International Journal of Communications, Network and System Sciences

ISSN: 1913-3715

Volume 2, Number 7, October 2009

www.scirp.org/journal/ijcns/

Scientific
Research

JOURNAL EDITORIAL BOARD

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

Int. J. Communications, Network and System Sciences, **2009, 7, 583-694**
Published Online October 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

TABLE OF CONTENTS

Volume 2 Number 7 October 2009

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

Journal Information

SUBSCRIPTIONS

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

E-mail: service@scirp.org

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

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

SERVICES

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

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

COPYRIGHT

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

PRODUCTION INFORMATION

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

Int. J. Communications, Network and System Sciences, 2009, 7, 583-591
doi:10.4236/ijcns.2009.27065 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

A Scalable Architecture Supporting QoS Guarantees Using Traffic Engineering and Policy Based Routing in the Internet

Priyadarsi NANDA[1], Andrew SIMMONDS[2]
[1]*School of Computing and Communications, Engineering and IT*
University of Technology, Sydney, Australia
[2] *Mathematics, Computing and Technology, the Open University, England, UK*
Email: *pnanda@it.uts.edu.au, ajs2244@tutor.open.ac.uk*
Received February 25, 2009; *revised April* 10, 2009; *accepted June* 21, 2009

ABSTRACT

The study of Quality of Service (QoS) has become of great importance since the Internet is used to support a wide variety of new services and applications with its legacy structure. Current Internet architecture is based on the Best Effort (BE) model, which attempts to deliver all traffic as soon as possible within the limits of its abilities, but without any guarantee about throughput, delay, packet loss, etc. We develop a three-layer policy based architecture which can be deployed to control network resources intelligently and support QoS sensitive applications such as real-time voice and video streams along with standard applications in the Internet. In order to achieve selected QoS parameter values (e.g. loss, delay and PDV) within the bounds set through SLAs for high priority voice traffic in the Internet, we used traffic engineering techniques and policy based routing supported by Border Gateway Protocol (BGP). Use of prototype and simulations validates functionality of our architecture.

Keywords: QoS, BGP, Policy Based Routing, Traffic Engineering, Bandwidth Broker (BB)

1. Introduction

The success of the Internet has brought a tremendous growth in business, education, entertainment, etc., over the last four decades. With the dramatic advances in multimedia technologies and the increasing popularity of real-time applications, end-to-end Quality of Service (QoS) support in the Internet has become an important issue, which in this paper we address using Traffic Engineering and Policy Based Routing using BGP (Border Gateway Protocol), the core routing protocol of the Internet.

The Internet can be considered as a connection of Autonomous System (AS) domains, where each AS domain controls traffic routing in their own domain based on their own policies. These policies are defined to benefit the AS domain without consideration of other AS domains, which may result in policy conflicts while establishing a flow to achieve a certain degree of QoS on an end-to-end basis. Traffic Engineering concerned with resource allocation mechanisms has been widely studied [8,11–13] and also by us with a proposal for an integrated architecture bringing routing and traffic engineering along with resource management to support end-to-end QoS in the Internet [1]. The novelty of our scheme is mapping traffic engineering parameters into QoS paths available in the network and using policy routing to support end-to-end QoS. This is discussed in terms of the architecture of Figure 1 in Section 2 and how our schemes can be used to achieve some well known QoS objectives such as Delay, Throughput and Packet Delay Variation (PDV) for high priority voice traffic in the Internet. We conducted simulations to validate our results.

We introduce our architecture in Section 2 in order to guide the reader in understanding where traffic engineering and policy routing are used. In Section 3 we highlight the use of a Bandwidth Broker (BB), which is also part of our proposed architecture, to manage interdomain resources. Section 4 discusses our traffic engineering model reflecting the objectives for end-to-end QoS. Policy routing using Border Gateway Protocol

(BGP) is presented in Section 5. Simulation results to validate our model are discussed in Section 6 and finally our conclusion is given in Section 7.

2. An Integrated Architecture

In order to achieve a better service oriented model for the Internet, we propose a three layer policy based architecture for the Internet. The main functions of the architecture are presented in Figure 1.

One of the key components of our architecture is to separate out the control plane from the data forwarding plane by hierarchically grouping network management functions.

In this architecture, layer 3 end-to-end QoS, would be responsible for policy based routing and traffic engineering to dynamically provision bandwidth between different domains. Having determined the route, the layer 3 policy agent would inform the layer 2 of the preferred route. This route provisioning provides a connectivity overlay on top of the normal IP routing, such that if the route from Domain *A* to Domain *B* changes at the IP layer it is not necessary to change the overlay routing. The fall back position for a null layer 3 is that routes will be statically provisioned between individual domains so as to carry the flow to the destination domain.

Layer-2, Network Level QoS. The management unit in this layer is a Bandwidth Broker (BB) [2,3,14]. This interfaces to layer 1 and 3 devices, but also supports inter-domain resource control functions in cooperating with BBs in neighboring domains. Note that the policy function is an add-on to the BB function, i.e. with a null policy to accept everything, BBs can support end to end QoS, but any domain which wishes to implement network policies can do so to its benefit without affecting the functionality of the BB layer.

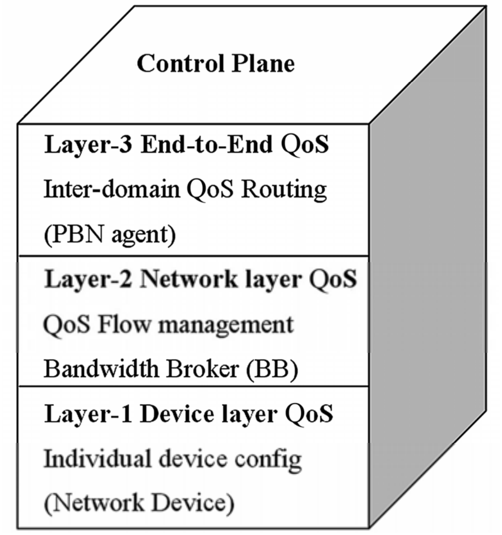


Figure 1. Logical view of the architecture.

The inclusion of null policies and layers is important to enable a gradual take-up of these tools in the Internet. It is not necessary for all domains to implement all levels before anything can work. We present the prototype of our BB design in Section 3 of this paper.

Layer 1, Device Level, is where network devices are configured to support the QoS levels agreed on in the higher levels, getting their instructions from higher layers in the architecture. One possible QoS mechanism being Differentiated Services (DiffServ) (RFC 2475) with Common Open Policy Service (COPS) [13] (RFC 2748, RFC 3084) and being used for signaling. Units in this layer are network devices such as routers and switches and the operation is purely intra-domain.

3. Bandwidth Broker (BB) Design

The conventional definition [2,3] of a Bandwidth Broker (BB) is an agent, running in an Autonomous System (AS), which manages resources within its own domain and with adjacent BB domains, to provide Quality of Service (QoS) support for traffic flows across multiple domains. BBs use hop-by-hop based routing to negotiate with other BBs (the inter-domain function) to provide agreed levels of service for selected traffic flows. Flows getting this preferential treatment will normally be expected to pay more, and this is expected to be a driver in sharing Internet resources as well as providing a revenue stream for Internet Service Providers (ISPs).

A BB controls the network devices in its own domain (the intra-domain BB function) which provide QoS functionality, such as routers and switches. Note that for scalability it is best if the core routers have as little to do as possible apart from forwarding packets, so there should be no interaction between a BB and the core routers. As no particular QoS mechanism is linked to the BB function, different domains can run different QoS mechanisms if they choose. As long as BBs can communicate with each other and agree on common definitions for the level of service required by different priority flows, then a consistent level of QoS support can be set up across different domains for a particular flow. When a new request for a particular QoS arrives, BBs pass the request from one to another, such that if resources are free all along the chain from source to destination then the request is allowed, else it is rejected.

We developed a prototype for a simpler BB architecture and signaling protocol which we believe can be implemented easily. A BB is a resource manager, the resource often being taken as simply bandwidth (BW), as in our prototype, but it could be high quality (e.g. low delay or low jitter or low loss links), buffers, or even low cost, low quality links. The six traffic classes we use for sake of example, in descending priority with binary values for the DiffServ field [15], are:

1) Network traffic – 11100000 (used for BB signaling)

2) Expedited Forward (EF) - 1011 10xx (used e.g. for VoIP)

3) Assured Forward Gold (AFg) - 0111 10xx, AF33

4) Assured Forward Silver (AFs) - 0101 10xx, AF23

5) Assured Forward Bronze (AFb) - 0011 10xx, AF13

6) Best Effort (BE) - 0000 00xx, default

RFC 2597 [16] defines the Assured Forward “Olympic” Per Hop Behavior (PHB) classes and RFC 3246 [17] the EF PHB class. A drop precedence of 3 was chosen for the AF values for compatibility with the deprecated TOS field of the IP packet header, giving flag settings for (D = 1) low delay, (T = 1) high throughput, and (R = 0) normal reliability.

The resources monitored in our implementation are simply additive, but statistical multiplexing could be used to carry more paying traffic over reserved links, as [18] suggests. Our current implementation is open loop, that is available resources are entered in a database (DB) and the BB subtracts resources from the available total as requests are granted, and adds resources when flows finish. Eventually the aim is to have closed loop control, by deploying a resource discovery mechanism to actually measure queue length, etc., e.g. as proposed by one of us using Fair Intelligent Admission Control (FAIC) [19].

The design philosophy we chose is one we believe is consistent with the design philosophy of the Internet: where we faced a design choice we chose the simplest solution, and we implement a minimum function set which can then be extended to provide added functionality.

4. Traffic Engineering Issues

An important objective of Internet traffic Engineering is to facilitate reliable network operations by providing proper QoS to different services through mechanisms which will enhance network integrity and achieve network survivability. The objective of traffic engineering measures in our architecture is to achieve load balancing between neighboring ASs using BGP parameters. By doing so, the architecture then optimizes resource utilization across multiple links, maps divergent QoS parameters to the paths which can support end-to-end service qualities, and avoids congestion hot-spots across the Internet.

In our architecture we used BGP routing to send traffic between domains. But BGP routing policies are not designed specifically to address traffic engineering issues in the Internet. Instead, they are designed to support routing policies determining network reachability between ASs. Obtaining a globally optimized routing path in the Internet is a difficult task due to different policy requirements. Our aim to achieve a scalable solution is based on the following assumptions while incorporating traffic engineering into the architecture:

1) The use of community attributes in policy routing to add extra policy information into the BGP path announcements, enabling traffic engineering to map different QoS parameters to the available paths computed using policy routing.

2) That load balancing traffic with different policies across multiple available routes to the same destination is performed only when the policy co-ordination algorithm for a specific path fails.

Hence our proposed traffic engineering solution can be stated as parameter mapping to different QoS paths available in the Internet, using a policy co-ordination algorithm to resolve any policy conflicts between different ASs while selecting a QoS routing path. In order to be more specific on the issue of parameter mapping, we identified three important parameters related to real-time services such as VoIP application:

a) Bandwidth: When different bandwidth capacities are available in different AS domains for a specific policy in an end-to-end QoS path, the BW allocated is the BW of the AS with the minimum available BW. This minimum bandwidth also needs to satisfy the performance requirements for VoIP traffic in order for the path to be selected.

b) Delay: Two components of end-to-end delay are important for VoIP traffic: delay due to codec processing and propagation delay. ITU-T recommendation G.114 [4] recommends one way delay values less than 150 ms for most user applications, 150 to 400 ms for international connections, with more than 400 ms deemed to be unacceptable. ASs can indicate end-to-end delay in their own domain between edge routers. Hence, complete end-to-end delay for a QoS path would be the sum of all the delays offered by individual AS provided that the sum satisfied the delay requirements specified by G.114. An AS receiving the path announcement along with the delay value from its neighbor adds its own delay and then announces the sum to other ASs further along.

c) Packet Delay Variation (PDV): as it is now properly called rather than jitter, affects real time services, e.g., voice and video traffic. For non real-time voice and video traffic PDV can be removed by a buffer in the receiving device. However if the PDV exceeds the size of the PDV buffer, the buffer will overflow and packet loss will occur. PDV is caused by queuing and serialization effects on the packet path, and is defined by the IETF (RFC 3393) as the difference in delay between successive packets, ignoring any delays caused by packet loss. The one-way delay being timed from the beginning of the packet being sent at the source to the end of the packet being received at the destination. To clarify further, if consecutive packets leave the source AS domain with time stamps t1, t2, t3, …, tn and are played back at the destination AS domain at times t1’, t2’, t3’, …, tn’, then

$$\textbf{Maximum PDV} = \text{Max} \{\text{Abs} [(t_n' - t_{n-1}') - (t_n - t_{n-1})], \ldots, \text{Abs}[(t_2' - t_1') - (t_2 - t_1)]\} = \text{Max} \{\text{Abs} [(t_n' - t_n) - (t_{n-1}' - t_{n-1})], \ldots, \text{Abs}[(t_2' - t_2) - (t_1' - t_1)]\}$$

PDV can also be signed, where a positive PDV indicates that the time difference between the packets at the destination is more than that at the source, and vice-versa.

Hence, while mapping QoS parameters such as bandwidth (BW), Delay (d), and PDV (j) for a specific QoS path, traffic engineering considers the following, where $1 \le i \le k$ are the ASs involved in the end to end path:

BW = Min {BW1, BW2, … BWk}
Delay = Sum (d1, d2, … dk)
PDV = Max{Abs (j1, j2, …., jk)}

And minimizing cost over all the announced path would be given by:

$$\textbf{Min } [C1|P_1 - A_1| + C2|P_2 - A_2| + \ldots \quad Ck|P_k - A_k|],$$

where P is the required policy parameter, A is the announced value of the policy parameter by a neighbor which exported the path and C is the cost associated with these parameters which determines the weight for them. Such costs are important to consider when different ASs have different QoS objectives to satisfy a given Service Level Agreement (SLA) for their customers. In a standard traffic engineering problem, the aim is to minimize the maximum utilization of links, whereas in our architecture it is to maximize the number of AS domains which support the above mentioned constraints. Hence traffic with different policies can be distributed among those paths, improving overall traffic engineering objectives by using the traffic engineering framework of Section 4 and the policy routing of Section 5.

4. Traffic Engineering Framework

The framework is based upon the fact that ASs must communicate with their neighbors to get a fair picture about which relationships they must hold with them in order to apply specific traffic engineering policies in their respective domains. At the same time, ASs must also protect themselves against route instabilities and routing table growth which may otherwise occur due to misconfigurations or problems in other ASs. Manually configuring routing will of course achieve optimum results if the routing is configured optimally. However, Internet routing is complicated so manually configuring routing will not achieve optimal routing in practice, and misconfigurations may well cause catastrophic failure to the Internet. Hence we seek an automatic solution. The components of our traffic engineering framework are presented in Figure 2.

The middle layer (network layer QoS) of our architecture presented in Section 2 has the necessary components for including network policies in traffic engineering.

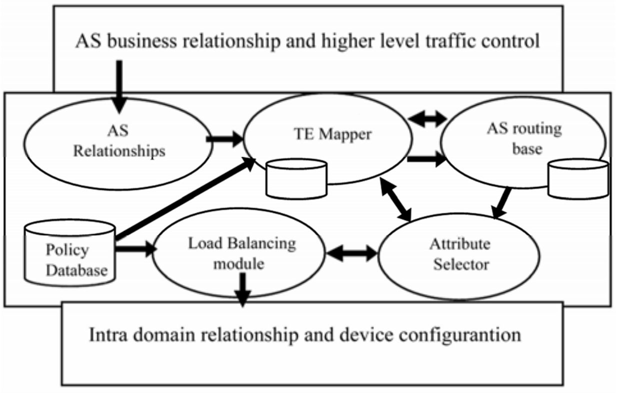

AS business relationship and higher level traffic control
AS Relationships
TE Mapper
AS routing base
Policy Database
Load Balancing module
Attribute Selector
Intra domain relationship and device configurantion

Figure 2. Framework of traffic engineering.

AS relationships play an important role supporting QoS in the Internet. But obtaining data on such relationship is a difficult task, as ASs such as ISPs may not reveal such data to their competitors. Hence we propose to use a measurement based approach where an ISP ranks ASs based on the frequency of their presence in the routing table. A heavily used AS in the path list is one where some kind of traffic engineering should be applied if selected for next hop forwarding. For example the decision of selecting local preference is very much local to an ISP in order to balance its outgoing traffic (selecting the path to forward packets to the next ISP). On the other hand, an AS which is used less frequently is less congested and has a better chance of providing QoS resources [5].

Traffic Engineering Mapper (TEM) has a repository that holds AS relationships and the hierarchy for interconnectivity between various ASs. TEM is responsible for directing those relationships to the Attribute Selector as well maintaining a list of those attributes once selected. Because the routing table holds information regarding import and export policy filters, as well the attributes associated with them, TEM also investigates their validity in the AS routing base. One of the export rules based on the business relationship between ASs is for the TEM to enforce the provider to send all routes (customer as well as provider routes) that the provider knows from its neighbors. Alternatively, TEM could ensure that peer or provider routes are not sent when sending routes to another provider (i.e. just send customer routes). TEM is an essential component of traffic engineering framework.

Finally, the decision on traffic engineering is taken by the Load Balancing module which receives necessary inputs regarding which attributes are to be applied and to which paths they must be applied. The policy database holds policy information regarding how the AS may change routing for its own domain. Also included in the policy database is information on a list of remote ASs which are also customers of this AS, and pricing structures imposed by the SLAs of its providers. Such information is given to the load balancing module which then takes a final decision on traffic engineering. The process is the same for both importing and exporting a route between neighboring ASs.

Several efforts on finding solutions to BGP based traffic engineering and AS relationships have been explored in the past [6–9]. While the authors described some drawbacks of BGP in the first instance and then proposed their schemes on better management of BGP for traffic engineering, our approach is different as we consider the relationship between ISPs as a central issue in defining necessary traffic engineering policies for the Internet, and add a community policy attribute to BGP to solve this issue. Hence our proposal builds on BGP to provide a solution. Policy routing using BGP is presented in the following section of this paper.

5. Policy Routing

Routing protocols play an important role in exchanging routing information between neighboring routers. Such information may be used to update routing tables and to share information about network status so that traffics to appropriate destinations will be set up quickly, efficiently and achieve the required QoS between end systems. Different types of routing protocols are in widespread use across the Internet. Apart from determining optimal routing paths and carrying traffics through the networks, these routing protocols should have additional functionalities such as resource discovery, policy mapping and policy negotiation mechanisms to support network policies, traffic engineering and security.

BGP is a path vector protocol that uses AS path information between neighboring routers in different AS domains to determine network reachability. Such network reachability information includes information on the list of ASs and the list of AS paths. One of the important features supported by BGP is policy routing, where an individual AS can implement network policies to determine whether to carry traffic from different users (mostly users from other ASs) with diverse QoS requirements. Such network policies are not part of BGP, but provide various criteria for best route selection when multiple alternative routes exist and help to control redistribution of routing information, resulting in a rich support by BGP for policy routing and traffic engineering in the Internet.

Current Internet Traffic Engineering depends heavily on both Intra and Inter Domain routing protocols using network policy in order to configure the routers across various domains. The support for policy based routing using BGP can provide source based transit provider selection, whereby ISPs and other ASs will then route traffic originating from different sets of users through different connections across the policy routers. Also QoS support for Diffserv networks can be supported using policy routing through the use of the DiffServ field in the IP packets. Hence, a combination of traffic engineering for load balancing across network links offered by destination based routing, and policy based routing, can enable implementation of policies that distribute traffic among multiple paths based on traffic characteristics.

Policy routing in the Internet can be based on the following principles:

1) Each AS to take action on routing based upon information received from neighbors. Such decision process is central within each AS.

2) Neighbors are free to negotiate any policy conflict by adjusting their traffic parameters and waiting for confirmations from all the domains involved in routing.

3) Incorporation of a direct relationship between network level flow management and traffic engineering objectives.

Routing traffic across several routers in the same domain to support QoS between the edges of the network is relatively easy to achieve, as we can gather knowledge on QoS paths and select edge routers administrated by a single network entity. But inter-domain QoS path selection is difficult to achieve and to demonstrate how we can approach such a problem, we present the policy routing framework in Figure 3. We assume that the intra-domain QoS path computations are already optimized based on the local knowledge of intra-domain routing protocol and this information is already stored in layer-2 of our architecture.

Standard BGP routing process involves applying an import policy onto routes received from neighbors, deciding the best route based on BGP routing decision process [10] and then applying export policy to the computed routes before announcing to neighbors. Such a process does not take all policy decisions into account, particularly while computing the routing paths in support of QoS in the Internet. The inter-domain route selector which is central to the routing module within an AS domain receives path announcements from the neighbors through the inbound route announcement. Apart from applying standard BGP decision process on selecting certain route advertisement from its neighbor, the route

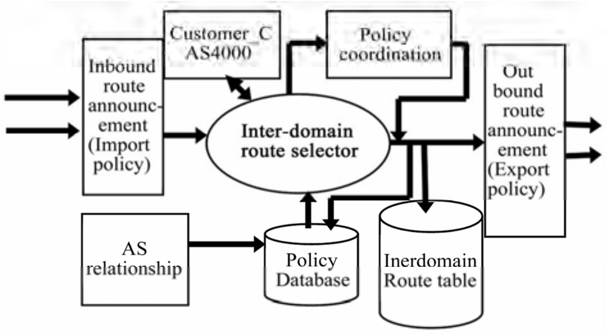

Customer_C AS4000
Policy coordination
Inbound route announcement (Import policy)
Inter-domain route selector
Out bound route announcement (Export policy)
AS relationship
Policy Database
Inerdomain Route table

Figure 3. Interaction between routing components for policy based routing.

selector needs further consultation for QoS path selection by interacting with the following components:

- It is important to decide which types of neighbor (e.g., provider, customer or peer) the route advertisement came from and based on that, the AS will then decide whether to announce the path to its neighbor. Such relationships are held in a policy database which then inputs the information to the route selector.
- The route selector gets path information within its own domain by communicating with the intra-domain QoS path repository. Actions such as changing values for LOCAL_PREF, MED, IGP Cost, and Pre-pending AS_PATH results in directing incoming traffic to a specific edge router.
- The decision process also needs to consider which QoS policies are supported by the AS domain which sent such path announcements. For this, each AS, which can support different policies in relation to QoS services (e.g., Premium, Gold, Silver, Bronze), adds a "COMMUNITY" policy attribute along with the path announcement.
- In case of policy mismatch i.e., advertised policies by neighbor does not match with the AS's own policy, the route selector will apply "policy co-ordination algorithm" (Subsection 5.1) to resolve such conflict.

Finally routes selected by either the route selector without any policy mismatch, or applying policy co-ordination algorithm in case of any policy conflict, are further announced to ASs through outbound route announcement. The announced route is stored in the AS's inter-domain routing table.

5.1 Policy Co-Ordination Algorithm

An algorithm performing such functions is presented below:

```
Get list of policies from neighbor
For each neighbor policy {
  Compare policy support with own policy list
  If match {
    Set values and put policy in End-to-End (E-E) list
  }
  Else (no match) {
      Tag policy as non-confirmed and put policy in Temp list
    }
  } (All policies checked)
  For all policies in the Temp list {
  Check if another route satisfies policy constraints
  If match {
      Set values and put policy in E-E list
    } (End the process of policy comparison)
    Else (no match) {
    For all policy mismatch {
      Adjust own policy and apply traffic engineering parameters for new policy
      Select the ones which contribute to maximum revenue
      Announce all paths to neighbors in the list
}
      Set values and put policy in E-E list
   } (End the process of policy adjustment)
} (Temp list emptied)
```

Finally in order to validate our algorithm and functional models, we conducted a series of experiments using OPNET based simulation to take into account the effect of traffic engineering and policy routing which are presented in the next section of this paper.

6. Simulation Results

In order to validate our algorithm and functional models we performed a series of experiments and obtained various statistics from the simulation. The topology and the default routing paths between customers A, B and C are presented in Figure 4 below:

A-B - - -, A-C ▬ and B-C -··-··

As presented in Figure 4, the network is created by configuring all default values into the devices and network reachability test is performed to ensure end-to-end connectivity between each AS domains in the network. Once these are performed, based on routing table entries, we performed our analysis on how BGP paths are recorded between different AS domains without any policy but with its default routing decision process. The complete network diagram is presented in Figure 5 below which also presents end-to-end connectivity between all the domains.

Our second scenario in Figure 6 demonstrates the effect of our proposed policy mechanism compared with the base-line scenario in Figure 4. The end-to-end path between customers now have different routes as a result of policy enforcements across all the AS domains.

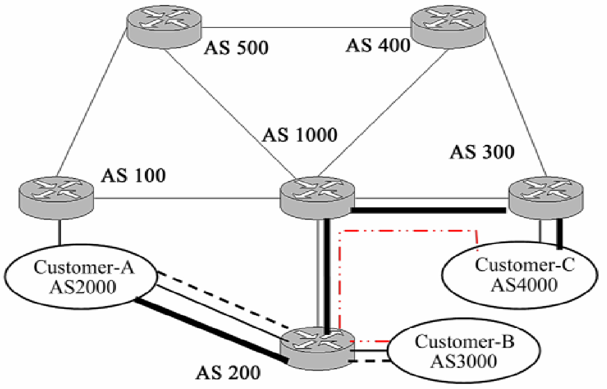
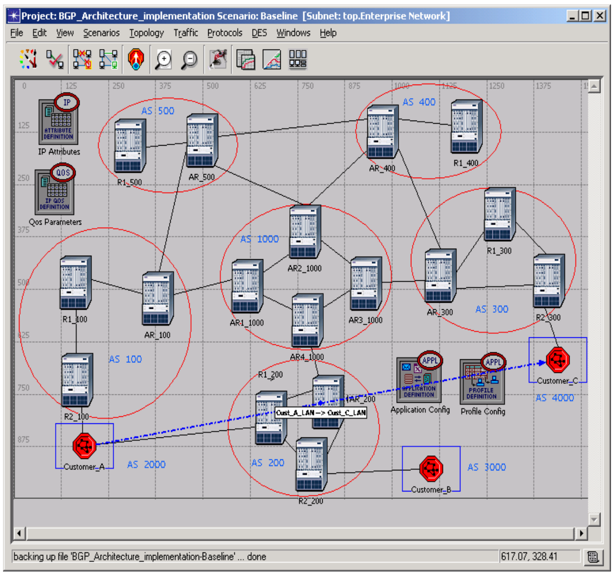


Figure 4. Simulation topology and default routing paths.



backing up file 'BGP_Architecture_implementation-Baseline' ... done

617.07, 328.41

Figure 5. End-to-end network configuration.

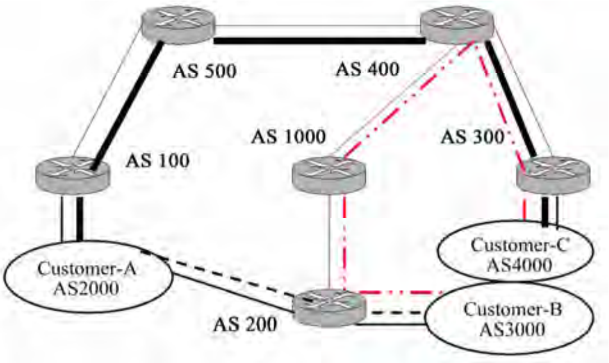


Figure 6. End-to-end path using policy.

The results show effect of our proposed Community attribute for selecting specific QoS domains using BGP routing process between end nodes. Such a scheme not only balances traffic distribution across inter-domain links but also fine tunes traffic engineering for better provisioning of QoS between end domains. However, the scheme does increases complexity in BGP decision process due to extra information involving the community attribute.

Traffic was generated from a G_711 interactive voice source with duration of 1 hour and several experiments were conducted to demonstrate the quality of voice traffic on an end-to-end basis. We assigned a DSCP value of B8 (EF=184) to the VoIP traffic which is then mapped to a BGP community value of 0x00640184 to ensure voice quality is maintained strictly between end domains. A series of graphs representing QoS parameters for VoIP applications are presented through Figure 7 (a-d).

While sending QoS aware applications in the Internet such as VoIP, we are mainly concerned about maintaining delay budget within the limit set for QoS assurance. The plots in Figure 7 (a-d) represent Packet Delay Variation (defined as jitter by OPNET), end-to-end delay, variance of the end-to-end delay and BGP updates, averaged over a 10 minute period for the scenario with policy routing enabled on all the routers running BGP. The actual VoIP traffic starts after 2 minutes and is deliberately set to make sure that BGP timer values are taken into account.

In our experiment, plot (a) demonstrates the variation in packet end-to-end delay (PDV) and shows that it is kept to low bounds (-0.3 µs to +0.1 µs), in spite of activating multiple QoS and routing policy configurations across the whole network. The PDV is influenced by packet scheduling and queuing strategy implemented across the routers (layer-1 functions) in order to support QoS within and across various domains. PDV is reported as the maximum absolute time difference between the instances when successive packets are received at the destination minus the time difference between the instances when these packets are sent at the source, averaged over 10 minutes, which is equivalent to the IETF definition assuming constant packet processing times at the destination.

The end to end delay for VoIP traffic is maintained at a value ≤ 50.4 ms (plot b), well within the SLA of 150 ms, while PDV converges to less than 0.1 µs (Plot a).

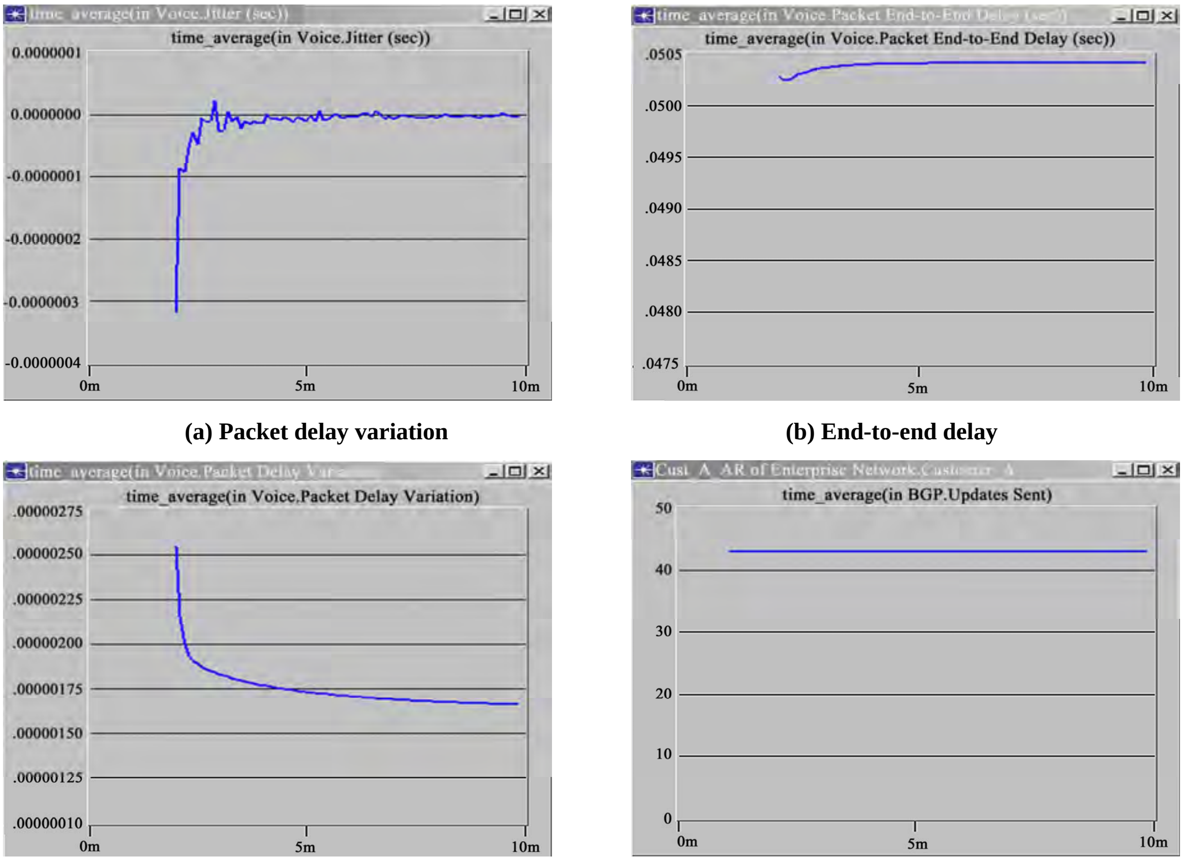
Plot c shows that the variance of the end-to-end delay falls to less than 1.75 µs after 5 min. This is confusingly defined as Packet Delay Variation (PDV) by OPNET, but we will use the IETF definition for PDV.

Plot d presents number of BGP updates. In our simulation the access router in Customer_A network (Customer_A_AR) is the one where most policies related to load balancing and traffic engineering are enforced. For this reason we collected the BGP updates sent by this router which contains either new routes or unfeasible routes or both in the system. In our case this access router sent 43 updates at 69 s due to strong policy enforcement.

As shown above, voice traffic sent between Customer_A network and Customer_C network experienced QoS parameters well within our design limits. However these parameters could be further improved by carefully selecting other QoS strategies within individual domains.

7. Conclusions

In this paper we demonstrated the effect of Internet traffic engineering and use of policy routing to achieve end-to-end QoS for high priority Voice traffic, in the context of our high level architecture of Figure 1. We also presented simulation results to demonstrate how we achieve automatic load balancing between different service providers using a BGP community policy attribute and the policy co-ordination algorithm of Subection 5.1.



(a) Packet delay variation

(b) End-to-end delay

(c) End-to-end delay variance

(d) BGP updates at Customer_A_AR

Figure 7 (a-d) VoIP QoS measurement.

This is substantially different from the default routing which does not select the AS domains based on QoS requirements for an application. Such results are evidence that our scheme improves end-to-end QoS requirements for high priority voice traffic particularly when many other applications are running simultaneously in the Internet.

The objective of our design is how BGP can be used to select QoS domains for QoS support. For this reason we are mainly concerned with AS domain traffic behavior contributing to policy routing and traffic engineering.

8. References

[1] P. Nanda and A. J. Simmonds, “Policy based QoS support using BGP routing,” International Conference on Communications in Computing, CIC’06, Las Vegas, Nevada, USA, CSREA Press, ISBN 1–60132–012 –4, pp. 63–69, June 26–29, 2006.

[2] B. Teitelbaum, “QBone bandwidth brokerarchitecture,” [Online]Available:http://qbone.internet2.edu/bb/bboutline 2.html.

[3] K. Nichols, V. Jacobson, and L. Zhang: “A two-bit differentiated services architecture for the internet,” IETF RFC 2638, July 1999.

[4] ITU–T Recommendation G.114, One way transmission time, 1996.

[5] A. D. Yahaya and T. Suda, “iREX: Inter–domain QoS automation using economics,” IEEE Consumer Communications and Networking Conference, CCNC’06, USA, January 2006.

[6] N. Feamster, J. Winick, and J. Rexford, “A model of BGP routing for network engineering,” SIGMETRICS/ Performance’04, New York, USA, June 12–16, 2004.

[7] D. O. Awduche, A. Chiu, A. Elqalid, I Widjaja, and X. Xiao, “A framework for internet traffic engineering,” Draft 2, IETF, 2002.

[8] B. Quoitin, S. Uhlig, C. Pelsser, L. Swinnen, and O. Bonaventure, “Internet traffic engineering with BGP,” Quality of Future Internet Services, Springer, 2003.

[9] G. Di Battista, M. Patrignani, and M. Pizzonia, “Computing the types of relationships between autonomous systems,” IEEE INFOCOM, 2003.

[10] Y. Rekhter and T. Li, “A border gateway protocol 4 (BGP-4),” Internet draft, draft-ietf-idr-bgp4-17.txt, work in progress, January 2002.

[11] B. Quoitin, C. Pelsser, O. Bonaventure, and S. Uhlig, “A performance evaluation of BGP-based traffic engineer-

ing," Int'l. J. Network Mgmt, 2005.

[12] R. Yavatkar, D. Pendarakis, and R. Guerin, "A framework for policy-based admission control," RFC 2753, January 2000.

[13] S. Salsano, "COPS usage for Diffserv resource allocation (COPS-DRA)," Internet Draft, October 2001.

[14] P. Nanda and A. Simmonds, "Providing end-to-end guaranteed quality of service over the Internet: A survey on bandwidth broker architecture for differentiated services network," CIT'01, 4th International Conference on IT, Berhampur, India, pp.211–216, 20–23 December 2001.

[15] K. Nichols, S. Blake, F. Baker, and D. Black, "Definition of the differentiated services field (DS Field) in the IPv4 and IPv6 headers," IETF RFC 2474, Dec. 1998.

[16] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, "Assured forwarding PHB group," IETF RFC 2597, June 1999.

[17] B. Davie, *et al*., "An expedited forwarding PHB (per-hop behavior)," IETF RFC 3246, March 2002.

[18] P. Pan, E, Hahne, and H. Schulzrinne, "BGRP: Sink-tree-based aggregation for inter-domain reservations," Journal of Communications and Networks, Vol. 2, No. 2, pp. 157–167, June 2000.

[19] M. Li, D. B. Hoang, and A. J. Simmonds, "Fair intelligent admission control over DiffServ network," ICON'03, 11th IEEE International Conference on Networks, Sydney, Australia, ISSN: 1531–2216, pp. 501–506, 28 Sept–1 Oct 2003.

Int. J. Communications, Network and System Sciences, 2009, 7, 592-599
doi:10.4236/ijcns.2009.27066 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/)

Performance Analysis of a Novel Dual-Frequency Multiple Access Relay Transmission Scheme

Javier DEL SER[1*], Babak H. KHALAJ[2]
[1]*TECNALIA-TELECOM*, 48170 *Zamudio-Bizkaia, Spain.*
[2]*Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran.*
Email: *jdelser@robotiker.es*, *khalaj@sharif.ir*

Received June 26, 2009; *revised July* 31, 2009; *accepted August* 27, 2009

ABSTRACT

In this paper we present the performance analysis of a novel channel assignment scheme where two non-cooperative independent users simultaneously communicate with their destination through a single relay by using only two frequency channels. The analytic derivation of the probability of symbol error for two main relay techniques will be provided, namely Amplify-and-Forward (AF) and Decode-and-Forward (DF). As shown by the obtained results, our switched-frequency approach results in a model that can achieve full-diversity by means of maximum-likelihood decoding at the receiver. Our results are especially important in the DF case, since in traditional techniques (such as half-duplex two-time slot approaches) two sources simultaneously transmit on the same channel through the first time slot, which necessitates some sort of superposition coding. However, since in our scheme both users transmit over orthogonal channels, such a coding scheme is not required. In addition, it is shown that the DF approach based on our novel channel assignment scheme outperforms the AF scheme, especially in scenarios where the relay is closer to the receiver.

Keywords: Multiple Access Relay Channel, Frequency Switching, Non-Cooperative Networks, Maximum Likelihood Decoding

1. Introduction

During the last years the use of relay nodes has attracted a lot of attention in practical areas such as cellular networks, especially in scenarios where multiple antennas cannot be installed in practice at any site. Deploying relay nodes between sender(s) and receiver(s) provides increased spatial diversity in the communication scenario under consideration. The research community has shown a great interest in this field: as to mention, for the single user scenario with relay channels, capacity bounds are computed for Detect-and-Forward (DF) based mechanism in a Rayleigh fading environment [1,2]. The outage capacity bounds for the case of a single user transmission with relay in low signal-to-noise ratio regime is considered in [3], in which frequency division model for the relay channel is assumed. The authors in [4] present the performance limits of Amplify-and-Forward (AF) relay channels for single user scenario, where a new transmission protocol is also proposed in order to achieve full diversity. Recently, the use of practical coding schemes at the relay has also been addressed in [5,6].

In this context, this paper will focus on the performance analysis of a non-cooperative two-source multiple access relay channel (MARC) [7]. In the MARC channel, two independent information sources transmit their data to a common destination aided by a shared relay node, which processes and combines the data from both sources. The achievable rate region for the MARC channel has been studied in [8] by employing a partial detect-and-forward strategy at the relay. In multiuser scenarios, cooperative ideas have also been proposed, so that users could interact with each other in order to improve each other's performance in fading environments [9–13]. In [14], the authors present an upper bound on the diversity-multiplexing trade-off for the single user relay channel. However, their proposed scheme does not achieve full diversity for the whole block transmission period, since samples in the second transmission slot are not protected by relay re-transmissions. The authors also extend their scheme to cooperative multiple access scenarios for the two user case in the absence of any additional relays. However, since such cooperative schemes rely on an inter-user channel which consumes network resources and therefore limits their performance by the condition of such a link [15], in this paper we will only focus on non-cooperative multiple access channels.

Traditionally, the case of multiple users and a relay has been addressed in a number of different ways. The most straightforward approach is to extend single-relay ideas to multiple users, in which each user uses a separate frequency and diversity is achieved by utilizing two time-slots. Such an approach is basically an extension of the well-known delay diversity schemes [16,17] and [18], which will naturally lead to a total of four orthogonal time-frequency channels. Another approach is to consider independent receive channels for signals coming from the users and the relay, also resulting in a total of four independent channels either in time or frequency. It should be especially noted that, although it is possible to extend half-duplex two-time slot Amplify-and-Forward schemes to MARC scenarios [19], the extension of Detect-and-Forward schemes to MARC requires the use of complex superposition-type multiuser coding strategies, so that the relay is able to detect multiple sources over the same multiple access channel [8,19,20,21].

The scheme proposed in this paper does not rely on any special coding scheme, since the users transmit to the relay over orthogonal channels. Our analysis will be focused on examining the error performance of the MARC scenario in the case of full multiplexing gain (two independent sources transmitting simultaneously in the same time slot), where diversity gain improvement can be verified through the slope of error probability curve as a function of received SNR in a log-log scale. As will be shown, by using the proposed frequency-switching channel assignment scheme at the relay (Figure 1), the signal model will be transformed from two independent scalar channels into a 2×2 Multiple-Input Multiple-Output (MIMO) model [22]. Consequently, Maximum Likelihood (ML) detection performed over the received vector will provide an overall diversity order of two for each user.

The structure of the paper is organized as follows. In Section 2, the signal model and the channel assignment scheme will be presented. In addition to Amplify-and-Forward scheme, Detect-and-Forward structure and its corresponding ML detector will be proposed. In Section 3, analytic probability of error computations for both AF and DF schemes will be derived. In Section 4, simulation results for both AF and DF mechanisms for different relay locations will be presented and compared with the analysis. As shown in these results, the DF approach outperforms the AF scheme, especially as the relay gets closer to the destination. Finally, Section 5 concludes the paper.

2. Signal Model

In this paper, we will assume the case of two independent and identically distributed (i.i.d.) random sources S_1 and S_2 which generate binary symbols b_1 and b_2 (b_1, $b_2 \in \{-1, 1\}$), respectively. Those symbols are transmitted in a wireless environment to the same destination. In between these sources and the destination a single dual-frequency relay is located. The distance between sources and destination is normalized to one, and the location of the relay is denoted as d ($0<d<1$). The channel is assumed to be block Rayleigh fading, i.e. the channel is assumed to be fixed within a block and varies independently from block to block. In addition to this Rayleigh model, a propagation loss as a function of distance is considered. This loss is a basic exponential model with exponent n=2, hence the power attenuation is assumed to be equal to K/d^2 at a distance d, where K is the propagation constant [23]. The channel state information (CSI) is only assumed to be known at the receiver locations (i.e. both relay and destination), whereas the transmitters are not assumed to have any knowledge of their forward transmission channels. Both users transmit their signals at two different frequency channels f_1 and f_2, respectively.

Instead of forwarding each received signal over the same incoming frequency channel, the relay of our proposed dual-frequency channel assignment switches the frequencies between the two transmitted signals. The proposed structure is shown in Figure 1, where the source S_1 transmits at frequency channel f_1, and the relay retransmits the same signal over frequency channel f_2. Similarly, the frequency channel of source S_2 is switched at the relay before retransmission to the destination. It can be easily verified that, without such frequency switching at the relay, no additional diversity can be achieved without resorting to delay diversity schemes, since the two signals coming from the source and the relay at each frequency are simultaneously combined at the receiver. As will be subsequently shown, the proposed channel assignment scheme will transform two independent scalar channels into a two dimensional vector channel which achieves a diversity of order two.

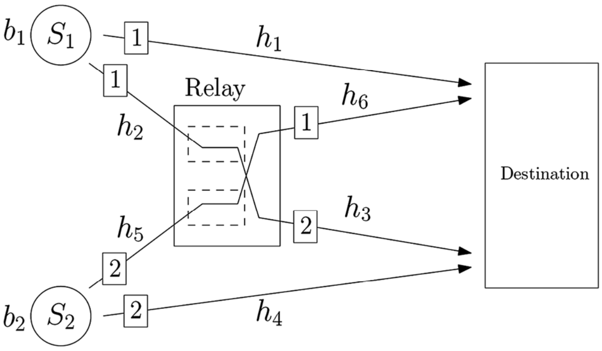


Figure 1. The proposed *switched* frequency assignment at the relay (the number over each link denotes the corresponding frequency channel used).

It should be noted that in the proposed scheme, the relay should perform in a full-duplex mode for each transmission path. In other words, two transceivers should operate simultaneously in the same time slot at the relay. However, since these two transceivers can be stationed separately in hardware and the input carrier frequency of each board is different from its output carrier frequency, the hardware complexity will be significantly less than the traditional full-duplex transceivers that operate on the same carrier frequency for both their input and output signals. In addition, the aforementioned assumption would lead to echo-interference in case of highly asymmetric receive and transmit power. Such an issue could be overcome by applying preprocessing and postprocessing techniques as done, for instance, in [24]. Nevertheless, the echo interference will be assumed to be negligible at the relay, since its suppression falls out of the scope of our contribution.

In the next subsections we present the two previously mentioned relaying schemes, DF and AF, particularized for our proposed setup.

2.1. Mplify-and-Forward ML Detector at the Relay

Assuming that signals transmitted by each source are denoted by b_1 and b_2, and that all information symbols from source and relay stations reach the common destination in the same time slot (as done, for instance, in [25, 5]), the received signal for each frequency will be given by

Frequency channel f_1:

$$y_1 = b_1 h_1 + \left(b_2 h_5 + n_2\right)\gamma_1 h_6 + n_3 \quad \textbf{(1)}$$

Frequency channel f_2:

$$y_2 = b_2 h_4 + \left(b_1 h_2 + n_1\right)\gamma_2 h_3 + n_4 \quad \textbf{(2)}$$

where h_1 and h_4 denote the channel coefficients between sources S_1 and S_2 and the destination, h_2 and h_5 denote channel taps between sources S_1 and S_2 and the relay, and h_3 and h_6 denote the channel weights between relay and destination at the two transmit frequencies f_1 and f_2, correspondingly. All channel coefficients $\{h_i\}_{i=1}^{6}$ are modeled as complex Gaussian random variables with zero mean and variance per dimension equal to K/d^2 (due to the propagation loss at the distance d). In addition, complex circularly symmetric zero mean additive white Gaussian noise $\{n_i\}_{i=1}^{4}$ are assumed for receive inputs at both relay and destination, with variance per dimension $\sigma^2 = {}^{N_0}\!/_{2}$. Also, the quantities γ_1 andγ_2 denote the gain of the relay for each frequency channel f_1 and f_2, whose values are chosen such that the relay transmits with unit average power.

A closer look at the above equations reveals that, by the proposed switching algorithm at the relay, the channel is transformed into a 2×2 MIMO channel model as given by

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} h_1 & \gamma_1 h_5 h_6 \\ \gamma_2 h_2 h_3 & h_4 \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} \gamma_1 h_6 n_2 + n_3 \\ \gamma_2 h_3 n_1 + n_4 \end{pmatrix} \quad \textbf{(3)}$$

As is well-known in MIMO literature, the above model can be solved in the context of standard spatial multiplexing MIMO systems, where the detection based on the Maximum Likelihood (ML) criterion yields a diversity of order two [26]. Consequently, in our approach we use a ML detector at the destination that will jointly estimate b_1 and b_2 from the received signals y_1 and y_2. It should also be observed that in our model, some entries of the channel matrix shown in Equation (3) are multiplications of Rayleigh variables, and therefore the resulting model is not roughly in the conventional form spatial multiplexing models. Nevertheless, the obtained results show that an increased diversity gain of order close to two will still be achieved for both users in the DF approach.

2.2. Detect-and-Forward ML Detector at the Relay

Instead of just amplifying and forwarding each signal, in this case the relay detects the source symbol from the received signal before retransmitting it. Consequently, in the DF approach the received signal at destination will be given by

Frequency channel f_1:

$$y_1 = b_1 h_1 + \hat{b}_2 \gamma_1 h_6 + n_3 \quad \textbf{(4)}$$

Frequency channel f_2:

$$y_2 = b_2 h_4 + \hat{b}_1 \gamma_2 h_3 + n_4 \quad \textbf{(5)}$$

where $\hat{b}_1$ and $\hat{b}_2$ denote the output of the detector at the relay for sources S_1 and S_2, respectively.

It should be noted that in this case, the maximum likelihood detector at the destination should also consider the effect of detection errors at the output of the relay. Such errors are mainly due to fading events in the source-relay links: when one of these links is affected by a deep fade, the detection errors committed at the relay are propagated to the destination. In order to account for both source-relay and relay-destination links, an end-to-end ML detector should be utilized. The associated end-to-end search criterion can be derived by first modeling the source-relay link as a binary symmetric channel (BSC) with probability of error equal to

$$P_e^1 = Q\left(\frac{|h_2|}{\sigma}\right) \tag{6}$$

$$P_e^2 = Q\left(\frac{|h_5|}{\sigma}\right) \tag{7}$$

where $Q(\bullet)$ denotes the standard Q-function. The estimated b_1 and b_2 values at the final destination, denoted by $\tilde{b}_1$ and $\tilde{b}_2$, are then computed by maximizing the likelihood function

$$\begin{aligned}(\tilde{b}_1,\tilde{b}_2) &= \underset{b_1,b_2\in\{-1,1\}}{\arg\max}\; p(y_1,y_2\,|\,b_1,b_2)\\ &= \underset{b_1,b_2\in\{-1,1\}}{\arg\max}\sum_{\hat{b}_1,\hat{b}_2} p(y_1,y_2\,|\,b_1,b_2,\hat{b}_1,\hat{b}_2)\,p(\hat{b}_1,\hat{b}_2\,|\,b_1,b_2\end{aligned} \tag{8}$$

where

$$p(\hat{b}_1,\hat{b}_2\,|\,b_1,b_2) = \begin{cases}(1-P_e^1)(1-P_e^2) & \textit{if } \hat{b}_1=b_1,\,\hat{b}_2=b_2.\\ P_e^1(1-P_e^2) & \textit{if } \hat{b}_1\neq b_1,\,\hat{b}_2=b_2.\\ (1-P_e^1)P_e^2 & \textit{if } \hat{b}_1=b_1,\,\hat{b}_2\neq b_2.\\ P_e^1P_e^2 & \textit{if } \hat{b}_1\neq b_1,\,\hat{b}_2\neq b_2.\end{cases} \tag{9}$$

It should also be remarked that, since the proposed scheme is working on a single-slot basis, it is assumed that the decoding delay at the relay is negligible with respect to symbol time intervals. Therefore, the relay is able to start the retransmission of the detected symbol after some small delay during the same time slot.

3. Analysis

In the following section a detailed derivation of the analytic probability of error for both schemes is provided. We begin by analyzing the AF approach.

3.1. Analysis of Amplify-and-Forward ML Detector at the Relay

In order to compute the analytic probability of error for the AF approach, we first rewrite Equation (3) as the sum of a signal term **u** and a noise term **w**, i.e. **y**=**u**+**w**, where **u** is given by

$$\mathbf{u} = \begin{pmatrix} h_1b_1+\gamma_1h_5h_6b_2\\ \gamma_2h_2h_3b_1+h_4b_2\end{pmatrix} \tag{10}$$

and the noise term **w** will be zero mean with a covariance expressed as[1]

$$\mathrm{E}\left\{\mathbf{w}\mathbf{w}^*\right\} = \begin{pmatrix} N_1 & 0\\ 0 & N_2\end{pmatrix} \tag{11}$$

where $N_1 \square N_0(1+\gamma_1^2|h_6|^2)$ and $N_2 \square N_0(1+\gamma_2^2|h_3|^2)$. Since the two terms of the noise vector **w** do not present the same variance, we will employ a scaling matrix M so as to transform **w** into a unit variance vector $\mathbf{w}' = M\mathbf{w}$ such that $\mathbf{E}\left\{\mathbf{w}'\mathbf{w}'^*\right\} = \mathbf{I}_{2\times 2}$, yielding

$$M = \begin{pmatrix}\frac{1}{\sqrt{N_1}} & 0\\ 0 & \frac{1}{\sqrt{N_2}}\end{pmatrix}. \tag{12}$$

As previously mentioned, the values of the relay gains γ_1 and γ_2 are set such that the average relay transmit power is normalized to one, yielding

$$\gamma = \frac{1}{\sqrt{|h_2|^2+|h_5|^2}} \tag{13}$$

We will henceforth denote the received vector corresponding to $(b_1,b_2)=(1,1)\square\,\mathbf{x}_A^{\mathrm{T}}$ by $\mathbf{u}_A$. Similarly, vectors $\mathbf{u}_B$, $\mathbf{u}_C$, and $\mathbf{u}_D$ correspond to $\mathbf{x}_B^{\mathrm{T}}\square(-1,1)$, $\mathbf{x}_C^{\mathrm{T}}\square(1,-1)$ and $\mathbf{x}_D^{\mathrm{T}}\square(-1,-1)$, respectively. Assuming that the possible transmit vectors (b_1,b_2) are equiprobable, the total probability of error is thus given by

$$\begin{aligned}P_e &= \frac{1}{4}\big(\Pr\{\text{error}|\,(b_1,b_2)\\ &= \mathbf{x}_A^{\mathrm{T}}\}+\Pr\{\text{error}|\,(b_1,b_2)\\ &= \mathbf{x}_B^{\mathrm{T}}\}+\Pr\{\text{error}|\,(b_1,b_2)\\ &= \mathbf{x}_C^{\mathrm{T}}\}+\Pr\{\text{error}|\,(b_1,b_2)=\mathbf{x}_D^{\mathrm{T}}\}\big)\\ &= \frac{1}{4}\sum_{i=A}^{D}\Pr\{\text{error}|\,(b_1,b_2)\\ &= \mathbf{x}_i^{\mathrm{T}}\}.\end{aligned} \tag{14}$$

Let us consider the error term corresponding to $\mathbf{x}_A^{\mathrm{T}}$. Given a set of channel coefficients $\{h_i\}_{i=1}^6$, we will use the union bound to compute the probability of error when $\mathbf{x}_A^{\mathrm{T}}$ was sent, resulting in

$$\begin{aligned}&\Pr\left\{\text{error}|\,(b_1,b_2)=\mathbf{x}_A^{\mathrm{T}},\{h_i\}_{i=1}^6\right\}\\ &\leq \Pr\left\{\mathbf{x}_A \;\rightarrow\; \mathbf{x}_B\;|\,\{h_i\}_{i=1}^6\right\}\\ &+\Pr\left\{\mathbf{x}_A \;\rightarrow\; \mathbf{x}_C\;|\,\{h_i\}_{i=1}^6\right\}\\ &+\Pr\left\{\mathbf{x}_A \;\rightarrow\; \mathbf{x}_D\;|\,\{h_i\}_{i=1}^6\right\}\end{aligned} \tag{15}$$

where $\Pr\left\{\mathbf{x}_i\rightarrow\mathbf{x}_j\,|\,\{h_i\}_{i=1}^6\right\}$ denotes the probability that, given a set of channel coefficients $\{h_i\}_{i=1}^6$, $\mathbf{x}_i$ is transmitted and $\mathbf{x}_j$ is detected at the receiver. Let us consider the term $\Pr\left\{\mathbf{x}_A\rightarrow\mathbf{x}_B\,|\,\{h_i\}_{i=1}^6\right\}$. This pairwise probability of

[1] $\mathrm{E}\{\cdot\}$ denotes mathematical expectation of a random variable.

error can be obtained by using the transformation in Expression (12), yielding

$$\begin{aligned}&\Pr\left\{\mathbf{x}_i \to \mathbf{x}_j \mid \{h_i\}_{i=1}^{6}\right\}\\ &= Q\left(\frac{\| M(\mathbf{u}_A - \mathbf{u}_B)\|}{\sqrt{2}}\right)\\ &= Q\left(\sqrt{2\left(\frac{|h_1|^2}{N_1}+\frac{\gamma^2 |h_2|^2 |h_3|^2}{N_2}\right)}\right)\end{aligned} \qquad \textbf{(16)}$$

where $\|\cdot\|$ denotes the Frobenius norm. Consequently, the union bound for the probability of error $\Pr\left\{\text{error} \mid (b_1, b_2) = \mathbf{x}_A^{\mathrm{T}}\right\}$ can be computed by adding the other terms in Expression (15) and taking the expectation over channel coefficients h_i, i.e.

$$\begin{aligned}&\Pr\left\{\text{error} \mid (b_1, b_2) = \mathbf{x}_A^{\mathrm{T}}\right\}\\ &\leq \mathrm{E}_{h_i}\left\{Q\left(\sqrt{2\left(\frac{|h_1|^2}{N_1}+\frac{\gamma^2 |h_2|^2 |h_3|^2}{N_2}\right)}\right)\right.\\ &+Q\left(\sqrt{2\left(\frac{\gamma^2 |h_5|^2 |h_6|^2}{N_1}+\frac{|h_4|^2}{N_2}\right)}\right)\\ &\left.+Q\left(\sqrt{2\left(\frac{|h_1+\gamma h_5 h_6|^2}{N_1}+\frac{|h_4+\gamma h_2 h_3|^2}{N_2}\right)}\right)\right\}\end{aligned} \qquad \textbf{(17)}$$

Finally, it can be easily verified that the other terms in Expression (14) can be upper-bounded by a similar expression to that corresponding to $\Pr\left\{\text{error} \mid (b_1, b_2) = \mathbf{x}_A^{\mathrm{T}}\right\}$. Therefore, the overall probability of error will be given by $P_e = \Pr\left\{\text{error} \mid (b_1, b_2) = \mathbf{x}_A^{\mathrm{T}}\right\}$, i.e. the same bound given in Equation (17) will also be valid for the end-to-end probability of error P_e. As verified by our simulation results, this expression results in a tight upper-bound of the end-to-end probability of error for the AF approach.

3.2. Analysis of Detect-and-Forward ML Detector at the Relay

Equations (4) and (5) show that the analysis of the DF method is quite different than that of the AF approach. In the DF case we have complex Gaussian noise of the same variance for both components of the received vector **y** and, since $\hat{b}_1, \hat{b}_2 \in \{-1,1\}$, the relay gain factor is set to $\frac{1}{\sqrt{2}}$, which is not a function of channel variables. However, the computation of the analytic probability of error for the former is more complicated since the detected bit values of $\hat{b}_1$ and $\hat{b}_2$ at the relay are random variables that depend on the condition of the channel from the sources to the relay. In fact, when the vector $\mathbf{x}_A^{\mathrm{T}} = (1,1)$ is transmitted from the sources, the noise-free signal $\mathbf{u}_{\mathrm{A}}$ received at the destination may be any of the following set with the corresponding probability,

$$\mathbf{u}_A = \begin{cases} \begin{pmatrix} h_1+\gamma h_6 \\ h_4+\gamma h_3 \end{pmatrix} \square\ \mathbf{u}_{A1},\ Pr\{\mathbf{u}=\mathbf{u}_{A1}\} = (1-P_e^1)(1-P_e^2) \\ \begin{pmatrix} h_1+\gamma h_6 \\ h_4-\gamma h_3 \end{pmatrix} \square\ \mathbf{u}_{A2},\ Pr\{\mathbf{u}=\mathbf{u}_{A2}\} = P_e^1(1-P_e^2) \\ \begin{pmatrix} h_1-\gamma h_6 \\ h_4+\gamma h_3 \end{pmatrix} \square\ \mathbf{u}_{A3},\ Pr\{\mathbf{u}=\mathbf{u}_{A3}\} = (1-P_e^1)P_e^2 \\ \begin{pmatrix} h_1-\gamma h_6 \\ h_4-\gamma h_3 \end{pmatrix} \square\ \mathbf{u}_{A4},\ Pr\{\mathbf{u}=\mathbf{u}_{A4}\} = P_e^1 P_e^2 \end{cases} \qquad \textbf{(18)}$$

where P_e^1 and P_e^2 are given in Expressions (6) and (7), respectively. Similarly, if the transmit vector $\mathbf{x}_B^{\mathrm{T}} = (-1,1)$ is sent, the possible received signal set will be

$$\mathbf{u}_A = \begin{cases} \begin{pmatrix} -h_1+\gamma h_6 \\ h_4-\gamma h_3 \end{pmatrix} \square\ \mathbf{u}_{B1},\ Pr\{\mathbf{u}=\mathbf{u}_{B1}\} = (1-P_e^1)(1-P_e^2) \\ \begin{pmatrix} -h_1+\gamma h_6 \\ h_4+\gamma h_3 \end{pmatrix} \square\ \mathbf{u}_{B2},\ Pr\{\mathbf{u}=\mathbf{u}_{B2}\} = P_e^1(1-P_e^2) \\ \begin{pmatrix} -h_1-\gamma h_6 \\ h_4-\gamma h_3 \end{pmatrix} \square\ \mathbf{u}_{B3},\ Pr\{\mathbf{u}=\mathbf{u}_{B3}\} = (1-P_e^1)P_e^2 \\ \begin{pmatrix} -h_1-\gamma h_6 \\ h_4+\gamma h_3 \end{pmatrix} \square\ \mathbf{u}_{B4},\ Pr\{\mathbf{u}=\mathbf{u}_{B4}\} = P_e^1 P_e^2 \end{cases} \qquad \textbf{(19)}$$

From the above definitions, the probability that $\mathbf{x}_A$ is transmitted and $\mathbf{x}_B$ is detected will be then given by the probability that one of the four signals corresponding to $\mathbf{x}_A$ is transmitted and one of the signals corresponding to $\mathbf{x}_B$ is detected. Since we have four different transmit pairs (b_1,b_2) (each of which can be transmitted in four different ways from the relay), an exact computation of the probability of error at the destination will involve complex scenarios of non-uniform vector constellations with non-uniform probabilities. To simplify this computation, we will derive an approximate probability of error by solely accounting for the most probable error events in our scenario.

Having said this, we will consider two main terms in the probability of error for the case when $\mathbf{x}_A$ is sent, i.e. $\Pr\left\{\text{error} \mid (b_1, b_2) = \mathbf{x}_A^{\mathrm{T}}\right\}$. The first component, denoted by $P_{e,\blacklozenge}$, corresponds to the case that no error has occurred at the relay, and therefore the signal $\mathbf{u}_{A1}$ with probability $\Pr\{\mathbf{u}_A = \mathbf{u}_{A1}\}$ is transmitted among all the points associated to $\mathbf{u}_A$. We will then compute the probability that this signal is detected erroneously as one of the three most likely points associated to $\mathbf{u}_B$ (i.e. ignoring the point $\mathbf{u}_{B4}$ with much smaller probability $\Pr\{\mathbf{u}_B = \mathbf{u}_{B4}\}$).

