Foreword

szabo@hit.bme.hu

Our journal is continuing with the practice of publishing English issues regularly, at present twice a year, in July and in January. As before, most part of the present issue contains English versions of reviewed research papers, selected from the preceeding five Hungarian issues. One of them has been substantially revised. We included also one paper from open call. The editors would like to encourage prospective authors to submit their results specifically for the English issues. Being a selection, the papers' topics span a wide range of issues of current interest as the reader can see from the short summaries below.

A. Mitcsenkov et al address adaptive protection methods. The motivation is that bandwidth requirements of modern integrated networks solidly grow, and at the same time the reliability plays an increasingly important role. Various methods are known and under development to ensure survivability. The main feature of the proposed protection rearrangement framework is that the protection paths can be adaptively rerouted (rearranged) as the traffic and network conditions change, since they do not carry any traffic until a failure occurs.

The paper by *L. Nagy* deals with deterministic indoor wave propagation modeling. Next generation mobile access network system design needs more precise characterization of the radio channel and sophisticated propagation models because of the decreasing cell sizes and of higher data rates. The author proposes a Finite Difference Time Domain method to analyze the 2- and 3-dimensional indoor wave propagation problems. The efficiency and flexibility of FDTD for curved tunnel, indoor office and special EMC cases is demonstrated.

The title of the paper by *L. Csurgai-Horváth and J. Bitó* is “Multipath Propagation Fade Duration Modeling of Land Mobile Satellite Radio Channel”. The propagation on a Land Mobile Satellite (LMS) radio link is highly influenced by the shadowing effects of buildings and vegetation, or by the multipath propagation. In this contribution a digital model with Markov chain will be introduced, which is also applicable to determine the statistical parameters of the fade duration. The model is based on the measurement data of a real LMS channel which has been used to calculate the model parameters.

A. Kőrösi et al deal with DSL access networks. They provide an exact data-layer model and mathematical analysis of priority queuing systems representing DSL access networks on packet level with pre-emptive option. The accuracy and the efficiency of the numerical analysis is demonstrated by presenting numerical results based on simulations and numerical analysis both for complete and partial rejections. The analysis could be applied for an in-depth packet-level performance evaluation of recent DSL systems.

B. Kovács and P. Fülöp investigate mobility management strategies from the point of view of their need of signaling and processing resources on the backbone network and load on the air interface. A method is proposed to model the serving network and mobile node mobility in order to be able to model the different types of mobility management algorithms. Different mobility approaches are analyzed and their performance is numerically compared in various network and mobility scenarios. The aim is to give general design guidelines for the next generation mobility managements on given network and mobility properties.

The paper by *J. Levendovszky et al* is concerned with developing new energy balancing protocols for wireless sensor networks (WSN) to maximize the life-span of the system by using rare event tools. Novel packet forwarding mechanisms from the nodes to the base station (BS) are proposed, which minimize the energy consumption of WSN. The tail distribution of the energy consumption is estimated by the tools of large deviation theory and the concept of generalized statistical bandwidth has been introduced to evaluate the energy need of the network. The new results demonstrate that the lifespan of WSN can significantly be increased by the new protocols.

The paper by *E. Udvary* provides an overview of the basics and application possibilities of the multifunctional Semiconductor Optical Amplifier (SOA) in Sub-Carrier Multiplexed (SCM) systems. The paper focuses on the linearity investigation of the device. It describes the frequency dependence of the modulation and the harmonic products, the effects of the bias current and the optical power, the mismatch between the light and the electrical signal, the temperature and optical reflection sensitivity. It is shown that by using SOA as an external modulator, the device provides acceptable nonlinear distortion for SCM telecommunication systems.

L. Bokor et al present a novel vertical handover mechanism which aims at assuring streaming media services in a heterogeneous network environment where the subscribers are roaming among different wired/wireless access systems including ADSL, WiFi, 2.5G and 3G cellular and WiMAX. The handover scheme provides seamless connectivity during roaming, with adapting the quality of the delivered media stream to the changes of the network characteristics and to the capabilities of a wide variety of devices.

László Zombory	*Csaba A. Szabó*
President of the Editorial Board	*Editor-in-Chief*

Adaptive protection methods

ATTILA MITCSENKOV, DIÁNA MESKÓ, TIBOR CINKLER

Budapest University of Technology and Economics
High-Speed Networks Laboratory, Department of Telecommunications and Media Informatics
{mitcsenkov, mesko, cinkler}@tmit.bme.hu

Keywords: adaptive, shared, protection, resilience, rearrangement

The bandwidth requirements of modern integrated networks solidly grow, and at the same time the reliability plays an increasingly important role. Various methods are known and under development to ensure survivability. Our methods deal with this issue. The main feature of the proposed protection rearrangement framework is that since the protection paths do not carry any traffic until a failure occurs, they can be adaptively rerouted (rearranged) as the traffic and network conditions change.

1. Introduction

Computer networks since its beginning has gone through strong evolution: in the late 1970's the network was used for communication between a few scientific institutions – and today it is part of our daily life, plenty of new services have arisen (e.g. peer-to-peer systems, grid computing, Video on Demand, Voice over IP, e-banking services, etc.), the number of users increases rapidly, at a guess it doubles in every year. The convergence of computer, telecommunication and broadcast networks also pose new challenges [10,11].

Therefore modern transport networks raise new problems, not only in the field of bandwidth requirements, but also the quality and the resilience of the offered services plays an increasingly important role. Service disruption is no longer tolerated by business or industry; therefore survivable services have to be provided. The failure of any part of the network has to remain invisible for the customers.

The network management can ensure survivability using various methods – the network operator should decide which one to use. The alternatives will be discussed in Section 2. Sections 3, 4 and 5 present the investigated solutions, designed for different conditions and offering different performance. Section 6 presents and evaluates the numerical results, Section 7 concludes our work.

2. Resilience Strategies

According to the previously mentioned requirements for survivability, the network should be prepared for failures, to be able to make them invisible for customers by eliminating their effects (e.g. service interruption, data loss, increasing delay). Various resilience strategies are known that deal with these requirements. A brief overview of them is needed for the subsequent description of our methods. Here we focus only on single failures, which is a widespread assumption in the literature as well. However, after some modifications our methods can deal with multiple failures [6].

Protection vs. Restoration: While using protection methods, protection (backup) paths are defined in advance, and in case of failure the traffic is immediately switched to the corresponding backup path. In case of restoration the protection paths are sought only when failures occur. It results in a thriftier operation, but it might fail when establishing backup paths due to insufficient resources.

In case of *Dedicated Protection* each working path has a separate backup path, with exclusively reserved resources. Conversely, *Shared Protection* means commonly used resources among backup paths of different working paths. It results in a thriftier but slower method with higher complexity.

Regarding the part of the network to be protected, the following classification can be made: (1) *path protection (or end-to-end protection),* means the working path is protected by one path that is totally disjoint from the working one [1,2]; (2) *link protection,* when all the traffic from the failed link is re-routed between the ends of that link; (3) *sub-network protection,* where the network is clustered into protection domains (sub-networks) that define the ends of protection segments; and (4) *segment protection* [3], when the working path is divided into segments, and protected by a few backup paths covering them. These backup paths (segments) should of course jointly cover the whole working path, and should be at least partially disjoint from the working path.

In case of *static* protection predefined working and protection paths are used for each node-pair. When these paths are redefined from time to time, we refer to this method as *dynamic* protection/restoration. *Adaptive* methods are able to alter the previously routed protection paths, and define new paths for every new demand, adapting to the varying network and traffic conditions. This is the slowest and most complex approach, but it offers the strongest control over the resources of the network.

According to the above definitions, our protection methods

– are shared protection methods
– are dynamic or adaptive
– offer path or segment protection
– guarrantee survival of any single failure,
but work for some multiple failure patterns as well.

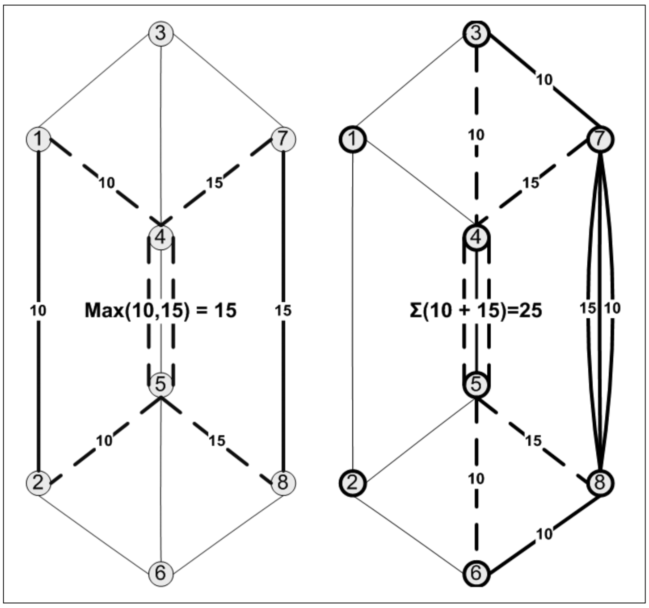
2.1. Spare Capacity Allocation

The main benefit of shared protection methods is the thrifty resource utilization. Resource sharing is allowed among different protection paths. However, we should avoid the case when different demands would switch over to the same protection path simultaneously. If only single failures occur in the network, then two disjoint (independent) working paths cannot fail simultaneously, therefore protection paths belonging to disjoint working paths can share resources. It results in a thrifty resource usage, without losing the survivability in case of any single failure [4].

Let us show a detailed example in *Figure 1*: two demands are given with capacities of 10 and 15 units. Working paths are denoted by solid lines and protection paths by dashed lines. In the first case (figure on the left side) these demands have to be routed between nodes 1-2 and 7-8 – it means the working paths have no common resources, i.e. the protection paths can not be simultaneously activated if only single failures occur. Therefore the protection paths can share resources among link 4-5, thus max (10;15) = 15 units of capacity is allocated. In the second case (figure on the right) these demands have to be routed between nodes 3-8 and 6-7, and both working paths use link 7-8. In this case, if link 7-8 fails, both demands will use their protection paths, therefore 10+15=25 units of capacity have to be allocated for protection paths on that link 7-8.

The capacity C_l of each link in the network is divided into three parts *(Figure 2)*: the allocated capacity for working paths (*C'*), the capacity used by protection paths, that can be used by another ones according to the above described criteria (*C"*), and the free, unused capacity: *C-C'-C"*. Different costs are assigned to these parts, while the re-use of capacity reserved for protection paths means no extra allocation, unlike the use of the free capacity-range. It results in a cost function with two linear segments.

Figure 1. Spare Capacity Allocation



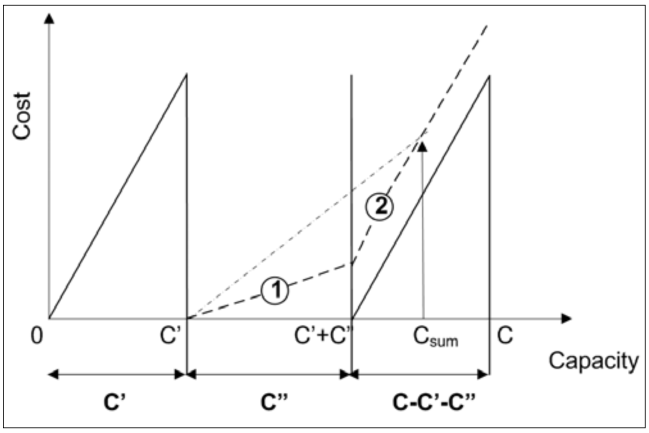


Figure 2. Different Capacity Domains

3. Dynamic Algorithms

Three different versions of a shared protection method were implemented as references and as the basic elements of more complex algorithms, with different restrictions on the protection paths, ensuring more and more flexibility in small steps.

The starting point was the Failure-Independent Shared Path Protection (F-I SPP), where after routing the working path, a single end-to-end disjoint protection path is needed, with the lowest cost in the sense of the above described capacity cost function *(Figure 2)*. In case of failure it re-routes the traffic to this single protection path regardless of the location and type of the failure.

The next step was the Failure-Dependent Shared Path Protection (F-D SPP), where after routing the working path, different protection paths are assigned for the failure of each link on the working path. The protection paths of the same working path are allowed to share links with each other and other working paths considering the above described criteria. However, all protection paths have to be end-to-end disjoint from their working ones.

Finally the failure-dependent Partially Disjoint Shared Path Protection (PDSP) was implemented. It assigns different protection paths for a single working path like the F-D SPP, but it has fewer restrictions for the protection paths. This method can be classified as link protection and also as segment protection, while it assigns different protection paths for the failures of any link on the working path. These protection paths have to substitute a certain segment of the working path; anyway they can use the rest of the links of the working path.

The requirement of the previous methods (F-I/F-D SPP) was that protection paths are disjoint from all links of the corresponding working path. PDSP requires disjointness only from the link of which failure it has to protect against. In this manner the protection paths can have common links with their working ones, except for only one link – and this is the difference between SPP and PDSP. Due to the fewer restrictions, ranging from F-I SPP through F-D SPP to PDSP the results are expected to improve in this particular order, especially for the protection path establishment in the network.

The detailed operation of the failure-independent SPP can be read in the next subsection, followed by algorithms F-D SPP and PDSP.

3.1. Failure Independent Shared Path Protection (F-I SPP)

The basis of the more sophisticated methods, the F-I SPP will be described now. It works as follows:

- **Step 1:** For any new demand (o_{new}):
  → Use Dijkstra's algorithm to find the shortest working path. If not successful, the demand will be blocked, go to *Step 1*.
  → Hide all the edges of the working path (*)
- **Step 2:** For all edges *l'* of the working path and for all edges *l"* of the network (except hidden edges) compute capacity $C_{l',l''}$, which represents the amount of available shared capacity for protection paths on link *l"* when link *l'* fails
- **Step 3:** Find the maximum of $C_{l',l''}$ for all *l'* found so far – this will be the value $C_{l''}$ (**). This is a failure-independent method; the same protection path will be used in case of any failure, therefore when determining the available shared capacity the worst case has to be considered to have a suitable solution for every failure.
- **Step 4:** Calculate the cost increment required for routing the protection path of demand o_{new} with bandwidth requirement b_o based on $C_{l''}$ along all the links *l"* of the network. It is the sum (C_{sum}) of the available shared capacity to be used (*C"*) with a lower linear cost (marked with 1 in *Figure 2,* and the amount of unused capacity (*C-C'-C"*) that has to be allocated with a higher linear cost (marked with 2 in *Figure 2*).
- **Step 5:** Use Dijkstra's algorithm to find the optimal protection paths, based on the cost increments described in *Step 4*.
- **Step 6:** If succeeded, allocate the new paths; otherwise de-allocate resources for terminated connections and update capacity allocations; demand o_{new} will be blocked. If there are more demands go to *Step 1*.

In essence, Failure-Independent Shared Path Protection (F-I SPP) is a really fast and simple shared protection method. Unlike the adaptive methods, it cannot reroute the previously allocated protection paths. SPP is a failure-independent method, thus in case of failure the traffic is switched to the same backup path, irrespective of the location and type of the failure.

The following two methods are modifications of F-I SPP.

3.2. Failure Dependent Shared Path Protection (F-D SPP)

The failure-dependent version differs in the number of protection paths to be allocated for a demand – the F-D SPP assigns separate protection paths for the case of failure of any links on the working path. Therefore, a modification of *Step 3* (marked with **) is needed: no maximization of the $C_{l',l''}$ values is necessary, and *Steps 4 and 5* must be applied for every links of the working path. If all executions of Dijkstra's algorithm in *Step 5* succeeded, the working and protection paths have to be allocated, otherwise the demand has to be blocked.

3.3. Partially Disjoint Shared Path Protection (PDSP)

This one is a shared protection method operating on segments of the network, i.e. segments of the working paths. It means a difference in the set of edges not available for the protection paths. It is a failure-independent method, working with separate protection paths for the case of the failure of any link on the working path. Therefore, it works like F-D SPP described above, but a modification is needed in *Step 1* (marked with *): not all the edges of the working path, but the only one link *e'* has to be hidden while establishing the protection path belonging to *e'*. It allows the algorithm to use any other links of the network for the protection path.

Table 1. Classification of algorithms

	Path Protection	Segment Protection
Dynamic	Shared Path Protection **(F-I/F-D SPP)**	Partially Disjoint Shared Path Protection **(PDSP)**
Adaptive	Shared Path Protection with Link Doubling **(SPP-LD)**	Partially Disjoint Shared Path Protection with Link Doubling **(PDSP-LD)**

4. Adaptive Methods

We presented the dynamic methods: F-I/F-D SPP and PDSP. In the next subsections, we explain the adaptive versions of these path and segment-protection methods. First the necessary modeling trick (Link Doubling – LD) to linearize the problem will be presented, followed by the Mixed Integer Linear Program (MILP) formulation of protection rearrangement and by our two proposed methods, namely Shared Path Protection with Link Doubling (SPP-LD) and Partially Disjoint Shared Path protection with Link Doubling (PDSP-LD).

4.1. Link Doubling (LD)

When rearranging the protection paths, a serious problem arises: the calculation of the available shared capacity ($C_{l',l''}$) hardly depends on the previously allocated protection paths. However, in case of adaptive methods multiple protection paths have to be determined or altered simultaneously, and it also affects the previously allocated protection paths. Therefore, the calculation of the available shared capacity cannot be based on the view of previously routed demands.

The *Minimal Cost Multi-commodity Flow* (MCMCF – [9,12]) problem deals with this problem. Numerous solutions can be found in the literature, e.g. some heuristics, iterative solutions or Integer Linear Programming (ILP) [13]. Our methods are based on ILP [14].

Our cost functions are not linear, but have two linear segments *(Figure 2)*. ILP needs linear cost functions, therefore auxiliary variables are needed. These extra variables can be illustrated by the Link Doubling modeling trick.

The composite cost function of the edge between nodes A and B actually belongs to two different parts of capacity, and two appropriate linear cost functions. These two can be separated: any edge of the network should be substituted by two parallel edges, one for each capacity range. The lower cost has to be assigned to the shared resources, the higher cost to the unused resources as shown in *Figure 2*. This way the number of edges doubles, therefore the complexity of the problem gets higher, but ILP can be used for the MCMCF problem.

4.2. MILP Formulation

Protection rearrangement means that a part of the previously allocated protection resources has to be deallocated, and then allocated again simultaneously with the new demand(s). A set of demands to be rearranged is selected for the case of any link failure in the network. This way the system of protection paths can be adapted to the altering network load.

To handle the above described special MCMCF problem a proper MILP (Mixed Integer Linear Program) formulation is needed. Its role is to find an optimal solution for simultaneous routing of different protection paths in the network based on the two-segment linear cost functions, considering the disjointness criteria for working and protection paths, and the flow conservation and link capacity constraints as well.

Objective:

$$\min \sum_{o \in T_e} \left\{ \sum_{l \in E^{free}} x_l^o w_l + \sum_{l \in E^{sh}} x_l^o \gamma_l \right\} \quad (1)$$

E^{sh} stands for the edges created by LD, representing the shared capacity of the links, and E^{free} stands for edges representing the unused capacity. As described above, a set of demands (T_e) is selected to be rearranged for the failure of any link (e) in the network. The amount of capacity needed by the protection path of demand o on link l is represented by variable x_l^o.

The two-segment capacity cost function is given by the cost-coefficients for shared (γ_l) and unused (w_l) capacity. By setting the γ_l/w_l proportion the priorities for using shared capacity can be affected. If γ_l is close to zero, protection sharing is forced. It leads to thriftier operation, but it may result in longer protection paths, avoiding the edges in E^{free}. If γ_l is close to w_l, the use of sharable capacity is not preferred over unused capacity. Extensive simulations have shown that the best results can be achieved by setting $\gamma_l/w_l \approx 0.1$, $\forall\, l \in E$.

Subject to:

$$\sum_{o \in T_e} x_l^o \le C_l - C'_l - C''_l \text{ for all links } l \in E^{free} \quad (2)$$

$$\sum_{o \in T_e} x_l^o \le C''_l \text{ for all links } l \in E^{sh} \quad (3)$$

$$\sum_{\forall j \in V, j=i} x_{ij}^o - \sum_{\forall k \in V, k=i} x_{ki}^o = = \begin{cases} 0 \text{ if } i \ne s_o \wedge i \ne d_o & \text{for all nodes } i \in V \\ b_o \text{ if } i = s_o & \text{and demands } o \in T_e \\ -b_o \text{ if } i = d_o & \end{cases} \quad (4)$$

$$0 \le x_l^o \le b_o \text{ for all links } e \in E \text{ and demands } o \in T_e \quad (5)$$

$$\sum_{\forall k \in V, k=i} x_{ki}^o = b_o \cdot z_i^o \text{ for all nodes } i \in V \text{ and demands } o \in T_e \quad (6)$$

$$z_i^o \in \{0,1\} \text{ for all nodes } i \in V \text{ and demands } o \in T_e \quad (7)$$

Equations (2) and (3) represent the capacity constraints for the free and sharable capacities, Eq. (4) is the well known flow conservation constraint. Equation (5) gives a constraint for the *x* variables, which may not exceed the bandwidth of the corresponding demand. Equation (6) allows flow splitting only between the parallel edges created by LD, using an auxiliary binary variable, z_i^o (Eq. 7).

4.3. Shared Path Protection with Link Doubling (SPP-LD)

This method is the adaptive extension of the failure-dependent shared path protection (F-D SPP).

The basic idea of these adaptive methods is the ability to rearrange the protection paths, from time to time. It allows us to exempt overloaded network elements by moving a segment of the load to other parts of the network. It allows higher control over the network load conditions. Considering that protection paths do not carry real traffic, they are just for backup purposes, altering them does not cause service disruptions, therefore it remains invisible for the customers.

The more protection paths are routed simultaneously, the better results are expected, however the complexity of the problem grows exponentially. Therefore to rearrange all of them at a time is not possible, but selecting a proper subset of protection paths makes it suitable.

New protection paths are assigned to the failure of any link in the network, whereas these are failure-dependent methods. When determining the protection paths for link *e*, every protection path using *e* is rearranged. This will be the T_e subset of rearranged paths.

The SPP-LD algorithm works as follows:

- **Step 1:** For any new demand (o_{new}):

→ Use Dijkstra's algorithm to find the shortest working path

- **Step 2:** For the edges e of the working path:

→ Hide temporarily link e.

→ Set T_e to contain the new demand o_{new} and all the demands using e as a part of their working paths.

- **Step 3:** De-allocate protection paths of T_e.

- **Step 4:** Execute the MILP with an added path diversity constraint for all demands in T_e to find the protection paths for the failure of link e

→ The path diversity constraint prevents protection paths to use resources of the working path of corresponding demand, in case of SPP-LD it is as follows:

$x_e^o = 0$ for all demands $o \in T_e$ and links $e \in WP_o$

→ If the ILP is feasible, it means in case of the failure of link e all the protection paths of the previously routed demands in T_e and the new demand can be allocated. If more links, go to *Step 2*, otherwise allocate the working and protection paths, and go to *Step 1*.

→ If the ILP is infeasible, the demand cannot be allocated, the previous state of the network will be restored and the demand o_{new} will be blocked.

4.4. Partially Disjoint Shared Path Protection with Link Doubling (PDSP-LD)

This method is the adaptive extension of the partially disjoint shared path protection (PDSP), described in 3.3.

It is also based on the rearrangement of protection paths, however differs from SPP-LD in its disjointness criteria. The working and protection paths do not have to be end-to-end disjoint, as it was in the case of path protection. It assigns different protection paths to any link of the working path for its failure, and as it was stated, these methods deal only with single failures. Therefore, it is enough to make the protection paths disjoint from the corresponding link, and allow them to use any other links, while two different links of the working path may not fail simultaneously.

Although the operation of the PDSP-LD is very similar to SPP-LD, the path diversity constraint of *Step 4* described in the previous paragraph differs:

$$x_e^o = 0 \text{ for all demands } o \in T_e$$

This prevents the protection paths to use the examined link of the working path, but allows them to use any other links of the network, as it was described above.

5. Semi-Adaptive Methods

Adaptive methods afford higher flexibility and more control over network load conditions, as the results will show. From the aspect of performance these methods perform well, as expected.

However there are some problems from the aspect of applicability. The complexity of integer linear programming results in seriously increasing computational times, depending on the size, connectivity and other attributes of the network. As the simulations show the required time for establishing working and protection paths in average for any single demand may be about 1.0-10.0 seconds of order of magnitude. And of course the larger the network is the longer times are needed *(Table 2)*.

It makes adaptive methods useless for certain services and conditions, therefore a trade-off between the performance and flexibility of adaptive algorithms and the low complexity of the dynamic methods is needed.

This trade-off can be realized in multiple ways, we introduce here a really obvious solution. The dynamic and adaptive algorithms use the same routine for working paths, but differ in the method for determining protection paths. Therefore, the working path of any new demand can be established in the common way, and then, first the dynamic routine will be used for the protection paths. In case of success the resources will be allocated.

The adaptive resource rearrangement mechanism will be applied only in case when the dynamic routine fails. Clearly it results in a faster algorithm, while in many cases the resource rearrangement mechanism is unnecessary, however it still has the ability of resource rearrangement.

The detailed operation of the Semi-Adaptive Shared Path Protection (S-A SPP) and Semi-Adaptive Partially Disjoint Shared Path Protection (S-A PDSP) is described in the following paragraphs.

5.1. Semi-Adaptive Shared Path Protection (S-A SPP)

The concurrent use of path- and segment-protection results in a complicated administration through the simultaneously established protection paths because different ILP-constraints are needed for path and segment protection. Furthermore in some cases path protection can be desired because its simpler management requirements, and a hybrid solution is not acceptable.

Therefore the investigation of a semi-adaptive path protection method is reasonable. It works as follows:

- **Step 1:** For any new demand (o_{new}):

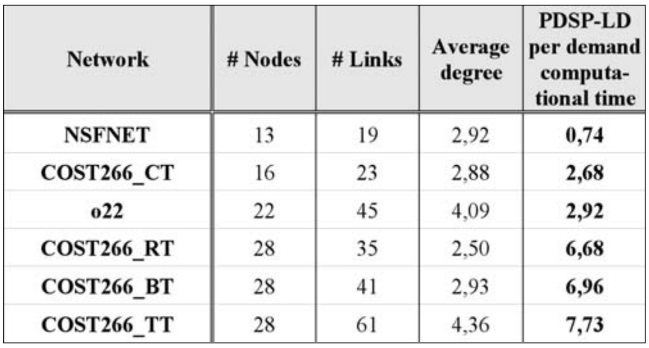
→ Use Dijkstra's algorithm to find the shortest working path

- **Step 2:** Try to find protection paths for any link failure as described for the dynamic, failure-dependent shared path protection (F-D SPP) algorithm in its *Steps 2-5*.

- **Step 3:** In case of success allocate resources for working and protection paths, go to *Step 1*, otherwise proceed to *Step 4*.

- **Step 4:** Try to find protection paths for any link failure as described for the adaptive shared path protection (SPP-LD) algorithm in its *Steps 2-4*.

- **Step 5:** If succeeded, allocate resources for the new demand, otherwise o_{new} has to be blocked. Go to *Step 1*.



Network	# Nodes	# Links	Average degree	PDSP-LD per demand computa-tional time
NSFNET	13	19	2,92	**0,74**
COST266_CT	16	23	2,88	**2,68**
o22	22	45	4,09	**2,92**
COST266_RT	28	35	2,50	**6,68**
COST266_BT	28	41	2,93	**6,96**
COST266_TT	28	61	4,36	**7,73**

Table 2. Network characteristics

5.2. Semi-Adaptive Partially Disjoint Shared Path Protection (S-A PDSP)

It is a kind of segment protection, as described for PDSP in 3.3. Of course it uses the protection path determining routine of PDSP and PDSP-LD instead of SPP and SPP-LD, otherwise it works as the above described semi-adaptive shared path protection.

6. Numerical Results

We have compared the performance of these algorithms on six well-known topologies consisting of 13-28 nodes and 19-61 edges. The three COST266 reference networks with the same 28 nodes and different sets of edges are of special interest from the aspect of the effect of variable graph connectivity on the performance. *Table 2* shows the characteristics of the networks. The traffic patterns consisted of sufficient amount of different traffic demands to eliminate the effect of the initial transient loading the empty network.

Blocking rates are an important aspect of the performance analysis; therefore at first the results for this topic will be presented. To investigate different blocking ranges we have scaled the link capacities, not the traffic. Note that increasing uniformly the capacity of each link is analogous to decreasing bandwidth of traffic offered to the network. For every network-traffic pair a ten-step simulation sequence was carried out, with capacity values resulting in blocking rates from roughly 90% to around 0%.

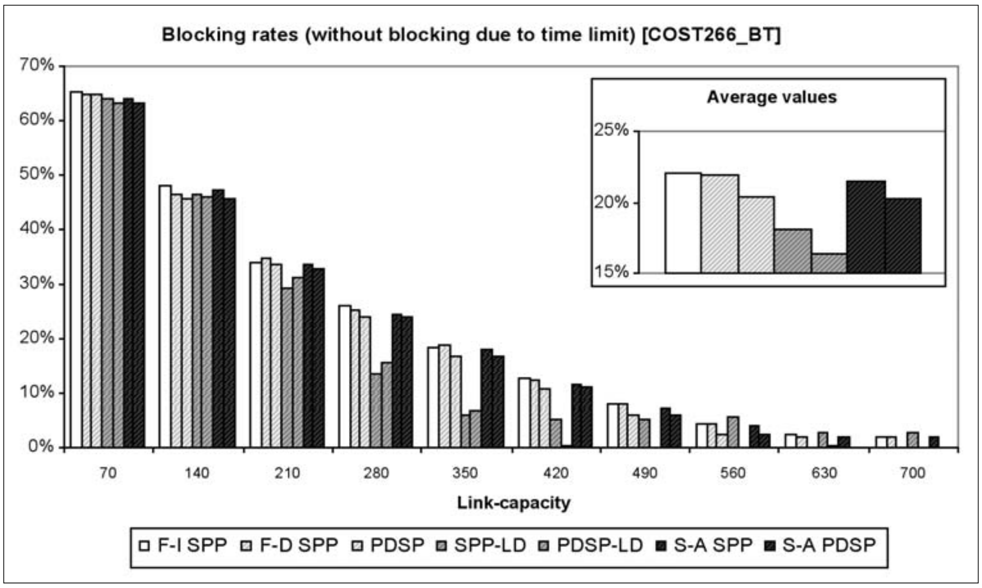
As we have previously mentioned, we face the problem of high complexity and computational time using ILP, and it can distort the results. When processing the incoming demands, the routing method cannot spend a long time to solve any single ILP-problem of a new demand, because it makes the following demands wait for it. Therefore a considerably strict time limit has to be taken into account, and it has a distortion effect: some ILP-problems that are feasible will not be calculated in time, and it means there will be some demands with enough capacity in the network to satisfy them, however due to this time limit, no protection paths will be found for them, and they will be blocked.

That is why we cannot count them as blocked demands, nor as routed ones, because the bandwidth-requirement of them is available, still not allocated in the network. Therefore these demands are simply left out from the statistics. It allows us to examine the potential of the adaptive methods, however in practical applications the limitations of computational time are not negligible, and these demands will count as blocked ones. It is the reason why semi-adaptive algorithms are needed.

The differences between the performances of the described algorithms become more apparent when using larger networks. Accordingly, if we focus on the results of the simulation based on one of our largest networks (COST266_BT – *Figure 3*), the following tendencies are noticeable:

- Failure-dependent SPP provides lower blocking rates than the failure-independent SPP
- The use of so-called “partially disjoint shared protection” results in significantly lower blocking rates than fully-disjoint shared path protection, due to more flexibility in the protection path-building phase (SPP vs. PDSP, SPP-LD vs. PDSP-LD and S-A SPP vs. S-A PDSP)
- Adaptive methods clearly perform better than the dynamic methods because of the protection rearrangement (SPP vs. SPP-LD, PDS vs. PDSP-LD)
- The semi-adaptive methods have lower blocking rates than dynamic versions, but higher than the adaptive algorithms, as expected (SPP < S-A SPP < SPP-LD, PDSP < S-A PDSP < PDSP-LD)

Figure 3. Blocking rates I.



The topology characteristics of network COST266_RT (low graph-connectivity, low average focal degree) make it suitable to demonstrate the weaknesses of path protection. If we take a look at *Figure 4*, and focus on the low-blocking capacity-range, a significant difference is noticeable between the segment and path-protection methods.

Path protection algorithms need two totally disjoint paths for every demand (for the whole source-destination route), and in sparse networks sometimes it is not suitable. At the same time, for segment protection a set of shorter path-duplications to protect different parts of the working path is sufficient. This is the reason why path protection methods cannot reach the 0% blocking rates in a sparse network like COST266_RT.

The adaptive algorithms have further benefits beyond the lower blocking rates. Due to the frequent rearrangement of resources used for protection paths it offers a more even resource usage. The simultaneous protection path establishment for a group of demands results in shorter protection paths, avoiding unnecessary long paths because of bottlenecks. *Figure 5* shows the length of the protection paths (in average): the adaptive algorithms offer clearly the shortest paths, then the semi-adaptive methods, and finally the dynamic versions. Furthermore, the adaptive methods need fewer resources for the protection paths *(Figure 6)*, which means the shorter paths are not the result of less resource sharing but the optimal reconfiguration.

Obviously by improving performance in finding the optimal solution for protection paths the required computational time grows. Its importance is that in some cases it makes the adaptive methods not applicable: for time-critical applications the permanent use of ILP is not acceptable. Furthermore, these algorithms are centralized routing methods, and therefore these cannot be used for large networks due to scalability problems – if not simplified. Table 2 shows the relationship between the network properties and the time needed in average for any single demand using adaptive algorithms: it strongly depends on the number of nodes (ten times longer for COST266_BT then NSFNet with two times more edges and almost the same average nodal degree), and slightly depends on the number of edges (the three COST266 reference networks with different amount of edges need almost the same time).

We do not have to lose all the benefits of adaptive methods neither for time-critical applications or large networks. The semi-adaptive methods could be the optimal choice for these situations: these have the ability to rearrange the protection paths, but for faster operation resource-reconfiguration is made only if a demand cannot be routed without it. Furthermore, these methods are adaptable to different requirements and compromises between speed and performance by altering the prevalence of rearrangement.

The following figures show the required average path computational times (in sec.) required for any single demand in the smallest and largest network tested (NSFNet – *Figure 7*, COST266_TT – *Figure 8*): the semi-adaptive versions remain around 0.01 seconds ever for large networks.

7. Conclusions

In this paper, a set of network protection algorithms have been proposed. A group of dynamic methods offered fast and simple solution, adaptive methods were presented to improve performance, and semi-adaptive versions as a trade-off between speed and performance. Adaptive methods are based on the idea of rearranging protection paths, since protection (backup) paths normally do not carry any traffic.

As the results show, due to this reconfiguration of protection paths the adaptive methods achieve better performance in the sense of throughput (lower blocking rates), network utilization and even traffic distribution. The drawback of these methods was the complexity of them: for time-critical applications and large networks these are poorly applicable. Therefore semi-adaptive versions were introduced to reduce the average per-demand computational time, and as the results show, these solutions still have some of the benefits of adaptive algorithms.

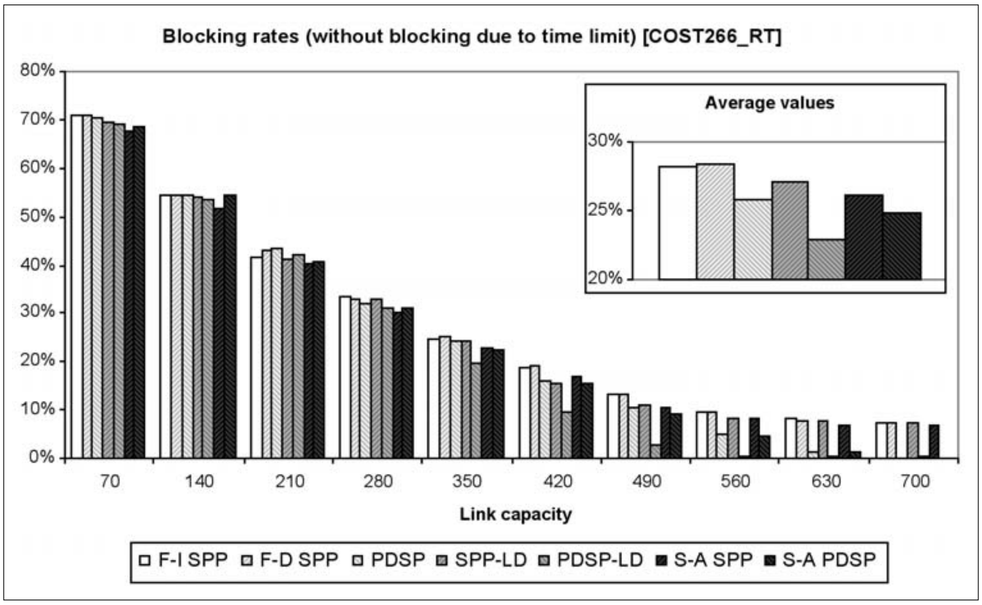


Figure 4.
Blocking rates II.

