Acta Universitatis Sapientiae

Electrical and Mechanical Engineering

Volume 2, 2010

Sapientia Hungarian University of Transylvania
Scientia Publishing House

Contents

Telecommunications

Signal Processing

Mechatronics & Industry Applications

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 5-15

Comparative Analysis of Model Predictive Control Structures

Katalin GYÖRGY, László DÁVID

Department of Electrical Engineering, Faculty of Technical and Human Sciences,
Sapientia University, Tg. Mureş,
e-mail: kgyorgy@ms.sapientia.ro, ldavid@ms.sapientia.ro

Manuscript received October 28, 2010; revised November 03, 2010.

Abstract: The industrial implementation of advanced multivariable control techniques like Model Predictive Control (MPC) is complex, time consuming and therefore it is expensive. Nowdays it is a popular research area to reduce the complexity of the MPC algorithm while preserving the control performance. This problem could be solvable with implementation of the MPC solution in a distributed way. The main idea of this work is to develop simple software agents that can be easily implemented in low cost embedded systems. Each one of these software agents solves the problem of finding one of the control actions with parallel computational facilities. This paper presents at first some general and theoretical considerations about centralized and distributed model predictive algorithms. The comparison between these algorithms is made using numerical simulation of these methods for a multiple input and multiple output theoretical linear discrete-time system. The comparision is possible to be made from the point of wiew of the normalized absolute reference tracking error. There is described a possible implementation of the distributed MPC algorithm using Matlab Simulink environment. It is important to notice that the algorithm to be solved by each software agent while computing the control action is much simpler than the one to be solved by the centralized algorithm.

Keywords: optimal control, cost function, model predictive control, distributed control, prediction horizon, control horizon, Jacobi over-relaxation method, linear constrained optimal programming problem.

1. Introduction

The model predictive control (MPC), – also called receding horizon control (RHC) – is the most important advanced control technique which has been very successful in practical applications, where the control signal can be obtained by solving a discrete-time optimal control problem over a finite horizon. The most important advantage of the MPC algorithms is the fact that they have the unique ability to take into account constraints imposed on process inputs, process state variables and outputs, which usually determine the quality, the efficiency, and safety of production. Implementation of centralized state space (SS) MPC algorithms is becoming an important issue for different multivariable industrial processes. The main idea of our work is to develop a multi-agent software that can be implemented in low cost embedded systems, with parallel computational facilities. These software agents are valid for a default model and can be multiplied and customized according to the control horizon. Each one solves the problem of finding one of the control actions. This procedure is repeated several times before the control action values are delivered to the final control elements. An agent, as an executive, has to know general information about the system and some others which are specific of his own department. It is important to notice that the algorithm to be solved by each agent while computing its control action is much simpler than the one to be solved by the centralized solution.

Previous works on distributed MPC [2], [3], [4], [6] use a wide variety of approaches, including multi-loop ideas, decentralized computation using standard coordination techniques, robustness to the actions of others, penalty functions, and partial grouping of computations. The key point is that, when decisions are made in a decentralized fashion, the actions of each subsystem must be consistent with those of the other subsystems, so that decisions taken independently do not lead to a violation of the coupling constraints. The decentralization of the control becomes more complex when disturbances act on the subsystems making the prediction of future behavior uncertain.

We will analyze how the overall performance of a distributed system is influenced if one or more agents – except the coordinating agent –, fail or obviously underperforms from some reasons. The objective is to solve SS-MPC problems with locally relevant variables, costs, and constraints, but without solving a centralized SS-MPC problem. The coordinated distributed computations solve an equivalent centralized SS-MPC problem. This means that properties that can be proven for the equivalent centralized MPC problem (e.g., stability, robustness) are valid to the above distributed SS-MPC implementation. The significance of the proposed distributed control scheme is that it reduces the computational requirements in complex large-scale systems and it makes possible the development of fault tolerant control systems.

2. Centralized State Space Model Predictive Control

All the MPC algorithms possess common elements and different options can be chosen for each one of these elements: prediction model, objective function and algorithms for obtaining the control law. In this paper the process model is a discrete input – state – output relationship:

$$\begin{aligned} \underline{x}_{k+1} &= \underline{\Phi} \cdot \underline{x}_k + \underline{\Gamma} \cdot \underline{u}_k \\ \underline{y}_k &= \underline{C} \cdot \underline{x}_k \end{aligned}, \tag{1}$$

where $\underline{x}_k$ is the state vector (*n x 1*), $\underline{u}_k$ is the input vector (*m x 1*), $\underline{y}_k$ is the output vector (*p x 1*), and $\underline{\Phi}$, $\underline{\Gamma}$ and $\underline{C}$ are the matrices of the system. If these matrices (parameters) are unknown, we have to implement a system identification module in the control algorithm.

The centralized model predictive algorithm looks for the vector $\Delta\underline{U}_k$ that minimizes a cost function represented by the scalar J, defined as:

$$J\left(\Delta\underline{U}_k\right) = \left(\underline{Y}_k - \underline{Y}_k^{ref}\right)^T \cdot \underline{Q} \cdot \left(\underline{Y} - \underline{Y}_k^{ref}\right) + \Delta\underline{U}_k^T \cdot \underline{R} \cdot \Delta\underline{U}_k , \tag{2}$$

where $\underline{Y}_k^{ref}$ is the vector with the future references, $\underline{Y}_k$ is the vector of the predictions of the controlled variables (output signals), $\Delta\underline{U}_k$ is a vector of future variations of the control signal, $\underline{Q}$ is a diagonal matrix with weights for set-point following enforcement, $\underline{R}$ is a diagonal matrix with weights for control action changes. If the prediction horizon is N and the control horizon is N_c these vectors and matrices are [1]:

$$\underline{Y}_k = \begin{bmatrix} \underline{y}_{k+1/k} \\ \vdots \\ \underline{y}_{k+N/k} \end{bmatrix}, \quad \underline{Y}_k^{ref} = \begin{bmatrix} \underline{y}_{k+1/k}^{ref} \\ \vdots \\ \underline{y}_{k+N/k}^{ref} \end{bmatrix}, \quad \Delta\underline{U}_k = \begin{bmatrix} \underline{\Delta u}_{k/k} \\ \vdots \\ \underline{\Delta u}_{k+N_c-1/k} \end{bmatrix}, \tag{3}$$

$$\underline{Q} = \begin{bmatrix} \underline{Q}_1 & 0 & \dots & 0 \\ 0 & \underline{Q}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \underline{Q}_N \end{bmatrix}, \qquad \underline{R} = \begin{bmatrix} \underline{R}_0 & 0 & \dots & 0 \\ 0 & \underline{R}_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \underline{R}_{N_c-1} \end{bmatrix}. \tag{4}$$

An incremental state space model can be used if the model input is the control increment $\Delta\underline{u}_k = \underline{u}_k - \underline{u}_{k-1}$. The following representation is obtained for predictions:

$$\underline{Y}_k = \underline{\Phi}^* \cdot \underline{x}_k + \underline{\Gamma}^* \cdot \underline{u}_{k-1} + \underline{G}_y \cdot \Delta\underline{U}_k \,, \tag{5}$$

where

$$\underline{\Phi}^* = \begin{bmatrix} \underline{C}\cdot\underline{\Phi} \\ \vdots \\ \underline{C}\cdot\underline{\Phi}^{N_c} \\ \underline{C}\cdot\underline{\Phi}^{N_c+1} \\ \vdots \\ \underline{C}\cdot\underline{\Phi}^{N} \end{bmatrix} \quad \underline{\Gamma}^* = \begin{bmatrix} \underline{C}\cdot\underline{\Gamma} \\ \vdots \\ \sum_{i=0}^{N_c-1} \underline{C}\cdot\underline{\Phi}^i\cdot\underline{\Gamma} \\ \sum_{i=0}^{N_c} \underline{C}\cdot\underline{\Phi}^i\cdot\underline{\Gamma} \\ \vdots \\ \sum_{i=0}^{N-1} \underline{C}\cdot\underline{\Phi}^i\cdot\underline{\Gamma} \end{bmatrix} \quad \underline{G}y = \begin{bmatrix} & & \\ \underline{C}\cdot\underline{\Gamma} & \cdots & \underline{0} \\ \vdots & \cdots & \vdots \\ \sum_{i=0}^{N_c} \underline{C}\cdot\underline{\Phi}^i\cdot\underline{\Gamma} & \ddots & \underline{C}\cdot(\underline{\Phi}\cdot\underline{\Gamma}+\underline{\Gamma}) \\ \vdots & \vdots & \vdots \\ \sum_{i=0}^{N-1} \underline{C}\cdot\underline{\Phi}^i\cdot\underline{\Gamma} & \cdots & \sum_{i=0}^{N-N_c} \underline{C}\cdot\underline{\Phi}^i\cdot\underline{\Gamma} \end{bmatrix}. \tag{6}$$

The cost function can be written as:

$$J\left(\Delta\underline{U}_k\right) = \frac{1}{2}\Delta\underline{U}_k{}^T \cdot \underline{H} \cdot \Delta\underline{U}_k + \underline{f}^T \cdot \Delta\underline{U}_k + const \,, \tag{7}$$

where

$$\begin{aligned} \underline{H} &= 2\cdot\left(\underline{G}_y{}^T \cdot \underline{Q} \cdot \underline{G}_y + \underline{R}\right), \\ \underline{f} &= -2\cdot\underline{G}_y{}^T \cdot \underline{Q} \cdot \underline{E}_k \,, \\ \underline{E}_k &= \underline{Y}_k^{ref} - \underline{\Phi}^* \cdot \underline{x}_k - \underline{\Gamma}^* \cdot \underline{u}_{k-1} \,. \end{aligned} \tag{8}$$

For problems without constraints the centralized model predictive control determines the vector $\Delta\underline{U}_k$ that makes

$$\frac{\partial J(\Delta\underline{U}_k)}{\partial(\Delta\underline{U}_k)} = 0 \Rightarrow \Delta\underline{U}_k^{opt} = \frac{1}{2}\cdot\left(\underline{H} + \underline{H}^T\right)^{-1}\cdot\underline{f} = \left(\underline{G}_y^T\cdot\underline{Q}\cdot\underline{G}_y + \underline{R}\right)^{-1}\cdot\underline{G}_y^T\cdot\underline{Q}\cdot\underline{E}_k \tag{9}$$

It is to be mentioned that only the first control action is taken at each instant, and the procedure is repeated for the next control decision in a receding horizon fashion.

3. Distributed State Space Model Predictive Control

The implementation of distributed model predictive control needs to search for $\Delta\underline{u}_{k/k}, \Delta\underline{u}_{k+1/k}, \ldots, \Delta\underline{u}_{k+N_c-1/k}$ [5], [7], that makes the

$$\frac{\partial J(\Delta\underline{U}_k)}{\partial(\Delta\underline{u}_{g/k})} = 0 \ . \tag{10}$$

for $k \le g \le k + N_c - 1$.

If the cost function is written as:

$$J(\Delta \underline{U}_k) = \Delta \underline{U}_k{}^T \cdot \left(\underline{G}_y^T \cdot \underline{Q} \cdot \underline{G}_y + \underline{R}\right) \cdot \Delta \underline{U}_k - 2 \cdot \underline{E}_k^T \cdot \underline{Q} \cdot \underline{G}_y \cdot \Delta \underline{U}_k + \underline{E}_k^T \cdot \underline{Q} \cdot \underline{E}_k \tag{11}$$

then the first order optimality condition can be determined in the following way:

$$\begin{aligned} \frac{\partial J(\Delta \underline{U}_k)}{\partial (\Delta \underline{u}_{g/k})} &= 2 \cdot \left[\underline{G}_y^T \cdot \underline{Q} \cdot \underline{G}_y + \underline{R}\right]_{g-k+1,g-k+1} \cdot \Delta \underline{u}_{g,k} - \\ &- 2 \cdot \sum_{i=1}^{N} \left(\left[\underline{Q} \cdot \underline{G}_y\right]^T_{i,g-k+1} \cdot \left[\underline{E}_k\right]_i \right) + \\ &+ \sum_{\substack{i=0 \\ i \neq g-k}}^{N_c-1} \left(\left([\underline{G}_y^T \cdot \underline{Q} \cdot \underline{G}_y + \underline{R}]^T_{i+1,g-k+1} + [\underline{G}_y^T \cdot \underline{Q} \cdot \underline{G}_y + \underline{R}]_{g,i+1} \right) \Delta \underline{u}_{k+i,k} \right) \end{aligned} \tag{12}$$

The variation of input signal is

$$\begin{aligned} \Delta \underline{u}_{g/k} = &\left(2 \cdot \left[G_y^T \cdot Q \cdot G_y + R\right]_{g-k+1,g-k+1}\right)^{-1} \cdot \left(2 \cdot \sum_{i=1}^{N} \left(\left[\underline{Q} \cdot \underline{G}_y\right]^T_{i,g-k+1} \cdot \left[\underline{E}_k\right]_i \right) - \right. \\ &\left. - \sum_{\substack{i=0 \\ i \neq g-k}}^{N_c-1} \left(\left([\underline{G}_y^T \cdot \underline{Q} \cdot \underline{G}_y + \underline{R}]^T_{i+1,g-k+1} + [\underline{G}_y^T \cdot \underline{Q} \cdot \underline{G}_y + \underline{R}]_{g,i+1} \right) \Delta \underline{u}_{k+i,k} \right) \right) \end{aligned} \tag{13}$$

The first value of every $\Delta \underline{u}_{g/k}$ is only an approximation since it depends on the other $\Delta \underline{u}_{i+k/k}$ values ($i \neq g-k$). It should be noticed that the computation burden to obtain $\Delta \underline{u}_{g/k}$ is much smaller than the one needed to compute the whole vector $\Delta \underline{U}_k$. As already discussed, in this distributed approach, the vector $\Delta \underline{U}_k$ is determined by software agents using a combination of repeated computation of $\Delta \underline{u}_{g/k}$ and exchange of information.

The equation (13) can be written in the following general form:

$$\Delta \underline{u}^n_{k+j/k} = \sum_{\substack{i=0 \\ i \neq j}}^{N_c-1} \left(\underline{A}_{j+1,i+1} \cdot \Delta \underline{u}^{n-1}_{k+i/k} + \underline{B}_{j+1} \right), \tag{14}$$

where $0 \leq j \leq N_c - 1$, the matrices $\underline{A}_{i,j}$ have the dimension $m \times m$ and vector $\underline{B}_i$ has the dimension $m \times 1$, where m is the number of inputs. Matrix $\underline{A}_{i,i}$ is zero. A centralized expression for $\Delta \underline{U}_k$ using equation (14) can be written as:

$$\begin{bmatrix} \underline{\Delta u}_{k/k} \\ \underline{\Delta u}_{k+1/k} \\ \vdots \\ \underline{\Delta u}_{k+N_c-1/k} \end{bmatrix}^n = \begin{bmatrix} \underline{0} & \underline{A}_{1,2} & \cdots & \underline{A}_{1,Nc} \\ \underline{A}_{2,1} & \underline{0} & \cdots & \underline{A}_{2,Nc} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{A}_{Nc,1} & \underline{A}_{Nc,2} & \cdots & \underline{0} \end{bmatrix} \cdot \begin{bmatrix} \underline{\Delta u}_{k/k} \\ \underline{\Delta u}_{k+1/k} \\ \vdots \\ \underline{\Delta u}_{k+N_c-1/k} \end{bmatrix}^{n-1} + \begin{bmatrix} \underline{B}_1 \\ \underline{B}_2 \\ \vdots \\ \underline{B}_{Nc} \end{bmatrix} \quad (15)$$

which, in a compact form becomes

$$\underline{\Delta U}_k^n = \underline{A} \cdot \underline{\Delta U}_k^{n-1} + \underline{B} \,. \quad (16)$$

The convergence of the $\underline{\Delta U}_k$ vectors to their true values has to be assured for a reliable application. For unconstrained applications the results obtained in the field of distributed computation can be used [5]. The Jacobi over-relaxation approach is adopted here by recomputing $\underline{\Delta U}_k$ as a linear combination of the value computed using equation (16) and the value obtained in the previous iteration,

$$\Delta \underline{U}_{k,\,filtered}^n = (I - diag(\underline{\alpha})) \cdot \Delta \underline{U}_k^n + diag(\underline{\alpha}) \cdot \Delta \underline{U}_k^{n-1} \quad (17)$$

where $\underline{\alpha}$ is a vector of the filter parameters. Applying the filter according to equation (16), it results:

$$\begin{aligned} \underline{\Delta U}_k^n &= \left(\left(\underline{I} - diag(\underline{\alpha})\right) \cdot \underline{A} + diag(\underline{\alpha})\right) \cdot \underline{\Delta U}_k^{n-1} + \left(\underline{I} - diag(\underline{\alpha})\right) \cdot \underline{B} = \\ &= \underline{A}(\underline{\alpha}) \cdot \underline{\Delta U}_k^{n-1} + \left(\underline{I} - diag(\underline{\alpha})\right) \cdot \underline{B} \end{aligned} \quad (18)$$

A sufficient condition for convergence of the iterative process is to have $\left\|\underline{A}(\underline{\alpha})\right\| < 1$ for $\alpha \in (0,1)$. The search for a filter vector $\underline{\alpha}$ which minimizes $\left\|\underline{A}(\underline{\alpha})\right\|$ can be reduced to a linear constrained optimal programming problem.

4. Numerical simulation

This section presents the application of the centralized and distributed model predictive algorithm to a multiple input and a multiple output theoretical system which is characterized by following state space model:

$$\begin{aligned} \underline{x}_{k+1} &= \begin{bmatrix} 0.7 & 0 & 0.1 & 0 \\ 0 & -0.5 & 0.2 & 0 \\ 0 & 0.01 & 0.1 & 0 \\ 0.01 & 0 & 0 & -0.5 \end{bmatrix} \cdot \underline{x}_k + \begin{bmatrix} 4 & 0 \\ 3 & 9 \\ -10 & 1 \\ 0 & 2 \end{bmatrix} \cdot \underline{u}_k \\ \underline{y}_k &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \underline{x}_k \end{aligned} \quad (19)$$

where

$$\underline{x}_k = \begin{bmatrix} x_{1,k} & x_{2,k} & x_{3,k} & x_{4,k} \end{bmatrix}^T, \underline{u}_k = \begin{bmatrix} u_{1,k} & u_{2,k} \end{bmatrix}^T \text{ and } \underline{y}_k = \begin{bmatrix} y_{1,k} & y_{2,k} \end{bmatrix}^T.$$

For both algorithms the Simulink models have been built and the following parameters were used for both simulations:

$$N = 4; \quad N_c = 3; \quad R = 0.1 \cdot I_2 \quad Q = 10 \cdot I_2 \tag{20}$$

The Simulink diagram of the centralized predictive control is shown in *Fig. 1*, where the "Centralized_MPC_control" subsystem contains one complex S-function module for centralized control algorithm. The Simulink model of the distributed predictive algorithm is shown in *Fig. 2*.

The "Distributed_MPC_control" subsystem is presented separately in *Fig. 3*, where the three interdependent modules for calculating $\Delta\underline{u}_{1/k}, \Delta\underline{u}_{2/k}, \Delta\underline{u}_{3/k}$ can be observed. The structure of these modules is one and the same, just the input signals and parameters are different.

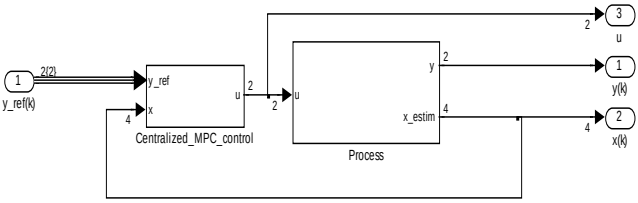


Figure 1: Simulink diagram for numerical simulation of the centralized predictive control.

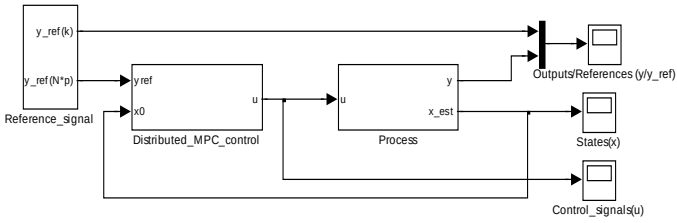


Figure 2: Simulink diagram for numerical simulation of the distributed predictive control.

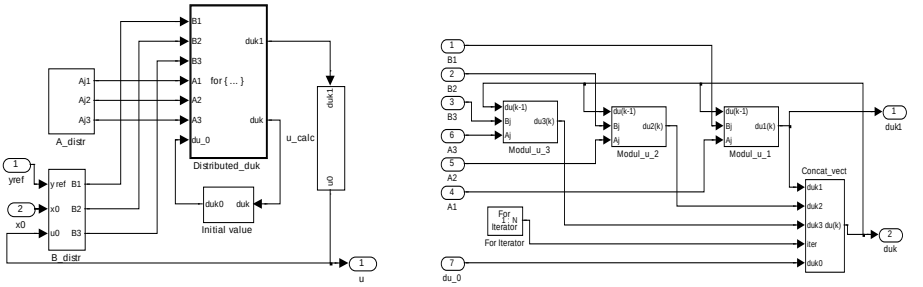


Figure 3: Subsystem diagram for distributed predictive algorithms (Nc=3).

The choice of alpha (α) provides all eigenvalues of matrix $A(\underline{\alpha})$ of equation (18) smaller than 1, which is sufficient to assure that the iterative method converges. These values were determined before the numerical simulation, and the one optimal constrained problem was solved in Matlab environment. It seems that the parameter tuning for the distributed algorithm does not need to be exactly the same as the one used for the centralized version. For the same amount of information exchange among agents, a faster reference filter improves the response.

The results obtained by numerical simulation for the centralized control algorithm using a variable reference signal are shown in *Fig. 4.* and results of numerical simulation of the distributed algorithm after 500 iterations are shown in *Fig. 5.*

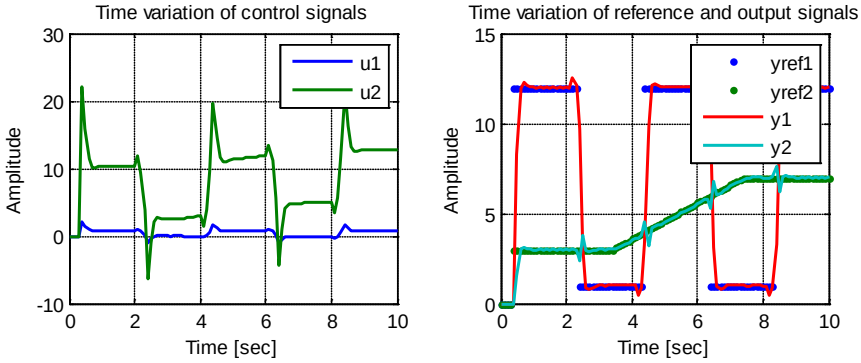
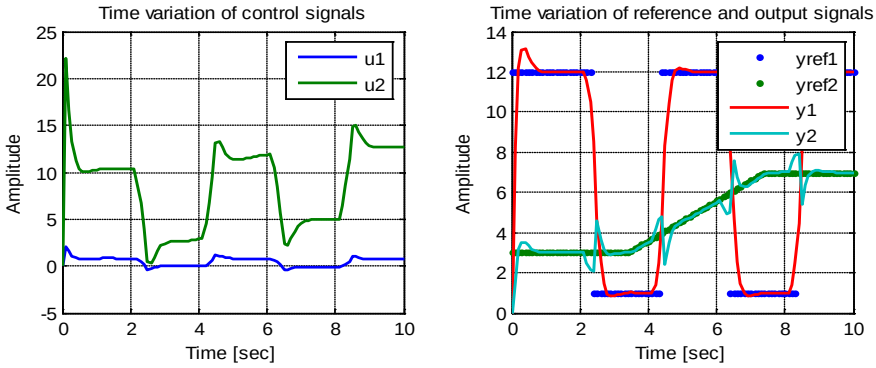


Figure 4: Time variation of the control signals (a) and outputs signals (b) in case of the centralized algorithm.



(a) (b)

Figure 5: Time variation of the control signals (a) and outputs signals (b) in case of distributed algorithm after 500 iterations.

It is noticeable that control signal obtained with the distributed algorithm is smoother than the control signal obtained with the centralized algorithm.

In order to have an idea on the number of information interchange iterations between agents needed for a certain performance, an analysis has been made based on the error betveen the outputs and reference signals. For both outputs (i=1,2) it was computed the error at every sample time k =1,…, N_t:

$$e_{i,k} = \frac{(y_{i,k} - y_{i,k}^{ref})}{y_{i,k}^{ref}} \tag{21}$$

A normalized absolute reference tracking error was computed using following relationship:

$$e_{abs} = \frac{1}{2}\sum_{i=1}^{2}\left(\frac{1}{N_t}\sum_{k=1}^{N_t}\left|e_{i,k}\right|\right) \tag{22}$$

There was estimated also the simulation time using the Matlab functions *clock* and *etime* for different numbers of iteration in case of the distributed model predictive control algorithms. *Fig. 6.a* presents the variation of the average errors and *Fig. 6.b* presents the estimated simulation time versus the number of iterations in case of the distributed algorithms. The simulation time is more important to be calculated for both algorithms (centralized version and distributed version) with different prediction horizon (N) values. This comparision is presented at *Fig. 7* in case of a fixed control horizon (Nc=3).

