International Journal of

Communications, Network and System Sciences

ISSN: 1913-3715

Vol. 1, No. 1, February 2008

www.SRPublishing.org/journal/ijcns/

Scientific
Research
Publishing

Journal Editorial Board

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

I. J. Communications, Network and System Sciences. **2008; 1: 1-103**
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

TABLE OF CONTENTS

Volume 1 **February 2008**

Multiband Scheduler for Future Communication Systems
K. DOPPLER, C. WIJTING, T. HENTTONEN, K. VALKEALAHTI…… 1

Adaptive Throughput Optimization in Downlink Wireless OFDM System
Y. FAKHRI, B. NSIRI, D. ABOUTAJDINE, J. VIDAL…..…… 10

A Discrete Newton's Method for Gain Based Predistorter
X.C. LIN, M.L. JIN, A.F. LIU…….……..…….... 16

Particle Filtering with Multi Proposal Distributions
F.S. WANG, Q.J. ZHAO, Y.B. ZHANG, L.Q. ZHANG……..……..... 22

PAPR Reduction Scheme with SOCP for MIMO-OFDM Systems
B. RIHAWI, Y. LOUET……..... 29

Any Resource Sharing via HWN* Routing
C. SHEN, S. REA, D. PESCH……..…….…. 36

Influence of the Limited Retransmission on the Performance of WLANs Using Error-Prone Channel
H.M. ALSABBAGH, J.P. CHEN, Y.Y. XU……..…... 49

Exploiting Geometric Advantages of Cooperative Communications for Energy Efficient Wireless Sensor Networks
I. AHMED, M.G. PENG, W.B. WANG……..…... 55

An Energy-aware Hierarchical Architecture Design Scheme
L.F. YUAN, Y. XU, X. DU, W.Q. CHENG……..…… 62

Efficient DPA Attacks on AES Hardware Implementations
Y.HAN, X.C. ZOU, Z.L. LIU, Y.C. CHEN……..……..... 68

QoS in Node-Disjoint Routing for Ad Hoc Networks
L. LIU, L. CUTHBERT……..…… 74

A Predicted Region based Cache Replacement Policy for Location Dependent Data in Mobile Environment
A. KUMAR, M. MISRA, A.K. SARJE……..……... 79

Advanced Localization of Mobile Terminal in Cellular Network
M. ANISETTI, C.A. ARDAGNA, V. BELLANDI, E. DAMIANI, S. REALE…….……..…. 95

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.SRPublishing.org) is published quarterly by Scientific Research Publishing, Inc. 3306 Apple Grove CT, Herndon, VA 20171, USA.

E-mail: jcnss@srpublishing.org

Subscription rates: Volume 1 2008

Print: $50 per copy.

Electronic: free, available on www.SRPublishing.org.

To subscribe, please contact Journals Subscriptions Department, E-mail: jcnss@srpublishing.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: jcnss@srpublishing.org

Reprints (minimum quantity 100 copies)

Reprints Co-ordinator, Scientific Research Publishing, Inc. 3306 Apple Grove CT, Herndon, VA 20171, USA.
E-mail: jcnss@srpublishing.org

COPYRIGHT

PRODUCTION INFORMATION

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

I. J. Communications, Network and System Sciences. 2008; 1: 1-103
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

Multiband Scheduler for Future Communication Systems

Klaus DOPPLER, Carl WIJTING, Tero HENTTONEN, Kimmo VALKEALAHTI
Radio Communications CTC, Nokia Research Center, Helsinki, Finland
P.O. Box 407, FI-00045 NOKIA GROUP, Finland
E-mail: {Klaus.Doppler, Carl.Wijting, Tero.Henttonen, ext-Kimmo.Valkealahti}@nokia.com

Abstract

Operation in multiple frequency bands simultaneously is an important enabler for future wireless communication systems. This article presents a new concept for scheduling transmissions in a wireless radio system operating in multiple frequency bands: the Multiband Scheduler (MBS). The MBS ensures that the operation in multiple bands is transparent to higher network layers. Special attention is paid to achieving low delay and latency when operating the system in the multiband mode. In particular, we propose additions to the ARQ procedures in order to achieve this. Deployment details and assessment results are presented for two multiband deployment scenarios. The first scenario is operation in a spectrum sharing context where multiple bands are used: one dedicated band for basic service and one shared extension band for extended services. In the second scenario we consider multiband operation in a relay environment, where the two bands have different propagation properties and relays provide extra coverage and capacity in the whole cell.

Keywords: Multiband Operation, Scheduling, Multiband Scheduler, IMT-Advanced, Relaying, Dynamic Spectrum Use, Flexible Spectrum Use, Spectrum Sharing

1. Introduction

Recently the World Radio Communications Conference (WRC-07) has allocated new spectrum for future radio communication systems in different frequency bands (including new allocations in the UHF and C band). Furthermore, spectrum that is currently allocated to second and third generation wireless communication systems may be reused. Consequently, it would be advantageous for such systems to be able to operate in multiple bands. They can use these multiple bands for balancing the load of the networks or for providing required quality of service levels. It is also predicted that some of these bands, here referred to as basic (B) bands, might be dedicated to specific services or operators, and that other bands, the extension (E) bands, might be shared between different operators and/or different services (e.g. mobile communications and fixed satellite services (FSS)). Sharing the spectrum with other radio technologies is seen as a promising technique for increasing the spectrum utilization. However, flexible and fast mechanisms for band transfers are required when the extension bands are (temporarily) unavailable. IMT-Advanced systems are mobile systems that include new capabilities that go beyond those of IMT-2000 (UMTS, WiMAX, etc.). Such systems provide access to a wide range of telecommunications services, including advanced mobile services, supported by mobile and fixed networks, which are increasingly packet-based. IMT-Advanced systems support low to high mobility applications and a wide range of data rates, in accordance with service demands in multiple user environments (100 Mbit/s for high mobility and 1 Gbit/s for low mobility were established as research objectives) [1]. Key innovation areas for these future wireless systems include new concepts such as spectrum sharing [2] and network relays [3]. Several research projects have already addressed these needs; for example the European WINNER project has developed a flexible and scalable radio interface, which covers different domains (local area, metropolitan area, and wide area) with the same radio interface [4].

In this article we introduce the multiband scheduler (MBS), which enables *simultaneously* high flexibility in terms of spectrum use *and* high spectral efficiency—two goals that are difficult to combine. The MBS is located in the medium access control (MAC) system layer, which controls the physical layer, including the radio resource allocation, the spatial processing and the packet scheduling [5]. Multiband operation is made possible by adding an MBS to single-band MACs. It schedules protocol data units (PDU) to the correct band, enables fast

switching to other bands and provides functions for load balancing between the bands. After introducing the concept of the MBS, we illustrate how to apply the MBS to a communication system that shares one band with another wireless system and to relay networks that operate in multiple bands.

Bands with lower center frequency have better propagation properties (larger ranges) than higher frequency bands, and relay nodes (RN) are a cost efficient way to extend the coverage area of the higher bands. RNs can be used in multiband operation for the purpose of balancing the coverage area of different frequency bands. Some of the RNs may operate in only one spectrum band, while the rest operate in several bands. In the latter case an MBS is needed at the base station (BS). In addition, a RN with an MBS can receive from the BS PDUs transmitted in the higher-capacity band and forward them to the user terminal (UT) using any band. However, relays add additional delay to the network, and thus a multiband operation with fast retransmissions and low delays is important.

The importance of delay can be seen from the following simplified delay budget calculation. Assuming a two-hop scenario where we want to achieve an end-to-end delay of up to 20 ms between two peer IP entities (this is the maximum delay for highly interactive services [6]) and assuming that a delay of maximum 10 ms is required over the air interface itself (as in 3GPP-LTE [7]). Allowing one retransmission over two hops and assuming an air interface transmission delay of 1 ms, the delay budget can then be determined as follows: one transmission on each link, plus one feedback message and a retransmission adds up to 4 ms [1] plus additional processing delay, leaving a margin of only 6 ms [8].

The remainder of this paper is organized as follows. Section 2 provides an introduction to the multiband scheduling concept. Section 3 focuses on the concept of hybrid ARQ context transfers. Sections 4 and 5 present the multiband scheduler in a spectrum-sharing context and in a relay context, respectively. Both sections provide numerical results in order to demonstrate the benefits of the multiband scheduler. Finally, section 6 presents the conclusions of the study.

2. Multiband Scheduling

The general framework in which the MBS operates is presented in Figure 1. Higher-layer protocol data units (PDU) arriving at the IP convergence layer (IPCL) are converted into Radio Link Control (RLC) service data units (SDU) after header compression and ciphering. In accordance with the decision of the scheduler, a certain amount of data is selected from the RLC SDU buffer and segmented and/or concatenated, depending on the size of the SDU. A user terminal (UT) identification code, a transmission sequence number, and optionally a CRC code are added. If RLC acknowledged operational mode is used, then an outer end-to-end ARQ is performed at the RLC level. The multiband scheduler schedules the RLC PDUs for transmission in the correct band. The resource scheduler (RS) of each band fetches the RLC PDUs from the buffer and constructs from them transport blocks (TB), which are scheduled for transmission. Hybrid ARQ (HARQ) can be used for improving the transmission of the TBs. The MAC adds a retransmission sequence number to TBs that use HARQ. The RS of each band operates independently and the coordination of the different bands is done exclusively by the MBS. The operation of the different ARQ schemes is depicted in Figure 2.

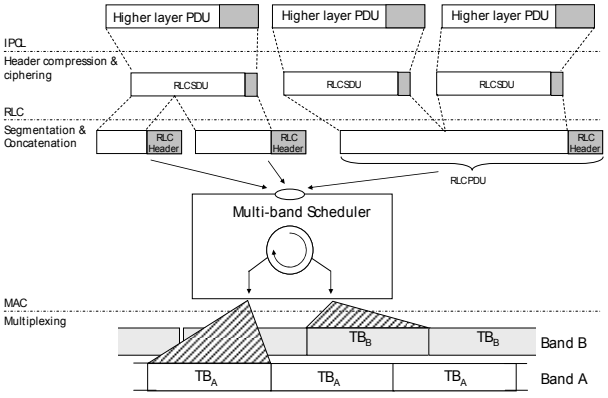


Figure 1. Illustration of the Multi Band Scheduler and the data unit it is working on.

Different Hybrid ARQ protocols exist and two possible approaches can be used: chase combining (retransmission of the whole TB and combining these at the receiver), or incremental redundancy (retransmission of additional redundancy bits, providing the receiver with more information about the TB).

In the case that a certain band is no longer available, the MBS provides a mechanism for transferring the user context from that band to any other band that is still being serviced (Context Transfer Unit). The user context comprises all the information necessary to continue the active services to the user in the other band. This transfer occurs in real time, and is transparent to the end user. During the context transfer, control parameters essential for the transmission are transferred. Adaptation and prioritization of the user data flows may be needed, since the new band might not be able to accommodate all traffic from the band that is no longer available. The timing of the transfer is an important issue; rapid changes when a band becomes unavailable are supported, as well as preparation of context transfers when information is obtained that the availability of a band will change in the (near) future. This is done by the so-called Band

[1] In [8] it was found that in a relaying scenario 95% of the users experience a delay of up to 3ms.

Monitoring functionality in the MBS.

In the multiband architecture considered here, many common functions for the operation in the different bands can be identified. These functions may be shared between the different bands. For example, the same flow ID, UT ID, etc. can be used in both bands, which simplifies the context transfer between the bands. However, some problems related to ARQ and synchronization remain, and need to be solved in order to allow fast switching. These will be addressed in the following sections. In the conceptual discussion above we spoke of switching between two spectrum bands, but a more generalized approach can also be used in the case of a spectrum that is more fragmented.

2.1. Handovers for MBS and Non-MBS Cases

We assume that the network supports mobility via normal inter- and intra-frequency handovers, and that the handover mechanism is hard handover (meaning that there is always a short break in the connection when a handover is done). Further, we assume that handover decisions are done by the network and that only UTs perform measurements and that the UTs keep the serving BS informed of the measured values and of triggered events (such when the signal of a neighbor BS has become stronger than that of the serving BS).

A UT performs periodical measurements, both for identifying neighboring BSs and for measuring the signal level of the identified BSs. A measurement report, containing filtered measurement results, is then sent to the serving BS, either periodically or when a network-configured event occurs. Based on the measurement reports, the serving BS decides whether a handover should be performed.

After a handover from the serving BS to a target BS is triggered, the serving BS sends a request to the target BS to confirm that the UT is allowed to do the handover. If the target BS allows the handover, the serving BS sends a handover command to the UT, identifying when and to which BS the handover should be done. Finally, the UT responds to the handover command by sending an acknowledgement to the serving BS, and then breaks the connection to the serving BS at the agreed time and connects to the target BS.

In the non-MBS case a UT needs to perform a handover when switching between the B and E bands. This means that there is always a break in the connection when moving from the B band to the E band, which causes the user throughput to drop to zero for a while. The overall delay caused by the handover is of the order of tens of milliseconds. In our simulations the delay has been assumed to be 20 ms.

In contrast, when an MBS is used the biggest delay is scheduling delay, which is only of the order or few milliseconds. A small additional delay arises from the time needed by the UT for switching bands. This allows for more seamless switching between resources and thus for better load balancing and higher user throughput.

3. Hybrid ARQ Context Transfer

The design of the automatic repeat request (ARQ) mechanism is very important with regard to the speed of context transfers between bands. If ARQ retransmissions have to be finished before switching to another band is allowed, then, depending on the ARQ design, switching delays of up to 20 ms are possible. In order to make fast context transfer possible, we propose the use of an ARQ mechanism that consists of an outer ARQ for RLC SDUs and an inner HARQ for retransmissions of transport blocks [8].

3.1. E2E ARQ (Outer ARQ)

The E2E ARQ is situated on a higher protocol plane than the MBS, and thus context transfers from one band to another do not affect the E2E ARQ process (Figure 2 illustrates this). After a context transfer the new band is briefly not in use, and the multiband scheduler takes this into account in its scheduling decisions.

3.2. HARQ (Inner ARQ)

Each band has an independent HARQ process, and thus a fast context transfer is required for these processes. In order to allow the HARQ process to continue uninterrupted, we propose a mechanism in which the HARQ buffer is transferred between the two bands, as illustrated in Figure 2. The Context transfer Unit of the MBS coordinates the exchange of the HARQ data and parameters.

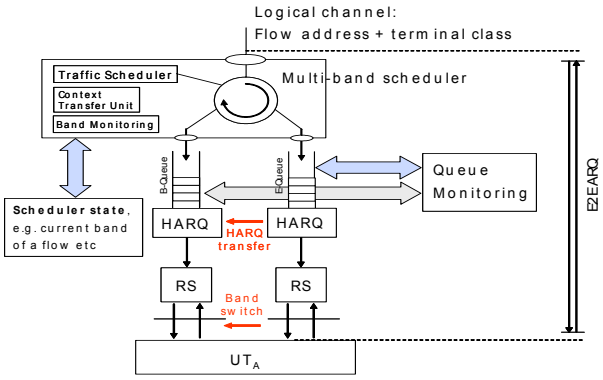


Figure 2. Main functions and operation of the Multband Scheduler. Fast context transfer during a switch to another band during an ongoing HARQ process.

When a UT switches for example from the extension (E) band to the basic (B) band, also the HARQ buffer is transferred, and as a result, the HARQ retransmissions can be continued on the B band. This requires the following:

1) The HARQ processes in the B queue and the E

queue use a common numbering scheme.

2) After the switch, the arriving data is directed to the scheduling queue of the new band and any data remaining in a buffer is transferred to the queue of the other band.

Next to the HARQ buffer, the number of performed retransmissions and information specific to the utilized HARQ scheme is exchanged.

3.3. Preparation for Band Switch

In many situations the timing of the band switch is known beforehand and preparations can be made before the switch. For example, if the BS knows that the E band will not be available any more after 5 ms, then it can command the UT to synchronize with the B band. The BS can assist the synchronization by sending, for example, the time shift to the beginning of the next frame on the B band, the frequency shift between the two bands, and other system information, such as the position of the resource allocation table. Furthermore, the UT can estimate the pathloss or an initial channel quality indicator and report it to the BS well before switching to the B band. When the UT switches bands, this information is forwarded from the E band to the B band scheduler. The anticipation of the band switch allows for a seamless handover, because contexts can be exchanged before the connection on the band is actually lost.

4. Use of MBS for Flexible Spectrum Sharing

The ITU-R studies [10][11] show that a considerable amount of new spectrum will be needed to provide the total capacity that is needed for delivering the predicted services and traffic in the future. However, spectrum for wireless networks is already a scarce resource and will become even scarcer in the future. Therefore sharing the spectrum with other radio systems is a possibility to access additional spectrum. It is based on the assumption that when one network operator or radio system is in demand of spectrum, another network operator might have spectrum available. Thereby the exploitation of available unused spectrum or sharing of spectrum between technologies leads to a better utilization of spectrum throughout a multi-operator or a multi-radio-network environment.

In the preparation phase towards WRC-07 several IMT candidate bands have been identified [12] and parts of these bands have been allocated as IMT bands for communication systems at WRC07. The newly identified spectrum comprises the following spectrum bands: 450-470 MHz was identified globally, 698-802 MHz was identified in the Americas and some Asian countries, 790-862 MHz was identified in Europe, Africa and most Asian countries, 2300 - 2400 MHz has been identified globally, and 3400 - 3600 MHz was identified in most countries in Europe and Africa, and in several countries in Asia.

Fixed Satellite Service (FSS) is the primary service deployed in large portions of the candidate bands. In the allocated C-band (3.4 to 3.6 GHz) dedicated spectrum will be available for IMT systems, but not world wide. To enable further deployment in the C-band sharing with FSS will be required. When bands are available on a sharing basis their availability cannot always be guaranteed. For example transmission exclusion zones around technologies with which the spectum is shared, might be defined. On the one hand the shared new bands should be accessible to guarantee high capacity, but on the other hand, dedicated and guaranteed bands are required to offer guaranteed network access. Therefore, the UTs have to operate in a multi band environment. A possible spectrum allocation for a system deploying one dedicated band between 3.4GHz and 3.6GHz and a shared band at higher frequency in the same band is illustrated in Figure 3.

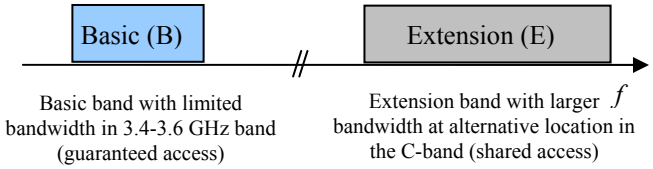


Figure 3. Possible spectrum allocation for IMT-Advanced in a multi band deployment.

Flexible Spectrum Use (FSU) between operators using the same technology is another possibility of dynamic spectrum use, enabling flexible deployments with a limited amount of available spectrum. Also in this case a shared band for enhanced capacity and a dedicated band with guaranteed access can be defined [13].

A fast context transfer is required when a user terminal served on the shared extension band moves into an area where it would interfere with the other system's transmissions, e.g. when it enters the transmission exclusion zone around a satellite earth station. The fast transfer to the basic band can be provided by the MBS, assuming that one BS handles both B and E band.

4.1. Case Study

Figure 4 depicts a simulation scenario for which we study a band transfer of a UT with and without MBS at the BS. The circle in the center marks a transmission exclusion zone, where UTs are not allowed to use the E band and have to switch to the B band.

In our simulations we use an event driven dynamic system simulator that simulates UL and DL directions simultaneously with OFDMA symbol resolution and uses an Exponential Effective SINR Mapping (EESM) link to system mapping [13].

We use the handover procedures described in section

2.1 to model BSs that are/are not equipped with an MBS. Figure 5 illustrates the instantaneous throughput of a UT when entering and leaving the exclusion zone and changing to the B band for the MBS and the non-MBS case. It clearly shows that the MBS and the anticipation of the band switch avoids periods with zero throughput when entering the exclusion zone. Nevertheless, in both cases the UT will experience lower throughput because of the lower capacity of the B band.

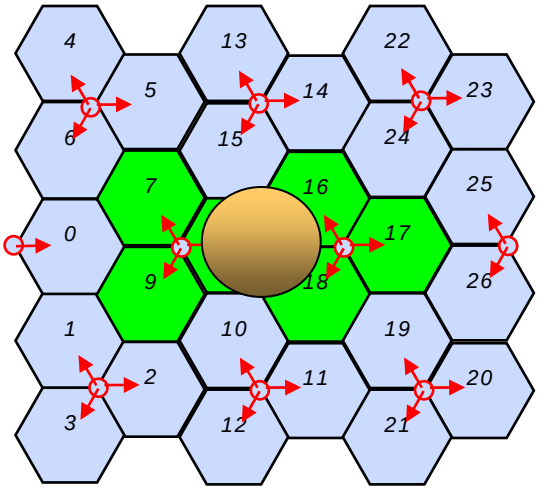


Figure 4. Macro-cellular simulation scenario with 27 sectors. Results are presented for the 6 sectors in the center. The circle in the center marks an exclusion zone. UTs in this zone are not allowed to use the E band.

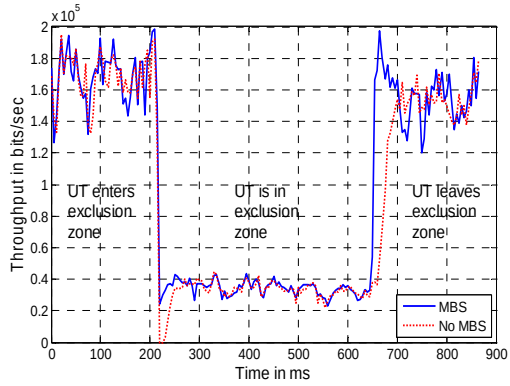


Figure 5. Instantaneous User Throughput averaged over 5ms with and without MBS when entering an exclusion zone.

If the B band is congested the MBS can relocate traffic from the B band to the E band or decide to drop or reduce the bandwidth of less important flows in the B band. When leaving the exclusion zone, the UT can immediately receive data on the E band and benefit from the increased capacity. However, without an MBS the UT has to go through a handover procedure, which typically involves some time to trigger the handover, pending HARQ transmissions are lost and the initial throughput is lower because the serving RAP does not have CQI information from the UT.

5. Application of MBS to Relaying

In a multiband operation, Relay nodes (RN) can be used to extend the coverage of the band with worse propagation characteristics. For the example spectrum allocation in Figure 6 a radio access point (RAP) is able to provide wide area coverage on the basic (B) band at 860MHz. However it cannot cover the same area with the shared extension (E) band at 3.4GHz because of the differences in the propagation loss due to the different carrier frequencies.

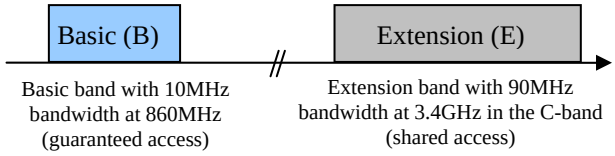


Figure 6. Possible spectrum allocation for IMT-Advanced in a multi band deployment.

In such a scenario RNs can extend the E band coverage to the areas of interest. The RNs do not require a backhaul connection and an E band radio interface is sufficient. Thus the RNs are less complex and cheaper than adding additional BSs. The RNs in the E band do not have to provide ubiquitous coverage but they should cover most of the area to make the high capacity E band available for most of the UT.

The BS uses the MBS for load balancing between the bands.

5.1. Case Study

We study the performance difference with and without RN in the E band for the scenario presented in Figure 4. The BSs are equipped with both B band and E band radio interface. The Inter-Site Distance is 3km, and the BS can provide the basic coverage for the B band at 860MHz using the pathloss model in [15].

Each BS sector has 6 RN to extend the coverage area of the E band at 3.4GHz, two of them are evenly distributed on a circle around the BS with a radius of 500m and the other 4 are on a circle with a radius of 1000m as illustrated in Figure 7. We assume a line-of-sight (LOS) link between RN and BS and for the BS-UT and RN-UT links we assume a non-LOS link. The corresponding channel and pathloss models can be found in [16]. The BS transmit power is 43dBm per sector and the RN transmit power is 37dBm. 2000 UTs move in the area at a speed of 3km/h and the scenario contains no exclusion zone.

Table 1 compares the average cell throughput of the two center cells. Adding the E band to the B band at the BS increases the cell throughput seven fold. However, the high capacity E band is not available for most of the users

in the cell and the increase does not fully scale with the increase in bandwidth. When using RNs to extend the E band coverage the throughput almost doubles. Further, the high capacity E band is available to most of the users in the cell.

Table 1. Cell throughput comparison with/without E band and RN.

Scenario	Average cell throughput [Mbps]
Only B band BS	22
BS with both bands	150
BS with both bands and RN with E band	270

For quality-of-service support the BS has to take additional criteria into account in the presence of relays, when deciding which data packets should be sent on the basic band and which packets on the extension band:

- Services with low delay requirements should be scheduled on the band/link that requires fewer hops (this will mainly be the B band as it has the better propagation conditions and therefore a wider coverage area).
- High speed users with an E band served by RNs should be transferred to the basic band (RN will have smaller coverage area and this policy will reduce the number of missed packets, because the UT has left the coverage area of the RN).

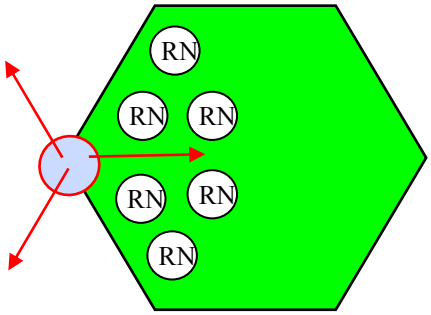


Figure 7. Each sector of Figure 4 is augmented with 6 RN. The BSs transmit on both B band and E band, whereas the RNs only transmit on the E band.

5.2. Relays with MBS

So far we have presented the multi-band operation in RECs for RNs that operate only in the E band. However some of the RNs might be equipped with both a B band and an E band radio interface.

Figure 8 illustrates the case where the RN closest to the BS is equipped with an MBS and therefore packets that should be transmitted on the B band to the UT by the RN can be received on the E band. Typically the E band offers higher capacity than the B band. Thus, it is beneficial to use the E band on the BS-RN link, even if the RN serves the UT on the B band. Here the MBS allows balancing the load of the two bands on the link between the BS and the RN.

To investigate the potential benefits of RNs equipped with an MBS and a B band radio interface for the spectrum allocation in Figure 6 we study the coverage for indoor users in the scenario presented in Figure 9. Each BS is equipped with two sectors and they form together with 3 RN in the same street a REC. The sectors to the right and down have 2 RNs wheras the RN closer to the BS is equipped with an MBS. The other RNs have only an E band radio interface.

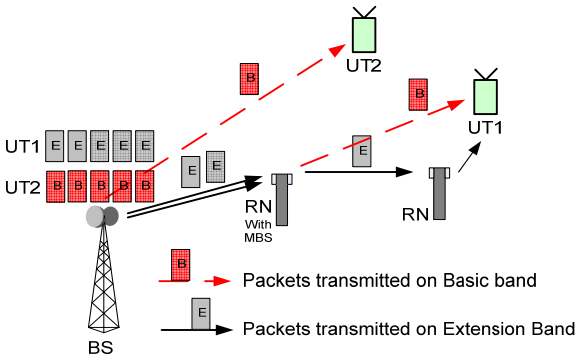


Figure 8. Relay enhanced cell with B band at lower frequency and E band at higher frequency. BS/RN with MBS can decide on which band to transmit packets. RN without MBS receive and forward packets only on extension band.

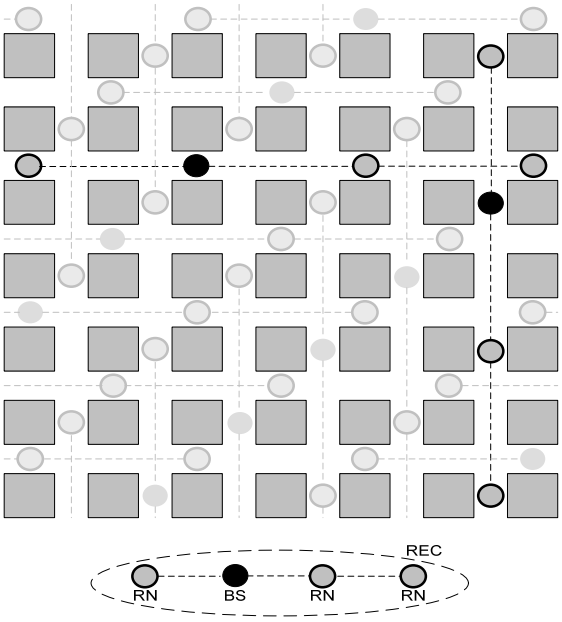


Figure 9. Relay based deployment in the Manhattan grid. The closest RN right of the BS in horizontal streets and down from the BS in vertical streets is equipped with a B and E band radio interface and an MBS.

Table 2 compares the coverage area of the different bands for this scenario. The coverage area has been

calculated using the pathloss models in [16]. In particular we applied the B1 LOS model for points in the same street than the RAP (BS or RN) and the B1 NLOS model for points in different streets. Inside the building blocks we use the B4 outdoor-to-indoor model, whereas we assume an indoor pathloss of 0.5dB/m for the E band (as specified by the model) and 0.3dB/m for the B band to take into account the lower indoor propagation losses at 860MHz than at 3.4GHz. In both cases the BS and RN use a transmit power of 30dBm and the BS is equipped with a 120 degree sector antenna having 11dBi antenna gain, whereas the RN is equipped with an omni-directional antenna and an antenna gain of 7dBi, following the assumptions in [17].

Table 2. Area with a spectral efficiency higher than 1bps/Hz.

E-band	60%
B band without RN	68%
B band with RN	83%

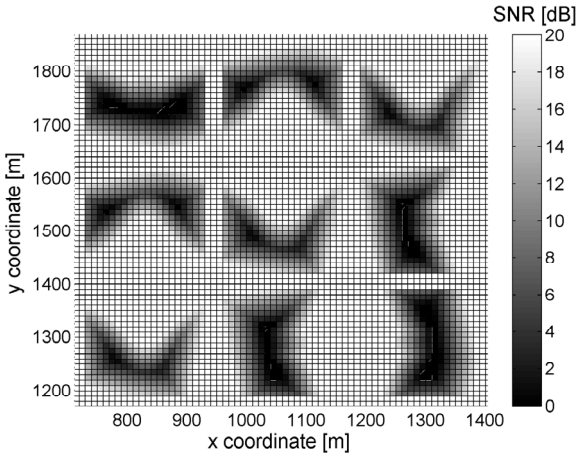


Figure 10. Coverage of B band with BS and part of the RN having an B band interface.

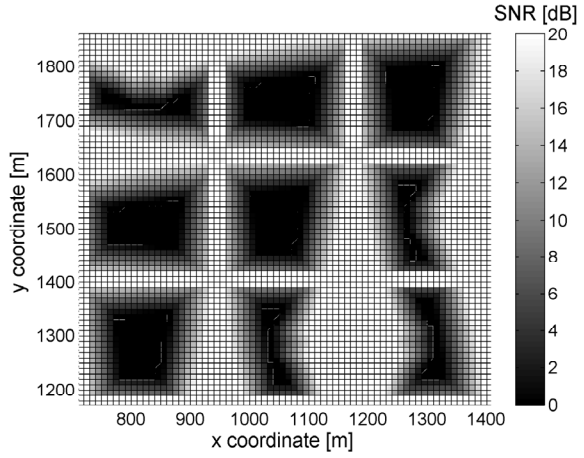


Figure 11. Coverage of the B band with only BS having a B band interface.

The E band can only provide a spectral efficiency of more than 1b/s/Hz in 60% of the area even though every RAP is equipped with an E band interface. In contrary, the BS alone can already provide this spectral efficiency for the same area on the B band. The coverage area can be further increased to 83% by equipping one third of the RNs with a B band interface.

The coverage for the B band in the center area of the scenario in Figure 9 is illustrated in Figure 10 and Figure 11 for the case when only BS have a B band radio interface and for the case where additionally one third of the RNs are equipped with a B band radio interface, respectively. This comparison clearly shows that the B band should be available at the RN as well to provide coverage. However, due to the lower bandwidth of the B band, the B band should only be used to serve UT that cannot be served on the E band but not for forwarding data to the RN. Therefore, the RN should be equipped with an MBS to be able to receive data on the E band and forward it on the B band to UTs that it serves on the B band.

On the other hand, for relay deployments with more than two hops the BS might be able to reach RNs via one hop on the B band and via multiple hops on the E band. In this case, it will be beneficial for delay sensitive traffic to send data on the B band to the RN, which forwards it then to the UT.

5.3. ARQ in relay network with MBS

Another important aspect of a REC is the handling of the outer ARQ (E2E ARQ) between BS and UT, sometimes also referred to as relay ARQ [8]. Without an outer ARQ between BS and UT, the BS does not know whether data sent to the RN is successfully transmitted to the UTs. Thus, in case of handovers, data that is still in the buffer of RNs might be lost, even if the handover destination is within the REC. Additionally, an outer ARQ (E2E ARQ) process is used on each hop to recover from residual inner ARQ (HARQ) errors, caused for example by a NACK that is misinterpreted as an ACK. A detailed description of the ARQ handling in RECs can be found in [8].

For relay deployments with more than two hops, an MBS offers additional degrees of freedom. Even though the BS initiates a handover to a RN in the E band, the BS or another RN with MBS might still be able to serve the UT on the B Band. In this case, outer ARQ retransmissions of delay sensitive traffic can be performed on the B band by these nodes.

5.4. MBS and cooperative relaying

Next to single path relaying, cooperative relaying can be integrated as an add-on to single path relaying as proposed for example in [8]. A receiving node combines signals from more than one transmitting node. Cooperative relaying and intelligent deployment reduce

the total cost of a multi-hop network by reducing the number of relays needed for a given performance [8]. Thus, it is important to show, that the MBS concept fits well with cooperative relaying. In particular we discuss it using the cooperative relaying concept in [8].

Multi-hop diversity is one example, in which a receiving node combines the signals received from previous nodes in the path. In a 2-hop DL path, the UT combines the signal from a RN with the signal from the BS that can be received on different resources. MIMO cooperative relaying is another example, where the BS first transmits the data to be forwarded in a cooperative transmission to the RN. Then the BS and the RN antennas form a virtual antenna array and perform a joint MIMO transmission on the same resources.

In both cases the BS allocates the resources for all the cooperative transmissions, i.e. even in a case where the RN is equipped with an MBS, the BS also decides on which band and resources the cooperative transmissions will take place.

For the multi-hop diversity case, cooperative transmissions could in principle be scheduled on two different bands requiring the UT to be associated with two bands at the same time which increases the UT complexity and power consumption. Thus, cooperative transmissions on different bands should be avoided and the BS has to balance the gain from utilizing cooperative relaying and from using different bands on the first and second hop. However, as described above, by using the HARQ context transfer, retransmissions can be performed on another band and additional diversity gain can be obtained without requiring the UT to operate simultaneously on two bands.

This is not an issue in the MIMO cooperative relaying case, where the BS can send for example data on the E band and then perform a joint MIMO transmission on the B band, whereas the UT only receives on the B band.

6. Conclusion

A new concept called Multi-Band Scheduler (MBS) for future communication systems was introduced. This scheduler allows for operation on multiple bands in a delay constrained environment. The proposed tight integration between multiple bands enables a fast and seamless switch between different bands. The Multi-Band scheduler also ensures that the PHY and MAC layer operation is abstracted from the higher layers. This means that the higher layers are not aware of the actual resources used, but only of the available capacity. The fast switch between multiple bands adds additional degrees of freedom for optimizing the network operation. Moreover, the MBS can be utilized to efficiently balance the load of the bands in the network or to provide required quality of service levels to the UTs.

The operation of the MBS was discussed in detail for two scenarios: Spectrum sharing and relays.

The spectrum sharing case comprises a multiband operation with a smaller band dedicated to the communication system and an extension band that is shared with another system but offers higher capacity. In this scenario the MBS can be applied in two ways, on the one hand the MBS enables simultaneous access to a guaranteed basic band and the band shared with another technology. On the other hand the MBS can be used for sharing spectrum between different operators of the same technology where a part of the band is dedicated to each operator and the rest of the band is shared between the operators. Our simulation results show that using the MBS, a seamless switch can be made if the shared band is no longer available.

The relay case illustrates a scenario with different propagation properties and thus different coverage area of the used bands. In this scenario RNs are used to extend the coverage of the high capacity extension band that has a higher propagation loss than the basic band. Our simulation results show that RNs are an effective way to balance the coverage of the bands. Thereby they greatly increase the overall capacity of the network and the high capacity band is available to most of the user terminals. Further, the MBS at both the BS and the RN enables to balance the network load on each hop by utilizing different bands.

7. Acknowledgement

The authors would like to thank Kennett Aschan for his valuable support and fruitful discussions during the preparation of this article.

8. References

[1] Recommendation ITU-R M.1645, “Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000”, 2003.

[2] K. Hooli, S. Thilakawardana, J. Lara, J-P. Kermoal, S. Pfletschinger, “Flexible Spectrum Use between WINNER Radio Access networks”, IST Mobile Summit, Greece, June 2006.

[3] K. Doppler, He Xiaoben, C. Wijting, A. Sorri, “Adaptive Soft Reuse for Relay Enhanced Cells”, IEEE Vehicular Technology Conference 2007 spring, April 2007, pp 758-762.

[4] https://www.ist-winner.org

[5] M. Sternard, T. Svensson, G. Klang, “The WINNER B3G System MAC Concept”, IEEE Vehicular Technology Conference 2006 fall, Sept 2006.

[6] IST-WINNER II, “D6.11.2 Key Scenarios and

Implications for WINNER II", Sept. 2006, available at https://www.ist-winner.org/deliverables.html

[7] 3GPP TR 25.913 V7.3.0 (2006-03) Requirements for Evolved UTRA (E-UTRA) (Release 7).

[8] IST-WINNER II, "D3.5.2 Assessment of relay based deployment concepts and detailed description of multi-hop capable RAN protocols as input for the concept group work", June 2007, available at https://www.ist-winner.org/deliverables.html

[9] K. Doppler, C. Wijting, J-P. Kermoal, "Multi-Band Scheduler for Future Communication Systems", WiCom 2007, P.R: China, Sept. 2007, pp 6738-6742.

[10] Recommendation ITU-R M.1768, "Methodology for calculation of spectrum requirements for the future development of the terrestrial component of IMT-2000 and systems beyond IMT-2000", 2006.

[11] Report ITU-R M.2078, "Spectrum requirements for the future development of IMT-2000 and IMT-Advanced", 2006.

[12] Report ITU-R M.2079, "Technical and operational information for identifying spectrum for the terrestrial component of future development of IMT-2000 and IMT-Advanced", 2006.

[13] M. Bennis, J. -P. Kermoal, P. Ojanen, J. Lara, S. Abedi, R. Pintenet, S. Thilakawardana and R. Tafazolli, "Advanced spectrum functionalities for 4G WINNER radio network", Wireless Personal Communications journal, Springer 2007 (accepted for publication).

[14] K. Brueninghaus, D. Astely, T. Saelzer, et al. "Link performance models for system level simulations for broadband radio access systems", Proceedings of the PIMRC 2005, Berlin, September 2005

[15] "Universal Mobile Telecommunications System (UMTS); Selection procedures for the choice of radio transmission technologies of the UMTS", TR 101 112 v3.2.0 (1997-11), UMTS30.03

[16] IST WINNER II "D1.1.2 WINNER II Channel Models Part I Channel Models", September 2007 available at https://www.ist-winner.org/deliverables.html

[17] IST WINNER II "D6.13.11 Final CG "metropolitan area" description for integration into overall System Concept and assessment of key technologies", October 2007 available at https://www.ist-winner.org/deliverables.html

I. J. Communications, Network and System Sciences. **2008; 1: 1-103**
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

Adaptive Throughput Optimization in Downlink Wireless OFDM System

Youssef FAKHRI[1], Benayad NSIRI[2], Driss ABOUTAJDINE[1], Josep VIDAL[3]
[1]*University Mohamed V-Agdal Faculty of Sciences, UFR Informatics and Telecommunication B.P. 1014 Rabat, Maroc.*
[2]*Faculty Ain Chok ,UFR Signal Processing ,University Hassan II, B.P. 5366 Casablanca, Maroc.*
[3]*University of Catalonia Dept. of Signal Theory and Communications Campus Nord, Modul D5, C/ Jordi Girona 1-3, Espagne.*
E-mail: yousseffakhri@yahoo.fr, benayad.nsiri@ieee.org, aboutaj@fsr.ac.ma, pepe@gps.tsc.upc.es

Abstract

This paper presents a scheduling scheme for packet transmission in OFDM wireless system with adaptive techniques. The concept of efficient transmission capacity is introduced to make scheduling decisions based on channel conditions. We present a mathematical technique for determining the optimum transmission rate, packet size, Forward Error Correction and constellation size in wireless system that have multi-carriers for OFDM modulation in downlink transmission. The throughput is defined as the number of bits per second correctly received. Trade-offs between the throughput and the operation range are observed, and equations are derived for the optimal choice of the design variables. These parameters are SNR dependent and can be adapted dynamically in response to the mobility of a wireless data terminal. We also look at the joint optimization problem involving all the design parameters together. In the low SNR region it is achieved by adapting the symbol rate so that the received SNR per symbol stays at some preferred value. Finally, we give a characterization of the optimal parameter values as functions of received SNR Simulation results are given to demonstrate efficiency of the scheme.

Keywords: Rate, Packet Lenght, FEC, Throughput, QoS, SISO-OFDM

1. Introduction

Orthogonal frequency division multiplexing (OFDM) is a promising technique for the next generation of wireless communication systems [1] [2]. OFDM divides the available bandwidth into N orthogonal sub-channels. By adding a cyclic prefix (CP) to each OFDM symbol, the channel appears to be circular if the CP length is longer than the channel length. Each sub-channel thus can be modelled as a time-varying gain plus additive white Gaussian noise (AWGN). Following the success of cellular telephone services in the 1990s, the technical community has turned its attention to data transmission. Throughput is a key measure of the quality of a wireless data link. It is defined as the number of information bits received without error per second and we would naturally like this quantity as to be high as possible. This paper looks at the problem of optimizing throughput for a packet based wireless data transmission scheme from a general point of view. The purpose of this work is to show the very nature of throughput and how it can be maximized by observing its response to certain changing parameters. There has been little previous work on the topic of optimizing throughput in general. Some things that have been investigated include choosing an optimal power level to maximize throughput [4][5]. Maximizing throughput in a direct sequence spread spectrum network by way of a link layer protocol termed the Transmission Parameter Selection Algorithm (TPSA) has also been discussed [3]. This provides real time distributed control of transmission parameters such as power level, data rate, and forward error correction rate. An analysis of throughput as a function of the data rate in a CDMA system has also been presented [6]. Most of the previous work found has taken a very specific look at throughput in different wireless voice systems such as TDMA, CDMA, GSM, etc. by taking into account many different system parameters in the analysis such as Parameter Optimization of CDMA Data Systems [7]. We have taken a more general look at throughput by considering its definition for a packet-based scheme and how it can be maximized based on the channel model being used. Unlike most of the work done on this topic, our research is focused on the transmission of data as opposed to that of voice. Most of the work done on data throughput analysis has been in wired networks (i.e. Ethernet, SONET, etc.). Even in this work, however, the analysis is mostly done with system specific parameters. Many

variables affect the throughput of a wireless data system including the packet size, the transmission rate, and the number of overhead bits in each packet, the received signal power, the received noise power spectral density, the modulation technique, and the channel conditions. From these variables, we can calculate other important quantities such as the signal-to-noise ratio γ, the binary error rate $P_e(\gamma)$, and the packet success rate $f(\gamma)$. Throughput depends on all of these quantities. The rest of this paper is organized as follows. In Section 2, our system model is introduced. In Section 3 and 4, we derive an optimal adaptation of individual design parameters in constant and varying receiver power. In Section 5, we conclude by describing future areas of research in multi-user throughput optimization.

2. Problem Formulation

Consider a communication link which consists of a transmitter, a receiver, and a communication channel with bandwidth W. The transmitter constructs packets of K bits and transmits the packets in a continuous stream. To ensure that bits received in error are detected, the transmitter attaches a C bit CRC to each data packet, making the total packet length K + C = L bits. This packet is then transmitted through the air and processed by the receiver. The CRC decoder at the receiver is assumed to be able to detect all the errors in the received packet. (In practice some errors are not detectable, but this probability is small for reasonable value of C and reasonable SNRs.) Upon decoding the packet, the receiver sends an acknowledgment, either positive (ACK) or negative (NACK), back to the transmitter. For ease of analysis we assume this feedback packet goes through a separate control channel, and arrives at the transmitter instantaneously and without error. If the CRC decoder detects any error and issues a NACK, the transmitter uses a selective repeat protocol to resend the packet. It repeats the process until the packet is successfully delivered. A packet is transmitted symbol by symbol through the channel, where each MQAM symbol has b bits in it and is modulated using fixed power MQAM. Thus, each packet corresponds to *L/b = Ls* MQAM symbols. We assume additive white Gaussian noise (AWGN) at the receiver front end, and no interference from other signals. The channel is narrowband (flat fading), so the power spectra of both the received signal and the noise have no frequency dependence, i.e., the channel is characterized by a single path gain variable.

2.1. SISO-OFDM Systems

This is the conventional system that is used everywhere. Assume that for a given channel, whose bandwidth is B, and a given transmitter power of P the signal at the receiver has an average signal-to-noise ratio of SNR0. Then, an estimate for the Shannon limit on channel capacity, C_p, is:

$$C_p \approx B\log_2(1+SNR_0) \quad \textbf{(1)}$$

It is clear from the formula that increasing the SNR, the channel capacity only increases following a logarithmic law (that is 1 more bit for a 3 dB increase of SNR), and the SNR cannot be increased as much as wished. This happens to be a big limitation, due to the strict power regulations, to the achievable throughput in wireless communication systems. In this paper we consider an OFDM system with only one antenna at the transmitter and the receiver, i.e., a Single-Input- Single-Output (SISO) channel. A N-carriers modulation is assumed, where:

$S_k(t)$, $0 \le t < N$. are the information symbols transmitted during the t^{th} time-block, the mean energy of which is normalized:

$E(|Sk(t)|^2) = 1$. and k is the carrier index. The OFDM modulation technique is generated through the use of complex signal processing approaches such as fast Fourier transforms (FFTs) and inverse FFTs in the transmitter and receiver sections of the radio. One of the benefits of OFDM is its strength in fighting the adverse effects of multipath propagation with respect to intersymbol interference in a channel. OFDM is also spectrally efficient because the channels are overlapped and contiguous. The basic principle of OFDM is to split a high-rate datastream into a number of lower rate streams that are transmitted simultaneously over a number of subcarriers. The relative amount of dispersion in time caused by multipath delay spread is decreased because the symbol duration increases for lower rate parallel subcarriers. The other problem to solve is the intersymbol interference, which is eliminated almost completely by introducing a guard time in every OFDM symbol. This means that in the guard time, the OFDM symbol is cyclically extended to avoid intercarrier interference. An OFDM signal is a sum of subcarriers that are individually modulated by using phase shift keying (PSK) or quadrature amplitude modulation (QAM). The symbol can be written as:

- If $t_s \le t < t_s + T$

$$S(t) = \mathrm{Re}\left(\sum_{i=-\frac{N}{2}}^{\frac{Ns}{2}-1} d_{i+\frac{Ns}{2}} \exp\left(j2\pi\left(f_c - \frac{i+0.5}{T}\right)(t-t_s)\right)\right) \quad \textbf{(2)}$$

- *else*

$$S(t)=0 \quad \textbf{(3)}$$

N_s is the number of subcarriers T is the symbol duration fc is the carrier frequency.

The use of channel estimation is a very interesting

function to be added to the receiver to make the system more resistant to fading and Doppler effects, over all, if it is going to be used aboard of cars in a highway. Wireless LAN is a very important application for OFDM and the development of the standard promises to have not only a big market but also application in many different environments.

2.2. Throughput Analysis

The total throughput of the system, *T*, is the sum of the individual throughputs of the sub-carriers *i* operating simultaneously in the system.

Since we are considering a OFDM system with *N* sub-carriers.

With the above simplifying assumptions, we define the throughput of a system as the number of payload bits per second received correctly [8],[9]:

$$T = \sum_{i=1}^{N} \frac{L-C}{L} * R_i * f(\gamma_i) \quad \textbf{(4)}$$

where R_i is the symbol rate assigned to sub-carriers i, $f(\gamma_i)$ is the packet success rate (PSR) defined as the probability of receiving a packet correctly, and γ_i is the SNR given by:

$$\gamma_i = \frac{P_i}{N_0 * R_i} \quad \textbf{(5)}$$

where N_0 the one-sided noise power spectral density, and P_i the received power in sub-carriers i.

3. Throughput Optimization

3.1. Optimal Symbol Rate

To find the symbol rate R_{is} that maximizes throughput, we differentiate (4) with respect to R_{is} and set it to zero to obtain the following condition.

$$\frac{dT}{dR_i} = \frac{L-C}{L} f(\gamma) + \frac{L-C}{L} R_i \frac{df(\gamma)}{d\gamma} \frac{d\gamma}{dR_i} \quad \textbf{(6)}$$

$$\frac{dT}{dR_i} = \frac{L-C}{L} (f(\gamma) + R_i \frac{df(\gamma)}{d\gamma} (\frac{-P_i}{N_0 R_i^2})) \quad \textbf{(7)}$$

Next we set the derivative to zero

$$f(\gamma) - (\frac{P_i}{N_0 R_i}) \frac{df(\gamma)}{d\gamma} = 0 \quad \textbf{(8)}$$

$$f(\gamma) = \gamma \frac{df(\gamma)}{d\gamma} \quad \textbf{(9)}$$

We adopt the notation $\gamma = \gamma^*$ for a signal to noise ration that satisfies equation (9). Since any symbol error in the packet results in a loss of the packet, the PSR f is given in terms of the symbol error rate P_e by:

$$f(\gamma) = [1 - P_e(\gamma^*)]^{\frac{L}{b}} \quad \textbf{(10)}$$

Combining these two, we arrive at an equation for obtaining the preferred SNR per symbol γ^* :

$$\gamma^* \frac{dP_e(\gamma)}{d\gamma}\Big|_{\gamma=\gamma^*} = -\frac{[1 - P_e(\gamma^*)]}{L/b} \quad \textbf{(11)}$$

where Pe of MQAM in AWGN channels is (approximately) given by [8]:

$$P_e(\gamma) = 4(1 - 2^{-\frac{b}{2}}) Q(\sqrt{\frac{3}{2^b - 1}\gamma}) \quad \textbf{(12)}$$

Once γ^* is determined, the optimal symbol rate is obtained from (5).

$$R_i^* = \frac{P_i}{\gamma^* N_0} \quad \textbf{(13)}$$

Note that the solution γ^* in equations (10) and (11) depends on only design parameters b and L, but is independent of the received power level. In essence, the adaptive system monitors γ , and upon deviation from its internally preset value γ , changes its symbol rate such that $\gamma = \gamma^*$. Figure 1 shows the spectral efficiency *T/W* versus received *SNR* for different symbol rates. Where W the bandwidth required for transmission of $b = \log_2(2^b)$ information bits. We see that the system can support high symbol rates at high SNR, but its throughput rapidly decreases at a certain SNR value below which the system should switch to a lower symbol rate to maintain the optimal throughput. The optimal curve is obtained by adapting R_i to keep $\gamma = \gamma^*$.

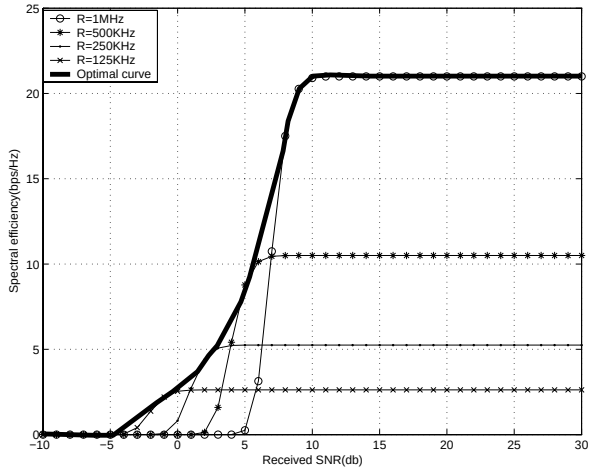


Figure 1. Throughput vs SNR with variable R,L=100,b=2 and N=50

3.2. Optimal Packet Length

To find an analytic solution for the optimal packet length L, we assume L to take continuous values. Differentiating (4) with respect to L and using (10) produce:

$$\frac{dT}{dL} = \frac{C}{L^2} R_i f(\gamma) + (1 - \frac{C}{L}) R_i f(\gamma) \ln(1 - P_e(\gamma)) \quad \textbf{(15)}$$

Setting this to zero produces a quadratic equation in L with the positive root:

$$L^* = \frac{C}{2} + \frac{1}{2}\sqrt{C^2 - \frac{4bC}{\ln(1 - P_e(\gamma))}} \quad \textbf{(16)}$$

Thus, the optimal packet length L depends on the constellation size 2^b, the SNR per symbol γ, and the probability of symbol error *Pe*. Fortunately, we observe in (4) that at high γ, $f(\gamma) \approx 1$ and the throughput is proportional to 1-C/L. Therefore, the throughput gain becomes negligible if we increase L beyond a certain point. In Rayleigh fading channels, L^* is much smaller and in fact asymptotically proportional to $\sqrt{\gamma}$.

Figure 2 shows the spectral efficiency of systems with various packet lengths under a fixed symbol rate and constellation size.

We see that large packet size gives high throughput at high SNR, whereas small packet size gives a better performance at low SNR. Thus, by adaptively changing the packet length, we can achieve both higher throughput and a wider operation range than using a fixed packet length. However, this doesn't necessarily mean that the packet length should always be variable. For example, if we adapt the symbol rate and the packet length simultaneously, there is a single $L^*(\gamma^*)$ that is optimal regardless of the received SNR value, because the symbol rate is adapted first to maintain $\gamma = \gamma^*$, eliminating the effect of any SNR change. In this case, one degree of adaptation (i.e., symbol rate) would be enough for the 2-D optimization problem [11].