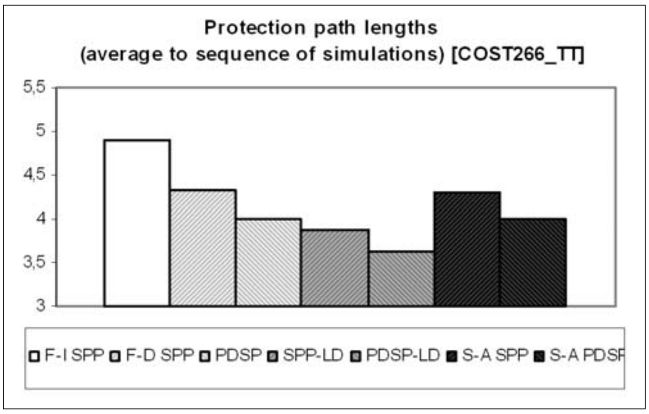


Figure 5. Average length of protection paths

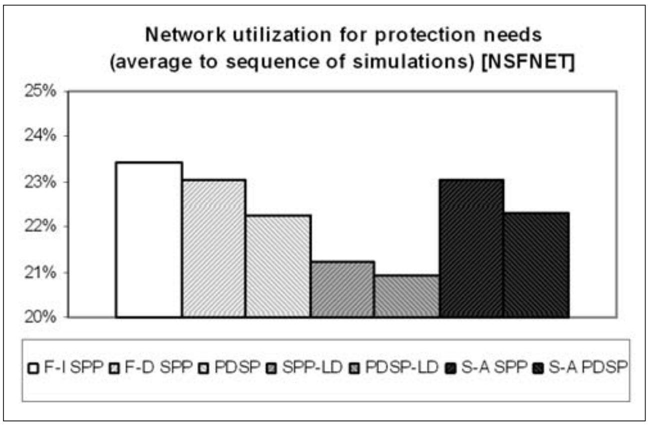


Figure 6. Network utilization for protection needs

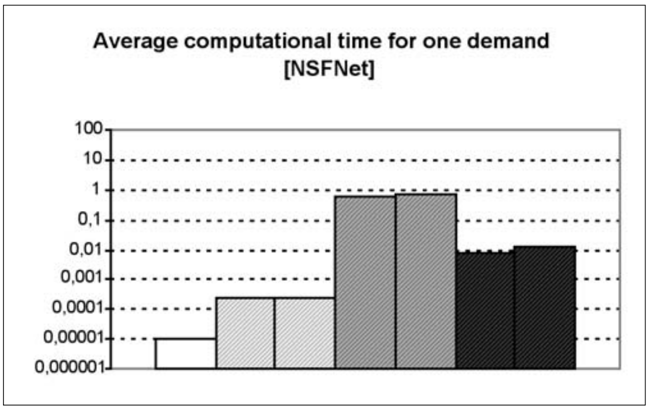


Figure 7. Computational times I.

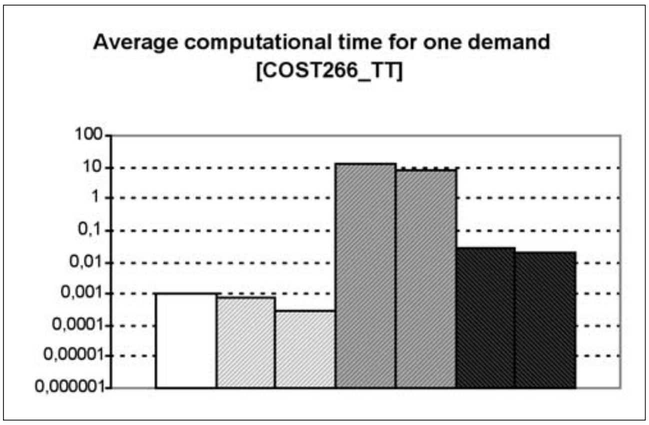


Figure 8. Computational times II.

Acknowledgements

This work has been done as a part of the
European 6th Framework Research Project IP NOBEL
(http://www.ist-nobel.org)

References

[1] J.W. Suurballe:
"Disjoint Paths in a network",
Networks, Vol. 4., pp.125–145., 1974.

[2] R. Bhandari:
"Survivable networks: algorithms for diverse routing"
Kluwer Academic Publishers,
ISBN 0-7923-8381-8, 1999.

[3] P.H. Ho, H.T. Mouftah:
"A framework for service-guaranteed
shared protection in WDM mesh networks",
IEEE Communication Magazine, Vol. 40., No.2,
pp.97–103., Februar 2002.

[4] T. Cinkler, P. Laborczi, Á. Horváth:
"Protection through thrifty configuration",
16th International Teletraffic Congress,
Edinburgh, UK, June 1999.

[5] T. Cinkler, D. Meskó, A. Mitcsenkov, G. Viola:
"Adaptive Shared Protection Rearrangement",
Design of Reliable Communication Networks,
Naples, October 2005.

[6] Zs. Pándi, M. Tacca, A. Fumagalli:
"A threshold based on-line RWA algorithm
with reliabilty guarantees",
Conference on Optical Network Design&Modelling,
Milan, Italy, 2005.

[7] J.-P.Vasseur, M. Pickavet, Piet Demeester:
Network Recovery, Elsevier,
pp.39–131, 203–297, 297–423., 2004.

[8] Wayne D. Grover:
Mesh-Based Survivable Networks, Prentice Hall,
pp.149–172, 173–268, 377–467., 2004.

[9] M.Pióro, D. Medhi:
Routing, Flow and Capacity Design
in Communication and Computer Networks,
Elsevier, 2004.

[10] Internet Software Consortium (ISC),
http://www.isc.org/

[11] London Internet Exchange (LINX),
http://www.linx.net/

[12] Ravindra K. Ahuja, T. L. Magnanti, J. B. Orlin:
Network Flows, Prentice Hall,
pp.108–113, 649–695., 1993.

[13] Mokhtar S. Bazaraa, John J. Jarvis, Hanif D. Sherali:
Linear Programming and Network Flows,
John Wiley and Sons, pp.587–601., 1977.

[14] Schrijver:
Theory of Linear and Integer Programming,
Wiley, 1998.

Deterministic indoor wave propagation modeling

LAJOS NAGY

Budapest University of Technology and Economics,
Department of Broadband Infocommunications and Electromagnetic Theory
lajos.nagy@mht.bme.hu

Keywords: indoor propagation, diffraction, FDTD, ray tracing

The next generation mobile access network system design needs more precise characterization of the radio channel and needs sophisticated propagation models, because of the decreasing cell sizes and of higher data rates. In particular, the planning of the coverage in tunnels and indoor spaces causes design problems without these models. Indoor propagation problems for wideband radio systems are widely investigated and one of the today applied approach of modeling is the ray-tracing, ray-launching method. These ray methods are efficient for parallel-perpendicular scenarios but there is a common problem when tracing the rays for curved surfaces. The other disadvantage of the ray methods is the difficulty in describing the diffraction for a complex scenario. The specific case of the straight circular tunnel can be modeled analytically as a waveguide with circular cross-section. Each of the two previous models has a disadvantage by modeling our problem: the ray tracing needs huge running time because on the curved surface reflection the number of rays in bundles has to be increased, whilst the analytical method is not able to handle the complex propagation problem in presence the vehicle. In our investigation the Finite Difference Time Domain method was proposed and used to analyze the 2 and 3 dimensional indoor wave propagation problems. We demonstrate the efficiency and flexibility of FDTD for curved tunnel, indoor office and special EMC cases [11-13].

1. Introduction

In the radio network design practice empirical, semi-empirical and deterministic radio wave propagation methods are used for field strength estimation. The designers of the 3rd and 4th generation systems need broadband characterization of the radio channel and coverage prediction, which is based on any deterministic propagation model. The receiver designers also need precise stochastic description of the radio channel to develop the equalizer and estimation of the receiver.

The indoor base station and radio access points are usually used to extend coverage to indoor areas where outdoor signals do not reach well, or to add network capacity in areas with very dense mobile device usage. These cells in the mobile structure are called as picocells. In the dense subscriber environments the precise coverage prediction has an increased importance, and the empirical and semi-empirical models are not able to guarantee the necessary accuracy.

The deterministic models generally based on ray tracing or on direct solution of the Maxwell's equations. The first step of the ray tracing methods is solving a pure geometrical problem, but in special cases this leads to very complex analysis especially for curved building geometries, highway and underground tunnels, and for highly reflective building media, like reinforced concrete. In such cases the tracing and storing of few million rays results in huge memory requirement and calculation time for multiple reflection, transmission and diffraction. The other disadvantage of the ray methods is the difficulty in describing the diffraction for a complex scenario. In related papers [1] ray tracing method introduced where bundles of rays are used to represent each “physical” wave. Monte Carlo techniques were used for the ray launching. Each bundle of rays was traced to a receiver position where reception spheres determined which rays are intercepted by the receiver.

The Maxwell's equations can be solved directly using parabolic type equations or in differential form using Finite Difference Time Domain (FDTD) method [5].

The main aim of this article is to introduce the FDTD method and to demonstrate its suitability in solving radio wave propagation problems. The application areas are individually introduced and demonstrated by examples, of which the memory and simulation time requirements are also analyzed. First the radio wave propagation physical mechanisms are summarized, which are also used by the ray tracing method for modeling wave material interaction. The next section summarizes the most important empirical and deterministic indoor propagation models, and the FDTD will be introduced briefly for general three-dimensional case in rectangular coordinate system and for two special two dimensional geometries. Section 4 discusses the data base requirements of building geometries for indoor radio wave propagation modeling, and the last part of the paper introduces the simulation results.

2. Wave propagation mechanisms

The effect of difficult and complex geometry of radio wave propagation environment can be simplified to simple physical models as direct, reflected, transmitted and diffracted paths. The ray tracing propagation modeling method means solving the geometrical problem first, and after partitioning the wave into rays the simple physical methods above are used to describe the interaction between the propagating waves and materials.

2.1. Direct path

The direct path means considering the propagation in Line of Sight (LOS), where the receiver is in direct visibility with the transmitter. In terms of radio wave propagation, being in LOS means having the Fresnel zones cleared all along the path.

If the transmitter antenna of gain G_A is fed by P_A input power, then the radiated power density at a distance r assuming LOS spherical wave propagation would be

$$S_o = \frac{P_A G_A}{4\pi r^2} \tag{1}$$

In the far field region of the transmitter antenna the electric and magnetic field strength vectors are mutually perpendicular to each other and of propagation direction, and are in phase. Therefore the power density can be described as

$$S_o = \frac{|\mathbf{E}|^2}{240\pi} \tag{2}$$

The electrical field strength magnitude can be derived from (1) and (2) as

$$E = \frac{\sqrt{60 P_A G_A}}{r} \tag{3}$$

Expression (3) shows that the electrical field strength magnitude is inversely proportional to the distance, for spherical waves, and the received power is inversely proportional to the square of the distance. For the two-dimensional problem, the electrical field strength dependence is $E \approx 1/\sqrt{r}$ in case of cylindrical wave.

2.2. Reflection

Reflection occurs when a propagating electromagnetic wave impinges upon an object which has very large dimensions when compared to the wavelength of the propagating wave. Reflections occur from the surface of the earth and from buildings and walls. The amplitude, phase and polarization of the reflected wave depend on material parameters of the reflecting medium and on the surface irregularity. If the interaction surface is plane and perfectly smooth then the specular reflection is observed and the energy flow is discrete in space. This ideal case can be modeled using the Snell-Descartes law extended for lossy dielectrics. In most cases of the radio wave propagation problems the medium is diamagnetic or nonmagnetic so its relative permeability is 1.

The reflection coefficient for plane waves is defined as the complex electric field strength ratio of the incoming and the reflected wave ($R = E_r / E_i$), which is decomposed into its perpendicular and parallel components *(Figure 1)*. The reflection coefficients for the two polarization states are:

$$R_\perp = \frac{\cos\vartheta - \sqrt{\varepsilon_r + \cos^2\vartheta - 1}}{\cos\vartheta + \sqrt{\varepsilon_r + \cos^2\vartheta - 1}}$$
$$R_\parallel = \frac{\varepsilon_r\cos\vartheta - \sqrt{\varepsilon_r + \cos^2\vartheta - 1}}{\varepsilon_r\cos\vartheta + \sqrt{\varepsilon_r + \cos^2\vartheta - 1}} \tag{4}$$

where ε_r is the ratio of the complex dielectric material parameters for the two media on the planar interface.

Figure 2 shows the stationary field excited by a ⊥ polarized point source in the upper half plane, with sinusoidal time dependence.

The reflection material's complex permittivity is ε_r = 3-1*j, and the area of investigation is 10λ*15λ. The typical interference picture shows wave front undulation and significant field strength decreasing in several directions.

Figure 2. Direct and reflected field component interference

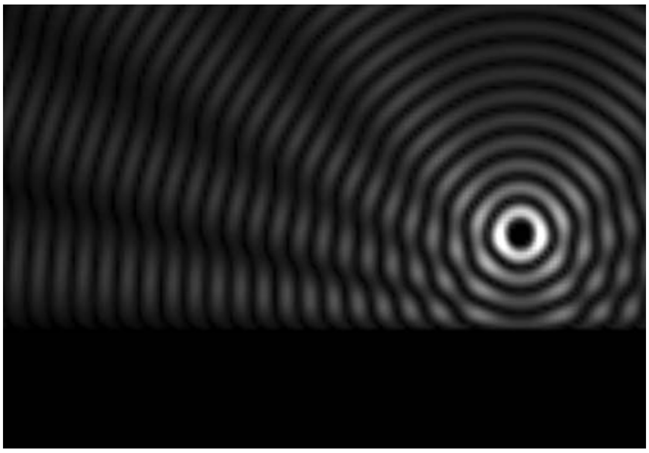
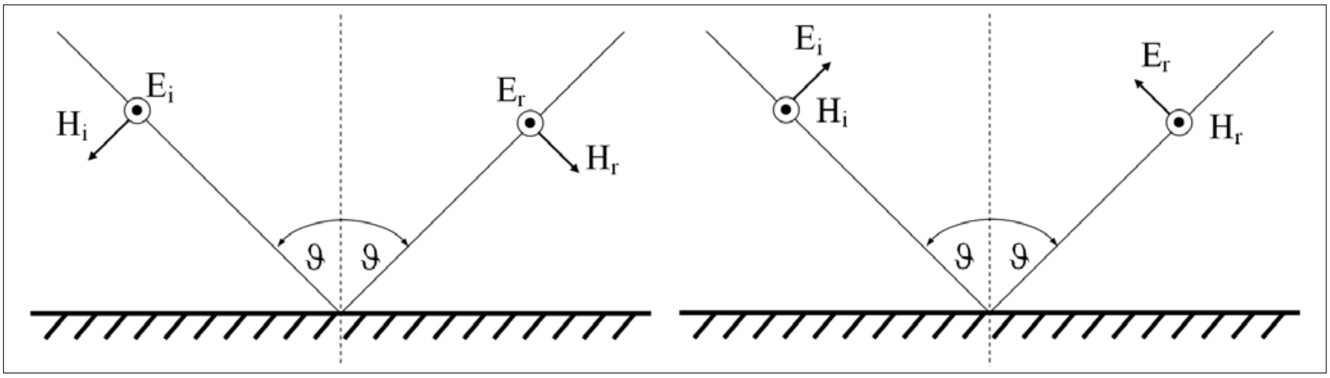


Figure 1. Perpendicular ⊥ (hard) and parallel || (soft) polarization



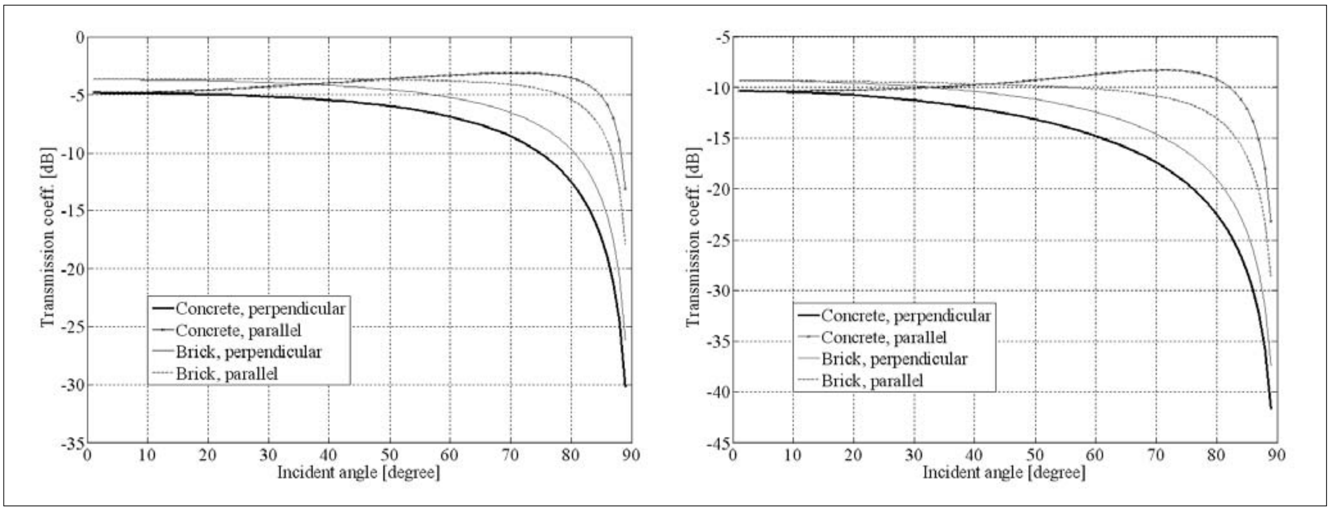


Figure 3. Transmission coefficients at 900 MHz and 2.4 GHz

2.3. Transmission

The transmission coefficient (throughput loss), $T = E_t / E_i$, is used to represent the electrical field strength ratio of the electromagnetic incident and transmitted wave at the interface of two media.

General formulation of the reflection and transmission coefficients of multiple material layers can be developed for uniform plane wave at oblique angle incidence using the transmission line theory. The model is based on the well known impedance transfer equation for transmission lines with different characteristic impedance [9]. The results in *Figure 3* present the simulated transmission coefficients as a function of the wave incidence angle for brick and concrete slabs bounded on both sides of air. The concrete and brick slab thicknesses are 12 cm with complex permittivity of ε_r = 9 -i *0.9 and ε_r = 2.8 -i*0.56 respectively.

The results show no significant difference in transmission coefficients for brick and concrete but this loss highly increases with frequency, therefore especially WLAN network areas are limited by this factor in multiple wall indoor environment. (The reinforced concrete iron layer produces additional reflection and transmission loss which can be modeled using FDTD simulation.)

The ray tracing method also uses the multiple layer transmission calculation but this simplification results in increasing error for short range radio environments where the material interfaces are in near field of the antennas or there are not bounded by plane surfaces.

The FDTD method is able to model also the previous cases, as well as the case of excitation by non sinusoidal time dependent source. *Figure 4* shows the field strength distribution for Gaussian pulse modulated sinusoidal excitation at different time with interaction of the wave by finite thickness lossy dielectric slab having permittivity of ε_r = 3-1*j. The wave decoupling into reflected and transmitted components is obvious.

2.4. Diffraction

Diffraction refers to the bending of waves around an edge of an object. Diffraction phenomenon depends on the size of the object relative to the wavelength of

Figure 5/a. Direct, reflection and diffraction regions

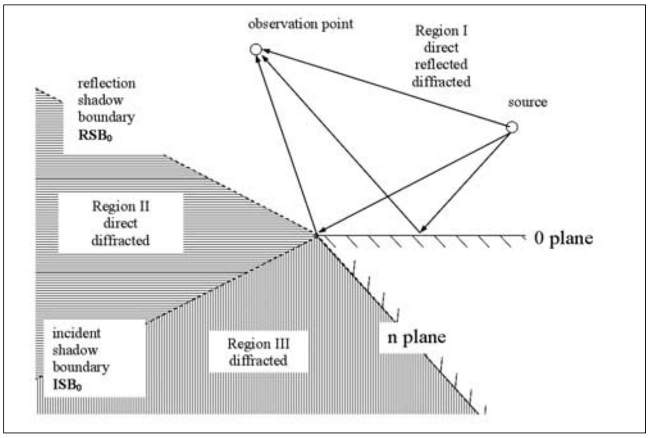
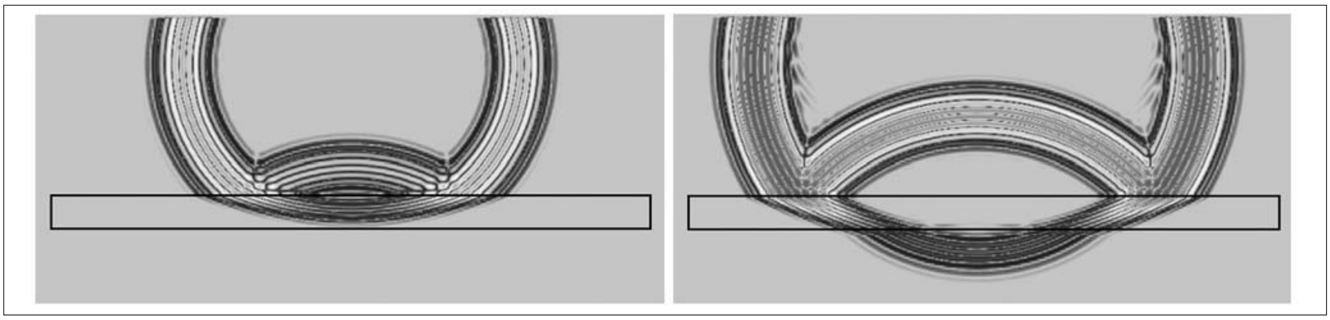


Figure 4. Wave transmission on finite thickness lossy dielectric excited by point source with ⊥ polarization



the wave. When the dimensions of the radiating object are large compared to the wavelength, high frequency asymptotic techniques can be used to analyze many, otherwise mathematically not treatable problems. Basically the application of the diffraction theory was started in the area of physics which deals with the description of the light wave propagation.

The basic concept of geometrical optics, or ray optics is in many situations inadequate to completely describe the behavior of the electromagnetic field in the shadow region, behind the diffraction objects. The diffracted field is added to calculate the field contribution in the shadow region, and that permits us to solve many practical radio wave propagation problems.

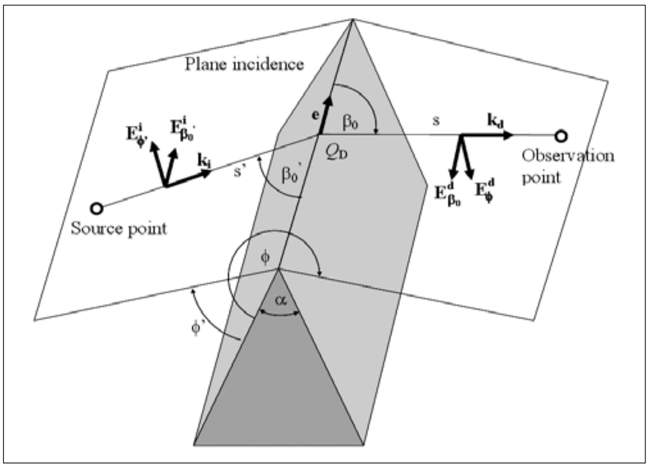


Figure 5/b. Diffraction geometry

The diffraction components are expressed as

$$\begin{bmatrix} \mathbf{E}^d_{\beta_0}(s) \\ \mathbf{E}^d_{\phi}(s) \end{bmatrix} = -\begin{bmatrix} D_s & 0 \\ 0 & D_h \end{bmatrix} \begin{bmatrix} \mathbf{E}^i_{\beta_0'}(Q_D) \\ \mathbf{E}^i_{\phi'}(Q_D) \end{bmatrix} A(s',s)e^{-jks} \qquad (5)$$

where

$\mathbf{E}^i_{\beta_0'}(Q_D)$ the component of incident electrical field parallel to the plane of incidence at the point of diffraction,

$\mathbf{E}^i_{\phi'}(Q_D)$ the component of incident electrical field perpendicular to the plane of incidence at the point of diffraction, D_s and D_h are the diffraction coefficients for soft and hard polarization.

$$A(s',s) = \begin{cases} \dfrac{1}{\sqrt{s \cdot \sin\beta_0}} & \text{for cylindrical incoming waves} \\ \sqrt{\dfrac{s'}{s(s+s')}} & \text{for spherical incoming waves} \end{cases}$$

The expression of diffraction coefficients are first derived by Keller [9] publishing the Geometrical Theory of Diffraction (GTD). The GTD diffraction coefficients possess singularities along the incident and reflection shadow boundaries and therefore in the neighborhood of these boundaries the model is inapplicable.

In the later work of Kouyoumjian and Pathak, the singularities were removed by introducing the Uniform Theory of Diffraction (UTD) and this approach is used in most wave propagation models. The regions in the neighborhood of the shadow boundaries are referred to as transition regions, and in these regions the fields undergo their most rapid changes. The diffraction coefficients are

$$D_h(\phi,\phi',n,\beta'_0) = \frac{e^{-j\pi/4}}{2n\sqrt{2\pi k}\,\sin\beta'_0}\left[D_0^{ISB} + D_n^{ISB} + R_0^h D_0^{RSB} + R_n^h D_n^{RSB}\right]$$
$$D_s(\phi,\phi',n,\beta'_0) = \frac{e^{-j\pi/4}}{2n\sqrt{2\pi k}\,\sin\beta'_0}\left[D_0^{ISB} + D_n^{ISB} + R_0^s D_0^{RSB} + R_n^s D_n^{RSB}\right] \qquad (6)$$

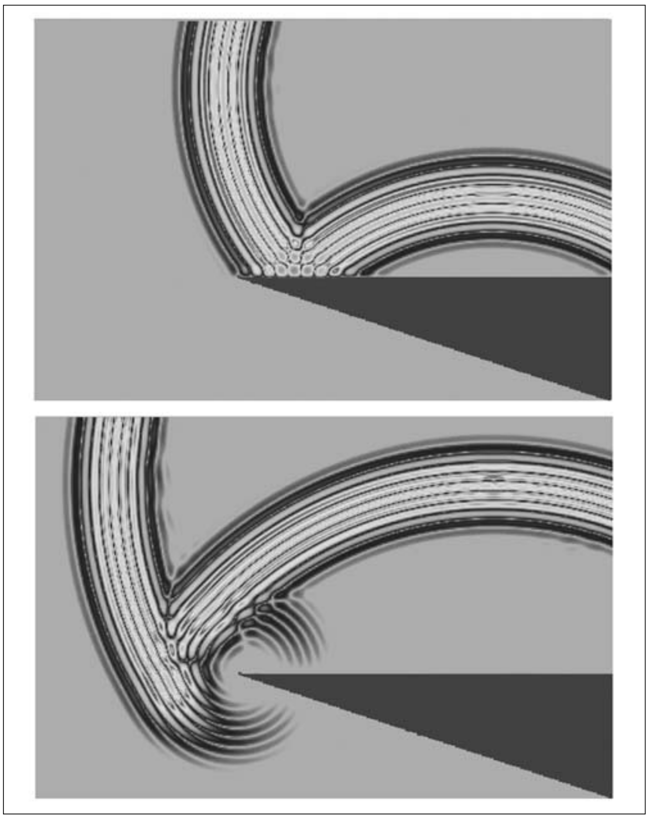
where $R_0^{h,s}$ and $R_n^{h,s}$ are the reflection coefficients on 0 and *n* planes represented in *Figure 5/a*, $D_{0,n}^{ISB,RSB}$ are the incident diffraction coefficient component at the Incident Shadow Boundary (ISB) and the reflected diffracted component at the Reflection Shadow Boundary (RSB), respectively.

Figure 6 shows a propagating wave before and after the diffraction on a lossy dielectric wedge for Gaussian modulated sinusoidal point source excitation and simulated by FDTD method.

2.5. Scattering

Rough surfaces and finite surfaces scatter the incident energy in all directions with a radiation diagram which depends on the roughness and size of the surface or volume. The dispersion of energy through scat-

Figure 6. Diffraction on a lossy dielectric wedge



tering means a decrease of the energy reflected in the specular direction. This simple view leads to account for the scattering process only by decreasing the reflection coefficient and thus, only by multiplying the reflection coefficient with a factor smaller than one, which, according to the Raleigh theory, depends exponentially on the standard deviation of the surface roughness.

3. Wave propagation models

To implement a mobile radio system, wave propagation models are necessary to determine propagation characteristics for signal and interference powers and for any arbitrary installation and any receiving positions. These models and results are the basis for the high-level network planning process. The narrow-band signals and simple building geometry make possible to apply empirical and semi-empirical models in the network design practice. The need for using deterministic models is mainly caused by the building complexity, and by the broadband or time dependent characterization of the radio channel.

3.1. Empirical and semi-empirical models

The various empirical and semi-empirical indoor propagation models use two approaches. The first is to model the propagation loss by a path loss law model determining the propagation exponent from measurements. The second one is a more successful approach to characterize indoor path loss by a fixed path loss exponent, plus additional loss factors related to the number of floors and walls intersected by the straight line between the access point and terminals. The two widely used models belonging to the second approach are the Motley-Keenan and the COST231 [10] models.

The path loss expression from the Motley-Keenan model for path distance *r* is:

$$L^{dB} = L_1 + 20\log r + n_f a_f + n_w a_w \tag{7}$$

where

L_1 is the loss at $r = 1$ m,

a_f and a_w are the attenuation factors (in dB) per floor and per wall respectively,

n_f and n_w are the number of floors and walls intersected by the radio path.

The path loss of the COST231 multi-wall model is:

$$L = L_F + L_c + \sum_{i=1}^{W} L_{wi} n_{wi} + L_f n_f^{\left[\left(n_f+2\right)\left(n_f+1\right)-b\right]} \tag{8}$$

where

L_F is the free space path loss for the straight line path,

L_c and b are empirically derived constants.

Table 1. Recommended parameters of the COST231 multi-wall indoor model for 1800 MHz

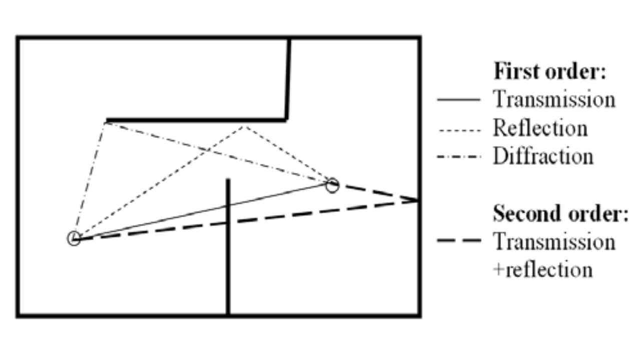
L_w	L_f	B
Light wall 3.4 dB Heavy walls 6.9 dB	18.3 dB	0.46

3.2. Deterministic models

Ray tracing

The ray tracing type radio wave propagation models are based on geometrical optics, instead of the entire domain field simulation. The method partitions the propagation waves into finite angular components, and these propagation components are traced independently and are applied to each the boundary conditions on material interfaces – reflection, transmission, diffraction. The solution on every observation points can be finally derived by summing the wave contributions.

Figure 7.
First and second order ray tracing components



The ray tracing method in practice uses either all possible first, second and third order combination of propagation mechanisms, or the ray components are traced till the field strength reaches a user defined threshold limit.

FDTD method

The FDTD (Finite Difference Time Domain) method is a time domain solution of the Maxwell's equations described in differential form and is widely used in circuit analysis because of its simplicity. The method divides the space investigated into finite grid elements and on the grid the time and space approximation of the electrical and magnetic field strength is performed [5].

There exist many various forms of the FDTD in one, two or three dimensions and for many coordinate systems or grids and material types. For the indoor wireless channel simulation the three dimensional rectangular coordinate system was chosen with linear lossy dielectric materials in volumes.

Table 2. Ray tracing and FDTD method comparison

	Ray tracing	FDTD
Features	Frequency domain solution; Narrowband, harmonic excitation	Time domain solution; Broadband, arbitrary excitation
Advantages	The problem can be split into parts	Simple programming; Simple data base structure with volume elements;

Disadvantages	Difficulty in programming;
	In case of complex or rounded geometry the significant divergence of traced rays;
	Data base pre processing needed – the volume element have to be dissolve into boundary surface elements;
	Tremendous running time;
		Tremendous running time and memory requirement

The method will be introduced briefly for the general three-dimensional case in rectangular coordinate system.

Starting from the generalized differential matrix operators, the Maxwell's curl equations can be expressed in the rectangular coordinate system as

$$\frac{\partial E_x}{\partial t}=\frac{1}{\varepsilon}\left[\frac{\partial H_z}{\partial y}-\frac{\partial H_y}{\partial z}-\left(J_{source_x}+\sigma E_x\right)\right]$$
$$\frac{\partial E_y}{\partial t}=\frac{1}{\varepsilon}\left[\frac{\partial H_x}{\partial z}-\frac{\partial H_z}{\partial x}-\left(J_{source_y}+\sigma E_y\right)\right] \qquad (9)$$
$$\frac{\partial E_z}{\partial t}=\frac{1}{\varepsilon}\left[\frac{\partial H_y}{\partial x}-\frac{\partial H_x}{\partial y}-\left(J_{source_z}+\sigma E_z\right)\right]$$

Then the Yee algorithm [5] is used for a discrete grid and, considering a substitution of central differences for the time ($\partial/\partial t$) and space ($\partial/\partial x$, $\partial/\partial y$, $\partial/\partial z$) derivatives in (9), one gets for the time marching solution of the following coupled equations. The algorithm defines the six (E_x, E_y, E_z, H_x, H_y, H_z) discretized field components in the FDTD rectangular unit cell (the Yee cell). This cell has dimensions of $\Delta x \Delta y \Delta z$ and the electric and magnetic field components locations are interleaved by half of the discretization length ($\Delta x/2$,$\Delta y/2$ and Δz).

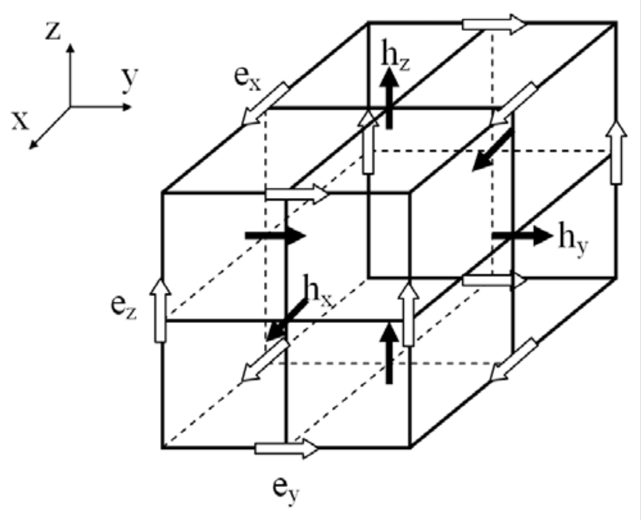


Figure 8. The 3 dimensional FDTD Yee cell with the electrical and magnetic reference vectors

In a similar manner calculating the fields every half-time step the centered difference for the time derivative is obtained.

The *x* directional electrical field strength component at the *n*+1/2 time step is:

$$E_x\Big|_{i,j+1/2,k+1/2}^{n+1/2}=\left(\frac{1-\dfrac{\sigma_{i,j+1/2,k+1/2}\Delta t}{2\varepsilon_{i,j+1/2,k+1/2}}}{1+\dfrac{\sigma_{i,j+1/2,k+1/2}\Delta t}{2\varepsilon_{i,j+1/2,k+1/2}}}\right)E_x\Big|_{i,j+1/2,k+1/2}^{n-1/2}+\left(\frac{\dfrac{\Delta t}{\varepsilon_{i,j+1/2,k+1/2}}}{1+\dfrac{\sigma_{i,j+1/2,k+1/2}\Delta t}{2\varepsilon_{i,j+1/2,k+1/2}}}\right)\cdot\left(\begin{array}{l}\dfrac{H_z\big|_{i,j+1,k+1/2}^{n}-H_z\big|_{i,j,k+1/2}^{n}}{\Delta y}\\[2ex] -\dfrac{H_y\big|_{i,j+1/2,k+1}^{n}-H_y\big|_{i,j+1/2,k}^{n}}{\Delta z}\\[2ex] -J_{source_x}\big|_{i,j+1/2,k+1/2}^{n}\end{array}\right) \qquad (10)$$

Similar finite difference equations can be expressed for the other five field strength components, E_y, E_z, H_x, H_y and H_z.

The discretization on the simulation volume is made by cubic lattice so $\Delta x = \Delta y = \Delta z = \Delta$ which results in a significant simplification of the finite difference equations.

The $\varepsilon_{i,j,k}$ and $\sigma_{i,j,k}$ are the permittivity and conductivity of the material at the *i,j,k* discretisation position.

Stability of the FDTD solution requires that the electromagnetic wave does not pass through more than one cell in one time step, i.e., the time step and the unit cell dimension satisfy the Courant condition.

The Δt time step was chosen in accordance of this magic time step

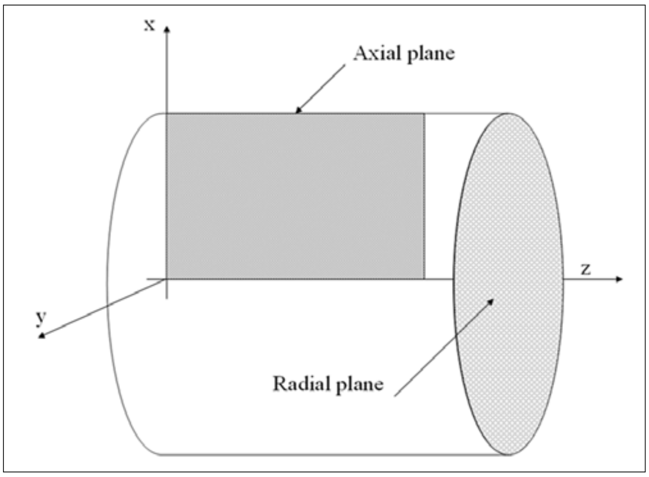
$$\Delta t \le \frac{1}{c\sqrt{\dfrac{1}{(\Delta x)^2}+\dfrac{1}{(\Delta y)^2}+\dfrac{1}{(\Delta z)^2}}} \qquad (11)$$

which results $\Delta t \le \Delta/(c\sqrt{3})$ for our cubic lattice.

Special two dimensional geometries

The first investigation showed that the determination of field strength distribution in tunnel using full 3D model extends our calculation possibilities. Therefore we decided to model our geometry in two cut planes, namely in the axial and in the radial plane. The two approximations differ basically because the axial cut plane can be applied in case of rotationally symmetric geometry and the radial plane is a cut of the tunnel waveguide which is assumed non changing cross section *(Figure 9)*.