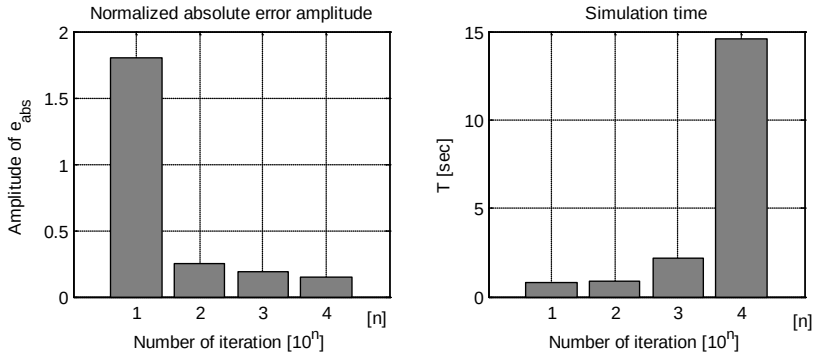


Figure 6: Variation of the reference tracking error (a) and of the simulation time (b) versus the number of iterations in case of the distributed algorithm.

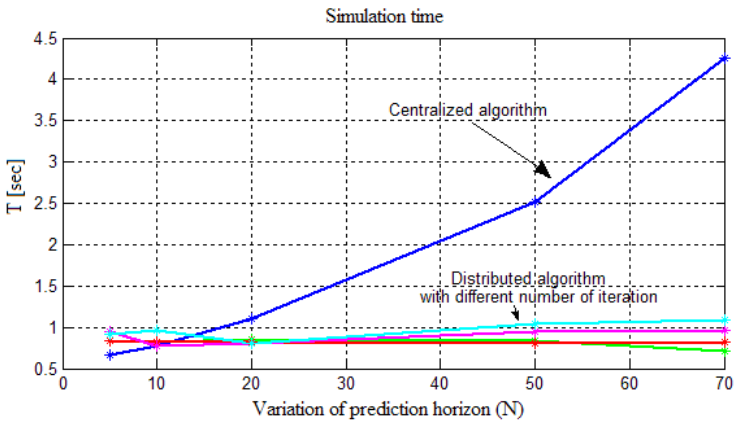


Figure 7: Comparison of the simulation times represented in function of the prediction horizon (N), obtained in case of the centralized respectively in case of the distributed algorithm.

5. Conclusion

The performance of the distributed control applied for the example discussed in the paper is comparable to that obtained with the centralized model predictive control. The computation power needed to solve the distributed problem is smaller than that is needed for the centralized case. This fact may allow the utilization of the model predictive control executed in distributed hardware with low computational power. The size of the centralized problem grows considerably with the number of inputs/outputs while the size of the

distributed problem remains the same for the same control horizon. One point to mention is that unlike in the case of the presented example, most of the multivariable problems do not have a complete interaction. In case of the distributed algorithms the problem is to choose the convenient sample time and the correct filter parameters' vector. The choice of the filter should be done off-line and the condition presented is enough to ensure the convergence of the algorithm. Future developments are needed to provide the best filter option (assuring the fastest convergence with robustness) and to introduce some constraints in the model predictive applications. The main benefit expected in case of the distributed MPC control is the improvement of the system's maitainability and the 'apparent' simplicity to the user.

References

[1] Camacho, E. F., "Model Predictive Control", Springer Verlag, 2004.

[2] Camponogara, E., Jia, D., Krogh, B. H. and Talukdar, S. N., "Distributed model predictive control", *IEEE Control Systems Magazine*, vol. 22, no. 1, pp. 44–52, February 2002.

[3] Venkat, A. N., Rawlings, J. B. and Wright, S. J., "Implementable distributed model predictive control with guaranteed performance properties", *American Control Conference Minneapolis, Minnesota, USA*, June 14-16, 2006, pp. 613-618.

[4] Mercangoz, M. and Doyle, F. J, "Distributed model predictive control of an experimental four tank system", *Journal of Process Control*, vol. 17, no. 3, pp. 297–308, 2007.

[5] Plucenio, A., Pagano, D. J., Camponogara, E., Sherer, H. F. and Lima, M., "A simple distributed MPC algorithm", Rio de Janerio, Brasil.

[6] Maestre, J. M., Munoz de la Pena, D. and Camacho, E. F., "Distributed MPC: a supply chain case study", *IEEE Conference on Decision and Control, Shanghai, China*, December 16-18, 2009, pp. 7099 – 7104.

[7] Venkat, A. N., Hiskens, I. A., Rawlings, J. B. and Wright, S. J. "Distributed MPC Strategies With Application to Power System Automatic Generation Control", *IEEE Transactions on Control Systems Technology*, vol. 16, no. 6, pp. 1192-1206, November, 2008.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 16-26

Synchronous Motor Drive at Maximum Power Factor with Double Field-Orientation

Csaba SZABÓ, Maria IMECS, Ioan Iov INCZE

Department of Electrical Drives and Robots, Faculty of Electrical Engineering,
Technical University of Cluj-Napoca,
e-mail: csaba.szabo@edr.utcluj.ro, imecs@edr.utcluj.ro, ioan.incze@edr.utcluj.ro

Manuscript received Oct 1, 2010; revised Oct 15, 2010.

Abstract: The paper presents a vector control structure for a wound-excited salient-pole synchronous motor, fed by a voltage-source converter, working at unity power factor. The variable exciting current is ensured by a DC chopper. Due to this additional intervention possibility the motor may have three degrees of freedom from the control point of view, and three control loops will be formed instead of two: one for the control of the mechanical quantities, and two for the magnetic ones. The three prescribed references are the rotor angular speed, the stator flux (both directly controlled by using PI regulators) and the power factor that is only imposed at its maximum value. In the control structure two types of orientation procedure are used: stator-field-orientation for power factor control, and rotor-orientation for computation of the voltage-control variables and self commutation. There is also presented a speed-computation procedure used in practical implementation regarding the signal processing of the incremental encoder position. The method is based on the derivation with respect to time of both sine and cosine functions of the rotor position. The angular speed is obtained then by computing the module of the two resulted sinusoidal signals. This method avoids the division by zero related issue that occurs at every zero crossing if the angular speed is computed by dividing the time based derivative of one signal with the other one. For validation of the presented control strategy simulations were carried out in Matlab/Simulink® environment.

Keywords: Vector control, unity power-factor, voltage-controlled drive, rotor speed identification, voltage-source inverter, stator-field orientation.

1. Introduction

For high performance dynamic applications the most suitable solution is the vector controlled AC drive fed by a static frequency converter (SFC). The wound-excited synchronous motor (Ex-SyM) is the only machine capable to operate at unity or leading power factor (PF). The structure of the vector control system is determined by the combination between the type of the SFC used including the pulse width modulation (PWM) procedure, the orientation field and its identification method [2], [8], [9].

The rigorous control of the PF can be made only with the resultant stator-field orientation. If the PF is maximum, there is no reactive energy transfer between the armature and the three-phase power source.

Some motor-control-oriented digital signal processing (DSP) equipments present on the market don't dispose over implementation possibility of the current-feedback PWM, suitable for current-controlled VSIs, consequently in the control structure it is necessary the computation of the voltage control variables from the current ones, imposed or directly generated by the controllers.

The proposed control structure is based on both types of orientation. The stator-field orientation is used for control of the unity power factor and stator-flux, and also for generation of the armature-current control variables. The orientation according to the rotor position (i.e. exciting-field orientation) is applied for self-commutation and for generation of the armature-voltage control variables for the inverter control. The transition between the two orientations is performed by using a coordinate transformation block (CooT), which rotates the stator-field oriented reference frame with the value of the load angle ($\delta = \lambda_s - \theta$).

2. The mathematical model of the Ex-SyM

The mathematical model (MM) of the Ex-SyM is suitable for simulation and also in implementation because, the computation of the control and feedback variables in the control structure is performed also based on the MM. Usually the equations are written in rotor-oriented rotating reference frame ($d\theta - q\theta$), where θ is the rotor position. The state equations based on the quasi-flux model may be written, as follows:

$$\begin{cases} \dfrac{d\Psi_{sd\theta}}{dt} = u_{sd\theta} - R_s \cdot i_{sd\theta} + \omega \cdot \Psi_{sq\theta}; \\ \dfrac{d\Psi_{sq\theta}}{dt} = u_{sq\theta} - R_s \cdot i_{sq\theta} - \omega \cdot \Psi_{sd\theta}; \\ \dfrac{d\Psi_e}{dt} = u_e - R_e \cdot i_e; \\ \dfrac{d\Psi_{A_d}}{dt} = u_{A_d} - R_{A_d} \cdot i_{A_d}; \\ \dfrac{d\Psi_{A_q}}{dt} = u_{A_q} - R_{A_q} \cdot i_{A_q}; \\ \dfrac{d\omega}{dt} = \dfrac{z_p}{J_{tot}} \cdot \left[\dfrac{3}{2} z_p (\Psi_{sd\theta} \cdot i_{sq\theta} - \Psi_{sq\theta} \cdot i_{sd\theta}) - m_L \right], \end{cases} \tag{1}$$

The integration of the state equations is made directly from the derivatives of the angular rotor speed and fluxes, then the currents are computed from the fluxes expressed according to the longitudinal $d\theta$ rotor axis:

$$\begin{cases} i_{sd\theta} = \dfrac{1}{L''_{sd}} \Psi_{sd\theta} - \dfrac{1}{L''_{m(sd\theta - A_d)}} \Psi_{A_d} - \dfrac{1}{L''_{m(sd\theta - e)}} \Psi_e \\ i_{A_d} = -\dfrac{1}{L''_{m(sd\theta - A_d)}} \Psi_{sd\theta} + \dfrac{1}{L''_{A_d}} \Psi_{A_d} - \dfrac{1}{L''_{m(e - A_d)}} \Psi_e \\ i_e = -\dfrac{1}{L''_{m(sd\theta - e)}} \Psi_{sd\theta} - \dfrac{1}{L''_{m(e - A_d)}} \Psi_{A_d} + \dfrac{1}{L''_e} \Psi_e \end{cases} \tag{2}$$

and according to the quadrature $q\theta$ rotor axis:

$$\begin{cases} i_{sq\theta} = \dfrac{1}{L''_{sq}} \left(\Psi_{sq\theta} - \dfrac{L_{mq}}{L_{A_q}} \Psi_{A_q} \right) \\ i_{A_q} = -\dfrac{1}{L''_{Aq}} \left(\Psi_{Aq} - \dfrac{L_{mq}}{L_{sq}} \Psi_{sq\theta} \right) \end{cases} \tag{3}$$

As state-variables were chosen the fluxes (i.e. the direct and quadrature axis components of the stator flux ($\Psi_{sd\theta}$ and $\Psi_{sq\theta}$) and of the damper winding flux (Ψ_{Ad} and Ψ_{Aq}), the exciting winding flux (Ψ_e) and the rotor electrical angular speed (ω).

3. Double field-oriented control of the Ex-SyM

In the proposed control structure, presented in *Fig. 1*, the salient pole Ex-SyM is fed by a voltage-source inverter (VSI). Three control loops are formed: two magnetical (for flux and PF control) and a mechanical one (for speed). The flux and the speed are controlled directly by PI controllers, and the PF is controlled indirectly. The PF is maximum, if the stator voltage and stator current are in phase. Consequently, the stator-current space phasor $\underline{i}_s$ results perpendicular onto the stator-flux vector $\underline{\Psi}_s$ [2], [8].

The perpendicularity can be achieved by canceling the stator-field-oriented longitudinal armature reaction ($i_{sd\lambda s}=0$). The reference armature-current components are oriented according to the direction of the resultant stator-flux phasor. The longitudinal component $i_{sd\lambda s}^{Ref}$ is an imposed value, and it is cancelled, while the quadrature component $i_{sq\lambda s}^{Ref}$ results at the output of the speed controller. However the stator-current control is recommended to be made in exciting field- (i.e. rotor-position-) oriented ($d\theta$–$q\theta$) rotating reference frame, because the self-commutation of the motor, based on the rotor position, is made inherently by means of the reverse Park transformation block. The re-orientation (from the resultant stator field to the exciting one) is made by means of a reverse CooT block, which rotates the stator-field-oriented reference frame with the value corresponding to the load angle $\delta=\lambda_s-\theta$ [2], [8], [9].

The voltage-control variables are generated by the two current controllers in the active and the reactive control loops. In the voltage-computation block UsC the electromagnetic cross-effect is taken into account, realizing the re-coupling of the two decoupled control loops, the active and the reactive ones, by means of the rotating EMF components [2], [8], as follows:

$$\begin{cases} u_{sd\theta}^{Ref} = v_{sd\theta}^{Ref} - \omega\Psi_{sq\theta}; \\ u_{sq\theta}^{Ref} = v_{sq\theta}^{Ref} + \omega\Psi_{sd\theta} \end{cases} \tag{4}$$

The inverter control is made by feed-forward voltage-PWM procedure with simple on-off controllers [5], based on the following PWM logic:

$$m^{\log} = \begin{cases} -1, \text{ if } u_{cr} < u^{Ref}; \\ 1, \quad \text{if } u_{cr} > u^{Ref}. \end{cases} \tag{5}$$

The exciting current is controlled also with voltage PWM by means of a DC chopper.

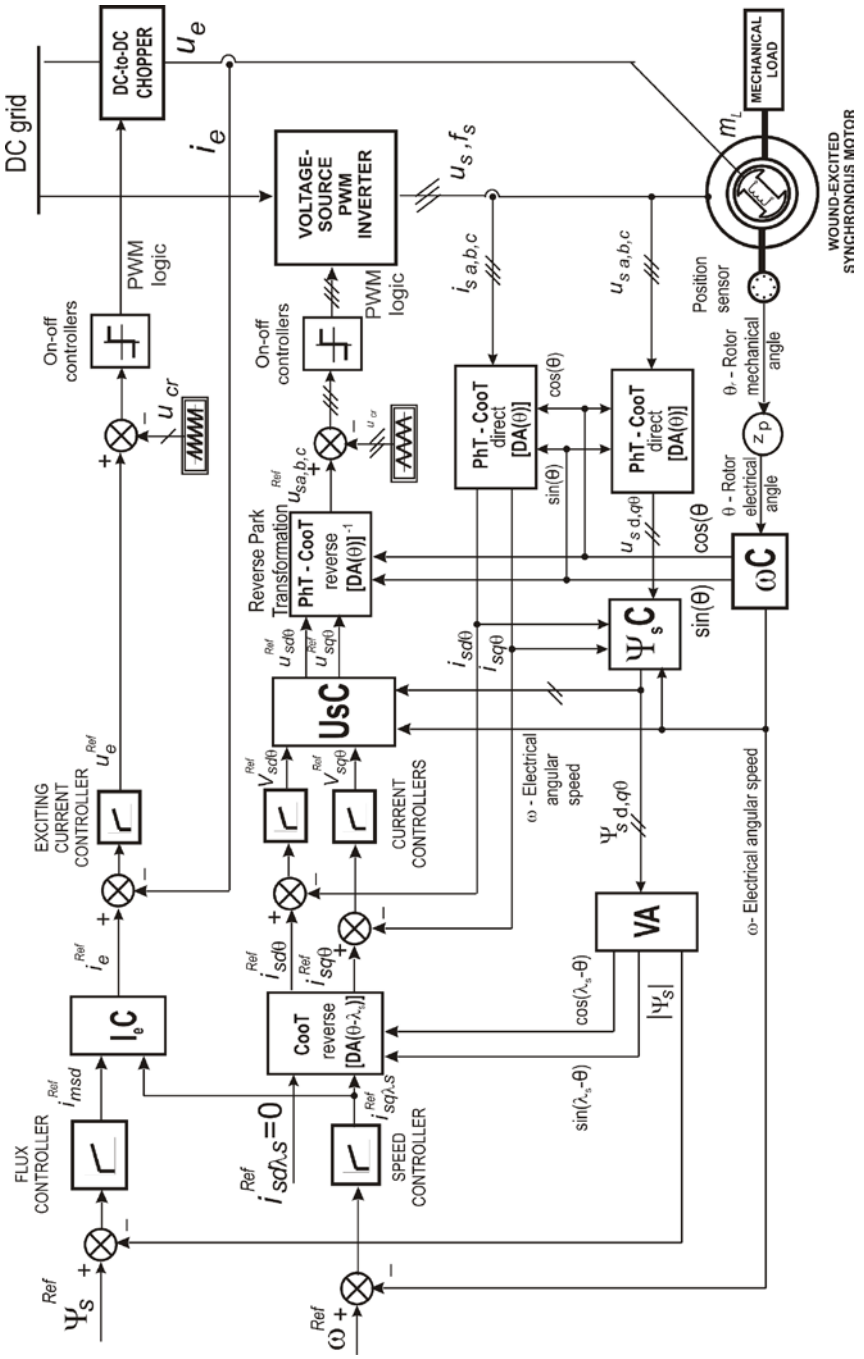


Figure 1: Vector control system of the adjustable excited synchronous motor fed by a static frequency converter with feed-forward voltage-PWM and double field orientation, operating with controlled stator flux and imposed unity power factor.

In the third control loop the resultant stator-flux is directly controlled with a PI controller, which outputs the i_{ms} magnetizing current, necessary for the computation of the excitation current in the IeC block [2], [8].

4. Angular speed computation

For the self-control of the Ex-SyM it is important an accurate information about the rotor position. This was realized using an incremental encoder, mounted on the motor shaft. The mounting is realized in a manner, that the encoder index signal is synchronized with respect to the rotor position. The encoder generates a number of pulses proportional to the angular position of the shaft. It gives information also about the sense of the rotating motion: positive values for direct and negative ones for reverse running. The counter resets it to zero at every full rotation. The amplitude of this signal will be equal to the number of the increments/revolution of the encoder. This position signal θ_{enc} provided by the encoder is processed in order to obtain a position signal θ between $[0, 2\pi]$ for direct, and $[0, -2\pi]$ for reverse rotation respectively, based on the following expression:

$$\theta = \frac{\theta_{enc}}{N_r} 2\pi , \tag{6}$$

where N_r is the number of increments per revolution of the digital encoder.

The electrical angular speed of the rotor is also required in the UsC block for the computation of the voltage control variable. The angular speed can be determined using the sine and cosine functions of the rotor position. The amplitude of these functions is equal to 1. The method presented in [1] is based on the derivation of either the sine or the cosine function of the position, and uses the following relation:

$$\omega = \frac{d\theta}{dt} = \left(\cos\theta\right)^{-1} \frac{d\left(\sin\theta\right)}{dt} \tag{7}$$

The drawback of this method is, that division by zero occurs at every zero crossing of the sinusoidal function, because the $\frac{d\left(\sin\theta\right)}{dt}$ is shifted by 90°, and is in phase with the $\cos\theta$. The resulting angular speed is inaccurate and presents infinite peaks at these moments, that cannot be processed correctly, as is shown in *Fig. 2*. In order to avoid this situation different methods were implemented. A known method is to filter the obtained signal, but this method leads to inappropriate results especially in transient operation when the speed is not constant. Another approach is based on avoiding the division by zero by

introducing a discontinuity in the sine function around zero, assigning instead of this a very small, but nonzero value.

In this paper a different approach is used, which consists in the derivation of both, the sine ($S = \sin\theta$) and cosine ($C = \cos\theta$) functions of the angular position θ, where $S = f(\theta)$, $C = f(\theta)$, and $\theta = f(t)$ [4], [15]:

$$\frac{dS}{dt} = \cos\theta \frac{d(\theta)}{dt} \tag{8}$$

$$\frac{dC}{dt} = -\sin\theta \frac{d(\theta)}{dt} \tag{9}$$

The angular speed is determined by the following relation [15]:

$$|\omega| = \sqrt{\left(\frac{dS}{dt}\right)^2 + \left(\frac{dC}{dt}\right)^2} = \sqrt{\cos^2\theta\left(\frac{d(\theta)}{dt}\right)^2 + \sin^2\theta\left(\frac{d(\theta)}{dt}\right)^2} = \left|\frac{d\theta}{dt}\right| \tag{10}$$

Using this procedure the zero crossing is avoided, and leads to an accurate result, as shown in *Fig. 3*.

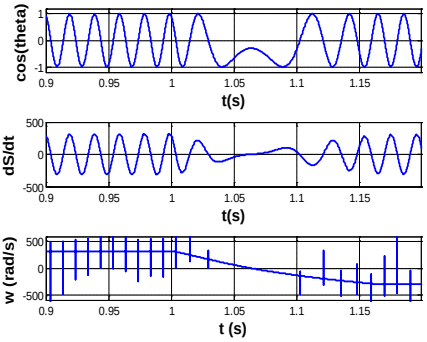


Figure 2: The rotor angular speed computed with the classical method, leading to an inaccurate result

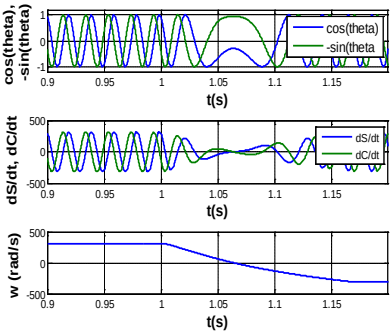


Figure 3: The rotor angular speed computed using the proposed method avoiding the division by zero issue.

The sign of the position signal θ gives the direction of the rotation. The computation of the angular speed is based on the following expression:

$$\omega = |\omega| sign(\theta) = \sqrt{\left(\frac{d(\sin\theta)}{dt}\right)^2 + \left(\frac{d(\cos\theta)}{dt}\right)^2} sign(\theta), \tag{11}$$

and its computation may be processed with the structure presented in *Fig. 4* [4].

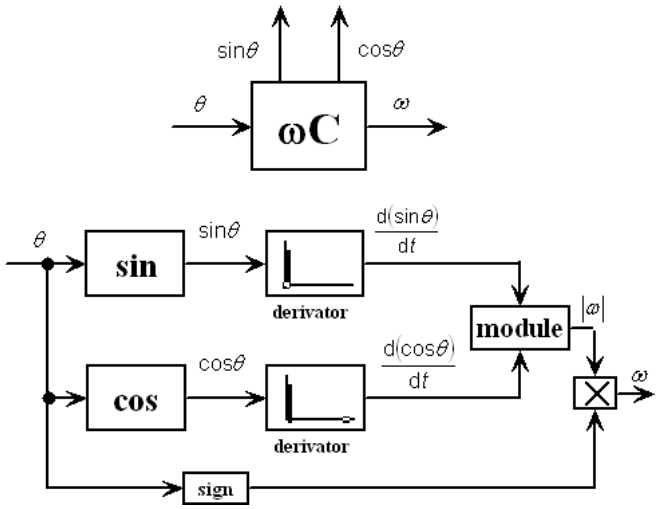


Figure 4: Block symbol and structure for computation of the rotor angular speed based on the encoder position signals.

5. Simulation results

Based on the structure from *Fig. 1* simulations were performed in Matlab-Simulink® environment. The rated data of the simulated salient pole Ex-SyM are: U_{sN} = 380 V, I_{sN} = 1.52 A, P_N = 800 W, f_N=50 Hz, n_N = 1500 [rpm], cosφ= 0.8 (capacitive).

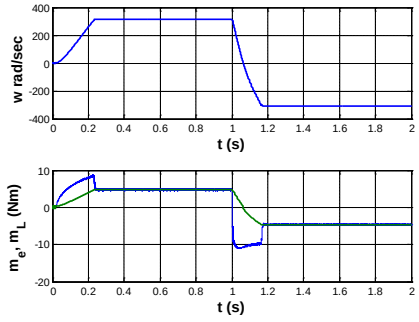


Figure 5: Electrical angular speed (w), electromagnetic (m_e) and load torque (m_L)

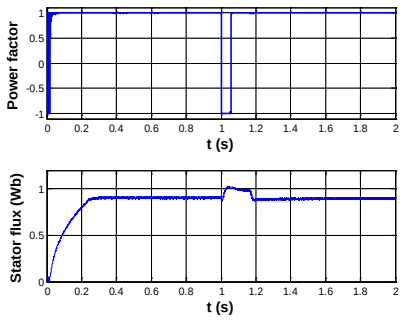


Figure 6: The power factor and the stator-flux amplitude versus time.

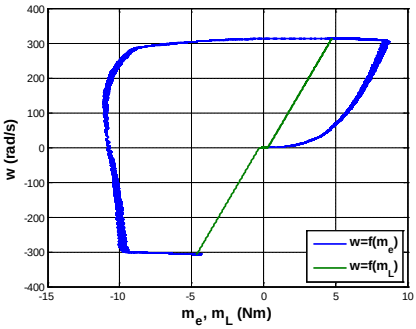


Figure 7: Mechanical characteristics*:* speed versus torque of the motor w=f(m_e) and of the mechanical load w=f(m_L).

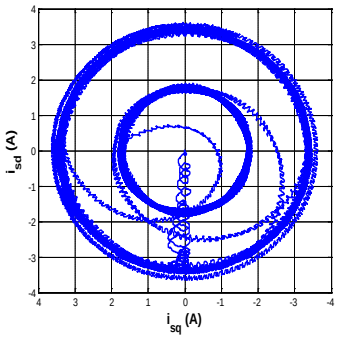


Figure 8: The trajectory of the armature-current space-phasor in natural stator-fixed coordinate frame.