Consequently, $P_{e,\diamond}$ can be approximated as

$$\begin{aligned} P_{e,\diamond} \square\ & \mathrm{E}\{\Pr\{\mathbf{u}=\mathbf{u}_{A1}\}\}\mathrm{E}\left\{ Q\left(\frac{\sqrt{|h_1|^2+|h_3|^2}}{\sqrt{2}\sigma}\right)\right. \\ & + Q\left(\frac{\sqrt{|h_1|^2}}{\sigma}+\frac{\sigma\ln\frac{\Pr\{\mathbf{u}=\mathbf{u}_{A1}\}}{\Pr\{\mathbf{u}=\mathbf{u}_{B2}\}}}{2\sqrt{|h_1|^2}}\right) \\ & + Q\left(\frac{\sqrt{|h_1|^2+\frac{|h_3|^2+|h_6|^2}{2}+\sqrt{2}\Re(h_1h_6^*)}}{\sigma}\right. \\ & \left.\left. + \frac{\sigma\ln\frac{\Pr\{\mathbf{u}=\mathbf{u}_{A1}\}}{\Pr\{\mathbf{u}=\mathbf{u}_{B3}\}}}{2\sqrt{|h_1|^2+\frac{|h_3|^2+|h_6|^2}{2}+\sqrt{2}\Re(h_1h_6^*)}}\right)\right\} \end{aligned} \tag{20}$$

where $\Re(\cdot)$ stands for the real part of a complex value, and σ^2 denotes the variance per dimension of the noise term n=$(n_3,n_4)^{\mathrm{T}}$ at the destination. Observe that the additional factors $\ln\frac{\Pr\{\mathbf{u}=\mathbf{u}_{A1}\}}{\Pr\{\mathbf{u}=\mathbf{u}_{B2}\}}$ and $\ln\frac{\Pr\{\mathbf{u}=\mathbf{u}_{A1}\}}{\Pr\{\mathbf{u}=\mathbf{u}_{B3}\}}$ in the second and third terms of the above equation are due to the non-equal probabilities of occurrence of the constellation vectors $\mathbf{u}_{A1}$, $\mathbf{u}_{B2}$ and $\mathbf{u}_{B3}$.

The second main component of the proposed approximation for $\Pr\left\{\mathrm{error}|(b_1,b_2)=\mathbf{x}_A^{\mathrm{T}}\right\}$, denoted as $P_{e,\square}$, is related to the case when the relay has wrongly detected the bit values (–1,1) corresponding to the point $\mathbf{u}_{A2}$. Therefore, we must compute the probability that this signal, which belongs to the set associated to $\mathbf{X}_A$, is erroneously detected as $\mathbf{X}_B$ at destination. In this case, such an error probability can be approximated by the probability that this signal is transmitted by the relay and is detected as the point $\mathbf{u}_{B1}$, which presents the highest probability among all the points associated to $\mathbf{u}_B$. In other words, an error event will occur if the received signal y=$\mathbf{u}_{A2}$+$\mathbf{n}$ is more likely to be detected as $\mathbf{u}_{B1}$ instead of $\mathbf{u}_{A2}$ or $\mathbf{u}_{A1}$. At this point it should be noted that, since we assume that the signal $\mathbf{u}_{A2}$ is transmitted from the relay, the probability of detecting the less probable constellation points $\mathbf{u}_{A3}$ and $\mathbf{u}_{A4}$ is negligible in comparison with the above mentioned probabilities. Thus the second error term $P_{e,\square}$ can be approximated by

$$\begin{aligned} P_{e,\square} \square\ & \mathrm{E}\Big\{\Pr\Big(-\|(\mathbf{u}_{A2}+\mathbf{n})-\mathbf{u}_{B1}\|^2 > -\|(\mathbf{u}_{A2}+\mathbf{n})-\mathbf{u}_{A1}\|^2, \\ & -\|(\mathbf{u}_{A2}+\mathbf{n})-\mathbf{u}_{B1}\|^2 > -\|(\mathbf{u}_{A2}+\mathbf{n})-\mathbf{u}_{A2}\|^2\Big)\Big\} \end{aligned} \tag{21}$$

The above joint probability can be computed by considering the projections of the noise term $\mathbf{n}$ on the two directions $\mathbf{u}_{B1}$–$\mathbf{u}_{A1}$ and $\mathbf{u}_{B1}$–$\mathbf{u}_{A2}$, and integrating over the joint probability distribution of these two projection terms. Let us denote the correlation factor of these two noise terms as

$$\begin{aligned} \rho\ \square\ & \frac{\langle \mathbf{u}_{B1}-\mathbf{u}_{A2},\mathbf{u}_{B1}-\mathbf{u}_{A1}\rangle}{\|\mathbf{u}_{B1}-\mathbf{u}_{A2}\|\cdot\|\mathbf{u}_{B1}-\mathbf{u}_{A1}\|} \\ & = \frac{4|h_1|^2}{2|h_1|\cdot 2\sqrt{|h_1|^2+\gamma^2|h_3|^2}}, \end{aligned} \tag{22}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product of two vectors. With this definition, Expression (21) reduces to [27]

$$P_{e,\square}\ \square\ \mathrm{E}\left\{\frac{1}{2\pi\sqrt{1-\rho^2}}\int_a^\infty\int_b^\infty e^{\frac{-(x^2-2\rho xy+y^2)}{2(1-\rho^2)}}dxdy\right\} \tag{23}$$

where the integration limits a and b are obtained by computing the distance of the received signal from the decision boundaries between $\mathbf{u}_{B1}$ and $\mathbf{u}_{A1}$ and between $\mathbf{u}_{B2}$ and $\mathbf{u}_{A2}$, respectively. Further geometric manipulations lead to

$$\begin{aligned} a & = \frac{\|\mathbf{u}_{B1}\|^2-\|\mathbf{u}_{A1}\|^2+2<\mathbf{u}_{A2},\mathbf{u}_{A1}-\mathbf{u}_{B1}>}{2\sigma\|\mathbf{u}_{B1}-\mathbf{u}_{A1}\|} \\ & = \frac{\sqrt{|h_1|^2+\gamma^2|h_3|^2}}{\sigma}-\frac{4\gamma^2|h_3|^2}{2\sigma\sqrt{|h_1|^2+\gamma^2|h_3|^2}} \end{aligned} \tag{24}$$

$$\begin{aligned} b & = \frac{\|\mathbf{u}_{A2}-\mathbf{u}_{B1}\|}{2\sigma}+\frac{\sigma}{\|\mathbf{u}_{A2}-\mathbf{u}_{B1}\|}\ln\frac{Pr\{\mathbf{u}=\mathbf{u}_{A2}\}}{Pr\{\mathbf{u}=\mathbf{u}_{B1}\}} \\ & = \frac{\sqrt{|h_1|^2}}{\sigma}+\frac{\sigma}{2\sqrt{|h_1|^2}}\ln\frac{\Pr\{\mathbf{u}=\mathbf{u}_{A2}\}}{\Pr\{\mathbf{u}=\mathbf{u}_{B1}\}}. \end{aligned} \tag{25}$$

It should be noted that, when computing $P_{e,\diamond}$, the value of $\Pr\{\mathbf{u}=\mathbf{u}_{B2}\}=P_e^1(1-P_e^2)$ should be obtained based on channel values that cause errors at the relay. Analogously, $P_{e,\square}$ should be computed by averaging over all values of h_2 that cause such errors at the relay, and not over the whole range of h_2. Finally, the approximate end-to-end probability of error P_{e} for the DF scheme can be obtained by adding these two main error terms, i.e.

$$P_e\ \square\ P_{e,\diamond}+P_{e,\square} \tag{26}$$

which, as shown in next section, is in close match with the Monte Carlo simulation results.

4. Simulation Results

In this section we provide Monte Carlo simulation results for the proposed algorithm in comparison with the previously derived analytic approximations. The simulations have been performed assuming BPSK signaling (i.e. $b_1,b_2 \in \{-1,1\}$). The transmitted signals at the sources and the relay are assumed to be of unit average power. Flat Rayleigh fading coefficients were generated based on a complex Gaussian distribution with variance per 2 dimensions equal to one. As mentioned earlier, an additional signal power attenuation equal to K/d^2 was assumed for signal propagation over a distance equal to d, where K is the propagation factor (set to 10^{-4} in our simulations). Fading coefficients have been assumed to be constant over a block of 100 samples, and are independently generated over different blocks. The Signal to Noise Ratio (SNR) is defined as the ratio of the average received power of the direct source-destination path of one of the sources (where average transmit power is assumed to be equal to one) to the noise variance per dimension. The noise variance used for computing the SNR has been assumed to be the same at the relay and destination sites. Furthermore, the relay gains γ_1 and γ_2 have been chosen such that average transmit power of the relay is also normalized to one. Finally, in order to investigate the effect of the relay location on the performance of the different considered schemes, the location of the relay varies over a range of d=0 to d=1. The results for d=0.1 (relay close to sources) and d=0.9 (relay close to destination) are shown in these plots.

Figure 2 depicts the probability of symbol error of the proposed DF and AF algorithms versus the SNR ratio at two different relay locations. First notice that in both approaches the obtained analytic approximation for the probability of error is in close match with the corresponding simulated curves. Also observe that the performance of the DF approach even improves slightly as the relay position is changed from a location close to sources (d=0.1) to a location close to the destination (d=0.9). However, the performance degrades considerably for the AF scheme as the distance between the relay and the sources increases. Based on these results it is foreseen that, in scenarios where the relay is close to the destination, the use of the proposed switched-frequency DF scheme will yield a significant performance enhancement (around 10 dB in our proposed setup) at the cost of a minor complexity increase.

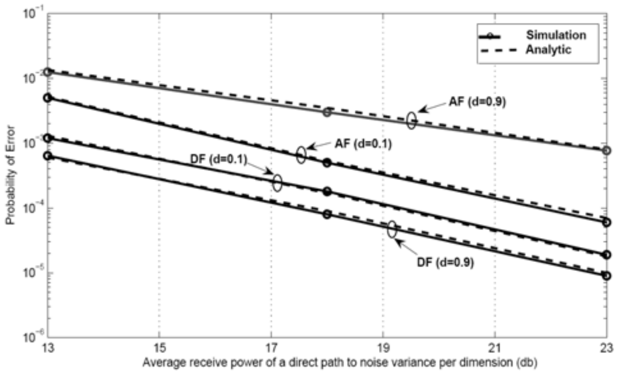


Figure 2. Analytic and simulated probability of symbol error for the proposed *frequency switching* AF and DF schemes at two different relay positions.

5. Concluding Remarks

In this paper, the error performance of a novel communication scheme for the two-user single-relay multiple access channel has been proposed, which achieves a diversity of order two by using only two frequency channels over all the links. The main advantage of the proposed scheme is to obviate the requirement of complex superposition-type coding schemes in Detect-and-Forward scenarios. The effect of the relay location for both AF and DF schemes has also been investigated, concluding in a superior performance of the DF scheme as the relay gets closer to the destination.

6. Acknowledgements

This work was supported in part by the Spanish Ministry of Science and Innovation through the CONSOLIDER-INGENIO 2010 programme (CSD2008–00010, *www.comonsens.org*), and by the Basque Government through the *Future Internet* project (ETORTEK programme).

7. References

[1] C. T. K. Ng and A. J. Goldsmith, “Capacity and cooperation in wireless networks,” in Proceedings of Information Theory and Applications (ITA) Workshop, February 2006.

[2] ——, “The impact of CSI and power allocation on relay channel capacity and cooperation strategies,” submitted to IEEE Transactions on Information Theory, January 2007.

[3] S. Avestimehr and D. Tse, “Outage capacity of the fading relay channel in the low SNR regime,” IEEE Transactions on Information Theory, Vol. 53, No. 4, pp. 1401–1415, April 2007.

[4] R. U. Nabar and H. Kneubuhler, F.W.; Bolcskei, “Performance limits of amplify-and-forward based fading relay channels,” in Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP), May 2004.

[5] C. Hausl and P. Dupraz, “Joint network-channel coding for the multiple-access relay channel,” in Proceedings of International Workshop on Wireless Ad Hoc and Sensor Networks (IWWAN), June 2006.

[6] J. Del Ser, P. M. Crespo, B. H. Khalaj, and J. Gutierrez-Gutierrez, “On combining distributed joint source-channel-network coding and turbo equalization in multiple access relay networks,” in Proceedings of 3rd IEEE International Conference on Wireless and Mobile

Computing, Networking and Communications (WiMob' 07), October 2007.

[7] G. Kramer and A. J. van Wijngaarden, "On the white gaussian multiple access relay channel," in Proceedings of the International Symposium on Information Theory, June 2000.

[8] L. Sankaranarayanan, G. Kramer, and N. B. Mandayam, "Capacity theorems for the multiple-access relay channel," in Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing, October 2004.

[9] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity – Part I: System description," IEEE Transactions on Communications, Vol. 51, No. 11, pp. 1927–1938, November 2003.

[10] T. E. Hunter and A. Nosratinia, "Cooperation diversity through coding," in Proceedings of the IEEE International Symposium on Information Theory, 2002.

[11] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," IEEE Transactions on Information Theory, Vol. 50, No. 12, pp. 3062–3080, December 2004.

[12] S. Valentin, H. S. Lichte, H. Karl, G. Vivier, S. Simoens, J. Vidal, A. Agustin, and I. Aad, "Cooperative wireless networking beyond store-and-forward: Perspectives for PHY and MAC design," in Proceedings of the 17th Wireless World Research Forum Meeting (WWRF 17), November 2006.

[13] P. Herhold, E. Zimmermann, and G. Fettweis, "On the performance of cooperative amplify-and-forward relay networks," in Proceedings of the 5th International ITG Conference on Source and Channel Coding (SCC), January 2004.

[14] K. Azarian, H. E. Gamal, and P. Schniter, "On the achievable diversity-multiplexing tradeoff in half-duplex cooperative channels," IEEE Transactions on Information Theory, Vol. 51, No. 12, pp. 4152–4172, December 2005.

[15] M. Yu and J. T. Li, "Is amplify-and-forward practically better than decode-and-forward or viceversa?" in Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing (ICASSP), Vol. 3, pp. 365–368, March 2005.

[16] N. Seshadri and J. H. Winters, "Two signaling schemes for improving the error performance of frequency-division-multiplex (FDD) transmission systems using transmitter antenna diversity," in Proceedings of the IEEE Vehicular Technology Conference, Vol. 1, pp. 508–511, May 1993.

[17] A. Wittneben, "A new bandwidth efficient transmit antenna modulation diversity scheme for linear digital modulation," in Proceedings of IEEE International Conference on Communications, Vol. 3, pp. 1630–1634, May 1993.

[18] V. Tarokh, N. Seshadri, and A. Calderbank, "Space-time codes for high data rate wireless communication: Performance analysis and code construction," IEEE Transactions on Information Theory, Vol. 44, No. 3, pp. 744–765, March 1998.

[19] D. Chen, K. Azarian, and J. N. Laneman, "A case for amplify-forward relaying in the block-fading multiaccess channel," submitted to IEEE Transactions on Information Theory, January 2007.

[20] L. Sankar, G. Kramer, and N. B. Mandayam, "Offset encoding for multiple access relay channels," IEEE Transactions on Information Theory, Vol. 53, No. 10, pp. 3814–3821, October 2007.

[21] L. Sankar, Y. Linang, H. V. Poor, and N. Mandayam, "Opportunistic communication in an orthogonal multiaccess relay channel," in Proceedings of the International Symposium on Information Theory, June 2007.

[22] B. H. Khalaj, J. Del Ser, P. M. Crespo, and J. Gutierrez-Gutierrez, "A novel dual-frequency multiple access relay transmission scheme," in Proceedings of IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (IEEE PIMRC), September 2007.

[23] B. Sklar, "Rayleigh fading channels in mobile digital communication systems Part I: Characterization," IEEE Communications Magazine, Vol. 35, No. 7, pp. 90–100, July 1997.

[24] H. Ju, S. Lee, K. Kwak, E. Oh, and D. Hong, "A new duplex without loss of data rate and utilizing selection diversity," in Proceedings of IEEE Vehicular Technology Conference, pp. 1519–1523, May 2008.

[25] C. Hausl, F. Schreckenbach, I. Oikonomidis, and G. Bauch, "Iterative network and channel decoding on a tanner graph," in Proceedings of 43rd Allerton Conference on Communication, Control, and Computing, September 2005.

[26] H. Jafarkhani, "Space-time coding, theory and practice," Cambridge University Press, 2005.

[27] H. Kuai, F. Alajaji, and G. Takahara, "Tight error bounds for nonuniform signaling over AWGN channels," IEEE Transactions on Information Theory, Vol. 46, No. 7, pp. 2712–2718, November 2000.

Int. J. Communications, Network and System Sciences, **2009, 7, 600-607**
doi:10.4236/ijcns.2009.27067 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

Research on the Active DDoS Filtering Algorithm Based on IP Flow

Rui GUO[1], Hao YIN[1], Dongqi WANG[2], Bencheng ZHANG[3]
[1]*Department of Computer Science and Technology, Tsinghua University, Beijing, China*
[2]*The Computing Center, Northeastern University, Shenyang, China*
[3]*Electronic Scouting and Commanding Department, College of Shenyang Artillery, Shenyang, China*
Email: *gr@tsinghua.edu.cn*

Received May 22, 2009; *revised July* 7, 2009; *accepted September* 5, 2009

ABSTRACT

Distributed Denial-of-Service (DDoS) attacks against public web servers are increasingly common. Countering DDoS attacks are becoming ever more challenging with the vast resources and techniques increasingly available to attackers. It is impossible for the victim servers to work on the individual level of on-going traffic flows. In this paper, we establish IP Flow which is used to select proper features for DDoS detection. The IP flow statistics is used to allocate the weights for traffic routing by routers. Our system protects servers from DDoS attacks without strong client authentication or allowing an attacker with partial connectivity information to repeatedly disrupt communications. The new algorithm is thus proposed to get efficiently maximum throughput by the traffic filtering, and its feasibility and validity have been verified in a real network circumstance. The experiment shows that it is with high average detection and with low false alarm and miss alarm. Moreover, it can optimize the network traffic simultaneously with defending against DDoS attacks, thus eliminating efficiently the global burst of traffic arising from normal traffic.

Keywords: DDoS Attack, Genetic Algorithm, IP Flow Statistics

1. Introduction

Denial-of-Service (DoS [1]) attacks use legitimate requests to overload the server, causing it to hang, crash, reboot, or do useless work. The target application, machine, or network spends all of its critical resources on handling the attack traffic and cannot attend to its legitimate clients. Both DoS and DDoS are a huge threat to the operation of Internet sites, but the DDoS [2,3] problem is more complex and harder to solve.

There are two main classes of DDoS attacks: bandwidth depletion and resource depletion. A resource depletion attack is an attack that is designed to tie up the resources of a victim system. This type of attack targets a server or process at the victim making it unable to legitimate requests for service. A bandwidth depletion attack is designed to flood the victim network with unwanted traffic that prevents legitimate traffic from reaching the victim system. And there are three main defense approaches: traceback [4]—with the increase of zombies this approach will be invalidated rapidly; filtrate [5]—because this method requires the participation of the communication company and many routers, the filter must be open at all times, and the approach is too costly; throttle [6]—legitimate data stream will be limited because too many data streams converge at a central point. Thus, based on these three methods, three distinct defense approaches emerge: gateway defense, router defense and computer defense.

This paper aims at discussing a low cost, high performance and easy-to-deploy approach [7] which selects five statistical features from IP flow is proposed on filtering DDoS attacks on routers. We use usual statistical traffic of IP flow to get the percentage of traffic of the upriver routers. Then we use the percentage to assign a weight for the router. When DDoS happens, we observe the traffic quota of several chosen routers. Our goal is to maximize goodput, with the weight that we figure out in normal state. At the same time we can calculate which routers should block the traffic. Then the victim server sends a filtering request to these routers to block all traffic from certain sources to the victim.

We present an implementation of these concepts, along with experimental results from our laboratory testbed. In the rest of this section we give a very brief overview of the filtering mechanism. Section 2 tells the reason of the IP Flow approach. Section 3 presents the architecture based on routers that can support filtering mechanism. Section 4 gives implementation and performance details. In Section 5, we conclude with a discussion of deployment options, as well as related work.

2. IP Flow Filtering Overview

IP flow is composed of IP packets arriving one after another. As the basic data carrying unit of Internet, IP packet holds the upper layer's information and can be easily caught and handled. In the following part of this section IP flow will be divided into the Micro-Flow and the Macro-Flow and we are going to research how to select effective IP flow based detecting features.

2.1. The Micro-Flow and Macro Flow

2.1.1. The Micro-Flow

A Micro-Flow is a packet set who is composed of packets belonging to the same time interval of Internet, and all these packets have the same specific characteristics. These same specific characteristics are called keys. A group of commonly used keys are <Protocol, SrcIP, SrcPort, DestIP, DestPort>. Protocol is the protocol used by the upper layer, SrcIP and SrcPort are the source IP address and the source port number separately. DestIP and SrcIP are the destination IP address and the destination port number separately.

The definition of Micro-Flow is helpful in two ways. First, each key group corresponds to one connection from SrcIP to DestIP, so keys can be used to describe DDoS connection. Second, a key group contains much information which can be used by routers and firewalls to operate each packet.

2.1.2. The Macro-Flow

All the packets belonging to one time interval compose a set which is called the Macro-Flow. Macro-Flow is pooled by Micro-Flows.

The definition of Macro-Flow is helpful in two ways too. First, Detecting features can be formed on the base of Macro-Flow. Second, the information contained in the Macro-Flow is the complementarities to keys.

In experiments, we intercept network traffic by time interval i=10s randomly. On one hand, in order to form the Micro-Flow based features, we classify packets by different keys. On the other hand, we abstract the Macro-Flow based features from the whole i directly.

2.2. IP Flow Based Features

2.2.1. Micro-Flow Based Features

1) Average Number of Packets in Per Flow (ANPPF)

Continuously and randomly generated "legitimate" IP are usually used in attack, so the generating speed of Micro-Flow is quickened, and the packet amount in per flow decrease. There are commonly 1~3 packets in per flow [9].

$$ANPPF = \left(\sum_{j=1}^{FlowNum} PacketsNum_j \right) / \ FlowNum$$

PacketsNumj is the quantity of packets in the jth flow of a time interval. FlowNum is the quantity of packets of the whole interval. Figure 1 shows the experimental com parison of ANPPF between normal traffic and DDoS traffic (110i~180i).The ANPPF of DDoS traffic which is near 1(attacking traffic is the mix of DDoS traffic generated by tfn2k and normal traffic of internet. ANPPF of tfn2k generating traffic is 1) differs from normal ANPPF (ruleless distribution) significantly.

2) Percentage of Correlative Flow (PCF)

During attack, though the victim still has capability to reply to attacking packets' "requests", the replying packets can not get to the zombies, because the attacking IP addresses are faked. If flow x is from SrcIPx=A to DestIPx=B, and flow y is from SrcIPy=B to DestIPy=A, then we call flow x and y is a pair of Correlative Flow.

$$PCF = CFNum \,/\, FlowNum$$

CFNum is two times of the pairs of Correlative Flow. PCF represents the "there is going-out but no coming-back" characteristic of DDoS. As is shown in Figure 2, when DDoS happens (110i~180i), PCF is near 0, while the PCF of normal traffic is 0.4~0.6. The difference between them is distinguishable.

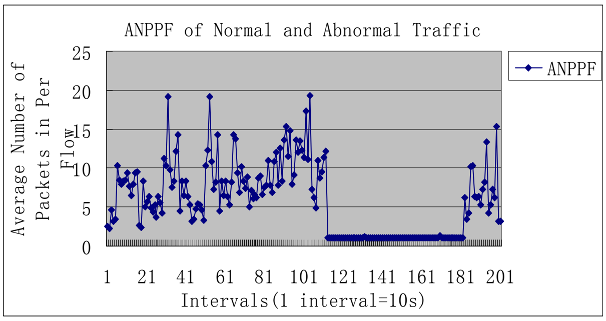


Figure 1. ANPPF of normal and abnormal traffic.

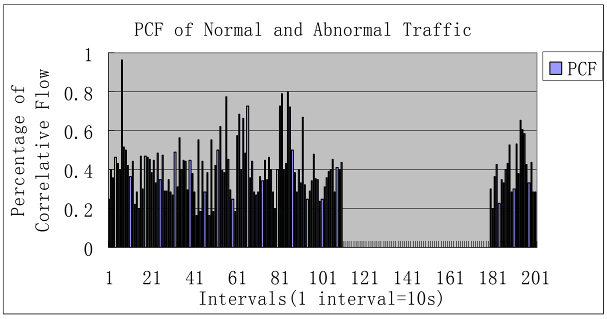


Figure 2. PCF of normal and abnormal traffic.

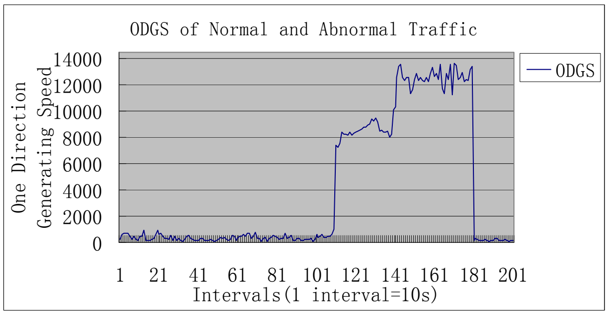


Figure 3. ODGS of normal and abnormal traffic.

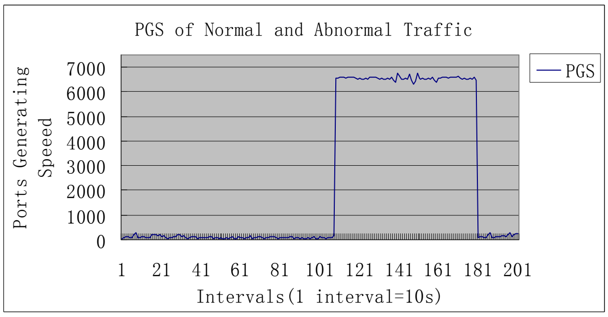


Figure 4. PGS of normal and abnormal traffic.

3) One Direction Generating Speed (ODGS)

Flow generating speed quickens when attack happens or busy time comes. In order to distinguish these two kinds of situations, ODGS is proposed.

$$ODGS = (FlowNum - CFNum) / interval$$

ODGS reflects the sudden increase of traffic when DDoS happens, and it also reflects the “there is going-out but no coming-back” characteristic of DDoS. Figure 4 gives the experimental comparison of ODGS between normal traffic (110i~180i) and abnormal traffic. ODGS’ order of magnitude in normal traffic (102) is much smaller than that in the abnormal traffic (104).

4) Ports Generating Speed (PGS)

$$PGS = PortsNum / interval$$

PortsNum is the number of distinct port in one time interval. Some researchers select the size of port [2] as a detecting feature, while we find that many newly emerged services and applications (such as famous p2p application BT) use port number bigger than 1024, so approach of [2] is not suitable anymore. Through deeper investigation, we realize that attackers continuously and randomly generate port too, so PGS is proposed. As is shown in Figure 4, the PGS of normal traffic is not bigger than 200, while PGS of attacking traffic (110~180i) is over thousands.

2.2.2. The Macro-Flow Based Feature

PAP (Percentage of Abnormal Packets)

In order to increase the efficiency of attacking, attacking packets’ content parts are usually unfilled or only filled with very few useless bytes (such as famous attacking tools tfn2k, trinoo). This kind of procedure results in the increase of abnormal small packets (for example, some TCP packets are only a little bigger than 40bytes, and UDP packets are only a little bigger than 28bytes). PAP presents this characteristic of DDoS attack by counting the percentage of abnormal packets in the one i(a Macro-Flow). Figure 5 is the comparison of PAP of normal traffic and abnormal traffic. As we can see, there is a significant change of PAP from near 0 to more than 0.9 when DDoS happens (110i~180i).

Defending against DDoS attacks often involves detection and response. There are a number of statistical approaches for detection of DDoS attacks, including the use of IP addresses and TTL [11] values and TCP SYN/FIN packets for detecting SYN flood attacks. Also entropy and Chi-Square statistics are used to differentiate between attack and normal packets. The D-WARD approach [8] uses, in addition to network and transport header statistics, application layer [10] knowledge to implement the filter policy. But all these method require the participation of many routers, the filter must be open at all times, so the approach is too costly.

3. The Design of Statistical Analysis Filtering System

From the Micro-Flow and Macro Flow, we can get the statistical result: $\sum_{j=1}^{n} S_{kj}$ are all connections form source IPk, $\sum_{i=1}^{n} S_{ik}$ are all connections which are routed to destination IPk. It is easy to create probability statistics of access records. Generally, DoS attacks launched by a large number of hosts which host never accessed the victim network before. Meaning during a DDoS attack most of the hosts to the victim are fresh new, which is so different to flash crowd [7]. So we can use history IP database by putting these IP of high frequency in a pool. Common algorithm is not efficient enough to catch up with the line rate of high speed at reasonable memory consumption.

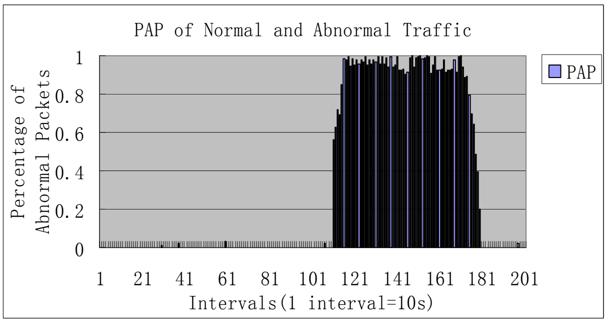


Figure 5. PAP of normal and abnormal traffic.

To address this limitation, one can use the Bloom filter. It reduces space/time complexity by allowing small degree of inaccuracy in membership representation Bloom Filter is chosen to generate the IP address white list. If a host exists in this IP Bloom Filter the router will route the packet to destination, if not it will pass though filter.

The conventional algorithm requires a memory of 1 G bits while our Bloom filter array requires a memory of only 50M bits, at the cost of losing 1% accuracy in membership representation.

A Bloom filter for representing a set S={x1,x2,…,xn} of n elements is described by an array of m bits, initially all set to 0. It uses k independent hash function h1, …, hk with range {1,…,m}. Here we have an assumption that hash functions are perfectly random, which means the hash functions map each item in the universe to a random number uniform over the range {1,…,m}. For each element $x\in S$, the bits hi(x) are set to 1 for $1\leq i\leq k$. Alocation can be set to 1 multiple times, but only the first change has an effect. For the membership query if $y\in S$, we check if $\forall$i, hi(y)=1. If $\exists$hi(i)≠1, then $y\notin S$. If $\forall$i, hi(y)=1 is true, we can assume $y\in S$ with a false positive rate as

$$p_{err}=\left(1-p_0\right)^k=\left(1-\left(1-\frac{1}{m}\right)^{kn}\right)^k\approx\left(1-e^{-\frac{kn}{m}}\right)^k$$

The construction of BF is shown in Figure 1. Initially, each bit in the element BF is set to 0 and the pointer list is set to null. Then each history Flow {S,A} *Si*, $i\in[1, n]$ is hashed by function *Hj*, $j\in[1, k]$ with corresponding hit bit in BF being set to 1. A new node of link list is created with the sum field being filled by the sum of previous value and the last 16 bits of the index value of the filter that are being set to 1. The *Si*, $i\in[1, n]$ are hashed *k* times. If the bit has already been set to 1, a new node of link list array is appended to the list. This design does not affect much accuracy because in all the experiments the false positive rates are the same (Figure 6).

As shown in Figure 7 our DDoS defense system has an Offline Training System (OTS) and an Online Filtering System (OFS) and is deployed between the source end and the victim end. From OTS we create whitelist and map the list in BF. The GA-Filter modules are deployed at the edge routers that are close to the attack. During DDoS attacks, if a flow matchs this bloom filter, it will be transmitted by routers, if not it will be filtered by GA-filter. The filtering routers can afford to selectively block traffic to the victim server. In that case legitimate traffic passing from that router is also unnecessarily filteredtogether with the attack traffic. We would like to filter out all attackers and allow all good traffic to reach the server. Unfortunately, in a DDoS attack, it’s hard to differentiate attack traffic and legitimate traffic. In this paper, we aim at designing a defense system that contains DDoS flooding attacks in high-speed networks. The objectives are to

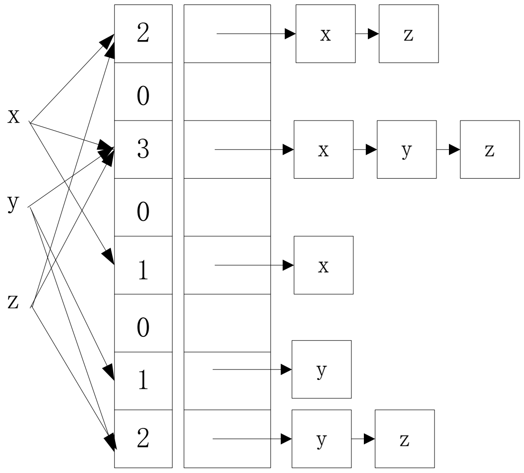


Figure 6. The construction of bloom filter.

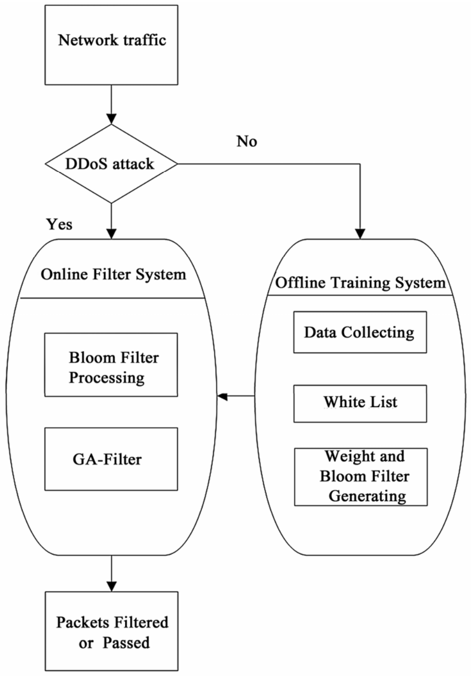


Figure 7. The Filter architecture.

1) Maximize friendly traffic throughput while reducing attack traffic as much as possible

2) Minimize the disturbance of the defense system on delay performance of friendly traffic

3) Achieve high compatibility to the original systems.

A router-based defense strategy: These routers are inserted in some important point of the network. We envision these routers that are deployed in the network to collaboratively perform the desired countermeasure functions, including detection of DDoS flooding attacks and access control of network traffic.

3.1. Combinatorial Optimization of Filtering Problem

The filtering problem is a combinatorial optimization of the traffic to victim server, which seeks for maximum legitimate traffic from all good or bad traffic. We assume that there are n distinct routers involved and the traffic in total transmit to victim server is $w_i \sum$, generally, each router j (j = 1, …, n) transmit traffic to victim server has assigned a profit Pi (i= 1, …, n) and the maximum throughput is C. When a router route stream i (i = 1, …, m) to victim server , we define Xi=1; if stream i (i =1, …, m) isn't routed to server, we define Xi =0. So the stream in total is $\sum_{i=1}^{n} w_i x_i$, but the good traffic is $\sum_{i=1}^{n} p_i x_i$,

The problem is to identify a subset of all traffic that leads to the highest possible total good traffic and does not exceed maximum throughput C. Formally, our filtering model can be stated as follows:

Maximize $\sum_{i=1}^{n} p_i x_i$

subject to $\sum_{i=1}^{n} w_i x_i \le c$

with $p_i \ge 0,\ w_i \ge 0,\ C \ge 0$

3.2. Genetic Algorithms for Filtering Bad Traffic

Genetic algorithms are stochastic iterative algorithms for search and optimization that find their origin and inspiration in the Darwinian theory of biological evolution. GA abstract and mimic some of the traits of the ongoing struggle in evolution in order to do a better job in problems that require adaptation, search and optimization. Since we are in fact dealing with artificial systems, we should also feel free to employ whatever device works well for a given class of problems, even if it has no direct biological origin. Genetic Algorithms are computer algorithms that search for good solutions to a problem from among a large number of possible solutions. Genetic Algorithms of our filtering can be stated as follows:

3.2.1. Initial Population

The algorithm begins by creating an initial population which contains M individuals; a mutation probability; a crossover probability; the length of every chromosome N, and the maximum generations. Randomly generate a population of N chromosomes. We randomly generated traffic to victim server and the percentage of bad traffic. Initial transmitting throughput by routers is more than the maximum throughput C which the server can handle.

3.2.2. Encoding of the Chromosomes

Encoding of the problem in a binary string, the length is n, X_i =1, meaning the traffic passes through to the server, X_i =0, meaning the router drops the traffic. Such as X={0，1，0，1，0，0，1} expressing that traffic is passing through router 2, 4, 7. Namely, router 2, 4, 7 will transmit traffic to victim server. We randomly select bits of a chromosome and set it to 0 or 1. For each of the chromosome, test whether the constraint is satisfied. If so, accept it to be a number of the population. If not, drop it and randomly create a new chromosome. The x-vector describes which of the routers that are chosen in each solution, for example, the vector 01001011 means that router NO. 2, 5, 7 and 8 are chosen to route data to server.

3.2.3. Fitness Function

Given a chromosome that represents which router filtrate the traffic, the corresponding fitness function is defined as follow: fitness function $f(X) = \sum_{i=1}^{n} X_i P_i$ subject to $\sum_{i=1}^{n} X_i W_i \le C$.

At first, we define stream *i* passes through a router to victim server , we set X_i =1; if stream *i* (*i* = 1; : : : ;m) drop, we set X_i =0. Considering about n routers, the throughput is $\sum_{i=1}^{n} W_i X_i$ in total, but the goodput is $\sum_{i=1}^{n} P_i X_i$, so how to optimize variable X_i (*i*=1,2,…,n)and maximize goodput. So this problem is subject to two formulas: at $\sum_{i=1}^{n} X_i W_i \le C$ maximize $\sum_{i=1}^{n} P_i X_i$ X_i=1 or 0 (i=1,2,…,n). after analyzing the problem, for the fitness function, it can be stated as follows: $f(x) = \sum_{i=1}^{n} P_i X_i$, X_i=1 or 0 (i=1,2,…,n).

3.2.4. Selection Functions

We choose chromosomes based on probability, and appoint the individual to be the first generation. In the implementation of the program, we tried roulette-wheel methods: the fitness value of each individual is f_i, the probability of *i* is chosen shown as follow: $P_{si} = f_i / \sum_{i=1}^{n} f_i$; For the initial population, first we figure out the fitness value of each chromosome, and then we calculate selection probability. After the comparison, the chromosome with low chosen probability is eliminated and the high chosen probability chromosome will be copied. This copied chromosome takes the place of the eliminated chromosome. Then the selection of population is over.

3.2.5. Crossover

We use single point crossover. The crossover point is determined randomly by generating a random number between 0 and 1. We perform crossover with a certain probability. If crossover probability is 100% then a whole new generation is made by crossover. If it is 0% then whole new generation is made by exact copies of chromosomes from old population. We decided upon crossover rate of P_c. This means that P_c of the new generation will be formed with crossover and 1-P_c will be copied to the new generation.

3.2.6. Mutation

Mutation is made to prevent GA from falling into a local extreme. We perform mutation on each bit position of the chromosome with 0.1 % probability.

4. Performance Evaluation and Comparison

For evaluating our system, we use Bell lab's [11,12] data. Bell lab's data is stored as pure text, and each row of the text is a packet composed of SIP, DIP, SPort, DPort, packet length and ACK (TCP packet) et. The attack launched in our own simulation is constant rate attack, so we choose the constant rate UDP attack data of Bell lab's as the attack samples. (Table 1).

We suppose to have a server that has a capacity of C bandwidth and several routers transmit traffic with different ratio. We want the greatest total benefit without overloading the constraint of the bandwidth. We use a data structure, called cell, with two fields (goodput and traffic) to represent every router (Table 2). Then we use an array of type cell to store all routers in it.

In our experiments, we measured filtering characteristics by the rate of false and rate of missed [13]. In Table 3 shows the sensitivity and accuracy of the Bloom Filter. The ROC curves in Figure 8 and Figure 9 show the sensitivity and accuracy of the neural network. A ROC curve is a plot with the false positive rate on the X axis and the true positive rate on the Y axis. The area below the curve reflects the sensitivity of the neural network. As we can see, the curve is close to both the Y axis and the point (0, 1) which means that we obtained low false positives and the classification capability is good.

Micro-Flow and Macro-Flow detection based detecting features that we described in Section 2 is used to allocate the weights for traffic routing. As indicated in Table 3, that our IP flow based filtering method achieves pretty high accuracy and precision. It's low cost, high performance and easy-to-deploy. It optimizes the web flow; enhance the network efficiency by precluding and dismissing the overall current abruptness of ordinary flow.

Table 1. DDoS traffic.

Country	DDoS Type I		DDoS Type II	
	% of Good Traffic	% of bad Traffic	% of Good Traffic	% of bad Traffic
USA	36.27	43.9	36.2	45.9
Korea	5.8	11.5	0	12
China	18.35	10.3	24.1	0
Taiwan	2.46	6.1	2.4	16.7
Canda	3.64	5.4	3.6	5.4
UK	6.74	5.2	6.7	5.3
Germany	8.4	5.1	8.4	5.2
Australia	2.5	4.3	2.5	1.1
Japan	13.91	4.2	14.2	0
Netherlands	1.93	4.1	1.9	8.4

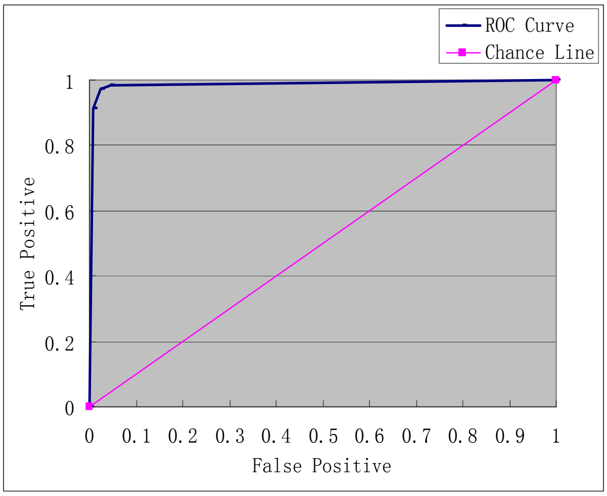


Figure 8. Our own data ROC curve.

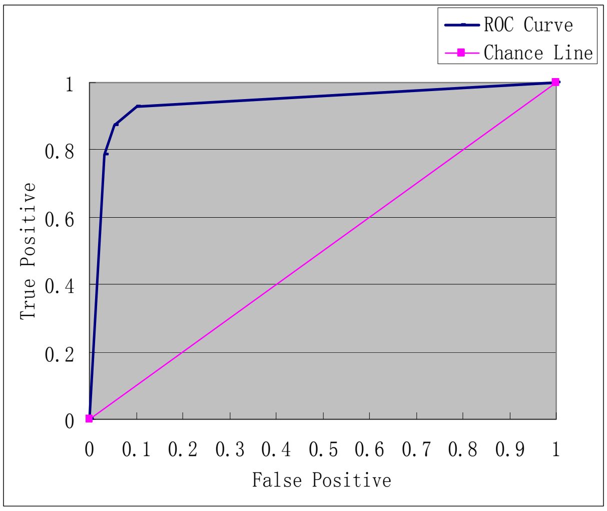


Figure 9. Bell ROC curve.

Table 2. Router information array.

Router 0		Router1		Router2		Router3		Router4	
36.27	80.17	5.8	17.3	18.35	28.65	2.46	8.56	3.64	9.04
Router 5		Router6		Router 7		Router 8		Router9	
6.74	11.94	8.4	13.5	2.5	6.8	13.91	18.11	1.93	6.03

Table 3. The detection results.

Bloom Filter	Rate of false Alarm(%)	Rate of missed Alarm(%)	Average detection Latency of attack (s)
1	0	7.6	12.9
2	1.1	4.5	11.2
3	1.3	2.3	10.3
4	2.6	0	9.8
5	2.9	0	7.6
6	3.6	0	7.5
7	4.9	0	7.1
8	5.9	0	6.9
9	8.6	0	6.5
10	12.8	0	6.3

5. Conclusions

The defense mechanism of DDoS attacks, particularly the multi-based, multi-approached and diversified flow method of offensive artifice, simulating the competition of legal users, inhabits a keystone and difficulty in the internet security arena. In this paper we present five effective detecting features base on the characteristics of IP flow: PAP, ANPPF, PCF, ODGS and PGS. These five features can exploit the abnormalities during DDoS attack. Byproducts of features generation are helpful for filtering. We prove the capabilities of these five features through experimental comparison between their normal values and values in attack.

Our mechanism is characteristically distinct from current methods:

1) Utilizes few resources and does not require participation from all ISP routers. In general, only requires several routers.

2) It's low-cost, high-performance and easy-to-deploy. It allows for simple and convenient updating of the Filter Algorithm.

3) Optimizes the IP flow; enhances the server's efficiency by precluding and dismissing the overall current abruptness of ordinary flow.

All in all, allocating the server and bandwidth resources to both the validation and service components with more efficiency, and applying the algorithm more accurate to filter flooding DDoS are seeking to be done in this sector of internet security.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 60673184 and Grand No. 60873254, in part by the National 863 program of China under Grant No.2007AA01Z400, in part by the National Basic Research Program of China under Grant No. 2008CB317101, and Tsinghua-ChinaCache CDN Program.

7. References

[1] J. Mirkovic, S. Dietrich, D. Dittrich, and P. Reiher, "Internet denial of service: Attack and defense mechanisms," Prentice Hall PTR, 2004.

[2] V. A. Siris and F. Papagalou, "Application of anomaly detection algorithms for detecting SYN flooding attacks In: Regency H, ed," Global Telecommunications Conf. (GLOBECOM'04). Dallas: IEEE, pp. 2050–2054, 2004.

[3] W. Li, L. F. Wu, and G. Y. Hu, "Design and implementation of distributed intrusion detection system NetNumen," Journal of Software, pp. 1723–1728, 2002.

[4] M. Sung and J. Xu, "IP traceback-based intelligent packet filtering: A novel technique for defending against Internet DDoS attacks," IEEE Trans. on Parallel and Distributed Systems, pp. 861–872, 2003.

[5] A. Chandra and P. Shenoy, "Effectiveness of dynamic resource allocation for handling Internet," University of Massachussets, 2003.

[6] F. Liang and D. Yau, "Using adaptive router throttles against distributed denial-of-service attacks," Journal of Software, pp. 1120–1127, 2002.

[7] A. B. Kulkarni, S. F. Bush, and S. C. Evans, "Detecting distributed denial-of-service attacks using kolmogorov complexity metrics," General Electric Research and Development Center, December 2001.

[8] J. Mirkovic, "D-WARD: Source-end defense against

distributed denial-of-service attacks," PhD thesis, University of California, Los Angeles, pp. 310–321, August 2003.

[9] C. Jin, H. Wang, and K. G. Shin, "Hop-count filtering: An effective defense against spoofed DDoS traffic," Proceedings of the 10th ACM Conference on Computer and Communication Security, ACM Press, pp. 30–41, October, 2003.

[10] Y. Chen, K. Hwang, and Y. K. Kwok, "Filtering of shrew DDoS attacks in frequency domain," lcn, pp. 786–793, The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05), Jan. 2005.

[11] C. Sangpachatanaruk, S. M. Khattab, T. Znati, R. Melhem, and D. Mosse', "A simulation study of the proactive server roaming for mitigating denial of service attacks," Proceedings of the 36th Annual Simulation Symposium (ANSS'03), pp. 1430–1441, March 2003.

[12] Bell Labs. Bell Labs Internet Traffic Research. http://stat.bell-labs.com/InternetTraffic/index.html.

[13] ICSI Center for Internet Research Traffic Generators for Internet Traffic. http://www.icir.org/models/trafficgenerators.html.

Int. J. Communications, Network and System Sciences, **2009, 7, 608-618**
doi:10.4236/ijcns.2009.27068 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

A New Fairness-Oriented Packet Scheduling Scheme with Reduced Channel Feedback for OFDMA Packet Radio Systems

Stanislav NONCHEV, Mikko VALKAMA
Department of Communications Engineering, Tampere University of Technology, Tampere, Finland
Email: {*stanislav.nonchev, mikko.e.valkama*}*@tut.fi*

Received June 26, 2009; *revised August* 13, 2009; *accepted September* 26, 2009

ABSTRACT

In this paper, we propose a flexible and fairness-oriented packet scheduling approach for 3GPP UTRAN long term evolution (LTE) type packet radio systems, building on the ordinary proportional fair (PF) scheduling principle and channel quality indicator (CQI) feedback. Special emphasis is also put on practical feedback reporting mechanisms, including the effects of mobile measurement and estimation errors, reporting delays, and CQI quantization and compression. The performance of the overall scheduling and feedback reporting process is investigated in details, in terms of cell throughput, coverage and resource allocation fairness, by using extensive quasi-static cellular system simulations in practical OFDMA system environment with frequency reuse of 1. The performance simulations show that by using the proposed modified PF approach, significant coverage improvements in the order of 50% can be obtained at the expense of only 10-15% throughput loss, for all reduced feedback reporting schemes. This reflects highly improved fairness in the radio resource management (RRM) compared to other existing schedulers, without essentially compromising the cell capacity. Furthermore, we demonstrate the improved functionality increase in radio resource management for UE's utilizing multi-antenna diversity receivers.

Keywords: Radio Resource Management, Packet Scheduling, Proportional-Fair, Channel Quality Feedback, Throughput, Fairness

1. Introduction

Development of new radio interface technologies for beyond 3G cellular radio systems with support to high data rates, low latency and packet-optimised radio access has led to the use of OFDM/OFDMA. One good example of such developments is e.g. the UTRAN long term evolution (LTE), being currently standardized by 3GPP [1–3]. In general, performance improvements over the existing radio systems are basically obtained through proper deployment of fast link adaptation and new packet scheduling algorithms, exploiting the available multi-user diversity in both time and frequency domains [4–6]. On the other hand, achieving such performance improvements typically requires relatively accurate channel state feedback in terms of CQI reports from mobile stations (MS) to the base station (BS) [6–12]. As a practical example, each mobile station can measure the effective signal-to-interference-plus-noise-ratio (SINR), per active subcarrier or block of subcarriers, and send back the obtained channel state to the base station for downlink radio resource management. This, in turn, can easily lead to considerable control signalling overhead if not designed and implemented properly. Thus in general, the amount of the feedback information needs to be limited and is also subject to different errors and delays, affecting the overall system-level performance. Another important aspect in scheduling and resource allocation process is fairness, implying that also users with less favourable channel conditions should anyway be given some reasonable access to the radio spectrum [4–6,13–18]. This is especially important in serving users at, e.g., cell edges in cellular networks.

In this paper, we address the packet scheduling and channel state reporting tasks in OFDMA-based cellular packet radio systems. Stemming from ordinary proportional fair (PF) scheduling principle, a modified PF scheduler is first proposed having great flexibility to tune the exact scheduling characteristics in terms of capacity, coverage and fairness. More specifically, the proposed scheduler can offer greatly improved fairness among the users in a cell, measured in terms of coverage and other established fairness measures, like Jain's index [19], without essentially compromising the overall cell capacity. This is verified using extensive quasi-static cellular system simulations, conforming to the current LTE downlink specifications [1–3]. In the performance studies, different realistic CQI reporting schemes are also addressed and incorporated in the system simulations.

In general, the research on novel packet scheduling algorithms and channel state reporting schemes has been very active in the recent years, see e.g. [8,10,11,13–18] and the references therein. Using [13–17] as starting points for LTE type packet radio systems, it has been reported that frequency domain packet scheduling (FDPS) algorithms are always a compromise between the overall cell throughput and resource fairness among users. Here we propose a modified proportional fair algorithm, which in general offers an attractive balance between cell throughput, coverage and user fairness. Compared to plain frequency domain scheduling, we extend the studies by deploying both time domain and frequency domain scheduling steps, together with proper metrics, that as a whole can more efficiently utilise the provided yet limited feedback information from all the user equipments (UEs). Furthermore, we apply different realistic CQI reporting schemes to thoroughly investigate the limits of achieved performance gains from enhanced scheduling. The cellular system model used for the performance evaluations is fully conforming to the 3GPP evaluation criteria [1–3]. The overall outcomes are measured in terms of average *cell throughput*, *coverage* and *fairness index*.

The rest of the paper is organised as follows: Section 2 reviews the reference proportional fair scheduler and proposes then a modified PF scheduling scheme. Section 3, in turn, addresses different feedback reporting schemes in the scheduling context. Section 4 presents then the overall system model and simulation assumptions, and the simulation results and analysis are presented in Section 5. Finally, the conclusions are drawn in Section 6.

2. Scheduling Process

2.1. General Scheduling and Link Adaptation Principles

In general, the task of a packet scheduler (PS) is to select the most suitable users to access the available radio spectrum at any given time window, in order to optimize the system performance in terms of 1) throughput, 2) resource fairness, and/or 3) delay [4–6]. Joint optimization of all the above features is generally known very difficult. In *fast* packet scheduling, new scheduling decisions are basically taken in each transmission time interval (TTI), which in LTE is 1ms.

To efficiently utilize the limited radio resources, the scheduler should consider the current state of the channel when selecting the user to be scheduled, by utilizing e.g. the ACK/NACK signalling information and CQI reports [4–6,8,10,11,14]. Depending on the selected CQI reporting scheme, the accuracy and resolution of the channel quality information can easily differ considerably. In OFDMA based radio systems, like LTE, the CQI information is not necessarily available for all the individual subcarriers but more likely for certain groups of subcarriers only [12,20–22]. In general, the channel state information is also used by link adaptation (LA) mechanisms to select proper modulation and coding scheme (MCS) for each scheduled mobile, and thereon to ensure that the individual link qualities conform to the corresponding target settings. This is typically measured in terms of block error rate (BLER) for the first transmission. Hybrid ARQ (HARQ) mechanisms are then commonly used to provide the necessary buffer information and transmission format for pending retransmissions [4–6,16]. A principal block-diagram of the overall RRM flow is given in Figure 1.

As a practical example of the available spectral resources, in the 10 MHz system bandwidth case of LTE [1–3], there are 50 physical resource blocks (PRB's or sub-bands), each consisting of 12 sub-carriers with subcarrier spacing of 15 kHz. This sets the basic resolution in frequency domain (FD) UE multiplexing (scheduling), i.e., the allocated individual UE bandwidths are multiples of the PRB bandwidth.

2.2. Ordinary Proportional Fair (PF) Scheduler

The well-known proportional fair scheduler [13,16] works in two steps: 1) time domain (TD) PF step and 2) frequency domain (FD) PF step. Such simplified sche-

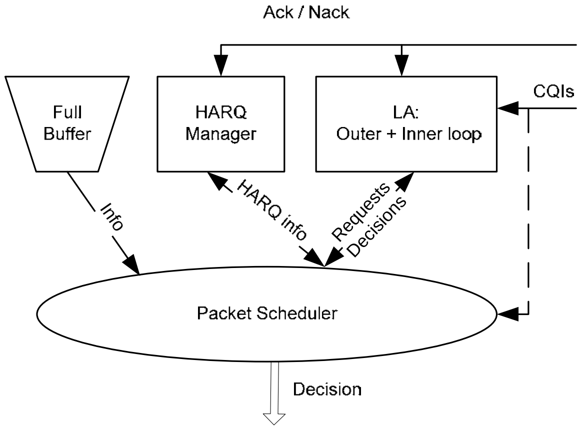


Figure 1. Principal RRM block diagram.

duling principle is beneficial from the complexity point of view, since the FD step considers a reduced number of UEs for frequency multiplexing in each TTI [17]. Thus in the first part, inside each TTI *n*, all the UE's are ranked according to the following priority metric

$$\gamma_i^{td}(n) = \frac{R_i(n)}{T_i(n)} \quad \textbf{(1)}$$

In above, the UE index $i = 1, 2, \ldots, I_{\mathrm{TOT}}$, $R_i(n)$ denotes the estimated throughput to the UE *i* over the *full bandwidth* (provided by link adaptation unit) [13,16], and $T_i(n)$ in turn is the corresponding average delivered throughput to the UE *i* during the recent past and can be obtained, e.g., recursively by

$$T_i(n) = \left(1 - \frac{1}{t_c}\right) T_i(n-1) + \frac{1}{t_c} R_i'(n-1) \quad \textbf{(2)}$$

In (2), t_c controls the averaging window length over which the average delivered throughput is calculated and $R'_i(n\text{-}1)$ denotes the actually realized throughput to the UE *i* at the previous TTI.

In the next step, out of this ranked list of UE's, the first I_{BUFF} (< I_{TOT}) UE's with highest priority metric are picked to the actual frequency domain multiplexing or scheduling stage. In the following, this subset is called scheduling candidate set (SCS), and is denoted by Ω(*n*). Then, for each physical resource block $k = 1, 2, \ldots, K_{\mathrm{TOT}}$, and for each *i* belonging to the SCS, the following final scheduling metric of the form

$$\gamma_{i,k}^{fd}(n) = \frac{R_{i,k}(n)}{T_i(n)} \quad \textbf{(3)}$$

is evaluated where now $R_{i,k}(n)$ denotes the *estimated* throughput to the UE *i* for the *k*-th PRB (provided by LA unit again), and $T_i(n)$ is again the corresponding average throughput *delivered* to the UE *i* during the recent past given in (2). Finally, the access to each PRB resource is granted for the particular user with the highest metric for the corresponding PRB.

2.3. Proposed Modified PF (MPF) Scheduler

In order to obtain a scheduler with yet increased fairness in the resource allocation, we proceed as follows. First the time domain priority metric is modified as

$$\overline{\gamma}_i^{td}(n) = CQI_i(n)\left(\frac{T_i(n)}{T_{tot}(n)}\right)^{-1} \quad \textbf{(4)}$$

where $CQI_i(n)$ denotes the full bandwidth channel quality report for UE *i* at TTI *n* and $T_i(n)$ is as defined in (2). $T_{tot}(n)$, in turn, denotes the averaged throughput over the past and over the scheduled users and can be calculated by

$$\begin{aligned} T_{tot}(n) &= \left(1 - \frac{1}{t_c}\right) T_{tot}(n-1) \\ &+ \frac{1}{t_c}\frac{1}{I_{BUFF}} \sum_{i \in \Omega(n-1)} R_i'(n-1) \end{aligned} \quad \textbf{(5)}$$

In (5), $R'_i(n\text{-}1)$ denotes the actual *delivered* throughput for UE *i* at the previous TTI.

Similar to the ordinary PF scheduler described in Subsection 2.2, this modified metric in (4) is used to rank the UE's inside each TTI, and the I_{BUFF} (< I_{TOT}) UE's with highest priority metric form a SCS. Ω(*n*) for the actual frequency domain resource allocation. Since estimated throughput in the link adaptation stage is based on reported CQI values, we assume that the substitution in (4) has the same weight in priority calculation. For mapping the users of the SCS into PRB's, the following modified frequency domain metric is then proposed:

$$\overline{\gamma}_{i,k}^{fd}(n) = \left(\frac{CQI_{i,k}(n)}{CQI_i^{avg}(n)}\right)^{s_1}\left(\frac{T_i(n)}{T_{tot}(n)}\right)^{-s_2} \quad \textbf{(6)}$$

Here s_1 and s_2 are adjustable parameters, and $CQI_{i,k}(n)$ is the channel quality report of user *i* for sub-band *k* at TTI *n* while $CQI_i^{avg}(n)$ is the corresponding average CQI over the past and over the sub-bands, and can be calculated using

$$\begin{aligned} CQI_i^{avg}(n) &= \left(1 - \frac{1}{t_c}\right) CQI_i^{avg}(n-1) \\ &+ \frac{1}{t_c}\frac{1}{K_{TOT}} \sum_{k=1}^{K_{TOT}} CQI_{i,k}(n) \end{aligned} \quad \textbf{(7)}$$

The access to each PRB resource is then granted for the particular user with the highest metric in (6) for the corresponding PRB.

Considering the re-transmissions, re-transmitting users are simply considered as additional users in the time domain scheduling part (step 1), and if qualified to the frequency domain SCS, the re-transmission users are given an additional priority to reserve exactly the same sub-bands used for the corresponding original transmissions. Even though this does not take the exact sub-band condition into account at re-transmission stage, the practical implementation is simplified, in terms of control signalling, and re-transmissions anyway always benefit from the HARQ combining gain [6].

Intuitively, the proposed scheduling metrics in (4) and (6) are composed of two elements, affecting the overall scheduling decisions. The first dimension measures the relative instantaneous quality of the individual user's radio channels against their own average channel qualities while the second dimension is related to measuring the achievable throughput of individual UE's against the corresponding average throughput of scheduled users. Consequently, by understanding the power coefficients s_1

and s_2 as additional adjustable parameters, the exact scheduler statistics can be tuned and controlled to obtain a desired balance between the throughput and fairness. This will be demonstrated in Section 5.

3. Feedback Reporting Process

The overall reporting process between UE's and BS is illustrated in Figure 2. Within each time window of length t_r, each mobile sends channel quality indicator (CQI) reports to BS, formatted and possibly compressed, with a reporting delay of t_d seconds [6,8,10,11]. Each report is naturally subject to errors due to imperfect decoding of the received signal. In general, the CQI reporting frequency-resolution has a direct impact on the achievable multi-user frequency diversity and thereon to the overall system performance and the efficiency of radio resource management (RRM), as described in general e.g. in [11]. In our studies here, the starting point (reference case) is that the CQI reports are quantized SINR measurements across the entire bandwidth (wideband CQI reporting), to take advantage of the time and frequency variations of the radio channels for the different users. Then also alternative reduced feedback schemes are described and evaluated, as discussed below.

3.1. Full CQI Reporting

In a general OFDMA radio system, the overall system bandwidth is assumed to be divided into v CQI measurement blocks. Then quantizing the CQI values to q bits, the overall full CQI report size is

$$S_{full} = q \times v \qquad \textbf{(8)}$$

bits which is reported by every UE for each TTI [1–3,11]. In case of LTE, with 10 MHz system bandwidth and grouping 2 physical resource blocks into 1 measurement block, it follows that v = 25. Assuming further that quantization is carried with q = 5 bits, then each UE is send ing 25x5 = 125 bits for every 1ms (TTI length).

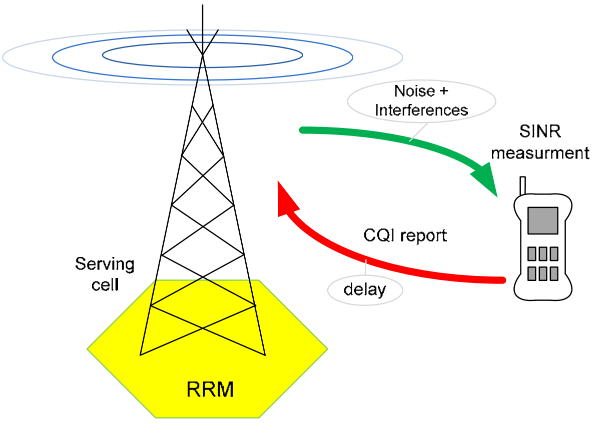


Figure 2. Reporting mechanism between UE and BS.

3.2. Best-m CQI Reporting

One simple approach to reduce the reporting and feedback signalling is obtained as follows. The method is based on selecting only m < v different CQI measurements and reporting them together with their frequency positions to the serving cell [8,11]. We assume here that the evaluation criteria for choosing those m sub-bands for reporting is based on the highest SINR values (hence the name best-m). The resulting report size in bits is then given by

$$S_{best-m} = q \times m + \left\lceil \log_2 \left(\frac{v!}{m!(v-m)!} \right) \right\rceil \qquad \textbf{(9)}$$

As an example, with v = 25, q= 5 bits and m = 10, it follows that $S_{best\text{-}m}$ = 72 bits, while S_{full} = 125 bits. Furthermore, on the scheduler side, we assume that the PRBs which are not reported by the UE are allocated a CQI value equal to the lowest reported one.

