International Journal of Communications, Network and System Sciences

ISSN: 1913-3715

Volume 2, Number 9, December 2009

www.scirp.org/journal/ijcns/

Scientific Research

JOURNAL EDITORIAL BOARD

ISSN 1913-3715 (Print) ISSN 1913-3723 (Online)
http://www.scirp.org/journal/ijcns/

Int. J. Communications, Network and System Sciences, **2009, 2, 805-926**
Published Online December 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

TABLE OF CONTENTS

Volume 2 Number 9 December 2009

International Journal of Communications, Network and System Sciences (IJCNS)

Journal Information

SUBSCRIPTIONS

The *International Journal of Communications, Network and System Sciences* (Online at Scientific Research Publishing, www.SciRP.org) is published monthly by Scientific Research Publishing, Inc.,USA.

E-mail: service@scirp.org

Subscription rates: Volume 2 2009
Print: $50 per copy.
Electronic: free, available on www.SciRP.org.
To subscribe, please contact Journals Subscriptions Department, E-mail: service@scirp.org

Sample copies: If you are interested in subscribing, you may obtain a free sample copy by contacting Scientific Research Publishing, Inc at the above address.

SERVICES

Advertisements
Advertisement Sales Department, E-mail: service@scirp.org

Reprints (minimum quantity 100 copies)
Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.
E-mail: service@scirp.org

COPYRIGHT

Statements and opinions expressed in the articles and communications are those of the individual contributors and not the statements and opinion of Scientific Research Publishing, Inc. We assumes no responsibility or liability for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained herein. We expressly disclaim any implied warranties of merchantability or fitness for a particular purpose. If expert assistance is required, the services of a competent professional person should be sought.

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: ijcns@scirp.org

Int. J. Communications, Network and System Sciences, **2009, 2, 805-821**
doi:10.4236/ijcns.2009.29094 published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Radio Access Selection in Integrated UMTS/WLAN Networks

I. MODEAS[1], A. KALOXYLOS[2], G. LAMPROPOULOS[1], N. PASSAS[1], L. MERAKOS[1]
[1]*Department of Informatics and Telecommunications, University of Athens, Athens, Greece*
[2]*Department of Telecommunications Science and Technology, University of Peloponnese, Tripoli, Greece*
E-mail: {*imodeas, glambr, passas, merakos*}*@di.uoa.gr*, *kaloxyl@uop.gr*

Received August 24, 2009; *revised September* 13, 2009; *accepted October* 31, 2009

Abstract

Heterogeneous networks combine different access technologies. An important problem in such networks is the selection of the most suitable radio access network. To perform this task efficiently, a lot of information is required, such as signal strength, QoS, monetary cost, battery consumption, and user preferences. These are well known issues and a considerable effort has been made to tackle them using a number of solutions. These efforts improve the performance of vertical handover but also add considerable complexity. In this paper, we introduce an enhanced algorithm for radio access network selection, which is simple, flexible and applicable to future mobile systems. Its main characteristics are the distribution of the radio access selection process among the mobile terminal and the core network, the evaluation of mobile terminal connections separately and the primary role of user preferences in the final decision. The performance of the algorithm is evaluated through simulation results, which show that the algorithm provides a high rate of user satisfaction. It decreases the messages required for the vertical handovers in the whole network and it alleviates the core network from the processing of unnecessary requests.

Keywords: Network Selection, Algorithm, Heterogeneous, Vertical Handover, WLAN, UMTS

1. Introduction

The expected evolution of mobile communications will offer several radio access technologies (RATs) with different characteristics served by a common core network. These networks try to combine RATs with different capabilities in a co-operating rather than in a competing manner. In this way, they combine complementary advantages of all RATs; they improve the overall network capacity and the supported quality of the service. Such heterogeneous networks can be seen as an evolution of 3G cellular networks, e.g., an integrated UMTS with WLAN coverage areas (hotspots). In these networks, the user can benefit from the universal coverage and the quality of service provision of UMTS, along with the higher bandwidth availability combined with the lower cost of the WLAN. At the same time, the network providers find an inexpensive way to increase the network capacity, to alleviate the UMTS radio interface from significant load and to provide more services at a lower cost for the users. In such environments, a mechanism is required for the user and the network to select the most appropriate RAT for a connection.

In a homogeneous network where only one RAT is available, the main factors upon deciding on the best access point to the network are the measured quality of the radio signal and the congestion of a cell. However, in a heterogeneous network, this procedure is far more complicated. If signal strength measurements at the mobile terminal (MT) suggest that more than one RAT is appropriate to serve a connection, additional parameters have to be evaluated to reach the final decision. Such parameters are the user preferences, the monetary cost, the battery consumption, the location/speed/direction of a user, the type of QoS support, as well as the current traffic load in a target RAT. Thus, the final decision should be reached as the result of a trade-off between different and sometimes contradicting criteria. Reaching this decision by keeping the user satisfied and not violating the network policies can be a very complex problem. Several proposals have been presented to tackle this issue. Some of the proposals are based on the numerical outcome of mathematical functions. Other proposed solutions are based in fuzzy logic, neural networks or a combination of them. Also policy-based schemes have been proposed to tackle the same issue. All these solutions are briefly described in the following paragraphs, while a more detailed description can be found in [1].

As mentioned earlier, one way to tackle this issue is to use mathematical functions (a.k.a. cost functions [2], score functions [3] and user utility or benefit functions [4]). Cost functions calculate the cost of using a specific RAT at the given time and the RAT with the least cost is selected. Score, benefit functions and user utility pick up the RAT with the higher result. The outcome of these functions is dependent on several parameters such as bandwidth, user preferences, power consumption etc. All parameters in such functions are normalised and the decision metric comes as a linear equation of all parameters with suitable weights. Each of these parameters may change dynamically over time, so it is necessary to recalculate these functions every time a decision needs to be taken. Such mathematical functions provide a simple way to select a RAT for a connection. Also, performance analysis in [2] shows that when compared to traditional mechanisms both throughput and effective bandwidth are improved. On the other hand, the different parameters in these functions have different units (e.g., dB in signal strength with dollars of using a network and hours of battery life) and there is a point to think about when mixing all these in a single equation. Furthermore, a certain unit does not directly measure some parameters, such as security and user preferences. So it is not always clear how they can be formed as mathematical equations and incorporated in such mechanisms.

Another solution is the use of fuzzy logic [5], of neural networks [6] or their combination [7]. As in the previous case, the solutions in this category consider many parameters, apart from the signal strength, in the heterogeneous environment to provide for solutions in the HO initiation and decision. These solutions can take into consideration both the user preferences as well as the operators' policies. This poses a quite complicated problem, where fuzzy logic systems and neural network classifiers can offer flexible solutions to cope with imprecise data. They can minimize the number of unnecessary HOs and maximize the percentage of satisfied users. The disadvantage of these solutions is that they increase the complexity of the decision process and that in the case of neural networks a pre-training session of the system is required [8].

A third solution to the same problem is the use of policy-based schemes. The term policy describes a rule-set that has to be enforced in the RAT selection. Policy-based schemes may involve several network entities and they can offer a simple or a more sophisticated solution based on rules that are mapped to actions taken when specific events occur. These events involve the change of various parameters, some of which are static (they do not change over time) and some dynamic (their values change), that have to be considered in the heterogeneous network. By keeping the rules simple, these solutions provide for a fast and easily implemented solution at the expense of non-optimal resource utilisation. In order to avoid this drawback, more sophisticated policies can be introduced, but the complexity of the system is increased. Special care is needed in order to avoid conflicts between different policies, especially when residing in different network nodes. These schemes may be combined with one of the previous mechanisms in order to make the final decision. It is important to mention that strict rules do not supply scalability and flexibility to cope with all contradicting parameters involved. There is always a tradeoff between the complexity of the network architecture and the performance of the system [9–11].

Apart from the disadvantages already presented in each one of the aforementioned category of solutions, most of them have been evaluated in a theoretical level and have been designed without any prior study on the required signalling exchange and the required calculations to be performed inside the network. Instead, the proposed mechanism has been designed with exactly these attributes in mind, trying to keep the overall procedure as simple as possible and offer a flexible and extendible solution. More specifically, the main aim of our proposal is to alleviate the core network from several calculations, to avoid certain unnecessary HO triggering and the corresponding signalling load in the radio interface, and at the same time to highly meet the user's preferences. A prior version of this algorithm has been introduced in [12]. Here we elaborate this work and evaluate the algorithm using simulation results.

The proposed algorithm deals with RAT selection in an integrated UMTS/WLAN heterogeneous network. There are three key points at the algorithm design:

1) It considers the user preferences in order to make the final decision. In other words, the first step is for the user to provide the network with a set of acceptable solutions.

2) It evaluates each connection of a MT separately and proposes the most suitable RAT for each one of them. The MT builds a prioritised list of target RATs per connection, based on a number of parameters (e.g., user profile, monetary cost, battery consumption). The network operator will decide based on this list and the values of another set of parameters such as the user speed and location, and the congestion of a target RAT. All these issues are explained in detail in the forthcoming sections.

3) It is split into two distinct and cooperating parts. The first runs on the MT while the second in the core network (CN). This architectural option aims at reducing the overall complexity of the system and the signalling exchange between entities by having the terminals to actively participate in complex operations. This is a valid option for us since next generation mobile terminals are expected to be equipped with more advanced processing and memory capabilities.

Although there is some work in the literature presenting the above key points, up to our knowledge, none of them take advantage of all of them at the same time.

Furthermore, there is no evaluation on the user satisfaction and the signalling load alleviation. The simulation results presented at the end of the paper justify the aforementioned design points. When compared with other proposal the key difference of our mechanism is that its main focus is not on a pure load-balanced system but rather on how to satisfy the preferences of the users.

The remainder of the paper is organised as follows. Section 2 elaborates the MT and network parts of the algorithm. Next, Section 3 presents a quantitative evaluation of the algorithm in an integrated UMTS/WLAN environment, through simulations. Finally, conclusions and future work are described in Section 4.

2. RAT Selection Algorithms

The algorithm involves several parameters in the decision process. Also, some assumptions were made for its design and functionality. Furthermore, as already mentioned, the proposed algorithm is the combination of two sub-algorithms: the one running at the MT and the second one at the CN. All these are described in the following subsections.

2.1. Assumptions and Parameters of the Algorithm

In the proposed algorithm, we assume that the MTs are multimode, i.e., they have multiple radio interfaces in order to support a number of connections via more than one RAT at the same time [13]. We focus on the problem of selecting the most suitable access network: 1) when a new call is to be initiated; 2) when a new alternative RAT becomes reachable by a MT having active connections (i.e., a vertical HO is imminent). The first case is simpler, since it only requires making the decision if the new call will be accepted and which of the available RATs will support it. In the second case, it is important to re-evaluate all active connections, given that another alternative RAT is now available. Since several parameters need to be taken into consideration, this task needs some time before reaching a decision. This processing time cannot be avoided, if sensible HO decisions are required (e.g., handing over a connection to a WLAN hotspot of a radius of 100 meters is not sensible for users moving with their vehicles in a speed of 80 Km/h since the connection will be handed over again to another cell and/or RAT in around 9 seconds).

When establishing a new connection or deciding a vertical HO, the algorithm evaluates the following five parameters:

1) The specific service requirements (i.e., service profile): Each service, even if it is adaptive to the bandwidth and QoS offered by each RAT, has always some minimum requirements from the radio connection in order to be successfully supported.

2) The MT specifications and capabilities (i.e., MT profile): Each MT may have a different set of radio interfaces, each one of them having particular requirements regarding the battery consumption, the CPU power, the available memory etc. Also, the battery duration and consumption are not constant and they depend heavily on the type and the number of connections and RATs that are active [14].

3) User's profile: The typical user is interested in neither the network technologies available nor the underlying difficulties to support seamless mobility. The user simply wants to get services easily, in a standard quality and at the least price possible. So, the user should be able to easily specify criteria in prioritised way, e.g. least cost, battery duration, QoS. This could be done via a graphical user interface on the MT, where the user could specify these criteria. Thus, a prioritisation of the alternative RATs based on user's preferences is feasible and can be part of the user's profile.

4) Network operator policies: The network operator wants to control the load of the proposed attachment points from the MT and also maximise if possible, the revenues. Though, it may be necessary to decide based on how to load balance the traffic between the different RATs, while at the same time taking into account the subscriber's preferences.

5) The MT location, speed and direction information: This is very important information, the knowledge of which may avoid the execution of unnecessary handovers. This could be the case of a fast moving user approaching a WLAN access point. There is no point in accepting this user to this WLAN, since in a few seconds he/she will be out of this coverage area.

All the above make quite clear that the selection of the radio interface to support a new call or a HO has to be based on several preferences and requirements, some of them conflicting with others. For example a user may prefer to pay the lowest price without sacrificing the quality of the received service, even in a congested network. So, this selection is mainly a trade-off between the user preferences and the operator's ones.

The algorithm proposed here is split in two cooperating parts. The first one runs in the MT while the second one in the core network. This approach has the advantage of easing the core network load on measurements and calculations for each HO case while minimizing the signalling exchange between terminals and network components. Thus, it leads to better utilisation of the precious resources at the radio interface. It looks like a mobile assisted handover case taken one step further, since the MT plays a more active and crucial role since it produces the set of acceptable RATs for each of its connections. Since the tendency in MT hardware characteristics is to be more powerful and having more battery autonomy, this approach does not stretch the MT. In the next two subsections, we present these two cooperating parts of the algorithm.

2.2. Algorithm Running in the Mobile Terminal

This part of the algorithm aims at prioritising the RATs for each connection separately. Its output is a list of user-preferred RATs for each one of the active connections of a MT. In order to accomplish this, it evaluates the first three of the parameters mentioned in the previous sub-section, i.e., the user, the terminal and the service profiles. The remaining two parameters are taken into account to the part of the algorithm running in the core network.

The algorithm running in the MT is shown in Figure 1. The whole procedure is initiated when one of the three following triggers occurs:

1) The MT detects a new alternative RAT with adequate signal strength/quality.

2) A new call is initiated from the MT user.

3) An "urgent" HO (regarding the time constraints) is imminent as a result of degradation of the radio link.

The algorithm treats the first two cases in the same way, since its main goal here is to create a prioritised list of the available RATs, based on the aforementioned parameters. A very important aspect affecting the RAT selection is when the various measurements are performed. These measurements may indicate the radio link degradation or the discovery of an alternative RAT. These are periodic measurements and very important to the whole procedure, but they are out of the scope of this paper.

The first possible trigger that is able to initiate the algorithm concerns the discovery of an additional alternative RAT in the vicinity of the MT, while the MT has active connections. This is described in the right part of Figure 1. If RSS measurements indicate that this new RAT has adequate signal strength, then the MT will create a list with the priorities of each RAT for each specific connection. If N is the number of active connections and M the number of available RATs, then this priority list takes the form of a two-dimensional matrix NxM, named pr_list in Figure 1. Then the MT reads the user profile and according to the user's preferences it constructs this matrix by giving a value to each cell (i, j) representing the priority of the j-th RAT for the i-th connection. The next step is to collect and average the values of all downlink (DL) measurements and evaluate them. The evaluation results may adjudicate that a certain RAT cannot fulfil the constraints that a specific type of service poses (stored in the service profile), such as bit-error rate and jitter. In such a case, this RAT will be eliminated from the priority list, by putting the value of zero in the specific cell of the pr_list. After the RAT elimination

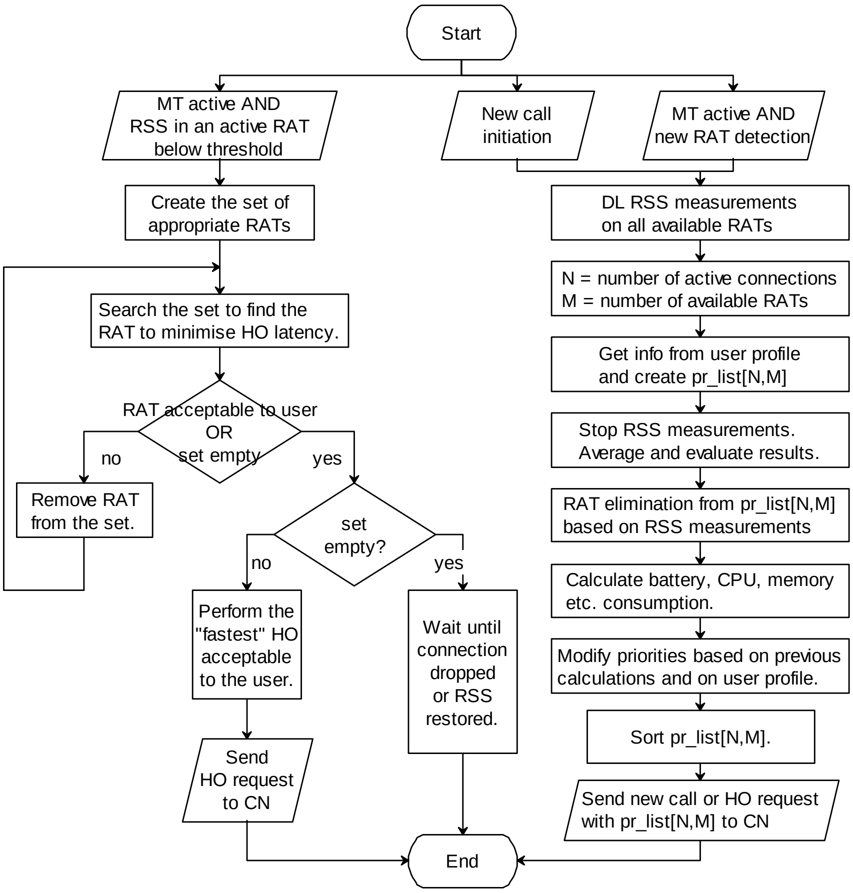

Start
MT active AND RSS in an active RAT below threshold
New call initiation
MT active AND new RAT detection
Create the set of appropriate RATs
Search the set to find the RAT to minimise HO latency.
RAT acceptable to user OR set empty
no
yes
Remove RAT from the set.
set empty?
no
yes
Perform the "fastest" HO acceptable to the user.
Wait until connection dropped or RSS restored.
Send HO request to CN
DL RSS measurements on all available RATs
N = number of active connections
M = number of available RATs
Get info from user profile and create pr_list[N,M]
Stop RSS measurements. Average and evaluate results.
RAT elimination from pr_list[N,M] based on RSS measurements
Calculate battery, CPU, memory etc. consumption.
Modify priorities based on previous calculations and on user profile.
Sort pr_list[N,M].
Send new call or HO request with pr_list[N,M] to CN
End

Figure 1. Algorithm running in the mobile terminal.

phase, the algorithm estimates all factors reflecting the cost of each candidate RAT selection to the MT's characteristics, i.e. the battery consumption, the CPU and memory requirements, etc. This evaluation takes place after the RAT elimination step, in order to avoid evaluations of a RAT that will be rejected. This estimation has to be combined with the user preferences. This means that according to the importance that each one of these factors has for the particular user, the algorithm treats it in a different way. The importance is indicated in the user profile stored in the MT. Thus, the pr_list is modified again. The final step of this part of the algorithm is to sort the matrix per connection, in descending order as far as priorities are concerned, starting from the one better satisfying the user. This puts pr_list in its final format and it is the final priority list sent to the CN. There, the corresponding part of the algorithm is executed, having as input this prioritised list, as described in the following sub-section.

In order to clarify how this part of the algorithm works, let us consider an example and follow each step of the algorithm. We consider a MT being able to simultaneously communicate via UMTS and WLAN. At a given time, it has three active connections, thus M=3. Also, it can communicate via either UMTS or two alternative WLANs, i.e. N=3. In this case, the information stored in the user profile is used to build the initial priority list, such as the one shown in Table 1. For connection 1, UMTS has the highest priority, whereas WLAN1 the lowest. For connection 2, WLAN2 is set to zero, indicating that this type of access network is not acceptable to the user for the specific service, for a reason such as monetary cost or QoS offered. The next step in the algorithm is the evaluation of radio signal measurements. Let us suppose that UMTS does not fulfil the service requirements of connection 1. So, it is eliminated from the list, as shown in Table 2. The next step is the evaluation of additional parameters, such as battery duration. In our example, the user wants to maximise the duration of the battery as long as possible. A simple solution could be to support all connections by WLAN1, since it is the only RAT adequate for all three connections. This means that the list will be modified, by finding the RAT with the maximum priority for each connection and add it to the corresponding priority of WLAN1. Table 3 shows the result of this step, where WLAN1 has the highest priority for all connections. The sorting of the list is the last step of this part of the algorithm. This is shown in Table 4, for our example. So, for connection 1 there are two alternative RATs, with WLAN1 having the highest priority. Connection 2 has again two RATs, while connection 3 has three. This is the final priority list sent to the CN.

Table 1. Priority list from user profile.

	UMTS	WLAN1	WLAN2
Con. 1	3	1	2
Con. 2	1	2	0
Con. 3	2	1	3

Table 2. Priority list after RAT elimination.

	UMTS	WLAN1	WLAN2
Con. 1	0	1	2
Con. 2	1	2	0
Con. 3	2	1	3

Table 3. Priority list (extra evaluation).

	UMTS	WLAN1	WLAN2
Con. 1	0	3	2
Con. 2	1	4	0
Con. 3	2	4	3

Table 4. Final priority list (pr_list).

Con. 1	WLAN1	WLAN2	-
Con. 2	WLAN1	UMTS	-
Con. 3	WLAN1	WLAN2	UMTS

The second possible trigger to start the algorithm running at the MT is a new call initiation. In this case the algorithm has to prioritise all RATs from the one providing best support for this specific connection type, to the one providing the worst, but still with adequate QoS. This trigger is handled in the same way as the first one, described in the previous paragraphs. The difference is that in this case, the prioritisation has to be done only for one connection. So, in this case, i.e. a new call, if N is the number of the active connections of the involved MT, then N=1.

Finally, the third trigger to initiate this part of the algorithm is an "urgent" HO case. The MT has at least one active connection and the radio signal strength measurements indicate that one link deteriorates under some specified threshold. This is an urgent HO case, where the HO latency becomes the most critical factor. Thus, no evaluation of the different parameters is performed, since this, along with the signalling introduced, increase the time required for HO completion. In this case, the HO type providing the least latency is chosen, if only it is acceptable to the user. This can be checked with the user profile. This is shown in the left part of Figure 1. When the MT realises that an "urgent" HO is imminent, it creates a set of all alternative RATs that can adequately support the particular connection. Then, it identifies within this set the RAT minimising the HO latency. This choice is dependant to the available RATs and the architecture of the heterogeneous network (such as loose or tight coupling). Then, the MT checks if the chosen RAT is acceptable according to the user preferences. If it is, a HO request message is send to the CN to execute a HO to this particular RAT. In a different situation the algorithm continues with the next RAT, until either one RAT acceptable to the user is found or there are no more RATs in the set. In the latter case, no HO is performed and the connection may be either terminated or normally continued in case that the RSS is restored back to acceptable levels.

2.3. Algorithm Running in the Core Network

This part of the algorithm starts when it receives the output from the corresponding part running at the MT. The CN takes the final decision, based on the last two parameters mentioned in subsection 2.1 and on the outcome of the algorithm at the MT. Thus, it is based on policies determined by the operator and on velocity, location and position of the MT. Uplink radio channel measurements indicating the quality of the uplink bearer are also taken into consideration as in any HO case. RAT specific parameters, such as the channel and/or UMTS Orthogonal Variable Spreading Factor (OVSF) codes availability play a role as well [15]. Some parameters change dynamically, so the core network has to acquire updated information either periodically, or after certain stimulus and message exchanging. This information gathering and/or message exchanging is an important issue, but out of the scope of this paper (a discussion on this issue can be found in [12]).

This part of the algorithm is shown in Figure 2 and it is executed either when a new call is going to be established or when there is a request for a HO from the MT. In a new call initiation, this algorithm is part of the call admission control procedure that is responsible for the load control in the entire heterogeneous network. In a HO request, the HO can be either "urgent", i.e., due to radio signal strength degradation, or initiated to better support the existing connections. Our focus is on the latter case, where a vertical HO is initiated to improve the satisfaction of user preferences, and is the result of changes in the number of RATs that the MT can reliably communicate with. Both HO types are indicated by a HO request message from the MT to the CN.

First, we consider the case of the "urgent" HO. This is shown in the right part of Figure 2 and it corresponds to the third and last trigger of the part of the algorithm at the MT, as described in the previous subsection. There, the outcome was a HO request message from the MT to the CN. This request indicates a HO due to radio link degradation along with the target RAT decided at the MT algorithm. This decision was based on the architecture of the heterogeneous network. Then, the CN reserves the appropriate resources and informs the MT about the HO execution.

In the case of a HO request due to a new RAT detection from the MT or a new call initiation, the time constraints are not as tight as in the "urgent" HO case.

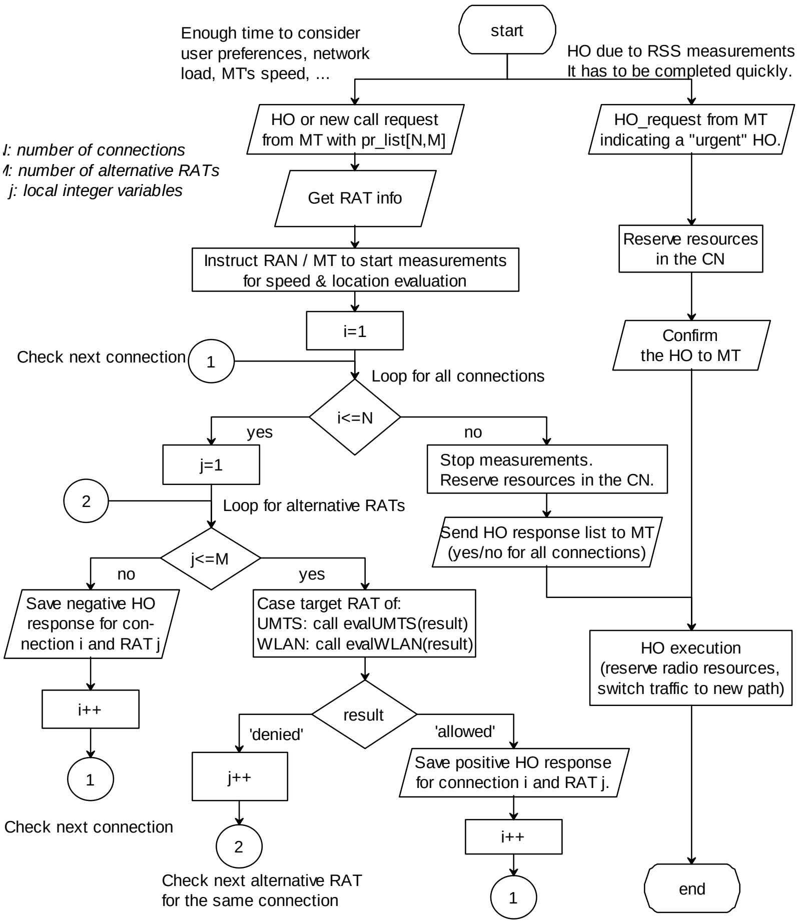

start
Enough time to consider
user preferences, network
load, MT's speed, ...
HO due to RSS measurements
It has to be completed quickly.
HO or new call request
from MT with pr_list[N,M]
HO_request from MT
indicating a "urgent" HO.
N: number of connections
M: number of alternative RATs
j: local integer variables
Get RAT info
Reserve resources
in the CN
Instruct RAN / MT to start measurements
for speed & location evaluation
i=1
Confirm
the HO to MT
Check next connection
1
Loop for all connections
i<=N
yes
no
j=1
Stop measurements.
Reserve resources in the CN.
2
Loop for alternative RATs
Send HO response list to MT
(yes/no for all connections)
j<=M
no
yes
Save negative HO
response for connection i and RAT j
Case target RAT of:
UMTS: call evalUMTS(result)
WLAN: call evalWLAN(result)
HO execution
(reserve radio resources,
switch traffic to new path)
i++
result
'denied'
'allowed'
1
j++
Save positive HO response
for connection i and RAT j.
Check next connection
i++
2
Check next alternative RAT
for the same connection
1
end

Figure 2. Algorithm running in the core network.

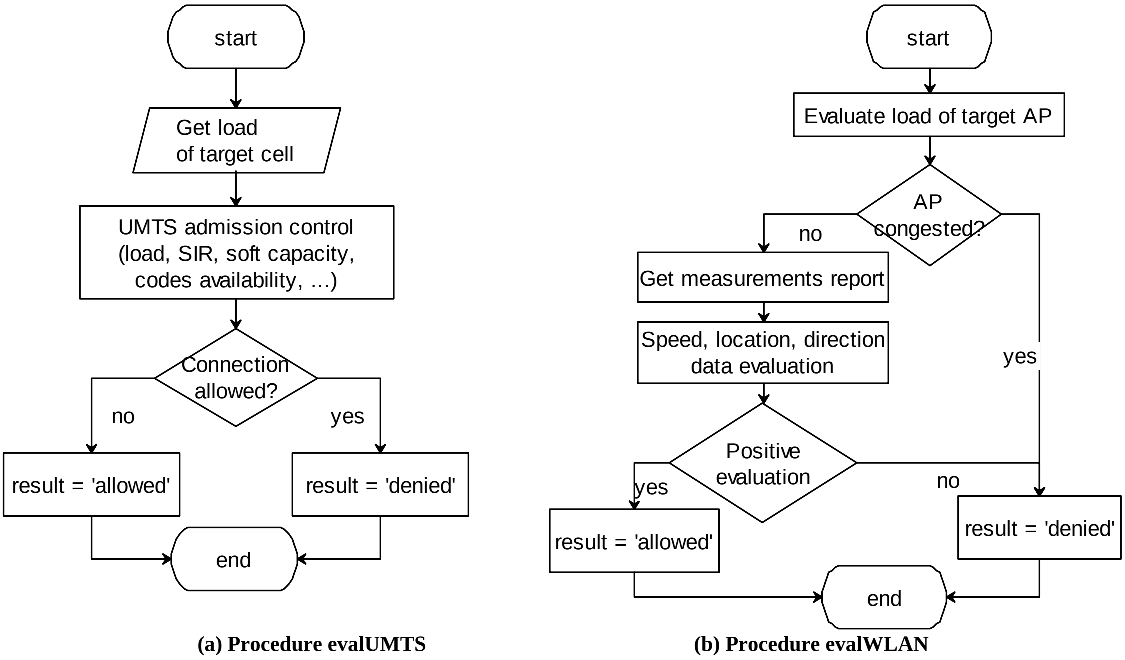


Figure 3. High level evaluation procedures.

This case is shown in the left part of Figure 2. There is enough time to evaluate network condition (e.g. congestion) and user's speed, location etc. First of all, the CN receives the HO request message including the RAT priority list, which was the outcome of the part of the algorithm in the MT. As described in the previous subsection, this priority list can be seen as a two-dimensional matrix NxM, where N is the number of the active connections of a MT and M the number of alternative RATs. In the case of a new call initiation N=1 and in the one of a HO request N≥1. Nevertheless, both cases are treated the same way by this part of the algorithm. Then, the CN gets all information related to the HO for all the involved RATs, such as the coverage area, the location of the access points (APs) and base stations the MT communicates with.

The next step is the initialisation of the procedures to support the MT's velocity and location estimation. These are important in a heterogeneous network, since they influence an inter-RAT HO decision. This is due to the fact that in some cases there may be no point for an inter-RAT HO, because of high speed, direction of movement, location of the MT or small coverage areas of a certain RAT. Thus, in these circumstances the MT will reside in the RAT coverage for a very short time, and then another HO, "urgent" this time will be required. So, some specific thresholds and rules have to be defined. These could have the form of simple rules such as "if velocity greater than z m/sec" or "the MT's distance from the AP is greater the x% of the cell's radius and it is moving away from it with velocity at least y m/sec" etc. Thus, it is clear that this kind of information will help the CN make a better decision and avoid useless HOs and thus reducing the total amount of signalling. There are several proposals for estimating the velocity of a MT. Some of them are based on estimations of the maximum Doppler frequency [16,17]. In 3GPP some work has been done for estimating the geographical position and optionally the velocity of the MT in UMTS, through radio signal measurements [18]. The particular method of velocity and location calculation is out of the scope of this paper. This kind of information is evaluated in the CN, since the appropriate data are not available to the MT.

The next step is a nested loop. The outer loop is for each one of the active connections and the inner loop for each alternative RAT for a specific connection i (I=1,…,N), where N is the number of active connections and M the number of alternative RATs for each connection. Thus, the algorithm evaluates the request for each connection separately. This evaluation is heavily dependent on the RAT type. As an example we consider UMTS and WLAN as alternative technologies. Figures 3(a) and 3(b) present some high level descriptions of UMTS and WLAN evaluation procedures.

In Figure 3(a), the CN has the information on the load of the target cell and the usual UMTS call admission control algorithm is executed. If the result is that the new connection can be supported then this procedure returns the result 'allowed', else the result 'denied'. In Figure 3(b), the high level procedure for evaluating a vertical HO to WLAN is shown. The CN has the information on the load of the target AP and if it is congested, the procedure returns the result 'denied'. Else, since the coverage area of an AP is rather small, the core network has to take into account the velocity, the location and the direction of the

MT to make a decision. So, it collects all measurement information, evaluates it and then, if the result indicates that the new connection can be supported the procedure returns the result 'allowed', else the result 'denied'.

After the evaluation of all M alternative RATs for each connection, the result of the corresponding procedures indicate if the HO will be executed and to which RAT. This is done for each one of the N connections (outer loop in Figure 2). When this is completed, any measurements to support speed and location estimation are stopped and the HO execution phase starts.

3. Simulations and Qualitative Analysis

3.1. Model, Assumptions and Parameters of the Simulations

In order to evaluate the performance of the proposed algorithm, a simulation model has been created using the Network Simulator-ns-2 [19]. Two alternative RATs are considered in this model. The first one represents the UMTS and it has global coverage, while the second one represents the WLAN and it covers a smaller portion, as shown in Figure 4. The general assumptions for the model used are that the MTs are uniformly distributed in the coverage area, their movements are not correlated and their direction is uniformly distributed. Also, all MTs are capable of having simultaneous active connections over UMTS and WLAN. Moreover, all MTs have both interfaces active throughout the whole simulation, so, for sake of simplicity, the battery consumption was not considered.

Each MT has a certain residency time in each of the RATs involved and in each one of the coverage areas shown in Figure 4, namely area 1 and area 2. In area 1 there is only UMTS coverage, whereas in area 2 there is both UMTS and WLAN coverage. The residency time is exponentially distributed. In the UMTS network all users have the same residence time, whilst in the WLAN the fast moving users have much lower residence time, due to the smaller coverage area. For each MT the mobility model shown in Figure 5 is used. This is a two state Markov process, representing the movement from the two coverage areas shown in Figure 4. Thus, the MT can be in an area having only UMTS coverage or in an area of double coverage. When the residence time expires, another state is chosen, according to the shown probabilities. These probabilities are related to the percentage of the WLAN coverage of the whole area. Only when a user changes coverage area the simulation model triggers a HO. In this way, we consider only the vertical HOs since these are important in our measurements.

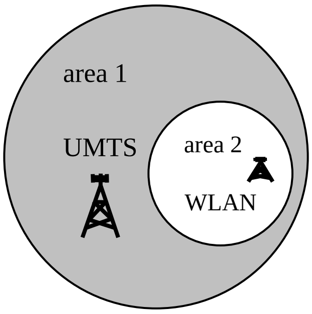


Figure 4. Assumed coverage area of RATs.

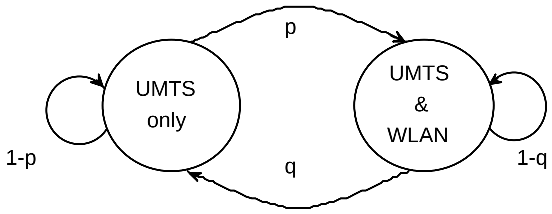


Figure 5. Mobility model.

The new calls arrive in the whole system as a Poisson process with an inter-arrival time that is exponentially distributed and a mean rate of λ calls per hour. Each one of the new connections belongs to a specific service type according to its requirements on bandwidth, call duration, delay, jitter etc. In the simulations, the four traffic classes of UMTS were considered to classify each new connection [20], but this can be easily adapted to any other classification:

1) TC1: QoS conversational, e.g. voice over IP
2) TC2: QoS streaming, e.g. video/audio streaming
3) TC3: QoS interactive, e.g. www browsing
4) TC4: QoS background, e.g. FTP downloading

When a new call enters the system, it is classified as TC1, TC2, TC3 or TC4 according to some respective probabilities p(TC1), p(TC2), p(TC3) or p(TC4), shown in Table 5, so that $\sum p(TC_i)=1$.

Furthermore, each traffic class poses different constraints on the simulation model regarding the bandwidth required and the mean duration of each call μ_i (i=1,2,3,4). The duration of each call has an exponential distribution with mean value μ_i (i=1,2,3,4). According to the traffic class that the new connection belongs to, there are the appropriate requirements on the bandwidth. For the simulation model, some typical mean values were considered for the bandwidth, namely BW_i (i=1,2,3,4). All these are shown in Table 5.

The user profile that describes the preferences of the user for each specific service and network is stored in the MT. Such information is semi-static, and does not change during a simulation run. For sake of simplicity three user profiles have been considered, namely UP_i (i=1,2,3). For the simulation model, three initial profiles have been considered: The first one aiming at the low cost of the supported services, the second one at the best quality of the offered services and the third one having the least energy requirements. Here we consider that:

1) WLAN has a lower cost per time unit or data unit, for all TCs.

2) UMTS has less power requirements in a "mixed" usage scenario, involving many TCs and connections [14].

3) UMTS offers guaranties for QoS for all TCs. Especially for TC1 and TC2, where time delay and jitter is

critical, UMTS should be considered as the first choice for QoS provisioning. For TC3 and TC4, where time constraints are not that strict, the higher bandwidth of WLAN makes it a better choice for these two TCs.

At the initialisation phase of each simulation run all users are distributed in one of the available user profiles, according to the probabilities shown in Table 5, so that $\sum p(UP_i)=1$.

In Table 5 we resume all the parameters that are common in every simulation run. In the following subsections, we present several scenarios executed in the simulation environment, in order to evaluate the algorithm performance. Firstly we evaluate the behaviour of the algorithm when we increase the new call rate arrival λ and the system becomes overloaded. Secondly, the available bandwidth is altered. As a third step, we increase the ratio of fast moving users over the total number of them. Finally, the last evaluation is done by changing the portion of the whole coverage covered by the smaller coverage RAT. All these are explained in detail in the following four subsections respectively. For each one of these four test scenarios, our focus is on three different metrics in order to evaluate the algorithm performance:

1) The first metric involves the probabilities for new call blocking $P(C_{block})$, handover blocking $P(HO_{block})$ and call dropping $P(C_{drop})$. A new call is blocked when a corresponding request is rejected. This results in a call not being initiated. A handover is blocked when a vertical handover request is rejected. This results to an abnormally terminated existing call. An outgoing call is dropped when an unsuccessful vertical handover occurs.

The three probabilities representing the first metric in our simulations are given by the Equations (1), (2) and (3).

Table 5. Simulation parameters.

Number of MTs	24
Maximum number of active connections per MT	4
p(TC1)	0.40
p(TC2)	0.15
p(TC3)	0.25
p(TC4)	0.20
area 2 residence time (slow users)	100 sec
μ_1	180 sec
μ_2	300 sec
μ_3	900 sec
μ_4	900 sec
BW_1	64 kbps
BW_2	384 kbps
BW_3	120 kbps
BW_4	120 kbps
p(UP1)	0.50
p(UP2)	0.25
p(UP3)	0.25

$$P(C_{block}) = \frac{c_r - c_s}{c_r} \tag{1}$$

$$P(HO_{block}) = \frac{h_r - h_s}{h_r} \tag{2}$$

$$P(C_{drop}) = \frac{c_d}{c_s} = \frac{h_r - h_s}{c_s} \tag{3}$$

where:

c_r is the number of new call requests in the whole system during a full simulation run,

c_s is the number of new call requests that has been successful and resulted in initiated calls,

c_d is the number of calls that have been started but during the connection they have been dropped due to an unsuccessful vertical HO,

h_r is the number of vertical HO requests sent from all MTs to the core network,

h_s is the number of successfully completed vertical HO.

2) The percentage of the new calls that have been accepted in the RAT indicated as the user's first preference, over the total number of new call requests, namely *nc (pref1)*. The corresponding percentage is measured for the connections that have been vertically handed over to another RAT, namely *ho (pref1)*. These two metrics are used as a guide to evaluate the user's satisfaction according to the preferences that each one of them stores in its profile. They are described by the two following Equations (4) and (5):

$$nc(pref1) = \frac{nc_1}{nc_1 + nc_2} \tag{4}$$

$$ho(pref1) = \frac{ho_1}{ho_1 + ho_2} \tag{5}$$

where:

nc_1 is the number of new call requests served by the RAT indicated as the first user preference,

nc_2 is the number of new call requests served by the RAT indicated as the second user preference,

ho_1 is the number of vertical HO requests served by the RAT indicated as the first user preference,

ho_2 is the number of vertical HO requests served by the RAT indicated as the second user preference.

3) The number of vertical HO requests messages sent from the MTs to the CN. Also, the number and the percentage of the vertical HO requests that have been dealt with in the MT, from the corresponding part of the algorithm running there. Thus, we try to evaluate the benefits of splitting the algorithm functionality in two parts, instead of one, not only as a measure to alleviate the core network from unnecessary processing, but more importantly, to see if we succeed to minimise the signalling load, especially at the radio interface, and if so, to quantitatively evaluate this.

Each simulation run was made for a simulation time of 10 days, so that it reaches a stable state. Then, 10 different runs with different seed each time were executed and all results from these runs were averaged in order to avoid any non-typical behaviour of the model.

3.2 Modifying New Call Arrival Rate λ

In this run all parameters mentioned in the previous subsection are maintain constant as shown in Table 5. The only parameter that changes is the total new call rate produced by the simulation model. This rate λ is measured in new calls per hour for the whole system. Let us see how the three metrics we focus in are influenced by the rate λ.

In Figure 6 the new call blocking, the HO blocking and the call dropping probabilities are shown as a function of λ. It is quite obvious that the system becomes overloaded for λ>80, as the new call blocking probability rises highly, for the chosen initial values. What is interesting here is to see the two other metrics, the user satisfaction related and the HO requests messages, even in these overloading conditions.

In Figure 7 we see the percentage of the new calls and of the HOs that have been served by the RAT indicated in the first user preference. The remaining calls/HOs have been served by the second preference of the user. Here, we observe that as far as the new calls are concerned, this percentage is extremely high, about 95%, and it does not significantly drop (less than 2%), even for very high load system cases (λ≥90). For the HO case though, the respective numbers are not that high, about 86%, which is still very high, and remains almost constant. The conclusion here it that the algorithm performs very well concerning the satisfaction of the first user preference for new calls and HOs, almost irrespectively of the load posed in the network. It performs very well even in an overloaded system, where the mean call rate leads to high new call blocking probability.

The last metric concerns the number of HO requests messages sent from the MTs to the core network and the number and percentage of HO cases dealt with in the MTs, as a function of λ. This is shown in Figures 8(a) and 8(b)

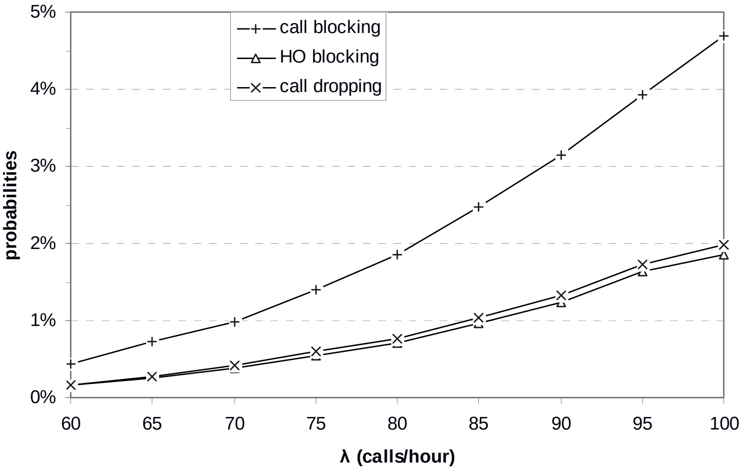


Figure 6. Call/HO blocking and call dropping probabilities.

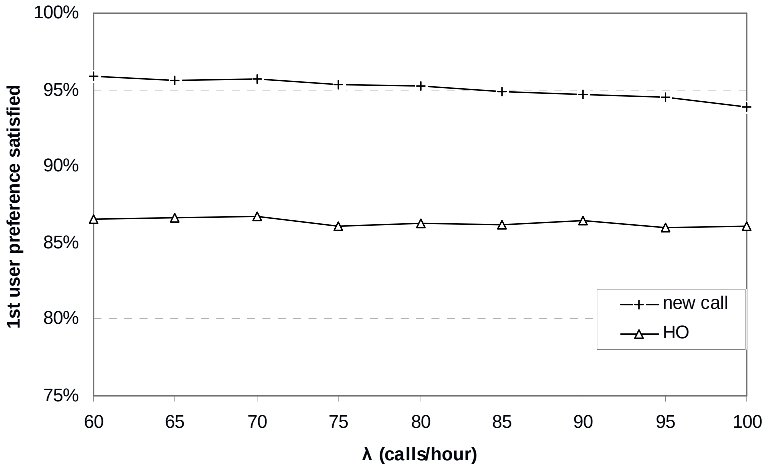
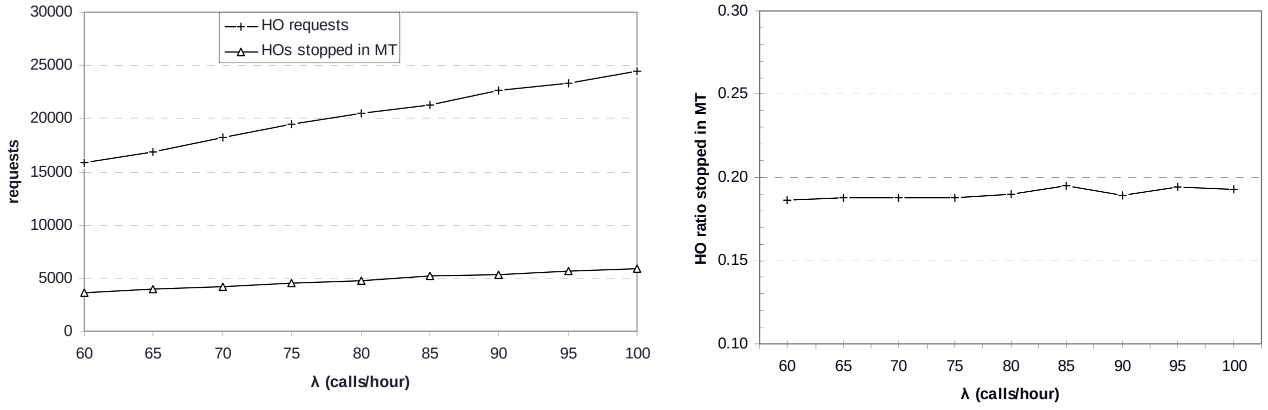


Figure 7. New calls and HOs served at the 1st user preference RAT.



(a) (b)

Figure 8. HO requests messages sent to the CN and stopped in the MT.

respectively. As seen in these figures, as λ increases so does the total number of HO requests sent to the core network and proportionally the number of HOs dealt with in the MTs (figure 8(a)). Though, it is really interesting that the percentage of HOs that are dealt with by the MTs is quite high and stable, around 19% for all λ values (figure 8(b)). This can be simplified by saying that almost one out of five HO requests along with the relative signalling is avoided in the radio interface, because of the splitting of the algorithm in two co-operative parts. This seems to justify this design choice. The small peak near the centre of Figure 8(b) is negligible (less than 0.5%).

3.3. Modifying the Available Bandwidth

In this scenario, we wanted to examine how the proposed algorithm performs regarding the available bandwidth of the RATs involved. In the simulation model, WLAN covers a smaller area and has more resources available. This makes UMTS the more stressed RAT regarding the available resources. Thus, for simplicity reasons, we do not alter the bandwidth of the RAT with the more available resources, i.e. WLAN. This is set equal to 11 Mbps. The available bandwidth for UMTS is changed from 2 up to 2.8 Mbps. Furthermore, we keep constant all the parameters shown in Table 5.

Considering the first metric which is the blocking probabilities, the results are shown in Figure 9. As expected, when the offered bandwidth is increasing, all blocking and dropping probabilities are reduced. What benefits more is the new call blocking probability which is decreased almost 10 times. Also, we see that for the given model, when the available UMTS bandwidth drops below 2.4 Mbps the whole system is highly stressed, given that WLAN bandwidth is constant. Nevertheless, all relevant probabilities drop significantly, when the UMTS available bandwidth supersets 2.4 Mbps. This first metric does not reveal anything new for the performance of the algorithm.

Considering the second metric, that is the percentage of new calls and HOs accepted in the first user preference RAT, the results are shown in Figure 10. Here we observe that the vast majority of new calls are served by the RAT indicated as the first user preference. The same applies to the HOs case, even though the percentages are not that high. Also, the more bandwidth UMTS has, the better the first user preference is satisfied. New calls seem to take more advantage of this increase, by increasing about 2%, whereas HOs are nearly constant around 86%. But the general picture is that this metric is not really influenced by the bandwidth available and remains in quite high values, showing the algorithm provides for a great user satisfaction, even when the network is stressed due to lack of resources, i.e. UMTS bandwidth less than 2.4 Mbps, as depicted from the previous metric.

Regarding the third metric, the results are shown in Figure 11. In Figure 11(a), we see that the number of HO messages sent to the core network increases about 10% whereas the available bandwidth of UMTS increases 40%. This is due to the fact that by increasing the available bandwidth, more new calls are accepted in the system, thus, more HOs are performed. On the other hand, the number of the HO messages that have not been sent to the core network, because the corresponding HOs have been dealt with in the MT, is relatively constant, about 19% as shown in Figure 11(b) (with a small fluctuation of 0.5%) and does not seem to be really influenced by the available bandwidth. This is quite positive if we consider that this percentage is quite important, since it means that about one fifth of the vertical HO requests are not sent to the core network even when the available bandwidth is quite limited and the system overloaded, such as in the case of UMTS having 2 Mbps as shown in Figure 9. That seems another point to justify the splitting of the functionality of the algorithm.

3.4. Modifying the Coverage of Alternative RATs

In this scenario, we investigate the algorithm behaviour according to the coverage relation between the two alternative RATs. Since UMTS is considered to have global

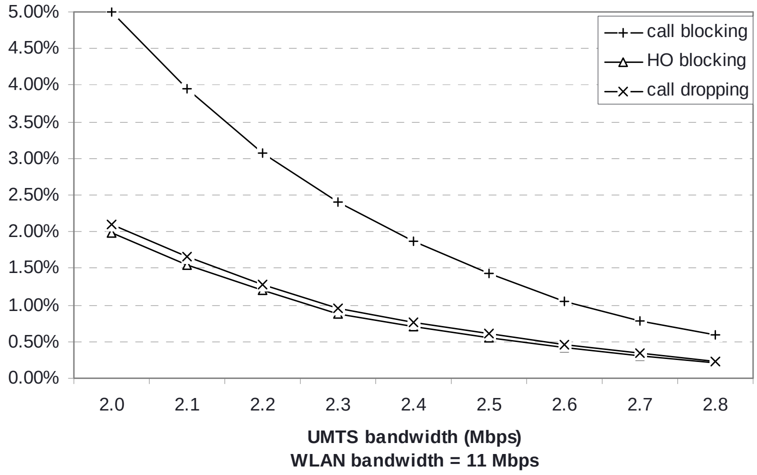


Figure 9. Call/HO blocking and call dropping probabilities.

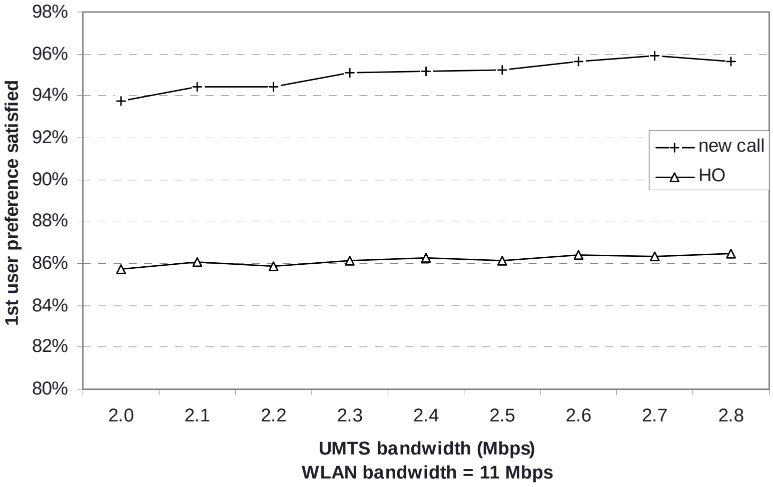


Figure 10. New calls and HOs served at the 1st user preference RAT.

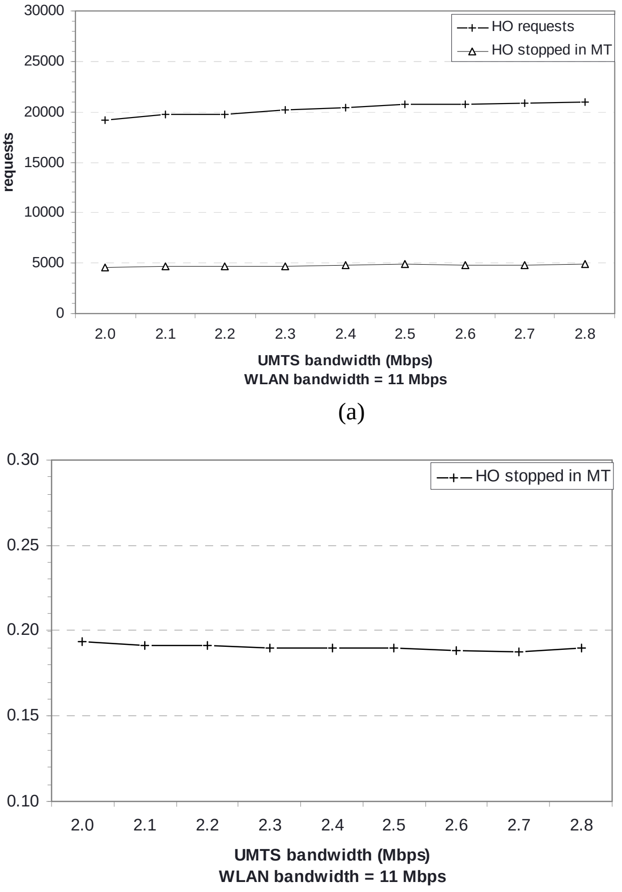
coverage, i.e. 100% of the geographical area, WLAN has a variable one from 5% up to 45% of the whole area. The new call arrival rate λ equals to 80 calls per hour. All other parameters of the simulation model remain as shown in Table 5.

In Figure 12 we see the results regarding the first performance metric, i.e. the new call blocking, the HO blocking and the existing call dropping probabilities. The new call blocking probability is decreasing with relation to the increase of the WLAN coverage availability. It drops quite below 2% when WLAN coverage exceeds 20%. The latter means that there are more resources available at a broader geographical region. Thus this is an expected behaviour, since the whole system’s load is constant. Another thing we notice is that the HO blocking probability is decreasing as well, but not with such a big slope as the call blocking. This can be explained after analysing the simulation results and is due to:

1) The number of new call requests is almost constant in all simulation runs but the number of the calls not being admitted to the system (blocked calls) is dropping according to the WLAN coverage. As long as the latter increases the former decreases. According to Equation (3), this means that the numerator is continuously decreasing while the denominator does not significantly change, presenting the result of new call blocking probability of Figure 12.

2) The number of vertical HO requests increases with the increase in WLAN coverage. After an initial big raise, this number keeps on growing but with a smaller rate. On the other hand, the number of the vertical HOs rejected due to the lack of resources is increasing as the WLAN coverage increases, but after 25% it tends to stabilise. According to Equation (4), this means that the nominator is relatively constant while the denominator increases, but with a decreasing rate. This explains why the HO blocking probability is decreasing with a smaller slope than the new call blocking probability.

Finally, in Figure 12 again, the probability of an ongoing call to be dropped remains quite small (below 1%) and lies in a small interval of 0.5%. Firstly it seems to



(a)

(b)

Figure 11. HO requests messages sent to the CN and stopped in the MT.

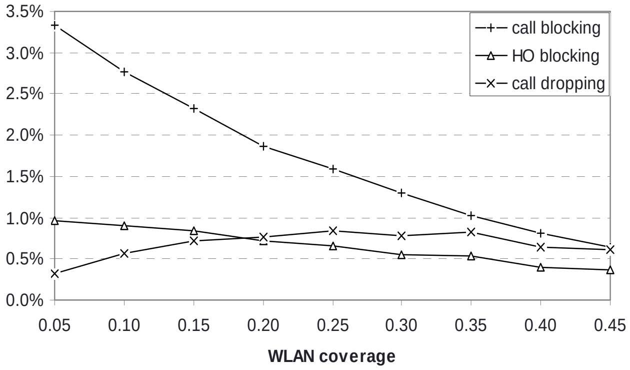


Figure 12. Call/HO blocking and call dropping probabilities.

augment but then it stabilises. This is explained if we check again with Equation (3) and see some numerical results from the simulations. When the coverage area of WLAN increases, so does the number of vertical HO requests, but with a big rate. This is not true though for the number of blocked HOs. This number is augmenting at the beginning (up to 0.25), due to the limited resources of the whole system. When WLAN coverage exceeds a certain point (about 0.25), the available resources in the simulation system are adequate, so that the number of rejected HOs tends to stabilise and then to decrease. If we consider this fact and Equation (3) we can understand the curve of call dropping probability in Figure 12.

In Figure 13, the second metric is depicted. We observe that as the coverage of WLAN increases, there is a slight rise in the percentage of new and of handed-over calls served by the RAT indicated as the first user preference. This means that when the area where only UMTS connection is available is decreasing, the same happens to the percentage of the new calls served by the first user preference. At a first thought, this might seem strange. It is explained if we consider Equation (4) and the numerical results from the simulations. The successful new calls entering the system are increasing along with the WLAN coverage. This means that both nc_1 and nc_2 in Equation (4) rise, but the denominator does so with a higher rate, as it is the sum of nc_1 and nc_2. In any case, the difference is not significant and lies in all cases below 3%.

As far as the third metric is concerned, we see the results in Figure 14. As the WLAN is available in a broader geographical region, the HO requests for a vertical HO are increased. This is very logical, since the probability p in the mobility model in Figure 5 increases

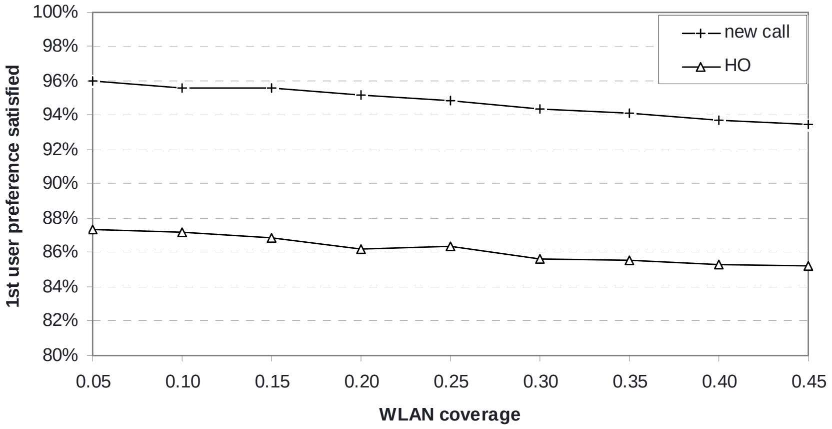
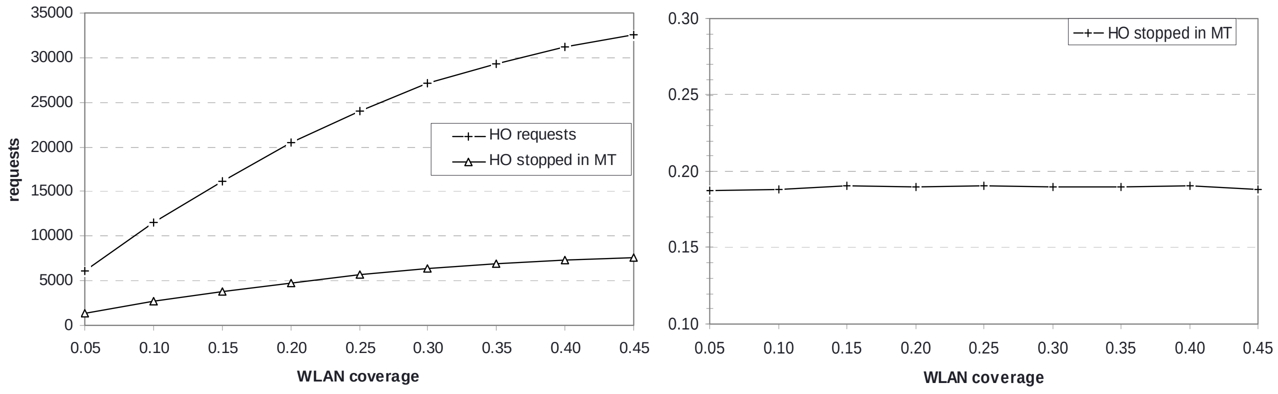


Figure 13. New calls and HOs served at the 1st user preference RAT.



(a) (b)

Figure 14. HO requests messages sent to the CN and stopped in the MT.

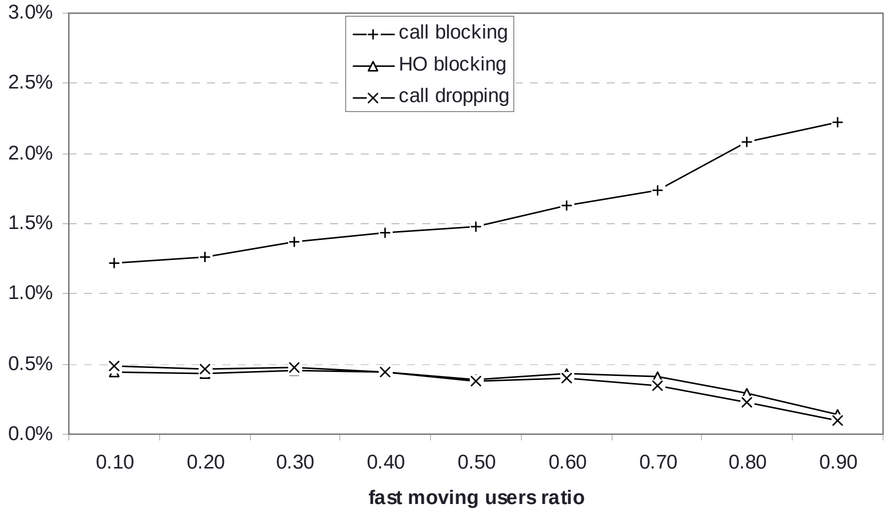


Figure 15. Call/HO blocking and call dropping probabilities.

according to the area of WLAN coverage. The same applies to the number of the HO requests that are dealt with in the part of the algorithm running at the MT. Relatively both numbers in Figure 14(a) augment by the same ratio, and this is the reason why the percentage of the HO requests stopped in the MT lies about 19% in Figure 14(b). Thus, as in the previous scenario, we see that this percentage is again both quite high and quite irrelevant to the coverage area.

3.5. Modifying the Fast Users Ratio

In this last scenario, we evaluate the algorithm according to the percentage of the mobile users that move faster than the specified velocity threshold beyond which a user is considered to move too fast to be eligible for service in WLAN. This means that the part of the algorithm running in the core network and specifically the WLAN evaluation function in Figure 3(b) will return a negative answer to either a new call or a vertical HO request. The coverage of WLAN is one fifth of the whole system area, the ratio of fast moving users fluctuate from 0.10 to 0.90–such as in the case of highway coverage–and all other simulation parameters are as shown in Table 5.

Let us have a look at the first metric of our evaluation. This is shown at Figure 15. The first observation is that the blocking probability of a new call is increasing with the increase in the fast users' ratio. This is expected since fast users are excluded from WLAN. This means that for all those users only UMTS is an option either for a new call or a vertical HO. Thus, UMTS has to serve an all increasing number of users thus it is highly stressed and this limits the number of new calls accepted in the whole system. On the other hand, the blocking probability of a vertical HO has the opposite tendency from the new call blocking probability. The same applies to the dropping probability of an ongoing call. Both remain relatively constant at the beginning and then tend to slightly decrease. The reason is that since the new call blocking probability is increasing, the number of new calls successfully entered the system lowers. This means less vertical HO requests and far less blocked HOs. From Equations (2) and (3), we then understand the HO blocking and call dropping curves.

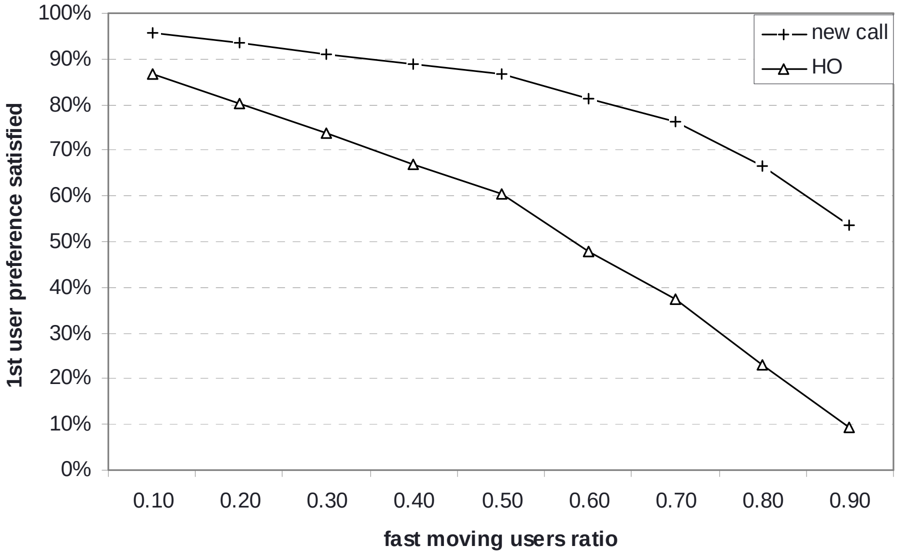
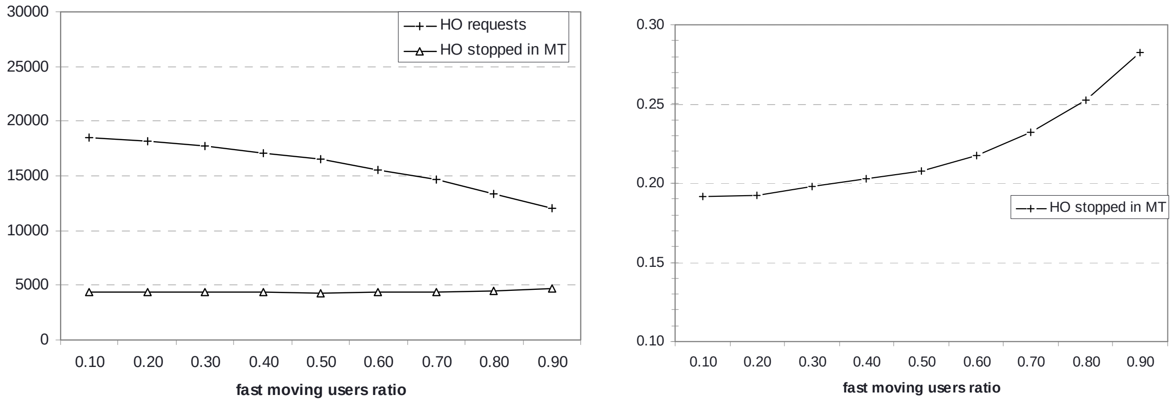


Figure 16. New calls and HOs served at the 1st user preference RAT.



(a) (b)

Figure 17. HO requests messages sent to the CN and stopped in the MT.

The next metric regarding the user satisfaction related to the number of new calls and HOs served by the RAT indicated in the first user preference, is shown in Figure 16. We see that both of them decrease when the ratio of fast users is increasing. This is because when more and more users are moving too fast to be accepted in WLAN, the latter is not an option for the majority of the users and only UMTS is accepting them. Thus, the algorithm cannot provide an alternative RAT and there is no option to choose from. Though, this seems a limitation of the number of alternative RATs and not of the algorithm itself. If more alternative RATs where available, we expect that the situation would be better.

The last metric regarding the number of HO requests sent at the core network and those dealt with in the MTs is shown in Figure 17. Here we see that when the ratio of fast moving users increases, the number of HO requests messages from the MTs to the core network is decreasing (Figure 17(a)). This, as explained earlier, is due to the decreasing number of new calls entering the system. On the other hand, what is really appealing here, is the fact that the number of HOs that are dealt with in the part of the algorithm running at the MTs is almost constant (Figure 17(a)) and as a percentage always increasing from 19-29% (Figure 17(b)). This high percentage is very important if we consider that all these could have been corresponding HO requests messages over the radio interface to the core network.

The simulation results show some very interesting results. First of all, in all tested scenarios they justify the splitting of the algorithm functionality in two parts. In every case, a significant number of the HO requests are treated by the corresponding part of the algorithm running at the MT, even in cases where the simulated system is overloaded. Secondly, the algorithm provides with high percentage of user satisfaction, with the exception of when the ratio of fast moving users is very high. On the other hand, the results also show that the first priority of the algorithm is not to provide for the best load balanced system. This is the reason why in some scenarios, the blocking probabilities may be quite high.

4. Conclusions and Future Work

This paper presents a network selection solution for integrated UMTS and WLAN mobile networks. It has three main differences in its design from related proposals in the field. Firstly, it is implemented as an algorithm that can be easily tested and implemented. Secondly, it evaluates each active connection of every MT separately, as a different HO case. This has the advantage that the more appropriate RAT according to specific parameters will support each connection. This decision though is a tradeoff between contradicting criteria, such as the user preferences and the network policies and management. The last difference is that the functionality is split in two distinct and co-operating parts. The first part is an algorithm running on the MT and produces a prioritised list of the preferred RATs per connection, taking into account the user preferences and the MT status. The second part is an algorithm running in the CN. It receives the prioritised list from the MT and based on that, it takes the final decision upon which HO is allowed or not, and in which RAT. This decision is made according to the RAT type, the network load conditions and the MT's movement characteristics. This split of the algorithm functionality alleviates the CN from some calculations and precipitates the HO decision. Also, because of the kind of pre-processing done in the MT some requests towards the core network are avoided, if considered not applicable. This results in reducing the signalling load at the radio interface.

A simulation model of a network comprising both UMTS and WLAN coverage areas has been implemented. Several scenarios have been run showing that the algorithm provides high user satisfaction, it decreases the messages required for the vertical handovers in the whole network, and it alleviates the core network from the processing of many vertical handover requests. This comes with the price of augmenting some blocking probabilities and thus allowing lower total traffic in the whole system. This is something we will try to ameliorate in a future version.

This algorithm is quite generic and easily extendable to cover a multitude of RAT in a heterogeneous network. This is something we plan to do as a next step. A step further is to specify in details all message and parameters exchange through Specification and Description Language–SDL [21]. This will show in details how this algorithm works in specific scenarios. Another interesting extension to this work is to map the functionality of the algorithm in specific network entities. Finally, a comparison and evaluation of this algorithm against other existing ones is on our future plans.

5. References

[1] A. Kaloxylos, I. Modeas, N. Passas, and G. Lampropoulos, "Radio resource management in 4G mobile systems," Encyclopedia of Wireless and Mobile Communications, ed. Borko Furht, CRC Press, Taylor & Francis Group, 2008.

[2] J. McNair and F. Zhu, "Vertical handoffs in fourth-generation multi-network environments," IEEE Wireless Communications, Vol. 11, pp. 8–15, June 2004.

[3] A. Hasswa, N. Nasser, and H. Hassanein, "Generic vertical handoff decision function for heterogeneous wireless networks," 2nd IFIP International Conference on Wireless & Optical Communications Networks, pp. 239– 243, 2005.

[4] O. Ormond, P. Perry, and J. Murphy, "Network selection decision in wireless heterogeneous networks," IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), September 2005.

[5] P. L. M. Chan, R. E. Sheriff, Y. F. Hu, P. Conforto, and C. Tocci, "Mobility management incorporating fuzzy logic for a heterogeneous IP environment," IEEE Communications Magazine, Vol. 39, No. 12, pp. 42–51, 2001.

[6] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J. P. Makela, R. Pichna, and J. K. Vallström, "Handoff in hybrid mobile data networks, " IEEE Personal Communications, Vol. 7, No. 2, pp. 34–47, April 2000.

