# 8$^{th}$ Scientific Conference of Young Researchers

# SCYR 2008

**Faculty of Electrical Engineering and Informatics**
**Technical University of Košice**

**Proceeding from Conference**

**May 28$^{th}$, 2008**
**Košice, Slovakia**

# General partners and sponsors

## General Partner

## Sponsors

**8th Scientific Conference of Young Researchers**
**of Faculty of Electrical Engineering and Informatics**
**Technical University of Košice**

Proceeding from Conference

Published: Faculty of Electrical Engineering and Informatics
Technical University of Košice
I. Edition, 163 pages, the number of CD Proceedings: 70 pieces

Editors: doc. Ing. Ľubomír Doboš, PhD.
Ing. Michal Forgáč
Ing. Jana Modrovičová

**ISBN 978-80-553-0036-8**

**Program Committee of 8th Scientific Conference of Young Researchers of Faculty of Electrical Engineering and Informatics Technical University of Košice**



**Organization Committee of 8th Scientific Conference of Young Researchers of Faculty of Electrical Engineering and Informatics Technical University of Košice**



**Contact address:**
Faculty of Electrical Engineering and Informatics
Technical University of Košice
Letná 9
042 00 Košice
Slovak Republic

## Foreword

Dear Colleagues,

The Organizing Committee welcomes you to the 8th Scientific Conference of Young Researchers at Faculty of Electrical Engineering and Informatics (FEI) Technical University of Košice (TUKE) (SCYR 2008). (Preceding title: PhD Student Conference and Scientific and Technical Competition of Students of Faculty of Electrical Engineering and Informatics Technical University of Košice).

8th Scientific Conference of Young Researchers at Faculty of Electrical Engineering and Informatics Technical University of Košice (SCYR 2008) is traditionally organized in the campus of Technical University of Košice.

8th Scientific Conference of Young Researchers at Faculty of Electrical Engineering and Informatics Technical University of Košice (SCYR 2008) is a local scientific meeting with eight year long tradition: it has been established by the dean on Faculty of Electrical Engineering and Informatics as an opportunity for PhD and graduating students use this event to train their scientific knowledge exchange. Nevertheless, the original goal to represent a forum for the exchange of information between young scientists from academic communities on topics related to their experimental and theoretical works in the very wide spread field of electronics, telecommunication, electrotechnics, computers and informatics, cybernetics and artificial intelligence, electric power engineering, remained unchanged.

8th Scientific Conference of Young Researchers at Faculty of Electrical Engineering and Informatics Technical University of Košice (SCYR 2008) is a guiding forum in university especially for young scientist as well for graduate students within their competition of the high schools and of various branches of industry.

The program of conferences this year includes two parallel sessions (both consist of oral and poster part):

- Electrical & Electronics Engineering
- Informatics & Telecommunications

with about 45 technical papers dealing with research results obtained mainly in university environment. We hope, that this day will be filled with a lot of interesting scientific discussions among the junior researchers and graduate students, and the representatives of the Faculty of Electrical Engineering and Informatics. This Scientific Network includes various research problems and education, communication between young scientists and students, between students and professors. Conference is also a platform for student exchange and a potential starting point for scientific cooperation.

We want to thank all participants for contributing to these proceedings with their high quality manuscripts.

It is our pleasure and honor to express our gratitude to our sponsors and to all friends, colleagues and committee members who contributed with their ideas, discussions, and sedulous hard work to the success of this event.

We also want to thank our session chairs for their cooperation and dedication throughout the whole conference.

Finally, we wish you all, the attendees of the conference, fruitful discussions and a pleasant stay in our event.

Ľubomír DOBOŠ
Roman CIMBALA
Alena PIETRIKOVÁ

Liberios VOKOROKOS

May 28$^{th}$ Košice

## Content

### 1st section: Electrical & Electronics Engineering

# 1st section: Electrical & Electronics Engineering

# Compensation of Wall Effect for Through Wall Moving Target Localization by UWB Radar

*Jana ROVŇÁKOVÁ*

Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

jana.rovnakova@tuke.sk

***Abstract*—Ultra wideband radar enables moving target localization also behind walls. In consequence of wall effect estimated trajectory can be considerably spatially shifted and distorted. In the paper, two different methods for compensation of this effect are described. Their effectiveness is evaluated at simulated as well as real radar data. Obtained results prove that proposed novel approach reach the best outcomes.**

***Keywords*—Moving targets, through wall localization, UWB radar, wall effect.**

## I. Introduction

Ultra wideband (UWB) radars which operate in a lower GHz-range base-band (up to 5 GHz) are characteristic with good penetration through various obstacles, e.g. through most common building materials including reinforced concrete, concrete block, sheet rock, brick, wood, plastic, tile, and berglass, as well as through ground or snow. There are a number of practical applications where such radars can be very helpful, e.g. through wall tracking during security operations, through wall imaging during fire, through rubble localization following an emergency (e.g. earthquake or explosion) or through snow detection after an avalanche, etc.

It is well known that higher permittivity of a wall results in slower velocity of the signal propagation inside the wall. This effect, which we refer to as wall effect, displaces targets from their true positions. The reason is that time of arrivals (TOA) obtained by processing raw radar data are recomputed in localization algorithms to distances by velocity of light, which responds to velocity of signal in free space [1]. In order to avoid described effect, target traces, which are formed by TOA of the signals reflected from moving targets in particular time instants of observation, should be time shifted so that they correspond with traces for the same scenario but without walls [2]. For realization of this operation, it is required to determine time difference caused by the wall in through wall TOA, that is in next called the delay time.

In the literature, several methods for exact or approximative computation of the above mentioned delay time can be found, e.g. in [3], [4], [5]. However, all of them are connected with radar imaging techniques, when the target locations are not calculated analytically but they are seen as radar blobs in gradually generated radar images. For these techniques it is possible to compute exactly the delay time caused by the wall by reason that for every pixel of radar image (which correspond with spatial point in scanned area) it is able to uniquely determine TOA as well as distance which signal flights through wall. Inverse assignment, that is needed for conventional localization algorithms, is not unique, i.e. a whole set of points belongs to known TOA and these points differ in regard to the delay time under consideration. As a result of this, effect of the wall can be compensated only by approximative methods.

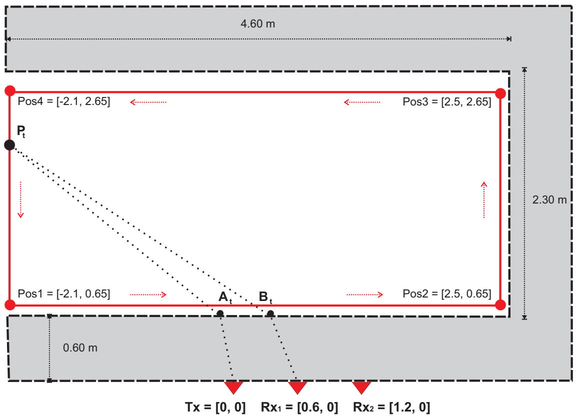



Fig. 1. A scheme of scanned area for given measurement scenario with illustration of distances required for $TOA_{TW,1}(P_t)$ computation.

In this paper, two of such methods are analyzed firstly at simulated and thereafter at real radar data. The first of the methods is described e.g. in [3], the other introduces novel approach based on knowledge of the delay time for radar images.

## II. Wall Effect and Its Consequences Showed on Simulated Data

To be possible to compare results of simulated and real data, we consider in both cases the same scenario: one person walks along perimeter of rectangular room with size 4.6 m $\times$2.3 m. The walls are concrete with relative permittivity 6.12 and thickness of 0.6 m. UWB radar system consists of one transmitting antenna $Tx$ and two receiving antennas $Rx_1$ and $Rx_2$. Their configuration is depicted in the scanned area scheme in Fig. 1.

We assumed ideal conditions for simulations, i.e. point target with uniform velocity and no additional sources of errors. Simulated target traces pertaining to rectangular trajectory were computed for scenario with and without wall. Their are depicted in Fig. 2, where horizontal axis represents time observation $t$ and vertical axis corresponds with $TOA$ of signal transmitted by $Tx$, reflected by target at position $P_t$ and received by antenna $Rx_i$, for $i = 1, 2$.

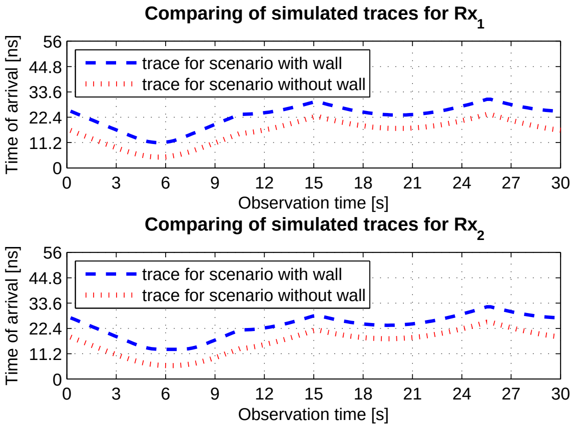



Fig. 2. Ideal simulated traces for through wall and no wall scenario.

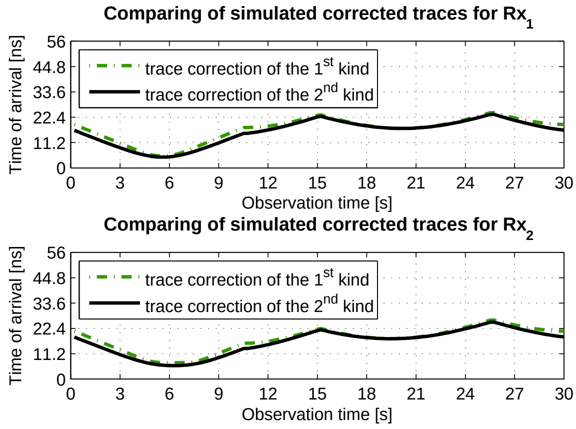



Fig. 4. Compensation of the wall effect by correction of traces.

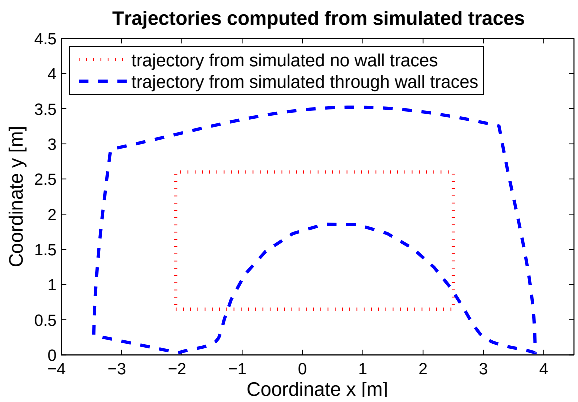



Fig. 3. Consequence of wall effect on estimated trajectory.

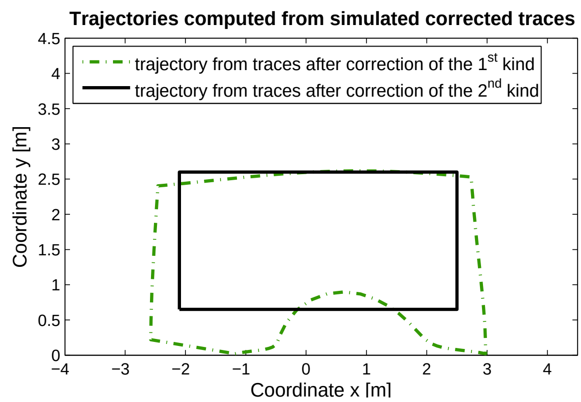



Fig. 5. Resultant trajectories after trace correction.

For scenario without wall, $TOA$ are calculated as follows

$$TOA_{noW,i}(P_t) = \frac{dist(Tx, P_t) + dist(P_t, Rx_i)}{c} \qquad (1)$$

where $dist(X, Y)$ represents distance between points $X, Y$ and $c$ is velocity of light. In case of through wall scenario, the computation is more complex:

$$TOA_{TW,i}(P_t) = \frac{dist(Tx, A_t) + dist(B_t, Rx_i)}{v_w} + \\ + \frac{dist(A_t, P_t) + dist(P_t, B_t)}{c} \qquad (2)$$

where $A_t$ and $B_t$ are refraction points at border line wall-air and $v_w$ is velocity in the wall (Fig. 1). It depends at relative permittivity of the wall and for concrete wall under consideration gets value

$$v_w = \frac{c}{\sqrt{\varepsilon_r}} = \frac{3.10^8}{\sqrt{6.12}} \doteq 1.2 \cdot 10^8 m/s \qquad (3)$$

The distances inside the wall, $d_w = dist(Tx, A_t) + dist(B_t, Rx_i)$, can be calculated on the base of minimalization task, the solution of that is described e.g. in [5].

As was mentioned before, one of the input variables to localization algorithms are $TOA$ obtained from radar data and recomputed to distances by utilization of velocity of light. In such way, we achieve true rectangular trajectory by combination of $TOA_{noW}$ from both receivers, but in case of through wall data, $TOA_{TW}$, the resultant trajectory is shifted and distorted (Fig. 3).

Simply said, in order to avoid consequences of wall effect on estimated trajectory, the through wall target trace should be approached as much as possible to trace corresponding with no wall scenario (i.e. to get from blue dashed to red dotted curve in Fig. 2).

## III. Methods of Wall Effect Compensation

The most easy way how to reduce effect of the wall is to assume simplified model in which signal flights through wall always perpendicularly. The delay time through wall compared to free space propagation is determined by

$$\tau_{delay} = \frac{d_w}{v_w} - \frac{d_w}{c} = \frac{\sqrt{\varepsilon_r} d_w}{c} - \frac{d_w}{c} = \frac{d_w}{c}(\sqrt{\varepsilon_r} - 1) \qquad (4)$$

For simplified model under consideration, the distance inside the wall $d_w$ gets value $2 \cdot 0.6$ m what corresponds with thickness of wall overflown in both directions. Target traces modified by $\tau_{delay} = \frac{2 \cdot 0.6}{3 \cdot 10^8}(\sqrt{6.12} - 1) \doteq 5.9$ ns are depicted in Fig. 4 as green dash-dot curves.

Distances $d_{i,t}$ between $Tx$, $P_t$ and $Rx_i$ for $i = 1, 2$, which enable to estimate target location in given observation time instant $t$, are computed now as follows

$$d_{i,t} = [TOA_{TW,i}(P_t) - \tau_{delay}] \cdot c \qquad (5)$$

Resultant trajectory is shown in Fig. 5 by the same color and type of line. As it can be seen, obtained location estimations much better correspond with true trajectory, but still some distortion is there visible, especially near antennas positions.

The reason is that delay time is not constant, but it changes in dependence on the target position relative to the antenna array.

Aforesaid defection can be solved by following proposed method. It computes the delay time for every time instant of observation particularly. At first we need to determine spatial grid in which individual points will be examined, e.g. for coordinate system used in examined scenario it was grid from $-4$ m to $4.5$ m in $x$ direction and from $0$ m to $4.5$ m in $y$ direction with step of $0.025$ m in both directions. For these points we know to calculate $TOA_{noW,i}$ and $TOA_{TW,i}$ according (1) resp. (2) for receiving antennas $Rx_i$, $i = 1, 2$. The next step rests in finding all points $P$ from matrix of $TOA_{TW,i}$ for which $TOA_{TW,i}(P) = TOA_{TW,i}(P_t)$, i.e. is equal to known value of through wall $TOA$ for given receiver $i$ and time observation $t$. Since we work with bistatic radar ($Tx$ and $Rx_i$ are not identic), a set of searched points form ellipse around $Tx$ and correspondent $Rx_i$. If on ellipse $E_1$ belonging to $Rx_1$ and on ellipse $E_2$ belonging to $Rx_2$ is possible to find the same points (i.e. intersections $I_t$), the delay time caused by wall is given by

$$\tau_{delay}(i,t) = mean[TOA_{TW,i}(I_t) - TOA_{noW,i}(I_t)] \quad (6)$$

The mean value of the difference is considered from this reason that with small step of spatial grid more points in vicinity can have the same value of given $TOA$. If such points is not possible to find, the delay time is estimated as follows

$$\tau_{delay}(i,t) = mean[TOA_{TW,i}(E_i) - TOA_{noW,i}(E_i)] \quad (7)$$

In such way we obtain not only the time value about which is needed to decrease through wall $TOA$, but we have available also primary estimation of searched target locations. These can be advantageously used as initialization values for more complex iterative localization algorithms or be helpful for target tracking [1].

For ideal simulated radar data proposed method totally remove effect of the wall, i.e. by localization the true rectangular trajectory is reached (black solid curves in Fig. 4 and Fig. 5). In the next, we refer to the first described method of wall effect elimination as through wall trace correction of the $1^{st}$ kind and to the second described method as correction of the $2^{nd}$ kind.

## IV. Evaluation of Method Performance on Real Radar Data

The real data were acquired by video impulse time-domain UWB radar system with one transmitting and two receiving channels. Vivaldi antenna was used as $Tx$ antenna excited with 50 ps mono-pulse and two dielectric wedge antennas as $Rx$ antennas. During measurement all antennas were placed 1.16 m elevation above the floor and there was no separation between the antennas and the wall.

Target traces were obtained after processing raw radar data by methods of pre-processing, background subtraction, detection and trace estimation [2]. They are shown in Fig. 6 by blue dashed curve, red dotted curve represents ideal trace for no wall scenario. As can be seen from this figure, through wall traces are now not only acquired with delay time caused by wall, but also they visible differ by shape from ideal traces. This fact is consequence of processing errors and shows itself also by localization. The traces in Fig. 8 and in

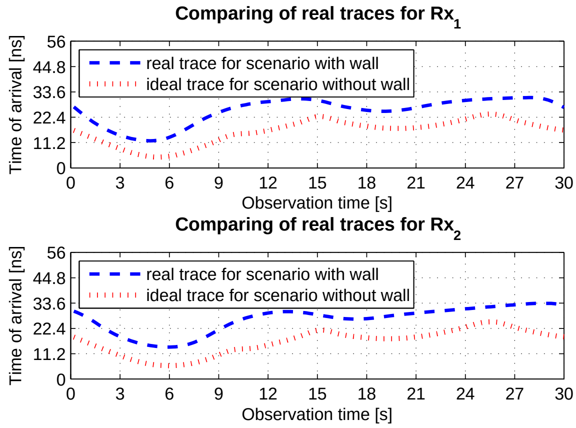



Fig. 6. Traces obtained from real through wall radar data.

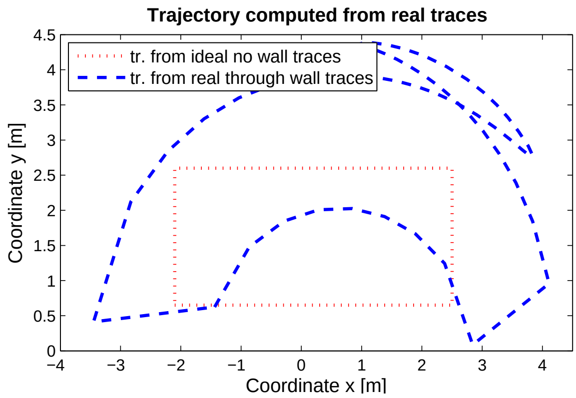



Fig. 7. Trajectory obtained from real through wall traces.

TABLE I
AVERAGE DIFFERENCE BETWEEN TRACES FOR SIMULATED DATA

| Trace in Scenario noW compared with | Average difference for $Rx_1$ | Average difference for $Rx_2$ |
|---|---|---|
| trace in Scenario TW | 7.06 ns | 7.06 ns |
| trace after Corr.1 | 1.16 ns | 1.17 ns |
| trace after Corr.2 | 0 ns | 0 ns |

Scenario noW - scenario without wall; Scenario TW - scenario with wall; Corr.1 - correction of the $1^{st}$ kind; Corr.2 - correction of the $2^{nd}$ kind.

TABLE II
AVERAGE DIFFERENCE BETWEEN TRACES FOR REAL DATA

| Trace in Scenario noW compared with | Average difference for $Rx_1$ | Average difference for $Rx_2$ |
|---|---|---|
| trace in Scenario TW | 9.12 ns | 9.93 ns |
| trace after Corr.1 | 3.23 ns | 4.04 ns |
| trace after Corr.2 | 2.67 ns | 3.37 ns |

Scenario noW - scenario without wall; Scenario TW - scenario with wall; Corr.1 - correction of the $1^{st}$ kind; Corr.2 - correction of the $2^{nd}$ kind.

Fig. 10 represent real target traces shifted about the delay time computed by two kinds of methods for wall compensation described in previous section.

Comparison of average difference between ideal no wall trace and trace for through wall scenario, respectively for its corrected version, is made in Table I for simulated data and in Table II for real data. In both data sets realized corrections markedly decreased average difference between

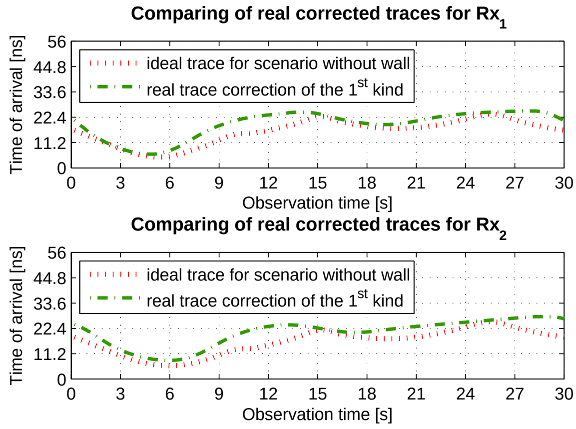



Fig. 8. Real through wall traces after correction of the $1^{st}$ kind.

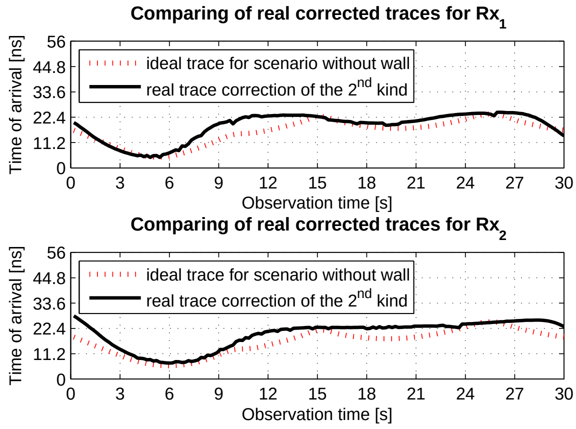



Fig. 10. Real through wall traces after correction of the $2^{nd}$ kind.

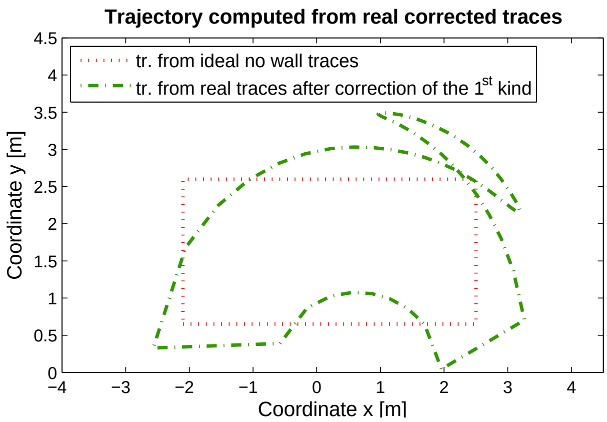



Fig. 9. Resultant trajectory after trace correction by the $1^{st}$ method.

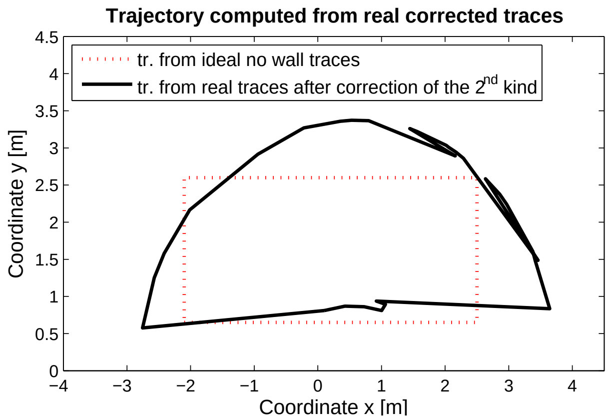



Fig. 11. Resultant trajectory after trace correction by the $2^{nd}$ method.

TABLE III
AVERAGE LOCALIZATION ERRORS FOR SIMULATED AND REAL DATA

| Trajectory from traces | Average localization error for simulated data | Average localization error for real data |
|---|---|---|
| for Scenario TW | 1.34 m | 2.75 m |
| after Corr.1 | 0.75 m | 2.20 m |
| after Corr.2 | 0 m | 1.36 m |

Scenario TW - scenario with wall; Corr.1 - correction of the $1^{st}$ kind; Corr.2 - correction of the $2^{nd}$ kind.

traces. Proposed method for compensation of wall effect (i.e. trace correction of the $2^{nd}$ kind) reached the best results in all examined cases.

Trajectory calculated from real uncorrected data is depicted in Fig. 7, trajectories from corrected traces are shown in Fig. 9 and in Fig. 11. In consequence of processing errors mentioned before, estimated locations form in one place of trajectory false loop. By proposed method of wall effect compensation this misleading movement was partially removed.

Comparison of average localization errors for simulated as well as for real data is made in Table III. Trace correction of the $1^{st}$ kind decreased quantity under consideration about almost 50% for simulated data and about 20% for real data in comparison with average localization error for uncorrected trajectory. Trace correction of the $2^{nd}$ kind removed average localization error totally in the case of simulated data and for real data this error was decreased about almost 50%.

## V. Conclusion

Results from previous sections clearly show that wall effect can cause considerable localization errors. For its compensation two different methods were analyzed. The first of them, which can be found also in literature, in spite of simplicity achieves very good results for simulated as well as real data. The second method, which introduces novel approach, is computational much more complex, but in addition to improved outcomes offers two other benefits - primary estimation of locations and better reformation of trajectory shape.

## Acknowledgment

This work was supported by the Slovak Research and Development Agency under the contract No. LPP-0287-06 and by European Commission under the contract COOPCT- 2006-032744.

## References


[1] K. Yu and I. Oppermann, "Performance of uwb position estimation based on toa measurements," in *Joint UWBST & IWUWBS*. Kyoto, Japan, 2004, pp. 400 – 404.

[2] J. Rovňáková, M. Švecová, D. Kocur, T. T. Nguyen, and J. Sachs, "Signal processing for through wall moving target tracking by m-sequence uwb radar." $18^{th}$ International Conference RADIOELEKTRONIKA, Apr. 2008, to be published.

[3] R. Tanaka, "Report on SAR imaging," internal Report, Oct. 2003.

[4] S. Gauthier, E. Hung, and W. Chamma, "Surveillance through concrete walls," in *Proceedings of SPIE 5403*, Dec. 2004, p. 597 608.

[5] M. Aftanas, J. Rovňáková, M. Drutarovský, and D. Kocur, "Fast computation of the real time of arrival for through the wall uwb radar systems." ICUWB, Hannover, Germany, Sept. 2008, submitted for publication.

# Dynamic Bandwidth Allocation Issues for Geostationary Interactive Satellite Systems

[1]*Peter PATLEVIČ*, [2]*Salvatore PULITANÒ*

[1]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic
[2]Dept. DIMET, Faculty of Engineering, University "Mediterranea" of Reggio Calabria, Italy

[1]peter.patlevic@tuke.sk, [2]salvatore.pulitano@unirc.it

***Abstract*—In the last few years there has been a growing interest for data transmission over geostationary (GEO) interactive satellite networks. This is due to the fact that GEO systems permit to deliver high-bandwidth services in geographical coverage areas. Obviously, in order to obtain an efficient usage of the satellite resources, while guaranteeing adequate Quality of Service (QoS) levels, it is a must to implement an efficient resource management strategy. This latter has to take into account traffic class QoS, reduce the waste of resources, and maximize the number of users who will access the system at the same time.**

***Keywords*—Return Channel, Digital Video Broadcasting, Fuzzy Logic, Neural Network, Traffic Prediction**

## I. Introduction

In the present paper, we focus our attention on a predictive and adaptive dynamic bandwidth allocation strategy and consider as a reference platform a DVB-RCS system (Digital Video Broadcasting Return Channel via Satellite) [1] [2] with multiple RCSTs (Return Channel Satellite Terminals) sharing the same return (uplink) channel. Goal of the proposed technique is to dynamically allocate the uplink bandwidth of a system by taking into account prediction about "future" traffic characteristics. In particular we will consider Rate Based Dynamic Capacity (RDBC) connections, each one with likely different QoS requirements. We are considering this class of capacity request because it is the most suitable for delivering IP traffic (internet interactive and multimedia traffic [3]) over DVB-RCS with no waste in resources.

## II. Proposed Approach

The proposed approach uses Neural Network for traffic prediction and several cooperating Fuzzy Logic blocks for prediction correction and requested bandwidth estimation, as depicted in Fig. 1. Our proposal is derived from some basic concepts proposed in a different environment and inspiring the interesting schemes proposed in [4] and [5]. The fuzzy-based dynamic allocation idea proposed in [4] can be useful adapted to satellite systems, because the scheme proposed is simple and relies on a very robust algorithm, although it shows a tendency to underestimate or overestimate the required bandwidth: in case of video and audio sources required bandwidth is underestimated, while in case of data source required bandwidth is overestimated. This problem, high significant in a satellite environment, can be reduced by applying a bandwidth predictor for next time slot. This is done by using a similar approach as the one in [5], where a neural network is used for direct estimation of the bandwidth required for the connection in next time slot.

The scheme proposed by the authors consists of several fuzzy systems and one neural network. The Admission Control Unit is based on a fuzzy system. Such a module needs both (i) system-related information, such as the available bandwidth CA and the Packet Loss Ratio PL and (ii) information relevant to the class of incoming connection, such as QoS parameters [6]. The traffic belonging to the accepted classes are, then, forwarded to the so called Class Buffer. The Bandwidth Estimator (BE), for each admitted connection, predicts the value of the needed bandwidth CE. The estimated bandwidth is the main input of the Bandwidth Range Reallocator (BRR). This is a fuzzy system that decides how much bandwidth (namely Bandwidth Expansion Factor - BEF) shall be added to or subtracted from the currently allocated classes of connections. The Allocated Bandwidth Range Controller (ABRC) has the task to determine if the allocated bandwidth in the preceding time slot was either underestimated or overestimated. The final module is the Dynamic Bandwidth Controller (DBC), which takes the last decision about increasing or decreasing the allocated bandwidth. Inputs to this unit are the outputs from the ABRC unit, the BRR unit, and total number of active connections AC. DBC may increase/decrease the bandwidth amount computed by BRR depending on the output from the ABRC unit. If ABRC detects that the connection behaviour was overestimated in the previous time slot, then the DBC unit makes the new value of bandwidth (or BEF) decrease to try to avoid overestimation in next time slot (or vice-versa).

In this paper we will specifically stress the concepts which drive the behaviour of the Bandwidth Estimator block. This latter plays a critical role. In fact, with reference to the typical MAC performance required in a satellite environment [7] [8] and to the main DVB-RCS standard features transmission protocol is based on three distinct steps:

- bandwidth request (via Capacity Request Message)
- resource scheduling (via TBTP Message)
- data traffic transmission

Fig. 2 shows the request-allocation-transmission procedure. In order to request enough resources to serve the traffic buffered up to the $(i-1)-th$ frame, a capacity request is forwarded to the NCC in the $i-th$ frame. Let d denote the Round-Trip-Time plus the capacity request processing time at the NCC (measured in frames), then the transmitting RCST can send the information traffic, after the reception of the

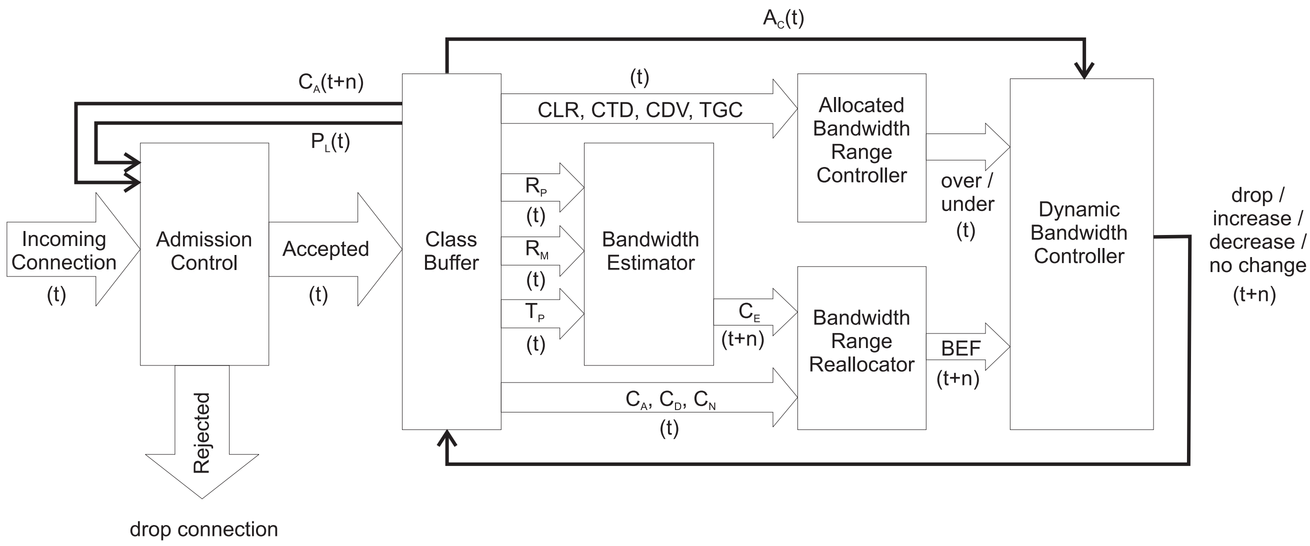



Fig. 1. Admission control and dynamic bandwidth allocation with fuzzy systems and neural networks

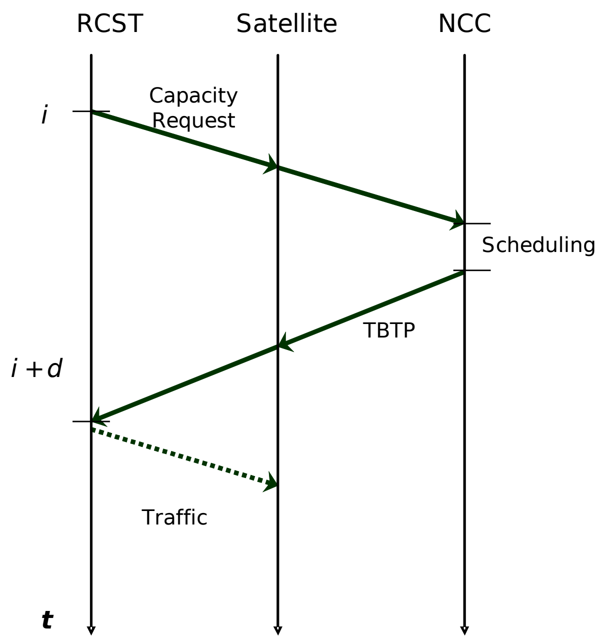



Fig. 2. The MAC signaling temporal sequence

TBTP message in the $(i + d) - th$ frame.

In so doing, the system doesnt take into account that variation in traffic production could happen. In our proposal, the BE will evaluate the bandwidth necessity considering a prediction of future traffic, taking into account the delay of d frames above outlined. Considering a Round Trip Time of $540ms$ and a frame duration of $60ms$ a value of $d = 9$ has been chosen.

### A. Bandwidth Estimator

Bandwidth Estimator consist of full connected feed-forward Neural Network with Back-propagation of Error learning algorithm. The network topology is 54-15-1, the activation function for neurons in both input and hidden layer is tangential sigmoid; for neurons in the output layer is linear. As a training method , gradient descent with momentum back-propagation of error has been used. The momentum parameter was set to 0.55 and the network was trained for max. 2000 epochs. The Bandwidth Estimator needs as input information belonging to a temporal window 50 frames long. In particular current bitrate (that is the data transmission rate at the current frame $i$), maximum and minimum bitrate are considered in order to estimate the needed resources for the frame $(i + d)$.

For training and testing purposes, 15000 frames long fragments from VBR MPEG2 encoded movies [9] have been used (the outputs relevant to “Jurassic Park” and “Silence of the Lambs” are shown in the present extended abstract, but a more exhaustive simulation campaign has been conducted).

In order to verify the performance of the suggested Connection Admission Control scheme, an exhaustive simulation campaign was conducted. The features of the considered system are listed in Table I.

TABLE I
SIMULATION FEATURES

| | |
|---|---|
| Return Link Capacity | 32 Mbps |
| Terminal Capacity | 10 Mbps |
| Frame Duration | 0.06 s |
| Time Slot per Frame | 32 |
| Time Slot Capacity | 198 Kbps |
| Round Trip Time | 0.270 s |
| Access Scheme | MF-TDMA |

## III. Simulation Results

Results of prediction are presented in Table II and depicted on Fig. 3 and Fig. 4. The reached average deviation was 8.55% for the jurrasic data and 10.82% for the silence data. This is the difference between average bitrate of original and predicted data. Most important is the deviation of the peaks (max and min bitrate). In this points moves the average deviation between 10 to 20%. The results shows, that the prediction is accurate enough to support the resource allocation scheme. In next step should be investigated the implementation of this block together with fuzzy systems (ABRC, BRR and DBC block) which roles is to correct the deviation of the prediction.

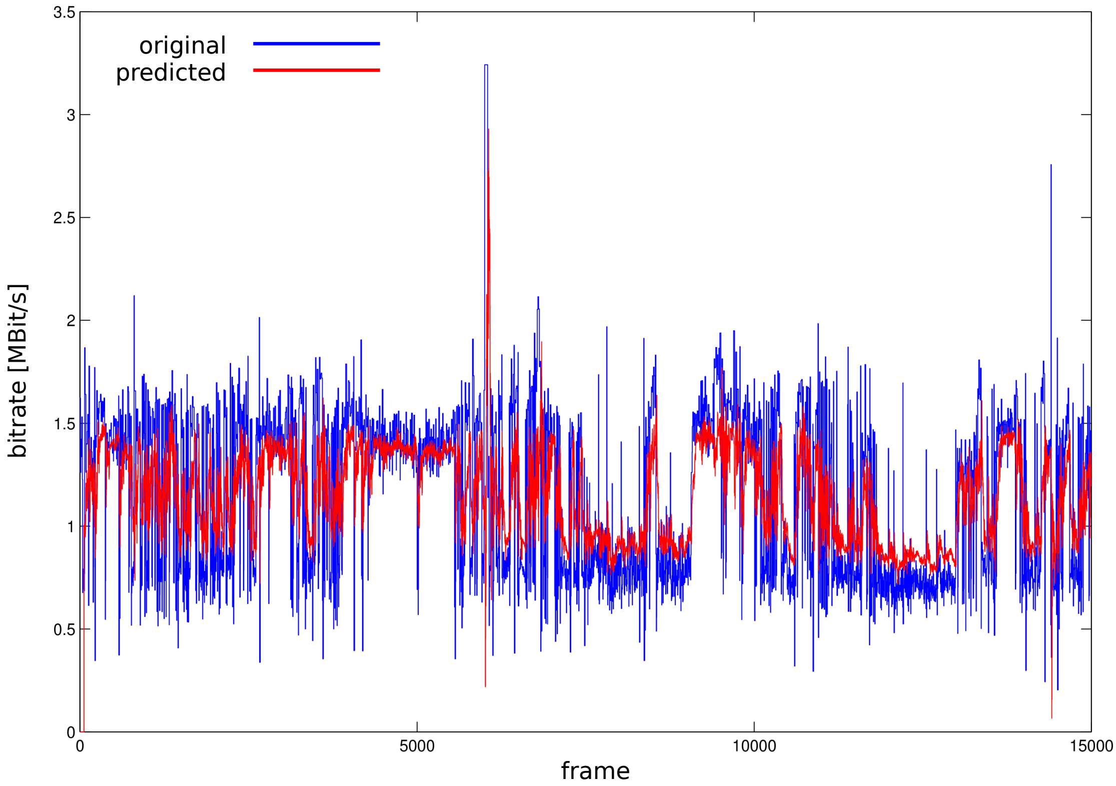



Fig. 3. Bitrate (requested bandwidth) prediction for the Jurassic data (red curve represent the predicted value, blue the original value)

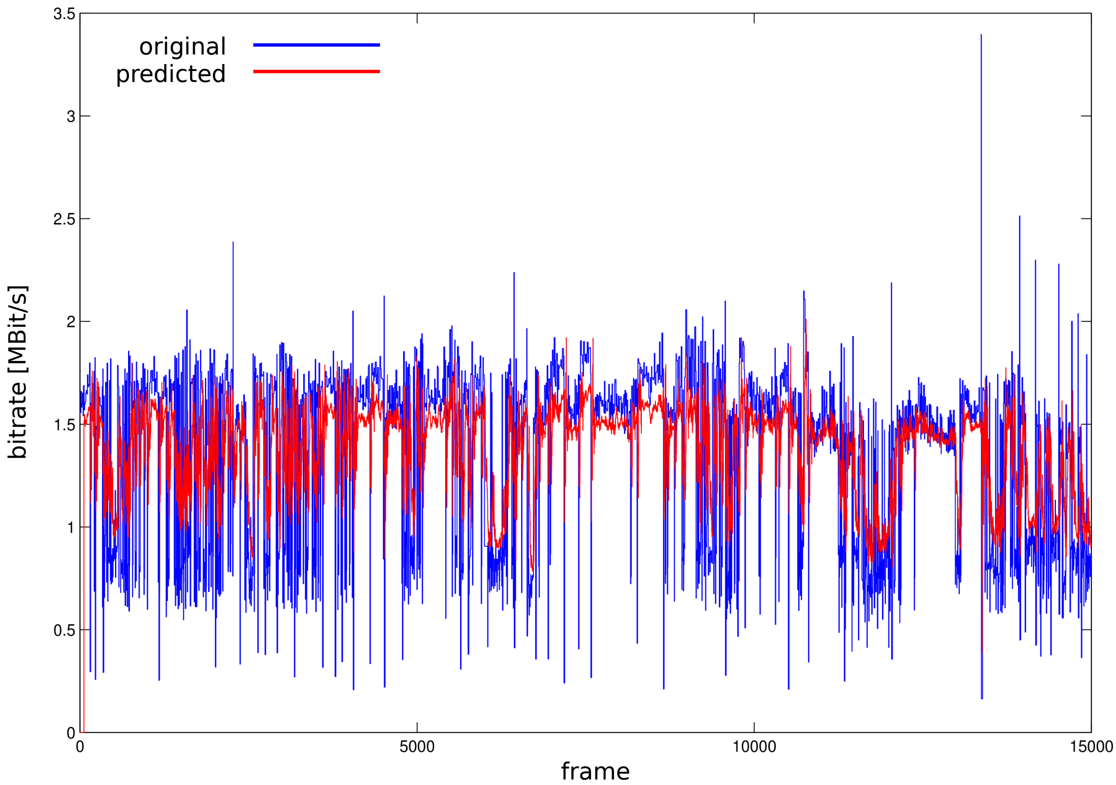



Fig. 4. Bitrate (requested bandwidth) prediction for the Silence data (red curve represent the predicted value, blue the original value)

TABLE II
PREDICTION RESULTS

| data | avg bitrate [MBit/s] | min bitrate [MBit/s] | max bitrate [MBit/s] | deviation [%] |
|---|---|---|---|---|
| original jurrasic | 1.493 | 0.263 | 3.281 | |
| predicted jurrasic | 1.315 | 0.170 | 2.891 | 8.55 |
| original silence | 1.119 | 0.205 | 3.447 | |
| predicted silence | 0.931 | 0.389 | 2.194 | 10.82 |

## ACKNOWLEDGMENT

Research described in the paper was financially supported by VEGA No.1/4054/07, and also by COST 2100 - Pervasive Mobile & Ambient Wireless Communications.

## REFERENCES

[1] "Etsi tr 101 790 v1.3.1 digital video broadcasting (dvb); interaction channel for satellite distribution systems; guidelines for the use of en301 790."

[2] L. Schwartz, V. Hottmar, and D. Trstensky, "End user box for interactive cable television," *IEEE Transaction on Consumer Electronics*, vol. 53, pp. 412–416, May 2007, iSSN 0098-3063.

[3] L. Schwartz, "Ip networks," *EDIS ZU v Ziline*, p. 176, 2006, iSBN 80-8070-504-6.

[4] S. Chandramathi and S. Shanmugavel, "Fuzzy-based dynamic bandwidth allocation for heterogeneous sources in atm networks," *Elsevier Science B.V. Applied Soft Computing*, vol. 3, pp. 53–70, 2003.

[5] X. L. N. Bourbakis, "Neural network training using genetic algorithms in atm traffic control."

[6] M. Krasnovsky and V. Wieser, "End-to-end delay, jitter and throughput in wireless ad-hoc networks,," pp. 135–138, 2007, iSBN 978-80-8070-693-7.

[7] H. Peyravi, "Medium access control protocols performance in satellite communications," *IEEE Communications Magazine*, vol. 37, pp. 62–71, 1999.

[8] M. Costabile, R.Ferraro, C. Follino, and A. Iera, "A mac protocol for the efficient access to the return channels of dvb-rcs satellite platforms," *Wireless Personal Multimedia Communications Conference*, 2003.

[9] [Online]. Available: http://www-tkn.ee.tu-berlin.de/research/trace/trace.html

# Data Security in Wireless Sensor Networks – the Task for Public Key Cryptography

*Michal VARCHOLA*

Dept. of Electronics and Multimedia Communications,
FEI TU of Košice, Slovak Republic

Miso.Varchola@tuke.sk

***Abstract*—The paper describes security algorithms and protocols provided by recent WSN stacks where symmetric-key schemes are commonly used. Using these schemes seems to be impractical in large scale networks; hence the paper intends their replacement by public-key schemes for low cost and low power MCU platforms discussed below. Moreover, the paper proposes implementation of Wireless Sensor Network (WSN) stacks embedded in common microcontroller (MCU) platforms – ARM7TDMI, x51, ColdFire, and HCS08 in order to compare it. Protocol stacks of proposed WSNs are based on the both – a proprietary and the ZigBee.**

***Keywords*—Elliptic Curve Cryptography, IEEE 802.15.4 Standard, Public Key Cryptography, Sensor Networks, ZigBee**

## I. Introduction

Wireless Sensor Networks (WSNs) have received considerable attention during last decade [1], [2], [3]. WSNs can be applied to a large number of areas, and its applications are continuously growing. They are expected to be used in a wide range of applications, from number military to various civil. Intentions of military are in target sensing or tracking in battlefields [4] or detection of biological or chemical weapons, or sensor nodes also could be deployed into enemy territory to observe it. WSNs penetrate into civil areas as well – biomedical, healthcare, building or home automation and environment from wildlife monitoring, early fire detection in forests or collecting microclimate data [5], [6], [7] to outdoor deployments of sensor networks to monitor storms, oceans, and weather events.

Paul Saffo from Institute for the Future says: [8] "Just as the personal computer was a symbol of the '80s, and the symbol of the '90s is the World Wide Web, the next nonlinear shift, is going to be the advent of cheap sensors." Let's add: secure sensors, because of besides the battlefield applications, security is critical in healthcare systems at hospitals or Home Automation (HA) too. WSNs are, in general, more vulnerable to attacks and unauthorized access than traditional (wired) networks. For example, an adversary can easily listen to the traffic and mislead communications between nodes.

WSNs are typically characterized by limited power supplies, low bandwidth, small memory sizes and limited energy. This leads to a very demanding environment to provide security. Because of that special characteristics and limitations of wireless sensor networks, desingers face an important challenge in security issue, particularly for the applications where WSNs are developed for use in a hostile environment or used for some crucial purposes. In order to establish the secure network, it is necessary to design secure protocols to deal with problems about key agreement and encryption in communications.

Many applications in the area of WSN would gain a lot from the availability of strong public-key cryptography (PKC). Recently a number of studies have been conducted to find out a practical way to use PKC in WSNs [9], [10], [11], [12].

The paper is organized as follows. Section 2 briefly presents the overview of recent platforms enabling to build WSNs. Next, Section 3 discusses security issues in WSNs. Section 4 deals with experimental result of implementation protocol stacks. Finally Section 5 presents conclusions and discusses future development.

## II. WSN Platform Overview

WSNs usually consist of a large number of ultra-small autonomous devices. Each device, called a node, is battery powered and equipped with integrated sensors, microcontroller (MCU) and Radio Frequency (RF) circuits. Nowadays, there are a lot of available MCUs and RF chips or their one-chip combination as well as various software (SW) protocol stacks. Selecting the most suitable platform is important decision in order to achieve better results. Aim of this section is to choose and to describe suitable platform for performing tests of cryptographic primitives and WSN stacks.

### A. Available hardware

Today's market offers various hardware (HW) platforms of different properties for WSN implementing. Manufacturers – Atmel [13], Ember [14], Freescale [15], Jennic [16], Microchip [17], Nordic [18] or Texas Instruments [19] produce wide variety of Radio Frequency (RF) transceivers for various frequency bands and standards from 433 MHz up to 2.4 GHz. Transceivers based on IEEE 802.15.4 standard are becoming most used in commercial area because of robust radio properties. This standard is a base of the ZigBee. Both standards are briefly described in the next sections. Manufacturers of RF chips offer various MCUs for running optimized network SW. Developers can choose between industry standard cores such as: 8051 clones (Texas Instruments - CC2431) and ARM7TDMI (Freescale - MC13225) or special vendor cores such as AVR (Atmel), HCS08 (Freescale), Coldfire (Freescale) or MPS430 (Texas Instruments). There is option to select 8-bit (AVR, HS08, 8051), 16-bit (MSP430) or 32-bit (Coldfire, ARM7TDMI) depending how powerful application has to be. Recent modern trend is to merge RF chip and MCU into one package for board area and silicon saving.

### B. Evaluation Hardware

Evaluation hardware was selected regarding to ambition of testing cryptographic protocols and interoperability between open and commercial WSN stacks. The industrial standard 8051 clone and ARM7TDMI cores was selected for testing open stacks and cryptographic protocols because of their general availability and simple porting assembly optimized SW form one clone to another. Analog Devices ADuC845 was chosen as 8051 clone. This MCU is based on modern single cycle x51 clone with 64 kB Flash and 2.3 kB Static RAM (SRAM). The most powerful peripheral in this MCU is a 24-bit sigma-delta Analog-to-Digital Converter (ADC) with programmable input gain amplifier in 1-128 gain range. The NXP (Phillips) LPC2138 was chosen as representative of the ARM7TDMI architecture. This chip provides large 512 kB Flash and 32 kB SRAM memory. There is possibility to clock it at 60 MHz, what could ensure good performance for time-critical tasks. The MC13203 chip ensures RF connectivity to MCUs. Each chip is placed on its own evaluation board designed at Department of Electronics and Multimedia Communications (DEMC) with rich connectivity options.

Freescale products were chosen for running commercial WSN stacks because of availability wide range of products. It is possible to use cheaper 8-bit HCS08 core or faster 32-bit Coldfire with mutual compatibility (Flexis series [15]). There will be an option to use proclaimed powerful ARM7TDMI based MC13225 as well. Developer can choose between one-chip (MC13214) or more-chip solution (MC13203 + Flexis). Each solution is available with various memory size or with different stack usage privilege (from simple to ZigBee).

### C. Available Software Stacks

Almost each of chip vendor mentioned above provides the ZigBee or a proprietary protocol stacks. Proprietary solutions are usually available free of charge and in open ANSI (American National Standards Institute) C form such as – Simple Media Access Controller (SMAC) protocol by Freescale, JenNET by Jennic, or MiWi Wireless Networking Protocol Stack by Microchip. The ZigBee stack is available either for free of charge (e.g. Mircochip or Texas Instuments) or not (e.g. Freescale) and either in open ANSI C form or in binary form respectively, depending on chip vendor. Freescale provides powerful and easy to use Graphical User Interface (GUI) – BeeKit Development Environment, in which the users can create, modify, save and update wireless networking solution. The Freescale platform was chosen because of supporting various MCU families, and supporting three levels of protocols – SMAC, IEEE 802.15.4 MAC and ZigBee by BeeKit sowtware. Latest two are described in next section. SMAC is set of functions for basic interfacing the IEEE 802.15.4 compliant radio. The SMAC is subset of the IEEE 802.15.4 compatible protocol and offers basic peer-to-peer connectivity only.

### D. IEEE 802.15.4 standard and ZigBee Stack

ZigBee is a low-cost, low-power, wireless mesh networking standard. The low cost allows the technology to be widely deployed in wireless control and monitoring applications, the low power-usage allows longer life with smaller batteries, and the mesh networking provides high reliability and larger range.

The ZigBee Alliance [20] selected the IEEE 802.15.4 standard [21], released in May 2003, as the base of ZigBee networking and applications. IEEE 802.15.4 defines three frequency bands: 868 MHz, 915 MHz and 2.45 GHz. The latest is used the most frequently thanks to worldwide availability and 250 kbps bit-rate. Except frequency bands, modulation and spreading methods IEEE 802.15.4 also define relatively simple protocol, based on CSMA/CA (Carrier Sense Multiple Access/Collision Avoidance) access method to the medium.

The ZigBee specification identifies three kinds of devices that incorporate ZigBee radios, with all three found in a typical ZigBee network:

- *Coordinator (ZC):* organizes the network and maintains routing tables,
- *Routers (ZR):* can talk to the coordinator, to other routers and to reduced-function end devices,
- *End devices (ZED):* can talk to routers and the coordinator, but not to each other.

ZC and ZR are defined as Full-Function Devices (FFD), which are powered on all the time where mains power is recommended. ZED is defined as Reduced Function Device (RFD) where the protocol stack is shorter without ability of routing but this device could be battery powered. Sensors and actuators could be connected to each of these three ZigBee devices. Except common used mesh topology, it is possible to use tree or star topology, which take less HW and SW resources of the MCU.

Network devices, whether wired or wireless, are commonly described by the Open Systems Interconnection (OSI) reference model by International Organization for Standardization (ISO). The adaptation ISO-OSI network reference model for ZigBee purposes is illustrated in the Fig. 1. ZigBee network model does not use presentation, session or transport layer and user application is directly tied into Application layer (APL). This figure shows also IEEE 802.15.4, ZigBee Alliance, and ZigBee product end manufacturer particular responsibility for ZigBee certified product as well as HW and SW proportion in ZigBee.

## III. Security Issues in WSNs

WSNs may have a few, hundreds or even thousand of nodes. As the networks grow, security and management begin to be

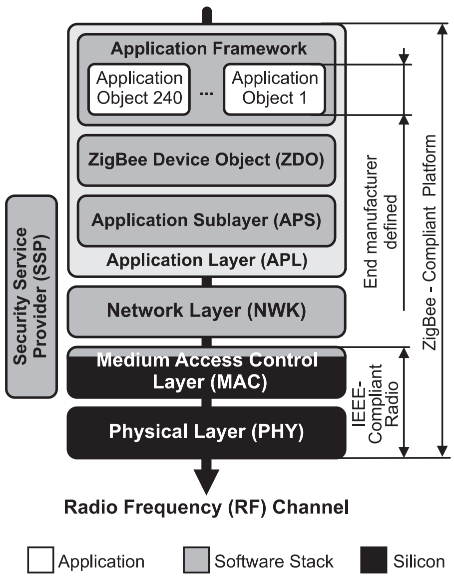



Fig. 1. Adaptation ISO-OSI to ZigBee standard where responsibility of IEEE 802.15.4, ZigBee Alliacne, and End Manufacturer is pointed out.

complicated. Implementing security protocols in WSNs is not easy proposition and systems often, for reasons of complexity, limited resources or implementation fail to deliver required levels of security. The data security and network integrity of such systems are essentially based on the safe distribution of encryption keys and device authentication. Main aim of cryptographic protocols in WSNs are – establish a key between all sensor nodes that must exchange data securely, node addition/deletion should be supported, it should works in undefined deployment environment, and unauthorized nodes should not be allowed to establish communication with network nodes.

### A. Security in the IEEE 802.15.4 standard

While the IEEE 802.15.4 standard goes into great detail when describing the functionality of Physical Layer (PHY) and Medium Access Controller Layer (MAC), security related issues received much less attention. The standard outlines some basic security services at the MAC that can be combined with advanced techniques from upper layers to implement comprehensive security solution. The IEEE 802.15.4 device can choose to operate in unsecured mode, secured mode, and Access Control List (ACL) mode. In unsecured mode, none of the services mentioned are available. In secured mode, the device may use one of security suites supported by standard, all of which use the Data Encryption Service. A device operating in ACL mode can maintain a list of trusted devices from which it expects to receive packets. While these services are useful they are by no means sufficient. In particular, procedures for key management, device authentication, and freshness protection are not specified by the IEEE 802.15.4 standard.

### B. Security Implementation in the ZigBee

Security of ZigBee is provided by Advanced Encription Standard (AES) [22]. This symmetric algorithm means communicating nodes use the same key to encrypt and decrypt the messages, but the two communicating nodes must find a way to agree symmetric key. Currently, ZigBee uses symmetric-key key establishment (SKKE) to establish keys between communicating nodes. This protocol defines the mechanism by which a ZigBee device may deliver a shared key (link key) with another ZigBee device. Key establishment involves two entities, an initiator device and respond device, and should be prefaced by trust provisioning step in which trust information (a master key) provides a starting point for establishing a link key. The master key maybe preinstalled during manufacturing, it may be installed by a trust center, or it may be based on user entered data. As has been proposed in [23] SKKE scheme is not immune to malicious attacks completely.

### C. Security Alternative (not only) for ZigBee

Distributed systems (as WSNs) are the ideal target to implement PKC where one key that only device knows binds the device to its identity on the network; and the second key, mathematically related to first is used by the network to verify that identity. This enables device identification to be performed rapidly, safety, and in cryptographically strong manner.

This property is useful for number of things – it greatly simplifies key exchange, as one example and it solves one critical problem secret-key cryptography (SKC) cannot solve – the problem of guaranteeing unique authentication. While personal computers have no computing limitation to implement well-known PKC algorithms such as RSA (Rivest-Shamir-Adleman), or Digital Signature Algorithm (DSA), WSNs nodes cannot use them due to constrains of used low cost and low power MCUs.

Elliptic Curve Cryptography (ECC) [24] offers secure and efficient alternative solutions for WSNs. ECC offers considerably greater security for given key size comparing to RSA. That smaller key size also makes possible much more compact implementations for a given level of security, which means faster cryptographic operations running on a smaller chips or more compact SW. This means less heat production, and less power consumption-all of which is of particular advantage in constrained WSN nodes.

An Elliptic Curve version of Menezes-Qu-Vanstone (ECMQV) protocol [24] was proposed as the key establishment mechanism and it may be suitable for ZigBee [25]. ECMQV is an efficient public-key agreement scheme that offers key authentication and key establishment. Like AES, ECMQV is fast, strong and could be inexpensively implemented in HW. In addition, by using elliptic curve methods, key sizes will be kept small even as security needs increase.

### D. Suitable HW platform

Implementation of ECC primitives in low performance MCU have been richly discussed in [26], [27], [28]. Computation of Elliptic Curve Digital Signature Algorithm (ECDSA) takes around 1.6 s in an 8-bit platform while the same computation takes around 100 ms in 32-bit platform that seems it can be usable in WSNs nodes. Moreover, Freescale announced the MC13225 chip, the composite of IEEE 802.15.4 radio and ARM7TDMI processor. This chip has only 20 mA current consumption in RX or TX mode, what ensures very long battery life for ZigBee end-device with appropriate active mode duty cycle. In addition this chip contains HW acceleration for both the IEEE 802.15.4 MAC and AES security and full set of MCU peripherals such as dual 12-bit analog-to-digital converter (ADC) or multiple serial channels. This chip do not need any passive matching parts to connect an antenna. The MC13225 chip, external crystal, onboard antenna, battery and optional sensor are all what application needs. In other words, using mentioned chip, PKC in WSN can became reality.

## IV. Experimental Results

Only the interoperability of SMAC based protocol between various MCUs was tried up to now. The SMAC is distributed in open (ANSI-C functions) form by Freescale for their HCS08 and Coldfire MCUs. This protocol was ported on ARM and x51 compatible microcontroller. A simple routing algorithm was written as extension for the SMAC. This routing algorithm allows star network building with 10,000 end devices.

A simple HA network (Fig. 2) was crated for interoperability demonstration. This network allows remote light switching, temperature regulation or detects door and window movements by accelerometers. There are Passive InfraRed (PIR) motion detector and Smoke Detector as well. Two options to visualize the network in computer are either Universal Serial Bus (USB) connection of coordinator or Ethernet connection provided by Freescale M52233DEMO board with web-server SW. Coverage radius of this network was about 40 m, what is enough for

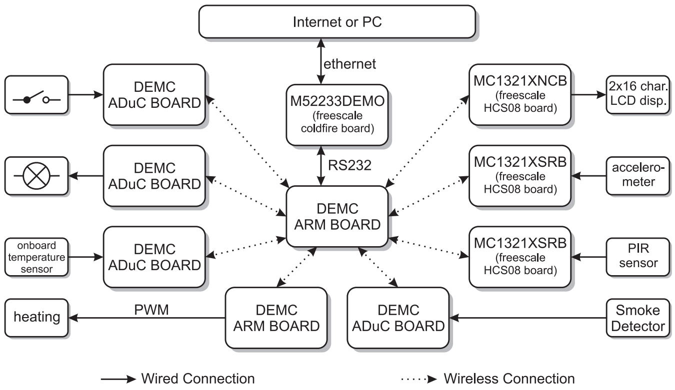



Fig. 2. Realized SMAC based Home Automation (HA) experimental network for testing hardware functionality of designed boards, HA sensors and actuators with proprietary routing algorithm and with Internet connection.

TABLE I
SMAC PROTOCOL AND ROUTING ALGORITHM MEMORY REQUIREMENTS OF VARIOUS MCUS

| MCU | HCS08 | ColdFire |
|---|---|---|
| SMAC ROM | 4491 B | 9512 B |
| Routing ROM | 462 B | 1608 B |
| SMAC RAM | 12 B | 197 B |
| Routing RAM | 413 B | 152 B |

flat or small house. A SMAC memory requirement for various used microcontrollers is shown in Table I.

## V. CONCLUSION

Paper describes implementation proprietary SMAC protocol by means of various MCU platforms. Simple routing protocol was created in order to a build simple HA network. It was compared resource requirements of implementation for both representative of Freescale family – HCS08 and ColdFire.

While ZigBee is a modern and powerful standard to creating WSN, its security lies upon AES and protocols such SKKE. However, while SKC has low requirements for processing power, it probably can not provide enough robustness and security in large scale networks. On the other hand, PKC provides availability of authentication and key exchange mechanisms that are more secure and reliable compared to SKC. Besides these advantages, the PKC has also one main disadvantage – it is computationally expensive. There are two strong limitations to implement PKC – first, to keep messages as short as possible because of each bit transmitted consumes about as much power as executing 800-1000 instructions [29], and as a consequence, any message expansion caused by security mechanisms comes at significant cost; and second, using low performance MCU is necessary in order to develop ever-cheaper sensor nodes. It is nowadays clear that it is possible to apply ECC based PKC, but the question that remains is how the application of strong public key cryptography affects the lifetime of the energy source and thus the lifetime of the sensor. This is why we would like to target next research to observe costs of PKC protocols in real low cost WSN platform (e.g. MC13225) and compare with the SKC ones in the terms of resource requirement, speed, network security, and global performance.

## ACKNOWLEDGMENT

This work has been done in the frame of the Slovak scientific projects VEGA 1/4088/07 of Slovak Ministry of Education and the grant – EU 6th Framework Programme, IST, MonAMI – Mainstreaming on Ambient Intelligence 035147. The MonAMI project is focused on Home Automation for elderly persons and persons with disabilities living at home (www.monami.info). Authors also thank to Freescale, Semiconductor, Inc. for providing software and hardware development tools used within this project.

## REFERENCES

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and CayirciE., “Wireless sensor networks: A survey,” *Computer Networks*, vol. 38, no. 4, pp. 393–422, 2002.
[2] A. Perrig, R. Szewczyk, V. Wen, D. Culler, and D. Tygar, “Spins: Security protocols for sensor networks,” in *Seventh Annual International Conference on Mobile Computing and Networks*, July 2001.
[3] D. Liu, P. Ning, and R. Li, “Establishing pairwise keys in distributed sensor networks,” *ACM Transactions on Information and System Security*, vol. 8, no. 1, pp. 41–77, February 2005.
[4] e. a. Burne, R., “Self-organizing cooperative sensor network for remote surveillance: improved target tracking results,” in *Proceedings of the SPIE*, 2001.
[5] B. Fulford, “Sensors gone wild,” Forbes Global, 2002, online. http://www.forbes.com/business/forbes/2002/1028/306.html.
[6] J. Polastre, “Design and implementation of wireless sensor networks for habitat monitoring,” Master’s thesis, University of California at Berkeley, 2003.
[7] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, “Wireless sensor networks for habitat monitoring,” in *First ACM Workshop on Wireless Sensor Networks and Applications*, 2002.
[8] P. Saffo, Homepage, online. www.saffo.com.
[9] G. Gaubatz, J. Kaps, and B. Sunar, “Public keys cryptography in sensor networks - revisited,” in *The Proceedings of the 1st European Workshop on Security in Ad-Hoc and Sensor Networks (ESAS)*, 2004.
[10] N. Gura, A. Patel, A. Wander, H. Eberle, and S. S. C., “Comparing elliptic curve cryptography and rsa on 8-bit cpus,” in *Proceedings of the Workshop on Cryptography Hardware and Embedded Systems (CHES)*, August 2004.
[11] D. J. Malan, M. Welsh, and S. M. D., “A public-key infrastructure for key distribution in tinyos based on elliptic curve cryptography,” in *The First IEEE International Conference on Sensor and Ad Hoc Communications and Networks*, October 2004, pp. 71–79.
[12] W. Du, R. Wang, and N. P., “An efficient scheme for authenticating public keys in sensor networks,” in *MobiHoc’05*, 2005, pp. 58–67.
[13] Atmel, Offcial Site, online. www.atmel.com.
[14] Ember, Offcial Site, online. www.ember.com.
[15] Freescale, Offcial Site, online. www.freescale.com.
[16] Jennic, Offcial Site, online. www.jennic.com.
[17] Microchip, Offcial Site, online. www.microchip.com.
[18] Nordic, Offcial Site, online. www.nordicsemi.com.
[19] T. Instruments, Offcial Site, online. www.ti.com.
[20] ZigBee Alliance, Official Site, online. www.zigbee.com.
[21] IEEE, “Ieee std 802.15.4-2003,” online. http://standards.ieee.org.
[22] NIST, *Federal Information Processing Standards Publication 197 – Advanced Encryption Standard*, October 2001, online. http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf.
[23] M. J., “Zigbee: A long way to go?” online. http://www.sciencedirect.com/.
[24] D. Hankerson, J. Menezes, and S. Vanstone, *Guide to elliptic curve cryptography*. New York: Springer-Verlag, 2004.
[25] M. Blaser, “Industrial-strength security for zigbee: The case for public-key cryptography,” *Embedded Computing Design*, May 2005.
[26] R. Roman and C. Alcatraz, “Applicability of public key infrastructures in wireless sensor networks,” 2007.
[27] M. Drutarovsky and M. Varchola, “Cryptographic system on a chip based on actel arm7 soft-core with embedded true random number generator,” in *Workshop on Design and Diagnostics of Electronic Circuits and Systems (DDECS)*, Bratislava, April 2008, accepted.
[28] M. Aydos, T. Yanik, and C. . K. Koc, “An high-speed ecc-basedwireless authentication protocol on an arm microprocessor,” in *The 16th Annual Computer Security Applications Conference*, 2000, pp. 401–409.
[29] J. Hill, R. Szewczyk, A. Woo, S. Hollar, D. Culler, and K. Pister, “System architecture directions for networked sensors,” in *Proceedings of ACM ASPLOS IX*, November 2000.

# Hybrid concealment mechanism

*Ján MOCHNÁČ*

Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

jan.mochnac@tuke.sk

***Abstract*—Video transmission over noisy channel leads to errors on video, which degrades the visual quality noticeably and makes error concealment an indispensable job. One approach that is especially suited for applications with real-time constraints, limited bandwidth, or multicast distribution, is to conceal the information loss at the receiver. In this paper I focused on hybrid concealment mechanism. Effectiveness of this algorithm was tried on several video sequences and the results are quite satisfying.**

***Keywords*—error concealment, block, video sequence, motion, videoconference.**

## I. Introduction

The base problem with packet communications system is information lost due to network congestion. Thus, with growth popularity of the Internet, error concealment becomes more important. The main goal of error concealment is to modify received information so that losses will be imperceptible for human eyes.

There have been many techniques proposed in the literature that conceal information loss from different angles. These methods can be grouped into: 1) forward error concealment [1], 2) interactive error concealment [1], 3) error concealment by postprocessing [1]. Each group of methods has it advantages and disadvantages. Examples of forward error concealment include FEC [2], joint source and channel coding, and layered coding. These techniques add a certain amount of redundancy at the source coder or transport coder. The main advantage of this group is good error resilience. We pay for it with increase in the transmission bandwidth. Examples of interactive error concealment include selective encoding [4] and retransmission without waiting [5]. This class of methods should give the best performance if backward channel from the decoder to encoder exist. The main disadvantage is significant increase of the transfer delay. In the third group, error concealment is performed by decoder. All postprocessing techniques make use of the correlation between a damaged macroblock and its spatial/temporal adjacent macroblocks [1], so no addition redundancy is needed. Resulting from these facts, it may not be possible to completely reconstruct the lost image blocks.

The loss concealment mechanism described in this paper fall into third group. Hence, it uses spatial and temporal redundancy presented at video signal. It is hybrid in the sense, that it masks a lost block using information from both the previous video frame (temporal concealment), and from the surrounding blocks of the current frame (spatial concealment), depending on the presence of motion in the area of the frame. Temporal concealment conceals loss of block with corresponding block from previous video frame. Spatial concealment expresses each lost pixel as a distant-dependant linear combination of all surroundings pixels [3]. This operation is performed for entire lost block with a single matrix multiplication. Performance of algorithm will be shown on several video sequences. The results of this algorithm are satisfactory for still and spatial smooth video sequences. In the areas with significant motion algorithm produce slightly blurred results.

## II. Hybrid Concealment Mechanism

A hybrid concealment mechanism was introduced in [3] at first. In the next, I have used algorithm description from [3].

Let F be the video frame on which I perform loss concealment, and $\mathbf{F_p}$ be the previous video frame, in which losses have been already concealed. Let **X** be a missing block in **F** and **E** the estimate of **X** that results from the loss concealment. **P** is the block of $\mathbf{F_p}$ which resides in the same location as **X**. The adjacent blocks with **X** are **U** for up, **D** for down, **L** for left, and **R** for right. Some of the blocks **U**, **D**, **L**, **R** may be also lost, without having been concealed yet. In that case, the missing adjacent block is temporarily replaced with the corresponding block from $\mathbf{F_p}$.

The described spatial concealment technique uses the pixels that are directly adjacent to **X** and that make up the four 8-by-1 surrounding vectors **u**, **d**, **l**, **r** (Fig. 1). If block **X** resides at the boundary of the image, vectors are predicted based on the closest known pixels. For example, if block **X** is at the top of the frame (but not at the corners), **u** is predicted as **u(i) = (l[1]+r[1]) / 2** for **i=1..8**. If it is at the top-left corner, **u** is predicted as **u(i) = r[1]** and **l** as **l(i)=d[1]** for **i =1, .., 8**.

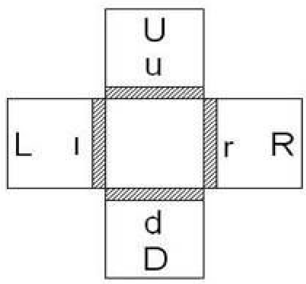



Fig. 1. The missing block X, its adjacent blocks U, D, L, R, and the corresponding surrounding columns u, d, l, r.

If motion is detected in one of the adjacent blocks **U**, **D**, **L**, **R**, lost block is masked using a smooth estimate $\mathbf{X}^S$ of **X** computed based on the adjacent vectors **u**, **d**, **l**, **r** (spatial concealment). Otherwise, if there is no significant motion in that area of **X**, lost block is replaced with the corresponding block **P** (temporal concealment).

### *A. Spatial Concealment*

Let **x** be a pixel of the spatially concealed block $\mathbf{X}^S$, and $\mathbf{x_u}$, $\mathbf{x_d}$, $\mathbf{x_l}$, $\mathbf{x_r}$ be the up, down, left, and right pixels, respectively (Fig. 2). It is required that for all pixels **x** of $\mathbf{X}^S$:

$$x = \frac{x_u + x_d + x_l + x_r}{4}$$

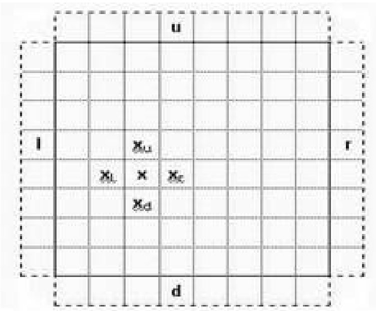



Fig. 2. Spatial concealment.

For an 8-by-8 block 64 such averaging equations can be written, one for each pixel **x**. The resulting system of linear equations has 64 unknowns, the pixels of **X**, while the constant terms depend on the surrounding vectors **u, d, l, r**. These 64 equations can be rewritten as a linear system:

$$\mathbf{A} \,.\, \mathbf{v} = \mathbf{c}$$

where the matrix **A** is the 64-by-64 block matrix:

$$\begin{bmatrix} L & -I_8 & 0 & 0 & 0 & 0 & 0 & 0 \\ -I_8 & L & -I_8 & 0 & 0 & 0 & 0 & 0 \\ 0 & -I_8 & L & -I_8 & 0 & 0 & 0 & 0 \\ 0 & 0 & -I_8 & L & -I_8 & 0 & 0 & 0 \\ 0 & 0 & 0 & -I_8 & L & -I_8 & 0 & 0 \\ 0 & 0 & 0 & 0 & -I_8 & L & -I_8 & 0 \\ 0 & 0 & 0 & 0 & 0 & -I_8 & L & -I_8 \\ 0 & 0 & 0 & 0 & 0 & 0 & -I_8 & L \end{bmatrix}$$

$\mathbf{I_8}$ is the 8-by-8 identity matrix and **L** is the Laplacian-like 8-by-8 matrix:

$$\begin{bmatrix} 4 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 4 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 4 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 4 \end{bmatrix}$$

The vector of unknowns **v** is the pixels of $\mathbf{X}^S$ in anti-lexicographic order (the anti-lexicographic order of a matrix is to first read all the elements of the first column, then of the second column, and so on). The constant term **c** is the 64-by-1 vector that results from the anti-lexicographic order of the matrix:

$$\begin{bmatrix} l_1+u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & r_1+u_8 \\ l_2 & 0 & 0 & 0 & 0 & 0 & 0 & r_2 \\ l_3 & 0 & 0 & 0 & 0 & 0 & 0 & r_3 \\ l_4 & 0 & 0 & 0 & 0 & 0 & 0 & r_4 \\ l_5 & 0 & 0 & 0 & 0 & 0 & 0 & r_5 \\ l_6 & 0 & 0 & 0 & 0 & 0 & 0 & r_6 \\ l_7 & 0 & 0 & 0 & 0 & 0 & 0 & r_7 \\ l_8+d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & r_8+d_8 \end{bmatrix}$$

The matrix A is invertible, so pixels of $\mathbf{X}^S$ can be directly computed as:

$$X^S = A^{-1}c$$

The inverse of **A** can be a priori calculated, and so the spatial concealment operation is just a matrix-vector multiplication for each lost block.

### *B. Temporal Concealment*

Motion in each of the surrounding blocks U, D, L, R is detected using the Pixel Difference Classification (PDC) method. The PDC distortion function compares each pixel of the target block with its counterpart in the candidate block and classifies each pixel pair as either matching or not matching. Pixels are matching if the difference between their values is less than some threshold:

$$\sum_{q=1}^{n}\sum_{p=1}^{m}[ord(|A[p,q]-B[p,q]| \geq T)]$$

Ord(e) evaluates to 1 if e is true and false otherwise.
The value of **T** has to be chosen based on the noise level of the image. A value around 10 lead to accurate motion detection most of the times.
Given two same-location blocks **A** and **B** from frames **F** and **Fp**, it is detected that there is motion present in the area of these blocks, and so it is switched to spatial concealment, if

$$PDC(A,B) > M_T$$

where $\mathbf{M_T}$ is the motion threshold. The value of $\mathbf{M_T}$ can be determined through experimentation, and it certainly depends on the specific video streams. Values around 16 to 24 (for 8-by-8 blocks) lead to the best results.

## III. Simulation

Effectiveness of introduced algorithm was tried on Foreman and Mobile video sequences. Simulations were performed by using MATLAB 7.0. Losses in video were generated by using of second order Markov model. After that, losses were concealed by described algorithm.

Fig. 3a shown 2nd frame of Foreman sequence, fig. 3b shown frame after random block losses and the frame resulting from the hybrid concealment mechanism is shown on fig. 3c. At the start of the video sequence is motion caused by vibration of video camera, also by move of man’s head and by his mimic. At the end of the video sequence is motion caused by move of camera from left to right. At these frames we can see that algorithm leads to very good concealment in the areas where motion is not noticeable, what was expected. In the areas where motion was detected, algorithm mask lost block with blurred block. It is result of spatial concealment.

But in the video sequence we can very hardly recognize these areas. Thus, I evaluate the effectiveness of the hybrid concealment

Fig. 3a. 2nd frame of Foreman sequence.

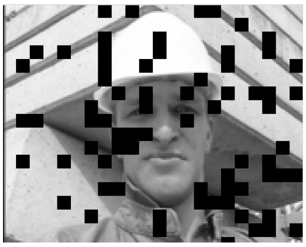


Fig. 3b. 2nd frame of Foreman sequence after random block losses.

Fig. 3c. 2nd frame of Foreman sequence resulting from the hybrid concealment mechanism.

performed on the Foreman video sequence as satisfactory and quality measures (Tab.1) confirm my opinion.

Error concealment in the Mobile video sequence is really challenging for all error concealment algorithms. At the bottom of this video sequence is train moving from the right to the left and in addition it pushes a ball, which, of course, rotates. Also, there is a calendar, which is moving to the right, above the train. In addition camera movement is not only from the right to the left but also from down to up. Very hard task is correctly concealing errors, which occurs in the area of numbers.

As we can see on the fig. 4a – 4c, algorithm failed in concealment of numbers. Some numbers we can not recognize, it is an effect of spatial concealment, which mask these errors with smooth estimation. Subjective perception confirm also quality measures (Tab. 2)

TABLE I
QUALITY MEASURES, FOREMAN, FRAME NO. 2

| | MSE | MAE | NMSE | SNR |
|---|---|---|---|---|
| Foreman, frame no.2 | 0.0012419 | 0.0095642 | 0.0026508 | 25.766 |

TABLE II
QUALITY MEASURES, MOBILE, FRAME NO. 50

| | MSE | MAE | NMSE | SNR |
|---|---|---|---|---|
| Mobile, frame no. 50 | 0.0026743 | 0.01306 | 0.0094311 | 20.2544 |

Fig. 4a. 50th frame of Mobile sequence.

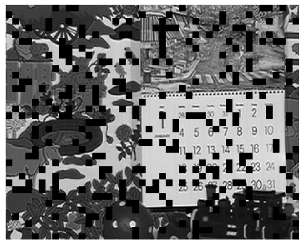


Fig. 4b. 50th frame of Mobile sequence after random block losses.

Fig. 4c. 50th frame of Mobile sequence resulting from the hybrid concealment mechanism.

## IV. CONCLUSION

Hybrid concealment mechanism described in this paper could be used in such applications as videoconference, thus, applications which have not very high quality requirements. That can we see in Foreman video sequence, where hybrid concealment provides good results. On the contrary, for applications with high quality requirements, such as IPTV, where we can tolerate some delay or channel bandwidth is not a limit, it is better to use forward error correction techniques or interactive error concealment.

## ACKNOWLEDGEMENT

The work presented in this paper was supported by Grand of Ministry of Education and Academy of Science of Slovak Republic VEGA under Grand No 1/4088/07.

## REFERENCES

[1] Y. Wang, Q.-F. Zhu, Error Control and Concealment for Video Communication: A Review, *Proc. IEEE*, vol. 86, no. 5, May 1998, pp. 974-997.
[2] E. Ayanoglu, R. Pancha, A. R. Reibman, S. Talwar, "Forward error control for MPEG-2 video transport in a wireless ATM LAN," *ACM/Baltzer Mobile Networks and Applications,* vol. 1, no. 3, December 1996, pp. 245-258.
[3] C. Dovrolis, D. Tull, P. Ramanathan, Hybrid Spatial/Temporal Loss Concealment for Packet Video, *Proceedings of the 9th International Packet Video Workshop (PVW '99)*, April 1999.
[4] W. Wada, "Selective recovery of video packet loss using error concealment," *IEEE J. Select. Areas Commun*., vol. 7, pp. 807-814, June 1998
[5] M. Ghanbari, "Postprocessing of late cells for packet video," *IEEE Trans. CAS for Video Tech.*, vol. 6, no. 6, pp. 669-678, Dec., 1996.

# Small Solar Charger With Maximal Power Point Tracking

*Marek PÁSTOR*

Dept. of Electrical Engineering, Mechatronics and Industrial Engineering, FEI TU of Košice, Slovak Republic

TE050987@tuke.sk

***Abstract*— This work presents basic principles of small NiCd/NiMH accumulators solar charger with Maximal Power Point Tracking (MPPT). A real model is also built. The basic part of charger is SEPIC (Single Ended Primary Inductance Converter). This step-up step-down DC/DC converter ensures possibility of charging accumulators with various nominal voltage. The charger is suitable for NiCd/NiMH accumulators with nominal voltage from 1,2 to 18V. Maximal charging current is limited to 1900mA and is set according to nominal capacity of the accumulator. DC/DC converter and charging process is controlled by microcontroller.**

***Keywords*— charger, Maximal Power Point Tracking (MPPT), photovoltaic panel, SEPIC (Single Ended Primary Inductance Converter).**

## I. Introduction

Electrical energy generated by photovoltaic technology for accumulator charging is very often used in nowadays off-line chargers. The simplest solar charger consists only of a photovoltaic (PV) panel and some wires. This is a small portable solar system with output power of several watts. On the other hand, large solar arrays are much more sophisticated. Maximal Power Point Tracking (MPPT) is the essential part of these solar systems with output power of several hundreds watts. MPPT can significantly improve the output power of a solar system independently on the amount of available power. MPPT allows to extract the maximum power from a PV panel. Why do not use MPPT in a small solar charger? This work tries to create such type of solar charger. In general this solution gives possibilities for charging of any accumulator with any photovoltaic panel (there are some limitations).

## II. Problem overview

A photovoltaic (PV) panel is the only source of energy for a solar charging system. A PV panel directly transforms sunlight into electricity. The output power of a PV panel is determined by many factors. First, the solar irradiance: higher solar irradiance means more input energy, which is converted by PV panel into output electric energy with some efficiency. We cannot change solar irradiance at given conditions without concentration of sun rays. However, the output power of a PV panel is also dependent on the output voltage of the PV panel, which we can change. The simplest way to change the output voltage of a PV panel in connection with a solar charger is to directly connect an accumulator to the PV panel. But this is definitely not the best way for any given accumulator.

Fig. 1 shows the basic parts of a solar charger which I designed. A solar charger consists of a PV panel, a DC/DC converter with charging control and an accumulator. Each part of the system has some efficiency. Due to low efficiency of commercially available PV panels (maximum approx. 20%) the overall efficiency of the system is less than 20%.


Fig. 1. Block scheme of solar charger.

Fig. 2 shows the characteristic power and volt-ampere curves of a PV panel. These describe attributes of the PV panel model, created from silicone diodes, which were used during measurements on the built model of the solar charger. From the power curve one can see that there is a point of maximal output power (MPP) at some output voltage ($U_{MPP}$). For this particular PV panel model it is $U_{MPP} = 6{,}4V$. The volt-ampere and power curves depend on several conditions: solar irradiance, temperature and, of course, type of the PV panel. Mathematical equations which describe these relations are not important for the design of a solar charger. The important things are only that the PV panel has the MPP and the voltage $U_{MPP}$ is not constant. If a PV panel has to work in the MPP, its output voltage must be equal to the $U_{MPP}$.

Now let us imagine that the accumulator is directly connected to the PV panel (see Fig. 2) and the voltage of the accumulator is equal to $U_{accu} = 2V$. From the power curve it can be seen that the operation point of the system created from the PV panel and the accumulator is far away from the MPP. The output power will be only about 0,6W. But the maximal output power at given irradiance in this case is 1,75W, which is the MPP of this PV panel. The power 1,15W is not generated because of non-adapted voltage of the accumulator. This PV panel is suitable only for charging accumulators with voltage approximately equal to 6,4V. Accumulator with a different voltage will not be charged with maximal power and accumulator with a voltage higher than 7,8V, which is the open circuit voltage of this PV panel, will not be charged at all. These are the basic disadvantages of a direct coupling of an accumulator to a PV panel.

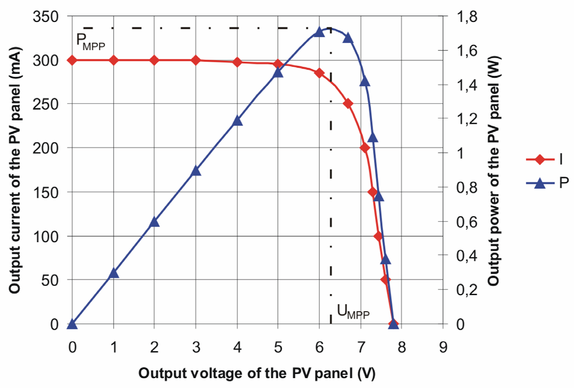



Fig. 2. Characteristic power and volt-ampere curves of PV panel.

## III. DC/DC Converter

### A. Connection between a PV panel and an accumulator

The essential part of a solar charger with MPPT is the DC/DC converter. This section describes the purpose of this converter in the solar charger. General connection between a PV panel and an accumulator can be described as a two-port network (or quadrapole) as is shown in Fig. 3. Each quadrapole can be described by cascade (chain) parameters which define the relation between output and input of the quadrapole. Cascade equations in the matrix form are:

$$\begin{bmatrix} U_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} U_2 \\ I_2 \end{bmatrix} \tag{1}$$

In connection with Fig.3, $U_1$ is the output voltage of a PV panel, $I_1$ is the output current of a PV panel, $U_2$ is the voltage of an accumulator and $I_2$ is the charging current of an accumulator.

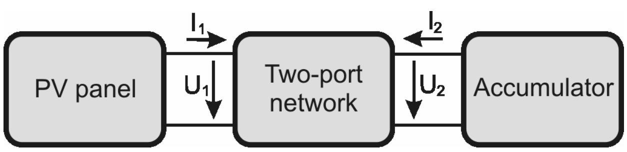



Fig. 3. Connection between a PV panel and an accumulator with general two-port network.

The goal is to find the cascade parameters A, B, C, D which ensure that for a given $U_2$ the voltage $U_1$ will be equal to the $U_{MPP}$ of the PV panel. In this case the PV panel will operate in the MPP and will supply maximal energy to the two-pole network. The two-pole network will transfer this energy with some efficiency to the accumulator. If the two-pole network has approximately constant efficiency, the charging current will be maximal (see Fig. 7.).

If the two-pole network represents direct connection between the PV panel and the accumulator with lossless wires, the cascade equations are following:

$$\begin{bmatrix} U_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} U_2 \\ I_2 \end{bmatrix} \tag{2}$$

Equation (2) means that the voltage of the accumulator is the output voltage of the PV panel and the output current of the PV panel, set according to the volt-ampere characteristic of the PV panel (Fig. 2.), is the charging current of the accumulator. There is no possibility to adapt the accumulator voltage to the PV panel because the cascade parameters are constants.

The DC/DC converter ensures that the cascade parameters are variables. An extrapolation of these parameters is not simple due to absence of a mathematical model of the DC/DC converter in my work. The cascade parameters change automatically with changes in the duty cycle.

### B. SEPIC

SEPIC (Single Ended Primary Inductance Converter) is step-up, step-down DC/DC converter. The scheme of SEPIC is shown in Fig. 4. The output voltage of SEPIC is controlled by the duty cycle of the transistor (MOSFET). Main advantages of this converter are: step-up, step-down possibilities, non-inverted output voltage, only one control component, high efficiency, negligible overshoots compared to flyback and no problem with leakage inductance. The basis of SEPIC is a boost converter which ensures step-up possibilities of SEPIC [4]. Design equations and components ratings can be found in [3], [4].