3.3. Threshold Based CQI Reporting

This reporting scheme is a further simplification and relies on providing information on only the average CQI value above certain threshold together with the corresponding location (sub-band index) information. First the highest CQI value is identified within the full bandwidth, which sets an upper bound of the used threshold window. All CQI values within the threshold window are then averaged and only this information is sent to the BS together with the corresponding sub-band indexes. On the scheduler side, the missing CQI values can then be treated, e.g., as the reported averaged CQI value minus a given dB offset (e.g. 5 dB, the exact number is again a design parameter). The number of bits needed for reporting is therefore only

$$S_{threshold} = q + v \qquad \textbf{(10)}$$

As an example, with v = 25 and q = 5 bits (as above), it follows that $S_{threshold}$ = 30 bits, while $S_{best\text{-}m}$ = 72 bits and S_{full} = 125 bits. The threshold-based scheme is illustrated graphically in Figure 3 [10].

4. System Simulation Model and Assumptions

In order to evaluate the system-level performance of the proposed scheduling scheme in a practical OFDMA-based cellular system context, a comprehensive quasi-static system simulator for LTE downlink has been developed,

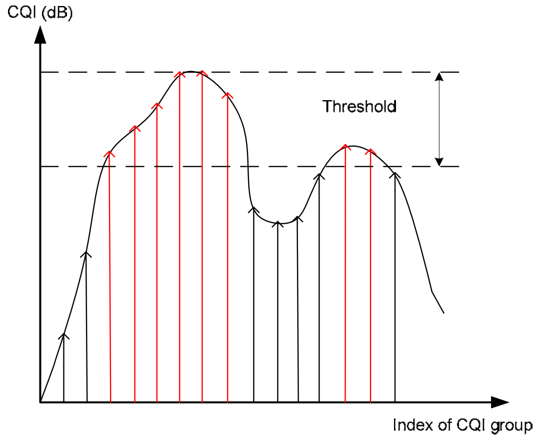


Figure 3. Basic principle of threshold-based CQI reporting.

Table 1. Basic simulation parameters.

Parameter	**Assumption**
Cellular Layout	Hexagonal grid, 19 cell sites, 3 sectors per site
Inter-site distance	500 m
Carrier Frequency / Bandwidth	2000 MHz / 10 MHz
Number of active sub-carriers	600
Sub-carrier spacing	15 kHz
Sub-frame duration	0.5 ms
Channel estimation	Ideal
PDP	ITU Typical Urban 20 paths
Minimum distance between UE and cell	>= 35 meters
Average number of UE's per sector	20
Max. number of frequency multiplexed UEs (I_{BUFF})	10
UE receiver type	2-Rx MRC, 2-Rx IRC
Shadowing standard deviation	8 dB
UE speed	3km/h
Total BS TX power (P_{total})	46dBm
Traffic model	Full Buffer
Fast Fading Model	Jakes Spectrum
CQI reporting schemes	Full CQI Best-m (with m=10) Threshold based (with 5dB threshold)
CQI log-normal error std.	1 dB
CQI reporting time	5 TTI
CQI delay	2 TTIs
CQI quantization	1 dB
CQI std error	1 dB
MCS rates	QPSK (1/3, 1/2, 2/3), 16QAM (1/2, 2/3, 4/5), 64QAM (1/2, 2/3, 4/5)
ACK/NACK delay	2ms
Number of SAW channels	6
Maximum number of retransmisions	3
HARQ model	Ideal chase combining (CC)
1st transmission BLER target	20%
Scheduler forgetting factor	0.002
Scheduling schemes used	Ordinary PF (for reference) Modified PF (proposed
Simulation duration (one drop)	5 seconds
Number of drops	10

conforming to the specifications in [1–3]. In the overall simulation flow, mobile stations are first randomly dropped or positioned over each sector and cell. Then based on the individual distances between the mobiles and the serving base station, the path losses for individual links are directly determined, while the actual fading characteristics of the radio channels depend on the assumed mobility and power delay profile. In updating the fading statistics, the time resolution in our simulator is set to one TTI (1ms). In general, a standard hexagonal cellular layout is utilized with altogether 19 cell sites each having 3 sectors. In the performance evaluations, statistics are collected only from the central cell site while the others simply act as sources of inter-cell interference.

As a practical example case, the 10 MHz LTE system bandwidth mode [1–3] is assumed. The main simulation parameters and assumptions are generally summarized in Table 1 for the so-called Macro cell case 1, following again the LTE working assumptions. As illustrated in Figure 1, the RRM functionalities are controlled by the packet scheduler and also link adaptation and HARQ mechanisms are modelled and implemented, as described in Table 1. As a practical example, the maximum number of simultaneously multiplexed users (I_{BUFF}) is set to 10 here. In general, we assume that the BS transmission power is equally distributed among all PRB's. In the basic simulations, 20 UE's are uniformly dropped within each sector and experience inter-cell interferences from the surrounding cells, in addition to path loss and fading. The UE velocity equals 3km/h, and the typical urban (TU) channel model standardized by ITU is assumed in modelling the power-delay spread of the radio channels. Infinite buffer traffic model is applied in the simulations, i.e. every user has data to transmit (when scheduled) for the entire duration of a simulation cycle. The length of a single simulation run is set to 5 seconds which is then repeated for 10 times to collect reliable statistics.

In general, every UE has an individual HARQ entry, which operates the physical layer re-transmission functionalities. It is based on the stop-and-wait (SAW) protocol and for simplicity, the number of entries per UE is fixed to six. HARQ retransmissions are always transmitted with the same MCS and on the same PRB's (if scheduled in TD step) as the first transmissions. The supported modulation schemes are QPSK, 16QAM and 64QAM with variable rates for the encoder as shown in Table 1.

Link adaptation handles the received UE reports containing the channel quality information for the whole or sub-set of PRB's as described in Section 3. The implemented link adaptation mechanism consists of two separate elements – the inner loop (ILLA) and outer loop (OLLA) LA's – and are used for removing CQI imperfections and estimating supported data rates and MCS. As a practical example, it is assumed that the CQI report errors are log-normal distributed with 1dB standard deviation.

The actual effective SINR calculations rely on estimated subcarrier-wise channel gains (obtained using reference symbols in practice) and depend in general also on the assumed receiver topology. Here we assume the single-input-multiple-output (SIMO) diversity reception case, i.e. a single BS transmit antenna and multiple UE receiver antennas. Considering now an individual UE *i*, the SINR per active sub-carrier *c* at TTI *n*, denoted here by $\xi_{i,c}(n)$, is calculated according to

$$\xi_{i,c} = \frac{\left|\mathbf{w}_{i,c}^{H}\mathbf{h}_{i,c}\right|^{2}\sigma_{sig,i}^{2}}{\mathbf{w}_{i,c}^{H}\sum_{noise}\mathbf{w}_{i,c} + \mathbf{w}_{i,c}^{H}\sum_{int,i}\mathbf{w}_{i,c}} \qquad \textbf{(11)}$$

where the time index *n* is dropped for notational simplicity. Here h_{ic} is an N_{RX} x 1 vector of the user *i* complex channel gains at subcarrier *c* from BS to N_{RX} receiver antennas and w_{ic} is the corresponding N_{RX} x 1 spatial filter used to combine the signals of different receiver antennas (more details below). $s^{2}_{sig,i}$, in turn, denotes the received nominal signal power per antenna while $\mathbf{\Sigma}_{noise}$ and $\mathbf{\Sigma}_{int,i}$ are the covariance matrices of the received (spatial) noise and interference vectors. The superscript $(.)^{H}$ denotes conjugate transpose. The noise covariance is assumed diagonal ($\mathbf{\Sigma}_{noise} = s^{2}_{noise}\mathbf{I}$) and independent of the user index *i*. The interference modeling, on the other hand, takes into account the interference from neighboring cells. Assuming a total of L_{int} interference sources, with corresponding path gain vectors $\mathbf{g}_{l,i,c}$, the overall interference covariance at receiving UE *i* is given by

$$S_{int,i} = \sum_{l=1}^{L_{int}} \sigma_{int,l,i}^{2}\mathbf{g}_{l,i,c}\mathbf{g}_{l,i,c}^{H} \qquad \textbf{(12)}$$

where $s^{2}_{int,l,i}$, denotes the received nominal interferer power per antenna and per interference source (*l*).

Concerning the actual UE receiver topologies (spatial filters), both maximum ratio combining (MRC) and interference rejection combining (IRC) receivers are deployed in the simulations. These are given by (see, e.g., [6] and the references therein)

$$\mathbf{w}_{i,c}^{MRC} = \frac{\mathbf{h}_{i,c}}{\left\|\mathbf{h}_{i,c}\right\|^{2}} \qquad \textbf{(13)}$$

and

$$\mathbf{w}_{i,c}^{IRC} = \frac{\sum_{tot,i}^{-1}\mathbf{h}_{i,c}}{\mathbf{h}_{i,c}^{H}\sum_{tot,i}^{-1}\mathbf{h}_{i,c}} \qquad \textbf{(14)}$$

where $\mathbf{\Sigma}_{tot,i}$ denotes the total noise plus interference covariance, i.e., $\mathbf{\Sigma}_{tot,i} = s^{2}_{noise}\mathbf{I} + \mathbf{\Sigma}_{int,i}$.

Using the above modeling and the selected UE receiver type, the effective SINR values are then calculated through exponential effective SINR mapping (EESM), as described in [1–3], for link-to-system level mapping purposes.

5. Results

In this section, we present the system-level performance results obtained using the previously described quasi-static radio system simulator. Both ordinary PF and modified (proposed) PF packet schedulers are used, together with the three different CQI reporting schemes. The system-level performance is generally measured and evaluated in terms of:

- Throughput statistics – the cumulative distribution function (CDF) of the total number of successfully delivered bits per time unit. Measured at both individual UE level as well as overall cell level.
- Coverage – the experienced data rate per UE at the 95% coverage probability (5% UE throughput CDF level).
- Jain's fairness index [19].

In addition to Jain's index, also the coverage and slope of the throughput CDF reflect the fairness of the scheduling algorithms.

With the proposed modified PF scheduler, different example values for the power coefficients s_1 and s_2 are used as shown in Table 2. To focus mostly on the role of the channel quality reporting, s_2 is fixed here to 1 and the effects of using different values for s_1 are then demonstrated. This way the impact of the different CQI reporting schemes is seen more clearly. For the cases of *Best–m* and *Threshold* based CQI reporting schemes, we fix the value of *m* equal to 10 and threshold to 5 dB, respectively. Similar example values have also been used by other authors in the literature earlier, see e.g. [11]. Complete performance statistics are gathered for both dual antenna MRC and dual antenna IRC UE receiver cases.

Table 2. Different power coefficient combinations used to evaluate the performance of the proposed scheduler.

Coefficient	Value						
s_1	1	2	4	6	8	10	20
s_2	1	1	1	1	1	1	1

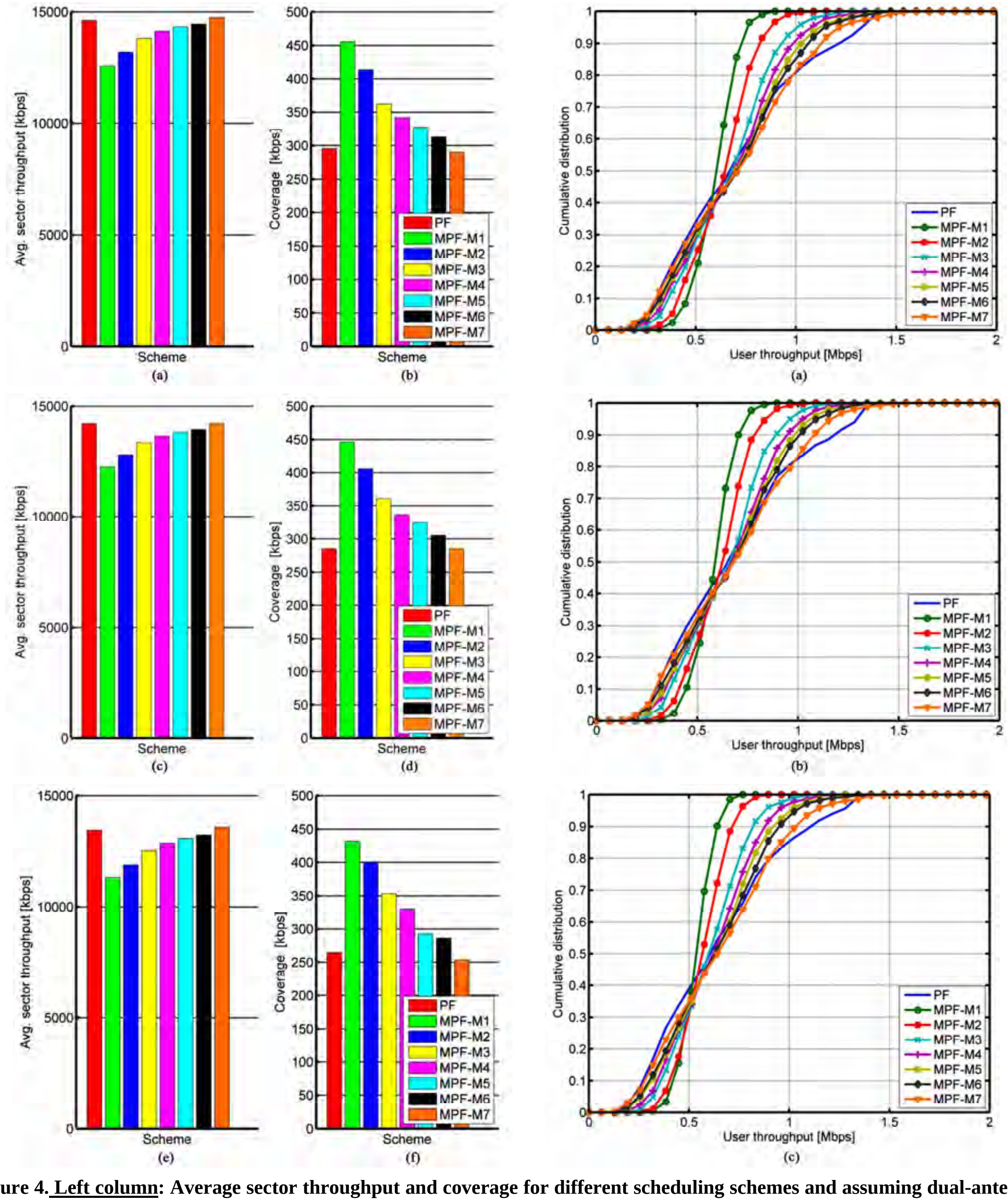


Figure 4. Left column: Average sector throughput and coverage for different scheduling schemes and assuming dual-antenna MRC UE receiver type with *full* CQI feedback (a, b), *Best -m* CQI feedback (c, d) and *Threshold* based CQI feedback (e, f). M1-M7 refer to the modified PF scheduler with power coefficient values as given in Table 2 (M1: s_1=1, s_2=1, etc.). Right column: CDF's of individual UE throughputs for different scheduling schemes and assuming dual-antenna MRC UE receiver type with *full* CQI feedback (a), *Best -m* CQI feedback (b) and *Threshold* based CQI feedback (c).

5.1. Dual Antenna MRC UE Receiver Case

Figure 4 (left column) illustrates the average sector throughput and coverage for the different schedulers, assuming dual antenna maximum ratio combining (MRC) UE receiver type. The power coefficient values from Table 2 are presented as index M, where M1 represents the first couple (s_1=1, s_2=1), etc, for the metric calculation of the modified PF scheduler. The used reference scheduler is the ordinary proportional fair approach. In the first coefficient case (M1), in combination with full CQI reporting scheme, we achieve coverage gain in the order of 50% at the expense of only 15% throughput loss as shown in Figure 4 (a) and (b). This sets the basic reference for comparisons in the other cases. In the case of best-m and threshold based reporting schemes presented in and (d), and Figure 4 (e) and (f), we have coverage increases by 57% and 63% with throughput losses of 16% and 19%, correspondingly.

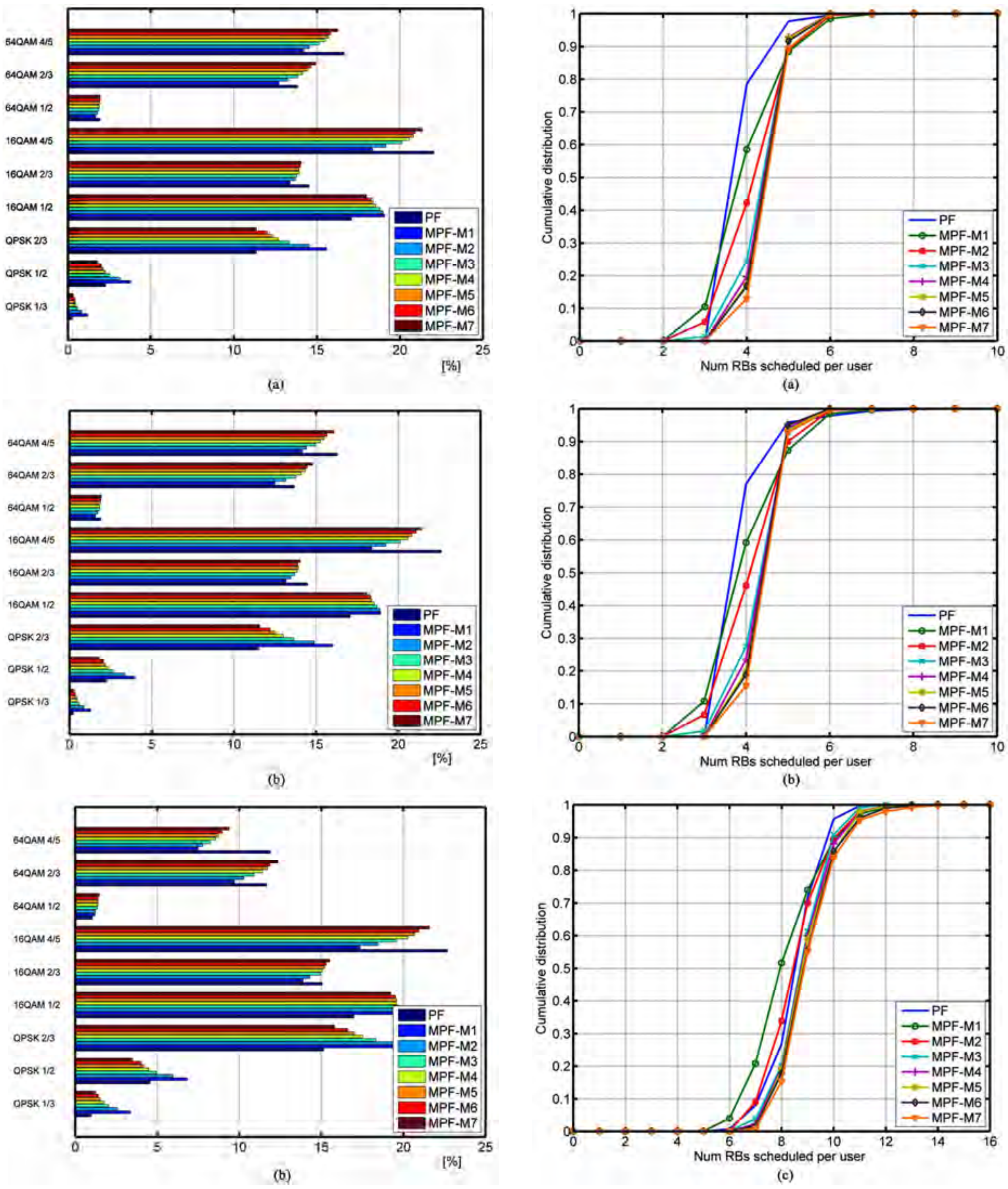


Figure 5. Left column: MCS distributions [%] for different scheduling principles with (a) *Full* CQI reporting, (b) *Best-m* CQI reporting, and (c) *Threshold* based CQI reporting assuming dual-antenna MRC UE receiver. Right column: CDF's of scheduled PRB's per user for different schedulers with (a) *Full* CQI reporting, (b) *Best-m* CQI reporting, and (c) *Threshold* based CQI reporting assuming dual-antenna MRC UE receiver.

Continuing on the evaluation of relative system performance using the modified PF scheduler, we clearly see a trade-off between average cell throughput and coverage for different power coefficient cases. The remaining power coefficient values shown in Table 2 are used for tuning the overall system behaviour together with the choice of the CQI reporting scheme. In the case of full CQI feedback and coefficient s_1 varying between 2 and 10 (M2–M6) the cell throughput loss is decreased to around 1%, while the coverage gain is reduced to around 6%. Similar behaviour is observed for the other feedback reporting schemes as well. The exact percentage values for the coverage gains and throughput losses are stated in Table 3 in the end.

Further illustrations on the obtainable system performance are presented in Figure 4 (right column) in terms of the statistics of individual UE data rates for the applied simulation scenarios. The slope of the CDF reflects generally the fairness of the algorithms. Therefore we aim to achieve steeper slope corresponding to algorithm fairness. This type of slope change behavior can clearly be established for each simulation scenario. Clearly, at 5% (coverage) point of the CDF curves, corresponding to users typically situated at the cell edges, we observe significant data rate increases indicated by shift to the right for all CQI feedback schemes when the coefficient s_1 is changed in the proposed metric. This indicates improved overall cell coverage at the expense of slight total throughput loss.

Figure 5 (left column) shows the modulation and coding scheme (MCS) distributions for different schedulers and with applied feedback reporting schemes, still assuming the case of 2 antenna MRC UE receiver type. The negligible decrease in higher order modulation usage (less than 3%) leads to the increase in the lower (more robust) ones for improving the cell coverage. In all the simulated cases, the MCS distribution behaviour has a relatively similar trend following the choice of the power coefficients in the proposed packet scheduling. In general, the use of higher-order modulations is affected mostly in the most coarse CQI feedback (threshold based) case while the other two reporting schemes behave fairly similarly.

Similarly, Figure 5 (right column) illustrates the CDF's of scheduled PRB's per UE for the different scheduler scenarios and reporting schemes. Clearly, the modified PF provides better resource allocation in the *full* and *best-m* feedback cases. Considering the 50% probability point for the resource allocation, and taking the case of M1, we have about 5% gain, while in case of M2 the gain is raised to 15% compared to ordinary PF. The average obtained improvement for the rest of the cases is about 33%. In the case of *threshold*-based feedback, the resource allocation is not as efficient, and even a small reduction in the RB allocation is observed with small power coefficients, compared to the reference PF scheduler. Starting from M3, the improvement is anyway noticeable and the achieved gain is about 20%.

Table 3. Obtained performance statistics compared to ordinary PF scheduler with different CQI reporting schemes and different power coefficients (M1-M7) for the proposed scheduler. Dual-antenna MRC UE receiver case.

	Coverage Gain [%]			Throughput Loss [%]		
	full	best-m	threshold	full	best-m	threshold
M1	54	57	63	16	16	19
M2	40	42	51	10	10	12
M3	23	26	33	6	6	7
M4	16	18	25	3	4	5
M5	11	14	11	2	3	3
M6	6	7	8	1	2	2
M7	-2	0	-4	0	0	0

Table 4. Obtained performance statistics compared to ordinary PF scheduler with different CQI reporting schemes and different power coefficients (M1-M7) for the proposed scheduler. Dual-antenna IRC UE receiver case.

	Coverage Gain [%]			Throughput Loss [%]		
	full	best-m	threshold	full	best-m	threshold
M1	56	58	64	15	15	18
M2	43	46	48	9	9	11
M3	26	30	32	6	6	8
M4	17	20	24	4	4	5
M5	10	12	13	2	3	3
M6	8	10	8	2	2	2
M7	-1	1	1	0	1	0

5.2. Dual Antenna IRC UE Receiver Case

Next similar performance statistics are obtained for dual antenna interference rejection combining (IRC) UE receiver case. Starting from the primary case M1, with full CQI, we obtain a 13% loss in throughput and 57% coverage improvement. For the reduced feedback reporting schemes – best-m and threshold based – we have 13% and 15% throughput losses and 58% and 62% coverage gains, respectively. Furthermore, resource allocation gains for full CQI feedback and best-m are 7% for M1 and 17% for M2 correspondingly. The average obtained improvement for the rest of the cases is about 34%. *Threshold* based reporting scheme leads to decrease of 12% for M1 and 7% for M2, and roughly 14% increase for the rest of simulated cases. The exact percentage read from the figures are again stated in table format in Table 4 in the end.

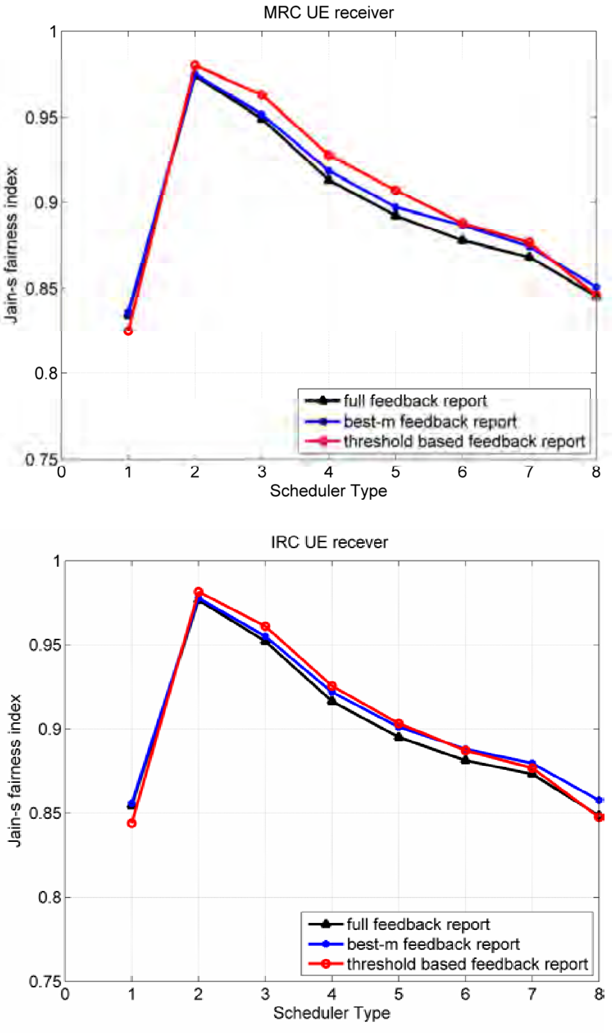


Figure 6. Jain's fairness index per feedback reporting scheme for dual-antenna MRC UE receiver case (up) and dual-antenna IRC UE receiver case (down). Scheduler type 1 means ordinary PF, while 2-8 means proposed modified PF with power coefficients as described in Table 2.

5.3. Fairness Index

Figure 6 illustrates the Jain's fairness index per scheduler for the applied feedback reporting schemes, calculated over all the I_{TOT} = 20 UE's using the truly realized UE throughputs at each TTI and over all the simulation runs. The value on the x-axis corresponds to the used scheduler type, where 1 refers to the reference PF scheduler and 2-8 refer to the proposed modified PF schedulers with different power coefficients. The Jain's fairness index defined in [19] is generally in the range of [0…1], where the value of 1 corresponds to all users having the same amount of resources (maximum fairness). Clearly, the fairness distribution with the proposed modified PF scheduler outperforms the used reference PF scheduler for both UE receiver types. The received fairness gains are in range of 2%-17% for the MRC receiver case, and 1%-14% for the IRC receiver case, respectively.

6. Conclusions

In this article, we have studied the potential of advanced packet scheduling principles in OFDMA type radio system context, using UTRAN long term evolution (LTE) as a practical example system scenario. A modified proportional fair scheduler taking both the instantaneous channel qualities (CQI's) as well as resource allocation fairness into account was proposed. Also different practical CQI reporting schemes were discussed, and used in the system level performance evaluations of the proposed scheduler. All the performance evaluations were carried out with a comprehensive quasi-static system level simulator, conforming fully to the current LTE working assumptions. Also different UE receiver types were demonstrated in the performance assessments. In general, the achieved throughput and coverage gains were assessed against more traditional ordinary proportional fair scheduling. In the case of fixed coverage requirements and based on the optimal parameter choice for CQI reporting schemes, the proposed scheduling metric calculations based on UE channel feedback offers better control over the ratio between the achievable cell/UE throughput and coverage increase. As a practical example, even with limited CQI feedback, the cell coverage can be increased significantly (more than 30%) by allowing a small decrease (in the order of only 5-10%) in the cell throughput. This is seen to give great flexibility to the overall RRM process and optimization.

7. Acknowledgments

Fruitful discussions with Markku Kuusela, Nokia Devices, Helsinki, Finland, and Dr. Toni Huovinen, Tampere University of Technology, Tampere, Finland, are greatly acknowledged.

8. References

[1] 3GPP RAN Technical Specification Group, "E-UTRA/ E-UTRAN Overall description, stage 2," Technical Report TR 36.300, ver. 9.0.0, June 2009.

[2] 3GPP RAN Technical Specification Group, "E-UTRA/ LTE physical layer—General description," Technical Report TR 36.201, ver. 8.3.0, March 2009.

[3] 3GPP RAN Technical Specification Group, "Physical layer aspects for evolved UTRA," Technical Report TR 25.814, ver. 7.1.0, Oct. 2006.

[4] N. D. Tripathi, *et al*., "Radio resource management in cellular systems," Springer, 2001.

[5] H. Holma and A. Toskala, Eds., "HSDPA/HSUPA for UMTS–High speed radio access for mobile communications," Wiley, 2006.

[6] E. Dahlman, *et al*., "3G evolution: HSPA and LTE for mobile broadband," Academic Press, 2007.

[7] S. Yoon, C. Suh, Y. Cho, and D. Park, "Orthogonal frequency division multiple access with an aggregated subchannel structure and statistical channel quality measurements," in Proc. IEEE Vehicular Technology Conference (VTC'04 Fall), Los Angeles, CA, September 2004.

[8] Y. Sun, *et al*., "Multi-user scheduling for OFDMA downlink with limited feedback for evolved UTRA," in Proc. IEEE Vehicular Technology Conference (VTC'06 Fall), Montreal, Canada, September 2006.

[9] I. Toufik and H. Kim, "MIMO-OFDMA opportunistic beamforming with partial channel state information," in Proc. IEEE International Conference on Communications, Instanbul, Turkey, pp. 5389–5394, June 2006.

[10] T. E. Kolding, F. Frederiksen, and A. Pokhariyal, "Low-bandwidth channel quality indication for OFDMA frequency domain packet scheduling," in Proc. ISWCS'06, Spain, September 2006.

[11] K. I. Pedersen, G. Monghal, I. Z. Kovacs, T. E. Kolding, A. Pokhariyal, F. Frederiksen, and P. Mogensen, "Frequency domain scheduling for OFDMA with limited and noisy channel feedback," in Proc. IEEE Vehicular Technology Conference (VTC'07 Fall), Baltimore, MD, pp. 1792–1796, Sept. 2007.

[12] P. Svedman, D. Hammarwall, and B. Ottersten, "Subcarrier SNR estimation at the transmitter for reduced feedback OFDMA," in Proc. European Signal Processing Conf., Florence, Italy, September 2006.

[13] C. Wengerter, J. Ohlhorst, and A. G. E Von Elbwert," Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA," in Proc. IEEE Vehicular Technology Conference (VTC' 05 Spring), Stockholm, Sweden, May 2005.

[14] S. Nonchev, J. Venäläinen, and M. Valkama, "New frequency domain packet scheduling schemes for UTRAN LTE Downlink," in Proc. ICT Mobile Summit, Stockholm, Sweden, June 2008.

[15] S. Nonchev and M. Valkama, "Efficient packet scheduling schemes for multiantenna packet radio downlink," in Proc. Fifth Advanced Int. Conf. Telecommunications (AICT'09), Venice, Italy, May 2009.

[16] T. E. Kolding, "Link and system performance aspects of proportional fair scheduling in WCDMA/HSDPA," in Proc. IEEE Vehicular Technology Conference (VTC'03 Fall), Orlando, FL, pp. 1717–1723, Oct. 2003.

[17] A. Pokhariyal, K. I. Pedersen, G. Monghal, I. Z. Kovacs, C. Rosa, T. E. Kolding, and P. E. Mogensen, "HARQ aware frequency domain packet scheduler with different degrees of fairness for the UTRAN long term evolution," in Proc. IEEE Vehicular Technology Conference (VTC' 07 Spring), Dublin, Ireland, April 2007, pp. 2761–2765.

[18] A. Pokhariyal, T. E. Kolding, and P. E. Mogensen, "Performance of downlink frequency domain packet scheduling for the UTRAN long term evolution," in Proc. IEEE Personal, Indoor and Mobile Radio Communications Conference (PIMRC'06), Helsinki, Finland, Sept. 2006.

[19] D. Chui and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," Computer Networks and ISDN Systems, 1989.

[20] P. Svedman, S. K. Wilson, L. J. Cimini, and B. Ottersten, "A simplified opportunistic feedback and scheduling scheme for OFDMA," in Proc. IEEE Vehicular Technology Conference (VTC'04 Spring), pp. 1878–1882, May 2004.

[21] P. Svedman, L. J. Cimini, and B. Ottersten, "Using unclaimed sub-carriers in opportunistic OFDMA systems," in Proc. IEEE Vehicular Technology Conference (VTC' 06 Fall), Montreal, Canada, September 2006.

[22] S. Sanayei, A. Nosratinia, and N. Aldhahir, "Opportunistic dynamic sub-channel allocation in multiuser OFDM networks with limited feedback," in IEEE Proc. Inform. Theory Workshop, San Antonio, TX, pp. 182–186, October 2004.

[23] P. Morgensen, *et al*., "LTE capacity compared to the Shannon bound," in Proc. IEEE Vehicular Technology Conference (VTC'07 Spring), Dublin, Ireland, April 2007.

Int. J. Communications, Network and System Sciences, **2009, 7, 619-626**
doi:10.4236/ijcns.2009.27069 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

A Hybrid ARQ System with Erasures Correction and Parity Retransmission

L. GOLDFELD, A. HOFMAN, V. LYANDRES
Department of Electrical and Computer Engineering, Communications Laboratory,
Ben-Gurion University of the Negev, Beer-Sheva, Israel
Email: *lyandres@ee.bgu.ac.il*

Received June 30, 2009; *revised August* 4, 2009; *accepted September* 8, 2009

ABSTRACT

A modified type of Hybrid ARQ system with erasures correction and parity bits retransmission is considered. Performance of the system is analyzed under assumption that the forward channel suffers from Nakagami common fading and additive white Gaussian noise. A good agreement between theoretical results and simulation is achieved. The proposed ARQ protocol is compared with other known Hybrid ARQ algorithms. It demonstrates significantly higher throughput efficiency in a range of SNR.

Keywords: Hybrid ARQ Scheme, Erasures Correction Decoding, Fading Channel

1. Introduction

Automatic Repeat ReQuest (ARQ) systems with error control are widely used for data transmission over noisy channels. Their performance is usually characterized by two parameters: Throughput Efficiency (*TE*) and Bit Error Rate (*BER*). So-called hybrid ARQ (HARQ) systems use two codes (inner and outer) [1–8]. In a HARQ-I system Forward Error Correction (FEC) is performed prior to error detection [2]. In more effective HARQ-II [3–5] system, parity-check digits for error correction are sent to the receiver only when they are needed. For example [3], at the initial step data blocks *D* with parity-check bits of the outer error detection code are transmitted. If errors in *D* are detected, the system begins not simple repetitions of *D*, but repetitions of a parity block *P*(*D*) of the inner code. *P*(*D*) as well as *D* itself are alternately stored in the receiver buffer for error correction until *D* will be recovered. As shown in [4], application of HARQ schemes can significantly improve *TE* in comparison with a pure ARQ scheme.

It is well known [1] that error correcting capability of block codes may be enhanced when soft decision approach is realized, for example, in the form of Soft Decision Erasures Correcting (SDEC) decoding [7–9]. In this case the number of erroneous bits that may be corrected is not less than d_H-1. It is assumed also that all error symbols are erased, while for FEC decoding the number of erroneous corrected bits is not less than $0.5(d_H-1)$, where d_H is the minimum Hamming distance of the block code.

In this paper, a modified type of HARQ-II is considered. Two linear block codes are used in the system: an *outer* systematic (*n*,*k*) block code C_2 and an *inner* half-rate invertible (2*k*,*k*) code C_1. At the receiver FEC and SDEC decoding are used alternately, according to the procedure described below. The system is named HARQ with Erasures Correction (HARQ-EC). Theoretical analysis and computer simulation of the proposed system are performed for the case of noiseless feedback, Nakagami common fading and Additive White Gaussian Noise (AWGN) in the forward channel[1]. Moreover, we consider the forward channel as memoryless, i.e. interleaving/de-interleaving assumed to be ideal. The obtained results show that HARQ-EC provides gain in *BER*, or gain in *TE* in comparison with parameters of HARQ-II for the same outer and inner codes [2,4–5].

The paper is organized as follows. In Section 2, we describe the HARQ-EC algorithm. In Section 3 the relevant *BER* and *TE* are analyzed. The comparison of HARQ-EC and HARQ-II characteristics is given in Section 4. Section 5 presents discussion of results and some conclusions.

2. Description of the HARQ-EC System

HARQ-EC scheme uses the outer (*n*,*k*) code C_2 with

[1] The assumption of noiseless feedback does not reduce the generality of the analysis, as we are interested in the performance comparison of HARQ-EC and HARQ-II system in the same conditions.

minimal Hamming distance $d_H^{(2)}$ and the inner $(2k,k)$ half-rate invertible code [2] C_1 with minimal Hamming distance $d_H^{(1)}$. It is called invertible since with the help of inversion of k parity-check bits k information bits can be uniquely determined. We assume that each transmitted message D consists of k information bits and that each encoded message, called a code word CW, consists of n bits. When D is ready for transmission, it is encoded by the outer encoder into the transmitted n-length code word $CW_2(D,Q)$, where Q denotes the vector of $n–k$ parity-check bits. In parallel the transmitter computes k bits of the parity-check block $P(D)$ of the half rate invertible $(2k,k)$ code C_1. The block $P(D)$ is not transmitted and is stored in the buffer for later use.

Let $CW_2(Dr\ ,Qr)$ denotes the received vector if $CW_2(Dr\ ,Qr)$ was transmitted. The received data block Dr is passed to the forward channel receiver of the HARQ-EC. Its key elements are Soft Decision Maximum Likelihood Demodulator (SDMLD), FEC and Errors Erasure Correction (EEC) decoders for the outer and inner codes respectively. In SDMLD the decision $A_l^{(r)}$ about symbol Al is obtained according to the following rule [7]:

$$A_l^{(r)} = \begin{cases} A_i & \text{if } \max\limits_k \Lambda_k \\ & = \Lambda_i \ \mathit{and}\ \left|\Lambda_i\right| > thr \quad \text{where} \quad i,k = 1,2,..,M \\ \text{Erasure} & \text{if } \max\limits_k \Lambda_k \\ & = \Lambda_i \ \mathit{and}\ \left|\Lambda_i\right| \le thr \end{cases} \qquad \mathbf{(1)}$$

where Λ_i is the log-likelihood ratio calculated for the i-th symbol, M is the dimension of the used constellation, and thr is the threshold level determining the width of the erasure zone in the decision space of the SDMLD. If the number of the erased bits $t_{er}^{(2)}$ in the code word is zero, the received vector feeds the outer code of the FEC decoder and after error correction the restored message Dr is sent to the user. If $t_{er}^{(2)} > 0$, the received vector is passed to the outer code of the EEC decoder. If this combination is identified by the EECD with only one of the codebook C_2, it is considered as the transmitted codeword and the message Dr is sent to the user. Otherwise, the ReQuest signal (RQ) is sent to the transmitter via the feedback channel. Simultaneously, the message Dr (with erased elements) is saved in the buffer of the receiver. Upon receiving this request, the transmitter encodes the k-th parity bits block $P(D)$ of the inner code C_1 into the n-length codeword $CW_2(P(D),Q^{(p)})$ of the code C_2 where $Q^{(p)}$denotes the $n–k$ parity-check digits for $P(D)$.

Let $CW_2(P_r(D),Q_r^{(p)})$ denotes the received vector corresponding to $CW_2(P(D),Q^{(p)})$. The SDMLD, according to (1), erases its unreliable symbols. If the number of the erasures $t_{er}^{(2)}$ in $CW_2(P_r(D),Q_r^{(p)})$ is equal to zero, the received vector is passed to the FEC decoder of the outer code. After error correction procedure, the message D is recovered from $P_r(D)$. by inversion and is sent to the user. Otherwise, the received vector is passed to the EECD of the outer code. If vector $CW_2(P_r(D),Q_r^{(p)})$is identified by the EECD, the message D is recovered with the help of inversion of $P_r(D)$. The message Dr that is stored in the receiver memory (after recovery of D from $P_r(D)$) is then discarded. If the combination $CW_2(P_r(D),Q_r^{(p)})$ is not identified by the EECD of the outer code, the received parity block $P_r(D)$ is integrated with Dr kept in the buffer. The code word $CW_1(D_r,P_r(D))$ of the code C_1 with $t_{er}^{(1)}$ erased symbols is formed and passed to EECD of the inner code. If this combination is identified by EECD with certain vector of the code book C_1 then it is considered to be correct and the recovered message D is passed to the user. Otherwise, the request signal is generated and transmitted via the backward channel. Simultaneously, the message D is discarded from the receiver buffer, and the parity block $P_r(D)$ (with erased symbols) is saved in the receiver buffer. Upon receiving the second request for the message D, the transmitter resends the code word $CW_2(D,Q)$ and the procedure described above is repeated. The block diagram in Figure 1 illustrates transmission and retransmission procedures in the proposed HARQ-EC.

3. Analysis of the HARQ-EC Performance in Fading Channel

The main characteristics of any ARQ systems are BER and TE, defined as

$$TE = \frac{E[N]}{E[V]} \cdot \frac{k}{n}, \qquad BER = 1 - \left(1 - P_{NC}\right)^k \qquad \mathbf{(2)}$$

where V is the total number of transmitted code words and N is the number of information messages sent during the transmission interval, $E[V]$, $E[N]$ denote the expectations of V and N respectively, and P_{NC} is the probability of an undetected word error. As follows from [2], for selective-repeat ARQ scheme with noiseless feedback channel, unlimited receiver buffer and maximal number of retransmissions the values of TE and P_{NC} may be written as

$$TE \ge \frac{k}{n}\left(P_{crd}^{(l)} + P_{erd}^{(l)}\right), \qquad P_{NC} = \frac{P_{erd}^{(l)}}{P_{erd}^{(l)} + P_{crd}^{(l)}} \qquad \mathbf{(3)}$$

where $P_{crd}^{(l)}$ and $P_{erd}^{(l)}$ are probabilities of correct and erroneous reception of the data block, respectively, at the

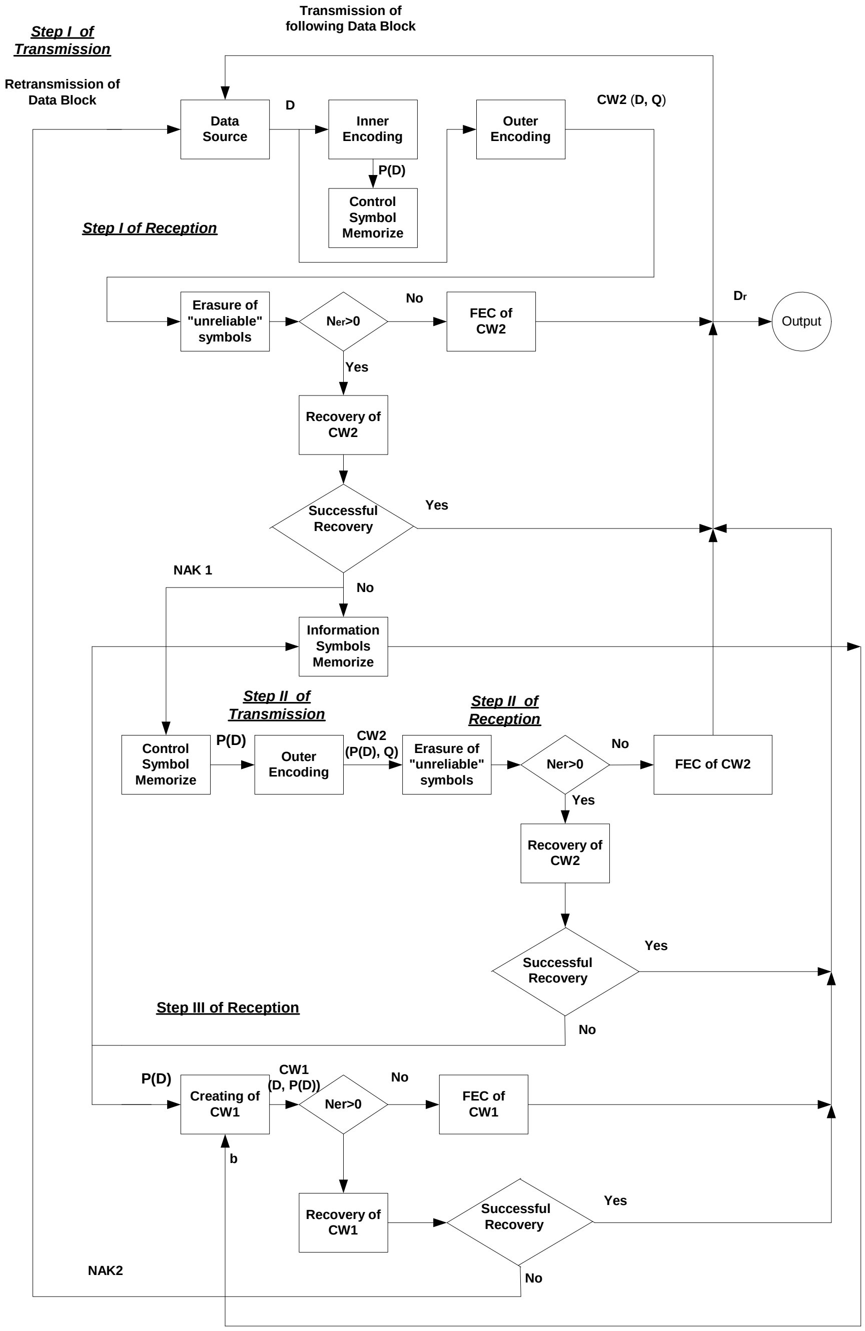


Figure 1. Transmission and retransmission procedures in HARQ-EC.

l-th[2] start of the procedure described above (see Figure 1).

We analyze performance of HARQ-EC for the case of a binary modulation. The transmitted signal within one symbol time duration T is represented by

$$s_m(t) = \mathrm{Re}\left\{A_m g(t - mT)\exp(j\omega_c t)\right\} \qquad \textbf{(4)}$$

where A_m is the information symbol, $g(t)$ is the impulse response of the transmitter filter and w_c is the carrier frequency[3]. As was mentioned earlier, we consider the case of the forward channel with additive white Gaussian noise and flat fading with Nakagami distribution [8]

[2] The index l will be omitted as the statistics of errors and erases in the received codeword do not depend on l.

[3] The kind of modulation and alphabet dimension M does not reduce the generality of the analysis, as we are interested in a comparison of the HARQ-EC performance to HARQ-II systems in the same conditions.

$$f\left(\mu_{ch}\right)=\frac{2}{\Gamma(m)}\left(\frac{m}{\mu_0^2}\right)^m \mu^{2m-1}\exp\left(-\frac{m\mu^2}{\mu_0^2}\right) \tag{5}$$

where $\Gamma(m)$ is the gamma function, $\mu_0^2=E\left\{\mu^2\right\}$, and m $\geqslant$0.5 is the fading depth parameter[4]. Moreover, it is assumed that fading is slow, which means that μ_{ch} may be considered constant, at least for one symbol interval *T*.

The signal $x_m(t)$ is demodulated by SDMLD which includes a Log-Likelihood Ratio (LLR) estimator followed by a Decision Device with Eraser (DDE) [1]. The output of LLR is

$$\Lambda_{12}=q_1-q_2,$$

where

$$q_i=\int_0^T x_m(t)s_i(t)dt,\quad i=1,2$$

for coherent SDMLD and

$$q_i=\sqrt{X_i^2+Y_i^2},\quad i=1,2$$

for noncoherent SDMLD.

Here $X_i=\int_0^T x_m(t)s_i(t)dt$, $Y_i=\int_0^T x_m(t)\hat{s}_i(t)dt$, $\hat{s}_i(t)$ is the Hilbert transform of $s_i(t)$, and *T* is the symbol duration. The vector $\mathbf{\Lambda}_{\mathbf{cw}}$=[$\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$,…$\mathbf{\Lambda}_l$…, $\mathbf{\Lambda}_n$] is passed to DDE which produces a version of the outer code word. The elements of the received vector $C_2^{(r)}$ are obtained according decision rule (1), which in our case can be written as

$$A_l^{(r)}=\begin{cases}1\text{ if }\Lambda_l>thr\\ 0\text{ if }\Lambda_l<thr\\ \text{erase if }1/thr\le\Lambda_l\le thr\end{cases} \tag{6}$$

Using (6) and results of BER analysis for binary orthogonal set of signals [1] probabilities of symbol error P_e and symbol erasure P_{ers} are written as (see Appendix A)

$$p_e=\frac{m^m\left(1+1/thr\right)\exp\left[-\left(thr-1\right)/m\right]}{\left(1+thr\right)\left[m\left(1+1/thr\right)+\gamma_0\right]^m} \tag{7}$$

$$p_{ers}=\frac{m^m\left(1+thr\right)^m}{\left(1+1/thr\right)\left[m\left(1+1/thr\right)+\gamma_0\right]^m}-\frac{m^m\left(1+1/thr\right)\exp\left[-\left(thr-1\right)/m\right]}{\left(1+thr\right)\left[m\left(1+1/thr\right)+\gamma_0\right]^m} \tag{8}$$

where $\gamma_0=E\left\{\mu^2E_s/\mu_0^2N_0\right\}$ and E_s is the energy of the transmitted signal element. With the help of (7) and (8) we estimate performance of the considered system.

Taking into account transmission and retransmission procedures in HARQ-EC, probabilities of correct and erroneous decoding of the code word can be evaluated with the help of the following expressions

$$\begin{aligned}P_{crd}&=P_{crd}^{(1)}+P_{rq}^{(1)}P_{crd}^{(2)}+P_{rq}^{(1)}P_{crd}^{(3)}\\ P_{erd}&=P_{erd}^{(1)}+P_{rq}^{(1)}P_{erd}^{(2)}+P_{rq}^{(1)}P_{erd}^{(3)}\end{aligned} \tag{9}$$

where

$$\begin{aligned}P_{crd}^{(1)}&=P_{cr}^{(FEC_2)}\left[CW_2\left(D_r,Q_r\right)\right]+P_{cr}^{(RC_2)}\left[CW_2\left(D_r,Q_r\right)\right],\\ P_{erd}^{(1)}&=P_{er}^{(FEC_2)}\left[CW_2\left(D_r,Q_r\right)\right]+P_{er}^{(RC_2)}\left[CW_2\left(D_r,Q_r\right)\right],\\ P_{rq}^{(1)}&=P\left(t_{er}^{(2)}\ge d_H^{(2)}\right)\end{aligned} \tag{10}$$

are probabilities of correct decoding, erroneous decoding and request of the code word $CW_2(D_r,Q_r)$ respectively, at the first stage of transmission,

$$\begin{aligned}P_{crd}^{(2)}&=P_{cr}^{(FEC_2)}\left[CW_2\left(P_r\left(D\right),Q_r^{(p)}\right)\right]\\ &\quad+P_{cr}^{(RC_2)}\left[CW_2\left(P_r\left(D\right),Q_r^{(p)}\right)\right]\\ P_{erd}^{(2)}&=P_{er}^{(FEC_2)}\left[CW_2\left(P_r\left(D\right),Q_r^{(p)}\right)\right]\\ &\quad+P_{er}^{(RC_2)}\left[CW_2\left(P_r\left(D\right),Q_r^{(p)}\right)\right]\end{aligned} \tag{11}$$

are probabilities of correct decoding and erroneous decoding of the codeword $CW_2(D_r,Q_r)$ respectively, at the second stage of transmission,

$$\begin{aligned}P_{crd}^{(3)}&=P_{cr}^{(FEC_1)}\left[CW_1\left(D_r,P_r\left(D\right)\right)\right]\\ &\quad+P_{cr}^{(RC_1)}\left[CW_1\left(D_r,P_r\left(D\right)\right)\right]\\ P_{erd}^{(3)}&=P_{er}^{(FEC_1)}\left[CW_1\left(D_r,P_r\left(D\right)\right)\right]\\ &\quad+P_{er}^{(RC_1)}\left[CW_1\left(D_r,P_r\left(D\right)\right)\right]\end{aligned} \tag{12}$$

are probabilities of correct and erroneous decoding, respectively, of the code word $CW_1(D_r,P_r(D))$, which is created from the block of data *Dr* extracted from the receiver memory and the parity block $P_r(D)$ of the inner code received at the second stage of transmission. Here $P_{cr}^{(FEC_j)}$, $P_{er}^{(FEC_j)}$ are probabilities of correct and erroneous reception of code words at the output of the FEC decoders, and $P_{cr}^{(RC_j)}$, $P_{er}^{(RC_j)}$ are probabilities of correct and erroneous reception of code words at the output of the errors erasure correction decoders, for the inner (*j*=1) and outer (*j*=2) codes respectively.

Bearing in mind the assumptions made above on the statistical properties of the channel, we consider that errors, as well as erasures in the received stream of code

[4] The Nakagami pdf includes, as a special case, the Rayleigh pdf for $m=1$, and can approximate both the Rice and log-normal pdf's [8].

symbols, are independent. In this case probabilities $P_{cr}^{(FEC_j)}$, $P_{er}^{(FEC_j)}$, $P_{cr}^{(RC_j)}$, $P_{er}^{(RC_j)}$ and $P_{rq}^{(j)}$ for the outer and inner codes can be determined (Appendix B) with the help of the following expressions:

$$
\begin{aligned}
&{}_{cr}^{(FEC_j)} = (1-p_{ers})^{n_j} \sum_{i=0}^{t_{cr_j}} \binom{n_j}{i} p_e^i (1-p_e)^{n_j-i},\\
&P_{er}^{(FEC_j)} = (1-p_{ers})^{n_j} \sum_{i=t_{cr_j}}^{n_j} \binom{n_j}{i} p_e^i (1-p_e)^{n_j-i},\\
&P_{rq}^{(j)} = \sum_{i=d_{H_j}}^{n_j} \binom{n_j}{i} p_{ers}^i (1-p_{ers})^{n_j-i},\\
&P_{cr}^{(RC_j)} = \sum_{t_{er_j}=1}^{d_{H_j}-1} \binom{n_j}{t_{er_j}} p_{ers}^{t_{er_j}} (1-p_{ers})^{n_j-t_{er_j}}\\
&\qquad \left\{ \sum_{t_{erd_j}=0}^{t_{er_j}} \binom{t_{er_j}}{t_{erd_j}} p_{ers}^{t_{erd_j}} (1-p_{ers})^{t_{er_j}-t_{erd_j}} P\left(cr/t_{er_j}, t_{erd_j}\right) \right\}\\
&P_{er}^{(RC_j)} = \sum_{t_{er_j}=1}^{d_{H_j}-1} \binom{n_j}{t_{er_j}} p_{ers}^{t_{er_j}} (1-p_{ers})^{n_j-t_{er_j}}\\
&\qquad \left\{ \sum_{t_{erd_j}=0}^{t_{er_j}} \binom{t_{er_j}}{t_{erd_j}} p_{ers}^{t_{erd_j}} (1-p_{ers})^{t_{er_j}-t_{erd_j}} P\left(er/t_{er_j}, t_{erd_j}\right) \right\}
\end{aligned}
\tag{13}
$$

where

$$
\begin{aligned}
&P\left(cr/t_{er_j}, t_{erd_j}\right) = \sum_{i=0}^{0.5\left(d_{H_j}-t_{erd_j}-1\right)} \binom{n_j-t_{er_j}}{i} p_e^i (1-p_e)^{n_j-t_{er_j}-i}\\
&P\left(er/t_{er_j}, t_{erd_j}\right) = \sum_{i=0.5\left(d_{H_j}-t_{erd_j}+1\right)}^{n_j-t_{er_j}} \binom{n_j-t_{er_j}}{i} p_e^i (1-p_e)^{n_j-t_{er_j}-i}
\end{aligned}
\tag{14}
$$

n_j is the length of code word, d_{H_j} is the minimum Hamming distance of the used code, t_{er_j} is a number of codeword symbols erased in the SDMLD, and t_{erd_j} is the number of the erased symbols taken from those that determine the Hamming distance of the original code $\left(0 \le t_{erd_j} \le t_{er_j}\right)$. Substituting (7), (8) into (13), (14), and the results into (9-12) and afterward into (2) and (3), we obtain values of *TE* and *BER*. For the given inner and outer codes they depend on the average SNR γ_0 and on the threshold *thr*.

4. Comparison of HARQ-EC and HARQ-II Performance

In this Section, we compare the theoretical results obtained above for HARQ-EC with those obtained by computer simulation. Then we will compare the performance of HARQ-EC to that of HARQ-II schemes [2, 4,5] for the same inner and outer codes[5]. The systematic linear block code with parameters n_2=15, k=11 and $d_{H_2} = 3$ is used as the outer code C_2, and the half invertible code with n_1=22, k=11 and $d_{H_1} = 7$ is used as the inner code C_1.

Figures 2, 3 show for HARQ-EC dependence of *BER* and *TE* (for the threshold values *thr*=1.5, *thr*=2, and *thr*=3), obtained as a result of analytical calculation and computer simulation respectively. Inspection of Figures 2 and 3 demonstrates good agreement between theory and simulation results. It follows that increase of the threshold level *thr* in HARQ-EC leads to decrease of both *BER* and *TE*.

Figure 4 shows *BER* for HARQ-EC (for the threshold values *thr*=1.5, *thr*=2, and *thr*=3), HARQ-II and FEC systems with the same code redundancy as a function of the average SNR while at Figure 5 dependence of *TE* of the compared systems from average SNR is presented.

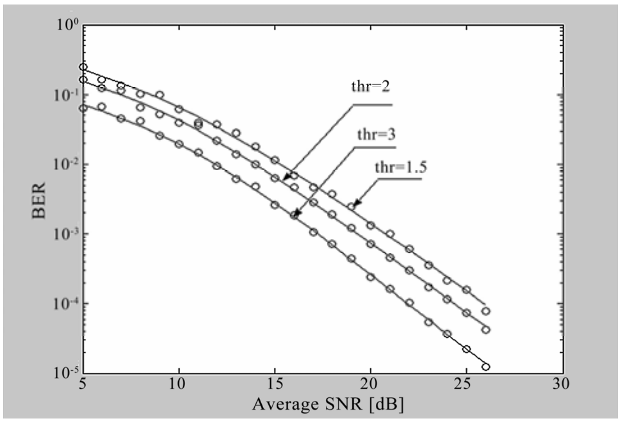


Figure 2. Average BER of HARQ-EC2 as a function of the average SNR (for the threshold values *thr*=1.5, *thr*=2, and *thr*=3); analytical calculation and computer simulation.

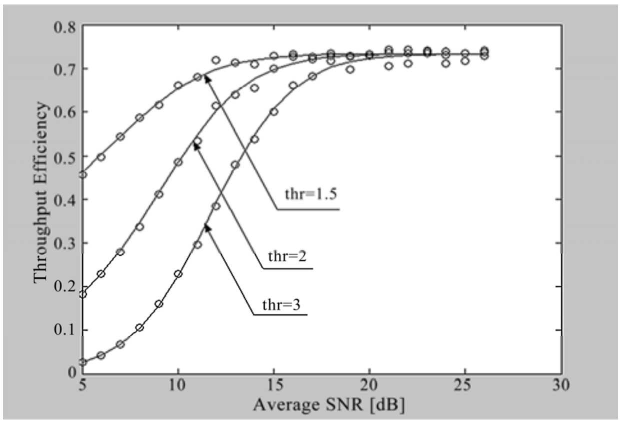


Figure 3. Throughput efficiency of HARQ-EC2 as a function of average SNR (for the threshold values *thr*=1.5, *thr*=2, and *thr*=3); analytical calculation and computer simulation.

[5] The kind of modulation and type of codes do not reduce the generality of the analysis, as we are interested in comparison of the HARQ-EC performance to HARQ-II systems in the same conditions.

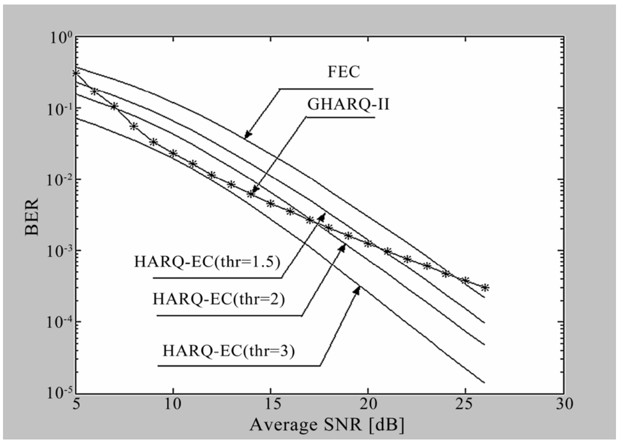


Figure 4. Average BER of HARQ-EC and HARQ-II systems as functions of average SNR.

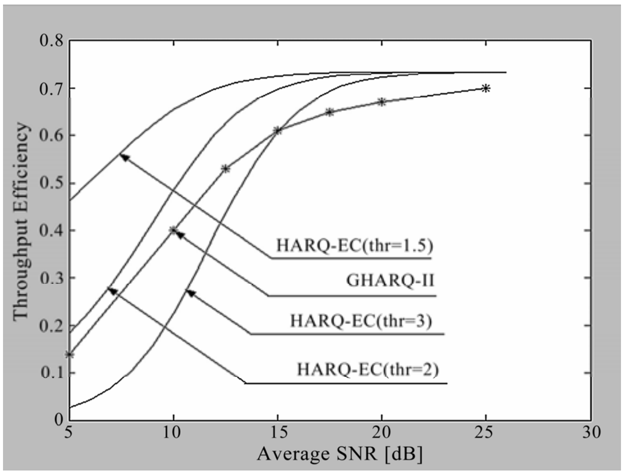


Figure 5. Throughput efficiency of HARQ-EC and HARQ–II systems as functions of average SNR.

Examination of these figures demonstrates the fact that by choice of *thr* value in SDMLD of HARQ-EC, gain in *BER* or *TE* can be achieved. For example, *TE* of HARQ-EC with *thr*=2 exceeds *TE* of HARQ-II for a roughly equal value of *BER* in a quite wide range of average SNR.

5. Conclusions

We propose the Hybrid ARQ system with erasure of unreliable symbols and retransmission of the code words (HARQ-EC). Its performance is considered for the case of flat Nakagami fading and AWGN in the forward channel. The obtained theoretical results are valid for any memoryless channel with common slow Nakagami fading, while the calculations and simulation were performed for Rayleigh fading. Good agreement between theoretical and simulation results is obtained. It has been shown that performance of HARQ-EC may be better than HARQ-II over a wide range of average SNR when the same codes are used.

6. References

[1] J. C. Proakis, "Digital communications," McGraw–Hill, New York, 1995.

[2] S. Lin and P. S. Yu, "A hybrid ARQ scheme with parity retransmission for error-control in satellite channels," IEEE Trans. Commun., Vol. COM-30, No. 7, pp. 1701–1719, 1982.

[3] Y. Wang and S. Lin, "A modified Selective-Repeat Type-II Hybrid ARQ system and its performance analysis," IEEE Trans. Commun., Vol. COM-31, No. 5, pp. 593–607, 1983.

[4] K. Q. Archer and J. A. Edwards, "Effect of channel fade rate on throughput of three GHARQ-II schemes over Rayleigh fading channel," Electronic Letters, Vol. 31, No. 16, pp. 1320–1322, 1995.

[5] C. Sunkyun and K. Shin, "A class of adaptive hybrid ARQ for wireless links," IEEE Trans. Veh. Tech., Vol. 50, No. 3, pp. 777–790, 2001.