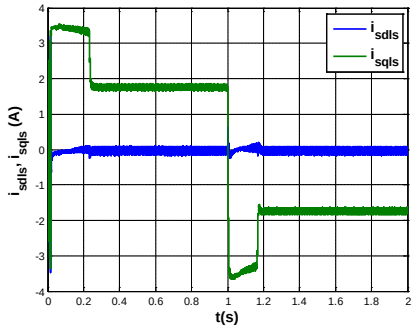


Figure 9. The armature-current two-phase components ($i_{sd\lambda s}$ and $i_{sq\lambda s}$) in stator-flux-oriented coordinate frame.

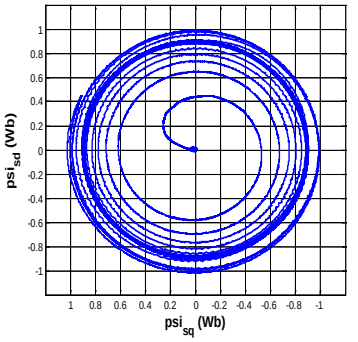


Figure 10. The trajectory of the stator-flux space-phasor in natural stator-fixed coordinate frame.

After the starting process the motor runs at the rated speed value corresponding to a frequency of 50 Hz. At t = 1 s under the full rated load a speed reversal is applied. The mechanical load has reactive character and it is linearly speed-dependent.

The simulation results show that the proposed control structure from *Fig. 4* is a viable one with improved performances with respect to the conventional vector control systems [2].

The results show a good performance of the drive also in transient operation at starting, and also at speed reversal (*Fig. 5*). The power factor is maximum also during the speed reversal, when the drive is in regenerative running for a short period of time, as is shown in *Fig. 6*. Unity power factor is realized by canceling the stator-field-oriented longitudinal armature reaction, as in *Fig. 9*.

6. Conclusion

The presented control structure uses two types of orientations: resultant stator-field and exciting-field, i.e. rotor-position orientation.

For a rigorous control of the power factor, stator-field orientation was applied, and in order to achieve unity power factor operation in the reactive control loop the stator-flux oriented longitudinal armature reaction was cancelled.

In order to obtain improved control performances, the computation of the control variables were made in rotor-oriented reference frame, so the self-control of the motor and the synchronization of the inverter trigger signal are made based on a directly measured value of the rotor position.

In the control structure of the voltage-controlled Ex-SyM drives the dual field-orientation combines the advantages of the two types of field-orientation procedure, on the one hand of the stator-field orientation suitable for power factor control and on the other hand of the exciting-field orientation for computing feedback- and control-variables based on the rotor-position-oriented classical MM, in order to ensure sophisticated calculus due to the geometry characteristics, i.e. two-axis symmetry of the salient-pole rotor.

The applied computation procedure of the synchronous angular speed corresponding to the rotating the stator-flux avoids the division by zero and gives an accurate result.

The applied computation procedure of the rotor angular speed, which avoids the division by zero, gives accurate results also in computation of the synchronous angular speed of the rotating resultant orientation flux in any field-oriented vector control structure, including induction motor drives, too.

The presented control structure was validated by simulation in Matlab/Simulink®, and the obtained result shows the reliability of this method.

The practical implementation was realized on an experimental rig based on the *dSpace DS1104* controller board. The results were published in [8].

References

[1] Kelemen, Á., and Imecs, M., “Vector Control of AC Drives, Volume 1: Vector Control of Induction Machine Drives”, OMIKK-Publisher, Budapest, Hungary, 1991.

[2] Kelemen, Á., Imecs, M., “Vector Control of AC Drives, Volume 2: Vector Control of Synchronous Machine Drives” Ecriture Budapest, Hungary, 1993.

[3] W. Leonhard, “Control of Electrical Drives”, Springer Verlag. Berlin, Heidelberg, New York, Tokyo, 1985.

[4] Szabó, Cs., “Implementation of Scalar and Vector Control Structures for Synchronous Motors (in Romanian)”, PhD Thesis, Technical University of Cluj-Napoca, Romania, 2006.

[5] Kazmierkowski, M. P., Tunia, H., "Automatic Control of Converter-Fed Drives", Elsevier, Amsterdam, 1993.
[6] Vauhkonen, V., "A cycloconverter-fed synchronous motor drive having isolated output phases", in *Proc. International Conference on Electrical Machines, ICEM '84, Lausanne, Switzerland*, 1984.
[7] Imecs, M., Szabó, Cs., Incze, I. I., "Stator-field-oriented control of the variable-excited synchronous motor: numerical simulation", in *Proc. 7th International Symposium of Hungarian Researchers on Computational Intelligence HUCI 200*6, *Budapest, Hungary*, 2006, pp. 95-106.
[8] Szabo, C., Imecs, M., Incze, I. I., "Vector control of the synchronous motor operating at unity power factor", in *Proc. 11th International Conference on Optimization of Electrical and Electronic Equipment, OPTIM 2008, Brasov, Romania,* 2008, vol. II-A pp. 15-20.
[9] Szabo, C., Imecs, M., Incze, I. I., "Synchronous motor drive with controlled stator-field-oriented longitudinal armature reaction", in *Proc The 33rd International Conference of the IEEE Industrial Electronics Society, IECON 2007, Taipei, Taiwan*, 2007, CD-ROM
[10] Davoine, J., Perret, R., Le-Huy H., "Operation of a self-controlled synchronous motor without a shaft position sensor", *Trans. Ind Applications*, IA-19, no. 2, pp. 217-222, March/April 1983.
[11] Shinnaka, S., Sagawa, T., "New optimal current control methods for energy-efficient and wide speed-range operation of hybrid-field synchronous motor," *IEEE Trans. Ind Electronics*, vol. 54, no. 5, pp. 2443-2450, Oct. 2007.
[12] Imecs, M., Incze, I. I., Szabó, Cs., "Double field orientated vector control structure for cage induction motor drive", *Scientific Bulletin of the „Politehnica" University of Timisoara, Romania, Transaction of Power Engineering*, Tom 53(67), Special Issue, pp. 135-140, 2008.
[13] Imecs, M., Incze, I. I., Szabó, Cs., "Dual field orientation for vector controlled cage induction motors", in *Proc. of the 11th IEEE Internat. Conference on Intelligent Engineering Systems, INES 2009, Barbados*, 2009, pp 143-148.
[14] Imecs, M., Szabó, Cs., Incze, I. I., "Stator-field-oriented vector control for VSI-fed wound-excited synchronous motor", in *Proc. International Aegean Conference on Electrical Machines and Power Electronics, ACEMP-ELECTROMOTION Joint Conference, Bodrum, Turkey*, 2007, pp. 303-308.
[15] Wallfaren, W, "Method and apparatus for determining angular velocity from two signals which are function of the angle of rotation", US Patent No.4814701, Mar. 21 1989.
[16] Bose K. B., "Modern Power Electronics and AC Drives", Prentice-Hall PTR. Prentice-Hall Inc., Englewood Cliffs, New Jersey, USA, 2002.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 27-39

Incremental Encoder Based Position and Speed Identification: Modeling and Simulation

Ioan Iov INCZE, Alin NEGREA, Maria IMECS, Csaba SZABÓ

Department of Electrical Drives and Robots, Faculty of Electrical Engineering,
Technical University of Cluj-Napoca,
e-mail: ioan.incze@edr.utcluj.ro, Cornel.NEGREA@edr.utcluj.ro,
imecs@edr.utcluj.ro, csaba.szabo@edr.utcluj.ro

Manuscript received Oct 1, 2010; revised Oct 15, 2010.

Abstract: Electrical drives frequently use incremental encoders as position sensor. The paper deals with the modeling and simulation of an incremental encoder and associated units for processing the information provided by the encoder. A mathematical model of the incremental encoder is presented. Based on encoder signals the direction of the rotation, the position and the speed are identified. The described procedure for determination of the direction of the rotation is able to identify the direction changing in all cases during a rotation equal to the minimal detectable rotation-angle-increment. The computing of the position is based on the algebraic summing of the number of the generated encoder signals. For the speed determination different methods are modeled and simulated: for high speed region the frequency measurement is used and for low speed domain the period measurement is appropriate. In case of a large speed variation the minimal-error-based switching between the two methods is suitable. Matlab-Simulink® simulation structures were realized for the encoder signals based on the identification of the direction of the rotation, for the position and speed computation. Experimental results performed on a DSP-based set-up (under development) are given. The presented simulation subsystems of the encoder, position and speed computation may be integrated in any Matlab-Simulink® structure.

Keywords: Angle transducer, position sensor, incremental encoder modeling, incremental encoder simulation, angular speed identification, electrical drive.

1. Introduction

The incremental encoder is a device which provides electrical pulses if its shaft is rotating [1], [2], [4]. The number of the generated pulses is proportional

to the angular position of the shaft. The incremental encoder is one of the most frequently used position transducers. The principle of an optical incremental encoder is presented in *Fig. 1*. Together with the shaft there is rotating a transparent (usually glass) rotor disc with a circular graduation-track realized as a periodic sequence of transparent and non-transparent radial zones which modulates the light beams emitted by a light source placed on one side of the disc on the fix part (stator) of the encoder. On the opposite side the modulated light beams are sensed by two groups of optical sensors and processed by electronic circuits. Each of the two outputs of the encoder (noted *A* and *B*) will generate one pulse when the shaft rotates an angle equal to the angular step of graduation θ_p, i.e. the angle according to one successive transparent and non-transparent zone. The number of pulses (counted usually by external electronic counters) is proportional to the angular position of the shaft. Due to the fact that the light beams are placed shifted to each other with an angle equal to the quarter of angular step of graduation $\theta_p/4$, the pulses of the two outputs will be also shifted, making possible the determination of the rotation sense. A third light beam is modulated by another track with a single graduation. The output signal (named *Z*) associated to this third beam provides a single pulse in the course of a complete (360°) rotation. The shaft position corresponding to this pulse may be considered as reference position. *Fig. 2* shows the output pulses of the encoder.

Usually for counter-clockwise (CCW) direction θ is considered as positive, and for clockwise (CW) direction it is considered negative.

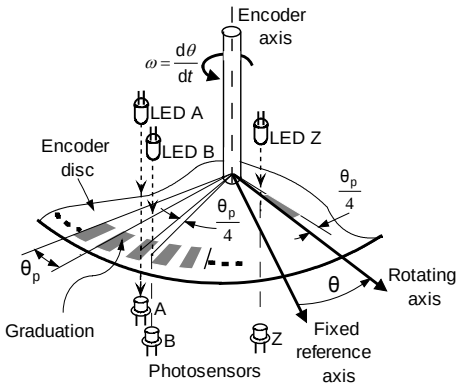


Figure 1: Construction principle of the incremental encoder: the gray surfaces are optically transparent.

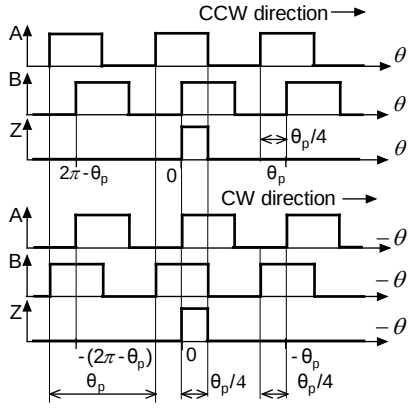


Figure 2: Diagram of the output signals for counter-clockwise (CCW) and clockwise (CW) rotation.

2. Incremental encoder modeling

The input signal of the incremental encoder is the angular position θ of its shaft with respect to the fixed reference axis. The output signals are the two pulses shifted by a quarter angular step $A(\theta)$ and $B(\theta)$, respectively the marker signal $Z(\theta)$. If θ_p is the angular step of the encoder, the outputs may be described by the following equations [2]:

$$\begin{aligned} A(\theta) &= \begin{cases} 1 & \text{if} \quad 0 \le \left(\theta \bmod \theta_p\right) \le \theta_p/2; \\ 0 & \text{if} \quad \theta_p/2 < \left(\theta \bmod \theta_p\right) \le \theta_p; \end{cases} \\ B(\theta) &= \begin{cases} 1 & \text{if} \quad 0 \le \left(\left(\theta - \theta_p/4\right) \bmod \theta_p\right) \le \theta_p/2; \\ 0 & \text{if} \quad \theta_p/2 < \left(\left(\theta - \theta_p/4\right) \bmod \theta_p\right) \le \theta_p; \end{cases} \\ Z(\theta) &= \begin{cases} 1 & \text{if} \quad \theta \bmod (2\pi) = 0; \\ 0 & \text{if} \quad \theta \bmod (2\pi) \ne 0. \end{cases} \end{aligned} \tag{1}$$

During a rotation angle of the shaft, equal to the angular step of graduation θ_p, there are four switching events in the output pulses; therefore the minimal rotation-angle-increment detectable by the encoder is $\theta_p/4$ [3]. The number of pulses, generated by the encoder in the course of a rotation, is equal with the number of angular steps of the graduation on the circular track on the rotor.

$$N_r = \frac{2\pi}{\theta_p} \tag{2}$$

Based on (1) a Matlab/Simulink® a simulation structure shown in *Fig. 3* was built. The outputs A, B and Z are computed by Simulink® function blocks. θp is defined by a constant block. The structure is saved as a subsystem. The simulation structure of the incremental encoder may be integrated in any other Simulink® structure.

3. Encoder based position identification

In an incremental-encoder-based system the angular position θ is measured with respect to a fixed reference axis (θ=0 rad.) and it is obtained by algebraic counting of the number of the generated encoder pulses according to the CCW (ΣN_i pulses) and CW direction (ΣN_j pulses) and multiplying it with the angular step θ_p of the encoder [2]. Mathematically:

$$\theta = \theta_p \left(\sum_i^{CCW} N_i - \sum_j^{CW} N_j \right) = \theta_p N \ . \tag{3}$$

In order to compute the algebraic number of pulses it is necessary to know the direction of the rotation.

A. Identification of the rotation direction

The two $\theta_p/4$ shifted output signals of the encoder contain implicitly also the direction information which may be obtained in different ways.

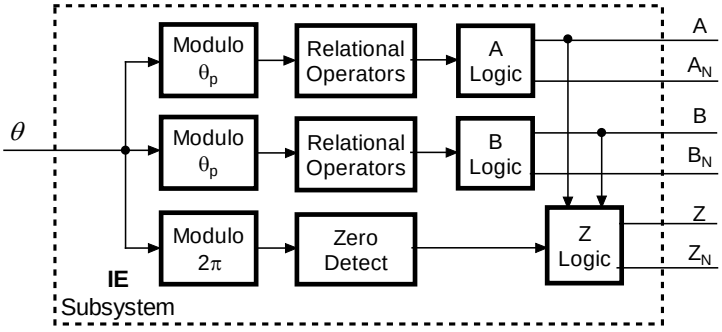


Figure 3: Simulation structure of the incremental encoder.
(Note: The subscript "N" denotes the negated logical variable.)

Taking into account the all four possible combinations of *A* and *B* signals for the reversals, it is possible to detect the direction changing in all cases during a rotation of the minimal detectable rotation-angle-increment $\theta_p/4$. *Table 1* shows the all combinations of signals which detect the reversal of the rotation sense.

Table 1. Combination of signals which detect the reversal of rotation.

From CCW to CW Q=1 to Q=0 (Triggered by R)		From CW to CCW Q=1 to Q=0 (Triggered by S)	
Occurs if		Occurs if	
A	*B*	*A*	*B*
0	0→1	0→1	0
0→1	1	0	1→0
1→0	0	1	0→1
1	1→0	1→0	1

Note: The 0→1 denote the raising-edge and the 1→0 the falling-edge of associated logic variable.

A trivial solution of the problem may be sampling at every rising edge of the *B* output pulses of the *A* output logic value. The resulted logic value will be 1 for counterclockwise (CCW) direction of rotation, and 0 for the clockwise (CW) direction. The method detects the direction changing only after a time interval according to a rotation of $3\,\theta_p/4 - 5\,\theta_p/4$.

B. The position-identification structure

Based on (3) and the above presented direction identification method, a Simulink® subsystem was built for position computation. Its structure is presented in *Fig. 4*.

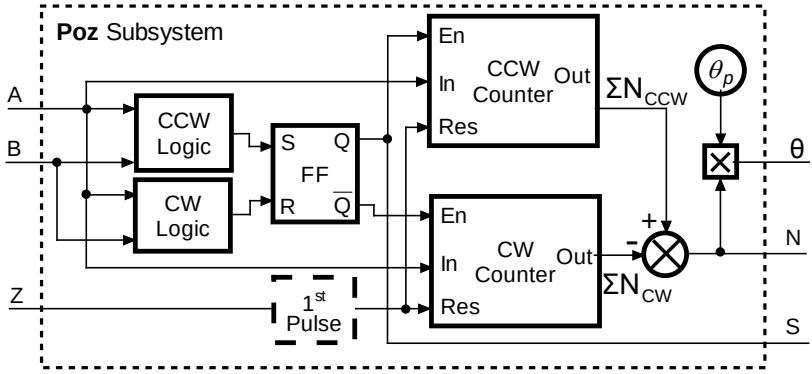


Figure 4: The simulation structure of the position computing subsystem.

The left side of the structure identifies the direction of the rotation. The direction signal S enables the appropriate (CCW or CW) counter. The structure has to be provided with the "1st Pulse" block (broken line in *Fig. 4*) in order to extend the position measurement to more rotations.

4. Encoder based speed identification

The signals generated by incremental encoder provide also information regarding the speed of rotation. A few basic methods are known which are discussed below.

A. Speed identification based on frequency measurement

The frequency of the encoder pulses is proportional to the angular speed. The number of pulses ΔN is counted during a known fix sampling period Ts (see *Fig. 5*) [5]. The angular speed is determined by the expression:

$$\omega = \frac{\mathrm{d}\theta}{\mathrm{d}t} \cong \frac{\mathrm{d}\theta}{T_s} \cong \frac{2\pi\Delta N}{N_r T_s} = \frac{\theta_p \Delta N}{T_s} \tag{4}$$

Due to the lack of synchronization between the sampling period and encoder pulses a quantization error occurs. The relative error of the procedure is given by

$$\varepsilon_f = \frac{1}{\omega}\frac{2\pi}{N_r T_s} \tag{5}$$

and depends on the reciprocal of the speed, the measuring interval and the resolution of the encoder [5].

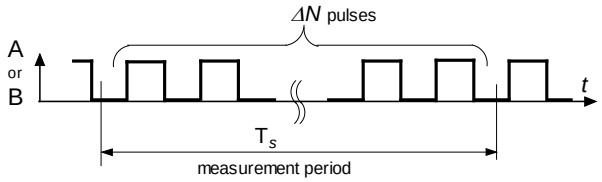


Figure 5: The principle of the speed identification based on the frequency measurement.

The speed calculation structure based on frequency measurement is presented in *Fig. 6*. In order to enhance the precision, the "Logic x4" block multiplies by 4 the frequency of the encoder signals. Two, alternatively resetted and enabled counters count the number of pulses. The content of the just disabled counter is used for speed computation.

The speed identification based on frequency measurement produces relatively small errors at high speed because the number of pulses from the encoder in the measurement-time interval is high.

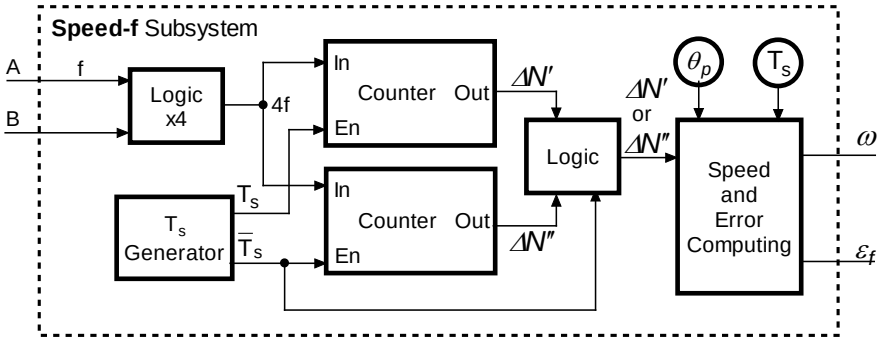


Figure 6: The simulation structure of the speed computing subsystem based on frequency measurement.

B. Speed identification based on period measurement

The speed identification based on frequency measurement at low speed is no longer an option. A better solution is the speed identification based on period measurement. The method consists of counting the pulses of a high frequency clock signal (having period Thf) during an encoder period, as is shown in *Fig. 7* [5].

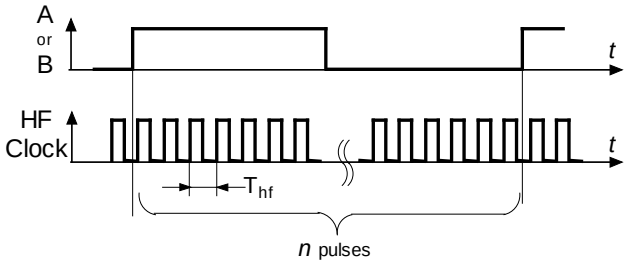


Figure 7: The principle of the speed identification based on the period measurement.

In this case the expression of the angular speed is

$$\omega = \frac{\mathrm{d}\theta}{\mathrm{d}t} \cong \frac{\mathrm{d}\theta}{nT_{hf}} \cong \frac{2\pi}{N_r n T_{hf}} \tag{6}$$

where n represents the counted number of the high frequency pulses. The relative error is increasing with the rotation frequency and is given by [5]

$$\varepsilon_p = \frac{N_r \omega T_{hf}}{2\pi} \tag{7}$$

The speed calculation structure based on period measurement is presented in *Fig. 8*.

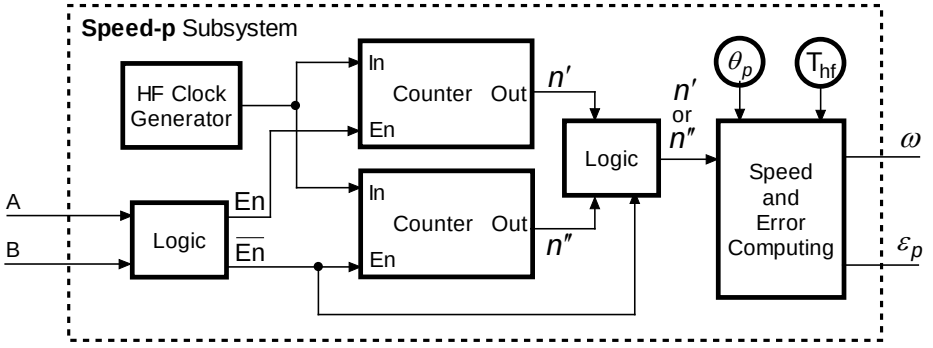


Figure 8: The simulation structure of the speed computing subsystem based on the period measurement.

The structure resembles the previous one; the main difference is that the high frequency pulses are counted during an encoder signal period.

C. Combined method for speed identification

In order to minimize the speed identification error it is suitable to use in low speed region the period measurement method $u(t)$ and as the speed is increasing to switch to the frequency measurement one. The moment of switching is given by the equality of the errors ($\varepsilon_p = \varepsilon_f$), that happens when the speed riches the value

$$\omega_s = \frac{2\pi}{N_r} \cdot \frac{1}{\sqrt{T_s \cdot T_{hf}}} \tag{8}$$

The subsystem presented in *Fig. 9* computes the speed based on above described method.

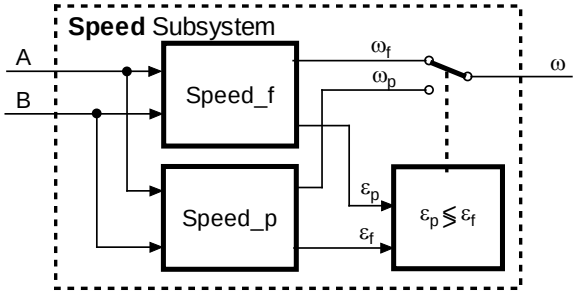


Figure 9: The simulation structure of the combined speed computing subsystem.

The compensation of the errors caused by the sampling times may enhance the precision of the measurement [6].

5. Simulation results

The structure of the interconnected functional units for simulation of position computing is shown in *Fig. 10*. The reference angular position θ_{ref} (the input signal of the encoder block "IE") is generated by a user programmable "Function generator" block. The encoder generates the A, B and Z signals. Based on these, the block "Poz" computes the position θ and the block "Speed" provide the computed angular speed. In order to test the structure, the function generator was programmed in order to start the simulation generating a positive ramp-reference angle, which is the input signal for the "IE" encoder block. At 0.2 s the ramp is switched to negative (equivalent to a reversal from CCW to CW), decreasing in time until 0.8 s, when it is again switched to positive.

