d^*_{K,i}+\\ &+\sum_{L=0}^{N/2-1}\sum_{K=0}^{N/2-1}\left(c_{2(L-M)}-c_{2(L-M)+1}\right)^*\left(CS_{2(K-M)}+CS_{2(K-M)+1}\right)d^*_{L,i}d_{K,i+1}+\\ &+\sum_{L=0}^{N/2-1}\sum_{K=0}^{N/2-1}\left(CF_{2(L-M)}+CF_{2(L-M)+1}\right)^*\left(c_{2(K-M)}+c_{2(K-M)+1}\right)d^*_{L,i-1}d_{K,i}+\\ &+2\mathrm{Re}\left\{W_{2M,i}\sum_{K=0}^{N/2-1}\left(CF_{2(K-M)}+CF_{2(K-M)+1}\right)^* d^*_{K,i-1}\right\}+\\ &+2\mathrm{Re}\left\{W_{2M,i}\sum_{K=0}^{N/2-1}\left(CS_{2(K-M)-1}+CS_{2(K-M)}\right)^* d^*_{K,i+1}\right\}+\\ &+2\mathrm{Re}\left\{W_{2M,i}\sum_{K=0}^{N/2-1}\left(c_{2(K-M)}-c_{2(K-M)+1}\right)^* d^*_{K,i}\right\}+\left|W_{2M,i}\right|^2.\end{aligned} \tag{A.2}$$

Because the expected value of all cross terms is zero, the expected value of the error term is equal to the variance of the noise. Similarly, we can obtain the power for the $(2M+1)$th subcarrier. Using $\left|d_{L,i}\right|^2 = 1$ for 4QAM, the subcarrier pair imbalance is then given by

$$\begin{aligned}
&\left|z_{2M+1,i}\right|^2 - \left|z_{2M,i}\right|^2 = \sum_{L=0}^{N/2-1} \left|\left(CF_{2(L-M)-1} + CF_{2(L-M)}\right)\right|^2 + \\
&+ \sum_{L=0}^{N/2-1} \left|\left(CS_{2(L-M)-1} + CS_{2(L-M)}\right)\right|^2 - \sum_{L=0}^{N/2-1} \left|\left(CF_{2(L-M)} + CF_{2(L-M)+1}\right)\right|^2 + \\
&- \sum_{L=0}^{N/2-1} \left|\left(CS_{2(L-M)} + CS_{2(L-M)+1'}\right)\right|^2 + \sum_{L=0}^{N/2-1} \left|\left(c_{2(L-M)-1} - c_{2(L-M)}\right)\right|^2 + \\
&- \sum_{L=0}^{N/2-1} \left|\left(c_{2(L-M)} - c_{2(L-M)+1}\right)\right|^2 + e_{2M+1,i} - e_{2M,i}\,.
\end{aligned} \tag{A.3}$$

It is shown in Appendix B that the combination of the first four summations in Eq. (A.3) equals zero. Thus we can obtain the subcarrier pair imbalance

$$F(\varepsilon) = \left|z_{2M+1,i}\right|^2 - \left|z_{2M,i}\right|^2 = S(\varepsilon) + e'_{M,i} \tag{A.4}$$

where $S(\varepsilon)$ is given by [6]

$$S(\varepsilon) = K \sin(\pi\varepsilon) \tag{A.5}$$

K is a factor defined in Eq. (11). $e'_{M,i}$ is the Mth total error, $e'_{m,i} = e_{2M+1,i} - e_{2M,i}$. In overlap PCC-OFDM, the error term is more complicated than in PCC-OFDM [8]. However, the dominant distribution in the error term is still Gaussian. As for PCC-OFDM, $e'_{M,i}$ can be approximated as AWGN with zero mean. The variance is larger than that of PCC-OFDM because of the overlapping components. Appendix C shows the variance of the coupling-crossing terms for $e'_{M,i}$. Thus, for a small frequency offset, the overall approximate variance is given by

$$e'_{M,i} \sim N(0, 8 + 8\sigma_n^2 + 2\sigma_n^4) \tag{A.6}$$

where "$\sim$" means distributed as. Note that we have used $\sigma_s^2 = 1$ for 4QAM in Eq. (A.6). The SPI estimator is therefore given by

$$\hat{\varepsilon} = \frac{1}{\pi} \sin^{-1}\left(\frac{1}{KM_l} \sum_{i=1}^{M_l} F_i(\varepsilon)\right). \tag{A.7}$$

Appendix B

The element of the first summation of Eq. (A.3) is given by

$$\begin{aligned}
&\left|\left(CF_{2(L-M)-1} + CF_{2(L-M)}\right)\right|^2 = \\
&= \frac{\sin^2\left(\frac{\pi\varepsilon}{2}\right)}{N^2 \sin^2\left(\frac{\pi(2L+\varepsilon)}{N}\right)} + \frac{\cos^2\left(\frac{\pi\varepsilon}{2}\right)}{N^2 \sin^2\left(\frac{\pi(2L-1+\varepsilon)}{N}\right)} + \\
&- \frac{\sin(\pi\varepsilon)}{N^2 \sin\left(\frac{\pi(2L+\varepsilon)}{N}\right) \sin\left(\frac{\pi(2L-1+\varepsilon)}{N}\right)} \sin\left(\frac{\pi}{N}\right).
\end{aligned} \tag{B.1}$$

The element of the second summation of Eq. (A.3) is given by

$$\left|\left(CS_{2(L-M)-1}+CS_{2(L-M)}\right)\right|^2 =$$

$$= \frac{\sin^2\left(\frac{\pi\varepsilon}{2}\right)}{N^2\sin^2\left(\frac{\pi(2L+\varepsilon)}{N}\right)} + \frac{\cos^2\left(\frac{\pi\varepsilon}{2}\right)}{N^2\sin^2\left(\frac{\pi(2L-1+\varepsilon)}{N}\right)} +$$

$$+ \frac{\sin(\pi\varepsilon)}{N^2\sin\left(\frac{\pi(2L+\varepsilon)}{N}\right)\sin\left(\frac{\pi(2L-1+\varepsilon)}{N}\right)}\sin\left(\frac{\pi}{N}\right). \tag{B.2}$$

Similarly the third and fourth terms in Eq. (A.3) can be derived.

Substituting the elements derived to the combination of the first four summations of Eq. (A.3) gives

$$\sum_{L=0}^{N/2-1}\left|\left(CF_{2(L-M)-1}+CF_{2(L-M)}\right)\right|^2+\sum_{L=0}^{N/2-1}\left|\left(CS_{2(L-M)-1}+CS_{2(L-M)}\right)\right|^2+$$

$$-\sum_{L=0}^{N/2-1}\left|\left(CF_{2(L-M)}+CF_{2(L-M)+1}\right)\right|^2-\sum_{L=0}^{N/2-1}\left|\left(CS_{2(L-M)}+CS_{2(L-M)+1}\right)\right|^2 =$$

$$= 2\sum_{L=0}^{N/2-1}\left\{\frac{\cos^2\left(\frac{\pi\varepsilon}{2}\right)}{N^2\sin^2\left(\frac{\pi(2L+\varepsilon-1)}{N}\right)}-\frac{\cos^2\left(\frac{\pi\varepsilon}{2}\right)}{N^2\sin^2\left(\frac{\pi(2L+1+\varepsilon)}{N}\right)}\right\}=0. \tag{B.3}$$

Appendix C

The coupling-cross terms $CC_{2M,i}$ in Eq. (A.2) is given by

$$CC_{2M,i} =$$

$$= \sum_{L=0}^{N/2-1}\sum_{K=0}^{N/2-1}\left(c_{2(L-M)}-c_{2(L-M)+1}\right)\left(CS_{2(K-M)}+CS_{2(K-M)+1}\right)^* d_{L,i}d^*_{K,i+1}+$$

$$+ \sum_{L=0}^{N/2-1}\sum_{K=0}^{N/2-1}\left(CF_{2(L-M)}+CF_{2(L-M)+1}\right)\left(c_{2(K-M)}-c_{2(K-M)+1}\right)^* d_{L,i-1}d^*_{K,i}+$$

$$+ \sum_{L=0}^{N/2-1}\sum_{K=0}^{N/2-1}\left(c_{2(L-M)}-c_{2(L-M)+1}\right)^*\left(CS_{2(K-M)}+CS_{2(K-M)+1}\right)d^*_{L,i}d_{K,i+1}+$$

$$+ \sum_{L=0}^{N/2-1}\sum_{K=0}^{N/2-1}\left(CF_{2(L-M)}+CF_{2(L-M)+1}\right)^*\left(c_{2(K-M)}-c_{2(K-M)+1}\right)d^*_{L,i-1}d_{K,i}. \tag{C.1}$$

Similarly, the coupling-cross terms $CC_{2M+1,i}$ for the $(2M+1)$th error term can also be obtained. Note the first term and the third term are mutually conjugated, the second term and the fourth term are mutually conjugated in Eq. (C.1). Thus, $CC_{2M+1,i} - CC_{2M,i}$ can be written as

$$CC_{2M+1,i} - CC_{2M,i} =$$

$$= -2\mathrm{Re}\left\{ d_{M,i} \sum_{K=0}^{N/2-1} \left(CS_{2(K-M)-1} + 2CS_{2(K-M)} + CS_{2(K-M)+1} \right)^* d^*_{K,i+1} \right\} +$$

$$- 2\mathrm{Re}\left\{ d^*_{M,i} \sum_{L=0}^{N/2-1} \left(CF_{2(L-M)-1} + 2CF_{2(L-M)} + CF_{2(L-M)+1} \right) d_{L,i-1} \right\}, \tag{C.2}$$

where $\mathrm{Re}(\bullet)$ represents the real part of a complex number. For the first summation in Eq. (C.2), considering the most significant complex coefficients CS_0, we get

$$d_{M,i} \sum_{K=0}^{N/2-1} \left(CS_{2(K-M)-1} + 2CS_{2(K-M)} + CS_{2(K-M)+1} \right)^* d^*_{K,i+1} \approx$$

$$\approx 2 d_{M,i} CS_0^* d^*_{M,i+1} \,. \tag{C.3}$$

Similarly, we can also simplify the second summation in Eq. (C.2). In addition, in the absence of frequency offset, the value of the most significant complex coefficient can be obtained from Eqs. (7) and (8), that is $CS_0 = CF_0 = 0.5$. Thus Eq. (C.2) can be written as

$$CC_{2M+1,i} - CC_{2M,i} \approx$$

$$\approx -2\mathrm{Re}\left\{ d_{M,i} d^*_{M,i+1} \right\} - 2\mathrm{Re}\left\{ d^*_{M,i} d_{M,i-1} \right\}. \tag{C.4}$$

The variance of the left side of Eq. (C.4) is given by

$$E\left[\left| CC_{2M+1,i} - CC_{2M,i} \right| \right]^2 \approx$$

$$\approx E\left[4\left(\mathrm{Re}\left(d_{M,i} d^*_{M,i+1} \right) \right)^2 + 4\left(\mathrm{Re}\left(d^*_{M,i} d_{M,i-1} \right) \right)^2 + \right.$$

$$\left. + 4\left(\mathrm{Re}\left(d_{M,i} d^*_{M,i+1} \right) \right) \left(\mathrm{Re}\left(d^*_{M,i} d_{M,i-1} \right) \right) \right] = 8\sigma_s^4 \,. \tag{C.5}$$

When all complex coefficients are considered, the variance will be slightly larger than above result. Please note that this result is obtained under the assumption of no frequency offset. In the presence of a small frequency offset, the variance could be slightly larger.

References

[1] ETSI, ETS 300 744, "Digital video broadcasting: frame structure, channel coding and modulation for digital terrestrial television", Aug. 1997.

[2] ETSI, DTS/BRAN 0023003, V0.k., "Broadband radio access networks; HiperLAN type 2 technical specification; Physical layer", Aug. 1999.

[3] J. Stott, "The effects of phase noise in COFDM", BBC Research and Development EBU Technical Review, 1998.

[4] J. Armstrong, "Analysis of new and existing methods of reducing intercarrier interference due to carrier frequency offset in OFDM", *IEEE Trans. Commun.*, vol. 47, no. 3, pp. 365–369, 1999.

[5] Y. Zhao and S. G. Haggman, "Sensivity to Doppler shift and carrier frequency errors in OFDM systems – the consequences and solutions", in *IEEE 46th Veh. Technol. Conf.*, Atlanta, USA, Apr. 1996, vol. 3, pp. 1564–1568.

[6] J. Armstrong, P. M. Grant, and G. Povey, "Polynomial cancellation coding of OFDM to reduce intercarrier interference due to Doppler spread", in *IEEE Globecom*, Nov. 1998, vol. 5, pp. 2771–2776.

[7] J. Shentu and J. Armstrong, "Effects of phase noise on the performance of PCC-OFDM", in *Proc. WITSP'2002*, Australia, Dec. 2002, pp. 50–54.

[8] J. Shentu, J. Armstrong, and G. Tobin, "MMSE frequency offset estimator for PCC-OFDM", in *IEEE MICC'2001*, Kuala Lumpur, Oct. 2001, pp. 304–309.

[9] J. Armstrong, J. Shentu, and C. Tellambura, "Frequency domain equalization for OFDM systems with mapping data onto subcarrier pairs and overlapping symbol periods", in *Proc. 5th Int. Symp. Commun. Theory & Appl.*, UK, July 1999, pp. 102–104.

[10] J. Shentu and J. Armstrong, "Frequency offset estimation for PCC-OFDM with symbols overlapped in the time domain", in *Proc. WITSP'2002*, Australia, Dec. 2002, pp. 72–79.

[11] J. Shentu and J. Armstrong, "Blind frequency offset estimation for PCC-OFDM with symbols overlapped in the time domain", in *Proc. IEEE ISCAS'2001*, Sydney, May 2001, vol. 4, pp. 570–573.

[12] J. Shentu and J. Armstrong, "A new frequency offset estimator for OFDM", in *Commun. Syst. Netw. Digit. Signal Proc.*, A. C. Boucouvalas, Ed., UK, July 2000, pp. 13–16.

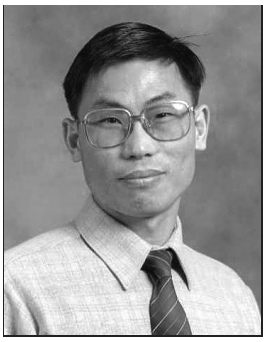
[13] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 1981.

[14] A. Ronald Gallant, *Nonlinear Statistical Models*. Wiley, 1987.

[15] J. Shentu, K. Panta, and J. Armstrong, "Effects of phase noise on the performance of OFDM systems using an ICI cancellations scheme" (accepted for publication by *IEEE Trans. Broadcast.*, 2003).

[16] A. G. Armada, "Phase noise and sub-carrier spacing effects on the performance of an OFDM communication systems", *IEEE Commun. Lett.*, vol. 2, no. 1, 1998.

[17] A. R. S. Bahai and B. R. Saltzberg, *Multicarrier Digital Communications: Theory and Applications of OFDM*. New York: Kluwer, 1999.



Jinwen Shentu received a B.Eng. and an M.Eng. from Lanzhou Jiaotong University (former Lanzhou Railway University), China in 1984 and 1991 respectively, and a Ph.D. from La Trobe University, Melbourne, Australia in 2001. From 1984 to 1988, he was a teacher at Neijiang Railway Mechanical School, Sichuan, China. From 1991 to 1997 he worked as an electronic engineer/senior engineer in Shanghai Railway Communication Equipment Factory, Shanghai, China. He is currently working as a research staff at the Department of Electronic Engineering, La Trobe University. His research interests include OFDM, synchronization, channel estimation, broadband wireless networking and signal processing.
e-mail: j.shentu@ee.latrobe.edu.au
Department of Electronic Engineering
La Trobe University
Victoria 3086, Australia

Jean Armstrong received a B.Sc. in electrical engineering from the University of Edinburgh, Scotland in 1974, an M.Sc. in digital techniques from Heriot-Watt University, Edinburgh, Scotland in 1980 and a Ph.D. in digital communications from Monash University, Melbourne, Australia in 1993. From 1974–1977 she worked as a design engineer at Hewlett-Packard Ltd., Scotland. In 1977 she was appointed as a Lecturer in electrical engineering at the University of Melbourne, Australia. Since 1977 she has held a variety of academic positions at the University of Melbourne, Monash University and La Trobe University. Her research interests include digital communications, engineering education and women in engineering and she has published over 50 research papers. She is currently an Associate Professor at La Trobe University.
e-mail: j.armstrong@latrobe.edu.au
Department of Electronic Engineering
La Trobe University
Victoria 3086, Australia

Paper

LPAR: an adaptive routing strategy for MANETs

Mehran Abolhasan, Tadeusz A. Wysocki, and Eryk Dutkiewicz

Abstract — This paper presents a new global positioning system (GPS)-based routing protocol, called location-based point-to-point adaptive routing (LPAR) for mobile ad hoc networks. This protocol utilises a 3-state route discovery strategy in a point-to-point manner to reduce routing overhead while maximising throughput in medium to large mobile ad hoc networks. In LPAR, data transmission is adaptable to changing network conditions. This is achieved by using a primary and a secondary data forwarding strategy to transfer data from the source to the destination when the condition of the route is changed during data transmission. A simulation study is performed to compare the performance of LPAR with a number of different exisiting routing algorithms. Our results indicate that LPAR produces less overhead than other simulated routing strategies, while maintains high levels of throughput.

Keywords — LPAR, GPS-based routing protocol, mobile ad hoc networks.

1. Introduction

Over the past decade the growing interest in mobile communication and the Internet has opened many new avenues of research in telecommunications. One research area is to provide Internet type applications over mobile ad hoc networks (MANETs). Unlike cellular networks, MANETs are made up of a number of end-user nodes, which are capable of determining routes to different parts of their networks, without using a base-station or a centralised administrator. This desirable feature of such networks makes them useful in many different applications, particularly in areas where an infrastructure is not available or cannot not be easily implemented. These areas include, the highly dynamic battlefields environment where rapid exchange of crucial information can give significant advantage to one side or in search and rescue operations where the rescue team can use these networks to coordinate their efforts to search more effectively.

Other areas where MANETs are useful are in exhibitions, conferences or concerts where a temporary network structure provided by MANETs can reduce implementation cost and time when compared to wired networks. However, MANETs have a number of limitations when compared to wired networks. These include, limitations in bandwidth, battery power and storage space. Other constraints include achieving different levels of QoS under a dynamic network topology and maintaining an acceptable level of data throughput as the number of users and traffic in the network increase.

A number of different routing protocols have been proposed for MANETs. They can be classified into three groups. These are proactive, reactive and hybrid routing protocols. The evolution in design of mobile ad hoc network routing protocols began from the traditional link state and distance vector algorithms, which are commonly used in wired networks. Routing protocols such as DSDV [1] and WRP [2] are among some of the early proactive protocols designed for MANETs [3]. However, due to the periodic updating strategy used in these protocols, they are not scalable in large networks, as the cost of maintaining full network topology will consume a significant part of the available bandwidth, power and storage space available at each node.

Reactive protocols were designed to reduce the cost of maintaining up-to-date routes in proactive protocols at a cost of introducing extra delays during route discovery. This is done by determining routes when they are required via a route discovery strategy, rather than periodically exchanging topology information. The route discovery for most on-demand protocols proposed to date, such as DSR [4], AODV [5] (recently expanding ring search was introduced to limit the scope of the search area) are based on pure-flooding. This means that every time a source requires a route to a particular destination, it will broadcast a route request (RREQ) packet throughout the network. This strategy lacks scalability, as the size of the network and the number of source/destination pairs increase under a dynamic network topology.

As a result, a number of hybrid protocols such as ZRP [6], ZHLS [7] and SLURP [8] were introduced to reduce the effect of flooding in the network. In ZRP, each node defines a zone radius in which the network topology is maintained proactively and routes to destination nodes outside of the zone radius are determined reactively by bordercasting [6]. In ZHLS and SLURP, the network is divided into a number of non-overlapping zones. The topology within each zone is maintained proactively and the routes to the nodes in the interzone are determined reactively. The main disadvantage of ZHLS and SLURP is that they rely on a static zone map, which must be defined for each node at the design stage.

In this study, we propose a number of different strategies to reduce the overheads during route discovery under a dynamically changing network topology, and minimise the power consumed at each node. The rest of this paper is organised as follows. In Section 2, we describe our routing strategy. Section 3, the simulation environment and parameters used are described. In Section 4, we present

a discussion on the results we obtained for our simulations and Section 5 presents the concluding remarks.

2. Location-based point-to-point adaptive routing protocol

Fundamentally, mobile ad hoc networks are dynamic in nature. These network may consists of a number of nodes with different levels of mobility, which may constantly create different node configurations and topologies. This means that during data transfer, source nodes may require a number of route recalculations to successfully transmit the data. As discussed earlier, determining routes proactively over the entire network may use significant amount of the networks available bandwidth. Furthermore, reactive route discovery strategies based on flooding lack scalability as the size of the network increases [9]. Previous work has been done in [5, 10, 11] to reduce the effects of flooding in source routing protocols. In this study, we propose different strategies to reduce overheads under point-to-point routing. In point-to-point routing, each node along the path to destination can make routing decision, which means that they are more adaptable to changing topology and reduce route recalculations at the source. In the following sections, we describe previous strategies proposed in the literature to reduce routing overheads in reactive routing and propose a number of new strategies to increase the performance of point-to-point routing.

2.1. Reactive route discovering strategies

The most common routing strategy used in on-demand routing protocol is pure flooding-based route discovery. In pure flooding the source node generates a route request packet which is broadcasted and propagated globally through the network. When a RREQ packet reaches the required destination or an intermediate node with knowledge about the destination, a route reply (RREP) is generated and sent back to the source. If the RREQ has travelled through bi-directional links, then link reversal can be used to send the reply back to the source, otherwise, the destination may piggy back the route (if source routing) in a route reply packet, which is also flooded to reach the source. Protocols such as DSR and AODV are based on the flooding algorithm. The main difference between the two is in the way routes are created and used. DSR is based on source routing, which means, each data packet carries the complete source to destination address. AODV is a point-to-point routing protocol, which means that the data packets only carry the next hop address and the destination address. A number of different strategies have been proposed to reduce the routing overheads of pure flooding. Two such strategies are, expanding ring search (ERS) and restricted search zones (RSZ). In ERS, the source node incrementally increases the search area until the entire network is searched or the destination has been found. For

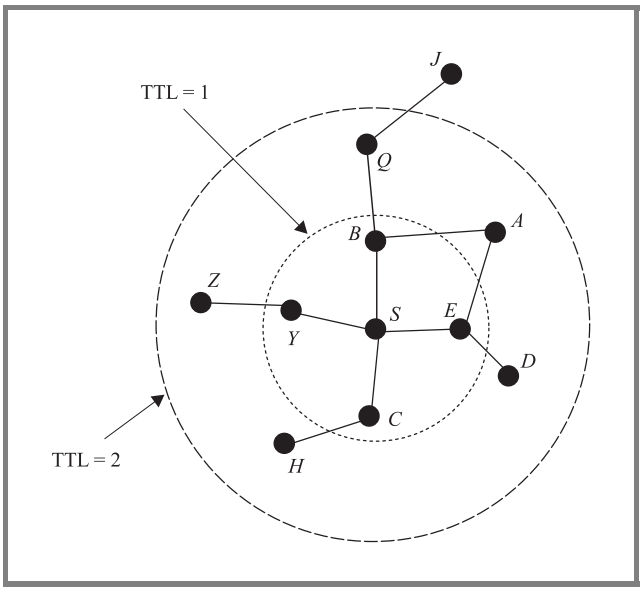


Fig. 1. Controlled flooding using expanding ring search.

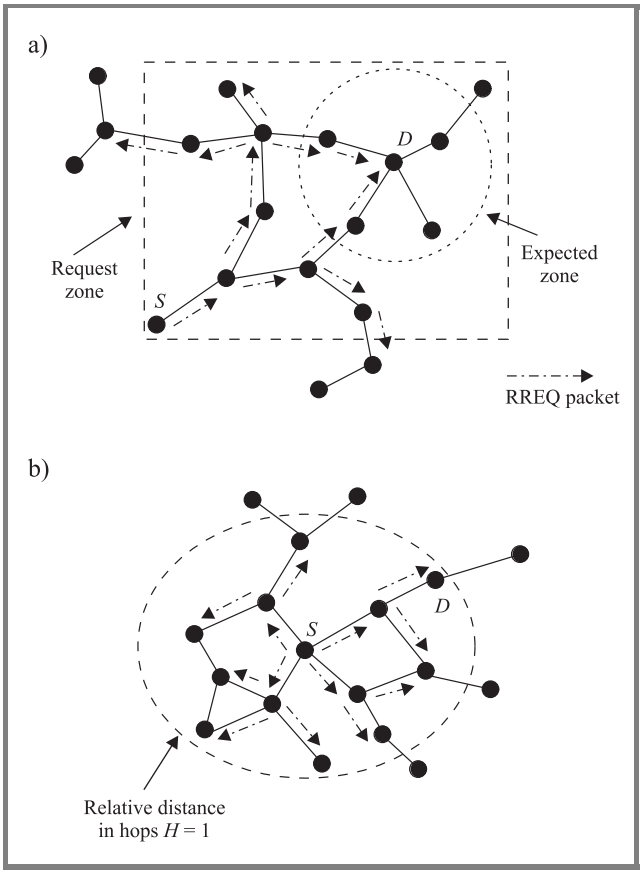


Fig. 2. Controlled flooding using restricted search zone: (a) localized RREQ propagation in LAR1; (b) localized RREQ propagation in RDMR.

example, if node *S* (Fig. 1) wants to find a route to node *A*, it will create a RREQ packet with a time to live (TTL) of one, which means that only the neighbouring nodes *Y*, *B*, *E* and *C* will see the packet. Now, since nodes *E* and *B*

have a link to node A, they can send back a RREP to node S. As a result, a route between node S and node A can be established without flooding the entire network. If node A was more that one hop away, then node S will timeout if no route reply is received and generate another RREQ packet with a higher TTL value. In RSZ, given that the source node has some idea of the current location of the destination or knows approximately how many hops away it is, it can calculate a region in which the destination node can currently reside and flood within that region only. Two such protocols which use RSZ are LAR1 and RDMAR. In LAR1, if the source node has a location information (through a GPS) about a particular node, it can calculate a region called the expected zone, in which the destination node can reside. If the source node is outside of the expected zone, a request zone (which is a region surrounding the expected zone) is also calculated. The source node will then restrict RREQ packet to the nodes within the request zone only (Fig. 2a). In RDMAR, the source nodes estimate the number of hops the destination is away from it (assuming a moderate velocity), thus restricting the route discovery within the calculated number of hops (Fig. 2b).

2.2. Tristate route discovering strategy

As discussed earlier, LPAR is a point-to-point based routing strategy, which has been built on the top of AODV. However, in LPAR, each node also exchanges location information (using GPS coordinates) in their hello message beaconing. In LPAR, if a node has location information for a required destination, it will use different route discovery strategies to determine a route, depending on the recored location and velocity of the destination. The aim of our 3-state route discovery strategy is to minimise routing overhead introduced into the network for each route discovery, while selecting relatively stable routes. The routing discovery strategies used in our 3-state algorithm are as follows:

- directed unicast route discovery (DURD),
- restricted search zone,
- expanding ring search.

When a node has data to send to a particular node and a location information is available, it will initiate the 3-state route discovery algorithm. Otherwise ERS route discovery will be used to determine a route. In our 3-state RD algorithm, the source node will first attempt to find a route to the required destination using our DURD algorithm. If the discovery was unsuccessful, RSZ strategy will be used to search over a wider scope. Finally, if the RZS strategy fails ERS will be used to determine a route. To illustrate how the 3-state algorithm works, suppose that node S (Fig. 3) wants to send data to node D and the known route had expired. Now assume that node S has recorded location information (x, y) and velocity information V for D at t_0, and the current time is t_1. Then, the possible migrating distance for D is $d_m = V(t_1 - t_0)$. Furthermore, a maximum

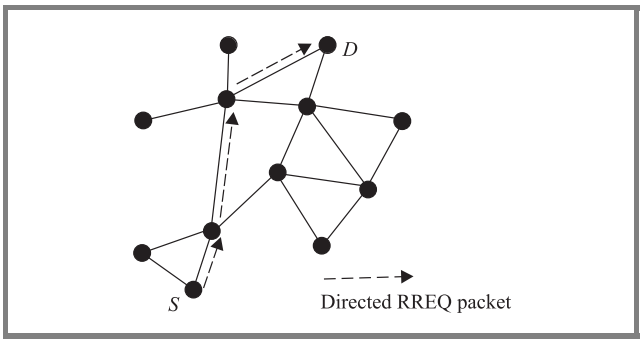


Fig. 3. Directed unicast RREQ propagation.

migration distance (MMD) is assigned[1], if $d_m \leq MMD$ and $d_m \leq d_s d$, then DURD will be initiated. The aim here is to increase the accuracy of the DURD algorithm, since only one packet is forwarded. Therefore, we will use DURD if the destination has not migrated too far from its known location and it has not migrated to the opposite side of the source. In DURD, the source node will attempt to send one packet through a selected node towards the destination. The selected node must lead towards the destination, have at least one outgoing link and meet the stability criterion (this is discussed later). Each intermediate node will follow the same procedure until the destination is reached. The DURD algorithm is outlined below[2].

Algorithm DURD
(∗ The DURD algorithm ∗)
1.
2. $N \leftarrow$ set of neighbours
3. $C_d \leftarrow 10000$ (∗ Closest distance found so far ∗)
4. $F_N \leftarrow NULL$ (∗ chosen forwarding neighbour so far ∗)
5. $D_d \leftarrow dist(node, destination)$
6. D_i distance between neighbour N_i and destination
7. D_f distance between neighbour N_i and this node
8. **for** $i \leftarrow 1$, $N_i \neq NULL$, $i++$
9. $D_i \leftarrow dist(N_i, destination)$
10. $D_f \leftarrow dist(N_i, node)$
11. **if** $Deg(N_i) > 1$ and $D_f < \tau$
12. **if** $D_i < D_d$ and $D_f < C_d$
13. $F_N \leftarrow N_i$
14. $C_d \leftarrow D_f$
15. **return** D_F

If DURD fails to find a route to the destination or if $d_m > MMD$, the source will calculate a RREQ propagation region (similar to LAR1), and attempt to find a route using RSZ. If unsuccessful, the source will increase the RSZ

[1]MMD is defined as a simulation parameter, we set $MMD = R/2$ where R is the maximum transmission range. Also, $d_s d$ is the distance between the source and the destination.

[2]τ = max allowable distance between two nodes.

and another localised route discovery is initiated. Finally, if DURD and RSZ both fail, or location information is not available, then ERS will be initiated (note that the radius of ERS will be adjusted to cover the previously calculated propagation region in RSZ, if RSZ was used prior to ERS).

2.3. Adaptive data forwarding

Another way to reduce routing overheads in the network is by reducing the effects of link breakage during data transmission. A number of different strategies have been proposed to reduce the overhead costs of link failure, these include:

- localised route maintainance (AODV, ABR),
- storing multiple routes (DSR, LAR1),
- backup routing using promiscuous overhearing (AODV-BR [12]).

Localised route maintainance, reduces routing overheads by repairing the route at the point of failure, by initiating a controlled flooding (similar to a RSZ) around the point of failure rather than initiating another route discovery at the source. Storing multiple routes (commonly used in source routing protocols such as DSR) can also be used to reduce the number of route recalculations at the source. However, this method still requires a RERR to be send back to the source. Furthermore, there is no guarantee that the source will have alternate route or whether it will still be valid. Link failure overhead can be also reduced by maintaining backup routes at every intermediate node in the route. For example, in AODV-BR, the node detecting the link failure broadcasts the data packets to the neighbours. The receiving neighbours with a route to the next hop unicast the data to the next hop. The disadvantage of this strategy is the redundancy, as multiple nodes maybe sending the same data to the next hop. Additionally, the forwarding nodes can alter the data, which introduces further security problems.

We propose a GPS-based alternate route selection (GARS) strategy, where each node can select another node as the secondary route, if the primary route fails. Similar to AODV-BR, in GARS, the alternate routes are calculated during a route reply phase. However, instead to building backup routes using promiscuous overhearing at each neighbouring node, the node sending the route reply also selects another neighbour, which can be used as a secondary route in case this node is no longer available. For example, during RREP, node B (Fig. 4) can select[3] node A as the secondary route to connect node L and E. This is done by calculating the distance between E and A and also L and A, if both these distances are less than the maximum allowable transmission range, then node B assign node A as an alternate path. Node L will accept node A as a secondary route if it forms a direct link with node A. Note that the RREP packet also contains the node id of the next node which leads to the destination. If a secondary route is used the node id of the second-hop is passed (using the IP options field at the moment), to the node in the secondary route. Therefore, the node in the secondary route can forward the data packet to the next hop which leads to the destination. For example, node L (Fig. 4) passes the node id, E, to node A, if the secondary route is used. Therefore, node A will then know that it should forward the data packet to node E unless it knows a better route.

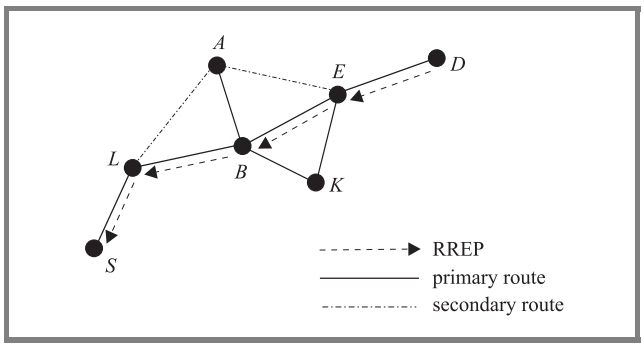


Fig. 4. Alternate route selection using GARS strategy.

The GARS algorithm is outlined below.

Algorithm GARS
(∗ The GARS algorithm ∗)
1.
2. $N \leftarrow$ set of neighbours
3. $SN \leftarrow NULL$ neighbour used as secondary route
4. $d_f \leftarrow 10000$
5. $d_r \leftarrow 10000$
6. $d_T prev \leftarrow d_f + d_r$
7. $d_T curr \leftarrow 0$
8. $T_x \leftarrow$ max transmission range
9. **for** $i \leftarrow 1$, $N_i \neq NULL$, $i{+}{+}$
10. $d_f \leftarrow dist(N_i, uplinknode)$
11. $d_r \leftarrow dist(N_i, downlinknode)$
12. **if** $d_f < T_x$ and $d_r < T_x$
13. **then** $d_T curr = d_f + d_r$
14. **if** $d_T curr < d_T prev$
15. $S_N \leftarrow N_i$
16. $d_T prev \leftarrow d_T curr$
17. **return** SN

The advantage of GARS compared to AODV-BR is that we eliminate data redundancy by specifying which node can be used as the secondary relay point if the primary relaying node is no longer available. Furthermore, security is increased since a known node is selected as a secondary relay point rather than relying on an unknown nodes to forward the data.

[3]Assuming all nodes have equal transmission range.

2.4. Stable route selection

Stable route selection can also contribute to reducing the total amount of routing overhead transmitted in the network. By selecting routes which last longer, the number of route recalculations due to link failure can be reduced. Most of the previous work done to provide stable routes in MANETs have been carried out with source routing protocols. ABR [13] and SSA [14] are two such protocols, which attempt to provide stable routes using source routing. In these protocols, the destination selects the route, which has travelled over the most stable links. We explore the effects of selecting stable routes in point-to-point routing. One way to select stable routes in a point-to-point manner is to restrict the flooding of RREQ packets over strong links only. To select strong link, we allow only the nodes which receive a RREQ packet over a strong link to further broadcast the packet. Therefore, the RREQ packets which reach the destination (or an intermediate node with a route to the destination) have travelled over strong route. This means that the destination (or the intermediate node) can send back a RREP over strong links, and a stable route between the source can be established. We define a link as being strong if the distance between the edges (nodes) in the link are less than a predefined threshold transmission range[4] (TTR).

3. Simulation model

The aim of our simulation study is to measure the performance of our routing strategy under changing network topology and investigate what levels of successful data delivery (and throughput) can be achieved under different network conditions. We compare the performance of LPAR under network scenarios, which have different levels of mobility, traffic and node density with a number of existing routing protocols and discuss how each protocol performs under each scenario.

3.1. Simulation environment and scenarios

The simulations were carried out using GloMoSim [15] simulation package. GloMoSim is an event driven simulation tool designed to carry out large simulations for mobile ad hoc networks. Our simulations were carried out for 50, 100, 200, 300, 400 and 500 node networks, migrating in a 1000 m $\times$ 1000 m boundary. IEEE 802.11 DSSS (direct sequence spread spectrum) was used with maximum transmission power of 15 dbm at 2 Mbit/s data rate. In the MAC layer IEEE 802.11 was used in DCF mode. The radio capture effects were also taken into account and the two-ray path loss characteristics was choosen for the propagation model. The antenna hight is set to 1.5 m and the radio receiver threshold is set to -81 dbm with the receiver sensitivity set to -91 dbm according to the Lucent's wavelan card specification [16]. A random way-point mobility model was used with the node mobility ranging from 0 to 20 m/s and pause time varied from 0 to 900 s. The simulation was run for 900 s for 10 different values of pause time and each simulation was averaged over eight different simulation runs using different seed values.
Constant bit rate (CBR) traffic was used to establish communication between nodes. Each CBR packet was 512 bytes and the simulation was run for 10 and 20 different client/server pairs with each session set to last for the duration of the simulation.

[4]TTR $<$ maximum possible transmission range.

3.2. Performance metrics

To investigate the performance of the routing protocols the following performance metrics were used:

- Packet delivery ratio (PDR): ratio of the number of packet sent by the source node to the number of packets received by the destination node.
- Control (O/H): the number of routing packets transmitted through the network for the duration of the simulation.
- Packet delivery ratio (vs) number of nodes: the percentage of packets successfully delivered as the number of nodes is increased for a chosen value of pause time.
- Control (O/H) (vs) number of nodes: the number of control packet introduced into the network as the number of nodes is increased for a chosen value of pause time.
- End-to-end delay: the average end to end delay for transmitting one data packet from the source to the destination.

The first metric is used to investigate the levels of data delivery (data throughput) achievable in each protocol under different network scenarios. The second metric will illustrate the levels of routing overhead introduced. The third and the forth metric are used to investigate the scalability of the protocols as the network grows in size. The last metric compares the amount of delay experienced by each data packet to reach their destination.

4. Results

This sections gives a discussion on the simulation results we obtained for our routing strategies. To investigate the performance of LPAR with and without stable link strategy (in Section 2.4), we ran two different versions of LPAR. These are: LPAR, which consists of Sections 2.2 and 2.3,

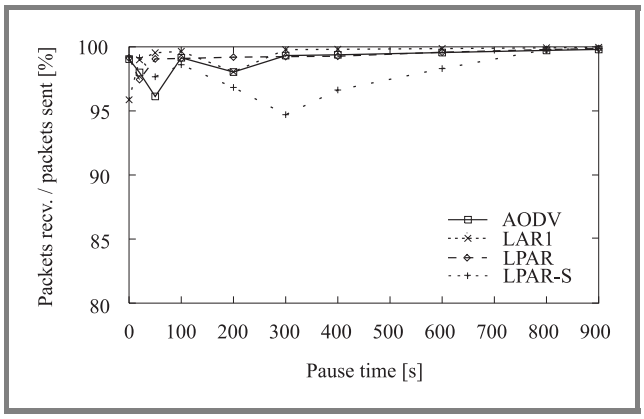


Fig. 5. PDR for 50 N and 10 s.

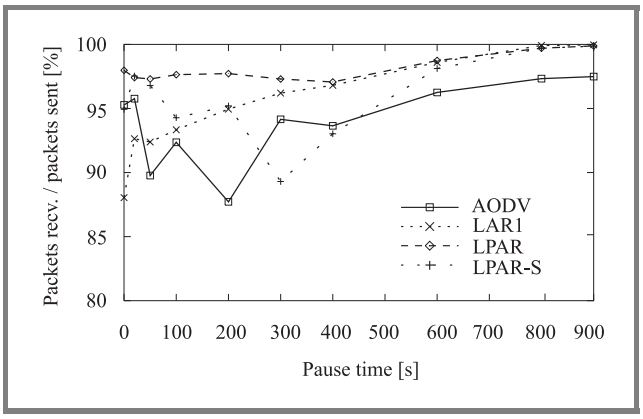


Fig. 7. PDR for 50 N and 20 s.

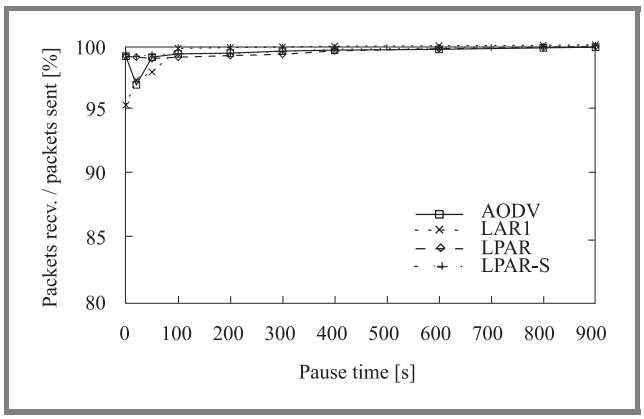


Fig. 6. PDR for 300 N and 10 s.

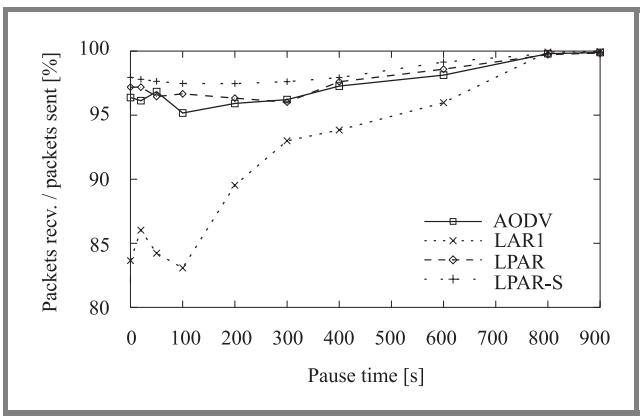


Fig. 8. PDR for 300 N and 20 s.

LPAR-S which Sections 2.2, 2.3 and 2.4. The performance of our LPAR strategies where compared with LAR1 and AODV.

4.1. Packet delivery ratio results

Figures 5 and 6 show the PDR achieved by each routing protocol as the number of nodes in the network was increased, for 10 CBR sources. In this scenario, for all node density levels, the PDR of all routing protocols are greater than 95%. The performance of each protocol converges to 100% when the mobility is reduced to zero (i.e. 900 s pause time). LAR1 has the highest level of PDR. This is more evident in Fig. 5 where the node density is lower than the other scenarios. This is because LAR1 stores multiple routes, where as the other protocols store a single route. The disadvantage of storing a single route when node density is low is that the nodes in the path to the destination have less chance of learning about a fresher route to the destination. This means that link failure between the intermediate nodes leading to the destination, may cause another route discovery. As a results some data packets maybe dropped, which means that PDR will be reduced. LPAR-S has the lowest delivery ratio in the 50 N scenario. However, as the number of nodes is increased, LPAR-S performs as well as LAR1. This is because when the node density is low, the number routes found (or available) is lower. Therefore, if route selection is done over strong links only, then the number of routes found will be lower and in some situations the RREQ packets may not reach the destination (or an intermediate node to the destination).

Figures 7 and 8 show the PDR for 20 CBR sources. In this scenario, LPAR shows the best performance under low node density (i.e. 50 node scenario), and as the node density is increased, LPAR maintains over 95% PDR. LPAR-S, still under performs in the 50 nodes scenario, however, as the node density is increased its performance increases and performs as well as LPAR and AODV. Furthermore, in the high node density scenario (i.e. Fig. 8) it begins to out perform the other routing protocols. This increase in performance is due to the availability of more stable routes when compared to the least dense scenarios. AODV also performs well across all ranges of node density. However, it starts to under perform LPAR and LPAR-S in the 300 node network scenario. LAR1 achieves the lowest levels of PDR in this scenario. This is more evident under the higher mobility (i.e. smaller pause times), where link failure rate is higher. Therefore, in this scenario, the point-to-point routing protocols clearly out performs the source routing protocol (i.e. LAR1).

4.2. Control overhead results

Figures 9 and 10 show the number of control packets introduced into the network by each routing protocol, for 10 CBR sources. In AODV, more overhead is introduced into the network than in the other routing strategies. This is because, AODV does not take any measurements to reduce the route discovery region if the source and the destination have recently communicated (or the source has location information about the destination). To the contrary, in LAR1 two factors contribute to reducing routing overhead. Firstly, nodes can have multiple routes to destinations, which may reduce the number of route discoveries initiated for each src/dest pair, whereas in AODV, each node only stores a single route. Secondly, in LAR1, if source nodes have location information about the required destination, they can use RZS, which minimises (or localises) the search area to a particular region. The advantage of this is that the number of nodes involved in broadcasting RREQ packets is reduced, which means that fewer control packets are transmitted. This also means more bandwidth to be available for the nodes that are not in the search area and reduce channel contention. LPAR and LPAR-S, which use the 3-state route discovery algorithm, produce less overhead than LAR1, despite only storing single routes. This is because in our 3-state route discovery algorithm, if unexpired location information is available, the source node will first attempt to discovery a route by unicasting rather than broadcasting. This means that fewer control packets are transmitted through the network. LPAR-S further reduces this overhead by flooding over links which have certain level of stability. The advantage of this is that route may last longer, which means fewer route recalculations will be required and fewer data packets will be dropped.

Figures 11 and 12 show the number of control packets introduced into the network by each routing protocol, for 20 CBR sources. It this scenario, it can be seen that LPAR-S continues to produce the least amount of overhead. Both LPAR and LAR1 show similar levels of overhead in low density scenario, with LPAR performing better under higher mobility and LAR1 performing better during mid range mobility. LPAR starts to out perform LAR1 at higher node density. This is because at higher node density the DURD algorithm has a better chance of forwarding the RREQ packet to the destination, which means that it will have a higher success rate for finding a route to the destination. Therefore, fewer control packets are transmitted when compared to using ERS or RSZ during route discovery. AODV continues to produce the highest level of control overhead in all scenarios. This is more evident during high mobility where AODV produces three times more control overhead than LPAR-S and two times more overhead than LPAR and LAR1. This result illustrates the importance of exploiting location information during route discovery.

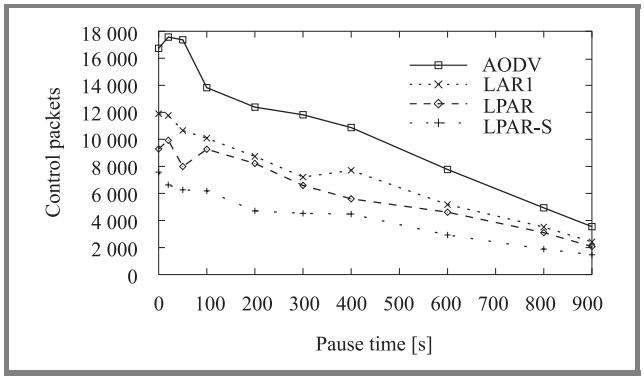


Fig. 9. CTRL packets for 50 N and 10 s.