[6] S. Kallel, R. Link, and S. Bakhtiyari, "Throughput performance of memory ARQ schemes," IEEE Trans. Veh. Tech., Vol. 48, No. 3, pp. 891–899, 1999.

[7] S. B. Wicker, "Reed–Solomon error coding for Rayleigh fading channels with feedback," IEEE Trans. Veh. Tech., Vol. 41, No. 2, pp. 124–133, 1992.

[8] L. Goldfeld, V. Lyandres, and D. Wulich, "ARQ with erasures correction in the frequency-nonselective fading channel," IEICE Trans. Commun., Vol. E80–B, No. 7, pp. 1101–1103, 1997.

[9] T. Hashimoto, "Comparison of erasure-and error threshold decoding schemes," IEICE Trans. Fundamentals, Vol. E76-A, pp. 820–827, 1993.

Appendix A

Let us assume that symbol "1" was transmitted. The probability of correct reception P_{cr} in SDMLD can be found as the probability that $\Lambda_i > thr$, and the probability of symbol erasure P_{ers} in SDMLD, in turn, can be obtained as the probability that $\frac{1}{thr} \le \Lambda_i \le thr$. From (6) we obtain.

$$\begin{aligned} p_{cr}(\mu) &= p(\Lambda_i > thr), \\ p_{ers}(\mu) &= p\left(\frac{1}{thr} \le \Lambda_i \le thr\right) \end{aligned} \qquad \textbf{(A1)}$$

where (see [1])

$$\begin{aligned} \Lambda_i &= \frac{U_1}{U_2}; \\ U_1 &= \left|2E_S \mu e^{-j\phi} + N_1\right| \\ U_2 &= \left|N_2\right| \end{aligned} \qquad \textbf{(A2)}$$

N_1 and N_2 are complex-valued Gaussian random variables with zero mean and variance $\sigma^2 = 2\mu Es$, U_1 and U_2 are mutually independent variables with distributions [1]

$$\begin{aligned} p(U_1) &= \frac{U_1}{2E_S N_0} \exp\left(-\frac{U_1^2 + 4\mu^2 E_S^2}{4E_S N_0}\right) I_0\left(\frac{\mu U_1}{N_0}\right), \\ \mathrm{p}(\mathrm{U}_2) &= \frac{U_2}{2E_S N_0} \exp\left(-\frac{U_2^2}{4E_S N_0}\right). \end{aligned} \qquad \textbf{(A3)}$$

Taking into account (A1), (A2) we obtain

$$\begin{aligned} p_{cr}(\mu) &= p(U_1 > thr \cdot U_2) \\ &= \int_0^\infty p(U_1) \cdot \left[\int_0^{U_1/thr} p(U_2)\, dU_2\right] dU_1; \\ p_{ers}(\mu) &= p\left(\frac{U_2}{thr} \le U_1 \le thr \cdot U_2\right) \\ &= \int_0^\infty p(U_1) \left[\int_{U_1/thr}^{U_1 \cdot thr} p(U_2)\, dU_2\right] dU_1 \end{aligned} \qquad \textbf{(A4)}$$

With the help of elementary algebra and tabulated integrals

$$p_{cr}(\mu) = 1 - \frac{thr}{1+thr} \exp\left(-\frac{1-thr}{\gamma(\mu) thr}\right) \exp\left(-\frac{\gamma(\mu)}{1+thr}\right) \qquad \textbf{(A5)}$$

$$\begin{aligned} p_{ers}(\mu) = {} & \frac{thr}{1+thr} \exp\left(-\frac{1-thr}{\gamma(\mu) thr}\right) \cdot \exp\left(-\frac{\gamma(\mu)}{1+thr}\right) - \\ & - \frac{1}{1+thr} \exp\left(-\frac{thr-1}{\gamma(\mu)}\right) \cdot \exp\left(-\frac{\gamma(\mu) thr}{1+thr}\right) \end{aligned} \qquad \textbf{(A6)}$$

where

$$\gamma(\mu) = \frac{\mu^2 E_S}{N_0} \qquad \textbf{(A7)}$$

Taking into account that

$$p_e = 1 - p_{cr} - p_{ers}$$

we have

$$p_e(\mu) = \frac{1}{1+thr} \exp\left(-\frac{thr-1}{\gamma(\mu)}\right) \exp\left(-\frac{\gamma(\mu) thr}{1+thr}\right) \qquad \textbf{(A8)}$$

Probabilities (A6) and (A8) are conditional probabilities of erasure and error reception in SDMLD respectively, given μ and therefore Pe and Per are

$$\begin{aligned} P_{er} &= \int_0^\infty p_{ers}(\mu) f(\mu) d\mu, \\ P_e &= \int_0^\infty p_e(\mu) f(\mu) d\mu, \end{aligned} \qquad \textbf{(A9)}$$

where $f(\mu)$ is defined by (5).

Appendix B

Let us determine probability of request P_{rq}, probabilities of correct and erroneous reception of a code word at the output of the ECE decoder ($P_{cr}^{(RC)}$ and $P_{er}^{(RC)}$), and also at the output of the FEC decoder ($P_{cr}^{(FEC)}$ and $P_{er}^{(FEC)}$) under the following conditions:

1) The code used is a linear block code (n,k) with the minimal Hamming distance d_H;

2) The encoded bit stream is represented by a codeword *CW* with length n, supplied from the SDMLD output to the FEC decoder, if *CW* does not contain the erased symbols. Otherwise, a codeword *CW* feeds the ECE decoder.

3) Errors and erasers in a sequence of code symbols *CW* are independent (the channel is memoryless). For Nakagami frequency -nonselective fading in the forward channel, the probabilities of symbol error *Pe* and symbol erasure *Pers* are defined by (7), (8).

First, we find probabilities $P_{cr}^{(FEC)}$ and $P_{er}^{(FEC)}$. Since symbol errors are independent events, the binomial law determines probabilities of correct and erroneous reception of a code word at the output of the FEC decoder. Keeping in mind that FEC decoding is used when the number of erased symbols in the received codeword $t_{er}=0$, we thus obtain

$$\begin{aligned} P_{cr}^{(FEC)} &= (1 - p_{ers})^n \sum_{i=0}^{t_{cr}} \binom{n}{i} p_e^i (1 - p_e)^{n-i} \\ P_{er}^{(FEC)} &= (1 - p_{ers})^n \sum_{i=t_{cr}+1}^{n} \binom{n}{i} p_e^i (1 - p_e)^{n-i} \end{aligned} \qquad \textbf{(B1)}$$

where t_{cr} is the number of correct errors per codeword.

Probabilities P_{rq} , $P_{cr}^{(RC)}$ and $P_{er}^{(RC)}$ may be obtained as follows. The ECE decoding is used when $t_{er}>0$, where t_{er} is a number of the erased symbols in the received code word. The erasure of t_{er} symbols from n creates a new shorter code with code word length $n^{(sh)}=n-t_{er}$ and the Hamming distance $d_H^{(sh)} = d_H - t_{erd}$, where t_{erd} is the number of erased symbols taken from those ones that determine d_H of the original code ($0 \leqslant t_{erd} \leqslant t_{er}$). Since symbol erasures are independent, the probability of the erasure of t_{er} symbols from n as well as probability of the erasure of t_{erd} symbols from d_H are determined by the binomial law. Taking this into account, we obtain P_{rq} as

$$P_{rq} = P\left(t_{er} \geq \ \ d_H \ \right) = \sum_{t_{er}=d_H}^{n} \binom{n}{t_{er}} p_{ers}^{t_{er}} \left(1 - p_{ers}\right)^{n-t_{er}} \qquad \textbf{(B2)}$$

Probabilities of correct and erroneous reception of a code word at the output of ECE decoder depend on the random variables t_{er} and t_{erd}, i.e. they have to be considered as conditional probabilities $P(cr/t_{er},t_{erd})$ and $P(er/t_{er}\,,t_{erd})$ written as

$$\begin{aligned} P\left(cr/t_{er}, t_{erd}\right) &= \sum_{i=0}^{0.5\left(d - t_{erd} - 1\right)} \binom{n-t_{er}}{i} p_e^i \left(1 - p_e\right)^{n-t_{er}-i} \\ P\left(er/t_{er}, t_{erd}\right) &= \sum_{i=0.5\left(d - t_{erd} + 1\right)}^{n-t_{er}} \binom{n-t_{er}}{i} p_e^i \left(1 - p_e\right)^{n-t_{er}-i} \end{aligned} \qquad \textbf{(B3)}$$

Twice averaging (B3) by the binomially distributed t_{er} and t_{erd}, we obtain the unconditional probabilities $P_{cr}^{(RC)}$ and $P_{er}^{(RC)}$ (see 13).

Int. J. Communications, Network and System Sciences, 2009, 7, 627-635
doi:10.4236/ijcns.2009.27070 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

Load Balanced Routing Mechanisms for Mobile Ad Hoc Networks

Amita RANI[1], Mayank DAVE[2]
[1]*Department of Computer Science & Engineering, University Institute of Engineering & Technology, Kurukshetra, India*
[2]*Department of Computer Engineering, National Institute of Technology, Kurukshetra, India*
Email: *amita26@rediffmail.com*

Received May 11, 2009; *revised July* 11, 2009; *accepted August* 24, 2009

ABSTRACT

Properties of mobile ad hoc networks (MANET) like dynamic topology and decentralized connectivity make routing a challenging task. Moreover, overloaded nodes may deplete their energy in forwarding others packets resulting in unstable network and performance degradation. In this paper we propose load-balancing schemes that distribute the traffic on the basis of three important metrics – residual battery capacity, average interface queue length and hop count along with the associated weight values. It helps to achieve load balancing and to extend the entire network lifetime. Simulation results show that the proposed load-balancing schemes significantly enhance the network performance and outperform one of the most prominent ad hoc routing protocols AODV and previously proposed load balanced ad hoc routing protocols including DLAR and LARA in terms of average delay, packet delivery fraction and jitter.

Keywords: Load Balanced Routing, Residual Battery Capacity, Hop Count, DLAR, LARA

1. Introduction

The proliferation of devices that do not depend upon centralized or organized connectivity has led to the development of mobile ad hoc networks (MANETs). These are the infrastructure-less networks where each node is mobile and independent of each other. Due to unorganized connectivity and dynamic topology, routing in MANET becomes a challenging task. Moreover, constraints like lower capacity of wireless links, error-prone wireless channels, limited battery capacity of each mobile node etc., degrade the performance of MANET routing protocols. Heavily-loaded nodes may cause congestion and large delays or even deplete their energy quickly. Therefore, routing protocols that can evenly distribute the traffic among mobile nodes and hence can improve the performance of MANETs are needed.

Routing protocols in MANETs are classified into three categories: proactive, reactive and hybrid routing protocols. Most of the prominent routing protocols like AODV [1], DSR [2] use hop count as the route selection metric. But it may not be the most efficient route when there is congestion or bottleneck in the network. It may lead to undesirable effects such as longer delays, lower packet delivery fraction and high routing overhead. Also some nodes that may lie on multiple routes spend most of their energy in forwarding of packets and deplete their energy quickly. Consequently they leave the network early. In this paper we present novel load-balancing mechanisms/schemes for MANETs that focus on distributing the traffic on the basis of combination of following three metrics:

- hop count
- residual battery capacity and
- average number of packets queued up in the interface queue of a node lying on the path from source to destination/traffic queue.

These three metrics along with associated weight values decide the path to be selected for data transmission. The results of simulations indicate that the proposed schemes outperform a prominent ad hoc routing protocol AODV and previously proposed load balanced ad hoc routing protocols including DLAR [3] and LARA [4] in terms of average delay, packet delivery fraction and jitter.

The rest of this paper is organized as follows. Section 2 discusses the work related to currently proposed load balanced ad hoc routing protocols. Section 3 details the proposed schemes in order to balance the load on various routes. Section 4 describes the methodology, performance metrics used and simulation results. Finally Section 5 concludes the paper.

2. Related Work

Load balanced routing aims to move traffic from the areas that are above the optimal load to less loaded areas, so that the entire network achieves better performance. If the traffic is not distributed evenly, then some areas in a network are under heavy load while some are lightly loaded or idle. There are various proposed algorithms for load balanced routing. In Dynamic Load Aware Routing (DLAR) protocol [3] routing load of a route has been considered as the primary route selection metric. The load of a route is defined as the summation of the load of nodes on the route, and the load of a node is defined as the number of packets buffered in the queue of the node. To utilize the most up-to-date load information when selecting routes and to minimize the overlapped routes, which cause congested bottlenecks, DLAR prohibits intermediate nodes from replying to route request messages.

Another network protocol for efficient data transmission in mobile ad hoc networks is Load Aware Routing in Ad hoc (LARA) [4] networks protocol. In LARA, during the route discovery procedure, the destination node selects the route taking into account both the number of hops and traffic cost of the route. The traffic cost of a route is defined as the sum of the traffic queues of each of the nodes and its neighbors and the hop costs on that particular route. Thus, the delay suffered by a packet at a node is dependent not only on its own interface queue but also on the density of nodes. In routing with load balancing scheme (LBAR) [5], the destination collects as much information as possible to choose the optimal route in terms of minimum nodal activity (i.e the number of active routes passing by the node). By gathering the nodes activity degrees for a given route the total route activity degree is found. Load Sensitive Routing (LSR) protocol [6] is based on DSR. In LSR the load information depends on two parameters: total path load and the standard deviation of the total path load. Since destination node doe not wait for all possible routes, the source node can quickly obtain the route information and it quickly responds to calls for connections. Correlated Load-Aware Routing (CLAR) [7] protocol is an on-demand routing protocol. In CLAR, traffic load at a node is considered as the primary route selection metric and depends on the traffic passing through this node as well as the number of sharing nodes. Alternate Path Routing (APR) protocol [8] provides load balancing by distributing traffic among a set of diverse paths. By using the set of diverse paths, it also provides route failure protection. Reference [9] gives a comparative study of some of the load balanced ad hoc routing protocol.

All The protocols discussed above concentrate on traffic balancing and do not emphasize on energy issues. A number of routing protocols that consider energy issues in MANETs have been proposed. On the basis of route selection criterion, there are mainly two categories of the energy efficient routing protocols. The first class [10–12] selects the path that consumes the least energy to transmit a single packet from source to destination, aiming at minimizing the total energy consumption along the path. The second one [13–15] intends to protect the overused nodes against breakdown, aiming at maximizing the whole network lifetime.

3. Proposed Schemes to Achieve Load Balancing

A number of routing protocols proposed for MANETs use shortest route in terms of hop count for data transmission. It may lead to quick depletion of resources of nodes falling on the shortest route. It may also result in network congestion resulting in poor performance. Therefore, instead of hop count a new routing metric is required that can consider the node's current traffic and battery status while selecting the route. The idea is to select a routing path that consists of nodes with higher residual battery power and hence longer life.

We define the required parameters, as follows: The terms used in this paper have been defined as follows:

1) Route Energy (RE): The route energy of a path is the minimum of residual energy of nodes (re_i) falling on a route. Higher the route energy, lesser is the probability of route failure due to exhausted nodes.

2) Traffic queue (tq): The traffic queue of a node is the number of packets queued up in the node's interface. Higher is its value, more occupied the node is.

3) Average Traffic Queue (ATQ): It is the mean of traffic queue of nodes from the source node to the destination node. It indicates load on a route and helps in determining the heavily loaded route.

4) Hop count (HC): The HC is the number of hops for a feasible path.

3.1. Scheme 1

The first scheme proposed in this paper tends to determine the routes in such a way that the routes consisting of nodes with lower residual battery capacity are avoided for data transmission even if they are short and less congested. This scheme tries to make a fair compromise between three route selection parameters i.e. hop count, residual battery capacity and traffic load.

A MANET can be represented as an undirected graph G(V, E) where V is the set of nodes (vertices) and E is the set of links (edges) connecting the nodes. The nodes may die because of depleted energy source and the links can be broken at any time owing to the mobility of the nodes. $\forall n|n\epsilon V$, n has an associated traffic queue tq(n) and residual battery energy re_i. A path between two nodes u and v is given as

$P(u, v) = (u, e(u, x), x, e(x, y), y, ..., e(z, v), v)$

It can be emphasized that a path between any two nodes is a set consisting of all possible paths between them. Formally, $P(u, v) = \{P_0 , P_1 , ..., P_n\}$ where each P_i is a candidate path between u and v.

Let $HC(P_i)$ be the hop count corresponding to path P_i between u and v. Weight of path P_i defined as:

$$W(P_i)= W_1 {}^* RE(P_i) - W_2 {}^* ATQ(P_i) - W_3 {}^* HC(P_i) \quad \textbf{(1)}$$

where $RE(P_i) = \min \{re_{n1}, re_{n2}, ..., re_{nm}\}$ and $n_{1,}$ n_2,..., n_m are the nodes making up the path.

$$ATQ(P_i) =(tq(n_1)+tq(n_2)+ ...+tq(n_m))/m-1 \quad \textbf{(2)}$$

The fields having adverse contribution to traffic distribution are built into negative coefficients in Equation (1). Also the weight values are calculated such that $W_1 + W_2 + W_3 = 1$.

The idea is to find a path from source to destination with maximum weight such that from the very beginning the path determined is energy efficient and there is a fair compromise between a short route and a light-loaded route. In this scheme RE has been given maximum weightage, i.e. W_1 is maximum and W_2 and W_3 are equal. We call this path Energy Aware Load-balanced Path (EALP).

Supposing that i ϵ {0,1,2,…,n}, $P(s,d) = \{P_0, P_1,…, P_n\}$ for given source s and destination d, we can define the problem mathematically as:

$EALP(s,d) = P_i$ with

$$W(P_i) = \max \{W(P_1), W(P_2),…, W(P_n)\} \quad \textbf{(3)}$$

$W_{1,}$ W_2 and W_3 are constants.

In proposed scheme routes are determined on demand. A source node initiates the route discovery process by broadcasting a route request (RREQ) packet whenever it wants to communicate with another node for which it has no routing information in its table. On receiving a RREQ packet, a node checks its routing table for a route to the destination node. If the routing table contains the latest route to the destination node, the intermediate node sends a RREP packet along the reverse path back to the source node also appending the weight value for the route. When a source node receives more than one RREP packet for a RREQ, it compares the weight values of the routes and selects the route with maximum weight. However, if an intermediate node has no information of the destination node, it adds its own traffic queue value, compares and finds the minimum of residual battery capacity field of RREQ packet with its own residual battery capacity and updates residual battery capacity field of RREQ packet, increments the hop count by one and rebroadcasts the route discovery packet. When destination node receives a route request packet, it waits for a certain amount of time before replying with a RREP packet in order to receive other RREQ packets. Then destination node computes ATQ and the weight value for each feasible path using Equation (2) and using weight function as given in Equation (1) respectively. The route with highest weight value is selected as the routing path and a RREP packet is sent back towards the source node on the selected path.

In the algorithm discussed above weight values are constant, which is its limitation as when route selection procedure starts there are more chances of network congestion because of flooding of many RREQ packets simultaneously. Moreover, nodes have maximum battery energy during initial phases. Therefore, the requirement is to change the above algorithm such that when the battery energy of nodes is high, emphasis is on selecting a short and light loaded route. As battery energy of nodes decreases we tend to conserve energy, compromising on short and lightly loaded route.

3.2. Scheme 2

Another scheme has been proposed in this paper in which weight values ($W_{1\,,}$ W_2 and W_3) are adaptive to the network status, instead of being constant. More weight age is given to find short and less congested routes during initial route discovery procedure, as the possibility of network congestion is high due to flooding of many RREQ packets simultaneously. Also, nodes have maximum battery energy during initial phases. However, as the time elapses battery energy of nodes decreases, therefore, we tend to conserve energy, compromising on short and lightly loaded routes. The adaptive behavior of the protocol has been implemented by computing the proportion of route energy and initial energy of nodes assuming that all nodes are similar with equal initial battery energy. Therefore, as per Scheme 2, weight value of a route is computed as:

$$W(P_i) = (1-\alpha) * RE(P_i) –\alpha/2*(ATQ(P_i) + HC(P_i)), \quad \textbf{(4)}$$

where,

$$\alpha =\min(RE(P_i))/ IE;0\leq \alpha\leq 1 \quad \textbf{(5)}$$

and gives the proportion of battery capacity left. Initially when nodes have high residual battery energy α is maximum, route selection is mainly done on the basis of hop count and average traffic load as can be seen from Equation (4). As nodes battery energy decreases with the passage of time α decreases and 1- α increases leading to more weightage to the route energy parameter.

3.3. Scheme 3

The scheme proposed next uses location information to limit the broadcast of RREQ packets. When an intermediate node receives a RREQ packet it uses the location information before broadcasting the RREQ packets further. Only the nodes that are closer to the destination than the source node are allowed to broadcast RREQ packets further. By doing so a broadcast storm can be avoided resulting in less congested routes. Flowchart given in Figure 3 gives the details of this algorithm.

A source node while starting a route discovery process, computes its distance w.r.t. the destination node, appends this value in the RREQ packet along with the fields as used in Scheme 2 and broadcasts it further. An intermediate node on receiving a RREQ packet, compares its distance to the destination node with the distance value stored in the RREQ packet. If its distance is longer, it drops the RREQ packet else it compares the energy value in the record of the RREQ packet with its own energy and assigns the lesser energy as the new energy value in the packet. It also adds its own traffic queue to the traffic queue already recorded in the packet and updates hop count by 1. It then broadcasts the packet further. By doing so only those nodes that are closer to the destination node than the source node participate in route selection procedure resulting in reduced routing overhead. This procedure has been explained with the help of Figures 1 and 2.

3.4. Example

As shown in Figure 1, we assume that there are three feasible paths from source node S and destination node D - Path I: (S,A,E,H,J,D), Path II: (S,B,F,K,D), Path III: (S,C,G,I,L,M).

Corresponding to Figure 1, the nodes on Path I (S,A, E,H,J,D), energies of intermediate nodes between source and destination are (450, 400, 433, 413); thus RE_1 = min(450, 400, 433, 413) = 400. Similarly, for Path II RE_2=410 and for Path III RE_3 = 420.

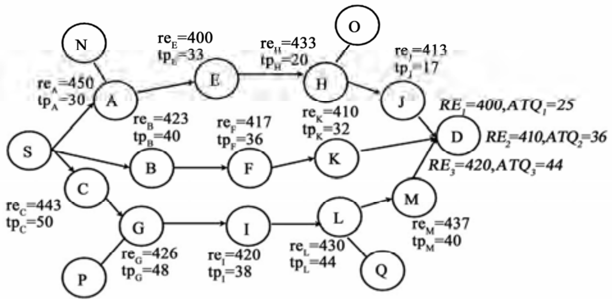


Figure 1. Route energy and average traffic queue of each feasible path for high residual battery capacity of nodes.

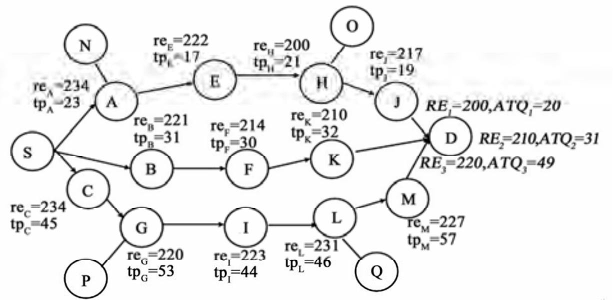


Figure 2. Route energy and average traffic queue of each feasible path for low residual battery capacity of nodes.

The traffic queue length of all the intermediate nodes between source and the destination as shown in Figure 1, for Path I (S,A,E,H,J,D), ATQ_1 = 25, HC_1 = 5. For Path II (S,B,F,K,D), ATQ_2 = 36, HC_2 = 4 and for Path III (S,C,G,I,L,M), ATQ_3 = 44, HC_3 = 6.

The destination node on receiving a RREQ packet waits for certain amount of time before replying with a RREP packet in order to receive more RREQ packets. According to first scheme the weight values are constant. After performing many simulations, we have determined that we get most favorable results for W_1=0.6, W_2=W_3= 0.2. On substituting these weight values and parameters as described above in Equation (1) we get W_3>W_2>W_1. Hence, Path III is the most suitable route and hence is selected for data transmission.

For the other two schemes we compute the value of α as per Equation (5). On substituting α in Equation (1), we get W_1>W_2>W_3 i.e. initially when the nodes have high residual battery capacity, more weightage is given to the short and lightly loaded route while route selection. However, a trade-off between hop count and ATQ is still maintained in order to avoid congested routes. As the battery energy of nodes diminishes, more emphasis is given on selecting the routes with high residual battery power. Although the routes selected may be longer. In this situation, Scheme 1 still results in W_3>W_2>W_1, however, for Schemes 2 and 3, the weight values have changed from W_1>W_2>W_3 for high residual battery capacity to W_3>W_2>W_1 for low energy nodes. This comparison has been illustrated in Table 1 and Table 2.

Table 1. Comparison of schemes for high vales of route energy.

Path	RE (IE=500)	ATQ	HC	Route weight	
				Scheme 1	Scheme 2 & 3
P1	400	25	5	W_3>W_2>W_1	W_1> W_2>W_3
P2	410	36	4		
P3	420	44	6		

Table 2. Comparison of schemes for low values of route energy.

Path	RE (IE=500)	ATQ	HC	Route weight	
				Scheme 1	Scheme 2 & 3
P1	200	20	5	W_3>W_2>W_1	W_3>W_2>W_1
P2	210	31	4		
P3	220	49	6		

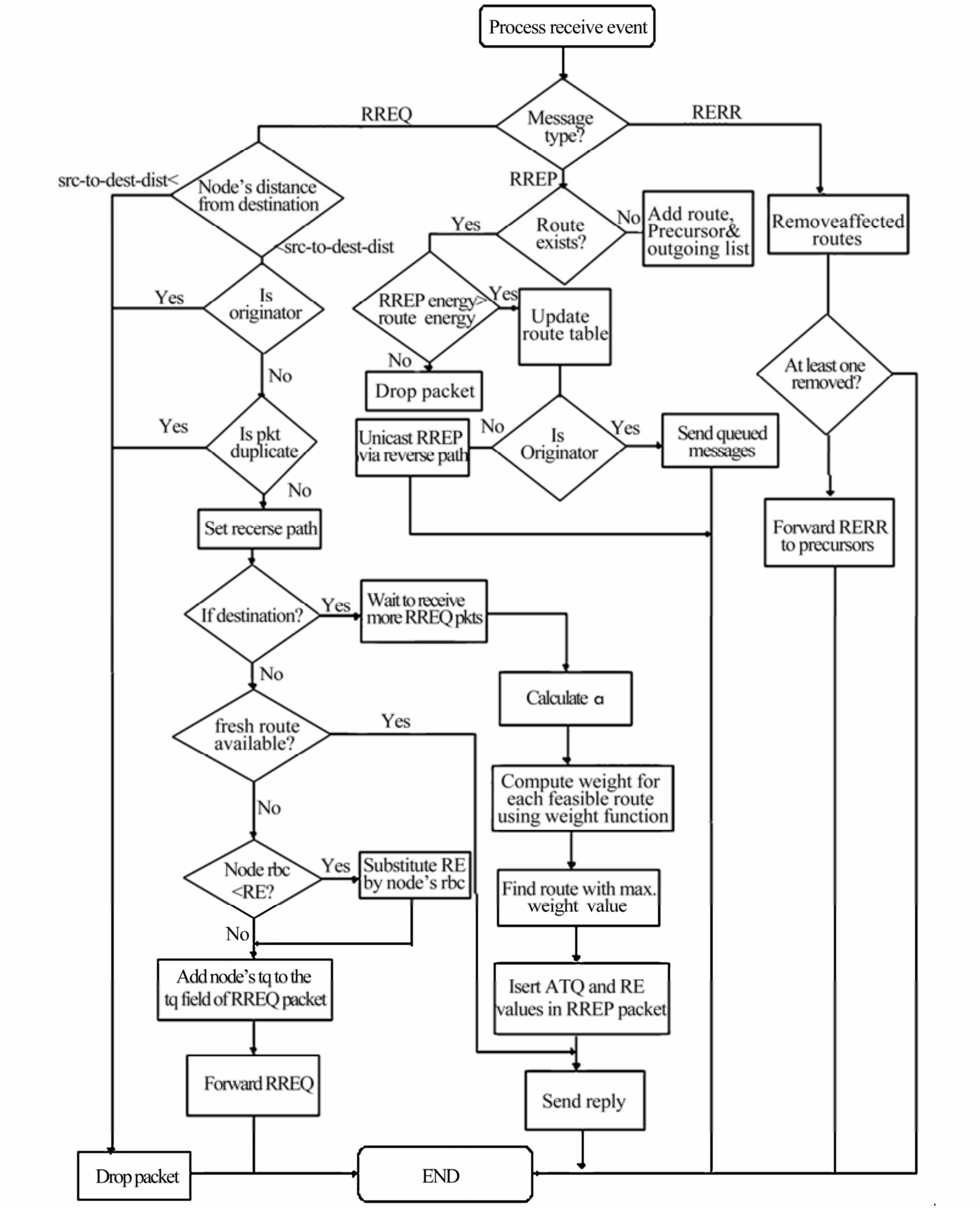

Process receive event
Message type?
RREQ
RERR
RREP
Node's distance from destination
src-to-dest-dist<
<src-to-dest-dist
Route exists?
Yes
No
Add route, Precursor& outgoing list
Removeaffected routes
Is originator
Yes
No
RREP energy> route energy
Yes
No
Update route table
Drop packet
At least one removed?
Is pkt duplicate
Yes
No
Unicast RREP via reverse path
Is Originator
No
Yes
Send queued messages
Set recerse path
Forward RERR to precursors
If destination?
Yes
No
Wait to receive more RREQ pkts
Calculate α
fresh route available?
Yes
No
Compute weight for each feasible route using weight function
Node rbc <RE?
Yes
No
Substitute RE by node's rbc
Find route with max. weight value
Add node's tq to the tq field of RREQ packet
Isert ATQ and RE values in RREP packet
Forward RREQ
Send reply
Drop packet
END

Figure 3. Flowchart depicting proposed algorithm.

4. Performance Evaluation

In this section we describe our simulation environment and performance metrics.

4.1. Performance Metrics

We have used ns-2 simulator version 2.29 to analyze the proposed algorithms. Our solution has been compared against AODV and two of the previously proposed load balanced ad hoc routing protocols - DLAR and LARA. We use the following performance metric to evaluate the performance of each scheduling algorithm:

- Packet Delivery Fraction: It gives the ratio of the data packets delivered to the destination to those generated by the sources, which reflects the degree of reliability of the routing protocol.
- Normalized Routing Load: The number of routing control packets per data packet delivered at the destination.
- Average End-to-End Delay: This is the average overall delay for a packet to traverse from a

source node to a destination node. This includes the route discovery time, the queuing delay at a node, the transmission delay at the MAC layer, and the propagation and transfer time in the wireless channel. As delay primarily depends on optimality of path chosen, therefore, this is a good metric for comparing the efficiency of underlying routing algorithms.

- Jitter: Jitter is defined as the delay variation between each received data packets. It gives an idea about stability of the routing protocol.
- Average Residual Battery Capacity: This metric depicts the amount of energy consumption of nodes with respect to time period.

4.2. Simulation Environment

Our simulation scenario consists of 50 nodes moving at maximum velocity of 20m/s in a 600m x 600m grid area with a transmission range of 100m with 25 and 37 TCP flows. Each source node transmits packets at a rate of four packets per second, with a packet size of 1024 bytes. We run simulation for pause times of 0, 100, 200, 300, 400, 500, 600, 700 and 900 seconds. The mobility of a node is defined by random waypoint model. This model forces nodes to move around with two predefined parameters, maximum velocity and pause time. Each node moves to a random destination at random velocity. They stay there for predefined time and then move to a new destination. Also it is the most widely used mobility model in previous studies. The size of the interface buffer of each node for simulation is taken as 50 packets. Each experiment is conducted four times and the average result has been considered.

4.3. Simulation Results

4.3.1. Packet Delivery Fraction

Figure 4 and Figure 5 show the packet delivery fraction of each protocol for 50 nodes with 25 and 37 sources respectively. The proposed schemes perform very well irrespective of the node's pause time and outperform AODV, DLAR and LARA. In high mobility scenarios, many route construction processes are invoked. When a source floods a RREQ packet to recover the broken route, many intermediate routes reply with the routes cached by overhearing packets during the initial route construction phase. A number of these cached routes overlap existing routes. Nodes that are part of multiple routes become congested and can not deliver the packets further resulting in poor performance of AODV. Although DLAR and LBAR also achieve a better performance than AODV, the effectiveness of load balancing is not salient compared with our schemes. The performance of proposed schemes is almost similar. However, the reason for lower packet delivery fraction at some points for third scheme is inability of the network to find out a route to the destination because of restricted number of RREQ packets. The results also show that the packet delivery fraction reduces with increase in load in the network.

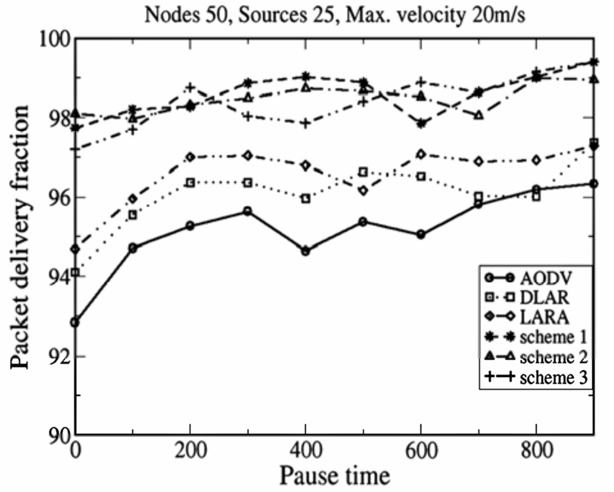


Figure 4. Packet delivery fraction vs. pause time for 25 sources.

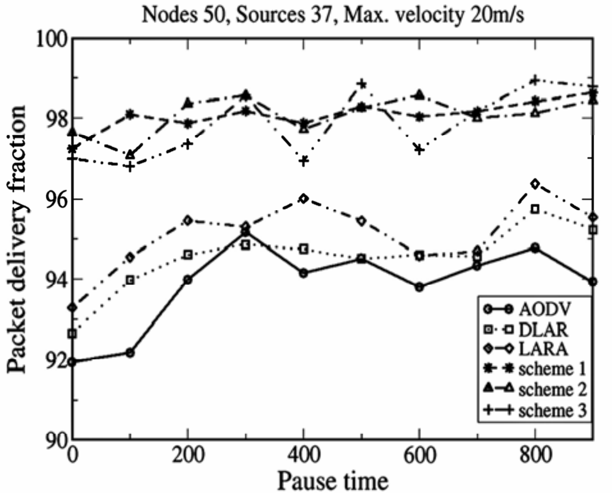


Figure 5. Packet delivery fraction vs. pause time for 37 sources.

4.3.2. Normalized Routing Load

Figure 6 and Figure 7 show normalized routing load of each protocol for 50 nodes with 25 and 37 sources respectively. Horizontal axis of the figures represent the pause times. As expected, normalized routing load for first two proposed schemes is comparatively higher than AODV protocol. However, in the third proposed algorithm

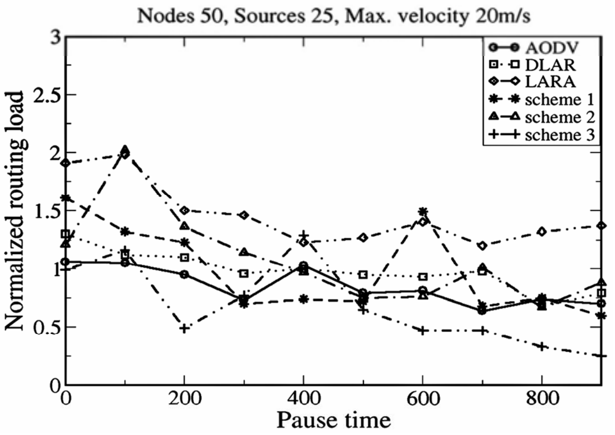


Figure 6. Normalized routing load vs. pause time for 25 sources

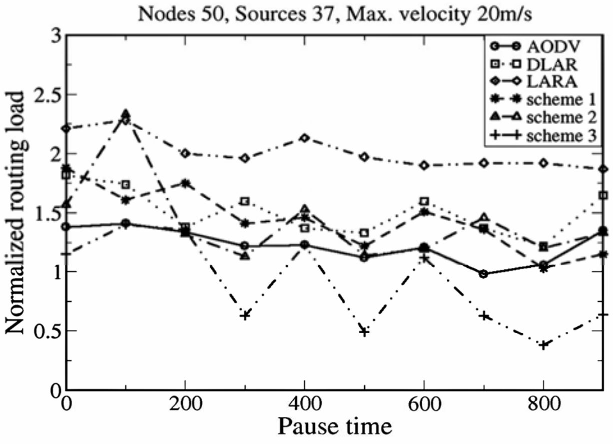


Figure 7. Normalized routing load vs. pause time for 37 sources.

we try to restrict the broadcast of RREQ packets, which results in lower routing load than the routing load of AODV, DLAR and LARA protocols. It has also been observed from Figure 6 and Figure 7 that normalized routing load increases with increase in number of sources in the network.

4.3.3. Average End-to-End Delay

Figure 8 and Figure 9 plot the average end-to-end delay for variations of node's pause time for 50 nodes with 25 and 37 sources respectively. Proposed algorithms have much improved average end-to-end delay than AODV and other two load balanced routing protocols i.e. DLAR and LARA. We can see that the end-to-end delay increases for all the protocols with increase in load as can be seen in Figures 8 and 9. The reason is the increased contention at MAC level due to increase in load. The packets now have to wait longer in the interface queue before being transmitted. Here, AODV suffers maximum delay as it often routes the packets around heavily loaded nodes. DLAR and LARA make better choice of routes than AODV. The proposed algorithms make best decision among all these protocols. The results are more noteworthy because even for highly dynamic topology (i.e. pause time = 0) and static topology (i.e. pause time = 900), proposed algorithms achieve significantly lower delay than rest three protocols. This is due to the effective routing strategy adopted for load balancing and their try to route packets along a less congested route to avoid overloading of some nodes.

4.3.4. Jitter

Figure 10 and Figure 11 show delay variation of received packets (jitter) versus pause time for 50 nodes with 25 and 37 sources respectively. It can be seen that jitter is considerably lower for proposed algorithms than AODV DLAR and LARA protocols, even for highly dynamic

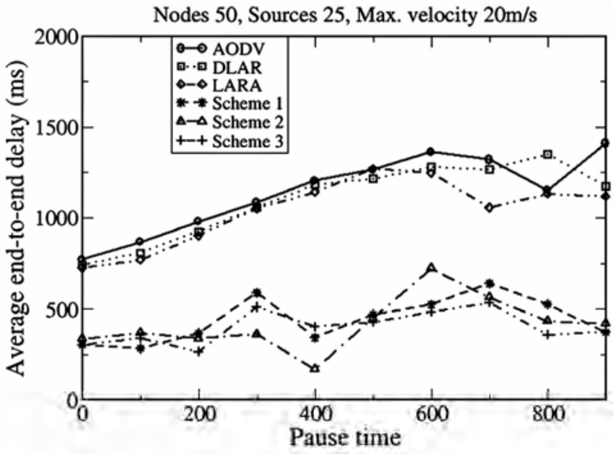


Figure 8. Average end-to-end delay vs. pause time for 25 sources.

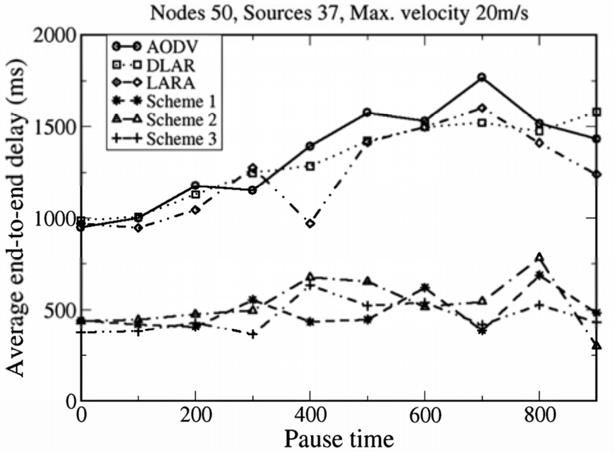


Figure 9. Average end-to-end delay vs. pause time for 37 sources.

topology (i.e. pause time = 0) and nearly static topology (i.e. pause time = 900) as well. This behavior is as anticipated because delays mainly occur in queuing and medium access control processing. These delays are reduced in proposed schemes by routing the packets towards nodes that are less occupied also taking into account more efficient nodes in terms of energy.

4.3.5. Average Residual Battery Capacity

Figure 12 compares the average residual battery capacity of nodes for AODV and the proposed schemes w.r.t. simulation time. It is evident from the figure that the rate of energy consumption is much higher for AODV than the proposed protocols. The reason is the energy aware load balancing behavior of proposed schemes. Initially when battery energy of nodes is high, energy consumption rate for the first proposed scheme is the least. This is due its behavior of energy considerations while balancing the load, even if the node energy is high. The performance of other two protocols improves with the reduction in battery energy, because as the battery capacity of nodes decreases, routes with higher residual battery capacity are considered irrespective of its length and load. As can be inferred from the Figure 12, a MANET employing third proposed strategy for routing has maximum residual battery capacity. It is due to restricting the broadcast of packets. As a result of which a proportion of energy spent by nodes in forwarding RREQ packets remains conserved.

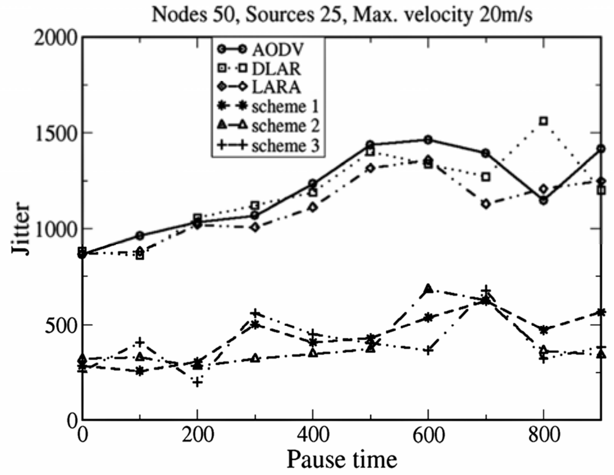


Figure 10. Jitter vs. pause time for 25 sources.

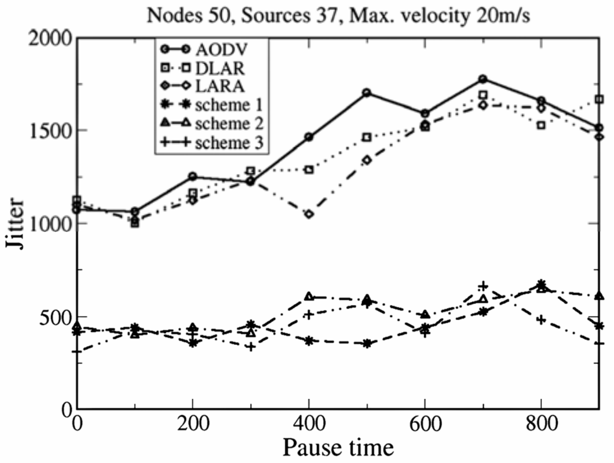


Figure 11. Jitter vs. pause time for 37 sources.

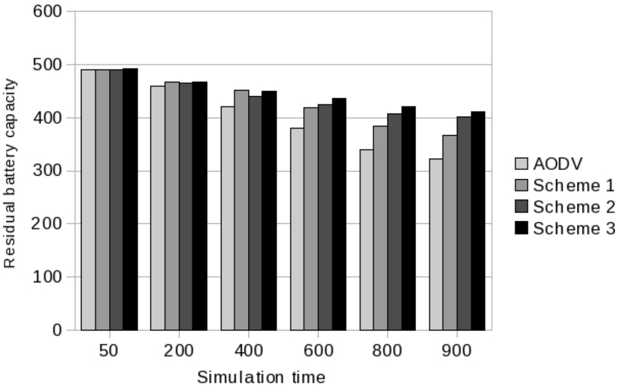


Figure 12. Average residual battery capacity of nodes.

5. Conclusions

In this paper, we presented some schemes for load balancing in mobile ad hoc networks. The proposed schemes are based on a new metric based on weighted combination of three parameters. The three parameters responsible for final route selection are - the average traffic queue, the route energy, and the hop count. And, the weights corresponding to these parameters may be fixed or adaptive to the network status, depending upon the load balancing scheme. By taking these three parameters together the traffic is deviated from high loaded routes towards routes possessing higher energy and less loaded. In proposed strategies a load balanced routing path is selected among all feasible paths on the basis of weight value calculated for each path. In a feasible path, the higher the weight value, the higher is its suitability for traffic distribution. The performance of the schemes is evaluated by simulation. The result of simulation indicates that, compared with previous load balanced routing schemes DLAR and LBAR, the proposed schemes exhibit a better performance in both moderately loaded and highly loaded situations. In addition, we have shown that the average residual battery capacity of nodes and hence network lifetime is higher in case of proposed schemes than AODV protocol.

6. References

[1] C. E. Perkins and E. M. Royer, and S. R. Das, “Ad hoc

on-demand distance vector routing," Internet Draft, draft-ietf-manet-aodv-05.txt, March 2000.

[2] D. B. Johnson and D. A. Maltz, "The dynamic source routing protocol for mobile ad hoc networks," IETF Draft, 1999.

[3] S. J. Lee and M. Gerla, "Dynamic load aware routing in ad hoc networks," Proc. ICC, Helinski, Finland, pp. 3206–3210, June 2001.

[4] V. Saigal, A. K. Nayak, S. K. Pradhan, and R. Mall, "Load balanced routing in mobile ad hoc networks," Elsevier Computer Communications, Vol. 27 pp. 295–305, 2004,.

[5] H. Hassanein and A. Zhou, "Routing with load balancing in wireless ad hoc networks," Proc. ACM MSWiM, Rome, Italy, pp. 89–96, July 2001.

[6] K. Wu and J. Harms, "Load sensitive routing for mobile ad hoc networks," Proc. IEEE ICCCN, Phoenix, AZ, pp. 540–546, Oct. 2001.

[7] J.-W. Jung, D. I. Choi, K. Kwon, I. Chong, K. Lim, and H.-K. Kahng, "A correlated load aware routing protocol in mobile ad hoc networks," ECUMN, LNCS 3262, pp. 227–236, 2004.

[8] M. R. Pearlman, Z. J. Hass, P. Sholander, and S. S. Tabrizi, "On the impact of alternate path routing for load balancing in mobile ad hoc networks," Proc. of First Annual Workshop on Mobile and Ad Hoc Networking and Computing, Mobihoc, Boston, MA, USA, pp. 3–10, August 2000.

[9] A. Rani and M. Dave, "Performance evaluation of modified AODV for LOAD balancing," Journal of Computer Science, Vol. 3, pp. 863–868, 2007.

[10] S. Singh, M. Woo, and C. Raghavendra, "Power-aware routing in mobile ad-hoc networks," Proceedings of the 4th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom), Dallas, TX, USA. New York, NY, pp. 181–190, Oct 25–30, 1998.

[11] A. Srinivas and E. Modiano, "Minimum energy disjoint path routing in wireless ad-hoc networks," Proceedings of the 9th Annual International Conference on Mobile Computing and Networking (MobiCom), San Diego, CA, USA. New York, NY, pp. 122–133, Sep 14–19, 2003.

[12] M. Subbarao, "Dynamic power-conscious routing for manets: An initial approach," Proceedings of the 50th IEEE Vehicular Technology Conference, VTC, Vol. 2, pp. 1232–1237, Sep 19–22, 1999.

[13] C. K. Toh, "Maximum battery life routing to support ubiquitous mobile computing in wireless ad-hoc networks," IEEE Communications Magazine, Vol. 39, pp. 138–147, 2001.

[14] N. Gupta and S. R. Das, "Energy-aware on-demand routing for mobile ad-hoc networks," Proceedings of the 4th International Workshop on Distributed Computing, IWDC, Capri, Italy, pp. 164–173, Sep 8–11, 2002.

[15] L. Y. Li, C. L. Li, and P. Y. Yuan, "An energy level based routing protocol in Ad-hoc networks," Proceedings of the IEEE/WIC/ACM International Conference of Intelligent Agent Technology (IAT'06), Hong Kong, China. Los Alamitos, CA, pp. 306–313, Dec 18–22, 2006.

Int. J. Communications, Network and System Sciences, **2009, 7, 636-640**
doi:10.4236/ijcns.2009.27071 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

Next Generation Optical Access Network Using CWDM Technology

Saba Al-RUBAYE[1], Anwer AL-DULAIMI[2], Hamed Al-RAWESHIDY[2]
Wireless Networks & Communications Centre, School of Engineering & Design,
Brunel University, London, UK
Email: [1]*sabaday17@yahoo.com*; [2]*anwer.al-dulaimi@brunel.ac.uk*

Received July 20, 2009; *revised August* 29, 2009; *accepted September* 23, 2009

ABSTRACT

We are developing a novel technology for the next generation optical access network. The proposed architecture provides FTTX high bandwidth which enables to give out 10Gbit/s per end-user. Increasing the subscribers in the future will cause massive congestion in the data transferred along the optical network. Our solution is using the wavelength division multiplexing PON (CWDM-PON) technology to achieve high bandwidth and enormous data transmission at the network access. Physical layer modifications are used in our model to provide satisfactory solution for the bandwidth needs. Thus high data rates can be achieved throughout the network using low cost technologies. Framework estimations are evaluated to prove the intended model success and reliability. Our argument that: this modification will submit a wide bandwidth suitable for the future Internet.

Keywords: Passive Optical Network (WDM-PON), Fibre-to-the-Home (FTTH), Optical Access Network, Next Generation

1. Introduction

Optical access network has attractive much attention [1], this is because of the low loss and enormous bandwidth of optical fibre, the increasing demand for capacity, coverage, and the benefits it offers in terms of low cost optical system ,all of which make it an ideal candidate for future access network.

The network and service providers are seeking to reduce their operational costs. The concept of using a passive optical network (PON) is an attractive option. In a PON there are no active components between the central office and customer’s premises, which can eliminate the need to power and manage active components in the cable system of the access network, and usually the PON has a tree topology in order to maximize their coverage with minimum network splits, thus reducing optical power loss [2].

Each PON terminates on an Optical Line Termination (OLT) in the head-end, or hub facility. The OLT connects through a Wave Division Multiplexing (WDM) coupler with a single fibre strand to the optical distribution network (ODN), and broadcasts an optical signal at 1490 nm that reaches each subscriber connected to that fibre through passive optical splitters. The OLT also receives signals at 1310 nm from each customer optical network user (ONU). OLTs are housed in a shelf that typically supports multiple OLTs, common control cards, and interfaces to voice and data services equipment [3]. Basically, fibre can deliver the information such as data, voice, and video from central office CO to the end of subscribers Figure 1. According to Heron [4], both FSAN-ITU and the IEEE have initiated projects to standardize a next generation of PON with 10Gbps bandwidth. Numerous options are under consideration. In anticipation of some of the potential options, FSAN-ITU is proposing a wavelength blocking filter for Gigabit PON (GPON ONT)s in order to allow for the potential coexistence of GPON ONTs with other wavelengths on the PON.

Dinan [5] argued that there are two alternatives for WDM metro networks dense WDM (DWDM) and CWDM. In high capacity environments, DWDM is used. In DWDM, the channel separation can be as small as 0.8 or 0.4 nm, for up to 80 optical channels at line rates up to 10Gbps. DWDM technologies is very expensive, so its application to access networks is difficult. Instead, CWDM is emerging as a robust and economical solution. The advantage of CWDM technology lies in its low-cost optical components. CWDM offers solutions for 850,

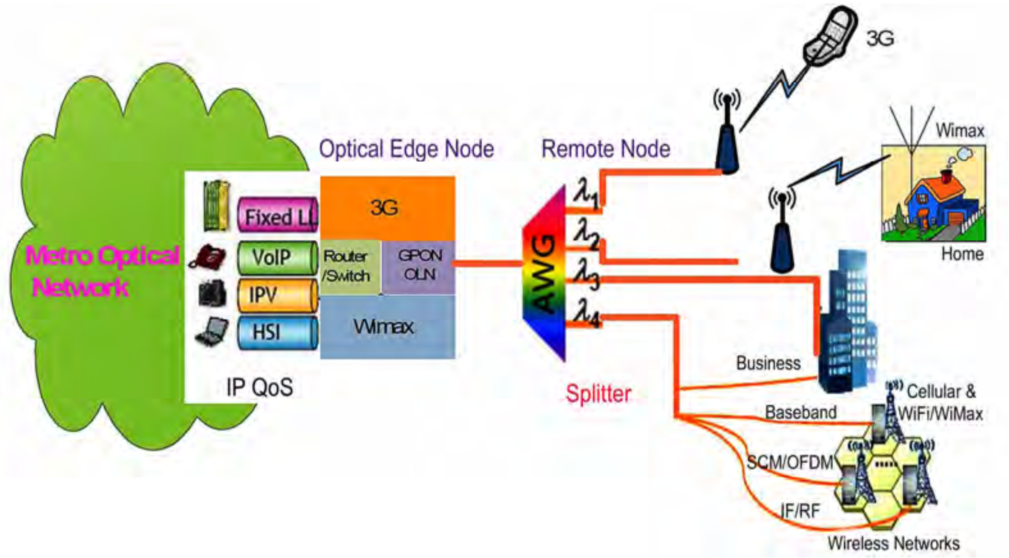


Figure1. Optical access network architecture.

1,300, and 1,500 nm applications at 10 and 40Gbps on up to 15 optical channels spaced 20 nm a part. Furthermore, the wavelength multiplexer with low channel crosstalk can be implemented easily for CWDM. It has been argued that the total system cost is 40% cheaper for the CWDM-PON [6].

2. WDM-Based Optical Access

2.1. Requirements of Next-Generation FTTH Architectures

A set of performance objectives was established by the members of Full Service Access Network (FSAN) for next-generation PONs that increase bandwidth and cost-effectively while safeguarding previous investments. These performance requirements can be summarized as follows [5]:

- Increase bandwidth by 4x.
- Respect similar optical distribution network parameters.
- Respect wavelength allocations of GPON.
- Keep changes of the media access control (MAC) layer to a minimum.
- Enable coexistence with GPON.

In addition, the IEEE has set like targets, but its focus is centred on the 10G time division multiplexing (TDM) Ethernet (EPON) solution. Any coexistence issues with EPON are addressed by using a different wavelength for the 10G EPON.

Hybrid four-wavelength PON is an approach that places four logical PONs on a single fibre using four discrete wavelengths. At 10 G bps, this increases the downstream bandwidth of GPON four fold. The existing downstream waveband of 1480-1500nm could be easily sub-divided into four bands permitting the cost effective use of four inexpensive medium-density lasers.

In the case of GPON and FSAN, the use of hybrid splitters is proposed. This would allow only one of the four GPON signals to pass to any particular ONT. Opportunely, the overall loss of a hybrid splitter would stay put similar or improved to that of a power splitter. A special dual-use splitter is being proposed that could be used firstly as a power splitter and later as a hybrid splitter, thus avoiding its replacement cost. In the upstream route, ONTs would contribute to the existing 1310nm wavelength and the upstream bandwidth would remain unaffected, resulting in an 8:1 ratio between downstream and upstream bandwidth.

2.2. Wavelength Options

Coarse (WDM) one of the next generation solution, in addition, require of opportunity networks to increase bandwidth with low cost available in WDM. Wavelength spacing of extra than 20 nm is generally called coarse WDM (CWDM). Optical interfaces, which have been standardized for CWDM, can be found in ITU G.695, at the same time as the spectral grid for CWDM is defined in ITU G.694.2. If the inclusive wavelength range of 1271 nm to 1611 nm, as defined in ITU G.694.2, is used with 20 nm spacing, then a total of 18 CWDM channels are accessible, as can be seen in Figure 2. [7,8].

3. Next Generation WDM-PON Networks Model Architecture

In our model, we assume a four channel C WDM (1490-1550nm), 2.5 GB/s directly modulated light wave system over a passive optical network. The source is DFB-LD

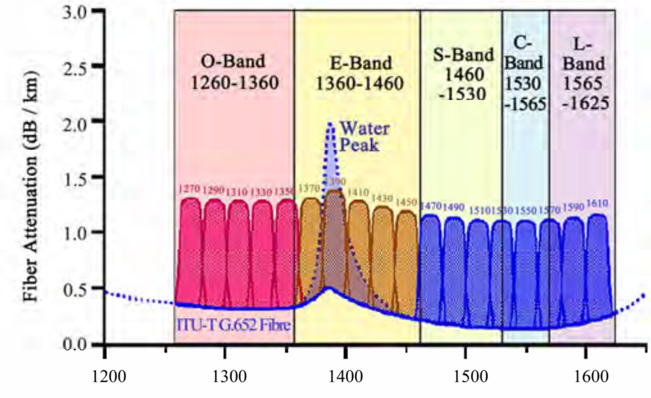


Figure 2. Metro CWDM wavelength grid as specified by ITU-T G.694.2.

modules with (1270 nm ~ 1610 nm) wavelength, and the bandwidth is 2.5Gbps.

The passive optical network utilizes one 2:2 splitter and four 1:16 splitters. Atypical topology for a CWDM metro network is shown in Figure 3. Metro network is linked via Central Office (CO) by PON. On the other hand, the CO consists of transmitter and receiver each of them has four lasers. These lasers have different wavelengths: 1490nm (λ1), 1510nm (λ2), 1530nm (λ3), 1550 nm (λ4) respectively for the upstream transmission. The receivers are consisted of four wavelengths: 1290(λ1) nm, 1310(λ2) nm, 1330(λ3) nm, 1350(λ4) nm respectively as the downstream transmission.

In this architecture, a single-mode optical fibre (SMF) connects the CO and the subscribers' site. The suggested distance for our estimations is 60 km. Four channels each of 2.5Gbps are multiplexed using OLT to achieve the suggested 10Gbit/s bandwidth. In addition long haul reach and narrow channel spacing are to be verified using the new arrangements.

4. Model Evaluation

4.1. BER versus SNR

The bit error rate (BER) is defined as the probability that a bit is inaccurately detected by the receiver, i.e., that a transmitted (0) is detected as a (1), or a (1) is detected as a (0). A theoretical bit error rate, as a function of signal to noise ratio (S/N), is known as a result of formula [9,10]:

$$\mathrm{BER}=\frac{1}{2}\left[1-\mathrm{erf}\left(\sqrt{\frac{I_s^2}{I_n^2}}\right)\right] \qquad (1)$$

In our CWDM system, we suppose the episode optical power on photodiode detector is P_r W, and the responsivity of the detector is l A/W, the signal current in photodiode is could be written:

$$(I_s)=P_r\ ,R_\lambda \qquad (2)$$

The noise originating in the detector is thermal noise current and generated within the photo detectors load resistor R_L, the thermal noise current is given by:

$$(I_{th}^2)=\frac{4KT\Delta f}{R_L} \qquad (3)$$

where k is the Boltzman constant (1.3805*10-23 J/K), T is the temperature is 300K Δf signal bandwidth is 2.5Gbps, R_L is load resistor in Ω.

Therefore the total current noise is:

$$I_n=I_{th}+I_d \qquad (4)$$

The signal to noise ratio(S/N) is given by:

$$\left[\frac{S}{N}\right]_{dB}=20\,\mathrm{Log}\left[\frac{i_s}{i_n}\right] \qquad (5)$$

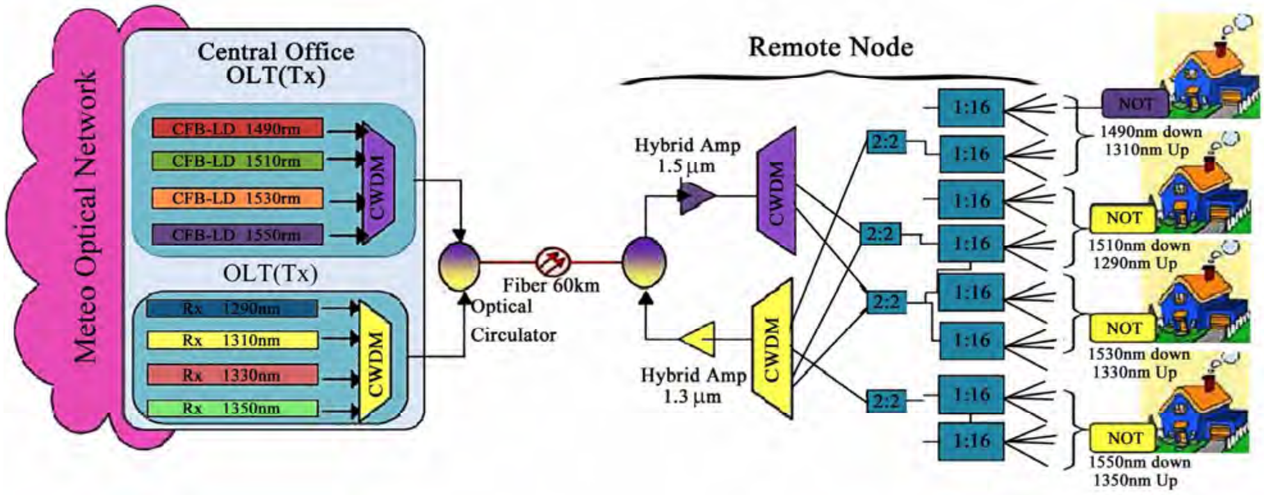


Figure 3. Model of next generation CWDM network architecture.

The signal to noise ratio of this model is about 41.4dB, as shown in Figure 4, then BER for the proposed system is about 5*10^{-8}.

Noticeably, the bit error rate (BER) decreases, as the signal to noise ratio (S/N) increases. Hence, data can be transmitted with high superiority as the expectation of error decreases.

4.2. Coverage of WDM-PON

For different splitter ratio the insert loss is different, with the insert loss of splitter is given by [2]:

$$L_{\text{splitter}} = 10\ \text{Log}_{10} \frac{1}{\text{splitter\_ratio}} \qquad \textbf{(6)}$$

The maximum coverage distance of the N remote node is given by:

$$D_{\text{N-Max}} = \left[\frac{\{(P_{\text{Tx}} - P_{\text{Rx-Min})} - [N*(2*L_{\text{TFF}} + L_{\text{TFF-other}}) + L_{\text{TFF}}] - +\text{Lspliter} - +\text{Lothers}\}}{\text{Fattenuation}} \right] \textbf{(7)}$$

where l is the average transmit power, P_{RX}-1 is the minimum receive optical power with error free in ONU/ONT, or it can be describe as the receive sensitivity of ONU/ONT, L_T is the insert loss of TFF, L_{TFF-ot} is some other attenuation connected to TFF, L_{oth} is other attenuations, such as the interface loss, $F_{attenuat}$ in different CWDM wavelengths have different attenuation.

As we show in the Figure 5. The relation between distance and nodes with different number of splitter, with increase in the splitter value the PON coverage it become less.

5. Conclusions

An innovative solution is presented to increase the bandwidth for optical access networks. The projected broadband access network is the key solution for point-to-multipoint optical communications. High data rates are achieved using low price infrastructures. In this paper CWDM is approved to be an extremely exceptional adjustment for the main optical core to edges providers. Reliable cabling can be achieved instantaneously using current modified technologies. Our evaluations show highly performance for the suggested model. The data bit rate achieved was 10Gbit/s resulting from four attached optical fibre wavelengths. The presented scenario used the passive optical network technology as the casting media between the central office and the end-users. Ultimately, further research may be done to the using of CWDM-PON in metro and long-haul fibre routes.