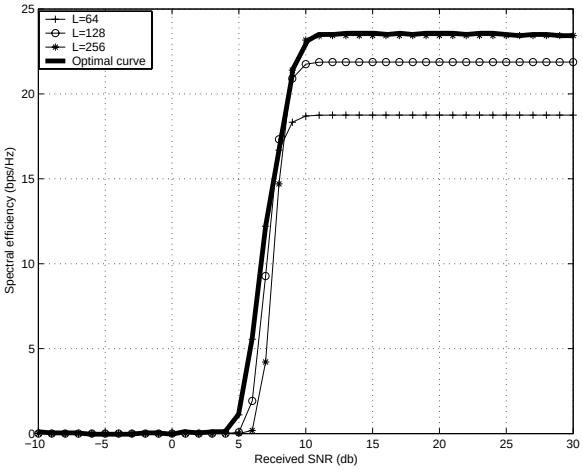


Figure 2. Throughput vs SNR with variable L,R=1MHz,b=2 and N=50

3.3. Optimal Constellation Size

Constellation size 2^b is another degree of freedom that can be adapted to variations in received SNR, to allow packing more bits per symbol when the channel gain is high By differentiating (4) with respect to b and setting it to zero, we obtain an equation for the optimal number of bits per MQAM symbol b^* as:

$$\frac{dP_e(\gamma, b)}{db}\Big|_{b=b^*} = -\frac{[1 - P_e(\gamma, b^*)]}{L} \quad \textbf{(17)}$$

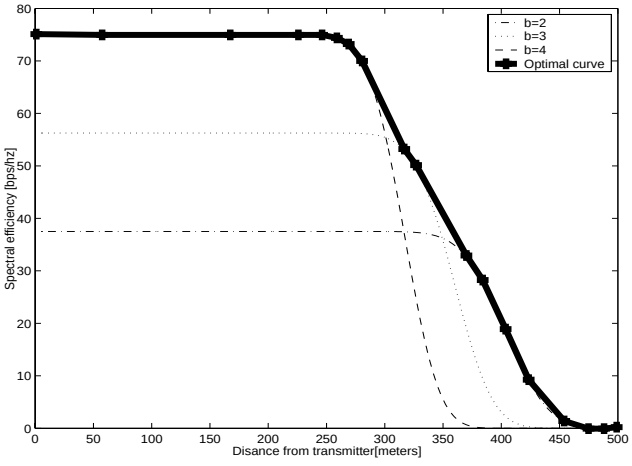


Figure 3. Throughput vs Distance from transmitter with variable b, L=64, R=1Msps

Figure 3 shows the spectral efficiency of systems with various constellation sizes under fixed symbol rate and packet length in AWGN channels. Higher level modulations using MQAM mainly increase the throughput at high low distance, but have shorter communication ranges. By adapting the constellation size, therefore, we can boost the throughput at low distance while maintaining the same performance at high distance.

4. Throughput Optimization using Forward Error Correction

4.1. Forward Error Correction Throughput Equations

We will denote the number of information bits in the packet as K, and the number of CRC bits as C. L is defined as K+C. Now instead of transmitting those L bits with no error correction capability, we will now add B error correcting bits and transmit a total of L+B bits. Using a block code forward error correction scheme, the minimum number of bits B required to correct t errors is given by [12]:

$$t = \max\left\{k \left| \log_2\left(\sum_{n=0}^{k}\binom{L+B}{n}\right)\right.\right\} \le B \qquad \textbf{(18)}$$

Now that we can correct t errors, our packet success rate, $f(\gamma)$ should be larger than its previous value with no error correction. Recall that $f(\gamma)$ with t=0 is given by:

$$f(\gamma) = [1 - P_e(\gamma_i)]^{\frac{L}{b}} \qquad \textbf{(19)}$$

where $P_e(\gamma)$ is the probability of a bit error as a function of the SNR. Now, with error correction capability, the packet success rate for some arbitrary value of t is [13]:

$$f(\gamma) = \sum_{n=0}^{t}\binom{L+B}{n} P_e^n(\gamma_i)[1 - P_e(\gamma_i)]^{L+B-n} \qquad \textbf{(20)}$$

Our new equation for the throughput as a function of the signal to noise ratio is:

$$T = \sum_{i=1}^{N}\frac{L-C}{L+B} * \frac{P/N_0}{\gamma_i} * f_t(\gamma_i) \qquad \textbf{(21)}$$

4.2. Optimum Error Correction

If we plot equation (21) for different values of t using MQAM in AWGN channels we get a very interesting result. Figure (4) shows the throughput for t=0, 20, 60,100. In each of these plots, P/N_0 is fixed at $5*10^6$ and C at 16 bits, N=50, b=1. Perhaps the most important thing to note in this graph is that there is an optimum value for t at which higher values of t will not produce higher throughput. From this graph the throughput appears to reach an optimum value somewhere around t=20.

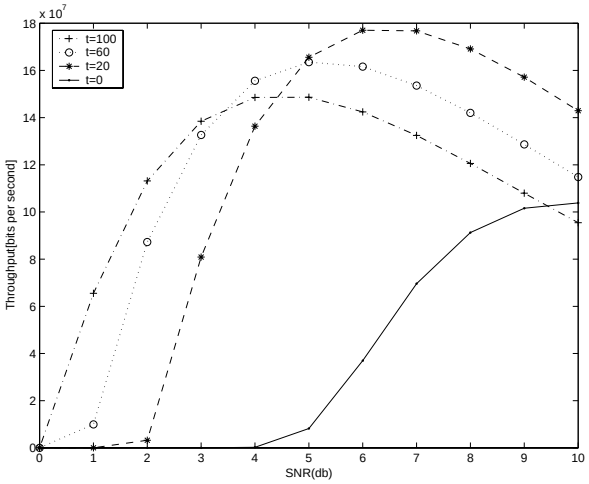


Figure 4. Throughput vs SNR (L=216,N=50)

5. Conclusion

Maximizing throughput in a wireless channel is a very important aspect in the quality of a voice or data transmission.

In this paper, we have shown that factors such as the optimum packet length and optimum transmission rate are all functions of the signal to noise ratio. These equations can be used to find the optimum signal to noise ratio that the system should be operated at to achieve the maximum throughput. The key concept behind this research is that for each particular channel (AWGN or Rayleigh) and transmission scheme $(P_e(\gamma))$, there exists a specific value for the signal to noise ratio to maximize the throughput. Once the probability of error, $P_e(\gamma)$ is known, this optimal SNR value can be obtained. The optimal values depend on the received signal strength. At low SNR, the throughput is maximized by adapting the symbol rate while using the smallest constellation size and some fixed packet length. Finally, we have characterized the optimal adaptation of the parameters in AWGN under restrictions on the values that the parameters can take. Our optimization framework is very general and can be applied to any systems where the combination of FEC, adaptive modulation, and packet length can be adapted to maximize throughput. The analysis and intuition in this paper, however, apply to single user systems. Typical multi-user systems are interference-limited and should be dealt with differently. For example, schemes that enable orthogonal channel sharing among users through frequency (variable symbol rate), time (time division multiplexing), or code division

(spread spectrum modulation) may have an advantage over the other schemes (variable packet length and/or adaptive FEC).

6. References

[1] H. Sampath, S. Talwar, J. Tellado, V. Erceg, and A. Paulraj, A Fourthgeneration MIMO-OFDM Broadband Wireless System: Design, Performance, and Field Trial Results, IEEE Communications Magazine, vol. 40, no. 9, pp. 143-149, 2002.

[2] T.S. Rappaport, A. Annamalai, R. M. Buehrer, and W. H. Tranter,
Wireless Communications: Past Events and a Future Perspective, IEEE
Communications Magazine, pp. 148-161, May 2002.

[3] J. A. C. Bingham, Multicarrier Modulation for Data Transmisson: an Idea whose Time Has Come, IEEE Communications Magazine, vol. 28, no. 5, pp. 5-14, May 1990.

[4] W. T. Webb and R. Steele, Variable rate QAM for mobile radio, IEEE Trans. on Commun., vol. 43, pp. 2223-2230, July 1995.

[5] Y.Fakhri,D.Aboutajdine,J.Vidal. Multicarrier power allocation for maximum throughput in delayed channel knowledge conditions. Proc. Of International Workshop on Wireless Communication in Underground and Confined Areas IWWCUCA2005, Val-dor, Quebec, Canada, Juin 6-7, 2005.

[6] J. Goldsmith and S.-G. Chua, Adaptive coded modulation for fading channels, IEEE Trans. Commun., vol. 46, pp. 595-602, May 1998.

[7] H. Matsuoka, S. Sampei, N. Morinaga, and Y. Kamio, Adaptive modulation system with variable coding rate concatenated code for high quality multi-media communication systems, IEICE Trans. Commun., vol. E79-B, pp. 328-334, Mar. 1996.

[8] M. B. Pursley and J. M. Shea, Adaptive non uniform phase-shift-key
modulation for multimedia traffic in wireless networks, IEEE J. Select.
Areas Commun, vol. 18, pp. 1394-1407, Aug. 2000.

[9] S. Catreux, P. F. Driessen, and L. J. Greenstein, Data throughputs using multiple-input multiple-output (MIMO) techniques in a noiselimited cellular environment, IEEE Trans. Wireless Commun., vol. 1, pp. 226-235, Apr. 2002.

[10] J. G. Proakis, Digital Communications, 4th Ed, New York: McGraw-Hill, 2000.

[11] S. T. Chung, A. Goldsmith, Degrees of freedom in adaptive modulation:a unified view, IEEE Trans. Commun., vol. 49, pp. 1561-1571, Sep. 2001.

[12] Clark,George C., Jr., Cain, J. Bibb, Error-Correction Coding for Digital Communications. Plenum Press, NY, 1981. p38.

[13] Clark,George C., Jr., Cain, J. Bibb, Error-Correction Coding for Digital Communications. Plenum Press, NY, 1981. p219.

I. J. Communications, Network and System Sciences. 2008; 1: 1-103
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

A Discrete Newton's Method for Gain Based Predistorter

Xiaochen LIN[1], Minglu JIN[2], Aifei LIU
School of Electronic and Information Engineering
Dalian University of Technology, Dalian, P.R.China
E-mail:[1] linxc0103@163.com,[2] mljin@dlut.edu.cn

Abstract

Gain based predistorter (PD) is a highly effective and simple digital baseband predistorter which compensates for the nonlinear distortion of PAs. Lookup table (LUT) is the core of the gain based PD. This paper presents a discrete Newton's method based adaptive technique to modify LUT. We simplify and convert the hardship of adaptive updating LUT to the roots finding problem for a system of two element real equations on mathematics. And we deduce discrete Newton's method based adaptive iterative formula used for updating LUT. The iterative formula of the proposed method is in real number field, but secant method previously published is in complex number field. So the proposed method reduces the number of real multiplications and is implemented with ease by hardware. Furthermore, computer simulation results verify gain based PD using discrete Newton's method could rectify nonlinear distortion and improve system performance. Also, the simulation results reveal the proposed method reaches to the stable statement in fewer iteration times and less runtime than secant method.

Keywords: Predistortion, Discrete Newton's Method, Power Amplifiers (PAs), Lookup Table (LUT)

1. Introduction

Radio frequency (RF) power amplifiers (PAs) play an important role in wireless communication systems, but are inherently nonlinear. To compensate for nonlinear of PAs, linearization is an indispensable technique today. Furthermore, among various linearization techniques, digital baseband predistortion is more attractive than others by virtue of its simplicity and ease of implementation with digital signal processor (DSP) equivalently in baseband. In this paper, we take into account the gain based predistortion [1] which employs a lookup table (LUT) block using random access memory (RAM).

PAs characteristics drift by reason of aging, temperature changing, channel switches, and source voltage variations, so a predistorter (PD) should have the ability of adaptation. And this paper focuses on adaptive techniques. The conventional adaptive algorithms including recursive least squares (RLS) and least mean squares (LMS) originate from adaptive filter theory. Based on above, [2] presents a broadcasting adaptive algorithm more efficient for updating PD. Moreover, [3] proposes a modified broadcasting adaptive algorithm, which does not require special form of training sequence. Meanwhile, in [1], J. K. Cavers generates an idea that the adaptation issue can be converted to the root finding problem on mathematics and presents secant method. In recent years, following J. K. Cavers, [4][5], and [6] present combining dichotomy with linear method, rapid secant method, and combing dichotomy with Newton's method, respectively. Methods above meet the requirement of convergence rate and are able to reach to the stable statement, but the drawback of these methods is of high computational load. Therefore, in this paper, we propose discrete Newton's method to adapt a PD which has the advantages of fast convergence and low computational load.

This paper is organized as follows. Section 2 analyses and deduces new method and formula. Section 3 demonstrates the virtue of new method by computer simulation. Finally, we conclude the paper in Section 4.

2. Adaptive Gain Based Predistortion

2.1. Gain Based Predistortion

Figure 1 shows the architecture of a communication system with gain based PD. PD, whose characteristic is opposite of PA's, is used in baseband circuit. The symbol stream v is transferred to a PD via a P/S converter to get predistorted signal w. Then w is converted to analog waveforms via a digital-to-analog (D/A) converter. Finally, the analog signals are quadrature modulated,

upconverted, amplified, and transmitted to the air in sequence.

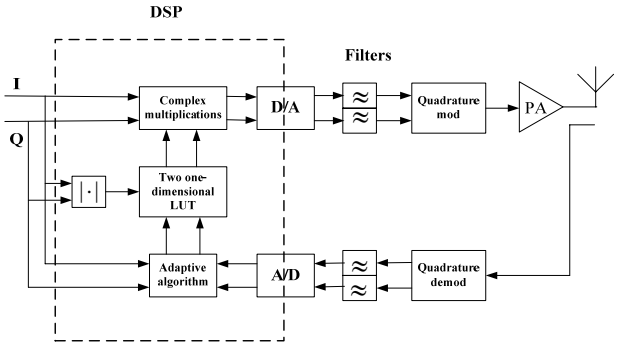


Figure 1. Block diagram for communication system with gain based PD

In Figure 2, we find the block diagram of a gain based PD [8]. The complex gains of PD are saved in a LUT whose entries are indexed. Supposed indexing function $Q(\cdot)$ maps each input signal amplitude to LUT entry index. Also, efficient $Q(\cdot)$ can improve PD performance. LUT entries are either uniform of amplitude or power, or nonuniform diversely due to $Q(\cdot)$. With mapping input signal to a certain LUT entry, adaptive algorithm begins to update the gain value memorized in LUT to get the optimum gain.

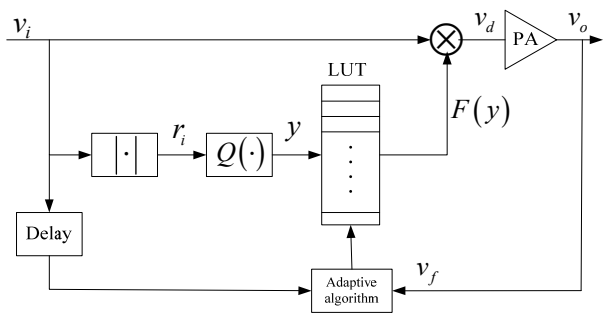


Figure 2. Block diagram of gain based PD

2.2. Discrete Newton's Method

As illustrated from Figure 2 v_d and v_o represent input complex signals and output complex signals of PA, respectively. Then PA is characterized by

$$v_o = v_d G\left(|v_d|^2\right) \quad \textbf{(1)}$$

where $G(\cdot)$ denotes the complex characteristic function of PA, and $|\cdot|^2$ is defined as squared amplitude of signals.

Describing input complex signals of PD as v_i , we can express PD characteristic function as

$$v_d = v_i F\left(|v_i|^2\right) \quad \textbf{(2)}$$

where $F(\cdot)$ denotes the complex characteristic function of PD.

By substituting (2) into (1), we can obtain the whole linear characteristic function of PD and PA in series

$$v_i F\left(|v_i|^2\right) G\left(\left|v_i F\left(|v_i|^2\right)\right|^2\right) = k v_i \quad \textbf{(3)}$$

where the whole linear gain k is a positive constant, and usually less than PA's midrange gain. Simplifying (3), we have

$$F\left(|v_i|^2\right) G\left(|v_i|^2 \left|F\left(|v_i|^2\right)\right|^2\right) - k = 0 \quad \textbf{(4)}$$

Hence, how to get the optimum gain F of LUT is transferred to the root finding problem for a nonlinear complex equation.

As follows, we decompose the complex equation (4) into two real equations separately representing amplitude and phase characteristic. $|(\cdot)|$ and $\angle(\cdot)$ label amplitude and phase of a complex number, respectively.

Then, amplitude and phase characteristic functions of PA are given as follows

$$|v_a| = |v_d| G_a\left(|v_d|^2\right) \quad \textbf{(5)}$$

and

$$\angle v_a = G_p\left(|v_d|^2\right) + \angle v_d \quad \textbf{(6)}$$

where $G_a(\cdot)$ and $G_p(\cdot)$ denote amplitude-modulated amplitude-distortion(AM/AM) and amplitude-modulated phase-distortion (AM/PM) of PA, respectively.

Similarly $F_a(\cdot)$ and $F_p(\cdot)$ represent AM/AM and AM/PM of PD, respectively. We obtain amplitude and phase characteristic functions of PD

$$|v_d| = |v_i| F_a\left(|v_i|^2\right) \quad \textbf{(7)}$$

and

$$\angle v_d = F_p\left(|v_i|^2\right) + \angle v_i \quad \textbf{(8)}$$

Inserting (7) into (5), we get

$$|v_a| = |v_i| F_a\left(|v_i|^2\right) G_a\left(|v_i|^2 \left|F_a\left(|v_i|\right)\right|^2\right) \quad \textbf{(9)}$$

Similarly, combing (8) to (6), hence,

$$\angle v_a = G_p\left(|v_i|^2 \left|F_a\left(|v_i|^2\right)\right|^2\right) + F_p\left(|v_i|^2\right) + \angle v_i \quad \textbf{(10)}$$

The whole gain of PD and PA in series is k, in other words, the whole amplitude gain is k, and the whole phase doesn't change at all. Therefore, we can obtain

$$|v_a| = |v_i| F_a\left(|v_i|^2\right) G_a\left(|v_i|^2 |F_a(|v_i|)|^2\right) = k|v_i| \quad (11)$$

$$\angle v_a = G_p\left(|v_i|^2 \left|F_a\left(|v_i|^2\right)\right|^2\right) + F_p\left(|v_i|^2\right) + \angle v_i = \angle v_i \quad (12)$$

Simplifying and composing (11) and (12), we write equations as follows

$$\begin{cases} F_a\left(|v_i|^2\right) G_a\left(|v_i|^2 \left|F_a\left(|v_i|^2\right)\right|^2\right) - k = 0 \\ G_p\left(|v_i|^2 \left|F_a\left(|v_i|^2\right)\right|^2\right) + F_p\left(|v_i|^2\right) = 0 \end{cases} \quad (13)$$

In the above system of equations, $F_a(\cdot)$ and $F_p(\cdot)$ are separately amplitude and phase gain memorized in LUT, as a result, the issue of updating LUT is converted to the roots finding problem for system of two element nonlinear equations. (13) is simply marked by H

$$H\left(F_a, F_p\right) = \begin{cases} h_1\left(F_a, F_p\right) = 0 \\ h_2\left(F_a, F_p\right) = 0 \end{cases} \quad (14)$$

We can apply Newton's method for system of equations (9) to (14)

$$\begin{pmatrix} F_a \\ F_p \end{pmatrix}^{k+1} = \begin{pmatrix} F_a \\ F_b \end{pmatrix}^{k} - J^{-1} H \quad (15)$$

where J is Jocabian matrix, defined as

$$J = \begin{pmatrix} \dfrac{\partial h_1}{\partial F_a} & \dfrac{\partial h_1}{\partial F_p} \\ \dfrac{\partial h_2}{\partial F_a} & \dfrac{\partial h_2}{\partial F_p} \end{pmatrix} \quad (16)$$

However, in fact, since the derivatives of $h_1(\cdot)$ and $h_2(\cdot)$ are not available, we can not apply Newton method. Instead, we make use of discrete Newton's method which substitutes differential entropy for derivative.

$$J' = \begin{pmatrix} \dfrac{h_1\left(F_a^k, F_p^k\right) - h_1\left(F_a^{k-1}, F_p^k\right)}{F_a^k - F_a^{k-1}} & \dfrac{h_1\left(F_a^k, F_p^k\right) - h_1\left(F_a^k, F_p^{k-1}\right)}{F_a^k - F_a^{k-1}} \\ \dfrac{h_1\left(F_a^k, F_p^k\right) - h_1\left(F_a^{k-1}, F_p^k\right)}{F_p^k - F_p^{k-1}} & \dfrac{h_1\left(F_a^k, F_p^k\right) - h_1\left(F_a^k, F_p^{k-1}\right)}{F_p^k - F_p^{k-1}} \end{pmatrix} \quad (17)$$

Only if J is a nonsingular matrix, iterative formula of discrete Newton's method is

$$\begin{pmatrix} F_a \\ F_p \end{pmatrix}^{k+1} = \begin{pmatrix} F_a \\ F_b \end{pmatrix}^{k} - \left(J'\right)^{-1} H \quad (18)$$

Moreover, we know of a simple way to get the inverse matrix of 2-D square matrix, it is that to say, supposed a 2-D square matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, hence, we obtain

$$A^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} \quad (19)$$

Compared discrete Newton's method for system of equations with secant method for an equation, they have the same order of convergence, super-linear convergence. However, the computational load of discrete Newton's method is lower than that of secant method, due to secant method finding a root of a complex equation, concretely, each iteration requiring 4 complex multiplications, 2 complex additions, and 2 complex by real divisions, in total, 20 real multiplications and 14 real additions. On the contrary, system of equations using discrete Newton's method are always in real number field, and each iteration only needs 6 real multiplications, 12 real additions and 4 real divisions. Obviously, real multiplications of the latter are fewer than those of former, so the computational load of discrete Newton's method is much lower than that of secant under large numbers of symbols.

From Figure 3, we can see the block diagram of a gain based PD with discrete Newton's method. The drawback of the method is one more R/P.

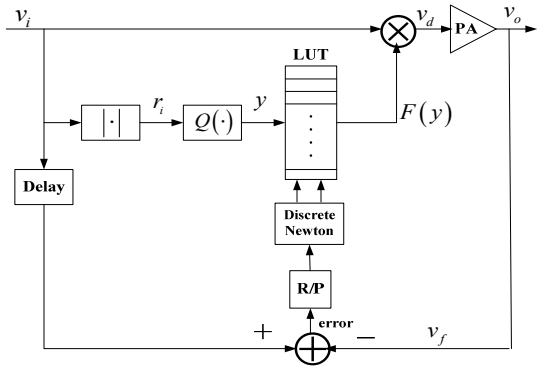


Figure 3. Block diagram of gain based PD with discrete Newton's method

3. Simulation Results

The baseband simulation model block diagram of adaptive PD is shown in Figure 4. In our simulation, we used 16QAM signal and square root raised cosine filter (SRRCF). The channel is assumed to be an additive white Gaussian noise (AWGN) channel. Input backoff (IBO) is 4dB. In this paper, the simulation is restricted to the memoryless distortion. Because PA widely used for

satellite transmission is Traveling Wave Tube Amplifier (TWTA), we apply TWTA to our model as well and simulate the Saleh's model [10]. In our simulation, input signal amplitude is normalized to vary from 0 to 1. We take for uniform amplitude distribution function as $Q(\cdot)$.

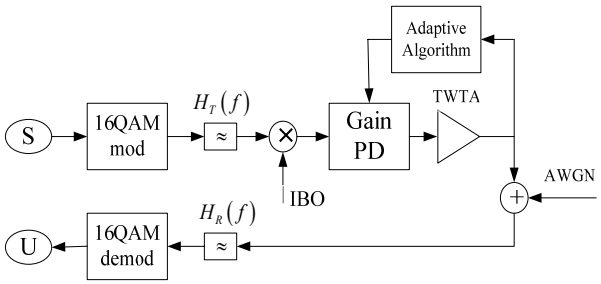


Figure 4. Block diagram of baseband simulation

It is apparent from Figure 5 that the iterative number of discrete Newton's method for one input signal is less by about 10 times than that of secant method. In addition, when signal source generates 2^{14} bits, runtime of secant is 74.8600s, while that of discrete Newton is 49.6410s; when there are 2^{16} bits, runtime of secant is 1536.40s, while that of discrete Newton is 858.7190s, about a half of secant's. And all the above time is on average. Thus we can conclude the proposed method could reduce runtime of adaptation. And with input data increasing, the superiority of discrete Newton's method is more outstanding.

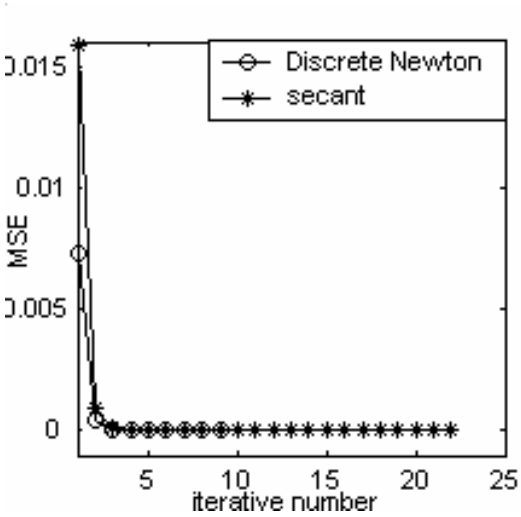


Figure 5. Compare for MSE curve

Figure 6 shows constellation for 16QAM modulation signals. Figure 7 shows constellation for signals distorted by PA. Further more, Figure 8 and Figure 9 illustrates constellation of receiver signals predistorted by the gain based PD with secant and discrete Newton's method, respectively. These simulation results prove the gain based PD with both the two adaptive method compensate nonlinear distortion, and their distorting effects are equally.

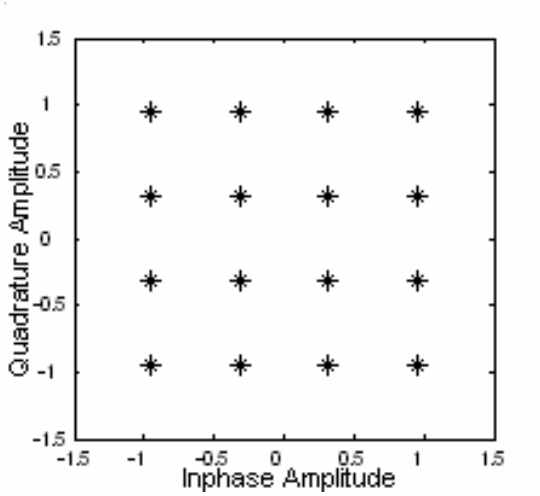


Figure 6. 16QAM modulation signals constellation

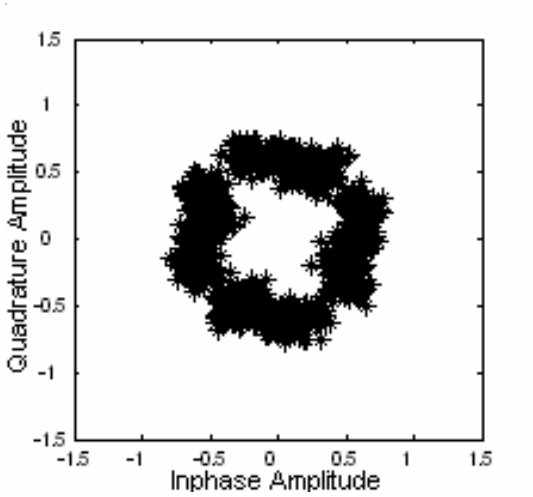


Figure 7. 16QAM signals constellation with PA

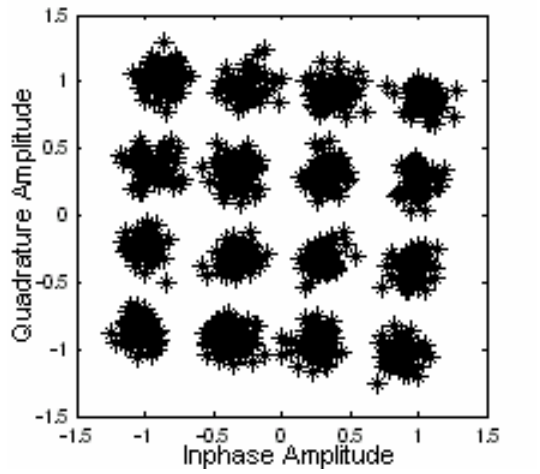


Figure 8. 16QAM receiver signals constellation with secant

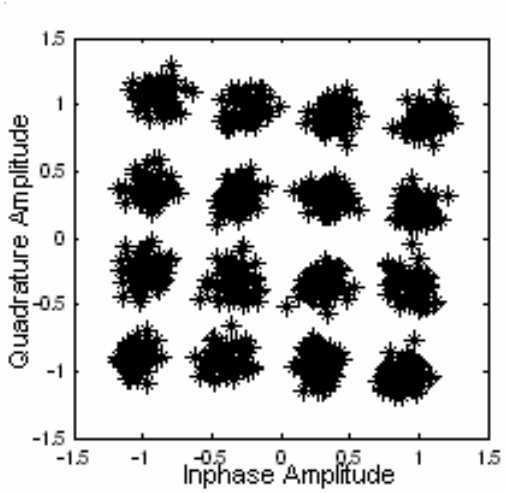


Figure 9. 16QAM receiver signals constellation with discrete Newton's method

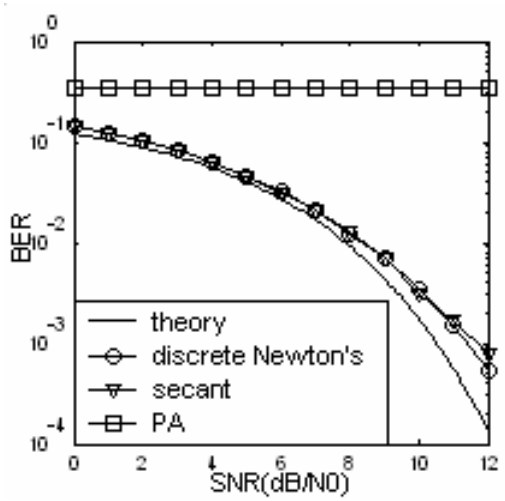


Figure 10. Compare for BER curve

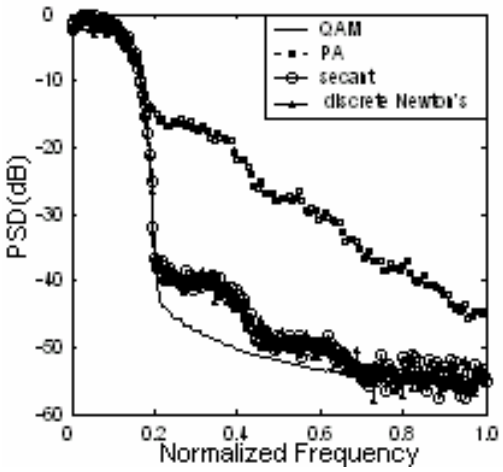


Figure 11. Comparison for PSD

According to Figure 10, at low signal noise rate (SNR), bit error rate (BER) of discrete Newton's method is similar to that of secant method, and they are much less than PA's and close to the theory value. However, with SNR increasing, BER of discrete Newton's method is a bit lower than that of secant method.

Figure 11 shows that the spctrums of 16QAM signals through PA have spectrum re-growth distortion and the gain based PD with secant and discrete Newton's method depress the spectrum re-growth.

TD performance in Figure 12 versus total degradation performance of distorted and compensated signals. There is about 1dB TD decrease with discrete newton's method compared with secant.

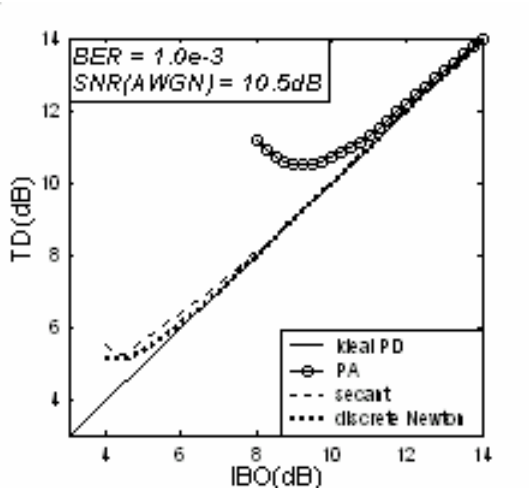


Figure 12. Comparison for TD curves

4. Conclusion

In this paper, we propose an improved adaptation technique using discrete Newton's method efficiently used in PD. We simplify and transfer adaptation to the roots finding problem for system of equations from nonlinear numerical analysis theory. The improved method can produce better convergence performance. In addition, due to computation processing completely being in real number field, the method requires fewer multiplications, lowers computational load. However, the shortcoming of proposed method is one more a R/P.

5. References

[1] J. K. Cavers, "Amplifier Linearization Using a Digital Predistorter with Fast Adaptation and Low Memory Requirements", IEEE Trans. Veh. Technol., Vol. 19, No. 4, Nov. 1990, pp. 374-382.

[2] Won Gi Jeon, Kyung Hi Chang, and Yong Soo Cho, "An Adaptive Data Predistorter of Compensation of Nonlinear Distortion in OFDM System", IEEE Trans. Commun., Vol. 45, No. 10, Oct. 1997, pp. 1167-1171.

[3] Minglu Jin, Sooyoung Kim, Doseob Ahn, Deock-Gil

Oh, and Jae Moung Kim, "A Fast LUT Predistorter for Power Amplifier in OFDM Systems", The 14th IEEE Intemational Sysmposium on personal Indoor and Mobile Radio Communication Proceedings , Beijing, China, Sep. 2003, pp. 1894-1897.

[4] Hongxin Zhao, Yiyuan Chen, and Wei Hong, "A Baseband Predistortin Linearizer for RF power amplifier Prototype Simulation and Experimentation", Journal of China Insttiute of Communications, Vol. 21, No. 5, May. 2000, pp. 41-47.

[5] Shuyue Wu, Xinguang Tian, and Ping Bao, "A Fast Algorithm of the Linearizing Technology of the Adaptive Power Amplifiers", Signal Processnig, Vol. 18, No.3, Jun.2002, pp. 254-256.

[6] Shuyue Wu, Wenming Guo, Xin Song, and Xinguang Tian, "A New Algorithm of the Linearizing Technology of the Digital Baseband Predistortion", J. Xiangtan Min. Inst., Vol. 19, No. 1, Mar. 2004, pp. 63-65.

[7] Kathleen J. Muhonen, and Mohsen Kavehrad, "Look-Up Table Techniques for Adaptive Digital Predistortion: A Development and Comparison", IEEE Trans. Veh. Technol., Vol. 49, No. 5, Sep.2000, pp.1995-2000.

[8] Chih-Hung Lin, Hsin-Hung Chen, Yung-Yi Wang, and Jiunn-Tsair Chen, "Dynamically Optimum Lookup-Table Spacing for Power Amplifier Predistortion Linearization", IEEE Trans. Micro. Theory Tech., Vol. 54, No. 5, May.2006, pp. 2118 - 2127.

[9] Rubiao Xie, and Peiqing Jiang, Nonlinear Numerical Analysis, Press of Shanghai Jiaotong University, Shanghai, China, 1984.

[10] A. A. M. Saleh, "Frequency-independent and frequency-dependent nonlinear models of TWT amplifiers", IEEE Trans. Commun., Vol. COM-29, Nov. 1981, pp. 1715-1720.

I. J. Communications, Network and System Sciences. 2008; 1: 1-103
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

Particle Filtering with Multi Proposal Distributions

Fasheng WANG[1,2], Qingjie ZHAO[2], Yingbo ZHANG[1], Liqun ZHANG[2]
[1]*Neusoft Institute of Information, Northeastern University, P.R.China*
[2]*Beijing Key Lab of Intelligent Information*
School of Computer Science and Technology, Beijing Institute of Technology, P.R.China
E-mail: {wangfasheng, zhangyingbo}@neusoft.edu.cn
{zhaoqj, wangfasheng, zhangliqun}@bit.edu.cn

Abstract

Particle filtering algorithm has been applied to various fields due to its capacity to handle nonlinear/non-Gaussian dynamic problems. One crucial issue in particle filtering is the selection of the proposal distribution that generates the particles. In this paper, we give a novel strategy for selecting proposal distribution. Firstly, divide-conquer strategy is used, in which the particles used are divided into several parts. Afterward, different parts of particles are drawn from different proposal distributions. People can flexibly adjust how many of the particles drawn from specific proposal distributions according to their idiographic requirements. We provide simulation results that show its efficiency and performance.

Keywords: Sequential Monte Carlo, Proposal Distribution, Multiple Proposal Distributions

1. Introduction

As is known to all, most important real world applications are nonlinear and/or non-Gaussian. Nonlinear filtering problems exist in many fields including statistical signal processing, bio-statistics, economics, and engineering such as communications, radar tracking, sonar ranging, target tracking, car positioning, robot localization, and so on [1]. There are a variety of solutions for these problems, among which the extended Kalman filter (EKF) is well known. This kind of filters is based on linearizing technique that linearize the process model and measurement model by using Taylor series expansions. However, it is likely to diverge when the linearizing approximation methods give poor representations of the nonlinear functions.

The unscented Kalman filter (UKF) is another kind of interesting solutions, which is founded on the fact that it is easier to approximate a Gaussian distribution than it is to approximate an arbitrary nonlinear function [12][15]. Unlike the EKF, the UKF does not use the approximated models of the nonlinear process and the observation, but approximates the distribution of the state random variable by using a set of samples. It is shown that the UKF can acquire more accurate estimation results than the EKF can, but might lead to notable errors for non-Gaussian distributions.

In recent years, particle filters have drawn much attention in adaptive processing of nonlinear and non-Gaussian problems. It has been proved that the performance of particle filter is much better than traditional nonlinear filtering methods, such as the EKF, UKF, etc. [2][3].

The basic idea of particle filtering algorithm is to represent the required probability density function (PDF) by a set of random samples with associated weights. In the past decades, this method has been applied with great success to a variety of nonlinear/non-Gaussian filtering problems that was raised in communication fields, such as channel equalization [4], problems in MIMO wireless communication [5][6], phase tracking [7], problems in digital communication [19] etc. In addition, it is also widely used in vision tracking [3], robot localization [8-10], positioning and navigating [11], and so on.

A key issue in particle filtering algorithm is the selection of the proposal distribution which is generally hard to design. There are many proposal distributions proposed in the literature, among which the EKF [12][13] and the UKF [15] are well-known, as well as the transition prior [16]. Using EKF and UKF as proposal distributions, we get the Extended Kalman Particle Filter (EKPF) and the Unscented Particle Filter (UPF) [12][15],

This work was supported by the National Natural Science foundation of China, Granted NO.60772063

respectively. The famous CONDENSATION algorithm uses transition prior as the proposal distribution [16]. But the EKPF can diverge for strong nonlinearity cases, while the UPF is not applicable to general non-Gaussian model and has very high time cost. The CONDENSATION algorithm does not use the latest incoming information which contains very valuable data.

In order to solve the problems that are encountered by several existing particle filters, in this paper, we give a multi-proposal-distribution based particle filter, which is based on the EKF, UKF, and the transition prior. The particles used in the experiments are firstly divided into two parts, with one part (*c* percent) drawn from the EKF and UKF, another part (1-*c* percent) drawn from the transition prior. It gives not only higher accuracy but also lower time with proper choice of *c*.

The remainder of the paper is organized as follows. In Section 2, we give a brief introduction of particle filtering algorithm. The multi-proposal-distribution based particle filter (MPD-PF) is given in Section 3. Section 4 shows the simulation results. Conclusions are drawn in Section 5.

2. Particle Filtering

We adopt the following dynamical models:

$$x_k = f_k(x_{k-1}, v_{k-1}) \quad \textbf{(1)}$$

$$z_k = h_k(x_k, u_k) \quad \textbf{(2)}$$

where $x_k \in R^{n_x}$ denotes the system state at time *k*, and $z_k \in R^{n_z}$ denotes the observations at time *k*. The functions f_k and h_k are the system transition model function and measurement model function respectively. The noise processes are v_k and u_k, the process noise and measurement noise, respectively.

According to the basic idea of particle filters, let $\{x_{0:k}^i, w_k^i\}_{i=1}^N$ denotes a random measure used to represent the posterior density function $p(x_{0:k}|z_{1:k})$. $\{x_{0:k}^i, i = 0,..., N\}$ is a set of support particles with associated weights $\{w_k^i, i=0,\ldots,N\}$, and $x_{0:k}$ show the states up to time *k*. When *N* tends to the infinity, the posterior density function can be approximated by:

$$p(x_{0:k} \mid z_{1:k}) \approx \sum_{i=1}^{N} w_k^i \delta(x_{0:k} - x_{0:k}^i) \quad \textbf{(3)}$$

where δ(.) denotes the Dirac delta function.

Many of the particle filters rely on the principle of importance sampling [17]. Because it is usually difficult to directly draw particles from the posterior density, we can generate particles from a proposal distribution density function $q()$, which is also known as an importance function, and the weights are assigned according to:

$$w_k^i = p(x_{0:k} \mid z_{1:k}) / q(\cdot) \quad \textbf{(4)}$$

So the choice of proposal distribution $q(.)$ is of great importance. Assume the particles $x_{0:k}^i$ are drawn from an importance density $q(x_{0:k}|z_{1:k})$. Considering the following factorization:

$$q(x_{0:k} \mid z_{1:k}) = q(x_k \mid x_{0:k-1}, z_{1:k}) q(x_{0:k-1} \mid z_{1:k-1}) \quad \textbf{(5)}$$

Suppose we have gotten the approximation of the posterior distribution at time *k*-1, that is, $p(x_{0:k-1}|z_{1:k-1})$ can be represented by the particle set $\{x_{0:k-1}^i, w_{k-1}^i\}_{i=1}^N$, which are drawn from $x_{0:k-1}^i \sim q(x_{0:k-1} \mid y_{0:k-1})$. The particle weights are computed by using the formula:

$$w_{k-1}^i = p(x_{0:k-1}^i \mid z_{1:k-1}) / q(x_{0:k-1}^i \mid z_{1:k-1}) \quad \textbf{(6)}$$

In the following step, we aim to obtain $\{x_{0:k}^i, w_k^i\}_{i=1}^N$ using the new observation z_k. So long as we get the particles x_k^i and augment them onto the old particle trajectory, the new trajectory can be acquired. The new particles are generated as follows:

$$x_k^i \sim q(x_k \mid x_{0:k-1}^i, z_{0:k}) \quad \textbf{(7)}$$

The recursive equation for calculating weights can be obtained using Bayesian rules:

$$w_k^i \propto w_{k-1}^i \frac{p(z_k \mid x_k^i) p(x_k^i \mid x_{k-1}^i)}{q(x_k^i \mid x_{k-1}^i, z_k)} \quad \textbf{(8)}$$

Generally, the generic particle filter uses the transition prior $p(x_k|x_{k-1})$as the importance density function [12][16]. It has been proved that the proposal distribution $q(x_k \mid x_{0:k-1}, z_{1:k}) = p(x_k \mid x_{0:k-1}, z_{1:k})$ is optimal [12].

One of the drawbacks of particle filter is degeneracy problems. After several iterations, all but one particle would probably have negligible weights. In order to solve the problem, the resampling method is introduced [17], [18]. There are several resampling methods, such as systematic resampling, residual resmapling, etc. In this paper, the residual resampling method is used for all the experiments.

The generic particle filtering algorithm can be shown as algorithm 1.

Algorithm 1: Generic Particle Filter

Step 1. Initialization. k=0

(1) FOR i=0, …, N

Draw the states x_0^i from the prior $p(x_0)$;

END FOR

Step 2. FOR k=1, 2, …

(1) FOR i=1,…,N

Draw $x_k^i \sim q(x_k^i \mid x_{k-1}^i, z_k)$;

Assign the particle a weight according to Equ.(8);

END FOR

(2) FOR i=1,…,N

Normalize the weights $w_k^i = w_k^i / \sum_{j=1}^{N} w_k^j$;

END FOR

(3) Resample

- Eliminate the samples with low importance weights and multiply the samples with high importance weights, to obtain N random sample $x_{0:k}^i$ which are approximately distributed according to $p(x_{0:k} \mid z_{1:k})$;
- FOR i=1, …, N, let w_k^i =1/N, END FOR

END FOR

Step 3. k=k+1, goto Step2 or end executing.

3. MPD-Based Particle Filter

The particles used are firstly divided into two parts – c percent and 1-c percent. The MPD-PF first uses a mixed Kalman filter, which combines the UKF and the EKF, to draw c percent the particles. Afterwards, it uses its transition prior for another part – 1-c percent. Choice of c can affect the performance of the MPD-PF greatly. But one can choose it flexibly according to practical requirement.

For the c percent particles, suppose we have obtained the estimates of the state and the corresponding covariance at time k-1, $\bar{x}_{k-1}^{(i)}$ and $\hat{P}_{k-1}^{(i)}$. At the next time step k, the UKF is firstly used to update the particles.

The sigma points used in this process are calculated using the equation (see [12]),

$$\chi_{k-1}^{(i)a} = [\, \bar{x}_{k-1}^{(i)a} \quad \bar{x}_{k-1}^{(i)a} \pm \sqrt{(n_a + \lambda) P_{k-1}^{(i)a}} \,] \quad \textbf{(9)}$$

The particles are propagated into the future through the nonlinear models (equation (1) and (2)),

$$\chi_{k|k-1}^{(i)x} = f(\chi_{k-1}^{(i)x}, \chi_{k-1}^{(i)v}) \quad Z_{k|k-1}^{(i)} = h(\chi_{k|k-1}^{(i)x}, \chi_{k-1}^{(i)u}) \quad \textbf{(10)}$$

The predicted state and corresponding covariance can be obtained,

$$\bar{x}_{k|k-1_{ukf}}^{(i)} = \sum_{j=0}^{2n_a} W_j^{(m)} \chi_{j,k|k-1}^{(i)x} \quad \textbf{(11)}$$

$$P_{k|k-1_{ukf}}^{(i)} = \sum_{j=0}^{2n_a} W_j^{(c)} [\chi_{j,k|k-1}^{(i)x} - \bar{x}_{k|k-1_{ukf}}^{(i)}][\chi_{j,k|k-1}^{(i)x} - \bar{x}_{k|k-1_{ukf}}^{(i)}]^T \quad \textbf{(12)}$$

where $W_j^{(m)}$ and $W_j^{(c)}$ are weights of sigma points (see [12] and [14]), n_a=n_x+n_v+n_u. And, the predicted measurement can be computed using,

$$\bar{z}_{k|k-1_{ukf}}^{(i)} = \sum_{j=0}^{2n_a} W_j^{(m)} Z_{j,k|k-1_{ukf}}^{(i)} \quad \textbf{(13)}$$

If a new measurement z_k is obtained, update the predicted estimates as follow,

$$\bar{x}_{k\ ukf}^{(i)} = \bar{x}_{k|k-1_{ukf}}^{(i)} + K_k (z_k - \bar{z}_{k|k-1_{ukf}}^{(i)}) \quad \textbf{(14)}$$

$$\hat{P}_{k\ ukf}^{(i)} = P_{k|k-1}^{(i)} - K_k P_{\tilde{z}_k \tilde{z}_k} K_k^T \quad \textbf{(15)}$$

where K_k is the Kalman gain, calculated with equation $K_k = P_{x_k z_k} P_{\tilde{z}_k \tilde{z}_k}^{-1}$,

$$P_{\tilde{z}_k \tilde{z}_k} = \sum_{j=0}^{2n_a} W_j^{(c)} [Z_{j,k|k-1}^{(i)} - \bar{z}_{k|k-1}^{(i)}][Z_{j,k|k-1}^{(i)} - \bar{z}_{k|k-1}^{(i)}]^T \quad \textbf{(16)}$$

$$P_{\tilde{x}_k \tilde{z}_k} = \sum_{j=0}^{2n_a} W_j^{(c)} [\chi_{j,k|k-1}^{(i)} - \bar{x}_{k|k-1}^{(i)}][Z_{j,k|k-1}^{(i)} - \bar{z}_{k|k-1}^{(i)}]^T \quad \textbf{(17)}$$

Here, we have obtained state and corresponding covariance estimates through UKF-update process. Consequently, we use the EKF to do the update process with the pre-computed state estimate $\bar{x}_{k\ ukf}^{(i)}$ as input.

First, predict the state and covariance using the following,

$$\bar{x}_{k|k-1_{ekf}}^{(i)} = f(\bar{x}_{k-1}^{(i)}) = f(\bar{x}_{k\ ukf}^{(i)}) \quad \textbf{(18)}$$

$$P_{k|k-1_{ekf}}^{(i)} = F_k^{(i)} P_{k-1}^{(i)} F_k^{T(i)} + G_k^{(i)} Q_k G_k^{T(i)} \quad \textbf{(19)}$$

The Kalman gain can be calculated,

$$K_k = P_{k|k-1_{ekf}}^{(i)} H_k^{T(i)} [U_k^{(i)} R_k U_k^{T(i)} + H_k^{(i)} P_{k|k-1_{ekf}}^{(i)} H_k^{T(i)}]^{-1} \quad \textbf{(20)}$$

So, the updated state and covariance estimates at time k are,

$$\hat{P}_{k\ ekf}^{(i)} = P_{k|k-1_{ekf}}^{(i)} - K_k H_k^{(i)} P_{k|k-1_{ekf}}^{(i)} \quad \textbf{(21)}$$

$$\bar{x}_{k\ ekf}^{(i)} = \bar{x}_{k|k-1\ ekf}^{(i)} + \hat{P}_{k\ ekf}^{(i)} H_k^{T(i)} R_k^{-1}(z_k - h(\bar{x}_{k|k-1\ ekf}^{(i)})) \quad \textbf{(22)}$$

where Q is process noise covariance and R is measurement noise covariance. $F_k^{(i)}$, $G_k^{(i)}$ and $H_k^{(i)}$, $U_k^{(i)}$ are the Jacobians of the process and measurement models, respectively. The estimates $\bar{x}_{k\ ekf}^{(i)}$ and $\hat{P}_{k\ ukf}^{(i)}$ are the estimates to be computed at time k.

For the other 1-c percent of the particles, we can directly compute the state estimates using the state transition model (2),

$$\bar{x}_{k|k-1}^{(i)} = f(x_{k-1}^{(i)}) \quad \textbf{(23)}$$

The MPD-PF algorithm can be shown as in Algorithm 2.

Algorithm 2: The MPD-PF algorithm

Step 1. Initialization: k=0

(1) FOR i=1…N, draw the particles $x_0^{(i)}$ from the prior $p(x_0)$ and set:

$\bar{x}_0^{(i)} = E(x_0^{(i)})$

$P_0^{(i)} = E[(x_0^{(i)} - \bar{x}_0^{(i)})(x_0^{(i)} - \bar{x}_0^{(i)})^T]$

$\bar{x}_0^{(i)a} = E[x_0^{(i)a}] = [(x_0^{(i)})^T, 0, 0]^T$

$P_0^{(i)a} = E[(x_0^{(i)a} - \bar{x}_0^{(i)a})(x_0^{(i)a} - \bar{x}_0^{(i)a})^T] = diag(P_0^{(i)} \quad Q \quad R)$

ENDFOR

Step 2. FOR k=1,2,…

(1). FOR i=1,…,cN,

- Update the particles using the UKF, and obtain the state estimate $\bar{x}_{k\ ukf}^{(i)}$ and covariance estimate $\hat{P}_{k\ ukf}^{(i)}$.
- Using the EKF to do the update process:
  - Let $x_{k-1}^{(i)} = \bar{x}_{k\ ukf}^{(i)}$, and compute one-step-ahead estimates of the state and the covariance with the equations (18-19).
  - Compute the updated state and covariance estimates with the equations (21-22).
  - Let $\bar{x}_k^{(i)} = \bar{x}_{k\ ekf}^{(i)}$, $\hat{P}_k^{(i)} = \hat{P}_{k\ ukf}^{(i)}$.
- Draw $\hat{x}_k^{(i)} \sim q(x_k^{(i)} \mid x_{0:k-1}^{(i)}, z_{1:k}) = N(\bar{x}_k^{(i)}, \hat{P}_k^{(i)})$

//c percent particles are drawn here.

- Assign the particle a weight, w_k^i according to Equ. (8).

ENDFOR

(2). FOR i=cN+1,…,N

- Directly compute the state estimate using Equ. (23).
- Draw $\hat{x}_k^{(i)} \sim p(x_{k|k-1}^{(i)} \mid x_{k-1}^{(i)})$

//1-c percent particles are drawn here

- Assign the particle a weight.

(3). FOR i=1,…,N, Normalize the weights,

$$w_k^i = w_k^i / \sum_{j=1:N} w_k^j$$

ENDFOR

(4). RESAMPLE:

- Eliminate the samples with low importance weights and multiply the samples with high importance weights, to obtain N random samples $x_{0:k}^i$ approximately distributed according to the posterior PDF $p(x_{0:k}|z_{1:k})$.
- FOR i=1,…,N let w_k^i =1/N. ENDFOR

Step 3. k=k+1, goto Step2 or end executing.

4. Experimental Results

In this section, we present the experimental results of the MPD-PF algorithm and compare the performance of several existing particle filtering algorithms. The system and the measurement models used in the experiment are:

$$x_k = 1 + \sin[0.04\pi(k-1)] + 0.5x_{k-1} + v_{k-1} \quad \textbf{(24)}$$

$$z_k = \begin{cases} 0.2x_k^2 + u_k & k \le 30 \\ 0.5x_k - 2 + u_k & k > 30 \end{cases} \quad \textbf{(25)}$$

where v_k denotes a Gamma random variable and $\zeta_a(3,2)$ modeling the process noise, and the measurement noise u_k is drawn from a Gaussian distribution $N(0,0.0001)$. 200 particles are used and the process is repeated 100 times for time-steps k=1,..,60. The output of the algorithm is the mean of samples set which can be computed using (22):

$$\hat{x}_k = \frac{1}{N}\sum_{j=1}^{N} x_k^j \quad \textbf{(26)}$$

The Mean Square Errors (MSE) of each run is defined as

$$MSE = \left(\frac{1}{T}\sum_{k=1}^{T}(\hat{x}_k - x_k)^2\right)^{1/2} \quad \textbf{(27)}$$

The program run on a computer with CPU: Celeron 2.66GHz and Memory: 1GB.

If c is set to be 0.3, Figure 1 shows the comparison of the estimates to the system state generated from a single run of different particle filters. It is shown that the estimates of the PF and the EKPF bias the true state very large at some time steps, but the UPF and the MPD-PF can improve the performance.

Figure 2 shows the comparison of the Root MSE of different particle filters generated over 100 runs. It is clearly shown that the MPD-PF gives the best performance compared to other particle filters, which also

can be seen in Table 1. Table 2 gives the average time of the UPF and the MPD-PF spent after 100 independent runs. The UPF spent longer running-time – 26.8 seconds, while the MPD-PF spent much shorter – 17.6 seconds. So, from Table I and II, we can clearly see that the MPD-PF gives us much higher accuracy with lower time cost, when the parameter *c* is set to be 0.3.