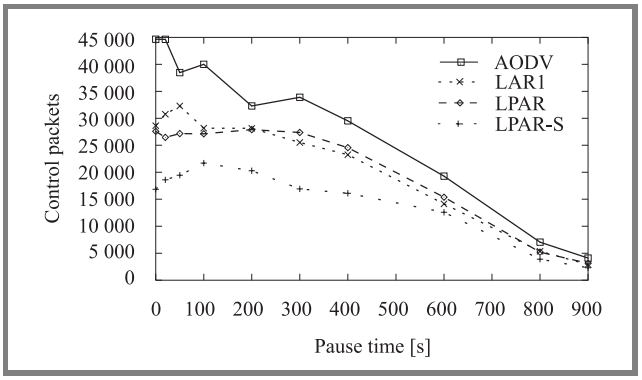


Fig. 11. CTRL packets for 50 N and 20 s.

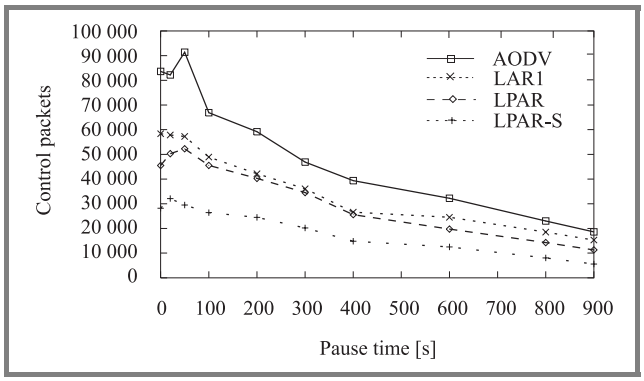


Fig. 10. CTRL packets for 300 N and 10 s.

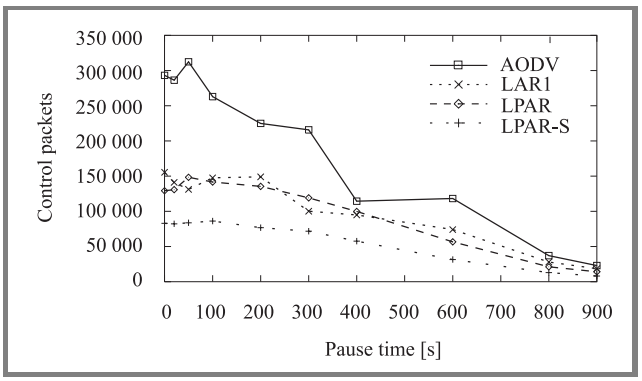


Fig. 12. CTRL packets for 300 N and 20 s.

4.3. Scalability results

To further investigate the scalability of each routing protocol, PDR and control overhead were recorded for the worst case network scenario (i.e. under constant node mobility, 0 pause time), for up to 500 nodes. Figures 13 and 14 show the PDR achieved for 10 src/dest pairs and 20 src/dest pairs respectively. For the 10 src/dest scenario, LPAR and LPAR-S achieve over 98% PDR for all node density levels. LAR1 achieves its highest PDR for up to 200 nodes, after this its performance begins to drop. AODV also performs well across all node density levels. In

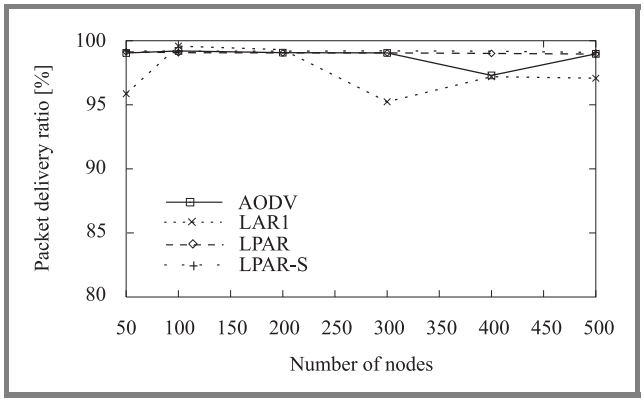


Fig. 13. PDR for pause time = 0 and 10 s.

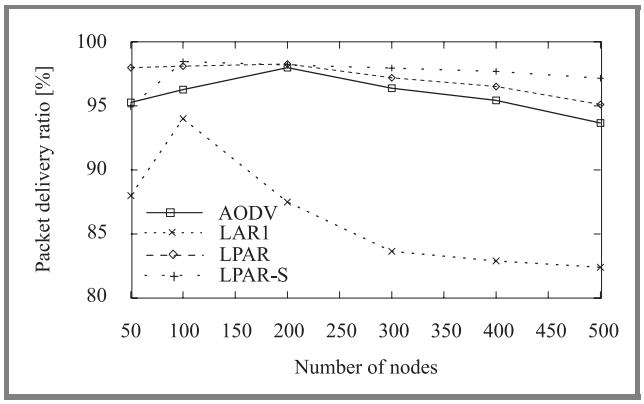


Fig. 14. PDR for pause time = 0 and 20 s.

the 20 src/dest scenario, LPAR, LPAR-S and AODV clearly out perform LAR1. LPAR-S shows the highest PDR and it maintains over 97% PDR. LPAR's performance is sightly less than LPAR-S during high node density. However, they both out perform AODV across whole range of node densities. Furthermore, AODV's performance starts to drop after 200 nodes. LAR1's highest PDR occurs at 100 nodes where it achieves 94%. However, after 100 nodes its performance continues to drop significantly, and by 500 nodes its performance has dropped to 83%.

Figures 15 and 16 show the number of control packets introduced into the network for 10 src/dest pairs and 20 src/dest pairs respectively. From these figures it can be seen that as the node density is increased the performance difference between each routing strategy becomes more significant. AODV has higher control overhead than LAR1, LPAR and LPAR-S for both the 10 src/dest scenario and the 20 src/dest scenario where it produces three times more overhead than LPAR-S and over two times more overhead than LPAR and LAR1. LPAR-S continues to produce the least amount of overhead for all node density scenario. LPAR also shows fewer overheads than LAR1 and AODV. Therefore, from these results it can be seen that both LPAR and LPAR-S are more scalable than AODV and LAR1 as the level of traffic and node density increases in the network.

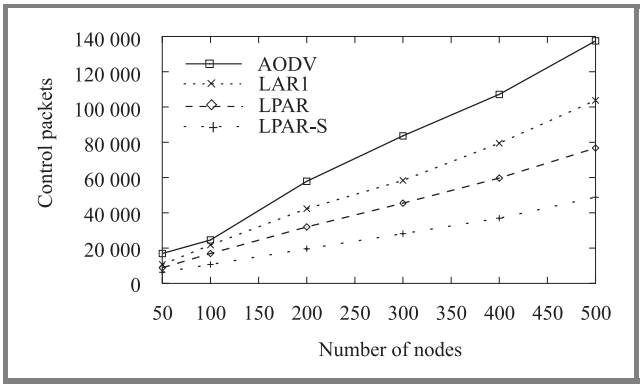


Fig. 15. CTRL for pause time = 0 and 10 s.

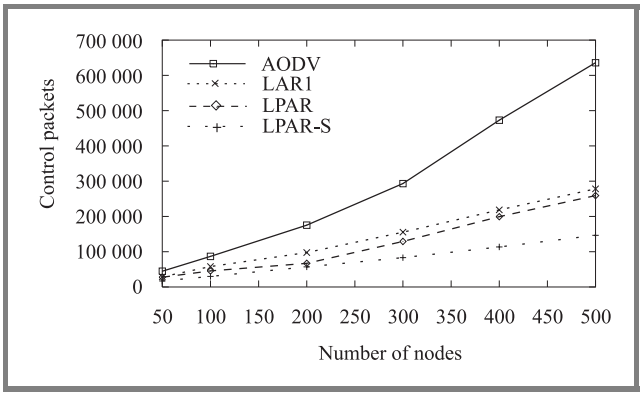


Fig. 16. CTRL for pause time = 0 and 20 s.

4.4. Delay results

Figures 17 and 18 show the average end-to-end delay experienced by each data packet for 10 src/dest pairs and 20 src/dest pairs in a 100 node network, respectively. As expected, all protocols experienced larger delays during high mobility, since more frequent link failures may cause route recalculations. This means than each packet may experience longer delays before they reach their destination. AODV has lowest end-to-end delay compared to the other protocols. This is because, AODV always uses the shortest route to the destination and it only maintains a single route, whereas LAR1 can store multiple route. This means that if

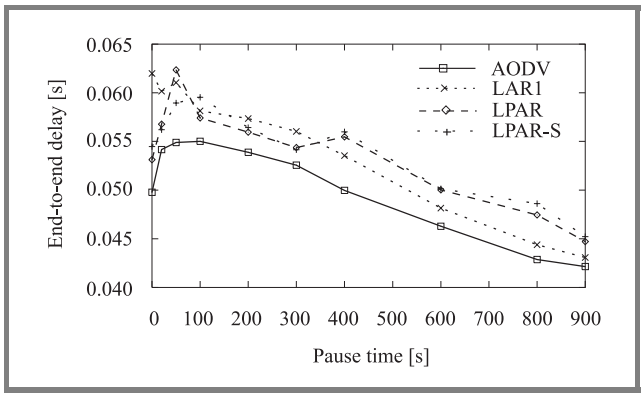


Fig. 17. Average end-to-end delay for 10 s.

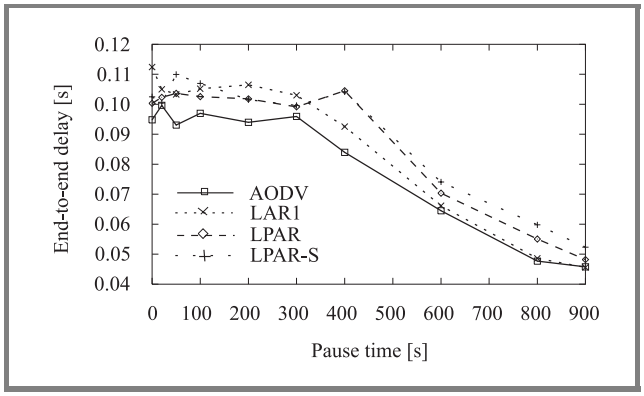


Fig. 18. Average end-to-end delay for 20 s.

optimal route fails (the one with the shortest src/dest path), an alternate route from the route cache may be use. As a result some packet may travel over longer routes to reach the destination. Similarly in LPAR and LPAR-S if the primary route fails, some packet may travel over the secondary route, which may be longer in length. Therefore, they may experience slightly longer delay. From the figures we can see that LPAR and LPAR-S have on average about 5 ms more delay across the whole range of mobility. However, by using a secondary route, LPAR and LPAR-S are able to successfully transmit more data packets, and reduce the number of route recalculations, which means fewer control packets.

5. Conclusions

This paper describes a new routing strategy for mobile ad hoc networks. We present LPAR routing protocol, which introduces a number of different strategies to reduce route discovery overhead and the power consumed by each node. We compared LPAR with LAR1 and AODV using simulations. Our results show that LPAR and LPAR-S produce lower overhead than LAR1 and AODV, while still maintaining high levels of data delivery when node density is low. In high node density both LPAR and LPAR-S produce fewer overheads and maintain higher levels of data throughput than AODV and LAR1.

References

[1] C. E. Perkins and T. J. Watson, "Highly dynamic destination sequenced distance vector routing (DSDV) for mobile computers", in *ACM SIGCOMM'94 Conf. Commun. Archit.*, London, UK, 1994.

[2] S. Murthy and J. J. Garcia-Luna-Aceves, "A routing protocol for packet radio networks", in *Mobile Computing and Networking*. 1995, pp. 86–95.

[3] M. Abolhasan, T. Wysocki, and E. Dutkiewicz, "A review of current on-demand routing protocols", in *Netw. – ICN 2001, First Int. Conf.*, Colmar, France, 2001.

[4] D. Johnson, D. Maltz, and J. Jetcheva, "The dynamic source routing protocol for mobile ad hoc networks", in *Internet Draft, draft-ietf-manet-dsr-07.txt*, work in progress, 2002.

[5] S. Das, C. Perkins, and E. Royer, "Ad hoc on demand distance vector (AODV) routing", in *Internet Draft, draft-ietf-manet-aodv-11.txt*, work in progress, 2002.

[6] Z. J. Hass and R. Pearlman, "Zone routing protocol for ad-hoc networks", in *Internet Draft, draft-ietf-manet-zrp-02.txt*, work in progress, 1999.

[7] M. Joa-Ng and I.-T. Lu, "A peer-to-peer zone-based two-level link state routing for mobile ad hoc networks", *IEEE J. Selec. Areas Commun.*, vol. 17, no. 8, 1999.

[8] S.-Ch. Woo and S. Singh, "Scalable routing protocol for ad hoc networks", accepted for publication in *Journal of Wireless Networks (WINET)*.

[9] S.-Y. Ni, Y.-C. Tseng, Y.-S. Chen, and J.-P. Sheu, "The broadcast problem in a mobile ad hoc network", in *Fifth Ann. ACM/IEEE Int. Conf. Wirel. Netw.*, 2002.

[10] Y.-B. Ko and N. H. Vaidya, "Location-aided routing (LAR) in mobile ad hoc networks", in *Proc. Fourth Ann. ACM/IEEE Int. Conf. Mob. Comput. Netw. (Mobicom'98)*, Dallas, USA, 1998.

[11] G. Aggelou and R. Tafazolli, "RDMAR: a bandwidth-efficient routing protocol for mobile ad hoc networks", in *ACM Int. Works. Wirel. Mob. Multimed. (WoWMoM)*, 1999, pp. 26–33.

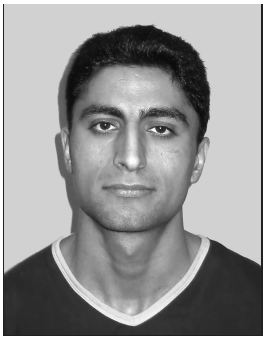
[12] S. Lee and M. Gerla, "Aodv-br: backup routing in ad hoc networks", in *Proc. IEEE Wirel. Commun. Netw. Conf. (WCNC)*, Chicago, USA, Sept. 2000.

[13] C. Toh, "A novel distributed routing protocol to support ad-hoc mobile computing", in *IEEE 15th Ann. Int. Pheonix Conf.*, 1996, pp. 480–486.

[14] R. Dube, C. Rais, K. Wang, and S. Tripathi, "Signal stability based adaptive routing (SSA) for ad hoc mobile networks", in *IEEE Pers. Commun.*, Feb. 1997, pp. 36–45.

[15] "Glomosim scalable simulation environment for wireless and wired network systems", http://pcl.cs.ucla.edu/projects/glomosim/

[16] "Orinoco pc-card", http://www.lucent.com/orinoco



Mehran Abolhasan recieved the B.E. computer engineering with honours from the University of Wollongong. He is currently undertaking a Ph.D. in School of Computer, Electrical and Telecommunications Engineering, University of Wollongong, Australia. His research interests are mobile ad hoc network scalability, MAC, routing protocols and QoS.

e-mail: mehran@titr.uow.edu.au

Telecommunication and Information Research Institute

University of Wollongong

Wollongong, NSW 2522, Australia

Tadeusz Antoni Wysocki received the M.Eng.Sc. degree with the highest distinction in telecommunications from the Academy of Technology and Agriculture, Bydgoszcz, Poland, in 1981. From then till the end of 1991, he was with the Academy of Technology and Agriculture. In 1984, he received his Ph.D. degree, and in 1990, was awarded a D.Sc. degree (habilitation) in telecommunications from the Warsaw University of Technology. In January 1992, Dr. Wysocki moved to Perth, Western Australia to work at Edith Cowan University. He spent the whole 1993 at the University of Hagen, Germany, within the framework of Alexander von Humboldt Research Fellowship. After returning to Australia, he was appointed a Project Leader, Wireless LANs, within Cooperative Research Centre for Broadband Telecommunications and Networking. In 1997, he became a Program Leader, Wireless Systems, within the same research center. Since December 1998 he has been working as an Associate Professor at the University of Wollongong, within the School of Electrical, Computer and Telecommunications Engineering. The main areas of Dr. Wysocki's research interest include: indoor propagation of microwaves, code division multiple access (CDMA), digital modulation and coding schemes, space-time-coding, as well as routing protocols for ad hoc networks. He is the author or co-author of four books, over 100 research publications and nine patents. He also chaired three International Symposia on DSP for Communication Systems, in 1996, 1999, and 2001, respectively, and is a Senior Member of IEEE.

e-mail: wysocki@uow.edu.au
Telecommunication and Information Research Institute
University of Wollongong
Wollongong, NSW 2522, Australia

Eryk Dutkiewicz received a B.E. degree in electrical and electronic engineering from the University of Adelaide in 1988, an M.Sc. degree in applied mathematics from the University of Adelaide in 1992 and a Ph.D. degree in telecommunications from the University of Wollongong in 1996. From 1988 to 1992 he worked at the Overseas Telecommunications Corporations in Sydney developing pioneering broadband multimedia telecommunications systems based on ATM networking. From 1992 to 1999 he conducted research and teaching in the School of Electrical, Electronic and Telecommunications Engineering at the University of Wollongong. In 1999 he joined Motorola Labs in Sydney where he is currently the manager of the WLAN Technologies Lab. His interests lie in modelling and analysis of next generation wireless networks.

e-mail: Eryk.Dutkiewicz@motorola.com
Motorola Australia Research Centre
12 Lord St, Botany, NSW 2525, Australia

Paper

Performance analysis of reactive shortest path and multi-path routing mechanism with load balance

Peter P. Pham and Sylvie Perreau

Abstract — Research on multi-path routing protocols to provide improved throughput and route resilience as compared with single-path routing has been explored in details in the context of wired networks. However, multi-path routing mechanism has not been explored thoroughly in the domain of ad hoc networks. In this paper, we analyze and compare reactive single-path and multi-path routing with load balance mechanisms in ad hoc networks, in terms of overhead, traffic distribution and connection throughput. The results reveals that in comparison with general single-path routing protocol, multi-path routing mechanism creates more overheads but provides better performance in congestion and capacity, provided that the route length is within a certain upper bound which is derivable. The analytical results are further confirmed by simulation.

Keywords — ad hoc networks, load balance, multi-path routing protocol, overheads.

1. Introduction

Mobile ad hoc networks (MANETs) are collections of wireless mobile nodes, constructed dynamically without the use of any existing network infrastructure or centralized administration. Due to the limited transmission range of wireless network interfaces, multiple hops may be needed for one node to exchange data with another one across the network. MANETs are characterized by limited power resource, high mobility and limited bandwidth. Routing in MANETs can be accomplished through either single path or multiple paths. When using single-path routing protocols, the traffic is distributed through one route and is therefore less flexible than in multi-path routing protocols.

The problem of two entities communicating using multiple paths has been considered widely in various contexts for wired networks [1–5]. It was shown that multi-path routing mechanism provides better throughput than single-path routing protocols [2, 3]. Although research on multi-path routing protocols has been covered quite thoroughly in wired networks, similar research for wireless networks is still in its infancy. Some multi-path routing protocols for MANETs have been proposed in [6–9]. However, the performance of these protocols are only assessed by simulations in certain limited scenario. Although some recent papers provide analytical models for multi-path routing [10, 11], they are limited on a single aspect of multi-path routing such as route discovery frequency or error recovery. To the best of our knowledge, there has been no paper which provides an analytical model which allows comparing the performance of reactive shortest single-path routing and multi-path routing with load balance.

In this paper, we propose models to analyze and compare reactive single-path and multi-path routing protocols in terms of overheads, traffic distribution and connection throughput. Thereafter, the terms "single-path routing" and "multi-path routing" are equivalent to "shortest single-path routing" and "multi-path routing with load balance" respectively. In addition, we focus our analysis only on reactive routing mechanism. The overhead analysis in this paper is only applicable for reactive routing mechanism. However, the results regarding the traffic distribution and connection throughput is also applicable for both proactive and hybrid routing mechanisms. The outcome from analytical models is further validated by simulation.

The remaining of this paper is organized as follows. Section 2 gives a detailed analysis of overhead for both single-path and multi-path routing techniques. In Section 3, we analyze the traffic distribution for both mechanisms and Section 4 concentrates on the capacity analysis. We finally conclude this study and discuss future research directions in Section 5.

2. Overheads analysis

2.1. Qualitative overheads analysis

Overheads in reactive routing protocols are caused in the following phases: *Route Discovery*, *Route Maintenance*, and *Data Transmission*. In this section, we describe these phases and also briefly comment on the amount of overhead they involve for both single-path and multi-path routing. A quantitative study, which provides numerical values is proposed in the next section.

2.1.1. Route discovery

In this phase, the source node broadcasts route request packets (RRQs) to find the route to the destination node. When a RRQs reach the destination, the node will response back by sending route reply packets (RRPs) to notify the source of the route path. *Route Discoveries* for single-path and multi-path routing mechanisms are shown in Fig. 1. Clearly shown, the number of broadcasted RRQs is the same for both single-path and multi-path routing. However, when

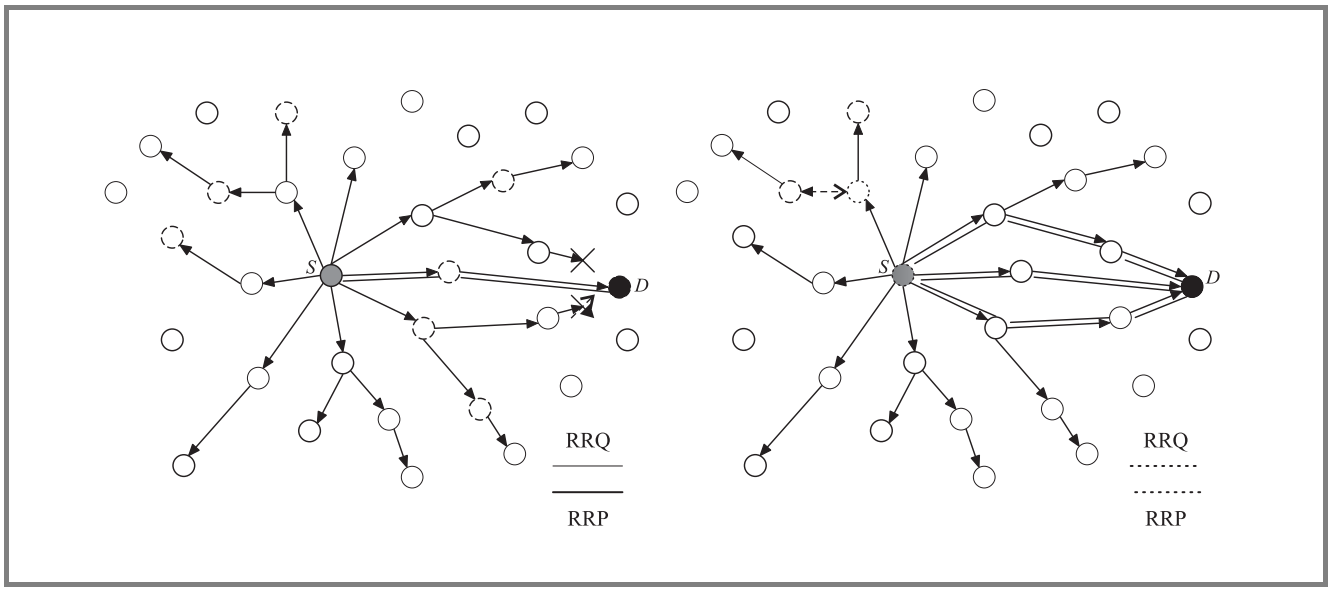


Fig. 1. Route discovery in single-path routing mechanism.

the destination sends the RRPs back to the source, because it has to send N_u (N_u being the number of multiple paths created in the *Route Discovery* phase) RRPs to correspond to N_u RRQs, the overheads of multi-path routing in *Route Discovery* phase is N_u times higher than that of single-path routing.

2.1.2. Route maintenance

In this phase, when a link is broken, an error packet (ERR) is sent back to the source to indicate the route breakage. In multi-path routing, since there are multiple paths for each source-destination pair, assuming the probability of link breakage and the route length for all the routes are the same, the number of route breakage is proportional to the number of paths. Therefore, it can be deduced that in multi-path routing, the number of ERRs is higher than in single-path routing which induces more overheads.

2.1.3. Data transmission

During this stage, the overhead portion contained in the data packets depend on the routing protocols themselves. For some protocols such as dynamic source routing protocol (DSR) [12], the complete route from the source to the destination is stored inside the overhead portion of the data packets. However, in other ones such as ad hoc on-demand distance vector routing protocol (AODV) [13], only the next node information is stored in the data packet which results in less overhead as compared with DSR.

2.1.4. Comment

In summary, we can clearly see that there is a trade-off between single-path and multi-path routing mechanisms. In multi-path routing, overheads in multi-path routing are expected to be higher than in single-path routing due to extra RRPs and ERRs. However, the frequency of route discoveries in multi-path routing is lower than in single-path routing as claimed in [11]. Hence, an analytical model is necessary to allow a better understanding of this trade-off.

2.2. Route creation frequency

Let us firstly review the results of [11]. This significant result indicates that the route creation rate for multi-path routing strategy is lower than it is for single-path routing. The link lifetimes are assumed to be independent and identically distributed (iid) exponential random variables with mean l. Since a route fails when any link in its path breaks, the lifetime of a route with L links is also an exponentially distributed random variable with a mean of l/L.

Theorem 1. *Denoting by $\mu_i = l/L_i$, the probability density function (pdf) of T, the time between successive route discoveries, is given by:*

$$f_T(t) = \prod_{i=1}^{N}(1-\exp(-\mu_i t))\sum_{i=1}^{N}\mu_i\frac{\exp(-\mu_i t)}{1-\exp(-\mu_i t)}. \qquad (1)$$

Comment. The expected value of T can be derived by knowing the hop-wise lengths of all the routes k_i, $i = 1,\ldots,N$. It was also shown in [11] that using multi-path routing can achieve 25% reduction in route discoveries rate for 3–4 hops routes as compared with single-path routing. This reduction is because in multi-path routing, route discovery is only initiated when all the routes to the destination are broken whereas in single-path routing, it is done when one single route is broken.

2.3. Overhead analysis using analytical model

2.3.1. Network model

We assume that mobile nodes are distributed uniformly with node density δ inside a circle of radius R. We also assume that there are N nodes in the network. N is related to the node density and the circle radius by the following expression $N = \pi R^2 \delta$. Each link has a link breakage rate of μ, i.e. a link has a average lifetime of $1/\mu$ seconds on average. Furthermore, we assume that the average route length (in terms of number of hops) for single-path routing is L_s and for multi-path routing is L_m. Since single-path routing mechanism uses shortest routes, we obviously have $L_m > L_s$. In addition, L_e is assumed to be the average length of the route from the source to the node where a link breakage occurs. For multi-path routing, N_u represents the number of paths for each source-destination pair. In addition, the number of active connections per node is denoted by A_c for both routing mechanisms. Furthermore, the size of route request packet, route reply packet and error packet are denoted as M_{rq}, M_{rp}, and M_e, respectively. Finally, a route discovery takes T seconds to find the routes to the destination. All the parameters are summarized in Table 1.

Table 1
Summary of parameters

Notation	Definition
N	Number of nodes
N_u	Number of routes per source-destination pair
L_e	Average length of error route
μ	Link breakage rate
L_s	Average length of a route for single-path routing
L_m	Average length of a route for multi-path routing mechanism
A_c	Number of active routes per node
M_{rq}	Size of the request packet
M_e	Size of error request packet
M_{rp}	Size of reply packet
ε	Inter-arrival rate
P	Overhead portion of a data packet
M_d	Size of the data packet
T	Average delay for route creation
λ_s	Route discovery frequency for single-path routing
λ_m	Route discovery frequency for multi-path routing

2.3.2. Overhead due to RRQs

- Single-path routing mechanism:
  Assuming that N nodes each broadcast a RRQ λ_s times per second, the total overhead created by RRQs is obviously $M_{rq}\lambda_s N^2$. λ_s (i.e the route discovery frequency) is related to link breakage as $\lambda_s = \mu L_s$. Hence, the amount of overheads due to the RRQs is $M_{rq}\mu L_s N^2$.
- Multi-path routing mechanism:
  Using a similar argument as above, the amount of overheads due to RRQs is $M_{rq}\lambda_m N^2$ where λ_m is the frequency of route discovery for multi-path routing algorithm. This parameter can be calculated using *Theorem 1*.

2.3.3. Overhead due to RRPs

- Single-path routing mechanism:
  Reply packets follow L_s hops to return back to the source. Since the rate of sending the RRPs is the same as the rate of sending RRQs, the overhead created by the RRPs, is $M_{rp}\mu L_s^2 N$.
- Multi-path routing mechanism:
  Since the destination node replies to N_u RRQs, the overhead due to RRPs is $M_{rp}\lambda_m L_m N N_u$. Note that the fact that λ_m is smaller than λ_s balances the fact that the number of RRPs are increased by a factor of N_u compared to single-path routing.

2.3.4. Overheads due to ERRs

When a link is broken, an error packet is sent back to the source to signal the link breakage. Recall that L_e is the average length of the path from the broken link to the source ($L_e < L_s < L_m$). Since the error packet has to travel L_e links to the source, this effectively produces L_e error packets per route broken.

- Single-path routing mechanism:
  Since the link breakage rate is μ, the route breakage rate for a route with L_s links is μL_s. For each node, the average number of active routes is A_c. Therefore, for a node, the route breakage rate is $\mu L_s A_c$. Therefore, in a N-node network, the average number of overheads due to error packets is $\mu L_s A_c N L_e M_e$.
- Multi-path routing mechanism:
  In multi-path routing, since each source-destination pair maintains N_u routes, the overhead due to error packets is $N_u \mu L_m L_e A_c N M_e$.

2.3.5. Overheads due to data transmission

The overheads created during data transmission are due to the overhead portion of data packets. We assume that the each route discovery is accomplished in T seconds on average. Furthermore, each mobile node is a simple source with data transmission rate of ε once the route discovery is completed.

- Single-path routing mechanism:
  Since the route discovery rate is λ_s, the interval between each route discoveries is on average $1/\lambda_s$. Each route discovery takes on average T seconds. Therefore, the actual time for data transmission is $(1/\lambda_s - T)$ seconds. The number of data packets sent during that interval is $(1/\lambda_s - T)\varepsilon$. Thus, data packets are sent with an average rate of $\lambda_s \varepsilon (1/\lambda_s - T)$

packets/sec. Since each data packet has to travel L_s hops to the destination, the total amount of overhead is $\lambda_s \varepsilon(1/\lambda_s - T)PL_s = \mu L_s \varepsilon(1/(\mu L_s) - T)PL_s$.

- Multi-path routing mechanism:
  Using a similar derivation as above, the total amount of overheads for multi-path routing is $\lambda_m \varepsilon(1/\lambda_m - T)PL_m$ where λ_m can be calculated using *Theorem 1* (we do not include the derivation of this calculation in this paper due to the lack of the space).

2.3.6. Summary

The total amount of overheads due to RRQs, RRPs, ERRs and data packets for single-path and multi-path respectively denoted by Ov_s and Ov_m can be expressed as:

$$Ov_s = M_{rq}\lambda_s N^2 + M_{rp}\lambda_s L_s N + \\ + \mu L_e L_s A_c N M_e + \mu L_s \varepsilon(1/(\lambda_s - T)PL_s, \quad (2)$$

$$Ov_m = M_{rq}\lambda_m N^2 + M_{rp}\lambda_m N L_m N_u + \\ + \mu L_e L_m A_c N M_e N_u + \mu \varepsilon(1/\lambda_m - T)PL_m. \quad (3)$$

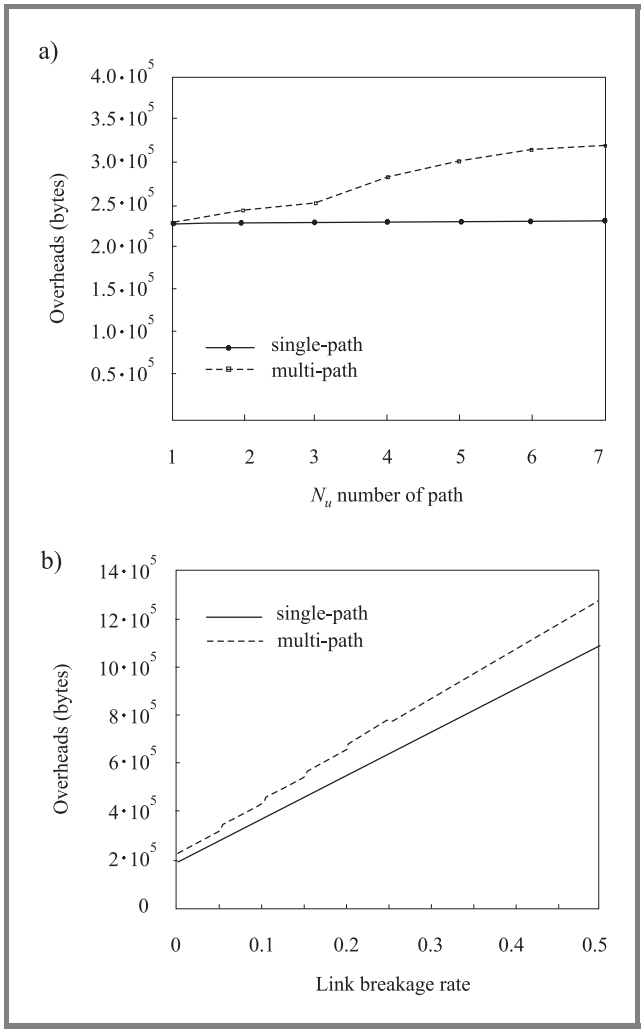


Fig. 2. Overhead comparison (a) versus N_u and (b) versus link breakage rate increases.

In Fig. 2, we have plotted Ov_s and O_m as functions of the number of paths N_u. One can see that there is no significant increase in overheads for N_u up to 3. This confirms the fact that in the literature, authors often mentioned that $N_u = 3$ provides an optimum trade off [3, 11]. This claim is usually based on simulation results and the study provided in this paper confirms this observation. In Fig. 2, $N_u = 3$ and Ov_s and Ov_m are compared as the link breakage is varied. It is interesting to note that the maximum increase in overheads is approximately 20% (for a link breakage rate of 50%). Otherwise, for link breakages lower than 10%, the increase in overhead is approximately 10%. One might argue that the figure is not insignificant. In fact, assessing whether this increase in overhead is acceptable or not really depends on the advantages brought out by multi-path routing. This is why a theoretical study in the following sections is necessary.

2.4. Simulation results

In the simulation, we choose dynamic source routing [12] and multi-path routing protocol with load balance (MRP-LB) [14] as typical candidates for shortest path and multi-path routing protocols respectively. The choice of these routing protocols does not limit the applicability of this result into the others. In other words, the result which is derived above is applicable to other reactive routing al-

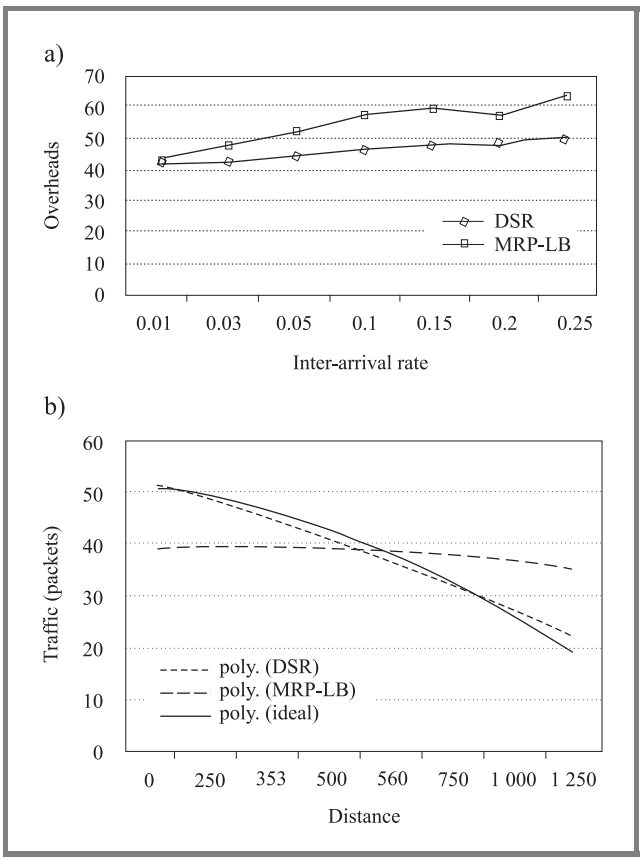


Fig. 3. (a) Overheads of DSR and MRP-LB; (b) traffic versus distance from circle centre.

gorithms such as ad hoc on-demand distance vector routing protocol [13], or temporally-ordered routing algorithm [15]. However, the result is not suitable for proactive and hydrid routing protocols.

Clearly seen from Fig. 3, MRP-LB exhibits higher overhead than DSR which once again confirms the correctness of our analytical model.

3. Traffic analysis of single shortest path and multi-path load balancing routing mechanisms

The following section compares the traffic distribution for the shortest-path and load-balancing routing mechanisms. We will be able to quantify the advantages in terms of congestion avoidance of the load-balancing routing mechanism over the shortest-path one. In particular, we will be able to determine the network conditions, i.e. network density, node-to-node transmission rate, and node processing rate, in which networks multi-path routing really present interest. We will also derive an upper bound for a certain parameter which will guarantee that when multi-path routing mechanism is worth considering, i.e. it results in congestion decrease.

3.1. Network model

In the model, we assume that mobile nodes are situated inside a circle with radius R. Furthermore, they are distributed uniformly with density δ. In addition, mobile nodes communicate with each other at a uniform rate λ. Each node is assumed to have the same processing power of η. Clearly, we can see that the traffic going through each node consists of two types, i.e. the common traffic which is defined as a point-to-point communication traffic between nodes and the relay traffic which is defined as the forwarding traffic caused by data packets travelling through multiple hops to the destination. The parameters to be used in the analysis are summarized in Table 2.

Table 2
Summary of parameters for traffic analysis

Notation	Definition
R	Radius of the circle
δ	Node density
λ	Node-to-node transmission rate
λ_m	Node-to-node transmission rate for multi-path routing
η	Node processing rate
r	Distance of the node of interest to the circle center
L_m	Average length of a route in multi-path routing

3.2. Analysis of the shortest path routing algorithm

It can be proven in Appendix A that the total traffic going through a node located at a distance r from the center of the circle, $\lambda(r)$ can be expressed as follows:

$$\lambda(r) = (\pi R^2\delta - 1)\lambda + \frac{\pi(R^2 - r^2)^2\delta^2\lambda\beta}{2}. \tag{4}$$

Therefore, according to Little theorem [16], the average number of packets in the queue for a node located at a distance r from the center of the circle is:

$$N_{pac}(r) = \frac{\lambda(r)}{\eta - \lambda(r)}. \tag{5}$$

From the above equation, the total number of congested packets in the circle is:

$$N_{pac_{total}} = \int_0^R 2\pi r\delta N_{pac}(r)dr. \tag{6}$$

Hence, the average number of packets in a queue can be evaluated as:

$$N_{pac_s} = \frac{1}{\pi R^2\delta}\int_0^R 2\pi r\delta N_{pac}(r)dr. \tag{7}$$

The exact calculation of N_{pac_s} is shown in the Appendix B. It is important to know that N_{pac_s} can be exactly evaluated by integration and is a good indicator of the general congestion of the network.

3.3. Analysis of the multi-path load balancing routing mechanism

A perfect load balancing multi-path routing mechanism distributes the traffic evenly among nodes in the network. As a consequence, "hot-spots" are eliminated. Therefore, packets are expected to experience lower average end-to-end delay. Suppose that L_m, λ_m and η are respectively the average length of a route in a network, the node to node traffic rate, and the processing rate. Let us evaluate the total traffic within the network. Since the number of nodes is $\pi R^2\delta$, it is easy to see that the total number of possible connections within the network is $(\pi R^2\delta - 1)\pi R^2\delta$. With an average route length between two nodes of L_m the total traffic within the network is $(\pi R^2\delta - 1)\pi R^2\delta\lambda_m L_m$. Therefore, the incoming traffic per node is $(\pi R^2\delta - 1)\lambda_m L_m$ and the average number of packets in the queue per node is:

$$N_{pac_m} = \frac{(\pi R^2\delta - 1)\lambda_m L_m}{\eta - \pi R^2\delta - 1)\lambda_m L_m}. \tag{8}$$

In order to ensure that the load balancing policy decreases the congestion level of the network, N_{pac_m} should be smaller than N_{pac_s}. One can see in the above equation that the key parameter which controls N_{pac_m} is the average length of a route. Indeed, in order to have $N_{pac_m} < N_{pac_s}$, L_m must satisfy:

$$L_m < \frac{N_{pac_s}\eta}{(N_{pac_s} + 1)(\pi R^2\delta - 1)\lambda_m} = L_{max}. \tag{9}$$

This result shows that if $L_m > L_{max}$, using a load balancing routing mechanism is no longer beneficial as compared with a shortest-path routing scheme. This can be easily implemented in practice: given a network characterized by its node density, its size and the traffic rate, one can evaluate N_{pacs}. This value can then be used to calculate the theoretical value for L_{max} which is interesting because the result of this section can be used as a criterion to select the route in multi-path routing mechanism.

3.4. Simulation results

Similarly to the previous section, DSR and MRP-LB are used to measure the traffic versus distance from the circle center. The results obtained from DSR and MRP-LB altogether with the result of ideal shortest path routing are shown in Fig. 3. Clearly shown, DSR demonstrates a consistent behavior to ideal shortest path routing in terms of traffic allocation. In addition, nodes closer to the circle center are experiencing more traffic intensity, i.e. more congestion. However, in MRP-LB, due to the load balancing policy, mobile nodes are experiencing approximately the same traffic.

In the next section, we will investigate another issue associated with a load balancing routing mechanism, namely the connection throughput of the network.

4. Connection throughput analysis

In this section, we compare how the resources for transmission are used within the network for single-path and multi-path routing protocols. In order to conduct this study, we define the concept of connection throughput as follows:

Definition. The connection throughput of a network is defined as the average transmission rate of a connection in the network.

Note that the higher is the connection throughput, packets are experienced lower delay during transmission. Therefore, the connection throughput is a good indicator of the average end to end delay in the network. Intuitively, we can see that congestion restricts the full usage of the available bandwidth. In other words, assuming that every route can support in theory a transmission at W bits/seconds, the actual transmission rate of a route is limited by the fact that the bandwidth has to be shared with other routes at the MAC layer of each node. Therefore, the transmission rate of a route will be limited by the bandwidth available at the most congested node of this route. A load balancing policy which relieves "hot-spot" congestion should improve the connection throughput of the network. However, one has to be cautious since while the transmission rate in "hot-spot" areas increases due to congestion avoidance, it also decreases elsewhere in the network where more traffic is distributed. There is therefore a trade-off needed to consider when applying multi-path routing mechanism. An interesting parameter characterizing the performance of multi-path routing is the average route length (calculated in number of hops). When this parameter increases, it results in more nodes in the network involved in connection, which means that more traffic is distributed across the network. In the following section, we propose an upper bound on the average length of a route in multi-path routing, which guarantees that the connection throughput is improved as compared to single-path routing.

4.1. Single-path routing

In this section, we use the same network model as in Section 4. According to Eq. (4), when a single-path routing mechanism is used, nodes closer to the circle center are experiencing more traffic, i.e. are more congested. Therefore, in terms of capacity, the total capacity of the network is limited by the capacity of the area close to the circle center. Considering a connection between nodes A_1 and A_2, let us denote by A, the orthogonal projection of the circle center O on the line A_1A_2. Assume that there is a node on the route between A_1 and A_2 very close to A. Since this particular node is closer to the circle center than any other nodes on the route on the line A_1 and A_2, it experiences the highest traffic. Therefore the data transmission rate on this particular route is limited by the congestion experienced by node close to A. From now on, we denote the node closed to A is node A for simplicity. It can be easily seen from Eq. (4) that the number of routes going through node A can be expressed as:

$$n(r) = (\pi R^2\delta - 1) + \frac{\pi(R^2 - r^2)^2\delta^2\beta}{2}. \qquad (10)$$

Assuming that we have a fair MAC layer, each route is allocated an equal bandwidth for data transmission. Therefore, each route going through node A will be allocated the bandwidth denoted by $W(r)$ expressed as:

$$W(r) = \frac{W}{(\pi R^2\delta - 1) + \frac{\pi(R^2 - r^2)^2\delta^2\beta}{2}}, \qquad (11)$$

where W is the total bandwidth allocated to the network. It can be recalled that $N = \pi R^2\delta$, the total number of nodes in the network. Because this number is large, we also assume that $\pi R^2\delta - 1 \approx N$.

Let us now evaluate the number of routes which transmission rate is limited by node A. Note that these routes have to be approximately perpendicular to OA and go through A. One can in Fig. 4 that these routes are such as their source and destination nodes are respectively in the areas R_1 and R_2, and vice versa. The number of nodes in each area can be expressed as:

$$N_{R_1}(r) = N_{R_2}(r) = (R^2 - r^2)\beta\delta. \qquad (12)$$

The derivation which leads to this results is very similar to the one leading to Eq. (4). We will therefore refer our reader to Appendix A for more details. From this, the number of routes which transmission rates are limited

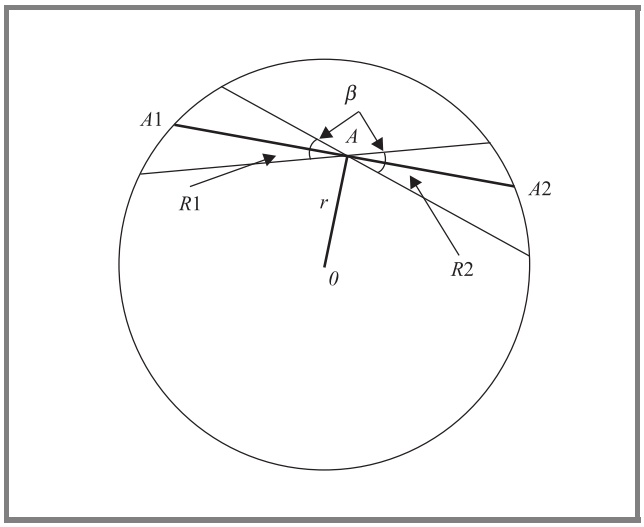


Fig. 4. Network model for connection throughput.

by $W(r)$ is simply $2N_{R_1}(r)N_{R_2}(r)$. Note that any node in the ring delimited by r and $r+dr$ with dr small enough will have the same traffic characteristics as $A(r)$. Therefore, it can be shown that W_{sp}, the total bandwidth used by the network will be expressed as:

$$\begin{aligned} W_{sp} &= \int_0^R W(r)2N_{R_1}(r)N_{R_2}(r)2\pi r\delta dr = \\ &= 2W\delta\int_0^R \frac{(R^2-r^2)^2\beta^2\delta^2}{(\pi R^2\delta-1)+\frac{\pi(R^2-r^2)^2\delta^2\beta}{2}}2\pi r dr = \\ &= 2W\sqrt{\frac{\beta N}{2\pi}}\left(\sqrt{\frac{\beta N}{2\pi}}-\arctan\sqrt{\frac{N\beta}{2\pi}}\right). \end{aligned} \tag{13}$$

Note that we have used the fact that $\pi R^2\delta = N$. The total number of possible connections being N^2, the connection throughput for this network using a single-path routing mechanism is $\lambda_{sp} = W_{sp}/N^2$.