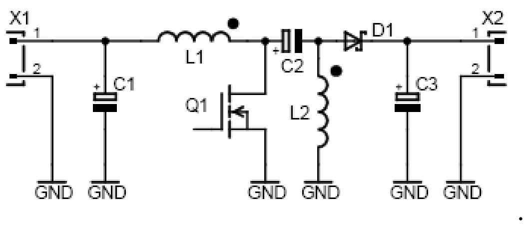



Fig. 4. Scheme of SEPIC.

The main advantage of SEPIC in solar charger is the fact that the output voltage of the converter can be greater than, equal to or less than the input voltage and can be controlled by only one component. There is no limitation in output voltage. The 1,2V accumulator can be charged with the same charger as well as the 18V accumulator.

### C. MPPT

Maximal Power Point Tracking (MPPT) ensures that a PV panel operates at maximal output power. MPPT means tracking of the operation point of a PV panel. In large PV systems there is also other tracking which involves following the Sun in the sky and ensures the optimal angle of sun rays. MPPT has nothing to do with a PV panel moving. MPPT just “moves” the operation point of the PV panel. If there is a demand for a maximal output power, MPPT is essential. MPPT has existed for several decades and many methods have been developed. However, there are several MPPT methods which are more common than others[1]:

1. Hill-Climbing,
2. Perturb and Observe,
3. Increment Conductance,
4. Fractional Open-Circuit Voltage and Current,
5. Fuzzy Logic Control,
6. Neutral Network.

A comparison of PV array MPPT techniques can be found in [1]. In my opinion the above methods are more suitable for large PV arrays or plants. Some of the techniques need many sensors or sophisticated control circuit. But there is one method which is very suitable for small solar chargers: Load Current or Load Voltage Maximization. The first variant is suitable for voltage-source type of load, the second one for current-source type of load. The voltage of an accumulator is not constant during charging but the accumulator can be considered as voltage-source type of load. If the DC/DC converter ensures that the output power of the PV panel is maximal, then for a given accumulator voltage the charging current will be maximal. Load Current Maximization is a very suitable MPPT method for solar chargers. System with Load Current Maximization may not operate at MPP [1] because there is no information about the output power of the PV panel. Only the output current of the DC/DC converter is available. However, if for the immediate given conditions the charging current is maximal, in my opinion this MPPT method is not inappropriate.

In Fig. 5. typical power curves of a solar charger with SEPIC are shown. The efficiency of the DC/DC converter in connection with Fig. 3. is defined as follows:

$$\eta_1 = \frac{P_2}{P_1} = \frac{U_2 \cdot I_2}{U_1 \cdot I_1} \tag{3}$$

Equation (3) describes the relation between the immediate input and output power. There is also other very important efficiency in connection with MPPT described below.

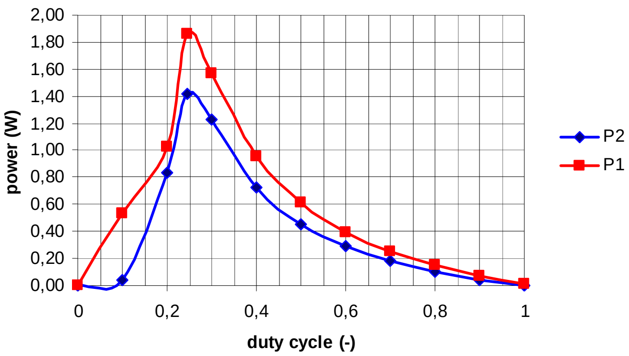



Fig. 5. Typical relation between duty cycle and input ($P_1$)/output ($P_2$) power of SEPIC in connection with a PV panel.

The relations between the duty cycle and the input/output current of the DC/DC converter are shown in Fig. 6. The output current $I_2$ (charging current) has a clear maximum. This happens when output voltage of the PV panel is equal to $U_{MPP}$ as can be seen in Fig. 6.

### *D. Working of MPPT*

The basic principle of the MPPT method called Maximization of Load Current is to change the duty cycle of the control transistor (Fig. 4.) according to changes of the output (in this case charging) current. This algorithm is described in Table I.

Fig. 7. illustrates the MPPT method Maximization Load Current. Let us imagine that the operation point is in A and the duty cycle is being increased. The operation point will be moving to B. After several iterations it reaches approximately the MPP (the current will be maximal). The control logic stays in mode of increasing of the duty cycle because up to the moment the current was increasing. But increasing of the duty cycle when operation point is in B will cause decreasing of the current. According to Table I. the next change of the duty cycle must be negative. The operation point is now moving from C to B. The converter approximately reaches B, the operation point will be moving to A and the whole cycle is repeated.

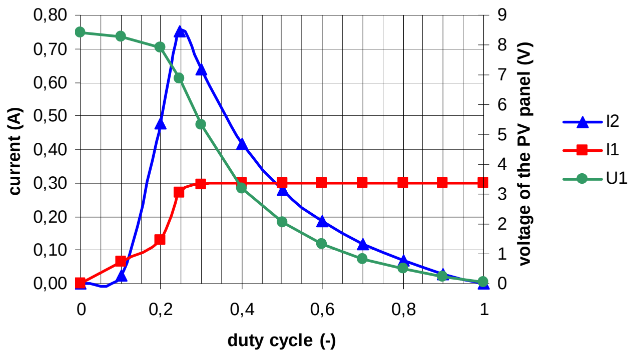



Fig. 6. Typical relations between duty cycle and input ($I_1$)/output ($I_2$) current of SEPIC and output voltage of the PV panel ($U_1$).

TABLE I
Summary of Maximization Load Current algorithm

| Change in duty cycle | Change in changing current | Next change in duty cycle |
|---|---|---|
| positive | positive | positive |
| positive | negative | negative |
| negative | positive | negative |
| negative | negative | positive |

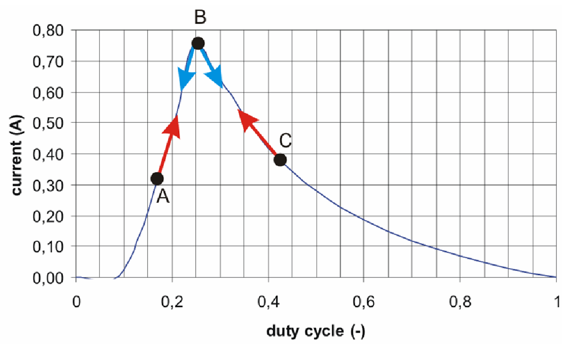



Fig. 7. Algorithm of Maximization Load Current.

Due to the continuing oscillation of the operation point this is not an ideal MPPT method. The efficiency of MPPT is defined as follows [2]:

$$\eta_{MPPT} = \frac{\int_0^t p_{actual}(t)dt}{\int_0^t p_{\max}(t)dt} \tag{4}$$

where $p_{actual}(t)$ is the real power produced by the PV panel in connection with MPPT and $p_{max}(t)$ is the maximal possible power which could be obtained if the PV panel has been working in MPP all the time (no oscillations).

The overall efficiency of the DC/DC converter consists of the efficiency of the DC/DC converter $\eta_1$ (3) and the efficiency of MPPT $\eta_{MPPT}$ (4):

$$\eta_{overall} = \eta_1 \cdot \eta_{MPPT} \tag{5}$$

## IV. Accumulator

Theory of accumulators charging is not described in this paper. NiCd/NiMH accumulators prefer constant current charging method. Charging current can be as high as 1C. This is done by current limit set in charger according to accumulator capacity. In general it is difficult to terminate charging in a solar charger when an accumulator is charged. Solar irradiance and output power of a PV panel are variables depending on weather. Charging current is changing over time and charging characteristic of accumulator has no typical constant current type. Charging termination methods such as Voltage Plateau, Timed Charge, Temperature Cutoff and others [5] cannot be used. It is impossible to detect completed charging by measurements of accumulator voltage or temperature. Accumulator voltage varies due to the changes in current not state of charge. Current version of control software uses Coulomb Counting method; Voltage and Temperature Cutoff will be added as backup methods.

## V. Real system

The built model of the solar charger consists of SEPIC and microcontroller with auxiliary circuits. The charger is controlled by AVR microcontroller ATtiny861. Operational frequency of the DC/DC converter is 125 kHz. The charger is designed for charging NiCd/NiMH accumulators with nominal voltage from 1,2V to 18V. Maximal charging current is 1,9A and is set according to the capacity of an accumulator. The final version of the charger will be capable of charging two 1,2V cells or one battery with maximal nominal voltage of 18V. User interface consists of the LCD and two buttons.

The efficiency of SEPIC in this solar charger depends on the input power and the output voltage. Fig. 7. shows the efficiency $\eta_1$ and the output current of the charger ($I_2$) at the output voltage ($U_2$) 7V. The charger is supplied by the PV panel shown in Fig. 2. The current version of the charger needs additional supply for control logic, so $\eta_1$ in Fig. 7. is the efficiency of SEPIC, not the whole charger.

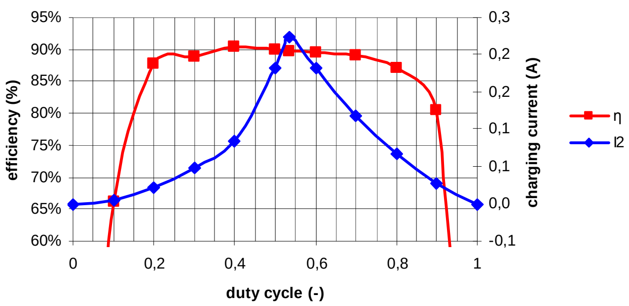



Fig. 7. Efficiency of SEPIC at output voltage 7V and charging current as function of duty cycle.

Fig. 8. shows current model of solar charger with MPPT. All previous characteristics are obtained by measurement on this model.

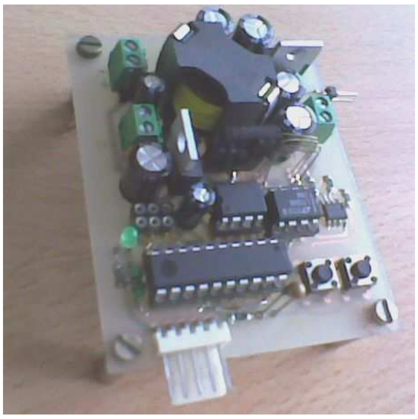


Fig. 8. Current model of solar charger without display

Table II describes the difference in supplied energy between direct coupling of the PV panel and load and coupling through SEPIC with MPPT. When the load voltage (accumulator voltage) is close to the $U_{MPP}$ the system has lower efficiency.

TABLE II
Supplied energy without and with MPPT controller

| U2 (V) | T (min.) | Energy (mWh) | | Difference |
|---|---|---|---|---|
| | | direct coup. | MPPT | |
| 1,8 | 10 | 89 | 211 | 137% |
| 5 | 10 | 238 | 260 | 9% |
| 7 | 10 | 312 | 267 | -14% |
| 9 | 10 | 0 | 273 | NA |

The final solar charger will be extended and improved towards higher efficiency, user comfort and no need of additional supply.

## Acknowledgment

Author gratefully thanks to supervisor Assoc. Prof. Ing. Jozef Ondera, CSc. from Dept. of Electrical Engineering, Mechatronics and Industrial Engineering, FEI TU of Košice, Slovak Republic.

## References

[1] T. Esram, P. L. Chapman, “Comparison of Photovoltaic Array Maximum Power Point Tracking Techniques,” [Online]. Available: http://energy.ece.uiuc.edu/chapman/papers/EC%202006%20in%20press.pdf

[2] D. P. Hohm, M. E. Ropp, “Comparative Study of Maximum Power Point Tracking Algorithm,” [Online]. Available: http://150.214.178.55/profesores/jaguilar/download/mate_docen/ficheros/lectura3_mppt.pdf

[3] National Semiconductor, Wei Gu , “Designing A SEPIC Converter,” Application Note 1484 [Online]. Available: http://wwwd.national.com/national/PowerMB.nsf/(approved-attachments)/73B027861F24B19E8825737100645E74/$File/AN-1484.pdf.

[4] Maxim Integrated Products, Dallas Semiconductor, “SEPIC Equations and components ratings,” Application Note1015 [Online]. Available: http://pdfserv.maxim-ic.com/en/an/AN1051.pdf

[5] Duracell, “Ni-MH Rechargeable Batteries,” [Online]. Available: http://www.duracell.com/oem/Pdf/others/TECHBULL.pdf

# Auxiliary circuit for PS-PWM DC/DC converter

[1]Vladimír Ruščin, [2]Marcel Bodor

Department of Electrical, Mechatronic and Industrial Engineering, FEI TU of Košice, Slovak Republic

[1]vladimir.ruscin@tuke.sk, [2]bodorm@orangemail.sk

***Abstract*— A novel auxiliary circuit for full-bridge PWM DC/DC converter with controlled secondary side rectifier is presented in this paper. Zero-current turn-on for all power switches of the inverter is achieved for full load range from no-load to short circuit by using controlled rectifier and auxiliary circuit on the secondary side. Phase shift PWM control strategy is used for the converter. The principle of operation is explained and analyzed and the simulation results are presented.**

***Keywords*— auxiliary circuit, soft switching, zero voltage zero current switching.**

## I. Introduction

The conventional phase shifted PWM converters are often used in many applications because their topology permits all switching devices to operate under zero-voltage switching by using circuit parasitics such as power transformer leakage inductance and devices junction capacitance.

However, because of phase-shifted PWM control, the converter has a disadvantage that circulating current flows through the power transformer and switching devices during freewheeling intervals.

This circulating current can be eliminated with the disconnection of the secondary windings, which can be done by using reverse bias application for the output rectifier or using controlled rectifier.

## II. Power scheme of the converter

The DC/DC converter shown in Fig.1 consists of high-frequency inverter, power transformer, controlled output rectifier and output filter. Circuit I. and Circuit II. represent the novel auxiliary circuits that eliminate the turn off looses of the secondary transistors, by taking over the current of the transistor at turn off moment. So the turn off in zero voltage can be achieved. A series combination of diodes and transistors was used to secure reverse capability of the MOSFET transistors.

The converter is controlled by phase shift PWM signals as shown in Fig. 2.

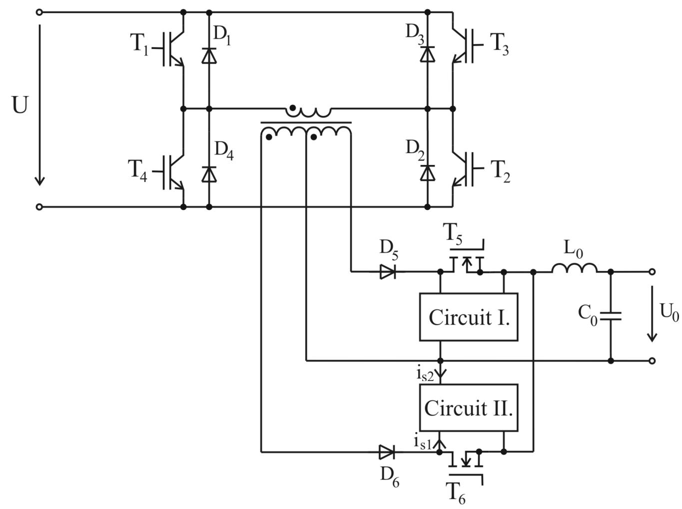



Fig. 1. Power scheme of the DC/DC converter.

## III. Operation principle

The switching diagram and operation waveforms are shown in Fig. 2.

Primary current begins to flow at the moment when the primary IGBTs are turned on ($t_3$). The rising slope both of the primary, as the secondary current is slowed down by the leakage inductance of the power transformer. This is the reason why the primary transistor can turn on in zero current. When the secondary MOSFET is turned off ($t_4$) the primary current begins to sink. At the turn off time the current of the secondary MOSEFT commutates with the auxiliary circuit. This commutation causes zero voltage turn off of the secondary MOSFET transistor. Energy that was obtained in this circuit is forced to the load when the MOSFET transistor turns on ($t_0$). At the time when the IGBT is turned off only a negligible magnetization current flows trough the primary winding of transformer. This causes only low turn off looses.

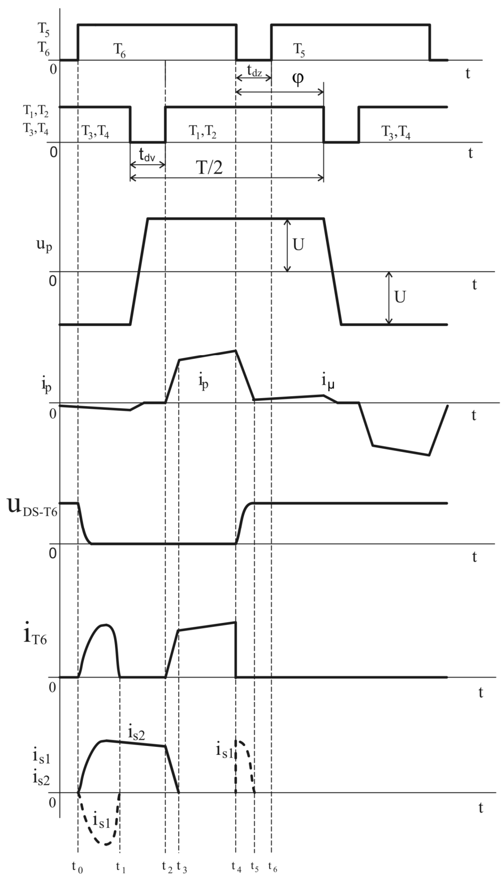



Fig. 2. Operation waveforms of the converter

## IV. SIMULATION RESULTS

A PSpice model of the converter was created to verify the properties of the proposed auxiliary circuit. Following waveforms were obtained at resistive load.

Fig. 3 shows the waveforms of the primary IGBT transistor (transistor voltage $u_{CE}$ and current $i_C$), where the current rise is slowed down with the leakage inductance of the power transformer. Primary current falls down when the secondary transistors are tuned off. IGBT transistor turns off at the interval when only the magnetization current flows trough primary winding of power transformer.

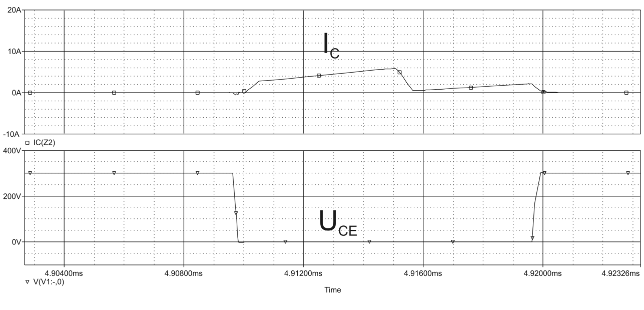



Fig. 3 Primary transistor waveforms.

In Fig. 4 the switching waveforms (Collector-emitter voltage $u_{DS}$ and collector current $i_D$) of the secondary MOSFET transistor are shown. The auxiliary circuit causes a zero voltage transistor turn off. The energy that was obtained at turn off interval is forced to flow trough the load at the turn on interval of this transistor. The value of this energy depends on the leakage energy of the power transformer.

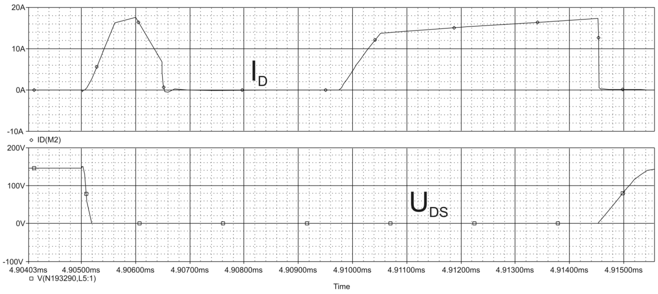



Fig. 4 Secondary transistor waveforms

## V. CONCLUSION

Operation principle of the DC/DC converter with auxiliary circuit that ensures soft switching of all main converter transistors is presented in the paper. For optimal utilization of the auxuliary circuit it is necessary to use a power transformer whose leakage inductance is minimized (planar transformer or coaxial transformer)..

## ACKNOWLEDGMENT

This work was supported by Research and Development Support Agency under project APVV-99-031205.

## REFERENCES


[1] Kim E. S.,Joe K. Y., Kye M. H., Kim Y. H., and Yoon B. D., " An Improved Soft Switching PWM FB DC-DC Converter for Reducing Conduction Losses," *in Record, IEEE PESC'96,* Vol. I., pp. 651-656.

[2] Rinne K. H., Theml K., McCarthy O., " An Improved Zero-Voltage and Zero-Current Switching Full Bridge Converter," *in Record, EPE' 95,* Vol. 2., pp. 725-730.

[3] Lacko M., Olejár M., Ruščin V., Dudrik J.: Converter system for renewable energy utilization with snubber circuit. In: EDPE'07 ; 16-th international conference on Electrical Drivers and Power Electronics, Proceedings: 24-26 September 2007, The High Tatras, Slovakia; ISBN 978-80-8073-868-6

[4] Tereň, A., Feňo, I., Śpánik, P: DC/DC Converters with Soft (ZVS) Switching. In Conf. Proc. ELEKTRO 2001, section - Electrical Engineering. Žilina 2001, Slovakia, pp. 82 – 90

[5] Dudrik, J., Špánik, P., Trip, N.-D. : Zero Voltage and Zero Current Switching Full-Bridge DC-DC Converter with Auxiliary Transformer. IEEE Trans. on Power Electronics, Vol.21, No.5, 2006, pp. 1328 – 1335.

[6] Tereň, A., Feňo, I., Śpánik, P: DC/DC Converters with Soft (ZVS) Switching. In Conf. Proc. ELEKTRO 2001, section - Electrical Engineering. Žilina 2001, Slovakia, pp. 82 – 90.

[7] Ruščin V., Olejár M., Lacko M., Dudrik J.,: ZVZCS DC-DC converter with controlled output rectifier. In: TRANSCOM 2007 : 7-th european conference of youg research and sience workers : Proceedings : Žilina June 25-27, 2007. Žilina : University of Žilina, 2007. s. 171-174. ISBN 978-80-8070-694-4

[8] Lee D. Y., Lee B. K., Hyun D.S.: a novel full-bridge zero voltage transition PWM DC/DC converter with zero-voltage/zero-current switching of auxioliary switches, PESC 98, Fukuoka, Japan, pp.961-968

[9] Dudrik J., Šepeľa J.: Soft-switching current-mode controlled DC-DC converter with secondary switches. In: EDPE 2005 : 13th international conference on Electrical Drives and Power Electronics, September 26-28, 2005, Dubrovnik, Croatia. Zagreb : KoREMA, 2005. 4 p. ISBN 953-6037-43-2

[10] Dudrik J., Dzurko P.: ARC welder with series-parallel resonant DC-DC converter. In: Acta Technica CSAV. vol. 51, no. 4 (2006), p. 415-426. ISSN 0001-7043.

[11] Lacko M., Olejár M., Ruščin V., Dudrik J.: Non-dissipative turn-off snubber for push-pull converter. In: TRANSCOM 2007 : 7-th european conference of youg research and sience workers : Proceedings : Žilina June 25-27, 2007. Žilina : University of Žilina, 2007. s. 139-142. ISBN 978-80-8070-694-4..

[12] Maxim, V., Židek, K., Lupták, M.: Spínač v nule napätia s minimálnym spätným vplyvom na napájaciu sieť. In: AT&P Journal plus. č. 1 (2007), s. 163-165. ISSN 1336-5010.

# Computation of Voltage Sags Propagation Through Two-Winding Three-Phase Transformers

[1]*Martin KANÁLIK*

[1]Dept. of Electric Power Engineering, FEI TU of Košice, Slovak Republic

[1]martin.kanalik@tuke.sk

***Abstract*—this paper deals about application of nodal analysis to computation of voltage sags propagation through two-winding three-phase transformers. There is a nodal analysis briefly described in this paper, on base of which the model of two-winding three-phase transformer is developed. Because there are results of some voltage sags propagation through two-winging three-phase transformers presented in this paper, also seven basic types of voltage sags are introduced in this paper.**

***Keywords*—Power quality, Voltage sags, Transformer modeling, nodal analysis.**

## I. Introduction

Voltage sag is one of the most important power quality indicators. Consequences of voltage sag occurrence for example at manufacturing plant's busbars can be enormous from economical or technical point of view. There are many publications dealing about impacts of voltage sags due to different types of faults in power systems. Because voltage sags description and classification can be found in many publications (for example [1-3]), this paper does not describe basics of voltage sags in detail. The aim of this paper is to apply proposed mathematical method for nodal analysis to evaluation of voltage sags propagation through two-winding three-phase transformers.

## II. Seven types of voltage sags

The voltage sags due to various types of faults are discussed in detail in [1]. For each type of fault, expressions have been derived for the voltages at the pcc. But this voltage is not equal to the voltage level at the equipment terminals. Equipment is normally connected at lower level than the level at which the fault occurs. The voltages at the equipment terminals, therefore, not only depend on the voltages at the pcc but also on the winding connection of transformers between the pcc and the equipment terminals. The voltages at the equipment terminals further depend on the load connection. [1]

Seven different types of voltage sags in equation form in pu are shown in Table I, where $V$ is voltage sag's magnitude and phasor-diagram form of these types of sags is shown in Fig. 1, where dotted lines shows phasors of pre-event voltages and solid lines shows phasors of voltages during the sag.

TABLE I
Seven types of sags in equation form in per unit

| Type A | Type B |
|---|---|
| $V_a = V$ | $V_a = V$ |
| $V_b = -\frac{1}{2}V - \frac{1}{2}jV\sqrt{3}$ | $V_b = -\frac{1}{2} - \frac{1}{2}j\sqrt{3}$ |
| $V_c = -\frac{1}{2}V + \frac{1}{2}jV\sqrt{3}$ | $V_c = -\frac{1}{2} + \frac{1}{2}j\sqrt{3}$ |
| **Type C** | **Type D** |
| $V_a = 1$ | $V_a = V$ |
| $V_b = -\frac{1}{2} - \frac{1}{2}jV\sqrt{3}$ | $V_b = -\frac{1}{2}V - \frac{1}{2}j\sqrt{3}$ |
| $V_c = -\frac{1}{2} + \frac{1}{2}jV\sqrt{3}$ | $V_c = -\frac{1}{2}V + \frac{1}{2}j\sqrt{3}$ |
| **Type E** | **Type F** |
| $V_a = 1$ | $V_a = V$ |
| $V_b = -\frac{1}{2}V - \frac{1}{2}jV\sqrt{3}$ | $V_b = -\frac{1}{3}j\sqrt{3} - \frac{1}{2}V - \frac{1}{6}jV\sqrt{3}$ |
| $V_c = -\frac{1}{2}V + \frac{1}{2}jV\sqrt{3}$ | $V_c = +\frac{1}{3}j\sqrt{3} - \frac{1}{2}V + \frac{1}{6}jV\sqrt{3}$ |
| **Type G** | |
| $V_a = \frac{2}{3} + \frac{1}{3}V$ | |
| $V_b = -\frac{1}{3} - \frac{1}{6}V - \frac{1}{2}jV\sqrt{3}$ | |
| $V_c = +\frac{1}{3} - \frac{1}{6}V + \frac{1}{2}jV\sqrt{3}$ | |

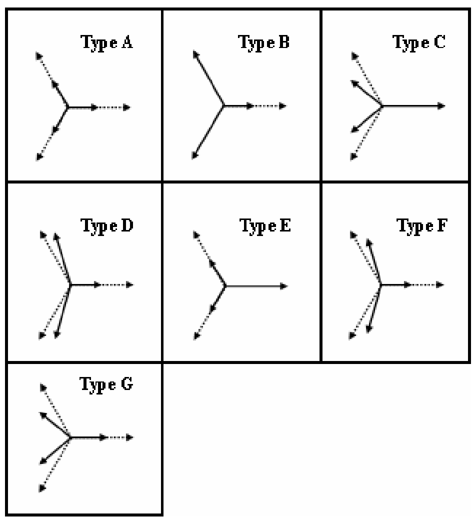



Fig. 1. Seven types of voltage sags in phasor-diagram form.

This classification of voltage sags will be used for evaluation of voltage sags propagation through two-winding three-phase transformers and the results will be compared with results in [1].

## III. Two-winding three-phase transformer modeling for nodal analysis

The model of two-winding three-phase transformer can be developed on the base of network's construction for nodal analysis method [4]. Fig. 2 shows an example of network with two-winding three-phase transformer. Consider that the transformer's connection is *Ygyg*, than the model can be represented on Fig. 3.

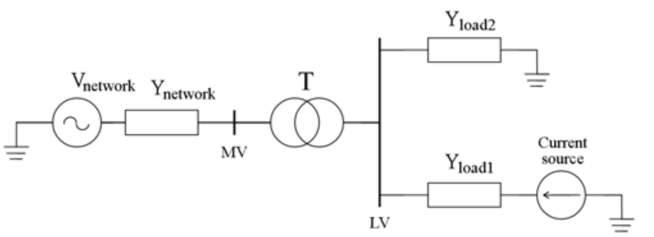



Fig. 2. Example of three-phase network with two-winding transformer and current source – single-phase equivalent

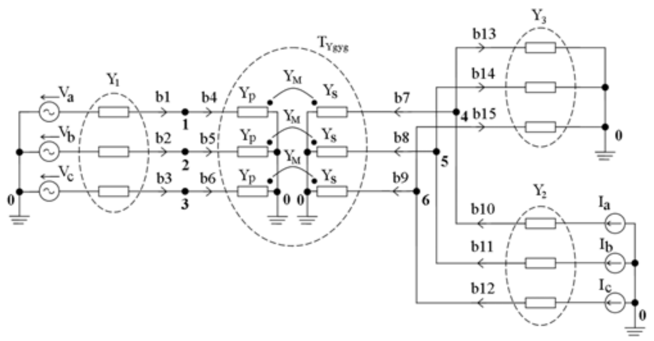



Fig. 3. Three-phase scheme of example on Fig. 2 with transformer connection Ygyg

It is clear from the Fig. 3 that the transformer provides galvanic separation of two circuits, which are connected through mutual couplings of transformer windings. Mutual couplings among circuit's branches are presented in the matrix of branch admittances as off diagonal elements at positions respected relevant branches with the mutual couplings. The three-phase admittance matrix of the circuit on Fig. 3 can be expressed as follows:

$$[Y_d]=\begin{bmatrix} Y_1 & & & & \\ & Y_p & M & & \\ & M & Y_s & & \\ & & & Y_2 & \\ & & & & Y_3 \end{bmatrix} \tag{1}$$

where $Y_i$ represents 3x3 admittances matrices of relevant three-phase elements and $M$ represents 3x3 matrices of mutual couplings among matrices $Y_p$ and $Y_s$.

The branch admittance matrix of given two-winding transformer can be written by equation (2), where $Y_{pw} = Y_{sw}$ are branch admittances of transformer windings in pu and $Y_M$ is admittance of the mutual couplings among transformer's windings. The incidence matrix $A_T$ of this transformer is given by equation (3).

$$[Y_T]=\begin{bmatrix} Y_p & M \\ M & Y_s \end{bmatrix}=\begin{bmatrix} Y_{pw} & & & Y_M & & \\ & Y_{pw} & & & Y_M & \\ & & Y_{pw} & & & Y_M \\ Y_M & & & Y_{sw} & & \\ & Y_M & & & Y_{sw} & \\ & & Y_M & & & Y_{sw} \end{bmatrix} \tag{2}$$

$$[A_T]=\begin{bmatrix} -1 & 0 & 0 & & & \\ 0 & -1 & 0 & & & \\ 0 & 0 & -1 & & & \\ & & & -1 & 0 & 0 \\ & & & 0 & -1 & 0 \\ & & & 0 & 0 & -1 \end{bmatrix} \tag{3}$$

The transformer's connection is represented by the form of $M$ matrices in the equation (2) and by elements in incidence matrix of the equation (3). This approach of two-winding transformer's modeling is described in detail in many publications, for example in [5], [6], [7] or [8], but on the other hand most of these publications provide often only approximations in expression of branch admittances and admittance of the mutual couplings of transformer's windings.

The following part of this paper briefly describes approach of transformer's branch admittances and admittance of the mutual coupling derivation.

### *Derivation of transformer's parameters for computation model*

The derivation of transformer's branch admittances and admittance of the mutual coupling is based on the theory of single-phase linear transformer [9]. The equivalent circuit of the single-phase transformer with parameters expressed in pu is shown in Fig. 4, where $Y_w$ represents winding admittance and $Y_m$ represents magnetization admittance.

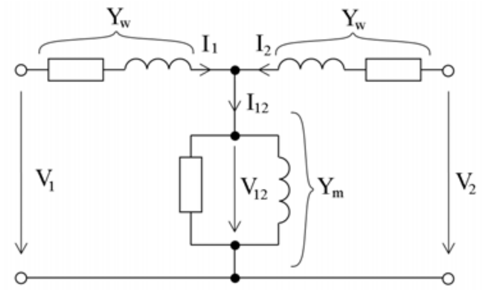



Fig. 4. Equivalent circuit of single-phase transformer with parameters expressed in pu.

The transformer's branch admittances $Y_{pw}$ and $Y_{sw}$ as well as admittance of the mutual coupling $Y_M$ can be expressed through following equations:

$$\Delta V_1=\frac{I_1}{Y_w},\quad \Delta V_2=\frac{I_2}{Y_w},\quad V_{12}=\frac{I_{12}}{Y_m}=\frac{I_1+I_2}{Y_m} \tag{4}$$

$$\begin{aligned} V_1&=V+\Delta V_1=\frac{I_1+I_2}{Y_m}+\frac{I_1}{Y_w} \\ V_2&=V+\Delta V_2=\frac{I_1+I_2}{Y_m}+\frac{I_2}{Y_w} \end{aligned} \tag{5}$$

By the expression of $I_1$ and $I_2$ from equations (5) we obtain:

$$I_1 = \frac{Y_w^2 + Y_w \cdot Y_m}{2 \cdot Y_w + Y_m} \cdot V_1 - \frac{Y_w^2}{2 \cdot Y_w + Y_m} \cdot V_2 = Y_{pw} \cdot V_1 + Y_M \cdot V_2$$
$$I_2 = -\frac{Y_w^2}{2 \cdot Y_w + Y_m} \cdot V_1 + \frac{Y_w^2 + Y_w \cdot Y_m}{2 \cdot Y_w + Y_m} \cdot V_2 = Y_M \cdot V_1 + Y_{sw} \cdot V_2 \qquad (6)$$

From the equation (6) is clear that:

$$Y_{pw} = Y_{sw} = \frac{Y_w^2 + Y_w \cdot Y_m}{2 \cdot Y_w + Y_m}$$
$$Y_M = -\frac{Y_w^2}{2 \cdot Y_w + Y_m} \qquad (7)$$

Table II shows branch admittance matrices $Y_p$ and $Y_s$ as well as admittance matrices of the mutual coupling $Y_M$ ($Y_{ps}$ and $Y_{sp}$)and incidence matrices for basic types of two-winding three-phase transformer's connection, whereby:

$$\left[Y_T\right] = \begin{bmatrix} Y_p & Y_{ps} \\ Y_{sp} & Y_s \end{bmatrix}, \qquad \left[A_T\right] = \begin{bmatrix} A_p & \\ & A_s \end{bmatrix} \qquad (8)$$

and

$$Y_u = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; Y_v = \frac{1}{3} \cdot \begin{bmatrix} 1 & 1 & -2 \\ -2 & 1 & 1 \\ 1 & -2 & 1 \end{bmatrix}, \; Y_w = \frac{1}{\sqrt{3}} \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix},$$

$$Y_k = \frac{1}{3} \cdot \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}, \; Y_l = \sqrt{3} \cdot Y_w = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

The superscript (*) behind the connection type in Table II indicates that the admittance of delta-connected winding is necessary to transpose to star-connected, because in these connection types the delta connection was considered as a star connection.

TABLE II
ADMITTANCE AND INCIDENCE MATRICES FOR TWO-WINDING THREE-PHASE TRANSFORMERS

| **T** | $Y_p$ | $Y_{ps}$ | $Y_{sp}$ | $Y_s$ | $A_p$ | $A_s$ |
|---|---|---|---|---|---|---|
| Ygyg | $Y_{pw} \cdot Y_u$ | $Y_M \cdot Y_u$ | $Y_M \cdot Y_u$ | $Y_{sw} \cdot Y_u$ | $-Y_u$ | $-Y_u$ |
| Ygy | $Y_{pw} \cdot Y_u$ | $Y_M \cdot Y_k$ | $Y_M \cdot Y_k^T$ | $Y_{sw} \cdot Y_u$ | $-Y_u$ | $-Y_k$ |
| Ygd1* | $Y_{pw} \cdot Y_u$ | $-Y_M \cdot Y_w$ | $-Y_M \cdot Y_w^T$ | $Y_{sw} \cdot Y_u$ | $-Y_u$ | $-Y_k$ |
| Ygd11* | $Y_{pw} \cdot Y_u$ | $-Y_M \cdot Y_w^T$ | $-Y_M \cdot Y_w$ | $Y_{sw} \cdot Y_u$ | $-Y_u$ | $-Y_k$ |
| Yyg | $Y_{pw} \cdot Y_u$ | $-Y_M \cdot Y_k$ | $-Y_M \cdot Y_k^T$ | $Y_{sw} \cdot Y_u$ | $-Y_k$ | $-Y_u$ |
| Yy | $Y_{pw} \cdot Y_u$ | $-Y_M \cdot Y_k$ | $-Y_M \cdot Y_k^T$ | $Y_{sw} \cdot Y_u$ | $-Y_k$ | $-Y_k$ |
| Yd1* | $Y_{pw} \cdot Y_u$ | $Y_M \cdot Y_w$ | $Y_M \cdot Y_w^T$ | $Y_{sw} \cdot Y_u$ | $-Y_k$ | $-Y_k$ |
| Yd11* | $Y_{pw} \cdot Y_u$ | $Y_M \cdot Y_w^T$ | $Y_M \cdot Y_w$ | $Y_{sw} \cdot Y_u$ | $-Y_k$ | $-Y_k$ |
| D1yg* | $Y_{pw} \cdot Y_u$ | $Y_M \cdot Y_w^T$ | $Y_M \cdot Y_w$ | $Y_{sw} \cdot Y_u$ | $-Y_k$ | $-Y_u$ |
| D1y* | $Y_{pw} \cdot Y_u$ | $-Y_M \cdot Y_w^T$ | $-Y_M \cdot Y_w$ | $Y_{sw} \cdot Y_u$ | $-Y_k$ | $-Y_k$ |
| D1d1 | $Y_{pw} \cdot Y_u$ | $Y_M \cdot Y_v^T$ | $Y_M \cdot Y_v$ | $Y_{sw} \cdot Y_u$ | $Y_l$ | $Y_l^T$ |
| D1d11 | $Y_{pw} \cdot Y_u$ | $Y_M \cdot Y_v$ | $Y_M \cdot Y_v^T$ | $Y_{sw} \cdot Y_u$ | $Y_l$ | $Y_l$ |
| D11yg* | $Y_{pw} \cdot Y_u$ | $-Y_M \cdot Y_w$ | $-Y_M \cdot Y_w^T$ | $Y_{sw} \cdot Y_u$ | $-Y_k$ | $-Y_u$ |
| D11y* | $Y_{pw} \cdot Y_u$ | $Y_M \cdot Y_w$ | $Y_M \cdot Y_w^T$ | $Y_{sw} \cdot Y_u$ | $-Y_k$ | $-Y_k$ |
| D11d1 | $Y_{pw} \cdot Y_u$ | $Y_M \cdot Y_v^T$ | $Y_M \cdot Y_v$ | $Y_{sw} \cdot Y_u$ | $Y_l$ | $Y_l$ |
| D11d11 | $Y_{pw} \cdot Y_u$ | $Y_M \cdot Y_v^T$ | $Y_M \cdot Y_v$ | $Y_{sw} \cdot Y_u$ | $Y_l$ | $Y_l^T$ |

## IV. COMPUTATION OF VOLTAGE SAG'S PROPAGATION

This chapter deals about the computation method of nodal analysis of three-phase power systems [10], for voltage sags propagation evaluation. The aim of this method is to obtain nodal voltages of the evaluated network for the pre-event case and for the case in voltage sag occurrence. The method is based the equation (9), which determines network's nodal voltages:

$$\left[V_{node}\right] = \left[Y_{node}\right]^{-1} \cdot \left[I_{node}\right] \qquad (9)$$

where $[Y_{node}]$ is network's nodal admittance matrix and $[I_{node}]$ is the vector of nodal currents. Elements of $[I_{node}]$ are determined as a sum of currents from all sources (internal branch currents and voltages) which incident with the relevant node, i.e. if the branch contains the current source, the branch contributes to the summation with the current of that source, and if the branch contains the voltage source, the branch contributes to the summation with the current given as a product of source's voltage and branch admittance. For example, in the case of network shown on Fig. 3, the $[I_{node}]$ has following form:

$$\left[I_{node}\right] = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{bmatrix} V_a \cdot Y_{1a} \\ V_b \cdot Y_{1b} \\ V_c \cdot Y_{1c} \\ I_a \\ I_b \\ I_c \end{bmatrix} \qquad (10)$$

Fig. 5. shows an example of network containing two two-winding three-phase transformers both connected to D1yg. Similar network was used as an example for different types of sags propagation through transformers in [1].

Because there are no parameters of network's elements mentioned in [1], parameters of transformers and loads were assimilated to voltage levels in Fig. 5, whereby $T_1$ is 110/22 kV, 40 MVA transformer; $T_2$ is 22/6 kV, 1.6 MVA transformer; $Load_1$ is 34 MVA with $\cos\varphi = 0.96$ and $Load_2$ is 1 MVA with $\cos\varphi = 0.94$.

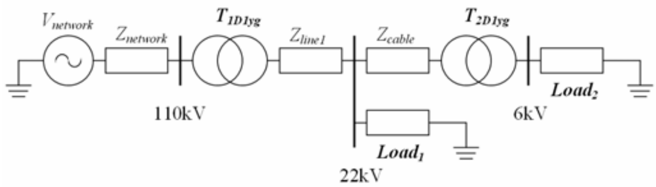



Fig. 5. Example of network containing two two-winding three-phase transformers, for voltage sags propagation evaluation

Propagation of voltage sags was evaluated based on two types of characteristic computations. The first computation type was executed for normal operating conditions and was common for all evaluations. The second computation type was executed with characteristic type of voltage sag, which was modeled by network voltage setup.

Table III sows results of computations in normal operating conditions and computations with characteristic voltage sag occurrence in the network voltage, where sag type in network voltage is characterized through its type and magnitude in percents.

TABLE III
COMPUTATION RESULTS FOR DIFFERENT SAG OCCURENCE IN NETWORK VOLTAGE

| Network voltage sag type | | Busbars | | | | | |
|---|---|---|---|---|---|---|---|
| | | 110 kV | | 22 kV | | 6 kV | |
| | | rms [kV] | angle [deg] | rms [kV] | angle [deg] | rms [kV] | angle [deg] |
| normal | a | 63.466 | 0.0 | 12.062 | -156.8 | 3.212 | 51.2 |
| | b | 63.466 | -120.0 | 12.062 | 83.2 | 3.212 | -68.8 |
| | c | 63.466 | 120.0 | 12.062 | -36.8 | 3.212 | 171.2 |
| A 50% | a | 31.734 | 0.0 | 6.031 | -156.8 | 1.606 | 51.2 |
| | b | 31.734 | -120.0 | 6.031 | 83.2 | 1.606 | -68.8 |
| | c | 31.734 | 120.0 | 6.031 | -36.8 | 1.606 | 171.2 |
| B 50% | a | 31.726 | 0.0 | 9.213 | -145.9 | 2.980 | 60.1 |
| | b | 63.477 | -120.0 | 12.062 | 83.2 | 2.980 | -77.8 |
| | c | 63.462 | 120.0 | 9.213 | -47.7 | 2.141 | 171.2 |
| C 50% | a | 63.466 | 0.0 | 10.873 | -170.7 | 2.124 | 32.1 |
| | b | 41.979 | -139.1 | 6.031 | 83.2 | 2.124 | -49.7 |
| | c | 41.979 | 139.1 | 10.873 | -22.9 | 3.212 | 171.2 |
| D 50% | a | 31.734 | 0.0 | 7.979 | -137.7 | 2.895 | 65.1 |
| | b | 57.197 | -106.1 | 12.062 | 83.2 | 2.895 | -82.7 |
| | c | 57.214 | 106.1 | 7.979 | -55.9 | 1.606 | 171.2 |
| E 50% | a | 63.473 | 0.0 | 9.213 | -167.7 | 1.930 | 37.2 |
| | b | 31.723 | -120.0 | 6.031 | 83.2 | 1.930 | -54.9 |
| | c | 31.737 | 120.0 | 9.213 | -25.9 | 2.676 | 171.2 |
| F 50% | a | 31.734 | 0.0 | 7.249 | -142.9 | 2.453 | 62.1 |
| | b | 48.473 | -109.1 | 10.052 | 83.2 | 2.453 | -79.7 |
| | c | 48.473 | 109.1 | 7.249 | -50.7 | 1.606 | 171.2 |
| G 50% | a | 52.888 | 0.0 | 9.213 | -167.7 | 1.930 | 37.3 |
| | b | 38.138 | -133.9 | 6.031 | 83.2 | 1.930 | -54.9 |
| | c | 38.138 | 133.9 | 9.213 | -25.9 | 2.676 | 171.2 |

Based on results in Table III and equations in Table I, classification of voltage sags through transformers connected in D1yg was made. This classification is shown in Table IV. The index a, b or c behind the sag's type indicates to which phase the relevant phasor-diagram from Fig. 1 is rotated.

TABLE IV
CLASSIFICATION OF VOLTAGE SAGS PROPAGATED THROUGH TRANSFORMERS D1YG FROM RESULTS IN TABLE III

| Network voltage sag type | Busbars | | |
|---|---|---|---|
| | 110 kV | 22 kV | 6 kV |
| A, 50% | A, 50% | A, 50% | A, 50% |
| $B_a$, 50% | $B_a$, 50% | $C_b$, 67% | $D_c$, 67% |
| $C_a$, 50% | $C_a$, 50% | $D_b$, 50% | $C_c$, 50% |
| $D_a$, 50% | $D_a$, 50% | $C_b$, 50% | $D_c$, 50% |
| $E_a$, 50% | $E_a$, 50% | $F_b$, 50% | $G_c$, 50% |
| $F_a$, 50% | $F_a$, 50% | $G_b$, 50% | $F_c$, 50% |
| $G_a$, 50% | $G_a$, 50% | $F_b$, 50% | $G_c$, 50% |

Similar classification was made in [1] but there were no relation ships between transformer's connection and voltage sag's rotation made in [1]. If we consider only type and magnitude of voltage sags, results obtained by proposed method are the same as in [1], what confirms validity of presented computation method.

Table V shows voltage sag's propagation through two-winding three-phase transformers given by [1], which does not include the phase rotation of voltage sags due to transformer's connection.

TABLE V
TRANSFORMATION OF SAG TYPE TO LOWER VOLTAGE LEVEL

| Transformer Connection | Sag on Primary Side | | | | | | |
|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G |
| Ygyg | A | B | C | D | E | F | G |
| Yy, Dd,Ygy,Yyg | A | D | C | D | G | F | G |
| Yd, Ygd,Dy, Dyg | A | C | D | C | F | G | F |

The sag rotation due to transformer connection is shown in Table VI, where "0" indicates no rotation, "+" indicates clockwise rotation and "-" indicates anticlockwise rotation.

TABLE VI
SAG ROTATION DUE TO TRANSFORMER'S CONNECTION

| Rotation | Transformer Connection |
|---|---|
| 0 | Ygyg, Ygy, Yyg, Yy, D1d1, D11d11 |
| + | Ygd11, Yd11, D1yg, D1y, D11d1 |
| - | Ygd1,Yd1, D11yg, D11y, D1d11 |

## V. CONCLUSION

The type of voltage sag occurred at equipment terminals depends on transformer connections between the fault and equipment terminals. Results of this paper were obtained by computation method for steady state computations of three-phase systems introduced in [4], whereby model of transformer described in chapter III of this paper was used. Results published in [1] were taken for verification of this method. Because results given by presented approach correspond to the results in [1] we can consider this method as suitable for evaluation of voltage sags propagation through two-winding three-phase transformers. Through the execution of computations with different voltage sags and different transformer's connections, voltage sags classifications in Table V and Table VI were obtained. These results can be helpful in fault type classification based on its manifestation at equipment terminals and known location.

## ACKNOWLEDGMENT

This work was supported by Scientific Grant Agency of the Ministry of Education of Slovak Republic and the Slovak Academy of Sciences under the contract No. 1/4075/07.

## REFERENCES

[1] Bollen, M. H. J: UNDERSTANDING POWER QUALITY PROBLEMS – Voltage Sags and Interruptions, John Wiley & Sons, Inc., Hoboken, New Jersey, 2000, ISBN 0-7803-4713-7
[2] Dugan, R. C. – McGranaghan, M. F. – Santoso, S. – Beaty, H. W: Electrical Power Systems Quality. McGraw-Hill, 2003, ISBN 0-07-138622-X
[3] STN EN 50160 (33 0121): 2002 Charakteristiky napätia elektrickej energie dodávanej z verejnej distribučnej siete.
[4] Kanálik, M – Kolcun, M: Computation of harmonic flows in three-phase systems. In: Acta Electrotechnica et Informatica. roč. 7, no. 3 (2007), s. 41-45. ISSN 1335-8243.
[5] Arrilaga, J. – Watson, N. R.: Power System harmonics, Second edition, John Wiley & Sons, Ltd, England, 2003, ISBN 0-470-85129-5
[6] Wakileh, G. J.: Power Systems Harmonics – Fundamentals, Analysis and Filter Design, Springer, 2001, ISBN 3-540-42238-2
[7] Arrilaga, J., Watson, N. R.– Computer Modelling of Electrical Power Systems, Second edition, 2001, John Wiley & Sons, Ltd, England, ISBN 0-471-87249-0
[8] Das J.C.: Power System Analysis – Short-Circuit Load Flow and Harmonics, 2002, Marcel Dekker, Inc, New York, ISBN 0-8247-0737-0
[9] Vladář J. – Zelenka J.: Elektrotechnika a silnoproudá elektronika, 1986, STNL Prague, Czech Republic, L25-C3-IV-51f/58663
[10] Schmidtmayer J.: Maticový počet a jeho použití v technice, 1967, STNL Prague, Czech Republic

# Effect of disturbances in Power Systems simulated in program EUROSTAG

[1]Michal VARGONČÍK, [2]Michal KOLCUN

[1]Dept. of Electronics and Electric Power Engineering, FEI TU of Košice, Slovak Republic
[2]Head of dept. of Electronics and Electric Power Engineering, FEI TU of Košice, Slovak Republic

[1]michal.vargoncik@tuke.sk, [2]michal.kolcun@tuke.sk

***Abstract*—This paper deals with clarifying the proper use of the terms power swing and out-of-step. The paper provides a brief discussion of these phenomena, how these phenomena affect the protective relaying on transmission lines, and explains many of the methods available that protective relays use to detect power swings and out-of-step conditions. This paper presents a simulation analyses of transient stability on simple power system in Eurostag.**

***Keywords*—power system stability, power swing, out-of-step, relay.**

## I. Introduction

Power system under steady-state conditions operate typically close to their nominal frequency. A balance between generated and consumed active and reactive powers exists during steady-state operating conditions and the sending and receiving and voltages are typically within 5%. The system frequency on a large power system will typically vary +/- 0,02 Hz on 50 Hz power system. Power system faults, line switching, generator disconnection, and the loss or application of large blocks of load results in sudden changes to electrical power, whereas the mechanical power input to generator remains relatively constant. These system disturbances cause oscillations in machine rotor angles and can result in severe power flow swings. Depending on the severity of the disturbance and the actions of power system controls, the system may remain stable and return to a new equilibrium state experiencing what is referred to as a stable power swing. Severe system disturbances, on the other hand, could cause large separation of generator rotor angles, large swings of power flows, large fluctuations of voltages and currents, and eventual loss of synchronism between groups of generators or between neighboring utility systems. Large power swings, stable or unstable, can cause unwanted relay operations at different network locations, which can aggravate further the power system disturbance and possibly lead to cascading outages an power blackouts.

A transmission line trip during a power swing may cause instability of a power system. It is necessary to recognize the power swing from fault.

## II. Definitions

Power swing: a variation in three phase power flow which occurs when the generator rotor angles are advancing or retarding relative to each other in response to changes in load magnitude and direction, line switching, loss of generation, faults and other disturbances.

Pole Slip: a condition whereby a generator, or group of generators, terminal voltage angles (or phases) go past 180 degrees with respect to the rest of the connected power system.

Stable power swing: a power swing is considered stable if the generators do not slip poles and the system reaches a new state of equilibrium, i.e. an acceptable operating condition.

Unstable power swing: a power swing that will result in a generator or group of generators experiencing pole slipping for which some corrective action must be taken.

Out-of-Step Condition (OOS): same as an unstable power swing

Electrical System Center or Voltage Zero: it is the point or points in the system where the voltage becomes zero during an unstable power swing.

## III. Effect of Power swings

A balance between generated and consumed active and reactive power exists during steady-state operating conditions. Any change in the power generated, load demand or in the transmission line network causes the power flow to change across the system until a new equilibrium is established between generation and load. These changes in power flow occur continuously, are automatically compensated for via control systems and normally have no detrimental effect on the power grid or its protective systems.

Power system faults, line switching, generator disconnection and the loss or application of large blocks of load result in sudden changes to electrical power, whereas the mechanical power input to generators remains relatively constant. These system disturbances cause oscillations in machine rotor angles and can result in severe power flow swings. Power swings are variations in power flow that occur when the internal voltages of generators at different locations of the power system slip relative to each other. Large power swings, stable or unstable, can cause unwanted relay operations at different network locations, which can aggravate the power-system disturbance and cause major power outages or power blackouts.

The loss of synchronism between power systems or a generator and the power system affects transmission line relays and systems in various ways. The required settings for the power swing blocking (PSB) and out-of-step tripping (OST) elements could be difficult to calculate in many applications. For these applications, extensive stability studies with different operating conditions must be performed to determine the fastest rate of possible power swings. In fact, some transmission line relays may operate for stable power swings for which the system should recover and remain stable.

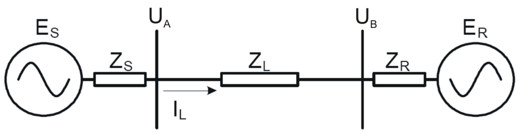



Fig. 1. Simple two machine system.

## IV. Impedance Measured by Distance Relays During Power swings

During a system OOS event, a distance relay may detect the OOS as a phase fault if the OOS trajectory enters the operating characteristic of the relay. To demonstrate this, look at the impedance that a distance relay measures during OOS condition for the simple two-source system.

Considering Figure 1, the current $I_L$ at bus A is computed as

$$I_L = \frac{E_S - E_R}{Z_S + Z_L + Z_R}. \tag{1}$$

The direction of current flow will remain the same during the power swing event. Only the voltages change with respect to one another. The impedance measured at a relay at bus A would then be:

$$I_L = \frac{U_A}{I_L} = \frac{E_S - I_L.Z_S}{I_L} = \frac{E_S}{I_L} - Z_S = \\ = \frac{E_S.(Z_S + Z_L + Z_R)}{E_S - E_R} - Z_S . \tag{2}$$

Assume that $E_S$ has a phase advance of $\delta$ over $E_R$ and that the ratio of the two source voltage magnitudes $|E_S|/|E_R|$ is *k*. Then

$$\frac{E_S}{E_S - E_R} = \frac{k(\cos\delta + j\sin\delta)}{k(\cos\delta + j\sin\delta) - 1} = \\ = \frac{k[(k - \cos\delta) - j\sin\delta]}{(k - \cos\delta)^2 + \sin^2\delta}. \tag{3}$$

For the particular case where the two sources magnitudes are equal or *k* is one, equation (3) can be expressed:

$$\frac{E_S}{E_S - E_R} = \frac{1}{2}(1 - j\cot\frac{\delta}{2}). \tag{4}$$

And finally the impedance measured at the relay will be:

$$Z = \frac{U_A}{I_L} = \frac{Z_S + Z_L + Z_R}{2}(1 - j\cot\frac{\delta}{2}) - Z_S . \tag{5}$$

$\delta$ is the phase angle between the sources, there is a geometrical interpretation to equation (5) that is represented in Figure 2a. The trajectory of the measured impedance at the relay during a power swing when the angle between the two source voltages varies, corresponds to the straight line that intersects the segment A to B at its middle point. This point is called the electrical center of the swing. The angle between the two segments that connect P to points A and B is equal to the

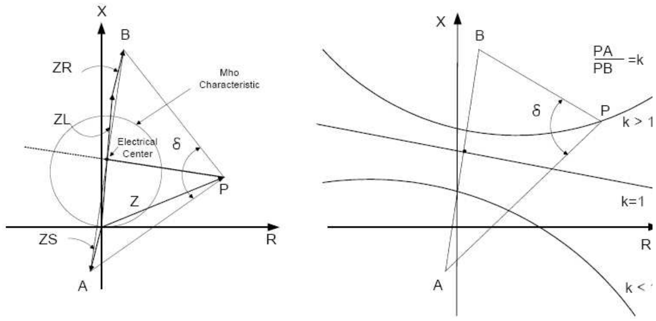



Fig. 2. Impedance trajectories at the relay during swing for different *k* values.

angle $\delta$. When the angle $\delta$ reaches the value of 180 degrees, the impedance is precisely at the location of the electrical center. It can be seen that the impedance trajectory during a power swing will cross any relay characteristic that covers the line, provided the electrical center falls inside the line.

## V. Simulation model

### A. Model system and initial condition

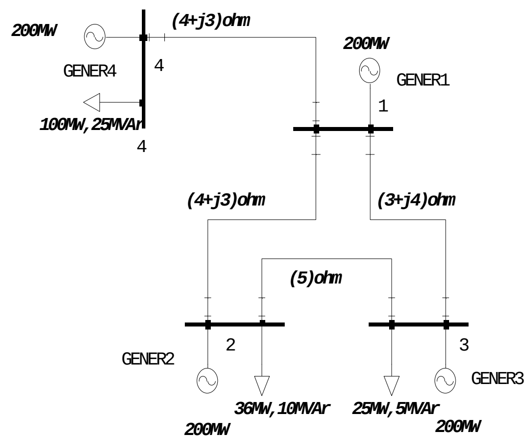
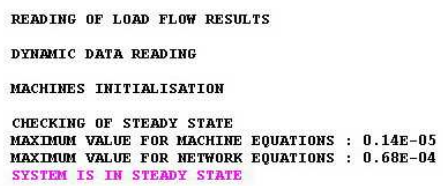



Fig. 3. Tested power system.

Figure 3 shows a four-machine model system which is used for the simulation analyses in program Eurostag, in which are four generators, four lines and four nodes where are connected three types of load. Figure 4 shows computation module in program Eurostag for steady-state operation of tested power system. This is the phase where program compute only steady-


READING OF LOAD FLOW RESULTS

DYNAMIC DATA READING

MACHINES INITIALISATION

CHECKING OF STEADY STATE
MAXIMUM VALUE FOR MACHINE EQUATIONS : 0.14E-05
MAXIMUM VALUE FOR NETWORK EQUATIONS : 0.68E-04
SYSTEM IS IN STEADY STATE

Fig. 4. Computation module process.

state without dynamics and transition phenomenas.

## B. *Transition phenomena*

In this phase of computation, program assuming the generators and theirs dynamics. For using this part of computation user must set automatic voltage regulator (AVR) and governors of all generators which are connected to the system. Because the most important part of a power system network for power swing are generators. When a disturbance occurs on a part of a power system, the generators being close to the disturbance or fault start swings. In the tested power

```
READING OF LOAD FLOW RESULTS

DYNAMIC DATA READING

MACHINES INITIALISATION

CHECKING OF STEADY STATE
MAXIMUM VALUE FOR MACHINE EQUATIONS : 0.18E-05
MAXIMUM VALUE FOR NETWORK EQUATIONS : 0.36E-14
SYSTEM IS IN STEADY STATE

10.0000 s - FAULT ON LINE 4-1-4 AT 50.00000 % OF NODE 4
            PHASE 1 : GROUNDED - IMPEDANCE :   0.00000  0.00000
            PHASE 2 : GROUNDED - IMPEDANCE :   0.00000  0.00000
            PHASE 3 : GROUNDED - IMPEDANCE :   0.00000  0.00000
            CONNECTION NODE : UNGROUNDED - IMPEDANCE :   0.00000  0.00000
10.5000 s - FAULT ELIMINATED ON LINE 4-1-4

TIME FOR END OF SIMULATION IS REACHED
```

Fig. 5. Next computation module process.

system was tried many types of disturbance for testing theirs influence on generators. Figure 5 shows computation module for the example of disturbance in Eurostag which assuming of dynamics of power system.

## C. *Influence of disturbances*

In this subsection is demonstration of disturbances which are influencing on power angle of generators and shows a changes in generators connected to tested power system.

Figure 6 shows a type of short circuit without transition impedance, which occurs on line 1 – 4 in power system with time duration 0,01s. Fault occurs in 50% of this line. This results in small changes in power angles of generators and all generators are in synchronism.

Figure 7 shows a type of short circuit without transition

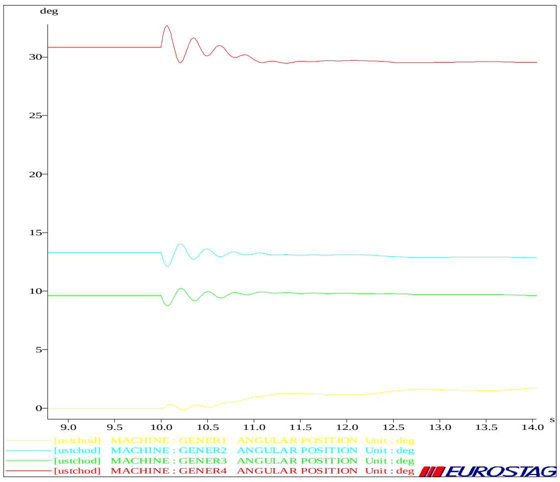



Fig. 6. Short circuit on line 1 – 4 with time duration 0,01s and distance 50% of line.

impedance, which occurs on line 1 – 4 in power system with time duration 0,1s. Fault occurs in 50% of this line. This results in big changes in power angles of generators but all generators still are in synchronism.

Figure 8 shows a type of short circuit without transition impedance, which occurs on line 1 – 4 in power system with time duration 0,5s. Fault occurs in 50% of this line. This results in huge changes in power angles of generators and all generators losing synchronism.

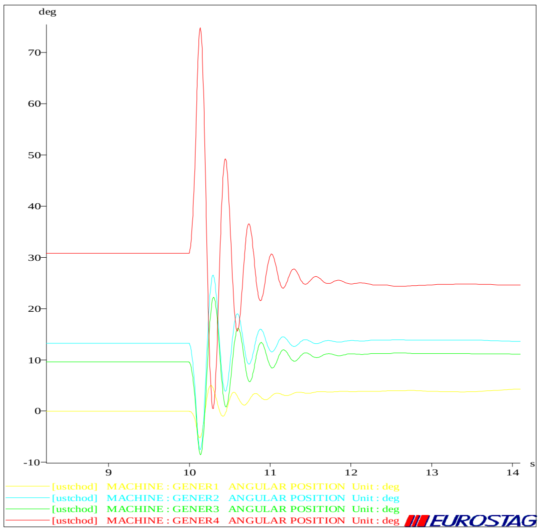



Fig. 7. Short circuit on line 1 – 4 with time duration 0,1s and distance 50%

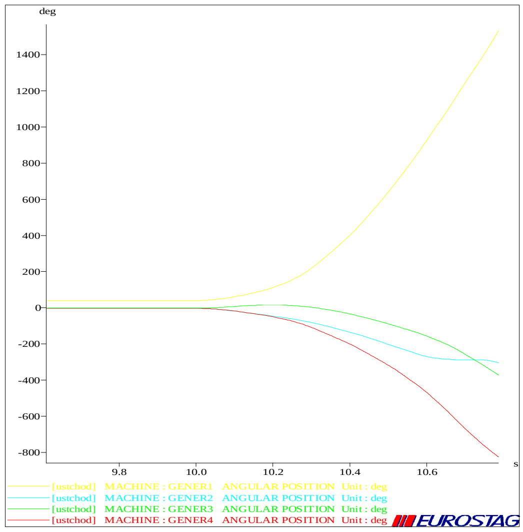



Fig. 8. Short circuit on line 1 – 4 with time duration 0,5s and distance 50% of line.

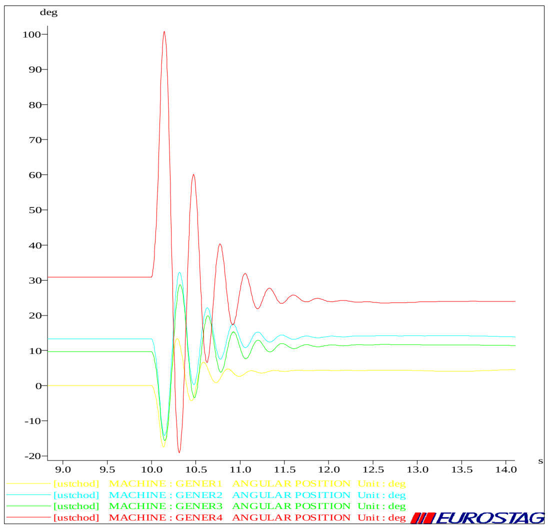



Fig. 9. Short circuit on line 1 – 4 with time duration 0,1s and distance 90% of line.

Figure 9 shows a type of short circuit without transition impedance, which occurs on line 1 – 4 in power system with time duration 0,1s. Fault occurs in 90% of this line near the node 1. This results in big changes in power angle of first generator and remaining generators trying to synchronize with the first generator and effect with opposing force. This example is comparison with example on Figure 7. The main change is in distance of disturbance.

Figure 10 shows a type of load change on node 2 in power system in time 10s. Increase in load is 200% of initial load. This results in big changes in power angle of second generator

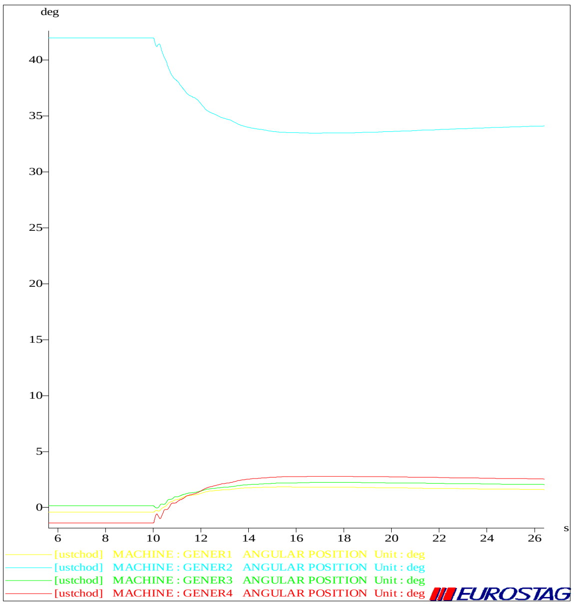



Fig. 10. Load change on node 2 in time 10s with increase 200% of initial values.

and remaining generators trying to synchronize with the second generator and effect with opposing force.

## VI. Conclusion

On the example of power system, which is described in this paper we can recognize how disturbances in power system influencing on generators and essentially on stability of power system. This types of faults was simulated in program Eurostag for a good visualization of theirs effect. This program uses Fortescue and balanced modeling. In section V subsection C are described individual faults which may occurs in power system. From results of analysis we can deduce that small disturbances causing small instability and swinging in rotors of generators however these generators stay in synchronism. Big disturbances causes big changes in power angle of generator rotors and causes big instability and loss of synchronism of all generators connected in power system.

## References

[1] KOLCUN, M. – CHLADNÝ, V. – VARGA, L. – BEŇA, Ľ. – ILENIN, S. – LEŠČINSKÝ, P. – MEŠTER, M. “*Anylýza elektrizačnej sústavy, časť: Ustálené chody v ES*“, Košice 2005, pp. 52–59.

[2] KUNDUR, P. “*Power System Stability and Control*“, The EPRI Power System Engineering Series, New York, 1994, ch.2.

[3] PAVELLA, M. – MURTHY, P. G. “*Transient Stability of Power Systems*“, Great Britain, 1994, ch. 6.

[4] TZIOUVARAS, D. – MCDONALD, M. “Power swing and Out-of-step considerations on transmission lines“, July 2005.

[5] TROJÁNEK, Z. – HÁJEK, J. – KVASNICA, P. “*Přechodné jevy v elektrizačních soustavách*“, Praha: 1.vydanie, vydavateľstvo SNTL/ALFA 1987.

[6] HELLER, B. “*Stabilita elektromechanických vyrovnávacích zjevů paralelně pracujících synchrónnich strojů”*, Elektrotechnický obzor, svazek 27, č.6, Praha 1938, pp. 86 – 90.

[7] Horowitz, S. H. – Phadke, A. G. *“Power System Relaying”*. 2nd Edition, John Wiley & Sons, 1996.

***This work was supported by Scientific Grant Agency of the Ministry of Education of Slovak Republic and the Slovak Academy of Sciences under the contracts No. 1/4072/07 and 1/4075/07.***

# Direct torque control of asynchronous motor with the help of fuzzy logic

*Ľubomír CIBUĽA*

Dept. of Electrical Engineering, Informatics and Industrial Engineering, FEI TU of Košice, Slovak Republic

lubomir.cibula@tuke.sk,

**Abstract— This contribution deals with the proposal of direct torque control (DTC) of asynchronous motor (AM) with the help of fuzzy logic. The whole structure of DTC is designed in software Matlab – Simulink, the fuzzy regulator is designed with the help of Fuzzy Toolbox. The results of DTC with fuzzy regulator are compared with DTC with the help of Depenbrock method and DTC with the help of Takahashi method.**

***Keywords*— direct torque control, asynchronous motor, fuzzy regulator**

## I. Introduction

Direct torque control is one of the most up to date ways of the control of asynchronous motor.

In DTC the link of converter's control with motor is very close. All changes of the status of switching elements of the inverter are derived from electro magnetic status of motor. DTC enables very fast and flexible control of induction motor.

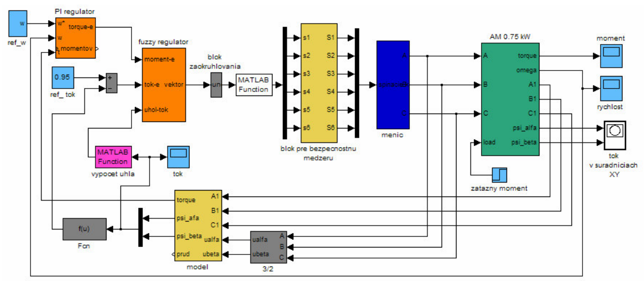
The calculation of necessary stator flux and moment is made based on the measurement of stator voltage and current and with the help of a block of the calculation of magnetic flux and moment. The aim of this work is to achieve better results as in case of DTC that was realized according to method of Depenbrock and Takahashi. The improvement of the results relates mainly to the increase of regulation's precision and decrease of undesired great vibration of moment.

## II. Structure of Direct Moment Control

The whole structure of DTC is comprised of several basic blocks:

- Block of the model of asynchronous motor (marked as AM 0.75 kW)
- Block of indirect frequency inverter with voltage inter-circuit, from which it is fed AM (inverter)
- Block of the calculation of magnetic flux and moment (model)
- Block of angle calculation (position) of the stator of magnetic flux (calculation of angle)
- Block PI of speed regulator (PI regulator) and fuzzy regulator of the moment and flux (fuzzy regulator)

Fig. 1. Structure DTC



## III. Fuzzy Regulator

Fuzzy regulator (FR) was designed and realized in software program Fuzzy Logic Toolbox (Matlab). As input value for FR are used: moment deviation (Fig. 2.), deviation of stator flux (Fig. 3.) and position of stator flux (Fig. 4.).

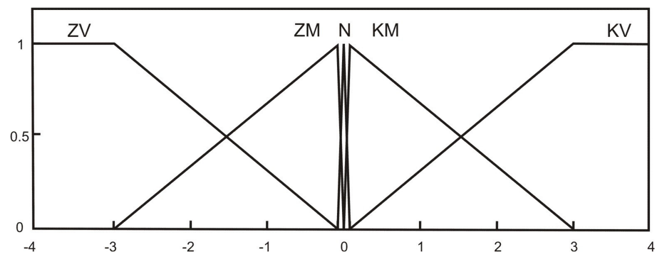



Fig. 2. The distribution of the functions of competences for deviation of the moment

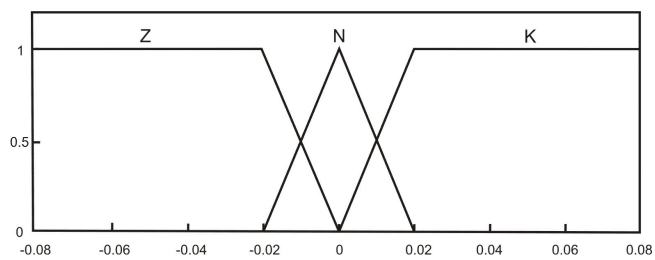



Fig. 3. Distribution of the functions of competences for deviation of stator flux

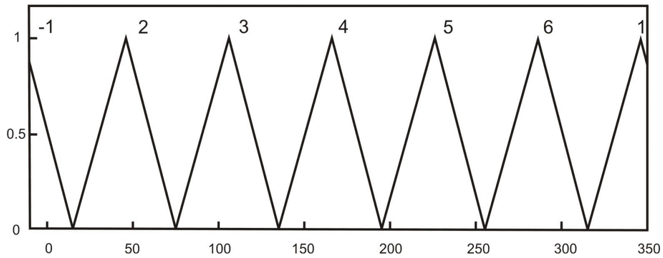



Fig. 4. Distribution of functions of competences for position of stator flux

The output from fuzzy regulator is the required vector of stator voltage (Fig. 5.). The resulting vector (we gain it from fuzzy regulator) does not represent directly the required vector (it is not a whole number) therefore this output has to be further modified (to round up) in block of Rounding function. This resulting value is then representing the vector of stator voltage, needed as an input into indirect frequency inverter with voltage of intermediate circuit.

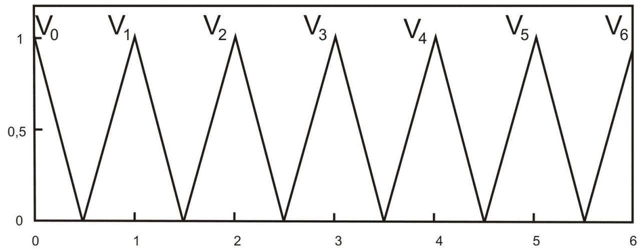



Fig. 5. Distribution of the function of competences for the vector of stator voltage

*Inferential rules*

The proper activities of fuzzy system - regulator is based on derivational (inferential) rules, similarly to expert systems. The benefit of such representation of knowledge is the transparency for users. These rules are of type IF-THEN and are illustrated with the help of Look-up table (Fig. 6.). The total number of rules after their reduction is 105.

-1(1)

| Mo \ To | K | N | Z |
|---|---|---|---|
| KV | 1 | 2 | 2 |
| KM | 2 | 2 | 3 |
| N | 1 | 0 | 4 |
| ZM | 6 | 5 | 4 |
| ZV | 6 | 5 | 5 |

2

| Mo \ To | K | N | Z |
|---|---|---|---|
| KV | 2 | 3 | 3 |
| KM | 3 | 3 | 4 |
| N | 2 | 0 | 5 |
| ZM | 1 | 6 | 5 |
| ZV | 1 | 6 | 6 |

3

| Mo \ To | K | N | Z |
|---|---|---|---|
| KV | 3 | 4 | 4 |
| KM | 4 | 4 | 5 |
| N | 3 | 0 | 6 |
| ZM | 2 | 1 | 6 |
| ZV | 2 | 1 | 1 |

4

| Mo \ To | K | N | Z |
|---|---|---|---|
| KV | 4 | 5 | 5 |
| KM | 5 | 5 | 6 |
| N | 4 | 0 | 1 |
| ZM | 3 | 2 | 1 |
| ZV | 3 | 2 | 2 |

5

| Mo \ To | K | N | Z |
|---|---|---|---|
| KV | 5 | 6 | 6 |
| KM | 6 | 6 | 1 |
| N | 5 | 0 | 2 |
| ZM | 4 | 3 | 2 |
| ZV | 4 | 3 | 3 |

6

| Mo \ To | K | N | Z |
|---|---|---|---|
| KV | 6 | 1 | 1 |
| KM | 1 | 1 | 2 |
| N | 6 | 0 | 3 |
| ZM | 5 | 4 | 3 |
| ZV | 5 | 4 | 4 |

Fig. 6. Look-up table with inferential rules

## IV. COMPARISON OF THE RESULTS OF SIMULATIONS OF DTC METHODS (DEPENBROCK AND TAKAHASHI) WITH DTC REALIZED WITH THE HELP OF FUZZY LOGIC

The courses were simulated for the start up of AM and reversal of angular velocity.

### A. *Courses of the start up of asynchronous motor*

In case of this simulation we start up AM to nominal speed 145 rad/s. We start up the motor from zero speed. The courses of angular velocity are on Fig. 7. Due to the fact that during the courses of angular velocities for individual methods there are no significant differences among the methods there is illustrated only the course of angular velocity in DTC with the help of fuzzy.

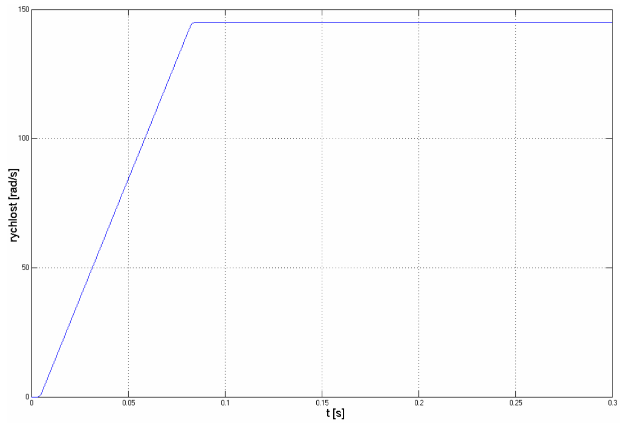



Fig. 7. Courses of angular velocity of DTC method with the help of fuzzy

More noticeable differences between the methods are represented by the course of moment. On Fig. 8 is the comparison of the courses of moment at Depenbrock method of DTC, at Takahashi method of DTC and at method of DTC with the help of fuzzy. We can see from the course that the moment at Depenbrock and Takahashi method are noticeably more oscillated as in case of DTC method with the help of fuzzy. This is one of the main advantages of this method compared to Depenbrock and Takahashi method.

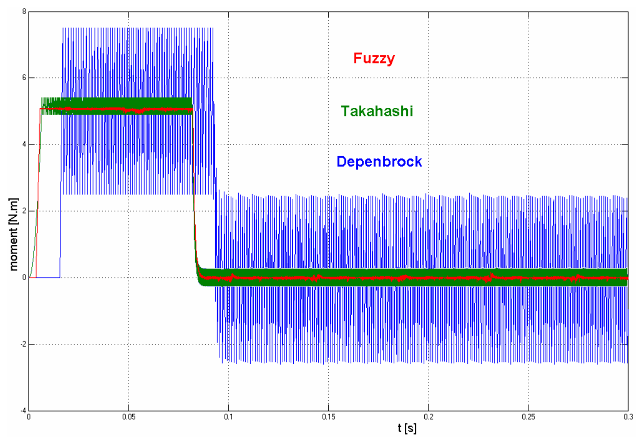



Fig. 8. Comparison courses of the moment for individual methods

In respect to the fact that in case of DTC with the help of fuzzy we try to maintain the constant magnetic flux at the required moment, it is obvious also the course of the vector of magnetic flux in coordinates X-Y on Fig. 9.

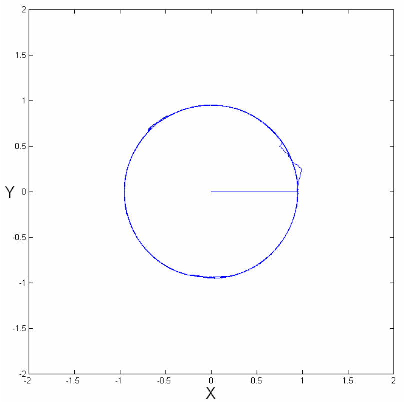



Fig. 9. Courses of the vector of magnetic flux in coordinates X-Y for method with the help of fuzzy

### B. *Courses in case of the reversal of angular velocity*

In simulation of these 3 methods we start up AM from zero value to value 145 rad/s. Until time of 0.15 s the speed is constant and the motor is reversed to value -145 rad/s. In time 0.4 s we again reverse into positive value 60 rad/s Fig. 10. At none of the methods occurred any problems with reversal.

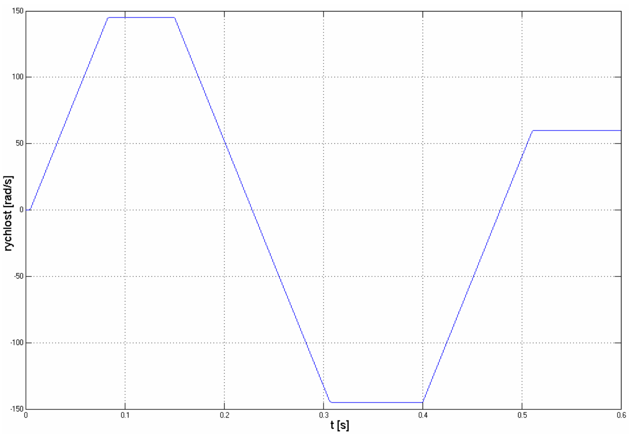



Fig.10. Courses of angular velocity at reversals at DTC with the help of fuzzy

For moment at reversals are characteristic the courses on Fig. 11. for all 3 methods. It is obvious again like in case of Depenbrock and Takahashi method the great oscillation of the moment. In case of fuzzy method occurs only a very slight oscillation and there are no overshoots here when is achieved the zero moment.

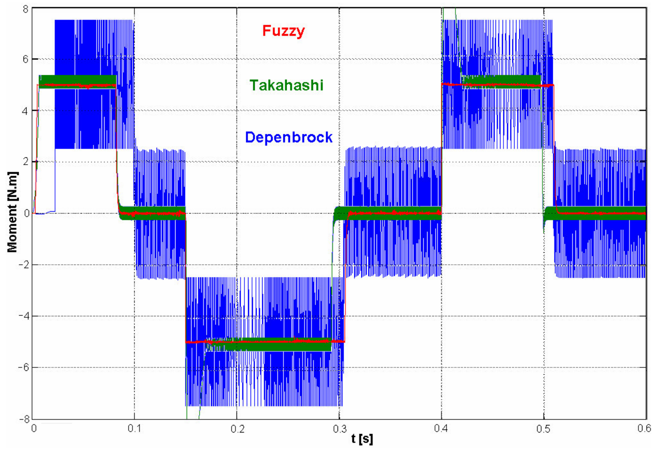



Fig.11. Courses of the moment at reversal

## V. Conclusion

This article deals with direct torque control of asynchronous motor with the help of fuzzy regulator and its comparison with other 2 methods of DTC. The comparison of these 2 methods (Depenbrock and Takahashi DTC) with DTC with the help of fuzzy is based on their courses. The courses were simulated for: a) the start up of asynchronous motor
b) reversal of angular velocity

In courses of Depenbrock´s method and Takahashi´s method of DTC is obvious the noticeable oscillation of the moment in comparison with fuzzy method. This high oscillation causes in operating conditions great noise and also excessive vibration not only of motor's shaft but also contingent vibration of the whole system. As far as concerns the required angular velocity, it can be achieved in case of all methods in approximately the same time and individual courses do not differ very much from each other.

## Acknowledgment

The assignment has been solved within the grant VEGA No. 1/4076/07

## References

[1] J. Timko, J. Žilková, P. Girovský, "Electrical drives," ISBN 80-8073-529–8, Košice.
[2] M. Žalman, I. Kuric, "Direct torque and flux control of induction machine and fuzzy controller," Journal of Electrical Engineering – No.1. Volume 5.2005.
[3] M. Depenbrock, "Direct Self Control (DSC) of Inverter-Fed Induction Machines," IEEE Trans. Power Electronics, Vol. PE-3, No.4, pp 420-429, 1988.
[4] C.D. Sousa, B.K Bose, "A Fuzzy Set Theory Based Control of a Phase-Controlled Converter DC Machine Drive," IEEE 1991.
[5] P. Brandštetter, L. Štěpanec, "Fuzzy Logic Control of Induction Motor Drive," IWCIT'01, VŠB-TU Ostrava, 2001, ISBN 80-7078-907-7.

# Low Power Low Cost Video Processing Methods for Driver Assistance Systems

*Martin Fifik*

Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

martin.fifik@tuke.sk

***Abstract*— This paper presents low power low cost video processing method for driver assistance system. These methods are here named, compared and described. Low power and low cost means, that we can use this system in every common vehicle with no expensive requirements to realize it. Potential of every method is discussed.**

***Keywords*—Driver assistance system, shape detection, color segmentation, neural networks.**

## I. Introduction

Video-based object recognition has many applications in different fields such as image retrieval, surveillance systems and driver assistance systems to name just some [1]. Every designed system in this field has some specific requirements on e.g. input data (image versus video), requirement on computational time, cost price and expectations on reliability. In general there is some basic idea: the input data are processed, and then from these data are extracted some features. Features depend on designed system. These features are used to recognition [1]. We focus ourselves in the field of driver assistance. We want here compare different methods in this field. The goal of this compare is:

- List different methods with their parameters,
- Point to methods that are perspective to future.

## II. Driver Assistance System

Driver assistance systems are systems to help the driver in its driver process. In driver assistance systems the most important thing is car safety and more generally road safety. The goal for developers is to create an ultimate system in car that can warm us or safe us before potential danger.

Preventive safety applications also help drivers to:

- Maintain a safe speed,
- Keep a safe distance,
- Drive within the lane,
- Avoid overtaking in critical situations,
- Safely pass intersections,
- Avoid collisions with vulnerable road users,
- Reduce the severity of an accident if it still occurs.

Driver assistance systems we can separate in two groups:

- Assistance systems for car support (safety),
- Assistance systems for driver support.

Assistance systems for car support: if these systems are activated, then they take control over the car, and driver can't affect this process. These systems must work fast and precisely. Assistance systems for car support involve Antiblock Braking System (ABS), Electronic Stability Program (ESP), Braking Assistant (BA), Electronic Brake-force Distribution (EBD), and etc.

Assistance systems for driver support: support driver indirectly, these systems inform driver about situation or warn before danger. These systems never take control over the car. The responsibility is in driver hands. Assistance systems for driver support involves Adaptive Cruise Control (ACC), Head Up Display (HUD), infrared night vision, Lane Departure Warning (LDW), Global Positioning System (GPS), Acoustic Parking System (APS), and etc.

Most often used sensors for these systems are optical cameras. For additional information are used e.g. Radar or Lidar. In this system can be used gray-scale or color cameras. Gray-scale cameras we can use only shape detection, but when used color camera, we can used color segmentation and shape detection, so we spare more time in terms of computational time (system with color camera will be faster).

Advanced driver assistance systems involve many approaches. The developed systems try to detect all possible dangers, so the systems must detect static objects or moving objects. In first case, to detect static objects only one image/frame is used, but in second case two or more image/frame is required. We focus on detecting static objects. Different static object recognition methods are compared in Table 1.

In traffic sign detection system described in [2], the input is taken from color camera. Input image is converted in to HSV color space, then are extracted red and blue color areas in the image. This process generates two binary images. To detect possible sign regions is used threshold method [9]. Obtained ROIs are tested for shape information. Different methods are used to detect circle, square, and triangle. After these steps is used a simple table classification methods to determine class type of traffic sign. Overall detection with this system was 94%. Average processing time was 2.285 seconds/image on Pentium 4 at 3GHz for 800x600 input images.

Another approach is described in [3]. Again color camera is used on input, and then image is processed in RGB color space. Three binary images for RGB components are generate. All pixels referred to same region are labeled together. For shape detection pattern matching is used. For classification neural network for each class type is used. At end it recognizes concrete traffic sign.

Different approach from previous two is described in [7]. Detection of ROI(s) is based on Bayesian classification for color information in the captured image. From each ROI(s) is build SIFT descriptors. These descriptors are used to detect possible sign.

Another advance in the traffic sign recognition is proposed in [8]. Detection is based on AdaBoost and Haar wavelet features. Classification is here based on a Gaussian probability density modeling. This approach is called a generic concept. The system error rate was 15%. Next static object worthwhile to detect are road lines [6]. First it generates a bird-eye view of road through a perspective transform. Second, it segments the pixels which belong to longitudinal road markings. Next, ego-lines are extracted by the Hough Transform and finally, the pitch angle is corrected, and the lane boundaries are classified.