[7] L. Giupponi, R. Augusti, J. Perez-Romero, and O. Sallent, "A novel joint radio resource management approach with reinforcement learning mechanisms," 24th IEEE International Performance Computing & Communications Conference (IPCCC), pp. 621–626, 2005.

[8] R. Augusti, *et al.*, "A fuzzy-neural based approach for joint radio resource management in a beyond 3G network," Proceedings 1st International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QSHINE), 2004.

[9] J. Pérez-Romero, O. Sallent, and R. Agustí, "Policy- Based initial RAT selection algorithms in heterogeneous networks," in Proceedings of Mobile and Wireless Communication Networks (MWCN), 2005.

[10] F. Zhu and J. McNair, "Multi-Service vertical handoff decision algorithms," in EURASIP Journal on Wireless Communications and Networking, 2006.

[11] W. Song, W. Zhuang, and Y. Cheng "Load balancing for cellular/WLAN integrated networks," IEEE Network, Vol. 21, No. 1, pp. 27–33, January–February 2007.

[12] I. Modeas, A. Kaloxylos, N. Passas, and L. Merakos, "An algorithm for radio resources management in integrated cellular/WLAN networks," IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), September 2007.

[13] A. Kaloxylos, G. Lampropoulos, N. Passas, and L. Merakos, "A flexible mechanism for service continuity in 4G environments," Elsevier Computer Communications Journal, special issue on end-to-end QoS provision advances, 2006.

[14] G. Lampropoulos, A. Kaloxylos, N. Passas, and L. Merakos, "A power consumption analysis of tight-coupled WLAN/UMTS networks," IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communica-

tions (PIMRC), September 2007.

[15] 3GPP TS 25.213 version 7.6.0 Release 7, Universal Mobile Telecommunications System (UMTS), Spreading and modulation (FDD), October 2008.

[16] J. M. Holtzman and A. Sampath, "Adaptive averaging methodology for handoffs in cellular systems," IEEE Transactions on Vehicle Technology, pp. 59–66, 1995.

[17] M. D. Austin and G. L. Stüber, "Velocity adaptive handoff algorithms for microcellular systems", IEEE Transactions on Vehicle Technology, Vol. 43, pp. 549–561, 1994.

[18] 3GPP TS 25.305 v7.4.0 Technical Specification, 3rd Generation Partnership Project, Technical Specification Group Radio Access Network, Stage 2 functional specification of User Equipment (UE) positioning in UTRAN (Release 7), September 2007.

[19] The Network Simulator-ns-2, http://www.isi.edu/nsnam/ns/.

[20] H. Kaaranen, A. Ahtiainen, L. Laitinen, S. Anghian, and V. Niemi, "UMTS networks, architecture, mobility and services," Second Edition, Wiley, 2005.

[21] International Telecommunication Union, "Specification and description language (SDL)," Recommendation Z.100, ITU-T Study Group 17, http://www.itu.int/ITU-T/studygroups/com17/languages/Z100.pdf.

Int. J. Communications, Network and System Sciences, **2009, 2, 822-826**
doi:10.4236/ijcns.2009.29095 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Adaptive Co-Channel Interference Suppression Technique for Multi-User MIMO MC DS/CDMA Systems

Prabagarane Nagaradjane[1], Arvind Sai Sarathi Vasan[2],
Lakshmi Krishnan[2], Anand Venkataswamy[1]
[1]*Department of Electronics and Communication Engineering, SSN College of Engineering, SSN Institutions, Chennai, India*
[2]*Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland, USA*
E-mail: *prabagaranen@ssn.edu.in*, {*arvind*88, *lakshmik*}*@umd.edu*, *anandv*267*@gmail.com*
Received September 2, 2009; *revised October* 10, 2009; *accepted November* 16, 2009

Abstract

In this paper, an adaptive co-channel interference suppression technique for multi-user MIMO MC DS/CDMA system is envisaged. MC DS/CDMA offers many advantages like flexibility, robustness, low PAPR and spectral efficiency. In spite of these advantages, performance of MC DS/CDMA system is greatly impaired by interference. Common interferences, which degrade the performance of the system, are MAI and CCI. Mitigating these interferences can directly increase the capacity of the system. In this work, an adaptive co-channel interference suppression technique based on single-stage and two-stage MMSE IC is considered for multi-user MIMO MC DS/CDMA system. Simulation results show that, at low SNR two-stage MMSE IC outperforms single-stage, while at high SNR, single-stage provides better BER performance. Based on this, a selection criterion has been propounded for improved system performance as a whole in interference limited environment. Also, adaptive selection criterion resulted in better error performance.

Keywords: CCI, MIMO, MC DS/CDMA, MMSE IC, ML Decoding

1. Introduction

Multi-carrier transmission has recently gained enormous attention for providing high data rate communications on both forward and reverse channels. Multi-carrier transmission is realized through OFDM.OFDM performs well in frequency selective channels [1]. One of the most promising single carrier transmission schemes is CDMA, which is robust to noise [2]. Also in the recent past, multi-input multi-output (MIMO) has proved to provide very high capacity without any increase in bandwidth or power [3]. Combining these techniques leads to MIMO MC DS/CDMA system, which is expected to meet the demands of future broadband (4G) wireless wide-area networks. Though MIMO MC DS/CDMA possesses many advantages, it still suffers from the traditional impairments of conventional CDMA systems like MAI and CCI [4–6]. Of late, myriad research concentration has been on proposing techniques for mitigating MAI (Multiple Access Interference) and multi-path, but very little work has been carried out on CCI (Co-Channel Interference) combating techniques. In this paper, we investigate the performance of single-stage and two-stage CCI cancellation techniques for a MIMO MC DS/CDMA system. MIMO is realized by employing space-time block codes. At the receiver, we have employed two-stage MMSE (Minimum Mean Square Error) IC (Interference Cancellation) with ML (Maximum Likelihood) decoding [4]. We have considered a multi-user MIMO MC DS/CDMA system, with k asynchronous co-channel users in the uplink. Each of the k asynchronous co-channel users is equipped with N_T transmit antennas and they communicate to a single base station, equipped with N_R receive antennas. In this multi-user environment, k x N_T interfering signals will be arriving at the base station. Conventional interference mitigation technique from k-1 co-channel users requires N_T x ((k-1) +1) receive antennas, so as to suppress the co-channel interferences. But by employing STBC the same can be achieved by using N_R receive antennas such that $N_R \geq k$, which exploits the spatial and temporal structure. Also in this work we consider two co-channel users, each equipped with two transmit antennas and the base station equipped with two receive antennas. The performance of MMSE IC and ML decoding algorithm over MC DS/CDMA system with each MC DS/CDMA transmitter in turn carrying multi-user data is assayed for both downlink and uplink. The paper is organized as follows: Section 2 introduces the system model we have considered throughout this work, Section 3 describes the improved MMSE IC with ML decoding algorithm, Section 4 describes the two-stage MMSE IC for MIMO MC DS/CDMA system, Section 5 expounds the performance results and Section 6 provides the conclusions.

2. System Models

Here the system model that we have considered comprises two asynchronous co-channel users in a MIMO MC DS/CDMA system in which each transmitting terminal is a MC DS/CDMA transmitter with MIMO support realized through Space Time Block Codes (spatial diversity). The two co-channel users communicate with two receive antennas at the receiver, which performs interference cancellation and then detects the transmitted signal of each user. Figure 1 shows the MIMO MC DS/CDMA transmitter .The user data at each transmitting terminal after appropriate constellation mapping is multiplied with a spreading code and the spread symbols of each data are multi-carrier modulated. Then the sum of all the carriers of the *k*th user composes the output of the *k*th user signal $S_k(t)$. The total output *S (t)* is the sum of all the user signals. After HPA (high power amplification), the final signal is transmitted. The transmitted signal corresponding to the *n*th data symbol of the *k*th user is

$$S_n^k(t) = \sum_{m=0}^{M-1}\sum_{q=0}^{Q-1} d_m^k C_{m,q}^k \, p_m(t - qT_c - nT)\, e^{j2\pi f_m t} \quad (1)$$

where, *k* is the user number, *m* is the carrier number, *q* is the chip number, T_c is chip duration and *T* is symbol duration and equals to $Q\ T_c$. *Q* is the length of user specific spreading code. d_m^k is the data of *m*th sub-carrier and *k*th user, $c_{m,q}{}^k$ is the *q*th spreading code of *m*th carrier of *k*th user, $p_m(t)$ is Root raised cosine pulse of *m*th carrier and f_m is the *m*th carrier frequency. When the total number of users is *K*, the total transmitted signal corresponding to the *n*th data symbol is

$$S_n(t) = \sum_{k=0}^{K-1}\sum_{m=0}^{M-1}\sum_{q=0}^{Q-1} d_m^k C_{m,q}^k e^{j2\pi m \Delta ft} \quad (2)$$

This modulated stream is then passed through a STBC encoder which groups the symbols according to a specific STBC pattern (G2) and then transmits the symbols through multiple transmit antennas. The received vector at the first receive antenna for the transmitted symbols $\boldsymbol{d_1}$ and $\boldsymbol{d_2}$ is expressed as

$$\begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} d_1 & d_2 \\ -d_2^* & d_1^* \end{bmatrix}\begin{bmatrix} h_{11} \\ h_{21} \end{bmatrix} + \begin{bmatrix} n_{11} \\ n_{21} \end{bmatrix} \quad (3)$$

where, $\boldsymbol{h_{11}}$ and $\boldsymbol{h_{21}}$ denotes the channel fading and $\boldsymbol{n_{11}}$ and $\boldsymbol{n_{21}}$ represents the additive white Gaussian noise (AWGN). Let $\boldsymbol{u_1}$, $\boldsymbol{u_2}$ and $\boldsymbol{v_1}$, $\boldsymbol{v_2}$ represent the symbols transmitted over two consecutive symbol durations by the corresponding two co-channel asynchronous users. As shown in Figure 2, let $\boldsymbol{h_{11}}$, $\boldsymbol{h_{21}}$, $\boldsymbol{g_{11}}$ and $\boldsymbol{g_{21}}$ represent the channel fading between all the transmit antennas and the first receive antenna at the base station respectively and $\boldsymbol{h_{12}}$, $\boldsymbol{h_{22}}$, $\boldsymbol{g_{12}}$ and $\boldsymbol{g_{22}}$ represent the channel fading between all the transmit antennas and the second receive antenna at the base station respectively. Now, the received vector at the receive antenna Rx1 over the two symbol time period is expressed as

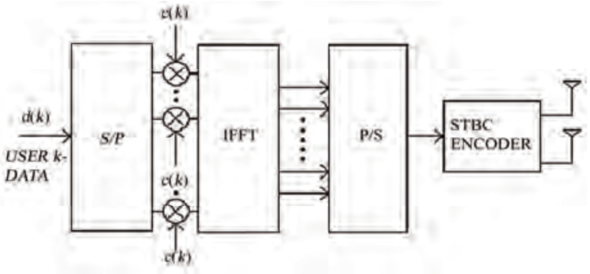


Figure 1. MIMO MC DS/CDMA transmitter.

$$r_{11} = h_{11}u_1 + h_{21}u_2 + g_{11}v_1 + g_{21}v_2 + n_{11} \quad (4)$$

$$r_{12} = -h_{11}u_2^* + h_{21}u_1^* - g_{11}v_2^* + g_{21}v_1^* + n_{12} \quad (5)$$

The received vector at the receiver antenna Rx2 is given by

$$r_{21} = h_{12}u_1 + h_{22}u_2 + g_{12}v_1 + g_{22}v_2 + n_{21} \quad (6)$$

$$r_{22} = -h_{12}u_2^* + h_{22}u_1^* - g_{12}v_2^* + g_{22}v_1^* + n_{22} \quad (7)$$

3. Improved MMSE IC ML Decoding

The overall received signal from each of the receive antenna is expressed as

$$r = \begin{bmatrix} R_1^T & R_2^T \end{bmatrix}^T \quad (8)$$

The channel fading coefficient between each transmitter and the receive antenna is rearranged for each co-channel user to form a generic channel matrix. For detecting the first co-channel user data, the corresponding channel matrix is represented as

$$H = \begin{bmatrix} H_1 & G_1 \\ H_2 & G_2 \end{bmatrix} \quad (9)$$

and for the second user, it is given by

$$\tilde{H} = \begin{bmatrix} G_1 & H_1 \\ G_2 & H_2 \end{bmatrix} \quad (10)$$

The MMSE weight matrix is calculated based on the generic channel matrix and the weight matrix value dif-

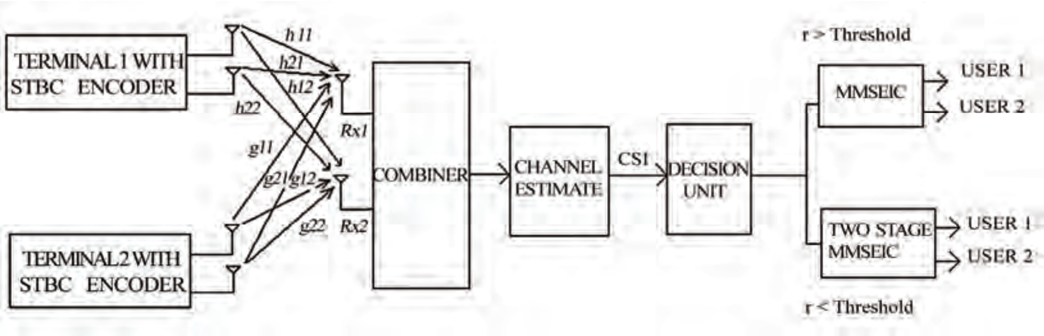


Figure 2. MIMO MC DS/CDMA with two co-channel asynchronous users and adaptive base station receiver system model.

fers for each co-channel user. For each of the two users considered here in this work, the MMSE weight matrix is computed from the expression [4,5]

$$M = HH^H + \sigma^2 I_4 \quad \text{and} \quad \tilde{M} = \tilde{H}\tilde{H}^H + \sigma^2 I_4 \qquad (11)$$

where, ***M*** is the weight matrix for the first user and $\tilde{M}$ is the weight matrix for the second user. H^H represent the Hermitian transpose of the channel matrix and $\boldsymbol{\sigma}^2$ is the noise variance. To suppress the interferences from other co-channel users, the inverse of the weight matrix is multiplied with the columns of the channel matrix that represents the channel fading for a particular co-channel user i.e.

$$w_1 = M^{-1}h_1 \quad \text{and} \quad \tilde{w}_1 = \tilde{M}^{-1}h_1 \qquad (12)$$

where,

h₁*=first column of H or* $\tilde{H}$

h₂*=second column of H or* $\tilde{H}$

The MMSE Interference cancellation receiver suppresses both co-channel interferences and noise components, which means that the mean square error or variance between the transmitted symbols and the estimate is reduced. The maximum likelihood (ML) detection is used to detect the transmitted symbols for the corresponding user. The ML decoding estimates the symbols by determining the minimum Euclidian distance of all possible transmitted symbols from the received constellation [4], given by

$$\hat{U} = \arg\min_{\hat{u}\in U} \{\|w_1^H r - \hat{u}_1\|^2 + \|w_2^H r - \hat{u}_2\|^2\} \qquad (13)$$

and

$$\hat{V} = \arg\min_{\hat{v}\in V} \{\|\tilde{w}_1^H r - \hat{v}_1\|^2 + \|\tilde{w}_2^H r - \hat{v}_2\|^2\} \qquad (14)$$

where, $\hat{U}$ and $\hat{V}$ represent the two co-channel users estimated data. $\hat{u}_1, \hat{u}_2$, $\hat{v}_1$ and $\hat{v}_2$ takes all possible values of the users signal constellation. The reliability of the decoded signals are computed by

$$\gamma_{c0} = \|w_1^H r - \hat{u}_1\|^2 + \|w_2^H r - \hat{u}_2\|^2 \qquad (15)$$

$$\gamma_{s1} = \|w_1^H r - \hat{v}_1\|^2 + \|w_2^H r - \hat{v}_2\|^2 \qquad (16)$$

4. Two Stage Interference Cancellation

The two stage interference cancellation proceeds with first decoding the data from the two terminals. Then, assuming each decoded value is correct, the data corresponding to the other user is estimated using Maximum likelihood estimation. i.e. first assuming the first terminal decoded data is correct based on single stage MMSE IC, the second user data is estimated from the decoded data. This is carried out by first calculating $\mathbf{x_1}$ and $\mathbf{x_2}$.

$$x_1 = R_1 - H_1.\hat{c}_0 \quad \text{and} \quad x_2 = R_2 - H_2.\hat{c}_0 \qquad (17)$$

where, $\hat{c}_0$ is the data decoded, corresponding to the first user. The second user data is estimated from the first user decoded data by using the maximum likelihood estimation given by

$$\hat{s}_0 = \arg\min_{\hat{s}_0\in S}\{\|x_1 - G_1 s\|^2 + \|x_2 - G_2 s\|^2\} \qquad (18)$$

where, ***S*** takes all possible values in the signal constellation. The reliability corresponding to the estimated data is given by the expression

$$\gamma_{s0} = \|x_1 - G_1\hat{s}_0\|^2 + \|x_2 - G_2\hat{s}_0\|^2 \qquad (19)$$

Similarly, the first user data based on single stage MMSE IC is estimated by computing $\mathbf{y_1}$ and $\mathbf{y_2}$.

$$y_1 = R_1 - G_1.\hat{s}_1 \quad \text{and} \quad y_2 = R_2 - G_2.\hat{s}_1 \qquad (20)$$

where, $\hat{s}_1$ is the data decoded corresponding to the second user. The first user data is estimated from the second user decoded data by using the expression [4]

$$\hat{c}_1 = \arg\min_{\hat{c}_1\in C}\{\|y_1 - H_1 c\|^2 + \|y_2 - H_2 c\|^2\} \qquad (21)$$

The corresponding value of reliability for the estimated data is

$$\gamma_{c1} = \|y_1 - H_1\hat{c}_1\|^2 + \|y_2 - H_2\hat{c}_1\|^2 \qquad (22)$$

The receiver computes the overall reliability for the two users i.e. $\gamma_0 = \gamma_{c0} + \gamma_{s0}$ and $\gamma_1 = \gamma_{c1} + \gamma_{s1}$ .The decision is made on the sets of symbols computed by comparing the two reliabilities. The comparison is made as [4]

$$\text{if } (\gamma_0 < \gamma_1)$$
$$(\hat{c}, \hat{s}) = (\hat{c}_0, \hat{s}_0)$$
$$\text{else}$$
$$(\hat{c}, \hat{s}) = (\hat{c}_1, \hat{s}_1) \quad (23)$$

The system illustrated in Figure 2, consists of a receiver with a switch between a single and double stage MMSE IC unit, CSI (Channel State Information) and a decision unit. When the channel is slowly varying, the receiver detects the symbols based on single stage technique. When the channel variation is rapid, two stage MMSE IC is employed to detect the symbols. At present, perfect channel knowledge is assumed at the receiver. The entire detection takes place based on the instantaneous SNR available at the receiver.

5. Performance Analysis

In this section, we present the performance of adaptive co-channel suppression technique for a multi-user MIMO MC DS/CDMA system. The simulation results for the single stage and two stage interference cancellation techniques are shown in Figures 3 and 4. The channel model considered is quasi-static Rayleigh fading channel, which is built on the classical understanding of Doppler shift and delay spread. The modulation scheme employed is BPSK, as it provides the best system throughput for MIMO realization based on STBC. The number of channel realizations considered for uplink and downlink is 5000 and 10000 respectively for each value for SNR. Table 1 summarizes the simulation parameters.

Simulation results divulge that, at low SNR value, two stage interference cancellation techniques perform well whereas at high SNR value single stage MMSE IC with ML decoding provides better BER performance. Hence, a trade off can be made in selecting the interference cancellation techniques at the receiver when the SNR dwindles. This can result in better performance of MIMO MC DS/CDMA system in an interference limited environment as switching of IC can be made in an adaptive manner.

Figure 3 shows the uplink performance of MIMO MC DS/CDMA system with single stage IC, two stage IC and adaptive IC for two co-channel users. Each user data is spread by a spreading factor of 32. Here in each MC DS CDMA system one user is accommodated. It can be discerned that adaptive IC outperforms both single stage and double stage. The same performance can also be realized over the downlink channel. Figure 4 elucidates

Table 1. Simulation parameters.

PARAMETERS		VALUES
Spreading code		Walsh Hadamard
Number of channel realizations	Uplink	5000
	Downlink	10000
FFT Size		128
Cyclic Prefix		1/8
Spreading Factor		16 or 32
Data Modulation		BPSK
Channel Model		Rayleigh
Number of Transmit antennas		2
Number of Receive antennas		2

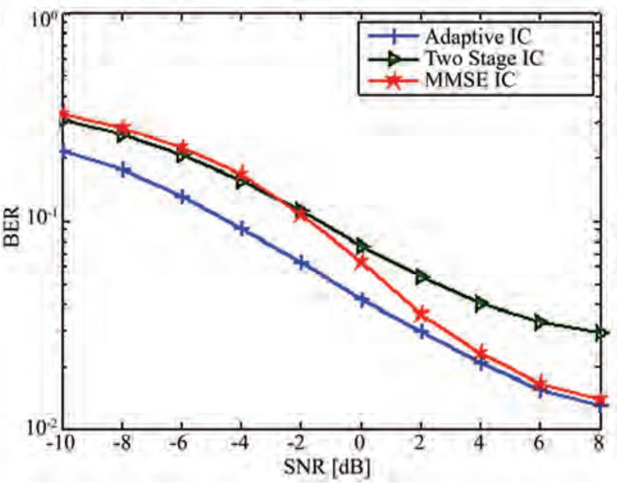


Figure 3. Performance of adaptive co-channel interference scheme for 2 co-channel users over an uplink communication channel for MIMO MC DS/CDMA system.

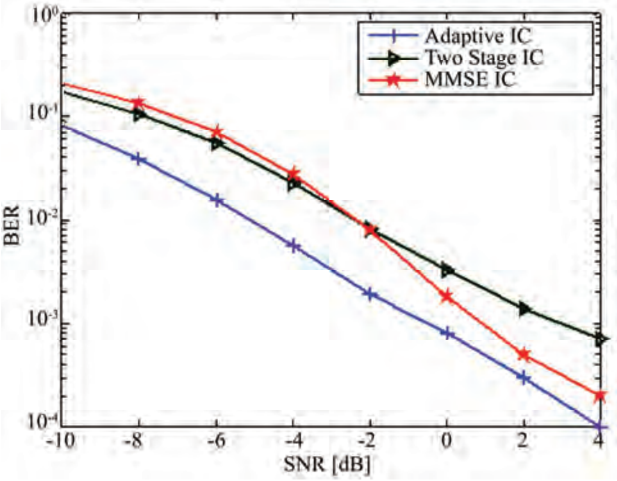


Figure 4. Performance of adaptive co-channel interference scheme for 4 co-channel users over downlink communication channel for MIMO MC DS/CDMA system.

the performance of the same system with four co-channel users, with each user spread by a spreading factor of 16 over downlink communication channel. Here also adaptive switching scheme provides better BER performance.

6. Conclusions

In this work, we considered a two stage MMSE co-channel interference cancellation receiver for MIMO MC DS/CDMA systems. MC DS /CDMA can be realized as a prominent air interface for 4G Broadband communications; however, capacity of such systems is limited by interference. Mitigating the various interferences can result in confronting the future generation wireless networks needs. In this paper we have analyzed a two stage IC technique for a multi-user environment. Results of our analysis reveal that a trade off could be made in selecting the IC techniques for mitigating CCI. It could be discerned that at low SNR values two stage has resulted in better performance because of its iterative nature while at high SNR values, single stage performs better. Also, it is expounded from our analysis that the adaptive interference cancellation receiver has resulted in better suppression of CCI.

7. References

[1] L. Hanzo, L-L. Yang, E-L. Kuna, and K. Yen, “Single and multi-carrier DS-CDMA multi-user detection, space-time spreading, synchronization and standards,” IEEE Press, 2003.

[2] M. K. Simon and M. S. Alouini, “BER performance of multi-carrier DS-CDMA systems over generalized fading channels,” IEEE International Conference, 1999.

[3] E. Bigieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. Vincent Poor, “MIMO wireless communications,” Cambridge University Press, 2007.

[4] Anand. V, Arvind. S, and Lakshmi Krishnan, “Investigations on the performance of MIMO assisted Multi Carrier DS/CDMA system with multi-user detection for 4G mobile communications,” Dissertation, SSN Institutions, 2009.

[5] Prabagarane Nagaradjane, Arvind Sai Sarathi Vasan, and Lakshmi Krishnan, “A robust space time co-channel interference mitigation and detection technique for multi-user MIMO multi-carrier DS/ CDMA systems,” Proceedings of IEEE International Conference, Wireless Vitae, 2009.

[6] S. Kondo and L. B. Milstein, “Performance of multi-carrier DS CDMA systems”, IEEE Transactions on Communications, Vol. 44, No. 2, February 1996.

Int. J. Communications, Network and System Sciences, **2009, 2, 827-835**
doi:10.4236/ijcns.2009.29096 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Fuzzy Timed Agent Based Petri Nets for Modeling Cooperative Multi-Robot Systems

Xingli HUANG, Hua XU, Peifa JIA
State Key Laboratory on Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing, China
E-mail: *hxl712@sina.com, xuhua@tsinghua.edu.cn*

Received September 1, 2009; *revised October* 11, 2009; *accepted November* 20, 2009

Abstract

A cooperative multi-robot system (CMRS) modeling method called fuzzy timed agent based Petri nets (FTAPN) is proposed in this paper, which has been extended from fuzzy timed object-oriented Petri net (FTOPN). The proposed FTAPN can be used to model and illustrate both the structural and dynamic aspects of CMRS, which is a typical multi-agent system (MAS). At the same time, supervised learning is supported in FTAPN. As a special type of high-level object, agent is introduced into FTAPN, which is used as a common modeling object in its model. The proposed FTAPN can not only be used to model CMRS and represent system aging effect, but also be refined into the object-oriented implementation easily. At the same time, it can also be regarded as a conceptual and practical artificial intelligence (AI) tool for multi-agent systems (MAS) into the mainstream practice of the software development.

Keywords: Petri nets, Multi Cooperative Robot Systems, Multi-Agent Systems

1. Introduction

As a kind of typical manufacturing equipment, cooperative multi-robot systems (CMRS) have been widely used in current industries [1]. The key solution for one CMRS is to realize the cooperation, which is different from generic control systems. So, currently, the CMRS modeling, analysis and refinement always meet with difficulties related to the cooperation problem. As one of the typical multi-agent systems (MAS) in distributed artificial intelligence [2], CMRS can be regarded as one MAS. In order to model MAS, some attempts to use object-oriented methodology have been tried and some typical agent objects have been proposed, such as *active object*, etc [3]. However, agent based object models still can not depict the whole structure and dynamic aspects of MAS, such as cooperation, learning, temporal constraints, etc [2].

On one hand, as one of the most proper and promised realization technologies, object-oriented concurrent programming (OOCP) methodology is always used to realize MAS [3]. In OOCP realization, one special object called *active object* is proposed [3], which can be used to model generic agent architectures and behaviors with OO methodology. Although OOCP can solve MAS realization problem favorably, the modeling problem mentioned above still exists.

On the other hand, as a kind of powerful formal description and analysis method for dynamic systems [4], Petri net (PN) has become a bridge between practical application and model theories [4]. Basic Petri nets lack temporal knowledge description, so they have failed to describe the temporal constraints in time critical or time dependent systems. Then in the improved models of Petri nets such as Timed (or Time) Petri nets (TPN) [5,6] etc al, temporal knowledge has been introduced, which has increased not only the modeling power but also the model complexity [7]. On the other hand, when Petri nets are used to analyze and model practical systems in different fields, models may be too complex to be analyzed. These years, object-oriented concepts have been introduced into Petri nets such as object Petri nets (OPN) [8], VDM++ [9], Object-Z [10], etc al. are suggested. Among the studies, the research on OPN has been focused on the extending Petri net formalism to OPN such as HOONet [11], OBJSA [12], COOPN/2 [13] and LOOPN++ [14], which are suggested on the base of colored Petri Net (CPN) [15]. Object-oriented Petri net (OPN) can model various systems hierarchically and the models can be analyzed even if they have not been completed. So the complexity of OPN models can be simplified at the beginning of modeling stage according to the analysis requirements. Although the results of such studies have shown promise, these nets do not fully support time critical (time dependent) system modeling and analysis, which may be complex, midsize or even small. When time critical systems with any sizes are modeled, it requires formal modeling and analysis method supporting temporal description and object-oriented concepts. Then for providing the ability of modeling time critical complex systems, timed hierarchical object oriented Petri net (TOPN) [16] is proposed on the base of HOONet [11]. It supports temporal knowledge description and object-oriented concepts. Modeling features in TOPN support describing and analyzing dynamic systems such as MAS and CMRS. Recently, some attempts have been conducted on modeling MAS by means of OPN [17].

Although Petri nets can be used to model and analyze different systems, they failed to model learning ability and the aging effects in dynamic systems. Recently, fuzzy timed Petri net (FTPN) [18] has been presented to solve these modeling problems. As a kind of reasoning and learning ability, fuzzy reasoning in FTPN can be considered as supporting autonomous judging or reasoning ability in MAS. In order to solve the reasoning ability and other modeling problems in large-scale MAS, fuzzy timed object-oriented Petri net (FTOPN) [19] is proposed on the base of TOPN [16] and FTPN [18]. In FTOPN, agent can be modeled as one FTOPN object with autonomy, situatedness and sociality. However, in FTOPN every agent should be modeled from common FTOPN objects and it needs generic FTOPN agent objects on the base of *active object*s.

This paper proposes a high level PN called fuzzy timed agent based Petri net (FTAPN) on the base of FTOPN [19]. As one of the typical active objects, ACTALK object is modeled by FTOPN and is introduced into FTAPN, which is used as generic agent object in FTAPN. The aim of FTAPN is to solve the agent or CMRS modeling ability problem and construct a bridge between MAS models and their implementations.

This paper is organized as the following. Section 2 reviews the relative preliminary notations quickly and Section 3 extends FTOPN to FTAPN on the base of ACTALK model. Section 4 discusses the learning which is important for representing dynamic behaviors in CMRS. Section 5 uses FTAPN to model one CMRS in the wafer etching procedure of circuit industry and makes some modeling analysis to demonstrate its benefits in modeling MAS. Finally, the conclusion and future work can be found in Section 6.

2. Preliminary Notations

In this section, the basic concepts of ACTALK are firstly reviewed, which is the relative concept in object-oriented concurrent programming. Then, the definitions of TOPN and FTOPN are introduced quickly, which are the basis of the research work in this paper.

2.1. ACTALK

2.1.1. Active Objects [3]

Object-oriented concurrent programming (OOCP) is one of the most appropriate and promising technologies for implementing or realizing agent based systems or MAS. Combining the agent concept and the object-oriented paradigm leads to the notion of agent-oriented programming [20]. The uniformity of objects' communication mechanisms provides facilities for implementing agent communication, and the concept of encapsulating objects or encapsulation support combining various agent granularities. Furthermore, the inheritance mechanism enables knowledge specialization and factorization.

The concept of an active object has been presented, which make it possible to integrate an object and activity (namely a thread or process). It also provides some degree of autonomy for objects in that it does not rely on external resources for activation. Thus, it provides a good basis for implementing agent based systems or MAS. However, similar to common objects in object-oriented systems, an active object's behavior still remains procedural and only reacts to message requests. More generally, the main feature of agent-based systems or MAS is autonomous. Agents should be able to complete tasks autonomously. That's to say, agents must be able to perform numerous functions or activities without external intervention over extended time periods. In order to achieve autonomy, adding to an active object a function that controls message reception and processing by considering its internal state is one of the effective realization methods [21,22].

On one hand, for modelling and realizing MAS, there are two basic questions regarding how to build a bridge between implementing and modelling MAS requirements [23,24]. On the other hand, the facilities and techniques OOCP provides [25]:

- How can a generic structure define an autonomous agent's main features?
- How do we accommodate the highly structured OOCP model in this generic structure?

The active-object (or actor) concept has been introduced to describe a set of entities that cooperate and communicate through message passing. This concept brings the benefits of object orientation (for example, modularity and encapsulation) to distributed environments and provides object-oriented languages with some of the characteristics of open systems [26]. Based on these characteristics, various active object models have been proposed [27], and to facilitate implementing active-object systems, several frameworks have been proposed. ACTALK is one example.

When ACTALK is used to model and realize the MAS, there still exist the following shortcomings:

ı Active object is not an autonomous agent. It only manifests the procedural actions.

ı Although active object can communicate, they do not own the ability to reduce the decision to communicate or order other active objects.

ı If one active object has not received the information from other active objects. It is still in none-active state.

In order to overcome the shortcomings mentioned above, the concept of active object has been proposed and a general agent framework [22,28]. A universal agent architecture has also been proposed so as to fulfil the modelling requirements of MAS [25], which can be used to model and analyze MAS deeply. For either agent-based systems or MAS, the method mostly is on the base of active objects. So in this chapter, the concept of active object is firstly reviewed quickly.

2.1.2. ACTALK [29]

One of the typical active objects is ACTALK. ACTALK is a framework for implementing and computing various active-object models into a single programming environment based on Smalltalk, which is an object-oriented programming language. ACTALK implements asynchronism, a basic principle of active-object languages, by queuing the received messages into a mailbox, thus dissociating message reception from interpretation. In ACTALK, an active object is composed of three component classes (see Figure 1), which are instances of the classes.

ı Address encapsulates the active object's mailbox. It defines how to receive and queue messages for later interpretation.

ı Activity represents the active object's internal activity and provides autonomy to the actor. It has a Smalltalk process and continuously removes messages from the mailbox, and the behavior component interprets the messages.

ı ActiveObject represents the active object's behavior—that is, how individual messages are interpreted.

To build an active object with ACTALK, the algorithm must describe its behavior as a standard Smalltalk (OOCP) object. The active object using that behavior is created by sending the message active to the behavior:

active

"*Creates an active object with self as corresponding behavior*"

^*self activity*: *self activityClass address*: *self addressClass*

The activityClass and addressClass methods represent the default component classes for creating the activity and address components (along the *factory method* design pattern). To configure the framework of ACTALK means to define the components of its sub-classes. That's

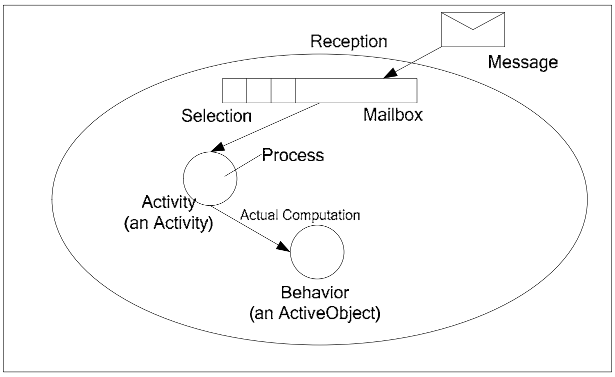


Figure 1. Components of an ACTALK active object.

to say, it allows users to define special active object models. So ACTALK is the basis to model agent based systems or MAS.

2.2. Timed Object-Oriented Petri Net (TOPN)

Formally TOPN is a four-tuple (OIP, ION, DD, SI), where (OIP, ION, DD) is an ordinary object Petri net—"HOONet" [30] and SI associates a static (firing) temporal interval SI: $\{o\} \rightarrow [a, b]$ with each object o, where a and b are rationals in the range $0 \leqslant a \leqslant b \leqslant +\propto$, with $b \neq +\propto$. The four parts in TOPN have different function roles. Object identification place (OIP) is a unique identifier of a class. Internal timed object net (ION) is a net to depict the behaviors (methods) of a class. Data dictionary (DD) declares the attributes of a class in TOPN. And static time interval function (SI) binds the temporal knowledge of a class in TOPN. There are two kinds of places in TOPN. They are common places (represented as circles with thin prim) and abstract places (represented as circles with bold prim). Abstract places are also associated with a static time interval. Because at this situation, abstract places represent not only firing conditions, but also the objects with their own behaviors. So, abstract places (TABP) in TOPN also need to be associated with time intervals. One problem to be emphasized is that the tokens in abstract places need to have two colors at least. Before the internal behaviors of an abstract place object are fired, the color of tokens in it is one color (represented as hollow token in this paper). However, after fired, the color becomes the other one

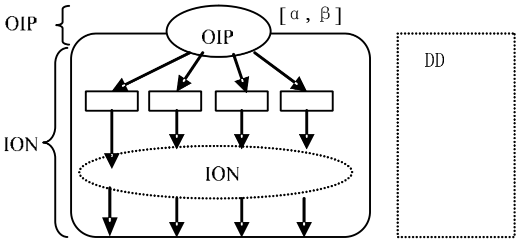
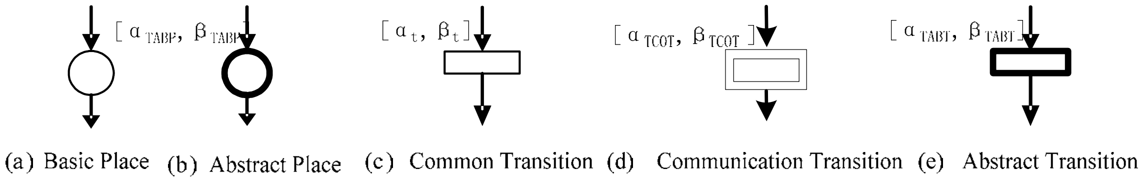


Figure 2. The general structure of TOPN.


[α_{TABP}, β_{TABP}]
[α_t, β_t]
[α_{TCOT}, β_{TCOT}]
[α_{TABT}, β_{TABT}]

(a) Basic Place (b) Abstract Place (c) Common Transition (d) Communication Transition (e) Abstract Transition

Figure 3. Places and transitions in TOPN.

(represented as liquid token in this paper). At this time, for the following transitions, it is just actually enabled. There are three kinds of transitions in TOPN. The timed primitive transition (represented as rectangles with thin prim) (TPIT), timed abstract transition (represented as rectangles with bold prim) (TABT) and timed communication transition (represented as rectangles with double thin prim) (TCOT).

Static time intervals change the behavior of TOPN just similar to a time Petri net in the following way. If an object o with SI(o) = [a, b] becomes enabled at time I_0, then the object o must be fired in the time interval [I_0+a, I_0+b], unless it becomes disabled by the removal of tokens from some input place in the meantime. The static earliest firing time of the object o is a; the static latest firing time of o is b; the dynamic earliest firing time (EFT) of t is I_0+a; the dynamic latest firing time (LFT) of t is I_0+b; the dynamic firing interval of t is [I_0+a, I_0+b].

The state of TOPN (Extended States—"ES") is a 3-tuple, where ES = (M, I, path) consists of a marking M, a firing interval vector I and an execution path. According to the initial marking M_0 and the firing rules mentioned above, the following marking at any time can be calculated. The vector—"I" is composed of the temporal intervals of enabled transitions and TABPs, which are to be fired in the following states. The dimension of I equals to the number of enabled transitions and TABPs at the current state. The firing interval of every enabled transition or TABP can be got according to the calculation formula of EFT and LFT in TOPN [16].

For enabling rules in TOPN, two different situations exist. A transition t in TOPN is said to be enabled at the current state (M, I, path), if each input place p of t contains at least the number of solid tokens equal to the weight of the directed arcs connecting p to t in the marking M. If the TABP object is marked with a hollow token, it is enabled. At this time, its ION is enabled. After the ION has been fired, the tokens in TABP are changed into solid ones.

An object o is said to be fireable in state (M, I, path) if it is enabled, and if it is legal to fire o next. This will be true if and only if the EFT of o is less than or equal to the LFT of all other enabled transitions. Of course, even with strong time semantics, o's being fireable in state (M, I, path) does not necessarily mean that t will fire in the time interval I.

2.3. Fuzzy Timed Object-Oriented Petri Net (FTOPN)

Similar to FTPN [19], fuzzy set concepts are introduced into TOPN [16]. Then FTOPN is proposed, which can describe fuzzy timing effect in dynamic systems.

Definition 1: FTOPN is a six-tuple, FTOPN = (OIP, ION, DD, SI, R, I) where

1) Suppose OIP = (oip, pid, M_0, status), where oip, pid, M_0 and status are the same as those in HOONet [30] and TOPN [16].

- oip is a variable for the unique name of a FTOPN.
- pid is a unique process identifier to distinguish multiple instances of a class, which contains return address.
- M_0 is the function that gives initial token distributions of this specific value to OIP.
- status is a flag variable to specify the state of OIP.

2) ION is the internal net structure of FTOPN to be defined in the following. It is a variant CPN that describes the changes in the values of attributes and the behaviors of methods in FTOPN.

3) DD formally defines the variables, token types and functions (methods) just like those in HOONet [30] and TOPN [16].

4) SI is a static time interval binding function, SI: {OIP}→Q*, where Q* is a set of time intervals.

5) R: {OIP} → r, where r is a specific threshold.

6) I is a function of the time v. It evaluates the resulting degree of the abstract object firing.

Definition 2: An internal object net structure of TOPN, ION = (P, T, A, K, N, G, E, F, M_0)

1) P and T are finite sets of places and transitions with time restricting conditions attached respectively.

2) A is a finite set of arcs such that $P \cap T = P \cap A = T \cap A = \Phi$.

3) K is a function mapping from P to a set of token types declared in DD.

4) N, G, and E mean the functions of nodes, guards, and arc expressions, respectively. The results of these functions are the additional condition to restrict the firing of transitions. So they are also called additional restricting conditions.

5) F is a special arc from any transitions to OIP, and notated as a body frame of ION.

6) M_0 is a function giving an initial marking to any place the same as those in HOONet [30] and TOPN [16].

Definition 3: A set of places in TOPN is defined as P

= PIP∪TABP, where

1) Primary place PIP is a three-tuple: PIP = (P, R, I), where

- P is the set of common places similar to those in PN [4,31].

2) Timed abstract place (TABP) is a six-tuple: TABP = TABP(pn, refine state, action, SI, R, I), where

- pn is the identifier of the abstract timed place.
- refine state is a flag variable denoting whether this abstract place has been refined or not.
- action is the static reaction imitating the internal behavior of this abstract place.
- SI, R and I are the same as those in Definition 1.

Definition 4: A set of transitions in TOPN can be defined as T = TPIT∪TABT∪TCOT, where

1) Timed primitive transition TPIT = TPIT (BAT, SI), where

- BAT is the set of common transitions.

2) Timed abstract transition TABT = TABT (tn, refine state, action, SI), where

- tn is the name of this TABT.

3) Timed communication transition TCOT = TCOT (tn, target, comm type, action, SI).

- tn is the name of TCOT.
- target is a flag variable denoting whether the behavior of this TCOT has been modeled or not. If target = "Yes", it has been modeled. Otherwise, if target = "No", it has not been modeled yet.
- comm type is a flag variable denoting the communication type. If comm type = "SYNC", then the communication transition is synchronous one. Otherwise, if comm type = "ASYN", it is an asynchronous communication transition.

4) SI is the same as that in Definition 1.

5) Refine state and action are the same as those in Definition 3.

Similar to those in FTPN [19], the object t fires if the foregoing objects come with a nonzero marking of the tokens; the level of firing is inherently continuous. The level of firing ($z(v)$) assuming values in the unit interval is governed by the following expression:

$$z(v) = (\overset{n}{\underset{i=1}{T}}(r_i \to x_i(v'))sw_i)tI(v) \qquad (1)$$

where T (or t) denotes a t-norm while "s" stands for any s-norm. "v" is the time instant immediately following v′. More specifically, $x_i(v)$ denotes a level of marking of the i^{th} place. The weight w_i is used to quantify an input coming from the i^{th} place. The threshold r_i expresses an extent to which the corresponding place's marking contributes to the firing of the transition. The implication operator (→) expresses a requirement that a transition fires if the level of tokens exceeds a specific threshold (quantified here by r_i).

Once the transition has been fired, the input places involved in this firing modify their markings that is governed by the expression

$$x_i(v) = x_i(v')t(1-z(v)) \qquad (2)$$

(Note that the reduction in the level of marking depends upon the intensity of the firing of the corresponding transition, $z(v)$.) Owing to the t-norm being used in the above expression, the marking of the input place gets lowered. The output place increases its level of tokens following the expression:

$$y(v) = y(v')sz(v) \qquad (3)$$

The s-norm is used to aggregate the level of firing of the transition with the actual level of tokens at this output place. This way of aggregation makes the marking of the output place increase.

The FTOPN model directly generalizes the Boolean case of TOPN and OPN. In other words, if $x_i(v)$ and w_i assume values in {0, 1} then the rules governing the behavior of the net are the same as those encountered in TOPN.

3. Agent Objects and Fuzzy Timed Agent Based Petri Nets

The *active object* concept [29] has been proposed to describe a set of entities that cooperate and communicate through message passing. To facilitate implementing *active object* systems, several frameworks have been proposed. ACTALK is one of the typical examples. ACTALK is a framework for implementing and computing various *active object* models into one object-oriented language realization. ACTALK implements asynchronism, a basic principle of *active object* languages, by queuing the received messages into a mailbox, thus dissociating message reception from interpretation. In ACTALK,

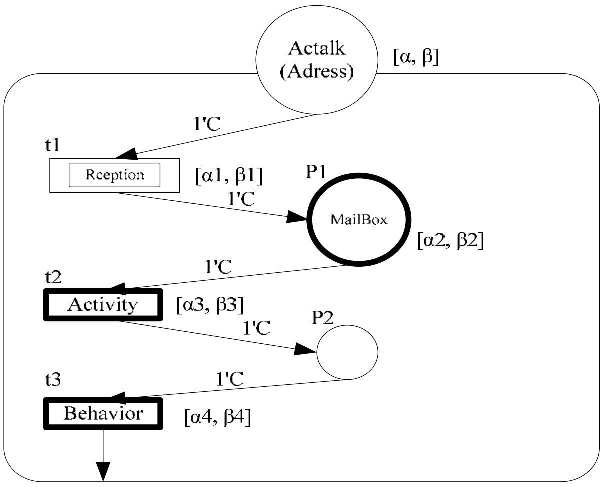


Figure 4. The FTOPN model of ACTALK.

an *active object* is composed of three component classes: *address*, *activity* and *activeObject* [29].

ACTALK model is the basis of constructing *active object* models. However, *active object* model is the basis of constructing multi-agent system model or agent-based system model. So, as the modeling basis, ACTALK has been extended to different kinds of high-level agent models. Because of this, ACTALK is modeled in Fig.4 by FTOPN.

In Figure 4, OIP is the describer of the ACTALK model and also represents as the communication address. One communication transition is used to represent as the behavior of message reception. According to the communication requirements, it may be synchronous or asynchronous. If the message has been received, it will be stored in the corresponding mail box, which is one "first in and first out queue". If the message has been received, the next transition will be enabled immediately. So mail box is modeled as abstract place object in FTAPN. If there are messages in the mail box, the following transition will be enabled and fired. After the following responding *activity* completes, some *active behavior* will be conducted according to the message.

Figure 4 has described the ACTALK model based on FTOPN on the macroscopical level. The detailed definition or realization of the object "*Activity*" and "*Behavior*" can be defined by FTOPN in its parent objects in the lower level. The FTOPN model of ACTALK can be used as the basic agent object to model agent based systems. That is to say, if the agent based model—ACTALK model is used in the usual FTOPN modeling procedure, FTOPN has been extended to *agent based modeling methodology*. So it is called *fuzzy timed agent based Petri net* (*FTAPN*).

4. Learning in Fuzzy Timed Agent Based Petri Nets

The parameters of FTAPN are always given beforehand. In general, however, these parameters may not be available and need to be estimated just like those in FTPN [19]. The estimation is conducted on the base of some experimental data concerning marking of input and output places. The marking of the places is provided as a discrete time series. More specifically we consider that the marking of the output place(s) is treated as a collection of target values to be followed during the training process. As a matter of fact, the learning is carried out in a supervised mode returning to these ***target*** data.

The connections of the FTOPN (namely weights w_i and thresholds r_i) as well as the time decay factors α_i are optimized (or trained) so that a given performance index *Q* becomes minimized. The training data set consists of (a) initial marking of the input places $x_i(0),\ldots, x_n(0)$ and (b) target values—markings of the output place that are given in a sequence of discrete time moments, that is target(0), target(1),…, target(K).

In FTAPN, the performance index Q under discussion assumes the form of Equation (4)

$$Q = \sum_{k=1}^{K}(t\arg et(k)-y(k))^2 \tag{4}$$

where the summation is taken over all time instants (k =1, 2,… , K).

The crux of the training in FTOPN models follows the general update formula in Equation (5) being applied to the parameters:

$$\mathbf{param}(\text{iter+1}) = \mathbf{param}(\text{iter}) - \gamma\nabla_{\mathbf{param}}\boldsymbol{Q} \tag{5}$$

where γ is a learning rate and $\nabla_{\mathbf{param}}\boldsymbol{Q}$ denotes a gradient of the performance index taken with respect to all parameters of the net (here we use a notation **param** to embrace all parameters in FTOPN to be trained).

In the training of FTOPN models, marking of the input places is updated according to Equation (6):

$$x_i^{\sim} = x_i(0)T_i(k) \tag{6}$$

where $T_i(k)$ is the temporal decay. And $T_i(k)$ complies with the form in Equation (7). In what follows, the temporal decay is modeled by an exponential function,

$$T_i(k) = \begin{cases} \exp(-a_i(k-k_i)) & if \quad k > k_i, \\ 0 & others \end{cases} \tag{7}$$

The level of firing of the place can be computed as Equation (8):

$$z = (\underset{i=1}{\overset{n}{T}} \quad ((r_i \to x_i^{\sim})\, sw_i)) \tag{8}$$

The successive level of tokens at the output place and input places can be calculated as that in Equation (9):

$$y(k) = y(k-1)\boldsymbol{s}z,\ x_i(k) = x_i(k-1)\boldsymbol{t}(1-z) \tag{9}$$

We assume that the initial marking of the output place y(0) is equal to zero, y(0) = 0. The derivatives of the weights w_i are computed as the form in Equation (9):

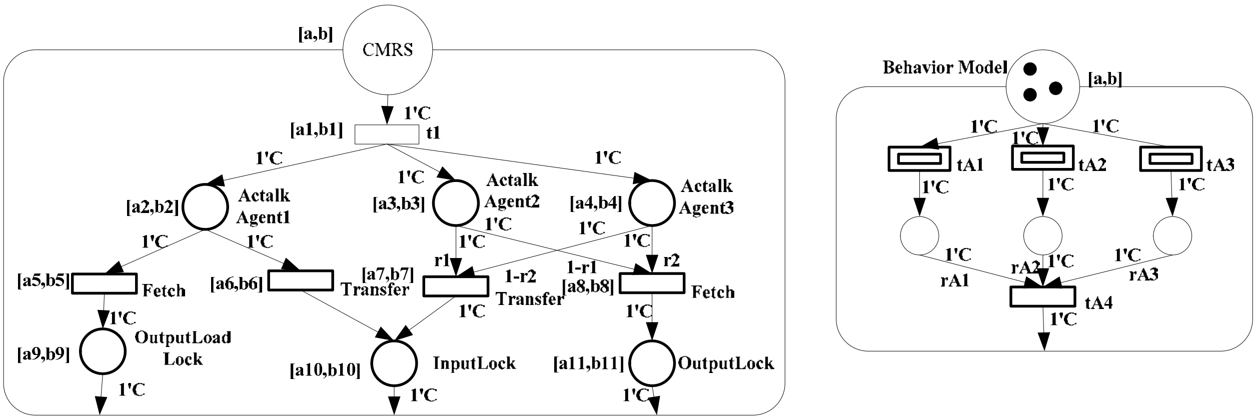
$$\begin{aligned} &\frac{\partial}{\partial w_i}(t\arg et(k)-y(k))^2 \\ &= -2(t\arg et(k)-y(k)\frac{\partial y(k)}{\partial w_i}) \end{aligned} \tag{10}$$

where i =1,2,…, n. Note that y(k+1) = y(k)***s***z(k).

5. A Modeling Example

5.1. A CMRS Model

In many typical integrated circuit manufacturing equipments such as etching tools, PVD, PECVD etc al., usually there is an EFAM platform which is made up of three Brooks Marathon Express (MX) [32] robots to transfer wafers to be processed. Among these three robots,



(a) The agent based FTAPN model

(b) The behavior model in every agent

Figure 5. The FTAPN model.

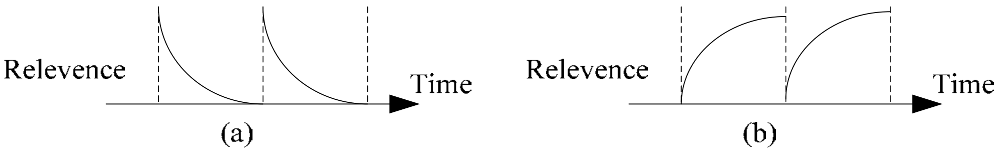


Figure 6. The relevance.

one is up to complete transferring wafers between atmospheric environment and vacuum environment, which is conducted in the atmospheric environment. In the vacuum environment, the other two robots are up to complete transferring one unprocessed wafer from the input lock to the chamber and fetch the processed wafer to the output lock in the vacuum environment. Any robot can be used to complete the transferring task at any time. If one robot is up to transfer one new wafer, the other will conduct the relative transferring or fetching task. They will not conflict with each other. Figure 5 depicts this CMRS FTAPN model, where three agent objects (ACTALK) are used to represent these three cooperative robots.

Figure 5(a) has depicted the whole FTAPN model. The agent object—“ACTALK” is used to represent every robot model. Different thresholds are used to represent the firing level of the behavior conducted by the corresponding robot (agent). They also satisfy the unitary requirements and change according to the fuzzy decision in the behavior of every agent in Figure 5(b). In the model of Figure 5(b), three communication transition objects are used to represent the behavior for getting different kinds of system states. These states include the state of the other robot, its own goal and its current state, which can be required by the conductions of the communication transitions tA1, tA2 and tA3. When one condition has been got, the following place will be marked. In order to make control decisions (transition object tA4) in time, all of these state parameters are required in the prescriptive time interval. However, the parameter arrival times complies with the rule in Figure 5(a). The other two kinds of information comply with that in Figure 5(b). After the decision, a new decision command with the conduction probability will be sent in this relative interval and it also affects which behavior (transfer or fetch) will be conducted by updating the threshold in Figure 5(a).

5.2. Application Analysis

Table 1 summarizes the main features of FTAPN and contrast these with the structures with which the proposed structures have a lot in common, namely MAS and FTOPN. It becomes apparent that FTAPN combines the advantages of both FTOPN in terms of their learning abilities and the glass-style of processing (and architectures) of MAS with the autonomy.

6. Conclusions and Future Work

CMRS is a kind of usual manufacturing equipments in manufacturing industries. In order to model, analyze and simulate this kind of systems, this paper proposes fuzzy timed agent based Petri net (FTAPN) on the base of FTOPN [19] and FTPN [18]. In FTAPN, one of the active objects—ACTALK is introduced and used as the basic agent object to model CMRS, which is a typical MAS. Every abstract object in FTOPN can be trained and reduced independently according to the modeling and analysis requirements for OO concepts supported in

Table 1. MAS, FTOPN and FTAPN: A comparative analysis.

Characteristics	MAS	FTOPN	FTAPN
Learning Aspects	Significant dynamic learning abilities. Dynamic learning and decision abilities are supported in every autonomous agent.	Significant learning abilities. Distributed learning (training) abilities are supported in different independent objects on various system model levels.	Significant dynamic learning abilities. Distributed dynamic learning and decision abilities are supported in every autonomous agent.
Knowledge Representation Aspects	Transparent knowledge representation (glass box processing style) the problem (its specification) is mapped directly onto the topology of the agent model. Additionally, agents deliver an essential feature of continuity required to cope with dynamic changes encountered in a vast array of problems (including autonomous decision tasks)	Glass Box Style (Transparent Knowledge Representation) and Black Box Processing is supported at the same time. The problem (its specification) is mapped directly onto the topology of FTOPN. Knowledge representation granularity reconfiguration reacts on the reduction of model size and complexity.	Glass Box Processing Style and Black Box Processing style are all supported. The problem (its specification) is mapped directly onto the topology of FTAPN, which can not only represent dynamic knowledge, but also deal with dynamic changes with well-defined semantics of agent objects, places, transitions, fuzzy and temporal knowledge.

FTOPN. The validity of this modeling method has been used to model Brooks CMRS platform in etching tools. The FTAPN can not only model complex MAS, but also be refined into the object-oriented implementation easily. It has provided a methodology to overcome the development problems in agent-oriented software engineering. At the same time, it can also be regarded as a conceptual and practical artificial intelligence (AI) tool for integrating MAS into the mainstream practice of software development.

State analysis needs to be studied in the future. An extended State Graph [16] has been proposed to analyze the state change of TOPN models. With the temporal fuzzy sets introduced into FTAPN, the certainty factor about object firing (state changing) needs to be considered in the state analysis.

7. Acknowledgement

This work is jointly supported by the National Nature Science Foundation of China (Grant No: 60405011, 60575057, 60875073) and National Key Technology R&D Program of China (Grant No: 2009- BAG12A08).

8. References

[1] Y. U. Cao, A. S. Fukunaga, A. B. Kahng, and F. Meng, "Cooperative mobile robotics: Antecedents and directions," Autonomous Robots, Vol. 4, pp. 7–27, 1997.

[2] N. R. Jennings, K. Sycara, and M. Wooldridge, "A roadmap of agent research and development," Autonomous Agents and Multi-Agent Systems, Vol. 1, pp. 7–38, 1998.

[3] Z. Guessoum and J. P. Briot, "From active objects to autonomous agents," IEEE Concurrency, Vol. 7, No. 3, pp. 68–76, July–September 1999.

[4] T. Murata, "Petri nets and properties, analysis and applications," Proceedings of IEEE, Vol. 77, pp. 541–580, 1989.

[5] Y. L. Yao, "A Petri net model for temporal knowledge representation and reasoning," IEEE Transactions on Systems, Man and Cybernetics, Vol. 24, pp. 1374–1382, 1994.

[6] P. Merlin and D. Farber, "Recoverability of communication protocols—Implication of a theoretical study," IEEE Transactions on Communication, Vol. 24, pp. 1036–1043, 1976.

[7] J. Wang, Y. Deng, and M. Zhou, "Compositional time Petri nets and reduction rules," IEEE Transactions on Systems, Man and Cybernetics (Part B), Vol. 30, pp. 562–572, 2000.

[8] R. Bastide, "Approaches in unifying Petri nets and the object-oriented approach," Proceeding of the International Workshop on Object-Oriented Programming and Models of Concurrency, Turin, Italy, June, 1995, http:// eprints. kfupm.edu.sa/26256/.

[9] D. Harel and E. Gery, "Executable object modeling with statechart," Proceedings of the 18th International Conference on Software Engineering, Germany, pp. 246–257, March 1996.

[10] S. A. Schuman, "Formal object-oriented development," Springer, Berlin, 1997.

[11] J. E. Hong and D. H. Bae, "Software modeling and analysis using a hierarchical object-oriented Petri net," Information Sciences, Vol.130, pp. 133–164, 2000.

[12] E. Battiston, F. D. Cindio, and G. Mauri, "OBJSA nets: A class of high-level nets having objects as domains," APN'88, Lecture Notes in Computer Science, Vol. 340, pp. 20–43, 1988.

[13] O. Biberstein and D. Buchs, “An object-oriented specification language based on hierarchical algebraic Petri nets,” Proceedings of the IS-CORE Workshop Amsterdam, September 1994, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.3092.

[14] C. Lakos and C. Keen, “LOOPN++: A new language for object-oriented Petri nets,” Technical Report R94-4, Networking Research Group, University of Tasmania, Australia, April 1994.

[15] K. Jensen, “Coloured Petri nets: Basic concepts, analysis methods and practical use,” Springer, Berlin, 1992.

[16] H. Xu and P. F. Jia, “Timed hierarchical object-oriented Petri net-part I: Basic concepts and reachability analysis,” Lecture Notes in Artificial Intelligence (Proceedings of RSKT2006), Vol. 4062, pp. 727–734, 2006.

[17] W. Chainbi, “Multi-agent systems: A Petri net with objects based approach,” Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology, Beijing, pp. 429–432, 2004.

[18] W. Pedrycz and H. Camargo, “Fuzzy timed Petri nets, fuzzy sets and systems,” Vol. 140, pp. 301–330, 2003.

[19] X. Hua and J. Peifa, “Fuzzy timed object-oriented Petri net,” Artificial Intelligence Applications and Innovations II-Proceedings of AIAI2005, Springer, pp. 155–166, September 2005.

[20] Y. Shoham, “Agent-oriented programming,” Artificial Intelligence, Vol. 60, No.1, pp. 139–159, 1993.

[21] J. P. Briot, “An experiment in classification and specialization of synchronization schemes,“ Lecture Notes in Computer Science, No. 1107, pp. 227–249, 1996.

[22] T. Maruichi, M. Ichikawa, and M. Tokoro, “Decentralized AI,” Modeling Autonomous Agents and Their Groups, Elsevier Science, Amsterdam, pp. 215–134, 1990.

[23] C. Castelfranchi, “A point missed in multi-agent, DAI and HCI,” Lecture Notes in Artificial Intelligence, No. 890, pp. 49–62. 1995.

[24] L. Gasser, ”An overview of DAI, ” Distributed Artificial Intelligence, N. A. Avouris and L. Gasser, eds., Kluwer Academic, Boston, pp. 1–25, 1992.

[25] L. Gasser and J. P. Briot, ”Object-oriented concurrent programming and distributed artificial intelligence,” Distributed Artificial Intelligence, N. A. Avouris and L. Gasser, eds., Kluwer Academic, Boston, pp. 81–108, 1992.

[26] G. Agha and C. Hewitt, “Concurrent programming using actors: Exploiting large scale parallelism,” Lecture Notes in Computer Science, S. N. Maheshwari, ed., Springer-Verlag, New York, No. 206, pp. 19–41, 1985.

[27] A. Yonezawa and M. Tokoro, “Object-oriented concurrent programming,” The MIT Press, Cambrige, Mass., 1987.

[28] Y. Shoham, “Agent-oriented programming,” Artificial Intelligence, Vol. 60, No.1, pp. 139–159, 1993.

[29] Z. Guessoum and J. P. Briot, “From active objects to autonomous agents,” IEEE Concurrency, Vol. 7, No. 3, pp. 67–76, 1999.

[30] J. E. Hong and D. H. Bae, ”Software modelling and analysis using a hierarchical object-oriented Petri net,” Information Sciences, Vol. 130, pp. 133–164, 2000.

[31] J. L. Peterson, “Petri net theory and the modeling of systems,” Prentice-Hall, N.Y., USA, 1991.

[32] J. H. Lee and T. E. Lee, “SECAM: A supervisory equipment control application model for integrated semiconductor manufacturing equipment, ” IEEE Robotics & Automation Magazine, Vol. 11, No. 1, pp. 41 – 58, 2004..

Int. J. Communications, Network and System Sciences, 2009, 2, 836-844
doi:10.4236/ijcns.2009.29097 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Performance Analysis of an Optimal Circular 16-QAM for Wavelet Based OFDM Systems

Khaizuran ABDULLAH, Seedahmed S. MAHMOUD*, Zahir M. HUSSAIN
School of Electrical & Computer Engineering, RMIT University, Melbourne, Australia
**Future Fiber Technologies Pty. Ltd., Mulgrave, Australia*
E-mail: *khaizuran.abdullah@rmit.edu.au, smahmoud@fft.com.au, zmhussain@ieee.org*
Received August 20, 2009; *revised September* 29, 2009; *accepted October* 29, 2009

Abstract

The BER performance for an optimal circular 16-QAM constellation is theoretically derived and applied in wavelet based OFDM system in additive white Gaussian noise channel. Signal point constellations have been discussed in much literature. An optimal circular 16-QAM is developed. The calculation of the BER is based on the four types of the decision boundaries. Each decision boundary is determined based on the space distance d following the pdf Gaussian distribution with respect to the in-phase and quadrature components nI and nQ with the assumption that they are statistically independent to each other. The BER analysis for other circular M-ary QAM is also analyzed. The system is then applied to wavelet based OFDM. The wavelet transform is considered because it offers a better spectral containment feature compared to conventional OFDM using Fourier transform. The circular schemes are slightly better than the square schemes in most SNR values. All simulation results have met the theoretical calculations. When applying to wavelet based OFDM, the circular modulation scheme has also performed slightly less errors as compared to the square modulation scheme.

Keywords: Performance OFDM, Fourier-Based OFDM, Wavelet-Based OFDM, Circular 16-QAM, Square 16-QAM

1. Introduction

Quadrature amplitude modulation (QAM) is one of the most popular modulation schemes used by orthogonal frequency division multiplexing. Some popular types of M-ary QAM are 4-QAM, 16-QAM and 64-QAM. The number of 4, 16 and 64 is corresponding to 2^2, 2^4 and 2^6 in which that the superscript number 2, 4 or 6 is the bit rate per OFDM symbol respectively. In this paper, the constellation points derivation and the BER analysis are focused on 16-QAM, which gives an intermediate result of BER performance between 4- and 64-QAM in an AWGN channel [1]. The 16-QAM is also one of the standard modulation schemes in OFDMs' applications such as terrestrial Digital Video Broadcasting (DVB), Digital Audio Broadcasting (DAB) and High Performance Radio LAN Version 2 (HIPERLAN/2) [2]. In the transmitter, an OFDM symbol is mapped from binary to complex signal with amplitude and phase represented in real and imaginary number. On the other hand, the signal is demapped or extracted from complex signal to OFDM symbol in the receiver. The decision boundary is needed to detect the correct symbols between the transmitter and receiver. The bit error rate (BER) performance is determined after performing the difference of errors between the transmitted bits with the received bits. The BER performance of M-ary QAM has been investigated by several authors. The exact BER expressions for QAM is presented in [3]. An extension of BER expressions considering of an arbitrary constellation size is discussed in [4]. Both works include the square constellation points.

In this paper, we propose an alternative BER expression using optimal circular constellation points. The calculation of probability of error is based on determining the decision boundary. We have proposed four types of decision boundaries.

Based on these types, the probability of error occurs when the receiver is making an incorrect decision. To the best of our knowledge, there is no work of the probability of error calculation for a circular 16-QAM with the application of wavelet-based OFDM, namely discrete wavelet transform (DWT). The principle feature of DWT is it has low pass and high pass filters satisfying perfect reconstruction property in the transmitter and receiver

[5-7]. The use of wavelet is significant since wavelet has a better spectral containment feature compared to conventional OFDM using Fourier filter [8,9]. To be specific, the application of Mband wavelet filters in wavelet-based OFDM, having the pulses for different overlapping data blocks in time, is designed to achieve a combination of subchannel spectral containment and bandwidth efficiency that is fundamentally better than with other forms of multicarrier modulation [9]. Other term of DWT is discrete wavelet multitone modulation (DWMT) or wavelet-OFDM (W-OFDM).

This paper is organized as follows. Determining the constellation points the circular 16-QAM is discussed in the next section followed by the calculation of an exact probability of error in Section 3 and the wavelet OFDM principles in Section 4. The system model of wavelet based OFDM is discussed in Section 5 and finally the BER results are obtained in Section 6.

2. The Derivation of an Optimal Circular Constellation Points

A circular signal point constellation has been discussed in [1]. However, the discussion is for M = 8 constellations, while M = 16 can be inferred as sub-optimal. We extend the work for an optimal circular 16-QAM. In this section, we discuss only the derivation for the circular 16-QAM since the derivation for a square 16-QAM is well known in many literatures. The number of circles and amplitudes for the circular scheme is different than those of the conventional square scheme. Let the number of circles define as S and the amplitude level associating with the diameter define as r. In this particular circular 16-QAM, we have S = 4 with 4 points on all circles with different diameter r_1, r_2, r_3 and r_4 with the derivation as follows

$$\begin{aligned} r_1 &= \sqrt{d^2+d^2} = \sqrt{2} \\ r_2 &= \sqrt{3d} = \sqrt{3} \\ r_3 &= \sqrt{(1+r_2)^2 + 2^2 - 4(1+r_2)\cos(P_h)} \\ r_4 &= \sqrt{d_s^2 + r_1^2 - 2d_s r_1 \cos\left(\frac{P_p}{2} + P_{si}\right)} \end{aligned} \tag{1}$$

where $P_h = \frac{\pi}{3} + P_0$, $P_0 = \tan^{-1}\left(\frac{1}{r_2}\right)$, $d_s = \sqrt{8d^2 - (8d)\cos(b)}$, $b = \pi - P_p$, $P_p = \phi - \frac{\pi}{3}$, $\phi = 2\pi - 2P_{si}$ and $P_{si} = \pi - \frac{\pi}{4} - P_0$.

Note that the minimum distance is d = 1. By rearranging (1) in vector representation, we have

$$V_c = d \times \left[r_1 \;\; 1+r_2 \;\; r_3 \;\; r_4 \right] \tag{2}$$

Since every 4 points share one diameter, we repeat every amplitude 4 times. Therefore

$$v_c = (V_c)^T \times [1 \; 1 \; 1 \; 1] \tag{3}$$

and the amplitude vector A_{vc} for all amplitudes of QAM constellation points becomes

$$A_{vc} = (v_c)^T \tag{4}$$

Subsequently we need to derive the rotating phase for the constellation points. Thus,

$$\begin{aligned} \theta_{c1} &= \frac{\pi}{4} \times [1 \; 3 \; 5 \; 7 \; 0 \; 2 \; 4 \; 6] \\ \theta_{c2} &= [h_0 \;\; h_1 \;\; h_2 \;\; h_3] \\ \theta_{c3} &= [g_0 \;\; g_1 \;\; g_2 \;\; g_3] \end{aligned} \tag{5}$$

where

$h_0 = \sin^{-1}\left(\frac{2}{r_3}\sin(p_h)\right)$, $h_1 = \pi - h_0$, $h_2 = \pi + h_0$, $h_3 = -h_0$,

$g_0 = \frac{\pi}{4} + \sin^{-1}\left(\frac{d_s}{r_4}\sin\left(\frac{P_p}{2} + P_{si}\right)\right)$, $g_1 = \pi - g_0$, $g_2 = \pi + g_0$, $g_3 = -g_0$. Rearranging (5), in vector representation, we obtain

$$\theta_c = [\theta_{c1} \;\; \theta_{c2} \;\; \theta_{c3}] \tag{6}$$

and θ_c has all angles of all constellation points. Combining the amplitude A_{vc} and the phase θ_c, the circular 16-QAM (Scir) is expressed as

$$S_{cir} = A_{vc}\cos(\theta_c) + jA_{vc}\sin(\theta_c) \tag{7}$$

The simulation result is obtained and shown in Figure 1.