4.2. Multi-path load balancing routing

Suppose that and A_c is the average number of active routes per node. Obviously, the number of active routes in the network is NA_c. Suppose L_m being the average number of hops involved in a route, the total number of connections in the whole network is NA_cL_m which means that the number of connections per node is A_cL_m. Assuming that the bandwidth available at each node is uniformly split among these connections, the bandwidth per connection is $W/(A_cL_m)$. Therefore, the total bandwidth used by this network is:

$$\begin{aligned} W_{mp} &= \text{number of active routes} \times \text{connection bandwidth} = \\ &= NA_cW/A_cL_m = NW/Lm. \end{aligned} \tag{14}$$

The connection throughput is $\lambda_{mp} = W_{mp}/N^2$.

This result shows that the capacity of the network is inversely proportional to the length of a route. This confirms our initial comment that increasing the route length means distributing more traffic across the network, therefore decreasing the average connection throughput. It is therefore useful to compute an upper bound on L_m which allows ensuring that:

$$\lambda_{mp} > \lambda_{sp}. \tag{15}$$

This leads to:

$$L_m < L_{max} = \frac{1}{2\left(\frac{\beta}{2\pi}-\sqrt{\frac{\beta}{2\pi N}}\arctan\left(\sqrt{\frac{\beta N}{2\pi}}\right)\right)}. \tag{16}$$

It is worth noticing that L_{max} is itself bounded as follows:

$$L_{max} > \frac{\pi}{\beta}. \tag{17}$$

Remember that β is a constant characterizing the fact that the routes between source and destination nodes are not perfect straight lines. This parameter, which only depends on the network density and node distribution, can be evaluated by geometric analysis. When the network density is high, β is typically small. Therefore, L_{max} will be a large number. For instance, for a network consisting of 100 nodes in 1 kilometer square, $\beta \approx \pi/16$. We therefore have $L_{max} > 16$. However, on average, simulations show that the average path length in multi-path routing is around 6 or 7 hops. This means that there is in fact no constraint on L_m as far as connection throughput improvement guarantee is concerned. In other words, using multi-path routing always improve the connection throughput of the network as compared to single-path routing. However, when the network density is low, β is bigger, the value L_{max} must be taken into account as an upper bound of the routes when performing the route discovery so that a better performance is guaranteed when using multi-path routing.

5. Conclusion

In this paper, we have analyzed and compared single-path and multi-path routing algorithms. We have first concentrated this study on the issue of overheads. We have shown how the amount of overheads increases with the number of multiple paths and we have seen that when this number exceeds three, the overheads increase significantly. This has confirmed many simulation results presented in the literature which state without any clear explanation that using three paths provides the best trade off. We have also derived an upper bound on the average length of the multi-path routes which guarantees a decrease of the network congestion. This upper bound depends on the traffic intensity, the processing power of each node and the number of nodes in the network, hence it is easy to compute in practice. Not only this bound allows to select routes that respect the upper bound constraint, but also, it can indicate in the first

place whether for a particular network, using load balancing will bring any improvement at all. Finally, we have shown that using multi-path routing always results in connection throughput improvement for high density networks.

Appendix A

Derivation of the traffic experienced by a node located at a distance r from the center of the network

Theorem 1. The traffic for a node located at a distance r from the center of the circle can be expressed as represented by the following expression:

$$\lambda(r) = (\pi R^2 \delta - 1)\lambda + \frac{\pi (R^2 - r^2)^2 \delta^2 \lambda \beta}{2}.$$

Proof. Consider Fig. 5 and let us denote by A, a node located at a distance r from the center of the circle. Let us also define the following notation: $x(i)$ is a point on the edge of the circle such as the angle between $(A, x(i))$ and the axis (O, A) is equal to i. Consider $S_\alpha d(\alpha)$, the portion of the circle (shadowed area on the Fig. 5) centered around $(A, x(\alpha))$ with a aperture of $d\alpha$. Our aim is to determine the amount of traffic originated by source nodes in $S_{alpha} d(\alpha)$ and going through node A.

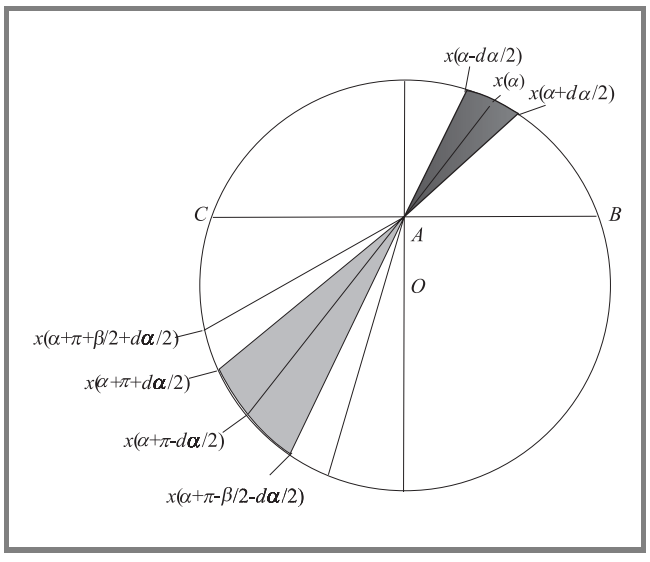


Fig. 5. Traffic analysis for shortest path mechanism.

Recall that we use a shortest path routing mechanism in this section. It is reasonable to assume that in this case, routes are "close" to straight lines. The problem is then to determine the "destination area" D, i.e. the portion of the circle containing all possible destination nodes corresponding with source nodes in $S_\alpha d(\alpha)$ through node A. If the routes were perfect straight lines, then obviously D would be the portion of the circle (dashed area in Fig. 5) centered around $(A, x(\alpha + \pi))$ with aperture $d\alpha$, i.e. $D = S_{\alpha+\pi} d(\alpha)$. However, since the routes are obviously not straight lines, D is in fact larger than this, i.e. $D = S_{\alpha+\pi}(d\alpha + \beta)$ with β being a small positive real number, independent of α and $d\alpha$ and typically small. The value of β depends on the network density and the nodes distribution. This value can be obtained by using graphical analysis of the nodes distribution or by simulation.

Let us now evaluate $S_\alpha d(\alpha)$ and $S_{\alpha+\pi}(d\alpha + \beta)$. Since $d\alpha$ can be reasonably assumed small enough so that $d\alpha^2 << d\alpha$, the following approximations hold:

$$\begin{aligned} sin(d(\alpha)) &= d(\alpha) \\ |A\,x(\alpha - d(\alpha))| &= |A\,x(\alpha)| \\ |A\,x(\alpha + d(\alpha))| &= |A\,x(\alpha)| \\ S_\alpha d(\alpha) &= \frac{|A\,x(\alpha - d(\alpha))| \times |A\,x(\alpha + d(\alpha))| sin(d(\alpha))}{2}, \end{aligned}$$

where the notation $|y\,z|$ stands for the distance between points y and z. From these, we can conclude that

$$S_\alpha d(\alpha) = \frac{|A\,x(\alpha)|^2 d(\alpha)}{2}. \tag{18}$$

Similarly:

$$S_{\alpha+\pi}(d\alpha + \beta) = \frac{|A\,x(\alpha + \pi)|^2 (d\alpha + \beta)}{2}. \tag{19}$$

Assuming a uniform distribution of nodes in the circle, the number of nodes in $S_\alpha d(\alpha)/2$ and $S_{\alpha+\pi}(d\alpha + \beta)$ will respectively be $S_\alpha d(\alpha)\delta$ and $S_{\alpha+\pi}(d\alpha + \beta)\delta$ and therefore, the number of routes going through node A will be:

$$\begin{aligned} N(\alpha) &= S_\alpha d(\alpha)\delta \times S_{\alpha+\pi}(d\alpha + \beta)\delta = \\ &= \frac{|A\,x(\alpha)|^2 * |A\,x(\alpha + \pi)|^2 \delta^2 (d\alpha^2 + d\alpha\beta)}{4}. \end{aligned} \tag{20}$$

Recalling that $d\alpha$ is very small, therefore $d\alpha^2 << d\alpha$, we have $(d\alpha^2 + d\alpha\beta) = \beta d\alpha$. Hence:

$$N(\alpha) = \frac{|A\,x(\alpha)|^2 |A\,x(\alpha + \pi)|^2 \delta^2 \beta d\alpha}{4}. \tag{21}$$

We need to evaluate $|A\,x(\alpha)|\ |A\,x(\alpha + \pi)|$. In order to solve the problem, we have to prove the following result:

For any line $(B_1 C_1)$ going through node A, (B_1 and C_1 located on the circle of radius R), we have: $|A\,C| * |A\,B| = |A\,C_1| * |A\,B_1| = (R^2 - r^2)$. Indeed, from Fig. 6, we can see that $\angle AB_1B = \angle ACC_1$ and $\angle ABB_1 = \angle AC_1C$ ($\angle AB_1B$ standing for the angle between lines (B_1, B) and (B_1, A)). Therefore, the triangle AB_1B is similar to the triangle ACC_1 and

$$\frac{|A\,B_1|}{|A\,C|} = \frac{|A\,B|}{|A\,C_1|}$$

or, $|A\,B_1| \times |A\,C_1| = |A\,B| \times |A\,C| = R^2 - r^2$. Since $Ax(\alpha)$ and $Ax(\alpha + \pi)$ are on the same straight line, we can apply the above result to the case where $x(\alpha) = B_1$ and $x(\alpha + \pi) = C_1$ which leads to

$$|A\,x(\alpha)| \times |A\,x(\alpha + \pi)| = R^2 - r^2.$$

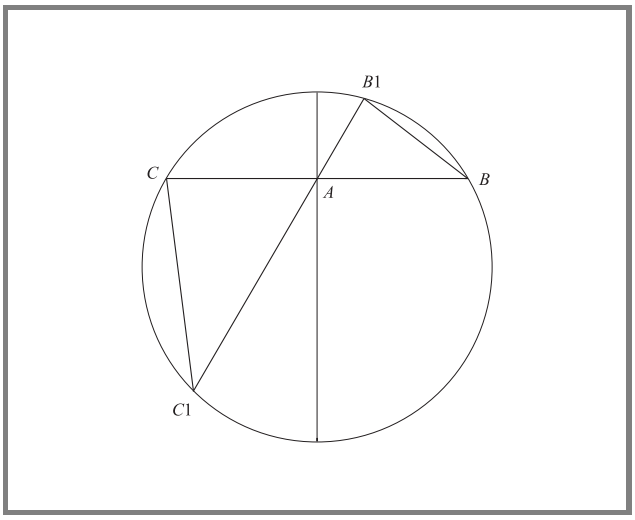


Fig. 6. Analysis of a line going through node *A*.

It is worth pointing out that $|A\,x(\alpha)| \times |A\,x(\alpha+\pi)|$ does not depend on α. We then have

$$N(\alpha) = \frac{\pi(R^2-r^2)^2\delta^2\beta\,d\alpha}{4}. \tag{22}$$

The total amount of traffic relayed by node A is obtained by integrating $N(\alpha)$ over α between 0 and π and multiplying it by the traffic rate λ:

$$\text{relay-traffic} = \frac{\pi(R^2-r^2)^2\delta^2\beta\lambda}{2}. \tag{23}$$

The traffic experienced by each node comprises relay traffic which has just been computed and traffic from others nodes. Since the circle is of radius R, the area πR^2. Therefore, the number of nodes in the circle are: $\pi R^2 * \delta$. Hence, there are $(\pi R^2\delta - 1)$ nodes communicating with the current node with traffic rate λ. The total traffic going through a node located at a distance r from the center is:

$$\begin{aligned}\text{traffic} &= \text{common-traffic} + \text{relay-traffic} = \\ &= (\pi R^2\delta - 1)\lambda + \frac{\pi(R^2-r^2)^2\delta^2\lambda\beta}{2}. \end{aligned}\tag{24}$$

Appendix B

Derivation of N_{pac_s}

From the derivation is Section 3.2, we have:

$$N_{pac}(r) = \frac{\lambda(r)}{\eta - \lambda(r)} = \frac{(\pi R^2\delta - 1)\lambda + \frac{\pi(R^2-r^2)^2\delta^2\lambda\beta}{2}}{\eta - (\pi R^2\delta - 1)\lambda - \frac{\pi(R^2-r^2)^2\delta^2\lambda\beta}{2}}$$

and

$$N_{pac_s} = \frac{1}{\pi R^2\delta} * \int_0^R 2\pi r\delta N_{pac}(r)dr.$$

Therefore,

$$\begin{aligned} N_{pac_s} &= \frac{1}{\pi R^2\delta}\int_0^R 2\pi r\delta N_{pac}(r)dr = \\ &= \frac{1}{\pi R^2\delta}\int_0^R 2\pi r\delta\frac{(\pi R^2\delta-1)\lambda + \frac{\pi(R^2-r^2)^2\delta^2\lambda\beta}{2}}{\eta - (\pi R^2\delta-1)\lambda + \frac{\pi(R^2-r^2)^2\delta^2\lambda\beta}{2}}dr = \\ &= \frac{1}{\pi R^2\delta}\int_0^R \frac{2Kr[A+B(R^2-r^2)^2]}{\mu - A - B(R^2-r^2)^2}dr, \end{aligned}\tag{25}$$

where: $K = \pi\delta$, $A = (\pi R^2\delta - 1)\lambda$, and $B = \pi\delta^2 sin(\beta)\lambda/2$. Hence

$$\begin{aligned} N_{pac_s} &= \frac{1}{\pi R^2\delta} * \frac{K\mu/B}{2\sqrt{\frac{\mu-A}{B}}}\ln\left(\frac{\sqrt{\frac{\mu-A}{B}}+R^2}{\sqrt{\frac{\mu-A}{B}}-R^2}\right) - \frac{1}{\pi R^2\delta}KR^2 = \\ &= \frac{\mu/B}{2R^2\sqrt{\frac{\mu-A}{B}}}\ln\left(\frac{\sqrt{\frac{\mu-A}{B}}+R^2}{\sqrt{\frac{\mu-A}{B}}-R^2}\right) - 1. \end{aligned}\tag{26}$$

References

[1] N. F. Maxemchuck, "Diversity routing", in *IEEE ICC'75*, San Francisco, USA, 1975, vol. 1, pp. 10–41.

[2] R. Krishan and J. A. Silvester, "Choice of allocation granilarity in multi-path source routing schemes", in *IEEE INFOCOMM'93*, 1993, vol. 1, pp. 322–329.

[3] R. Rom, I. Cidon, and Y. Shavitt, "Analysis of multi-path routing", *IEEE/ACM Trans. Netw.*, vol. 7, no. 6, pp. 885–896, 1999.

[4] R. C. Ogier and V. Ruthenburg, "Minimum-expected-delay alternate routing", in *INFOCOMM'92*, Florence, Italy, 1992, pp. 617–625.

[5] S. V. R. Nageswara and S. G. Batsell, "QoS routing via multiple paths using bandwidth reservation", in *INFOCOM (1)*, 1998, pp. 11–18.

[6] S. J. Lee and M. Gerla, "AODV-BR: backup routing in ad hoc network", in *IEEE WCNC 2000*, 2000, pp. 1311–1316.

[7] L. Wang *et al.*, "Multipath source routing in wireless ad hoc network", in *Can. Conf. Elec. Comp. Eng.*, 2000, vol. 1, pp. 479–483.

[8] S. J. Lee and M. Gerla, "Split multi-path routing with maximally disjoint paths in ad hoc networks", in *ICC'01*, 2001.

[9] M. R. Pearlman *et al.*, "On the impact of alternate path routing for load balancing in mobile ad hoc network", in *MobiHOC*, 2000, p. 150.

[10] A. Tsirigos and Z. J. Haas, "Multi-path routing in the presence of frequent topological changes", *IEEE Commun. Mag.*, Nov. 2001.

[11] A. Nasipuri and S. R. Das, "On-demand multi-path routing for mobile ad hoc networks", in *IEEE ICCCN'99*, 1999, pp. 64–70.

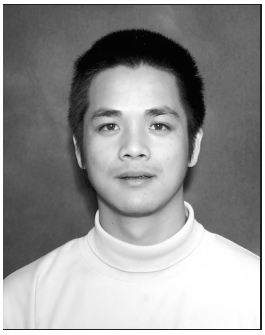
[12] D. B. Johnson and D. A. Maltz, "Dynamic source routing in ad hoc wireless networks", in *Mobile Computing*, T. Imielinski and H. Korth, Eds. Kluwer, 1996, vol. 353.

[13] C. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing", in *IEEE Works. Mob. Comput. Syst. Appl. WMCSA*, 1999, pp. 90–100.

[14] P. Pham and S. Perreau, "Multi-path routing protocol with load balancing policy in mobile ad hoc network", in *IEEE MWCN'2002*.

[15] V. D. Park and M. S. Corson, "Temporally-ordered routing algorithm (tora) version 1: functional specification", Internet-Draft, Nov. 1997, draft-ietf-manet-tora-spec-00.txt.

[16] D. Bertsekas and R. Gallager, *Data Networks*. Prentice-Hall, 1992.



Peter P. Pham received the B.E. in computer system engineering (honour) from the University of Adelaide, Adelaide, Australia, in December 2000. After graduation, he worked as a software engineer for Motorola for 6 months in Singapore. Since August 2001, he received a President scholarhip and started as a Ph.D. candidate at Institute for Telecommunications Research, the University of South Australia. His area of interests are performance analysis and coding techniques for ad hoc networks.
e-mail: ppham@spri.levels.unisa.edu.au
Institute for Telecommunications Research
University of South Australia
Mawson Lakes, SA 5095, Australia

Sylvie Perreau received the engineering diploma and master degree (Diplome d'Etudes Approfondies) from the Ecole Nationale Superieure de l'Electronique et de ses Applications (ENSEA), France, in 1993 and the Ph.D. degree from the Ecole Nationale Superieure des Telecommunications de Paris, in 1997. She was research associate at the University of Connecticut (USA) in 1997 and later that year joined the Centre for Sensor Signals and Information Processing (CSSIP) in Adelaide, Australia, as a post-doctoral fellow. Since May 1998, she has been with the Institute for Telecommunications Research where she is now senior research fellow. Her research interests are in physical layers areas (equalisation, power control, turbo-equalisation) and in networks (routing protocols for ad hoc networks, congestion management).
e-mail: sylvie@spri.levels.unisa.edu.au
Institute for Telecommunications Research
University of South Australia
Mawson Lakes, SA 5095, Australia

Paper

Linear quadratic power control for CDMA systems

Michael D. Anderson, Sylvie Perreau, and Langford B. White

Abstract — In this paper, we present a robust decentralized method for jointly performing channel estimation and closed loop power control for the reverse link of CDMA networks. Our method, based on linear quadratic Gaussian (LQG) control systems theory and Kalman filtering, does not require any training symbols for channel or signal to interference ratio (SIR) estimation. The main interest of this new scheme is that it improves the performance of current SIR based power control techniques while avoiding the problem of power escalation, which is often observed in current systems.

Keywords — CDMA systems, power control, Kalman filtering, channel estimation, linear quadratic control systems.

1. Introduction

Up-link power control (PC) is a crucial element in multi user CDMA systems. In order to maximize capacity and quality of service (QoS), all mobile station (MS) transmissions should be received at the base station (BS) with equal power [1, 2]. Since CDMA systems are interference limited, much work concentrates on the signal to interference ratio as a measure to control MS powers [3–6].

In theory, assuming perfect knowledge of the SIR, SIR based PC outperforms signal strength based PC. This is because the power of the received signal is in fact the sum of the power of the desired and the interference signals. Therefore, in a situation where the received signal power is strong because the interference is significant, a signal strength based power control algorithm would wrongly instruct the MS to decrease its transmit power. Therefore, SIR seems to be a more natural parameter to control interference limited systems. However, SIR based PC is associated with 2 major drawbacks. Firstly, SIR is difficult to estimate accurately [7]. UMTS [8] specifies simultaneous transmission of the DPCCH control channel with DPDCH data channel. SIR is estimated using pilot bits transmitted on DPCCH. System capacity is obviously reduced accordingly. Centralized schemes, where the BS has information for all MS, can determine SIR more accurately than distributed systems [5]. However, centralized methods are difficult to implement due to their high computational complexity.

The other major downfall of SIR based PC is the problem of power escalation. SIR, as the name suggests, is a ratio between signal and interference. As one MS increases its power to compensate for interference from other MS's, its signal interferes more on all other MS's which will in turn increase their transmit power. Instability and power escalation (also defined as positive feedback) can result while the SIR for each MS remains the same. This is particularly prevalent when the system is operating at or near the capacity limit. Therefore, an SIR based PC scheme should be used in conjunction with a perfect call admission control mechanism, which is very difficult to guarantee in real systems.

Zhang et al. [9] address the power escalation problem with a joint signal strength and SIR based PC scheme that compares both quantities to thresholds and adjusts power in the MS with a simple adaptive step size algorithm. The positive feedback potential is of course eliminated with the use of the signal strength constraint. This approach may stop the escalation problem but it still does not minimize the MS transmit power. Ratanamahatana et al. [10] propose a simple method for extracting from the received signal, the desired and interfering signal strengths using pilot symbols. This method, however, reduces system capacity and has potential problems when PC bits are in error on the forward channel.

Qain and Gajic [11] take the power escalation problem further by applying stochastic control systems theory to SIR estimation and PC. SIR error and MS transmit power are jointly minimized and an $H\infty$ filter/estimator is used to track the channel variations. They extend their work by applying constrained optimization techniques to include maximum and minimum allowed MS transmit power. This work is followed up in [12] where the SIR error variance and sum of the variance of MS transmission power are jointly minimized. Linear quadratic control theory is applied. However, this method is not suitable for tracking channel variations due to the mobile speed, which is a crucial issue to be addressed for fading channels.

In this paper, we present a robust decentralized method for jointly performing channel estimation and close loop power control for the reverse link of CDMA networks. We base our approach on optimal control systems theory. The novelty of our approach is that while it aims at maintaining the SIR of each MS close to the SIR targets, its implementation does not rely on the actual calculation of the SIR. In other words, our approach takes *implicitly* into account the interference component of each signal but is not affected by a positive feedback. Another important feature of our proposed method is that it does not rely on any pilot or training sequence thus increasing the system capacity. The general structure of this algorithm is similar to the one already implemented in the IS95 system. It is therefore very easy to implement in practical systems. Finally, being an adaptive method, it allows taking into account fading characteristics of wireless channels. This adaptivity to the

channel conditions is further improved by a multi step size approach, which allows compensating for deep fades. In order to achieve this, a quantized 3-bit PC command on the feedback (forward) channel is proposed. The paper is organized as follows: we first present the problem formulation and the general linear quadratic gain controller. Then, the channel estimation is addressed using a Kalman filter. Finally, simulations comparing the conventional IS95 power control device with our proposed scheme are presented.

2. Problem formulation

In this paper, we use the following notations and assumptions:

- $p_k(n)$ is the transmit power of mobile user k during the frame n.
- $\Gamma_k(n)$ is the squared absolute value of the average (over frame n) of the up-link channel gain (for user k).
- The spreading codes are long random sequences such that the cross correlations between users averaged over a frame are approximately $1/N$, where N is the spreading gain.
- σ^2 is the variance of the thermal noise process (modeled as a zero mean Gaussian random variable).
- $P_k(n) = p_k(n)\Gamma_k(n)$ is the power of the signal, received at the base station, after despreading for user k.

3. Perfect power control

Using the above notations, we can write the signal to interference (plus noise) ratio for user k over frame n:

$$SIR_k(n) = \frac{P_k(n)}{\frac{1}{N}\sum_{j\neq k} P_j(n) + \sigma^2}. \tag{1}$$

Assuming that perfect power control is feasible, we want to find $P_k(n)$ for every user k in the system which satisfies

$$SIR_k(n) = \beta, \tag{2}$$

where β is the SIR target, assumed to be identical for all users in the system (this condition can be relaxed).
This will be achieved when all user signals will be received with the same power denoted by P^* expressed as:

$$P^* = \frac{\beta\sigma^2}{1 - \frac{\beta(K-1)}{N}}, \tag{3}$$

where K is the total number of users in the system. In this paper, we perform power control so that the received powers for all users approach this optimal value P^*. Note that in order to obtain P^*, it is necessary that $1 - \frac{\beta(K-1)}{N}$ be positive which leads to the classical condition on the capacity:

$$K_{\max} < 1 + \frac{N}{\beta}. \tag{4}$$

As mentioned in the introduction, it is well known that this capacity limit must be ensured for the power control scheme implemented in IS95 to be stable. Our proposed method can be more flexible: when the call admission control momentarily allows more users than $K_{\max}$ in the system, our algorithm assumes that the capacity has reached its limit ($K = K_{\max}$) and this does not affect the stability of the system.

4. Robust power control formulation on the logarithmic scale

Let us denote by $w_k(n)$ the average over frame n of the power of the CDMA signal despread by the spreading sequence of user k. Then $w_k(n)$ is written as:

$$w_k(n) = P_k(n) + \frac{1}{N}\sum_{j\neq k} P_j(n) + \sigma^2 + v_k(n). \tag{5}$$

Here, $v_k(n)$ is the measurement noise due to the limited number of samples involved in the average operation.
In the decentralized case, we do not have access to the received powers $P_j(n)$ for $j \neq k$. However, it is reasonable to assume that for each j we have:

$$P_j(n) = P^* + e_j(n). \tag{6}$$

In other words, even though each user's signal is power controlled so that Eq. (3) is respected, the power control is not perfect and each user's signal is received with a power which differs from P^* by a value $e_j(n)$.
Therefore Eq. (5) can be rewritten as:

$$w_k(n) = P_k(n) + \frac{K}{N}P^* + \sigma^2 + v'_k(n), \tag{7}$$

where $v'_k(n) = v_k(n) + \frac{1}{N}\sum_{j\neq k} e_j(n)$. Note that $\frac{K}{N}P^* + \sigma^2 = \frac{1}{\beta}P^*$. Also, since $P_k(n) = P^* + e_k(n)$, we can write:

$$\begin{aligned} w_k(n) &= P_k(n) + \frac{1}{\beta}P^* + v'_k(n) = \\ &= \left(\frac{1}{\beta} + 1\right)P_k(n) - \frac{1}{\beta}e_k(n) + v'_k(n) = \\ &= \left(\frac{1}{\beta} + 1\right)P_k(n) + v''_k(n), \end{aligned} \tag{8}$$

where $v''_k(n) = v'_k(n) - \frac{1}{\beta}e_k(n)$.

Let us recall that the usual IS95 power control algorithm is performed on a logarithmic scale. It would be thus useful to write this observation equation on a logarithmic scale. Let us denote by $X_{dB_k}(n)$ the value of $X_k(n)$ on the logarithmic scale, i.e. $X_{dB_k}(n) = 10\log_{10}(X_k(n))$.
We then have:

$$\begin{aligned} w_{dB_k}(n) &= 10\log_{10}\left(\frac{1}{\beta}+1\right) + \\ &\quad +10\log_{10}\left(P_k(n)(1+\frac{v_k''(n)}{P_k(n)})\right) = \\ &= 10\log_{10}\left(\frac{1}{\beta}+1\right)+10\log_{10}(P_k(n)\Big) + \\ &\quad +10\log_{10}\left(1+\frac{v_k''(n)}{P_k(n)}\right) = \\ &= P_{dB_k}(n)+\lambda^2+\xi_k(n)\,, \end{aligned} \tag{9}$$

where $\lambda^2 = 10\log_{10}\left(\frac{1}{\beta}+1\right)$ and the measurement noise $\xi_k(n) = 10\log_{10}\left(1+\frac{v_k''(n)}{P_k(n)}\right)$.
We thus have the following observation equation on the logarithmic scale:

$$w_{dB_k}(n) = P_{dB_k}(n)+\lambda^2+\xi_k(n)\,. \tag{10}$$

Recall that the received power $P_{dB_k}(n)$ is in fact written as:

$$P_{dB_k}(n) = \Gamma_{dB_k}(n)+p_{dB_k}(n)\,, \tag{11}$$

where $\Gamma_{dB_k}(n)$ is the channel gain and $p_{dB_k}(n)$ the transmit power at frame n.
The aim of this paper is to properly adjust $p_{dB_k}(n)$ so that the SIR of each user is close to its target value (i.e. so that the receive power of each user is close to P^*). We propose to do so by deriving the infinite horizon linear quadratic Gaussian controller as detailed in the next section. More precisely, our aim is to design $u_k(n)$ such that:

$$p_{dB_k}(n+1) = p_{dB_k}(n)+u_k(n)\,. \tag{12}$$

5. The linear quadratic controller

From now on, for simplification, we will omit the subscripts k and dB, bearing in mind that all quantities are expressed in dBs and that each user performs the same operations in a decentralized way.
In this section, we assume that the channel gain $\Gamma(n)$ is exactly known to the base station. In the next section, we will explain how to estimate these quantities using the Kalman filter.
The aim of this section is to design the control command $u(n)$ in an LQG framework, i.e. which minimizes the following linear quadratic cost function:

$$J = E\Big\{\lim_{N\to\infty}\sum_{n=0}^{N} q_c\|P(n)-P^*\|^2+r_c\|u(n)\|^2\Big\}\,, \tag{13}$$

where q_c and r_c are quantities to be determined. This cost function is a weighted combination of the squared error between the received power and the optimum received power P^*, and the power of the control command $u(n)$. Indeed, while the ultimate aim is to meet the SIR requirements, it is also important to keep the control command as small as possible for implementation purposes. It is obvious that in some circumstances, imposing too much constraint on the control command will make it impossible to achieve the main objective (i.e. minimize $\|P(n)-P^*\|$). There is hence a trade off in the choice of the cost minimization weighting factors r_c and q_c. A discussion on the respective importance of these two parameters is provided in the next section.
It is well know that a static feedback law determined by the solution to an algebraic Ricatti equation (ARE) gives the solution to the minimization of the cost function J.
In other words, the optimal control $u(n)$ is given by:

$$u(n) = K^pP(n)+K^rP^*, \tag{14}$$

where

$$K^p = -(P^{(1)}+r_c)^{-1}P^{(1)} \tag{15}$$

$$K^r = -(P^{(1)}+r_c)^{-1}P^{(2)} \tag{16}$$

with $P^{(1)}$ given by the solution to the ARE:

$$P^{(1)} = P^{(1)}-P^{(1)}(P^{(1)}+r_c)^{-1}P^{(1)}+q_c \tag{17}$$

and $P^{(2)}$ obtained by:

$$\begin{aligned} P^{(2)} &= \frac{q_c}{K^p} = \\ &= -\frac{q_c(P^{(1)}+r_c)}{P^{(1)}}\,. \end{aligned} \tag{18}$$

One can easily see that

$$K^r = \frac{q_c}{P^{(1)}} = -K^p \tag{19}$$

and that

$$P^{(1)} = \frac{1+\sqrt{1+4r_c}}{2}\,. \tag{20}$$

Finally, the command $u(n)$ is computed as

$$\begin{aligned} u(n) &= K^p(P(n)-P^*) = \\ &= -K^r(p(n)+\Gamma(n)-P^*)\,. \end{aligned} \tag{21}$$

As mentioned previously, the control command $u(n)$ depends on the channel gain $\Gamma(n)$ which is usually not known at the base station. In the next section, we show how to estimate this quantity using the Kalman filter.

6. Kalman filtering estimation of the channel gain

Recall that the observation process on the logarithmic scale is written as:

$$w(n) = p(n) + \Gamma(n) + \lambda^2 + \xi(n), \tag{22}$$

where the transmit power $p(n)$ given by

$$p(n) = p(n-1) - K^r(p(n-1) + \Gamma(n-1) - P^*) \tag{23}$$

is known to the receiver.
In this section, we assume that due to the Doppler effects, the channel coefficients are correlated in time. We therefore model the (unknown) channel gain on the logarithmic scale as an auto regressive (AR) process as:

$$\Gamma(n) = \left[\Gamma(n-1)\,\Gamma(n-2)\,\cdots\,\Gamma(n-L)\right]\mathbf{h} + b(n), \tag{24}$$

where $\mathbf{h} = [h_1\ h_2\ \cdots\ h_L]^T$ is supposed to be known at the base station. A method for estimating these coefficients in conjunction with the method presented in this paper can be found in [13].
Using this time dependency, we can easily write a state equation for vector $\underline{\Gamma}(n) = [\Gamma(n)\,\Gamma(n-1)\,\cdots\Gamma(n-L+1)]$ as follows:

$$\underline{\Gamma}(n) = A\underline{\Gamma}(n-1) + [b(n)\,0\,\cdots\,0]^T, \tag{25}$$

where

$$A = \begin{bmatrix} h_1 & h_2 & \cdots & h_L \\ 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{26}$$

The observation Eq. (22) can be written as a function of the model state $\underline{\Gamma}(n)$ as:

$$w(n) = C\underline{\Gamma}(n) + p(n) + \lambda^2 + \xi(n) \tag{27}$$

with $C = [1\,0\,\cdots 0]$.
Using Eqs. (25) and (27), we can easily derive the Kalman filter estimate of state $\underline{\Gamma}(n)$, i.e. $\hat{\underline{\Gamma}}(n|n)$:

$$\hat{\underline{\Gamma}}(n|n) = \mathcal{F}A\hat{\underline{\Gamma}}(n-1|n-1) + M(w(n) - p(n) - \lambda^2), \tag{28}$$

where $\mathcal{F} = I - MC$. The filter gain M is given by:

$$M = \Sigma C^T (C\Sigma C^T + R_0)^{-1} \tag{29}$$

and Σ is obtained by solving the Ricatti equation:

$$\Sigma = A(\Sigma - \Sigma C^T (C\Sigma C^T + R_0)^{-1} C\Sigma)A^T + Q_0, \tag{30}$$

where $R_0 = cov(b(n))$ and $Q_0 = cov(\xi(n))$.

7. The proposed PC algorithm

Let us recall the major steps involved in our proposed method. At frame n:

1. Evaluate $w(n)$, the logarithm of the average over the frame of the users' signal power.
2. Compute the Kalman filtering estimation of channel gain state $\Gamma(n)$ using Eq. (28).
3. Using the Kalman estimate of the channel gain for frame n, compute $u(n+1)$, the control command for next frame using the LQG controller (Eq. (21)).
4. Quantize and encode the control command and transmit the control bits on the feedback channel.
5. At the mobile station, upon reception of the control command bits, compute the transmit power $p(n+1)$ using Eq. (12).

8. Numerical issues

We have seen that the scheme derived in the previous sections depends on the 4 following parameters: r_c and q_c for the LQG controller and R_0 and Q_0 for the Kalman filter. In this section, we provide a preliminary discussion on the choice of these parameters. A more exhaustive study is left for future work.

Choice of q_c. In the particular framework we deal with in this paper, there is no other constraint on q_c than $q_c > 0$. In fact, what really matters is the ratio between q_c and r_c. Therefore, we will assume that $q_c = 1$ and emphasize on the choice of r_c.

Choice of r_c. It is well known that by taking r_c close to zero, the LQG performs a *loop transfer recovery*. In this case, it can be shown that the LQG behaves like a Kalman filter predictor. The advantage of loop transfer recovery is that we do know its good robustness property to modelization errors. While this is desirable, the drawback is that there is no constraint on the amplitude of the control command $u(n)$ (the controller only minimizes the variance of the receive power). In the case of power control, it is important to keep $u(n)$ as small as possible. Indeed, errors can occur during the transmission of the control bits on the feedback channel (these bits are not protected by any error control scheme). Therefore, by keeping $u(n)$ as small as possible, this also ensures that if an error occurs on the control command bits, this error will also be small. Another determining factor in the choice of r_c is the velocity of the mobile: indeed, for high speeds, deep fades are likely to occur. Therefore, $u(n)$ needs to be big enough to properly track those fades. To summarize, r_c must be chosen as small as possible consistent with limiting the control action. An analytical study is needed in order to properly determine this parameter.

Choice of Q_0. By definition, $Q_0 = var(\xi(n))$ where $\xi(n)$ is the measurement noise. Clearly, $\xi(n)$ depends on the number of users in the system.

Choice of R_0. R_0 is by definition the variance of the input noise to the state model. This parameter in fact depends on the velocity of the mobile: for large speeds, R_0 is large and vice versa. The choice of R_0 is important for the tracking performance of the Kalman filter: if chosen too small, the Kalman estimate is not able to track the variations of the channel. On the other hand, if chosen to large, the Kalman estimate will be too noisy. However, with R_0 depending only on the mobile velocity, it is possible to derive it (this is not addressed in the paper and is left for future work). An interesting point is that this parameter will help determining analytically r_c which also depends on the mobile speed.

9. Simulation results

We developed a simulation environment corresponding to the proposed UMTS guidelines. For simplicity we assigned a single up-link channel per user at a continuous data rate of 60 kbit/s (uncoded). All users' signals passed through a fast fading Rayleigh channel. The Rayleigh fading channel was computed using the Jakes method [14]. Table 1 shows the simulation parameters. We used 3-bit quantized PC commands and optimized step sizes for both our proposed method and the IS95 power control device.

Table 1
Simulation parameters

Parameter	Value
PC command rate	1 500 Hz
Frame length	10 ms
Slots per frame	15
Channel bandwidth	5 MHz
Chip rate	3.84 Mc/s
Processing gain	64
Data rate (uncoded)	60 kbit/s
Filter length	10

In Fig. 1, we show how the transmit power, derived with our proposed scheme, follows the variations of the channel.

Figure 2 highlights the main advantage of our proposed scheme as compared to the IS95 power control device: our proposed approach is not subject to positive feedback, which occurs when all users in the system unnecessarily increase their transmit power. Given the spreading gain of 64 and the SIR target of 7 dB, one can easily see that the maximum number of users allowed in the system is 14. In our simulation, the number of users was set to $K = 13$. Therefore, we operated just under the capacity limit. One can see that the IS95 power control device is subject to instability from frame number 50. One can see that even though the capacity limit is not exceeded, the transmit power is unnecessarily high, compared to our scheme.

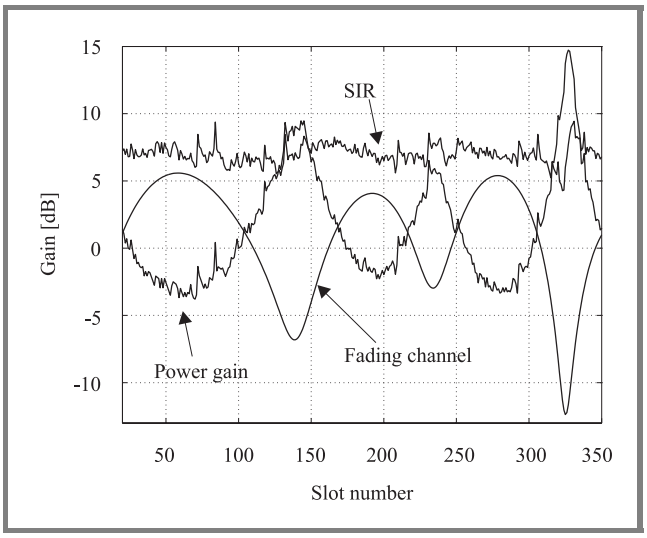


Fig. 1. Transmit power and channel gain variations.

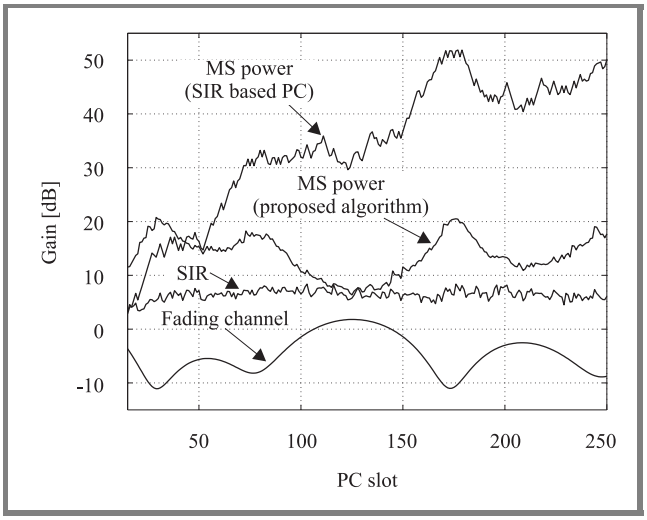


Fig. 2. Example of power escalation – power escalation in SIR based power control.

In Fig. 3, we show that the performance of our scheme in terms of bit error rate (BER) is slightly better than the IS95 performance. In this simulation, the number of users was set to 10 (which eliminates the power escalation issue of the IS95 scheme) and we assumed that the SIR has been estimated using the technique in [7]. It is important to recall the fact that SIR based power control techniques outperform signal strength based schemes. Our scheme, even though based on signal strength, implicitly takes into account the interference component. This explains why, even though based on signal strength, it does not perform badly compared to SIR based methods. The fact that it actually performs better than IS95 is due to the fact that the SIR estimate used in the IS95 scheme does

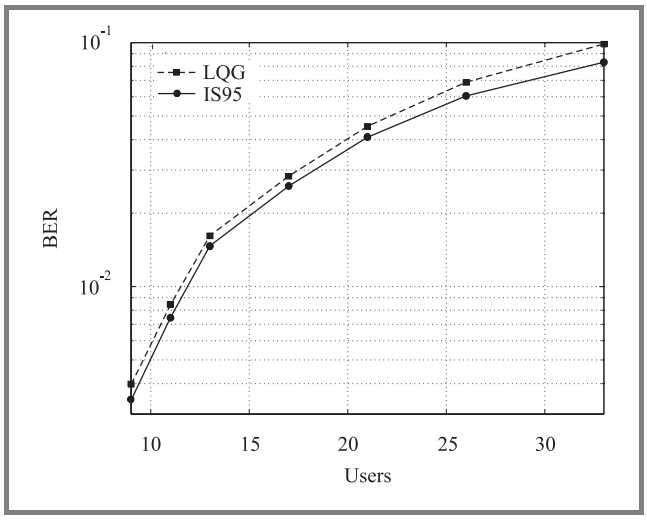


Fig. 3. Bit error rate versus users.

not match exactly the actual SIR. Also, since the Kalman filter estimate takes explicitly into account the channel gain dynamics, the control commands showed better tracking performance, especially in deep fade situations.

10. Conclusion

In this paper, we have proposed a new scheme for power control in a CDMA system, based on a linear quadratic Gaussian controller and using Kalman filtering for channel estimation purposes. The main feature of this method is that it ensures that the SIR requirements for each user are met without the usual drawback of SIR based power control techniques, i.e. positive feedback. Our method does not require any SIR estimation, which is usually difficult to accurately perform without training or pilot symbols. Essentially based on signal strength, it implicitly takes into account the interference component of the signal, without any knowledge of the other users' signal strength and without any training sequence for estimation purposes. Simulations show that this method provides a better performance than SIR based techniques, without the risk of power escalation, which can occur in IS95 based systems when approaching the capacity limit. We have shown in this paper that our method relies on a set of parameters that are crucial for the robustness and the tracking abilities of the Kalman filter estimate and the controller. We have qualitatively discussed the choice of these key parameters. However, further analytical work is needed in order to clarify how they should be expressed as a function of the mobile velocity.

Acknowledgement

This research is carried out with the financial support from the Commonwealth of Australia through the Cooperative Research Centers program.

References

[1] J. G. Proakis, *Digital Communications*. McGraw-Hill, 1995, 3 ed.

[2] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communications*. Addison-Wesley, 1995.

[3] S. Nourizadeh, P. Taaghol, and R. Tafazolli, "A novel closed loop power control for UMTS", in *First Int. Conf. 3G Mob. Commun. Technol.*, 2000, pp. 56–59.

[4] F. Ye, Q. T. Zhang, and C. C. Ko, "A simple adaptive up-link power control algorithm for DS/CDMA system", in *IEEE/AFCEA EUROCOMM'2000, Inform. Syst. Enhan. Publ. Saf. Secur.*, 2000, pp. 16–19.

[5] F. C. M. Lau and W. M. Tam, "Novel predictive power control in a CDMA mobile radio system", in *IEEE 51st Veh. Technol. Conf. Proc.*, 2000, vol. 3, pp. 1950–1954.

[6] S. Gunaratne, S. Nourizadeh, T. Jeans, and R. Tafazoli, "Performance of SIR-based power control for UMTS", in *Second Int. Conf. 3G Mob. Commun. Technol.*, 2001, pp. 16–20.

[7] A. Kurniawan, "SIR-estimation in CDMA systems using auxiliary spreading sequence", *Mag. Electr. Eng.*, vol. 5, no. 2, pp. 9–18, 1999.

[8] H. Holma and A. Toskala, Eds., *WCDMA for UTMS*. Wiley, 2001.

[9] D. Zhang, Q. T. Zhang, and C. C. Ko, "A novel joint strength and SIR based CDMA reverse link power control with variable target SIR", in *IEEE Int. Conf. Commun.*, 2000, vol. 3, pp. 1502–1505.

[10] S. Ratanamahatana and H. M. Kwon, "Channel estimation for power controlled 3G CDMA", in *IEEE 51st Veh. Technol. Conf. Proc.*, 2000, vol. 3, pp. 2429–2433.

[11] L. Qian and Z. Gajic, "Joint optimization of mobile transmission power and SIR error in CDMA systems", in *Proc. 2001 Am. Contr. Conf.*, 2001, vol. 5, pp. 3767–3772.

[12] L. Qian and Z. Gajic, "Variance minimization stochastic power control in CDMA systems", in *IEEE Int. Conf. Commun.*, 2002, vol. 3, pp. 1763–1767.

[13] S. Perreau, M. Anderson, and L. White, "Adaptive power control for CDMA systems using linear quadratic Gaussian control", in *Thirty Sixth Asil. Conf. Sig., Syst. & Comput.*, Nov. 2002.

[14] W. C. Jakes, Ed., *Microwave Mobile Communications*. IEEE Press, 1994.

Michael David Anderson was born in Adelaide, Australia, in 1970. He obtained an Associate Diploma in Electronic Engineering in 1994. He worked for the Bureau of Meteorology, Adelaide for 5 years as a technical officer (electronics) before obtaining his engineering degree (electronics) in 2000 (first class honours) from the University of South Australia. In 2000 he joined the Institute for Telecommunication Research at the University of South Australia as a Ph.D. candidate. He research interests cover the area of mobile communications, particularly at the physical layer.

e-mail: mikea@spri.levels.unisa.edu.au
Institute for Telecommunications Research
University of South Australia
Mawson Lakes, SA 5095, Australia

Langford B. White graduated from the University of Queensland, Brisbane, Australia with the degrees of B.Sc. (maths), B.E. (hons) and Ph.D. (electrical eng.) in 1984, 85 and 89, respectively. From 1986–1999, he worked for the Defence Science and Technology Organisation, Salisbury, South Australia. Since 1999, he has been Professor in the School of Electrical and Electronic Engineering, The University of Adelaide where he is also Director of the Centre for Internet Technology Research. Prof. White's research interests include signal processing, control, telecommunications and Internet engineering.

e-mail: Lang.White@adelaide.edu.au
Department of Electrical and Electronic Engineering
Adelaide University
Adelaide, SA 5000, Australia

Sylvie Perreau – for biography, see this issue, p. 47.