## III. Conclusion

The static object detection for driver assistance systems has many possibilities how to realize it.

All methods described here are based on two step detection scheme: Hypothesis Generation (HG) and Hypothesis Verification (HV). At HG are used different methods like AdaBoost detection, Haar wavelet, SIFT descriptors, Perspective transform, Hough transform, binarization, Fourier analysis, symmetry detection, color segmentation and etc. For these named methods the main goal is extract features from input image. After this extraction, features are used to HV step. HV involves classification like: neural networks, template matching, table classification and Bayesian classifier. All these proposed systems can be realized with no special expensive computer requirements, so it can be used in every common vehicle.

None of these methods is 100% correct. Studying, modification, improving and upgrading all these methods will be subject to further work.

## References


[1] Laika,A. - Stechele,W.: A review of different object recognition methods for the application in driver assistance systems, Image Analysis for Multimedia Interactive Services, 2007. WIAMIS apos;07. Eighth International Workshop on Volume, Issue, 6-8 June 2007, pp 10 - 10

[2] Paulo, F.P. - Correia, L.P.: Automatic detection and classification of traffic signs, Image Analysis for Multimedia Interactive Services, 2007. WIAMIS apos;07. Eighth International Workshop on Volume, Issue, 6-8 June 2007, pp 11 – 11

[3] Broggi,A. - Cerri,P. - Medici,P - Porta,P.P.: Real Time Road Signs Recognition, Intelligent Vehicles Symposium, 2007 IEEE Volume , Issue , 13-15 June 2007, pp 981 - 986

[4] Turán, J.: Fast translation invariant transforms and their applications, Elfa, Kosice, 1999

[5] Turán, J. et al.: Trace transform and KLT based invariant features and image recognition systém, In Acta electrotechnica et informatica, vol. 6, nº3, 2006, pp. 5-15

[6] Collado J.M. – Hilario C. – Armingol J.M. – Escalera de la A. : Adaptive road lines detection and classification, Lecture Notes in Computer science, 2006, vol. 4179, pp. 1151-1162

[7] Farag A.A. – Abdel-Hakim E.A.: Detection, categorization and recognition of road signs for autonomous navigation, Advanced concepts for intelligent vision systems, September 2004,

[8] Bahlmann C. – Zhu Y. – Ramesh V. – Pellkofer M. – Koehler T.: A system for traffic sign detection, tracking, and recognition using color, shape, and motion information, IEEE Intelligent Vehicles Symposium, June 2005, pp. 255-260

[9] Otsu. N.: A threshold selection method from gray-level histogram, IEEE Transactions on Pattern Analysis and machine Intelligence, vol. 25, nº8, August 2003, pp. 959- 973

[10] Turán, J. – Ovseník, Ľ. – Turán, J. Jr.: Transform based invariant feature extraction, IWSSIP 2006, Budapest, 2006, pp. 79-82


TABLE 1: DIFFERENT STATIC OBJECT RECOGNITION SYSTEMS

| Name, Title and References | Recognition Rate | Computing Cost | Methods | Classification |
|---|---|---|---|---|
| Automatic Detection and Classification of Traffic Signs[2] | 94 | 2.285 | Shape detection, color segmentation, symmetry | Table classification |
| Real Time Road Signs Recognition[3] | not specified | 0.1 | Shape detection, color segmentation | Neural networks |
| Adaptive Road Lines Detection and Classification[6] | not specified | 0.1 @ P4 | Perspective transform, Hough transform, Fourier analysis, | Table classification |
| Detection, Categorization and Recognition of Road Signs for Autonomous Navigation[7] | not specified | not specified | Shape, color detection, SIFT descriptors | Bayesian classifier |
| A System For Traffic Sign Detection, Tracking, and Recognition Using Color, Shape, and Motion Information[8] | 85 | 0.1 @ 2.8GHz Xeon | AdaBoost detection, Haar wavelet | Bayesian classifier |

# Possibilities of using the methods for multicriterial decision making in electricity supply system's asset management

*Peter FRÁK*

[1]Dept. of Power Engineering, FEI TU of Košice, Slovak Republic
[2]Grid strategy, Východoslovenská distribučná, a.s., Košice, Slovak Republic

frak_peter@vse.sk

***Abstract*** **The paper deals with current trends of asset management within utilities. Distribution companies, which supply electricity via electric grid, became managed by private capital here in Slovakia 5 years ago. On the one side these companies are motivated to make profit, on the other side there are requirements from the customers respectively regulatory to ensure reliable and safe electricity supplying. To be more closely to these two issues it is necessary to realize comprehensive and sophisticated analyses and consequential optimizations for responsible decision making. One of the ways to get answers for e.g. future financial development influenced by improvement of reliability is to know condition of grid – its state.**

***Keywords*** **Asset management, reliability, asset state, multicriterial analysis.**

## I. INTRODUCTION

Electricity supply system means the technology chain consisted of following parts: generation – transmission – distribution – consumption of electricity. Each of these components is characterized by relevant signs. Hereby I will focus on distribution part which is represented mostly by electric lines, substations and related devices.

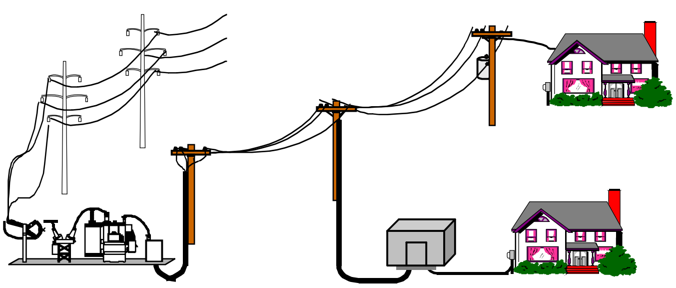


Fig. 1. Visual diagram of distribution grid consisted of electric lines and substations of various voltage levels.

## II. ASSET MANAGEMENT

Targets of distribution companies' asset management are especially to:

• Ensure that the required standard service levels are met, including reliability of supply to customers.
• Provide a safe environment for operating personnel and the general public.
• Proactively manage environmental issues.
• Manage the assets to achieve the required functionality, performance and value of assets to enable the continuation of a viable network business.

Asset management gives an overview and solutions to utilities approach to maintenance, development, asset renewal and replacement.

## III. RELIABILITY AND ASSET STATE

There are a lot of definitions and evaluations of electricity supply reliability. Here I want to focus on reasons of unreliability or worsened reliability. Well known general reliability bath tube curve shows that when asset is getting older then there is higher probability of device malfunction. This statement relates with age of asset. However an empirical practice shows that, concerning system devices malfunctions, not only age is crucial but also asset (devices) state. Based on real experiences from operation, it is no exception when aged asset shows better function compared to new ones because of their condition. Here the questions appeared what the asset state is and how to evaluate it, how to measure and influence it?

State is qualitative indices of device. It is possible to influence asset state mostly by doing asset replacement, proper maintenance, inspections and so on – all issues related with asset management. Of course it is necessary to find the way what will be the indicators for right asset state description. Then it is necessary to make wide overview of existing grid in some form and to collect these data for using them for analyses. Mostly due to numerous lengths and counts of devices in distribution grid it seems to be the substantial challenge to do such work from the point of required tools and databases and also from the point of time.

Multicriterial analysis seems to be good approach to find solutions for dealing with knowing asset state.

## IV. ASSET STATE

### A. Structure of distribution grid

Distribution grid consists of devices separated mostly by voltage levels (high voltage, middle voltage, low voltage). There are various types of overhead and cable lines divided by cross-section, type of conductor, type of insulation, supporting points, insulators and so on in distribution grid. Transformer stations or substations include also the quantity of various types of devices like circuit breakers, disconnectors, bus bars, electric protections and more others.

One of first steps to make analysis in asset state is to create particular asset groups and to describe their properties which are important concerning asset state.

Additional issue is to make the state evaluation methodology from the point of particular states definitions as well as their assessment, acquisition and interpretation.

### B. State of distribution grid

In ideal situation the asset state relates only with asset age what can be seen like example in *Fig. 2*. There are chosen three asset states - green line represents asset in good condition, yellow is bad and red is critical asset. Violet line is total amount of asset.

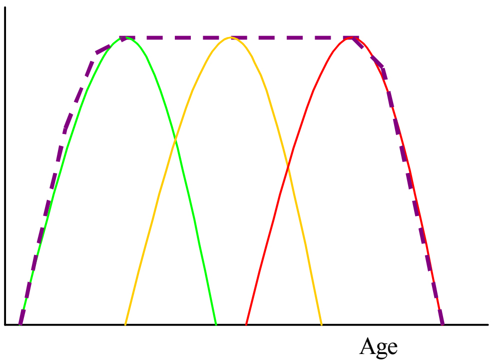



Fig. 2. Example demonstration of ideal situation - asset state depends only on asset age.

How mentioned above the asset state doesn't correspondent considerably to the asset age profile in linear dependence in reality. It is due to a lot of factors like device installation, operation, maintenance, loading, external impacts and more. The sound indicator of good or wrong asset state is reliability or unreliability. From this point the known reliability bathtub curve should be built in that way that outage probability will not depend on asset age, but on its state.

If we chose only three states of some asset type, e.g. as shown in *Fig.2*, then the next formula is valid in any instant of time.

$$\sum total\ \ asset = \sum_i state\,1 + \sum_j state\,2 + \sum_k state\,3 \quad (1)$$

Due to enormous number of distribution asset groups and devices types it is suitable to evaluate the state of small parts of concrete device what will influence whole device asset state. It is necessary to say that due to each element's different properties, the impact on device will be also different – here the new term appeared – weight.

## V. MULTICRITERIAL ANALYSIS

### A. Multicriteria decision making

$$\max\{f_1(a);\ f_2(a);\ \dots;\ f_k(a) : a \in A\} \qquad (2)$$

A is a set of n possible decision alternatives, $f_1$, $f_2$... $f_k$ are criteria by which the alternatives are evaluated. If all the criteria are not equally important, their weights can be marked by $w_1$, $w_2$... $w_k$. The basic information for this kind of decision problem can be shown in the form of an evaluation table:

| | $f_1(.)$ | $f_1(.)$ | … | $f_j(.)$ | … | $f_k(.)$ |
|---|---|---|---|---|---|---|
| | $w_1$ | $W_2$ | … | $w_j$ | … | $w_k$ |
| $a_1$ | $f_1(a_1)$ | $f_2(a_1)$ | … | $f_j(a_j)$ | … | $f_k(a_1)$ |
| $a_2$ | $f_1(a_2)$ | $f_2(a_2)$ | … | $f_j(a_j)$ | … | $f_k(a_2)$ |
| … | … | … | … | … | … | … |
| $a_i$ | $f_1(a_i)$ | $f_2(a_i)$ | … | $f_j(a_i)$ | … | $f_j(a_i)$ |
| … | … | … | … | … | … | … |
| $a_n$ | $f_1(a_n)$ | $f_2(a_n)$ | … | $f_j(a_n)$ | … | $f_k(a_n)$ |

Table 1. Evaluation table

To solve the problem (2) means to choose the best alternative or to rank all the alternatives. This kind of decision problem solving, including a discrete set of explicitly described alternatives, is different from the multicriteria optimization problem solving where the set of alternatives has been determined implicitly by constraints. To solve the second problem the vector optimization was used. The efficient solutions or Pareto optimal solutions of the vector optimization problem can be characterized in a few ways which enable their identification, but the concept is not usable to solve the problem (2). To solve the problem (2) it is necessary to use the procedure which meets the following criteria: (i) aberrations in evaluation of alternatives according to the individual criteria should be considered, (ii) the scaling effect, which occurs as a result of the use of different measuring scales used for evaluation of alternatives according to different criteria, must be eliminated and (iii) there must exist a possibility for a clear interpretation of the weight of the
criteria. It is necessary to stress that the method must enable that for each pair of alternatives *a; b* ϵ *A* during their comparing, one of the following statements can be chosen:

- *aPb* or *bP a* - *a* is preferred to *b* or vice versa
- *aIb* - *a* and *b* are indifferent
- *aRb* - *a* and *b* are incomparable.

The main concepts for problem solving (2) satisfying the mentioned conditions are:

(a) multiattribute utility function (value function) for deterministic case,

(b) scalarizing of the problem (2),
(c) introduction of an outranking relation into the set $A$.

(a) It is necessary to assess a value function $v(.)$ of the form

$$v(a) = w_1v_1(f_1(a)) + w_2v_2(f_2(a)) + \ldots + w_kv_k(f_k(a)) \qquad (3)$$

where $v_j(f_j(a))$ are assumed to be strictly positive, and maximise it over the set $A$. The Saaty method known as the Analytic Hierarchy Process (AHP) has been shown as an example of the method explicable by this theoretical approach.

(b) A few methods have been developed which characterize the efficient solutions of the multicriteria programming problem by its transformation into a scalar form. The compromise programming i.e. the search of the solution closest to the ideal solution, is the method which can also be used for discrete problem solving. The example of this concept for problem solving can be recognised in the TOPSIS method.

(c) According to the information from the evaluation table, the relation which ranks alternatives partially or totally has been introduced into the set $A$. The examples can be mentioned methods ELECTRE (I and II) and PROMETHEE (I and II).

### B. *Analytic hierarchy process*

The Analytic Hierarchy Process (AHP) developed by T. Saaty [1] is useful as a decision making methodology when multiple costs and multiple benefits are relevant for determining the priorities of the alternatives.

The basic steps in constructing and examining an AHP model are: (1) decompose the problem into a hierarchical structure, (2) perform judgments to establish priorities for the elements of the hierarchy, (3) synthesis of the model, (4) perform a sensitivity analysis.

There exist different types of AHP hierarchies, but for this paper it is enough to understand the basic AHP model which includes the goal, criteria and alternatives (*Fig. 3*).

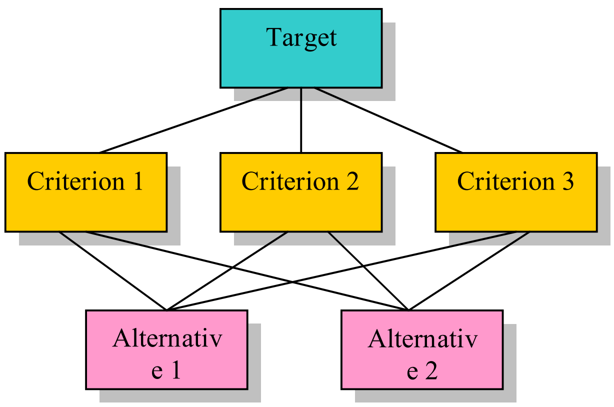



Fig. 3. Basic AHP model with target, criteria and alternatives

After constructing AHP hierarchy, it is necessary to perform judgments to establish priorities for the elements of the hierarchy. The decision maker's judgments about the relative importance between two elements in each pair of all elements on the same level of the hierarchy are expressed by the help of e.g. the nine-point intensity scale (the fundamental scale).

Once all judgments have been performed, they are all synthesized by the help of a related mathematical model (4) which is briefly described here. Let $n$ be the number of criterion (or alternatives) for which we want to find the weights $w_i$. Let $a_{ij} = w_i/w_j$ , where $w_i$ is the weight of $i$-th criterion (or priority of $i$-th alternative) be the element of matrix $A$. The pairwise comparison matrix $A$ and vector $w$ satisfy the equation

$$Aw = nw \qquad (4)$$

Because of a special form of the matrix $A$ (each row is a constant multiple of the first row, all elements are positive and $a_{ij} = 1/a_{ji}$), rank of $A$ is one, all eigenvalues are zero, except for one, and nonzero eigenvalue has a value of $n$.

If the matrix $A$ contains inconsistencies, the vector $w$ of the weights can be obtained using the equations

$$(A - \lambda_{max}I)w = 0$$
$$\Sigma w_i = 1$$

where $\lambda_{max}$ is the largest eigenvalue of matrix $A$. Because of the characteristics of the matrix $A$, $\lambda_{max} \geq n$ and the difference , $\lambda_{max} - n$ can be used for measuring inconsistencies. A consistency index $CI$ has been constructing

$$CI = \frac{\lambda \max - n}{n - 1} \qquad (5)$$

and consistency ratio is defined as $CR = CI/RI$ , where $RI$ is a random index (random index is the consistency index of many randomly generated pairwise comparison matrices of size $n$). If the value of $CR$ is less or equal to 0; 10, the pairwise comparisons are considered to be acceptable. Otherwise, the comparisons must be repeated in order to resolve the inconsistencies.

## VI. Conclusion

One of the ways how to make asset state evaluation is described in this paper. Methodology Analytic Hierarchy Process (AHP) like a part of multicriterial analysis seems to be useful tool for analysis of asset state in distribution grid's asset management. Applying such analysis it is possible to evaluate the state of small parts of grid and consequently to assess asset state in general. Knowing the state of asset it is rather to estimate potentially required measures within grid over the next time.

## References

[1] J. Endrenyi, "Reliability modeling in electric power systems" J. Wiley and Sohns, N.Y., 1978.
[2] R. Brown, "Aging distribution infrastructure at Pacific and Electric Company", San Francisco, July 2005, pp. 3-18.
[3] T. Hunjak , "Mathematical foundations of the methods for multicriterial decision making", Mathematical Communications 2 (1997).

# Voltage Control Based on Reactive Power in the Transmission System and Energy Losses

*Daniel HLUBEŇ*

Department of Electric Power Engineering, FEI TU of Košice, Slovak Republic

daniel.hluben@tuke.sk

***Abstract*—This paper entitled "Voltage control based on reactive power in the transmission system and energy losses" deals with a problem of the losses in the transmission system in case of voltage control by control of reactive power of a generator.**

***Keywords*—reactive power, voltage control,**

## I. Introduction

The main goal of each operator, who is operating a transmission system, is the operation of a network with minimal losses. There are several possibilities, how to optimize electric energy losses in a transmission system.

In this article is described voltage control based on reactive power in transmission system in order to show the point, in which energy losses are the lowest.

### *A. Basics*

In my model the transmission system of the Slovak Republic is simulated. The voltage in control node is regulated by generator and the main goal is to find the value of reactive power, at which the voltage is in required range, but with lowest losses.

For this case was selected a node, in which the generator with the following parameters is placed:

Pmin = 20 MW
Pmax = 110 MW
Plom = 100 MW
Qlom = -37 MVAr
Q2min = -40 MVAr
Q1mi = -25 MVAr
Q1max = 35 MVAr
Q2max = 80 MVAr

These parameters define the range, in which it is possible to operate our generator. Parameters of the generator are shown on the following figure.

The simulation is made for 3 cases:

1. Active power of generator equals 20 MW
   Range of reactive power: -40 ÷ 80 MVAr
2. Active power of generator equals 100 MW
   Range of reactive power: -37 MVAr ÷ Q1lom.
3. Active power of generator equals 110 MW
   Range of reactive power: -25 ÷ 35 MVAr

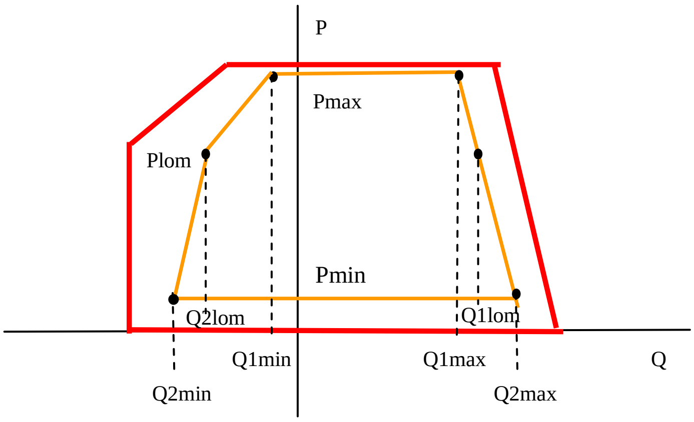



Fig 1. Operating chart of the generator

On the following figure (fig. 2) is shown figure 1 with the description of the possible cases:

- Active power
  - Generation
- Reactive power
  - Generation
  - Consumption

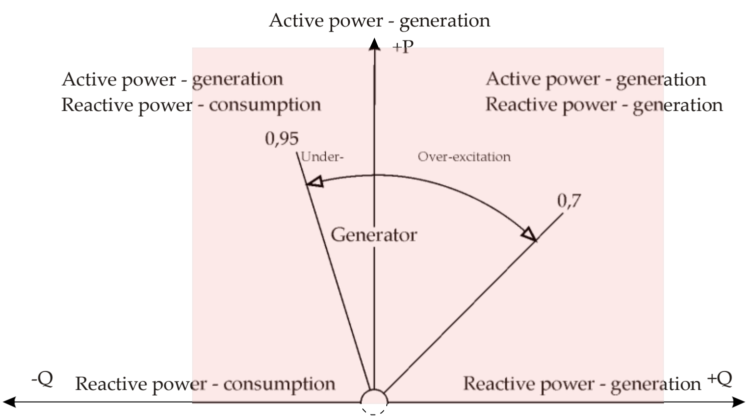



Fig 2. Operational diagram of the generator [2]

### B. *Results*

On the following figure can be seen results of these simulations. In this figure are 3 dashed lines and 3 solid lines.

The dashed line defines dependency of energy losses in windings of transformers and lines on reactive power generated by the generator.

The solid line defines dependency of voltage in the control node on reactive power generated by the generator.

## II. Conclusion

This model was built for voltage regulation in one node at the same time. But, the transmission system is a more complicated system and therefore it is necessary to develop other methods for optimal real-time voltage control.

The main goal of this article was to show dependencies of losses and voltages in case of voltage control based on reactive power.

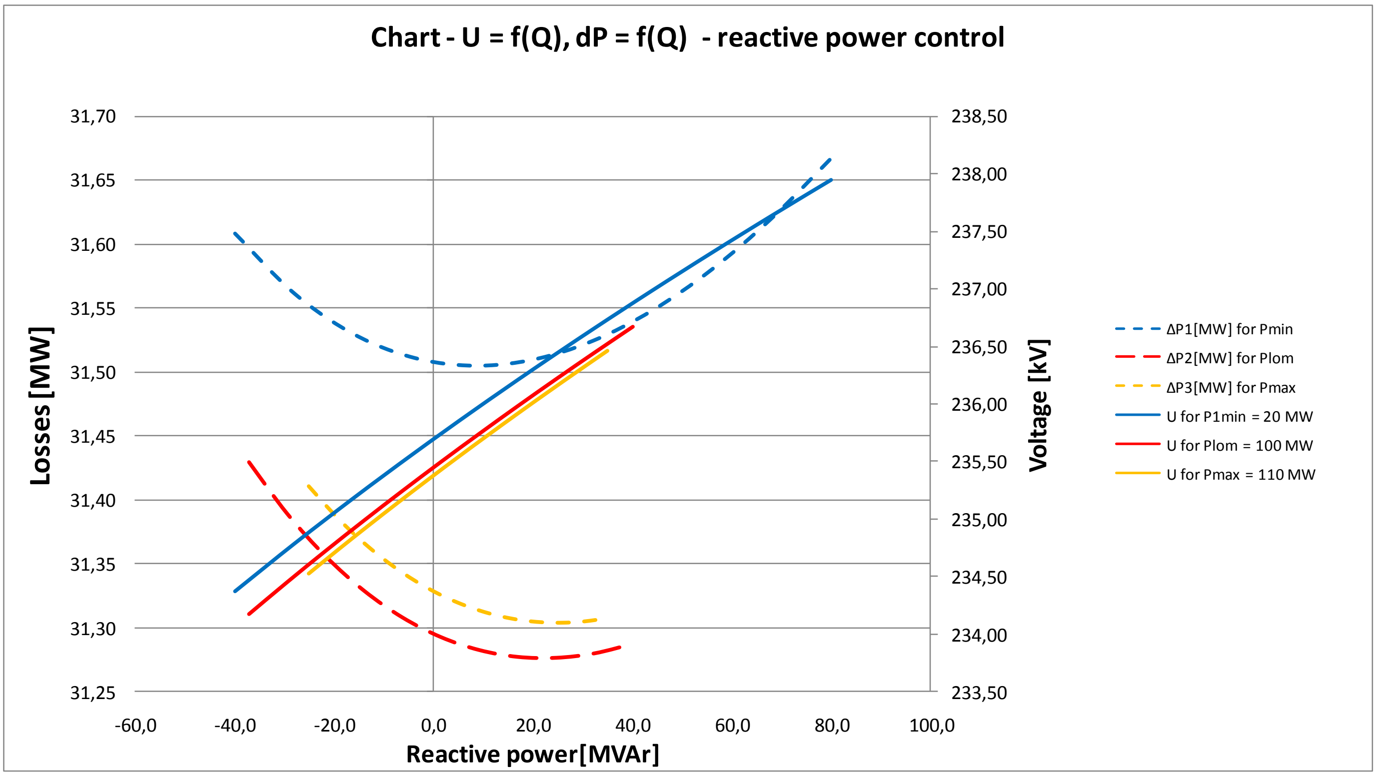



Fig 3 Voltage and losses dependency on reactive power in the regulation node

Table 1: Table of results

| | Pmin | Plom | Pmax |
|---|---|---|---|
| dPmin [MW] | 31.5 | 31.3 | 31.3 |
| Q [MVAr] for dPmin | 8.0 | 24.6 | 23.0 |
| U [kV] for dPmin | 235.9 | 236.2 | 236.1 |
| dPmax [MW] | 31.7 | 31.4 | 31.4 |
| dP [MW] if Q = 0 MVAr | 31.5 | 31.3 | 31.3 |
| Umin [kV] for Qmin/lom/max | 234.4 | 234.2 | 234.5 |
| Umax [kV] for Qmin/lom/max | 237.9 | 236.7 | 236.5 |
| U [kV] for Q = 0 MVAr | 235.7 | 235.4 | 235.4 |

## References


[1] Kolcun M., Chladný V., Mešter M., Cimbala R., Tkáč J., Hvizdoš M., Rusnák J.: Power plants, Košice 2006, pp. 111-112. (in Slovak)

[2] Kolcun M., Griger V., Beňa Ľ., Rusnák J.: Opertion of transmission system, Košice 2007, pp. 149, 185,

[3] Ministry of Industry and Trade of the Czech Republic: Regulation of Ministry of Industry and Trade, which determine details of efficiency of electricity usage in case of transmission, distribution and indoor electrical distribution, Collection of Laws No. 153/2001, part 60, pp. 3491 – 3506 (in Czech)


This work was supported by Scientific Grant Agency of the Ministry of Education of Slovak Republic and the Slovak Academy of Sciences under the project VEGA No. 1/4072/07.

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-0385-07.

# Connectivity of decentralized generation (DG) into distributive network

[1]*Peter Horos,* [2]*Marek Dzivák*

[1]Dept. of Electric Power Engineering, FEI TU of Košice, Slovak Republic
[2]Dept. of Electric Power Engineering, FEI TU of Košice, Slovak Republic

[1]peter.horos@gmail.com, [2]marek.dzivak@gmail.com

***Abstract*—Distributed generation (DG) is expected to grow rapidly in near future generation. The most important advantages comparing with traditional generation like coal plants is environmental aspects, location near by users, also the shorter construction period. However, there are also some severe barriers before wider employment. One of these is the connection to the network, especially distribution networks, because there are practically no production units connected into them. For this purpose new rules and recommendations are needed.**

***Keywords*—distributive network, decentralized generation, distributed generation, connectivity.**

## I. Introduction

Electricity was originally generated at remote hydroelectric dams or by burning coal in the city centers, delivering electricity to nearby buildings and recycling the waste heat to make steam to heat the same buildings. Rural houses had no access to power. Over time, coal plants grew in size, facing pressure to locate far from population because of their pollution. Transmission wires carried the electricity many miles to users with a 10 to 15 percent loss. [9] For all that, new ways of generation of heat and electricity has to be developed. Distributed generation appeared as a practical solution, so we can say that, the share of distributed generation (DG) is expected to grow rapidly in near future. In addition to the environmental aspects, location near by users, also the shorter construction period can be considered as a key driving force accelerating the development. However, there are also some severe barriers before wider employment of DG is possible. One of them is the connection to network or grid. Especially when we consider distribution networks the situation is totally new since traditionally there are practically no production units connected in the distribution network. For this purpose new rules and recommendations are needed. Furthermore, new technical solutions are necessary to make DG economically more viable. From distributed generation technologies especially wind power has already been largely applied, but also other technologies, such as gas, diesel and biomass fired micro and mini turbines and CHP devices, solar cell systems etc., are in the phase of commercialisation. Flexibility for the user and environmental aspects are often mentioned as the benefits of distributed energy generation. Connection of lot of small-distributed power generation units to the distribution network will have consequences related both to technological and legal matters. When DG becomes more common and the unit sizes increase, the effects on distribution network planning and operation will also increase. From the technical point of view the connection of DG units to distribution network or grid (i.e. the DG interconnection) is a challenging task. The basic reason for this is that the distribution networks are typically not planned for having some generation. E.g., the traditional relay protection system of a medium voltage (MV) network relies on the fact that the fault current is fed from one direction only. [2]

## II. DISTRIBUTED GENERATION

*Definition*

There is no single one definition for the distributed generation. Generally mentioned criteria
are:

- Not centrally planned and dispatched
- Usually connected to the distribution network

The rating of DG units is usually said to be small but the given MW limits vary considerably the highest values being even up to 100 MW. A good discussion about the terminology is presented in [4].

*The interconnection*

DG units can be grid independent or grid parallel as well as a combination of the both. In the latter case a grid failure means that the DG unit is disconnected from the grid and continues to operate independently from the grid and thus creates an “island” (islanding, island mode operation). A typical arrangement for the DG interconnection to the medium voltage network is depicted in Figure 1. Connection and disconnection of the generator is made by the circuit breaker at the generator side of the main power transformer (main breaker). Depending on the size of the plant the disconnector on the grid side of the transformer may be replaced by circuit breaker. The general scheme presented in Figure 1 illustrates interconnection of DG Technologies based on synchronous (or asynchronous) generator. Other DG technologies apply slightly different interconnection arrangements. In all cases the voltage level at the interconnection point determines the need for a transformer. Smaller units can be directly connected to the low voltage network.

*Types of sources*

Considering only the electrical characteristic there are three different DG types:

- synchronous generator (alternator)
- asynchronous generator
- invertor (frequency inverter)

The first two types represent traditional technology based on rotating electrical machines. The last type refers here various arrangements applying modern power electronic converters. From the interconnection point of view these three types have different impacts on the distribution network.

## III. WIND POWER

In a wind turbine the kinetic energy of streaming air is converted to electric power. The size of wind turbines has increased rapidly during the past two decades, the largest units being now about 4 MW. For the smaller units the typical construction is fixed speed stall regulated turbine. Units larger than 1 MW are equipped with variable speed system in order to withstand the increased mechanical stresses. Single units are typically connected to a medium voltage (MV) network (in Nordic countries typically 10 or 20 kV). For larger wind parks connection to the high voltage (HV) grid is necessary. [5] The most common generator type is the asynchronous generator. If an asynchronous generator is directly connected to the generator a soft-starter is needed in order to minimize the large currents at the generator startup. During normal operation directly connected generator may still cause some increase in the flicker levels and in the variation of the active power flow. Recently various arrangements applying modern inverter technology have come into market allowing variable speed systems, where the power output can be held relatively constant independently from the wind speed variations. For these kinds of systems both synchronous and asynchronous generators can be applied. [6]

## IV. RECIPROCATING ENGINES

Reciprocating engines, developed more than 100 years ago, were the first among DG technologies. Both Otto (spark ignition, SI) and Diesel cycle (compression ignition, CI) engines have gained widespread acceptance in almost every sector of the economy. They are 4 used on many scales, ranging from small units of 1 kVA to large several tens of MW power plants. Smaller engines are primarily designed for transportation and can usually be converted to power generation with little modification. Larger engines are most frequently designed for power generation, mechanical drive, or marine propulsion. Reciprocating engines are usually fuelled by diesel or natural gas, with varying emission outputs. Co-generation configurations are also available with heat recovery from the gaseous exhaust. Heat can also be recovered from the cooling water and the lubrication oil [7]. Typically, synchronous generators are applied with internal combustion engines although some examples can be found where induction generator is applied.

## V. PHOTOVOLTAIC

Photovoltaic (PV) systems convert the sunlight directly to electricity. PV technology is well established and widely used for power supplies to sites remote from the distribution network [7]. Photovoltaic systems are commonly known as solar panels. PV solar panels are made up of discrete cells that convert light radiation into electricity connected together in series or parallel. Current units have efficiencies of 24% in laboratory conditions and 10% in actual use [5]. The maximum theoretical efficiency that can be attained by a PV cell is 30% [8]. PV units are connected to the network applying inverter. This kind of arrangement will potentially cause harmonics unless they have been filtered properly. On the other hand, the inverters of PV system could operate, in the future, as active filters to reduce low order harmonics in the distribution system.

## VI. MICRO-TURBINES

Distributed generation with micro-turbines is a new and fast growing business. The market is worldwide. In the Nordic countries micro-turbines are expected to be operated in combined heat and power mode. The reason for this is that the cost of power is close to the cost of heat. For each produced kilowatt-hour of electricity the micro-turbines will produce two kilowatthours of heat. The micro-turbines could also be used for peak shaving, stand-by power, capacity addition, stand-alone generation and others. In the case of capacity addition the short time from decision and order to operation will be a heavy argument for DG with microturbines in the future. [5] Most micro-turbines use a turbine mounted on the same shaft as the compressor and a highspeed generator rotor. The rotating components can be mounted on a single shaft that spins up to 96 000 rpm. The high frequency AC current from the generator is first rectified and then converted to AC at grid frequency. [5]

## VII. OTHER DG TECHNOLOGIES

Above only four different DG technologies were briefly introduced. In addition to these there are number of technologies that are still in early development phase, such as the fuel cell technology, or technologies relying on the traditional steam cycle process, but applying some renewable energy source (e.g., biofuels). Overview of different technologies is presented in the following table.

*Power Quality*

a) ***Relative change of voltage***- it depends on short-circuit power of network at the point of connection and power of source.

*b)* ***Voltage level***- Voltage quality problems may also arise during normal operation of DG when considering the voltage level. When a distribution feeder is designed to carry a certain power flow from the primary substations to the loads a generation unit along the feeder may cause a reversed power flow and a voltage rise. This is schematically illustrated in the following figure. Some practical examples may be found, e.g., in [1].

c) ***Flicker effect***- Feeling of labile optical perception caused of twinkling lights. It is caused by voltage fluctuations. Sources of flicker are performance electronic, frequency changers, rectifiers, etc.

d) ***Harmonics***- When using inverter based applications the increased level of harmonics may become a problem. The amount of harmonics as well as their spectrum depends on the type of inverter applied. A PWM inverter generates harmonics at multiples of its switching frequency. There have traditionally been no limits or recommendation to the harmonics at switching frequencies. Older thyristor converters generate 5th, 7th, 11th, 13th (and so on) harmonics of the 50 Hz base frequency. The total harmonic distortion (THD) of the current is usually high and above the standard limits without proper filtering.

e) ***Rapid voltage changes***- Voltage control and quality problems may arise when generators embedded within the distribution network start/stop generating. This might cause other network users to suffer voltage fluctuation, dips and steps outside of the statutory limits. Rapid voltage changes are defined as a single, rapid change of the voltage RMS value where the voltage change is of certain duration. They may occur, e.g., at switchings in a wind farm.

## REFERENCES


[1] Repo, S., Laaksonen, H., Järventausta, P., Huhtala, O., Mickelsson, M., *A case study of a voltage rise problem due to a large amount of distributed generation on a weak distribution network,* PowerTech, Bologna, Italy, 2003

[2] Kauhaniemi, K., Nordac 2004 et al paper.pdf.

[3] Kauhaniemi, K. (ed.), CODGUNet WP4 - *Technical analysis of network connection of different types of generation units*, Technobothnia, Vaasa, 22.04.2003B. Smith, “An approach to graphs of linear forms (Unpublished work style),” unpublished.

[4] Ackermann, T., Andersson, G., Söder, L., *Distributed Generation: a Definition*, Electric Power Systems Research, 57, 2001, pp. 195–204

[5] Häggmark, S., Neimane, V., Axelsson, U., Holmberg, P., Karlsson, G. Kauhaniemi, K., Olsson, M., Liljegren, C., *Aspects of Different Distributed Generation Technologies – CODGUNet WP 3*, Vattenfall Utveckling Ab, 2003-03-14.

[6] Slootweg, J.G., de Haan, S.W.H., Polinder, H., Kling, W.L., *Aggregated Modelling of Wind Parks with Variable Speed Wind Turbines in Power System Dynamics Simulations*, 14th PSCC, Sevilla, 24-28 June 2002.

[7] Jenkins, N., Allan, R., Crossley, P., Kirschen, D., Strbac, G., *Embedded Generation*, The Institution of Electrical Engineers, 2000

[8] *DG Technologies Summary*, WWW-site: http://www.distributed-generation.com/technologies.htm

[9] CASTEN, R. Thomas. – DOWNES, Brennan : *The Case for Decentralized Generation of Energy*. Published on 20 Jan 2005 by The Skeptical Inquirer. Archived on 20 Jan 2005.

# Dielectric Spectroscopy of Insulation Systems

[1]František KOVAĽ, [2]Peter SEMANČÍK
[1]Department of Electro-Power Engineering, FEI TU of Košice, Slovak Republic
2Department of Electro-Power Engineering, FEI TU of Košice, Slovak Republic

[1]frantisek.koval@tuke.sk, [2]peter.semancik@tuke.sk

***Abstract*— This article describes insulation systems, degradation of this systems, temperature stress of Remikaflex and natural ester fluid insulation and calculation dissipation factor in frequency domain.**

***Keywords* — dielectric spectroscopy, insulation systems, relaxation current, dissipation factor.**

## I. INTRODUCTION

Insulation system is the most sensitive part of every electrical mechanism. Damage of this system, result in secondary impact to entire manufacturing processes mainly in big production factories.

Insulation degrades over a period of time because of various stresses affect upon it during its normal working life. The basic initiators for insulation degradation are electrical stress, mechanical stress, chemical attack, thermal stress and environmental contamination. Normal cycles of operation lead to aging through these mechanisms [1]. Interaction various factors together may significantly speed up the degradation processes.

## II. DIAGNOSTIC OF INSULATION SYSTEMS

No insulation is perfect, therefore a certain magnitude of current does flow through it. This current may be insignificantly small but for most practical purposes it is the basis of insulation testing [1].

There are changes of the material structure due to degradation. This changes lead to change of the insulation properties. No equation or relationship exists to determine exactly the rest life of the insulation. However, in present day, many methods exist through which we can observe changes in the insulation structure due to ageing and following the results we can estimate insulation rest life.

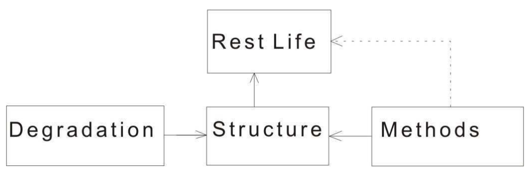



Fig. 1. Degradation of material

This article deal with degradation processes of natural ester fluids and Remikaflex 45.004 insulation which are used in electrical equipments. For this purpose we used two methods: measurement and analysis of the relaxation currents in the time domain (PDC, Polarisation and Depolarisation Current), of the dissipation factor tanδ in the frequency domain (FDS, Frequency Domain Spectroscopy).

## III. DIELECTRIC SPECTROSCOPY OF POLARISATION SYSTEMS

Insulation system can be described by an RC network model (fig.2) consisting of an ohmic resistor $R_0$ due to the conductivity of the insulation system, the capacitance $C_0$ (geometric capacitance of the insulation system) and the RC series branches $R_i$, $C_i$, which represent the time-depend polarization of the insulation system. $R_d$, $C_d$ represents dipole polarization, $R_{Ir}$, $C_{Ir}$ represents ion relaxation polarization, $R_m$, $C_m$ represents migration polarization, $R_s$, $C_s$ represents spontaneous polarization, $R_r$, $C_r$ represents resonance polarization and capacitances $C_e$, $C_i$ represents electronic and ionic polarization. There are no resistances $R_e$, $R_i$ because this type of polarization has no dielectric losses [2].

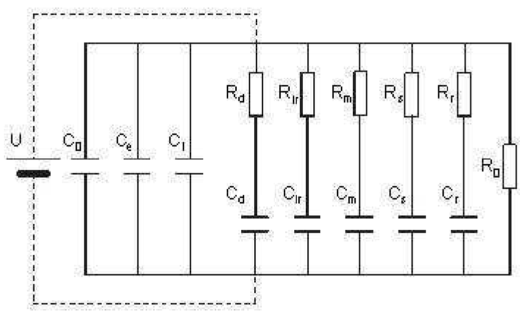



Fig. 2 Equivalent circuit to model a linear dielectric material (the Maxwell-Wagner model).

The insulation ageing status can be determined by measuring the relaxation currents in the time domain and in combination with intelligent software we can get elements of dielectric substitution diagram (fig.2) according to [3,4]:

$$R_i = \frac{U_0}{I_{mi}}, \qquad C_i = \frac{T_i}{R_i} \qquad (1)$$

In insulation system can occur more than 7 polarization processes how is describes in fig.2. However, we did just 1000 sec. measurements and in that time we could record by measuring of relaxation currents only 7 types of polarizations.

Polarization current presents equation (2).

$$i_t(t) = \frac{U_0}{R_0} + \sum_{i=1}^{n} \mathrm{I}_{mi} \exp\left(\frac{-t}{T_i}\right) \qquad (2)$$

Where: $U_0$ – applied DC voltage (100V),
$I_{mi}$ – currents in single branches,
$T_i$ – relaxation times constant

## IV. Measurement of Insulation Systems

Figure 3 shows network scheme for measuring of relaxation currents. Samples of natural ester fluids and Remikaflex 45.004 insulation was placed in electrode system. Then the electrode system was connected to test voltage (100V) by Keithley 617 electrometer and the sample was subjected to affect of this electrical field for 1000 seconds. Keithley 617 was connected to computer through IEEE 488.2 interface. By means of software we got the currents $I_{mi}$ in single branches of substitution diagram and the times $T_i$ which represents single actions. RC values we can calculation according equation (1).

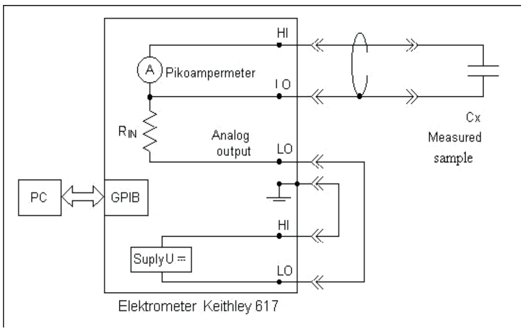



Fig. 3. Measuring equipment.

Figures 4 and 5 shows diagrams of time and currents (for Remikaflex samples) when software calculates with 6 types of polarizations. Diagrams compare changes of the times $T_1$ and currents $Im_{11}$, $I_0$ depending on degradation time which was zero hour (new sample); 24; 76.2; 183.6; 404.6; 859.1 and 1794.3 hours. Remikaflex samples were subjected to temperature stress (186°C) during those times. Conducting current represents $I_0$.

We didn't compare seventh approximation because we've got negative $I_{7i}$ at 24 hours aged sample and at 859.1 hours aged sample what wasn't right values

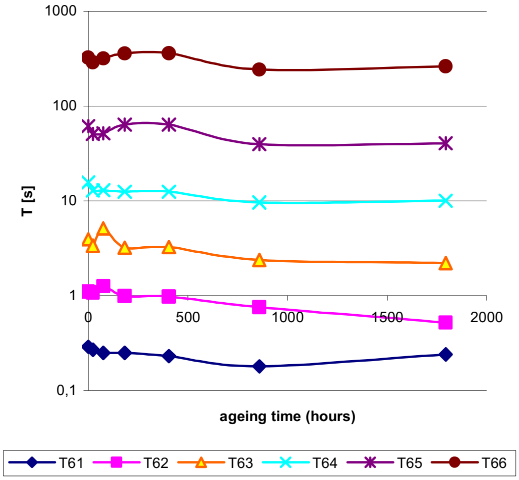



Fig. 4. Sixth approximation of $T_{6i}$ (Remikaflex)

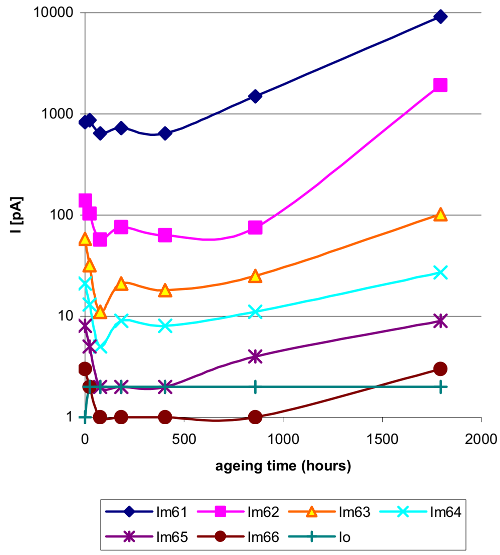



Fig. 5. Sixth approximation of $Im_{6i}$, $I_0$ (Remikaflex)

## V. Relation Between Time and Frequency Domain

### A. *Dissipation factor Tanδ (Remikaflex samples)*

The admittance Y of the RC model of a dielectric is:

$$\underline{I}(\omega)=\underline{Y}.\underline{U}(\omega) \qquad (3)$$

Then the dissipation factor can be easily derived out from the admittance according to the equations (3)

$$\tan\delta = \frac{\mathrm{Re}\{Y\}}{\mathrm{Im}\{Y\}} = \frac{\frac{1}{R_0} + \sum_{i=1}^{N} \frac{(\omega R_i C_i)^2}{R_i.\left(1+(\omega R_i C_i)^2\right)}}{\omega C_0 + \sum_{i=1}^{N} \frac{\omega C_i}{1+(\omega R_i C_i)^2}} \qquad (4)$$

By equation (4) we can do conversion tanδ into frequency domain. Parameters $R_i$, $C_i$, $R_0$ and $C_0$ we've got by measuring of relaxation currents and ω=2πf. If we will change the frequency in equation 4, then we can observe how does dissipation factor is changing in dependence on frequency. Figure 6 shows diagram of dissipation factor new Remikaflex sample at 0.001; 0.1; 0.2; 0.3; 0.4; 0.6; 0.8 and 1 hertz.

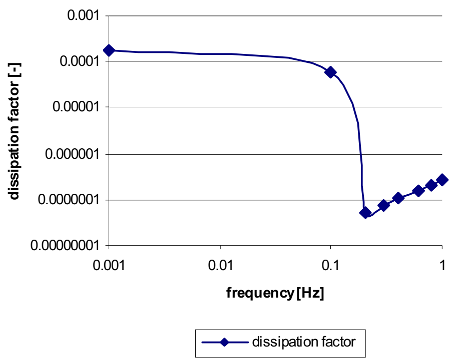



Fig. 6. Diagram of dissipation factor at sixth approximation

### *B. Hamon's approximation (natural ester fluids sample)*

The information obtained in either frequency or time domain is theoretically equivalent if the dielectric material can be described as a linear system (Fig. 2) and complex permittivity can be calculated by the Fourier transformation as [5]:

$$\hat{\varepsilon} = \varepsilon' - j\varepsilon'' \tag{5}$$

$$\hat{\varepsilon} = \varepsilon_\infty + (\varepsilon_s - \varepsilon_\infty)\int_0^\infty \varphi(x)e^{(-j\omega)}dx \tag{6}$$

Where $\varepsilon_\infty$ is the optical permittivity, $\varepsilon_s$ is the static permittivity, φ(x) is the regress function define by polarization current.

Conversion of polarization current into dissipation factor tanδ and into complex permittivity ε'' can be made by the Hamon's approximation [5].

$$\varepsilon'' = \frac{i(t)}{2\pi f C_0 U_c} \tag{7}$$

Where i(t) is measured polarization current of dielectric, $U_c$ is test voltage, $C_0$ is capacity calculated of empty vessel without oil. Measurements were done under temperatures varied from 20 °C to 100 °C.

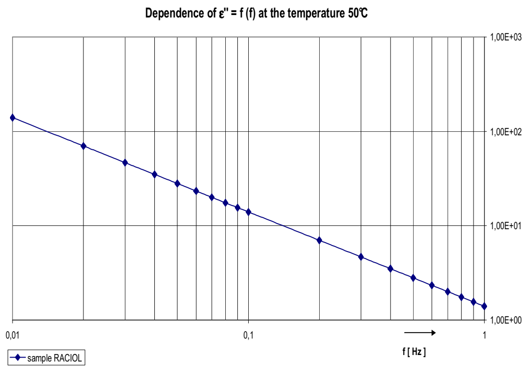



Fig. 5 Complex permittivity as a function of frequency domain for sample natural ester at the temperature 50°C.

## VI. Conclusion

Reason why our diagrams of Remikaflex samples are unsymmetrical in a beginning is because new solid insulation materials needs some time for curing. After this time the material starts degradation. It invalid for fluids insulation systems.

We can conversion polarization current into dissipation factor tanδ (complex permittivity ε″) by the Hamon's approximation.

The parameters what we've got are able to monitoring degradation processes in insulation systems due to temperature stress and results are suitable for next investigation.

*This work was supported by scientific Grand Agency of the ministry of Education of Slovak Republic and the Slovak Academy of Sciences under the projects VEGA No. 1/3142/06 and APVV-20-006005.*

## References

[1] Hanif, M : Principles and applications of insulation testing with DC, IEP-SAC Jourmal 2004-2005, Saudi Arabia.
[2] W.S. Zaengl; Dielectric Spectroscopy in Time and Frequency Domain for HV Power Equipment (Transformers, Cables etc.); 12th ISH, Bangalore India, Aug. 2001
[3] M. Beigert, H.-G. Kranz, "Destruction Free Ageing Diagnosis of Power Cable Insulation Using the Isothermal Relaxation Current Analysis," IEEE ISEI, Pittsburgh, USA, pp. 94
[4] G. Hoff, H.-G. Kranz, "Isothermal Relaxation Current Analysis:A New Non-destructive Diagnostic Tool for Polymeric Power Distribution Cables" April 1999 IEEE / PES Panel on Diagnostic Measurement Techniquesfor Power Cables, New Orleans, USA
[5] Palai-Dany T.: Možnosti využití elektrometru Keithley 617 v dielektrické relaxační spektroskopii. In: Diagnostika '07 : Mezinárodní konference, Nečtiny 11.-13. září 2007. Plzeň : Západočeská univerzita v Plzni, 2007. p. 66-69. ISBN 978-80-7043-557-1.

# Energy savings from high efficiency distribution transformers

*František Lizák*

Department of electric power engineering, FEI TU of Košice, Slovak Republic.

frantisek.lizak@tuke.sk

*Abstract*—This paper describes the potential for high effciency distribution transformers, as a technology to improve network losses. There are several good reasons, why to do a research in this area.

***Keywords***—distribution transformer, distribution network, losses, international standard efficiencies, voltage, gas emission

## I. Introduction

Transformers convert electrical energy from one voltage level to another. They are an essential part of the electricity network. After generation in power stations, electrical energy needs to be transported to the areas where it is consumed. This transport is more effcient at higher voltage, which is why power generated at 10 { 30 kV is converted by transformers into typical voltages of 220 kV up to 400 kV, or even higher [1].

Since the majority of electrical installations operate at lower voltages, the high voltage needs to be converted back close to the point of use.

In this way, electrical energy passes through an average of four transformation stages before being consumed. A large number of transformers of different classes and sizes are needed in the transmission and distribution network, with a wide range of operating voltages. Large transformers for high voltages are called system transformers [1]. The last transformation step into the consumer mains voltage (in Europe 400/230 V) is done by the distribution transformer.

## II. Why focusing on distribution transformers?

Energy losses throughout the world's electrical distribution networks amount to 1 279 TWh. They vary from country to country between 3.7% and 26.7% of the electricity use, which implies that there is a large potential for improvement [2].

After lines, distribution transformers are the second largest loss-making component in electricity networks. Transformers are relatively easy to replace, certainly in comparison with lines or cables, and their effciency can fairly easily be classifed, labeled and standardized. Moreover, modern technology exists to reduce losses by up to 80%.

The worldwide electricity savings potential of switching to high effciency transformers is estimated to be at least 200 TWh, equivalent to the Benelux electricity consumption. This savings potential is not only technically advantageous, but also brings economic and environmental (benefits reduction in greenhouse gas emissions).

| Country | 1980 | 1990 | 1999 | 2000 |
|---|---|---|---|---|
| Finland | 6.2 | 4.8 | 3.6 | 3.7 |
| Netherlands | 4.7 | 4.2 | 4.2 | 4.2 |
| Belgium | 6.5 | 6.0 | 5.5 | 4.8 |
| Germany | 5.3 | 5.2 | 5.0 | 5.1 |
| Italy | 10.4 | 7.5 | 7.1 | 7.0 |
| Denmark | 9.3 | 8.8 | 5.9 | 7.1 |
| United States | 10.5 | 10.5 | 7.1 | 7.1 |
| Switzerland | 9.1 | 7.0 | 7.5 | 7.4 |
| France | 6.9 | 9.0 | 8.0 | 7.8 |
| Austria | 7.9 | 6.9 | 7.9 | 7.8 |
| Sweden | 9.8 | 7.6 | 8.4 | 9.1 |
| Australia | 11.6 | 8.4 | 9.2 | 9.1 |
| United Kingdom | 9.2 | 8.9 | 9.2 | 9.4 |
| Portugal | 13.3 | 9.8 | 10.0 | 9.4 |
| Norway | 9.5 | 7.1 | 8.2 | 9.8 |
| Ireland | 12.8 | 10.9 | 9.6 | 9.9 |
| Canada | 10.6 | 8.2 | 9.2 | 9.9 |
| Spain | 11.1 | 11.1 | 11.2 | 10.6 |
| New Zealand | 14.4 | 13.3 | 13.1 | 11.5 |
| Average | 9.5 | 9.1 | 7.5 | 7.5 |
| European Union | 7.9 | 7.3 | 7.3 | 7.3 |

Fig. 1. Transmission and distribution losses in selected countries (% of electricity use).

## III. Basic principles of distribution transformers

A distribution transformer consists of an iron core with a limb for each of the phases (figure 2).

Around each limb, there are two windings: one with a large number of turns connected to the higher voltage side, and another with a lower number connected to the low voltage.

The windings are separated by insulating material. A change in voltage in one winding induces a change in the other. The result is that an alternating voltage applied to one windings produces a voltage with the same frequency at the terminals of the other one, with the voltage ratio equal to the ratio of the number of turns (Faraday's law) [4].

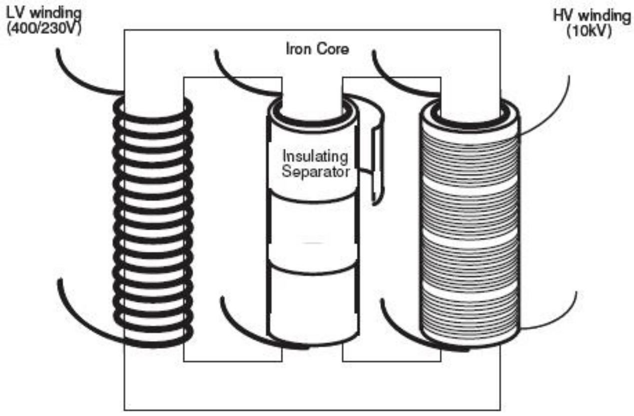



Fig. 2. Schematic diagram of the inside of a three-phase distribution transformer

*A. Oil cooled versus air cooled transformers*

One of the main subdivisions in distribution transformers is the way they are cooled. Most transformers are placed in an oil filled filled tank. The oil cools the coils and at the same time functions as electrical insulation.

In the past, polychlorinated biphenyl (PCB) was regarded as one of the most convenient insulation liquids for transformers, because of its high fire resistance and its excellent electrical qualities. PCBs are however very dificult to decompose, they can accumulate in the food chain and may be a danger for public health. In addition, when burning PCBs, emissions might contain dioxins. Therefore, most countries imposed a program to take all PCB-filled transformers out of use. Today, nearly all of the PCB insulation oil has been replaced by mineral or silicon oil, if no dry transformers are used.

Oil-cooled transformers have the highest efficiency, but are not allowed in environments with a high fire risk. In those places, air-cooled (or 'dry') transformers are used. Air cooling can be combined with an epoxy resin or impregnated paper for electrical insulation. If a dry transformer is installed in a building, the heat must be dissipated. Natural convection may have to be supplemented by forced cooling, e.g. a fan [4].

## IV. LOSSES IN TRANSFORMERS

There are three different types of losses [1]:

1. No-load loss (also called iron loss or core loss): Caused by the hysteresis and eddy currents in the core. It is present whenever the transformer is connected, and independent of the load. It represents a constant, and therefore significant, energy drain.
2. Load loss (or copper loss or short circuit loss): Caused by the resistive losses in the windings and leads, and by eddy currents in the structural steelwork and the windings. It varies with the square of the load current.
3. Cooling loss (only in transformers with fan cooling): Caused by the energy consumption of a fan. The bigger the other losses, the more cooling is needed and the higher the cooling loss.

An estimation of the total energy loss can be calculated from:

$$E_{loss}[kW] = (P_0 + P_k * I^2) * 8760$$

In which:
$P_0$ is the no-load loss [kW].
$P_k$ is the load loss [kW].
I is the rms-average load of the transformer.
8 760 is the number of hours in a year.

*A. Improving efficiency*

To reduce losses in transformers, two elements can be adapted: core and windings. Transformer design is complex, with many of the characteristics of distribution transformers specified in national or international standards.

*1) No-load losses*

The no-load losses can be reduced by selecting a high performance steel for the core (Figure 3.). Over the years, better steels for transformer cores have been developed.

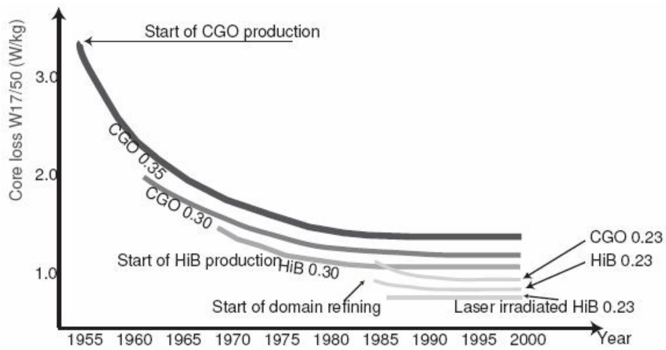



Fig. 3. Different types of magnetic steel

1. Around 1900, hot-rolled steel became the basic material for the core, made up of individual sheets separated by insulating layers to reduce no-load losses. Cold-rolled steel and more sophisticated insulation techniques were progressively developed for improving the performance.
2. Cold-rolled grain oriented silicon steels ('CGO') became available in the 1950's and were the first big leap forward in the reduction of no-load losses.
3. Various processing and coating techniques and a reduced silicon content led to the creation of high permeability grain oriented steels ('HiB'). They remain the current standard material for manufacturing distribution transformers in Europe.
4. During the 80's, techniques were introduced to refine the domains of the iron crystals by laser etching.
5. More recently, the development of amorphous iron introduced a significant new evolution for reducing iron losses.

Next to the choice of the steel, the way in which distribution transformer cores are designed, cut, fabricated and assembled, plays an important role in energy efficiency. Increasing the size of the core reduces the density of the magnetic field, and in this way improves energy efficiency.

Amorphous iron deserves a special mention. Distribution transformers built with amorphous iron cores can have more than 70% lower no-load losses compared to the best conventional designs, and achieving up to 99.7% efficiency for 1000 kVA units. Amorphous iron became commercially available in the early 1980's. These transformers have cores wound with amorphous ribbon made from a ferric metal alloy produced by very rapid quenching to avoid crystallisation. This technology has been used in several hundred thousand distribution transformers in the US, Japan, India and China.

Amorphous technology has been demonstrated for transformer sizes up to 10 MVA, and its application range is expanding.

*2) Load losses*

Load losses are proportional to the square of the load current, so one should always consider how the unit will be loaded over time. Load losses can be reduced by increasing the cross section of the windings. This reduces the current density and consequently the loss, although at a higher construction cost.

The materials for windings have not experienced the same significant improvements inrecent years as the core steels. However, the continuous cold rolling process that is now being introduced for strip production, can lead to more consistent quality.

The process of winding the conductor coils and then fitting them into the assembled core has a very large influence on the energy efficiency of a transformer. It is a labour-intensive process that requires skilled workers. Mechanised winding, under operator control, is increasingly used, especially for smaller sizes.

Another interesting technology in terms of efficiency is the transformer with superconducting windings, cooled with nitrogen. A number of such distribution transformers have been built. They remain much more expensive than conventional types however, and seem to be only promising for specialized applications.

## V. Transformer efficiency standards

In addition to the main division in transformers is between oil-immersed and air-cooled (or dry) types, further subdivisions can be made according to location { pole or ground mounted; single or three phase or ownership { utility or privately owned transformers. Among these, different energy fficiency standards apply. They can be expressed in terms of electrical efficiency, at a certain load level, or in terms of maximum values for no load and load loss. Most standards are voluntary. Table 1 presents overview of international standards [3].

Standards are not limited to efficiency, or loss levels, but may also include total cost of ownership or cost capitalisation formulae. Separate documents define testing procedures and conditions. Reference standards on testing are NEMA TP-2 and IEC 60076, acting as the basis for national equivalents.

| Country / Region | Standard | Subject |
|---|---|---|
| USA | Guide for Determining Energy Efficiency for Distribution Transformers (TP1- 1996). National Electrical Manufacturers Association. 1996. | Efficiency standards and TOC formula |
| | Standard Test Method for Measuring the Energy Consumption of Distribution Transformers (TP2-1998). National Electrical Manufacturers Association. 1998. | Efficiency testing methodology |
| International | Power transformers - Application guide, 60076-8, IEC:1997 | Design, calculation aspects including measurement of losses |
| Europe | Cenelec 1992, Harmonisation documents HD 428, HD538 oil and dry type transformers | Efficiency standards and cost capitalisation formula |
| Variety of country standards defining efficiency levels; MEPS in Australia, Canada, China, Japan, Mexico, proposed in India and New Zealand, non mandatory in Europe | | |

Tab. 1: Main transformer efficiency standards

In figure 4 international standard efficiencies at 50% of load are compared. Some important highlights are:

- In 1997, Oak Ridge National Laboratory [5, 6] performed extensive studies to determine whether energy conservation standards for distribution transformers would offer significant energy savings, be technically achievable and economically justified. This has led to the definition of the NEMA TP-1 standard, which became the basis for the rule making process on minimum standards. NEMA TP-1 has been used as a guideline by Canada, Australia, New Zealand and (partially) Mexico.
- In Europe, CENELEC Technical Committee 14 has published standards HD 428 and HD 538 classifying losses for oil (428) and dry type (538) transformers (Tables 2 & 3). Country standards should be in line with CENELEC documents but, since the standard allows many possibilities, national approaches in Europe widely differ. Efficiency standards are high in Benelux, Germany, Austria and Switzerland, but low in France, Italy and Spain. Utilities in certain European countries have policies exceeding national standards, e.g. Endesa in Spain purchases HD 428 CC'- level distribution transformers, while the country standard is equivalent to HD 428 AA'. Swiss utilities have been commissioning transformers with efficiencies in excess of the HD428 classification scheme for already many years [5].

| Rated power | Oil-filled (HD428) up to 24 kV | | | Dry type (HD538) |
|---|---|---|---|---|
| | List A' | List B' | List C' | 12 kV primary |
| kVA | W | W | W | W |
| 50 | 190 | 145 | 125 | n/a |
| 100 | 320 | 260 | 210 | 440 |
| 160 | 460 | 375 | 300 | 610 |
| 250 | 650 | 530 | 425 | 820 |
| 400 | 930 | 750 | 610 | 1 150 |
| 630 (4%) | 1 300 | 1 030 | 860 | 1 500 |
| 630 (6%) | 1 200 | 940 | 800 | 1 370 |
| 1000 | 1 700 | 1 400 | 1 100 | 2 000 |
| 1600 | 2 600 | 2 200 | 1 700 | 2 800 |
| 2500 | 3 800 | 3 200 | 2 500 | 4 300 |

Tab. 2: No-Load losses for distribution transformers according to CENELEC HD428 and HD538.

| Rated power | Oil-filled (HD428) up to 24 kV | | | Dry type (HD538) |
|---|---|---|---|---|
| | List A | List B | List C | 12 kV primary |
| kVA | W | W | W | W |
| 50 | 1 100 | 1 350 | 875 | n/a |
| 100 | 1 750 | 2 150 | 1 475 | 2 000 |
| 160 | 2 350 | 3 100 | 2 000 | 2 700 |
| 250 | 3 250 | 4 200 | 2 750 | 3 500 |
| 400 | 4 600 | 6 000 | 3 850 | 4 900 |
| 630 (4%) | 6 500 | 8 400 | 5 400 | 7 300 |
| 630 (6%) | 6 750 | 8 700 | 5 600 | 7 600 |
| 1000 | 10 500 | 13 000 | 9 500 | 10 000 |
| 1600 | 17 000 | 20 000 | 14 000 | 14 000 |
| 2500 | 26 500 | 32 000 | 22 000 | 21 000 |

Tab. 3: Load losses for distribution transformers according to CENELEC HD428 and HD538

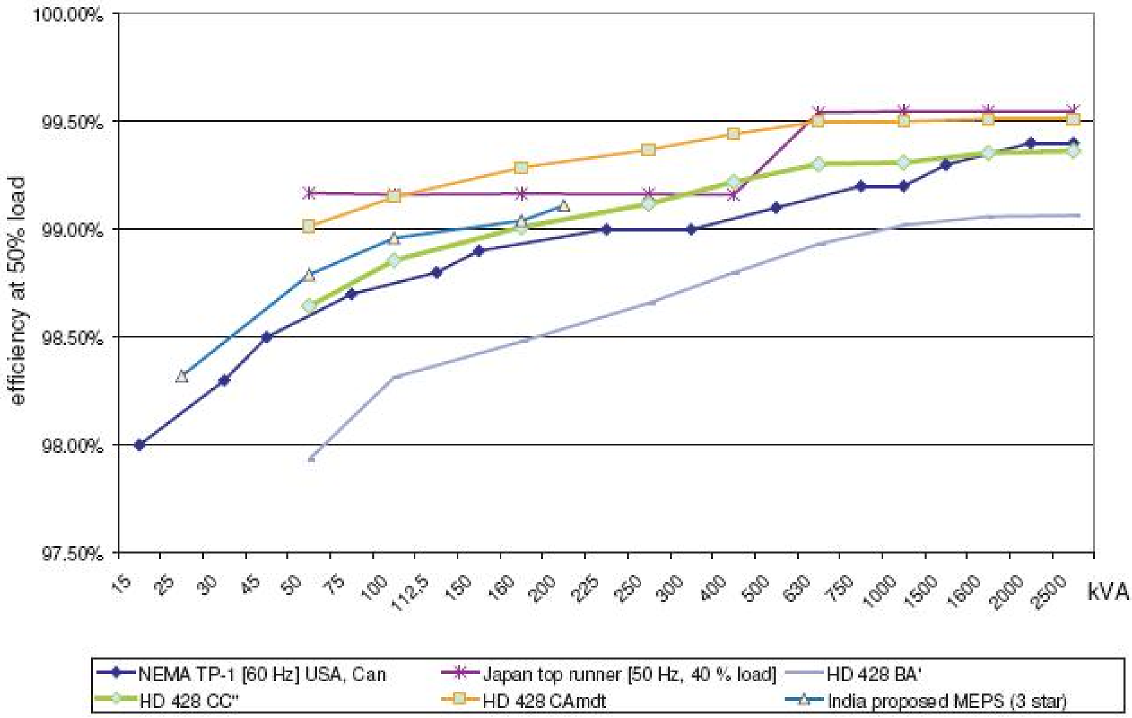



Fig..4: Comparison of international standard efficiencies at 50% of load. 'C-AMDT' refers to an amorphous-core transformer with HD 428 C-class of load losses

In China, the current standard is S9, and a new standard (S11) is being introduced, which has losses slightly below Europe's AC' level. The standard defines allowable levels for no-load and load losses [7].

The Indian Bureau of Energy Efficiency (BEE), acting under a mandate from the Indian Ministry of Power, has analyzed the feasibility of a distribution transformer minimum efficiency standard. BEE classifies distribution transformers up to 200 kVA into 5 categories from 1 Star (high loss) to 5 Stars (low loss). 5 Stars represents world-class performance. 3 Stars is being proposed as a minimum efficiency standard, and is being widely followed by utilities [6].

Japan has a different type of distribution system, with the last step of voltage transformation much closer to the consumer. The majority of units are pole mounted single phase transformers. The driver for setting up minimum efficiency standards was the Kyoto commitment. Transformers, together with 17 categories of electrical equipment, should meet minimum effciencies. In the case of transformers, the effciency is defined at 40% load. Target average efficiency has been defined for the year 2006 (oil) or 2007 (dry type), based on the best products on the market in 2003. This Japanese standard is currently the most demanding compared to other regulated standards [7].

## VI. Conclusion

Replacing all distribution transformers by energy efficient types could save 200 TWh a year, equivalent to 130 million tonnes of $CO_2$ emissions. An advantage large enough to justify the effort.

High-efficiency transformers are a mature technology with their economic and environmental benefits clearly demonstrated. While their higher initial cost can be more than recovered by reduced running costs, many distribution transformers are still chosen on the basis of the purchasing price. The paper suggests that a new regulatory framework is required to stimulate change to allow the capture of the available benefits.

This work was supported by scientific Grant Agency of the ministry of Education of Slovak Republic and the Slovak Academy of Sciences under the project VEGA No. 1/4072/07.

## References

[1] Electricity distribution losses - a consultation document. Technical report, Office of Gas and Electricity Markets, 2003.

[2] *World Energy Investment Outlook*. International Energy Agency, 2003.

[3] E Shibata. Amorphous transformers in japan. *Production and Electricity*, September 2003.

[4] B Kennedy. *Energy efficient transformers*. Mc Graw - Hill, 1998.

[5] M McDermott. Energy efficient distribution transformers { utility initiatives. Technical report, European Copper Institute, 2000. Available from http://transformers.copperwire.org.

[6] P Barnes et al. Determination analysis of energy conservation standards for distribution transformers. Technical Report ORNL/6847, Oak Ridge National Laboratory, 1996.

[7] Y Zhao. The current situation of distribution transformers in china. Report prepared for the Promotion Partnership for High Efficiency Transformers, 2003.

# Remote Measurement in Automobile

[1]*Ján Molnár*

[1]Dept. of Theoretical Electrotechnics and Electrical Measurement, FEI TU of Košice, Slovak Republic

[1]jan.molnar@tuke.sk

**Abstract - The article deals with the proposal of the measuring chain assigned for remote measuring in the automobile industry. The designed solution should be able to do automatic measurement of all required parameters and to send obtained data to the remote centre, where they could be analyzed by the telemetric based expert system.**

***Keywords* - GSM (Global System of Mobile communication), WiFi connection (Wireless Fidelity), remote, measuring, Turbo Lite, SIM Toolkit.**

## I. PROBLEM DESCRIPTION

Due to big progress of automotive industry the remote measuring system could help to solve many problems. Various automobile failures could be eliminated by prevention, early malfunction detection or failure detection. The detection and prevention would be based on selected data measuring, data collecting and transporting to the remote center for its further analyzing and evaluation.

Requirements for such a measure chain depend on the data type, which have to be measured. For instance, it is a big difference between a battery voltage and engine revolution measuring. In first case, it is sufficient to collect data few times per hour. In the other case, there is necessary to take a data few times per second. The sampling time is then some tens of miliseconds. In the case of automobile measuring, there exist various data of various types to be measured.

Another problem is created by the information transport to the remote center, where the data are analyzed, evaluated and also the decision about the failure state is made. Today's situation is based on preventive service inspections at regular intervals, where the car is connected to the PC and all diagnostic methods are done. It would be more suitable, if it would be possible to do at the moment of optimal value deviation or in any failure detection. All required data would be transferred to the center and evaluated by the expert system.

Expert system should decide, if the failure state exists or not. In the case of failure state appearance, the driver should receive the information about failure from center. Measurement should be running during the automobile performance. Such a way discovering of incidental and a periodic failures will be easier, because such failures are the most difficult to identify. By designed method should be possible to transport measured data at the moment, when the failure appears.

## II. SOLUTION DESING

Two types of net seem to be the most appropriate for data transport. The first is GSM (Global System of Mobile communication) and the second one is WiFi connection (Wireless Fidelity). Both nets have own advantages and disadvantages. The most important advantage of WiFi (compared with GSM) is bandwidth. It means that the bigger data amount in shorter time period can be transferred by WiFi using. The most important advantage of GSM (compared with WiFi) is that the great area is covered by GSM signal. Except that, it is possible to use various tools of GSM communication such a SMS etc. For remote measurement problem solving it is possible to use both types of the net. The nets could complement each other, depending on situation.

*GSM (Global System of Mobile communication):*

The Global System for Mobile Communications (GSM) is the most popular standard for mobile phones in the world. Most GSM networks operate in the 900 MHz or 1800 MHz bands. The channel data rate is 270.833 kbit/s, and the frame duration is 4.615 ms. The longest distance between mobile modem and the component of GSM network is 35 km or 22 miles in practical use. There are also several implementations of the concept of an extended cell, where the cell radius could be double or even more, depending on the antenna system, the type of terrain and the timing advance. Structure of GSM network is shown in Fig.1.

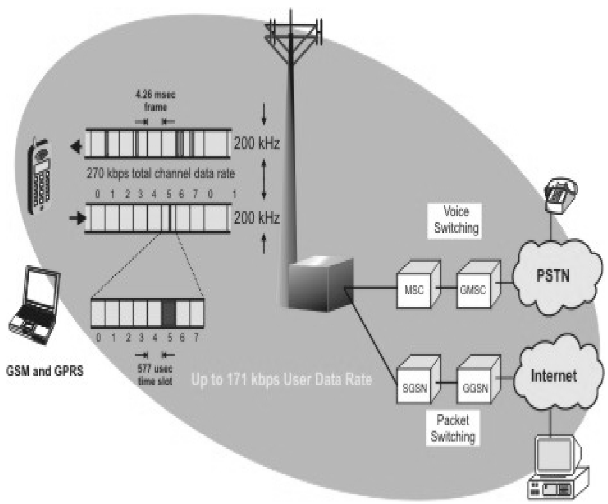



Fig. 1. Structure of GSM network

*WiFi connection (Wireless Fidelity):*

Wi-Fi networks use radio technologies called IEEE802.11b, 802.11a or 802.11g to provide wireless connectivity in local area networks. A Wi-Fi network can be used to connect computers to each other, to the Internet, and to wired networks using IEEE 802.3 or Ethernet. Wi-Fi networks operate in the unlicensed 2.4 and 5 GHz radio bands, with an 11 Mbps (802.11b) or 54 Mbps (802.11a or 802.11g) data rate or with products that contain both bands (dual band), so they can provide real-world performance similar to the basic 10BaseT wired Ethernet networks used in many offices. Unfortunately, 802.11a is not compatible with 802.11b/g. Wi-Fi is a standard developed by the Wi-Fi alliance that certifies vendor products to ensure 802.11 products on the market follow the various 802.11 specifications. Example of such WiFi network shows Fig 2.

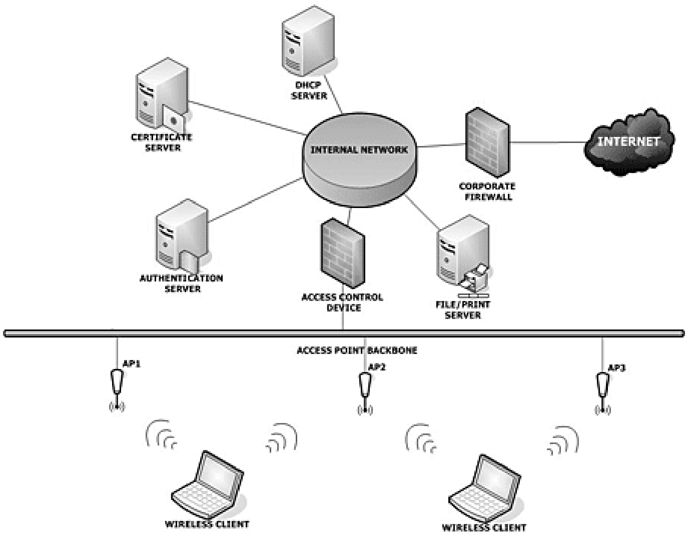



Fig.2. WiFi network.

In praxis, if every automobile would be connected to the net and every one car would continuously sending a data, it would be inefficient and the net would be very soon overloaded. The net should be used only for diagnostic and optimization case.

During diagnostic, there would be monitoring of all automobile processes and values. Measured date would be compared with optimal values. Evaluation could be done by some local expert system or using artificial neural network etc. In case of deviation from standard state, automobile should connect the expert system and measured data would be transferred to the remote expert system. Remote expert system should be more sophisticated and frequently updated, so the evaluation by remote center should be more exact. For such communication a GSM network would be sufficient. Similar local expert system is nowadays implemented in advanced automobiles of higher class. But connection with central expert system will enable much complex and reliable diagnostic. It is much effective to update the central expert system than individual local systems in each automobile.

If the car will be inside the area covered by WiFi signal and transfer capacity is sufficient, then the optimizing of engine setup, the automobile software update and various modifications would be enabled which were done only in car service yet. The basic condition of such modification is given by sufficiently fast connection existence, which would be able to transfer the fast changed parameters (such engine revolution, gas consumption etc).

## III. ARCHITECTURE PROPOSAL

Architecture proposal of such connection is displayed in figure Fig.3. Architecture doesn't deal directly with measuring. Measuring problems are mostly solved nowadays. Architecture deals with problem of data transport. The "bus" of the measure chain is responsible for data transport to the evaluation system.

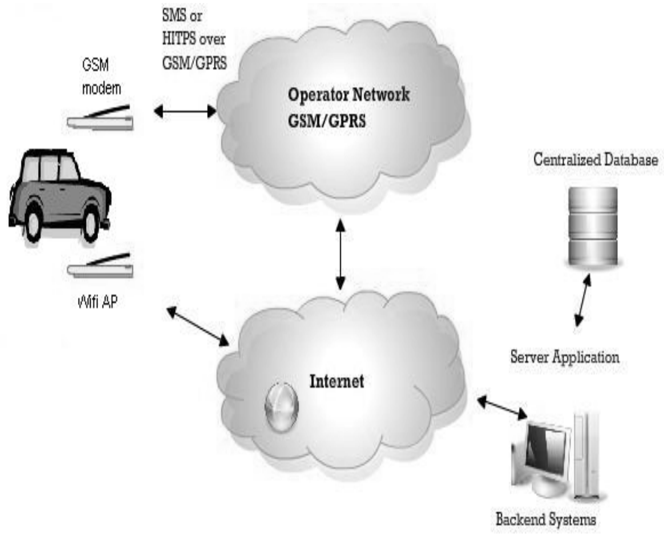



Fig.3. The architecture proposal for remote car measurement

As the figure shows, the automobile should have integrated GSM modem and also device for WiFi net connection. Using the GSM or WiFi net, the automobile would have access to the Internet. Via Internet the car would be connected with remote expert system.

Local expert system would not just watch the automobile parameters, but it could also control the communication. It would decides, which data and how often would be send to the remote system in the case of failure state. It would also decide which data would be sending via GSM and which via WiFi. It would be also responsible for optimizing data transfer.

Expert system would offer suitable solution for driver. In case of failure, it would recommends the driver to stop the car immediately, drive the car to the next service or ignore the problem (problem is temporary). It would be possible also give attention to the driver, in the case when he often uses the high engine revolution etc. Connection to the Internet would have much more advantages like as an on-line help (how to change electrical fuse – and which one, change a wheel, weather forecast). But the most important is possibility of emergency call. Using GSM net it would be possible to find out the car location etc.

Such a way created system, the part of which is remote measurement chain utilizing the Internet and connection to the expert system, would be great benefit for automobile reliability and safety.

## IV. TURBO LITE 2

Turbo Lite 2 is an open and price sensitive solution for remote control, measurement and security GSM applications. It works with any SIM Toolkit enabled mobile phone (all mobile phones produced since 1999) and it is ideal to use with

abandoned or recycled mobile phones for SMS home automation, alarm systems and remote data gathering.
Turbo Lite 2 follows on from original Turbo Lite with new 2+2 opto-isolated inputs/outputs and Turbo Adapter compatible ports including SPI and I2C interfaces.
It can be connected to Turbo Programmer and used for application debugging and SIM – Mobile Equipment communication tracing. Real device is shown in Fig 4.

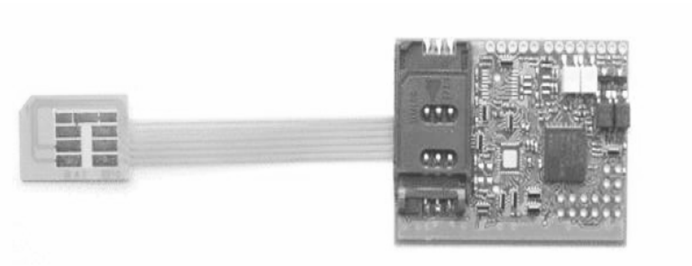


Fig.4. Turbo Lite 2

*Features of Turbo Lite 2:*

- 2+2 opto-isolated I/O
- Turbo Lite compatible 13 digital I/O, 4 of them usable as ADC inputs
- Turbo Adapter compatible ports, SPI, I2C interfaces
- SIM Toolkit based, compatible with any SIM Toolkit enabled GSM mobile phone or module
- Comes with preinstalled open source Pager v2 application for SMS control and measurement
- Easy to load applications with mobile phone data cable
- Allows application debugging and SIM-ME communication tracing
- Upgradable with freely available firmware
- Open, free and well documented application development
- Open source and free development tool chain

*Software:*

Turbo Lite 2 software (both application and firmware) is compatible with family of Turbo SIM Toolkit Adapter products produced by BLADOX. The preinstalled Pager v2 application is a ready-to-use flexible and easy for configuring of SMS remote measurement and control pager. Applications can be loaded and removed via data cable with AT commands for SMS handling. The application is best used with Turbo Lite 2, but it is compatible with all products in the Turbo SIM Toolkit Adapter family – including the original Turbo Lite. Pager v2 follows on from version 1 with the inclusion of most of the users' wishes and resolves several problems experienced with original Pager application deployment.
The firmware (kernel) can be upgraded by bootloader procedure with the help of Turbo Programmer.

Example of such Turbo Lite 2 opto-isolated inputs and outputs is shown in Fig. 5. These inputs and outputs we are able to use for measuring or controlling of various peripherals. For instance the measuring of an analog voltage we can do by utilizing of the inputs labeled IN0+ and IN0-. The limiting condition is the fact that the measuring voltage range must be preproceessed into input voltage range 0 - 2,56V (Vref = 2,56 V is voltage reference of internal A/D converter). The corresponding digital output value is then in the range of 0-1023. The concrete value depends on measured voltage according the following relation:

$$output\ value = 1023.\left(\frac{U}{2,56}\right) \qquad [V]$$

where the *output value* represents integral number. Based on above mentioned, in the case of automobile application we have to correctly configure the voltage divider.

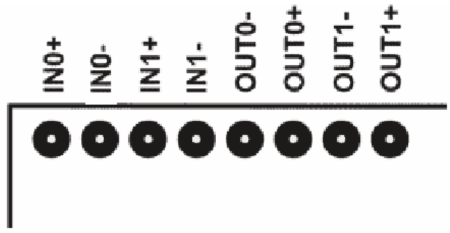



Fig.5. Turbo Lite 2 opto-isolated inputs and outputs

## V. Conclusion

Using of worldwide network – Internet seems to be optimal for remote data measurement inside the car. The GSM and WIFI connection is suitable for Internet connection. The biggest advantage of WiFi is bandwidth. It is able to transfer data faster than GSM. The biggest advantage of GSM net is that the GSM signal covers majority of area. Realization such a remote measure chain will serve for future exploration of remote measure chain using Internet as a bus, also in real-time applications. That way would increase possibilities to integrate various measurement systems in the whole complex system.

## Acknowledgment

The paper has been prepared by the support of Slovak grant projects VEGA No. 1/4174/07, VEGA No. 1/0660/08, KEGA 3/5227/07, KEGA 3/6386/08 and KEGA 3/6388/08.

## References

[1] Savol, T.: "Použitie GPS a GSM v bezpečnostných systémoch".http://fel.utc.sk/~nagy/BS/PDF/GPS_GSM_Savol.pdf
[2] Yang, S.H. - Tan, L.S. - Chen, X.: "Requirements Specification and Architecture Design for Internet-based Control Systems", Proceedings of the 26th Annual International Computer Software and Applications Conference (COMPSAC'02), 2002.
[3] Rizzoli A.E: A Collection of Modelling and Simulation Resources on the Internet. Web page: ttp://www.idsia.ch/~andrea/simtools.html
[4] Web page: http://en.wikipedia.org/wiki/Ethernet
[5] Kurose, J. F. - Ross, K. W.: "Computer Networking, A Top-Down Approach Featuring the Internet". 2001, Addison Wesley, ISBN 0-201-47711-4
[6] Kweon, S. K - Cho, M. - Shin, K. G.: "Soft Real-Time Communication over Ethernet with Adaptive Traffic Smoothing", IEEE Transactions on parallel and distributed systems, VOL. 15, NO. 10, October 2004.
[7] El-Marakby, R. - Hutchison, D.: "Delivery of Real-time Continuous Media over the Internet". Proceedings of the 2nd IEEE Symposium on Computers and Communications (ISCC '97) 1997

# Chaotic sequences using in DS – CDMA system

[1]Henrieta PALUBOVÁ

[1]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]henrieta.palubova@tuke.sk

***Abstract*—The chaotic sequences in DS – CDMA system is presented in this paper. The simulation results in Matlab are presented here. The Rake receiver is describe in this paper too.**

***Keywords*—DS – CDMA, Chaos, Chaotic sequences, Rake receiver.**

## I. Introduction

Chaos is a deterministic, random-like process found in non-linear, dynamical system, which is non-period, non-converging and bounded. Moreover, it has a very sensitive dependence upon its initial condition and parameters. Chaotic signals can be used in communication systems [7].

The paper is organized as follows: section 2 presents the system model, section 3 describes the direct sequence Code Division Multiple Access (CDMA), section 4 describes the simulation results and finally conclusion is given.

## II. System Model

Chaotic signals can be used in communication. A chaotic map is a discrete-time dynamical system

$$x_{k+1} = f(x_k), \quad 0 \langle x_k \langle 1, \quad k = 0,1,2,.... \qquad (1)$$

running in chaotic state. The chaotic sequence

$$\{x_k : k = 0,1,2,....\} \qquad (2)$$

can be used as spread-spectrum sequence in place of Pseudorandom number (PN) sequence in conventional Direct-Sequence Spread-Spectrum (CDMA DS/SS) communication systems. Chaotic sequences are uncorrelated when their initial values are different, so in chaotic spread-spectrum systems, a user corresponds to an initial value.

Improved logistic-map is defined by:

$$x_{k+1} = f(x_k) = 1 - 2(x_k)^2, \quad x_k \in (-1,1) \qquad (3)$$

The chaotic sequence:

$$\{x_k : k = 0,1,2,....\} = \{f^k(x_0) : k = 0,1,2,....\}, \quad (4)$$

generated by improved logistic-map, is neither periodic nor converging, and sensitively dependent on initial value. The chaotic sequence is random-like, so probability and statistics can be used in discussing their characteristics [7].