3. The Exact BER Calculation

Each decision boundary in Figure 2 is determined by the space distance d following the pdf Gaussian distribution with respect to the in-phase and quadrature components

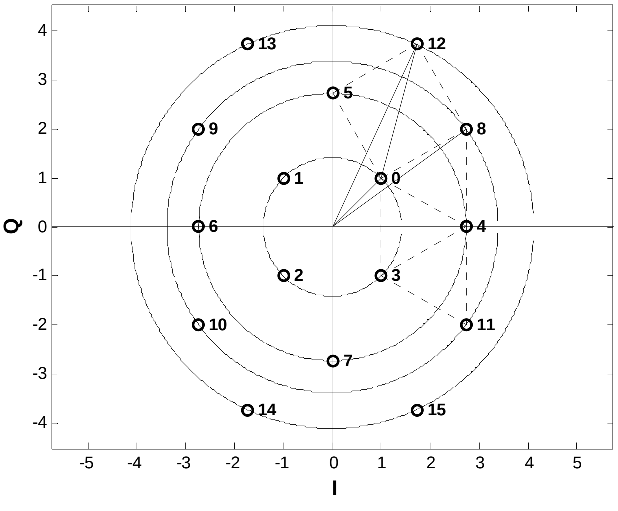


Figure 1. Circular 16-QAM constellation points.

n_I and n_Q with the assumption that they are statistically independent to each other. Thus, four types of boundary regions can be determined accordingly from the figure. Type 1 in part (a) of Figure 3 is for the points related to the most inner circles in Figure 2. The probability of a correct decision is

$$\begin{aligned} P_{c1} &= \left[1-\left\{P\left(n_I \le -\frac{0.5d}{\sigma}\right)+P\left(n_I \ge \frac{0.5d}{\sigma}\right)\right\}\right] \\ &\quad \times \left[1-\left\{P\left(n_Q \ge \frac{0.5d}{\sigma}\right)+P\left(n_I \le -\frac{0.5d}{\sigma}\right)\right\}\right] \\ &= \left[1-2Q\left(\frac{0.5d}{\sigma}\right)\right]\left[1-2Q\left(\frac{0.5d}{\sigma}\right)\right] \end{aligned} \tag{8}$$

where $\mathrm{Q}(\mathrm{x}) = \frac{1}{\sqrt{2\pi}}\int_{\mathrm{x}}^{\infty} \mathrm{e}^{\frac{\mathrm{x}^2}{2}}\,\mathrm{dx}$ and the probability of error is

$$P_{e1} = 4[1-P_{c1}] = 4\left[1-\left(1-2Q\left(\frac{0.5d}{\sigma}\right)\right)^2\right] \tag{9}$$

The next type is associated with 2 points, $\{5, 7\}$. The boundary region is shown in part (b) of Figure 3. The probability of correct decision can be expressed as

$$\begin{aligned} P_{c2} &= \left[1-\left\{P\left(n_I \le -\frac{0.5d}{\sigma}\right)+P\left(n_I \ge \frac{0.5d}{\sigma}\right)\right\}\right] \\ &\quad \times \left[1-P\left(n_Q \le -\frac{0.232d}{\sigma}\right)\right] \\ &= \left[1-2Q\left(\frac{0.5d}{\sigma}\right)\right]\left[1-Q\left(\frac{0.232d}{\sigma}\right)\right] \end{aligned} \tag{10}$$

and the probability of error for Type 2 is given by

$$\begin{aligned} P_{e2} &= 2[1-P_{c1}] \\ &= 2\left[1-\left\{\left[1-2Q\left(\frac{0.5d}{\sigma}\right)\right]\left[1-Q\left(\frac{0.232d}{\sigma}\right)\right]\right\}\right] \end{aligned} \tag{11}$$

Considering the BER analysis for Type 3, six points {4,6,8,9,10,11} are involved. The decision boundary related to this type is shown in part (c) of Figure 3. Then probability of correct decision is given by

$$\begin{aligned} P_{c3} &= \left[1-\left\{P\left(n_Q \ge \frac{0.5d}{\sigma}\right)+P\left(n_Q \le -\frac{0.5d}{\sigma}\right)\right\}\right] \\ &\quad \times \left[1-P\left(n_I \le -\frac{1.232d}{\sigma}\right)\right] \\ &= \left[1-2Q\left(\frac{0.5d}{\sigma}\right)\right]\left[1-Q\left(\frac{1.232d}{\sigma}\right)\right] \end{aligned} \tag{12}$$

and the probability of error is

$$\begin{aligned} P_{e3} &= 6[1-P_{c3}] \\ &= 6\left[1-\left\{\left[1-2Q\left(\frac{0.5d}{\sigma}\right)\right]\left[1-Q\left(\frac{1.232d}{\sigma}\right)\right]\right\}\right] \end{aligned} \tag{13}$$

Next, the decision boundary for Type 4. It is associated to the points {12,13,14,15}. The probability of correct decision is

$$\begin{aligned} P_{c4} &= \left[1-P\left(n_I \le -\frac{0.232\,d}{\sigma}\right)\right]\left[1-P\left(n_Q \le -\frac{1.232\,d}{\sigma}\right)\right] \\ &= \left[1-Q\left(\frac{0.232\,d}{\sigma}\right)\right]\left[1-Q\left(\frac{1.232\,d}{\sigma}\right)\right] \end{aligned} \tag{14}$$

and the probability of error for Type 4 is expressed as

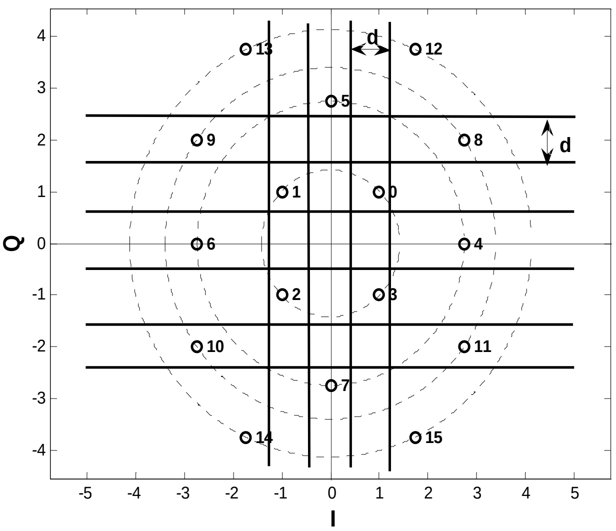
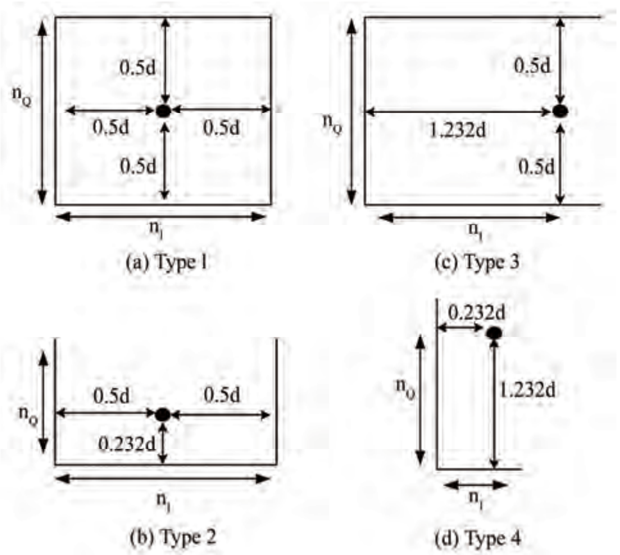


Figure 2. Signal-space diagram for circular 16-QAM.



(a) Type 1 (c) Type 3 (b) Type 2 (d) Type 4

Figure 3. All types of decision boundary associated to Figure 2. Note that 0.5, 0.232 and 1.232 are the results obtained from (7) due to variations of A_{vc} and θ_c.

$$\begin{aligned}P_{e4} &= 4\left[1-P_{c4}\right]\\ &= 4\left[1-\left\{\left[1-Q\left(\frac{0.232d}{\sigma}\right)\right]\left[1-Q\left(\frac{1.232d}{\sigma}\right)\right]\right\}\right]\end{aligned} \quad (15)$$

Combining and rearranging Equations (9), (11), (13) and (15), the average probability of error for the circular 16-QAM scheme is given by

$$\begin{aligned}P_{cir} &= \frac{1}{16}\left[P_{e1}+P_{e2}+P_{e3}+P_{e4}\right]\\ &= \frac{1}{8}\left[8A\left(2-A-\frac{1}{4}\left(B+3C\right)\right)+2B\left(\frac{3}{2}-C\right)+5C\right]\end{aligned} \quad (16)$$

where $\mathrm{A} = \mathrm{Q}\left(\frac{0.5\mathrm{d}}{\sigma}\right)$, $\mathrm{B} = \mathrm{Q}\left(\frac{0.232\mathrm{d}}{\sigma}\right)$ and $\mathrm{C} = \mathrm{Q}\left(\frac{1.232\mathrm{d}}{\sigma}\right)$. Using $\mathrm{d} = \sqrt{\frac{3\log_2 \mathrm{M.E}_\mathrm{b}}{2\left(\mathrm{M}-1\right)}}$ from [4] where M=16 and $\sigma = \frac{1}{2}\sqrt{\frac{\mathrm{N}_0}{5}}$, Equation (16) can be expressed as in term of energy per bit over noise density ratio $\left(\frac{\mathrm{E}_\mathrm{b}}{\mathrm{N}_0}\right)$.

Thus parameters A, B and C can be rewritten as

$\mathrm{A} = \mathrm{Q}\left(0.5\left(\gamma\right)\right)$, $\mathrm{B} = \mathrm{Q}\left(0.232\left(\gamma\right)\right)$

and

$\mathrm{C} = \mathrm{Q}\left(1.232\left(\gamma\right)\right)$ where $\gamma = 4\sqrt{\frac{1}{2}\frac{E_b}{N_0}}$.

The analysis can also determine the exact BER for other circular M-ary QAM. Note that the process of obtaining BER analysis for the square scheme is excluded since it is available in much literature.

4. Wavelet OFDM Principles

A wavelet is normally assigned the square integrable function ψ(t) to illustrate the wavelet fundamental definition [10]. In other literature [7], it is also indicated by $\psi(\mathrm{t}) \in L^2(R)$ where L is a Lebesque integral and 2 signifies the integral of the square of the modulus of the function, and R denotes the real number for integration of the independent variable t. In this section, we discuss two principles of wavelet transforms, orthogonal and biorthogonal wavelet as follows.

4.1. Orthogonal Wavelets

The Fourier transform has exponential parts consisting of cosine and sine signal bases. These bases are orthogonal to each other. The wavelet transform also has orthogonal bases. Its bases are low pass and high pass filters which are associated with the scaling and wavelet functions respectively. Among orthogonal wavelets are Daubechies, Coiflets, Morlet and Meyer [6].

Orthogonal wavelet functions can be generated by scaling and shifting properties as follows [10]:

$$\psi_{ab}\left(t\right) = \frac{1}{\sqrt{2}}\psi\left(\frac{t-b}{a}\right) \quad (17)$$

where *a* and *b* are the scaling and shifting real parameter values. According to [11], the wavelet transform is called continuous if *a* and *b* are continuous. The drawbacks of a continuous wavelet transform are redundancy and impracticality. To avoid these problems, those parameters have to be discredited as follows [10,12]:

$$\begin{aligned}a &= a_0^m\\ b &= nb_0 a\end{aligned} \quad (18)$$

where *m* and *n* indicates the exponential integers. From (17) and (18), the basis of the DWT can be formed as

$$\psi_{mn}\left(t\right) = a_0^{-\frac{m}{2}}\psi\left(a_0^{-m}t-nb_0\right) \quad (19)$$

Using $a_0 = 2$ and $b_0 = 1$, we can have the signal function

$$\begin{aligned}U\left(t\right) &= \sum_{n=-\infty}^{\infty} C_{L,n}\phi\left(2^{-L}t-n\right)\\ &+ \sum_{m=1}^{L}\sum_{n=-\infty}^{\infty} D_{mn}2^{-\frac{m}{2}}\psi\left(2^{-m}t-n\right)\end{aligned} \quad (20)$$

where the scaling coefficient $C_{L.n}$ is

$$\begin{aligned}C_{L.n} &= \left\langle U\left(t\right),\phi_{L,n}\left(t\right)\right\rangle\\ &= 2^{-\frac{L}{2}}\int U\left(t\right)\phi\left(2^{-L}t-n\right)dt\end{aligned} \quad (21)$$

where $\phi_{\mathrm{L,n}}\left(t\right) = 2^{-\frac{L}{2}}\phi\left(2^{-L}t-n\right)$ and the wavelet coefficient D_{mn} is

$$\begin{aligned}D_{mn} &= \left\langle U\left(t\right),\psi_{mn}\left(t\right)\right\rangle\\ &= 2^{-\frac{L}{2}}\int U\left(t\right)\psi\left(2^{-L}t-n\right)dt\end{aligned} \quad (22)$$

In (20), the time domain signal *U*(*t*) is DWT transformed to scales in which all the coefficients are denoted as the scales [10]. *U*(*t*) can also be called the finite resolution wavelet representation [12]. The sum of scaled *φ*(2*t*) can make up the parent scaling function, and can be expressed as [7,10]:

$$\phi\left(t\right) = \sqrt{2}\sum_{n} h_n\phi\left(2t-n\right) \quad (23)$$

where the coefficients h_n are a sequence of real or perhaps complex numbers called the scaling function (or scaling vector or filter). The use of $\sqrt{2}$ is to maintain the norm of the scaling function with the scale of 2. This scaling function in (23) can also be used for the multiresolution analysis (MMRA) [13]. A fundamental wavelet function can be expressed as a linear combination of translates of the scaling function as follows [10,12]:

$$\psi(t) = \sqrt{2}\sum_n g_n \phi(2t-n) \tag{24}$$

where the wavelet coefficients g_n are related to the scaling coefficients h_n by

$$g(n) = (-1)^n h_{1-n} \tag{25}$$

An example of the application of (23) is the Haar scaling function which is given by [7] as follows:

$$\phi_H(t) = \phi(2t) + \phi(2t-1) \tag{26}$$

It can be seen that $\phi(2t)$ can be used to construct $\phi_H(t)$. It also can be noted that (26) is the result of (23) for the first 2 sequence of discrete samples of n with coefficients $\mathrm{h}(0) = \frac{1}{\sqrt{2}}, \mathrm{h}(1) = \frac{1}{\sqrt{2}}$ [7]. Examples of Haar scaling and wavelet functions are shown in Figure 4.

The Haar wavelet can be categorised as an orthogonal wavelet. All Daubechies wavelet families are categorised as orthogonal wavelets [6]. Another figure of a Daubechie wavelet such as db2 is shown in Figure 5.

4.2. Biorthogonal Wavelets

Biorthogonal wavelets are different than orthogonal wavelets because they have biorthogonal bases. Their bases have symmetric perfect reconstruction properties with compactly support. They also have two duality functions for each scaling and wavelet functions which are ϕ and $\hat{\phi}$ for the scaling filters, and ψ and $\hat{\psi}$ for the wavelet filters accordingly. In MATLAB, we have built-in functions such as bior1.1, bior2.2, bior5.5, rbio1.1, rbio2.2 and rbio5.5. The number next to the wavelet name refers to the length of the filter in the decomposition and reconstruction filters respectively.

Biorthogonal wavelets can be constructed from orthogonal wavelets by considering the duality concept. Let ϕ and $\hat{\phi}$ be two scaling functions and let ψ and $\hat{\psi}$ be two wavelet functions, then we can express the biorthogonal scaling and wavelet functions as follows [5,14]:

$$\begin{aligned} \langle \phi(t), \hat{\phi}(t) \rangle &= \delta_n \\ \langle \psi(t), \hat{\psi}(t) \rangle &= \delta_k \\ \langle \psi(t), \hat{\phi}(t) \rangle &= 0 \\ \langle \hat{\psi}(t), \phi(t-n) \rangle &= 0 \end{aligned} \tag{27}$$

where $\hat{\phi}(t) = \sqrt{2}\sum_n \hat{h}_n \hat{\phi}(2t-n)$ and $\hat{\psi}(t) = \sqrt{2}\sum_n \hat{h}_n \hat{\psi}(2t-n)$ with δ_n and δ_k are the results of biorthogonal bases. The last two equations in (27) satisfy the orthogonality properties. One advantage of using biorthogonal wavelets is that the scaling and wavelet functions are symmetric due to the duality concept [5,7], therefore, biorthogonal wavelets provide an advantage over orthogonal

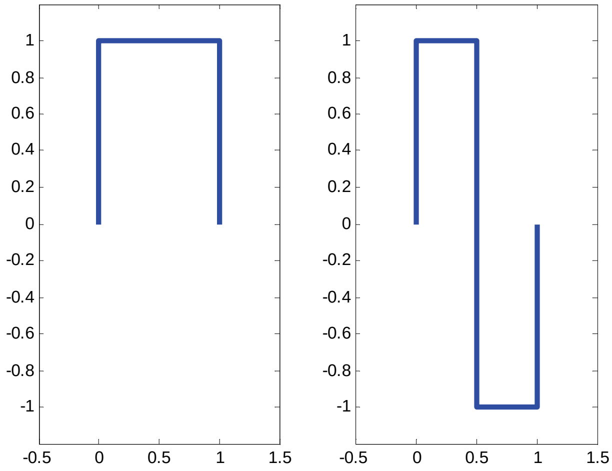


Figure 4. Haar (db1) scaling function $\phi(t)$ (left) and wavelet function $\psi(t)$ (right).

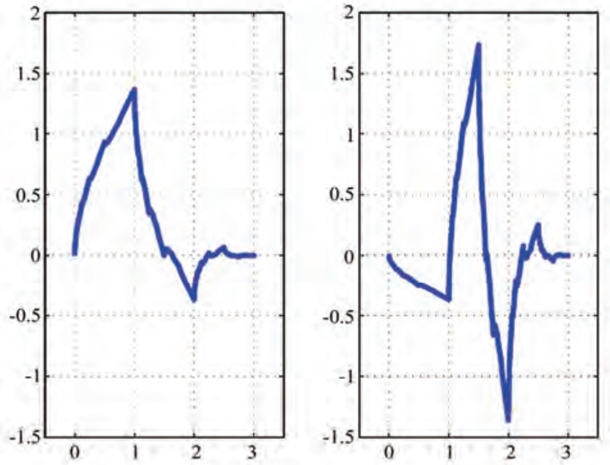


Figure 5. Db2 scaling function $\phi(t)$ (left) and wavelet function $\psi(t)$ (right). Note that this plot is similar to [5] p. 197 and [7] p. 81.

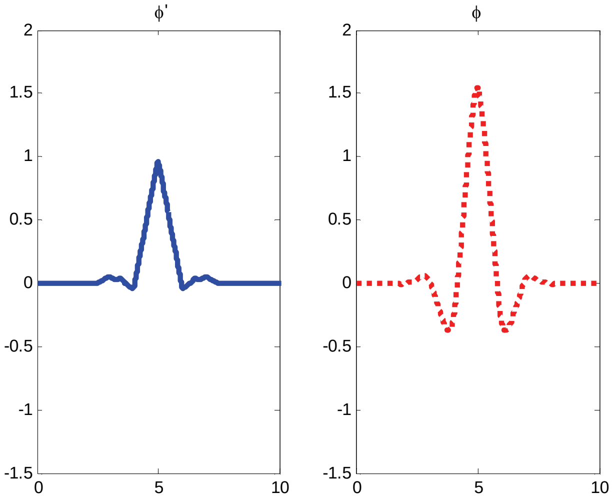


Figure 6. Bior5.5 shows duality concept with two scaling functions, $\hat{\phi}$ (left) and ϕ (right). Note that this plot is similar to [5] p. 280.

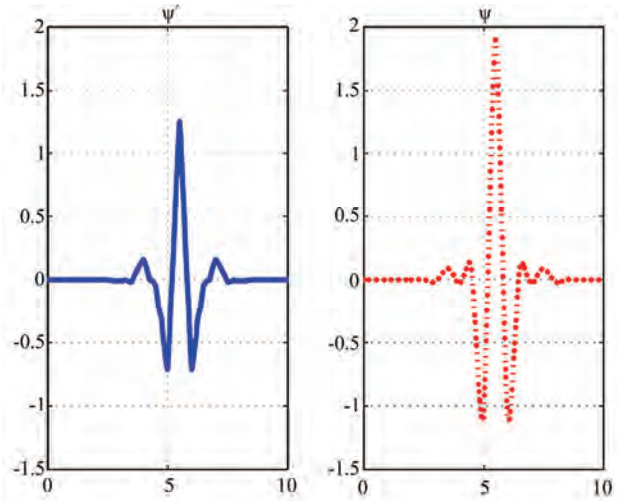


Figure 7. Bior5.5 shows duality concept with two wavelet functions, $\hat{\psi}(t)$ (left) and $\psi(t)$ (right). Note that this plot is similar to [5] p. 280.

wavelets because they offer not only orthogonality but also symmetry. In [15], comparing orthogonal transforms, biorthogonal transforms relax some of the constraints on the mother wavelet (or filters) and allow the mother wavelet to be symmetric and have linear phase. The plots of biorthogonal scaling and wavelet functions are shown in Figure 6 and Figure 7.

5. System Model of Wavelet-Based OFDM

The wavelet transform blocks comprise of an inverse discrete wavelet transform (IDWT) at the transmitter and a discrete wavelet transform (DWT) at the receiver as shown in Figure 8. Due to the overlapping nature of wavelets, the wavelet-based OFDM does not need a cyclic prefix to deal with the delay spreads of the channel. As a result, it has a higher spectral containment than in Fourier based OFDM [8,9]. The DWT-OFDM system model comprise of low pass as LPF filter coefficients and h as HPF filter coefficients, the orthonormal bases are satisfied by four possible ways as follows [6]:

$$\langle g, g^* \rangle = 1 \tag{28}$$

$$\langle h, h^* \rangle = 1 \tag{29}$$

$$\langle g, h^* \rangle = 0 \tag{30}$$

$$\langle h, g^* \rangle = 1 \tag{31}$$

where (28) or (29) is related to the normal property and (30) or (31) is for orthogonal property accordingly. The commas and star symbols in Equations (28) to (31) above are referring to the dot product and transpose vector accordingly. Both filters are assumed having perfect reconstruction property. This means that the input and output of the two filters are expected to be the same. The g and h coefficients can be further described as having convolution operations to perform as orthonormal wavelets which can be represented as [16]

$$\begin{aligned} \alpha_i(n) &= h\left(\frac{n}{2^i}\right) * g\left(\frac{n}{2^i}\right) * \ldots * g\left(\frac{n}{2^{i-j}}\right) * \ldots * g(n) \\ \alpha_{N-1}(n) &= g\left(\frac{n}{2^{N-2}}\right) * g\left(\frac{n}{2^{N-1}}\right) * \ldots \\ &\quad \ldots * g\left(\frac{n}{2^{i-j}}\right) * \ldots * g(n) \end{aligned} \tag{32}$$

where $(i-j)$ is a positive integer for $i, j \in 0,1,\ldots, N-2$.

The signal is up-sampled and filtered by the LPF coefficients or namely as approximated coefficients.

The system model in Figure 8 is assumed that there is no frequency offset so that the DWT itself acts as a matched filter at the receiver. To determine the data in sub-channel k, we match the transmitted waveform with carrier i [17]:

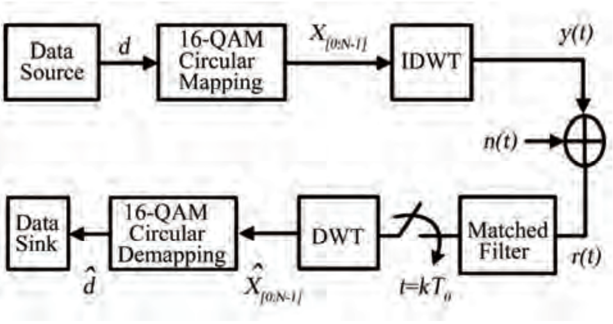


Figure 8. The system model of wavelet based OFDM transceiver.

$$\left\langle y(t), W_i(\ )\right\rangle = \sum_{k=0}^{N-1} d_k \left\langle W_k(t), W_i(t)\right\rangle \tag{33}$$

where $y(t)$ is the transmitted data via IDWT, d_k is the data projected onto each carrier, $W_k(t)$ are the complex exponentials or it can be written as $e^{j2\pi 2\frac{m}{N}}$, $\left\langle W_k(t), W_i(t)\right\rangle$ equals 1 when $k = i$ and 0 when $k \neq i$.

In a typical communication system, data is transmitted over a dispersive channel. The impulse response of a deterministic (and possibly time-varying) channel can be modeled by a linear filter $h(t)$:

$$\begin{aligned} r(t) &= y(t) * h(t) + n(t) \\ &= \sum_{k=0}^{N-1} d_k W_k^{'} + \sum_{l=0}^{g-1}\sum_{k=0}^{N-1} d_{k,l} W_k^{'}(t - lk) + n(t) \end{aligned} \tag{34}$$

where $\mathrm{W}_\mathrm{k}^{'} = W_k(t) * h(t)$, and g (g > 1) is the wavelet *genus* so that Ng is the filter order (number of taps in that subband), and $n(t)$ is an additive white Gaussian noise. Due to the overlapped nature of the wavelet-based OFDM, it requires g symbol periods, for a genus g system, to decode one data vector [17]. This is the reason of having the wavelet transforms of g-1 other data vectors in the second term of (34).

After matched-filtering with carrier i, the signal becomes

$$\begin{aligned} \left\langle r(t), W_i(t)\right\rangle &= \sum_{k=0}^{N-1} d_k \left\langle W_k^{'}(t), W_i(t)\right\rangle \\ &\quad + \sum_{l=1}^{g} \sum_{k=0}^{N-1} d_{k,l} \left\langle W_k^{'}(t - lk), W_i(t - lk)\right\rangle \\ &\quad + \left\langle n(t), f_i(t)\right\rangle \\ &= \sum_{k=0}^{N-1} d_k \rho_{k,0}(0) + \left\langle\left\langle n(t), f_i(t)\right\rangle\right\rangle \\ &= d_k \rho_{i,i}(0) + \sum_{\substack{k=0 \\ k\neq i}}^{N-1} d_k \rho_{k,i}(0) \\ &\quad + \sum_{l=1}^{g} \sum_{\substack{k=0 \\ k\neq i}}^{N-1} d_{k,l} \rho_{k,i}(l) + n''(t) \end{aligned} \tag{35}$$

where $d_k \rho_{i,i}(0)$ is the recovered data with correlation term $\rho_{i,i}(0)$. The second term which is $\sum_{k=0,k\neq i}^{N-1} d_k \rho_{k,i}(0)$ is the interference due to the distorted filters that are no longer orthogonal to one another with correlation terms $\rho_{k,i}(0)$, and $\sum_{l=1}^{g} \sum_{k=0,k\neq i}^{N-1} d_{k,l} \rho_{k,i}(l)$ is the interference term with correlation $\rho_{k,i}(l)$ due to the overlapped nature of the wavelet transform. If the channel has no distortion, only the first and last terms would appear, which result that the decoder would obtain almost the correct signal.

6. System Performance

The performance between the square and circular for 16-QAM for unfiltered constellation points is discussed in Subsection 6.1. A filtered version, which is the processed signal through the matched filter and DWT block shown in Figure 8 at the receiver, is then considered. The result is obtained and discussed in Subsection 6.2.

6.1. Square versus Circular

In this subsection, two main parts are discussed. The first part is to obtain the simulation result for circular 16-QAM from (16) and compare with the square 16-QAM provided by (17) in [4] which is written as follows

$$\begin{aligned} P_{sq} &= \frac{\sqrt{M} - 1}{\sqrt{M}\log_2\left(\sqrt{M}\right)} Q\left(\sqrt{\frac{3\log_2\left(\sqrt{M}\right).E_b}{2(M-1)N_0}}\right) \\ &\quad + \frac{\sqrt{M} - 2}{\sqrt{M}\log_2\left(\sqrt{M}\right)} Q\left(\sqrt{\frac{3\log_2\left(\sqrt{M}\right).E_b}{2(M-1)N_0}}\right) \end{aligned} \tag{36}$$

Note that the square 16-QAM curve in Figure 9 is also approximately similar to the theoretical 16-QAM plot if one uses the Bit Error Rate Analysis Tool (bertool) from Matlab. From the figure, it is shown that the circular scheme slightly outperforms the counterpart scheme at most SNR values.

The second part is to obtain the result for other M-ary QAM. The exact BER analysis for other circular M-ary QAM is performed by changing the value of M in $\mathrm{d} = \sqrt{\frac{3\log_2 \mathrm{M.E_b}}{2(\mathrm{M}-1)}}$ and fix σ accordingly. When M is changed, the parameters A, B and C are consequently affected. Then, they are substituted into (16). Table I shows the summary of the arbitrary parameters due to varying M. Figure 9 also shows the BER results for other circular schemes with comparisons of other square QAM. The circular of other M-ary QAM are also slightly better than the square schemes in most SNR values. The simulation results show that they met the theoretical analysis.

6.2. Wavelet Based OFDM

To simulate the system using wavelet based OFDM (W-OFDM), the orthogonal wavelet family such as Daubechies,

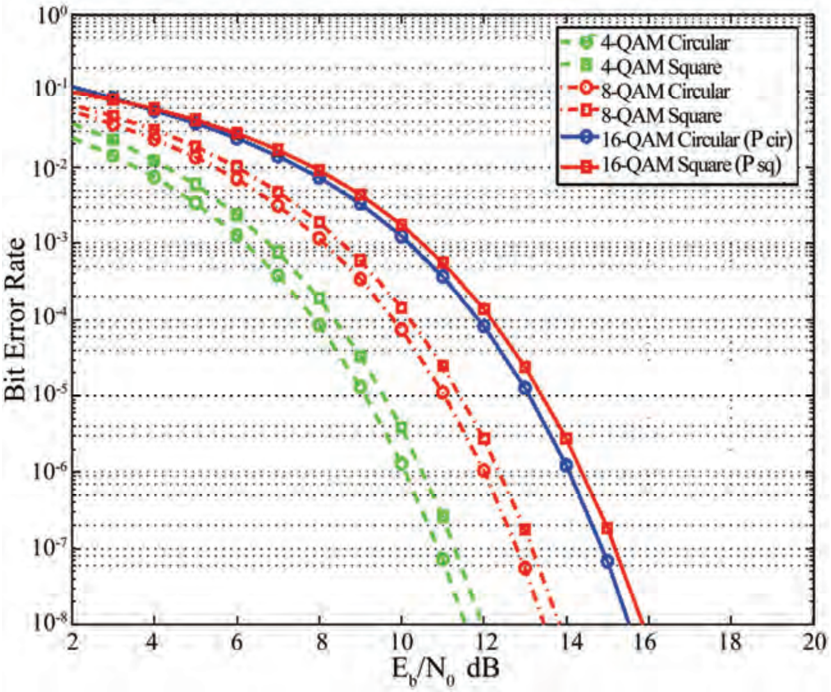


Figure 9. Exact BER of circular and square M-ary QAM.

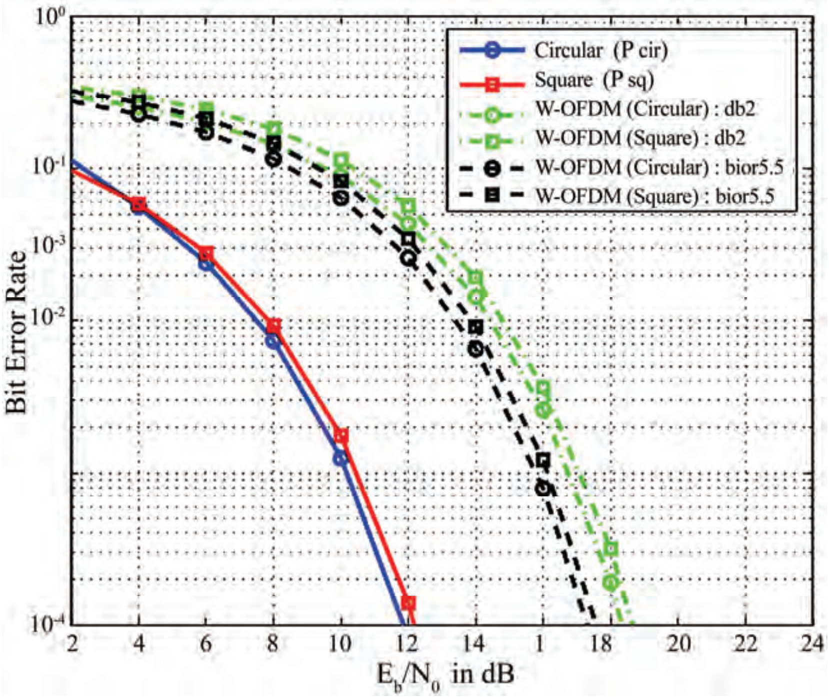


Figure 10. Comparison of circular and square of 16-QAM wavelet (db2 and bior5.5) and Fourier based OFDM.

db2 with comparison of the biorthogonal wavelet family, bior5.5 are considered. From Figure 10, it is shown that circular 16-QAM has better outperformed the square scheme in most SNR values. It is interesting to observe that the W-OFDMs results have less BER performance compared to P_{cir} and P_{sq}. The result is the effect of the filtered version that have been through the imperfect functional components in the receiver such as distorted filters and an additive white Gaussian noise channel as indicated in the previous section.

Table 1. Summary of parameters for circular *M*-Ary (*M* ≤ 16) Qam.

	M=4	M=8	M=16
d	$\sqrt{E_b}$	$\sqrt{\frac{9}{14}E_b}$	$\sqrt{\frac{2}{5}E_b}$
γ	$2\sqrt{5\frac{E_b}{N_0}}$	$2\sqrt{\frac{45}{14}\frac{E_b}{N_0}}$	$2\sqrt{2\frac{E_b}{N_0}}$
A	$Q\left(\sqrt{5\frac{E_b}{N_0}}\right)$	$Q\left(\sqrt{\frac{45}{14}\frac{E_b}{N_0}}\right)$	$Q\left(\sqrt{2\frac{E_b}{N_0}}\right)$
B	$Q\left(b\sqrt{5\frac{E_b}{N_0}}\right)$	$Q\left(b\sqrt{\frac{45}{14}\frac{E_b}{N_0}}\right)$	$Q\left(b\sqrt{2\frac{E_b}{N_0}}\right)$
C	$Q\left(c\sqrt{5\frac{E_b}{N_0}}\right)$	$Q\left(c\sqrt{\frac{45}{14}\frac{E_b}{N_0}}\right)$	$Q\left(c\sqrt{2\frac{E_b}{N_0}}\right)$

Note: b=0.464, c=2.464 and Q(.)=erfc(.)

7. Conclusions

The optimal circular 16-QAM constellation points and the analysis of its exact BER calculation have been derived. The work has also been applied to wavelet based OFDM systems to compare the circular and square schemes. The results showed that the circular were slightly better than the square counterparts. When applying wavelet based OFDM using different wavelet families (orthogonal and biorthogonal), the same results were also obtained that the circular has slightly outperformed the square.

8. References

[1] J. G. Proakis, "Digital communications," Fourth edition, New York: McGraw-Hill, 2001.

[2] R. V. Nee and R. Prasad, "OFDM for wireless multimedia communications," Boston: Artech House, 2000.

[3] M. P. Fitz and J. P. Seymour, "On the bit error probability of QAM modulation," International Journal of Wireless Information Networks, Vol. 1, No. 2, pp. 131–139, 1994.

[4] K. Cho and D. Yoon, "On the general BER expression of one and two dimensional amplitude modulations," IEEE Transactions on Communications, Vol. 50, No. 7, pp. 1074–1080, July 2002.

[5] I. Daubechies, "Ten lectures on wavelets," Philapdelphia: Society for Industrial and Applied Mathematics, 1992.

[6] M. Weeks, "Digital signal processing using matlab and wavelets," Infinity Science Press LLC, 2007.

[7] C. S. Burrus, R. A. Gopinath, and H. Guo, "Introduction to wavelets and wavelet transforms," Upper Sadle River, NJ: Prentice-Hall, 1998.

[8] R. Mirghani and M. Ghavami, "Comparison between wavelet-based and Fourier-based multicarrier UWB systems," IET Communications, Vol. 2, No. 2, pp. 353–358, 2008.

[9] S. D. Sandberg and M. A. Tzannes, "Overlapped discrete multitone modulation for high speed copper wire communications," IEEE Journal on Selected Areas in Communications, Vol. 13, No. 9, pp. 1571–1585, 1995.

[10] F. Xiong, "Digital modulation techniques," Second edition, Boston: Artech House, 2006.

[11] I. Daubechies, "Orthonormal bases of compactly supported wavelets," Communications in Pure and Applied Math., Vol. 41, pp. 909–996, 1988.

[12] A. N. Akansu, "Wavelets and filter banks: A signal processing perspective," Tutorial in Circuit and Devices, November 1994.

[13] L. Cui, B. Zhai, and T. Zhang, "Existence and design of biorthogonal matrix-valued wavelets," Nonlinear Analysis: Real World Applications, Vol. 10, pp. 2679–2687, 2009.

[14] R. M. Rao and A. S. Bopardikar, "Wavelet transforms: Introduction to theory and applications," MA: Addison-Wesley, 1998.

[15] R. K. Young, "Wavelet theory and its applictions," Massachusetts: Kluwer Academic, 1993.

[16] B. G. Negash and H. Nikookar, "Wavelet based OFDM for wireless channels," Vehicular Technology Conference, 2001.

[17] N. Ahmed, "Joint detection strategies for orthogonal frequency division multiplexing," Dissertation for Master of Science, Rice University, Houston, Texas, pp. 1–51, April 2000.

Int. J. Communications, Network and System Sciences, **2009, 2, 845-851**
doi:10.4236/ijcns.2009.29098 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Frequency-Domain Receivers for Rate-1 Space-Time Block Codes

Mário Marques da Silva[1,2,3], Rui Dinis[1,4], Américo M. C. Correia[1,5]
[1]*Instituto de Telecomunicações*, Lisbon, *Portugal*
[2]*Centro de Estudos de Sistemas de Informação e Tecnologias Informáticas, Portugal*
[3]*Universidade Autónoma de Lisboa, Lisboa, Portugal*
[4]*Universidade Nova, Lisboa, Portugal*
[5]*Instituto Superior de Ciências do Trabalho e da Empresa, Instituto Universitário de Lisboa, Lisboa, Portugal*
E-mail: *marques.silva@ieee.org, rdinis@netcabo.pt, americo.correia@iscte.pt*

Received September 24, 2009; *revised October* 28, 2009; *accepted November* 30, 2009

Abstract

This paper considers iterative frequency-domain receivers for block transmission techniques with rate-1 Space Time Block Coding (STBC) for two and four transmit antennas using both Orthogonal Frequency Division Multiplexing (OFDM) and Single-Carrier (SC) schemes. The proposed receiver includes an interference canceller which enhances the performance of the non-orthogonal STBC scheme with 4 transmit antennas, allowing performances close to those of orthogonal codes. Our performance results show that combining STBC with block transmission techniques allows excellent performances.

Keywords: SC-FDE, Turbo Equalization, STBC, OFDM

1. Introduction

Block transmission techniques, with appropriate cyclic prefixes and employing FDE techniques (Frequency-Domain Equalization), have been shown to be suitable for high data rate transmission over severely time-dispersive channels [1,2]. OFDM (Orthogonal Frequency Division Multiplexing) is the most popular modulation based on this technique.

Single Carrier modulation using FDE is an alternative approach based on this principle. As with OFDM, the data blocks are preceded by a cyclic prefix, long enough to cope with the overall channel length. Due to the lower envelope fluctuations of the transmitted signals, and implicitly a lower PMEPR (Peak-to-Mean Envelope Power Ratio), Single Carrier – Frequency Domain Equalization (SC-FDE) schemes (also named as Single Carrier-Frequency Domain Multiple Access (SC-FDMA) are especially interesting for the uplink transmission (i.e., the transmission from the mobile terminal to the base station) [1,2].

OFDM transmission technique has been selected for the downlink of Long Term Evolution (LTE) in Release 8 of Third Generation Partnership Project (3GPP), as opposed to WCDMA which is the air interface technique that has been selected by European Telecommunications Standard Institute (ETSI) for UMTS. Moreover, SC-FDE technique has been selected for the uplink of LTE in Release 8 of 3GPP, to be deployed in 2010.

A promising Iterative Block–Decision Feedback Equalization technique (IB-DFE) for SC-FDE was proposed in [3] and extended to other scenarios in [4] and [5]. These IB-DFE receivers can be regarded as iterative DFE receivers where the feedforward and the feedback operations are implemented in the frequency domain, enhancing the performance as compared to non-iterative methods [3–5].

Transmit Diversity (TD) techniques are particularly interesting for fading channels where it is difficult to have multiple receive antennas (as in conventional receiver diversity schemes). A possible scenario is the downlink transmission where the base station uses several transmittal antennas and the mobile terminal has a single one [6,7].

The application of Alamouti like transmit diversity in OFDM schemes is more-or-less straightforward [8]. With respect to SC-FDE schemes, [9] proposed a way of combining it with a linear FDE. This technique was extended to SC-FDE with IB-DFE in [10].

In this paper, we consider transmit diversity schemes for both OFDM and SC-FDE schemes, specifically the STBC with two [6,7] and four antennas [11,12]. The same concept can be used in STBC based Multiple Input Multiple Output (MIMO) schemes by adopting receive diversity. For OFDM schemes we consider conventional receiver and for SC-FDE schemes we consider IB-DFE receivers. For non-orthogonal codes (i.e., with more than two transmit antennas), we also consider iterative receivers with cancellation of the residual interference (for SC schemes with IB-DFE receivers, this means a negligible increase on the receiver complexity).

This paper is organized as follows. The system considered in this paper is introduced in Section 2 and Section 3 describes the proposed iterative receiver structure for SC-FDE systems with transmit diversity. A set of performance results is presented in Section 4 and Section 5 contains the conclusions of this paper.

2. System Characterization

2.1. Space Time Block Coding for Two Antennas

We consider block transmission schemes and the *l*th transmitted block has the form

$$s_l\left(t\right)=\sum_{n=-N_G}^{N-1} s_{n,l}h_T\left(t-nT_S\right) \tag{1}$$

with T_s denoting the symbol duration, N_G denoting the number of samples at the cyclic prefix and $h_T(t)$ is the adopted pulse shaping filter. For a single transmit antenna system, the signal $S_l(t)$ is transmitted over a time-dispersive channel and the signal at the receiver input is sampled and the cyclic prefix is removed, leading to the time-domain block $\left\{y_{n,l};n=0,1,\ldots,N-1\right\}$, which is then subject to the frequency domain equalization. For SC-FDE schemes the *l*th time-domain block to be transmitted is $\left\{s_{n,l};n=0,1,\ldots,N-1\right\}$, where $S_{n,\,l}$ is the *n*th data symbol, selected from a given constellation (e.g., a QPSK constellation) under an appropriate mapping rule (it is assumed that $s_{-n,l}=s_{N-n,l}$, $n=-N_G,-N_G+1,\ldots,-1$); the frequency-domain blocks associated with the data are $\left\{S_{k,l};k=0,1,\ldots,N-1\right\}=DFT\left\{s_{n,l};n=0,1,\ldots,N-1\right\}$. For OFDM schemes, the data symbols are transmitted in the frequency domain, i.e., $S_{k,\,l}$ are selected according to an appropriate constellation. At the output of the FDE we have the samples $\tilde{A}_{k,l}=Y_{k,l}H_{k,l}^{*}\Big/\left(\alpha+\left|H_{k,l}\right|^{2}\right)$. In the OFDM case this equalization process is simply accomplished through $\tilde{A}_{k,l}=Y_{k,l}H_{k,l}^{*}$.

If we employ Alamouti's transmit diversity we need some processing at the transmitter. The Alamouti's coding can be implemented either in the time domain or in the frequency domain. In this paper we consider time-domain coding, although the extension to frequency domain coding is straightforward. By considering the Space Time Block Coding with two transmit antennas, the time-domain blocks to be transmitted by the *m*th antenna (*m* = 1 or 2) are $\left\{s_{n,l}^{(m)};n=0,1,\ldots,N-1\right\}$, with

$$\begin{aligned} s_{n,2l-1}^{(1)} &= a_{n,2l-1} \\ s_{n,2l-1}^{(2)} &= -a_{n,2l}^{*} \\ s_{n,2l}^{(1)} &= a_{n,2l} \\ s_{n,2l}^{(2)} &= a_{n,2l-1}^{*} \end{aligned} \tag{2}$$

Considering the matrix-vector representation, this is equivalent to

$$\mathbf{A}_{n,[1,2]}=\begin{bmatrix} a_{n,1} & a_{n,2} \\ -a_{n,2}^{*} & a_{n,1}^{*} \end{bmatrix} \tag{3}$$

Assuming that the cyclic prefix is longer than the overall channel impulse response of each channel, the *l*th frequency-domain block after the FDE block (i.e., the DFT of the *l*th received time-domain block, after removing the cyclic prefix) is

$\left\{y_{n,l};n=0,1,\ldots,N-1\right\}=IDFT\left\{Y_{k,l};k=0,1,\ldots,N-1\right\}$,

with

$$Y_{k,l}=S_{k,l}^{(1)}H_{k,l}^{(1)}+S_{k,l}^{(2)}H_{k,l}^{(2)}+N_{k,l} \tag{4}$$

where $\left\{H_{k,l}^{(m)};k=0,1,\ldots,N-1\right\}=DFT\left\{h_{n,l}^{(m)};n=0,1,\ldots,N-1\right\}$ denotes the channel frequency response for the *k*th subcarrier and the *m*th transmit antenna (the channel is assumed invariant in the frame) and $N_{k,\,l}$ is the frequency-domain block channel noise for that subcarrier and the *l*th block. Assuming, for now, the conventional linear FDE for SC schemes, the Alamouti's post-processing for two antennas (denoted in this paper STBC2) comes,

$$\begin{aligned} \tilde{A}_{k,2l-1} &= \left[Y_{k,2l-1}H_{k,l}^{(1)*}+Y_{k,2l}^{*}H_{k,l}^{(2)}\right]\beta_k^{(2)} \\ \tilde{A}_{k,2l} &= \left[Y_{k,2l}H_{k,l}^{(1)*}-Y_{k,2l-1}^{*}H_{k,l}^{(2)}\right]\beta_k^{(2)} \end{aligned} \tag{5}$$

where $\left\{A_{k,m},k=0,1\ldots,N\right\}=DFT\left\{a_{n,m},n=0,1\ldots,N\right\}$ and where $\beta_k^{(2)}=\left(\alpha+\left(\left|H_{k,l}^{(1)}\right|^2+\left|H_{k,l}^{(2)}\right|^2\right)\right)^{-1}$. This leads to $\tilde{A}_{k,2l-j}=A_{k,2l-j}\sum_{m=1}^{M}\left|H_{k,l}^{(m)}\right|^2\beta_k^{(2)}+N_{k,2l-j}^{eq}\quad j=0,1$. In addition, we define $\alpha=E\left[\left|N_{k,l}\right|^2\right]/E\left[\left|S_{k,2l-j}\right|^2\right]$. $N_{k,l}^{eq}$ denotes the equivalent noise for detection purposes, with $E\left[\left|N_{k,l}^{eq}\right|^2\right]=\left[2\sigma_N^2\left(\sum_{m=1}^{M}\left|H_{k,l}^{(m)}\right|^2\right)\right]\beta_k^{(2)}$, and with $\sigma_N^2=E\left[\left|N_{k,l}\right|^2\right]\Big/2$.

The Alamouti's post-processing for OFDM signals is the same as defined in (5) but without multiplying by the $\beta_k^{(2)}$ component.

2.2. Space Time Block Coding for Four Antennas

Using unspecified complex valued modulation, such an improvement is possible only for the two antenna scheme. Higher schemes with 4 and 8 antennas with code rate one exists only in the case of binary transmission [13]. The proposed STBC4 scheme has *M*=4 transmit antennas, presenting a code rate one. The symbol construction can be generally written as [11–12]

$$\mathbf{A}_{n,[1,4]} = \begin{bmatrix} \mathbf{A}_{n,[1,2]} & \mathbf{A}^*_{n,[3,4]} \\ \mathbf{A}_{n,[3,4]} & -\mathbf{A}^*_{n,[1,2]} \end{bmatrix} \qquad (6)$$

where $\mathbf{A}_{n,[3,4]}$ is the same as $\mathbf{A}_{n,[1,2]}$, by replacing the subscripts 1 by 3 and 2 by 4. Similarly to (2), considering the Space Time Block Coding with four transmit antennas, the time-domain blocks to be transmitted by the *m*th antenna (*m* = 1, 2, 3 or 4) are $\left\{s^{(m)}_{n,l}; n=0,1,\ldots,N-1\right\}$, with

$$\begin{array}{llll}
s^{(1)}_{n,4l-3} = a_{n,4l-3} & s^{(1)}_{n,4l-2} = a_{n,4l-2} & s^{(1)}_{n,4l-1} = a_{n,4l-1} & s^{(1)}_{n,4l} = a_{n,4l} \\
s^{(2)}_{n,4l-3} = -a^*_{n,4l-2} & s^{(2)}_{n,4l-2} = a^*_{n,4l-3} & s^{(2)}_{n,4l-1} = -a^*_{n,4l} & s^{(2)}_{n,4l} = a^*_{n,4l-1} \\
s^{(3)}_{n,4l-3} = a^*_{n,4l-1} & s^{(3)}_{n,4l-2} = a^*_{n,4l} & s^{(3)}_{n,4l-1} = -a^*_{n,4l-3} & s^{(3)}_{n,4l} = -a^*_{n,4l-2} \\
s^{(4)}_{n,4l-3} = -a_{n,4l} & s^{(4)}_{n,4l-2} = a_{n,4l-1} & s^{(4)}_{n,4l-1} = a_{n,4l} & s^{(4)}_{n,4l} = -a_{n,4l-3}
\end{array} \qquad (7)$$

The *l*th frequency-domain block after the FDE block (i.e., the DFT of the *l*th received time-domain block, after removing the cyclic prefix) is $\left\{y_{n,l}; n=0,1,\ldots,N-1\right\} = IDFT\left\{Y_{k,l}; k=0,1,\ldots,N-1\right\}$, with

$$Y_{k,l} = S^{(1)}_{k,l}H^{(1)}_{k,l} + S^{(2)}_{k,l}H^{(2)}_{k,l} + S^{(3)}_{k,l}H^{(3)}_{k,l} + S^{(4)}_{k,l}H^{(4)}_{k,l} + N_{k,l} \qquad (8)$$

Assuming, for now, the conventional SC-FDE decoding (i.e., no IB-DFE receiver), the post-processing STBC for four antennas (*M*=4) comes,

$$\begin{aligned}
\tilde{A}_{k,4l-3} &= \left[Y_{k,4l-3}H^{(1)*}_{k,l} + Y^*_{k,4l-2}H^{(2)}_{k,l} - Y^*_{k,4l-1}H^{(3)}_{k,l} - Y_{k,4l}H^{(4)*}_{k,l}\right]\beta^{(4)}_k = \\
&= \overbrace{A_{k,4l-3}\sum_{m=1}^{M}\left|H^{(m)}_{k,l}\right|^2}^{Desired\ Symbol} - \overbrace{C_k A_{k,4l}}^{\mathrm{Re}sidual\ Interference} + N^{eq}_{k,4l-3} \\
\tilde{A}_{k,4l-2} &= \left[Y_{k,4l-2}H^{(1)*}_{k,l} - Y^*_{k,4l-3}H^{(2)}_{k,l} - Y^*_{k,4l}H^{(3)}_{k,l} + Y_{k,4l-1}H^{(4)*}_{k,l}\right]\beta^{(4)}_k = \\
&= A_{k,4l-2}\sum_{m=1}^{M}\left|H^{(m)}_{k,l}\right|^2 + C_k A_{k,4l-1} + N^{eq}_{k,4l-2} \\
\tilde{A}_{k,4l-1} &= \left[Y_{k,4l-1}H^{(1)*}_{k,l} + Y^*_{k,4l}H^{(2)}_{k,l} + Y^*_{k,4l-3}H^{(3)}_{k,l} + Y_{k,4l-2}H^{(4)*}_{k,l}\right]\beta^{(4)}_k = \\
&= A_{k,4l-1}\sum_{m=1}^{M}\left|H^{(m)}_{k,l}\right|^2 + C_k A_{k,4l-2} + N^{eq}_{k,4l-1} \\
\tilde{A}_{k,4l} &= \left[Y_{k,4l}H^{(1)*}_{k,l} - Y^*_{k,4l-1}H^{(2)}_{k,l} + Y^*_{k,4l-2}H^{(3)}_{k,l} - Y_{k,4l-3}H^{(4)*}_{k,l}\right]\beta^{(4)}_k = \\
&= A_{k,4l}\sum_{m=1}^{M}\left|H^{(m)}_{k,l}\right|^2 - C_k A_{k,4l-3} + N^{eq}_{k,4l}
\end{aligned} \qquad (9)$$

with $\beta^{(4)}_k = \left(\alpha + \sum_{m=1}^{M}\left|H^{(m)}_{k,l}\right|^2\right)^{-1}$, where is defined as above (*j*=1,2,3,4), and where $C_k = 2\mathrm{Re}\left\{H^{(1)*}_{k,l}H^{(4)}_{k,l} - H^{(2)}_{k,l}H^{(3)*}_{k,l}\right\} \Big/ \left\{\left(\sum_{m=1}^{M}\left|H^{(m)}_{k,l}\right|^2\right)\right\}$ which stands for the residual interference coefficient generated in the STBC decoding process. In the following we will show how we can remove this residual interference.

3. Receiver Design

In this section we describe an IB-DFE receiver for Space Time Block Coding with four antennas considering SC-FDE signals. The frequency-domain block at the output of the receiver is $\left\{\tilde{A}^{(i)}_{k,4l-j}; k=0,1,\ldots,N-1\right\}$, with

$$\begin{aligned}
\tilde{A}^{(i)}_{k,4l-3} &= \overbrace{Y_{k,4l-3}F^{(i)(1)}_{k,l} + Y^*_{k,4l-2}F^{(i)(2)}_{k,l} - Y^*_{k,4l-1}F^{(i)(3)}_{k,l} - Y_{k,4l}F^{(i)(4)}_{k,l}}^{STBC4\ decoding\ plus\ IB-DFE\ feedforward} \\
&+ \quad \overbrace{+C_k\widehat{A}^{(i)}_{k,4l}}^{Cancellation\ of\ r\mathrm{e}sidual\ interference} \quad - \overbrace{B^{(i)}_{k,l}\overline{A}^{(i-1)}_{k,4l-3}}^{IB-DFE\ feedback} \\
\tilde{A}^{(i)}_{k,4l-2} &= Y_{k,4l-2}F^{(i)(1)}_{k,l} - Y^*_{k,4l-3}F^{(i)(2)}_{k,l} - Y^*_{k,4l}F^{(i)(3)}_{k,l} + Y_{k,4l-1}F^{(i)(4)}_{k,l} \\
&-C_k\widehat{A}^{(i)}_{k,4l-1} - B^{(i)}_{k,l}\overline{A}^{(i-1)}_{k,4l-2} \\
\tilde{A}^{(i)}_{k,4l-1} &= Y_{k,4l-1}F^{(i)(1)}_{k,l} + Y^*_{k,4l}F^{(i)(2)}_{k,l} + Y^*_{k,4l-3}F^{(i)(3)}_{k,l} + Y_{k,4l-2}F^{(i)(4)}_{k,l} \\
&-C_k\widehat{A}^{(i)}_{k,4l-2} - B^{(i)}_{k,l}\overline{A}^{(i-1)}_{k,4l-1} \\
\widehat{A}^{(i)}_{k,4l} &= Y_{k,4l}F^{(i)(1)}_{k,l} - Y^*_{k,4l-1}F^{(i)(2)}_{k,l} + Y^*_{k,4l-2}F^{(i)(3)}_{k,l} - Y_{k,4l-3}F^{(i)(4)}_{k,l} \\
&+C_k\widehat{A}^{(i)}_{k,4l-3} - B^{(i)}_{k,l}\overline{A}^{(i-1)}_{k,4l}
\end{aligned} \qquad (10)$$

where C_K is as defined for (9). The feedforward coefficients are $\left\{F^{(i)(m)}_{k,l}; k=0,1,\ldots,N-1; m=1,2,\ldots,M\right\}$ and the feedback coefficients are $\left\{B^{(i)}_{k,l}; k=0,1,\ldots,N-1\right\}$. The block $\left\{\widehat{A}^{(i-1)}_{n,4l-j}; n=0,1,\ldots,N-1\right\} = DFT\left\{\widehat{a}^{(i-1)}_{n,4l-j}; n=0,1,\ldots,N-1\right\}$, and denotes the DFT transform of the data estimates associated to the previous iteration, i.e., the Hard Decisions associated to the time-domain block at the output of $\left\{\tilde{a}^{(i-1)}_{n,4l-j}; n=0,1,\ldots,N-1\right\} = IDFT\left\{\tilde{A}^{(i-1)}_{k,4l-j}; k=0,1,\ldots,N-1\right\}$. $\left\{\overline{A}^{(i-1)}_{k,4l-j}; k=0,1,\ldots,N-1; j=0,1,2,3\right\}$ denotes the average signal conditioned to the FDE output for the previous iteration $\left\{\overline{a}^{(i-1)}_{n,4l-j}; n=0,1,\ldots,N-1\right\}$ from (19). It is worth noting that since $\tilde{A}^{(i)}_{k,4l-j}$ presents residual interference, the detection of $A^{(i)}_{k,4l-j}$ should be accompanied by the detection of $A^{(i)}_{k,4l-p}$ (with *p*=3-*j*) to allow the cancellation of the residual interference generated in the STBC4 decoding process.

In case of a SISO system, (10) takes the form $\widehat{A}_{k,l}^{(i)} = Y_{k,l}F_{k,l}^{(i)} - B_{k,l}^{(i)}\overline{A}_{k,l}^{(i-1)}$, i.e., there is a single branch (there is no STBC4 decoding) and there is no cancellation of the residual interference. In case of STBC2 (two transmit antennas), there is no residual interference component.

To further improve performance with STBC4 the residual interference to be subtracted (which is a function of the estimate of the symbol that generates interference), we consider an Iterative Interference Cancellation (IIC) that can be implemented as follows:

1) Compute $\widehat{A}_{k,4l-j}^{(i)(q)}$ using (10) without cancelling the residual interference.

2) Based on $\widehat{A}_{k,4l-j}^{(i)(q)}$ from i., compute $\widehat{A}_{k,4l-p}^{(i)(q)}$ after cancelling the corresponding residual interference.

3) Based on $\widehat{A}_{k,4l-p}^{(i)(q)}$ from ii., compute $\widehat{A}_{k,4l-j}^{(i)(q+1)}$ after cancelling the residual interference ($C_k\widehat{A}_{k,4l-p}^{(i)}$).

4) Repeat steps ii. and iii. iteratively to improve the accuracy of $\widehat{A}_{k,4l-p}^{(i)}$ (cancellation of the residual interference), which will finally be used to improve the accuracy of $\widehat{A}_{k,4l-j}^{(i)}$.

It can be shown that the optimum feedback coefficients are described by [3–4].

It can be shown that the optimum feedback coefficients are described by [3–4].

$$B_{k,l}^{(i)} = \sum_{m=1}^{M} F_{k,l}^{(i)(m)} H_{k,l}^{(m)} - 1 \tag{11}$$

and the feedforward coefficients are given by

$$F_{k,l}^{(i)(m)} = \frac{Q_{k,l}^{(m)}}{\left[\alpha + \left(1 - \left(\rho_l^{(i-1)}\right)^2\right)\sum_{m=1}^{M}\left|H_{k,l}^{(m)}\right|^2\right]\gamma_l^{(i)}} \tag{12}$$

with $Q_{k,l}^{(m)} = H_{k,l}^{(m)*}$ for m=1 or 4 and $Q_{k,l}^{(m)} = H_{k,l}^{(m)}$ for m=2 or 3. In the particular case of SISO we only have m=1 (with M=1) and $Q_{k,l} = H_{k,l}^{*}$. In case of STBC of order two (i.e., STBC2), we have $Q_{k,l}^{(m)} = H_{k,l}^{(m)*}$ for m=1 and $Q_{k,l}^{(m)} = H_{k,l}^{(m)}$ for m=2. The parameter $\gamma_l^{(i)}$ is defined as

$$\gamma_l^{(i)} = \frac{1}{N}\sum_{m=1}^{M}\sum_{k=0}^{N-1} F_{k,l}^{(i)(m)} H_{k,l}^{(m)} \tag{13}$$

and the correlation factor $\rho_{4l-j}^{(i-1)}$ is defined as

$$\rho_{4l-j}^{(i-1)} = \frac{E\left[\hat{a}_{n,4l-j}^{(i-1)} a_{n,4l-j}^{*}\right]}{E\left[\left|a_{n,4l-j}\right|^2\right]} \tag{14}$$

It can be shown that, for the QPSK modulation, the correlation coefficient is given by [14]

$$\rho_{4l-j}^{(i)} = \frac{1}{2N}\sum_{n=0}^{N-1}\left(\rho_{n,4l-j}^{I(i)} + \rho_{n,4l-j}^{Q(i)}\right) \tag{15}$$

($\rho_{4l-j}^{(i)}$ is almost independent of l for large values of N, provided that $H_{k,l}^{(m)}$ is constant for the frame duration), as

$$\begin{aligned} \rho_{n,4l-j}^{I(i)} &= \left|\tanh\left(\frac{L_n^{I(i)}}{2}\right)\right| \\ \rho_{n,4l-j}^{Q(i)} &= \left|\tanh\left(\frac{L_n^{Q(i)}}{2}\right)\right| \end{aligned} \tag{16}$$

The LLRs (Log Likelihood Ratios) of the ”in-phase bit” and the ”quadrature bit”, associated to $a_{n,4l-j}^{I(i)}$ and $a_{n,4l-j}^{Q(i)}$, respectively, are given by

$$\begin{aligned} L_n^{I(i)} &= \frac{2}{\sigma_i^2}\tilde{a}_{n,4l-j}^{I(i)} \\ L_n^{Q(i)} &= \frac{2}{\sigma_i^2}\tilde{a}_{n,4l-j}^{Q(i)} \end{aligned} \tag{17}$$

respectively, with

$$\sigma_{i,4l-j}^2 = \frac{1}{2}E\left[\left|a_{n,4l-j} - \tilde{a}_{n,4l-j}^{(i)}\right|^2\right] \approx \frac{1}{2N}\sum_{n=0}^{N-1}\left|\hat{a}_{n,4l-j}^{(i)} - \tilde{a}_{n,4l-j}^{(i)}\right|^2 \tag{18}$$

(as with $\rho_{4l-j}^{(i)}$, $\sigma_{i,4l-j}^2$ is almost independent of l for large values of N, provided that $H_{k,l}^{(m)}$ remains constant for the frame duration).

The conditional average values associated with the data symbols are given by

$$\overline{a}_{n,4l-j}^{(i)} = \tanh\left(\frac{L_{n,4l-j}^{I(i)}}{2}\right) + j\tanh\left(\frac{L_{n,4l-j}^{Q(i)}}{2}\right) \tag{19}$$

Therefore, the several symbols of order jth (j=0,1,2,3) that comprise the STBC4 block need to be decoded independently by the IB-DFE receiver, with the exception of the symbol estimates that originate the residual interference generated in the STBC4 decoding process, as shown in (10). The *IB-DFE with soft decisions* described above does not need to perform the channel decoding in the feedback loop. As an alternative, we can define a *Turbo FDE* that employs the channel decoder outputs, instead of the uncoded “soft decisions” in the feedback loop of the IB-DFE. The main difference between *IB-DFE with soft decisions* and the *Turbo FDE* is in the decision device: in the first case the decision device is a symbol-by-symbol soft-decision (for QPSK constellation this corresponds to the hyperbolic tangent, as in (19));

for the *Turbo FDE* a Soft-In, Soft-Out channel decoder is employed in the feedback loop. The Soft-In, Soft-Out block, that can be implemented as defined in [15], provides the LLRs of both the “information bits” and the ”coded bits”. The input of the Soft-In, Soft-Out block are LLRs of the ”coded bits” at the FDE output, given by (17) and (18).

The receiver for OFDM schemes with STBC2 is straightforward. For OFDM schemes with STBC4, (10) also applies with the difference that there is no feedback component, and the feedforward component only have the numerator of (12). It is worth noting that these STBC schemes can easily be extended to multiple receive antennas.

4. Performance Results

In this section we present a set of performance results concerning the proposed receivers, for both SC-FDE and OFDM schemes with two and four-antenna STBC schemes. We consider both Bit Error Rate (BER) and Block Error Rate (BLER) performances, which are expressed as a function of E_b / N_0, where N_0 is the one-sided power spectral density of the noise and E_b is the energy of the transmitted bits (i.e., the degradation due to the useless power spent on the cyclic prefix is not included).

Each block has N = 256 symbols selected from a QPSK constellation under a Gray mapping rule (similar results were observed for other values of N, provided that $N >> 1$). The pulse shaping filter is raised cosine with roll-off 0.1. The results shown in this paper considers the Pedestrian A propagation environment [16].

The channel is assumed to be invariant during the block. The duration of the useful part of the blocks (N symbols) is 1μs and the cyclic prefix has duration 0.125μs. For SC-FDE systems we considered the *IB-DFE receiver with soft decisions* and the *Turbo FDE*, both with five iterations. Beyond this number the performance improvement was almost negligible.

Linear power amplification is considered at the transmitter and perfect synchronization is assumed at the receiver. The channel encoder is a convolutional code with generators $1+D^2+D^3+D^5+D^6$ and $1+D+D^2+D^3+D^6$, and the coded bits associated to a given block are interleaved and mapped into the constellation points.

Figure 1 considers uncoded BER results for the SC-FDE and a linear FDE receiver (i.e., just the first iteration of the IB-DFE receiver) versus the IB-DFE receiver with soft decisions (i.e., without channel decoding in the feedback loop), in this case with five iterations. Clearly, the increased diversity due to STBC schemes leads to significant performance improvements relatively to the SISO case. From this figure, it is also clear that the IB-DFE performs always better than the linear FDE receiver. It can also be observed that the STBC4 with the linear FDE receiver performs very badly, due to the residual interference (generated in the STBC4 decoding process). However, when we add the IB-DFE with soft decisions to the STBC4, we have a significant performance improvement, namely due to the ability to mitigate the residual interference. It is worth noting that, with the IB-DFE receiver, the STBC4 achieves a performance improvement over the STBC2. It happens because the proposed receiver cancels the interference generated in the STBC4 decoding process. This residual interference is, in fact, the reason why this STBC4 scheme is considered as non-orthogonal. In this case, we have seen that the non-orthogonality is not a reason for loss of performance.

Figure 2 concerns the coded results for the SC-FDE. In this case, the Linear FDE and the Turbo FDE receiveris considered. For the linear FDE receiver, the STBC4 performs worse than the STBC2, due to the residual interference. However, for the Turbo FDE (i.e., the proposed iterative frequency-domain receiver that employs the channel decoder outputs), the STBC4 outperforms

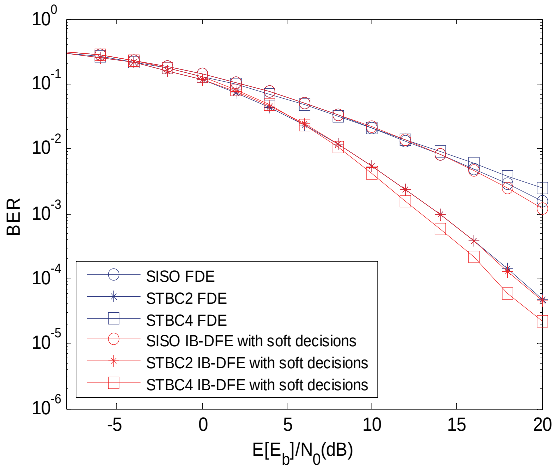


Figure 1. Uncoded BER results for the SC-FDE.

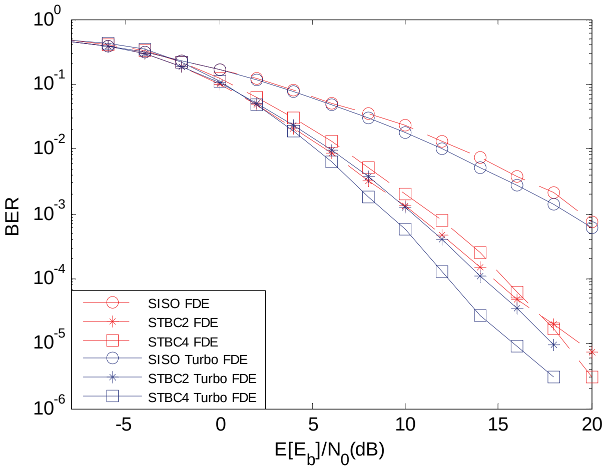


Figure 2. Coded BER results for the SC-FDE.

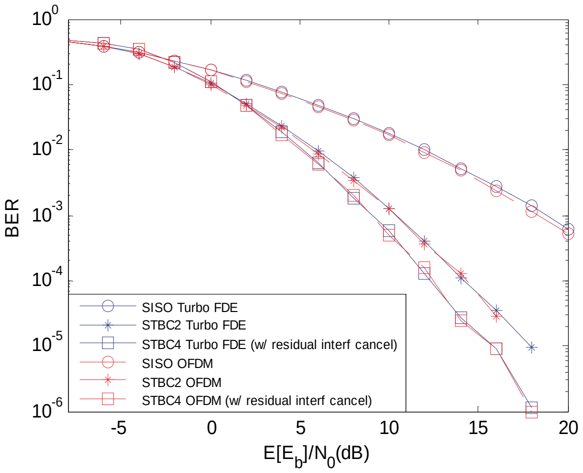


Figure 3. Coded BER for SC-FDE and OFDM.

the STBC2 (and the SISO, as expected). This is a consequence of the additional diversity order and the effective residual interference cancellation inherent to the proposed receiver. Therefore, although using a higher number of antennas leads to an increase in the system complexity, its advantage is clear as long as the proposed iterative receiver is adopted.

Figure 3 shows a performance comparison between SC-FDE and OFDM when channel coding is considered (it is well-know that uncoded performances are very poor for OFDM schemes). Note that the OFDM receiver for the STBC4 also includes a residual interference canceller, similar to the one included and described in the IB-DFE that was considered for the SC-FDE STBC4. The proposed Turbo FDE receiver for SC-FDE signals allows similar or better performance than coded OFDM signals for the STBC schemes considered. However, OFDM technique presents much more demanding requirements in terms of PMEPR, as compared to SC-FDE technique.

Figure 4 shows the uncoded BER performance of STBC4 with and without residual interference cancellation for both SC-FDE (in this case the IB-DFE receiver is considered) and OFDM. From this figure it is seen that, when the residual interference cancellation is considered, SC-FDE with the proposed iterative receiver achieves better results than those achieved with OFDM. Moreover, when we focus on the results without the residual interference cancellation, it is clear the much better results achieved with the SC-FDE due to the inherent ability of the iterative frequency domain SC-FDE receiver to cancel generic interference. In this case, SC-FDE without the residual interference cancellation achieves approximately the same performance than that achieved with the OFDM scheme with the interference cancelled. Finally, it is noticeable the very bad performance obtained with the OFDM technique when the residual interference is not cancelled.

Figure 5 shows the coded BER performance of STBC4 with and without residual interference cancellation for both SC-FDE and OFDM. From this figure it is observed that, when the residual interference cancellation is considered, SC-FDE with the proposed iterative receiver (i.e., the Turbo FDE receiver) achieves similar results to those achieved with OFDM. However, when we focus on the results without the residual interference cancellation, as before, it is clear the better results achieved with the SC-FDE, for higher values of E_b / N_0, due to the inherent ability of the iterative frequency domain receiver (Turbo FDE) to cancel generic interference. Figure 6 presents results similar to Figure 3, but in terms of BLER, instead of the BER. As before, for the same diversity order, SC-FDE schemes achieve similar results as those obtained with the OFDM. The BLER results confirm the advantage of the STBC4 over lower diversity orders.

5. Conclusions

In this paper we considered iterative frequency-do main

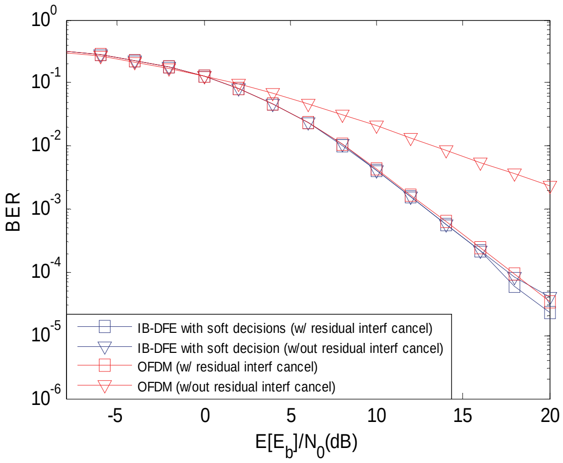


Figure 4. Uncoded BER performance for STBC4 (w/ and w/out residual interference cancellation).

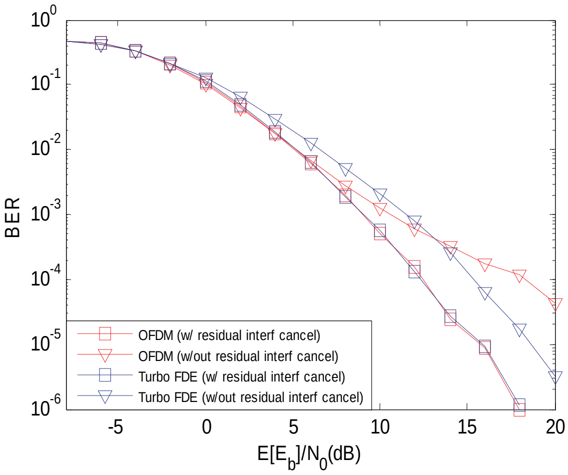


Figure 5. Coded BER performance for STBC4 (w/ and w/out residual interference cancellation).

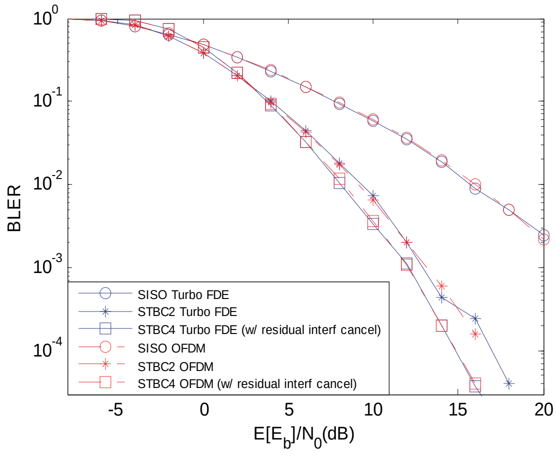


Figure 6. Coded BLER for SC-FDE and OFDM

receivers for SC-FDE technique with code rate-1 STBC using two or four transmit antennas. OFDM technique was also considered in system description and performance results.