Paper

Queuing models for cellular networks with generalised Erlang service distributions

Aruna Jayasuriya

Abstract — Providing seamless handover is one of the major problems in mobile communication environments. Careful dimensioning of the network and the underlying teletraffic analysis plays a major role in determining the various grade of services (GoSs) that can be provided at various network loads for handover users. It has been shown that the channel holding time of a cell, one of the important parameters in any teletraffic analysis, can be accurately modelled by Erlang distributions. This paper focuses on solving queuing systems with generalised Erlang service distributions and exponential arrival distributions. We present the quasi-birth-death (QBD) process, which characterises the queuing models with generalised Erlang service and exponential interarrival distributions. We then use the properties of Erlang distributions and characteristics of channel allocation process of cellular networks to simplify the queues used to model cellular networks. The use of these simplifications provide a significant reduction in computation time required to solve these QBDs.

Keywords — cellular networks, phase-type distributions, generalised Erlang.

1. Introduction

One of the major problems that needs to be addressed in mobile communication networks is the continuity of a service during a handover without any data loss, as the user moves from cell to cell. This is called seamless handover [1]. The blocking probability encountered at handover is an important grade of service parameter for mobile users. It is of utmost importance to carefully dimension the network to provide the guaranteed GoS levels.

The channel holding time of a mobile user is an important parameter in the analysis of communication networks. It was shown in [2] that channel holding time in cellular networks can be accurately modelled as a generalised Erlang phase-type distribution. However the resulting queuing models are not tractable using common matrix manipulation techniques. In this paper we propose to use a simplification technique based on the properties of Erlang distributions and characteristics of channel allocation procedures in cellular networks to create a tractable queuing model for cellular networks.

Rest of the paper is organised as follows. Section 2 briefly describes the proposed QBD processes resulting from the queuing models associated with the cellular network. This work has been presented in [2] and has been included here for completeness. The proposed simplification techniques are presented in the following section. Matrix equations that describe the stationary probabilities of the system are presented in Section 3. Sections 4 and 6 describe the techniques that were used to solve for the blocking probabilities of the system. We conclude the paper staying that the proposed simplification leads to the creation of standard matrix equation from a seemingly un-tractable queuing system. Due to space limitations we only intend to describe the simplification techniques in this paper. Readers who are interested in final results from the study should refer to [2, 3].

2. Queuing model for cellular network channels

A phase-type (PH) distribution of generalised Erlang form with 2 phases has been proposed to model channel holding times in cellular networks [2]. Phase-type distributions can be used to approximate virtually any renewal process, with the dimensionality of the phase-type distribution increasing with the complexity of the particular process being modelled [2, 3]. Furthermore they provide an accurate description of the channel holding time distribution in cellular networks, while retaining the underlying Markovian properties of the distribution. These Markovian properties are essential in generating tractable queuing models for cellular networks. The parameters of this distribution can be estimated from experimental data by using the expectation maximisation (EM) algorithm [2].

References [2] and [3] describe methods used to derive the channel holding time distribution for cellular networks through network simulation models. Use of EM algorithm to approximate the actual distribution with a generalised Erlang distribution with 2 phases is presented in [3] and [4]. Arrivals of new and handover users are modelled with exponential distributions with appropriate parameters.

Assuming an exponential interarrival distribution and a phase-type service distribution, a cellular network with n channels per cell can be modelled as an $M/PH/n/n$ queue [5]. The resulting queuing system may not be tractable even for moderate values of n. Equation (1) shows the rate transition matrix or **Q** matrix for an $M/PH/n/n$

queue, which models a cell with n channels. In Eq. (1) it can be observed that $\mathbf{Q}$ is of block tri-diagonal form [5]:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_1 & 0 & 0 & \cdots & & \\ \mathbf{B}_1 & \mathbf{A}_{11} & \mathbf{A}_{01} & 0 & \cdots & & \\ 0 & \mathbf{A}_{22} & \mathbf{A}_{12} & \mathbf{A}_{02} & 0 & & \\ \vdots & \ddots & \ddots & \ddots & & & \\ \vdots & \vdots & \mathbf{A}_{2i} & \mathbf{A}_{1i} & \mathbf{A}_{0i} & & \\ & & & \ddots & \ddots & \ddots & \\ & & & 0 & \mathbf{A}_{2,n-1} & \mathbf{A}_{1,n-1} & \mathbf{A}_{0,n-1} \\ & & & 0 & 0 & \mathbf{A}_{2,n} & \mathbf{A}_{1,n} \end{bmatrix} \tag{1}$$

Assuming an exponential arrival process with the rate λ the matrices $\mathbf{B}_0$, $\mathbf{B}_1$ and $\mathbf{B}_2$ are defined as [2, 5]:

$$\mathbf{B}_0 = \lambda \tau, \tag{2}$$

$$\mathbf{B}_1 = \lambda, \tag{3}$$

$$\mathbf{B}_2 = \mathbf{t}. \tag{4}$$

Where τ is the initial probability vector of the PH service distribution and $\mathbf{t}$ is a column vector partition of the rate transition matrix $\mathbf{Q}_{PH}$ of the PH distribution, given in Eq. (5):

$$\mathbf{Q}_{PH} = \begin{bmatrix} 0 & 0 \\ \mathbf{t} & \mathbf{T} \end{bmatrix}. \tag{5}$$

The construction of matrices $\mathbf{A}_{0i}$, $\mathbf{A}_{1i}$ and $\mathbf{A}_{2i}$ for a general phase type distribution is rather complex and requires lengthy and tedious computations. However, we have identified some properties of generalised Erlang distributions which simplify these calculations to a great extent. The next section of this paper describes the proposed simplifications and algorithms leading to the construction of tractable $\mathbf{Q}$ matrix for the queuing process. We assume the service distribution given by the following equation throughout the rest of the paper:

$$\mathbf{T} = \begin{bmatrix} -\mu_1 & \mu_1 \\ 0 & -\mu_2 \end{bmatrix}. \tag{6}$$

Row sums for any rate transition matrix are zero [5], resulting in,

$$\mathbf{t} = \begin{bmatrix} 0 \\ \mu_2 \end{bmatrix}, \tag{7}$$

where $\mathbf{T}$ and $\mathbf{t}$ are the partition matrices of $\mathbf{Q}_{PH}$ as describes in Eq. (5).

3. $M/PH/n/n$ server with a generalised Erlang service distribution

In the previous section we selected the $M/PH/n/n$ queue to model a cell with n channels. A drawback in using the above model to represent mobile network cells is that the size of the rate transition matrix depends on the number of channels available in a cell. As the size of the block matrix at level i is on order $\binom{i+2}{i}$ for a service distribution of 2 phases [6], the $\mathbf{Q}$ matrix becomes unmanageable for networks with large number of channels available per cell. Future mobile networks intend to provide a large number of channels per cell to support the high data rate services that will be available. Therefore the methods available in [5] cannot be used to find the blocking probabilities experienced by new and handover users in the mobile network environment.

A simplification can be made to the $\mathbf{Q}$ matrix by observing some properties of the Erlang distribution and the behaviour of servers at mobile cells. In servers with Erlang distributions, the users always start the service in the first phase and move on to the next phase with probability 1 once the sojourn in that phase is over. Users who finish the sojourn in the last phase depart the system. When there are m users in the system it is irrelevant which of these m users are at which server and similarly who finishes service first. This leads us to combine all the users in the same service phase to a single server with the service rate equivalent to the combined rate of all the servers. These simplifications allow to represent the system with a reduced state space. The new state space can be defined as follows:

$$\{\text{number of users is phase 1 } (n_1), \\ \text{number of users in phase 2 } (n_2)\}.$$

Using this simplification and a service distribution of two phases, the number of different states possible in the system when there are m users in the system are given in Table 1.

Table 1
Allowed states when there are m users in the system

State	Number of users in phase 1	Number of users in phase 2
1	m	0
2	$m-1$	1
$\vdots$	$\vdots$	$\vdots$
M	1	$m-1$
$m+1$	0	m

This reduces the size of the matrix at level m to $m+1$. The whole system can be arranged into two-dimensional

continuous time Markov chain with the following state space:

0	no users in the system
(1,0)	1 user in the system
(0,1)	1 user in the system
⋮	⋮
(n_1,n_2)	n_1+n_2 users in the system with n_1 users in phase 1 and n_2 users in phase 2
⋮	⋮

The events, which change the state of the system and rates of leaving the current state at those events are shown in Table 2. Figure 1 shows the transitions listed in Table 2. The first few states illustrating the construction of the rate transition matrix, **Q**, are given in Fig. 2. The rate transition matrix for the $M/PH/n/n$ QBD process can be constructed by observing the transitions between states given in Figs. 1 and 2, and using Table 2 to get the transition rates between different states. In order to obtain the characteristic tri-diagonal form of the **Q** matrix, it is necessary to perform a linear ordering of the states. In this case it is the simple ordering $\{0,(1,0),(0,1),(2,0),(1,1),(0,2),\ldots,(n,0),(n-1,1),\ldots,(1,n-1),(0,n)\}$. We can define levels where level m is the combination of all states when the number of users in the system are m. This ordering allows to generate the **Q** matrix of the form given in Eq. (1).

Table 2
Transition rates between different states

From	To	Rate	Event	Range
0	1,0	λ	Arrival	
(n_1,n_2)	(n_1+1,n_2)	λ	Arrival	$n_1 \geq 1$
(n_1,n_2)	(n_1,n_2-1)	$n_2\mu_2$	Departure	$n_2 \geq 1$
(n_1,n_2)	(n_1-1,n_2+1)	$n_1\mu_1$	Phase change	$n_1,n_2 \geq 1$

Elements of **Q** correspond to the transition rates for all the allowed transitions in the queuing system. A rate of zero means that the particular transition is not allowed in the system. **Q** can be divided into several row levels. Row level i corresponds to the matrices which describe the system when there are i users in the system. Matrices $\mathbf{A}_{0i}$, $\mathbf{A}_{1i}$ and $\mathbf{A}_{2i}$ make up the ith row level of **Q** given in Eq. (1). Similarly the columns can also be divided into different levels resulting column levels $(i-1)$, i and $(i+1)$ for matrices $\mathbf{A}_{0i}$, $\mathbf{A}_{1i}$ and $\mathbf{A}_{2i}$ respectively. Therefore matrix $\mathbf{A}_{2i}$ corresponds to the transitions which result in the number of users in the system being decreased from i to $(i-1)$, namely departures. Similarly the matrix $\mathbf{A}_{1i}$ represents the transitions which do not change the total number of users in the system, which corresponds to phase changes and self transitions in the system. Finally the matrix $\mathbf{A}_{0i}$ corresponds to arrivals which increase the number of users from i to $(i+1)$.

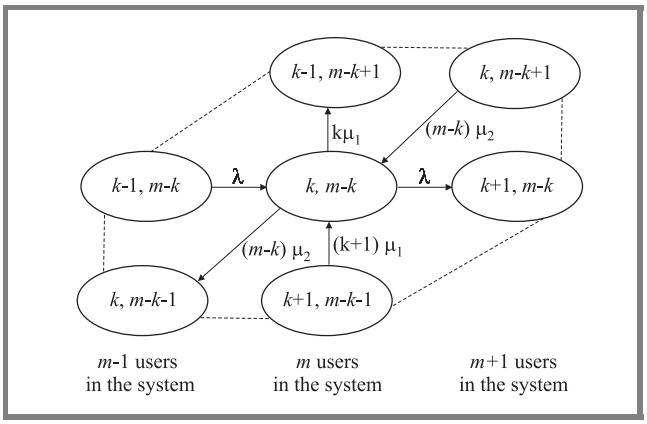


Fig. 1. State transitions for $M/PH/n/n$ QBD process.

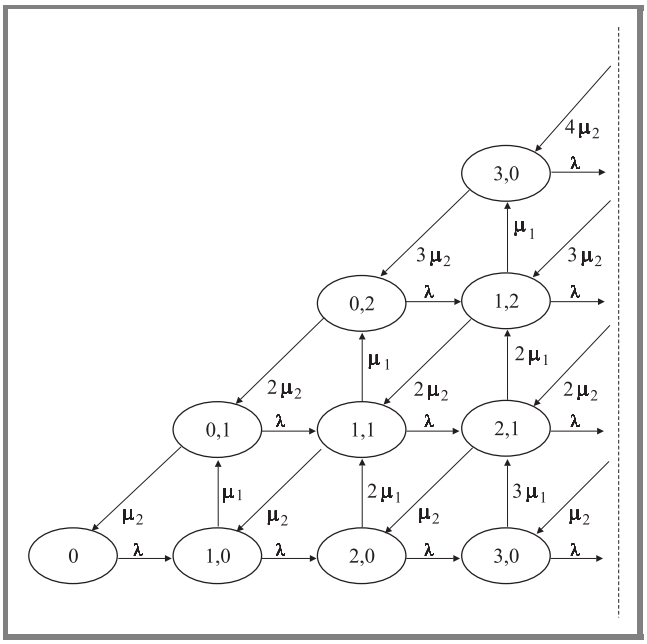


Fig. 2. First few states of the $M/PH/n/n$ QBD process.

3.1. Creating matrix $\mathbf{A}_{2i}$

It was explained earlier that the **Q** matrix represents all the allowed transitions in the system. Furthermore, as matrix $\mathbf{A}_{2i}$ is a sub matrix of **Q**, matrix $\mathbf{A}_{2i}$ corresponds to the all the departures from the system with i users. Assume that the number of users in phases 1 and 2 are given by n_1 and n_2 respectively. Then the allowed transitions in matrix $\mathbf{A}_{2i}$ are from states of the form (n_1,n_2) to states of the form (n_1,n_2-1). In other words, the transitions that leave the number of users in phase 1 unchanged while decrease the number of users in phase 2 by 1. The transition rate for these transitions are the combined rate of all n_2 users in phase 2, which results in $n_2\mu_2$ as the appropriate transition rate. To construct matrix $\mathbf{A}_{2i}$ it is necessary to identify all

the allowed departures from the system with i users and then calculate the effective rate for these departures. The following algorithm is used to achieve this by identifying all the departures from the system with i users. Firstly we define matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ as follows:

$$\mathbf{S}_1 = \begin{bmatrix} i & 0 \\ i-1 & 1 \\ i-2 & 2 \\ \vdots & \vdots \\ 1 & i-1 \\ 0 & i \end{bmatrix} \quad \mathbf{S}_2 = \begin{bmatrix} i-1 & 0 \\ i-2 & 1 \\ \vdots & \vdots \\ 1 & i-2 \\ 0 & i-1 \end{bmatrix}$$

Matrix $\mathbf{S}_i$ contains the linearly ordered set of all possible states with i users in the system, while $\mathbf{S}_2$ contains the state space for system with $i-1$ users. In matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ the first and second columns correspond to the number of users in phase 1 and phase 2 respectively. The two state spaces $\mathbf{S}_1$ and $\mathbf{S}_2$ correspond to the state spaces along rows and columns of $\mathbf{A}_{2i}$ respectively. For example, the transition resulted by a move from first row of $\mathbf{S}_1$, $\{i,0\}$, to first row of $\mathbf{S}_2$, $\{i-1,0\}$, corresponds to the element $\{1,1\}$ of matrix $\mathbf{A}_{2i}$. We use the variable *row* to represent rows of matrix $\mathbf{A}_{2i}$ and variable *column* to represent columns of $\mathbf{A}_{2i}$. The algorithm then loops through the elements of $\mathbf{S}_1$ and $\mathbf{S}_2$ (effectively compares the transitions formed by moving from a state in $\mathbf{S}_1$ to a state in $\mathbf{S}_2$) and identify which transitions are allowed in the system. For allowed transitions the transition rate is created and it is entered into the position $\{row, column\}$ of $\mathbf{A}_{2i}$. All other transitions are given a value of zero to represent that they are not allowed in this system. The algorithm described is given below:

For $row = 1$ to $i+1$
  For $column = 1$ to i
    If $\mathbf{S}_2\ (column, 1) == \mathbf{S}_1(row, 1)$
    AND $\mathbf{S}_2(column, 2) == \mathbf{S}_1(row, 2) - 1$
      $\mathbf{A}_{2i}(row, column) = \mathbf{S}_1(column, 2)\mu_2$
    Else
      $\mathbf{A}_{2i}(row, column) = 0.0$
    End If
  End *column* loop
End *row* loop

Using the above algorithm $\mathbf{A}_{2i}$ (of size $(i+1) \times i$) is calculated and presented in Eq. (8):

$$\mathbf{A}_{2i} = \begin{bmatrix} 0 & \cdots & \cdots & \vdots \\ \mu_2 & & & \vdots \\ \vdots & 2\mu_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \vdots & \cdots & \cdots & i\mu_2 \end{bmatrix} \tag{8}$$

3.2. Creating matrix $\mathbf{A}_{0i}$

The concept used to derive the algorithm described in the previous section can also be used to derive an algorithm to calculate $\mathbf{A}_{0i}$. Matrix $\mathbf{A}_{0i}$ represents the arrivals into a system with i users. Using the previous notations the allowed transitions in this matrix are from states of the form (n_1, n_2) to states of the form (n_1+1, n_2). Alternatively, the transitions that increase the number of users in phase 1 by 1, while leaving the number of users in phase 2 unchanged. As the arrival rate into the system is independent of number of users in the system, all transitions have rate λ.
We define matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ as follows:

$$\mathbf{S}_1 = \begin{bmatrix} i & 0 \\ i-1 & 1 \\ \vdots & \vdots \\ 1 & i-1 \\ 0 & i \end{bmatrix} \quad \mathbf{S}_2 = \begin{bmatrix} i+1 & 0 \\ i & 1 \\ i-1 & 2 \\ \vdots & \vdots \\ 1 & i \\ 0 & i+1 \end{bmatrix}$$

The notation used here is the same as explained earlier in Section 3.1. Matrix $\mathbf{S}_1$ contains the linearly ordered set of all possible states with i users in the system while $\mathbf{S}_2$ contains the state space for system with $i+1$ users. Then the following algorithm can be used to identify the allowed transitions in $\mathbf{A}_{0i}$. As explained in Section 3.1 this algorithm loops through all the possible transitions in the system by traversing through the state spaces of $\mathbf{S}_1$ and $\mathbf{S}_2$ and finds the transitions allowed for this particular queuing model:

For $row = 1$ to $i+1$
  For $column = 1$ to $i+2$
    If $\mathbf{S}_2\ (column, 1) == \mathbf{S}_1(row, 1) + 1$
    AND $\mathbf{S}_2(column, 2) == \mathbf{S}_1(row, 2)$
      $\mathbf{A}_{0i}(row, column) = \lambda$
    Else
      $\mathbf{A}_{0i}(row, column) = 0.0$
    End If
  End *column* loop
End *row* loop

Using this algorithm matrix $\mathbf{A}_{0i}$ (of size $(i+1) \times (i+2)$) can be given as follows:

$$\mathbf{A}_{0i} = \begin{bmatrix} \lambda & 0 & \cdots & \vdots \\ 0 & \ddots & & \vdots \\ \vdots & 0 & \lambda & 0 \end{bmatrix} \tag{9}$$

3.3. Creating matrix $\mathbf{A}_{1i}$

Matrix $\mathbf{A}_{1i}$ represents all the transitions which do not change the number of users, i, in the system, namely the phase changes and the self-transitions. Using the previous notations, the allowed transitions in this matrix are from

states (n_1, n_2) to (n_1-1, n_2+1) (phase changes) and from (n_1, n_2) to (n_1, n_2) (self-transitions). The rate for the phase transitions are given by the combined rate of $n_1$users in the first phase before the transitions. The self-transition rates can be found easily by observing that all the row sums are zero for any $\mathbf{Q}$ matrix [5].

The following algorithm can be used to create all the elements except the diagonal elements of $\mathbf{A}_{1i}$. The diagonal elements are initially given zero in this algorithm. Then the correct diagonal elements become the negative of row sums of the $\mathbf{Q}$ matrix as the row sums are zero for any rate transition matrix. We define matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ as follows to obtain all the elements except the diagonal elements of $\mathbf{A}_{1i}$:

$$\mathbf{S}_1 = \mathbf{S}_2 = \begin{bmatrix} i & 0 \\ i-1 & 1 \\ \vdots & \vdots \\ 1 & i-1 \\ 0 & i \end{bmatrix}$$

Matrix $\mathbf{S}_1$ and $\mathbf{S}_2$ contain the linearly ordered set of all possible states with i users in the system. Then we can use the following procedure to derive $\mathbf{A}_{1i}$:

For $row = 1$ to $i+1$
  For $column = 1$ to $i+1$
    If $\mathbf{S}_2\ (column, 1) == \mathbf{S}_1(row, 1) - 1$
    AND $\mathbf{S}_2(column, 2) == \mathbf{S}_1(row, 2) + 1$
      $\mathbf{A}_{1i}(row, column) = \mathbf{S}_1(row, 1)\mu_1$
    Else
      $\mathbf{A}_{1i}(row, column) = 0.0$
    End If
  End *column* loop
End *row* loop

The diagonal elements are given by:

$$-\left(\sum_{row}\mathbf{A}_{0i} + \sum_{row}\mathbf{A}_{1i} + \sum_{row}\mathbf{A}_{2i}\right),$$

where $\sum_{row}$ represents the row sum of the particular matrix. $\mathbf{A}_{1i}$ (of size $(i+1)\times(i+1)$) is given in Eq. (10):

$$\mathbf{A}_{2i} = \begin{bmatrix} -\zeta_0 & i\mu_1 & 0 & \cdots & & \cdots & \vdots \\ 0 & -\zeta_1 & (i-1)\mu_1 & & & & \\ \vdots & 0 & \ddots & \ddots & & & \\ & & & -\zeta_k & (i-k)\mu_1 & & \vdots \\ & & & & \ddots & \ddots & 0 \\ & & & & & -\zeta_{i-1} & \mu_1 \\ \vdots & \cdots & & & & & \zeta_i \end{bmatrix} \tag{10}$$

where for $k = 0, 1, \ldots, i$

$$\zeta_k = \lambda + (i-k)\,\mu_1 + k\mu_2\,. \tag{11}$$

However, due to the extra boundary condition present when the number of users in the system is equal to the number of servers, ζ_k's for $k = 0, 1, \ldots, n$ in the matrix $\mathbf{A}_{1n}$ have to be modified as follows:

$$\zeta_k = (i-k)\,\mu_1 + k\mu_2\,. \tag{12}$$

4. Stationary probabilities of the system

Sections 3.1–3.3 explained how to create all the sub matrices required to generate the rate transition matrix representing the $M/PH/n/n$ QBD process with generalised Erlang service distribution. The objective of this study is to analyse the performance of a cellular network, focusing on the blocking probabilities for handover and new users as the main performance analysis parameter. To calculate the blocking probabilities, it is necessary to find the stationary probabilities of the system. In other words the probability of having i users in the system, with i ranging from 0 to n, need to be found. The phase-type service distribution introduces $i+1$ *substates* within each of these i states. In this section we will present methods to calculate the stationary probabilities corresponding to all the sub states and then show how we can combine these sub state stationary probabilities to calculate the probability of finding i users in the system at any instance.

The relationship between the stationary or equilibrium probabilities and the rate transition matrix for a time-homogeneous continuous time Markov chain (CTMC) is given by Eq. (13). It has been shown in [5] that Eq. (13) is valid for a wide variety of systems involving phase-type distributions. The stationary distribution vector, $\mathbf{x} = [x_1, x_2, \ldots, x_{S_{n+1}}]$ of the $M/PH/n/n$ queuing system satisfies the following equations:

$$\mathbf{xQ} = 0, \qquad x_i \geq 0, \quad \sum_{i=1}^{S_{n+1}} x_i = 1\,, \tag{13}$$

$$\mathbf{xP} = \mathbf{x}, \qquad x_i \geq 0, \quad \sum_{i=1}^{S_{n+1}} x_i = 1\,. \tag{14}$$

Where $\mathbf{P} = \mathbf{Q} + \mathbf{I}$ is a stochastic matrix and $\mathbf{I}$ is an identity matrix. The blocking probability of the system (i.e., the probability that a new user joining the system finds all the channels occupied) is given by the following equation[1]:

$$P_{block} = \sum_{j=S_n+1}^{S_{n+1}} x_j\,. \tag{15}$$

[1]State space of a queue with a single-phase service distribution can be expressed as the number of users at service. With a 2 phase generalised Erlang service distribution the state space of the resulting QBD process is 2 dimensional and can be expressed as {number of users in service phase 1, number of users in service phase 2}. Therefore the probability of having j users in the system is calculated by adding all the sub states such that,

number of users in phase 1 + number of users in phase 2 = j.

For small values of n, Eq. (14) can be solved by finding an eigenvector of $\mathbf{P}^T$ associated with unit eigenvalue and then normalising it such that the sum of the entries of the normalised eigenvector equals one. This method becomes computationally very expensive even for moderate values of n.

5. Stochastic complementation

Stochastic complementation provides a computationally cheaper mechanism to solve Eq. (14) by decoupling an irreducible large Markov chain into smaller irreducible Markov chains [7, 8]. Reference [8] states that an irreducible large Markov chain $\mathbf{P}$ with m states can be uncoupled into k smaller chains, say $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k$, containing $r_1, r_2, \ldots, r_k$ states, where $\sum_{i=1}^{k} r_i = m$. It further states that it is possible to determine the solutions for these smaller chains, $\mathbf{s}_i s$, completely independent of each other. Therefore instead of solving the larger matrix a number of smaller matrices can be solved. Solutions of these smaller chains can then be combined appropriately to generate the solution for the larger system.

A divide-and-conquer approach can be used to systematically simplify the system until the individual matrices become small enough to solve directly. The $S_{n+1} \times S_{n+1}$ matrix $\mathbf{P}$ can be partitioned roughly in half as given in Eq. (16):

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{bmatrix} \tag{16}$$

From this partition two *stochastic complements* of $\mathbf{P}$, $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ can be derived as follows:

$$\mathbf{S}_{11} = \mathbf{P}_{11} + \mathbf{P}_{12}(\mathbf{I} - \mathbf{P}_{22})^{-1}\mathbf{P}_{21}, \tag{17}$$

$$\mathbf{S}_{21} = \mathbf{P}_{22} + \mathbf{P}_{21}(\mathbf{I} - \mathbf{P}_{11})^{-1}\mathbf{P}_{21}. \tag{18}$$

These are each irreducible stochastic matrices of order approximately $S_{n+1}/2$, and the combination of $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ is equivalent to the irreducible Markov chain $\mathbf{P}$ [8]. The most time consuming operation in calculating $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ is the inversion of matrices of size $S_{n+1}/2$, which is far simpler than solving for eigen vectors of a square matrix of size S_{n+1}.

If $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ are small enough to be solved directly (through eigen vectors) then the solutions of $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$, $\mathbf{s}_1, \mathbf{s}_2$, can be combined to generate the solution for $\mathbf{P}$, in terms of the *coupling factors*, $\varepsilon_1 \varepsilon_2$ between the two matrices [8]. Where

$$\varepsilon_1 = \frac{\mathbf{s}_2\mathbf{P}_{21}\mathbf{e}}{\mathbf{s}_1\mathbf{P}_{12}\mathbf{e} + \mathbf{s}_2\mathbf{P}_{21}\mathbf{e}}, \tag{19}$$

$$\varepsilon_2 = \frac{\mathbf{s}_1\mathbf{P}_{12}\mathbf{e}}{\mathbf{s}_1\mathbf{P}_{12}\mathbf{e} + \mathbf{s}_2\mathbf{P}_{21}\mathbf{e}}. \tag{20}$$

Then the solution for $\mathbf{P}$, x, is given by

$$\mathbf{x} = \begin{bmatrix} \varepsilon_1\mathbf{s}_1 & \varepsilon_2\mathbf{s}_2 \end{bmatrix}. \tag{21}$$

Where $\mathbf{s}_1$ and $\mathbf{s}_2$ are the stationary distribution vectors for $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ respectively, and $\mathbf{e}$ is a column vector of ones. If the matrices $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ are too large to solve for $\mathbf{s}_1$ and $\mathbf{s}_2$ directly, they are roughly partitioned in half again to get four stochastic complements, $(\mathbf{S}_{11})_{11}, (\mathbf{S}_{11})_{22}, (\mathbf{S}_{22})_{11}(\mathbf{S}_{22})_{22}$. Size of these stochastic complements is approximately $s_{n+1}/4$. This process can be continued until the resulting stochastic complements are small enough to be solved using a direct method.

Once the stochastic complements are solved using a direct method coupling factors for that level can be found using Eqs. (19) and (20). Then the coupling factors and solution for that level are combined to get the solutions for the matrices at the level above using Eq. (21). This combination process at each level continues until the solution for $\mathbf{P}$ is obtained.

Once the stationary probabilities of the system, $\mathbf{x}$, have been obtained such that $\mathbf{x}.\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of 1's, we can find the probability of having i users in the system $p(i)$ as follows:

$$p(0) = x(1)$$

$$p(i) = \sum_{j=S_i+1}^{S_i+1+i} x(i) \quad \text{for} \quad i \geq 1. \tag{22}$$

If the system does not distinguish between new and handover users no priority will be given to one class over the other. In such a system the blocking probability for any user, P_{block}, will be given by

$$P_{\text{block}} = p(n) \tag{23}$$

and the average load of the system, L, at this blocking probability is given by:

$$L = \sum_{i=1}^{n} ip(i). \tag{24}$$

6. Conclusions

Accurate methods have been derived to model cellular networks in recent times [9, 10]. However due to their complexity, these model do not result in tractable queuing systems. In [2] we proposed to model the channel holding time in cellular networks with a 2 phase generalised Erlang distribution. We showed that this distribution accurately approximates the distribution of channel holding time in a cellular network. In this paper we used some properties of the generalised Erlang distribution to derive a simplified queuing system to model the collective channels of a cell in cellular networks. These simplifications enabled us to derive a tractable queuing model from a seemingly

untractable QBD. We then presented the methods used to calculate the blocking probabilities for handover and new users. Due to space limitations of the paper, we only presented the simplification techniques we used in this study that can be applied to similar problems. Interested readers should refer to [2] and [3] for final results.

References

[1] E. D. Re, R. Fantacci, and G. Giamabene, "Handover and dynamic channel allocation techniques in mobile cellular networks", *IEEE Trans. Veh. Techn.*, vol. 44, no. 2, pp. 397–405, 1995.

[2] A. Jayasuriya, J. Asenstorfer, and D. Green, "Modelling service time distributions in cellular networks using phase-type service distributions", in *Proc. Int. Conf. Commun.*, Helsinki, Finland, June 2001, vol. 2, pp. 440–444.

[3] A. Jayasuriya, "Improved handover performance through mobility predictions". Ph.D. thesis, University of South Australia, 2001; URL – http://www.itr.unisa.edu.au/∼aruna/papers/thesis.pdf

[4] S. Asmussen, O. Nerman, and M. Olsson, "Fitting phase-type distributions via the EM algorithm", *Scand. J. Stat.*, vol. 23, pp. 419–441, 1996.

[5] M. Neuts, *Matrix-Geometric Solutions in Stochastic Models*. John Hopkins, 1981.

[6] V. Naoumov, U. Krieger, and D. Wagner, "Analysis of a multiserver delay-loss system with a general Markovian arrival process", in *Proc. 1st Int. Conf. Matrix-Analyt. Meth. (MAM) Stoch. Mod.*, Flint, USA, Aug. 1995, pp. 44–66.

[7] C. Meyer, "Stochastic complementation, uncoupling Markov chains and the theory of nearly reducible systems", *SIAM Rev.*, vol. 31, no. 2, pp. 240–272, 1989.

[8] A. Jayasuriya and K. Dogancay, "Stochastic complementation applied to the analysis of blocking probabilities in cellular networks", in *Proc. IASTED Int. Conf., Wirel. & Opt. Commun.*, Banff, Alberta, Canada, July 2002, pp. 607–610.

[9] P. Orlik and S. Rappaport, "A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions", *IEEE J. Selec. Areas Commun.*, vol. 16, pp. 788–803, 1998.

[10] Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modelling of PCS networks", *IEEE Trans. Commun.*, vol. 47, pp. 1062–1072, 1999.

Aruna Jayasuriya received his bachelor of engineering from University of Adelaide in 1997 and Ph.D. from University of South Australia in 2001. Currently he is research fellow with Institute for Telecommunication Research of University of South Australia. His research interests include traffic modelling in cellular and IP networks, mobility prediction mechanisms for cellular and wireless LAN networks, policy- based quality of service control and routing in ad hoc networks. He has published numerous papers in the above areas.

e-mail: Aruna.Jayasuriya@unisa.edu.au
Institute for Telecommunications Research
University of South Australia
Mawson Lakes, SA 5095, Australia

Paper

Adaptive handover control in IP-based mobility networks

Taeyeon Park and Arek Dadej

Abstract — In this paper, we propose framework for an adaptive handover control architecture (AHCA), which aims at enhancing overall IP handover performance while maximising utilisation of resources in wireless access networks. The IP handover procedures in the AHCA adapt dynamically to network conditions, as well as to a wide range of user profiles and application quality of service (QoS) requirements. To confirm our expectations that the AHCA will bring performance benefits in heterogeneous mobile IP networking environment, we have investigated basic performance characteristics of different handover mechanisms. The preliminary simulation results demonstrate that the AHCA will bring significant performance improvements as compared with non adaptive IP handovers.

Keywords — mobile networking, mobile IP, handover performance, adaptive handover control.

1. Introduction

Mobile IP (Internet Protocol) [11, 12] provides network layer transparent mobility support to mobile nodes (MNs) roaming across different IP subnetworks. Among many deployment issues of mobile IP, the support for micro (local) mobility and seamless handover have been in focus of many research activities over a number of recent years. While many different proposals such as in [8] have been published thus far to address these issues, it is generally accepted that one solution can not fit all situations and requirements, especially in environments where various mobility mechanisms and quality of service models are mixed together [1] in heterogeneous wireless access networks [9, 13].
There are several reasons why a smart, adaptive handover control is needed:

- With adaptive handover control, various handover strategies can be mixed to take advantage of what each technique/strategy can offer, depending on the availability of the technique in a given access network and the network, user and application preferences.

- Adaptive handover control will improve resulting handover performance as the handover procedures selected will best reflect the dynamically varying network operating conditions.

- A number of mobility mechanisms have been proposed to achieve effective global and local mobility management. As a consequence, there is a strong need to harmonise the use of these different mechanisms and promote interoperability across the entire network.

- Normally, some coupling between layer-3 and layer-2 is required (layer-2 support) to achieve best handover performance with the different access network technologies.

- Heterogeneous wireless access technologies require specific handover strategies suited for each wireless access network, resulting in a need for common framework to make handover across the different access technologies seamless.

To best adapt to the current operating conditions and the access network environment where a MN has just moved into, it would be preferable if a smart (adaptive) handover control mechanism [3] could provide flexible service depending on dynamically varying requirements of each traffic flow and application session involved [2]. For this purpose, we have designed the adaptive handover control architecture. As a core part of the architecture, the adaptive handover engine takes inputs from several input pre-processing modules, e.g. network resource information from the network resource prober, traffic QoS attributes from the traffic classifier, user preferences information from the user input handler, and policy information from the policy input handler. Then, it selects the best combination of handover mechanisms using a handover adaptation algorithm, so that the chosen handover strategy produces the best performance for the user, while minimising the use of shared network resources. The architecture has been inspired by the related research in the field of mobile IP handoff control, such as programmable handoffs [4], policy-enabled handoffs [14], and many other adaptive or feedback-based control approaches [5–7].
As an example, the AHCA can be applied to the environment where interoperation between terrestrial and satellite wireless mobile networks is required [10]. In a simple scenario of a satellite-to-terrestrial handover case, the optimal handover control would force handover as soon as an available terrestrial mobile network can be found, thus increasing user satisfaction in terms of both performance and cost. The details of our example scenario would change according to varying conditions surrounding the MN, thus would require some form of adaptation which can be accomplished within the AHCA.
The design goals for the AHCA can be summarised as follows:

- Seamless (both low-loss and low-latency) handover, adaptive in respect to specific requirements of traffic type and its explicit or implicit QoS attributes.
- Microflow based handover control, supporting both user and terminal mobility.
- Fairness- or priority-based usage of resources (e.g. bandwidth, buffer memory, power consumption etc.) while providing reasonable level of QoS.
- Graceful degradation of QoS in cases of resource shortages or unavailability of required capability.
- Dynamic adaptation in pace with varying conditions of operating environments and MN itself – automatic or interactive change of operating parameters.
- Backward compatibility with existing standard or de facto standard protocols.
- Extensibility to cover proposed and future handover algorithms and micro-mobility mechanisms.
- Deployability across a wide range of mobility networks including 802.11 WLAN (wireless local area network) and next generation IP-based cellular networks.

The organisation of this paper is as follows: in the next section, we describe the details of the AHCA. In Sections 3 and 4 we explain the simulation setup and present the example network topology used in the simulation study. We then follow with some preliminary simulation results and their analysis. Finally, we give some concluding remarks and comments on future directions in this research.

2. Adaptive handover control architecture

Figure 1 shows the basic concept of adaptive handover control (components and flows). The handover adaptation algorithm produces optimal set of handover strategies according to various inputs. Various inputs – probed network information, traffic type and QoS attributes of a traffic flow, policy control information, and user preference – are fed to the adaptation algorithm to reflect the environment within which the handover is to occur. Besides these regular inputs, there can be two other possible inputs from the feedback loop and re-adaptation loop. Feedback loop provides performance measures to the adaptation algorithm for the purpose of fine-tuning of future handovers. Re-adaptation loop could be used as a calibration path due to short term changes of surrounding network conditions. To speed up the operation, re-utilisation path can be used to save time and resources by utilising a hashed cache table, which is updated during a few previous iterations of the control algorithm. That could save repetitions of control computations

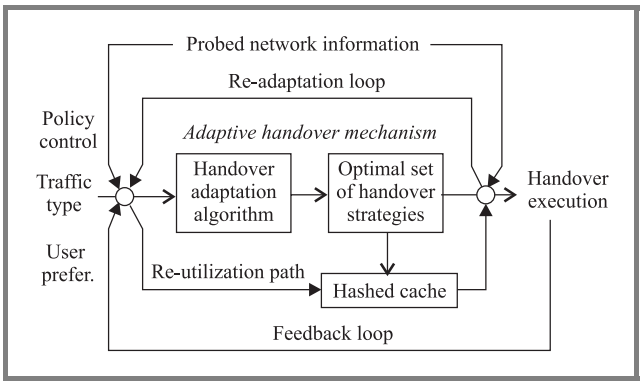


Fig. 1. The concept of adaptive handover control (AHC).

and reduce the time overhead added to the handover by the handover control procedures.

In accordance with the basics described above, we have constructed the AHCA as shown in Fig. 2.

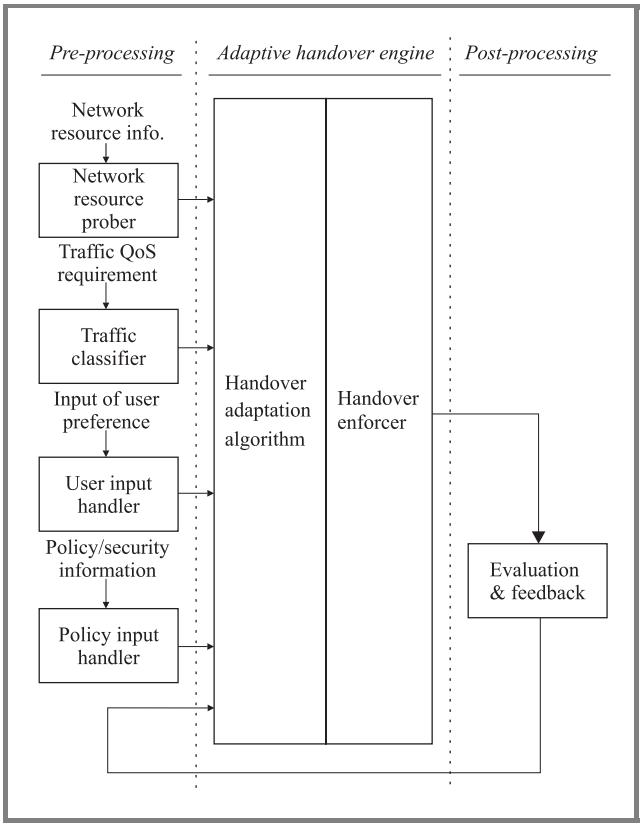


Fig. 2. Adaptive handover control architecture.

The basic operation of the AHCA is as follows. The AHCA:

a) gathers input information;

b) processes inputs to choose the best set of handover mechanisms, and the best parameters for the mechanisms selected;

c) controls the execution of the chosen handover by the MN and mobility agents (MAs);

d) (optionally) feeds back some performance information into the handover adaptation engine.

The component processes of the AHCA reside mainly in MAs and co-operate with components of the AHCA residing within the MN. In most cases, some form of communication needs to occur between MA and MN (or between MAs) to control the handover execution, and to exchange information that will aid handover process. This communication may take the basic form of handshaking messages and is described in detail in Sections 2.1 and 2.4. The AHCA is designed to be an open architecture so that the internal details of its component modules can be substituted as long as the basic interfaces between modules are maintained. In this way, new or more enhanced mechanisms can be used to increase the performance benefits, or mechanisms not available in the given access network environment may be substituted with available ones at the expense of some performance degradation.
Below, we give brief descriptions of the AHCA component modules, outlining the major inputs and outputs and the main functionality of each module.

Network resource prober (NRP): Probe available network resources, using the dynamic network resource probing protocol (DNRPP), in the neighbouring access network and the MN's home network.

Traffic classifier (TC): Get QoS attributes via signaling protocol related to specific microflow and/or sample data traffic to determine the type of traffic and associated QoS attributes.

User input handler (UIH): Process user preferences input interactively or via a built-in static interface.

Policy input handler (PIH): Query network policy/ security/ AAA (authentication, authorisation and accounting) control information and manage local policy information (in the form of configuration table or by dynamic gathering).

Handover adaptation algorithm (HAA): Determine the optimal set of handover strategies in respect to the obtained input criteria, and feed them to the handover enforcer.

Handover enforcer (HEnf): Enforce handover according to the given set of strategies.

Evaluation and feedback processor (EnF): Obtain performance metrics, evaluate against predefined threshold, and feed back to the engine.

In the following subsections, the component modules of the AHCA are explained in more detail.

2.1. Dynamic network resource probing protocol

The dynamic network resource probing protocol can be considered a kind of network resource discovery protocol used by the network resource prober module of the AHCA.
The objective of the DNRPP is to probe network resource information dynamically and in co-operation between MN and MA. Its operation mode can be passive or active. In passive mode, some information is advertised periodically from MA to nearby MNs in an unsolicited manner. In active mode, MN solicits network resource information from nearby MAs. The MAs receiving the request should respond with requested resource information unless security association between MA and MN has not been established or is broken.
The network resource information will be used as an input to the HAA as well as to the re-adaptation loop of the AHCA. It may also be used in prediction and preparation of future handovers. To effectively aid the various uses of network resource information, it is important to select the information items most useful to the AHCA and define efficient format of the information as to not consume too much network bandwidth in the process of probing. A few candidates for the components of the network resource information are delay-distance measure between probe initiating node (e.g. MN, FA – foreign agent) and probe responding node (e.g. FA, HA – home agent, CN – correspondent node), and capabilities supported by the MN or the MA(s).

2.2. Traffic classifier

The traffic classifier consists of four component modules; three of them are input processor modules, and remaining one is QoS level classifier module. One of the input processor modules, the QoS signal handler, examines explicit QoS signaling information from various QoS signaling protocols (e.g. resource reservation protocol Path/Resv) for the specific microflow concerned, and feeds QoS attributes specified for the microflow to the traffic QoS level classifier module. The other two input processors are the header examiner and the payload examiner, which respectively examine some IP header fields and a few starting sequences of payload traffic to get traffic type information and associated QoS attributes, and then feed this information to the traffic QoS level classifier module. Finally, according to traffic type and associated QoS requirements attributes, the Traffic QoS level classifier module produces quantised level of QoS requirements (a value selected from a range of predefined QoS levels) and this output is fed to the adaptive handover control engine.

2.3. Handover adaptation algorithm

In general, the essence of the first stage of fast handover is finding appropriate MA(s) to be in charge of mobility support in the access network area where MN is expected to

move or has just arrived. Then, MN has to decide the most appropriate time to effect the seamless handover. Once handover decision is made, the next step is to choose the best handover strategy i.e. both the handover mechanisms (algorithms) and the related set of parameters. These steps are listed below.

1. Select the best MA (FA) to support the handover.
2. Decide the best time to execute the handover.
3. Choose the best set of handover mechanisms available for this handover.
4. Select the best set of parameters for the chosen handover mechanisms.

The handover adaptation algorithm focuses on the last two steps of the fast handover, choosing best handover mechanism(s) and selecting the best parameter set for the selected handover mechanism. The first two steps, dealing with movement detection and handover decision, are not directly covered by the HAA itself. Figure 3 shows the basic operation of the HAA in respect to the last two steps of the fast handover.

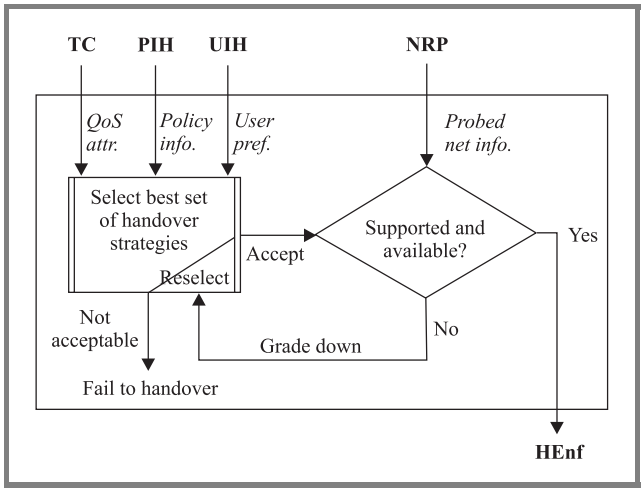


Fig. 3. Operation of the handover adaptation algorithm.

2.4. Handover enforcer

The handover enforcer module provides direct handover control services for the actual handover execution occurring between a MN and one or more of MAs. Depending on the chosen set of handover mechanisms, more than one MA could be involved in the process of exchanging handover control messages.
The messages exchanged between the MAs and the MN can be similar to those used in basic handover control, and thus can be combined with, or substituted for, these basic messages as needed. This may help reduce the overall handover signaling load incurred by the adaptive handover.

2.5. Evaluation and feedback

The evaluation and feedback process is a key component in the *closed-loop* AHCA control system. Without this process, the AHCA becomes merely an *open-loop* control system that has no ability to self-adjust and optimise its own performance. An open-loop AHCA could never directly utilise the measures of its performance, normally collected while the system operates. However, for the purpose of handover control, we can still call the open-loop AHCA adaptive, since it adapts the handover execution according to varying inputs collected from its network environment; in such case the adaptive handover engine would be adjusted manually rather than automatically through the use of feedback component.
In order to achieve effective, fast and dynamic control of the system, while maintaining acceptable stability and overall system efficiency, it is important to make a careful selection of the performance measures that are collected and fed back to the control algorithm. While some conventional performance measures may include packet loss rate, end-to-end transmission delay, delay variance (jitter), throughput, success/failure rates (per call or per handover) and resource usage levels, we can also consider the following second-order performance measures: signaling load, user satisfaction level and (handover and/or network access) cost function.