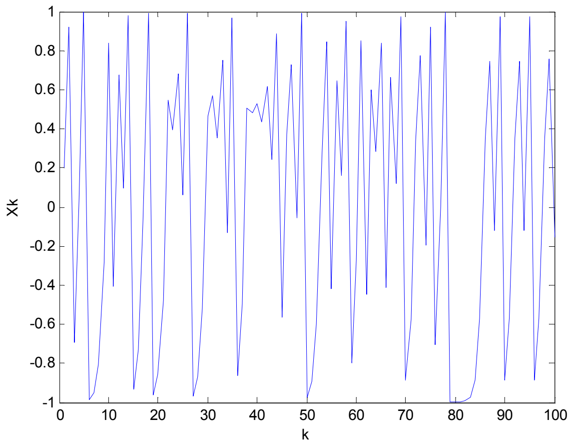



Fig. 1. Chaotic sequence

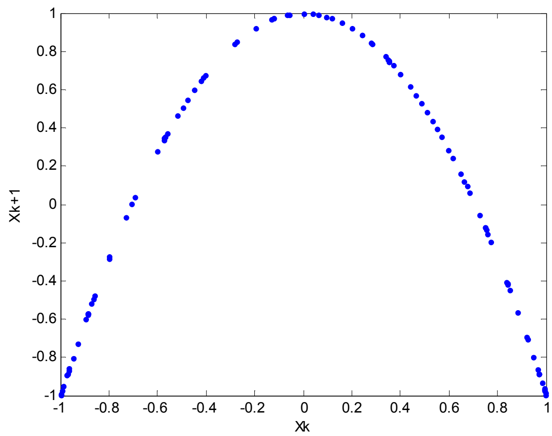



Fig. 2. Chaotic sequences ($x_k$ in first step is set to 0.2)

To generate the chaotic sequence with biased values, is used the chaotic map shown in Fig.2. This map was made from the Bernoulli shift map. This map is described by:

$$x_{k+1} = \begin{cases} \dfrac{(1+r)x_k + q + r}{1-q} & (-1 \le x_k \le -q) \\ \dfrac{(1-r)x_k + q}{q} & (-q \langle x_k \le 0) \\ \dfrac{(1-r)x_k - q}{q} & (0 \langle x_k \le q) \\ \dfrac{(1+r)x_k - q - r}{1-q} & (q \langle x_k \le 1) \end{cases} \tag{5}$$

The slopes of the map can be changed by deciding parameters $q$ ($0.5 \le q \le 1.0$) and $r$ ($0.0 \le r \le 1.0$)

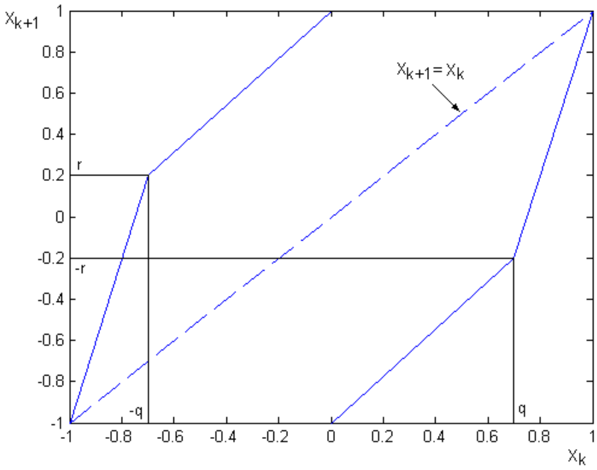
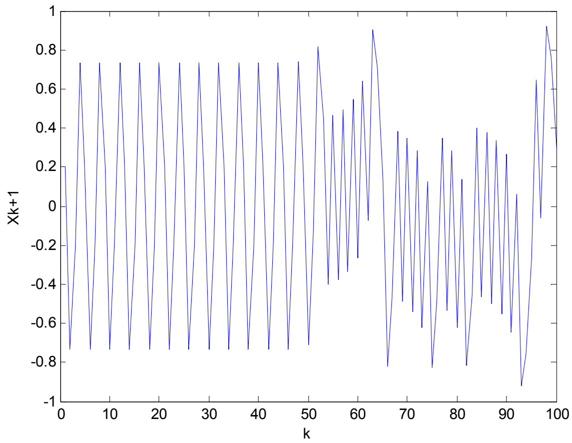



Fig. 3. Chaotic map with different slopes[5].

As an example, chaotic sequence with biased values is shown in Fig.3. From these figure, it can be observed that the chaotic sequence includes many values near 1.0 (or -1.0) as both $q$ and $r$ approach 1.0, namely, it can be said that deviation was made to the values of the chaotic sequence [5].




(q,r) = (0.6, 0.2), Xk=0.2

Fig. 4. Chaotic sequences with biased values

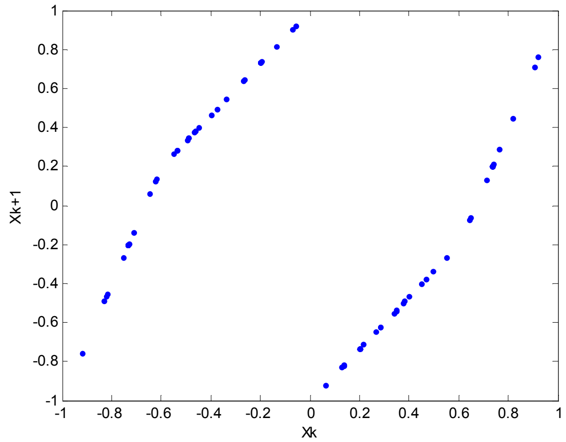



Fig. 5. Chaotic sequences ($x_k$ in first step is set to 0.2)

## III. Direct Sequence CDMA

In Direct Sequence (DS) spread spectrum transmission, the user data signal is multiplied by a code sequence. Mostly, binary sequences are used. The duration of an element in the code is called the "chip time". The ratio between the user symbol time and the chip time is called the spread factor. The transmit signal occupies a bandwidth that equals the spread factor times the bandwidth of the user data.

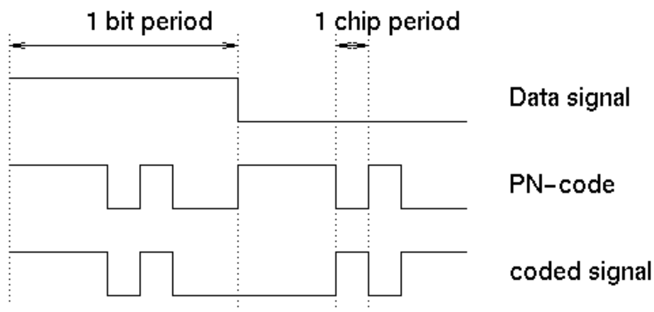



Fig. 6. A DS-CDMA signal is generated by multiplication of a user data signal by a code sequence[1].

In the receiver, the received signal is again multiplied by the same code. This operation removes the code, so we recover the transmitted user data.

$$\sum_{n=1}^{N} c_1^2(nT_C + t_d) = \sum_{n=1}^{N} c_1^2(nT_c) = N \tag{6}$$

where $c_1$ is the code sequence used by user 1, $T_c$ is the chip duration, $t_d$ is a common time offset, shared between transmitter and receiver and $N$ is the length of the code sequence. The receive code must be perfectly time aligned with the transmit code.

The used code sequences are:

- Maximum length or Pseudo Noise (PN) sequences
- Walsh Hadamard Codes
- Gold codes
- Kasami codes

A major difficulty in Direct Sequence transmission is the Near-Far effect. If more than one user is active, the incoming interference power is suppressed by the cross correlation between the code of the reference user and the code of the interferer. In the event that the interferer is closer to the receiver then the reference user, the interference components may not be sufficiently attenuated by the despreading process. In cellular CDMA systems, (adaptive) power control is needed to avoid this problem

In a multipath fading channel, delayed reflections interfere with the direct signal. However, a DS-CDMA, [9] signal suffering from multipath dispersion can be detected by a rake receiver. This receiver optimally combines signals received over multiple paths. Multipath self-interference is attenuated, because one can choose codes such that [1].

$$\sum_{n=1}^{N} c_1(nT_c)c_1(nT_c + t_d) \cong 0 \qquad (7)$$

## IV. Simulation Results

The simulations in Matlab are used in this section. Fig. 6,7,8,9 shows the Bit Error Rate (BER) dependency from Eb/No. Different sequences (Walsh, Golay, Gold, Chaotic sequence, generated by improved logistic-map (4)) is used in Multi Carrier (MC) – CDMA system [8], which consists from DS-CDMA and from Orthogonal Frequency Division Multiple Access (OFDM). The simulation parameters as follows: number of users = 1, number of transmitted bits = 8000, Fast Fourier Transformation (FFT) = 128, length of sequence = 31, IBO = 3. Additive White Gaussian Noise (AWGN) channel and 16 Quadrature Amplitude Modulation (QAM) are used in this simulation. The High Power Amplifier (HPA) is using the Saleh nonlinear model. Sigila *bmf* in figures means that: as a receiver is used conventional receiver-bank of matched filters. *mmse* is minimum mean square error estimation and *msf* means that Microstatistic Detector is used in this system. From these simulation we can see that the BER is lower, when microstatistic detector is used in system. The results of chaotic sequence are comparable with walsh sequence.

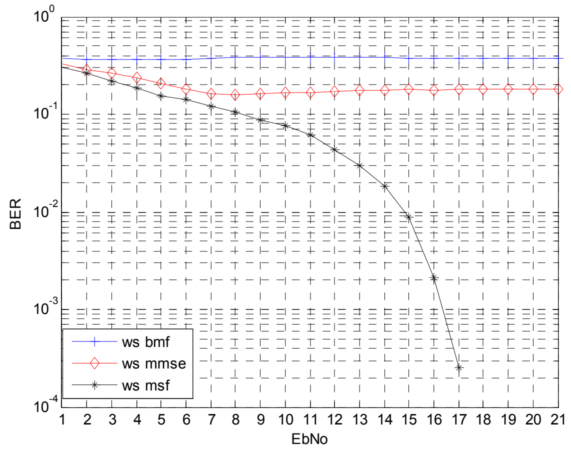



Fig. 7. BER (Bit Error Rate) – Walsh sequences

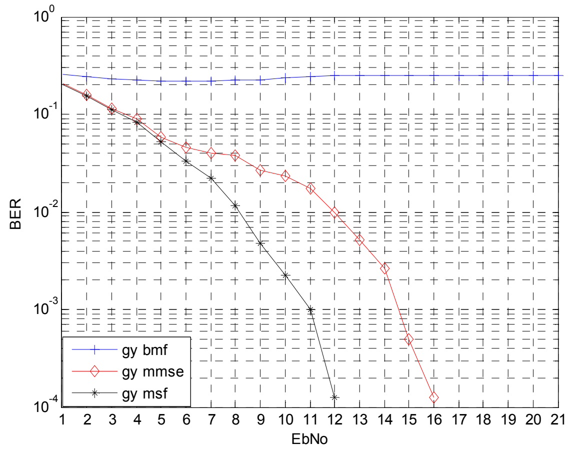



Fig. 8. BER (Bit Error Rate) – Golay sequences

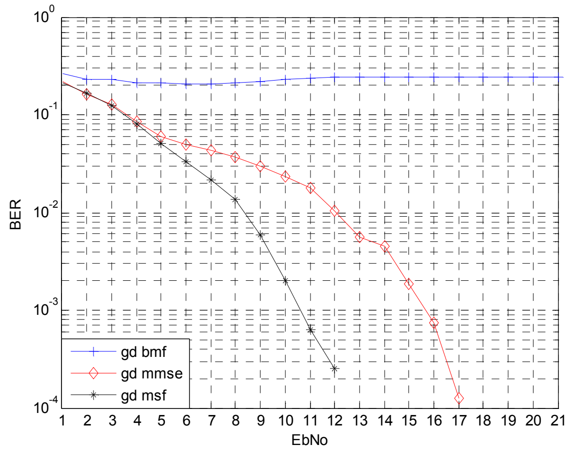



Fig. 9. BER (Bit Error Rate) – Gold sequences

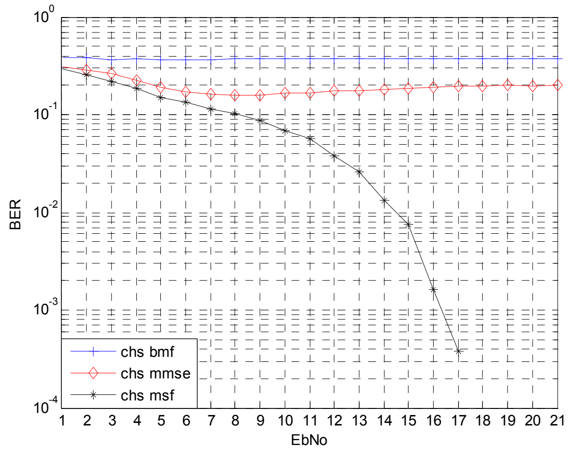



Fig. 10. BER (Bit Error Rate) – Chaotic sequences

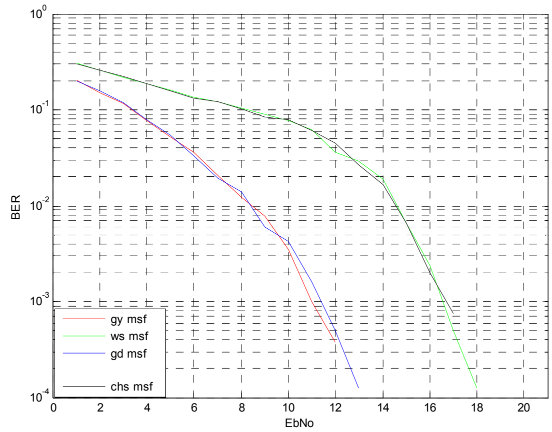



Fig. 11. BER (Bit Error Rate) – ber_msf

Using AWGN channel in real is insufficient. The aim of my next work is using otherwise type of channel, for example Rayleigh channel, and the Rake receiver. In Rayleigh channel we can see the Doppler shift, multipath fading and time delay. The BER in this type of channel can be lower, when the rake receiver is used. Chaotic sequences have good autocorrelation characteristics, when the Rake receiver is used in the system. The rake receiver consists of multiple correlators, in which the receive signal is multiplied by time-shifted versions of a locally generated code sequence. The intention is to separate signals such that each finger only sees signals coming in over a single path. The spreading code is chosen to have a very small autocorrelation value for any nonzero time offset. This avoids crosstalk between fingers. In practice, the situation is less ideal. It is not the full periodic autocorrelation that determines the crosstalk between signals in different fingers, but rather two *partial* correlations, with contributions from two consecutive bits or symbols. It has been attempted to find sequences that have satisfactory partial correlation values, but the crosstalk due to partial (non-periodic) correlations remains substantially more difficult to reduce than the effects of periodic correlations [2].

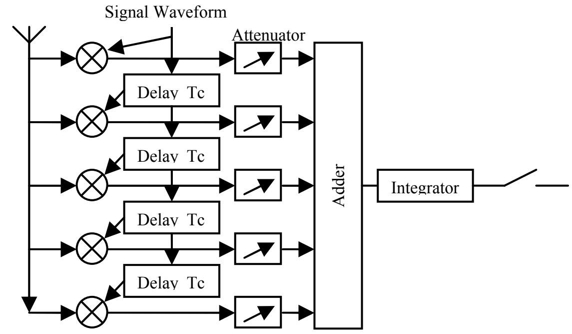



Fig. 12. 5 – finger Rake receiver[3].

## V. Conclusion

The using of chaotic sequence in MC – CDMA system is study in this work. BER is simulated in Matlab. As we can see from Figs. 8, 9 to using *mmse* we can achieve low BER. But using *mmse* for walsh or for chaotic sequences is insufficient. For these sequences are apply *msf*. As we can see from Figs. 7, 10 using of *msf* have can improved BER. Using only *bmf* for all of these sequences is uselessly. As we can see from simulation results *bmf* is the system not available. The parameters of simulation can be better, if we can changed for example modulation, coding, system receiver. I will change the system receiver in my next work.

## Acknowledgment

This work was supported by the project VEGA 1/4088/07 and COST 297.

## References

[1] http://wireless.per.nl/reference/chaptr05/cdma/dscdma.htm
[2] http://wireless.per.nl/reference/chaptr05/cdma/rake.htm
[3] T. Heikkilä, *Rake Receiver*. Postgraduate Course in Radio Communications, 2004
[4] I. Vasilache, D. Grecu, C. Grozea, B. Cristea, "Random Sequence Generator Based on Nonlinear Function" Bucharest, Romania.
[5] S. Arai, Y. Nishio, "Research on Differential Chaos Shift Keying Changing Deviation of Chaotic Sequence", 2007 RISP International Workshop on Nonlinear Circuits and Signal Processing, Shanghai Jiao Tong University, Shanghai, China, Mar. 2007.
[6] G. Mazzini, V. Tralli, "Performance Evaluation Methods for DS – CDMA Systems with Rake Receivers," *IEEE* 1995.
[7] W. Hai, H. Jiandong, "Chaotic Spread – Spectrum Communication Using Discrete – Time Synchronization", The Journal of China Universities of Posts and Telecommunications, Vol. 4, No. 1, Jun 1997.
[8] L. Hanzo, M. Munster,"OFDM and MC-CDMA for broadband Multi-User Communications", IEEE Press, UK, 2003
[9] J. Čížová, "Mnohoúčastnícka detekcia v DS-CDMA penosových systémoch", TU-FEI Písomná práca k dizertačnej skúške 2003

# Node Localization Methods in UWB Wireless Sensor Networks: A Review

*Mária ŠVECOVÁ*

Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

maria.svecova@tuke.sk

***Abstract*—Localization of the node in UWB wireless sensor networks (WSN) is very interesting topic, because the knowledge of the node location is the presumption for the large number of applications (e.g. rescue, security, medical, military, logistic, etc.). The first solution task of the node localization problem is to estimate some signal parameters, for example, angle of arrival, received signal strength, time of arrival or time difference of arrival. The coordinates of the node are then computed using these parameters by means some localization methods. In this paper will be shown, that the time-based localization techniques are the most suitable for UWB WSN. Therefore, a review of node localization methods based on time measurements is presented.**

***Keywords*—AOA, localization, RSS, TDOA, TOA, UWB, WSN.**

## I. Introduction

Localization of a node in WSN is an important area that attracted considerable interest of research. It was subjected to recent advances in WSN theory and applications [1]. Wireless communications [2] and microelectronics technologies have enabled the development of low-cost, low-power and multi-functional sensors that are small in size and communicate in short distances. Cheap, smart sensors, networked through wireless links and deployed in large numbers, provide unprecedented opportunities for monitoring and controlling homes, cities and the environment. In addition, networked sensors have a broad spectrum of applications including security (localization authorised persons in high-security areas or unauthorised persons), medical applications (monitoring of patients), logistics (package tracking), family communications (supervision of children), search and rescue works (detecting people buried by avalanche or earthquake), military applications, etc.

Localization methods estimate the locations of nodes with initially unknown location information by using knowledge of the positions of a few nodes and inter-nodes measurements such as direction and distance-based measurements. Nodes with known location are called anchors and their locations can be obtained by using e.g. a global positioning system (GPS), or by installing anchors at points with known coordinates. In applications requiring a global coordinate system, these nodes will determine the location of the node within network in the global coordinate system. In applications where a local coordinate system suffices (e.g. smart homes), these anchors define the local coordinate system to which all other sensors are related. Because of constraints on the cost and size of sensors, energy consumption, implementation environment (e.g. GPS is not accessible in some environments) and the deployment of sensors (e.g. sensor nodes may be randomly scattered in the region), most sensors do not know their locations. These sensors with unknown location information are called ordinary nodes and their coordinates will be estimated by the localization method [3].

Ultra wideband (UWB) technology [4] has proven to be useful in short range, high data rate, robust and low power communications. These features make UWB systems ideal candidates for reliable data communications between sensors of WSN. Since, UWB signals contains a wide range of frequency components, they have a higher probability of passing through or around obstacles. Moreover, time-frequency duality implies that a UWB signal has a very high time resolution, which facilitates accurate time measurement, hence distance estimation. UWB signals are therefore very suitable for location applications. In this paper we will focus on node localization in UWB WSN in 2D.

The problem of the node localization in UWB WSN is the estimation of the unknown node coordinates by given set of anchors. The first task of the estimation is measure various signal parameters, such as angle of arrival (AOA), received signal strength (RSS), time of arrival (TOA) or time difference of arrival (TDOA) between the node and the anchors. Then, the position of the node is estimated by using these signal parameters [5].

This paper will have the following structure. In the Section II, we will analyse the measurement techniques based on AOA, RSS, TOA and TDOA measurements from UWB perspective. Based on this analyse it will be concluded that time-based techniques are the most suitable for the node localization in UWB WSN. Therefore, in the Section III we will provide a review of the basic localization methods based on time measurements. Conclusion encloses the paper.

## II. Measurement techniques

Measurement techniques in UWB WSN can be broadly classified into four categories: AOA measurements, RSS measurements, TOA measurements and TDOA measurements.

### A. Angle of arrival measurements

AOA of the signal between the node and the anchor is measured by steering the main lobe of a directional antenna or an adaptive antenna array [6], as is outlined in Fig. 1. Each AOA measurement forms a bearing line between the anchor and the node to be positioned. Ideally, in the error-free case, all bearing lines will intersect at single point that determines a location of the node. Otherwise, more than two bearing lines need not intersect at a single point (Fig. 1). However, the

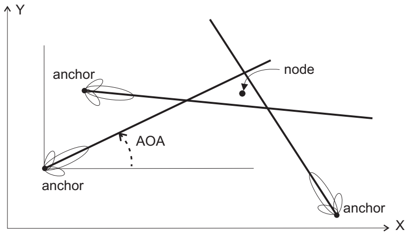



Fig. 1. AOA-based localization.

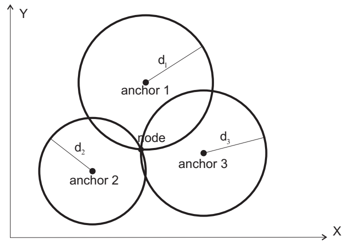



Fig. 2. RSS or TOA-based localization.

appropriate methods can solve this problem and they provide the estimate of the node location [1], [3], [7], [8].

The AOA technique has the advantages of not requiring synchronization of the nodes nor an accurate timing reference. The use of the antenna array increases the system cost, annulling the main advantage of a UWB radio equipped with low-cost transceivers. In additional, due to the large bandwidth of a UWB signal, the number of transmission channels may be very large, especially in indoor environments. Therefore, accurate angle measurement becomes very challenging due to scattering from obstacles in the environment [9]. With regard to these facts, the AOA technique is not suited for node localization in UWB WSN.

### *B. Received signal strength measurements*

The distance between the node and the anchor can be calculated by measuring the power, or energy of the received signal. Each RSS measurement in 2D will provide a circle of radius representing the distance $d$ between the node and the anchor that made this RSS measurement, centred at the anchor. The node position is then given by the intersection of these circles (Fig. 2). To determine the distance from RSS measurements, the characteristics of the channel must be known.

A signal travelling from the node to another node is affected by slow or fast fading, shadowing and path-loss. The effect of multipath fading and shadowing can be diminished by using averaged RSS measurements over a sufficiently long time interval. The averaged received power $\bar{P}(d)$ in dB at a distance $d$ can be then modelled as [10]

$$\bar{P}(d) = P_0 - 10n\log_{10}(d/d_0) \tag{1}$$

where $n$ is the path loss exponent that measures the rate at which the RSS decreases, and $P_0$ is the known received power in dB at the reference distance $d_0$.

In the real scenario, the observation interval is not long enough to mitigate the effects of shadowing. Therefore, the received power is commonly modelled to include both path-loss and shadowing effects. The shadowing effect is modelled as a zero mean Gaussian random variable with a variance of $\sigma_{sh}^2$ in the logarithmic scale. Then, the received power $P(d)$ in dB at a distance $d$ can be expressed as $P(d) \sim N\left(\bar{P}(d), \sigma_{sh}^2\right)$ where $\bar{P}(d)$ is as given by (1).

The unbiased estimate of the distance is then given by

$$\hat{d} = d_0 \left(P/P_0\right)^{\frac{-1}{n}} e^{-\frac{\sigma_{sh}^2}{2\eta^2 n^2}} \tag{2}$$

where $\eta = \frac{10}{\ln 10}$ and $P$ is RSS measurement between the node and the anchor. The Cramer-Rao lower bound (CRLB) for unbiased distance estimator $\hat{d}$ can be expressed as

$$\sqrt{var\left(\hat{d}\right)} \geq \frac{\ln 10}{10}\frac{\sigma_{sh}}{n}d. \tag{3}$$

It can be observed from (3) that the RSS measurements are more accurate as the standard deviation of the shadowing variance $\sigma_{sh}^2$ decreases. Also a larger path-loss exponent results in a smaller lower bound, as the average power becomes more sensitive to distance for larger $n$. Finally, the accuracy is deteriorated as the distance between the node and the anchor increases. Note that the best achievable limit depends on the channel parameters and the distance between the node and the anchor. Therefore, the unique characteristic of a UWB signal, namely the very large bandwidth, is not exploited by the RSS measurements to increase the best achievable accuracy [9].

### *C. Time of arrival measurements*

TOA measurement is the summary name for one way propagation time (OWPT) and round trip propagation time (RTPT) measurements.

OWPT measurements measures the difference between the time instant of the signal transmitted by the transmitter and the time instant of the signal received by the receiver. The signal can be transmitted from the anchor as well as from the node and then received by the node or by the anchor, respectively. Each OWPT measurement will provide a circle of radius representing the distance $d$ between the transmitter and the receiver that made this OWPT measurement, centred at the receiver. The transmitter position is then given by the intersection of these circles (Fig. 2).

This OWPT technique is easy to implement, however, it requires knowledge of the time instant of the transmitted signal as well as synchronization of the node and anchor clock. Otherwise, huge position errors can occur.

RTPT measurement measures the difference between the time instant when a signal is transmitted by an anchor and the time instant when the signal is received by the same anchor or another anchor. The signal is reflected from the node. In the case, that signal transmitting anchor and signal receiving anchor are the same, the RTPT measurement will provide a circle of radius representing the half RTPT measurement, centred at the anchor (Fig. 2). In the case, that signal transmitting anchor and signal receiving anchor are different, the RTPT measurement will provide a ellipse with the focuses at anchors (Fig. 3). The node position is then given by the intersection of these circles or ellipses, respectively [11]. Conventionally, an optimal TOA estimate can be obtained using a correlator [12] or matched filter [13] receivers.

For a single-path additive white Gaussian noise channel, the best achievable accuracy of an unbiased distance estimate $\hat{d}$ derived from TOA measurement satisfies the following CRLB lower bound

$$\sqrt{var\left(\hat{d}\right)} \geq \frac{c}{2\sqrt{2}\pi\sqrt{SNR}\beta}, \tag{4}$$

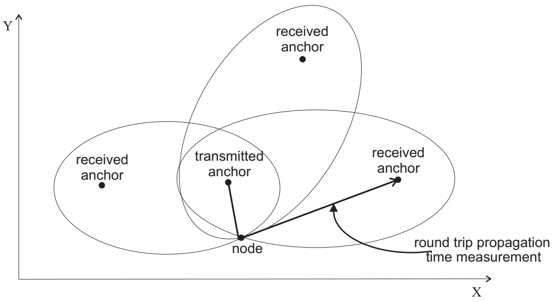



Fig. 3. RTPT-based localization.

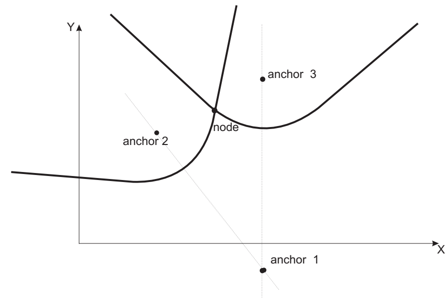



Fig. 4. TDOA-based localization.

where $c$ is the speed of light , SNR is the signal-to-noise ratio and $\beta$ is the effective (root mean square) signal bandwidth defined by

$$\beta = \left( \int_{-\infty}^{\infty} f^2 \left| S\left(f\right) \right|^2 df \bigg/ \int_{-\infty}^{\infty} \left| S\left(f\right) \right|^2 df \right)^{1/2} \tag{5}$$

and $S\left(f\right)$ is the Fourier transform of the transmitted signal $s(t)$ [10].

It is observed form (4) that unlike the RSS estimation, the accuracy of the TOA estimation can be improved by increasing the SNR and/or the effective signal bandwidth. Since UWB signals have very large bandwidth, this property facilitates very accurate distance information and so extremely accurate localization of UWB devices using time-based techniques.

### D. Time difference of arrival measurements

In the case of absence of synchronization between the node and the anchors and if there is synchronisation among the anchors, the TDOA measurements can be performed. The difference between the arrival time instants of two signals travelling between the node and the two anchors is measured. This TDOA measurement determines the position of the node on a hyperbola with a constant distance difference equal to TDOA measurement between the anchors, with foci at this anchors (Fig. 4) [14].

The measurement of TDOA can be made using two approaches. The first one is to firstly estimate TOA for each signal travelling between the node and the anchor, and then to obtain the difference between the two measurements. If $\hat{\tau}_i$ for $i = 1, 2$, denotes the TOA measurements for the signal travelling between the node and $i$th anchor the TDOA estimate can be then obtained as

$$\tau_{TDOA} = \hat{\tau}_1 - \hat{\tau}_2. \tag{6}$$

Another approach how to implement TDOA measurements is to perform cross-correlations of the signal travelling between the node and the anchors, and to calculate the delay corresponding to the largest cross-correlation value. If $r_i(t)$, for $i = 1, 2$, represents the signal travelling between the node and the $i$th anchor and $T$ is the observation interval, the cross-correlation function is given by

$$\varphi_{12}\left(\tau\right) = \frac{1}{T} \int_0^T r_1\left(t\right) r_2\left(t+\tau\right) dt. \tag{7}$$

The TDOA estimation is calculated from (7) as

$$\tau_{TDOA} = \arg\max_{\tau} \left| \varphi_{12}\left(\tau\right) \right|. \tag{8}$$

The TDOA measurement based on cross-correlation (8) works well for single path channels and white noise models, however, its performance degrade considerable over multipath channels and/or colored noise. Therefore generalised cross-correlation techniques have been proposed [10].

Note, that for these TDOA measurements, the position accuracy limits can be deduced from the expression (4) and (5). The accuracy of TDOA estimation increases as SNR increase and/or effective bandwidth increases.

These direction-based and distance-based measurement techniques can be used individually or in a number of different configurations to determine the node location. Examples of such hybrid schemes include TOA/AOA [15], TOA/RSS [16], TDOA/AOA [17] and TOA/TDOA [18] localization.

## III. Time-based localization methods

It was shawn in previous sections, that time-based localization techniques are the most suitable for the node localization in UWB WSN. Therefore, we will introduce localization methods based on time measurements. Localization methods using time measurements between node and the anchors in WSN are classify into two broad classes: non-iterative and iterative methods.

The most straightforward non-iterative localization method is the direct calculation method [19], [20], which directly solves a set of non-linear equations with two anchors for 2D positioning using one way propagation time measurements or three anchors for 2D positioning using time difference of arrival measurements. This method, however, may not effectively exploit additional measurements from another anchors to improve position accuracy. In order to exploit the additional measurements, several algorithms have been proposed, including the least-squares method (LS), constrained weighted least-squares method (CWLS), spherical-interpolation method (SI), spherical-intersection method (SX), two-stage maximum likelihood method (TSML) and Friedlander's method (FM).

The LS [15], [21], [22] linearises the set of non-linear equations obtained from time measurements and solves them by using least-squares method. The CWLS [22], [23] is an improved version of the LS with the use of weighting matrix and constraint. The SI [24], [25] and SX [26] reorganise the non-linear equations into a set of linear equations by introducing an intermediate variable, which is a function of the node position. These SI and SX solve the linear equations via least-squares method or directly without using the known relation between the intermediate variable and the node coordinates. The TSML [27] and FM [28] reorganise a set of non-linear equations into another set of equations and solve them.

Also several iterative methods have been investigated for the purpose of node localization. The Taylor series method [29],

[30] linearises the set of non-linear equations by using Taylor series and the solution looks in any iterative way. However, this method requires an initial estimate and cannot guarantee convergence to the correct solution unless the initial guess is close it. The method that firstly linearises the equations and then perform gradient searches for the minimum suffer from initial condition sensitivity and convergence difficulty is approximate maximum likelihood algorithm [31], [32]. A different iterative method for positioning comes from non-linear optimization theory. Gauss-Newton method, Levenberg-Maquardt method and quasi-Newton method, including the DFP formula (Fletcher and Powell) and the BFGS formula (Broyden, Fletcher and Powell) [33] can all be employed for position estimation iteratively.

The particular methods differ in their performance properties and computational complexity. The selection of the particular methods for the final application will depend on the required accuracy of node localization and allowed computational complexity.

## IV. Conclusion

In this paper, we have dealt with node localization methods in sensor networks with stress on UWB WSN. Firstly, we analysed the localization techniques with regard to the radio signal parameters applied for node localization. Based on short analyses it have been shown that the techniques based on time measurements (TOA, TDOA) are more suitable and therefore convenient for node localization in UWB WSN than techniques based on AOA or RSS measurements. With regard to this fact, we have provided a review of node localization time-based methods. The selection of the particular methods depends on the required accuracy of node localization and allowed computational complexity.

## Acknowledgment

This work was supported by the Slovak Research and Development Agency under the contract No. LPP-0287-06 and by European Commission under the contract COOP-CT-2006-032744.

## References

[1] A. H. Sayed, A. Tarighat, and N. Khajehnouri, "Network-Based Wireless Location: Challenges faced in developing techniques for accurate wireless location information," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 24–40, July 2005.

[2] J. Foerster, E. Green, S. Somayazulu, and D. Leeper, "Ultra-Wideband Technology for Short- or Medium-Range Wireless Communications," *Intel Technology Journal*, vol. 5, no. 2, 2001.

[3] G. Mao, B. Fidan, and B. D. O. Anderson, "Wireless Sensor Network Localization Techniques," *Computer Networks: The International Journal of Computer and Telecommunications Networking*, vol. 51, no. 10, pp. 2529–2553, July 2007.

[4] J. D. Taylor, *Introduction to Ultra-Wideband Radar Systems*. CRC Press, 1995.

[5] M. Švecová, "Node Localization in UWB Wireless Sensor Networks," December 2007, thesis to the dissertation examination, Technical University of Kosice, Department of Electronics and Multimedia Communications, Slovak Republic.

[6] J. C. Chen, K. Yao, and R. E. Hudson, "Source Localization and Beamforming," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 30–39, March 2002.

[7] D. Niculescu and B. Nath, "Ad Hoc Positioning System (APS) Using AOA," *Proc. IEEE Conference on Computer and Communications Societies, INFOCOM*, vol. 3, pp. 1734–1743, March 2003.

[8] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero III, R. L. Moses, and N. S. Correal, "Locating the Nodes: Cooperative Localization in Wireless Sensor Networks," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 54–69, July 2005.

[9] S. Gezici, Z. Tian, G. B. Giannakis, H. Kobayashi, A. F. Molisch, H. V. Poor, and Z. Sahinoglu, "Localization via Ultra-Wideband Radios: A Look at Positioning Aspects for Future Sensor Networks," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 70–84, July 2005.

[10] S. Gezici, "A Survey on Wireless Position Estimation," *Wireless Personal Communications*, vol. 44, no. 3, pp. 263–282, 2003.

[11] K. Yu, J.-p. Montillet, A. Rabbachin, P. Cheong, and I. Oppermann, "UWB Location and Tracking for Wireless Embedded Networks," *Signal Processing*, vol. 86, no. 9, pp. 2153–2171, September 2006.

[12] B. Sklar, *Digital Communications: Fundamentals and Applications*. Prentice Hall PTR, January 2001.

[13] G. Turin, "An Introduction to Matched Filters," *IEEE Transactions on Information Theory*, vol. 6, no. 3, pp. 311–329, Jun 1960.

[14] G. A. Mizusawa, "Performance of Hyperbolic Position Location Techniques for Code Division Multiple Access," Master's thesis, Faculty of the Virginia Polytechnic Institute and State University, Blacksburg, Virginia, August 1996.

[15] K. Yu, "3-D Localization Error Analysis in Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 10, pp. 3472–3481, October 2007.

[16] A. Catovic and Z. Sahinoglu, "Hybrid TOA/RSS and TDOA/RSS Location Estimation Schemes for Short-Range Wireless Networks," *Bechtel Telecommunication Technical Journal (BTTJ)*, vol. 2, no. 2, pp. 77–84, September 2004.

[17] L. Cong and W. Zhuang, "Hybrid TDOA/AOA Mobile User Location for Wideband CDMA Cellularsystems," *IEEE Transactions on Wireless Communications*, vol. 1, no. 3, pp. 439–447, July 2002.

[18] R. I. Reza, "Data Fusion for Improved TOA/TDOA Position Determination in Wireless Systems," Master's thesis, Faculty of the Virginia Polytechnic Institute, State University, Virginia, July 2000.

[19] B. T. Fang, "Simple Solutions for Hyperbolic and Related Position Fixes," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 26, no. 5, pp. 748–753, September 1990.

[20] K. Yu and I. Oppermann, "Performance of UWB Position Estimation Based on Time-of-Arrival Measurements," *Proc. Ultra Wideband Systems, Joint conf. on UWBST & IWUWBS*, pp. 400–404, May 2004.

[21] F. K. W. Chan and H. C. So, "Accurate Sequential Weighted Least Squares Algorithm for Wireless Sensor Network Localization," *Proc. 2006 European Signal Processing Conference, Italy*, September 2006.

[22] K. W. Cheung, H. C. So, W.-K. Ma, and Y. T. Chan, "Least Squares Algorithms for Time-of-Arrival-Based Mobile Location," *IEEE Transactions on Signal Processing*, vol. 52, no. 4, pp. 1121–1130, April 2004.

[23] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time Passive Source Localization: A Practical Linear-Correction Least-Squares Approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, November 2001.

[24] J. O. Smith and J. S. Abel, "Closed-Form Least-Squares Source Location Estimation from Range-Difference Measurements," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 12, pp. 1661–1669, December 1987.

[25] ——, "The Spherical Interpolation Method of Source Localization," *IEEE Journal of Oceanic Engineering*, vol. 12, no. 1, pp. 246–252, January 1987.

[26] H. C. Schau and A. Z. Robinson, "Passive Source Localization Employing Intersecting Spherical Surfaces from Time-of-Arrival Differences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 8, pp. 1223–1225, August 1987.

[27] Y. T. Chan and K. C. Ho, "A Simple and Efficient Estimator for Hyperbolic Location," *IEEE Transactions on Signal Processing*, vol. 42, no. 8, pp. 1905–1915, August 1994.

[28] B. Friedlander, "A Passive Localization Algorithm and Its Accuracy Analysis," *IEEE Journal of Oceanic Engineering*, vol. 12, no. 1, pp. 234–245, January 1987.

[29] W. H. Foy, "Position-Location Solutions by Taylor-Series Estimation," *IEEE Transaction on Aerospace and Electronic Systems*, vol. 12, no. 2, pp. 187–194, March 1976.

[30] K. Yu, H. Saarnisaari, J.-p. Montillet, A. Rabbachin, I. Oppermann, and G. T. F. de Abreu, *Ultra-Wideband Wireless Communications and Networks*. John Wiley, February 2006, ch. Localization.

[31] Y. T. Chan, W.-Y. Tsui, and H.-C. So, "Time-of-Arrival Based Localization Under NLOS Conditions," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 1, pp. 17–24, January 2006.

[32] Y.-T. Chan, H. Y. C. Hang, and P.-c. Ching, "Exact and Approximate Maximum Likelihood Localization Algorithms," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 1, pp. 10–16, Jan. 2006.

[33] I. Oppermann, M. Hamalainen, and J. Iinatti, *UWB: Theory and Applications*. John Wiley & Sons, Ltd, England, 2004.

# Internet as a bus for mechatronic system remote control

*Tibor VINCE*

Dept. of Theoretical Electrotechnics and Electrical Measurement, FEI TU of Košice, Slovak Republic

tibor.vince@tuke.sk

***Abstract*—The article presents abilities of Internet as a bus for mechatronic system remote control. The Internet protocols divided in 7 layers OSI model could handle most of the requirements for mechatronic system's bus. Most focus is paid to TCP and UDP protocols and comparing the protocols from the controlling point of view. The paper handles the advantages and disadvantages of Internet as a control and communication bus in different level of the information hierarchy. The article also compares more buses used in remote control. It contains also such a remote control system proposal.**

***Keywords*—Internet, Internet protocols, distance remote, mechatronic systems.**

## I. Introduction

The Internet begins to playing a very important role in industrial processes manipulation, not only in information retrieving. New concept of controlling, which has been paid much attention in these years is distance remote via Internet, or other words, Internet-based control. Such a type of control bus allows remote monitoring or regulation of plants or single devices over the Internet. With the progress of the Internet it is possible to control and regulate from anywhere around the world at any time. The design process for the Internet-based control systems includes requirement specification, architecture design, control algorithm, interface design and possibly safety analysis. Due to the low price and robustness resulting from its wide acceptance and deployment, Ethernet has become an attractive candidate for real-time control networks. Most difficulties cause the Ethernet MAC (Medium Access Control) protocol - persistent CSMA/CD (Carrier Sense Multiple Access with Collision Detection) protocol may cause unpredictable access delay. The goal of the article is to explore now days possibilities for Internet based controlling of mechatronic systems, eventual trends, compare UDP and TCP protocols from the controlling point of view, review of advantages and disadvantages of distance remote via Internet in different level of information hierarchy and possible solutions. Many requirements validation techniques involve building prototypes or executable specifications or waiting until the system is constructed and then testing the whole system. It could be too late and too expensive by that time to make any change in specification for control systems although certainly much can be learned by "testing" a specification.

## II. Network performance

Network performance is referred in more parameters in mutual relationship. One of performance parameters is Latency. Latency means a time required to transfer an empty message between relevant computers. Another parameter is Data transfer rate. Data transfer rate is the speed at which data can be transferred between sender and receiver in a network. The unit of this parameter is Bits/sec. For message transfer time calculating is equation 1. A third parameter of network performance is Bandwidth. Bandwidth is a total volume of traffic that can be transferred across the network. Maximal data rate formula is shown in equation 2. This maximum is only theoretical, not reachable in practice [5].

$$MTT = Lt + (Lm) / (DTR) \tag{1}$$

where *MTT* is Message transfer time, *Lt* is latency, *Lm* is Length of message and *DTR* is Data transfer time.

$$MDR = CB * \log2 (1 + (signal/noise)) \tag{2}$$

where *MDR* is maximal data rate (bps) and *CB* is carrier Bandwidth

The all parameters are pointing on the main disadvantage of controlling via the Internet – packets delivery delay. When packets are concurrently transported over an ordinary Ethernet, packets may experience a large delay due to contention with other packets in the local node where they originate and collision with other packets from the other nodes. By data transmission, four sources of delay spring up at each hop: nodal processing, queuing, transmission delay and propagation delay. The most significant part of total delay belongs to queuing. By queuing is considered the following equation 3:

$$TI = L * A/W \tag{3}$$

where TI is traffic intensity, L is packet length (bits), A is average packet arrival rate, and W is link bandwidth (bps).

If ratio L*A/W will be very small almost 0, average queuing delay is small. If ratio L*A/W rise up to 1, delays become large (exponentially) and if ratio L*A/W is bigger than 1 average delay is infinite, more "work" arriving than can be serviced. Today, there is a term Quality of Service (QoS). Quality of Service is the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow. For example, a required bit rate, delay, jitter, packet dropping probability and/or bit error rate may be guaranteed. Quality of Service guarantees are important if the network capacity is insufficient, especially for real-time streaming multimedia applications such as voice over IP and IP-TV or also real-time controlling applications, since these often require fixed bit rate and are delay sensitive, and in networks where the capacity is a limited resource, for example in cellular data communication.

## III. UDP AND TCP - INTERNET PROTOCOLS

All the transport-layer protocols (including TCP and UDP) use IP for their basic delivery services. The IP is an unreliable protocol, providing no guarantees that datagrams or packets will reach their destination intact. There are two standard transport protocols that applications use to communicate with each other on an IP network. These are the User Datagram Protocol (UDP), which provides a lightweight and unreliable transport service, and the Transmission Control Protocol (TCP), which provides a reliable and controlled transport service.

The Transmission Control Protocol (TCP) provides a reliable, connection-oriented transport protocol for transaction-oriented applications to use. This reliability is achieved through the use of a virtual circuit that TCP builds whenever two applications need to communicate.

The User Datagram Protocol provides a low-overhead transport service for application protocols that do not need (or cannot use) the connection-oriented services offered by TCP. UDP is more appropriate for any application that has to issue frequent update messages or that does not require every message to get delivered.

TCP protocol is not suitable for transport of monitoring data (where data is transferred many times per second). For this transfer is much suitable UDP protocol. For other type of data transfers (for instance transfer of commands) is much suitable protocol TCP [3].

## IV. INFORMATION ARCHITECTURE

It is becoming increasingly necessary to think in terms of integration of information and control, across the entire plant site. The term of integration, especially integration of information and control across the entire plant site, becomes more and more significant. In the manufacturing industries this is often referred to as "Computer Integrated Manufacturing" (CIM). On the surface, it would seem that the increasing use of microprocessor-based plant level devices such as programmable controllers, distributed digital control systems, smart analyzers, personal computers, etc., would make this easy. After all, most of these devices have "RS232" connectors, which enable connection to computers. Unfortunately, the real world situation is somewhat more challenging. If we began to hook all these RS232 ports together, there would soon be an unmanageable mess of wiring, custom software and little or no communication. This problem solution results in integration these devices into a meaningful "Information Architecture". This Information Architecture can be separated into 4 levels with the sensor/actuator level as shown in Fig. 1, which are distinguished from each other by"4Rs" principle criteria.

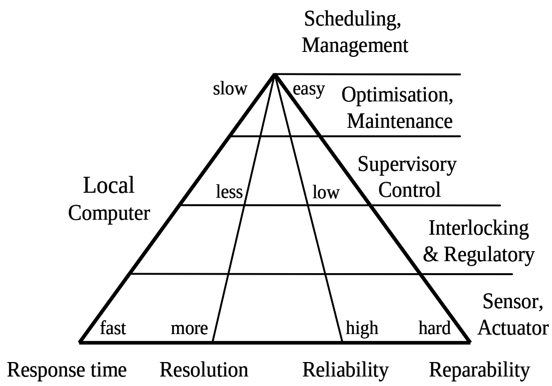



Fig.1 Information Architecture

The 4Rs criteria are: Response time, Resolution, Reliability and Reparability.

*Response time*: as one moves higher in the information architecture, the time delay, which can be tolerated in receiving the data, increases. Conversely, information used at the management & scheduling level can be several days old without impacting its usefulness.

*Resolution*: an Abstraction level for data varies among all the levels in the architecture. The higher the level is, the more abstract the data is.

*Reliability*: Just as communication response time must decrease as one descends through the levels of the information architecture, the required level of reliability increases. For instance, host computers at the management & scheduling level can safely be shut down for hours or even days, with relatively minor consequences. If the network, which connects controllers at the supervisory control level and/or the regulatory control level, fails for a few minutes, a plant shutdown may be necessary.

*Reparability*: The reparability considers the ease with which control and computing devices can be maintained.

Local computer on supervisory control level is able communicate with higher levels of information architecture via Internet, but there is also possibility to use the Internet also in lower levels of the Information architecture. The Internet can be linked with the local computer system at any level in the information architecture, or even at the sensor/actuator level. These links result in a range of 4Rs (response time, resolution, reliability, and reparability). For example, if a fast response time is required a link to the control loop level should be made. If only abstracted information is needed the Internet should be linked with a higher level in the information architecture such as the management level or the optimization level.

## V. INDUSTRY'S BUSES

There are various types of buses using in Industry to connect of different devices, which are supported on the type of interface/bus [6]:

*AS-Interface* – the Actuator Sensor Interface offers many of the benefits of more powerful and expensive fieldbuses, but at much lower cost and as a much simpler installation. Transmission of analog signals via time multiplex procedure. Data and power via the same line. No termination necessary. Address setting automatically from the master or via service tool. ASI conforment power supply required. It supports 62 nodes, max 300m with 3 repeaters. Date rate is 167kbit/s. Addressing is Master/Slave.

*CANopen* – is a CAN-based higher layer protocol. It was developed as a standardized embedded network with highly flexible configuration capabilities. Node removal without severing the network is possible. Provisions for the typical request/response orientated network communications. Provisions for the effiecient movement of data framentation for moving larger bodies of information It supports 127 nodes, 25-5000m (depending on boudrate). Data rate is from 10kbit/s to 1Mbit/s, addressing is Master/Slave, Peer-to-Peer, Multi-cast and Multi-master.

*ProfiBus* – is a Multi-Master System and makes possible the mutual operation of several automation, engineering or visualizing systems at a Bus. The Masters, also designated as active devices, define the data traffic on the Bus. When in possession of the access permission (Token), they can send data without external requests. The Slaves, designated as passive devices, have no Bus access permission. They can only confirm received messages or send messages when requested by a Master. Baud rates from 9.6 kBaud up to 12 MBaud are supported. A maximum of 126 devices can be operated at the Bus. Profibus also supports Broadcast and Multicast communication. Network length is possible 100-1200m, addressing is DP: Master/Slave, Cyclic, Polling, DPV1: Cyclic, Polling + acyclic data transfer.

*EthernNet/IP* – The Industrial Ethernet Protocol (Ethernet/IP) has been developed by ODVA with strong support from Rockwell Automation. It uses the Control & Information Protocol (CIP) which is already well known from ControlNet and DeviceNet. Network size (number of nodes) is scalable and nearly unlimited. Network size can be 100m (10/100 Base-T) or 35-2000m (fiber optic). Standard layers 1-4 providing Ethernet data transmisson, bus access, internet protocol (IP) and TCP & UDP protocols. CIP "implicit" and "explicit" messaging with encapsuslation technology. Message routing between EtherNet/IP, DeviceNet & ControlNet.

## VI. REMOTE SYSTEM PROPOSAL

Convenient way for demonstrating remote architecture could be architecture for remote controlling of mechatronic system. Efficient way of monitoring or controlling a Mechatronic system is to use appropriate visualization software. Unfortunately, the slow rate and instability of the Internet connection may restrict the real-time control and feedback of remote tasks. To efficiently implement the teleoperation of mechatronic system, most systems apply visualization tools to simulate the environment. The main advantage is the achievement of fast response to the operator's actions because smaller data package is required to update the virtual mechatronic system and its environment. It provides the operator with a "live" virtual representation of the scene instead of the delayed video images.

It can also increase the efficiency of the operator performance because the operator can choose appropriate points of view, zoom the scenes and make some objects transparent or semitransparent, etc. The augmented reality can also be used in the system to get a better visualization and help the operator get more immersion of the virtual environment.

In the architecture design, a distance remote mechatronic system generally includes three major parts: client, server and controlled mechatronic system. The general remote mechatronic system architecture is shown in Figure 2. The client part is the interface for the operations. It includes computers, visualization software with user interface for operators. Client computer receives state information of remote mechatronic system via Internet. Received information will be processed and evaluated in remote computer.

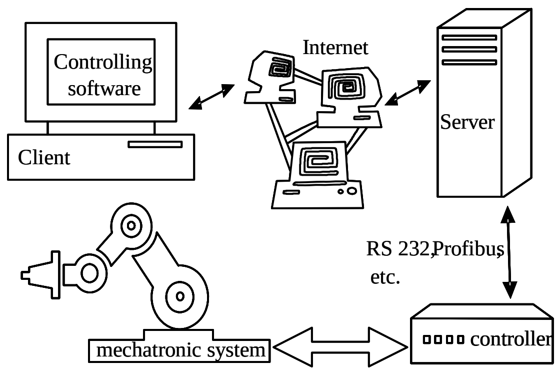



Fig.2 Remote system architecture

The server part contains a server computer, which is connected to the real mechatronic system. The server software includes different models of data processing such as sensor data acquisition and processing. Server contains all required drivers and devices for communication with mechatronic system. Communication of server with mechatronic system could be based on several ways (RS232, Profibus,CAN, MPI etc.). Server can be connected with mechatronic system directly (server is directly connected with all sensors and actuators), or indirectly (server is connected with control unit of mechatronic system – PLC, microcontroller, etc., but these days is also Ethernet based connection possible). The third part of system architecture is mechatronic system itself. It contains all sensors, actuators and all devices with specific purpose.

Common way for distance remote is, when remote client computer has limited functions – only start/stop of mechatronic system, or choosing specific program for mechatronic system to run on server site. But Internet speed progress open possibility for real-time control from client site.

So there is possibility that server would serve only as an interface between remote client and mechatronic system. This way of distance remote is possible, if the mechatronic system is “sufficient slowly” and server-client connection is sufficient fast. Between client computer and server could be thousands of kilometers, or they could be in the same room. The difference is in the communication delay, but generally the system is the same. The communication service (the bus) can be achieved by wired connection, mostly Ethernet, or wireless – very popular WiFi.

## VII. CONCLUSION

It is become to be a standard, that many control elements have been embedded with Internet-enabled functions, for example, PLC with TCP/IP stack, smart control valves with a built-in wireless communication based on TCP/IP protocol, and process control computer (DCS) with an Internet gateway. There possibility that some mechatronic system could be connected directly to the Internet (without a necessity of a server computer). On the basis of done analysis it is evident that the existence of server as a gate to the Internet for mechatronic system is still highly recommended (because of capriciousness of Internet, computer crime and many other reasons). By utilizing of UDP Internet protocol it is possible to regulate real-time systems with tenths milliseconds of feedback. When compare Ethernet as a bus with other standard types of industrial bus, there are more advantages and disadvantages. The most powerful advantage is nearly unlimited size of bus, possible huge distance, open system of the internet protocols and accessibility of the Internet.

## ACKNOWLEDGMENT

The paper has been prepared by the support of Slovak grant projects VEGA

No. 1/4174/07, VEGA No. 1/0660/08, KEGA 3/5227/07, KEGA 3/6386/08 and KEGA

3/6388/08.

## REFERENCES

[1] Yang, S.H., Tan, L.S., Chen X: Requirements Specification and Architecture Design for Internet-based Control Systems, proceedings of the 26th Annual International Computer Software and Applications Conference (COMPSAC’02), 2002.

[2] [2] Kweon, S. K, Cho M., Shin K. G.: Soft Real-Time Communication over Ethernet with Adaptive Traffic Smoothing, IEEE Transactions on parallel and distributed systems, VOL. 15, NO. 10, October 2004.

[3] [3] Hall E.: Internet Core Protocols: The Definitive Guide, O'Reilly & Associates (February, 2000) USA, ISBN: 1-56592-572-6.

[4] [4] Yang, X., Chen Q.: Virtual Reality Tools for Internet-Based Robotic Teleoperation, proceedings of the 8th IEEE International Symposium on Distributed Simulation and Real-Time Applications (DS-RT’04).

[5] [5] Fonda C., Postogna F.: Computer networking basics, ICTP workshop on telecommunications: science, technology and applications. Trieste, 15th September - 3rd October 1997

[6] http://www.anybus.com/technologies/technologies.shtml, 10.2.2008

# PS-PWM Soft Switching DC-DC Converter

*Vladimír RUŠČIN*

Dept. of Electrical, Mechatronic and Industrial Engineering, FEI TU of Košice, Slovak Republic

vladimir.ruscin@tuke.sk

**Abstract—Soft switching DC-DC converter is described in this article. Full bridge inverter has the snubber capacitors parallel connected to switches to improve switching properties and to reduce turn-off looses. Power center tapped transformer rectifier is designed for switching frequency of 50 kHz. Each leg of the rectifier consists of series connection of the MOSFET transistor and diode to ensure reverse ability.**

## I. Introduction

Many of pulse-width modulated converters are controlled by phase shift. Main types of soft switching DC-DC converters have a simple structure, but their conduction losses are high as a result of circulating current that flows trough primary winding of transformer. In some of the zero voltage and zero current switching (ZVZCS) converters this circulating current can be eliminated with a complicated auxiliary circuits.

Another way how to keep low circulating current is to use a controlled rectifier on the secondary side of power transformer. We do not need to generate special signals for the output rectifier. It is possible to utilize the signals used by Phase-Shifted Pulse Width Modulation (PS-PWM). Rectifier switches need to have reverse capability. This can be solved as a combination of the switch without reverse capability and series diode. But in the converters with low output voltage and high output current the series combination of the semiconductor devices causes high conduction losses.

### A. Operating principle

The schematic of the converter with secondary switches is displayed in Fig. 1. Full-bridge inverter consists of four IGBT transistors $T_1$ - $T_4$ and their anti-parallel freewheeling diodes $D_1$ – $D_4$. The reduction of the transistors turn off losses is ensured by parallel connected snubber capacitors $C_1$ - $C_4$. The secondary switches are created by MOSFET transistors $T_5$ - $T_6$ and in series connected diodes $D_5$ - $D_6$. Output power is controlled by phase shift of control pulses for MOSFET switches on the secondary side of power transformer compared to pulses for IGBT transistors $T_1$-$T_4$ in inverter. (Fig. 2)

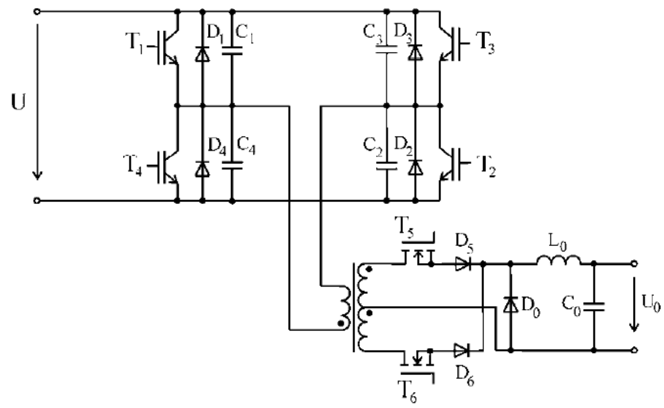



Fig. 1. Scheme of the DC-DC converter

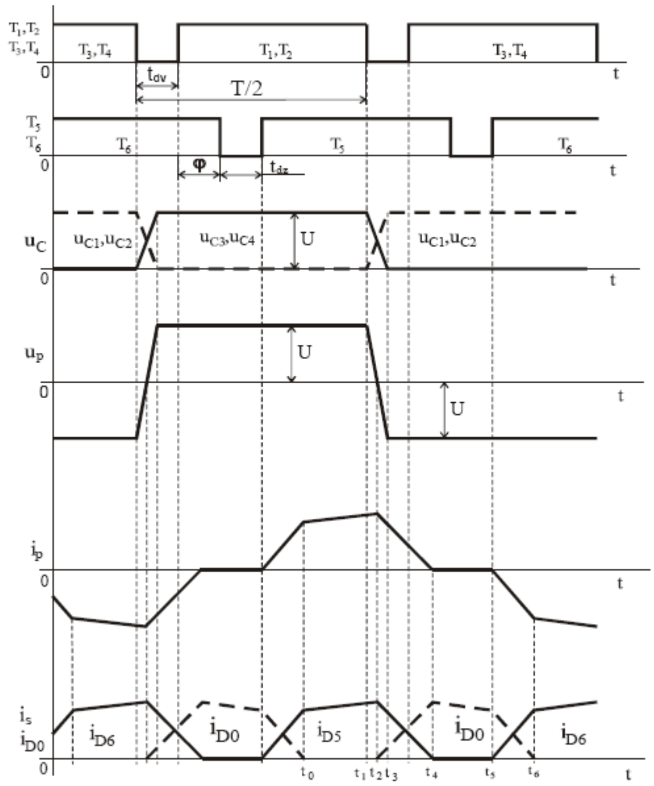



Fig. 2. Basic converter waveforms

### B. Results

The operation of the converter was verified by simulation and on laboratory model. Control pulses for individual semiconductors components were generated by microcontroller 8051 series, which is slave to other microcontroller from the same series. The output voltage of the converter is regulated by phase shift, what is displayed in Fig. 3. The signals from microcontroller are adapted by the driver UC3708 for inverter switches and by driver TC429 for rectifier switches.

Waveforms of the collector emitter voltage and collector current of transistor $T_1$ are shown in Fig. 4. the detail at turn off is displayed in Fig. 5. Capacitors $C_1$ - $C_4$ form a snubber network that causes the reduction of turn off losses of the inverter transistors.

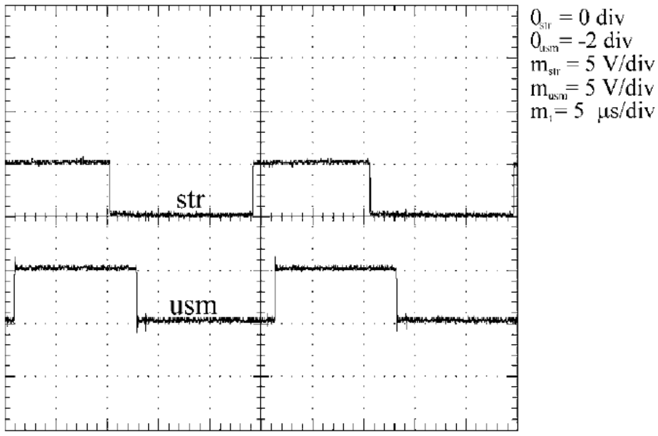



Fig. 3. Slave processor output

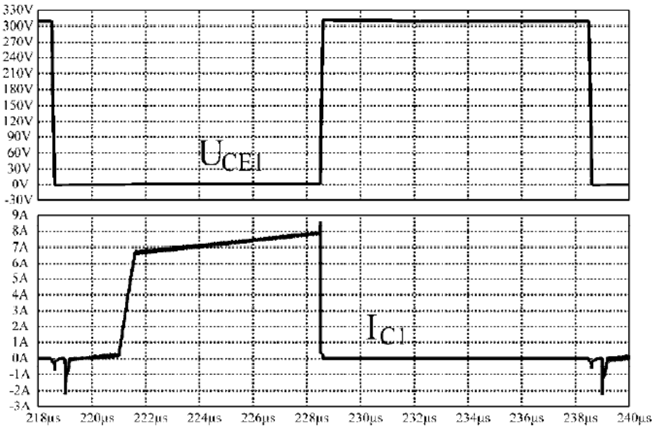



Fig. 4. Transistor $T_1$ switching

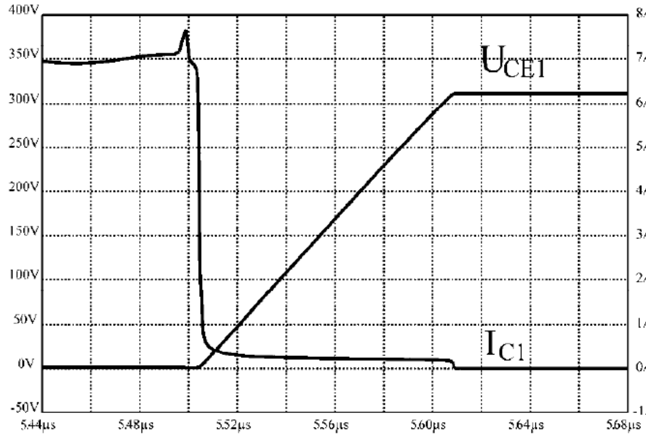



Fig. 5. Transistor $T_1$ turn off detail

The rectifier transistors are turned on and turned off at zero current (Fig. 6).

## II. Conclusion

The DC-DC converter with phase shift and soft switching is described in this article. The converter is controlled with two microcontrollers 8051 series, where the width of the phase shift is calculated in the master of those two processors and the equivalent pulses for transistors are generated by slave microcontroller. Soft switching of the IGBT transistors of the inverter is insured by the nondissipative snubbers which are

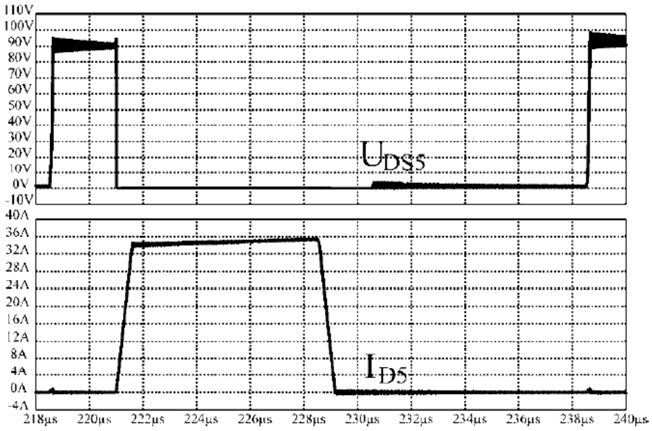



Fig. 6. Transistor $T_5$ switching

displayed in the Fig. 1. The rate of rise of the drain current of the transistors on the secondary side is slowed down by the leakage inductance of high-frequency power transformer. The leakage inductance causes a reduction of turn on losses. Using high frequencies allows a remarkable reduction of the converter volume and mainly its weight.

## Acknowledgment

This work was supported by Scientific Grant Agency of the Slovak Republic under contract VEGA No. 1/2178/05.

## References

[1] Masakazu Michihira: *A novel quasi-resonant DC-DC converter using phase-shift modulation in secondary side of high-frequency transformer*, IEEE PESC record, 1996 pp. 670-675G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] J. Dudrik: *Current Source for Arc Welding, Elektro* (journal), 1993, No.1, pp. 450-455 (in Slovak).

[3] J. Dudrik, *Soft Switching PWM DC-DC Converters for High Power Applications*, Proc.of the Int. Conf. IC-SPETO 2003, Gliwice-Niedzica, Polen, 2003, pp.11-11a-11f-12.

# 2$^{nd}$ section: Informatics & Telecommunications

# Building ontologies using Coloured Petri Nets

*Iveta ADAMUŠČÍNOVÁ*

Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

iveta.adamuscinova@tuke.sk

***Abstract*—Nowadays, the knowledge-based systems present one of the most observed area in scientific research and related trend is, among others, an effort to find ways how to capture, model and represent the knowledge about specific domain within ontologies. In this work, a concept of powerful formalism Coloured Petri Nets is used to model the knowledge and subsequently, there is presented an outline about how this concept can be used to build an ontology using the ontological environment Protégé2000.**

***Keywords*—Coloured Petri Nets, knowledge modeling, knowledge representation, ontology.**

## I. Introduction

Knowledge about application domain of used software system are important for software engineering processes related to management, maintenance and modification of these software systems. Searching, processing, maintaining and using the knowledge belong to the main activities in all aspects of life, not excluding the software engineering.

A great problem presents a choice of the standards, methods or tools in context of integration of the future wide and complex, often distributed, goal systems. An effective representation of knowledge and its integration with a version of some software systems' categories presents a crucial aspect of software systems' development [1].

Although, there exist many methods that support modeling of knowledge in structural way, there are considerably much less methods that are able to model and represent the knowledge about behaviour of concurrent, discrete-event dynamic systems, often of distributed nature. Petri nets are one of the most widely used methodologies for modeling that kind of systems supported by powerful theoretical background [2].

Many works present how they are able to depict the important knowledge not only about the structure, but also about the dynamic behaviour of the studied system, mainly focusing on using the concepts of Generalized or Fuzzy Petri nets [2], [3]. However, even though they offer a sufficient way of representing these aspects of complex system, they face some disadvantages that can inhibit their further use in knowledge-based systems. The construction of Generalized Petri nets of complex distributed systems is often very challenging and time-consuming [4] and Fuzzy Petri nets lack accessible tools that would ease its integration into the standard knowledge building process.

On the other hand, Coloured Petri Nets (CPN) present a widely used formalism to model complex systems and offer some well developed tools that support the construction of CPN and offer its portability using the extensible markup language format.

In this work, there is also presented a way how to profit from the mentioned advantages and build up a new ontology using the created CPN, taking into account its possible integration into the concept of Semantic Web.

## II. Knowledge modeling and representation

### A. Knowledge models

Modeling of knowledge in terms of knowledge-based systems consists of its representation through the concept of some well known knowledge model.

A thorough understanding of different knowledge representations is a vital part of Knowledge Engineering, since the ease of solving a problem is almost completely determined by the way the problem is conceptualised and represented. The same is true for the task of communicating knowledge. A well-chosen analogy or diagram can make all the difference when trying to communicate a difficult idea to someone, especially a non-expert in the field [5].

Knowledge models are structured representations of knowledge using symbols to represent pieces of knowledge and relationships between them. Knowledge models include:

- Symbolic character-based languages, such as logic
- Diagrammatic representations, such as nets
- Tabular representations, such as matrices
- Structured text, such as hypertext [5].

### B. Knowledge representation

The knowledge about certain domain (e.g. distributed systems) are usually represented and built using ontologies. An *ontology* defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them.

**Definition***:* Ontology is description of problem domain, where entities of the domain, its properties and its relations are described [6].

Ontology has become a very important aspect in many applications to provide a semantic framework for knowledge management. The huge advantage of ontology is not in processing, but in sharing meaning, emergence and discovery of gaps and for improving a tacit knowledge transfer.

Ontology may contain information in a specified declarative language, but it may also include unstructured or unformalized information expressed in a natural language or a procedural code.

If knowledge is represented by ontology, naturally there are several ways how to present it - by object trees, graphs and ontology languages [6].

Ontology can be represented by *UML*, any object oriented language, *RDF, DAML+OIL, OWL* or any other representation which can define objects, properties and its relations, so the representation concept by *Coloured Petri Nets* may be also suitable.

The biggest problem of using this concept presents the portability of constructed CPNs, as there don't exist any profiles or plugins in existing accessible ontology environments which would support direct construction of CPNs and therefore their integration into an existing (or new) ontology.

## III. Abstraction of Coloured Petri nets into ontology Concepts

### A. Coloured Petri nets

Coloured Petri Nets are one of the most accepted formal methods for modelling and analysing of complex (e.g. distributed) systems. They are based on the concepts and terminology taken from the modern programming languages and so it's relatively simple to learn to use it, especially for people who have some programming practice.

**Coloured Petri Nets** (CPNs) are basically an elaboration of ordinary Petri nets. In a CPN, each place is associated with a 'colour', which is a type. Places can contain a multiset of tokens of their declared type. Each input arc to a transition is annotated with an expression (possibly containing variables) that represents a multiset of tokens.

For a transition to be enabled, it must be possible to match the expression on each input arc to a sub-multiset of the tokens in the associated input place. This may involve binding variables. In addition, a Boolean expression associated with the transition (its guard) must evaluate to true, taking into account the variable bindings [4].

### B. Fundamental ontology concepts

In previous paragraph an ontology was defined as a *formal explicit description of concepts in a domain of discourse* (**classes** - sometimes called *concepts*), where *properties* of each concept describe various features and attributes of the concept (**slots** - called also roles or properties), and also *restrictions* on slots (**facets** or role restrictions)). An ontology together with a set of individual **instances** of classes constitutes a knowledge base.

*Classes* are the focus of most ontologies. Classes describe concepts in the domain. A class can have subclasses that represent concepts that are more specific than the superclass.

*Slots* describe properties of classes and instances. In practical terms, developing an ontology includes:

- defining classes in the ontology,
- arranging the classes in a taxonomic (subclass–superclass) hierarchy,
- defining slots and describing allowed values for these slots,
- filling in the values for slots for instances.

A knowledge base can be created afterwards by defining individual *instances* of these classes filling in specific slot value information and additional slot restrictions [7].

### C. Analysis and abstraction of CPN concepts into ontology concepts

With using some level of abstraction and theory of Coloured Petri Nets, it's possible to realize that ontology *classes* present the same concept as the CPN **colour sets** related to net's places, with **tokens** and **places** as the abstractions of ontological *instances*.

Using this generalized *abstract* concept, it is rather simple to think of CPN **transition** nodes as of *operations* of different complexity with help of which it's possible to create relations to other net's place nodes of same colour set (and therefore representing only the change of the instance's state) or relations to the places of different colour sets which in ontological terms can represent a *relationship* between two different classes, for example a ***„subclass-of"*** type of relationship. Another standard ontological relationship ***„instance-of"*** is represented directly by relating a specific colour set to chosen net's places.

The *attributes* of ontology classes, also called slots were easily recompensed by CPN's colour sets properties, as colour sets can be designed as complex sets (and even subsets), that consist of several colour sets of various types and hence the properties.

*Restrictions* on the slots can be additionally presented in CPNs by implementing the **arc** and **guard inscriptions**.

Obviously, to use the concept of Coloured Petri Nets only as the abstraction for ontological purposes, as there is no theoretically proved background of this concept yet, it's inevitable to follow several specific rules. An initial analysis and following creation of the *terminology glossary* is crucial, so as the setting of *conventions* of naming the CPN's nodes and arcs. It's important to mention that the final CPN is often processed and evaluated by ontology environments in its representation by markup language thereby the accent on the naming and using the specified structures is necessary [2].

In Fig. 1, there is an example of CPN, which models the processing of particular jobs by server, using the names and inscriptions by agreement with respect to terminology of designed ontology.

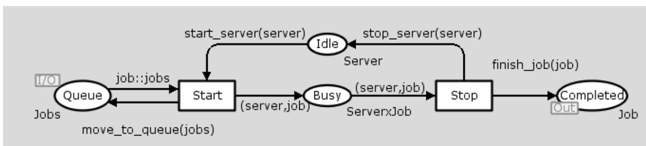


Fig. 1. Example of CPN model

By using the presented concept and strictly following the given rules of naming and using the CPN structures, it's possible to model almost any kind of knowledge about complex systems, as the presented abstraction doesn't limit the modeling and formal power of CPN, just suggests some outlines of working with CPN to ease integrating the modeled knowledge into the ontological terms and environments.

## IV. Building a New Ontology

In previous section, the abstract connection between the concept of CPNs and the concept of ontologies' knowledge representation was presented. However, the problem presents the way how to transform the existing, by CPN modeled knowledge into a representation understandable to an ontological environment.

For purposes of this work, a well known ontology tool, called *Protégé2000* was used. It is a free, open source ontology editor and a knowledge acquisition system. Protégé is being developed at Stanford University in collaboration with the University of Manchester. It allows creating, editing, customizing and maintaining the ontologies for various domains, supporting a concept of a so-called Semantic Web [8].

In this paper, a plan for using a concept of Semantic Web as a framework for designed knowledge-based system was proposed, so during the research on possible ontology representation of modeled knowledge, the support of this concept was highly considered.

### A. Semantic Web from perspective of ontology languages

The *Semantic Web* is an evolving extension of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content.

The languages supporting this concept are following:

**XML** provides an elemental syntax for content structure within documents, yet associates no semantics with the meaning of the content contained within.

**XML Schema** is a language for providing and restricting the structure and content of elements contained within XML documents.

**OWL** (Web Ontology Language) adds more vocabulary for describing properties and classes: among others, relations between classes, cardinality, equality, richer typing of properties, characteristics of properties, and enumerated classes [9], [10].

There exist many other ontology languages supporting the concept of Semantic Web, their description is however beyond the scope of this paper. Especially, ontology language OWL is the most widely used knowledge representation supporting also by tool Protégé2000, so it makes OWL a suitable candidate to fit into this work's proposed ideas [10].

### B. Transformation of CPN knowledge models into OWL

For modeling the knowledge by Coloured Petri Nets, a modeling tool developed by University of Aarhus, called *CPN Tools* was used [11]. It was chosen not only for its high level modeling features, but also because of fact, that all constructed CPN models are stored in specific XML format, called CPNXML, that can be after some modifications transformed into an OWL format [12].

To proceed the needed modifications, a special *CPNXML to XML format parser* was constructed. This parser modifies the CPNXML file by getting rid of some obsolete data (as CPN Tools environment settings), modifying the existing tags and also creating additional tags into the final XML file that can be correctly processed by Protégé2000. This environment contains built-in parser that transforms XML format files into the OWL language (Fig. 2), which creates a solid base for building and modifying proposed ontology about the studied domain, which is dynamic behaviour of studied complex systems. Following this concept, it's possible to expect a successful integration of built ontology within Semantic Web.

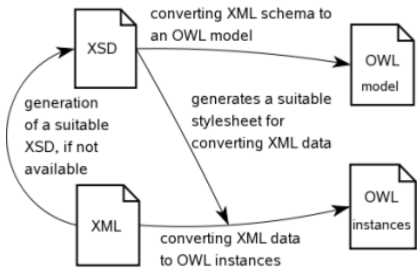



Fig. 2. Conversion of modified XML file into OWL

## V. Conclusion

In this paper, a possibility of modeling the knowledge with support of Coloured Petri Nets was presented, so as the way of transforming the final CPN model and a possibility of its use to build a base of new ontology about the domain of interest.

The main motivation behind the use of this approach was an effort to take advantage of a powerful formalism of CPN and also of well developed modeling tool. The fact that the CPN models are stored in XML-like format which after some parsing processes can be converted into a semantically richer representation of OWL ontology language, presented a main advantage. In this form, the modeled knowledge can be almost directly processed and interpreted by various ontology-based software applications with no additional plugins or profiles.

The presented concept is however still under development and further support for large and even hierarchical CPNs is considered for future research. The aim of the future direction is to build a complex ontology about the domain of distributed systems and its integration within Semantic Web.

## Acknowledgment

The paper was prepared within the project "Life cycle and architectures of program systems based on the knowledge", No. 1/0350/08, 2008-01-01 2010-12-31 with the support of VEGA.

## References

[1] Z. Havlice, M. Paralic, et al. „Knowledge-Based Software Life Cycle and Architectures“, *Computer Science and Technology Research Survey*, Košice, elfa, 2007, pp. 47-68.
[2] M. Peleg, I. Gbashvili „*Integrating bio-ontologies with a workflow/Petri Net model to qualitatively represent and simulate biological systems*“, Stanford University, 2006.
[3] J. Knybel, V. Pavliska „*Representation of Fuzzy IF-THEN rules by Petri Nets*“, Institute for Research and Applications of Fuzzy Modeling, University of Ostrava, *Research report No. 84*, 2005.
[4] K. Jensen „Coloured Petri Nets - Basic Concepts, Analysis Methods and Practical Use: Volume 1“. Springer-Verlag, 1997.
[5] *Knowledge Models*, <http://www.epistemics.co.uk/Notes/90-0-0.htm/>
[6] M. Laclavik „*Ontology and Agent based Approach for Knowledge Management*“, Slovak Academy of Sciences, Bratislava 2005.
[7] N. Noy, D. McGuinness „*Ontology Development 101*“, Stanford University, 2004.
[8] Protégé Ontology Editor, <http://protege.stanford.edu/>.
[9] T. Berners-Lee, J. Hendler, O. Lassila "*The Semantic Web*". Scientific American Magazine, May 2001.
[10] W3C „*OWL Web Ontology Language Guide*“, W3C Recommendation
[11] *CPN Tools*, <http://wiki.daimi.au.dk/cpntools/>.
[12] H. Bohring, S. Auer „*Mapping XML to OWL Ontologies*“, University of Leipzig, Leipzig Germany, 2004

# Reachability analysis of Time Basic Petri nets

*Attila N.Kovács*

Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

attila.n.kovacs@tuke.sk

***Abstract*—Several extensions of Petri nets that are dealing with time have been proposed. In this paper we deal with a special type of high-level timed Petri nets called Time Basic Petri nets. Aim of this paper is to put a light on the possibility of solving the reachability problem for this kind of Petri nets. We will introduce some methods and constructions that help us to solve this crucial problem.**

***Keywords*—M$_\omega$ automaton, Reachability analysis, Stabile loops, Time basic nets.**

## I. INTRODUCTION

The systems whose functionalities are defined with respect to time and whose correctness can only be assessed by taking time into consideration are called time-critical systems. As an example can serve a typical computer system with a mouse input device where making two clicks on the mouse (the first and the second click at the beginning and at the end of the time interval duration 2 seconds respectively), has a quite different meaning than a double click within the time interval duration an half of second.

Solving the reachability problem for a Petri net opens the door to examine other important properties like liveness or coverability. Reachability problem takes in the worst case exponencial time to solve it [4] and it is one of the crucial problems linked with Petri Nets (P.N.).

In this paper we introduce Time Basic Petri nets that can model time-critical systems. In section III. we introduce the reachability problem and summarize some well known methods and approaches to solve this problem in ordinary Petri nets. Section IV. deals with the time reachability analysis of Time Basic Petri nets and reveals a possible solution for the reachability problem.

## II. TB NETS

Time Basic nets (TB nets) are an extension of Petri nets (PNs). TB nets have been introduced in [1]. In this paper, we introduce them using a slightly different notation than in [1].

### *A. TB nets definition*

A TB net is a 6-tuple $<P,T,\Theta,F,tf,m_0>$ where

*P, T* and *F* are, respectively, the sets of places, transitions, and arcs of nets. The preset of transition *t*, i.e., the set of places connected with *t* by an arc entering *t*, is denoted by $^\bullet t$.

Θ (a numeric set) is the set of values (timestamps), associated with the tokens. A timestamp represents the time at witch the token has been created. In the following, we assume Θ to be the set of non-negative real numbers, i.e., time is assumed to be continuous.

*tf* is a function that associates a function $tf_t$ (called time-function) with each transition t. Let *en* denote a tuple of tokens, one for each place in $^\bullet t$. Function $tf_t$ associates with each tuple *en* a set of value $\theta$ ( $\theta \subseteq \Theta$ ), such that each value in $\theta$ is not less than the maximum of the timestamps associated with the tokens belonging to tuple *en*.

### *B. Enabling tuple, enabling, firing time, enabling time*

Given a transition *t* and a marking *m*, let *en* be a tuple of tokens, one for each input place of transition *t*. If $tf_t(en)$ is not empty, *en* is said to be an enabling tuple for transition *t* and the pair $x = <en, t>$ is said to be an *enabling*. The triple $y = <en, t, \tau>$ where $<en, t>$ is an enabling and $\tau \in tf_t(en)$, is said to be a firing. τ is said to be the firing time. The maximum among the timestamps associated with tuple *en* is the enabling time of the *enabling* $<en, t>$.

The dynamic evolution of the net (its semantics) is defined by means of firing occurrences, which ultimately produce firing sequences. Markings represent the states of the modeled system; transitions represent events; and firing sequences represent evolution of the modeled system.

Axioms presented in [2] tell us the following facts: time never decreases; if the system does not stop, time eventually progresses. The set of all firing sequences that satisfy these axioms defines the *Monotonic Weak Time Semantics* (MWTS). Adding further restrictions to the firing possibilities we get *Strong Time Semantics* (STS).

### *C. Time Interval Semantics of TB Nets*

Interval semantics of TB nets give us the opportunity to assign to each token a time interval (TI) in which the token can be created. Using TI instead of timestamps gives us a bigger modeling power. Any token (chronos) $\tau$ in TI $\tau = [\tau_i, \tau_a] \subseteq \mathbb{R}^+$. In TI semantics we replace any enabling tuple $en = (m(p_1)...m(p_{|^\bullet t|}))$ with a corresponding collection of TIs that is called *Time Interval Profile* (TIP). Besides the set operation $\cap$, $\cup$, $(\ )^c$ a new operation "+" is defined. Given a constant $c \in \mathbb{R}^+$ and TIs τ, τ' and τ'' we have: $c + \tau = [\tau_i + c, \tau_a + c]$, $c \,.\, \tau = [\tau_i \,.\, c, \tau_a \,.\, c]$; $\tau' + \tau'' = \tau \Leftrightarrow \tau = [\tau_i, \tau_a]$, $\tau' = [\tau'_i, \tau'_a]$, $\tau'' = [\tau''_i, \tau''_a]$, $\tau_i = \tau'_i + \tau''_i$, $\tau_a = \tau'_a + \tau''_a$.

Given TB net $N_0 = (P, T, \Theta, pre, post, tf, q_0)$, then tft(en) has for given *en* the unique representation

$$tf_t(en) = \tau en + tf_t(0) \qquad (1)$$

where $\tau en$ is a TI that depends on *en*, $tf_t(0)$ is a TI that does not depend on *en*. To put it another way, any t-generated TI $\tau_t$ can be represented as a sum of two TIs: $\tau$en- the determinate TI that depends on TIP *en* in question and on *t* (or $tf_t$) and a constant TI $tf_t(0)$, which depends only on the structure of the TB nets in question.

## III. Reachability Analysis of Ordinary Petri Nets

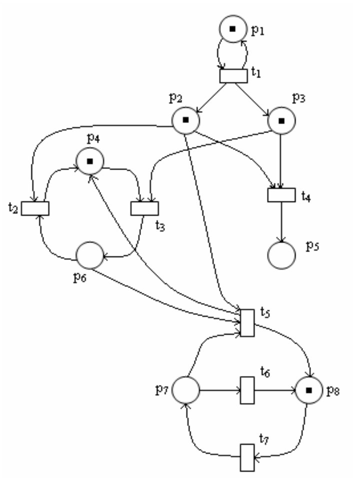



Fig. 1. Ordinary Petri net of a voice station

Given Petri net $N_0$ = *(P, T, pre, post, m$_0$)* and a *k*-dimensional nonnegative integer vector *q*; the problem whether $q \in \Re(N_0)$ is called (the instance of) the reachability problem (RP) of PN $N_0$ for state *q,* where $\Re(N_0)$ is the set of all reachable markings in $N_0$.

Time-critical systems represent an important class of discrete systems for which the reachability analysis with respect to time is an inevitable task to be done during design and analysis of such systems. We call the problem in that case *time reachability problem* or *time reachability analysis* (TRA).

Given a PN $N_0$ = *(P, T, pre, post, m$_0$)*, from the initial marking $m_0$, we can obtain as many "new" markings as the number of the enabled transitions. From each marking, we can again reach more markings. This process results in a tree representation of the markings. Nodes represent markings generated from $m_0$ and its successors, and each arc represents a transition firing, which transforms one marking to another. The above tree representation, however, will grow infinitely large if the net is unbounded. To keep the tree finite, we use a special symbol ω, which can be thought of as "infinity." It has the properties that for each integer *n*, $\omega > n$, $\omega \pm n = \omega$ and $\omega \geq \omega$ . This means, that if we have some marking $m_i$ = (1,0,1,1) and another marking reachable from this marking $m_j$= (1,0,2,1) we replace the marking $m_j$ with the marking in the form (1,0,$\omega$,1).

The *coverability graph* of a PN is a labeled directed graph *G = (V, E)*. Its node set *V* is the set of all distinct labeled nodes in the coverability tree, and the arc set *E* is the set of arcs labeled with single transition $t_k$ representing all possible single transition firings such that $m_i[t_k>m_j$, where $m_i$ and $m_j$ are in *V*.

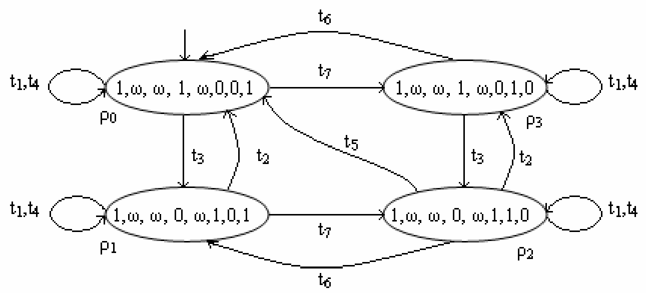



Fig. 2. Finite state automaton of type $M_\omega$ for the voice station

From the coverability tree we can create finite state automaton (fsa) of type $M_\omega$ (in the case of TB nets we call it $M_{\tau\omega}$) [4]. If we have markings like mentioned above, than the automaton will have just one state (marking) labeled as (1,0,$\omega$,1). In the Fig. 2 is the $M_\omega$ automaton for the voice station illustrated in the Fig. 1.

As it can be seen from the Fig. 2, using the fsa of type $M_\omega$ alone we can not decide, whether some marking is reachable or not. This is because we have no information telling us what values can $\omega$ represent.

In [4] the Modified Integer Linear Programming (MILP) interpretation of RP was proposed. It was proved that ρ – simple loops of the fsa $M_\omega$ connected with the MILP play profound role in the reachability analysis of the ordinary PN.

Another approach to solve RP is to use a so called de/composition of the ordinary PN. Given a PN *N = (P, T, pre, post, m$_0$)* we are looking for some method how to split *N* into two subnets $N_i = (P_i, T_i, pre_i, post_i, m_{0i})$ i = 1, 2 such that there is clear relation between languages of Petri nets L(*N*), L($N_i$), reachable states $\Re(N), \Re(N_i)$, fsa of type $M_\omega$ *M*, $M_i$ respectively and also there would be splendid if

$$q \in \Re(N) \Leftrightarrow q_1' \in \Re(N_1) \wedge q_2' \in \Re(N_2) \qquad (2)$$

provided that $q_1', q_2'$ are corresponding images of *q* under the operation that we are looking for. The operations that reliably do this are called T-JUNCTION, P-JUNCTION and PT-JUNCTION [4].

## IV. Time Reachability Analysis of TB Nets

In this section we show some crucial problems connected with the reachability analysis of TB nets.

### A. Loop-spectral Time Reachability Analysis

In fsa $M_\omega$ we can distinguish two subclasses of loops: *selffeeded* and *stabile*. The loop is selffeeded if in *t*-firing (*t* belongs to the loop) *t* "consumes" only tokens that were created solely by firings of loop's transitions. We call a loop stabile if at any *t*-firing (*t* belongs to the loop) all tokens at precondition places are uniformly generated, i.e. at any *t*-firing at each repetition *t* consumes tokens from the same generators, i.e. transitions that generates tokens consumed by *t*.

After defining initial TIP for any loop on a path we can define the TIs of any chronos (token) in which it can exist in the future.

To demonstrate how are loops helpful to solve RP problem we will use the example of the voice station (Fig. 1). Firstly we introduce some relation lately used in the example.