Since our STBC with 4 transmit antennas is not orthogonal, our receiver includes the cancellation of the residual interference.

The proposed Turbo FDE receiver for SC-FDE signals allows similar or better performance than coded OFDM signals with the same diversity order. However, OFDM technique presents much more demanding requirements in terms of PMEPR, as compared to SC-FDE technique, limiting its applicability. In this sense, SC-FDE is a good alternative to OFDM transmission technique, especially for the uplink.

It was shown that the best overall performance is achieved with STBC4 schemes, as long as the receiver includes the described residual interference cancellation system. It is worth noting that by adding *N* order receive diversity (*N* receive antennas instead of a single one), the proposed SC-FDE STBC4 receiver keeps being valid and the system can be seen as a $4 \times N$ MIMO system.

6. Acknowledgements

This work was supported by the Portuguese Foundation for the Science and Technology (FCT).

7. References

[1] A. Gusmão, R. Dinis, J. Conceicão, and N. Esteves, “Comparison of two modulation choices for broadband wireless communications,” Proceedings IEEE VTC Spring, pp. 1300–1305, May 2000.

[2] D. Falconer, S. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson, ”Frequency domain equalization for single-carrier broadband wireless systems,” IEEE Communications Magazines, Vol. 4, No. 4, pp. 58–66, April 2002.

[3] N. Benvenuto and S. Tomasin, “Block iterative DFE for single carrier modulation,” IEE Electronic Letters, Vol. 39, No. 19, September 2002.

[4] R. Dinis, A. Gusmão, and N. Esteves, “On broadband block transmission over strongly frequency-selective fading channels,” Wireless 2003, Calgary, Canada, July 2003.

[5] R. Dinis, R. Kalbasi, D. Falconer, and A. Banihashemi, “Iterative layered space-time receivers for single-carrier transmission over severe time-dispersive channels,” IEEE Communication Letters, Vol. 8, No. 9, pp. 579–581, September 2004.

[6] S. M. Alamouti, “A simple transmitter diversity scheme for wireless communications,” IEEE JSAC, pp. 1451–1458, October 1998.

[7] Tarokh, *et al.*, “Space-time block codes from orthogonal designs,” IEEE Transactions on Information Theory, pp. 1456–1467, July 1999.

[8] J. Wang, O. Wen, S. Li, R. Shu, and K. Cheng, “Capacity of alamouti coded OFDM systems in time-varying multipath rayleigh fading channels,” IEEE VTC’06 (Spring), May 2006.

[9] N. Al-Dhahir, “Single-carrier frequency-domain equalization for space-time block-coded transmission over frequency-selective fading channels,” IEEE Communications Letters, Vol. 5, July 2001.

[10] R. Dinis, A. Gusmão, and N. Esteves, “Iterative block-DFE techniques for single-carrier-based broadband communications with transmit/receive space diversity,” IEEE ISWCS’04, Port Louis, Mauritius, September 2004.

[11] M. M. Silva and A. Correia, “Space time block coding for 4 antennas with coding rate 1,” IEEE International Symposium on Spread Spectrum Techniques and Application (ISSSTA), Prague, Check Republic, 2–5 September 2002.

[12] M. M. Silva, A. Correia, and R. Dinis, “Wireless systems on transmission techniques for multi-antenna W-CDMA systems”, European Transactions on Telecommunications, Wiley Interscience, DOI: 10.1002/ett.1252, http://dx.doi.org/10.1002/ett.1252, published on-line in advance of print in 16 November 2007.

[13] B. Hochwald, T. Marzetta, and C. Papadias, “A transmitter diversity scheme for wideband CDMA systems based on space-time spreading,” IEEE Journal on Selected Area in Communications, Vol. 19, No. 1, pp. 48–60, January 2001.

[14] A. Gusmão, P. Torres, R. Dinis, and N. Esteves, “A class of iterative FDE techniques for reduced-CP SC-based block transmission,” International Symposium on Turbo Codes, April 2006.

[15] B. Vucetic and J. Yuan, “Turbo codes: principles and applications,” Kluwer Academic Publication, 2002.

[16] ETSI, “Channel models for hiperLAN/2 in different indoor scenarios,” ETSI EP BRAN 3ERI085B; pp. 1–8, 1998.

Int. J. Communications, Network and System Sciences, **2009, 2, 852-856**
doi:10.4236/ijcns.2009.29099 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Adaptive Channel Estimation in OFDM System Using Cyclic Prefix (Kalman Filter Approach)

P. V. NAGANJANEYULU[1], K. SATYA PRASAD[2]
[1]*Department of ECE, Guntur Engineering College, Guntur, India*
[2]*JNTU, Kakinada, India*
E-mail: {*pvnaganjaneyulu, prasad_kodati*}*@yahoo.co.in*
Received August 31, 2009; *revised October* 1, 2009; *accepted October* 29, 2009

ABSTRACT

OFDM is a promising technique for high data rate transmission and the channel estimation is very important for implementation of OFDM. In this paper, cyclic prefix (CP) can be used as a source of channel information which is originally used to reduce inter symbol interference (ISI). Based on this CP observation, we propose two cross coupled dual Kalman filters to track the channel variations without additional training sequences. One Kalman filter AR parameter estimation and another for fading channel estimation.

Keywords: Cyclic Prefix, Kalman Filter

1. Introduction

In OFDM systems, due to user mobility, each carrier is subject to Doppler shifts resulting in time-varying fading. Thus, the estimation of the fading process over each carrier is essential to achieve coherent symbol detection at the receiver. In that case, training sequence/pilot aided techniques and blind techniques are two basic families for channel estimation. Training based methods require the transmission of explicit pilot sequences followed by suitable filtering. This paper focuses on estimation of fading wireless channels for OFDM, using the ideas of Cyclic Prefix (CP) based estimation and adaptive filtering.

The time-varying fading channels are usually modelled as zero-mean wide-sense stationary circular complex Gaussian processes, whose stochastic properties depend on the maximum Doppler frequency denoted by f_d. According to the Jakes model [1], the theoretical Power Spectrum Density (PSD) of the fading process, is band-limited. Moreover, it exhibits twin peaks at ± f_d. The fading wireless channel statistics can be directly estimated by means of the Least Mean Square (LMS) and the Recursive Least Square (RLS) algorithms as in [2]. Alternatively, Kalman filtering algorithm combined with an Autoregressive (AR) model to describe the time evolution of the fading processes and it provides superior performance over the LMS and RLS based channel estimators in [3]. In addition, when the AR model parameters are unknown, dual filtering algorithms are used to estimate the fading channels.

In this paper, for the channel estimation of OFDM, a system model and architecture over fading channels are presented. In the next section a CP based model and the different channel estimation algorithms Kalman and Dual-Kalman are discussed. The performance results are discussed in the next section, finally simulation results are presented.

1.1. Existing Methods for Channel Estimation

Different Channel Estimation methods are proposed based on training sequence, blind and semi-blind. In practice we either assume the channel is invariant and use the initial training to get the channel estimation are periodically employ training sequence to trap the channel variations. These will cause performance loss or increase the overhead of the system. So, we present that the CP in OFDM which is used to reduce ISI and normally discarded at the receiver can be viewed as a training sequence for channel estimation. In paper [3], channel estimation is proposed by two Kalman filters based on noisy data as training sequence. In this paper, we present two cross coupled Kalman by using CP as training sequence and their performances are compared.

2. System Model

In the following, we consider a low to moderate Doppler environment, which allows for a block fading (quasi-static) channel assumption. This implies that the channel

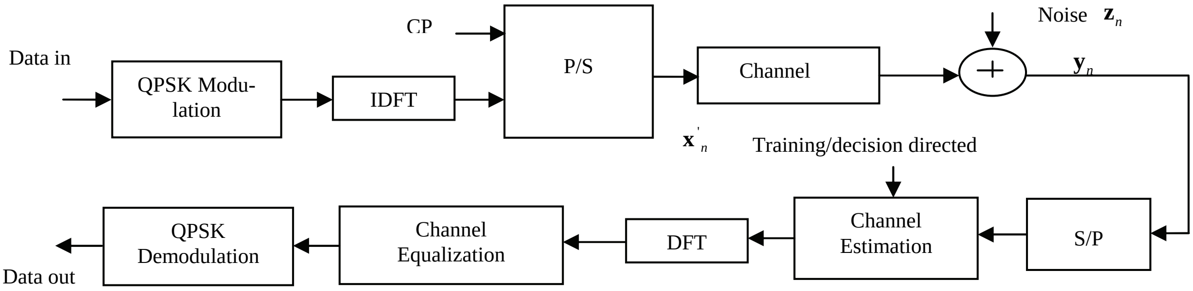


Figure 1. Block schematic of the OFDM system.

tap variations within OFDM symbol duration are negligible, and hence we may define an L×1 channel tap vector for each OFDM symbol as

$$\mathbf{h}_n = [h_n(0)h_n(1)\ldots h_n(L-1)]^T \quad (1)$$

where $h_n(l)$ is the l^{th} channel tap for the n^{th} OFDM symbol.

The classical Doppler spectrum for each of the L channel taps is approximated by an independent AR-2 process [4].

For the l^{th} channel tap at n^{th} OFDM symbol, we have

$$h_n(l) = -a_1 h_{n-1}(l) - a_2 h_{n-2}(l) + v_n(l) \quad (2)$$

where a_1 and a_2 are the AR-2 coefficients are defined in [5] and $v_n(l)$ is the modelling noise for l^{th} tap at symbol n.

2.1. OFDM Architecture over Fading Channel

We consider an OFDM system as in Figure 1 with N data subcarriers. Input data are buffered, converted to a parallel stream and modulated to i.i.d. equi-probable symbols $X_n(k)$, where $X_n(k)$ denotes the k^{th} symbol of the n^{th} OFDM symbol. Each symbol mapped to some complex constellation points, $X_n(k)$, k=0,1,. . ,N-1 at each n. The modulation is implemented by N-point inverse discrete Fourier transform (IDFT) for the symbol vector

$$\mathbf{X}_n = [X_n(0)\ X_n(1)\ldots X_n(N-1)]^T \quad (3)$$

is

$x_n(m+gi) = \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1} X_n(k)e^{j2\pi mk/N}$, $0 \le \mathrm{m} \le N$-1CP of length gi is appended to form the transmitted vector as

$$\mathbf{x}'_n = [x_n(0)x_n(1)\ldots x_n(gi-1) \vdots x_n(gi)x_n(gi+1)\ldots x_n(gi+N-1)]^T \quad (4)$$

where

$$x_n(m) = x_n(N+m), \quad 0 \le m \le gi-1$$

The received symbol corrupted by fading channel and AWGN becomes

$$y_n(m) = \sum_{l=0}^{L-1} h_n(l)x_n(m-l) + z_n(m),\ 0 \le \mathrm{m} \le N+gi+L-1 \quad (5)$$

where n is the OFDM symbol index,

$zn(m)$ is an AWGN sample with zero mean and variance σ^2 at instant m in the n^{th} OFDM symbol.

Demodulation involves removing the cyclic prefix and taking N-point DFT of the received vector to get

$$\mathbf{Y}_n = [Y_n(0)\ \mathrm{Y}_n(1)\ldots \mathrm{Y}_n(N-1)]^T \quad (6)$$

In frequency domain, we have over each subcarrier

$$Y_n(k) = X_n(k)H_n(k) + Z_n(k) \quad (7)$$

where $H_n(k)$ is the channel frequency response at subcarrier k given by

$$H_n(k) = \frac{1}{\sqrt{N}}\sum_{l=0}^{L-1} h_n(l)e^{-j2\pi lk/N},\quad 0 \le k \le N-1 \quad (8)$$

and $z_n(k)$ is the noise on k^{th} subcarrier of n^{th} OFDM symbol i.e.,

$$Z_n(k) = \frac{1}{\sqrt{N}}\sum_{m=0}^{N-1} z_n(m)e^{-j2\pi mk/N},\ 0 \le k \le N-1 \quad (9)$$

At the receiver, the channel estimator is followed by frequency domain equalizer. After equalization, the estimated symbol at the k^{th} symbol becomes

$$\hat{X}_n(k) = \frac{Y_n(k)}{\hat{H}_n(k)} = \frac{X_n(k)H_n(k)}{\hat{H}_n(k)} + \frac{Z_n(k)}{\hat{H}_n(k)} \quad (10)$$

where $\hat{H}_n(k)$ is the estimate of $H_n(k)$ defined in Equation (8). The estimated symbols $\hat{X}_n(k)$ are then demapped to output bits.

3. CP Based Channel Estimation Techniques

This section describes the use of various adaptive filtering algorithms in CP based frame work for channel estimation in OFDM systems. From Equation (5), we know that

$$y_n(m) = h_n(0)x(m) + h_n(1)x(m-1) + \ldots + h_n(L-1)x(m-L+1) + z_n(m) \quad (11)$$

Gathering the received samples of the n^{th} received OFDM symbol for time instants 0≤m≤gi –1, we obtain a gi×1 vector

$$\mathbf{y}_{n,CP} = [y_n(0)y_n(1)y_n(2)\ldots y_n(gi-1)]^T \quad (12)$$

which is the CP of the received OFDM symbol, and

$$\mathbf{z}_{n,CP} = \left[z_n(0)z_n(1)z_n(2)...z_n(gi-1)\right]^T \qquad (13)$$

is the $gi\times1$ vector of AWGN samples affecting the CP part of the n^{th} received OFDM symbol.

3.1. Kalman Filtering (KF) Algorithm

When operating in a non-stationary environment, Kalman filter [6] is known to yield an optimal solution to the linear filter problem. This subsection describes the application of KF to the channel estimation problem in OFDM. For this purpose, the system is formulated as a state-space model, with unknown channel taps comprising the state of the system. We assume that the state $\mathbf{S}_{\text{n}}$, to be estimated at OFDM symbol index n, comprises of channel taps at two consecutive OFDM symbols [7].

$$\mathbf{s}_n = \left[\mathbf{h}_{n-1}\ \mathbf{h}_n\right]^T_{2L\times1} \qquad (14)$$

From Equation (1) and Equation (2) we have

$$\mathbf{h}_n = \left[h_n(0)h_n(1)\ldots h_n(L-1)\right]^T_{L\times1}$$

$$\mathbf{h}_{n-1} = \left[h_{n-1}(0)h_{n-1}(1)\ldots h_{n-1}(L-1)\right]^T_{L\times1}$$

and

$$\mathbf{h}_n = a_1\mathbf{h}_{n-1} + a_2\mathbf{h}_{n-2} + \mathbf{v}_n \qquad (15)$$

From above equations we get

$$\begin{bmatrix}\mathbf{h}_{n-1}\\ \mathbf{h}_n\end{bmatrix} = \begin{bmatrix}\mathbf{0}_{L\times L} & \mathbf{I}_L\\ a_2\mathbf{I}_L & a_1\mathbf{I}_L\end{bmatrix}\begin{bmatrix}\mathbf{h}_{n-2}\\ \mathbf{h}_{n-1}\end{bmatrix} + \mathbf{v}_n - \qquad (16)$$

We observe that Equation (16) provides the basis for forming the process equation as

$$\mathbf{s}_n = \mathbf{B}\mathbf{s}_{n-1} + \mathbf{v}_n \qquad (17)$$

Here, transition matrix

$$\mathbf{B} = \begin{bmatrix}\mathbf{0}_{L\times L} & \mathbf{I}_L\\ a_2\mathbf{I}_L & a_1\mathbf{I}_L\end{bmatrix}_{2L\times2L} \qquad (18)$$

$\mathbf{0}_{L\times L}$ denotes the $L\times L$ matrix of all zeros and $\mathbf{I}_L$ is the $L\times L$ identity matrix.

Process noise vector

$$\mathbf{v}_n = \left[\mathbf{0}_{1\times\text{L}} \vdots v_n(0)v_n(1)...v_n(L-1)\right]^T_{2L\times1} \qquad (19)$$

where $v_n(l)$ is the modelling noise as in (2)

From Equation (11), we have

$$y_n(m) = h_n(0)x(m) + h_n(1)x(m-1) + ... + h_n(L-1)x(m-L+1) + z_n(m)$$

$$\begin{bmatrix}y_n(0)\\ y_n(1)\\ \vdots\\ y_n(gi-1)\end{bmatrix} = \begin{bmatrix}x_n(0) & x_{n-1}(N+gi-1) & \ldots & x_{n-1}(N+gi-L+1)\\ x_n(1) & x_n(0) & x_{n-1}(N+gi-1) & \ldots\\ \vdots & \vdots & \vdots & \vdots\\ x_n(gi-1) & x_n(gi-2) & \ldots & x_n(gi-L)\end{bmatrix}\begin{bmatrix}h_n(0)\\ h_n(1)\\ \vdots\\ h_n(L-1)\end{bmatrix} + z_n(m)$$

where $0 \le m \le gi-1$

We observe from above that following provides the basis for forming measurement equation as

$$\mathbf{y}_{n,CP} = \bar{\mathbf{A}}_n\mathbf{s}_n + \mathbf{z}_{n,CP} \qquad (20)$$

where the measurement matrix $\bar{\mathbf{A}}_n$ in Equation (20) is formed from the matrix $\mathbf{A}_n$ by augmenting it with a null matrix as

$$\bar{\mathbf{A}}_n = \left[\mathbf{0}_{gi\times\text{L}} \vdots \mathbf{A}_n\right]_{gi\times2L} \qquad (21)$$

Here $\mathbf{A}_n$ is a $gi\times L$ matrix of transmitted symbols that determine the CP of the received OFDM symbol.

$$\mathbf{A}_n = \begin{bmatrix}x_n(0) & x_{n-1}(N+gi-1) & \ldots & x_{n-1}(N+gi-L+1)\\ x_n(1) & x_n(0) & x_{n-1}(N+gi-1) & \ldots\\ \vdots & \vdots & \vdots & \vdots\\ x_n(gi-1) & x_n(gi-2) & \ldots & x_n(gi-L)\end{bmatrix}_{gi\times L}$$

Considering that the CP appended to an OFDM symbol is a replication of the last gi values of that symbol, we may write $\mathbf{A}_n$ in terms of transmitted CP value as,

$$\mathbf{A}_n = \begin{bmatrix}x_n(0) & x_{n-1}(gi-1) & x_{n-1}(gi-2) & \ldots & x_{n-1}(gi-L+1)\\ x_n(1) & x_n(0) & x_{n-1}(gi-1) & \ldots & x_{n-1}(gi-L+2)\\ \vdots & \vdots & \vdots & & \vdots\\ x_n(gi-1) & x_n(gi-2) & \cdots & \ldots & x_n(gi-L)\end{bmatrix}_{gi\times L}$$

$\mathbf{A}_n$ has gi rows corresponding to gi consecutive time instants of the CP. The L elements of each row are the transmitted symbol values affecting the received CP value at that instant. This matrix structure assumes that the CP length is at least equal to the number of taps in the channel impulse response, i.e. no inter block interference.

The measurement noise vector $\mathbf{Z}_{n,CP}$, in Equation (20), comprises the $gi\times1$ vector of AWGN samples affecting the cyclic prefix part of the OFDM symbol.

We observe that Equation (17) and Equation (20) provide the basis for forming the process equation and measurement equation, respectively for the state space model, as follows

$$\mathbf{s}_n = \mathbf{B}\mathbf{s}_{n-1} + \mathbf{v}_n$$

$$\mathbf{y}_{n,CP} = \bar{\mathbf{A}}_n\mathbf{s}_n + \mathbf{z}_{n,CP} \qquad (22)$$

A Kalman filter is employed to estimate the unknown state of the system. Cyclic prefix of the received OFDM symbol $\mathbf{y}_{n,CP}$ is given as input observation to Kalman algorithm, the following estimation equations are given by [3]

$$[\boldsymbol{P}_{n|n-1}]_{2L\times 2L} = \mathbf{B}\boldsymbol{P}_{n-1|n-1}\mathbf{B}^H + \mathbf{Q}_1 \tag{23}$$

$$[\boldsymbol{\alpha}_n]_{gi\times 1} = \left[\mathbf{y}_{n,CP} - \bar{\mathbf{A}}_n\hat{\mathbf{s}}_{n-1}\right] \tag{24}$$

$$[\mathbf{C}_n]_{gi\times gi} = \bar{\mathbf{A}}_n\boldsymbol{P}_{n|n-1}\bar{\mathbf{A}}_n^H + \mathbf{Q}_2 \tag{25}$$

$$[\mathbf{K}_n]_{2L\times gi} = \boldsymbol{P}_{n|n-1}\bar{\mathbf{A}}_n^H\mathbf{C}_n^{-1} \tag{26}$$

$$[\hat{\mathbf{s}}_n]_{2L\times 1} = \mathbf{B}\hat{\mathbf{s}}_{n-1} + \mathbf{K}_n\boldsymbol{\alpha}_n \tag{27}$$

$$[\hat{\mathbf{h}}_n]_{L\times 1} = R\hat{\mathbf{s}}_n\,,\quad R = [\mathbf{0}_{L\times L}\ \ \mathbf{I}_L]_{L\times 2L} \tag{28}$$

$$[\boldsymbol{P}_n]_{2L\times 2L} = [\mathbf{I}_{2L} - \mathbf{K}_n\bar{\mathbf{A}}_n]\boldsymbol{P}_{n|n-1} \tag{29}$$

where $\mathbf{K}_n$ is the $2L\times gi$ Kalman gain matrix , $\hat{\mathbf{s}}_n$ is the state estimate at the n^{th} OFDM symbol, $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are the covariance matrices of $\mathbf{v}_n$ and $\mathbf{Z}_{n,CP}$ respectively, $\boldsymbol{P}_{n|n\text{-}1}$ is the priori covariance matrix of estimation error , and $\boldsymbol{P}_n$ is the current covariance matrix of estimation error. When the channel taps are modelled as a zero mean random process, the algorithm is initialized with an all-zero state vector. Besides this, the assumption of uncorrelated scattering (US) causes the different channel taps to be i.i.d., and the error covariance matrix is initialized as an identity matrix.

$$\mathbf{s}_0 = \hat{\mathbf{s}}_0 = \mathbf{0}_{2L\times 1}$$

$$\boldsymbol{P}_0 = E\left[(\mathbf{s}_0 - \hat{\mathbf{s}}_0)(\mathbf{s}_0 - \hat{\mathbf{s}}_0)^H\right] = \mathbf{I}_n$$

The receiver operates in training and decision directed modes. In training mode the known transmitted CP ($\mathbf{x}_{n,CP}$) and CP part of the received OFDM symbol ($\mathbf{y}_{n,CP}$) form the input to the above Kalman filter algorithm, and get the channel estimation $H_n(k)$, we get

$$\hat{X}_n(k) = \frac{Y_n(k)}{\hat{H}_n(k)} \tag{30}$$

In decision directed mode the receiver uses the estimated channel vector from the previous OFDM symbol to demodulate the received symbol and generate an estimate of transmitted CP ($\hat{X}_{n,CP}(k)$). Here the transmitted CP part can be estimated by previous estimated channel i.e.,

$$\hat{X}_{n,CP}(k) = \frac{Y_{n,CP}(k)}{\hat{H}_{n-1}(k)} \tag{31}$$

This estimated CP and CP of the received OFDM symbol ($\mathbf{y}_{n,CP}$) helps to estimate the channel.

The equations from (23) to (29) can be carried out by providing the AR parameters that are involved in the transition matrix **B** and the driving process variances are available. In case, these are unknown, for estimating these parameters Dual-Kalman filtering technique is used.

3.2. Dual-Kalman Filtering Algorithm

To estimate the AR parameters $\boldsymbol{\theta}_n$ from the estimated fading process $\hat{\mathbf{h}}_n$, Equation (28) is firstly represented as an AR-2 model to express the estimated fading process as a function of $\boldsymbol{\theta}_n$ (AR parameter vector).

$$\hat{\mathbf{h}}_n = \left[\hat{\mathbf{h}}_{n-1}\ \ \hat{\mathbf{h}}_{n-2}\right]\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \mathbf{w}_n \tag{32}$$

$$\begin{bmatrix} \hat{h}_n(0) \\ \hat{h}_n(1) \\ \vdots \\ \hat{h}_n(L-1) \end{bmatrix}_{L\times 1} = \begin{bmatrix} \hat{h}_{n-1}(0) & \hat{h}_{n-2}(0) \\ \hat{h}_{n-1}(1) & \hat{h}_{n-2}(1) \\ \vdots & \vdots \\ \hat{h}_{n-1}(L-1) & \hat{h}_{n-2}(L-1) \end{bmatrix}_{L\times 2} \begin{bmatrix} a_1 \\ \\ a_2 \end{bmatrix}_{2\times 1} + \mathbf{w}_n$$

$$\mathbf{r}_n = \mathbf{H}\boldsymbol{\theta}_n + \mathbf{w}_n \tag{33}$$

where $\mathbf{r}_n$ is the estimated channel vector, $\boldsymbol{\theta}_n$ is the AR parameter vector defines as $\boldsymbol{\theta}_n = [a_1\ \ a_2]^T$.

and $\mathbf{w}_n$ is the L×1 noise vector as in Equation (19).

When the channel is assumed to be stationary, the AR parameters are time-invariant and satisfy the following relationship

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} \tag{34}$$

As Equation (33) and Equation (34) define a state-space representation for the estimation of the AR parameters, a second Kalman filter can be used to recursively estimate $\boldsymbol{\theta}_n$ as follows [3]

$$[\boldsymbol{P}_{\theta_{n/n-1}}]_{2\times 2} = \boldsymbol{P}_{\theta_{n-1/n-1}} \tag{35}$$

$$[\boldsymbol{\alpha}_{\theta_n}]_{L\times 1} = [\mathbf{r}_n - \mathbf{H}\hat{\boldsymbol{\theta}}_{n-1}] \tag{36}$$

$$[\mathbf{C}_{\theta_n}]_{L\times L} = \mathbf{H}\boldsymbol{P}_{\theta_{n/n-1}}\mathbf{H}^H + \mathbf{Q}_3 \tag{37}$$

$$[\mathbf{K}_{\theta_n}]_{2\times L} = \boldsymbol{P}_{\theta_{n/n-1}}\mathbf{H}^H\mathbf{C}_{\theta_n}^{-1} \tag{38}$$

$$[\hat{\boldsymbol{\theta}}_n]_{2\times 1} = \hat{\boldsymbol{\theta}}_{n-1} + \mathbf{K}_{\theta_n}\boldsymbol{\alpha}_{\theta_n} \tag{39}$$

$$[\boldsymbol{P}_{\theta_{n/n}}]_{2\times 2} = [\mathbf{I}_2 - \mathbf{K}_{\theta_n}\mathbf{H}]\boldsymbol{P}_{\theta_{n/n-1}} \tag{40}$$

where $\mathbf{Q}_3$ is the covariance matrix of the $\mathbf{w}_n$, the error covariance matrix and the initial AR parameter vector are defined as

$$\hat{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0 = \mathbf{0}_{2\times 1}$$

$$\boldsymbol{P}_{\theta_{0/0}} = E\left[(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_0)^H\right] = \mathbf{I}_2$$

3.3. Noise parameters estimation

Apart from estimating the AR parameters, we also need to estimate the noise parameters for the fading channel

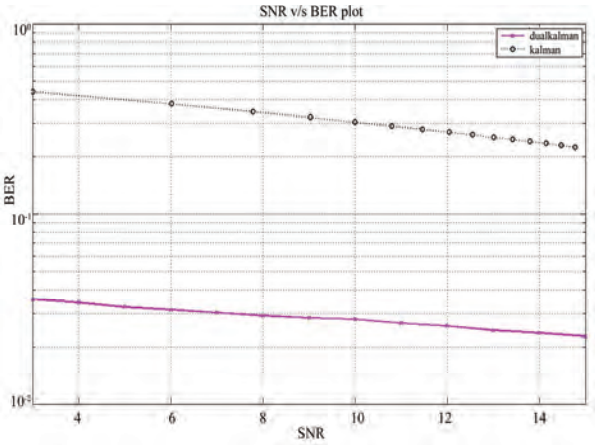


Figure 2. BER v/s SNR for Kalman and Dual Kalman.

environment i.e., variance of $\mathbf{v}_n$ and $\mathbf{Z}_{n,CP}$. This can be done by using the error covariance matrices. From Equation (23) and Equation (29) we can write the noise variances recursively as

$$[\boldsymbol{L}_n]_{2L\times 2L} = \boldsymbol{P}_n - \mathbf{B}\boldsymbol{P}_{n-1|n-1}\mathbf{B}^H + \mathbf{K}_n\boldsymbol{\alpha}_n\boldsymbol{\alpha}_n^H\mathbf{K}_n^H \quad (41)$$

$$\hat{Q}_1(n) = \frac{n-1}{n}\hat{Q}_1(n-1) + \frac{1}{n}\boldsymbol{D}\boldsymbol{L}_n\boldsymbol{D}^T \quad (42)$$

$$\boldsymbol{D} = \left[1\,0\,.....\,0\right]_{1\times 2L}$$

$$[\mathbf{M}_n]_{gi\times gi} = \boldsymbol{\alpha}_n\boldsymbol{\alpha}_n^H - \overline{\mathbf{A}}_n\boldsymbol{P}_{n|n-1}\overline{\mathbf{A}}_n^H \quad (43)$$

$$\hat{Q}_2(n) = \frac{n-1}{n}\hat{Q}_2(n-1) + \frac{1}{n}\boldsymbol{D}_1\mathbf{M}_n\boldsymbol{D}_1^T \quad (44)$$

$$\boldsymbol{D}_1 = \left[1\,0\,.....\,0\right]_{1\times gi},$$

where $\hat{Q}_1(n)$ and $\hat{Q}_2(n)$ are the estimated variances of the process noise $\mathbf{v}_n$ and modelling noise $\mathbf{Z}_{n,CP}$ respectively.

4. Results and Conclusions

In this analysis we compare CP based dual Kalman and Kalman. At SNR of 10db, BER of Kalman is $10^{-0.8}$ where as for dual Kalman; this value is $10^{-1.81}$. For Overall SNR is concerned, dual Kalman gives better performance and doesn't require any additional training sequence like training bits and noise. (Figure 2)

5. References

[1] R. Steele, "Mobile radio communications," New York: IEEE Press, 1992.

[2] X. W. Wang and K. J. R. Liu, "Adaptive channel estimation using cyclic prefix in multicarrier system," IEEE Communication Letters Magazines, Vol. 3, pp. 291–293, October 1999.

[3] A. Jamoos, D. Labarre, E. Grivel, and Najim, "Two cross coupled Kalman filters for joint estimation of MC-DS-CDMA fading channels and their corresponding autoregressive parameters," Proceedings of EUSIPCO, Antalya, Turkey, September 4–8, 2005.

[4] M. K. Tsatsanis, G. B. Giannakis, and G. Zhou, "Estimation and equalization of fading channels with random coefficients," Signal Process, Vol. 53, pp. 211–229, 1996.

[5] A. Jamoos, J. Grolleau, and E. Grivel, "Kalman vs H_∞ algorithms for MC-DS-CDMA channel estimation with or without a priori AR modelling," IEEE Multicarrier Spread Spectrum, Springer Verlag, pp. 427–436, 2007.

[6] I. R. Petersen and A. V. Savkin, "Robust Kalman filtering for signals and systems with large uncertainties," Boston, MA: Birkhäuser, 1999.

[7] Y. Li, L. J. Cimini, and N. R. Sollenberger, "Robust channel estimation for OFDM systems with rapid dispersive fading channels," IEEE Transactions on Communications, Vol. 46, pp. 902–915, July 1998.

Int. J. Communications, Network and System Sciences, 2009, 2, 857-873
doi:10.4236/ijcns.2009.29100 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Forensic Investigation in Communication Networks Using Incomplete Digital Evidences

Slim REKHIS, Jihene KRICHENE, Noureddine BOUDRIGA
CN&S Research Lab, University of the 7th of November, Carthage, Tunisia
E-mail: *slim.rekhis@isetcom.rnu.tn, jkrichene@gmail.com, nab@supcom.rnu.tn*
Received July 16, 2009; *revised September* 10, 2009; *accepted October* 21, 2009

Abstract

Security incidents targeting information systems have become more complex and sophisticated, and intruders might evade responsibility due to the lack of evidence to convict them. In this paper, we develop a system for Digital Forensic in Networking, called DigForNet, which is useful to analyze security incidents and explain the steps taken by the attackers. DigForNet combines intrusion response team knowledge with formal tools to identify the attack scenarios that have occurred and show how the system behaves for every step in the scenario. The attack scenarios construction is automated and the hypothetical concept is introduced within DigForNet to alleviate missing data related to evidences or investigator knowledge. DigForNet system supports the investigation of attack scenarios that integrate anti-investigation attacks. To exemplify the proposal, a case study is proposed.

Keywords: Formal Digital Investigation, Incident Response Probabilistic Cognitive Map, DigForNet, Anti-Forensic Attacks Investigation, Attack Scenarios Reconstruction

1. Introduction

Considering the state of digital security incidents which has dramatically increased in terms of complexity, number, and sophistication, it becomes evident that the traditional ways of protecting information systems (e.g., Firewalls, IDSs) are no longer sufficient. Faced to this situation, security experts have started giving a great interest to a novel discipline called the digital investigation of security incidents, which is defined by the literature as preservation, identification, extraction, documentation and interpretation of computer data [1]. Digital investigation aims to perform a post-incident examination on the compromised system while achieving several objectives including evidence collection, attack scenarios construction, and results argumentation with non refutable proofs.

Performing a digital investigation is a very complex task for many reasons. First, attacks may use multiple sources and become difficult to trace using the available trace-back techniques. Second, systems may not be initially prepared for investigation, leading to the absence of effective logs or alerts to be used during the analysis of the incident. Third, the attackers may use a number of techniques to obfuscate their location or to hide traces on the system that could be used to prove their malice. Fourth, attack scenarios may use several automated tools that create intensive damaging activities on the compromised systems. A large amount of data should thus be analyzed and several evidences need to be identified and extracted.

To face the above complexity, the digital investigation should, first, be well structured by reconciling both the expertise of the incident response team (IRT) and the use of formal reasoning techniques about security incidents. This reconciliation allows to: 1) better filter the data to be analyzed and source of evidences to be explored, based on the skills developed by the IRT; 2) validate the results of the formal techniques, by the IRT, before presenting them and also use them to improve the content of the attacks library. Second, digital investigation should integrate the use of formal techniques that are useful to develop non-refutable results and proofs, and avoid errors that could be introduced by manual interpretations. Moreover, it should consider the development of tools to automate the proof provided by these formal methods. Third, since the collected evidences may be incomplete and describing all potential malicious events in advance is impractical, hypotheses need to be put forward in order to fill in this gap.

Despite the usefulness of formal methods and approaches, digital investigation of security incidents remains scarcely explored by these methods. Stephenson [2] took interest to the root cause analysis of digital incidents and used

Colored Petri Nets as formalism for modeling occurred events. The methodology may become insufficient to deal with sophisticated attack scenarios, where there is a lack of information on the compromised system that requires some hypotheses formulation. Stallard and Levitt [3] proposed a formal analysis methodology entitled semantic integrity checking analysis. It is based on the use of an expert system with a decision tree that exploits invariants relationship between existing data redundancies within the investigated system. To be usable with highly complex systems, it is imperative to have a prior list of *good state* information; otherwise, the investigator has to complete its analysis in ad hoc manner. Gladychev [4,5] provided a Finite State Machine (FSM) approach to reconstruct potential attack scenarios discarding scenarios that disagree with the available evidences. Since investigation may occur on systems that could not be completely described due to their complexity, if unknown system transitions are ignored, the event construction may freeze or its accuracy may be severely affected. Carrier and Spafford [6] proposed a model that supports existing investigation frameworks. It uses a computation model based on a FSM and the history of a computer. A digital investigation is considered as the process that formulates and tests hypotheses about occurred events or states of digital data. Additionally, the model allows defining different categories and classes of analysis techniques. A key idea in the proposed approach is that every computer has a history, which is not fully recorded or known in the practice. A digital investigation is considered as the process that formulates and tests hypotheses about occurred events or states of digital data. Willanssen [7] took interest in enhancing the accuracy of timestamp evidences. The aim is to alleviate problems related to the use of evidences whose timestamps were modified or referred to an erroneous clock (i.e., which was subject to manipulation or maladjustment). The proposed approach consists in formulating hypotheses about clock adjustment and verifying them by testing consistency with observed evidences. Later, the testing of hypotheses consistency is enhanced by constructing a model of actions affecting timestamps in the investigation system [8]. In [9], a model checking-based approach for the analysis of log files is proposed. The aim is to search for pattern of events expressed in formal language. Using this approach logs are modeled as a tree whose edges represent extracted events in the form of algebraic terms. P. Sencar and Memon [10] proposed a methodology to recover files from unallocated space of disk without the assistance of meta-data or file system table. The proposed technique assumes that files may be initially fragmented and several contiguous blocks may be scattered around the storage area. To enhance the effectiveness of file recovery, the technique looks for detecting the point of fragmentation of a file, using a sequential hypothesis testing (SHT) procedure. Peisert [11] proposed to determine what data are necessary to perform investigation and basis its idea on the use of the requires/provides model, which is commonly used for intrusion detection.

We develop in this paper, a system for Digital Forensic in Networking called DigForNet. It integrates the analysis performed by the Incident Response Team on a compromised system, through the use of a new Cognitive Map [12,13] called the Incident Response Probabilistic Cognitive Map (IRPCM), which extended the Cognitive Map proposed in [14]. DigForNet uses formal approach to identify potential attack scenarios using a formal specification language entitled, I-TLA. The formal approach allows specifying different forms of evidences. It identifies an attack scenario as a series of elementary actions, retrieved from a used library, which, if executed on the investigated system, would produce the set of available evidences. We developed in DigForNet the concept of executable specification of attack scenarios, which shows with details how an attack is performed progressively on the system and how the latter behaved during the attack. DigForNet uses I-TLC, an automated verification tool for I-TLA specifications. To alleviate any missing evidences or details related to attack scenarios, DigForNet integrates a technique for generating hypothetical actions to be appended to the scenario under construction.

Our contribution is three-fold. First, DigForNet reconciles in the same framework conclusions derived by the incident response team and theoretical and empirical knowledge of digital investigators. To the best of our knowledge, it is the first investigation system which supports such feature. Second, we proposed a new IRPCM which integrates the temporal aspect. In fact, during IRPCM construction, the appending of anti-investigation relations between concepts could make other concepts inactive. Several IRPCM snapshots could thus be obtained depending of time. Third, using the concept of hypothetical actions, DigForNet stands out from the other existing approaches and allows generating sophisticated and unknown attack scenarios. The new generated hypothetical actions could be used to extend the content of the library of attacks. Fourth, the formal techniques used by DigForNet allow supporting a collaborative working between the IRT members, and generating a formal specification useful for conducting an investigation, where a model checking-like technique could be used to automate the generation of executable specification of attack scenarios.

This paper is organized as follows. Section 2 defines the important concepts related to the digital investigation of security incidents and describes the DigForNet's methodology for reasoning about security incidents. The use of the IRPCM technique to represent the intrusion response team members' view about the security incidents is shown in Section 3. Section 4 describes I-TLA as logic for specifying evidences and identifying potential attack scenarios that satisfy them. It also shows how to pass from IRPCM to I-TLA specification. Section 5 introduces I-TLC as an automated verification tool for I-TLA specifications, which allows generating executable specification of attack scenarios. Section 6 illustrates an example of the use of DigForNet in investigating a real security incident. Finally, Section 7 concludes this paper.

2. Methodology of Structured Investigation

We start this section by introducing the need for digital investigation and then we describe the DigForNet's methodology.

2.1. Need for Digital Investigation

Focusing merely on restoring the system, which is the simplest and easiest method, is disadvantageous. In fact, valuable information and traces that allow understanding the attack could be removed if the compromised system is straightforwardly formatted or reinstalled. The above mentioned weaknesses in the response point up the need for conducting a post-incident digital investigation [15]. The latter can be considered as the process that allows to: 1) determine how the computer attack was performed and what are the security weaknesses and design mistakes that let the incident succeeds; 2) trace the attackers to their source to identify their identities; 3) build a proof from the collected information to bring a prosecution against attackers who committed the attack; 4) argument and underline the results with well-tested and proved methods and techniques; 5) Study the attackers' trends and motives, and take the accurate security measures to prevent future similar attack scenarios.

Since digital investigation [16,17] focuses on the investigation of an incident after it has happened, a digital evidence should be gathered from the system to support or deny some reasoning an investigator may have about the incident. Digital evidence is defined as any data stored or transmitted that support or refute a theory of how an offence occurred or that address critical elements of the alibi [16].

2.2. DigForNet Methodology

DigForNet integrates the incident response team contributions under the form of Incident Response Probabilistic Cognitive Maps (IRPCMs). An IRPCM is built with a collaborative fashion by the IRT members based on the information collected on the system. IRPCMs provide a foundation to mainly investigate and explain occurred security attacks.

DigForNet provides a formal way for reconstructing potential attack scenarios. It defines a novel logic entitled Investigation-based Temporal Logic of Actions (I-TLA), and its logic-based language entitled I-TLA$^+$. DigForNet methodology is composed of six steps organized in a waterfall model as shown in Figure 1. They are described as follows.

The first step collects evidences available within three different sources, namely the operating systems, networks, and storage systems. The second builds the IRPCM, which is nothing but a directed graph representing security events, actions and their results along with the relations between these concepts. The third step consists in extracting the sets of evidences and actions from the cognitive map for the formal specification of the potential attack scenarios. The fourth step generates a formal specification. A formal approach is necessary for this purpose. DigForNet uses logic, referred to as I-TLA, to generate a specification containing a formal description of the set of extracted evidences and actions, the set of elementary attack scenario fragments retrieved from the library of attacks, and the initial system state. During this step, DigForNet uses I-TLA to prove the existence of potential attack scenarios that satisfy the available evidences. To be able to generate a variety of attack scenarios, DigForNet considers the use of a library of elementary actions supporting two types of actions: legitimate and malicious. Malicious actions are specified by security experts after having assessed the system or appended by investigators upon the discovery of new types of attacks.

The fifth step generates executable specification [18,19] of potential attack scenarios using a model checker tool associated with the formal specification. DigForNet builds Investigation-based Temporal Logic model Checker called I-TLC composed of two sub-steps. The first works to rebuild the attack scenarios in forward and backward chaining processing, showing details of all intermediate system states through where the system progresses during the attack. The second sub-step provides a tolerance to the incompleteness of details regarding the investigated incident and the investigator knowledge. It interacts with a library of hypothetical atomic actions to generate hypothetical actions, append them to the scenarios under construction, and efficiently manage them during the whole process of generation. The library of hypothetical atomic actions is composed of a set of entries showing interaction between a set of virtual system components and a set of rules used to efficiently create hypothetical actions as a series of hypothetical atomic actions.

The sixth step uses the generated executable potential attack scenarios to identify the risk scenario(s) that may have compromised the system, the entities that have originated these attacks, the different steps they took to conduct the attacks, and the investigation proof that confirms the conclusion. These results are discussed with the IRT members in order to check the hypotheses added by I-TLC and update the initial IRPCM by: 1) omitting some concepts because they do not present an interest for the attack scenario construction, and/or 2) adding other concepts, corresponding to the hypothetical actions, to the IRPCM and linking them to the other concepts. Links in

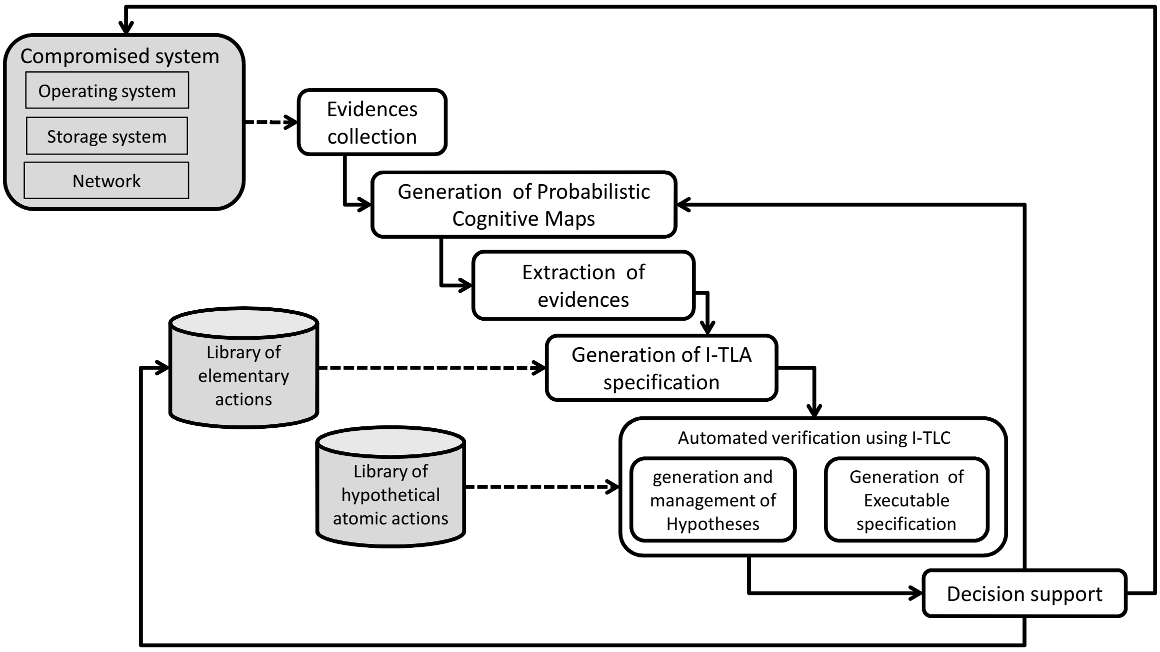

Compromised system
Operating system
Storage system
Network
Evidences collection
Generation of Probabilistic Cognitive Maps
Extraction of evidences
Library of elementary actions
Generation of I-TLA specification
Library of hypothetical atomic actions
Automated verification using I-TLC
generation and management of Hypotheses
Generation of Executable specification
Decision support

Figure 1. DigForNet methodology.

the IRPCM are deleted in the case where the concepts originating from or going to them are omitted. New links are also added to the IRPCM to link the newly added concepts. Hypothetical actions are also added to the attack library. In addition, tools collecting the evidences are enhanced to be able to detect the newly discovered vulnerabilities.

3. Intrusion Response Probabilistic Causal Maps

We have defined in [20] a category of cognitive maps to support the intrusion response activity. In this paper, we provide an extension to these cognitive maps referred to as Incident Response Probabilistic Cognitive Maps (IRPCMs) by introducing the notions of probability and activation degree of concepts, and integrating links with complex semantics. IRPCMs provide the foundation to investigate and explain security attacks which have occurred in the past and to predict future security attacks against the system. These aspects are important for negotiation or mediation between IRT members solving thus disparities which are generated by the difference in their view points and which can lead to conflicting decisions.

3.1. IRPCM Definition

An Incident Response Probabilistic Cognitive Map (IRPCM) is a directed graph that represents intrusion response team members' experience-based view about security events related to an incident. The nodes of the graph represent concepts belonging to the network security field and a set of edges representing relationships between the concepts.

IRPCM concepts can be symptoms, actions, and unauthorized results related to network security field. Symptoms are signs that may indicate the occurrence of an action. System crashes or the existence of new user accounts or files are examples of symptoms. An action is a step taken by a user or a process in order to achieve a result. Probes, scans and floods are samples of actions. An unauthorized result is an unauthorized consequence of an event (defined by an action directed to a target). For instance, an authorized result can be an increased access or a disclosure of information or a theft of resources. IRPCM concepts are labeled by values in the interval [0,1] informing about the activation of the correspondent concepts. They are also labeled by a value indicating their occurrence time.

IRPCM edges link concepts to each others. An edge e_{ij} linking concept c_i to concept c_j is labeled as (π_{ij} , q_{ij}) where π_{ij} is the predicate expressing the relation between the two nodes (examples include $<_t$, I/O , CE) and q_{ij} (taking values in]0,1]) the probability expressing the certitude degree that the relationship π_{ij} really occurred between the concepts c_i and c_j. Quantitative values are given by security experts. Notice that the semantic of the predicate π_{ij} depends on the nature of the concepts c_i and c_j. For the sake of simplicity, we consider seven cases of relationships in this paper. They are described here after.

1) Input/output relation: Let c_i be a symptom and c_j be a symptom or an action. An input/output relation, which

is expressed using the predicate $\pi_{ij}=I/Q$, means that part of output of the concept c_i is the input of the concept c_j.

2) Temporal relation: Let c_i and c_j be two actions. A temporal relation, which is expressed using the predicate $\pi_{ij}=<t$, means that c_i is an action that precedes c_j.

3) Cause/Effect relation: Let c_i be an action and c_j be an unauthorized result. The cause/effect relation, which is expressed using the predicate $\pi_{ij}=CE$, means that the effect produced by concept c_i is visible through concept c_j.

4) Concealment relation: Let c_i and be an action and c_j be an action, a symptom or an unauthorized result. The concealment relation, expressed using the predicate $\pi_{ij}=$ *conceal*, means that concept c_i when it happens, leads to the hiding of concept c_j. This corresponds to the situations where the attackers execute some actions on the compromised system to hide information revealing their access to this system, or to hide the results on this system.

5) Destruction relation: Let c_i be an action and c_j be an action, a symptom or an unauthorized result. The destruction relation, which is expressed using the predicate π_{ij}=*destroy*, means that the occurrence of concept c_i wipes out the existence of concept c_j. This corresponds to the situation where the attacker executes some actions to destroy any trace that may inform about his access to the compromised system.

6) Forgery relation: Let c_i be an action and c_j be an action, a symptom or an unauthorized result. The forgery relation, which is expressed using the predicate π_{ij}=*forge*, means that the occurrence of concept c_i creates a new forged concept c_j with random time of occurrence. This corresponds to the situation where the attacker tries to deceive the investigation activity.

7) Replacement relation: Let c_i be a forged action, symptom, or unauthorized result and c_j be a concept belonging to the same category as c_i. The replacement relation expressed using the predicate π_{ij}=replace, means that the concept c_j is replaced by concept c_i when the latter has been forged by an action.

Notice that relations *conceal*, *destroy*, *forge* and *replace* corresponds to an anti-investigation activity.

3.2. Appearance-Period

Let c_i and c_j be two concepts belonging to the IRPCM having respectively occurrence times equal to t_i and t_j. Appearance period of c_j, say A_{cj} is determined as follows:

- If $\pi_{ij} = conceal, destroy, replace$ then $A_{c_j} = [t_j, t_i]$
- If $\pi_{ij} = I/O, <_t, C/E, forge$ then $A_{c_j} = [t_j, \infty]$

3.2.1. Snapshot Function

An IRPCM may vary as anti-forensic actions and relations are appended. Therefore, some concepts in the IRPCM may be invalid at a given time, and the analysis of the IRPCM becomes complex. To make the analysis simple, we need a snapshot of the IRPCM for different instants. To this end, we introduce the snapshot function. The main feature of this function is to show a sub-view of the IRPCM which hides the concepts in the IRPCM that are invisible at that time. To do so, the appearance period of concepts is exploited.

Formally, let *Vsisible*(c,t) be the function that returns the Boolean value *True* if the time t is within the appearance period of the concept c. The IRPCM snapshot at time instant t is created by deleting any concepts c in that IRPCM, such that *Vsisible*(c,t)=*false*, and all edges which are connected to c_i.

3.3. Building IRPCMs

The IRT members are responsible for building the IRPCM (second step in the DigForNet methodology). The basic elements needed in this activity are the events collected on the information system. These events may be IDS alerts; compromises of services offered by the network, or any sign indicating the occurrence of malicious or suspect actions against the network. IRT members analyze these signs and define the appropriate symptoms, actions and unauthorized results and assign the appropriate probabilities and relationships to the edges linking the defined concepts. The process of building an IRPCM has two properties: completeness (if an attack has occurred and a sufficient number of events are collected to identify this attack, then we can find an IRT able to build an IRPCM allowing to identify the attack) and convergence (if an IRPCM is built and is large enough to collect all the events related to a given attack, then the IRT must build in a finite time an IRPCM allowing to provide the right solution to protect against this attack).

The building of an IRPCM follows a methodology based on the iterative process described in the following steps:

1) Collect a first set of security events observed in the compromised system or detected by security tools.

2) Build an IRPCM based on the collected events.

3) Continue to collect security events.

4) Update the IRPCM based on the new recollected events. Events which do not belong to the previous IRPCM are added. Probable links related to the newly considered concepts are also added to the IRPCM.

5) Refine the IRPCM.

6) Update the probabilities of the links and the activation degree of the concepts.

7) If the stopping criterion is satisfied, stop the IRPCM building process; else, return to step 4.

In the second step of the above methodology, the generation and building of the IRPCM is the duty of the Incident Response Team (IRT). Two main tasks should be handled within this step. First, the IRT members should collaborate to append concepts based on their knowledge and skills, and negotiate between them to classify the appended concepts into necessary and unnecessary concepts. Concepts can be considered as unnecessary if they are duplicated, do not cope with the properties of the attack or the system under investigation, or are erroneous. The unnecessary concepts will thus be deleted from the IRPCM under construction. Second, the IRT members collaborate to locate concepts in the IRPCM that could be linked together and append edges. Obviously, such activity is subject to discussion and negotiation in order to correct or delete erroneous edges.

In the fifth step of the methodology, the refinement of the IRPCM is done through the analysis of the semantic of concepts. Two forms of modifications on the IRPCM take place during the refinement. The first consists in substituting a concept by a more accurate one, merging some concepts together, or segmenting a concept into many other ones to make relations more significant. While attack scenarios look different, they, in most cases, reuse techniques of attacks and actions. The IRT, which is always in charge of constructing IRPCMs for the investigated scenarios, could exploit IRPCMs related to previously resolved incidents to complete and update the current IRPCM under construction. To do so, it suffices to define patterns that allow detect similarities between similar fragments of IRPCMs. The IRT has just to detect patterns in the IRPCM under construction and find IRPCM fragments that could be integrated from the previously constructed ones.

Two criteria can be considered to decide about the end of the IRPCM building process. The first is when all the candidate actions in the library (those which have a relationship with the collected events) are present in the IRPCM. The second is based on the decision of the IRT members. If the latter agree that the IRPCM is large enough, then the building process is stopped. The IRT decision can be shared by all the members or it can be taken by a mediator (a member of the IRT in charge of coordinating activity helping solving conflicts and terminating the process).

3.4. Activation Degree of a Concept

IRPCM concepts values give indications about their activation. These values, referred to as activation degrees, belong to the interval [0,1]. We define the function *dac* assigning activation degrees to the concepts as follows:

$$\begin{aligned} dac:\ & C \times S \rightarrow [0,1] \\ & c,s \mapsto dac(c,s) \end{aligned} \qquad (1)$$

where *C* represents the set of concepts in the IRPCM and *S* stands for the set of snapshots. A concept is said to be dac-activated if its activation degree is equal to 1. In the following, we show how to build a *dac* function based on a given set of selected concepts in the IRPCM for a snapshot *s*. Let *I* be the set of concepts related to collected events of involvement in attack with respect to detected intrusions. $I = \{c_1 \cdots c_n\} \subseteq C$.

1) Let $dac(c_i, s) = 1, i = 1 \cdots n$.

2) Compute iteratively the remaining activation degrees as follows: Let *F* be the set of the concepts for which we have already computed the activation degree. *F* is initially equal to the set *I*.

3) Let *G* be the set of concepts that have a relation with one or more concepts belonging to *F*. $G = \{c \in C / \exists d \in F, (d,c) is an edge\}$ .Then, $dac(c,s) =$ $sup_{d \in G}\{q_{dc} dac(d\ \ )\}$ where q_{dc} is the probability expressing the certitude degree that there is a relationship between the concepts *d* and *c*.

4) $F := F \cup G$ and return to step 3 if $F \neq \varnothing$.

In the case where the IRT members have detected malicious actions against the secured system, they start constructing the IRPCM corresponding to this situation. The concepts that represent the collected events are activated and will form the set *I* in this case. The activation degree of the remaining concepts is determined according to the previous algorithm.

The *dac* function is used in the third step of the DigForNet methodology to extract set of evidences. Nodes having a *dac* degree greater than a predefined threshold are extracted as evidences for the formal specification.

Notice that the activation degree of concepts may vary from one snapshot to another if some concepts are deleted or, in the contrary, added to the current snapshot. In the first case, a concept c_m is deleted from a given snapshot of the IRPCM. If c_f is a concept to which c_m is directly linked, then the activation degree of the concept c_f is reduced if the activation degree of c_m is the most important over the set of concepts directly linked to c_f . In the second case, if c_m is a concept which is added to the current snapshot, we distinguish three sub-cases: 1) the new concept is not evidence and has no concepts directly linked to it. In this case the *dac* value of the new concept is unknown and must be set by the investigation team; 2) the new concept is evidence. In this case its *dac* value is set to one; 3) the new concept is not evidence and there are concepts directly linked to it. In this case, the *dac* value of the new concept is calculated according to the previous algorithm. Having determined the *dac* value of the new concept c_m and if we represent by *P* the set of concepts to which c_m is directly linked, then the for every concept *c* in *P*, the activation degree increases if c_m has the highest activation degree over those directly linked to *c*.

During IRPCM construction, it may happen that in some snapshots, some concepts constitute evidences, while they did not in the preceding IRPCM snapshot. In this case, we set to one the activation degree of these concepts and, using the previous algorithm, we update the *dac* values of the concepts to which these evidences are directly or indirectly linked. Conversely, if some anti-forensic relations appear in the new snapshot showing that some data sources were affected by attacks to alter the stored evidences, the dac values of concepts related to evidences collected from these sources should be reduced. Consequently, the algorithm is re-executed to update the dac values in the new IRPCM snapshot.

4. Generation of a Formal Specification of Attack Scenarios

The Investigation-based Temporal Logic of Actions, I-TLA [21], is a logic for the investigation of security incidents. It is a extension to the Temporal Logic of Actions (TLA [22]). I-TLA defines a theoretical framework for: 1) modeling and specifying evidences left by intruders further to the occurrence of a security incident; 2) supporting advanced description and specification of potential happened attack scenarios as a series of elementary attacks, extracted from the library of attacks, that if assembled together, would satisfy the available evidences. Similarly to TLA, I-TLA allows to reason about systems and their properties in a uniform logic formalism. I-TLA is provided with I-TLA$^+$, a highly expressive formal language that defines a precise syntax and module system for writing I-TLA specifications. I-TLA will be used in this paper to generate a specification describing potential attack scenarios that satisfy the available set of evidences.

In the sequel, we focus on describing the different forms of evidences supported by I-TLA, showing how they can be specified and how they should be satisfied by the expected attack scenario. The reader is referred to [21] for more details of I-TLA and I-TLA$^+$ and a complete semantic and syntactic description.

4.1. Modeling Scenarios and Evidences in I-TLA

I-TLA is typeless and state-based logic. It allows the description of states and state transitions. A state, while it does not explicitly appear in a I-TLA specification formula, is a mapping from the set of all variables names to the collection of all possible values. An I-TLA specification ϕ generates a potential attack scenario in the form of: $\omega = \langle s_0, s_1, ..., s_n \rangle$, as a series of system states s_i (i = 0 to n) that satisfies all available evidences. I-TLA supports four different forms of evidences, namely history-based, non-timed event-based, timed event-based, and predicate-based evidences. A state-based representation of attack scenarios allows a security expert to observe how its system progresses during the attack and how it interacts with the actions executed in the scenario.

4.1.1. History-Based Evidences

Typically, security solutions do not have direct access to all system components. Some of them are able to provide evidences as histories of the values of the monitored system variables, during the spread of an attack scenario. These security solutions cannot realize that the system has progressed or not from a state to another if the value of the monitored component is either blind, or does not change. I-TLA encodes a history-based evidence, say E, as an observation over a potential attack scenario ω, generated by $Obs(\omega)$ ($Obs()$ is the observation function that characterizes the ability of a security solution to monitor the history of the system during an attack scenario). It uses a labeling function that allows a third party to only see limited information about states of an execution. Since a state is a valuation of all system variables, a labeling function allows to either:

- Totally observe the content of variable value. Variable v is visible and its value is interpretable by the observer. It represents a system component whose modification is monitored by some security solution.
- Observe a fictive value instead of the real variable value. Variable v is visible but not interpretable by the observer, meaning that its variation does not bring any supplementary information to an observer. It can represent an encrypted data whose decryption key is unknown by the observer.
- Observe empty value. Variable v is completely invisible, such that none information regarding its value could be determined. It represents a system component which is not monitored by any security solution.

$Obs(\omega)$ is obtained by following two steps:

1) Transform each state s_i to $\hat{s}_i$, by hiding some of the details it provides. $\hat{s}_i$ is obtained from s_i by making the value of every system variable v in s_i to be:

a) Unmodified. In this case the variable is visible and its value is interpretable by the observer. It represents a system component whose modification is monitored by some security solution;

b) Equal to a fictive value fictive value. In this case the variable is visible but not interpretable by the observer, meaning that its variation does not bring any supplementary information to an observer. It can represent an encrypted data whose decryption key is unknown by the observer;

c) Equal to an empty value, denoted by ε. In this case, the variable is completely invisible, such that none information regarding its value could be determined. It represents a system component which is not monitored any security solution.

2) Delete any $\hat{s}_i$ which is equal to null value (i.e., all values are invisible) and then collapse together each maximal sub-sequence $\langle \hat{s}_i, ..., \hat{s}_j \rangle$ such that $\hat{s}_0 = ... = \hat{s}_i$, into a single $\hat{s}_i$.

Taking into consideration the availability of a history-based evidence E, consists in generating, an attack scenario ω such that $Obs(\omega)$=E.

4.1.2. Ordering of Observations

A step in the scenario may not change all the values of the system variables. As the scope of the observations differs, they may not allow noticing that the system has progressed during the attack at the same time. I-TLA allows to specify for two given observations, which one will vary first (respectively last) when the attack scenario starts (respectively finishes). Consider the following example involving an attack scenario ω and two observations $OBS = [e_1,...,e_n]$ and $OBS' = [e_1^{'},...,e_m^{'}]$, generated by observation functions $Obs()$ and $Obs'()$, respectively. OBS is said to be an observation that allows to notice the occurrence of an incident before observation OBS', if and only if: $\exists \omega_x$ such that: $\omega = \omega_x \omega_y$ $\wedge$ $obs_1(\omega_x) = e_1$ $\wedge$ $obs_2(\omega_x) = [e_1^{'},...,e'_j]$ for some j ($1 < j \le m$).

4.1.3. Non-Timed Events-Based Evidences

As the length of observations is different from the length of an attack scenario, reconstructed attack scenarios may differ by the manner in which observations are stretched and aggregated together to generate intermediate states of the execution. I-TLA defines non-timed events based evidences in the form of predicates over I-TLA executions, which specify the modification pattern of variables values through an execution. For instance, the execution predicate *AtSameTime*, states that state predicate p_1 switches to true at the same time the state predicate p_2 switches to false.

$$\begin{gathered} AtSameTime(p_1, p_2) \triangleq \forall \langle s_i, s_{i+1} \rangle \in \omega : \\ (s_i \nvDash p_1 \wedge s_{i+1} \vDash p_1) \Rightarrow s_i \vDash p_2 \wedge s_{i+1} \nvDash p_2) \end{gathered} \qquad (2)$$

Taking into consideration the availability of a non-timed event-based evidence E, is amount to generate an attack scenario ω such that $\omega \vDash E$.

4.1.4. Timed Events-Based Evidences

Starting from a set of available alerts, an investigator can extract some indications related to occurred events. I-TLA defines timed event-based evidence $E = [A_0,...,A_m]$ as a set of ordered actions (A_0 to A_m) that should be part of an expected execution. While the order in which events appear should be respected, there is no need that these events be contiguous. Given a timed event-based evidence $E = [A_0,...,A_m]$, an execution $\omega = \langle s_0,...s_n \rangle$ satisfies evidence E if and only if: $\forall (A_x, A_{x+1}) \in E : \exists (s_i, s_{i+1}) \in \omega$ such that $A_x(s_i, s_{i+1}) =$ *true* and $A_{x+1}(s_j, s_{j+1}) = true$ for some $j \ge i+1$.

4.1.5. Predicate-Based Evidences

With regards to the security response team's members, an unexpected system property is a preliminary argument supporting the incident occurrence (e.g., the integrity of a file was violated). I-TLA defines predicate-based evidence as a predicate, say E, over system states, that characterizes the system compromise. An execution ω satisfies evidence E if E divides ω into two successive execution fragments ω_1 and ω_2 (ω can thus be written as ω=$\omega_1\omega_2$). ω_1 is composed of secure states ($\forall s \in \omega_1$: $s \nvDash pr$), while ω_2 is composed of insecure system states ($\forall s \in \omega_2$: $s \nvDash pr$).

4.1.6. Illustrative example

The following example clarifies the use of I-TLA in digital investigation, and illustrates the mechanism of handling evidences during the construction of potential attack scenarios. We consider a system under investigation which is specified by three variables *x*, *y*, and *z*. The initial system state, described in advance, is the state defined by variables *x*, *y*, and *z* are all equal to *0*. The library of elementary actions contains two actions A_1 and A_2 that can be executed by the system.

$$A_1 \triangleq \begin{aligned} x' &= x \\ y' &= y + 1 \\ z' &= z + 2 \end{aligned} \qquad (3)$$

$$A_2 \triangleq \begin{aligned} x' &= x + 1 \\ y' &= y \\ z' &= z / 2 \end{aligned} \qquad (4)$$

Action A_1, for instance, keeps the value of variable *x* in the new state unchanged with respect to the previous state, and sets the values of *y*, and *z* in the new state 1 and 2 higher than its values in the old state, respectively.

Three different evidences are provided. The two first ones represent history-based evidences, defined as E_1= $\langle 0\varepsilon\varepsilon, 1\varepsilon\varepsilon, 2\varepsilon\varepsilon \rangle$ and E_2= $\langle \varepsilon 0 \varepsilon, \varepsilon 1 \varepsilon, \varepsilon 2 \varepsilon, \varepsilon 3 \varepsilon \rangle$, where ε stands for the invisible value. These evidences are generated by observation functions $obs_1()$ and $obs_2()$, respectively. The first observation function $obs_1()$, allows a security solution to only monitor variable *x*, meaning that, when it is applied to a state s, makes the value of y and *z* both equal to ε, and keeps the values of variable *x* unchanged. The second observation function $obs_2()$ allows a security solution to only monitor variable *y*. The ordering of observations indicates that observation provided by $obs_2()$ allows to notice the occurrence of an incident before the observation provided by $obs_1()$. The third evidence E_3, is provided as a predicate-based evidence defined as $E_3 \triangleq z \ge 1$. The fourth evidence E_4, defined as $E_4 \triangleq$ $\forall \langle s_i, s_{i+1} \rangle \in \omega : (s_i \nvDash p_1 \wedge s_{i+1} \vDash p_1) \Rightarrow s_i \vDash p_2)$, is an non-timed evidence, stating that predicate $p_1 \triangleq x = 1$, false in a state s_i, could not switch to true in the next state

s_{i+1}, unless predicate $p_2 \triangleq z \neq 4$ is true in that state. Finally, evidence E_5, indicates that sequences of events (A_1, A_2) is part of the attack scenario.

Figure 2 shows how I-TLA guarantees the satisfaction of evidences during the construction of the potential attacks. Two potential attack scenarios satisfying the available evidences are provided by I-LA, namely ω_1 and ω_2. The first scenario ω_1 is described as ω_1=<s_1, s_3, s_4, s_7, s_{12}, s_{15}>, and consists in consecutively executing the five following actions $A_1 \rightarrow A_1 \rightarrow A_1 \rightarrow A_2 \rightarrow A_2$. The second scenario ω_2 is described as ω_1=<s_1, s_3, s_5, s_9, s_{11}, s_{18}>, and consists in a consecutive execution of the five actions $A_1 \rightarrow A_2 \rightarrow A_1 \rightarrow A_1 \rightarrow A_2$.

Starting from state s_1, I-TLA cannot execute action A_2 as it moves the system to a state that does not satisfy the ordering of observations. In fact, the sub-scenario <s_0, s_1> is observed by obs_1() as <$0\varepsilon\varepsilon$,$1\varepsilon\varepsilon$> and by obs_2() as <$\varepsilon 0 \varepsilon$>. Thus, the event A_2 is detected by E_1 but not by E_2. Starting from state s_4, I-TLA does not execute action A_2 as it moves the system to a state that violates evidence E_4. State s_8 could not be considered in the construction process as it violates predicate E_3. In fact, the predicate $p1$ has become already true in state s_3 and should not change to false in state s_8 again. I-TLA discards states s_{13}, s_{14}, and s_{19} as each one of them would create an execution that violates evidence E_2 if appended to the scenario under construction. In the same context, state s_{16} is also not added to the scenario under construction as it creates an execution that violates E_1.

4.2. Handling Anti-Investigation Attacks

To elude the process of digital investigation, a seasoned attacker may try to conduct an anti-investigation attack [23, 24] to remove, hide, obfuscate, or alter available evidences after breaking into the system. Available techniques include deletion of relevant log entries, installation of root-kits, steganography, and even wiping of disks [25] to disable any further recovery.

Let *obs*(-) be an available observation function, and *OBS* be a history-based observation which corresponds to the output of *obs*(-) when executed on the attack scenario under progression, say ω. Formally, an anti- investigation attack represents any action which moves the system to some state, say s_j in ω, and does not only append *obs*(s_j) to *OBS*, but also affects *OBS* to modify any content related to *obs*(s_0,…,s_{j-1}).

We remind that I-TLA reconstructs attack scenarios by executing an action unless it satisfies all the available history-based evidences. Let s_i be the current state reachable from the initial state s_0 through the execution <s_0,…,s_{i-1}>. If an I-TLA action A is executed from state s_i to produce state s_{i+1}, the execution obtained after reaching the new state should satisfy the available observation. Formally, $obs(\langle s_0,...,s_{i+1}\rangle) \subseteq OBS$. We demonstrate in the following the impact of the anti-investigation attack on the process of attack scenarios reconstruction in I-TLA using the regular definition of observation functions.

4.2.1. Example

We consider a system modeled using two variables *p* and *l* which are related to the user granted privilege and the content of the system log file, respectively. Variable *p* can take three values: 0, 1, and 2 which stand for no access, unprivileged access, and privileged access, respectively. As the log file is typically accessed in append mode, variable *v* takes a series of values representing the commands executed on the system. These values are included in chronological order of their execution. $\omega = \langle s_0,..,s_5 \rangle$ represents an attack scenario composed of five states, where actions A_1, A_2 and A_3 stand for the execution of arbitrary commands. Every one of these actions appends an entry to the tail of the log file. A_4 consists in exploiting vulnerability on the system to get a privileged access. Action A_5 is an anti-investigation attack. It consists in getting a privileged access on the system and altering the content of the log file by deleting the entry corresponding to the execution of action A_2. The attack scenario ω is described as a series of six states, where every state is a valuation of the two variables, and edges are labeled by the executed action.

$$[1,\langle - \rangle] \xrightarrow{A_1} [1,\langle -,Act_1 \rangle] \xrightarrow{A_2} [1,\langle -,Act_1,Act_2 \rangle]$$

$$\xrightarrow{A_3} [1,\langle -,Act_1,Act_2,Act_3 \rangle] \xrightarrow{A4}$$

$$[2,\langle -,Act_1,Act_2,Act_3 \rangle] \xrightarrow{A_5} [2,\langle -,Act_1,Act_3 \rangle]$$

We consider a security solution which is modeled by the observation function *obs*(). The latter allows observing the current executed commands on the system by looking for new entries appended to the log file. Formally, *obs*(*s*)=*tail*(*s*(*log*)), where *tail*(*x*) returns the last entry in *x*. Using the regular definition of observation function *obs*(), the history-based evidence generated by the security solution further to the execution of the attack scenario is given by:

$$\begin{aligned} & obs(\omega) = \\ & \langle obs(s_1),obs(s_2),obs(s_3),obs(s_4),obs(s_5),obs(s_6) \rangle \quad (5) \\ & = \langle -,Act_1,Act_2,Act_3 \rangle \end{aligned}$$

Starting from this definition, it is expected that the provided observation content will be in the form of <-, Act_1, Act_2, Act_3,>. However, since this history-based evidence is provided by the content of the log file, which is retrieved after the attack, only the content <-, Act_1, Act_3,> will be visible. The difference between the expected and the available output is due to the execution of the anti-investigation attack.