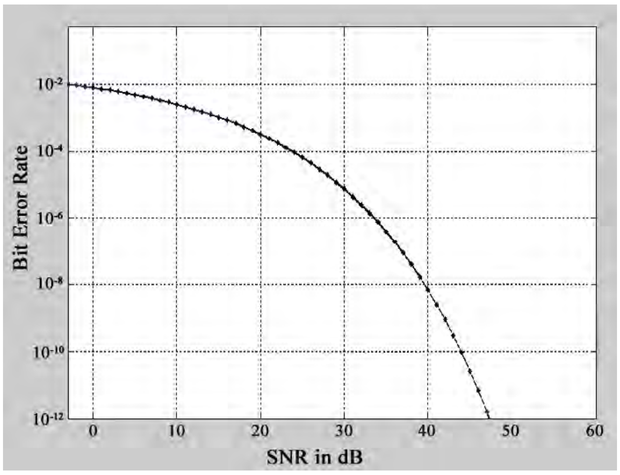


Figure 4. BER as function of SNR of CWDM.

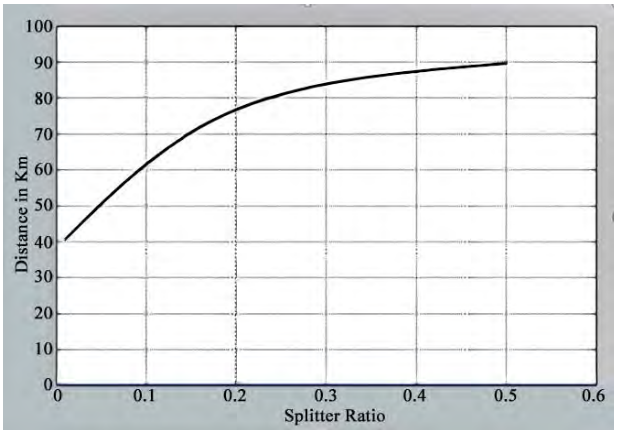


Figure 5. Illustrate the coverage distance of remote node with different splitter.

6. References

[1] H. Al-Raweshidy and S. Komaki, "Radio over fiber technologies for mobile communication networks," Artech House Publisher, ISBN 1–58053–148–2, 2002.

[2] Z. Peng, "The hybrid CWDM/TDM-PON architecture base on pointto-multipoint wavelength multiplex/demultiplexer," Proceedings of the SPIE, Vol. 6784, pp. 678420, 2007.

[3] A. Banerjee, Y. L. Park, F. Clarke, H. Song, S. Yang, G. Kramer, K. Kim, and B. Mukherjee, "Wavelength-division-multiplexed passive optical network (WDM-PON) technologies for broadband access," Journal of Optical Networking, Vol. 4, No. 11, pp. 737–758, 2005.

[4] Fiber Optics for Government and Public Broadband, "A feasibility study prepared for the city and county of San Francisco," Jan. 2007.

[5] R. Heron, T. Pfeiffer, D. van Veen, J. Smith, and S. Patel, "Technology innovations and architecture solutions for the next-generation optical access network," Bell Labs Technical Journal, Vol. 13, No. 1, pp. 163–182, 2008.

[6] E. Dinan, "Survivable FTTP network architectures with metro WDM and access PON," Bechtel Telecommunications Technical Journal, Vol. 2, No. 2, pp. 53–60, 2004. www.adc.com.

[7] I. Adam, M. Ibrahim, N. Kassim, A. Mohammad, and A. Supa, "Design of arrayed waveguide grating (AWG) for DWDM/CWDM applications based on BCB polymer," Elektrika Journal, Vol. 10, No. 2, pp. 18–21, 2008.

[8] S. A. AL-Rubaye, A. A. AL-Dulaimi, and H. J. Gubashi, "Calculations of SNR for free space optical communication systems," DJAR Journal, Vol. 3, No. 2, pp. 76–83, 2007.

[9] D. G. Baker, "Fibre optic design and applications," Prentic-Hall, 1985.

Int. J. Communications, Network and System Sciences, 2009, 7, 641-644
doi:10.4236/ijcns.2009.27072 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

An Identifier-Based Network Access Control Mechanism Based on Locator/Identifier Split

Rui TU[1], Jinshu SU[1], Ruoshan KONG[2]
[1]*School of Computer Science, National University of Defense Technology, Changsha, China.*
[2]*International School of Software, Wuhan University, Wuhan, China*
Email: *ruitu@nudt.edu.cn, krs*1024*@*126*.com*
Received January 5, 2009; *revised May* 15, 2009; *accepted July* 12, 2009

ABSTRACT

Legacy IP address-based access control has met many challenges, because the network nodes cannot be identified accurately based on their variable IP addresses. "Locator/Identifier Split" has made it possible to build a network access control mechanism based on the permanent identifier. With the support of "Locator/Identifier Split" routing and addressing concept, the Identifier-based Access Control (IBAC) makes network access control more accurate and efficient, and fits for mobile nodes' access control quite well. Moreover, Self-verifying Identifier makes it possible for the receiver to verify the packet sender's identity without the third part authentication, which greatly reduces the probability of "Identifier Spoofing".

Keywords: Access Control, Locator/Identifier Split, IBAC, Self-Verifying Identifier, Identifier Spoofing

1. Introduction

In the current TCP/IP architecture, IP address has dual semantic functions, which indicates both the network node's routing locator and its endpoint identifier [1]. It means that the IP address is a variable label related to the location. Because of the "IP Overload" [1], IP address-based access control has met many challenges.

Firstly, IP address-based access control limits the resource access when a node changes its location. Network services often distinguish users by their IP addresses, so many services are bound with the clients' locations. As a result, when a user of an authorized organization moves to another location (and so the IP address is changed.), he will lose the access ability of the service.

Secondly, "IP Overload" makes IP address-based access control even more complex, and greatly affects its defense efficiency:

1) Because IP address is a variable label, it can't be used as an accurate identifier of the nodes. Moreover, "IP Spoofing" has made it even more critical. So it is difficult to identify the access source in the network layer, and the attackers can anonymously attack the network devices and services.

2) IP address can't match users precisely [2]. One IP address can represent different nodes at different time. On the other hand, one IP address can also represent multiple nodes simultaneously (e.g. NAT). As a result, the attacker can hide his true identity easily.

For the above reasons, the efficiency of IP address-based access control is greatly declined, and some misuses will harm the valid users.

Finally, the changes of the network topology and the ISP policies will lead to the reconfiguration of the IP addresses. Thus, many access control rules and configurations based on IP addresses have to be modified. Undoubtedly, this will make the access control management more complex.

The reason of the above drawbacks lies in that there is no accurate, unique and permanent identifier to describe a network node. So the key problem is to resolve the "IP Overload" problem. IAB announced that in order to resolve the "IP Overload", two name spaces should be introduced to denote a network node's locator and identifier separately, which is called "Locator/Identifier Split" [3]. The communication session is based on the permanent Identifier, and the routing is based on the variable Locator.

In this paper, we propose LISA Network Access Control (LISA-NAC) which is a new network access control mechanism based on the Locator Identifier Separation Architecture (LISA) [4]. The main contributions of LISA-NAC are the Identifier Based Access Control (IBAC) model and the Self-Verifying Identifier, which will make network access control more efficient.

The rest of this paper is organized as follows. Section 2 presents an Overview of LISA Architecture. Section 3 describes some new characters of LISA-NAC, including IBAC model and Self-Verifying Identifier. Section 4 gives an outline of our future work. Finally, we conclude with a summary of the main research result in Section 5.

2. LISA Overview

LISA is a network-based “Locator/Identifier Split” naming and addressing architecture, which borrowed come ideas of LISP [5]. As Figure 1 shows, the network is divided into two parts: kernel network and edge network. The kernel network uses Locator name space, while the edge network uses permanent Identifier name space. The communication session is built on permanent Identifier, but the mapped Locator is variable.

LISA adopts “Mapping + Encapsulation” method to process packets. LISA Router (Edge router) maps the Identifier space into Locator space by querying distributed mapping service system based on one-hop hash (LISA-Mapping). Moreover, LISA Router can update the mapping record in the LISA-Mapping. The Identifier space is a new name space (see Subsection 3.2). The Locator space can reuse the legacy IP address space (IPv4/v6), which will avoid updating network devices in the kernel network.

When a LISA Route receives the packet from host, it queries the LISA-Mapping for the matched Locator according to the packet’s Identifier. After receiving the mapped Locator, the LISA Router adds a new packet header (including the Locator) to the original packet. So in the encapsulated packet, the inner address is an Identifier, and the outer address is a Locator. LISA uses Identifier to denote the node identity, and uses Locator to forward packet in the kernel network. When the encapsulated packet arrives at the destination (the LISA Router), the LISA Router decapsulates the packet, and forwards the original packet to the destination host according to the Identifier.

3. LISA-NAC

In order to improve the efficiency of network access control, network accountability should be mentioned. Network accountability is the capability to identify network entity (user, host and device) and distinguish mal-traffic. However, limited by the “dumb” network infrastructure, it is difficult to achieve accountability in the Internet. There is no accurate, unique and permanent identifier to identify network entity. IP header is too simple, more state information (e.g. identifier) should be added to satisfy the needs of security, QoS and network management.

In the LISA, LISA-NAC runs on the permanent Identifier name space, and provides an accurate and efficient fine-grained access control mechanism for the edge network. The main features of LISA-NAC are the IBAC model and the Self-Verifying Identifier.

3.1. IBAC Model

Different from the traditional network access control, IBAC makes access control policies based on the network node’s true permanent Identifier, not IP address or device port.

IBAC includes three entities: Identifier (I), Object (O) and Permission (P). There are two types of Identifiers: Individual Identifier (I^2) and Identifier Affiliation (IA).

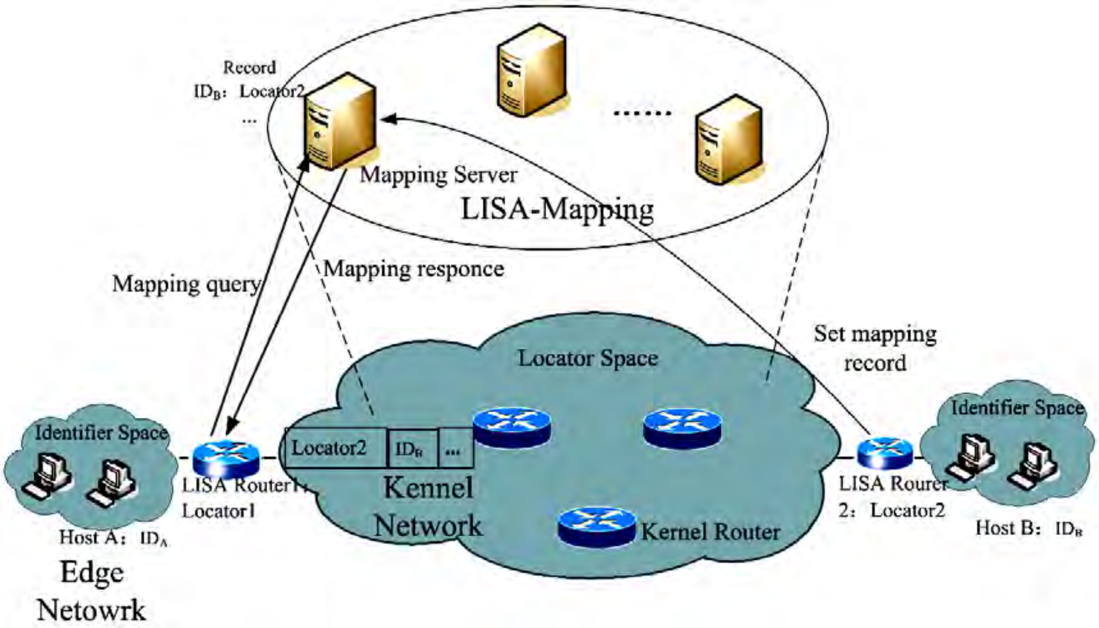


Figure 1. LISA architecture.

I^2 denotes the single network node, and IA denotes a group of network nodes.

IBAC uses three-tuple (I, O, P) to describe an authority. If there exists a (I, O, P), it indicates that I can perform P on the O. Particularly, (I^2, O, P) indicates that single I can perform P on the O, and (IA, O, P) indicates that a group of I can perform P on the O.

IBAC provides end to end security mechanism and fine-grained access control. For example, if several users share a locator (e.g. IP address), IBAC can make independent security policy for everyone. In order to simplify the format of the access control policy and reduce the ACL's size, IBAC uses the IA to classify Identifiers, and adopts unified operation on the Identifiers which have the same IA. IA is not directly in the packet header, and is stored in the LISA-Mapping system. The destination should query the LISA-Mapping system for the matched IA.

IBAC guarantees the access control policy's long term stability. Although the network entities' Locators are variable, the access control policies based on the permanent Identifier are unchanged, so the valid users can always use their services. So IBAC can fit for the mobile node's access control. IBAC avoids the policy updates due to the Locators' changes, and greatly reduces the workload of maintaining the access control policy.

In current network, in order to achieve end to end authority control, network access control should collaborate with the access control mechanisms of the system or application software. Since IBAC guarantees the end to end access control and provides network accountability, it is possible to simplify the upper layer's access control. If the Identifier can be combined with the user's biology properties in the future, the network will be aware of the user's identity and behaviors, and thus no more needs of user's accounts and passwords.

3.2. Self-Verifying Identifier

True Identifier is the basis of IBAC. Similarly, IBAC also meets the potential threat of "Identifier Spoofing". So we introduce "Self-verifying Identifier" in the LISA-NAC. With Self-verifying Identifier, the receiver can verify the sender's identity based on the packet's Identifier without the participation of third part authentication.

In the LISA, every network node gets a pair of asymmetry keys from the CA. The node holds the private key, and makes the public key as the node's globe unique identifier. In other words, the identifier name space is a public key space. LISA-NAC ensures the consistency between the Identifier and the node's identity through the digital signature mechanism.

Self-verifying Identifier simplifies packet's source Identifier verification, and strengthen the scalability because there is no need for the third part authentication. At present, we adopt 160-bit Self-verifying Identifier.

Since the Identifier is actually a public key, we should choose an appropriate asymmetry keys generation algorithm. Traditional asymmetry keys algorithms such as RSA, DSA and Diffie-Hellman often choose long keys to guarantee the key's safety. For example, a normal RSA key is 1024-bit. However, such long key is unfit for the Identifier. Firstly, long identifier increases the packet's size, which may lead to packet fragment and consumes additional bandwidth. On the other hand, since 128-bit Identifier space is enough for current IPv6 network size, it is useless to make a huge Identifier name space.

In the LISA-NAC, we use ECC (Elliptic curve cryptography) algorithm to create a pair of 160-bit asymmetry keys for every network node. ECC's advantages lie in:

1) ECC offers security equivalent to RSA using much smaller key size. For example, ECC 160-bit key offers security equivalent to RSA 1024-bit key [6]. This property will reduce the engineering challenges brought by long key.

2) ECC generates asymmetry keys pair faster than RSA does for the comparable length [7]. Considering the signature generation and verification, ECC's processing speed is much faster than that of RSA [8]. This makes it possible to implement packet digital signature verification with limited packet delay.

At present, 109-bit ECC key has knocked over with brute force. However, the secure 160-bit ECC key is approximately one hundred million times harder to crack than 109-bit ECC key [9]. So we think that 160-bit ECC key can fit the Identifier length, as well as satisfy the basic security requirements.

Figure 2 shows the verification process of Self-verifying Identifier. ID_s and ID_d denote the packet's source and destination Identifier separately. In fact, ID_s and ID_d are the sender and receiver's public key. Dig is the packet's digest. Sig is the digital signature. The receiver identifies the true sender though verifying packet's signature.

If an attacker disguises as the sender and sends a packet, he must have the sender's private key to generate

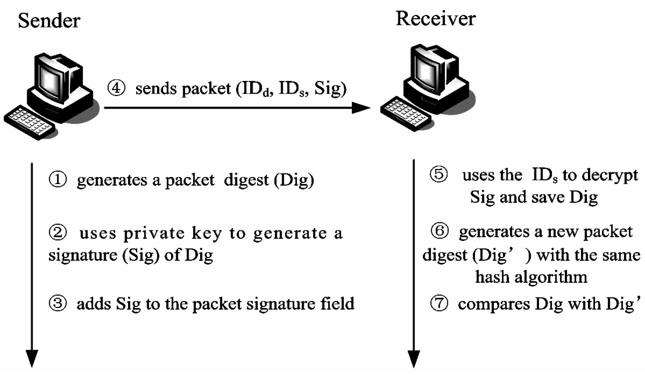

Sender
Receiver
④ sends packet (ID_d, ID_s, Sig)
① generates a packet digest (Dig)
② uses private key to generate a signature (Sig) of Dig
③ adds Sig to the packet signature field
⑤ uses the ID_s to decrypt Sig and save Dig
⑥ generates a new packet digest (Dig') with the same hash algorithm
⑦ compares Dig with Dig'

Figure 2. Self-verifying identifier verification.

the correct encrypted signature. Since the attacker doesn't have the sender's private key, when the receiver generates a new packet digest (Dig'), it must be different from the decrypted original packet digest (Dig). So the "Identifier Spoofing" can be detected.

The packet carries the public key, and there is no key exchange during the node identity verification. Obviously, it will simplify the identity verification process. Since network access control is deployed to protect the important services, it is unnecessary to include signature verification in the general packet processing. Most of the network nodes can choose the packet signature verification as an option, but the packet signature is imperative. Moreover, a node can publish its Identifier to the DNS so that all the other nodes can get its public key to encrypt data.

4. Future Work

In the LISA-NAC, verifying signature on every packet will undoubtedly add packet delay. The transmission performance degradation is what we are concerning about. A prototype is under development, and we will measure the main transmission performance (delay, loss and throughput) changes to test the feasibility of LISA-NAC.

At present, Identifier only indicates the network node's property not including the user's property. Next step, we will try to combine the Identifier with the user's biology properties. Then the network will be aware of users' identity and behaviors.

5. Conclusions

LISA separates the network node's identity from location, which makes it possible to build a network access control mechanism based on the identifier. IBAC makes network access control more accurate and efficient. Moreover, IBAC fits for the mobile node's access control. Since true Identifier is the basis of IBAC, "Identifier Spoofing" must be avoided. Self-verifying Identifier makes it possible for the receiver to verify packet sender's identity without the third part authentication, which simplifies the packet source verification. We think that LISA-NAC is a concrete step to strengthen network security through the "Locator/Identifier Split".

6. Acknowledgements

This work was supported by China "863" Project (No. 2008AA01A325) and China National Grand Fundamental Research "973" Project (No. 2009CB320503).

7. References

[1] J. Scudder, "Routing/addressing problem solution space," 2007, http://www.arin.net/meetings/minutes/ARIN_XX/PDF/wednesday/SolutionSpace_Scudder.pdf

[2] R. Tu, J. S. Su, Z. W. Meng, and F. Zhao, "UCEN: User centric enterprise network," in Proceedings IEEE ICACT'08, Phoenix Park, Korea, pp. 66–71, Feb 2008.

[3] D. Meyer and K. Fall, "Report from the IAB workshop on routing and addressing," Internet Draft, 2006.

[4] R. Tu and J. S. Su, "A hash-based locator/ID mapping mechanism," The Computer Engineering and Science, No. 1, pp. 9–12, 2009.

[5] D. Meyer, "The locator identity separation protocol (LISP)," The Internet Protocol Journal, Vol. 11, No. 1, pp. 23, 2008

[6] A. J. Menezes, "Elliptic curve public key crytosystems," Kluwer International Series in Engineering and Computer Science, 1993.

[7] N. Jansma and B. Arrendondo, "Performance comparison of elliptic curve and RSA digital signature," Technical Report, 2004. http://www.nicj.net/files/498termpaper.pdf.

[8] Certicom Corp, "The elliptic curve crypto system for smart cards," Certicom White Paper, 1998, http://www.comms.scitech.susx.ac.uk/fft/crypto/ECC_SC.pdf.

[9] W. Chou and Laerence, "Elliptic curve cryptography and its applications to mobile device," Project Report, University of Maryland, 2003, http://www.cs.umd.edu/Honors/reports/ECCpaper.pdf.

Int. J. Communications, Network and System Sciences, 2009, 7, 645-651
doi:10.4236/ijcns.2009.27073 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

Efficient Techniques and Algorithms for Improving Indoor Localization Precision on WLAN Networks Applications

Antonio del CORTE-VALIENTE[1], Jose Manuel GÓMEZ-PULIDO[2], Oscar GUTIÉRREZ-BLANCO[2]
[1]*Computer Engineering Department; University of Alcalá, Alcalá de Henares, Madrid, Spain*
[2]*Computational Science Department, University of Alcalá, Alcalá de Henares, Madrid, Spain*
Email: {*antonio.delcorte, jose.gomez, oscar.gutierrez*}*@uah.es*
Received May 26, 2009; *revised June* 28, 2009; *accepted August* 1, 2009

ABSTRACT

This paper proposes efficient techniques that allow the deploying of high precision location applications for indoor scenarios over Wireless Local Area Networks (WLAN). Firstly, we compare the use of radio frequency (RF) power levels and relative time delays based on ray-tracing as detection methods to estimate the localization of a set of mobile station using the fingerprint technique. Detection method play an important role in applications of high frequencies techniques for locations systems based on current and emerging standards such as Wi-Fi (802.11x) and Wi-Max (802.16x). The localization algorithm computes the Euclidean distance between the samples of signals received from each unknown position and each fingerprint stored in the database or radio-map obtained using the FASPRI simulation tool. Experimental results show that more precision can be obtained in the localization process by means of relative delay instead of RF power detection method. Secondly, the Euclidean distance has been compared with others similarity distance measures. Finally, an interpolation algorithm between the fingerprinting weighing based on the distances has been implemented in order to eliminate those fingerprints that do not contribute to the improvement in the accuracy. These techniques allow obtaining more precision in the localization of indoor mobile devices over WLAN networks.

Keywords: Localization, Euclidean Distance, Interpolation, Mahalanobis Distance, Faspri, Fingerprinting, GTD

1. Introduction

In this work the problem of indoor localization based on the signals available from the wireless devices [1,2] that comprise the Wi-Fi and Wi-Max standards is presented. The localization process is done by using the fingerprinting technique [3,4] that operate the relationship between the power levels and the relative delays between signals due to multipath reflections. In comparison with other techniques, such as angle of arrival (AOA) or time of arrival (TOA) that present several challenges due to multipath effects and non-line-of-sight (N-LOS) [2], the fingerprinting technique is relatively easy to implement. In traditional indoor localization systems based on Wi-Fi networks, the Euclidean distance is used as a metric in the localization process and the fingerprinting technique is based on the power levels detected by means of the received signal strength indicator (RSSI) parameter available on the 802.11x standard between the radio frequency (RF) power levels of the received signals and the samples stored in the database or radio-map. However, due to the development of new radio access standards, such as Wi-Max, it is necessary to explore new techniques to improve the precision by using alternative detection methods. In a previous work [8] the fingerprinting technique has been implemented using the relative delays as fingerprint and the Euclidean distance also as metric. More precision compared to the power detection technique was obtained. However, the localization precision can still be increased by using different techniques within the fingerprinting matching process [7,8]. Firstly, the Euclidean distance metric has been replaced by others metrics such as Manhattan, Bray-Curtis, Chi-Squared and Mahalanobis distances. Secondly, an interpolation between the fingerprinting weighing based on the distance was implemented. The results obtained demonstrate the accuracy of these techniques.

2. Ray-Tracing Model

The ray-tracing model can be obtained with the FASPRI [4] simulation tool, that is able to make a 3D indoor propagation analysis by means of deterministic methods [5,6]. FASPRI (Figure 1) is a ray-tracing code based on geometric optic (GO) and the uniform theory of diffraction (UTD).

In order to optimize the program computing time, ray-tracing algorithms such as the angular zeta-buffer (AZB) or the space volumetric partitioning (SVP) [5,6] have been implemented. These algorithms make it possible to simulate a great number of case-studies in a reasonable amount of time. These results can be used to examine the effect of varying certain sensing parameters on the precision of the system such as the number of antennas, the position of the antennas and the number of tracks. The electric field levels can be obtained using the direct, reflected, transmitted, diffracted ray or combinations of these effects. Figure 2 shows the scene where the simulations take place as well as the multipath raytracing effects.

An advantage of using the ray-tracing techniques is that, besides obtaining the power level of a series of points, information can also be obtained about the multipath effects, such as the relative delays between rays and the directions of arrival. This information can be used as a fingerprint in the fingerprinting method with the purpose of improving the efficiency of the localization system.

3. Fingerprinting Technique

The fingerprinting technique can be divided in two phases [2]. In the first one, it obtains the radio map or fingerprinting database. The radio-map of fingerprints are obtained by performing an analysis of the relative ray-tracing delay (Figure 2) and signal strength (Figure 3) from multiple APs over a defined grid. The vector of received signal of power and relative ray delay samples at a position on the grid is called the location fingerprint of that point. In the second phase, it analyzes the accuracy obtained in the localization process. For this purpose, the developed technique places a significant number of mobile stations into the area covered by the radio map and it obtains the vector of received samples from different APs [7]. The location estimation is made by an algorithm that computes the distance between the measured samples and each fingerprint stored in the radio map. The X and Y coordinates associated with the fingerprint that results in the smallest distance are returned as the position estimation. Figure 3 shows the relative delays between the detected rays in a fingerprint and their contribution to the total field due to the different ray-tracing effects. Figure 4 shows the power levels available for all the fingerprints in a regular grid. In both cases the grid example correspond with 72x72 points (30x30 meters area size) being 2.4GHz the radiation frequency of the antenna.

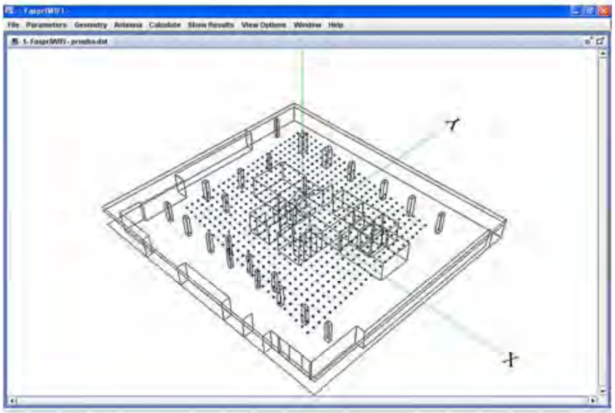


Figure 1. Multipath ray-tracing effects.

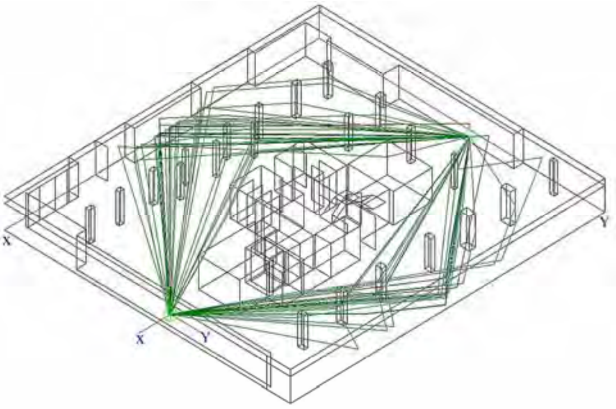


Figure 2. Multipath ray-tracing effects.

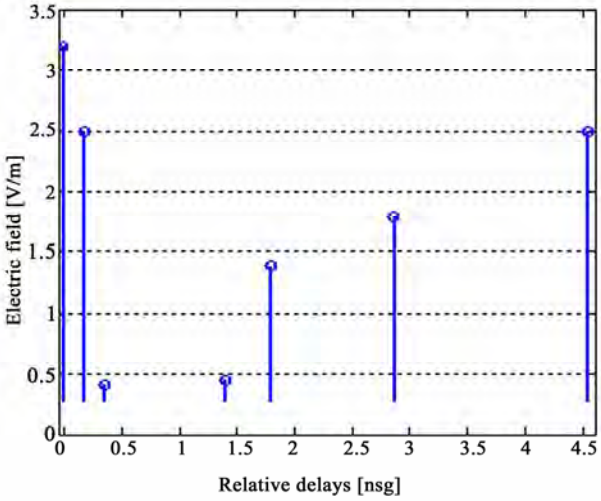


Figure 3. Relative delays between rays in a fingerprint vs. electrical field.

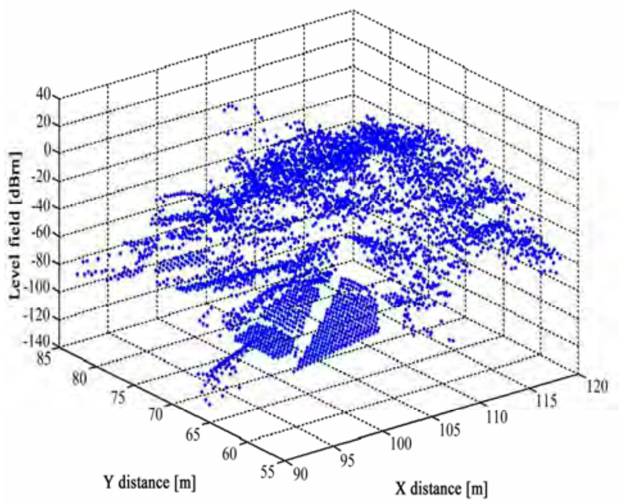


Figure 4. Power levels available for all the fingerprints in a regular grid.

4. Distance Metrics

Distance Metric is the key component used by the fingerprinting technique. By this reason it is important to explore different similarity measures to find the best distance metric that minimizes the positioning average error. The method implemented in the localization algorithm is based on to compute the distance metric between the vectors of received signals X and the vector of samples stored in the radio map Y. Then it determines the points of the grid that corresponds with the position of the mobile stations. The coordinates X and Y that corresponds with the vector Y that has a smaller distance with the vector of samples X for a certain position of the mobile stations are selected as solution. Five equations have been implemented to explore which will improve more localization accuracy.

$$D_E(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \qquad \textbf{(1)}$$

$$D_M(x,y) = \sum_{i=1}^{N}\left|x_i - y_i\right| \qquad \textbf{(2)}$$

$$D_{BC}(x,y) = \sum_{i=1}^{N}\frac{\left|x_i - y_i\right|}{x_i + y_i} \qquad \textbf{(3)}$$

$$D_{CHI}(x,y) = \sum_{i=1}^{N}\frac{(x_{i-}y_i)^2}{x_i + y_i} \qquad \textbf{(4)}$$

$$D_{Mah}(x,y) = \sqrt{(x-y)^{'}Cov(x)^{-1}(x-y)} \qquad \textbf{(5)}$$

In the Euclidean metric Equation (1) the mobile station will be more similar to the fingerprint radio map if the distance is smaller. More moderate approach implemented in Manhattan metric Equation (2) is by using sum of the absolute differences rather than their squares, as the overall measure of dissimilarity. On the other hand, in Bray-Curtis and Chi-Squared metrics Equations (3) and (4) the numerator signifies the difference and denominator normalizes the difference. In Mahalanobis metric equation (5) the (x-y)' term denotes the (x-y) transpose vector and the Cov term denotes the covariance matrix, where retrieval performance is sensitive to the sample topology.

5. Interpolation Algorithm

In our next experiment we have implemented an interpolation between four fingerprinting weighing based on its metric distance. For it, the coordinate of the wished point cannot correspond to the fingerprinting (Figure 5). The coordinates of the point where it is considered locating the mobile station, DP, it is determined by means of the Expression (6), where Nf is the fingerprinting number, X_j and FP(x,y) are respectively, the value of the corresponding distance and the coordinate corresponding to the j fingerprinting. In order to increase the precision when the mentioned interpolation is applied we made the weighing with the four fingerprinting that presented the smaller distance.

$$DP(x,y) = \frac{\sum_{j=1}^{Nf}\left[FP(x,y)/X_j\right]}{\sum_{j=1}^{Nf}\left(1/X_j\right)} \qquad \textbf{(6)}$$

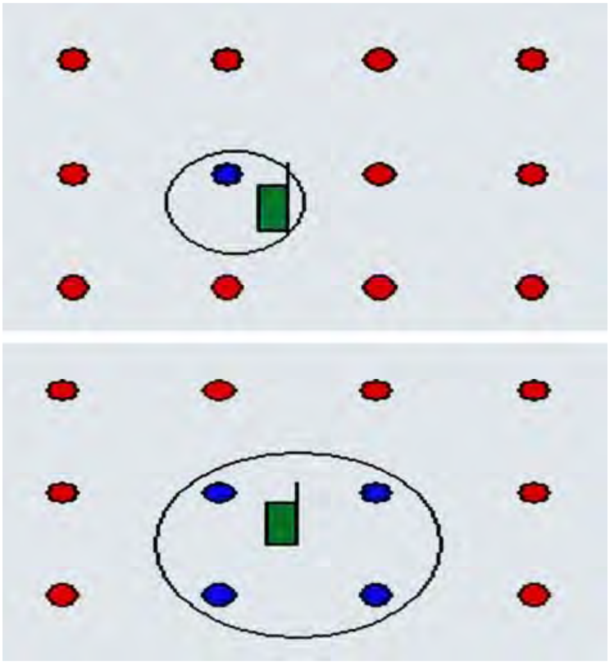


Figure 5. Location algorithm without interpolation and with 4 points of interpolation.

In this way we will eliminate those fingerprints that, due to the great distance to the objective point, present a higher distance and therefore do not contribute to the improvement in the accuracy.

6. RF Power vs. Relative Delay Detection Method

In this case we compare the use of radio frequency (RF) power levels and relative time delays based on ray-tracing as detection methods to estimate the localization of set of mobile stations using the fingerprint technique. The metric considered was the Euclidean distance. The information provided by the simulation tool is stored in four vectors. Two of them, Ph and Th, correspond to the information at every fingerprint. The first vector contains the power level from the N access points at the fingerprint h and the second one contains the relative delay at the same point. The Pm and Tm vectors contain the same information at the mobile station m (Expressions 7 to 10).

$$Ph = \left[Ph_1, Ph_2, Ph_3, ..., Ph_N\right] \quad \textbf{(7)}$$

$$Th = \left[Th_1, Th_2, Th_3, ..., Th_N\right] \quad \textbf{(8)}$$

$$Pm = \left[Pm_1, Pm_2, Pm_3, ..., Pm_N\right] \quad \textbf{(9)}$$

$$Tm = \left[Tm_1, Tm_2, Tm_3, ..., Tm_N\right] \quad \textbf{(10)}$$

These vectors are calculated at the beginning of the process and are stored in a database. The Euclidian distance between each mobile station and every fingerprinting is calculated using the Expression (11) where the parameter *v* is a weighting factor that indicates the correlation ratio between the powers and delays. This factor is set to 0 to find the distance by using the power levels, set to 1 to find the distance by using the relative delays and set to 0.25 to find the distance with a hybrid method of the power and delays. The position of the mobile corresponds with the fingerprint whose Euclidean distance is smaller.

$$D(x,y) = \sqrt{\sum_{i=1}^{N}\left[(1-\upsilon)(Pm_i - Ph_i)^2 + \upsilon(Tm_i - Th_i)^2\right]} \quad \textbf{(11)}$$

The number of fingerprints and the frequency are parameters that would affect the precision of the results. For this reason the experiment considers two grids: one consisting of 72x72 fingerprints and another composed of 36x36, as well as two different frequencies: 2.4GHz and 5.2GHz. The distance between the fingerprints in the first and second grids are 40cm and 80cm, respectively. The simulation also placed 9 AP's at the above-mentioned frequencies and 99 mobile stations randomly distributed over the grids. Figure 6 shows the area of 28.8x28.8 meters where the simulations have been done.

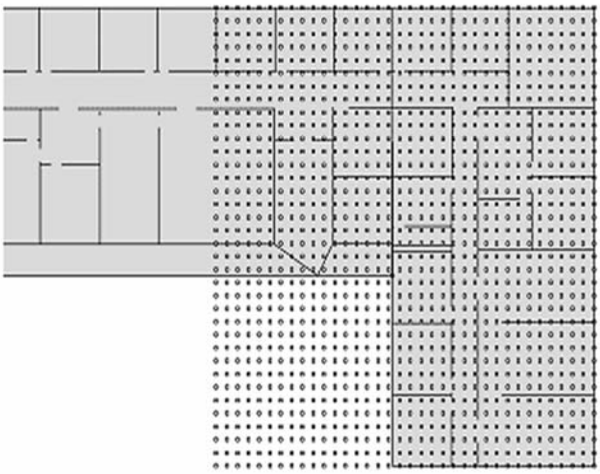


Figure 6. Regular grid example.

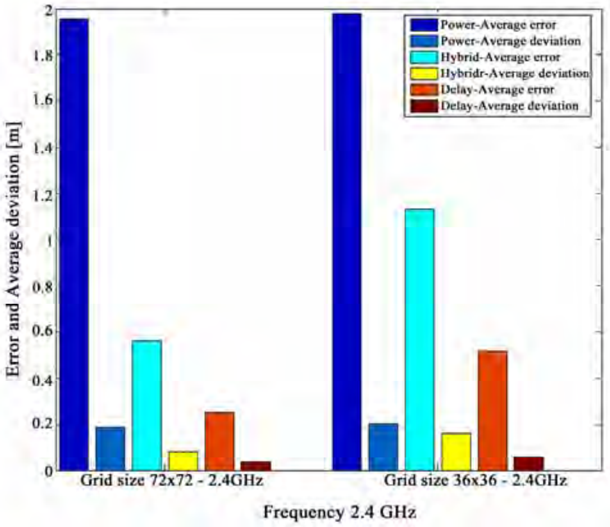


Figure 7. Detection methods comparison – Frequency of 2.4GHz.

In order to evaluate the benefits of the detection method two statistical indicators have been used. These indicators are the total mean error and the total mean deviation (standard deviation). Data obtained after the FASPRI simulations have been analyzed under MATLAB. Figure 7 shows the results obtained at 2,4GHz and Figure 8 the results obtained at 5,2GHz. In both graphs the three detection methods analyzed are compared by means of the statistical indicators. With these results, we can confirm that the relative delay detection method provides better results than the power detection method for any grid size and frequency. The hybrid detection, although it provides worse results than using delays, can reduce the detection ambiguity at the cost of increasing the mean error. However, it should be noted that the mean error is reduced when the number of fingerprints is increased for the same frequency, independent of the detection method used. In addition, as far as the impact of the frequency used with the detection method is concerned, we can observe that, the mean error is lower at 2,4GHz using power detection; however, it stays constant when using delay and hybrid detections. Therefore, the grid density is a more critical factor than the frequency. Table 1 shows the mean error and standard deviation values obtained when comparing the three detection methods. The two last columns of the table present the percentage of the mean error that is possible to improve when using either hybrid versus power detection and delay versus power detection. Finally, the results obtained have been analyzed with a statistical point of view. The probability error was calculated by using the position error of the 99 mobiles. Figures 9, 10 and 11 shows the probability error distribution for the three methods analyzed in the case of 72x72 grid size for 2.4GHz. It should be noted that the error distribution function reduces extremely by using delay detection compared with power or hybrid detection methods.

7. Distance Metric Comparison and Interpolation Algorithm Effect

Distance Metric is the key component used by the fingerprinting technique. By this reason, it is important to explore different similarity measures to find the best distance metric. Five equations have been implemented to explore which will improve more localization accuracy: Euclidean, Manhattan, Bray-Curtis, Chi-Squared and Mahalanobis distance. In this case the relative delay detection method was been implemented in the fingerprinting algorithm. Finally an interpolation of the four best distances has been added. An irregular geometry of

Table 1. Mean error and typical deviation detection methods comparison.

	Power Detection	Hybrid Detection	Delay Detection	Hybrid vs. Power	Delay vs. Power
Grid size and Frequency	Mean error [m] Typical deviation [m]			Mean error Improvement (%)	
72x72 2.4GHz	1,9554 0,1871	0,5621 0,0821	0,2504 0,0366	71.28	87.17
72x72 5.2GHz	2,0844 0,1843	0,5207 0,0841	0,2504 0,0366	75.00	87.98
36x36 2.4GHz	1,9801 0,2029	1,1337 0,1610	0,5155 0,0572	32.82	74.24
36x36 5.2GHz	2,5275 0,2548	0,9497 0,1142	0,5155 0,0572	62.69	75.48

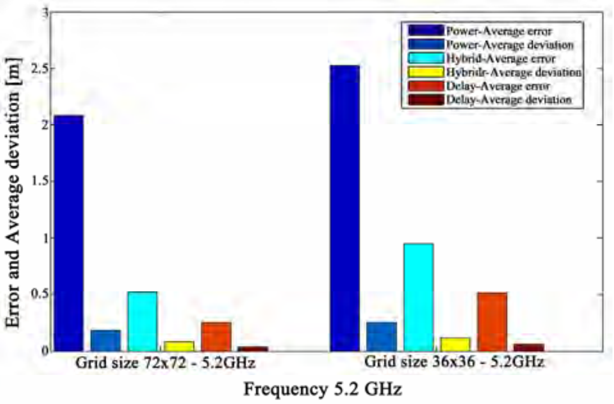


Figure 8. Detection methods comparison–frequency of 5.2GHz.

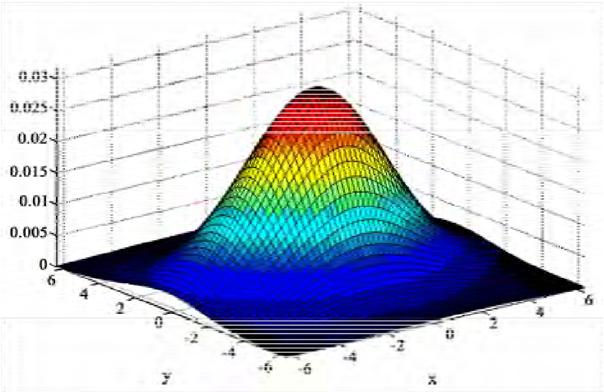


Figure 9. Probability error distribution – power detection.

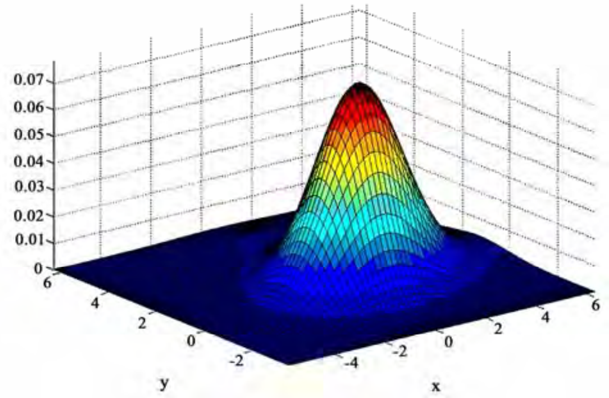


Figure 10. Probability error distribution – hybrid detection.

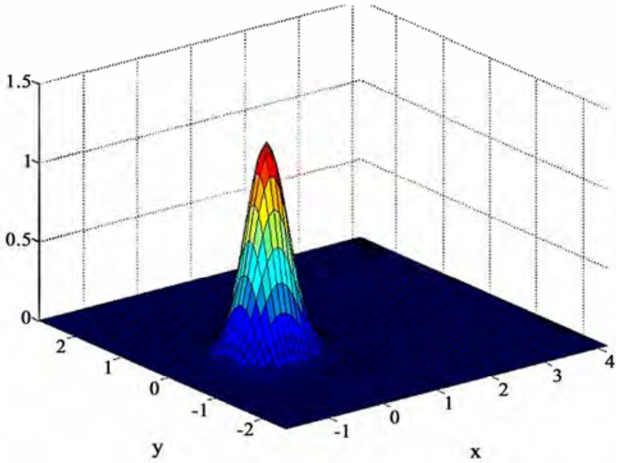


Figure 11. Probability error distribution – relative delay detection.

36x36 meters that corresponds with a section of the polytechnic building has been analyzed (Figure 12). In it, 9 antennas and 5184 fingerprints has been tested being the frequency of the antenna 2.4 GHz and 100 the number of mobiles stations to detect.

Running the experiment we are able to find which distance metric gives best result. For this purpose Matlab tool has been used. Figure 13 show a detailed comparison of the accuracy obtained using the five similarity measures, with and without adding the interpolation algorithm. Two statistical indicators have been used, the total mean error and the total mean deviation to evaluate the benefits. It is clear that conventional distances metrics like Euclidean or Manhattan does not perform the best results. The average localization error obtained with these metrics is very similar. Therefore, we can affirm that to calculate the sum of the absolute differences or their squares in each fingerprinting iteration process has the same effect. On the other side, Bray-Curtis and Chi-Squared distances present better results than the previous metrics due to the normalization realized in their expressions. Finally, the Mahalanobis distance, where the covariance matrix is sensible to the topology of the radio-map, presents results slightly better than the rest of metrics, due to the irregularity of the topology. Similar conclusions are obtained in the case of being applied the technique of interpolation, but being observed, in this case, a reduction in the two statistical indicators (Table 2).

8. Conclusions

In this work alternative detection methods that can be used in the fingerprinting technique for mobile localization has been presented. These methods make possible the analysis and design of indoor localization services over WLAN networks. A comparative study between detection methods based on RF power and relative delays has firstly been implemented. Secondly, we have presented a detailed comparison of five different similarity metrics to test the performance of the algorithm. Finally, an interpolation between the fingerprinting weighing based on its distance has been tested. We can conclude that relative delay detection technique, which is used in emerging standards such as WiMax (802.16x), presents better results in the indoor localization process than the power detection technique used in traditional Wi-Fi systems (802.11). On the other side, conventional distance metrics like Euclidean or Manhattan does not perform necessarily the best accuracy. On opposite, Mahalanobis distance metric improve the results when the geometry has irregularities that can been modeler between measures by the covariance matrix. Finally, we conclude that the interpolation technique eliminate those fingerprints that do not contribute to the improvement in the accuracy.

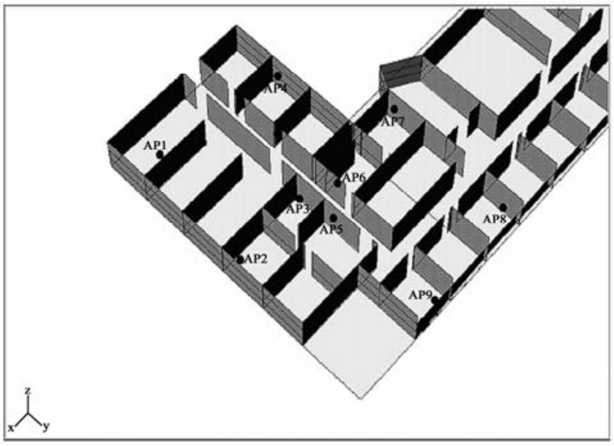


Figure 12. Irregular section of polytechnic building.

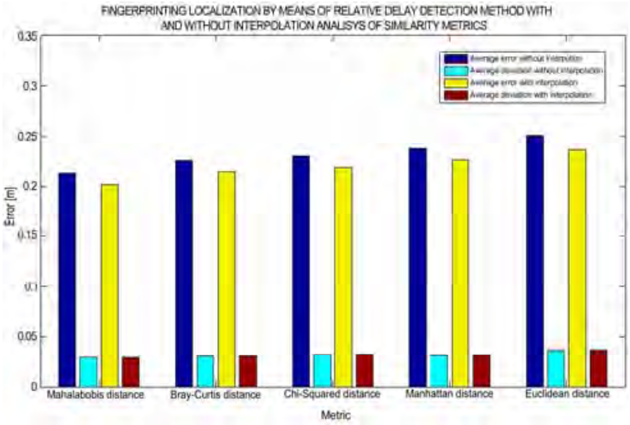


Figure 13. Similarity metrics comparison and interpolation effect.

Table 2 Mean error [m] - metrics and interpolation comparison

Distance	Without Interpolation	With Interpolation
Mahalanobis	0.2128	0.2021
Bray-Curtis	0.2253	0.2140
Chi-Squared	0.2303	0.2187
Manhattan	0.2378	0.2259
Euclidean	0.2504	0.2367

9. Acknowledgement

This work has been financed by the Community of Madrid, project S-0505/TIC/0255, and by the Ministry of Education and Science, project TEC-2006-03140.

10. References

[1] Cisco Wireless Location Appliance-Products. Datasheet 2006.

[2] P. Bahl, *et al*, “Enhancements of the radar user location and tracking system,” Microsoft Research Technology, February 2000.

[3] K. Kaemarungsi and P. Krishnamurthy, “Properties of indoor received signal strength for WLAN location fingerprinting,” Proc. First Annual International Conf. on Mobile and Ubiquitous Systems: Networking and Services, pp. 14–23, August 2004.

[4] F. Sáez de Adana, *et al*. “Propagation model based on ray tracing for the design of personal communication systems in indoor environments,” IEEE Trans. on Vehicular Technology, Vol. 49, pp. 2105–2112, November 2000.

[5] M. F. Cátedra and J. Pérez-Arriaga, “Cell planning for wireless communications,” Artech House Publishers, Boston, 1999.

[6] K. Kaemarungsi, “Distribution of WLAN received signal strength indication for indoor location determination,” Proc. First International Symposium on Wireless Pervasive Computing, CD-ROM, pp. 6, January 2006.

[7] A. del Corte-Valiente, J. M. Gómez-Pulido, O. Gutiérrez -Blanco, and M. F. Cátedra-Pérez, “Algoritmos eficientes de localización en interiores basados en técnicas de trazado de rayos,” XXIII Simposium Nacional de la Unión Científica Internacional de Radio (URSI). Madrid (Spain), September 2008.

[8] A. del Corte-Valiente, J. M. Gómez-Pulido, O. Gutiérrez-Blanco, and M. F. Cátedra-Pérez, “Efficient Techniques for indoor localization based on WLAN networks,” Second International Workshop on User-Centric Technologies and Applications (MADRINET), Salamanca (Spain), October 2008.

Int. J. Communications, Network and System Sciences, **2009, 7, 652-656**
doi:10.4236/ijcns.2009.27074 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

Optical Network Traffic Control Algorithm under Variable Loop Delay: A Simulation Approach

Manoj Kr DUTTA, Vinod Kumar CHAUBEY
Electrical and Electronics Engineering Department, Birla Institute of Technology & Science,
Pilani, Rajasthan, India
Email: *mkdutta*13*@gmail.com, vkc@bits.pilani.ac.in*

Received May 9, 2009; *revised July* 16, 2009; *accepted August* 22, 2009

ABSTRACT

In this paper we present a concept of new architectural model consisting of multiple loop delay to increase the throughput. The simulated behavior of an optical node has been realized by using an n x m optical switch and recirculating optical delay lines. This investigation infers the scaling behaviors of the proposed architecture to maintain efficient use of the buffer under Poisson traffic loading. The analysis also reports the traffic handling capacity for the given complexity of the node architectural design.

Keywords: Fiber Delay Line, Recirculation, Traffic, Throughput

1. Introduction

ALL-OPTICAL communication has been proposed as a promising candidate for providing high-speed networking [1–4] owing to huge bandwidth of optical channels. Wide bandwidth available in low attenuation window in the optical fiber can be divided into a number of independent wavelength channels as per network standard and specification leading to SONET, SDH or wavelength division-multiplexing (WDM) based all optical network system [5–7]. Evidently to support such all optical control in these networks several technologies have been proposed for efficient networking viz., broadcast and select, wavelength routing, optical packet switching (OPS), and optical burst switching [8–10]. In an OPS network, optical interconnect (or optical switch) forwards the packets to their destinations involving programmable switch fabric and control circuitry and thereby support in packet contention resolution. However in a WDM interconnect, output contention arises when more than one packet on the same wavelength are destined to the same output fiber at the same time. To resolve this contention one will have to either temporarily store some of the packets in a buffer, or to convert wavelengths to some available wavelengths by wavelength converters [11–13]. Obviously optical buffers, wavelength converters add the complexity by enhancing the installation and recurring cost of the system. However allocating some dedicated buffers for each output fiber which can share a common optical delay line (ODL) buffer pool [5–6] will essentially reduce the cost and complexity as well. In a WDM a packet that cannot be directly sent to the output fiber is sent back to one of the delay lines for recirculation and after being delayed by some specific time, that packet will come out of the delay line to compete for throughput with the newly arriving packets. In case of unsuccessful throughput it gets back into the delay line for the next round trip with additional delays.

In the proposed model a node has been considered with more input channels than output channels and the maximum capacity of this node is decided by the available output channel. It is assumed that arriving packets are destined to their respective destinations based on First Come First Serve (FCFS) scheduling policy. In this way we can avoid the continuous recirculation of some packet in the delay line. Packets that arrive in the meantime are also sent to delay line. The node includes finite capacity buffer and multiple delay lines arranged in synchronized mode.

2. Node Architecture Design and Modeling

The packet switching has its own (unique) issues in optical networks. In an optical packet-switched network, contention occurs due to unavailability of free output wavelength. In electrical packet-switched networks, contention is resolved with the store-and-forward technique, which requires the packets losing the contention to be stored in a memory bank and to be sent out at a later time

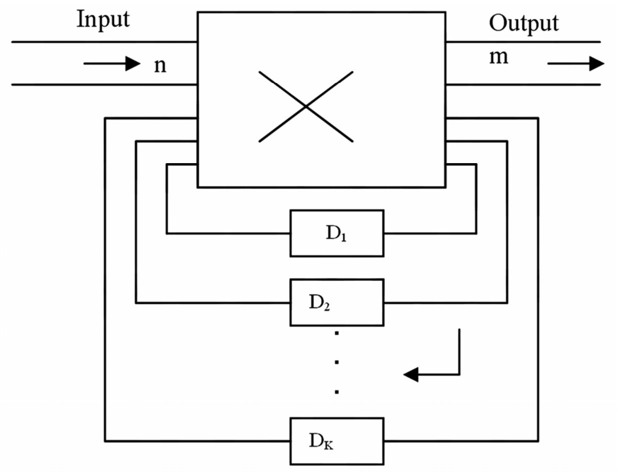


Figure 1. Recirculating delay line optical.

when the desired output port becomes available. This is possible because of the availability of electronic random-access memory (RAM). There is no equivalent optical RAM technology; therefore, the optical packet switches need to adopt different approaches for contention resolution. Meanwhile, WDM networks provide one new additional dimension namely wavelength, for contention resolution. There have been studies in literature for utilizing the three dimensions of contention-resolution schemes: wavelength, time, and space.

In this paper we explore the contention resolution, based on time and propose a new scheduling algorithm for prioritizing the packets within the node. The optical buffers basically delay the incoming signal by making it to travel a small distance, so as to provide some time to the processor for serving them in case the service is not available initially. Now this delay can be provided in fixed quanta's only. This unique feature of optical buffers (unlike their electronic counterpart which 'store' a packet) makes it necessary to have a minimum fixed delay once the packet has entered into the fiber delay line (FDL). Traditionally the buffer is implemented such that once the packet has entered into the FDL it suffers the delay and comes out after that time. The packet might be served had necessary arrangements been made or otherwise dropped. This architecture provides a single chance to server to serve it thus resulting in high packet loss. Ideally the packet should be available at all times at output after having entered the FDL (like equivalent electronic memory) so that it can be served whenever the resources are available.

Our new buffer architecture attempts to realize this objective by giving delays in steps of small granularity D (µSec) which allows the packets to be processed if the resource at output is available otherwise reflected back to the FDL for multiple reflections as per the control algorithm.

It is already assumed that the number of output channels (m) is less than the number of input channels (n) and therefore the queuing system has a fair chance of packet contention. The buffer works with a first-come first-served (FCFS) scheduling policy and is implemented by means of FDL's with reflection.

In the proposed buffer architecture, when the packet arrives, it will be sent to the output node but if all output nodes are busy then it will be placed back in the first loop of the FDL having a delay of $D_{1,}$ after completion of the delay the packet competes for output port, failing this it will again be reflected back into the second delay of D_2 and so on. The maximum delays that can be provided by using FDL's are assumed to have different values of delay such as a constant, arithmetic or a geometric progressive delay.

The flow chart for the packet servicing algorithm involving multiple delays in the proposed node architecture is presented in Figure 2. Obviously as a packet arrives at the node and the server is idle it is served immediately but these are queued if the server is busy. Usually the delays are kept finite by means of the FDL's, due to the limited time resolution related to the granularity and the new packet is going to be delayed at least by an amount of D for one loop circulation. Also the packet can't be made to reflect infinite number of times due to loss of energy at each reflection and hence is limited by accepted SNR.

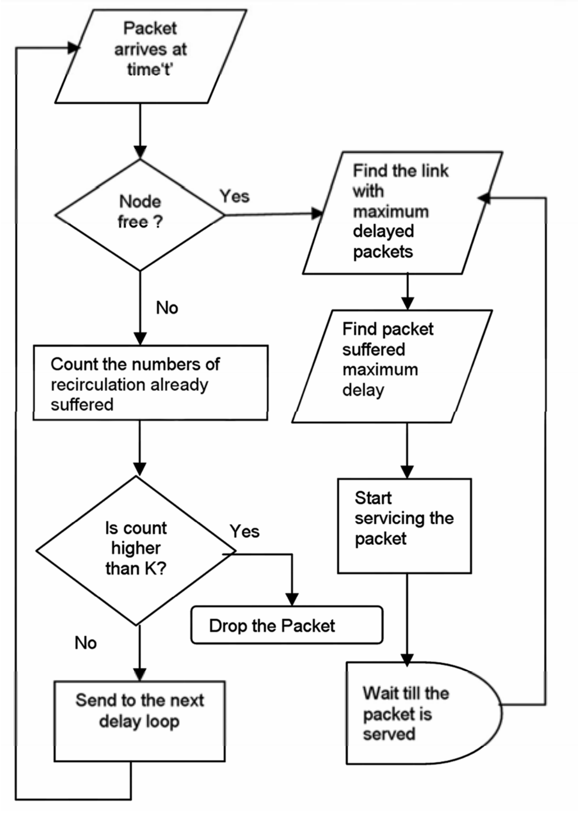

Packet arrives at time't'
Node free ?
Yes
No
Find the link with maximum delayed packets
Find packet suffered maximum delay
Count the numbers of recirculation already suffered
Start servicing the packet
Is count higher than K?
Yes
No
Drop the Packet
Send to the next delay loop
Wait till the packet is served

Figure 2. Flow chart for node performance analysis.

Thus the packet is dropped after K reflections, which is modeled in terms of acceptable quality q and reflection loss α as a function of log (q-α). Considering the evolution of buffer contents over time, we can identify three important variables viz. order of bursts arrival, the packet inter arrival time (IAT) having Poisson distributed (T_k) and the intermittent time between the k^{th} arrival and the next one.

This system is modeled for a random input, having an exponential service with N servers, an infinite number of prospective customers and a maximum queue length of L. System probability for j^{th} call is expressed in term of packet arrival rate λ and packet length t_m as:

$$P_j(A) = P_0(A)\frac{A^j}{j!},\ for \quad 0 \le j \le N \tag{1}$$

$$P_j(A) = P_0(A)\frac{A^j}{N!N^{j-N}} \quad for \quad N \le j \le N+L \tag{2}$$

where P_0 (A) is used to make the sum of P's to unity assuming A as λt_m. Further $P_0(A)$ can be written as:

$$P_0(A) = \left[\sum_{j=0}^{N}\frac{A^j}{j!} + \frac{A^N}{N!}\sum_{j=1}^{L}\frac{A^j}{N^j}\right]^{-1} \tag{3}$$

In the proposed algorithm an incoming packet will be blocked if all the servers are busy & queue is full. However the packet will be delayed if the servers are busy but queue is not completely full. The probability that (N+L) incoming packet is delayed can be written as

$$P_1 = \sum_{j=N}^{N+L-1} P_j(A) \tag{4}$$

Further a packet will be serviced immediately if there are less than N packets in the system and the probability of immediate service of packet is expressed as

$$P_{I\ S} = \sum_{j=0}^{N-1} P_j(A) \tag{5}$$

The waiting time distribution for the incoming traffic can be expressed using the standard equation [14] as

$$P = P_N(A)\sum_{j=0}^{L-1}\frac{\rho^j}{j!}\int_{Nt/t_m}^{\alpha} x^j e^{-x}dx \tag{6}$$

These equations have been used in throughput simulation in the MATLAB environment under the appropriate node and traffic assumptions.

3. Simulation and Results

Traffic throughput of the offered traffic that gets processed through the node has been estimated under various node design parameter constraints. This traffic has been evaluated using Equations (2-6) for the proposed node operated under traffic resolution algorithm. Figure 3 presents the carried traffic corresponding to incoming offered traffic with the variation of number of delay lines (N) involved. The simulated curve shows a linear dependence of the carried traffic on the offered traffic only upto a specific input load but beyond that it deteriorates owing to the rise in the blocking probability. Moreover increased incoming traffic results a crowded node forcing to reject the excess traffic. This qualitative behavior is also supported by the simulation curve showing a rejection beyond a critical offered traffic viz. for N=6 beyond 2 Erlang.

The Figure 3 reveals better throughput is available if the delay is varied for different passes instead of keeping it constant for all passes. Basically if the delay is increased in every recirculation by a certain amount then it requires less number of recirculation comparing the fixed delay case to achieve a same particular amount of delay. As we have already discussed that recirculation of optical signal in the fiber delay line causes attenuation of signal power, insertion of different noises which ultimately affects the throughput of the network so it is better to have less number of recirculation to achieve better output. It may also be inferred from Figure 3 that the region of offered traffic for which the throughput is very high or the length of the high throughput region is greater in case of fixed delay network comparing to the variable delay system.

The Figure 4 depicts that, as the holding time increases the throughput decreases for all types of delay systems. Holding time corresponds to the processing speed and it increases for slower processing speed. Delay line will provide an amount of delay to the signals which are in the queue of getting served. Fast servicing will provide lesser processing time which in turn reduces the number

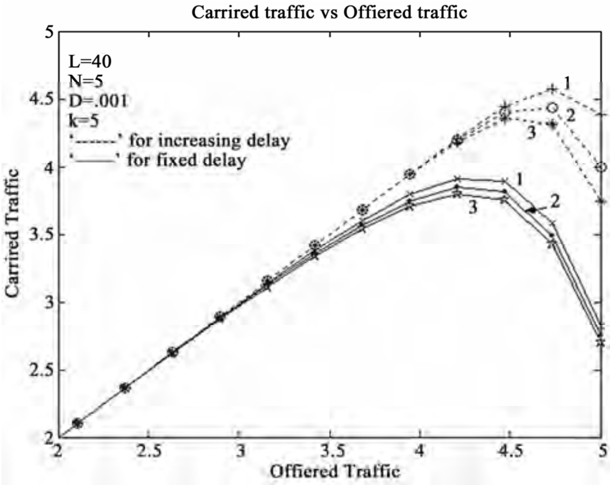


Figure 3. Plot of carried traffic vs offered traffic for different values of N.

of recirculation in the delay loop. From Figure 4, it is also seen that the spreading of the linear region is greater in case of fixed delay loop comparing to the variable delay loop.

Figure 5 shows the variation of throughput for different numbers of recirculation loop. It is observed that as the number of recirculation loop increases the amount of carried traffic increases for both types of delay system i.e, for fixed delay as well as for increasing delay system. This is because the lesser loop increases the holding probability of the packet in the delay line. This in turn reduces the packet dropping probability. Throughput improvement is significant increasing in case of increasing delay system; this is obvious because the amount of delay achieved during recirculation increases gradually. It may be noted that the amount of region with maximum throughput is available in case of fixed delay system.

The analysis has been made more general by including a geometrical progressing delay loop in addition to arithmetic and constant delay lines. The corresponding throughputs have been presented in Figure 6 The fig reveals that the throughput improves as the delay increases which is expected but the increment of throughput will sustain up to a certain value of incoming traffic, after which the output decreases, means the packets which are coming further are being completely rejected.

From Figure 6 we can also infer that the insertion of more delay in the loop will increase the cost and complexity of the system as well and it is tolerable up to a certain limit. Thus this investigation will help the network designer to take a decision on the possible maximum throughput and the complexity of node architecture design.