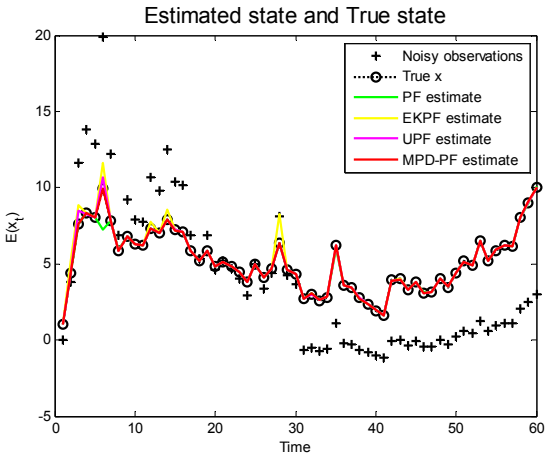


Figure 1. Estimates generated by different particle filters

Table 1. The means and the variances of the MSE over 100 independent runs (*c*=0.3).

Algorithm	MSE	
	Mean	variance
PF	0.21374	0.052091
EKPF	0.28551	0.02258
UPF	0.054599	0.004701
MPD-PF	0.016846	5.8999e-006

Figure 3 shows changing trend of the MSE of different particle filters, while the value of parameter c is increasing. From this figure, we can see that MPD-PF gives the best performance whatever the value of c. As to executing time of the particle filters, Figure 4 gives the changing trend of the executing time of the particle filters, from which we can see that, when the value of parameter c is turning bigger, the time cost of MPD-PF is increasing at the same time. At the point of *c*=0.8, it exceeds the UPF. For users with different requirements, they can flexibly adjust the value of parameter c according to their practical needs.

5. Conclusion

In this paper, multi-proposal-distribution based particle filter is introduced. It first takes a divide-conquer strategy, which draws the required particles using different proposals, and can give a better performance than several particle filters. The simulation results show that it can give much higher accuracy than the other particle filters, especially that, it can save a lot of executing time. With proper choice of *c*, the MPD-PF can supply us a fast and efficient way in dealing with some nonlinear filtering problems in real world applications, such as signal processing problems and nonlinear situations raised in wireless communications, etc. For people with different practical requirements, they can flexibly adjust the value of parameter *c* in order to obtain ideal results.

Table 2. The average time of UPF and MPD-PF used in the experiment

Algorithm	Time (seconds)
UPF	26.8
MPD-PF	16.62

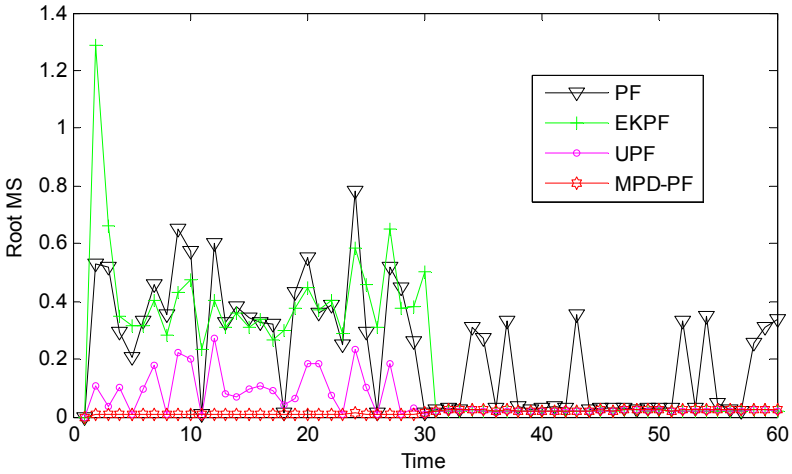


Figure 2. Root MSE of different particle filters after 100 independent runs (c=0.3).

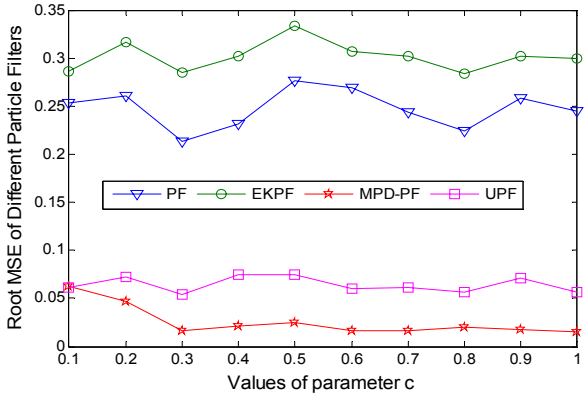


Figure 3. Changing trendline of MSE

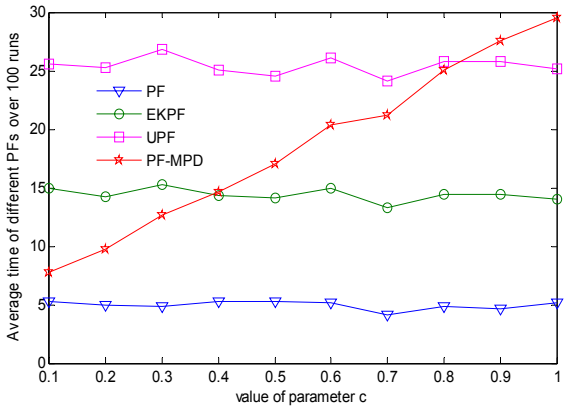


Figure 4. Changing trendline of execution time.

6. References

[1] J. H. Kotecha and P. M. Djuric, "Gaussian Particle Filtering", IEEE Transactions on signal processing. vol.51, no.10, 2003, pp. 2592-2601.

[2] N.J. Gordon, D.J. Salmond, A.F.M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation", IEE. proceedings-F, vol.140, no.2, 1993, pp.107-113

[3] Rui Y, Chen Y., "Better proposal distributions: Object tracking using unscented particle filter", IEEE Conf. on Computer Vision and Pattern Recognition, 2001, pp. 786−793.

[4] H. Kamel, W. Badawy. "Adaptive equalization of a communication channel in a non-Gaussian noise environment", Proc. of 3rd International IEEE-NEWCAS Conference, Jun. 19-22 2005. pp.395-398

[5] Y.M. Liang, H.W. Luo, X.X. Zhao, H.B. Zhang, C.G. Y. "Nonlinear Channel Estimation Based on Particle Filtering for MIMO-OFDM Systems", Proc. of International Conference on Communications, Circuits and Systems, Vol. 1. Jun. 2006. pp.347-351.

[6] S. Haykin, K. Huber, Zhe Chen. "Bayesian Sequential State Estimation for MIMO Wireless Communications". Proc. of the IEEE. Vol. 92 Issue 3. Mar. 2004. pp.439-454.

[7] Anis Ziadi, Gerard Salut. "Non-overlapping deterministic Gaussian particles in maximum likelihood non-linear filtering- phase tracking application". Proc. of International Symposium on Intelligent Signal Processing and Communication Systems. 2005. pp. 645-648

[8] FANG Zheng, TONG Guo-Feng, XU Xin-He, "A Robust and Efficient Algorithm for Mobile Robot Localization". ACTA Automatic Sinica, Vol. 33, NO. 1. Jan. 2007. pp. 48-53.

[9] Cody Kwok, Dieter Fox, and Marina Meila, "Real-Time Particle Filters", Proceedings of the IEEE, vol.92, no. 3, Mar. 2004, pp.469-484.

[10] Christian Plagemann, Dieter Fox, Wolfram Burgard, "Efficient Failure Detection on Mobile Robots Using Particle Filters with Gaussian Process Proposals", proceedings of International Joint Conference on Artificial Intelligence, Hyderabad, India. Jan. 6-12, 2007. pp. 2185-2190.

[11] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, J. Nordlund, "Particle filters for positioning, navigation, and tracking", IEEE Transactions on Signal Processing, 2002, pp.425−437.

[12] Rudolph van der Merwe, Arnaud Doucet, Nando de Freitas, Eric Wan, "The unscented particle filter", Technical report. Cambridge University. Engineering Department. 2000.

[13] Greg Welch, Gary Bishop, "An Introduction to the Kalman Filter", Technical Report, TR 95-041, University of North Carolina at Chapel Hill, 2004.

[14] Simon J. Julier and Jeffrey K. Uhlmann, "Unscented Filtering and Nonlinear Estimation", proceedings of the IEEE, vol. 92. no. 3, 2004, pp.401-422.

[15] Eric A. Wan and Rudolph van der Merwe, "The Unscented Kalman Filter for Nonlinear Estimation", proceedings of ASSPCC, 2000, pp.153-158.

[16] Michael Isard, Andrew Blake, "Condensation – conditional density propagation for visual tracking", International Journal of Computer Vision, 1998. pp. 5~28

[17] M. Sanjeev Arulampalam, Simon Maskell, N. Gordon and T. Clapp, "A tutorial on particle filters for On-line Nonlinear/Non-Gaussian Bayesian Tracking." IEEE Transactions on signal processing, vol.50, no.2, 2002, pp.174-188.

[18] Arnaud Doucet, "On Sequential Simulation-Based Methods for Bayesian Filtering", Technical report. Signal Processing Group, Department of Engineering, University of Cambridge, 1998.

[19] Tanya Bertozzi, Didier Le Ruyet, Gilles Rigal and Han Vu-Thien, "On Particle Filtering for Digital Communications", Proc. of 4th IEEE Workshop on Signal Processing Advances in Wireless Communications, 2003, pp570-574.

I. J. Communications, Network and System Sciences. **2008; 1: 1-103**
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

PAPR Reduction Scheme with SOCP for MIMO-OFDM Systems

Basel RIHAWI, Yves LOUET
SUPELEC - IETR
Avenue de la Boulaie CS 47601 35576 Cesson Sévigné, France
E-mail: {basel.rihawi,yves.louet}@supelec.fr

Abstract

Combination of multiple-input multiple-output (MIMO) with orthogonal frequency division multiplexing (OFDM) has become a promising candidate for high performance wireless communications. However one major disadvantage of MIMO-OFDM systems lies in a prohibitively large peak-to-average power ratio (PAPR) of the transmitted signal on each antenna. In this paper we extend from SISO to MIMO systems a method based on allocating dedicated subcarriers for PAPR mitigation. These subcarriers are located on unused subcarriers of OFDM spectrum under the assumption they all fall under the power mask. This is originally implemented with a SOCP optimization algorithm applied before space time coding scheme. This jointly mitigates PAPR on each MIMO branch scheme. This approach does not degrade the bit-error-rate (BER) and the data bit rate and no side information (SI) transmission is required. Simulation results are presented in the IEEE 802.16 WiMAX standard contexts: an Alamouti space time code with two transmitted antennas and 256 OFDM subcarriers are considered where 56 of which are unused and allocated for PAPR reduction. PAPR gains up to 7dB are obtained depending on mean power increase limitation. Moreover, with a spectrum mask constraint, this method is standard compliant.

Keywords: MIMO-OFDM, PAPR, SOCP

1. Introduction

MIMO radio systems have attracted considerable interest and have been studied intensively since Telatar [3] due to its potential of achieving high data rates in multipath channel. It is shown that when multiple transmit and receive antennas are used to form a MIMO system, the system capacity can be improved by a factor given by the minimum between transmitter and receiver antennas number, compared to a single-input single-output (SISO) system with flat Rayleigh fading or narrowband channels [4] For high data rate wireless wideband applications, MIMO combined with orthogonal frequency division multiplexing (OFDM) is being considered in a large number of current technology applications as in IEEE 802.16 WiMAX standard for example.

OFDM system has great advantages of having simple equalization and capability of transmitting high data bit rates over frequency selective channels. Nevertheless, its associated signal exhibits high PAPR. To counteract this well known problem, many PAPR reduction techniques have been proposed in the literature [5].

In the same way, OFDM and MIMO association (known as MIMO-OFDM) has faced the similar difficulties. In that case, large peaks can occur in any transmitted branch of the MIMO system. Some recent works have investigated MIMO-OFDM PAPR reduction via selective mapping [6] cross-antenna rotation and inversion [7], unitary rotation [8], polyphase interleaving and inversion [9] or optimal PAPR reduction [10]. All these methods imply SI transmission so as to recover useful data at the receiver which has to be inevitably modified if one of these methods is implemented. To avoid SI transmission, dedicated subcarriers have been generated for PAPR mitigation. In [2], unused subcarriers of OFDM based standards have been allocated and optimized for PAPR reduction under the assumption that they all fall under spectrum mask requirements. The receiver does not take these subcarriers into account and so needs no modification. It has been applied in a single input single output (SISO) context and with a SOCP optimization algorithm, what constitutes a natural extension of the well know Tone Reservation (TR) method [1].

In this paper, we apply SOCP algorithm (discussed in [2] for SISO-OFDM) to MIMO-OFDM systems. The objective is to mitigate the PAPR without any BER and date bit rate degradation and without sending any SI.

These dedicated subcarriers reveal mean power increase which should be controlled associated with spectrum mask requirements.

This paper is organized as follows. Section 2 defines OFDM signal, MIMO-OFDM systems and PAPR for MIMO-OFDM used in this study. Section 3 shows how the problem of PAPR reduction can be formulated as an optimization problem for MIMO-OFDM systems. Section 4 is devoted to simulation results in the IEEE 802.16 WiMAX standard context and exposes the constraints used for SOCP. Finally, Section 5 summarizes and concludes the paper.

2. OFDM Signal and PAPR Definitions for MIMO-OFDM

Before proceeding further, let us define the notations used throughout the paper. Time and frequency domain vectors are denoted by small and capital bold case letters respectively. The matrices are denoted by capital italic bold case letters.

In OFDM modulation, a block of N data symbols $X_k (k = 0,1,..., N-1)$, of vector **X**, will be transmitted in parallel such that each symbol (X_k) modulates a subcarrier $f_k (k = 0,1,..., N-1)$. The N subcarriers are orthogonal if $f_k = k\Delta f$, where $f_k = 1/T$ and T is the OFDM symbol period. The resulting analog OFDM base band signal $x(t)$ can be expressed as

$$x(t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j2\pi f_k t}, \quad t \in [0, T]. \tag{1}$$

In practice, the frequency complex symbol vector **X** is transmitted into a discrete-time signal **x** = $[x_0, x_1, ..., x_{N-1}]$ via an inverse discrete Fourier transform (IDFT) i.e.

$$\mathbf{x} = IDFT(\mathbf{X}). \tag{2}$$

Unfortunately, because of the statistical independence of all subcarriers, time-domain samples are approximately complex Gaussian distributed what results in high amplitude values. This is characterized by PAPR of signal x(t) defined by

$$PAPR\{x(t)\} = \frac{\max_{t\in[0,T]} |x(t)|^2}{E\{|x(t)^2|\}}, \tag{3}$$

where E{.} denotes the expectation operation. To compute easily PAPR, signal *x(t)* is sampled to get

$$x_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j2\pi \frac{kn}{NL}}, \quad n = 0,..., LN-1, \tag{4}$$

where *L* is an oversampling factor. It has been proved that *L=4* is sufficient for capturing the continuous-time peaks [11]. Then, the resulting temporal signal x is

$$\mathbf{x} = Q_L \mathbf{X}, \tag{5}$$

where Q_L is the IDFT matrix of size *NL* scaled by *L* (see Appendix 6.2). Subsequently, PAPR of **x** is defined as

$$PAPR\{\mathrm{x}\} = \frac{\max_{0\le k\le NL-1} |x_k|^2}{E\{|\mathrm{x}|^2\}}. \tag{6}$$

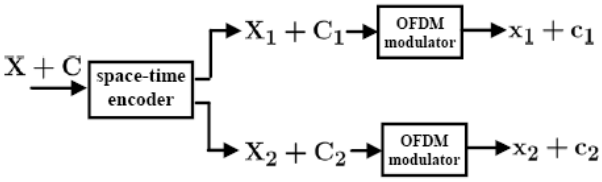


Figure 1. Structure of the two antennas MIMO-OFDM system

In the rest of this paper, all vectors are supposed to be oversampled by factor L. Moreover, we assume a space-time block code (STBC) for the studied MIMO-OFDM system (see Figure 1) that employs Alamouti scheme [12]. In this case,

$$\mathbf{X}_1 = [X_0 \; -X_1^* \ldots X_{\frac{LN}{2}-1} \; -X_{\frac{LN}{2}}^* \ldots X_{LN-2} \; -X_{LN-1}^*]$$

and

$$\mathbf{X}_2 = [X_1 \; X_0^* \ldots X_{\frac{LN}{2}} \; X_{\frac{LN}{2}-1}^* \ldots X_{LN-1} \; X_{LN-2}^*]$$

are the two outputs of Alamouti STBC (see Appendix 6.1). **C** is the corrective signal added to **X** in frequency domain in order to reduce PAPR (see Figure 2). $\mathbf{C} = [C_0 C_1 \ldots C_{\frac{LN}{2}-1} C_{\frac{LN}{2}} \ldots C_{LN-2} C_{LN-1}]$, $\mathbf{C}_1$ and $\mathbf{C}_2$ are generated by the same STBC. To avoid in and out of band distortions due to non linear power amplification, PAPR of all transmit signals should be simultaneously as small as possible. Since performance is governed by the worst-case PAPR, we define MIMO-OFDM PAPR $PAPR_{MIMO}$ as the maximum of all PAPR related to all N_T MIMO paths [14]. Subsequently,

$$PAPR_{MIMO} = \max_{i=1,\ldots,N_T} PAPR\{\mathbf{x}_i\}. \tag{7}$$

PAPR being a random variable, an appropriate description is its complementary cumulative distribution function (CCDF). It gives the probability P_0 of exceeding a given threshold $PAPR_0$, i.e.,

$$P_0 = \mathbf{Pr}(PAPR > PAPR_0). \tag{8}$$

The proposed strategy for PAPR reduction is to search an additive corrective signal c in order to verify PAPR(x+c) < PAPR(x). One way to find c is to use an optimization approach, what is detailed in next section.

3. Minimizing PAPR with SOCP Technique

3.1. General Principle of the Propose Method

PAPR mitigation principle used through out this paper is to add artificial subcarriers instead of unused ones. The signal addition is achieved in frequency domain as presented in [2] and illustrated in Figure 2.

Thus, PAPR of the resulting signal (**x**+**c**) is given by

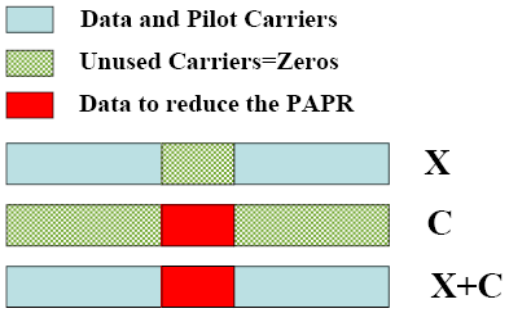


Figure 2. Principle of adding corrective subcarriers instead of unused ones for PAPR mitigation

$$PAPR\{\mathrm{x}+\mathrm{c}\}=\frac{\max\limits_{0\le k\le NL-1}\left|x_k+c_k\right|^2}{E\left\{\left|\mathrm{x}+\mathrm{c}\right|^2\right\}}. \tag{9}$$

Ideal computing would be to mitigate PAPR by minimizing the maximum peak of the combining signal (**x**+**c**=**x**+Q_{L}**C**) while keeping constant its average power. This objective can be formulated as

$$\min_{C}\max_{k}\left|x_k+q_{k,L}^{row}\mathbf{C}\right|, \tag{10}$$

where $q_{k,L}^{row}$ is the k-th row of Q_{L}.

Such a minimization problem given by Eq. 10 can be formulated as SOCP much easier than in semi definite program (SDP) or quadratically constrained quadratic program (QCQP) [13]. SOCP is a convex optimization problem class that minimizes a linear function over the intersection of an affine set and the product of second-order (quadratic) cones [13].

3.2. SOCP Approach

Tone reservation (TR) is an efficient technique to reduce the PAPR of a multicarrier signal [1]. This method is based on adding a data-block-dependent time domain signal to the original multicarrier signal to reduce its peaks. This time domain signal can be easily computed at the transmitter and stripped off at the receiver.

For this technique, the transmitter optimizes a given subset of subcarriers for PAPR reduction. The objective is to find the time domain signal to be added to the original time domain signal x. If we add a frequency domain vector $\mathbf{C}=[C_0,C_1,...,C_{LN-1}]$ to **X**, the new time domain signal can be represented as **x**+**c** = IDFT(**X**+**C**), where **c** is the time domain signal due to **C**. The TR technique restricts the data block **X** and peak reduction vector **C** to lie in disjoint frequency subspaces, i.e.

$$\begin{cases}X_n=0, & n\in\{i_1,i_2,...,i_v\}\\ C_n=0, & n\notin\{i_1,i_2,...,i_v\}.\end{cases}$$

The *v* nonzero positions in **C** are called peak reduction carriers. Due to orthogonality, these additional subcarriers cause no distortion on the useful data subcarriers. To find the value of $C_n, n\in\{i_1,i_2,...,i_v\}$, we must solve a convex optimization problem that can easily be modeled as a linear programming (LP) problem. To reduce the computational complexity of LP, a simple gradient algorithm is proposed in [1]. Nevertheless, TR proposed in [1] allocates subcarriers among those which carry useful information. This results in a data rate loss. In [2] TR has been extended to unused subcarriers of standards with SOCP algorithm. Significant PAPR reduction gains have been obtained, under spectrum mask constraints (all corrective subcarriers must fall under the spectrum mask) and with mean power increase control.

3.3. Application of MIMO-OFDM

From Figure 1, signals $\bar{x}_1$ and $\bar{x}_2$ at the outputs of the OFDM modulators are expressed as

$$\begin{aligned}\bar{\mathbf{x}}_1&=\mathbf{x_1}+\mathbf{c_1}\\&=\mathbf{x_1}+Q_L(\mathbf{C}\bullet\mathbf{A}_1-\mathbf{C}^*\bullet\mathbf{A}_2),\end{aligned} \tag{11}$$

$$\begin{aligned}\bar{\mathbf{x}}_2&=\mathbf{x_2}+\mathbf{c_2}\\&=\mathbf{x_2}+Q'_L(\mathbf{C}^*\bullet\mathbf{A}_1+\mathbf{C}\bullet\mathbf{A}_2),\end{aligned} \tag{12}$$

where $(.)^*$ denotes complex conjugate, $\bullet$ the element wise (dot) product, Q'_L the matrix deduced from Q_L by pair and odd rows permutations of Q_L (see Appendix 6.2), $\mathbf{A}_1=[1010...10]^T$ and $\mathbf{A}_2=[0101...01]^T$ NL×1 vectors (T denotes the transpose of the vectors).

Now PAPR of signals to be transmitted are

$$PAPR\left\{\bar{\mathrm{x}}_1\right\}=\frac{\max\limits_{k}\left|x_{1k}+c_{1k}\right|^2}{E\left\{\left|\mathrm{x}_1+\mathrm{c}_1\right|^2\right\}}, \tag{13}$$

$$PAPR\left\{\bar{\mathrm{x}}_2\right\}=\frac{\max_k\left|x_{2k}+c_{2k}\right|^2}{E\left\{\left|\mathrm{x}_2+\mathrm{c}_2\right|^2\right\}}. \quad (14)$$

According to Eq.(7), PAPR of MIMO system is

$$PAPR_{MIMO}=\max\left\{PAPR\left\{\bar{\mathrm{x}}_1\right\},PAPR\left\{\bar{\mathrm{x}}_2\right\}\right\}. \quad (15)$$

The objective of this method is to reduce $PAPR_{MIMO}$ by jointly reducing $PAPR\{\bar{\mathbf{x}}_1\}$ and $PAPR\{\bar{\mathbf{x}}_2\}$. At first, SOCP algorithm does not take into account the mean power increase. So, the problem of PAPR minimization can be formulated as

$$\begin{aligned}&\min_{\mathbf{c}_1}\max_k\left|x_{1k}+c_{1k}\right|^2=\min_{\mathbf{C}_1}\left\|\mathbf{x_1}+\mathbf{c}_1\right\|_\infty^2\\&=\min_{\mathbf{C}}\left\|\mathbf{x_1}+Q_L(\mathbf{C}\bullet\mathbf{A}_1-\mathbf{C}^*\bullet\mathbf{A}_2)\right\|_\infty^2,\end{aligned} \quad (16)$$

$$\begin{aligned}&\min_{\mathbf{c}_2}\max_k\left|x_{2k}+c_{2k}\right|^2=\min_{C_2}\left\|\mathbf{x_2}+\mathbf{c}_2\right\|_\infty^2\\&=\min_{\mathbf{C}}\left\|\mathbf{x_2}+Q'_L(\mathbf{C}^*\bullet\mathbf{A}_1+\mathbf{C}\bullet\mathbf{A}_2)\right\|_\infty^2.\end{aligned} \quad (17)$$

Eqs.(16) and (17) are convex optimization problems formulated as one with SOCP :

$$\begin{aligned}&\min\beta\\&subject\ to\quad\left\|\mathbf{x_1}+Q_L(\mathbf{C}\bullet\mathbf{A}_1-\mathbf{C}^*\bullet\mathbf{A}_2)\right\|_\infty\le\beta\\&\qquad\left\|\mathbf{x_2}+Q'_L(\mathbf{C}^*\bullet\mathbf{A}_1+\mathbf{C}\bullet\mathbf{A}_2)\right\|_\infty\le\beta.\end{aligned} \quad (18)$$

SOCP algorithm provides optimized solution $\mathbf{C}_{opt}$ according to Eq.(18). Nevertheless, the two resulting signals $\mathbf{x_i}+Q_L\mathbf{C}_{opt}(i=1,2)$ have larger mean powers compared to x what has to be taken into account for practical applications. Moreover, $\mathbf{C}_{opt}$ components must fall under the spectrum mask requirements to make the signal standard compliant. These two points are now discussed.

4. Constraints Statement and Simulation Results

4.1. Mean Power Increase Constraint

Figure 3 shows PAPR CCDF of original and optimized signals for randomly generated QPSK symbols. MIMO-OFDM system uses two antennas with N=256 subcarriers per antenna. 56 unused subcarriers are used as the corrective signal. This system is inspired from IEEE 802.16 WiMAX standard. We set the oversampling factor L to a value of 4. As we can see, PAPR is reduced by 7dB for a 10^{-3} exceed probability level without any mean power constraint.

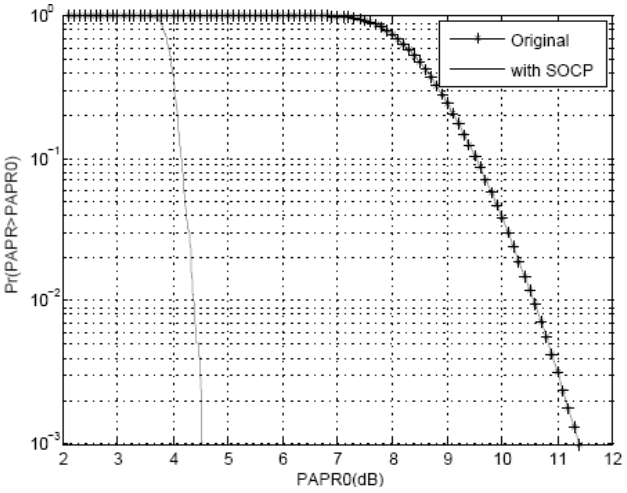


Figure 3. PAPR reduction using SOCP without any constraint in IEEE 802.16 context

However, adding a signal c to an original signal **x** to reduce its PAPR increases the mean transmit power. The relative mean increase power ΔE is defined as [1]

$$\Delta E=10\log_{10}\frac{E(\left|\mathbf{x}+\mathbf{c}\right|^2)}{E(\mathbf{x}^2)}. \quad (19)$$

This parameter must be as small as possible in order to avoid power amplification saturation. Indeed, it is easy to understand that if average power increases indefinitely **x**+**c** would obviously have a resulting PAPR of 0dB but the signal cannot be transmitted. Thus, the relative mean power must be upper bounded and can be written as

$$\Delta E_{dB}\le\gamma, \quad (20)$$

what implies

$$E(\left|\mathbf{x}+\mathbf{c}\right|^2)\le\lambda\,E(\left|\mathbf{x}\right|^2), \quad (21)$$

where $\lambda=10^{\frac{\gamma}{10}}$.γ is a constant related to the power amplifier characteristics. The above condition can be added as a constraint in the optimization problem. By taking these conditions into account, SOCP algorithm becomes

$$\begin{aligned}&\min\beta\\&subject\ to\quad\left\|\mathbf{x_1}+Q_L(\mathbf{C}\bullet\mathbf{A}_1-\mathbf{C}^*\bullet\mathbf{A}_2)\right\|_\infty\le\beta\\&\qquad\left\|\mathbf{x_2}+Q'_L(\mathbf{C}^*\bullet\mathbf{A}_1+\mathbf{C}\bullet\mathbf{A}_2)\right\|_\infty\le\beta\\&\qquad\left\|\mathbf{x_1}+Q_L(\mathbf{C}\bullet\mathbf{A}_1-\mathbf{C}^*\bullet\mathbf{A}_2)\right\|_\infty\le\sqrt{\lambda K_1}\\&\qquad\left\|\mathbf{x_2}+Q'_L(\mathbf{C}^*\bullet\mathbf{A}_1+\mathbf{C}\bullet\mathbf{A}_2)\right\|_\infty\le\sqrt{\lambda K_2},\end{aligned} \quad (22)$$

where $K_i=NL.E(\left|\mathbf{x}_i\right|^2),(i=1,2)$.

Then, it has been shown that if γ=0.2 dB, the additional allocated power for PAPR reduction represents 4.7% of the total amount. Figure 4 plots CCDFs of PAPR for several γ values. As we can see, PAPR decreases as γ increases.

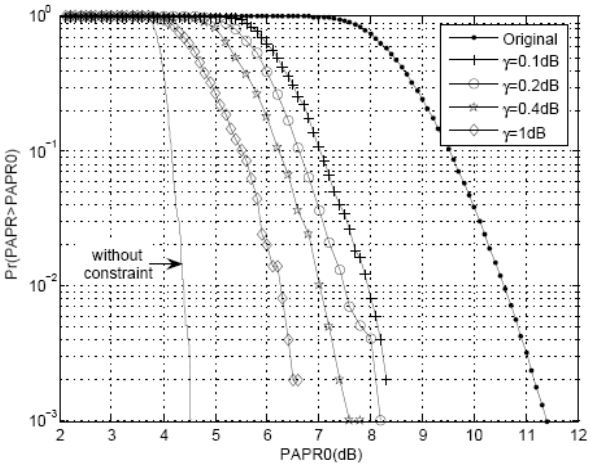


Figure 4. PAPR reduction with SOCP with mean power constraint

Nevertheless, even if mean power control is mandatory as explained above, optimized subcarriers for PAPR reduction has to fit spectrum requirements referring to standard masks. It is clear that in some cases, **x**+**c** values exceed spectrum mask as shown in Figure 5 for γ=0.2dB. Subsequently, an additional constraint has to be included in SOCP formulation related to spectrum mask, which is considered in next subsection.

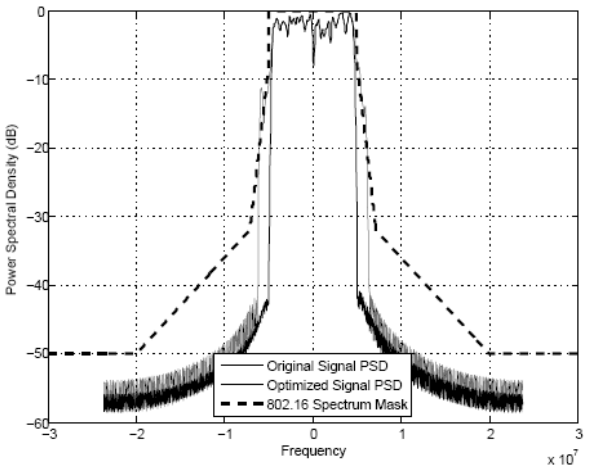


Figure 5. OFDM power spectral density (PSD) and IEEE 802.16 mask under mean power constraint (γ=0.2dB)

4.2. Spectrum Mask Constraint

A second constraint is added to SOCP. It concerns spectrum mask and is formulated as

$$\left|C_{i,n}\right| \le \delta_n, n \in \Omega, \tag{23}$$

where $i = \{1,2\}$, δ refers to spectrum mask values and Ω is the set of all allocated subcarriers indexes.

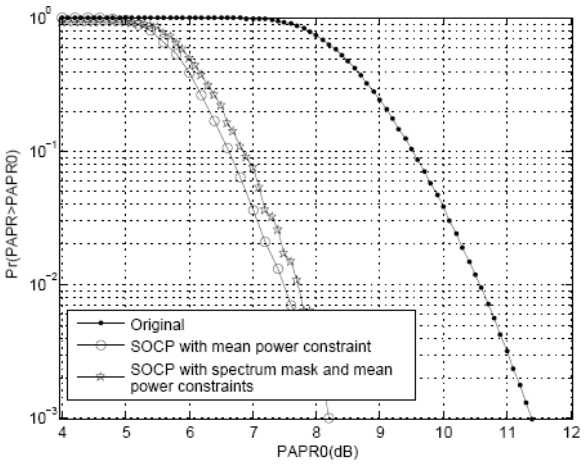


Figure 6. PAPR reduction with mean power and spectrum mask constraints compared to that with only mean power constraints γ=0.2dB in both cases

Figure 6 compares performance of two SOCP schemes: the first one considers only mean power constraint and the second one considers both mean power and spectrum mask constraints. In both cases the mean power limitation is the same (γ=0.2dB). As a result, it is shown that this additional mask constraint slightly degrades PAPR compared to the case where only mean power constraint is considered but with the gain that spectrum requirements are fulfilled. Figure 7 shows the power density function of the OFDM signal with IEEE 802.16 spectrum mask when the two constraints are considered. As expected, the resulting spectrum respects the mask.

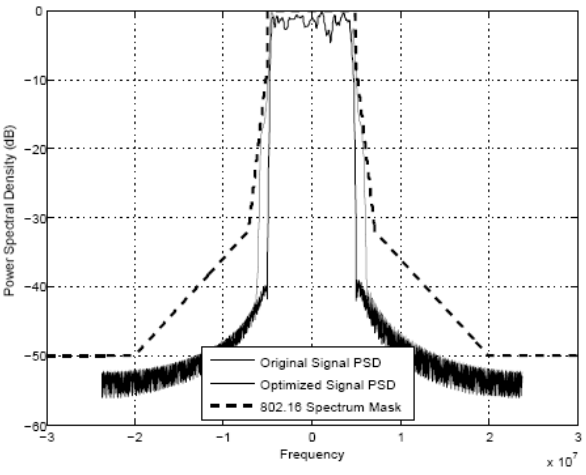


Figure 7. OFDM power spectral density (PSD) and IEEE 802.16 mask under mean power (γ=0.2dB) and spectrum mask constraints

5. Conclusion

In this paper, we have proposed a novel PAPR

reduction approach for MIMO-OFDM. It is based on SOCP and provides significant PAPR reduction gains without transmitting any side information and without degrading BER and data rate at the same time. To do so, we have used the unused subcarriers of standards to generate the corrective signal that jointly mitigates the PAPR of the MIMO-OFDM scheme. We have shown that PAPR can be reduced by 7dB at the 10^{-3} probability exceed level. To avoid an uncontrolled increase of the relative mean power, we have included a mean power constraint in SOCP algorithm. Finally, to be standard compliant, the allocated subcarriers for PAPR reduction have to fall under the spectrum mask, which has been done in IEEE 802.16 WiMAX standard context.

6. Appendices

6.1. Alamouti STBC

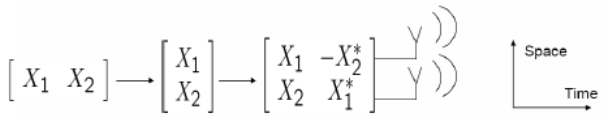


Figure 8. Alamouti STBC for 2 antennas

6.2. Fourier matrices Q_L and Q'_L

$$Q_L = \frac{1}{\sqrt{LN}} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{j\frac{2\pi}{LN}1.1} & \cdots & e^{j\frac{2\pi}{LN}1.(LN-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{j\frac{2\pi}{LN}(LN-2).1} & \cdots & e^{j\frac{\pi}{LN}(LN-2)(LN-1)} \\ 1 & e^{j\frac{2\pi}{LN}(LN-1).1} & \cdots & e^{j\frac{\pi}{LN}(LN-1)(LN-1)} \end{bmatrix}$$

$$Q'_L = \frac{1}{\sqrt{LN}} \begin{bmatrix} 1 & e^{j\frac{2\pi}{LN}1.1} & \cdots & e^{j\frac{2\pi}{LN}1.(LN-1)} \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{j\frac{2\pi}{LN}(LN-1).1} & \cdots & e^{j\frac{\pi}{LN}(LN-1)(LN-1)} \\ 1 & e^{j\frac{2\pi}{LN}(LN-2).1} & \cdots & e^{j\frac{\pi}{LN}(LN-2)(LN-1)} \end{bmatrix}$$

7. References

[1] J. Tellado-Mourelo, "Peak to Average Power Ratio Reduction for multicarrier modulation", PhD thesis, Stanford University, Sept. 1999.

[2] S. Zabre, J. Palicot, Y. Louet, and C. Lereau, "SOCP approach for OFDM peak-to-average power ratio reduction in the signal adding context", IEEE ISSPIT 06, Vancouver, Canada, Aug. 2006.

[3] I. E. Telatar, "Capacity of multi-antenna Gaussian channels", AT T Bell Labs Internal Tech. Memo., Jun. 1995.

[4] G. J. Foschini, and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas", Wireless Personal Communication, vol.6, Mar. 1998, pp. 314-335.

[5] S. H. Han, and J. H. Lee, "An overview of peak-toaverage power ratio reduction techniques for multicarrier transmission", IEEE Wireless Communication, vol.12, no.2, Apr. 2005, pp.56-65.

[6] Y. Lee, Y. You, W. Jeon, J. Paik, and H. Song, "Peak-to-average power ratio in MIMO-OFDM systems using selective mapping", IEEE Commun. Letters, vol.7, no.12, Dec. 2003, pp. 575-577.

[7] M. Tan, Z. Latinovic, and Y. Bar-Ness, "STBC MIMO-OFDM Peak-to-Average Power Ratio Reduction by Cross-Antenna Rotation and Inversion", IEEE Commun. Letters, vol.9, no.7, Jul. 2005.

[8] H.Lee, D. N. Liu, W. Zhu, and M. P. Fitz, "Peak power reduction using a unitary rotation in multiple transmit antennas", in Proc. IEEE International Conference on Communication, vol.4, Seoul, South Korea, May 2005, pp.2407-2411.

[9] Z. Latinovic, and Y. Bar-Ness, "SFBC MIMO-OFDM Peak-to-Average Power Ratio Reduction by Polyphase Interleaving and Inversion", IEEE Commun. Letters, vol.10, no.4, Apr. 2006.

[10] Aggarwal, A. Sauffr, and E.R. Meng T.H, "Optimal Peak-to-Average Power Ratio Reduction in MIMO Systems", IEEE International Conference on Commnication (ICC 06), vol.7, Istanbul, Turkey, Jun. 2006, pp.3094-3099.

[11] G. Wunder, and H. Boche, "Peak value estimation of bandlimited signals from their samples, noise enhancement, and a local characterization in the neighborhood of an extremum", IEEE Trans. Signal Processing, Mar. 2003, pp.771-780.

[12] Siavash M. Alamouti, "A simple Transmit Diversity Technique forWireless communications", IEEE Journal on Select Areas in Communications, vol.16, no.8, Oct. 1998.

[13] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second order cone programming", Linear Algebra and its Applications, 284:193-228, Special Issue on Linear Algebra in Control, Signals and Image Processing, Nov. 1998.

[14] R.F.H. Fischer, and M. Hoch, "Peak-to-Average Power Ratio Reduction in MIMO OFDM", IEEE International Conference on Communications (ICC 2007), Glasgow, United Kingdom, Jun. 2007.

[15] M. Grant, S. Boyd, and Y. Ye, "CVX – Matlab Software for Disciplined Convex Programming", http://www.stanford.edu/ boyd/cvx/

I. J. Communications, Network and System Sciences. **2008; 1: 1-103**
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

Any Resource Sharing via HWN* Routing

Chong SHEN, Susan REA, Dirk PESCH
Centre for Adaptive Wireless Systems, Department of Electronic Engineering
Cork Institute of Technology Ireland
E-mail: {chong.shen, susan.rea, dirk.pesch}@cit.ie

Abstract

In this paper we present an adaptive distributed cross-layer routing algorithm (ADCR) for hybrid wireless network with dedicated relay stations (HWN*). To support versatile multimedia communication for Mobile Terminals (MTs), the HWN* integrates a cellular network, an ad hoc network and fixed relay nodes (RNs). A MT may borrow cellular data channels that are available thousands mile away via secure multi-hop RNs, where RNs are placed at flexible locations in the network. The MT can also communicate with each other or access internet ubiquitously. We discuss cross media access and network layers routing issues. The ADCR establishes routing paths across RNs or cellular network while providing appropriate QoS (quality of service). Through simulation, we verify the routing performance benefits of HWN* over conventional cellular systems and other hybrid network frameworks. It is anticipated that the simulation results reported in this paper will serve as a guideline for research based on distributed source routing involving heterogeneous wireless technologies.

Keywords: Routing, QoS, Cross-layer, Traffic Management, HWN*

1. Introduction

The recent widespread uptake of high-data rate and multimedia services has led to a shift in how cellular networks are being used. This is motivating wireless providers to develop innovative infrastructure and accompanying routing and radio access algorithms. Due to the small size of the cells of micro cellular networks and the uneven nature of the time varying spatial distribution [1], network performance metrics such as overall throughput, Mobile Terminal (MT) throughput, call blocking probability, handover rate, end-to-end delay, etc. are not sufficient for today's wireless network where more ad hoc features are being introduced. The traffic in heterogeneous environments is a mixture of real-time packets with widely varying support on Quality of Service (QoS) levels. Recently, Channel Access Scheduling (CAS), Medium Access Control (MAC) and distributed routing schemes were investigated for Mobile Ad Hoc Networks (MANETs) with scenarios involving military networks, emergency services, and inexpensive deployment of sensor networks [2]. But these are energy constrained networks with limitations on capacity and QoS support since as the ad hoc network topologies can rapidly change.

To effectively manage problems stated above, we propose to combine the advantages of different networks so that the MTs can utilise an optimised MANET or the Base Station Oriented Network (BSON) and packet relay

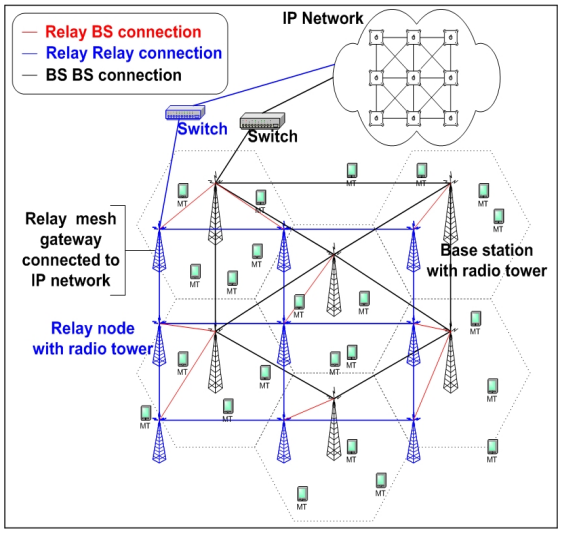


Figure 1. The hybrid wireless network with fixed relay stations and IP network

services. Figure 1 presents our Hybrid Wireless Network with Relay Nodes (HWN*) connected to the Internet Protocol (IP) networks where relay nodes (RNs) compose a mesh like structure through RN gateways, while Base Stations (BSs) are connected to the IP networks via switches. Other than the improvement of downlink power efficiency, the primary goal of the RN incorporation in HWN* is to provide uniform coverage and support the hybrid network relay. Two MTs may communicate directly, or through an intermediate node. This node can

be a MT, a fixed RN or a group of RN matrices. When a MT transmits packets to a BS through RNs, the RNs extend the signaling coverage of BSON thus we can expect an enhanced resource sharing performance. A further discussion on the HWN* topology description can be found in [3].

We propose an Adaptive Distributed Cross-layer Routing (ADCR) algorithm in HWN* to (a) route packet using the minimal number of relay hops to reduce latency

Section 3, we present the cellular network, ad hoc network and relay network formation algorithm. The practical RN positioning strategy is also investigated to boost the HWN* performance. We then discuss ADCR algorithm design criteria with three traffic transmission modes in ***Section 4***. A novel MT mobility model is presented in ***Section 5***. In ***Section 6***, we verify the ADCR performance benefits of the HWN* over conventional cellular system, ad hoc network and other hybrid wireless networks in terms of network capacity, per MT capacity,

Table 1. Recent Hybrid Networks Classification and Summary

Project	**HWN***	**WINNER**	**SOPRANO**
Basic Infrastructure	Incorporate a MANET to increase system capacity while realising differentiated QoS services	Novel interface technologies for ubiquitous networks	Test CDMA and connect a cellular network with an IP and MANET network
Main objectives	BSON, BSON with RN, MANET, MANET with RN	Unified radio access technology, RN and BS may change transmission range	RNs transfer the traffic from a hot spot to the neighbouring cooler
Fixed node Antenna	Not investigated	Smart antenna / directional antenna	Not investigated
Traffic handover	Heterogeneous network handover with QoS support	Ubiquitous network handover	Not investigated
Call admission	Coordinated admission controlled by BS	Coordinated admission controlled by both BS and RN	Central admission controlled by BS
Routing issues	BS switch and RN assisted traffic diversion	BS and RN switch	BS switch
Load Balance	Multihop based load balancing considering QoS	Not investigated	Multihop load balancing

preserve communications, (b) deliver good overall throughput and per node throughput and (c) enable low-complexity router implementation. A cross-layer network design [4] that seeks to enhance the performance of a system by jointly designing MAC and NETWORK layers is desirable. In the design stage of the ADCR algorithm development, we analysed the theoretical cellular network media access capacity, multi-hop traffic relaying issues and inter network traffic handovers. The cascaded ADCR scheme includes three sub packet transmission modes labeled as One-Hop Ad-Hoc Transmission (OHAHT) for point to point ad hoc direct communication, Multi-Hop Combined Transmission (MHCT) for cellular resource relaying using fixed RNs and Cellular Transmission (CT) for traditional cellular service. Our approach allows upper layers to better adapt their strategies to varying link and network conditions while the resulting flexibility helps to improve access speed, end-to-end delay and dynamic resource management performance.

This paper begins with a hybrid wireless networks literary review, including up to date network deployment scenarios and infrastructure dimensioning analysis. In

access speed and end-to-end delay. Finally in ***Section 7***, we present conclusions and discuss future directions for this work.

2. Hybrid Wireless Networks

Based on hop distance, we can classify hybrid wireless networks as single-mode, where MTs only perform single hop communication, or multi-mode, where either multi-hop or single-hop MT communications occur. Recently, the IST WINNER project proposed a novel hybrid relay network [5] to setup new 4G standards in Europe. This work mainly focuses on specific radio interface technologies, which are needed for a ubiquitous radio system. The RNs share the same RAT with BSs and MTs to realise dynamic spectrum usage.

Multi-hop Cellular Network (MCN), Multi-Power Architecture for Cellular network (MuPAC), integrated Cellular and Ad hoc relaying system (iCAR), Self-organising Packet Radio Networks with Overlay (SOPRANO) [6] and Hybrid Wireless Network (HWN) [7] are the architecture designs that have been proposed

for hybrid networks. The iCAR is derived from existing cellular networks and enable the network to achieve theoretical capacity through adaptive traffic load balancing. Excessive bandwidth in surrounding cells can be borrowed for the congested cell through RNs with primary, secondary and cascaded orders. The SOPRANO is a scalable architecture that assumes the use of asynchronous Code Division Multiple Access (CDMA) with spreading codes to support high data rate internet & multimedia traffic. It is similar to iCAR other than IP network support and cross network connection methods. The HWN, without RN support, requires Global Positioning System (GPS) capability to extract geographical location. Each cell operates either in cellular structures or ad hoc structures depending on the MTs topology information from the GPS. Table 1 presents the main features for the HWN*, WINNER and SOPRANO architectures. A comparison of the iCar, MuPAC, HWN and MCN can be found in [8].

Table 2 identifies the technologies used and the features considered for each of the four HWN* communication structures. This table specifies the physical layer mode; media access method, spectrum usage, mobility characteristics and data rate. Time Duplex Division (TDD) is implemented on all four modes: BSON, MANET, RN supported BSON and RN supported MANET. Typical physical layer parameters are used [9]. The log-normal standard deviation σ is set as 10 dB, shadowing correlation distance χ_s is set to 50 m, and the mean *SIR* value r_d is set to 17 dB. The default energy mode provided by OMNET++ [10] is implemented, specifically, for a 250m transmission range the transmit power used is 0.282W. The transmit power used for a transmission range of d is proportional to d^4. For the MANET and RN supported MANET, we implemented the CSMA/CA Distributed Coordination Function (DCF) of the IEEE 802.11 standard. The air interface can be adapted for TDD/CDMA of 3GPP as described in [11], where multiple data rates are achieved by using various spreading codes.

In Time Division Multiplex Access (TDMA) cellular systems, when an MT is involved in a call admission or handover process in a congested cell, but is not able to find an admission channel or alternative channel, respectively, the data transmission will be terminated. For example, consider a scenario in Figure 2 where MT A is currently connected to MT B and is moving out of Cell 1 into Cell 6. A request for a BS handover will be sent as soon as the power level by MT A goes below a certain threshold, its trajectory is indicated by the red line. A successful handover will take place within a few hundred milliseconds depending on speed before the received power from BSs reaches an unacceptable level. When MT A arrives in Cell 6, if the congestion persists for a period of time during which the MT moves farther away from the other neighbouring cell border, thus causing the

Table 2 HWN* Four Communication Structures

	BS cellular network	Ad hoc network	RN supported cellular network	RN supported MANET
Physical layer	Time division duplex	Time division duplex	Time division duplex	Time division duplex
Media access layer	TDMA	CSMA/CA	TDMA	CSMA/CA
Spectrum regulation	Licensed	Unlicensed	Proposed to use same spectrum as cellular network	Proposed to use same spectrum as cellular network
Mobility speed	Low, Medium and High	Low	Low, Medium and High	Low and Medium
Transmission data rate	Low, Medium and High	Low	Low, Medium and High	Low and Medium

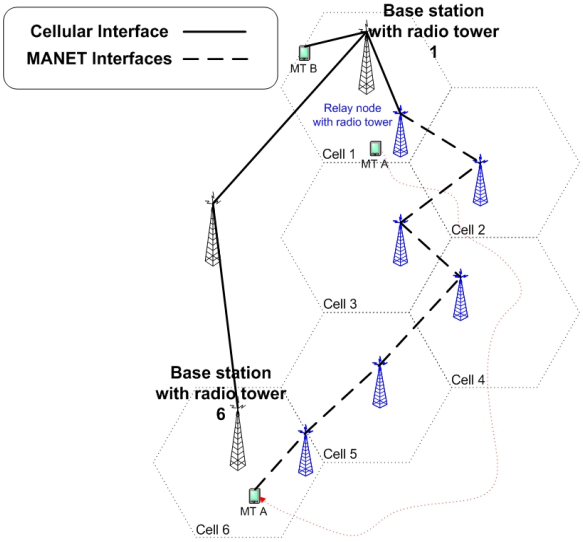


Figure 2. Multi-Hop Combined Transmission Example of cellular resource relaying using fixed RNs

received power level from BS A to fall below the acceptable level, handover will fail and the call will be permanently terminated.

However in MHCT mode of HWN*, the data session does not have to be dropped even though the congestion in Cell 6 persists. For example, when MT A moves into the congested Cell 6, apart from trying cellular connections, it also associates itself with a RN using ad hoc frequencies, then the RN may continue the data session with any BS in the network via the multi-hop relaying structure and the relaying path can be also extended to the area with no cellular coverage.

Since the ultimate goal of this framework is to improve the resource sharing performance in terms of data admission, traffic handover and routing optimization, in addition, OHAHT of point-to-point ad hoc communications is added to the routing strategy to further balance traffic load while sharing media resources. The analytical as well as simulation results we experienced have proven that inter-network traffic management can significantly improve the grade of service, reduce the traffic blocking probability, while maintaining the individual MT QoS.

The spectrum for each RN could be also allocated dynamically. Multiple non-interfering relay frequencies operate in parallel through the use of intelligent radios. The spectrum where a RN operates can be leased for a limited time depending on the network status. The spectrum on which it is operating is reclaimed when network performance improves. For example, two RNs operating on non-interfering spectrums form a network relay link with multiple orthogonal bands. Multiple nodes within range of each other may also transmit simultaneously on different channels without relying on a media access protocol or distributed scheduling algorithm to resolve contention.

Although the large scale deployment of dual-interfaced RNs could be costly, the use of RNs extends the BSON service range, optimizes cell capacity, minimizes transmit power, covers shadowed areas, supports inter network load balancing and supports MANET routing. Theoretically, both the HWN* system capacity and the transport capacity per MT, when compared to a cellular network, should be improved because the RNs provide relay capability as the substitution of a poor quality single-hop wireless link with a better-quality link being encouraged whenever possible. Also a higher end-to-end data rate could be obtained if a MT had two simultaneously communicating interfaces.

Using three scaling approaches proposed in [12], we can implement simulation dimensioning and estimate how many RNs should be deployed when the number of MTs changes. The three parameters are the number of RNs **m**, the number of MTs **n** and the system capacity **C**. The asymptotic scaling for the per user throughput as **n** becomes large is:

$$m \le \sqrt{n/\log n} \tag{1}$$

The per user throughput is of the order $C\sqrt{n/\log n}$ and can be realised by allowing only ad hoc communications which does not necessarily need RN support, when:

$$\sqrt{n/\log n} \le m \le n/\log n \tag{2}$$

The order for the per user throughput is Cm/n therefore the total additional bandwidth provided by **m** RNs is effectively shared among **n** MTs. Finally, when:

$$n/\log n \le m \tag{3}$$

the order of the per user throughput is only $C/\log n$ which implies that further investments in relay nodes will not lead to an improvement in throughput and bandwidth optimization.

3. HWN* Formation Algorithm

In this section we present the basis for our HWN* fixed RN placement plan and related formation algorithm. The positioning and formation scheme are addressed where RNs operate in the same ad hoc frequency band.

Various plans have already been proposed to select sites for fixed RNs, the solutions depend on the varied network performance objectives. The network spectral efficiency was taken by [13] as the objective to optimize RN positioning. For HWN* we adopt a bandwidth oriented criterion to investigate the proper RN positioning, the RN positioning is formulated as a constrained optimization problem, of which the goal is to maximize the overall network throughput and per node throughput so that 95% of MTs are better served with guaranteed quality than by the conventional cellular network without relaying. By imposing such a constraint we can improve the performance for MTs at disadvantaged locations and enlarge network dimensioning and help deliver a more uniform communication experience. In searching for the proper RN locations, [13] made the assumption that the quality on the links connecting BS and RN is always better than the link between RN to from RN. Therefore the authors stated that this assumption can be satisfied by establishing Line of Sight (LOS) links between the BS and RN or by designing links that enhance the antenna gains. But the former solution imposes extra difficulty on network planning, while the latter complicates the transceiver design. Therefore, it is of importance to investigate the RN positioning with limited change on existing radio technologies to save the infrastructure cost.