If this history-based evidence is considered during the reconstruction of the attack scenario, starting from state s_2,

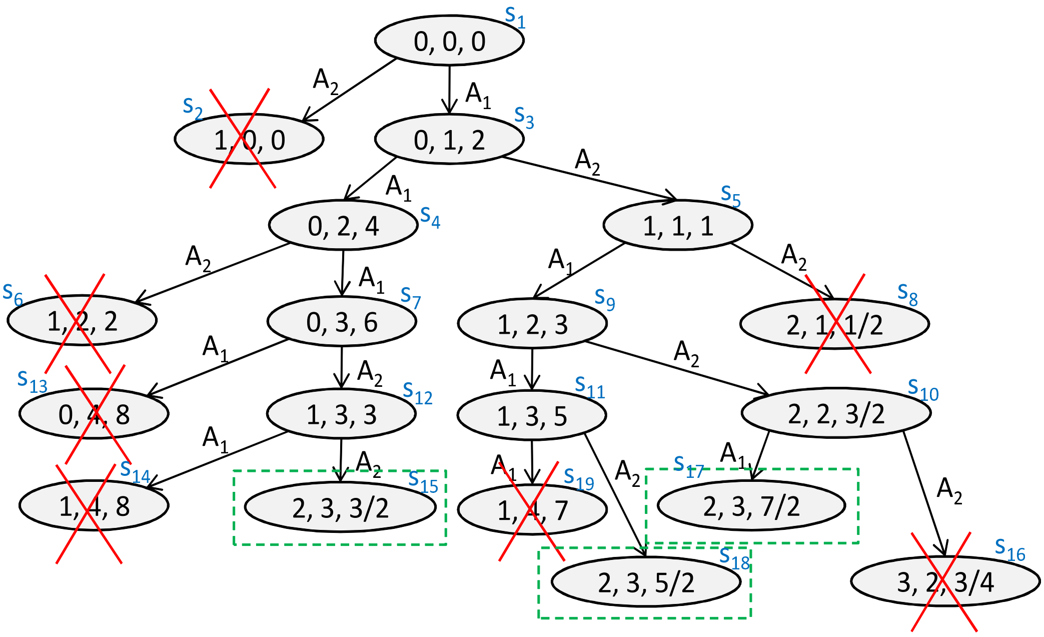


Figure 2. I-TLA attack scenario specification: an illustrative example.

action A_2 cannot be executed since it does not provide a state whose observation is included in the evidence. In other words, there is no entry in the log file after the one corresponding to the execution of A_1, which shows an indication regarding the executed action A_2.

Starting from this statement, we describe in the sequel a new observation function which allows coping with anti-investigation attacks.

4.3. Observations

Let $\{v_1,...,v_n\}$ represents the set of system variables. A variable in this set could represent a system component which is accessed in append mode (e.g., log or alert file, raw traffic capture) and takes a value or a series of values. Let $v^x(s)$ and $Card(v(s))$ represents the x^{th} value, and the number of values, in the series $v(s)$, respectively. We denote by $v(s_i)\circledast v(s_{i-1})$ the operation which consists in superimposing the series $v(s_i)$ on the series $v(s_{i-1})$ while keeping elements in $v(s_i)$ and discarding those situated beyond the limit of the intersection. Formally, $v(s_i)\circledast v(s_{i-1}) = [v^1(s_i),...,v^y(s_i)]$ where $y = min[Card(v(s_i)), Card(v(s_{i-1}))]$.

We denote by $obs^*(\omega)$ a new observation function over the executed attack scenario which allows capturing the situation where an anti-investigation attack has been conducted. Formally,

$$obs^*(\omega) = \langle obs(s_n \circledast ... \circledast s_0), obs(s_n \circledast ... \circledast s_1), ..., obs(s_n)\rangle \qquad (6)$$

where

$$s_i \circledast s_{i-1} = [v_1(s_i)\circledast v_1(s_{i-1}),...,v_n(s_i)\circledast v_n(s_{i-1})] \qquad (7)$$

By applying $obs^*(-)$ function on the scenario ω provided in the example of subsection 4.2, we obtain $obs^*(\omega)$= <-, A_1, A_3>. The output of this function is equal to the content retrieved from the system after the execution of the attack scenario which included an anti-investigation attack.

Theorem 1: Given an executed attack scenario ω, and an observation function $obs(-)$. If $obs^*(\omega) \neq obs(\omega)$, the attack scenario ω includes an anti-investigation attack.

Proof: We suppose that $obs^*(\omega) \neq obs(\omega)$ and there is no anti-investigation attack in the scenario ω. In the following we will disapprove this proposition.

Let v be an observable variable (with regard to $obs(-)$ function). Typically, since the variable v is in append mode, and the modifications are introduced to the tail of the series, the x^{th} value in $v(s_i)$ should be the one corresponding to the x^{th} value in $v(s_{i-1})$. In the absence of anti-investigation attack, none action executed from state s_i would modify the x^{th} value in $v(s_{i-1})$. Formally, the following condition should be satisfied:

$$\forall x \in [1..Card(s_i)] : v^x(s_{i-1}) = v^x(s_i) \qquad (8)$$

Therefore, $v(s_i)\circledast v(s_{i-1}) = v(s_{i-1})$. Assuming that the investigated system is modeled using only variable v and the attack scenario is composed of two states s_i and s_{i-1}, we obtain $obs(s_i \circledast s_{i-1}) = obs(s_i)$ and $obs^*(\omega) = obs(\omega)$. The proposition is therefore disapproved.

4.4. From IRPCM to I-TLA Specification

Starting from the IRPCMs built by the IRT, useful information, in the form of symptoms, unauthorized results, or actions, will be extracted and used to formally describe different type of evidences with I-TLA. We denote by useful information, any concept in the IRPCM having a degree of activation value that exceeds some predefined threshold, denoted by extraction threshold.

Symptoms are typically extracted from log files, traffic capture, or even keystrokes. They can be traduced to history-based evidences by transforming the whole content of the log file (including the record indicating the symptom itself) into an I-TLA history-based evidence. Symptoms extracted from alerts files indicate the occurrence of an events whose position in the reconstructed attack scenario cannot be determined. They will typically be transformed to non-timed I-TLA based evidence.

Actions selected from an IRPCM represent steps taken by a user or a process in order to achieve some result. A well intentioned reader has noticed that actions in the I-TLA library and actions in the IRPCM may not have the same form, and are not of the same granularity. In fact, an IRPCM action can be traduced to one or several consecutive I-TLA actions. In this context, for every selected IRPCM action an investigator has to extract sequence of elementary actions from the I-TLA library. The different obtained sequences will represent timed events-based evidences.

Unauthorized results represent unauthorized consequence of events. They are traduced to I-TLA predicate-based evidences. An investigator has to identify the system variable affected by the unauthorized consequence and then use it to describe the evidence.

Since the attack scenario may integrate anti-investigation attacks, the investigator has to locate in the IRPCM the set of concepts that are linked by anti-investigation relations (i.e., conceal, destroy, forge, and replace). Since anti-investigation attacks are executed to compromise evidences, the investigator has to determine which of the system variables, specified with I-TLA, describe the content of these compromised evidences. After that, it has to select observation functions that are defined to observe the content of these affected variables. This feature is highly essential for the reconstruction of the attack scenarios.

5. Executable Scenarios Generation Using I-TLC

To automate the proof in the context of digital investigation and generate executable attack scenarios showing with details how the attack was conducted and how the system progressed for each action part of the scenario, I-TLC [21], a model checker for I-TLA$^+$ specifications can be used. I-TLC is somehow an extension to TLC, the model checker of TLA$^+$ specification. It works by generating an optimized directed graph of states representing the space of possible scenarios generated from the I-TLA$^+$ specification. Despite checking that some types of computation are impossible as they violate safety properties, I-TLC aims to reconstruct execution (i.e., potential attack scenarios) that satisfy each form of evidences supported by I-TLA. I-TLC provides a novel concept entitled hypothetical action, defines techniques for its generation and management, and improves the representation of states. The directed graph is built by ensuring that a given node is reachable under optimal sets of hypothetical actions.

5.1. I-TLC's States Representation

I-TLC represents a node in the graph as a tuple of two information: *node core* and *node label*. The core of a node represents a valuation of the entire system variables, and the node label represents the potential sets of hypothetical actions under which the node core is reached. A reading of the node label indicates a) the state of the system in the current node, and b) the alternatives (hypothetical action sets) under which the system state is reachable.

$[(1,3),\{\{H_1,H_4\},\{H_2,H_3\}\}]$ represents an example of a node which can be represented by the graph generated by I-TLC. (1,3) is the node core, $\{H_1,H_4\}$ and $\{H_2,H_3\}$ represent the set of hypothetical actions under which the node core is reachable, and $\{H_1, H_4\}$, $\{H_2, H_3\}$ is the node label. This representation means that the system state (1,3), representing a valuation of the two variables x and y, respectively, is reachable under one of the two sets of hypothetical actions $\{H_1,H_4\}$ or $\{H_2,H_3\}$.

5.2. Generation of Hypothetical Actions

Generation of potential attack scenarios may fail if the library of actions is incomplete. In fact, for the particular case of attack scenarios that involve the use of unknown techniques, the system may come at some state while being unable to reach another state that if appended to the scenario under construction, will make it satisfy all the available evidences. To alleviate this issue, I-TLC tries to generate a hypothetical action and append it to the graph under construction, whenever available evidences are not completely satisfied.

The idea behind the generation of hypothetical actions is based on the fact that unknown actions can be generated if additional details about internal system components (i.e., those abstracted by the specification) are available. This detail involves a description of how these internal system components are expected to behave (if atomic actions are executed on them) and how they depend on each other. These internal system components are modeled by a specific set of variable denoted by internal variables. The other variables specified by I-TLA are denoted by external variables.

Semantically, a hypothetical action is true or false for a pair of states <*s*, *t*>. Syntactically, a hypothetical action is modeled as a series of hypothetical atomic actions, executed one after the other from state *s* to move the system to state *t*. It is defined in the following form $H = m_{ie}h_0 \rightarrow ... \rightarrow h_n m_{ei}$. m_{ie} defines a mapping from the external variables values to the internal variables values in state *s* and m_{ei} defines a mapping from the internal variables to the external variables in state *t*. The set of h_i (*i* from 0 to *n*) represents executed hypothetical atomic actions. A hypothetical atomic action h_i only modifies a single internal variable, and represents a relation between two consecutive internal system states. During hypothetical actions generation, I-TLC needs access to the library of hypothetical atomic actions. This library describes all the potential hypothetical atomic actions that can be executed on the investigated system.

During scenarios generation, several hypothetical actions may be appended whenever needed. I-TLC manages hypotheses following the two key ideas. First as hypotheses are not completely independent from each others and some hypotheses are contradictory, I-TLC avoids reaching a state under a contradictory situation. In this context, the library of hypotheses indicates potential contradictory sequence of hypothetical atomic actions. Second, in order to ensure that generated hypothetical actions are at the maximum close to real actions performed on the system, I-TLC defines techniques to refine the selection of hypothetical atomic actions.

5.3. Generation of Anti-Investigation Attacks

Typically, when an I-TLA action is to be executed, I-TLC verifies whether it satisfies all available evidences, especially history-based observations. Similarly to the case of unknown actions (as discussed in the previous subsection), the generation of potential attack scenarios may fail if the history-based observations were compromised using an anti-investigation attack. To cope with such an issue, I-TLC handles separately actions which modify the compromised evidences (with regard to history-based observations detected by I-TLA in the previous phase to be compromised using anti-investigation attacks).

Let $\mathcal{O}$ be the set of observations compromised by anti-investigation attacks, and *V* be the set of variables affected by these attacks. I-TLC could execute, during the reconstruction of the attack scenario, an action, say *A*, which do not create states satisfied by the available observations in $\mathcal{O}$, provided that the executed action modifies at least a variable in *V*. In the sequel, an action which satisfies the above conditions will be entitled *Prep-anti-investigation* action.

To support the generation of *Prep-anti-investigation* actions, heuristics can be used so that only accurate actions will be integrated to the attack scenario under construction. These heuristics exploit the values that could be taken by some other variables in the execution. For instance:

- Execute an action *A* if one of the collected evidences, namely the history-based evidences, shows that later the user will get a privileged system access. Such condition would mean that an anti-investigation attack can potentially be executed.
- Discard the attack scenario under construction if the number of states between the first generation of action *A* and the execution of the anti-investigation action has exceeded a threshold.

While staring from *the first generation of Prep-anti -investigation* action, the generated attack scenario will no longer satisfy the available observations, I-TLC should verify later that, further to the execution of some I-TLA action, which will typically be an anti-investigation attack (included in I-TLA as evidence), the attack scenario under construction, say ω, becomes satisfied by all evidences. For a completely generated attack scenario, say ω, which included, at some step in the execution, an anti-investigation action, I-TLC should verify that $obs(\omega) \neq obs^*(\omega)$.

5.4. Inferring Scenarios with I-TLC

To generate potential scenarios of attacks, DigForNet uses I-TLC Model Checker, which follows three phases. The reader is referred to [21] for a detailed description of I-TLC algorithms.

5.4.1. Initialization Phase

During this step, the generated scenarios graph is initialized to empty, and each state satisfying the initial system predicate is computed and then checked whether it satisfies the system invariants and the set of evidences. In that case, it will be appended to the graph with a pointer to the *null* state, and a label equal to $\varnothing$ (as no hypothetical action is generated).

5.4.2. Forward Chaining Phase

The algorithm starts from the set of initial system states, and computes in forward chaining manner the entire successor states that form scenarios satisfying evidences described in I-TLA. Successor states are computed by executing an I-TLA action or by generating a hypothetical action or Prep-anti-investigation action, and executing it.

When a new state is generated, I-TLC verifies if another existing node in the graph has a node core equal to that state. If the case is false, a new node, related to the generated state, is appended to the graph under construction, and linked to its predecessor state. If the case is true, the label of the existing node is updated so that it embodies the set of hypothetical actions under which the new system state is reachable. During label update, I-TLC ensures that each node label is provided with the following properties: soundness (a node holds the set of hypothetical actions under which its core is reachable), consistency (none set of hypothetical actions in the node label is an inconsistent or contradictory one), completeness (every set of hypothetical actions in the node is a superset of some other hypothetical actions), and minimal (none set of hypothetical actions is a proper subset of any other). If the scenario yielding the new generated state satisfies all the evidences, the system state is considered as a terminal state.

5.4.3. Backward Chaining Phase

All the optimal scenarios that could produce terminal states generated in forward chaining phase and satisfy the available evidences, are constructed. This helps obtaining potential and additional scenarios that could be the root causes for the set of available evidences. During this phase, the algorithm starts with a queue holding the set of terminal states generated in forward chaining phase. Afterwards, and until the queue becomes empty, the tail of the queue is retrieved and its predecessor states is computed. The new generated states are managed and appended to the graph under construction with the same manner followed in forward chaining phase.

All potential scenarios are supposed to be generated by I-TLC. The only exception may occur due to the lack of actions in the library of elementary actions. Nonetheless, the use of hypothetical actions concepts allows alleviating this problem.

6. Case Study

To demonstrate how DigForNet works, we provide in the following a case study related to the investigation of a compromised Linux Red Hat 7.2 operating system, which was deployed as a Virtual Honeypot in a VMWare session. The compromised system was suspended with VMWare immediately after the attack and a live image was created and posted by the Honeynet Project[1] for investigation. This case study deals with an investigation of a live system, the attack is highly complicated and requires advanced digital investigation skill knowledge, and the conducted scenario integrates several anti-investigation actions. In this case study, we will start by describing the attack. Then, we will show the use of DigForNet to investigate such incident.

6.1. Attack Description

First, the attacker probed the HTTP server from the machine identified by the IP address 213.154.118.219. Then, the attacker tried to exploit the Apache SSL handshake bug. Using this vulnerability, he gained a remote access as the Apache user. After that, he escalated his privilege and gained root access. At this level, the attacker conducted many attempts to install a rootkit. Only one of these attempts has succeeded. The following paragraphs describe the rootkits installation attempts.

The attacker has downloaded the tarball rk.tar.gz from geocities.com/mybabywhy/rk.tar.gz. Obviously, he then installed the rk.tar.gz. This install operation infected some binary files on the system, including ifconfig, ls, netstat, ps and top, and saved their original version in /usr/lib/libshtift. When the install script of rk.tar.gz finished the installation process, some system files (such as /bin/ps) have been replaced; mails with information about the system have been sent to mybabywhy@yahoo.com and buskyn17@yahoo.com; new unknown processes have been run as daemon; and the log file has been deleted to hide the attacker actions. After this, the hacker downloaded other tools including abc.tgz, an installation script for the current SSH server; and mass2.tgz, which is an exploit used to hack the server. However, the attacker has failed to stop the SSH daemon and has installed an SSH server under the file name “smbd -D”. The attacker does not even know the backdoor password. So, he carried out a novel attempt. He downloaded adore rootkit and tried to install it. But the install operation failed.

The attacker did not give up. He again gained a root access. This time, the attacker used the program gods (a shell script from izolam.net), to download adore LKM and an SSH server. After this, the attacker has installed the SucKIT rootkit using the installation script inst. This time, the rootkit installation has succeeded. The attacker also run xopen and lsn programs and moved /lib/.x/.boot from /var/tmp/.boot. After this, the attacker has connected to the FTP server identified by the IP address 63.99.224.38. Then, the file /root/sslstop.tar.gz has been moved from /lib/.x/s.tgz. It contains the sslstop program which modifies httpd.conf to disable the SSL support. The program sslport modifies httpd.conf to change the default SSL port (443) to another port (3128 in this case). The primitive HAVE_SSL has been replaced by HAVE_SSS in /etc/httpd/conf/httpd.conf. This indicates that sslstop has been run. In addition, the attacker downloaded psyBNC from www.psychoid.lam3rz.de/ psybnc and installed it. This program is used to hold open an IRC connection and run a proxy IRC in order to hide the user’s IP address. Using psyBNC, the user *sic* has connected from sanido-09.is.pcnet.ro to fairfax.va.us. undernet.org, an IRC server. He has created an account named redcode.

6.2. IRPCM Construction

The generated IRPCM is given by Figure 3. Actions, symptoms, and unauthorized results in this IRPCM are depicted by plain, dashed, and dotted ellipses, respectively.

To construct the IRPCM, we used the evidences which

[*] Honeynet Project-Scan of the Month #29 http://old.honeynet.org/scans/scan29/

were collected on the compromised system. Three concepts in the form of symptoms, namely S_1, S_2, and S_3, are initially appended to the IRPCM. Symptom S_1 indicates that the httpd log file contains suspicious entries showing potential exploit of ssl. Symptom S_2 indicates the existence of suspicious connections to the web server from 213.154.118.218 in the /var/log/ messages. Symptom S_3 shows that the web server banner is indicating a vulnerable version of Apache/OpenSSL. These three symptoms are linked to the concept A_1 which represents the action "Execution of mod_ssl/OpenSSL exploit". The latter action leads to the creation of the following unauthorized result, denoted by U_1 "Unauthorized access to the system with privileged rights" meaning that the intruder can execute some commands on the compromised system. Therefore, an edge is appended to the IRPCM from the concept A_1 to the concept U_1, creating a Cause/Effect relation.

The attacker reconnected to the system from the IP address 213.154.118.218. In the IRPCM this action is defined by the concept A_2, succeeds the action A_1, and precedes the action A_3 which represents a tentative to install the Adore rootkit. Action A_3 is vindicated by the content of log files and some mails from Apache indicating a failed installation of the Adore rootkit. In this context, the symptom S_5 is linked to the action A_3.

Always using the privileged access, the attacker downloaded rk.tar.gz from geocites and installed the rootkit it contains. This is shown by the action A_4 in the IRPCM. The latter is vindicated by the content of swap-colon.txt (symptom S_8 in the IRPCM). This rootkit leads to the unauthorized result U_2 which indicates that an unauthorized installation of programs on the system was performed. The attacker succeeded to install other programs such as a port scanner called sl2. Such activity is represented by action A_5 in the IRPCM. It is vindicated by some entries in the swap-colon.txt file too. By the completion of the installation, the attacker erased the log installation history of the tools contained in rk.tar.gz. Since this behavior represents an anti-forensic attack, the concept A_4 is linked to the concept S_9 using a destruction relation. The investigation process did not show any further use of this rootkit.

After this, the attacker used his privileged access to run the /lib/.x/hide script (action A_6 in the IRPCM) in order to destroy the symptoms "gods is running" and "inst is running". Two destruction relations are therefore created from the action A_6 to the symptoms S_{10} and S_{11}, respectively. The two latter symptoms give a proof regarding the execution of actions "Install SucKIT" and "Download SucKIT", respectively. In reality, the attacker installed the SucKIT rootkit as proven by the / partition analysis which shows the use of gods, a script used to download SucKIT, and inst, a script used to install this rootkit. SucKIT installation led to the unauthori-

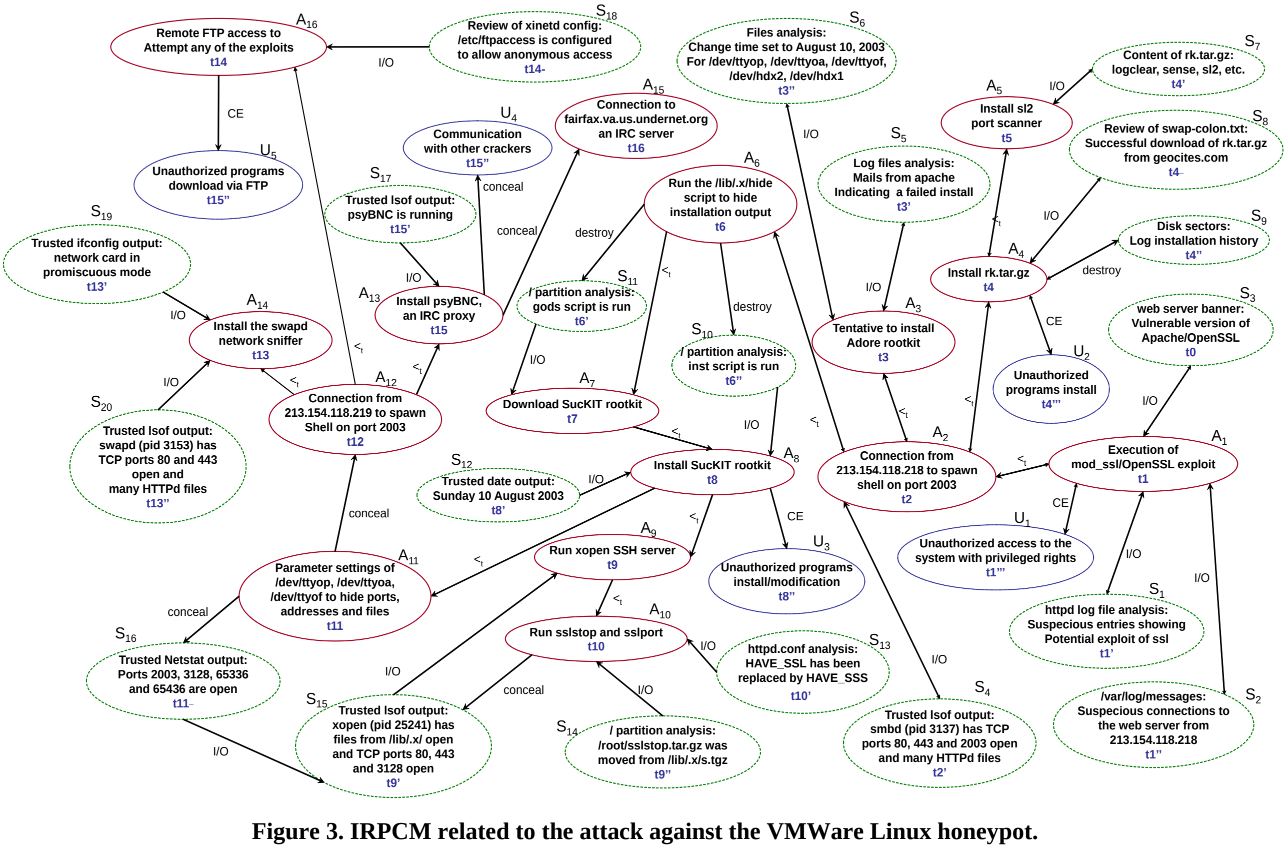

A16 Remote FTP access to Attempt any of the exploits t14
S18 Review of xinetd config: /etc/ftpaccess is configured to allow anonymous access t14-
U5 Unauthorized programs download via FTP t15''
U4 Communication with other crackers t15''
A15 Connection to fairfax.va.us.undernet.org an IRC server t16
S6 Files analysis: Change time set to August 10, 2003 For /dev/ttyop, /dev/ttyoa, /dev/ttyof, /dev/hdx2, /dev/hdx1 t3''
S7 Content of rk.tar.gz: logclear, sense, sl2, etc. t4'
A5 Install sl2 port scanner t5
S8 Review of swap-colon.txt: Successful download of rk.tar.gz from geocites.com t4
S17 Trusted lsof output: psyBNC is running t15'
A6 Run the /lib/.x/hide script to hide installation output t6
S5 Log files analysis: Mails from apache Indicating a failed install t3'
S19 Trusted ifconfig output: network card in promiscuous mode t13'
S9 Disk sectors: Log installation history t4''
A4 Install rk.tar.gz t4
A13 Install psyBNC, an IRC proxy t15
S11 / partition analysis: gods script is run t6'
A14 Install the swapd network sniffer t13
S3 web server banner: Vulnerable version of Apache/OpenSSL t0
A3 Tentative to install Adore rootkit t3
S10 / partition analysis: inst script is run t6''
U2 Unauthorized programs install t4'''
A12 Connection from 213.154.118.219 to spawn Shell on port 2003 t12
A7 Download SucKIT rootkit t7
S20 Trusted lsof output: swapd (pid 3153) has TCP ports 80 and 443 open and many HTTPd files t13''
A2 Connection from 213.154.118.218 to spawn shell on port 2003 t2
A1 Execution of mod_ssl/OpenSSL exploit t1
S12 Trusted date output: Sunday 10 August 2003 t8'
A8 Install SucKIT rootkit t8
U1 Unauthorized access to the system with privileged rights t1'''
U3 Unauthorized programs install/modification t8''
A9 Run xopen SSH server t9
A11 Parameter settings of /dev/ttyop, /dev/ttyoa, /dev/ttyof to hide ports, addresses and files t11
A10 Run sslstop and sslport t10
S1 httpd log file analysis: Suspecious entries showing Potential exploit of ssl t1'
S16 Trusted Netstat output: Ports 2003, 3128, 65336 and 65436 are open t11
S13 httpd.conf analysis: HAVE_SSL has been replaced by HAVE_SSS t10'
S15 Trusted lsof output: xopen (pid 25241) has files from /lib/.x/ open and TCP ports 80, 443 and 3128 open t9'
S14 / partition analysis: /root/sslstop.tar.gz was moved from /lib/.x/s.tgz t9''
S4 Trusted lsof output: smbd (pid 3137) has TCP ports 80, 443 and 2003 open and many HTTPd files t2'
S2 /var/log/messages: Suspecious connections to the web server from 213.154.118.218 t1''
I/O
CE
conceal
destroy
<t

Figure 3. IRPCM related to the attack against the VMWare Linux honeypot.

zed result “unauthorized programs install/ modification”, namely U_3. This rootkit was installed on 10 August 2003 as indicated by a trusted version of the date command. This information was provided by symptom S_{12} which is linked to the action A_8. The action representing the SucKIT installation was followed by action A_9 denoted by “run xopen SSH server”. This action is vindicated by the symptom S_{12} which is provided by the output of a trusted version of the lsof command. The latter shows that xopen is running on the compromised system. This output shows also that SSL is using port 3128 instead of 443. After this, the attacker executed action A_{10} to run sslstop and sslport programs. The content of the concept A_{10} is vindicated by symptom S_{10}. The latter is indicated by the analysis of /etc/httpd/conf/httpd.conf which shows that the primitive HAVE_SSL was replaced by HAVE_SSS. The script sslstop modifies httpd.conf to disable the SSL support. sslport modifies httpd.conf to change the default SSL port (443) to something else (3128 in this case) and then to conceal any port scanner output that can provide the symptom informing about the use of SSL. A concealment relation is appended from Action A_{10} to symptom S_{14} in the IRPCM.

After installing the SucKIT rootkit, the attacker closed the first connection and reconnected to the same web server from 213.154.118.219. This action, namely A_{12}, succeeds the action A_{11} in the IRPCM, which consists in setting the parameters of /dev/ttyop, /dev/ttyoa and /dev/ttyof to hide processes, addresses and files, respectively and then to conceal symptoms such as NetStat output. Action A_{11} conceals also the action of connecting to the server 213.154.118.219. A concealment relation is created from the concept A_{11} to the concept A_{12}.

Using the new shell, the attacker conducted three other actions. He first installed the swapd network sniffer (Action A_{14} in the IRPCM). This action is supported by the two symptoms S_{19} and S_{20}. S_{19} indicates that a trusted version of ifconfig showed that the network card was in promiscuous mode. S_{20} represents the output of a trusted version of the lsof command indicating that swapd was running using the pid 3153. Second, the attacker executed a remote FTP access (Action A_{16} in the IRPCM). This action is vindicated by the concept S_{18} representing the analysis of xinetd configuration. It shows that /etc/ftpaccess is configured to allow anonymous access. Action A_{16} leads to the unauthorized U_5 showing that an unauthorized downloading of programs via FTP was performed. Third, the attacker installed psyBNC (Action A_{13} in the IRPCM) which is an Internet Relay Chat (IRC) proxy. This action conceals the unauthorized result U_4 which shows a communication with other crackers. By concealing U_4, the IRC program allows communications without revealing the intruder identity. Action A_{13} is vindicated by the symptom S_{17}. In fact the execution of a trusted lsof command on the compromised system shows that psyBNC is running. Using psyBNC, the attacker connected to the IRC server fairfax.va.us.undernet.org (Action A_{15} in the IRPCM). An edge labeled by a concealment relation is created from the concept A_{13} to the concept S_{17}.

6.3. Extracting Evidences from IRPCMs

We model the investigated system using six variables; namely *Pr*, *hhtplog*, *port*2003, *ConAddr*2003, *SorftLog*, and *AppSoft*. They represent the system privilege granted to the remote user (i.e., the attacker), the tail of the content of the web service http log file, the service running on port 2003, the IP address connected to port 2003, the content of residual software installation logs, and the additional software installed on the system.

The evidences extracted from the IRPCM in conjunction with the library of elementary actions are then used by the I-TLA logic to specify the set of potential attack scenarios. For the sake of space, we will only consider a specific part of the IRPCM and we will describe the related I-TLA specification. Concepts in the IRPCM having a degree of activation that exceeded a pre-defined threshold are traduced into I-TLA evidences.

The concept “Disk sectors: Log installation history” is traduced to history-based evidence in I-TLA. This evidence, which is provided by the log installation file, allows monitoring the content of variable *SoftInsLog*. The provided evidence represents an observation over such variable. Since it was targeted by an anti-investigation attack, it is equal to <-> showing that none log file, which could be left by the installed software, is on the system. Some concepts from the IRPCM, in the form of actions, are mapped to I-TLA actions. For instance, the two actions “Execution of mod_ssl/OpenSSL exploit” and “Install rk.tar.gz” are traduced to the two following I-TLA actions, say *A*1 and *A*2, respectively:

$$A1 \triangleq \begin{array}{l} \wedge Pr' = 1 \\ \wedge httplog' = \langle modsslattack \rangle \\ \wedge port2003 = "/bin/sh" \\ \wedge \langle ConAddr2003, SoftInsLog, AppSoft \rangle \end{array} \tag{9}$$

$$A2 \triangleq \begin{array}{l} \wedge Pr = 2 \\ \wedge port2003 = <-> \\ \wedge ConAddr2003 = <-> \\ \wedge \langle Pr, httplog, port2003, ConAddr, AppSoft \rangle \end{array} \tag{10}$$

Action *A*1 can be executed to compromise the web service using the SSL vulnerability. It consists in inducing the system to grant a shell on port 2003. Further to the execution of such action, an entry is appended to the HTTP log file showing that a suspicious behavior has occurred. Since, the exploitation of the OpenSSL vulnerability gives access using the privilege of the Apache user, variable *Pr* gets value 1. This value means that the access level is more privileged that the user access but less privileged than the root one.

Action *A*2 cannot be executed unless variable *Pr* is equal to 2 to mean that the user should have gained a root privilege on the system. The action consists in hiding the execution of services on port 2003, and the connection of suspicious hosts on port 2003, if the Linux commands ps and netstat are used. Actions *A*1 and *A*2 stand for timed event-based evidence showing that *A*1 is executed before *A*2 and both of them are part of the conducted attack scenario.

6.3.1. Executable Scenarios Generation by I-TLC

I-TLC was used to generate executable specification [18], of the potential attack scenarios from I-TLA specification. One potential attack scenario is generated and is shown by Figure 4. The scenario is composed of nine states where every state shows the value of the modeled variables. Edges linking states, are labeled by the name of the executed action.

The system starts with an empty HTTP and system log file. The attacker first connects to the web service and runs the modssl exploit to get system access with the Apache user privilege. After that, it connects to the shell granted on port 2003. Thus, variable *ConAddr*2003 gets the value of the IP address of the remote user. After that, the attacker makes a tentative to install the rk rootkit. The operation is logged to the installation log of this tool. However, since it fails, no files were integrated to the system directory and variable *AppSoft* remains unchanged. Later, the attacker conducts a storage-based anti-investigation attack to hide the content of the installation log file. It reconnects from another host identified by the IP address 213.154.118.218, escalates its privilege, installs the suckit rootkit, and configures it to hide the execution of the shell on port 2003 and connected the IP address.

In this execution, it can be noticed that the execution of action *A*2 creates state *s*3 which does not satisfy the content of the history-based evidence. In fact, since an anti-investigation attack was executed and detected during the IRPCM construction, I-TLC has allowed the execution of *A*2 because it modifies the content of variable *SoftInsLog* which was affected by the anti-investigation attack.

6.3.2. Hypothetical Actions Generation

I-TLC has generated some hypothetical actions. For the lack of space, we only kept one hypothesis among those generated. Starting from state s_6, I-TLC could not find an action described in I-TLA specification, which, if executed, lets obtaining a potential attack scenario that satisfies the history-based evidence. I-TLC looks within the

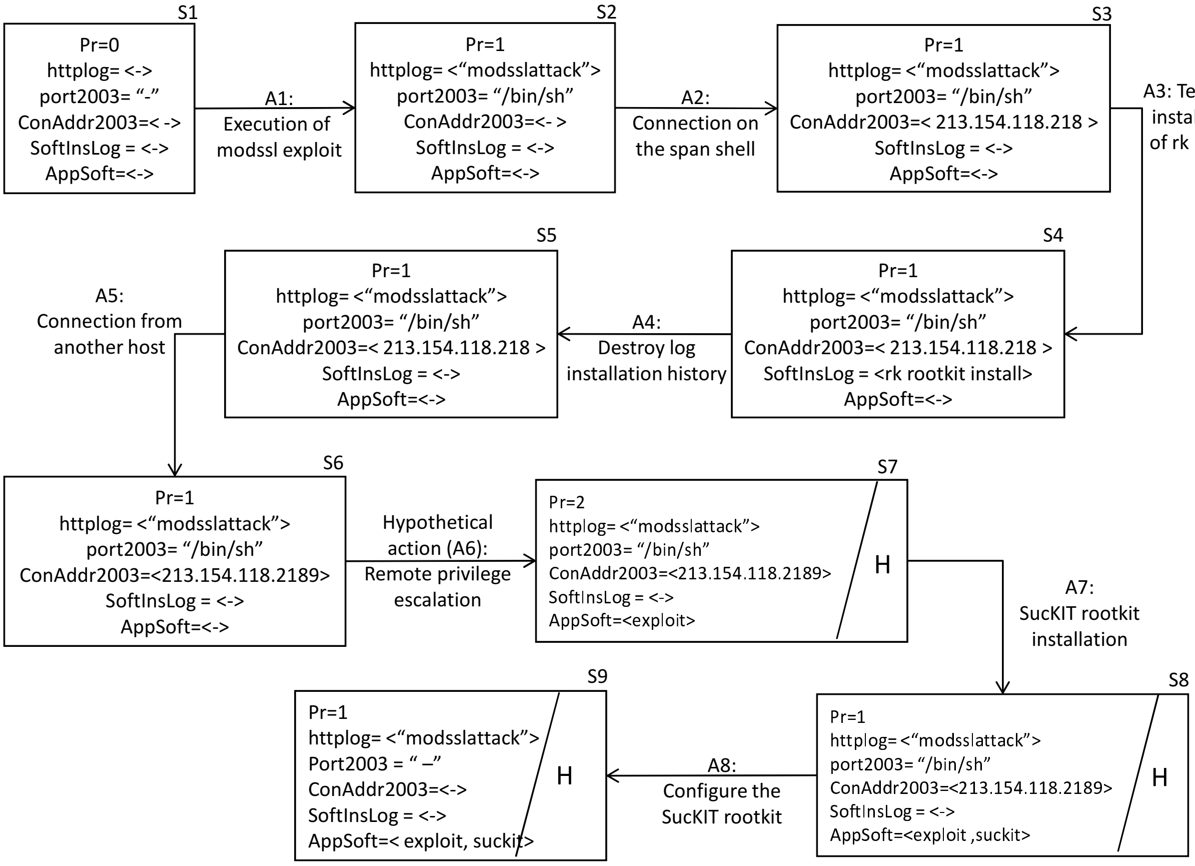


Figure 4. Fragment of the generated executable specification by I-TLC.

library of hypotheses and generates, a, hypothetical action H, and executes it to move the system to state s_7. The hypothetical action consists in uploading an exploit to the compromised system and executing it to get a root privilege. Further to the execution of the hypothetical action, variable *pr* gets value 2, and the value *exploit* is appended to the content of variable *AppSoft*. Later actions $A7$ and $A8$ are executed. I-TLC specifies that states s_7, s_8 and s_9 are reachable under the hypothetical action H by setting their label equal to the singleton H.

7. Conclusions

In this paper, we have developed a system for digital investigation of networks security incidents. This system uses formal techniques as well as the IRT members' knowledge to analyze the attacks performed against the networks. We have introduced the intrusion response probabilistic cognitive maps that are constructed by the IRT upon the occurrence of the attack. A formal language has been introduced to help specifying the attack scenarios based on the cognitive map. A model checker was built to automatically extract the attack scenarios and a hypothetical concept is introduced here to help in the construction process. To illustrate the proposed system, we used it in a real case of security attack.

8. References

[1] P. D. Dixon, "An overview of computer forensics," IEEE Potentials, Vol. 24, No. 5, pp. 7–10, 2005.

[2] P. Stephenson, "Modeling of post-incident root cause analysis," International Journal of Digital Evidence, Vol. 2, No. 2, pp. 1–16, 2003.

[3] T. Stallard and K. Levitt, "Automated analysis for digital forensic science: Semantic integrity checking," Proceedings of the 19th Annual Computer Security Applications Conference, Las Vegas, USA, 2003.

[4] P. Gladyshev, "Finite state machine analysis of a blackmail investigation," International Journal of Digital Evidence, Vol. 4, No. 1, 2005.

[5] P. Gladyshev and A. Patel, "Finite state machine approach to digital event reconstruction," Digital Investigation journal, Vol. 1, No. 2, pp. 130–149, 2004.

[6] B. D. Carrier and E. H. Spafford, "Categories of digital investigation analysis techniques based on the computer history model," Digital Investigation Journal, 3(S), pp. 121–130, 2006.

[7] S. Willassen, "Hypothesis-Based investigation of digital timestamps," Proceedings of Fourth Annual IFIP WG 11.9 International Conference on Digital Forensics, Kyoto, Japan, 2008.

[8] S. Y. Willassen, "Timestamp evidence correlation by model based clock hypothesis testing," Proceedings of the 1st International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia, 2008.

[9] A. R. Arasteha, M. Debbabi, A. Sakhaa, and M. Saleh, "Analyzing multiple logs for forensic evidence," Digital Investigation, Vol. 4, No. 1, pp. 82–91, 2007.

[10] A. Pal, H. T. Sencar, and N. Memon, "Detecting file fragmentation point using sequential hypothesis testing," Digital Investigation, Vol. 5, No. 1, pp. S2–S13, 2008.

[11] S. P. Peisert, "A model of forensic analysis using goal-oriented logging," PhD thesis, University of California, San Diego, 2007.

[12] A. S. Huff, "Mapping strategic thought," John Wiley & Sons, 1990.

[13] J. Krichene, M. Hamdi, and N. Boudriga, "Collective computer incident response using cognitive maps," IEEE International Conference on Systems, Man and Cybernetics, Hammamet, Tunisia, pp. 1080–1085, 2004.

[14] S. Rekhis, J. Krichene, and N. Boudriga, "Dig for net: Digital Forensic in networking," In Proceedings of the 3rd International Information Security Conference (SEC), Milan, Italy, 2008.

[15] B. D. Carrier and E. H. Spafford, "An event-based digital forensic investigation framework," Proceedings of Digital Forensic Research Workshop, 2004.

[16] B. Mangnes, "The use of Levenshtein distance in computer forensics," Master's thesis, Gjovik University College, 2005.

[17] E. Casey, "Digital evidence and computer crime," Second Edition, Academic Press, 2004.

[18] D. Drusinsky and J. L. Fobes, "Executable specifications: Language and applications," The journal of Defense Software Engineering, Vol. 17, No. 9, pp. 15–18, 2004.

[19] Y. Guan and A. K. Ghose, "Executable specifications for agent oriented conceptual modelling," Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT), France, pp. 475–478, 2005.

[20] M. Hamdi, J. Krichene, and N. Boudriga, "Collective computer incident response using cognitive maps," Proceedings of IEEE conference on Systems, Man, and Cybernetics (IEEE SMC 2004), The Hargue, Netherland, 2004.

[21] S. Rekhis and N. Boudriga, "A formal approach for the reconstruction of potential attack scenarios," Proceedings of the International Conference on Information & Communication Technologies: From Theory to Applications (ICTTA), Damascus, Syria, 2008.

[22] F. Kröger and S. Merz, "Temporal logic and state systems," Springer, 2008.

[23] S. Rekhis and N. Boudriga, "Formal forensic investigation eluding disk-based anti-forensic attacks," Proceedings of Workshop on Information Security Applications, Jeju Island, Korea, 2005.

[24] S. Garfinkel, "Anti-forensics: Techniques, detection and countermeasures," Proceedings of the 2nd International Conference on I-Warfare and Security, Monterey, USA, 2007.

[25] G. C. Kessler, "Anti-forensics and the digital investigator," Proceedings of 5th Australian Digital Forensics Conference, Perth, Australia, 2007.

Int. J. Communications, Network and System Sciences, **2009, 2, 874-878**
doi:10.4236/ijcns.2009.29101 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

A Velocity-Adaptive Handover Scheme for Mobile WiMAX

Caiyong HAO, Hongli LIU, Jie ZHAN
Department of Electronic Information Engineering, Hunan University, Changsha, China
Email: haocaiyong@gmail.com, hongliliu@vip.sina.com, jiezhanwl@163.com

Received March 7, 2009; *revised July* 27, 2009; *accepted September* 12, 2009

Abstract

Mobile WiMAX is a wireless networking system based on the IEEE 802.16e standard. In order to support mobile, some kinds of handover schemes must be adopted, and the hard handover is defined as mandatory. Since the data transmission should be paused during the hard handover process, it causes handover delay in mobile communication. The handover delay makes severe degradation in system performance when implemented in real-time applications such as IPTV and VoIP. The existing draft standard considers only the received signal strength when deciding handover initiation. However, the velocity factor also has an important influence on handover initiation and can not be neglected. To deal with these problems, this article proposes a velocity-adaptive handover scheme. This scheme adopts dynamic handover threshold according to different velocity to skip some unnecessary handover stages**,** reduces handover delay and enhances the network resource utilization. The simulation result and performance analysis validate the efficiency of the proposed scheme.

Keywords: Mobile WiMAX, Handover, Delay, Velocity-Adaptive

1. Introduction

In order to meet the demand of high data rate in wireless service for anytime, anywhere, and anyone, the Mobile WiMAX (Worldwide Interoperability for Microwave Access) based on the IEEE 802.16e [1] standard is developed for broadband wireless access as a promising technology. The IEEE 802.16e is the new, mobile version of the old WiMAX specification known as IEEE 802.16-2004 [2], which is a wireless, but fixed, data transmission scheme for providing broadband connection for metropolitan areas.

Support for handover (HO) is the most important amendment in IEEE 802.16e to embrace mobility. The HO is performed to maintain a continuous data transmission service for all applications when a Mobile Station (MS) is moving across cell borders of the BSs (Base Station). The IEEE802.16e defines three basic types of HO [3]: Hard handover (HHO), Macro Diversity Handover (MDHO) and Fast Base Station Switching (FBSS). MDHO and FBSS are soft handover. HHO is mandatory in WiMAX system while the others are optional ones. HHO adopt break-before-make scheme, the MS stops its radio link with the serving BS before establishing its radio link with the target BS. This is a rather simple HO but causes long HO delay and service disruption for some applications, especially when the MS is in high velocity. Thus it is unsuitable for services requiring low latency. In MDHO or FBSS scheme, a MS is registered to several BSs at the same time. For MDHO, a MS communicates with two or more BSs in a given interval, while for FBSS, a MS communicates with only one BS. Since both the MDHO and FBSS adopt the make-before-break scheme, they can improve link quality and provide better performance for users. In MDHO and FBSS, the Diversity Set shall be maintained by the MS and the BS. Since the MS and the BS have to scan and modify the Diversity Set periodically, these schemes demand more capacity and multiple channels in terms of bandwidth efficiency, which give rise to wireless resource waste.

Up until now, a few papers have proposed several schemes to deal with the research about HO in IEEE 802.16e. In [4], fast handover scheme for real-time downlink services using fast DL_MAP_IE is suggested, which allow forward data transmission before the establishment of the MS registration and authorization. However, it is not adopted by IEEE 802.16e while some fast handover schemes such as soft handover are adopted. In [5], it proposed to associate only one neighboring BS instead of several BSs. But, [5] hasn’t given an effective method to select only one neighboring BS. In [4] and [5], both neglected the HO threshold value’s influence on HO process.

In this paper, a velocity-adaptive handover scheme is proposed to reduce the HO delay and the waste of the wireless network resource. In WiMAX system, the Received Signal Strength Indicator (RSSI) is typically used as a measure of signal quality. And as soon as the RSSI form the current serving BS is lower than a threshold and the RSSI from a potential target BS reaches a threshold, HO is executed. The most important factor to initiate HO is the received signal strength and MS mobility. However, the IEEE802.16e standard considers only the former. In our scheme, the HO threshold is set variably according to the MS's velocity, which can reduce the HO delay and wireless network resource waste.

The remainder of this paper is organized as follows. Section 2 describes the HO process in IEEE 802.16e. Then Section 3 proposes the velocity-adaptive handover scheme. And the simulation and the performance analysis of the proposed HO scheme are given in Section 4. Finally the conclusion is provided in Section 5.

2. Handover Process and Analysis

2.1. Handover Process

The HO process defined in the IEEE 802.16e consists of two phases. In the first phase, network topology acquisition is carried out before HO initiation. Then the actual HO process is performed, which includes HO decision, HO initiation, ranging and network re-entry stage. The detail explanations of the HO process are given as follows.

2.1.1. Network Topology Acquisition

The BS periodically broadcasts the network topology information using MOB NBR-ADV messages, which contains channel information of neighboring BSs such as the BS ID (Identifier), radiation power, and their UCD (Uplink Channel Descriptor) and DCD (Downlink Channel Descriptor) information. Thus, the MS is able to synchronize with neighboring BSs without listening their DCD / UCD broadcast messages. Once the MS synchronize with the neighbor BSs, it can start a scanning and association procedure in order to select a candidate BS for HO. The scanning procedure is done through exchanging MOB SCAN-REQ/RSP messages with the serving BS. During the scanning process, all downlink and uplink transmissions are paused and the MS can optionally perform association with the neighbor BSs by performing initial ranging. To acquire ranging parameter and service availability information for the purpose of selecting a potential future HO target BS, the association procedure is performed between the MS and the candidate target. Initial ranging process performed during MS's scanning interval is optional. The MS can reuse this information for the future HO through the initial ranging values of associated BS.

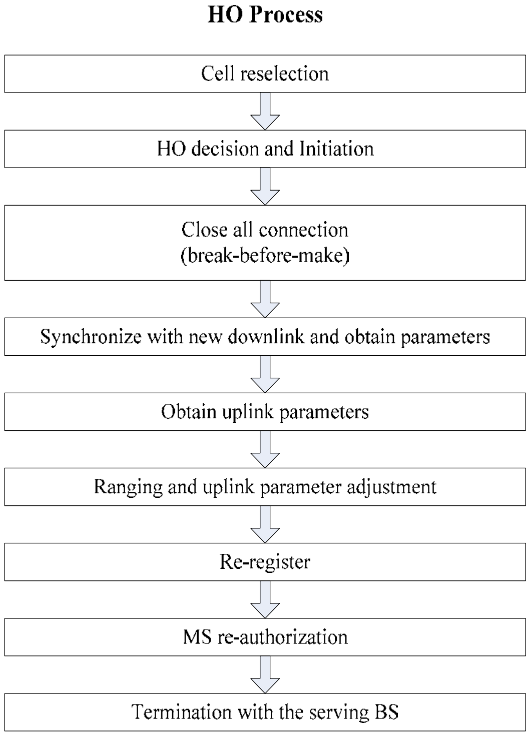

HO Process
Cell reselection
HO decision and Initiation
Close all connection
(break-before-make)
Synchronize with new downlink and obtain parameters
Obtain uplink parameters
Ranging and uplink parameter adjustment
Re-register
MS re-authorization
Termination with the serving BS

Figure 1. HO Process

2.1.2. Handover Process

HO is executed when a MS migrates from the serving BS to the target BS. The HO process consists of the following stages and is illustrated in Figure 1.

2.2. Handover Analysis

At first, the MS conducts cell reselection with the information acquired from network topology acquisition to evaluate the potential target BSs. If the network topology has not changed during the process, this stage can be abbreviated. Such procedure does not terminate the existing connection to the Serving BS. Then HO Decision and Initiation is performed, which is the beginning of the actual HO. The HO Decision consummates with a notification of MS intent to handover through MOB_MSHO-REQ or MOB_BSHO-REQ message. The HO is initialized, then the MS synchronize with target BS’s downlink to obtain the downlink and uplink parameters. If the MS has received a neighbor advertisement earlier, the synchronization procedure can be faster. And if a HO notification was sent by the serving BS and received by the target BS via backbone connection, non-contention-based initial ranging can be assigned, which shortens the HO delay.

A good handover scheme should minimize the HO delay and reduce wireless channel resource waste. However, some factors degrading the system performance exist in the HO process. Suppose that the MS moves in a certain velocity, If the velocity is low, the network topology architecture may maintain the same in a long time, thus in the cell reselection stage, the MS can use the same network topology information and skip this stage. Also, since the received neighbor advertisement (include BS ID, DCD and UCD) do not change, the MS could synchronization to target BS downlink by performing non-contention-based initial ranging. Therefore, the HO delay is able to shorten. However, for a high velocity, the channel condition change frequently, which makes the pre-obtained information become useless. So, during the actual HO process, the neighboring BSs scanning and contention-based ranging operation must be performed, which causes a long HO delay and wireless channel resource waste. To deal with these problems the handover scheme should be adjusted with the velocity.

Once the alternate target BS has been successfully selected, before going into normal operation, network reentry process is initiated. It includes MS authorization and new BS registration. After the success registration with the target BS, the MS sends MOB_HO_IND message and notifies the serving BS that the HO is completed.

3. Velocity-Adaptive Handover Scheme

In this section, we propose the velocity-adaptive handover scheme. Firstly, we mainly focus on the HO decision. The Mobile WiMAX specification [6] defines the procedures during the HO, but does not include the HO decision. Typically, the MS makes the HO decision according to signal quality, which can be measured by the RSSI. Figure 2 shows a simple case involving two BSs and an MS moving away from base station A (serving BS) toward base station B (target BS).

The $\text{Th}_{\text{handover}}$ represents the HO threshold. And the Th_{drop} is the point below which the quality of the link becomes unacceptable, and will lead to excessive packet loss and the session being dropped. The hysteresis value (ΔH) is used to eliminate thrashing effect. The idea is that the MS choose those BSs with high RSSI value which results in a better link-level communication with the target BS and a lower bit error rate. Typically, HO procedures are initiated when the RSSI drops below the $\text{Th}_{\text{handover}}$ (1). Also, HO is executed only if there is another BS for which the RSSI is at least ΔH higher than the Th_{drop} (2). These can be described as fallows:

$$RSSI_{cur} < Th_{handover} \tag{1}$$

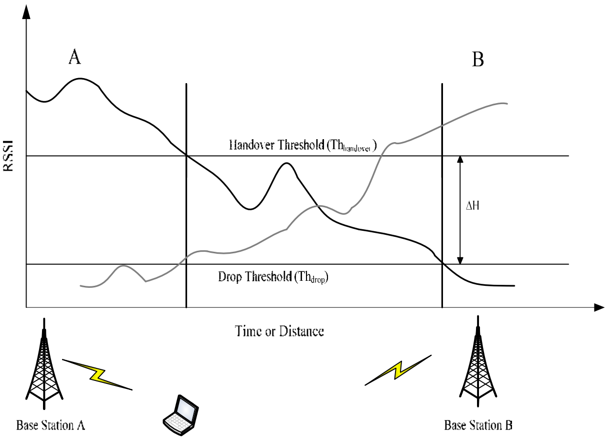

A
B
RSSI
Handover Threshold ($\text{Th}_{\text{handover}}$)
ΔH
Drop Threshold (Th_{drop})
Time or Distance
Base Station A
Base Station B

Figure 2. Handover decision based on RSSI.

$$RSSI_{candidate} > Th_{drop} + \Delta H \tag{2}$$

We have known that when the MS moves in a high velocity, the network channel information will change frequently, so that some unnecessary stages must be performed, which makes severe degradation in system performance such as HO delay and network resource waste. Form Equation (1) and (2), it can be known that if the $\text{Th}_{\text{handover}}$ is set higher, the frequency of HO initiated will be higher, thus the network channel information can be acquired quickly. It means that the probability of pre-obtained information used in the HO process (cell reselection and synchronization to target BS downlink) will be improved. So it can reduce the HO delay. However, the frequently performing HO causes a great wireless network resource waste. If the $\text{Th}_{\text{handover}}$ is set lower, the HO frequency becomes lower. Since the pre-obtained information do not update with the changes of channel condition, the neighboring BSs scanning and contention-based ranging operation must be performed. Therefore, the HO delay will be longer, but the wireless channel resource is consumed less. If the $\text{Th}_{\text{handover}}$ maintains a constant value regardless of the velocity, for a high $\text{Th}_{\text{handover}}$, the wireless channel resource waste will increase in the low velocity condition, and for a low $\text{Th}_{\text{handover}}$, the HO delay will increase in the high velocity condition.

To keep the balance of the two facets and cope with HO delay and wireless channel resource waste problems, we propose the velocity-adaptive handover scheme. The scheme changes the HO threshold by dynamically adapting the $\text{Th}_{\text{handover}}$ value based on velocity. It can be analyzed as follows:

$$Th_{handover} = Th_{drop} + \Delta H \tag{3}$$

According to the analysis above, ΔH should become larger with the velocity increase. In the stationary state (v=0), the network channel topology information keeps constant, and the HO delay is least, to avoid the unnecessary network waste, the $Th_{handover}$ should be setted as low as possible, so it might be the same value of the Th_{drop}, thus the ΔH=0(when v=0). In the highest velocity situation (v tends to infinity), the network channel topology information changes so quickly that the improvement of the $Th_{handover}$ have nearly no effect on the HO delay, for some unnecessary stages must be performed. Thus the ΔH will maintain a nearly constant value when in a very high velocity. In other situation, the ΔH would be set larger with the velocity becomes higher. In order to reflect this dynamic character, an influence factor r is proposed, which is the index associated with threshold value and velocity.

$$\Delta H = r \cdot Th_{drop} \tag{4}$$

$$r = \log_2（\mathrm{v}+1) \tag{5}$$

where v is the velocity of the MS. It can be acquired from the received SINR [7]. To make the simulation simple, the Th_{drop} is set as a constant value 2dB. Applying Equations (1), (2) and (3), the initial HO threshold $Th_{handover}$ is given by

$$Th_{handover} = Th_{drop}(1+\log_2（\mathrm{v}+1) \tag{6}$$

From the Equation (6), we can acquire that the $Th_{handover}$ can be set different value according to the different velocity of the MS. When the MS moves in a higher velocity, the $Th_{handover}$ is set higher, and thus the HO delay is reduced. When the MS moves in a low velocity, the $Th_{handover}$ is set lower. Since the unnecessary HO is reduced, the wireless channel resource waste becomes little.

4. Simulation and Performance Analysis

In order to evaluate the performance of the proposed handover scheme, a simulation platform is built. The HO threshold has an influence on HO delay, which depends on the velocity of the MS. Table 1 shows the main parameters and the default values used in the simulation.

The simulation was done with MS speeds between 1 and 40 m/s with 1 m/s step. The 40 m/s equals to 144km/h, which is above the 100 km/h limit described in IEEE 802.16e for a seamless handover. The Okumura-Hata model for small or medium city [8] was used for evaluating path losses.

When the MS is moving to the border of one BS in a certain speed, the signal quality of the Serving BS begins to degrade, if the signal level meets the Equation (5), initial HO process would be performed, and if the condition meets the Equation (6), the actual process of HO would be executed.

In the HO based on constant threshold scheme, we use the constant threshold value (Drop-threshold 2dB, Default handover-threshold 4dB) when the velocity varied. And, we propose velocity-adaptive HO scheme which adopts different threshold value according to the different velocity (6).

Table 1. Simulation parameters.

Simulation Parameters	Value
Handover type	HHO
Propagation model	Okumura-Hata
BS to BS distance	800m
Cell radius	5000m
Number of MS	1
Number of BS	2
Height of MS	2m
Height of BSs	30m
Frequency	5G
Drop-threshold	2dB
Default handover-threshold	4dB
Step of simulation	1 m/s
Speed range	1—40 m/s

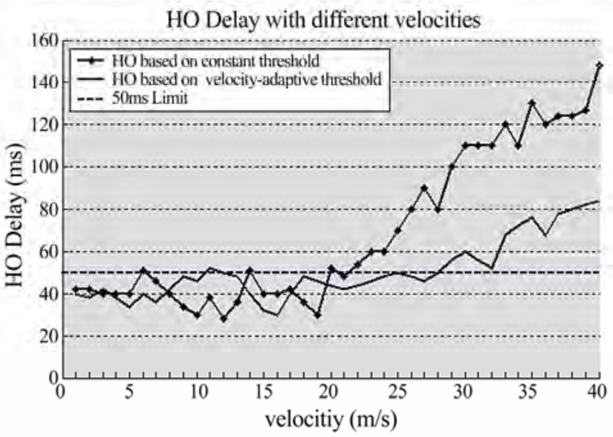


Figure 3. HO delay changes with the velocity.

The simulation results of constant threshold HO scheme and velocity-adaptive HO scheme are shown in Figure 3.

From Figure 3, we know that when the MS is in a low velocity (below the 20m/s), there is little influence when implementing our scheme. However, when the MS moves in a high velocity (above the 20m/s), the HO delay is obviously reduced in using velocity-adaptive scheme.

In HO based on constant threshold scheme, the HO delay remained below the 50ms limit (the WiMAX Forum defines that the Mobile WiMAX supports the HO delay should be less than 50ms) until the velocity rises up to 20m/s, apart from a few exceptions that exceeded the limit only few milliseconds. As the velocity rises, the HO delay is growing up to 150ms.

In HO based on velocity-adaptive scheme, the HO delay is below the 50ms limit until the velocity is up to 28m/s. After that, the delay increase to 82ms region with the 40m/s MS velocity. Since the high velocity causes a rapid changes of the network topology and wireless channel condition and the RSSI is not stable, some steps such as cell reselection and synchronize to downlink of the target BS may consume more time, so we can not always make the HO delay less than 50ms in a high velocity, but in this condition our scheme reduces the HO delay greatly compared with the HO based on constant threshold.

5. Conclusions

In this paper, a velocity-adaptive handover scheme is presented. According to the existing draft version of 802.16e standard, the HO initiation should be performed if the RSSI of the serving BS is lower than the threshold. However, it does not consider the velocity's influence on the HO process, and the HO threshold is set as a constant. So the velocity has a bad effect on the HO performance. To cope with this problem, our scheme is proposed, the HO threshold is set variably according to the MS mobility. The simulation results show that HO delay below 50ms limit when the velocity vary from 20m/s to 28m/s. And HO delay is greatly reduced when the velocity exceeds 28m/s. Therefore, this velocity-adaptive handover scheme can provide seamless communication for Mobile WiMAX in delay-sensitive and high velocity applications.

6. References

[1] IEEE P802.16e/D12, "Air interface for fixed and mobile broadband wireless access systems: amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands," 2005.

[2] IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001), "IEEE standard for local and metropolitan area networks–Part16: Air interface for fixed and mobile broadband wireless access systems," 2004.

[3] WiMAX Forum, Mobile WiMAX-Part I: "A technical overview and performance evaluation," 2006.

[4] S. Choit, G. H. Hwangt, *et al.*, "Fast handover scheme for real-time downlink services in IEEE 802.16e BWA systems," IEEE Vehicular Technology Conference (VTC 2005), Stockholm, Sweden, Vol. 3, pp. 2028–2032, May 2005.

[5] D. H. Lee and K. Kyamakya, "Fast handover algorithm for IEEE 802.16e broadband wireless access system," IEEE Wireless Pervasive Computing Conference, pp. 16–18 January 2006.

[6] IEEE 802.16e-2005, IEEE Standard for local and metropolitan area networks-Part16: "Air interface for fixed and mobile broadband wireless access systems-amendent2: physical and medium access control layers for combined fixed and mobile operation in licensed band," 2006.

[7] M. D Austin and G. L Stuber, "Velocity adaptive handoff algorithms for microcellular systems," IEEE Transactions, Vol. 43, pp. 549–561, August 1994.

[8] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," Wireless Communicaitons and Mobile Computing (WCMC), Vol. 2, pp. 483–502, May 2002.

Int. J. Communications, Network and System Sciences, **2009, 2, 879-887**
doi:10.4236/ijcns.2009.29102 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Fast Switching Fractional-N Frequency Synthesizer Architecture Using TDTL

Mahmoud A. AL-QUTAYRI, Saleh R. AL-ARAJI, Abdulrahman Al-HUMAIDAN
College of Engineering, Khalifa University of Science, Technology and Research, Sharjah, UAE
E-mail:{*mqutayri, alarajis*}*@kustar.ac.ae, humaidan*2*@gmail.com*

Received July 30, 2009; *revised August* 2, 2009; *accept September* 19, 2009

ABSTRACT

This paper presents an efficient indirect fractional frequency synthesizer architecture based on the time delay digital tanlock loop. The indirect type frequency synthesis systems incorporate a low complexity high performance adaptation mechanism that enables them to remain in a locked state following the division process. The performance of the proposed fractional-N synthesizer under various input conditions is demonstrated. This includes sudden changes in the system input frequency as well as the injection of noise. The results of the extensive set of tests indicate that the fractional-N synthesizer, proposed in this work, performs well and is capable of achieving frequency divisions with fine resolution. The indirect frequency synthesizer also has a wide locking range and fast switching response. This is reflected by the system ability to regain its lock in response to relatively large variations in the input frequency within a few samples. The overall system performance shows high resilience to noise as reflected by the mean square error results.

Keywords: Fractional, Synthesizer, Time Delay Tanlock Loop, Register Adaptation

1. Introduction

Frequency synthesizers, particularly the fractional type, are a fundamental component of the many types of modern wireless communication systems in use now a day. These complex wireless systems, be it for the ever growing mobile communications market or other applications, need to support a multitude of different wireless standards for disparate applications with their own data transfer requirements. However, irrespective of the wireless standard in use, the data to be transferred will somehow need to be modulated on a radio frequency (RF) carrier, and the modulated signal is then transmitted over the air, and received and demodulated at the receiving end. At both the transmitting and the receiving end, an accurate RF carrier signal must be generated. Therefore, a radio frequency synthesizer is always required in order to perform frequency translation and channel selection [1–5].

The design of frequency synthesizers continues to present major challenges due to the stringent RF requirements as well as the demands for high speed in digital transceivers supporting the drive for convergence. Conventional frequency synthesis techniques [1,2] in use today may be broadly classified into the following types:

- Phase-locked loop (PLL) based, or ‘indirect’
- Mixer / filter / divide, or ‘direct analog’
- Direct digital synthesis (DDS)

Each of these methodologies has its merits and limitations. Direct analog synthesis uses the functional elements of multiplication, division and other mathematical manipulation to produce the desired frequency, but this method is a very expensive one. DDS uses logic and memory elements to digitally construct the desired output signal. On the output side, a digital-to-analog (D/A) converter is used to convert the digital signal to analog domain. The main limitations of DDS are the limited bandwidth and spurious harmonic generation. PLL-based frequency synthesis has been widely used in industry. However, one of the major difficulties associated with the PLL-based technique is that a PLL with a wide frequency range cannot be achieved easily. In addition, fast switching is difficult to achieve. Typically, the output frequency step size of this method is the reference frequency. With fractional-N synthesis technique, finer frequency control can be achieved; however, these systems typically have very narrow bandwidth [7–10].

Today’s advanced communication systems demand frequency synthesizers of high resolution, wide bandwidth and fast switching speed. The fractional frequency synthesizer proposed in this paper is designed to achieve the above mentioned requirements. It is of the indirect

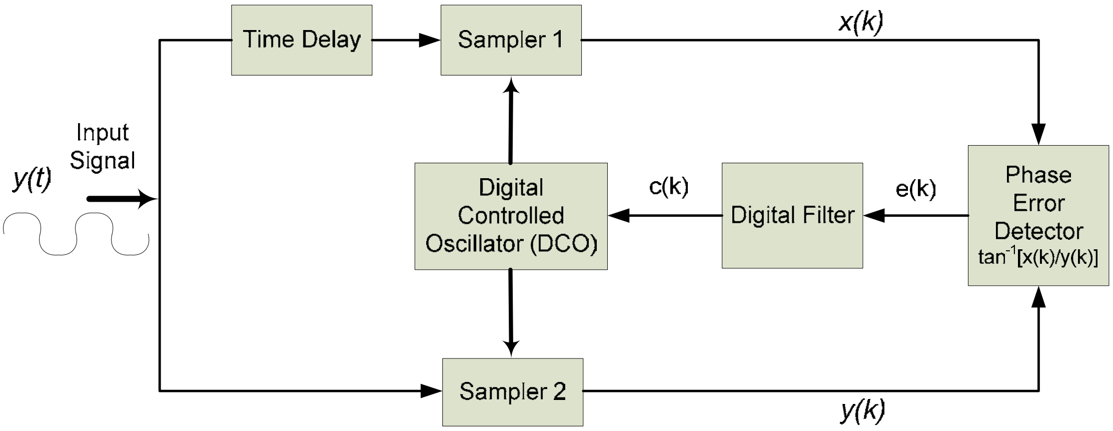


Figure 1. Architecture of time-delay digital tanlock loop.

type that uses time delay digital tanlock loop (TDTL) as the phase locking element, a divider, and an efficient adaptive control structure.

The paper is organized as follows. Section 2 presents the architecture and system equations of the time-delay digital tanlock loop. This also includes the locking range of the TDTL which is a major aspect of a phase lock loop. Section 3 discusses the process and the challenges of using the TDTL as a frequency synthesizer. The adaptation process introduced to enable the TDTL based frequency synthesizer to achieve locking for the fractional frequency division is detailed in Section 4. The results of the new TDTL based fractional frequency synthesizer (TDTL-FFS) for various division factors and under different input conditions are detailed in Section 5. An assessment of the performance of the TDTL-FFS is presented in Section 6. The conclusions of this work are presented in Section 7.

2. TDTL Architecture and System Equations

The architecture of the TDTL is shown in Figure 1. It consists mainly of two samples and hold blocks, a phase detector, a low pass filter, a digitally controlled oscillator, and a time-delay block [11]. Being comprised of these components, the TDTL lends itself for implementation in various digital systems technologies. The TDTL offers an inexpensive implementation and improved performance compared with other synchronization techniques. Compared with the conventional digital tanlock loop in [12], the TDTL in Figure 1 does not preserve the linearity of the phase characteristics due to the existence of input dependant phase shift caused by the time delay. However, this disadvantage is considered relatively minor due to the significant advantages offered by the TDTL, which include wider locking range and fast acquisition behavior. An in depth comparison of the conventional digital tanlock loop and TDTL with extensive results and discussion is given in [13]. The mathematical analysis of the TDTL under noise free conditions is detailed below. All of the signal notations are chosen in reference to the block diagram shown in Figure 1. The analysis follows a similar line to that given in [13,14].

The TDTL receives a continuous time sinusoid *y(t)* which is given by (1).

$$y(t) = A\sin\left[w_o t + q(t)\right] + n(t) \tag{1}$$

where A is the amplitude of the signal, ω_0 is the free running frequency of the DCO, $\theta(t)=(\omega-\omega_0)t+\theta_0$ is the information-bearing phase and $n(t)$ is the additive white Gaussian noise (AWGN). The signal is assumed not to have a DC component. Usually the phase process $\theta(t)$ is a translation of frequency or phase steps. w is the radian frequency of the input signal and θ_0 is a constant. A phase lag $\psi=\omega\tau$ is induced to the input signal after it passes through the time delay block. Therefore, $x(t)$ is generated, which is a phase shifted version of the input signal *y(t),* this signal is given by (2).

$$x\ (t) = A\sin\left[w_o t + q(t) - \Psi\right] + n^{/}(t) \tag{2}$$

where $n'(t)$ is the time-delayed AWGN due to τ. The aforementioned continuous time signals pass to the sample and hold blocks, and thereby get transformed to the discrete time signals in (3) and (4).

$$y(k) = A\sin\left[w_o t(k) + q(k)\right] + n(k) \tag{3}$$

$$x(k) = A\sin\left[w_o t(k) + q(k) - \Psi\right] + n^{/}(k) \tag{4}$$

where $q(k) = q[t(k)]$.

The sampling interval between the sampling instants $t(k)$ and $t(k$-1) is given by (5).

$$\mathrm{T}(k) = T_0 - c(k-1) \tag{5}$$

where $T_0=2\pi/\omega_0$ is the nominal period of the DCO and $c(i)$ is the output of the digital filter at the ith sampling instant. Assuming $t(0)=0$, the total time $t(k)$ elapsed up to the kth sampling instant is given by (6).

$$t(k)=\sum_{i=1}^{k}T(i)=kT_o-\sum_{i=0}^{k-1}c(i) \quad (6)$$

$$y(k)=A\sin\left[q(k)-w_o\sum_{i=0}^{k-1}c(i)\right]+n(k) \quad (7)$$

$$x(k)=A\sin\left[q(k)-w_o\sum_{i=0}^{k-1}c(i)-\Psi\right]+n^{/}(k) \quad (8)$$

and therefore, the phase error between the input signal and the DCO can be also defined as:

$$j(k)=q(k)-w_o\sum_{i=0}^{k-1}c(i)-\Psi \quad (9)$$

Having defined the phase error, Equations (7) and (8) can be rewritten as

$$y(k)=A\sin[f(k)+\Psi]+n^{/}(k) \quad (10)$$

$$x(k)=A\sin[f(k)]+n^{/}(k) \quad (11)$$

These signals are applied to the phase detector producing the error signal *e*(*k*) given in (12).

$$e(k)=f\left[\tan^{-1}\left(\frac{\sin[f(k)]}{\sin[f(k)+\Psi]}\right)\right]+z(k) \quad (12)$$

where $f(g)=-p+[(g+p)\bmod 2p]$, $z(k)$ is a random phase disturbance due to AWGN. The error signal *e*(*k*) represents a nonlinearly mapped version of the phase error. However, the effect of the nonlinearity is minimum and *e(k)* can be approximately linear if ψ is equal to, or in the vicinity of π/2. The digital filter, which has a transfer function given by *D*(*z*) receives the error signal *e(k)* and produces the signal *c*(*k*) that drives the DCO. Therefore, the system difference equation can be derived from (6) and (9) as

$$f(k+1)=f(k)-wc(k)+\Lambda_o \quad (13)$$

where $\Lambda_0=2p(w-w_0)/w_0$, and the AWGN terms are neglected since noise-free analysis is assumed. In the case of the conventional digital tanlock loop, the linear characteristic function of the phase detector enables the description of the loop as a linear difference equation, and hence finding the lock range using the stability criterions of its Z-transformed transfer function [12]. However, the nonlinear characteristic function of the TDTL phase detector results in a nonlinear difference equation, which can only be solved by numerical analysis. The lock range of the TDTL was analyzed in [11], using fixed-point theorems [15]. The digital filter of the first order loop is simply a gain block *G₁*, and the system equation is given by

$$f(k+1)=f(k)-K_1^{/}h[f(k)]+\Lambda_o \quad (14)$$

where $K_1^{/}=wG_1$. Defining *K*1 as ω_oG_1 will result in $K_1^{/}=K_1/W$, where $W=w_o/w$. The nominal phase lag Ψ_o induced by the time delay units on the input can be initially arranged by manipulating the parameters *ω*o and *τ* in the manner given by $\Psi_o=\omega\tau_o$. Therefore, the locking range can be acquired by numerically solving the inequality

$$2|1-W|<K_1<2W\frac{\sin^2(a)+\sin^2(a+\Psi_o)}{\sin(\Psi_o)} \quad (15)$$

where $a=\tan^{-1}(b)$, $b=\frac{\sin(\Psi)\tan(h)}{1-\cos(\Psi)\tan(h)}=\frac{\sin\ (\Psi)}{\cot(h)-\cos(\Psi)}$

and $h=\frac{\Lambda_o}{K_1^{/}}$

One of the properties of the first order TDTL is that it converges to a nonzero steady state phase error, which is translated with a phase offset between the pulses of the DCO and the zero crossings of the input signal. The steady-state value of the phase error is given by $f_{ss}=s+jp$ where $j\in\{1,0,-1\}$. Figure 2 shows the locking range of the first-order TDTL for different values of Ψ_o as well as the conventional digital tanlock loop locking region [11]. Note that the region enclosed by (1), (2) and (3) is for the conventional digital tanlock loop; the region enclosed by (1), (2) and (4) is for the TDTL when $\Psi_0=\pi/2$; and the region enclosed by (1) and (5) is for the TDTL when $\Psi_0=\pi$.