Let $p_1 = \tau_1 = 18$, $p_2 = \tau_2 = 18$, $p_3 = \tau_3 = 10$, $p_4 = \tau_4 = 18$, $p_8 = \tau_8 = 10$;

TABLE I
SOME RELATIONS DEFINED FOR TI VARIABLES

| Representation: Graphical | Representation: Symbolic | Formal characterization |
|---|---|---|
| { [ } ] | $\tau <_{sq} \tau'$ | $(\tau i < \tau i') \wedge (\tau i' < \tau a < \tau a')$ |
| { } [ ] | $\tau \le \tau'$ | $(\tau a < \tau i')$ |

$$tf_{t_1}(p_1) = (\tau = p_1 + 10) \tag{3}$$

$$tf_{t_2}(p_2, p_6) = (\tau = p_2 + 10 \wedge \tau \ge p_6) \tag{4}$$

$$tf_{t_3}(p_3, p_4) = (\tau = \max(p_3, p_4) + 0.2 \wedge \tau < p_3 + 10) \tag{5}$$

$$tf_{t_4}(p_2, p_3) = (\tau = p_3 + 10 \wedge p_2 = p_3) \tag{6}$$

$$tf_{t_5}(p_2, p_6, p_7) = (\tau = p_7 + 0.1 \wedge \tau \ge p_2 \wedge \tau \ge p_6) \tag{7}$$

$$tf_{t_6}(p_7) = (\tau = p_7 + 0.15) \tag{8}$$

$$tf_{t_7}(p_8) = (p_8 \le \tau \le p_8 + 8) \tag{9}$$

In TABLE I are depicted two relations between TIs $\tau = \{\tau i, \tau a\}$ and $\tau' = \{\tau i', \tau a'\}$ shown as curly and square brackets, respectively.

We have to explore the following: chain approach to estimate interval limits is well founded; selffeedable loop $\sigma_t$ guaranties the stabile regime, i.e. if $\sigma_t = tt_1t_2...t_r$ selffeedable loop then

$$\tau_t^{(n)} = [\tau en_\alpha i + n \cdot \tau eni(tf_{\sigma_t}(0)), \tau en_\alpha a + n \cdot \tau ena(tf_{\sigma_t}(0))] \tag{10}$$

where $\tau eni(tf_{\sigma_t}(0))$ expresses the contribution to the left limit of $\tau en$ by $\sigma_t$ loop. The same meaning wrt the right limit of $\tau en$ is attached to the $\tau ena(tf_{\sigma_t}(0))$ and $tf_{\sigma_t}(0) = tf_t(0) + tf_{t_1}(0) + ... + tf_{t_r}(0)$.

Let us compute TI $\tau_4$ that can be $t_2$ or $t_5$ generated. We have

$$\tau_4 = \tau en(\tau_2, \tau_6, \tau_7) + tf_{t_5}(0) = [\tau_6 i, \tau_7 a + 0.1] \tag{11}$$

$$\tau en(\tau_2, \tau_6, \tau_7) = [\tau_6 i, \tau_7 a]$$

$$tf_{t_5} = \overrightarrow{0.1}$$

TI $\tau en$ will be determined (in the case of $t_5$ generation of $\tau_4$) $\tau_6 i$ and $\tau_7 a$ as far as the lower and upper bound of $\tau en$ respectively is concerned.

We are able to create a chain of transitions leading to creation of $\tau_6$. Then chain will be $\{t_1t_3\}$. To generate $\tau_7$ $t_7$ has to be fired so the chain for $\tau_7$ is $\{t_7\}$.

The loop $\sigma_t = t_7t_1t_3t_5$ has to be executed repeatedly in the stable regime to make creation of $\tau_6$ and $\tau_7$ respectively. From $\sigma_t$ only the part $t_1t_3t_5$ can be considered to create the lower bound of $\tau en$ of t5 and the part $t_5t_7$ is the only chain usable to create the upper bound of $\tau en$ of $t_5$. The part $t_1t_3t_5$ will be denoted as $iL(t_5)$ and the part $t_5t_7$ as $aL(t_5)$.

Let $\tau en_i$ be the initial value of the determinate TI of $t_i$ prior to its first firing as the member of the stable loop σ, i.e.

$$\tau_{t_i}^{(1)} = \tau en_i + tf_{t_i}(0) \tag{12}$$

We denote by $\tau_{t_i}^j$ $t_i$ generated TI after j-th firing of $t_i$ in the stable loop $\sigma$ and

$$\tau_{t_i}^{(j)} = \tau en_{t_i}^{(j-1)} + tf_{t_i}(0) \tag{13}$$

$$\tau_{t_i}^{(j)} \prec \tau_{t_i}^{(j+1)}, \tau en_i^{(j)} \prec \tau en_i^{(j+1)} \text{ where } \prec \in \{\le_{sq}, \le\}.$$

Lets have $\tau en_t = [\tau_{t'} i, \tau_{t''} a]$. To calculate $\tau en_t^{(j+1)}$ in the case of the stable loop, we have to consider the contribution of $\sigma$ to the lower and upper bound of $\tau en$. The contribution to the lower and upper bound is denoted by

$$tf_{iL(t)} \text{ and } tf_{aL(t)}. \tag{14}$$

For $t_5$ generated $\tau_4$ in our example we have

$$\tau_4^{(n+1)} = \tau_{t_5}^{(n+1)} = \tau en_{t_5}^{(n+1)} + tf_{t_5}(0) \tag{15}$$

$$\tau en_{t_5}^{(n+1)} = [\tau_6^{(n)} i, \tau_7^{(n)} a].$$

We start with calculation of initial values for $\sigma = u\sigma\sigma...$ provided $\sigma = 1375$ and $u = 3726$

$$\tau_{t_3} = tf_{t_3}(0) + \tau en_3 = 0.2 + \max(10,18) \tag{16}$$

$$= 0.2 + 18 = 18.2$$

$$\tau'_{t_7} = tf_{t_7}(0) + \tau en_7 = \vec{8} + 10 = [10,18] \tag{17}$$

$$\tau_{t_2} = tf_{t_2}(0) + \tau en_2 = 10 + [\max(\tau_6 i - 10, \tau_2 i), \tau_2 a] \tag{18}$$

$$= 10 + [\max(18.2 - 10, 18), 18] = 10 + [18,18] = 28$$

$$\tau_{t_6} = tf_{t_6}(0) + \tau en_6 = 0.15 + \tau_{t_7} \tag{19}$$

$$= 0.15 + [10,18] = [10.15, 18.15]$$

$$\tau''_{t_7} = tf_{t_7}(0) + \tau en_7 = \vec{8} + [10.15, 18.15] = [10.15, 26.15] \tag{20}$$

$$\tau_{t_5} = tf_{t_5}(0) + \tau en_5 = \overrightarrow{0.1} + [\max(\tau_6 i, \tau_2 i, \tau_7 i + 0.1), \tau_7 a] \tag{21}$$

$$= \overrightarrow{0.1} + [\max(18.2, 18, 10.25), 26.15] = [18.2, 26.25].$$

TABLE II
TRANSITION TIME INTERVALS

| Transition time | Lower bound (i) | Upper bound (a) |
|---|---|---|
| $tf_{t_1}(0) = 10$ | 10 | 10 |
| $tf_{t_2}(0) = 10$ | 10 | 10 |
| $tf_{t_3}(0) = \overrightarrow{0.2}$ | 0 | 0.2 |
| $tf_{t_5}(0) = \overrightarrow{0.1}$ | 0 | 0.1 |
| $tf_{t_7}(0) = \vec{8}$ | 0 | 8 |
| $tf_{t_6}(0) = 0.15$ | 0.15 | 0.15 |

Now we calculate

$$tf_{iL(t_5)}(0) = (tf_{t_1}(0))i + (tf_{t_3}(0))i + (tf_{t_5}(0))i = 10 + 0 + 0 \quad (22)$$

$$tf_{aL(t_5)}(0) = (tf_{t_7}(0))a + (tf_{t_5}(0))a = 8 + 0.1 = 8.1 \quad (23)$$

The following formula calculates the value of $\tau_4$

$$\tau_4 = [18.2 + n \cdot tf_{iL(t_5)}(0), 26.25 + n \cdot tf_{aL(t_5)}(0)] \quad (24)$$

$$\tau_4 = 18 + [0.2 + n \cdot 10, 8.25 + n \cdot 8.1].$$

The expression for $\tau_4$ has a very interesting feature. After each iteration of $\sigma_t = t_7t_1t_3t_5$ the lower bound of TI $\tau_4$ will be increased by 10, while the upper bound just by 8.1. This mean that after a few iteration $\tau_4$ becomes a dummy interval with lower bound grater then the upper bound.

### *B. Problems connected with time reachability analysis*

As we stated in the previous section, reachability analysis can be made by constructing a fsa of type $M_\omega$ and using Modified Integer Programming. The main problem for PNs that are dealing with time is, how to construct $M_\omega$ automaton that will keep the information about the time associated with every token?

## V. Conclusion

As it was shown the selffeeded and stabile loops play profound role in the reachability analysis of TB nets. This approach reveals the behavior of any token and discovers perhaps a moment when it disappears because of the emptiness or dummy feature of its TI. For any TI $\tau_t$ (*t* generated TI) we can construct a formula for the TI to be calculated.

In [5] we can exactly see how can be process algebra connected with PNs. This connection revealed a nice way for inspecting the reachable markings in ordinary PN. The question is, whether it is possible to use process algebra also on the TB nets and inspect reachable markings? Our future work will be focused to get answers to the questions raised above.

## References


[1] C. Ghezzi, D. Mandrioli, S. Morasca, M. Pezzé,"A unified high-level Petri net formalism for time critical systems (Book style with paper title and editor)," IEEE Transition on Software Engineering, 2nd ed. vol. SE-17, 1991, pp. 10-150.

[2] M. Felder, C. Ghezzi, M. Pezzé, *Analyzing refinements of state based specification: the case of TB nets*. New York: Springer-Verlag, 1994, ch. 7.

[3] Š. Hudák, Time Interval Semantics of TB Nets, Proceeding of international conference RSEE'96, Romania, 1996, pp. 1-12.

[4] Š. Hudák,"Reachability Analysis of Systems Based on Petri Nets (Book style)," Košice, elfa s.r.o., 1999, pp. 47-180.

[5] Š. Hudák, S. Šimoňák, Š. Korečko, A. N.Kovács, Formal Specification and de/composition approach to design and analysis of discrete systems, Košice, elfa, 2007. pp. 8-46.

# Connection Admission in Mobile Ad-Hoc Networks: A Brief Survey

*Jozef RATICA*

Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

jozef.ratica@tuke.sk

*Abstract*—**Wireless devices are becoming very popular because of their ability to provide mobile communication, which can be represented by Mobile Ad-hoc Networks (MANETs). MANETs are kind of self-configuring wireless networks, which do not need any centralized coordination. However, due to changeable properties of MANETs Quality of Services (QoS) vary and therefore Connection Admission Control (CAC) algorithms are applied. In this paper will be presented CAC overview based on which future work aims will be established.**

*Keywords* — **CAC, MANET, QoS, throughput.**

## I. Introduction

Wireless devices are becoming prevalent because of their ability to provide mobile communication (voice or multimedia). Mobile Ad-hoc Networks (MANETs) [1] represent one form of mobile wireless communication, which utilize IEEE 802.11 wireless technology [2]. However, in this type of communication we run into trouble, which didn't exist in wired counterparts. Examples of these troubles are: self-interference due to multipath, lower bandwidth (transmission rate) availability and hence limited network throughput, higher jitter, delay, limited service coverage, higher data loss rates due to interference, frequent disconnection caused by node mobility, etc.

Mobile ad-hoc networks are kind of networks where particular nodes arbitrarily configure network by themselves. Mobile terminals move freely and randomly hence network topology changes rapidly and unpredictably. Due to these properties Quality of Services (QoS) of MANETs change frequently. In order to MANETs could be applied for real-time applications, they have to meet hard constraints for such type of services. Real-time applications require low packet loss and delay/jitter in the presence of mobility. These requirements become much more important when network load unduly grows and achieves congestion point. Wireless medium have to be kept from reaching congestion point, if QoS requirements are to be met.

For the reasons mentioned above there are Connection Admission Control (CAC) algorithms applied to ensure that particular connections didn't lose their QoS requirements. That means, no real-time connection will suffer from insufficient bandwidth for its data-flow because there will be admitted as much real-time connections as it can be to meet QoS requirements of particular real-time connections.

CAC algorithms are closely associated with routing algorithms, which are needed for connections longer than one hop. Most frequently used routing algorithms are Dynamic Source Routing (DSR) [3] and Ad Hoc On-Demand Distance Vector (AODV) [4].

The rest of the paper will be organized by follows: in section II. will be resumed IEEE 802.11 MAC technology which is utilized in MANET. Section III. will address CAC algorithms and finally section IV. concludes this paper with future work goals.

## II. IEEE 802.11 MAC for Ad-Hoc Mode

IEEE 802.11 standard specifies both MAC layer and Physical Layer (PHY). The MAC layer offers two different types of service: a contention service provided by Distributed Coordination Function (DCF), and a contention-free service implemented by Point Coordination Function (PCF). DCF provides basic access method of the 802.11 MAC protocol and is based on a Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) scheme. PCF is implemented on top of DCF and is based on a polling scheme. It uses a Point Coordinator that cyclically polls stations, giving them the opportunity to transmit. PCF cannot be adopted in the ad hoc mode.

According to DCF, if medium is found to be idle for an interval longer than Distributed Inter-Frame Space (DIFS) before transmitting a data frame, the station continues with its transmission. On the other hand, transmission is deferred until the end of the ongoing transmission. A random interval - backoff time - is then selected. The backoff timer is decreased for as long as channel is sensed as idle, stopped when a transmission is detected on the channel, and reactivated when channel is sensed as idle again for more than a DIFS. Station is enabled to transmit its frame when backoff timer reaches zero. The backoff time is slotted and it is an integer number of slots uniformly chosen in the interval (0, CW-1). CW is defined as the Backoff Window, also referred to as the Contention Window (CW). At the first transmission attempt, $CW = CW_{min}$, and it is doubled at each retransmission up to $CW_{max}$. Values of the CW depend on employed PHY.

Because of collision that may occur, there is an immediate positive acknowledgement (ACK) used. Upon reception of a data frame, destination station initiates transmission of acknowledgement frame (ACK) after time interval called Short InterFrame Space (SIFS). SIFS is shorter than DIFS in order to give higher priority to receiving station over other possible stations waiting for transmission. If ACK is not received by source station, data frame is presumed to have been lost, and retransmission is scheduled. ACK is not transmitted if the received packet is corrupted.

There is also Request To Send (RTS) and Clear To Send (CTS) control frames used due to alleviate hidden and exposed terminal problem (on transmission level). Since RTS/CTS are control frames they also have higher priority and therefore there must be SIFS used in their transmission. To guarantee fair access to shared medium for all stations one that has just transmitted a packet and has another packet ready for transmission must perform the backoff procedure before initiating the second transmission [1][2].

## III. Connection Admission Control Algorithm Overview

### A. CAC background

CAC algorithms could be divided according to QoS parameters, however due to QoS parameters assurance, available sources is to be known. Hence, this part of paper will address CAC algorithms based on available sources measurement. Connection admission is mostly based on available sources measurement used and still available for new user. There is one key question. How to measure available sources? There are several ways how to measure network sources but most common way is to measure utilization, $U$. Let us assume that $U$ is network utilization and $B_{max}$ is maximum available bandwidth then available bandwidth $B_{avail}$ is estimated by follows:

$$B_{avail} = (1 - U) \times B_{\max} \tag{1}$$

where $0 \le U \le 1$. There are many techniques to measure the network utilization, for example:

- MAC Layer Congestion Window
- Queue Length
- Number of collisions
- Delay
- Channel Busy Time

The first three methods provide little or no information regarding network utilization if a node is not actively transmitting packets. For example, a collision only occurs if a packet fails to send. If a node does not send any packets, it cannot determine the current state of the channel. The same holds true for the MAC layer congestion window and the queue length. Since these techniques are not adequate for determining the available bandwidth, there must be used delay or channel busy time measurement method [5].

### B. CAC based on delay measurement

There are some CAC algorithms, which measure delay experienced on MAC layer. Based on this measurement there were proposed different CAC algorithms. In [6], there was proposed CAC algorithm, which compares results from VS (Virtual Source) and VMAC (Virtual MAC) mechanisms with service requirements. Based on this comparison is new connection admitted or rejected. VS and VMAC make use of delay measurements. Other CAC algorithm [7] exploits service curve, which reflect status of network. Service curve is based on total delay measurement, which occurs by exploring packet transmissions. New connection is granted if the service curve is bounded below by some non-decreasing deterministic function, which is called the universal service curve. Universal service curve acts as a worst-case reference curve and it is also boundary for QoS support.

### C. CAC based on busy time measurement

Channel busy time is a direct measure of the channel utilization. In wireless networks, carrier sensing enables nodes to detect four states: transmitting (Tx), receiving (Rx) busy or carrier sensing (CS) and idle. By measuring the amount of time when channel is in first three states we get channel busy time, i.e. busy time is the total time within an interval that a node is transmitting, receiving or sensing packet transmissions.

The transmission of each data packet in IEEE 802.11 MAC takes following time:

$$\begin{aligned} T_{busy} &= T_{rts} + T_{cts} + \frac{L}{B} + T_{PLCP} + T_{ack} + \\ &\quad + T_{difs} + 3T_{sifs} + T_{backoff} \end{aligned} \tag{2}$$

where $L$ is the size of the data packet including MAC header. $B$ denotes link rate used by the source node. $T_{PLCP}$, $T_{rts}$, $T_{cts}$, and $T_{ack}$ represent the time for transmitting PLCP header, RTS, CTS, and ACK packets, respectively. $T_{sifs}$ and $T_{difs}$ denote the inter-frame space SIFS and DIFS and $T_{backoff}$ is backoff time applied before transmission [8].

**Call admission and rate control (CARC) algorithm** [9] is CAC algorithm, which utilize CAC for real-time (RT) and streaming flows and Rate Control (RC) for non real-time flows, also called Best Effort (BE). Channel capacity is partitioned for RT flows (80%) and BE flows (20%). RT flows with variable bit-rate can lower channel capacity of BE flows due to QoS assurance. CAC employ three parameters $TR$, $TR_{peak}$, $len$, where $TR$ is mean rate, $TR_{peak}$ is peak rate (both in bit/s) and $len$ is mean packet length. For Constant Bit Rate (CBR) flow is $TR_{peak} = TR$. We can compute channel utilization by follows:

$$U = \frac{TR}{len} \times T_{suc} \tag{3}$$

where $T_{suc}$ is mean time needed for successful packet receiving. Total bandwidth occupied by all admitted flows is recorded in to parameters, $U_{aggr}$ and $U_{peak_{aggr}}$, which are continually updated. For new connection admission, there must be conditions in (4) and (5) executed.

$$U_{aggr} + U < B_M \tag{4}$$

$$U_{peak_{aggr}} + U_{peak} < B_U \tag{5}$$

where $B_M$ is 80% of $B_U$. $B_U$ is maximal channel utilization. Traffic rate of BE flows, $TR_b$, is given by (6):

$$TR_{b_{new}} = TR_{b_{old}} \times \frac{R_{bt} - R_{br}}{R_b - R_{br}} \tag{6}$$

where $TR_{b_{new}}$ and $TR_{b_{old}}$ are values of $TR_b$ after and before the adjustment, and $R_{bt}$ is a threshold of channel busyness ratio, $R_b$, and is set to 95% × $B_U$ . $R_{br}$ is RT flow contribution to $R_b$. Node increases $TR_b$ if $R_b < R_{bt}$ and decreases it otherwise.

**Contention-aware admission control (CACP)** [10] is another CAC algorithm, which at the admission decisions execution consider not only local available resources but also neighboring available resources (i.e. in CSR). It utilize (2) for local available bandwidth measurement, $B_{local}$, by following equation:

$$B_{local} = \frac{T_{idle}}{T_p} B_{channel} \tag{7}$$

$$T_{idle} = 1 - T_{busy} \tag{8}$$

where $T_p$ is measurement interval and $B_{channel}$ is channel capacity. Same equation is used for neighboring available bandwidth, $B_{nbh}$ , however there is applied $T_{idle}^{neighbor}$ , idleness on Carrier-Sensing Neighbors (CSN). CSN are neighboring nodes in Carrier-Sensing Range (CSR). CACP uses on-demand route discovery with source routing, similar to DSR.

Admission is granted if two conditions are met. First, $B_{local}$ must be greater than application bandwidth requirement. Second, $B_{nbh}$ must be greater than $B_{req}$, which is given by (9).

$$B_{req} = CC \times R \times T_{busy} \times B_{channel} \tag{9}$$

$R$ is number of packets generated by application in one second and $CC$ is contention count. CC at a particular node is the number of nodes on the multihop path that are located within carrier-sensing range of the given node.

In [11], there is CAC algorithm very similar to CACP, but it furthermore considers time wasted at MAC layer by collisions and channel access delays.

**Perceptive admission control (PAC)**; the core idea of PAC [12] is to allow nodes to depend on their own estimation of the available bandwidth to make correct admission decisions. The communication range of the available bandwidth measurement will change so that all nodes will be able to make admission decision without communication with any other nodes. CSR will change to 2 × $L$ + $RID$ (minimum distance required for two simultaneous transmission to operate without collision) and therefore each node can itself make admission control decisions and it only need to check its locally available bandwidth. $RID$ (Receiver Interference Distance) is the distance between a receiver (Rx) and another sender ($A_{TX}$), such that the receiver can successfully receive Tx's packets and $A_{TX}$ can simultaneously send a packet to another receiver (Fig.1). $L$ is the distance between receiver-sender pair.

PAC uses small portion of bandwidth (reserved bandwidth) to prevent the channel before congestion. To admit a new flow, it must be executed following condition:

$$B_{local} - B_{rsv} > B_{rq} \tag{10}$$

where $B_{rq}$ is bandwidth required for new data flow, $B_{rsv}$ is reserved bandwidth mentioned above and $B_{local}$ is amount of local available bandwidth, computed according to (7).

The reception power threshold, propagation model and capture factor must be known to determine RID. The capture factor defines the minimum power ratio between the received power of two packets such that the packet with the higher power can be successfully received.

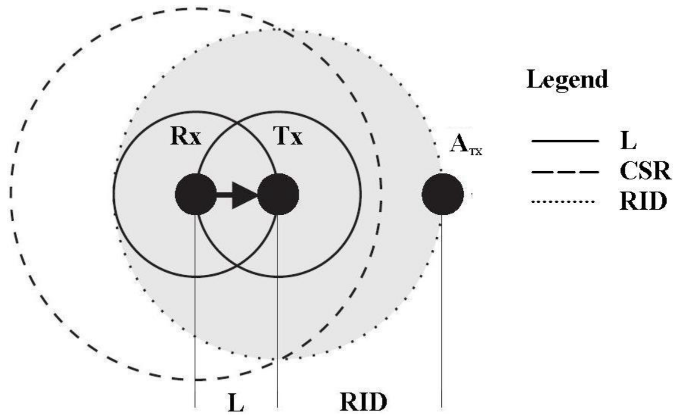



Fig. 1 Example of communication ranges: RID, CSR

**QoS and CAC for 802.11 Ad Hoc Networks.** CAC mechanism in this QoS scheme [8] consist of link utilization prediction, channel availability estimation and routing protocol extension.

Link utilization prediction, $U$, go out from following equation:

$$U = \sum_{i=1}^{N_{count}+1} R \times \left( \frac{L}{B_i} + T_{oh} \right), \tag{11}$$

where $R$ is packet transmission rate, $L$ is data packet length including MAC header, $B_i$ is link bitrate between two nodes, $T_{oh}$ includes time needed to RTS, CTS, ACK packets, PLCP header transmission and also backoff time. $N_{count}$ is number of nodes in CSR except destination node.

Channel availability estimation relate with carrier sensing monitoring. Monitoring level was lowered to NCSR (Neighbor CSR). NCSR covers all CSN nodes within CSR of transmitting node. CSN nodes are neighboring nodes located in CSR (Fig. 2).

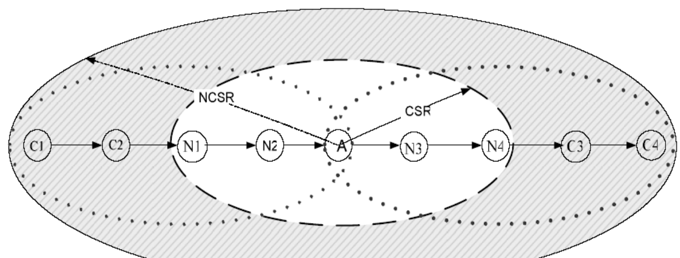



Fig.2 Example of neighbor-carrier sensing range NCSR and carrier sensing range CSR.

The novelty of CAC in [8] is parallel transmission influence consideration. Transmissions outside CSR also influence channel capacity of particular transmitter. This impact is expressed in (12) and it is parallel transmission contribution:

$$U_{par} = \frac{T_B^{CSN} - T_B^{local}}{T_p} \times R. \left( T_{oh} + \frac{L}{B_a} \right), \tag{12}$$

where $R \cdot (T_{oh} + L/B_a)$ represent transmission of particular sender and $B_a$ is the link rate at this sender. $T_B^{CSN}$ is busy time measured in NCSR and $T_B^{local}$ is busy time measured in CSR. For new data flow admission must be executed condition:

$$U_{aggr} \leq 1 - U_{CSN} + U_{par} \tag{13}$$

where $U_{aggr}$ represents aggregate link utilization required by flow. $U_{CSN}$ is channel utilization at nodes in NCSR and is given by follows:

$$U_{CSN} = \frac{T_B^{CSN}}{T_p} \tag{14}$$

### D. CAC based on other type of measurement

Another way how to measure network utilization is to measure throughput. In [13], authors measure throughput based on following equation:

$$Th = \frac{S}{t_r\text{-}t_s}, \tag{15}$$

where $S$ represent packet length, $t_s$ is time-stamp at which packet is ready at MAC layer and $t_r$ is time-stamp at which ACK was received. In time interval, $t_r$-$t_s$, have to be busy time and also contention time included. Throughput estimation is executed for each node independently. It is important in this algorithm in order to throughput measurement was made independently from packet length. This can be done by packet length normalization.

In [14], CAC utilize two techniques. First one is noise level measurement on active nodes. Second one is sufficient bandwidth allocation. The interference is not handled uniformly but it is classified based on position estimation of interfering nodes. One of innovation in this algorithm is division of time when node is in CS state. Carrier sensing was divided into two states:

- BSB – Bandwidth considering Sensing as Busy
- BSI – Bandwidth considering Sensing as Idle

BSB state includes busy period measurement for transmitting, receiving and carrier sensing, while in BSI state busy period includes transmitting and receiving, but carrier sensing is considered as idle. Available bandwidth is somewhere between BSB and BSI.

Based on BSB, BSI and noise level on receiver (NR) or transmitter (NS), position estimation of interfering transmitters was divided into three categories:

1. all interfering senders are located within Tx range of transmitter S
2. all interfering senders are located beyond Tx range of transmitter S, but within its CS range:
   a) all interfering senders are located within Tx range of receiver R and outside the transmission range of S
   b) some interfering senders are located beyond Tx range of receiver R
3. Interfering receivers are beyond and within Tx range of transmitter S

For each category is available bandwidth computable based on parameters gathered based on BSB, BSI and node locations.

## IV. Conclusion

In this paper was introduced connection admission control problem. In part III was analyzed different type of CAC algorithms. Based on this analyze I want to suggest new CAC algorithm, what is topic of my PhD. thesis, which will get better results than previous one. There are still some problems, which was not investigated. For example, how will be measured results changed if we try to use RID instead of CSR in [14]. The available resources measurement in CSR is challenging problem, due to exposed terminal problems. I prefer available results measurement as it was listed in III.C. Therefore I want to lead my investigation in this direction.


## Acknowledgment

Research described in the paper was financially supported by VEGA No.1/4054/07, and also by COST 2100 - Pervasive Mobile & Ambient Wireless Communications.


## References


[1] S. Basagni, M. Conti, S. Giordano, I. Stojmenović: *Mobile Ad-Hoc Networking*. Wiley-IEEE Press, August 2004, 480s. ISBN: 978-0-471-37313-1

[2] IEEE Draft International Standards, *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, ISO/IEC 8802-11, IEEE P802.11/D10, Jan.1999.

[3] D.B. Johnson, D.A. Maltz: Dynamic Source Routing in Ad Hoc Wireless Networks. Mobile Computing, vol. 353, 1996.

[4] C. Perkins: Ad-Hoc On-Demand Distance Vector Routing. Proc. Military Comm. Conf., 1997.

[5] I. D. Chakeres, E.M. Belding-Royer: *PAC: perceptive admission control for mobile wireless networks*. Proceedings of the 1st. international conference on Quality of service in heterogeneous wired/wireless networks, QSHINE 2004, Oct. 2004, pp. 18-26.

[6] M. Barry, A. T. Campbell, A. Veres: *Distributed control algorithms for service differentiation in wireless packet networks,* Proceedings of IEEE INFOCOM 2001. Vol. 1, pp. 582-590.

[7] S. Valaee, B. Li: *Distributed call admission control for ad hoc networks,* Proceedings of IEEE Vehicular Technology Conference, (VTC 2002-Fall.) 2002, Vol.2, pp. 1244 – 1248.

[8] L. Luo, M. Gruteser, H. Liu, D. Raychaudhuri, K. Huang, S. Chen: *A QoS routing and admission control scheme for 802.11 ad hoc networks*. International Conference on Mobile Computing and Networking, Proceedings of the 2006 workshop on Dependability issues in wireless ad hoc networks and sensor networks, Los Angeles, CA, USA, pp. 19 – 28.

[9] H. Zhai, X. Chen, Y. Fang: *A call admission and rate control scheme for multimedia support over IEEE 802.11 wireless LANs*. Wireless Networks, Vol. 12, July 2006, pp. 451-463.

[10] Y. Yang, R. Kravets: *Contention-Aware Admission Control for Ad Hoc Networks*. IEEE Transaction on mobile computing, vol.4, No.4, July/August 2005, pp. 363-377

[11] L. Hanzo, R. Tafazolli: *Quality of Service Routing and Admission Control for Mobile Ad-hoc Networks with a Contention-based MAC Layer*. Proceedings of IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS), Oct. 2006, pp. 501-504.

[12] I. D. Chakeres, E. M. Belding-Royer: *PAC: perceptive admission control for mobile wireless networks*. Proceedings of the 1st. international conference on Quality of service in heterogeneous wired/wireless networks, QSHINE 2004, Oct. 2004, pp. 18-26

[13] S. H. Shah, K. Chen, K. Nahrstedt: *Dynamic Bandwidth Management for Single-Hop Ad Hoc Wireless Networks.* Proceedings of IEEE PerCom, 2003, pp. 195-203

[14] K. N. Sridhar, Chan Mun Choon: *Interference based Call Admission Control for Wireless Ad Hoc Networks.* Mobile and Ubiquitous Systems - Workshops, 2006. 3rd Annual International Conference on**,** 17-21 July 2006 Nat. Univ. of Singapore, pp. 1-10.

# MANIAC: Mobile Ad-Hoc Networks Interoperability and Cooperation: THE LIVE AND LET LIVE STRATEGY

*Ivan KLIMEK, Vladimír SIDIMÁK*

Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

ivan.klimek@cnl.tuke.sk, vladimir.sidimak@cnl.tuke.sk

*Abstract*—**Mobile Ad-Hoc networks are certainly the most flexible computer networks that currently exist. They can be built anytime, anywhere, using any wireless-enabled devices and provide end-to-end connectivity. The benefits of such a decentralized technology are quite clear. That is why there is a vast number of interesting implementations. But why are MANETs not used in everyday life then? This paper describes a solution that combines the strength and simplicity of existing implementations of the Optimized Link State Routing Protocol (OLSR) with a strategy originally presented and awarded with the Strategy Award at the MANIAC Challenge 2007 (Mobile Ad Hoc Networks Interoperability and Cooperation) held in conjunction with IEEE Globecom 2007. This strategy aims to solve the two most important question that hinder MANETs from real-life use, first the minimization of traffic on nodes with limited battery capacity (e.g. mobile phones) and second the detection and isolation of uncooperative nodes while keeping the algorithm as simple and fast as possible so it could be run practically on any device without large overhead. Simplicity was the key factor during the whole developing process that should guarantee easy portability and maximal compatibility.**

*Keywords*—**MANET, OLSR, wireless, network**

## I. INTRODUCTION

Mobile wireless devices provide a comfortable way of communication for their ease of use and mobility. As a price for this comfort, there is a limiting factor, the battery life. Wireless transfers of information on these devices usually result in huge energy consumption, thus rapidly decrease the time that devices can be used without recharging. Any group of wireless enabled devices should be able to create a MANET, in the most cases a full-mesh topology is used, which is too demanding as for already mentioned power consumption concerns. Presented approach minimizes that by reducing the number of active connections between nodes by a "minimal active neighbor topology calculation" algorithm. Another problem that makes the use of a MANET less attractive, are uncooperative nodes. Those are the nodes that try to use the network for their own data transfers but do not forward traffic destined to, or originating from other nodes. A neighbor forwarding traffic sensing mechanism is applied to check if the neighbors are really forwarding the data as they should be. The strategy is called Live and Let Live.

## II. THE LET AND LIVE STRATEGY

### A. Minimal active neighbor topology calculation

The general idea is that no node in any network really wants to forward the data of other nodes at its own expenses. If every element would treat others this way, the network would not work. On the other hand, every node needs to communicate. That is why it became part of the network. As MANETs are decentralized, in order to keep the network working, some nodes have to forward foreign data too. Our minimal active neighbor topology algorithm reduces the number of communication links between the nodes while keeping them all connected and being able to communicate with each other. The first step is to find the best next-hop for forwarding the local node's data (e.g. best next hop to the gateway). Then the direct connectivity to all other one-hop neighbors is temporarily terminated. After the changes in topology propagate, the local node checks if its one-hop neighbors are reachable thru the only enabled node, the best next hop. If not, one of the original neighbors is randomly chosen and enabled. The check for accessibility of other nodes is performed again. If they have any other way to access the network except of the local node, the local node will be able to hear of them thru one of the enabled nodes in their routing updates. This algorithm continues until all of the original nodes are accessible again. In each iteration the nodes that become reachable as a result of the actions during that iteration are deleted from the list of nodes that need to be checked. Algorithm is rerun whenever a topology change is detected. Topology changes are propagated extremely fast thanks to the OLSR MPR concept. The minimal neighbor topology calculation is only applicable to battery powered mobile devices. Other nodes with no power saving concern act as "center of topology", ideally a next-hop node for several mobile devices. Such approach creates a topology in which every battery powered device minimizes its number of connections and traffic load, but also acts as an entry point for all other nodes that have no other means of connecting to the network, keeping the network operational.

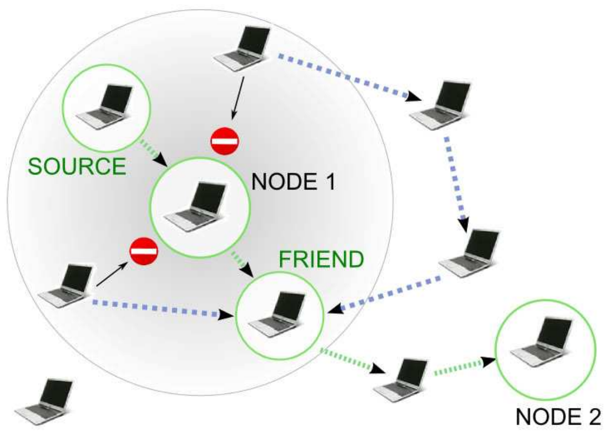



Fig. 1. An example MANET topology. Node 1 has chosen the next hop (FRIEND) for delivering the traffic to its destination (Node 2). Two other nodes that are in the range are blocked as there is an alternate route to the network for their traffic. All nodes keep connectivity to each other.

### B. Unicasting Multicasts and ARP filtering

To minimize the number of connections, the combination of ARP filtering and unicasting of the originally multicast OLSR update messages is used, so that the nodes are visible only to the nodes they want to be visible to and want to create links with them. By using this approach a logical topology inside the physical topology is created. Despite the fact that the nodes have mutual Layer 1 connectivity (e.g. they are in range of each other wireless adapter) they won't know of each other as of a direct neighbor unless the Minimal active neighbor topology calculation doesn't indicate otherwise.

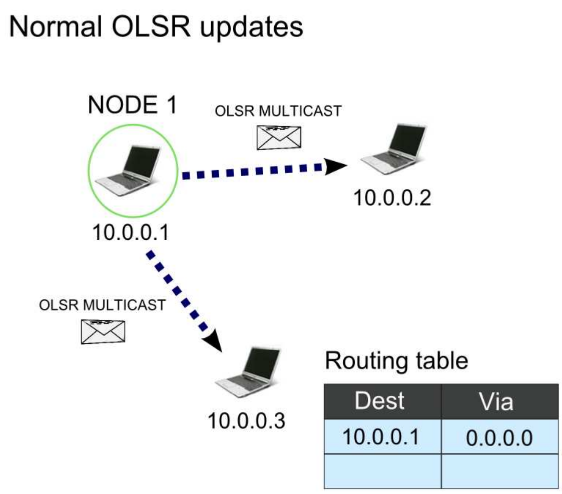



Fig. 2. OLSR updates as they are normally propagated to all nodes in range.

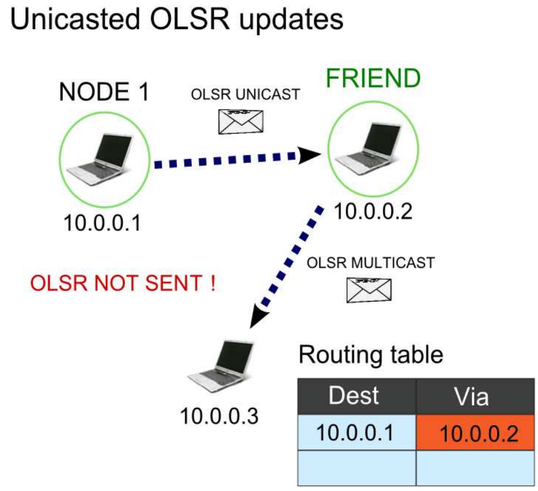



Fig. 3. OLSR updates unicast to chosen locations. Excluded nodes do not recognize NODE 1 as a neighbor node, even though it is in their range.

### C. Neighbor forwarding traffic sensing mechanism

There are cases when technology is not used exactly like how it was supposed to be used. In an MANET network, there might be nodes unwilling to participate on common interest, which is reliable network with end-to-end connectivity for all elements. These uncooperative nodes would accept the traffic destined to them, but not forward and drop other traffic in order to conserve their battery. To prevent these rogue nodes from exploiting the network, each node (node A) checks its neighbor whose data it is going to forward (node B) by simulating another node (node C). That node then tries to connect to node whose data the local node is forwarding /A is simulating node C that wants to connect to node B/. Because node A is the only possibility for node B to send its data to the network, node A has to be able to hear node C fake data. This can be done by generating fake traffic (node C traffic) with different source MAC and IP address than the original interface. To make this check mechanism even more realistic and harder to detect, the interface can be temporarily configured with different L1 parameters as for example signal strength. This would effectively avoid possible recognition that node A and node C are the same nodes. If node B fails in any stage of this check, node A simply does not forward node's B traffic because there is a reason to believe node B is a rogue node. Another possible scenario is that the attacker would become the OLSR MPR and would drop other nodes traffic. To avoid that, it is needed that each node monitors its neighbors and finds out how cooperative they are by seeing its own traffic being forwarded by the neighbor by listening for packets with the source MAC address of the forwarding next-hop, and the original source and destination IP addresses and/or other packet identifiers as packet length etc. At this stage of algorithm, the topology is already minimized and most of the traffic is forwarded through the one chosen next-hop node. That fact makes the monitoring process much easier. Uncooperative nodes should be blocked and not included in further topology calculations.

## III. Real Life Example

The following figure shows an example of real implementation of the strategy - limitation of the unnecessary traffic and keeping all nodes connected. PC1 and PC2 represent devices with no battery concerns, while NODEs 1 – 3 are mobile battery driven devices. NODE 2 does not need to establish the connection to NODE 1 as it can connect to the network via PC1, which is not limited by battery. On the other hand, NODE3 is fully dependant on NODE 2 and that is why NODE 2 provides the connection.

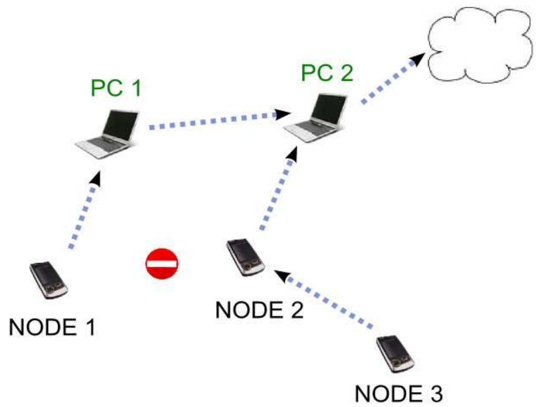


Fig. 4. Real life example for the Live and let live strategy

## IV. Conclusion

The presented solution is a simple MANET implementation based on the popular Optimized Link State Routing Protocol which is enhanced in a way that it can provide a fair-use environment for the network users without rogue nodes and optimizes the topology so that the traffic directed thru battery driven nodes is minimized. Parts of this solution are built using the MANIAC Challenge 2007 API, and are still under development. First practical tests with promising results were done during the MANIAC Challenge 2007, which finally led to the winning of the Strategy Award. Simulations and practical measurements are planned for the near future.

## References

[1] IETF, "OLSR Routing Protocol (RFC3626)," 2005, [Online; accessed 19-March-2008], [Online], Available: http://www.ietf.org/rfc/rfc3626.txt

[2] Wikipedia, "Optimized Link State Routing protocol," 2008, [Online; accessed 28-March-2008], [Online], Available: http://en.wikipedia.org/wiki/OLSR

[3] MANIAC Challenge, "MANIAC Challenge API," 2007, [Online; accessed 15-March-2008], [Online], Available: http://www.maniacchallenge.org/maniac_release01.tar

# Artificial intelligence in mobile robotics

[1]*Ing. Juraj EPERJEŠI*, [2]*Ing. Marek BUNDZEL PhD.*, [3]*Prof. Ing. Peter SINČÁK CSc.*
[1]Dept. of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic
[2]Dept. of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic
[3]Dept. of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic

[1]juraj.eperjesi@tuke.sk, [2]marek.bundzel@tuke.sk, [3]peter.sincak@tuke.sk

***Abstract*—This article offers review of several problems connected to the mobile robotics with possible solutions from artificial intelligence domain.**

***Keywords*—mobile robotics, artificial intelligence, control, design.**

## I. Introduction

Mobile robotics offers wide area of opportunities for artificial intelligence. From the design of robot itself, its shape, possible movements and also executable actions, thus on all levels of design, to the “life” of robot and its applying to real work, methods of artificial intelligence have potential to be used. This includes areas like optimization, control, scheduling or information collection and processing.

## II. Adaptivity

Robots can of course be controlled by manmade rules, or algorithms based on different areas of mathematics, for example statistics. Manually treat all possible error states is viable only in the simulated environment, or in environment with rigorously set conditions. But in real world such situations can emerge, with which are impossible to take into account beforehand. Robot can encounter such conditions which weren’t considered during design and development stages, or unpredictable failure can occur. Therefore the robot should be adaptive to be able to cope with such situations. Many methods of artificial intelligence offer acceptable results even in formerly unencountered situations.

In case that robot can be operated remotely, there is possibility that in such situations human operator intervene, or the control software can be reprogrammed. But if this possibility doesn’t exist and robot should be fully autonomous, the need for adaptivity rises again.

## III. Design

Robot’s design leads from its supposed application. If desired attributes and design are defined beforehand, robot is constructed based on them, but in case of prototype they must be specified. Fundamentally it is just optimization problem where typical tool of artificial intelligence are evolutionary computations, which search for optimal values of set of parameters. Each set is evaluated and its fitness is monitored. It reflects the degree of fulfillment of requirements placed on current task. If automated evaluation of fitness is not possible, methods of interactive evolution can be applied, where the task to evaluate sets of parameters is taken by the user. The result then corresponds with user’s requirements even in case, that they can’t be exactly specified or described.

Design can concern also movement of robot. In the simplest case, when robot’s movement is based on wheels this problem is nearly non existent, but in other cases is pretty essential. In area of legged robots, method of walk must be designed, which requires to describe movement of many joints. Such description can be based on observation of living animals, or human and measurement of their gait, or the results can be obtained through methods of artificial intelligence, for example before mentioned evolutionary computation. Such approach has one big merit, that this way acquired movement is made precisely for that robot and therefore it takes into consideration its construction characteristics, especially the power of its motors. Beside the wheeled or legged robots, some robots have serpentine movement, or are swimming, or diving. All that was said about the legged robots can be applied to these robots also, because their movement is mostly nature inspired.

## IV. Environment Perception

### A. Vision

If the robot had to be autonomous it must be able to perceive its environment. Either it has its own sensors, or gets the information about its position and environment from outside. But in both cases it’s necessary to obtain such information and process it for later use.

Easiest way of information gathering about obstacles, or objects on proximity of robot in general is scanning the surroundings of robots by one or more sonars, or infrared sensors, or by other similar sensors. Their problem is, that they scan the distance to the obstacle in just one direct line and between these lines the robot is, it can be said, blind. On the other hand they offer direct information of distance towards the obstacle, which is problem when using different methods. If the camera is used for gaining the information about the surroundings of robot, it scans the arc section of the whole visible environment. In contrast with sonars, camera can also which items it sees, so not only presence or absence of obstacle. But the object detection itself is big challenge, because it requires image processing, which means image segmentation and subsequent object recognition. Here it is also possible to use tools of artificial intelligence, for example artificial neural networks. Another problem is determining the distance to the objects. This problem can be solved by using two cameras, similar to the two eyes of human or many animals. This approach brings also new problems, where main is how to decide which point from first image corresponds with which point from second image.

### B. *Sound*

Beside visual information robot should react on and therefore sense also sound impulsions. This way it can avoid damage if this sound is coming from direction in which the robot isn't currently looking, or scanning and if such sound means danger, like loud noise. Another possible utilization is giving orders to robot simply by telling him. In this case also speech recognition is involved and this is also one of areas of interest for artificial intelligence. This way of giving orders is most natural for human and it allows giving orders without use of keyboard and therefore without some computer and robot itself can process such order.

## V. Control

### A. *Environment*

Robot control requires, that robot knows where in the environment it is and where it have to go, or what to do in current situation.

Information about position in the environment can robot gain from what it sees in the surroundings, or it gets it from outside viewer. From outside it can obtain the information if there is for example camera set above some space and image is processed and the robot gets only the information about its own position and orientation as well as position and orientation of obstacles. Another possibility is getting position information from GPS and the rotation can be determined from compass. If the robot have to get information about its position from what it sees, it must know where in the environment are which objects, so it must know how is the environment organized, which means it must have map of the environment of some kind.

Map of the environment must be first created. In most cases robot roams through the environment where it can and creates the map based on information gained from its sensors. Later if it is placed in the random place of this environment, it should scan the surroundings, compare it with the map and decide where it is. Problem can occur if the environment is maze of some kind and if all corridors look almost the same.

In each moment, or with some frequency the robot makes decisions what action it should perform next. The way it decides and the set of possible actions define the robots behaviour. Based on construction and robots employment is the amount and kind of these actions. In most situations it is just choosing next movement, but it also can be action like picking up the object, recognizing objects in proximity etc. In many cases two different rules of action can collide, like when robot knows that the objective of the movement is straight ahead, so it should move straight on, but there is also some obstacle. Then it must turn away from the objective, but it means, that it is not going to fulfill the goal. When the robot is for example in the maze, in many situations it would be forced to even move away from the objective due to go around the wall.

### B. *Controller*

Controller can be based on different methods of artificial intelligence, for example artificial neural networks, fuzzy systems, or reinforcement learning, or their combination. Each tool has its advantages and disadvantages.

In fuzzy systems expert must create membership functions, but then the system is fast and also the rules can be extracted and are in the form comprehensible by human.

Neural networks include many different learning algorithms and many topologies. Most common is network with backpropagation learning algorithm, when using this network we need learning set, to set the weights, but when the network is learned it reacts really quickly and precisely.

Reinforcement learning is approach which assigns some punishment or reward to the action performed in certain state of system, in this case, robot. So it is learning gradually.

## VI. Ability to learn incrementally

One of the most essential abilities, which autonomous robot should have is the ability to learn even during its activity and not only in the beginning, before it is put into use. Such robot would then increase its autonomy, because it would be able to learn to react to new situations which it didn't encountered before, or which were not taken into consideration during design process.

## VII. Groups of robots

Robots often work in groups, which brings in many advantages, but of course also new problems. Significant improvement is possibility to keep working on the task even in case when one or few robots gets damaged or destroyed. Another option is to work on several tasks at the same time, when each robot handles one of them.

Such groups are called multiagent. Important element of multiagent systems is communication between robots, or agents. This can run between robots directly, or through some central computer. Without the communication, cooperation between agents would be impossible. Local communication have sense mainly if two or more agents are cooperatively working on some smaller local task. In case that agents are working on a larger task, or when each of them is working on different task, but it is reasonable to save information about their activities for future utilization, or for learning, communication through central computer is more appropriate. This computer then accumulates all gained knowledge and in case of need offers them to individual agents.

Beside communication, different problems must be solved, like cooperation of collective movement of the whole formation. This problem is pretty well solved by group of algorithms from area of artificial life called boids. Here, based on few simple rules it is possible to keep the whole formation together and they ensure also obstacle avoidance.

## VIII. Information processing

Information, which robots collect during their activity needs to be processed. Whether it is just collecting or also sending information to the robots in multiagent system, or if it is also complex processing with the goal of getting new information, by for example generalizing, information processing is really important element of robotics.

## IX. Simulators

Simulators have considerable application in robotics, mainly in design stage. Thanks to them it is possible to test the robot construction before construction itself, or even for example design the movement. Mainly in the beginnings of the designing it is possible that certain order for movement can cause damage to robot. Possibly the robot is moving in such way, that it is falling constantly and it must be still picked up, which isn't problem in simulator. Another problem is when robot is powered from batteries, because they must be recharged, or changed which interrupts the work. In simulator this problem isn't present, but the virtual model shouldn't correspond exactly with the characteristics of real robot and then behaviour or movement designed in simulator, when applied to real robot can cause him damage or the robot will act differently as in simulator. Therefore the simulator can be well utilized in the beginnings of the design.

## X. Artificial intelligence

This whole effort leads of course to the creation of artificial intelligence, which would be able to work similarly as human mind, so it should be able to learn and use this new knowledge for its activity. Such intelligence would be able to really autonomously, maybe even creatively solve new problems without the need for human intervention.

## XI. Conclusion

This is only general overview where and which methods of artificial intelligence can be used. Many of these problems were already solved and many still wait for suitable solution.

My later work will involve robotic dog Aibo, which is quadruped with camera and also 2 IR sensors, therefore my interest is aimed in this direction.

## References

[1] D. Kortenkamp, R. P. Bonasso, R. Murphy, Artificial Intelligence and Mobile Robots, AAAI Press, Menlo Park, California, Cambridge, Massachusetts, London, England, 1998.

[2] J. L. Jones, D. Roth, Robot Programming, A Practical Guide to Behavior-Based Robotics, McGraw-Hill, 2004.

# Transformation of Functionality with Utilization of Metaprogramming and Reflection

[1]Michal FORGÁČ, [2]Michal VAGAČ

[1]Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic
[2] Dept. of Informatics, FPV UMB in Banská Bystrica, Slovak Republic

[1]michal.forgac@tuke.sk, [2]vagac@fpv.umb.sk

**Abstract— Effective software evolution needs to be supported by appropriate execution environment. Program can be viewed as a sequence of statements that are aimed to produce some result. The execution is done by a platform that interprets the program's sequence of statements. The new result of a computation can be achieved by transformation of program, interpreter or both. Software evolution as long-term process can be supported by adaptive language and by environment, which offers reflective possibilities. One of the candidates for this purpose should be Smalltalk-like environment Squeak, which offers several reflective possibilities.**

***Keywords*—software evolution, program transformation, interpreter transformation, metaprogramming, reflection.**

## I. Introduction

Modification of complex software systems after their delivery means difficult process. The most often reasons for such modification are on the one hand detected faults, which have to be fixed or on the other hand requirements for new functionality, which systems have to include (e.g. replacement of the system from one computational environment into new environment). Such modifications require additional costs. Even implementation of required changes takes in some cases longer than implementation of the first operational software version. Thus there is significant demand for reaching optimal methods in order to achieve effective modification of software systems.

Under modification we can mean software maintenance or software evolution. These terms are often used as synonyms. We incline rather to the claim (according to e.g. [19]), that these terms do not express similar meaning. The purpose of maintenance is mostly in removal of some faults, the purpose of evolution is adaptation of a system according to the new external or internal requirements.

This paper is structured as follows: section II presents information about software evolution. Basic distinction of transformation of interpreted functionality is presented in section III. Section IV deals with main principles of metaprogramming and reflection. Section V presents main reflective characteristics of Smalltalk environment with regard to its possible utilization in our research. Finally, section VI concludes this paper.

## II. Software Evolution

According to [20], software evolution is defined as a collection of all programming activities intended to generate a new version from an older and operational version.

Two main types of software evolution can be named as static and dynamic evolution [11]. Static evolution consists in evolving the code of an application while it is stopped whereas dynamic evolution consists in evolving an application during its execution, without stopping it. The advantage of the former is that there is no need for state transfer or active thread to solve, whereas main disadvantage is that application is stopped and thus its services are stopped too (there is temporary unavailability). The advantage of the latter is no unavailability but there are some uncertain technical issues. Furthermore, evolution can be anticipated or unanticipated [11]. Anticipated evolution is an evolution that has been foreseen by the programmer while unanticipated evolution consists in evolution that has not been foreseen.

There is general agreement for the set of several important challenges on software evolution presented in [1]. We suppose that significant role for our research play mainly three ideas: evolution as a language construct, support for multi-language systems, and post-deployment runtime evolution. The challenge about evolution as a language construct states, that programming languages should provide more direct and explicit support for software evolution. The second challenge is about support for multi-language systems, and claims that software evolution techniques must provide more and better support for multi-language systems. The last challenge is about post-deployment runtime evolution and calls for an urgent need for proper support of runtime adaptations of systems while they are running, without pause or stop them. But these challenges need to be supported by appropriate software evolution platform (platform should be common, proper solution for this purpose could be Smalltalk [9]), which is another challenge from the set of challenges presented in [1].

## III. Transformation of Interpreted Functionality

Program P can be viewed as a sequence of statements that are aimed to produce some result R. This result R is obtained through the execution of the program P. The execution is done by a platform that interprets the program's sequence of statements. The result R of a computation depends on both a program P and interpreter I. Interpreter may be any virtual machine or in general even CPU. Different result may be obtained by changing, at least, one of the elements of the couple <P, I> [12].

### A. Program transformation

This approach is based on transformation of the program $P_0$ of the couple <$P_0$, $I_0$>. The interpreter $I_0$ is kept unchanged. A new program $P_1$ is built from both the application aspects and the program $P_0$. Just aspect-oriented programming (AOP) [7] is based on this principle and transformation of program $P_0$ into program $P_1$ is performed using transformation process which name is weaving. Originally this approach is named as weaving through program transformation [12].

AOP supports means for separation of crosscutting concerns and thus allows separation of design and implementation of a large complex software system into several parts. Base functionality can be expressed by appropriated base-level language and crosscutting concerns can be expressed by one or more domain-specific languages. Basic functionality and crosscutting concerns are then composed according to assigned rules.

Weaving can be invasive or non-invasive and is performed in compile-time, load-time or run-time. Compile-time weaving is usually named as static weaving, run-time weaving as dynamic weaving. If load-time weaving is performed only before execution of a program, it is a case of static weaving, if load-time weaving is performed during execution of a program, it is a case of dynamic weaving.

### B. Interpreter transformation

This approach is based on transformation of the interpreter $I_0$ of the couple <$P_0$, $I_0$>. The program $P_0$ is kept unchanged. Originally this approach is named as weaving through interpreter transformation [12] since aspect weaving consists in applying these transformation rules to the initial interpreter $I_0$ and thus new interpreter $I_1$ is created. A new interpreter $I_1$ is built from both the application aspects and the interpreter $I_0$.

Our approach is related to this group of transformation because it is based mainly on non-program transformation (but it is possible to extend it into transformation from <$P_0$, $I_0$> to <$P_1$, $I_1$>). Our idea of interpreter transformation is widen by transformation of various components of the execution mechanism. Thus, according to our approach, execution is a synonym for transformation in general, such as translation, type checking, code generation, loading, interpretation, modelling, algebraic specification, and even for informal but constructive thinking about algorithmic problems.

Depending on the result of interpretation, the adaptive language should be changed, and the next interpretation follows different semantics, i.e. potentially different result of the same source expression.

## IV. Metaprogramming and Reflection

Metaprogramming is about writing programs that represent and manipulate other programs or themselves, i.e. metaprograms are programs about programs [15]. The impact of metaprogramming is obvious in traditional development processes, by sorting existing programs as transformational processes with inputs and outputs.

Reflection is an entity's integral ability to represent, operate on and otherwise deal with itself in the same way that it represents, operates on, and deals with its primary subject matter [15]. A metalevel provides information about selected system and makes the software self-aware. A base level includes the application logic.

There are two aspects of reflection [18]: introspection and intercession (Fig. 1). Introspection is the ability of a program to observe and therefore reason about its own state. Intercession is the ability of a program to modify its own execution state or alter its own interpretation or meaning. Both aspects require a mechanism for encoding execution state as data, providing such an encoding is called reification [8, 16, 18]. Thus with introspection it is possible to read and access reifications whereas intercession allows also modify these reifications [17].

Reflection can be divided into structural and behavioural reflection [16]. Structural reflection represents the ability of a program to access a representation of its structure, as it is defined in the programming language. For instance, in an object-oriented language, structural reflection gives access to the classes in the program as well as their defined members.

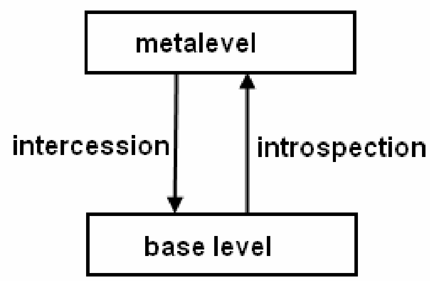



Fig. 1. Introspection and intercession between base level and metalevel

Behavioural reflection represents the ability of a program to access a dynamic representation of itself, that is to say, of the operational execution of the program as it is defined by the programming language implementation (processor). In an object-oriented language, behavioural reflection could for instance give access to base-level operations such as method calls, field accesses, as well as the state of the execution stack.

A programming language is said to be reflective if it provides an explicit representation (i.e., reification) of entities that either represent program building blocks (e.g., classes, methods) or are involved in program execution (e.g., stack, garbage collector) [16]. Developers thus can define system (software) functionalities and also new program building blocks or execution mechanisms (how functionalities will be performed).

From the object-oriented point of view, objects that define program functionalities are called base level objects or base-objects, objects defining program building blocks or execution mechanisms are called meta-level objects or meta-objects [15].

Meta-Object Protocol (MOP) [16] offers possibility to extend the programming language and adapt the execution mechanisms. Using a reflective language it is possible to implement both applications functionalities using objects and non-functional concerns using meta-objects.

Using reflection and MOP is the most appropriate for interpreted languages. An interpreter is the ideal place for metalevel information about running program.

## V. Smalltalk Possibilities For Software Adaptability

Many wide-spread languages such as C++ and Java do not have appropriate reflective properties. C++ does not have any reflective features included, Java was successively extended to support some reflective mechanisms such as structural introspection, but its support for behavioural reflection is still limited [17]. The implementation of Smalltalk [8, 9] itself is structured as an object-oriented program, expressed in Smalltalk and organized around meta-level objects representing the classes, methods, lexical closures, processes. Moreover, a Smalltalk environment implementation offers a compiler, debugger and others tools that can be easily adapted [2].

Squeak [10, 21] is a Smalltalk implementation that improves reflection possibilities of its ancestor. Squeak handles methods as first class entities that can be modified and replaced on the fly, and also allows for method execution reification, exposing dynamic information related to method execution as objects [2].

According to [6] the point of behavioural reflection is to modify the syntax, the semantics and the implementation of the language. Compiler extensions and automatic compiler generation appear as tools of choice to directly achieve such modifications. A modification to the syntax is achieved by modifying the lexer and the parser. Modifications to the semantics are implemented by modifications to the code generator. The runtime system can also be modified to match the needs of an application. The semantics of the language can be incrementally modified with utilization of reflection.

Our approach of adaptive language implementation [4, 5] (which is background of our adaptive execution idea) is based on the fact, that depending on the result of interpretation, an adaptive language should be changed, and the next interpretation should follow different semantics, i.e. potentially different result of the same source expression. Smalltalk offers fully reified compilation process. Just Smalltalk implementation Squeak offers appropriate tools for this adaptive language implementation by means of reflective compilation process. There are several possibilities how to compile program with utilization of reflection. One of them consists of the following parts: Scanner/Parser (built with Smalltalk compiler compiler – SmaCC [22]), Semantic Analysis, Translation to so called intermediate form, and Bytecode generation.

An extension of Smalltalk named MetaclassTalk [13, 14] provides explicit metaclasses and allows the method evaluation process to be transparently controlled and redefined. Within MetaclassTalk, classes play the role of metaobjects which define execution mechanisms for their instance. Another extension, which name is Bytesurgeon [3] can provide access to the values pushed on the stack, when needed, using so called meta-variables. We suppose, that possibilities of reflective compilation process, explicit metaobject protocol and other available solutions including Bytesurgeon can allow realisation of our idea on adaptive language implementation as a part of Squeak environment. Explicit metaobject protocol can be utilized for parameterization of compilation process and thus can allow building of metalevel compiler. Basic ideas about metalevel compiler infrastructure are presented in [5].

## VI. Conclusion

In this paper we have presented basic principles of program and interpreter transformation, metaprogramming and reflection in order to achieve software evolution.

Next step of our work will be proposal and implementation of the execution environment, which will support language transformation (in sense of interpreter transformation). Utilization of Squeak reflective possibilities, modification of Squeak language or environment and adding new domain specific languages could improve our solution.


## Acknowledgment

This work was supported by VEGA Grant No. 1/4073/07 Aspect-oriented Evolution of Complex Software Systems.


## References


[1] T. Mens, M. Wermelinger, S. Ducasse, S. Demeyer, R. Hirschfeld, and M. Jazayeri, “Challenges in Software Evolution“, In Proceedings of the Eighth International Workshop on Principles of Software Evolution (IWPSE'05), pp. 13-22, 2005.

[2] A. Bergel and M. Denker, “Prototyping Languages, Related Constructs and Tools with Squeak” Proceedings of the Workshop on Revival of Dynamic Languages (co-located with ECOOP'06), July 2006.

[3] M. Denker, S. Ducasse and É. Tanter, “Runtime Bytecode Transformation for Smalltalk“, Journal of Computer Languages, Systems and Structures, vol. 32, no. 2-3, July 2006, pp. 125-139.

[4] J. Kollár, J. Porubän, P. Václavík, J. Bandáková, and M. Forgáč, “Adaptive Language Approach to Software Systems Evolution“, International Multiconference on Computer Science and Information Technology: 1st Workshop on Advances in Programming Languages (WAPL'07), Wisla, Poland, October 15-17, Polish Information Processing Society, 2007, 2, pp. 1081-1091, ISSN 1896-7094.

[5] J. Kollár, J. Porubän, P. Václavík, J. Bandáková, and M. Forgáč, “Adaptive Compiler Infrastructure“, Komunikačné a informačné technológie, Tatranské Zruby, October 2007, pp. 4-5, ISBN 978-80-8040-324-9.

[6] J. Malenfant, M. Jacques and F. Demers, “A Tutorial on Behavioral Reflection and its Implementation“, Proceedings of Reflection 96, San Francisco, 1-20 (1996).

[7] R. Filman, and D. Friedman, “Aspect-oriented programming is quantification and obliviousness“, Workshop on Advanced Separation of Concerns (OOPSLA 2000), October 2000.

[8] F. Rivard, "Smalltalk: a Reflective Language", Proceedings of Reflection '96 Edited by G. Kiczales San Francisco, April 1996, pp 21-38.

[9] A. Goldberg, D. Robson, “Smalltalk-80: The Language“, Addison Wesley, 1989.

[10] A.P. Black, S. Ducasse, O. Nierstrasz, and D. Pollet, “Squeak by Example“, Square Bracket Associates, Switzerland, 2007.

[11] M. Oriol, “An Approach to the Dynamic Evolution of Software Systems“, Ph.D. Thesis, University of Geneva, Geneva, Switzerland, April 2004.

[12] N. Bouraqadi, and T. Ledoux. “How to weave?“, ECOOP 2001 Workshop on Advanced Separation of Concerns, June 2001.

[13] N. Bouraqadi, “Concern Oriented Programming using Reflection”, In Workshop on Advanced Separation of Concerns – OOSPLA 2000.

[14] N. Bouraqadi, and T. Ledoux, “Aspect-oriented programming using reflection”, Technical Report 2002-10-3, Ecole des Mines de Douai, October 2002.

[15] K. Czarnecki, and U. Eisenecker, “Generative Programming: Methods, Tools, and Applications“, Addison Wesley (2005), 832 pp.

[16] E. Tanter, "From Metaobject Protocols to Versatile Kernels for Aspect-Oriented Programming", PhD thesis, University of Nantes, France, and University of Chile, Chile. November 2004.

[17] D. Röthlisberger, “Geppetto: Enhancing Smalltalk's Reflective Capabilities with Unanticipated Reflection”, Masters thesis, University of Bern, January 2006.

[18] I. D. G. Bobrow, R. G. Gabriel, and J.L. White, “CLOS in Context - The Shape of the Design Space“, In Object Oriented Programming - The CLOS Perspective. MIT Press, 1993.

[19] M. Lehman, Laws of Software Evolution Revisited, EWSPT96, Oct. 1996, LNCS 1149, Springer Verlag, 1997, pp. 108-124.

[20] M. Lehman and J. Ramil: Towards a theory of software evolution - and its practical impact. Proceedings of the International Symposium on Principles of Software Evolution, Nov. 2000, Japan, pp. 2-11.

[21] Squeak homepage, http://www.squeak.org, 25. 3. 2008.

[22] Smalltalk Compiler Compiler home page, http://www.refactory.com/Software/SmaCC, 25. 3. 2008.

# Using gender-dependent acoustic models in speech recognition

*Marek PAPCO*

Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

marek.papco@tuke.sk

***Abstract*—The formant frequences in man and woman voice significantly differ. This article focus on the training set of acoustic models for men and women speakers separately. Acoustic models are 3-state continuous density HMM models with Gaussian mixture to model acoustic units. Features used to training models were MFCCs with delta and acceleration coefficients. Models were trained with HTK toolkit using flat-start method, method of isolated units and method of triphones with tied states. All methods were used to train models with different number of Gaussian used. Training and testing of acoustic models were on SpeechDat-E database. Tests were executed on application words, isolated digits, own names, phonetically rich words and connected digits. The results of speech recognition using gender-dependent acoustic models were compared with the gender-independent models. Comparing these results is seen that gender-dependent acoustic models increase speech recognition accuracy.**

***Keywords*—gender-dependent acoustic model, hidden Markov model, speech recognition, phoneme.**

## I. INTRODUCTION

The task of speech recognition (decoding) is conversion of speech signal into corresponded sequence of words in text. To make this conversion there are needed some complex steps [1], [2]. The decoding process include using of acouctic model, language model, phoneme set and dictionary as desribed in [3] and [4]. Acoustic model is achieved by training on SpeechDat speech database [5]. Most of speech recognition systems use hidden Markov models for modelling speech elements (phonemes, graphemes). The phonemes are basic elements of the speech and are used to compose words of the language. Their acoustic representation depends on the speaker vocal tract. The vocal tract determine distribution of energy on different frequences-formant frequences. This causes the different sounds of speech elements, that is used for speaker identifiaction, accent identification or gender identification [6], [7]. Distribution of formant frequences are highly different between men and women or between the languages, that is used to language recognition or crosslingual recognition [8], [9], where the phonemes mapping is used. In common speech recognition systems is used just one acoustic model that covers all speakers from training. But there are significant differences between men and women voices and these differences influence a parametres of acoustic models a lot of. Nowadays exist a lot of methods to increase the accuracy of systems for speech recognition and to make system insensible to different kind of noises or speakers. Ideal system for automatic speech recognition is speaker independent and imunne to noise. The main idea of this work is to improve accuracy of speech recognition by training two sets of acoustic models, one for men speakers only and second only for women speakers. To perform this was used HTK toolokit [10], that is primarily designed for building HMM-based speech processing tools. The training and testing procedures follow the MASPER initiative that was formed as a part of the COST 278 Action [11]. Trained gender dependent acoustic models then could be used in running Spoken Language Dialogue system IRKR [12], where from few initial speech sequences would system recognize the gender of the speaker and gender-dependent acoustic model would be used for the rest of the session.

## II. ACOUSTIC MODEL TRAINING

### A. Data preparation

As was mentioned before to train acoustic models was used Speechdat database. This database contains speech records by 1000 speakers, half are men speakers and other half are women speakers. Firstly was trained one set of models for all speakers without splitting to men and women speakers. In next step the records from 800 speakers were splitted for training (400 men, 400 women) and from 200 speakers (100 men, 100 women) were choosed for testing. To choose records the Perl programming language was used. Speechdat database was recorded in the office environment that does not contain high level of noises. The database contain records of aplication words, isolated digits, questions, own names, phonetically rich words and connected digits from each speaker. Therefore were created lists of records accord to gender and to purpose, either for training or testing. The other steps of data preparation not covered in detail are preparation of pronunciation dictionary, phonemes list, grammar and the transcriptions. The last stage of data preparation is parametrisation of speech waveforms into sequences of feature vectors. The speech waveforms in Speechdat database are stored without header in A-Law format, 8bit to sample with sampling 8000Hz and due to this the configuration file has to consider this. Here is the content of configuration file used to extract features:

```
TARGETKIND = MFCC_D_A_Z_0
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
ENORMALISE = F
SOURCEKIND = WAVEFORM
```

```
SOURCEFORMAT = NOHEAD
SOURCERATE = 1250
```

As is it seen the MFCC koefficients with delta, acceleration, static and zero mean static koefficients were used. The important parametres used to set up are *SOURCERATE*, that means sampling frequency. The value 1250 because HTK toolkit use 100ns format. *SOURCEKIND* and *SOURCEFORMAT* desribes type of the speech file, in this case waveform with no head.