4. Conclusions

The problem of wavelength contention in packet switched WDM networks using recirculation optical delay lines has been developed. The proposal is based on putting priority to the packets which have suffered maximum delay on the link in processing by using a proposed contention resolution algorithm. The performance of the algorithm has been evaluated using MATLAB simulation to establish a better contention resolution using a varied delay lines at the nodes. The analysis presented here is useful to predict the traffic throughput range of a processing node with given number of FDL's and relevant design parameters.

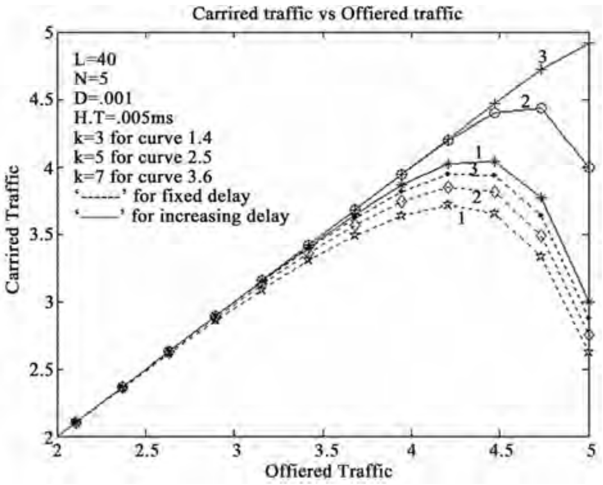


Figure4. Carried traffic vs offered traffic for different values of holding time.

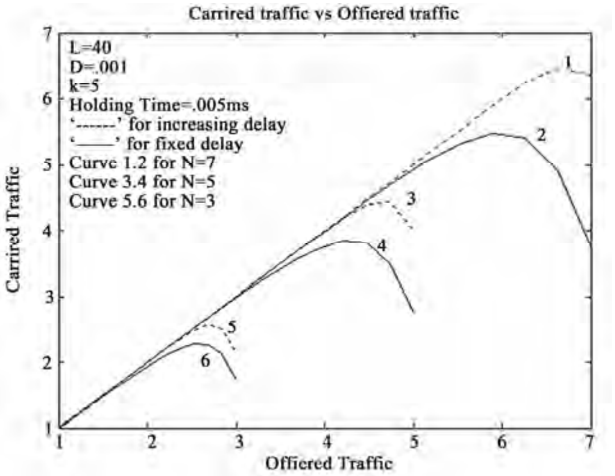


Figure 5. Plot of carried traffic vs offered traffic for different values of k.

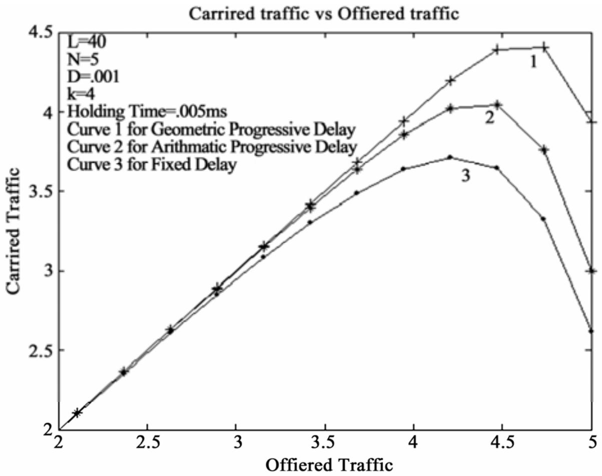


Figure 6. Plot of carried traffic vs offered traffic for different types of delay.

5. References

[1] L. Xu, H. G. Perros, and G. Rouskas, "Techniques for optical packet switching and optical burst switching," IEEE Comm. Magazine, pp. 136–142, 2001.

[2] S. L. Danielsen, *et al.*, "Analysis of a WDM packet switch with improved performance under bursty traffic

conditions due to tunable wavelength converters," J. Lightwave Technology, Vol. 16, No. 5, pp. 729–735, 1998.

[3] G. Shen, *et al.*, "Performance study on a WDM packet switch with limited-range wavelength converters," IEEE Comm. Letters, Vol. 5, No. 10, pp. 432–434, 2001.

[4] Y. Yang, J. Wang, and C. Qiao, "Nonblocking WDM multicast switching networks," IEEE Trans. Parallel and Distributed Systems, Vol. 11, No. 12, pp. 1274–1287, Dec. 2000.

[5] D. K. Hunter and I. Andronovic, "Approaches to optical internet packet switching," IEEE Comm. Magazine, Vol. 38, No. 9, pp. 116–122, 2000.

[6] C. Develder, M. Pickavet, and P. Demeester, "Assessment of packet loss for an optical packet router with recirculating buffer," Proc. Conf. Optical Network Design and Modeling, pp. 247–261, 2002.

[7] W. Lin , R. S. Wolff and B. Mumey, "A markov-based reservation algorithm for wavelength assignment in all-optical netwrks," Journal of Lightwave Technology, Vol. 25, No. 7, pp. 1676–1683, July 2007.

[8] V. K. Chaubey and K. Divakar, "Modeling and simulation of WDM optical networks under traffic control protocols," Optical Fiber Technology (Elsevier), Vol. 15, pp. 95–99, Jan. 2009.

[9] J. Li , C. Qiao, J. Xu and D. Xu, "Maximizing throughput for optical burst switching networks," IEEE/ACM Transactions on Networking, Vol. 15, Issue 5, pp. 1163–1176, October 2007.

[10] R. Kundu, V. K. Chaubey, "Analysis of optical WDM network topologies with application of LRWC under symmetric erlang - C traffic," Novel Algorithms and Techniques in Telecommunications, Automation and Industrial Electronics, pp. 468–473, Aug. 2008.

[11] R. Ramamurthy and B. Mukherjee, "Fixed-alternate routing and wavelength conversion in wavelength-routed optical networks," IEEE/ACM Transactions on Networking, Vol. 10, No. 3, pp. 351–367, June 2002.

[12] Y. Zhou and G. S. Poo, "Multicast wavelength assignment for sparse wavelength conversion in WDM networks," INFOCOM'06, 25th IEEE International Conference on Computer Communications Proceedings, pp. 1–10, April 2006.

[13] P. RajlakshmiN and A. Jhunjhunwala, "Anlytical tool to achieve wavelength conversion performance in no wavelength conversion optical WDM networks," IEEE International Conference on Communications, ICC'07. pp. 2436–2441, 24–28 June 2007.

[14] A. O. Allen, "Probability, statistics, and queueing theory with computer science applications," Academic Press INC, USA, 1990.

Int. J. Communications, Network and System Sciences, **2009, 7, 657-663**
doi:10.4236/ijcns.2009.27075 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

Performance Evaluation of Signal Strength Based Handover Algorithms

Sanjay Dhar ROY
Affiliation1 National Institute of Technology, Durgapur, India
Email: *s_dharroy@yahoo.com*

Received November 27, 2008; *revised March* 21, 2009; *accepted June* 7, 2009

ABSTRACT

Performance evaluation of handover algorithms has been studied for mobile cellular network. Effects of averaging, hysteresis margin and shadow fading are investigated for different handoff algorithms. Probability of outage, handover delay and average number of handovers are considered as performance metrics. Different handover algorithms considered here are based on relative signal strength with hysteresis, relative signal strength with hysteresis and threshold, absolute signal strength and combined relative and absolute signal strength. Both analytical and simulation methods have been used in this paper. This study is important as performance analyses of cellular system, in presence of handoff, will be important for future generation wireless networks, for example, WiMAX, UMTS.

Keywords: Handoff, Algorithm, Averaging, Outage, Handoff Delay

1. Introduction

Signal strength at a Mobile Station (MS) depends upon path loss, shadow fading and multipath fading. Path loss depends on the distance of MS from a Base Transceiver Station (BTS). It increases with the distance from BTS. Between BTS and MS, there are many obstacles e.g., trees, buildings, vehicles. Those obstacles create variation of signal strength over the mean path loss. This variation is known as shadow fading which follows log normal distribution i.e. standard deviation of shadow fading(σ) in dB follows normal or Gaussian distribution [3]. MS receives line of sight (LOS) signal from BTS and signals reflected from different places. Those multipath components result multipath fading. Multipath fading is found to follow Rayleigh distribution [1]. Signal averaging can filter out multipath fading. When MS moves from one BTS to another, on the way signal from current BTS get reduced whereas signal from other BTS increased. So, MS should be served by the new BTS when signal from serving BTS reduced below a specified level. This process of transferring control of MS from one BTS to another BTS without interruption of service is known as Handover. Handover or handoff is mainly of two types, hard handoff or soft handoff. Hard handoff is also referred to as “Break before Make connection”. MS is connected to only one BTS at a time. Soft handoff refers to as “Make before Break connection”. MS may be in connection with more than one BTS at a time. We have investigated performance evaluation for hard handoff case. Handover may also be classified as horizontal and vertical handover. Horizontal handoff takes place when MS moves from one cell to another cell of the same system e.g., GSM. Vertical handoff takes place when MS moves from one cell of a system to another cell of a different system e.g., GSM and WLAN. Handover process can be divided into mainly Initiation and Execution phase. In initiation phase based on some criteria viz., Received Signal Strength indicator (RSSI), BER, SIR, distance, velocity, it is checked if MS receives signal from BTS other than serving BTS then QoS will be better or not. Ideally, handover should depend on path loss and to some extent on shadow fading. To make handover decision independent of Raleigh fading, both uplink and downlink measurements are taken over a interval of 480 milliseconds time (sampling time) for averaging of fast fading effects in case of GSM. In practice, diversity techniques such as frequency hopping, antenna diversity and signal processing such as convolution coding, equalizers are used to handle Rayleigh fading. Long term shadow fading is compensated by increasing power budget margin increasing transmit power and co-channel reuse distance. If handover does not take place at right time then an ongoing call may be dropped. To prevent call drop before handoff due to unavailability of channel, several handoff prioritization scheme are proposed e.g., Guard channel scheme, Queuing of handoff [4]. In execution phase, once the need of handoff is detected, MS receives new channel in association with Base station controller (BSC) and Mobile switching center (MSC). Several Handover analyzes have been made so far [1,2, 5,9]. Handoff in cellular systems was summarized in [4]. Description about macro cell, micro cell, corner effects were also provided in [4]. Vijayan *et al*. provides a framework considering level crossing analysis for performance evaluation of handoff algorithms [1]. Effects of correlation for shadow fading were investigated based on measurements [6]. A closed form expression for handoff rate was proposed in [8]. Handover initiation can be based on various approaches viz., relative signal strength, relative signal strength with threshold, relative signal strength with hysteresis, relative signal strength with hysteresis and threshold, prediction technique, distance, velocity, combined relative and absolute signal strength [7]. Our approach considers relative signal strength with hysteresis and absolute signal strength. Performance of handover algorithm can be determined based on criteria viz., number of unnecessary handoffs, probability of outage, average number of handoffs, handover delay, and probability of blocking [7]. We have analyzed the performance of the algorithm based on probability of outage, handover delay, and number of handovers. Effects of shadow fading, averaging interval, hysteresis margin (h) are considered on these parameters. Singh *et al.* [10] suggested that h = e × σ where e =1.3 – 1.6 and over the coverage area, h must be dynamically adjusted as a function of σ. Number of handovers is traded off against handover delay (HO delay) in several papers [1,9]. Simulation study [2] is performed to find the effect of type and length of averaging window on handover performance. In all cases number of handovers should be small as it would reduce switching load of MSC.

Organization of this paper: In Section 2, system model is presented. Then probability of outage, P_{out} and probability of handover or assignment, P_{assn} are calculated using simple analytical model considering handover based on absolute signal strength measurements. Effect of shadow fading on P_{out} is also analyzed. Section 3 describes simulation model to obtain handover delay and number of handovers considering averaging of signal strength and other parameters. In Section 4, numerical results are presented. Finally in Section 5, conclusion is stated.

2. System Model

Two base stations, BTS1 and BTS2 are separated by D meters [1,10]. Mobile station (MS) is moving from BTS1 to BTS2 with constant speed. The signal level received from two BTSs (in dB) at a distance, d from BTS1 can be expressed as follows:

$$P_{rx1}(d) = K_1 - K_2 \log_{10}(d) + x_1(d) \quad d \in (0,D) \text{ meter.} \quad \textbf{(1)}$$

$$P_{rx2}(d) = K_1 - K_2 \log_{10}(D-d) + x_2(d) \quad \textbf{(2)}$$

$P_{rx1}(d)$ and $P_{rx2}(d)$ are received signal from BTS1 and BTS2 respectively at a distance d meters from BTS1. Rayleigh fading is neglected since it has shorter correlation distance compared to shadow fading. K_1 and K_2 are due to path losses. K_2 is actually 10n, where n is path loss component. We assume that $K_1 = 0$ and $K_2 = 30$. $x_1(d)$ and $x_2(d)$ are two independent zero mean stationary Gaussian processes. Hence received power from BTSs may also be considered to be Gaussian processes with mean, $\mu_1 = K_1 - K_2 \log(d)$ and $\mu_2 = K_1 - K_2 \log(D-d)$ respectively. $x_1(d)$ and $x_2(d)$ are assumed to have exponential correlation proposed by Gudmundson[6] based on experimental results. That is, $E\{x_1(d_1)\, x_1(d_2)\} = E\{x_2(d_1)\, x_2(d_2)\} = \sigma^2 \exp(-d_s/d_0)$. Where d_0 is correlation distance which determines the decaying factor for correlation.

2.1. Handover Algorithm for Absolute Signal Strength Method

When received signal from BTS1 is less than a specified value and at the same time received signal from BTS2 is more than minimum value of received signal for continuation of a call then handover (HO) will take place from BTS1 to BTS2. Similarly condition for handover from BTS2 to BTS1 can be stated as follows.

$P_{rx1}(d) < P_{rho}$ and $P_{rx2}(d) > P_{rmin}$: HO: BTS1→BTS2

$P_{rx2}(d) < P_{rho}$ and $P_{rx1}(d) > P_{rmin}$: HO: BTS2→BTS1

where P_{rho} = Absolute value of received power from any BTS after which handover should take place. P_{rmin} = Minimum value of received power for which call is possible. If signal strength becomes less than P_{rmin} then there will be call drop for ongoing call and new call will not be possible.

At a distance, d from BTS1 if received signal strengths from both BTSs go below P_{rmin} then call will not be possible i.e, there will be outage. Probability of outage,

$$P_{out} = prob\left(P_{rx1}(d) \le P_{r\min} and P_{rx2}(d) \le P_{r\min}\right)$$

$$P_{out} = \Pr ob\left(P_{rx1}(d) \le P_{r\min}\right) and \Pr ob\left(P_{rx2}(d) \le P_{r\min}\right)$$

(Since these two events are statistically independent)

$$P_{out} = Q\left(\frac{(\mu_1 - P_{r\min})}{\sigma}\right) \times Q\left(\frac{(\mu_2 - P_{r\min})}{\sigma}\right) \quad \textbf{(3)}$$

$Q(x)$ is Q-function. $P(X \ge x) = Q(x)$ for $X \sim N(0,1)$

If $X \sim N(\mu,\sigma)$ then $P(X \ge x) = Q\left(\frac{(x-\mu)}{\sigma}\right)$ and

$$P(X \le x) = Q\left(\frac{(\mu - x)}{\sigma}\right) = 1 - Q\left(\frac{(x-\mu)}{\sigma}\right).$$

Mean of received powers are distance dependent. Using a computer program, varying d, we have plotted Figure 1. Keeping distance fixed, varying σ, we have plotted Figure 2 and Figure 3. When received signal from serving BTS will be less than P_{rho} then there will be handover to other BTS. Current BTS should be able to serve the MS i.e., received power from it should be more than P_{rmin}. So probability of assignment (or handover) to any BTS can be obtained as follows: Probability of assignment to BTS1,

$$P_{assn1} = prob(P_{rx1}(d) \leq P_{rho} and P_{rx2}(d) \geq P_{r\min})$$

(Since these two events are statistically independent)

$$P_{assnl} = Q\left(\frac{(\mu_1 - P_{rho})}{\sigma}\right) \times Q\left(\frac{(P_{r\min} - \mu_2)}{\sigma}\right) \quad (4)$$

Similarly, probability of assignment (or handover) to BTS2 can be obtained as follows:

$$P_{assn2} = Q\left(\frac{(\mu_2 - P_{rho})}{\sigma}\right) \times Q\left(\frac{(P_{r\min} - \mu_1)}{\sigma}\right) \quad (5)$$

Using the above equation we have plotted Figure 4. It is noticed that at 1000 meter Probability of handover is maximum, where MS can be assigned to any BTS. Shadow fading effect will be maximum there (Figure 3).

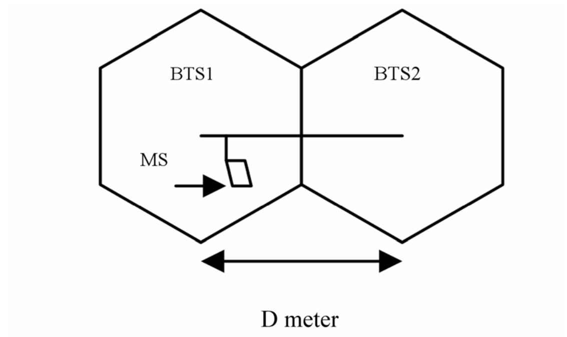


Figure 1. System model.

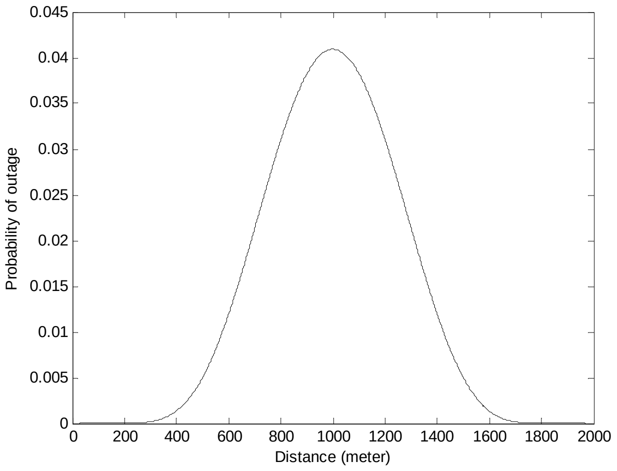


Figure 2. Probability of outage vs. distance.

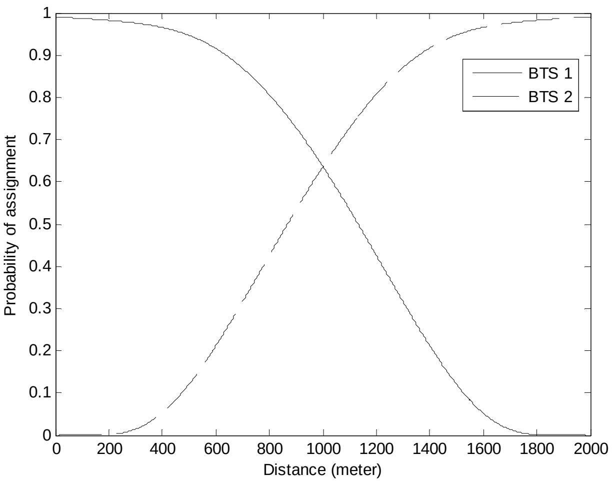


Figure 3. Probability of assignment to a BTS vs. distance.

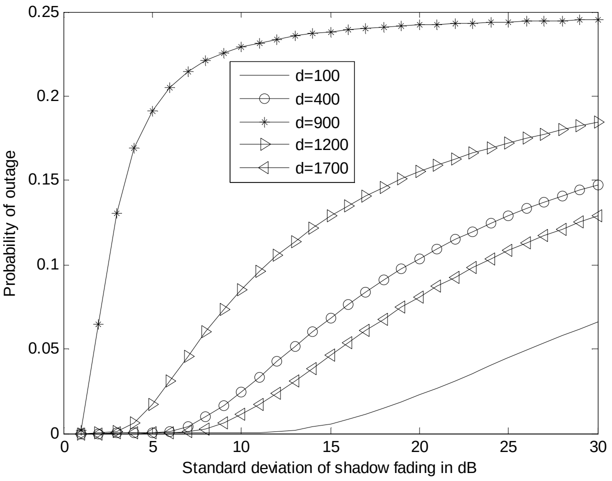


Figure 4. Probability of outage vs. standard deviation of shadow fading.

3. Simulation Model

Received signal strength is sampled at discrete time instants, $t_i = kt_s$. t_s is sampling time. And corresponding sampling interval in distance is $d_s = vt_s$. Here v is constant velocity of the mobile. t_s is 480ms (nearly equal to 0.5 sec) in case of GSM. We assume v = 2 m/sec so that d_s is 1 meter. Received signal strengths from both BTSs are averaged using exponential averaging window. Received signal strengths from BTSs sampled at kd_s distance corresponding to t_i are respectively as follows:

$$P_{rx1}(kd_s) = K_1 - K_2 \log_{10}(kd_s) + x_1(kd_s) \quad (6)$$

$$P_{rx2}(kd_s) = K_1 - K_2 \log_{10}(D - kd_s) + x_2(kd_s) \quad (7)$$

These received signal strengths are averaged using discrete time counterpart of exponential averaging window [8] as shown below. To generate the shadow fading component, correlation of shadow fading is considered.

Using recursive relations fading components have been generated and the same have been used for simulation purpose.

$$P_{rx1,avg}(k) = e^{-\left(\frac{d_s}{d_{av}}\right)} P_{rx1,avg}(k-1) + \left(1 - e^{-\left(\frac{d_s}{d_{av}}\right)}\right) P_{rx1}(k) \quad \textbf{(8)}$$

$$P_{rx2,avg}(k) = e^{-\left(\frac{d_s}{d_{av}}\right)} P_{rx2,avg}(k-1) + \left(1 - e^{-\left(\frac{d_s}{d_{av}}\right)}\right) P_{rx2}(k) \quad \textbf{(9)}$$

where d_{av} is length of averaging window and h is hysteresis margin. $P_{rx1,avg}(k)$ is the averaged received signal from BTS1 at kd_s distance. $P_{rx1,avg}(k-1)$ is the averaged received signal from BTS1 at $(k-1)d_s$ distance. $P_{rx2,avg}(k)$ is the averaged received signal from BTS2 at kd_s distance.

3.1. Handover Algorithm for Relative Signal Strength with Hysteresis

If received signal from BTS1 is less than received signal from BTS2 by a margin, h then handover will be there from BTS1 to BTS2. Similarly, if received signal from BTS2 is less than received signal from BTS1 by a margin, h then handover will be there from BTS2 to BTS1. We can express these using following simple relations.

$[P_{rx2,avg}(k) - P_{rx1,avg}(k)] > +h$ HO: BTS1→ BTS2

$[P_{rx2,avg}(k) - P_{rx1,avg}(k)] < -h$ HO: BTS2→ BTS1

3.2. Handover Algorithm for Relative Signal Strength with Threshold

If received signal from BTS1 is less than received signal from BTS2 by a margin, h and received signal from BTS2 is greater than a threshold value then handover will be there from BTS1 to BTS2. Similarly, if received signal from BTS2 is less than received signal from BTS1 by a margin, h and received signal from BTS1 is greater than a threshold value then handover will be there from BTS2 to BTS1. We can express these using following simple relations. For simplicity, we consider threshold value equal to P_{rmin}.

$[P_{rx2,avg}(k) - P_{rx1,avg}(k)] > +h$ and $[P_{rx2,avg}(k) > P_{rmin}]$

HO: BTS1→ BTS2

$[P_{rx1,avg}(k) - P_{rx2,avg}(k)] > +h$ and $[P_{rx1,avg}(k) > P_{rmin}]$

HO: BTS2→ BTS1

3.3. Handover Algorithm for Combined Relative and Absolute Signal Strength Method

If received signal from BTS1 is less than received signal BTS2 by a margin, h and received signal from BTS1 is less than a threshold value (P_{rho}) then handover will be there from BTS1 to BTS2. Similarly, if received signal from BTS2 is less than received signal BTS1 by a margin, h and received signal from BTS2 is less than a threshold value then handover will be there from BTS2 to BTS1. We can express these using the following simple relations.

$[P_{rx2,avg}(d) - P_{rx1,avg}(d)] > +h$ and $[P_{rx1,avg}(d) < P_{rho}]$:

HO: BTS1→BTS2

$[P_{rx1,avg}(d) - P_{rx2,avg}(d)] > +h$ and $[P_{rx2,avg}(d) < P_{rho}]$:

HO: BTS2→BTS1

Using these algorithms after large number of iterations average number of handovers (N_{ho}), handover delays are calculated and plotted against hysteresis margin, standard deviation of shadow fading component.

4. Numerical Results

Following values are chosen for the analysis purpose:

1) Standard deviation of shadow fading, σ = 6 dB
2) Distance between BTSs, D = 2000 meter.
3) Correlation distance, d_0 = 20 meter.
4) Length of averaginvg window, d_{av} =10 meter and 20 meter.
5) Velocity of mobile station, v = 2 meter/sec.
6) Sampling time, t_s = 0.5 sec
7) Sampling distance, $d_s = vt_s$ = 1 meter.
8) P_{rmin} = -95 dB 9. P_{rho} = -85 dB

Analytical results are shown in Figure 2 to Figure 4. In Figure 2, P_{out} is plotted against distance from BTS1. P_{out} is large near the boundary of the cells and it is zero near to any of the BTSs. Figure 3 illustrates where handoff will take place and corresponding assignment probability is shown in this figure.

Figure 4 shows effects of shadow fading on probability of outage are shown. Figure 4 considers d = 100, 400, 1700 i.e, very near to either of BTSs. Naturally probability of outage is very less, almost zero for small σ, but for large values of σ, outage is possible for the specified values. Figure 4 also considers distance near the cell boundary (d = 900, 1200) where the signal from either of BTSs is very low, so we see that P_{out} largely depends on σ. Near to the cell boundary, probability is very large.

Figure 5 to Figure 14 shows simulation results. Average number of handoffs and handover delay are plotted against h for different values of d_{av}. We consider delay or handover delay to be the distance where first handover occurs. Actually, handover delay is total of averaging delay and hysteresis delay. Hysteresis time: it is the time needed when MS moves some distance away after measurements. It can be noticed from the figures Figure 5 to Figure 7 that handover delay increases with increasing h. And handover delay increases when averaging distance

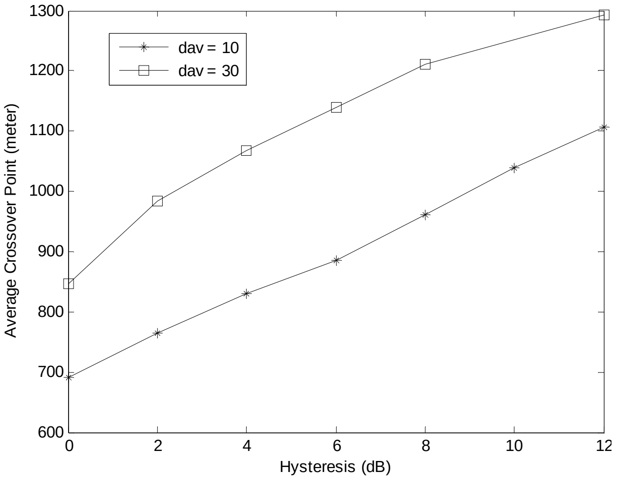


Figure 5. Handover delay vs. hysteresis margin for relative signal strength with hysteresis.

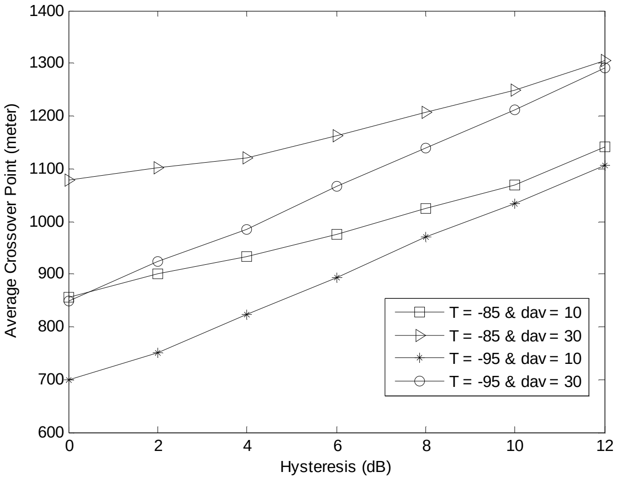


Figure 6. Handover delay vs. hysteresis margin for relative signal strength with threshold and hysteresis.

is increased. Since more averaging consumes more time. There is no significant change in handover delay (with d_{av} =30) with respect to σ after considering its correlation. That is due to averaging of signal strength. Averaging of signals filter out multipath component and to some extent shadow fading variation. For this reason delay vs. σ plot is not shown. Figure 5 shows variation of handoff delay for relative RSS for d_{av} =10 and 30. Handoff occurs near to cell boundary for d_{av} = 10 and h = 12. Hystersis value is to be very large for avoiding ping-pong for this algorithm. Figure 6 shows variation of crossover point for relative signal strength with hystersis and threshold. Four curves have been plotted for different threshold and d_{av} values. Handover delay will not change with T for large values of h and d_{av}. For example, at h =12 and d_{av} =30, handoff delays for T = -85 and -95 are same.

Figure 7 shows handover delay for combined relative and absolute signal strength based algorithm (CSS). Handover delay is more for T = -95 dB as this allows MS to be connected to the serving BTS for long compared to that with T = -85 dB. Crossover point is more with d_{av} = 30 than with d_{av} = 10 meter because of large averaging.

It is observed that numbers of handoffs are less for large values of hysteresis. For d_{av} =30, number of handoffs is almost equal to one i.e, no unnecessary handovers for the specified values. For d_{av} =10, number of handoffs is large. That means less averaging may lead to unnecessary handoffs. Figure 8 shows number of handoff vs. hysteresis for relative signal strength basesd algorithm. Number of handoff is less for d_{av} = 30 meter due to large averaging window length. Figure 9 shows variation of N_{ho} with h for CSS algorithm for different values of T and d_{av}. Number of handoff is lowest for T = -95 dB and d_{av} = 30. This happens because the current BTS keeps control of MS for longer time. Figure 10 shows variation

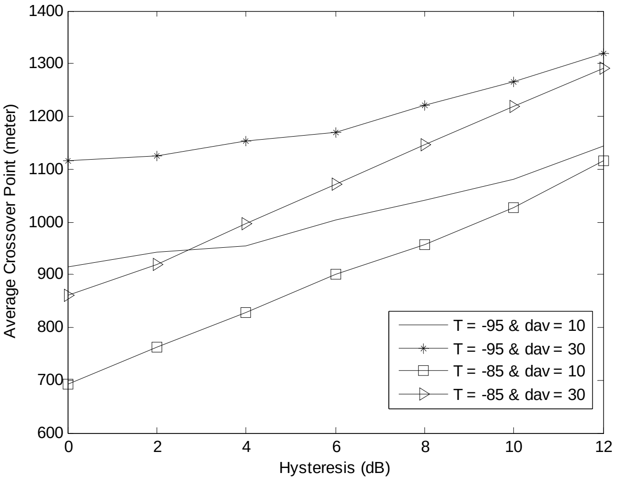


Figure 7. Handover delay vs. hysteresis margin for combined relative & absolute signal strength with hysteresis method.

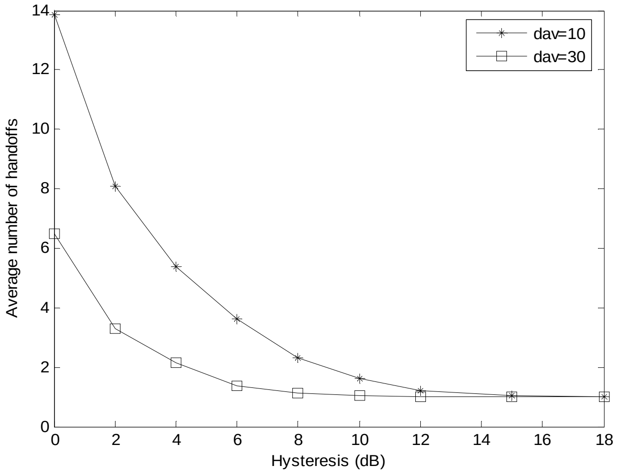


Figure 8. Average number of handoff vs. hysteresis margin for relative RSS.

of N_{ho} with h for relative signal strength with threshold and hysteresis based algorithm for different values of T and d_{av}. Number of handoff is lowest for T = -85 dB and d_{av} = 30. This happens because the current BTS keeps control of MS for longer time and control is transferred to candidate BTS only after it provides very good signal strength to sustain good quality and avoid ping pong. We have analyzed three different handover algorithms. Results suggest that effect of σ, h and d_{av} similar for all different algorithms.

Next three figures show tradeoff curves for all three different handoff algorithms considered in this paper. Fig 11 show tradeoff for relative signal strength based handoff algorithm. Tradeoff curve provides an idea for choosing handover design parameter. Handoff parameter may be chosen for point where N_{ho} and Cross over point both are low. Figure 12 shows tradeoff for CSS algorithm for different values of d_{av} and T. N_{ho} is one when cross over point is almost 1100 meter. Hence CSS algorithm provides very good balance between two conflicting parameters N_{ho} and handoff delay. Figure 12 shows tradeoff for relative signal strength with threshold and hysteresis based algorithm for different values of d_{av} and T. N_{ho} is one when cross over point is almost 1200 meter. Finally, Figure 14 shows tradeoff curve for all three different algorithms. Two algorithms other than relative signal strength based handoff algorithm, provides almost same performance with proper choice of hysteresis value. Crossover point can be around 1200 meters with proper setting of hystersis value for N_{ho} = 1 (one).

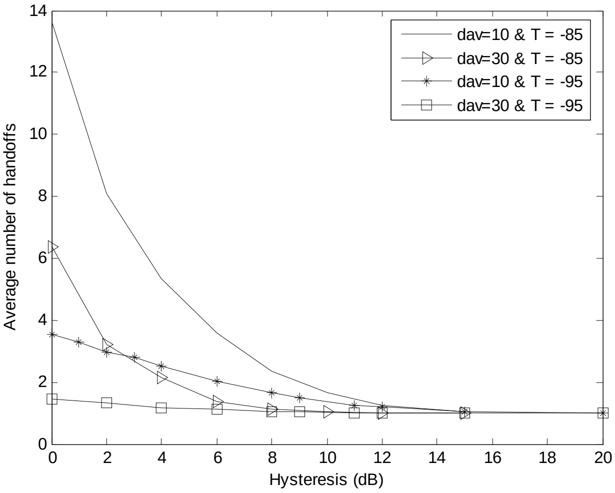


Figure 9. Average number of handoff vs. hysteresis margin for CSS.

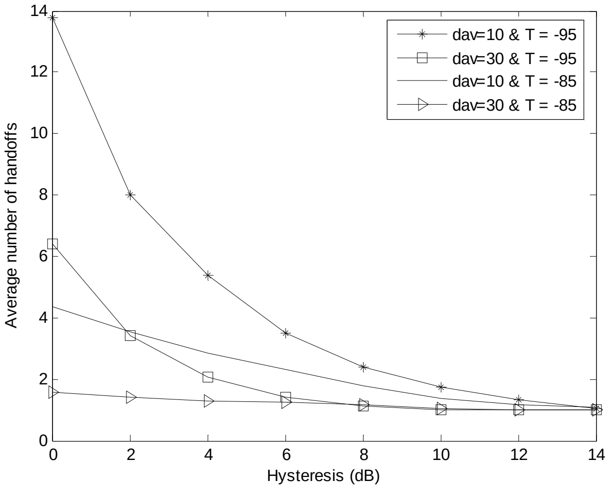


Figure 10. Average number of handoff vs. hysteresis margin for relative signal strength with threshold and hysteresis.

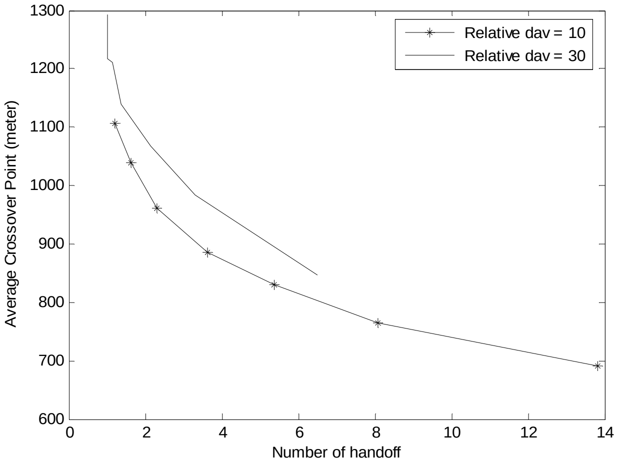


Figure 11. Tradeoff for relative RSS based algorithm.

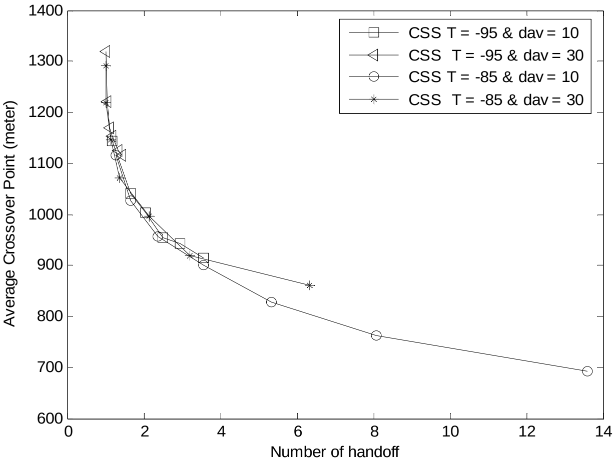


Figure 12. Tradeoff for relative CSS based algorithm.

5. Conclusions

This paper presents very simple method to choose handover design parameters (e.g., averaging window length, hysteresis margin, standard deviation of shadow fading) for Mobile Cellular system. It uses analytical method for finding probability of outage and it uses simulation method for finding handover delay and average number of handoffs. Analysis and simulation results are obtained for three different algorithms. Absolute signal strength based algorithm has to be considered for intersystem handoff, as relative measurements are not possible for different cellular systems because of their different power requirement and other criteria. If candidate signal strength is not large enough then there may be ping-pong effect. Relative signal strength with hysteresis and threshold takes care of this. If serving BS strength is enough to provide good quality of service then a handoff to candidate BS may be considered as unnecessary. Hence, a combined absolute and relative signal strength based handoff algorithm can take care of this problem. Both of these two algorithms prevent unnecessary handoff by increasing handoff delay to some extent. Handoff decision criteria can be critical when cell splitted (microcellular) to increase capacity and decrease power requirements of MS. When there is very small hysteresis margin or no hysteresis there may be ping-pong effect. Due to dynamic behaviour of propagation environment MS very close to BTS may be in deep fade for very short duration (e.g., street corner effect). Handover should not occur in such cases. To avoid this, averaging of signal strength may be done over short duration while keeping large hysteresis margin. Overlay macro cell may also be employed to overcome this problem. For microcellular systems short averaging time and large hystersis margin is more reliable and reverse for macro cellular systems. Probability of outage increases with increase in shadow fading. Since designer has almost no control over the shadow fading component, hysteresis margin can be dynamically varied to compensate the effect of shadow fading.

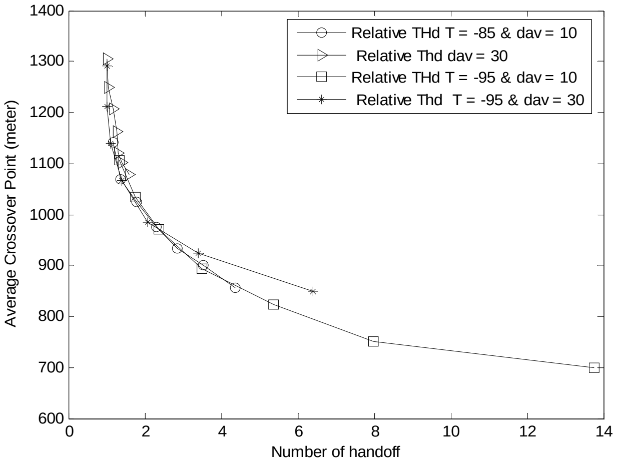


Figure 13. Tradeoff for relative RSS with threshold based algorithm.

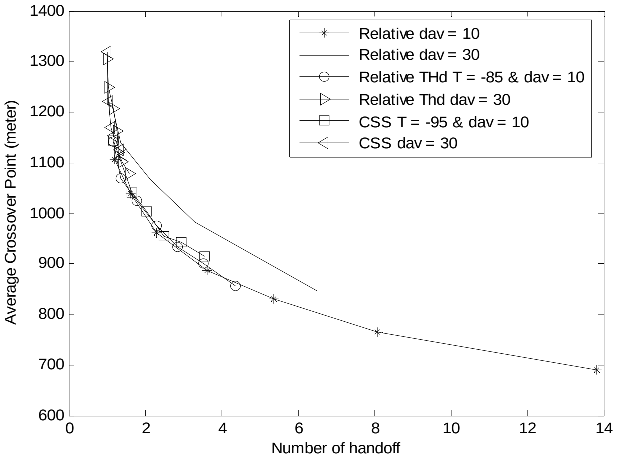


Figure 14. Tradeoff for all three algorithms together.

6. Acknowledgments

The author would like to thank Mr. A. Chandra, Dr. S. Kundu and Dr. S. K. Datta of NIT, Durgapur, India for their valuable suggestions.

7. References

[1] R. Vijayan and J. Holtzman, “A model for analyzing handoff algorithms,” IEEE Transactions on Vehicular Technology, Vol. 42, No. 3, pp. 351–356, August 1993.

[2] G. E. Corazza, *et al.*, “Characterization of handover initialization in cellular mobile radio networks,” IEEE VTC’ 94, pp. 1869–1872, 1994.

[3] T. S. Rappaport, “Wireless communications,” Prentice Hall, 1996.

[4] D. Nisith, *et al.*, “Handoff in cellular systems,” IEEE Personal Communications, Dec. 1998.

[5] M. Gudmundson, “Analysis of Handover algorithm”, IEEE VTC, May 1991, pp 537–42

[6] M. Gudmundson, “Correlation model for shadow fading in mobile radio systems,” Electronics Letters, Vol. 27, No. 23, pp. 2145–2146, November 1991.

[7] G. P. Pollini, “Trends in handover design,” IEEE Communication Magazine, Vol. 34, pp. 82–90, March 1996.

[8] G. P. Pollini, “Handover rate in cellular systems: Towards a closed form approximation,” Global Telecommunications Conference, GLOBE COM’97, IEEE, Vol. 1, pp. 711–715, November 1997.

[9] N. Zhang and J. M. Holtzman, “Analysis of handover algorithms using both absolute and relative measurements,” IEEE VTC’94, pp. 82–86, 1994.

[10] B. Singh, *et al.*, “Sensitivity analysis of handover performance to shadow fading in microcellular systems,” IEEE ICPWC, 2005.

[11] S. D. Roy, “Effects of averaging, shadow fading and hysteresis margin on handover performance,” CD proceedings, ICEMC, Pesit, Bangalore, Aug. 2007.

Int. J. Communications, Network and System Sciences, **2009, 7, 664-668**
doi:10.4236/ijcns.2009.27076 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

Space-Time-Frequency Coded for Multiband-OFDM Based on IEEE 802.15.3a WPAN

Kamal MOHAMED-POUR, Hossein EBRAHIMI
Department of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran
Email: *kmpour@kntu.ac.ir*

Received September 9, 2008; *revised February* 23, 2009; *accepted April* 15, 2009

ABSTRACT

In this paper, Multiband-OFDM UWB system based on IEEE 802.15.3a standard is studied and simulated with spatial, time and frequency (STF)coding scheme. The using of STF coding method can guarantee both full symbol rate and full diversity advantages. The simulation results show that the STF code uses multipath-rich and random-clustering characteristics of UWB channel environment on the performance of MB-OFDM system.

Keywords: IEEE 802.15.3a, UWB, MB-OFDM, MIMO, Space-Time-Frequency Coding

1. Introduction

Ultra wideband (UWB) systems are the first nomination for future wireless personal area networks (WPANs). The enormous band with availability provides the potential for very high data rates. The ultra wide bandwidth of UWB enables various WPAN applications such as high-speed wireless universal serial bus (WUSB) connectivity for personal computers and their accessories, high-quality real-time video and audio transmission, and cable replacement for home entertainment systems.

Currently, the multi-band orthogonal frequency division multiplexing (MB-OFDM) [1] is an important candidate for the physical layer within IEEE 802.15.3a standard.

On the other hand, the rich scattering multipath channel in UWB indoor environment provides an ideal transmission scenario for multiple antenna configurations. In this paper, we use a space-time-frequency coding (STFC) method [2] that can guarantee both full symbol rate and full diversity for performance improvement of MB-OFDM UWB systems over CM 1-CM 4 environment with 2ISO and 2I2O MIMO configurations.

The paper is organized as follows: In Section 2, an overview of STFC MB-OFDM UWB system is given. Section 3 gives the simulation results, and finally Section 4 concludes the paper.

2. STFC MB-OFDM UWB System

2.1. UWB Channel Model

The channel model is based on Saleh-Valenzuela model [3] according to IEEE 802.15.3a standard. The channel impulse response can be expressed as

$$h_i(t) = X_i \sum_{l=0}^{L_c-1} \sum_{k=0}^{K_c-1} \alpha_{k,l}^i \delta\left(t - T_l^i - \tau_{k,l}^i\right) \qquad \textbf{(1)}$$

where i represents the realization of the i-th impulse response, $\alpha_{k,l}^i$ is the multipath gain coefficients, T_l^i is the delay of the l-th cluster, $\tau_{k,l}^i$ is the delay of k-th ray, and X_i represents the log-normal shadowing. The cluster arrivals and the path arrivals within each cluster are modeled by Poisson processes

$$\begin{aligned} p\left(T_l \mid T_{l-1}\right) &= \Lambda \exp\left[-\Lambda\left(T_l - T_{l-1}\right)\right], \quad l > 0 \\ p\left(\tau_{k,l} \mid \tau_{(k-1),l}\right) &= \lambda \exp\left[-\lambda\left(\tau_{k,l} - \tau_{(k-1),l}\right)\right], \quad k > 0 \end{aligned} \qquad \textbf{(2)}$$

where Λ and λ (where λ >Λ) are the cluster arrival rate and ray arrival rate, respectively. Four set of channel model (CM) parameters for different measurement environments were defined, namely CM 1, CM 2, CM 3, and CM 4. Table 1 provides the model parameters of CM 1-CM 4 [4].

2.2. STFC MB-OFDM Structure

We consider a UWB multiband OFDM system with fast band-hopping rate that signal is transmitted on a frequency-band during one OFDM symbol interval, and then

Table 1. The IEEE UWB channel parameters.

Parameters	CM 1	CM 2	CM 3	CM 4
Condition	LOS 0-4m	NLOS 0-4m	NLOS 4-10m	NLOS 4-10m
Λ (1/nsec)	0.0233	0.4	0.0667	0.0667
λ (1/nsec)	2.5	0.5	2.1	2.1
cluster decay factor	7.1	5.5	14	24
ray decay factor	4.3	6.7	7.9	12
N_{path} (10 dB)	12.5	15.3	24.9	41.2
N_{path} (85%)	20.8	33.9	64.7	123.3

moved to a different frequency-band at the next interval. In Table 2 you can see the simulation parameters of MB-OFDM UWB system. The data is encoded by STF code words across M_t transmit antennas, N OFDM subcarriers, and K OFDM blocks. We suppose a frequency-selective fading channels based on S-V model, between any pair of transmit and receive antennas. Figure 1 represents the structure of system. Because of small wavelength in UWB environment and fast frequency hopping, consideration of independency between MIMO channel elements is reasonable. In this case, according to [2,5] the maximum achievable diversity is at most min { M_tM_r, $rank(R_T)$, KNM_r}, where L is the number of delay path and R_T is the temporal correlation matrix.

We use repetition coded STF code [2] that is a full diversity code as follows:

$$D_{STF} = \mathbf{1}_{K\times 1} \otimes D_{SF} \tag{3}$$

where $1_{K\times 1}$ is an all one matrix, $\otimes$ is tensor product, and D_{SF} is a full diversity SF code of size $N\times M_t$ which have been proposed in [6]. At the transmitter, the information is jointly encoded across M_t transmit antennas, M OFDM subcarriers, and K OFDM blocks. Each STF codeword is a $KN\times M_t$ matrix that can be expressed as a

$$\boldsymbol{D}_k = \left[\boldsymbol{G}_{k,1}^T \quad \boldsymbol{G}_{k,2}^T \quad \cdots \quad \boldsymbol{G}_{k,P}^T \quad \mathbf{0}_{(N-P\Upsilon M_t)\times M_t}^T\right] \tag{4}$$

where $P = \left\lfloor N / \Upsilon M_t \right\rfloor$ and Υ is an integer smaller than N, which determines the number of jointly encoded subcarriers. Also

$$\boldsymbol{G}_{k,P} = \left(\boldsymbol{I}_{KM_t} \otimes 1_{\Upsilon\times 1}\right)\begin{pmatrix} x_{p,1} & x_{p,2} \\ -x_{p,2}^{*} & x_{p,1}^{*} \end{pmatrix} \tag{5}$$

where $x_{p,k}s$ are selected from QPSK or BPSK constellations. As mentioned earlier, we use repetition STFC which is based on Alamouti's structure. After STFC encoder, we add some preambles and headers for channel estimation and frame and packet synchronization. The baseband OFDM signal to be transmitted by i-th transmit antenna at the k-th OFDM block can be expressed as [7]

$$x_i^k(t) = \sqrt{\frac{E}{M_t}}\sum_{n=0}^{N-1} d_i^k(n)\exp\left\{\left(j2\pi n\Delta f\right)\left(t-T_{CP}\right)\right\} \tag{6}$$

where $d_i^k(n)$ represents the complex symbol to be transmitted over n-th subcarrier by i-th transmit antenna during the k-th OFDM symbol period. Finally, after filtering, up conversion and band hopping, the trans-mitted signal over i-th antenna is

$$s_i(t) = \sum_{k=0}^{K-1}\mathrm{Re}\left\{x_i^k(t-kT_{SYM})\exp(j2\pi f_c^k t)\right\} \tag{7}$$

In the receiver, after frequency dehopping, down converting and filtering, we have received signal at in matrix form as [2]

$$\boldsymbol{Y} = \sqrt{\frac{E}{M_t}}\boldsymbol{DH} + \boldsymbol{Z} \tag{8}$$

where $\boldsymbol{D}$ is the STF coded data, $\boldsymbol{H}$ is the MIMO channel matrix, and $\boldsymbol{Z}$ is complex baseband noise. Because of channel estimation pilots, we can determine $\boldsymbol{H}$, so we have

$$\boldsymbol{W}\times\boldsymbol{Y} = \sqrt{\frac{E}{M_t}}\boldsymbol{WDH} + \boldsymbol{WZ} = \sqrt{\frac{E}{M_t}}\boldsymbol{D} + \boldsymbol{WZ} \tag{9}$$

where $\boldsymbol{W} = \left(\boldsymbol{H}^H\boldsymbol{H}\right)^{-1}\boldsymbol{H}^H$. The receiver exploits a maximum likelihood detector over received signal matrix.

$$\hat{\boldsymbol{D}} = arg\ min\left\{\left\|\boldsymbol{Y} - \sqrt{\frac{E}{M_t}}\boldsymbol{DH}\right\|^2\right\} \tag{10}$$

Therefore, the error probability will be as

$$P_e\,|_{\boldsymbol{H}} = Q\left(\sqrt{\frac{\rho}{2M_t}\sum_{j=1}^{M_r}\left\|\left(\mathsf{D}-\hat{\mathsf{D}}\right)\boldsymbol{H}\right\|^2}\right) \tag{11}$$

Table 2. The MB-OFDM UWB parameters.

Information rate	200 Mbps
Number of Subcarriers	128
Channel coding	5/8 rate convolutional
Constellation	QPSK/BPSK
Data tones	100
T_{FFT}	242.4 nsec
T_{CP}	60.6 nsec
T_{GI}	9.5 nsec
T_{SYM}	312.5 nsec
Decoder	Hard viterbi

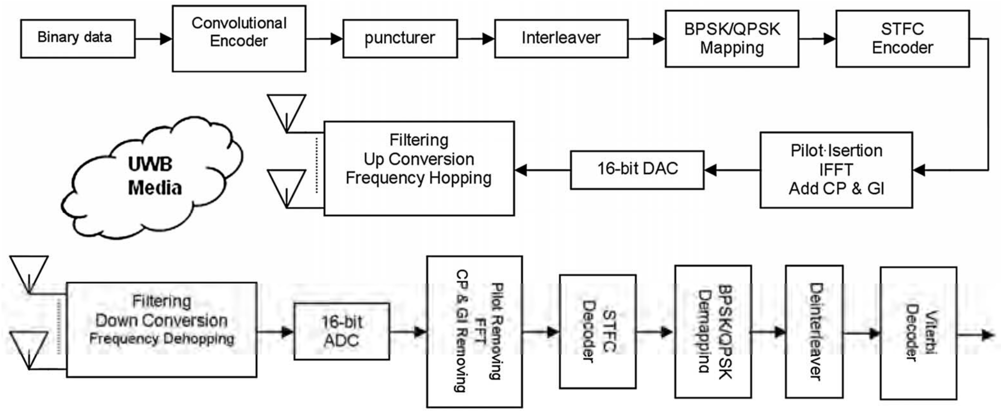
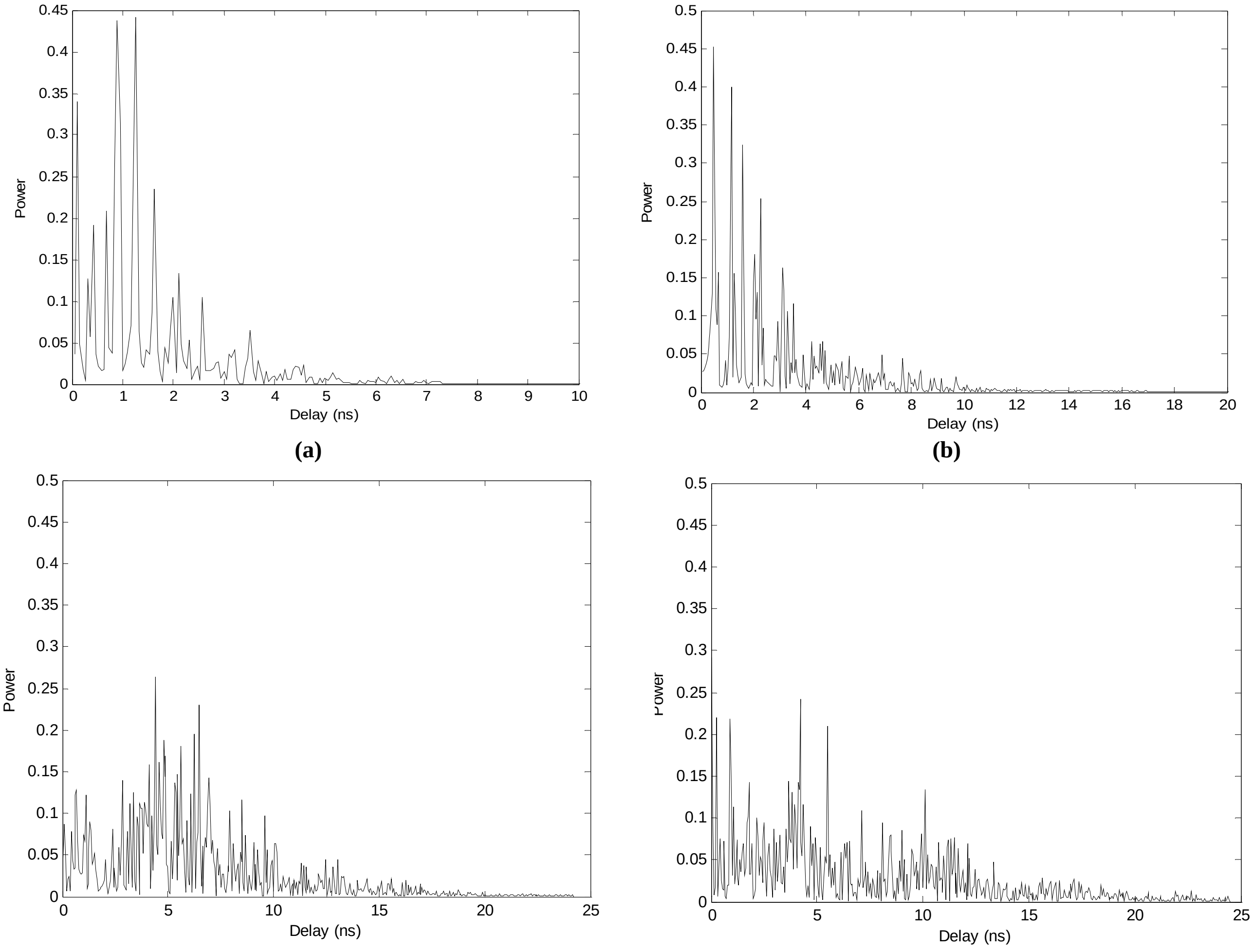


Figure 1. STFC MB-OFDM UWB system.



(a) **(b)**

(c) **(d)**

Figure 2. Simulated channel response; a) CM 1, b) CM 2, c) CM 3, d) CM 4.

3. Simulation Results

We performed simulations for a multiband UWB system with N = 128 subcarriers and the subband bandwidth of 528 MHz. Each OFDM symbol was of duration 242.42 *n*s. After adding the cyclic prefix of length 60.61 ns and the guard interval of length 9.47 ns, the symbol duration became 312.5 ns. Figure 2 gives the simulated channel impulse responses for CM 1-CM 4. We used a repetition STFC based on Alamouti's structure that can guarantee full diversity [2]. Also we used 5/8-rate convolutional coding for improving performance. So our pure data rate was 200 Mbps. We first simulated UWB channel based on IEEE 802.15.3a standard. CM 2 is 0-4 m, non line of sight channel, so it is reasonable to consider CM 2 for realistic application. Figure 3 gives the BER performance of MB-OFDM UWB as a function of SNR for CM 2 channel model, as frame length is 4200 QPSK symbols. In each frame, 600 symbols were preamble pilot for channel estimation. 100 channel realizations of IEEE 802.15.3a channel model (CM 1, 2, 3 and 4) were considered for the transmission of each symbol.

Figure 4 gives the BER performance of STFC coded MB-OFDM UWB as a function of SNR for CM 2 channel model without channel coding with the data jointly encoded across two subcarriers. The simulation results show that for CM 2 scenario, when K=1 the 2ISO and 2I2O configurations are almost 8.5dB and 16 dB better than MB-OFDM, respectively. For K=2, the 2ISO and 2I2O configurations are almost 11.5dB and 17.5 dB better than MB-OFDM, respectively.

Figure 5 gives the BER performance of STFC coded MB-OFDM UWB as a function of SNR for CM 2 and CM 4 for 2ISO and 2I2O configurations with the data jointly encoded across two subcarriers, when K=1. In conventional MB-OFDM, performance for CM 4 is worse than other scenarios, but it can be seen that the simulated performance for CM 4 is better than CM 2, when repetition STFC is employed. In coded system under CM 4 the coding gain is larger. It seems that space time frequency coding yields the MB-OFDM system can gain the multipath clustering property of UWB environments. In fact, when repetition STFC is employed, in comparison with other scenarios, CM 4 has the minimum correlation among OFDM subcarriers.

4. Conclusions

In this paper, MIMO-MB-OFDM has been studied. The simulation results indicate that the 2I2O STFC-MB-OFDM scheme for UWB system shows much better performance compared with un-coded MB-OFDM. On the other hand, the performance of STF coded system can be improved by increasing the number of antenna, regardless of the random clustering behavior of UWB channels.

5. References

[1] A. Batra, *et al.*, "Multi-band OFDM physical layer proposal for IEEE 802.15 task group 3a," IEEE 802.15–03/268r3, Texas Instruments, March 2004.

[2] W. Su, Z. Safar, and K. J. R. Liu, "Towards maximum achievable diversity in space, time and frequency: Per-

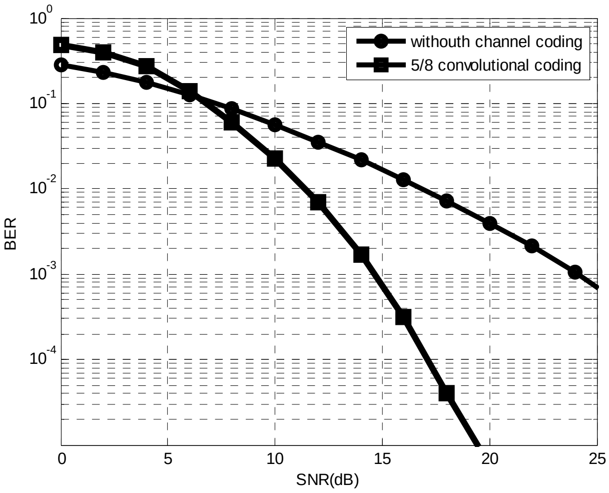


Figure 3. The performance of MB-OFDM UWB.

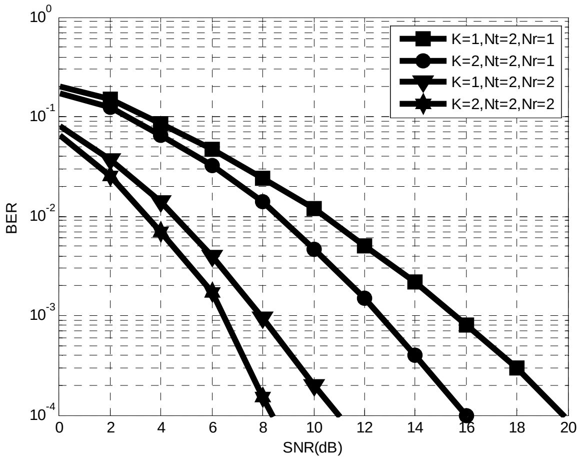


Figure 4. Performance of MB-OFDM for different MIMO configurations.

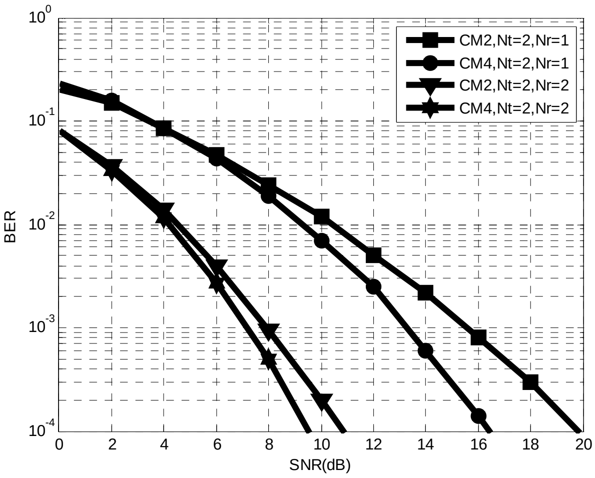


Figure 5. Performance comparison between CM 2 and CM 4.

formance analysis and code design," IEEE Trans. on Wireless Commun., Vol. 4, No. 4, pp. 1847–1857, Jul. 2005.

[3] IEEE 802.15WPAN High Rate Alternative PHY Task Group 3a (TG3a). Internet: www.ieee802.org/15/pub/TG3a.html

[4] M. Ghavami, L. B. Michael, and R. Kohno, "Ultra wideband signals and systems in communication engineering," John Wiley & Sons, Ltd, 2004.

[5] B. Lu, X. Wang, and K. R. Narayanan, "LDPC-based space-time coded OFDM systems over correlated fading channels: Performance analysis and receiver design," IEEE Trans. Commun., Vol. 50, No. 1, pp. 74–88, Jan. 2002.

[6] W. Su, Z. Safar, M. Olfat, and K. J. R. Liu, "Obtaining full-diversity space-frequency codes from space-time codes via mapping," IEEE Trans. Signal Processing, Vol. 51, pp. 2905–2916, Nov. 2003.