When considering pre-engineered RN deployment, it

is well known from planar geometry that to cover a two dimensional district with equal sized circles, the best possible packing solution can be obtained by surrounding each circle by six circles as shown in Figure 3. But to have connections between the RNs, we need an overlap between relay cells. We therefore consider a situation where the location of the RNs is pre-computed and the best deployment strategy is chosen. The deployments shown in Figure 3 are two examples of such pre-engineered approach with a specific number of RNs in HWN*. The first deployment tries to cover the entire area, without considering MT mobility and service requirement. The second one tries to cover densely populated regions of MTs near the edge of the cell. Our MT mobility model is based on a city layout that consists of **n** probability based attractor points that MTs will move towards.

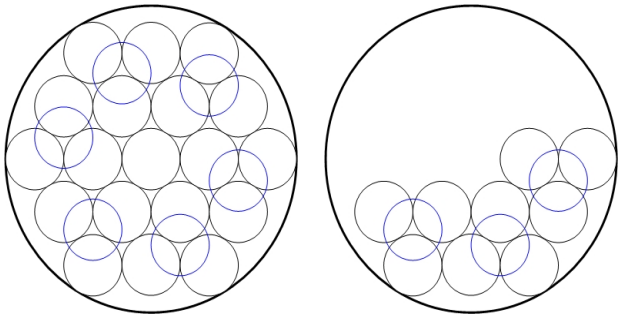


Figure 3. Pre-engineered placement of relay nodes

We propose to form the relay network of HWN* in two stages, which are Association and Route identification. At the former stage, each MT begins a searching process to associate itself with one or more RNs based on Signal-to-Interference Ratio (SIR). If the BS has spectrum available for one or more RNs, this information is broadcast over the cellular control channels so all nodes within the cell receive it simultaneously. The transmission radius of a node on the relay network is small compared to the cellular coverage. Thus, a RN only connected to a group of MTs and group members of a RN usually overlap with other RN's group members. To select a serving RN, reduce computation complexity and isolate groups, every MT with multiple membership broadcasts a neighbour advertisement (NADV) message with a Time-To-Live (TTL) value. The NADV message contains the identification of the source node and its received signal quality from different RNs. When a node receives NADV message back from multiple RNs, it compares various signal qualities. The RN with the best SIR is chosen as the serving RN and connections with other RNs are terminated.

In stage two, if required, MTs join the relay network by forming a path through serving RN using MHCT transmission mode. We consider several methods, which will be described in the next section to overcome high contention or overload at the serving RN, which may occur if several MTs attempt to join the relay network simultaneously. The MHCT uses a modified version of label routing as the main strategy to find the path from the MT via RNs to BS. When an intermediate RN forwards a route request (RREQ) message, it appends its identification and its distance from the BS to the message. Therefore, the serving RN can learn of the nodes involved in the relay path.

4. Adaptive Distributed Cross-layer Routing

The heterogeneous network media resources are shared by multiple traffic classes. Service classes that deliver QoS to applications have priority over others that do not. In such a network, routing decisions for high priority QoS sessions can affect what resources are available for lower priority traffic. Poor route selection can result in congestion for, or even starvation of, lower priority traffic. Many studies [14] have focused on routing algorithms that optimise the network throughput for individual service classes, minimise call blocking rate and improve per-session throughput. Little effort has been devoted to routing algorithms that address interclass resource sharing. But the fact is routing in a multi-class network is not as simple as selecting an “optimal” path for each flow, assuming that the traffic class of this flow is the only service class in the network. Such an “optimal” path can adversely affect the congestion condition of flows in other traffic classes. For example, in a network that supports both QoS traffic requiring bandwidth guarantees and best-effort traffic, which resources are available to best-effort traffic depends on how QoS flows are routed. QoS flows can consume all the bandwidth on certain links, thus creating congestion for, or even starvation of best effort sessions. Statically partitioning the link resources can result in low network throughput if the traffic mix changes over time. Thus, a mechanism that dynamically distributes link resources across traffic classes based on the current load conditions in each traffic class is critical for performance.

By proposing a cascaded Adaptive Distributed Cross-layer Routing (ADCR) for HWN*, we discourage applications from using any route that is heavily loaded with low priority traffic. Traditional routing strategies that use global state information are not considered. Problems associated with maintaining global state information and the staleness of such information are avoided by having individual MTs infer the network states based on route discovery statistics collected locally, and perform traffic routing using this localised view of the network QoS state. Each application, categorised by service class with the choice of three possible transmission modes, maintains a set of candidate paths to each possible destination and routes flows along these paths. The selection of the candidate paths is a key issue in localised routing and has a considerable impact on how the ADCR performs. The high priority traffic is given high priority in accessing comparatively expensive cellular resource, while low priority traffic tries to access

low cost ad hoc resource. Per MT bandwidth is used as the only metric for route local statistics collection since it is one of the most important metrics in QoS routing, furthermore, important metrics such as end-to-end delay, jitter can be expressed as a function of the bandwidth.

We first divide all traffic sessions into three simple service classes to fairly share spectrum between end user and provider equipment, further reduce the cost implementation, monitoring and management. The three service classes are:

- High profile users (HPUs)
- Normal profile users (NPUs)
- Low profile users (LPUs)

The HPUs have the highest access priority among the three communication modes in HWN* and traffic admission of NPUS and LPUs are considered based on current HPUs sessions. The NPUs are configured to have a higher probability than LPUs in terms of resource acquisition and this probability is decided by an Association Level (AL) set. In the case of network congestion, CT mode may temporarily become unavailable to NPUs when HPUs are not fully accommodated, while LPUs sessions may be only granted MHCT and OHAHT mode access to mitigate network congestion, reduce transmission delay and improve per MT throughput.

A MT locally links each application with a service class based on particular QoS requirement. The choices are flexible as one subscriber may link business voice calls with the HPUs class and Voice over IP (VOIP) calls to the LPUs class. Another subscriber may link all voice and data services to HPUs class when moving at high speed and then change them to the LPUs class when walking on the street. For the purpose of simulation evaluation of the HWN* applications are temporarily fixed to specific classes. For example, real time collaboration, wireless gaming, and geographic real time datacast applications are associated with the HPUs class. Interactive multimedia, media telephony and rich data applications are linked to the NPUs class. Lightweight browsing, LAN access and file exchange applications are classed as LPUs. As HPUs are liable to require more channels than NPUs and LPUs, applications such as large volumes of file exchanges are not suitable for the ad hoc structures.

Secondly, as radio resource management on fixed RNs directly affects the performance of the media access layer and network layer we adopt a policy based management approach and two policies are proposed to realise effective packet transmission. We define:

The RN has the right to reserve QoS guaranteed free channels for packet transmission and it maintains a status table that's refers to be other RNs and it provides information on change busy conditions or relay failure.

The purpose of bandwidth reservation is to let RNs that receive the relaying discovery command check if they can provide the bandwidth required for the connection. Meanwhile, to decrease queuing delay, high profile traffic should be given higher priority than low profile traffic. Thus, the routing packets and content packets for high profile applications have less waiting time in queues. The other way to raise priority is that the packets related to a high profile have shorter back-off time to increase the probability of early medium access. To more accurately guarantee queuing delay and to avoid having high profile traffic being influenced by other data traffic, we place a limitation on queues. For example, if the transmission time of LPUs exceeds the starting time of HPUs, the transmission of LPUs is suspended. As for the status table maintenance, information flooding is restricted to a limited scope. Once a positive acknowledgment message is confirmed by a requesting RN, the relay paths will not be changed unless resource contention happens. Given the fact that maintaining global RNs channel status in each RN slows down RN response time, we only require each RN update neighbouring RNs' information, periodically.

The transmission model defines the methods that nodes employ for communicating with one another. For ADCR based equipment, several commands are defined as:

1) ***ACK/ACCEPT/REJECT/REJHO*** for the message delivery acknowledgment, packet acceptance, packet rejection and after rejection handover request.
2) ***SEARCH/SETUP/DATA/BREAK*** for destination node finding, new connection establishment, packet delivery and connection teardown.
3) ***MOS*** for MT to choose adaptive transmission mode.
4) ***FAIL*** used to acknowledge any failure on RN or MT.
5) ***LREQ*** to request a label during the routing. The label is a short, fixed-length identifier. Multiple labels can identify a path or connection from the source MT to the destination MT. The structure of a label message contains flag, label, cost and TTL.
6) ***LREP*** to request a label replay during the label routing in MHCT model.

Time-sensitive multimedia applications have restrictions on end-to-end transmission delay, while FTP data transfers need a minimum guarantee on packet losses. Further actions such as channel transfer and re-routing are required before service termination. The routing algorithms in HWN* should allow subscribers seamlessly move without dropping their communications and considers differentiated QoS issues, for example, the QoS guarantee for HPUs require that they agree to pay more

than NPUs and LPUs.

The cascaded ADCR scheme includes three sub packet transmission models, which are the OHAHT for point-to-point ad hoc communications, the MHCT for cellular resource relaying using fixed RNs and the CT for traditional cellular service, as illustrated in Figure 4 We further define the HPU has having the highest priority to access any packet transmission models, in the order of CT, MHCT then OHAHT. However, due to the high priority of premium traffic, the global network behaviour as a consequence of this service class, including routing and scheduling of premium packets, may impose significant influences on traffic of other classes as we explained in a previous section. These negative influences, which could degrade the performance of low-priority classes with respect to some important metrics such as the packet loss probability and the packet delay, are often called the inter-class effects. To reduce the inter-class effects, the premium-class routing algorithm must be carefully selected, must work correctly under the hop-by-hop routing paradigm, and minimize the congestion resulting from the traffic of premium classes over the network. We also proposed mechanisms for load balancing of high-priority traffic on DiffServ networks. The Association Level (AL) calculation was proposed to differentiate between user classes. The AL is a set of parameters monitoring channel availabilities, an AL that scores higher than the threshold means that the channels are already occupied by ongoing sessions. Our extensive simulations clearly demonstrated that the proposed methods distribute the premium bandwidth requirements more efficiently over the whole network. We also present corresponding computerised process of MT's association with a serving BS and RN, and simplified ADCR algorithm with three transmission models in Figure 4.

4.1. One-Hop Ad-Hoc Transmission

In this model, the requesting MT first broadcasts ***SEARCH*** messages to every node in its transmission range including its associated RN and BS. For example, MT A in Figure 4 broadcasts ***SEARCH*** messages, if the destination MT B is within its transmission range and there is no ad hoc based media contention between MT A and MT B, MT B can respond to MT A with an ***ACK*** message. Once MT A confirms the acknowledgement, it starts a connection ***SETUP*** session immediately. MT A continues transmitting directly until the SIR falls to a certain level, where traffic re-routing or handover process will be initiated.

4.2. Multi-Hop Combined Transmission

The multi-hop combined transmission model involves RNs acting as intermediate nodes for message relaying. Figure 4 shows the connection setup process for communication between MT A and MT B via the RN infrastructure. MT A first broadcasts ***SEARCH*** messages to every node to find MT B. After the ***SEARCH*** session, MT A may find that the cellular resources can be

```
BS & RN Association()
{
For N MTs in a cellular cell, 0 < i < N;
SIRi = Received signal quality evaluation of MT i from the serving BS;
For N MTs in a cellular cell, 0 < i < N;
TTL = 1;
/*Receive neighbouring information from surrounding RNs*/
For N MTs, 0 < i < N , M RNs, 0 < j < M;
SIRij = Received signal quality evaluation of MT i from surrounding RNs;
Sort SIRij in descending order from high SIR to Low SIR;
Associated RN = the RN with Max(SIRij);
}
ADCR Routing()
{
DB(i) = Node i's distance from serving BS;
DR(i) = Node i's distance from serving RN;
/* Identify which traffic class the packet belongs to, HPU, NPU or LPU*/
Traffic_Class_Discovery ();
/* Individual packet routing with three sub models, OHAHT, MHCT and CT*/
One_hop_adhoc_transmission();
Multi_hop_combined_transmission();
Cellular_transmission();
Node i is scheduled to initiated a packet transmission at time T(k);
switch(service_class)
{
case HPU():
/* Evaluate QoS requirement and urgency based on weighted calculations */
Evaluation();
/*Check media access constraints for three transmission models*/
Media_check();
/*Try to use the transmission models in order of CT, MHCT then OHAHT*/
HPU_routing();
break;
case NPU():
Evaluation();
Media_check();
/*Try to use the transmission models in order of MHCT, CT then OHAHT*/
NPU_routing();
break;
case LPU():
Evaluation();
Media_check();
/*Try to use the transmission models in order of OHAHT, MHCT then CT*/
LPU_routing();
break;
}
```

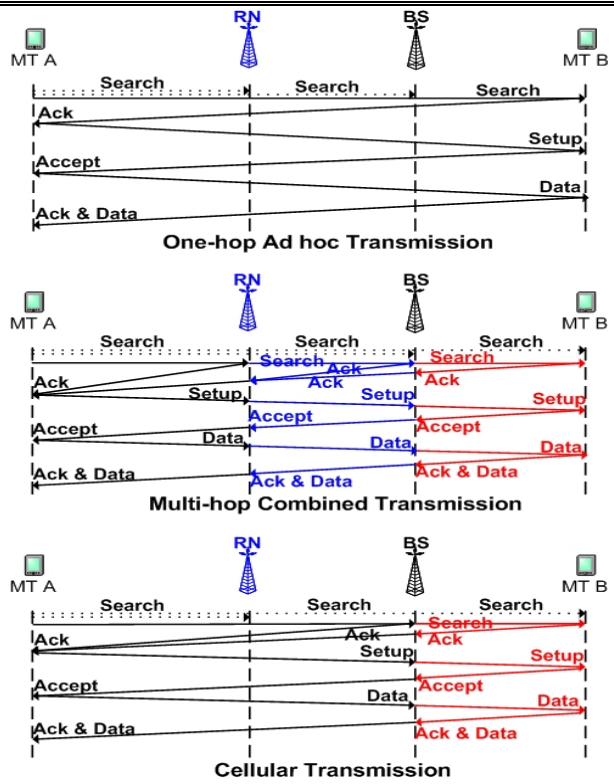


Figure 4. ADCR transmission

borrowed through RNs by receiving three ***ACK*** messages from the serving BS of MT B, RNs and the MT B. The positive acknowledgement requires MT B to send an ***ACK*** to its serving BS, then the serving BS sends an ***ACK*** to the RN infrastructure and finally the RNs feedback the ***ACK*** to MT A. Once the positive ***ACK*** is confirmed, the MT A starts a connection ***SETUP*** from MT A to RN, then RN to BS, and finally BS to MT B. The ***DATA*** transmission process follows the same packet delivery route, and further route discovery is prohibited to reduce the signaling overhead. MT A continues transmitting directly until the SIR falls to certain level, where traffic re-routing or a handover process will be to setup, configure, and maintain a path between two MTs. Media resources are largely saved when compared to other traditional routing algorithms, where paths are initiated.

Different from MANET nodes, which have limited energy and processor resources, protocols and algorithms running on the RNs can have a high level of efficiency with reasonable complexity. We introduce the label routing concept [15] in MHCT. The connectionless label routing is a distributed routing protocol that is able maintained regularly. The path from source node works as a tunnel identified by multiple labels, which requires source MT creates a ***LREQ*** message in which the packet contains IDs, sequence number and service class of the source MT. This packet also contains a broadcast ID and a hop count that is initialized to zero. All RNs that receive this message will increment the hop count. If a RN does not have any information about the destination node, it will record the neighbour's ID where the first copy of ***LREQ*** is from and send this ***LREQ*** to its neighbours. ***LREQs*** from the same node with the same broadcast ID will not be processed more than once. Figure 5 gives an example of label routing in MHCT. In this example, there are eight nodes with duplex connection link.

The MT A first creates a ***LREQ*** message and sends it out to its associated RN. Figure 5 illustrates the propagation of ***LREQ*** across the RNs and the reverse path at every RN. The reverse path entry is created for the transmission of the reserved label for this path. This label is embedded in the label reply message ***LREP***. The reserve path entry will be maintained long enough for the ***LREQ*** to traverse the path and for RNs to send a ***LREP*** to the source MT. Once a path is found in the relay structure, the source MT will check the sequence number (SEQ) of the destination MT in the current path in order

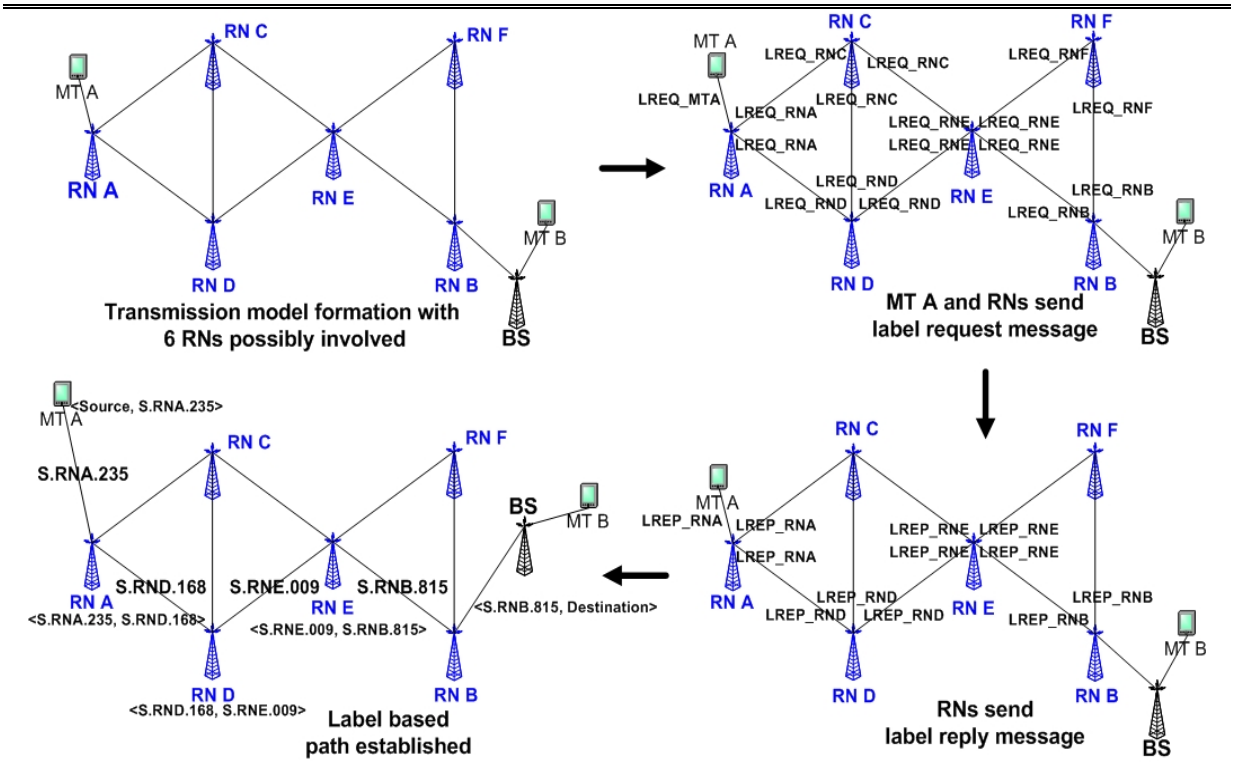


Figure 5. Label Routing illustration

information from both the media access layer and NETWORK layer. With label routing, intermediate RNs can provide fast and efficient forwarding without checking the Internet Protocol (IP) address and accessing a large routing table in the memory of the host RN.

In order to find a path to the destination MT, the to avoid old path information. It should be at least as great as the value entry in the ***LREQ***. Otherwise the existing path in the table will be discarded. If $SEQ \geq SEQ_{LREQ}$, it will also check in the current path whether the QoS requested by the source MT has been satisfied. If not, this request will be discarded. If the

source MT still can not find the destination MT B, MT A will increment the hop count in the ***LREQ*** by one and then broadcast it to its neighbors. Any duplicated ***LREQ*** with same source node ID and same broadcast ID will be discarded. Normally relay based label routing should have a maximum hop count. However, there is no energy constraints and node motilities issues in our relay infrastructure, thus theoretically any hop count threshold can be possible. We specify the hop court in ***LREQ*** as not being larger than 10 as a simulation limitation to avoid computation complexity and if the sender of a ***LREQ*** does not receive the reply message, each node only re-sends the ***LREQ*** once for each connection request.

The RN only creates a ***LREP*** with the total hop count of this path if hop count, sequence number, and path QoS are all acceptable, the new sequence number of the destination MT is the largest one between SEQ and SEQ_{LREQ}, the best QoS, and a label from its label pool. Then this ***LREP*** will be sent back to the source MT along the reverse path entry. The third plot in Figure 5 shows the propagation of the ***LREP*** along the reserve paths. Note that both RN C and RN F fail to send the ***LREP*** due to hop count, sequence number or QoS issues. The path between the source MT and the destination MT is composed of multiple segments and all data packets are relayed by these segments. Each segment is a real connection between two nodes and labeled by the sending-side node of the ***LREP*** in this segment. For example, in the path MT A to RN A to RN D to RN E to RN B to MT B showed in the last plot of Figure 5, RNs A, D, E and B set up the labels of the segments between A and D, D and E, and E and E respectively. MT A and RN A, MT B, RN B and its associated BS are the other two segments. Since the topology of the relay structure is meshed, the source MT can receive more than one ***LREP***. There is a hop count field in the ***LREP***. This field records the total number of hops of the path. The source MT will choose the smallest hop count from the ***LREPs*** in the specific limited time. All ***LREPs*** that are received after this time threshold will be ignored. And if some available ***LREPs*** have the same hop count, the path that has the largest destination sequence number, which means it is the latest path, will be chosen.

4.3. Cellular Transmission

The cellular transmission model applies when both the transceiver MT and the receiver MT are within the transmission range of a BS or BSs. The last plot in Figure 4 shows the connection setup process for communication between MT A and MT B via cellular BSs. MT A first broadcasts ***SEARCH*** messages to every node to find MT B. After the ***SEARCH*** session, the MT A finds that it is able to communicate with MT B directly via BSs, while the connection can be setup through a virtual wireless backbone. The positive acknowledgement of a connection requires MT B to send an ***ACK*** to its serving BS, then the serving BS informs the serving BS of MT A or the BS feedbacks the ***ACK*** to MT B when both MT A and MT B share the same serving BS. Once the positive ***ACK*** is confirmed, MT A starts connection ***SETUP*** from MT A to BS, then BS to BS, and finally BS to MT B. The ***DATA*** transmission process follows the same packet switched delivery route. MT A continues transmitting directly until the SIR falls to certain level, where traffic re-routing or an inter network handover process will be initiated.

Dynamic channel allocation in each base station can be managed in a distributed manner given that the channel usage does not break the two cellular network channel interference constrains [16] which are cosite constraint where there are minimum channel separations within a cell and non-cosite constraint where there is a minimum channel separation between two adjacent base stations.

5. Mobile Terminal Mobility Model

The HWN* system should be tested under realistic mobility conditions. To characterise the mobility pattern, we implement a HWN* Attractor Point Mobility Model (HPMM) based on the random waypoint mobility model and attractor point mobility model [17].

At the simulation start, a MT schedules an **ACK** message to itself before it determines a new two dimensional <x, y> position. After messages saving, the MT sends a ***MOVE*** message to the physical layer and reschedules the ***ACK*** to be delivered in $move^2$ seconds. The layout of a metropolitan environment consists of **n** probability based attractor points which MTs will move towards. We provide an approach which influences user mobility in a distributed manner based on attractor points, instead of grouping MTs with macro mobility, each MT selects a destination using micro mobility. We consider two ***Behaviours*** options when a MT approaches the simulation environment map borders. These are ***Rebounding Behaviour*** that makes the nodes reverse direction according to the elastic impact theory and ***Toroidal behaviour***, which makes the nodes leave the map. ***Speed Control*** is introduced on each MT and the speed continuously increases or decreases. For each step, a new MT speed sample $v(t_{k+1})$ is calculated:

$$v(t_{k+1}) = v(t_k) + a^*(t_{k+1})(t_{k+1} - t_k) \quad (4)$$

where $v(t_k)$ is the current speed, the Acc. speed a^* is a non-linear variable associated with the distance between a MT and its destination point and $\Delta t = (t_{k+1} - t_k)$ is the sampling period. Two limits are introduced, which are v_{upper} and v_{lower} and **Equation 4** applies to a MT only if $v_{lower} \le v(t_{k+1}) \le v_{upper}$, otherwise at next sampling

point, the speed will remain unchanged. The current speed $v(t_{k+1})$ is correlated with the previous speed value $v(t_k)$. A small sampling time leads to a smoother speed change.

6. Simulation and Results

The HWN* system and routing algorithm are evaluated by extensive discrete event computer simulations. Both the Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) are used. As more than 90 percent of web services consist of TCP flows it is reasonable to presume that traffic due to a MT web application will not deviate from this behavior if we only use TCP for web services in simulations. The H.263 codec is imported into the OMNET++ simulator with video file downloaded from the internet. We generalize and refer to all video streaming as real-time services, while all web transports are referred to as non real-time services. Table 3 shows the default QoS profile used consisting of 30% 64 kbps streaming video, 45% general voice calls and 25% non real-time web services according to the 3GPP [11] specification. The service request RN & BS registration, routing path discovery, transmission mode selection and data delivery. 100, 200, 300, 400 and 500 MTs are placed gradually in the system to study the impact of QoS support. The implementation of service classification will guarantee the delay of real-time packets.

The per cell capacity in the planned HWN* should be greater than that in random HWN* and traditional cellular network, especially when the number of MTs is larger and the cellular service percentage is lower. This is because MTs not serviced in a cell or outside of any cell coverage can use the carefully planned relaying path to access nodes with cellular coverage and communicate with other nodes far away. Therefore, when the number of nodes is large enough, the HWN* can achieve complete connectivity regardless of the cellular service percentage. But the capacity and topology connection of a network are very much influenced by infrastructure support and traffic volume. On the other hand, one hop ad hoc transmission is a part of three HWN* transmission modes. Therefore, the capacity in any HWN* and cellular system should be much larger than that in any pure ad hoc network, so we do not compare ADCR based novel infrastructure with MANET capacity.

Table 3. Characteristics of QoS differentiated users

	Low profile user	**Normal profile user**	**High profile user**
Portion of arrival request	20% Voice, 10% Web and 5% Video	15% Voice, 8% Web and 10% Video	10% Voice, 7% Web and 15% Video

Voice Dwell / Session time:	**Web Dwell / Session time:**	**Video Dwell / Session time:**
60s / 120s	120s / trace	120s/240s

portion is distributed and shared among these three user classes.

We randomly distribute the MTs in 13 regular hexagonal cells (1km length, $2.6\,km^2$) in an 8 km X 8 km grid. The BS is located in the centre of each cell, and each cell owns a RN located at a random position. From 300 to 1300 MTs are scattered in the grid to simulate varied scenarios. To ensure that the same cellular frequencies are repeatedly used in the cellular network, 7 frequencies are allocated to each cell and 128 channels are available on each BS. The MT travels from 0 to 80 km/hour since a relative speed higher than 160 KM/hour is not suitable for the 802.11 radio propagation model, which has limited compensation for channel fading. Typical simulation parameters are used [9] as illustrated in Section 2.

6.1. HWN* Capacity Analysis

Our first objective is to show the pre-engineered RN positioning strategy influence on the HWN* capacity under various network conditions. Operations of the proposed ADCR are simulated, including the process of

All the simulations run up to 1000 simulated seconds. The first 100 seconds is a simulation warm-up time thus traffic intensity remains unchanged. Figure 6 records per cell capacity performance of three scenarios when the cell's traffic load is increasing. The capacities of both the planned HWN* and the random HWN* go up till maximum throughputs reach around 5.6 Mbps and 4.7 Mbps respectively. As we can see from the trend of the capacity lines, when the traffic load becomes higher the pre-engineered HWN* outperforms the random HWN* in terms of network fairness and its maximum capacity gets near to the theoretical gain with a more uniform communications experience.

For packet delivery ratio in the HWN*, the System Throughput (ST) is defined as the delivery ratio:

$$ST = \frac{Total\_number\_of\_data\_received}{Total\_number\_of\_data\_sent} 100\% \qquad (5)$$

In this scenario, we only implement UDP traffic (without handshaking mechanism) on each MT instead of the default QoS traffic profile, and operations of the proposed ADCR are also simulated. The packets are sent

at constant bit rate (CBR) with a packet size of 1500 bytes but the MTs are added from 0 to 500 gradually as an input parameter to increase the offered load on the system. Figure 7 shows the impact of increased traffic on the packet delivery ratio. Results show that whichever traffic load is used, the proposed ADCR with pre-engineered RNs placement gives a higher throughput than the HWN* with random RN placement and cellular system. We can also see that the packet delivery ratio decreases when the UDP traffic load increases, this is mainly due to the congestion. However, the pre-engineered HWN* outperforms random HWN* or TDMA network by 12% and 26% respectively, when the maximum traffic load is achieved.

6.2. Packet Transmission Delay Analysis

The average packet transmission end-to-end delay of a traffic flow should be directly proportional to the number of hops traversed by the flow, and inversely proportional

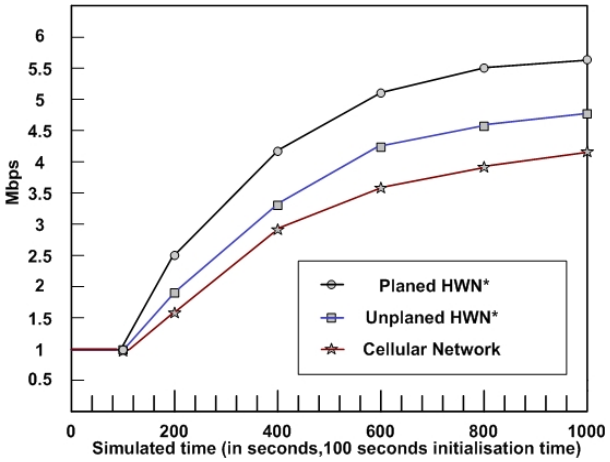


Figure 6. Average capacity per cell per second

to the flow's end-to-end throughput, this is an interesting metric to study since the HWN* system itself has complicated transmission models, which can be seen as hybrid traffic migration of both ad hoc and cellular network with RNs. We first define Average End-to-end Delay (AED) as:

$$AED = \frac{Total\_number\_of\_data\_received}{Total\_delivery\_time} \quad (6)$$

The cellular network is only packet-switched and the multi-hop ad hoc network routes traffic in unstable topology using various routing algorithms, thus the packet transmission delay should not be compared between HWN* and any single layer network. We therefore simulate simplified WINNER and SOPRANO hybrid network infrastructures with traffic routing functionality, respectively. For WINNER, through cooperative relaying algorithm, the RN operates full resource management functions like a cellular BS does. However, for distributed SOPRANO, the routing calculation is the sole responsibility of the local MT, thus, we implemented minimum energy routing protocol as recommended in [18]. Figure 8 shows the average end to end delay versus quantized load offered for three hybrid networks. We observe that there is a significant improvement in the delay performance of HWN* ADCR, compared to the cooperative relaying in WINNER and the minimum energy routing in the SOPRANO. In Figure 8, at 15 erlangs, the corresponding average end-to-end days are 0.10, 0.21, and 0.033. The delay in other systems is two or three times larger than that of ADCR, respectively. This shows that the proposed ADCR adaptively selects paths with better quality, and it can prevent to wasting transmission time.

6.3. QoS Based Routing Analysis

Experiments are also conducted to verify that the proposed ADCR algorithm meet the goal of providing

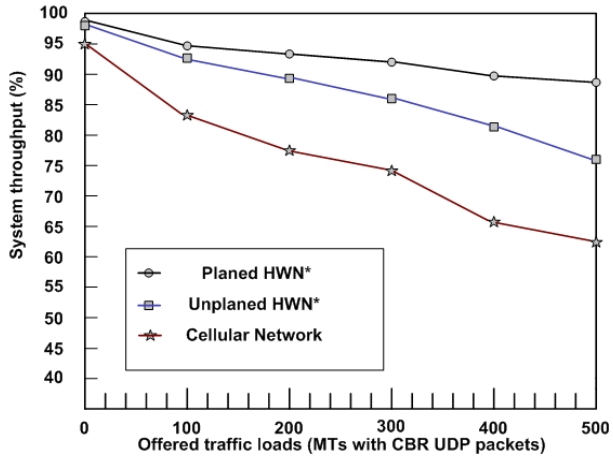


Figure 7. System throughput versus offered load

QoS differentiation among different users based on their class profile. To setup a comparison benchmark, we simulated a simple HWN* without any dedicated resource management and routing algorithms. In this network, each session has the same privileges when accessing the media resource. The arriving packets are accommodated on a first-come-first-serve basis until all available channels have been used. A MT terminates the routing process when it can not find an alternative route. Figure 9 shows the route acquire success ratio with increasing of the traffic load. The ADCR has the best success ratio since it always returns a path on-demand and performance improvement is marginal when the system is heavily loaded. The simple routing algorithm in the HWN* performs worse than the ADCR due to its limitation on route selection and lack of alternate paths.

Table 4 shows the individual class successful path acquiring probabilities against traffic loads offered for differenct user classes. It can be seen that different results are experienced by user applications in different service classes and for unclassified users in the simple HWN*. Under low and medium traffic intensities, the success rates are similar among HPUs, NPUs and LPUs, since

sufficient routes are available and LPUs are not largely affected by HPUs and NPUs communications. However, in the high traffic intensity case, HPUs and NPUs application encounter large resource competition in the MAC layer, which consumes a considerable fraction of the radio resource. This may adversely affect the route finding performance of LPUs, in particular when a HPU and NPU traffic hot spot occurs; LPUs are pushed to use the ad hoc communications modes, where the routing process are comparatively unstable compared against the infrastructure based modes.

7. Conclusion

Routing is an essential part for realizing HWN* environment. In this article we first present a review of current hybrid networks addressing the issue of system infrastructure, radio technology, media access methods, routing strategies and QoS based algorithm constraints, merits and deficiencies. We have devised a self-organising routing scheme, ADCR. In order to manage traffic in a combined ad hoc, cellular and relay system, ADCR employs three sub-transmission modes, respectively, to effectively reduce its resultant communication overhead to acquire cost-effective delay-constrained paths. ADCR also employs a service class based approach to discourage applications from using any route that is heavily loaded with low priority traffic. Simulation results demonstrate that the performance of ADCR meets our design goals. Moreover, this cascaded routing algorithm can also be used to support cost-effective routing with differentiated user classes, referred to as High profile users, Normal profile users and Low profiles users, which are based on different service requirements.

The routing in HWN* is challenging. Although some work has been carried out to address this critical issue for other types of hybrid networks, research in this area is far from adequate. The scalability issue is always critical in designing large networks. 1). The performance of an algorithm for selecting routes primarily depends on hybrid network layers structures, domain (cell) definition as well as proper information gathering method, since there exists a tradeoff between efficient data dissemination and communication overhead. 2).The route selection method should not only consider NETWORK layer constrains, but also constrains posed by the Media access layer and Physical layer, to meet individual application QoS requirements. For example, in a cross-layer co-design approach, if MAC layer issues are not

Table 4. Success route acquiring ratio comparison between different user classes

	HPUs	NPUs	LPUs	Simple HWN*
5 Erlangs/cell	100.0%	100.0%	98.0%	98.5%
10 Erlangs/cell	98.1%	93.2%	91.9%	86.2%
15 Erlangs/cell	97.3%	87.0%	86.7%	77.7%
20 Erlangs/cell	96.1%	84.4%	82.5%	57.5%
25 Erlangs/cell	95.0%	77.1%	72.1%	45.1%

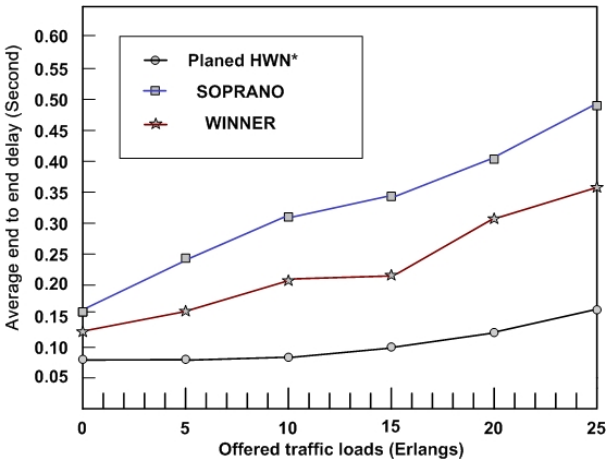


Figure 8. Average end to end delay versus offered load

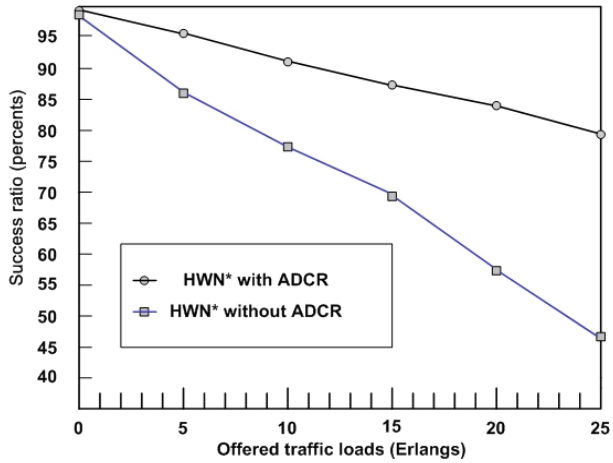


Figure 9. Comparisons of route acquire ratio vs. traffic

considered, then the route selection might ignore bandwidth requirements over time slot or frequency slot. This would pose a big problem for radio resource management. 3). Future work on relay positioning will involve using evolutional strategies or genetic algorithms to dynamically allocate the RNs.

8. References

[1] R. Beck and H. Panzer, "Strategies for Handover and Dynamical Channel Allocation in Micro-Cellular Mobile Radio Telephone Systems", IEEE Trans., on Vehicular Tech., vol.1, pp.178-185, 1989.

[2] A. Goldsmith and S. B. Wicker, "Design Challenges for Energy-constrained Ad hoc Wireless Networks", IEEE Wireless Comms., vol. 9, No. 4, Aug. 2002.

[3] C. Shen, D. Pesch and J. Irvine, "A Framework for Self-Management of Hybrid Wireless Networks Using Autonomic Computing Principles", In Proc. IEEE CNSR, Halifax, Canada, May, 2005.

[4] E. Setton and T. Yoo et. all, "Cross-Layer Design of Ad Hoc Networks for Real-Time Video Streaming," IEEE Wireless Comms. Magazine, pp. 59-67 Aug., 2005.

[5] WINNER, D4.3: Identification, definition and assessment of cooperation schemes between RANs, Final deliverable, IST-2003-50758, June 2005.

[6] C. Murthy and B. Manoj, Ad Hoc Wireless Networks: Architectures and Protocols, PRENTICE HALL, New Jersey, 2004.

[7] H.Y. Hsieh and R. Sivakumar, "Performance Comparison of Cellular and Multi-hop Wireless Networks: A Quantitative Study", in Proc., ACM SIGMETRICS, June, 2001.

[8] B.S. Manoj and K.J. Kumar, "On the Use of Multiple Hops in Next Generation Wireless Systems", Springer Science Wireless Networks, Dec., 2006.

[9] G. Stuber, Principlse of Mobile Communications, KAP, 1996.

[10] Omnet++ Discrete Event Simulation System, http://www.omnetpp.org

[11] 3GPP, UTRAN overall description, Technical Specification TS 25.401, http://www.3gpp.org

[12] A. Zemlianov and G. Veciana, "Capacity of Ad Hoc Wireless Networks With Infrastructure Support", IEEE Journal on Comms., Vol. 23, No. 3, Mar. 2005.

[13] H. Hu and K. Yanikomeroglu, "Performance Analysis of Cellular Networks with Digital Fixed Relays", Carleton University, 2006.

[14] B. Zhang and H. Mouftah, "QoS Routing for Wireless Ad Hoc Networks: Problems, Algorithms, and Protocols", IEEE Comms., Magazine, Oct., 2006.

[15] A. Acharya, A. Misra and S. Bansal, "A Label-switching Packet Forwarding Architecture for Multi-hop Wireless Lans", In proc. ACM. WoWMoM, Atlanta, Sep. 2002.

[16] C. Shen, D. Pesch and J. Irvine, "Distributed Dynamic Channel Allocation with Fuzzy Model Selection", ITT Conference, Limerick, Ireland, Oct. 2004.

[17] K. Murray, Admission and Handover Management fro Multi-Service Heterogeneous Wireless Access Networks, Ph.D Thesis, Cork Institute of Technology, Ireland, 2005.

[18] Ali N. Zadeh et all, "Self-Organizing Packet Radio Ad Hoc Networks with Overlay (SOPRANO)", IEEE Coms. Magazine, June, 2002.

I. J. Communications, Network and System Sciences. 2008; 1: 1-103
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

Influence of the Limited Retransmission on the Performance of WLANs Using Error-Prone Channel

Haider M. ALSABBAGH[1], Jianping CHEN, Youyun XU
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, P.R.China
E-mail: [1] haidermaw@gmail.com

Abstract

In WLANs, stations sharing a common wireless channel are governed by IEEE 802.11 protocol. Many conscious studies have been conducted to utilize this precious medium efficiently. However, most of these studies have been done either under assumption of idealistic channel condition or with unlimited retransmitting number. This paper is devoted to investigate influence of limited retransmissions and error level in the utilizing channel on the network throughput, probability of packet dropping and time to drop a packet. The results show that for networks using basic access mechanism the throughput is suppressed with increasing amount of errors in the transmitting channel over all the range of the retry limit. It is also quite sensitive to the size of the network. On the other side, the networks using four-way handshaking mechanism has a good immunity against the error over the available range of retry limits. Also the throughput is unchangeable with size of the network over the range of retransmission limits. However, the throughput does not change with retry limits when it exceeds the maximum number of the backoff stage in both DCF's mechanisms. In both mechanisms the probability of dropping a packet is a decreasing function with number of retransmissions and the time to drop a packet in the queue of a station is a strong function to the number of retry limit, size of the network, the utilizing medium access mechanism and amount of errors in the channel.

Keywords*:* IEEE802.11 DCF, WLAN, MAC Protocol, Throughput, Error-prone Channel.

1. Introduction

Next generation communications networks are being investigated thoroughly to satisfy its ultimate goal: accessible at any time and anywhere. One key point to satisfy such aim is to increase both data rate and processing speed [1][2]. The Wireless local area networks (WLANs) are able to comply with such demands and have become one of the fastest growing segments in the communications industry, especially with utilizing the new version of its protocol IEEE 802.11n [3]. The worldwide shipments of the WLAN equipment products reach $5.9 billion in 2006, and it is expected that WLAN equipment will continue to grow in 2007 to reach around $6.5 billion level as new IEEE 802.11n and VoWi-Fi equipment is introduced and the infrastructure for traditional Wi-Fi expands [3][4]. In 1997 IEEE's committee standardized 802.11 protocol for WLANs [5]. Since that time several versions of this protocol have been made. The physical media in the WLANs is shared between all stations and has limited connection range compared with its wired counterpart. The standard defines three PHY technologies and a unified MAC protocol to support 1 and 2 Mbps transmission over wireless media. The MAC protocol has two functions, namely distributed coordination function (DCF) and the optional point coordination function (PCF). DCF has superior attractiveness over PCF in many aspects [6], therefore this study is conducted to investigate WLANs utilizing DCF. DCF defines two mechanisms to access transmission medium: the basic access scheme, which is the default scheme and the request to send/clear to send (RTS/CTS) scheme, also known as four-way handshaking scheme [4][7][8]. Recently, considerable studies have been concentrated on modeling the IEEE 802.11 DCF medium access method. Bianchi in [9] modeled the idealistic assumption of collision only errors and packet retransmits are unlimited; a packet is being transmitted continuously until its successful reception. Wu in [10] extended Bianchi's analysis to include the finite packet retry limits as defined in the standard. Both studies used Markov chain model to analyze DCF operation and calculated the saturated throughput of 802.11 protocol. Periklais et al. [11] extended the work in [9] and [10] by taking into account both: transmission errors and packet retry limits for basic access of the IEEE802.11a protocol. X. Wang et al. [12] evaluate the impact of transmission error rate on the contention and the system throughput in

WLAN's protocol. However, [11] and [12] considered the probability of bit errors appearing on the transmission channel is the same in the two access mechanisms. Z. Tang et al. [13] presented an analytical model to evaluate the performance of the DCF in the case of bit errors appearing on the transmission channel and taking type of the used mechanism into account. However, Tang's study was under assuming of unlimited retransmitting. In this paper we extend Tang's work by taking into account influence of retransmissions and investigate its impacts on the performance of the WLANs. The results show that the throughput is insensitive to the number of the retransmissions when it exceeds the maximum number of backoff stages in both mechanisms. This insensitive trend does not change with amount of BER in the utilizing channel. However, adopting four-way handshaking mechanism show that the throughput is more immunes than its counterpart mechanism when WLAN's channel suffers from much error. This paper is organized as follows: Section 2 devoted to explain the used model in this study. Explanations for the achieved results are given in Section 3 and then our conclusions are drawn to Section 4.

2. The Model

The analysis employs the Markov chain model in [8] and [10] and makes use of the same assumptions as in [13]: all stations always have a packet available for transmitting (saturation case) into an error prone-channel.

2.1. A Brief Description of the Backoff Process in IEEE 802.11 DCF

When stations sense that the medium is idle for a period, more than DCF, the backoff timer value for each station is uniformly chosen within the interval $[0,W_i - 1]$, where W_i is the current contention window (*cw*) size and *i* is the backoff stage $i \in [0,m]$ and *m* represents the station's retry limit. The backoff counter for every station depends on the collision and on the successful packet transmissions experienced by the station in the past. At the beginning: $W = W_o$ and after each retransmitting due to a packet collision or error, W_i is doubled up to a maximum value, $W_{m'} = 2^{m'} W$, where *m*' is the maximum number of the backoff stages. When W_i reaches $W_{m'}$, it will stay at this value until it is reset to W_o again either after the successful data the counter reaches to its limit.

2.2. The analytical modeling[1]

As in [10], the probability of a station to transmit a packet in a randomly chosen slot time is:

$$\tau = \frac{1-p^{m+1}}{(1-p)}.\Psi_{0,0} \tag{1}$$

where:

$$\Psi_{0,0} = \begin{cases} \dfrac{2(1-2p).(1-p)}{W(1-(2p)^{m+1}.(1-p)+(1-2p).(1-p^{m+1}))} & m \le m' \\ \dfrac{2(1-2p).(1-p)}{W(1-(2p)^{m'+1}).(1-p)+(1-2p).(1-p^{m+1})+W.2^{m'}\,p^{m'+1}.(1-2p).(1-p^{m-m'})} & m > m' \end{cases}$$

τ does not depend on the type of the mechanism adapted by a station: basic access or four-way handshaking, *P* is the unsuccessful probability when a transmitted packet encounters a collision with at least one of the $n-1$ remaining stations in a time slot. So:

$$p = 1-(1-\tau)^{n-1}(1-p_c) \tag{2}$$

Influence of errors in the transmitting channel is included through the parameter P_C as [13]: In the case of basic access mechanism:

$$p_C = 1-(1-BER)^L \tag{3-a}$$

where $L = PHY_h + MAC_h + P + ACK$. In the case of four-way handshaking:

$$P_c = 1-(1-BER)^{RTS} * (1-BER)^{CTS} * (1-BER)^{P} * (1-BER)^{ACK} \tag{3-b}$$

When a station transmits and the remaining ***n-1*** stations defer their transmissions, the packet would be arriving successfully with probability P_S. Considering the probability that a random slot is empty $(1-P_{tr})$, probability of successful transmission is $P_{tr}P_s$ and probability of the collision is $P_{tr}(1-P_S)$, the average length of a slot time is:

$$E[slot] = (1-p_{tr}).\sigma + p_{tr}.p_s.T_s + p_{tr}.(1-p_s).T_c \tag{4}$$

Consequently, the system throughput,, can be expressed by dividing the successfully transmitted payload data over a slot time. The probability of dropping a packet when the retry limit is reached is known as the packet drop probability and given as:

$$p_{drop} = p^{m+1} \tag{5}$$

A packet is dropped when it reaches the last backoff stage and experiences another collision or an error.

3. Performance Analysis

3.1. Influence of the Retry Limits on the Network Characteristics

This analysis is based on the model in [13] and we have included influence of limited number of retransmissions. To validate our evaluation and highlight influence of limited retries we compare our results, *after* taking influence of limited retries, with results that have been gotten by using the model in [12] and with that using model given in Ref. [13]. The results are illustrated in Figure 1. The same parameters values in [13] have

[1]Definition and values of all the rest parameters that do not mentioned here are as in Ref. [13].

been used in this study in order to facilitate the comparison purpose. In Figure 1 our results denoted as (a), results of Ref. [13] denoted as (b) and the results that obtained by using the model in [12] denoted as (c). The system throughput has been estimated for three different network sizes: 5 (small), 20 (middle), and 50 (large). Figure 1 shows that: for networks utilizing the basic access, results (a) are much closer to results (b) in the small and middle network sizes and much close to results (c) for the large networks over all the range of BERs. This can be justified since on the small and middle networks the rate of collisions is relatively small compared with that in large networks, which indicates that influence of retransmission at small and middle networks is also small. For networks utilizing the four-way handshaking, all results: (a), (b) and (c) show that the throughputs are insensitive to size of the networks over all the examined range of the BERs. While almost all the results are mostly closed when BERs less than 10^{-5}, there is a difference between the results (a) and (b) in one side and results (c) on the other side when $10^{-5} < BER < 10^{-4}$. This difference is due to the effect of the parameter $\boldsymbol{P_C}$ on $\boldsymbol{P}$, which could be much sensible at higher level of BERs. Considering errors in the channel, $\boldsymbol{P_C}$ in the Ref. [12] did not distinguish the used access mechanism as it did in [13] and in ours as shown in Eq. 3.

Figure 2 illustrates variation of the network throughput with number of retransmissions at low level of BERs = 10^{-5} (Figure 2a) and at relatively high level of BERs 10^{-4} (Figure 2b). It is obvious that trend of the network throughput is almost unchangeable with level of errors in the channel when a network uses the four-way handshaking scheme. In the contrary side, when the networks use basic access, the throughput level is obviously decreased with increasing errors in the channel over all the retransmission range. Figure 3 shows that probability to drop a packet is exponentially decreasing with number of retransmissions and has undistinguishable differences between the two access mechanisms. However, this probability is increasing function of errors in the channel. Probability of dropping a packet is ignorable when the network exceeds the maximum number of backoff stage[2] with low level of BER (see Figure 3 a). This leads to enforce packets stay at queue of the transmitter for a long time and cause much delay, which could be unacceptable for some delay-sensitive applications. Furthermore, Figure 4 a and b show that the average time to drop a packet when adapting four-way handshaking is lower than that when using basic access at low level of BER and it becomes more obvious at high level of BER (as in Figure 4-b). These are beneficial of using RTS/CTS packets by reserving the channel in advance before starting to transmit a long data packet and hence supporting reduction the rate of collisions.

3.2. Network Characteristics with Specific Number of Retransmissions

In this section, the model that developed in the section 2 will be used to investigate some characteristics of WLANs. IEEE 802.11b is considered with retry limits = 5. Figure 5 shows packet delay with channel capacity C= 1, 5.5 and 11 Mbps against the number of contending stations. Results indicate that packet delay increases linearly with increasing number of stations due to increase rate of collisions and hence increasing number of retry, which also causes increase of packet drop probability as clarified in Figure 6. However, Figure 7 shows that as transmission rate increases the packet drop time decreases remarkably. This is due to the time spent for packet transmission decreases as the data rate increases [11]. For large network sizes, the probability of collision increases and hence number of retransmitting increases for each sharing station. As a result, a station to get a chance to establish the new transmitting at large networks is lower than that in the small networks. Thus, the time to drop a packet in small networks is much suppressed, compared with that in large networks. For instance, for an access point in WLAN covering an area of 50 contending stations, the drop time is 7.57 s at 1Mbps bit rate and is 1.12 s at 11 Mbps bit rate. Thus, from Figure 5, the corresponding packet delays are 573 ms and 85 ms, respectively. Figures 8 and 9 indicate that using long packet payload leads to large packet delay and packet drop times. This is due to the collision time, $\boldsymbol{T_C}$ which is proportional to the packet size as: $T_C = L + DIFS + \delta$, where L= (PHY header + MAC header) + packet payload, and δ is the propagation delay. The collision time is defined as the average time the channel is sensed busy by the stations during a collision. On the other side, the delay of a successful packet depends on the average number of slots required for a packet to experience $m+1$ collisions in the (0,1,……,m) stages and also depends on average slots time, which is proportional to the number of contending stations, and channel bit rates. Results in figures 8 and 9 investigate three different WLAN sizes: 5, 40, and 70 with three different channel bit rates: 1, 5.5 and 11 Mbps, the packet delay and packet drop time is linearly increasing with size of payload.

4. Conclusion

In this paper, an analytical model is developed to include influence of limited number of retransmissions on the main characteristics of WLANs that use error-prone channels. The model has been applied on the two available mechanisms in the DCF functions. We validate our evaluations by comparing our results with the results in two considered references. This study shows that: the networks using four-way handshaking mechanism have a

[2] Extensive calculations for different values have been done, and all the obtained results were valid and assist the result stated in this paper: Probability of dropping a packet is ignorable when the network exceeds the maximum number of backoff stage.

good immunity against the increasing error in the transmitting channel over the range of retransmissions, also WLAN's throughput is unchangeable with size of the network over the range. On the other side, for the networks using basic access mechanism, the throughput is suppressed with increasing amount of errors in the transmission channel over all the available range of the retry limits as well as it is sensitive to the size of the network. Further, the throughput does not change with retry limits when it exceeds the maximum number of backoff stages in both DCF's mechanisms. In the both mechanisms probability of dropping a packet is decreasing function with number of retransmissions and the time to drop a packet in the queue of a station is strong function to the number of retry limits, size of the network, the utilized mechanism and the amount of errors in the transmitting channel. Also, the study investigates characteristics of the IEEE 802.11b with a specific number of retry limits. The achieved results indicate that packet delay increases linearly with increasing number of stations due to increasing rate of collisions and hence increasing number of retry, which also causes increase of packet drop probability. Also, the network performance is strongly dependent on channel bit rate and the used packet length with adapting specific number of retransmission.

5. References

[1] G.-i. Ohta, F. Kamada, N. Teramura, and H. Hojo, "5GHz W-LAN verification for public mobile applications -internet newspaper on train and advanced ambulance car," First IEEE Consumer Communications and Networking Conference. CCNC 2004, 2004, pp. 569-575.

[2] T. B. Zahariadis, K. G. Vaxevanakis, C. P. Tsantilas, and N. A. Zervos, "Global Roaming in Next-Generation Networks," IEEE Communications Magazine, February 2002, pp. 145-151.