Figure 9. Axial and radial cut planes



Axial plane

For the derivation of the cylindrical FDTD equations starting from the generalized differential matrix operators, we express the Maxwell's curl equations in the cylindrical coordinate system as

$$\nabla\times\mathbf{E}=\begin{vmatrix}\mathbf{e}_r & \mathbf{e}_\varphi & \frac{1}{r}\mathbf{e}_z\\ \frac{\partial}{\partial r}r & \frac{\partial}{\partial\varphi} & \frac{\partial}{\partial z}\\ E_r & E_\varphi & \frac{1}{r}E_z\end{vmatrix}=-\frac{\partial\mu\mathbf{H}}{\partial t}+\sigma^m\mathbf{H} \qquad (12)$$

$$\nabla\times\mathbf{H}=\begin{vmatrix}\mathbf{e}_r & \mathbf{e}_\varphi & \frac{1}{r}\mathbf{e}_z\\ \frac{\partial}{\partial r}r & \frac{\partial}{\partial\varphi} & \frac{\partial}{\partial z}\\ H_r & H_\varphi & \frac{1}{r}H_z\end{vmatrix}=\frac{\partial\varepsilon\mathbf{E}}{\partial t}+\sigma^e\mathbf{E} \qquad (13)$$

where

ε is the permittivity,

μ is the permeability,

σ^e is the electric conductivity,

σ^m is the magnetic conductivity.

The φ variation of **E** and **H** in the cylindrical coordinates system will have the following form

$$\mathbf{E},\mathbf{H}=\sum_{m=0}^{\infty}\left[(\mathbf{e}_u,\mathbf{h}_u)\cos m\varphi+(\mathbf{e}_v,\mathbf{h}_v)\sin m\varphi\right] \qquad (14)$$

where

m is the mode number.

Using the cylindrical symmetry of the geometry the 3D equations are reduced to 2D in the (x-z) r-z plane.

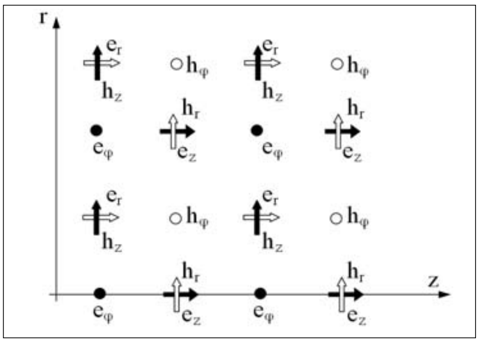


Figure 10. Electric and magnetic field components on the axial plane

After discretizing the equations 12-14 and applying the finite difference approximations to yield the updating equations for each field components [5]:

$$E_r\Big|_{i,k}^{n+1}=\left(\frac{1-\dfrac{\sigma_r^e\Delta t}{2\varepsilon_0\varepsilon_r}}{1+\dfrac{\sigma_r^e\Delta t}{2\varepsilon_0\varepsilon_r}}\right)E_r\Big|_{i,k}^{n}-\left(\frac{\dfrac{\Delta t}{\varepsilon_0\varepsilon_r}}{1+\dfrac{\sigma_r^e\Delta t}{2\varepsilon_0\varepsilon_r}}\right)\left[\frac{H_\varphi\Big|_{i,k}^{n+1/2}-H_\varphi\Big|_{i,k-1}^{n+1/2}}{\Delta z}-\frac{m}{(i)\Delta r}H_z\Big|_{i,k}^{n+1/2}\right] \qquad (15)$$

$$E_\varphi\Big|_{i,k}^{n+1}=\left(\frac{1-\dfrac{\sigma_\varphi^e\Delta t}{2\varepsilon_0\varepsilon_\varphi}}{1+\dfrac{\sigma_\varphi^e\Delta t}{2\varepsilon_0\varepsilon_\varphi}}\right)E_\varphi\Big|_{i,k}^{n}+\left(\frac{\dfrac{\Delta t}{\varepsilon_0\varepsilon_\varphi}}{1+\dfrac{\sigma_\varphi^e\Delta t}{2\varepsilon_0\varepsilon_\varphi}}\right)\left(\begin{matrix}\dfrac{H_r\Big|_{i,k}^{n+1/2}-H_r\Big|_{i,k-1}^{n+1/2}}{\Delta z}\\ -\dfrac{H_z\Big|_{i,k}^{n+1/2}-H_z\Big|_{i-1,k}^{n+1/2}}{\Delta r}\end{matrix}\right) \qquad (16)$$

where

$$i=r/\Delta r,\ k=z/\Delta z,\ n=t/\Delta t$$

Radial plane

The main steps will be introduced briefly for the two-dimensional rectangular coordinate system. Starting from the generalized differential matrix operators, the Maxwell's equations can be express in the rectangular coordinate system as

For TM$_z$ case:

$$\frac{\partial E_z}{\partial t}=\frac{1}{\varepsilon}\left[\frac{\partial H_y}{\partial x}-\frac{\partial H_x}{\partial y}-J_{source_z}\right]$$

$$-\mu\frac{\partial H_x}{\partial t}=\frac{\partial E_z}{\partial y} \qquad (17)$$

$$-\mu\frac{\partial H_y}{\partial t}=\frac{\partial E_z}{\partial x}$$

For TE$_z$ case similar expressions can be introduced.

Then the Yee algorithm is used for a discrete grid and considering a substitution of central differences for the time and space derivatives in (17) one gets for the time marching solution of the following coupled equations [5]:

$$H_x\Big|_{j+1/2,k}^{n+1/2}=H_x\Big|_{j+1/2,k}^{n-1/2}-\frac{\Delta t}{\mu}\frac{E_z\Big|_{j+1/2,k+1/2}^{n}-E_z\Big|_{j+1/2,k-1/2}^{n}}{\Delta y} \qquad (18)$$

$$H_y\Big|_{j,k+1/2}^{n+1/2}=H_y\Big|_{j,k+1/2}^{n-1/2}+\frac{\Delta t}{\mu}\frac{E_z\Big|_{j+1/2,k+1/2}^{n}-E_z\Big|_{j-1/2,k+1/2}^{n}}{\Delta x} \qquad (19)$$

$$\begin{aligned}E_z\Big|_{j+1/2,k+1/2}^{n+1}&=\frac{1-\xi}{1+\xi}E_z\Big|_{j+1/2,k+1/2}^{n}+\\&+\frac{1}{1+\xi}\frac{\Delta t}{\varepsilon}\frac{H_y\Big|_{j+1/2,k+1}^{n+1/2}-H_y\Big|_{j+1/2,k}^{n+1/2}}{\Delta x}-\\&-\frac{1}{1+\xi}\frac{\Delta t}{\varepsilon}\frac{H_x\Big|_{j+1/2,k+1}^{n+1/2}-H_x\Big|_{j+1/2,k}^{n+1/2}}{\Delta y}\end{aligned} \qquad (20)$$

where

$$\xi=\frac{\sigma\Delta t}{2\varepsilon}$$

The discretization is made by cubic lattice on the simulation volume, and $\Delta x=\Delta y=\Delta$.

4. The data base requirement of the indoor wave propagation models

The geometrical description of the indoor scenarios are based on the same concept both for ray tracing and for FDTD methods The walls have to be partitioned to surrounding closed polygons and every such polygons are characterized by its electric material parameters.

The data base for the ray tracing method in our applications can not contain cut-out surfaces directly, such as windows, doors. Therefore the cut-out surface description is based on surface partitioning of the geometry as can be seen in *Figure 11*. The FDTD algorithm, on the contrary, allows application of simple ordinal database with overlapping polygons and the Yee cell parameter will be chosen by the simple decision that whichever object is higher on the list will overwrite the lower object in the mesh.

We prepared the indoor data base for a typical office building V2 at the Budapest University of Technology and Economics (BUTE), which has 7 floors and is made partially from concrete and brick. This data base for the sixth floor is used to make verification calculations and measurements of indoor wave propagation models and to develop new models. The floor plane is shown in *Figure 18*, and the floor view and polygonal partitioning in *Figure 13*, which is based on the previously described concept.

The data base contains the polygon coordinates, multiple and single layer wall types on the basis of *Table 3 (on the next page)*, for which the electrical material parameters of *Table 4* are used.

These parameters are partly results coming from literature, and from own material parameter measurements [14].

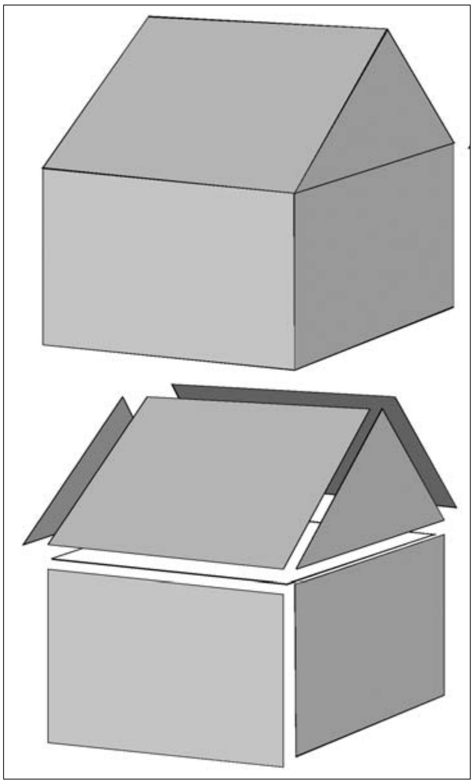


Figure 11. Polygon representation of building structure

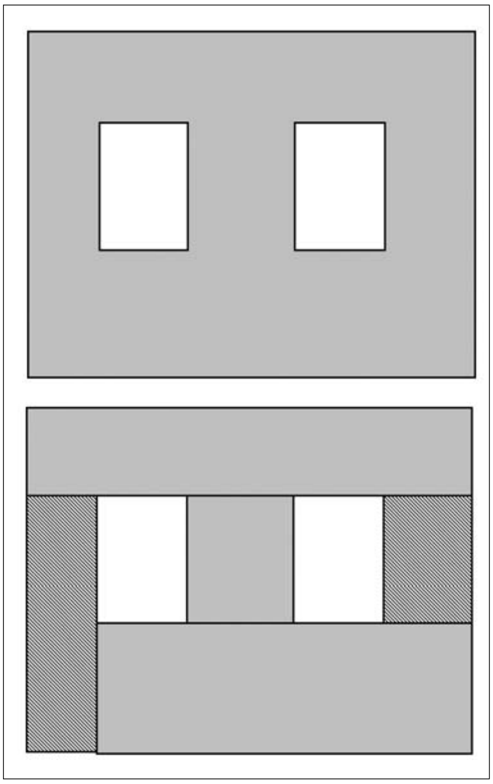


Figure 12. A possible polygonal partitioning of windowed walls for ray tracing method

Figure 13. Floor view and polygon data base of V2 building at BUTE

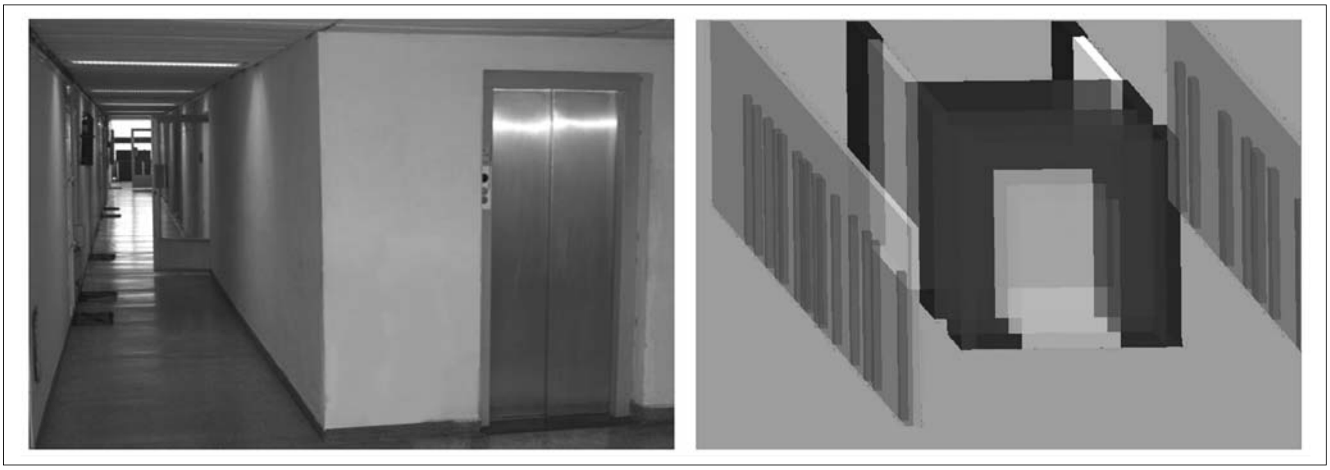


Table 3. Main wall types of indoor data base

Type	Nr. of layers	Layer thicknesses
Brick	1	Brick – 6 cm
Brick	1	Brick – 10 cm
Brick	1	Brick – 12 cm
Brick + Concrete	3	Brick – 6 cm, Concrete – 20 cm, Brick – 6 cm
Brick + Concrete	3	Brick – 10 cm, Concrete – 12 cm, Brick – 10 cm
Brick + Concrete	3	Brick – 10 cm, Concrete – 10 cm, Brick – 10 cm
Brick	1	Brick – 15 cm
Concrete	1	Concrete – 15 cm
Concrete	1	Concrete – 20 cm
Concrete	3	Concrete – 15 cm, Air – 2 cm, Concrete – 15 cm
Glass	3	Glass – 3 mm, Air - 10 cm, Glass – 3 mm
Plasterboard	1	Plasterboard – 5 cm
Wood	1	Wood – 6 cm
Wood	1	Wood – 6 cm

These parameters of materials are specified either by the permittivity and loss tangent or directly by the complex permittivity, where $\varepsilon_r = \varepsilon'_r + j \cdot \varepsilon''_r = \varepsilon'_r - j \cdot (\tan\delta) \cdot \varepsilon'_r$.

Table 4. Electrical parameters of building materials

	Permittivity ε'_r	Loss tangent $\tan\delta$
Wood	3.5	0.01
Paper	3	0.008
Glass	5.5	0.001
Brick	2.8	0.2
Concrete	9	0.1

5. Application, results

Two ANSI C code were generated using the previous theory of general three-dimensional and of special two-dimensional FDTD methods for the two main cut of the cylindrical geometries. The methods are verified and results are presented for wave propagation problems.

5.1. Field strength distribution in the main two cut planes of tunnel

The application of the theory presented in Section 3 makes possible to investigate the mobile radio coverage in tunnels [13]. The axial plane field strength distribution in a tunnel with radius 2 m at the distance of 1 m from the axis of symmetry for sinusoidal excitation source by frequency of 900 MHz. The results are in good agreement with analytical results of literature [4].

The gradient of the linear regression to the FDTD simulation results in *Figure 14* is 9 dB/decade which indicates the waveguide nature of the tunnel at the frequency of investigation.

Figure 14. Electrical field strength vs axial distance in tunnel

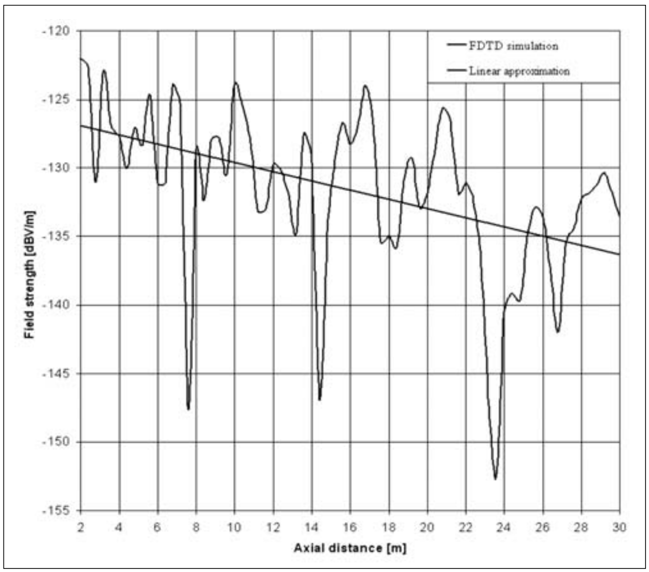
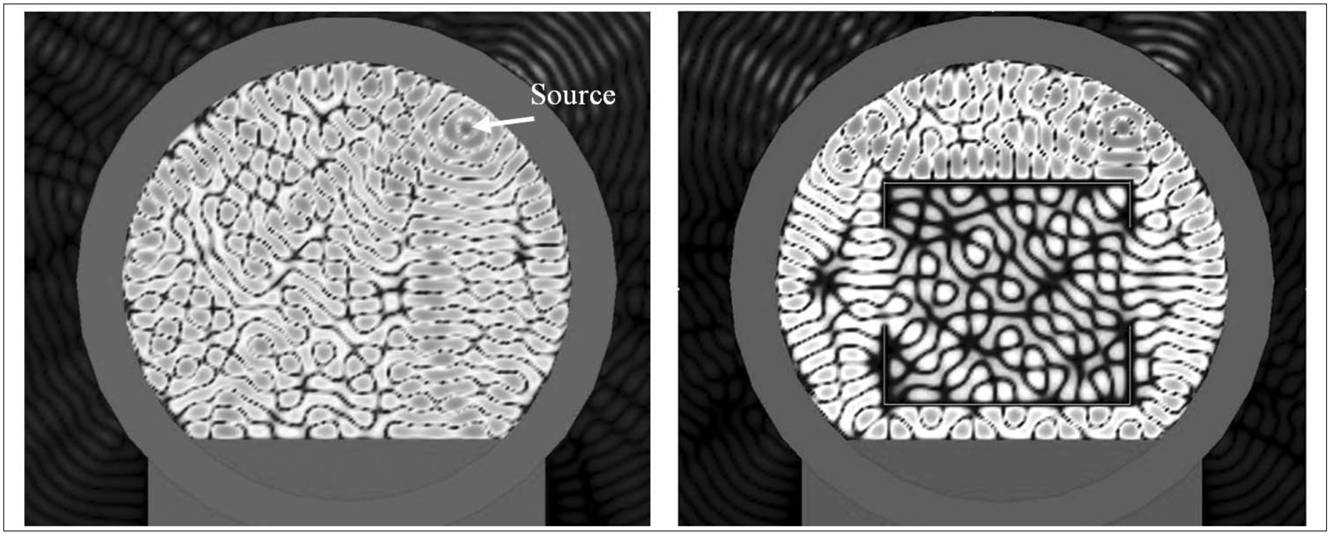


Figure 15. Field strength distribution of radial plane in tunnel without and with vehicle



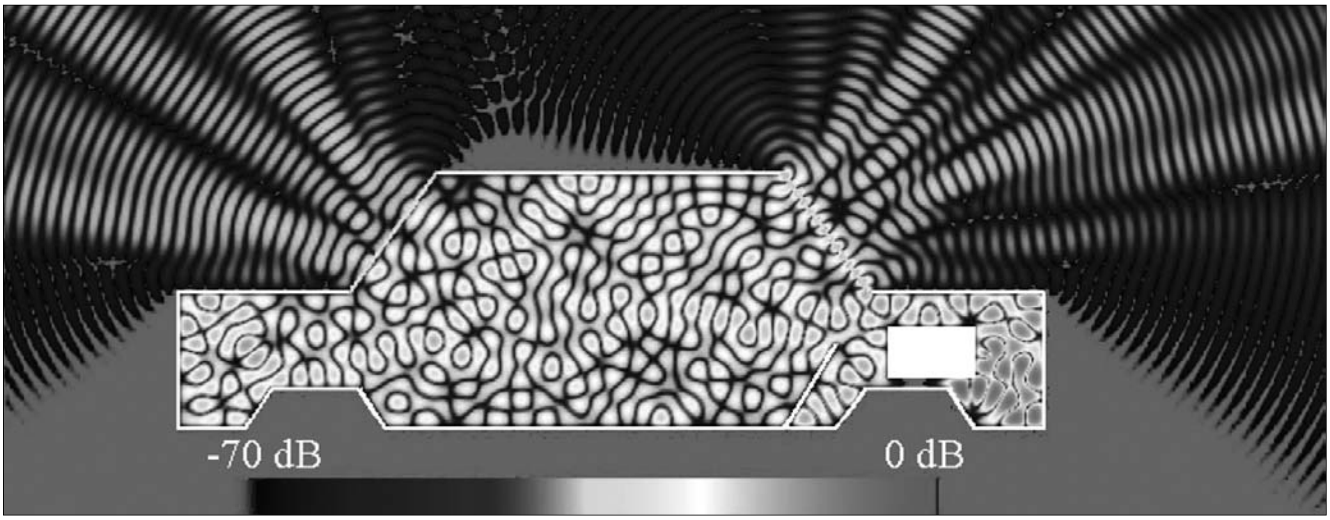


Figure 16. Stationary field strength distributions inside cars and radiated outside, caused by disturbing source from engine room at 1800 MHz

Our results in radial plane are shown as a two-dimensional electric field strength distribution in the tunnel and inside of the carriage for a point source with sinusoidal excitation.

The results illustrate very well how the field strength changes in presence of the underground carriage and how the propagating energy flows inside. The radial plane investigation gives a good opportunity for the optimization of the leaky cable for tunnel coverage and to determine the base station transmission power requirement for proper coverage inside of the vehicles.

Our next application areas are the EMC/EMI problems and additionally health risk analysis of electromagnetic waves. We demonstrate the stationary field strength distribution inside cars, caused by disturbing source from engine room *(Figure 16)*. Our next example is the induced electromagnetic field inside buildings generated by GSM base station *(Figure 17)*.

The GSM base station coverage area at frequency of 900 MHz was investigated with FDTD using two-dimensional grid of 1 cm discretization and 1500x1700 grid size of area of simulation. The time step is specified as 19 ns and 3000 steps were calculated using PC with Centrino Duo processor working at 1.83 GHz. The simulation took 20 minutes and required 140 MB of operational RAM memory to store the field strengths. The walls are modeled as brick walls and as concrete ceiling, each having of 10 cm thickness.

In *Figure 17* in the interior space the field strengths at points 2 and 3 are less by a factor of 10^{-4} and 10^{-6} compared to the value at point 1. Using and improving the investigations outlined above the field levels of safety standards can be validated.

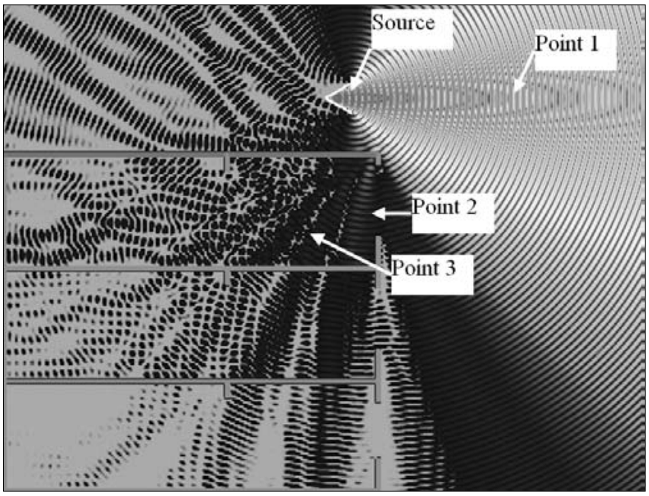


Figure 17. Stationary field strength distribution excited by a GSM base station at 900 MHz

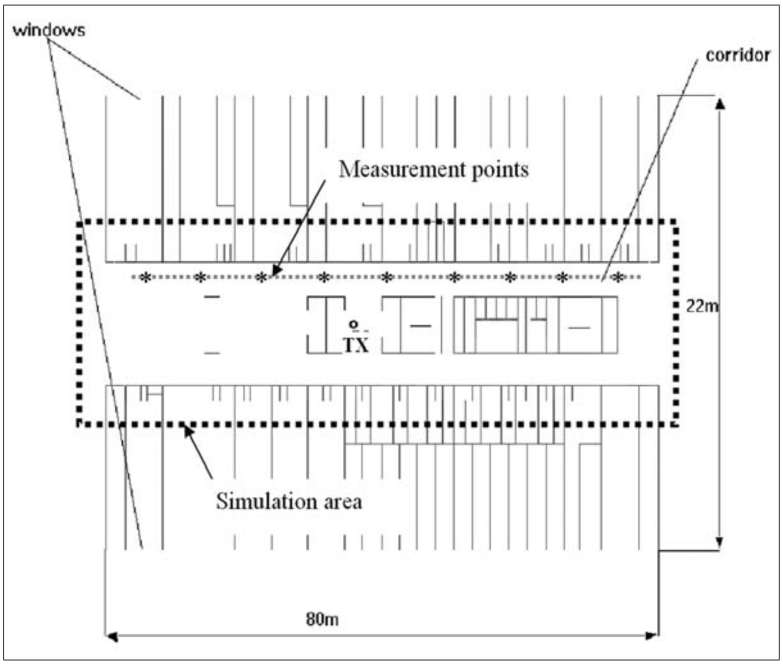


Figure 18. Indoor measurement scenario, floor plan with simulation area and measurement route

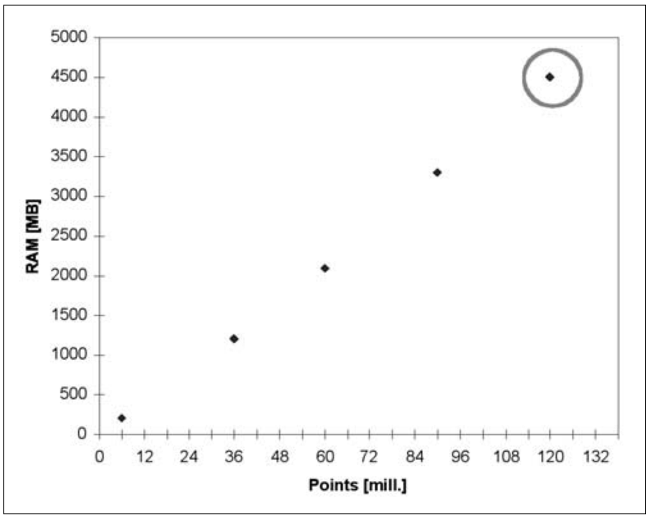


Figure 19.
Operational memory requirement of simulation

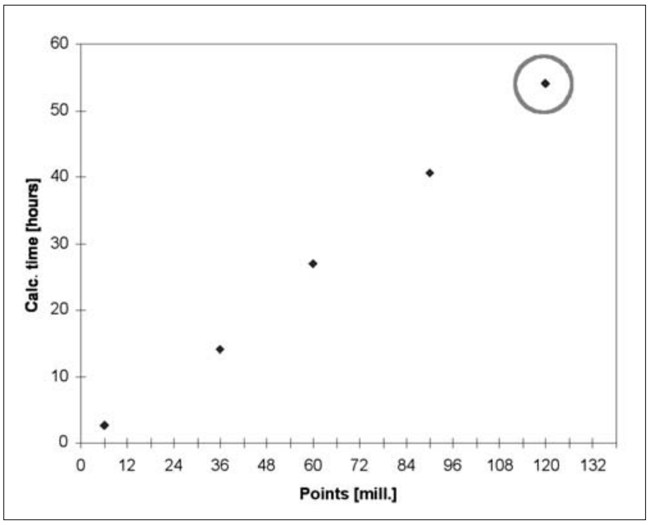


Figure 20. Running time requirement of simulation

Our last example will be the qualification of LPD short range radio links in indoor environment. The measurement and calculation results are shown at frequency of 433 MHz.

The building floor plan with measurement points on corridor and the area of simulation are presented in *Figure 18*. The simulation volume for FDTD has rectangular 90x11x3 cubic meter three dimensional size, and λ/20≈30mm resolution of the FDTD results 120 million of Yee cells.

The running time and memory requirements of the simulation program is demonstrated in *Figures 19 and 20*, the values for present simulation are indicated.

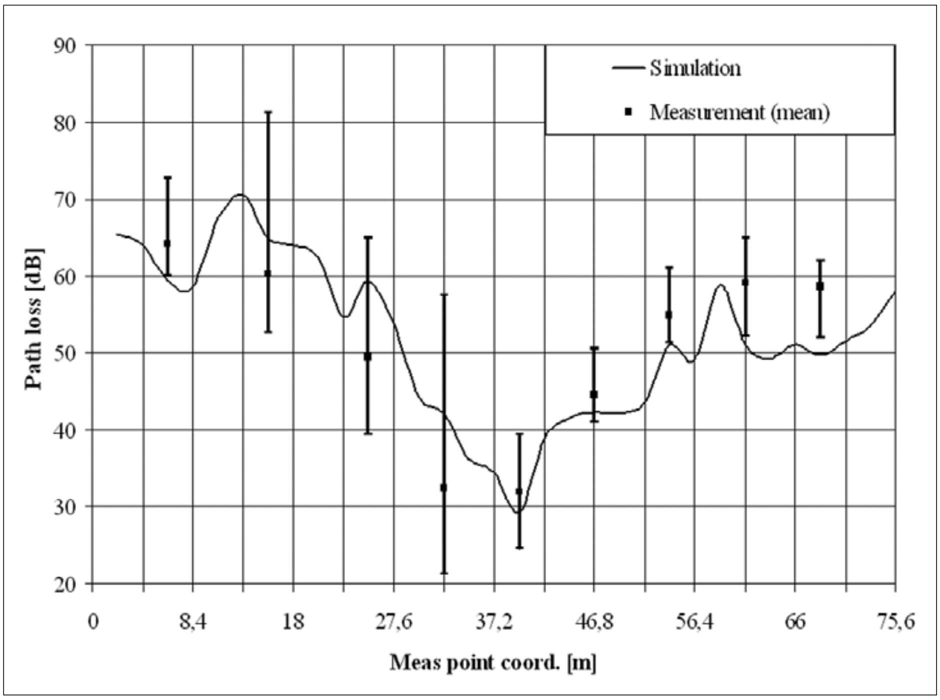
Two quarter wavelength dipole was manufactured and vertically placed to perform the wireless channel measurements, spectrum analyzer was used for the received power and the channel attenuation measurements. At each spatial measurement points 500 received strength levels were measured and buffered for later processing. *Figure 21* compares measured minimum, maximum and average levels to simulated ones.

Excitation with sinusoidal time dependence is used to the FDTD simulation and at each spatial measurement points the recorded time function is transformed to frequency domain using Fourier transform. The received field level and radio link loss at frequency of 433 MHz was taken from this spectral distribution.

The comparative analyzis of simulations and measurements results in average difference of -1.74 dB and standard deviation of 15.5 dB. The average difference shows a fair agreement but the difficulty in channel modeling is indicated by the notable deviation and therefore the received field level estimation at individual spatial points only possible with high probability of error.

The most important parameter of the radio network design is the path loss exponent of the distance parameter, which is introduced for our investigation using the measurement results. The free space and two rays propagation models have path loss exponents of n=2 and n=4 respectively. On the contrary, our indoor short distance measurements have exponent of n=4.65 at frequency of 433 MHz, which was derived using linear regression to measurement path loss results presented in *Figure 22*.

Figure 21.
Indoor path loss level comparison at frequency of 433 MHz (min, mean and max values for measurements are indicated)



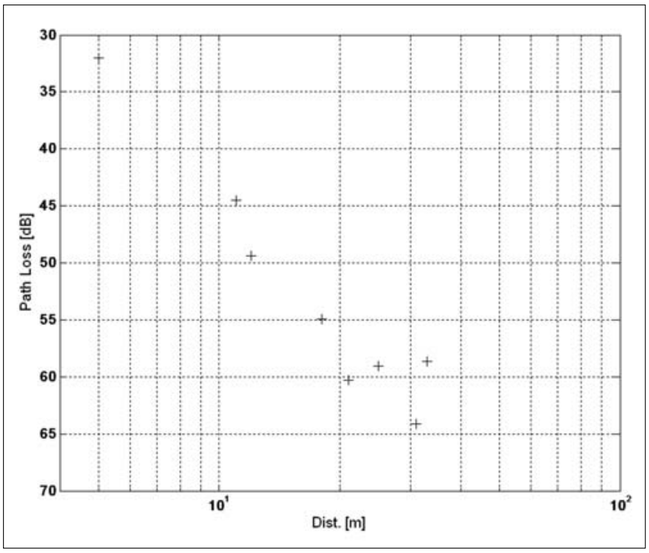


Figure 22.
Path loss dependence of measurement result at frequency of 433 MHz

6. Summary

Figures 19 and 20 present memory requirement of 4.5 GB and running time of 55 hours for the last 3 dimensional short range radio simulations. The resource requirement of simulation is increasing approximately by a factor of *frequency*^{3} for three-dimensional, and of *frequency*^{2} for two-dimensional geometry, and therefore the FDTD method can be applied only with significant limitations in simulation volumes, especially for simulations at higher frequencies .

The pure FDTD can improve by combining with ray tracing method. In this concept the ray tracing is performed to the border of a volume of detailed investigation and in this surrounding space of the receiver antenna FDTD method is applied using the ray traced field as excitation. Specific such problem is the characterization of MIMO (Multiple Input Multiple Output) radio channel, for which the combined method can be introduced. [12]

Acknowledgment

This work was carried out in the framework of Mobile Innovation Center, Hungary (Mobil Innovációs Központ).

References

[1] A. von Hippel,
Dielectric Materials and Applications,
Artech House, Boston, 1995.

[2] Lukas Müller, Walter Vollenweider,
Measurements of Radio Propagation in Buildings,
LPRA Conference, Birmingham, England,
October 29-31, 1996.

[3] Lambertus J.W. van Loon,
Mobile In-Home UHF Radio Propagation for
Short-Range Devices,
IEEE Antennas and Propagation Magazine,
Vol. 41., No.2, April 1999.

[4] Donald G. Dudley,
Wireless Propagation in Circular Tunnels,
IEEE Trans. Antennas Propagat.,
Vol. 53., pp.435–441., 2005.

[5] Allen Taflove, Susan C. Hagness,
Computational Electrodynamics:
The finite-difference time-domain method,
Artech House, Norwood, 2005.

[6] V. Rodrigez-Pereyra, A.Z. Elsherbeni, C.E. Smith,
A Body of Revolution Finite Difference
Time Domain Method with Perfectly Matched Layer
Absorbing Boundary, PIERS 24, pp.257–277., 1999.

[7] Yee, K.S.,
Numerical Solution of Initial Boundary Value Problems
Involving Maxwell's Equations in Isotropic Media,
IEEE Trans. Ant. Prop., Vol. 14., No.3, p.302., 1966.

[8] H. L. Bertoni,
UHF Predictions for Wireless Personal
Communications, Proceedings of the IEEE,
Vol. 82., No.9, pp.1333–1356., 1994.

[9] Constantine A. Balanis,
Advanced Engineering Electromagnetics,
John Wiley & Sons, 1989.

[10] Simon R. Saunders,
Antennas and Propagation for
Wireless Communication Systems, Wiley, 1999.

[11] Lajos Nagy,
FDTD Field Strength Prediction for Mobile Microcells,
ICECOM 2005 – 18th International Conference
on Applied Electromagnetics and Communications,
Dubrovnik, Croatioa, 12-14. October 2005.

[12] Lajos Nagy,
MIMO cube in realistic indoor environment,
EuCAP 2006 – The European Conference on
Antennas and Propagation,
Nice, France, 6-10. November 2006.

[13] Lajos Nagy,
Propagation modeling in subway tunnel using FDTD,
EuCAP 2006 – The European Conference on
Antennas and Propagation,
Nice, France, 6-10. November 2006.

[14] Lajos Nagy,
An Improved TDR Method for Determining Material
Parameters, XXIII. General Assembly of the URSI,
Prague, 1990.

Multipath propagation fade duration modeling of Land Mobile Satellite radio channel

LÁSZLÓ CSURGAI-HORVÁTH, JÁNOS BITÓ

Budapest University of Technology and Economics, Faculty of Electrical Engineering and Informatics
Department of Broadband Infocommunications and Electromagnetic Theory
{csurgai, bito}@mht.bme.hu

Keywords: multipath propagation, fade duration, LMS, partitioned Markov chain, Fritchman model

This contribution presents a modeling method of the fade duration caused by multipath propagation on a land mobile satellite channel. The model is based on the measurement of a satellite channel and applied to calculate the model parameters. The proposed model is based on a partitioned Fritchman's Markov chain which is applicable to calculate the complementary cumulative distribution function of the fade duration process. The dependency of the model parameters on the attenuation threshold will be also shown. Therefore the model will be available to calculate the fade duration distribution for any threshold what can be applied later in attenuation time series synthesis.

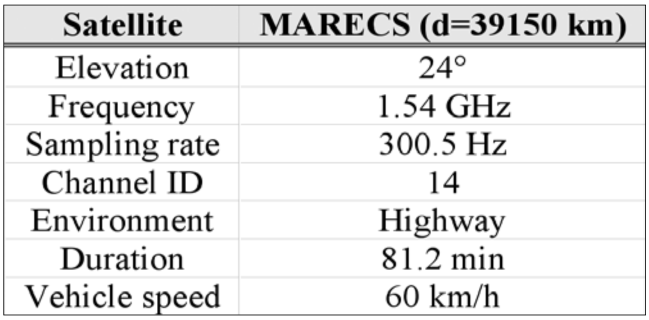
1. Introduction

The propagation on a Land Mobile Satellite (LMS) radio link is highly influenced by the shadowing effects of buildings and vegetation, or by the multipath propagation. This kind of fading arises due the multiple reflexions of the radio waves on the surrounding objects; therefore not only the direct signal is received. The characteristics of fading highly depend on the surroundings. During the design of LMS radio links one can apply the distribution function of the attenuation or the fade duration statistics to determine the fluctuation of the received signal. The fade duration is an important dynamic parameter of the path attenuation which gives the duration of fading higher than a given attenuation threshold. Therefore the fade duration is always calculated for multiple threshold levels.

In our contribution a digital model with Markov chain will be introduced, which is also applicable to determine the statistical parameters of the fade duration. The model is based on the measurement data of a real LMS channel what has been used to calculate the model parameters.

The proposed model is a partitioned Fritchman's Markov chain which is applicable to describe the stochastic fade duration process and also to calculate the Complementary Cumulative Distribution Function (CCDF) of the fade duration. The expressions to calculate the model parameter dependency on the threshold level will be also introduced. Therefore the model will be applicable to calculate the CCDF of fade duration for any desired threshold level which may lead us to the synthesis of attenuation time series in the future.