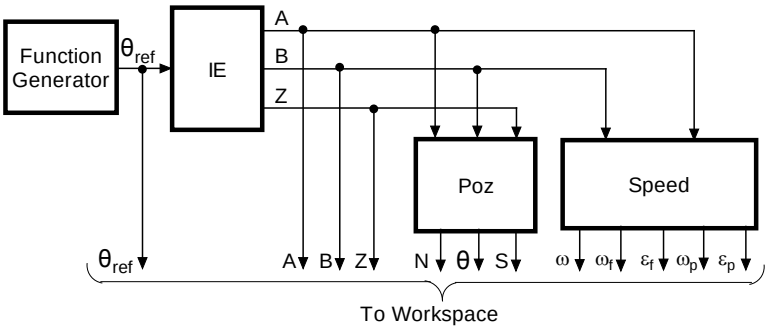
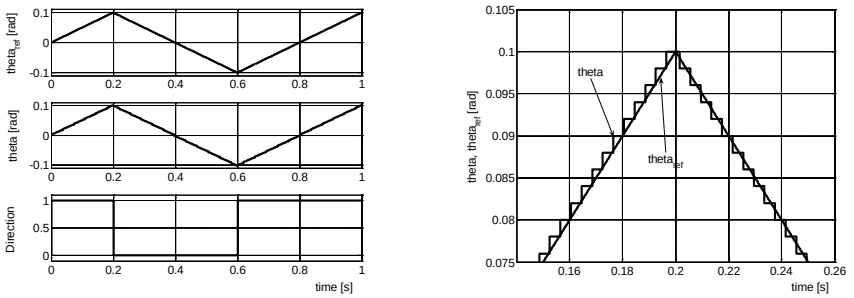


Figure 10: The structure of the interconnected functional units for simulation of the position and speed computation:
IE – incremental encoder, Poz – position computing block,
Speed – speed computing block.

The time profile of the generated reference angle is presented in *Fig. 11 a)* (top trace). The block "Poz", using the encoder output signals, determines the direction of the rotation (in *Fig. 11 a*) bottom trace) and computes the position (shown in *Fig. 11 a)* middle trace). The computed position follows very well the reference one. *Fig. 11 b)* presents an enlarged detail of superposed reference and computed angle before and after the reversal at 0.2 s. The incremental character of the computed position is evident.



a) b)

Figure 11: The simulation results representing the reference angle and computed angle during reversal.
a) From top to bottom: reference angle θ_{ref}, computed angle θ, direction signal S.
b) Detail of the reference angle θ_{ref} and computed angle θ versus time.

There was analyzed the reversal process. The simulated results are presented in *Fig. 12. a)–d)*. The parameters of the function generator were selected in such a manner, that all possible combinations of signals A and B at reversal (presented in *Table 1*) were captured. In all cases the sensing of the reversal is done in a quarter of angular step, as is shown in *Fig. 12.*

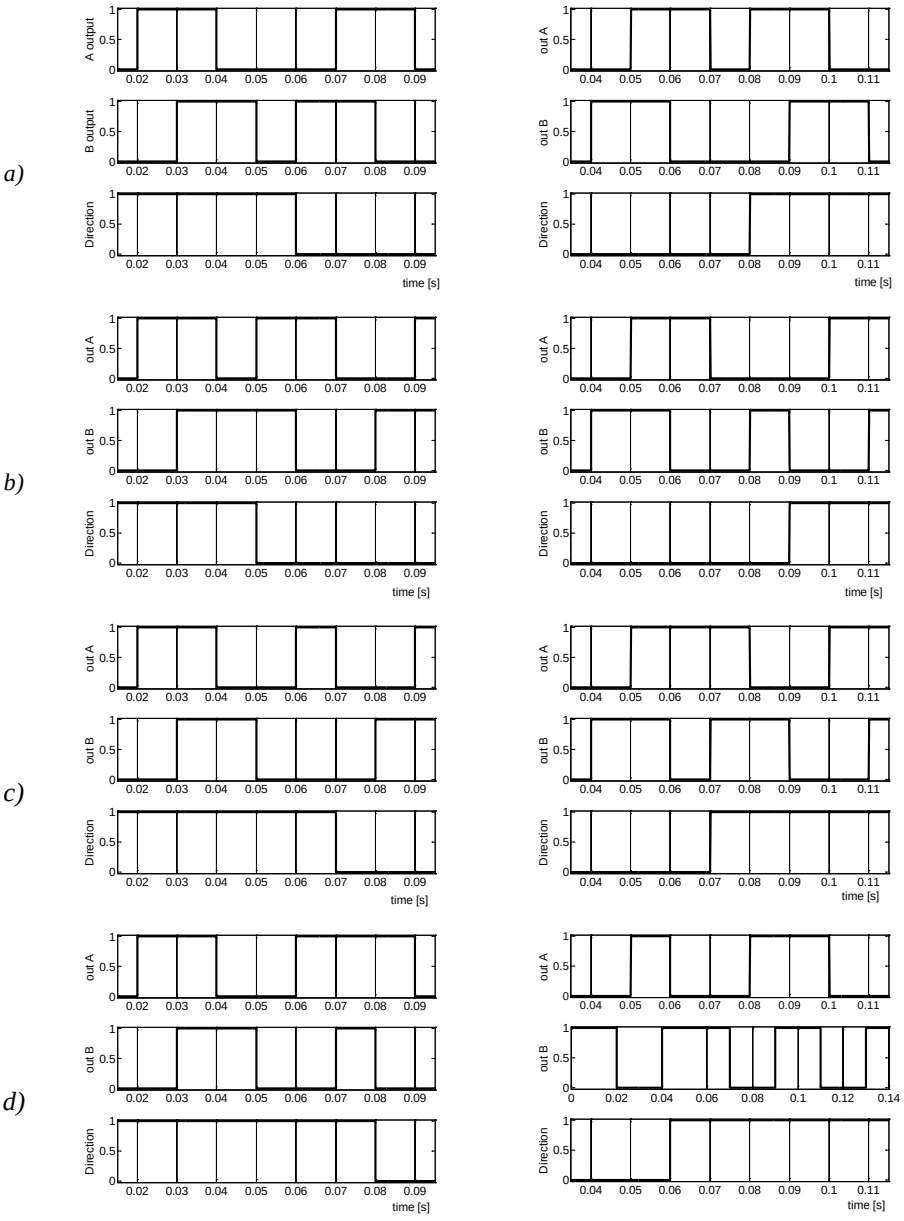


Figure 12: The simulation results showing all combinations of the reversal process. Left column: Reversal from CCW to CW, Right column: Reversal from CW to CCW, top trace: output A, middle trace: output B, bottom trace: direction signal. Reversal occurs at: a) A=0, B=0; b) A=0, B=1; c) A=1, B=0; d) A=1, B=1.

Fig. 13 presents the A, B and Z signals at crossing the reference position ($\theta = 0$) in CW and CCW direction.

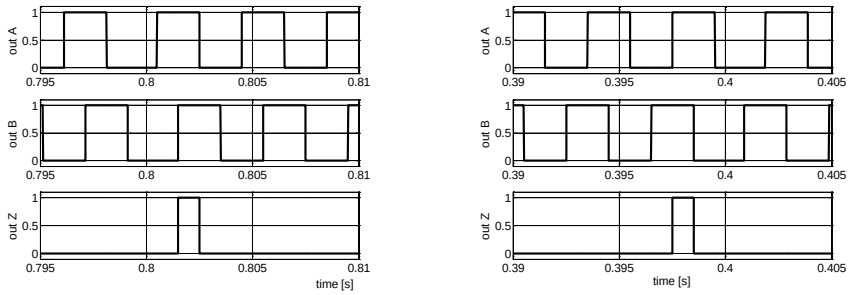


Figure 13: The A, B and Z signals of the encoder at crossing the reference position: a) in CCW direction, b) in CW direction.

In order to test the speed identification, the function generator was programmed for a linearly increasing and decreasing speed profile. *Fig. 14* shows the theoretical speed profile and the computed speed. In *Fig. 15* is presented the variation of the errors ε_f and ε_p. The switching between the two methods occurs at moment 0.2 s and 0.8 s, respectively.

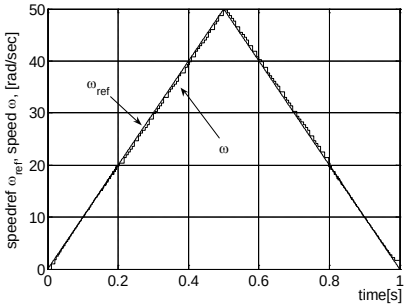


Figure 14: The simulation results showing reference and calculated speed.

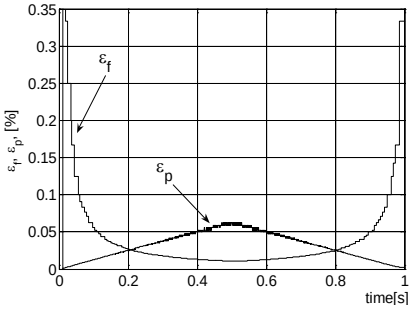


Figure 15: Variation of the error versus time of the two speed calculation methods.

The structure presented in *Fig. 10* may be integrated in the simulation structures of electrical drives [2], [5]. In this case the input signal of the encoder – i.e. the angular position – will be provided by the mathematical model of the electrical machine. The computed position and speed is used as position feed-back signal by the control system of the drive.

The conditions used in simulations are: N=500, T_{hf} = 4 μs, T_s = 6 ms, the simulation step was taken 1 μs.

6. Experimental results

In order to investigate different position and speed determination algorithms an experimental set-up is under construction (see *Fig. 16*). The incremental encoder (type 1XP8001-1) is mounted on the shaft of an induction motor driven by a static frequency converter. (Micromaster, *Siemens*). The encoder signals are processed by an experimental board built around a DSP based development board from Spectrum Digital.

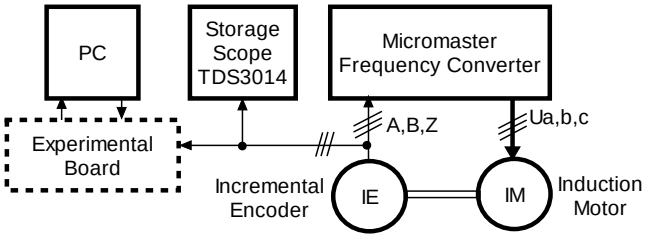
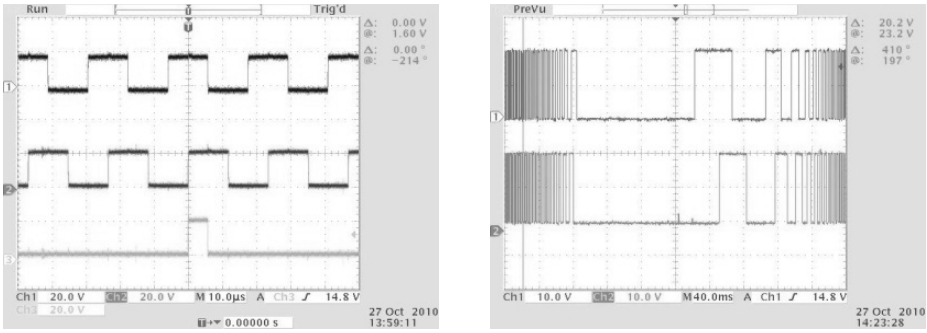


Figure 16: Block scheme of the experimental rig.

Fig. 17 a) shows the captured encoder signals for CCW rotation, and *Fig. 17 b)* represents the A and B signals during the CW to CCW reversal process. The reversal occurs for A=0 B=0.



a) b)

Figure 17: Captured encoder output signals
a) for CCW direction versus time;
Top: Signal A, Middle: Signal B, Bottom: Marker signal Z;
b) for direction reversal from CW to CCW direction versus time;
Top: Signal A, Bottom: Signal B.

A comparison of the above figures to *Fig. 13 a)* and *Fig. 12 a)* shows that the captures are very closed to the simulated results.

7. Conclusion

The information provided by the incremental encoders is inherently digital.

The angular position of the encoder shaft is obtained by algebraic summing of the number of pulses provided by the encoder according to CCW and CW rotation.

The direction of the rotation may be determined by a digital decoding scheme using the two quadrature signals. The direction changes are detected in an angular interval equal to a quarter of the angular step of the graduation.

The frequency of the pulses generated by the encoder is proportional to the speed of the rotation. The error of the measurement is inversely proportional to the speed, therefore the procedure is appropriate for high speed region.

At low speeds the measurement of the period of the encoder pulses is recommended. The measurement error is decreasing with the decreasing of the speed.

In case of large speed variations – in order to minimize the errors – a switching between the two described methods is suitable.

The presented simulation structure of the incremental encoder, position and speed computation may be integrated in any Matlab-Simulink® structure.

References

[1] Incze, J. J., Szabó, Cs., and Imecs, M., "Modeling and simulation of an incremental encoder used in electrical drives", in *Proc. of 10th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics CINTI 2009, Budapest, Hungary*, 2009, pp. 97-109.

[2] Incze, I. I., Szabó, Cs., and Imecs, M., "Incremental encoder in electrical drives: modeling and simulation" in *Studies in Computational Intelligence* Editors: I. J. Rudas, J. Fodor, J. Kacprzyk, Springer Verlag, Germany, under press.

[3] Koci, P., and Tuma, J., "Incremental rotary encoders accuracy", in *Proc. of International Carpathian Control Conference ICCC 2006, Roznov pod Rashosten, Czech Republic*, 2006, pp. 257-260.

[4] Lehoczky, J., Márkus, M., and Mucsi, S., „Szervorendszerek, követő szabályozások", Műszaki Kiadó, Budapest, Hungary, 1977.

[5] Petrella, R., Tursini, M., Peretti, L., and Zigliotto, M., "Speed measurement algorithms for low resolution incremental encoder equipped drives: comparative analysis", in *Proc. of International Aegean Conference on Electrical Machines and Power Electronics, ACEMP-ELECTROMOTION Joint Conference, Bodrum, Turkey*, 2007, pp. 780-787.

[6] Miyashita, I., and Ohmori, Y., "A new speed observer for an induction motor using the speed estimation technique", in *Proc. of European Power Electronics Conference EPE´93, Brighton, United Kingdom*, 1993, vol. 5, pp. 349-353.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 40-50

Aluminium Electrolytic Capacitor Research and Development Time Optimization Based on a Measurement Automation System

Dénes FODOR

Pannon University, Veszprém, Institute of Mechanical Engineering, Department of Applied Mechanics, Automotive System Engineering Group, Egyetem st. 10.; 8200 Veszprém, Hungary, e-mail: fodor@almos.uni-pannon.hu

Manuscript received October 20, 2010; revised November 7, 2010.

Abstract: The aim of this paper is to present partly the Measurement Automation System (MAS) of an Electrolytic Capacitor Development Laboratory at EPCOS Hungary. The main role of the MAS is to facilitate the electrolyte and capacitor research and development, through the automation of the related measurement tasks, and to provide a powerful database system background for data retrieval and research decision support. The paper introduces only a few applications of the entire system. More than 27 different electrolyte and capacitor measurements were automated. All the measurements have been implemented in a similar manner. During the process the user initializes the measurement, sets the measurement environmental parameters, and launches the execution. The program runs on its own, sending automatically the results of the measurement to a database system, from where the data can be retrieved in a predefined or a non-predefined way. For the realization of the above requirements the National Instruments measurement, data acquisition and LabVIEW software development tools were chosen as implementation and development platform. After validation of the system, there are many advantages as making the measurements more precise, more reliable, fault tolerant, parallel running, which all contribute to speed up the research and development of new components and devices.

The developed measurement system controls and harmonizes the different devices and supervises their work. The developers do not have to encroach. The user can simply check the measurement phase by a glance on the screen. The programs estimate and display the running time of the experiments, allowing for the researchers working on the laboratory to manage the instrumental resources in time and to schedule in advance (for hours, weeks and months) the new measurements. Another big advantage is the database system behind, which stores the result of each measurement in an easy searchable way. Different measurement reports and statistical diagrams can be made automatically and the results can be reused in later research. The effectiveness of the

system was tested also via the inner gas pressure measurement of electrolytic capacitors to estimate the life-span of the capacitors. According to the experiments the introduction of the MAS system in the Lab the research and development time for new electrolytes and capacitors has been decreased considerably.

Keywords: Measurement automation, test automation, aluminium electrolytic capacitors, data acquisition, data mining.

1. Introduction

Capacitors play a very important role in our world [1], [2]. They can be found in every electronic device around us, they are widespread all over the world used as energy storage elements, filters and decouplers.

The main features of capacitors are: capacity (1pF-1F), operational voltage (from 1,5 V up to some kV), operational temperature (from -55 °C to 125 °C), loss factor, size and shape.

The most frequently used capacitor types in the industry nowadays are: ceramic, foil, aluminium and tantalum capacitors. The four most important application fields for capacitor technologies are radio techniques, electrical power processing, energy storage and power electronics. Except for the first application field electrolyte-capacitors can be used, so this type of capacitor is prevalent.

The main advantage of the electrolytic capacitors is the high capacity and voltage value, which can be attributed to the dielectric layer with very small thickness, but with very large surface. Their disadvantage is the over voltage sensitivity.

The main characteristics of the electrolytic capacitors are determined by the electrolyte, the anode foil and the paper separator.

The electrolyte generally consists of the following components:

- solvent: e.g.: ethylene glycol,
- acids and bases: usually organic,
- different additives.

The electrolytes are characterized by two major features: conductivity and breakdown potential, both of them dependent on the temperature. The change of conductivity as a function of temperature decisively affects the electric parameters of the capacitor. The chemical reactions, which take place inside the electrolyte, are in direct relationship with the conductivity value at different temperatures and the quantification of this relationship is important.

The conductivity and breakdown potential of the electrolyte influences the maximum operating condition of the capacitors. Electrolytes with high

conductivity are used in the low voltage capacitors, while the electrolytes with low conductivity are used in the high voltage capacitors.

The paper is organized as follows. The short descriptions of the measurement types, which must be automated, are given in Section 2. The architecture of the proposed measurement system is presented in Section 3. In Section 4 a characteristic measurement "*Conductivity* (*T*)" is presented in detail in order to demonstrate the program structure and some implementation issues. Some characteristic measurement results are given in Section 5, and the conclusions are presented in Section 6.

2. Measurement types

There are two groups of measurements used in electrolyte-capacitor research and development. The first main measurement group is related to electrolytes, while the second main measurement group is related to capacitors. The electrolyte measurement consists of six measurement programs as follows:

- "*Conductivity* (*T*)": measurement on the temperature dependence of conductivity. This is one of the most important measurements, and is presented in details in the next section.
- "*Ph* (*T*)": measurement of the pH value as a function of temperature. The structure of the program is completely similar with the above-mentioned one, with a difference that pH meter is used instead of conductivity meter. The measurement is important because the pH value of the used electrolyte in the electrolytic-capacitor must be in a specified range.
- "*Mixing* (*pH with single temperature*)": measurement of the pH value as a function of the concentration of an electrolyte composition at a specified temperature. As a matter of fact we use this measurement in order to set up the pH value of the electrolyte.
- "*Mixing* (*conductivity with single temperature*)": measurement of the conductivity as a function of the concentration of an electrolyte composition at a specified temperature.
- "*Mixing* (*conductivity with multi temperature*)": measurement of the conductivity as a function of the concentration of an electrolyte composition at several temperatures. This and the former measurement are used to set up the conductivity value of the electrolyte.
- "*Spark detector*": measurement of the breakdown potential of the electrolyte.

The main capacitor measurements are as follows:

- "*Spark detector*": this measurement program is similar to the structure of the program used for electrolyte measurement. The only difference is that the object of the measurement is the winding impregnated with the electrolyte. In that case we want to know the breakdown potential of the paper impregnated with different electrolytes.
- "*ESR* (*Equivalent Serial Resistance*)": This measurement is mostly used in order to determine the resistance of the capacitor at different frequencies and temperatures.
- "*Gas pressure*": measurement of the internal gas pressure of the capacitor in various operating conditions.

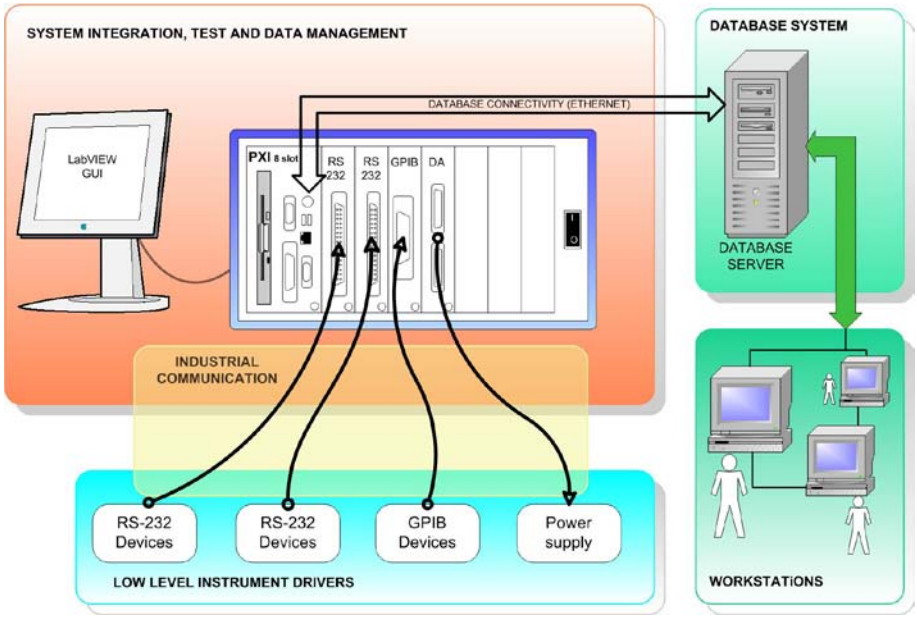


Figure 1: System architecture of the automated measurement system.

3. Architecture of the measurement system

The core of the automated system is the NI PXI-1042 chassis [3], with five modules such as NI PXI-8185 controller [3] (embedded PC), NI PXI-6723 D/A card [3] for controlling the power supplies, NI PXI-8420 and NI PXI-GPIB for communicating with the measurement instruments via RS-232, RS-485 and GPIB [3] (*Fig.1*). This compact form system gives equivalent performance as a PC-based data acquisition and measurement control system but has some added functions, such as trigger buses, increased bus speed, rugged and modular

packaging. The different measuring instruments, which can communicate through RS-232, RS-485 and GPIB have been integrated easily in the system, with the help of the above mentioned communication cards (NI PXI-GPIB, NI PXI-8420). In this way also a common development tool, the LabVIEW (for details see [5]), is available to develop the related communication routines (drivers) and the measurement programs too. The measurement results are migrated into a database from where data can be retrieved in non-predicted and predicted way for evaluation and decision support. (For more information about PXI see [4])

There are five electrolytic measurements out of six whose architecture show the same structure (see: *Fig. 2)*. The object of the measurement, the electrolyte, is located inside a double-jacketed vessel. In the external part of the vessel the water is circulated by the thermostat, while the electrolyte is inside the vessel.

The most important parameters are provided by the pH meter and the conductivity meter. Each instrument has got an electrode that can be used to measure the temperature too. Only one of them is used during an experiment. The experimental setup requires the simultaneous operation of four individual instruments such as Thermostat [6], pH meter [7], [8], Conductivity meter [7], [9], Burette. Controlling and following up the temperature precisely is essential during the measurements so the Thermostat device is connected to the system every time. The Burette is only used when the variation of the pH or conductivity value as a function of a composition must be known. The mixing type measurements give the relation of the conductivity or pH to the concentration of the examined component.

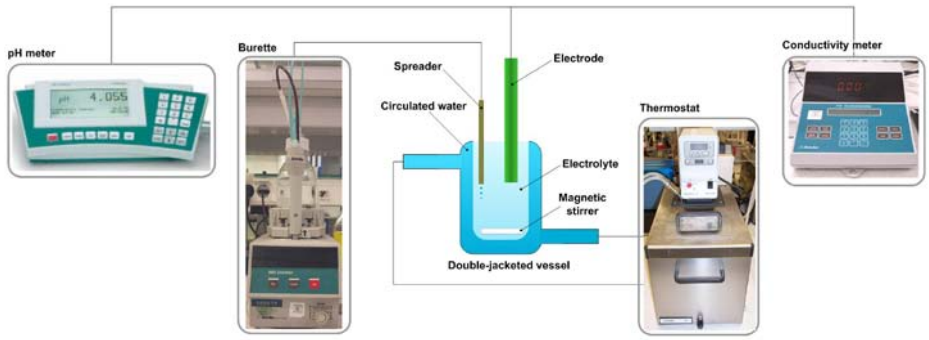


Figure 2: Diagram of the electrolytic measurements.

4. “Conductivity(T)” measurement

The base of the whole software system is a framework which was originally designed to provide a common user interface for the different measurement programs. In spite of the fact that all the measurements have an individual character, they were integrated into the above mentioned framework, in order to manage the communication ports and instruments, and to provide the parallel run of the programs.

The measurement system includes at least 27 different measurements, whose presentation can not be done here. All the measurements have been implemented in a similar manner. Firstly the user initializes the measurement, sets the measurement parameters, launches the execution and leaves the program to run on its own, sending the results of the measurement to a database system. This process is presented afterwards through the “*Conductivity* (*T*)” measurement.

Two instruments are involved in this measurement: the thermostat and the conductivity meter.