[3] www.bbwexchange.com/publications/page1421-4666.asp

[4] M. Etoh, "Next Generation Mobile Systems 3G and Beyond," John Wiley & Sons, Ltd, 2005.

[5] "IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications ,ISO/IEC," 1999, pp. 8802-11.

[6] H. Zhu and I. Chlamtac, "An Analytical Model for IEEE 802.11e EDCF Differential Services," The 12th international conference on Computer comunications and networks, ICCCN 2003, Preceedings, 2003, pp. 163-168.

[7] Y. F. Y. Kwon, and H. Latchman, "Design of MAC Protocols with Fast Collision Resolution for Wireless Local Area Networks," IEEE Transactions on wireless communications, vol. 3, MAY 2004, pp. 793-807.

[8] A. C. B. P. Chatzimisios, and V. Vitsas, "Performance analysis of the IEEE 802.11 MAC protocol for wireless LANs," Int. J. Commun.Syst., vol. 2, 2005, pp. 545-569.

[9] G.Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," IEEE Journal on Selected Area in Communications, vol. 18, 2000, pp. 535-547.

[10] H. Wu, Y. Peng, K. Long, and J. Ma, "Performance of Reliable Transport Protocol over IEEE 802.11 Wireless LAN: Analysis and Enhancement," Proc. of IEEE INFOCOM, vol. 2, 2002, pp. 599-607.

[11] P. Chatzimisios, A. C. Boucouvalas, and V. Vitsas, "Performance Analysis of IEEE 802.11 DCF in Presence of Transmission Errors," IEEE International Conference on Communications, 2004, pp. 2854-2858.

[12] X. Wang, J. Yin, and D. P. Agrawal, "Impact of channel conditions on the throughput optimization in 802.11 DCF," Wirel. Commun. Mob. Comput., vol. 5, 2005, pp. 113-122.

[13] Z. Tang, Z. Yang, J. He, and YanweiLiu, "Impact of bit error on the performance of DCF for wireless LAN," Communications, Circuits and Systems and West Sino Expositions, IEEE 2002 International Conference, 2002, pp. 529-533.

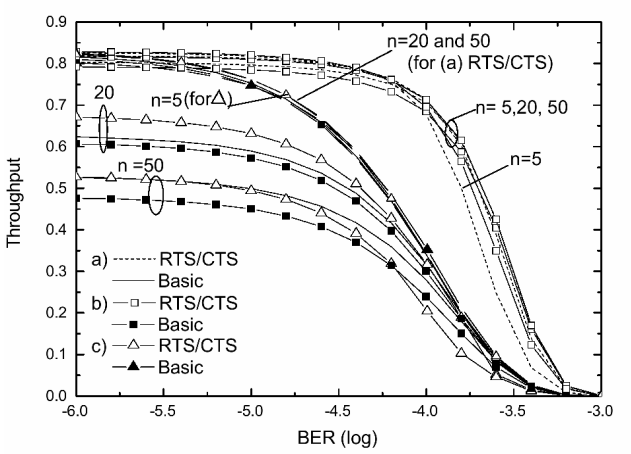
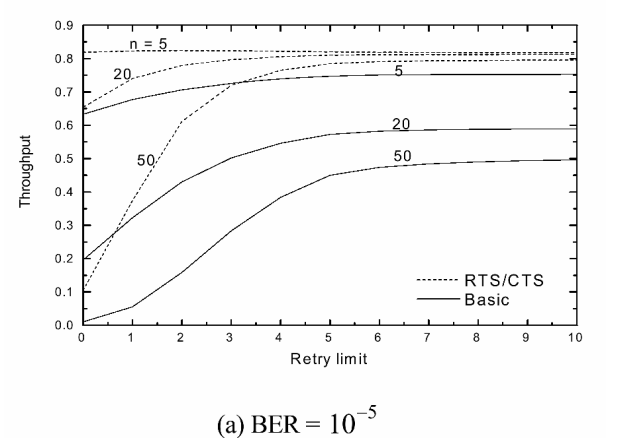
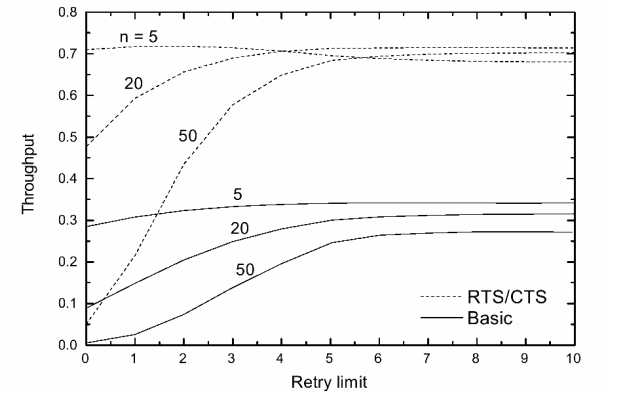


Figure 1. Throughput comparison for different network size (n=5, 20 and 50): (a) our results, (b) results from Ref. [13], and (c) results obtained using the model in Ref.[12]

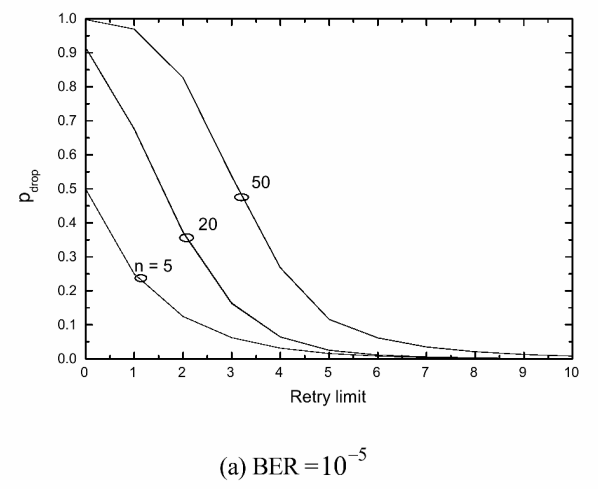


(a) BER = 10^{-5}

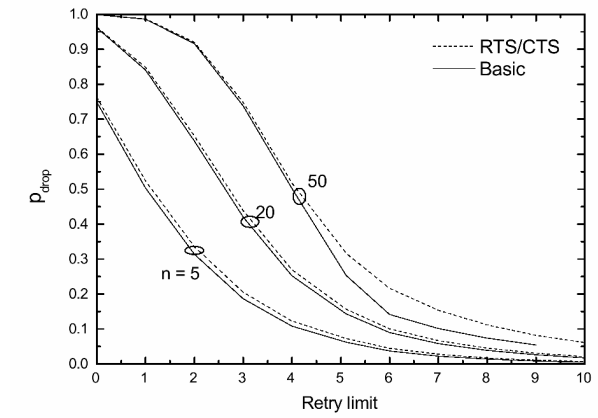


(b) BER = 10^{-4}

Figure 2. Throughput efficiency of the two mechanisms as a function of retry limit for different network sizes

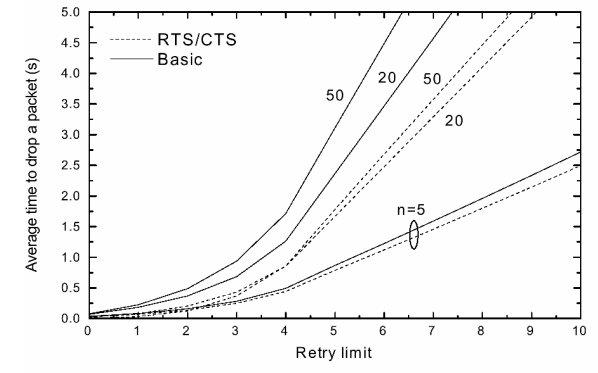


(a) BER = 10^{-5}

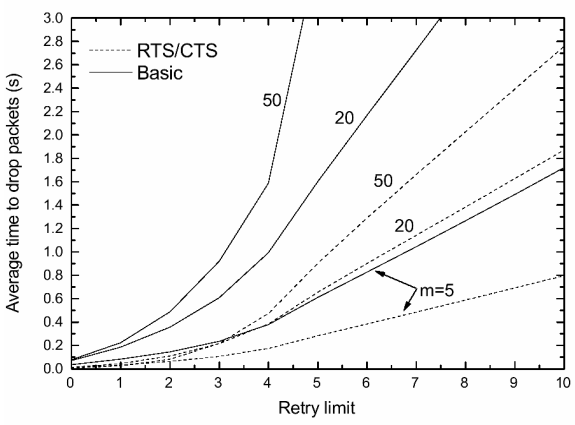


(b) BER = 10^{-4}

Figure 3. Impact of retry limit on the drop probability for different network sizes



(a) BER = 10^{-5}



(b) BER = 10^{-4}

Figure 4. Influence of retry limit on the time to drop a packet for different network sizes

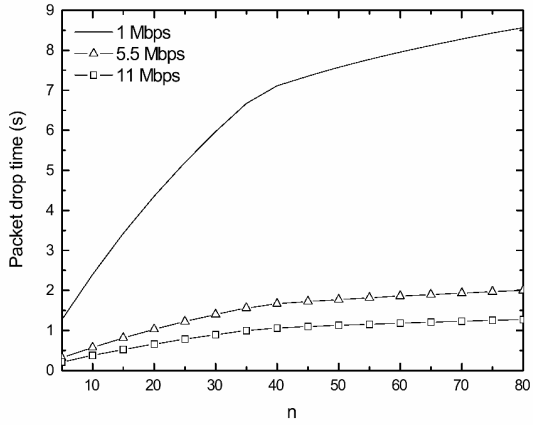


Figure 7. Packet drop time with different channel bit rates versus number of contending stations.

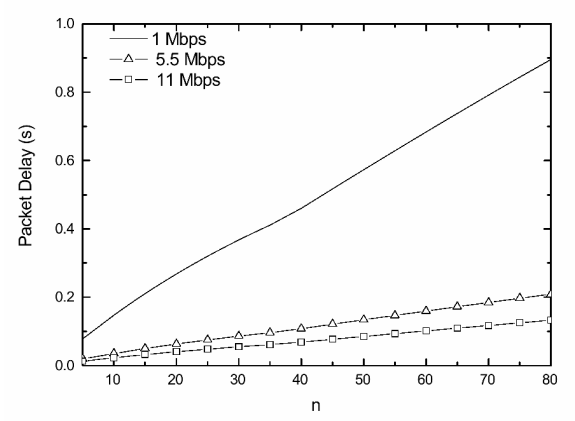


Figure 5. Packet delay with different channel capacities as a function of number of contending stations.

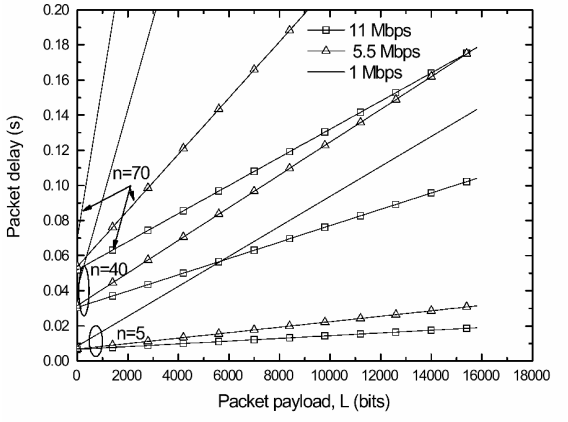


Figure 8. Packet delay against packet payload size for different contending stations and channel bit rates.

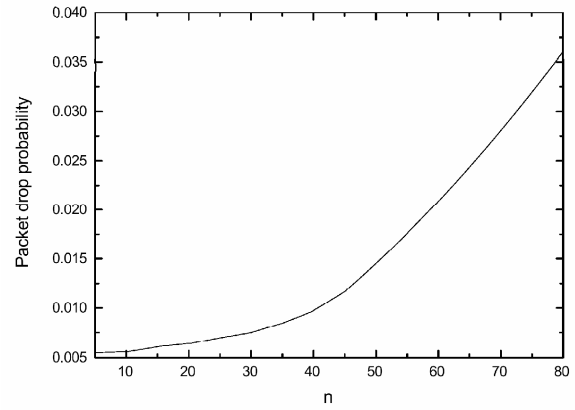


Figure 6. Packet drop probability against number of contending stations.

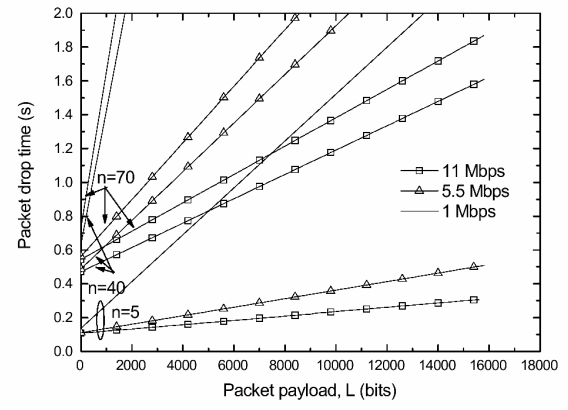


Figure 9. Packet drop time against packet payload length for different WLANs sizes and channel bit rats.

I. J. Communications, Network and System Sciences. 2008; 1: 1-103
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

Exploiting Geometric Advantages of Cooperative Communications for Energy Efficient Wireless Sensor Networks

Irfan AHMED[1], Mugen PENG[2], Wenbo WANG[2]
Wireless Signal Processing and Network Lab
Beijing University of Posts and Telecommunications, Beijing, P.R.China
E-mail: [1]irfanahmed44@gmail.com, [2]{pmg, wbwang}@bupt.edu.cn

Abstract

Energy efficiency and enhanced backbone capacity is obtained by exploiting the geometric orientation of cooperative nodes in wireless sensor network. The cooperative communication in wireless sensor networks (WSN) gives us leverage to get the inherent advantages of its random node's locations and the direction of the data flow. Depending on the channel conditions and the transmission distance, the number of cooperative nodes is selected, that participate in an energy efficient transmission/reception. Simulation results show that increasing the cooperative receive diversity, decreases the energy consumption per bit in cooperative communications. It has also been shown that the network backbone capacity can be increased by controlled displacement of antennas at base station at the expense of energy per bit.

Keywords: Cooperative Communication, Energy Efficiency, Capacity, Receive Diversity Sensor Network.

1. Introduction

Wireless sensor network consists of hundred or thousand of small, inexpensive wireless nodes responsible for monitoring a physical activity and reporting to the base station (BS), where end user can access the reported data. Cooperative communication has gain an obdurate place in wireless networks. It has been shown [1] that multiple-input multiple-output (MIMO) systems require less transmission energy than the single-input single-output systems. It's still intricate to build multiple antennas on low-cost, small-sized sensor nodes; therefore MIMO techniques cannot be applied directly on wireless sensor networks. However it is possible to implement MIMO techniques in WSN without physically having multiple antennas at the sensor nodes via cooperative communication techniques [2][3], such distributed MIMO techniques can offer considerable energy saving even after allowing some extra circuit power, communication and training overheads.

The use of cooperative communications in wireless sensor networks allows for energy savings through spatial diversity gains [2]. Traditional MIMO utilizes fixed antenna arrays in transmitter/receiver. These arrays are of definite geometric shapes, e.g., one dimensional array, circular array, rectangular array etc, but in cooperative MIMO communications there is no pre-defined array of antennas, the cooperative nodes cooperate with each other at run time to send/receive data.

There is already a lot of works related to cluster formation [4-7], capacity improvement [8] and energy efficient cooperative communication [2][3][9] in WSN, but little attention has been given to cooperative nodes selection on the basis of their location in clusters to improve the energy efficiency and reliability.

Most of the recent research work in wireless sensor networks, modeled the wireless channel with rich scattered or Rayleigh fading channel model [2][3][10], which is suitable channel model for wireless communications in urban areas where dense and large man made buildings act as rich scatterers. Sensor networks are usually deployed in the areas far away from human population, e.g., in plain desert areas for surveillance, on volcano hills for early alerts and near sea-shore for storm alerts etc. In all these situations, Ricean fading channel model works well, because it contains both non line-of-sight (NLOS) and line-of-sight (LOS) components.

In this paper, we have divided the WSN into two functional model parts, i) whole sensor network except the backbone link is modeled with Ricean fading channel and ii) the backbone link (the link between base station and the

cooperative nodes near the base station) is modeled by pure deterministic channel.

Remainder of this paper is organized as follows. System model is presented in section 2. Section 3 exploits the inherent advantages of wireless sensor network. Performance analysis in terms of energy efficiency and capacity of sensor network is presented in section 4. We show simulation results in Section 5, and conclude in Section 6.

2. System Model

The system model shown in figure 1, it consists of a cluster based sensor network. Sensor nodes are grouped for cooperative communications. The selection of group nodes is described in next section. In typical wireless sensor network scenarios as narrated above, large number of nodes are randomly deployed in an unattended area. Due to the presence of scatterers as well as line-of-sight component in those areas, Ricean fading channel model has been used to modeled the WSN, Ricean fading channel has both LOS and NLOS components [1]

$$\mathbf{H}=\sqrt{K/(1+K)}\overline{\mathbf{H}}+\sqrt{1/(1+K)}\mathbf{H}_w \qquad (1)$$

where $\sqrt{K/(1+K)}\overline{\mathbf{H}}$ is LOS or deterministic component of the channel, $\sqrt{1/(1+K)}\mathbf{H}_w$ is NLOS or scattered component

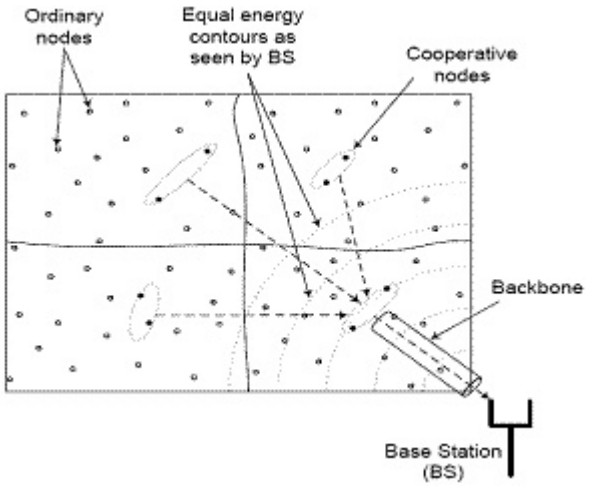


Figure 1. System Model

of the channel and K is the Ricean factor, it is the ratio of the LOS power to the scattered power. $K=0$ in the presence of rich scattered or pure Rayleigh fading, and $K\to\infty$ in case of non-fading channel.

We restrict our analysis to the case of frequency-flat fading. The elements of $\mathbf{H}_w$ are zero-mean circularly symmetric complex Gaussian (ZMCSCG) random variable with unit variance [1]. We have assumed binary phase shift keying (BPSK) modulation with Alamouti scheme [11] for cooperative MIMO communications. We have furthermore assumed that the channel is unknown to the transmit nodes and is perfectly known to the nodes at receive side.

3. Advantages of Cooperative Communications in WSN

3.1. Exploiting the WSN Architecture

1) WSN is deployed with hundred or thousand of structural or randomly placed nodes making it ideal for cooperative communications. Selections of nodes in a cooperative group which are approximately at same distance from the intended receive nodes, results in equal energy consumption per bit in cooperative communications.

2) By selecting closest nodes within a cooperative group, we can decrease energy consumption per bit in intra-cooperative node communications.

3) In WSN the data flow direction is from sensor network to BS i.e., most of the time BS act as a receiver (neglecting small amount of signaling data from BS to network). By increasing the number of antennas at BS, we can take advantages of receive diversity at BS, because BS has no energy constraint.

4) Usually the base station is located at some height in order to get reliable communication with the network, which results in a dominant LOS component. By exploiting this LOS communications we try to get maximum capacity in this backbone link.

3.2. Selection of Cooperative Nodes in a Network

Cooperative communications in WSN can be more beneficial if transmit cooperative nodes are at equal distance from the intended receive nodes. This strategy evenly distributes the transmission energy in LOS component because in LOS component the power loss is inversely proportional to the square of the distance between the transmitter and receiver. In addition to this the cooperative nodes should be at minimum distance from each other to save energy in local communication within the cooperative nodes group. These measures reduce the required bit energy for desired BER. How these measures can be taken?

3.2.1. Algorithm

After the cluster formation [5] and selection of cluster-heads in each cluster, following steps ensure the energy efficient selection of cooperative nodes:

Step I: BS broadcasts RSSI (Receive Signal Strength Indication) beacon signal to neighbor cluster nodes.

Step II: Each node $\chi(i)$ measures receive signal strength, calculate $d_{\chi}(i)$ and send back an ACK beacon signal to BS, where $d_{\chi}(i)$ is the distance between node $\chi(i)$ and BS.

Step III: Upon receiving the ACK beacons from all nodes, BS calculates their distances from itself $d_{BS}(i)$. Final estimate $d_{\chi\_BS}(i)$ is obtained by averaging $d_{\chi}(i)$ and $d_{BS}(i)$.

Please note that $d(i)$ is not an actual physical distance, it is equal-power-distance.

Step IV: Based on these estimates BS allocates an equal-power group identity to those nodes which lies within a certain range $\min\left(d_{\chi\_BS}(i)\right)+\aleph\Delta\Xi$, where $\aleph$ is an integer and $\Delta\Xi$ is small distance depends upon $\min\left(d_{\chi\_BS}(i)\right)$ and $\max\left(d_{\chi\_BS}(i)\right)$.

Step V: Each node in equal-power group $\square_P$ multicasts a RSSI beacon signal within its group. It will know its distance from other nodes in the group with the help of ACK receives.

Finally the cooperative communications group $\square_C, \square_C \subseteq \square_P$. Nodes selection preference order in $\square_C$ is

Proximity > Residual node energy

Node within a cooperative communication group with highest residual energy becomes the cooperative-group-head, it responsible for data aggregation and the communication within the cooperative group.

In the same way cooperative-group-head plays the role of BS for neighbor clusters and follows from *Step I* to *V*.

4. Performance Analysis

4.1. Network Performance in Ricean Fading Channel

We have modeled the sensor network with Ricean fading channel model. In the particular arrangement of cooperative nodes shown in figure 1, where all nodes (antennas) are at same distance from the intended receivers, we use following equation [2] for total energy required per bit per hop,

$$E_{bit} = (1+\alpha)\rho N_0 \frac{(4\pi d)^2}{G_t G_r \lambda^2} M_l N_f + \frac{P_{tx\_elect}}{R_b} n_T + \frac{P_{rx\_elect}}{R_b} n_R \quad (2)$$

where α is depends upon the drain efficiency of power amplifier in transmitter circuitry, ρ is the signal power (from n_T transmit nodes) to noise ratio at each of the receive node, N_0 is AWGN power spectral density, d is the distance between transmitter and receiver, λ is the carrier wavelength, G_t, G_r are antenna gains of transmitter and receiver respectively, M_l is the link margin compensating hardware process variations and other noise or interference, N_f is the receiver noise figure, R_b is transmission rate and P_{tx_elect}, P_{rx_elect} are power dissipated in transmitter and receiver circuits, respectively. The system parameters related to above equation are taken from [2, Table I], and the power consumption values of transmitter and receiver circuits (P_{tx_elect} =38mW, P_{rx_elect} =41mW) are of TelosB mote[1] [12]. Expression for BER as a function of SNR in Ricean fading channel is given as [1]

$$\overline{P}_b \le \left(\frac{1+K}{1+K+\rho d_{\min}^2/4n_T}\right)^{n_T \times n_R} e^{-n_T \times n_R \left(\frac{\frac{\rho d_{\min}^2}{4n_T} K \|\mathbf{H}\|_F^2}{1+K+\frac{\rho d_{\min}^2}{4n_T}}\right)} \quad (3)$$

where $d_{\min}$ is the minimum distance of separation of underlying scalar constellation and $\|\mathbf{H}\|_F$ is the Frobenius norm of channel matrix $\mathbf{H}$. Using upper bound in above expression, we obtain table I, which has been used for evaluation of energy per bit in equation (2).

4.1.1. Discussion

From equation (2) and (3), we can see that the transmission energy per bit for a desired BER is functions of the number of transmit/receive nodes, underlying constellation size and the channel condition. Table I shows the values of ρ for different combinations of transmit and receive nodes at $\overline{P}_b = 10^{-3}$. It can be seen that required SNR at receiver nodes decreases with increasing number of cooperative nodes for a fixed BER.

4.2. Backbone Link Performance in Deterministic or LOS Channel

4.2.1. Energy Consumption

The backbone link is modeled by LOS channel, for which the total energy required per bit is given by equation (2). The SNR (ρ) per receive node in equation (2) is given as [1, equation 5.60]

$$\overline{P}_b = \exp\left(-4\rho \|\mathbf{H}\|_F^2 / n_T\right) \quad (4)$$

then the equation (2) becomes

[1] With very low power TelosB mote and power loss exponent of two, we can neglect the inter-node cooperation overhead [14].

$$E_{bit} = (1+\alpha)\frac{n_T \ln\left|1/\overline{P}_b\right| N_0}{4\left\|\mathrm{H}\right\|_F^2}\frac{(4\pi d)^2}{G_t G_r \lambda^2} M_l N_f + \frac{P_{tx\_elect}}{R_b} n_T + \frac{P_{rx\_elect}}{R_b} n_R \quad (5)$$

4.2.2. Capacity

The instantaneous capacity (in bps/Hz) for n_T transmit and n_R receive cooperative nodes under an average transmit power constraint is given by [1]

$$C = \log_2\left[\det\left(\mathbf{I}_{n_R} + \frac{\rho}{n_T}\mathbf{H}\mathbf{H}^H\right)\right] \quad (6)$$

where $\mathbf{I}_{n_R}$ is $n_R \times n_R$ identity matrix, $\mathbf{H}$ is $n_R \times n_T$ channel matrix and $\mathbf{H}^H$ is the Hermitian transpose of $\mathbf{H}$.

For LOS propagation, and a narrow-band channel at fixed carrier frequency $f_c = c/\lambda$, ray tracing from transmitter to receiver gives following channel matrix $\mathbf{H}$ with complex scalar entries

$$H_{jk} = \frac{\exp\left(-j2\pi\left|T_{node,j} - R_{node,k}\right|/\lambda\right)}{\left|T_{node,j} - R_{node,k}\right|} \quad (7)$$

where $\left|T_{node,j} - R_{node,k}\right|$ is the distance between j th transmit node and k th receive node.

Let $n_R = n_T = n$ be the number of nodes participating in cooperative communication on either side. With $d \gg d_r, d_t$ in figure 2, the path lengths between any pair of transmit and receive nodes are approximately the same to within the array size i.e., $d \approx \left|T_{node,j} - R_{node,k}\right|$, and the complex scalars H_{jk} all have same magnitude but different phases θ_{jk}. Mathematical example of this particular type of normalized $\mathbf{H}$ is given as [13]

$$H_{jk} = \exp\left(j\theta_{jk}\right) \quad (8)$$

where $\theta_{jk} = \frac{\pi}{n}\left[(i-i_0)-(k-k_0)\right]^2$, i_0, k_0 are arbitrary integers.

Consider the case of $n = 2$ and $i_0, k_0 = 0$, the small path length difference

$$\delta = \left|T_{node,1} - R_{node,1}\right| - \left|T_{node,1} - R_{node,2}\right| = \lambda/4 \quad (9)$$

guaranteed that the channel matrix is orthogonal and of full rank, 2 in this case. Practically this can be realized by the slight movement of antennas at base station. Since base station is mostly act as a receiver in the wireless sensor network, positioning of antennas based on received RSSI at base station can be easily achieved.

With these values orthogonal channel matrix become

$$\mathbf{H}_{orthognal} = \begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix} \quad (10)$$

and the capacity from equation (6)

$$C = n\log_2\left[1+\rho\right] \quad (11)$$

Advantage of cooperative MIMO in terms of capacity can be obtained if the channel matrix $\mathbf{H}$ is orthogonal.

4.2.3. Discussion

Interestingly, from equation (5) we see that the distance dependant part of above equation is inversely proportional to the square of Frobenius norm of channel matrix, and, it linearly dependent on the number of transmit nodes. Therefore the receive diversity techniques provide more energy efficient communication.

In clear LOS environment, capacity of wireless channel depends upon the rank of channel matrix **H** as depicted by equation (6). Capacity becomes maximum, when the channel matrix is of full rank (or orthogonal).

5. Simulation Results

We determine the impact of the selection of cooperative nodes and the physical orientation of antennas of BS on the total energy required per bit transmission in WSN and the capacity of the backbone link. The simulations are performed using MATLAB® by The MathWorks.

Figure 3 shows an amount of extra energy consumption per bit compared to our selection algorithm. The transmission distance ratio along x-axis is the ratio of the distance of intra cooperative group nodes without our node selection algorithm and with our algorithm. Since intra cooperative group communication is non-cooperative 1 to 1 communication, therefore more energy consumption per bit is observed in the case of severe fading, *K*=1, it rises with square of the distance and consumes 33% more energy when the nodes separation is 3.5 time to that of closest nodes of our algorithm. With increasing values of *K* (i.e., less fading) the slope of extra energy consumption decreases. Therefore in fading channels our algorithm saves a sufficient amount of energy.

In figure 4 we have total energy consumption per bit as a function of long-hual distance in a rich scattering environment i.e., *K*=1. Among different combinations of transmit and receive cooperative nodes, the largest receive diversity combination ($n_T = 1, n_R = 4$) gives lower energy consumption per bit in a wide range of transmission distance (from 60m to onwards).

Figure 5 shows an improvement in all energy curves as compared to the energy curves in figure 4, this is due to

less severe fading environment, K=2.

When the deterministic or LOS component is dominant K=10, (i.e., the channel is more close to AWGN channel) as shown in figure 6 the lowest energy consumption per bit is given by the case when there is one transmit node and two receive cooperative nodes, it remains with lowest energy per bit up to the transmission range of 100m., after that ($n_T = 1, n_R = 4$) combination becomes the lowest energy. This confirms the effectiveness of large receive diversity in long haul transmissions because at large distances the 1st term (which depends upon d, n_T, and n_R) in equation (2) becomes more dominant than 2nd and 3rd terms.

In pure deterministic or LOS communications with BS, we have neglected the receive circuit energy consumption in the BS, because BS has no energy constraint. Figure 7, shows that the receive diversity case ($n_T = 1, n_R = 4$) renders the least energy consumption per bit in the entire transmission range. Due to the LOS BS antennas orientation the energy consumption per bit is smallest among all other Ricean cases. But the cost we have given for this, is the degradation in spectral efficiency (bps/Hz).

Figure 8 shows a comparison of LOS orthogonal channel capacity ($n_T = 2, n_R = 2$) with ($n_T = 1, n_R = 4$). We can see that there is a trade off between energy consumed per bit and the capacity. In most of the cases capacity is not the main issue, and we are more concerned about the energy consumption.

Finally, figure 9 shows how cooperative energy consumption depends on the constellation size b for various cooperative MIMO schemes. It is clear from Figure 9 that there is an optimal constellation size for each cooperative MIMO scheme for which the total transmission energy per bit is minimized. It again strengthens our argument that receive diversity give more energy efficient cooperative communication. We have used following approx. expression to obtain relation between BER and SER in (3) for K=1,

$$P_b \approx \frac{P_M}{\log_2 M}, \qquad M = 2^b \tag{12}$$

Table 1.

$\bar{P}_b = 10^{-3}$ / ρ	$n_T = 2, n_R = 1$	$n_T = 1, n_R = 2$	$n_T = 2, n_R = 2$	$n_T = 1, n_R = 4$	$n_T = 2, n_R = 4$
K=1	16.6	13.6	8.8	5.8	4
K=2	14.7	11.7	7.9	4.9	3.5
K=10	9.9	6.9	6	3	2.7

6. Conclusion

Geometric orientation based selection of cooperative nodes and the specific placement of antennas at BS, greatly impact the total transmission energy consumption per bit and the capacity of the backbone link. Simulation results show that, using the geometric advantages in selecting the cooperative nodes, we come across an energy efficient transmission for different transmission ranges. Since the sensor network is data centric and most of the time there is unidirectional data flow (from network to BS), therefore cooperative communication with minimum number of transmit nodes and maximum optimal number of receive nodes result in low energy per bit transmission. An optimal constellation size for various cooperative schemes also reveals that receive cooperative diversity is near optimal. A trade off exists between energy consumption per bit and the capacity. Since sensor networks are usually a low data rate networks, therefore we usually opt low energy cooperative communication.

7. Acknowledgement

This work was supported by National advanced technologies researching and developing programs. (China 863 programming, NO.: 2006AA01Z257).

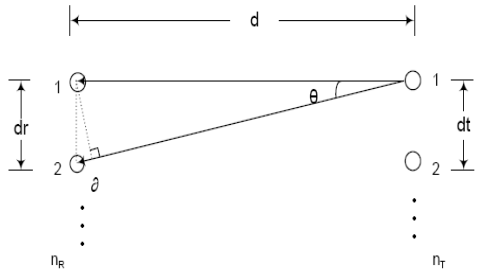


Figure 2. n_T transmit and n_R receive cooperative nodes LOS channel model

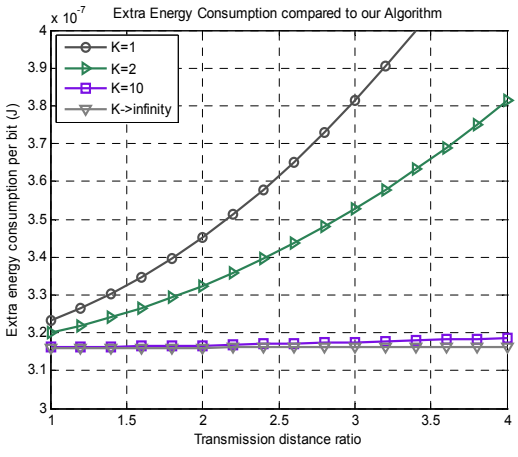


Figure 3. Energy consumption as compared to our algorithm

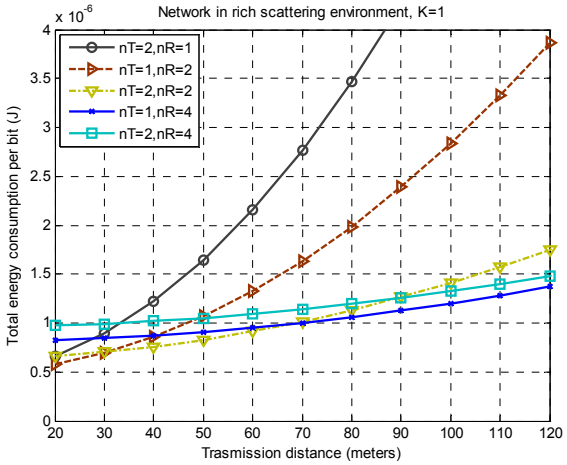


Figure 4. Total energy per bit in Rayleigh fading, K=1

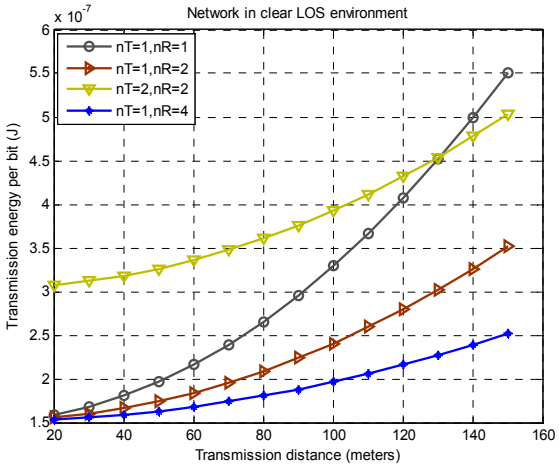


Figure 7. Total energy consumption in backbone link

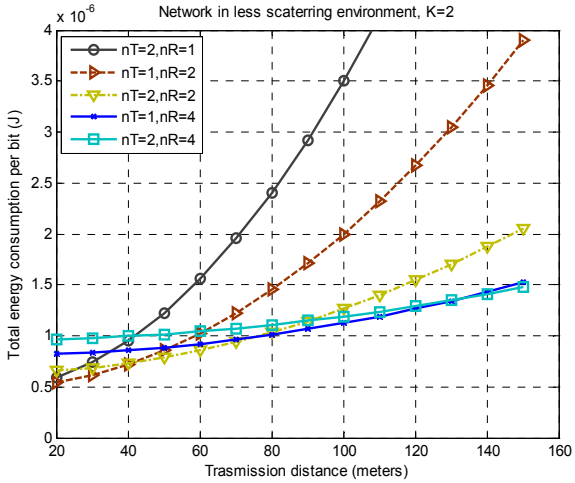


Figure 5. Total energy per bit with K=2

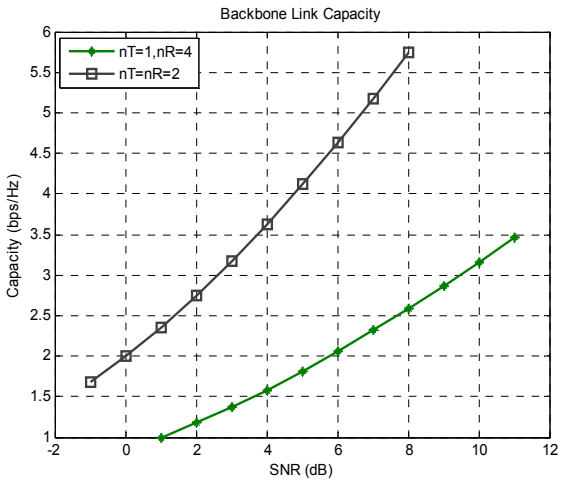


Figure 8. Capacity of backbone link Vs SNR

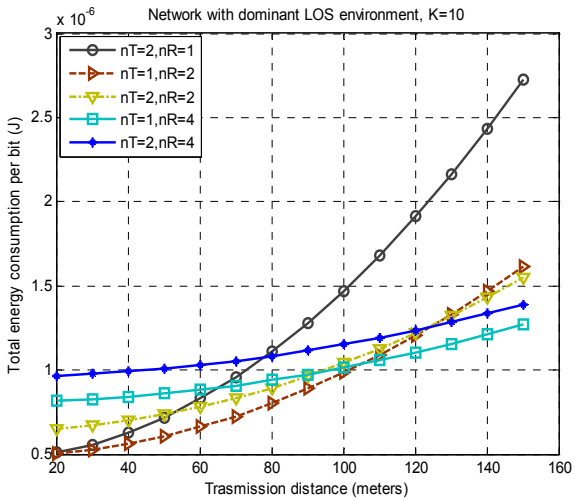


Figure 6. Total energy per bit with K=10

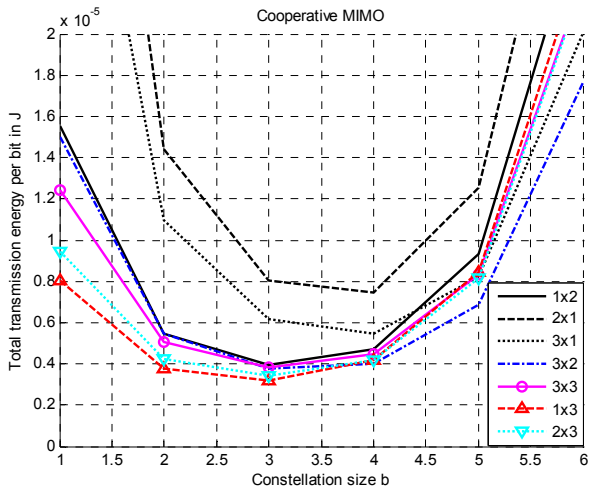


Figure 9. Total transmission energy consumption over b

8. References

[1] A. Paulraj, R. Nabar, and D. Gore, Introduction to Space-Time Wireless Communications, Cambridge Univ. Press, 2003.

[2] C. Shuguang and A. Goldsmith, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," IEEE J. Select. Areas in Comm., vol. 22, no. 6, pp. 1089–1098, Aug.2004.

[3] Wenyu Liu Xiaohua Li Mo Chen, "Energy efficiency of MIMO transmissions in wireless sensor networks with diversity and multiplexing gains," Acoustics, Speech, and Signal Processing, 2005. Proc. (ICASSP'05). IEEE International Conference, Vol.: 4, On page(s): iv/897- iv/900 Vol. 4 March 2005

[4] Wenfeng Li; Weike Chen; Xinzhu Ming, "A Local-centralized Adaptive Clustering Algorithm for Wireless Sensor Networks," Proc. -15th Int'l Conf. on Computer Comm. and Networks, Oct. 2006 Page(s):149 – 154

[5] SangHak Lee; JuneJae Yoo; TaeChoong Chung; "Distance based energy efficient clustering for wireless sensor networks," 29th Annual IEEE International Conference on Local Computer Networks, 2004, 16-18 Nov. 2004 Page(s):567 – 568

[6] O. Younis and S. Fahmy, "Distributed Clustering in Adhoc Sensor Networks: A Hybrid, Energy-Efficient Approach", IEEE INFOCOM 2004, Mar. 2004.

[7] Y. Xu, J. Heidemann, and D. Estrin, "Geography-Informed Energy Conservation for Ad Hoc Routing," MOBICOM 2000, Rome, Italy, July 2001, pp. 70-84.

[8] Sankaranarayanan, L.; Kramer, G.; Mandayam, N.B.; "Hierarchical sensor networks: capacity bounds and cooperative strategies using the multiple-access relay channel model," First Annual IEEE Comm. Society Conference on Sensor and Ad Hoc Comm. and Networks 2004.

[9] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," IEEE Trans. on Information Theory, vol. 22, no. 12, pp. 3062–3080, Dec 2004.

[10] Yong Yuan, Min Chen, Tackyong, "A Novel Cluster-based Cooperative MIMO Scheme for Multihop Wireless Sensor Networks", EURASIP Journal on Wireless and Networking, Vol 2006, Article ID 72493, page 1-9.

[11] S. Alamouti, "A simple transmit diversity technique for wireless communications," IEEE J. Select. Areas Commun., vol. 16, pp. 1451–1458, Oct. 1998.

[12] Joseph Polastre, Robert Szewczyk, and David Culler, "Telos: Enabling Ultra-Low Power Wireless Research," Proceedings of the 4th Int'l symposium on Information processing in sensor networks, Los Angeles, California, Apr.2005

[13] P.F. Driessen, G.J. Foschini, "On the capacity of multiple- input multiple-output wireless channels: a geometric interpretation," IEEE Trans. Commun., vol. 47, No.2, pp.173-176, Feb. 1999.

[14] S. K. Jayaweera, "Virtual MIMO-based cooperative communications for energy-constraint wireless sensor networks", IEEE Trans. on wireless comm., Vol. 5, No. 5, May 2006

I. J. Communications, Network and System Sciences. 2008; 1: 1-103
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

An Energy-aware Hierarchical Architecture Design Scheme

Linfeng YUAN[1], Yang XU[1], Xu DU[2], Wenqing CHENG[2]
[1] *Wuhan Maritime Communication Research Institute, Wuhan, P.R.China*
[2] *Huazhong University of Science and Technology, Wuhan, P.R.China*
E-mail: yuanlf@163.com

Abstract

Compared with the flat architecture in the design of sensor networks, the hierarchical architecture gains much attractive for the reason of scalability, management and energy efficiency. In order to distribute the energy evenly, nodes act the cluster head in some orders. The existing approaches don't pay a critical attention to the overhead during the role rotations. And the duration of a round is a priori, which is very application-specific. An energy-aware hierarchical architecture design scheme is put forward in this paper, namely, Adaptive Minimum Rotational Cost (AMRC) cluster formation scheme. The decision of beginning a new round is made adaptively by the cluster head itself. It combines the dynamic and static advantages in the clustering architecture. The simulation results demonstrate AMRC outperforms some other clustering protocols in many aspects.

Keywords: Sensor Network, Hierarchical Architecture, Energy-Aware, Role Rotation

1. Introduction

Sensor network technology, based on sensing, communication and computing research results, is a key technology for the future. Many future applications will rely on the embedded sensor network [1]. Sensor networks have spurred much interest in the networking research community recently.

Many questions must be dealt with in the design of the sensor network. Efficient routing protocol is one of those significant items. There are two main routing topologies in the sensor network: flat and hierarchical. The hierarchical (clustering) architecture gains much attractive for the reason of scalability, management and energy efficiency.

In the hierarchical architecture, each cluster includes one cluster head (CH) and several common member nodes (MNs). MNs send the data to the CH during the communication and the CH send the gathered data to the sink node or to the other CHs. Usually, the MNs in one cluster will not communicate with each other directly, and the MNs in different clusters communicate with each other via the CHs in their clusters. So the CHs act the roles of "router" in the hierarchical architecture, they take more responsibility than the MNs and will consume more energy than the MNs.

Based on the role assigned to sensor nodes in a system, a hierarchical scheme can be grouped as a static one or a dynamic one [2]. In a static system [3][4], the CHs and their corresponding nodes will not change. So it is hard to minimize the system's energy consumption for non-optimized distances between a CH and its MNs or distribute energy cost among all the nodes. Dynamic mechanisms [5][6][7] can provide better fairness while remaining simplicity. In these dynamic schemes, clusters are periodically regenerated, and nodes take responsibility as the CHs in rotation. However, the regeneration of clusters is energy consuming. Therefore, it is desirable to design a hybrid scheme combining the advantages of both static and dynamic scheme [2].

In order to distribute the energy evenly, nodes act the CH in some orders. The existing approaches have solved the problems in many aspects [6][8][9][10], but they have two main disadvantages. First, they all don't pay a critical attention to the overhead during the role rotations. If the energy cost for the role rotations has added up to a certain degree, it will kill the benefits of hierarchical architecture. Second, these hierarchical protocols decide the beginning of a new round by some constant experimental periodic time. If the round time is too short, the rotational overhead will increase much. Or if the round time is long enough, when the CH has depleted its energy before the time of a new round, the system is "dead" in fact. Therefore, the selection of the duration time is closely dependent on the specific applications,

which will greatly affect the scalability and management of the system.

In this paper, we put forward an adaptive minimum rotational cost (AMRC) energy efficient cluster formation scheme to minimize the rotational times and the rotational overhead. In AMRC, in order to minimize the rotational times, the node with the biggest residual energy will act the CH during the role rotations; in order to minimize the rotational overhead, one round will last for a period until the CH finds its residual energy has dropped to a certain threshold which is only decided by the CH's own energy. Therefore, the decision of beginning a new round is made adaptively by the CH. It needs no information of the other nodes and needs no global information. Each cluster's rotation is independent and the rotation in one cluster will not affect the other clusters. So it combines both the dynamic advantages and static advantages in the hierarchical architecture. And, it is totally application-independent. It can not only distribute energy consumption in the network, but also reduce the overhead to prolong the system's lifetime.

The main contributions of this paper are as follows:

- Minimum rotational cost to reduce the overhead.
- Adaptive round time to meet any application's requirement.
- Combination of dynamic and static advantages in the hierarchical architecture.

The rest of the paper is organized as follows. Section 2 discusses some related work. Section 3 describes the AMRC scheme. Section 4 presents some experiments and compares AMRC with other schemes. The last section concludes this paper.

2. Related Work

In the flat routing architecture, each node typically plays the same role and sensor nodes collaborate to perform the sensing task. In hierarchical routing architecture, higher-level nodes (CH) can be used to process and send the information, while low-level nodes (MN) can be used to perform the sensing in the proximity of the target [11]. The hierarchical techniques can greatly contribute to overall system scalability, lifetime, and energy efficiency for load balancing and efficient resource utilization [2][5][6][7][11].

The CH will cost more energy than its MNs. In order to distribute energy consumption evenly among all the nodes including CHs and MNs, role rotations will be used to let all nodes take turn to act the CH [2][5][6][7]. The selection of CH can be weight-associated [4][12][13][14] or weight-independent [6][7]. The latter is simpler to implement and more robust in the presence of network dynamics while the former can produce more optimal clusters, which is more essential to the sensor networks. In weight-associated selection methods, the selection of CHs can be based on node ID [12][13], nodal degree [14], residual energy [4][6] or a combination of several parameters [5][12]. Since the energy consumption is more sensitive to the sensor network, the residual energy is often used as the criterion for CH selection. They all take the ratio, the node's residual energy to the sum of all nodes' residual energy, as the selection principle. Every node will need to know the sum either by a central controller or by computing on its own, it is hard or costly to get the computation results. In AMRC, each node is restricted within one cluster forever, so the selection of CH is very easy to be decided.

Many times of role rotations will take place in clustering formation, which will cost lots of energy too. But none of the existing cluster-based routing schemes have paid a critical attention to the overhead during the role rotations. Some have tried to reduce the rotation times by setting appropriate round period time. For example, LEACH [6] sets the round time to enable each node has enough energy to act CH once and act MN several times throughout the simulation lifetime. It gives no analysis about the number of role rotations and the energy consumption for these rotations. Though the network operation interval (T_{no}) is forced to be much bigger than the clustering process interval (T_{cp}) in HEED [5], the decision of its privilege, it will become the CH again if it has more energy than the other neighboring nodes. So this rotation is not necessary, and it will cost some energy among all the nodes in this cluster. In AMRC, each role rotation will happen in the same cluster and it will not affect any other clusters, different clusters have different number of role rotations and different duration time.

3. AMRC Scheme

3.1. Connectivity Subnet

It will efficiently decrease the overhead incurred by the interfaces among the clusters if each role rotation will only happen in its own cluster and it will not affect other clusters. So in AMRC, we restrict that the members in each cluster will not change throughout the whole transmission lifetime.

After the system runs in a steady state, the position of the sink node is relatively fixed. In order to make fully use of the position information in this protocol, the sink node lets all related nodes know its coordinates before the route discovery process. We need to establish a Connectivity Subnet (or Connectivity Set, CS). The CS is organized as follows. The sensing area is divided into sub-areas according to the position information, which must ensure the nodes in either sub-area can directly communicate with any node in the neighboring sub-area.

GAF [15] has also put forward the concept of virtual grid to try to ensure the direct communication, but in fact, it cannot ensure the direct communication between the diagonal neighboring sub-area, as shown in Figure 1.

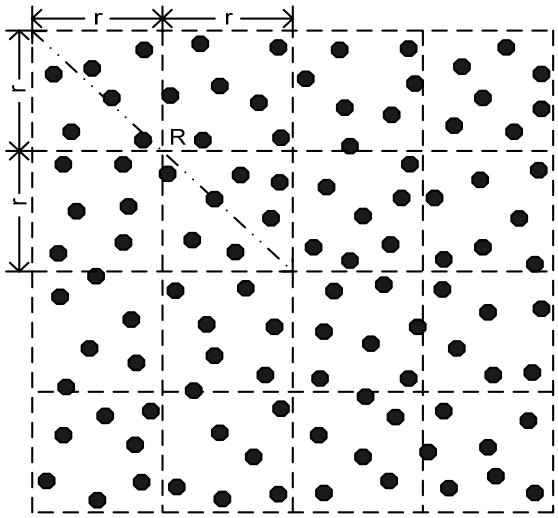


Figure 1. Connectivity subnet

We will establish the CS as follows. In Figure 1, the dashed line surrounds each sub-area and sixteen sub-areas are involved here. If the communication radius of each node is R and the square's length of each sub-area is r, in order to ensure direct communication for any nodes in the neighboring sub-areas, R and r must meet the following requirement:

$$(2r)^2 + (2r)^2 \le R^2$$

That is

$$r \le \sqrt{2}R/4$$

In this case, whichever node is selected as the CH in each cluster, it can directly communicate with the other CHs in the neighboring sub-areas. The restriction ensures that each role rotation can only take place in the same cluster, and it will not affect the other clusters. It also implies that different clusters can have different number of role rotations and different interval time.

3.2. Decisive Rotational Threshold

Since the role rotation in one cluster has no relationship with the other clusters, we first consider the rotation process within one cluster. Suppose the *i*th node is the CH in a cluster at some time. If the CH finds its residual energy has decreased to some threshold, it will tell all the nodes in this cluster to begin a new round.

Denote $E_i(CH)$ as the energy when the node acted as the CH and $E_i(i)$ as its residual energy at present. In AMRC, the threshold to begin a new round is $\alpha \cdot E_i(CH)$, where α is a coefficient parameter of the threshold and it is determined before the system works. That is to say, if the energy of the CH has dropped to the α times of the energy when it acted the CH in this round, it needs to initiate a new round selection, as shown in the following term,

$$E_i(i) \le \alpha \cdot E_i(CH)$$

Since the energy $E_i(CH)$ is different for each node and it is also different for each node in different round, the time to begin a new round is changeable. It means the role rotations will change adaptively.

3.3. Cluster Formation

There are two main issues in the design of the hierarchical architecture: the number of clusters and the cluster formation. We have discussed the first one in above subsection, now we consider the second issue.

In AMRC, because the nodes in a specific cluster will not change throughout the lifetime, the main problem in the cluster formation is to select a new CH within this cluster. If the old CH decides to begin a new round, it will broadcast an energy-request message (*ReqEn*) using a non-persistent carrier-sense multiple access (CSMA) MAC protocol. Each MN transmits its residual energy with an acknowledgement message (*ACK*) back to the CH. The CH compares all the MNs' residual energy and informs the node with the highest one to become the new CH in the coming round with an indication message (*IND*). The new CH sets up a TDMA schedule and transmits this schedule to the MNs in the cluster. After all nodes in the cluster know the TDMA schedule, the set-up phase is complete and the steady-state network operation phase (data transmission) can begin. Figure 2 gives the AMRC flowchart.

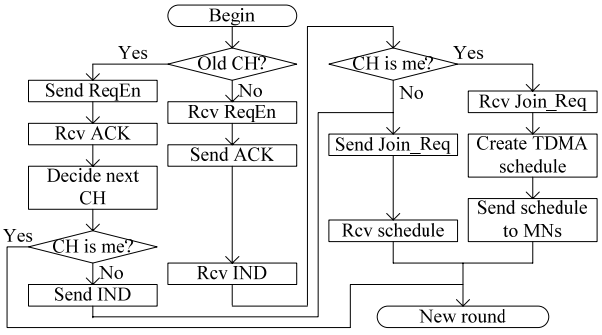


Figure 2. AMRC flow chart

3.4. Some Discussions

The design of hierarchical architecture can gain many benefits from the AMRC's solutions.

First, the process of AMRC ensures that the node with the highest residual energy will act the CH in the local area during the role rotations. Second, it is a prior for each node to know the parameters *k* (number of clusters) and *N* (total number of nodes) in LEACH-like protocols [5], while in AMRC, the decision to begin a

new round is made adaptively by the CH own and it needs no information of the other nodes or any global information. Third, the working status is independent for each cluster and each node in one cluster will act the CH in some turns, so it combines the static and dynamic advantages successfully. Fourth, the duration of each round is determined adaptively, so it is not application-specific. The parameters α can be set with a bit smaller value when there are many packets should be transmitted after establishing the source-destination pair, so as to reduce the number of rotations. Or else, it can be set with a bit larger value when only a few packets need to deliver from the source node to the destination node, so as to balance the energy consumption more efficiently.