Table 1. Parameters of the measured LMS channel



Satellite	MARECS (d=39150 km)
Elevation	24°
Frequency	1.54 GHz
Sampling rate	300.5 Hz
Channel ID	14
Environment	Highway
Duration	81.2 min
Vehicle speed	60 km/h

2. Description of the measured LMS channel

To investigate and model the LMS radio channel we applied real measurement data as a starting point. The measurements have been performed by the DLR (Deutsches Zentrum für Luft- und Raumfahrt) between 1984 and 1987 [1], the parameters are detailed in *Table 1*.

The connection is the radio channel of the geostationary satellite Marecs operating at 1.54 GHz in the L-band. The measurement has been performed on highway on the board of a vehicle moving with 60 km/h speed, the measurement duration was 81.2 minutes. During the measurement the received power has been sampled with 300.5 Hz frequency and the data were recorded after normalization. The normalization was so performed that the average received power of 0 dBm represents the level of the fading-free signal.

During the measurement described above, due the movement of the receiver the receiving path has been crossed with different objects and as a result of the changing in the reflexion environment the signal arrived to the receiver on multiple paths. The above effects result in multipath fading which can be stochastically modeled as it will be described in the next sections.

3. Fade duration on the radio channel

The fade duration is one of the most important dynamic parameter of the attenuation on radio connections; it gives the time length when the attenuation is higher than the given threshold. The precise estimation of fade duration is essential when designing wireless communication systems like BFWA, B3G, 4G mobile systems or LMS channels. In calculations of the system outage or availability time or when sharing the resources or selecting different coding methods, the measured or modeled fade duration statistics plays an important role. The interfade duration is of similar importance, which is the duration between two consecutive fadings and its calculation and modeling is very similar to the method what we apply in the case of fade duration.

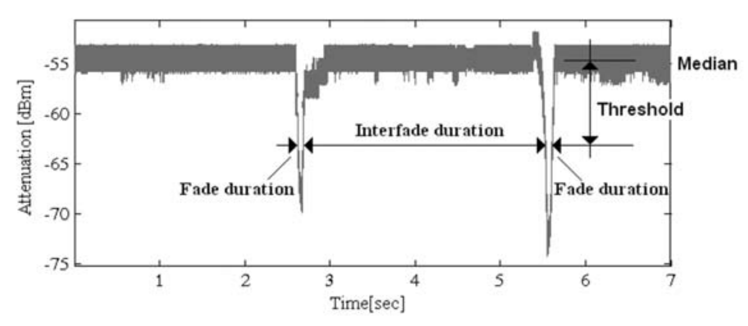


Figure 1.
Measured attenuation time series with fading and inter-fading

Figure 1 shows a typical attenuation time series with multiple fade events indicating the fade and interfade durations, respectively.

The Complement Cumulative Distribution Function (CCDF) of the received power on a radio channel is a first order statistics what is often depicted to qualify the channel or rather the radio connection. The fade duration is determined usually relatively to the median level of the received power for different thresholds, then the complementary distribution of the number of fade events are depicted as the function of fade duration.

4. Modeling with partitioned Markov chain

The ITU-R (International Telecommunication Union, Radio communication Sector) proposes a two-component model for fade duration [2], which models the fast fading with log-normal distribution and the slow fading with power-law one, ensuring the smooth transition between the two stages.

To model the fading process caused by the multipath propagation we propose a Markov model, which can be applied not only to the stochastic modeling of the fading process but it allows the exact calculation of the fade duration distribution for different thresholds. Comparing with the ITU-R model, this digital model handles uniformly the short and long fading, respectively. In the model we apply an *N*=5 state partitioned Fritchman's Markov chain [3], where 4 states are representing the fading and one state the inter-fading events *(Figure 2)*.

The transition probabilities p_{ij} of the Markov chain are depicted in *Figure 2* and its feature is that there are no transitions between the states in the same partition. This is the simplification in the Fritchman's model and it can be applied because the states in a partition are representing same type but different length of events – in our case fade and interfade events – therefore it can be supposed that there are no transitions between them.

The transition matrix of the model can be written according to the Equation (1), where we can observe the absence of transitions at the specific places.

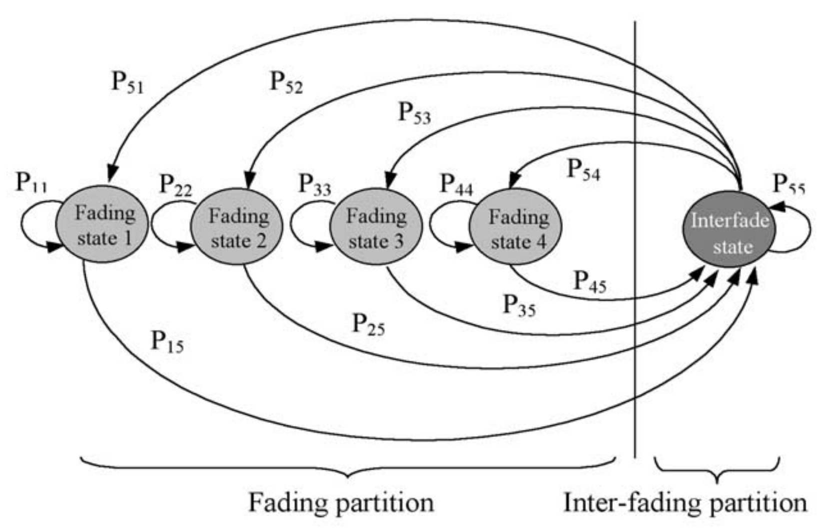
$$\overline{P} = \begin{pmatrix} p_{11} & 0 & 0 & 0 & 1-p_{11} \\ 0 & p_{22} & 0 & 0 & 1-p_{22} \\ 0 & 0 & p_{33} & 0 & 1-p_{33} \\ 0 & 0 & 0 & p_{44} & 1-p_{44} \\ p_{51} & p_{52} & p_{53} & p_{54} & 1-\sum_{i=1}^{4} p_{5i} \end{pmatrix} \qquad (1)$$

This model – unlike to the multiple-state models, where every attenuation level corresponds to a state in the Markov model – can describe the stochastic behavior of the fading process with the distribution functions of the partitions. This model has been developed by Fritchman to characterize burst errors on binary communication channels, what we have been adapted to the fading process.

The Fritchman's model can be applied to calculate the complementary cumulative distribution function of the fade and interfade duration [3] according to the Equations (2) and (3):

$$F_F^C(n) = \sum_{i=1}^{N-1} \frac{p_{Ni}}{p_{ii}} p_{ii}^n \qquad (2)$$

Figure 2.
5 state partitioned Fritchman's Markov model



$$F_I^C(n) = (\sum_{j=1}^{k} \frac{Z_j p_{jN}}{Z_F p_{NN}}) p_{NN}^n \qquad (3)$$

where N=5 is the number of states and p_{ij} is the probability of state transitions.

In Equation (3), Z_i denotes the steady state probabilities, and Z_F is the fading partition probability. They can be calculated using Equations (4) and (5):

$$Z_N = \frac{1}{1+\sum_{i=1}^{k} \frac{p_{Ni}}{p_{iN}}} \qquad (4)$$

$$Z_i = \frac{p_{Ni}}{p_{iN}} Z_N \qquad (5)$$

Expression (2) gives the probability of a fade event being longer than the given duration. Earlier investigations, described in [5], showed that the modeling of the inter-fade duration and calculating its complementary distribution with the Equation (3) can not be applied with proper accuracy because there is only one state representing it in the Markov chain. Therefore, to model the interfade duration with appropriate precision, a different Fritchman's model must be applied, a similar one as the one depicted on *Figure 2*, where multiple states are assigned to the interfade duration and one state to the fade duration.

5. Model parameterization

In case of Markov models the parameterization is usually the process of determination of the transition matrix elements. The Fritchman's model is widely used due its relatively simple parameterization and the correct representation of the modeled process. We should also take into account that the Fritchman's model with a single error state can be applied only to model communication channels with renewing feature [6].

To determine the parameters of the Markov chain (see Figure 2) we apply the gradient method as described in [4]. The point of the method is that the logarithmical CCDF of the measured fade duration can be approximated with linear according to the equation (6), afterwards the transition probabilities of the Markov chain can be determined from the line parameters.

$$\log(F_F^C(n)) = \log(\sum_{i=1}^{N-1} \frac{p_{Ni}}{p_{ii}} p_{ii}^n) \approx \left\{ \begin{array}{l} \text{for small n:} \\ n\log(p_{N-1N-1}) + \log(\frac{p_{NN-1}}{p_{N-1N-1}}) \\ \text{for large n:} \\ n\log(p_{11}) + \log(\frac{p_{N1}}{p_{11}}) \end{array} \right\} \qquad (6)$$

One can see that the expressions on the right side of the equation correspond to the equation of lines; the gradient and crossing with the abscissa give the transition matrix elements.

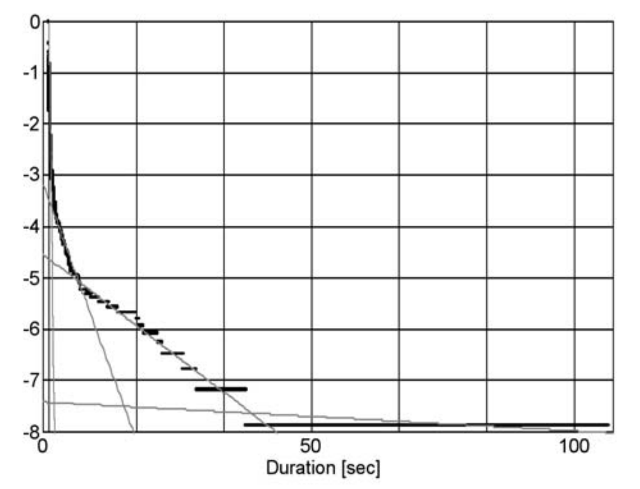


Figure 3.
Linear regression of the logarithmic complementary fade duration distribution at 5 dB

The parameterization process is depicted on the *Figure 3* in case of 5 dB threshold level.

The number of regression lines determines the state numbers in the Markov chain and their number is depending on the required lines to properly approximate the original logarithmical CCDF. In our case four numbers of lines are sufficient, which results in four fading states in the Markov chain.

Figure 4.
Measured and modeled fade duration distribution at 5 dB

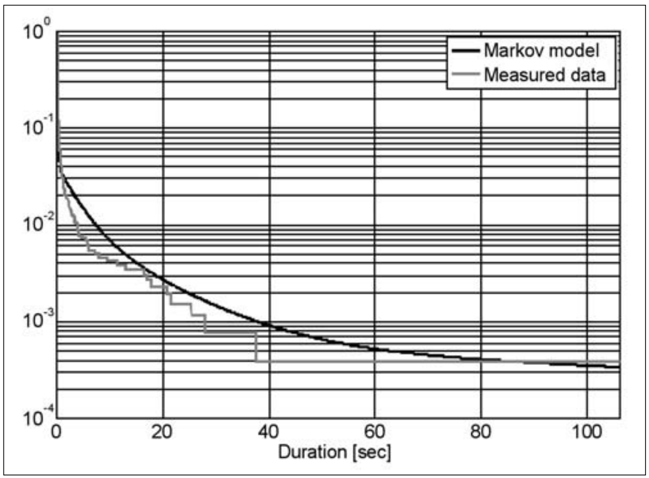


Table 2.
Parameters for Equations (7 and 8)

Transition probability	a_{ii}	b_{ii}
p_{11}	-1.849e-007	1.0000000
P_{22}	-8.646e-007	0.9995000
P_{33}	-2.949e-006	0.9990000
p_{44}	-8.963e-006	0.9795000
	a_{Ni}	b_{Ni}
p_{51}	1.037e-007	0.0003791
p_{52}	3.340e-007	0.0044070
p_{53}	3.652e-006	0.0317600
p_{54}	1.093e-005	0.5377000

After the determination of the transition matrix elements using Equation (2), the complementary distribution function of fade duration can be calculated and by depicting it together with the original measurement one can see the proper approximation (see *Figure 4*).

The method described above can be also applied to model the fade duration for different threshold levels, usually in the range of 1-30 dB.

6. The threshold dependency of the model

If we perform the modeling process for other different thresholds and depict the transition matrix elements of p_{ii} and p_{5i}, one can see that with the cubic equations given in (7 and 8) the threshold *A* dependency of the matrix can be well approximated:

$$p_{ii}(A) = a_{ii} * A^3 + b_{ii} \tag{7}$$

$$p_{Ni}(A) = a_{Ni} * A^3 + b_{Ni} \tag{8}$$

In *Table 2* we can see the parameters necessary to calculate the transition matrix threshold dependency.

With applying the above constants it is possible to calculate the elements of the transition matrix for any desired threshold level and the CCDF of the fade duration can be also computed.

In *Figures 5 and 6* one can see the threshold dependency of the model parameters and the approximations applying the Equations (7 and 8).

By this method we can calculate the transition probabilities for the 5 state Fritchman's Markov chain of the fade duration at different threshold levels. It allows computing the CCDF of fade duration which is depicted in *Figure 7* for 2-10 dB thresholds.

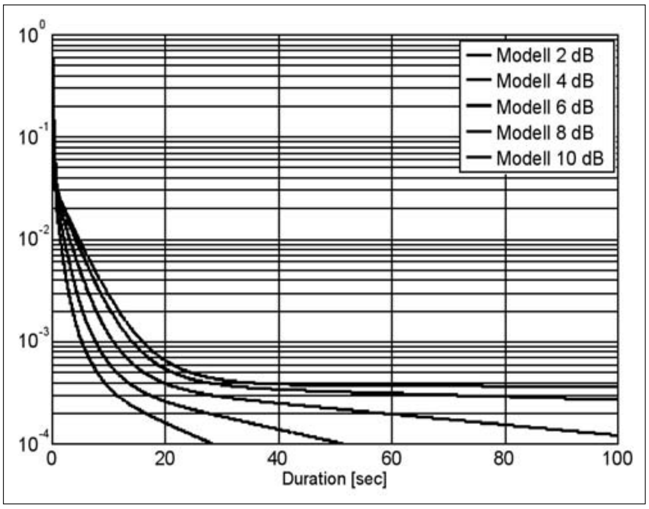


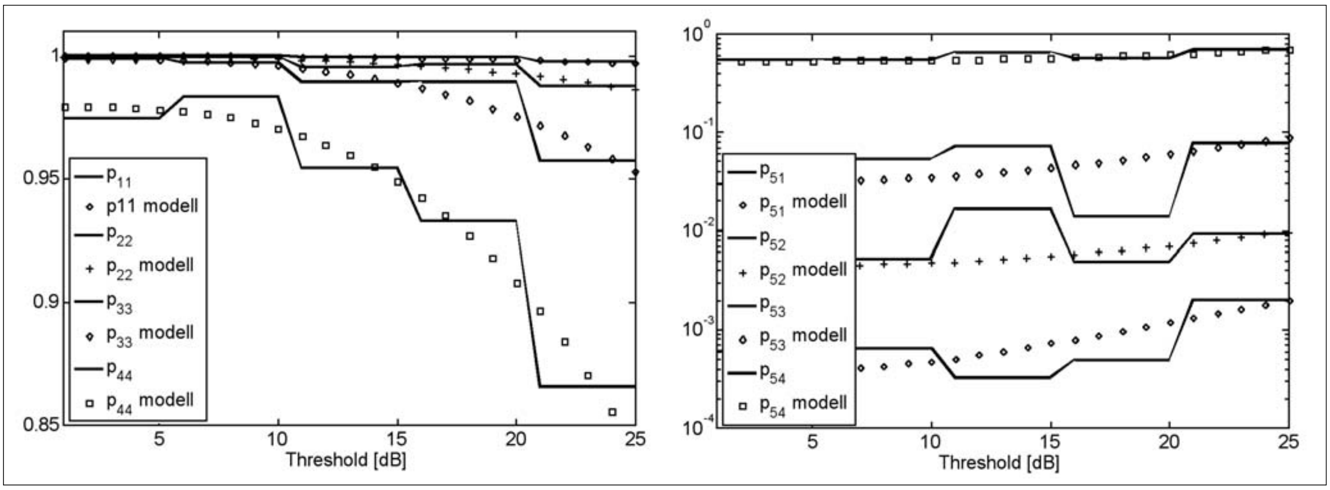
Figure 7.
Modeled fade duration complementary distributions for 2-10 dB thresholds

The cumulative complementary fade duration distribution functions can be also applied to calculate the statistics of a radio channel for a desired time period. Depicting of the total number of fade events longer than a given duration is also a common graphic method. To create this kind of statistics the modeled fade duration CCDF functions must be multiplied with the total number of fade events on the actual channel. This parameter is available from measurements and statistics and it is also applied by the earlier mentioned ITU-R model [2].

7. Summary

In our contribution we proposed a partitioned Fritchman's Markov chain to model the fade duration process caused by the multipath propagation on a Land Mobile Satellite radio channel. The parameterization of the model is the determination of the transition matrix elements of the Markov chain what can be performed from the original measurement data of the channel. This kind of Markov model is applicable to calculate the CCDF of fade duration which is an important statistical data for the radio channel designers.

Figure 5-6. Threshold dependency of pii and p5i transition probabilities



The threshold dependency of the Markov model parameters are also shown, which results that the complementary distribution functions can be calculated for any desired thresholds. This may lead us to develop attenuation time series generators and synthesize measurement data for any required duration.

Acknowledgement

This work was carried out in the framework of IST FP6 Satnex NoE project and Mobile Innovation Center, Hungary (Mobil Innovációs Központ).

References

[1] E. Lutz, D. Cygan, M. Dippold, F. Dolainsky, W. Papke: "The Land Mobile Satellite Communication Channel-Recording, Statistics and Channel Model", IEEE VT-40, pp.375–386., May 1991.

[2] ITU-R Rec. P.1623, "Prediction method of fade dynamics on Earth-space paths", ITU, 2003.

[3] B.D. Fritchman, "A binary channel characterization using partitioned Markov chains," IEEE Trans. Information Theory, Vol. 13., pp.221–227., April 1967.

[4] J.-Y. Chouinard, M. LeCours, G.Y. Delisle, "Estimation of Gilbert's and Fritchman's models parameters using the gradient method for digital mobile radio channels," IEEE Trans. Veh. Technology, Vol. 37., pp.158–166., August 1988.

[5] László Csurgai-Horváth, János Bitó: "Fade Duration Modeling of Satellite Links Applying Markov Chain", 3rd Advanced Satellite Mobile Systems Conference, Herrsching am Ammersee, Munich, Germany, pp.76–83., May 2006.

[6] Cecilo Pimentel, Ian F. Blake, "Modeling Burst Channels Using Partitioned Fritchman's Markov Models", IEEE Trans. Veh. Technology, Vol. 47., No.3, August 1998.

Modelling packet queuing of DSL access lines for the case of complete and partial rejections

ATTILA KŐRÖSI, BALÁZS SZÉKELY
Budapest University of Technology and Economics, Dept. of Stochastics, Institute of Mathematics
{akorosi, szbalazs}@math.bme.hu

CSABA LUKOVSZKI, TRANG DANG DINH
Department of Telecommunications and Media Informatics, High Speed Networks Laboratory
{lukovszki, trang}@tmit.bme.hu

Keywords: DSL, queueing models, packet scheduler

In this paper we provide an exact data-layer model and mathematical analysis of priority queuing systems representing DSL access networks on packet level with pre-emptive option. We demonstrate the accuracy and the efficiency of our numerical analysis by presenting numerical results based on simulations and numerical analysis both for complete and partial rejections. Consequently, this analysis could be applied for an in-depth packet-level performance evaluation of recent DSL systems.

1. Introduction

Nowadays the mostly used protocols in the access network are those from the family of Digital Subscriber Line (DSL) [1,2]. A wide range of DSL technologies is available providing different sets of maximum available capacities and physical reach. From the aspect of available resources it is well-known that the edge of the network is less developed than its middle level. Therefore, the access capacity often appears to be the bottleneck of the network connection. The usual method to confront successfully this bottleneck, as also proposed by 3GPP and ITU-T, is to classify packet flows into four classes that cover applications with the same order of magnitude of Quality of Service (QoS) requirements. Packets of each class are stored in separate buffers and usually served by strict priority scheduler [3].

This paper is motivated by the performance evaluation study of DSL based access network supporting QoS. The related data-layer model leads to the analysis of priority queuing system with finite buffers and bursty arrivals, where at the inlet of a common DSL line a strict priority scheduling is applied on the fragmented upper layer data units, while depending on the actual implementation complete or partial rejection could be applied.

The study of priority queuing systems today is also an actual topic in the field of queuing research. However, the exact description of such a system is not yet available. Instead, several approximate solutions can be found in the related literature, which are not sufficient enough in practical performance analysis. The modelling approach presented in this paper overcomes the requirements of performance evaluation of both types of rejection rules.

2. DSL access architecture

Architecturally, the DSL customers are connected to the access network by using DSL modems that are aggregated into DSL Access Multiplexer (DSLAM) through DSL access lines as depicted in *Figure 1*.

The latest standards of DSL [1,2] offer two options of packet transport. In the ATM-based DSL technologies ATM Adaptation Layer 5 (AAL5) is used to encapsulate higher layer packets. A packet entering the DSL modem or DSLAM output port is first converted to an AAL5 Protocol Data Unit (PDU) then the whole PDU is segmented into ATM cells. In the other case when Ethernet-based DSL technology is considered, Packet Transfer Mode – Transmission Convergence (PTM-TC)

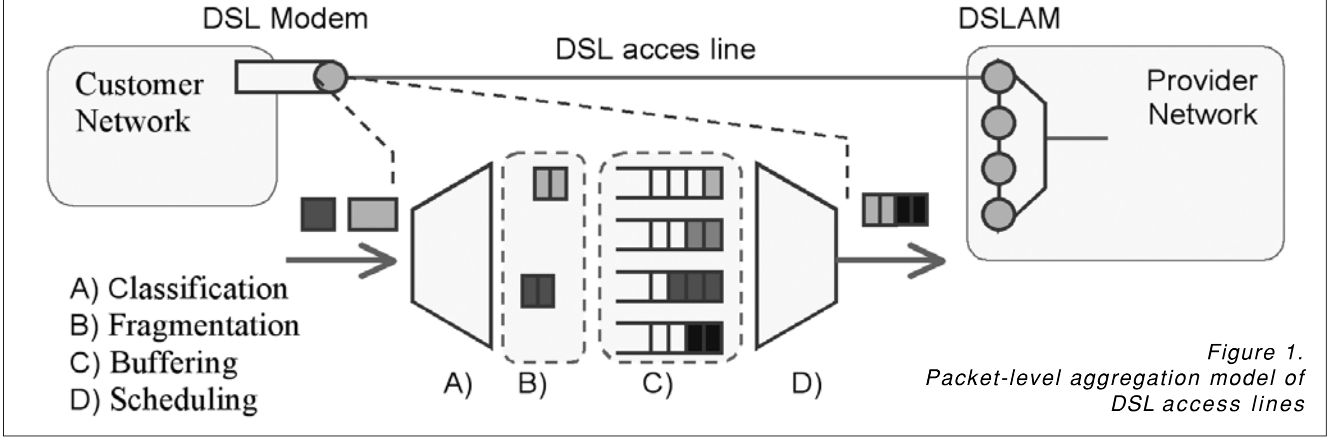


Figure 1. Packet-level aggregation model of DSL access lines

is introduced that supports transmission of higher layer packets by applying the 64/65 byte encapsulation method of High-Level Data Link Control (HDLC) framing.

At the data plane, when a packet arrives at an empty queue, even if it has higher priority than the packet has which is currently under service, it has to wait for the service completion. This mechanism is particularly problematic for low-rate transmissions. The service time of a full length Ethernet frame at the lower priority class introduces a considerable delay in the first priority queue, especially at the line rates of today's access networks. In order to reduce this kind of delay, pre-emption mechanisms such as ATM AAL5 encapsulation or PTM-TC with pre-emption option enabled, since fragmenting packets into small pieces would lower the additional delay introduced by serving these large packets from different classes.

The above described data-layer model including segmentation of user traffic and pre-emption option leads us to the model of priority queuing system with batch arrivals. Besides, when in the real system congestion occurs, at the buffer of user traffic two options are implemented. When complete rejection rule is implemented, the whole higher-layer data unit is dropped in the case of congestion, while partial rejection first fills the free slots in the buffer, and only the remaining segments are dropped. During the proposed analysis both options are considered.

3. Related work

A number of papers have been published regarding the analysis of priority queuing systems since the first initial results of Takács [12]. Although the study of priority queuing systems is also an actual topic in the field of queuing research nowadays, the exact description of such a system is not yet available. First, we summarize the works in which infinite buffers are assumed. The problem is less complicated and some nice and explicit formulae can be provided in this case. Besides, the results for systems with infinite buffers are good approximations of finite, large buffer systems in some certain conditions. These papers, e.g. [6], often apply generating functions, Laplace transform, or matrix geometric methods to determine the distribution of waiting time.

Assuming the arrival process is Poissonian, Takács [7] gave necessary and sufficient conditions for the existence of a stationary limit distribution of the waiting time. He also provided the Laplace transform and the first three moments of the limit distribution. In [8] two priority classes are considered. The arrival processes are assumed to form four mutually independent renewal processes determined by general distributions. Limit theorems are obtained for the low priority waiting time and for the total uncompleted service time of unfinished work in the system at time t.

Non-preemptive priority queues with MAP (Markovian Arrival Process) arrivals were considered in Takine's paper [9]. The service times of each priority class are i.i.d. random variables with a general distribution function. Using both the generating function technique and the matrix analytical method, they derived various formulas for the marginal queue length distribution of each priority class. Furthermore, they provided the delay cycle analysis of the waiting time distribution of each class and characterized its Laplace-Stieltjes transform.

Xue and Alfa [10] assumed BMAP (Batch MAP) arrivals of the high priority class and MAP arrivals of the low priority class in the case of two queues. A sufficient condition under which this tail probability has asymptotically geometric property was derived. If the asymptotically geometric property holds, a method was designed to compute the asymptotic decay rate. Alfa, Liu and He [11] used the matrix geometric method to study the MAP/PH/1 general pre-emptive priority queue with multiple classes of jobs. They determined the stationary behaviour of the system. Next, the distribution of the number of waiting packets and their waiting time are easily calculated.

Reducing the amount of the necessary computation is the goal of the work of Van der Heijden et al. [12]. The idea of their approximation method for N classes of customers was the following: for each class, aggregate the remaining customers into one class and evaluate the performance of the system with these two classes. This method leads to the analysis of N two-class systems instead of the analysis of one N-class system. The service time of the aggregated class is approximated by a hyperexponential distribution.

Finally, we summarize some further works in which priority queuing system with finite buffers were analyzed. In the case of finite buffers the packets may be lost if the buffer is overloaded. Sharma and Virtamo [13] investigated a priority system with two buffers, Poisson arrivals, and general service time. An algorithm was given to calculate the distribution of the waiting time and the rejection probability. Gómez-Corral et al. [14] used a continuous-time Markov chain to describe the state of the system at arbitrary times, constituting a finite QBD process. Computationally convenient formulas were derived for various performance measures: the blocking probability, the stationary distribution of state at pre-arrival epochs, post-departure epochs, and loss epochs.

4. The queuing model

Let us consider a priority queue with a single server with constant service rate V [bps]. When the server turns to the high priority queue, all high priority packets are served before any of those from lower priority classes. The server applies non-preemptive service principle (NPRP), which means that a low-priority packet is not interrupted if a high-priority one comes along while it is under service.

The system has I priority classes and assumes that the class of lower priority index value has higher priority. Each priority class has its own queue of finite length

$b^{(i)}$, $i=1,2,\dots,I$. Packets in each class are served according to first-come first-served (FCFS) order. Batch of packets of different priority classes arrives to the system according to the Poisson process. Denote by $\lambda^{(i)}$, $i=1,2,\dots,I$ the incoming traffic intensity of a given class *i*. The number of packets in each arrived batch follows a discrete random variable *X*. In general *X* can be different for each class. In addition, packets of batches of all traffic classes have the same constant size of *L* [bits].

Since the buffers are finite, in case of overload two cases of rejection rules are analyzed. The first case, when the arriving batch of packets could not fully get into the queue of that certain class the whole batch will be lost, is called complete rejection. Alternatively, the partial rejection could be used, which means if there is no room in the right queue for an arriving batch, the batch will fill the buffer with packets, and the rest will be lost. In the calculation of rejection in this case, the whole batch of packet is considered as lost. An illustration of the considered priority queuing model is shown in *Figure 2*.

5. Analysis of the queuing system

In this section we outline the mathematical analysis of the queuing system presented in Section 4. The proposed model precisely describes the system behaviour without any approximation. Let us see the following systematic steps.

First of all the analysis of the presented queuing system with constant service rate (*V*) is converted into the simpler problem of I $^{M_x/G/1/b}$ queues. Let us see the system from the *i*th priority queue point of view. The batch of *L*-sized packets arrival follows a Poisson process with $\lambda^{(i)}$ parameter, while the queue size is $b^{(i)}$. The service time for a packet in this queue, however, differs from the time needed by the server to serve the packet itself (*L/V*).

Instead, with the selfish respect to the queue *i*, the service time begins when the server starts to serve a packet of class *i* and then finishes when it is ready to serve the next packet of the same queue. It includes the operation time to serve the possible higher priority packets which arrived in the mean time. This service time is denoted by $S^{(i)}$. The basis of our analysis is a recursive calculation of $S^{(i)}$ based on the distribution of the service time of the previous queues and their busy periods $T^{(i)}$.

$$S^{(i+1)} = S^{(i)} + T^{(i)}[N^{(i)}(S^{(i)})], \qquad i = 1,2,\dots,I.$$

Note that there is another similar recursive formula for the distribution of $T^{(i)}$.

We still need to calculate the distribution of the special service time S^* of the first packet of the batch that arrives to the empty queue. This random variable depends on the state of the other queues. Handling this issue a much more sophisticated recursion is provided.

So we decompose the system into different queues but they are not independent so we encode the dependence of the queues into the service time and special service time. Next, we determine the long-run average distribution of a single $^{M_x/G/1/b}$ queue with special first service time. The probability that there are *j* packets in the queue, p_j, is defined as the limit of the fraction of time the system spends in state with *j* packets in the buffer over the operation. To determine this probability we use the theory of regenerative processes. More precisely, we divide the whole service time into the expected values of the time that the queue spends with *j* packets in the queue.

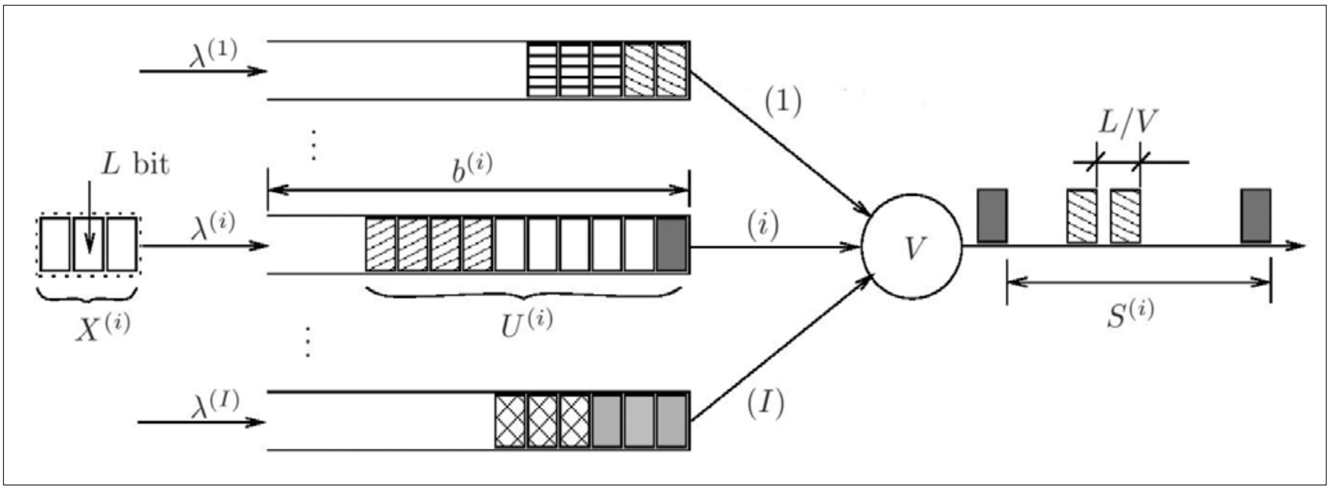
$$p_j = \frac{d_0 t_{0,j} + d_1 t_{1,j} + \cdots d_b t_{b,j} + d_{x_1} t_{x_1,j} + \cdots + d_{x_l} t_{x_l,j}}{d_0 \dfrac{1}{\lambda} + (d_1 + \cdots d_b)\mathbf{E}S + (d_{x_1} + \cdots + d_{x_l})\mathbf{E}S^*}$$

Using the above results the waiting time distribution is given by the following formula:

$$W = R + \sum_{k=1}^{U} S_k + \sum_{k=U+1}^{U+X-1} S_k + \frac{L}{V}.$$

The batch with *X* packets arrives into the queue during a service time, such that, there are *U* packets in the queue. The waiting time includes the remaining time *R*,

Figure 2. The queuing model



that is the time while the service of the first packet in the queue starts, the first sum which is the time is needed for the batch to wait for the service of the packets in the queue, and the second sum regarding as the service times of the packets in the batch itself, except the last one that needs only time of *L*/*V*. Since these times are independent simple convolution can be applied. The calculation of the distribution of *R* is very similar to the calculation of p_j above.

The rejection probability can be easily calculated using the previous paragraphs. Let *X* be the number of arriving packets in a batch after a sufficiently long time then the probability of the rejection can be formulated into the following form:

$$P_{rej} = \sum P(U = j)\, P(X > b - j)$$

Regarding the rejection rules, we have to emphasize two important differences in the analysis. One of them appears in the matrix of the probabilities which tells us how the number of packets changes in the queue when a batch arrives. This matrix is used for the calculation of $S^{(i)}$ and the stationary distribution of the aforementioned Markov chain. The other difference is in $t_{i,j}$ and *R*. Both of them concern an exponential random variable that describes how long the system has to wait until the change of the queue length if there are *j* packets in the queue.

The intensity of this time is different for different rejection rules. Namely, if partial rejection is considered then the queue length always changes if a batch arrives while with complete rejection the queue length changes only if the batch fully fits in the buffer.

6. Numerical results

The presented numerical algorithms have been implemented in *C* and for the justification of our analysis and to investigate the different behaviour of complete and partial rejections we provided long run packet-level simulations as well. The system parameters are chosen so that they meet closely the real DSL based access conditions. The load is set to achieve 50%, 70% and 90% utilizations, and the ratios between traffic classes are 4%, 12%, 24%, and 60%, respectively. The packet arrival time is chosen to fit to voice traffic in the first class and to internet-like traffic based on the simple IMIX model of the low-priority classes. Note that the infinite sums and continuous distributions in the numerical calculations are approximated to have the errors less than 10^{-6}.

The tail probability distribution of the service time of class-4 is shown in *Figure 3*. Remind that the service time of a priority packet is the time difference between the service of two consecutive packets in a given priority buffer. The curves show the results of our numerical analysis and simulations of 90% utilization. Simulation results are done for 10^8, 10^7 and 10^6 packet arrivals. It can be observed that the results are almost the same. The only difference between the two curves is that the numerical method can also provide those probabilities where the simulation is less feasible. The other observation is that the service time seems to follow geometrical distribution since the tail distribution is almost a straight line in the log-log scale.

In *Figure 4*, the differences between complete and partial rejections could be seen in terms of the queue length Probability Density Function (PDF) of class-2 under link utilization of 50%. There is a significant difference in the results near the capacity limit of the queue, which cause the difference at the rejection probability.

Our last results in *Figure 5 and 6* show the variance in Cumulative Density Function (CDF) of whole batch of packets waiting times. The difference is not significant compared with the queue length distribution. However, if we rescale the graph to finer the grid we could realize that the probabilities of partial rejection are always below the one of complete rejection.