2.6. Security considerations

In the AHCA, interactions between MNs and MAs are essential part of dynamically probing network resources and of enforcing/coordinating the actual handover. The fundamental importance of these to the network operation means that some kind of security association must be formed between the interacting agents to avoid security attacks. This paper assumes that the security mechanisms specified for the standard mobile IP protocol [11] can be used as part of the AHCA.

3. Simulation setup

3.1. Network topology

Figure 4 shows the network topology we have used as a basis for our simulations under OPNET network simulation environment, to investigate the basic characteristics of mobile IP handover mechanisms. In the figure, the *R_x* denotes border routers in each subnetwork that connect the subnetwork to the Internet. For the home subnetwork, the HA functionality may be incorporated in the border router *R_h*. Similarly, for the foreign subnetwork, the gateway FA functionality that resides in FA1 may be integrated in the border router *R_v*. In hierarchical terms, the FA1 can act as a gateway FA. Otherwise, it acts as

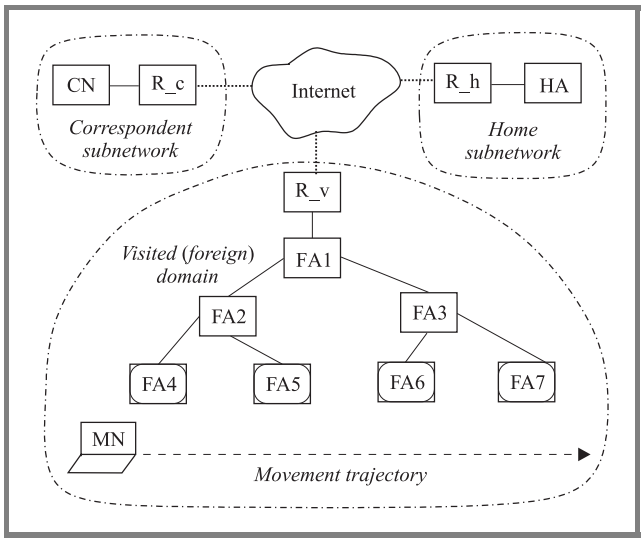


Fig. 4. Network topology used in the simulation.

a normal router or normal FA depending on the functionality implemented and the specific needs of the network. The FA hierarchy constructed this way may be used for the purpose of regional registration, or as a flat FA topology/structure in other cases. For FAs acting as leaf access routers (FA4 – FA7), it is assumed that the FAs have also been equipped with BS (base station, in 802.11 terms, access point) functionality. The coexistence of FA and BS functionalities in the same node also implies that any number of layer-2 handovers may occur as long as layer-3 IP address (a care-of address in mobile IP sense) has not been changed.

3.2. FA-HA path delay emulation

In order to emulate FA-HA path delay between the foreign subnetwork (in the visited domain) and the home subnetwork, we have set appropriately the "delay" attribute of the point-to-point link between the border router *R_v* and the Internet cloud. In the subsequent sections, we use *DD* to denote the FA-HA path delay between MN/FA and HA/CN. The combinations of nodes, like MN/FA and HA/CN, mean that we assume that MN moves typically around FA, and CN resides in the vicinity of HA, unless otherwise mentioned. The unit of *DD* is ms.

3.3. Wireless LAN configuration

WLAN is configured as IEEE 802.11, with 11 Mbit/s data rate and no RTS/CTS and fragmentation used. Each WLAN radio coverage is set to 250 metres; that ensures non-overlapping radio coverage of separate access points (BSs), eventually requiring a sort of hard handoff upon crossing the coverage boundaries.

3.4. Movement model

Mobility pattern of the MN is characterised by a horizontal linear path with constant ground speed of 30 km/h (the speed has been varied from 1 to 30 km/h when needed to observe the impact of moving speed on various performance measures). The moving speed (30 km/h) implies that MN moves faster than typical pedestrians but also slower than typical passenger vehicles in a metropolitan area. Consequently, this choice of mobility pattern results in a moderate handover rates.

3.5. Traffic model

The application traffic exchanged between the CN and MN is configured to represent either voice or data. For real-time voice traffic running on top of UDP transport protocol, we have configured it as IP telephony using voice over IP techniques where CN and MN act as clients to each other. The voice traffic exchanged between the MN and CN can start and stop in each direction in random manner, and continue until simulation stops. Unless otherwise mentioned in the relevant sections, almost all simulation results for this chapter are obtained using IP telephony voice traffic as the application traffic type. For data traffic using TCP transport protocol, we have used Ftp application (file download). Acting as client, MN requests a download of a data file from CN which is acting as a file server.

4. Simulation results

4.1. Definition of user/network satisfaction index

We define a new performance metric that is used for the evaluation of user satisfaction level. We call it user satisfaction index (USI) and use it to compare the performance of the adaptive handover control against non-adaptive handover methods.

The USI is denoted by $\mathscr{U}$. In general, $\mathscr{U}$ is defined as follows:

$$\mathscr{U} = \omega_1 \mathscr{A}_1 + \omega_2 \mathscr{A}_2 + \ldots + \omega_n \mathscr{A}_n = \sum_{i=1}^{m} \omega_i \mathscr{A}_i, \qquad (1)$$

where m is the number of application scenarios used to compute $\mathscr{U}$ and ω_i is the weighting factor for each application scenario i:

$$0 \leq \omega_i \leq 1, \qquad \sum_{i=1}^{m} \omega_i = 1 \qquad (2)$$

Table 1
Comparison of user and network satisfaction indices for handover strategies

Handover strategy		Voice (IP telephony)				Ftp data	$\mathscr{U}$, $\mathscr{N}$
		DD = 0	DD = 100	DD = 200	DD = 300	DD = 0	
NBa	USI_A	9.730	5.668	8.470	4.622	7.207	7.139
	NSI_A	10.000	10.000	10.000	10.000	10.000	10.000
NBu	USI_A	6.072	7.578	6.647	7.114	9.698	7.421
	NSI_A	9.994	9.995	9.996	9.995	9.988	9.993
NBi	USI_A	8.570	6.374	8.616	8.811	9.981	8.470
	NSI_A	9.979	9.979	9.979	9.980	9.979	9.979
RBa	USI_A	9.622	6.605	9.495	7.882	N/A	8.401
	NSI_A	7.504	7.503	7.504	7.502	N/A	7.503
RBu	USI_A	4.009	6.463	7.868	7.941	9.997	7.255
	NSI_A	7.500	7.501	7.500	7.500	7.497	7.499
RBi	USI_A	9.946	9.778	6.431	6.156	9.997	8.461
	NSI_A	7.493	7.493	7.490	7.493	7.494	7.492
AHC	USI_A	9.730	9.778	9.495	7.882	9.981	9.373
	NSI_A	10.000	7.493	7.504	7.502	9.979	8.495

$\mathscr{A}$ for a specific type of user application scenario[1] is defined as:

$$\mathscr{A}_i = \alpha_{i1}\mathscr{S}_{i1} + \alpha_{i2}\mathscr{S}_{i2} + \ldots + \alpha_{in}\mathscr{S}_{in} = \sum_{j=1}^{n} \alpha_{ij}\mathscr{S}_{ij}, \quad (3)$$

where α_{ij} is the weighting factor for each performance metric j (e.g. packet loss, delay, jitter, ...). The value of α_{ij} resides between 0 and 1, and the sum of α_{ij} for all j values should be 1:

$$0 \le \alpha_{ij} \le 1, \qquad \sum_{j=1}^{n} \alpha_{ij} = 1. \quad (4)$$

Score value for performance metric type j, $\mathscr{S}_{ij}$ (for a specific application scenario i) is defined as the fraction of performance achievement against pre-defined level of perfect performance for each performance metric i (e.g. packet loss, delay, jitter, ...). The range of value should be between 0 and 10:

$$0 \le \mathscr{S}_{ij} \le 10. \quad (5)$$

From (3) – (5), we can easily derive the possible value range of $\mathscr{A}$ as follows:

$$0 \le \mathscr{A}_i \le 10. \quad (6)$$

From Eqs. (1) and (3), we can get the general form of USI, $\mathscr{U}$ in terms of scores of performance measures, $\mathscr{S}_{ij}$ as

$$\mathscr{U} = \sum_{i=1}^{m} \omega_i \left(\sum_{j=1}^{n} \alpha_{ij}\mathscr{S}_{ij}\right) = \sum_{i}\sum_{j} \omega_i \alpha_{ij}\mathscr{S}_{ij} \quad (7)$$

[1]A user application scenario may be constructed to account for many factors, such as specific type of application traffic, end-to-end transmission delay etc.

and from Eqs. (2) and (6) the possible value of $\mathscr{U}$ falls into the range of 0 to 10:

$$0 \le \mathscr{U} \le 10. \quad (8)$$

To show how to use the USI performance metric, we give an example definition of user satisfaction index for "voice over IP" application traffic type as follows:

$$\mathscr{A}_{voip} = 0.4 * \mathscr{S}_{delay} + 0.4 * \mathscr{S}_{jitter} + 0.1 * \mathscr{S}_{loss} + 0.1 * \mathscr{S}_{thru}. \quad (9)$$

While the USI is oriented towards satisfaction level from the user's perspective, another metrics, the network satisfaction index (NSI), focuses on the satisfaction level from the network perspective.

The NSI can be thought of as a kind of *cost function*, which defines necessary cost for the use of network resources to manage operation of specific mobility mechanism. The computed value of NSI increases as the total cost of using network resources (the value of cost function) decreases.

One possible candidate of network resources to be accounted for in the NSI is available network bandwidth, normally shared among many users and thus valuable, especially in the bandwidth-limited wireless network environment. To measure the efficiency of network bandwidth usage, we express it as *signalling overhead*. The signalling includes control messages that are exchanged in the course of performing various mobile IP operations.

Similarly to the definition of USI, $\mathscr{U}$ in Eq. (1), we can define NSI, $\mathscr{N}$ as follows:

$$\mathscr{N} = \omega_1\mathscr{A}_1 + \omega_2\mathscr{A}_2 + \ldots + \omega_n\mathscr{A}_n = \sum_{i=1}^{m} \omega_i\mathscr{A}_i, \qquad (10)$$

where m is the number of application scenarios used to compute $\mathscr{N}$ and ω_i is the weighting factor for each application scenario i. The weighting factor ω_i and the definition of $\mathscr{A}$, which is network satisfaction index specific to an application scenario, can be reused as defined in Eqs. (2) and (3) respectively.

Similarly to Eq. (8), the possible range of values for $\mathscr{N}$ is as follows:

$$0 \leq \mathscr{N} \leq 10. \qquad (11)$$

4.2. Comparison of handover strategies using satisfaction indices

In this section, we illustrate the benefits of adaptive handover control against various non-adaptive handover strategies. To compare the mechanisms, we have calculated USI and NSI values for the simulation results obtained for both voice traffic and Ftp data traffic.

Throughout the rest of this section we use following notation to distinguish between the different handover strategies used in the simulations:

- NBa – basic mobile IP handover,
- NBu – mobile IP handover with buffering,
- NBi – mobile IP handover with pre-registration and bicasting,
- RBa – mobile IP handover with regional registration,
- RBu – mobile IP handover with regional registration and buffering,
- RBi – mobile IP handover with regional registration, pre-registration and bicasting.

Table 1 summarises USI and NSI values calculated for each handover strategy including AHC. We have used notations USI_A and NSI_A to indicate $\mathscr{A}$ of USI and $\mathscr{A}$ of NSI respectively for each scenario case. The calculation of user satisfaction index $\mathscr{U}$ and network satisfaction index $\mathscr{N}$ based on Eqs. (1) and (10) respectively is carried with weighting factor $\omega_i = 1/m$ assuming that each application scenario contributes equal amount to the overall satisfaction of user or network. If we assume differently, i.e. modify the contribution factors for the scenarios of choice, we may get results for $\mathscr{U}$ and $\mathscr{N}$ different from those in Table 1. From the values of $\mathscr{U}$ and $\mathscr{N}$ as in Table 1, we can conclude that AHC outperforms the other, non-adaptive handover strategies at least in respect to user satisfaction index. In respect to network satisfaction index, the AHC shows better results than handover strategies using regional registration (i.e. RBa, RBu, and RBi). However, it becomes worse than handover strategies not using regional registration (NBa, NBu, and NBi) due to additional control overhead contributed by chosen handover strategies in certain scenarios. If the network satisfaction index is our main concern (e.g. within policy framework favouring the network operator's perspective), we may obtain better values for $\mathscr{N}$ by changing the handover adaptation algorithm to select handover strategies optimised for minimum use of network resources rather than for maximum user-perceived performance.

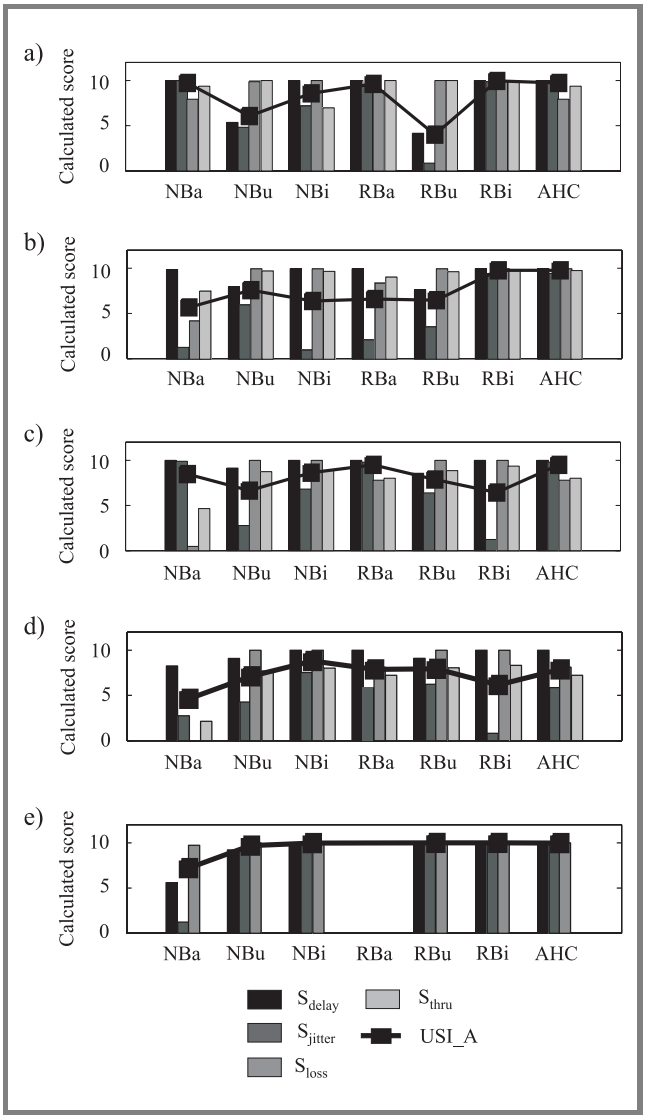


Fig. 5. Comparison of user satisfaction index for each scenario: (a) scenario 1 (VoIP, DD = 0 ms); (b) scenario 2 (VoIP, DD = 100 ms); (c) scenario 3 (VoIP, DD = 200 ms); (d) scenario 4 (VoIP, DD = 300 ms); (e) scenario 5 (Ftp, DD = 0 ms).

Using data in Table 1, we have compared user satisfaction index of $\mathscr{A}$ across all application scenarios in Fig. 5. In the figure, we have illustrated comparison of the value of $\mathscr{A}$ calculated for each simulated handover strategy. The his-

tograms in each figure represent the score values of selected performance measures, which are then used for the calculation of corresponding $\mathscr{A}$ value for each handover strategy. For scenario 5, which uses Ftp data traffic type and the network topology with delay (distance) measure $DD = 0$, we could not get satisfactory results for the RBa handover strategy. Thus, we have considered only four application scenarios to calculate the value of $\mathscr{U}$ for the RBa handover strategy. As expected, the simulation results shown in the figure confirm that one handover strategy cannot fit all scenarios. In other words, we need to select handover strategy specific to each application scenario, case by case, to maximise user satisfaction level across all cases. This justifies the need for adaptive handover control proposed in this article.

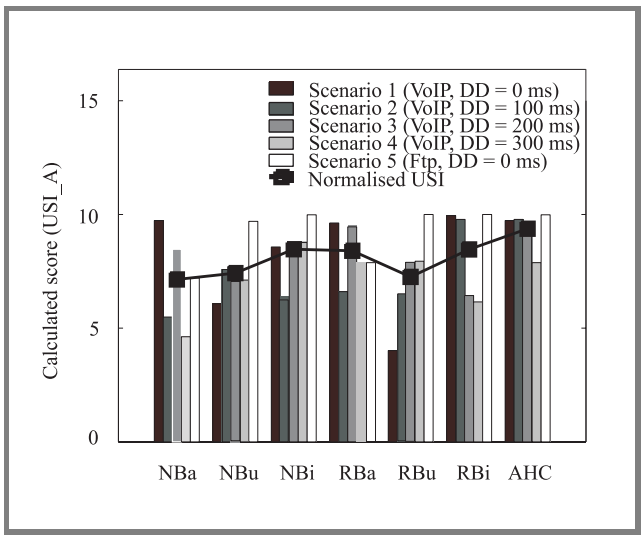


Fig. 6. Comparison of user satisfaction index over all the scenarios.

Figure 6 shows the comparison of user satisfaction index for various handover strategies, including the AHC proposed in this article. The values of $\mathscr{U}$ are obtained so as to cover all five handover scenarios (except the RBa case) with equal weighting factors for all scenarios. The AHC can be seen as outperforming the non-adaptive handover control cases. When we use the score functions as defined in Section 4.1, the estimated increase in user satisfaction index attributed to the use of AHC is about 31.3% measured against the worst-performing NBa, and about 10.7% measured against the best performing fixed handover strategy NBi.

5. Conclusions

While mobile IP protocol is generally considered to be a reasonable solution for mobility across IP subnetworks, many works available in the subject literature indicate that mobile IP alone (as specified by IETF) is not sufficient to provide seamless IP mobility, especially for time-critical (real-time or delay-sensitive) applications. The same argument can be applied to applications with other QoS attributes.

Inspired by the realisation that one solution can not suit all situations equally well, we have proposed a smart handover control framework, called the AHCA. The AHCA was designed to be flexible and open to changes of the design details such as the number of inputs and the specific handover adaptation algorithm. The component modules of the architecture can be freely substituted or modified as desired depending on the network operating conditions and characteristics of the application services and network users.

The possible extension of the AHCA could be the incorporation of modern control theory into some component modules of the architecture, as well as dynamic policy-based handover control. The implementability of the AHCA has been already confirmed through detailed functional specifications of its component modules and interfaces between them. Both qualitative and quantitative study of the benefits from using AHCA as compared to non-adaptive handover strategies is currently in progress. This extensive simulation study involves multiple network and user scenarios, as well as multiple component mechanisms of the adaptive handover.

Acknowledgements

This work was supported by the Commonwealth of Australia through its Cooperative Research Centres Program.

References

[1] H. Chaskar, Ed., "Requirements of a QoS solution for mobile IP", Draft-ietf-mobileip-qos-requirements-01.txt, work in progress, Aug. 2001.

[2] K. Ishibashi, K. Shimizu, and S. Seno, "Behavior of a mobility agent in mobile IP in order to manage the flow", Draft-ishi-mobileip-behavior-ma-00.txt, work in progress, Oct. 2001.

[3] D. B. Johnson and D. A. Maltz, "Protocols for adaptive wireless and mobile networking", *IEEE Pers. Commun.*, vol. 3, no. 1, pp. 34–42, 1996.

[4] M. E. Kounavis, A. T. Campbell, G. Ito, and G. Bianchi, "Design, implementation and evaluation of programmable handoff in mobile networks", *Mob. Netw. Appl.*, vol. 6, no. 5, pp. 443–461, 2001.

[5] K. Lee, "Adaptive network support for mobile multimedia", in *Proc. First Ann. Int. Conf. Mob. Comput. Netw. (MOBICOM 95)*, Berkeley, USA, Nov. 1995, pp. 62–74.

[6] R. R.-F. Liao and A. T. Campbell, "A utility-based approach for quantitative adaptation in wireless packet networks", *Wir. Netw.*, vol. 7, no. 5, pp. 541–557, 2001.

[7] Ch. Lu, T. F. Abdelzaher, J. A. Stankovic, and S. H. Son, "A feedback control approach for guaranteeing relative delays in web servers", *IEEE Real-Time Technol. Appl. Symp.*, TaiPei, Taiwan, June 2001.

[8] MIPv4 handoffs design team, "Low latency handoffs in mobile IPv4", Draft-ietf-mobileip-lowlatency-handoffs-v4-03.txt, work in progress, Nov. 2001.

[9] T. Park, "Seamless handoffs in heterogeneous wireless network environments", in *Proc. CRCSS Conf. 2001*, Newcastle, Australia, Feb. 2001, p. 53.

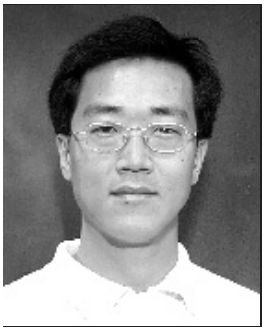
[10] T. Park and A. Dadej, "Adaptive handover between terrestrial and satellite wireless networks", in *Proc. CRCSS Conf. 2002*, Canberra, Australia, Feb. 2002, p. 46.

[11] C. Perkins, Ed., "IP mobility support for IPv4", RFC 3344, Aug. 2002.

[12] J. Solomon, *Mobile IP: The Internet Unplugged*. Englewood Cliffs: Prentice Hall, 1998.

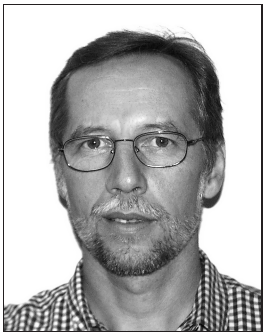
[13] M. Stemm and R. H. Katz, "Vertical handoffs in wireless overlay networks", *Mob. Netw. Appl.*, vol. 3, no. 4, pp. 335–350, 1998.

[14] H. J. Wang, R. H. Katz, and J. Giese, "Policy-enabled handoffs across heterogeneous wireless networks", in *Proc. WMCSA'99*, New Orleans, Louisiana, Feb. 1999.



Taeyeon Park received his B.Sc. and M.Sc. degrees in electronic engineering from Seoul National University, Seoul, Korea, in 1985 and 1987. He is currently working toward his Ph.D. degree at the Institute for Telecommunications Research, University of South Australia, Adelaide, Australia. From 1987 to 1998, Taeyeon Park was a senior engineer with the Computer Division of Samsung Electronics Co., Ltd., Seoul, Korea, working in the field of computer communication protocols, including TCP/IP, X.25 and OSI, as well as network cards for enterprise servers and high performance multiprocessor servers. From 1987 to 1991, he was a visiting senior engineer with the Electronics and Telecommunications Research Institute, Taejon, Korea, where he developed a computer communication system as part of the national administrative information system project funded by the Korean government. During that period, he was also an active member of the OSIA and TTA in Korea, developing a set of national specifications for OSI protocol standards. In 1999, he joined the Institute for Telecommunications Research and subsequently became one of the key researchers in the wireless data research consortium d*mobility project during 1999–2001. His current research interests include IP mobility with QoS guarantees in the wireless mobile Internet and next generation cellular network environment.

e-mail: typark@spri.levels.unisa.edu.au
Cooperative Research Centre for Satellite Systems
Institute for Telecommunications Research
University of South Australia
Mawson Lakes, SA 5095, Australia



Arek Dadej is an Associate Professor and leader of the telecommunication networks and services research group in the Institute for Telecommunications Research (ITR), University of South Australia. Over the years, he developed and delivered many undergraduate and postgraduate level courses in the areas of telecommunication networks and computer systems engineering, as well as led major industry-sponsored research projects in the area of telecommunication networks and protocols. The most recent industry-sponsored projects include studies of high-capacity 802.11 WLAN design with guaranteed QoS, a study of multicasting and scheduling techniques in satellite broadcast systems, and a study of ad hoc networking technologies. In 1999–2001, Dr. Dadej led a project sponsored by Nortel Networks, which focused on the network architecture, signalling, QoS and application session control in new generation wireless Internet. Part of the project involved 6 months visiting researcher placement with Nortel Networks research laboratories in Richardson, Texas. Dr. Dadej also led two projects within the Cooperative Research Centre for Satellite Systems, focusing on the network architectures, protocols and on-board processing for ATM and IP-based service delivery via networks of small satellites. In the past, Dr. Dadej led two major projects on self-organising tactical packet radio networks and their interoperation with fixed broadband infrastructure. Dr. Dadej's current research interests include protocols for mobility support in IP networks, QoS control in wireless/mobile network environment, and routing as well as network configuration, service discovery and user access control aspects of ad hoc networking.

e-mail: arek.dadej@unisa.edu.au
Cooperative Research Centre for Satellite Systems
Institute for Telecommunications Research
University of South Australia
Mawson Lakes, SA 5095, Australia

Paper

A concept of Differentiated Services architecture supporting military oriented Quality of Service

Marek Kwiatkowski

Abstract — This paper presents a concept of IP Differentiated Services (DiffServ) architecture in conjunction with bandwidth brokerage and policy based network management, all aimed at efficient and flexible provision of the military oriented Quality of Service (M-QoS) features in the Australian Defence (strategic) wide area network and its satellite trunk interconnections with the tactical domain. Typical DiffServ functions are analysed in the paper with regard to their roles in offering M-QoS. Some preliminary simulation results of applying these mechanisms to achieve traffic policing and differentiation for (UDP) video traffic streams, are also presented. Finally, the paper proposes the use of bandwidth brokerage in each DiffServ domain to facilitate automatic Service Level Specification (SLS) arrangements with end-user applications, and policy based network management to support the flexible implementation of bandwidth brokerage.

Keywords — Quality of Service, military networks, Differentiated Services, bandwidth brokerage, policy based network management.

1. Introduction

The term military oriented Quality of Service, introduced in [1], represents commercial QoS in conjunction with the following features. Firstly, in military packet networks, when not enough network resources are available to support QoS for all traffic flows, the flows carrying mission critical information should get preference (i.e., higher priority) over less important flows. Secondly, in overloaded networks, it is preferable to gracefully "step down" the hard QoS[1] of less important military flows instead of automatically tearing down these flows. Finally, higher flow priorities should be given for a restricted time defined by an enterprise policy.

IP differentiated services is a promising new technology that could facilitate implementation of M-QoS in the Australian Defence strategic and tactical packet communication environment [2]. This is mainly because this technology is scalable, can provide both hard[2] and soft QoS as well as graceful degradation in hard QoS to IP flows. However, DiffServ does not specify a standardised user network interface to negotiate service level specification in an automated fashion.

[1] *Hard QoS* offers an absolute reservation of resources for specific traffic, while *soft QoS* provides to some traffic a statistical preference over other traffic.

[2] DiffServ can offer hard QoS through the use of an appropriate flow admission control and queueing mechanisms in routers.

This paper presents a novel concept of IP DiffServ architecture in conjunction with bandwidth brokerage and policy based network management[3], all aimed at efficient and flexible provision of the M-QoS features in an IP-oriented subset of the long-distance Australian Defence Core communication environment (further called Defence Core for short). This subset is composed of: (1) packet oriented strategic (terrestrial) networking infrastructure composed of the IP-based routing backbone network and the ATM-based Defence Corporate Backbone Network (DCBN); and (2) Geo-synchronous Earth Orbit (GEO) satellite infrastructure used to: (a) interconnect the strategic network with tactical trunk networks; and (b) provide back-up connectivity for the strategic network.

It is stressed that currently the considered environment only offers best effort service. On the other hand, it is expected that the same environment will soon carry bulk defence multimedia (i.e., voice, video and data) traffic of different importance. It is vital to provide the M-QoS features not only in bandwidth-impoverished parts of the Defence Core (such as satellite links), but also as broadly as possible. The reason for the latter is the need to maintain the ability to transmit mission critical information even if the environment is partially destroyed.

The paper is structured as follows. Section 2 presents a general concept of the proposed architecture. DiffServ functions, bandwidth brokerage and policy based network management supporting bandwidth brokerage are described in more detail in Sections 3, 4 and 5, respectively. Conclusions and future work are given in Section 6.

2. General concept

Following the analysis provided in [2], Fig. 1 presents a combination of transmission technologies proposed to support M-QoS in the Defence Core. IPv4/IPv6 will generally be used for end-to-end communication across the (terrestrial/satellite) Defence Core between end-user applications to transfer multimedia information. An open question is whether Voice over IP (VoIP) can be carried over relatively slow satellite links. Note that DSTO is currently investigating this problem.

DiffServ will provide both hard and soft QoS as well as graceful degradation in hard QoS to IP flows. Both the strategic and tactical trunk networks will be divided into

[3] The term management refers here to longer time frame (e.g., hours, days) operations.

DiffServ domains. DiffServ will be augmented by the use of bandwidth brokerage. The latter will mainly be responsible for communication with end-user applications and flow admission control.

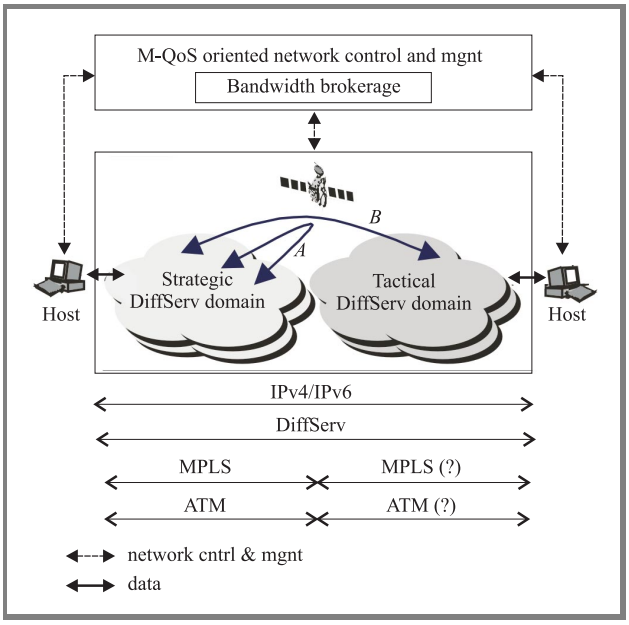


Fig. 1. General concept.

MPLS will be used to provide traffic engineering, mainly in the terrestrial part of the Defence Core. It seems to be desirable to use MPLS over a satellite to provide back-up links to the terrestrial Defence Core (see *A* in Fig. 1), thus increasing its survivability. The use of MPLS between the strategic and tactical trunk domains (see *B* in Fig. 1) requires further study.

ATM will still be used in DCBN, firstly to support MPLS switching, and secondly to continue carrying voice traffic until VoIP is implemented on a large scale. ATM may be required to transport voice over slow satellite links if IPv4/IPv6 and DiffServ do not satisfy the low jitter requirements.

Figure 2 presents the proposed DiffServ architecture in more detail. The Defence Core routing environment will be divided into a number of DiffServ domains. Routers in each domain implement a number of Per-Hop behaviours (PHBs), each characterising the externally observable forwarding treatment applied at a DiffServ-compliant router to a collection of packets each having a distinct DiffServ Code Point (DSCP) value [3]. A maximum number of 64 PHBs can be created this way. Although, the parameters such as the number of PHBs, their characteristics and groupings will be decided in a relatively static fashion through an enterprise policy, we strongly suggest that packets representing different traffic types (e.g., file transfer, interactive, voice, video) and different military precedence levels (e.g., routine, flash) should belong to separate PHBs. This approach should facilitate dimensioning of network resources (e.g., buffers in routers) and flow admission control.

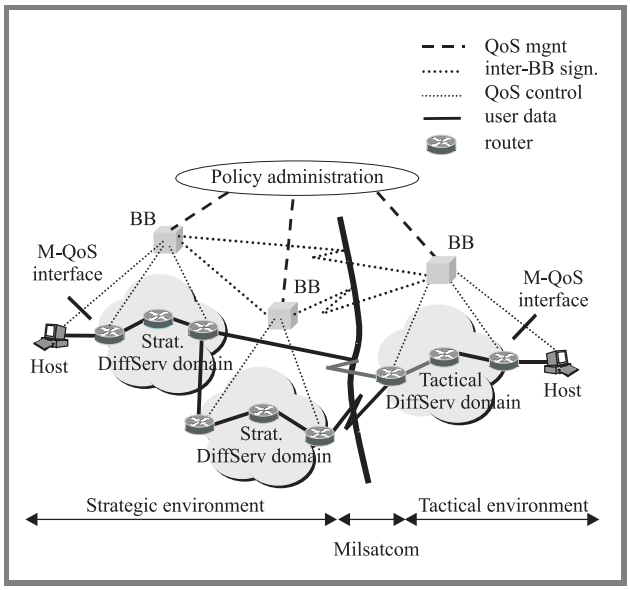


Fig. 2. Proposed concept of DiffServ architecture.

Each domain will be equipped with a single bandwidth broker (BB) entity. It will be responsible for automatic admitting to particular PHBs flows requiring (soft/hard) QoS and traversing the domain. To achieve this goal, the BB will communicate with:

- Local end-user applications (or their proxies) via a standard interface – to obtain information about parameters specifying the flows.
- Other BBs – to coordinate admission of flows that need to traverse a number of domains.

Note that BBs will not be involved in admitting best effort flows. In order to implement BB functions in a flexible and coordinated way amongst a number of BBs, policy administration will impose a single policy or a set of coherent policies onto BBs. It is assumed that these policies can often change, thus reflecting the dynamics of the battle space.

The next sections will present in more detail DiffServ functions, bandwidth brokerage and supporting it policy framework.

3. Differentiated services functions

To implement a PHB, Defence Core routers will use typical DiffServ functions such as packet classification, marking, metering, policing (dropping), shaping and queueing [3]. Below, we present how these functions can support M-QoS features in a generic way. Since primarily Cisco routers are used in the considered Defence Core environment, we will also discuss how these functions can be implemented using Cisco routers. It is noted that DSTO is currently conducting experiments with various configuration arrangements of DiffServ functions in Cisco routers.

3.1. Packet classification

Packets entering a router will be classified to one of the specified PHBs using filters. Our concept assumes that packets sent from end-user applications to ingress routers (IRs) will be classified based on the 5-tuple (source IP address, source port number, destination IP address, destination port number, and the transport protocol). The rules for classifying a flow in IRs will be delivered by the domain's bandwidth broker after deciding to admit the flow. All other (transit/egress/boundary) routers will have statically configured filters, which will classify packets based on their DSCP value set during packet marking (see below) in IRs.
We propose the use of Access Lists [4] to implement packet classification in Cisco routers.

3.2. Packet metering

Packet metering is used to measure temporal properties of a flow (flow aggregates) selected by the classifier against a traffic profile specified in a service level specification and/or against any relevant policy requirements. In our concept, packet metering will be required when implementing policing, shaping and queueing functions. As for Cisco routers, packet metering is in-built in the latter functions.

3.3. Packet marking

Packet marking is the process of setting the DSCP value in a packet based on defined rules. In our approach, packets are marked by IRs based on results of packet classification and metering. The rules for marking are specified by BBs at the time of admitting flows.
As for Cisco routers, marking will be implemented as a part of the policing mechanism (see below).

3.4. Packet policing

This process aims at discarding packets based on information provided by meters, and according to the rules specified by BBs.
In our concept, policing in IRs will be applied to all (military-essential) flows admitted by BBs. BBs will be responsible for sending to the routers a specification of dropping rules. This policing will be crucial to assure conformance of end-user application traffic to the previously negotiated SLS. It is stressed that individual best-effort flows will not be policed.
We propose the use of Two Rate Three Color Marker [5] to carry out packet dropping. In Cisco routers it can be implemented by Two-Rate Policer, which, as indicated above, also covers packet metering and packet marking.

3.5. Packet shaping

Packet shaping is a process of delaying packets within a packet stream to conform to some defined traffic profile. We expect that this function will be performed mainly by border routers to shape whole PHBs.
Cisco routers offer Generic Traffic Shaping (GTS) [4] to do the shaping.

3.6. Packet queueing

From the perspective of our architecture, the following packet queueing features are desirable:

- Use of up to 64 queues representing different PHBs.
- Group PHBs into PHB Groups, each having a separate output queue.
- Differentiate between PHBs using the same queue.
- Allocate a minimum guaranteed bandwidth per each PHB Group, thus preventing bandwidth starvation of any PHB Group.
- Automatically re-allocate unused bandwidth to other PHBs that need it, thus providing efficient use of bandwidth.
- Offer absolute priority to some chosen PHBs – a feature crucial to implement real-time, low jitter traffic (e.g., voice).

With regard to Cisco routers, the following queueing mechanisms [4] can potentially fulfil the above features:

1. Class Based Weighted Fair Queueing (CBWFQ). This scheduling discipline enables the definition of up to 64 PHBs. PHBs can be grouped into classes (i.e., PHB Groups), each having assigned a minimum guaranteed bandwidth during congestion, weight and maximum length. The weight of a packet belonging to a specific class is derived from the minimum bandwidth assigned to the class. If a queue reaches its configured queue limit, enqueueing of additional packets to the class causes tail drop.
2. Weighted Random Early Detection (WRED). This mechanism is a combination of random early detection (RED) and DSCP-based precedence. Typical for RED lower/upper thresholds and the dropping probability for the upper threshold can separately be set for different DSCPs.
3. Low Latency Queueing (LLQ). When used with CBWFQ, LLQ allows delay-sensitive packets (e.g., carrying voice) to be sent first before packets in other queues, thus giving delay-sensitive traffic preferential treatment over other traffic. A single strict priority queue is maintained for the LLQ traffic.

We are currently considering an approach where PHBs representing the same traffic type are allocated to the same queues, and WRED is used to reflect military precedence mentioned in Section 2. CBWFQ will be used to differentiate between different traffic types (e.g., data bulk transfer, video, formal messaging).

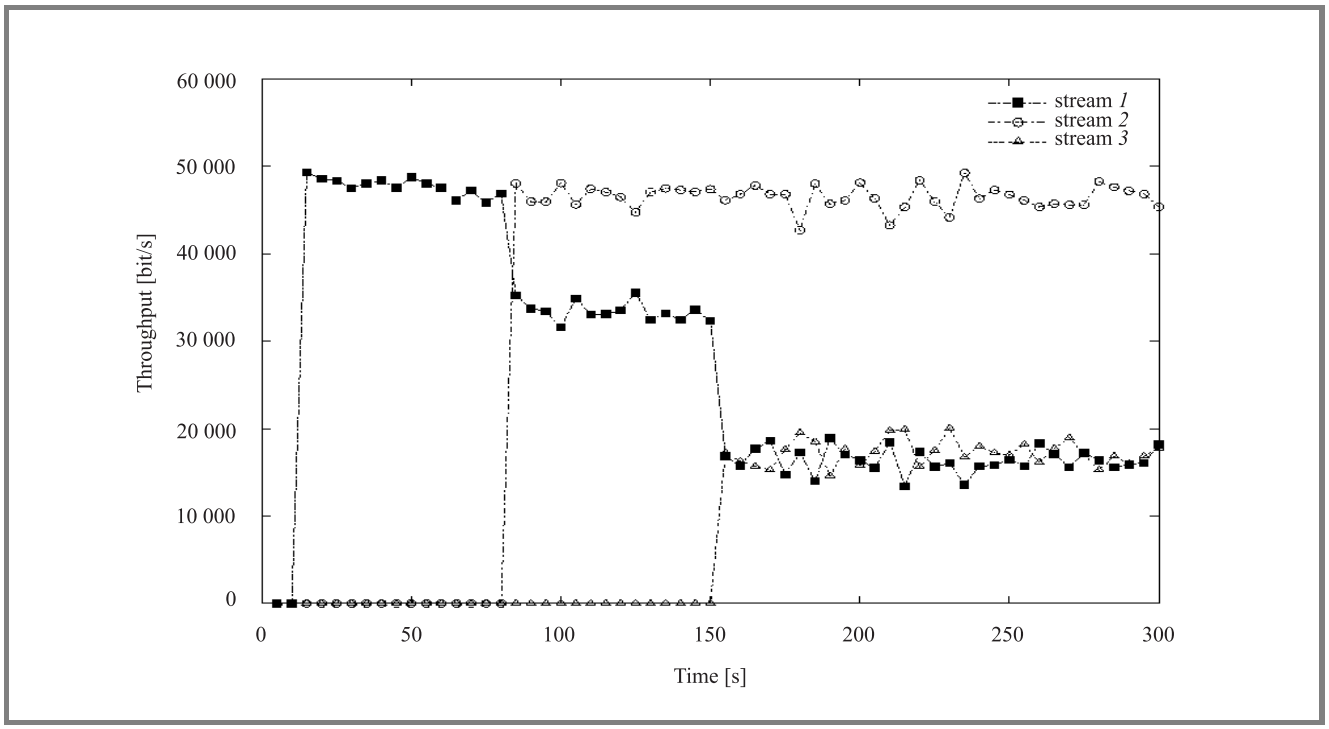


Fig. 3. Throughput versus time for (UDP) NetMeeting streams.

We have already done some preliminary simulation studies, using OPNET [6], to validate the above approach. Figure 3 shows throughput in five-second intervals obtained for three (UDP) NetMeeting [7] streams competing for the bandwidth of a 100 kbit/s link. Each stream, representing a "talking head", was first recorded on a real network and then sent in the simulation environment to a two-rate policer whose both peak rate and committed rate token buckets were set to the token replenishment rate equal to 50 kbit/s and the maximum burst size equal to 2000 bytes.

The non-conforming packets were lost. The policers also classified (using the DSCP byte) the conforming packets of streams *1* and *3* to a low priority, and of stream *2* to a high priority. Then, all the conforming packets were sent to a module simulating the WRED mechanism with two priority levels corresponding to the ones set by policers. The lower and upper thresholds for these levels were set to (1, 5) and (6, 10) packets, respectively. The thresholds were chosen to impose minimal queuing delay, and, at the same time, to enable traffic differentiation. For both priority levels, the packet dropping probability at the upper threshold was set to 0.1 and exponential weighting constant set to 2. The streams were initiated at approximately 1 min intervals to analyse the transients. It is visible in Fig. 3 that WRED gives consistent preference to the high priority stream, with no packets being dropped by WRED during the simulation run. The same figure also shows that WRED fairly divides the available bandwidth between the two low priority streams. This is confirmed by the average throughputs for streams *1* and *3* between times 170 s and 300 s being 16 234 bit/s and 17 439 bit/s, respectively. Note however, that as for TCP streams (e.g., see [8, 9]), WRED increases variability of the considered UDP traffic as well. For example, the coefficient of variation (i.e., standard deviation divided by mean) for stream *1* increases from 0.31 to 0.5, for stream *2* – from 0.31 to 0.54, and for stream *3* – from 0.32 to 0.49. More simulation/trial studies are planned for UDP and TCP streams to fully validate the presented approach.

Finally, note that a different approach to traffic differentiation, solely based on CBWFQ (i.e, with no WRED being used) is also being considered.

4. Bandwidth brokerage

We argue that the standard M-QoS interface described in [10] can be used to provide communication between an end-user application and its local BB. In brief, this interface enables the end-user application to specify for an IP flow a set of Commercial QoS Specific Parameters (e.g., peak rate, error rate, jitter) and a set of Military Specific Parameters (i.e., mission identification, precedence capturing both importance and timeliness, as well as user perceived priority). The same interface can also be used to inform the end-user application about any problems in delivering the requested/promised QoS. Finally, the M-QoS interface can be used by the end-user application to provide all the information required to perform authentication functions. Note that DSTO is currently building a prototype of an IP-based M-QoS interface using Java and Corba.

Our approach assumes that the commercial QoS parameters and military specific parameters are used by BBs involved

in the admission process of the flow to evaluate the flow's ultimate priority, which corresponds to a particular PHB. Based on this evaluation, BBs decide whether to admit the flow to the PHB or not.
The ultimate priority evaluation is based on an algorithm defined by the policy implemented in the domain. Once the flow is admitted by all the involved BBs (and possibly by the receiving end-user application), the source BB orders the source IR to invoke appropriate classification, marking and policing functions (cf. Section 3).
It is stressed that in our approach, admission of a flow may result in degradation of other, less important flows, thus reflecting the idea of graceful degradation of QoS. This may also apply to hard QoS flows. For example, a flow carrying voice with precedence level flash can partially or completely preempt another voice flow having precedence level routine.
Other expected BB functions include:

- Evaluation of time restrictions related to the military precedence level of a flow. Such restrictions may trigger a change in the flow's classification (e.g., from flash to routine) at the flow's IR.
- Modification (if required) of reservations for pending flows.
- Organising monitoring of domain's resources and tracking SLSs of active flows.

Some form of inter BB communication is necessary to perform the above functions for cross-domain flows (cf. Fig. 2). We propose to base this communication on the Simple Interdomain Bandwidth Broker Signalling (SIBBS) protocol being finalised within the QBone Project [11, 12]. Note that DSTO is currently investigating the ability of this protocol to fully support the M-QoS requirements.

5. Policy framework

The proposed approach to policy-based M-QoS management uses the IETF policy framework [13], and comprises the following generic components:

a. Policy administration (PA) – responsible for consistent DiffServ offerings across all Defence Core domains. It controls multiple BBs, automatically distributes changes to the policy, and correlates feedback from BBs regarding the health of their domains. Policy administration retrieves policies from a policy repository(ies).

b. Bandwidth broker/Policy Decision Point (BB/PDP) – plays a dual role, firstly acting as a PDP in relation to policy administration, and secondly performing typical bandwidth broker functionality.

c. Other policy servers – examples of such servers include an authentication server(s) and an accounting server(s).

d. Policy Enforcement Points – these are mainly DiffServ-enabled routers capable of enforcing QoS policy rules.

As depicted in Fig. 2, policy administration needs to cover both strategic and tactical DiffServ domains to provide end-to-end M-QoS. A complete centralisation of this administration in the strategic environment may not be desirable since a substantial amount of policy information may refer to these BB functions, which are strictly related to tactical domains operating in isolation. In such a case, sending to these domains all the policy details from a single strategic repository via relatively low bandwidth and unreliable satellite links may create performance and reliability issues. Therefore, it seems to be beneficial to distribute policy management. There are a number of possible approaches to such distribution. A plausible one is presented in Fig. 4, where a single strategic policy administrator (S-PA) controls all fixed DiffServ domains using a policy stored on its strategic policy repository (S-PR). In addition, each tactical DiffServ domain has its own PA

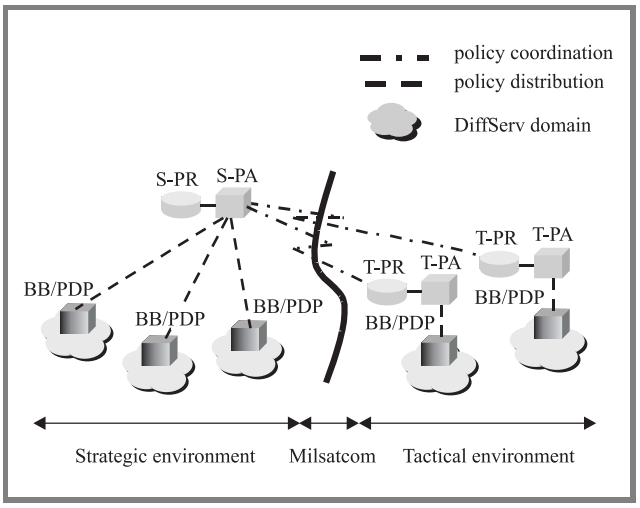


Fig. 4. Considered approach to policy distribution/coordination supporting bandwidth brokerage.