From these benefits, AMRC can reduce the number of rotations and minimize the rotational overhead significantly.

4. Performance Simulation

In this section, we evaluate the performance of AMRC protocol in the ns-2 simulator. Some parameters are set as seen in Table 1.

Table 1. Parameters setting

Parameter	Meaning	Value
$M*M$	sensing region	200m*200m
N	total number of nodes	100
R	radio propagation range	50m
E_{elec}	electronics energy	50nJ/bit
ε_{fs}	free space coefficient	10pJ/bit/m^2
ε_{mp}	multipath fading coefficient	10pJ/bit/m^2
L_{data}	data packet size	100bytes
$E_i(max)$	initial energy of each node	1Joules

We compare AMRC with LEACH-like protocols and evaluate the following performance metrics:

- The number of role rotations.
- Energy consumption in the transmission.

4.1. The Number of Role Rotations

We first examine the number of role rotations in LEACH-like protocols. In LEACH-like protocols, there are two steps in each round: the cluster process time T_{cp} (set-up phase) and the network operation time T_{no} (steady-state phase). Along with the differences of the network operation time T_{no}, the number of role rotations is also different. In our simulation, the network operation time T_{no} are set as 20, 40, 60, 80 seconds, the number of role rotations is calculated every 100 seconds and the system lasts for 700 seconds. Figure 3 shows the simulation results.

From the figure, we can find that at the time of 700 second, the number of role rotations is 24 when Tno equals to 100 second, but it reaches 132 when Tno equals to 20 second. It indicates that Tno has a great influence to the number of role rotations in LEACH-like protocols.

While in AMRC protocol, the network operation time is not decided by manual operation and it is not application-specific, so it cannot influence the number of role rotations directly. It is the parameter of α that can decide the number of role rotations. With the same settings in LEACH-like protocol, we set α as 0.8, 0.85, 0.9, 0.95, and also examine the number of role rotations when the system works in the time from 100 second to 700 second. Figure 4 shows the simulation results.

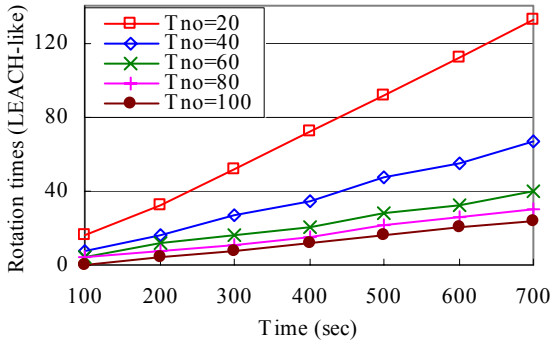


Figure 3. The number of role rotations in LEACH-like protocol

From Figure 4, when we compare the number of role rotations in the same times, we can find along with the increasing number of α, the number of role rotations gets larger. At the time of 700 second, the number of role rotations is 4 when α equals to 0.8, and it is 22 when α equals to 0.95. This implies that the decisive rotation threshold is more likely satisfied when α is smaller.

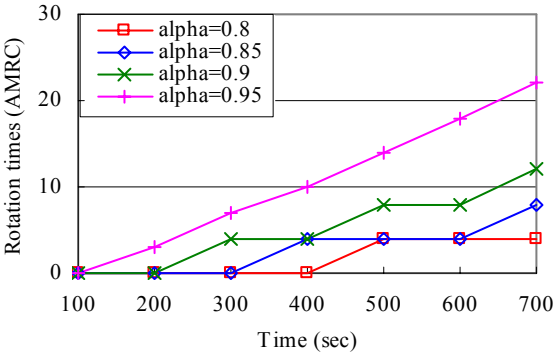


Figure 4. The number of role rotations in AMRC protocol

Comparing Figure 3 and Figure 4, we can also find that at the same time, all the number of role rotations in AMRC protocol is smaller than those in LEACH-like protocols, even that of the worst curve in Figure 4 (α=0.95) is smaller than that of the best curve in Figure 3 (T_{no} =100). In fact, α equals to 0.95 means that the CH will begin a new round when it detects that its

current energy has dropped 5%, but 5% is not necessary for the CH to give up its CH's privilege, and that CH can continue to do more jobs at that time.

4.2. Energy Consumption in the Transmission

We set the network operation time as 20, 40, 60, 80 seconds in LEACH-like protocol and set α as 0.9 in AMRC protocol, Figure 5 shows the energy consumption of all the nodes during the transmission per 100 seconds.

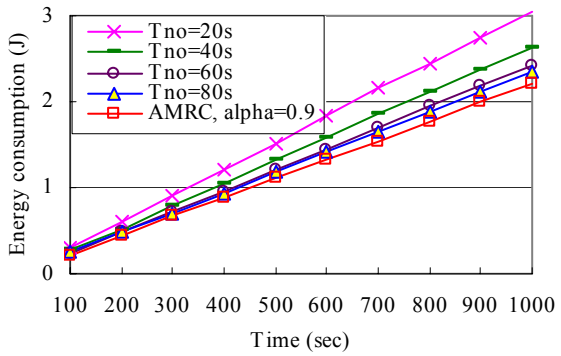


Figure 5. Energy consumption

In LEACH-like protocols, the energy consumption gets smaller when the network operation time is longer. The differences of energy consumption come from the different number of the role rotations. This implies that from the view of the energy efficiency, the role rotations cannot be neglected in the hierarchical architecture system. In AMRC protocol with α equals to 0.9, the energy consumption is smaller than those in LEACH-like protocols at each corresponding time. It is for the reason that decreasing the number of role rotations that the energy consumption is reduced too. It can be expected that when α gets smaller, the energy consumption for all nodes will reduce further.

5. Conclusion

The LEACH-like hierarchical protocols suffered from the overhead for the role rotations and they are application-specific. In this paper, the AMRC energy efficient scheme is put forward to minimize the number of role rotations and the rotational overhead. Each cluster can take its own rotation without affecting the other clusters. And the adaptive round time makes it widely used in any application scenarios.

6. Acknowledgement

This work is supported by the National Natural Science Foundation of China (No.60572049) and the Natural Science Foundation of Hubei Province, P.R.China (No. 2005ABA264).

7. References

[1] John A. Stankovic, Tarek F. Abdelzaher, Chenyang Lu, Lui Sha, Jennifer C. Hou, "Real-Time Communication and Coordination in Embedded Sensor Networks", Proceedings of the IEEE, July 2003, Vol.91, No.7, pp.1002-1022.

[2] Quanhong Wang, Hossam Hassanein, Glen Takahara. "Stochastic Modeling of Distributed, Dynamic, Randomized Clustering Protocols for Wireless Sensor Networks", Proceedings of the 2004 International Conference on Parallel Processing Workshops, pp.1-8.

[3] C. F. Chiasserini, I. Chlamtac, P. Monti, and A. Nucci, "Energy Efficient Design of Wireless Ad Hoc Networks", Proc. of Networking 2002, Lecture Notes in Computer Science (LNCS), Italy, May 2002, No.2345, pp.387-398.

[4] S. Ghiasi, a. Srivastava, X. Yang, and M. Sarrafzadeh, "Optimal Energy Aware Clustering in Sensor Networks", Sensors Magazine, 2002, Vol.19, No.2, pp.258-269.

[5] O. Younis, and S. Fahmy. "Distributed Clustering in Ad-hoc Sensor Networks: A Hybrid, Energy-Efficient Approach", IEEE INFOCOM 2004, March 2004, Vol.1, pp.629-640.

[6] W.B.Heinzelman, A.P.Chandrakasan, H.Balakrishnan. "An Applicationspecific Protocol Architecture for Wireless Microsensor Networks", IEEE Transactions on Wireless Communications, Oct. 2002, Vol.1, No.4, pp.660-670.

[7] S. Bandyopadhyay and E. J. Coyle, "An energy efficient hierarchical clustering algorithm for Wireless Sensor Networks", IEEE INFOCOM 2003, Vol.3, pp.1713-1723.

[8] Mark Perillo and Wendi Heinzelman, "DAPR: A Protocol for Wireless Sensor Networks Utilizing an Application-based Routing Cost", IEEE WCNC 2004, Vol.3, March 2004, pp.1540-1545.

[9] Yan Yu, Ramesh Govindan, Deborah Estrin. "Geographical and Energy Aware Routing: a recursive data dissemination protocol for wireless sensor networks", In Seventh Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'00), August 2000.

[10] Fan Ye, Haiyun Luo, Jerry Cheng, et al. "A Two-tier Data Dissemination Model for Large-scale Wireless Sensor Networks". In Proc. of the 8th Annual International Conf. on Mobile Computing and Networking (MobiCom'02), September 2002, pp.148-159.

[11] Jamal N. Al-karaki, Ahmed E. Kamal, "Routing techniques in wireless sensor networks: a survey", Wireless Communications, IEEE [see also IEEE Personal Communications], vol: 11, issue: 6, Dec 2004, pp. 6-28.

[12] D. J. Baker, and A. Ephremides, "The architecture organization of a mobile radio network via a distributed algorithm", IEEE Tran. On Communications (Legacy, pre 1988), vol.29, issue: 11, Nov. 1981, pp.1694-1701.

[13] A. Ephremides, J.E.Wieselthier, and D.J. Baker, "A

design concept for reliable mobile radio networks with frequency hopping signaling", Proceedings of IEEE, vol.75, Issue: 1, Jan. 1987, pp. 56-73.

[14] M. Gerla, and J. T. C. Tsai, "Multi cluster, mobile, multimedia radio network", Wireless Networks, vol.1, issue: 3, 1995, pp.255-265.

[15] Xu Y, Heidemann J, and Estrin D. "Geography-informed Energy Conservation for Ad Hoc Routing". In Proc. 7th Annual International Conference on Mobile Computing and Networking (MobiCOM 2001). July 2001, pp. 70-84.

I. J. Communications, Network and System Sciences. 2008; 1: 1-103
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

Efficient DPA Attacks on AES Hardware Implementations

Yu HAN[1], Xuecheng ZOU, Zhenglin LIU, Yicheng CHEN
Department of Electronic Science & Technology, Huazhong University of Science & Technology, Wuhan, P.R.China
E-mail: [1]husthyt@gmail.com

Abstract

This paper presents an effective way to enhance power analysis attacks on AES hardware implementations. The proposed attack adopts hamming difference of intermediate results as power mode. It arranges plaintext inputs to differentiate power traces to the maximal probability. A simulation-based AES ASIC implementation and experimental platform are built. Various power attacks are conducted on our AES hardware implementation. Unlike on software implementations, conventional power attacks on hardware implementations may not succeed or require more computations. However, the method we proposed effectively improves the success rate using acceptable number of power traces and fewer computations. Furthermore from experimental data, the correlation factor between the hamming distance of key guesses and the difference of DPA traces has the value 0.9233 to validate power model and attack results.

Keywords: Security, AES, Differential Power Analysis (DPA), Power Model, Correlation Factor

1. Introduction

The security in mobile applications [1] is of crucial importance because a large number of nodes may be exposed in a hostile environment. And if only one node is captured by attackers, the impact to the whole network can be devastated. Therefore, various cryptographic services required for these applications involve not only solutions for data protection but also self-implementation concerns. Mobile nodes are usually equipped with hardware coprocessors which are used to perform security protocol. If they are captured by attackers, side-channel information leakages, such as timing, power consumption and electromagnetic radiation, may be monitored for cryptanalysis. Among them, differential power analysis (DPA) [2] poses a serious threat to the security of different cryptographic implementations because it is practical, non-invasive, and easy to repeat.

Power analysis attacks exploit the correlation [3] between the data and the instantaneous power consumption of cryptographic devices. As this correlation is usually very small, statistical methods should be used to exploit it efficiently. In a power analysis attack, an attacker first creates a hypothetical power model of the cryptographic device at a very abstract level. In practice, each cryptographic algorithm designed operates only small parts of the secret key, called subkey, at certain period. Thus, the attacker can write a simple computer program that executes the algorithm at that period. The program calculates the intermediate result of this part for all possible subkey guesses. These values allow for predicting the power consumption, which is related to the inputs of cryptographic algorithms and subkey guesses. Next, the attacker feeds the same inputs to the real cryptographic device and measures its power consumption. Then the attacker correlates the predictions of the power model with real power consumption. For all wrong key guesses, the predictions will not correlate with the real measurements, but for the correct key guess, there will be a visible peak for the power analysis traces. In order to set up the correlation, predictions from different power models and statistical methods must be tested.

AES [4] is a new symmetric block cipher standard, which was issued by the National Institute of Standards and Technology (NIST) on November 26, 2001. AES has special particularities suitable for area- and power-constrained applications. Hence, the secure AES implementation can greatly affect the nodes in severely resource-constrained networks. AES is a round-based symmetric block cipher and can be implemented efficiently on all kinds of platforms. The standard key size is 128 bits. But for some applications, 192 and 256-bit keys can be supported as well. The round consists of four different operations, namely, *SubBytes*, *ShiftRows*, *MixColumn*, and *AddRoundKey*. Each operation maps a 128-bit input state into a 128-bit output state. The state is represented as a 4×4 matrix of bytes. The number of rounds depends on the key size. For a 128-bit key (AES-

128 is our concern), the round starts with a single *AddRoundKey* operation followed by 9 identical computation rounds. And, a slight difference is that the final round has no *MixColumns* operation. Figure 1 shows an AES-128 encryption diagram, and more details can be found in [4].

In this paper, we conduct a successful DPA attack on an AES hardware implementation to examine its security. The remainder of the article is organized as follows. We review related work in section 2. Section 3 addresses the principle of our presented attack. A simulation-based experimental environment and power acquisition are presented in section 4. Experimental results and discussions are provided in section 5. Finally, we conclude in section 6.

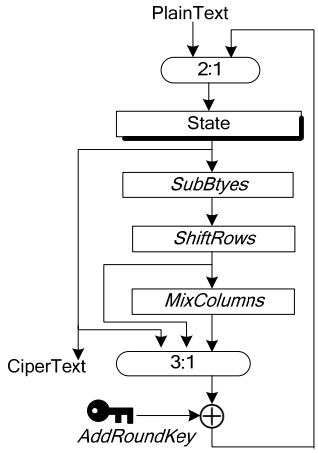


Figure 1. AES-128 encryption flow

2. Related Work

One of the most typical targets of power analysis attacks is the smart card, which is capable of performing secure computations. It consists of a (typically, 8-bit) processor, together with ROM, EEPROM, and a small amount of RAM. The cryptographic software basically operates on 8-bit data blocks because of the 8-bit architecture. In the original paper [2], P. Kocher et al. announced a DPA attack against the DES implementation in smart card microprocessors. In [5], T. S. Messerges et al. extended the research and presented experimental data and attack details. E. Brier [6] enhanced power analysis attacks using correlation factor between power samples and hamming weight of the handled data. All of these attacks have been extensively proved to be effective on symmetric and public-key encryption schemes in smart cards.

Contrasted to software implementations in smart cards, hardware implementations in FPGAs and ASICs are usually required for their ability to deal with high throughput. They allow parallel computing and have a more flexible architecture (as they are under the control of designers). Due to different processing behaviors and physical characteristics, a simple hamming power model can not be used to predict the power consumption of hardware implementations. Hardware implementations may leak less data-dependent power information to resist DPA attacks than software implementations. Recent publications [7, 8] show that DPA attacks has effectively defeated hardware implementations on cryptographic circuits as well. However, it requires a number of power measurements and a high computational complexity. To retrieve a secret key, an attacker must know more implementation details to deduce more precise power models. And he has to perform more complex statistical analyses to pre-process power data.

An important improvement has come with the appearing of high-order DPA [9]. This type of attacks generalizes DPA attacks by simultaneously considering multiple samples that correspond to several intermediate values within the same power trace. Template attack is presented as a new variant of power analysis attack. According to [10], this is the strongest form of side channel attack possible in an information theoretic sense. However, these new attacks require a deeper knowledge of the experimental device and more time consuming to mount. In this sense, it is therefore much less general. We have no comparison with attack results of these new methods in this paper.

According to above discussions, our work aims to improve first-order DPA attacks by developing a simple attack strategy against AES hardware implementations with fewer computations and more generality.

3. Principle of Improved DPA Attack

Current power analysis attacks [2][5][6][7][8] exploit the fact that the power consumption of a device executing a algorithm depends on the intermediate results handled. In these DPA attacks, random plaintext inputs and hamming model are used. Further, we assume that the differences between two different power measurements at the same sampling time are also related to the differences of the intermediate results at least to a certain degree. Thus, we deduce an improved power model with hamming difference of intermediate results, not hamming weight or hamming distance power model. In addition, partial plaintext inputs are fixed to set up DPA traces.

Power attacks can be divided into single-bit DPA and multi-bit DPA. In a single-bit DPA attack, a certain bit of intermediate result is predicted. It is used to split the power measurements into two sets, of which the means are computed and subtracted. For multi-bit DPA, multiple bits of intermediate results are predicted. In the two contexts, we have to verify the peak of bias signal by observing DPA traces. It is often subjective. The CPA is presented when a correlation factor between the outputs of power model and real power traces is shown. Correlation factors can be directly compared in different

CPA traces at cost of computational complexity. Our improved DPA approach overcomes these drawbacks in the following analyses.

3.1. Improved Power Model

We suppose that at time t an intermediate result, $I(x, t, k)$, which is an n-bit word, only depends on plaintext x and key k. Therefore, the general hamming weight model [6] based on a zero reference state for power consumption can be defined as

$$P(t) = aH[I(x,t,k)] + b. \quad (1)$$

Here a denotes a scalar gain between the hamming weight H and the instantaneous power consumption $P(t)$, and b is a hardware-dependent constant. Considering a plaintext x_1 with corresponding intermediate result I_1, we can obtain $P_1(t) = aH[I_1(x_1,t,k)] + b$. Similarly, when another plaintext x_2 results in the opposite intermediate result I_2, the corresponding instantaneous power consumption is $P_2(t) = aH[I_2(x_2,t,k)] + b$. Then, we can get the maximal difference of the power model in an absolute value:

$$\begin{aligned} &|\Delta P(t)| = |P_1(t) - P_2(t)| \\ &= \alpha |H[I_1(x_1,t,k)] - H[I_2(x_2,t,k)]| = \alpha \times n. \end{aligned} \quad (2)$$

where n is 128 for AES. This means that the power consumption of the whole circuit tends to be maximal. There is still the possibility even though an 8-bit subkey is concerned.

SubBytes is the sole nonlinear operation of AES, which consumes a majority of the area and power of AES. In addition, any intermediate result that occurs after *MixColumns* depends on 32-bit of the round key. This lead to a large number of subkey guesses needed to be tested, which is impractical. Furthermore, the subkey used for guessing should be the original key without key expansion. Accordingly, the intermediate results to predict power consumption target the output byte of the initial *AddRoundKey* or the output byte of following *SubBytes*. Each of them is a function of the plaintext byte and corresponding subkey guess. If the first byte subkey (denoted as *Ks*) of the first round key is targeted, and x_1, x_2 are the two corresponding plaintext bytes in two encryptions. We use the hamming difference of intermediate results under two different plaintext inputs as our power model, which shown as Figure 2. The predictions of the power model are given as follows:

$$P = aH(I_1) - aH(I_2). \quad (3)$$

For the target after *AddRoundKey*, I_1=*Ks* $\oplus$ x_1 and I_2=*Ks* $\oplus$ x_2. Whereas for the target after *SubBytes*, I_1=*SubBytes*(*Ks* $\oplus$ x_1) and I_2=*SubBytes*(*Ks* $\oplus$ x_2). We consider two cases: I_1=*0x00* and I_2=*0xff* regardless of which target. Here plaintext byte x_1 is for the former and x_2 for the latter. Thus, we can derive the corresponding plaintext bytes for each subkey guess *Ks* from the maximal prediction of the improved power model.

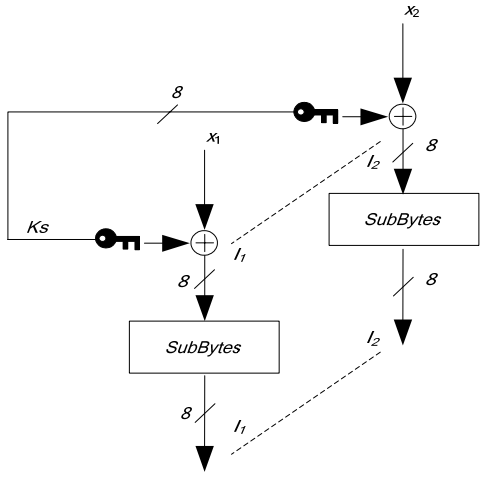


Figure 2. The improved hamming difference power

3.2. Improved DPA Traces

The generic power analysis attacks pay little attention to the choices of plaintexts during power sampling. The plaintexts are usually assumed random. But, for our improved attack, we first set plaintext bytes as x_1, x_2 obtained from the hypothetical power model, separately. Then, keeping other plaintext bytes random can result in a uniform distribution of remained partial bits of the intermediate results. And their influence on real power measurements can be eliminated if the number of random plaintext inputs is enough. Finally, for each subkey guess, we can prepare two plaintext sets as follows, each of which contains m plaintexts.

$$\begin{aligned} S_1(Ks) &= \{S_1[Ks,i] : (x_1, PTi[119:0]) \mid 1 \le i \le m\} \\ S_2(Ks) &= \{S_2[Ks,i] : (x_2, PTi[119:0]) \mid 1 \le i \le m\} \end{aligned} \quad (4)$$

where *PTi*[119:0] denotes 120 random plaintext bits, and they are the same in S_1 and S_2. Therefore, we can derive the plaintext inputs from the hypothetical power model for every subkey guess. Since there are 256 AES subkey guesses, the total plaintext number of two sets is 512×m.

For each subkey guess *Ks*, we perform AES encryptions using the above two plaintext sets in real power acquisition stage. And two power trace sets can be obtained, denoting $E(S_1(Ks), t)$ and $E(S_2(Ks), t)$. DPA trace between the two cases is presented as follows for the subkey guess *Ks*.

$$\Delta E(Ks,t) = \sum_{i=1}^{m} E(S_1[Ks,i],t) - \sum_{i=1}^{m} E(S_2[Ks,i],t) \quad (5)$$

when the correct subkey is assumed, a peak can be identified since it is obviously higher than other subkey guesses. In addition, we do not need to use the averaging statistics since the numbers of power traces in two sets are equal.

4. Power Acquisition in Simulation-Based Environment

There is now a demand to evaluate the DPA resistance of AES circuits. Thus, a typical hardware implementation of AES has been developed. The experimental conditions are shown in TABLE 1.

Table 1. Experimental Conditions

Description- languages	Verilog-HDL
Design technology	UMC CMOS 0.25μm 1.8v
Logic synthesizer	Synopsys DesignCompiler v200509
Power simulator	Synopsys PrimePower v200406sp1
PC spec.	CPU: Ultra SPARC Ⅱ 450MHz, Memory: 4GB, OS: Solaris9

Our AES implementation [11] is unfolded, including 16-byte registers for storing the intermediate results for each round of operation. All registers are set zeros when starting the encryption. The remainder combinatorial circuits perform *SubBytes*, *ShiftRows*, *MixColumns* and *AddRoundKey* operations. Each round operates during one clock cycle. And, we implemented the *SubBytes* unit with a classic GF architecture [12], which is especially suitable for resource-constrained applications.

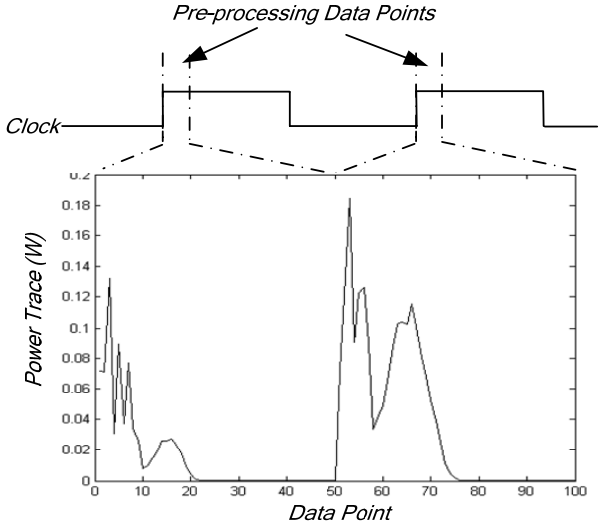


Figure 3. The power trace of the first two clock cycles measured for one AES encryption

During the power acquisition stage, the clock frequency applied to the system was 2.5 MHz and the sampling frequency was 1 GHz. The initial key addition operation occurred during the first clock cycle. And the result of this operation was written into register at rising edge of the second clock cycle. Hence, in the simulation-based power acquisition environment, we only measured the power consumption of the target period (the first two clock cycles of every encryption operation) during an AES encryption. 800 data points for two clock cycles during each encryption were acquired. One power trace is shown in Figure 3. The main component of power consumption is dynamic power consumption caused by data switching. The power trace in Figure 3 approaches the minimum from the 30th data point after the rising edge of the clock cycle, which is due to the critical path delay. To reduce computational complexity, a pre-processing technique was necessitated to eliminate the last 350 data points during every clock cycle. We considered those remaining 100 data points for two clock cycles as the instantaneous power consumption related to the target intermediate results.

5. Experimental results and discussions

5.1. Experimental results

We first conducted original power attacks on our AES implementation in the experimental environment, which involved singe-bit DPA, multi-bit DPA and CPA. We could not retrieve the right subkey from single-bit DPA and multi-DPA using 6000 power measurements. A CPA attack on the intermediate results of *AddRoundKey* revealed the correct subkey based on 4000 power measurements. As to the intermediate results of *SubBytes*, none of these attacks was successful. Successful CPA results are shown in Figure 4. The black plot denotes the

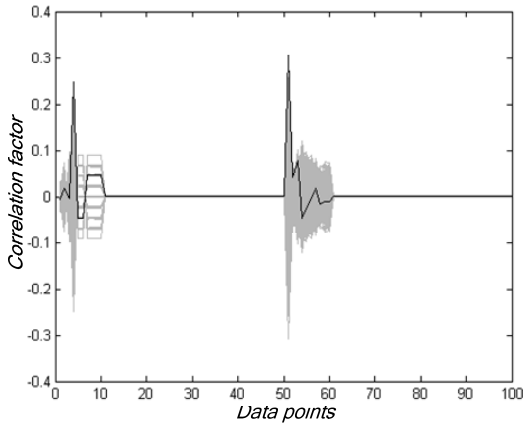


Figure 4. Correlation coefficients for 256 subkey guesses during two clock cycles

correlation coefficient traces for the correct subkey guess 0x74, whose peak is a little higher than the second highest point for incorrect subkey guess 0x16. In addition, an attacker can learn the moment of time when the instantaneous power consumption has a maximal correlation with the intermediate results of *AddRoundKey* by observing Figure 4. It also means that the AES hardware implementation has a maximal probability to leak data-dependent power at 5ns and 454ns during its encryptions. These two moments are closely related to the first *AddRoundKey* operation and the affine transformation of *SubBytes* operation, respectively. Thus, we conclude that these linear operations in the AES implementation result in more data-dependent power

leakages than other round operations.

We performed the improved power attack on the intermediate results of *AddRoundKey* and *SubBytes*, respectively. Like CPA, the correct subkey was extracted only for the intermediate results of *AddRoundKey*, and we took m = 10, denoting 5120 power measurements. Figure 5 shows the results of our improved attack. The peak for our improved DPA traces also occurs at about 4ns after starting AES encryption. That means the chosen plaintext inputs according to the improved power model generate

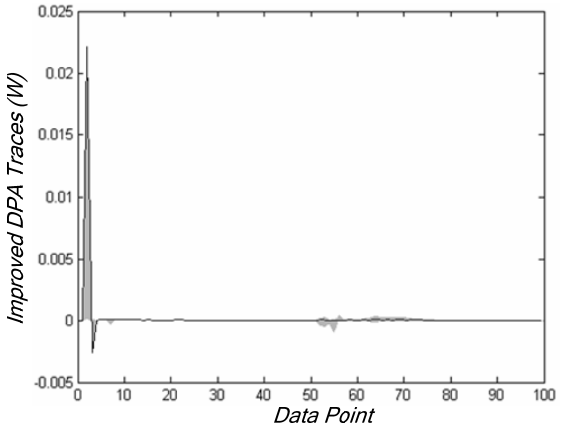


Figure 5. Improved DPA traces for 256

the maximal DPA signal when attacking the first *AddRoundKey*. In addition, the bias signal of the improved DPA trace between the peak and the second highest point is about 1.1mW, which provides a more effective comparison to the above results of CPA.

5.2. Discussions

From the results of our improved DPA, we compute the hamming distance *HD* between the subkey guess *Ks* and the extracted subkey *Kr* as following.

$$HD(Ks) = H(Ks \oplus Kr) \tag{6}$$

Since each power trace for attacking *AddRoundKey* contains 50 data points, we assume the sum of these samples as interesting DPA traces, computing as follows.

$$\Delta E(Ks) = \sum_{t=1}^{50} \Delta E(Ks, t) \tag{7}$$

Then, we also compute the difference $|\Delta P|$ between the two interesting DPA traces corresponding to subkey guesses as follows.

$$|\Delta P(Ks)| = \big||\Delta E(Ks)| - |\Delta E(Kr)|\big| \tag{8}$$

The correlation factor between *HD* and $|\Delta P|$ is 0.9233, which shows that $|\Delta P|$ has a perfect linear relation with *HD*. Therefore, we confirm that hamming power model is valid for our proposed method, and the correct subkey has been retrieved.

Single-bit DPA attacks have been successfully conducted on cryptographic software implementations in smart cards, but it is often not the fact for hardware implementations. It is due to their differences in processing behaviors and physical characteristics. Multi-bit DPA can upgrade the peak level by encrypting a large amount of random plaintexts. But sometimes it is impractical. CPA exploits the data-dependent power leakage in a statistical way, which has been proved to be effective on both software and hardware implementations. However, in comparison with our improved DPA, CPA uses the correlation coefficient which needs to compute a mass of expectations, variances and square roots. Only summing and subtracting are required in our improved DPA. Further, the proposed DPA peak can be verified more objectively than original DPA attack.

6. Conclusion

In this paper, an effective DPA method to retrieve the secret key from an AES hardware implementation is presented. Based on the improved power model, we can prepare the corresponding input plaintext bytes for every subkey guess in advance. In addition, our DPA traces can be built through simple summing and subtracting operations instead of complex statistical techniques. As the partitioning criterions of single- and multi-bit DPA are usually abstract and simple, these two DPA methods can not retrieve any useful information even with 6000 power measurements. Although the CPA attack can extract the right subkey based on 4000 power measurements, its computational complexity sometimes exhibits a bottle-neck. Compared with the methods mentioned above， our proposed DPA excels them in both effectiveness and computation requirements. Furthermore, the perfect linear relation validates the improved DPA attack and the power model by analyzing experimental data.

7. Acknowledgement

The research described in this paper has been supported by High technology Research and Development Program of China under grant 2006AA01Z226 and by Scientific Research Foundation of Huazhong University of Science and Technology under grant 2006Z001B.

8. References

[1] A. Perrig, R. Szewczyk, V. Wen, D. Culler, and J. D. Tygar, Spins: Security protocols for sensor networks, Wireless Networks, Vol. 8, pp. 521-534, 2002.

[2] P. Kocher, J. Jaffe, and B. Jun, Differential power analysis, in Advances in Cryptology—CRYPTO 99. Heidelberg, Germany: Springer-Verlag, 1999, vol.

1666, Lecture Notes in Computer Science, pp. 398–412.

[3] J.M.Rabaey, A.Chandrakasan, and B.Nikolic, Digital Integrated Circuits, A Design Perspective, Second Edition, Prentice-Hall, Upper Saddle River, NJ, 2003.

[4] J. Daemen, V. Rijmen: AES Proposal: Rijndael, Document Version 2, 1999.

[5] T.S. Messerges, E.A. Dabbish, and R.H. Sloan. Examining Smart-Card Security under the Threat of Power Analysis Attacks. IEEE Transactions on Computers, 51(5), 2002.

[6] E. Brier, C.Clavier, F.Oliver: Correlation Power Analysis with a Leakage Model, In proceedings of CHES 2004, LNCS 3156, pp. 16-29.

[7] F.X. Standaert, S. B. Ors, J.J. Quisquater and B. Preneel Power analysis attacks against FPGA implementations of the DES. In Field Programmable Logic and Application. Heidelberg, Germany: Springer-Verlag, 2004, vol. 3203, Lecture Notes. in Computer Science, pp. 84–94.

[8] S.B.Ors, F.Gurkaynak, E. Oswald, B. Preneel. Power-Analysis Attack on an ASIC AES implementation. In the proceedings of ITCC 2004, Las Vegas, April 5-7 2004.

[9] Jason Waddle and David Wagner. Towards Efficient Second-Order Power Analysis. In Cryptographic Hardware and Embedded Systems–CHES 2004, 6th International Workshop, Cambridge, MA, USA, August 11-13, 2004, Proceedings, volume 3156 of Lecture Notes in Computer Science, pages 1–15. Springer, 2004.

[10] Suresh Chari, Josyula R. Rao, and Pankaj Rohatgi. Template Attacks. Proceedings of CHES 2002, volume 2535 of LNCS, pages 13-28. Springer, 2003.

[11] Http://www.opencores.org.

[12] J. Wolkerstorfer, E. Oswald, and M. Lamberger, An ASIC Implementation of the AES S-boxes, The Cryptographer's Track at the RSA Conference, CT-RSA 2002, LNCS 2271, pp. 67-78, 2002.

I. J. Communications, Network and System Sciences. **2008; 1: 1-103**
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

QoS in Node-Disjoint Routing for Ad Hoc Networks

Luo LIU[1], Laurie CUTHBERT[2]
Department of Electronic Engineering, Queen Mary, University of London, Mile End Road, E1 4NS, UK
E-mail: [1] luo.liu@elec.qmul.ac.uk

Abstract

Ad hoc network (MANET) is a collection of mobile nodes that can communicate with each other without using any fixed infrastructure. To support multimedia applications such as video and voice MANETs require an efficient routing protocol and quality of service (QoS) mechanism. Node-Disjoint Multipath Routing Protocol (NDMR) is a practical protocol in MANETs: it reduces routing overhead dramatically and achieves multiple node-disjoint routing paths. QoS support in MANETs is an important issue as best-effort routing is not efficient for supporting multimedia applications. This paper presents a novel adaptation of NDMR, QoS enabled NDMR, which introduces agent-based SLA management. This enhancement allows for the intelligent selection of node-disjoint routes based on network conditions, thus fulfilling the QoS requirements of Service Level Agreements (SLAs).

Keywords: MANET, Node-Disjoint, Multipath, Agent-based SLA Management

1. Introduction

Mobile ad hoc networks are infrastructureless networks that can be rapidly deployed. They are characterized by multihop wireless connectivity, frequently changing network topology and the need for efficient dynamic routing protocols [1]. There are no static nodes such as base stations in the network. Each mobile node operates not only as a host but also as a router, forwarding packets to other mobile nodes in the network that may not be within direct wireless transmission range of each other. The design of efficient and reliable routing protocols in such a network is a challenging issue. On-demand routing protocols are widely used because they use much lower routing load than proactive protocols [2]. Ad Hoc on-demand Distance Vector (AODV) [3] and Dynamic Source Routing (DSR) [4] are the two most widely studied on-demand ad hoc routing protocols. The limitation of both of them is they build and rely on a unipath route for each data transmission. Whenever there is a link break on the route, both of the two protocols need to initiate a new route discovery process. This results in a high routing load. On-demand multipath routing protocols can alleviate these problems by establishing multiple routes between the source node and destination node during one route discovery process. A new route discovery is initiated only when all the paths fail or only one path is available. This paper presents an approach built on the Node-Disjoint Multipath Routing Protocol (NDMR) [5]. NDMR has two novel aspects compared to the other on-demand multipath protocols: it reduces routing overhead dramatically and achieves multiple node-disjoint routing paths [5].

Because of the rising popularity of multimedia applications in the commercial environment and the ever-growing requirements of mission-critical applications in the military arena, a best-effort service cannot meet all requirements in most situations. QoS support in mobile ad hoc networks has become an important area of research. Compared to the demands of traditional data-only applications, these new requirements generally include high bandwidth availability, high packet delivery ratio and low delay rate.

Software agents have been demonstrated to provide effective QoS support in networks [6, 7]. The main rationale for using intelligent agents in ad hoc networks is to give greater autonomy to the mobile nodes (since they act as router as well as host). That autonomy, plus the flexibility associate with agents [8] allows the system to meet different QoS requirements as network conditions, eg traffic load [9], change.

2. Node-disjoint multipath routing protocol (NDMR)

Node-disjoint multipath routing protocol (NDMR) is a new protocol developed by Xuefei Li [5], modifying and extending AODV to enable the path accumulation feature of DSR in route request packets. It can efficiently

discovery multiple paths between source and destination nodes with low broadcast redundancy and minimal routing latency.

In the route discovery process, the source creates a route request packet (RREQ) containing message type, source address, current sequence number of source, destination address, the broadcast ID and route path. Then the source node broadcasts the packet to its neighbouring nodes. The broadcast ID is incremented every time that the source node initiates a RREQ, forming a unique identifier with the source node address for the RREQ.

Finding node-disjoint multiple paths with low overhead is not straightforward when the network topology changes dynamically. NDMR routing computation has three key features that help it to achieve low broadcast redundancy and avoid introducing a broadcast flood in MANETs: Path accumulation, Decreasing multipath broadcast routing packets (using shortest routing hops), Selecting node-disjoint paths.

In NDMR, AODV is modified to include path accumulation in RREQ packets. When the packets are broadcast in the network, each intermediate node appends its own address to the RREQ packet. When a RREQ packet finally arrives at its destination, the destination is responsible for judging whether or not the route path is a node-disjoint path. If it is a node-disjoint path, the destination will create a route reply packet (RREP) which contains the node list of whole route path and unicasts it back to the source that generated the RREQ packet along the reverse route path. When an intermediate node receives a RREP packet, it updates the routing table and reverse routing table using the node list of the whole route path contained in the RREP packet.

When receiving a duplicate RREQ, the possibility of finding node-disjoint multiple paths is zero if it is dropped, for it may come from another path. But if all of the duplicate RREQ packets are broadcast, this will generate a broadcast storm and dramatically decrease the performance. In order to avoid this problem, a novel approach is introduced in NDMR recording the shortest routing hops to keep loop-free paths and decrease routing broadcast overhead. When a node receives a RREQ packet for the first time, it checks the node list of the route path calculates the number of hops from the source node to itself and records the number as the shortest number of hops in its reverse routing table. If the node receives a duplicate RREQ packet again, it computes the number of hops and compares it with the shortest number of hops in its reverse routing table. If the number of hops is larger than the shortest number of hops in the reverse routing table, the RREQ packet is dropped. Only when it is less than or equal to the shortest number of hops, the node appends its own address to the node list of the route path in a RREQ packet and broadcasts it to neighbouring nodes again.

The destination node is responsible for selecting and recording multiple node-disjoint paths. When receiving the first RREQ packet, the destination records the list of node IDs of the entire route path in its reverse route table and sends a RREP packet along the reverse route path. When the destination receives a duplicate RREQ, it compares the whole node IDs of the entire route path in the RREQ to all of the existing node-disjoint paths in its reverse routing table. If there is no common node (excepting the source and destination node) between the node IDs from the RREQ and node IDs of any node-disjoint path in the destination's reverse table, the route path in current RREQ is a node-disjoint path and is recorded in the reverse routing table of the destination. Otherwise, the current RREQ is discarded.

3. Quality of service (QoS) in NDMR

Differentiated Services (DiffServ) is a standard approach to achieve QoS in any IP network and could potentially be used to provide QoS in MANETs. DiffServ provides QoS by dividing traffic into a small number of classes and allocating network resources on a per-class basis. The class is marked directly on the packet, in the 6 bit DiffServ Code Point (DSCP) field. The DSCP field is part of the original type of service (ToS) field in the IP header. The IETF redefine the meaning of the little-used ToS field, splitting it into the 6-bit DSCP field and a 2- bit unused field. The unused field is being allocated to the Explicit Congestion Notification (ECN) mechanisms, as shown in Figure 1.

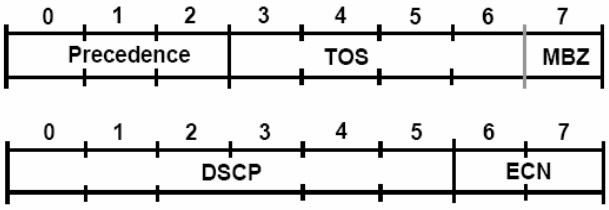


Figure 1. DSCP and ECN

The basic goal of the Differentiated Services architecture is to meet the performance requirements of the users. It classifies traffic into different priority levels and applies priority scheduling and queuing management mechanisms to obtain QoS support.

DiffServ is a fully distributed and stateless model. No state information is required to be maintained at any node. The model aims at pushing the complexity to the edge nodes of the network so that the process in intermediate nodes can be as simple and fast as possible. Instead of providing QoS at per flow granularity, DiffServ differentiates the traffic into a fixed number of classes.

The notion of QoS is a guarantee by the network to satisfy some predetermined service performance constraints for the users in terms of the end-to-end delay,

available bandwidth, probability of packet loss, and so on [10]. Future ad hoc mobile networks will carry increasing levels of diverse multimedia applications such as voice, video and data. This has resulted in an increasing focus on guaranteeing the QoS for such eg delay sensitive applications such as voice, as specified to the customer in a Service Level Agreement (SLA). This section introduces a novel approach to QoS in MANETS: QoS enhanced NDMR, which combines the advantages of NDMR and DiffServ and makes it suitable for the environment of MANETs to support end-to-end QoS solutions

In NDMR, after deciding a path is a node-disjoint path, the destination will create a route reply packet (RREP) which contains the node list of whole route path and unicasts it back to the source. However, since an RREP only currently contains the route path, it cannot provide effective QoS support for MANETs. It is proposed that RREP packets should carry more information such as delay time (queue length) in order to meet certain SLA requirements. When each intermediate node receives a RREP packet, it adds the queue length of this node to the "queue_length" field in RREP packet. Thus, when the source node receives the RREP from the destination node it knows the exact queue length along the path.

Each source keeps three node-disjoint paths for a particular destination. With the "queue_length" field in RREP packet, it chooses the path with the minimum queue length. This allows it to minimise the delay time thus providing higher QoS.

Figure 2 shows queue length along the multiple paths. Assume source node ***S*** first receives the RREP from route 2 (**R2**) (***s-a-b-d***). In standard NDMR, ***S*** will always transmit data on that route so long as no link break happens, even though route 3 (**R3**) (***s-g-h-i-d***) has a smaller queue length and hence a lower rate of delay. With the introduction of the "queue_length" field in RREP, ***S*** will initially choose route 2 (**R2**) (***s-a-b-d***) to transmit data as it receives an RREP from that route fqairst. But after receiving the RREP from route 3 (**R3**) (***s-g-h-i-d***), it will compare the queue length of the existing routes, then change to route 3 (**R3**) (***s-g-h-i-d***) to continue transmitting data. Using this approach, it can reduce the transmission delay rate and meet the SLA requirements.

As an RREP is generated only in the route discovery process, the protocol currently cannot frequently refresh the queue length of each path. As part of the enhancement to NDMR, the need for a similar packet, RUP, route update packet, containing the "queue_length" field used in an RREP packet that performs more frequent updates has been identified. The destination node will periodically unicast RUP packet containing up-to-date queue length to the source node. The source will be able to choose the best path according to the change of queue length in real time.

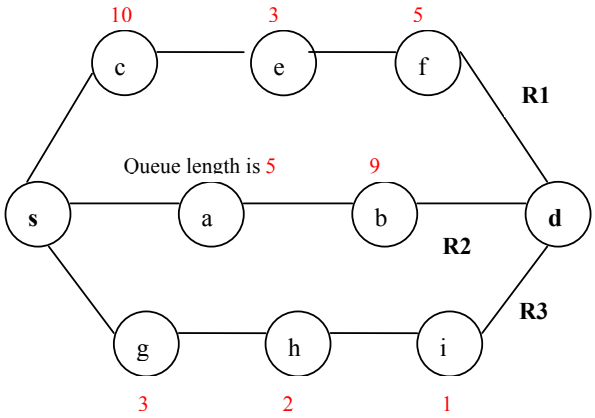


Figure 2. Queue length in multiple node-disjoint paths

Figure 3 shows that the simulation results by OPNET of packet average delay for QoS enabled NDMR give better performance than that of NDMR. The delay time for all mobile velocities tends to be equal. The reason is that with RREP packets carrying real-time delay time back and RUP, the data packets will always be transmitted along the lowest congestion path.

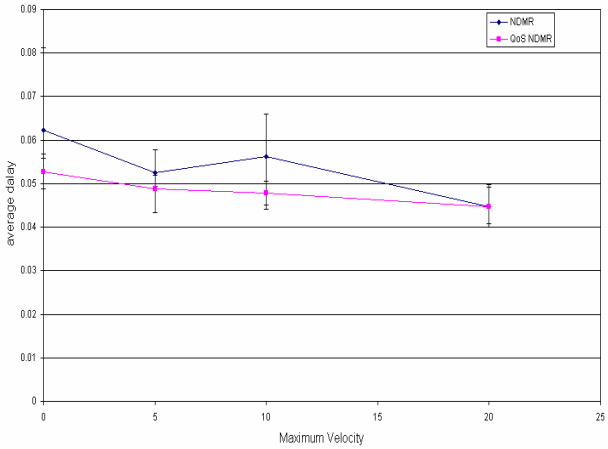


Figure 3. Average delay – CBR

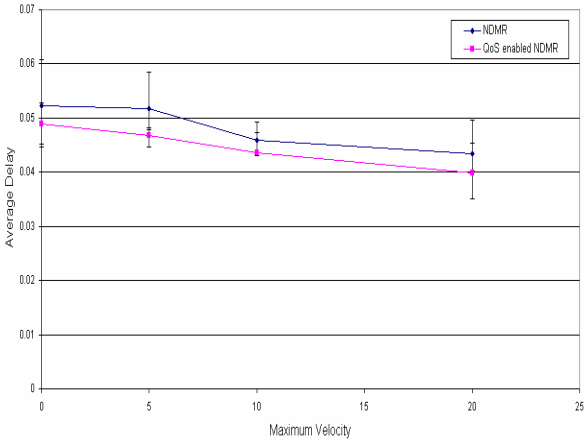


Figure 4. Average delay - exponential source

It can be seen from Figure 4 that QoS enabled NDMR gives better performance when the source generates packets exponentially as well as in constant bit rate (CBR).

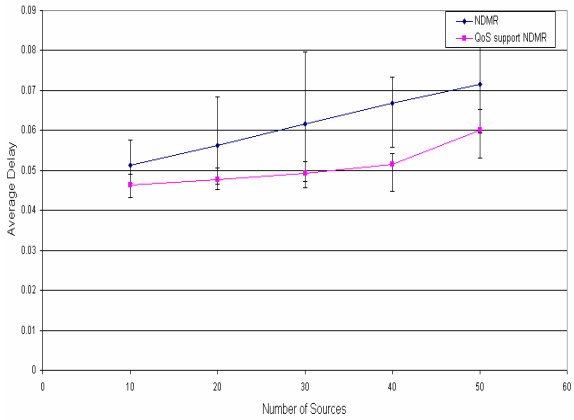


Figure 5. Average delay -different source, CBR

Figure 5 shows the average delay of different number of sources that generate packets. QoS enabled NDMR keep the packet delay time lower as well.

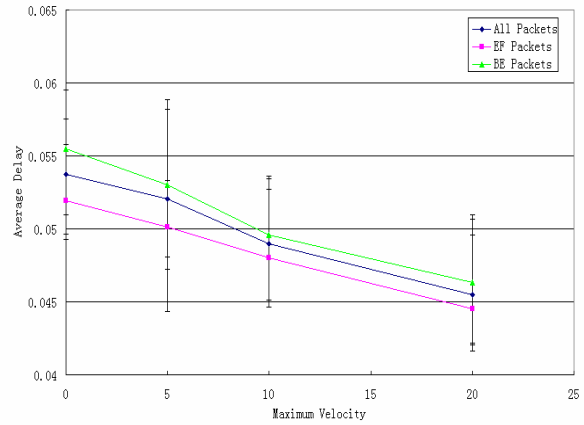


Figure 6. Average delay - different priority

It is necessary to let the Expedited forwarding (EF) traffic which often requires low packet delay time transmit on lower delay time path and Best effort (BE) traffic transmit on other node-disjoint paths. With QoS enabled NDMR, source is able to choose the best path for EF traffic. Figure 6 shows the average delay time of different priority traffic. EF traffic gets the lower delay than BE traffic.

4. Agent-based SLA management

An agent is a computational entity, such as a software program or a robot [11]. It acts upon its environment and is autonomous in that the behavior depends on its own experience.

There are four main properties according to [12]: autonomy, reactiveness, proactiveness and social ability. *Autonomous* means an agent must be able to work without direct command from a programmer or user. *Reactiveness* means agents are capable of controlling the environment and can efficiently move from current situation to the goals. *Proactiveness* means agents can instigate actions to move towards the goals. *Social ability* means agents can communicate with other agents directly and act on information from other agents to make their own decision.

The main reason to use intelligent agents is to give greater autonomy to the mobile nodes. The autonomy increases the flexibility to deal with new situations to acquire QoS to meet different SLAs.

A major feature of the QoS enabled NDMR proposed in this paper is the application of intelligent software agents for SLA management. Employing intelligent agents provides greater autonomy to the mobile nodes, allowing for the essential flexibility to respond to the dynamic nature MANETs. This flexibility will allow the system to meet the QoS requirements agreed in SLAs.

The delay time for each path is calculated from queue length or buffer length. These two parameters are very important in queue management and should be taken into consideration to meet the requirements of any SLAs. A technique to keep the queue length short in a long buffer is necessary. It is proposed that this technique be under the control of an intelligent agent.

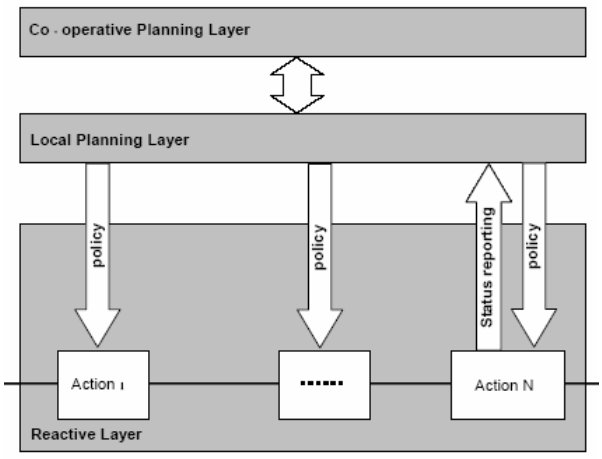


Figure 7. General agent structure [6]

The general agent structure is shown in Figure 7. It uses three layers – reactive, local planning and co-operative – allowing it to take action and make decisions in different timescales. The reactive layer is designed for quick response in real-time. More complex and slower acting functions are implemented in the two planning layers. Generally the local planning layer is concerned with long-term actions within its own node and the co-operative planning layer is concerned with long-term

actions with other agents. Future work will develop the communication and cooperation with agents in other nodes.

In QoS enhanced NDMR, the co-operative planning layer is used for deciding whether to change path or not (according to the queue length of this node and other nodes); the local planning layer is for choosing which path to transmit data (according to the queue length of this node). As illustrated in Figure 8, the packet is transmitted on the reactive layer and the parameters critical to decision making (such as queue length) are passed up to the planning layers. After calculating delay and choosing the appropriate path, the packet will be routed out along this path.

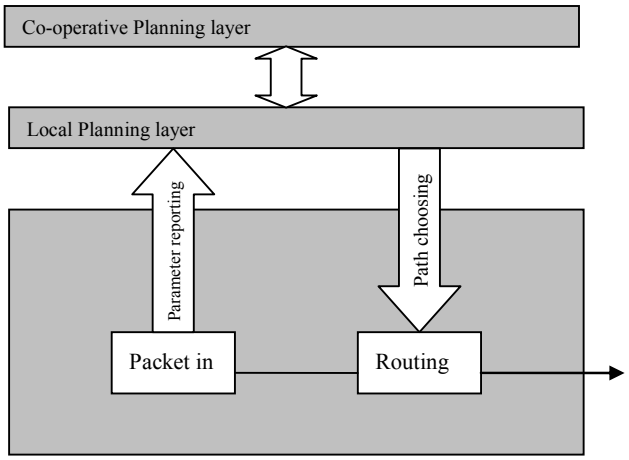


Figure 8. NDMR agent structure

5. Conclusion

This paper has presented architecture for guaranteeing QoS based on Node-Disjoint Multipath Routing Protocol (NDMR) in mobile ad hoc networks. The issue of QoS provision is highly challenging for MANETs and necessarily different from traditional fixed networks. Due to the growth in demand for diverse multimedia applications, fulfilling the QoS guarantees in SLAs requires solutions that are responsive to network state. The use of multiple node-disjoint paths gives the opportunity for allocating packets to paths in an optimum way to meet instantaneous constraints. This paper has compared performance in different situation of NDMR and found a means of developing NDMR – through the queue length field and additional route update packets – to allow for QoS measurement along such node-disjoint paths. By using intelligent agents it will be possible to distribute this optimisation at the planning layer, thus allowing very fast processing to occur at the reactive layer while still taking into account the needs of all nodes.

6. References

[1] Charles E. Perkings, Elizabeth M.Royer and Samir R.Das, Performance Comparison of Two On-Demand Routing Protocols for Ad Hoc Networks, IEEE Personal Communications, Feb 2001.

[2] Z.J. Haas and S. Tabrizi, "On Some Challenges, and Design Choices in Ad hoc Communcations", Proceedings of the IEEE Military Communications Conference (MILCOM), Bedford, MA, October 1998, pp.187-192.

[3] Charles E. Perkings, Elizabeth M. Belding-Royer, Samir R.Das, Ad Hoc On-Demand Distance Vector (AODV)Routing, http://www.ietf.org/internetdrafts/draft-ietf-manet-aodv-13.txt, IETF Internet draft, Feb 2003

[4] J.Broch, D.Johnson, and D. Maltz, The Dynamic Source Protocol for MobileAd hoc Networks, http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-10.txt, IETF Internet draft, 19 July 2004.