Figure 3.
The numerical and simulation results of the service time

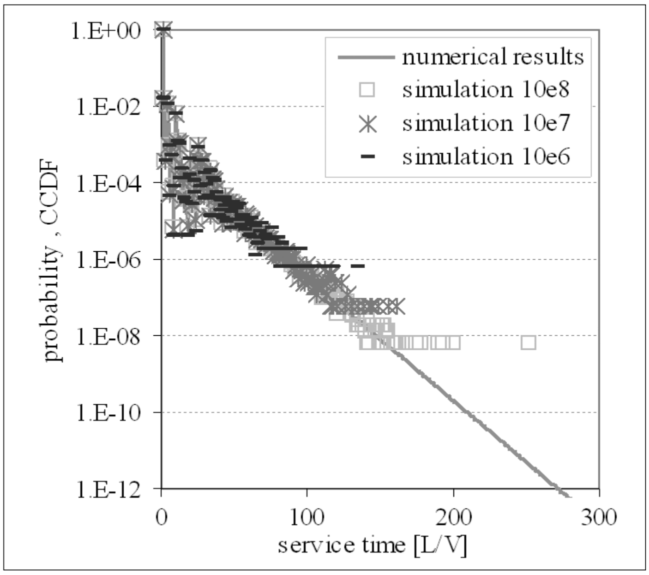
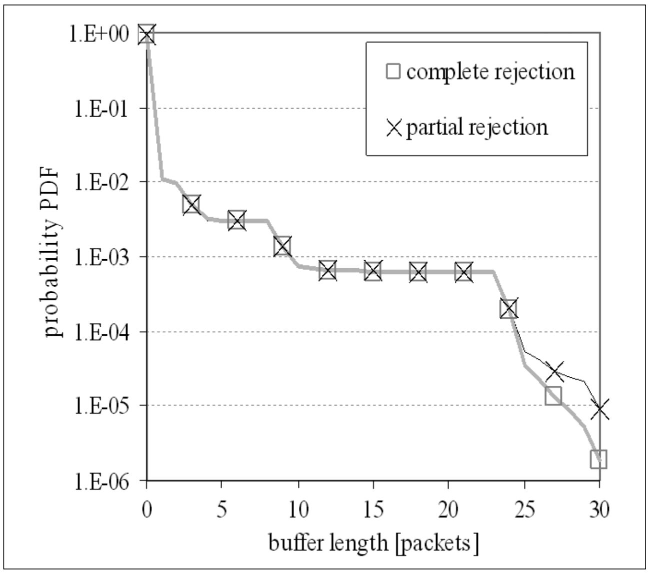


Figure 4.
Queue length distribution for complete and partial rejection rules



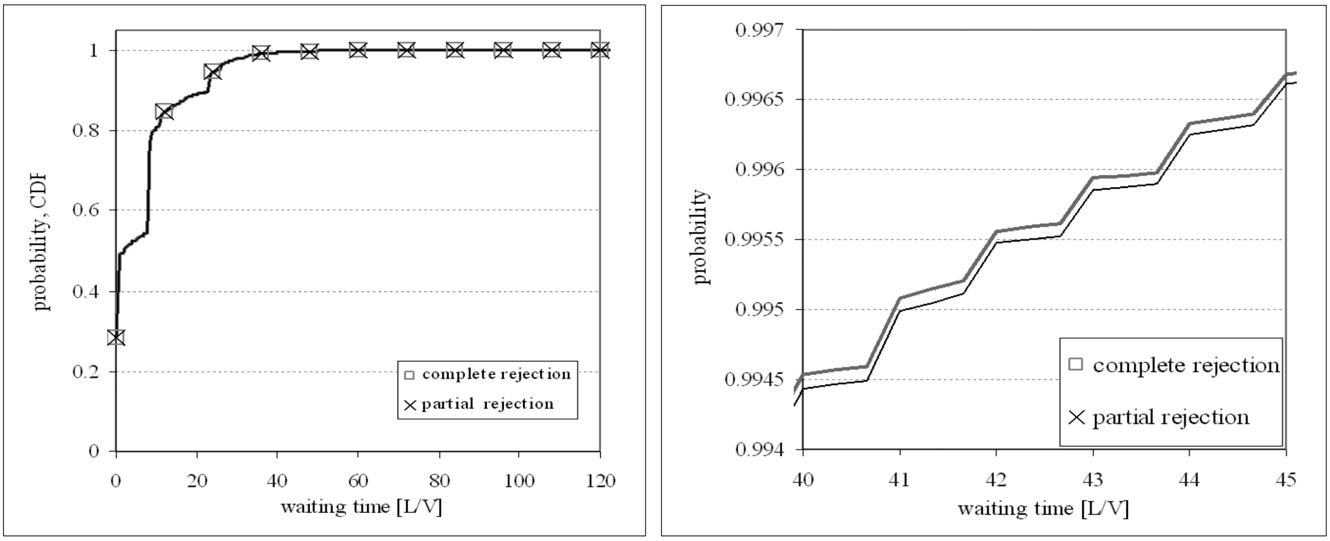


Figure 5-6. Delay distribution in complete and partial rejection rule

7. Conclusions

In this paper an exact analysis of finite buffer priority queuing system with Poisson batch arrivals is provided. We have investigated both the complete and partial rejection rules. The main step of the analysis was the encoding of the dependence structure of the whole system into the service and the special service time in each queue.

The derived results show the practical difference between the partial and the complete rejection: the rejection probabilities are significantly different while the delays almost equal. Besides, the feasibility of the numerical analysis model has been proven, comparing it with long-term simulation results.

References

[1] “Asymmetric digital subscriber line (ADSL) transceivers – extended bandwidth adsl2 (ADSL2+)”, ITU-T, Recommendation G.992.5., 2005.

[2] “Very high speed digital subscriber line transceivers 2 (VDSL2)”, ITU-T, Recommendation G.993.2., 2006.

[3] T. Orphanoudakis, S. Perissakis, K. Pramataris, N. Nikolaou, N. Zervos, M. Steck, C. Baumhof, D. Verkest, C. Ykman-Couvreur, G. Doumenis, F. Karoubalis, I. Theologitou, D. Reisis, G. Konstantoulakis, N. Vogiatzis, “Hardware Architectures for the Efficient Implementation of Multi-Service Broadband Access and Multimedia Home Networks”, Telecommunication Systems, Springer Netherlands, 23(3-4), pp.351–367.

[4] Kleinrock, L., “Queueing Systems”, Vol. II., Computer Applications, Wiley, NY, 1976.

[5] Tijms, H. C., “A First Course In Stochastic Models”, John Wiley & Sons, 2003.

[6] Daigle, J. N., “Queueing Theory With Applications To Packet Telecommunication”, Springer, 2005.

[7] Takács, L., “Priority queues”, Operations Res. 12., pp.63–74., 1964.

[8] Hooke, J. A., “Some heavy-traffic limit theorems for a priority queue with general arrivals”, Operations Res. 20., pp.381–388., 1972.

[9] Takine, T., Hasegawa, T., “The non-preemptive priority MAP/G/1 queue”, Operations Research, 47(6), pp.917–927., 1999.

[10] Xue, J., Alfa, A. S., “Tail probability of low-priority queue length in a discrete-time priority BMAP/PH/1 queue”, Stoch. Models, 21(2-3), pp.799–820., 2005.

[11] Alfa, A., Liu, B., HE, Q., “Discrete-time analysis of MAP/PH/1 multiclass general preemptive priority queue”, Naval Research Logistics 50, pp.662–682., 2003.

[12] van der Heijden, M.; van Harten, A., Sleptchenko, A., “Approximations for Markovian multi-class queues with preemptive priorities”, Oper. Res. Letters, 32(3), pp.273–282., 2004.

[13] Sharma, V., Virtamo, J. T., “A finite buffer queue with priorities”, Performance Evaluation 47, pp.1–22., 2002.

[14] A. Gomez-Corral, A. Krishnamoorthy, V.C. Narayanan, “The impact of self-generation of priorities on multi-server queues with finite capacity.” Stoch. Models , 21(2-3), pp.427–447., 2005.

Numerical analysis of mobility management algorithms

BENEDEK KOVÁCS, PÉTER FÜLÖP

Budapest University of Technology and Economics, Department of Telecommunications
{bence, fepti}@mcl.hu

Keywords: mobility management, modeling, handover, Mobility Access Points, Mobile Ad-hoc Network

This paper investigates mobility management strategies from the point of view of their need of signaling and processing resources on the backbone network and load on the air interface. A method is proposed to model the serving network and mobile node mobility in order to be able to model the different types of mobility management algorithms. To obtain a good description of the network we calculate parameters from given topologies that we think are the most important ones. Mobility approaches derived from existing protocols and other possible mobility scenarios are analyzed and their performance is numerically compared in various network and mobility scenarios. The aim is to give general design guidelines for the next generation mobility managements on given network and mobility properties.

1. Introduction

Information mobility has become one of the most common services in the modern world with the penetration of the portable phones and other mobile equipments. The wireless multimedia and other services have many requirements and the resources in the serving network are often expensive and limited.

In the first mobility protocol designs the main scope was to create a well-functioning mobility. For example the Global System for Mobile Communication (GSM) network uses a cellular approach to save bandwidth on the air interface but does not really focus on the problem of signaling load on the wired serving network. In the Mobile IP (MIP) structure the IP mobility is in the main scope. There are many enhancements of MIP to optimize the original protocol and introduces for example hierarchy, location tracking to obtain a cheaper solution. However, Host Identity Protocol (HIP) is drastically different from MIP: their mobility approaches are similar but implemented on different network layer levels. Wireless Local Area Networks (WLAN) are constructed like the original Local Area Networks (LAN) and provide mobility only within the radio interface and use Dynamic Host Configuration Protocol (DHCP). Future protocols might use different media and technological background to provide mobility. For this reason it is appropriate to treat mobility as an abstract problem regardless of actual technical solutions.

The advantage of our work is that we do not focus on a selected technology – not even on a given network generation – but discuss mobility in general within the modern computer and telecommunication networking technologies. We compare selected mobility approaches and show how the network properties affect the usability of each. The aim is to find the suitable one for different scenarios or at least to give guidelines how to construct the network for a protocol or adjust the protocol to the network.

2. Abstract mobility management

In this paper, the mobility management is discussed generally regardless of the very technology used. One will see that the approaches discussed here could be applied for various types of mobility management protocols on different technology levels. We try to grab the most significant properties of the mobility that is worth to discuss within the scope of the modern mobility protocols.

We define *Mobility Management System* as an application running on network nodes that helps to locate the mobile equipment towards its unique identifier.

- *Mobile Nodes* (MN) are the mobile equipments who want to communicate to any other mobile or fixed partner.
- There are *Mobility Access Points* (MAP) as the only entities that are capable to communicate with the Mobile Equipments.
  (Note: mobility does not necessary imply radio communication. It means only that the Mobile Node changes its Mobility Access Points and when it is attached to one, communication between them can be established.)
- *Mobility Agents* (MA) are network entities running the mobility management application.
- There is a core network that provides communication between the Mobility Access Points and has a structure that can be described with a graph.
  Vertices are either Mobility Access Points or Mobility Agents or other serving nodes who are not part of the mobility management application and the edges can be any kind of links (even radio links) for the data communication between the vertices.

With this definition, one can see that most of the functionalities of the current mobility protocols and others under development can be generally described.

However, this model is too general and we should restrict the discussion with some practical assumptions:

- A Mobility Access Point is always a Mobility Agent. (In our discussion, there can always be a sub-network of multiple physical access points under a single Mobility Access Point. We do not discuss the lowest (micro) level of mobility.)
- Mobile equipment can communicate with multiple Access Points at a time but one connection is necessary and enough to maintain the correct communication. The mobile can also attach and detach from any Mobility Access Points.
  At this point we assume that the mobile node is administrated only at one agent.
  This means that the problem of finding the mobile node is the same as to find the correct agent.
- The nodes in the core network communicate and find each other with a given protocol or method (for example via IP routing). For this reason this part of the mobility protocols is not discussed.

Now mobility management is simplified to a protocol that finds the correct, marked Mobility Access Point where the MN is attached. This suits to our aim to investigate the properties of various management strategy approaches since the number of messages sent and the number of tasks completed can be calculated. With cost parameters one will be able to adopt the model to exact solutions and can analyze them.

We derived the strategies into the *Centralized-, Hierarchical-, Tracking- and Cellular-like* approaches. There can be some other special approaches but mostly they can be classified into these categories. It is also common that the mixture of applications is used on different mobility layers. By our investigations we believe that design guidelines for new generation network mobility protocols can be given.

3. Network graph and node mobility parameters

In this section, we introduce how we will handle the networks on which the mobility management algorithms work. To derive the main parameters we will have to model the behavior of the mobile nodes first. There will be general and algorithm-specific parameters introduced.

Secondly, the three cost dimensions we want to handle in this paper will be introduced. These are the "signaling on the links" (C^{signal}) as a bandwidth and interworking equipment usage, the "processing in the nodes" ($C^{process}$) which are taken into account on the nodes running the mobility protocol, the "air interface usage" (C^{air}) containing explicitly battery consuming as well.

3.1. Modeling the network

In many works the network is modeled in order to emphasize the properties of a single protocol compared to another one. This approach has the disadvantage of inflexibility since new protocols can not be included in the comparison and also it is difficult to follow little modifications in the protocols.

An approach of the network modeling uses the given network structure that is essential to make an appropriate examination in those cases but limits the scope of discussion. For example, when a GSM cell structure is used, no vertical handovers are taken into account: another mobility protocol might have a different structure to cover the same geographical region. One can see that in these cases, the graph, describing the network might not be drawn on a plane.

For this kind of reasons, many works describe a network using single parameters, for example by a general average distance between nodes. With this approach, any kind of network could be described. However, introducing these parameters is not enough to compare most of the protocols.

Summing up the requirements, we introduce a method to model the given networks to get the benefits of the first approach and we provide a method how the protocol-specific parameters and also additional ones can be derived in order to generalize the discussion just like in the case of the second approach.

3.2. Deriving parameters of a given network

Let us have a given network topology with a given MN behavior. The network is modeled with a graph just like the possible movements of the mobile. The behavior of the mobile node that is the frequency of some kind of a handover between two mobility access points will be modeled with Poisson processes like in [5].

Let us assume that the behavior of the MN can be modeled with a Markov chain, given with a rate matrix. In this matrix, all the possible (in practice: the practically possible) MAs are listed where the Mobility application runs. (These MAs can also denote single access points, bigger networks or the Home Agent if desired.)

The number of MAs is n and so the matrix will be an $n \times n$ matrix where each element denotes how frequent the movement of the mobile is from MAPi→MAPj. (If an MA is not a MAP then there are 0 values in its row and column.) From the rate matrix the transition matrix can be determined easily. We assume that the matrix, without the non-MAP nodes, is practically irreducible and aperiodic that implies that the chain is stable and there exists a stationary distribution. This will be denoted by a density vector. In this vector, the ith element denotes the probability of the MN being located under the ith MAP. (For MA nodes that does not support access point functionality, there exists an element in the vector with 0 value.)

Let us have the corresponding network graph given with its adjacency matrix A. This matrix should include all the nodes in the network where the mobility application runs (all the MAs again) so has the same $n \times n$ size as matrix. With the Floyd algorithm the optimal distances between the nodes can be calculated (even with weighted or directed edges as well). The distance between nodes will be the sum of weights on the shortest path from one to the other. Let this result matrix be given. In the *i*th row of the matrix, the distances from FAi are listed. Let the distances from the HA, – a special FA – be given with the vector *a*.

We will have parameter *w* to denote the average of the weights in the network. It can be calculated by summing up the elements of the matrix and dividing it by n^2.

3.2.1. Determining m

Parameter *m* will denote the average depth level, that is the average number of edges on the shortest path from the MN to the HA. Clearly, the average number of vertices among the path is *m+1*. We will use matrix A_d and vector $\underline{a}$ to calculate this parameter. Both have to be normalized with the average weight of edges in the network (*w*). Now *mw* can be calculated by determining the weighted average of the distances where the weights are the probabilities that the node is under a given MAP:

$$m = \frac{\underline{a} * \underline{b}}{w} \qquad (1)$$

where * stand for the scalar product. One can see that the nodes which are not MAPs have a 0 multiplier and do not count in the average distance as expected.

3.2.2. Parameter g_T

We will have another parameter like *m* that is the average distance between two nodes who handle the MNs handovers. They might be connected, but they can also be quite far from each other logically due to different technologies especially in the case of vertical handovers. So as we see this parameter has to denote the weighted average value of the length between every two neighboring MAs where the mobile can attach. Then it is calculated as follows:

$$g_T = \frac{\underline{b} * tr(A_d \cdot B_\Pi)}{w} \qquad (2)$$

Our notation indicates that this parameter will have the most effect on the Tracking-like management solutions as we will see.

3.2.3. Parameter g_H

This parameter denotes how far is the nearest hierarchical junction to register at in the average, if we consider the optimal covering tree of the network with the HA in the root. The junction node is the nearest common node of the paths from HA to the old and the new FA of the MN. (In most cases, it is not possible to achieve the optimal tree structure since the different service providers will not mesh their networks: approximate values can be used instead.) About determining of parameter g_H can be read in [6].

3.2.4. Parameter g_C

This parameter will denote the average distance of MAPs from the main MA of a Location Area in the Cellular-like approaches. It is an NP full problem to calculate the optimal cell structure, but there are algorithms approximating it very well in some sense. We have run the algorithms developed and published in [12].

3.3. Modeling the mobile node

As we have seen, matrix B_Q describes the movement behavior of the MN, handover-wise. Summing up the *i*th row in this matrix we get a rate of how frequently the MN moves from the *i*th MA (MAP) with a Poisson-process. Let λ denote the average parameter of the Poisson-process (at each MAP) and so denote the rate of handovers for a general MN anywhere in the network.

The other parameter that can be introduced in a similar manner is the rate of receiving a call: μ. This parameter can also be time- or location-dependent. We take its average value like we did it in the case of λ and we assume it is constant in the examined very small time interval just like we did in the case of matrix B_Q and through the whole modeling.

Using the achievements in [6], let us introduce ρ as the “mobility ratio” meaning the probability that the MN changes its FA before a call arrives:

$$\rho = \frac{\lambda}{\lambda + \mu} \qquad (3)$$

3.4. Definitions of cost constants

The three main classes of cost types will be introduced here. One will see in Section 4 that modifying the ratio of some parameters (for example the registration and packet forwarding costs) will have strong effect on the results.

If one tries to design a mobility management algorithm and also wants to implement and use it he has to decide the network level he wants to use. Also the equipments might be different. It is possible to modify the parameters we will introduce and then to have a relevant calculation on the expected costs.

3.4.1. Link related constants

cu: The unit cost of one update on a link.
cd: The unit cost of one delivery on a link.

3.4.2. Node related constants

cr: Registration cost, as the cost of the process in the MAP that has to run in the case when a MN node wants to attach. This can include the generation cost of a temporary ID, database handling, agent discovery etc.

cf: Forwarding cost at a MA. If a signaling message reaches a MA it has to decide if there is some process has to be executed with the package and where to forward it. (This can be really low for a number of protocols but also high as well.)

cm: This is the constant cost of modifying some node related records in a MA.

cec: The cost of building up a message. For example to encapsulate a message when a corresponding node wants to communicate with the MN in the MIP structure.

crc: The cost of recapsulating or rebuilding a message.

cdc: The cost of decapsulate or open the message at an endpoint.

3.4.3. Mobil equipment connection related constants

cau: The cost of uplink messaging between the MN and the MAP.

cad: The cost of downlink messaging between the MN and the MAP.

4. Modeling the existing approaches

In this section, the selected five main types of mobility management protocols are described and modeled with their signaling-, processing-, and air interface cost functions [1]. One will see that these main protocols could be applied to most of the existing mobility approaches.

4.1. Centralized approaches

In this management structure the mobile always sends location update messages in case of handover to a centralized management node, which maintains a database to contain the location of Mobile Nodes. Because of this the central agent is always able to forward the packets to the MN (Mobile IP [10]), or to send back the reachability of the MN (SIP).

$$C_{CENT}^{signal} = \rho m c_u + (1-\rho) m c_d$$

$$C_{CENT}^{process} = \rho(c_r + (m-1)c_f + c_m) + (1-\rho)(c_{ec} + (m-2)c_f + c_{dc}) \quad (4)$$

$$C_{CENT}^{air} = \rho c_{au} + (1-\rho) c_{ad}$$

One can see that the cost functions are obvious and simple. The second main advantage of this protocol is its simplicity: these approaches can be installed by setting up a Central Agent in the network and by running an IP-level software module on the MN. There is no need to change any other entity in the network, therefore it is cheap and easy to install.

On the other hand, centralized mobility puts extraordinary high overload on the bearer network and uses non-optimal routing, which is unacceptable. However this solution is far from the optimal, still, most of the mobility implementations use the same kind of this centralized approach.

4.2. Hierarchical solutions

Instead of the global management node regional management system can be used to reduce the signaling traffic by maintaining the location information locally. For this reason we can use the MAPs and MAs as local agents that have database to store the actual IP addresses of MN. So we can consider this hierarchical network structure as a tree of MAP, MA and other network node with Central Agent in the root of the tree.

Because the location information is sent only to the nearest MA, the costs function changes compared to the centralized solution. The advantage of this method is the more optimal functionality, and smaller load on the bearer network. However the change of some other entity is needed in the network, therefore the solution is more expensive.

An example for such solution is the Hierarchical Mobile IP (HMIP) [4]:

$$C_{HIERARCH}^{signal} = \rho g_H c_u + (1-\rho) m c_d$$

$$C_{HIERARCH}^{process} = \rho(c_r + (g_H - 1)c_f + c_m) + \quad (5)$$

$$+ (1-\rho)(c_{ec} + (m - g_H - 1)c_f + c_{rc} + (g_H - 1)c_f$$

$$C_{HIERARCH}^{air} = \rho c_{au} + (1-\rho) c_{ad}$$

4.3. Cellular-like solutions

For mobility problem there are cellular-like solutions as well, whose idea comes from the GSM protocol.

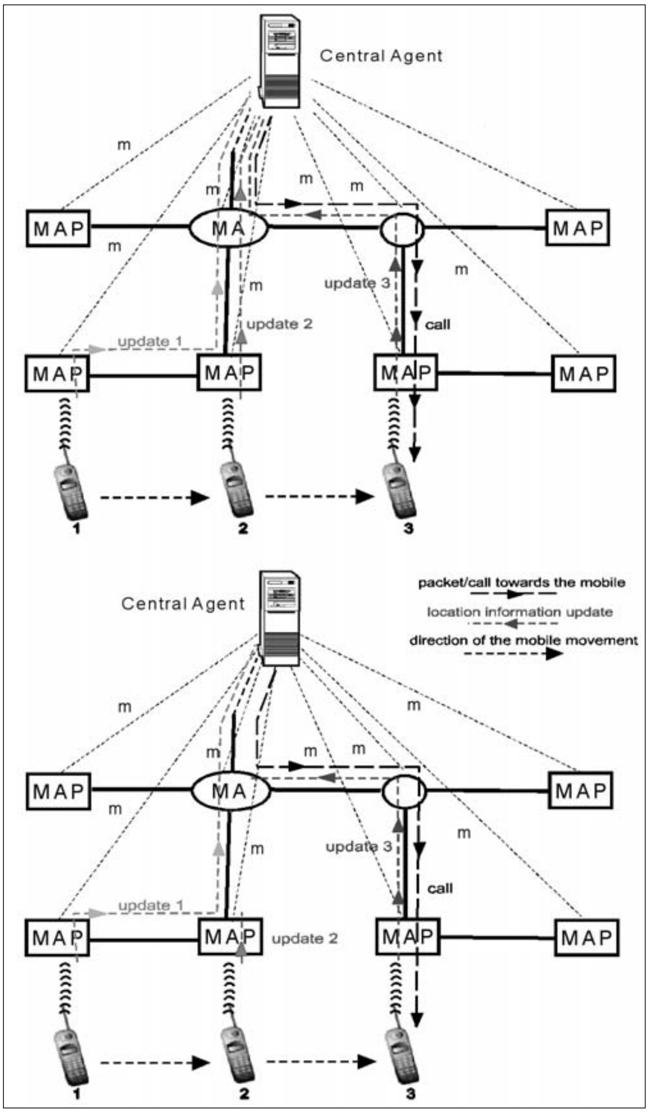
The advantages of these approaches are the quick handover mechanism in lower layer and cheap passive connectivity as it can be seen through the cost functions in *Figure 1* as well. The disadvantage is that the building of the network has to be done carefully and too many paging messages will cause an extreme increase in the costs. In cellular like solutions two constants related to the network topology are very important:

nc: The average number of MAPs in a page.

nd: The number of pages in the whole network.

Three main subtypes could be distinguished, which are introduced in the next sections.

Figure 1. Centralized and hierarchical strategies



4.3.1. Standard cellular

For mobility problem there are other cellular-like solutions. One well known example is Cellular IP (CIP) [3].

The solution builds strongly on the fact, that from the large number of mobile nodes only a small percentage is receiving data packets. For this reason we can introduce well-defined optimized areas, called paging areas, and it is enough to know in which paging area the idle mobile are moving. In this case the hop-by-hop manner routing leads the packet only to the domain border of the paging area.

From this point of the network to the mobile, the nodes in the paging area do not store any information about the idle mobiles, accordingly in case of a packet addressed to an idle mobile the paging area is flooded with the packet by broadcast message (6):

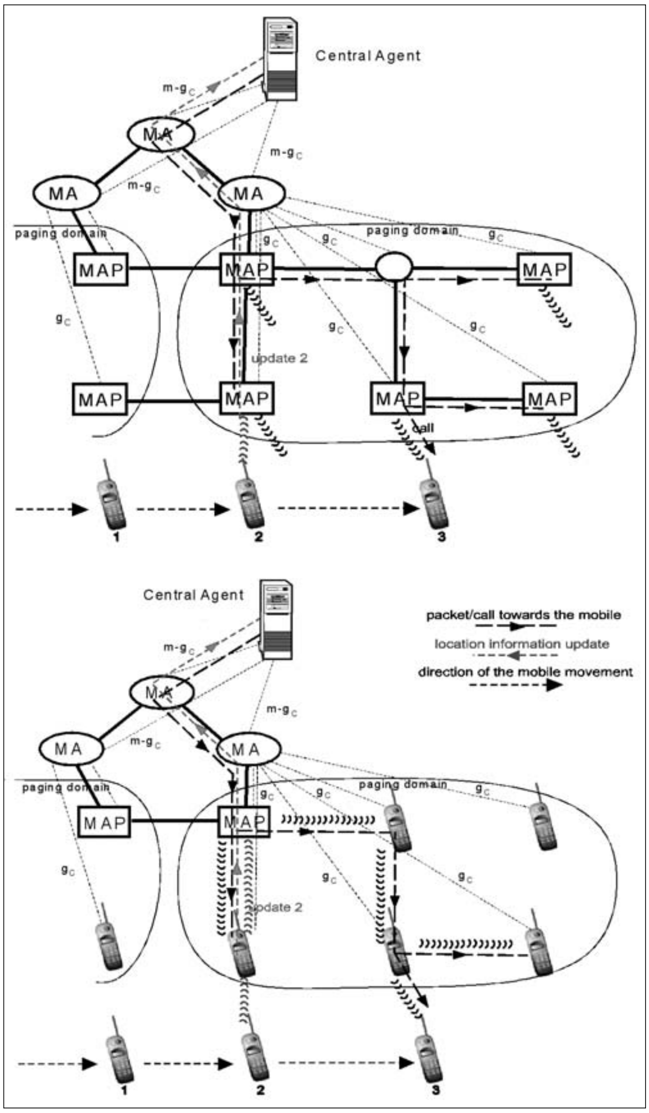
$$C_{CELLULAR}^{signal} = \rho(1-P_c)g_H c_u + (1-\rho)((m-g_C)+(n_C g_C c_d)+g_C c_u$$

$$\begin{aligned}C_{CELLULAR}^{process} = {} & \rho((1-P_c)c_r + g_H c_f + c_m) + \\ & +(1-\rho)(c_{ec} + (m-g_C-1)c_f + c_{rc} + (g_C-1)n_C c_f + c_{dc}c_r)\end{aligned}$$

$$C_{CELLULAR}^{air} = \rho((1-P_c)c_{au}) + (1-\rho)(n_C c_{ad} + c_{au})$$

where P_C the probability of entering a new page.

Figure 2. Standard cellular and MANET-like solutions



4.3.2. Hierarchical paging

The main idea behind the Hierarchical Paging [8] is that not only the lower layer network is flooded with the packet but broadcast message is used to find the paging controller MA in the higher layer as well. With this functionality signaling costs could be saved because update messages are not sent to HA, but only to the MA which controls the page.

But in case of calling the multilevel flooding causes high network load (7).

$$C_{HP}^{signal} = \rho(1-P_c)g_C c_u + (1-\rho)((m-g_C)+(n_C g_C c_d)+g_C c_u$$

$$\begin{aligned}C_{HP}^{process} = {} & \rho((1-P_c)c_r + g_C c_f + c_m) + \\ & +(1-\rho)(c_{ec} + (m-g_C-1)n_d c_f + c_{rc} + (g_C-1)n_C c_f + c_{dc}c_r)\end{aligned}$$

$$C_{HP}^{air} = \rho((1-P_c)c_{au}) + (1-\rho)(n_C c_{ad} + c_{au})$$

4.3.3. MANET in the page areas

The MANET [9] in the page areas solutions introduced by us could be the best solution when we would like to save the infrastructure cost and the air interface using is cheaper. In this management system it is assumed that all MN could be reached via other MNs. Paging areas are defined like in other cell-like solutions, but only one MAP exists in one page, through this the packets are routed using an optimal MANET algorithm. Advantage of this solution also is that signaling cost can be saved with correct MANET protocol in a page.

However, in the suboptimal case some mobiles could not be reached, and aggregate air interface cost can be high *(Figure 2)*. (8)

$$C_{MANET}^{signal} = \rho(1-P_c)g_C c_u + (1-\rho)((m-g_C+1)+(P_M n_C g_C c_d)+g_C c_u$$

$$\begin{aligned}C_{MANET}^{process} = {} & \rho((1-P_c)c_r + g_H c_f + c_m) + \\ & +(1-\rho)(c_{ec} + (m-g_C-1)c_f + c_{rc} + P_M n_C g_C c_f + c_{dc}c_r)\end{aligned}$$

$$C_{MANET}^{air} = \rho((1-P_c)(g_C-1)c_{au}) + (1-\rho)(P_M n_C g_C c_{ad} + c_{au})$$

In MANET like solutions at ad-hoc mobility level the request have to be sent via P_M percent of mobile nodes in order to be delivered it to the destination mobile node in a page.

4.4. Tracking-like Solutions

In the tracking-like approaches each mobile node has an entry in a Central Agent like in other solutions. This CA stores the address where it received location update message from. It is the address of an MAP, and is a next-hop towards the mobile node. The mobile node is either still connected to that MAP, or that MAP knows another next-hop MAP towards the mobile.

Finally the mobile node can be found at the end of a chain of MAPs. One can read more about these protocols in [2,6,11].

4.4.1. Wireless tracking

In case of tracking handover of wireless tracking the mobile sends the address of the new MAP node to the old MAP node over the air interface.

$$C_{WLESSTR}^{signal} = \rho P_H g_H c_u + (1-\rho)(g_H c_d + M[h_r]g_T c_d + (1-P_0)g_H c_u \qquad (9)$$

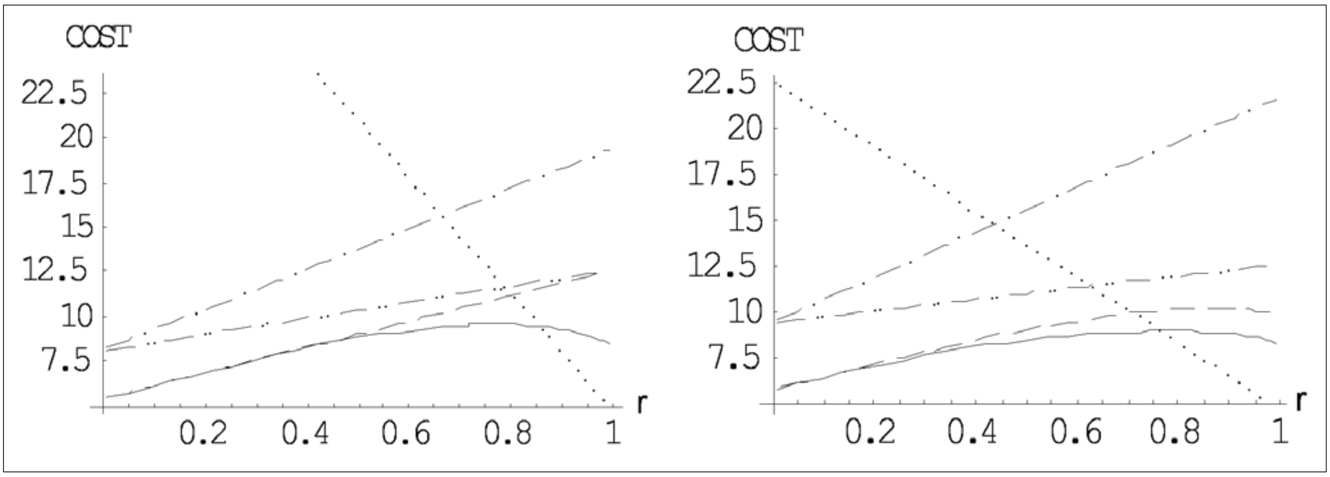


Figure 3.
One can see the summed cost functions of centralized-like (one-dot-dash), hierarchical-like (two-dot-dash), wireless (dashed) and wired (solid) tracking-like, cellular-like (dotted) approaches here with the vary of the mobility ratio. The two figures show the costs on different networks.

$$C^{process}_{WLESSTR} = \rho((1-P_H)(c_r+c_m)+P_H(c_r+(g_H-1)c_f+c_m)+$$
$$+(1-\rho)(c_{ec}+(m-1)c_f+P_0c_{dc}+$$
$$(1-P_0)(M[h_r]((g_T-1)c_f+c_{rc})+c_{dc}+(g_H-1)c_f+c_m))$$

$$C^{air}_{WLESSTR} = \rho c_{au}+(1-\rho)c_{ad},$$

where the $M[h_r]$ the number of tracking handovers after a normal handover, P_H the probability of that the Markovian model is in state H [6], which means a normal handover in the next step.

4.4.2. Wired tracking

Wired tracking differs from the wireless in the method of the tracking handover. In this case the MN sends the address of the new MAP node to the old MAP node through the wired network.

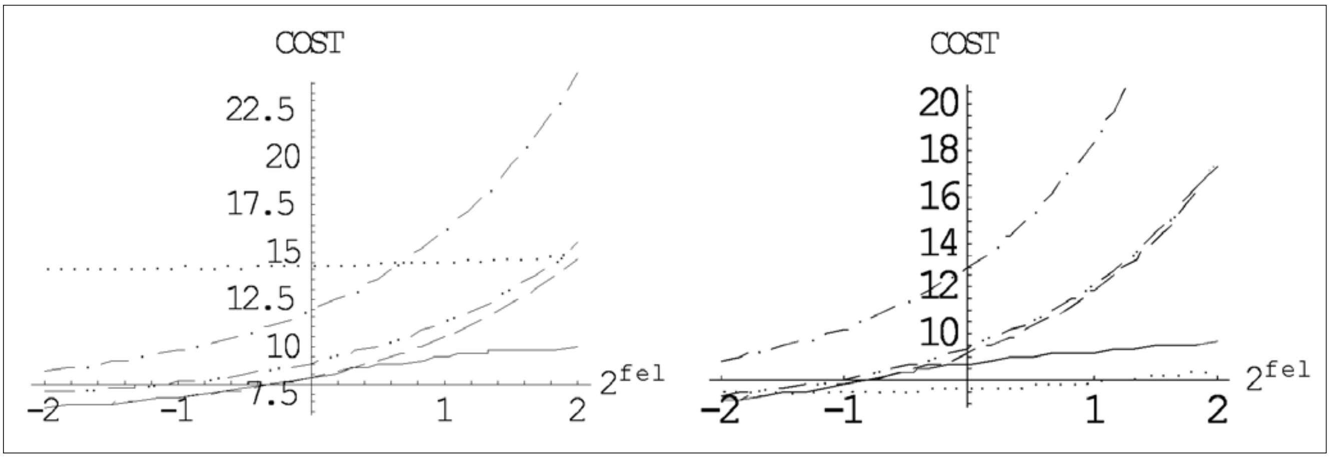
$$C^{signal}_{WTRACK} = \rho(g_T(1-P_H)+g_HP_H)c_u+(1-\rho)(mc_d+M[h_r]g_Tc_d+(1-P_0)g_Hc_u$$
$$C^{process}_{WTRACK} = \rho(c_r+(g_T-1)c_f+c_m)+(1-\rho)(c_{ec}+(m-1)c_f+P_0c_{dc}+$$
$$(1-P_0)(M[h_r]((g_T-1)c_f+c_{rc})+c_{dc}+(g_H-1)c_f+c_m)) \qquad (10)$$
$$C^{air}_{WTRACK} = \rho c_{au}+(1-\rho)c_{ad}$$

5. Numerical results

We do not attempt to give an exhausting numerical analysis with our method here since this paper focuses on the modeling framework itself. However, we give a very few examples for the type of investigations that could be performed using our model. The exact numerical values of the results are not important, we focus on the behavior of mobility with the change of the parameters.

In *Figure 3* one can see the difference between the approaches considering all the cost types (signaling, processing, air). It is clear that with the bigger frequency of handovers (ρ) the cost is bigger for the centralized-like, hierarchical-like and wired tracking-like approaches since each handover gives more signaling on the network. In the wireless tracking-like case if the number of handovers increases between the incoming calls, it starts saving the costs of the rerouting of the packets. In the centralized-like ones, it is clear that the rarer there is an incoming call the lower load the network has. The cost

Figure 4.
The uplink/downlink vary dependency with the same notation at ρ=0.7 and ρ=0.9.



is obviously high in this case. The same case is printed in both figures, but the values of g_T; g_C network parameters are significantly less than g_H (more meshed network). One can see that the wired tracking-like solution is getting cheaper as well and begins to behave as its tracking-like pair.

In *Figure 4* the mobility ratio is fixed to ρ = 0.7 and ρ = 0.9, respectively. On the other hand, the cost of a single upload (c_u) to a single download (c_d) is exponentially changing from the half to the twice on the horizontal axis. Most of the solutions are more expensive if the upload is higher but it can be seen that the wireless tracking cuts this cost as expected.

In *Figure 5* the different cellular strategies can be seen as a function of the air interface costs. One can see that the most optimal solution is MANET with low air costs. The larger the air interface cost is, the lower the difference is between the Hierarchical Paging diagram and Standard Cellular diagram.

One is able to perform further examinations using our Mathematica program.

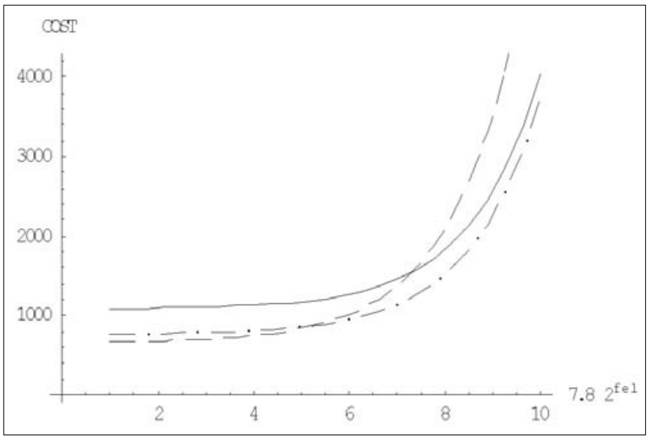


Figure 5.
Cost of cellular strategies as a function of air interface costs
(dashed: MANET, one-dot-dash: Standard Cellular, solid: Hierarchical Paging)

6. Conclusion and future work

Our primary aim was to develop an abstract modeling method for mobility managements. In this paper, we grabbed numerous significant parameters of mobility and modeled the mobile node behavior as well as the network and some general management strategies. Using our results, it can be shown which mobility management gives the best solution in different network scenarios and which aspect of resources could be a bottleneck in each case. One can use our achievements to analyze various mobility managements.