### B. Creating flat-start monophones

The first step in training acoustic models was to create prototype to define HMM model. The values are not important, these are used as a starting point and they are estimating during the training. Example of the prototype HMM:

```
~o <VecSize> 39 <MFCC_D_A_Z_0>
~h "proto"
<BeginHMM>
 <NumStates> 5
 <State> 2
    <Mean> 39
      0.0 0.0 ....
    <Variance> 39
      1.0 1.0 ....
 <State> 3
    <Mean> 39
      0.0 0.0 ....
    <Variance> 39
      1.0 1.0 ....
 <State> 4
    <Mean> 39
      0.0 0.0 ....
    <Variance> 39
      1.0 1.0 ....
 <TransP> 5
0.0 1.0 0.0 0.0 0.0
0.0 0.6 0.4 0.0 0.0
0.0 0.0 0.6 0.4 0.0
0.0 0.0 0.0 0.7 0.3
0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

From prototype of HMM can be seen that used HMM model is 3-state HMM with used 39 length vector. The number 39 include 13 static coefficients plus 13 delta and 13 acceleration coefficients. Tool *HCompV* was used to compute global mean and variances from data files so the zero means and variances are replaced by global. Computed means and variances are used to create Master Macro File (MMF) and variance floor macro. Next, the tool *HERest* is used to re-estimation of flat-start monophones. Other tools used to training acoustic models are *HLEd* (to create phone-level transcription), *HHEd* (fixing silence model) and *HVite* (decoding).

### C. Method of isolated units

The variances and means are computed for initialized HMM model separately. Inicialization is provided by the tool *HInit* and the next reestimation is performed by *HRest*. *HHEd* add silence between words and in the next step is *HERest* used to re-estimation.

### D. Creating tied-state triphones

Context-dependent triphones can be made by simply cloning monophones and then re-estimating using triphone transcriptions. That is, executing tool *HLEd* were converted monophones transcriptions to equivalent set of triphone transcriptions and for cloning models in case for biphones was used tool *HHEd*. The *HDMan* was used to create the dictionary with triphone transcriptions and list of triphones. The outcome of the previous stage is a set of triphone HMMs with all triphones in a phoneme set sharing the same transition matrix. When estimating these models, many of the variances in the output distributions will have been floored since there will be insufficient data associated with many of the states. The last step in the model building process is to tie states within triphone sets in order to share data and thus be able to make robust parameter estimates.

The training process is ilustrated in Fig. 1. Same process of training acoustic models was used for men and women utterances separately.

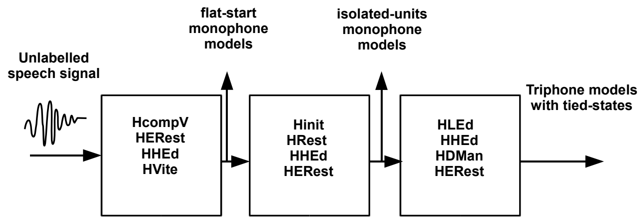



Fig. 1. Training process

In this manner were trained sets of gender-independent and sets of gender-dependent acoustic models with different methods and different number of Gaussians used. To adding Gausians is designated tool *HHEd*.

## III. Testing and Results

The tests were performed on the Speechdat-E database on application words, isolated digits, own names, phonetically rich words and connected digits. For testing were used records from 200 speakers, 100 men speakers and 100 women speakers. Acoustic models trained on men speakers were tested on men speakers, alike for women speakers. The recognition accuracy of the model was advised by the WER (Word Error Rate).

Word error rate was computed by equation:

$$WER = \frac{INS + SUBS + DELS}{Num. of\, words} \quad (1)$$

where *INS - inserted words, SUBS - substituted words, DELS - deleted words*

In the graphs are compared the WER values for gender–dependent models and gender–independent models. In the graphs GI_AM represents gender independent acoustic models, GD_AM_F represents gender dependent models trained and tested for women and GI_AM_M are men models.

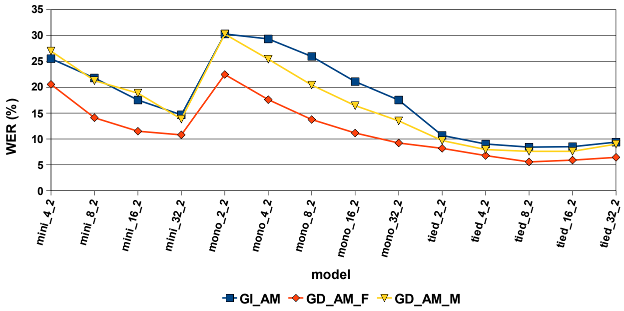



Fig. 2. Own names

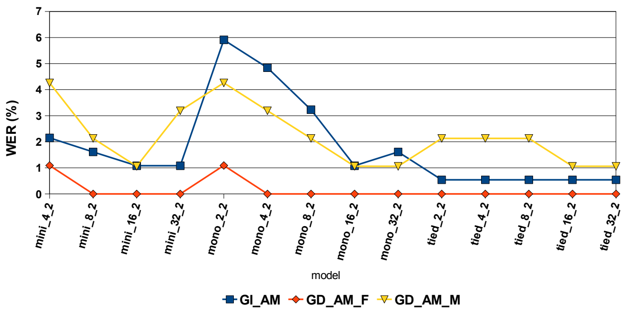



Fig. 3. Isolated digits

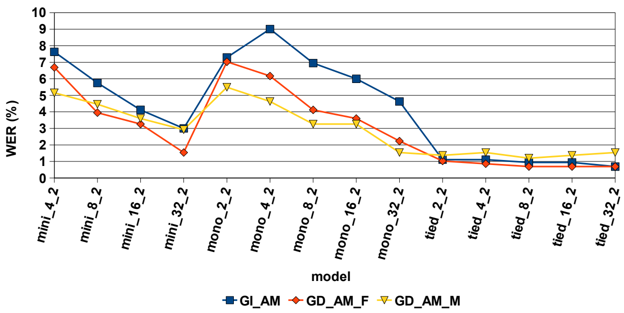



Fig. 4. Application words

## IV. Conclusions

The graphs in the figures 2 to 6 gives an overview of results obtained by testing acoustic models on several test scenarious. In summary the accuracy of recognition was increased with using gender dependent acoustic models in compare with gender independent models. The most increase of accuracy was achieved with the female models. The best accuracy in compare with gender independent models was achieved in the acoustic models trained with method of isolated units (mono) and flat-start method (mini). The least improvement was achieved in triphone models with tied states (tied). The biggest

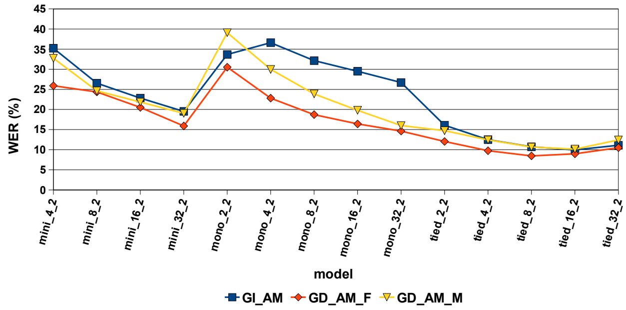



Fig. 5. Phonetically rich words

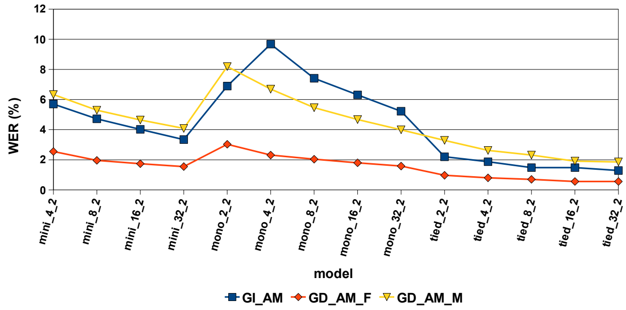



Fig. 6. Connected digits

increment in accuracy was achieved in phonetically rich words in model mono_4_2, that was 13.79% in GD_AM_F and 10.7% in model mono_8_2 in GD_AM_M.
In few models occured decrease in accuracy, but only for men models. I suppose this could solve the more training data from men speakers.

## Acknowledgment

Work pesented in this paper was promoted by the Ministry of education in Slovak republic within AV 4/0006/07 and VEGA num. 1/4054/07 projects.

## References

[1] X. Huang, A. Acero, H.W. Hon, Spoken language processing – A Guide to Theory, Algorithm, and System Developement. Prentice Hall, 2001.
[2] J. Psutka, J. Matoušek, L. Muller, V. Radová, Mluvíme s počítačem česky, Academia Praha, 2006.
[3] M. Mirilovič, System of automatic speech recognition for Slovak language. master thesis, FEI TU Košice, 2004.
[4] Lihan, S. Automatic speech recognition using HMM models, dissertation thesis, FEI TU Koice, (2004)
[5] M. Rusko, S. Daržagín., M. Trnka, M. Petriška, SpeechDat-E, the First Slovak professional-quality telephone speech database. In: Research Advances in Cybernetics. Košice : ELFA Publishing House, 2000.
[6] Jin, Q. Schultz, T. and Waibel A.: Phonetic Speaker Identification. Proceedings of the International Conference of Spoken Language Processing (ICSLP-2002), Denver, CO, USA, Sept. 16-20, p. 1345-1348
[7] Ghesquiere P. J., Compernolle D. V., Flemish Accent identification based on formant and duration features, Acoustics, Speech, and Signal Processing, Proceedings. (ICASSP ’02). IEEE International Conference, Orlando, USA, May 13-17, 2002, p. 749-751. ISBN: 0-7803-7402-9.
[8] Lihan,S.-Juhár,J.-Čižmár,A.: Crosslingual and Bilingual Speech Recognition with Slovak and Czech Speechdat-E Databases. Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, Sept. 4-8, 2005, p.225-228, ISSN: 1018-4074.
[9] Lihan,S.-Juhár,J.-Čižmár,A.: Comparison of Slovak and Czech Speech Recognition Based on Grapheme and Phoneme Acoustic Models. In: Proc. Interspeech 2006, Pittsburg, USA, Sept. 17-21, 2006, p. 149-152. ISSN 1990-9772.

[10] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book, Dec. 2006.

[11] Žgank, A. Juhár, J. et. al. The cost 278 masper initiative crosslingual speech recognition with large telephone databases, International Conference On Language Resources & Evaluation, no. 4, 4 2004.

[12] Juhár ,J. Ondáš, S. Čižmár, A. Jarina,R. Rusko,M. Rozinaj,G.: Development of Slovak GALAXY/VoiceXML Based Spoken Language Dialogue System to Retrieve Information from the Internet. In: Proc. Interspeech 2006, Pittsburg, USA, Sept. 17-21, 2006, p. 485-488. ISSN 1990-9772.

# Speech Emotion Recognition

*Ing. Matúš PLEVA, Ing. Lenka MACKOVÁ, Roman FRIČ*

Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

matus.pleva@tuke.sk, lenka.mackova@tuke.sk, romano@klacno.sk

***Abstract*—The aim of this article is to present an introduction to the speech emotion recognition. The first tests of emotional speech recognition and the process of the emotional database recording are described, too. The purposes of emotion detection are depicted and possible algorithms of emotion detection from human speech communication are listed. The results of the first tests on the database are presented and compared with other institutions.**

***Keywords*—emotion recognition, Hidden Markov Models - HMM, nonverbal communication, short-time analysis**

## I. Introduction

Nonverbal communication covers great area of different signals which follow human verbal communication. Nonverbal communication includes gestures, head and body movement, facial gaze, voice intonation and others.

Even when human emotions are hard to characterize and categorize the effort of its recognition increased in recent years due to the wide variety of applications that benefit from such technology.

Emotion recognition has lots of useful applications, for example in Human-Robotic Interfaces where robots can be taught to interact with humans and recognize human emotions (for example robotic pets could be able to understand to human commands), in call-centers where the „smart" robotic system can replace human operators, in intelligent spoken tutoring systems to fill the gap between human and computer tutors and others.

Emotion recognition solutions depend on which emotion is wanted to be recognized by a machine and for what purpose. In general there are six basic emotional states: *neutral, happiness, fear, sadness, anger and disgust (or surprise).* In this article we focus in four emotional states: *neutral, happiness, sadness and anger.*

## II. Speech Emotion Recognition

The main task of speech emotion recognition is appropriate voice processing. Human speech includes much more information than verbal text only. Each human voice is unique and in each human speech are coded emotions of the very speaker (different people use different tone in case of talking in anger, happiness or when even whisper). Besides human voice is changing, in cases of illness or even in the morning sounds, the human voice is a bit different than during the day.

Automatic emotion recognition of speech can be viewed as a pattern recognition problem [1]. The results of emotion recognition are characterized by: a) the features that are involved in the speaker's emotional state, b) the type of required emotions; c) the type of classifier used in the experiments and finally d) the database used for training and testing the classifier. The final results are obtained by comparison of classifiers in which case the same dataset and set of emotions is used.

Dellaert et al. [2] used in comparison three classifiers: the maximum likelihood Bayes classification, kernel regression, and k-nearest neighbor (K-NN) methods and was interested particularly in emotions of sadness, anger, happiness, and fear. The features used in this experiment were the pitch contour and the reached accuracy was 60%-65%.

In the experiment of Lee et al. [3] the linear discrimination, k-NN classifiers, and support vector machines (SVM) where used to distinguish two emotions states: negative and non-negative. In this case the maximum accuracy was 75%.

Navas et al. [4] provided two experiments of recognizing of joy, sadness, anger, disgust, surprise and fear. In the first one the short-term spectral features where used. The short-time statistics refer to the features obtained from the frame of the speech in short time intervals. The team of Navas used 18-MFCC and their first derivates. For each emotion in the database was built a GMM with 256 Gaussian mixtures.

In the second experiment long-term prosodic feature where used. The long-time statistics refer to the features calculated from speech parameters during a long time interval. The long-term features consist of different statistics calculated the pitch curve and it's first and seconds derivates, as well as from the first and second derivates of the power curve. The mentioned statistics where mean, variance, minimum range, skewness and kurtosis and finally jitter and shimmer values where also estimated and append to the final vector.

The results for emotional states differed approximately from 100% to 94%, in case of short-time analysis and from 97% to 83% in case of long-term analysis.

## III. Emotional Database Recording

The recording of posed emotions in 16 neutral sentences – speech utterances (four different type of sentences) was done for only 30 speakers for now (male/female, different ages, accents), using 4 emotional states (angry, neutral, happiness, sad) and 4 background conditions (home, quiet, public place, office) using notebook as a recording device.

The four different neutral sentences without an emotional meaning were spoken using four emotional speech states.

Recordings were done using notebook microphone in 48kHz 16bit WAV PCM format. Every speaker records 16 sentences in one background environment (home, quiet, public place or office). Then the recordings were converted using sox [5] tool to RAW format recordings. Next the recordings were labeled according to the emotional state posed in the speech segment.

After the Master Label File (MLF) [6] was generated using automated Perl scripts from the whole database, the configuration files for HTK tools were prepared (grammar file, dictionary, list of all speech files, list of emotions, HMM prototype, etc.).

## IV. Feature Extraction and Training

Only a short-time analysis [4] was done using the HMM models. The MFCC (Mel-Frequency Cepstral Coefficient) coefficients were used for feature extraction from the speech segments. More feature extraction configurations were tested. The energy (0), log-energy (E), delta (D), delta-delta - acceleration (A) coefficients and cepstral mean subtraction (Z) parameter were tested and the recognition results were compared. The best configuration was when computing thirteen MFCC log-energy, delta and acceleration coefficients with cepstral mean subtraction (E_D_A_Z) as we can see on the Figure 1 below.

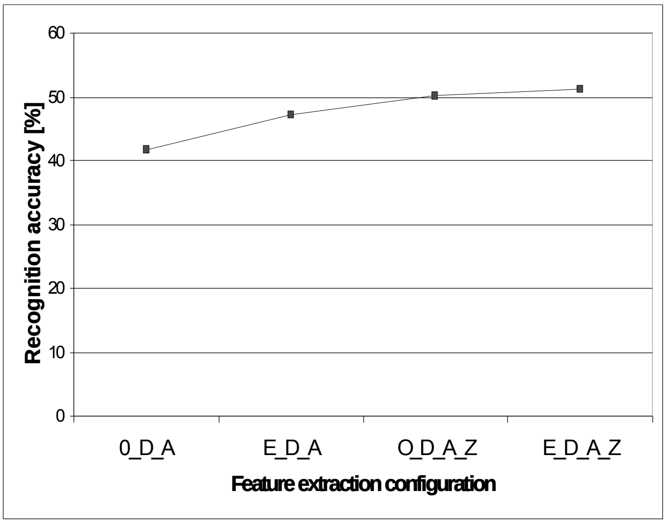



Fig. 1. The emotional states of the spoken segment recognition accuracy changed according to the feature extraction configuration.

The next step was testing different types of HMM prototypes (Fig 2). The left-right, and different ergodic prototypes from 3 to 7 states were trained and than tested using the same feature extraction configuration. The best result was reached using 5-state ergodic model with a small possibility to transit between all states (except the entering and emitting state).

Also a testing of different numbers of the PDF (Power Density Function) mixtures was done using from 2 to 256 PDF on the state. The recognition accuracy raised together with the number of the PDF mixtures on the state (Fig. 3). So for the first recognizer configuration the 256 PDF mixtures on the state of the HMM model was used.

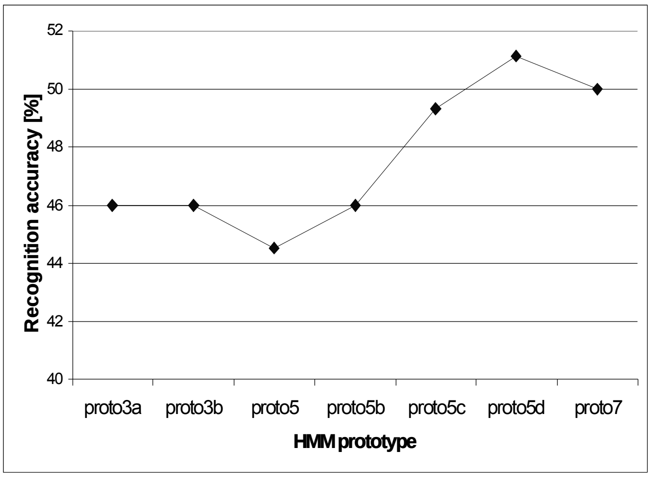



Fig. 2. The emotional states of the spoken segment recognition accuracy changed according to configuration of the HMM prototype and the number of the states.

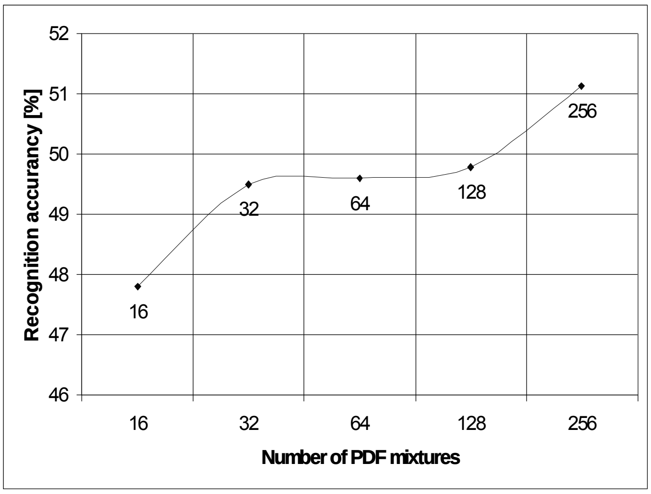



Fig. 3. The emotional states of the spoken segment recognition accuracy raised according to the number of PDF mixtures on the HMM model state.

## V. Testing

The testing was realized using Viterbi decoding algorithm from HTK Tools (HVite). The recorded speech utterances were divided to training and testing part in the rate of 2:1. The training part was used only for training and the testing part was used only for automatic recognition. The HMM model produced sometimes more than one result for one speech utterance. For example sadness was recognized for the first half and neutral for the second half of the speech utterance sentence. That's why the sum of the recognized emotions in the anger column is more than one hundred percent. The complete confusion matrix is shown below in the Table 1.

TABLE I
CONFUSION MATRIX [% OF RECOGNIZED EMOTION TO ORIGINAL]

| *Recognized:* | *Original Emotion* | | | |
|---|---|---|---|---|
| | *Anger* | *Sadness* | *Happiness* | *Neutral* |
| *Anger* | *66* | *8* | *17* | *6* |
| *Sadness* | *5* | *50* | *11* | *33* |
| *Happiness* | *26* | *11* | *54* | *8* |
| *Neutral* | *14* | *34* | *11* | *40* |

As we can see the anger and the happiness emotional speech is better recognized. The problem is to recognize the difference between neutral and sadness emotional speech. Both are uncertain also for humans, because there are differences not only between the men and women expression [7] of the emotions, but also there are differences between emotion expressions in the various cultures.

Next the verification of the results using emotional speech utterances recognized by humans will be done. All speech utterances will be recognized by humans and only utterances where the human recognized emotion corresponds with the original emotion will be used for training the automatic emotional speech recognition engine.

Then also testing will be done on the filtered utterances of the posed emotional speech. This process will filter the error, which corresponds to a bad acting of the emotions of the non-professional speakers in the recorded database. After the testing will be done the comparing of the results will answer the question how important is a good acting in posed emotional database recording.

## VI. Comparing The Results

So finally only a short-time analysis [4] using the 5-state ergodic HMM models was done with overall 51.13% accuracy using MFCC_E_D_A_Z parameterization, 256 PDF on each of the 5-state ergodic HMM prototype. In the future will this emotional speech recognition algorithm trained and tested using a bigger spontaneous emotional database with also the background conditions annotated (using F-measures [8]) after the annotation process will be done from the previously recorded live TV discussions.

This non-verbal information about the emotion of the spoken speech segment could give the hearing impaired users of the recognized speech utterances database a better understanding of the automatic transcribed texts, which could lead to better information efficiency [11].

The 51.13% recognition accuracy for recognizing one of four possible emotional states is not clearly comparable with the tests of the other institutions mentioned before. For example Dellaert et al. [2] reached 60% overall accuracy but recognizing 6 emotional states in speech utterances. Navas et al. [4] from Basque university reached 95% overall accuracy for recognizing six emotions, but on various bigger emotional speech databases as Idoia (665 emotional speech utterances) [9], Karolina and Pello (700 sentences) [10].

## VII. Conclusion

The interesting application for speech utterance emotion recognition is the automatic transcription of the speech with some emotional text labels. This additional information could give the user of the transcribed texts database the additional information about the meaning of the sentence. This additional emotional state information is useful when searching some special meaning of the words or presenting the texts without the audio data for hearing impaired or when the audio channel is not available.

Also a very useful application for emotional speech recognition engine is the area of the human machine interaction and virtual agents. There are some experiments with virtual agents, interacting on the emotional state of the human and trying to also place the speech synthesis [12] in the dialog and the whole virtual agent behavior in the same emotion, and make the communication friendlier for the end-users.

These preliminary results give us the knowledge for building the spontaneous emotional database, and spontaneous emotional speech recognition engine.

In the future we would like to tune the algorithms of training and the feature extraction process. Also the posed emotional speech database is not useful for spontaneous emotional speech recognition, so the annotated spontaneous speech database will be built from previously recorded live TV discussions.

## Acknowledgment

Research described in the paper was supported by the Slovak Ministry of Education under research projects AV 4/0006/07, AV 4/2016/08, VEGA 1/4054/07 and MVTS COST2102/07.

## References

[1] S. Yacoub, S. Simske, X. n Lin, and J. Burns, “Recognition of Emotions in Interactive Voice Response Systems, “ HP Laboratories Palo Alto, 2003, pp.2

[2] F. Dellaert, T. Polzin, and A. Waibel, “Recognizing Emotion in Speech,” in *Proc. of ICSLP 1996*, Philadelphia, PA, pp. 1970 -1973, 1996

[3] C. Lee, S. Narayanan, and R. Pieraccini, “Classifying Emotions in Human-Machine Spoken Dialogs”, in *Proc .of International Conference on Multimedia and Expo*, Lausanne, Switzerland, August 2002

[4] E. Navas, I. Hernáez, I. Luengo, I. Sainz, I. Saratxaga, and J. Sanchez, “Meaningful Parameters in Emotion Characterisation,” In: *Verbal and Nonverbal Communication Behaviours*, Lecture Notes in Computer Science 4775, Springer Verlag 2007, pp. 74 - 84, Editors: A. Esposito et al., Revised selected and invited papers from COST Action 2102 international workshop in Vietri sul Mare, Italy, March 2007

[5] http://sox.sourceforge.net/ - cached March 2008

[6] S. Young, “ATK: An application Toolkit for HTK”, version 1.3, Cambridge University, January 2004

[7] T. Vogt, and E, André, “Improving automatic emotion recognition from speech via gender differentiation,” in *Proc. Language Resources and Evaluation Conference (LREC 2006)*, pp. 1123 - 1126

[8] J. Žibert, F. Mihelič, J.P. Martens, H. Meinedo, J. Neto, L. Docio, C. Garcia-Mateo, P. David, J. Zdansky, M. Pleva, A. Čižmár, A. Žgank, Z. Kačič, C. Teleki, and K. Visci, “COST278 broadcast news segmentation and speaker clustering evaluation”, In: Interspeech, Lisboa, 2005. Bonn, Universität Bonn, 2005. p. 629-632. ISSN 1018-4074

[9] E. Navas, A. Castelruiz, I. Luengo,, J. Sanchez , and I. Hernáez, “Designing and recording an audiovisual database of emotional speech in Basque,” in Proc. of *Language Resources and Evaluation Conference (LREC 2004)*, pp. 1387 - 1390

[10] I. Saratxaga, E. Navas, I. Hernáez,and I. Luengo, “Designing and recording emotional speech database for corpus based synthesis in Basque,” in Proc. of *Language Resources and Evaluation Conference (LREC 2006)*, pp. 2127 - 2129

[11] M. Pleva, A. Čižmár, J. Juhár, S. Ondáš, and M. Mirilovič, “Towards Slovak Broadcast News Automatic Recording and Transcribing Service”, In: *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Selected papers from COST Action 2102 International Workshop. Volume Editor(s): A. Esposito, N. Bourbakis, N. Avouris, and I. Hatzilygeroudis. Lecture Notes in Computer Science, to be published in Springer Verlag 2008, pp. 165-176

[12] J, Pribil, and A. Pribilova, “Emotional style conversation in the TTS system with cepstral description,” In: *Verbal and Nonverbal Communication Behaviours*, Lecture Notes in Computer Science 4775, Springer Verlag 2007, pp. 65 - 73, Editors: A. Esposito et al., Revised selected and invited papers from COST Action 2102 international workshop in Vietri sul Mare, Italy, March 2007

# Streaming of multimedia to mobile and desktop users using Videoserver platform

*Miroslav MICHALKO, Vladimír SIDIMÁK*

Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

miroslav.michalko@tuke.sk, vladimir.sidimak@cnl.tuke.sk

***Abstract*—Recent advances in computer networking combined with powerful home computers and modern operating systems made streaming media practical and affordable for most end users. This paper deals with implementation of this new streaming solutions into video streaming platform (Videoserver) developed by Computer Networks Laboratory at Technical University of Košice. Videoserver provides services also for mobile users using latest streaming technologies. Moreover this solution provides HDTV streaming as multicast and unicast service for clients capable receiving high bandwidth data streams.**

***Keywords*—streaming, flash, video-on-demand, platform, mobile**

## I. Introduction

Videoserver is platform developed to stream multimedia content to wide variety of end users and platforms. Platform provides powerfull and easy-accessible services for sharing video and audio resources using Flash technology. For mobile clients there are several possibilities how to deliver content. Content is streamed in optimized data stream. Videoserver allows also sharing of audio files and pictures. One of powerful service provided by Videoserver is possibility to create multimedia content in supporting environment. After accessing archive, Videoserver initiates streaming of video stored in server directly to your computer over the network.

Videoserver also contains several technologies which supports access to formats using mobile devices. Implementation of RSS feeds allows simple notification and WAP page provides also short description of last content added into platform. Accessing video streams is possible using Darwin Streaming Server which streams video in on-demand mode for devices capable receiving rtsp streams.

Services provided by Videoserver streaming platform are preferably targeting school and universities by allowing them to share educational content by simple way. Thanks to open source and widely used technologies behind, platform is easily accessible and there are no special requirements to end users.

## II. Realisation of Videoserver platform

### A. Desktop streaming technology

Selecting the proper multimedia format means to be successful or not. End users are bored to install always new extensions and applications in order to receive some special content. Based on several studies Flash is the most installed technology on mobile and desktop operating systems. For this reason Flash technology was selected [1].

Videoserver streams data flows using Flash coded video and HTTP protocol, most simple and cheapest method. HTTP streaming is enough for our purposes, but for larger video portals it's recommended to use serious streaming technology. As client JW FLV is used and for streaming purposes Lighttpd server instead on Apache.

In case of higher number of clients Red5 technology will override HTTP streaming. Red5 is a free, open source Flash server that supports streaming and recording audio/video, live stream publishing and Flash remoting. Big advantage of this Java based application is cooperation with FFMPEG software package. Videoserver use for flash streaming Red5 as the main streaming technology [2]. Red5 is easy to install and maintain. Unfortunately still lack some important features in compare to Adobe Flash Streaming server.

### B. Automatic video processing

Videoserver allows automatic video processing using FFmpeg. FFmpeg is a collection of software libraries that can record, convert and stream digital audio and video in numerous formats. It includes libavcodec, an audio/video codec library used by several other projects, and libavformat, an audio/video container mux and demux library. After submitting video using portal Upload form ffmpeg is used to automatically convert video into flash format. During converting phase user is allowed to see partial video output of encoded video [3].

### C. HDTV Streaming

There are only few streaming technologies available with HDTV streaming capability. Most known are Microsoft Media Services with its own proprietary HDWMV codec, Helix based solution provided by Real, QuickTime and VLC. Two most suitable technologies VLC and Helix are implemented and provide multicast and unicast on-demand streaming. For unicast streaming special module was developed and implemented into IP streaming platform running on CMS Drupal.

Due to limitations of multicast streaming content is accessible only within local intranet of Technical University in Kosice - TUNET. In order not to overload network with unnecessary data stream platform allows also editing TTL value in TCP packet header. After reaching value equal to zero packet is destroyed by any active layer 3 packet device. HDTV video streams are propagated using SAP playlist what makes system more user friendly in compare to use only IP multicast address to remember. For extranet viewers interested in multimedia content there is only on-demand streaming available using specially developed web interface. Due to higher data bandwidth requirements only users with capable connection are ready to access content.

### D. Web Interface

If considering to build complex streaming portal there are two possibilities. First, most time consuming one is to make code from scratch. Second, more simple and preferred by many users is to use existing Content Management Systems (CMS) available at the market. One of the greatest advantages of web content management systems is their modular structure which allows system to be expanded of required features. The base of web content management systems allow to include different modules developed for actual web CMS.

Drupal is a free modular content management system written in PHP. It has a basic layer, or core, which supports pluggable modules which enable additional behaviors. The integration between the core and the modules is achieved via a system of hooks, or call-backs, to allow modules to insert functions into Drupal's path of execution. Drupal core provides protection against many of the usual security problems, like SQL injection. Drupal offers many advantages to its users. For its ease installations and configuration and very few requirement of special technical acknowledge it is very popular between administrators. After creation of web site or portal it offers simple and clear navigation trough the site and organization of the content [4].

One of the Drupal greatest advantages and popular features which does not have any other CMS is taxonomy module. It is a core module, it means, taxonomy module is automatically installed when installing Drupal. The great power of taxonomy module comes from its ability to organize content by type. The taxonomy module can automatically classify new content on-the-fly. The taxonomy module also allows user to define vocabularies or sets of categories which are used to classify content.

## III. Mobile Streaming

### A. Streaming servers

Streaming for mobile devices present numerous challenges, such as how to provide efficient content delivery for various mobile services for different user devices. Most known are Darwin Streaming Server, Mobile Flash, Microsoft Windows Media and RealNetworks. Videoserver uses Darwin Streaming Server for its properties and simple way of editing streaming parameters.

There are also other technologies for video content delivery. For TV or IPTV streaming solutions with broadcast delivery method is successor of DVB-T named DVB-H. DVB-H is developed for digital terrestrial television to the specific requirements of handheld, battery-powered receivers. DVB-H can offer a downstream channel at high data rates which can be used as standalone or as an enhancement of mobile telecoms networks which many typical handheld terminals are able to access anyway [5].

### B. Problems in mobile streaming

There are several problems related with mobile streaming. For example its bandwith limitations, low computing power of mobile devices, packet loss, delay, billing, etc. There are some possibilities how to improve some parameters. For better bandwith parameters active devices like routers can utilize QoS features. Streaming servers rely on network conditions and calculates its own parameters. Based on results video bitrate has to adapt.

TCP-Friendly Rate Control Protocol (TFRC) is useful framework for multimedia streaming servers [6]. Protocol adjusts its transmission rate on server side in response to the level of congestion, as estimated based on the calculated loss rate. TRFC sender uses the following formula for response evaluation:

$$T(p, RTT, RTO) = \frac{1}{RTT\sqrt{\frac{2p}{3}} + RTO(3\sqrt{\frac{3p}{8}})p(1+32p^2)} \tag{1}$$

where $p$ is the steady-state loss event rate and RTO is the retransmission timeout value [7]. Equation (1) enforces an upper bound on the sending rate *T*. Also utilization of Scalable Streaming Video protocol together with Loss Discrimination Algorithm should be an option.

One of the possible technique how to improve streaming parameters is to measure actual network conditions on client side and based on results to modify video parameters on server side. It could be achieved by changes in TCP window sizes. The Transmission Control Protocol receive window size is the maximum amount of received data, in bytes, that can be buffered at one time on the receiving side of a connection. The sending host can send only that amount of data before waiting for an acknowledgment and window update from the receiving host.

For future development in mobile streaming 3GPP2 standard will be implemented. The chosen protocols are fully compliant with existing standards. Videoserver through its HTTP protocol provides access to static content using TCP connection and upgrade requires only implementation of streaming server with 3GPP2 support. For improving quality an extra proxy server should be considered as an option. Proxy handles problems with continuously changing conditions within providers network environment [8].

## IV. Future plans for Videoserver

Videoserver provides powerful platform for content creating and sharing in supporting environment. After accessing archive, Videoserver initiates streaming of video stored in server directly to your computer over the network. Current development is now focused on content searching and user interactivity in order to make this system more flexible and interesting for students what makes portal different to other portals. Sharing of video materials, web-based curriculums and on-line exams is part of e-learning and blended learning. Implementation of these learning techniques is powerful and effective strategy how to increase quality of education process at schools.

In order to extend platform functionality and to improve quality of offered services following activities are planned in near future:

Customizability - next generation services will be oriented directly on user needs. Videoserver will allow creating of user defined playlists supporting various formats of video, audio and picture files. Using simple and user defined web links will make playlist easy to share over the network. Sharing of pictures will be divided into categories or albums with different level of access privileges defined by owner.

Mobile streaming - as rapidly growing segment demands for implementation of new 3GPP2 scenario. To enable interoperability between servers and mobile devices, especially

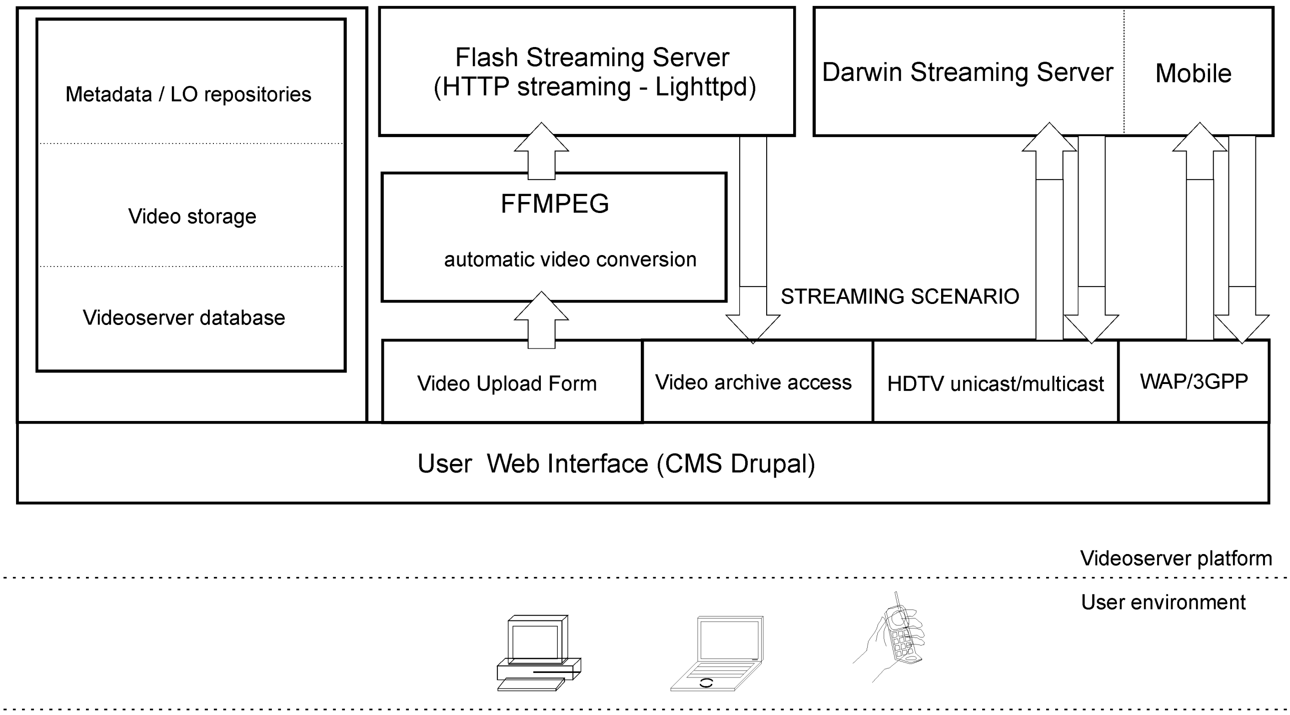



Fig. 1. Architecture of the Videoserver platform.

when using MMS standard specifies MPEG4 as an optional file format but this is currently problem. User uploaded videos are converted only into flash video format, additional conversion requires high system resources on server side. The 3GPP streaming standard offers the possibility of creating presentations in which several media elements such as video, audio, images, and formatted text play at the same time. SMIL, an XML-based presentation language developed by the World Wide Web Consortium, is the glue that combines these different elements to create an interactive multimedia presentation. SMIL is HTML with additional notions of time and temporal behavior. Thus, it can describe a media screen and control the placement of media elements in space and time. The 3GPP streaming client interprets the SMIL scene description and uses it to control the spatial layout and synchronization in the multimedia presentation.

IPTV - as next logical step. Delivering of content is planned by using of existing streaming scenario based primarily on flash technology. Developer snapshots are already available [9].

## V. Conclusion

Technologies presented in this paper show only few examples of utilization of new streaming and processing technologies connected together in order to provide services with added value. Videoserver platform provides complex service for sharing multimedia content using latest streaming technologies. Implementation of mobile streaming extension is unique in compare to similar web portals. Next development will focus on higher level of user interactivity, streaming adaptability based on continuously changing conditions within providers network environment, distributed access to video content and automatic metadata recognition.

## References


[1] G. O'Driscoll, *Next Generation IPTV Services and Technologies*. Hoboken, New Jersey: Wiley-Interscience, Jan. 2008.

[2] Red5, "Open source flash server," 2007, [Online; accessed 15-March-2008]. [Online]. Available: http://osflash.org/red5

[3] Wikipedia, "Ffmpeg," 2008, [Online; accessed 16-March-2008]. [Online]. Available: http://en.wikipedia.org/wiki/FFmpeg

[4] Drupal, "An open source content management platform," 2008, [Online; accessed 16-March-2008]. [Online]. Available: http://www.drupal.org

[5] I. Wong and B. Evans, *Resource Allocation in Multiuser Multicarrier Wireless Systems*. New York: Springer, Nov. 2007.

[6] M. Handley, S. Floyd, J. Padhye, and J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification," Internet Engineering Task Force, RFC 3448, Jan. 2003. [Online]. Available: http://www.rfc-editor.org/rfc/rfc3448.txt

[7] S. Floyd, M. Handley, J. Padhye, and J. Widmer, "Equation-based congestion control for unicast applications," in *SIGCOMM 2000*, Stockholm, Sweden, August 2000, pp. 43–56. [Online]. Available: citeseer.ist.psu.edu/floyd00equationbased.html

[8] S. Wee, J. Apostolopoulos, W. Tan, and S. Roy, "Research and design of a mobile streaming media content delivery network," 2003. [Online]. Available: citeseer.ist.psu.edu/wee03research.html

[9] F. Jakab, M. Michalko, M. Binas, and J. Baca, "Flash and Mobile Solutions for Video Streaming Platform," *5th International Conference on Emerging e-Learning Technologies and Applications, Conference Proceedings*, pp. 163–170, Sept. 2007.

# Integration process of security as QoS parameter via Security Service Vector in MANET

*Ján Papaj*

Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

jan.papaj@tuke.sk

**Abstract—Mobile ad-hoc network (MANET) is a collection of communication devices or nodes that communicates without any fixed infrastructure and pre-determined organization of available links. The MANET provides new challenges in area of commercial services for end users. The main challenge is provide ability to configure own level of security and QoS. This article describes basic techniques of integration security as a Quality of Service (QoS) parameter by modified concept of security service vector (SSV) to the MANET. Nowadays in MANET, security and QoS have been considered separately. Standard QoS models and architectures are network-oriented and the parameters don't reflect user's requirements to providing of new services. The security mechanism provides end-to-end security but user has no chance to change level of protection or level of security itself. In this article, we show how to integrate security as a QoS parameter by SSV. The SSV is implemented to the MANET via Dynamic Source Routing Protocol (DSR). We show basic principle and operations of this process.**

***Keywords*— DSR, QoS, Security, Security Service Vector**

## I. INTRODUCTION

MANET is a collection of communication devices or nodes that communicate without any fixed infrastructure and predetermined organization of available links. The nodes in MANET themselves are responsible for dynamically discovering other nodes to communicate [1].

Specific services have specific QoS and security needs. Although the ongoing trend is to adopt ad hoc networks for commercial uses due to their certain unique properties, the main challenge is provide ability to configure own level of security and QoS. The QoS and security have become a primary concern in order to provide protected communication between mobile nodes in a hostile environment with defined QoS and security [2].

The unique characteristics of mobile ad hoc networks pose a number of nontrivial challenges to security design, such as open peer-to-peer network architecture, shared wireless medium, stringent resource constraints, and highly dynamic network topology. These challenges clearly make a case for building multifence security solutions that achieve both broad protection and desirable network performance [3].

In MANET, security and QoS have been considered separately. Both have different objectives and implementation architectures and schemes. Nowadays no protocols, algorithms or models have been designed and implemented so far to integrate security as a one QoS parameter to the MANET. This implies that users have no opportunity to configure their own level of security a QoS parameters [4].

This article deals with possibility of security integration as a QoS parameter in MANET environment. The integration provides user's new abilities of configuration QoS and security requirements.

This article provides extension of concept of security service vector **(*SSV*)**, which is used in wired network [2]. With SSV, users will get the possibility to specify own requirements for QoS and security.

First step of integration is modification of routing protocol. For testing implementation, Dynamic Source Routing **(*DSR*)** was selected for testing process of adding SSV. This article does not deal with process of finding of security and QoS related information and we will not investigate the security as a architecture.

## II. INTEGRATION SECURITY AS A QOS PARAMETER – OVERVIEW

Security and QoS are very important area of research. It's interesting that area of QoS is not as old as security but has been extensively addressed by researchers and practitioners. In QoS literature, there was security interpreted as a dimension QoS, but process of integration has not been studied.

The concept of security as a dimension of QoS has been suggested as a concept called ***variant security*** [5]. The idea in this concept is that security mechanisms and services are considered to have a security range and a set of measurable security variables have been identified, which can be used to quantify a security attribute. The term ***Quality of Security Service*** (QoSS) has been provided by authors Irvine at al. [6].

The ***Explicit Endpoint Admission Control*** (*EEAC*), is proposed, with the introduction of a new concept, namely the service vector. The EEAC and service vector scheme allows in order enhancing the end-to-end QoS granularity in the network [7].

A ***security service vector*** (*SSV*) has been presented for describing functional requirements of security policies. SSV was proposed to represent the level of services within the range of security services and mechanisms. The attributes of their security vector include security components, security services, level of security, and service area [8].

## III. SECURITY SERVICE VECTOR (SSV)

Concept of Security service vector (SSV) is introduced in [8]. The main concept is based on EEAC [7]. Security vector (*SV*) was proposed to determine a number of customizable Security Services (SSs) with choices of customizable Service Degrees (*SDs*). Let $SSV = \|SSV\|_J$ be a security service vector, consisting of *j* security service vector portions, where each vector portion is dedicated to every nodes. The *SSV* portions of each intermediate node are in the generic form of

$$\|SSV\| = \bigcup_{j=1}^{J} \|SSV\|_j \equiv \{(SS^1, SD_m),...(SS^N, SD_m),[x]\}_1,..,\quad (1)$$
$$\{(SS^1, DS_m),...,(SS^N, SD_m),[x]\}_J$$

where *x* denotes other information that can be attached, for example estimated cost, time and data length. There are two communication phases taken place for data transmission: **probing phase** and **data transmission.**

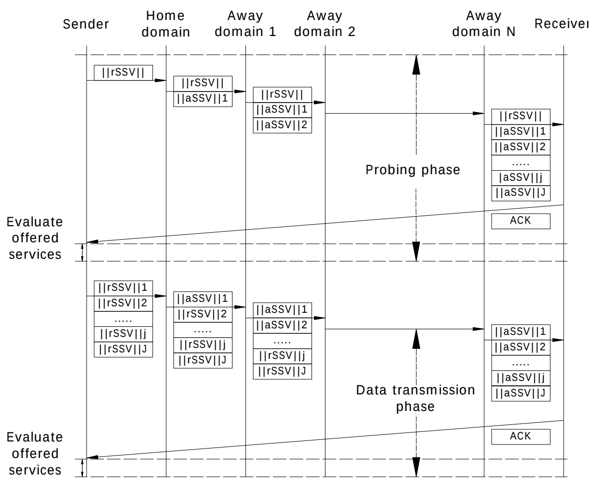



Fig. 1. The transmiton diagram of SSV

**The probing phase** happens during a connection establishment and the data phase starts after the connection has been set up. During the probing phase, the sender who wants to exercise security service options sends the probing packet through all domains along the path.

The probing packet verifies whether satisfied security services can be offered along with sufficient resources for the following data packets.

This probing packet contains the same requested SSV (*rSSV*) for every domain. The *rSSV* portion in the probing phase is denoted as

$$\|rSSV\| = \{(SS^1, SD_m),...,(SS^N, SD_m),[data\_lenght]\}\quad (2)$$

Any edge router performs the following basic tasks: verifying the sender's identity; examining the *rSSV*; verifying whether the sender has a privilege to request the services; checking available resources; writing down its aSSV portion; and forwarding the probing packet to the next hop. The receiver replies with an *ACK* packet containing all available *SSV* (*aSSV*) portions to the sender.

The sender then evaluates all services offered and concludes whether to proceed to the data phase or to drop this connection and try again later. At the end of the probing phase, the querying user retrieves information from all *aSSV* portions carried in the *ACK* packet is denoted as

$$\|aSSV\| = \bigcup_{j=1}^{J} .\|aSSV\|_j \equiv$$
$$(SS^1, SD_m),...,(SS^N, SD_m)\{(SS^1, SD_m),...,(SS^N, SD_m),$$
$$[delay, time\_process, \cos t]estimated\}_1,...,\quad (3)$$
$$(SS^1, SD_m),...,(SS^N, SD_m)\{(SS^1, SD_m),...,(SS^N, SD_m),$$
$$[delay, time\_process, \cos t]estimated\}_J$$

Satisfied with the evaluation result, the user starts the **data transmission phase** during which the data flow is attached with security-related information and sent through the network. In other words, the *rSSV* portions, one for each intermediate router, are attached into each data flow [9]. The *rSSV* portions in the data transmission phase are denoted as

$$\|rSSV\| = \bigcup_{j=1}^{J} \|rSSV\|_j \equiv \{(SS^1, SD_m),...(SS^N, SD_m)]\}_1,..,\quad (4)$$
$$\{(SS^1, DS_m),...,(SS^N, SD_m)\}_J$$

Upon an arrival at each router, a router picks up its associated *rSSV* portion and executes the security services requested individually. The requested services may be rejected if the service degree is downgraded from the one chosen by the querying user or if the service is entirely unavailable due to insufficient resources. After the security services were served, each router records the results by replacing the corresponding *rSSV* portion with the *aSSV* portion to report the querying user. Upon an arrival of the data packet, the receiver may reply either immediately upon an arrival of a data packet or after a delay for several data packets with an *ACK* packet.

## IV. DYNAMIC SOURCE ROUTING PROTOCOL (DSR)

The Dynamic Source Routing (DSR) protocol is a simple and robust routing protocol designed for use in multi-hop wireless ad-hoc networks of mobile node. The key features of DSR are [10]:

- **Source Routing** – The sender of a data packet knows the complete hop-by-hop route to the destination. These routes are stored in a route cache. Data packets sent by the source node carry the complete route in the packet header. Intermediate nodes forward the packet based on the route in its header. In most cases, the only modification that an intermediate node may make to the header of a packet is to the hop count field. The fact that all data packets are routed from the source has widely perceived security benefits. This may make DSR a strong contender as a routing protocol for many applications.
- **On-Demand** – DSR attempts to reduce routing overhead by only maintaining routes between nodes taking part in data communication. The source discovers routes on-demand by initiating a route discovery process only when it needs to send a data packet to a given destination. As a result, a significant amount of routing overhead is eliminated.

To send data to another node, if a route is found in its route cache, the sender puts this route (a list of all intermediate nodes) in the packet header and transmits it to the next hop in the path. Each intermediate node examines the header and retransmits it to the node indicated after its id in the packet route. If no route is found, the sender buffers the packet and obtains a route using the route discovery process described below.

A node that needs to send a packet to another node in the network but does not have a route to the destination in its cache initiates the route discovery procedure [11]. The source broadcasts a route request packet to all nodes within wireless transmission range of it. In addition to the addresses of the source and the destination nodes, a route request packet contains a route record, which is an accumulated record of the sequence of hops taken by the route request packet as it is propagated through the ad-hoc network during this route discovery. When a node receives a route request packet, it does the following:

- If it has seen the same request before (determined by the source address and the request sequence number), the request is ignored.
- If the destination address of the request matches its own address, then the route record in the packet contains the route by which the request reached this node from the source. A copy of this route is sent back to the source in a route reply packet by following the same route in reverse order.
- Otherwise, if the node has a route to the destination in its cache table, it creates a route reply packet with the route from its cache, and sends it back to the source using the path in the route request packet header, in reverse order. Such replies are called intermediate node replies.
- Otherwise, it appends its own address to the route record, increments the hop counts by one, and rebroadcasts the request.

## V. INTEGRATION OF SVV TO THE MANET

The SSV will be implemented into MANET network through DSR. Process of integration SSV is divides to following phases:

- *Specification of requested requiremenst for QoS and security and process of collecting QoS and security related information*
- *Process of adding SSV and Routing process*
- *Data transfer process*

In first phase, the user (source node) define own requirements for QoS and security. The information is collecting to the DSR header (Fig. 2). In this article we try to show only the way of adding and routing of modified header. In next work will by implemented new method of collecting requested information and will present decision model to the selection of the optimal way.

In second phase, the user (source node) who wants to make security-related service with another user (destination node) or server, sends the modified route request packet (Fig. 2c) to the all-intermediate nodes along the source node (Fig. 3). Request packet contains specified requirements for QoS and security.

If intermediate node is not destination node, adds its information about possibilities of provisioning requested services. This information is added to the relevant aSSV and then is send to another intermediate node (Fig. 3).

This process is called route discovery and correspond with Probing phase. Process of collecting of aSSV information is executed till destination node is found. When destination node is found, route is written to the source node route cache memory and destination node rewrites all aSSV information to the rSSV information. This packet is then called ACK packet. Then ACK packet is send back to the source node. The main reason of this operation is authenticity of intermediate nodes and it is necessary to select optimal way for transition of the data packet.

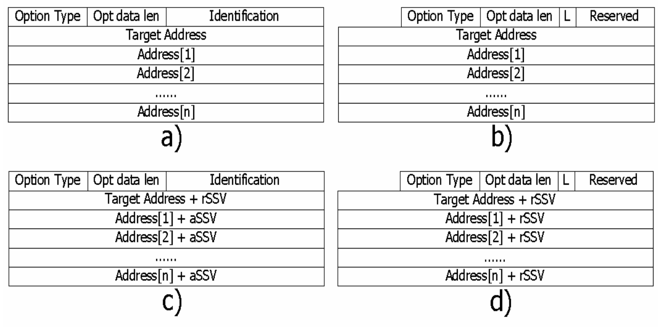



Fig. 2. Header packet: a) DSR Route Request RREQ, b) DSR Route Replay RREP, c) Modified Route Request RREQ with SSV, d) Modified Route Replay RREP with SSV.

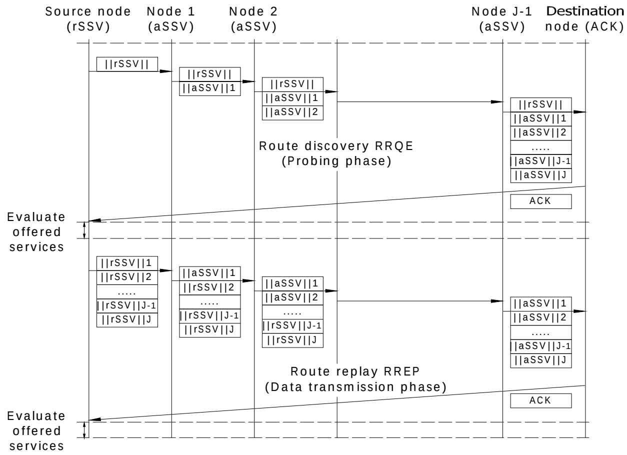


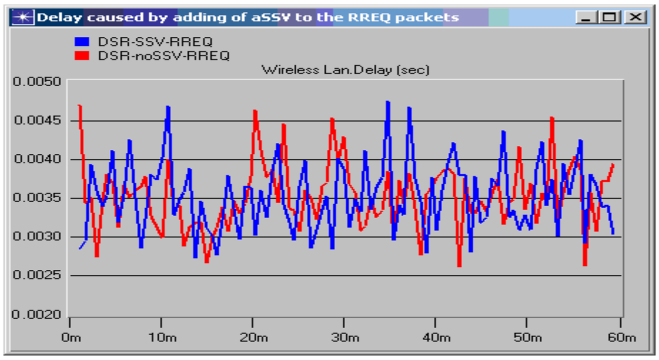

Fig. 3. Implementation of SSV to the DSR

## VI. SIMULATION AND RESULTS

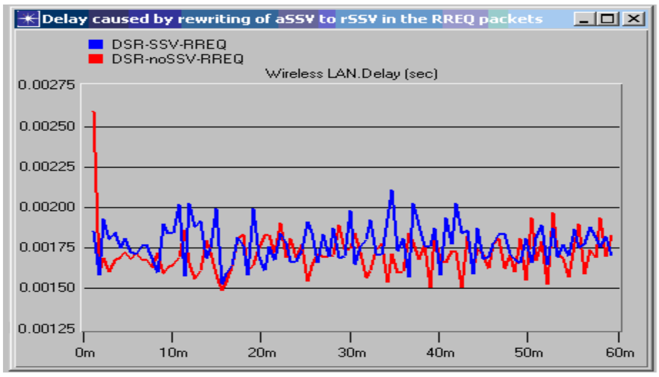
The simple simulation model was designed for testing of modification. The simulation was being implemented in OPNET simulator. A regular well-behaved DSR protocol was used as a reference for the performance analysis of the protocol modification. Two types of scenario, was considered (Fig.3):

- ***probing phase (route discovery)***
- ***data transmition phase (route replay).***

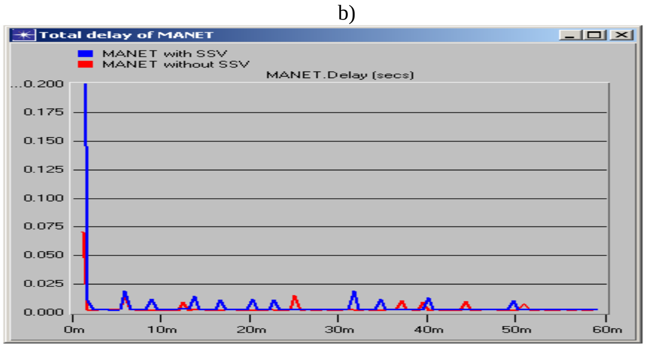
The delay caused by adding process of aSSV and rSSV was observed parameter of simulation. First scenario simulates only probing phase (route discovery process-RREQ). For comparison, scenario is separated to standard DSR route discovery algorithm and to modified SSV process. We didn't simulate information gathering process but only adding of aSSV to the RREQ packets (Fig. 2c). For simplification of simulations, random information about QoS and security was generated and then added to corresponding aSSV. The packet was then send to the another nodes. Every node adds information to the corresponding part of RREQ (Fig. 2 c,d). In second scenario, the transmition process was simulated. The corresponding information stored in aSSV was rewriting to the rSSV. Process of modification is described in Section V. The results are displayed on Fig. 4.




a)




b)




c)

Fig. 4. SSV delay: a) Delay of aSSV to the RREQ packets, b) Delay of rSSV to the RREP packets, c) Total delay of MANET

## VII. Conclusion

In this article, modification of the routing protocol DSR with SSV was presented. Integration SSV to DSR, is necessary for our next work, because it is good way how to add secure-related information from the nodes.

Two processes were simulated: process of adding aSSV and rewriting process of rSSV. The standard delay is compared with delay caused by SSV. As we can see, processes of adding information (aSSV) do not significantly affect delay of transfers the packet. Delay is caused by modification of DSR RREQ and RREP packet, where QoS and security related information are added.

In next work, we will create classification of security services and security degrees. We will specify methods for extending architectures include security and QoS and will proposing definitions of security that reflect the need of the users and QoS. We will analyze and create classification of user-oriented services.

In future work we will try to design the enhanced model integrated security as one of QoS parameters. We will define new evaluate paradigm of integration security as one of QoS parameters. Users will have ability to select level of QoS that reflect their actual requirement and new model will reflect user demand to QoS and security.

## Acknowledgment

Research described in the paper was financially supported by MobilTel No.APVT-20-029004 and also by COST 2100 - Pervasive Mobile & Ambient Wireless Communications.

## References

[1] S. Chakrabarthi, A. Mishra, *QoS Issues in Ad Hoc Wireless Networks*, IEEE Communications Magazine, Feb. 2001. pp 142 –148.

[2] P. Sakarindr, N. Ansari, R. Rojas-Cessa, S. Papavassiliou: *Security-enhanced quality of service (SQoS) networks: a network analysis*,IEEE Military Communications Conference, October 2005.

[3] E. Crawley, R. Nair, B. Rajagopalan, H. Sandrick: *A Framework for QoS Based Routing in the Internet*. RFC 2386, August 1998..

[4] M. Krasnovský, V. Wieser.: A Performance of Wireless Ad-Hoc Network Routing Protocol. In: Proceedings of 17th International Conference "Radioelektronika 2007", Brno University of Technology, 24.-25.4.2007, Brno, pp.173-175, ISBN: 978-80-214-3390-8.

[5] E. Spyropoulou, T. E. Levin, and C. E. Irvine: *Calculating costs for quality of security service*. In Proceedings of the 16th Annual Computer Security Applications Conference (ACSAC'00), pages 334–343, New Orleans, Louisiana, USA, December 11–15, 2000.

[6] C. E. Irvine and T. E. Levin: *Toward quality of security service in a resource management system benefit function*. In Proceedings of the 2000 Heterogeneous Computing Workshop (HCW'00), pages 133–139, Cancun, Mexico, May 1, 2000

[7] J. Yang, J. Ye, S. Papavassiliou, and N. Ansari: *A flexible and distributed architecture for adaptive end-to-end QoS provisioning in next-generation networks*, IEEE Journal on Selected Areas on Communications, Vol. 23, Issue 2, February 2005, pp: 321-333.

[8] P. Sakarindr, N. Ansari, R. Rojas-Cessa, S. Papavassiliou: *Security-enhanced quality of service (SQoS) networks*, IEEE Sarnoff Symposium on Advanced in Wired and Wireless Communications, April 2005, pp. 129-132.

[9] J. Yang, J. Ye, and S. Papavassiliou: *A new differentiated service model paradigm via explicit endpoint admission control*, in Proc. SCC2003, Jun. 2003, pp. 299–304.

[10] D. Johnson, D. Maltz, and J. Broch: *DSR: The Dynamic Source Routing Protocol for Multihop Wireless Ad Hoc Networks in Ad Hoc Networking*, Addison-Wesley, 2001, pp. 139–172.

[11] M. Krasnovský, V. Wieser.: A Performance of Wireless Ad-Hoc Network Routing Protocol. In: Proceedings of 17th International Conference "Radioelektronika 2007", Brno University of Technology, 24.-25.4.2007, Brno, pp.173-175, ISBN: 978-80-214-3390-8.

# Cross – devices testing of efficiency in data – parallel applications

*Ján PERHÁČ*

Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

jan.perhac@tuke.sk

***Abstract*** **– With the development of new technologies it is natural to perform cross – device testing. One reason for doing this is the rise of GPGPU solutions in data – parallel applications. GPGPU is a new area of informatics which utilizes the computation power of GPU processors for non – graphic tasks. This paper describes the GPGPU problematic and presents results of cross – devices testing of efficiency in data – parallel applications.**

***Keywords*** **— GPGPU, CUDA, Brook**

## I. Introduction

GPU or graphic processor units are representing nowadays a massive – parallel stream SIMD – type processor, which computation power is as high as hundreds of GLOPS. The high number of parallel computational units in GPUs, together with fastest memory units with high bandwidth represents in present a device for both graphical and non graphical operations. The increase of computational power of GPU units is several fold higher than the increase of computational power of central processor units (or CPU). GPU processors are suitable for effective solving of problems, which can be represented as data-parallel computations. This means that the same program is parallel executed on multiple data elements with high arithmetical intensity, what means that the number of arithmetical operations is higher than the number of memory – related operations. Because the same program is executed on every data element, the need of sophisticated management of operations flow is lower. The need of dealing with latency of access to the memory through big cache registers is possible to hide / compensate exactly by computations with high arithmetical intensity. Data – parallel executing is mapping data elements into threads of parallel execution (or parallel processing threads). Many applications, which are handling big data sets (for example arrays) can use model of data – parallel programming for speeding of these computations. In 3D rendering big sets of vertexes and pixels are mapped into parallel threads. Similarly, applications for audio or video processing are also mapped into parallel threads. In fact, many algorithms outside of area of computer graphics are speeded by using of data – parallel execution.

## II. Analysis of GPGPU problematic

The programmability of graphical hardware in the level of GPGPU application is a subject of intensive research of many institutions. First and for most, there are producers of graphical hardware. Many of them are rivals to each other in the GPGPU area in contrast to their cooperation in developing and / or enhancing of graphical standards and application programming interfaces (or APIs). Main consequences of this rivalry are proprietary solutions in GPGPU area which means that there is no possibility of application of solution of one producer on hardware of other producers. The primary research goal of many academic and scientific institutions (but also research and development laboratories of many companies from private sector) is to minimize or directly to annul the proprietary meanings of graphics vendors solutions by creating of special GPGPU solutions based on the creation of meta-languages to existing graphical standards and their APIs. But this kind of solution is less effective than native programming of graphical hardware. The arrival of new technologies on the level of hardware has meant also the arrival of new possibilities on level of software. The number of (meta) languages for direct GPU programming from their producers or other scientific institutions has emerged into high numbers of solutions over time. This increase, together with difficulty or different effectiveness of individual languages / meta - languages is representing for GPU developers also the need of analysis of choosing the right language for concrete GPGPU project. For a long time, the utilization of graphical hardware for GPGPU applications was problematic in that fact, that the architecture of GPU processor is strongly oriented on execution of graphical operations, what means that the GPU architecture is focused on manipulating with graphical objects in graphical process. So to fully utilize the computation – power – potential of graphical hardware for GPGPU, developers had to think about problem solution in steps mostly typical for computer graphics than for classical programming. But the progress in area of increasing of GPU effectiveness in graphical operations advanced into that state, that the internal architecture of graphical hardware was significantly changed not only to take advantage of GPU processing power in graphical operations but also in general – purpose – computations on graphical hardware.

## III. THE POSSIBILITIES OF GPGPU PROGRAMMING

GPGPU applications are possible to develop with the arrival of third generation of graphical hardware. First GPGPU applications were based on the use of graphical standards Direct3D or OpenGL, including pixel or vertex shader languages for computation of first data – parallel tasks like dynamic fluids simulation. The development process of this generation of graphical adapters requested the transformation of concrete numerical solution into graphical terminology and elements (for example texture = array, rendering = computing, shader unit = computation unit). These necessary transformations resulted into slow development of GPGPU applications and that is why many research groups, mostly from academic area, tried to help developers in programming of graphical adapter with creating of meta – languages with syntax similar to language C or C++. These meta – languages are representing some kind of extension to existing graphical standards, see Fig 1.

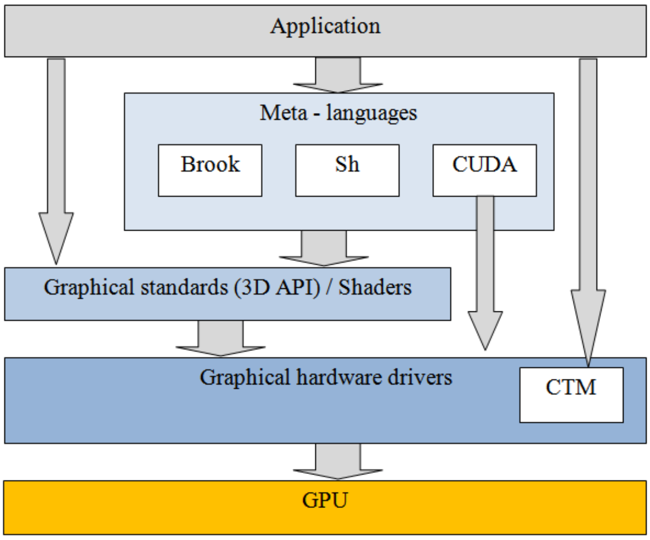



**Fig. 1 Existing possibilities of programming GPGPU applications**

Enumerated chronologically, basic possibilities of programming GPGPU applications are:

- **Graphical standards OpenGL and DirectX** – programming by using of these standards is requiring execution of arithmetical operations by using difficult numerical – operations – mapping onto graphical objects. During this process, the array is considered as texture, parallel loop as quadroon, ALU as shader, computation as rendering. This kind of GPU programming has tendency to be time – consuming, code – un – transparent with many difficulties in distribution and execution of code instructions or data.

- **Shading languages –** are representing a middleware between hardware and concrete graphical standard (OpenGL or DirectX). Shading languages are often used in the fourth generation of graphical hardware and they are designated only to program vertex or pixel shader units and are heavy - graphical – standard dependent. Examples of shader languages are Cg language, HLSL language and OpenGL shading language. Their semantics is quite the same as of language C but is containing many elements from graphical standards.

- **Metalanguages –** are representing a middleware between application and graphical standard (OpenGL or DirectX). Their primary task is to simplify the programming process with the help of transforming of C-like or C++like syntax into concrete shader language or into code of graphical standard. This transformation is done by CPU processor. The meta - language Brook is developed on Standford University and is available for free.

- **Metalanguage nVidia CUDA** – is representing a proprietary possibility to program GPU processor respectively unified shader units by using of package of functions and supporting libraries. CUDA represents a C++ compiler which also contains many extensions to the C++ language, which allows writing code in C++ language that is possible to be executed on GPU processor. These extensions are supported by software libraries and specialized OS – related device driver. CUDA is a product of nVidia and provides whole bunch of tools for code debugging like compiler, debugger and tester.

## IV. CROSS – DEVICES TESTING OF EFFICIENCY IN DATA – PARALLEL APPLICATIONS

Considering information characterized in previous chapters, there were done several performance tests to gain real results of difference between graphics processor and central processor unit in solving of data – parallel tasks. On the basis of test performed in [3], the available source code was modified and performed also with using the meta – language Brook on the same hardware, but also on less powerfull graphical hardware. Characteristics of tested components are in Tab. 1. Main goals of these tests were:

- **Test no.1 –** to experimentally confirm or disconfirm the superiority of numerical / performance potential of graphical hardware in the process of solving data – parallel tasks against performance potential of general central processor units

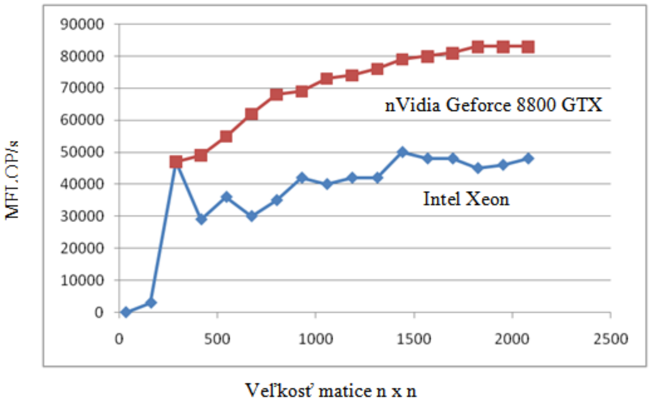
- **Test no.2** – to determine numerical / performance difference between two graphical cards, one from high – end category of fifth generation and one from mid – end category of fourth generation

- **Test no.3** – to confirm / disconfirm the presumption, that using of proprietary meta - language will be faster and more effective than using of non – proprietary meta-language (in this case the meta - language Brook)

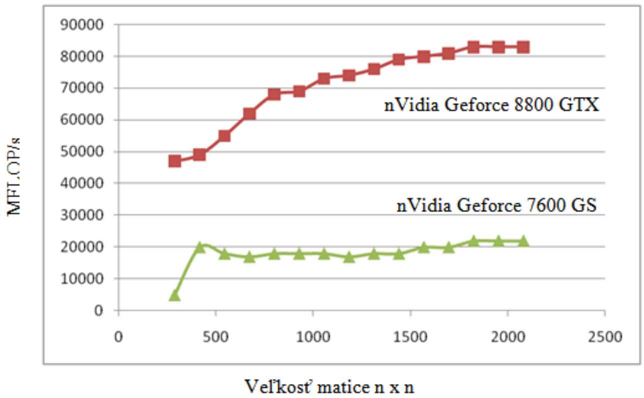
These tests are part of research done on department of computers and informatics, FEI TU Košice, Slovakia.

| | Intel Xeon X5355 | nVidia Geforce 8800 GTX | nVidia Geforce 7600 GS |
|---|---|---|---|
| No. of ALUs | 4 | 128 | 12 |
| Processor frequency | 2660 MHz | 575 MHz | 400 MHz |
| GFLOP/s | 84 | 500 | 22 |
| Processor cache | L2: 8MB | shared: 256KB | ? |
| No. of transistors (milions) | 582 | 681 | 178 |

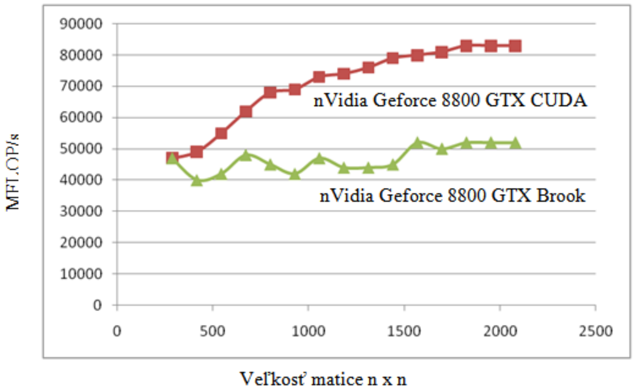
**Tab. 1 Characteristics of tested devices**




**Graph 1 Results of test no. 1**




**Graph 2 Results of test no. 2**




**Graph 3 Results of test no. 3**

## CONCLUSION

Results of tests no. 1, 2 and 3 fulfilled expected assumes. Correctly loaded graphical hardware is in data – parallel tasks more efficient and powerful than CPU processor, high – end g graphical hardware is more powerful than mid - end graphical hardware and the use of proprietary solution resulted in better performance and results than the use of non – proprietary solution

## REFERENCES

[1] Pharr M., Fernando R.: GPU Gems 2: Programming Techniques for High-Performance Graphics and General-Purpose Computation, Addison-Wesley Professional, 2005, ISBN 0321335597

[2] NVIDIA Corporation: NVIDIA CUDA Compute Unified Device Architecture - Programming Guide, 2007

[3] Heinecke A.: A Short Introduction to nVidia's CUDA, Ferienakademie, 2007, Dostupné na internete: http://home.in.tum.de/~heinecke/docs/heinecke_fa2007.pdf

[4] Vokorokos L.: Digital Computers Principles, Typotex 2004, Budapest, ISBN 9639548 09

[5] Vokorokos L. a kolektív: Princípy počítačového inžinierstva, Elfa, Košice, Slovakia, 2008, ISBN 978-80-8086-075-2

**SUPPORTED BY VEGA PROJECT No. 1|4071|07**

**AND APVV 0073-07**

# Consideration of Shared Redundancy in Networked Control Systems

*Ján GALDUN*

Dept. of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic

jan.galdun@tuke.sk

***Abstract*—The paper presents common cascade control architecture where specific kind of redundancy could be considered. There are different approaches how to increase the reliability of networked control systems. Common approach uses redundant components in control system i.e. passive or active redundancy. However, in case of cascade control there is advantage proposed by its structure which allows profit from another type of redundancy – shared redundancy (quasi-redundant components). This type of redundancy offers several important advantages such as minimizing the number of components as well as reliability increasing.**

***Keywords*— Networked control systems, Reliability, Shared redundancy, Quasi-redundant components.**

## I. Introduction

Several different approaches exist in order to increase the systems reliability. A classical technique consists in designing a fault-tolerant control [6] where the main aim is to propose a robust control algorithm. Guenab and others in [3] deal with this approach and reconfiguration strategy in complex systems, too. Another approach uses active or passive redundancy to increase reliability of the systems.

The question of reliability is more important when networks are used for transmission of required values among control system components as sensors, controllers and actuators – thus in case of distributed control systems (DCS) or networked control systems (further only NCS). There are new phenomena which appear such as random delays, data losses, asynchronisms, problems of components initialization [4] which have influence on the reliability of the NCS, too.

Some of the conventional control structures could be used not only to accomplish a required control function of the control system, but also to profit from their structure in term of reliability increasing. Suitable control structure is a cascade control where its structure naturally proposes advantages which could be used for reliability increasing. The main advantage which we would like to describe and profit from is shared redundancy [6] or quasi redundant components.

Further is paper organised as follows. In first part is introduced the basic CC structure as well as different types of redundant components are presented. In the next part are presented networked cascade control topology considered with one control network. In final part the results of the simulations are presented - influence of using the shared redundancy (quasi-redundant components) [7] to system's reliability increasing. At the end is proposed a short conclusion.

## II. Control Structure Implementation

### A. Conditions and description of application of cascade control

With using cascade control structure there are several constraints. In order to application this control structure must be fulfilled one of these conditions [1]:

- Controlled system must contain subsystem (secondary subsystem) that directly affect to primary system.
- The gain of the secondary subsystem (encapsulate in actuator) is non-linear.
- The disturbance is directly measured.

Usually for secondary subsystems there is a condition of faster dynamics than primary process. This condition must not be fulfilled [1]. However, when dynamics of the primary process is faster after that some modifications of conventional cascade structure and control laws must be provided. Primarily we suppose that the secondary subsystem has faster dynamics than the primary system.

As we can see from conditions of cascade control the system must contain a primary system $F_P(s)$ and a secondary subsystem $F_S(s)$ that affects directly the primary system. In Fig. 1 is shown the equivalent of a cascade structure with the control components included (sensors, controllers and actuators) which are connected with the network – NCCS.

In cascade control structure there are inner and outer control loops. Inner loop contains a secondary controller (SLAVE – CS) and the secondary subsystem $F_S(s)$. The main function of this control loop is to eliminate disturbances through secondary measured state value with faster dynamics. The outer loop encapsulates the inner loop, the primary controller (MASTER - CM) and the primary process $F_P(s)$.

Thus, using the cascade control structure has the main objective to eliminate the influence of the disturbances on the primary process through the secondary subsystem and in consequence to improve the parameters of quality of control. However, we will profit from other features provided by this control structure.

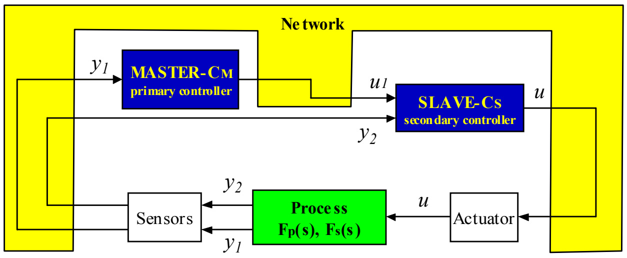



Fig. 1 Structure of the networked cascade control system (NCCS) with one network and two quasi-redundant controllers (primary Master and secondary Slave)

### *B. Active and passive redundant subsystems*

The primary aim of redundant components is to compensate the component failure in order to achieve its mission. There are two basic types of redundancy [5]:

- Passive (cold),
- Active (warm/hot).

Passive redundant component starts operate at the moment when the primary component has failed. This redundancy is often conditioned by the component initialization (in the worst case by the system restart) which could not be applied in some critical control system application.

On the contrary, active redundant components are in operational mode at the same time the primary component is working. Two cases of life time of the system with active redundant components can be envisaged:

- the first case is when the primary component failed whereas the redundant component continues accomplishes its mission,
- the second case is when the redundant component failed first.

We can suppose that life mission time of the system with passive redundant components could be longer than life time of the system with active redundancy. However, the mentioned problem with component initialization could decrease the number of application where this type of redundancy could be implemented, this depends on the PFD (Probability of Failure on Demand).

In further text we considered another type of redundant components which are not primary determined as redundant but they are able to replace some mission if it is urgently required. This type of redundancy is referred as shared redundancy [7].

### *C. Quasi-redundant components*

In the similar sense as shared redundancy [7] we use term quasi-redundant components in networked cascade control structure. These parts of the system are not primary redundant. Each quasi-redundant part accomplishes its primary mission when the system is in its nominal state. However its functionality allows compensate the failure of another subsystem with similar mission.

The quasi-redundant components are not primary determined as active redundant subsystem because each one has its own mission which must be accomplished. Only in case of failure this type of redundancy should be used.

In order to profit from quasi-redundant architecture the existing hardware components are used. This paper presents a simple networked cascade structure which is naturally composed of potential quasi-redundant components as controllers. When a networked control structure is composed of more than one network then we could use shared redundancy approach too.

### *D. Quasi-redundant components in NCC structure*

In figure 1 is shown a networked implementation of the basic cascade structure. There are two controllers which could be considered as quasi-redundant components. The controllers follow their primary mission, stabilization and performance optimization of the controlled system. Therefore, in regard to the same hardware it allows share the computing capacity and executes different tasks. Thus, in both controllers could be implemented both control tasks – for primary and secondary subsystems.

In non-failure mode the primary task is executed in both controllers. However, in case of controller's failure (primary or secondary) non-failed controller starts execute both tasks and computes actuating value for primary as well as secondary subsystem. In this case we can suppose two scenarios.

The first one supposes that controller is able to execute all necessary algorithms within required sample periods. In this case the behavior of quasi-redundant component is identical as in case of active redundant components. Thus, in case of failure one of the components the second one takes care about its mission until its failure.

The second case when time to execute both necessary tasks is grater than required sample period the controller will cause the delays which have significant influence to system stability [2], [4]. Thus, we can suppose decreasing of the component's reliability by specified value $d_R$ (decrease factor) from its nominal reliability $R_n$ considered at beginning.

The results of the both cases are shown in next part of the paper.

## III. Simulations And Results

Presented networked cascade control architecture (Fig. 1) was modelled by using Petri nets. This tool was chosen thanks to its ability to model different types of complex systems and dependencies within them. To provide the reliability analysis the Monte Carlo simulation (further only MCS) method was used.

The reliability R(T) all of the components included by control system (Sensors S1, S2, Controllers CM, CS, Actuator and Networks 1) within one sample period was seted up to $R_n(T) = 0.999$. Thus, unreliability or probability of failure of the components is $10^{-3}$ for one sample period T. Sample period is represented by 10 simulation units. The results of the simulations thus reliability are shown in figure 2 and table 1.

In figure 2 there are shown the results of the simulation networked cascade control structure with two quasi-redundant controllers. The simulation were provided also with different values of decrease factor $d_R$. Thus, infuence of this factor to final system reliability is shown as well. In figure 2 there are shown the reliability curves of system with quasi-redundant subsystems ($d_R = 0$) and without them ($d_R = 0.999$). Between these two values of decrease factor are shown the three curves which represent final reliability of the system with decrease factors as $2.10^{-3}$, $10^{-2}$, $59.10^{-3}$. We can see that decreasing the component's nominal reliability by decrease factor equal to $59.10^{-3}$ which represents approximately 6% of nominal reliability $R_n$ has significant influence to decreasing the final

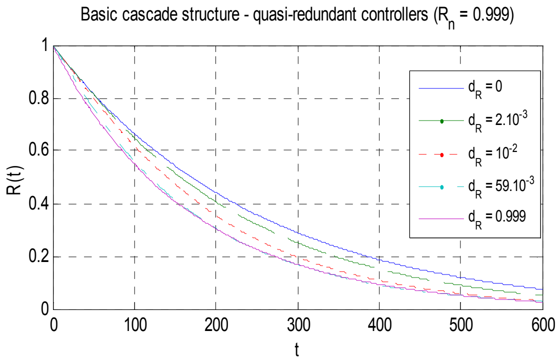



Fig. 2 Simulation results of the basic cascade control architecture with two quasi-redundant controllers

TABLE I
MEANS TIME TO FIRST FAILURE OF THE NCCS

| $d_R$ | MTTFF [T] |
|---|---|
| 0 | 236.4 |
| $2.10^{-3}$ | 213.07 |
| $10^{-2}$ | 186.76 |
| $59.10^{-2}$ | 169.04 |
| $R_n = 0.999$ | 167.03 |

reliability of the system. The results are a little bit better than in case of system without redundant components ($d_R$ = $R_n$ = 0.999), but we could say that almost the same.

In table 1 there are shown the values of MTTFF (Means Time to First Failure) of the system for all values of the decrease factor. We can see that influence of the decrease factor to life time of the system until first failure is very significant.

## IV. Conclusion

The paper shows the advantages of using NCC structure in control from the reliability point of view. The conventional cascade control structure is shown within conditions of networked control systems as naturally suitable to profit from quasi-redundant subsystems as networks, controllers and potentially sensors if physical process allows it. Despite of some constraints for using this type of control is cascade architecture widely used in industrial control applications. Hence, only the reconfiguration algorithm should be implemented to profit from quasi-redundant subsystems.

Despite of advantages shown in this paper it is needed to remind that cascade control structure can be applied only for some processes which accomplish several basic conditions described above.

The main advantages of the quasi-redundant components could be summarized as follows:

- The system composes only of necessary components (parts) for following the primary mission of the system whereas higher system reliability is ensured without using any additional active redundant components.
- Following the first point we could suppose less number of the components used for safe the control mission. Thus, economic aspect could be very significant.
- Prevention of system's critical failure also when subsystems have not sufficient hardware capacities.

## References

[1] C. Brosilow, J. Babu, "Techniques of Model-Based Control", Chapter 10, Prentice Hall, ISBN: 978-0-13-028078-7, 2002

[2] J. Galdun, R. Ghostine, J. M. Thiriet, J. Liguš, J. Sarnovský, "Definition and modelling of the communication architecture for the control of a helicopter-drone", 8th IFAC Symposium on Cost Oriented Automation, 2007

[3] F. Guenab, D. Theilliol, P. Weber, Y.,M. Zhang, D. Sauter, "Fault-tolerant control system design: A reconfiguration strategy based on reliability analysis under dynamic behaviour constraints", 6th IFAC Symposium on Fault Detection, 2006, pp. 1387-1392.

[4] J. Ligušova, J.M. Thiriet, J. Liguš, P. Barger, "Effect of Element's Initialization in Synchronous Network Control System to Control Quality", RAMS/IEEE conference Annual Reliability and Maintainability Symposium, 2004.

[5] J.R., Pimentel, M. Salazar, "Dependability of Distributed Control System Fault Tolerant Units", The 28th Annual Conference of the IEEE Industrial Electronics Society, vol. 4, IECON '02, ISBN: 0-7803-7474-6, 2002, pp. 3164- 3169

[6] J. T. Spooner, K.M., Passino, "Fault-Tolerant Control for Automated Highway Systems", IEEE Transactions on vehicular technology, vol. 46, no. 3, 1997, pp. 770-785

[7] J. Wysocki, R. Debouk, K. Nouri, "Shared redundancy as a means of producing reliable mission critical systems", 2004 Annual Symposium – RAMS - Reliability and Maintainability, 2004, pp. 376-381

# Model of the Autonomous Tracked Vehicle

*František BANÍK*

Dept. of Electrotechnical, Mechatronic and Idustrial Engineering, FEI TU of Košice, Slovak Republic

fero@banik.sk

***Abstract*—Autonomous navigated vehicles represent a special class of robots. Many theories for autonomous navigation were designed in the last decade. Some of them were tested on real vehicles. The others were only simulated. The article describes the model of the autonomous vehicle, in which these theories can be applied. Dimensions of the vehicle were designed for the motion inside administration buildings. Possibilities for the construction and dynamic control are main purposes of this article. Personal computer was used for navigation computations. Communication between the computer and the vehicle was designed through the wireless system.**

***Keywords*—Autonomous vehicle, Drive control, Autonomous navigation**

## I. Introduction

In recent years big boom was reached in the employment of robots in industry. These robots are often static machines. Operational areas are known before programming of these robots. Static robots do not need any distance sensors or vision systems (safety systems against to the reaction with other machines), because there is guaranteed non reacting environment. On the other hand there are moving machines, e.g. for transfer production material. The accident danger must be considered in the construction and programming of these robots. The induction line is usually used for the navigation of the machine inside production hall. New methods of the navigation do not need induction lines. But for machine employment is very important to test these methods on the real machine. Described model has been real constructed. The weight of the vehicle requires speed and torque control loops for the motion. Programmable unit allows testing a few types of control loops. Superior control system is realized in personal computer, which enables usage of standard computing equipments. Connection between the computer and the vehicle is created through the wireless system $2,4GHz$. Additional measurement modules can be mounting in the back side of the vehicle construction. Dimensions of the vehicle was adapted for use inside laboratories and classrooms.

## II. Construction of the Model

### A. Chassis

Chassis is the main part of the autonomous vehicle. This part is important for the size of the vehicle and the motion in space. It is important to define basic environment in which the vehicle will be used. Basic parameters of this environment are the surface, ramp angle, the surface of obstacles and maximal obstacle height. Motion systems there have been designed for different environments based on rotary wheels, arms and other special systems. Basic environment for this vehicle was defined as the inside space of administration buildings. The surface is linoleum or industrial carpet. Ramp angle is permanently zero. Obstacles are walls and furnitures. Obstacles which need not limit motion are door sills and stairs with low elevation. Chassis moving on two independent tracks was designed for this environment. This motion system was chosen mainly for simply control of motion direction and zero radius needed for turn back of the vehicle. The space needed for turn back is the diagonal of chassis. Diagonal length is main parameter of chassis, which is limited by width of doors and free space inside buildings. Chassis constructed on these premises is shown in figure 1.

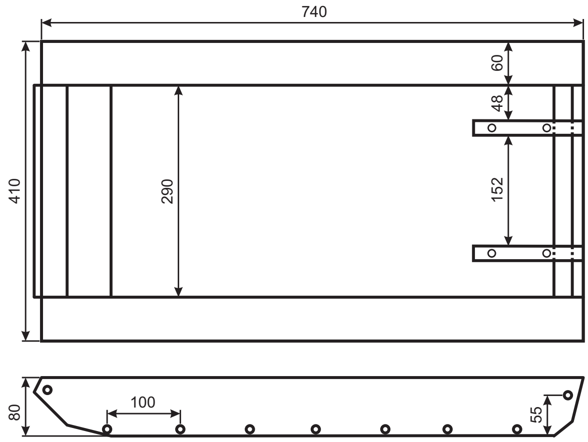



Fig. 1. Chassis dimensions

Chassis of German tank Leopard 2 A5 was used as a model. Scaling factor is 1:10. There are seven wheels in each track. Each of wheels has an independent suspension. Motor torque is transmitted to the track by the rosette in the backside of the track. The track consists of three sort of segments. The first segment interlocks to the rosette and transmits the power from the rosette to the track. The second segment transfers the vehicle weight between wheels and floor. The third segment keeps the track into wheels direction.

### B. Drives Dimensioning

Requirements for drives result from the construction of chassis. Independent speed, torque and direction are required for each track. Dimension of the drive results from the weight and dimensions of the vehicle. Required properties for the vehicle are in table I. For vehicles with the weight about 40 kilograms can be used gas engine with mechanical jack in the box. But exhausts disallow their usage inside buildings. For mentioned vehicle has been used direct current (DC) motor with permanent magnets.

Additional modules have been required for the measurements and next research in this vehicle. The overload capacity is limited by acceptable overload of commutator for DC

TABLE I
REQUIRED PROPERTIES OF THE VEHICLE

| | | | |
|---|---|---|---|
| Weight of the single vehicle | $m$ | 40 | $[kg]$ |
| Weight of the vehicle with additional module | $m_p$ | 50 | $[kg]$ |
| Maximal speed for zero elevation | $v_c$ | 1.2 | $[ms^{-1}]$ |
| Maximal speed for maximal elevation | $v_\alpha$ | 0.4 | $[ms^{-1}]$ |
| Maximal elevation angle | $\alpha$ | 60 | $[°]$ |

motors. Motor without compensating winding can work with max torque $M_{Pmax} = 2M_n$. For one drive, in system with two independent drives for each track, there has been considered only half of the vehicle weight from the table I. In the table II are parameters for one drive dimensioning. These independent drives are connected in two points. The first is mechanical connection through the floor. The second is the electrical (program) connection in programmable logic controller (PLC).

TABLE II
ADDITIONAL PARAMETERS FOR ONE DRIVE DIMENSIONING

| | | | |
|---|---|---|---|
| Coefficient of rolling friction (total) | $f$ | 0.02 | |
| Frontal area of vehicle for one track | $S_{half}$ | 0.094 | $[m^2]$ |
| Coefficient of aerodynamic resistance | $c_W$ | 0.4 | $[kgm^{-3}]$ |
| Necessary time for accel. to max speed | $t_{max}$ | 2 | $[s]$ |
| Acceleration in max speed achieving | $a_{max}$ | 0.6 | $[ms^{-2}]$ |
| Acceleration in moving up speed achieving | $a_\alpha$ | 0.2 | $[ms^{-2}]$ |
| Efficiency of motor | $\eta_m$ | 0.9 | |
| Efficiency of gear box and gears (total) | $\eta_p$ | 0.8 | |
| Acceleration of gravity | $g$ | 9.81 | $[ms^{-2}]$ |

Equations used for dimensioning are described in [1]. Elevation speed is lower than straight way speed. Power for elevation is bigger than total power needed for straight way motion. For this reason there was considered only the worst state, when the vehicle is moving up with additional module. Total motor power was calculated as:

$$P_{M_\alpha} = \frac{v_\alpha(F_{v_\alpha} + F_{w_\alpha} + F_{g_\alpha} + F_{a_\alpha})}{\eta_p \eta_m} \tag{1}$$

Where $F_{v_\alpha}$ is friction power from rolling friction of all seven wheels:

$$F_{v_\alpha} = \frac{mgf}{2}(1 + 0.022v_\alpha) \tag{2}$$

$F_{w_\alpha}$ is aerodynamic power of air reacting against to the vehicle:

$$F_{w_\alpha} = c_W S_{half} v_\alpha^2 \tag{3}$$

$F_{g_\alpha}$ is elevation power of gravity:

$$F_{g_\alpha} = \frac{mg}{2}\sin\alpha \tag{4}$$

$F_{a_\alpha}$ is power needed for acceleration from zero to maximal elevation speed $v_\alpha$:

$$F_{a_\alpha} = 1.1\frac{m}{2}\frac{dv}{dt} = 1.1\frac{m}{2}a_\alpha \tag{5}$$

Power calculated for this state is $P_{M_\alpha} = 124W$. Energy recuperation was required for the drive construction. MCC 24MP4N type of the drive was chosen from [2] according to this specification. Parameters are described in Appendix A. Gear box is the type of maintenance free closed system. Rosette shaft is placed in back side of chassis. Direct gear was needed for connection gear box to rosette shaft, because there is size collision between chassis and gear box. Efficiency of this gear was included in to parameter $\eta_p$ in table II. Motors mounting is shown in figure 2.

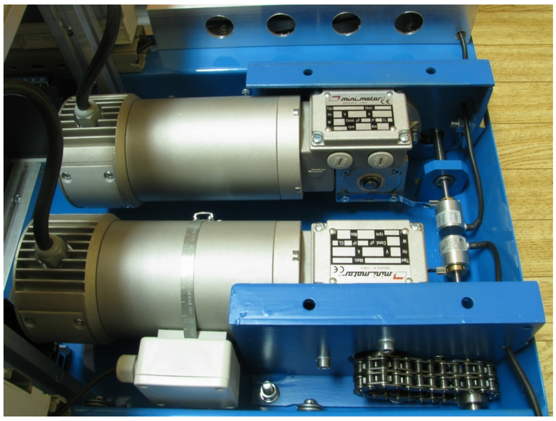


Fig. 2. Motors mounting in back side of chassis

## C. Accumulator and Power Supply

Many kinds of accumulators can be used for power supply of the vehicle. Standard automation circuits were used for electrical construction, where $24V$ power supply is used. This voltage is also used by dimensioned drives. Lead acid accumulator $U_{cc} = 24V$ was chosen considering required power supply for drives and other circuits. Practically all lead acid accumulators can be divided in to two groups. Starter accumulators with possibility to short big currents represent the first group. Discharging cycles are from 100 to 300 times for deep discharge. Traction accumulators are the second group. These accumulators enable long time small current discharging. They usually consist of the positive pipe electrodes. Maximum discharging is 80% of capacity. Definition of the max vehicle range was needed for dimensioning of accumulator. Range was defined as $R = 10km$ and calculated as:

$$R = 0,8v_c t_{dis} \tag{6}$$

Where $v_c$ is maximal straight way speed from table I. $t_{dis}$ is discharging time from 100% to 0% capacity of accumulator. Calculated time for defined range $R$ is $t_{dis} = 2,89h$. Total power $P_{total} = 158W$ was considered in this case. This consists of power for drives $P_{v_c} = 108W$ and power for other circuits $P_c = 50W$. Total current $I_{total}$ was calculated as:

$$I_{total} = \frac{P_{total}}{U_{cc}} \tag{7}$$

Ratio 8 for discharging time $t_{dis}$ has been deducted from discharging characteristic:

$$I_{total} : I_5 \tag{8}$$

Where $I_5$ is rated current for five hours discharging. Current $I_5 = 3,35A$ is direct depending on battery capacity $C_{cc}$:

$$C_{cc} = 5I_5 \tag{9}$$

Panasonic 24/17 type of accumulator was chosen considering to this $C_{cc}$. It is $17Ah$ maintenance free accumulator. Power supply of circuits was distributed to individual power lines. Each power line is controlled by PLC. Also PLC measures accumulator voltage and then controls power mode.

Three power modes were defined. Power management can be expanded to complex system with theory from [3].

*1) Stand by:* Mode for the lowest discharge power. E.g. for the vehicle waiting for the next use.

*2) Normal:* Mode when the vehicle is carrying its task. Every power lines can be switched on.

*3) Power safety:* Mode for low accumulator capacity (voltage). Unnecessary power lines are switched off and cannot be again switched on. E.g. temperature sensors, fans are not important for the safety comeback to user.

## III. Drive Control

### *A. Requirements*

Independent speed and torque was required for each track as was described in chapter Drive Dimensioning. But the cooperation of both drives has been needed for achieving motion in space. Six states were defined for basic motion. The first group represents forward motion with turns to left and right. The second group represents backward motion with left and right turns. Turn radius $r_T[m]$ must be adjustable from superior control system. The last group represents motion without change the global position data. The global position angle $\delta_G[rad]$ change is required in this group. Practically vehicle stays on one point. Tracks have a same speed with opposite direction. Also $\delta_G$ can be changed positively or negatively.

### *B. Control Devices*

Devices in combination shown in figure 3 were chosen according to required motion states.

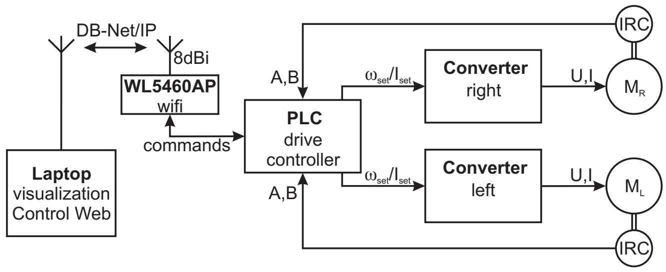



Fig. 3. Control devices scheme

Where converter L and R are types of Maxonmotor ADS 50/10. DC converters were designed considering drive and accumulator dimensioning. It is powerful pulse wide modulation (PWM) servoamplifier for permanent magnet-activated DC motors from 80 to approx. 500 watts output power. Switching frequency is $50kHz$. Concrete electrical and mechanical data are described in [4]. Module housing used in the vehicle is shown in figure 4.

Incremental sensors (IRC) were mounted on rosettes shafts. On rosette shaft mounting was chosen for elimination of gear clearance. IRC is 1024 impulses per rotate. Only A and B signals have been used in PLC. Zero point of IRC is not needed for the track control. Other signals have been converted from 24 volt level to 5 volt transistor-transistor level (TTL). TTL level is required for converter input, when this control loop is used. PLC includes program for control converters. Also turn radius commands $[v; r_T]$ are accepted by PLC. Command format consist of speed $v$ and turn radius $r_T$. Format is changed only when $\delta_G$ change was required. Zero speed condition must be satisfied. Then format is $[0; \delta_G]$. Also PLC is a gateway between laptop and vehicle. Wireless module in figure 3 is standard IEEE 802.11g access point. Included dynamic host control protocol (DHCP) server enables easy connection for any clients.

Fig. 4. Used DC converter housing

### *C. Control Loops*

*1) Speed and Torque:* Six types of speed and torque control loops can be used. Torque on rosette correspond to motor current. Torque loop is one of the converter function and cannot be eliminated. The converter has a four operation modes: IxR compensation, DCtacho/encoder speed control, current control. Gear is percentage of the speed, set from superior laptop. No feedback information is used in the first type, figure 5. It can be used for manual drive without controlling required speed.

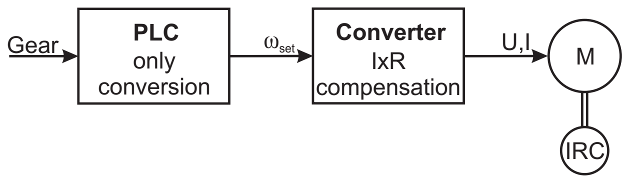



Fig. 5. IxR compensation, speed set without feedback

PID speed controller is included in the second type, figure 6. Direction and value of speed are calculated from impulse signals A and B of IRC. The third type is also scheme from fig. 6. But P,I,D components are adjusting by artificial neural network. The fourth type is same than the second, but PID controller has been changed to fuzzy controller. Derivation $de$ of actual speed excentricity $e$ must be calculated for this case.

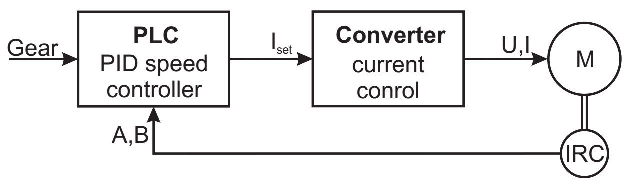



Fig. 6. Current control, PID speed control loop

Converters speed controller is used in the fifth type, in fig. 7. Minimally $10kHz$ IRC signals frequency is required for this loop. Construction of the drive is not with IRC on motor shaft, but with IRC on rosette shaft. This is the unstable point in control loop for this case.

The last type is used in vehicle most frequently. IxR compensation is set in converter. Then speed and torque

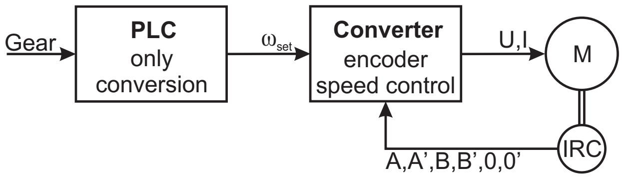



Fig. 7. Speed and current controller in converter

are controlled by converter without feedback. Right vehicle direction is controlled by PLC. Vehicle direction is received from superior laptop. Also steady state of vehicle is controlled by PLC. Vehicle position controller is very suitable for the global navigation.

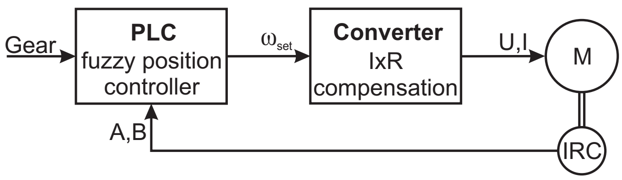



Fig. 8. IxR compensation, fuzzy position controller

*2) Global Motion:* Global motion control is set from laptop, according to figure 8. Zero point is defined after motion control start. Shifts on X and Y global axis are calculated by PLC every $200ms$. Also changes $\delta_G$. Calculation is based on IRC signals A, B. Global parameters are transfered to superior software through wireless connection. Definition for the next motion $[v; r_T]$ is feedback for this parameters. Formated data are sequenced to Front In Front Out (FIFO) buffer. Interaction system for obstacles can consists of laser, ultrasonic, infrared or other systems for scanning space. Additional module is useful for this purpose [5].

### *D. Safety System*

Safety system was designed for the prevention of crashes. Four infra distance sensors located in corner of chassis represent the basic part of the safety system. Reaction distance is $100mm$. System automatically brakes the vehicle in front of the obstacle. Then the vehicle is waiting for new commands with limited motion direction.

## IV. Communication Busses

Synchronous RS485 bus was used for communication between modules in the vehicle. ARION communication protocol was used in this bus. Network consists of five slave modules and one master represented by PLC. HalfDuplex mode was preset for this network. Communication is described in [6]. DB-Net/IP protocol was used for transferring data between PLC and laptop. Data path consists of Ethernet and wireless section. DB-Net/IP protocol has been specially designed for Ethernet automation networks. WPA encryption with Pre-Shared key was chosen for wireless security.

## V. Visualization

Standard MS Windows interface based on desktop application was used for visualization. Application was created in developing software Control WEB 5. DDBNET32 driver was parametrized for the connection Control WEB 5 to DB-Net network. This driver is described in [7]. Visualization allows

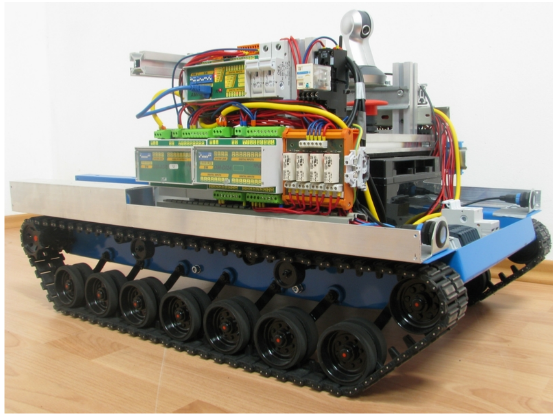


Fig. 9. Constructed model without cover

control of all vehicle functions. Object programing in Control WEB 5 is the best tool for application methods of autonomous vehicle navigation [8], [9].

## VI. Conclusion

Education based on the real model is much more interesting for students than theoretical. Constructed vehicle can be used inside classrooms. The electrical structure enables using of six types of control loops. Ramp obstacle can be used for motion testing. Current and speed characteristics can be compared for each type. These facilities are useful for educational purposes. Also superior laptop structure enables testing of navigation theories in real environments. Additional module can be used for mapping of environment and the other applications for the next research.

## Appendix A

| | | |
|---|---|---|
| Type | MCC 24MP4N | |
| Supply voltage | 24 | $[V]$ |
| Current | 9.6 | $[A]$ |
| Input power | 230 | $[W]$ |
| Output power | 150 | $[W]$ |
| Input rotation speed | 2800 | $[rpm]$ |
| Gear ratio | 7.5 | |
| Output rotation speed | 373 | $[rpm]$ |
| Rated torque | 3.4 | $[Nm]$ |
| Isolation class | F | |
| Covering | IP65 | |
| Weight | 6.26 | $[kg]$ |

## References


[1] F. Ďurovský, "Sylaby z prednášok Elektrovýzbroj vozidiel," 2004, dimenzovanie elektromobilu.
[2] TEA Technik Czech Republic, "Převodové motory," 2007, catalogue.
[3] C. C. O'Gorman, D.Ingersoll, R.G.Jungst, and T.L.Paez, "Artificial Neural Network Simulation of Battery Performance," *Hawaii International Conference on System Science*, vol. 5, pp. 115–121, Jan. 1998.
[4] Maxon motors, "Maxon motors control," 2007, catalogue.
[5] P. Hoppen, T. Knieriemen, and E. von Puttkamer, "Laser-Radar based Mapping and Navigation for an Autonomous Mobile Robot," *IEEE International Conference on Robotics and Automation 1990*, pp. 948 – 953, May 1990.
[6] L. Frýbort and M. Vosáhlo, *ARION Protokol pro ovládání vzdálených V/V*, 1st ed., AMIT spol s.r.o., Chlumova 17, 130 00 Praha 3, Czech Republic, Aug. 2001.
[7] D. Křivský, *Ovládač DDBNET32*, 1st ed., AMIT spol s.r.o., Chlumova 17, 130 00 Praha 3, Czech Republic, Feb. 1998.
[8] A. Vitko and J. Markusek, "Autonomous navigation and control of mobile robot," *Inf. Conf. Mechanical Eng. Bratislava, Slovakia 1999*, pp. 178 – 182, 1999.
[9] S.-Y. Yi and B.-W. Choi, "Autonomous navigation of indoor mobile robots using a global ultrasonic system," *Robotica*, vol. 22, no. 4, pp. 369–374, 2004.

# Fault Tolerance and Reaching Agreement in Faulty Systems

*Gabriel BOCEK*

Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

Gabriel.Bocek@gmail.com

***Abstract*—This paper defines the term fault tolerance and explains the differences between the failures, errors and faults and their occurrence in systems. It presents two traditional fault models, fail-stop and Byzantine, and new fail-stutter fault model. Finally, it describes the Two Army as well as Byzantine Generals problem with the deeper explanation of Lamport's algorithmic solution of Byzantine Generals problem.**

***Keywords*—fault tolerance, Byzantine generals, fail-stop, fail-stutter, fault models, agreement**

## I. Introduction

We are increasingly dependent on services provided by computer systems and our vulnerability to computer failures is growing as a result. Reliability and availability have become more and more important in today's computer dependent world. In many applications where computers are used, outages or malfunction can be expensive or even disastrous. Just imagine the computer systems controlling nuclear equipments or aircrafts. To achieve reliability and availability, we need fault-tolerant computer systems.

## II. Fault Tolerance

*Fault tolerance* is the ability of a system or component to continue normal operation despite the presence of hardware or software faults [6].

System is able to perform its function correctly following the system specification even in the presence of internal faults. The purpose of fault tolerance is to increase the dependability of a system.

A system is k-fault tolerant if it can survive faults in k components and still meets its specification.

## III. Failures, Errors and Faults

A *failure* occurs when an actual running system deviates from this specified behavior. The cause of a failure is called an *error*. It represents an invalid state of system, one that is not allowed by the system specification. The error itself is the result of a defect in the system or fault. *Fault* is the root cause of a failure. It is a malfunction, possibly caused by a design error, manufacturing error, programming error, physical damage, deterioration in the time, unexpected input, operator error and many other causes. That means that an error is merely the symptom of a fault. A fault may not necessarily result in an error, but the same fault may result in multiple errors. Similarly, a single error may lead to multiple failures [2][7].

Lee et al. [5] expanded the definition of failure to meet also the Quality of Service (QoS) criteria in GRID systems as follows:

It is a failure if and only if one of the following two conditions is satisfied:

a. Resources stops due to resource crash
b. Availability of resource does not meet the minimum levels of QoS

They recognize three types of failures related to the previous definition with the reason of failure:

a. Process failure – process stops or starves
b. Processor failure – processor crashes or processor throughput decreases due to burst jobs
c. Network failure - network disconnection and partition or a decrease of network bandwidth due to communication traffic

## IV. Fault Classification

Based on duration, faults can be classified as transient, permanent or intermittent [2][7].

*Transient faults* occur once and then disappear without any apparent intervention. If the operation is repeated, the fault goes away.

*Permanent fault* is one that continues to exist until the faulty component is repaired.

*Intermittent fault* occurs, then vanishes of its own accord, then reappears and so on. It behaves often unpredictably which makes that kind of fault hard to diagnose and handle.

A different way to classify faults is by their underlying *cause* [7].

*Design faults* are the result of design failures. It may appear that in a carefully designed system all such faults should be eliminated through fault prevention, this is usually not realistic in practice. For this reason, many fault-tolerant systems are built with the assumption that design faults are inevitable.

*Operational faults* are that occur during the lifetime of the system and are invariably due to physical causes, such as processor failures or disk crashes.

Finally, the classification of faults can be made on how failed component behaves once its failure occurs. That classification provides following groups of faults [7]:

a. *Crash faults* - the component either completely stops operating or never returns to a valid state

b. *Omission faults* – the component completely fails to perform its service

c. *Timing faults* – the component does not complete its service on time

d. *Byzantine faults* – these are faults of an arbitrary nature

## V. Fault Models

There are countless ways in which computing systems and applications may fail. These failures can be categorized by abstract models that describe how a system will behave in the presence of faults. A fault tolerance technique will assume a certain model of failure when making claims about the types of faults it can handle. We present the two most common failure models here, as well as a relatively new model which is a more accurate, but more complex representation of how real systems work.

The *fail-stop fault* model allows any node to fail at any time, but when the failure occurs, it stops producing output and interacting with the rest of the system. Furthermore, all other nodes automatically know that the node has failed. This fault model represents common modes of failure such as a system hang or crash. It is also called *fail-silent* model [2][4].

The *Byzantine fault* model represents the most adversarial model of failure. This fault model allows failed nodes to continue interacting with the rest of the system. Behavior can be arbitrary and inconsistent, and failed nodes are allowed to collude in order to devise more malicious output. Correctly operating nodes cannot automatically detect that any nodes have failed, nor they know which nodes in particular have failed if the existence of a failure is known. This model can represent random system failures as well as malicious attacks by a hacker. Dealing with the Byzantine faults is much more difficult than dealing with fail-stop one [2][4].

The Byzantine fault model is extremely broad, perhaps unrealistically so, and is therefore extremely difficult to analyze or tolerate. The fail-stop model is commonly used when presenting fault-tolerance techniques, but it is often criticized as overly simplistic due to its failure to represent many types of real-world failures. The *fail-stutter fault* model is an attempt to provide a model between these two extremes. The fail-stutter model is an extension of the fail-stop model. It attempts to maintain the tractability of that model while expanding the set of real-world faults that it includes. The fail-stutter model includes all provisions of the fail-stop model, but it also allows for *performance faults*. A performance fault is an event in which a component provides unexpectedly low performance and doesn't meet the performance specification, but it continues to function correctly with regard to its output. This extension allows the model to describe failures in systems much more realistic such as poor latency performance of a network switch when suddenly hit with a very high traffic load [3].

## VI. Agreement in Faulty Systems

There is a need in many distributed systems to have processes agree on something. Examples are electing a coordinator, deciding whether to commit a transaction or not, dividing up tasks among workers, synchronization and many others. When the communication and processors work correctly, reaching such agreement is often straightforward. The problems arise when they are not perfect and some faults appear in system. The general goal of distributed agreement algorithms is that all properly working processors reach consensus on some issue and do that within finite number of steps.

Let us have a case of perfect processors but communication lines are not reliable and can lose messages. There is a famous problem known as the *Two Army* also called *Two Generals* or the *Coordinated Attack* problem. It illustrates the difficulty of getting even two perfect processors to reach agreement.

The red army is encamped in a valley. Two blue armies are encamped on the surrounding hills overlooking the valley. Two blue armies will be victorious just in case they will coordinate their attacks on the red army. The goal of the blue armies is to reach the agreement about the attacking. The catch is that they can only communicate using an unreliable channel – sending the messenger who can be caught by the enemy army.

Now assume that the commander of blue army 1, General Alexander, sends a message to the commander of blue army 2, General Bonaparte, with following contents: "Let's attack at dawn tomorrow". The messenger goes through and Bonaparte sends him back with a note saying: "Perfect idea. See you at dawn tomorrow." The messenger arrives back to his base safely. Later Alexander realizes that Bonaparte does not know if a messenger got back safely and not knowing this, may not dare to attack. Therefore Alexander sends the messenger again to go and tell Bonaparte that his message arrived and the battle is set. Once again the messenger gets through and delivers the acknowledgement, now Bonaparte worries whether Alexander knows that the acknowledgement was received. He assumes that if Bonaparte thinks that the messenger was captured, he will not be sure about the plans and may not risk the attack so he sends the messenger back again. Even the messenger makes it though every time it is easy to show that Alexander and Bonaparte will never reach agreement, no matter how many acknowledgements will be sent. It is clear that the sender of last message does not know if the last message arrived. Even with non-faulty processors (generals), agreement between two processes is not possible in a system with unreliable communication [2].

## VII. BYZANTINE GENERALS PROBLEM

Now let us assume that the communication is reliable but the processors may occur like faulty. The classical problem here also occurs in a military setting and is called the Byzantine generals problem which was expressed by Lamport et. al [1].

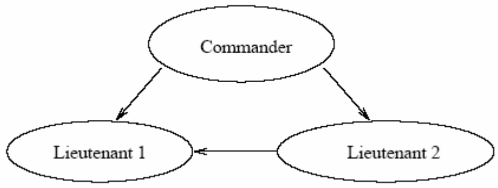



Figure 1: Byzantine Generals model

The army consists of several divisions. Each of them is commanded by its own general. The divisions are camped outside the enemy city and the generals can communicate each other by sending a messenger. The generals must decide on the course of action. This must be a collective decision based on all available facts. However in some cases there may be a traitor. Traitor generals try to prevent loyal general from reaching agreement. They do this by sending inconsistent messages regarding the order to other generals.

The generals must have an algorithm to guarantee that

a. All loyal generals decide upon the same plan of action.
b. A small number of traitors cannot cause the loyal generals to adopt a bad plan.

Condition a. means that the loyal generals will all obey what the algorithm say, but the traitors may do anything they wish. The algorithm must guarantee condition a. regardless of what the traitors do.

Condition b. is hard to formalize. Instead, we consider how the generals reach a decision. Let $v_i$ be the information communicates by the i-th general. Each general knows how to combine the values $v_1$ .. ,$v_n$ into a single plan of action, where n is the number of generals. Condition a. can be reached by having all generals use the same way how to combine the information. Let the possible values for $v_i$ be attack or retreat. Then $v_i$ can be General i's opinion which opinion is the best. The final decision can be made on the majority vote among them. To satisfy the condition a., the following must be true:

1. Every loyal general must obtain the same information $v_1$.., $v_n$
2. If the i-th general is loyal, then the value that he sends must be used by every loyal general as the value of $v_i$

A commanding general must send and order to his n-1 lieutenant generals such that

IC1. All loyal lieutenants obey the same order.

IC2. If the commanding general is loyal, then every loyal lieutenant obeys the order he sends.

Conditions IC1 and IC2 are called *interactive consistency* conditions.

It has been shown that for a solution to the Byzantine Generals problem using oral messages to identify m traitors, there must be at least 3m+1 generals [1]. The definition of an oral message is:

1. Every message that is sent is delivered correctly.
2. The receiver of a message knows who sent it.
3. The absence of a message is detectable.

Assumption 1 and 2 prevent traitors from interfering with the communication between two other generals. This is prevented by the fact that assumption 1 does not allow for messages sent to be interfered with. With assumption 2 we can assert that communications can not be inserted to confuse the network with spurious information. Assumption 3 foils a traitor who tried to prevent a decision by simply not sending messages. The following algorithms require for each general to be able to send messages directly to every other general.

A traitorous commander may decide not to send any order. Each lieutenant is required to presume that an order is to be communicated across the network, so a default order must be specified and it's RETREAT.

Lamport [1] defines the Oral message algorithm OM($m$) for all positive integers m, such that each commander sends an order to $n-1$ lieutenants. OM($m$) solves the problem for $3m+1$ or more generals in the presence of at most m traitors.

Lamport's algorithm presumes a function of majority with the property that if a majority of values $v_i$ equals $v$, then *majority* $(v_1, \ldots, v_{n-1})$ equals $v$. There are 2 natural choices for the value of the sequence $(v_1, \ldots, v_{n-1})$:

1. The majority value among the $v_i$ if it exists, otherwise the value RETREAT
2. The median of the $v_i$ assuming that they come from an ordered set.

The following algorithm presented by Lamport [1] requires only a majority.

Algorithm OM(0):

1. The commander sends his value to every lieutenant.
2. Each lieutenant uses the value he receives from the commander or uses the value RETREAT if he receives no value.

Algorithm OM($m$), $m > 0$

1. The commander sends his value to every lieutenant
2. For each $i$, let $v_i$ be the value Lieutenant $i$ receives from the commander or else be RETREAT if he receives no value. Lieutenant $i$ acts as the commander in Algorithm OM($m-1$) to sent the value $v_i$ to each of the $n-2$ other lieutenants.
3. For each $i$, and each $i \neq j$, let $v_i$ be the value lieutenant $i$ received from lieutenant $j$ in step (2) (using Algorithm OM($m-1$)) or else RETREAT if he received no such value. Lieutenant $i$ uses the value of the majority ($v_i, \ldots, v_{n-1}$).

We will now work through an example, so as to understand how the algorithm works. Let us consider the case with one traitor m = 1, n = 4. As we can see from Figure 2 the messages received by Lieutenant 2 when the General sends the value v and Lieutenant 3 is the traitor. To summarize OM(1):

1. General sends v to all Lieutenants
2. Lieutenant 1 sends v to Lieutenant 2 using OM(0)
3. Lieutenant 3 sends x to Lieutenant 2
4. Lieutenant 2 has $v_1 = v_2 = v$ and $v_3 = x$
5. Value is v = majority(v, v, x)

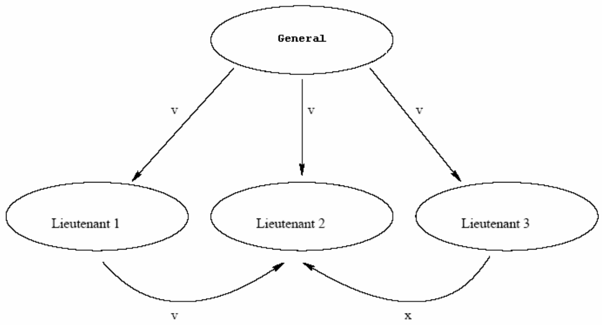



Figure 2: Lieutenant 3 is a traitor

## VIII. Conclusion

This paper is focused on the problems related to the fault tolerance. At the start, the term fault tolerance is defined as well as the distinctions between the failure, fault and error. Furthermore, the classification of failures from several points of view is presented. After that, the paper continue describing two standard fault models (fail-stop and Byzantines) as well as one recently introduced fail-stutter fault model. The second part attends to reaching agreement in faulty systems with deeper explanation of two army problem and Byzantine Generals problems and its Lamport's algorithmic solution.

## References

[1] L. Lamport, R. Shostak, M. Pease, "The Byzantine Generals Problem" , ACM Transactions on Programming Languages and Systems, Vol.4, No.3, July 1982, pp 382 - 401

[2] A.S. Tanenbaum, M. van Steen, "Distributed Systems: Principles and Paradigms", Prentice Hall, 2002, pp. 212 - 223

[3] R.H. Arpaci-Dusseau, A.C.Arpaci-Dusseau, "Fail-Stutter Fault Tolerance", In The 8th Workshop on Hot Topics in Operating Systems 2001

[4] M. Treaster, "A Survey of Fault-Tolerance and Fault-Recovery Techniques in Parallel Systems", ArXiv Computer Science e-prints 2005

[5] H.M. Lee, K.S. Chung, S.H. Chin, D.W. Lee, S.Y.Jung, J.Ch. Yu, "Grid Fault Tolerance Service for Quality of Service", The 3rd IEEE/ACM International Symposium on Cluster Computing and the GRID, 12-15 May 2003

[6] Institute of Electrical and Electronics Engineers, "IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries". New York, NY: 1990.

[7] B. Selic, "Fault Tolerance Techniques for Distributed Systems", IBM, 27th July 2004

# Steganography in Colour Images

[1]Gabriel BUGÁR

Supervisor: Prof. Ing. Dušan LEVICKÝ, CSc.

[1]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic

[1] gabriel.bugar@tuke.sk, dusan.levicky@tuke.sk

***Abstract*—Steganography is the art and science of writing hidden messages in such a way that no one apart from the intended recipient knows the existence of the message. Recently, many digital techniques have been developed that perform steganography on electronic media, most notably sound and image files. This paper presents the basic idea of steganography in colour images by using spread spectrum technique. The discrete cosines transformation and its major properties are briefly described too. Finally, some experimental results verify the effectiveness of designed algorithm.**

***Keywords*—Steganography, DCT, Spread Spectrum Systems, Colour spaces.**

## I. Introduction

Criminals, spies, and terrorists are always looking for better ways to communicate with one another without detection by police and intelligence officers. These officials fear their foes might turn to steganography, a technique that hides a small message within the code representing an image or other standard computer file. While classical cryptography is about concealing the content of messages, steganography is about concealing their existence. Classical steganography concerns itself with ways of embedding a secret message which might be a copyright mark, or a covert communication, or a serial number in a cover message such as an image, video film, an audio recording, or computer code. The embedding is typically parametrised by a key; without knowledge of this key or a related one it is difficult for a third party to detect or remove the embedded material. Once the cover object has material embedded in it, it is called a stego object. Thus, for example, we might embed a mark in a covertext giving a stegotext; or embed a text in a cover image giving a stego-image; and so on. This terminology was agreed at the First International Workshop on Information Hiding [1][2].

## II. The Discrete Cosine Transform (DCT)

The Discrete Cosine Transformation (DCT) helps separate the image into parts or spectral sub-bands of differing importance with respect to the image's visual quality. The DCT is similar to the discrete Fourier transform. It transforms a signal or image from the spatial domain to the frequency domain.

The general equation for a 1D ($N$ data items) DCT is defined by the following equation:

$$F(u)=\alpha(u)\sum_{x=0}^{N-1} f(x).\cos\left[\frac{\pi(2x+1)u}{2N}\right], \tag{1}$$

for $u$=0,1,2,…,N-1. Similarly, the inverse transformation is defined as:

$$f(x)=\sum_{x=0}^{N-1}\alpha(u)F(u).\cos\left[\frac{\pi(2x+1)u}{2N}\right], \tag{2}$$

for $x$=1,2,…,N-1. In the both equation (1) and (2) $\alpha(u)$ is defined as:

$$\alpha(u)=\begin{cases}\sqrt{\dfrac{1}{N}} & for \quad u=0\\[2mm] \sqrt{\dfrac{2}{N}} & for \quad u\neq 0\end{cases} \tag{3}$$

It is clear from (1) that for $u$=0, $F(u=0)=\sqrt{\frac{1}{N}}\sum_{x=0}^{N-1} f(x)$. Thus, the first transform coefficient is the average value of the sample sequence. This value is referred to as the *DC coefficient*. All other transform coefficients are called the *AC coefficients*.

In this paper all works has been doing with 2-D discrete cosine transformation. The 2-D DCT is a direct extension of the 1-D case and is given by the following way:

$$F(u,v)=\alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y) \cos\left[\frac{\pi(2x+1)u}{2N}\right]\cos\left[\frac{\pi(2y+1)v}{2N}\right], \tag{4}$$

for $u$,$v$=0,1,2,…,N-1 and $\alpha(u)$ and $\alpha(v)$ are defined in (3). The inverse transform is defined as:

$$f(x,y)=\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\alpha(u)\alpha(v)F(u,v) \cos\left[\frac{\pi(2x+1)u}{2N}\right]\cos\left[\frac{\pi(2y+1)v}{2N}\right], \tag{5}$$

for $x,y$=0,1,2,…,N-1. The 2-D basis functions can be generated by multiplying the horizontally oriented 1-D basis functions with vertically oriented set of the same functions. The basis functions for $N$ = 8 are shown in Fig. 1.

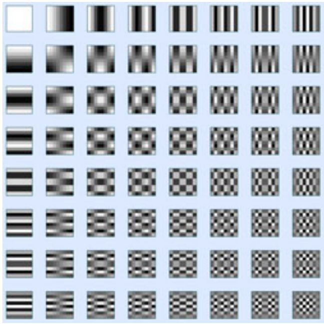


Fig. 1. The basic functions for N=8

Again, it can be noted that the basis functions exhibit a progressive increase in frequency both in the vertical and horizontal direction. The top left basis function of results from multiplication of the DC component in Fig. 1. with its transpose. Hence, this function assumes a constant value and is referred to as the DC coefficient [3].

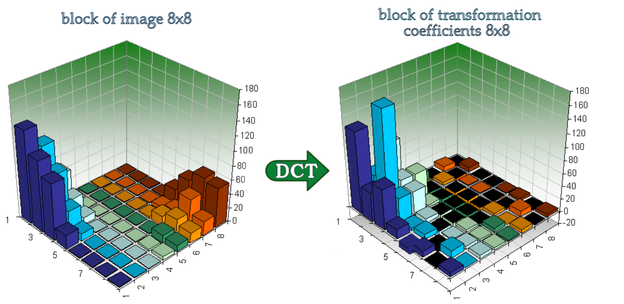



Fig. 2. Transformation from the spatial domain to the frequency domain

## III. The Spread Spectrum Technique

Spread spectrum uses wide band, noise-like signals. Because Spread Spectrum signals are noise-like, they are hard to detect. Spread Spectrum signals are also hard to intercept or demodulate.

Spread spectrum techniques are methods by which energy generated in a particular bandwidth is deliberately spread in the frequency domain, resulting in a signal with a wider bandwidth. These techniques are used for a variety of reasons, including the establishment of secure communications, increasing resistance to natural interference and jamming, and to prevent detection.

In this paper the new algorithm uses the Direct Sequence Spread Spectrum (DSSS) is presented. Direct sequence is the best known Spread Spectrum Technique. The data signal is multiplied by a Pseudo Random Noise Code (PN code). A PN code is a sequence of chips valued -1 and 1 (polar) or 0 and 1 (non-polar) and has noise-like properties. This results in low cross-correlation values among the codes and the difficulty to jam or detect a data message. Several families of binary PN codes exist; a usual way to create a PN code is by means of at least one shift-register. This work uses a Hadamard PN codes. In direct sequence systems the length of the code is the same as the spreading-factor. This can also be seen from Fig. 3; where we show how the PN code is combined with the data [4].

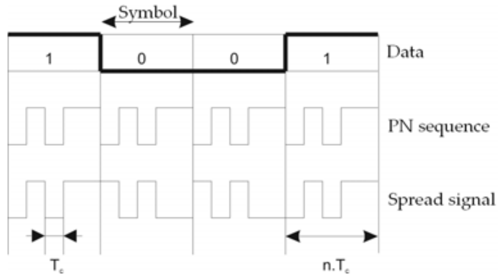



Fig. 3. Direct Sequence spreading

## IV. System Design

The basic communication techniques with the spread spectrum systems and the processing technique of the static colored images are combined with aim to achieve the Spread Spectrum Image Steganography (SSIS). The basic idea of SSIS includes the embedding a secrete message to a noise, that is subsequently embedded to a digital cover image. As far as the power of the noise is constrained at a low level, the embedded informations are not visible either to a human person or even to computational resources that have no prior knowledge of the original cover object.

### A. *Basic activity of the designed method*

The designed method is based on embedding a secrete message $m$ to the transformation coefficients 2D DCT of a cover image $c$. It is necessary to know the original image at the receiving side to achieve successful extraction.

### B. *Embedding a secrete message to a cover image*

The first step of this procedure is based on the choosing of the particular cover image and secret message. Afterwards, cover image is decomposed into single colored component that belongs to the colored model that is distributed to 8x8 blocs. These single blocs are transformed into the transformation coefficients with using the DCT transformation. Each single matrix 8x8 of transformation coefficients are rearranged to the vectors with the "cik-cak" method in such a way that always the first coefficient is the DC element and the other coefficients consist of AC coefficients.

In this paper the variable offset defines a number of coefficients in a single block that are not modified. In this block, if the offset equals e.g. 8 then the first 8 coefficients are not modified. This protective procedure causes the fact, that the coefficients with the greatest energy are not changed.

The secrete message is decomposed at first to a single colored component in correspondent colored model. It is necessary to change all values of secret message to binary values (1,0) and due to increase the error robustness they are modified to values (-1,1). Subsequently, the secrete message is spread by Pseudo Random Noise Code (PN code). Each information bit of secret message is spread by the PN code. The computer simulation of this algorithm uses Hadamard PN codes with period 8 bits for spreading the information signal. The following spread data are multiplied by the scale factor α and then it is added to transformation coefficients. The next step is the inverse arrangement of coefficients and application of the inverse cosine transformation (IDCT). The block diagram of the designed method of embedding the secrete message to the cover image is shown on Fig. 4.

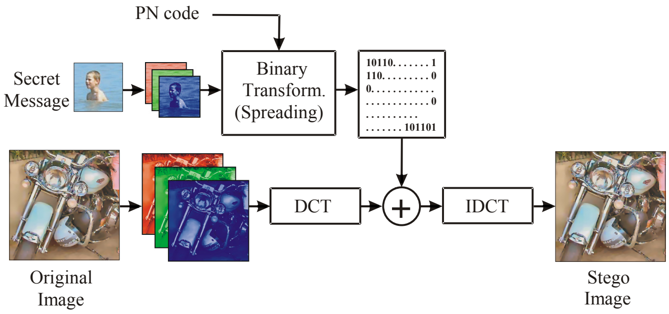



Fig. 4. Embedding of images to messages

### C. Extraction

For the successful extraction of a secrete message it is needed to know the original image, stego image and the kind of PN code on the other side. The procedure of the extraction is analogical to embedding. It starts with the DCT transformation of original and stego image. After the DCT transformation of the original and stego image are the consequent transformation coefficients subtracted from each other. It makes the spread data. The secrete message is obtained if the spread data are multiplied with the correct PN code. The process of extraction is shown on the fig. 5.

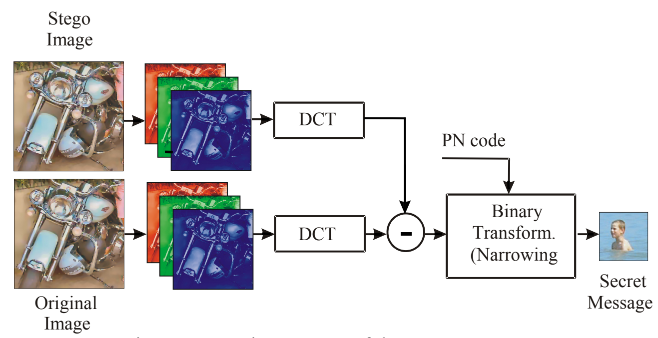



Fig. 5. Extraction process of the secrete message

## V. Experimental Result

In experiments two static colour cover images *Motorka246.bmp* and *Lietadlo256.bmp* with spatial resolution 256x256 pixels and bit depth 24 bit/pixel were tested in my experiments. These two images have different dynamic details and different colour contrast. Two static colour images with different spatial resolution and colour contrast were tested as secret messages. Three colour models, the RGB, $YC_BC_R$ and HSV were employed in experiments.

The important properties in steganography systems are the undetectability, the secrete message imperceptibility and the capacity, as well as the quantum of the embedded secrete message. All of these properties are tested in experiments and some of the results are written in Tables I, II and III. The general indicator of undetectability is the value of PSNR (Peak Signal to Noise Ratio). The most highly achieved values of PSNR are highlighted in these tables.

TABLE I
Demonstration Of The Highest Value Of PSNR For RGB

| **RGB** scale factor **α** | 0,35 | 0,40 | 0,45 | 0,50 | 0,55 |
|---|---|---|---|---|---|
| **PSNR**$_{RGB}$ [dB] | 49,143 | 48,431 | 47,759 | 47,084 | 46,454 |
| **Extraction** [%] | 99,979 | 99,988 | 99,998 | 100 | 100 |

TABLE II
Demonstration Of The Highest Value Of PSNR For $YC_BC_R$

| $YC_BC_R$ scale factor **α** | 0,35 | 0,40 | 0,45 | 0,50 | 0,55 |
|---|---|---|---|---|---|
| **PSNR**$_{YCbCr}$ [dB] | 48,971 | 47,245 | 46,879 | 45,917 | 45,032 |
| **Extraction** [%] | 99,966 | 99,972 | 99,986 | 99,992 | 100 |

TABLE III
Demonstration Of The Highest Value Of PSNR For HSV

| **HSV** scale factor **α** | 0,35 | 0,40 | 0,45 | 0,50 | 0,55 |
|---|---|---|---|---|---|
| **PSNR**$_{HSV}$ [dB] | 48,984 | 48,239 | 47,554 | 46,881 | 46,250 |
| **Extraction** [%] | 99,979 | 99,994 | 99,997 | 100 | 100 |

We can see that the most highly achieved values of PSNR presented for the $YC_BC_R$ colour model as well; however it is impossible to apply them for steganography using the designed method on the ground of visible degradation of the stego image, i.e. sensibility of embedding the secret message into the cover image. Therefore, we can only use RGB and HSV colour models. If we compare the values for these two colour models in tables I and II, then we can state that the steganography using the designed method in the RGB colour model is more efficient as in the HSV colour model. Regarding the capacity of embedding data, it is the same for both colour models, because the method of embedding and extraction of the secret message into or from the cover image is identical.

## VI. Conclusion

In this paper, there has been presented some possibilities of steganography exploitation by systems of spread spectrum in colour images with different colour models. The main idea is the embedding the spreaded secret message by PN code to 2D DCT coefficients of the cover image. By increasing the volume of secret data, the number of modified transformation coefficients is increasing as well, and so do the degradation of the stego images. Therefore, it is necessary to choose a volume of secret data which would not threaten the safety of the communication in the chosen cover image.

## Acknowledgment

The work presented in this paper was supported by Ministry of Education of Slovak republic VEGA Grant No. 1/4054/07.

## References

[1] R. J. Anderson, F. A. P. Petitcolas, “On limits of steganography,” *IEEE Journal of Selected Areas in Communications*, 16(4):474-481, May 1998.

[2] S. Katzenbeisser, F. A. P. Petitcolas, *Information Hiding Techniques for Steganography and Digital Watermarking*. Artech House, Boston, 2000.

[3] Syed Ali Khayam, *The Discrete Cosine Transform (DCT):Theory and Application*. Michigan State University, March 2003. http:\\www.egr.msu.edu/waves/people/Ali_files/DCT_TR802.pdf

[4] E. Strom, T. Ottosson, A. Svensson, *An Introduction to Spread Spectrum Systems*. Chalmers University of Technology, Goteborg, Sweden, 2002.

# Effects of Spreading Codes and Convolution Coding on the Performance of MC-CDMA System with Nonlinear Model of HPA

*Juraj GAZDA, Peter DROTÁR*

Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

peto.drotar@tuke.sk, juraj.gazda@tuke.sk

***Abstract*—In this paper, the performance analysis of uplink perfectly synchronized multi-carrier code division multiple access (MC-CDMA) transmission system model for different spreading sequences (Walsh code, Gold code, orthogonal Gold code, Golay code and Zadoff-chu code) and different types of receivers (minimum mean-square error receiver (MMSE-MUD), bank of matched filters (BMF) and microstatistic multi-user receiver (MSF-MUD)) at the Saleh model of high power amplifier (HPA) is made. The results of our analyses based on computer simulations will show very clearly, that the application of MSF-MUD in combination with Golay codes can provide significant better results than the other tested spreading codes and receivers. Taking advantage of convolutional coding even highlights the performance improvement provided by MSF-MUD. Besides, a failure of Walsh code at the Saleh model of HPA will be presented. This effect will be explained by using constellation diagram symbol migration depending on input back-off of HPA.**

***Keywords*—MC-CDMA, HPA, Saleh model, multi-user detection, convolution coding**

## I. Introduction

MC-CDMA is a modulation technique that exploits the advantages of spread spectrum and orthogonal frequency division multiplexing (OFDM). One of the major disadvantages of multi-carrier systems based on OFDM is the high sensitivity to nonlinear amplification, which requires large back-off in the transmitter amplifier and, as a consequence, inefficient use of HPA. On the other hand, using low back-offs leads to signal distortion and, as a result, increased performance degradation.

Several techniques can be found in the literature to reduce the sensitivity of MC-CDMA systems to nonlinear amplification. Most common solutions at the transmitter side include predistortion and $PAPR$ reduction [1]. Receiver side strategies usually combine iterative decoding [2] and multi-user detection so both HPA nonlinearity compensation and multiple access interference (MAI) are taken into account [1], [3], [4].

In this paper, the performance analysis of uplink MC-CDMA transmission system for the different spreading codes, different input back-off (IBO) parameters, Saleh model of HPA and different receiver types (bank of matched filters (BMF), MMSE-MUD, MSF-MUD) will be done. Moreover, for the enhanced performance of MC-CDMA system, convolution coding is applied in presented MC-CDMA system model.

The results of our analyses obtained by computer simulations will show, that the best performance properties of MC-CDMA transmission system can be obtained by Golay codes application in conjuction with MSF-MUD and convolutional coding at Saleh model of HPA. Besides, a failure of Walsh codes at the Saleh model of HPA will be presented. This Walsh code behaviour will be explained by using constellation diagram symbol migration depending on input back-off of HPA.

## II. Nonlinear effects in MC-CDMA

The basic scenario of our study of nonlinear effects in MC-CDMA systems is represented by uplink MC-CDMA transmission system performing through AWGN transmission channel, at $K$ active users and under condition of perfect synchronization.

A block diagram of the base-band model of MC-CDMA transmitter is given in the Fig. 1. It can be seen from this figure, that the information bits to be transmitted by a particular user are firstly feed into the block of convolutional coder, followed by QAM base-band modulator. Then they are spreaded by using a specific spreading code $c_m$. In the case of MC-CDMA, as the spreading codes Walsh codes, Gold codes, orthogonal Gold codes, polyphase Zadoff-Chu codes and complementary Golay codes are used mostly [5]. The spreaded symbols are modulated by multi-carrier modulation implemented by IFFT operation. The IFFT block outputs after parallel-to-serial conversion represents the input signal of a HPA. In the transmitter model according to the Fig. 1, the traditional block of cyclic prefix insertion is not included because of AWGN channel assumption.

The basic structure of receivers considered in this paper is sketched in the Fig. 2. It can be seen from this figure, that the receiver consists of serial-to-parallel converter (S/P), blocks of FFT, BMF, block of linear or non-linear transformation (labelled as T), QAM demapper block, followed by block of convolutional decoder, which is represented by Viterbi algorithm. A deeper description of MC-CDMA transmission systems can be found e.g. in [6] or [7].

In order to extend the BMF into multi-user receiver, the T-transformation block is included in the receiver structure. In this paper, the linear MMSE-MUD [7] as well as nonlinear MSF-MUD for MC-CDMA [4] are considered. MSF-MUD is based on complex-valued multi-channel microstatistic filter aplication (C-M-CMF). C-M-CMF is minimum mean-square error estimator based on the estimation of desired signals by using a linear combination of vector element obtained by treshold decomposition of the filter input signals. The detailed description of MMSE-MUD and MFS-MUD structure, design and performance is beyond of this paper and can be found e.g. in [4], [7].

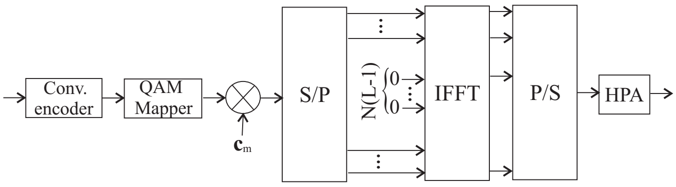



Fig. 1. MC-CDMA transmitter

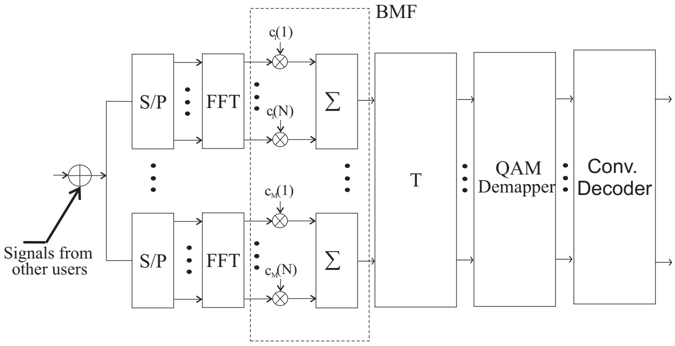



Fig. 2. MC-CDMA receiver

It is well-known, that the output of multi-carrier modulator based on IFFT operation are characterized by the wide dynamic range or so-called large fluctuation of signal envelope [1], [7]. Measure of this fluctuation is usually expressed by $PAPR$. The $PAPR$ of signal $s(t)$ is defined as follows:

$$PAPR = \frac{max|s(t)|^2}{E[|s(t)|^2]} \tag{1}$$

where $E[\ .]$ is the expectation operator. It can be shown (e.g. [5]), that in the case of MC-CDMA systems, $PAPR$ parameter strongly depends on the applied spreading code.

Because of the large envelope fluctuations of the input signal of HPA, the real HPAs have to be modeled as nonlinear amplifiers. The most widely used models of HPAs applied for this purpose are Saleh and Rapp model and models known as soft- and hard-limit [1]. Because of the limited range of this paper, we will focus on Saleh model. The Saleh model is the most complicated model from above introduced list of nonlinear HPA model. The Saleh model is defined by so-called amplitude-to-amplitude (AM/AM) and amplitude-to-phase (AM/PM) characteristics given by

$$AM/AM : G_{u_x} = \frac{\kappa_G u_x}{1+\chi_G u_x^2} \tag{2}$$

$$AM/PM : \Phi_{u_x} = \frac{\kappa_\Phi u_x^2}{1+\chi_\Phi u_x^2} \tag{3}$$

The HPA operation in the region of its nonlinear characteristic causes a nonlinear distortion of transmitted signal, that subsequently results in higher $BER$ and out-of-band energy radiation. The nonlinear HPA operating point is defined by $IBO$ which corresponds to the ratio between the saturated and average input power[4]:

$$IBO_{dB} = 10\log_{10}(\frac{P_{max}}{P_x}) \tag{4}$$

The measure of nonlinear effects due to nonlinear HPA could be decreased by the selection of relatively high value of $IBO$ parameter. However, this approach will result in inefficient HPA performance. Another solutions intent on decreasing of nonlinear HPA influence are representing by a number of methods implemented in transmitter or/and receiver side of MC-CDMA transmission system [3]. It has been proposed in [4] that MC-CDMA performance improvement under the nonlinear HPA application could be reached also by the application of nonlinear receivers.

It follows from the above outlined analyses that the MC-CDMA performance expressed by $BER$ depends strongly on spreading codes, HPA model and the applied receiver. The spreading codes selection depends on the $PAPR$ parameter (the lower $PAPR$ the better performance is expected) as well as on the level of MAI due to cross-correlation properties of the particular spreading codes. It will be shown in the next, that in the case of the spreading code selection, the HPA model including $IBO$ parameter should be taken into account. Because of non-linear distortion due to HPA and MAI inherency, it is expected that the application of non-linear and multi-user receiver will overcome linear and single user receivers.

## III. MC-CDMA PERFORMANCE ANALYSIS

MC-CDMA performance analysis presented in this chapter is based on computer simulations. The basic scenario of our simulation is represented by the uplink MC-CDMA transmission system performing through AWGN transmission channel, at 16-QAM based-band modulation, for $K$ active users ($K = 8$), under condition of perfect synchronization of transmitter and receiver, at the presence of Saleh model of HPA.

As the spreading codes, Walsh codes, Gold codes and orthogonal Gold codes with period of 32 chips as well as complementary Golay codes and polyphase Zadoff-chu codes with period of 31 chips have been applied. In our simulations, the total number of sub-carriers has been set to $N = 128$. In order to avoid aliasing and the out-of-band radiation into the data bearing tones, the oversampling rate of 4 has been applied [1]. Then, $N_u = 32$ (Walsh codes, Gold codes and orthogonal Gold) and $N_u = 31$ (complementary Golay codes and Zadoff-chu codes) carriers per transmitted block have been used for the spreaded 16-QAM symbol transmission. Non-recursive, non-systematic convolution encoder with rate $R = 1/3$ is applied at the trasmitter side. For convolution decoding, Viterbi algorithm is employed in presented MC-CDMA system model. For the precise specification of the Saleh model of HPA, the parameters $\kappa_G = 2$, $\chi_G = \chi_\Phi = 1$ and $\kappa_\Phi = \pi/3$ have been chosen.

It follows from M-ary QAM theory that symbols generated by M-ary QAM modulator are located on finite number of circles of signal constellation. For example in the case of 16-QAM, there are 3 cycles where the particular symbols are located. The symbols located on the same circle have the same magnitude. Therefore if we assume nonlinear HPA modeled by Saleh model at single user scenario under noiseless condition all M-ary QAM symbols located on the same circle will be mapped by nonlinear HPA following (2) from the original circle into another circle called image circle. Because the nonlinear HPA will cause also phase distortion, the particular symbols at the receiver side will be rotated along the image circle in an angle given by (3). It follows from (2), that the image circle radius (i.e. the magnitudes of the symbols located on the image circles) depends on the particular symbols and $IBO$ parameter expressing of HPA operating point.

As we declared above, the aim of simulations is to analyze MC-CDMA performance for the different spreading codes, $IBO$ parameter values, receiver types (BMF, MMSE-MUD and MSF-MUD) at Saleh model of HPA. In the Fig. 3 and Fig. 4, the original signal constellations and symbol constellations at the output of BMF for $E_b/N_0 = 40dB$ are given. In these figures, small empty circles represent original signal constellation of the 16-QAM symbols and other small filling blue circles represent signal constellation of the 16-QAM symbols at the BMF output for $IBO = 0dB$. The black solid curves illustrate the symbol constellation migration due to changing $IBO$ from $0\ dB\ to\ 20\ dB$. The direction of the symbol migration is outlined by the arrow. It will be seen from these figures, that if $IBO = 20dB$ then the original signal constellation and constellation at the BMF output is the same. It simply means, that if $IBO = 20dB$ then HPA operates in linear region of its characteristics.

If Walsh codes are applied as the spreading codes and Saleh model of HPA is used, very interesting phenomenon can be identified in the symbol constellations given in the Fig. 3. As it is clearly seen from this figure, all symbols of symbol constellation of the signal at the BMF output (red asterisks) are located on the same image circle for $IBO = 5dB$. It means, that the trajectories of all migrated symbols for increasing $IBO$ intersect the same red-dash circle for $IBO = 5dB$. This special phenomenon can be explained as the result of combination of relatively high fluctuation of the envelope of signal at the HPA input for Walsh codes (i.e. relatively high $PAPR$) and AM/AM characteristic shape of the Saleh model of HPA. This critical $IBO$ values are varied in dependance on Walsh sequences that are considered for transmission. Following [7] this feature is critical for all conventional receivers and as we can see in the Fig. 5 and Fig. 6, even for MSF-MUD. This phenomenon results in significant high $BER$ that is not acceptable in conventional communication systems.

It can be seen from the Fig. 4, that the Golay codes do not generate an effect similar to that of Walsh codes. Other codes (Gold, ortogonal Gold, Zadoff-chu) generate similar figures than Golay codes do and therefore corresponding constellation diagrams are not depicted here. Fig. 4 indicates that the application of Golay codes, Gold codes, ortogonal Gold codes and Zadoff-chu codes should provide better performance of MC-CDMA system than Walsh code using. This conclusion is confirmed by the simulation results given in the Fig. 5 and Fig. 6.

In the Fig. 5, $BER\ vs.\ IBO$ for $E_b/N_0 = 17dB$ is illustrated. It can be seen from this figure, that the worst performance of MC-CDMA is provided by the Walsh code application what is due to the effect outlined in the Fig. 3. On the other hand, the best performance can be provided by the application of Golay codes in combination with MSF-MUD. This performance can be explained by Golay code $PAPR$ and the MSF-MUD ability to decrease nonlinear distortion influence of HPA. It follows from this figure, that there is the interval values of $IBO$ for which the application of MSF-MUD can provide meaningful improvement of the MC-CDMA performance in comparison with that of BMF and MMSE-MUD application.

The performance of the MC-CDMA expressed by $BER\ vs.\ E_b/N_0$ at $IBO = 3\ dB$ is illustrated by Fig. 6. The simulation results confirms the previous conclusions i.e. the best performance can be obtained by using Golay codes in combination with MSF-MUD.

Fig. 7 and Fig. 8 show the impact of convolution coding on the performance of MC-CDMA system. These figures expose the fact, that by applying convolution coding, MC-CDMA system becomes more robust against HPA influence as well as against AWGN. These results confirms the error-correction coding theory and therefore, the combination of relatively low $IBO$, Golay codes, MSF-MUD and convolutional coding provides perfect performance of MC-CDMA system in selectected scenario depicted in Fig. 7 and Fig. 8. Note that higher computing complexity of convolutional encoder and decoder has to be taken into account.

## IV. Conclusions

In this paper, we have investigated performance for MC-CDMA for the different spreading codes, $IBO$ parameter values and receiver types at the Saleh model. It has been found that the application of Walsh codes provides the worst performance of MC-CDMA what is due to Walsh code $PAPR$ and HPA model according to Saleh. Therefore, the simulation results indicate that the Walsh codes are unsuitable for the scenario where HPA is described by the Saleh model. On the other hand, the application of the Golay codes as the spreading codes in combination with MSF-MUD and convolution coding can overcome clearly the other investigated spreading codes and receivers. These properties of MSF-MUD are reached at the cost of its the higher complexity [4].

## V. Acknowlegments

This work has been funded by VEGA 1/4088/07 ”Rekonfigurovateľné platformy pre širokopásmové bezdrôtové telekomunikačné siete” and COST 297: “High Altitude Platforms for Communications and other Services”.

## References


[1] M. Deumal, “On OFDM-based Multicarrier Communication Systems with Low Sensitivity to Nonlinear Amplification,” *Thesis for the Advanced Studies Diploma*, Barcelona, 2006

[2] J. Tellado, L. M. C. Hoo, J. M. Cioffi, “Maximum likelihood detection of nonlinearly distorted multicarrier symbols by iterative decoding,” *IEEE Transactions on Communications, vol. 51, no. 2, pp. 218-228, Feb. 2003.*

[3] M. Deumal, A. Behravan, T. Eriksson, J. L. Pijoan, “Evaluation of Performance Improvment Capabilities of PAPR-reductig mehods,” *Submited to IEEE Transactions on Wireless Communications*, 2006

[4] J. Krajňák, M. Deumal, P. Pavelka, D. Kocur, P. Galajda, J. L. Pijoan, “Multi-user Detection of Nonlinearly Distorted MC-CDMA Symbols by Microstatistic Filtering,” *Wireless Personal Communications*, October 2007, DOI:10.1007/s11277-007-9398-5

[5] S. Nobilet, J. Helard, D. Mottier, “Spreading Sequences for Uplink and Downlink MC-CDMA Systems: PAPR and MAI Minimization,” *European Transactions on Telecommunications*, 2002

[6] S. Hara, R. Prasad, “Overview of Multicarrier MC-CDMA,” *IEEE Communication Magazine*, 1997, vol. 35, p. 126–133

[7] L. Hanzo, M. Munster, B. Choi, T. Keller, *OFDM and MC-CDMA for Broadband Multi-User Communications, WLANs and Broadcasting*, Wiley, 2004, England

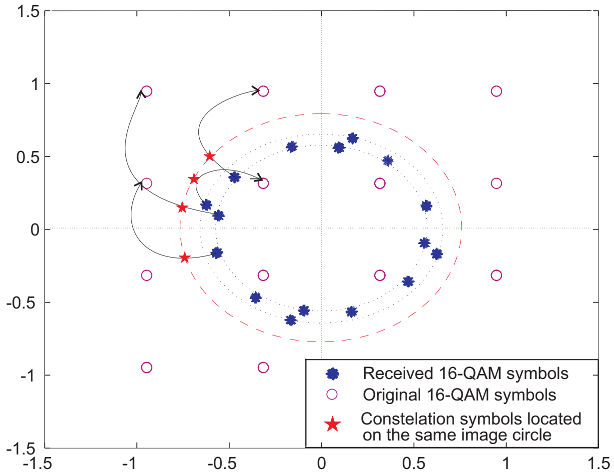



Fig. 3. Original and received symbol constelation at BMF output. Spreading sequences: Walsh codes. Number of active users: 1

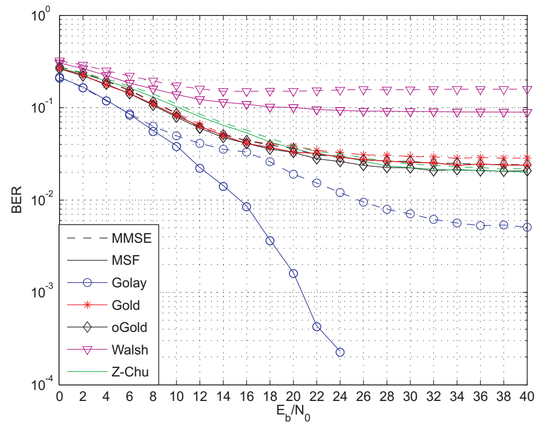



Fig. 6. $BER$ vs. $E_b/N_0$ for MC-CDMA transmission system for different spreading sequences. $IBO = 3dB$.

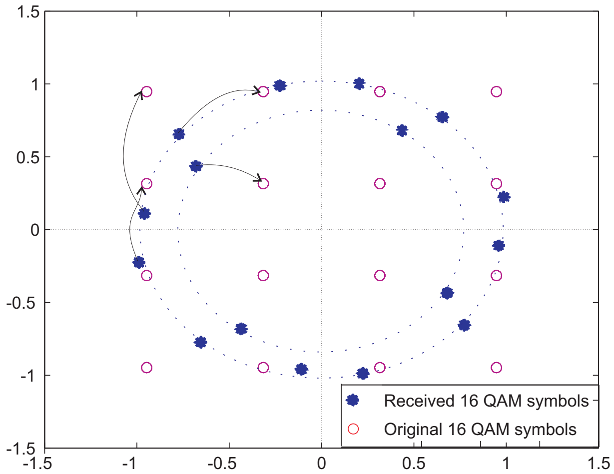



Fig. 4. Original and received symbol constelation at BMF output. Spreading sequences: Golay codes. Number of active users: 1

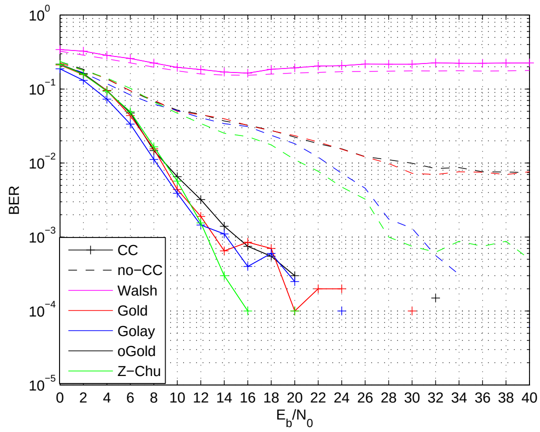



Fig. 7. $BER$ vs $E_b/N_0$ for MC-CDMA transmission system with convolution coding for different spreading sequences. $IBO = 3dB$. Receiver: MMSE-MUD

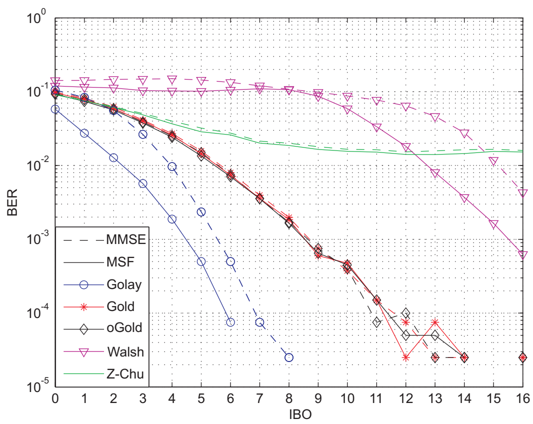



Fig. 5. $BER$ vs. $IBO$ for MC-CDMA transmission system for different spreading sequences. $E_b/N_0 = 17dB$.

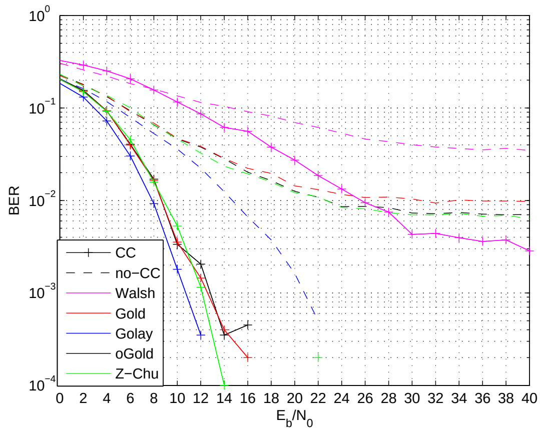



Fig. 8. $BER$ vs $E_b/N_0$ for MC-CDMA transmission system with convolution coding for different spreading sequences. $IBO = 3dB$. Receiver: MSF-MUD.

# Use of Security Features for Computer Network Security

[1]*Alžbeta Kleinová*

[1]Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

[1]alzbeta.kleinova@tuke.sk

***Abstract*—In the present, there is the innumerable amount of various types of attacks, which it is necessary to know. By the analysis of the security solutions it is possible to prevent an attack, if not completely, at least to restrain its. So the article dwells on the analysis of the security solutions which should contribute to the computer network security effectively and by that it refers to the way how it is possible to prevent these attacks or precede them. The research team deals also with problems of use of security solution in the computer network under the VEGA project titled: "Intrusion tolerant security architectures of heterogeneous distributed and parallel computing systems and dynamic computer networks."**

***Keywords*—Firewall, IPS, 802.1x, Security.**

## I. Introduction

The process of computer network security is the serious problem for all types of networks. It is the continuing problem with constant evolution and changes. It is necessary to come from fact that every computer network is permanently under the attack. Whether this attack is known or not, computer network is permanently scanned. After the term security, it is understood the protection against malicious attacks which originate from the intruder. Security also controls the result of the mistakes and device failure [6]. In the article there are analyzed different security solutions which it is possible to implement to the computer network. The choice of the concrete security solutions depends on the structure and topology of the computer network as well as on its analysis.

## II. Security Goals

Before it is implemented any security solution to the computer network, it is important to know what it is necessary to protect in the computer network, or rather, before what we want to protect the computer network. So, it is needful to determine the security goals.

The goal of network security is to keep confidentiality, integrity and availability of data. There are three main goals of network security [2]:

- Confidentiality,
- Integrity,
- Availability.

### A. Confidentiality

Confidentiality relates to data protection before unauthorized revelation by the third side. Whether this data are customer data or internal data of a company, the company is responsible for the privacy protection of this data. All customers have the right to their privacy information be protected. In many cases it is a legal demand. In the best concern of the company, it is needful to keep a relation with its customers by respecting their rights for privacy information. Company information protected by the law should also keep confidential. Moreover, the transmission of this information should be executed by the secure way which would precede whatever unauthorized access.

### B. Integrity

Integrity relates to certainty that data are not changed or destroyed by the unauthorized way. For example, integrity will be preserve, if the sent message will be consistent with the received message. Even for data which are not confidential, the protective treatment measures must be accepted. These measures assure data integrity.

### C. Availability

Availability is defined as the continuing operation of the computer systems. To the application be available, all components must offer the uninterrupted services. These components include application and database servers, store devices and transport end-to-end network.

These three primary goals seem to be very simple. But the task of a secure network can represent very difficult task, if it regards the needs of the society.

## III. Security Solutions

In the present, there is a huge amount of devices that carry out the task of firewall that protect the computer network before outside attacks. Firewall can include the list of various access rights to separate the computer network from the Internet. There are not only the hardware solutions, for example ZyWALL5, ASA or PIX but also the software solutions which are included for example with Netfilter.

*A. ZyWALL 5*

The ZyWALL5 device represents the “gate” for all data which are situated in the network traffic between the Internet and LAN. It is a security solution which can effectively protect the computer network. Its advantage lies in integration of Network Address Translation (NAT), firewall, contents filtering, certificates and Virtual Private Networks (VPN) ability in one device. [4]

*B. ASA*

Cisco Adaptive Security Appliance (ASA) is security product with the wide use. There are integrated the functions of firewall, VPN and Intrusion Prevention System (IPS) and antivirus and antispyware protection of e-mail in one device [5]. It allows standardizing the security functions.

*C. PIX Firewll*

Cisco PIX firewalls are secure, reliable and high-performance hardware state firewalls which have the wide use. They offer the application inspection, network address translation, support VPN, IPS, user authentication, dynamical routing, backup and so on [5]. The creation of security rules is assigned via the creation of objects.

*D. IPS*

Security of information should be based on the influence of the technology stratifying in the whole organization to provide protective shield which restrains the risk and so minimizes the threat. One of these layers is also IPS (Fig. 1).

Intrusion prevention systems represent the security layer, which was set up to combine the protection of firewalls with the ability to monitor the network using Intrusion Detection System (IDS).

In the last twenty years the security technologies were divided into different IDS, firewalls, routers, switches and so on. Each of them works on the different segment of the network, although together provides restraining of the threats and minimizing the risks by using collection of records, rules, policies and configurations [7].

Accessibility of IPS provides simplification of network management problem [1]. IPS offers the ability to identify an intrusion, suitability, effect, direction and appropriate analysis of events and then to send an appropriate information and commands to the firewalls, switches and other network devices in order to reduce the risk of events.

To the network which is suggested on the Department of Computers and Informatics it is not possible to physically implement models such as IDS. This fact follows from the character of the computer network, which is not modular. But, there is possibility for passive monitoring of the computer network. System can only warn against suspicious activities, but it can not prevent them. In the present, there are implemented programs which partially secure the prevention against intrusions to the network. This analysis follows from the suggestion of the VEGA project No. 1/4071/07. In the future the research will focus on the improvement of this problem by using security model.

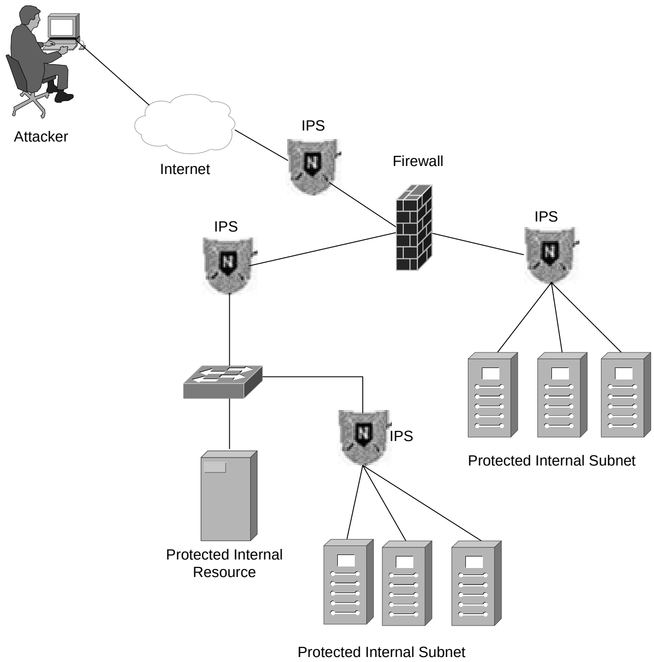


Fig. 1. Intrusion Prevention System

*802.1x*

It is an access control based on the client-server and authentication protocol that restricts an access for unauthorized devices to a LAN through publicly accessible ports [3]. It is a software solution of protection on the routers. The main benefits of 802.1x are following:

- Uses technology based on standards to allow controlling network access.
- Extends authentication to other security areas, for example - authorization, access control, and policy enforcement.
- Controls exercised at link layer, so all services running on it can use link layer services.
- Interoperates in wired, wireless, switching scenarios.
- Reduces overall IT costs by preventing external and internal threats.
- Enables and performs centralized user administration.

## IV. Computer Network on the DCI

The research team deals with all variants of security solution which were mentioned above within the VEGA project No. 1/4071/07. With the help of the analysis of the computer network the research team tries to constitute the model of security solution which will be applied to the computer network on the Department of Computers and Informatics (DCI), Fig. 2, as well as eliminate the defects in the protection of computer networks against the possible attacks from the Internet which it is necessary to daily solve and eliminate.

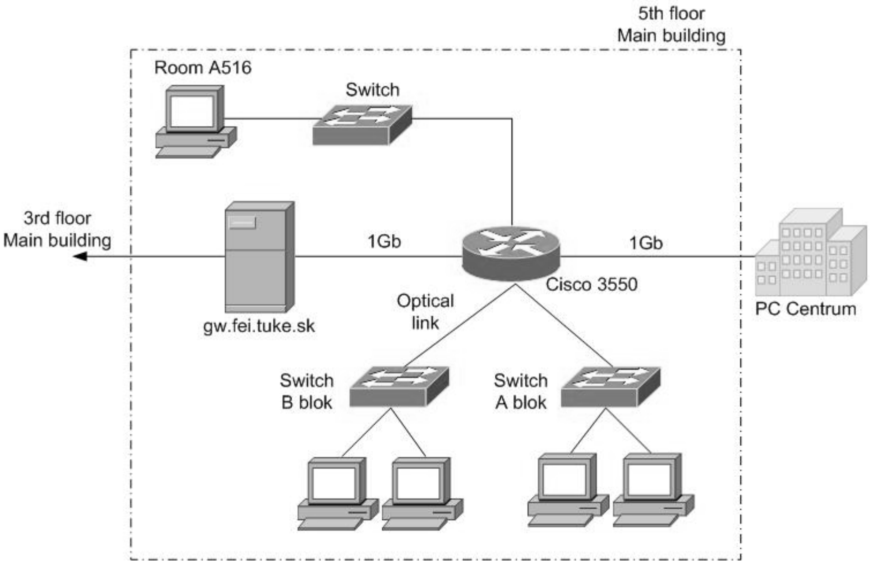



Fig. 2. Computer Network Topology on the DCI

## V. Conclusion

The choice of the suitable security solution depends on the network administrator to determine which security features can implement to the computer network. This is based on the network analysis. It is not possible to implement the same security solution into many networks. Every computer network has own particularities which determine how to proceed by the implementation of the security solution into the computer network. For that reason, it is needful by the suggestion of the security solution to come out from the detailed analysis of computer network which reveals the network defects in the relation to the threats and attacks from the side of the intruder. The research team deals with the detailed analysis as well as the suggestion of the security solution for the network on the DCI within the VEGA project: "Intrusion tolerant security architectures of heterogeneous distributed and parallel computing systems and dynamic computer networks."

Supported by VEGA project No. 1/4071/07

## References

[1] Intrusion Prevention System. SecurityFocus, 2005. Available on the Internet: http://www.securityfocus.com/infocus/1670
[2] M. Oliver, Database privacy: balancing confidentiality, integrity and availability. SIGKDD Explor. Newsl. 4, 2 (2002), pp. 20–27
[3] N. Sudhir: 802.1x Overview, Presentation. 2006
[4] ZyWALL 5 – Device for sexure connection to the Internet. Available on the Internet: http://www.wificentrum.sk/buxus/docs/download/_ ZyWALL5_Manual_SK.pdf
[5] Cisco Systems, ASA/PIX 7.x: Redundant or Backup ISP Links Configuration Example. Sep. 2007. Available on the Internet: http://www.cisco.com/en/US/products/hw/vpndevc/ps2030/ products_configuration_example09186a00806e880b.shtml
[6] L. Vokorokos, Digital Computer Principles, 1.vydanie, Typotex Publish House, Budapest, 2004, ISBN 9639548 09 X
[7] L. Vokorokos, N. Ádám, M. Hauliš, J. Trelová, A. Baláž, M. Chovanec, A. Kleinová, O. Látka, Parallel Computing Systems applied to Secure Architectures and Computer Networks, Computer Science and Technology Research Survey, Košice, Elfa s.r.o., 2007, pp. 25-29, ISBN 978-80-8086-046-2

# Selected Congestion Control Algorithms

[1]Pavol KOCAN

[1]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]pavol.kocan@tuke.sk

***Abstract*— The main reason for implementing the TCP congestion control algorithm is the effective utilizing of network based on the philosophy that a sender should not introduce additional traffic onto a network too rapidly, and that it should respond to network congestion, detected by packet losses, by significantly reducing its transmit rate. In the article presented I introduce the congestion control problematic, how the congestion collapse is occuring in telecommunications networks and how can be handle that. Elementary theory and principles is decribed by several example on typical undesirable failures. General forms of congestion control is offered and binomial algorithm is described more closely.**

***Keywords*—congestion collapse, congestion control, fairness, bottleneck link**

## I. Introduction

Generally several links are sharing the same connection by bottleneck link, which offers specified transfer rate. The packetdrop can occur in the case, when the overall requested transfer rate is a little bit higher then the offered one. Even if buffer is used for regulation, incoming flows packet delays will increase until the congestion collapse. For a short time this can be solved be increasing buffer size, else it is undesirable case for regular and reliable connection. Flow control is neccesary to regulate incoming packet flows and prevent congestion collapse within the network. Flow control enables a source to match its transmission rate to the available service rate at a receiver and in the network. The TCP protocol provides flow control in the Internet. More frequently the term „congestion control” is used.

## II. The Objective Of Congestion Control

Let’s start with example of a simple network exhibiting some inefficiency where sources are not limited by some feedback from the network. We assume that only resources to allocate is link bit rates. Consider first the network illustrated on figure 1. Sources 1 and 2 send traffic to destination nodes D1 and D2 respectively, and are limited only by their access rates. There are five links labeled 1 through 5 with capacities shown on the figure. Assume sources are limited only by their first link, without feedback from the network. Call $\lambda_i$ the sending rate of source i, and $\lambda'_i$ the outgoing rate ( $\lambda_1$ = 100kb/s, but only $\lambda'_1 = \lambda'_2$ =10kb/s, and the total throughput is 20kb/s !). Source 1 can send only at 10 kb/s because it is competing with source 2 on link 3, which sends at a high rate on that link; however, source 2 is limited to 10 kb/s because of link 5. If source 2 would be aware of the global situation, and if it would cooperate, then it would send at 10 kb/s only already on link 2, which would allow source 1 to send at 100 kb/s, without any penalty for source 2. The total throughput of the network would then become $\theta$ = 110kb/s. In complex network scenarios, this may lead to a form of instability known as congestion collapse. [2]

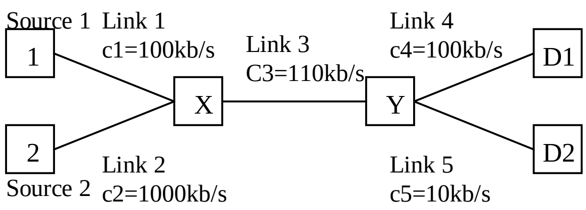



Fig. 1. Simple network

One objective of congestion control is to avoid such inefficiencies. As we will see in the next section, a secondary objective is fairness [2].

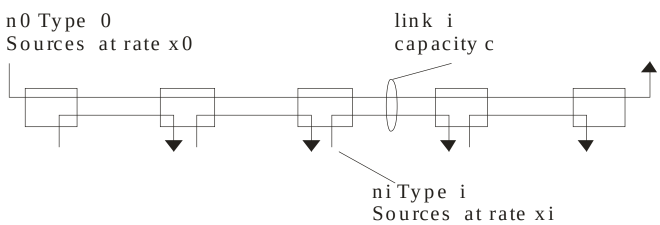



Fig. 2. Simple network for fairness ilustrating

Assume that we want to maximize the network throughput. Consider the network example in Figure 2, where source i sends at a rate $x_i$, i = 0, 1 . . . , I, and all links have a capacity equal to c. We assume that we implement some form of congestion control and that there are negligible losses. Thus, the flow on link i is $n_0x_0 + n_ix_i$. For a given value of $n_0$ and $x_0$, maximizing the throughput requires that $n_ix_i = c - n_0x_0$ for i = 1,..., I. The total throughput, measured at the network output, is thus $Ic - (I - 1)n_0x_0$; it is maximum for $x_0 = 0$ ! The example shows that maximizing network throughput as a primary objective may lead to gross unfairness; in the worst case, some sources may get a zero throughput, which is probably considered unfair by these sources. In summary, the objective of congestion control is to provide both efficiency and some form of fairness. In a simple vision, fairness simply means allocating the same share to all. In the simple case of Figure 2 with $n_i = 1$ for all i, this would mean allocating $x_i$ = c 2 to all sources i = 0, ..., I. However, in the case of a network, such a simple view does not generally make sense.

## III. End-to-end Congestion Control

There are three families of solutions for congestion control: rate based, hop by hop and end-to end. Take a look closer at end-to-end form. A source continuously obtains feedback from all downstream nodes it uses. The feedback is piggybacked in packets returning towards the source, or it may simply be the detection of a missing packet. Sources react to negative feedback by reducing their rate, and to positive feedback by increasing it. The difference with hop-by-hop control is that the intermediate nodes take no action on the feedback; all reactions to feedback are left to the sources. In the example of the previous section, node Y would mark some negative information in the flow of source 2 which would be echoed to the source by destination D2; the source would then react by reducing the rate, until it reaches 10 kb/s, after which there would be no negative feedback.

### *A. Slow start and Congestion avoidance*

In early implementations of TCP before 1988, the sender injected multiple segments up to the window size at once into the network. This immediate burst of segments can cause buffer overflow, hence, packet loss in intermediate queues of routers. Even worse, it can put the connection into a persistent failure mode of continuous retransmissions. Beside the advertised window, the slow start algorithm imposes a second window, called the congestion window W, to the sender. The amount of data allowed to send is the minimum of the advertised window and the congestion window. At connection setup, W = 1 maximum segment size (MSS). Each received acknowledgment increases the congestion window by 1. After 1 round trip time (RTT), 1 ack is received and W = 2, which allows the sender to send 2 segments and causes 2 acks to be returned. After 2 RTT, the congestion window is W = 4. In general, the congestion window size doubles every RTT. Hence, the slow start algorithm provides an exponential increase regulated by the return of acks. The exponential increase in W continues until the slow start threshold (sstresh) is reached. Crossing sstresh makes the sender to enter the congestion avoidance phase. The best sender's policy in the congestion advoidance phase is an additive increase, but multiplicative decrease of his congestion window.

### *B. Additive increase, multiplicative decrease*

As long as there are no losses, the TCP congestion window increases approximately linearly in the congestion avoidance phase. Upon detection of a loss, the fast retransmit algorithm halves the window effectively. If the channel is stable and the loss probability p is not too large, TCP exhibits a sawtooth behavior. This window update rule has been introduced as "Additive Increase, Multiplicative Decrease" (AIMD). The "additive increase" refers to the congestion avoidance phase when signals of congestion (i.e. losses) are absent, while the "multiplicative decrease" reflects the reaction to a congestion signal. Beside the control theoretic argument based on queueing theory, the AIMD rule also leads to inherent fairness among TCP sources: AIMD results in an equal division of the available capacity among identical, competing TCP sources.

Consider a single bottleneck link with capacity C [packets/s] that is shared by two TCP flows with congestion windows $W_1$ and $W_2$ respectively. Assume that the TCP flows are completely synchronized: their round-trip times T are equal and constant, and losses at the bottleneck link occur for both flows at the same time. TCP's self-clocking property and the conservation of packets when no loss occurs implies that, during a round-trip time T, the maximum number of packets arriving at the bottleneck link, $W_1 + W_2$, equals the maximum number of packets transmitted by that link, CT. Hence, the maximum throughput is $CT = W_1 + W_2$. Assume that initially $W_2$ is larger than $W_1$, $W_2(t_0) > W_1(t_0)$. In the congestion avoidance phase, both windows inerease linearly in time with same slope because the round-trip times are assumed to be equal. Hence, $W_i(t) = t/T + W_i(t_0)$ for either i = 1 and i = 2. In the $W_1W_2$ plane, the additive increase corresponds to a trajectory of the system state moving parallel to the diagonal $W_1 = W_2$. When the line $W_1 + W_2 = C.T$ is reached, both flows will experience congestion. The fast retransmit algorithm reduces both windows by a factor of 2. This corresponds to halving the vector **W** with components $(W_1,W_2)$. Subsequently, a new cycle of congestion avoidance begins and the trajectory of the system state will again move parallel to the diagonal, until congestion occurs again, etc.

### *C. Binomial congestion control algorithms*

TCP is not well-suited for several emerging applications including streaming because its reliability and ordering semantics increases end-to-end delays and delay variations. The protocols used by such applications as audio and video streaming must implement "TCP-compatible" congestion control algorithms, which interact well with TCP and maintain the stability of the Internet.[5] Ideal for this are binomial congestion control algortihms a class of nonlinear congestion control algorithms.

They generalize TCP-style additive-increase by increasing inversely proportional to a power k of the current window and also they generalize TCP-style multiplicative-decrease by decreasing proportional to a power l of the current window. There are infinite number of TCP-compatible binomial algorithms with rule k + l = 1 which converge to fairness on the term $k+l>0$; $k,l\geq 0$. Binomial algorithms interact well with TCP across a RED gateway, implicative-increase additive decrease (IIAD) and square root (SQRT) are well-suited to applications that do not react well to large TCP-style window reductions. We call the first IIAD (k=1, l=0) because its increase rule is inversely proportional to the current window, and the second SQRT (k =1/2, l=1/2) because both its increase is inversely proportional and decrease proportional to the square-root of the current window.

The parameter k represents the aggressiveness of probing, while l represents the conservativeness of congestion response of a binomial control algorithm. Small value for k implies that the algorithm is more aggressive in probing for additional bandwidth, while a large value of l implies that the algorithm displays large window reductions on encountering congestion. Thus, it would seem that there is a trade-off between k and l in order for for a binomial protocol to achieve a certain throughput at some loss-rate. Binomial algorithm is TCP-compatible if and only if k+l =1 and l≤1. We call this the k+l rule.

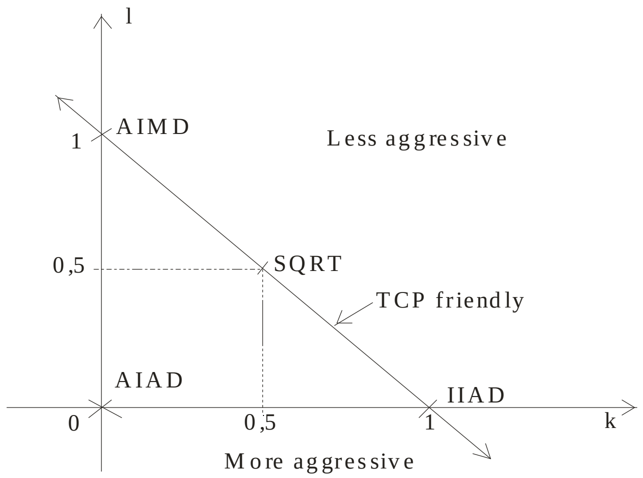



Fig. 3. The (k,l) space of nonlinear controls

Algorithm should be converge to fair and efficient operating point $x_1=x_2=1/2$. When k=0, the increase is additive, when k>0, the increase curve lies below the k=0. Increase improve efficiency, since $x_1+x_2$ increases, and moves towards a fairer allocation.

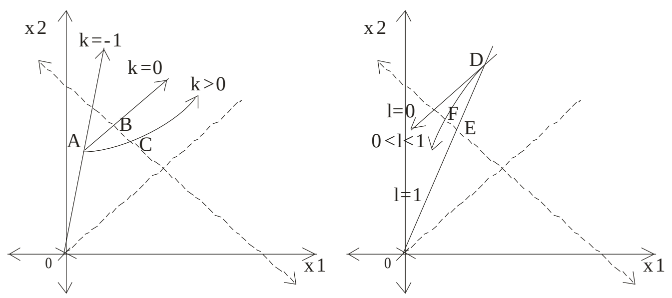



Fig. 4. Path showing the convergence to fairness

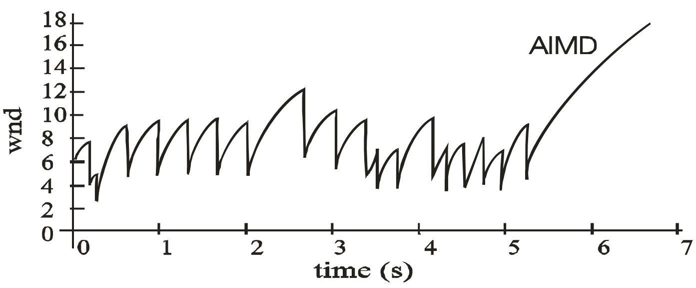
When l=0, additive decrease, the window reduction occurs along the 45°-line, worsering fairness. When l=1, the decrease is multiplicative and moves along the line to the origin without altering fairness. For 0<l<1, the window reduction occurs along a curve with the shape shown in the middle picture of Figure 5, this curve evolve to an under-utilized region. The key to convergence argument is to observe that the successive points of intersection of a binomial evolution curve with the maximum-utilization line always progress toward the fair allocation point, when $x_1 > x_2$, this continually moves downwards, and when $x_1 < x_2$, it continually moves upwards, towards the $x_1=x_2=1/2$ point. Once $x_1 = x_2$, a binomial algorithm behaves like a linear algorithm, moving on the equi-fairness line, though with different magnitude of oscillation depending on k and l.

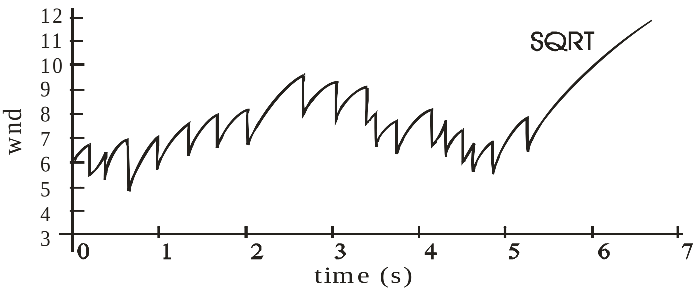
According to simulation results done by Bansal and Balakrishnan [5], where one connection is TCP and the other a binomial algorithm parametrized by k, IIAD or SQRT obtain higher long-term throughput than TCP across a drop-tail bottleneck gateway, because of a higher average buffer occupancy. Active queue management schemes like RED allow binomial algorithms and TCP to interact well with each other, which may be viewed as another among many important reasons to eliminate drop-tail gateways from the Internet infrastructure.

IIAD connections are fair to TCP and to themselves across a wide range of loss-rates. At very high loss-rates, TCP gains because of its more aggressive probing for bandwidth, and it is able to recover faster from burst losses. Similar results were obtained with SQRT algorithm that gave faster convergence to fairness and hence was fairer to TCP over smaller time scales than IIAD. AIMD mechanism causes the noticeable amount of oscillation in transmission rate even when the loss rate is constant. On the posittive side, it responds fast to changes in loss rate. IIAD have a negligible oscillations, but is slower to respond to bandwith changes. SQRT congestion control mechanism have a smaller magnitude of oscillations than what AIMD would observe.

## IV. Simulations

Three congestion control algorithms was simulated [6]. SQRT and IIAD algorithms will be compared with the basic AIMD congestion control algorithm in this section. Important in all cases is that the sender adjusts its sending rate according to receiver's feedback information. Look closer for the first five seconds in graphs.



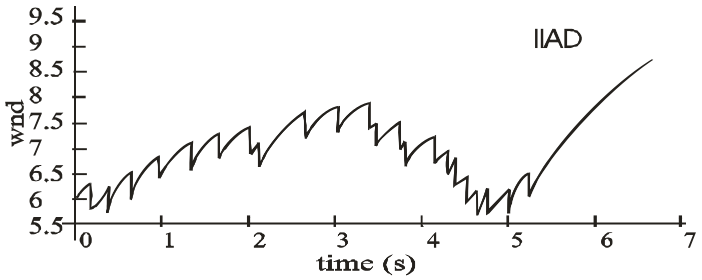



Fig. 5. Increase/decrease of the TCP sending window (wnd)

Sending rates starts at the same value, the six times of the basic sending window. AIMD have a fluent congestion regulation, after detecting the congestion the decrease is multiplicative, usually the transfer rate gets down to the half rate. IIAD and SQRT congestion control algorithms achieves little more transfer rate at start, but greater window increases, which is more suitable for streaming transfer where the variable data rate is used. Simulations also shows that SQRT algorithm has globally more transfer rate as IIAD.

## V. Conclusion

Problematic of TCP Congestion control is much more extensive so only brief overview was described in this article. In streaming media aplications, sent over the internet with TCP, can suffer from various fluctuations and low utilization due to TCP's AIMD congestion control mechanism that was found unsuitable for video applications. Therefore smoother congestion control methods such as binomial algorithm has been developed. Improved methods and apparatuses are provided in which logarithm-based rate control algorithms are employed to better utilize available bandwidth or control and smooth the sending rate of the streaming media. Such a algorithm is binomial congestion control implemented in new H.264 codec, with faster transmission rate it eliminating the loss retransmission by removing the feedback mechanism for each packet. The next development of the binomial algorithms is focused right at this ubiquitous codec.

## References

[1] H.J.Chao,X. Guo: Quality of service control in high-speed networks, John Wiley & Sons, 2002, ISBN: 0-471-22439-1
[2] J.Y. Le Boudec: Rate adaptation, Congestion Control and Fairness: A Tutorial, 2007, Ecole Polytechnique Federale de Lausanne (EPFL)
[3] P. Gevros,J. Crowcroft,P. Kirstein,S. Bhatti: Congestion Control Mechanisms and the Best Effort Service Model, University College London, IEEE Network, May/June 2001
[4] P.V. Mieghem: Data Communications Networking, Techne Press, Amsterdam 2006, ISBN: 90-8594-008-7
[5] D. Bansal, H. Balakrishnan: Binomial Con-gestion Control Algorithms, MIT Laboratory for Computer Science, Cambridge, IEEE Infocom 2001
[6] Y.Xiaogang, L.Lei: End-to-end Congestion Control for H.264/SVC. Sixth International Conference on Networking (ICN ’07) IEEE, 2007

# Utilization of UML models for the ease of maintenance process of software systems

*Ján Kunštár*

Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

jan.kunstar@tuke.sk

***Abstract*—The need for software changes after delivery is here from the beginning of electronic calculations. These changes are a consequence of the nature of software and the changing environment in which it is used. The change of software system after delivery is called software maintenance. Development in the area of software maintenance is still falling behind the development in software systems. This is the reason why the interest of a research is growing up in the maintenance area in the present. The research tries to find methods and tools which could streamline and speed up maintenance process. This article analyzes present state in the software maintenance, with emphasis on current problems in this area which need to be eliminated to decrease the great coast of software maintenance (and through it also a total cost of software life cycle) in the future. Knowledge about software system is essential in the maintenance process. The easiest and the most rapid way how to reach them are abstract models. Therefore article also analyzes UML models which could be used as easy readable and understandable source of essential knowledge about software system.**

***Keywords*—maintenance, models, software artifacts, software system, UML.**

## I. Introduction

During the past few decades, there has been a proliferation of software systems in a wide range of working systems. Among the sectors of society which exploit these systems are manufacturing industries, financial institutions, information services and construction industries. There is an increasing reliance on software systems and correct use and functioning of software system can be a matter of life and death. Therefore the factors like functionality, flexibility, continuous availability, correct operation and security are required from software systems.

According to the first Lehman law of the software evolution [9]: "A program that is used must be continually adapted else it becomes progressively less satisfactory."

The need of these changes is either a result of changing environment in which the system is used or changing of user requirements.

Maintenance is undoubtedly the most expensive stage of software system life cycle. By this way software maintenance influences total software expense, and therefore solutions that could improve maintenance productivity are bound to have a dramatic impact on software costs and the overall profitability of companies. It is interesting to note that advancements in software technology over the last two decades, as well as much research on maintenance methods, have not improved the situation. The cost of maintenance, rather than dropping, is on the increase [8].

In the present, the research demands lead to development of strong maintenance mechanism, which will be fast, cheap, reliable and secure, and also will be able to exploit knowledge and experiences.

The paper is organized as follows. First, essential information about software maintenance are briefly analyzed in Section 2. Then, Section 3 introduces importance of knowledge in the maintenance process. Section 4 introduces UML models and the importance of visualization in maintenance. Finally, Sections 5 presents the conclusions.

## II. Software Maintenance

### A. Definition

Traditionally we use the term "software maintenance" for naming the discipline concerned with changes related to software system after delivery. An appreciation of this discipline is important especially because the cost is now extremely high. Safety and cost of software maintenance mean that there is an urgent need to find ways of reducing or eliminating maintenance problems [6].

Many definitions of software maintenance exist. Some take a specific view and others take more general view. The IEEE Standard for Software Maintenance [7] defines maintenance as: "Modification of a software product after delivery to correct faults, to improve performance or other attributes, or to adapt the product to a modified environment."