[5] Xuefei Li and Laurie Cuthbert, On-demand Node-Disjoint Multipath Routing in Wireless Ad hoc Networks, In Proceedings of the 29th Annual IEEE Conference on Local Computer Networks, LCN 2004, Tampa, Florida, U.S.A., November 16-18, 2004

[6] L.G. Cuthbert, D. Ryan, L. Tokarchuk, J. Bigam and E. Bodanese, Using intelligent agents to manage resource in 3G Networks, Journal of IBTE, 2(4), 2001

[7] Alex Hayzelden & John Bigham Heterogeneous Multi-Agent Architecture for ATM Virtual Path Network Resource Configuration, in Intelligent Agents for Telecommunications Applications (IATA '98), LANAI 1437, S.Albayrak & F.J.Garijo (eds), Springer-Verlag, 1998, ISBN 3-540-64720-1

[8] M.Wooldridge & N.R. Jennings, Agent Theories, Architectures and Languages: a Survey in Wooldridge & Jennings eds. Intelligent Agents, Springer-Verlag, Berlin, 1995

[9] Soamsiri Chantaraskul, An Intelligent-Agent Approach for Managing Congestion in W-CDMA Networks, PhD thesis, University of London, August 2005

[10] S. Chakrabarti, A. Mishra, QoS Issues in Ad hoc Wireless Networks, IEEE Communications Magazine, February 2001.

[11] G. Weiss, Multiagent System: A Modern Approach to Distributed Artificial Intelligence, the MIT Press, Cambridge, Massachusetts, London, England, 1999.

[12] M. Luck, R. Ashri and M. d'Inverno,Agent-Based Software Development, Artech House, Inc., 2004.

I. J. Communications, Network and System Sciences. 2008; 1: 1-103
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

A Predicted Region based Cache Replacement Policy for Location Dependent Data in Mobile Environment

Ajey KUMAR[1], Manoj MISRA[2], Anil K. SARJE[2]
Department of Electronics and Computer Engineering,
Indian Institute of Technology Roorkee, Roorkee, India
E-mail: [1]ajeykumar7@gmail.com, [2]{manojfec,sarjefec}@iitr.ernet.in

Abstract

Caching frequently accessed data items on the mobile client is an effective technique to improve the system performance in mobile environment. Proper choice of cache replacement technique to find a suitable subset of items for eviction from cache is very important because of limited cache size. Available policies do not take into account the movement patterns of the client. In this paper, we propose a new cache replacement policy for location dependent data in mobile environment. The proposed policy uses a predicted region based cost function to select an item for eviction from cache. The policy selects the predicted region based on client's movement and uses it to calculate the data distance of an item. This makes the policy adaptive to client's movement pattern unlike earlier policies that consider the directional / non-directional data distance only. We call our policy the Prioritized Predicted Region based Cache Replacement Policy (PPRRP). Simulation results show that the proposed policy significantly improves the system performance in comparison to previous schemes in terms of cache hit ratio.

Keywords: Mobile Computing, Cache Replacement, Location Dependent Data, Valid Scope, Location Dependent Information Services.

1. Introduction

Recent advances in wireless technology have ushered the new paradigm of mobile computing. With the advent of new mobile infrastructures providing higher bandwidth and constant connection to the network from virtually everywhere and advances in the global positioning technologies , a new class of services referred to as Location Dependent Information Services (LDIS) has evolved and is gaining popularity among mobile users.

LDIS provide users with the ability to access information related to their current location. By including location as a part of user's context information, service carriers can provide better services to many value-added applications such as travel and tourist information system, assistance and emergency system, nearest object searching system and local information access system, which specifically target the mobile users. Hence, the need for LDIS arises frequently. For example, imagine you are on a business trip in a foreign city and you do not know the city very well, you have no idea where to go. In this situation, with the help of your portable device you can query your personal interest like nearest restaurant, nearest gas station, nearest ATM, nearest theater etc. and can get the response on your device.

There are many challenges in providing LDIS services to users. These challenges include limited bandwidth, limited client power and intermittent connectivity [1][2][4][6]. Caching helps to address some of these challenges. Caching of frequently accessed data item on client side is an effective technique to improve data accessibility and to reduce access cost. However, due to limitations of cache size on mobile devices, it is impossible to hold all accessed data items in the cache. Thus, there is a need of efficient cache replacement algorithms to find out suitable subset of data items for eviction from the cache. Good cache performance heavily depends on these replacement algorithms. Also, for wide area mobile environments due to its distributed nature, the design of an efficient cache replacement policy becomes very crucial and challenging to ensure good cache performance.

Most of the existing cache replacement policies use cost functions to incorporate different factors including access frequency, update rate, size of objects etc. Temporal-based traditional cache replacement strategies,

such as Least Recently Used (LRU), Least Frequently Used (LFU) and LRU-k [13] have been studied widely in the past. These polices are based on the assumption that client's access patterns exhibit temporal locality (i.e. the objects that were queried frequently in the past will continue to be queried frequently in the future). These algorithms replace data items based on recency or frequency of access. LRU is the most commonly used recency based algorithm. LFU, which maintains a reference count for each cached objects, is the most commonly used frequency based algorithm. Frequency-based algorithms are well suited for skewed access patterns in which a large fraction of accesses go to a disproportionately small set of hot objects. Frequency and recency based algorithms form the two ends of a spectrum of caching algorithms. LRU-k cache replacement algorithm tries to balance both recency and frequency. However, in mobile networks where clients utilize location dependent information services, clients access pattern do not only exhibit temporal locality, but also exhibit dependence on location of data, location of the client and direction of the client's movement [4][5]. As a result, the aforementioned policies are unsuitable for supporting location dependent services because they do not take into account the location of data objects and the movement of mobile clients. Hence, relying solely on temporal locality when making cache replacement decisions will result in poor cache hit ratio in LDIS. To overcome this problem, several location-aware cache replacement policies [5][7][9][10] have been proposed for location dependent information services.

Manhattan Distance-based cache replacement policy [10] supports location dependent queries in urban environments. Cache replacement decisions are made on the basis of distance between a client's current location and the location of each cached data object. Objects with the highest Manhattan distance from the client's current location are evicted at cache replacement. While the Manhattan based policy accounts for the distance between clients and data objects, the major limitation of this approach is that it ignores the temporal access locality of mobile clients and the direction of client movement while making cache replacement decisions.

Furthest Away Replacement (FAR) [9] policy uses the current location and movement direction of mobile clients to make cache replacement decisions. Cached objects are grouped into two sets, viz., in-direction set and the out-direction set. Data objects in the out-direction set are always evicted first before those in the in-direction set. Objects in each set are evicted in the order based on their distance from the client. Similar to the Manhattan approach, FAR also neglects the temporal properties of clients' access pattern. It also becomes ineffective when mobile clients change direction frequently due to frequent change in the membership of objects between the in-direction and out-direction sets.

In Probability Area Inverse Distance (PAID) [5] policy, the cost function of data item i takes into account the access probabilities (P_i) of data objects, area of its valid scopes $A(vs_i)$ and the distance $D(vs_i)$ between the client's current position and the valid scope of the object concerned (known as data distance). The cost function of PAID is given by $P_i A(vs_i)/D(vs_i)$. It neither takes into account the size of the data object nor does it give priority to the data objects in cache that are near to the mobile client. Mobility Aware Replacement Scheme (MARS) [7] policy is also a cost based policy, which comprises of temporal score, spatial score and cost of retrieving an object. Unlike PAID, it takes into account the updates of data objects. But as far as location dependent data (LDD) is concerned, their update rate (if exist) is negligible as compared to temporal data. Thus, for LDD, only spatial score dominates which consists of area of valid scope, data distance from current client location and data distance from future client location. The impact of client's anticipated location or region in deciding cache replacement still remains unexplored.

None of these cache replacement policies are suitable if client changes its direction of movement quite often. Existing cache replacement policies only consider the data distance (directional/undirectional) but not the distance based on the predicted region or area where the client can be in near future. Very few of these policies [5][7] account for the location and movement of mobile clients.

In this paper, we predict an area in the vicinity of client's current position, and give priority to the cached data items that belong to this area irrespective of the client's movement direction. PPRRP calculates the data item cost on the basis of access probability, valid scope area, data size in cache and data distance based on the predicted region, which has not been considered in any of the existing policies.

The rest of the paper is organized as follows. Section 2 briefly describes mobile system model used in our work. Section 3 details the proposed new cost based replacement policy PPRRP. Section 4 and section 5 deal with simulation model, and performance evaluation and comparison simultaneously. Section 6 concludes the paper.

2. Mobile System Model

We assume a cellular mobile network that is similar to the model discussed by D. Barbara [1] as mobile computing infrastructure. A mobile system [1][4][5][6] is usually made up of a server, moving clients, and a wireless connection between them (see Figure 1). The geographical area is divided into small regions, called cells. Each cell has a B*ase Station* (BS) or *Mobile Support Station (MSS)* augmented with wireless interfaces

and a number of *Mobile Clients* (MCs). Inter-cell and intra-cell communications are managed by the MSSs. The MCs communicate with the MSS by wireless links within its radio coverage area. An MC can move freely from one location to another within a cell or between cells while retaining its network connection. An MC can either connect to a MSS through a wireless communication channel or disconnect from the MSS by operating in the *doze* (power save) mode. The MC queries the database servers that are connected to a wired network. The wireless channel is logically separated into two sub channels: *uplink channel* and d*ownlink channel*. The uplink channel is used by MCs to submit queries to the server via an MSS, while the downlink channel is used by MSSs to disseminate information or to forward the answers from the server to the target client.

The mobile computing platform can be effectively described under the *client/server* paradigm [19]. A data item is the basic unit for update and query. MCs only issue simple requests to read the most recent copy of a data item. There may be one or more processes running on an MC. These processes are referred to as clients (we use the terms MC and client/users interchangeably). In order to serve a request from a client, the MSS needs to communicate with the database server to retrieve the data items. Since the communication between the MSS and the database server is through wired links and is transparent to the clients (i.e., from the client's point of view, the MSS is the same as the database server), we also use the terms MSS and server interchangeably.

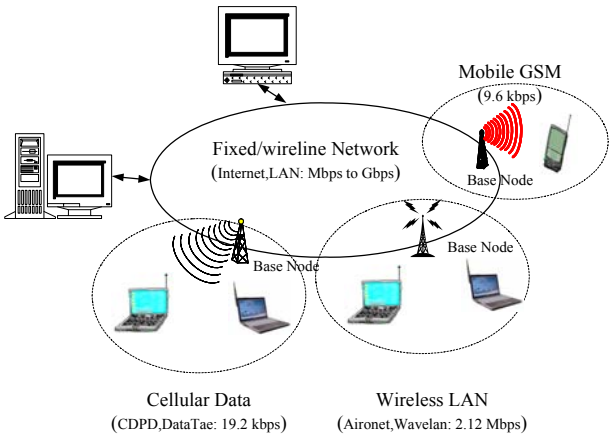
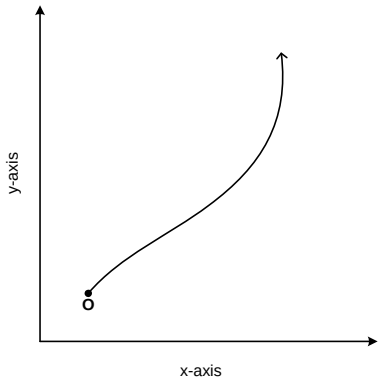
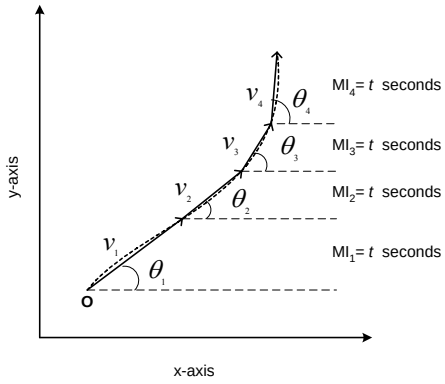


Figure 1. Mobile computing system model

The information system provides location dependent services to mobile clients. The geographical area covered by the information system is referred as the service area. In this paper, we assume a geometric location model, i.e., a location is specified as a two-dimensional coordinate. However, it can be easily extended to 3-dimension space by including the third dimension. Mobile clients can identify their locations using systems such as the Global Positioning System (GPS) [3]. The data item value is different from data item, i.e., data item value for a data item is an instance of the item valid for a certain geographical region. Moreover, the data item value is different from data item. Data item value for a data item is an instance of the item valid for a certain geographical region. So, a data item can show different values when clients at different locations query it. For example, "restaurant" is a data item, and the data values for this data item vary depending on the location of query i.e. point at which the query " *Tell me the nearest restaurant*" was issued by a mobile client. The valid scope of a data item value is defined as the region within which the data item value is valid. In a two-dimensional space, a valid scope (*vs*) can be represented by a geometric polygon $p(e_1, \ldots, e_n)$, where e_i 's are endpoints of the polygon. A mobile client can cache data on its local disk or in any storage system that survives power-off. In this paper, data values are assumed to be of fixed size and read–only so that we can omit the influence of data sizes and updates on cache performance and concentrate on the impact caused by the unique properties of location-dependent data.



(a) Abstract model



(b) Discrete model

Figure 2. Client's movement path

We also assume an unconstrained network, where mobile clients move freely inside the geographical region covered by the mobile network (without any restrictions).In the abstract model, the path of a moving client is represented by a curve in 2-dimension (x-y plane), as shown in Figure 2(a). Though abstract model is

simple, from computer implementation point of view, discrete model is preferred [8]. In discrete model, the path traveled is modeled as a sequence of line segments, each associated with fixed velocity and direction, as shown in Figure 2 (b). Length of the line segment depends on the rate of change of direction and velocity. For random movement this duration between change in direction and velocity is small and for regular movement and highway users this duration is large. This duration is known as *Moving Interval* (MI) [5][7][17]. Figure 2(b) shows the discrete movement of a mobile user with MI of t seconds. The distance between any two locations or points is the length of a straight line connecting the two points (i.e. Euclidean distance).

3. Prioritized Predicted Region Based Cache Replacement Policy (PPRRP)

3.1. Motivation

LDIS, being spatial in nature needs that the distance of data item from client's current position and its valid scope area should also be taken into account for cache replacement. Greater the distance of valid scope of data from the user's current position lower is the chance that client will enter in the valid scope area of the data in near future. Thus, it is better to eject the farthest data value when replacement takes place. Moreover, because the client is mobile, its position at the time of next query will be different from its current position. Therefore the client's movement should also be taken into account. Locations in the opposite direction of client's movement have very low chance of being revisited, though they may be very close to it. Based on this reasoning, existing cache replacement schemes like FAR and PAID (directional) assign higher priorities to data items in the client's direction of movement giving very low priority to the items in the opposite direction of user's movement. However, with random movement patterns of clients, it is not always necessary that client will continue moving in the same direction. Therefore, evicting data values which are in the opposite direction of client's movement but are very close to client's current position may degrade the overall performance.

3.2. Basic Idea

When client movement pattern is random, retaining the data items in the direction of user movement and discarding the data items that are in the opposite direction of user movement may not improve the performance. Therefore, our cache replacement policy considers the predicted region of user presence in near future (rather than considering the direction of user movement only) while selecting an item for replacement. The predicted region is based on the client's current movement pattern. We show that it is useful to calculate the data distance with respect to this region so that the data items in the vicinity of client's current position are not purged from cache. Valid scope area of the data item and the amount of space required to store the data item in cache are also used to select an item for replacement. This is because the client has higher chance of being in large region rather than small regions and keeping smaller size data items in cache helps to accommodate a large number of data items in the cache. Hence, our cache replacement policy selects a victim with low access probability, small valid scope area falling outside the predicted region and large data size.

3.3. Approach

We make use of discrete model for client's movement path as described in Section 3.2 and used in [5][7][14]. Assuming a predefined path of mobile user or a predefined destination is generally not possible unless we are dealing with a case where the user is moving in a train or a ship and the entire path of the user is known well in advance. For discrete model, the direction and velocity of the user are known for current MI. At the end of each MI, direction is selected randomly between 0° to 360° degrees and the velocity between minimum (v_{min}) and maximum speed (v_{max}) of the client. This motivates us in predicting a region instead of predicting the path.

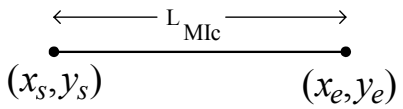


Figure 3. Current moving interval

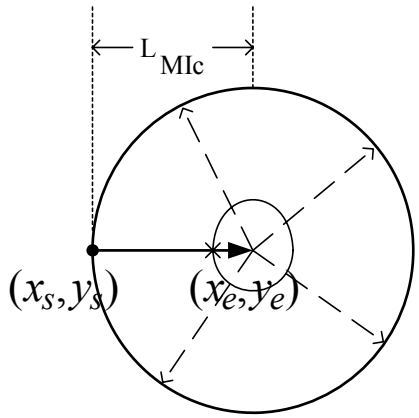


Figure 4. Predicted region

Let v_c be the velocity in current moving interval MI_c, L_{MIc} be the length of MI_c along direction θ_c and (x_s, y_s) and (x_e, y_e) be the starting and end point of MI_c respectively (see Figure 3). Assuming that the velocity v_c remains same (generally the mobile user does not changes its velocity significantly over a long period) in the next MI also, we can predict the region of user presence in near future by the circle with radius L_{MIc} and centre (x_e, y_e) as shown in Figure 4. Our cache replacement policy uses

this region to calculate the distance of data items in cache as follows:

- The distance of data items outside the predicted region is calculated from the centre of the circle
- The distance of data items inside the predicted region is calculated as the minimum of {L_{MIc}, distance of the valid scope from the current position of the user}.

Calculating the distance of data items in this way ensures that

- Items outside the predicted region always have the lower priority than the items inside the predicted region.
- Items inside the predicted region, close to the user have higher priority.

One of the advantages of using predicted region is that it dynamically changes with speed of client and MI and also takes into account the random movement of client.

Now, we define cost function for our cache replacement policy PPRRP that considers access probability, predicted region based data distance, valid scope area and size of the data in cache. Associated with each cached data object is the replacement cost. When a new data object needs to be cached and there is insufficient cache space, the object(s) with lowest replacement cost is (are) removed until there is enough space to cache new object. The cost of replacing a data value *j* of data item *i* in client's cache is calculated as:

$$Cost_{i,j} = \begin{cases} \frac{P_i . A(vs'_{i,j})}{S_{i,j}} . \frac{1}{minimum\left\{L_{MIc}, D(vs'_{i,j})\right\}} & if \quad vs'_{i,j} \in Pred\_Reg \\ \frac{P_i . A(vs'_{i,j})}{S_{i,j}} . \frac{1}{D'(vs'_{i,j})} & if \quad vs'_{i,j} \notin Pred\_Reg \end{cases} \quad \textbf{(1)}$$

where P_i is the access probability of data item i , $A(vs'_{i,j})$ is the area of the valid scope $vs'_{i,j}$ for data value j, $S_{i,j}$ is the size of data value j and valid scope $vs'_{i,j}$, $D(vs'_{i,j})$ is the distance of the valid scope $vs'_{i,j}$ from the current user position , $D'(vs'_{i,j})$ is the distance of the valid scope $vs'_{i,j}$ from the centre of the predicted region and *Pred_Reg* is the predicted region.

4. Simulation Model

This section describes the simulation model used to evaluate the performance of the proposed location-dependent cache invalidation methods. Our Simulator is implemented in C++ and setup is similar and in accordance with those used in earlier studies [5][14].

4.1. System

Since seamless hand-off from one cell to another is assumed, the network can be considered a single, large service area within which the clients can move freely and obtain location-dependent information services. In our simulation, the service area is represented by a rectangle with a fixed size of *Size*. We assume a "wrapped-around" [5][14][107] model for the service area. In other words, when a client leaves one border of the service area, it enters the service area from the opposite border at the same velocity.

The database contains *ItemNum* items. Every item may display *ScopeNum* different values for different client locations within the service area. The size of data value varies from S_{min} to S_{max} and has following three types of distributions [6]:

- **IncreasingSize:** The size S_i of data item i grows linearly as i increases, and is given by:

$$S_i = S_{\min} + \frac{(i-1)*(S_{\max} - S_{\min})}{ItemNum - 1}, i = 1, ..., ItemNum; \quad \textbf{(2)}$$

- **DecreasingSize:** The size S_i of data item i decreases linearly as i increases, and is given by:

$$S_i = S_{\max} - \frac{(i-1)*(S_{\max} - S_{\min})}{ItemNum - 1}, i = 1, ..., ItemNum; \quad \textbf{(3)}$$

- **RandomSize:** The size S_i of data item i falls randomly between S_{min} and S_{max}, given by:

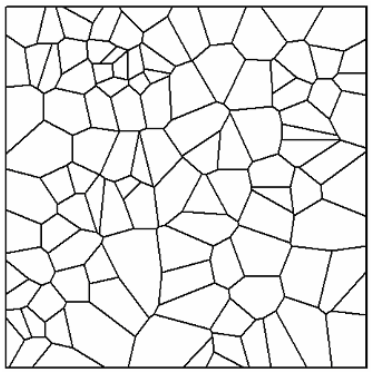
$$S_i = S_{\min} + \lfloor prob() * (S_{\max} - S_{\min}) \rfloor, i = 1, ..., ItemNum; \quad \textbf{(4)}$$

where, $prob()$ is a random function with uniformly distributed value between 0 and 1. The choice of the size distributions are based on previously published trace analysis [6]. Though, some researchers have shown that small data items are accessed more frequently than large data items, but recent web trace analysis shows that the correlation between data item size and access frequency is weak and can be ignored [16]. Combined with the skewed access pattern, IncreasingSize and DecreasingSize represent client's preference for frequently querying smaller items and larger items respectively. In other words, with IncreasingSize setting, the clients access the smallest item most frequently and with DecreasingSize setting, the clients access the largest item most frequently. RandomSize, models the case where no correlation between the access pattern and data size exists.

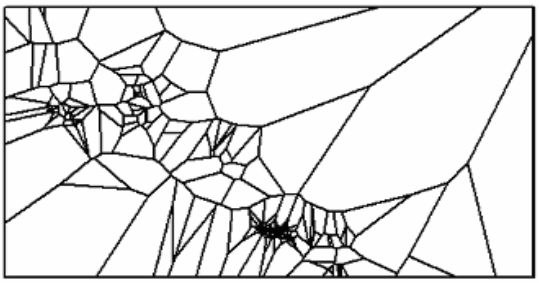
In the simulation, the scope distributions of the data items are generated based on voronoi diagrams (*VD*s) [12][20] because valid scopes of nearest neighbor queries is defined by VD. Formally, given sets of point

$O=\{o_1,o_2,.....,o_n\}$, $V(o_i)$, the Voronoi Cell (*VC*) for o_i, is defined as the set of points q in the space such that $dist(q,o_i) < dist(q,o_j)$, $\forall\ j \neq i$. That is, $V(o_i)$ consists of set of points for which o_i is nearest neighbor. In our simulation, first data set Scope Distribution 1 (Figure 5 (a)) contains 110 points randomly distributed in a square Euclidean space. The second data set, Scope Distribution 2 (Figure 5 (b)), contains the locations of 215 hospitals in California area, which is extracted from the point data set available at [18].

This model assumes that two floating-point numbers are used to represent a two-dimensional coordinate and one floating-point number to represent the radius of circle. The size of a floating-point number is *FloatSize*. The wireless network is modeled by an uplink channel and a downlink channel. The uplink channel is used by clients to submit queries, and the downlink channel is used by the server to return query responses to target clients. The communication between the server and a client makes use of a point-to-point connection. It is assumed that the available bandwidth is *UplinkBand* for the uplink channel and *DownlinkBand* for the downlink channel.



(a) Scope distribution 1 (*ScopeNum*=110)



(b) Scope distribution 2 (ScopeNum=215)

Figure 5. Scope distributions for performance evaluation

4.2. Client

The mobile client is modeled with two independent processes: *query process* and *move process*. The *query process* continuously generates location-dependent read-only queries for different data items. After the current query is completed, the client waits for an exponentially distributed time period with a mean of *QueryInterval* before the next query is issued. The client access pattern over different items follows a *Zipf* distribution with skewness parameter θ, which is shown to be a realistic approximation of skewed data access and are frequently used to model non-uniform distribution [5][6][16]. In the Zipf distribution, the access probability of the i^{th} ($1 \leq i \leq N$) data item is represented as follows

$$p_i = \frac{\frac{1}{i^{\theta}}}{\sum_{j=1}^{N} \frac{1}{j^{\theta}}} \qquad \textbf{(5)}$$

where N is the number of items in the database and $0 \leq \theta \leq 1$.

When $\theta = 0$, the access pattern is uniform. As θ value is increased the skewness increases. When $\theta = 1$, it is the strict Zipf distribution. Large θ results in more "skewed" access distribution. To answer a query, the client first checks its local cache. If the data value for the requested item with respect to the current location is available, the query is satisfied locally. Otherwise, the client submits the query and its current location to the server and retrieves the data through the downlink channel. The *move process* controls the movement pattern of the client using the parameter *MovingInterval*. After the client keeps moving at a constant velocity for a time period of *MovingInterval*, it changes velocity for next MI. The next speed is selected randomly between *MinSpeed* and *MaxSpeed*. Similarly, the next moving direction (represented by the angle relative to the x-axis, counter clock wise taken as positive) is selected randomly between 0° to 360°. If the difference between *MinSpeed* and *MaxSpeed* is low the mobile users move with almost same velocity. The client is assumed to have a cache of fixed size, which is a *CacheSizeRatio* ratio of the database size.

4.3. Server

The server is modeled by a single process that services the requests from clients. The requests are buffered at the server if necessary, and an infinite queue buffer is assumed. The FCFS service principle is assumed in the model. To answer a location-dependent query, the server locates the correct data value with respect to the specified location. Since the main concern of this paper is the cost of the wireless link(i.e. transmission time, receiving time and disconnections), which is more expensive than the wired-link and disk I/O costs(i.e. disk access time), the overheads of request processing and service scheduling at the server are assumed to be negligible in the model.

5. Performance Evaluation

This section describes the performance parameters and measures used for simulation and analyze the results of the simulation.

5.1. Performance Parameters

The default values of different parameters used in the simulation experiments are given in Table 1. They are chosen to be the same as used in earlier studies [5][7][14].

Experiments are performed using different workloads and system settings. In order to get the true performance for each algorithm, we collect the result data only after the system becomes stable, which is defined as the time when the client caches are full [5][6]. Each simulation runs for 20,000 client issued queries and each result obtained in the experiment is taken as the average of 10 simulation runs with Confidence Interval of 96 percent.

For simulation purpose, we assume that all data items follow the same scope distribution in a single set of experiments. Two scope distributions with 110 and 215 valid scopes are used (see Figure 5). Since the average valid scope areas differ for these two scope distributions, different moving speeds are assumed, i.e., the pair of (*MinSpeed*,*MaxSpeed*) is set to (1,2) and (5,10) for Scope Distribution 1 and Scope Distribution 2, respectively. The Caching-Efficiency-Based (CEB) [5] cache invalidation policy is employed for cache management. For calculating data distance between valid scope (either a polygon or a circle) and current location we select a reference point for each valid scope and take the distance (Euclidean distance) between the current location and the reference point. For polygonal valid scope, the reference point is defined as the endpoint that is closest to the current location and for circular valid scope, it is defined as the point where the circumference and the line connecting the current location and the center of the circle meet. Access probability for each data item is estimated by using exponential ageing method [5][6]. Two parameters are maintained for each data item *i*: a running probability P_i and the time of the last access to item t^l_i. P_i is initialized to 0. When a new query is issued for data item *i*, P_i is updated using the following formula:

$$P_i = \alpha/(t_c - t^l{}_i) + (1-\alpha)P_i \qquad \textbf{(6)}$$

where, t_c is the current system time and α is a constant factor to weigh the importance of most recent access in the probability estimate. Note that the access probability is maintained for each data item rather than for each data value. If the database size is small, the client can maintain these parameters (i.e., P_i and t^l_i for each item *i*) for all items in its local cache. However, if the database size is large, these parameters will occupy a significant amount of cache space. To alleviate this problem, we set an upper bound to the amount of cache used for storing these parameters (5 percent of the total cache size in our simulation) and use the Least Frequently Used (LFU) policy to manage the limited space reserved for it.

5.2. Performance Metric

Our primary performance metric is *cache hit ratio*. Other performance metrics can be derived from the cache hit ratio. Cache hit ratio can be defined as the ratio of the number of queries answered by the client's cache to the total number of queries generated by the client. Specifically, higher the cache hit ratio, higher is the local data availability, less is the uplink and downlink costs and the battery consumption [5][6][14].

5.3. Comparison of Location-Dependent Cache Replacement Schemes

This subsection examines the performance of different location-dependent cache replacement policies, namely, PPRRP with PAID, FAR and Manhattan. Figures 6 to 15 show the cache hit ratio for both scope distributions (see Figure 5) under various query intervals, moving intervals, cache sizes, client's speed and Zipf's distribution.

5.3.1. Effect of Query Interval

Query interval is the time interval between two consecutive client queries. In this set of experiments, we vary the mean query interval from 20 seconds to 200 seconds. Figures 6 and 7, show cache performance for both scope distributions and for the data distributions: IncreasingSize, RandomSize and DecreasingSize.

Results show that, when the query interval is increased, almost every scheme shows a worse performance. This is because, for a longer query interval when a new query is issued the client has a lower probability of residing in one of the valid scopes of the previously queried data items. When different cache replacement policies are compared, the proposed policy substantially outperforms the existing policies. Figure 6, compares the performance of cache replacement policies versus query interval for Scope Distribution 1. PPRRP, which prefers object within the predicted region over the objects outside the predicted region and gives priority to the data objects that are nearer to the client's current position within the predicted region, obtains better performance than PAID. Average improvement of PPRRP over PAID is 28%, 21% and 19% for IncreasingSize, RandomSize and DecreasingSize respectively for Scope Distribution 1. Figure 7, shows the effect of change in query interval on the performance of cache replacement policies for Scope Distribution 2. It can be observed that the PPRRP show similar gains in performance for Scope Distribution 2 also as they were for Scope Distribution1. The average improvement of PPRRP over PAID for Scope Distribution 2 is 8%, 12.2%, and 11% for IncreasingSize, RandomSize and DecreasingSize respectively .

5.3.2. Effect of Moving Interval

This subsection examines the performance of the replacement policies when the client's moving interval is varied. Longer the moving interval, less frequent is the changes in velocity of the client and hence, there is lesser

randomness in the client's movement. The performance results for IncreasingSize, RandomSize and DecreasingSize of data distribution are shown in Figures 8 and 9.

We can see that when the moving interval is varied from 50 seconds to 400 seconds, the cache hit ratio decreases drastically. The reason for this is as follows. For a relatively longer moving interval, there is a high probability of the client leaving one valid region and entering another. Consequently, the cached data are less likely to be re-used for subsequent queries. This leads to a decreased performance with increase in MI.

Figure 8, compares the performance of cache replacement policies over changing moving interval for Scope Distribution 1. Though for small MI, the randomness in client movement is more as compared to larger MI but PPRRP perform better than all existing policies for both small and large MI. The predicted region in PPRRP helps to keep the data items within the influence of client's movement, thereby reducing the effect of randomness in client's movement. Average improvement of PPRRP over PAID is 27%, 21% and 12% for IncreasingSize, RandomSize and DecreasingSize respectively. Figure 9, compares the performance of cache replacement policies over change in moving interval for Scope Distribution 2. For Scope Distribution 2 also, we get similar improvement in performance of PPRRP as they were for Scope Distribution 1. The average improvement of PPRRP over PAID for Scope Distribution 2 is 7%, 6.3%, and 11% for IncreasingSize, RandomSize and DecreasingSize respectively.

5.3.3. Effect of Cache Size

In this set of experiments, we intend to investigate the robustness of the proposed replacement schemes under various cache sizes. Figures 10 and 11, show the results when CacheSizeRatio is varied from 5% to 20%. As expected, the performance of all replacement schemes improves with increasing cache size. This is because the cache can hold large number of data items which increases the probability of getting a cache hit. Moreover, replacement occurs less frequent in comparison to the case when cache size is low. Figure 10, shows the performance for Scope Distribution 1. PPRRP consistently outperforms the existing policies from small size cache to large size cache. Average improvement of PPRRP over PAID is 25%, 24% and 18% for IncreasingSize, RandomSize and DecreasingSize respectively. Figure 11, compares the performance of cache replacement policies under varied CacheSizeRatio for Scope Distribution 2. Results show similar performance gains for all proposed policies for Scope Distribution 2 also. The average improvement of PPRRP over PAID for Scope Distribution 2 is 6.7%, 9.4%, and 10% for IncreasingSize, RandomSize and DecreasingSize respectively.

5.3.4. Effect of Client's Speed

This subsection examines the effect of change in client's speed on the performance of the proposed cache replacement policies. Client's cache hit ratio is shown against client speed from Figures 12 to 13. Four speed ranges [15], 1~5m/s, 6~10m/s, 16~20m/s, 25~35m/s, corresponding to the speed of a walking human, a running human, a vehicle with moderate speed and a vehicle with high speed, respectively are used. It can be seen that very high cache hit ratio can be achieved for walking human. For higher speed range, the cache hit ratio drops as client spends less time at each geographic location and the valid scope of each data item stored in cache becomes less effective. In PPRRP, higher the speed of client, greater is the predicted region and hence more data items stored in the cache are held in that region. Average improvement of PPRRP over PAID for different speed ranges for Scope Distribution 1 and Scope Distribution 2 are given in Table 2 and Table 3 respectively.

5.3.5. Effect of Client's Access pattern

This subsection examines the performance of the replacement policies under various clients' access pattern. Client's access pattern is modeled by Zipf's Distribution [16]. The Zipf parameter θ determines the "skewness" of the access pattern over data items. When θ=0, the access pattern is uniformly distributed. When θ increases, more access is focused on few items (skewed). Figures 14 and 15, show the impact of access pattern on performance of replacement policies for both scope distributions. As desired, performance of PPRRP along with other replacement policies, increases with increase in θ for both Scope distributions over all the three data size distributions. Moreover, proposed policy shows an edge over other policies.

6. Conclusion

In this paper, we presented a cache replacement policy, PPRRP, for location-dependent data that uses predicted region based cost function for selecting data items to be replaced from the cache. In order to decide which data items to replace from cache, an attempt must be made to predict what items will be accessed in the future.

We emphasized on predicting a region around mobile client's current position apart from considering only user's direction or distance. Predicted region plays an important role in improving the system performance. Using the predicted region of user influence, the data items in the vicinity of client's current position are not purged from cache, which increases the cache hit. Proposed policy takes into account factors like access probability, data distance from predicted region, valid scope and data size in cache. In PPRRP, data distance is calculated such that the data items within the predicted region are given higher priority than the data items

outside the predicted region. In addition to giving highest priority to the data items within the predicted region, data items nearer to the client's current position are also favored over other data items in the same predicted region. A number of simulation experiments have been conducted to evaluate the performance of the PPRRP. The results show that PPRRP, with different system settings, give better performance (improves cache hit ratio) than PAID. PPRRP achieves an average improvement of more than 25% for Increasing Size, more than 20 % for Random Size and more than 15 % for DecreasingSize as compared to existing replacement policy PAID.

We compared all policies under various system parameters and for two scope distributions. But in real-world, there can be lots of scope distributions varying from regions to countries. Moreover, we used Euclidean distance to calculate the distance between two points. However, in the real-world, this distance cannot represent the real distance that a user has to cover in order to reach to an object. For example, in City model the distance between two points is calculated using Manhattan distance. Hence, there is a need to explore proposed policies under different real-world conditions. Also, LDIS have been referred to as some of the most promising technological inventions that may impact consumer behavior over the next decade. User's expectations from mobile networks are becoming more demanding and this trend is expected to intensify in the future. Hence, the user need/profile is essential to develop better cache replacement policies. Also, existing location dependent cache management schemes consider only location-dependent data. Investigation of location dependent cache management schemes, which includes temporal dependent update frequencies, is also required. Future schemes for cache management should consider the above facts to over come the challenges posed by LDIS.

7. References

[1] D. Barbara, "Mobile Computing and Databases: A Survey," IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 1, pp. 108-117, January/February 1999.

[2] D. Barbara and T. Imielinski, "Sleepers and Workaholics: Caching Strategies in Mobile Environments," In the Proceedings of the ACM SIGMOD Conference on Management of Data, Minneapolis, USA, pp. 1-12, 1994.

[3] I. A. Getting, "The Global Positioning System," IEEE Spectrum, Vol. 30, No. 12, pp. 36-47, December 1993.

[4] D. L. Lee, W.-C. Lee, J. Xu and B. Zheng, "Data Management in Location-Dependent Information Services," IEEE Pervasive Computing, Vol. 1, No. 3, pp. 65-72, July 2002.

[5] B. Zheng, J. Xu and D. L. Lee, "Cache Invalidation and Replacement Strategies for Location-Dependent Data in Mobile Environments," IEEE Transactions on Computers, Vol. 51, No. 10, pp. 1141-1153, October 2002.

[6] L. Yin, G. Cao and Y. Cai, "A Generalized Target-Driven Cache Replacement Policy for Mobile Environments," In the Proceedings of the IEEE Symposium on Applications on the Internet, pp. 14-21, January 2003.

[7] K. Lai, Z. Tari and P. Bertok, "Location–Aware Cache Replacement for Mobile Environments," IEEE Global Telecommunication Conference (GLOBECOM 04), Vol. 6, pp. 3441-3447, 29th November- 3rd December 2004.

[8] M. Erwig, R. H. Guting, M. Schneider and M. Vazirgiannis, "Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases," GeoInformatica, Vol. 3, No. 3, pp. 269-296, 1999.

[9] Q. Ren and M. H. Dhunham, "Using Semantic Caching to Manage Location Dependent Data in Mobile Computing," In the Proceedings of 6th ACM/IEEE Mobile Computing and Networking (MobiCom), Boston, USA, pp. 210-221, 2000.

[10] S. Dar, M. J. Franklin, B. T. Jonsson, D. Srivastava and M. Tan, "Semantic Data Caching and Replacement," In the Proceedings of the 22nd International Conference on Very Large Databases(VLDB), pp. 330-341,1996.

[11] A. Balamash and M. Krunz, "An Overview of Web Caching Replacement Algorithms," IEEE Communications Surveys & Tutorials, Vol. 6, No. 2, pp. 44-56, 2004.

[12] J. O' Rourke, Computational Geometry in C, chapter 5, Univ. of Cambridge Press, 1998.

[13] E. O'Neil and P. O'Neil, "The LRU-k page replacement algorithm for database disk buffering", In the Proceedings of the ACM SIGMOD, Vol. 22, No. 2, pp. 296-306, 1993.

[14] Il-dong Jung, Young-ho You, Jong-hwan Lee and Kyungsok Kim, "Broadcasting and caching policies for location-dependent queries in urban areas," In the Proceedings of the 2nd International Workshop on Mobile Commerce**,** Atlanta, USA, pp. 54-60, September 2002.

[15] C. Lu, G. Xing, O. Chipara and C. L. Fok, "MobiQuery: A Spatio Temporal Data Service for Sensor Networks," In the Proceedings of the 2nd

International Conference on Embedded Networked Sensor System (ACM SenSys'04), Baltimore, USA, pp. 320-334, 2004.

[16] L. Breslau, P. Cao, L. Fan, G. Phillips and S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications," In the Proceedings of the 18th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'99), New York, USA, Vol. 1, pp. 126-134, March 1999.

[17] T. Camp, J. Boleng and V. Davies, "A Survey of Mobility Model for Ad Hoc Network Research," Wireless Communication & Mobile Computing (WCMC): Special Issue on Mobile AdHoc Networking: Research, Trends and Applications, Vol. 2, No. 5, pp. 483-502, 2002.

[18] Spatial Datasets. Website at http://www.rtreeportal.org, 2005.

[19] J. Jing, A. Helal and A. Elmagarmid, "Client-Server Computing in Mobile Environments," In the Proceedings of the ACM Computing Surveys, Vol. 31, No. 2, pp. 117-157, June 1999.

[20] M. Berg, M. Kreveld M. Overmars and O. Schwarzkopf, "Computational Geometry: Algorithms and Applications," chapter 7, New York, NY, USA, Springer-Verlag, 1996.

Table 1. Configuration parameters and default parameter settings for simulation model

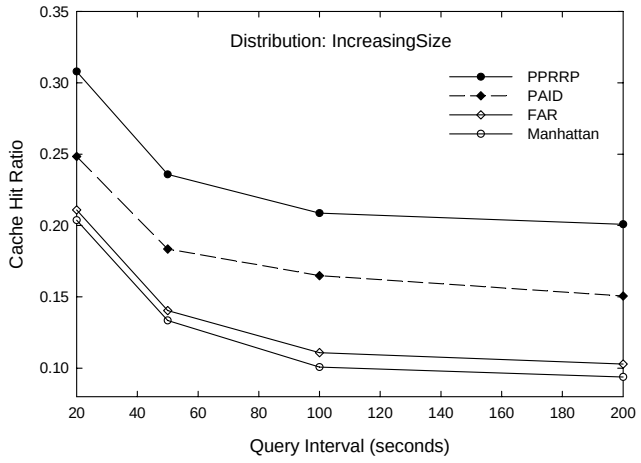
Parameter	Description	Setting
Size	size of the rectangle service area	4000m*4000m, 44000m*27000m
ItemNum	number of data items in the database	500
ScopeNum	number of different values at various locations for each item	110, 215
S_{min}	minimum size of a data value	64 bytes
S_{max}	maximum size of a data value	1024 bytes
UplinkBand	bandwidth of the uplink channel	19.2 kbps
DownlinkBand	bandwidth of the downlink channel	144 kbps
F loatSize	size of a floating-point number	4 bytes
QueryInterval	average time interval between two consecutive queries	50.0 s
MovingInterval	time duration that the client keeps moving at a constant velocity	100.0s
MinSpeed	minimum moving speed of the client	1m s^{-1},5 m s^{-1}
MaxSpeed	maximum moving speed of the client	2m s^{-1}, 10 m s^{-1}
CacheSizeRatio	ratio of the cache size to the database size	10 %
θ	skewness parameter for the Zipf access distribution	0.5

Table 2. Improvement of PPRRP over PAID on different speed ranges (Scope distribution 1)

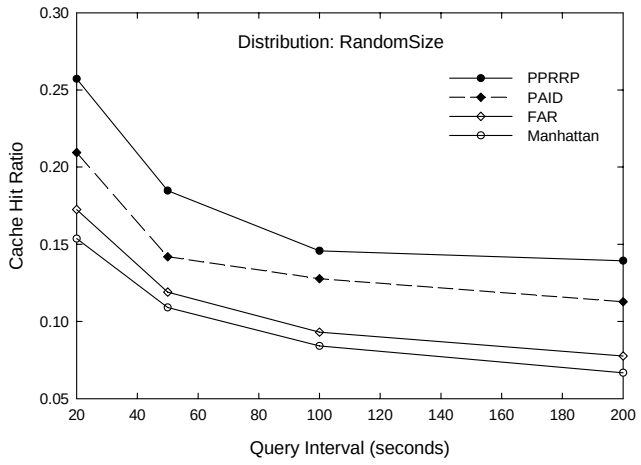
Speed Ranges (m/s)	IncreasingSize (%)	RandomSize (%)	DecreasingSize (%)
1~5	43	29	21
6~10	41	31	25
16~20	40	30	24
25~35	37	29	35

Table3. Improvement of PPRRP over PAID on different speed ranges (Scope distribution 2)

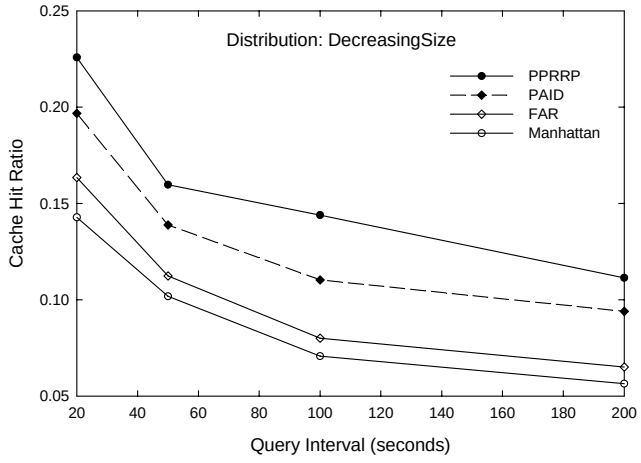
Speed Ranges (m/s)	IncreasingSize (%)	RandomSize (%)	DecreasingSize (%)
1~5	2.2	15.3	16.2
6~10	5.8	8.5	13.5
16~20	4.7	10.9	6.3
25~35	1.2	7.4	11.3



(a)

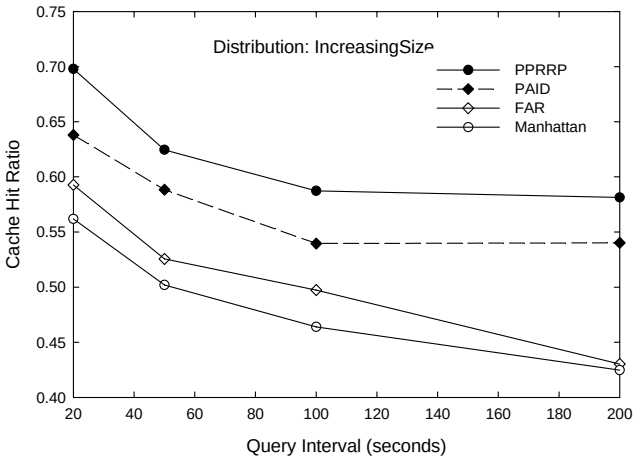


(b)

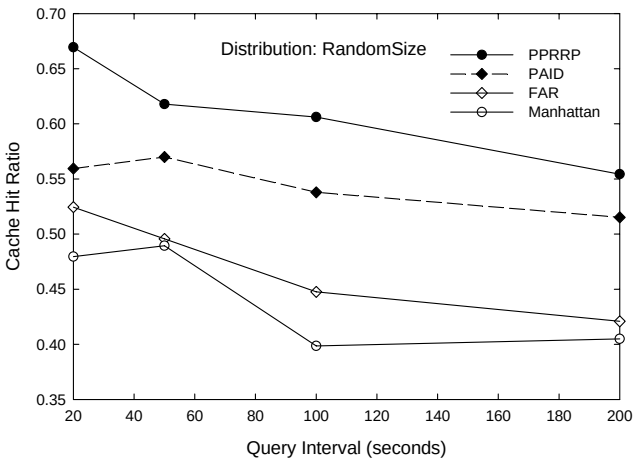


(c)

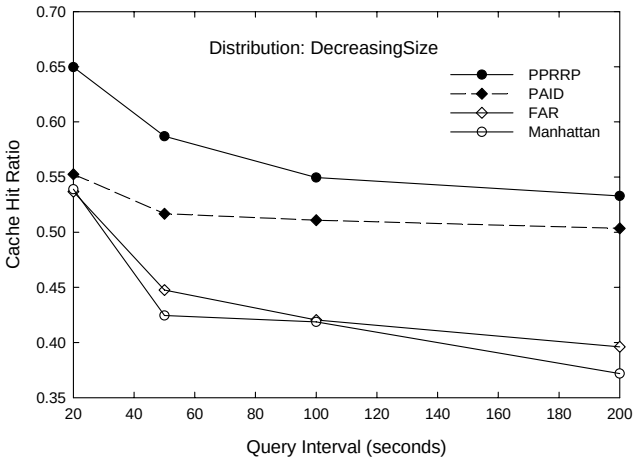
Figure 6. Cache hit ratio vs query interval for scope distribution 1



(a)

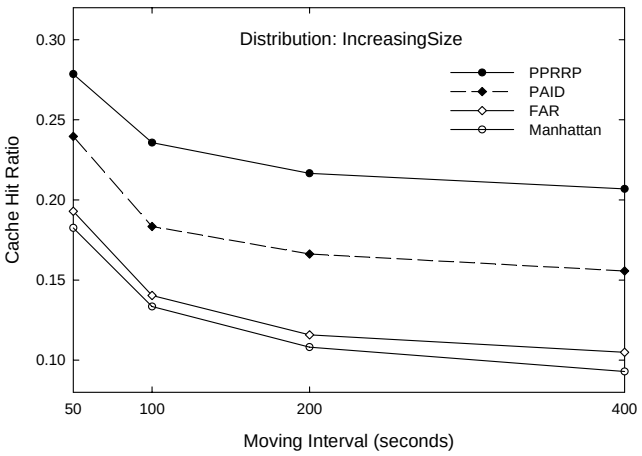


(b)

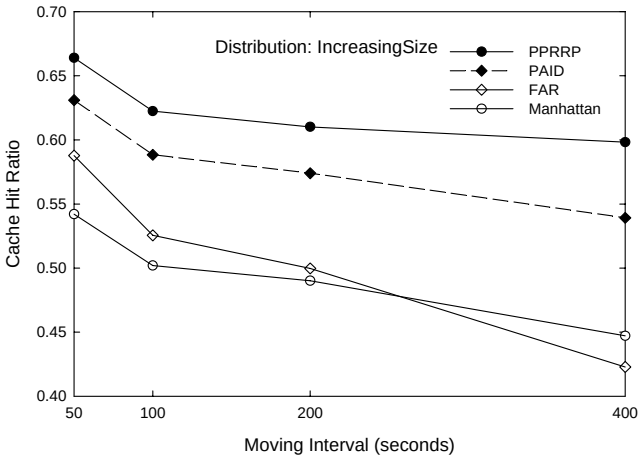


(c)

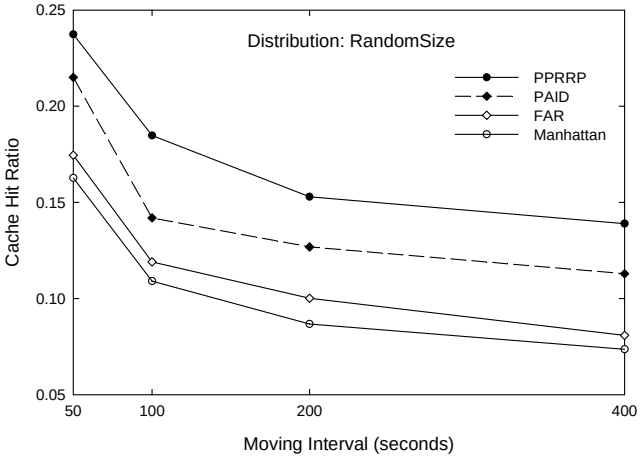
Figure 7. Cache hit ratio vs query interval for scope distribution 2



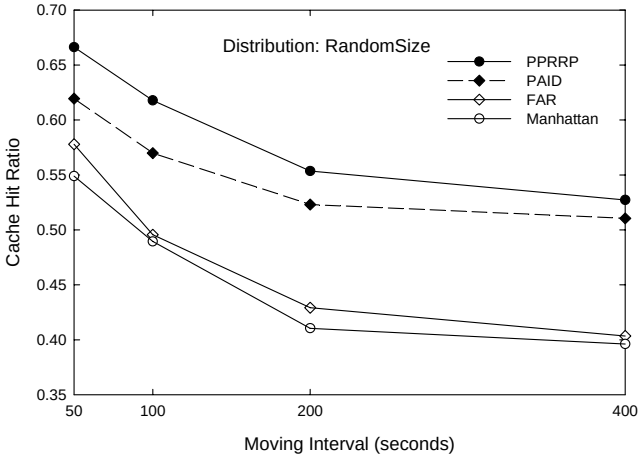
(a)



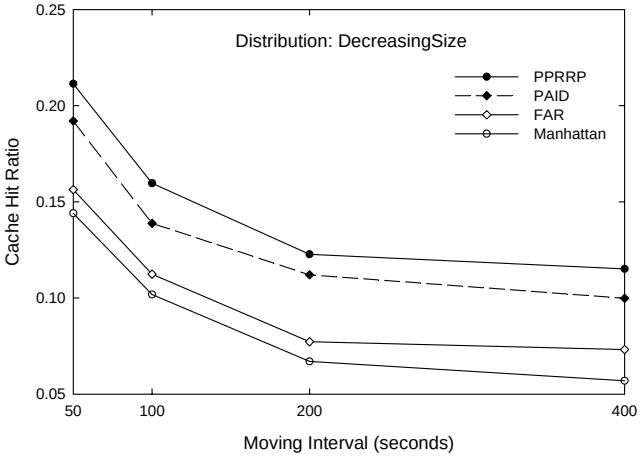
(a)



(b)

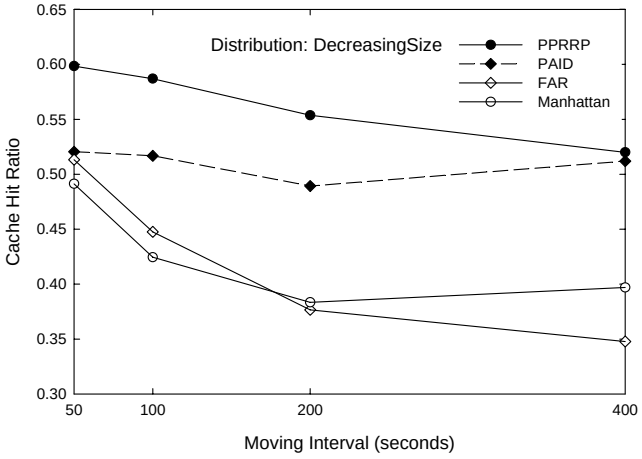


(b)



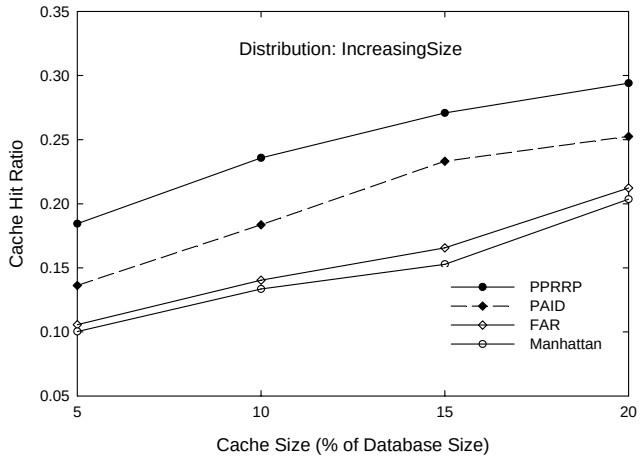
(c)

Figure 8. Cache hit ratio vs moving interval for scope distribution 1

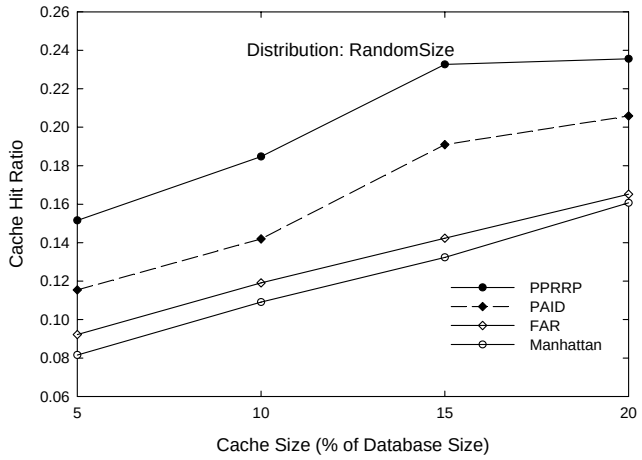


(c)