The range of independent locking of the TDTL and the effect of initial phase error are studied in depth in [11, 13]. Since the TDTL has non-linear characteristic function numerical analysis were used to determine the range of independent locking. The analysis shows that the TDTL offers an advantage on the conventional digital tanlock loop in this regard.

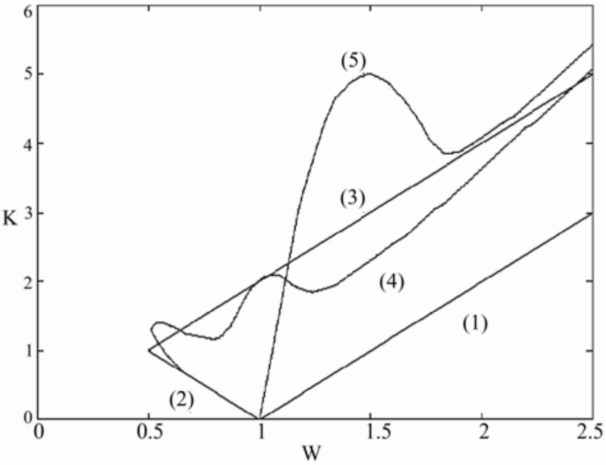


Figure 2. Major locking range of the first-order TDTL for different values of $Y_0=w_0t$.

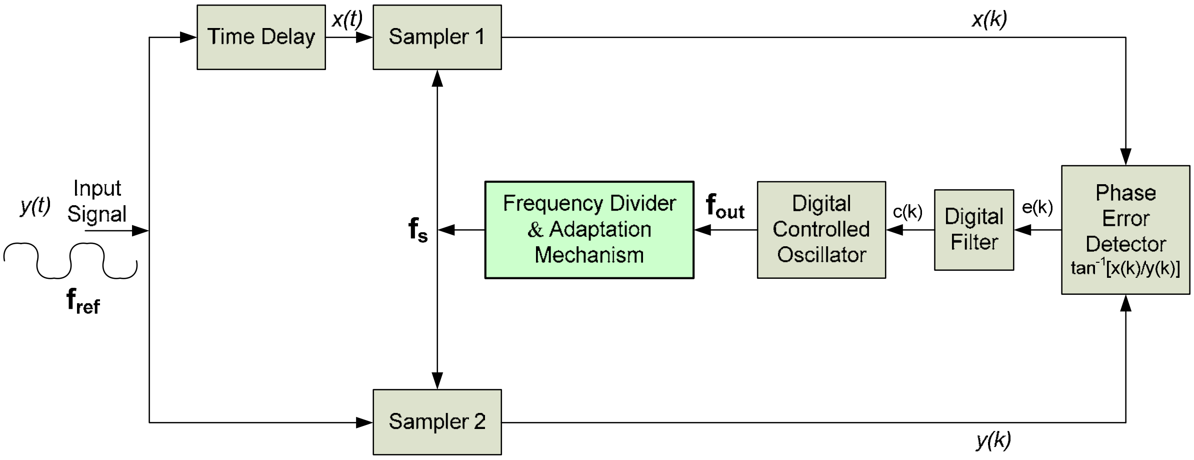


Figure 3. Basic architecture of TDTL-based frequency synthesizer.

3. TDTL-Based Frequency Synthesizer Architecture

In order to utilize the TDTL system, described in the previous section, for indirect frequency synthesis the original system architecture needs to be modified as shown in Figure 3. The additional block in Figure 3 is a composite frequency divider with adaptation control mechanism. This is slightly more involved than a conventional indirect phase lock loop based synthesizer, which would normally only include a divider in its basic form. However, in the TDTL-based frequency synthesizer (TDTL-FS) structure in Figure 3 the adaptation mechanism is essential for the proper operation and stability of the system, as explained below. Once the system operates correctly it results in dividing the DCO output frequency by the desired factor specified by the composite divider block.

The necessity to incorporate an adaptation mechanism is dictated by the direct bearing that a divider block, after the DCO, has on the overall locking state of the TDTL-FS. If the system parameters of the original TDTL in Figure 1, without the divider, are selected such that it operates at optimum point “A” within the first-order loop locking range depicted in Figure 4, then such a system will have a stable behavior, within bounds, as demonstrated by the results in Section 4 as well as in [14]. However, it was observed that the moment a division of the TDTL DCO output frequency is attempted the complete system gets driven out to a point outside the lock range, such as point “B”, of Figure 4. Once the system moves to such points outside the lock range, it will be in an unstable state and hence cannot be used. The reason for this is that the divider block affects both the DCO free running frequency and the error signal at the input of the digital filter block. Both of these parameters have direct effect on the system locking.

Therefore, incorporating an adaptation mechanism with the divider in Figure 3 is necessary to overcome the critical locking problem outlined above. It is to be noted that point “A” in Figure 4, at which W=1and K_1=1, is considered an optimum point within the lock range of the first-order TDTL with π/2 delay, because it allows for maximum symmetrical variation in both the input frequency and the loop gain while maintaining the system locked state. In this study it is always assumed that the loop is operating at this point prior to switching the divider for frequency synthesis.

4. TDTL Fractional-N Frequency Synthesizer

The complete block diagram of the TDTL-FFS is depicted in Figure 5. In addition to the TDTL, the system includes a fractional divider block that uses the pre- scaler technique, and an efficient adaptation mechanism that utilizes registers in order to maintain coherence between the input signal frequency (f_{ref}) and the divider output frequency (fs). As the TDTL-FFS involves multiple divisions, fast system response that counteracts the effect of the division and keeps the complete system in lock is highly desirable.

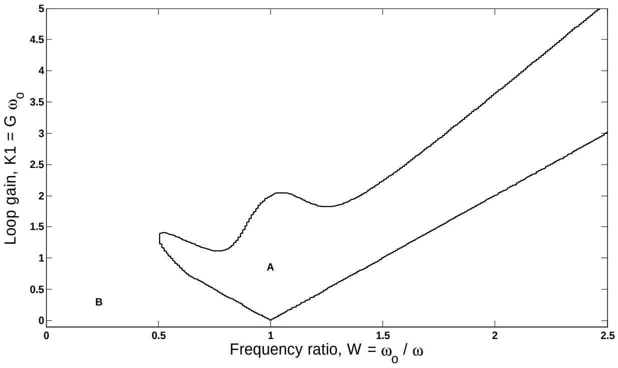


Figure 4. Lock range of 1^{st} order TDTL with $\pi/2$ delay.

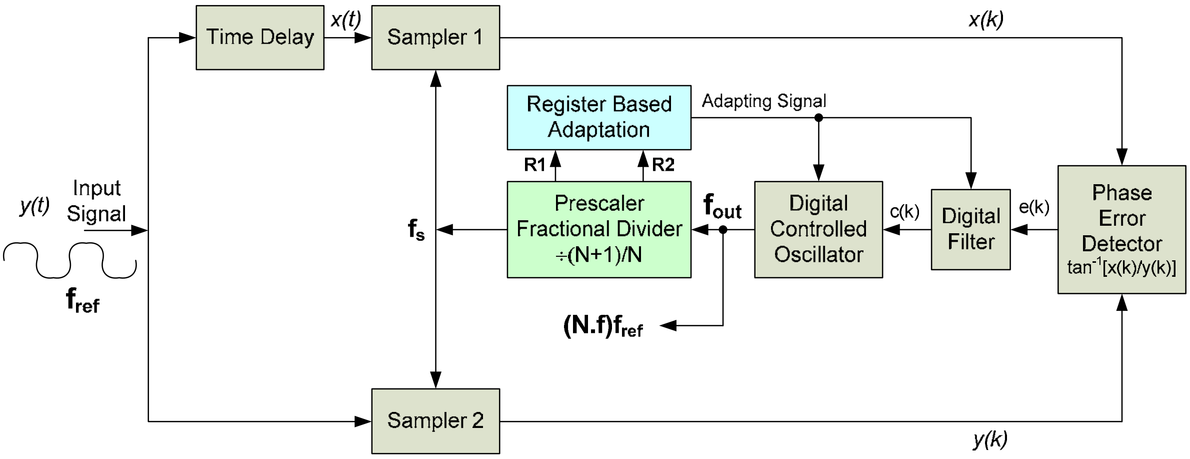

Time Delay
x(t)
Sampler 1
x(k)
Register Based Adaptation
Adapting Signal
R1
R2
Input Signal
y(t)
f_s
Prescaler Fractional Divider ÷(N+1)/N
f_out
Digital Controlled Oscillator
c(k)
Digital Filter
e(k)
Phase Error Detector tan^-1[x(k)/y(k)]
f_ref
(N.f)f_ref
Sampler 2
y(k)

Figure 5. TDTL-FFS with register adaptation.

In the TDTL-FFS in Figure 5, the output frequency (f_{out}) varies by a fraction of the input reference frequency (f_{ref}). This is achieved by realizing an equivalent fractional division ratio, such as N.f, where N and f are the integer and fractional parts of the division respectively. Figure 6 shows the combined structure of the prescaler fractional divider (PFD), that achieves N.f, and the register-based adaptation (RBA). The shaded blocks in Figure 6 make up the RBA while the rest form the PFD module. The PFD divides by N or N+1 according to the control unit. If the PFD divides by N for P output pulses of the DCO and N+1 for Q output pulses, then the equivalent division ratio will be as given in (16).

$$\frac{P+Q}{P/N+Q/(N+1)} \tag{16}$$

The RBA part of Figure 6 uses two registers and a multiplexer. Registers 1 and 2 store the division outputs of dividers N and N+1 respectively. The storage process is controlled by pulses R1 and R2. The outputs of the registers are fed back to the DCO and the digital filter to compensate for the effect of the division and hence keep the overall system in lock.

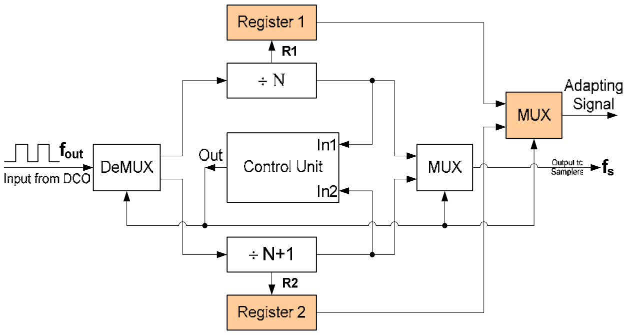


Figure 6. Adaptation mechanism for fractional divider using register approach.

5. Simulation Results of TDTL-FFS

The TDTL-FFS architecture described in the previous section was modelled and simulated using MATLAB/ Simulink. The performance of the complete system was evaluated for two criterions at this stage. The first is the ability of the synthesiser to stay locked or regain locked state should it lose that due to the division process. The second is achieving the correct frequency division ratio with respect to the system input signal frequency. The testing process involved subjecting the system to sudden changes in the input signal frequency by applying positive and negative frequency steps to the input, and changing the division factor. It is assumed that the system is stable prior to the application of step inputs and the division. The subsections below show the results of the first-order TDTL without the presence of a divider, and the TDTL-FFS.

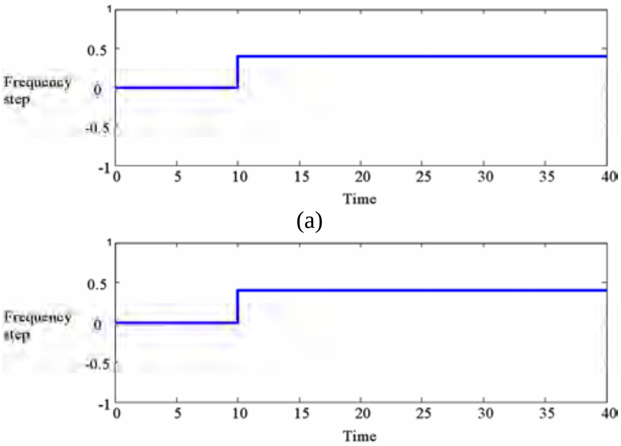
5.1. First-Order TDTL

The first-order TDTL was simulated and its operation verified by applying both positive and negative frequency steps to the input. The steps represent the sudden change in the input signal frequency. The reason behind testing the TDTL by itself is to form a kind of a reference or a signature that can be used for further assessment of the performance of the TDTL synthesizers. This necessity is dictated by the fact that, as discussed in the previous section, utilizing the TDTL as a synthesizer affects the architecture as well as the stability of the loop. A loop that is unstable obviously will not be useful.

The loop was set to operate at optimum point "A" in Figure 4, where $W=\omega_o/\omega_{in}=1$, $K_1=G\omega_o=1$, and $\psi_o=\omega_o\tau = \pi/2$. An input was applied to the TDTL with a positive frequency step of 0.4, this means that the operating point will shift to the point where W=0.71. Although the input shifted the loop to another operating point, the new point is still within the locking range of the loop. Hence, the loop did not go out of lock and stabilized indicating that the loop has locked onto the new frequency. However, the TDTL response does not converge to zero error in the steady state due to the limitations of the first order loop. The limitation is an inherent feature of the first-order TDTL because the filter block in Figure 1 is only a gain block of G. The response of the system to a positive frequency step is shown in Figure 7. The error signal or the system response is taken from the output of the phase error detector. The result of applying a negative frequency step is shown in Figure 8. In this case W=1.42, but the system manages to stabilize and converge to none-zero steady state error.

5.2. Fractional-N TDTL Frequency Synthesizer

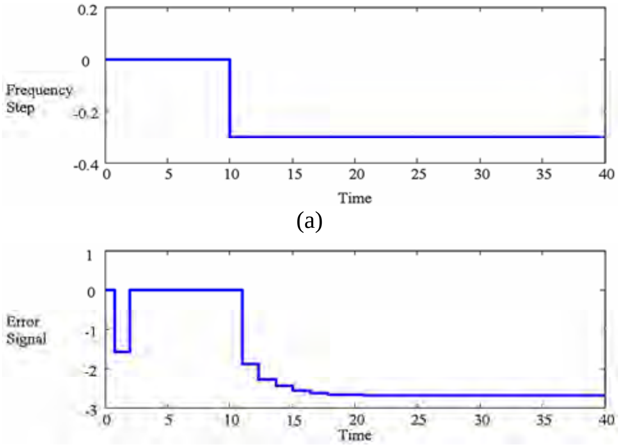
The performance of the fractional-N TDTL frequency synthesizer architecture with register based adaptation mechanism, as described in Section 3, was evaluated through an extensive set of tests. The tests were primarily concerned with assessing the ability of the TDTL-FFS, which includes the composite divider block, to remain in lock following the division of the DCO output frequency and achieve various frequency division ratios.



(a)

(b)

Figure 7. TDTL response to a positive frequency step of 0.4. (a) Frequency step input; (b) Output of the phase error detector.



(a)

(b)

Figure 8. TDTL response to a negative frequency step. (a) Frequency step input; (b) Output of the phase error detector.

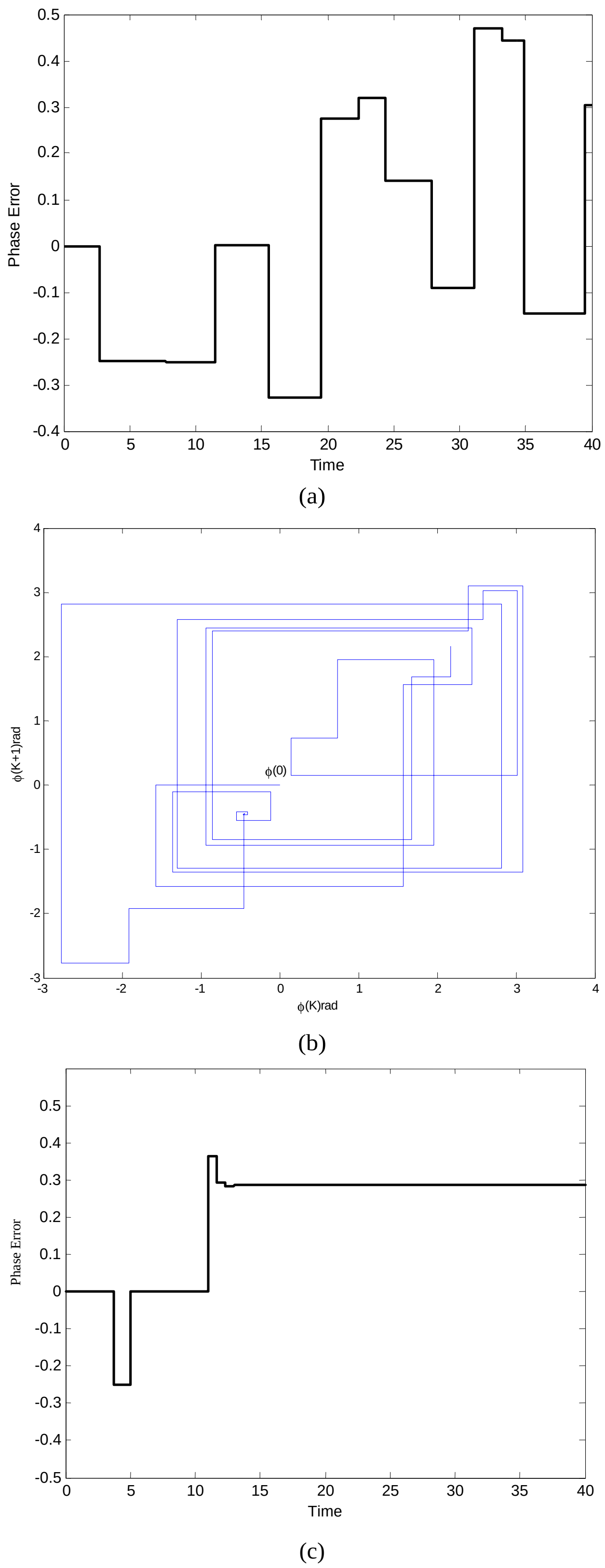
The extensive tests conducted on the TDTL-FFS indicate that it performs very well with respect to the evaluation criterions stated above. This section details some of the results that were achieved. In all the test cases it is assumed that the basic TDTL parameters were selected so that it operates at the optimum point within the locking range of Figure 4 prior to the activation of the composite divider block that converter the system to a TDTL-FFS. It is also assumed that sudden changes in the input signal frequency are not severe so as to drive the system out of lock. These assumptions are applied in order to focus on proving that the new architecture is capable of performing fractional frequency synthesis. TDTL system architectures that deal with wide variations in the input signal frequency through extended locking range are discussed in [13,14].

The importance and effectiveness of the register based adaptation mechanism of the composite divider block of the TDTL-FFS is illustrated in Figure 9. In this case the TDTL-FFS was subjected to the positive frequency step in Figure 7(a) and a pre-scaler divider block with a division ration of 4 and no adaptation was included at the DCO output. Figure 9(a) shows that the system lost its locked state. This is further illustrated by the phase plane plot depicted in Figure 9(b). Replacing the divider block with one that also includes the register adaptation elements enabled the TDTL-FFS system to regain its locked for the same positive step frequency input as shown in Figure 9(c).

The effect of a division ration of 3.5 and a positive step input, the same one in Figure 7(a), is illustrated in Figure 10. The figure shows the DCO output as well as the divider output which clearly indicates the impact of the division.

The response of the TDTL-FFS to a negative frequency step and a division ratio of 3.8 are illustrated in Figure 11 and Figure 12 respectively. The output of the phase error detector in Figure 11 indicates that the system regain locking following the application of the negative step within a relatively small number of samples. The fractional divider output in Figure 12 indicates a frequency that is 3.8 with respect to that of the DCO output.

In many applications, such as adjacent cells of a wireless network, fine divisions are required. The ability of the TDTL-FFS to achieve such fine fractional division ratios is demonstrated in Figure 13. The division ratio at output of the divider block with respect to the DCO output is 2.0714285. The TDTL-FFS was also tested for other division factors and input frequency steps. The system behaved consistently by achieving the required divisions provided that the assumptions stated earlier are maintained.

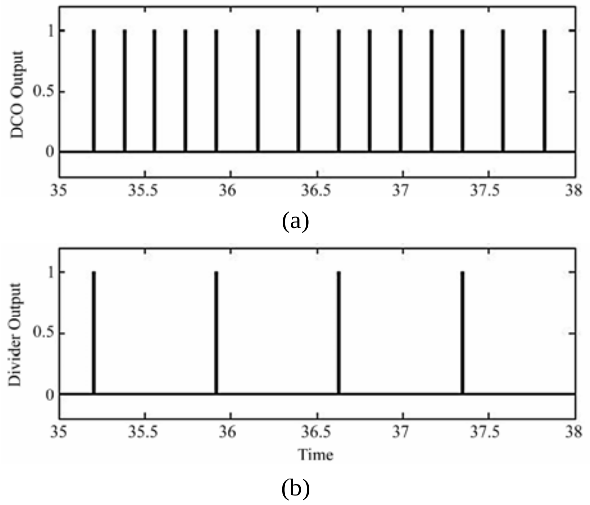


(a)

(b)

(c)

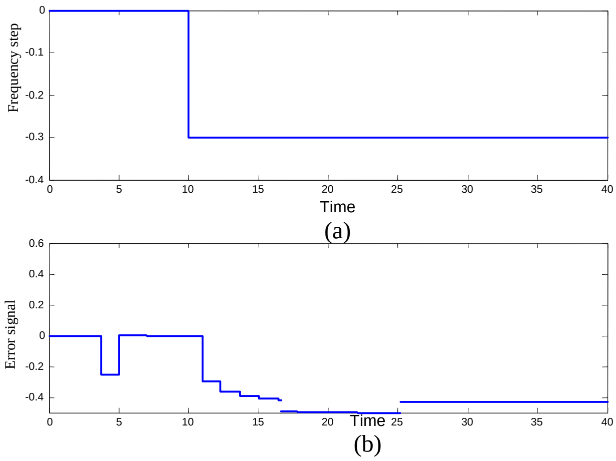
Figure 9. TDTL-FFS response to a positive step. (a) Without adaptation; (b) Phase plane without adaptation; (c) Response with adaptation.



(a)

(b)

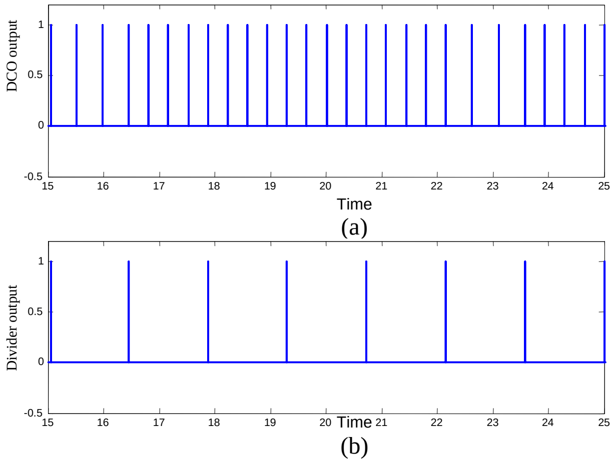
Figure 10. TDTL-FFS outputs for 3.5 division factor. (a) DCO output; (b) Frequency divider output.



(a)

(b)

Figure 11. TDTL-FFS system response to a negative frequency step with RBA. (a) Frequency step input; (b) Phase error detector output.



(a)

(b)

Figure 12. TDTL-FFS outputs for 3.8 division factor. (a) DCO output; (b) Frequency divider output.

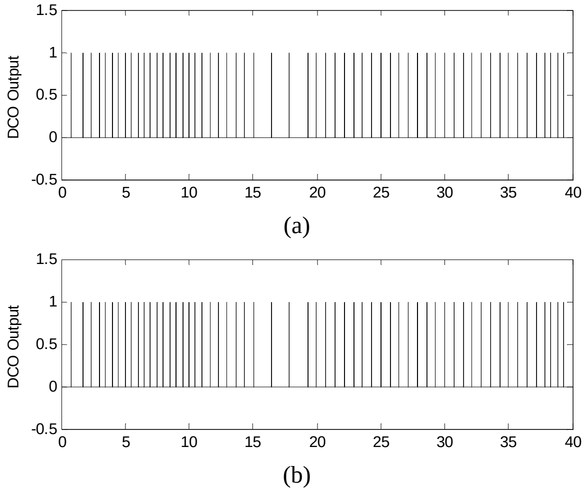


Figure 13. TDTL-FFS outputs for 2.0714285 division factor. (a) DCO output; (b) Frequency divider output.

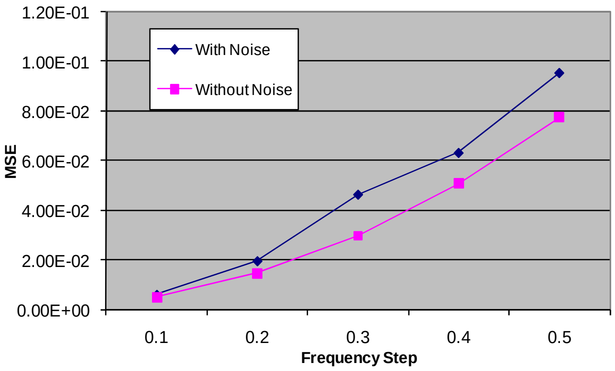


Figure 14. Relationship between frequency step and MSE.

6. TDTL-FFS Speed and Noise Performance

Section 5 above demonstrated the ability of the TDTL-FFS to perform fractional division while maintain the state of locking. Extensive set of test showed that the TDTL-FFS is able to stay in lock when the input frequency varies over a wide range. This is a highly desirable feature of frequency synthesizers as it gives the device the ability to support a variety of wireless communication standards that will invariably operate at different frequencies.

The simulation results of the TDTL-FFS also show clearly that the device support fast switching. The various plots of the synthesizer response, for positive as well as negative input frequency steps, show that the system settles to steady state within a few samples. This was achieved for a variety of division factors including very fine ones. This characteristic of the TDTL-FFS makes it useful in wireless networks with frequency hopping requirements. The performance of the TDTL-FFS under noise conditions was also evaluated. This was done by measuring the mean square error (MSE) for various frequency steps under both noise free and noisy conditions. The plots in Figure 14 illustrate the resilience of the TDTL-FFS to noise. The system kept its locked state and the increase in MSE as a result of the injection of noise is acceptable. A similar study was also conducted to evaluate the effect of noise at different division ratios. The results also indicated that the TDTL-FFS performs very well under noisy conditions.

7. Conclusions

This paper presented a novel fractional frequency synthesizer architecture based on the TDTL. A major challenged faced in utilizing the TDTL for frequency synthesis is keeping the system in locked state following the division process. This problem was overcome by including an adaptation block that compensated for the effects the division had on the loop. The compensation mechanism is an efficient one as it has low complexity and enables fast locking.

The composite adaptor divider block in the TDTL-FFS consisted of a pre-scaler divider and a register based adaptation mechanism. The registers utilize the division information to force the complete system to operate within the locking region by controlling the loop filter gain and the DCO.

The results showed that the TDTL-FFS is capable of performing fractional-N divisions with fine resolution. It can also deal with both positive and negative frequency steps over a wide range of frequencies. The TDTL-FFS has fast switching capabilities and high resilience to noise.

8. References

[1] R. Staszewski and P. Balsara, “All-digital frequency synthesizer in deep-submicron CMOS,” Wiley, 2006.

[2] A. Lacaita, S. Levantino, and C. Samori, “Integrated frequency synthesizers for wireless systems,” Cambridge University Press, 2007.

[3] C. Mishra, *et al.*, “Frequency planning and synthesizer architectures for multiband OFDM UWB radios,” IEEE Transactions on Microwave Theory and Techniques, Vol. 53, pp. 3744–3756, December 2005.

[4] C. Gaudes, M. Valkama, and M. Renfors, “A novel frequency synthesizer concept for wireless communications,” Proceings of International Symposium Circuits and Systems (ISCAS), Vol. 2, pp. 85–88, 2003.

[5] T. Lin, J. Kaiser, and J. Pottie, “Integrated low-power communication system design for wireless sensor networks,” IEEE Communication Magazine, Vol. 42, pp. 142–150, December 2004.

[6] F. Agnelli, *et al.*, “Wireless multi-standard terminals:

System analysis and design of a reconfigurable RF front-end," IEEE Circuits and Systems Magazine, Vol. 6, pp. 38–59, 2006.

[7] A. Chenakin, "Frequency synthesis: Current solutions and new trends," Microwave Journal, Vol. 50, pp. 256–260, May 2007.

[8] S. Moon, A. Valero-Lopez, and E. Sanchez-Sinencio, "Fully integrated frequency synthesizer: A tutrial," International Journal High Speed Electronics and Systems, Vol. 15, pp. 353–375, June 2005.

[9] J. Vankka, "Digital synthesizers and transmitters for software radio," Springer, 2005

[10] A. Bellaouar, M. O'brecht, M. Fahim, and M. Elmasry, "Lowpower direct digital frequency synthesis for wireless communications," IEEE Journal of Solid-State Circuits, Vol. 35, pp. 385–390, March 2000.

[11] Z. Hussain, B. Boashash, M. Hassan-Ali, and S. Al-Araji, "A time-delay digital tanlock loop," IEEE Transactions on Signal Processing, Vol. 49, No. 8, pp. 1808–1815, 2001.

[12] C. Lee and C. K. Un, "Performance analysis of digital tanlock loop," IEEE Transactions on Communications, Vol. COM-30, pp. 2398–2411, October 1982.

[13] S. R. Al-Araji, Z. M. Hussain, and M. A. Al-Qutayri, "Digital phase lock loops: Architectures and applications," Springer, 2006.

[14] M. Al-Qutayri, S. Al-Araji, and N. Al-Moosa, "Improved first-order time-delay tanlock loop architectures," IEEE Transactions on Circuits and Systems Part-I, Vol. 53, No. 9, pp. 1896–1908, 2006.

[15] A. Granas and J. Dugundji, "Fixed point theorem," Springer, 2003.

[16] A. Al-Humaidan, S. Al-Araji, and M. Al-Qutayri, "Frequency synthesizer for wireless applications using TDTL," IEEE APCCAS, pp. 1518–1521, December 2006.

Int. J. Communications, Network and System Sciences, 2009, 2, 888-894
doi:10.4236/ijcns.2009.29103 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Optimizing WiMAX: A Dynamic Strategy for Reallocation of Underutilized Downlink Sub-Frame to Uplink in TDD Mode

Abdul Qadir ANSARI[1], Abdul Qadeer K. RAJPUT[2], Adnan Ashraf ARAIN[2], Manzoor HASHMANI[2]
[1]*Wireless Core Network, Pakistan Telecommunication Company Limited, Pakistan*
[2]*CREST-Research Group, IICT, Mehran University of Engineering and Technology, Jamshoro, Pakistan*
E-mail: *qadir.ansari@ptcl.net.pk, aqkrajput@muet.edu.pk, adnanlooking@ieee.org, mhashmani@yahoo.com*
Received August 20, 2009; *revised September* 22, 2009; *accepted November* 1, 2009

Abstract

WiMAX networks experience sporadic congestion on uplink when applications running at subscriber stations need more bandwidth to transmit than allocated. With the fast proliferation of mobile Internet, the wireless community has been looking for a framework that can address the issue of impediment on uplink. Due to asymmetric behavior of Internet applications downlink sub-frame is expected to have longer duration as compared to uplink. According to IEEE 806.16 standard for WiMAX the segmentation of TDD frame between uplink and downlink can be dynamically redefined even at runtime. Research contributions so far lack in addressing an optimal strategy for readjustment of uplink and downlink sub-frame boundaries; based on traffic statistics. In this paper, we introduce a mechanism that allows uplink sub-frame to grow, borrowing resources from the downlink sub-frame, if the uplink utilization is high and the downlink is being underutilized. We present here, a framework to dynamically demarcate the TDD frame-duration between uplink and downlink. Proposed algorithm takes into account the present utilization of downlink and reallocates a certain quantum of free resources to uplink. This occurs when uplink observes shortage of bandwidth to transmit. We simulate some test scenarios using OPNET Modeler with and without dynamic reallocation capability. The results of our simulation confirm the effectiveness of proposed algorithm which observes a remarkable decrease in end-to-end packet delay. Also, we observe an improvement in throughput at uplink such that, the performance of downlink remains unaffected.

Keywords: WiMAX, Time Division Duplex (TDD), Quality of Service (QoS), End-to-End Packet Delay, Network Throughput

1. Introduction

The IEEE 802.16 family of standards specifies the air interface of fixed and mobile broadband wireless access (BWA) systems that support multimedia services. The IEEE 802.16-2004 standard, which was previously called 802.16d or 802.16-REVd, was published for fixed access in October 2004. Good reviews of the standard can be found in [1–3]. The standard has been updated and extended to the 802.16e standard for mobile access, Mobile WiMAX, as of October 2005 [4]. Mobile WiMAX is a broadband wireless solution that enables convergence of mobile and fixed broadband networks. Mobile WiMAX technology is designed to be able to scale to work in different channelizations from 1.25 to 20 MHz to comply with varied requirements.

The fundamental premise of the IEEE 802.16e MAC architecture is QoS on the move. With fast air interface, asymmetric downlink/uplink configuration capability, fine resource granularity and a flexible resource allocation mechanism, Mobile WiMAX can meet QoS requirements for a wide range of data services and applications [5].

IEEE 802.16e standard includes QoS support framework; however, it left undefined the details to ensure QoS guarantees, scheduling algorithms, uplink (UL) and downlink (DL) sub-frame allocation; for vendors as a motivation to device effective scheduling and resource allocation mechanisms to deliver QoS guarantees, especially for the real-time traffic.

WiMAX networks support two types of duplexing modes to separate UL and DL communication; i.e. Time Division Duplex (TDD) and Frequency Division Duplex (FDD). In this paper we have focused on TDD mode where both UL and DL share same frequency and to separate downlink and uplink, time division multiple access (TDMA) is used.

The duration of DL and UL sub-frames may be decided once based on average traffic statistic expectations. However, it is also possible to tune the network configuration through real-time monitoring, and may readjust the uplink and downlink boundaries.

According to the standard, this segmentation can be dynamically adjusted even at runtime. Unfortunately, research contributions so far lacks in addressing an optimal strategy towards readjustment of UL and DL boundaries dynamically; while keeping the current traffic statistics in account. It is important to remember that asymmetric behavior of Internet applications intuitively ask for more duration of DL sub-frame as compared to UL. Also DL traffic behaviors could be controlled at serving Base Station (BS); but it is not true for UL.

We have introduced a mechanism in order to allow the UL sub-frame to "grow", borrowing resources from the DL sub-frame, if the UL utilization is high and the DL utilization is low. Our strategy is to keep the performance graph of downlink traffic unaffected by monitoring the downlink utilization and requirement. Moreover resources borrowed from DL will be relinquished as and when required at DL. The mechanism is tested in a controlled environment for it effectiveness. Simulation results confirmed the positive impact of this new capability on throughput and packet end-to-end delay on UL.

2. WiMAX-Time Division Duplex

The IEEE 802.16e-2005 supports both time division duplexing (TDD) and frequency division duplexing (FDD) modes. In TDD mode, the uplink and downlink transmission share the same frequency but do not transmit simultaneously. The frame, in Figure 1, is flexibly divided into a downlink sub-frame and an uplink sub-frame. The downlink sub-frame used to transmit data from a BS to SS. The uplink sub-frame carries SS traffic to the BS. The sub-frame is divided into mini slots, which is the minimum unit of data transfer in this level. Frames are broadcasted and during the downlink sub-frame, the SS picks up the data addressed to it. Media Access Protocol (MAP) messages are broadcasted at the beginning of each downlink sub-frame. There are two types of MAP messages, DL-MAP and UL-MAP. The DL-MAP described the usage of the downlink sub-frame whereas the UP-MAP tells which mini slot and how many mini slots are allocated to the specified SS for its transmission during the uplink sub-frame.

Mobile WiMAX profiles only consider the TDD mode of operation for the following reasons:

1) It allows dynamic reallocation of DL and UL radio resources to effectively support asymmetric traffic pattern that is common in Internet applications.

2) The allocation of radio resources in DL and UL is determined by the DL/UL switching point(s).

3) Both DL and UL are in the same frequency channel to yield better channel reciprocity and to better support link adaptation.

4) A single frequency channel in DL and UL can provide more flexibility for spectrum allocation.

As shown in Figure 2, the connection between a SS and BS is identified by a unique connection identifier (CID). One CID can correspond to an individual application or a number of applicants bundled together such as a group of users in the same building. CID also specifies polling schemes provided to the connection by the BS, which will result in QoS for the connection who owns this CID.

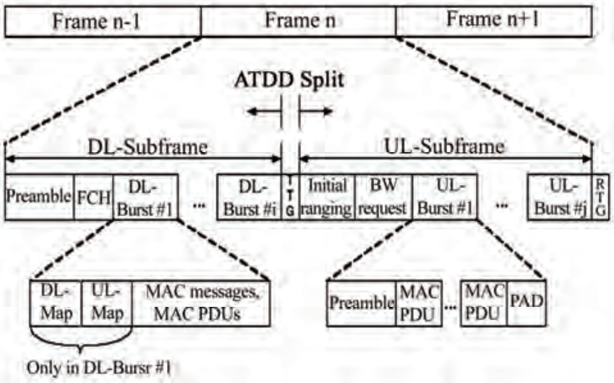


Figure 1. WIMAX TDD frame.

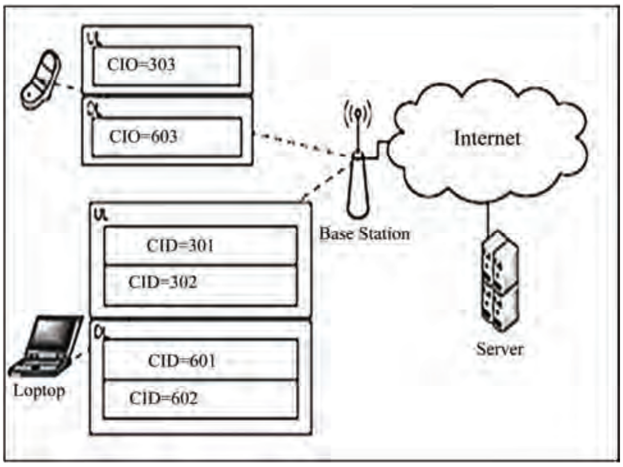


Figure 2. Connection (CID)-based WiMAX MAC layer [6].

3. Quality of Service Support in WiMAX

WiMAX with provisioned quality of service (QoS) for digital multimedia applications to mobile end users over wide area networks is the new frontier of telecommunications industry.

Before providing a certain type of a service, the base station and user-terminal first establish a unidirectional logical link between the peer MACs called a connection. The outbound MAC then associates packets traversing the MAC interface into a service flow to be delivered over the connection. The QoS parameters associated with the service flow define the transmission ordering and scheduling on the air interface. The connection-oriented QoS therefore, can provide accurate control over the air interface. Since the air interface is usually the bottleneck, the connection-oriented QoS can effectively enable the end-to-end QoS control. The service flow based QoS mechanism applies to both DL and UL to provide improved QoS in both directions.

The service flow parameters can be dynamically managed through MAC messages to accommodate the dynamic service demand. IEEE 802.16 MAC is connection-oriented. BS controls the access to the medium, bandwidth is granted to SS on demand. At the beginning of each frame, the BS schedules the uplink and downlink grants to meet the negotiated QoS requirements. Each SS learns the boundaries of its allocation under current uplink sub-frame via the UL-MAP message. The DL-MAP delivers the timetable of downlink grants in the downlink sub-frame [7].

4. Related Research Review

During last couple of years, many proposals for QoS service support in WiMAX networks were published [8–14]. Most of them are the solutions on the bandwidth allocation (with preset allocation of UL and DL sub-frame size [13]), flow scheduling and Adaptive Modulation and Coding Schemas [14] to optimize the performance of WiMAX network. In [15] a three-tier QoS framework is introduced where a Pre-scale Dynamic Resource Reservation (PDRR) is proposed to allocate frame bandwidth to UL sub-frame and DL sub-frame with pre-scaled bounds. In [16] the presented baseline network model is examined for fixed and dynamic reallocation, but under dynamic reallocation the resources are relinquished completely from UL and allocated back to DL when a certain preset threshold for DL occupancy is met. We have modified the algorithm and proposed a step-by-step allocation and similarly a step-by-step deallocation of UL resources back when certain threshold of DL occupancy is met.

5. Proposed Strategy to Enable Reallocation Capability

We provide here a basic mechanism that allows uplink sub-frame to grow, borrowing resources from the DL sub-frame, if the uplink utilization is high and the downlink is underutilized. Our strategy ensures that resources borrowed should be relinquished as and when required by DL in order to ensure that with this introduced capability the DL performance should not be affected. Our strategy is to observe the utilization of DL resources and if DL is underutilized and UL is starving for bandwidth; DL sub-frame duration may be reduced and UL sub-frame duration may be increased by a certain quantum of time.

5.1. The Proposed Algorithm

1) Baseline network scenario is simulated using OPNET Modeler 14.5 (Wireless Suite) to test the impact of new capability to redefine the boundaries of UL and DL sub-frames.

2) Observe the frame allocation information and the traffic behavior prior to the addition of the new capability. For example; a) Observe sub-frame utilization, and b) Application Performance in terms of throughput and end-to-end packet delay on UL

3) Implement a mechanism to change the uplink and downlink partition dynamically.

4) Implement an algorithm that performs re-allocation of the sub-frames.

5) Ensure the performance of DL traffic should remain unaffected with introduction of new capability.

6) Observe and analyze the results with the new capability.

7) Compare, analyze and summarize the simulation results with the baseline network results.

5.2. Flowchart of our Proposed Algorithm

Here, we set certain thresholds for UL and DL sub-frame utilization to decide whether or not the sub-frame duration need to be dynamically adjusted. The same is shown here, in the flowchart of our proposed algorithm in Figure 3.

The above flow chart shows how the evaluation of the sub-frames utilization is used to determine whether increase the UL sub-frame size or prohibit any readjustment of sub-frame duration.

5.3. Pseudo Code of Proposed Algorithm

```
1) Check DL Sub-Frame Utilization.
    a) Is DL already reduced
```

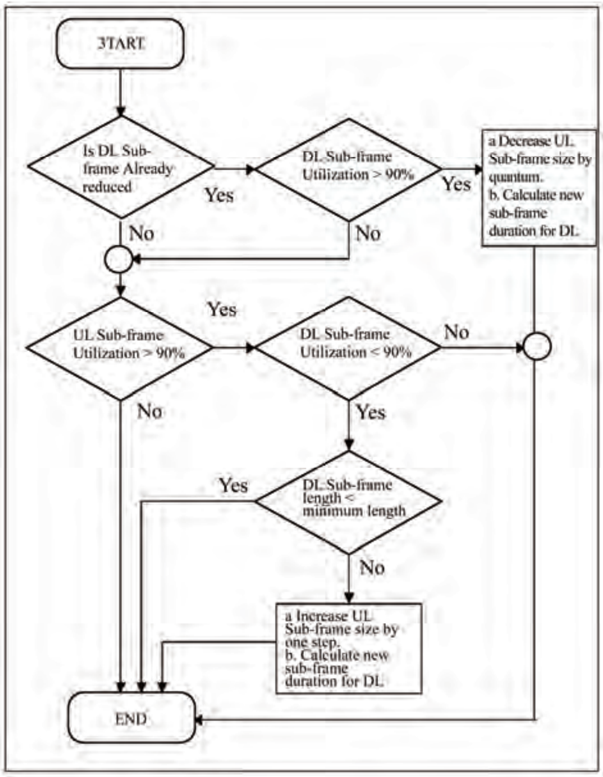

START
Is DL Sub-frame Already reduced
Yes
DL Sub-frame Utilization > 90%
Yes
a Decrease UL Sub-frame size by quantum.
b. Calculate new sub-frame duration for DL
No
No
UL Sub-frame Utilization > 90%
Yes
DL Sub-frame Utilization < 90%
No
No
Yes
Yes
DL Sub-frame length < minimum length
No
a Increase UL Sub-frame size by one step.
b. Calculate new sub-frame duration for DL
END

Figure 3. Proposed scheme- flow chart.

b) Is DL utilization is below threshold

c) If DL is already reduced but utilization is still under threshold go to Step 2.

d) Else revert back last reduction in DL

e) Reset the UL and DL sub-fame boundaries to the previous values for each respectively and go to Step 4.

2) Check UL Sub-Frame Utilization.

a) Is UL sub-frame utilization above threshold?

b) Is DL under utilized?

c) If both conditions a and b are TRUE then check condition in d.

d) Check DL sub-frame length < minimum allowable length

e) If condition fails in d then go to Step 4.

f) Else go to Step 3.

3) Increase UL Frame Size by One Step.

a. Update UL and DL sub-fame boundaries

4) EXIT.

6. Our Proposed Setup for Simulation

The baseline network is composed of one WiMAX cell with four SS nodes. All SS nodes have an uplink application load of 250 Kbps for a total of 1 Mbps. At specific times, the Server generates 600 Kbps of application traffic directed to SS-0 and SS-1; this creates a total downlink application load of 1.2 Mbps. The cell uses Scalable OFDMA frame with 512 sub-carriers and of 5 milliseconds duration. Uplink sub-frame is set to 12 symbol times (i.e. frame columns). Downlink sub-frame is assigned 34 symbol times. For QPSK ½, the capacity expected is Uplink: ~ 0.6 Mbps. Downlink: ~2.5 Mbps.

6.1. Preset Sub-frame Allocation between UL and DL

6.1.1. DL Traffic Behavior

Following graph (Figure 4) shows MAC load and throughput for DL. It can be easily observed that DL MAC load and throughput is same i.e. 1.2 Mbps. Total DL capacity is 2.5 Mbps; thus there is enough capacity at DL to successfully transport the DL load.

Moreover the ETE packet delay remains between 6~10 milliseconds which is fairly under acceptable range, as shown in Figure 5.

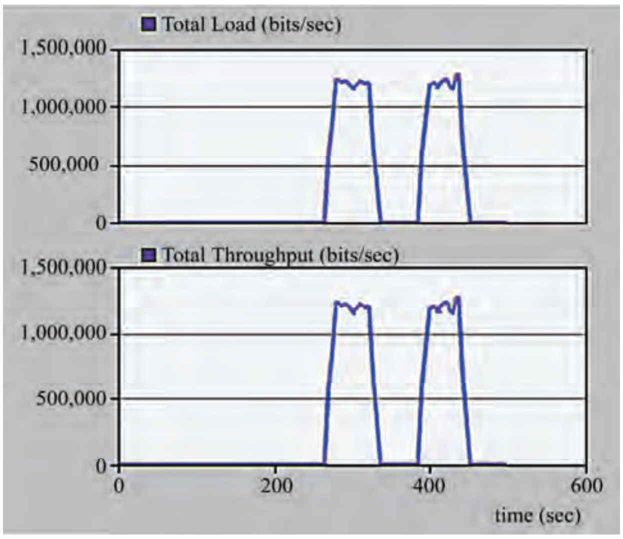


Figure 4. DL traffic load and throughput.

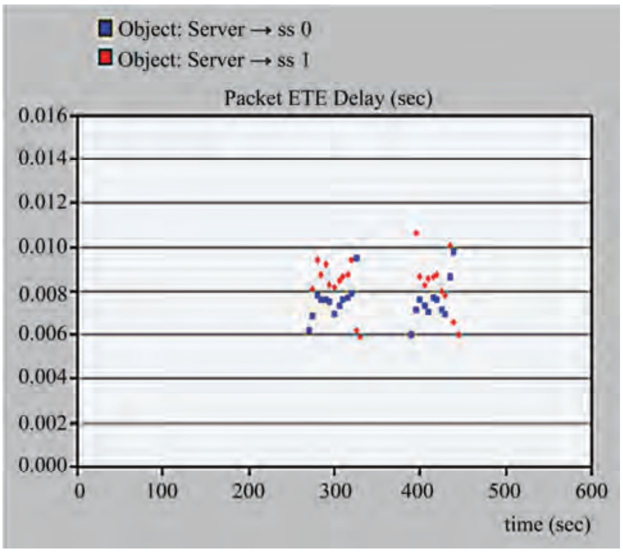


Figure 5. End-to-end packet delay (DL).

6.1.2. UL Traffic Behavior

Following graph (in Figure 6) shows MAC load and throughput for UL. We observe that UL MAC load is ~ 1 Mbps and throughput is 0.56 Mbps. UL capacity is 0.6 Mbps; thus the offered load is exceeding the capacity; which results in high application delays ranging from 3.5 to 3.75 seconds (Figure 7). Thus UL is running out of resources and do not have enough capacity to accommodate any further load.

6.1.3. Statistics of Usage and Usable Sizes (UL and DL)

Statistics results are also collected, in Figure 8, for data burst percentage utilization and sub-frame usable size for UL and DL.

From above diagram it is evident that DL usage is about 60%; however UL usage is 100%. Also usable sub-frame size for UL is merely ~3.3 K symbols and same for DL is nearly 11.3 K Symbols.

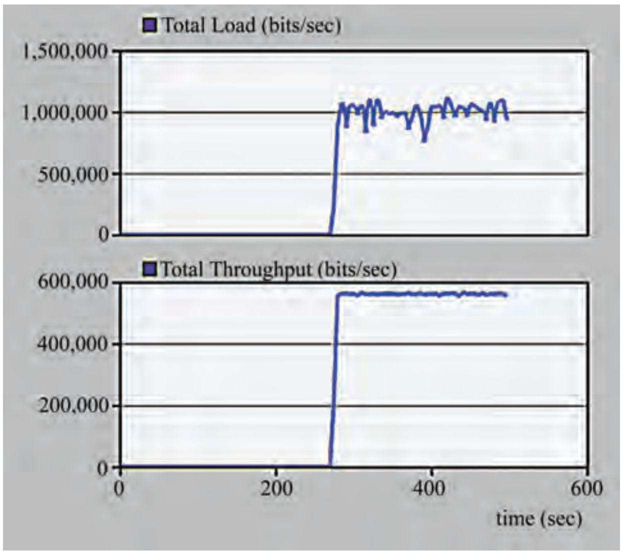


Figure 6. UL traffic load and throughput.

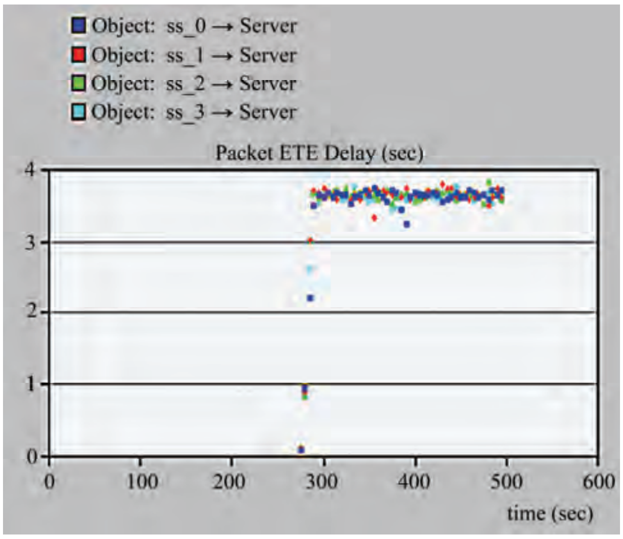


Figure 7. End-to-end packet delay (UL).

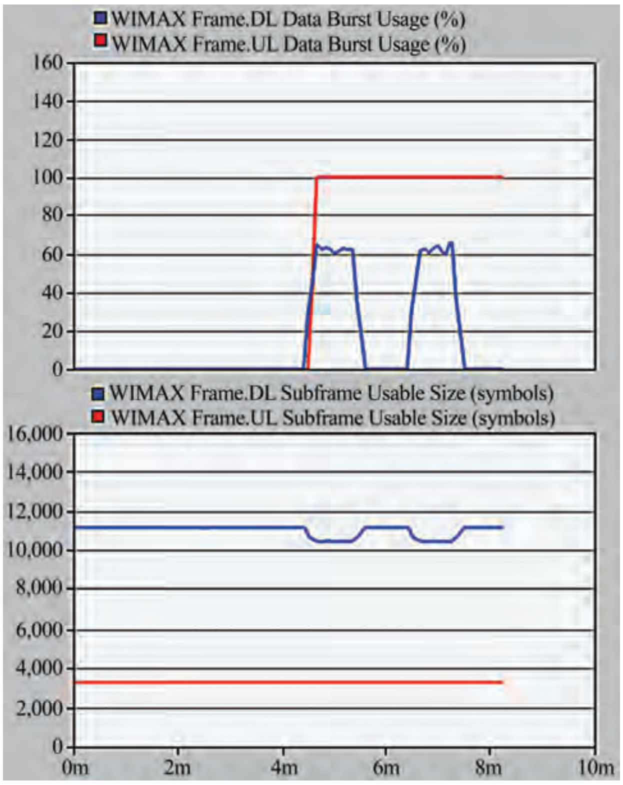


Figure 8. Data burst usage and sub-frame usable size.

6.2. Dynamic Sub-Frame Allocation between UL and DL

6.2.1. DL Traffic Behavior

The performance of DL traffic remained unaffected as desired; i.e. 1.2 Mbps MAC load and throughput (Figure 9). Comparison could also be found in second frame. Application end-to-end packet delay is observed to remain under ~8 milliseconds (Figure 10).

6.2.2. UL Traffic Behavior

Here we can confirm the major improvement in throughput at UL. Comparison is presented between constant and adaptive schemas. It is evident that load and throughput are almost aligned on UL as well (Figure 11). Now UL transports almost all the traffic originated form MS and directed to BS.

This improvement has also resulted in remarkable decrease in packet end-to-end delay on UL; which has now reduced to ~7 milliseconds from 3~4 seconds (Figure 12).

6.2.3. Statistics of Usage and Usable Sizes (UL and DL)

From UL and DL data burst usage and Usable size (Fig-

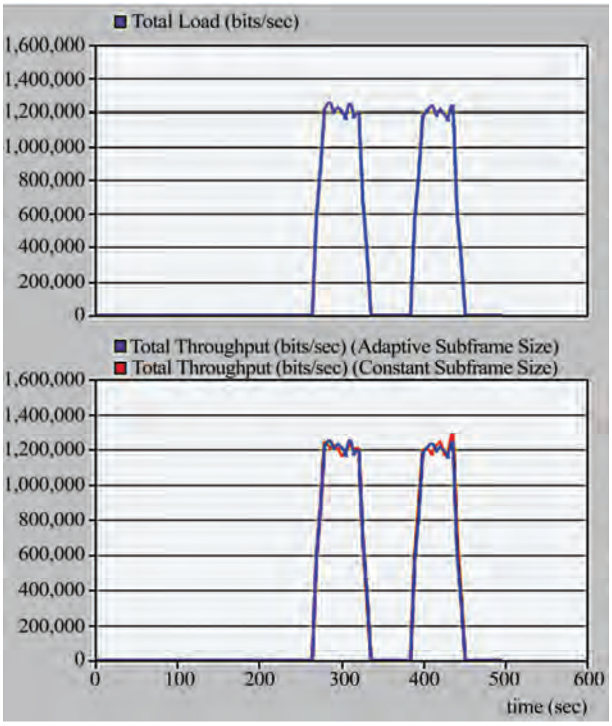


Figure 9. DL traffic load and throughput.

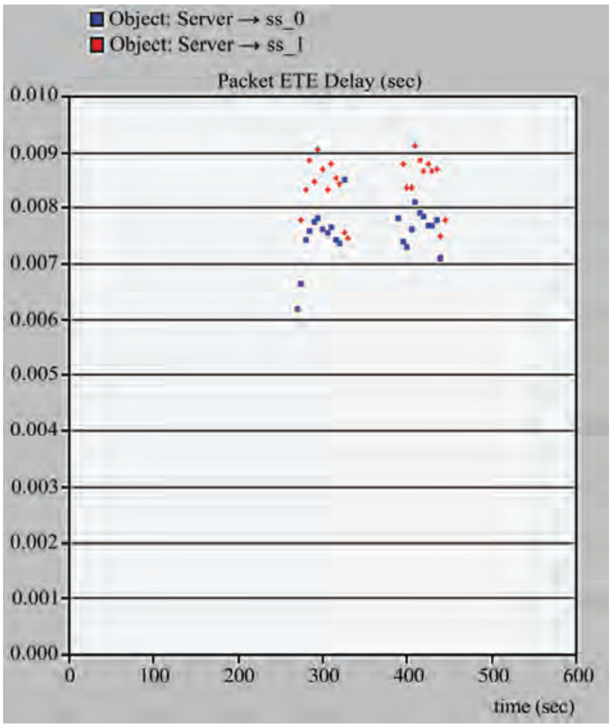


Figure 10. End -to -end packet delay (adaptive).

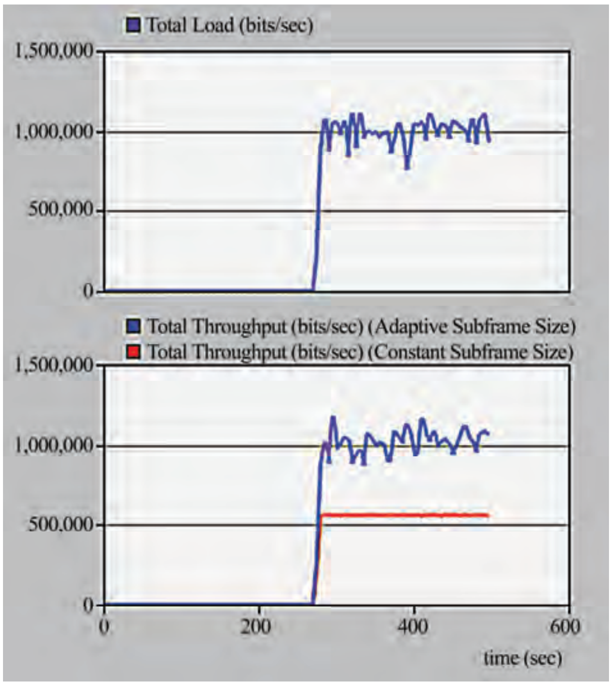


Figure 11. UL traffic load and throughput.

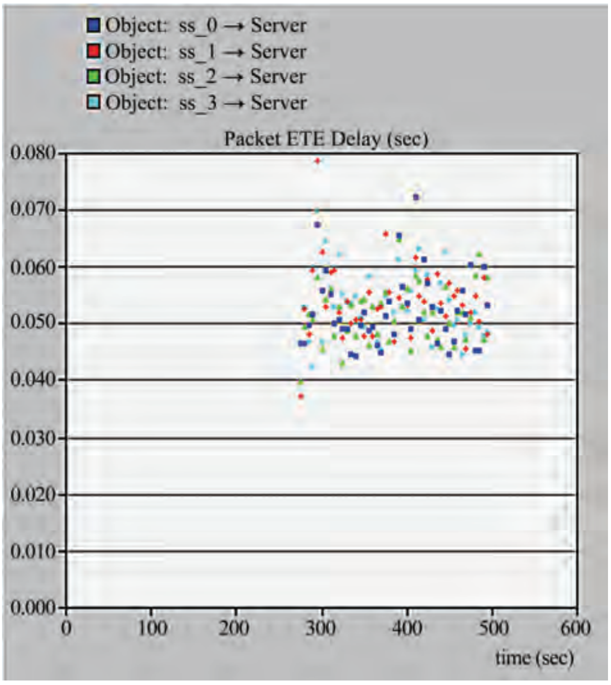


Figure 12. End-to-end packet delay (UL).

ure 13) statistics it is evident that our scheme worked well and effectively utilized the free DL bandwidth at UL. With our strategy UL usage has improved because the UL usable capacity increases to maximum when DL utilization is lowest.

7. Conclusions

This paper presents a dynamic strategy that takes advantage of underutilized DL sub-frame, allowing the UL sub-frame to acquire temporarily free resources of DL

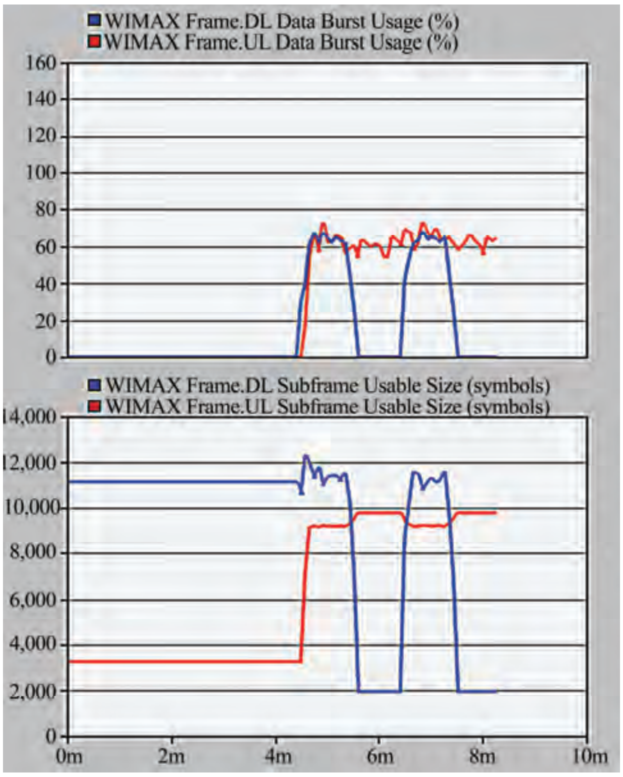


Figure 13. Data burst usage and sub-frame usable size.

when UL observes shortage of resources. Our proposed mechanism ensures the performance of DL to remain unaffected by continuous monitoring of the DL utilization. In case of requirements, the sub-frame boundaries could be readjusted and DL will be prioritized. The proposed mechanism is tested under controlled environment using OPNET Modeler for correctness and effectiveness.

In our proposed work, an observable improvement is seen in both throughput and end-to-end packet delay at UL without affecting the performance of DL.

8. References

[1] C. Eklund, *et al.*, "IEEE standard 802.16: A technical overview of the wireless MAN air interface for broadband wireless access," IEEE Communication Magazines, pp. 98–107, June 2002.

[2] A. Ghosh, *et al.*, "Broadband wireless access with WiMAX/802.16: Current performance benchmarks and future potential," IEEE Communication Magazine, pp. 129–36, February 2005.

[3] 802.16d Task Group, "IEEE standard for local and metropolitan area networks part 16: Air interface for fixed broadband wireless access systems," IEEE, IEEE802-16-2004 Version, 802.16d or Fixed WMAN, June 2004.

[4] 802.16e Task Group, "IEEE standard for local and metropolitan area networks part 16: Air interface for mobile broadband wireless access systems," IEEE, Active International Standard, 802.11e or Mobile WMAX, 2005.

[5] WiMAX Forum: Mobile WiMAX – Part I: A Technical Overview and Performance Evaluation, June 2006.

[6] X. Yang, M. Venkatachalam, and S. Mohanty, "Exploiting the MAC layer flexibility of WiMAX to systematically enhance TCP performance," IEEE Mobile WiMAX Symposium, 2007.

[7] K. C. Chen and J. Roberto B. de Marca, "Mobile WiMAX," IEEE PRESS, IEEE Communications Society, Sponsor, John Wiley & Sons, Ltd, April 2008.

[8] J. Sun, Y. Yao, and H. Zhu, "Quality of service scheduling for 802.16d broadband wireless access systems," Proceedings of IEEE 63rd International Conference of Vehicular Technology 2006, VTC 2006-Spring, Vol. 3, pp. 1221–1225, 2006.

[9] G. Chu, D. Wang, and S. Mei, "A QoS architecture for the MAC protocol of IEEE 802.16d BWA system," Proceedings of IEEE International Conference on Communications, Circuits and Systems, Vol. 1, pp. 435–439, June 2002.

[10] K. Wongthavarawatn and A. Ganz, "Packet scheduling for QoS support in IEEE 802.16d broadband wireless access systems," International Journal of Communication Systems, Vol. 16, pp. 81–96, 2003.

[11] J. Chen, W. Jiao, and H. Wang, "A service flow management strategy for IEEE 802.16d broadband wireless access systems in TDD mode," Proceedings of IEEE International Conference on Communications 2005, Vol. 5, pp. 3422–3426, May 2005.

[12] R. Iyengar, P. Iyer, and B. Sikdar. "Delay analysis of 802.16 based last mile wireless networks," IEEE Globecom, USA, 2005.

[13] R. Pries, D. Staehle, and D. Marsico, "IEEE 802.16 capacity enhancement using an adaptive TDD split," Vehicular Technology Conference, pp. 1539–1543, May 11–14, 2008.

[14] A. Q. Ansari, A. Q. K. Rajput, and M. Hashmani; "WiMAX network optimization-analyzing effects of adaptive modulation and coding schemes used in conjunction with ARQ and HARQ," Proceedings of 7th International Conference on Communication Networks and Services Research, pp. 6–13, 2009.

[15] M. Ma, J. Lu, S. Kumar, and B. Chong, "A three-tier framework and scheduling to support QoS service in WiMAX," Proceedings of ICICS, 2007.

[16] "Understanding WiMAX model internals and interfaces," Opnetwork, 2008.

Int. J. Communications, Network and System Sciences, **2009, 2, 895-902**
doi:10.4236/ijcns.2009.29104 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Balanced Topology NEMO Construction for the Internet-Based MANET

Long-Sheng LI[1], Gwo-Chuan LEE[2], and Li-Keng KANG[1]
[1]*Department of Computer Science and Information Engineering, National Chiayi University, Chiayi, Taiwan, China*
[2]*Department of Computer Science and Information Engineering, National United University, Miaoli, Taiwan, China*
E-mail: {*sheng, s*0950292}*@mail.ncyu.edu.tw, gclee@nuu.edu.tw*

Received September 1, 2009; *revised October* 13, 2009; *accepted November* 21, 2009

ABSTRACT

A mobile ad hoc network (MANET) is a wireless network without any fixed infrastructure. All nodes must communicate with each other by a predefined routing protocol. Most of routing protocols don't consider binding internet addresses to mobile nodes. However, in network mobility (NEMO), all mobile nodes can't only communicate with Internet using Bi-directional tunneling but also can be allocated an Internet address. In this paper, we propose two algorithms with the nested NEMO topology to reconstruct the Internet-based MANET. Additionally, a novel load balancing solution is proposed. The Mobile Router (MR) which acts as a central point of internet attachment for the nodes, and it is likely to be a potential bottleneck because of its limited wireless link capacity. We proposed a load-information in the route advertisement (RA) message. The simulation results show that the proposed solution has significantly improved the connection throughput.

Keywords: MANET (Mobile Ad Hoc Network), NEMO (Network Mobility), RA (Route Advertisement), MR (Mobile Router)

1. Introduction

In recent years, as a result of wireless network development, the number of mobile devices increases, e.g. Notebook, Personal Digital Assistant (PDA) and Cellular Phone etc, the request for ubiquitous Internet access is violently blooming. The user can access the Internet by the mobile devices when they are moving. The present wireless network technology can be divided into two kinds: 1) with the foundation network construction 2) without the foundation network construction. In former, mobile nodes can communicate with each other and access the Internet by connecting to access point (AP) or base station (BS). Compared with on former, the latter is characterized by the lack of any fixed network infrastructure. Mobile ad hoc network (MANET) is one kind of the letter. MANET is composes by a crowd of mobile nodes. Without the support of fixed network infrastructure, every node will play the role of router. In other words, every node in MANET has ability to forward the packets. Any two nodes in communication can transmit the packets in direct or through other nodes in indirect. The nodes in MANET can not access the Internet directly. If the node in MANET wants to communicate with the node in the Internet, the packets sent to the AP or BS, and then into the Internet. As we know, the routing issue in MANET is focus on the packets transmit in the MANET. Many researches discuss the performance about the routing protocol algorithms. But they do not discuss the routing problems between the MANET and Internet.

When more and more electronic devices become be capable of wireless communications, the original static-routing Internet protocol (IP) address is not enough to support the mobility of the devices. The Mobile IP Working Group within the Internet Engineering Task Force (IETF) has proposed the Mobile IP protocol to support the mobility of the devices in IP-based networks [1]. A network which is viewed as a single unit and moved around we call a Mobile Network (MONET). The concept that implements Mobile Network is called Network Mobility (NEMO) [2]. The IETF NEMO working group has been setup in an effort to address issues related to NEMO [3]. The working group has standardized the NEMO Basic Support protocol recently [4]. The NEMO protocol is a way of managing the mobility of an entire network which changes its point to the global Internet [5]. The NEMO protocol is based on mobile IPv6 [6]. Figure shows the basic operation of the NEMO Basic Support protocol. The MONET is composed of one or more mobile routers (MRs) with the communication devices called mobile nodes (MNNs). All MNNs can connect to Internet via an MR. The access router (AR) connects to an MR via a wireless link. MR and its home agent (HA) keep communicating by a bi-directional tunneling. When the MNN moves around, the MR sends a router advertisement (RA) to the MNN periodically. After receiving the RA, MNN gets the care-of-address (CoA) and sends the binding update (BU) message includes the CoA to the HA. CoA is a new and temporary address that is obtained on each visited MONET. After this process, the HA can forward the packets destined for the home address (HoA) of the MR to this CoA. MR has two interfaces, the Egress Interface and Ingress Interface. The MR transmits the packets to MNN via Ingress Interface. On the contrary, MR uses the Egress Interface to forward the packets from MONET.

It is an important issue for the mobile router to guarantee that stable communications and a high data rate. According to the NEMO support protocol, users can dispose their own mobile devices into a mobile network but only the MR needs to consider that the task of mobility management. As we know, the MR is attached to the AR through a wireless interface, and all packets in/out the mobile network must pass through it. So the MR is the single point of failure. However, the NEMO protocol is a hierarchical construction, the ARs and MRs could be the overloaded points with the most traffic load. According to the characteristic of wireless channels, i.e. limited bandwidth and high jitter, we hope that the traffic load of whole system could be balanced between ARs and MRs. To achieve this goal, we propose a dynamic load balancing scheme. We add some information in BU and RA. According to this information, MRs/MNNs can connect to the AR/MR with the less traffic load.

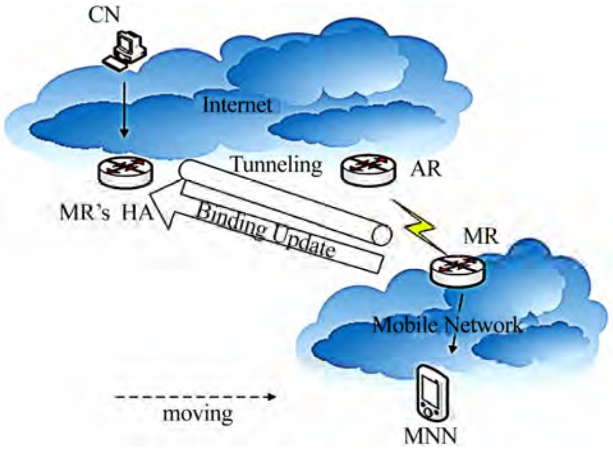


Figure 1. A NEMO architecture.

The rest of the paper is organized ad follows. Section 2 introduces two cluster algorithms in MANET and the load balance index which we use to evaluate our scheme. Section 3 describes two ways to reconstruct MANET to NEMO and a dynamic load balancing scheme. Section 4 shows our simulation results and analyzes performance evaluation. Finally, Section 5, we conclude our results and present some future research issues.

2. Related Work

There are several heuristics algorithms have been proposed to solve ad hoc networks clustering problem. We use Lowest-ID (LID) algorithm [7] and Highest-Degree algorithm (HD) [8] to achieve our goal. In LID, each node has a unique ID. Every node periodically broadcasts its ID through a ‘Hello’ message. By receiving the message, nodes know the neighbor ID. The lowest-ID node in the same neighborhood is selected as a cluster head (CH); nodes which can ‘listen’ two or more CHs become cluster gateway (CG), while other nodes are cluster member (CM). In HD, the highest degree node in a neighborhood is selected as a CH. The degree value means that the number of neighbors. In contrast with LID, HD uses location information for the cluster composition. Like LID, node periodically broadcasts its degree through a ‘Hello’ message.