Before launching the program the user initializes the measurement. During the measurement the user can choose between two main tab controls ① in *Fig.3*. The “*Set parameters*” tab contains the parameters set of the measurement, while the “*Measurement*” tab shows the state flow of the measurement, graphs and displays. The program executes the same steps cyclically. The state diagram ③ shows the current state of the measurement and the remaining time before the next phase. Firstly the program sets up the temperature. On the temperature graphs ② the temperature of the measuring probe and of the thermostat can be followed up. During the *stabilization time* the temperature of the electrolyte becomes the same as the thermostat’s. Through *measuring & saving* phase the important conductivity values ⑤ are measured and stored locally. *The remaining time before the next measurement* is indicated with a progress bar ④. After the *measuring & saving* phase the program calculates the mean value from the stored conductivity data and only the result is migrated into the database. On the ⑦ and ⑧ graphs the conductivity and the temperature as a function of the number of measurements can be seen. These are the most important graphs, because the electrolyte forming can be directly correlated with the temperature. The measurement is ended after the conductivity is measured at all of the settled temperatures. The remaining time before the end of the experiment is shown during the measurements ⑥. The program can be stopped with the *STOP* button ⑨.

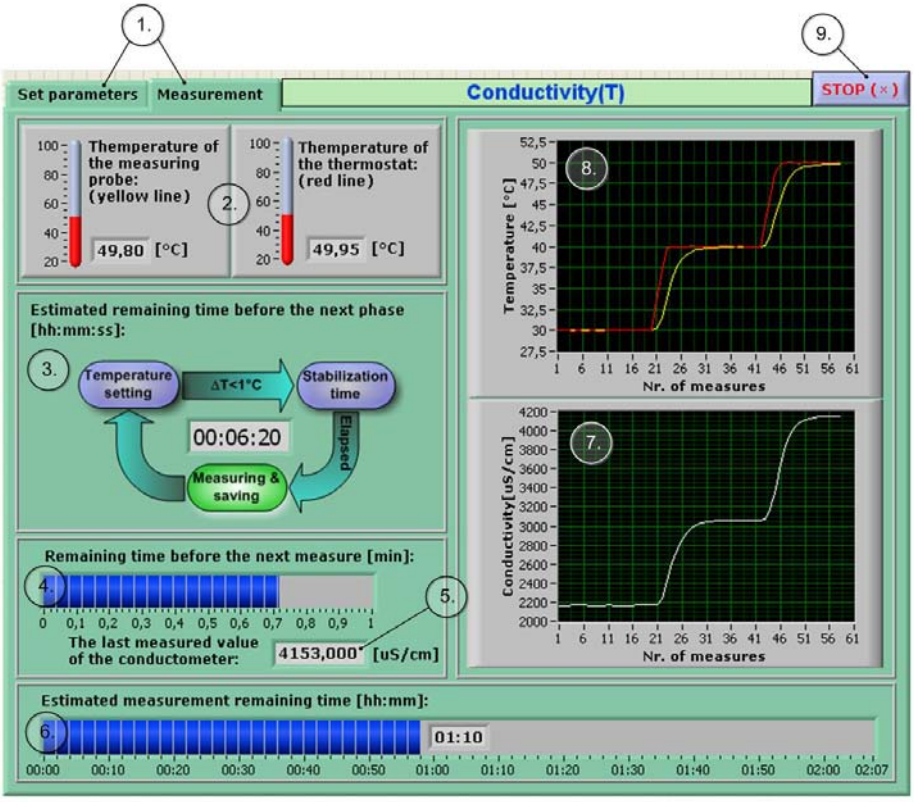


Figure 3: Graphic User Interface (GUI) of the "Conductivity(T)" measurement program.

5. Results

In *Fig. 4* it is demonstrated how decisions are supported by the automated measurement system. Each dot represents a separate "*Conductivity* (*T*)" and "*Sparkling voltage*" measurement at 85 degree Celsius- result obtained by the automated system. By graphical evaluation electrolytes with specific conductivity or sparkling voltage can be selected for further research.

With the help of the currently automated system the forming of an electrolyte as a function of the temperature can be followed up. In *Fig. 5* measurement results of Electrolyte 1 and Electrolyte 2, which has been obtained using the "*Conductivity* (*T*)" measurement program are demonstrated. The conductivity of the electrolyte is measured more than once (usually 10 times) at the specified temperatures. A mathematical mean is calculated from the

measurement values at one temperature, which is indicated by dots in *Fig. 5*. As mentioned, the conductivity of the electrolyte significantly influences the electric parameters of the capacitor. Conductivity between 900 and 3000 μS/cm at 30 degree Celsius is considered low, and conductivity values exceeding 10000μS/cm in similar thermal conditions is considered high. As shown in *Fig. 5*, the conductivity of Electrolyte 1 measured at 30 degree Celsius is 1500 μS/cm and the Electrolyte 2 has 2300 μs/cm at the same temperature which indicates low conductivity in both cases, resulting in high breakdown potential. From this reason both electrolytes can be used in high voltage capacitors.

In *Fig. 6* the results of a capacitor type measurement – named *"ESR"*, mentioned shortly in the second chapter can be seen. The resistance values of a capacitor as a function of frequency at different temperatures are given.

During capacitor development we aspire to obtain as low serial resistance as possible. To achieve this goal a series of experiments have to be done using different types of electrolyte and paper constructions.

As can be seen in the *Fig. 6* the capacitor resistance below zero degrees Celsius is linearly dependent on frequencies up to 1 KHz. For positive temperatures it can be observed that the resistance values have linear variation on frequencies above 1 KHz.

The application areas of the capacitors are based on the frequency dependence of the resistance value in different temperature ranges.

The "*Mixing* (*conductivity with multi temperature*)" measurement is an extended version of the "*Conductivity* (*T*)" measurement, which is completed with one measurement device: a burette. With the "*Mixing*" measurement the conductivity value as a function of a key component concentration of the used electrolyte can be observed on *Fig. 7*. First of all the conductivity of the electrolyte is measured at 25 C°, 40 C°, 60 C° and 85 C° many times (usually 10), without dosing the conductive salt solution. Each dot on *Fig. 7* represents a mathematical mean which is calculated from the measurement values at each temperature. After that a predefined dosage from the conductive salt solution is added to the electrolyte and the measurements of the conductivity at different temperatures are repeated. The measurements are repeated 16 times adding each time a new predefined conductive salt solution dosage to the electrolyte. Finally the "*Mixing*" measurement results in 17 dots at each temperature.

According to the readings the increase of the conductivity of the electrolyte is in direct relationship to the quantity of the conductive salt solution.

The results can be applied in order to set up the conductivity value of an electrolyte.

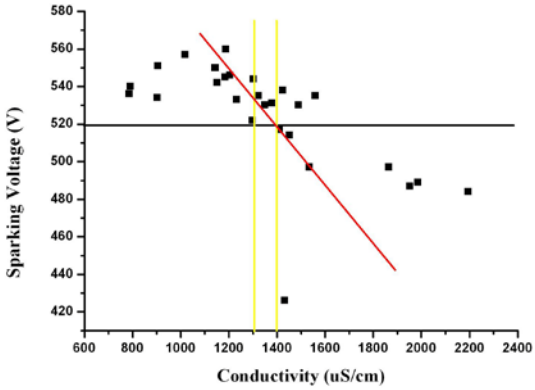


Figure 4: Conductivity vs. Sparkling Voltage of various electrolyte at 85 degrees Celsius.

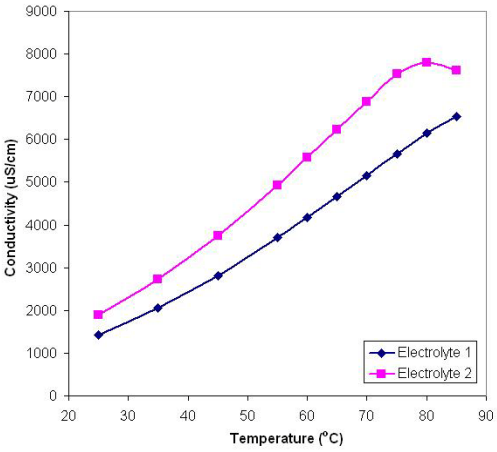


Figure 5: Result of the measurement of conductivity as a function of temperature of Electrolyte 1 and Electrolyte 2.

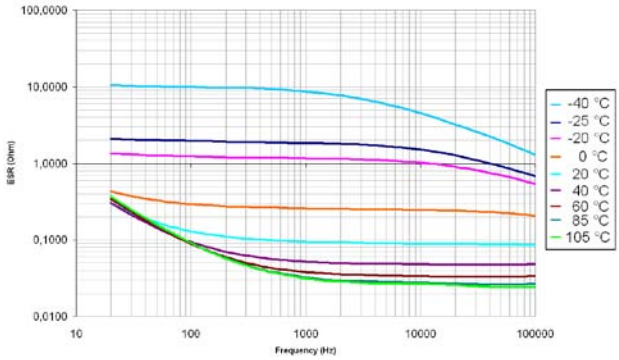


Figure 6: Result of the capacitor resistance (ESR) as a function of frequency at different temperatures.

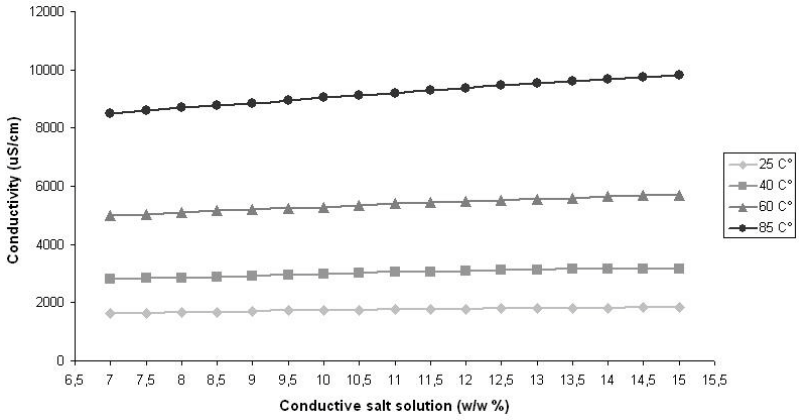


Figure 7: Variation of conductivity of Electrolyte 1 as a function of the conductive salt solution concentration.

6. Conclusions

More than 27 different electrolyte and capacitor measurements, have been automated. All the measurements have been implemented in a similar manner. Firstly the user initializes the measurement, sets the measurement parameters, launches the execution and leaves the program to run on its own, sending the results of the measurement to a database system, from where the data can be retrieved in a predefined or a non-predefined way. After validation of the MAS, there are many advantages as making the measurements more precise and more reliable, fault tolerant (i.e.: monitoring functions are implemented like detection

of missing of line voltage, open gate etc.), running multiple measurements in parallel, which all contribute to speed up the research and development of new components and devices.

The developers do not have to encroach. The developed measurement system controls and harmonizes the different devices and supervises their work. The user can simply check the measurement phase by a glance at the screen. The program estimates and displays the running time of the experiment, allowing for the researchers working on the Lab to manage the instrumental resources in time and to schedule in advance (for hours, weeks and months) the new measurements. Another big advantage is the database system, which stores the result of each measurement in an easy searchable way. Measurement reports and diagrams can be made automatically and the results can be reused in later research.

Acknowledgements

The author fully acknowledges the support of the National Office for Research and Technology via the Agency for Research Fund Management and Research Exploitation (KPI), under the research grants GVOP-3.2.2.-2004-07-0022/3.0 (KKK), GVOP-3.1.1.-2004-05-0029/3.0 (AKF) and more recently for *TAMOP-4.2.1/B-09/1/KONV-2010-0003:* Mobility and Environment: Researches in the fields of motor vehicle industry, energetics and environment in the Middle- and West-Transdanubian Regions of Hungary. The Project is supported by the European Union and co-financed by the European Regional Development Fund.

References

[1] Theisbürger, K. H., “Der Elektrolyt-Kondensator”, FRAKO Kondensatoren- und Apparatbauen G.m.b.H Teningen, third edition, EPCOS internal document.
[2] Per-Olof Fägerholt, “Passive components” (internal document), 1999.
[3] National Instruments, “Product guide”, 2004.
[4] ***, National Instruments Hompage, http://zone.ni.com/devzone/conceptd.nsf/webmain/5D1A4BAB15C82CC986256D3A0058C66C?OpenDocument&node =1525_us.
[5] ***, National Instruments Hompage, http://www.ni.com/labview/whatis/.
[6] ***, “Haake DC30 Circulator- Technical Manual”, Thermo Electron Corporation, 2002.
[7] ***, Metrohm AG Hompage, www.metrohm.com.
[8] ***, “780 pH Meter Instruction Manual”, Metrohm Ltd., Switzerland, 2002.
[9] ***, “712 Conductometer Instruction Manual”, Metrohm Ltd., Switzerland, 2002.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 51-62

Framework for Modeling, Verification and Implementation of Real-Time Applications

Liviu HAȚEGAN [1], Piroska HALLER [2]

[1]Technical University of Cluj-Napoca, Cluj-Napoca, România,
e-mail: liviu.hategan@yahoo.com,
[2]"Petru Maior" University of Tîrgu Mureş, Tîrgu Mureş, România,
e-mail: phaller@upm.ro

Manuscript received October 01, 2010; revised October 30, 2010.

Abstract: This paper proposes a framework for modeling, simulation and formal verification of embedded real-time applications running over a real-time multitasking kernel. We extend a simple real time kernel (RTOS) with synchronous and asynchronous message passing interface to communicate between tasks and drivers. In the same time some embedded system's specific drivers have been added, allowing unified resource access through these interfaces. The process engineer defines the control system as a set of tasks interacting with events occurring irregularly in time (alarms, user commands, communication) and regularly in time (sampled sensor data and actuator control signals). Taking into consideration both non-preemptive and preemptive scheduling, we propose two models consisting of networks of timed automata. Using a model-checker tool (UPPAAL), one can verify the timing and logical properties of an application, changing the time constrains and priorities. In a priority-based scheduling scheme, tasks interact both through the scheduler and through the mutual exclusion mechanism, but there are hidden from the engineer by the framework. The framework also offers a solution for generating the source code skeleton of the modeled application. This reduces the risk of errors do to error-prone human coding and most importantly ensures that the task will have the same behavior as described in the model.

Keywords: Formal verification, timed automata, real-time applications.

1. Introduction

Real-time embedded systems have become widely used in a large number of fields, especially in the industrial environment, playing an increasing role in modern society and are rapidly evolving, growing in complexity. Moreover they are often used not only by themselves, but in clusters and networks. An

embedded real-time system is in close relationship with the physical environment it interacts with, it is involved in monitoring and control of complex physical processes. The applications running on such systems face a set of constrains like memory, processing power, energy consumption, but mainly timing constraints. Consequently, a real-time system must exhibit a predictable behavior, under a given set of conditions the designer must be able to know if the system will meet its requirements. Developing and running embedded applications with predictable and controllable behavior requires a real-time operating system (RTOS). This allows for an application to be constructed as a collection of tasks managed by the RTOS according to the scheduling policy, the timing requirements being mapped as task deadlines.

But the engineers that write embedded software are rarely computer scientists or experts in operating systems. It should be necessary to create an integrated framework for modeling, simulation, verification and code generation. On the other hand timeliness, concurrency, bounded response time, and heterogeneity need to be an integral part of the programming abstractions.

The timed automata formalism is widely used and well-proven in the description and verification of real-time systems. In [1] timed automata are proposed for the description of task arrival patterns. The authors present a unified model for finite control structures, concurrency, synchronization, and tasks with combinations of timing, precedence and resource constraints.

Another work [2] approaches the problem of formal modeling based on timed automata of a multitasking application running under a RTOS. The described model considers an operating system, the application tasks and the behavior of the controlled environment. In this approach the authors also model the internal structure, allowing for the verification of not only task schedulability, but other complex properties like safety and bounded liveness.

In [3] the authors present a framework for modeling and verification of real-time embedded applications running under a multitasking RTOS kernel. The authors propose a model of a minimal operating system defined as a network of timed-automata developed in order to use it as a framework for simulation and verification of mini real-time applications. The authors chose and analyzed the FreeRTOS [4] mini real-time kernel either in a non-preemptive (cooperative) or a preemptive configuration. Access to resources is done using a unified resource access interface.

In this paper we propose a framework for modeling, simulation and formal verification of embedded real-time applications and a solution for automatic source code generation of the modeled application. As in [3], the simulation and verification can be done using the UPPAAL [5] model checker.

We extended the model described in [3] taking into consideration in more detail the internal structure of the tasks [6], adding synchronous and

asynchronous message passing interface to communicate between tasks and drivers, allowing the model to be more expressive. Access to resources is done using a unified resource access interface, fact that simplifies the task structure and facilitates automatic source code generation based on the model. The existing drivers were extended with these interfaces and new drivers were added for the SAM7-EX256 development board specific devices. For the implementation we chose the SAM7-EX256 development board. We improved the unified resource access interface, now consisting of three functions: *Request()*, *Read()* and *Write()*. This allows for a simplified structure of the modeled tasks and is necessary for the purpose of translating the tasks from the model into the corresponding source code.

This paper is organized as follows: section 2 will describe the general properties and behavior for the task templates; in section 3 we describe the properties of the resources and the unified resource access interface; section 4 presents the cooperative model (tasks, cooperative scheduler, resources); the model constructed for the preemptive version of the FreeRTOS kernel is detailed in section 5; section 6 addresses the issue of simulation and formal verification of the applications; in section 7 we describe the process of automatically generating the source code of the application from the model and finally, in section 8 we state our conclusions.

2. The application tasks

Each task instance is modeled by a timed automaton that is synchronized using channels. In general, embedded tasks present the following behavior [1], [3]:

```
   INFINITE_LOOP
- Request access to a resource (blocking call)
- Perform a read/write operation
- Perform a computation
- Request access to another resource (blocking call)
- Perform a read/write operation
   ……
   END_INFINITE_LOOP
```

Resources are accessed by tasks through blocking request calls. The desired resource is explicitly specified through its RID (resource ID) [7]. After making a request call, the task enters in blocking state, where it waits for the resource to become available.

The FreeRTOS tasks are prioritized.

The computations performed by the tasks are characterized by a worst case execution time (WCET) and a best case execution time (BCET). This is a con-

sequence of branching instructions in the computations. Also, the running task will be delayed by the occurrence of interrupts generated by the resources. The ISR execution time will be added to the current task's WCET and BCET [2].

3. The resources

In constructing the general resource model, we considered that the following: resources are reusable and can be shared (but only one task can have access to a resource at any given time), a task can request a single resource at one time and in the request call the resource is explicitly specified through its resource ID. Every resource has a minimal inter-arrival time (the MIAT). A resource can unblock a waiting task and provide data at any time after its MIAT expires.

3.1 The unified resource access interface

In our implementation on the SAM7-EX256 development board, resources provide interrupts that are managed by drivers. Depending on the peripheral, data is read from registers and stored in queues when the interrupt occurs for an input peripheral or data is sent when an output peripheral is ready to accept it. A task that is waiting on the resource's queue will immediately be unblocked.

As we mentioned in the previous sections, access to a particular resource is performed trough the unified access interface. Each resource has a state variable associated. The read and write operations are performed using the state variables (or any other user-defined data structure that corresponds). The interface is composed of the following functions:

- Request (RID)
- Read (RID, var_rid, nr)
- Write (RID, var_rid, nr)

When requesting a resource, it must be explicitly specified through its RID. If the resource is not available (is being held by another task) the calling task will block until the requested resource becomes available. The *Read()* and *Write()* are nonblocking calls, they will be performed after the access to the requested resource is granted and provide the means for read/write operations. Besides these resource RIDs, the function calls must also include the state variables associated with the resource (*var_rid*) and the number of bytes to be read or written. The functions implemented in the unified access module are also present in the model and permit a simple description of concurrent access to the desired resources and read/write operations.

4. The cooperative model

4.1 The cooperative application tasks

Considering the general task form described in the previous section, we present a simple task model that requests access to a resource (RID), performs a read operation and executes a computation. The task in pseudocode is:

```
Task1{
      Loop
      { Request(RID);
        Read(RID, var_rid, nr);
        computation();
      }}
```

The task automaton (*Fig. 1*) is synchronized with the models of the scheduler and resources via channels.

The RUN locations are characterized by a best-case and worst-case execution time (WCET, BCET). This is modeled by the location's invariant ($y<=W$), and the guard on the outgoing transition ($y>=B$). *REQ_BCET[RID]*, *REQ_WCET[RID]*, *READ_BCET[RID]* and *READ_WCET[RID]* are constants defined in the model and they are the BCET and WCET for requesting and reading data in the case of the resource represented by *RID*. While being in a running state, a task can be interrupted by a resource's interrupt service routine (ISR). The execution time of the ISR is added to the *W* and *B* variables.

BLOCKED: the state is entered when the task requests a particular resource through the *request[pid][RID]!* channel. The state is left when the resource becomes available (the *event_or_timer[RID]?* channel is activated).

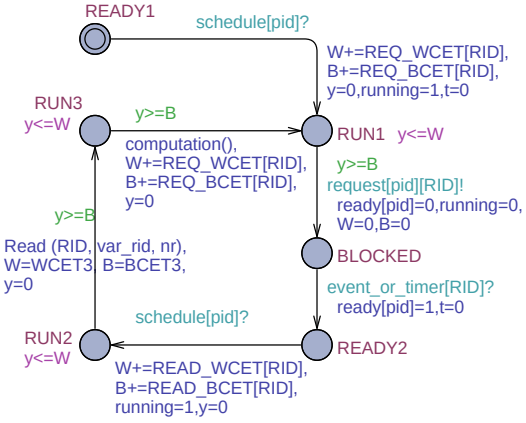


Figure 1: Simple cooperative task model.

If there is no task running or ready, the FreeRTOS kernel schedules the Idle Task (*Fig. 2*), which is always available for scheduling. Following the general task form, the Idle Task periodically requests the *NULL* resource, yielding processor control.

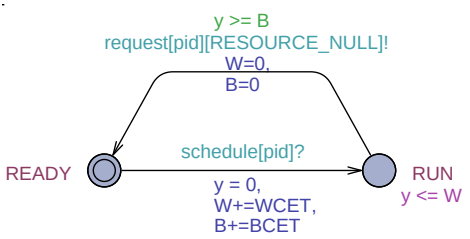


Figure 2: The Idle Task model.

4.2 The cooperative scheduler

In the cooperative behavior the running task has full control of the processor, regardless of its priority, until it makes a blocking call. The task can explicitly invoke the scheduler by calling the *taskYIELD()* macro or by requesting access to a resource (the *Request(RID)* function).

The cooperative scheduler is described by a timed automaton that presents three states: INIT, SELECT and IDLE (*Fig. 3*).

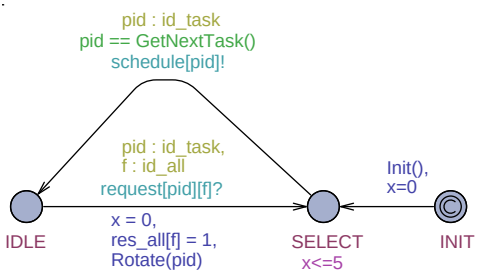


Figure 3: The cooperative scheduler model.

INIT: the necessary hardware settings and initialization of task priorities and data structures take place. Because the state is committed, the scheduler will leave this state immediately at startup.

SELECT: the ready task with the highest priority is chosen for scheduling (the *GetNextTask()* function). The invariant *x<=5* specifies the time needed to select the next task; the value can be changed to match the actual physical time, which is hardware-dependant.

IDLE: the previously chosen task is running. In order to have an equal chance at scheduling for the tasks with the same priority, the *Rotate()* function is called when the scheduler exits from the IDLE state.

4.3 The system resources

Fig. 4 illustrates the general resource model.

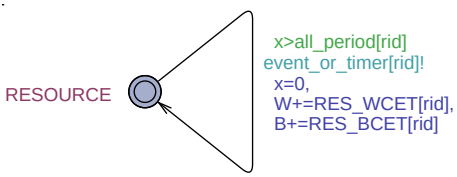


Figure 4: Resource model.

The MIAT values for all of the system's resources are stored in the array *all_period[NR_RESOURCES]*. The MIAT is modeled by the guard *x>all_period[rid]*. The waiting task is unblocked via the *event_or_timer[rid]!* channel. The *RES_WCET[pid]* and *RES_BCET[pid]* constants are used to delay the task interrupted by the resource ISR.

In case a task must execute an action periodically, at strict interval, it can utilize the timer resource (*Fig. 5*). The timer unblocks a waiting task when the *x* clock has the same value as *all_period[tid]*. The constant *tid* represents the timer's RID.

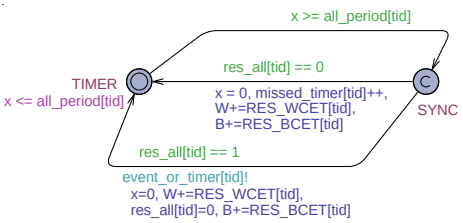


Figure 5: Timer model.

The initial state TIMER is left when the predefined period expires. The SYNC state is committed so it is left immediately, the automaton unblocking any waiting task. In order to avoid system deadlock, the timer is allowed to expire even if none of the tasks are blocked waiting for it.