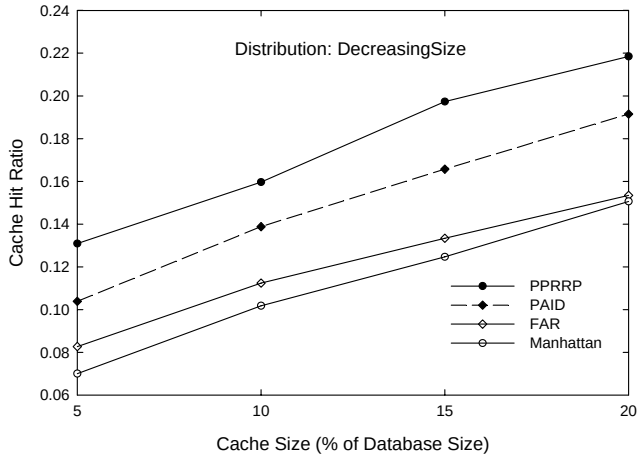
Figure 9. Cache hit ratio vs moving interval for scope distribution 2



(a)

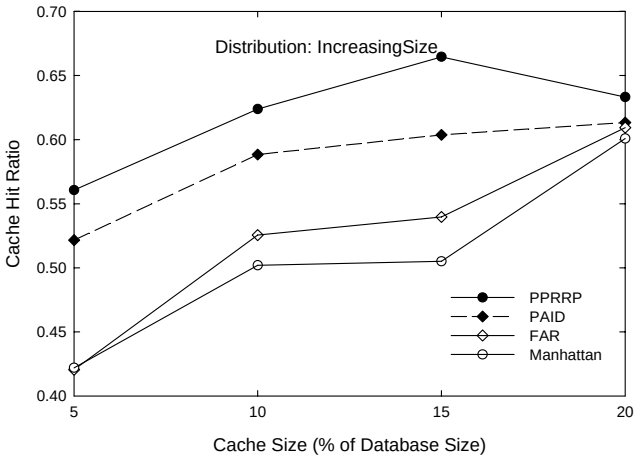


(b)

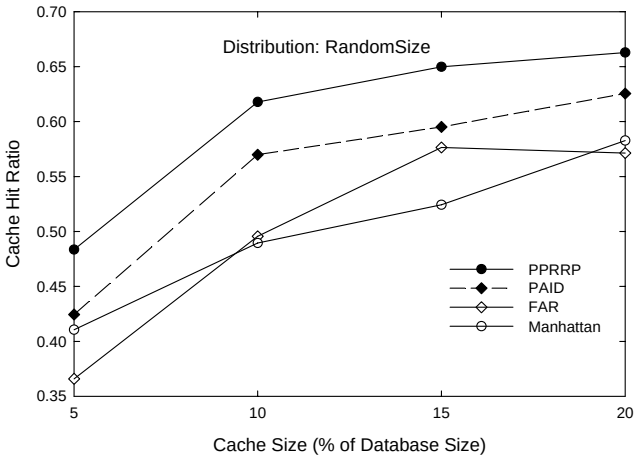


(c)

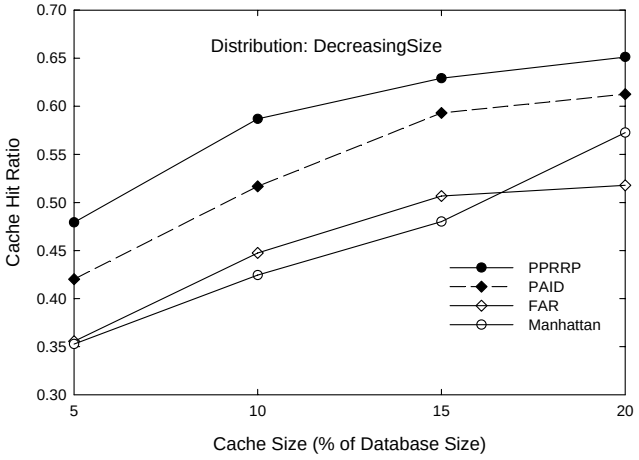
Figure 10. Cache hit ratio vs cache size for scope distribution 1



(a)

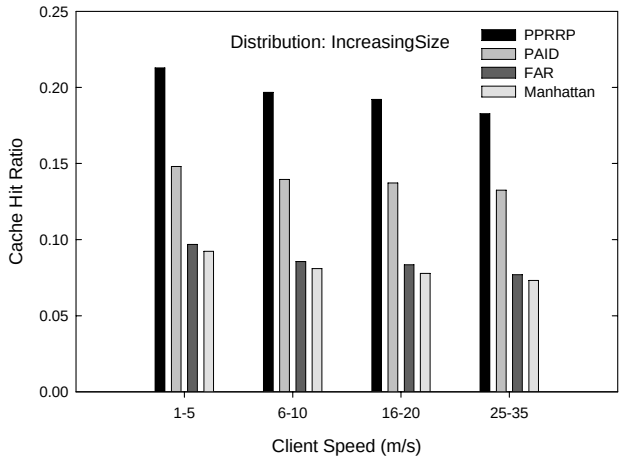


(b)

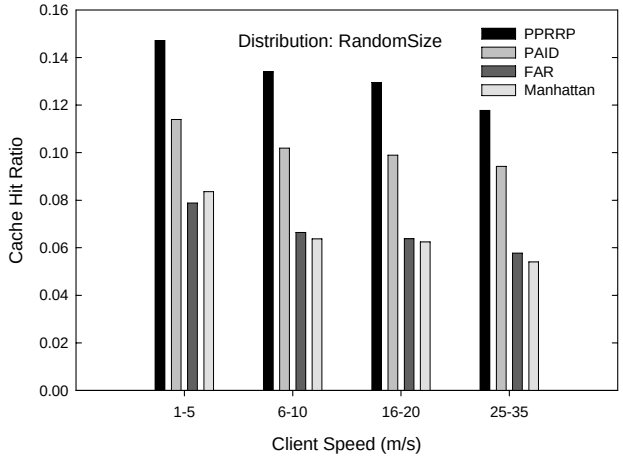


(c)

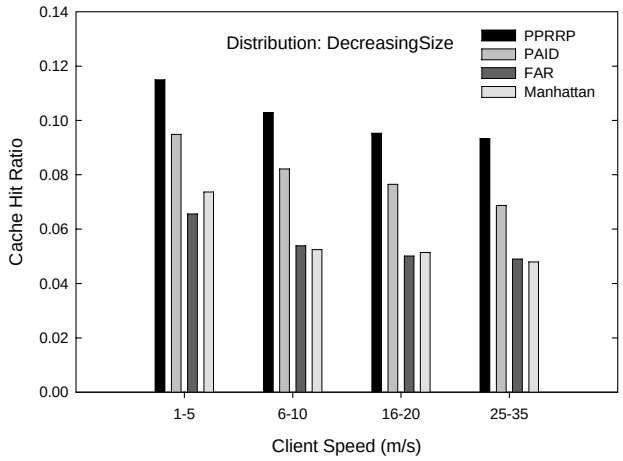
Figure 11. Cache hit ratio vs cache size for scope distribution 2



(a)

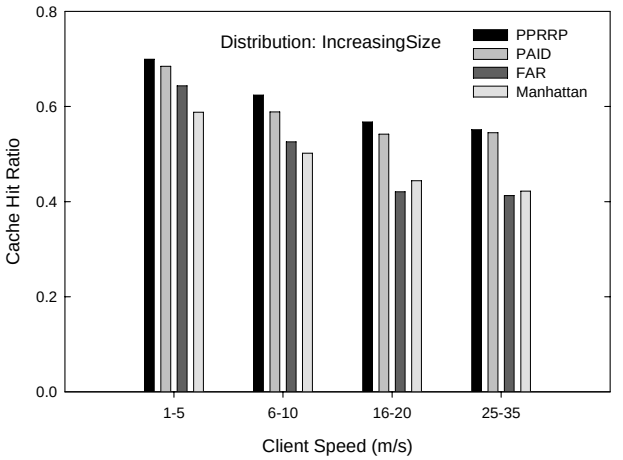


(b)

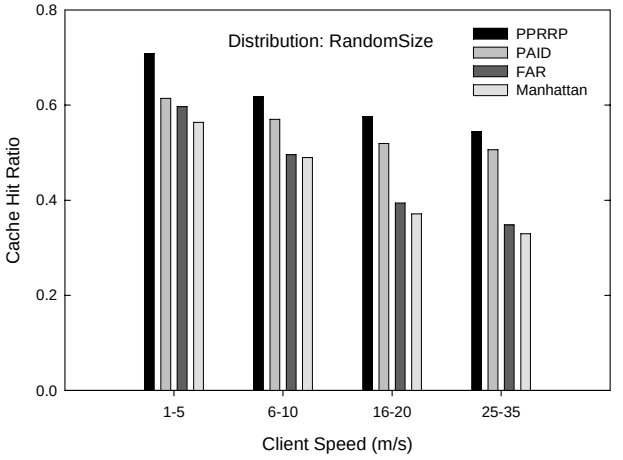


(c)

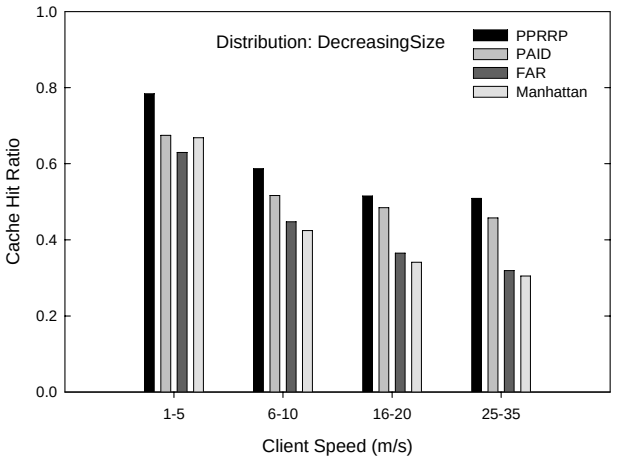
Figure 12. Cache hit ratio vs client speed for scope distribution 1



(a)

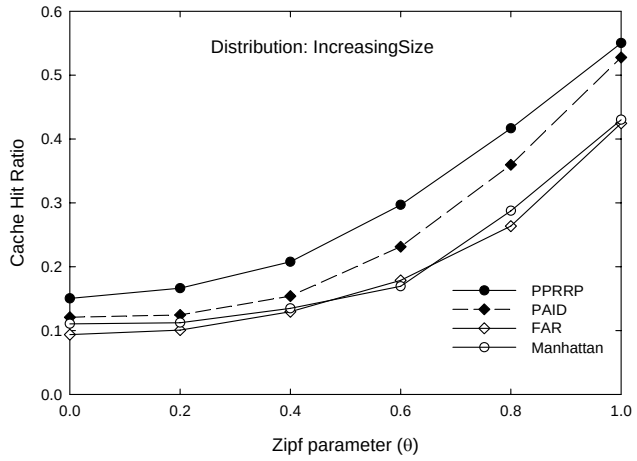


(b)

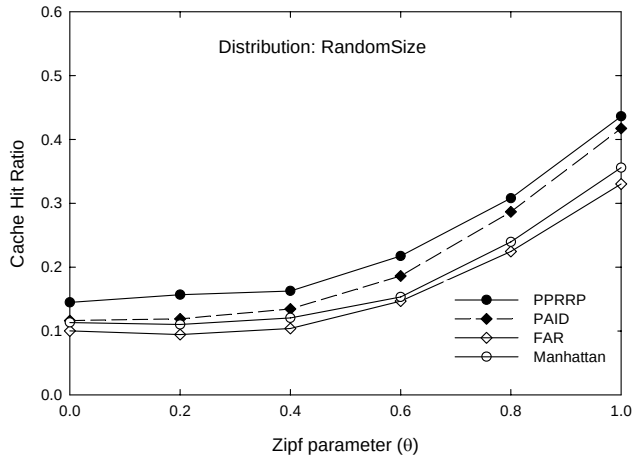


(c)

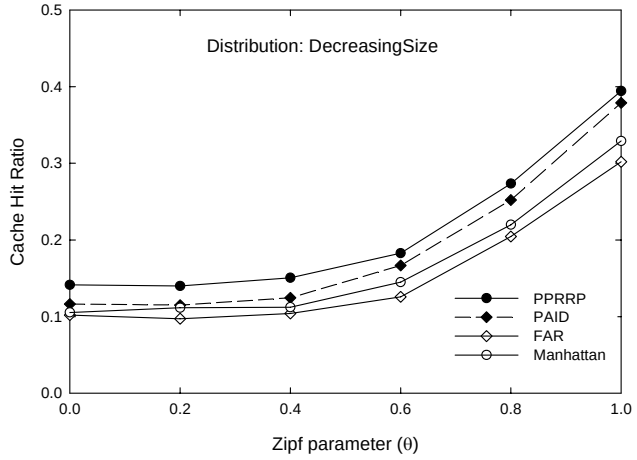
Figure 13. Cache hit ratio vs client speed for scope distribution 2



(a)

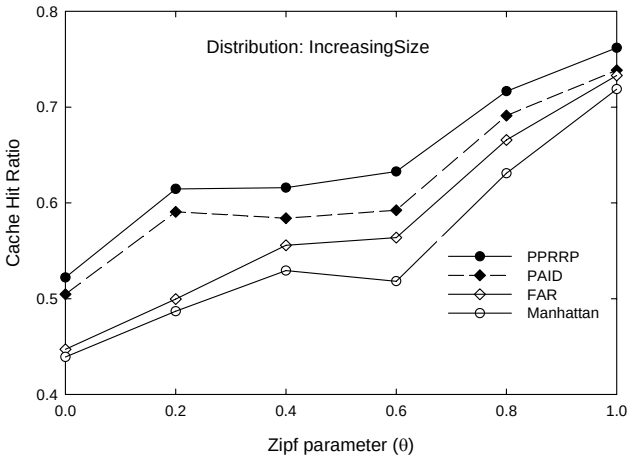


(b)

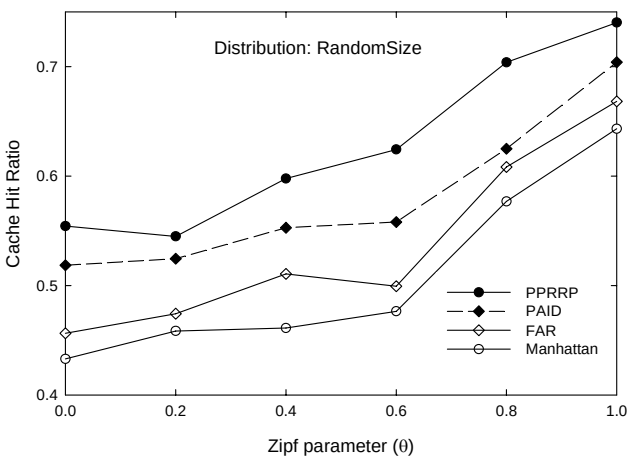


(c)

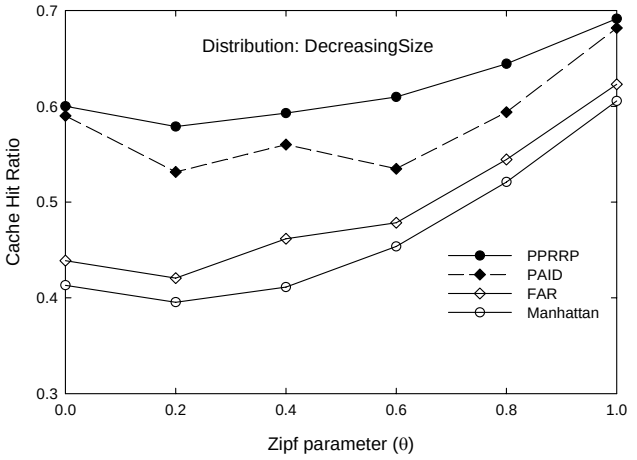
Figure 14. Cache hit ratio vs zipf parameter for scope distribution 1



(a)



(b)



(c)

Figure 15. Cache hit ratio vs zipf parameter for scope distribution 2

I. J. Communications, Network and System Sciences. 2008; 1: 1-103
Published Online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

Advanced Localization of Mobile Terminal in Cellular Network

Marco ANISETTI[1], Claudio A. ARDAGNA[1], Valerio BELLANDI[1]
Ernesto DAMIANI[1], Salvatore REALE[2]
[1]*Department of Information Technologies, University of Milan, via Bramante 65, Crema (CR), Italy*
[2]*Nokia Siemens Networks, via Monfalcone 1, Cinisello Balsamo (MI), Italy*
Email: {anisetti, ardagna, bellandi, damiani}@dti.unimi.it, salvatore.reale@nsn.com

Abstract

The growing diffusion of pervasive collaboration environments and technical advancement of sensing technologies have fostered the development of a new wave of online services whose functionalities are based on users' physical position. Thanks to the widespread diffusion of mobile devices (e.g. cell phones), many services can be greatly enriched with data reporting where people are, how they are moving, or whether they are close by specific locations. Geolocation of mobile terminals relies on the cellular network infrastructure and protocols to provide a reliable and accurate estimate of mobile terminals' position, without the need of global positioning systems, such as GPS. In this paper, we present a novel lookup table correlation technique for geolocation, with multiple position estimations and optimal location techniques. Our approach provides high precise location and tracking of mobile terminals by exploiting advanced propagation models for mobile radio networks design, and by querying Geographical Information Systems (GIS), in conjunction with Kalman predictive filtering.

Keywords: Mobile Phone, Geolocation, Kalman Filter

1. Introduction

In the field of mobile networks, the term "geolocation" is used for denoting a variety of techniques aimed at *mobility prediction*, which is, computing and tracking the position of a mobile terminal. Mobility prediction can be exploited both at *network* and *service level*. At *network level*, mobility prediction and location support several crucial tasks, such as handoff management [1], efficient code division in 3G network [2], orthogonality factor prevision for WCDMA network [3], wireless routing management [4], system for QoS support [5][6] and resource allocation [7]. At *service level*, mobility prediction and location are tightly coupled with number of location-based applications, such as, navigation, instant messaging [8], friend finder and point of interest services [9][10], emergency rescue [11], and many other safety and security services [12][13][14][15][16]. Although some of the above applications need only rough position estimation, many of them need precise tracking of mobile terminal position within cells to provide an adaptable quality of service.

A first branch of research on mobile geolocation focuses on satellite-based positioning, i.e., GPS (Global Positioning System) [17], which is an interesting option for high-end applications, where location precision represents a critical requirement. However, standard GPS geolocation is not well suited for all contexts, as for instance in dense urban areas or inside buildings, where satellites are not visible from the mobile terminals. For these reasons, we claim that even leaving costs and impact on battery consumption aside, GPS techniques are not likely to be the key technologies for a number of interesting Location Based Services (LBSs), such as mobile tracking and path certification. Also, GPS systems are not suitable within urban areas, due to the high costs of their adaptation to urban settings. By contrast, our solution is based on traditional GSM/3G networks and does not involve any change to existing mobile network infrastructure being based on data collected by the cellular network. Our general purpose, low cost *Positioning and Motion Tracking System* (PMTS) for mobile networks provides mobility prediction both at *network* and *service level* [18].

In this paper, we discuss the importance of GIS information in Electro Magnetic Field (EMF) prediction for PMTS and propose a novel technique for high reliable mobile terminals location and tracking. Our proposal relies on additional information extracted from a GIS database covering the area of interest, used in conjunction with advanced predictive filtering. In general GIS maps can include information about the roadways (*Type 1*), to

improve tracking and trajectory prevision, and about the environment (*Type 2*), used especially for EMF or *time delay* prediction (for a complete overview of location techniques see Section 2). Regarding *Type 1*, we classify the information extracted from GIS maps in three layers with increasing level of detail: *i) layer 1* provides a coarse-grained subdivision of the mapped area into regions (e.g., pedestrian-only areas), *ii) layer 2* provides information such as streets width and precise street conformation, *iii) layer 3* provides highly detailed information (e.g., distance from crossroads, one-way street) and, when available, information on traffic and speed limits. Concerning environment-related information (*Type 2*), we consider two possible levels of detail: the absence of information (really diffuse in many GIS map), and the presence of terrain or buildings information.

Our approach takes advantages of both from type 1 and type 2 information of GIS maps; the more information is available, the more accurate the location will be. The level of location accuracy, in fact, depends on maps information and time constraints.

Our novel contributions can be summarized as follows.

- *Improved database technique for multiple candidates' localization.* Our technique is based on a LookUp Table (LUT) signal strength approach where the lookup table is filled with path loss previsions of each antenna. These previsions take advantage from environmental information extracted by GIS map (Type 2), and consider the antenna's shape to better fit real environments. The lookup table is then used to perform a multi-candidates selection. A local minimum management strategy is included to improve the precision in multi-candidates selection process.
- *Time-Forwarding Tracking (TFT).* Our technique exploits GIS map (Type 1) information and predicts motion model to select one among all candidates' locations. Each candidate is previously projected on the road to check whether the mobility model[1] is compatible with the actual movement. TFT is also able to deal with EMF fluctuation building a time forwarding graph.
- *Constrained Advanced Filtering.* We perform an advanced filtering for better enforcing other map constraints, such as, one-way streets.

Finally, we validate our algorithm providing real experiments carried out in a complex urban environment, that is, the city center of Milan, Italy.

[1]We consider two basic types of motion models: pedestrian and vehicle. Other models could be introduced for special applications.

2. Related Work

Mobile location techniques are the topic of several studies in the area of mobile applications. Among the solutions used by GSM/3G technologies for location purposes, the most important and already standardized are propagation time and signal strength techniques.

Propagation time-based methods (e.g. ToA [19], TDoA [20], E-OTD [21]) rely on time measurement. However, they have the main disadvantage of producing acceptable data only when the Line-Of-Sight (LOS) between terminal and based stations is guaranteed. Generally speaking, the main limiting factor of this class of techniques is that the accuracy of the estimated position depends mainly on the number of measurements done and on the geometric configuration. This problem is even worse in urban environments, where multipath propagations lead to very complicated scenarios without LOS between the mobile terminal and the base stations.

Signal strength-based techniques instead are based on Received Signal Strength Indication (RSSI), which measures signal attenuation, assuming free space propagation and omnidirectional antennas, i.e., signal level contours around a base station are concentric circles, where smaller circles enjoy more powerful signals [22][23]. Although the same principle works well also for directional antennas, signal level contours are not concentric circles, but more complex geometrical shapes. Exploiting this assumption mobile antenna location problem is reduced to the well-known triangulation position problem which is similar to time-based and angular approaches. As a consequence, RSSI metric is also not well-suited for urban areas and the signal strength calculated with this approach is not more reliable than the one obtained by time-based approaches. The lack of precision in urban environments is then due to the fact that free space propagation assumption does not hold because of multipath propagation and shadowing, leading to complex signal shapes. Physical phenomena influencing radio propagation are mainly four: reflection, diffraction, penetration, and scattering. To solve these problems, deterministic (ray-tracing, IRT [24][25]), empirical (Hata-Okumura, Walfisch-Ikegami [26][27]), and hybrid techniques (Dominant Path [28][29][30]) for EMF prediction have been developed, for various environments. EMF prediction methods using signal strength can be profitable for location purposes. Even if path loss prediction techniques seem to achieve the best results, many problems still remain unsolved, including the intrinsic error of EMF prediction algorithms and fluctuations due to environmental changes. In this context, recently, an interesting approach has been proposed to deal with EMF fluctuations, by using support vector regression [31]. In a nutshell, regression techniques model the location problem as a checkpoint location, which can be solved as a machine learning problem. Our

approach, however, does not rely on a training phase, since real data sampling are not available everywhere. Furthermore, we focus on tracking and on dense localization, where checkpoint and neural network based strategies are not suitable. Specially, we rely on the basic techiniques outlined below.

Database correlation. This technique relies on a database built on RSSI predictions or measurements. Mobile terminals positions are determined by evaluating RSSI measurements, which are performed by individual terminals or even by the BSs. The measurements are compared with the entries in the database. The corresponding correlation calculations find out the database entry that better matches the measurements lead to a location estimation [32][33]. Our approach basically relies on the principle of database correlation. Much in line with the RADAR system developed by Microsoft research [34] for wireless networks, our method uses a LUT for multiple candidates' selection. Differently from RADAR, however, we select the number of candidates depending on the sensibility of the region and on GIS information.

Filtering. Using triangulation and database correlation, the intrinsic location error of RSSI is never taken into account. Many recent works are aimed at overcoming this problem using Kalman filtering techniques with mobile motion model and RSSI triangulation. One interesting work [39] treats the problem of mobility in ATM network. It develops a hierarchical user mobility model that closely represents the movement behavior of a mobile user, and uses pattern matching and Kalman filtering, yielding to an accurate location prediction algorithm. Another work [40] proposes two algorithms for real-time tracking, location, and dynamic motion of a mobile station in a cellular network. This method is based on pre-filtering and two Kalman filters (one to estimate the discrete command process and the other to estimate the mobility state). The mobility model is built on a dynamic linear system driven by a discrete command process that was originally developed for tracking maneuvering targets in tactical weapons system [35][36]. The command process is modeled as a semi-Markov process over a finite set of acceleration levels, as in [37]. The filtering technique, presented in our work does not filter the location of mobile using RSSI [37][38], but it takes the most probable position filtered by our Time Forwarding Tracking (TFT) technique and tries to enforce the map and motion model constraints. Therefore, our filtering technique is well distinct from the ones introduced above, even if it exploits a similar motion model.

3. EMF Prediction with Antennas' Shape

EMF prediction is a crucial task for those location strategies relying on triangulation or lookup table over RSSI. The amount of GIS information available (Type 2 from GIS map) drives the choice among a number of applicable algorithms. In this paper, we adopt Hata-Okumura [26] for prediction in absence of environmental information and COST231 Wallfisch-Ikegami [27] to take into account the shapes of the buildings. The statistic prediction model COST231 is only suitable for simulated environment where antennas are supposed to be omni-directional. In our previous works, we adopted this approach for EMF prediction, obtaining very good results in a simulated urban environment [39][40]. In real environment, however, this omni-directionality cannot be assumed without a loss in EMF estimation quality, since real antennas' shape have a big impact on EMF prediction. There are two ways to deal with real antennas' shapes: *i)* use a deterministic ray-tracing, *ii)* introduce shapes in statistical EMF prediction. Here, we propose a variation of COST231 that includes a simple notion of antenna shape to better estimate the EMF. The deterministic ray tracing techniques in fact suffer of several problems, as for instance burdensome time-complexity need for a high precision antenna database, which make them not suitable for our approach. We model the shape of the antenna as a simple function S_a of the direction angle, S_a:*[0…360]* → *[0…Maxloss]*, where *Maxloss* is the maximum loss defined for a certain type of antenna. Function $S_a(\alpha)$ maps degree α to the loss due to the shape of the antenna *a*; such loss depends upon the angle of the point where the field is measured with respect to the antenna's main axes.

As an example, using our shape function inside COST231, the path loss of point *p* over the map, with respect to antenna *a*, is defined as follow:

$$PATHLOSS = l_{add} + l_b \quad (1)$$

where $l_{add}=S_a(\alpha)$ represents the additional loss produced by the shape of the antenna *a,* with α the angle identified by point *p* and the principal direction of antenna *a*, and l_b is the component of canonical COST231 (i.e., either *LOS* or *NLOS*). Fig. 1 shows a comparison between EMF predicted with buildings structure and omni-directional antennas and EMF predicted with buildings structure and directional antennas.

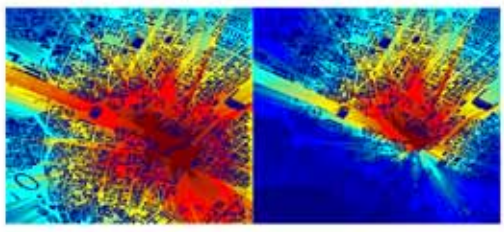


Figure 1. Comparison between omni-directional antennas and directional antennas.

The availability of information about the shapes and directions of the antennas, as well as more information on the environment, allows of significantly improving the

accuracy of the prevision. Also, the knowledge of buildings heights and sizes greatly improves the quality of prediction. To assess this improvement, we performed several tests described in the experimental results in Section 7. Fig. 2 shows a comparison between real RSSI and the predicted ones using shape and COST231. Fluctuation is a typical problem of real environments; nevertheless the predicted trend is very close to the real one.

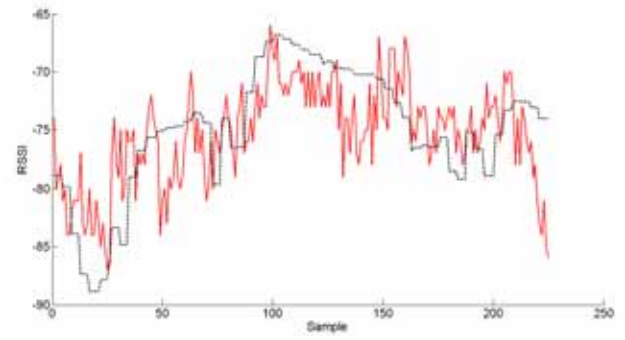


Figure 2. Comparison between real RSSI in red and the estimated ones in black.

4. Multiple Candidate Lookup Table Geolocation

The core of our solution is based on a database correlation approach [33], where the position of a mobile terminal is determined by comparing the measurements performed by the mobile terminal itself (assuming it knows the signal strengths of the six bestserving antennas) with the entries in the lookup table. Our lookup table is defined in the area of interest, starting from the predicted path loss for all the antennas in the area. Using these prediction values, we obtain a matrix with the structure shown in Table 1. Every row in the matrix represents a single point within the coverage area (expressed in *x, y* coordinates, if 2D cartesian representation of area is used, or as a latitude and longitude pair, if GPS representation is used). The path loss predictions from the *r*-th base stations to each given point *p* are stored as entries in the row corresponding to point *p*.

Table 1. Lookup table structure

Location		Cell			
x	y	1	2	i	r
x_1	y_1	$E_{1,1}$	$E_{2,1}$	$E_{i,1}$	$E_{r,1}$
x_2	y_2	$E_{1,2}$	$E_{2,2}$	$E_{i,2}$	$E_{r,2}$
⋮	⋮	⋮	⋮	⋮	⋮
x_m	y_m	$E_{1,m}$	$E_{2,m}$	$E_{i,m}$	$E_{r,m}$

Of course, lookup table filling and updating can be done only once in a while, when major changes on the area of interest occur. In principle, all EMF prediction models can be used to compute the lookup table, depending on application requirements. Generally speaking, computing the lookup table for a given area consists in super-imposing a grid where field levels are quantized. The grid does not need to be uniform; rather, its sparseness can be controlled on the basis of the characteristics of the area of interest and of the cost constraints.

During the process of locating a mobile device, the observed path loss on terminal is compared with all entries in the table. This comparison can be done with many different criteria (interesting examples can be found in [33]). Specifically, we used a sum of squared errors between the measured path loss M_j for each antenna *j* and the path loss defined by entry *i* in the lookup table $E_{i,j}$ (see Equation (2)). Formally, we introduce the following equation.

$$e = \sum_{j=1}^{r} \left(M_j - E_{i,j} \right)^2 \tag{2}$$

The location of the mobile terminal is defined as the coordinates of the entry in the table that produces the smallest error *e*. Of course, this single-point location technique suffers of an intrinsic error; indeed, the estimated mobile terminal position using this kind of minimization hardly ever gives the correct geolocation due to discrepancy between real field strength and the predicted one. The main causes of error in predicting field strength are:

i) intrinsic model error,
ii) imprecision in geographical database,
iii) variation in the antenna features,
iv) variation in weather conditions.

Again, these errors are spread over all the area of interest. For these reasons, our approach produces a variable number *n* of position estimates, depending on a sensibility map analysis. Our sensibility map, built on map information and antennas positions, represents the error sensitivity of our multiple candidates geolocation method [40]. As expected, the higher the candidates number, the higher the probability of obtaining a better location. Nevertheless there is an unwanted side effect in increasing the number of candidates: the number of candidates taken into account increases the complexity of the candidate selection process described in the following section.

5. Time-Forwarding Tracking (TFT)

For validating candidates selected using the techniques introduced in the previous section, we developed a tracking method based on a time-forwarding algorithm. This algorithm uses *m* time position estimates and *n* nodes candidates at each time to define a directed acyclic graph, called Time Forwarding Graph (TFG). Every node *p* in the graph represents one of the possible positions of the mobile terminal, while edges, defined by the bounded

nodes (source and destination nodes), represent motion between them. Each edge has associated a weight that is computed based on reachability and map constraints. In our previous work [40], this weight was defined by taking into account the estimated velocity and acceleration at the source node, and by evaluating the reachable velocity and acceleration considering the position of the destination node. The obtained values were then compared with information inferred from the map. Of course, this type of advanced function works well when a reasonable upper bound to the position error can be assumed. Since in this paper we deal with real (as opposed to lab) environments, an acceptable error for every position cannot be guaranteed. Therefore the aforesaid weighting techniques cannot be applied. We then adopt a simpler weighting technique postponing the refinement to the filtering stage. At this stage, we only evaluate the edge weight to exclude the unreachable nodes. Therefore the weight function W is defined over an edge $e = (p_{i,t}, p_{j,t+k})$ as follows:

$$W(e,map)=\begin{cases}\mu(e,map) & if \quad \mu(e,map)\le Th(k)\\ +\infty & otherwise\end{cases} \tag{3}$$

where $i,j \in [1,\ldots,n]$ and $k \in [1,\ldots,m]$. The function μ calculates the real distance between two nodes $p_{i,t}$ and $p_{j,t+k}$ taking into account the map (presence of buildings, street curves and so on). The threshold $Th(k)$ defines the maximum acceptable distance between each node and it is a function of the time variation k. Using this approach the weights of the edges are in linear relation with the nodes distances. Considering the real environment fluctuation and the fact that the sampling interval is very short (≈500 *ms*), this assumption is realistic enough for TFT filtering purpose[2]. Summarizing each edge of the graph has a weight that defines the reachability between the bounded nodes. All maps and motion constraints are enforced by the weight function. By searching the minimum path on the graph (i.e., the path from time t to time $t+m$ with minimum weight), we obtain a preliminary filtered position for the mobile (the nodes included in the obtained path). In our previous work, we used a different weight function that worked using $k = 1$, meaning that there were no edge between non-temporal consecutive nodes.

Once again, we remark that, in real applications, the assumption of a bounded error is too restrictive, and a set of candidates' nodes can be completely unreachable due to the fluctuation effect. Therefore our TFG needs to take into account also this noise effect. Every node p of the TFG is then associated with a set of nodes S_t at certain time t as follows:

$$S_t = \left[p_{1,t}, p_{2,t}, \cdots, p_{n,t}\right] \tag{4}$$

S_t represents the set of the n candidates positions for time t. The distance from a node $p_{i,t}$ to a node $p_{j,t+k}$ is the same as the weight W among them. The distance can be also defined from a node $p_{i,t}$ and a set of nodes S_{t+k} as follow:

$$M\ (p_{i,t}, S_{t+k}, map) = \min_{j\in[1\cdots n]}(\mu(p_{i,t}, p_{j,t+k}, map)) \tag{5}$$

where M is a distance function from node $p_{i,t}$ to set S_{t+k} according to the map. In our TFG graph, two types of edges exist:

- edges e between two consecutive nodes, if $W \neq +\infty$.
- forwarding edges e between two non temporal consecutive nodes. $p_{i,t}$ and $p_{j,t+k}$, if $M(p_{i,t}, S_{t+k}, map) \neq +\infty$ and $min_{h=1..k-1}(M(p_{i,t}, S_{t+h}, map) = +\infty$.

Fig. 3 shows the TGF. Note that in our approach, the choice of parameter m becomes critical. On the one hand, using a high m (i.e., long time prevision), we obtain a "strong trend" prevision that filters out any out-of-trend movement. This result is not acceptable in pedestrian-only areas like a city square, where motion can well be chaotic without any prevalent movement trend. On the other hand, a value too small for m could cause an error dependency, which in turn could produce bad results in high trend-correlation areas like motorways or one-way streets. However, since layer 1 map information permits to distinguish between pedestrian-only areas and motorways, it becomes adequately possible to tune parameter m.[3] In this way, we obtain a map correlation of our first-level movement estimates, and we can substantially improve the tracking quality of our lookup table technique (see Section 7 for details).

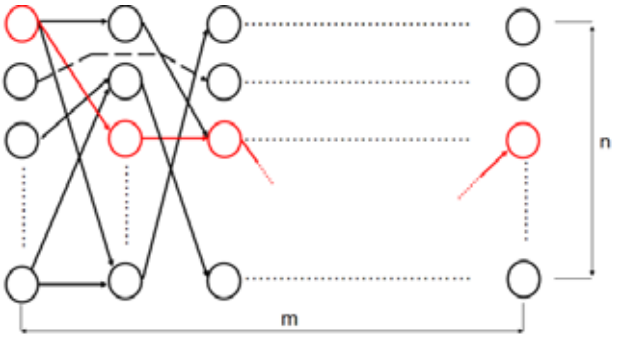


Figure 3. Graph through example for time-forwards tracking. In red the selected position nodes and the selected path, in black the possible position nodes. The dotted line shows an example of an edge between two non consecutive nodes.

[2] To the best of our knowledge the probability of having an overlapped estimation, over time, is high only in the case of stationary mobile device.

[3] A similar dynamic tuning of parameter m can be based on the characteristics of user's vehicle.

Our time-forwarding tracking technique takes full advantage from all available GIS information, such as area classification, so that validation of candidates' locations depends on map constraints. In many cases, the additional information provided by the map is poor or absent; when this happens, our technique is able to dynamically build an information database by estimating all relevant knowledge including speed and acceleration.

6. Tracking with Constrained Kalman Filtering

Using the time-forwarding tracking technique explained in the previous section, we obtain a trusted location measurement z_k at time k, which complies with all map constraints taken into account by TFT. This location z_k can be associated with a state $X_k=[x,y,x',y']^T$ of our mobile device at certain time k, where x and y are the position coordinates and x' and y' define the velocity vector. In general, this state defines a movement trend that already has high confidence and low error, if compared with the actual mobile antenna path (see Fig. 5). To increase its quality, this movement trend is filtered with constrained Kalman filter (CKF) [41], obtaining a robust error and time-deep prevision tracking (Fig. 4 shows our CKF architecture).

The Kalman filter module (the white box in Fig. 4) has the following input: *i)* state position measure z_k from TFT that use map layer 1, ii) the u_{k-1} control input provided by HSMM (Hidden Semi-Markov Model) module, used also in [40], and defined using also map layer 3, iii) constrained function defined on map layer 2. One of the main advantages of this filtering is that it takes into account both measurement error and system error; besides, this one previous-state dependency filtering becomes very usable in real time tracking environment. The Kalman state equation shown in Fig. 4 is defined using the following fundamental matrixes:

$$A=\begin{bmatrix}1 & 0 & T & 0\\ 0 & 1 & 0 & T\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix} \quad (6)$$

$$B=\begin{bmatrix}\frac{T^2}{2} & 0\\ 0 & \frac{T^2}{2}\\ T & 0\\ 0 & T\end{bmatrix} \quad (7)$$

$$C=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix} \quad (8)$$

where T is the time interval between two consecutive observations.

The work [40] proposes to use simple Kalman filtering with HSMM module for location in wireless network. The main difference between this approach and ours is that we filter only motion errors and inertia oversight, but we do not filter the error introduced by signal strength prevision. Indeed, the position estimate used to correct the Kalman prevision is already compatible with map constraints and motion characteristics of the mobile antenna. Using this filtering, we improve the accuracy of the mobile terminal's location.

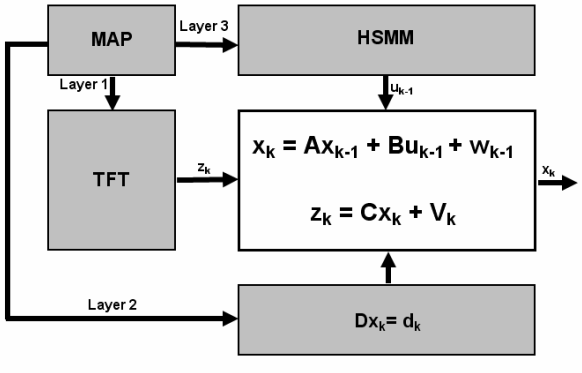


Figure 4. Constrain Kalman filtering architecture.

7. Experimental Results

Using three different cellular phones and a GPS antenna, we performed five trips in downtown Milan, over a month of experimentation. The city area we chose, which is near the railway station called Milano Centrale, includes a non-uniform urban environment, with a park and some skyscrapers. All trips were performed both by car and on foot, and had duration varied from 5 minutes to 1 hour. During these trips, information related to serving and neighboring cells coupled with GPS latitude and longitude were collected every 480 *ms*. We performed two different types of testing, one for evaluating EMF prediction quality and another for assessing geolocation quality with respect to the actual position of a moving cellular phone. Specially, we emphasize the relation between the amount of available environmental information and the quality of predicted EMF. Environmental information is costly to collect and is not likely to be available everywhere.

Since our aim is to locate mobile phones, we also investigated the relation between EMF prediction and the amount of filtering. More specifically, in the first type of experiments, we computed EMF using different combinations of information taken from our Type 2 GIS

map. Selected combinations are: *i)* antenna's shape with no environmental information (LOS), *ii)* some environmental information (buildings structure, with fixed elevation) but no antenna shape (NLOS), *iii)* environmental information and antenna's shape (NLOS+S), and *iv)* all information, including buildings elevation and shape (NLOS+S+E).

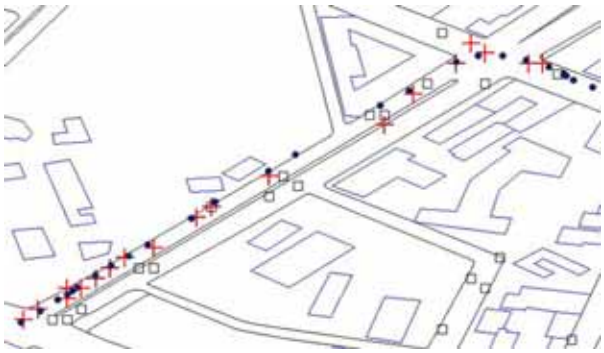


Figure 5. A comparison between some of lookup table candidate(square symbol), geolocation after filtering (cross symbol) and real position presented in black dot.

The results presented in Table 2 show that the quality in EMF prediction depends on the amount of type 2 GIS information available. When all information is available, the mean and variance of the error in EMF prediction is highly reduced.

Table 2. Comparison of Mean Error in EMF Prediction. For sake Of Conciseness, Mean and Variance (in Db) are presented considering a subset of our experimens also number of involved antennas (Ant.) and duration of the trip (Dur) are presented.

Exp.	Dur.	Ant.	LOS	NLOS	NLOS+S	NLOS+S+E
Exp1	3000s	54	49.3-7.0	17.4-6.3	13.0-6.0	11.6-6.6
Exp2	4000s	39	48.3-6.9	18.5-7.1	12.8-7.2	9.4-11.8
Exp3	1089s	22	42.9-5.4	22.4-6.3	14.6-7.9	11.3-6.3
Exp4	1443s	37	48.8-7.1	20.5-8.2	13.1-7.8	9.9-8.6
Exp5	959s	20	41.9-9.9	21.1-9.2	14.9-5.5	13.0-4.7

In the second type of experiments, we present our results in terms of location quality, showing their relation with the available information and thus with the predicted EMF. In our previous work [43], we investigated the quality of our location approach in a simulated environment achieving encouraging results. Here, we evaluate the relation between quality of location and amount of information available. Table 3 shows our results; it is clear that our filtering strategy, which relies on map information, can be fruitfully applied only when all environmental information is available.

Table 3. Comparison of Mean Error Using our Location algorithm

LOS	NLOS	NLOS+S	NLOS+S+E
451m	141m	104m	54m

To conclude, the overall precision of our location techniques is suitable for many location-based services and applications even in a highly diverse urban environment.

8. Conclusion

We have investigated the problem of mobile terminals location in urban environments, analyzing some existing location algorithms, and proposing a novel way to improve location accuracy. Our approach employs a time-forward tracking algorithm with GIS map constraints and a constrained Kalman filtering for error correction purposes. A complete experimentation confirms that detailed GIS information can produce highly precise location even using a simple statistical EMF prediction.

9. References

[1] H. Lin, R. Juang, and D. Lin, "Validation of an improved locayionbased algorithm using gsm measurement data," IEEE Trans. Mobile Computing, vol. 4, pp. 530 – 536, September - October 2005.

[2] H. Buddendick, G. WÖlfle, S. Burger, and P. Wertz, "Simulator for performance analysis in umts fdd networks with hsdpa," 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Barcelona, Spain, September 2004.

[3] S. Burger, H. Buddendick, G. WÖlfle, and P. Wertz, "Location dependent cdma orthogonality in system level simulations," 61st IEEE Vehicular Technology Conference (VTC), Stockholm, Sweden, May 2005.

[4] L. Blazevic, J. L. Boudec, and S. Giordano, "A location-based routing method for mobile ad hoc networks," IEEE Trans. Mobile computing, pp. 97–110, March - April 2005.

[5] W. S. Soh and K. Kim, "QoS provisioning in cellular networks based on mobility prediction techiniques," IEEE Comm.Magazine, vol. 41(1), pp. 86–112, January 2003.

[6] A. Aljadhai and T. Znati, "Predictive mobility support for qos provisioning in mobile wireless environments," IEEE Journal Selected Areas in Communications, pp. 1915 – 1930, October 2001.

[7] J. Chan and A. Seneviretne, "A pratical user mobility prediction algorithm for supporting adaptive qos in wireless networks," Proc. IEEE International Conference Networks (ICON'99), pp. 104 – 111, 1999.

[8] A. J .H. Peddemors, M. M. Lankhorst, J. de Heer, "Presence, location, and instant messaging in a

context-aware application framework," in Proc. of the 4th International Conference on Mobile Data Management (MDM), Melbourne, Australia,, 2003.

[9] CellSpotting.com, http://www.cellspotting.com.

[10] World-tracker.com, http://www.world-tracker.com/.

[11] (14 Jan 2004) Us fcc enhanced 911 fac sheet 2001. [Online]. Available: http://www.fcc.gov/911/enhanced/release/factsheet requirements 012001.pdf

[12] MapAmobile, http://www.mapamobile.com/.

[13] N. Marmasse and C. Schmandt, "Safe & sound a wireless leash," in Proc. of CHI 2003, extended abstract, 2003.

[14] N. Marmasse and C. Schmandt, "User-centered location awareness," IEEE Computer, vol. 37, no. 10, pp. 110–111, 2004.

[15] Teen Arrive Alive, http://www.teenarrivealive.com.

[16] C.A. Ardagna, M. Cremonini, E. Damiani, S. De Capitani di Vimercati, and P. Samarati, "Supporting location-based conditions in access control policies," in Proc. of the ACM Symposium on Information, Computer and Communications Security (ASIACCS'06), Taipei, Taiwan, March 2006.

[17] E. Kapplan, Understanding GPS: Principles and Applications. Amsterdam, Netherlands: Elsevier, 1996.

[18] N. Samaan and A. Karmouch, "A mobility prediction architecture based on contextual knowledge and spatial conceptual maps," IEEE Trans. Mobile Computing, vol. 4, pp. 537 – 550, November - December 2005.

[19] C. Drane, M. Macnaughtan, and C. Scott, "Positioning gsm telephones," IEEE Comm. Mag., vol. 46, pp. 46 – 55, Apr 1998

[20] J. J. Caffery and G. L. Stüber, "Overview of radiolocation in cdma cellular systems," IEEE Communications Magazine, April 1998.

[21] M. Spirito, "On the accurancy of cellular mobile station location estimation," IEEE Trans. Veh. Technol., vol. 50, pp. 3674 – 3685, May 2001

[22] W. Figel, N. shepherd, and W. Trammell, "Vehicle location by a signal attenuation method," IEEE Tans. Veh. Technol., vol. VT - 18, pp. 105 – 110, November 1969.

[23] H. song, "Automatic vehicle location in cellular communications systems," IEEE Tans. Veh. Technol., vol. 43, pp. 902 – 908, November 1994.

[24] R. Hoppe, G. WÖlfle, and F. Landstorfer, "Fast 3-d ray tracing for the planning of microcells by intelligent preprocessing of the data base," 3rd European Personal and Mobile Communication Conference (EPMCC), Paris, March 1999.

[25] G. WÖlfle, R. Hoppe, and F. Landstorfer, "A fast and enhanced ray optical propagation model for indoor and urban scenarios, based on an intelligent preprocessing of the database," 10th IEEE PIMRC, Osaka, September 1999.

[26] Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda, "Okumura-hata propagation prediction model for uhf range, in the prediction methods for the terrestrial land mobile service in the vhf and uhf bands," ITU-R Recommendation P.529-2, Geneva: ITU, 1995.

[27] E. Damosso, "Cost 231: (digital mobile radio towards future generation systems)," Final report, European Comission, Bruxelles, 1999.

[28] G. WÖlfle, R. Wahl, P. Wertz, P. Wildbolz, and F. Landstorfer, "Dominant path prediction model for indoor scenarios," German Microwave Conference (GeMIC), April 2005.

[29] G. WÖlfle, R. Wahl, P. Wildbolz, and P. Wertz, "Dominant path prediction model for indoor and urban scenarios," Duisburg (Germany), September 2004.

[30] R. Wahl, G. WÖlfle, P. Wertz, and P. Wildbolz, "Dominant path prediction model for urban scenarios," 14th IST Mobile and Wireless Communications Summit, Dresden (Germany), June 2005.

[31] Z. li Wu, C. hung Li, J. K.-Y. Ng, and K. R. Leung, "Location estimation via support vector regression," IEEE Transactions on Mobile Computing,vol. 6, March 2007.

[32] D. Zimmermmann, P. Wertz, J. Bauknecht, and F.Landstorfer, "Performance of location methods in an urban mobile communication network," Technical report, 2004.

[33] D. Zimmermmann, J. Baumann, M. Layh, F. Landstorfer, R. Hoppe, and G. WÖlfle, "Database correlection for positioning of mobile terminals in cellular networks using wave propagation models," IEEE Vehicoular technology conference (VTC), Los Angeles USA, October 2004.

[34] P. Bahl and V. N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," IEEE INFOCOM, 2000.

[35] R. A. Singer, "Estimating optical tracking filter performance for manned maneuvering targets," IEEE Trans. Aerosp. Electron. Syst., 1970.

[36] R. L. Moose and H. F. Vanlandingham, "Modeling

and estimation for tracking maneuvering targets," IEEE Trans. Aerosp. Electron. Syst., 1979.

[37] T. Liu, P. Bahl, and I. Chlamtac, "Mobility modeling, location tracking, and trajectory prediction in wireless atm networks," IEEE Journal on selected areas in communications, vol. 16, no. 6, August 1998.

[38] Z. R. Zaidi and B. L. Mark, "Real-time mobility tracking algorithms for cellular networks based on kalman filtering," IEEE Transaction on Mobile Computing, vol. 4, no. 2, March - April 2005.

[39] M. Anisetti, V.Bellandi, E. Damiani, and S. Reale, "Accurate localization and tracking of mobile terminals," 2th International Conference on Wireless Communications, Networking and Mobile Computing, Whuan, Cina, 2006.

[40] M. Anisetti, V. Bellandi, E. Damiani, and S. Reale, "Localization and tracking of mobile antenna in urban environment," International Symposium on Telecomminications, pp. 353 – 358, September 2005.

[41] D. Simon and T. Chia, "Kalman filtering with state equality constraints," IEEE Transactions on Aerospace and Electronic Systems, vol. 39, January 2002.

Call for Papers

International Journal of

Communications, Network and System Sciences (IJCNS)

ISSN 1913-3715 (Print) ISSN 1913-3723 (Online)

Http://www.SRPublishing.org/journal/ijcns/

IJCNS is an international refereed journal dedicated to the latest advancement of communications and network technologies. The goal of this journal is to keep a record of the state-of-the-art research and promote the research work in these fast moving areas.

Editors-in-Chief

Subject Coverage

This journal invites original research and review papers that address the following issues in wireless communications and networks. Topics of interest include, but are not limited to:

MIMO and OFDM technologies	Sensor networks
UWB technologies	Ad Hoc and mesh networks
Wave propagation and antenna design	Network protocol, QoS and congestion control
Signal processing and channel modeling	Efficient MAC and resource management protocols
Coding, detection and modulation	Simulation and optimization tools
3G and 4G technologies	Cognitive Radio

We are also interested in short papers (letters) that clearly address a specific problem, and short survey or position papers that sketch the results or problems on a specific topic. Authors of selected short papers would be invited to write a regular paper on the same topic for future issues of the *IJCNS*.

Notes for Intending Authors

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Paper submission will be handled electronically through the website. All papers are refereed through a peer review process. For more details about the submissions, please access the website.

Website and E-Mail

http://www.SRPublishing.org/journal/ijcns/ Jcnss@SRPublishing.org

I. J. Communications, Network and System Sciences. **2008**; **1**: **1**-**103**
Published online February 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

TABLE OF CONTENTS

Volume 1 **February 2008**

Multiband Scheduler for Future Communication Systems
K. DOPPLER, C. WIJTING, T. HENTTONEN, K. VALKEALAHTI..... 1

Adaptive Throughput Optimization in Downlink Wireless OFDM System
Y. FAKHRI, B. NSIRI, D. ABOUTAJDINE, J. VIDAL..... 10

A Discrete Newton's Method for Gain Based Predistorter
X.C. LIN, M.L. JIN, A.F. LIU..... 16

Particle Filtering with Multi Proposal Distributions
F.S. WANG, Q.J. ZHAO, Y.B. ZHANG, L.Q. ZHANG..... 22

PAPR Reduction Scheme with SOCP for MIMO-OFDM Systems
B. RIHAWI, Y. LOUET..... 29

Any Resource Sharing via HWN* Routing
C. SHEN, S. REA, D. PESCH..... 36

Influence of the Limited Retransmission on the Performance of WLANs Using Error-Prone Channel
H.M. ALSABBAGH, J.P. CHEN, Y.Y. XU..... 49

Exploiting Geometric Advantages of Cooperative Communications for Energy Efficient Wireless Sensor Networks
I. AHMED, M.G. PENG, W.B. WANG..... 55

An Energy-aware Hierarchical Architecture Design Scheme
L.F. YUAN, Y. XU, X. DU, W.Q. CHENG..... 62

Efficient DPA Attacks on AES Hardware Implementations
Y.HAN, X.C. ZOU, Z.L. LIU, Y.C. CHEN..... 68

QoS in Node-Disjoint Routing for Ad Hoc Networks
L. LIU, L. CUTHBERT..... 74

A Predicted Region based Cache Replacement Policy for Location Dependent Data in Mobile Environment
A. KUMAR, M. MISRA, A.K. SARJE..... 79

Advanced Localization of Mobile Terminal in Cellular Network
M. ANISETTI, C.A. ARDAGNA, V. BELLANDI, E. DAMIANI..... 95

191337150101