(depicted as T-PA in Fig. 4) responsible for M-QoS delivery within the domain according to a policy stored on the tactical policy repository (T-PR). To achieve consistent M-QoS across strategic/tactical domains, all policies have to be coordinated using communication between S-PA and T-PAs across satellite links (Fig. 4).

6. Conclusions and further work

In this paper we have proposed a flexible and scalable solution to implement M-QoS within the Australian Defence terrestrial/satellite Defence Core using differentiated services in conjunction with bandwidth brokerage and supporting it policy-based network management. Some encouraging preliminary simulation results of applying DiffServ mechanisms to achieve traffic policing and dif-

ferentiation for (UDP) video traffic streams have been presented.
More simulation/trial studies are planned for UDP and TCP streams to fully validate the two approaches (i.e., with and without WRED) proposed in the paper. In addition, a number of other issues require further investigation, including:

- Design of viable flow admission control algorithm(s) and performance monitoring.
- Performance aspects (e.g., consumed bandwidth) related to inter-domain brokerage signalling over slow satellite trunks.
- Efficient and reliable distribution of policy administration for dispersed strategic/tactical trunk communications involving satellite communication.

Note that DSTO is currently investigating the first two groups of issues. It is also noted that DSTO is conducting research under the aegis of The Technical Cooperation Program (TTCP) [14] on the applicability of the proposed DiffServ architecture to a coalition environment.

Acknowledgments

The author would like to thank Mathew Elliot for providing the simulation results.

References

[1] M. Kwiatkowski and P. George, "A network control and management framework supporting military Quality of Service", in *Proc. IEEE MILCOM'99*, Atlantic City, USA, 1999, vol. 2, pp. 1161–1165.

[2] M. Kwiatkowski, "A Concept of Defence Core Communication Infrastructure Supporting M-QoS", (unclassified) DSTO Techn. Rep., DSTO-TR-1220, Oct. 2001.

[3] S. Blake *et al.*, "An Architecture for Differentiated Services", IETF, RFC 2475, Dec. 1998.

[4] Cisco IOS Release 12.2, 2001.

[5] Heinanen *et al.*, "A Two Rate Three Color Marker", IETF, RFC 2968, Sept. 98.

[6] OPNET, Version 8.0, OPNET Technologies Inc., 2001.

[7] NetMeeting, Version 3.01 for Windows, Microsoft Corporation, 1999.

[8] Th. Bonald, M. May, and J. Bolot, "Analytic evaluation of RED performance", in *Proc. IEEE INFOCOM'00*, 2000, vol. 3, pp. 1415–1424.

[9] S. H. Low *et al.*, "Dynamics of TCP/RED and a scalable control", in *Proc. IEEE INFOCOM'02*, New York, USA, 2002, vol. 1, pp. 239–248.

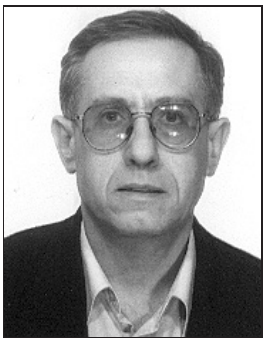
[10] P. Blackmore, P. George, and M. Kwiatkowski, "A Quality of Service interface for military applications", in *Proc. IEEE MILCOM'00*, Los Angeles, USA, 2000, vol. 1, pp. 470–474.

[11] B. Teitelbaum *et al.*, "Internet2 QBone: building a testbed for differentiated services", *IEEE Network*, vol. 13, no. 5, pp. 8–16, 1999.

[12] QBone Signaling Design Team, http://qbone.internet2.edu/bb/index.shtml

[13] IETF, Policy Charter, http://www.ietf.org/html.charters/policy-charter.html

[14] The Technical Cooperation Program (TTCP), http://www.dtic.mil/ttcp



Marek Kwiatkowski received the M.Sc. degree from the Silesian Technical University, Gliwice, Poland in 1979 (computer science), and the Ph.D. degree from AGH University of Technology, Cracow, Poland, in 1990 (telecommunications). From 1991 until 1998, he worked at the Teletraffic Research Centre, University of Adelaide, Australia, first as a Post Doctoral Fellow, and from 1995 as a Research Fellow. Since June 1998 he has been working at the Communications Division of the Defence Science and Technology Organisation (DSTO), Adelaide, as a Senior Research Scientist. His main research interests include control and management aspects of narrowband and broadband networks.
e-mail: marek.kwiatkowski@dsto.defence.gov.au
Defence Science and Technology Organisation
PO Box 1500, Edinburgh, SA 5108, Australia

Paper

Bandwidth broker extension for optimal resource management

Shaleeza Sohail and Sanjay Jha

Abstract — Bandwidth broker (BB), resource manager of differentiated services domain cannot provide per domain behavior (PDB) attribute information to customers and neighboring domains at the time of service level agreement (SLA) negotiation. Extending BB's functionality to calculate PDB attributes can help it to negotiate SLAs dynamically and efficiently. Using current measurements or historic data about PDB attributes, bandwidth broker can perform off-line analysis to evaluate the range of quality of service (QoS) parameters that its domain can offer. Using these values BB can perform optimal capacity planning of the links and provide better QoS guarantees.

Keywords — bandwidth broker, per domain behavior, resource management.

1. Introduction

In order to support quality of service in the network, new architectures such as IntServ and DiffServ have been proposed in the IETF. These architectures support diverse service levels for multimedia and real-time applications. DiffServ architecture is capable of providing well defined end-to-end service over concatenated chains of separately administered domains by enforcing the aggregate traffic contracts between domains [2]. At the interdomain boundaries, service level agreements specify the transit service to be given to each aggregate [11]. SLAs are complex business related contracts that cover a wide range of issues, including network availability guarantees, payment models and other legal and business necessities. SLA contains a service level specification (SLS) that characterizes aggregates traffic profile and the per hop behavior (PHB) to be applied to each aggregate. PHB is the treatment that a packet receives in a DiffServ domain at any router. All traffic belonging to a particular class experiences same PHB. To automate the process of SLS negotiation, admission control and configuration of network devices correctly and to support the provisioned QoS, each DiffServ network may be added with a new component called a bandwidth broker [13].

Bandwidth broker is a complex entity that might need integration of several technologies such as standard interface for inter/intra domain communication, protocol entity for communication, standard protocol and database. Organizational policies can be configured by using the mechanism provided by BB. On the inter domain level BB is responsible for negotiating QoS parameters and setting up bilateral agreements with neighboring domains. On intra domain level BB's responsibilities include configuration of edge routers to enforce resource allocation and admission control. With the help of simulation [6], it has also been suggested that bandwidth broker in DiffServ architecture can be effectively used to provide QoS to real time applications like VoIP. Moreover these studies also indicate that admission control mechanism of BB improves the profit for the ISPs by improving network resource utilization.

Currently BB keeps no information about values of QoS parameters that it can offer. Some time critical applications or their users may need to know the exact treatment that their application will get in terms of delay, jitter, packet loss etc. For example in case of multi-party tele-conferencing, a user may need guarantee that his/her application's traffic will not suffer end-to-end delay more than 50 ms. The Internet service provider (ISP), using DiffServ in its domain can only guarantee that the user's traffic will be assigned to a particular behavior aggregate (BA) and PHB. ISP can guarantee the PHB that the aggregate traffic will experience but cannot guarantee the QoS parameters like delay, jitter and packet loss etc. To know these attributes ISP needs to know the per domain behavior of the domain. PDB is the edge-to-edge treatment that traffic receives in a DiffServ domain [7]. In order to efficiently negotiate SLS in this scenario and satisfy user's demands an ISP can use BB to calculate these QoS parameters for different classes of traffic. BB can perform off-line analysis on the current results or historic data and find out the QoS values that it can offer. In this manner BB will have a complete knowledge about the range of QoS parameters supported by the domain at any particular load condition. In order to improve the QoS values BB can negotiate SLAs dynamically with the neighboring domains.

The rest of the paper is organized as follows: Section 2 has a brief description of BB. Section 3 describes per domain behavior and related works are mentioned in Section 4. Section 5 relates BB with PDB. Section 6 reports on the simulation studies that we performed. Section 7 elaborates the impact of calculating PDBs at BB. Section 8 concludes the paper and give some ideas for future work.

2. Bandwidth broker

The main resource management entity in DiffServ domain is a BB. The BB maintains policies, and negotiates SLAs with customers and neighboring domains. The interaction of BB with other components of DiffServ domain as well as the end-to-end communication process in DiffServ domain is shown in Fig. 1. The figure shows that when a flow needs

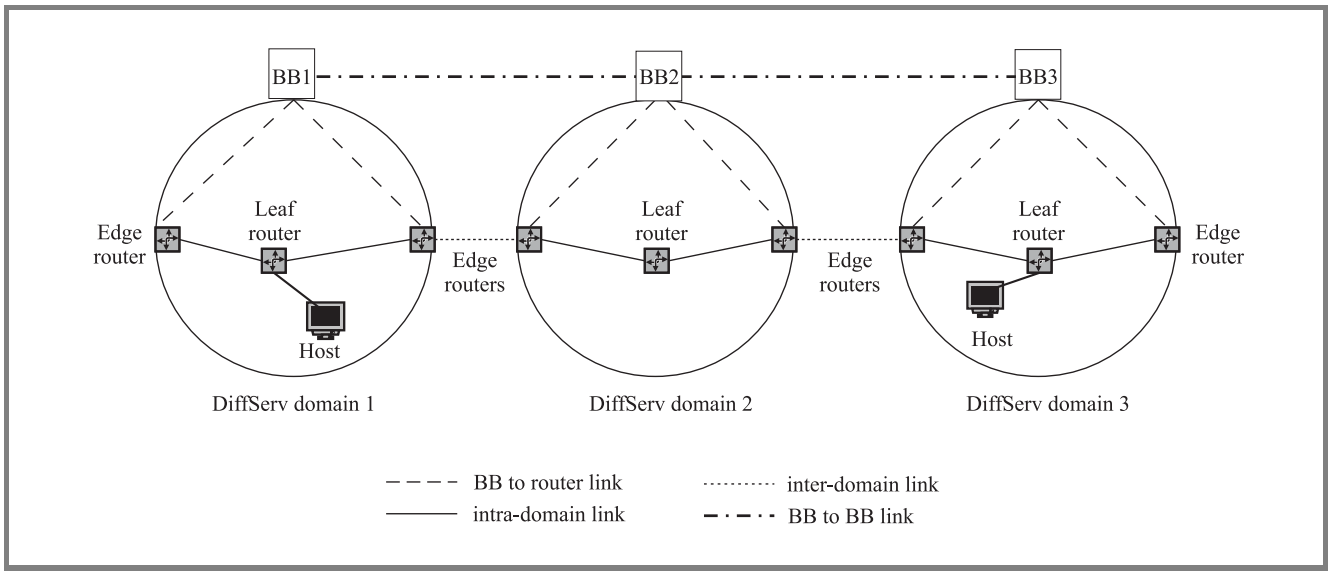


Fig. 1. Role of bandwidth broker in DiffServ.

to enter the DiffServ domain or a local user wants to send some traffic, BB is requested to check related SLA. BB is responsible for admission control as it has global knowledge of network topology and resource allocation. BB decides as to allow the traffic or not on the basis of previously negotiated SLAs. In case of a new flow, a BB might have to negotiate a new SLA with the neighboring domain depending upon the traffic requirements. Once BB allows the traffic, the edge router or the leaf router needs to be reconfigured by BB. SLA negotiation is a dynamic process due to the ever changing requirements of the network traffic.

3. Per domain behavior

PDB consists of measurable attributes that define the treatment that each PHB will experience from edge-to-edge in a particular domain [7]. For example the PDB may specify the edge-to-edge delay that the traffic belonging to assured forwarding (AF) class may experience in the domain. PDB depends upon the PHB as well as the load conditions and some domain specific parameters like domain topology, links used to transfer traffic etc. The sum of same type of PDB parameters of all the domains from which the flow will pass gives the end-to-end QoS parameters for the particular flow. The attributes that can be part of the PDB are like delay, packet loss and throughput etc. The network specific parameters need to be specified for the measurement of these attributes [7].

4. Related work

The basic BB model is extended in virtual private network (VPN), supported by DiffServ to implement and negotiate range based SLAs [16, 17]. The resource wastage is reduced by using range based SLAs as the mechanism provides better resource utilization when user is unable to specify the exact resource requirement [15, 17]. IETF has defined PDB and the rules for its specification [7]. Multiple types of PDB are also defined; assured rate [9], virtual wire [5] and lower effort [8] are some of the examples. However ISPs can define their own PDBs according to their network requirements. Different research groups are studying the QoS attributes relation with the network parameters [1, 6].

5. Bandwidth broker calculating PDB

The bandwidth broker is a management entity that has a complete and up-to-date picture of the topology of the domain. Hence, the BB is the best possible entity that can be extended to calculate PDBs. In general the areas about which BB maintains information are policy, SLA, network management, and current resource allocation status [12]. Adding the functionality in the BB to calculate PDB and advertise them at the time of SLA negotiation can result in better user satisfaction. Moreover by knowing the PDB experienced by different PHBs, the BB can efficiently and optimally allocate resources.

The BB may choose to define a range of the QoS attributes supported by its domain by calculating maximum, minimum and average values of these attributes at various load conditions. BB can use these values to indicate the QoS treatment that any traffic may receive. To support particular value of QoS parameter BB uses this information for admission control as well as for SLA negotiation. For example BB may need to provide 50–100 ms of delay to any particular PHB. However from previously performed analysis BB knows that it is not possible at the present load conditions of the network. The solution is to negotiate the increase of bandwidth with the neighboring domains and considering the QoS requirement before accepting new

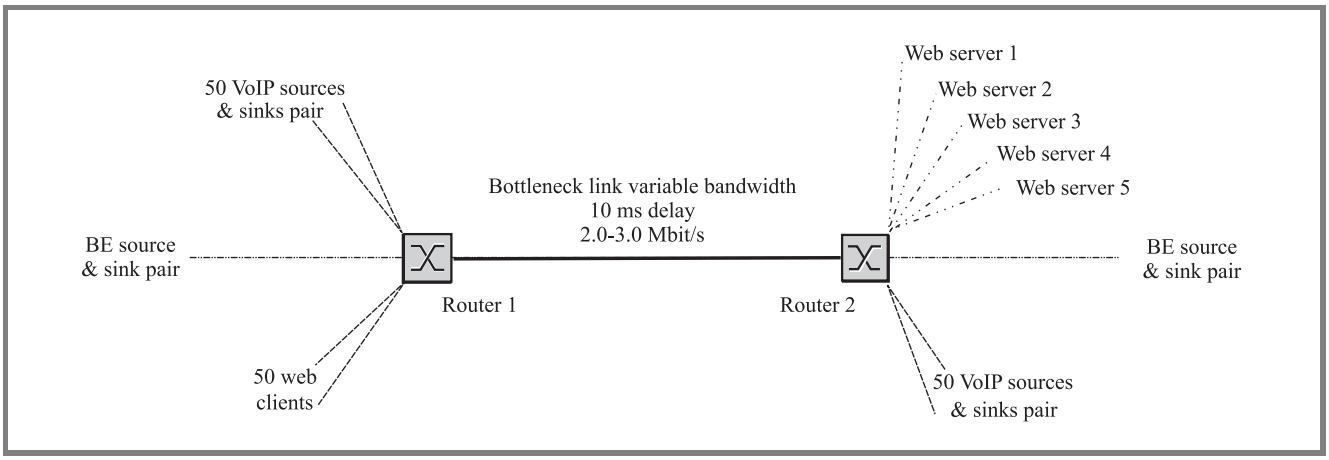


Fig. 2. Topology. All access links have 10 Mbit/s bandwidth and 0.1 ms delay.

connections. In this manner BB can optimally perform capacity planning of the links of the domain.
The simulation study in the next section calculates different values for some QoS attributes by changing few parameters. This simulation study shows that by using simple mechanism a BB can be extended to monitor different attributes of PDB.

6. Simulations

The simulations are performed using the network simulator (NS) [10]. Some of the simulation parameters are taken from the simulation study of DiffServ [1], however the scheduler used is weighted fair queuing [14]. In the simulation, the sources are generating traffic at constant rate and the bandwidth of the link changes for each simulation run. The values of QoS parameters change with the change of link capacity and the minimum link capacity can be found in this way that can support some particular QoS value. The impact of capacity on the attributes can help BB to decide what link capacity to use to transfer traffic, if certain QoS requirements like delay, packet loss etc; at a particular load condition, are to be fulfilled.

6.1. Simulation topology and parameters

The network is a simple dumb bell shape as shown in Fig. 2. There is one bottleneck link which has varying bandwidth with 10 ms delay. On one side of bottleneck link there are 50 web clients and 50 voice sources/sinks. On the other side there are 5 web servers and 50 voice sources/sinks. There are two best effort sources and sinks to produce congestion on the bottleneck link. There is minimum bandwidth reserved for the BE sources but these sources always send at the rate higher than the rate allocated to them.

Following three types of traffic are used in the network:

1. **Voice traffic**. The voice traffic is modeled as VoIP and there is no compression and silence suppression [1]. There are 50 voice source/sink pairs at each side of bottleneck link. The VoIP sources are actually UDP ON/OFF sources. The inter call gap is 15 minutes and the mean rate of traffic is 86.4 kbit/s. The 80% of the calls are short calls and rest are long calls. The on time for short calls is 3 minutes and that for long calls is 8 minutes. The VoIP traffic is assigned expedited forwarding (EF) PHB. EF PHB provides low latency, low loss, low jitter, assured bandwidth service through DiffServ domains [4].
2. **Data traffic**. The data traffic is web traffic generated by the request and reply interaction of HTTP/1.1 between web servers and clients. There are 50 clients requesting to 5 web servers [1]. The number of objects requested are random. This traffic is assigned to assured forwarding PHB. AF PHB provides forwarding assurance to the packets belonging to this PHB [3].
3. **Best effort**. The best effort (BE) traffic is a simple UDP source generating at the rate higher than the rate it is allowed. The traffic assigned to BE PHB has no assurance from DiffServ domain.

6.2. Simulation results

The end-to-end delay and packet loss for different classes are measured. These values vary with the change of the capacity of the bottleneck link. The results are shown in two different ways. There are tables in Section 6.2.1 that show the packet loss for traffic belonging to all PHBs. The graphs in Section 6.2.2 show the average end-to-end delay.

6.2.1. Packet statistics

Tables 1 and 2 show the packet loss statistics of the traffic of different classes. In Tables CP is the DiffServ code point of the packet. TotPkts and TxPkts are the counters of packets received and packets transmitted respectively. The ldrops are the packets that are dropped due to link overflow. Edrops mean the packets dropped due to random early detection (RED) early dropping mechanism. The code

point 10, 20 and 30 are for the traffic belonging to EF, AF and BE classes respectively. The code point 21 is assigned to out-of-profile packets of AF class.

Table 1
Packets statistics at router 2: bandwidth 2.0 Mbit/s

CP	TotPkts	TxPkts	ldrops	edrops
All	832 524	341 706	490 818	0
10	80 543	73 125	7 418	0
20	173 118	172 115	1 003	0
21	78 863	0	78 863	0
30	500 000	96 466	403 534	0

Table 2
Packets statistics at router 2: bandwidth 3.0 Mbit/s

CP	TotPkts	TxPkts	ldrops	edrops
All	842 908	462 496	380 412	0
10	82 617	82 615	2	0
20	171 882	171 672	210	0
21	88 409	0	88 409	0
30	500 000	208 209	291 791	0

Tables 1 and 2 show the packet statistics of the traffic at router 2 whereas the bottleneck link capacity is 2.0 Mbit/s and 3.0 Mbit/s respectively. By comparing the tables it is obvious that the number of dropped packets reduces considerably with the increase of the link capacity. If same tables are to be used by BB to define PDB then BB may interpret those in the following manner:

1. For the specified load conditions the link with bandwidth of 2.0 Mbit/s has packet drop of almost 10% for EF traffic.
2. For the same load conditions the packet drop for EF traffic for the link with bandwidth of 3.0 Mbit/s is less than 1%.
3. If the SLA with user requires packet loss less than 1% then link capacity should be 3.0 Mbit/s.
4. BB can indicate during SLA negotiation that the packet loss for EF traffic is less than 1%.

6.2.2. End-to-end delay

The graphs presented in this section show end-to-end delay for VoIP and best effort traffic during the simulation. In Figs. 3 and 4 along *x*-axis is the time in seconds and *y*-axis has the average end-to-end delay in seconds. From the graphs, it can be observed that in the beginning of the simulation, the delay is lower but as more and more sources start sending traffic the average delay increases. The end-to-end delay mentioned here is the average of all the sources belonging to that particular PHB.

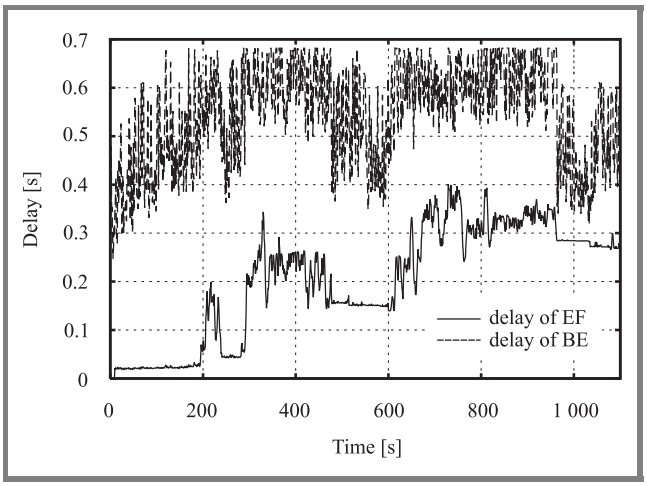


Fig. 3. End-to-end delay of EF and BE PHB when the capacity of the bottleneck link is 2.0 Mbit/s.

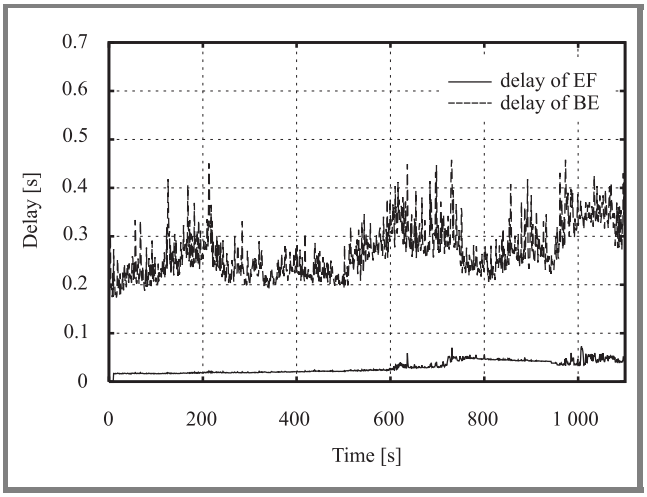


Fig. 4. End-to-end delay of EF and BE PHB when the capacity of the bottleneck link is 3.0 Mbit/s.

Figures 3 and 4 show the average end-to-end delay of EF and BE traffic when the capacity of the bottleneck link is 2.0 Mbit/s and 3.0 Mbit/s respectively. By comparing Figs. 3 and 4 it can be seen that the average end-to-end delay reduces considerably from 280 ms to less than 50 ms with increase of bandwidth. BB may interpret these results in order to calculate PDB in the following manner:

1. For the specified load conditions, the link with bandwidth of 2.0 Mbit/s has average end-to-end delay of almost 280 ms for EF traffic.
2. For the same load conditions the average end-to-end delay for EF traffic for the link with bandwidth of 3.0 Mbit/s is less than 50 ms.
3. If user SLA requires average delay less than 50 ms then link capacity should be 3.0 Mbit/s.
4. BB can indicate during SLA negotiation that the edge-to-edge average delay for EF traffic is less than 50 ms in its domain.

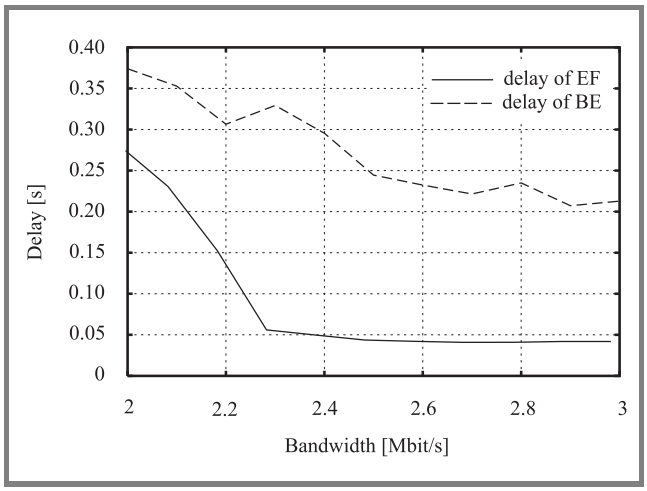


Fig. 5. End-to-end delay of EF and BE PHB with the variation of capacity of bottleneck link.

Figure 5 shows the variation of end-to-end delay with the variation of capacity of bottleneck link. Along *y*-axis is the average delay in seconds and along *x*-axis is the link capacity in Mbit/s. This type of graph can give an idea as how much delay can be accepted when the traffic passes through a specified link at a particular load.

6.3. Discussion

It is evident from the graphs and the tables presented in the previous subsections that by using a simple approach like this, BB can find QoS attributes for PDB of different PHBs. BB may choose to specify the range of these QoS parameters that can be supported by the domain.

The packet statistics tables show the number of packets lost for every type of PHB. These values can be used to perform off-line analysis by BB to find out the minimum bandwidth required to support some specific packet loss value for particular PHB. BB may get these packet loss statistics at different time of the day or month. These statistics can perform important role in performing future capacity planning.

The end-to-end delay is a very important QoS parameter for optimal performance of some applications. Calculating it with a simple mechanism used in the simulations can greatly reduce the overhead. We have only calculated the average delay however calculating maximum or minimum delay with the same mechanism is a trivial task. BB may use these values to specify the range of delay that particular PHB can suffer. BB can perform an efficient analysis of these values for future capacity planning as well as efficient QoS guarantees.

7. The impact

There is a great concern about the scalability issues regarding bandwidth broker. There is a rapid growth of QoS applications like VoIP and real time content delivery, which require dynamic QoS control and management. The ability of BB to handle large volumes of flows is debatable. The badly designed BB can become the bottleneck to allocate the network resources effectively even in the scenarios when the network is underutilized.

The extension of BB to optimize the resource allocation by using PDB values can increase the BB's complexity exponentially. The approach discussed in this paper greatly depends upon the ability of BB to perform efficient offline analysis of the PDB information. By adding the ability to BB to analyze PDB information offline, the already overburden BB does not have to perform any additional processing for optimal SLA negotiation and resource allocation.

The scalability issues of BB are dealt by introducing the concept of multiple BB architecture [18] and multi layer BB architecture [19]. In the multiple BB architecture there is one central BB and number of edge BBs in the domain. The total domain resources are divided among edge BBs and it is the responsibility of the edge BB to efficiently manage the resources allocated to it. The central BB is responsible for overall resource management and coordination among edge BBs. The multi layer BB architecture distributes the functionality of BB among multiple entities and some of those entities are structured hierarchically to further break down the scalability problem. The BB's extension to optimize the resource allocation by using PDB information can be added to multi layer and multiple BB architecture. However, in both the architectures the information analysis should be performed in a way that the analysis results are updated at all resource management entities at the same time.

8. Conclusion and future work

An idea of using BB to measure and calculate attributes of PDB for dynamic SLA negotiation is proposed in this paper. Simulation was performed to give idea about the mechanism that can be used to relate these attributes to parameters of the network.

Introducing this type of mechanism in BB can increase its complexity, however the magnitude of this complexity entirely depends upon the ISPs. During SLA negotiation these attributes of PDB for different PHBs can give ISP an edge over others in defining their services better and in the terms that are better understood by users. Moreover ISPs can use this mechanism in their domain's BB to provide extra motivation to the user to select their services.

The DiffServ working group has defined PDBs but how, when and where to calculate and advertise these are the topics for future research. We have presented a simulation study to elaborate our idea of adding the ability of calculating PDBs in BB. We are planning to do more simulation studies in this area using complex topologies and calculating more PDB attributes.

References

[1] U. Fiedler, P. Huang, and B. Plattner, "Towards provisioning DiffServ intranets", in *Lecture Notes in Computer Science*, 2001, vol. 2092, pp. 27–43.

[2] M. Fine *et al.*, "An architecture for differentiated services", Internet request for comments RFC2475, IETF, Dec. 1998.

[3] J. Heinanen *et al.*, "Assured forwarding PHB group", Internet request for comments RFC2597, IETF, June 1999.

[4] V. Jacobson *et al.*, "An expedited forwarding PHB", Internet request for comments RFC2598, IETF, June 1999.

[5] V. Jacobson *et al.*, "The virtual wire per-domain behavior", Internet draft, IETF, July 2000.

[6] G. Kim, P. Mouchtaris, S. Samtani, R. Talpade, and L. Wong, "QoS provisioning for VoIP in bandwidth broker architecture: a simulation approach", in *Commun. Netw. Distrib. Syst. Model. Simul. Conf. CNDS* '01, Phoenix, USA, Jan. 2001.

[7] K. Nichols and B. Carpenter, "Definition of differentiated services per domain behaviors and rules for their specification", Internet request for comments RFC3086, IETF, Apr. 2001.

[8] K. Nichols *et al.*, "A lower effort per-domain behavior for differentiated services", Internet draft, IETF, June 2002.

[9] N. Seddigh *et al.*, "An assured rate per-domain behavior for differentiated services", Internet draft, IETF, Feb. 2001.

[10] Network simulator NS-2, Jan. 2003, http://www.isi.edu/nsnam/ns/

[11] S. Sohail and S. Jha, "The survey of bandwidth broker", Tech. Rep. UNSW CSE TR 0206, School of Computer Science and Engineering, University of New South Wales, Sydney, Australia, May 2002.

[12] B. Teitelbaum and P. Chimento, "Qbone bandwidth broker architecture", Work in Progress, June 2002, http://qbone.ctit.utwente.nl/deliverables/1999/d2/bboutline2.htm

[13] B. Teitelbaum and R. Geib, "Internet2 QBone: a test bed for differentiated service", in *INET '99*, *Internet Glob. Sum.*, San Jose, USA, June 1999.

[14] "WFQ scheduler for NS-2", June 2002, http://www.tik.ee.ethz.ch/fiedler/provisioning.html

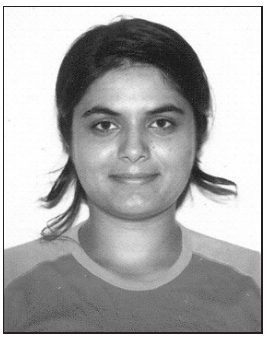
[15] T. Braun, M. Gunter, I. Khalil, R. Balmer, and F. Baumgartner, "Virtual private network and quality of service management implementation", Tech. Rep. IAM-99-003, CATI, July 1999.

[16] T. Braun and I. Khalil, "Edge provisioning and fairness in VPN-DiffServ networks", in *9th Int. Conf. Comput. Commun. Netw. ICCCN 2000*, pp. 424–433.

[17] T. Braun and I. Khalil, "A range based SLA and edge driven virtual core provisioning in DiffServ-VPNs", in *26th Ann. IEEE Conf. Loc. Comp. Netw. LCN'2001*, Tampa, USA, 2001.

[18] Z. Zhang, Z. Duan, and Y. Hou, "On scalable design of bandwidth brokers", in *IEICE Trans. Commun.*, E84-B (8), 2001.

[19] G. Politis, P. Sampatakos, and I. Venieris, "Design of a multi-layer bandwidth broker architecture", in *Interworking*, 2000, pp. 316–325.



Shaleeza Sohail is a Ph.D. student in the School of Computer Science and Engineering at the University of New South Wales (UNSW), Sydney, Australia. She has her masters degree from UNSW in 2000. She is a member of Network Research Laboratory of UNSW and currently being supervised by dr Sanjay Jha. Currently she is working on the bandwidth broker's extensions to provide network level QoS to next generation applications.
e-mail: sohails@cse.unsw.edu.au
School of Computer Science and Engineering
University of New South Wales
Sydney, Australia

Sanjay Jha is an Associate Professor (Networks) at the School of Computer Science and Engineering (CSE) at the University of New South Wales. He has a Ph.D. degree from the University of Technology, Sydney, Australia. He is the founder of the Network Research Laboratory at CSE, UNSW. His research activities cover a wide range of topics in networking including quality of service, mobile/wireless Internet, and active/programmable network. He is the principal author of the book *Engineering Internet QoS* (Artech House, 2002). He has been working as an industry consultant for major organizations such as Canon Research Lab (CISRA), Lucent and Fujitsu. In his previous job, he was a lecturer at the School of Computing Sciences, University of Technology (UTS), Sydney. He also worked as systems engineer for the National Informatics Centre, New Delhi. He was a visiting scholar at the Distributed Computing and Communications Laboratory, Computer Science Department, Columbia University, New York, and Transmission Systems Development Department of Fujitsu Australia Ltd., Sydney.
e-mail: sjha@cse.unsw.edu.au
e-mail: jhask@ieee.org
School of Computer Science and Engineering
University of New South Wales
Sydney, Australia

Paper

Manipulation of compressed data using MPEG-7 low level audio descriptors

Jason Lukasiak, David Stirling, Shane Perrow, and Nick Harders

Abstract — This paper analyses the consistency of a set of MPEG-7 low level audio descriptors when the input audio stream has previously been compressed with a lossy compression algorithm. The analysis results show that lossy compression has a detrimental effect on the integrity of practical search and retrieval schemes that utilize the low level audio descriptors. Methods are then proposed to reduce the detrimental effects of compression in searching schemes. These proposed methods include improved searches, switched adaptive scalar and vector prediction, and other prediction schemes based on machine learning principles. Of the proposed schemes the results indicate that searching which incorporates previous and future frames combined with machine learning based prediction best nullifies the effects of compression. However, future scope is identified to further improve the reliability of the MPEG-7 audio descriptors in practical search environments.

Keywords — MPEG-7, metadata, multimedia description, machine learning, multimedia retrieval.

1. Introduction

With the ever increasing volume of multi-media (MM) data available via shared networks, such as the Internet or even large organizational intranets, meaningful and efficient storage, retrieval, archiving and filtering of the available MM data is becoming increasingly difficult. Current text based search and retrieval schemes rely on meaningful textual tags being associated with every MM item. Such text based methods are limited in usefulness by the quality (content) of the tag used. For example, it may be possible to find a particular MM item via Authors name or title, but it is extremely unlikely that content specific features such as colour, melody or frequency structure would be identifiable using a text based tag. Overcoming this limitation is the realm of the new MPEG-7 standard [1]. This standard provides a structured framework for describing MM content in a platform independent environment [1, 2].

At its lowest level, the MPEG-7 standard specifies a set of low level descriptors that are calculated directly from the MM content [2]. Examples of the low level descriptors generated are colour space [2] and audio spectral envelope [3]. A detailed description of the entire MPEG-7 standard can be found in [1] and an excellent overview in [2].

Having descriptors associated with MM data that describe the actual content of the data, provides the potential for powerful manipulation of content supply and consumption. The manipulation could involve finding all MM items in a database that have a blue background or selecting the audio segments that represent specific sources (such as dogs barking) [4, 8]. Such searching could feasibly return all of the images in a database that contain "Sampras playing tennis" [5].

Whilst the proposed MPEG-7 standard offers a powerful new scheme for control of MM data, there are a number of issues that may limit practical application of the standard. Examples of such limitations are the complexity involved and the integrity of the descriptors in compressed environments. As a substantial amount of MM data is compressed before storage using various lossy compression schemes, such as MP3 for audio and JPEG for images, the perceptually redundant information from the signal is substantially removed. Consequently, the effect of compression on the integrity of the descriptors is extremely important for practical applications. For example, can we find an MP3 audio file (or an audio segment previously compressed by MP3) that matches our target song, using the low level descriptors generated from a CD?

The effect of compression on the MPEG-7 low level audio descriptors is the focus of this paper. A thorough analysis into the effect of compression on five of the seventeen low level audio descriptors is presented in Section 2. These five descriptors were selected due to space constraints in presenting results for the full set of descriptors, and also, due to these descriptors being frame based (as opposed to entire file based). The combination of the prescribed descriptors also presents a compact description of the underlying audio data.

Once identified some preliminary methods for reducing the effects of compression on the low level descriptors are detailed in Section 3. Finally the major points are summarized in Section 4.

2. Effect of compression on low level descriptors

The five audio low level descriptors selected for the analysis were: audio power (AP), audio waveform (AW), audio spectrum ennvelope (ASE), audio spectrum centroid (ASC)

and audio spectral spread (ASS). A full description of these can be found in [3].

To determine the effect of compression on the audio low level descriptors, two well known audio compression algorithms, MP3 [6] and WMA [7], were used to compress and uncompress 90 16-bit 44.1 kHz sampled audio files. Each of the audio files was of approximately 10 seconds duration, with 50 files representing instrumental only signals (inst) and 40 representing combinational signals (comb) such as pop music.

The MP3 encoder was operated at 128, 160 and 192 kbit/s and the WMA coder was operated at both 128 and 160 kbit/s.

The MPEG-7 low level audio descriptors defined above, were then calculated for both the files which had been compressed/uncompressed and the original files. A frame size of 20 ms was used to calculate the descriptors.

2.1. Objective measures

To give an objective indication of the effect of compression on the descriptors, the segmented signal to noise ratio (SegSNR) was used. This was calculated as:

$$\text{SegSNR} = \frac{1}{N}\sum_{x=0}^{N-1} 10\,\log_{10}\left(\frac{\sum_{n=1}^{r} x_n^2}{\sum_{n=1}^{r}(x_n-\overline{x_n})^2}\right), \qquad (1)$$

where x_n is the descriptor from the original file, $\overline{x_n}$ is the descriptor from the compressed file, r is the dimension of the descriptor per frame and N is the number of frames in the file. The SegSNR for each descriptor and compression configuration was calculated for each file, with the results summarized in Sections 2.3–2.8.

2.2. Practical measures

Determining the effect of compression noise on the descriptors requires not only objective measures, but also analysis of the performance in a practical searching scheme. To this end, a simple searching scheme was utilized that attempts to locate a specific frame in the compressed file, using the descriptors generated for the target frame from uncompressed data. This is a simplified version of search/retrieval schemes outlined in [8, 9]. The searching scheme selects the frame in the compressed file that minimizes the mean squared error (MSE) defined as:

$$\text{MSE} = \frac{1}{r}\sum_{n=1}^{r}(x_n-\overline{x_n})^2, \qquad (2)$$

where x_n, $\overline{x_n}$ and r are as defined for Eq. (1). It should be noted that when the descriptors are generated from original uncompressed data, the above scheme returns the correct frame for all individual (and combinations of) descriptors.

2.3. Results for AP

The AP describes the instantaneous power of each input frame [3]. AP consists of only a single scalar value per frame. The average SegSNR values for the instrumental files, combinational files and the overall average for all files, for each compression scheme, are shown in Table 1.

Table 1
SegSNR values for AP [dB]

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	33.84	36.75	37.68	36.12	38.53
Inst.	40.13	43.2	44.46	42.3	44.78
Average	37.33	40.33	41.44	39.55	42

The results in Table 1 indicate that the AP descriptor achieves a very high SegSNR value of over 40 dB for all of the compression schemes. This high value would indicate that compression noise has little effect on the value of the descriptor.

It is interesting to note that there is approximately a 5–7 dB improvement in SegSNR for the instrumental files when compared to the combination files. This distinct difference is most likely attributed to the increased masking effect present in the spectrally rich combinational files. This increased masking allows the compression schemes to remove more redundant information from the combinational files than the instrumental files, and thus, the objective difference between the compressed and original files is greater for the combinational files.

The average search results achieved when using only the AP to identify the target frame are shown in Table 2. The results in Table 2 indicate the percentage of incorrect frames identified for the first 5 instrumental and combination files.

Table 2
Percentage incorrect frames for AP search

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	84.3	84.1	86.8	80.6	75.7
Inst.	85.2	84.5	85.2	84.3	81.1
Average	84.75	84.3	86	82.45	78.4

The results in Table 2 indicate that despite the very high SegSNR values reported in Table 1, the AP descriptor is a very unreliable search mechanism. This inability to adequately match the target frame is due to both the AP descriptor having very similar values across adjacent frames and the fact that the logarithmic amplitude quantisation is employed in most audio encoders.

2.4. Results for AW

The AW descriptor provides a low resolution representation of the time domain envelope [3]. It consists of 2 scalar values (max and min) per frame. The average SegSNR values for the same files used in Table 1 are shown in Table 3.

Table 3
SegSNR values for AW [dB]

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	39.51	39.57	38.83	41.49	44.36
Inst.	45.28	46.15	46.55	47	49.83
Average	42.71	43.22	43.12	44.55	47.4

The results in Table 3 indicate that whilst the AW descriptor has slightly lower SegSNR values than those for AP shown in Table 1, the values are still very high. The worst case is 33.84 dB for MP3 using the combination files.
The average search results achieved using the AW descriptor for the same files used in Table 2 are shown in Table 4.

Table 4
Percentage incorrect frames for AW search

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	50.4	40.7	35.1	40.5	30.8
Inst.	62.1	52.1	47.5	55.1	46.3
Average	56.25	46.4	41.3	47.8	38.55

The results presented in Table 4 indicate that despite the AW exhibiting lower SegSNR values than the AP descriptor, it provides a more reliable searching mechanism. This improvement is due to the fact that the AW descriptor has two values for each frame and the probability of two frames having very similar descriptors is reduced. However, the AW descriptor still finds incorrect frames on approximately 40 to 60 percent of occasions.

2.5. Results for ASC

The ASC descriptor represents the center of gravity of the frequency spectrum [3]. The descriptor is a single scalar value per frame that indicates the octave shift from 1 kHz of the centroid value. The average SegSNR values for the same files used in Table 1, are shown in Table 5.
The results in Table 5 indicate that in an objective sense, compression has very little effect on the ASC descriptors. This result is clearly evidenced by the smallest value for SegSNR being in excess of 40 dB.

Table 5
SegSNR values for ASC [dB]

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	41.09	44.06	44.29	44	46.81
Inst.	47.46	50.64	51.96	50.79	53.51
Average	44.63	47.72	48.55	47.77	50.53

The average search results achieved using the ASC descriptor for the same files used in Table 2 are shown in Table 6.

Table 6
Percentage incorrect frames for ASC search

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	84.6	77.6	76.4	78.2	77.6
Inst.	88.5	87.4	85.9	88.1	87.4
Average	86.55	82.5	81.15	83.15	82.5

The results in Table 6 indicate that due to the relative stability of the centroid value across frames, the minor effects of compression evident in Table 5 cause the ASC descriptor to be unsuitable for frame identification in compressed environments.

2.6. Results for ASS

The ASS descriptor describes the RMS deviation from the centroid value (ASC) for a given frame [3]. The descriptor consists of a single scalar value per frame that represents the octave spread from the ASC value. The average SegSNR values for the same files used in Table 1 are shown in Table 7 and the average search results achieved using the ASS descriptor for the same files used in Table 2, are shown in Table 8.

Table 7
SegSNR values for ASS [dB]

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	45.47	47.17	47.11	48.21	51.06
Inst.	49.01	50.76	52.09	52.82	55.8
Average	47.43	49.15	49.87	50.77	53.69

As for the ASC descriptor, despite achieving high SegSNR values (as shown in Table 7) the ASS provides a very unreliable search mechanism in compressed environments.

Table 8
Percentage incorrect frames for ASS search

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	88.6	79.2	87.5	83.6	79.2
Inst.	92.5	92.8	92.1	88.1	87.5
Average	90.55	86	89.8	85.85	83.35

The descriptor often finds the incorrect frame in over 90% of instances, as shown in Table 8. This poor search performance is again attributed to the fact that the ASS values may vary only marginally between frames, and hence, the relatively small amount of compression noise introduced is sufficient to cause ambiguity when searching for absolute matches to the ASS values.

2.7. Results for ASE

The ASE descriptor provides a representation of the power spectrum for each frame of the audio file. The descriptor consists of a vector of values for each frame, with each vector component representing the magnitude of a particular frequency band. The number of frequency bands (and hence the length of the ASE descriptor) is variable according to a predetermined set of user parameters. These parameters include loEdge, hiEdge and resolution [3]; where loEdge represents the lowest edge (frequency) of the frequency bands, hiEdge represents the highest edge (frequency) of the frequency bands and resolution defines the width of the frequency bands (in octaves with respect to 1 kHz) between loEdge and hiEdge. The ASE also contains two additional values representing 0 Hz-loEdge and hiEdge-Sampling_freq/2.

An important note in the standard [3], is that for fine resolutions (i.e. $< \frac{1}{4}$ octave) the window length (length of frame) restricts the minimum value for loEdge such that at least 1 FFT frequency coefficient is present in each band. The result of this restriction is that for resolutions less than $\frac{1}{4}$ octave, the resultant ASE descriptor becomes biased. This bias is due to the fact that the value representing 0 Hz-loEdge has many FFT coefficients lumped into it, and thus, becomes very large, whilst the neighbouring bands contain only a single coefficient. The net result of this effect is that resolutions less than $\frac{1}{4}$ octave are not searching or identifying complex audio signals that have significant low frequency content (such as speech). To alleviate this problem the ASE descriptor generated for this work used a resolution of $\frac{1}{4}$ octave, which produced 32 frequency bands for each frame.

The average SegSNR values for the same files used in Table 1 are shown in Table 9.

The values in Table 9 indicate that the ASE descriptor produces lower objective results in the presence of compression than the other spectral descriptors, ASS and ASC.

Table 9
SegSNR values for ASE [dB]

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	31.33	33.85	37.49	33.88	36.63
Inst.	38.37	42.17	43.66	40.55	43.32
Average	35.24	39.05	40.36	37.58	40.35

This should be expected as the ASE produces much finer resolution than those other descriptors, and hence, the effects of removing masked components in the compression scheme produces more visible objective distortions. It is also clearly evident that as the bit rate of the compression schemes increases, the SegSNR also increases. This effect is due to the compression schemes using the additional bits available to better represent the spectral envelope of the signal.

The average search results achieved using the ASE descriptor for the same files used in Table 2 are shown in Table 10.