The resource model is identical for both preemptive and cooperative systems.

5. The preemptive model

5.1 The preemptive application tasks

In addition to the READY, RUN and BLOCKED states presented by the cooperative tasks, the preemptive versions also have a SUSPENDED state. This state is entered when the task is preempted (via the *suspend[pid]?* channel). The invariant *y'==0* will stop the *y* clock while the automaton is in a suspended location. Upon rescheduling, the clock is restarted (the *y'==1* expression). *Fig. 6* presents the preemptive version of the simple read-request-computation task model described in section 4.1.

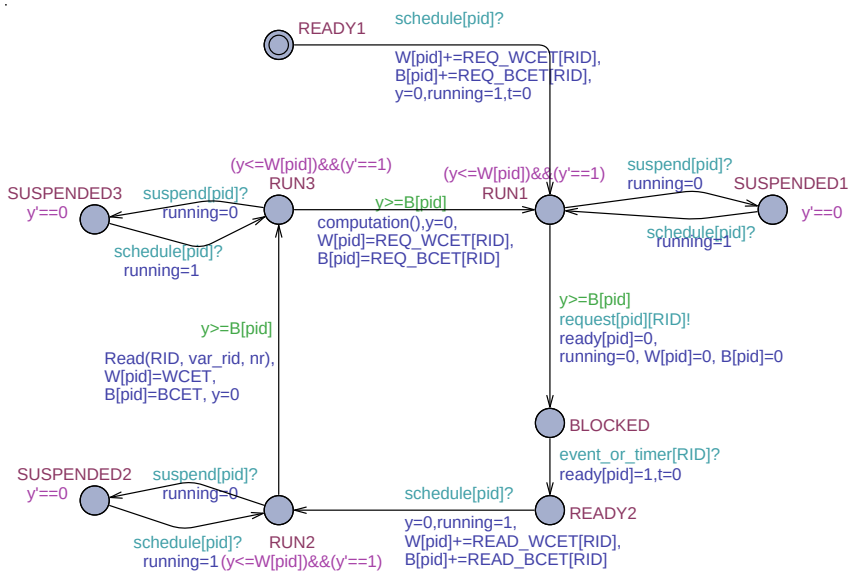


Figure 6: Simple preemptive task model.

The Idle Task is the same as in the case of the cooperative model, except for the fact that it doesn't suspend itself by requesting the *NULL* resource, but it is preempted by the scheduler.

5.2 The preemptive scheduler

The preemptive scheduler model is illustrated in *Fig. 7*.

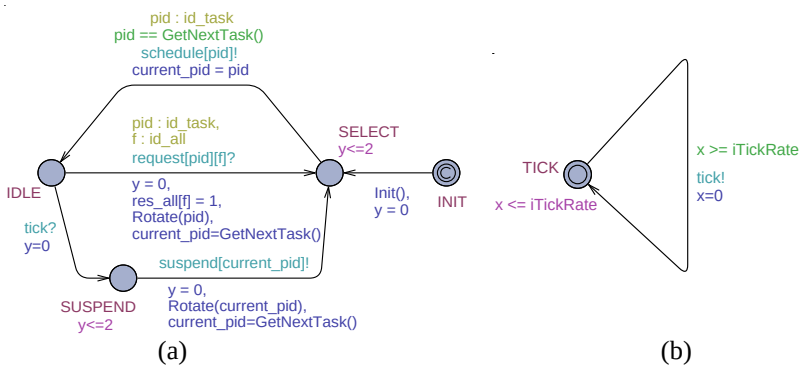


Figure 7: The preemptive scheduler model (a) and the tick interrupt model (b).

The preemptive scheduler can periodically perform a context switch, temporarily suspending the running task in favor of an equal or higher priority one. It does so by using the tick interrupt. Each time the interrupt occurs, the kernel determines if a context switch must take place. This action is performed in the SUSPEND state.

6. Simulation and formal verification

Simulation and formal verification can be performed using the UPPAAL [8] integrated simulation and verification tools. For formal verification, the properties required for an application to function according to requirements must be expressed in UPPAAL's CTL subset [9]. The verifiable properties are: reachability, safety, liveness, bounded liveness and deadlock-freeness.

Reachability properties are checked to see if it is possible to reach a state where a formula *p* is satisfied, for example:

- E<> Task1.RUN1 - checks if *Task1* can ever reach the *RUN1* state;
- E<> Task1.SUSPENDED1 - verifies if there is a possibility that *Task1* will ever be suspended, a context switch taking place.

Safety properties require that a formula *p* is satisfied in all reachable states or, that there is a path in which *p* is always true:

- A[] not (Task1.running and Task2.running) - two different tasks can not be running at the same time;

- E[] Task1.READY2 imply Task1.t<=500 - there is a path where *Task1* will not spend more than 500 time units in the *READY2* state.

Liveness properties require that, in all cases, the system will eventually reach a state where a formula *p* is true. Another form is that if a formula *p* is true, another formula *q* will become true eventually:

- A<> Task1.RUN1 - *Task1* will inevitably be in the *RUN1* state at some point;
- Timer(1).SYNC --> Task5.RUN2 - considering a blocked task waiting for a timer, if *Timer(1)* expires then *Task5* will inevitably be scheduled.

Bounded liveness properties can be formulated "whenever *p* becomes true, *q* becomes true within the time limit *t*":

- Timer(1).SYNC --> (Task5.RUN2 and Task5.t<=200) - when *Timer(1)* expires *Task5* will be scheduled within 200 time units.

Deadlock-freeness [7]: A[] not deadlock.

To verify the time constrains, the execution time of the elements regarding our implementation for the SAM7-EX256 development board (task scheduling time, resource ISR execution time, resource access and read/write operations time, etc.) were measured using the system's physical timers and were introduced in the models.

The interrupt service routine execution time:

- Analog/digital converter: RES_BCET=RES_WCET=4,8 μs;
- Key pressed: RES_BCET=5,2 μs; RES_WCET=10,2 μs;
- Joystick: RES_BCET=5,5 μs; RES_WCET=10,4 μs;
- Timer0 ÷ Timer2: RES_BCET=RES_WCET=2.5 μs.

Driver unified resource access:

- Analog/digital converter: RES_BCET=RES_WCET=26,5 μs;
- Display: RES_BCET=RES_WCET=30 μs;
- External storage: RES_BCET=RES_WCET=30 μs;
- Keys: RES_BCET=RES_WCET=15,4 μs;
- Joystick: RES_BCET=RES_WCET=15,4 μs.

Write operation:

- Display: RES_BCET=390 μs (1 char.); RES_WCET≈7 ms (20 char.);
- External storage: RES_BCET=54 μs (1 char.) ;RES_WCET ≈100 μs(128 char.).

Read operation:

- Analog/digital converter: RES_BCET=RES_WCET=26,5 μs;
- Keys: RES_BCET=27,2 μs; RES_WCET=28 μs;
- Joystick: RES_BCET=27,2 μs; RES_WCET=28 μs.

7. Source code generation

We have developed a code generator application that, based on the model presented as input (XML file), recognize the states and functions calls of each task and output the corresponding source code skeleton. In the case of the simple request-read-computation task presented in the previous sections the following code is generated:

```
void Task1( void *pvParameters )
 {
        for( ;; )
        {
            Request(RID);
            Read(RID, var_rid, nr);
            //!!computation();
        }
 }
```

The generator creates a header file containing the declarations for all the tasks and a C source code file with their implementation. These resulting files can be compiled in a project along with the FreeRTOS source code. Before compilation, all that remains is to add the computational blocks containing the algorithms that manipulate the data. The task code is identical in both cases, cooperative and preemptive.

Alongside the models and source code generator, we constructed a source code project that contains the FreeRTOS source, drivers including interrupt mechanisms for the system's resources and the module for unified peripheral access.

The source code project also includes a special task that can be used to directly measure the time necessary for a sequence of code to execute on this physical system (for use in the model). Also, we included functions to facilitate the conversion of data from the type specific to a particular resource to another resource's type. For example, to convert and copy the data from the state variable associated with the system's analog-to-digital converter to the state variable of the LCD display, one can use the function *ADCtoLCD(res_adc, res_lcd)*. The available functions and they execution time is:

- INTtoLCD(Integer, res_lcd); RES_BCET=7,8 μs; RES_WCET=44,1 μs;
- INTtoSD(Integer, res_sd); RES_BCET=66,5 μs; RES_WCET=305 μs;
- ADCtoSD(res_adc, res_sd); RES_BCET=45,3 μs; RES_WCET=333 μs;
- ADCgetINT(res_adc); RES_BCET=RES_WCET=31,4 μs;
- ADCgetSTR(res_adc); RES_BCET=17,5 μs; RES_WCET=350 μs;
- LCDgetSTR(res_lcd); RES_BCET=RES_WCET=18 μs;
- SDgetSTR(res_sd); RES_BCET=RES_WCET=47,2 μs.

These functions are also present in the model, allowing it to be more expressive and further simplifying the code of application tasks.

8. Conclusions

This paper presents a framework that can be used to model, verify and implement real-time multitasking applications. The operating system, resources and application tasks are modeled by timed automata. This approach allows for the system's simulation and verification before the actual implementation, permitting the early detections of any undesirable behavior. The unified resource access interface and the code generator make possible the automatic generation of the modeled (and verified) application's source code, avoiding most of the error-prone human coding. Because the method is susceptible to state space explosion, the model must be abstract as much as possible, making a compromise between model complexity and its state space size.

References

[1] Fersman, E., "A generic approach to schedulability analysis of real-time systems", *Ph.D. Thesis*, Faculty of Science and Technology, Uppsala University, November 2003.

[2] Waszniowski, L., and Hanzalek, Z., "Formal verification of multitasking applications based on timed automata model", *Real-Time Systems*, vol. 38, no. 1, Springer-Verlag, pp. 39-65, 2008.

[3] Zaharia, T., and Haller, P., "Formal verification and implementation of real time operating system based applications", *in Proc. of the 4th IEEE International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania*, pp. 299-302, 2008.

[4] FreeRTOS – portable, open source, mini Real Time Kernel; http://www.freertos.org

[5] UPPAAL – tool box for modeling and verification of real-time systems modeled as networks of timed automata; http://www.uppaal.com

[6] Liu, J.W., "Real-time systems", Prentice-Hall, Inc., Upper Saddle River, New Jersey 2000.

[7] Li, P., Ravindran, B., Suhaib, S., and Feizabadi, S., "A formally verified application-level framework for real-time scheduling on POSIX real-time operating systems", *IEEE Trans. Software Eng.* vol. 9, no. 30, pp. 613-629, 2004.

[8] Hessel, A., Larsen, K. G., Mikucionis, M., Nielsen, B., Pettersson, P., and Skou, A., "Testing real-time systems using UPPAAL", *Formal Methods and Testing*, Springer-Verlag, pp. 77-117, 2008.

[9] Behrmann, G., David, A., and Larsen, K. G., "A tutorial on UPPAAL", *In Proceedings of the 4th International School on Formal Methods for the Design of Computer, Communication, and Software Systems* (*SFM-RT'04*). *LNCS 3185*, Springer-Verlag, 2004.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 63-72

Application Development in Database-Driven Information Systems

Marius MUJI

Department of Electrical Engineering, Faculty of Engineering, Petru Maior University of Tîrgu Mures, Tg. Mureş, e-mail: marius_muji@yahoo.com

Manuscript received October 1, 2010; revised October 18, 2010.

Abstract: The relational model provides extensive support for data integrity constraints (i.e. business rules) specification, as an integral part of the data model. Current Relational Database Management Systems (RDBMS), however, cover just partially the various categories of data integrity constraints, mostly those directly related with the database structure (e.g. entity integrity, referential integrity). The rest of them are delegated to the application languages. Consequently, they are usually defined in a function-oriented approach (e.g. the object-oriented technology), loosing their direct link with the data model – with all the negative consequences in terms of system scalability and logical data independence. The present paper proposes a data-oriented approach for the development of the external level of database systems. Under the proposed model, the external data is structured only by means of ordered sets of tuples (i.e. arrays of tuples), and the corresponding business rules (i.e. the presentation rules) are treated as external schema integrity constraints. Consequently, the application developer is able to define the user views of the system in a declarative fashion, similar to the relational database design. The immediate advantage is that he or she gains a data designer perspective, rather than one of a programmer. The essentiality (i.e. the unique data constructor) of the model facilitates a seamless integration with the relational model, an entity-relationship graphical representation, and the complete automation of the user interface development.

Keywords: Logical design– data models, schema and subschema.

1. Introduction

Database-driven information systems are developed around an *integrated* and *shared* source of data. The integration is important when somebody needs a general view of the system: for example, a manager who wants to track an item from the supplier to the end-client, spanning the procurement, production and sales activities of a company. This is why, regardless how many individual views we have about an organization's data, there is always needed an integrated, general view of the entire database. On the other hand, it is also important for initial system development *and* for long-term data management purposes to work with data representations which are not dependent on the physical storage equipment.

These requirements led to the ANSI/SPARC three levels architecture [2, 3] (see *Fig. 1*), which makes a clear distinction between the *physical* and the *conceptual* (i.e. logical) representation of the system, and between the general, integrated *community view* and the *individual views* of the system, respectively. The physical-logical separation provide *physical data independence*, which basically means that the applications would not be affected by changes at the physical data representations (for hardware upgrade purposes, for example); the community-individual views separation provides *logical data independence*, which means that the system could grow (through some new user views or modification of the existent ones) without affecting the applications corresponding to the user views that remain unchanged.

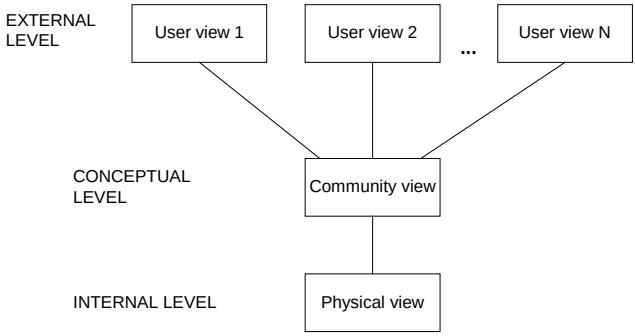


Figure 1: The three levels architecture.

The relational model provides the theoretical support for the development of information systems in accordance with the three levels architecture. Thus, Relational Database Management Systems are currently the technology of choice for the development of the physical and conceptual level, sharing with the application languages the development of the external level.

In this context, the database professionals are traditionally responsible for:

- the data structures at the conceptual and physical level;
- some of the integrity constraints at the conceptual level (i.e. the *database rules*), like: type constraints, entity integrity constraints, referential integrity constraints;
- some of the data structures of the external level (e.g. relational views, parameterized relation-valued operators [7]/stored procedures).

The application professionals are, in turn, responsible for:

- the remaining part of the integrity constraints for the conceptual/community data (i.e. the *application rules*);
- all the data structures of the external views – even when the DBMS provides a layer of data at the external level (e.g. relational views), the application languages need to redefine the entire external view using their own data constructs;
- the *presentation rules* [6] implementation, i.e. the end-user interface, including CRUD (create, retrieve, update, and delete) operations, and display customization (e.g. field labels, field alignment, background and foreground colors, etc.).

The current trend in application development is determined by a significant pressure coming from the programming community, which promotes an object oriented approach for the entire architecture of the information system. Consequently, the data structures and the business rules are usually defined in *a function-oriented approach* [12], loosing their direct link with the *data model* [6] – with all the negative consequences in terms of *system flexibility* and *logical data independence* [7].

By contrast, we propose a data-oriented approach for the development of the information systems, including the external level of the ANSI/SPARC architecture. Thus, we defined a *presentation model*, which preserves the *essentiality* of the relational model [4], i.e. the existence of a *unique data constructor* (in our case, the array of tuples), and prescribes a declarative solution for the presentation rules specification, perceived as *external view integrity constraints*.

The model introduces a clear separation between the display-related presentation rules (e.g. field labels, field alignment, background and foreground colors, etc.), and data-related presentation rules (e.g. data filtering, master-detail navigation, data ordering). CRUD operations are accomplished through the standard behavior of the array constructor. For any CRUD operation initiated by the end user, the system initiates automatically the invocation of some operators from the underlying levels, which actually realize the *mapping* between the presentation level and the lower levels of the system (i.e. the lower external sub-levels and/or the conceptual level – see *Fig. 2*).

Section 2 provides a discussion about the external level of a database-driven system. Section 3 presents our presentation level modeling approach, followed by an example in Section 4. We conclude with the advantages of the proposed approach, and some possible applications.

2. The external level - a closer look

While the external level of the three levels architecture is usually split in multiple sub-levels [3] (see *Fig.2*), *the presentation level* of the system is actually the outermost sublevel, which contains the external data as seen by the end user (i.e. the *external views* of the system).

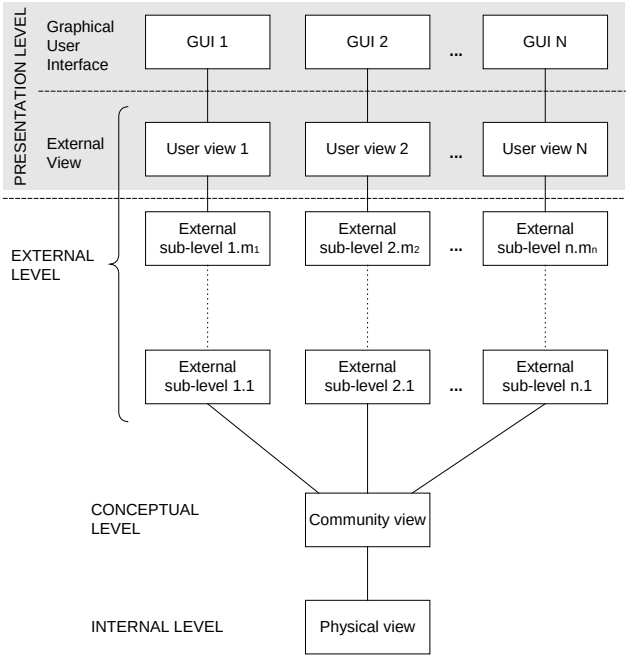


Figure 2: The external sub-levels of the system.

Some of the external sub-levels, i.e. those closer to the conceptual level, are usually implemented under the relational model, through relational views and/or relational operators (e.g. stored procedures). The external sub-levels closer to the end user are built under the theoretical model employed by the application languages (in most cases, object-oriented). The well-known *impedance mismatch* issue is in fact a measure for the lack of compatibility between the two theoretical models. The major difference is determined by the switch of focus from data to function: the data constructs defined by the database

designers are spread among multiple function-oriented software constructs by the application developers [13].

The majority of the mapping solutions employed today to overcome the impedance mismatch have the aim to provide the application developer with the means for accessing the lower (relational) levels of the system transparently, using only the concepts and tools specific to the application languages. Even when some specific concepts of the conceptual level models are introduced (e.g. data entities and relationships) [1], the main purpose is to 'push' the mapping layer as 'low' as possible.

We follow the opposite approach, which considers that the relational model is better suited not only to design the persistent data structures of the conceptual level, but also to build *and* manipulate the data structures of the external level.

However, the end user's perception of data often implies the existence of a *current element* and a certain *inspection order* for a given set of data. It follows that, at least for the presentation level, there is a need for some non-relational features. At the same time, we consider that the essentiality of the relational model (i.e. the existence of a unique, *essential*, data constructor [4]) would provide, also, at the presentation level important advantages related to impedance mismatch and interface automation. Consequently, our model considers the array of tuples as its *unique* data constructor. It could have been a list, or any other collection type, as well – to cite from reference [7], any preference is just "a purely psychological decision – there is no logical reason for preferring (say) an array over a list".

3. The user view from a data oriented perspective

From the end user's point of view, the general behavior of a typical application consists on a limited set of actions related to data entities. In fact, there are just four basic actions, or data-function interfaces [12], classically known as CRUD operations: create, retrieve, update, and delete.

Since it requires a more complex analysis, we'll discuss first the issues related to *data retrieval*. In this regard, the end user can take the following typical actions:

1. **to identify one element in a set**, by means of some unique property or set of properties which distinguishes that particular element from all the rest in the set;
2. **to determine a subset of a set**, based on some *filtering criteria*, namely some common properties of the subset elements;
3. given one element in a set A, and an existing relationship defined from A to B, **to identify all the related elements** in B, under the rule that defines the relationship (e.g. master-detail navigation);

4. **to display the elements of a set in a particular order.**

Let us consider that *all* the data 'seen' by the end user through one particular user view, is composed at any given time by ordered sets of tuples (i.e. arrays of tuples). The user may also be aware about some existing relationships between two sets, under the definition provided on the reference [5]: "Let A and B be sets, not necessarily distinct. Then the relationship from A to B is a rule pairing elements of A with elements of B." Note that we discuss about *directed* relationships, so a relationship defined from A to B will be different from another relationship defined from B to A.

If we consider that any filtering value, which is to be applied to the set X, is seen as an element of another set Y, when a relationship was defined from Y to X, then *the rule which defines the relationship is the filter itself*. Similarly, a change of the display order of a set A may be also accomplished by changing the current element of another set B (which contains the ordering sequences of choice), when a relationship is defined from B to A. Based on the relationship definition, and on the current element of B, the system will reorder the set A accordingly (more accurate: the ordered collection representing A is *(re)created* based on the relationship definition).

Under this approach, we are able to design the entire presentation level only by means of (ordered) sets and relationships between sets. The processing of all the data requests at the presentation level is hidden inside the defining rules for set relationships. The only functionality kept at the presentation level is related to the automatic enforcement of the relationship rules, i.e. the automatic recall of the defining operator attached to the dependent array, when the current element of the parent array changed its value.

Considering the **update operations** (i.e. insert, update, and delete), our presentation model does not require special features, other than the existing data access solutions employed by the application languages. However, in order to preserve the uniformity of the model, and to increase the level of logical data independence, the recommended solution implies the existence of a level of update operators (e.g. stored procedures, or any other application procedures), at the interface with the underlying levels of the system. At the presentation level, we'll have to declare the procedure's name, and the name and type of its parameters.

4. An example

The following example is inspired from the chapter about presentation rules in reference [6]. Some details were added to enable a better presentation of our approach (see *Fig. 3*).

Suppose that we have a user view that exposes to the end user data about customers, orders, and order details. Suppose that the user will have to be able to see at any time all the customers which simultaneously satisfy the following conditions:

- they have a credit limit less than a certain value;
- they are located in a specific region;
- they can be ordered by name, by credit limit, or by the total value of their orders;
- customers whose accounts are overdue must be displayed in red.

Likewise, the user should be able to see, also, at any time, the orders which simultaneously satisfy the following conditions:

- they belong to the current customer;
- their issuing date is in a certain period, say after a start_date and before an end_date, specified by the user;
- they can be ordered by date, value-ascending, or value-descending;
- rush orders must be displayed before regular orders.

When the user inspects a specific order, the system should provide all the order_details that belong to that particular order. Those details should be displayed in their part number order.

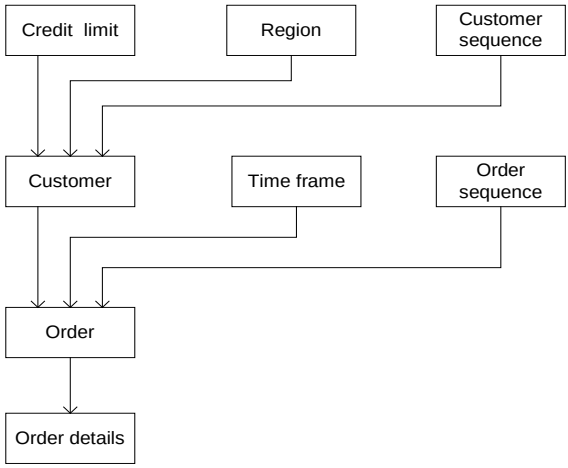


Figure 3: A *user view* example.

In Figure 3, *all* data structures are arrays. Some of them represent application data (i.e. filtering and/or ordering conditions), like *credit limit*, *customer sequence*, *time frame*, *order sequence*. They are not dependent on other data, so their defining functions don't have parameters.

The array named *region* takes its values from the conceptual level (possibly through a relational view), but its content doesn't depend on any other data structure from the user view.

The customer data contained by the *customer* array depends on the current region chosen by the user from the *region* array, on the current customer sequence chosen by the user in the *customer sequence* array, and also on the value provided by the user in the *credit limit* array (the credit limit array will be a special case of an array with one tuple and one attribute, but still an array and not a simple scalar variable, in order to preserve the *essentiality* of the external view model). This is why the defining operator of the customer array should have three parameters, which will automatically take their values at run time from the current tuples in the *region*, *customer sequence*, and *credit limit* arrays, respectively, at any refresh of the customer data.

The list of customer orders exposed to the user at a given moment, contained by the *order* array, depends on the current elements of the *customer* array, the *time frame* array, and the *order sequence* array. Consequently, the defining operator of the order array should have at least three parameters, one for every parent array. In fact, for the present example, we may consider four parameters: one for the link with the customer array (e.g. customer_id), two for the link with the time frame array (e.g. start date, and end date), and one for the link with the order sequence array (e.g. order_sequence_no).

As required, the *order details* array will contain at any moment all the details of the current order from the *order* array. The rule that the details should always be ordered by their part number is specified *inside* the defining function of the order details array, and will remain transparent at the user view design level.