[7] W. P. Siriwongpairat, *et al*, " Multiband-OFDM MIMO coding framework for UWB communication systems," IEEE Trans. Signal Processing, Vol. 54, No. 1, Jan. 2006.

Int. J. Communications, Network and System Sciences, 2009, 7, 669-674
doi:10.4236/ijcns.2009.27077 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

Q-Learning-Based Adaptive Waveform Selection in Cognitive Radar

Bin WANG, Jinkuan WANG, Xin SONG, Fulai LIU
Northeastern University, Shenyang, China
Email: wangbin_neu@yahoo.com.cn

Received June 11, 2009; *revised August* 12, 2009; *accepted September* 20, 2009

ABSTRACT

Cognitive radar is a new framework of radar system proposed by Simon Haykin recently. Adaptive waveform selection is an important problem of intelligent transmitter in cognitive radar. In this paper, the problem of adaptive waveform selection is modeled as stochastic dynamic programming model. Then Q-learning is used to solve it. Q-learning can solve the problems that we do not know the explicit knowledge of state-transition probabilities. The simulation results demonstrate that this method approaches the optimal waveform selection scheme and has lower uncertainty of state estimation compared to fixed waveform. Finally, the whole paper is summarized.

Keywords: Waveform Selection; Q-Learning; Space Division; Cognitive Radar

1. Introduction

Radar is the name of an electronic system used for the detection and location of objects. Radar development was accelerated during World War Ⅱ. Since that time it has continued such that present-day systems are very sophisticated and advanced. Cognitive radar is an intelligent form of radar system proposed by Simon Haykin and it has many advantages [1]. However, cognitive radar is only an ideal framework of radar system, and there are many problems need to be solved.

Adaptive waveform selection is an important problem in cognitive radar, with the aim of selecting the optimal waveform and tracking targets with more accuracy according to different environment. In [2], it is shown that tracking errors are highly dependent on the waveforms used and in many situations tracking performance using a good heterogeneous waveform is improved by an order of magnitude when compared with a scheme using a homogeneous pulse with the same energy. In [3], an adaptive waveform selective probabilistic data association algorithm for tracking a single target in clutter is presented. The problem of waveform selection can be thought of as a sensor scheduling problem, as each possible waveform provides a different means of measuring the environment, and related works have been examined in [4,5]. In [6], radar waveform selection algorithms for tracking accelerating targets are considered. In [7], genetic algorithms are used to perform waveform selection utilizing the autocorrelation and ambiguity functions in the fitness evaluation. In [8], Incremental Pruning method is used to solve the problem of adaptive waveform selection for target detection. The problem of optimal adaptive waveform selection for target tracking is also presented in [9].

In this paper, the problem of adaptive waveform selection in cognitive radar is viewed as a problem of stochastic dynamic programming and Q-learning is used to solve it.

2. Division in Radar Beam Space

The most important parameters that a radar measures for a target are range, Doppler frequency, and two orthogonal space angles. However, in most circumstances, angle resolution can be considered independently from range and Doppler resolution. We may envision a radar resolution cell that contains a certain two-dimensional hypervolume that defines resolution.

Figure 1 is abridged general view of range and Doppler. Range resolution, denoted as ΔR, is a radar metric that describes its ability to detect targets in close proximity to each other as distinct objects. Radar systems are normally designed to operate between a minimum range $R_{\min}$, and maximum range $R_{\max}$. Targets seperated by at least ΔR will be completely resolved in range. Radars use Doppler frequency to extract target radial velocity (range rate), as well as to distinguish moving and stationary

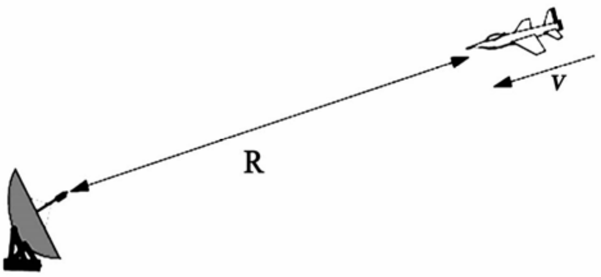


Figure 1. A closing target.

targets or objects such as clutter. The Doppler phenomenon describes the shift in the center frequency of an incident waveform.

In general, a waveform can be tailored to achieve either good Doppler or good range resolution, but not both simultaneously. So we need to consider the problem of adaptive waveform scheduling. The basic scheme for adaptive waveform scheduling is to define a cost function that describes the cost of observing a target in a particular location for each individual pulse and select the waveform that optimizes this function on a pulse by pulse basis.

We make no assumptions about the number of targets that may be present. We divide the area covered by a particular radar beam into a grid in range-Doppler space, with the cells in range indexed by t=1,…,N and those in Doppler indexed by v=1,…,M. There may be 0 target, 1 target or NM targets. So

$$C_{NM}^{0}+C_{NM}^{1}+C_{NM}^{2}+...+C_{NM}^{NM-1}+C_{NM}^{NM}=2^{NM} \quad \textbf{(1)}$$

The number of possible scenes or hypotheses about the radar scene is 2^{NM}. Let the space of hypotheses be denoted by X. The state of our model is $X_t=x$ where $x\in X$. Let Y_t be the measurement variable. Let u_t be the control variable that indicates which waveform is chosen at time t to generate measurement Y_{t+1}, where $u_t\in U$. The probability of receiving a particular measurement $X_t=x$ will depend on both the true, underlying scene and on the choice of waveform used to generate the measurement.

We define $a_{x'x}$ is state transition probability where

$$a_{x'x}=P(x_{t+1}=x'\mid x_t=x) \quad \textbf{(2)}$$

We define $b_{x'x}$ is the measurement probability where

$$b_{x'x}(u_t)=P(Y_{t+1}=x'\mid X_t=x,u_t) \quad \textbf{(3)}$$

Assume the transmitted baseband signal is $s(t)$, and the received baseband signal is $r(t)$. The matched filter is the one with an impulse response $h(t)=s^*(-t)$, so an output process of our matched filter is

$$x(t)=\int s^*(\lambda-t)r(\lambda)d\lambda \quad \textbf{(4)}$$

In the radar case, the return signal is expected to be Doppler shifted, then the matched filter to a return signal with an expected frequency shift v_0 has an impulse response

$$h(t)=s^*(-t)e^{j2\pi v_0 t} \quad \textbf{(5)}$$

The output is given by

$$x(t)=\int s^*(\lambda-t)e^{-j2\pi v_0(\lambda-t)}r(\lambda)d\lambda \quad \textbf{(6)}$$

where v_0 is an expected frequency shift.

The baseband received signal will be modeled as a return from a Swerling target:

$$r(t)=As(t-\tau)e^{j2\pi v_d t}I+n(t) \quad \textbf{(7)}$$

where $s(t,\tau,v_d)=s(t-\tau)e^{j2\pi v_d t}$ is a delayed t and Doppler-shifted v_d replica of the emitted baseband complex envelope signal $s(t)$; I is a target indicator. A approaches a complex Gassian random variable with zero mean and variance $2\sigma_A^2$. We assume $n(t)$ is complex white Gaussian noise independent of A, with zero mean and variance $2N_0$.

At time t the magnitude square of the output of a filter matched to a zero delay and a zero Doppler shift is

$$|x(t)|^2=\left|\int_0^t r(\lambda)s^*(\lambda-t)d\lambda\right|^2 \quad \textbf{(8)}$$

When there is no target

$$r(t)=v(t) \quad \textbf{(9)}$$

So

$$x(\tau_0)=\int_0^{\tau_0} n(\lambda)s^*(\lambda-\tau_0)d\lambda \quad \textbf{(10)}$$

The random variable $x(\tau_0)$. is complex Gaussian, with zero mean and variance given by

$$\sigma_0^2=E\left\{x(\tau_0)x^*(\tau_0)\right\}=2N_0\xi \quad \textbf{(11)}$$

ξ is the energy of the transmitted pulse.

When target is present

$$r(t)=As(t-\tau)e^{j2\pi v_d t}I+n(t) \quad \textbf{(12)}$$

$$x(\tau_0)=\int_0^{\tau_0}\left[As(\lambda-\tau)e^{j2\pi v_d\lambda}+n(\lambda)\right]s^*(\lambda-\tau_0)d\lambda \quad \textbf{(13)}$$

This random variable is still zero mean, with variance given by

$$\begin{aligned}\sigma_1^2&=E\left\{x(\tau_0)x^*(\tau_0)\right\}\\&=\sigma_0^2(1+\frac{2\sigma_A^2\xi^2}{\sigma_0^2}A(\tau_0-\tau,v_0-v))\end{aligned} \quad \textbf{(14)}$$

$A(t,v)$is ambiguity function, given by

$$A(\tau,v)=\frac{1}{\left(\int|s(\lambda)|^2d\lambda\right)^2}\left|\int s(\lambda)s^*(\lambda-\tau)e^{j2\pi v\lambda}d\lambda\right|^2 \quad \textbf{(15)}$$

Recall that the magnitude square of a complex Gaussian random variable $x\sim N(0,\sigma_i^2)$ is exponentially distributed, with density given by

$$y = x^2 \sim \frac{1}{2\sigma_i^2} e^{-\frac{y}{2\sigma_i^2}} \quad \textbf{(16)}$$

We consequently have that the probability of false alarm P_f is given by

$$P_f = \int_D^\infty \frac{1}{2\sigma_0^2} e^{-\frac{x}{2\sigma_0^2}} dx = e^{-\frac{D}{2\sigma_0^2}} \quad \textbf{(17)}$$

And the probability of detection P_d by

$$P_d = \int_D^\infty \frac{1}{2\sigma_1^2} e^{-\frac{x}{2\sigma_1^2}} dx = e^{-\frac{D}{2\sigma_0^2(1+\frac{2\sigma_A^2\xi^2}{\sigma_0^2}A(\tau_0-\tau,\nu_0-\nu))}} \quad \textbf{(18)}$$

In the case when a target is present in cell (τ,υ), assuming its actual location in the cell has a uniform distribution

$$P_d = \frac{1}{|A|}\int_{(\tau_a,\upsilon_a\in A)} e^{-\frac{D}{2\sigma_0^2(1+\frac{2\sigma_A^2\xi^2}{\sigma_0^2}A(\tau_0-\tau,\nu_0-\nu))}} d\tau_a d\upsilon_a \quad \textbf{(19)}$$

where A is the resolution cell centred on (*t*,*v*) with volume |*A*|.

3. Q-Learning-Based Stochastic Dynamic Programming

A target for which measurements are to be made will fall in a resolution cell. Another target, conceptually, does not interfere with measurements on the first if it occupies another resolution cell different from the first. Thus, conceptually, as long as each target occupies a resolution cell and the cells are all disjoint, the radar can make measurements on each target free of interference from others.

Define $\pi = \{u_0, u_1, ..., u_T\}$ where *T*=1 is the maximum number of dwells that can be used to detect and confirm targets for a given beam. Then π is a sequence of waveforms that could be used for that decision process.

We can obtain different π according to different environment in cognitive radar. Let

$$V_t(X_t) = E[\sum_{t=0}^{T}\gamma^t R(X_t, u_t)] \quad \textbf{(20)}$$

where R（X_t,u_t） is the reward earned when the scene X_t is observed using waveform u_t and γ is discount factor. Then the aim of our problem is to find the sequence π^* that satisfies

$$V^*(X_t) = \max_\pi E[\sum_{t=0}^{T}\gamma^t R(X_t, u_t)] \quad \textbf{(21)}$$

However, knowledge of the actual state is not available. Using the method of [10], we can obtain that the optimal control policy π^* that is the solution of (21) is also the solution of

$$V^*(\mathbf{p}(0)) = \max_\pi E[\sum_{t=0}^{T}\gamma^t R(\mathbf{p}_t, u_t)] \quad \textbf{(22)}$$

where $\mathbf{P}_t$ is the conditional density of the state given the measurements and the controls and $\mathbf{P}_0$ is the a priori probability density of the scene. **P** is a sufficient statistic for the true state X_t. So we need to solve the following problem

$$\max_\pi E[\sum_{t=0}^{T}\gamma^t R(\mathbf{p}_t, u_t)] \quad \textbf{(23)}$$

The refreshment formula of $\mathbf{P}_t$ is given by

$$\mathbf{p}_{t+1} = \frac{\mathbf{BAp}_t}{\mathbf{1'LAp}_t} \quad \textbf{(24)}$$

where **B** is the diagonal matrix with the vector $(b_{x'x}(u_t))$ the non-zero elements and **1** is a column vector of ones. **A** is state transition matrix.

If we wanted to solve this problem using classical dynamic programming, we could have to find the value function $V_t(\mathbf{p}_t)$ using

$$V_t(\mathbf{p}_t) = \max_{u_t}(R_t(\mathbf{p}_t, u_t) + \gamma E\{V_{t+1}(\mathbf{p}_{t+1}) \mid \mathbf{p}_t\}) \quad \textbf{(25)}$$

It can also be written in probability form

$$V_t(\mathbf{p}_t) = \max_{u_t}(R_t(\mathbf{p}_t, u_t) + \gamma\sum_{\mathbf{p}'\in\mathbf{P}} P(\mathbf{p}'|\mathbf{p}_t, u_t)V_{t+1}(\mathbf{p}')) \quad \textbf{(26)}$$

However, in radar scene, explicit knowledge of target state-transition probabilities are unknown. So directly using Bellman's dynamic programming is very hard. The Q-leaning algorithm is a direct approximation of Bellman's dynamic programming, and it can solve the problem that we do not know explicit knowledge of state-transition probabilities. For this reason, Q-learning is very suitable to be used in the problem of adaptive waveform selection in cognitive radar.

We define a Q-factor in our problem. For a state-action pair $(\mathbf{p}_t, u_t)$,

$$Q(\mathbf{p}_t, u_t) = \sum_{\mathbf{p}'\in\mathbf{P}} P(\mathbf{p}'|\mathbf{p}_t, u_t)[R_t(\mathbf{p}'|\mathbf{p}_t, u_t) + \gamma V_{t+1}] \quad \textbf{(27)}$$

According to (26), (27) we can derive

$$V_t^* = \max_{u_t} Q(\mathbf{p}_t, u_t) \quad \textbf{(28)}$$

The above establishes the relationship between the value function of a state and the Q-factors associated with a state. Then it should be clear that, if the Q-factors are known, one can obtain the value function of a given state from above fomula.

So Q form of Bellman equation is

$$\begin{aligned} &Q(\mathbf{p}_t, u_t) \\ &= \sum_{\mathbf{p}'\in\mathbf{P}} P(\mathbf{p}'|\mathbf{p}_t, u_t)[R_t(\mathbf{p}'|\mathbf{p}_t, u_t) + \gamma\max_{u_{t+1}} Q(\mathbf{p}_{t+1}, u_{t+1})] \end{aligned} \quad \textbf{(29)}$$

Let us denote the *i*th independent sample of a random variable *X* by S^i and the expected value by *E*(*X*). X^n is the estimate of *X* in the *n* th iteration. So

$$E(X)=\lim_{n\to\infty}\frac{\sum_{i=1}^{n}s^i}{n} \quad \textbf{(30)}$$

$$X^n=\frac{\sum_{i=1}^{n}s^i}{n} \quad \textbf{(31)}$$

We can derive

$$X^{n+1}=(1-\alpha^{n+1})X^n+\alpha^{n+1}s^{n+1} \quad \textbf{(32)}$$

where

$$\alpha^{n+1}=\frac{1}{n+1} \quad \textbf{(33)}$$

So

$$Q(\mathbf{p}_t,u_t)=E[R_t(\mathbf{p}'\big|\mathbf{p}_t,u_t)+\gamma\max_{u_{t+1}}Q(\mathbf{p}_{t+1},u_{t+1})] \quad \textbf{(34)}$$

where *E* is the expectation operator. We could use this scheme in a simulator to estimate the same Q-factor. Using this algorithm, Equation (29) becomes:

$$\begin{aligned}&Q^{n+1}(\mathbf{p}_t,u_t)\leftarrow(1-\alpha^{n+1})Q^n(\mathbf{p}_t,u_t)\\&+\alpha^{n+1}[R_t(\mathbf{p}'\big|\mathbf{p}_t,u_t)+\gamma\max_{u_{t+1}}Q^n(\mathbf{p}_{t+1},u_{t+1})]\end{aligned} \quad \textbf{(35)}$$

Obviously, we do not have the transition probabilities in it.

Our Q-learning algorithm is as follows:

Step 1. Initialize the Q-factors to 0. Set *n*=1.

Step 2. For *t*=0,1,…*T*,do step 3-step 6.

Step 3. Simulation action u_t. Let the curren state be $\mathbf{P}_t$, and the next state be $\mathbf{P}_{t+1}$.

Step 4. Find the decision using the current Q-factors:

$$u_t=\arg\max_{u_t}Q_t^{n-1}(\mathbf{p}_t^n,u_t) \quad \textbf{(36)}$$

Step 5. Update Q($\mathbf{P}_t$,u_t) using the following equation:

$$\begin{aligned}&Q^{n+1}(\mathbf{p}_t,u_t)\leftarrow(1-\alpha^{n+1})Q^n(\mathbf{p}_t,u_t)\\&+\alpha^{n+1}[R_t(\mathbf{p}'\big|\mathbf{p}_t,u_t)+\gamma\max_{u_{t+1}}Q^n(\mathbf{p}_{t+1},u_{t+1})]\end{aligned} \quad \textbf{(37)}$$

Step 6. Find the next state:

$$\mathbf{p}_{t+1}=\frac{\mathbf{BAp}_t}{\mathbf{1'BAp}_t} \quad \textbf{(38)}$$

Step 6. Increment *n*. If *n*<*N*, go to step 2.

Step 7. For each **P** *t*∈P, select

$$d(\mathbf{p}_t)\in\arg\max_{u_t}Q(\mathbf{p}_{t+1},u_{t+1}) \quad \textbf{(39)}$$

The policy generated by the algorithn is $\hat{d}$. Stop.

4. Simulation

In this section, we make three experiments. In order to explain the necessity of waveform selection, we make the curve of measurement probability versus SNR of three waveforms. Curve of uncertainty of state estimation demonstrates validity of our proposed algorithm. We also plot the figure of Q value space versus state and waveform.

We consider a simple situation. The state space is 4× 4. We consider 5 different waveforms where for each waveform *u*, and each hypotheses for the target x, the distribution of x' is given in Table 1. The discount factor γ=0.9. State transition matrix **A** is given by

$$\mathbf{A}=\begin{bmatrix}0.96 & 0.02 & 0.01 & 0.04\\0.01 & 0.93 & 0.03 & 0.04\\0.02 & 0.03 & 0.95 & 0.02\\0.01 & 0.02 & 0.01 & 0.9\end{bmatrix} \quad \textbf{(40)}$$

Following the approach described in [11,12], linear form of reward function will be adopted:

$$R(\mathbf{p},u)=\mathbf{p}'\mathbf{p}-1 \quad \textbf{(41)}$$

The formula *E*(–*R*) can be considered as the uncertainty in the state estimation. In other words, it can be considered as the tracking errors.

Figure 2 is curve of measurement probability versus SNR of three waveforms. From this curve we can see that with the same SNR, different waveforms correspond to different measurement probability. Generally speaking, the waveform with wide pulse duration corresponds to high measurement probability. From this point of view, the waveform with wide pulse duration is better. However, wide pulse duration means large energy of the transmitted pulse. So we should improve measurement

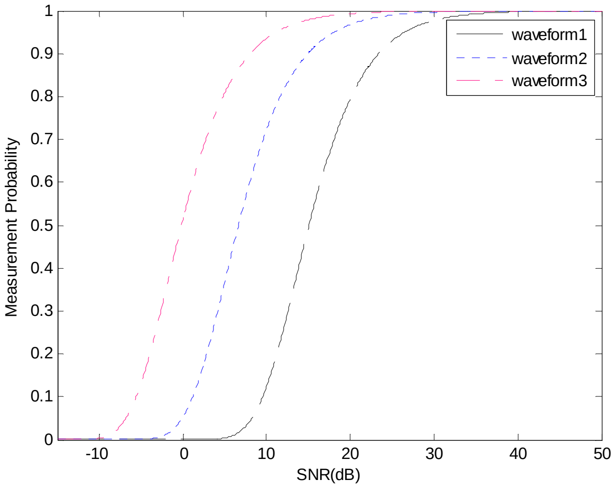


Figure 2. Curve of measurement probability versus SNR of three waveforms.

Table 1. Measurement probabilities for the example scenario.

	x=1 x’=1,2,3,4	x=2 x’=1,2,3,4	x=3 x’=1,2,3,4	x=4 x’=1,2,3,4
u=1	0.97,0.01 0.01,0.01	0.01,0.01 0.96,0.02	0.01,0.02, 0.01,0.96	0.96,0.01, 0.01,0.02
u=2	0.96,0.01 0.02,0.01	0.02,0.95 0.01,0.02	0.01,0.01, 0.01,0.97	0.02,0.96, 0.01,0.01
u=3	0.94,0.02 0.03,0.01	0.02,0.02 0.01,0.95	0.02,0.96, 0.01,0.97	0.01,0.02, 0.95,0.02
u=4	0.96,0.01 0.01,0.02	0.01,0.02 0.96,0.01	0.97,0.01, 0.01,0.01	0.03,0.95, 0.01,0.01
u=5	0.95,0.02 0.01,0.02	0.01,0.97 0.01,0.01	0.02,0.01 0.96,0.01	0.04,0.94 0.01,0.01

probability through changing waveforms according to different environment and make a balance between the width of pulse duration and the energy of the transmitted pulse. We can also derive measurement probabilities for the example scenario from this curve, as is shown in Table 1.

Figure 3 is curve of uncertainty of state estimation. From this curve we can see that for all the cases, the uncertainty of state estimation is decreasing with time, no matter how the state is changing with time. Compared to a fixed waveform, Q-learning algorithm we proposed has lower uncertainty of state estimation. That means our algorithm will reduce uncertainty in locating targets. Meanwhile our algorithm approaches the optimal waveform selection scheme even though explicit knowledge of state-transition probabilities is unknown.

Figure 4 is the figure of Q value space versus state and waveform. Q value of different state-waveform pair can be obtained in this figure. We can see that the proposed algorithm has lower computational cost.

5. Conclusions

Adaptive waveform selection is an important problem in cognitive radar and the problem of adaptive waveform scheduling can be viewed as a stochastic dynamic programming problem. In this paper, Q-learning-based waveform selecting algorithm is proposed. The advantages of Q-learning over fixed waveform have been shown with simulations. The Q-learning algorithm can minimize the uncertainty of state estimation compared to fixed waveform and approaches the optimal waveform selection scheme. Meanwhile, Q-learning can solve the problems in which explicit knowledge of state-transition probabilities are unknown. Reasearch on alogorithms which approach the optimal waveform selection scheme and has lower computational cost is an important problem.

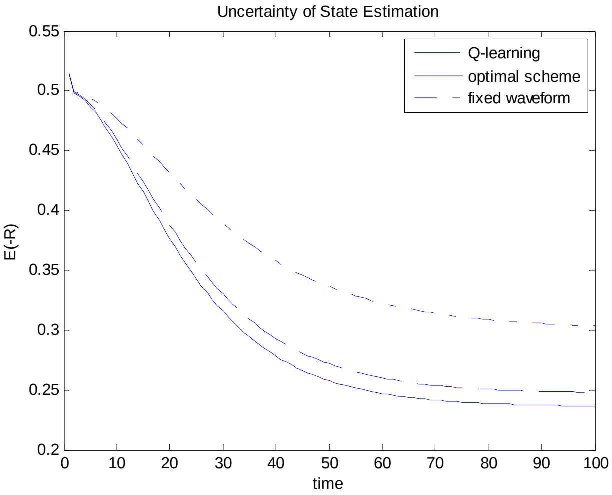


Figure 3. Curve of uncertainty of state estimation.

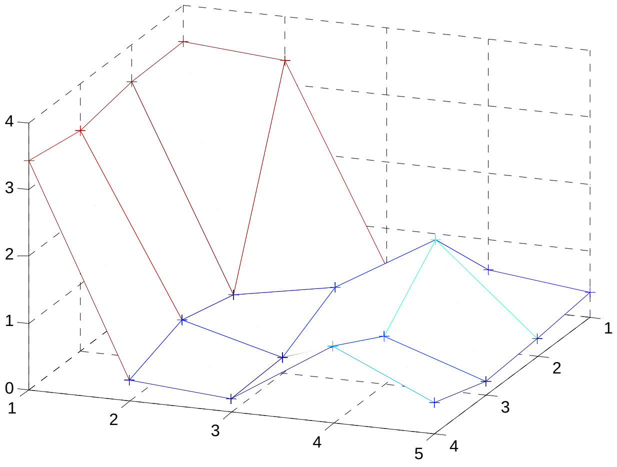


Figure 4. Q value space versus state and waveform.

6. References

[1] S. Haykin, “Cognitive radar: A way of the future,” IEEE Signal Processing Magazine, Vol. 23, No. 1, pp. 30–40, 2006.

[2] C. Rago, P. Willett, and Y. Bar-Shalom, “Detecting-tracking performance with combined waveforms,” IEEE Transactions on Aerospace and Electronic Systems, Vol. 34, No. 2, pp. 612–624, 1998.

[3] D. J. Kershaw and R. J. Evans, “Waveform selective probabilistic data association,” IEEE Transactions on Aerospace and Electronic Systems, Vol. 33, No. 4, pp. 1180–1188, 1997.

[4] Y. He and E. K. P. Chong, “Sensor scheduling for target tracking in sensor networks,” 43rd IEEE Conference on Decision and Control, Paradise, Island, Bahamas, pp. 743–748, 2004.

[5] V. Krishnamurthy, “Algorithms for optimal scheduling of hidden Markov model sensors,” IEEE Trans. on Signal Processing, Vol. 50, No. 6, pp.1382–1397, 2002.

[6] C. O. Savage, and B. Moran, “Waveform selection for maneuvering targets within an IMM framework,” IEEE Transactions on Aerospace and Electronic Systems, Vol. 43, No. 3, pp. 1205–1214, 2007.

[7] C. T. Capraro, I. Bradaric, G. T. Capraro, and T. K. Lue,

"Using genetic algorithms for radar selection," 2008 IEEE Radar Conference, Inc., Utica, NY, pp. 1–6, May 2008.

[8] B. F. La Scala and R. J. Moran Wand Evans, "Optimal adaptive waveform selection for target detection," The International Conference on Radar, Adelaide, SA, Australia, pp. 492–496, Sept. 2003.

[9] La Scala, Rezaeian, and Moran, "Optimal adaptive waveform selection for target tracking," International Conference on Information Fusion, pp. 552–557, 2005.

[10] D. Bertsekas, "Dynamic programming and optimal control," Athena Scientific, Second Edition, Vol. 1, 2001.

[11] V. Krishnamurthy, "Algorithms for optimal scheduling of hidden Markov model sensors," IEEE Transactions on Signal Processing, Vol. 50, No. 6, pp. 1382–1397, 2002.

[12] W. S. Lovejoy, "Computationally feasible bounds for partially observed Markov decision processes," Operations Research, Vol. 39, No. 1, pp. 162–175, 1991.

Int. J. Communications, Network and System Sciences, 2009, 7, 675-686
doi:10.4236/ijcns.2009.27078 Published Online September 2009 (http://www.SciRP.org/journal/ijcns/).

TCP-R with EPDN: Handling out of Order Packets in Error Prone Satellite Networks

Arjuna SATHIASEELAN
Electronics Research Group, University of Aberdeen, Aberdeen, United Kingdom
Email: *arjuna@erg.abdn.ac.uk*

Received May 28, 2009; *revised July* 8, 2009; *accepted September* 10, 2009

ABSTRACT

Studies have shown that packet reordering is common, especially in satellite networks where there are link level retransmissions and multipath routing. Moreover, traditional satellite networks exhibit high corruption rates causing packet losses. Reordering and corruption of packets decrease the TCP performance of a network, mainly because it leads to overestimation of the congestion in the network. We consider satellite networks and analyze the performance of such networks when reordering and corruption of packets occurs. We propose a solution that could significantly improve the performance of the network when reordering and corruption of packets occur in a satellite network. We report results of our simulation experiments, which support this claim.

Keywords: TCP, Packet, Reorder, Throughput, Drops, Congestion, Multipath, Satellite

1. Introduction

Transmission Control Protocol (TCP) is the commonly used transport protocol in the Internet. The TCP protocol provides a reliable, connection oriented, in-order delivery of data between any two hosts in the Internet [1]. To ensure the data is delivered from the sender to receiver correctly and in-order, the TCP sender uses sequence numbers to each octet of data that is transmitted and the TCP receiver acknowledgements (ACKs) the receipt of these transmitted bytes that are received correctly and in-order. Since ACKs are cumulative, the receipt of out-of-order packets generates duplicate ACKs that are sent to the TCP sender. TCP assumes congestion in the network to be the cause of loss of packets. Thus when a TCP sender receives three successive duplicate ACKs, it assumes a packet has been lost and that this loss is an indication of network congestion and reduces its sending rate [2].

Networking using satellites began by using individual satellites in Geostationary Earth Orbit (GEO), where the signals were amplified and then up-linked to the GEO satellite. The satellite then frequency-shifts the signal and broadcasts it down to a large ground area. These GEO satellites acted as simple transparent ‘bent-pipe’ repeaters [3]. A Lower Earth Orbit (LEO) network such as Iridium [4] has several satellites connected together to form a network. When there is congestion in a particular path, the satellite routes the packets through a different path. This introduces reordering of packets [5]. Satellite networks such as GEO and LEO have high round trip times (RTTs), typically in the order of several hundred milliseconds. In order to keep the pipe full, link-layer retransmission protocols send subsequent packets while awaiting an ACK or negative acknowledgement (NAK) for a previously sent packet. Here, a link-layer retransmission is *reordered* by the number of packets that were sent between the original transmission of that packet and the return of the ACK or NAK [6].

In satellite networks, the packet loss is mainly due to corruption. These corrupted packets could be dropped either in the routers (in case of LEO network) or in the receiver when the header checksum fails. If the link layer at the receiver detects any errors and if there is enough redundancy transmitted in the code, then the errors can be corrected using the Forward Error Correction (FEC) algorithms without requesting for a retransmission. Therefore, the link layer could pass the error free packet to the top layers. If the link layer detects an error but cannot correct it (i.e. the cyclic redundancy check (CRC) fails, the link layer drops the corrupted packet and the link layer at the receiver requests the link layer at the sender to retransmit the packet. These link level retransmissions only make a limited attempt to recover the lost packet. If the link layer cannot recover the lost packet, it will be left to the higher layers to recover the packet.

Thus when packets are lost due to corruption, link layer protocols that do not attempt in-order delivery across the link cause packets to reach the TCP receiver in out-of-order. This leads to the generation of duplicate ACKs by the TCP receiver, which causes the sender to invoke fast retransmission and recovery [7].

There have been many proposals for extending TCP to improve the performance when the losses are mainly due to corruption. Some of the proposed mechanisms are the Explicit Loss Notification (ELN) mechanism [8], Explicit Transport Error Notification (ETEN) [9], Indirect-TCP (I-TCP) [10], TCP PEACH [11] etc. These proposals improve the performance of TCP when packets get lost due to corruption in the network but do not consider the effects caused due to reordering of packets.

Several methods to detect the needless retransmission due to the reordering of packets have also been proposed:

The DSACK (Duplicate Selective Acknowledgement) option in TCP, allows the TCP receiver to report to the sender when duplicate segments arrive at the receiver's end. Using this information, the sender can determine whether a retransmission was spurious [12]. If the retransmission was spurious, then the slow start threshold (*ssthresh*) is set to the previous congestion window (*cwnd*). Their proposal does not specify any mechanisms to proactively detect reordering of packets. We term this mechanism as DSACK-R (DSACK with Recovery).

In [13], the authors propose mechanisms to detect and recover from false retransmits using the DSACK information. They propose several algorithms for proactively avoiding false retransmits by adaptively varying the duplicate threshold (*dupthresh*) value. The DSACK-FA algorithm (DSACK-R + fixed FA ratio), the *dupthresh* value is chosen to avoid a percentage of false fast retransmits, by setting the *dupthresh* value equal to the percentile value in the reordering length cumulative distribution. The percentage of reordering the algorithm avoids is known as FA ratio algorithm. In the DSACK-FAES algorithm (DSACK-FA + Enhanced RTT sampling), the DSACK-FA algorithm is combined with a RTT sampling algorithm which samples the RTT of retransmitted packets caused by packet delays. The DSACK-TA algorithm (DSACK-FA + Timeout Avoidance) uses cost functions that heuristically increase or decrease the FA ratio such that the throughput is maximized for a connection experiencing reordering. The FA ratio will increase when false retransmits occur and the FA ratio will decrease when there are significant timeouts. In the DSACK-TAES algorithm (DSACK-TA + Enhanced RTT sampling), the DSACK-TA algorithm is combined with a RTT sampling algorithm which samples the RTT of retransmitted packets caused by packet delays. According to [19], the DSACK-TA algorithm performed the best when compared with the other algorithms for various delay distributions. DSACK-TAES performed better than DSACK-TA for large packet delays that exceeded the one second minimum RTO. For other delay distributions, both DSACK-TA and DSACK-TAES perform similarly.

In [14], we proposed a novel method to enable the TCP senders to distinguish whether a packet has been dropped or reordered in the network by using the gateways to inform the 'receiver' about the dropped packets. This mechanism was called the Explicit Packet Drop Notification version 2.0 (EPDNv2). The receiver then uses this information to inform the sender about which packets have been reordered by setting a *drop-negative* bit. If the packets had been dropped in the network, the TCP sender retransmits the lost packets after waiting for three duplicate ACKs. If the packets are assumed to be reordered in the network, the TCP sender waits for '3+k' duplicate ACKs (*k*,1) before retransmitting the packets. We termed this new version of TCP as Reorder Notifying TCP (RN-TCP). Although using EPDNv2 requires modifications to all routers along the path, when combined with RN-TCP it gave good performance improvements. The computational and storage costs, and other implementation issues were analyzed in detail in [14].

The proposals mentioned to alleviate the TCP performance in the presence of packet reordering do not consider error prone networks. It would be interesting to find out the performance of these protocols when reordering happens in a network that is error prone. If a packet had been actually dropped due to corruption, having an increased value of *dupthresh* may require a timeout to detect the packet loss. Thus increasing the *dupthresh* value to more than three when a packet has been assumed not to be dropped may have serious implications in the performance, if the packet had actually been dropped due to corruption. The EPDNv2 mechanism proposed by us in [14], informs the sender/receiver about dropped packets. This could be dropped due to congestion in the network or due to corruption in the network. RN-TCP retransmits the lost packet and reduces the transmission rate even if the packets had been dropped due to corruption. If a packet had actually been lost due to corruption, the performance of TCP can be improved, if the TCP sender does not reduce the *cwnd* upon a retransmission of the lost packet. Moreover, when networks exhibit high RTT, unnecessary reduction of the *cwnd* requires large number of RTTs to retrieve back to the previous *cwnd*. This reduces the throughput performance. Thus, it was imperative for us to propose a new TCP protocol, that can improve the throughput performance when packets experience reordering and corruption.

In this paper, we propose extending the SACK protocol to enable TCP senders to recognize whether a received duplicate ACK means that a packet has been dropped or corrupted/reordered. The extended protocol also requires a modification to EPDNv2. We call the modified mechanism as Explicit Packet Drop Notification Version 3.0 (EPDNv3) mechanism to infer which packets have been dropped due to congestion. The TCP sender uses this information to take an appropriate action. We term this new TCP protocols as Robust TCP (TCP-R).

The remainder of this paper is organized as follows. Section 2 presents the details of our proposed solution. In Section 3 discusses the simulation environment. Section 4 discusses the simulation results. We conclude the paper with a summary of our work and a short discussion of the further research in Section 5.

2. Implementation Details

2.1. EPDNv3

We propose a mechanism similar to EPDNv2 (proposed by us in [14]), by maintaining information about packets dropped due to congestion in the gateways and not by header checksum error which occurs mainly due to corruption[1]. Corrupted packets dropped by the MAC layer are not maintained. Each gateway maintains information about dropped packets in the gateways. This is by having a hashtable that maintains for each flow the maximum sequence number and minimum sequence number of the packets that get dropped due to congestion in the gateway. When the next data packet of flow *i* passes through that gateway, the gateway inserts the maximum sequence number and the minimum sequence number of the dropped packets in the data packet and the entry is deleted from the data structure. We term this mechanism of explicitly informing the receiver about the dropped information as Explicit Packet Drop Notification Version 3.0 (EPDNv3.0)[2].

2.2. TCP-R: Robust TCP

We propose extending the TCP SACK protocol to enable TCP senders to recognize whether a received duplicate ACK means that a packet has been dropped or corrupted/reordered in the network. The extended protocol uses the EPDNv3.0 mechanism mentioned above to infer which packets have been dropped. The TCP-R receiver maintains two lists: the reorder/corruption list and the drop list. The TCP-R receiver uses the algorithm mentioned in [14] to determine which packets have been dropped due to congestion and which have been reordered or corrupted, and puts those sequence numbers into the corresponding lists[3]. Informing the sender is done by setting the *drop-negative* bit in corresponding duplicate ACKs if the packet has been assumed to be reordered or corrupted[4]. If the packets are assumed to be reordered or corrupted in the network, the TCP-R sender retransmits the packet after receiving three duplicate ACKs with the *drop-negative* bit set and enters our modified fast recovery mechanism where the procedure of reducing the *ssthresh* and the *cwnd* are bypassed i.e. we do not reduce the *ssthresh* and the *cwnd*. In order to prevent TCP-R from falsely misjudging drops due to congestion and not reducing the sending rate, we ensure the modified fast recovery mechanism is executed only in the absence of ECN messages. If the packets had been dropped in the network, the TCP-R sender retransmits the lost packet after waiting for three duplicate ACKs (fast retransmit) and reduces the cwnd by half (fast recovery)[5].

In an environment where out-of-order packets can happen due to corruption and reordering, the throughput performance can be improved if we do not reduce the *cwnd* upon the occurrence of a reordering or a corruption event. Even though TCP-R unnecessarily retransmits a reordered packet, it reduces the instances of timeouts by not delaying the retransmission of corrupted packets. Moreover it does not reduce the *cwnd* unnecessarily, thus improving the throughput of the sender.

2.2.1. Sender Side: Implementation Details

When an ACK is received the TCP sender does the following,

- If none of the three duplicate ACKs received have their *drop-negative* bit set, then the sender assumes that the packet has been dropped. So the sender retransmits the lost packet after receiving three duplicate ACKs and enters fast recovery.
- If all the three duplicate ACKs received have their *drop-negative* bit set and the ECE bit is set to zero (i.e no ECN information has been received off lately), then the sender assumes that the packet has been reordered or corrupted in the network and retransmits the packet immediately. The procedure of reducing the *ssthresh* and the *cwnd* in the fast recovery procedure are bypassed. In order to avoid multiple retransmits, we ensure that a packet assumed to be reordered or corrupted would be retransmitted only once in a RTT.

3. Simulation Environment

Figure 1 presents the topology used for our simulations. The simulated network has a source and destination node

[1] EPDNv2 maintains information about drops caused by both corrupted and congested packets. The implementation details are similar to EPDNv2.

[2] The implementation details are similar to EPDNv2. The implementation issues and costs have been thoroughly analyzed in [14].

[3] TCP-R uses the reorder list specified in [14] to store out-of-order packet sequence numbers caused by losses due to corruption and reordering. We denote the list as reor-der/corruption list.

[4] The drop-negative bit is set to 1.

[5] The drop-negative bit is set to 0.

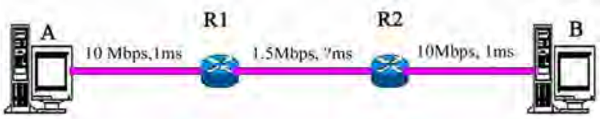


Figure 1. TCP-R: Simulated network.

connected to two intermediate routers. The nodes are connected to the routers via 10 Mbps Ethernet having a delay of 1 ms. The routers are connected to each other via long delay link with a fixed link capacity (set to 1.5 Mbps) and variable delay. Our simulations use 1500 byte packets. We used the RED queuing strategy with a queue size set to the bandwidth delay product. All routers were ECN enabled. Reordering and packet drops are introduced at the bottleneck link (R1,R2). The experiments were conducted using a single long lived FTP flow traversing the network topology, except otherwise noted. The maximum window size of the TCP flow was also set to the bandwidth delay product. The TCP flow lasts 1000 seconds.

Our simulations consider the case of both LEO and GEO satellite links. The lack of flexibility of the current ns-2 simulator to delay a fraction of packets and to test for various average packet delays for satellite links caused us to model the satellite links by representing a wired link with the same capacity and delay as a GEO or a LEO satellite link, similar to [15]. The GEO satellite link has roughly a delay of 300 ms (one way). The LEO satellite link has a one way delay that varies [40,400] ms depending on whether the LEO network has one satellite hop or multiple hops and how far each of these satellite hops are placed [15]. In GEO networks, we consider the reordering to be caused only due to link level retransmissions and in LEO networks we consider the case of link level retransmissions and multipath routing. Reordering is caused when a delayed packet with a higher sequence number is scheduled to traverse the link later than an un-delayed packet with a lower sequence number. To simulate packet reordering, we delay a percentage of packets traversing the link by delay distributions. Our experiments can be classified into the following scenarios.

1) **Scenario 1:** The link layer at the receiver could detect and correct *some* of the corrupted packets using link level retransmissions. Those packets that cannot be recovered at the link layer will not arrive at the TCP, causing out-of-order packets. In addition to this, some corrupted packets may have been dropped by the satellite nodes (incase of LEO networks). Due to retransmission at the link layer, some of the packets could have been reordered. Thus the TCP receiver will receive out-of-order packets caused by both reordering and corruption. In order to simulate packet loss due to corruption, the experiments were performed for a BER of 10^6.

2) **Scenario 2:** The link layer at the receiver could detect and correct *all* corrupted packets using link level retransmissions, but these link level retransmissions could cause reordering. In this scenario, we assume all packets are correctly retrieved in the link layer after retransmissions. In addition to this, we also assume that the intermediate satellite nodes do not drop any corrupted packets. In order to simulate no loss of packets, we set the BER to zero. Thus the TCP receiver will receive out-of-order packets caused by reordering only.

3) **Scenario 3:** Reordering of packets caused due to multipath routing (in LEO networks) and some packets getting dropped due to corruption. Thus the TCP receiver will receive out-of-order packets caused by reordering and corruption. In order to simulate packet loss due to corruption, the experiments were performed for a BER of 10^6.

4) **Scenario 4:** Reordering of packets caused due to multipath routing (in LEO networks) and no packets getting dropped due to corruption. In order to simulate no loss of packets, we set the BER to zero. Thus the TCP receiver will receive out-of-order packets caused by reordering.

We compare TCP-R with RN-TCP(EPDNv2) i.e. when RN-TCP operates with EPDNv2, RN-TCP (EPDNv3) i.e. when RN-TCP operates with EPDNv3, DSACK-TA, DSACK-R and SACK.

4. Results

4.1. Reordering due to Link Level Retransmissions

In this section, we consider the case when packets get reordered due to link level retransmissions in GEO and LEO networks. The propagation delay was set to 300 ms. To introduce severe packet delays in the order of multiple of RTTs, we used a mean packet delay of yP s (P is the propagation delay) and standard deviation of $y/3\ P$ s, such that the delay introduced varied from 0 to $2yP$ s. The packet delay rate was fixed at 4%. We varied the value of y from 1.0 to 6.0.

4.1.1. Throughput: Large Reordering Delays (Scenario 1)

In this section, we compare the throughput performance of TCP-R, RN-TCP(EPDNv2), RN-TCP(EPDNv3), DSACK-TA, DSACK-R and SACK when the packets experience reordering and losses due to corruption. In order to simulate packet losses due to corruption, we set the BER to 10^6. Figure 2 presents the results of the simulations. As we increase y, the RTT of the packet that gets delayed exceeds the one second minimum RTO causing large number of timeouts. TCP-R outperforms RN-TCP (EPDNv2), RN-TCP(EPDNv3), DSACK-TA, DSACK-R and SACK for all tested mean packet delays from 0.3 s to 1.8 s. For example, when the average packet delay is 0.9 s, TCP-R gives a three fold throughput improvement

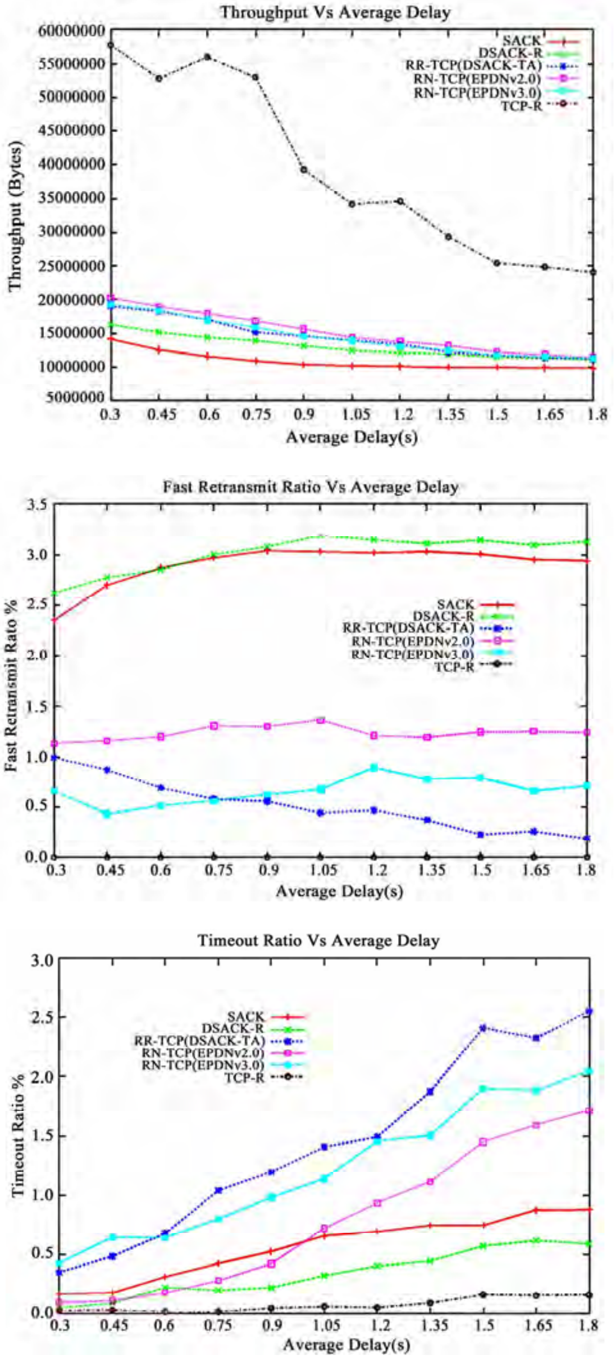


Figure 2. GEO link: Propagation delay of 300 ms, BER of 10.

over the other protocols. When the average packet delay is 1.8 s (3 *x RTT*), TCP-R gives more than a two fold throughput improvement over the other protocols.

When packets get dropped in the gateways due to header checksum error caused by corruption, EPDNv2 informs the sender about the dropped information. RN-TCP(EPDNv2) on receipt of this information, retransmits the corrupted packet immediately and reduces the *cwnd* after receiving three duplicate ACKs. Thus RN-TCP(EPDNv2) encounters more fast retransmissions than RN-TCP(EPDNv3) but lesser incidence of timeouts compared to RN-TCP(EPDNv3) and DSACK-TA. On the other hand, EPDNv3 informs the sender only about packets dropped due to congestion. Thus RN-TCP (EPDNv3) would assume the packet to be reordered and delays the fast retransmission procedure. This could cause the timer to expire leading to timeouts. Unlike RN-TCP and DSACK-TA, TCP-R on detecting the packet has not been dropped due to congestion in the network, retransmits the packet after receiving three duplicate ACKs without reducing the *cwnd*. This reduces the incidence of timeouts and unnecessary reduction of the *cwnd* due to fast retransmissions, leading to an improved throughput performance.

4.1.2. Throughput: Large Reordering Delays (Scenario 2).

Figure 3 presents the results of the simulations when packets just experience reordering and no loss of packets due to corruption. RN-TCP(EPDNv2) and RN-TCP (EPDNv3) perform similarly as both EPDN versions store only packets that get dropped due to congestion as there are no drops due to corruption. Initially, TCP-R performs slightly less compared to RN-TCP and DSACK-TA for average packet delays of 0.3 s and 0.45 s. From further investigation, as TCP-R sends more packets compared to RN-TCP and DSACK-TA (TCP-R does not prevent retransmission of corrupted or reordered packets), the ECN enabled routers mark the packets of the TCP-R flow by setting the CE bit. This causes TCP-R to reduce the *cwnd* more often. When the average packet delay is more than 0.6 s, TCP-R performs better than the other protocols. For example, when the average packet delay is set to 0.9 s, TCP-R gives a 4% throughput improvement over RN-TCP, 9% improvement over DSACK-TA, four times more than DSACK-R and six times more than SACK. Similarly when the average packet delay is set to 1.8 s, TCP-R gives a two fold throughput performance over RN-TCP, 80% improvement over DSACK-TA, three times more than DSACK-R and SACK. Moreover DSACK-TA undergoes large number of *cwnd* reductions due to fast retransmissions and has a higher timeout ratio compared to RN-TCP and TCP-R. TCP-R has almost zero fast retransmit and timeout ratios.

4.2. Reordering Due to Multipath Routing

In this section, we consider the case when packets get reordered due to multipath routing in LEO networks. The propagation delay was set to 400 ms. Multipath routing produces modal delays i.e. when successive packets are sent on paths with different RTTs, these packets would be reordered proportionally to the RTT difference of the path. 50% of the data packets were delayed. We performed

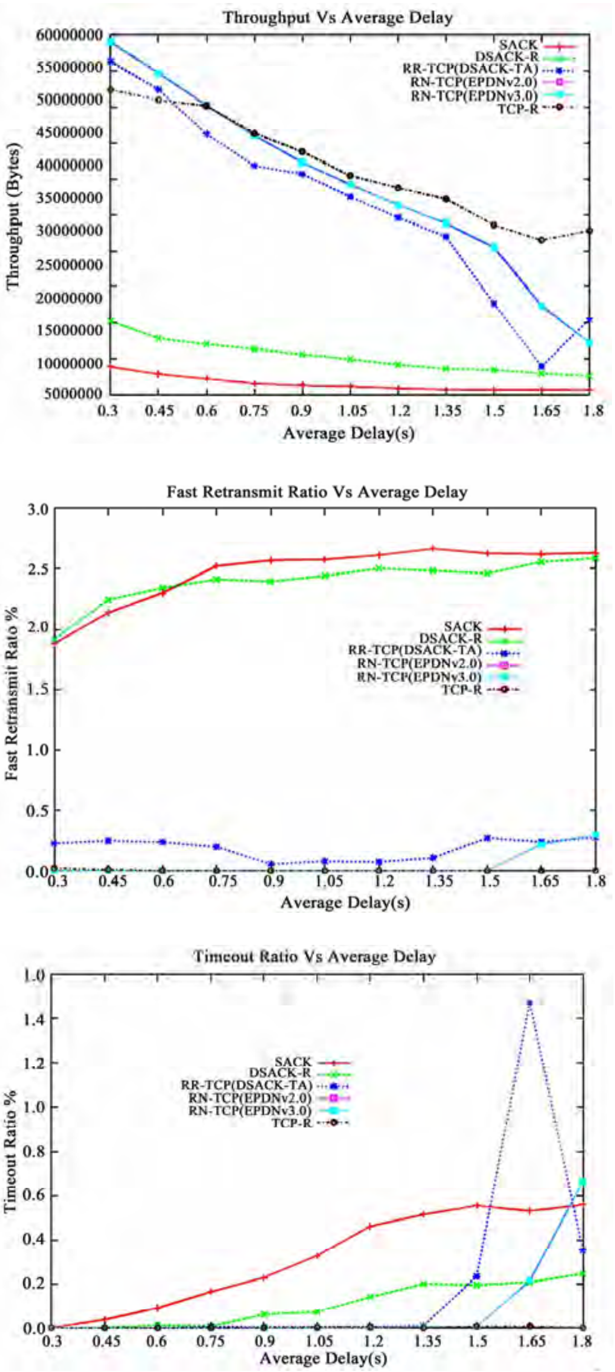


Figure 3. GEO link: Propagation delay of 300 ms.

experiments by gradually increasing the average packet delay from 0.0 s to 0.8 s (2 *x RTT*). When the average packet delay is 0.0 s, the packets are routed through the same path without the packets are routed through the same path without any reordering events.

4.2.1. Throughput: Multipath Routing (Scenario 3).

In this section, we compare the throughput performance of the simulated network using SACK, DSACK-R, DSACK-TA, RN-TCP and TCP-R when packets experience losses due to corruption and reordering. The experiments were performed for a BER of 10^6.

Figure 4 presents the results of the simulations when the propagation delay was set to 400 ms. For all tested average packet delays, TCP-R outperforms the other protocols. For example, when the average packet delay is 0.4 s, TCP-R performs four times more than RN-TCP (EPDNv2), five times more than DSACK-TA and RN-

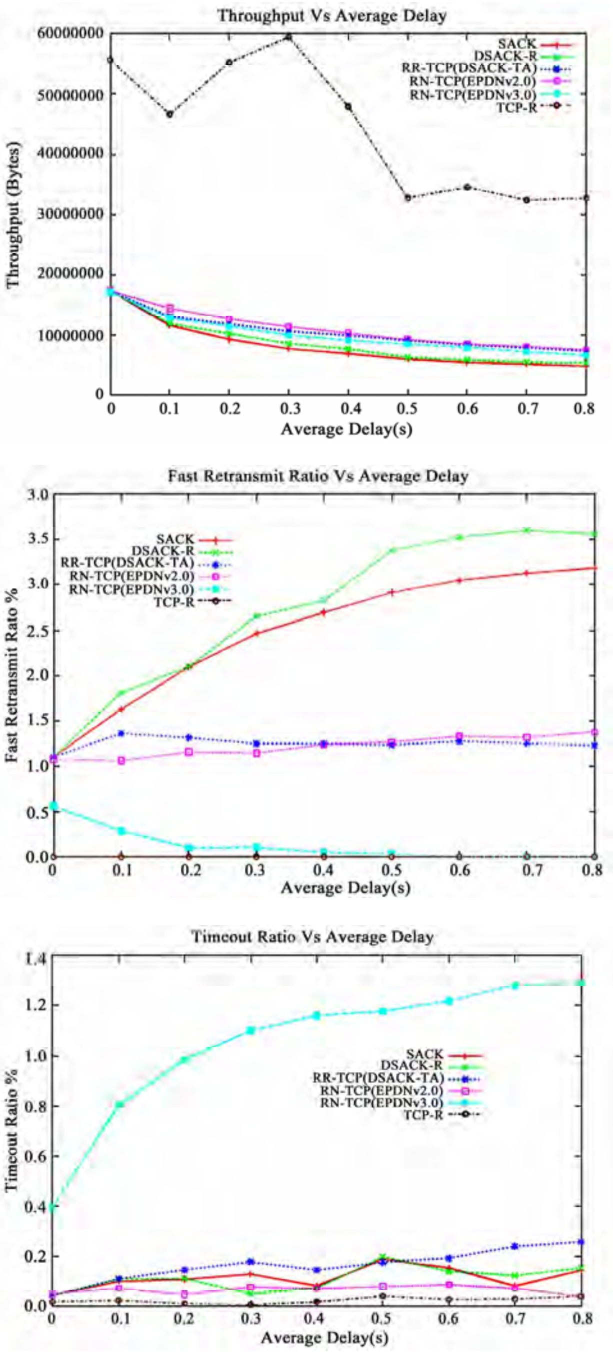


Figure 4. LEO link: Multipath performance, propagation delay of 400 ms, BER of 10^6.

TCP(EPDNv3), six times more than DSACK-R and almost seven times more than SACK. Similarly, when the average packet delay is 0.8 s, TCP-R performs four times more than RN-TCP(EPDNv2), RN-TCP(EPDNv3), DSACK-TA, six times more than DSACK-R and almost seven times more than SACK. RN-TCP(EPDNv2) encounters fewer timeouts compared to RN-TCP(EPDNv3) and DSACK-TA. TCP-R maintains a low incidence of timeouts and maintains a zero fast retransmit ratio compared to the other protocols, thus giving a better throughput performance.

4.2.2. Throughput: Multipath Routing (Scenario 4).

In this section, we compare the throughput performance of the simulated network using SACK, DSACK-R, DSACK-TA, RN-TCP and TCP-R when packets experience reordering due to multipath routing and no drops due to corruption.

Figure 5 presents the results of the simulations when packets experience reordering. TCP-R performs almost similar to RN-TCP for all tested average packet delays. TCP-R gives a better throughput performance when compared to DSACK-TA for majority of the average packet delays. For example when the average packet delay is 0.2 s, TCP-R gives a 5% throughput improvement over DSACK-TA and when the average packet delay is 0.8 s, TCP-R gives a 10% improvement in throughput performance when compared to DSACK-TA. TCP-R outperforms DSACK-R and SACK for all tested average packet delays. None of the protocols exhibit any timeouts. TCP-R and RN-TCP maintain a zero fast retransmit ratio. DSACK-TA's fast retransmit ratio hovers around zero. DSACK-R and SACK have a higher fast retransmit ratio.

4.3. Varying Delay with Constant Bandwidth

In order to analyze the performance of the protocols over the range of propagation delays exhibited by LEO networks, we compare the throughput performance of the protocols when the propagation delay varied from [40,400] ms. The bandwidth was fixed at 1.5 Mbps. 7% of packets were delayed using uniform distribution [0; 4*P*] where *P* is the propagation delay. The BER was set to 10^7.

As shown in the Figure 6, it is evident that TCP-R outperforms SACK, DSACK-R RN-TCP and DSACK-TA irrespective of the propagation delay. For large propagation delays, TCP-R gives an improved throughput performance. For example when the propagation delay was set to 80 ms, TCP-R gives a 10% throughput improvement over RN-TCP, 20% improvement over DSACK-TA, 71% more than DSACK-R and a two fold throughput performance over SACK. When the propagation delay was set to 200 ms, TCP-R gives a 50% improvement over RN-TCP, 66% improvement over DSACK-TA, three fold improvement over DSACK-R and four times more than SACK.

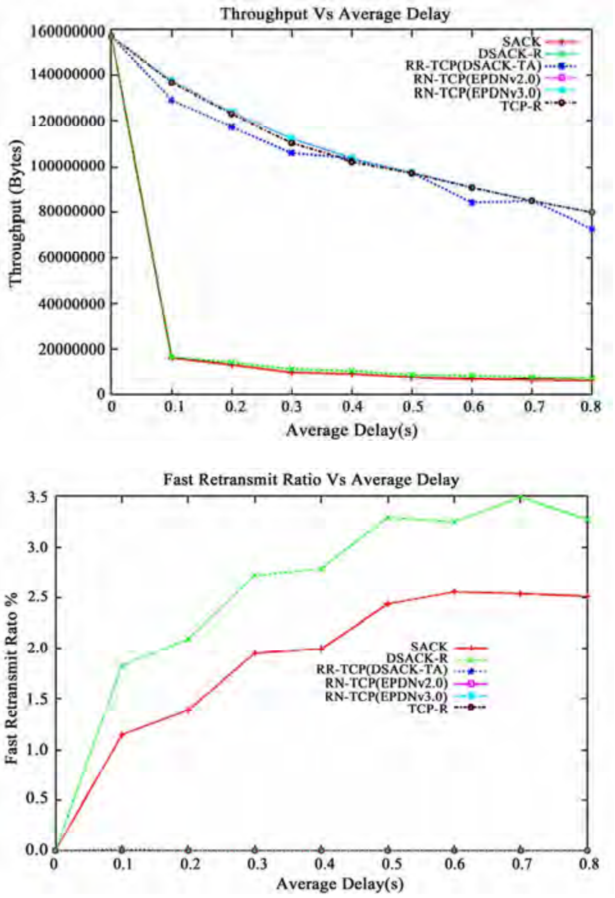


Figure 5. LEO link: Multipath performance, propagation delay of 400 ms.

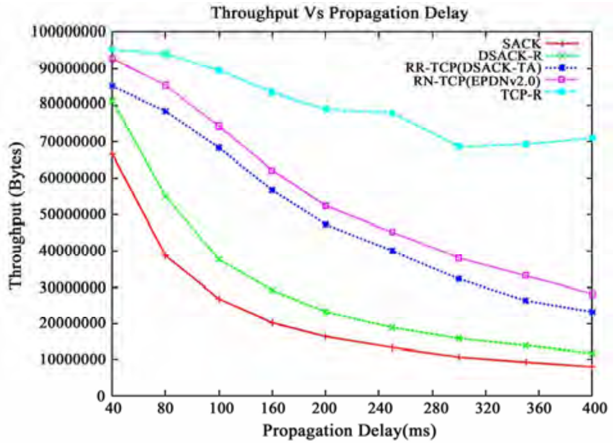


Figure 6. Throughput versus propagation delay.

4.4. Varying Bit Error Rates

In this section, we compare the throughput performance of the protocols for various BERs. The bandwidth was fixed at 1.5 Mbps. The propagation delay was fixed at 300 ms. 5% of packets were delayed using normal distribution with a mean packet delay of 0.6 s and a standard deviation of 0.2 s such that the packets were delayed between 0 and 1.2 s. We tested the protocols for various BERs from a relatively low BER of 10^{13} to a relatively high BER of 10^{5}.

From the Figure 7, it is evident that TCP-R gives a better throughput performance compared to the other protocols for different BERs from 10^{13} to 10^{6}. When the BER is 10^{5}, all protocols give a similar throughput performance. For BERs from 10^{13} to 10^{8}, TCP-R, DSACK-TA, RN-TCP(EPDNv2) and DSACK-R have almost a zero timeout ratio, whereas SACK encounters more timeout event and thus has a large timeout ratio. For BERs of 10^{7} and 10^{6}, TCP-R still maintains a zero timeout ratio whereas the other protocols start encountering more timeout events and have a larger timeout ratio compared to TCP-R. When the BER is large as 10^{5}, all protocols undergo lots of timeout events and thus have a large timeout ratio. Even though RN-TCP (EPDNv2) has a large fast retransmit ratio compared to DSACK-TA, DSACK-TA experiences more timeout events compared to the RN-TCP(EPDNv2). Thus RN-TCP(EPDNv2) performs better compared to DSACK-TA. TCP-R has a zero fast retransmit ratio and thus gives an overall improvement in throughput compared to the other protocols.

4.5. Throughput: Packet Drops due to Congestion.

In this section, we compare the throughput performance of the protocols when the link experiences both packet delays and packet drops due to congestion in the network. We also compared the performance of SACK and TCP-R with packet drops only. The propagation delay was set to 300 ms. 4% of the packets were delayed with a mean packet delay of 0.6 s and a standard distribution of 0.2 s such that the packets were delayed between 0 s to 1.2 s. The packet drop rate varied from 0% to 1%. We assumed that there were no loss of packets due to corruption and thus we set the BER to zero.

Figure 8, reveals that the throughput of all the protocols reduce considerably when packets get dropped. When packet drops occur, the throughput of any TCP variant would reduce drastically even when there is no reordering in the network. This is evident from the graph, where the performance of SACK with no delay reduces drastically with increasing packet drops. Moreover, our TCP-R with no delay performs similar to SACK with no delay. Thus, it is clear that given there are no reordering events, TCP-R performs similar to SACK. When packets get dropped, TCP-R performs almost similar to DSACK-TA. RN-TCP is able to achieve a better throughput when compared to TCP-R and DSACK-TA. But when the packet drop rate is increased from 0.4%, the performance of TCP-R, RN-TCP and DSACK-TA are similar. It is to be noted that all protocols experience similar throughput performance when large number of packets get dropped in the network due to congestion in the network.