Our secondary aim, that is part of our future work, is to use the measurements to provide guidelines for the design of new mobility management algorithms and to propose solutions for different requirements.

References

[1] Kovács B., Fülöp P., Imre S.,
"Study on mobility management modeling methods",
MoMM2006, Yogyakarta, Indonesia, December 2006.

[2] Abondo, C., Pierre, S.,
"Dynamic Location and Forwarding Pointers
for Mobility Management",
Mobile Information Systems,
IOS Press, 2005., pp.3–24.

[3] A. T. Campbell, J. Gomez, A. G. Valkó,
"An Overview of Cellural IP",
IEEE, 1999., pp.29–34.

[4] C. Castelluccia,
"A Hierarchical Mobile IP Proposal",
INRIA Technical Report, 1998.

[5] Y. Fang, Y. Lin,
"Portable Movement Modelling for PCS Networks",
IEEE Transactions on Vehicular Technology, 2000.
pp.1356–1362.

[6] Kovács B., Szalay M., Imre S.,
"Modelling and Quantitative Analysis of LTRACK –
A Novel Mobility Management Algorithm",
Mobile Information Systems,
Vol. 2., No.1, 2006., pp.21–50.

[7] W. Ma, Y. Fang,
"Dynamic Hierarchical Mobility Management
Strategy for Mobile IP Networks",
IEEE Journal of Selected Areas in Comm., 2004.

[8] Szalay M., Imre S.,
"Hierarchical Paging –
A novel location management algorithm",
ICLAN'2006 International Conference on Late
Advances in Networks, 6-8. December 2006, Paris.

[9] Ashwini K. Pandey, Hiroshi Fujinoki,
"Study of MANET routing protocols
by GloMoSim simulator",
International Journal of Network Management, 2005.

[10] C.E. Perkins,
"Mobile IP",
IEEE Communications Magazine, 1997.

[11] R. Ramjee, T.La Porta, S. Thuel,
K. Varadhan, L. Salgarelli,
"A Hierarchical Mobile IP Proposal",
Inria Technical Report, 1998.

[12] Simon V., Imre S.,
Location Area Design Algorithms for Minimizing
Signalling Costs in Mobile Networks,
International Journal of Business Data
Communications and Networking (IJBDCN), 2007.

[13] Wolfram Research Inc.,
"Mathematica",
http://www.wolfram.com/

Energy balancing by combinatorial optimization for wireless sensor networks

JÁNOS LEVENDOVSZKY, ANDRÁS OLÁH*, CSEGŐ OROSZ, TIVADAR PÁPAI, THAN L. TRAN

Budapest University of Technology and Economics, Department of Telecommunications
{levendov, oroszcs}@hit.bme.hu, {pteddy, ttl}@cs.bme.hu
**Pázmány Péter Catholic University, Faculty of Information Technology*
olah@itk.ppke.bme.hu

Keywords: energy balancing, WSN, clusterhead, LEACH protocol, lifespan

The paper is concerned with developing new energy balancing protocols for wireless sensor networks (WSN) to maximize the life-span of the system by using rare event tools. When developing these new protocols, the statistical traffic characteristics of the sensed quantities are taken into account and some novel packet forwarding mechanisms from the nodes to the base station (BS) are proposed, which minimize the energy consumption of WSN. The tail distribution of the energy consumption is estimated by the tools of large deviation theory and the concept of generalized statistical bandwidth has been introduced to evaluate the energy need of the network. Furthermore, the clusterhead (CH) selection of "LEACH-like" protocols have been optimized by using spanning tree design and improved Li-Silvester bounds. The new results have been tested by extensive simulations which demonstrated that the lifespan of WSN can significantly be increased by the new protocols.

1. Introduction

Due to the recent advances in electronics and wireless communication, the development of low-cost, low-power, multifunctional sensors have received increasing attention [1]. These sensors are compact in size and besides sensing they also have some limited signal processing and communication capabilities. However, these limitations in size and energy make WSNs different from other wireless and ad-hoc networks [2]. As a result, new protocols must be developed with special focus on energy balancing in order to increase the lifetime of the network, which is crucial in applications where recharging of the nodes is out of reach (e.g. military field observations, living habitat monitoring etc., for more details see [4]).

The paper addresses energy balancing in WSN and develops novel packet forwarding mechanisms to increase the lifetime of the system. First a random class of protocols will be investigated, where the sensor nodes randomly select other nodes for packet forwarding, subject to a probability distribution. For example, node i can choose to forward to the neighbouring node closer to the base station (labeled as i-1) with probability 1–a_i, or send the packet directly to the BS with probability a_i. The optimal p.d.f. a_i, i=1,...,N is found which maximizes the tail of life-time distribution, based on large deviation theory by extending the concept of statistical bandwidth.

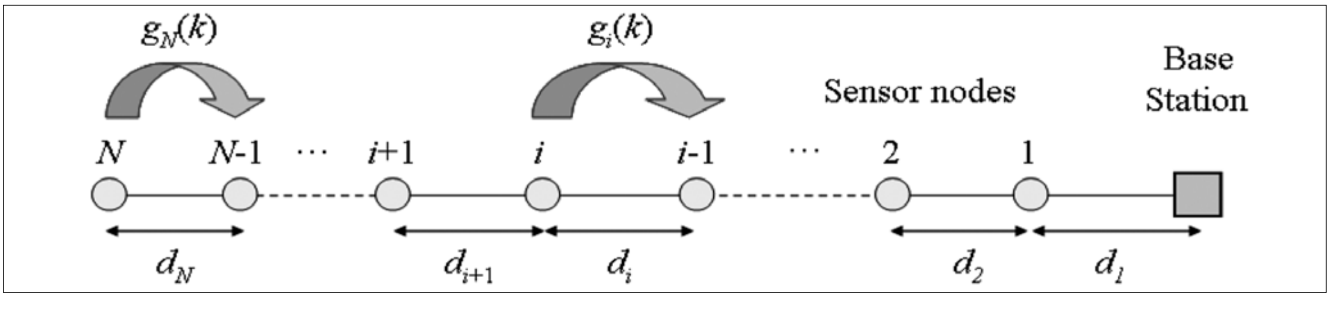
Then a LEACH-type protocol is analyzed. In this case, the active nodes select a cluster-head to which all the generated packet are sent and then the CH re-transmits the received packets to the BS. However, as opposed to the random CH selection of the traditional LEACH protocol (detailed in [6]), we select the CH by using an optimal spanning tree model. This spanning tree statistically optimizes the minimum remaining energy over all possible random traffic state vectors. Since the design of such a spanning tree is of exponential complexity, we develop a modification of the Li-Silvester bounds (which is known in statistical reliability analysis for reliability measure estimation) to optimize the protocol.

The new protocols can ensure longer WSN lifespan than the traditional packet forwarding mechanisms which is also demonstrated by extensive simulations.

2. The model

After the routing protocol (e.g. LEACH [6-8] or PEDAP [7]) has found the path to the base station, the subsequent nodes participating in the packet transfer can be regarded as a one dimensional chain labeled by i= 1,...,N and depicted by *Figure 1*.

Figure 1.
One dimensional chain topology of WSN packet forwarding



The system is characterized as follows:

- the topology is uniquely defined by a distance vector $\mathbf{d}=(d_1,...,d_N)$, where $d_i, i=1,...,N$ denotes the distance between node i and i-1, respectively;
- the energy needed to transmit packet over distance d is given as $g=\frac{d^\alpha \Theta \sigma_Z^2}{-\ln p_r}+g_{Elec}$ dictated by the Rayleigh model, where d is the distance, α depends on the propagation type, p_r is the reliability of correct reception, Θ is the modulation coefficient, σ_Z^2 is the noise energy, while g_{Elec} represents the consumption of the electronics during transmitting and receiving;
- the initial battery power on each node is the same and denoted by C;
- we assume that each sensor generates packets subject to an On/Off model, i.e. packet generation occurs with probability $P(y_i=1)=p_i$, whereas the node does not generate packet with probability $P(y_i=0)=1-p_i$;
- the traffic state of the network is represented by an N dimensional binary vector $\mathbf{y} \in \{0,1\}^N$ and the corresponding probability of a traffic state is given as $p(\mathbf{y})=\prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}$ assuming independence among the sensed quantities;
- the nodes operate in a time synchronous manner where the discrete time (clock signal) is denoted by $k=0,1,2,...$

As a result, a WSN is fully characterized by vectors **g**, **p** and **c**, respectively.

When analyzing the lifespan of the network, the following packet forwarding mechanisms are taken into account:

1. *Chain protocol:*
   Each node transmits packet to its neighbour laying closer to the BS. In this way, each node consumes minimal energy being engaged with short range energy transmission. However, each packet is traversing toward the BS, thus a packet consumes energy on each node along its path to the BS.
2. *Random shortcut protocol:*
   Node i can choose to forward the packet to its neighbouring node closer to the base station (labeled as i-1) with probability $1-a_i$, or directly send the packet to the BS with probability a_i.
3. *Single-hop protocol:*
   Each node sends its packet directly to the BS.
4. *CH protocol:*
   Each active node forwards its packet to a selected cluster-head and CH re-transmits them to the BS.

The paper is concerned with evaluating the lifetime of these protocols. Furthermore, our aim is to optimize probability vector $\mathbf{a}=(a_1,...a_N)$ and the CH selection in order to minimize energy consumption and thus maximizing the lifespan for WSNs operating with the random shortcut protocol.

3. Lifespan estimation by large deviation theory

Let assume that the chain protocol is in effect. The energy consumed by sending a packet generated on node i to the BS is given as

$$G_i := \sum_{j=1}^{i} g_j \,, \tag{1}$$

and the average energy consumption up to time instant K is given as

$$\sum_{k=1}^{K} \frac{1}{N} \sum_{i=1}^{N} y_i(k) G_i \,. \tag{2}$$

The lifespan of node denoted by $\tilde{K}$ is defined as

$$\tilde{K} : P\left(\sum_{k=1}^{K} \frac{1}{N} \sum_{i=1}^{N} y_i(k) G_i < C \right) = e^{-\alpha} \,, \tag{3}$$

where $e^{-\alpha}$ is close to one and
α is a reliability parameter.

By using the complementary probability

$$P\left(\sum_{k=1}^{K} \frac{1}{N} \sum_{i=1}^{N} y_i(k) G_i > C \right) = 1-e^{-\alpha} \,, \tag{4}$$

life time evaluation is cast as a tail estimation problem, where bounds like the Chernoff inequality can be used as (5):

$$P\left(\sum_{k=1}^{K} \frac{1}{N} \sum_{i=1}^{N} y_i(k) G_i > C \right) \le \exp\left(\sum_{i=1}^{N} \mu_i(\hat{s}, G_i) - \frac{\hat{s}NC}{K} \right)$$

$$\text{Here } \mu_i(s,G_i) := \log\left(\mathrm{E}\left[e^{s y_i G_i} \right] \right) = \log\left(1-p_i+p_i e^{sG_i} \right)$$

$$\text{and } \hat{s} : \min_s K \sum_{i=1}^{N} \mu_i(s,G_i) - \frac{sNC}{K} \,.$$

By using the estimation above, one obtains

$$e^{\sum_{i=1}^{N} \mu_i(\hat{s},G_i) - \frac{\hat{s}NC}{K}} = 1-e^{-\alpha} \tag{6}$$

and the lifespan of the simple chain protocol can finally be estimated by the following formula:

$$\tilde{K} = \frac{\hat{s}NC}{\sum_{i=1}^{N} \mu_i(\hat{s},G_i) + \log\left(1-e^{-\alpha}\right)} \,. \tag{7}$$

If the random shortcut protocol is in effect, then the packet generated by node i will travel in the chain down to the fits shortcut to BS. Let the node in which the shortcut takes place is denoted by λ_i. The distribution of λ_i is given as

$$P(\lambda_i = l_i) = a_{i-l_i} \prod_{j=i-l_i+1}^{i} (1-a_j) \,. \tag{8}$$

In this case the packet consumes $V_i := \sum_{j=i-l_i+1}^{i} g_j + \gamma_{i-l_i}$ energy, where γ_{i-l_i} is the shortcut energy from node $i-l_i$ (i.e. the energy required to transmit the packet from node $i-l_i$ directly the BS). As a result, the average energy consumption is given as

$$\sum_{k=1}^{K} \frac{1}{N} \sum_{i=1}^{N} y_i \left(\sum_{j=i-\lambda_i+1}^{i} g_j + \gamma_{i-\lambda_i} \right) \cdot \tag{9}$$

Thus the lifespan is defined as follows:

$$\tilde{K} : P\left(\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N} y_i\left(\sum_{j=i-\lambda_i+1}^{i} g_j + \gamma_{i-\lambda_i}\right) > C\right) = 1 - e^{-\alpha} \quad (10)$$

The probability in equation (10) can be rewritten as

$$P\left(\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N} y_i\left(\sum_{j=i-\lambda_i+1}^{i} g_j + \gamma_{i-\lambda_i}\right) > C\right) =$$

$$= \sum_{l_1}\cdots\sum_{l_N} P\left(\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N} y_i\left(\sum_{j=i-\lambda_i+1}^{i} g_j + \gamma_{i-\lambda_i}\right) > C \,\middle|\, \lambda_1 = l_1,\ldots,\lambda_N = l_N\right) P(\lambda_1 = l_1,\ldots,\lambda_N = l_N) =$$

$$= \sum_{l_1}\cdots\sum_{l_N} P\left(\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N} y_i\left(\sum_{j=i-l_i+1}^{i} g_j + \gamma_{i-l_i}\right) > C\right)\prod_{i=1}^{N} P(\lambda_i = l_i) \le$$

$$= \sum_{l_1}\cdots\sum_{l_N} e^{\sum_{i=1}^{N}\mu_i(s,V_i) - \frac{sNC}{K}} \prod_{i=1}^{N}\left(a_{i-l_i}\prod_{j=i-l_i+1}^{i}(1-a_j)\right) = e^{-\frac{sNC}{K}}\sum_{l_1}\cdots\sum_{l_N}\prod_{i=1}^{N} e^{\mu_i(s,V_i)}\left(a_{i-l_i}\prod_{j=i-l_i+1}^{i}(1-a_j)\right)$$

$$= \sum_{l_1}\cdots\sum_{l_N} e^{\sum_{i=1}^{N}\mu_i(s,V_i) - \frac{sNC}{K}} \prod_{i=1}^{N}\left(a_{i-l_i}\prod_{j=i-l_i+1}^{i}(1-a_j)\right) = e^{-\frac{sNC}{K}}\prod_{i=1}^{N}\sum_{l_i}\left(e^{\mu_i(s,V_i)}\left(a_{i-l_i}\prod_{j=i-l_i+1}^{i}(1-a_j)\right)\right),$$

where $\mu_i(s,V_i) := \log\left(\mathrm{E}\left[e^{sy_iV_i}\right]\right) = \log\left(1 - p_i + p_i e^{sV_i}\right)$.

Introducing the extended logarithmic moment generation function as

$$\beta_i(s,V_i) := \log\left(\sum_{l_i}\left(e^{\mu_i(s,V_i)} a_{i-l_i}\prod_{j=i-l_i+1}^{i}(1-a_j)\right)\right). \quad (11)$$

one can write

$$P\left(\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N} y_i\left(\sum_{j=i-\lambda_i+1}^{i} g_j + \gamma_{i-\lambda_i}\right) > C\right) \le e^{-\frac{sNC}{K}}\prod_{i=1}^{N} e^{\beta_i(s,V_i)} = e^{\sum_{i=1}^{N}\beta_i(s,V_i) - \frac{sNC}{K}}. \quad (12)$$

Comparing the bound with $1-e^{-\alpha}$, we obtain

$$e^{\sum_{i=1}^{N}\beta_i(\hat{s},V_i) - \frac{\hat{s}NC}{K}} = 1 - e^{-\alpha}, \quad (13)$$

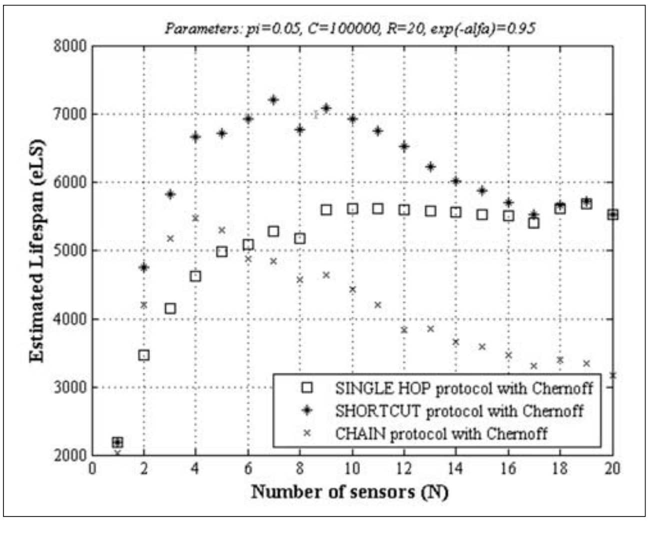
where

$$\hat{s} : \min_s \sum_{i=1}^{N}\beta_i(s,V_i) - \frac{sNC}{K}.$$

The lifespan is the solution of the following equation:

$$\tilde{K} : \sum_{i=1}^{N}\beta_i(\hat{s},V_i) = \frac{\hat{s}NC}{K} + \log\left(1 - e^{-\alpha}\right). \quad (14)$$

Figure 2.
Estimated lifespan in the function of the number of sensors



As one can see the equation above determines the lifespan as a function of vector **a**, the components of which represent the probabilities of shortcut on a given node. This relationship is denoted by $\tilde{K} = \Psi(\mathbf{a})$.

Using equations (11) and (14) to evaluate $\Psi(\mathbf{a})$ for a given **a** vector, protocol optimization can take place by searching in the space of **a**-vectors to find the optimal shortcut probabilities. This can be done by gradient descent type of optimization given as follows (15):

$$a_i(n+1) = a_i(n) - \Delta\,\mathrm{sgn}\left\{\frac{\Psi(\mathbf{a}(n)) - \Psi(\mathbf{a}(n-1))}{a_i(n) - a_i(n-1)}\right\}, i = 1,\ldots,N$$

As a result, protocol optimization has been carried out in the following steps:

Given: N - number of nodes,
p - packet generate probability vector,
g - energy vector,
c - initial battery power vector;

Step 1. select an initial $\mathbf{a}(0)$ shortcut probability vector;

Step 2. evaluate the value of $\tilde{K} = \Psi(\mathbf{a})$ by solving the equation $\tilde{K} : \sum_{i=1}^{N}\beta_i(\hat{s},V_i) = \frac{\hat{s}NC}{K} + \log\left(1 - e^{-\alpha}\right)$;

Step 3. Perform the gradient search
$a_i(n+1) = a_i(n) - \Delta\,\mathrm{sgn}\left\{\frac{\Psi(\mathbf{a}(n)) - \Psi(\mathbf{a}(n-1))}{a_i(n) - a_i(n-1)}\right\}$,
where $i = 1,\ldots,N$;

Step 4. Check the stopping criterion (i.e. $\|\mathbf{a}(n+1) - \mathbf{a}(n)\| \le \varepsilon$) and go back to Step 2. if it is not met.

In the case of single-hop protocol we have $a_i = 1$, $i = 1\ldots N$. Thus,

$$\tilde{K} : P\left(\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N} y_i(k)\gamma_i < C\right) = e^{-\alpha} \quad (16)$$

which leads to the following life span

$$\tilde{K} = \frac{\hat{s}NC}{\sum_{i=1}^{N}\mu_i(\hat{s},\gamma_i) + \log\left(1 - e^{-\alpha}\right)}, \quad (17)$$

where $\mu_i(s,\gamma_i) := \log\left(\mathrm{E}\left[e^{sy_i\gamma_i}\right]\right) = \log\left(1 - p_i + p_i e^{s\gamma_i}\right)$

is the energy required by the shortcut.

3.1. Performance analysis and numerical results

In this section a detailed performance analysis is given using the chain, the shortcut and the single-hop protocols. The aim is to evaluate the lifespan of a sensor network containing N number of sensors placed in an equidistant manner. *Figure 2* shows how the lifespan changes as the function of the number of nodes (*N*) in the case of the three methods described above. The distance between the base station and the farthest node was 20 meters and the nodes were located randomly subject to a Poissonian distribution.

One can see that there is a maximum lifespan in the cases of chain and random shortcut protocols with the optimal number of nodes $N_{Chain}=4$ and $N_{Shortcut}=7$, respectively.

Figure 2 shows that when the network is sparsely installed, both methods result in almost the same lifespan, while departing form the optimal number of nodes (either decreasing or increasing the number of sensors), the shortcut model definitely gives much higher relative lifespan (it is more than 37% in the case of *N*=7).

Figure 3 demonstrates the accuracy of lifespan estimation at the different protocols. One can see that the Chernoff bound yields a relatively sharp estimation.

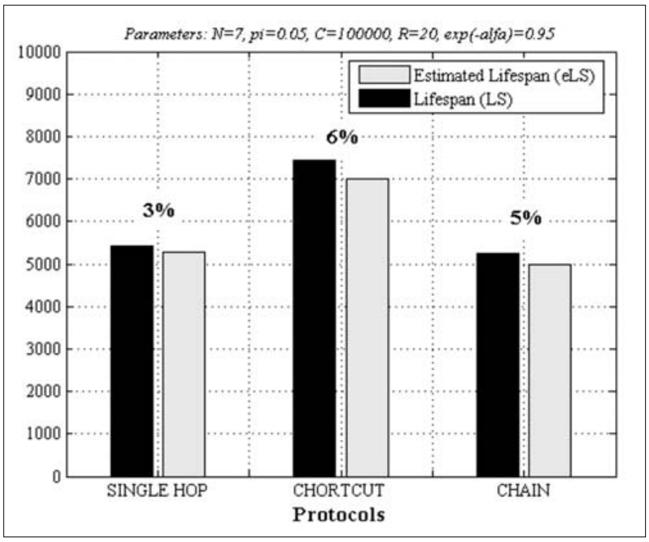


Figure 3. Lifespan and estimated lifespan values achieved by different protocols

4. Spanning tree design for optimal clusterhead selection

In this case, the network elects a CH (being the root of the spanning tree), the index of which is denoted by ξ. The nodes where packets are generated communicate with the CH via a single hop communication and then the CH re-transmits these packets to the BS. This gives rise to the following model:

- the WSN is in an energy state $\mathbf{c}(k)=\left(c_1(k),...,c_N(k)\right)$ where $c_i(k)$ denotes the available energy at node i
- a traffic vector $\mathbf{y}$ occurs with probability $p\left(\mathbf{y}\right)=\prod_{i=1}^{N} p_i^{y_i}\left(1-p_i\right)^{1-y_i}$
- the energy consumption of conveying the traffic to the BS on node i is given as
$$\eta_i=\begin{cases} G_{i\xi} & \text{if } i\neq\xi \\ w\left(\mathbf{y}(k)\right)G_{\xi 0} & \text{if } i=\xi \end{cases}$$
- the new energy state is $\mathbf{c}(k+1)=\left(c_1(k+1),...,c_N(k+1)\right)$ where $c_i(k+1)=c_i(k)-\eta_i$

In order to maximize the lifespan, we are concerned with finding $\xi_{opt}:\max_{\xi}\min_{i} c_i(k+1)$, which guarantees the longest lifespan of the bottleneck node. We will refer to this objective as $\xi_{opt}:\max_{\xi}\psi\left(\xi\right)$ where $\psi\left(\xi\right):=\min_{i} c_i(k+1)$.

If vector $\mathbf{y}(k)$ is known, then this optimum can easily be calculated in polynomial time, due to the single hop nature of the spanning tree. In this case, one can evaluate the function $\psi\left(\xi,\mathbf{y}(k)\right):=\min_{i} c_i(k+1)$ "$w\left(\mathbf{y}(k)\right)$-times"

by placing the root node in different positions and ξ_{opt} can be selected.

However, this calculation cannot be carried out, as neither the BS nor the nodes are aware of the current traffic vector $\mathbf{y}(k)$.

Thus location of the root node can only be optimized in the mean sense, by finding:

$$\xi_{opt}:\max_{\xi}\sum_{\mathbf{y}\in\{0,1\}^N}\psi\left(\xi,\mathbf{y}\right)p\left(\mathbf{y}\right)$$

This optimization can be reformulated as follows:

$$\xi_{opt} : \max_{\xi} f(\xi) \quad \text{where } f(\xi) := \sum_{\mathbf{y} \in \{0,1\}^N} \psi(\xi, \mathbf{y}) p(\mathbf{y}) \;. \tag{18}$$

Of course the optimal index ξ_{opt} depends on the current energy state. Thus this optimization must be carried out each time when transmitting a traffic vector $\mathbf{y}(k)$ to the BS.

To carry out this optimization, one needs an efficient estimation of

$$g(\xi) \approx f(\xi) := \sum_{\mathbf{y} \in \{0,1\}^N} \psi(\xi, \mathbf{y}) p(\mathbf{y})$$

in order to circumvent the exponentially large summation in (18).

In the next section we develop a powerful approximation of $f(\xi)$ based on the modification of the Li-Silvester bounds (the original LS bound can be found in [10]).

4.1. Modified LS bounds to estimate the minimum remaining energy

In this section we propose a novel approach to estimate $f(\xi) := \sum_{\mathbf{y} \in \{0,1\}^N} \psi(\xi, \mathbf{y}) p(\mathbf{y})$ which is an improvement of the LS bounds. The original bounds are given as follows:

Let us assume that the values of $\psi(\xi, \mathbf{y})$ can be lower and upper bounded as

$0 \le \psi(\xi, \mathbf{y}) \le \psi_{\max}(\xi)$

And we pick the first K relevant vectors $Y_1 := \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(K)}\}$ for which $p(\mathbf{y}^{(1)}) \ge \ldots . \ge p(\mathbf{y}^{(K)})$ and $p(\mathbf{y}^{(K)}) > p(\mathbf{y}) \;\; \forall \mathbf{y} \notin Y_1$. Then

$$\sum_{\mathbf{y} \in Y_1} \psi(\xi, \mathbf{y}) p(\mathbf{y}) \le \sum_{\mathbf{y} \in \{0,1\}^N} \psi(\xi, \mathbf{y}) p(\mathbf{y}) = \sum_{\mathbf{y} \in Y_1} \psi(\xi, \mathbf{y}) p(\mathbf{y}) + \sum_{\mathbf{y} \in Y_2} \psi(\xi, \mathbf{y}) p(\mathbf{y}) \le \sum_{\mathbf{y} \in Y_1} \psi(\xi, \mathbf{y}) p(\mathbf{y}) + \psi_{\max}(\xi) P(Y_2),$$

where $Y_1 \cup Y_2 := \{0,1\}^N$

The estimation error can be upper bounded with $\psi_{\max}(\xi) P(Y_2)$ which is minimal due to the fact that set Y_1 contains the most probable elements.

The bound given above can be modified as follows:

Let us define the binary vector $\tilde{\mathbf{y}}$ as a "descendant" of binary vector **y** if (i) $w(\tilde{\mathbf{y}}) > w(\mathbf{y})$ and (ii) $y_i = 1$ implies that $\tilde{y}_i = 1$ (e.g. one of the descendants of vector $\mathbf{y} = (010010)$ is $\mathbf{y} = (111010)$). In the forthcoming discussion this relationship is denoted as $\mathbf{y} \prec \tilde{\mathbf{y}}$. It is noteworthy that if $\mathbf{y} \prec \tilde{\mathbf{y}}$ then $\psi(\xi, \mathbf{y}) \ge \psi(\xi, \tilde{\mathbf{y}})$ (i.e. in the case of a descendant vector there are some additional packets to be transmitted to the BS).

Let us divide the set $\{0.1\}^N$ into three disjoint subsets as follows:

$\{0.1\}^N = Y_1 \cup Y_2 \cup Y_3$ in such a manner that $Y_1 := \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(K)}\}$, $p(\mathbf{y}^{(1)}) \ge \ldots . \ge p(\mathbf{y}^{(K)})$, $p(\mathbf{y}^{(K)}) > p(\mathbf{y}) \;\; \forall \mathbf{y} \notin Y_1$, $Y_2 = \{\tilde{\mathbf{y}} : \tilde{\mathbf{y}} \succ \mathbf{y}, \mathbf{y} \in Y_1\}$ and $Y_3 = \{\hat{\mathbf{y}} : \hat{\mathbf{y}} \succ \tilde{\mathbf{y}}, \tilde{\mathbf{y}} \in Y_2\}$.

Let us define $A_{\tilde{\mathbf{y}}} := \{\mathbf{y} : \mathbf{y} \prec \tilde{\mathbf{y}}, \mathbf{y} \in Y_1\}$, $\alpha(\tilde{\mathbf{y}}) := \mathbf{y}^* : \min_{\mathbf{y} \in A_{\tilde{\mathbf{y}}}} \psi(\xi, \mathbf{y})$ and $B_{\mathbf{y}^*} := \{\tilde{\mathbf{y}} : \alpha(\tilde{\mathbf{y}}) = \mathbf{y}^*\}$.

In a similar fashion $A_{\hat{\mathbf{y}}} := \{\mathbf{y} : \mathbf{y} \prec \hat{\mathbf{y}}, \mathbf{y} \in Y_1\}$, $\beta(\hat{\mathbf{y}}) := \mathbf{y}^{**}$ is selected to be the first element in $A_{\hat{\mathbf{y}}}$ and $B_{\mathbf{y}^{**}} := \{\hat{\mathbf{y}} : \beta(\hat{\mathbf{y}}) = \mathbf{y}^{**}\}$. One must note that while $\alpha(\tilde{\mathbf{y}}) := \mathbf{y}^* : \min_{\mathbf{y} \in A_{\tilde{\mathbf{y}}}} \psi(\xi, \mathbf{y})$ involves an optimization over a smaller set, $\beta(\hat{\mathbf{y}}) := \mathbf{y}^{**}$ can be obtained automatically. As a result, the size of Y_2 in the partition of $\{0.1\}^N = Y_1 \bigcup Y_2 \bigcup Y_3$ is determined by the available time and computational resources to carry out $\alpha(\tilde{\mathbf{y}}) := \mathbf{y}^* : \min_{\mathbf{y}\in A_{\tilde{\mathbf{y}}}} \psi(\xi,\mathbf{y})$ for each $\tilde{\mathbf{y}} \in Y_2$.

Then $\sum_{\mathbf{y}\in\{0,1\}^N} \psi(\xi,\mathbf{y})p(\mathbf{y}) = \sum_{\mathbf{y}\in Y_1} \psi(\xi,\mathbf{y})p(\mathbf{y}) + \sum_{\tilde{\mathbf{y}}\in Y_2} \psi(\xi,\tilde{\mathbf{y}})p(\tilde{\mathbf{y}}) + \sum_{\hat{\mathbf{y}}\in Y_3} \psi(\xi,\hat{\mathbf{y}})p(\hat{\mathbf{y}})$, where the second term and third terms can be upper bounded as follows:

$$\sum_{\tilde{\mathbf{y}}\in Y_2} \psi(\xi,\tilde{\mathbf{y}})p(\tilde{\mathbf{y}}) \le \sum_{\tilde{\mathbf{y}}\in Y_2} \psi(\xi,\mathbf{y}^*)p(\tilde{\mathbf{y}}) = \sum_{\mathbf{y}^*} \psi(\xi,\mathbf{y}^*)p\left(B_{\mathbf{y}^*}\right)$$

$$\sum_{\hat{\mathbf{y}}\in Y_3} \psi(\xi,\tilde{\mathbf{y}})p(\tilde{\mathbf{y}}) \le \sum_{\hat{\mathbf{y}}\in Y_3} \psi(\xi,\mathbf{y}^{**})p(\hat{\mathbf{y}}) = \sum_{\mathbf{y}^{**}} \psi(\xi,\mathbf{y}^{**})p\left(B_{\mathbf{y}^{**}}\right) \cdot$$

In this way a new upper bound can be obtained as

$$\sum_{\mathbf{y}\in\{0,1\}^N} \psi(\xi,\mathbf{y})p(\mathbf{y}) \le \sum_{\mathbf{y}\in Y_1} \psi(\xi,\mathbf{y})p(\mathbf{y}) + \sum_{\mathbf{y}^*} \psi(\xi,\mathbf{y}^*)p\left(B_{\mathbf{y}^*}\right) + \sum_{\mathbf{y}^{**}} \psi(\xi,\mathbf{y}^{**})p\left(B_{\mathbf{y}^{**}}\right). \qquad (19)$$

This upper bound is sharper than the original LS bound due to the fact that

$\psi(\xi,\mathbf{y}^{**}) \le \psi_{\max}(\xi)$, $\psi(\xi,\mathbf{y}^*) \le \psi_{\max}(\xi)$ and $\sum_{\mathbf{y}^*} p\left(B_{\mathbf{y}^*}\right) = P(Y_2)$, $\sum_{\mathbf{y}^{**}} p\left(B_{\mathbf{y}^{**}}\right) = P(Y_3)$.

For the sake of notational simplicity, the bound in (19) will be denoted as

$$g(\xi) := \sum_{\mathbf{y}\in Y_1} \psi(\xi,\mathbf{y})p(\mathbf{y}) + \sum_{\mathbf{y}^*} \psi(\xi,\mathbf{y}^*)p\left(B_{\mathbf{y}^*}\right) + \sum_{\mathbf{y}^{**}} \psi(\xi,\mathbf{y}^{**})p\left(B_{\mathbf{y}^{**}}\right).$$

4.2. Computational model to find the optimal CH

Based on the discussion above, the modified LS bound will be used to estimate $f(\xi)$, which gives rise to the following two protocol optimization models.

In the first case, we assume that $\mathbf{y}(k)$ is known prior to the transmission and then optimization is carried out as indicated by *Figure 4*.

However, this case will only serve as a reference for the performance analysis, as the traffic vector $\mathbf{y}(k)$ cannot be known prior to the transmission.

Therefore, the protocol optimization takes place according to (18), which yields the algorithm depicted in *Figure 5*.

4.3. Numerical results

In *Figure 6* the lifespan obtained by the chain, single-hop, and CH protocols are plotted as a function of the number of the nodes. The results were obtained on the same WSN as described in Section 3.1.

In the case of adaptive CH protocol, the clusterhead is selected as a function of the current traffic vector $\mathbf{y}(k)$, whereas in the case of average CH protocol the clusterhead was selected by maximizing the expected value of the minimum remaining energy and the expectation was taken over the whole traffic state space $\{-1,1\}^N$. On the other hand, the chain protocol forwarded packets node-by-node to the BS, while the single hop protocol transmitted packets directly to the BS from each active nodes.

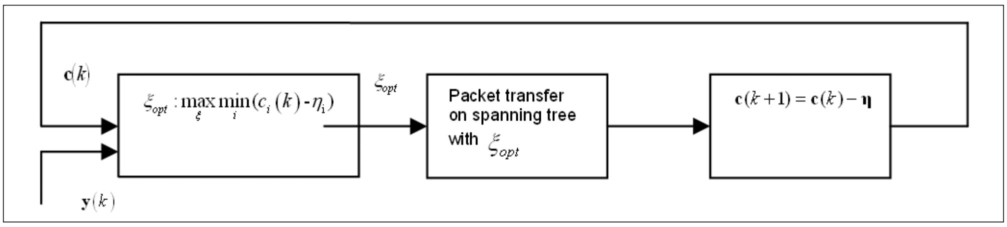

c(k)
ξ_opt : max_ξ min_i (c_i(k) - η_i)
ξ_opt
Packet transfer on spanning tree with ξ_opt
c(k+1) = c(k) − η
y(k)

Figure 4. Adaptive CH optimization assuming known traffic vector $\mathbf{y}(k)$

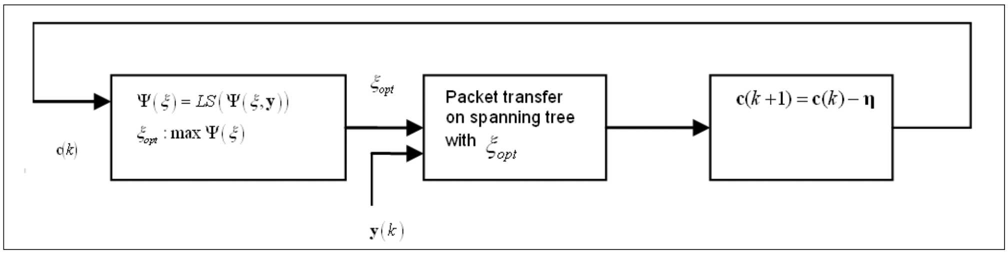


Figure 5. CH optimization on the basis of energy vector $\mathbf{c}(k)$

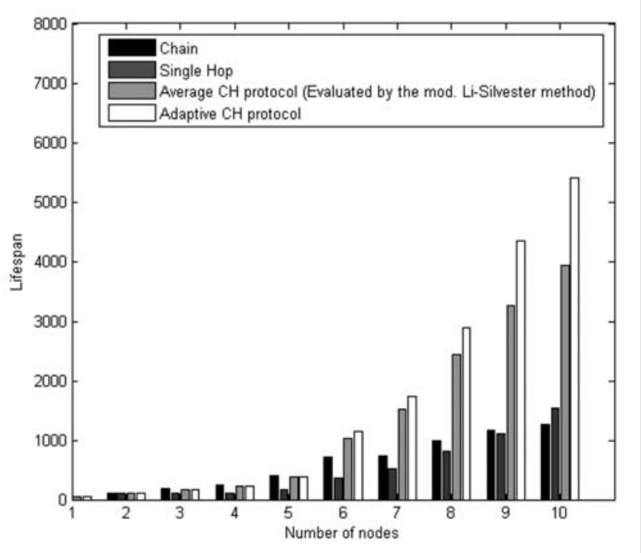


Figure 6.
Comparing the lifespan of CH type protocols to the traditional ones

One can see that the CH type protocols outperform the chain and single-hop communication as far as the lifespan is concerned. In the case of ten nodes the lifespan has been increased to three or four times longer than the lifespan obtained by traditional methods. This strongly motivates the use of CH type protocols.