Another important issue that we consider is load balancing. We use the index β firstly introduced in [9] and used in [10]. Let B_i be the throughput at AR i , then the β is defined as

$$b = \frac{\left(\sum Bi\right)^2}{\left(n\sum Bi^2\right)} \tag{1}$$

where n is the number of neighboring ARs. The index is 1 when all ARs have the same throughput. In this paper, our target is to get maximum index.

3. Proposed Schemes

In this section, we proposed two schemes to reconstruct the nested NEMO from MANET. We adopt two cluster algorithms which are used in MANET, but we do not discuss their drawbacks. We use the two cluster algorithms in the initial construction. When the network topology changes, we adopt the NEMO basic support protocol. However, the NEMO support protocol is a hierarchical architecture; we modify the way of cluster. Figure 2 shows that AR is the CH of Cluster a and Node u is the CH of Cluster b and Node v is the CH of Cluster c. Then, Node u and Node v are both CMs for AR. So the AR is the root node of the tree topology. We construct the network topology in a tree base to avoid the loop route. In our cluster algorithm, the Node u and Node v will be the roles of MR. First, we define the “Level” value. In Figure 3,

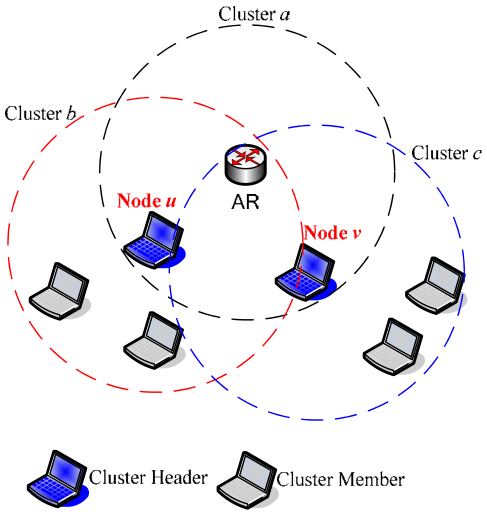


Figure 2. The clustering scheme.

the Level value of the MNN is defined as the shortest hop counts to AR. Level 1 means that the shortest hop counts between node and AR is 1.

3.1. Distributed Scheme (DS)

First, we assume that all nodes have at least two network interfaces. Because every node has the opportunity to be elected as an MR. Therefore, in DS, every node uses the LID algorithm to choose its CH. For example, in Figure 4, node 111 is the node of Level 2. It can 'hear' from node 225 and node 32. Because node 32 is the lowest id for node 111, node 32 is node 111's CH, node 111 is node 32's CM. However, node 200 is the node of Level 4. We will choose the node of Level 3 to be the CH. So, node 1 is node 200's CH, node 200 is node 1's CM. Then, after cluster completes, each node has its own cluster. So every node will be the CH or CM. Now, the network architecture is reconstructed to a tree topology. So the CHs play the role of MRs and the CMs play the role of MNNs. We successfully reconstruct the MANET to NEMO.

3.2. Centralized Scheme (CS)

In DS, each node is unable to know the entire network topology condition; it could unable to select suitable nodes. The construction is independently completes by each Node. In CS, we use the top-down approach. First, in Figure 6, we choose that AR is the Root CH and AR's neighbors are CMs. So node 225, node 32, node 80, node 44 and node 144 are AR's neighbors. Then, AR use HD algorithm to choose the CH of the Level 1. In Figure 7, the node 32 and node 44 are both the CMs of AR and CHs of the Level 2. Additionally, we find the node 25 in the overlapping place between node 32 and node 44. We can not guarantee node 25 is assigned to the most

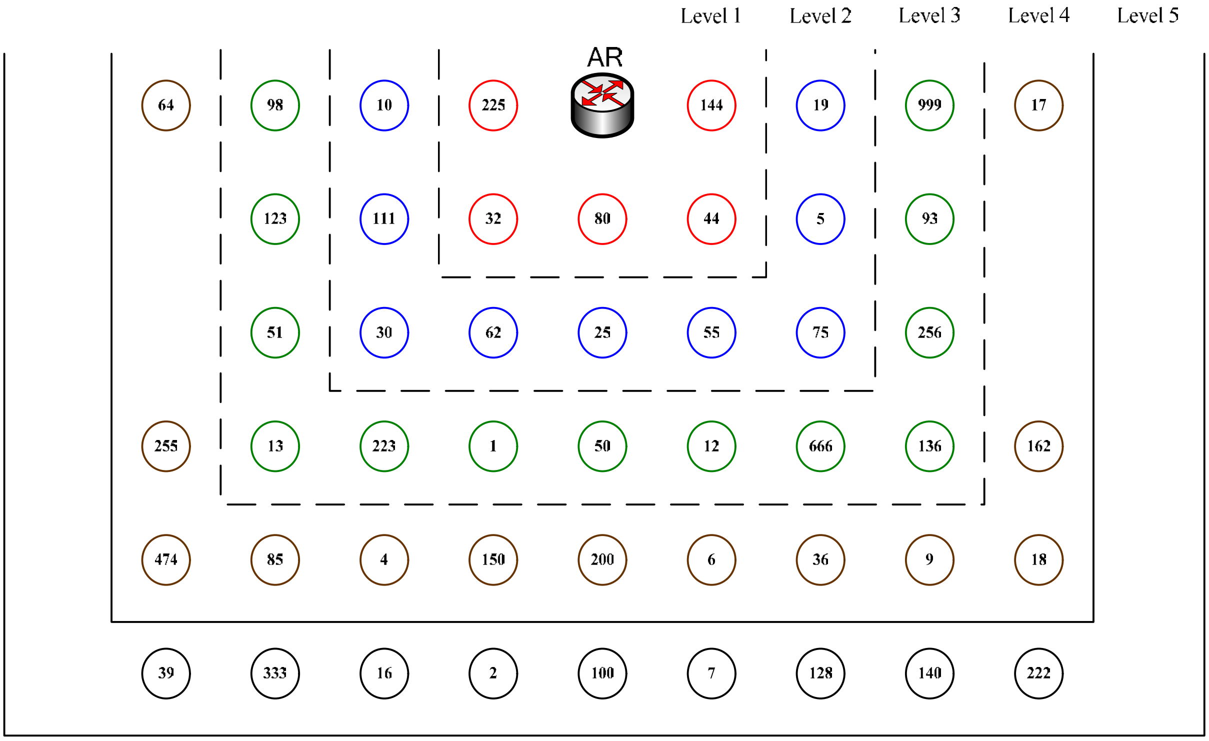


Figure 3. Level of all nodes.

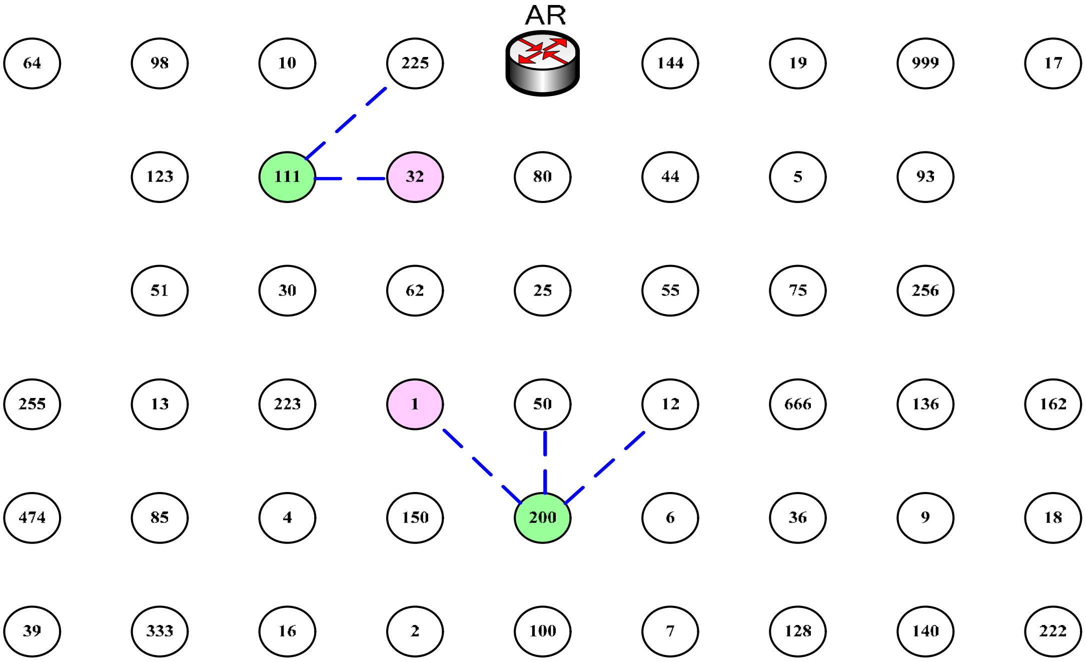


Figure 4. The distributed scheme.

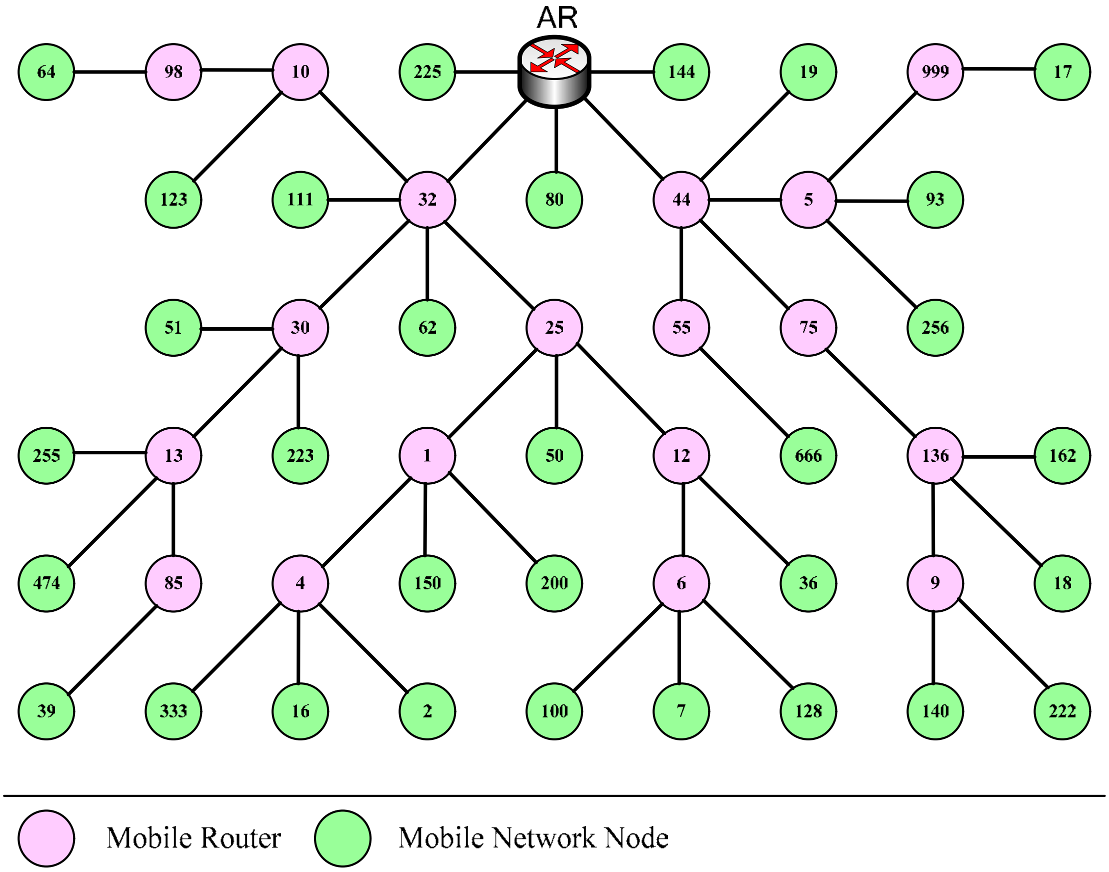


Figure 5. The distributed scheme (complete).

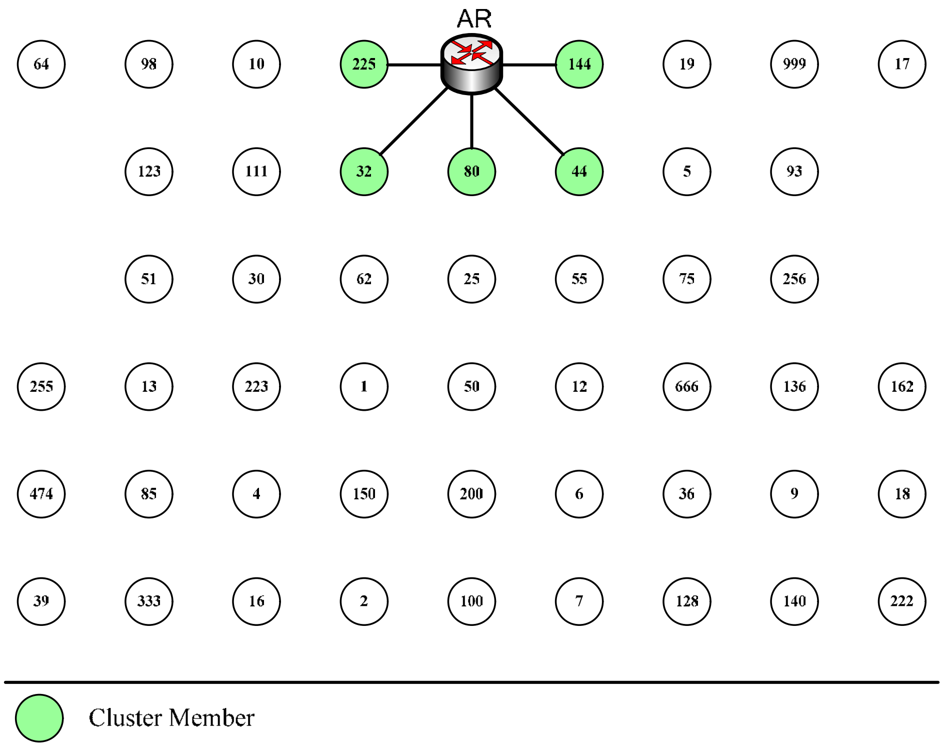


Figure 6. Centralized scheme (step 1).

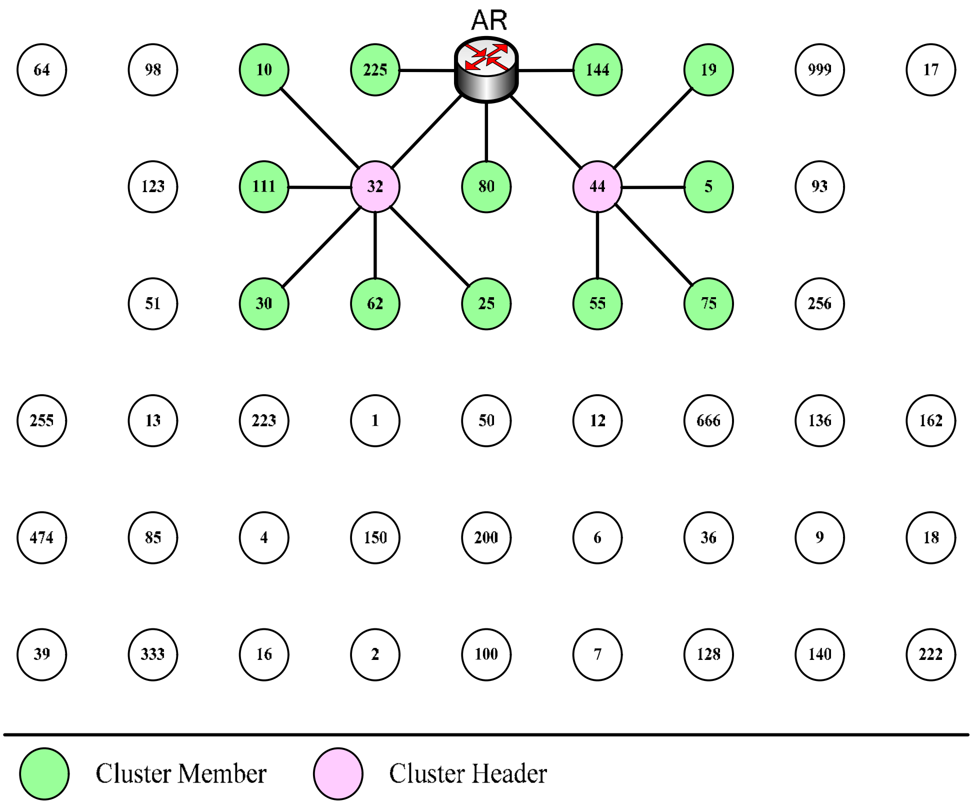


Figure 7. Centralized scheme (step 2).

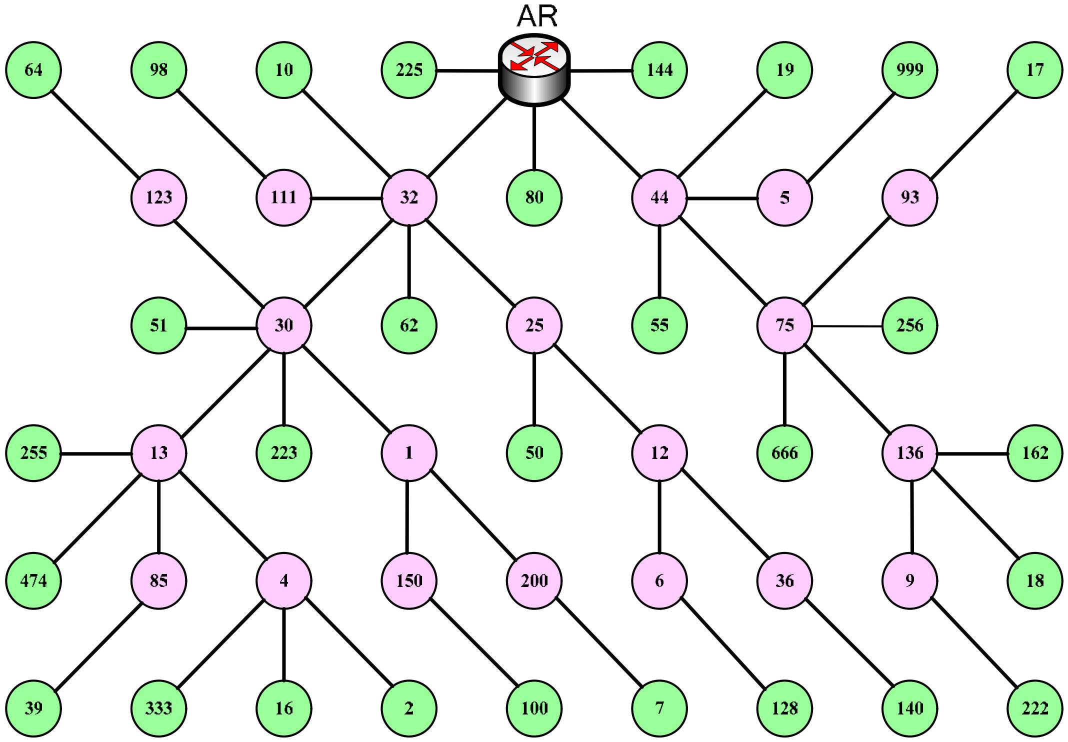


Figure 8. Centralized scheme (complete).

suitable cluster. This will cause the load balance issue. We will discuss the Sub-Section 3.3. From top to down, every node is selected as a CM or CH. However, CMs play the roles of MNNs and CHs play the roles of MRs.

3.3. Load balancing Scheme (LBS)

In Sub-Section 3.1 and Sub-Section 3.2, when the connection of the AR and the Internet fails, the communication of a number of nodes in MONET is disconnected. For large MONET, which includes many MRs and MNNs, it will become a critical issue. In this section, we propose a scheme to achieve dynamic load balancing. In Figure 9, the load in/out in AR1 is 100kb/s, the load in/out in AR2 is 500kb/s. In the point of view for load balancing, the cluster-overlay node should be connecting to AR1. There are smaller packet load and nodes behind AR1 than AR2. But in NEMO basic support protocol, the cluster-overlay node receives the RA1 and the RA2 simultaneously. It will send BU to it's HA through AR1 or AR2. If the cluster-overlay node sends BU via AR2, the cluster-overlay node will communicate to CN with 100Kb/s, the total load in/out in AR2 is 600Kb/s. The load between AR1 and AR2 will be exceedingly unbalanced. In Figure 10 is the RA message format defines in Mobile IPv6. There is no useful information to solve this problem. So we modify the RA message format to achieve our goal.

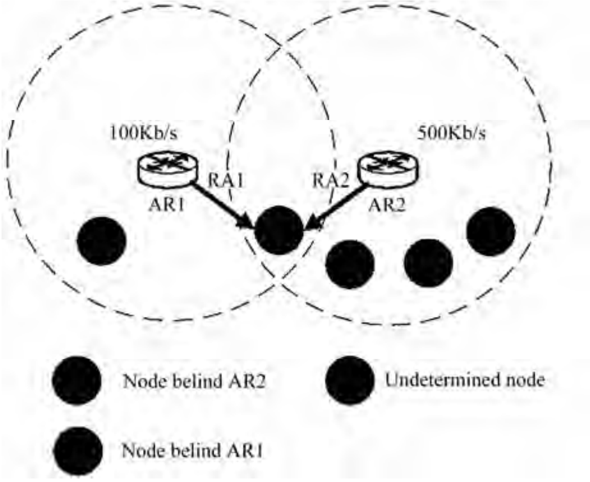


Figure 9. The node between two Ars.

<table>
<tr><td>Type</td><td colspan="4">Code</td><td>Checksum</td></tr>
<tr><td>Cur Hop Limit</td><td>M</td><td>O</td><td>H</td><td>Reserved</td><td>Router Lifetime</td></tr>
<tr><td colspan="6">Reachable Time</td></tr>
<tr><td colspan="6">Retrans Timer</td></tr>
<tr><td colspan="6">Options</td></tr>
</table>

Figure 10 . RA message format.

Type	Code				Checksum
Cur Hop Limit	M	O	H	Node_Num	Router Lifetime
Reachable Time					
Retrans Timer					
Options					

Options Type	Length	…
…		

Options Type=11	Length	Current Load Information

Figure 11. RA message format that we proposed.

In Figure 11, we add some information in "Reserved" and "Option". The "Node_Num" is the number of nodes behind the AR/MR. The Current Load Information is the current packets load through the AR/MR. For Example, in Figure 9, the AR1 sends RA1 with "Node_Num = 1" and "Current Load Information = 100", the AR2 sends the RA2 with "Node_Num = 3" and "Current Load Information = 500". So the cluster-overlay node receives the RA1 and RA2 simultaneously, it will send the BU to it's HA through AR1. Because the packets load of AR1 is smaller than AR2.

3.3.1. Traffic Load Balancing

Two load balancing schemes are proposed. One is the traffic load balancing, another is node number balancing. Of course, the traffic load is the most important criterion in our scheme. In Figure, MR1 will send BU to MR1's HA through AR1. We hope that the traffic load of AR1 and the traffic load of AR2 are equal. But it is impossible. So our goal is to reduce the load difference between two ARs. So that is why MR5 choose the AR with smaller load.

3.3.2. Node Number Balancing

According to the load information, the node can choose the most appropriate AR/MR to connect. In Figure 13, the traffic load of AR1 is 110Kb/s and the traffic load of AR1 is 105Kb/s. The difference between two ARs is too small. Now, another criterion in our scheme is the number of node. AR2 has three nodes and AR1 has only node. According to 3.3.1, the MR1 should connect to AR2. So AR2 will have six nodes and AR1 has only one node. If the MNN2, MNN3 and MNN4 will not increase their traffic load, it is not a matter. But the MNN2 increases 50Kb/s, MNN3 increases 50Kb/s and MNN4 increases 50Kb/s. The total traffic load of AR2 is 255Kb/s. In this situation, we hope that the MR1 connects to the AR with fewer nodes behinds it. So we propose the Node Number Balancing approach to solve this problem. If the difference of two AR is less than 10% of total load, the node will connect to the AR/MR with fewer nodes behind it.

4. Simulation Results

To evaluate the performance of the proposed schemes, we design a simulation program to calculate the performance. The simulation program is designed by Visual C++ in Windows XP. Table is the simulation parameters in our network topology.

In Figure 14, the x-axis is the number of nodes; the y-axis is the value of LBI. According to the formula of LBI, the most ideal situation is the value of the LBI is 1. In our simulation, we assume that the total packet numbers are proportionally increase by the node number. So the numbers of nodes increase the traffic load increase. But the LBI does not change. The reason is that the total packets proportionally increase, according to the formula of LBI, the LBI will not change. Additionally, the LBI of

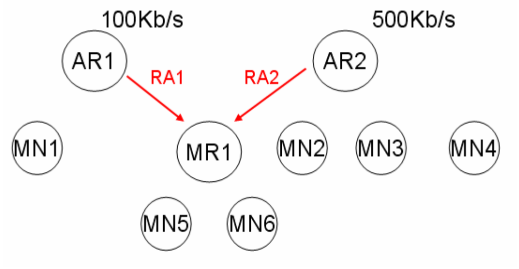


Figure 12. Traffic Load Balancing.

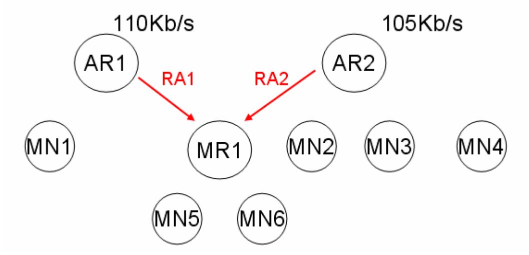


Figure 13. Node balancing.

Table 1. Simulation parameters.

Room Size	500 (m) * 500 (m)
The number of the Access Router	2
The position of the Access Router	(125,125), (375,375)
The number of nodes	50~200
The transmission range of nodes	180 (m)
The average speed of nodes	10 (m/sec)
The direction of nodes	360, Random
The pause time of nodes	10 (sec)
The interval of Router Advertisement	3 (sec)
The interval of Hello Message	3 (sec)
The transmission time of the data packet	Exponential distributions, 300ms in average
The simulation time	900 (sec)

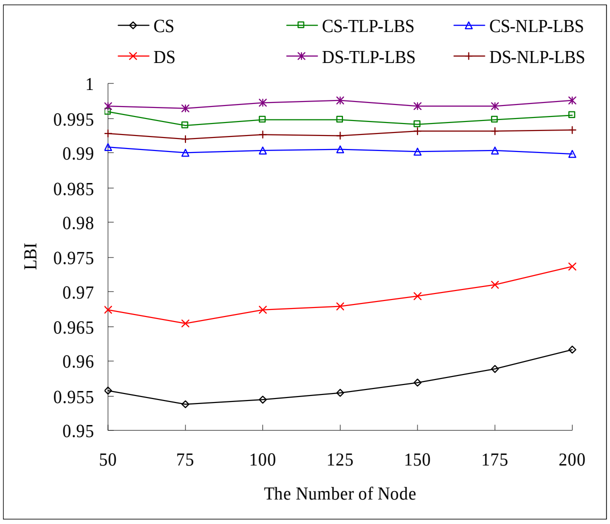


Figure 14. The LBI in different number of nodes (Speed = 10 m/s).

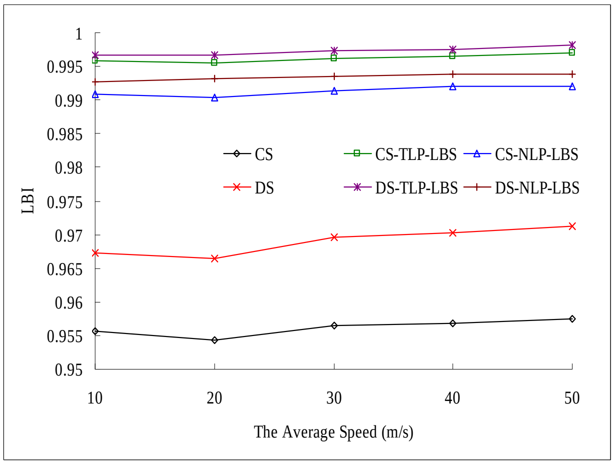


Figure 15. The LBI in different speed (Node =100).

CS (Centralized Scheme) and DS (Distributed Scheme) get improvement when the number of nodes increases. In fact, the unbalancing situation does not get improvement. The reason is that the throughput of one AR is getting down by the packet loss which causes by the limited bandwidth. The packet loss of one AR will cause the difference of throughput of two AR becoming small. That will cause the LBI to get improvement.

In Figure 15, the x-axis is the average speed; the y-axis is the value of LBI. In the original schemes, the nodes can connect to the router with less traffic load. It will cause the unbalancing situation. As the nodes moving faster, the situation does not change more. The reason is the unbalancing situation will alleviate by the moving speed of nodes. In TLP-LBS, we can get the better value of the LBI. The reason is that the nodes connect to the router with less traffic load. In NLP-LBS, the nodes connect to the router with fewer nodes. The router maybe has more traffic load. In this situation, that will cause the router to get more traffic load afterwards.

5. Conclusions and Future Work

In this paper, we propose the two algorithms of reconstruction from MANET to NEMO and the dynamic load balancing scheme. We adopt two cluster algorithms which used in MANET. In the NEMO basic support protocol, the node receives the RA message without any useful information and then sends the BU to it's HA. The node can not choose the most suitable point to connect. In LBA, we put some useful load information into the RA message. According this information, the node in overlapping place between two routers, it will connect to the less loaded router. We hope that the packets load between routers is balanced. The simulation results show that our approach can achieve large performance improvement.

The evaluation performed in this paper is just preliminary. NEMO provides ubiquitous communications with network access. Therefore, all groups of communication nodes can move as a single unit. We must evaluate the proposed scheme with more realistic condition. e.g., the nodes with the different movement speed. In the future work, we will evaluate the end-to-end delay and packet delivery ratio.

6. References

[1] D. Johnson, C. Perkins, and J. Arkko, "Mobility support in IPv6," Network Working Group, RFC3775, June 2004.

[2] T. Ernst and H. Y. Lach, "Network mobility support terminology," Internet Draft drafr-ietf-nemo-terminology-03.txt, February 2005.

[3] The IETF NEMO working group, As of April 2005. http://www.ietf.org/html.charters/nemocharter.html.

[4] V. Devarapalli, *et al.*, "Network mobility (NEMO) basic support protocol," Request for Comments: 3963, IETF, January 2005.

[5] T. Ernst and H. Lach, "Network mobility support terminology," Internet draft, February 2005. [Online]. Available: http://ietfreport.isoc.org/idref/draft-ietf-nemo-terminology

[6] D. Johnson, *et. al.*, "Mobility support in IPv6," Internet-Draft, draft-ietf-mobileip-ipv6-24.txt, Internet Engineering Task Force (IETF), Work in Progress, June 2003.

[7] C. R. Li and G. Mario, "Adaptive clustering for mobile wireless networks," IEEE Journal of Selected Areas in Communications, Vol. 15, No. 7, pp. 1265–1275, September 1997.

[8] M. Gerla and J. T. C. Tsai, "Multicluster, mobile, multimedia radio network," Wireless Networks, Vol. 1, No. 3, pp. 255–265, 1995.

Int. J. Communications, Network and System Sciences, **2009, 2, 903-911**
doi:10.4236/ijcns.2009.29105 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Efficient Time/Frequency Permutation of MIMO-OFDM Systems through Independent and Correlated Nakagami Fading Channels

Khodr A. SAAIFAN, Emad K. AL-HUSSAINI
Department of Electronics and Communications, Cairo University, Giza, Egypt
E-mail: *khedrs@hotmail.com, emadh@eng.cu.edu.eg*

Received August 16, 2009; *revised September* 28, 2009; *accepted November* 4, 2009

Abstract

Space-Time Frequency (STF) codes for MIMO-OFDM over block-fading channel can achieve rate M_t and full-diversity $M_t\ M_r\ M_b\ L$ which is the product of the number of transmit antennas M_t, receive antennas M_r, fading blocks M_b and channel taps L. In this article, time permutation is proposed to provide independent block-fading over Jake's Doppler power spectrum channel. Moreover, we show the performance variations of STF code as channel delay spread changes. Therefore, we introduce a frequency/time permutation technique in order to remove the frequency correlation among sub-carriers, which subsequently increases the coding gain and achieves maximum diversity. Finally, the symbol error rate (SER) performance of the proposed time/frequency permuted STF codes over independent and correlated MIMO antenna branches under Nakagami fading channel is simulated. We show that the proposed systems provide better performance and more robust to large values of antennas correlation coefficients in comparison with the un-interleaved one.

Keywords: MIMO, OFDM, Space-Time Frequency Coding, Nakagami Fading Channel, Time/Frequency Interleaving

1. Introduction

Achieving high data rate, full diversity gain and higher network capacity becomes the major requirements of wireless system providers. MIMO-OFDM system is one of the most attractive techniques to provide these capabilities.

Recently, some attention has been devoted to design STF codes for MIMO-OFDM system with M_t transmit antennas, M_r receive antennas, and N-OFDM tones through L multi-path fading channel. There are several papers, which discussed the code structure to provide full diversity gain and high data rate. In [1], W. Su *et al.* proposed the design of full diversity space frequency block code (SFBC) with rate-1 for any number of transmit antennas and arbitrary power delay profiles. The rate-M_t full diversity SFBC was proposed in [2] for any arbitrary number of transmit antennas. However, because a zero-padding matrix has to be used when N is not an integer multiple of $M_t L$, the symbol transmission rate M_t cannot be always guaranteed.

In [3], better diversity gains through block-fading channels can be obtained, that was done by spreading the coding across multiple fading blocks. In [4], they studied the error performance results of STF codes in MIMO-OFDM systems for a variety of system configurations and channel conditions. The maximum diversity is the product of time diversity, frequency diversity and space diversity as shown in [5]. Recently in [6], W. Zhang *et al.* proposed a systematic design of high-rate STF codes for MIMO frequency-selective block-fading channels. By spreading the algebraic coded symbols across different OFDM sub-channels, transmit antennas and fading blocks, the proposed STF codes can achieve a rate-M_t and a full diversity of $M_t\ M_r\ M_b\ L$, where M_b is the number of independent fading blocks in the code-words. To achieve the full-diversity performance of STF code, maximum-likelihood (ML) decoding must be employed. In order to decrease the large complexity of ML decoding, sphere decoder can be considered to achieve near-ML performance [7,8]. For block-fading channels, the performance of STF-coded OFDM is much better than SF coding as demonstrated in [9].

In MIMO-OFDM systems, the DFT operation introduces correlation into the channel frequency response at different sub-carriers [10,11], making its performance varies as the delays between paths vary.

The outline of the paper is as follows. Section 2 describes the channel statistics and system model. The suggested time/frequency permutations of high rate STF codes structure proposed in [6] for independent and correlated Nakagami fading are introduced in Section 3. In Section 4, we provide simulation results for the performance of the proposed scheme. Finally, some conclusions are made in Section 5.

2. Channel Statistics and System Models

Before investigating permutation schemes for MIMO-OFDM systems equipped with M_t transmit antennas, M_r receive antennas in mobile radio channels, we briefly describe the channel statistics, emphasizing the separation property of mobile wireless channels, which is crucial for simplifying our time/frequency permutation. In this section we also briefly describe a MIMO-OFDM system.

2.1. Statistics of Mobile Radio Channels

The channels between each pair of transmit and receive antennas are assumed to have L independent delay paths and the same power delay profile. The channel impulse response between $m_t{}^{th}$ transmit antenna and $m_r{}^{th}$ receive antenna can be modeled as

$$h_{m_t,m_r}(t;\tau) = \sum_{l=0}^{L-1} \alpha^l_{m_t,m_r}(t)\delta(t-\tau_l) \qquad (1)$$

where τ_l is the delay of the l^{th} path, and $\alpha^l_{m_t,m_r}(t)$ is complex amplitude of the l^{th} path between $m_t{}^{th}$ transmit antenna and $m_r{}^{th}$ receive antenna. $\alpha^l_{m_t,m_r}(t)$'s are modeled as a complex random fading signals with Nakagarni-m distributed fading amplitudes and uniform phases. Nakagami m-distribution fading model [12] is one of the most versatile, in the sense that it has greater flexibility and accuracy in matching some experimental data than Rayleigh, log-normal, or Rician distributions. The Rayleigh distribution is a special case when the fading parameter m=1. It can approximate Rice distribution for $m>1$. Moreover, it is assumed that all path gains between any pair of transmit and receive antennas follow the same power profile, i.e., $E\left[\left|\alpha^l_{m_t,m_r}(t)\right|^2\right] = \sigma_l^2 > 0$ for any given (m_t, m_r, l). The powers of the paths are normalized such that $\sum_{l=0}^{L-1}\sigma_l^2 = 1$. Using Equation (1), the frequency responses of the time-varying radio channel at time t is

$$H_{m_t,m_r}(t,f) = \sum_{l=0}^{L-1} \alpha^l_{m_t,m_r}(t)\exp(-j2\pi f\tau_l) \qquad (2)$$

The MIMO channel is assumed to be spatially correlated for any (m_t, m_r), where m_t=1,…M_t, m_r=1,…M_r, and independent for any l where, l=0,…L-1. Let $\rho^{TX}_{m_t,m_t'}$ denotes the spatial correlation coefficient between $\alpha^l_{m_t,m_r}(t)$ and $\alpha^l_{m_t',m_r}(t)$ defined as

$$\rho^{Tx}_{m_t,m_t'} = \left\langle \alpha^l_{m_t,m_r}(t), \alpha^l_{m_t',m_r}(t) \right\rangle \qquad (3)$$

The spatial correlation coefficient observed at the receiver has also been extensively studied in the literature and is given as

$$\rho^{Rx}_{m_r,m_r'} = \left\langle \alpha^l_{m_t,m_r}(t), \alpha^l_{m_t,m_r'}(t) \right\rangle \qquad (4)$$

Given Equations (3) and (4), the symmetrical correlation matrices at transmitter and the receiver can be defined respectively as

$$\mathbf{R}_{Tx} = \begin{bmatrix} \rho^{Tx}_{11} & \rho^{Tx}_{12} & \cdots & \rho^{Tx}_{1M_t} \\ \rho^{Tx}_{21} & \rho^{Tx}_{22} & \cdots & \rho^{Tx}_{2M_t} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{Tx}_{M_t1} & \rho^{Tx}_{M_t2} & \cdots & \rho^{Tx}_{M_tM_t} \end{bmatrix}_{M_t\times M_t} \qquad (5)$$

and,

$$\mathbf{R}_{Rx} = \begin{bmatrix} \rho^{Rx}_{11} & \rho^{Rx}_{12} & \cdots & \rho^{Rx}_{1M_r} \\ \rho^{Rx}_{21} & \rho^{Rx}_{22} & \cdots & \rho^{Rx}_{2M_r} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{Rx}_{M_r1} & \rho^{Rx}_{M_r2} & \cdots & \rho^{Rx}_{M_rM_r} \end{bmatrix}_{M_r\times M_r} \qquad (6)$$

The spatial correlation matrix $\mathbf{R}$ of the MIMO radio channel is the Kronecker product of the spatial correlation matrix at the transmitter and the receiver and is given by [13]

$$\mathbf{R} = \mathbf{R}_{Tx} \otimes \mathbf{R}_{Rx} \qquad (7)$$

where $\otimes$ denotes the Kronecker product.

The correlation function of the frequency response for different times and frequencies is

$$\begin{aligned} \Phi_{m_t,m_t'}(\Delta t,\Delta f) &= E\left[H^*_{m_t,m_r}(t,f)H_{m_t',m_r}(t+\Delta t, f+\Delta f)\right] \\ &= \sum_{l=0}^{L-1} E[\alpha^{l*}_{m_t,m_r}(t)\alpha^l_{m_t',m_r}(t+\Delta t)]\exp(-j2\pi\Delta f\tau_l) \end{aligned} \qquad (8)$$

Assume Jake's Doppler power spectrum [14], therefore the correlation of the l th path is given by

$$E[\alpha^{l*}_{m_t,m_r}(t)\alpha^l_{m_t',m_r}(t+\Delta t)] = \rho^{Tx}_{m_t,m_t'}\sigma_l^2 J_0(2\pi f_D\Delta t) \qquad (9)$$

where σ_l^2 represents the power of l th path, f_D is the Doppler frequency, and $J_O(x)$ is the zero order Bessel function of the first kind. Substitute Equation (9) in Equation (8), then Equation (8) can be rewritten as

$$\begin{aligned}\Phi_{m_t,m_t'}(\Delta t,\Delta f) &\\ = \rho^{Tx}_{m_t,m_t'} J_0(2\pi f_D \Delta t)\sum_{l=0}^{L-1}\sigma_l^2 \exp(-j2\pi\Delta f\tau_l) & \quad (10)\\ = \rho^{Tx}_{m_t,m_t'}\Phi_t(\Delta t)\Phi_f(\Delta f) &\end{aligned}$$

where $\Phi_t(\Delta t)$ is the time domain correlation function and $\Phi_f(\Delta f)$ is the frequency domain correlation function. From Equation (10), the time-frequency domain channel correlation function of $H_{mt,mr}(t,f)$ can be separated as the product of the spatial correlation coefficient, the time domain channel correlation, and the frequency domain channel correlation, which are dependent on the antenna separation, the Doppler frequency, and multi-path delay spread respectively.

For an OFDM system with block length $\boldsymbol{T}$ and tone spacing (sub-channel spacing) Δf=1/T, the correlation function for different blocks and tones can be written as

$$\Phi_{m_t,m_t'}(\Delta t,\Delta f) = \rho^{Tx}_{m_t m_t'}\Phi_t(kT)\Phi_f(n/T) \qquad (11)$$

2.2. MIMO-OFDM System Model

Consider a STF-coded MIMO-OFDM system with M_t transmit antennas, M_r receive antennas and N sub-carriers operating over a frequency-selective multi-path fading channel. The MIMO-OFDM system with code permutations considered in this paper is shown in Figure 1.

The source **S** generates N_s=N $M_t M_b$ information symbols from the discrete alphabet $\boldsymbol{A}$, which are quadrature amplitude modulation (QAM) normalized into the unit power. Using a mapping f: **S**→**C**, an information symbol vector $\mathbf{S}\in A^{Ns}$ is parsed into blocks and mapped onto a STF codeword to be transmitted over the M_t transmit antennas and M_b OFDM blocks. Each STF codeword **C** can be expressed as a $N \times M_b M_t$ matrix.

$$\mathbf{C} = \begin{bmatrix}\mathbf{C}^1 & \mathbf{C}^2 & \cdots & \mathbf{C}^{M_b}\end{bmatrix} \qquad (12)$$

where the N×M_t matrix $\mathbf{C}^{m_b} = \begin{bmatrix}\mathbf{c}_1^{m_b} & \mathbf{c}_2^{m_b} & \cdots & \mathbf{c}_{M_t}^{m_b}\end{bmatrix}$ for m_b=1,...M_b denotes the sub-codeword ready to be sent during the time epoch m_b. The m_tth (m_t=1,...M_t) column of $\mathbf{C}^{mb}$ denoted by $\mathbf{c}^{m_b}_{m_t}$ is sent to the OFDM block at the m_tth transmit antenna during the time epoch m_b. After inverse fast Fourier transform (IFFT) modulation and cyclic prefix (CP) insertion, OFDM symbols are sent from all transmit antennas simultaneously.

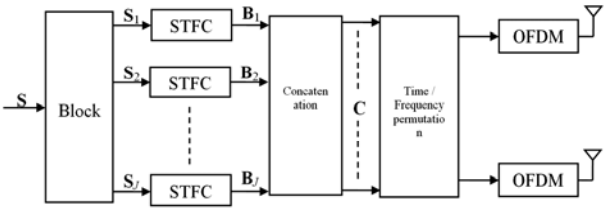


Figure 1. MIMO-OFDM system with code permutation to combat channel correlation.

At the receiver, after matched filtering, removing the cyclic prefix, and applying FFT, the received signal at the received signal at the m_rth receive antenna during the time epoch m_b is given by

$$\mathbf{Y}^{m_b}_{m_r} = \sqrt{\frac{\rho}{M_t}}\sum_{m_t=1}^{M_t}\mathrm{diag}(\mathbf{c}^{m_b}_{m_t})\mathbf{H}^{m_b}_{m_t,m_r} + \mathbf{Z}^{m_b}_{m_r} \qquad (13)$$

where

$$\mathbf{H}^{m_b}_{m_t,m_r} = \begin{bmatrix}H^{m_b}_{m_t,m_r}(0) & H^{m_b}_{m_t,m_r}(1) & \cdots & H^{m_b}_{m_t,m_r}(N-1)\end{bmatrix}^T$$

is the m_bth OFDM block channel frequency response vector between m_tth transmit antenna and m_rth receive antenna and $\mathbf{Z}^{m_b}_{m_t}$ denotes the complex discrete AWGN process with zero mean and unit variance at the m_rth receive antenna. The factor $\sqrt{\rho/M_t}$ in Equation (13) ensures that the average SNR at each receive antenna is independent on the number of transmit antennas.

3. Time/Frequency Permuted STF Codes

STF coding proposed in [6] can achieve rate of M_t and full diversity for any number of transmit antennas and any arbitrary channel power delay profiles. It was constructed by applying the layering concept along with algebraic code components, which was introduced in the design of threaded algebraic space-time (TAST) code [15]. The STF code structure spreads the algebraic code components in adjacent sub-carriers and adjacent time slots that suffer from high correlation introduced by DFT operation and time correlation respectively. In this section, time/frequency permuted STF code structure is introduced into STF code structure of [6] in order to remove the effect of channel correlation among the code components and achieve better diversity order.

3.1. STF Codes Structures

Let $N_p = 2^{\lceil \log_2 L\rceil}$, $N_q = 2^{\lceil \log_2 M_t\rceil}$, and $K = N_p \cdot N_q$, then a block of N_s transmitted information symbols S=[S$_1$,S$_2$,...S_{NMtMb}]^T are parsed into J(J=N/K) equal size sub-blocks. Each sub-block $\mathbf{S}_j\in AK^{MtMb}$ (j=1,2…, J) is respectively encoded into an STF code matrix **B***j* of size K×M_tM_b through the following steps:

1) Each subblock $\mathbf{S}_j$ (j=1,2…, J) are parsed into N_q information vector $\mathbf{s}_{n_q} \in A^{N_p M_t M_b}$ $(n_q = 1,2,\cdots,N_q)$.

2) Generate algebraic code sub-block $\overline{\mathbf{X}}_{n_q}$ by applying a fully-diverse unitary transformations $\mathbf{\Theta}$ into each information vector $\mathbf{s}_{n_q}$ $(n_q = 1,2,\cdots N_q)$ to generate N_q threads by

$$\begin{aligned}\overline{\mathbf{X}}_{n_q} &= \left[\overline{\mathbf{X}}^{1}_{n_q,1} \quad \cdots \quad \overline{\mathbf{X}}^{1}_{n_q,N_L} \quad \cdots \quad \overline{\mathbf{X}}^{M_b}_{n_q,1} \quad \cdots \quad \overline{\mathbf{X}}^{M_b}_{n_q,N_L}\right] \\ &= \left[X_{n_q}(1) \quad X_{n_q}(2) \quad \cdots \quad X_{n_q}(\overline{N})\right] \\ &= \mathbf{\Theta}\mathbf{s}_{n_q}\end{aligned} \tag{14}$$

where $\overline{N} = N_p M_t M_b$, and $\mathbf{\Theta}$ is the first principal $\overline{N} \times \overline{N}$ unitary matrix of the following matrix

$$\mathbf{\Psi} = \mathbf{F}^{H}_{\overline{M}}\, diag\left(1, \varphi, \cdots, \varphi^{\overline{M}-1}\right) \tag{15}$$

where $\overline{M} = 2^{\lceil \log_2 \overline{N} \rceil}$, $\mathbf{F}^{H}_{\overline{N}}$ is the $\overline{M} \times \overline{M}$ discrete Fourier transform (DFT) matrix, and $\varphi = \exp\left(j2\pi/4\overline{M}\right)$.

3) Applying the layering concept to construct the encoder sub-matrices $\overline{\mathbf{X}}^{m_b}_{n_p}$ ($n_p = 1,\cdots N_p$ and $m_b = 1,\cdots M_b$).

$$\overline{\mathbf{X}}^{m_b}_{n_p} = \left[\overline{\mathbf{X}}^{m_b\;T}_{1,n_p} \quad \varphi\overline{\mathbf{X}}^{m_b\;T}_{2,n_p} \quad \cdots \quad \varphi^{N_q-1}\overline{\mathbf{X}}^{m_b\;T}_{N_q,n_p}\right]$$
$$= \begin{pmatrix} X_1(k^{m_b}_{n_p}+1) & \varphi X_2(k^{m_b}_{n_p}+1) & \cdots & \varphi^{N_q-1}X_{N_q}(k^{m_b}_{n_p}+1) \\ X_1(k^{m_b}_{n_p}+2) & \varphi X_2(k^{m_b}_{n_p}+2) & \cdots & \varphi^{N_q-1}X_{N_q}(k^{m_b}_{n_p}+2) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(k^{m_b}_{n_p}+M_t) & \varphi X_2(k^{m_b}_{n_p}+M_t) & \cdots & \varphi^{N_q-1}X_{N_q}(k^{m_b}_{n_p}+M_t) \end{pmatrix} \tag{16}$$

where $\phi = \varphi^{1/N_q}$ and $k^{m_b}_{n_p} = (n_p - 1)M_t + (m_b - 1)N_p M_t$.

4) Re-arrange the elements of $\overline{\mathbf{X}}^{m_b}_{n_p}$ by $\overline{\mathbf{X}}^{m_b}_{n_p}(m'_t, n'_q)$ $= \overline{\mathbf{X}}^{u}_{m}(m_t, n_q)$: $n'_q = \left\{\left(m_t + n_q - 2\right)_{\bmod M_t} + 1\right\}$, and $m'_t = \left\{m_t + \left\lceil \frac{n_q}{M_t} \right\rceil - 1\right\}$, for $1 \le m_t < M_t$, $1 \le n_q \le N_q$

$$\begin{pmatrix} X_1(k^{m_b}_{n_p}+1) & \varphi X_2(k^{m_b}_{n_p}+1) & \cdots & \varphi^{M_t-1}X_{M_t}(k^{m_b}_{n_p}+1) \\ \varphi^{N_q-1}X_{N_q}\left(k^{m_b}_{n_p}+1\right) & X_1(k^{m_b}_{n_p}+2) & \cdots & \varphi^{M_T-2}X_{M_t-1}(k^{m_b}_{n_p}+2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi X_2(k^{m_b}_{n_p}+M_t) & \varphi^2 X_3(k^{m_b}_{n_p}+M_t) & \cdots & \varphi^{N_q-1}X_{N_q}(k^{m_b}_{n_p}+M_t) \end{pmatrix} \tag{17}$$

then, the $K \times M_t M_b$ code matrix $\mathbf{B}_i$ is constructed as

$$\mathbf{B}_i = \begin{pmatrix} \overline{\mathbf{X}}^{1}_{1} & \overline{\mathbf{X}}^{2}_{1} & \cdots & \overline{\mathbf{X}}^{M_b}_{1} \\ \overline{\mathbf{X}}^{1}_{2} & \overline{\mathbf{X}}^{2}_{2} & \cdots & \overline{\mathbf{X}}^{M_b}_{2} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\mathbf{X}}^{1}_{N_L} & \overline{\mathbf{X}}^{2}_{N_L} & \cdots & \overline{\mathbf{X}}^{M_b}_{N_L} \end{pmatrix} \tag{18}$$

The STF coding applies the same coding strategy to every sub-block $\mathbf{B}_j\,(j = 1,2,\cdots,J)$, then the rate-$M_t$ STF code $\mathbf{C} \in C^{N \times M_t M_b}$ is of the form

$$\mathbf{C} = \left[\mathbf{B}^T_1 \quad \mathbf{B}^T_2 \quad \cdots \quad \mathbf{B}^T_J\right]^T \tag{19}$$

It is clear that, each thread of codeword $X_{n_q}(n_{\overline{n}})$ ($n_q = 1,2,\cdots N_q$ and $n_{\overline{n}} = 1,2,\cdots \overline{N}$) is spread over space, time and frequency dimensions. Therefore, the STF code structure is not optimum in spreading the code components of each thread on adjacent sub-carriers that suffer from high correlation introduced by DFT operation. However, if the power delay profile of the channel is available at the transmitter side, further improvement can be achieved by developing an interleaving strategy (can reduce the correlation between adjacent sub-carriers) which explicitly considers the power delay profile. In addition, since the STF code structure maintains its diversity gain from sending the OFDM blocks through independent fading blocks, we shall introduce time permutation to achieve independent fading blocks through MIMO channels that suffer from high correlation introduced by Doppler power spectrum.

3.2. Time/Frequency Permutation Schemes

The assumption of independent fading at the branches is acceptable if the antennas are spaced sufficiently apart with respect to the radio frequency (RF) carrier wavelength. In this case, $\rho^{TX}_{m_t m'_t} = 0, \forall m_t \neq m'_t$, and $\rho^{TX}_{m_t m'_t} = 1, \forall m_t = m'_t$,then Equation (11) will be reduced to the autocorrelation function [10]

$$\Phi_{m_t,m_t}(k,n) = J_0(2\pi f_D kT) * \sum_{l=0}^{L-1} \sigma^2_l \exp(-j2\pi\tau_l n/T) \tag{20}$$

Obviously, the sources of channel correlation are caused by the time domain channel correlation, and the frequency domain channel correlation. Our objective is to find the separation parameters *k* and *n* for MIMO-OFDM system which produce zero time and frequency correlations then permute the algebraic code components of $\mathbf{B}_j$ (*j*=1,…*J*) at zero time frequency correlation to maximize the diversity gain.

$$K_c = \min_k\left[J_0(2\pi f_D kT)\right] \tag{21}$$

$$N_c = \min_{0 \le n \le N-1}\left[\sum_{l=0}^{L-1} \sigma^2_l \exp(-j2\pi\tau_l n/T)\right] \tag{22}$$

The zeros of the Bessel functions (Equation (21)) play a dominant role in our applications. The Bessel functions have infinite number of zeros. The maxima and minima of J_0 steadily decrease in absolute value as *k* increases.

The first five zeros of J_0 are 2.4048, 5.5201, 8.6537, 11.7915, and 14.9309. The interval between the last two is 3.1394, which is already close to π. The larger roots are approximately $\left(v-\frac{1}{4}\right)\pi$, where v is the number of the root. To break the time correlation of the channel, verify independent fading block and realize high-rate full-diversity STC of [6], the M_b-OFDM blocks of STC matrix **B**$_j$ (j=1,…J) should be transmitted at time difference of $K_c = \left\lceil \frac{2.4048}{2\pi f_D T} \right\rceil$. For large coherence time or equivalently low Doppler spread of the fading, high interleaving size is required to break the memory of the channel.

The optimum sub-carriers separation factor N_c (see Equation (22)) can be easily found via low-complexity computer search. However, closed-form solutions for specific cases are reported in [1].

Based on the knowledge of channel separations factors N_c and K_c, time/frequency permuted STF code can be introduced using the following steps:

1) Distribute the STC blocks over independent fading blocks by permuting the u-OFDM blocks of STC matrix **B**$_j$ (j=1,…J) with those blocks at time uK_c, $(u = 2,\cdots M_b)$.

2) Apply frequency permutation into each pair of code matrices **B**$_j$ and **B**$_{j'}$, where $j' = j + N_b$, $N_b = N_c/K$, $j = [1,\cdots,N_b] + 2(n_b - 1)N_b$ and $n_b = 1,2,...,J/2N_b$ by permuting rows $K/2+1,\cdots,K$ of **B**$_j$ with the rows $1,\cdots,K/2$ of **B**$_{j'}$.

3) Further permutation should be done to break the rest of channel frequency correlation by permuting each pair of rows (n_1,n_2), where $n_1 = 2,\cdots K/2$ and $n_2 = K/2+2,\cdots K$ for all code matrices **B**$_j$ $(j = 1,\cdots J)$ with the corresponding pair of rows at block distances $u(M_b - 1)K_c$ where $(u = 2,\cdots,K/2)$.

By performing the above steps as shown in Figure 2, the code components $X_{n_q}(n_{\bar{n}})$ ($n_q = 1,2,\cdots N_q$ and $n_{\bar{n}} = 1,2,\cdots \overline{N}$) of each thread of code matrix **B**$_j$ are affected by independent fading blocks which subsequently achieve maximum diversity gain.

Examples of STF codes and permuted STF codes for M_t=2, L=2 are shown Figures 3 and 4. For M_b=1, STF codes will be, in fact, the SF codes of [16]. The rate-2 SF code structure and the suggested time/frequency permutation (antenna 1 is shown only) are shown in Figure 3.

The rate-2 STF code structure and the suggested time/frequency permutation for M_b=2 are shown in Figure 4.

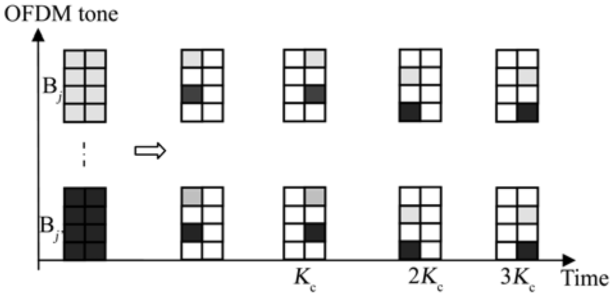


Figure 2. The suggested time/frequency permutation of STF codes.

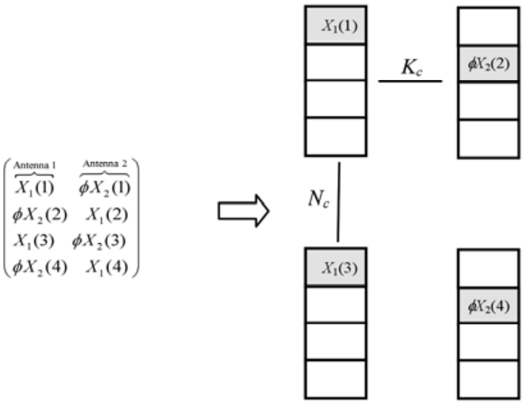


Figure 3. Rate-2 time/frequency permuted SF code (T/FP- SF).

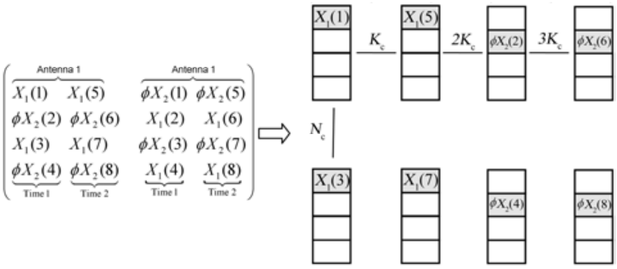

Antenna 1
Antenna 1
$X_1(1)$ $X_1(5)$ $\phi X_2(1)$ $\phi X_2(5)$
$\phi X_2(2)$ $\phi X_2(6)$ $X_1(2)$ $X_1(6)$
$X_1(3)$ $X_1(7)$ $\phi X_2(3)$ $\phi X_2(7)$
$\phi X_2(4)$ $\phi X_2(8)$ $X_1(4)$ $X_1(8)$
Time 1
Time 2
Time 1
Time 2
$X_1(1)$
K_c
$X_1(5)$
$2K_c$
$\phi X_2(2)$
$3K_c$
$\phi X_2(6)$
N_c
$X_1(3)$
$X_1(7)$
$\phi X_2(4)$
$\phi X_2(8)$

Figure 4. Rate-2 time/frequency permuted STF code (T/FP-STF).

4. Simulation Results

In this section, we simulated the proposed permutation scheme and compared with the non-permuted STF codes for different power delay profiles of the channel. We present average symbol-error rate (SER) curves as functions of the average SNR. Then we illustrate the performance of the proposed permutation for SF codes through correlated Nakagami fading channels. To investigate the performance of the proposed time/frequency permutation of STF codes over frequency-selective fading channels, we perform the simulation experiments and compare with the STF codes [6] for MIMO-OFDM systems. In the simulation, we use a 2×2 system with 128 OFDM tones and 4QAM transmission scheme, thus the spectral efficiency is 4 bit/s/Hz, ignoring the cyclic prefix. The bandwidth of OFDM system is 1 MHz and the length of the cyclic prefix is 32, i.e., 32μs. Hence the duration of one OFDM symbol (cyclic prefix excluded) is T=128μs. A two-ray Nakagami fading channel statistics model is considered with the equal gain, Doppler spread f_D=200Hz, and fading depth m = 0.5, 1 and 2.

It is to be noted that m = 0.5 represents the worst fading situation that can be represented by Nakagami distribution. This case can be countered in bad urban mobile radio. When m=1, we obtain Rayleigh fading channel. Finally, m=2 represents the best considered situation in which the fading is less than that of Rayleigh.

4.1. Performance Comparison for Different Delay Spreads

The first set of experiments is conducted to compare the performance of the proposed scheme with STF codes for different path delay of the two-ray model. A simple two-ray, equal-power delay profile, with a delay τ microseconds between the two rays is assumed. Simulation is carried out for two cases: 1) 8μ sec (optimum permutation N_c=8) and 2) 20μ sec (optimum permutation N_c=16). For Doppler spread f_D=200Hz the optimum time separation is 14 OFDM symbols to ensure independent fading blocks, therefore the interleaved STF code is spanned over 56 OFDM symbols.

Figures 5, 6 and 7 depict the improvement in SER performance offered by the proposed time/frequency permutations through independent Nakagami fading channel with different m. The values of the fading depth considered are m = 0.5, 1, and 2 respectively.

It can be observed from these figures that the SER performance of STF codes [6] varied as the delay spread of the channel changed. The SER performance of STF codes is further improved as delay spread of the channel increased. Such an improvement is attributed to the large coding gain induced by multi-path fading channels with a larger delay spread. The performance of the STF code degraded significantly from the 20μs case to the 8μs case, whereas the performance of the STF code using time/ frequency permutation was almost the same for the two delay profiles.

We can see that the T/FP-STF codes have better SER performance than the non-permuted STF codes. For τ=8μs case, there is an improvement of about 3.2 dB for SF codes and an improvement of about 1.8 dB for the STF codes at a SER of 10^{-4} when m=1. Therefore; the proposed interleaving method offering higher code gains making it more robust to small delay spread. This confirms that by careful interleaver design, the performance of the STF codes can be significantly improved.

From Table 1, it is clear that the SNR decreases with the increase of m. The performance of the interleaved codes is not sensitive to the variation in the channel time delay spread. In all of cases considered, the required SNR of the time/frequency interleaved codes is lower than that needed for the un-interleaved one to achieve the same SER.

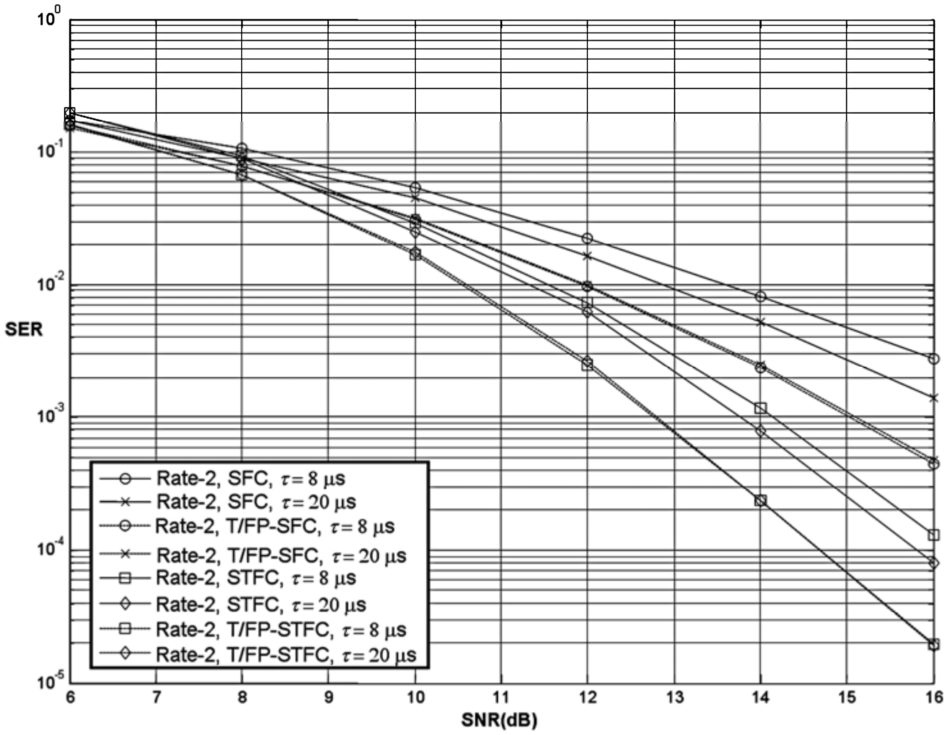


Figure 5. Average SER versus SNR of 2×2, MIMO-OFDM system through independent Nakagami fading channel *m*=0.5 with different delay spread.

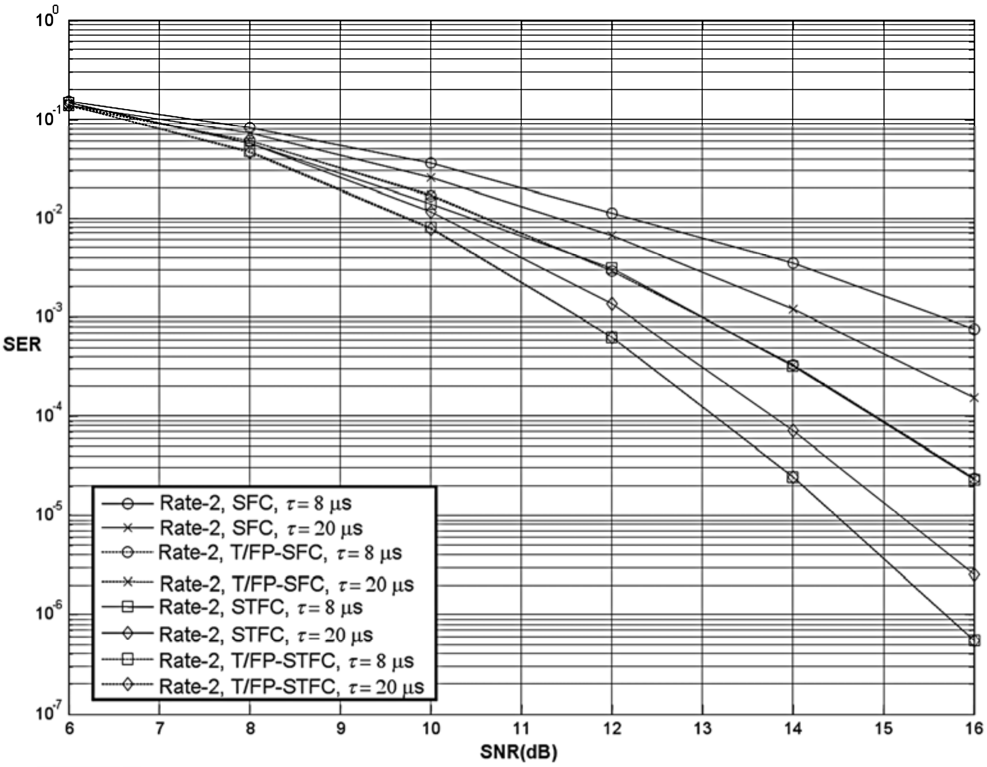


Figure 6. Average SER versus SNR of 2×2, MIMO-OFDM system through independent Nakagami fading channel *m*=1 with different delay spread.

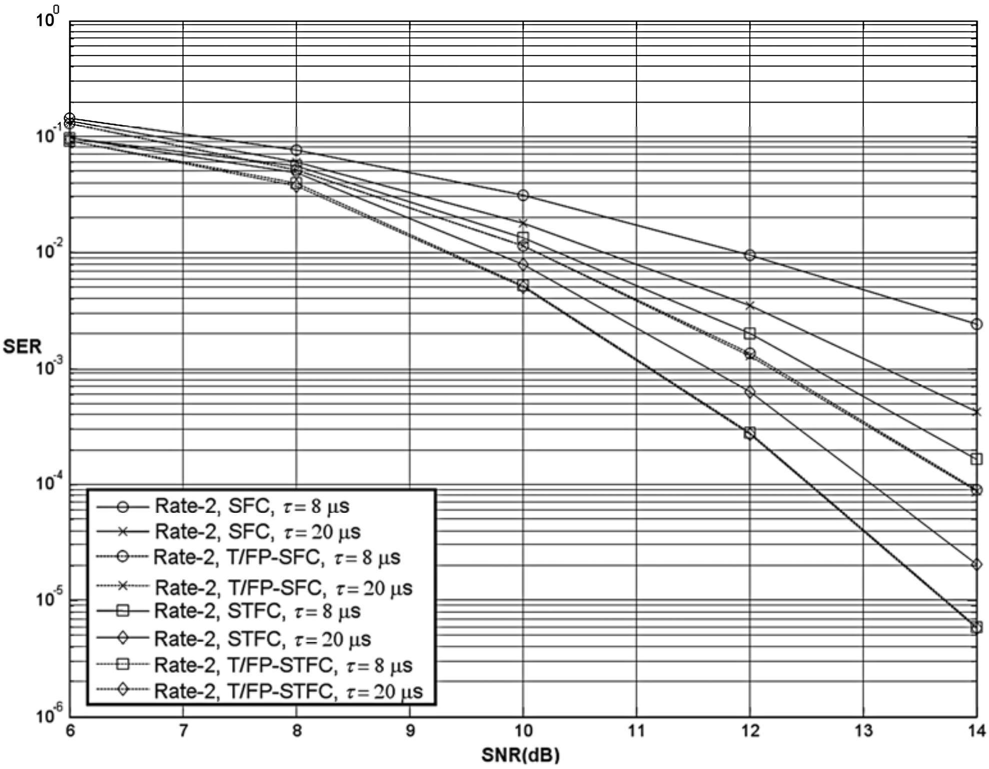


Figure 7. Average SER versus SNR of 2×2, MIMO-OFDM system through independent Nakagami fading channel *m*=2 with different delay spread.

Table 1. SNR required to obtain a SER=10^{-4} for STFC and T/FP-STFC at different time delay spread.

M	SFC		T/FP-SFC		STFC		T/FP-STFC	
	8μsec	20μsec	8μsec	20μsec	8μsec	20μsec	8μsec	20μsec
0.5	20.8 dB	19.4 dB	18 dB		16.3 dB	15.8 dB	14.7dB	
1	18.1 dB	16.2 dB	14.9 dB		14.9 dB	13.8 dB	13.13 dB	
2	17.4 dB	15.2 dB	13.9 dB		14.3 dB	13.1 dB	12.53 dB	

4.2. Performance Comparisons over Correlated Nakagami Fading Channels

MIMO system with closely spaced antenna elements is considered here. Our aim is to analyze the influence of the Nakagami-m fading parameter and the effect of antenna correlation on the SER performance of the rate-2 SF code, and the proposed T/FP-SF code depicted in Figure 3.

Figure 8 shows the SER degradation as the correlation coefficients between the transmitting antenna branches ρ vary from 0 up to 0.8. Similar correlation is assumed between receiving antenna branches. Simulation is carried out for two cases: 1) Transmitter correlated Nakagami MIMO fading channel case: $\mathbf{R}_t = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, and $\mathbf{R}_r = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and 2) Doubly correlated Nakagami MIMO fading channel case: $\mathbf{R}_t = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, and $\mathbf{R}_r = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. The values of the fading depth considered are m=0.5, 1, and 2 respectively. It is clear that the SER increases with the increase of correlation coefficient ρ. At ρ=0, the received signals are independent and the codes practically achieves full diversity reception gain. It is clear that the probability of error decreases with the increase of *m*, which is with the decrease of the severity of fading.

From these figures, it is clear that the systems under consideration appreciably dominate the systems considered in [6].