We should also be able to provide solutions for the presentation rules that are not related with *relationship definitions*:

- "customers whose accounts are overdue must be displayed in red" – for this rule, we need to introduce an attribute in the *customer* array, which would allow the distinction of the 'red' customers, so that, at the display level, while defining the graphical object (e.g. the grid, or the list) which displays the customers data, we'll be able to incorporate this presentation rule in a straightforward manner (i.e. declaratively, if possible);
- "rush orders must be displayed before regular orders" – this rule is implemented *inside* the defining function of the *order* array (which is completely transparent for our model) .

So, under the proposed model, the developer is able to design the presentation level declaratively, just specifying:

- the declaration of all the array structures: array name, attribute names, data types;

- the defining operator of every array;
- the link between every parameter of any defining operator and its corresponding attribute from the parent array;
- the update procedures, their triggering events, and the links of their parameters with the corresponding attributes.

Our user view's dependency graph was represented graphically in *Fig. 3* using arrows, oriented from parent to child, but it could have been used any other entity relationship graphical notation (e.g. crow foot, IDEF1X, IE, etc.). Thus, the user views will have the same (E-R like) graphical representation as the conceptual level. The only difference is that instead of *foreign key relationships*, we have pairs of defining operator *parameters* and *attributes* of the parent array(s).

5. Conclusions

There is a clear need for a data-oriented approach in application engineering. The software engineering field is now dominated by the new trend introduced by the OMG's Model Driven Architecture [14], which has a strong object oriented bias. The position sustained by this paper is that the application development should be *not only model-driven, but data-model-driven* [10, 11]. The paper introduces a data-oriented model for the development of the external level of database systems, which considers the *presentation level* as the only required data layer above the relational data model. Moreover, this should be a thin layer, with the unique purpose of data presentation, which doesn't need to address any business logic other than the *presentation rules* [6].

The *standard behavior* and the *essentiality* of our model enable the automation of the presentation level development. At the same time, the *mapping operators* (defined at the lower levels and called at the presentation level to promote the CRUD operations to the conceptual level) are the key for the provision of logical data independence at the presentation level. This constitutes the major step forward from the previous attempts to automate the interface, which failed to provide an appropriate degree of logical data independence at the external level of the system. Trying to generate the interface based on various entity-relationship patterns existent at the conceptual level, and assuming that the user views are just sub-schemas of the conceptual level [15, 16, 18], they become useless as soon the external level has multiple sublevels, i.e. the presentation data is obtained from the conceptual data through a series of complex operations – which is always the case for large, integrated information systems.

The foreseen applications of the presentation model are related primarily to the application development for database-centric systems (e.g. enterprise

resource planning systems, e-commerce systems, etc.). CASE tools which support entity-relationship diagrams represent, also, an important area for our model implementation.

Future work will concentrate primarily on the development of interface automation tools, designed in an object-oriented approach, and implemented with general purpose third-generation languages (e.g. Java, C#). In a long term vision, the presented model could be used in data-model driven methodologies for declarative development of database-centric applications.

References

[1] Adya, A., Blakeley, J. A., Melnik, S., and Muralidhar, S., "Anatomy of the ADO.NET entity framework", *ACM SIGMOD International Conference On Management Of Data. Beijing, China*, 2007, pp. 877-888.

[2] ANSI/X3/SPARC Study Group on Data Base Management Systems. "Interim Report", *ACM SIGMOD Bulletin*, no. 2, 1975.

[3] Date, C. J., "An Introduction to Database Systems (8th edition)", Addison-Wesley, 2003.

[4] Date, C. J., "Date on Database: Writings 2000-2006", Apress, 2006.

[5] Date, C. J., "Logic and Databases: The Roots of Relational Theory", Trafford Publishing, 2007.

[6] Date, C. J., "What Not How: The Business Rules Approach to Application Development", Addison-Wesley, 2000.

[7] Date, C. J., and Darwen, H., "Foundation for Future Database Systems: The Third Manifesto (2nd Edition)", Addison-Wesley, 2000.

[8] Halle, B., "Business Rules Applied: Building Better Systems Using the Business Rules Approach", Wiley, 2001.

[9] Hay, D. C., "Data Model Views", *The Data Administration Newsletter - TDAN.com*, Apr. 2000.

[10] Lewis, B., "Data Lineage: The Next Generation", *The Data Administration Newsletter - TDAN.com*, Aug. 2008.

[11] Lewis, B., "Data-Oriented Application Engineering: An Idea Whose Time Has Returned", *The Data Administration Newsletter - TDAN.com*, Jan. 2007.

[12] Lewis, W. J., "Data Warehousing and E-Commerce", Prentice Hall PTR, 2001.

[13] Lewis, W. J., "E-Commerce Vs. Data Management", *The Data Administration Newsletter – TDAN.com*, Jan. 2002.

[14] Model Driven Architecture. http://www.omg.org/mda/

[15] Pizano, A., Yukari, S., and Atsushi, I., "Automatic generation of graphical user interfaces for interactive database applications", *Conference on Information and Knowledge Management, Washington, D.C.*, 1993, pp. 344-355.

[16] Rollinson, S. R., and Roberts, S. A., "A mechanism for automating database interface design, based on extended E-R modelling", *Advances in Databases. s.l. : Springer Berlin / Heidelberg*, 1997, pp. 133-134.

[17] Ross, R. G., "Principles of the Business Rule Approach", Addison-Wesley Professional, 2003.

[18] Rowe, L. A., and Shoens, K. A., "A form application development system", *ACM SIGMOD International Conference On Management Of Data., Orlando, Florida*, 1982, pp. 28-38.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 73-86

Exambrev - Integrated System for Patent Application

Attila ASZALOS, József DOMOKOS, Tamás VAJDA,
Sándor Tihamér BRASSAI, László DÁVID

Department of Electrical Engineering, Faculty of Technical and
Human Sciences, Sapientia University, Tîrgu Mureș,
e-mail: aszi.atti@hotmail.com; domi@ms.sapientia.ro;
vajdat@ms.sapientia.ro; tiha@ms.sapientia.ro; ldavid@ms.sapientia.ro

Manuscript received October 1, 2010; revised October 30, 2010.

Abstract: In this paper we present the design and development of a patent application, examination and evaluation system based on the JEE platform.
In the Introduction part of the paper we present the necessity of a patent application system, and various organizations dealing with patent data and classification.

The second section of the paper presents the architecture of our system and its functionalities, which include online patent request registration, reduction of the time required for the application examination by automatic and semiautomatic verification of the formal aspect of a patent application. The system helps the inventors in the International Patent Code assignment procedure and provides the possibility for the domain experts to search for similar technical solutions in the Romanian State Office for Inventions and Trademarks (OSIM), the European Patent Office (EPO) and the World Intellectual Property Organization (WIPO) databases and using the Google Patent Search engine. This considerably speeds up the patent examination process.
The final part of the paper presents the results of the IPC suggestion algorithm, both execution time and qualitative evaluation of it. It also includes the resulting execution times of the search in the above presented patent databases.

In this article we have made an overall description of the system and we have focused on the description and results of the IPC suggestion algorithm and on the search process in the above mentioned patent databases.

Keywords: Patent application, patent search, evaluation and examination system, IPC suggestion.

1. Introduction

Worldwide patent applications are growing at an average rate of 4.7% per year, according to the 2007 edition of the World Intellectual Property Organization (WIPO)'s Patent Report [1]. The patent examination procedure has

two stages: formal verification which follows all the formal procedural steps and verifies if applications are patentable and the evaluation stage which checks the grade of novelty and innovation of the patents [2], [3]. To reduce the patent examination time and increase the quality of the evaluation, despite that the number of the patent applications are growing, there are two possibilities: to increase the number of employments of the State Office for Invention and Trademarks (OSIM) or to reduce the amount of work required for registration, formal verification and evaluation by using an online integrated system. The following paragraphs of this section present similar existing systems.

OSIM [4] is a specialized government body that has exclusive authority in Romania in the field of protection of industrial property. Taking into consideration the special economic importance of the industrial property and the need of a competitive management of information in the field of industrial property, the OSIM has developed a system of services by which offers to the large public useful information concerning industrial property, processed by highly competent specialists such as to facilitate correct economic decisions to be taken. It pays special attention to the promotion of the industrial property.

From 2006, OSIM offers the possibility to register on-line to the epoline® system, for the following types of patents:

- patents filed according to the European Patent Convention (CBE/EPC), through OSIM as the national office;
- patents filed according to the Patent Cooperation Treaty (PCT), through OSIM as reception office;

In the present it is not possible to register online the Romanian national patent. On the OSIM web page you can find important information about on-line registration for the above mentioned patent application such as: important announcements, details about the services, information about how to register on-line, software for registration of the patent request at OSIM, recommendations, assistance for clients who want to register on-line an invention and some details about this page services.

EPO [5] provides a uniform, coherent application procedure for individual inventors and companies from 38 European countries. It is the executive body of the European Patent Organization and is supervised by the Administrative Council. The main role of the EPO is to grant European patents.

The EPO carries out researches and substantive examinations on a continuously growing number of European patent applications and international applications filed according to the Patent Cooperation Treaty. In the case of European patent applications, the Office gives the option of an accelerated procedure. The Office examines also oppositions against already granted European patents.

Publication of the invention is very important to the European patent system. The public can obtain copies of the patent documents from the European Publication Server. The European Patent Register provides details of the status of patent procedures at the EPO. All the EPO's patent documents are available to the public through the free Esp@cenet service on the Internet. The EPO also provides a wide range of other products for searching patent databases.

The epoline® is an EPO package of software with on-line services that allows users to create and apply electronically for patents at the Intellectual Property Office and other national and international offices, including the EPO and WIPO. The epoline® is a high security system based on smart cards.

To join the system, one should make two steps: first, to get an EPO smart card, then to fill in on-line an IPO (Intellectual Property Organization) enrolment form and submit.

This service has a series of advantages. It is a user-friendly application that helps inventors to build their applications and forms with a validation option that helps users to make applications and forms right even when doing this for the first time. Security of sending the documents is ensured, there are no postal delivery delays or postage costs. The sender receives an immediate filing receipt after sending the forms.

WIPO [1] is a specialized agency of the United Nations. It is dedicated to develop a balanced and accessible international intellectual property system, which rewards creativity, stimulates innovation and contributes to economic development in regard to the public interest.

WIPO was established by the WIPO Convention in 1967 for the protection of intellectual property worldwide by collaboration with other international organizations and cooperation among states. Its headquarters are in Geneva, Switzerland. WIPO considers that intellectual property is essential to the economic, social and cultural growth of all countries. Thus its objective is to promote the effective use and protection of intellectual property (IP) worldwide.

WIPO provides services for the owners and users of intellectual property, such as international registration services, thus a single application has to be filed that is valid in multiple countries. IP classification systems of WIPO are used for registering IP and making it easy to search in IP databases and registries. WIPO's Arbitration and Mediation Centre offers resolution services for private parties involved in international intellectual property disputes [1].

PATENTSCOPE® Search Service is a service that makes possible for users to search in all international patent applications published, starting from the first one that was published in 1978 to nowadays, and has a special part for the latest information and documents available online.

In this article we present an Integrated Patent Examination Expert System (EXAMBREV) developed as a web application considering the Java EE

platform, which helps the applicants to accomplish the entire registration procedure of a patent request and verifies the formal correctness of the patent application form. Our system's other functionalities include the management of civil servants and expert users, the presentation of the application for evaluation through a web interface. Our translator unit is a helping hand for experts for searching similar technical solutions in a wide range of different language patent databases and the Romanian Patent Database. Our system will also help the management of the OSIM to see in which field to employ new domain experts in the future.

The outline of the paper is as follows. Section 2 presents our system architecture and describes the tasks and functionalities of each subsystem. Section 3 presents the results of the execution time of various algorithms used by the system, Section 4 contains the conclusions. The final part of the paper presents the acknowledgements and references.

2. Technical information

A. System overview

The architecture of our system is presented in *Fig. 1*. Our system has two main modules divided in multiple subsystems. The first module is called *Interfaces and data preparation module* which manages the patent requests, common users (UCOM), expert users (UEX), civil servants (UFUNC), applicants (UAPP), administrators (UADM), civil servant managers (UFUNCM) and expert managers (UEXPM) and also prepares some initial data for the *Expert system module* (SIEXP). The second module is the *Expert system module* (SIEXP) which gives the world wide novelty of a technical solution proposed by an inventor and contains the legal and procedural database. In this paper the *Interfaces and data preparation module* and especially the search methods for similar technical solutions in the online patent databases are presented.

The deployment diagram shown in *Fig. 2* illustrates the connections between the different subsystems of the *Interfaces and data preparation module* and their deployment on the used servers. As we can see, all the subsystems communicate with the system database through JPA (Java Persistence Application Programming Interface) which communicates with the database through the JDBC API (Java Database Connectivity API).

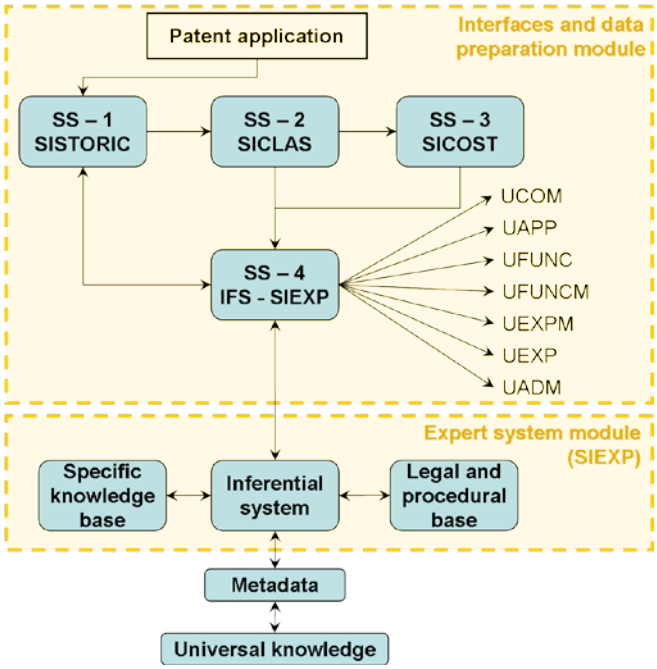


Figure 1: EXAMBREV system architecture.

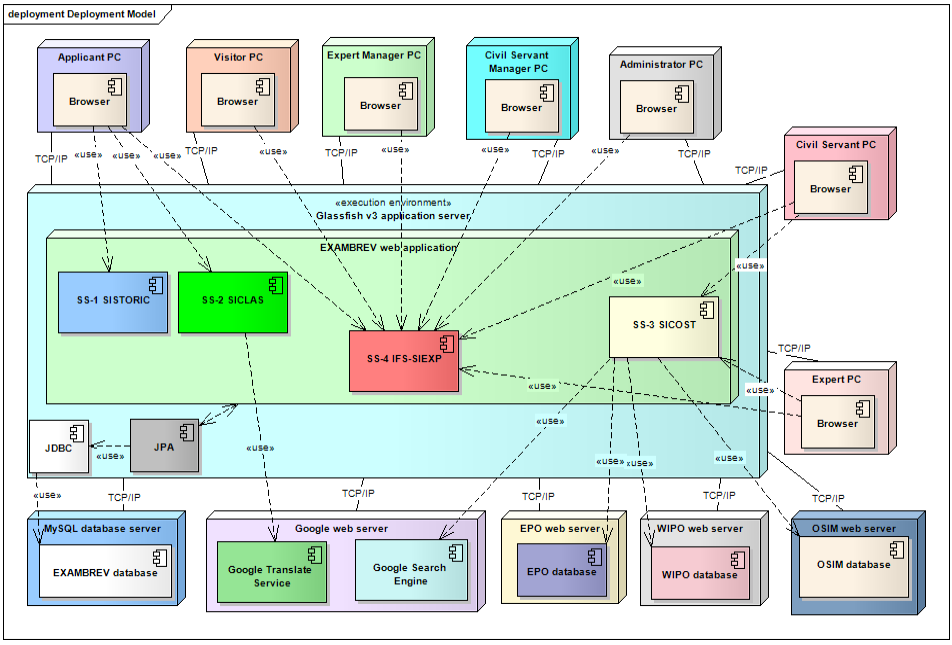


Figure 2: The deployment diagram of the EXAMBREV system.

B. The SISTORIC subsystem

The SISTORIC (SS-1) subsystem has two main functionalities [6]:

- the management and storage of the information regarding a technical solution proposed by an inventor;
- the automatic formal verification of the patent application.

According to Romanian State office for Inventions and Trademarks we had to deal with three types of patents:

- EPC (European Patent Convention);
- PCT (Patent Cooperation Treaty);
- Romanian national patent.

We have implemented for now the management system for the Romanian national patent. The application is accessible from the Internet through a web browser and makes possible to submit online patent requests. The data of a patent application is stored in a relational database management system, a MySQL database.

The software subsystem provides the following features [7]:

- Applicant user registration;
- Civil servant user registration;
- Expert user registration;
- Account activation for the above users;
- Online patent application;
- Editing patent application information;
- Editing account information;
- Patent application list;
- Semiautomatic IPC code assignment.

In the registration process the user starts the registration and fills in all the required personal information and specific user information. In the case of the expert users the specific information is the list of the IPC categories in which he has knowledge. After the data validation process the system sends an activation email to the registered person, containing a link to the activation page. The account of the user will be accessible after clicking the link in the received email and activating his account. In the case of the civil servant and expert users after the activation procedure their account must be confirmed. This can be done by the civil servant manager or the expert manager. These manager type users are promoted from the civil servant users by the administrator. If a civil servant is promoted to manager his account gets confirmed automatically. The user's personal information is saved in the database.

The login process is different for almost every user type, although there is a common part too. In the common part of the login process the system checks

the username, password and the activation status of the account. If the username and password are matching and the account is activated the applicant and expert users can login. If the expert user's account isn't confirmed he has limited accessibility in the system and can only change the list of the categories he is expert at. The civil servant users can log in only if their account is confirmed.

The patent application process consist in filling of the online application form which is the same used at OSIM in present. This process is divided into 4 steps, because this way the amount of required data on one page isn't too large and if there are validation errors, the user can correct them more easily.

C. The SICLAS subsystem

The SICLAS (SS-2) subsystem contains the IPC suggestion algorithm which can be used by the applicants to find the appropriate IPC categories for their invention [8], [9]. After the patent application is submitted by the applicant he is asked to select the IPC categories in which the invention should be categorized. At this step the applicant can describe his invention by keywords in Romanian or English and ask for a list containing the suggested IPC categories. If the keywords are given in Romanian, the subsystem's translator unit automatically translates them into English. The API used by the translator unit is the Google Translate API.

Before we describe the logic of the IPC suggestion algorithm, we present the database diagram extract containing the tables used for the storage of the IPC categories. *Fig. 3* shows the tables representing the 5 IPC levels: sections, classes, subclasses, main groups and subgroups. These tables contain the IPC codes and their description of every IPC level. The one-to-many relationships between the tables illustrate the hierarchical organization of the IPC levels.

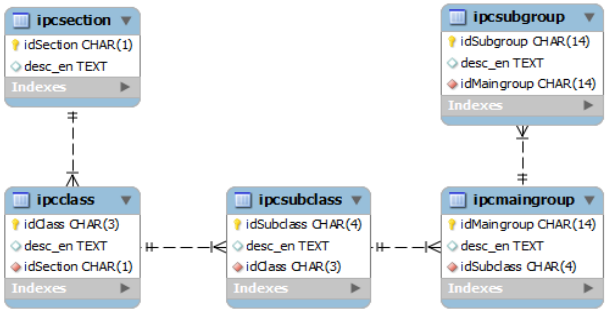


Figure 3: Database diagram extract of the IPC category tables.

We can define the goal of the IPC suggestion algorithm as determining a list of appropriate IPC categories for an invention described by keywords. The algorithm can be split into 3 sections: initialization, search and result

propagation. The list of coefficients initialized in the first section and their initialization values are shown in Table 1.

Table 1: The IPC suggestion algorithm coefficients and their initialization values.

Section *(ps)*	**Class** *(pc)*	**Subclass** *(psc)*	**Main group** *(pmg)*	**Subgroup** *(psg)*
30	25	20	15	10

The search section of the algorithm contains 6 steps. In the following these will be presented in detail. For a better understanding of the algorithm we introduce two result sets, A and B, which will contain the temporary and final search results. In the first step we search on all IPC levels for IPC categories of which description contains at least one of the given keywords. These categories are inserted into result set A. This step is shown in *Fig. 4*.

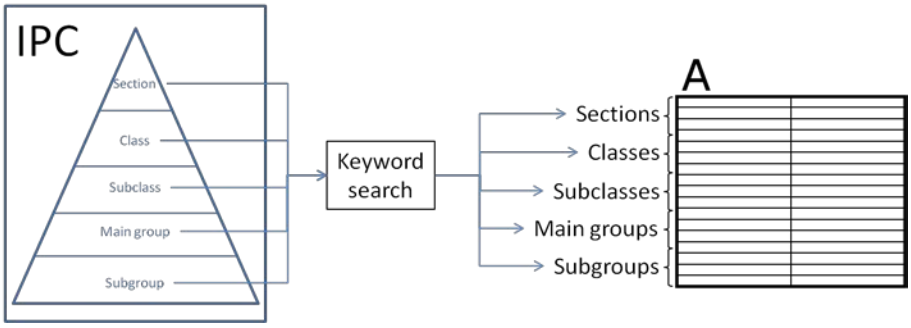


Figure 4: Keyword search mechanism on all IPC levels and building of the result set A.

In the 2nd step we take the subgroups from result set A, insert them into result set B and calculate a suggestion value for them. The general form of the calculation formula is given in *Eq. 1*:

$$SgVal = NoOfKwdsInIPC\_Desc \cdot IPCLevCoef + OldSgVal \quad (1)$$

In this step the old suggestion value is 0, the IPC level coefficient is 10, and the number of keywords in the IPC category description is calculated for every record. *Fig. 5* shows the graphical representation of this step.

In the 3rd step we advance upwards in the IPC's hierarchical organization to the level of the main groups. We take those main groups from A, which does not have subgroups in result set A, insert them into result set B, and calculate their suggestion values with the *Ec. 1*. This is illustrated on *Fig. 6* as the "i" substep. The second substep "ii" consists of updating the suggestion values of those subgroups in result set B, which belong to the main groups in the result set A. The updated value is calculated using the formula given in the *Ec. 1*.

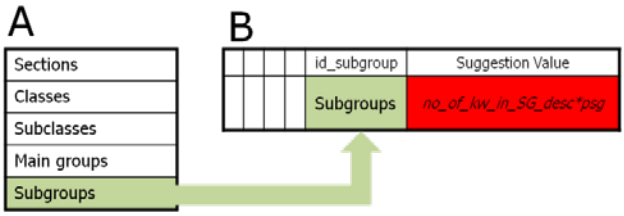


Figure 5: 2nd step of the search in the IPC suggestion algorithm.

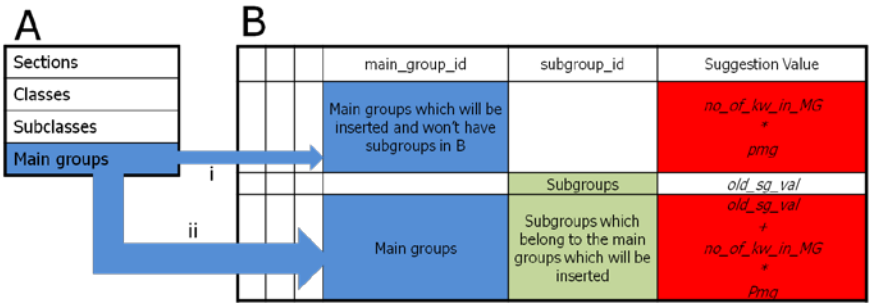


Figure 6: 3rd step of the search in the IPC suggestion algorithm.

The following steps from 4 to 6 are similar to the steps already presented. We continue to advance upwards in the IPC's hierarchical levels, insert the IPC categories without subcategories in result set B along with their suggestion value, and finally update the suggestion value of those categories in result set B which had parent categories in result set A. All of the steps and the algorithmic language of the suggestion algorithm were presented in detail in [9]. One more step, the 4th, of the algorithm is presented in *Fig. 7*.

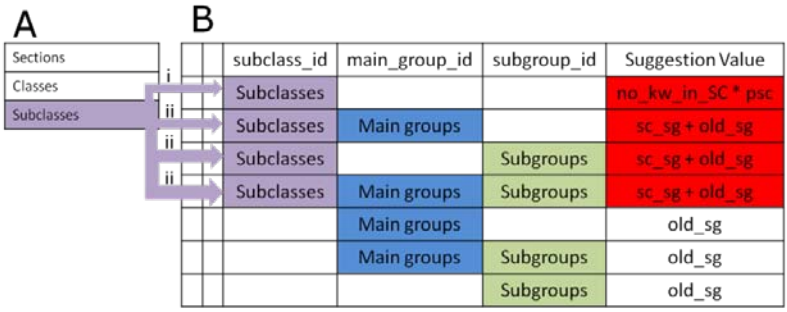


Figure 7: 4th step of the search in the IPC suggestion algorithm.