5. Conclusions

In this paper, energy balancing of WSN has been studied by statistical tools. A novel “random shortcut” protocol has been introduced and the optimal probability distribution for selecting destination for packet forwarding has been found.

Identifying the optimal CH has also been considered by using a spanning tree model and a novel bound to estimate the means of the remaining energy function. Both protocols can significantly increase the lifespan of WSN. The performance of the methods have been tested by extensive simulations which also demonstrated the improvement on the lifespan.

References

[1] C.Y. Chong, S.P. Kumar,
“Sensor networks:
Evolution, opportunities and challenges”.
IEEE Proceedings, pp.1247–1254., August 2003.

[2] A. Goldsmith, S. Wicker,
“Design challenges for energy-constrained
ad hoc wireless networks”.
IEEE Wireless Communications Magazine, 9:8-27.
August 2002.

[3] “Self-healing Mines”
http://www.darpa.mil/ato/programs/SHM/

[4] A. Mainwaring, J. Polastre, R. Szewczyk,
D. Culler, J. Anderson,
“Wireless sensor networks for habitat monitoring”,
First ACM Workshop on Wireless Sensor Networks
and Applications, Georgia, Atlanta,
September 2002.

[5] D. Puccinelli, M. Haenggi,
“Wireless Sensor Networks-Applications and
Challenges of Ubiquitous Sensing”.
IEEE Circuits and Systems Magazine, 5:19-29.
August, 2005.

[6] W. Heinzelman, A. Chandrakasan, H. Balakrishnan,
“Energy-Efficient Communication Protocols for
Wireless Microsensor Networks”.
Proc. Hawaaian Int’l Conf. on Systems Science,
January 2000.

[7] W. Heinzelman, A. Sinha, A. Wang, A. Chandrakasan,
“Energy-scalable algorithms and protocols for
wireless microsensor networks”.
Proc. International Conference on Acoustics,
Speech and Signal Processing (ICASSP‘00).
June, 2000.

[8] W. Heinzelman, A. Chandrakassan, H. Balakrishnan,
“An application-specific protocol architecture for
wireless microsensor networks”.
IEEE Trans. on Wireless Comm., 1 (4), 2002.

[9] Huseyin Ozgur Tan, Ibrahim Korpeoglu,
“Power Efficient Data Gathering and Aggregation
in Wireless Sensor Networks”.
ACM SIGMOD Record, 32 (4), pp.66–71.
December 2003.

[10] V.O. Li, J.A. Silvester,
“Performance Analysis of Networks with
Unreliable Components”.
IEEE Trans. on Comm., COM-32 10, pp.1105–1110.
October 1984.

Linearity and chirp investigations on Semiconductor Optical Amplifier
as an external optical modulator

ESZTER UDVARY

Budapest University of Technology and Economics, Dept. of Broadband Infocom Systems
udvary@mht.bme.hu

Keywords: amplitude and phase modulation, chirp, nonlinearity, SOA, subcarrier multiplexing

This paper provides an overview of the basics and application possibilities of the multifunctional Semiconductor Optical Amplifier (SOA) in Sub-Carrier Multiplexed (SCM) systems. The paper focuses on the linearity investigation of the device. It describes the frequency dependence of the modulation and the harmonic products, the effects of the bias current and the optical power, the mismatch between the light and the electrical signal, the temperature and optical reflection sensitivity. It is shown by numerical simulation and measurements that by using SOA as an external modulator, the device provides acceptable nonlinear distortion for SCM telecommunication systems. Finally, the frequency chirping in external SOA modulator is treated for different operation conditions.

1. Introduction

Optical sub-carrier multiplexing (SCM) is a scheme where multiple signals are multiplexed in the radio frequency (RF) domain and transmitted by a single optical wavelength. The sub-carriers usually are in the range of the microwave and millimeter waves, because the optically transmitted channels are converted into/from the RF domain. There are combined system, which utilize both baseband and subcarrier signals.

The literature suggests the application of SCM in several systems for transmission and distribution the microwave or millimeterwave signals. A popular application of SCM technology in fiber optic systems is analog cable television (CATV) distribution. Typical application is the remote antenna feeding in radar systems, where the high frequency signal must be transmitted to the antenna with low loss.

SCM has also been proposed to transmit multi-channel digital optical signals using direct detection for local area optical networks, microwave signal distribution in picocell-based communication systems, combined wireless data communication systems. SCM is used in the picocellular wireless (possibly mobile) telecommunication systems, where several radio channels are needed in certain cells. In the fiber-radio systems the huge bandwidth of the optical transmission allows the radio frequency carriers to be directly transported over the optical fiber without the need for frequency conversion or multiplexing/demultiplexing functions. Therefore, complex processing equipment can be located in a local exchange, thus simplifying field installation and maintenance procedures. The system is very flexible, it can easily be extended to contain more terminals, the number and frequency of subcarriers can be modified according to the traffic.

In the complex systems the baseband signal is transmitted in parallel with the subcarrier information. The photonic switched networks with label on subcarrier utilize both baseband and subcarrier information. A baseband digital label is modulated onto a RF subcarrier and then multiplexed (electronically or optically) with the baseband packet on the same wavelength.

Current technologies utilize several separate optical elements in the presented systems. It would be beneficial if a multi-functional optical element were provided to reduce number of the components, size, maintenance, production costs and complexity. However, the inherent design trade-off between different functions demands more advanced design. The special devices have better parameters than the multifunctional device. The degradation has to be minimized, hence the study of the potential multifunctional devices is very important.

2. The Semiconductor Optical Amplifier

The Semiconductor Optical Amplifier is based on the same technology as a semiconductor laser diode, but the cavity reflections are blocked by using antireflection techniques. So, the SOA is a semiconductor based, small size, potentially cheap, electrically pumped device, which has large optical bandwidth. Moreover, the semiconductor technology offers a wide flexibility in the choice of the operating wavelength by just appropriately choosing the material composition of the active layer. The small size and compatibility with semiconductor laser sources and semiconductor detectors offer the possibility of photonic integration with other active or passive optical components.

It amplifies the incoming weak optical signal directly in the optical regime without any optical-electrical, electrical-optical conversion. The stimulated emission provides the amplification, the absorption means optical loss, the spontaneous emission is the source of noise. It is a random process, which is statistically stationary and will cause fluctuations in both amplitude and phase of optical signal. The operation can be described by the multimode rate equations, like semiconductor lasers. However the total carrier density is time and spatial dependent and a term for the optical injection is added.

The operation of the multifunctional SOA-modulator is based on the following phenomenon. The electrical bias current of the SOA is modulated, therefore the material gain is modulated, and consequently in case of continuous wave input the intensity of the output power is also modulated [1]. If small signal sinusoidal current modulation is considered, the electrical signal consists of an invariant and a sinusoidal modulation parts, hence the number of carriers and photons are also time dependent and the shape of these parameters are similar to the shape of the modulation [2]. The device amplifies the incoming optical signal and adds an intensity modulated component. The intensity modulated optical signal can be detected by traditional pin photodiode. The magnitude and purity of the signal depend on the modulation signal, the bias current, the input power and the operation parameters of the SOA [3].

The SOA modulator requires low modulation power, the detected electrical power is high because of the optical gain of the SOA in contrary to the optical insertion loss of other external modulators. In SOAs the gain dynamics are determined by the carrier recombination lifetime (few hundred picoseconds), hence the modulation bandwidth is limited by the electrical circuits. However, the SOA has remarkable optical noise and the optimal operation demands more advanced amplifier-modulator working state planning.

The SOA can provide the branching function in the SCM systems. It operates as a modulator to add a new channel, as a detector to drop the needed channel and as an in-line amplifier to amplify the other channels, simultaneously. It realizes a compact, small size and cost-effective radio repeater for signal distribution [4]. The achieved functions are similar in Fiber-to-the-Home Networks, where simple optical network unit is needed for the customer [5].

The compact SOA-modulator can solve the optical sub-carrier label swapping problem in sub-carrier label packet switched all optical systems. The wavelength conversion and all-optical regeneration can be achieved through cross-phase modulation (XPM) performed in a SOA based active Mach-Zehnder interferometer. Current modulation of the SOA in one or both arms of the wavelength converter is used to add the new label [6].

3. Linearity investigation

Cascadability is critically important in optical SCM networks where several electrical subcarriers are transmitted on the same optical signal. Degradation of the transmission system will occur due to the crosstalk between the subcarriers (nonlinearity) and noise expansion (ASE) [7].

The traditionally used electro-optical modulator shows high nonlinearity, because it has a cosine type characteristic. The photo-detector and the optical fiber can be treated as near linear device. The SOA-modulator can improve the nonlinear behavior of the system, if it provides lower nonlinear distortion than the electro-optical external modulators.

The second and third order intermodulations will be considered, because of the crosstalk between the channels and the partial up-conversion of the baseband payload into the subcarrier. As the number of subcarriers increases the linearity becomes a more and more serious problem because many third order mixing products appear in the used band.

3.1. Simulation results

The SOA model uses a pair of coupled partial differential equations, the wave and the rate equations. The model takes into account the detailed nonlinear carrier recombination rate:

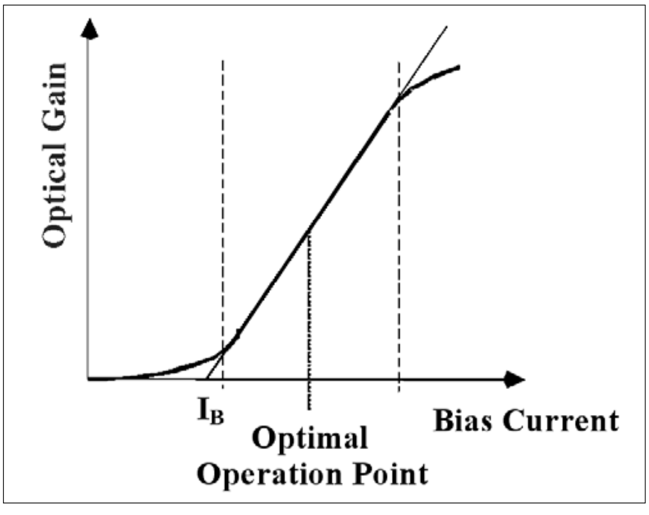
$$R(N) = A \cdot N + B \cdot N^2 + C \cdot N^3 \qquad (1)$$

Here *N, A, B,* and *C* are the spatial dependent carrier density, the non-radiative recombination rate, the radiative recombination coefficient and the Auger recombination coefficient, respectively.

The carrier density is obtained by solving the spatial dependent rate equation, and the propagation of the electromagnetic field inside the amplifier is governed by solving the wave equation. The time dependent amplifier's output power is calculated by solving numerically the coupled rate and wave equations.

There are two types of the nonlinear distortion of the SOA [8]. The static distortion is caused by the nonlinearity of the amplifier output power-current curve under continuous wave condition. The dynamic distortion is caused by signal-induced carrier density modulation. During the simulation the nonlinearity of the amplifier is characterized by using a single tone modulation. The static distortion is calculated directly from the optical gain – current curve [9] shown in *Figure 1*.

Figure 1. Static distortion, optimal bias point of SOA



The main objective is to select the most linear region of the curve over a wide bias current range, and then to place the dc operating point roughly at the middle of this region. It is strongly dependent on the input optical power.

With the optimal operation conditions, the calculated values of the static nonlinear distortions are less than the dynamic distortions, hence static distortions will not be taken into account.

The dynamic nonlinearity is calculated by numerical analysis of the output optical power. *Figure 2* represents the optical power (Pdc, broken line), the signal levels for the fundamental (P1), the second (P2) and the third (P3) order harmonic products versus the bias point. The operation is strongly nonlinear near the threshold. As the bias current increases the modulation product becomes constant, but the value of the harmonic products decrease significantly.

Figure 2. Dynamic distortion products versus bias point

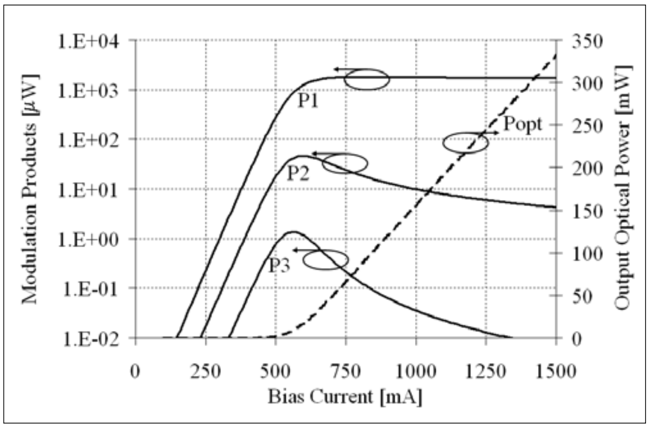
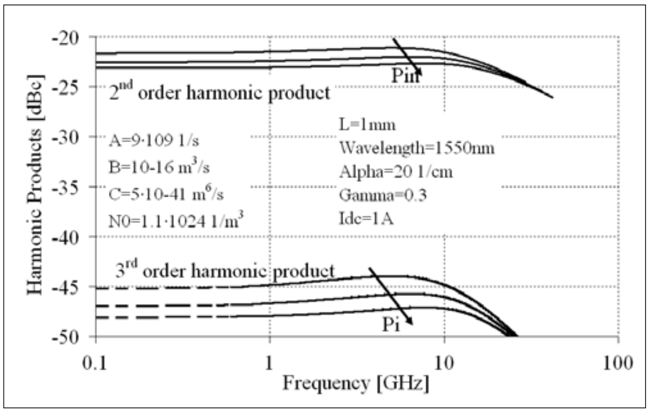


Figure 3 shows the relative second and third order harmonic distortion as a function of the modulation frequency for various input optical powers. The input optical power will not affect the relative value of harmonic products when the level of the input optical power is very low. The nonlinearity can be improved when the input optical power increases, because of the saturation effect *(Figure 4)*. The simulation results show that the nonlinearity can be improved, but the modulation efficiency decreases in the saturation regime.

Figure 3. Second and third order harmonics



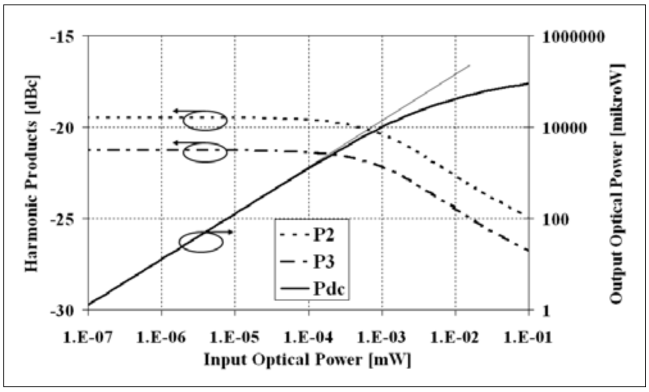
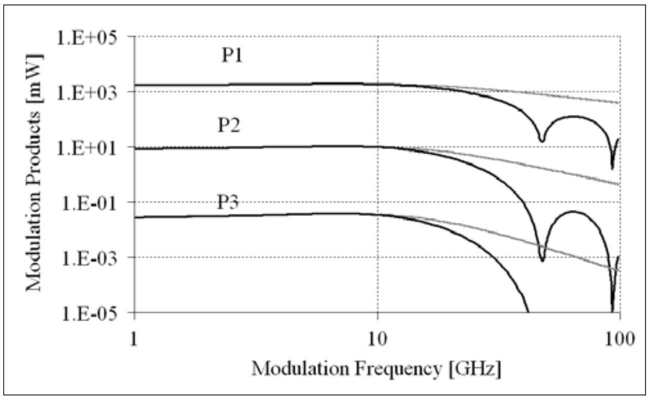


Figure 4. Saturation effect

The previous model assumed that the velocity of the traveling microwave signal was matched exactly with that of the optical signal. The next model applies to a more realistic situation where the current modulation propagates with a speed different from the optical signal. The phase velocity of the microwave is in the range of 7-12% of the velocity of light in vacuum for frequencies in the range of 5-40 GHz [10].

Thus the phase index for the microwave propagation on the electrode (n_μ) is in the range of 14.3-8.3. *Figure 5* shows a calculation for harmonic products in case of the typical co-propagating effect (n_μ=10) compared with the matched situation. The mismatch leads to dips in the modulation response and reduces the modulation bandwidth, but the bandwidth remains in the range of 10 GHz because of the SOA's rapid response time.

*Figure 5.
Mismatch between the microwave and
the light propagation velocities*



3.2. Experimental results

In the two-tone inter-modulation experiments the SOA was biased and modulated by the sum of two microwave signals. The output noise (P_{noise}) and signal levels were measured for the fundamental (P1), the second (P2) and the third (P3) order mixing products. For characterizing the level of third order nonlinearity, the third order intercept point, IP3, or the spurious suppression in dBc is used. When the nonlinearity is investigated together with noise, the figure of merit is the spurious free dynamic range, SFDR. The determination of SFDR, IP2 and IP3 are presented in (2) and *Figure 6*.

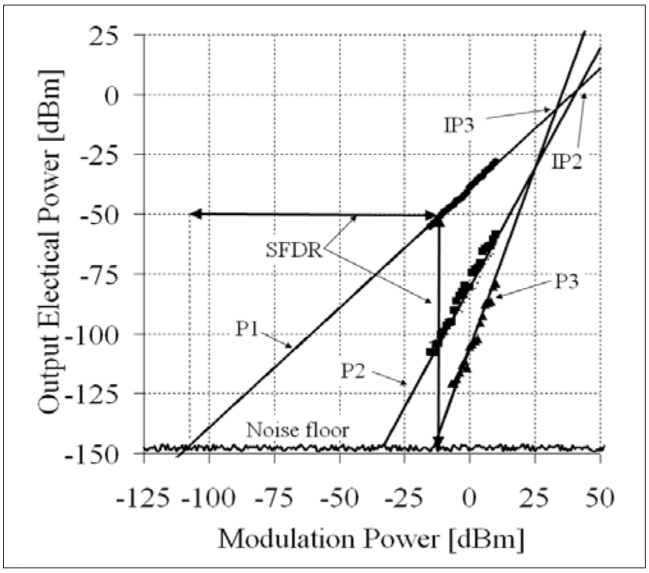


Figure 6. Determination of SFDR, IP2, IP3

$$SFDR = \frac{P_{in}(P_3 = noise)}{P_{in}(P_1 = noise)} = \frac{P_1(P_3 = noise)}{P_{noise}}$$

$$IP2[dBm] = 2 \cdot P_1[dBm] - P_2[dBm] \tag{2}$$

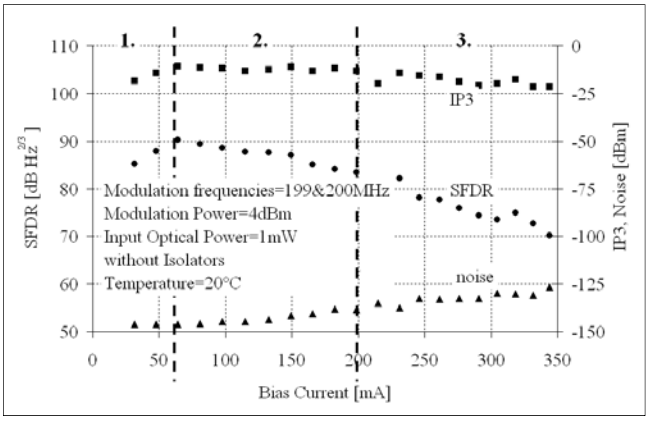
$$IP3[dBm] = \frac{1}{2} \cdot \left(3 \cdot P_1[dBm] - P_3[dBm]\right)$$

$$SFDR[dB] = \frac{2}{3} \cdot \left(IP3[dBm] - P_{noise}[dBm]\right)$$

In the linear regime the SOA modulator shows low, not measurable nonlinearity because the noise generated by the SOA will dominate in the system. The intermodulation products overcome the noise floor in case of high modulation indices. The device ensures efficient SFDR for the general optical networks (>90 dB).

Figure 7 shows the noise level, IP3 and SFDR versus SOA working state. The results show that in the first part of the graph the device is strongly nonlinear. The IP3 and the SFDR improve versus the bias current. In the second part the modulation and inter-modulation products do not change significantly but the noise level rises, hence the SFDR decreases. Finally, the inter-modulation products also start rising and the degradation of the SFDR is faster.

Figure 7. Nonlinear behavior of SOA modulator



The nonlinear behavior is also temperature sensitive, because the operation of semiconductor devices depends on the temperature. *Figure 8* shows the SFDR and the IP3 versus temperature. From the measurement results, it is clear that the linearity decreases when the temperature increases, hence temperature control is needed.

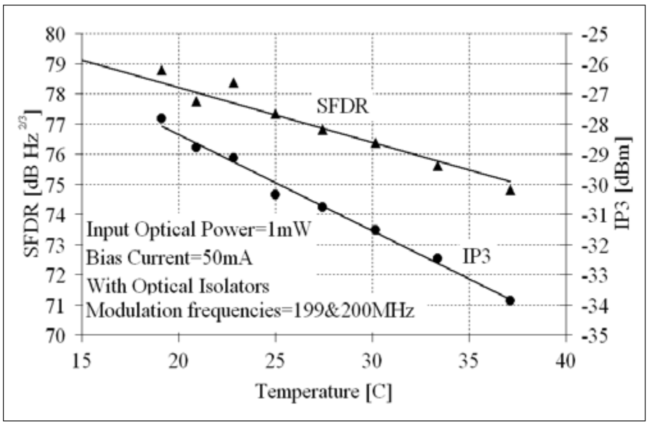


Figure 8. Nonlinearity depends on the Temperature

The noise effect and the nonlinear distortion products are more significant in case of strong optical reflection level, i.e. without optical isolators. The system will be more instable in case of strong optical reflection, and larger SFDR degradation can be observed as seen in *Figure 9 (on the next page).* The change of the SFDR is caused by two different effects. First the noise level of the device increases as a function of the bias point, the degradation is more significant without optical isolator *(Figure 10).* On the other hand the level of the nonlinear product will fluctuate in case of strong optical reflection *(Figure 11).*

4. Chirp investigation

Frequency chirping, that is the change in the instantaneous frequency of the optical signal, is produced by semiconductor devices under pulsed or modulated operating conditions

In the direct modulation of a semiconductor laser, the frequency chirping is caused by the refractive index change of the active layer due to the carrier density modulation. The change of the optical cavity modifies the frequency of the generated optical signal. In case of SOA-modulator the fluctuation of the bias current modifies the value of the carrier density (and the refractive index) and changes the transmission speed. Therefore, it causes phase variation of transmitted light through the modulator together with intensity modulation.

The refractive index can be modeled using the chirp parameter (Linewidth Enhancement Factor = LEF = Henry factor = α factor) approximation. The LEF was originally defined as the ratio of the changes of the real to the imaginary part of the material refractive index [11].

In case of small signal modulation, assuming that the carrier density change (ΔN) is uniform in SOA, for a pure

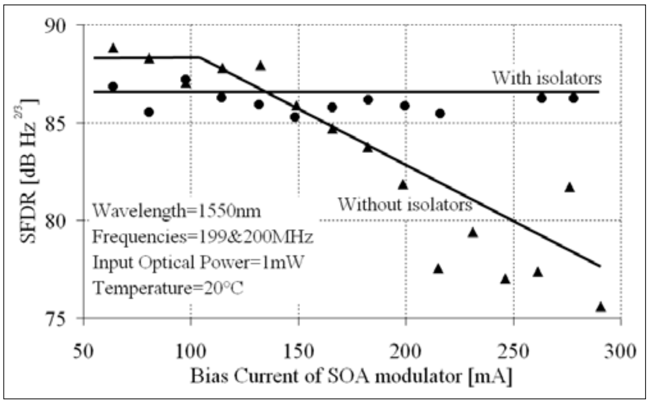


Figure 9. SFDR depends on the optical reflection

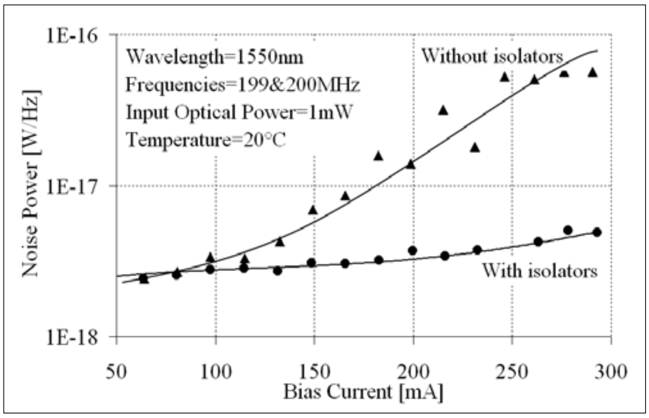


Figure 10. Noise level depends on the optical reflection

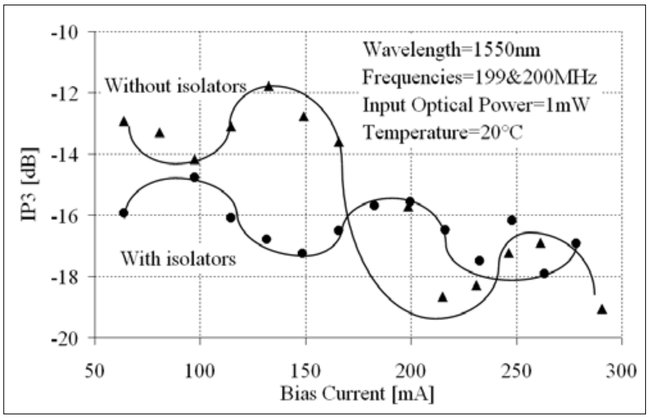


Figure 11. Nonlinearity depends on the optical reflection

traveling-wave amplifier (the facet reflectivity is ignored) the relative AM response becomes independent of LEF, the PM response becomes proportional to LEF, and the ratio of PM to AM reduces to LEF /2 [12].

$$AM = \frac{\Delta G}{G} = \frac{dg}{dN} \cdot L \cdot \Delta N$$

$$PM = \Delta\Phi = -\frac{dk}{dN} \cdot L \cdot \Delta N = \frac{LEF}{2} \cdot \frac{dg}{dN} \cdot L \cdot \Delta N \qquad (3)$$

$$LEF = -2 \cdot \frac{2 \cdot \pi}{\lambda_{in}} \cdot \left(\frac{dn}{dN}\right) \Bigg/ \left(\frac{dg}{dN}\right)$$

where G, ΔG, $\Delta\Phi$, L, g, N, n, λ_{in} and k are the optical gain, the perturbation of the optical gain, the perturbation of the output phase, the length, the material gain per unit, the carrier density, the refractive index, the wavelength of the input optical signal and the wave number, respectively.

Measurements of LEF can be found in the literature and show that LEF is not a constant factor but it is for instance a function of bias current, wavelength and input optical power. To obtain the total phase variation of the beam in a long SOA, we have to take the longitudinal variation of LEF into account. To solve it we can divide the active region into a large number of short sections. It means a quasi ideal situation: constant carrier density along the active region of the section length. The total amplitude and phase modulation can be calculated from the modulation product of the sections.

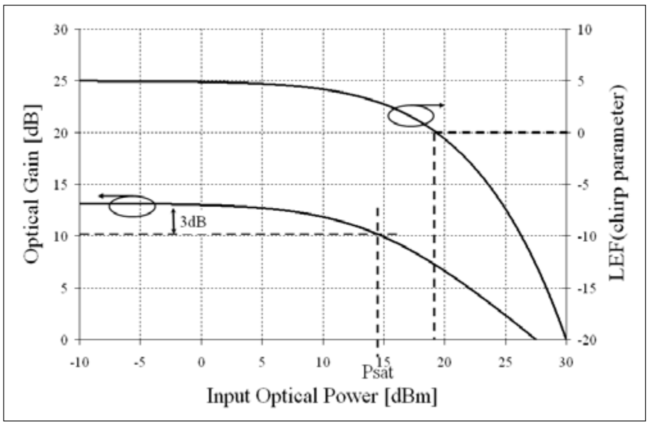
In the unsaturated region the LEF value ranges from 2 to 7 for GaAs and GaInAsP conventional lasers and from 1.5 to 2 for quantum well lasers [13]. However, as the optical input power increases, carrier depletion occurs in SOA and this induces gain saturation. In optical amplifiers under saturation conditions, an increasing input intensity causes a decrease in the amplifier gain ($dG/dP_{in}<0$). In this case LEF can be calculated from the unsaturated *LEF* value *(LEF*$_{unsat}$*)*: (4)

$$LEF = LEF_{unsat} \cdot \frac{dG}{dP_{out}} = LEF_{unsat} \cdot \frac{dG / dP_{in}}{1 + (dP_{out} / dP_{in})}$$

Due to this reason, the chirping parameter which is positive for light sources and unsaturated optical amplifiers is negative for saturated amplifiers [14].

Figure 12 represents the optical gain saturation and the LEF dependence on the optical power. When the input power becomes larger than the saturation value, the chirp parameter of the SOA rapidly falls to a negative value.

Figure 12.
Optical gain saturation and the calculated chirp parameter



The amplitude and phase modulation indices are presented in *Figure 13*. Based on these results the modulation of the laser amplifier can also be made in such a way that, the PM response is suppressed. Beside frequency modulation, this method does also reduce the amplitude of intensity modulation of the SOA. Thus, near-pure AM can be obtained.

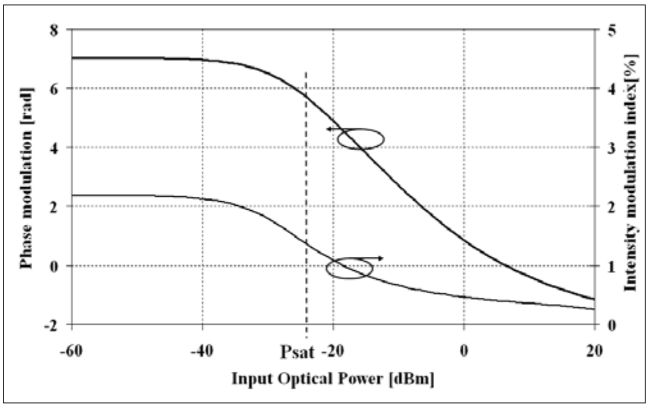


Figure 13.
Phase modulation and intensity modulation indices

5. Conclusion

The numerical simulation and experimental results show that SOA provides acceptable nonlinear distortion and the frequency chirp can be eliminated.

Applying the nonlinear carrier recombination rate, the simulation describes the frequency dependence of the modulation and the harmonic products, and the effects of the bias current and the input optical power. The model can take the mismatch between the light and the electrical signal into consideration, then the modulation bandwidth decreases. From the measurements it is clear that the dynamic range is temperature and optical reflection sensitive.

The modulation efficiency decreases and the nonlinearity can be improved when the input optical power increases, because of the saturation effect. The unwanted phase modulation decreases, because the line enhancement factor falls to the negative value.

The optimal operating point must be selected cautiously. The SOA is efficiently used as an external modulator in optical SCM systems.

References

[1] J. Mork, A. Mecozzi, G. Eisentein,
The modulation response of
a Semiconductor Laser Amplifier,
IEEE J. on Selected Topics in Quantum Electronics,
Vol. 5., No.3, May/June 1999, pp.851–860.

[2] M. J. Connelly,
Wideband Semiconductor Optical Amplifier Steady-State Numerical Model,
IEEE Journal of Quantum Electronics,
March 2001, pp.439–447.

[3] Gerhátné Udvary Eszter,
Félvezető optikai erősítők alkalmazása segédvivős optikai hálózatokban,
PKI Napok, 2004. november 29-30., pp.173–184.

[4] E. Udvary, T. Berceli,
Branching Function by Semiconductor Optical Amplifier in Subcarrier Multiplexed Optical Systems,
MICROCOLL'03, Budapest, 10-11. September 2003.

[5] J. Prat, C. Arellano, V. Polo, C. Bock,
Optical Network Unit Based on a Bidirectional Reflective Semiconductor Optical Amplifier for Fiber-to-the-Home Networks,
IEEE Photonics Technology Letters,
Vol. 17., No.1, 2005., pp.250–252.

[6] E. Udvary, T. Berceli,
Optical subcarrier label swapping by semiconductor optical amplifiers,
Journal of Lightwave Technology,
Vol. 21., No.12, 2003., pp.3221–3225.

[7] T. Marozsák, E. Udvary, T. Berceli,
Transmission Characteristics of All Semiconductor Fiber OpticLinks Carrying Microwave Channels,
30th EuMC, Paris, France, 3-5 October 2000,
Vol. 2., pp.52–55.

[8] M. A. Ali, G. Metivier,
Performance Analysis of Multichannel 16/64-QAM CATV Distribution Network Using Semiconductor Optical Amplifier,
IEEE Photon. Tech. Letters, May 1997, pp.690–692.

[9] R. Olshansky, V. Lanzisera, P. Hill,
Subcarrier Multiplexed Lightwave Systems for Broadband Distribution,
Journal of Lightwave Techn., 1989, pp.1329–1342.

[10] D. Tauber.et al.,
Distributed Microwave Effects
in High Speed Semiconductor Lasers,
IEEE MTT-S Intern. Microwave Symposium Digest,
1994, pp.49–51.

[11] L. Occhi, L. Schares, G. Guekos,
Phase Modeling Based on the α Factor
in Bulk Semiconductor Optical Amplifiers,
IEEE J. of Selected Topics in Quantum Electronics,
2003, pp.788–797.

[12] L. Gillner,
Modulation properties of a near travelling-wave semiconductor laser amplifier,
Optoelectrics, October 1992, pp.331–338.

[13] F. Koyama, K. Iga,
Frequency chirping in external modulators,
IEEE J. of Lightwave Technology,
January 1988.

[14] T. Watanabe et al.,
Transmission Performance of Chirp-Controlled Signal by Using Semiconductor Optical Amplifier,
IEEE Journal of Lightwave Technology,
August 2000, pp.1069–1077.

A novel vertical handover mechanism for media streaming
in heterogeneous wireless architectures

LÁSZLÓ BOKOR, LÁSZLÓ LOIS, CSABA A. SZABÓ, SÁNDOR SZABÓ

Budapest University of Technology and Economics, Department of Telecommunications
{bokorl, lois, szabo, szabos}@hit.bme.hu

Keywords: heterogeneous wireless environment, streaming media delivery, WiFi, 2.5G and 3G mobile

The paper presents a novel vertical handover mechanism which aims at assuring streaming media services in a heterogeneous network environment where the subscribers are roaming among different wired/wireless access systems including ADSL, WiFi, 2.5G and 3G cellular and WiMAX. The handover scheme provides seamless connectivity during roaming, with adapting the quality of the delivered media stream to the changes of the network characteristics and to the capabilities of a wide variety of devices. The paper presents general design considerations, focuses on introducing the operational behaviour of the novel vertical handover method, its implementation and evaluation in a heterogeneous access network testbed, as well as discusses some aspects of gaining higher level features based on our proposal.*

1. Introduction

The key issue of any handover management mechanism designed for heterogeneous architectures is the efficient management of all kind of transitions between access networks, called briefly mobility (we are using this term in a broader sense, denoting movement from one access network to another, regardless of the velocity of the user, thus including real or nomadic mobility, etc.). Traditional mobility management is hedged in providing *terminal mobility*. This kind of mobility allows a mobile node to maintain ongoing communication or commence/receive incoming session requests independently from its network point of attachment. (Note that there is a subset of terminal mobility which ensures handling only new sessions after changes of networks. Providing this subset requires only dynamic DNS and DHCP functions.) The development, deployment and convergence of different wired and wireless technologies introduced several new mobility types which can be grouped into two main categories. On one hand, there is a device-centric, low-level mobility including *ad hoc mobility* (mobile devices are routable in ad hoc networks) and *mode mobility* (devices can switch between ad hoc and infrastructure modes). On the other hand, we can talk about user-centric, high-level mobility consisting of *personal mobility* (users are globally reachable at different scenarios: one address for many different devices or many addresses reaching one device), *session mobility* (active sessions are switchable between terminals) and *service mobility* (personalized services can be maintained while moving and/or changing devices/ISPs).

Supporting mobility between different types of access networks (e.g. UMTS to WLAN) is called “vertical handover” in order to distinguish it from the usual “horizontal handover” (the migration of mobile nodes between homogeneous networks, e.g. UMTS to UMTS) often occurring in a mobile operator’s network whenever a user leaves the radio cell of a base station and enters a neighboring cell [15]. (Note that the difference between the terminology of horizontal and vertical handover is vague. For example, a handover from an 802.11b WLAN AP to an 802.11g AP link may be considered as either a vertical or a horizontal handover, depending on the point of view.)

In a media streaming delivery architecture, built on a heterogeneous architecture, managing vertical handovers during mobility is necessary for three main reasons. The first is obvious: while moving, the user approaches the boundary of the coverage area of the actual network or part of the network (e.g. a radio cell) so that the bit error rate, packet loss or any other QoS parameter becomes too high and an *automatic* (unplanned) vertical handover is to be performed. The second reason is when the user wants to change the current manner of network attachment intentionally because there is a better way available to support better QoS parameters for the media (consequently, often for higher price). This latter case is called *user-initiated* vertical handover. Third, when a load balancing mechanism of an overlay network is able to distribute the overall network load in order to optimize the performance of each individual network. This is called *network management-initiated* vertical handover which allows more efficient resource utilization for service providers.

The main novelty of our handover management approach lies in transparently and effectively supporting the close incorporation and tight integration of advanced

** Based on “Testbed of a Novel Media Streaming Architecture for Heterogeneous Wireless Environment”, by László Bokor, László Lois, Csaba A. Szabó, Sándor Szabó, published in the Proceedings of the 3rd International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TridentCom2007), Orlando, Florida, USA.*

vertical mobility management, multimedia processing and transmission, quality of service, adaptation to different network and device capabilities, digital rights management and flexible quality based billing. Using our vertical handover mechanism, all of these features and enhancements can easily be merged into a single integrated and transparent architecture by granting the capability of moving across a wide range of loosely-coupled access networks.