Several authors disagree with this view and affirm that software maintenance should start well before a system becomes operational. The view that maintenance is strictly a post-delivery activity is one of the reasons that make maintenance hard. Therefore Thomas Pigoski[1] [11] formed new definition:

"Software maintenance is the totality of activities required to provide cost-effective support to a software system. Activities are performed during the pre-delivery stage as well

[1] THOMAS M. PIGOSKI is Senior Software Engineer at TECHSOFT, Project Editor and primary author of ISO's proposed International Standard on Software Maintenance, and General Chair of the IEEE International Conference on Software Maintenance.

as the post-delivery stage. Pre-delivery activities include planning for postdelivery operations, supportability, and logistics determination. Post-delivery activities include software modification, training, and operating a help desk."

According to Pigoski's definition of software maintenance, it is very important to prepare software system for its modifications still during the development of the system and not only after delivery. Erasure of the border between software development and maintenance will make maintenance process more effective and will decrease the amount of costs needed for maintenance process.

### *B. Problems of Maintenance*

Software maintenance task is the most expensive part of the software system life cycle. The surveys [4], [12] indicate that software maintenance consumes 60% to 90% of the total life cycle costs. Survey also shows that around 75% of the maintenance effort is on the enhancements, and error correction consumes about 21% (Fig. 1).

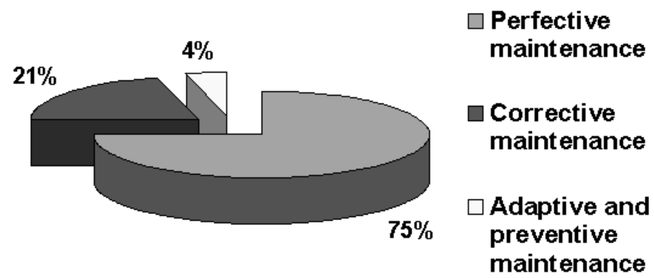



Fig. 1. Costs of maintenance activities.

This proves that the maintenance cost is not only a function of poor design, but mostly a function of the changing customer or environmental requirements and the manner in which the system was constructed [1].

Since we aren't able to stop development of the software we can't eliminate the need of maintenance. According to survey the maintenance costs are largely due to enhancements (75%), rather than corrections. Therefore even if we develop systems without errors, the maintenance will be always needed for improving these systems, for supplying the requirements for system changes of users and constantly changing environment. This is the reason why the research in the future would aim to improvement of maintenance, instead of trying to eliminate it.

The most challenging problems of software maintenance are [3]:

- program comprehension
- impact analysis
- regression testing

**Program comprehension** is a well known problem of software maintenance, which increases the costs for maintenance a lot. Whenever a change is made to a piece of software, it is important that the maintainer gains a complete understanding of the structure, behavior and functionality of the system being modified. As a consequence, maintainers spend a large amount of their time (from 50% up to 90% of total maintenance time) reading the code and the accompanying documentation to comprehend its logic, purpose, and structure. Program comprehension is frequently compounded because the maintainer is rarely the author of the code. On the other side many times also a programmer who developed system isn't able change system without studying documentation and the code due to understanding its own work.

**Impact analysis** is one of the major challenges in software maintenance in the present. It tries to determine the effects of a proposed modification on the rest of the system. Impact analysis [5] is the activity of assessing the potential effects of a change with the aim of minimizing unexpected side effects. The task involves assessing the appropriateness of a proposed modification and evaluating the risks associated with its implementation, including estimates of the effects on resources, effort and scheduling. It also involves the identification of the system's parts that need to be modified as a consequence of the proposed modification.

**Regression testing** is a retest of the system after a change has been implemented to gain confidence that it will perform according to the (possibly modified) specification. The process of testing the system after it has been modified is called regression testing. The aim of regression testing is twofold: (1) to establish confidence that changes are correct and (2) to ensure that unchanged portions of the system have not been affected.

An inconsistent state of the software's artifacts markedly contribute to three mentioned problems. Each software system consist of artifacts which describes only a limited part of the software and the actual system is their composite (Fig. 2). When a software system is changed, each artifact which is influenced by this change has to be modified for preservation of software maintainability[2].

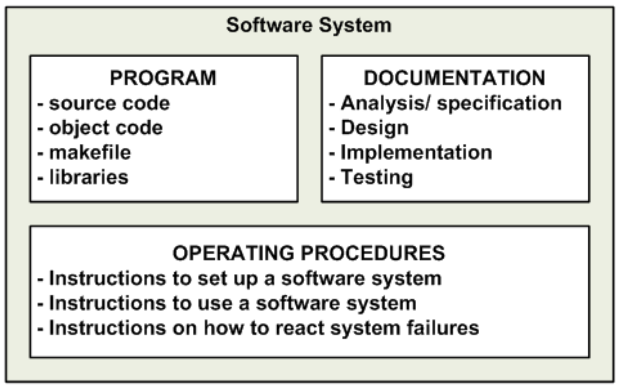



Fig. 2. Artifacts of software system.

## III. KNOWLEDGE IN THE MAINTENANCE PROCESS

For each member of maintenance personnel is very important to know and understand to software system which

[2] MAINTAINABILITY–the ease with which a software system or component can be modified to correct faults, improve performance, or other attributes, or adapt to a changed environment [7].

he wants to change - he needs knowledge about certain software system. For the maintenance process is important especially knowledge about the artifact which need to be changed and also knowledge about the effects of proposal changes to another artifacts of software system.

Essential knowledge for maintenance process is critical knowledge and general knowledge.

**Critical knowledge** – is knowledge related with the certain software system (are useful only for maintenance of this system). Next knowledge belongs to this category:

- *relationships and dependences among software system artifacts*
- *relationships and dependences among components of diverse software system artifacts*
- *knowledge about made changes on previous versions of software system*
- *knowledge about connectivity between software system and environment in which system is used*
- *knowledge about the data representation*

The main problem of the software systems in the present is, that critical knowledge are stored separately from the source code, which lead to longer time required for accessing and processing this knowledge. Therefore is important to find their common representation for ease of software maintenance in the future.

**General knowledge** – is knowledge useful for all or certain category of software systems. On the basis of the feedback from wide group of software systems which use this knowledge, knowledge could react on the changes in the design and implementation and could be updated at the basis of new trends and technologies.

Knowledge about software system could be acquired from artifacts of this system. Although two artifacts contain the same knowledge, the times needed for acquiring them from one and from another may differ a lot. Therefore the artifacts from which programmers could obtain knowledge in the shortest time are important for maintenance process. Appropriate artifacts for obtaining knowledge during maintenance process are abstract models. They are easy readable and understandable source of essential knowledge about software system and therefore are regarded as future of maintenance of software systems.

## IV. VISUALIZATION AND UML MODELS

### A. UML Models

The Unified Modeling Language (UML) is a general purpose visual modeling language for systems. Although UML is most often associated with modeling OO software systems, it has a much wider application than this due to its inbuilt extensibility mechanisms. UML was designed to incorporate current best practice in modeling techniques and software engineering [2].

When we are talking about UML, it is important to realize that it does not give any kind of modeling methodology, it just provides a visual syntax that could be used to construct models. A methodology is the Unified Process (UP), which tells us the workers, activities, and artifacts that we need to utilize, perform or create in order to model a software system.

In UML the models are used for describing structural or behavioral parts of the software system. The **model** is the repository of all things (structural, behavioral, grouping, annotational) and relationships that you have created to help describe the required behavior of the software system you are trying to design [2].

**Diagrams** are windows or views into the model and also are the most used device for inserting new things to UML models. There are thirteen different types of diagrams in UML 2 (Fig. 3).

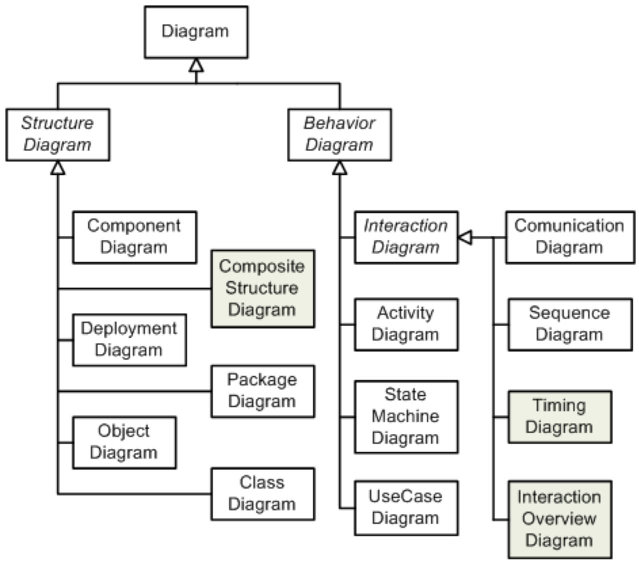



Fig. 3. UML 2.0 diagrams. Each box represents a type of diagram. When the text in the box is italic, it represents an abstract category of diagram types. Shaded boxes indicate concrete diagram types that are new in UML 2.

In the present UML diagrams are widely used in software engineering. Popularity of UML diagrams result from the fact that these diagrams are human-readable and yet are easily rendered by computers.

### B. Utilization of UML Models in Maintenance Process

As was mentioned in Section 3, acquiring of essential knowledge for maintenance process consumes great portion of time and with it increases costs of software system maintenance. One way how to decrease the time required for maintenance process and by this way also decrease total cost of software maintenance is visualization. The role of visualization is during maintenance more important than during development [10]. The proof for this statement is a lot of tools which use visualization for supporting software maintenance, e.g. xVue [14], VIFOR 2 [13]. In this case visualization helps to understand to software system in shorter time.

UML models could be also utilized in maintenance process in different way. As was already mentioned before, uncovering the impacts of made changes (required by users or changing environment) is another challenging problem of software maintenance in the present.

In ideal case all impacts of required changes would be uncovered even before their implementation. It would allow implementation of required changes together with the changes eliminating impacts of these required changes. For this purpose the UML models could be used. On the basis of dependencies among various components of UML models is possible to uncover the chain reaction of the impacts of made changes.

In my research I would like to engage in Model Driven Maintenance (MDM), in which the impacts of required changes will be identified by comparison of abstract models (before and after implementation of required change) and at the basis of knowledge obtained from UML models the recommendations for maintenance personnel will be generated.

Since UML models in this case represent artifacts of software systems, which are source of critical knowledge, is important to store them together with source code in extended source files.

## V. Conclusion

We can often encounter with an opinion, that software maintenance is still inefficient and ineffective. This opinion has the basis in high costs of software maintenance (the highest of all stages of software life cycle). The costs of software maintenance still arise in spite of new technologies and solutions in area of software maintenance were built. Therefore the importance of research in the area of software maintenance doesn't terminate and constantly are finding solutions how to decrease maintenance costs, speed up and increase efficiency of maintenance process.

Utilization of the UML models could streamline maintenance process and decrease the time needed for maintenance of certain software system. Therefore in my future research I will work on Model Driven Maintenance based on UML models.

## Acknowledgment

This work was supported by VEGA Grant No. 1/0350/08 Knowledge-Based Software Life Cycle and Architectures.

## References

[1] Allen C., "Software maintenance – an overview", in British Computer Society, Programming & Software Articles, en-GB 3rd, February 2006.
[2] Arlow J., Neustadt I., "UML 2 and the unified process", Second Edition, Addison-Wesley, 2005, ISBN 0-321-32127-8.
[3] Canfora G., Cimitile A., "Software Maintenance", in Handbook of Software Engineering and Knowledge Engineering, volume 1,World Scientific, ISBN: 981-02-4973-X, December 2001.
[4] Erlikh, L., "Leveraging Legacy System Dollars for E-Business", in IEEE IT Pro, May–June 2000, pp. 17–23.
[5] Fyson, M. J., Boldyreff, C., "Using Application Understanding to Support Impact Analysis", in Journal of Software Maintenance – Research and Practice, 10(2):93-110, 1998.
[6] Grubb P.A., Takang A.A., "Software Maintenance: Concepts and practice", SE, ISBN 978-981-238-425-6, World Scientific 2003.
[7] IEEE Std. 610.12, "Standard Glossary of Software Engineering Terminology", IEEE Computer Society Press, Los Alamitos, CA, 1990.
[8] Jarzabek S., "Effective Software Maintenance and Evolution: A Reuse-Based Approach", in Auerbach Publicatio, Taylor & Francis Group, ISBN: 0-8493-3592-2, 2007.
[9] Lehman M. M.: "Lifecycles and the Laws of Software Evolution", in Proceedings of the IEEE, Special Issue on Software Engineering, 19:1060-1076, 1980.
[10] Oman Paul, "Maintenance Tools", IEEE Software, vol. 07, no. 3, pp. 59-65, May/Jun, 1990.
[11] Pigoski T. M., "Practical Software Maintenance – Best Practices for Managing Your Software Investment", John Wiley & Sons, New York, NY, 1997.
[12] Port, O., "The Software Trap — Automate or Else", in Business Week, 3051(9), 1998, pp.142–154.
[13] Rajlich V., Adnapally S.R., "VIFOR 2: a tool for browsing and documentation", icsm, p. 296, 12th International Conference on Software Maintenance (ICSM'96), 1996.
[14] "Telcordia Software Visualization and Analysis Toolsuite User's Manual", Telcordia Technologies, Inc., 1998.

# Fuzzy logic application by Real Time system

*Peter MACKO*

Dept. of Electrotechnical, Mechatronics and Industrial Engineering, FEI TU of Košice, Slovak Republic

peter.macko@tuke.sk

***Abstract*—This paper deals with a modeling and control of an induction motor (IM) by the fuzzy logic. There is designed the IM fuzzy model using the data obtained by the real system and the fuzzy inverse model based control of it. The results obtained by the simulation show that the proposed control structure can be applied to the real asynchronous motor too. This control strategy is applied by the Real System RT-LAB.**

***Keywords*—fuzzy logic, inverse model based control, Real Time system.**

## I. PROBLEM DESCRIPTION

Controlled system consists of an induction motor and a frequency converter (FC) as is shown in Fig. 1. This system is nonlinear. Fuzzy inverse model based control is designed in order to compensate the system's non-linearity.

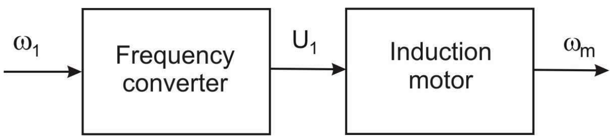



Fig. 1. Controlled system

## II. FUZZY MODEL

In Fig.2 is a fuzzy model structure, where $\omega_1$ is a stator angular frequency, $\omega_m$ is a mechanical angular frequency and $M_m$ is a mechanical torque of a IM.

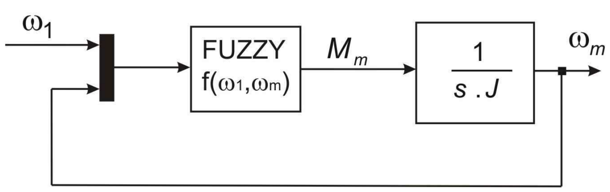



Fig. 2. Fuzzy model of controlled system

On the real system were made experimental measurements in order to obtain its qualitative information. The qualitative information is a simplified speed torque characteristic. In order to identification, we need to know several points on the characteristic. In Fig. 3 is characteristic surface describing fuzzy system properties.

## III. FUZZY INVERSE MODEL BASED CONTROL

In Fig. 4 is shown control structure based on the inverse fuzzy model. The function $f(\omega_1, \omega_m)$ is a controlled system and function $g(M_m, \omega_m)$ is a controller. The surface of

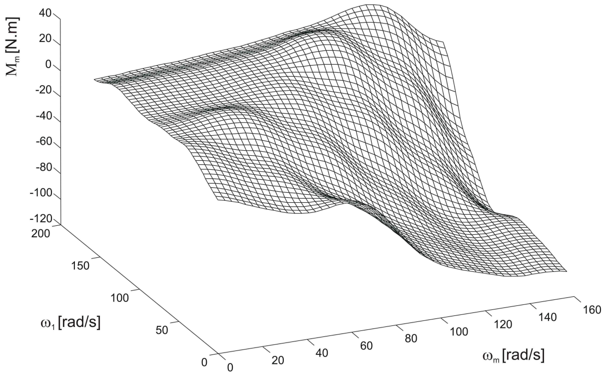



Fig. 3. Surface described controlled system

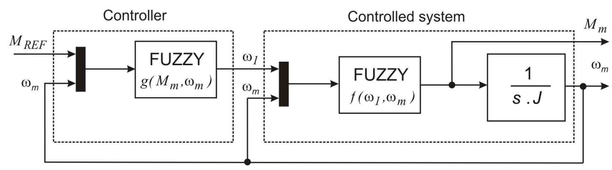



Fig. 4. Inverse control structure

function $g(M_m, \omega_m)$ was created by axis exchange ($M_m \leftrightarrow \omega_1$), which is shown in Fig. 5

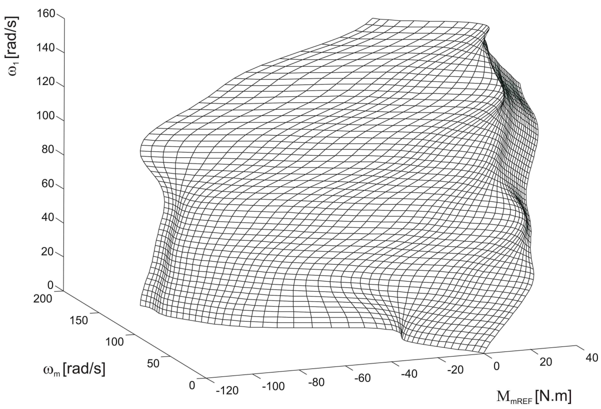



Fig. 5. Surface described controller

The data for fuzzy system $g(M_m, \omega_m)$ were obtained from surface in Fig. 5 by the function *griddata* (Matlab Package). The obtained database of points is consistent, what is presented in Fig.6.

The fuzzy system (Fig.7) was created from consistent points database by the tool *anfisedit* (Matlab Package).

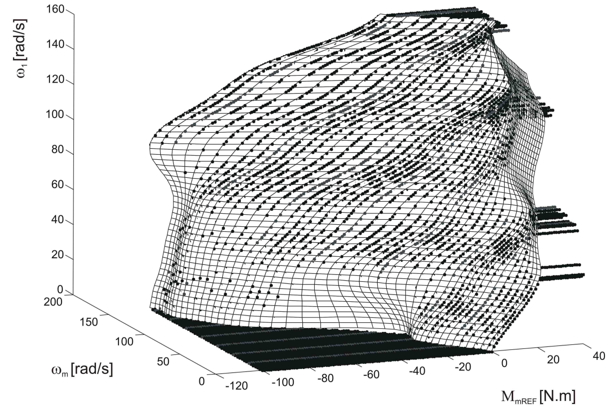



Fig. 6. Points database for function *g*

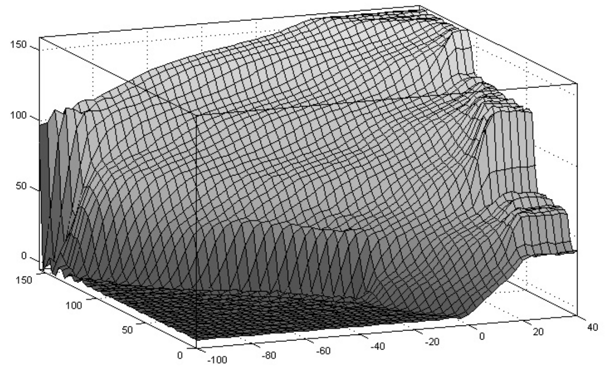


Fig. 7. Surface of fuzzy controller

## IV. REAL TIME SYSTEM

For this application was used Real Time system by OPAL-RT. RT-LAB allows to readily convert Simulink or SystemBuild models to real-time simulations over one or more PC processors, particularly for Hardware–in–the–Loop (HIL) applications. It runs on a networked PC configuration consisting of a Command Station, Target Node, Communication Links (real-time and Ethernet) and I/O boards. **Command station** allows users to prepare the model for distributed real-time execution, control the compilation and execution, and interact with the simulation at run-time. For real-time simulation in **target node** is required a Real-Time Operating System such as QNX or RedHawk Real-Time Linux. Target node connect to the real world through I/O boards.

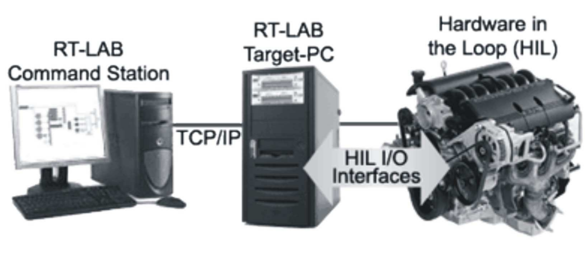



Fig. 8. Single target configuration

### A. Model preparation for distributed Real-Time execution in Simulink

The model must be grouped into two subsystems. **The Console Subsystem** contains all user interface blocks. There can only be one console block in the system. **The Master Subsystem** contains the computational elements of the model that will be run on the target processor. There can only be one master block in the system. In Fig.9 is shown fuzzy control configuration applied in RT-LAB. Output of the console subsystem is the reference value of the motor torque. In the console (Fig.10) we can see quantities as actual value of angular velocity, controller actuating quantity and other special signals.

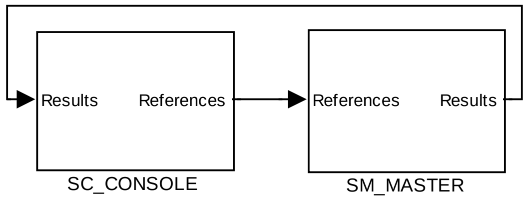



Fig. 9. Fuzzy control configuration in RT-LAB

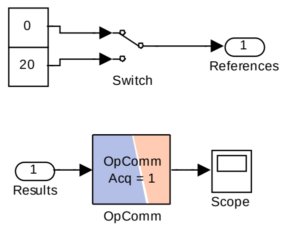



Fig. 10. The console subsystem

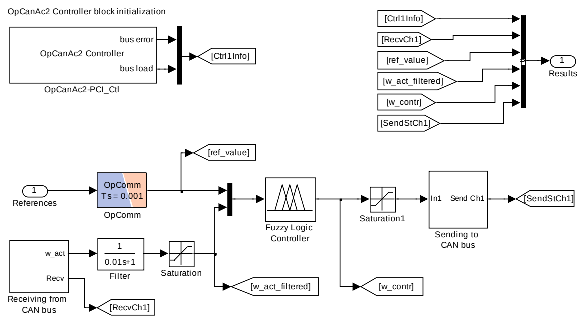



Fig. 11. The master subsystem

In the master subsystem (Fig.11) is control structure with the fuzzy logic controller, saturations and filter for actual value of angular velocity.

For the transfer data between Real-Time system and the frequency converter is used CAN bus. In the Real-Time system is CAN bus card CAN-AC2-PCI by SOFTING. In the frequency converter is special CAN communication card. A data transfer is asynchronously, it means that both nodes transmit a frames periodically. The transmit period is 1 millisecond. In one word, which is transmitted by Real-Time system is saved the controller actuating quantity. In one word, which is transmitted by frequency converter is saved the actual value of motor angular velocity.

## V. Experimental Results

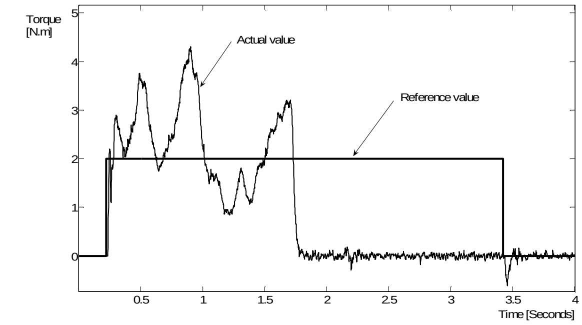



Fig. 12. Reference and actual value of the motor torque

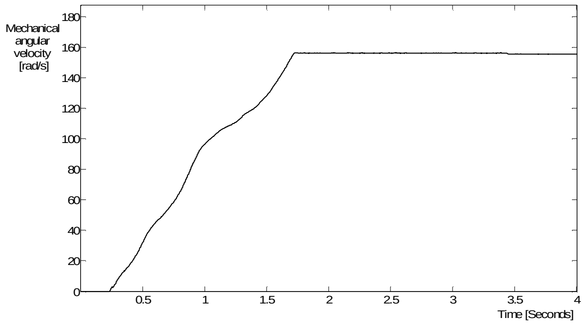



Fig. 13. Mechanical angular velocity

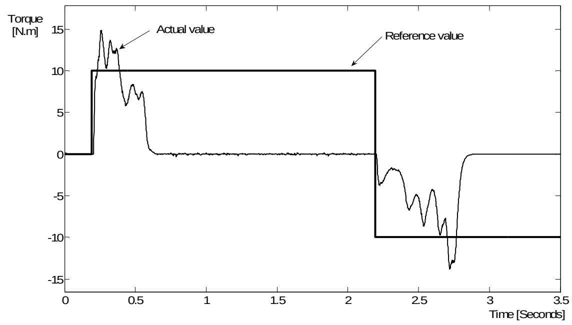



Fig. 14. Reference and actual value of the motor torque

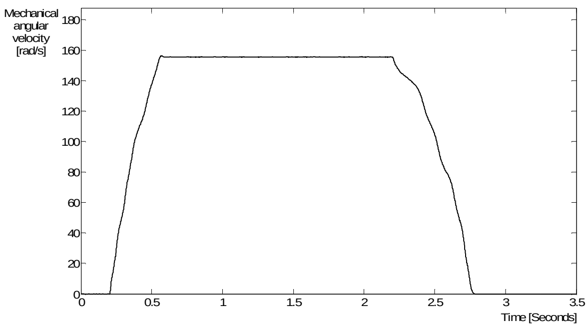



Fig. 15. Mechanical angular velocity

Required motor torque (reference value) *2N.m* is shown in Fig.12. The actual value of motor torque is changing from *1N.m* to *4N.m* up to the time *1,8s.* The problem presents fuzzy logic controller which has been created from an inverse fuzzy model of controlled system. It means that an error has been created during the fuzzy model designing and especially during the real system identification. From the time *1,8s* the actual value of motor torque is zero, what is caused by achieving of maximal value of angular velocity (Fig.13). The similar working states are shown in Fig.14 and Fig.15. The required value of motor torque is *10N.m* and then *-10N.m.* Behavior of the controller is similar as in previous case. In one part the motor torque is moving about required value and then is equal zero. The zero is caused by achieving of limited value of angular velocity.

## VI. Conclusion

A fuzzy inverse model based control strategy was supposed in this paper. As is shown in experimental results there is a big difference between required and actual value of motor torque. But the aim of this paper was to show a basic functionality of this method. Improving presented control strategy will be the main aim of my dissertation work.

## References

[1] Macko P. : Fuzzy modelovanie a riadenie mechatronických systémov. Písomná práca k dizertačnej skúške. KEMPI FEI TU Košice.

[2] Fedor, P.: Aplikácie fuzzy logiky pri riadení procesov. Mercury-Smékal, Košice 2001, s. 86, ISBN 80-89061-32-X.

[3] Babuška R.: Promising Fuzzy Modeling and Control Methodologies for Industrial Applications, Delft University of Technology, Faculty of Information Technology and Systems Control Engineering Laboratory.

[4] Andri Riid - Transparent Fuzzy Systems: Modeling and Control, Tallinn, April-September 2001, December 2001, February-March 2002.

[5] Babuška R. - Fuzzy Modeling for Control. Kluwer Academic Publishers, Boston, 1998

[6] Babuška R.- Fuzzy Systems, Modeling and Identification, Delft University of Technology, Faculty of Information Technology and Systems Control Engineering Laboratory

[7] Abonyi J. - Fuzzy Model Identification for Control. University of Veszprém, Hungary, January 2003 / 288 pp. / 132 ill. / Hardcover ISBN 0-8176-4238-

[8] Abonyi J., Andersen H., Nagy L., Szeifert F. - Inverse Fuzzy-Process-Model Based Direct Adaptive Control, Department of Chemical Engineering Cybernetics, University of Veszprém, Veszprém, H-8201, POB. 158, Hungary

[9] Garcia, C.E. and M. Morari (1982). Internal model control. 1. a unifying review and some new results. Ind. Eng. Chem. Process Des. Dev. 21, 308–323

[10] Chenging Xu: Intelligent multivariable control system, Dissertation work, Faculty of Purdue University, West Lafayette, Indiana, December 2006

[11] Fedor P., Perduková D.: - Použitie II. Ljapunovovej metódy v regulovaných elektrických pohonoch. ISBN 80-967 636-3-6

# Securing Data in Information Environment

*František MRÁZIK*

Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

fm@ist.sk

**Abstract—Electronical information are represented by different forms: stored in different file types, databases, XML documents. Achievement of integrity, uniqueness and access to informations is realized by environment where informations are stored, e.g.: operating system access privileges (read, execute, write to files), database privileges (table or record creation, drop database ) or access gived direct by application (access to information system, its modules and specific action over data). In these environment information is „covered" or „drived" by operating system, database system or application. Information protection is fully under system control. System: authentify user / data processor, authorize specific processor activities , gives user access and limit access by time or quantity. To guarantee the confidentiality, integrity, and availability of data, it is needed to ensure that all components that have access to data are secure and well behaved.**

## I. Securing Application Environment

### A. Confinement

Process confinement is used to restrict the actions of a program. Simply put, process confinement allows a process to read from and write to only certain memory locations and resources. The operating system, or some other security component, disallows illegal read/write requests. If a process attempts to initiate an action beyond its granted authority, that action will be denied. In addition, further actions, such as logging the violation attempt, may be taken. Systems that must comply with higher security ratings most likely record all violations and respond in some tangible way.

### B. Bounds

Each process that runs on a system is assigned an authority level. The authority level tells the operating system what the process can do. In simple systems, there may be only two authority levels: user and kernel. The authority level tells the operating system how to set the bounds for a process. The bounds of a process consist of limits set on the memory addresses and resources it can access. In most systems, these bounds segment logical areas of memory for each process to use. It is the responsibility of the operating system to enforce these logical bounds and to disallow access to other processes. More secure systems may require physically bounded processes.

### C. Isolation

When a process is confined through enforcing access bounds, that process runs in isolation. Process isolation ensures that any behavior will affect only the memory and resources associated with the isolated process. These three concepts (confinement, bounds, and isolation) make designing secure programs and operating systems more difficult, but they also make it possible to implement more secure systems.

### D. Controls

To ensure the security of a system, you need to allow subjects to access only authorized Common Security Models, Architectures, and Evaluation Criteria uses access rules to limit the access by a subject to an object. Access rules state which objects are valid for each subject. There are both mandatory and discretionary access controls. With mandatory controls, static attributes of the subject and the object are considered to determine the permissibility of an access.

Both mandatory and discretionary access controls limit the access to objects by subjects. The primary goals of controls are to ensure the confidentiality and integrity of data by disallowing unauthorized access by authorized or unauthorized subjects.

## II. Securing Information In Heterogenous Environment

Information / data flow doesn't end at the gates of operating systems or databases. Data flow between different information system, applications, platforms. System, where data / document was created loose control over authentification, authorization and operations over document. Information flow in different forms, e.g.: ftp, http, prílohy email, tcp/ip data-flow, etc.

Here we ask: How to ensure confidentality, integrity and availability - information safety in (unsecured) heterogenously systems? Main ways how to protect these data is to crypt and/or sign them.

## III. Asymetric Cryptography

Asymmetric cryptosystems is using pairs of public and private keys to facilitate secure communication. The security of these systems relies upon the difficulty of reversing a one-way function.

Cryptosystems rely on pairs of keys assigned to each user of the cryptosystem. Every user maintains both a public key and a private key. As the names imply, public key cryptosystem users make their public keys freely available to anyone with whom they want to communicate.

The private key, on the other hand, is reserved for individual. It is never shared with any other cryptosystem user. The sender encrypts the plaintext message with the recipient's public key to create the ciphertext message. When the recipient opens the ciphertext message, they decrypt it using their private key to re-create the original plaintext message. Once the sender encrypts the message with the recipient's public key, no user (including the sender) can decrypt that message without knowledge of the recipient's private key (the second half of the public-private key pair used to generate the message).

## IV. Digital Signatures

Digitally signed messages assure the recipient that the message truly came from the claimed sender and enforce nonrepudiation and the recipient that the message was not altered while in transit between the sender and recipient. Document signation is done by creation of message digest of the original plaintext message using one of the cryptographically hashing algorithms using private key. Recepient decrypts the message digest using sender's public key and compares decrypted message digest he received from sender with the message digest receiver computed himself.

## V. Public Key Infrastructure

### *A. Certificates*

Digital *certificates* provide communicating parties with the assurance that they are communicating with people who truly are who they claim to be. Digital certificates are essentially endorsed copies of an individual's public key. This prevents malicious individuals from distributing false public keys on behalf of another party and then convincing third parties that they are communicating with someone else. Certificates are made by c*ertificate authorities* which are neutral organizations offer notarization services for digital certificates.

## VI. Conclusion

Creating and processing electronical documents from the security point of view bring up these security risks: author unambiguous, modification, unambiguous, data unambiguous and data lifetime. In homogenous data environment, information system can hold and fix these threats by its functions. Heterogenous environment offers data cryptography and signing by unique senders and recepient keys. Question to future is how to protect data lifetime and limit data access in heterogenous environment.

## References


[1] DEPARTMENT OF DEFENSE STANDARD, *Trusted Computer System Evaluation Criteria*, DoD 5200.28-STD, Dec 1985.

[2] Gary McGraw: *Software Security: Building Security In.* Addison Wesley Professional, January 2006, 448 pp. ISBN 1580533701.

[3] Torsten Lodderstedt, David A. Basin, Jürgen Doser: *SecureUML: A UML-Based Modeling Language for Model-Driven Security*, Springer-Verlag, 2002, 441 pp. ISBN:3-540-44254-5B

[4] Information technology - Security techniques, Secretariat: DIN, Germany, ISO/IEC FDIS 17799:2005, Library No. S225,7ll, p. 128, Feb 2005

[5] ACM and IEEE: Computing Curricula 2005. FoundStone, September 2005, 62 pp. ISBN: 1-59593-359-X.

# Some aspects about extending toposes

[1]Viliam SLODIČÁK, [2]Anita VERBOVÁ

Dept. Computers and Informatics, FEI TU of Košice, Slovak Republic

[1]viliam.slodicak@tuke.sk, [2]anita.verbova@tuke.sk

***Abstract*—Category theory provides possibilities to model many important features of computer science. We used symmetric monoidal closed category for the construction of model of linear type theory. Toposes as specific categories enable to model theories over types. In this contribution we prove that topos is symmetric monoidal closed category. This fact will allow us to use topos in the rôle of symmetric monoidal closed category - for construction of the models of the type theory.**

***Keywords*—linear logic, model, subobject, symmetric monoidal closed category, topos**

## I. Introduction

A program consists of data structures and algorithm. Data structures always have some type. Because types are complex structures, we use category theory for them. In our previous work [1] we defined the basic types as the startpoint of scientific problem solving by help of logically and mathematically founded programming of mathematical machines. In [2] we followed the development of our approach of regarding programming as logical reasoning in intuitionistic linear logic and we constructed linear type theory over linear Church's types involving linear calculus with equational axioms. The interpretation of linear type theory we concluded in symmetric monoidal closed category.

In our approach we also use toposes. Toposes are special categories - cartesian closed categories with some extra structures which produce an object of subobject for each object. This structure makes toposes more like the category of sets than cartesian closed categories generally are [3]. Toposes have proved attractive for the purpose of modelling computation. We constructed category of higher-order theories as topos [4] and found correspondence between tree automata and group as a mapping that assigns to each group an automaton which works over that groups [5]. In this contribution we will follow the ideas from [6]. We want to prove that topos is symmetric monoidal closed category, because in our next work we want to work with one category that has properties of symmetric monoidal closed category and also properties of topos.

## II. Basic notions

In this section we briefly introduce notions of symmetric monoidal closed categories and toposes.

### A. *Symmetric monoidal closed categories*

A monoidal category $\mathbf{C} = (\mathbf{C}, \otimes, I, a, l, r)$ consists of

- a category $\mathbf{C}$;
- a tensor functor $\otimes : \mathbf{C} \otimes \mathbf{C} \to \mathbf{C}$;
- natural isomorphisms $a$, $l$, $r$

$$
\begin{aligned}
a_{A,B,C} &: (A \otimes B) \otimes C \to A \otimes (B \otimes C) \\
l_A &: I \otimes A \to A \\
r_A &: A \otimes I \to A
\end{aligned}
$$

where $A$, $B$, $C$ are objects of the category $\mathbf{C}$. The first isomorphism expresses associativity of tensor functor, the two latter left and right neutral element of it. They have to satisfy the coherence axioms expressed by the following diagrams.

$$
\begin{tikzcd}
(A \otimes I) \otimes B \arrow[rr, "a"] \arrow[dr, "r \oplus id"'] & & A \otimes (I \otimes B) \arrow[dl, "id \oplus l"] \\
& A \otimes B &
\end{tikzcd}
$$

Fig. 1. Triangle - coherence axiom for isomorphisms $l$ and $r$ in monoidal category

To achieve commutativity of tensor product we add to monoidal category a natural isomorphism

$$c_{A,B} : A \otimes B \to B \otimes A$$

which satisfies coherence axiom in Fig 2 and Fig. 3.

$$
\begin{tikzcd}
& A \otimes B \arrow[dl, "c"'] \arrow[dr, "id"] & \\
B \otimes A \arrow[rr, "c"'] & & A \otimes B
\end{tikzcd}
$$

Fig. 2. Coherence axiom for $c$ isomorphism in monoidal category

$$
\begin{tikzcd}
I \otimes A \arrow[rr, "c"] \arrow[dr, "l"'] & & A \otimes I \arrow[dl, "r"] \\
& A &
\end{tikzcd}
$$

Fig. 3. Coherence axiom for $c$, $l$ and $r$ isomorphisms in monoidal category

A symmetric monoidal category $\mathbf{C}$ is closed, if for every object $A$ in $\mathbf{C}$ the functor $- \otimes A$ has a specified right adjoint - the hom functor $\mathrm{Hom}\,(A, -)$

$$- \otimes A \dashv \mathrm{Hom}\,(A, -)$$

that is, there exist natural transformations

$$\begin{aligned} \varepsilon_{A,B} &: \operatorname{Hom}(A,B) \otimes A \to B \\ \delta_{A,B} &: A \to \operatorname{Hom}(B, A \otimes B) \end{aligned}$$

which satisfy the triangle identities for an adjunction

$$1 = \varepsilon(\delta \otimes 1) : A \otimes B \to \operatorname{Hom}(B, A \otimes B) \otimes B \to A \otimes B$$

and

$$1 = \operatorname{Hom}(1, \varepsilon)\,\delta :$$

$$\operatorname{Hom}(A,B) \to \operatorname{Hom}(A, \operatorname{Hom}(A,B) \otimes A) \to \operatorname{Hom}(A,B)$$

In [2] we defined linear type theory and its interpretation. The interpretation of linear type theory we defined in symmetric monoidal closed category as a pair of functions $(i,j)$:

$$(i,j) : LinTT(B, \mathcal{F}, E) \to (\mathbf{C}, \otimes, I, a, l, c, \hom(-,-))$$

where $B$ is the set of basic types, $\mathcal{F}$ is the set of function symbols and $E$ is the set of axioms. Then $i : B \to \operatorname{Ob}(\mathbf{C})$ is *type interpretation function* and $j(f) : i(A) \to i(B)$ is a *function interpretation mapping* (for function symbol $f \in F$ of the form $f : A \to B$).
For example, if $\mathbf{C}$ is a category with finite products then the tensor product '$\otimes$' is given by cartesian category product, $I$ is a terminal object of the category $\mathbf{C}$ and natural isomorphisms are given by appropriate combinations of projection morphisms and pairing.

### B. Toposes

Toposes are special kind of categories defined by axioms saying roughly that certain constructions one can make with sets can be done in category theory [1]. A *topos* is a category $\mathcal{E}$ which satisfies the following properties:

1) $\mathcal{E}$ has a *terminal object* 1, and for every corner of morphisms $X \to Z \leftarrow Y$ in $\mathcal{E}$ there is a pullback:

$$\begin{tikzcd} P \arrow[r] \arrow[d] \arrow[dr, phantom, "\lrcorner", very near start] & Y \arrow[d] \\ X \arrow[r] & Z \end{tikzcd}$$

2) $\mathcal{E}$ has a *subobject classifier*: an object (traditionally denoted) $\Omega$ with a monomorphism $true : 1 \rightarrowtail \Omega$ such that for any monomorphism $m : M \rightarrowtail X$ in $\mathcal{E}$ there is a unique morphism $\chi_m : X \to \Omega$ such that the following diagram is a pullback:

$$\begin{tikzcd} M \arrow[r] \arrow[d, tail, "m"'] & 1 \arrow[d, "true"] \\ X \arrow[r, "\chi_m"'] & \Omega \end{tikzcd}$$

3) $\mathcal{E}$ has *power objects*: for each object $X$ in $\mathcal{E}$, an object $\Omega^X$ and a morphism $ev_X : X \times \Omega^X \to \Omega$ such that for any morphism $f : X \times Y \to \Omega$ in $\mathcal{E}$ there is a unique morphism $\lambda_X.f : Y \to \Omega^X$ such that the following diagram commutes:

$$\begin{tikzcd} X \times \Omega^X \arrow[r, "ev_X"] & \Omega \\ X \times Y \arrow[u, "1_X \times \lambda_X.f"] \arrow[ur, "f"'] & \end{tikzcd}$$

.

By 1) a topos has all finite limits. In particular, the product $X \times Y$ of two objects $X$ and $Y$ is the pullback of the corner of morphisms $X \to 1 \leftarrow Y$. For each object $X$, the unique morphism to the terminal object 1 we denote $!_X : X \to 1$. The morphism $\chi_m : X \to \Omega$ of 2) is called the *characteristic* (or *classifying*) *morphism* of the monomorphism $m : M \rightarrowtail X$. In 3), the morphism $ev_X : X \times \Omega^X \to \Omega$ is called *evaluation*, and $\lambda_X.f : Y \to \Omega^X$ is called the $X$*-transpose* of $f : X \times Y \to \Omega$. The object $\Omega^X$ is the exponential of $\Omega$ by $X$. For each object $X$ we have an exponential functor $\operatorname{Hom}(X,-) : \mathcal{E} \to \mathcal{E}$ which is right adjoint to the functor $- \times X$, so the topos is cartesian closed category [7].

## III. Topos is symmetric monoidal closed category

We are interested in toposes. Topos has a subobject classifier. It is a special object $\Omega$ of a category together with monomorphism $true$. A subobject classifier classifies subobjects of a given object according to which elements belong to the subobject. Because of this rôle, the subobject classifier is also referred to as the truth value object [8]. In fact the way in which the subobject classifier classifies subobjects of a given object, is by assigning the values true to elements belonging to the subobject in question, and false to elements not belonging to the subobject. This is the way the subobject classifier is widely used in the categorical description of logic.

In [6] we constructed model of predicate logic and linear logic as symmetric monoidal closed category. Here we formulate the proof, that topos is also symmetric monoidal closed category. This property is important when we want to construct model of linear logic as topos. So we formulate proposition: *A topos is symmetric monoidal closed category.*

*Proof.* Let's have the topos $\mathcal{E}$. We have to prove that $\mathcal{E}$ is symmetric monoidal closed category. So we have to prove that $\mathcal{E}$ satisfies the definition of symmetric monoidal closed category.

Category $\mathbf{C}$ in definition of symmetric monoidal closed category is cartesian closed category [9]. According the definition of topos is $\mathcal{E}$ also cartesian closed category.

Tensor product '$\otimes$' has a form $\otimes : \mathbf{C} \otimes \mathbf{C} \to \mathbf{C}$. Terminal object $I$ of category $\mathbf{C}$ is neutral element of tensor product.
Product '$\times$' in topos corresponds to tensor functor:

$$\times : \mathcal{E} \times \mathcal{E} \to \mathcal{E}$$

The product has neutral element 1 which is the terminal object of topos $\mathcal{E}$ [7].
Tensor product is associative [9]. The category $\mathbf{C}$ has the isomorphism expressing associativity of tensor product $a_{A,B,C}$. As the topos is cartesian closed category, from the properties

of product, we define associativity in topos as an isomorphism $assoc_{X,Y,Z}$

$$assoc_{X,Y,Z} : (X \times Y) \times Z \to X \times (Y \times Z)$$

Left and right neutral element of tensor product is given by isomorphisms $l_A$ and $r_A$. In topos $\mathcal{E}$ from the properties of product we define isomorphisms

$$\begin{aligned} left_X &: 1 \times X \to X \\ right_X &: X \times 1 \to X \end{aligned}$$

The following equations also hold:

$$\begin{aligned} left_X &= \pi_2 \left(1 \times X\right) \\ right_X &= \pi_1 \left(X \times 1\right) \end{aligned}$$

where $\pi_1 \left(-\right)$ and $\pi_2 \left(-\right)$ are the projection morphisms.

The commutativity in the category $\mathbf{C}$ is given by the isomorphism $c_{X,Y}$. For any pair of objects $X, Y \in \mathrm{Ob}\left(\mathcal{E}\right)$ we define an isomorphism $change_{X,Y}$ as follows: while topos is cartesian closed category, we define an object $X \times Y \in \mathrm{Ob}\left(\mathcal{E}\right)$ together with projections $\pi_1 \left(X \times Y\right) = X$ and $\pi_2 \left(X \times Y\right) = Y$. While $dom \left(\pi_1\right) = dom \left(\pi_2\right) = X \times Y$, we formulate $change_{X,Y}$ isomorphism as a product function

$$change_{X,Y} = \langle \pi_2, \pi_1 \rangle : X \times Y \to Y \times X$$

which expresses the commutativity in topos.

The category $\mathbf{C}$ is closed if for every object $A \in \mathrm{Ob}\left(\mathbf{C}\right)$ the functor $- \otimes A$ has a specified right adjoint, the hom-functor $\mathrm{Hom}\left(A, -\right)$

$$- \otimes A \dashv \mathrm{Hom}\left(A, -\right)$$

with the natural transformations $\varepsilon_{A,B}$ and $\delta_{A,B}$.
Similarly we define closedness for topos: for every object $X \in \mathrm{Ob}\left(\mathcal{E}\right)$ the functor $- \times X$ has the right adjoint $\mathrm{Hom}\left(X, -\right)$:

$$- \times X \dashv \mathrm{Hom}\left(X, -\right)$$

with the natural transformations

$$\begin{aligned} ev_X &: \Omega^X \times X \to \Omega \\ d_X &: X \to \mathrm{Hom}\left(\Omega, X \times \Omega\right) \end{aligned}$$

which also satisfy the triangle identities for an adjunction at Fig. 4 and Fig. 5:

$\mathrm{id} = ev_{X \times \Omega} \left(d \times \mathrm{id}_\Omega\right)$ :
$X \times \Omega \to \mathrm{Hom}\left(\Omega, X \times \Omega\right) \times \Omega \to X \times \Omega$
and

$\mathrm{id} = \mathrm{Hom}\left(\mathrm{id}_X, ev_X\right) d_{\mathrm{Hom}(X,\Omega)}$ :

$$\mathrm{Hom}\left(X, \Omega\right) \to \mathrm{Hom}\left(X, \mathrm{Hom}\left(X, \Omega\right) \times X\right) \to \mathrm{Hom}\left(X, \Omega\right)$$

We showed how the topos satisfies the definition of symmetric monoidal closed category. So we can conclude, that any topos $\mathcal{E}$ is symmetric monoidal closed category given as a structure

$$\left(\mathcal{E}, \times, 1, assoc, left, right, \mathrm{Hom}\left(-,-\right)\right)$$

$\square$

$$\begin{tikzcd}
& X \times \Omega \arrow[dl, "d \times \mathrm{id}_\Omega"'] \arrow[dr, "\mathrm{id}"] & \\
\mathrm{Hom}\left(\Omega, X \times \Omega\right) \times \Omega \arrow[rr, "ev_{X \times \Omega}"'] & & X \times \Omega
\end{tikzcd}$$

Fig. 4. Triangle identity for adjunction

$$\begin{tikzcd}
& \mathrm{Hom}\left(X, \Omega\right) \arrow[dl, "d_{\mathrm{Hom}(X,\Omega)}"'] \arrow[dd, "\mathrm{id}"] \\
\mathrm{Hom}\left(X, \mathrm{Hom}\left(X, \Omega\right) \times X\right) \arrow[dr, "\mathrm{Hom}(\mathrm{id}_X, ev_X)"'] & \\
& \mathrm{Hom}\left(X, \Omega\right)
\end{tikzcd}$$

Fig. 5. Triangle identity for adjunction

## IV. Conclusion

In our contribution we formulated basic aspects about topos theory needed for our approach of scientific problem solving. We presented a proof, that topos has properties of symmetric monoidal closed categories. Our next goal is to specify how to construct model of linear type theory in categorical terms by using that property of topos.

## Acknowledgment

This work was supported by VEGA Grant No.1/0175/08: Behavioral categorical models for complex program systems.

## References

[1] V. Novitzká, "Logical reasoning about programming of mathematical machines," in *Acta Electrotechnica et Informatica*. Košice, March 2005, pp. 50–55.
[2] V. Novitzká, D. Mihályi, and V. Slodičák, "Linear logical reasoning on programming," in *Acta Electrotechnica et Informatica*. Košice, 2006, pp. 34–39, iSSN 1335-8243.
[3] M. Barr and C. Wells, *Category Theory for Computing Science*. Prentice Hall International, 1990, pp. 1–432.
[4] V. Novitzká, V. Slodičák, and A. Verbová, "On modeling higher-order logic," in *Ivan Plander (ed.): Proceeding from International Scientific Conference Informatics 2007, $21^{st}$-$22^{nd}$ June*, 2007.
[5] V. Novitzká and V. Slodičák, "Some aspects about topos theory and tree automata," in *Romanian Journal of Information Science and Technology*, 2007, (accepted).
[6] V. Novitzká, D. Mihályi, and V. Slodičák, "Categorical models of logical systems in the mathematical theory of programming," in *MaCS'06 6th Joint Conference on Mathematics and Computer Science, Book of Abstracts*. University of Pcs, Hungary, 2006, pp. 13–14.
[7] M. Barr and C. Wells, *Toposes, Triples and Theories*. Springer-Verlag, 2002, pp. 1–326.
[8] P. Taylor, *Practical Foundations of Mathematics*. Cambridge University Press, 1999, pp. 1–572.
[9] S. Ambler, "First order linear logic in symmetric monoidal closed categories," Ph.D. dissertation, The University of Edinburgh, 1991.

# Turbo Codes and Channel Coding for Next Generation Mobile Systems

*Ján Šterba*

Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

jan.sterba@tuke.sk

***Abstract*—The astonishing performance of turbo codes have attracted many attention among researchers since their discovery in 1993. Not only did they approached closest to the Shannon capacity limit than any other code before, but they also did not actually arise from applying pre-existing and well explored theory pioneered by Claude Shannon, exhibiting new unexplored area for both theoretical and practical research. This paper is a short overview of turbo codes and their main components, and sheds some light on history and on the future of turbo codes.**

***Keywords*—turbo codes, channel coding, concatenation, interleaving, puncturing, decoding, SISO, SOVA**

## I. Introduction to Coding Theory

The basics principles of information theory has been founded by Claude Shannon. He showed in his article from 1946, that for every channel with a channel capacity *C* and transmission rates less than *C*, there exists codes able to transmit information over noisy channel with as few errors as we wish [1]. However, transmission capacity obtained usually in practical applications is much less than this capacity limit, what highlights the potential gains and leads to quest for techniques that could achieve this capacity limit in practice. In fact, Shannon also showed the principle how to achieve the capacity – with error correcting code, or channel code, which maps incoming data of length $k$ to corresponding codewords of length $n$, and then transmitting codewords over noisy channel. The capacity could be achieved by completely random code, but the drawback is that this performance could be achieved only when $k$ and $n$ tends to infinity [1].

A lot of codes have been developed for the last 45 years, but none of them could approach Shannon capacity limit. Taking into account that there must be infinitely many of such a codes, it was even remarked, that infinite numbers of good codes must exist, but none of those that were already developed were good. The introduction of turbo codes in 1993, which approached the Shannon limit within $0.7dB$, therefore created a great deal of interest, and led to sudden boom in both research and practical applications of turbo codes.

In this paper, the basic introduction to turbo codes is provided. The main components and primary building blocks of turbo codes are shortly explained, and the history behind the turbo codes and possible research trends are also mentioned.

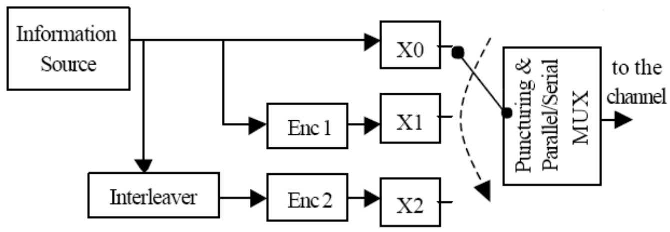



Fig. 1. Basic turbo code encoder block scheme

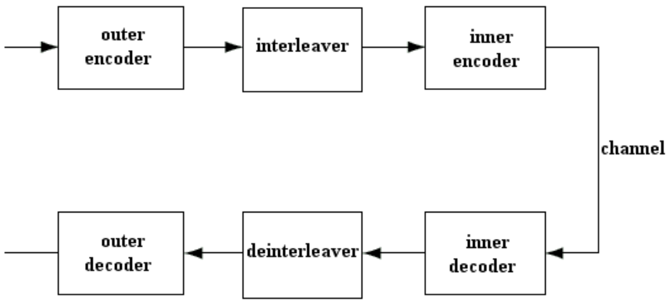



Fig. 2. Serial concatenation encoder (upper branch) and decoder (lower branch)

## II. Turbo Codes

To create the codes which could approach the Shannon capacity limit, the researchers have stood over awkward dilemma – to provide code that is infinitely long and random. This dilemma has been solved by proposal of Berrou, Glavieux, and Thitimajshima, who in their work, published in 1993 introduced the Parallel-Concatenated Convolutional Code (PCCC), also called as 'turbo code' [2]. Turbo codes solve the dilemma of infinite length and randomness by sophisticated combination of concatenation and interleaving. The basic scheme for turbo codes encoder is shown in Fig. 1.

## III. Concatenation

Whereas long codes cannot be designed easily because of the increased complexity with decoding of the long codes, it would be desirable to build a long, complex code, from shorter component codes, which can be decoded much more easily. *Concatenation* is a very straightforward way

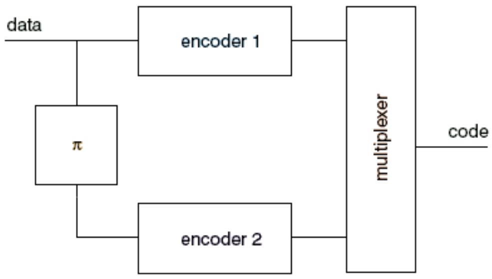



**Fig. 3.** Parallel concatenation

of achieving this. Basically, concatenation can be parallel or serial. In serial concatenation, the principle is to feed the output of one encoder, labeled as output encoder, to the input of another encoder, labeled as inner encoder, as is shown in Fig. 2.

In final, the resulting code seems to have properties of much longer code than codes in fact used, but such a code can be easily decoded, as two - or more if more complex code is used - simple decoders will be used. The interleaver between encoders is used to spread out the errors occurring in bursts, so they can be repaired more easily. The more detailed view on interleaver will be given in the next section.

In parallel concatenation, two encoders are working parallely, which means they are fed by the same input, as is shown in Fig. 3, where π denotes the block of interleaver.

## IV. Interleaving

An *interleaver* is a device that rearranges the ordering of symbol sequence in a deterministic manner. Associated with the interleaver is a *deinterleaver,* that applies the inverse permutation to restore the original sequence. Interleaver performs operation of permutation, thus the block of interleaver is often referenced to as $\pi$, and to inverse operation as $\pi^{-1}$.

The interleaver is very important part in the design of turbo code, as the size and map of interleaver has an significant effect on the code performance. Conventionally, interleaving is used to spread out the errors occurring in burst, but for turbo codes, the interleaver has more functions. Interleaving is used to feed the encoders with permutations so that the generated redundancy sequences can be assumed independent. Another key role of the interleaver is to shape the weight distribution of the code, which ultimately controls its performance. This is because the interleaver will decide which word of the second encoder will be concatenated with the current word of the first encoder, and hence what weight the complete codeword will have.

Issue worth considering in the design of interleaver is the termination of the trellis of both convolutional encoders. By properly designing the map of the interleaver, it is possible to force the two encoders to the all-zero state with only *m* bits (where *m* is the memory length of the convolutional encoder, assuming the same convolutional code is used in both encoders). The two main issues in the interleaver design are the interleaver size and the interleaver map.

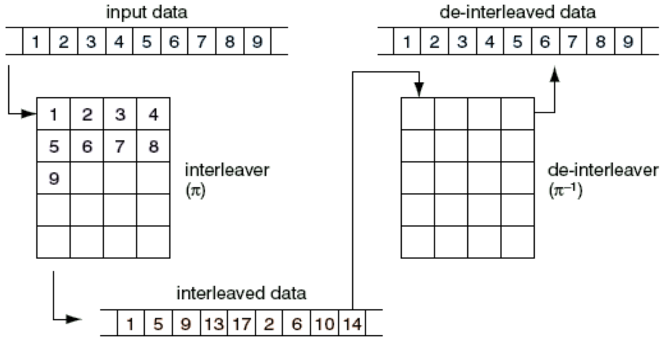



Fig. 4. Rectangular interleaver

The size of the interleaver plays an important rule in the trade off between performance and delay since both of them are directly proportional to the size.

The most simple interleaver is probably the block interleaver (also known as rectangular interleaver) is shown in Fig. 4. The data are being written into the array in rows, and once filled, the data are being read out by columns.

## V. Puncturing

The encoder, shown in Fig. 1, will have a code rate of 1/3, a relatively low rate. The code rate can be increased by *puncturing* the two parity streams. For example, one bit can be deleted from each of the parity streams in each turn, so that one parity bit remains for each data bit, and thus reducing the coding rate, resulting in a rate ½ code. Other rates are also possible by puncturing different proportions of the parity streams.

Puncturing can also be described as process of deleting some bits from the codeword according to a puncturing matrix. The puncturing matrix (*P*) consists of zeros and ones where the zero represents an *omitted* bit and the one represents an *emitted* bit. It is usually used to increase the rate of a given code. Puncturing can be applied to both block and convolutional codes. An example of the puncturing matrix to go from rate 1/2 to rate 2/3 is given by matrix

$$P=\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}.$$

This matrix implies that the first bit is always transmitted while every other second bit is omitted.

In the case of turbo codes, the same coder may serve for various coding rates by means of puncturing, allowing the same silicon chip to be used in different applications [3]. When the redundant information of a given encoder is not transmitted, the corresponding decoder input is set to zero. Of course, the decoder needs to know the current puncturing table. This function is performed by the DEMUX/INSERTION block in the turbo decoder (Fig. 7). The DEMUX will demultiplex the stream between the decoders and the INSERTION will insert an analog zero if the corresponding bit is omitted. When the code is punctured, the branch metric corresponding to the punctured bits need not be computed.

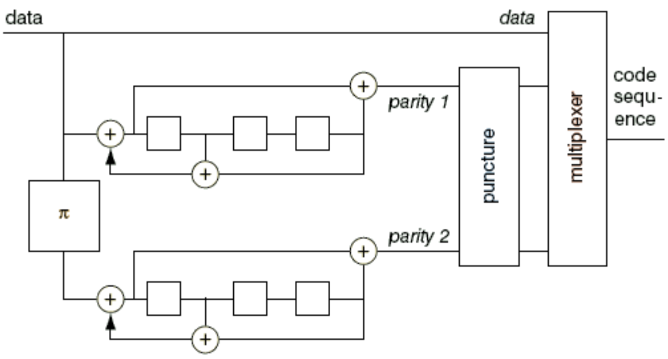



Fig. 5. Turbo code encoder

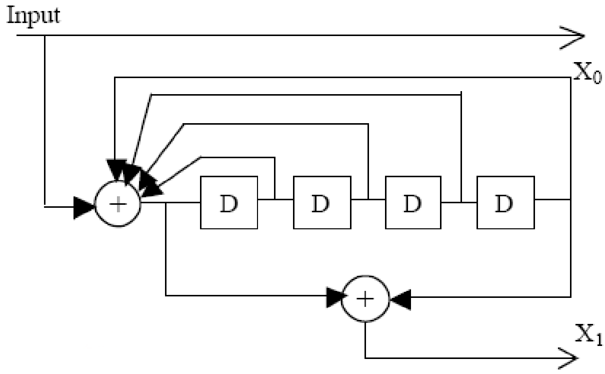



Fig. 6. Recursive Systematic Code

Determining the best puncturing pattern for turbo codes is still an open problem. Berrous suggests that systematic bits should not be punctured (Berrous puncturing) [3]. The paper demonstrated, by simulation, that it is always better to avoid puncturing systematic bits. However, in [10] a new puncturing is suggested, named UKL puncturing. It concludes that nonsystematic version of the designed punctured code is advantageous at high values of $E_b / N_o$ which are related to low BERs. This result was tested over AWGN and fully interleaved flat Rayleigh fading channels.

Puncturing is a trade off between rate and performance, but fortunately, punctured codes have only $0.1dB - 0.2dB$ lower performance than the original code, as was reported for the convolutional codes [11]. Convolutional codes and their punctured alternatives together with their difference in performance are tabulated in [8]. It is suggested that using unequal puncturing (puncturing one encoder outputs different than the other) improves the performance. In [9] Caire shows that to achieve best performance, the puncturing and the interleaver should be designed jointly. In [10], Caire suggested to define the puncturing pattern on the interleaver map. Moreover, adaptive puncturing can be used depending on the channel noise, allowing adaptability to the channel conditions.

## VI. Encoder

The basic scheme for turbo encoder is shown in Fig. 1, respectively in Fig. 5 where general encoder blocks have been replaced with convolutional encoders.

In the next section we will describe the block diagram, which is shown in Fig. 5, containing three branches. In the first branch, the information bits are fed directly to the input of the multiplexer - this code is therefore *systematic* code. A systematic code is such a code, in which the input data is embedded in the encoded output. The codes used in concatenated coding are usually systematic. The second branch of the Fig. 5 contains the first encoder, typically the convolution one. The same input data which are fed to the second branch are also fed to the third branch, where the information bits are first scrambled in the interleaver before entering the second encoder. The turbo encoders typically convolution coder.

Viewed from the output of the multiplexer, the input bits are then followed by the parity check bits from the first encoder, and then the by parity bits from the second encoder.

To deal with the delay of the interleaver in the third branch, the delay line is in general placed before the first encoder to keep both branches work simultaneously.

The turbo encoder block diagram in Fig. 1 and Fig. 5 shows only two branches. In general, we can use turbo encoders with more than two branches. The convolutional code at every branch is called the *constituent code*. The *constituent code*s can be generated by similar or different encoders. In this paper, we will be assuming the usual configuration with two branches having the same *constituent code*.

Puncturing can be also used in encoder, as shown in Fig. 1, to increase the rate of the convolutional code beyond the speed resulting from the basic structure of the encoder. Separate sections are devoted for interleaving and puncturing, where they are being analyzed more closely.

The convolution codes can be labeled as *recursive* or *non-recursive*, and *systematic* or *non-systematic*.

In the case of turbo codes Recursive Systematic Convolutional (RSC) codes are proved to perform better than the non-recursive ones [2], [11]. An RSC encoder, shown in Fig. 6, can be obtained from an non-systematic encoder by setting one of the outputs equal to the input and obtained from non-recursive encoder by using a feedback.

## VII. Iterative decoding

The turbo decodes use iterative decoding for data decoding, which is often referenced to as 'turbo principle'. The block diagram of turbo decoder is shown in Fig. 7. The iteration stage is shown with doted lines, and the initialization stage is shown with solid line. Only one loop is performed at a time.

The term turbo principle, after which the turbo codes has been named, and which in fact refers to the iterative decoding, was taken from the turbo engine principle, in which part of the engine output is fed back to the input of the system to increase the overall performance. The same approach is used in turbo code decoder – the first decoder will decode the sequence and then passes the hard decision together with a reliability estimate of this decision to the next decoder. Now, the second decoder will have additional information for the decoding - the result of the first iteration and the original sequence. The interleaver between the decoders is responsible for making the two decisions of the decoders uncorrelated. The exact procedure in what information to pass to the next decoder or next iteration stage is a subject of research.

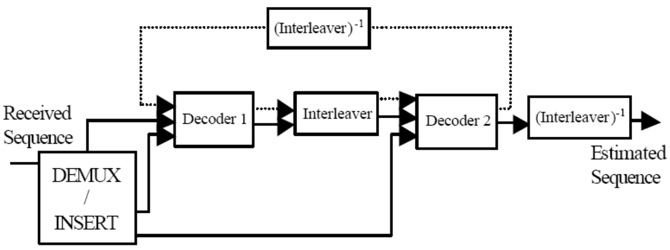



Fig. 7. The block scheme of turbo decoder

The turbo code decoders employs *soft-in, soft-out* decoding (SISO). The utilization of SISO decoding greatly improves the overall performance of decoders. In addition to the hard decision, the SISO demodulators are using additional soft information, which is passed to the decoder and which measures the reliability of the decision.

For example, when received signal is passed to the comparator, at the output of the comparator in addition to the output level of the signal ‚0' or ‚1', the signal measuring the level of reliability is also produced. If the input signal is close to the decision level of comparator, this signal will have low value, and when the input signal is far from the decision level, this signal will have higher value.

To decode turbo coded data stream, the algorithms typically used to decode other codes have been modified to decode turbo codes. One of such a mostly used algorithm used to decode turbo codes is a Soft Output Viterbi Algorithm (SOVA). Viterbi algorithm is a decoding method that was created for decoding of convolutional code. It is standardly used decoding method for its optimal algorithm, allowing to minimize the probability of sequence error. To utilize Viterbi algorithm for turbo codes, the algorithm has been modified with soft outputs to include reliability option to decoding decision. This modified version of Viterbi algorithm is introduced in [6], [7]. SOVA has only twice the complexity of Viterbi algorithm.

## VIII. Performance of Turbo Codes

The turbo codes can produce remarkable results when comparing to other channel coding methods, especially in areas of relative low $E_b / N_o$. For example, the turbo codes can achieve $BER{=}10^{-5}$ over Rayleigh channel at only $E_b / N_o = 4.3dB$ [7]. Comparing with classical convolutional coder, the turbo codes brings gain of $2.3dB$. Over AWGN channel, the turbo codes achieve $BER{=}10^{-5}$ over at only $E_b / N_O{=}0.7dB$, which is very close to Shannon limit [2], and it is closer to the limit bound than most of the researchers had even imagined possible. The performance of turbo codes is highly dependent on number of iterations. increase significantly with increase of decoder iterations. At higher number of iterations, the decoder show significant improvement, but higher number of iterations also means more processing time and possibly delay inserted into communication channel. Also, the turbo codes use an interleaver both in the transmitter and receiver. Since the data must be stored in the memory of interleaver in order to be processed, this means latency inserted to the communication system. As different applications are differently delay dependent, the optimization of system design is necessary to provide satisfying latency to the system.

## IX. Conclusion

The turbo codes has attracted a lot of attention from their introduction in 1993, as they can provide for the way the communication researchers has been trying to solve for decades – to attack the Shannon limit bound for capacity of communication channel with practical applications. The first appearance of turbo codes led to renaissance in the field of coding and inspired other areas of research to look for 'turbo principle' and try to use it to improve older applications and processes.

The practical applications of turbo codes has emerged immediately after their introduction and promptly took their place in the real world systems. Still, to achieve excellent performance of turbo codes, deep understanding of system and optimization of parameters are required, and further investigation needs to take place to inspect the effect of individual ingredients on the overall performance of system.

## X. Acknowlegments

This work has been funded by VEGA 1/4088/07 "Rekonfigurovateľné platformy pre širokopásmové bezdrôtové telekomunikačné siete".

## References

[1] C. E. Shannon, "A mathematical theory of communication", Bell Syst. Tech. J., 1948, pp. 379–423 and 623–656.
[2] C. Berrou, A. Glavieux, P. Thitimajhima, "Near Shannon limit error correcting coding and decoding : turbo-codes," Proc. Of ICC '93, Geneva, May 1993, pp. 1064-1070.
[3] C. Berrou, A. Glavieux, "Near optimum error correcting coding and decoding: Turbo-Codes", IEEE, Transactions on Communications, 1996, vol. 44, pp. 1261-1271.
[4] P. Jung, J. Plechinger, "Performance of rate compatible punctured Turbo-codes for mobile radio applications," Electronics Lettes, 1997, vol. 33, No.25, pp. 2102-2103.
[5] P. Sweeney, " Error Control Coding: an Introduction", UK: Prentice Hall International Ltd. 1991.
[6] J. Hagenauer, P. Hoeher, "A Viterbi algorithm with soft-decision outputs and its applications", Proceedings of GLOBECOM '89, Texas, 1989, pp. 47.11-47.17.
[7] J. Hagenauer, L. Papke, "Decoding turbo codes with the soft-output Viterbi algorithm (SOVA)", International Symposium On Information Theory, Norway, 1994, pp.164.
[8] G. Clark, J. Cain, "Error-Correction Coding for Digital Communications", Plenum Publishing Corporation, New York, 1988.
[9] G. Caire, E Biglieri, "Parallel concatenated codes with unequal error protection," IEEE Transactions on Communications, vol. 46, No. 5, 1998, pp. 565-567.
[10] G. Caire, G. Lencher. ,"Turbo codes with unequal error protection", Electronic Letters, Mar. 1996, vol. 32, No. 7, pp. 629-631.
[11] S. Benedetto, G. Montors, "Design of parallel concatenated convolutional codes," IEEE, Transactions on Communications, vol. 44, No. 5, 1996, pp. 591-600.
[12] A. H. Mugaibel, M. A. Kousa, "Understanding Turbo Codes", King Fahd University of Petroleum and Minerals, Saudi Arabia
[13] A. Burr, "Turbo-codes: the ultimate error control codes?", Electronics & Communication Engineering Journal, 2001

# Interaction categories and $\pi$-calculus as a description of asynchronous processes

[1]*Anita VERBOVÁ*, [2]*Viliam SLODIČÁK*

[1]Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic
[2]Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

[1]anita.verbova@tuke.sk, [2]viliam.slodicak@tuke.sk

***Abstract*—In our research we are intrested in comparing existing methods for describing the behaviour of the model of asynchronous processes. We represent two approaches and write out interaction categories which involves more complex data structures than $pi$-calculus.**

***Keywords*—concurrent processes, interaction categories, linear logic, $pi$-calculus.**

## I. Introduction

We can model asynchronous processes in many different ways. In this paper we describe them first by $\pi$-calculus then on the contrary by interaction categories.

The $\pi$-calculus [1] provides a simple yet powerful framework for specifying communication systems with evolving communication structures. There exist many formalisms for describing processes. $\pi$-calculus is one of these formalisms for describing parallel processes, which communicate by transition of messages. It is not the best choice for describing abstract data types and for describing states with complex data structures.

## II. $\pi$-calculus

Joachim Parrow in [2] defined $\pi$-calculus as a process algebra. $\pi$-calculus is a mathematical model of processes, which links are changing dynamically during their interaction. *Communication link* is a medium, which names are sent through. Basic computational step is *the transfer* of communication link between two processes. After that receiver can use this communication link for further interaction with other participants. Its expressiveness derives mainly from the possibility of passing communication channels, restricting the scope of channels and scope extrusion. It is suitable for modelling systems, where resources change in time.

At the first sight it seems that $\pi$-calculus is only a special form of process algebra with data transfer, where data could be links. In this case $\pi$-calculus would be very weak, as it does not define any data types or functions. The reason, that $\pi$-calculus is despite of this considered to be more expressive, is that it enables *migration* of communication links and data through another communication links. This feature expresses mobility of processes or in other words this means modification of process structure. The name passing feature alone (input prefix) in $\pi$-calculus would yield infinite branching transition systems.

## III. Interaction categories

Fragment of linear logic can be considered as process calculus, where linear terms represent processes and proofs represent execution of processes. *Proof nets* [3] are deductive systems of linear logic. In this case interaction can be interpreted as computational behaviour of proof nets. The connection between proof nets and processes is interpreted by Abramsky in [4]. Traced monoidal categories describe the semantics of asynchronous communication of concurrent processes.

Roughly speaking we can describe synchronous communication as parallel sending and receiving messages, which requires so-called "handshake" between sender and receiver. Communication is asynchronous if messages pass through a communication medium with possible delay. Therefore sender does not have to wait for the presence of receiver.

There exist many semantical models of networks of communicating processes, but only a few of unifying principles for all these models. We use category theory to separate some common and useful properties of models of asynchronous communication. These properties could form the basis of classification and connection of some existing models.

We begin with the observation of a concrete model of asynchronous communication, which was interpreted in [5]. In this model, which is based upon labeled transition systems, are asynchronous channels described explicitly as unbounded capacity containers. Process is asynchronous if its input and output behaves as if it passed through such a container. These asynchronous processes are essential morphisms in the traced monoidal category **Buf**. We take this category as a prototype and by abstraction from its properties we derive categorical model of asynchronous processes.

### A. Feedback operation

We define the type of the transition system $X \rightarrow_B Y$ as the set $InX + OutY + B + \{\tau\}$. Labeled transition system $\mathcal{S}$ of type $X \rightarrow_B Y$ is called *labeled transition system with input and output* or *agent*. If $B$ is an empty set, then we omit the subscript in $X \rightarrow_B Y$.

*Container* $\mathcal{B}_X$ has states from multiset from the commutative monoid $X^{**}$, initial state is $\emptyset$, and transitions are $w \stackrel{in\ x}{\rightarrow} wx$ and $xw \stackrel{out\ x}{\rightarrow} w$ for all $w \in X^{**}$ a $x \in X$.

In asynchronous communication model we assume, that messages pass through a communication medium. Sometimes we suppose, that this medium is a FIFO (queue), sometimes as

in the case of asynchronous $\pi$-calculus, messages are received in arbitrary order (container). The basic idea is that we use infinite containers for modelling asynchronous behaviour of a process. Asynchronous agent is such agent, which output and/or input behaves as if it passed through container $\mathcal{B}$.
Agent $\mathcal{S} : X \to Y$ is

- *output-containered*, if $\mathcal{S} \approx \mathcal{S}; \mathcal{B}_Y$
- *input-containered*, if $\mathcal{S} \approx \mathcal{B}_X; \mathcal{S}$
- *containered* or *asynchronous*, if $\mathcal{S} \approx \mathcal{B}_X; \mathcal{S}; \mathcal{B}_Y$.

Asynchronous agents are closed under composition. Containers $\mathcal{B}_X : X \to X$ represent identity morphisms for this composition. Therefore classes of weak bisimulations of asynchronous agents form morphisms of the category $\mathbf{Buf}$.
Category $\mathbf{Buf}$ is defined as the following structure:

- Objects are sets of states.
- Morphisms are classes of weak bisimulations of containered agents $\mathcal{S} : X \to Y$.
- Identity morphism on $X$ is given by container $\mathcal{B}_X$.
- Composition is given by composition on agents.

There also exists a natural operation of feedback on agents, i.e. connecting output of agent back to its input. This operation enables us to construct networks with loops.
Agent $\mathcal{S} : X + Z \to Y + Z$ can be illustrated as a box with input and output links as on the following figure:

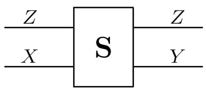



Fig. 1. Agent with inputs and outputs

We define next agent $Tr_Z\mathcal{S} : X \to Y$ by connecting the output link of type $Z$ from $\mathcal{S}$ to the appropriate input link as on the Fig. 2:

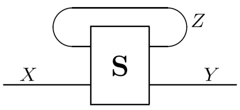



Fig. 2. Agent with feedback

Formally, agent $Tr_Z\mathcal{S}$ is defined as follows: its states and the initial state are the same as those of $\mathcal{S}$. Transitions are given by two rules (1).

$$\frac{s \xrightarrow{\alpha}_{\mathcal{S}} s' \quad \alpha \neq in\ z, out\ z}{s \xrightarrow{\alpha}_{Tr_Z\mathcal{S}} s'} \qquad \frac{s \xrightarrow{out\ z}_{\mathcal{S}} s' \quad s' \xrightarrow{in\ z}_{\mathcal{S}} s''}{s \xrightarrow{\tau}_{Tr_Z\mathcal{S}} s''} \tag{1}$$

Asynchronous agents are closed under the operation of feedback. This operation defines traced monoidal structure on the category $\mathbf{Buf}$.

### B. Traced monoidal categories

Traced monoidal categories are extension of symmetric monoidal categories with the notion of *loop*. In computer science traced monoidal categories are used for modelling feedback in proces algebra [6] and cyclic graphs of data flow [7]. Notation $\mathcal{C}(A, B)$ denotes homset of morphisms from $A$ to $B$ in the category $\mathcal{C}$. In symmetric monoidal category we denote the symmetric morphism by $c_{XY} : X \otimes Y \to Y \otimes X$. Unit element of tensor is denoted by $I$.

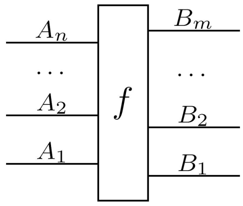



Fig. 3. Morphism with many inputs and outputs

We present morphism $f : A_1 \otimes \ldots \otimes A_n \to B_1 \otimes \ldots \otimes B_m$ as a box with links in the following way:
Fig. 3 illustrates the process $f$ with input links $A_1, \ldots, A_n$ and output links $B_1, \ldots, B_m$. Operations of composition, tensor and trace together with identity and symmetry morphism are written as on Fig. 4.

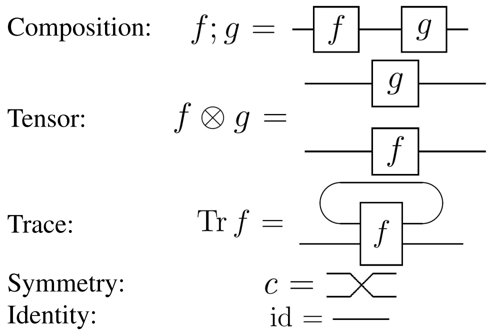



Fig. 4. Operations on processes

*Traced monoidal category* $(\mathcal{C}, \otimes, Tr)$ is symmetric monoidal category $(\mathcal{C}, \otimes)$ together with the class of operations $Tr_X : \mathcal{C}(A \otimes X, B \otimes X) \to \mathcal{C}(A, B)$, which satisfy the following four axioms:

(i) Naturality. $Tr_X(g \otimes id_X; f; h \otimes id_X) = g; Tr_X f; h$
(ii) Strength. $Tr_X(g \otimes f) = g \otimes Tr_X f$
(iii) Symmetry sliding. $Tr_Y(Tr_X(f; id_B \otimes c_{XY})) = Tr_X(Tr_Y(id_A \otimes c_{XY}; f))$
(iv) Yanking. $Tr_X(c_{XX}) = id_X$

On Fig. 5 we show four axioms of traced monoidal categories.

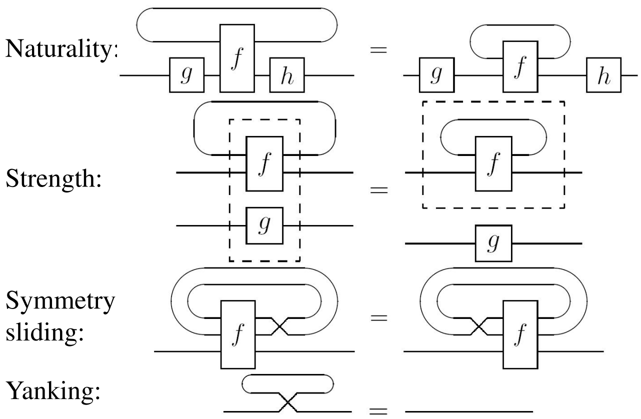



Fig. 5. Axioms of traced monoidal categories

### C. Classification of traced monoidal categories

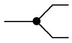
Monoidal category with diagonal is a symmetric monoidal category together with two morphisms:

$$\Delta_A : A \to A \otimes A, \ \Diamond_A : A \to I$$

such that $(A, \Delta_A, \lozenge_A)$ is a symmetric comonoid for each $A$, and these morphisms are compatible with symmetric monoidal structure in the following sense:

$$\lozenge_I = \mathrm{id}_I : I \longrightarrow I \qquad \begin{tikzcd} & I \otimes I \arrow[dd, "\cong"] \\ A \otimes B \arrow[ur, "\lozenge_A \otimes \lozenge_B"] \arrow[dr, "\lozenge_{A \otimes B}"'] & \\ & I \end{tikzcd} \qquad \begin{tikzcd} & A \otimes A \otimes B \otimes B \arrow[dd, "id_A \otimes c \otimes id_B"] \\ A \otimes B \arrow[ur, "\Delta_A \otimes \Delta_B"] \arrow[dr, "\Delta_{A \otimes B}"'] & \\ & A \otimes B \otimes A \otimes B \end{tikzcd}$$

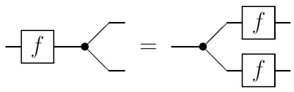
The first diagram presents that by applying weak terminal morphism $\lozenge$ to unit object of tensor product $I$, then we get the identity on this object. The second diagram indicates the closure of weak terminal morphism according to tensor. The third diagram illustrates the closure of diagonal morphism according to tensor.
Morphisms $\Delta_A$ and $\lozenge_A$ are called *diagonal* and *weak terminal morphism* respectively. Their graphical representation is:

 and 

In monoidal category with diagonals the morphism $f : A \to B$ is called *copyable* or *deterministic*, if the following diagram commutes

$$\begin{tikzcd} A \arrow[r, "\Delta"] \arrow[d, "f"'] & A \otimes A \arrow[d, "f \otimes f"] \\ B \arrow[r, "\Delta"] & B \otimes B \end{tikzcd}$$

graphically,



Morphism $f : A \to B$ is called *discardable* or *total*, if the following diagram commutes

$$\begin{tikzcd} A \arrow[dd, "f"'] \arrow[dr, "\lozenge"] & \\ & I \\ B \arrow[ur, "\lozenge"'] & \end{tikzcd}$$

graphically,

In the context of communicating processes the diagonal morphism means duplicating messages on the channel. Such deterministic morphism represents process, which two independent copies with the same input generate the same output. This is exactly the case of deterministic process. Weak terminal morphism accepts, but ignores all inputs. Total morphism represents process, which is defined for all inputs, i.e. which never refuse any input. For such reason it is called total morphism.

## IV. Conclusion

In this contribution we presented two ways of describing asynchronous processes. We have extended symmetric monoidal categories with feedback and formed traced monoidal categories. We showed on a specific category **Buf** the properties of asynchronous categories, which are described by traced monoidal categories. Our future work will contain extension of interaction categories for more complex structures.


## Acknowledgment

This work was supported by VEGA Grant No.1/0175/08: Behavioral categorical models for complex program systems.


## References


[1] R. Milner, J. Parrow, and D. Walker, “A calculus of mobile processes (parts i and ii),” *Information and Computation*, vol. 100, pp. 1–77, 1992.
[2] J. Parrow, *An Introduction to the $\pi$-Calculus, in The Handbook of Process Algebra*. Amsterdam: Elsevier, 2001, p. 479.
[3] A. Verbová, V. Novitzká, and V. Slodičák, “From linear sequent calculus to proof nets,” in *Informatics 2007, Proceedings of the Ninth International Conference on Informatics*. Slovak Society for Applied Cybernetics and Informatics Bratislava, 2007, pp. 100–107.
[4] S. Abramsky, “Proofs as processes,” in *MFPS ’92: Selected papers of the conference on Meeting on the mathematical foundations of programming semantics, part I : linear logic*. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 1994, pp. 5–9.
[5] P. Selinger, “First-order axioms for asynchrony,” in *International Conference on Concurrency Theory*, 1997, pp. 376–390. [Online]. Available: citeseer.ist.psu.edu/selinger97firstorder.html
[6] S. Abramsky, “Retracting some paths in process algebra.” in *CONCUR*, 1996, pp. 1–17.
[7] M. Hasegawa, *Models of Sharing Graphs: A Categorical Semantics of Let and Letrec*. London, UK: Springer-Verlag, 1999.

# Ontologies and Web Services as a Building Blocs of the Semantic Web (March 2008)

[1]Jozef VRANA, Kristina MACHOVÁ, [2]Martin DŽBOR

[1]Dept. of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic
[2]Knowledge Media Institute, The Open University, United Kingdom

[1]jozef.vrana@tuke.sk, kristina.machova@tuke.sk, [2]m.dzbor@open.ac.uk

**Abstract— Web content growing very fast and it started to be a rather complicated to find relevant information. Semantic web is an initiative to make web more usable, acceptable and comprehensible for users. In order to make Semantic Web very much practical, is necessarily to provide new functionalities and additional information about web content. Those may be interesting not only for users but also for web agents. So it suppose to be provided in form, processable by both. The key technologies for the Semantic Web support are Ontologies and Web Services. The ontologies represent resource of meta information about objects, properties of these objects and relations among them. On the other hand, web services are deliver further functionalities upon ontologies and web content to the user. Together they are the most significant technologies for support of semantic web.**

***Keywords*—semantic web, ontology, web service.**

## I. Introduction

Size of the information on the web is increasing very fast. In 2000 was available about 7 million of unique web pages. Contemporary estimations say about 30 billion pages and 1 billion of users. Biggest searching engine have indexed only small number of these documents so far. Most of the users get used to on fact, that everything is available via Google. If it is not true, how we can get an access to this information? The answer would be: We can not! What is the point of having a huge amount of data if you can not use them?

Poor content aggregation is another important issue of the web. Very important assumption for data aggregation is to have it in form comprehensive for machines. Apparently, web pages are readable by humans. But web agent is struggling with this type of content because there is no data about meaning. From computer point of view is web page only a set of words but words could have several meanings and computers are not able to distinguish among them.

Described problems represent only small part of amount of issues of current web. Obviously, there are more approaches how to solve this unpleasing situation. In this paper I would like to clarify one of such ideas which is Semantic web.

Tim Bernes-Lee has a two-part vision for the future of the Web. The first part is to make the Web a more collaborative medium. The second part is to make the Web understandable, and thus processable, by machines. It basically means captured relationships between information items. This could be done by ontologies. Ontology is formal description of objects from real word. It captures objects, properties and relationships from some particular domain. On base of such a formal description, a new data can be inferred from existing data by following logical rules. These data would be application-independent and part of the larger information ecosystem.

## II. Ontologies

### A. Definition

One of the possible definitions of ontology: Ontology is a formal, unambiguous delimitation of shared terms. It means that it provides shared dictionary which describes the chosen domain, the types of objects and terms, their attributes and relations among them. Categories, their attributes and relations among them create the ground of ontology. It is necessary to define them; it means to note the category, determine its attributes as well as its relation to other categories. After the definition of categories it is possible to assign them some instances, which will represent the objects of these categories. For the reason that the relations among objects are known, it is possible to derive some new facts.[1]

### B. Ontology and Semantic Web

One of the fundamental problems in semantic web is that one and the same term can be labelled by different codes. Agent in such case must have a possibility to differentiate, what meaning has information, for given data source it can get in touch with.

It is just ontology that was designed for this problem solving. The term ontology represents the summary of formally defined relations among the objects. In ontology so called derivational rules are used. They serve for derivation of other relations among entities.

Ontology increases the functionality of web. In this way that it allows to specify the web retrieval. With help of ontology, program should be able to display, only relevant pages for current query. This is the most direct utilization of ontology.

There exist also more sophisticated accesses, for example:

- the using of association of knowledge structure
- derivation of ontology rules

Such a use could lead to the complete automatic retrieval of all the relevant references.

### C. Ontological issues

As long as web contend rise very dramatically, the amount of ontologies are increasing too. There are lots of institutions, particularly in academic sector, which are interested in ontology engineering. It means up there is available huge set of documents which describes our world by formal manner. Obviously, this data can be used on various purposes but first the problem is to find such an ontology, which perfectly matches with domain of yours particular resource. Better to say: How to find perfect ontology for web page?

One way how to do it is via some semantic web gateway for instance Watson. Watson is a tool developed by KMI. Ontologies can be found easily by keywords but still it is rather tricky to choose relevant ontological model for user.

Another important point is that most likely ontology and web resource were created in different conditions so it could cover different parts of particular domain.

Discussed issues haven't been solved so far, but for further development of these technologies, it is necessary to do that.

## III. Web Services

### A. Definition

Above content of the web pages and ontologies stand web services. It is software services that are defined accessed using web protocols. The main goal of web services is to provide additional functionalities by using ontologies and web pages.

Web services are software applications that can be discovered, described, and accessed based on XML and standard Web protocols over intranets, extranets and the Internet. [1].

From the definition, web services consist form several layers:

- web service are software applications available on the Web that perform specific functions,
- discover: integrations UDDI (Universal Description, Discovery and Integration) and ebXML registers allow applications to discover information about web services,
- describe: Web service is described via WSDL (Web Service Description Language)
- access: SOAP (Simple Object Access Protocol), is the XML-based message protocol for communicating with Web services,
- communication layer: consist from under layered protocol for instance HTML, SMTP.

### B. How does it fit to the Semantic Web

The first way that Web services fit into the Semantic Web is by furthering the adoption of XML, or more smart data. An adoption of XML by Web services is fairly important point as well as enabling Web services to interact with other Web services. Advanced Web services will require Semantic Web technologies for such interactions to be automated.

This technology provides interoperability solutions, making application integration and transacting business easier. Because it is now obvious that monolithic, proprietary solutions are barriers to interoperability, industry has embraced open standards. The widespread support and adoption of Web services, in addition to cost-saving advantages of Web services technology, make the technologies involved very important to understand.[1]

## IV. Applications

The Semantic Web is not sole new thing but practical aspects of this technology haven't been recognized with many projects so far. The users are not really familiar with technologies like ontologies and web services. It is because these are rather complicated and not very comprehensible for them.

However there are still couple of interesting tools which take advantages of those technologies. Magpie is one of the very clever tools based on mentioned technologies. It is fairly unique product in several ways. It is basically first attempt to create semantically enriched browser. What actually Magpie does is that automatically identifies entities on the web page which may have concept in ontology. For those entities are available services on demand via contextual menu. This approach could be occasionally helpful for users. The strength of the Magpie is in idea which allows users to take an advantage of semantically enriched content. On the other hand, Magpie is able to work with very well defined problems. It is because it works with one ontology in same time. This constraint is very strong because it is hard to find perfect ontology for each problem. Thus rather than being isolated may be we should bring more information in there and using set of ontology instead. [2]

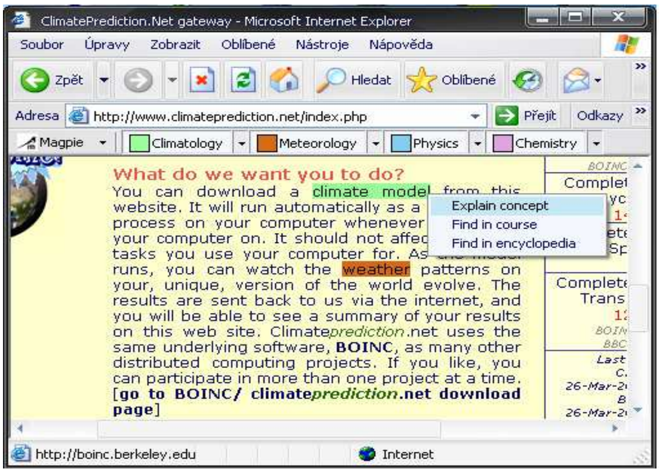



Fig. 1. Window IE with running Magpie

Another interesting tool is PowerAqua. PowerAqua is a multi-ontology-based Question Answering (QA) system, which takes as input queries expressed in natural language and is able to return answers drawn from relevant distributed resources on the Semantic Web. In contrast with any other existing natural language front end, PowerAqua is not restricted to a single ontology and therefore provides the first comprehensive attempt at supporting open domain QA on the Semantic Web.[3]

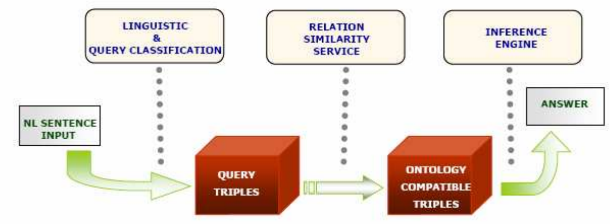



Fig. 2 The AquaLog Data Model

## V. Adoption of the Semantic Web

The semantic web hasn't been widely adopted by the users so far. There would by probably several reasons why it is like that:

- Simplicity: Only small number of users is able to create ontology without any assistance or past experience with ontological engineering. It is because the knowledge from logic and artificial intelligence are required. Obviously, users most often do not have such an experience.
- High threshold value of users: Currently, we have working alternative to Semantic Web. Well, the problem is how to convince users to switch from well-known technology to new one. The important point is: there is now real motivation for users to do that yet.
- Slow process of standardization: So far, it has been standardized only low level technologies as XML, URI and RDF. Those are obviously very important ones and create base for higher level technologies but from users point of view they are useless.
- Semantic web applications: As long as Web exists there were successful applications which make it popular because brought new efficient capabilities. For example Google or Netscape are this kind of tools. But there is no such a tool for Semantic web and this is probably most significant reason why is current state so unpleasing.

## VI. Conclusion

In this paper were brought a brief overview of current state of Semantic Web and related technologies. The main motivation for research in this scope is: How can Semantic Web make task I'm doing easier, faster, simpler … After that will be this technology interesting not only for academic sector but also for private companies. And this is very important for widely adoption of these technologies.

The question is: how to achieve it? May be we should focus on using existing web contend as an asset instead of pushing new technology. We need to improve the overall use experience

## Acknowledgment

The work presented in this paper was supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic within the 1/4074/07 project "Methods for annotation, search, creation, and accessing knowledge employing metadata for semantic description of knowledge"..

## References

[1] M. C. Daconta, Leo J. Obrst, Kevin T. Smith, "The Semantic Web," in Indiana: Indiana Polis, 2003, pp. 57–237.
[2] J. Vrana, M. Dzbor, K. Machova: Software Means for the Support of Semantic Web, The 6th Slovakian - International Symposium on Applied Machine Intelligence 2008, Slovakia.
[3] V. Lopez, E. Motta, V. Uren: PowerAqua: Fishing the Semantic Web. Proceedings of the European Semantic Web Conference, ESWC 2006, Montenegro.
[4] M. Dzbor, E. Motta: Semantic Web Technology to Support Learning about the Semantic Web. In 13th Intl. Conf. on Artificial Intelligence in Education (AIED). July 2007, California, US.

# Author`s Index

**8th Scientific Conference of Young Researchers**
**of Faculty of Electrical Engineering and Informatics**
**Technical University of Košice**

Proceeding from Conference

Published: Faculty of Electrical Engineering and Informatics
Technical University of Košice
The number of CD Proceedings: 70 pieces

**ISBN 978-80-553-0036-8**