Table 10
Percentage incorrect frames for ASE search

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	0.15	0.15	0.18	0.18	0.18
Inst.	2.5	0.8	1	2.5	1.9
Average	1.325	0.475	0.59	1.34	1.04

The results in Table 10 indicate that the ASE produces fairly reliable search results for all compression schemes. This is obviously due to the fine resolution present in the descriptor and thus the likelihood of two similar frames existing (even in the presence of compression noise) is lower than for the more generic descriptors. However, the search still fails approximately 2% of the time for the instrumental files.

2.8. Results for combined descriptor searches

To ascertain if improved search results could be achieved by combining multiple descriptors together into meta-descriptors, we formed two meta-descriptors. The first of these meta-descriptors combined all five of the specified descriptors together and the second combined ASC and ASS to produce a compact representation of frequency content. The search results for these two meta-descriptors using the same files used in Table 2 are shown in Tables 11 and 12 respectively.

Comparing the results in Table 11 to those for the ASE only in Table 10 indicates that including the additional descriptors into the search actually degrades the searching

Table 11
Percentage incorrect frames for meta-descriptor using all of the original descriptors

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	1.8	0.74	0.3	0.5	0.1
Inst.	16.3	8.7	4.8	7.8	4.4
Average	9.05	4.72	2.55	4.15	2.25

Table 12
Percentage incorrect frames for the ASS/ASC meta-descriptor

	MP3			WMA	
Bit rate [kbit/s]	128	160	192	128	160
Comb.	24.5	16.2	13.2	13.5	8.8
Inst.	44.7	37	31.2	34.2	24.1
Average	34.6	26.6	22.2	23.85	16.45

performance. This result indicates, that because the additional descriptors may have larger absolute values than the individual ASE components, the ambiguity introduced into these additional descriptors by compression is sufficient to degrade the unweighted search performance used in Eq. (2). Better results may be achieved by introducing a weighting function into Eq. (2).

When compared to the results for the descriptors ASC and ASS in isolation (Tables 6 and 8 respectively), the results in Table 12 indicate that combining the descriptors together reduces the incorrect results by between 50 and 66%. This improved result is encouraging and supports the finding that implementing more sophisticated search (weighted) mechanisms for meta-descriptors may improve performance in compressed environments.

2.9. Summary of results

The results presented for all descriptors indicate that an objective measure of the effects of compression gives little indication of the actual performance degradation in a practical search situation. The results indicate that compression noise is a significant problem for practical applications of the low level audio descriptors.

The primary reason for the modification of the MPEG-7 audio descriptors when using compressed input data, is the redundancy removal performed by the compression algorithms. For audio compression, redundancy is removed (compression achieved) by hiding quantisation noise in sections of the spectrum that are masked (inaudible to the human ear). This procedure may result in the actual spectral shape being significantly modified by the compression algorithm. As many of the frame based MPEG-7 low level audio descriptors are based on representations of the spectrum, this modification of the spectral shape directly affects the values of the MPEG-7 descriptors calculated from the compressed input stream. Also, as most audio compression schemes quantise amplitude values on a logarithmic scale that replicates the response of the human ear, the linear representation of the audio power provided by the AP descriptor suffers from quite severe quantisation noise for large amplitudes. These effects in combination with the fact that many of the descriptors analysed vary only marginally between frames, cause the presented MPEG-7 audio descriptors to produce very unreliable searching parameters.

It should be noted that on average, across all of the results presented in Sections 2.3–2.8, the WMA coder had less effect on the descriptors than the MP3 encoder. This result is most likely attributed to the fact that the WMA algorithm is significantly more modern than MP3, and thus, exploits more sophisticated signal processing and psycho-acoustic techniques.

3. Methods for improving performance in compressed environments

The methods examined for improving the search performance in compressed environments can be grouped into two categories: 1) adaptive signal processing (ASP) and 2) machine learning.

3.1. ASP techniques

A number of signal processing techniques were examined. The first of these was a simple fixed predictor that attempts to predict the original descriptors from the compressed descriptors. The predictor coefficient was calculated as:

$$a = \frac{\sum_{n=0}^{N-1} x(n)\overline{x(n)}}{\sum_{n=0}^{N-1} x(n)^2}, \tag{3}$$

where a is the predictor coefficient and x_n, $\overline{x_n}$ and N are as defined for Eq. (1).

The more sophisticated technique of vector linear prediction (VLP) [10] was also employed. This technique uses a matrix of coefficients to predict the original frame of descriptors from the compressed frame. The VLP is calculated as [10]:

$$\text{VLP} = C_{01}(C_{00})^{-1},$$

where:

$$\begin{aligned} C_{01} &= \tfrac{1}{N}\sum_{n=0}^{N-1} x(n)\overline{x(n)}^T, \\ C_{00} &= \tfrac{1}{N}\sum_{n=0}^{N-1} x(n)x(n)^T. \end{aligned} \tag{4}$$

Table 13
Percentage incorrect frames for the prediction schemes, compared to no prediction

Descriptor	AP	AW	ASC	ASS	ASE	ASS/ASC	ALL
Non-predicted	92.7	72.2	89.7	90.3	0	37.5	4.1
VLP	93.1	87.4	96.8	92.9	0	63.1	26.2
Fixed pred	92.3	71.8	91.7	89.3	0	40	4.5

Table 14
SegSNR for prediction schemes, compared to no prediction

Descriptor	AP	AW	ASC	ASS	ASE	ASS/ASC	ALL
Non-predicted	39.6	33.5	36.9	46.2	30.2	39.5	38.2
VLP	35	28.1	31.7	38.6	25.7	33.2	32.5
Fixed pred	39.3	33.6	36.6	46	30.2	39.5	38.2

To augment the performance of Eqs. (3) and (4), both methods were utilized in conjunction with switched adaptive prediction [10]; a method that has been shown to operate well for quantisation of speech spectral envelop parameters [10]. This method trains classifiers for the compressed vectors offline using a vector quantiser (VQ) [11]. Separate predictors, Eqs. (3) and (4), are then trained for each classification. In the prediction process the input vector of descriptors is firstly classified (using the VQ) and then the original descriptors are estimated from the compressed descriptors, using the predictor that corresponds to the classification. For this paper the vectors were classified into 4 classes.

The search results using the proposed predictors in conjunction with the MP3 coder operating at 128 kbit/s are shown in Table 13 and the SegSNR results for the same test files are shown in Table 14. The results in Tables 13 and 14 represent utilizing the prediction schemes on test vectors not included in the training set. If vectors from within the training set were used, better results are achieved however, this would require training predictors for specific files. We are searching for a more generic solution.

The results in Tables 13 and 14 indicate that of the two prediction schemes, the simple fixed predictor achieved the best performance. As the VLP relies upon strong inter and intra-frame correlation, this result indicates that the inter and intra-frame correlation between the descriptor vectors is quite low. One could infer that the entropy of the descriptor vectors actually increases as the descriptors are grouped together, which supports the results of Section 2.8 where the search performance deteriorated when additional descriptors were added to the ASE to produce a meta-descriptor.

Whilst the fixed predictor achieved the best results of the two predictors, it achieved little or no improvement over using the non-predicted vectors. This result indicates that either the compression noise is non-stationary and thus not predictable, or alternately, that more sophisticated prediction techniques (non-linear, multi-tap) or sophisticated heuristic algorithms are required.

3.2. Machine learning techniques

A number of alternative approaches, emanating from the field of machine learning (ML) have been considered. Data mining is, in one aspect, a specialized application of machine learning, and as it essentially embodies much of the same key features, they are considered here as the same. The practical focus of these endeavours is to detect useful, but often-implicit patterns in empirical data, and to further construct descriptive models of these. Whilst there are several plausible techniques that could be considered in improving the performance of the low level descriptors, the impact of each is governed primarily by how the various aspects and concepts are both represented and modelled. A useful overall guide to this technology is found in [12].

The approach taken in this paper is to learn rule-based predictive models, based on compression-affected descriptors that can essentially predict the former compression free state. An overview of this process is shown in Fig. 1.

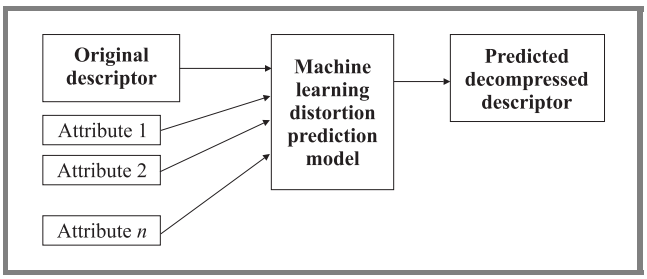


Fig. 1. Overview of ML techniques.

Several trials were undertaken to develop a range of non-parametric regression models that would map these

descriptors. As such, all of the available descriptors on one side, say those derived from the compressed file, are used to form a prediction model for each of the non-effected descriptors. Thus this technique can be viewed as being closest to that of the VQ method of the previous section. However, despite the availability of employing all descriptors to build a suitable predictor, the machine learning algorithm, Cubist [13], used only a small number of these in each case (selected via various entropy-based metrics).

The results shown in Table 15 compare the search performance for the meta-descriptor comprising all of the descriptors when using, ML prediction, the ASP predictors of Section 3.1 and no prediction. The meta-descriptor was selected to evaluate the performance of the ML predictor, as the 37 dimensional input and output vectors present the most challenging scenario. As can be seen in Table 15, the ML approach produces the most reliable search performance. The ML predictor significantly improves the missed-framing rate by 0.9% compared to the simple non-predicted method. This actually represents a 22% improvement in searching performance.

Table 15
Comparative improved error rate in matching frames

Descriptor	ALL
Non-predicted	4.1
VLP	26.2
Fixed pred	4.5
ML: Cubist	3.2

This improvement is readily explained by the nature of the machine learning approach. As such algorithms seek to learn specific concepts, they create and grow a model to fit or explain the training data. In contrast, traditional signal processing techniques often apply the reverse situation, where a fixed model is employed and data fitted to the model structure. Under the ML regime, several virtual contexts are formed and utilized, compared to the VQ predictor of the last section that used 4.

The type of ML algorithm used here, produces what could be best described as a combined decision tree with a series of linear regression models, these reside at the leaf nodes of the decision tree. The internal nodes and branch structure assist in segmenting the descriptor data into several regions of similar magnitudes (at the leaves) where afterwards, suitable linear regression representations finally model any residual variance within each leaf.

Whilst the ML prediction provides a significant improvement in search performance, the reliability of the search is still not sufficiently high for practical applications (a failure rate of approximately 0.1% may be deemed necessary). To improve the overall searching reliability a more sophisticated searching algorithm and ML predictor are proposed and analysed in Section 3.3.

3.3. Improved search algorithm and ML predictor

This section proposes an extension to the simple MSE searching method proposed in Section 2.2. The new method is still based on MSE minimization but now employs the calculation across the previous, current and future (PCF) frames. Incorporating a larger time scale (adjacent frames) into the calculation allows more accurate searching, as the evolution, pattern or trend is identified as opposed to the previous method of matching only a single value. An example of this improved matching performance is illustrated in Fig. 2. If we consider the nine samples illustrated in Fig. 2 as separate individual samples, and, if the target sample is that second from the left, it is easy to visualise how a number of the other samples could be incorrectly located in the presence of ambiguous noise. However, when groups of three consecutive samples are used, it is easy to identify that the smooth curve represented by the left-hand group is easily distinguished from the other groups.

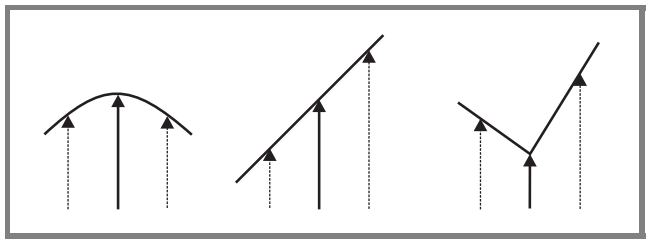


Fig. 2. Examples trends searched for in descriptor data using the PCF MSE method.

The PCF method essentially provides a localised context that increases the frame's resolution by a factor of three. Therefore singular dimensioned descriptors (AP, ASC, ASS) are now 3 dimensions wide, AW is 6 dimensions wide and the ASE descriptor now has an effective resolution of 96. The trade off of the proposed PCF scheme is that the temporal resolution of the search is reduced. By extending the MSE method, the algorithm remains consistent with those discussed in [8, 9].

The frame that minimizes Eq. (5) is considered the closest matched frame for the PCF:

$$\mathrm{PCF}_{\mathrm{MSE}} = \frac{1}{3r}\left(\begin{array}{l}\sum_{n=1}^{r}(x_{pn}-\overline{x}_{pn})^2+\sum_{n=1}^{r}(x_{cn}-\overline{x}_{cn})^2+\\ \sum_{n=1}^{r}(x_{fn}-\overline{x}_{fn})^2\end{array}\right), \tag{5}$$

where x_n, $\overline{x_n}$ and r are as defined for Eq. 2 and the additional p, c and f stand for the previous, current and future frames respectively. Table 16 displays an average representation for the percentage of unsuccessfully matched frames using MP3 – 128 kbit/s encoding.

Comparing the results in Table 16 to Tables 2, 4, 6, 8 and 10, indicates that adding the previous and future frames into the search criteria, results in a dramatic improvement

for all descriptors. Even the least reliable descriptor for searching compressed files, ASS, has its incorrect matches reduced from 90.6% to 15.7%. The ASE descriptor now produces incorrect matches on 0.4% of occasions. This is approximately a third of the error rate of the simple search and is approaching our target of 0.1%. However, the trade off for the improved performance is increased complexity and reduced temporal resolution.

Table 16
Percentage incorrect frames for the PCF search

	AP	AW	ASC	ASS	ASE
Inst.	10.3	4.8	10.4	13.9	0.4
Comb.	9.3	4.8	13.3	17.4	0.4
Average	9.8	4.8	11.9	15.7	0.4

We subsequently modified our ML algorithm to additionally incorporate the past and future frames into the range of attributes available for building the various prediction models. This resulted in a modest improvement in search reliability of approximately 0.3% for the single value descriptor AP. Whilst this improvement in performance may be able to be translated to the ASE descriptor, this was not tested due to the dramatic increase in complexity and storage required to build predictors for each of the 32 ASE elements.

4. Conclusion

An extended analysis of lossy compression effects on the MPEG-7 low level audio descriptors was conducted. This analysis exposed a distinct degradation in the performance of simple practical searching schemes when lossy compression has been used to modify the MM files. Methods to reduce the effects of compression in practical searching were then investigated. Of the proposed methods, prediction schemes based on machine learning were found to offer the greatest reduction in distortion. However, these prediction schemes did not completely nullify the effect of compression. A more sophisticated search mechanism that employs multiple frames in its calculation was then proposed. Generally this scheme significantly improved the reliability of searching, however, even the best performing descriptor combination still did not provide adequate performance.

A possible method for improving the search performance of the MPEG-7 audio descriptors with compressed input data would be to develop new compression algorithms that maintain the integrity of at least some of the descriptors. However, due to the prevalence of existing audio compression schemes (such as MP3) it is highly unlikely that such a new compression algorithm would be widely adopted. The authors instead propose that future work should focus on developing better search mechanisms such as those based on maximum likelihood. In addition and more importantly, new audio descriptors that incorporate the characteristics of existing compression schemes into their structure should be developed.

References

[1] ISO/IEC JTC1/SC29/WG11/N4031, "Overview of the MPEG-7 Standard (version 5)", International Organisation for Standardisation, Singapore, March 2001.

[2] S. Chang, T. Sikora, and A. Puri, "Overview of the MPEG-7 standard", *IEEE Trans. Circ. Syst. Video Technol.*, vol. 11, no. 6, pp. 688–695, 2001.

[3] ISO/IEC FDIS 15938-4, "Information technology multimedia content description interface", Part 4: "Audio", International Organisation for Standardisation, Singapore, March 2001.

[4] M. Casey, "MPEG-7 sound recognition tools", *IEEE Trans. Circ. Syst. Video Technol.*, vol. 11, no. 6, pp. 737–747, 2001.

[5] M. Hu and Y. Jian, "MD2L: content description of multimedia documents for efficient process and search/retrieval", in *Proc. IEEE Forum Res. & Technol. Adv. Dig. Libr.*, 1999, pp. 200–213.

[6] ISO/IEC JTCI/SC29, "Information technology-coding of motion pictures and associated audio for digital storage media upto about 1.5 Mbit/s – IS 11172", Part 3: "Audio", 1992.

[7] Microsoft, "Windows media encoder", July 2002, available at http://www.microsoft.com/windows/windowsmedia/WM7/encoder/whitepaper.asp

[8] S. Quackenbush and A. Lindsay, "Overview of MPEG-7 audio", *IEEE Trans. Circ. Syst. Video Technol.*, vol. 11, no. 6, pp. 725–729, 2001.

[9] E. Allamanche, "Robust matching of audio signals using spectral flatness features", in *IEEE Worksh. Appl. Signal Proc. Audio Acoust.*, 2001, pp. 127–130.

[10] M. Yong, G. Davidson, and A. Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction", in *Proc. ICASSP*, 1988, vol. 1, pp. 402–405.

[11] A. Gersho and R. M. Gray, *Vector Quantisation and Signal Compression*. Kluwer, 1992.

[12] I. H. Witten and E. Frank, *Data Mining Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2000.

[13] Cubist (version 1.13), Rulequest Research, www.rulequest.com

Jason Lukasiak is currently a lecturer in the School of Electrical, Computer and Telecommunications Engineering at the University of Wollongong. He received his B.E. (Hons.) and Ph.D. from the University of Wollongong in 1998 and 2002, respectively. Prior to commencing his Ph.D. studies in 1999, he was employed by BHP steel from 1987. During his employement with BHP his positions ranged from computer network technician to electrical project engineer. The topic of his Ph.D. thesis was "Techniques for low-rate scalable speech compression" and other current research interests include description and adaptation of multimedia objects, linked audio-visual modeling and transcoding of speech signals. In relation to the

MPEG-7 work detailed in this paper, a web based implementation allowing calculation of the MPEG-7 low level audio descriptors for any input audio file has been completed.

e-mail: jasonl@elec.uow.edu.au
School of Electrical, Computer
and Telecommunications Engineering
University of Wollongong
Wollongong, NSW 2522, Australia

David Stirling has developed considerable expertise in data analysis and knowledge management with skills in problem solving, statistical methods, visualization, pattern recognition, data fusion and reduction, and programming and is widely experienced in applying these to organizations requiring solutions to complex concept relationship problems. He has applied machine learning and data mining techniques of specialised classifier designs for noisy multivariate data to medical research, exploration geoscience, and financial markets, as well as to industrial primary operations. Before setting up his own consultancy company in 1998, Dr. Stirling was a principal research scientist with BHP research (and before that, with John Lysaght) and has over 15 years experience working in the Port Kembla steelworks. He has recently taken up a position as senior lecturer in the School of Electrical, Computer and Telecommunications Engineering at the University of Wollongong.

e-mail: stirling@uow.edu.au
School of Electrical, Computer
and Telecommunications Engineering
University of Wollongong
Wollongong, NSW 2522, Australia

Shane Perrow graduated from the University of Wollongong with a B.E. (Hons.) in December 2002. The work contributed by Shane to the paper formed the majority of his final year honours thesis topic.

e-mail: Perrow@ali.com.au
School of Electrical, Computer
and Telecommunications Engineering
University of Wollongong
Wollongong, NSW 2522, Australia

Nick Harders is a final year undergraduate student in maths/electrical engineering at the University of Wollongong. He completed his contribution to the paper whilst working on a summer university scholarship program.

e-mail: nharders@bigpond.com
School of Electrical, Computer
and Telecommunications Engineering
University of Wollongong
Wollongong, NSW 2522, Australia

Paper

Fully spatial and SNR scalable, SPIHT-based image coding for transmission over heterogenous networks

Habibollah Danyali and Alfred Mertins

Abstract — This paper presents a fully scalable image coding scheme based on the set partitioning in hierarchical trees (SPIHT) algorithm. The proposed algorithm, called fully scalable SPIHT (FS-SPIHT), adds the spatial scalability feature to the SPIHT algorithm. It provides this new functionality without sacrificing other important features of the original SPIHT bitstream such as: compression efficiency, full embeddedness and rate scalability. The flexible output bitstream of the FS-SPIHT encoder which consists of a set of embedded parts related to different resolutions and quality levels can be easily adapted (reordered) to given bandwidth and resolution requirements by a simple parser without decoding the bitstream. FS-SPIHT is a very good candidate for image communication over heterogenous networks which requires high degree of scalability from image coding systems.

Keywords — wavelet image coding, scalability, SPIHT, progressive transmission, multiresolution.

1. Introduction

The main objective of traditional image coding systems is optimizing image quality at given bit rate. Due to the explosive growth of the Internet and networking technology, nowadays a huge number of users with different network access bandwidth and processing capabilities can easily exchange data. For transmission of visual data on such a heterogenous network, efficient compression itself is not sufficient. There is an increasing demand for scalability to optimally service each user according to the available bandwidth and computing capabilities. A scalable image coder generates a bitstream which consists of a set of embedded parts that offer increasingly better signal-to-noise ratio (SNR) or/and greater spatial resolution. Different parts of this bitstream can be selected and decoded by a scalable decoder to meet certain requirements. In the case of an entirely scalable bitstream, different types of decoders with different complexity and access bandwidth can coexist.

Over the past decade, wavelet-based image compression schemes have become increasingly important and gained widespread acceptance. An example is the new JPEG2000 still image compression standard [1,2]. Because of their inherent multiresolution signal representation, wavelet-based coding schemes have the potential to support both SNR and spatial scalability. Shapiro [3] pioneered embedded wavelet-based image coding by introducing the embedded zerotree wavelet (EZW) coding scheme based on the idea of grouping spatially related coefficients at different scales to trees and efficiently predicting zero coefficients across scales. The scheme provides an output bitstream that consists of data units ordered by their importance and that can be truncated at any point without degradation of the coding efficiency. Many researchers have since worked on variations of the original zerotree method [4–10]. An important development of EZW, called set partitioning in hierarchical trees algorithm by Said and Pearlman [7] is one of the best performing wavelet-based image compression algorithms. This coder uses the spatial orientation trees shown in Fig. 1 and partitions them as needed to sort wavelet coefficients according to magnitude. Further improvements of SPIHT have been published in [11–16]. Although almost all of the state-of-the-art zerotree-based image compression methods are SNR scalable and pro-

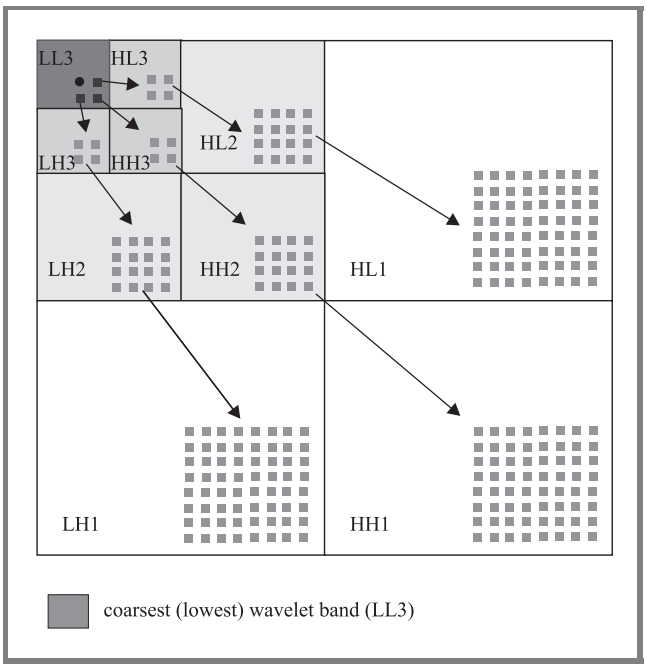


Fig. 1. Orientation of trees across wavelet bands.

vide bit streams for progressive (by quality) image compression, they do not explicitly support spatial scalability and do not provide a bitstream which can be adapted easily according to the type of scalability desired by the decoder.

An improvement of the EZW algorithm called predictive EZW (PEZW) was reported in [8]. The PEZW improves the EZW through better context modelling for arithmetic coding and an improved symbol set for zerotree encoding. It also uses proper syntax and markers for the compressed bitstream to allow extracting bitstreams that represent various qualities and resolutions of the original image. However the decoder needs some additional side information to decode these bitstreams. Tham et al. [17] introduced a new zerotree structure called tri-zerotree and used a layered coding strategy with the concept of embedded resolution block coding to achieve a high degree of scalability for video coding. A spatially scalable video coding scheme based on SPIHT was reported by Kim et al. in [13]. Their coder produces a two-layer bitstream; the first layer is used for low resolution, and the second one adds the extra information required for high resolution. Although the first layer of this method is rate scalable, the bitstream is not fully embedded for high resolution. Moreover, it is not possible to easily reorder the encoded bitstream to arbitrary spatial resolutions and SNR's. However, the ability to reorder the bitstream is an important requirement for access to images through heterogeneous networks with a large variation in bandwidth and user-device capabilities.

In this paper, a fully scalable image coding scheme based on the SPIHT algorithm is presented. We modify the SPIHT algorithm to support both spatial and SNR scalability features. The encoder creates a bitstream that can be easily parsed to achieve different levels of resolution or/and quality requested by the decoder. A distinct feature of the presented coder is that the reordered bitstreams for different spatial resolutions, which are obtained after parsing the main bitstream, are fully embedded (SNR scalable) and can be truncated at any point to obtain the best reconstructed image at the desired spatial resolution and bit rate. In other words, our modified SPIHT algorithm provides spatial scalability without sacrificing SNR scalability in any way.

The rest of this paper is organized as follow. The next section, Section 2, describes the FS-SPIHT algorithm. The bitstream formation and parsing are explained in Section 3. Section 4 shows some results on the rate-distortion performance for our codec and provides comparisons with the SPIHT coder. Finally some conclusions are presented in Section 5.

2. Fully scalable SPIHT (FS-SPIHT)

In this section, we first give a brief description of the SPIHT algorithm, then explain our modification of SPIHT (FS-SPIHT) for fully supporting SNR and spatial scalabilities.

The SPIHT algorithm consists of three stages: initialization, sorting and refinement. It sorts the wavelet coefficients in three ordered lists: the list of insignificant sets (LIS), the list of insignificant pixels (LIP), and the list of significant pixels (LSP). At the initialization stage the SPIHT algorithm first defines a start threshold due to the maximum value in the wavelet coefficients pyramid, then sets the LSP as an empty list and puts the coordinates of all coefficients in the coarsest level of the wavelet pyramid (i.e. the lowest frequency band; LL band) into the LIP and those which have descendants also into the LIS. Figure 1 shows the parent-child relationships used within the wavelet tree. The pixels in the coarsest level of the pyramid are grouped into blocks of 2×2 adjacent pixels, and in each block one of them has no descendants. In the sorting pass, the algorithm first sorts the elements of the LIP and then the sets with roots in the LIS. For each pixel in the LIP it performs a significance test against the current threshold and outputs the test result to the output bitstream. All test results are encoded as either 0 or 1, depending on the test outcome, so that the SPIHT algorithm directly produces a binary bitstream. If a coefficient is significant, its sign is coded and then its coordinate is moved to the LSP. During the sorting pass of LIS, the SPIHT encoder carries out the significance test for each set in the LIS and outputs the significance information. If a set is significant, it is partitioned into its offspring and leaves. Sorting and partitioning are carried out until all significant coefficients have been found and stored in the LSP. After the sorting pass for all elements in the LIP and LIS, SPIHT does a refinement pass with the current threshold for all entries in the LSP, except those which have been moved to the LSP during the last sorting pass. Then the current threshold is divided by two and the sorting and refinement stages are continued until a predefined bit-budget is exhausted.

In general, an N level wavelet decomposition allows at most $N+1$ levels of spatial resolution. To distinguish between different resolution levels, we denote the lowest spatial resolution level as level $N+1$. The full image then becomes resolution level 1. The three subbands that need to be added to increase the resolution from level k to level $k-1$ are referred to as level $k-1$ resolution subbands. An algorithm that provides full spatial scalability would encode the different resolution levels separately, allowing a parser or the decoder to directly access the data needed to reconstruct with a desired spatial resolution. The original SPIHT algorithm, however, encodes the entire wavelet tree in a bitplane by bitplane manner and produces a bitstream that contains the information about the different spatial resolutions in no particular order.

In [18] we modified SPIHT to support both spatial and SNR scalability by adding a new list to the SPIHT lists and modifying the SPIHT sorting pass. The FS-SPIHT algorithm proposed in this paper solves the spatial scalability problem through the introduction of multiple resolution-dependent lists and a resolution-dependent sorting pass. For each spatial resolution level we define a set of LIP, LSP

and LIS lists, therefore we have LIP_k, LSP_k, and LIS_k for $k = k_{max}, k_{max}-1, \ldots, 1$ where k_{max} is the maximum number of spatial resolution levels supported by the encoder. In each bitplane, the FS-SPIHT coder starts encoding from the maximum resolution level (k_{max}) and proceeds to the lowest level (level 1). For the resolution-dependent sorting pass of the lists that belong to level k, the algorithm first does the sorting pass for the coefficients in the LIP_k in the same way as SPIHT and then processes the LIS_k list. During processing the LIS_k, sets that lie outside the resolution level k are moved to the LIS_{k-1}. After the algorithm has finished the sorting and refinement passes for level k it will do the same procedure for the lists related to level $k-1$. According to the magnitude of the coefficients in the wavelet pyramid, coding of higher resolution bands usually starts from lower bitplanes. The total number of bits belonging to a particular bitplane is the same for SPIHT and FS-SPIHT, but FS-SPIHT arranges them according to their spatial resolution dependency.

In the following we first define the sets and symbols required by FS-SPIHT. These are the same as for the original SPIHT algorithm. Then we list the entire FS-SPIHT coding algorithm.

Definitions:

- $c(i,j)$: wavelet transformed coefficient at coordinate (i,j).
- $O(i,j)$: set of coordinates of all offspring of node (i,j).
- $D(i,j)$: set of coordinates of all descendants of node (i,j).
- $L(i,j)$: set of coordinates of all leaves of node (i,j). $L(i,j) = D(i,j) - O(i,j)$.
- H: set of coordinates of all nodes in the coarsest level of wavelet coefficients pyramid.
- $S_n(i,j)$: significance test of a set of coordinates $\{(i,j)\}$ at bitplane level n

$$S_n(i,j) = \begin{cases} 1 & \text{If } \max_{\{(i,j)\}}\{|c(i,j)|\} \geq 2^n \\ 0 & \text{otherwise} \end{cases}$$

- Type A sets: for sets of type A the significance tests are to be applied to all descendants $D(i,j)$.
- Type B sets: for sets of type B the significance tests are to be applied only to the leaves $L(i,j)$.
- n_{max}: maximum bitplane level needed for coding $n_{max} = \lfloor \log_2(\max_{\{(i,j)\}}\{|c(i,j)|\}) \rfloor$
- k_{max}: maximum level of spatial scalability to be supported by the bitstream ($1 \leq k_{max} \leq N+1$).
- β_k: A set of subbands in the decomposed image that belong to spatial resolution level k ($1 \leq k \leq k_{max}$) of the image.

FS-SPIHT coding steps:

1. **Initialization**
   - $n = n_{max}$, and output n;
   - $\mathrm{LSP}_k = \emptyset$, $\forall k, 1 \leq k \leq k_{max}$;
   - $\mathrm{LIP}_k = \begin{cases} \emptyset & \text{for } 1 \leq k < k_{max} \\ \{(i,j)\},\ \forall (i,j) \in H & k = k_{max} \end{cases}$
   - $\mathrm{LIS}_k = \emptyset$, $\forall k, 1 \leq k < k_{max}$;
   - $\mathrm{LIS}_{k_{max}} = \{(i,j)\}$ as type A, $\forall (i,j) \in H$ which have descendants;
   - $k = k_{max}$.
2. **Resolution-dependent sorting pass**
   - SortLIP(n,k);
   - SortLIS(n,k).
3. **Refinement pass**
   - RefineLSP(n,k).
4. **Resolution scale update**
   - if $(k > 1)$
     - $k = k-1$;
     - go to step 2;
   - else, $k = k_{max}$.
5. **Quantization-step update**
   - if $(n > 0)$
     - $n = n-1$;
     - go to step 2;
   - else, end of coding.

Pseudo code:

SortLIP$(k,n)\{$

- for each entry (i,j) in the LIP_k do:
  - output $S_n(i,j)$;
  - if $(S_n(i,j) = 1)$, then move (i,j) to the LSP_k, output the sign of $c(i,j)$;

$\}$

SortLIS$(n,k)\{$

for each entry (i,j) in the LIS_k

- if the entry is of type A
  - if $(\forall(x,y) \in D(i,j) : (x,y) \notin \beta_k)$, then move (i,j) to LIS_{k-1} as type A;
  - else
    - output $S_n(D(i,j))$;
    - if $(S_n(D(i,j)) = 1)$ then for each $(p,q) \in O(i,j)$
      - output $S_n(p,q)$;
      - if $(S_n(p,q) = 1)$, add (p,q) to the LSP_k, output the sign of $c(p,q)$;
      - else, add (p,q) to the end of the LIP_k;
    - if $(L(i,j) \neq \emptyset)$, move (i,j) to the end of the LIS_k as an entry of type B;
    - else, remove entry (i,j) from the LIS_k;
- if the entry is of type B
  - if $(\forall(x,y) \in L(i,j) : (x,y) \notin \beta_k)$, then move (i,j) to LIS_{k-1} as type B;
  - else
    - output $S_n(L(i,j))$;
    - if $(S_n(L(i,j)) = 1)$
      - add each $(p,q) \in O(i,j)$ to the end of the LIS_k as an entry of type A;
      - remove (i,j) from the LIS_k.

$\}$

RefineLSP$(n,k)\{$

- for each entry (i,j) in the LSP_k, except those included in the last sorting pass (i.e. the ones with the same n), output the n^{th} most significant bit of $|c(i,j)|$.

$\}$

Note that the total storage requirement for all lists LIP_k, LSP_k, and LIS_k for $k = k_{max}, k_{max}-1, \ldots, r$ is the same as for the LIS, LIP, and LSP used by the SPIHT algorithm.

To support bitstream parsing by an image server/parser, some markers are required to be put into the bitstream to identify the parts of the bitstream that belong to the different spatial resolution levels and bitplanes. This additional information does not need to be sent to the decoder.

3. Bitstream formation and parsing

Figure 2 shows the bitstream structure generated by the encoder. The bitstream is divided into different parts according to the different bitplanes. Inside each bitplane part, the bits that belong to the different spatial resolution levels are separable. A header at the beginning of the bitstream identifies the number of spatial resolution levels supported by the encoder, as well as information such as the image dimension, number of wavelet decomposition levels, and the maximum quantization level. At the beginning of each bitplane there is an additional header that provides the information required to identify each resolution level.

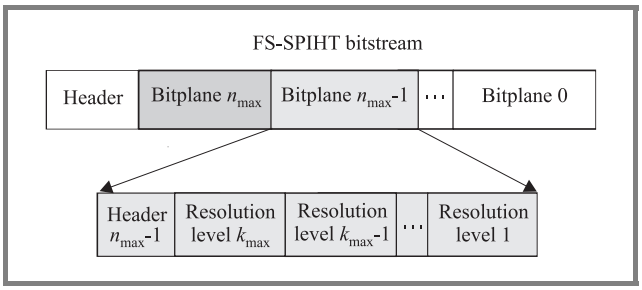


Fig. 2. Structure of FS-SPIHT encoder bitstream which is made up of different parts according to spatial resolution and quality.

A single encoded bitstream for the full-resolution image is stored on an image server. Different users with different requirements send their request to the server and the server or a parser within the network provides them with properly tailored bitstreams that are easily generated by selecting the related parts of the original bitstream and ordering them in such a way that the user requests are fulfilled. Figure 3 illustrates the principle. To carry out the parsing process, the image server or parser does not need to decode any parts of the bitstream.

Figure 4 shows an example of a reordered bitstream for spatial resolution level r. In each bitplane only the parts that belong to the spatial resolution levels greater or equal to the requested level are kept and the other parts are removed. Note that all header information for identifying the individual bitplanes and resolution levels are only used by the image parser and does not need to be sent to the decoder.

The decoder required for decoding of the reordered bitstream follows the encoder with the output command replaced by an input command, similar to the original SPIHT algorithm. It needs to keep track of the various lists (LIS, LIP, LSP) only for resolution levels greater or equal to the required one. It can recover all information for updating the lists during the sorting pass of each quantization level (bitplane) at each spatial resolution level. The only additional information required by the decoder is the maximum number of spatial scalability levels (k_{max}) supported by the encoder.

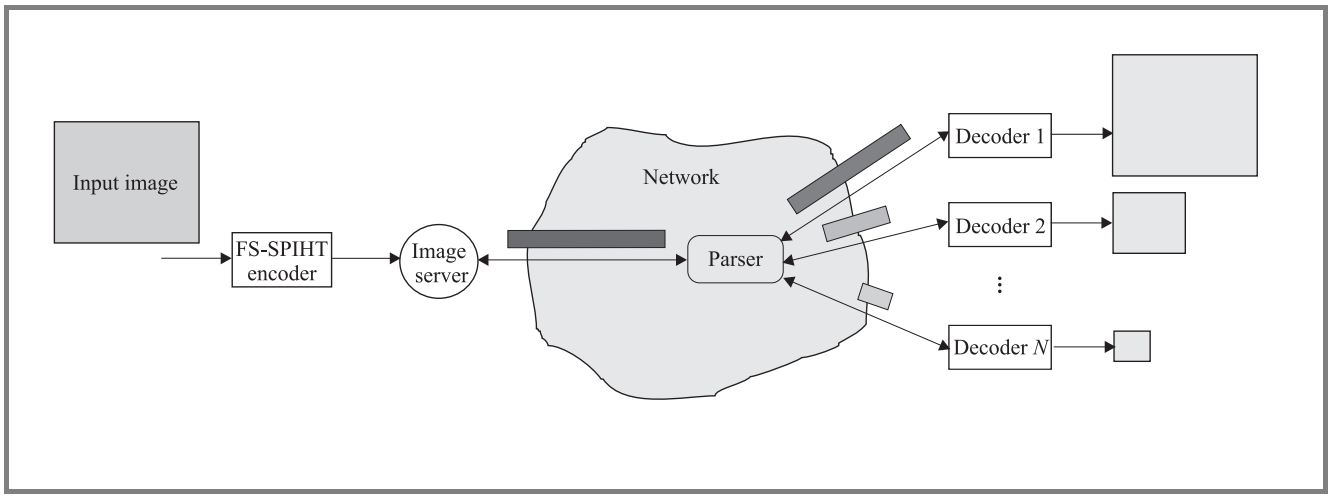


Fig. 3. An example of an image server and a parser in a network for providing various bitstreams for different resolutions or/and quality levels requested by different users.

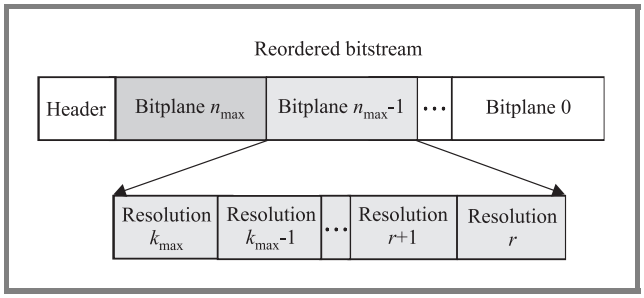


Fig. 4. Reordered FS-SPIHT bitstream for spatial resolution level r decoding.

4. Experimental results

In this section we present some numerical results for the FS-SPIHT algorithm. All results were obtained with 8 bit per pixel (bpp) monochrome images of size 512×512 pixels. We first applied five levels of wavelet decomposition with the 9/7-tap filters of [19] and symmetric extension at the image boundaries. The FS-SPIHT encoder was set to produce a bitstream that supports six levels of spatial scalability.

After encoding, the FS-SPIHT bitstream was fed into a parser to produce progressive (by quality) bitstreams for different spatial resolutions. The bitstreams were decoded with different rates and the fidelity was measured by the peak signal-to-noise ratio (PSNR). The bit rates for all levels were calculated according to the number of pixels in the original full size image.

Figures 5 and 6 compare rate-distortion results of FS-SPIHT and SPIHT at different spatial resolution levels for test images. For spatial resolution level 1, the bitstream needed by the FS-SPIHT decoder can be obtained by simply removing the bitplane headers from the encoder output bitstream. The results clearly show that the FS-SPIHT completely keeps the progressiveness (by SNR) property of the SPIHT algorithm. The small deviation between FS-SPIHT and SPIHT is due to a different order of coefficients within the bitstreams. For resolution levels 2 and 3, the FS-SPIHT decoder obtained the proper bitstreams tailored by the parser for each resolution level while for the SPIHT case we first decoded the whole image at each bit rate and then compared the requested spatial resolutions of the reconstructed and original images. All bits in the reordered FS-SPIHT bitstream for a particular resolution belong only to that resolution, while in the SPIHT bitstream, the bits that belong to the different resolution levels are interwoven. Therefore, as expected, the performance of FS-SPIHT is much better than for SPIHT for resolution levels greater than one. As the resolution level increases, the difference between FS-SPIHT and SPIHT becomes more and more significant. All the results are obtained without extra arithmetic coding of the output bits. As shown in [7], an improved coding performance (about 0.3–0.6 dB) for SPIHT and consequently for FS-SPIHT can be achieved by further compressing the binary bitstreams with an arithmetic coder.

5. Conclusions

We have presented a fully scalable SPIHT algorithm that produces a bitstream which supports spatial scalability and can be used for multiresolution parsing. This bitstream not only has spatial scalability features but also keeps the full SNR embeddedness property for any required resolution level after a simple reordering which can be done in a parser without decoding the bitstream. The embeddedness is so fine granular that almost each additional bit improves the quality, and the bitstream can be stopped at any point to meet a bit budget during the coding or decoding process. The algorithm is extendable for combined SNR, spatial and frame-rate scalable video coding and also for fully scalable

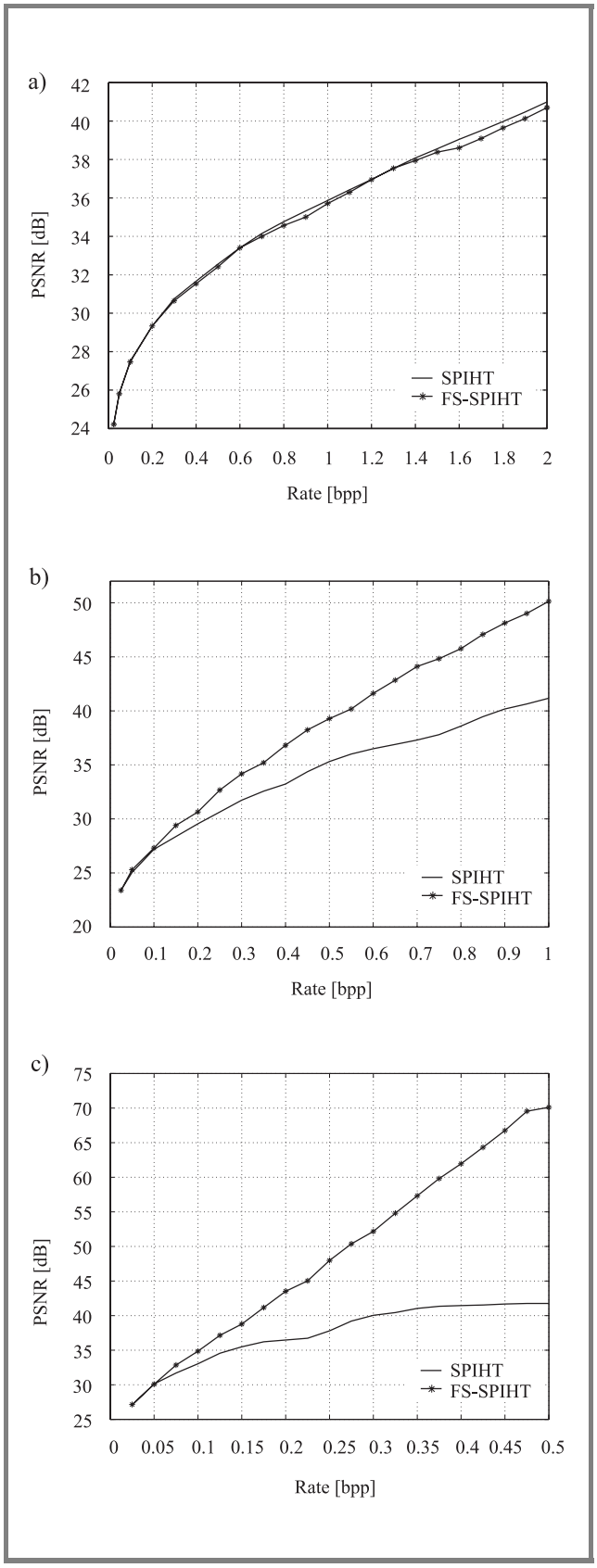


Fig. 5. Comparison of rate-distortion results for the Goldhill test image at different spatial resolution levels: (a) level 1 (original image size 512×512); (b) level 2 (256×256); (c) level 3 (128×128).

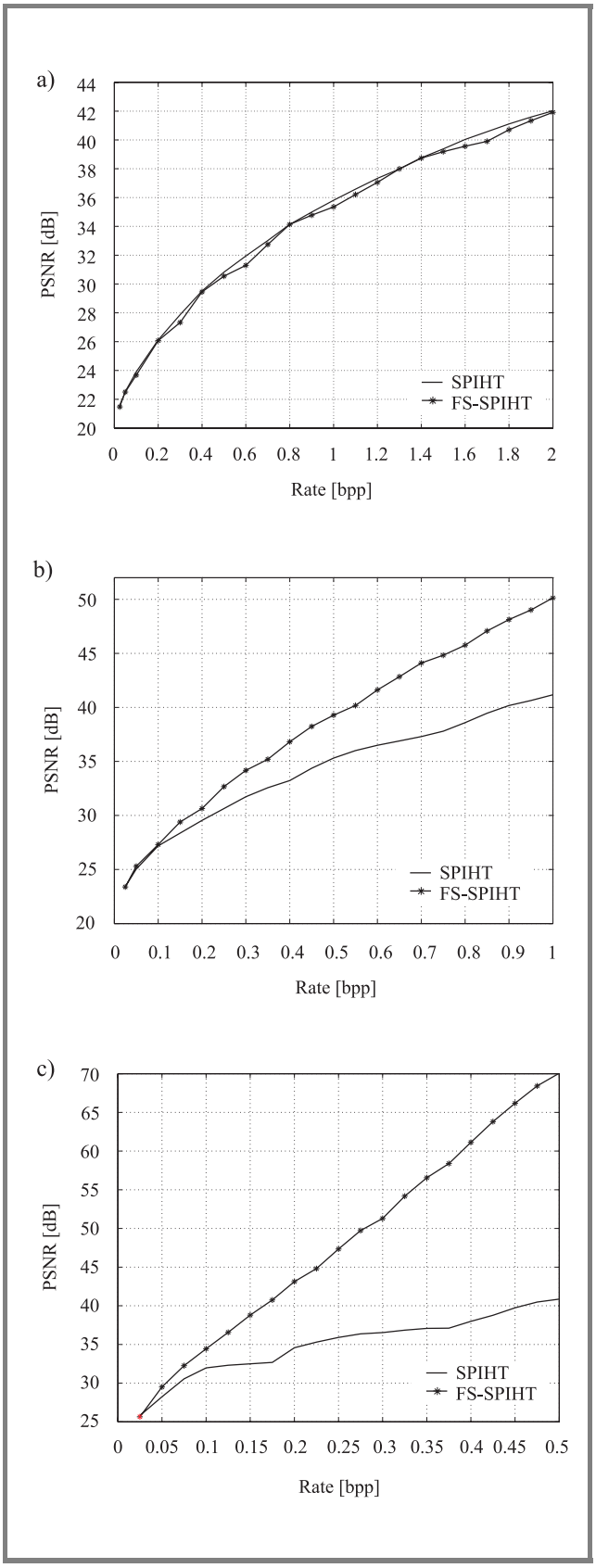


Fig. 6. Comparison of rate-distortion results for the Barbara test image at different spatial resolution levels: (a) level 1 (original image size 512×512); (b) level 2 (256×256); (c) level 3 (128×128).

coding of arbitrarily shaped still and video objects. The proposed multiresolution image codec is a good candidate for multimedia applications such as image storage and retrieval systems, progressive web browsing and multimedia information transmission, especially over heterogenous networks where a wide variety of users need to be differently serviced according to their network access and data processing capabilities.