The paper presents an overview of the design, implementation and evaluation of this novel handover mechanism for media streaming in heterogeneous wireless architectures providing the features introduced above. The paper is organized as follows. Our main design choices are described in Section 2. In Section 3, the mechanisms to implement and test vertical handover management across heterogeneous access networks are described and test results are presented and evaluated. Section 4 deals with the integration possibilities of different features into a single media streaming architecture based on our vertical handover scheme. The paper concludes with a summary in Section 5.

2. Conceptual overview

The main requirements towards a mobility handling mechanism supporting vertical handovers for media streaming in heterogeneous architectures are as follows:

a) Efficient location management.

The mobile device should always be accessible by a static identifier regardless of its current location. The following challenges should be overcome:

– Reduction of signaling overhead and latency
– Guaranteed QoS in different access networks
– Intelligent decision algorithms controlling the MN's behavior in overlapping areas of heterogeneous wireless networks

b) Handling wide variety of mobility.

The mobility management solution should allow as many mobility types (terminal, personal, etc.) as possible.

c) Transparent and seamless handovers.

The change between different networks should not cause considerable data loss, the transition itself should not last long and the long-term connection-oriented protocols should run in a seamless way. The following challenges are to be solved:

– Reduction of signaling overhead and latency
– Guaranteed QoS during the handover procedures
– Scalability, efficient utilization of resources, reliability, robustness

d) Infrastructure-less solution.

The mobility management solution should be located at the boundaries of the network so that no or only minor changes are required in the service providers' networks.

We have evaluated all of these requirements in [2], in order to choose an appropriate handover management approach fitting into the proposed streaming media architecture. We have assessed our criteria and compared the parameters/attributes of mobility management methods at different layers of the TCP/IP architecture. We have concluded, that network and application layer mobility (Mobile IPv6 [13] and Session Initiation Protocol [6]) are the most promising approaches for our purposes. Considering these two methods we have pointed out that in case of terminal mobility Mobile IPv6 is the most general and effective solution but requires more complex infrastructure than SIP-based mobility.

Personal mobility cannot be achieved using only network layer methods; however SIP forking proxies in the application layer easily bypass this issue. Session mobility is not supported by Mobile IPv6 (however IPv6 anycasting could have the potential to provide this feature [3]) in contrast to SIP, which allows session mobility by explicitly transferring a session to another destination. Service mobility is achievable based on Mobile IPv6 (by using subsidiary components to keep service definitions updated) but SIP offers built-in mechanism for synchronizing service definitions and other configuration elements.

Besides that SIP is able to provide terminal, personal, session and service mobility, it supports the widest range of applications from VoIP, Internet conferencing and presence to instant messaging and event notification. SIP was originally proposed by IETF to establish, modify and release sessions in all-IP networks and then gained support by 3GPP to perform signaling tasks in IP Multimedia Subsystem (IMS) [7,8].

Now SIP is the basic protocol of the 3GPP IMS, and IMS is also incorporated in ETSI's NGN architecture. The SIP-based IMS defines an overlay architecture on the top of any 3G packet switched core network comprising the key technologies of the future's converged, service and application oriented networks [9,10].

With regard to the current trends in telecommunications, it is obvious that all players of the future mobile communication market need an easy-to-use instrument for quick integration and examination of new services and applications even from third parties. IMS seems to be that generic instrument and that was our most important reason to choose SIP and the application layer-located approach as the basis of our handover management proposal, despite the fact that an integrated and/or hybrid MIPv6-SIP architecture could combine the benefits of each protocol [4,5].

3. The proposed handover management mechanism

A. Components and basic operation

To provide a generic solution, the proposed mechanism separates media functions and mobility functions by segregating the following operational components *(Figure 1):*

Media Player, SIP Streaming Client, Media Server, SIP Streaming Server, SIP Server.

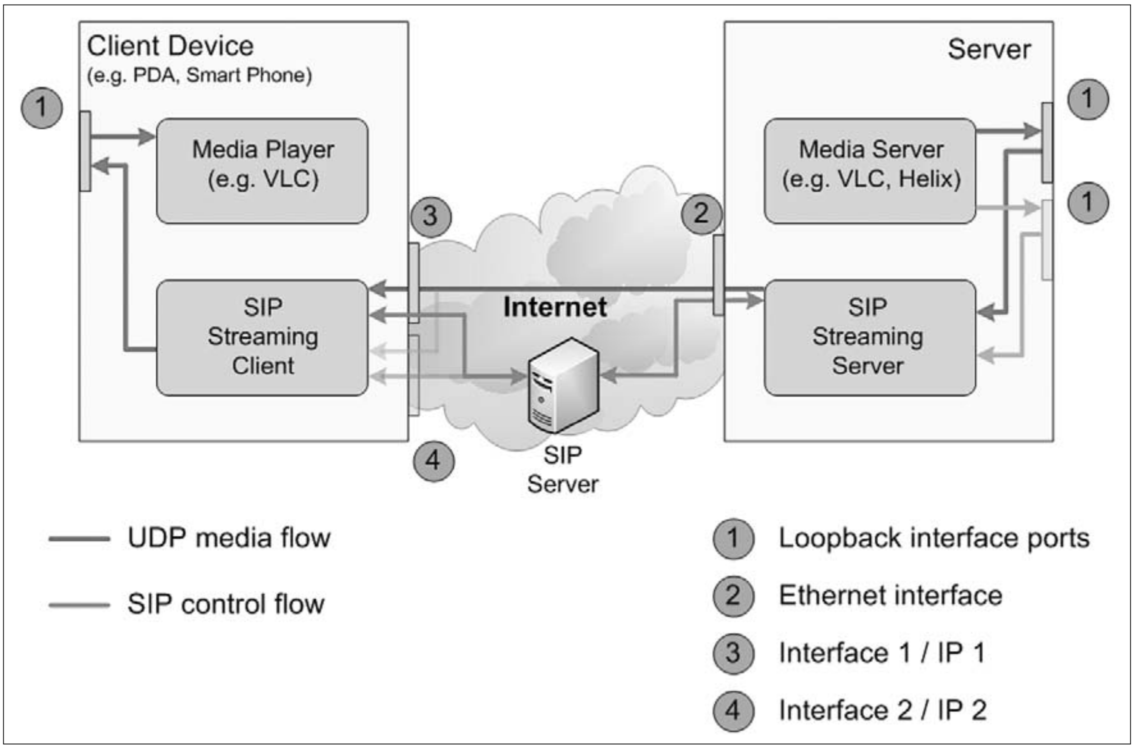


Figure 1.

SIP-based streaming architecture for heterogeneous environment

Any type of media server (e.g. HELIX, VLC) and player (e.g. VLC) can be used, which is capable of handling UDP/RTP network streams. The only task of the media server-player modules is to create and play back the media content. Any other function is completely independent from them, making their functions transparent and the implementation portable and open. The separate module-pair of SIP Streaming Client and Server is responsible for transporting the UDP/RTP media stream across various networks, by utilizing a SIP based user authentication and mobility management protocol. The SIP Server handles the standard functions of controlling the SIP-based communication.

The basic idea is the following: the mobile terminal (Client Device) registers its current IP address on an access network, and after successful authentication and registration, the SIP Streaming Server forwards the stream of the Media Server to the mobile terminal's registered IP address. The SIP Streaming Client captures the UDP/RTP media stream, and forwards it to a local UDP port on the mobile terminal. The Media Player – which should be able to accept streams from a defined UDP port – connects to this local port, and plays back the content. Upon handover, the mobile terminal registers its new IP address (using the functions of SIP Streaming Client) to the SIP Streaming Server, and the server transfers the media stream to the new address. The media player running on the client device is not aware of the handover event.

Notice the fact of separating the UDP/RTP based media and SIP signaling messages, according to NGN and IMS concepts [11,12]. We have to highlight that one of the most important advantages of utilizing SIP protocol in the proposed manner is the ease of integration with IMS system, as an AS (Application Server).

B. SIP streaming client and server

The main component of the described handover management method is the SIP Streaming Client. This module integrates the following functions *(Figure 2)*:

- *Forwarder:* Transmits the media stream from a given interface to the loopback port where the Media Player can reach and play back.
- *SIP Communicator:* Handles the SIP communication by managing SIP requests and responses.
- *Interface:* Monitors active interfaces/networks (listed in the user's preference file given by the GUI), and measures the QoS of present connections based on packet loss and round-trip time. Every interface owns a state machine describing its actual status (see Figure 7).
- *Connection Manager:* Selects the active connection based on the QoS measurements made on the interfaces, and upon a given threshold, initiates the handover and all concerning SIP procedures automatically. In addition to the architectural support, an efficient decision algorithm is also implemented for fast selection of connection to perform seamless, or nearly seamless vertical handover. (Note that in our testbed the threshold values can be predefined or dynamically calculated by the decision algorithm.)
- *Utility:* A little application which allows NAT traversal.
- *Graphical User Interface:* The Graphical User Interface allows the users to define the precedence of the potential network interfaces, presents information about the current connection and active interfaces, and helps to execute user-initiated handovers.

The main task of the decision algorithm operating in the SIP Streaming Client is to pick out the most appropriate connection (interface) from the user-designated

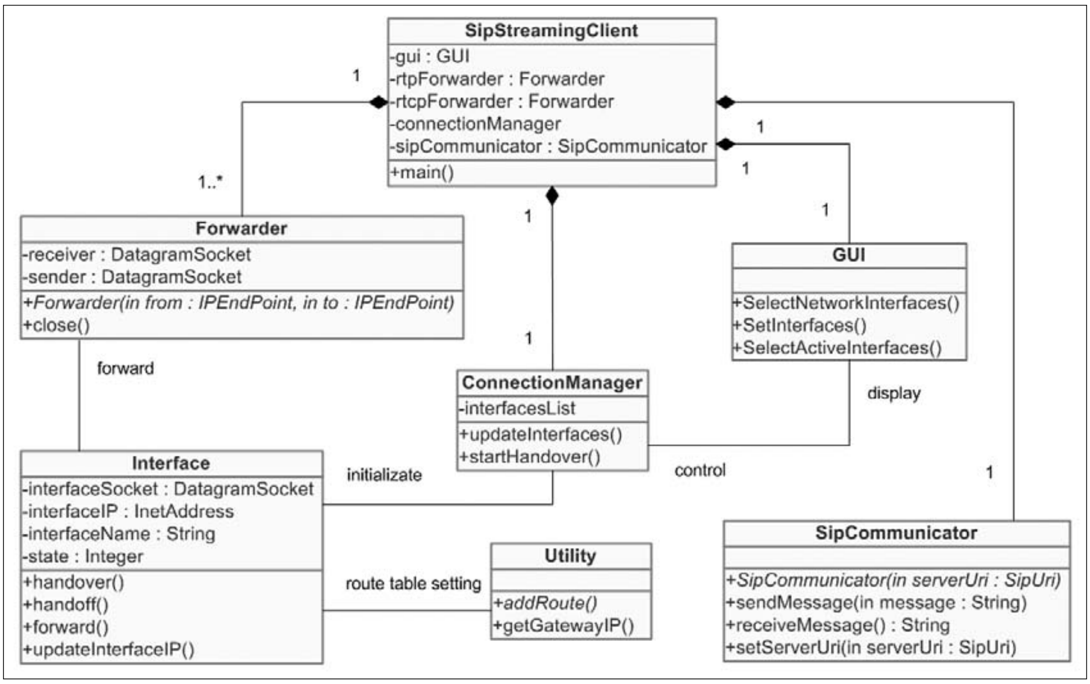

SipStreamingClient
-gui : GUI
-rtpForwarder : Forwarder
-rtcpForwarder : Forwarder
-connectionManager
-sipCommunicator : SipCommunicator
+main()
1
1..*
1
1
1
1
1
1
Forwarder
-receiver : DatagramSocket
-sender : DatagramSocket
+Forwarder(in from : IPEndPoint, in to : IPEndPoint)
+close()
GUI
+SelectNetworkInterfaces()
+SetInterfaces()
+SelectActiveInterfaces()
forward
display
ConnectionManager
-interfacesList
+updateInterfaces()
+startHandover()
Interface
-interfaceSocket : DatagramSocket
-interfaceIP : InetAddress
-interfaceName : String
-state : Integer
+handover()
+handoff()
+forward()
+updateInterfaceIP()
initialize
control
1
SipCommunicator
+SipCommunicator(in serverUri : SipUri)
+sendMessage(in message : String)
+receiveMessage() : String
+setServerUri(in serverUri : SipUri)
Utility
+addRoute()
+getGatewayIP()
route table setting

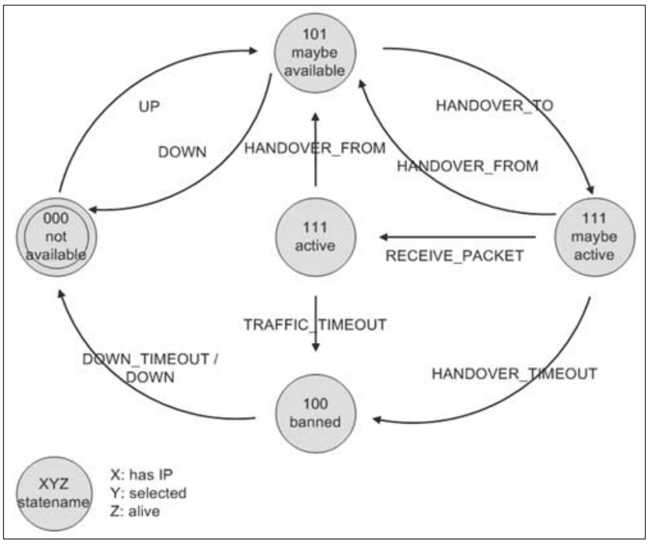
Figure 2.

UML class diagram of SIP Streaming Client

ones. In order to achieve this function, the decision engine continuously gets and evaluates jitter and packet loss measurement results (referenced among the SIP Streaming Server and the active interfaces), and continuously calculates and updates the threshold value of beginning a transition. At this phase no physical layer information is used as an input parameter for the algorithm: the decision is based only on IP level attributes, calculated/predefined threshold values, different states of interfaces and on user-preferences. This enables simpler implementation and build-up while keeping the effective operation and adequate performance as well.

Because the need of a handover transition doesn't depend only on calculated or predefined threshold values but on different states of interfaces and on user-preferences as well, a state machine was defined to maintain the altering conditions of every interface and to control the process *(Figure 3)*.

Figure 3. State machine of an interface



The SIP Streaming Server is responsible for managing the multimedia sessions by relaying between a media player and server and for transmitting the media streams always to the actual addresses of the clients by parsing incoming SIP signaling messages and forwarding the media content from the local port towards the mobile node's actual location.

Compared to the SIP Streaming Client, this software module does not contain any special intelligence: mainly it is a simple SIP interpreter driven by SIP commands in order to achieve a dynamic, transparent and media server independent source of media content.

C. Experimental results

To evaluate our vertical handover method designed for media streaming in heterogeneous environment, we have set up an experimental testbed consisting of a loosely-coupled UMTS-WLAN-LAN architecture *(Figure 4)*. This experimental testbed combines several independent IP-based networks, including T-Mobile Hungary's UMTS architecture, IEEE 802.11a/b/g WLAN network, and BME-HT's* LAN. All of these technologies are integrated by a common, IP-based operation, however below IP each access network has its own protocol stack, differing characteristics and attributes.

The 3G/HSDPA cellular UMTS network and all related infrastructure are organic part of T-Mobile Hungary's production network. This component enables services at maximum 1.5 Mbps download and 384 Kbps upload speeds with RTTs around 80 ms. The WLAN connectivity is a separate sub-network of our testbed. There are IEEE 802.11a/b/g compatible Linksys WAP55AG Access Points interconnected. This exclusive, local wireless access enables Mobile Nodes to transmit/receive data up to 54 Mbps with RTTs around 10 ms.

** Budapest University of Technology and Economics, Department of Telecommunications*

The LAN component is an IEEE 802.3 compatible Ethernet connection established at BME-HT. This connectivity is used by cable link enabling high-speed wired networking up to 100 Mbps with negligible RTTs (around 1 ms).

For accessing this heterogeneous network setup, Mobile Nodes have been equipped with separate interfaces for every access component, and with a Java implementation of our SIP Streaming Client software. A MN's hardware is based on Fujitsu-Siemens Lifebook C1320D (1.73 GHz Pentium M processor, 256 MB RAM) with BroadCom NetXtreme Gigabit Ethernet interface, Edimax ZD1201 USB WLAN adapter, and Globetrotter 3G+ PCCard Modem for UMTS connectivity.

The testbed's Media Server and SIP Server is a standard PC with a 3 GHz Pentium 4 processor and 512 MB RAM, running a Java implementation of our SIP Streaming Server and SIP Proxy/Registrar.

Based on this heterogeneous testbed, experimental surveys can be performed in order to observe Mobile Nodes while they perform seamless roaming among different access technologies, and maintain ongoing media connections. The first experiments were focused on vertical handover management, the results obtained in these scenarios using the basic testbed topology are presented in *Figures 5, 6, 7 and 8*.

The amount of time required by the handover among different network types, and the resulting packet loss were measured to present the performance characteristics of our vertical handover-enabled media streaming architecture. We gathered measurement data using preset values based and decision algorithm aided connection detection, and performed several trials in the same operating conditions for three different classes of media traffic (48, 192, 385 Kbps) in three different upward vertical handover scenarios (LAN to WLAN, LAN to UMTS, WLAN to UMTS). The values in the tables represent average values obtained on the mobile node (moving client of a streaming application) across five repetitions of each test. The handovers were initiated by simulated total link failure (e.g. disconnection of the physical medium); in more realistic situations even better results are expected.

In case of user initiated handover the system provides no packet loss during the transition (soft-handoff), thus no visible deterioration of the media stream can be observed unlike in case of automatic (unplanned) handovers where the packet loss of an ongoing media stream depends on three main factors:

- Bit rate of ongoing media traffic;
- Timeout of detecting the network failure on the current link;
- RTT of the new network (i.e. delay of SIP signaling messages managing the handover procedures).

The comparison of packet loss measurement results presented in *Figure 5 and 6* shows that the decision algorithm aided vertical handover outperforms the preset values based method by dynamically and effectively adapting the SIP Streaming Client to the continuously varying network conditions. Results of vertical handover latency measurements in *Figure 7 and 8* confirm the

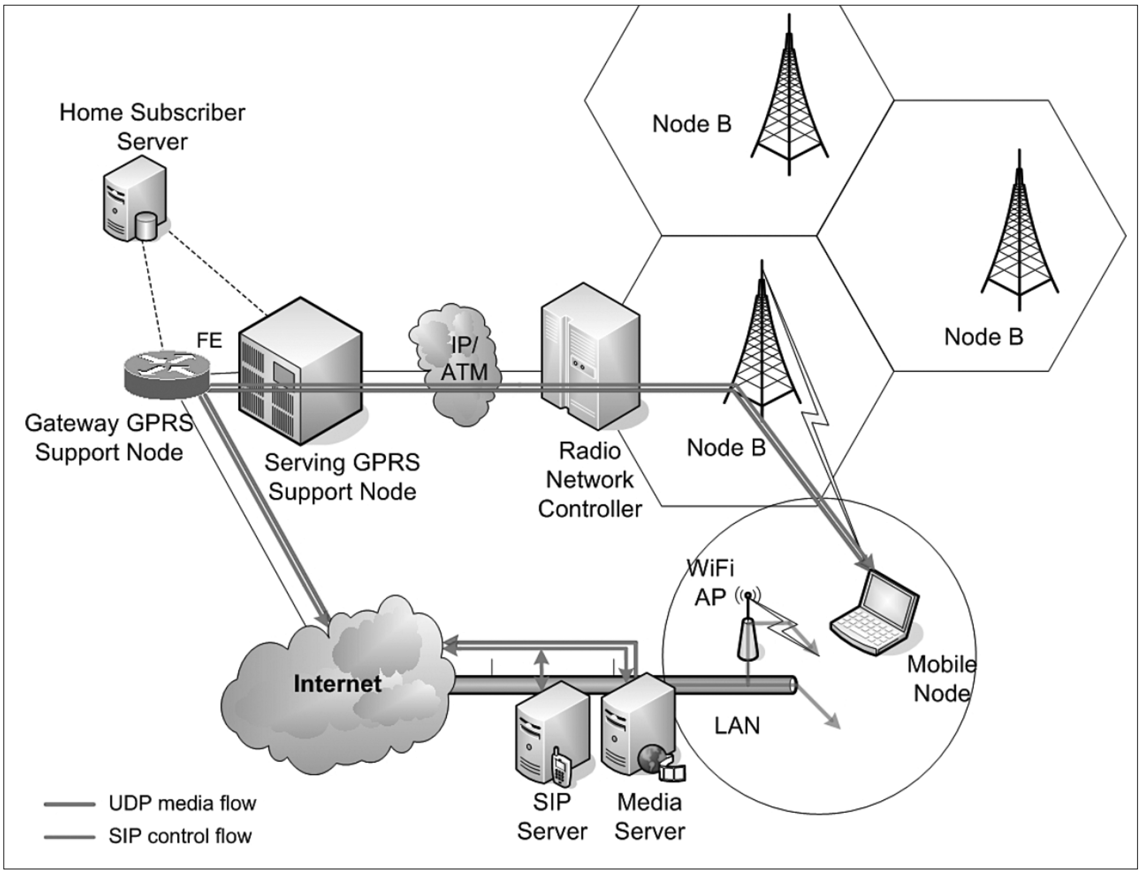
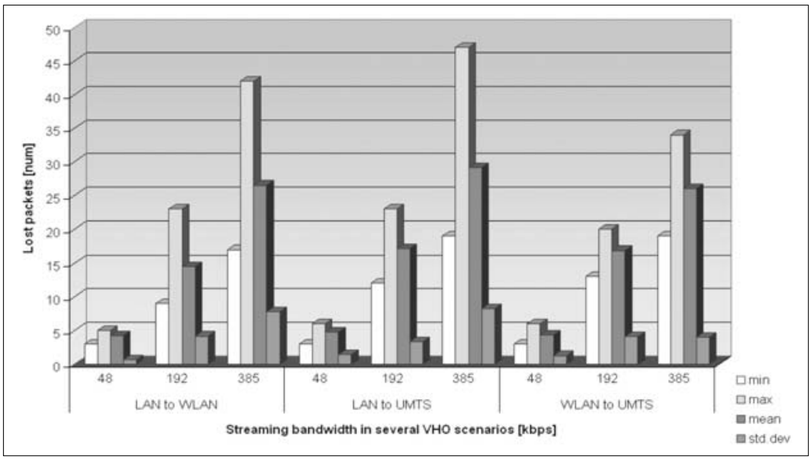


Figure 4.

Basic topology for the measurements



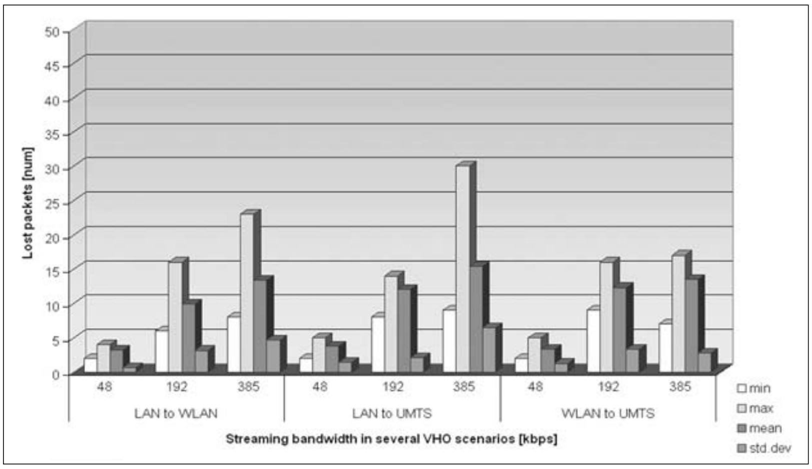


Figure 5.
Number of packets lost
(Preset values based connection detection)

Figure 6.
Number of packets lost
(Decision algorithm aided connection detection)

above statement and show that the VHO latency will be less than 2.5 s in all the examined scenarios even without using adaptive clients based on decision algorithm aided connection detection.

4. Higher level features

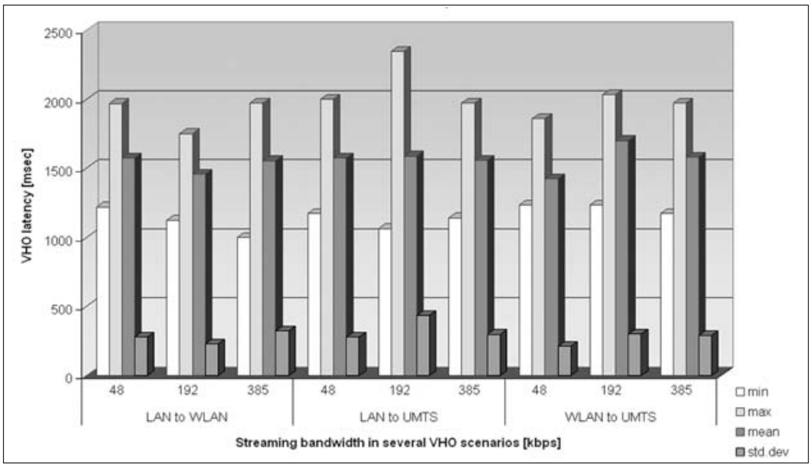
In a multi-platform access network environment, the user has several physical connections to access the Internet, hence the same resource could be even accessed simultaneously via different wireless networks. Based on a transparent and effective vertical handover mechanism, this opportunity could be utilized to achieve higher quality service, i.e. faster download or higher quality media streaming solution by using all access networks simultaneously or selecting the best access network(s) dynamically. The main components of such an integrated media streaming architecture for heterogeneous environment are the following:

- Media delivery subsystem
- Network access and handover module
- Bit-rate and resolution selection and ranking module
- Security and accounting module

However, all of these modules should cooperate with each-other to achieve a higher level service. The cooperation of the main modules of such a media streaming system based on our transparent vertical handover mechanism is shown in *Figure 9 (on the next page)*.

The media delivery subsystem provides important information such as:

- decoded video and audio quality and buffer fullness,
- available bit-rates and resolutions of the current media content,
- aviable bit-rate switching positions.



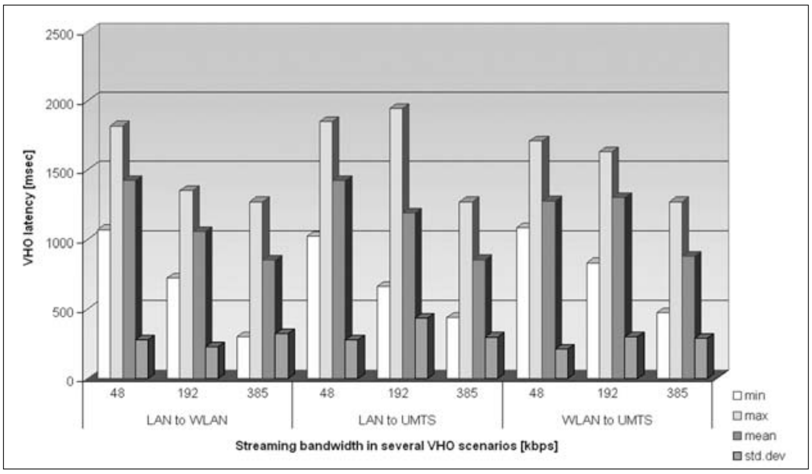


Figure 7.
VHO latency
(Preset values based connection detection)

Figure 8.
VHO latency
(Decision algorithm aided connection detection)

While the media server knows when the bit-rate switching could be performed without any transient effects in the reconstructed video, the network access module performs the switching from one access network to another in case of need. By communicating the new destination IP address, the new bit-rate and resolution to the media server and client, the media presentation becomes continuous (in exception of an abrupt network loss) when the client sends retransmission requests and uses enough long look-ahead buffer at the input of the media decoder. Since the network access units have different IP addresses, the network access and handover client forwards the media stream arriving on the active port to a fixed virtual port referring to the media client.

The different client capabilities are the input of the bitrate selection and ranking module, and the other input of this module is the current state of each access network. The client capabilities can be evaluated at startup, but the state of the access network is reported regularly by the network access and handover module. The selection of the new access network and the new bitrate is done by the bit-rate selection and ranking module, but the bit-rate could also be changed while the access network remains the same, i.e. in case of remarkable alteration of the packet loss statistics.

In a system based on our vertical handover management scheme, the decision of the stream switching can be based on the client's measurements, since the available network connections are handled by the SIP client and the server has little to do with the possible connections of every client. Based on the measured parameters (current packet loss rate and the access network type), the optimal bandwidth can be estimated and the ranking of the access networks are made. Based on the ranking, the optimal bandwidth of the best connection and the decoder properties (i.e. screen size, decoding speed), the best bandwidth/quality version of the content is determined and the switching is carried out in case of need.

The ranking of the connections can be based on several measurements or pre-defined values, such as:

- the availability of the connection, even by measuring the field strength at the receiver's front-end if possible, or upper layer parameters (e.g. packet loss rate),
- the expected available/achievable bandwidth of the connection,
- the expected or actual packet loss rate for the idle or active connections, respectively,
- the expected video and audio quality of the media streaming over the connection,
- cost of the connection.

The ranking and the selection of the best connection is re-evaluated upon:

- degradation of parameters of the active connection,
- improvement of parameters of an idle connection.

With the proper ranking method, the media streaming system can operate on the active connection close to the optimal single-connection configuration on that connection. Since our proposal handles the handovers almost seamlessly and allows the media streaming parameters to be changed to the optimal ones of a new active connection, we can achieve a sub-optimal best-effort single connection scheme.

5. Conclusions and future work

We have presented a novel, transparent, efficient and open vertical handover mechanism which can be easily integrated with media coding and delivery methods that allow for receiving media streams optimized to network and user device capabilities. Ongoing work includes real integration of these components into a single test environment.

Acknowledgment

This research was funded within the framework of a project financed by the Hungarian National Research and Technology Office under contract No. 2/027/2004. The authors would like to thank Csaba Balogh and András Kocsis for their essential work on this project. Also thanks to Péter Patkó, Zoltán B. Kirchner and other colleagues of the industrial partners in the project consortium.

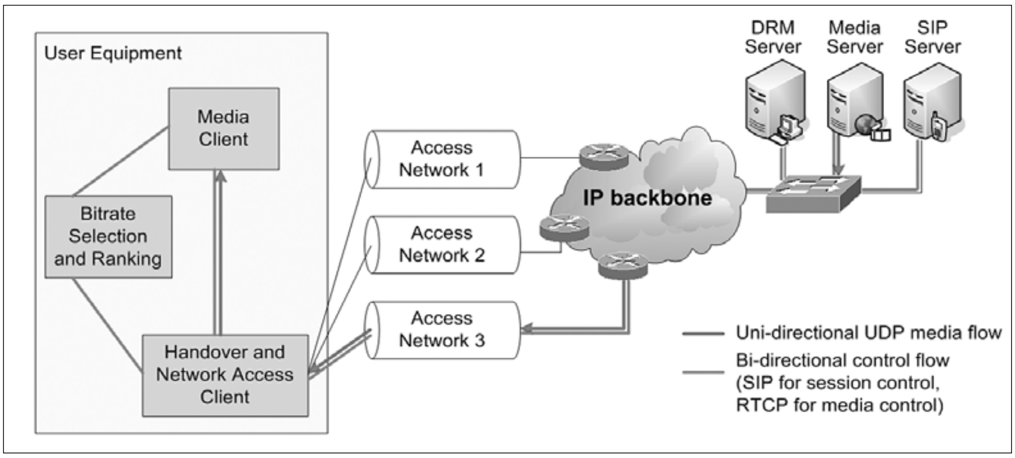

User Equipment
Media Client
Bitrate Selection and Ranking
Handover and Network Access Client
Access Network 1
Access Network 2
Access Network 3
IP backbone
DRM Server
Media Server
SIP Server
Uni-directional UDP media flow
Bi-directional control flow (SIP for session control, RTCP for media control)

Figure 9.

The interconnection of the three main components

References

[1] J. Manner, M. Kojo,
“Mobility Related Terminology”,
IETF RFC 3753, June 2004.

[2] Cs. A. Szabó, S. Szabó, L. Bokor,
“Design Considerations of a Novel Media Streaming Architecture for Heterogeneous Access Environment”,
BWAN‘06, September 2006.

[3] S. Doi, S. Ata, H. Kitamura, M. Murata,
“Design, Implementation and Evaluation of Routing Protocols for IPv6 Anycast Communication”,
AINA‘05, Taiwan, 2005.

[4] Bi-Lynn Ong, Suhaidi Hassan,
“A Survey of IPv6 Mobility Management in Real-time Communications”,
Networks‘05, v.2, p.6, November 2005.

[5] Q Wang, M. A. Abu-Rgheff,
“Integrated Mobile IP and SIP Approach for Advanced Location Management”,
3G ‘03, London, Vol. 494, pp.205–209., June 2003.

[6] J. Rosenberg, H. Schulzrinne, G. Camarillo, R. Sparks, A. Johnston, J. Peterson, M. Handley, E. Schooler,
“SIP: Session Initiation Protocol”,
IETF RFC 3261, June 2002.

[7] Gonzalo Camarillo, Miguel-Angel Garcia-Martin,
“The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds”,
John Wiley & Sons, 2004.

[8] 3GPP TS 24.229,
“Internet Protocol (IP) multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP) Stage 3, (Release 7)”, v7.6.0, December 2006.

[9] Zarri, M.,
“Future service capabilities offered by the 3GPP system”, 3G Conference 2003,
25-27. June 2003., pp.354–358.

[10] Hector Montes, Gerardo Gomez, Renaud Cuny, Jose F. Paris,
“Deployment of IP Multimedia Streaming Services in Third Generation Mobile Networks”,
IEEE Wireless Communications, Vol. 9., Nr.5, October 2002., pp.84–92.

[11] DES/TISPAN-02007-NGN-R1,
“Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); NGN Functional Architecture Release 1, Overall Architecture”, August 2005.

[12] 3GPP TS 23.228,
“3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; IP Multimedia Subsystem (IMS) Stage 2 (Release 7)”,
v7.6.0, December 2006.

[13] D. Johnson, C. Perkins, J. Arkko,
“Mobility Support in IPv6”,
IETF RFC 3775, June 2004.

16th IST Mobile and Wireless Communications Summit in Budapest

PROF. ISTVÁN FRIGYES

General Chair of the conference

The 16th "edition" of a series of annual conferences was held in Budapest, 1-5 July, 2007. (At the time of this writing, more precisely: will be held in that period.) IST Mobile Summits were initiated by the European Commission (EC). The person, who founded it 15 years ago, was Dr João da Silva, director of Converged Networks and Services of the European Commission. He has been the main promoter of this event since then. (Note: IST stands for Information Society's Technologies.)

During these years this conference grew to the greatest one in its field in Europe, both in terms of its size and scientific significance. Its main aim was and is to report on the progress of EU-sponsored projects in the field of mobile and wireless communications. However, it is by no means restricted either to these projects or to EU member countries or even to Europe. It is a usual scientific conference with invited and submitted papers, with at least 3-fold reviewing of submitted ones and with a broad professional audience.

It was a very great honor for us but also a great challenge when we, the Budapest University of Technology and Economics, Department of Broadband Communications and Electromagnetic Theory, were offered to organize the 2007 event here in Hungary. Of course, we accepted this offer, without any doubt. The conference was organized by us and co-organized by HTE. The organizers got big professional, financial and moral support by the EC in the framework of FP6 project SPECTRUM.

The number of participants was some 550, quite a large number; they came from 6 continents, 42 countries. Of course, most of them from Europe but there were somewhat more than 10% from Asia, about 15 people from North America, and a few also from Brazil (3), from Africa (2) and from Australia (1). It is also interesting to mention: Hungary was only second in the number/country of participants; the first in this list was Germany and the third the UK, preceeding Italy as fourth.

The Summit was composed of basically two different types of publications: panel sessions and papers.

Panel sessions were held in plenary (except one special panel session). Topics of panel sessions covered discussions about the foreseeable future of four very important points in communications. These were:

(i) next generation of mobile communications (called 4G, 4th Generation, in some countries and B3G, Beyond 3rd Generation, in others);
(ii) the changing role and appearance of media in the Internet age;
(iii) the future of Internet while communication is mainly mobile; and
(iv) technical, security and legal problems related to Near Field Communication (NFC). Moderators and panelists came from the most important European, Asian service and technology providers and also Americans.
Companies represented included – to list some of them only – Motorola, Intel, Huawei, NTT-DoCoMo, Samsung, Vodafone, NSN, Joost.

Among paper presentations, the keynote talk of Dr. Steve B. Weinstein is first to mention (on broadband wireless and optical-wireless communications); there were two invited special sessions: a North-American session and one of ITC in healthcare. The rest, i.e. the 300 contributed papers were organized in 30 oral sessions (60%) and 5 poster sessions (40%). Interestingly enough, the distribution (60%-40%) was the same between project-related and individual papers.

It is of some interest to look at the distribution of paper subjects. Of course, there is no one-to-one relationship between this distribution and the distribution of main problems in this segment of science – however these are not very far from each other. By far the most papers covered problems related to networks, about one third of all. It was even more typical that but-most papers dealt with applications, business models and services. In similar conferences the highest number of papers deals usually with problems of the physical layer; in this conference it was of the lowest interest – except one individual topic, i.e. that of MIMO and space-time techniques. Taking into account that more than half of the papers were related to IST projects, we can discover an interesting correlation/decorrelation between these projects and practical vs. theoretical studies.

The Summit has a follow-up life as well, in the form of a post-conference publication; as far as known by the author this was non-existing in previous-year Summits. It is foreseen that a significantly deepened and more detailed version of the best papers – 7-10% of all – will appear in the form of a book to be published by Springer. This is foreseen as the first issue of a new series called Lecture Notes in Electrical Engineering (LNEE). Appearance of the book is foreseen for early 2008.

As this paper is written four days before the Summit inauguration the author, not being an oracle, has no knowledge about its success or failure. Help us God!