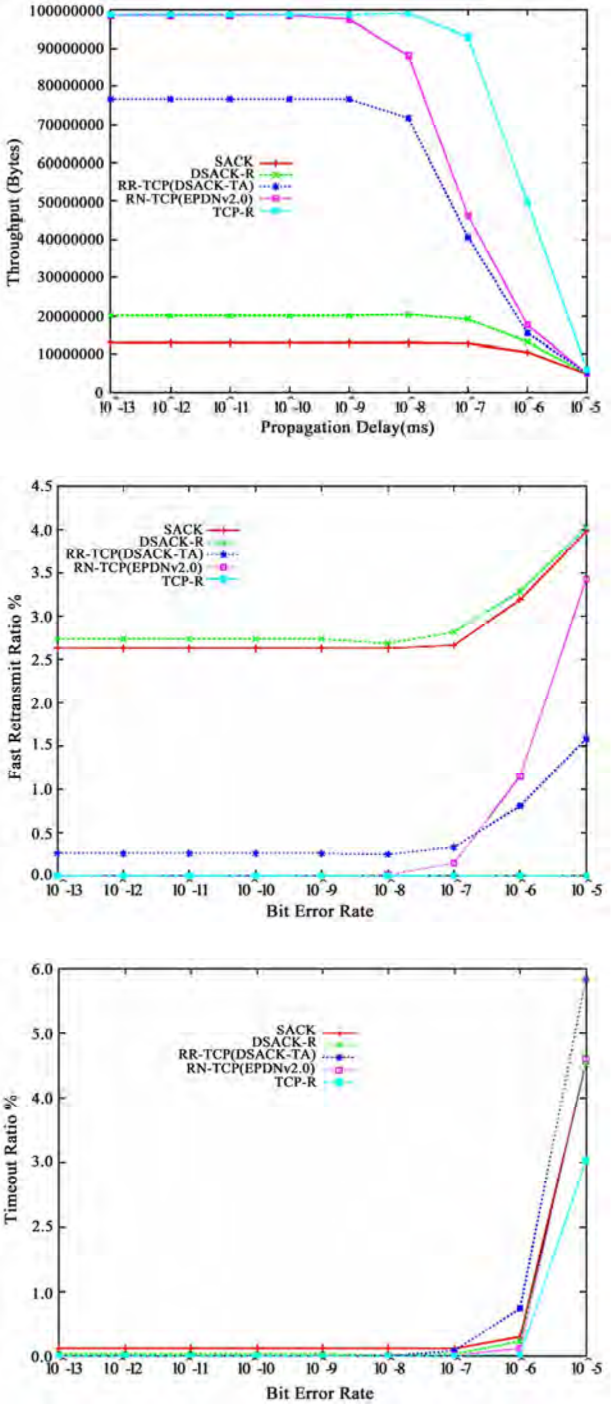


Figure 7. Performance for various BERs.

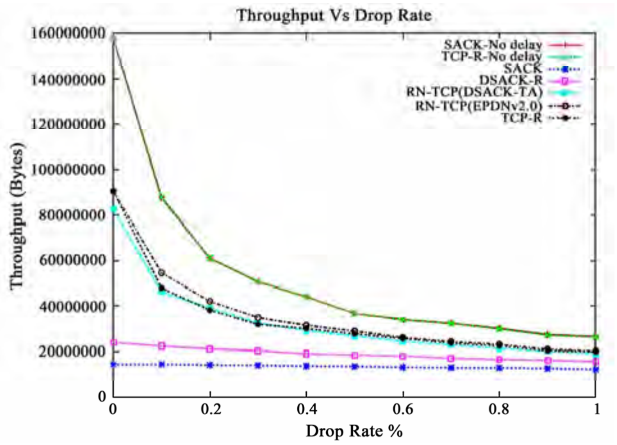


Figure 8. Throughput versus packet drop rate.

4.6. Performance: ECN Enabled Multiple FTP Flows.

In this section, we conducted a simulation on the same topology as in Figure 1, but used RED queues instead of *drop-tail* queues. The gateways were ECN enabled with a queue size set to 300 packets. The link delay was set to 300 ms. To introduce severe packet delays, we used a mean of *yP* s (*P* is the propagation delay) and standard deviation of *y/3 P* s, such that the delay introduced varied from 0 to 2*yP* s. The packet delay rate was fixed at 4%. We varied the value of *y* from 1.0 to 8.0. The BER was set to 10^{i6}. The total number of flows traversing the network was increased to fifty flows, in which the sender was configured to send ten SACK flows, ten DSACK-R flows, ten DSACK-TA flows, ten RN-TCP flows and ten TCP-R flows. All the flows were enabled with ECN. However, when the number of flows id increased, congestion will be caused in the bottleneck queue causing large number of packet drops. The graphs in Figure 9 present the results of the average throughput of SACK, DSACK-R, DSACK-TA, RN-TCP and TCP-R flows. It can be seen from the graph, that for all average packet delays from 0.3 s to 2.4 s, TCP-R outperforms the other protocols. For example, when the average packet delay is 0.3 s, TCP-R gives a 16% improvement over RN-TCP, 29% improvement over DSACK-TA, 36% improvement over DSACK-R and a 62% improvement ver SACK. Similarly when the average packet delay is 2.4 s, TCP-R gives a 26% improvement over RN-TCP, 29% improvement over DSACK-TA, 45% improvement over DSACK-R and a 77% improvement over SACK.

4.7. Conservative RTO Estimation

DSACK-TA and RN-TCP avoids sampling RTTs for all packets that have been retransmitted by timeouts or fast retransmit, in accordance with Karn's algorithm [9], since the sender cannot infer whether an ACK matches an original transmission or the retransmission of a data packet. In the case of paths that predominantly delays packets it is the delayed packets that are most likely to provoke retransmissions. Thus they are not included hile estimating the RTO value. This produces RTO estimates that are too short leading to false timeouts. This is evident from the poor throughput performance achieved by RN-TCP and DSACK-TA for large average packet delays in the previous sections. Thus it is imperative to implement a conservative RTO sampling algorithm that considers both transmission and retransmission of data packets. DSACK-TAES is a combination of DSACK-TA and a conservative RTO sampling algorithm and [19] proves that for large packet delays, DSACK-TAES performs much better compared to DSACK-TA. In order to verify the performance of DSACK-TAES with our proposed protocols, we also ensure that our RN-TCP algorithm is also enhanced with this new RTO sampling algorithm.

When packets are falsely retransmitted either by fast retransmit or by a timeout, two ACKs return, the second of which is a DSACK. When the sender receives both these ACKs, it can calculate the RTTs experienced by the packets by pairing the first ACK with the first transmission, and the second ACK with the second transmission. It can then calculate the time elapsed for each and take the mean of these two values as a single RTT sample for the RTO estimator. The scoreboard data structure used by SACK can hold this time information associated with the packet's transmission and retransmission. Incase no DSACK arrives at the sender, then either the original or retransmitted packet was lost due to packet drops, and we do not sample that packets RTT. This complies with the Karns Algorithm for retransmissions caused by packet drops, but includes additional RTT samples for retransmissions caused by packet delays. We term this RTT sampling extension to RN-TCP as RN-TCP-ES

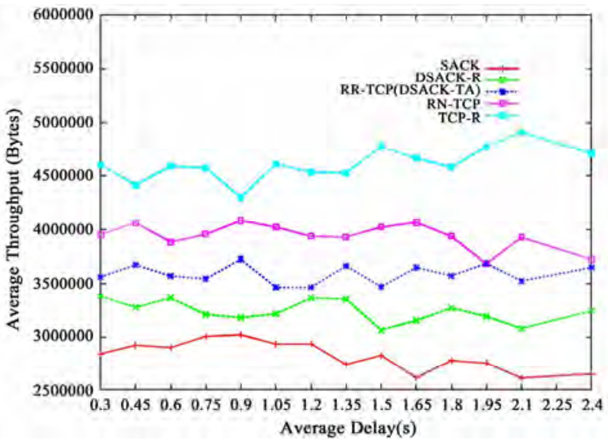


Figure 9. Average throughput versus average packet delay.

(Reorder Notifying TCP with Enhanced RTT Sampling).

4.7.1. Throughput: Large Reordering Delays (Scenario 1)

In this section, we analyze the throughput performance of TCP-R over RN-TCP, RN-TCP-ES, DSACK-TA and DSACK-TAES when packets experience both reordering due to link level retransmissions and packet losses due to corruption. In order to drop packets based on corruption, the BER was set to 10^6. Figure 10 presents the results of the simulations when the propagation delay was set to 300 ms. To introduce severe packet delays, we used a mean packet delay of *yP* s (*P* is the propagation delay) and standard deviation of *y/*3 *P* s, such that the delay introduced varied from 0 to 2*yP* s. The packet delay rate was fixed at 7%. We varied the value of *y* from 1.0 to 7.0. As we increase *y*, the RTT of the packet that gets delayed exceeds the one second minimum RTO causing large number of timeouts.

For all tested average packet delays from 0.3 s to 0.9 s, the performance of RN-TCP-ES and DSACK-TAES is almost similar to RN-TCP and DSACK-TA respectively. When the average packet delay exceeds the one second minimum RTO, the performance of RN-TCP and DSACK-TA begin to decrease whereas RN-TCP-ES and DSACK-TAES maintain a steady throughput. For all tested average packet delays, TCP-R outperforms the other protocols. For example, when the average packet delay is 0.9 s, TCP-R gives a three fold throughput improvement over the other protocols. Similarly, when the average packet delay is 1.8 s, TCP-R gives a 62% throughput improvement over RN-TCP-ES, 81% throughput improvement over RN-TCP, 77% throughput improvement over DSACK-TAES and almost a 86% improvement over DSACK-TA. TCP-R maintains a low timeout ratio and a zero fast retransmit ratio compared to the other protocols. Even though RN-TCP-ES has a higher fast retransmit ratio compared to DSACK-TAES, RN-TCP-ES has a lower timeout ratio compared to DSACK-TAES leading to a better throughput improvement.

4.7.2. Throughput: Large Reordering Delays (Scenario 2)

In this section, we analyze the throughput performance of TCP-R over RN-TCP, RN-TCP-ES, DSACK-TA and DSACK-TAES when packets experience reordering due to link level retransmissions and no packets get dropped due to corruption. The packet delay rate was fixed at 7%. We varied the value of *y* from 1.0 to 7.0.

It can be seen from the Figure 11, that when the average packet delay is less than 1.0 s, the throughput performance of RN-TCP-ES and DSACK-TAES is similar to RN-TCP and DSACK-TA. When the average packet delay is more than 1.0 s, the throughput of RN-TCP and DSACK-TA drop rapidly whereas the throughput of RN-TCP-ES and DSACK-TAES is steady. It is evident from the timeout ratio graph, that both RN-TCP and DSACK-TA have a large timeout ratio unlike RN-TCP-ES and DSACK-TAES which have timeout ratio that hovers around zero. TCP-R performs much better compared to the other protocols. For example, when the average packet delay is 0.9 s, TCP-R gives a 12% throughput performance over RN-TCP and RN-TCP-ES.

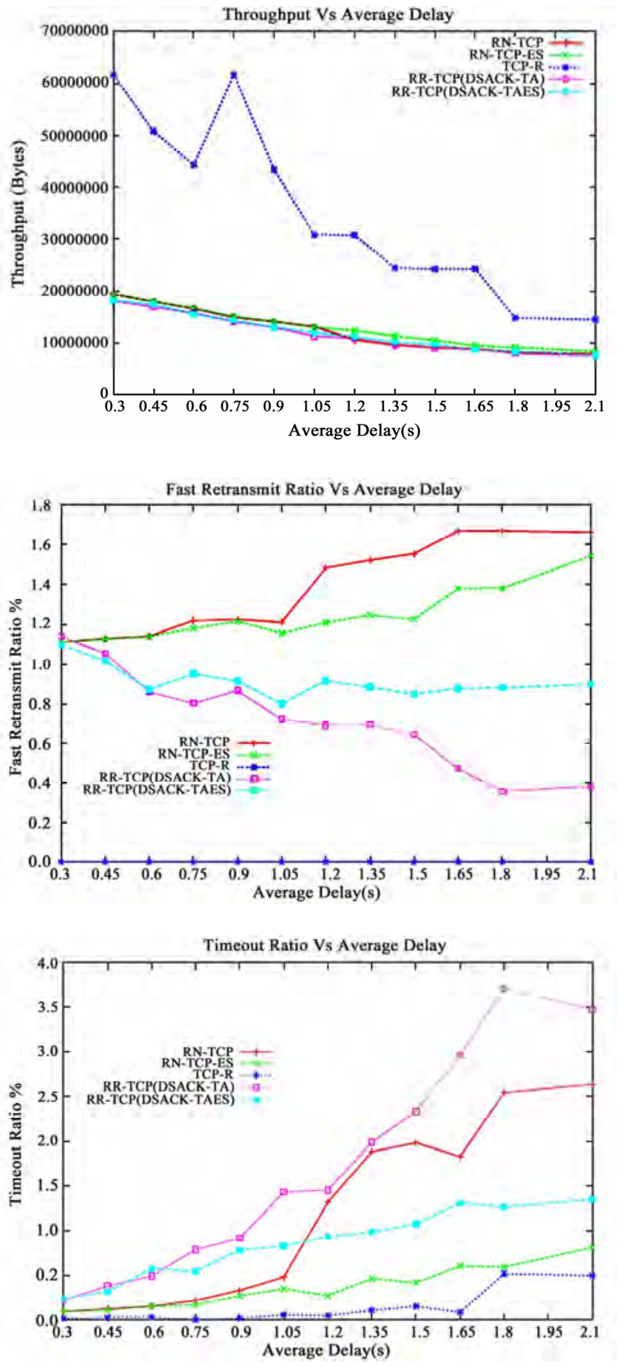


Figure 10. Propagation delay of 300 ms, BER of 10^6.

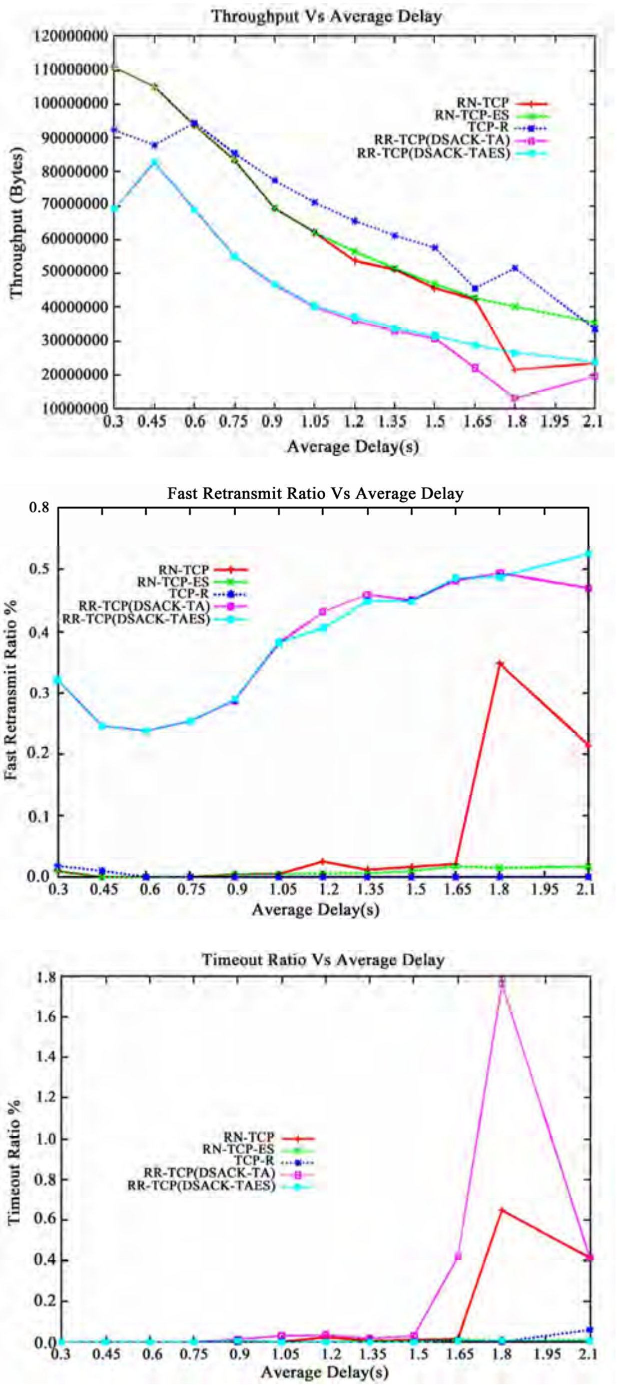


Figure 11. Propagation delay of 300 ms.

Moreover TCP-R gives a staggering 65% improvement over RR-TCP(both DSACK-TA and DSA- CK-TAES). When the average packet delay is 1.65 s, TCP-R gives a 7% throughput improvement over RN-TCP-ES, 8% throughput improvement over RN-TCP, 58% throughput improvement over DSACK-TAES and almost a two fold improvement over DSACK-TA. TCP-R has a zero fast retransmit ratio and almost a zero timeout ratio.

5. Conclusions and Future Work

In this paper, we proposed a solution that allows the TCP sender to distinguish whether a packet has been lost or reordered in the satellite network and perform actions accordingly. This was done by maintaining information about dropped packets (only due to congestion in the network) in the gateway and using this information to notify the sender, whether the packet has been dropped or reordered/corrupted in the network. We termed this mechanism as Explicit Packet Drop Notification (EPDNv3). We also proposed an extension to SACK protocol called Robust TCP (TCP-R). If the TCP-R sender assumes the packets to be reordered or corrupted in the network, it immediately retransmits the packet after receiving three duplicate ACKs and enters our modified fast recovery mechanism where the procedure of reducing the *ssthresh* and the *cwnd* are bypassed i.e. we do not reduce the *ssthresh* and the *cwnd*. We compared TCP-R with other protocols namely SACK, DSACK-R, RR-TCP and RN-TCP and showed TCP-R gave performance improvement over these protocols. We believe the gateway could be modified to send the dropped information in an ICMP message to the sender. This requires further study and testing. Further simulations and testing needs to be carried out to find the efficiency of the protocol when there is an incremental deployment i.e. when there are some routers in a network which have not been upgraded to use our mechanism.

6. References

[1] J. Postel, “Transmission control protocol,” RFC 793, 1981.

[2] V. Jacobson, “Symposium proceedings on communications architectures and protocols,” California, pp. 314–329, 1988.

[3] G. Maral, “VSAT networks,” J. Wiley and Sons, 1995.

[4] R. J. Leopold, “Low-earth orbit global cellular communications network,” Proceedings of ICC’91, pp. 1108–1111, 1991.

[5] L. Wood, G. Pavlou, and B. G. Evans, “Effects on TCP of routing strategies in satellite constellations,” IEEE Communications Magazine, special issue on Satellite-Based Internet Technology and Services, Vol. 39, No. 3, pp. 172–181, 2001.

[6] C. Ward, H. Choi, and T. Hain, “A data link control protocol for LEO satellite networks providing a reliable datagram service,” IEEE/ACM Transactions on Networking, Vol. 3, No. 1, 91103, Feb. 1995.

[7] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. H. Katz, “A comparison of mechanisms for improving TCP performance over wireless links,” IEEE/ACM Transactions on Networking (TON) Archive, Vol. 5 , No. 6, pp. 756–769, December 1997.

[8] H. Balakrishnan and R. H. Katz, “Explicit loss notification and wireless web performance,” Proceedings of the IEEE Globecom Internet Mini-Conference, Sydney, Australia.

[9] R. Krishnan, P. G. Sterbenz, W. M. Eddy, C. Partridge, and M. Allman, “Explicit transport error notification for error-prone wireless and satellite networks,” Computer Networks journal, Elsevier, 2004.

[10] A. Bakre and B. R. Badrinath, “I-TCP: Indirect TCP for mobile hosts,” Proceedings of the 15th International Conference on Distributed Computing Systems (ICDCS), 1995.

[11] I. F. Akyildiz, G. Morabito, and S. Palazzo, “TCP-peach: A new congestion control scheme for satellite IP networks,” IEEE/ACM Transactions on Networking, Vol. 9, No. 3, pp. 307–321, 2001.

[12] S. Floyd, J. Mahdavi, M. Mathis, and M. Podolsky, “An extension to the selective acknowledgement (SACK) option for TCP,” RFC 2883, 2000.

[13] M. Zhang, B. Karp, S. Floyd, L. Peterson, “RR-TCP: A reordering robust TCP with DSACK,” 11th IEEE International Conference on Network Protocols (ICNP’03), Georgia, 2003.

[14] A. Sathiaseelan and T. Radzik, “Reorder notifying TCP (RN-TCP) with explicit packet drop notification (EPDN),” International Journal of Communication Systems, Wiley, Vol. 19, No. 6, pp. 659–678, 2005.

[15] T. R. Henderson and R. H. Katz, “Transport protocols for internet-compatible satellite networks,” IEEE Journal on Selected Areas in Communications, Vol. 17, No. 2, pp. 345–359, February 1999.

Int. J. Communications, Network and System Sciences, **2009, 7, 687-694**
doi:10.4236/ijcns.2009.27079 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

Augmented Reality for Realistic Simulation Using Improved Snake and Picking Algorithm by Proportional Relational Expression

JeongHee CHA[1], GyeYoung KIM[2], HyungIl CHOI[3]
[1]*Division of Computer Information, School of Computing, School of Media,*
[2]*BaekSeok Culture University,Anseo Dong, DongNam Gu, Cheonan, Korea,*
[3]*Soongsil University , Sangdo 5 Dong , DongJak Gu, Seoul, Korea,*
Email: pelly@bscu.ac.kr, {gykim11,hic}@ssu.ac.kr

Received May 8, 2009; *revised June* 29, 2009; *accepted August* 17, 2009

ABSTRACT

In realistic simulation of mobile Augmented Reality, essential point is how to best depict occluded area in such a way that the user can correctly infer the depth relationships between real and virtual objects. However, if the constructed 3D map of real world is not accurate or the density is not sufficient to estimate the object boundary, it is very difficult to determine the occluded area. In order to solve this problem, this paper proposes a new method for calculating the occlusion area using the improved snake algorithm and picking algorithm by the proportional relational expression. First, we generated the wireframe by the DEM in the experimental region and mapped to CCD real image using visual clues. And then, we calculated the 3D information at the point where occlusion problem for a moving virtual target by the proposed method. Experimental results show the validity of the proposed approach under the environment in which partial occlusions occur.

Keywords: Snake, Picking, DEM, Occlusion, AR, Simulation

1. Introduction

This paper studied on the development of a realistic simulated training model through the display of virtual targets in the input images of CCD camera mounted in a vehicle and the determination of occlusion areas generated from the creation and movement of virtual objects through a movement path according to a scenario. Augmented reality has three general characteristics: image registration, interaction, and real time [1,2]. Image registration refers to the matching of the locations of the real world object that user watch and the related virtual object. Interaction implies that the combination of virtual objects and the objects in real images must be harmonized with surrounding environment in a realistic manner, and refers to the determination of occlusion areas according to the changed location or line of sight of the observer or the re-rendering of virtual objects after detection of collisions. Real time refers to the real time image registration and interaction. A key problem in the AR field is how to best depict occluded objects in such a way that the viewer can correctly infer the depth relationships between different real and virtual objects. For occlusion processing such as the hiding of farther virtual objects by closer real objects and the covering of objects in real images by other virtual objects, the two worlds must be accurately coordinated and then the depth of the actual scene must be compared with the depth of virtual objects [3,4]. But if the accuracy or density of the created map is insufficient to estimate the boundary of occlusion area, it is difficult to determine the occlusion area. To solve this problem, first, we created a 3D wireframe using the DEM of the experiment area and then coordinate this to CCD camera images using visual clues. Second, to solve the problem of occlusion by accurately estimating the boundary regardless of the density of map, this paper also proposed a method to obtain the reference 3D information of the occlusion points using the improved Snake algorithm and the Picking algorithm and then to infer the 3D information of other boundaries using the proportional relations between 2D and 3D DEM. Third, for improving processing speed, we suggest a method by comparing the MER (Minimum Enclosing Rectangle) area of the real object in the camera's angle of view and the MER of the virtual target.

We briefly review related work in Section 2. In Section 3, we outline the framework of our proposed algorithm. The methodology of Wireframe modeling, improved snake algorithm for extracting image boundary and picking algorithm for acquiring 3D information are explained in Section 4. Section 5, we show the experimental results and the validity of the proposed approach. Finally, in Section 6 we discuss drawbacks of the algorithm and propose possible future work.

2. Previous Work

A basic design decision in building an AR system is how to accomplish the combining of real and virtual. Toward this purpose, many researchers make efforts to minimize virtual objects registration error and to increase the realness of virtual objects [5]. Drastic and Milgram list a number of cues that a user may use to interpret depth, including image resolution and clarity, contrast and luminance, occlusion, depth of field and accommodation [6]. We can use one of two technologies to see the real world in AR, one is optical see-through and the other is video see-through. These technologies can present occluded objects, and each has a variety of challenges [7]. Blurring also can help compensate for depth perception errors [8]. Koch uses computer vision techniques to infer dense and accurate depth maps from image pairs, and uses this information to construct 3D graphical representations of the restricted static environment [9]. Several authors observe that providing correct occlusion of real objects by virtual objects requires a scene model. Correct occlusion relationships do not necessarily need to be displayed at all pixels. The purpose of many applications is to see through real object. We introduce here mobile vehicle-mounted display system capable of resolving occlusion between real and virtual objects. We restricted the real environment to some area and constructed that area's scene Model using 3D information. The heart of our system is boundary extraction algorithm and 3D information inference algorithm of object boundary. Figure 1 shows our monitor-based training vehicle configuration.

In our experimental vehicle configurations, we send steering, acceleration, brake data to car driving controller through Bluetooth using remote car controller. Vehicle can be controlled by transmitted data and we can get feedback of present car location data by mounted sensor system. RS232 communicator is interface between vehicle driving controller and sensor fusion system. And it receives instructions from sensor system. CCD camera views the environment. The camera may be static or mobile. In mobile case, the camera might move around by being attached to a vehicle, with their locations tracked by GPS and INS. The image of real world and the virtual images generated by a scene generator are combined.

3. System Flow Description

Figure 2 outlines the framework of our proposed system. System has two stages. First stage is environment map construction stage. It consists of registration of two world using visual clues and object contour extraction and acquiring 3D information. Second stage is virtual object rendering stage. It has creation of virtual target path and selection of candidate occlusion object and occlusion processing.

4. Methodology

4.1. Formation of Wireframe Using DEM and Registration with Real Images Using Visual Clues

The topographical information DEM (Digital Elevation

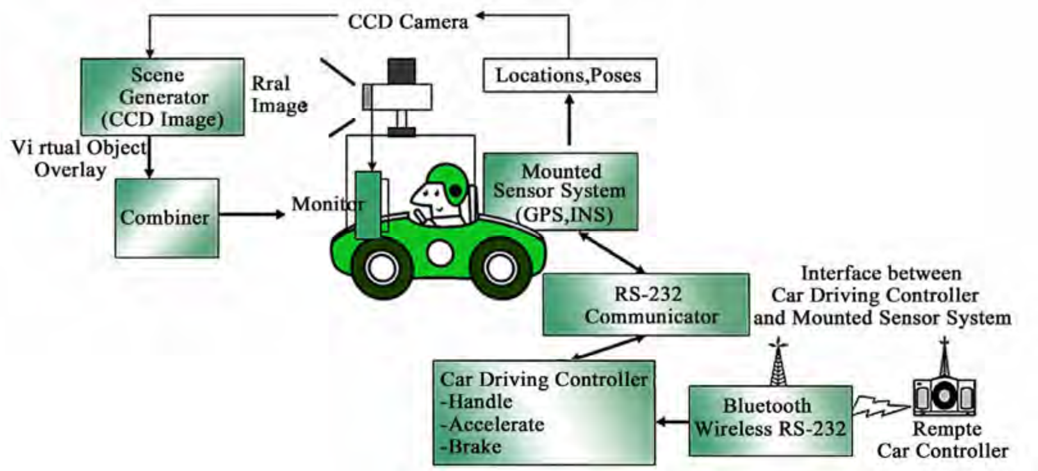


Figure 1. Training vehicle configuration.

Model) is used to map the real world coordinates to each point of the 2D CCD image. DEM has information on the latitude and longitude coordinates expressed in X and Y and heights in fixed interval. The DEM used for this experiment is a grid-type DEM which had been produced to have the height information for 2D coordinates in 1M interval for the limited experiment area of 300 m x 300 m. The DEM data are read to create a mesh with the vertexes of each rectangle and a wireframe with 3D depth information [10,11] as Figure 3. This is overlaid on the sensor image to check the coordination, and visual clues are used to move the image to up, down, left or right as shown in Figure 4, thus reducing error. Based on this initial coordinated location, the location changes by movement of vehicles were frequently updated using GPS (Global Positioning System) and INS (Inertial Navigation System).

4.2. Extracting the Contour of Objects in Image by Enhanced Snake Algorithm

4.2.1. Edge Map Using Gradient Vector Flow

The Snake algorithm [12,13] is a method of finding the outline of an object by repeatedly moving to the direction of minimizing energy function from the snake vertex input by user. But existing snake algorithm cannot accurately extract the contour information when the object form is complex because as shown in Figure 5, the direction of the energy function appears as a composite vector of the current, previous, and the next snake points, and shrinks toward the center of these points. To solve this problem, this paper proposes a method to form an edge map using the Gradient Vector Flow (GVF) algorithm [14,15,16], and add a new energy term that indicates the distance between the searched edge point and snake point so as to extract an accurate contour.

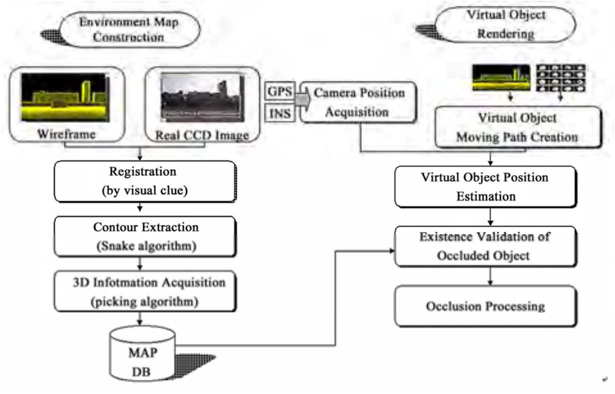

Environment Map Construction
Virtual Object Rendering
Wireframe
Real CCD Image
GPS
INS
Camera Position Acquisition
Virtual Object Moving Path Creation
Registration (by visual clue)
Virtual Object Position Estimation
Contour Extraction (Snake algorithm)
Existence Validation of Occluded Object
3D Infotmation Acquisition (picking algorithm)
Occlusion Processing
MAP DB

Figure 2. Proposed system framework.

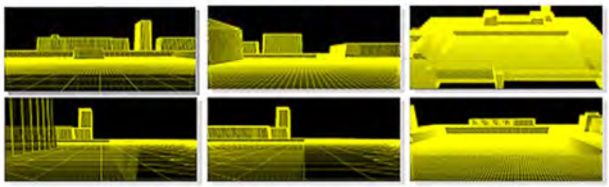


Figure 3. Wireframe creation using DEM.

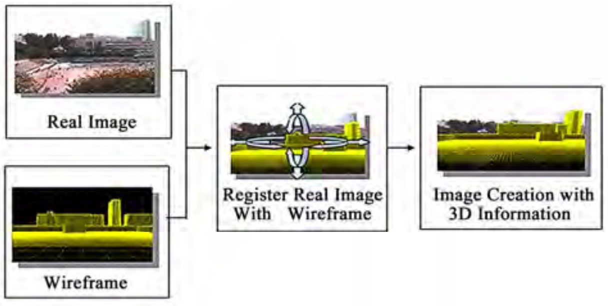

Real Image
Wireframe
Register Real Image With Wireframe
Image Creation with 3D Information

Figure 4. Registration of two worlds using visual clues.

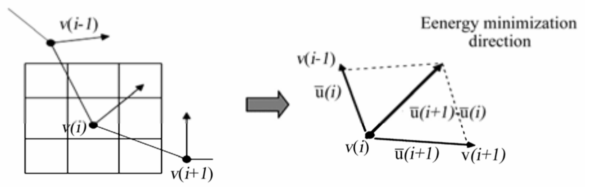


Figure 5. The direction of energy minimization search in snake algorithm.

The GVF algorithm can measure the contour of complex objects using the gradient of edge, and move to the concave contour regardless of initialization. Further, the gradient vector of the edge map has a larger value as it is near edge, and approaches zero as it is farther. This paper uses the edge information of the gradient vector flow to search the proximal edge point, and when there is an edge, adds a new energy term($E_{edge\text{-}distance}$) that shows the dis tance from the reference point to the searched edge as Equation (1). Here, α β, and γ are all set to 1 without exhaustive adjustment .

$$E_{snake} = \int_0^1 E_{continuity}(v(s)) + E_{curvature}(v(s)) + E_{image}(v(s)) + E_{edge-dis\tan ce}(v(s))ds \qquad \textbf{(1)}$$

4.2.2. The Proximal Edge Search Method

Figure 6 shows a proposed proximal edge search method in this paper.

First it searches edge points while rotating around the axis *d* which is the connection between current and previous snake points v_i and v_{i-1}. In other words, if the angle formed by the three points v_i, v_{i-1}, and v_{i-2} is ϕ, to prevent the situation where the axis meets with or passes by v_{i-2} and meets v_i again, it searches the edge point $v_i^{'}$ where the image strength ∇I is greater than the threshold while rotating only by $\frac{\phi}{2}$ and adds a new energy term using the value of the distance *d'* between v_i and $v_i^{'}$ to the existing algorithm. This paper determined the rotation direction for accurate search by assuming the following two facts: First, it was assumed that the initial snake points form a closed curve that encloses the object. Second, it was assumed that the points were arranged sequentially in one direction. The reason for this was because to search proximal edge, it must move inside the contour, but the direction may be wrong due to the diversity of object forms if simply the direction to the object center is set. Figure 7 is an example of setting the rotation direction of the snake points.

4.2.3. Calculation of $E_{edge\text{-}distance}$

Figure 8 is an example of calculating the distance between an arbitrary snake point v_i and the edge $v_i^{'}$ around it. If we surround the arbitrary point v_i with a 9×9 window and assume that its distance with a new edge is $d_{mn}^{'}$, the height and width of the window are s, and the horizontal and vertical positions of the snake point in the window are m and n, the $d_{mn}^{'}$ can be obtained with the Equation (2) by the Euclidean theorem, and the energy term to be added can be defined as the Equation (3) by applying the distance value instead of the brightness value of the image term.

$$d_{mn}^{'} = \sqrt{\left(\frac{2(\left|v_x - v_x^{'}\right| + m) - s + 1}{2}\right)^2 + \left(\frac{2(\left|v_y - v_y^{'}\right| + n) - s + 1}{2}\right)^2} \quad \textbf{(2)}$$

$$\begin{aligned} E_{edge-distance} &= (\left|v_i - v_i^{'}\right| - d_{\min}^{'}) / (d_{\max}^{'} - d_{\min}^{'}) \\ &= (d_{mn}^{'} - d_{min}^{'}) / (d_{max}^{'} - d_{min}^{'}) \end{aligned} \quad \textbf{(3)}$$

Added new energy term $E_{edge\text{-}distance}$ is expressed together with continuity and curvature energy terms in Figure 9. When only the two terms of the existing algorithm were considered, the minimum point of energy was in line 3, column 5, but the location changed to line 4, column 6 when the energy value in consideration of the distance between proximal edges was included. In conclusion, the flow of the enhanced snake energy function to which the proximal edge energy function is added can extract the edge exactly in complex situations by approaching the edge more closely.

Table 1 shows the pseudo codes of the proposed algorithm using proximal edge search method.

4.3. Acquisition of 3D Information Using the Picking Algorithm

In order to acquire the 3D information of the extracted vertexes, this paper used the Picking algorithm which is a well-known 3D graphics technique [17]. It finds the collision point with the 3D wireframe created by DEM that corresponds to the points in 2D image and provides the 3D information of the points. The picking search point is the lowest point of the vertexes of the objects

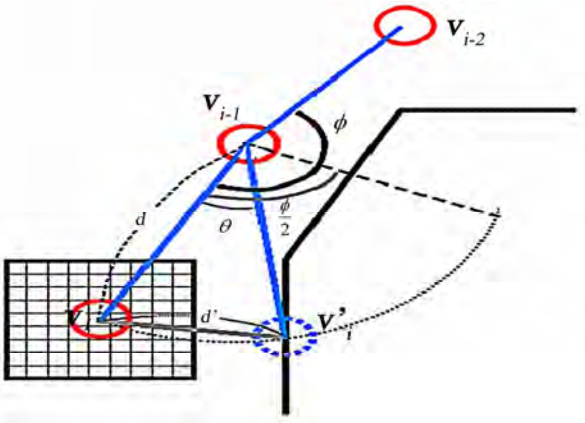


Figure 6. The proximal edge search method.

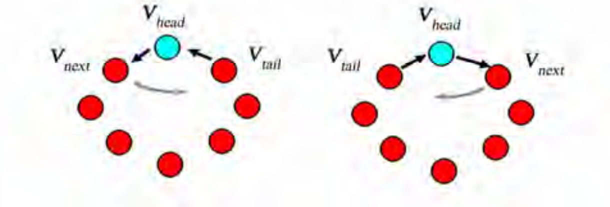


Figure 7. Snake's rotation direction.

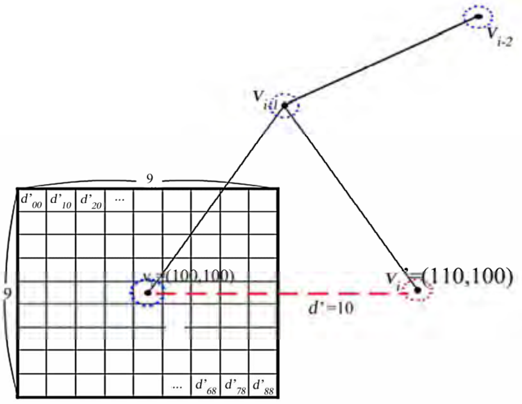
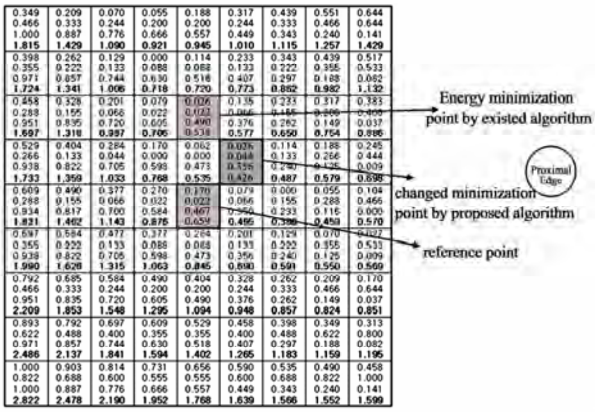


Figure 8. Distance between a point of snake and edge.



Energy minimization point by existed algorithm
Proximal Edge
changed minimization point by proposed algorithm
reference point

0.349 0.466 1.000 **1.815**	0.209 0.333 0.887 **1.429**	0.070 0.244 0.776 **1.090**	0.055 0.200 0.666 **0.921**	0.188 0.200 0.557 **0.945**	0.317 0.244 0.449 **1.010**	0.439 0.333 0.343 **1.115**	0.551 0.466 0.240 **1.257**	0.644 0.644 0.141 **1.429**
0.398 0.355 0.971 **1.724**	0.262 0.222 0.857 **1.341**	0.129 0.133 0.744 **1.006**	0.000 0.088 0.630 **0.718**	0.114 0.088 0.518 **0.720**	0.233 0.133 0.407 **0.773**	0.343 0.222 0.297 **0.862**	0.439 0.355 0.188 **0.982**	0.517 0.533 0.082 **1.132**
0.458 0.288 0.951 **1.697**	0.328 0.155 0.835 **1.318**	0.201 0.066 0.720 **0.987**	0.079 0.022 0.605 **0.706**	0.026 0.022 0.490 **0.538**	0.135 0.066 0.376 **0.577**	0.233 0.155 0.262 **0.650**	0.317 0.288 0.149 **0.754**	0.383 0.466 0.037 **0.886**
0.529 0.266 0.938 **1.733**	0.404 0.133 0.822 **1.359**	0.284 0.044 0.705 **1.033**	0.170 0.000 0.598 **0.768**	0.062 0.000 0.473 **0.535**	0.026 0.044 0.356 **0.426**	0.114 0.133 0.240 **0.487**	0.188 0.266 0.125 **0.579**	0.245 0.444 0.009 **0.698**
0.609 0.288 0.934 **1.831**	0.490 0.155 0.817 **1.462**	0.377 0.066 0.700 **1.143**	0.270 0.022 0.584 **0.876**	0.170 0.022 0.467 **0.659**	0.079 0.066 0.350 **0.495**	0.000 0.155 0.233 **0.388**	0.055 0.288 0.115 **0.458**	0.104 0.466 0.000 **0.570**
0.697 0.355 0.938 **1.990**	0.584 0.222 0.822 **1.628**	0.477 0.133 0.705 **1.315**	0.377 0.088 0.598 **1.063**	0.284 0.088 0.473 **0.845**	0.201 0.133 0.356 **0.690**	0.129 0.222 0.240 **0.591**	0.070 0.355 0.125 **0.550**	0.027 0.533 0.009 **0.569**
0.792 0.466 0.951 **2.209**	0.685 0.333 0.835 **1.853**	0.584 0.244 0.720 **1.548**	0.490 0.200 0.605 **1.295**	0.404 0.200 0.490 **1.094**	0.328 0.244 0.376 **0.948**	0.262 0.333 0.262 **0.857**	0.209 0.466 0.149 **0.824**	0.170 0.644 0.037 **0.851**
0.893 0.622 0.971 **2.486**	0.792 0.488 0.857 **2.137**	0.697 0.400 0.744 **1.841**	0.609 0.355 0.630 **1.594**	0.529 0.355 0.518 **1.402**	0.458 0.400 0.407 **1.265**	0.398 0.488 0.297 **1.183**	0.349 0.622 0.188 **1.159**	0.313 0.800 0.082 **1.195**
1.000 0.822 1.000 **2.822**	0.903 0.688 0.887 **2.478**	0.814 0.600 0.776 **2.190**	0.731 0.555 0.666 **1.952**	0.656 0.555 0.557 **1.768**	0.590 0.600 0.449 **1.639**	0.535 0.688 0.343 **1.566**	0.490 0.822 0.240 **1.552**	0.458 1.000 0.141 **1.599**

Figure 9. Changed energy minimization point by proposed algorithm.

Table 1. Pseudo codes of proposed algorithm.

```
Do    /* loop for proposed algorithm */
  For i=0 to n-1 /* n is number of snake points */
      Angle = $(\angle v_{i-2} v_{i-1} v_i)/2$ ; /* search limit determination */
      for j = 0 to Angle
      if $v_i^{'}$ is Edge then bFind = true;
      $E_{\min} = BIG$ ;
      for j = 0 to m-1   /* m is size of neighborhood */
         if bFind is True then
            $E_j = E_{cont,j} + E_{curv,j} + E_{image,j} + E_{edge-distance,j}$ ;
         Else $E_j = E_{cont,j} + E_{curv,j} + E_{image,j}$ ;
         If $E_j \langle E_{min}$ then
            $E_{min} = E_j$ ;
            $j_{\min} = j$ ;
      move point $v_i$ to location $j_{min}$ ;
      if ( $j_{min}$ != current location) cnt_movedpoint += 1;
      /* following process determines where to allow corners */
      For i=0 to n-1
            $c_i = \left| \overrightarrow{u_i} / \left| \overrightarrow{u_i} \right| - \overrightarrow{u_{i+1}} / \left| \overrightarrow{u_{i+1}} \right| \right|^2$ ;
      For i=0 to n-1
      If $c_i \langle c_{i-1}$ and $c_i \rangle c_{i+1}$ ;
      /* if c_i(curvature) is larger than neighborhood's */
        and $c_i \rangle$ threshold1 ;/* if c_i is larger than threshold1 */
        and $mag\ (v_i)\rangle$ threshold2;
         /* if edge strength is larger than threshold2 */
        Then $\beta_i$ =0; /* relax curvature at point i */
        Until cnt_movedpoint < threshold3;
```

extracted from the 2D image. The screen coordinate system that is a rectangular area indicating a figure that has been projection transformed in the 3D image rendering process must be converted to the viewport coordinate system in which the actual 3D topography exists to pick the coordinate system where the mouse is actually present. First, the conversion matrix to convert viewport to screen is used to obtain the conversion formula from 2D screen to 3D projection window, and then the ray of light is lengthened gradually from the projection window to the ground surface to obtain the collision point between the point to search and the ground surface. Figure 10 is an example of picking the collision point between the ray of light and DEM. The lowest point of the occlusion area indicated by an arrow is the reference point to search, and this becomes the actual position value of 2D image in a 3D space.

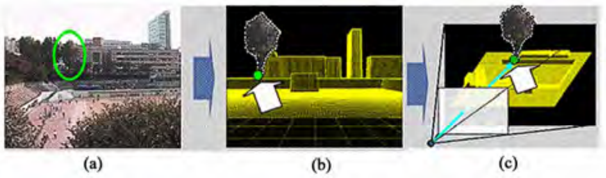


Figure 10. 3D information extraction using collision point of ray and DEM. (a)occlusion candidate (b)matching ref.point and DEM (c)3D information extraction.

4.3.1. Creation of 3D Information Using Proportional Relational Expression

The collision point, or reference point, has 3D coordinates in DEM, but other vertexes of the snake indicated as object outline cannot obtain 3D coordinates because they don't have a collision point. Therefore, this paper suggested obtaining a proportional relation between 2D image and 3D DEM using the collision reference point and then obtaining the 3D coordinates of another vertex. Figure 11 shows the proportional relation between 2D and 3D vertexes. In Figure 11, S_m is center of screen, S_B is reference point of snake vertexes (lowest point), $\Delta S_B = (\Delta S_{x_B}, \Delta S_{y_B})$ is a distance from S_m to S_B, S_k is a optional point except reference point of snake vertexes, $\Delta S_k = (\Delta S_{x_k}, \Delta S_{y_k})$ is a distance from S_m to S_k. P_m is a projection point of straight line of P_B in 3D, which is through the center of screen. P_B is a 3D correspondence point of S_B, $\Delta P_B = (\Delta P_{x_B}, \Delta P_{y_B}, \Delta P_{z_B})$, P_k is a optional point except reference point, $\Delta P_k = (\Delta P_{x_k}, \Delta P_{y_k}, \Delta P_{z_k})$, $t = \overrightarrow{P_O P_B}$, $t_m = \overrightarrow{P_o P_m}$, $\theta_B : \angle tt^{'}$, $\phi_B : \angle t^{'} t_m$. $t^{'}$: a projected vector of t to xz plane.

To get p_m in 3D that passes the center of the screen using the coordinates of the reference point obtained above, $t^{'}$ must be obtained first. As the t value is given by the picking ray, the given t value and y_B are used to get θ_B and $t^{'}$ is obtained using thisθ_B in Expression (4).

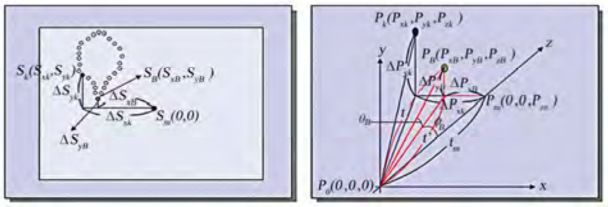


Figure 11. Proportional relation of the vertexes in 2D and 3D.

$$\theta_B = sin^{-1}(\frac{\Delta P_{y_B}}{t}), t^{'} = |t_B| cos(\theta_B), (t^{'} = |t^{'}|) \qquad \textbf{(4)}$$

To get t_m, Φ_B which is the angle between $t'$and t_m is obtained, t_m can be obtained using Φ_B from Expression (5)

$$\varphi_B = tan^{-1}(\frac{\Delta P_{x_B}}{t^{'}}), t^{'} = |t_m| cos(\varphi_B) |t_m| = \frac{|t^{'}|}{cos(\varphi_B)} t_m = |t_m| \qquad \textbf{(5)}$$

Because t_m=PZ_m, we can define P_m=(0,0, t_m).

Now, we can present the relation between the 2D screen view in Figure 11 and the 3D space coordinates, and this can be used to get P_k, which corresponds to the 2D optional snake vertex.

$$\begin{aligned}
&\Delta S_B : \Delta P_B = \Delta S_k : \Delta P_k, \\
&\Delta S_{x_B} : \Delta P_{x_B} = \Delta S_{xk} : \Delta P_{x_k} \\
&\Delta P_{x_k} = \frac{\Delta P_{x_B} \times \Delta S_{x_k}}{\Delta S_{x_B}}, \Delta S_{y_B} : \Delta P_{y_B}, \qquad \textbf{(6)} \\
&\Delta S_{yk} : \Delta P_{y_k}, \Delta P_{y_k} = \frac{\Delta P_{y_B} \times \Delta S_{y_k}}{\Delta S_{y_B}}
\end{aligned}$$

Consequently, we can get $\Delta P_k = (\Delta P_{x_k}, \Delta P_{y_k})$, which is the 3D point corresponding to each snake vertex to search.

4.3.2. Creation of Virtual Target Path and Selection of Candidate Occlusion Objects Using MER (Minimum Enclosing Rectangle)

To test the proposed occlusion-resolving algorithm, we created the movement path of a virtual target, and determined the changes of the direction and shape of the target as well as the 3D position of the target. First, the beginning and end points of the target set by instructor were saved and the angle of these two points was calculated, and the direction and shape of the target were updated in accordance with the change of the angle. Further, the remaining distance was calculated using the speed and time of the target, and the 3D coordinates corresponding to the position after movements were determined. We also suggest a method of improving processing speed by comparing the MER (Minimum Enclosing Rectangle) area of the object in the camera's angle of vision and the MER of the virtual target because the relational operations between all objects extracted from the image for occlusion processing and the virtual target take much time. The MER (Minimum Enclosing Rectangle) of an object refers to the minimum rectangle that can enclose the object and determines the object that has an overlapping area by comparing the objects in the camera image and the MER of the virtual target. In addition, the distance between object and virtual target is obtained using the fact that the determined object and virtual target are placed more or less in a straight line from the camera, and this value was used to determine whether there exists an object between the virtual target and the camera.

5. Experimental Results

Figure 12(up) shows movement path of the virtual target which trainee sets. Also, (down) shows the various virtual targets created to display the targets changing with movement on the image.

Figure 13, Figure 14 compares the search results, accuracy

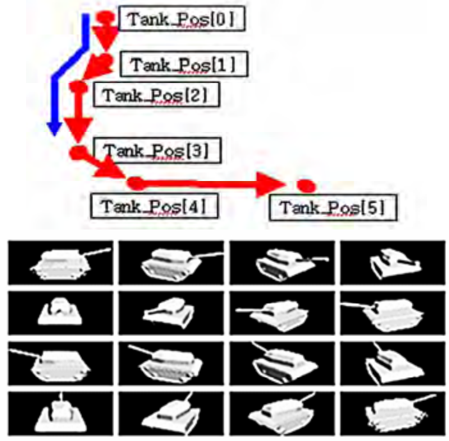


Figure 12. Moving route creation(up) and appearance of virtual object as it moved(down).

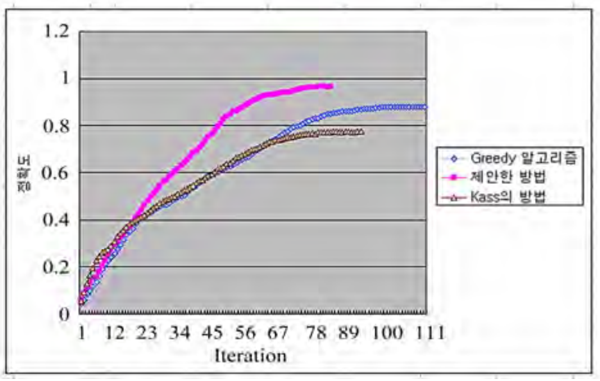


Figure 13. Accuracy comparison(leaf).

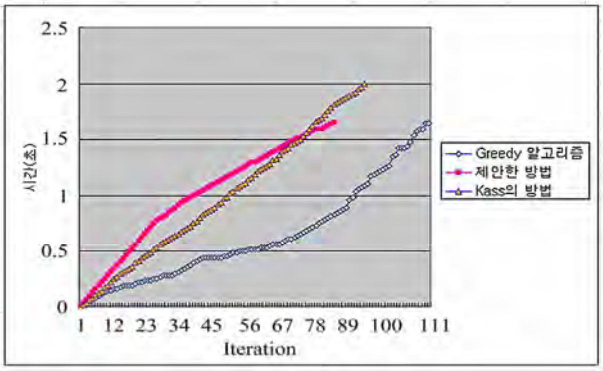


Figure 14. Speedy comparison(leaf).

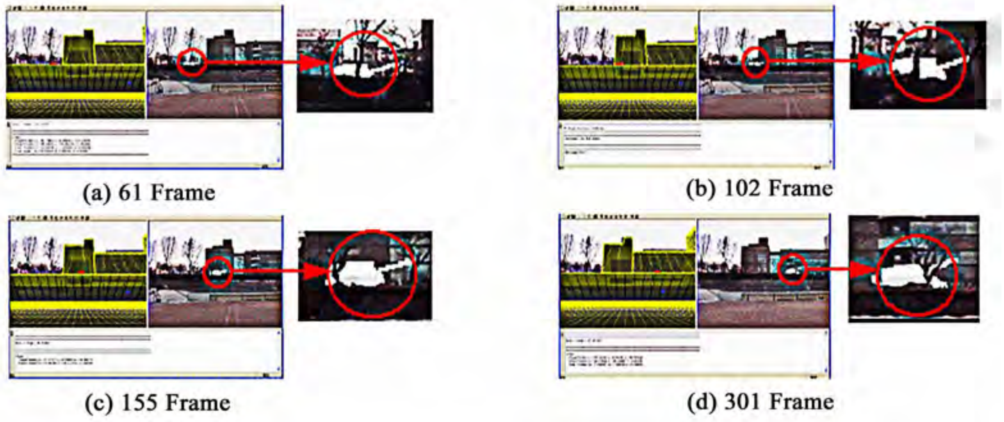


Figure 15. Experimental results of moving and occlusion.

and speed for more complex leaf. As shown in the figures and graphs, we can see that the proposed algo rithm has much higher accuracy and less repetition counts, and the speed is equal to greedy algorithm.

As shown in Figure 13, the proposed algorithm stopped search at the 80th round, and the accuracy was 0.96 while the Kass and greedy algorithms showed the search count 96 and 150 and the accuracy 0.78 and 0.84, respectively. Therefore, we can conclude that the proposed algorithm has higher performance than existing algorithms. The search speed of the proposed algorithm was 1.65 seconds, which is equal level to the greedy algorithms.

Figure 15 shows the virtual images moving along the path by frame. We can see that as the frames increase, it is occluded between the tank and the object.

Table 3 compares between the case of using snake vertexes to select objects in the image to compare with virtual targets and the case of using the proposed MER. With the proposed method, the processing speed decreased by 1.671, which contributed to performance improvement.

6. Conclusions

To efficiently solve the problem of occlusion that occurs when virtual targets are moved along the specified path over an actual image, we created 3D virtual world using DEM and coordinated this using camera images and visual clues. Moreover, the enhanced Snake algorithm and the Picking algorithm were used to extract an object that is close to the original shape to determine the 3D information of the point to be occluded. To increase the occlusion processing speed, this paper also used the method of using the 3D information of the MER area of the object, and proved the validity of the proposed method through experiment. In the future, more research is required on a more accurate extracting method for occlusion area that is robust against illumination as well as on the improvement of operation speed. We also hope to study more in real time environment and to overcome complicated factors that were beyond our control, such as sensor error in the current settings, the brightness difference of same image.

Table 3. Speed comparison.

Method	Total frame	Used object	Speed(sec)	Frame per sec.
Snake vertexes	301	10	112	2.687
MER(proposed)	301	10	67	4.492

7. Acknowledgement

This Work was supported by the Korea Research Foundation Grant (KRF-2006-005-J03801) Funded by Korean Government.

8. References

[1] R. Azuma, “A survey of augmented reality,” in ACM SIGGRAPH’95 Course Note #9-Deveoping Advanced Virtual Reality Applications, August 1995.

[2] O. Bimber and R. Raskar, “Spatial augmented reality: A modern approach to augmented reality,” Siggraph, Los Angeles, USA, 2005,

[3] J. Y. Noh and U. Neumann. “Expression cloning,” In SIGGRAPH’01, pp. 277–288, 2001.

[4] E. Chen. “QuickTime VR—An image-based approach to virtual environment navigation,” Proc. of SIGGRAPH, 1995.

[5] A. Ronald and G. Bishop. “Improving static and dynamic registration in an optical see-through HMD,” Proceedings of SIGGRAPH’94, Orlando, Florida, In Computer Graphics Proceedings, Annual Conference Series pp. 197–204, July 24-29, 1994.

[6] D. Drastic and P. Milgram, “Perceptional issues in aug-

mented reality," In M. T. Bolas, S. S. Fisher, and J. O. Merritt, editors, SPIE Volume 2653: Stereoscopic Displays and Virtual Reality Systems *///*, pp. 123–134, January-February 1996.

[7] J. P. Rolland and H. Fuchs, "Optical versus video see-through head-mounted displays in medical visualization." Presence: Teleoperators and Virtual Environments, Vol. 9, No. 3, pp. 287–309, June 2000.

[8] A. Fuhrmann, G. Hesina, F. Faure, and M. Gervautz, "Occlusion in collaborative augmented environments," Computers and Graphics, Vol. 23, No. 6, pp. 809–819, 1999

[9] K. Reinhard, "Automatic reconstruction of buildings from stereoscopic image sequences," In R. J. Hubbold and R. Juan, editors, Eurographics'93, Eurographics, Blackwell Publishers, Oxford, UK, pp. 339–350, 1993..

[10] S. Growe, P. Schulze, and R. Tnjes, "3D visualization and evaluation of remote sensing data," Computer Graphics International'98 Hanover, Germany, June 22–26, 1998.

[11] E. Chen. "QuickTime VR—An image-based approach to virtual environment navigation," Proc. of SIGGRAPH, 1995.

[12] L. L. Ji and H. Yan, "Attractable snakes based on the greedy algorithm for contour extraction," Pattern Recognition 35, pp. 791–806, 2002.

[13] C. C. H. Lean, A. K. B. See, and S. A. Shanmugam, "An enhanced method for the snake algorithm," First International Conference on Innovative Computing, Information and Control (ICICIC'06), Vol. 1, , pp. 240–243, 2006

[14] C. Xu and J. L. Prince, "Gradient vector flow: A new external force for snakes," Proc. IEEE Conf. on Comp. Vis. Patt. Recog. (CVPR), Los Alamitos: Comp. Soc. Press, pp. 66–71, 1997.

[15] C. Y. Xu and J. L. Prince, "Snakes, Shapes, and Gradient vector fow," IEEE Transactions in Image Processing, Vol. 7, No. 3, Mar. 1998.

[16] C. Y. Xu and J. L. Prince, "Generalized gradient vector flow external frces for active contours," Signal Processing, Vol. 71, No. 2, pp. 131–139, Dec. 1998.

[17] S.-T. Wu, M. Abrantes, D. Tost, and H. C. Batagelo, "Picking and snapping for 3D input devices," In Proceedings of SIBGRAPI'03, pp. 140–147, 2003.

WICOM 2010

The 6th International Conference on Wireless Communications, Networking and Mobile Computing

September 23–25, 2010, Chengdu, China
http://www.wicom-meeting.org/2010

Call For Papers

WiCOM serves as a forum for wireless communications researchers, industry professionals, and academics interested in the latest development and design of wireless systems. In 2010, **WiCOM** will be held in **Chengdu**, China. You are invited to submit papers in all areas of wireless communications, networking, mobile computing and applications.

Wireless Communications

- B3G and 4G Technologies
- MIMO and OFDM
- UWB
- Cognitive Radio
- Coding, Detection and Modulation
- Signal Processing
- Channel Model and Characterization
- Antenna and Circuit

Network Technologies

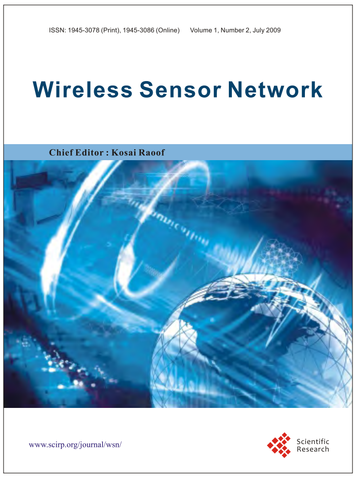
- Ad hoc and Mesh Networks
- Sensor Networks
- RFID, Bluetooth and 802.1x Technologies
- Network Protocol and Congestion Control
- QoS and Traffic Analysis
- Network Security
- Multimedia in Wireless Networks

Services and Application

- Applications and Value-Added Services
- Location based Services
- Authentication, Authorization and Billing
- Data Management
- Mobile Computing Systems

IMPORTANT DATES

Paper due:	March 10, 2010
Acceptance Notification:	May 10, 2010
Camera-ready due:	May 31, 2010


ISSN: 1945-3078 (Print), 1945-3086 (Online) Volume 1, Number 2, July 2009
Wireless Sensor Network
Chief Editor : Kosai Raoof
www.scirp.org/journal/wsn/
Scientific Research

Wireless Sensor Network (WSN)

Call For Papers

http://www.scirp.org/journal/wsn

ISSN 1945-3078 (Print) ISSN 1945-3086 (Online)

WSN is an international refereed journal dedicated to the latest advancement of wireless sensor network and applications. The goal of this journal is to keep a record of the state-of-the-art research and promote the research work in these areas.

Editor-in-Chief

Dr. Kosai Raoof , GIPSA LAB, University of Joseph Fourier, Grenoble, France

Subject Coverage

This journal invites original research and review papers that address the following issues in wireless sensor networks. Topics of interest are (but not limited to):

- Network Architecture and Protocols
- Self-Organization and Synchronization
- Quality of Service
- Data Processing, Storage and Management
- Network Planning, Provisioning and Deployment
- Integration with Other System
- Software Platforms and Development Tools
- Routing and Data Dissemination
- Energy Conservation and Management
- Security and Privacy
- Developments and Applications
- Network Simulation and Platforms

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 WSN.

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. Authors are responsible for having their papers checked for style and grammar prior to submission to WSN. Papers may be rejected if the language is not satisfactory. For more details about the submissions, please access the website.

Website and E-Mail

http://www.scirp.org/journal/wsn

Email: wsn@scirp.org

Call for Papers

International Journal of

Communications, Network and System Sciences (IJCNS)

ISSN 1913-3715 (Print)   ISSN 1913-3723 (Online)

http://www.scirp.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

Prof. Huaibei Zhou	Advanced Research Center for Sci. & Tech., Wuhan University, China
Prof. Tom Hou	Department of Electrical and Computer Engineering, Virginia Tech., USA

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
UWB technologies
Wave propagation and antenna design
Signal processing and channel modeling
Coding, detection and modulation
3G and 4G technologies
Sensor networks
Ad Hoc and mesh networks
Network protocol, QoS and congestion control
Efficient MAC and resource management protocols
Simulation and optimization tools
Network security

We are also interested in:

- Short reports—Discussion corner of the journal:
  2-5 page papers where an author can either present an idea with theoretical background but has not yet completed the research needed for a complete paper or preliminary data.
- Book reviews—Comments and critiques.

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.scirp.org/journal/ijcns   ijcns@scirp.org

TABLE OF CONTENTS

Volume 2 Number 7 October 2009

ISSN: 1913-3715

9771913371005 11