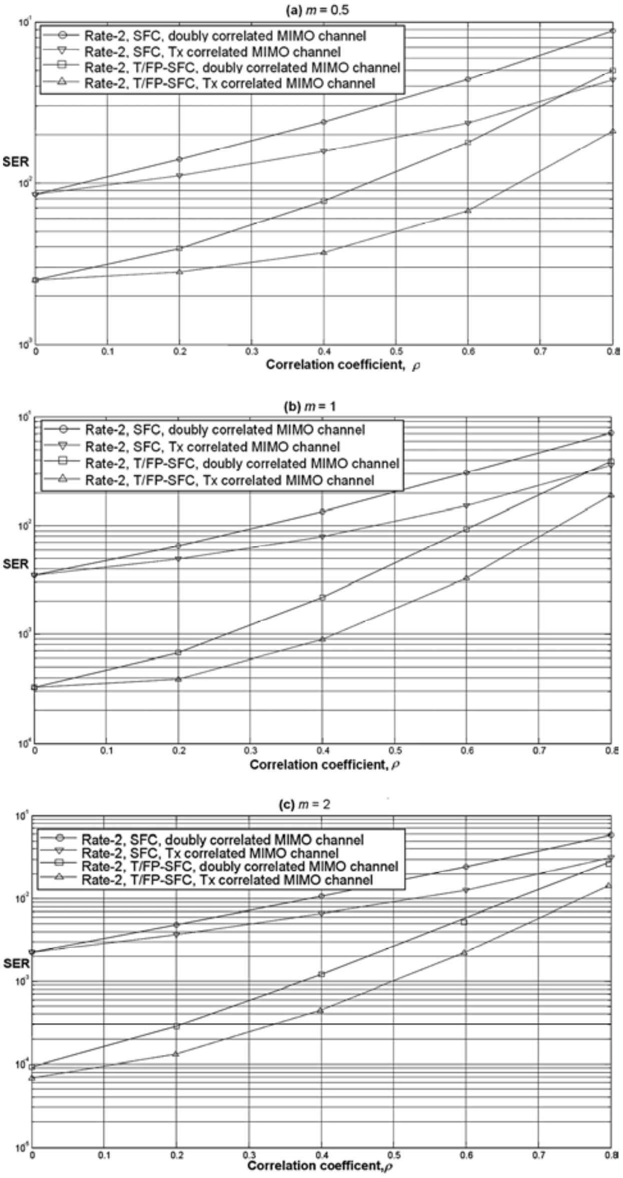


Figure 8. Average SER versus correlation coefficients for 2×2 MIMO-OFDM systems at SNR=14dB.

5. Conclusions

In this paper, the limitation for achieving full-diversity of STF-coded OFDM is introduced. The limitation arises due to the fact that the algebraic code components are spread in adjacent sub-carriers that suffer from high correlation introduced by DFT operation. Assuming that the power delay profile of the channel is available at the transmitter, we proposed an efficient time-frequency interleaving scheme to further improve the performance. Based on simulation results, we can draw the following conclusions.

First, the proposed time/frequency permutations STF codes offer considerable performance improvement over previously reported results. Second, the applied interleaving scheme can have a significant effect on the overall performance of the STF code through correlated and independent Nakagami fading channels.

6. References

[1] W. Su, Z. Safar, and K. J. R. Liu, "Full-rate full-diversity space: Frequency codes with optimum coding advantage," IEEE Transactions on Information Theory, Vol. 51, pp. 229–249, January 2005.

[2] T. Kiran and B. S. Rajan, "A systematic design of high-rate full-diversity space frequency codes for MIMO-OFDM systems," in Proceedings IEEE International Symposium Information Theory, pp. 2075–2079, September 2005.

[3] H. E. Gamal and A. R. Hammons Jr., "On the design of algebraic space-time codes for MIMO block fading channels," IEEE Transactions on Information Theory, Vol. 49, pp. 151–163, January 2003.

[4] M. Fozunbal, S. W. McLaughlin, and R. W. Schafer, "On space-time-frequency coding over MIMO-OFDM systems," IEEE Transactions on Wireless Communication, Vol. 4, pp. 320–331, January 2005.

[5] W. Su, Z. Safar, and K. J. R. Liu, "Towards maximum achievable diversity in space, time, and frequency: Performance analysis and code design," IEEE Transactions on Wireless Communication, Vol. 4, pp. 1847–1857, July 2005.

[6] W. Zhang, X. G. Xia, and P. C. Ching, "High-Rate full-diversity space-time-frequency codes for broadband MIMO block fading channels," IEEE Transaction on Communication, Vol. 55, pp. 25–34, January 2007.

[7] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," IEEE Transactions on Information Theory, Vol. 45, No. 5, pp. 1639–1642, July 1999.

[8] M. O. Damen, A. Chkeif, and J. C. Belfiore, "Lattice code decoder for space-time codes," IEEE Communication Letters, Vol. 4, No. 5, pp. 161–163, May 2000.

[9] W. Zhang, X. G. Xia, and K. B. Letaief, "Space-time/ frequency coding for MIMO-OFDM in next generation broadband wireless systems," IEEE Wireless Communications Magazine, Vol. 14, No. 3, pp. 32–43, June 2007.

[10] Y. Li, L. J. Cimini, and N. R. Sollenberger, "Robust channel estimation for OFDM systems with rapid dispersive fading channels," IEEE Transactions on Communication, Vol. 46, No. 7, pp. 902–915, July 1998.

[11] X. Wang and K. J. R. Liu, "Channel estimation for multicarrier modulation systems using a time-frequency polynomial model," IEEE Transactions on Communication, Vol. 50, No. 7, pp. 1045–1048, July 2002.

[12] M. Nakagami, "The m-distribution: A general formula of intensity distribution of rapid fading", in W. C. Hoffman (ed.), Statistical Methods in Radio Wave Propagation, Pergamon Press, New York, pp. 3–36, 1960.

[13] W. C. Jakes (2nd), "Microwave mobile communications," IEEE Press, New York, 1994.

[14] K. I. Pedersen, J. B. Andersen, J. P. Kermoal, and P. E. Mogensen, "A stochastic multiple-input multiple-output radio channel model for evaluation of space-time coding algorithms," In Proceedings of Vehicular Technique Conference, pp. 893–897, September 2000.

[15] H. E. Gamal and M. O. Damen, "Universal space-time coding," IEEE Transactions on Information Theory, Vol. 49, pp. 1097–1119, May 2003.

[16] W. Zhang, X. G. Xia, P. C. Ching, and H. Wang, "Rate two full-diversity space-frequency code design for MIMO-OFDM," Proceedings of IEEE Workshop Signal Process, Advanced Wireless Communications, New York, pp. 321–325, June 2005.

Int. J. Communications, Network and System Sciences, **2009, 2, 912-916**
doi:10.4236/ijcns.2009.29106 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

A New Multicast Wavelength Assignment Algorithm in Wavelength-Converted Optical Networks

Anping WANG, Qiwu WU, Xianwei ZHOU, Jianping WANG
Department of Communication Engineering, School of Information Engineering,
University of Science and Technology Beijing, Beijing, China
E-mail: *wuqiwu700@163.com*

Received September 14, 2009; *revised October* 25, 2009; *accepted November* 23, 2009

Abstract

In this paper, we propose a new multicast wavelength assignment algorithm called NGWA with complexity of $O(N)$, where N is the number of nodes on a multicast tree. The whole procedure of NGWA algorithm is separated into two phases: the partial wavelength assignment phase and the complete wavelength assignment phase. It tries to minimize the total number of wavelength conversions of the multicast tree. Meanwhile, the number of different wavelengths used is minimized locally. Through illustrative example and simulation experiments, it is proved that the NGWA algorithm works well and achieves satisfactory performance in terms of the average number of wavelength conversions and the average blocking probability.

Keywords: WDM Network, Multicast, Wavelength Assignment, Wavelength Conversion

1. Introduction

Multicast is an efficient way to implement one-to-many communication. The problem of finding a multicast tree and allocating available wavelength for each link of the tree is known as the Multicast Routing and Wavelength Assignment (MC-RWA) problem, which plays a key role in supporting multicasting over WDM networks [1].

Since improper wavelength assignment will cause low network capacity and high connecting blocking probability, the problem of how to assign wavelengths for the multicast tree becomes an important problem in a WDM optical network. According to the different number of multicast connecting requests, the multicast wavelength assignment (MC-WA) problem can be divided into two categories: MC-WA for single multicast which was studied in [2–5] and MC-WA for multiple multicasts which was studied in [6–8]. But to our best knowledge, few studies have been done on the multicast wavelength assignment (MC-WA) problem to minimize both the total number of wavelength conversions of the multicast tree and the number of different wavelengths used. Based on the above, we will propose a greedy algorithm to solve the MC-WA problem.

The rest of the paper is organized as follows. Section 2 introduces the network model and the problem specification. Section 3 proposes a new multicast wavelength assignment algorithm, and an illustrative example is given. The simulation results are shown in Section 4. Finally, the paper is concluded in Section 5.

2. Problem Formulation

2.1. Network Model

The assumptions for the MC-WA problem in this paper are given as follows:

1) The WDM network is an arbitrary connected graph.

2) All links in the network are equipped with the same set of wavelengths.

3) All nodes are provided with full wavelength conversion capacity and light splitting capacity.

Let a directed graph $G=(V, E, M)$ is used to represent a WDM network where V is the vertex set with $|V|=n$, E represents the set of links and $M=\{\lambda_1, \lambda_2,\ldots\lambda_k,\}$ is the set of wavelengths supported by each link with $|M|=k$. Meanwhile, let $M(e)\subset M$ be the set of available wavelengths on link e. In the graph $G=(V, E, M)$, each vertex node $v\in V$ or each edge $e\in E$ is associated with the following costs:

Wavelength usage cost, $C_w(e, \lambda_i)$, the cost of using wavelength λ_i on link e, which is used to computer the multicast tree in multicast routing algorithm.

Wavelength conversion cost, $C_c(v, \lambda_p, \lambda_q)$, the wavelength conversion cost from input wavelength λ_p to output wavelength λ_q at node v, and if $\lambda_p=\lambda_q$, then $C_c(v, \lambda_p,$

λ_q)=0. Otherwise, if either λ_p or λ_q is not available, then $C_c(v, \lambda_p, \lambda_q)=\infty$. Note that the wavelength conversion cost between any two different available wavelengths is the same in this paper, i.e., $C_c(v, \lambda_p, \lambda_q)=1$. Hence, it's obvious that the total cost of wavelength conversions can be reduced to the number of wavelength conversions.

Let $r(s:D)$ be a multicast request, where s is the source node and D is the set of all destination nodes. And the route from the source to each of the destinations is represented to be a multicast tree $T(V_T, E_T)$. In the tree T, let e_v be the input link of node v, $A(e_v)$ be the available wavelength set on the input link of node v excepting the rout node, $Out(v)$ be the set of output links of node, and $Q(v)$ be the set of child nodes of node v.

Given a multicast tree T and the available wavelength set $A(e_v)$ of all links in the tree, a wavelength assignment of T is defined as a function F:E **a** M, such that, for each $e_v \in E_T$, $F(e_v) \in A(e_v)$. Therefore, the multicast wavelength assignment is used to assign one appropriate wavelength $F(e_v) \in A(e_v)$ on each link of the tree.

For each non-root node v in the tree T, we define a cost function $C_c(T_v, \lambda)$ to be the number of wavelength conversions needed in the sub-tree rooted at v, assuming wavelength λ is assigned on the input link of node v. For each leaf node v, we set $C_c(T_v, \lambda)=0$. Hence, the total number of wavelength conversions of the tree can be defined as follows:

$$C_T(F) = \sum_{v \in Q(s)} C_v(l),\ l \in A(v) \qquad (1)$$

2.2. Problem Specification

Based on the above, the MC-WA problem in this paper can be described as follows: given a multicast tree $T(V_T, E_T)$ rooted at node s and available wavelength set $A(e_v)$ on the input link of each non-root node v, the multicast wavelength assignment problem is to assign the wavelength set $F(e_v) \in A(e_v)$ on the input link for each non-root node of the tree, while the total number of wavelength conversions for the tree T, $C_T(F)$, is minimized. Based on Equation (1), the MC-WA problem can be formulated as follows:

$$\text{Min}\, C_T(F) = \text{Min} \sum_{v \in Q(s)} C_v(l),\ l \in A(v) \qquad (2)$$

According to Equation (2), we can find that the total number of wavelength conversions of the tree T is the summation of the number of wavelength conversions of sub-trees that rooted at each child node of the root node.

3. The Proposed Algorithm

3.1. The NGWA Algorithm

In this subsection, we will propose a new multicast wavelength assignment algorithm called NGWA. The objective of the algorithm aims to minimize the total number of wavelength conversions of the multicast tree as few as possible; meanwhile, the number of different wavelengths used is minimized locally. The main parameters in our proposed algorithm are given in Table 1.

Table 1. Parameter and definition.

Parameter	Definition
T	Multicast tree
$r(s:D)$	Multicast request, where s is the source and D is the set of all destination nodes
e_v	Input link of node v
$Out(v)$	Set of output links of node
$Q(v)$	Set of child nodes of node v
$A(e_v)$	Set of available wavelengths on the input link of node v
$H(e_v)$	Candidate wavelength(s) on e_v in the tree
$F(e_v)$	Assigned wavelength on the input link of node v in the tree
$MU(\lambda_i)$	Counter of assigned wavelength λ_i
$Pare(v)$	Parent node of the node v
$C_c(T_v, \lambda)$	Number of wavelength conversions needed in the sub-tree rooted at v
$C_T(F)$	Total number of wavelength conversions for the tree T

The basic steps of the NGWA algorithm are given below.

```
Input: Multicast tree $T$ and available wavelength set $A(v)$
Output: Wavelength assignment for the multicast tree $T$.
Begin
    Let the N nodes of $T$ have a topological order 0,
  1,…, N-1, beginning from the root node.
  //Partial wavelength assignment phase:
  For (i=1 up to N-1) Do
      $v = v_i$.
    If node $v$ is not a leaf node, Then
        Computer the candidate wavelength(s):
        $H(e_v) = A(v'_1) \cap A(v'_2) \cap K \cap A(v'_{|Q(v)|})$,
        where $v'_i \in Q(v)$.
        If $H(e_v) \geq 2$ Then
          Save the set $H(e_v)$ and the node $v$ is
          marked as "uncompleted"
    Else
          $F(e_v) = H(e_v)$, $MU(F(e_v)) = MU(F(e_v)) + 1$,
        the node $v$ is marked as "completed"
      Endif
    Else
        If state of $Pare(v)$ is already marked as "completed" and $|A(e_v)| = 1$ Then
          $F(e_v) = A(e_v)$.
      Endif
    Endif
Endfor
  // Complete wavelength assignment phase:
    For (i= N-1 up to 1) Do
        $v = v_i$ .
        If state of node $v$ is already marked as "uncompleted" Then $F(e_v) = l'$, where $l' \in H(e_v)$, and $MU(l')$ is maximum among all wavelengths currently. $MU(F(e_v)) = MU(F(e_v)) + 1$ .
        Endif
    Endfor
End
```

The whole procedure of the NGWA algorithm is separated into two phases: the partial wavelength assignment phase and the complete wavelength assignment phase. They are depicted as follows, respectively.

1) In the partial wavelength assignment phase, the local optimality strategy is used to computer the set of candidate wavelengths on e_v which is available on the maximum number of output links. If there are more than two candidate wavelengths, the corresponding node $\boldsymbol{v}$ is marked as "uncompleted." Otherwise, it assigns the only wavelength on e_v. Meanwhile, the node $\boldsymbol{v}$ is marked as "completed" and the counter of the corresponding wavelength increases by one. For each leaf node $\boldsymbol{v} \in \boldsymbol{D}$ if the wavelength of input link of parent node of the leaf node $\boldsymbol{v}$ is assigned and the number of available wavelength on e_v is only one, then the only available wavelength is assigned to link e_v.

2) In the complete wavelength assignment phase, the main task is to deal with the nodes that are marked as "uncompleted" according to the order of bottom-up in the tree. Similar to the method of Most-Used [9], it chooses the wavelength that is the most-used in the multicast tree from the candidate wavelengths so as to make full use of the overall wavelength utilization situation on the tree and reduce the number of different wavelengths used.

Theorem 1: The time complexity of NGWA algorithm is no more than $O(N)$, where $N=|V_T|$.

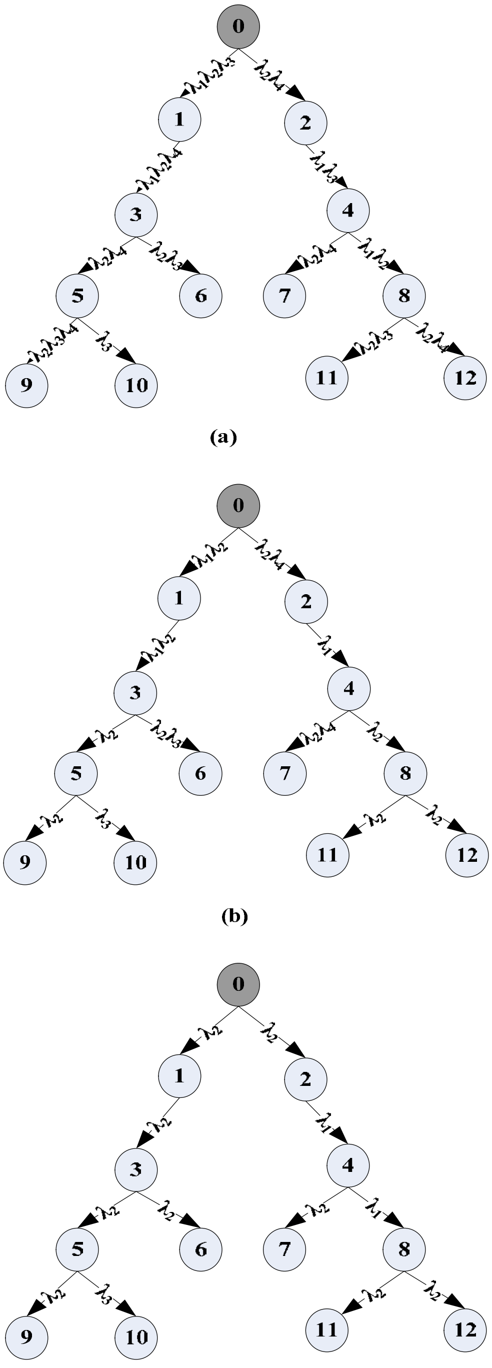
Proof: The time complexity is obvious. In the first phase, the time of spanning all non-root nodes in the tree is $O(N\text{-}1)$, where $N=|V_T|$. In the second phase, the time complexity is same as the first phase. Therefore, the time complexity of NGWA algorithm is no more than $O(N)$, where $N=|V_T|$.

3.2. Illustrative Example

To help further illustrate how the NGWA algorithm works, a multicast tree is given in Figure 1(a). Figures 1(b) and 1(c) depict the two phase's executive results of the NGWA algorithm, respectively. It's clear that the frequency of each wavelength λ_i(i=1,2,3,4) used in the tree is 2, 9, 1, and 0, respectively. And the total number of wavelength conversions is 4.

4. Simulation Results

We carry out a simulation study to see how well the proposed algorithm works, and compare the performance of our proposed NGWA algorithm to the old greedy algorithm proposed in [2] in terms of the average number of wavelength conversions and the average blocking probability. In view of briefness, the old greedy wavelength assignment algorithm is abbreviated to OGWA.



(a)

(b)

(c)

Figure 1. The illustrative examples of (a). A given multicast tree (b). The result of the first phase (c). The result of the second phase.

Our simulation works are carried out on the platform of Network Simulator version 2 (NS2) [10]. The network model and various parameters are set as follows: 1) The network graphs used in the simulations are constructed by using the approach proposed by [11]. Each link is assumed to consist of |*M*| wavelengths. 2) While the multicast trees are built for the fixed multicast connecting requests by using Dijkstra's shortest path algorithm, the multicast trees of the multicast request arriving at random are built dynamically.

For simplicity, the max number of available wavelengths and the multicast group size are abbreviated to *L* and *G* respectively. Note that the multicast group size is used to represent the fraction of nodes that are destinations. If there is no specific declaration, the simulation parameters are configured as follows: |*V*|=200 |*M*|=8 *G*=0.4 and *L*=12.

The first experiment aims at assessing the effect of the max number of available wavelengths and the multicast group size (*G*) on the average number of wavelength conversions of all the multicast trees for each algorithm. The results of the experiments are depicted in Figures 2 and 3. As can be seen, our NGWA algorithm outperforms the OGWA algorithm. This also shows that more available wavelengths imply that it will result in less wavelength conversions.

The second experiment aims at assessing the average blocking performance by varying the multicast group size (*G*). Figure 4 shows the result of the experiment. It can be seen that with the increase of multicast group size, the difference between these algorithms in the average blocking performance is slight. But, compared with the OGWA algorithm, the NGWA algorithm achieves better average blocking performance.

5. Conclusions

In this paper, we study the multicast wavelength assignment (MC-WA) problem in WDM networks with full wavelength conversion capability, and propose a new multicast wavelength assignment algorithm consisting of two phases called NGWA with complexity of *O*(*N*), where *N* is the number of nodes on a multicast tree. Through simulation experiments, it's proved that the proposed algorithm works well and achieves satisfactory performance in terms of the total number of wavelength conversions and the average blocking probability.

6. Acknowledgements

This research was supported by the National High Technology Research and Development Program of P. R. China (No.2009AA01Z217, No.2009AA01Z209) and also supported by the National Natural Science Foundation of P. R. China (No.60872047, No. 60773074).

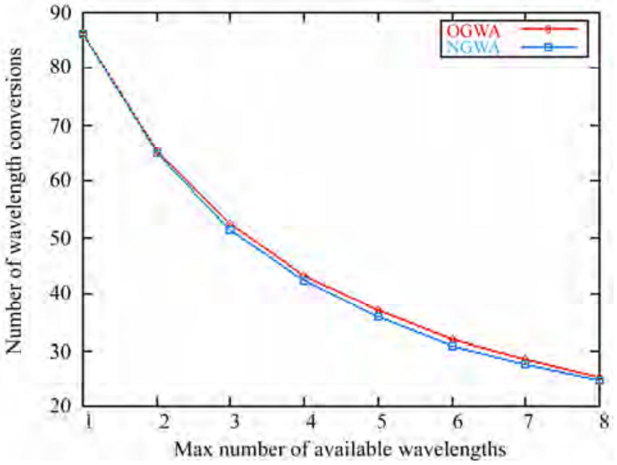


Figure 2. Number of wavelength conversions vs. max number of available wavelengths.

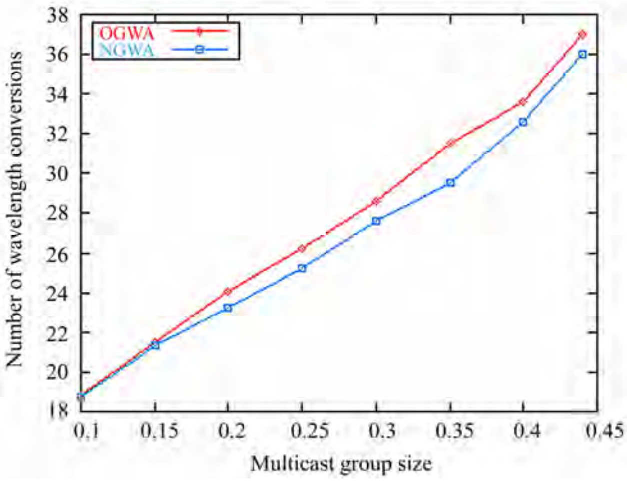


Figure 3. Number of wavelength conversions vs. multicast group size.

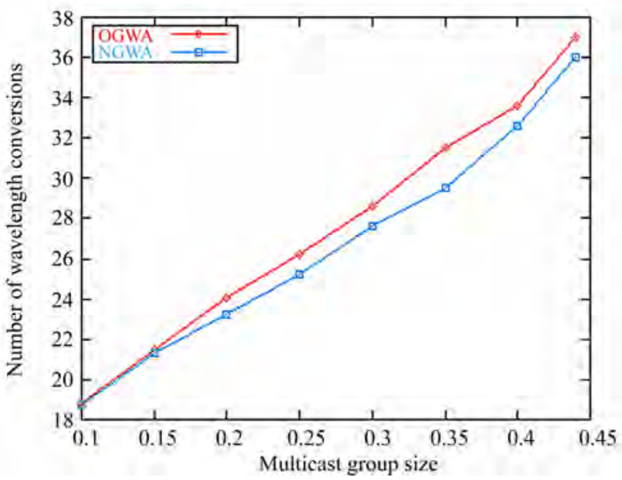


Figure 4. Blocking probability vs. multicast group size.

7. References

[1] Y. Z. Zhou and G. S. Poo, "Optical multicast over wavelength-routed WDM network: A survey," Optical Switching and Networking, Vol. 2, No. 3, pp. 176–197, November 2005.

[2] B. Chen and J. Wang, "Efficient routing and wavelength assignment for multicast in WDM networks," IEEE Journal of Selected Areas Communication, Vol. 20, No. 1, pp. 97–109, January 2002.

[3] G. S. Poo and Y. Zhou, "A new multicast wavelength assignment algorithm in wavelength-routed WDM networks," IEEE Journal of Selected Areas Communication, Vol. 24, No. 4, January 2006.

[4] R. Libeskind-Hadas and R. Melhem, "Multicast routing and wavelength assignment in multi-hop optical networks," IEEE/ACM Transactions on Networking, Vol. 10, No. 5, October 2002.

[5] J. Wang, B. Chen, and R. N. Uma, "Dynamic wavelength assignment for multicast in all-optical WDM networks to maximize the network capacity," IEEE Journal of Selected Areas Communication, Vol. 21, No. 8, pp. 1274–1284, October 2003.

[6] X. H. Jia, D. Z. Du, X. D. Hu, *et al.*, "Optimization of wavelength assignment for QoS multicast in WDM networks," In Proceedings of IEEE Transactions on Communication, Vol. 49, No. 2, pp. 341–350, February 2001.

[7] I. S. Hwang, S. N. Lee, and Y. F. Chuang, "Multicast wavelength assignment with sparse wavelength converters to maximize the network capacity using ILP formulation in WDM mesh networks," Photonic Network Communication, Vol. 12, No. 2, pp. 161–172, August 2006.

[8] Y. W. Chen, and I. H. Peng, "Study of multicast wavelength arrangement for maximizing network capacity in WDM networks with sparse wavelength converters," Photonic Network Communication, Vol. 15, No. 2, pp. 141–152, April 2008.

[9] M. Saad and Z. Luo, "On the routing and wavelength assignment in multi-fiber WDM networks," IEEE Journal of Selected Areas Communication, Vol. 22, No. 9, pp. 1708–1717, June 2004.

[10] The Network Simulator version 2, http://www.isi.edu/nsnam/ns/.

[11] B. M. Waxman, "Routing of multipoint connections," IEEE Journal of Selected Areas Communication, Vol. 6, No. 9, pp. 1617–1622, December 1988.

Int. J. Communications, Network and System Sciences, 2009, 2, 917-926
doi:10.4236/ijcns.2009.29107 Published Online December 2009 (http://www.SciRP.org/journal/ijcns/).

Enhanced Spectrum Utilization for Existing Cellular Technologies Based on Genetic Algorithm in Preview of Cognitive Radio

K. SRIDHARA[1], Aritra NAYAK[2], Vikas SINGH[2], P. K. DALELA[2]
[1]*Member Technology, Government of India, New Delhi, India*
[2]*C-DOT, New Delhi, India*
E-mail: *pdalela@gmail.com*

Received September 2, 2009; *revised October* 21, 2009; *accepted November* 29, 2009

Abstract

This paper attempts to find out the distributed server-based dynamic spectrum allocation (DSA) within liberalized spectrum sharing regulation concept as an alternative to existing regulation based on fixed frequency spectrum allocation schemes towards development of cognitive radio for coverage-based analogy. The present study investigates a scenario where a block of spectrum is shared among four different kinds of exemplary air interface standards i.e., GSM, CDMA, UMTS and WiMAX. It is assumed to offer traffic in an equally likely manner, which occupy four different sizes of channel bandwidths for different air interfaces from a common pooled spectrum. Four different approaches for spectrum pooling at the instance of spectrum crunch in the designated block are considered, viz. channel occupancy through random search, existing regulation based on fixed spectrum allocation (FSA), FSA random and channel occupancy through Genetic Algorithm (GA) based optimized mechanism to achieve desired grade of service (GoS). The comparisons of all the approaches are presented in this paper for different air interfaces which shows up to 55% improvement in GoS for all types of air interfaces with GA-based approach in comparison to existing regulations.

Keywords: DSA, GA, GSM, CDMA, UMTS, WiMAX, Cognitive Radio

1. Introduction

The sophistication possible in a software-defined radio (SDR) [1–3] has now reached the level where each radio can conceivably perform beneficial tasks that help the user, network, and minimize spectral congestion. Radios are already demonstrating one or more of these capabilities in limited ways [4,5]. A simple example is the adaptive digital European cordless telephone (DECT) wireless phone, which finds and uses a frequency within its allowed plan with the least noise and interference on that channel and time slot [6]. Of these capabilities, conservation of spectrum is already a national priority in international regulatory planning. As on date, there are certain rules [4] by which a fixed spectrum is allocated to designated technology, and other technology/service provider cannot use this spectrum. We are interested to investigate this hypothesis in the case of cognitive radio [5,6] i.e., in case of availability of spectrum anywhere, any technology/service provider user can use that to accommodate maximum subscribers within limited spectrum.

As an example of the potential for utilizing the time varying nature of the traffic, we consider four different radio networks: GSM, CDMA, UMTS and WiMAX. We also assume that these radio networks might be used to support different services, e.g. voice telephony on GSM [7], CDMA, broadband internet access along with video streaming on UMTS (for individual subscribers) and WiMAX (mainly for corporate connections). The traffic pattern (and therefore demand for frequency spectrum) seen on each of these networks would vary throughout the day. Example traffic patterns are shown in Figure 1 based on the assumption that voice telephony and corporate connection demands will be high during office time while individual broadband internet subscriber demand will be high before and after office hours. Here GSM traffic variation has been drawn with the help of reference [7] whereas the traffic variation of CDMA, UMTS and WiMAX are drawn based on above assumption.

Here we assumed that a block of spectrum is shared among four different kinds of exemplary air interface standards i.e., GSM, CDMA, UMTS and WiMAX which

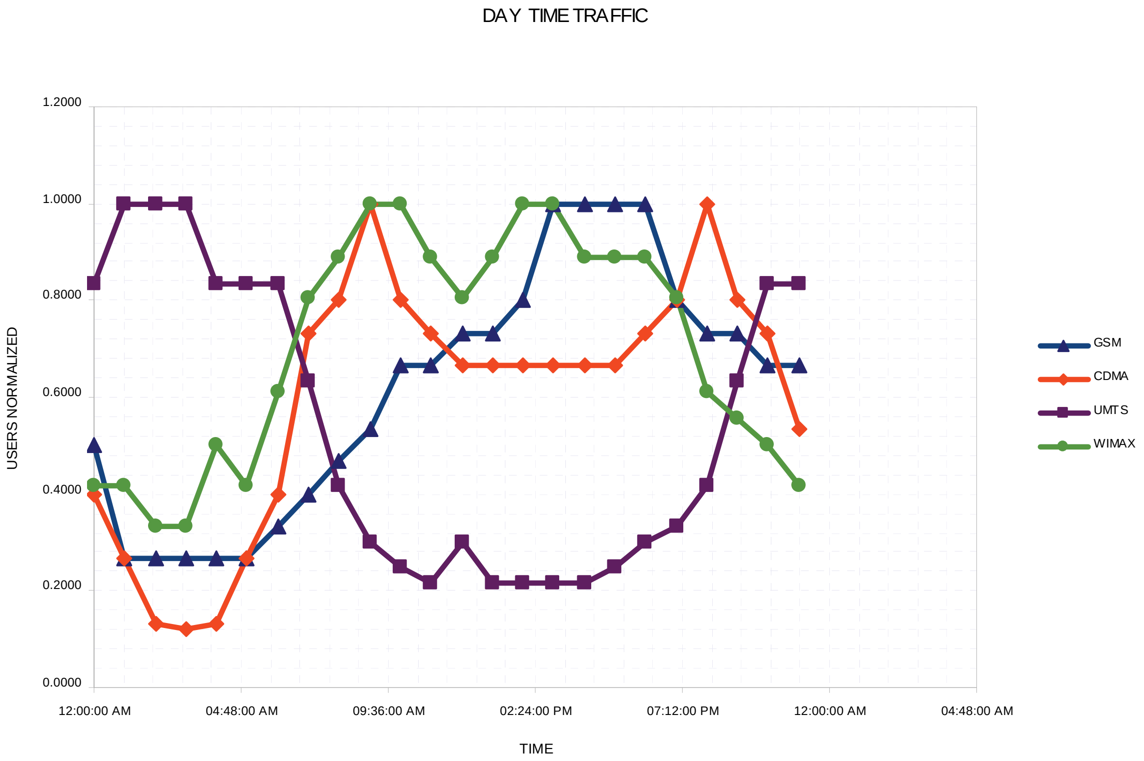


Figure 1. The peak traffic variations of the four types of technologies that share the spectrum over a 24-hour period.

occupies 200 kHz, 1.25 MHz, 5 MHz and 10 MHz frequency bandwidths respectively. The relation among different technologies is spectrum occupancy related for coverage-based analogy that in case of secondary users, i.e., if UMTS allocated spectrum is fully occupied and UMTS subscriber (which will be secondary user for other technology spectrum) wants to grab channel in some other allocated technology spectrum e.g. GSM, then it would also require full 5 MHZ instead of 200 KHz for GSM. The distributed server instead of centralized server-based approach has been taken to reduce computing time. The fixed frequency spectrum has been allocated to these different technologies. Initially, some traffic patterns based on actual traffic load have been assumed for all these four technologies during a day. As per present regulations, initially as traffic (number of users) increases, the specific technology user tries to grab channel within its allocated frequency spectrum slot and in case of unavailability of frequency resources user would be dropped. The fixed spectrum allocation (FSA) does have some disadvantages. For example, most communication networks are designed to cope with a certain maximum amount of traffic. The dimensioning of the network is based on the "busy hour", which is the time of the peak use of the network. If this network uses its allocated spectrum fully during this hour, then the rest of the time the spectrum is not fully utilized. A similar pattern is also seen with other services, hence, with the help of dynamic spectrum allocation the dropped users can be reduced and hence GoS can be enhanced. This paper leads through the technologies and regulatory considerations to support spectrum management and optimizations that raise SDR's capabilities and make it a cognitive radio. Many technologies have come together to result in the spectrum efficiency and cognitive radio technologies may be considered as an application on top of a basic SDR platform. In the present paper, biologically inspired Genetic Algorithm (GA) [8–10] based dynamic spectrum access (DSA) [4,11,12], with distributed server based approach [13], as one of its intended applications have been proposed to reduce blocked users i.e., GoS by utilizing unutilized spectrum. This paper is organized as follows. In Section 2, simulation model for GoS of existing regulation based on FSA and other approaches which includes GA-based optimized mechanism of channel grabbing has been explained along with traffic model. In Section 3, a brief review to GA and its applicability in simulation has been explained. In Section 4, simulation results are shown and Section 5 concludes this study.

2. Simulation Model

In this section, the concepts behind simulation of GoS with time for 4 different scenarios i.e. Fixed Spectrum Allocation (FSA), FSA random (FSA_RAND), total random (TOT_RAND) and GA optimized mechanism have been elaborated.

The random spectrum allocation situation is analogous to road traffic control for multiple lanes dedicated to a particular type of vehicle philosophy i.e., where allocation of frequency spectrum is fixed for each technology within a certain frequency range. The different sizes of vehicles can be compared to different channel bandwidth requirements for the different air interface standards. A deregulated regime is analogous to having a traffic circle at the road junction wherein every vehicle finds a suitable slot proportional to its size in the circle; the circle itself represents the available spectrum pool. It is assumed that a pool of F=120MHz of spectrum is available for four different bandwidths, viz. B1=0.2MHz, B2=1.25MHz, B3=5MHz and B4=10MHz, all operating in Time Division Duplex mode, i.e., pairing of frequencies for uplink and downlink is not considered. The entire band of 120MHz is quantized in steps of f=0.05MHz for simulation purposes.

The above analogy is repeated with fixing slots of frequency spectrum for different technologies. At the time of congestion pertaining to one technology, the additional amount of required frequency spectrum can be borrowed from other technology slots if they have spare frequency spectrum at that moment. The optimization of bandwidth borrowing and lending is proposed by GA i.e., introducing regulations based on DSA with GA. Four different approaches for spectrum pooling at the instance of spectrum crunch in the designated block are considered. In first approach, channel occupancy through random search in complete pooled frequency spectrum is simulated. This is done by allocating chunks of the quantized spectrum to the various users of the four technologies. For example, a GSM user gets four blocks of 50 KHz i.e., 200 KHz for a call but this allocation is done on the basis of a random channel grabbing where the channel to be grabbed is generated by a random generator. In second approach, the channel occupancy through existing regulations based on fixed spectrum allocation (FSA) is simulated. This is done by allocating users in their fixed spectrums one after the other is a sequential manner until space runs out for new users on which we simply do a sequential scan to find if there is any empty space to accommodate the new user else the call is dropped. In third approach, FSA random i.e., allocation of resources to different technologies in the designated slots only through randomized search is simulated. For this scheme of allocation, the users are allocated space only in their respective spectrums just like FSA the difference being that the users grab channels within the spectrums allocated with the help of a randomly generated channel number. Lastly the channel occupancy in designated slots through Genetic Algorithm (GA) based optimized mechanism is simulated to achieve the desired grade of service (GoS). The scheme of which will be explained in Section 3. The comparisons of all the four approaches for individual and combined traffic are presented in Section 4.

Traffic Model: For the simulation purpose we assume a perfect channel (either idle or busy). Poisson random process [14] is used to model the arrival traffic with rate λ and the inter arrival time is negative-exponentially distributed with mean $1/\lambda$. The hold time duration is generated from Gaussian random process [14] which is negative-exponentially distributed with mean $1/\mu$. It is assumed that there is some time spent for spectral scanning and this is required to be less than the inter arrival rate.

3. A Brief Review to GA and its Applicability in Simulation

A Genetic Algorithm (GA) [10] is a search algorithm based on the principles of evolution and natural genetics. It combines the exploitation of past results with the exploration of new areas of the search space. By using survival of the fittest techniques combined with a structured yet randomized information exchange, a GA [10] can mimic some of the innovative flair of human search.

In our case, the GA [10] is used to maximize the spectral utilization by using the least bandwidths to create spectrum opportunities for competing users in the spectrum. We can describe GA [10] to find solution for blocked users as follows. For example, in a network with d users, the number of different ways (Γ) [15] these users can be allocated to a spectrum bandwidth (assuming reuse factor =1) without repetition can be computed using the following equation.

$$\Gamma = \sum_{r=1}^{d} \binom{d}{r} \tag{1}$$

$$\binom{d}{r} = \frac{d!}{r!(d-r)!} \tag{2}$$

This would lead to a total of $\mathbf{r}^{m}$ [15] different combinations for the m spectrum bands. For instance, a system with blocked users d=15 and four spectrum bands for different technologies (m=4) would have approximately 2^{56} possible allocations and to find the optimal solution, exhausting all combinations would not be efficient, as a processor checking one billion solutions per second requires approximately 2.3 years to analyze all permutations. Therefore analytical methods may not be suitable for such type of problems. Thus, GA [10] has been successfully applied to this class of combinatorial optimization problems.

Simplicity of operation and power of effect are two main attractions of the GA [10] approach. The effectiveness of the GA [10] depends upon an appropriate mix of exploration and exploitation. Three operators to achieve this are selection, crossover, and mutation [10]. GA has parameters and variables to control the algorithm. There are evolution operation, genetic operations and parameter settings in GA. First evolution operation is selection. Typical methods for selection [10] are encoding scheme, fitness function and seeding. Second genetic operations mainly have crossover and mutation operations. The selection parameters are defined as follows:

1) Encoding Scheme: Encoding scheme [10] is one of the important and crucial aspects to control the performance of Genetic Algorithm [10]. It refers to the method of mapping the problem parameters into a chromosome [10], which decides the nature of being a weak or strong coding. Encoding scheme can be strong or week in terms of its capability to explore the search space and strong encoding scheme exploits more features of the solution domain.

The encoding method is a global approach to the problem. It is global because any chromosome has enough information to describe a set of channel- borrowings for the entire network.

A chromosome is composed in the following way. For every slot of technology in the spectrum, there is a major chromosome slot or super-gene. Within each super-gene, there are four actual genes. These genes represent the four technologies within a spectrum [8].

A Gene is an array of length 10. At the first location of the array, we keep the number of blocked slots and second location has the number of free slots which were formed by quantization of the spectrum. Next four locations contain the data about number of slots borrowed from other technologies and last four locations contain information about slots lent to other technologies. Thus the super-gene is formed as matrix of order 4 X 10 where each row of the gene represents one particular technology.

Gene Structure

1	2	3	4	5	6	7	8	9	10

The allocated spectrum for different technologies in the present study is in the ratio of 1:2:4:5 for GSM, CDMA, UMTS, and WiMAX respectively. However, to show results independent from allocated frequency, spectrum slots for individual technologies GoS has been evaluated.

2) Fitness Function: The fitness function [10] is measuring mechanism to rank the quality of a chromosome. It serves the only link between the problem and algorithm to search the optimal solutions.

$$\begin{aligned} fitniess = \sum \alpha * accommodated + \\ \sum \beta * unserviced + \\ \sum \mu * (\exp ected\ congestion * serviced) \end{aligned} \qquad (3)$$

where α, β, μ are constant values to suit the environment; such that, α ε {w1, w2, w3} and w1>w2>w3; β<0 and μ<0 for all cases.

accommodated is the measure of whether the given number of users can be serviced by the free space and depending upon this α takes values from w1, w2 and w3.

unserviced is the number of users blocked during the congestion and *expected congestion* is calculated by working out possibly how much traffic is going to arrive in each band and finding a ratio to this expected traffic and the maximum capacity of the band.

3) Seeding: We have presented the solution having initial gene to zero which is the first step of seeding [10]. While with channel-borrowing heuristic, the initial gene can be stated with one of the possible solutions and then the genetic flow is operated to get the best solution. Definitely with the later approach, the numbers of iterations to converge to a solution are decreased. This approach is applied with the improved pluck operation [9].

The genetic operations are defined as follows:

1) Crossover: Crossover [10] is the most important function in GA [10], which produces children as new chromosomes from the process of combining two chromosomes (parents). The operation of crossover may gives the children (new chromosomes) with better fitness as it takes best attributes from both the parents [10] we have used single point crossover as shown below.

```
First matrix
         | cut point
 1 2 3 4   5 6 7 8 9 5
 2 3 4 3   4 8 7 3 5 8
 5 6 7 5   6 4 3 1 4 9
 9 9 3 7   2 6 1 5 3 2

Second matrix
         | cut point
 4 5 6 5   7 4 6 5 7 9
 4 3 5 8   7 9 4 7 9 8
 1 2 2 6   6 3 4 8 9 3
 5 3 9 3   7 6 6 2 8 1
```

The offspring or children [10] generated from two parental matrices are:

```
offspring 1
 4 5 6 5 7 6 5 8 5 9
 4 3 5 8 7 4 8 3 5 8
 1 2 2 6 6 3 4 1 4 9
 5 3 9 3 6 2 2 5 3 1

offspring 2
 1 2 3 4 5 4 6 7 7 9
 2 3 4 3 4 9 8 7 9 7
 5 6 7 5 6 4 3 8 3 9
 9 9 3 7 7 2 6 6 8 1
```

This crossover function facilitates not to copy entire second half of the second matrix to the first matrix, while it retains the common elements of the parental matrix, which is essential for borrowing policy to preserve necessary information or the parent characteristics of the children. However if row wise second half of the parental matrices are entirely different then this crossover function serves as normal crossover function discussed initially.

2) Mutation: The mutation [10] operator is responsible for diversity into a population. In the first phase of results, we are using the normal mutation operator with high probability to provide diversity in genes as we have started with initial gene zero in seeding. However, before swapping the child to next population, we check the

needy technologies that require borrowing with correct mutations applied by means of a simple probability.

3.1. GA Application in the Simulation

In this simulation, whenever the users find that their respective allocated spectrum are fully occupied, then they try to grab empty spectrum in the other allocated technology spectrum bands. This is where GA is employed to select an empty space among the many available by repeated iterations based on crossovers and seeding so that due to the borrowing of spectrum by another technology user i.e., the secondary users [6] quality of service (QoS) of primary users [6] are not affected. The allocation is done by first randomly seeding the population with all possible solutions and then applying fitness function to select only the healthy children [10] or fit solutions. The crossover and mutation is applied to further consolidate the results and possible scheme of allocation of unutilized spectrum based on the fitness function mentioned earlier, which keeps in mind the possible traffic pattern and future trends to evolve an optimal solution. Also before the application of GA, it is checked if a specific technology spectrum which has been full and cannot accommodate the primary users, a spectrum search scanning [6] has been performed to find out the secondary users residing in this specific spectrum slot and terminates them to create space for the primary users before the application of GA. The termination is done on the basis of least number of calls to be terminated and in order of time of residence in the spectrum.

The flowchart shown in Figure 2 describes the steps employed to obtain the optimal solutions using GA [10] for DSA by using its basic operators. As shown in the flowchart, the results are obtained by following a number of iterations which is fixed as a constant named MAX ITERATION. In each of the iterations, all the basic steps are followed until we obtain an optimal solution.

4. Results

This section shows the comparison of GoS among fixed frequency spectrum allocation (FSA) which is existing regulation, FSA random (FSA_RAND), total randomized allocation (TOT_RAND) and GA-based DSA scheme which is proposed regulation for GSM, CDMA, UMTS and Wi-MAX along with mixed traffic of all these technologies.

Figure 3 shows the case of GoS of GSM for 120 MHz common spectrum for all the technologies during a span

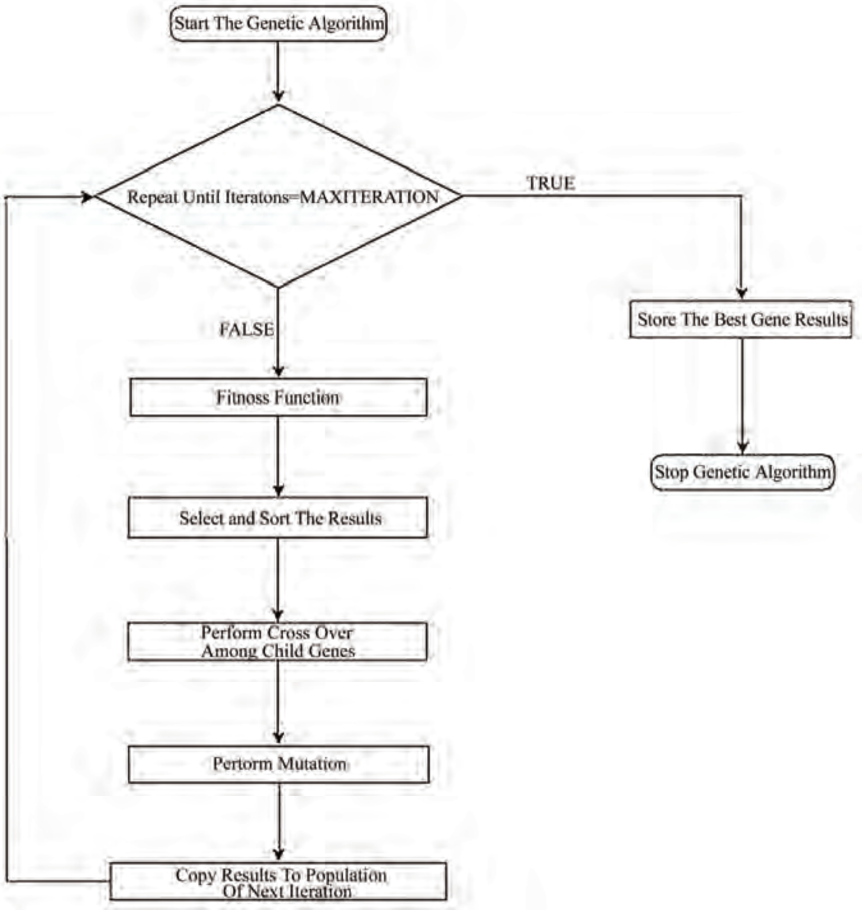

Start The Genetic Algorithm
Repeat Until Iteratons=MAXITERATION
TRUE
FALSE
Fitnoss Function
Select and Sort The Results
Perform Cross Over
Among Child Genes
Pertorm Mutation
Copy Results To Population
Of Next Iteration
Store The Best Gene Results
Stop Genetic Algorithm

Figure 2. Flow chart of GA applicable to DSA in the present case.

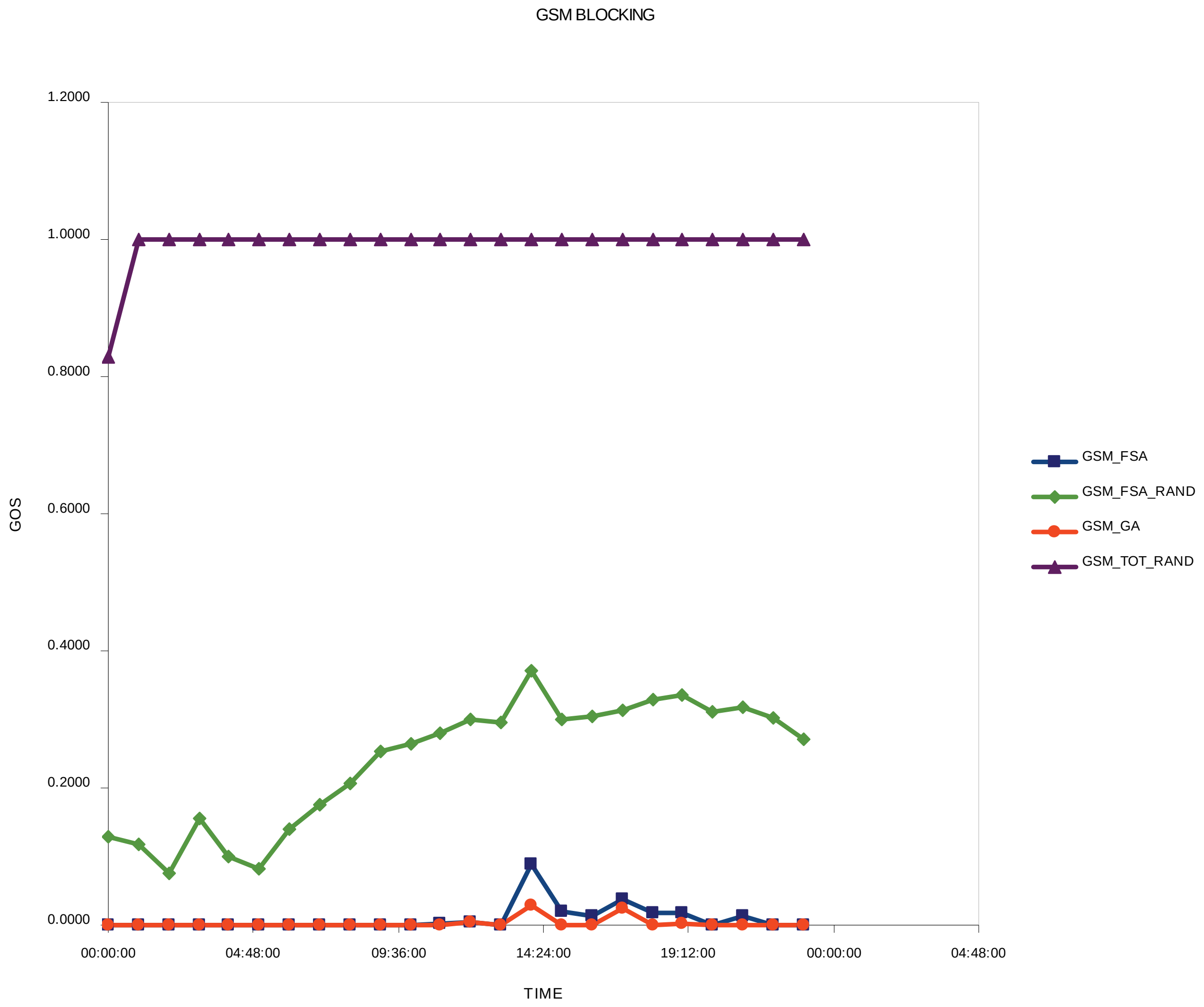


Figure 3. Comparison of GoS for GSM (0.2 MHz) users among FSA, FSA_RAND, GA-based DSA scheme and TOT_RAND over duration of 24 hours of the day.

of one day. The GSM_FSA graph shows GoS for GSM when different technologies users grab the channel slots in the allocated frequency spectrum slots only by using FSA which is existing regulation. The GSM_TOT_RAND graph shows the GoS for GSM user when there is no regulation and any technology user can grab the channel at random position if it is available. The GSM_FSA_RAND graph is of the randomized channel grabbing FSA where the spectrum is divided among the various technologies but the channel grab is at any random position provided it is vacant. The GSM_GA graph shows the GoS for GSM user with GA optimized DSA algorithm which is proposed regulation. It is seen that GA optimized DSA algorithm is always better than all other cases. Also it has approximately up to 72% improvement in comparison to FSA which is the best performing algorithm amongst the three. If the mean GoS is taken for the given time then a maximum of 67% improvement is noticed neglecting the cases where DSA gives zero blocking.

Figure 4 shows the case of GoS of CDMA for 120 MHz common spectrum for all the technologies during a span of one day. The CDMA_FSA graph shows GoS for CDMA when different technologies users grab the channel slots in the allocated frequency spectrum slots only by using FSA which is existing regulation. The CDMA_TOT_RAND graph shows the GoS for CDMA user when there is no regulation and any technology user can grab the channel at random position if it is available. The CDMA_FSA_RAND graph is of the randomized channel grabbing FSA where the spectrum is divided among the various technologies but the channel grab is at any random position provided it is vacant. The CDMA_GA graph shows the GoS for CDMA user with GA optimized DSA algorithm which is proposed regulation. It is seen that GA optimized DSA algorithm is always better than all other cases. Also it has approximately up to 72% improvement in comparison with FSA which is the best performing algorithm amongst the three. If the mean GoS is taken for the given time then a maximum of 95% improvement is noticed neglecting the cases where DSA gives zero blocking.

Figure 5 shows the case of GoS of UMTS for 120 MHz common spectrum for all the technologies during a span of one day. The UMTS_FSA graph shows GoS for UMTS when different technologies users grab the channel slots in the allocated frequency spectrum slots only by using FSA which is existing regulation. The UMTS_TOT_RAND graph shows the GoS for UMTS user when there is no regulation and any technology user can grab the channel at random position if it is available. The UMTS_FSA_RAND graph is of the randomized channel grabbing FSA where the spectrum is divided among the various technologies but the channel grab is at any random position provided it is vacant. The UMTS_GA graph shows the GoS for UMTS user with GA optimized DSA algorithm which is proposed regulation. It is seen that GA optimized DSA algorithm is always better than all other cases. Also it has approximately up to 30% improvement in comparison to FSA which is the best performing algorithm amongst the three. If the mean GoS is taken for the given time then a maximum of 60% improvement is noticed neglecting the cases where DSA gives zero blocking.

Figure 6 shows the case of GoS of WiMAX for 120 MHz common spectrum for all the technologies during a span of one day. The WiMAX_FSA graph shows GoS for UMTS when different technologies users grab the channel slots in the allocated frequency spectrum slots only by using FSA which is existing regulation. The WiMAX_TOT_RAND graph shows the GoS for WiMAX user when there is no regulation and any technology user can grab the channel at random position if it is available. The WiMAX_FSA_RAND graph is of the randomized channel grabbing FSA where the spectrum is divided among the various technologies but the channel grab is at any random position provided it is vacant. The WiMAX_GA graph shows the GoS for WiMAX user with GA optimized DSA algorithm which is proposed regulation. It is seen that GA optimized DSA algorithm is always better than all other cases. Also it has approximately up to 15% improvement in comparison with FSA which is the best performing algorithm amongst the three. If the mean GoS is taken for the given time then a maximum of 70% improvement is noticed neglecting the cases where DSA gives zero blocking.

Figure 7 shows the case of GoS of all four technologies for 120 MHz common spectrum for all the technologies

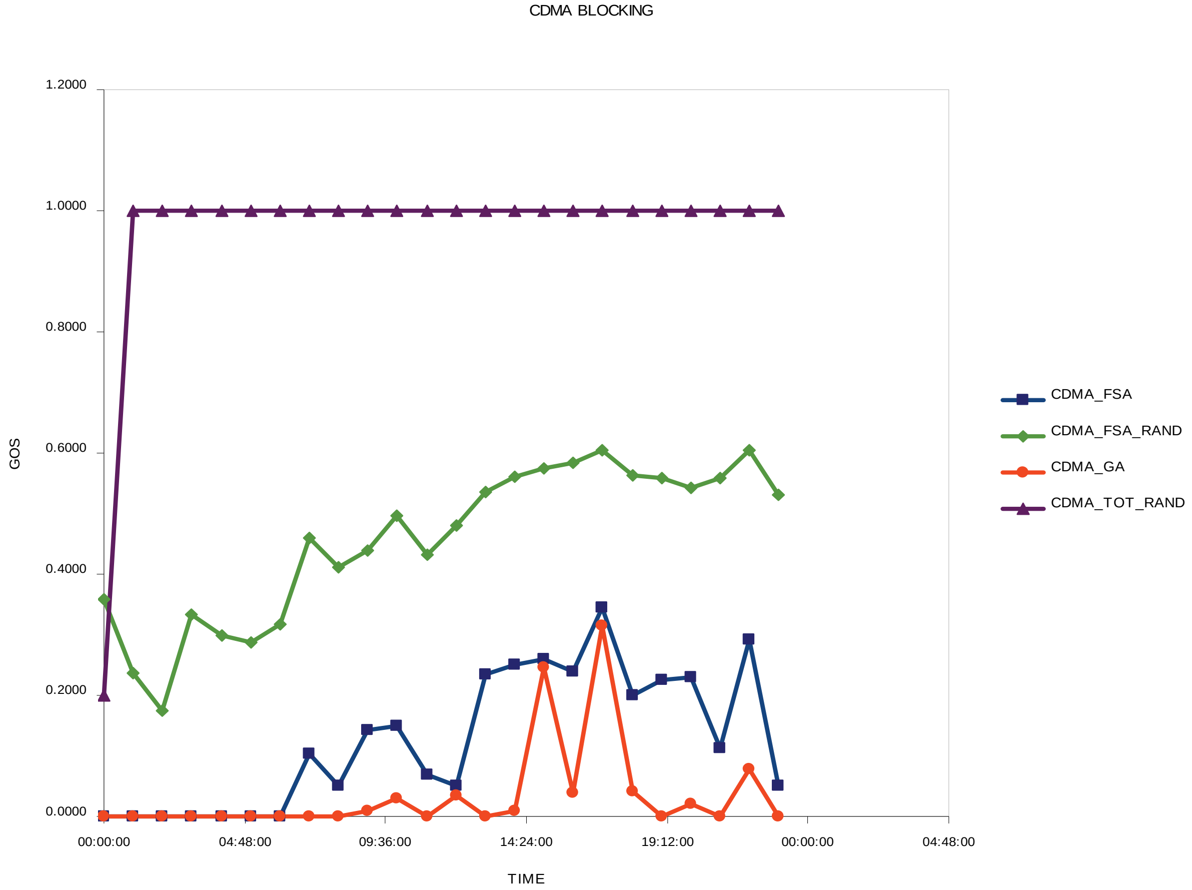


Figure 4. Comparison of GoS for CDMA (1.25 MHz) users among FSA, FSA_RAND, GA-based DSA scheme and TOT_RAND over duration of 24 hours of the day.

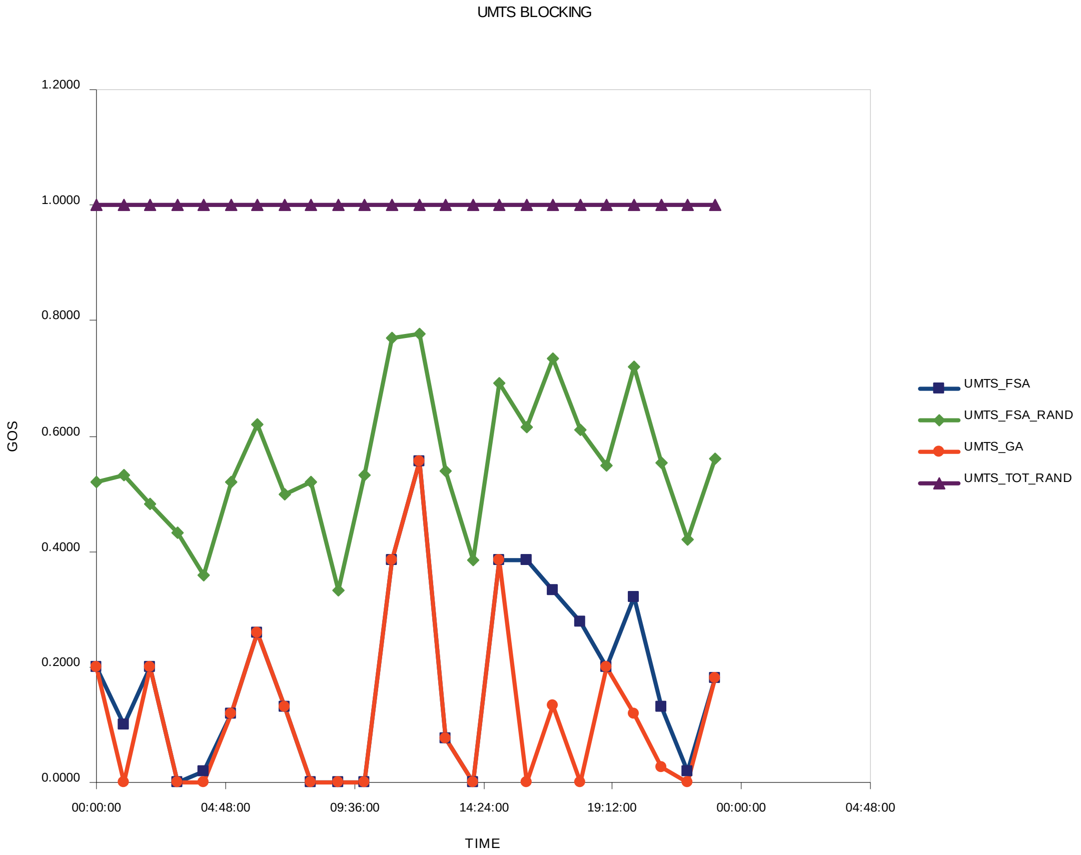


Figure 5. Comparison of GoS for UMTS (5 MHz) users among FSA, FSA_RAND, GA-based DSA scheme and TOT_RAND over duration of 24 hours of the day.

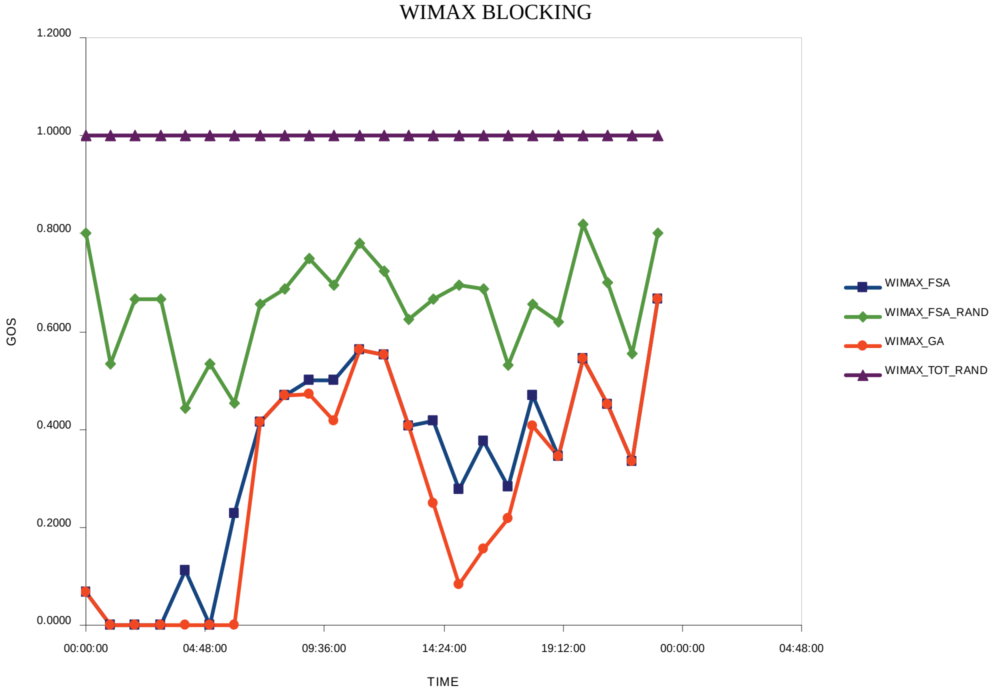


Figure 6. Comparison of GoS for WiMAX (10 MHz) users among FSA, FSA_RAND, GA-based DSA scheme and TOT_RAND over duration of 24 hours of the day.

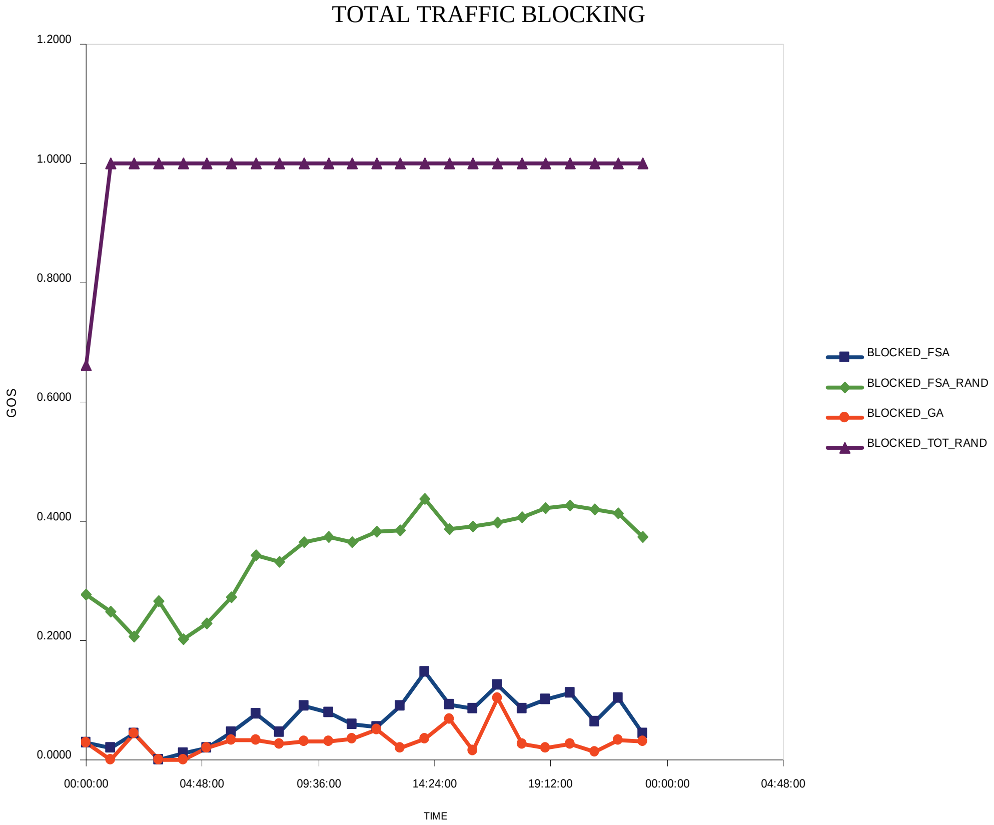


Figure 7. Comparison of GoS for all technologies mixed traffic users among FSA, FSA_RAND, GA-based DSA scheme and TOT_RAND over duration of 24 hours of the day.

during a span of one day. The BLOCKED_FSA graph shows GoS for UMTS when different technologies users grab the channel slots in the allocated frequency spectrum slots only by using FSA which is existing regulation. The BLOCKED_TOT_RAND graph shows the GoS for the combined user traffic when there is no regulation and any technology user can grab the channel at random position if it is available. The BLOCKED_FSA_RAND graph is of the randomized channel grabbing FSA where the spectrum is divided among the various technologies but the channel grab is at any random position provided it is vacant. The BLOCKED_GA graph shows the GoS for combined user traffic with GA optimized DSA algorithm which is proposed regulation. It is seen that GA optimized DSA algorithm is always better than all other cases. Also it has approximately up to 55% improvement in comparison to FSA which is the best performing algorithm amongst the three. If the mean GoS is taken for the given time then a maximum of 80% improvement is noticed neglecting the cases where DSA gives zero blocking.

5. Conclusions

In this study, an attempt is made to analyze the impact of our proposed GA based optimized DSA mechanism for spectrum utilization by comparing GoS for existing FSA based regulation with proposed mechanism. Four prominent commonly used technologies are used for the simulations which occupy different bandwidths to analyze the impact of our proposed Genetic Algorithm based solution which uses aspects of cognitive radio for improving the overall GoS of a shared cellular spectrum scenario. The simulation results can be utilized to justify the need of regulatory approach in case of liberalized spectrum sharing in the present cellular network spectrum in the preview of cognitive radio. As it can be seen from results in Figures 3 to 7 that GA based algorithm enhances the GoS when compared to the present allocation scheme. The maximum value of improvement in GoS for GSM, CDMA, UMTS, WiMAX and mixed traffic are 72%, 72%, 30%, 15% and 55% respectively. The improvement in GoS for GSM, CDMA, UMTS, WiMAX and mixed traffic around mean values are 67%, 95%, 60%, 70% and 80% respectively. This study shows some sort of regulatory based approach within liberalized spectrum sharing concept is required to allow all users to get equal chance of utilizing the spectrum and getting a better GoS for the overall traffic.

6. References

[1] M. P. Olivieri, G. Barnett, A. Lackpour, A. Davis, and P.

Ngo, “A scalable dynamic spectrum allocation system with interference mitigation for teams of spectrally agile software defined radios,” New Frontiers in Dynamic Spectrum Access Networks, pp. 170–179, November 2005.

[2] J. Mitola III, “Cognitive radio an integrated agent architecture for software defined radio,” Dissertation, Doctor of Technology, Royal Institute of Technology (KTH), Sweden, May 2000.

[3] Paul Burns, “Software defined radio for 3G,” Artech House, Inc., 2003.

[4] J. Hwang and H. Yoon, “Dynamic spectrum management policy for cognitive radio: An analysis of implementation feasibility issues,” New Frontiers in Dynamic Spectrum Access Networks, 3rd IEEE Symposium, Digital Object Identifier, pp. 1–9, October 2008.

[5] S. Haykin, “Cognitive radio: Brain-Empowered wireless communications,” IEEE Journal on Selected Areas in Communications, Vol. 23, No. 2, February 2005.

[6] Bruce A. Fette., editor, “Cognitive radio technology,” Elsevier Inc., 2006.

[7] S. Almeida, J. Queijo, and L. M. Correia, “Spatial and temporal traffic distribution models for GSM,” Vehicular Technology Conference, Vol. 1, pp. 131–135, September 1999.

[8] A. Y. Zomaya, Senior Member, IEEE, and Michael Wright, “Observations on using genetic-algorithms for channel allocation in mobile computing,” IEEE Transactions on Parallel and Distributed Systems, Vol. 13, No. 9, September 2002.

[9] S. S. M. Patra, K. Roy, S. Banerjee, and D. P. Vidyarthi, “Improved genetic algorithm for channel allocation with channel borrowing in mobile computing,” IEEE Transactions on Mobile Computing, Vol. 5, No. 7, July 2006.

[10] M. Melanie, “An introduction to genetic algorithm,” MIT press, Cambridge, 1999.

[11] D. Maldonado, B. Le, A. Hugine, T. W. Rondeau, and C. W. Bostian, “Cognitive radio applications to dynamic spectrum allocation: A discussion and an illustrative example,” New Frontiers in Dynamic Spectrum Access Networks, First IEEE International Symposium, pp. 597–600, November 2005.

[12] P. Leaves, S. Ghaheri-Niri, R. Tafazolli, L. Christodoulides, T. Sammut, W. Staht, and J. Huschke, “Dynamic spectrum allocation in a multi-radio environment: Concept and algorithm,” 3G Mobile Communication Technologies, Second International Conference on (Conference Publication No. 477), pp. 53–57, March, 2001.

[13] J. Zhao, H. T. Zheng, and G. H. Yang, “Distributed coordination in dynamic spectrum allocation networks,” New Frontiers in Dynamic Spectrum Access Networks, pp. 259–268, November 2005.

[14] Viswanathan and Thiagarajan, “Telecommunication switching systems and networks,” Prentice-Hall, New Delhi, 1992.

[15] D. Thilakawardana, K.Moessner, and R.Tafazolli, “Darwinian approach for dynamic spectrum allocation in next generation systems,” IET Communications, Centre for Communication Systems Research, University of Surrey, Guildford, UK, 2008.

Int. J. Communications, Network and System Sciences, **2009, 2, 805-926**
Published Online December 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

TABLE OF CONTENTS

Volume 2 Number 9 December 2009

ISSN: 1913-3715

9771913371005 13