Acknowledgment

The first author would like to acknowledge the financial support provided for him by the Ministry of Science, Research and Technology (MSRT), Iran and Kurdistan University, Sanandaj, Iran during doing this research as a part of his Ph.D. study at the University of Wollongong, Australia.

References

[1] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards, and Practice*. Boston: Kluwer, 2002.

[2] C. Christopoulos, A. Skordas, and T. Ebrahimi, "The JPEG2000 still image coding system: an overview", *IEEE Trans. Consum. Electron.*, vol. 46, no. 4, pp. 1103–1127, 2000.

[3] J. M. Shapiro, "Embedded image coding using zerotree of wavelet coefficients", *IEEE Trans. Signal Proc.*, vol. 41, pp. 3445–3462, 1993.

[4] A. Zandi, J. D. Allen, E. L. Schwartz, and M. Boliek, "CREW: compression with revesible embedded wavelet", in *Proc. IEEE Data Compres. Conf.*, 1995, pp. 212–221.

[5] Y. Chen and W. A. Pearlman, "Three-dimensional subband coding of video using the zero-tree method", in *Proc. SPIE 2727-VCIP'96*, 1996, pp. 1302–1309.

[6] S. A. Martucci and I. Sodagar, "Entropy coding of wavelet coefficientsfor very low bit rate video", in *IEEE Int. Conf. Image Proc.*, 1996, vol. 2, pp. 533–536.

[7] A. Said and W. A. Pearlman, "A new, fast and efficient image codec based on set partitioning in hierarchical trees", *IEEE Trans. Circ. Syst. Video Technol.*, vol. 6, pp. 243–250, 1996.

[8] J. Liang, "Highly scalable image coding for multimedia applications", in *Proc. ACM Multimedia 97*, 1997, pp. 11–19.

[9] Q. Wang and M. Ghanbari, "Scalable coding of very high resolution video using the virtual zerotree", *IEEE Trans. Circ. Syst. Video Technol.*, vol. 7, no. 5, pp. 719–727, 1997.

[10] S. A. Martucci, I. Sodagar, T. Chiang, and Y.-Q. Zhang, "A zerotree wavelet video coder", *IEEE Trans. Circ. Syst. Video Technol.*, vol. 7, no. 1, pp. 109–118, 1997.

[11] B.-J. Kim and W. A. Pearlman, "An embedded video coder using three-dimensional set partitioning in hierarchical trees (SPIHT)", in *Proc. IEEE Data Compres. Conf.*, 1997, pp. 251–260.

[12] J. Karlenkar and U. B. Desai, "SPIHT video coder", in *Proc. IEEE Reg. 10 Int. Conf. Glob. Connect. Ener., Comput., Commun. Contr., TENCON'98*, 1998, vol. 1, pp. 45–48.

[13] B.-J. Kim, Z. Xiong, and W. A. Pearlman, "Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT)", *IEEE Trans. Circ. Syst. Video Technol.*, vol. 10, no. 8, pp. 1374–1387, 2000.

[14] J. Zho and S. Lawson, "Improvements of the SPIHT for image coding by wavelet transform", in *Proc. IEE Sem. Time-Scale & Time-Freq. Anal. Appl.*, 2000, pp. 24/1–24/5.

[15] E. Khan and M. Ghanbari, "Very low bit rate video coding using virtual SPIHT", *IEE Electron. Lett.*, vol. 37, no. 1, pp. 40–42, 2001.

[16] H. Cai and B. Zeng, "A new SPIHT algorithm based on variable sorting thresholds", in *Proc. IEEE Int. Symp. Circ. Syst.*, 2001, vol. 5, pp. 231–234.

[17] J. Y. Tham, S. Ranganath, and A. A. Kassim, "Highly scalable wavelet-based video codec for very low bit-rate environment", *IEEE J. Select. Areas Commun.*, vol. 16, no. 1, pp. 12–27, 1998.

[18] H. Danyali and A. Mertins, "Highly scalable image compression based on SPIHT for network applications", in *IEEE Int. Conf. Image Proc.*, Rochester, USA, 2002, pp. 217–220.

[19] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform", *IEEE Trans. Image Proc.*, vol. 1, pp. 205–220, 1992.

Habibollah Danyali received the B.E. and M.E. degrees in electrical engineering respectively from the Isfahan University of Technology, Isfahan, Iran, in 1991 and the Tarbeyat Modarres University, Tehran, Iran, in 1993. From 1994 to 2000 he was with the Department of Electrical Engineering, Kurdistan University, Iran, as a lecturer. In 2000 he received a scholarship from the Ministry of Science, Research and Technology, Iran to work towards his Ph.D. at the University of Wollongong, Australia. He is currently a Ph.D. candidate in the School of Electrical, Computer and Telecommunication Engineering, University of Wollongong. His research interests include scalable image and video coding and multimedia delivery over heterogeneous networks.
e-mail: hd04@uow.edu.au
Telecommunication and Information Technology
Research Institute (TITR)
School of Electrical, Computer
and Telecommunications Engineering
University of Wollongong
Wollongong, NSW 2522, Australia

Alfred Mertins – for biography, see this issue, p. 16.

Regular paper

On a method to improve correlation properties of orthogonal polyphase spreading sequences

Beata J. Wysocki and Tadeusz A. Wysocki

Abstract — In this paper, we propose a simple but efficient method for improving correlation properties of polyphase spreading sequences for asynchronous direct sequence code division multiple access (DS CDMA) applications. The proposed method can be used to reduce the mean square value of aperiodic crosscorrelation or the mean square value of aperiodic autocorrelation, the maximum value of aperiodic crosscorrelation functions, merit factor or other properties of the sequence set. The important feature of the method is that while it modifies correlation properties of the sequence set, it preserves sequence orthogonality for perfect synchronization, if this is the property of the original sequence set.

Keywords — spread spectrum communication, orthogonal sequences, wireless communication, wireless LAN, correlation.

1. Introduction

Walsh-Hadamard bipolar spreading sequences are generally used for channel separation in direct sequence code division multiple access systems, e.g. [1]. They are easy to generate, and orthogonal [2] in the case of perfect synchronization. However, the crosscorrelation between two Walsh-Hadamard sequences can rise considerably in magnitude if there is a non-zero delay shift between them [3]. Unfortunately, this is very often the case for up-link (mobile to base station) transmission, due to the differences in the corresponding propagation delays. As a result, significant multi-access interference (MAI) [4] occurs which needs to be combated either by complicated multi-user detection algorithms [5], or reduction in bandwidth utilization.

Another possible solution to this problem can be use of orthogonal complex valued polyphase spreading sequences, like those proposed in [6], which for some values of their parameters can exhibit a reasonable compromise between autocorrelation and crosscorrelation functions. However, in most cases the choice of the parameters is not simple. In addition, improving one of the characteristics is usually associated with significant degradation of the others [7].

In the paper, we propose a method to optimize correlation properties of polyphase sequences which allows to use standard optimization techniques, like the Nelder-Meade simplex search [8] being implemented in several mathematical software packages, e.g. MATLAB. By using a standard optimization technique, one can choose the penalty function in a way, which takes to account all the important correlation characteristics. The numerical example shows application of the method to optimize properties of the orthogonal sequence set of the length 31 from the family of sequences proposed in [6]. The results show that significant changes in sequence characteristics can be achieved. Another example illustrates application of the method to change the characteristics of the orthogonal bipolar sequences, i.e. Walsh-Hadamard sequences of length 32.

The paper is organized as follows. In Section 2, we introduce the method used later to optimize correlation characteristics of the spreading sequences. Section 3 introduces optimization criteria, which can be used for DS CDMA applications. The numerical example of optimization applied to orthogonal polyphase sequences is given in Section 4. Section 5 deals with application of the proposed modification method in case of bipolar sequences, i.e. Walsh-Hadamard sequences, and Section 6 concludes the paper.

2. Modification method

Sets of spreading sequences used for DS CDMA applications can be represented by $M \times N$ matrices $\mathbf{S}_{MN}$, where M is the number of sequences in the set and N is the sequence length. The sequences are referred to as orthogonal sequences if, and only if the matrix $\mathbf{S}_{MN}$ is orthogonal, i.e.

$$\mathbf{S}_{MN}\mathbf{S}_{MN}^{H} = k\mathbf{I}_{M}, \tag{1}$$

where k is a constant, $\mathbf{S}_{MN}^{H}$ is the Hermitian transposition of matrix $\mathbf{S}_{MN}$, (i.e. transposition and taking complex conjugate of the elements of matrix $\mathbf{S}_{MN}$), and $\mathbf{I}_{M}$ is an $M \times M$ unity matrix.

There are a few families of orthogonal spreading sequences proposed in literature, e.g. [2, 6, 9, 10, 11]. Out of them, the most commonly applied are Walsh-Hadamard sequences. Some of the proposed sequence families are designed in a parametric way, which allows for some manipulation of parameters to change the desired correlation characteristics. However, those changes are usually of a limited magnitude, and very often while improving the crosscorrelation functions, a significant worsening of the autocorrelation functions is experienced, e.g. [7].

Here, we propose to modify correlation properties of the set of orthogonal spreading sequences by multiplying the matrix $\mathbf{S}_{MN}$ by another orthogonal $N \times N$ matrix $\mathbf{D}_{N}$. Hence,

the new set of spreading sequences is represented by a matrix $\mathbf{W}_{MN}$

$$\mathbf{W}_{MN} = \mathbf{S}_{MN}\mathbf{D}_N \tag{2}$$

that, of course, is also orthogonal. Hence, the matrix $\mathbf{D}_N$ satisfies the condition:

$$\mathbf{D}_N\mathbf{D}_N^H = c\mathbf{I}_N\,, \tag{3}$$

where c is a real constant. In addition, if $c = 1$, then the sequences represented by the matrix $\mathbf{W}_{MN}$ are not only orthogonal, but possess the same normalization as the original sequences represented by the matrix $\mathbf{S}_{MN}$. However, other correlation properties of the sequences defined by $\mathbf{W}_{MN}$ can be significantly different to those of the original sequences.

To this point, it is not clear how to chose the matrix $\mathbf{D}_N$ to achieve the desired properties of the sequences defined by the $\mathbf{W}_{MN}$. A simple class of orthogonal matrices are diagonal matrices with their elements $d_{m,n}$ fulfilling the condition:

$$|d_{m,n}| = \begin{cases} 0 & \text{for} \quad m \neq n \\ c & \text{for} \quad m = n \end{cases} \;;\; m,n = 1,\ldots,N\,. \tag{4}$$

To achieve the same signal power as in the case of spreading by means of the original sequences defined by $\mathbf{S}_{MN}$, the elements of $\mathbf{D}_N$, being in general complex numbers, must be of the form:

$$d_{m,n} = \begin{cases} 0 & \text{for} \quad m \neq n \\ \exp(j\phi_m) & \text{for} \quad m = n \end{cases} \;;\; m,n = 1,\ldots,N\,, \tag{5}$$

where the phase coefficients $\phi_m;\ m = 1,2,\ldots,N$, are real numbers taking their values from the interval $[0, 2\pi)$, and $j^2 = -1$. The values of $\phi_m;\ m = 1,2,\ldots,N$, can be chosen to improve the correlation and/or spectral properties, e.g. reduce out-off-phase autocorrelation or value of peaks in aperiodic crosscorrelation functions.

3. Optimization criteria

In order to compare different sets of spreading sequences, we need a quantitative measure for the judgment. Therefore, we introduce here some useful criteria, which can be considered for such a purpose. They are based on correlation functions of the set of sequences, since both the level of multiaccess interference and synchronization amenability depend on the crosscorrelations between the sequences and the autocorrelation functions of the sequences, respectively. There are, however, several specific correlation functions that can be used to characterize a given set of spreading sequences [4, 7, 12].

One of the first detailed investigations of the asynchronous DS CDMA system performance was published in 1969 by Anderson and Wintz [13]. They obtained a bound on the signal-to-noise ratio at the output of the correlation receiver for a CDMA system with hard-limiter in the channel. They also clearly demonstrated in their paper the need for considering the aperiodic crosscorrelation properties of the spreading sequences. Since that time, many additional results have been obtained (e.g. [4] and [14]), which helped to clarify the role of aperiodic correlation in asynchronous DS CDMA systems.

For general polyphase sequences and $\{s_n^{(i)}\}$ and $\{s_n^{(k)}\}$; $n = 1,2,\ldots,N$, of length N, the discrete aperiodic correlation function is defined as [12]:

$$c_{i,k}(\tau) = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{n=0}^{N-1-\tau} s_n^{(i)}[s_{n+\tau}^{(k)}]^*, & 0 \leq \tau \leq N-1 \\ \dfrac{1}{N}\displaystyle\sum_{n=0}^{N-1+\tau} s_{n-\tau}^{(i)}[s_n^{(k)}]^*, & 1-N \leq \tau < 0 \\ 0, & |\tau| \geq N \end{cases} \tag{6}$$

where $[\bullet]^*$ denotes a complex conjugate operation. When $\{s_n^{(i)}\} = \{s_n^{(k)}\}$, Eq. (6) defines the discrete aperiodic autocorrelation function.

Another important parameter used to assess the synchronization amenability of the spreading sequence $\{s_n^{(i)}\}$ is a merit factor, or a figure of merit [15], which specifies the ratio of the energy of autocorrelation function mainlobes to the energy of the autocorrelation function sidelobes in the form:

$$F = \frac{c_i(0)}{2\displaystyle\sum_{\tau=1}^{N-1}|c_i(\tau)|^2}\,. \tag{7}$$

In DS CDMA systems, we want to have the maximum values of aperiodic crosscorrelation functions and the maximum values of out-of-phase aperiodic autocorrelation functions as small as possible, while the merit factor as great as possible for all of the sequences used.

The bit error rate (BER) in a multiple access environment depends on the modulation technique used, demodulation algorithm, and the signal-to-noise power ratio (SNR) available at the receiver. Pursley [4] showed that in case of a BPSK asynchronous DS CDMA system, it is possible to express the average SNR at the receiver output of a correlator receiver of the ith user as a function of the average interference parameter (AIP) for the other K users of the system, and the power of white Gaussian noise present in the channel. The SNR for ith user, denoted as SNR_i, can be expressed in the form:

$$\mathrm{SNR}_i = \left(\frac{N_0}{2E_b} + \frac{1}{6N^3}\sum_{\substack{k=1 \\ k\neq i}}^{K}\rho_{k,i}\right)^{-0.5}\,, \tag{8}$$

where E_b is the bit energy, N_0 is the one-sided Gaussian noise power spectral density, and $\rho_{k,i}$ is the AIP, defined for a pair of sequences as

$$\rho_{k,i} = 2\mu_{k,i}(0) + \mathrm{Re}\{\mu_{k,i}(1)\}. \tag{9}$$

The crosscorrelation parameters $\mu_{k,i}(\tau)$ are defined by:

$$\mu_{k,i} = N^2\sum_{n=1-N}^{N-1} c_{k,i}(n)[c_{k,i}(n+\tau)]^*. \tag{10}$$

However, following the derivation in [16], $\rho_{k,i}$ for polyphase sequences may be well approximated as:

$$\rho_{k,i} \approx 2N^2 \sum_{n=1-N}^{N-1} |c_{k,i}(n)|^2. \tag{11}$$

In order to evaluate the performance of a whole set of M spreading sequences, the average mean-square value of crosscorrelation for all sequences in the set, denoted by R_{CC}, was introduced by Oppermann and Vucetic [7] as a measure of the set crosscorrelation performance:

$$R_{CC} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{k=1 \\ k\neq i}}^{M} \sum_{\tau=1-N}^{N-1} |c_{i,k}(\tau)|^2. \tag{12}$$

A similar measure, denoted by R_{AC} was introduced in [7] for comparing the autocorrelation performance:

$$R_{AC} = \frac{1}{M} \sum_{i=1}^{M} \sum_{\substack{\tau=1-N \\ \tau\neq 0}}^{N-1} |c_{i,i}(\tau)|^2. \tag{13}$$

The measure defined by (13) allows for comparison of the autocorrelation properties of the set of spreading sequences on the same basis as the crosscorrelation properties. It can be used instead of the figures of merit, which have to be calculated for the individual sequences.

For DS CDMA applications we want both parameters R_{CC} and R_{AC} to be as low as possible [7]. Because these parameters characterize the whole sets of spreading sequences, it is convenient to use them as the optimization criteria in design of new sequence sets. Therefore, we will use them for optimizing the values of the phase coefficients ϕ_m; $m = 1,2,\ldots,N$, in the considered numerical examples. We will also look into the maximum value of aperiodic crosscorrelation functions since this parameter is very important when the worst-case scenario is considered. Optimization criteria, not based on the correlation characteristics, can be envisaged as well.

4. Optimization of orthogonal polyphase sequences

Oppermann and Vucetic introduced in [7] a new family of polyphase spreading sequences. The elements $u_n^{(k)}$ of these sequences $\{u_n^{(k)}\}$ are given by:

$$u_n^{(k)} = (-1)^{kn} \exp\left[\frac{j\pi(n^m k^p + k^s)}{N}\right], \quad 1 \le n \le N, \tag{14}$$

where k can take integer values being relatively prime to N such that $1 \le k < N$, and the parameters m, p, s can take any real values. They showed there that – depending on the choice of the parameters m, p, and s – the sequences could have a wide range of the correlation properties. However, no clear method for selecting the appropriate values of the parameters depending on the desired correlation characteristics was given in [7]. Later in [6], Oppermann showed that the sequences defined by (14) were orthogonal if $p = 1$ and m is a positive nonzero integer.

In this section, we apply the developed method to improve the properties of the spreading sequence set belonging to the family defined by (14), with $N = 31$, $p = 1$, and $m = 1$. Since N is a prime number, k can take any nonzero integer value lower than 31, i.e. $k = 1,2,\ldots,30$, and the maximum number of sequences in the set is 30. To select the appropriate value for s, we plotted in Fig. 1 the values of R_{CC}, R_{AC} and the value of the maximum peak in all aperiodic crosscorrelation functions $C_{\max}$ as the functions of s. From the plots we choose $s = 2.5$, for which $R_{CC} = 0.9803$, $R_{AC} = 0.5713$, and $C_{\max} = 0.4546$.

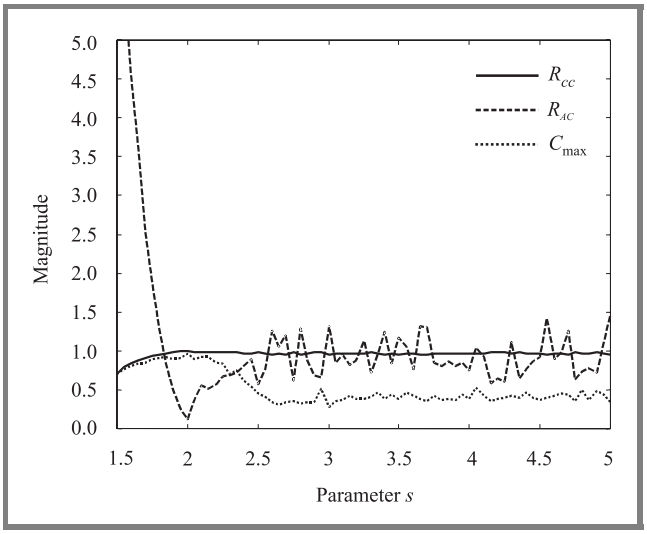


Fig. 1. Plots of the values of R_{CC}, R_{AC}, and $C_{\max}$ as functions of the parameter s for the sequences $\{u_n^{(k)}\}$, with $N = 31$, $p = 1$, $m = 1$, and $k = 1,2,\ldots,30$.

To illustrate the modification method, we first applied it to reduce the value of R_{CC} for this set of sequences. The process was performed using the standard "*fmin*" function of MATLAB [8] with the optimized function being $R_{CC}(\Phi)$, where

$$\Phi = [\phi_m;\ m = 1,2,\ldots,31] \tag{15}$$

and the phase coefficients ϕ_m; $m = 1,2,\ldots,31$, being used to define the elements of the modification matrix $\mathbf{D}_N$ (see Eq. (5)).

The function $R_{CC}(\Phi)$ is very irregular and may have several local minima. Therefore, depending on the starting point, different local minima can be reached. To illustrate this, we selected randomly 6 vectors $\Phi_1, \Phi_2, \ldots, \Phi_6$, and run "*fmin*" for each of them chosen as a starting point. The results of the obtained values of R_{CC}, R_{AC} and $C_{\max}$ are given in Table 1.

Next we repeat the procedure, this time finding the sequences to reduce the value of R_{AC}, and finally to achieve reduction in the value of $C_{\max}$. The results of R_{CC}, R_{AC} and $C_{\max}$ are given in Tables 2 and 3, respectively.

Table 1
Values of R_{CC}, R_{AC} and $C_{\max}$ obtained for the sequences optimized to achieve minimum R_{CC}

Coefficients	R_{CC}	R_{AC}	$C_{\max}$
$\Phi_{1,\text{opt}}$	0.3262	18.1835	0.3262
$\Phi_{2,\text{opt}}$	0.3199	18.9537	0.3199
$\Phi_{3,\text{opt}}$	0.3774	17.6342	0.3334
$\Phi_{4,\text{opt}}$	0.3583	18.3313	0.3233
$\Phi_{5,\text{opt}}$	0.3990	17.8249	0.3561
$\Phi_{6,\text{opt}}$	0.4314	16.8953	0.3872

Comparison of the results listed in Tables 1, 2, and 3, indicates that the best compromise amongst the values of R_{CC}, R_{AC} and $C_{\max}$ were obtained while searching for the lowest value of $C_{\max}$. The values of the phase coefficients leading to the values listed in Table 3 are presented in Table 4.

Table 2
Values of R_{CC}, R_{AC} and $C_{\max}$ obtained for the sequences optimized to achieve minimum R_{AC}

Coefficients	R_{CC}	R_{AC}	$C_{\max}$
$\Phi_{1,\text{opt}}$	0.9965	0.1006	0.5310
$\Phi_{2,\text{opt}}$	0.9974	0.0749	0.3795
$\Phi_{3,\text{opt}}$	0.9969	0.0910	0.3774
$\Phi_{4,\text{opt}}$	0.9963	0.1081	0.3583
$\Phi_{5,\text{opt}}$	0.9974	0.0762	0.3990
$\Phi_{6,\text{opt}}$	0.9972	0.0824	0.4314

Table 3
Values of R_{CC}, R_{AC} and $C_{\max}$ obtained for the sequences optimized to achieve minimum $C_{\max}$

Coefficients	R_{CC}	R_{AC}	$C_{\max}$
$\Phi_{1,\text{opt}}$	0.9743	0.7475	0.2817
$\Phi_{2,\text{opt}}$	0.9592	1.1846	0.2871
$\Phi_{3,\text{opt}}$	0.9763	0.6872	0.2855
$\Phi_{4,\text{opt}}$	0.9748	0.7323	0.2933
$\Phi_{5,\text{opt}}$	0.9720	0.8141	0.2781
$\Phi_{6,\text{opt}}$	0.9692	0.8940	0.2730

To show that the modified sequences are still orthogonal, in Fig. 2, we plotted the function $C_{\max}(\tau)$ for the sequences $\{w_n^{(k)}\}$ obtained from the original sequences $\{u_n^{(k)}\}$ by modifying them using the vector $\Phi_{6,\text{opt}}$ from Table 4. For the comparison, we plotted there also the function $C_{\max}(\tau)$ for the original sequences $\{u_n^{(k)}\}$ using a dashed line. It is clearly visible that both sets of sequences are orthogonal, and the values of $C_{\max}(\tau)$ are significantly lower for the new sequence set than for the original one around zero, which corresponds to the point of perfect synchronization.

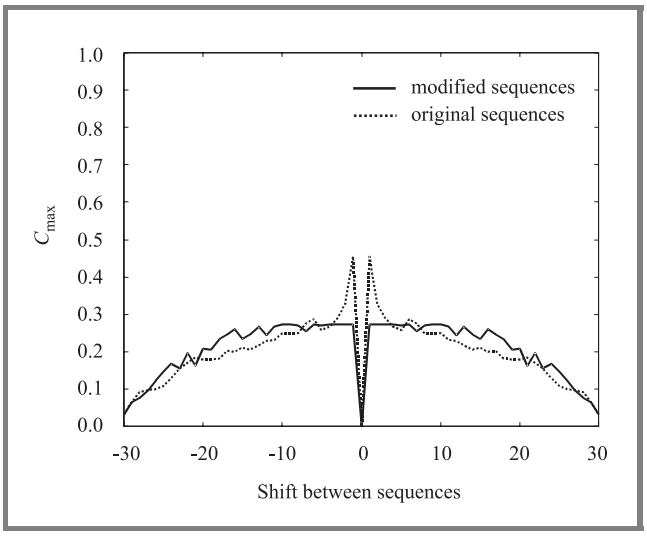


Fig. 2. Plots of the maximum peaks in the crosscorrelation functions versus the relative shift between the sequences, $C_{\max}(\tau)$.

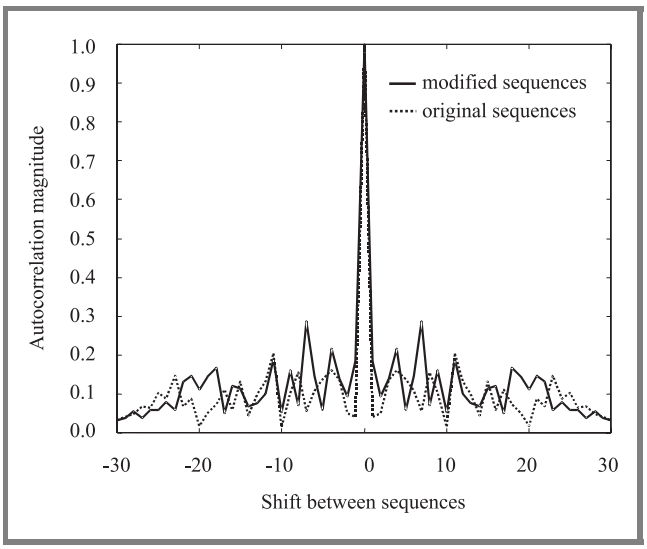


Fig. 3. Plots of the maximum magnitudes of the autocorrelation functions, $A_{\max}(\tau)$.

To compare the synchronization amenability of the original sequences $\{u_n^{(k)}\}$ and these new sequences $\{w_n^{(k)}\}$, we plotted the maximum magnitudes of the autocorrelation functions $A_{\max}(\tau)$

$$A_{\max}(\tau) = \max_i c_{i,i}(\tau), \quad i = 1,2,\ldots,31 \tag{16}$$

for both sets of sequences in Fig. 3. It is possible to notice that the maxima in the off-peak autocorrelation are slightly higher for the set of sequences $\{w_n^{(k)}\}$ than for the set of sequences $\{u_n^{(k)}\}$. However, in both cases the peak at zero shift, corresponding to the perfect synchronization, is very significant.

Table 4

Values of the phase coefficients for which the results listed in Table 3 were obtained

Coefficients	$\Phi_{1,opt}$	$\Phi_{2,opt}$	$\Phi_{3,opt}$	$\Phi_{4,opt}$	$\Phi_{5,opt}$	$\Phi_{6,opt}$
Φ_1	0.3925	5.2150	4.6452	4.0811	3.9235	1.4734
Φ_2	0.5345	4.2788	4.6753	5.3314	1.5435	2.0594
Φ_3	3.7864	4.4254	5.8496	5.0765	4.7242	2.8292
Φ_4	1.1751	1.8707	2.5273	3.7142	2.8908	0.1822
Φ_5	5.0800	1.0640	3.8517	3.0107	5.5125	3.9440
Φ_6	1.0535	1.0444	5.4272	0.7436	2.8254	2.1154
Φ_7	1.1270	1.1725	1.1420	4.0309	5.9926	3.3442
Φ_8	0.4683	2.7861	5.0271	2.0431	0.4931	3.6686
Φ_9	3.0181	5.1669	3.8160	0.8948	3.1154	0.9194
Φ_{10}	2.5626	3.1003	4.0349	3.4789	2.1240	4.5375
Φ_{11}	2.0366	0.5176	1.3245	4.4366	5.5740	5.7438
Φ_{12}	1.8242	2.6811	3.1068	4.3561	2.1648	3.7034
Φ_{13}	2.0952	2.6034	0.3976	5.8205	3.1619	0.1789
Φ_{14}	3.5153	2.5592	2.1515	5.7127	0.3111	5.0597
Φ_{15}	0.7190	2.1976	3.8016	0.4326	2.2879	3.6221
Φ_{16}	0.2276	6.1675	2.3824	2.4168	3.4585	4.7125
Φ_{17}	2.5854	0.0353	4.5556	3.9018	2.6660	0.5854
Φ_{18}	4.3384	1.8769	4.5526	1.6916	3.7406	2.7495
Φ_{19}	0.8591	0.3220	2.8490	3.7723	3.8776	2.3331
Φ_{20}	1.5947	4.3013	2.7105	0.2738	0.6973	1.3127
Φ_{21}	1.0851	4.4499	3.9996	3.5643	5.4335	5.3541
Φ_{22}	4.1919	5.2382	4.8776	5.4238	5.6328	4.9300
Φ_{23}	1.6175	2.6795	2.0416	0.2190	5.3110	3.1792
Φ_{24}	4.3589	2.5970	2.7735	5.0224	4.1288	5.7822
Φ_{25}	0.5043	1.3012	3.7371	4.9092	4.1911	0.9306
Φ_{26}	4.0159	4.2658	3.8720	3.0306	1.0407	5.7265
Φ_{27}	5.2438	4.7754	3.6737	3.7614	1.8494	4.9512
Φ_{28}	0.0640	2.6637	2.5367	5.2247	4.2399	2.6627
Φ_{29}	0.9041	0.0558	4.4353	0.6034	3.6006	3.9809
Φ_{30}	4.8797	2.6827	3.3236	4.9620	0.3715	6.0747
Φ_{31}	2.7785	5.3831	1.1931	5.8136	0.5456	3.9853

5. Application to bipolar sequences

From the implementation point of view, the most important class of spreading sequences are bipolar or biphase sequences, where the ϕ_m; $m = 1, 2, \ldots, N$, can take only two values 0 and π. This results in the elements on the diagonal of $\mathbf{D}_N$ being equal to either "+1" or "−1". Even for this bipolar case, we can achieve significantly different properties of the sequences defined by the $\mathbf{W}_N$ than those of the original bipolar sequences of the same length.
As an example, let us compare some properties of Walsh-Hadamard sequences, with the properties of the sequence set defined by the bipolar matrix $\mathbf{W}_{32}$, with the diagonal of the $\mathbf{D}_{32}$ represented for the simplicity by a sequence of "+" and "−" corresponding to "+1" and "−1", respectively:

$$\{++++--++--+-+-+--++-+---+----+\}. \tag{17}$$

For the unmodified set of 32 Walsh-Hadamard sequences of length $N = 32$, the maximum in the aperiodic crosscorrelation function $C_{\max}$ reaches 0.9688, and the mean square out-of-phase aperiodic autocorrelation R_{AC} is equal to 6.5938. That high value of R_{AC} indicates the possibility of significant difficulties in the sequence acquisition process, and the high value of $C_{\max}$ means that for some time shifts the interference between the different DS CDMA channels can be unacceptably high. On the other hand, for the set of sequences defined by the matrix $\mathbf{W}_{32}$ considered here, we have $C_{\max} = 0.4375$, and $R_{AC} = 0.8438$. This means lower peaks in the instantaneous bit-error-rate due to the MAI and a significant improvement in the sequence acquisition process.

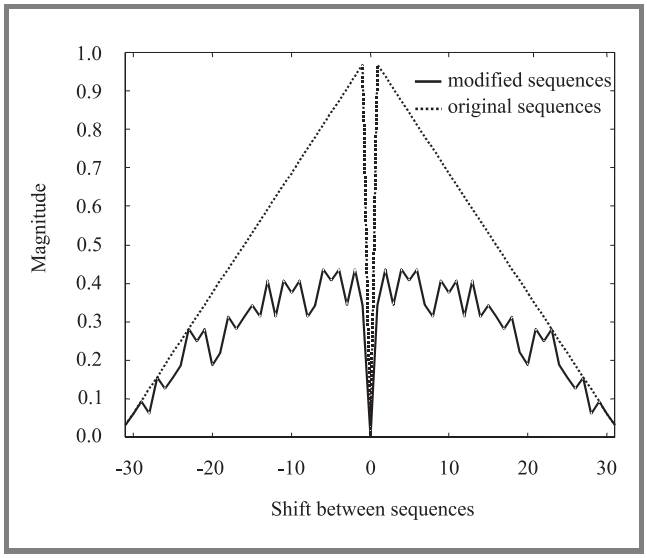


Fig. 4. Plots of the maximum value of the aperiodic crosscorrelation $ACC_{\max}$ for all possible pairs of the sequences versus the relative shift between them.

In Figure 4, we present the plot of the upper limits for the aperiodic crosscorrelation functions for the set of

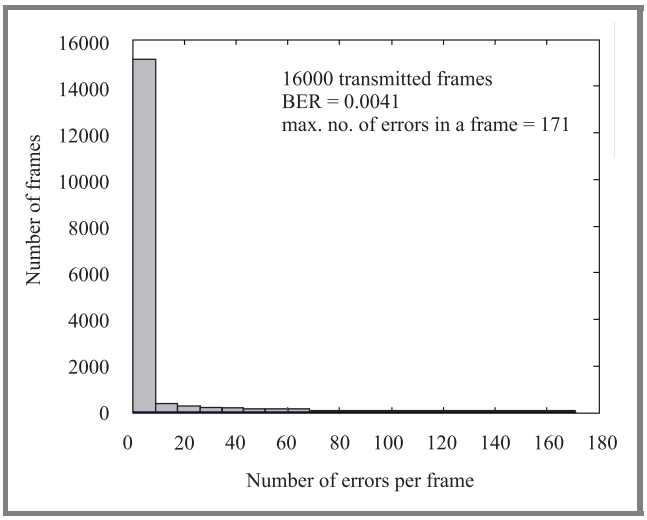


Fig. 5. Histogram of a number of errors in a transmitted frame for the DS CDMA system utilizing Walsh-Hadamard spreading sequences.

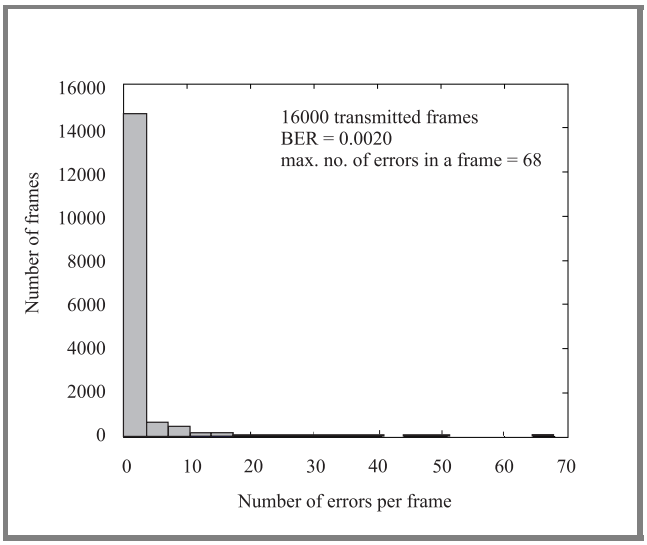


Fig. 6. Histogram of a number of errors in a transmitted frame for the DS CDMA system utilizing spreading sequences defined by the matrix $\mathbf{W}_{32}$.

Walsh-Hadamard sequences and the set of sequences defined by our matrix $\mathbf{W}_{32}$. The presented plots illustrate the improvement that can be achieved by our modification. This, of course, translates directly on the level of BER caused by MAI, which has been confirmed by simulations involving transmission of 500 frames of 524-bites over 32-channel asynchronous DS CDMA systems. One of those systems utilized 32-chip original Walsh-Hadamard sequences and another one employed 32-chip modified sequences defined by the matrix $\mathbf{W}_{32}$. In both cases, the number of simultaneously active users was equal to 8. From the simulations, we achieved an average BER over all 32 channels equal to 0.0041 for the original Walsh-Hadamard sequences, and 0.0020 for the modified se-

quences, respectively. In addition to the reduction in BER, the maximum number of errors in any frame is much lower for the system utilizing the modified sequences, i.e. equal to 68 compared to 171 for the system utilizing the unmodified sequences. This is illustrated in Figs. 5 and 6 showing the histograms of the number of errors per received frame for the system utilizing the original Walsh-Hadamard sequences and the modified sequences, respectively.

6. Conclusions

In the paper, we presented a simple method to modify orthogonal spreading sequences to improve their correlation properties for asynchronous applications, while maintaining their orthogonality for perfect synchronization. The method leads, in general, to complex polyphase sequences but can also be used to obtain real bipolar sequences. In the case of polyphase sequences, the phase coefficients can be chosen to achieve the required correlation/spectral properties of the whole set of sequences. The presented numerical example illustrates how different correlation characteristics can be successfully modified or even optimized for the set of polyphase orthogonal sequences. Of course, different search criteria can be used depending on the particular application. Another presented example shows that for practical applications, where bipolar sequences are preferred, the method can also yield a significant improvement in the properties of the sequence set over those of pure Walsh-Hadamard sequences. Simulation results indicated that the asynchronous DS CDMA system utilizing our sequences has lower BER and significantly smaller number of errors per frame than that can be achieved when the system utilizes unmodified Walsh-Hadamard sequences.

References

[1] R. Steele, "Introduction to digital cellular radio", in *Mobile Radio Communications*, R. Steele and L. Hanzo, Eds., 2nd ed. New York: IEEE Press, 1999.

[2] H. F. Harmuth, *Transmission of Information by Orthogonal Functions*. Berlin: Springer-Verlag, 1970.

[3] B. J. Wysocki and T. A. Wysocki, "Orthogonal binary sequences with wide range of correlation properties", in *Proc. 6th Int. Symp. Commun. Theory Appl., ISCTA'01*, Ambleside, Lake District, U.K., 2001, pp. 483–485.

[4] M. B. Pursley, "Performance evaluation for phase-coded spread-spectrum multiple-access communication – Part I: System analysis", *IEEE Trans. Commun.*, vol. COM-25, pp. 795–799, 1977.

[5] A. Duel-Hallen, J. Holtzman, and Z. Zvonar, "Multiuser detection for CDMA systems", *IEEE Person. Commun.*, pp. 46–58, Apr. 1995.

[6] I. Oppermann, "Orthogonal complex-valued spreading sequences with a wide range of correlation properties", *IEEE Trans. Commun.*, vol. COM-45, pp. 1379–1380, 1997.

[7] I. Oppermann and B. S. Vucetic, "Complex spreading sequences with a wide range of correlation properties", *IEEE Trans. Commun.*, vol. COM-45, pp. 365–375, 1997.

[8] MATLAB, "Reference guide", The Math Works, 1996.

[9] A. W. Lam and S. Tantaratana, "Theory and applications of spread-spectrum systems", IEEE/EAB Self-Study Course, IEEE Inc., Piscataway, 1994.

[10] B. J. Wysocki, "Signal formats for code division multiple access wireless networks". Ph.D. thesis, Curtin University of Technology, Perth, Western Australia, 1999.

[11] J. E. Hershey, G. J. Saulnier, and N. Al-Dhahir, "New Hadamard basis", *Electron. Lett.*, vol. 32, no. 5, pp. 429–430, 1996.

[12] P. Fan and M. Darnell, *Sequence Design for Communications Applications*. New York: Wiley, 1996.

[13] D. R. Anderson and P. A. Wintz, "Analysis of a spread-spectrum multiple-access system with a hard limiter", *IEEE Trans. Commun. Techn.*, vol. COM-17, pp. 285–290, 1969.

[14] J. L. Massey and J. J. Uhran, "Sub-baud coding", in *Proc. Thirteenth Ann. Allert. Conf. Circ. Syst. Theory*, Oct. 1975, pp. 539–547.

[15] M. J. E. Golay, "The merit factor of long low autocorrelation binary sequences", *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 543–549, 1982.

[16] K. H. Kärkkäinen, "Mean-square cross-correlation as a performance measure for spreading code families", in *Proc. IEEE 2nd Int. Symp. Spread Spectr. Techn. Appl. ISSSTA'92*, pp. 147–150.

Beata Joanna Wysocki graduated from Warsaw University of Technology receiving her M.Eng. degree in electrical engineering in 1991. In 1994, she started the Ph.D. study in the Australian Telecommunications Research Institute at Curtin University of Technology. In March 2000, she was awarded her Ph.D. for the thesis: "Signal Formats for Code Division Multiple Access Wireless Networks". During the Ph.D. studies, she was involved in a research project "Wireless ATM Hub" at the Cooperative Research Centre for Broadband Telecommunications and Networking, and worked as a research assistant at Edith Cowan University, within the ARC funded "CDMA with enhanced protection against frequency selective fading", and "Reliable high rate data transmission over microwave local area networks". Since October 1999 she has been with the Telecommunications & Information Technology Research Institute at the University Wollongong as a research fellow. Her research interests include sequence design for direct sequence (DS) code division multiple access (CDMA) data networks and error control strategies for broadband wireless access (BWA) systems.
e-mail: wysocki@uow.edu.au
School of Electrical, Computer
and Telecommunications Engineering
University of Wollongong
Wollongong, NSW 2522, Australia

Tadeusz Antoni Wysocki – for biography, see this issue, p. 37.

INFORMATION FOR AUTHORS

The *Journal of Telecommunications and Information Technology* is published quarterly. It comprises original contributions, both regular papers and letters, dealing with a broad range of topics related to telecommunications and information technology. Items included in the journal report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

The *Journal* is dedicated to publishing research results which advance the level of current research or add to the understanding of problems related to modulation and signal design, wireless communications, optical communications and photonic systems, speech devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology. Suitable research-related manuscripts should hold the potential to advance the technological base of telecommunications and information technology. Tutorial and review papers are published by invitation only.

Papers published by invitation and regular papers should contain up to 15 and 8 printed pages respectively (one printed page corresponds approximately to 3 double-space pages of manuscript, where one page contains approximately 2000 characters).

Manuscript: An original and two copies of the manuscript must be submitted, each completed with all illustrations and tables attached at the end of the papers. Tables and figures have to be numbered consecutively with Arabic numerals. The manuscript must include an abstract limited to approximately 100 words. The abstract should contain four points: statement of the problem, assumptions and methodology, results and conclusion, or discussion, of the importance of the results. The manuscript should be double-spaced on only one side of each A4 sheet (210 × 297 mm). Computer notation such as Fortran, Matlab, Mathematica etc., for formulae, indices, etc., is not acceptable and will result in automatic rejection of the manuscript. The style of references, abbreviations, etc., should follow the standard IEEE format.

References should be marked in the text by Arabic numerals in square brackets and listed at the end of the paper in order of their appearance in the text, including exclusively publications cited inside. The **reference entry** (correctly punctuated according to the following rules and examples) **has to contain.**

From journals and other serial publications: initial(s) and second name(s) of the author(s), full title of publication (transliterated into Latin characters in case it is in Russian, possibly preceded by the title in Russian characters), appropriately abbreviated title of periodical, volume number, first and last page number, year. E.g.:

[1] Y. Namihira, "Relationship between nonlinear effective area and modefield diameter for dispersion shifted fibres", *Electron. Lett.*, vol. 30, no. 3, pp. 262–264, 1994.

From non-periodical, collective publications: as above, but after title – the name(s) of editor(s), title of volume and/or edition number, publisher(s) name(s) and place of edition, inclusive pages of article, year. E.g.:

[2] S. Demri, E. Orłowska, "Informational representability: Abstract models versus concrete models" in *Fuzzy Sets, Logics and Reasoning about Knowledge*, D. Dubois and H. Prade, Eds. Dordrecht: Kluwer, 1999, pp. 301–314.

From books: initial(s) and name(s) of the author(s), place of edition, title, publisher(s), year. E.g.:

[3] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1986.

Figure captions should be started on separate sheet of papers and must be double-spaced.

Illustration: Original illustrations should be submitted. All line drawings should be prepared on white drawing paper in black India ink. Drawings in Corel Draw and Postscript formats are preferred. Colour illustrations are accepted only in exceptional circumstances. Lettering should be large enough to be readily legible when drawing is reduced to two- or one-column width – as much as 4:1 reduction from the original. Photographs should be used sparingly. All photographs must be gloss prints. All materials, including drawings and photographs, should be no larger than 175 × 260 mm.

Page number: Number all pages, including tables and illustrations (which should be grouped at the end), in a single series, with no omitted numbers.

Electronic form: A floppy disk together with the hard copy of the manuscript should be submitted. It is important to ensure that the diskette version and the printed version are identical. The diskette should be labelled with the following information: a) the operating system and word-processing software used, b) in case of UNIX media, the method of extraction (i.e. tar) applied, c) file name(s) related to manuscript. The diskette should be properly packed in order to avoid possible damage during transit.

Among various acceptable word processor formats, TEX and LATEX are preferable. The *Journal's* style file is available to authors.

Galley proofs: Proofs should be returned by authors as soon as possible. In other cases, the article will be proofread against manuscript by the editor and printed without the author's corrections. Remarks to the errata should be provided within two weeks after receiving the offprints.

The copy of the "Journal" shall be provided to each author of papers.

Copyright: Manuscript submitted to this journal may not have been published and will not be simultaneously submitted or published elsewhere. Submitting a manuscript, the authors agree to automatically transfer the copyright for their article to the publisher if and when the article is accepted for publication. The copyright comprises the exclusive rights to reproduce and distribute the article, including reprints and also all translation rights. No part of the present journal may be reproduced in any form nor transmitted or translated into a machine language without permission in written form from the publisher.

Biographies and photographs of authors are printed with each paper. Send a brief professional biography not exceeding 100 words and a gloss photo of each author with the manuscript.

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Editorial Office

tel. +48(22) 872 43 88
tel./fax: +48(22) 512 84 00
e-mail: redakcja@itl.waw.pl
http://www.itl.waw.pl/jtit