From the first version of the algorithm to the final version there were made some changes in the technical realization of it, in order to achieve better execution times. The necessity of this optimization was presented in [10], but

there weren't presented the results of it. Section 3 presents a comparison between the execution times of the non-optimized and optimized version of the suggestion algorithm.

D. The SICOST subsystem

The SICOST (SS-3) subsystem is responsible for Expert user (UEXP) data management [6], [8]. This subsystem also takes the IPC code given to a patent application by SS-2, and outputs it to the civil servant manager helping him to choose 3 expert users for this domain.

This subsystem also provides a web interface for the expert users to search the following online patent databases:

- WIPO database;
- EPO database;
- Romanian Patent Database.

All of the above databases can be searched online from the esp@canet webpage. In our system we give the option to the expert to search these databases by keywords or by IPC code.

When the expert searches by keywords we build a URL with the required request parameters for the esp@canet webpage and execute it. The response given by esp@canet has a fixed structure containing an HTML table with the results of the search. This HTML table is parsed by us using a HTML parser and the results are organized in a list containing the following information: the title of the invention, the link to the page presenting the invention in detail, the inventors, the applicants and the IPC code of the invention.

If the expert chooses to search in the patent databases by IPC code the process is similar to the search by keywords, the difference is in the construction of the URL. It takes different request parameters with the value of the IPC code given by the expert.

Our system also provides the possibility for the experts to search with the Google Patent Search Engine.

The results of these searches are presented to the expert immediately after the expert submits the search and help him in finding similar technical solutions.

The detailed description of these search mechanisms was presented in [10], where wasn't presented the fact, that these mechanisms are implemented in two versions: a single- and multi-threaded version. The multi-threaded version of the search mechanism was implemented as an optimization of the execution time of the search mechanisms. In section 3 we present the results and comparison of the execution time of the two versions of the search mechanisms.

E. IFS-SIEXP subsystem

The IFS-SIEXP (SS-4) subsystem is the special interface for the SIEXP module [6], [8], [11]. It makes data transfer between the Interfaces and data preparation module and Expert system module. It also communicates with UEXP, UCOM and UINV via a Web interface. This is the login point to the Web application for registered users.

3. Results

In the first part of this section there is presented a comparison between the execution times of the optimized and the non-optimized IPC suggestion algorithms. This is followed by the qualitative results of the previously mentioned algorithm. The execution time of the similar invention search mechanism and the comparison of the two versions (single and multi-threaded) of the search are presented in the second part of this section.

A. Results of the IPC Suggestion Algorithm

Table 2 shows the execution times of the IPC suggestion algorithm. The execution times were measured with the SQLyog software, which measures the execution time of every executed query.

Table 2: Execution times of the IPC Suggestion Algorithm.

Keywords	**Non-optimized (sec)**	**Optimized (sec)**
METHOD SYSTEM COMPUTER CONTROLLED BICYCLE GEAR SHIFTING	19,312	6,813
BICYCLE GEAR SHIFTING METHOD APPARATUS	20,984	5,203
MOUSE RODENT TRAP	2,813	2,860
ELECTRIC MOUSE RODENT TRAP	6,968	2,718
HAIR CUTTING DEVICE	26,563	6,562
EDGE DETECTION IMAGE PROCESSING	2,812	2,047
Average:	13,2420	4,3672

The second and third columns of the table contain the non-optimized and optimized algorithm's execution times. If we have a look at the difference between the average execution times of the two versions of the algorithm, it is evident that there is a 67.02% decrease in the execution time of the algorithm so the optimized algorithm performs more than 3 times faster.

Table 3: Qualitative results of the IPC Suggestion Algorithm.

No.	Section	Class	Subclass	Main group	Subgroup	Suggestion value
1	G	G06	G06T	G06T0009000000	G06T0009200000	75
2	A	A22	A22C	A22C0029000000	A22C0029020000	70
3	A	A22	A22C	A22C0029000000	A22C0029040000	70
4	G	G06	G06T	G06T0001000000		70
5	G	G06	G06T	G06T0003000000	G06T0003200000	65
6	G	G06	G06T	G06T0003000000	G06T0003400000	65
7	G	G06	G06T	G06T0003000000	G06T0003600000	65
8	G	G06	G06T	G06T0005000000	G06T0005500000	65
9	G	G06	G06T	G06T0007000000	G06T0007600000	65
10	G	G06	G06T	G06T0011000000	G06T0011800000	65

Table 3. shows the suggested IPC categories and suggestion values for an existing invention, an algorithm suitable for edge detection in image processing. The following keywords were given as an input for the suggestion algorithm: "edge detection image processing algorithm". The IPC main group in which the existing invention is categorized is shown in the gray cell of the table. As for the evaluation of the algorithm's quality, we can state, that having a look at the suggestion values, the correct IPC category is located on the 3rd place. If we have a look at the hierarchical structure of the IPC, we can see that the algorithm determined correctly the section, class and subclass level of the invention even in the first result.

B. Results of the Similar Invention Search Mechanism

Table 4 and *Table 5* contain the execution times of the single- and multi-threaded version of the similar invention search mechanisms and the number of results.

It is important to mention that the table contains only the execution time of the search mechanism, measured with the functions provided by the Java language for execution time measurement and does not include the search setup time. *Table 3* shows us that the single-threaded version of the search mechanism was approximately 2 times slower in average, comparing to the multi-threaded version.

The difference between the test cases shown in *Table 3* and *Table 4* is in the used search providers and search languages. In *Table 3* there were used three search providers provided by Esp@cenet plus the Google Patent Search with English keywords. In *Table 4* the tests were conducted on the Esp@cenet search providers in English and Romanian languages. Having a look at the average execution times in *Table 4* we can conclude that the multi-threaded version of

the search mechanism is approximately 2 times faster than the single-threaded version.

Table 4: Execution times of the two versions of the Similar Invention Search Mechanisms with Esp@cenet and Google Patent Search with English keywords.

Keywords	Single (sec)	Multi (sec)	No. of results
METHOD SYSTEM COMPUTER CONTROLLED BICYCLE GEAR SHIFTING	6,233	2,926	96
	6,154	2,443	
	5,925	2,326	
BICYCLE GEAR SHIFTING METHOD APPARATUS	7,127	4,612	126
	7,293	3,088	
	7,785	3,010	
HAIR CUTTING DEVICE	14,471	8,346	276
	12,908	6,950	
	13,835	7,396	
Average:	9,08	4,57	

Table 5: Execution times of the two versions of the Similar Invention Search Mechanisms with Esp@cenet using English and Romanian keywords.

Keywords	Single (sec)	Multi (sec)	No. of results
MOUSE RODENT TRAP	5,545	3,115	6
	5,520	2,365	
	5,549	2,393	
EDGE DETECTION IMAGE PROCESSING	6,728	3,971	36
	6,960	3,322	
	6,643	3,191	
HOUGH TRANSFORMATION	6,691	2,963	30
	6,356	3,175	
	6,296	3,200	
Average:	6,25	3,08	

4. Conclusion

We designed and developed a JEE based integrated system for patent examination. The system will help applicants to make online patent application registration for all three patent types discussed (EPC, PCT and Romanian national patent type). The system also helps OSIM patent evaluator experts management, employers management and patent management.

The main results obtained are the UCOM, UEXP, UINV, UFUNC and patent application registration interfaces. The interfaces were developed considering Java Server Faces technology and PrimeFaces 2.0 technology.

We have developed an algorithm for semiautomatic IPC code assignment for helping the applicants and also the evaluator experts and a patent database search mechanism which speeds up the similar technical solutions search.

The focus in this paper was on the presentation of the results of the optimized IPC suggestion algorithm and the multi threaded similar invention search mechanisms.

Acknowledgements

This project is developed under Partnership in Anterior Domains Program of National Authority for Scientific Research in Romania, project code: 11-076/2007.

References

[1] WIPO webpage: http://www.wipo.int/

[2] Implementing regulations to the patent law no. 64/1991, as republished in *Official Gazette of Romania*, Part I, No. 456/18 June 2008.

[3] Patent law No. 64/1991, as republished in *Official Gazette of Romania*, Part I, no. 541/8 August 2007.

[4] OSIM webpage: http://www.osim.ro/

[5] EPO webpage: http://www.epo.org/

[6] Radu, M., "Elaborarea strategiei de cercetare privind examinarea cererilor de brevet de invenție și studiu critic asupra procedurilor de examinare aflate în uz", *Technical report EXAMBREV, stage I, PNII – Parteneriate, no. 11-076/2007*, Centrul de Cercetări pentru Materiale Macromoleculare și Membrane, București, 2007.

[7] Domokos, J., Vajda, T., Brassai, S. T., Dávid, L., "Realizarea, implementarea în faza de laborator și testarea sistemului informatic de examinare a cererilor de brevet de invenție", *Technical report for stage III, EXAMBREV, PNII – Parteneriate, no. 11-076/2007*, Sapientia University, Tîrgu Mureș, 2009.

[8] Brassai, S. T., Dávid, L., Domokos, J., Vajda, T., "Technical report for stage II", *PNII – Parteneriate, no. 11-076/2007*, Sapientia University, Tîrgu Mureș, 2008.

[9] Vajda, T., Domokos, J., Brassai, S. T., Dávid, L., Aszalos, A., "Developement of EXAMBREV Integrated System for Patent Application", in *Proceedings of the 4th edition of The INTER-ENG International Conference*, Tîrgu Mures, România, 12-13 November, 2009, pp. 309-314.

[10] Aszalos, A., Domokos, J., Vajda, T., Brassai, S. T., Dávid, L., "EXAMBREV Integrated System for Patent Application", in *Proceedings of the 2nd International Conference on Mechatronics, Automation, Computer Science and Robotics* (*MACRo 2010*)*, Tîrgu Mureş, Romania*, 14-15 May 2010, pp. 55-62.

[11] Domokos, J., Vajda, T., Brassai, S., T., Dávid, L., "Integrated System for Patent Application Examination (EXAMBREV)", in *Proceedings of 17th International Conference on Control Systems and Computer Science* (*CSCS17*)*, Bucureşti, Romania*, 26 - 29 May 2009, pp. 135-139.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 87-98

Inter-Domain Traffic Engineering for Balanced Network Load

Levente HUSZÁR, Csaba SIMON, Markosz MALIOSZ

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Magyar tudósok krt. 2., 1117 Budapest, Hungary,
e-mail: {huszar | simon | maliosz}@tmit.bme.hu

Manuscript received October 01, 2010; revised October 30, 2010.

Abstract: Inter-domain cooperation at control plane level offers the possibility to balance the traffic in a much more flexible way compared to the single domain Traffic Engineering (TE) methods. This solution is especially effective if the offered traffic and the network load permits to avoid congestion without recomputing the routes. Current networking technologies and the emerging trends in aggregation-access-core network architectures provide means to realize such a cooperative control plane. We consider a dual opto-electronic network model where traffic grooming is also possible. In the particular case of the layer 2 switched aggregation network, which feeds a core network using a GMPLS control plane, this solution means redistributing the traffic between the spanning trees of the aggregation domain. We propose an inter-domain control plane cooperation and investigate it by means of simulations on several core network topologies deploying IP over WDM technology. In the simulations we analyze the effects of the traffic redistribution rate on the throughput, on the number of lightpaths and on the number of opto-electronic conversions. We show that if congestion occurs in the core, we can eliminate the congestion just with a proper coordination between the control planes of the aggregation and core domains, redistributing the traffic prior entering the core.

Keywords: Traffic engineering, knowledge plane, GMPLS, MSTP, CSPF.

1. Introduction

Emergent technologies like GMPLS, WDM and carrier-grade Ethernet will replace legacy ones in future Internet domains. Combining of these different data plane technologies and different services at different layers into an efficient interworking environment is a challenging task. The resulting system should offer a trade-off for service providers to operate their networks.

A common trend in communication networks is the widespread of fiber technologies. The optical networks are favored by the operators because they offer higher network capacities. At the same time they also come with more advanced control and management features. However, these, combined with the new services demanded by the market, increase the complexity of the networks and advanced network optimization mechanisms become a must. The most common optimization mechanisms deal with the traffic, several Traffic Engineering (TE) proposals are known in the literature. Typically such TE solutions deal with a single domain, only [1], [2], [3]. Alternatively a joint optimization of traffic over a cascade of core networks has been proposed [4], [5], but they consider the same TE algorithm all over the domains. Nevertheless, the traffic originated by the end users reaching the core networks should cross the aggregation and access domains, which use different TE mechanisms. This paper investigates the effect of cooperation at control plane level between the different TE mechanisms of the core and access or aggregation domains. The advantage of this cooperation is that it allows much more flexibility in maintaining balanced core network usage. Additionally, other optimization goals can be satisfied, if we can redistribute the traffic among the edge nodes connecting the access and the core. In our paper we investigate the impact of network load optimization on the efficiency of a dual opto-electronic network model [6], [7].

In the followings, we will investigate the networking technologies covered by the paper and summarize the trends in aggregation-access-core network architectures. Then we will present our solution on the joint access-core network optimization and we investigate it by means of simulations. Finally we conclude our paper.

2. Networking technologies

A. Data and control plane technologies

Most of the applications deployed over the info-communication networks are based on IP, while the access networks are predominantly Ethernet-based. From the access network the Ethernet traffic is concentrated at the edge nodes and is forwarded to packet-based transport in the core domain. The trends evolve towards the wide deployment of WDM (Wavelength Division Multiplexing) network devices in the core, which enables the transmission over the established connection-oriented lightpaths in the optical domain.

These lightpaths form a virtual topology over the physical topology that can be reconfigured dynamically in response to traffic changes and/or network

planning. The combination of IP directly with WDM results in an efficient assignment of optical network resources to forward IP traffic [8], [9].

The versatile Generalized MultiProtocol Label Switching (GMPLS) protocol enables the integrated control of both IP and WDM. Integrated routing and wavelength assignment based on GMPLS is currently the most promising technology, which also delivers effective TE capabilities. The typical routing protocol in such networks is the Constrained Shortest Path First (CSPF) [10], [11].

B. Dual opto-electronic model

In the above model the lower layer is an ’optical’ one, the upper layer is typically an ’electronic’ one, capable of performing joint time and space switching. Paths of the lower layer correspond to a single link in the upper layer. Lightpaths are special routes: they arise and terminate in the electronic layer. The upper, electrical layer can perform multiplexing of different traffic streams into a single wavelength path (λ-path) or lightpath via simultaneous time and space switching. Similarly it can demultiplex different traffic streams of a single lightpath. Furthermore, it can perform re-multiplexing as well: some of the de-multiplexed traffic will be again multiplexed into some other wavelength paths and handled together along this other path. This is often referred to as traffic grooming and we will refer to it as grooming [7].

The question is how these layers can be operated together. Both IETF and ITU-T propose models and solutions how to operate these two (or more) layers together [21], [22]. We consider the case when both layers are handled via a distributed control plane to ensure full and joint on-line adaptivity of both of the layers. By using dynamic optical layer, it is possible to create an adaptive set of lightpaths that satisfies emerging traffic demands. Those two physical nodes that are connected by a lightpath are seen as adjacent by the upper layer. Multiplexing and demultiplexing the traffic of a lightpath is impossible by applying only optical devices. In these cases lightpaths have to be torn down, their traffic has to be taken up to the electronic layer that increases the number of lightpaths. In addition, the number of applicable opto-electronic converters per node is limited.

C. Traffic Engineering mechanisms

The traffic in the network domains is engineered in the control plane. The major goal of such a Traffic Engineering (TE) mechanism is the enhancement of the performance of an operational network, at both the traffic and resource levels. Traffic engineering optimizes the use of network resources to achieve specific goals, such as to avoid congestion, to minimize delay, to differentiate

services, etc. Most of these methods are valid on intra-domain level, because internal network information is typically limited outside of administrative domains [17]. Consider the domain as a closed entity, with a given traffic matrix. In such solutions the TE affects only the output, but it does not take into consideration the possibility to influence the input. [2], [3], [18], [19]. Even if it does, it considers that all domains – the one that provides the traffic, and the one that conveys the traffic – use the same TE method (e.g. Path Computation Element based TE - SPF) [4], [5].

3. Network Models

A. Reference network model

Our proposal is based on the presumption that both the aggregation and access domains are, or in the near future are expected to be, based on switched layer 2 (L2) technologies [12], which offer lower bit costs [13]. L2 switched networks deploy Spanning Tree Protocol variants (STP, MSTP, etc.) [14] to convey the traffic towards the core.

Fig. 1 gives a picture of the reference network model developed within the CELTIC TIGER2 project [15]. The reference network [16] reflects the view of major service providers and vendors on the evolution of networking infrastructure and the way it will assimilate the new technologies. As seen in *Fig. 1,* the networks are divided in three segments: the Access, the Metro and the Core. Depending of the country and/or local geographic specificities as well as the Internet Service Provider (ISP) choices, part of the sub-segments depicted in *Fig. 1* may be missing, but based on the current practices and medium-term forecasts, this generic model describes all networks.

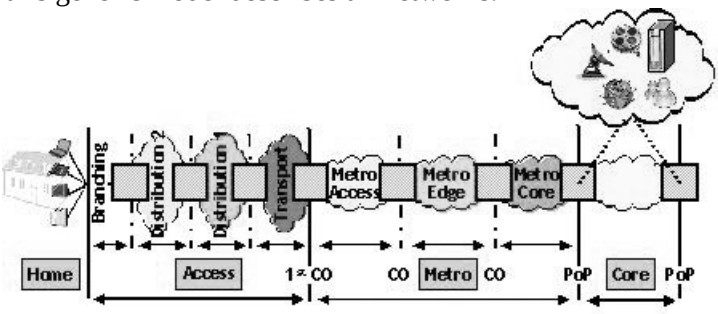


Figure 1: TIGER2 generic network reference model.

The Access network is a local area network, and is widely studied in the literature. It connects the end users to the first Central Office (CO). Typically they have a tree-based topology, which aggregates the traffic to the COs. Core

networks are also well studied and in this model we define it as the national or wider area domain. Typically they have a meshed topology. As seen in *Fig. 1*, the metro network, which links the access to the core, is split into three sub-segments. In legacy infrastructures, these sub-segments together form a hierarchy. Access areas may be connected to any metro sub-segment by COs, and each metro is connected to the core through a Point-of-Presence (PoP).

The roles of the sub-segments should be specified in the context of the deployed technologies. In this paper we assume that carrier-grade Ethernet-based L2 technologies become dominant not only in the aggregation, but also in the access [12]. Based on the above assumptions we obtain the reference network used in this paper, derived from the generic network model [16]. In this specific model the metro sub-segments use L2 switching in the access and edge, while the core deploys L2/L3 TE mechanisms. Thus, the first two sub-segments of the metro represent successive aggregation levels of the user traffic. In the metro-access, the first aggregation level, the traffic from multiple COs is aggregated in Concentration Nodes (CN). In the metro-edge, the second level of aggregation, traffic from different CNs is processed by a L3/L2 metro node, and the PoPs at L2/L3 boundary are handling several tens of thousands users. As a summary, we consider that the metro-access and metro-edge networks form an aggregation domain, while the metro-core segment is a meshed distribution.

B. Core network models

A specific core network model has been proposed [16], starting from current ring topologies, widely deployed in optical networks. The model is a Double Rings with Dual Attachments (DRDA) and it can be used in core networks. In such topologies two rings, (the inner and the outer metropolitan rings) are interconnected in such a way, that every node in the outer ring is directly connected with its associated node in the inner ring, via double links (dual attachment). These provide high connectivity and multiple back-up paths for restoration purposes while reusing current network fiber deployments.

C. Investigated network topologies

Based on the reference networks presented in the previous section we designed a network that was used for our simulation based investigations, and its topology is presented in *Fig. 2* (left). This network is divided in two main parts, an aggregation network using Multiple Spanning Tree Protocol (MSTP) [14] and a core part with Constraint based Shortest Path First (CSPF) TE [10].

The traffic sources are depicted on the left-most part of the figure, the aggregation network conveys the packets to the core network. At the boundary we have only three edges. In real-life networks the number of edges is kept as

low as possible for reasons of costs. The network has six destination nodes (sinks) represented by the exit points of the core network on the right side. The main function of the aggregation domain is to channel end-user traffic towards the core, thus its nodes are connected to two neighboring devices, at most.

The core domain has a meshed topology, with a 3 hop shortest distance between the ingress and the egress. The nodes of the core have a degree of connectivity of 3 or 4. This is a trade-off between cost effectiveness and the assurance of alternative paths. The aggregation domain uses Ethernet-switched technology, and the core uses WDM extended with an electronic control layer.

Apart from investigating the efficient network capacity usage and balanced load of the core domain, we also investigated the possibility to minimize the operations in the electronic layer and the usage of longer optical paths. These last two parameters are characteristics of the dual opto-electronic models.

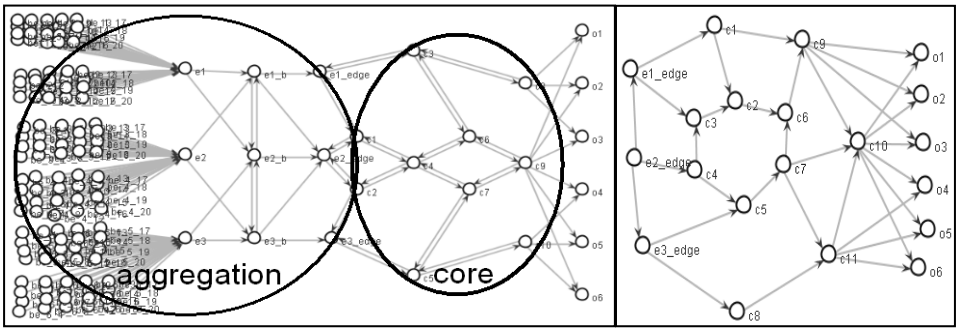


Figure 2: Topologies of aggregation and meshed core (left) and dual ring core (right).

Our proposal supposes that the domains have a control plane that apart of running TE and other control functions are capable of communicating/ cooperating with the control planes of the neighbouring domains. Such a control plane model is the Knowledge Plane [23] that can use MSTP in the aggregation and CSPF in the core domains.

We also investigated the behavior of the core if it deploys a dual ring topology (see *Fig. 2* - right). We have kept the edge nodes and the output nodes from the previous topology in order to use the same aggregation network and to be able to compare the two results. In the following we refer to the first topology as *meshed core*, while to the latter one as *dual ring core*.

4. Inter-domain TE cooperation

Our proposal is to use shared intelligence between control planes, where the core intra-network functions are unchanged and only the inter-network control planes co-operate which enhances the performance.

In *Fig. 2* the traffic reaches the core network through the aggregation domain. In case of any event (congestion on a link, link failure, etc.) the classical TE works with the assumption that the traffic matrix remains unchanged and it has to re-distribute the traffic volume relying on load redistribution inside the core. Our proposal is to use the Knowledge Plane and re-arrange the input traffic distribution outside the core edge routers. This means that – from the point of view of the core – we change the traffic matrix, since the load on the edges will be different.

Let us take the topology presented in the left-side of *Fig. 2*. Now in the situation when the aggregation domain directs all the traffic to the *e1_edge* (the "northern" one), while *e2_edge* (the "middle" edge) and *e3_edge* (the "southern" edge) do not feed any traffic to the core. This is the worst case situation to overload the core and corresponds to the situation when only the tree rooted in *e1_edge* is used to collect the traffic in the aggregation domain. Now, if we take the opposite situation, when we use each of the trees in the aggregation domain to forward the same amount of traffic, then the aggregation domain distributes the traffic evenly among the three ingresses. In this case all regions of the core will be evenly loaded.

It is the task of the Knowledge Plane to map the traffic sources among the trees. In our simulations we used small individual flow throughputs. Each tree is collecting such individual demands and the sum of these represents the traffic load at the edges. Practically the granularity of the traffic is small enough to allow us to finely balance the load. In what follows we will use the term load balancing as the operation of load redistribution in the aggregation domain as described above. The goal of load balancing will be to decongest a certain area of the core network with a minimal redistribution of the original load.

5. Simulation results

A. Traffic model

During the simulations the traffic flows originated from the sources have the same bandwidth. We considered that we know the traffic matrix and the paths in the core are computed by a PCE using CSPF protocol. Additionally we generated background traffic, as well, which enter the core at the edge nodes and sink on the most right-hand side destination nodes. The links of the core networks had 200 Mbps capacity, which defines the load region where the core network is congested, but not overloaded of 400 Mbps to 800 Mbps.

In our investigations we used the *e2_edge* node where we directed all the traffic and tried to serve it using CSPF. The resulting paths were called the *main branch*. If the demand is high enough, the traffic demand cannot be served. If

we apply our solution to this situation that means that some part of the traffic will be shifted to the other two edges, *e1_edge* and *e3_edge*. The paths that follow the flows entering on these two edges are called *secondary branches.*

We used the background traffic to "fill" the network up to the point where congestion might start to develop. We sent 200 Mbps background traffic on the main branch. Then we started to add new traffic demands until we reached the total one, which was set differently from case to case: all our simulations were run with the 500Mbps, 600Mbps, 700Mbps and 800Mbps total traffic demand. These are the situations when we can test the usefulness of our proposal and evaluate its impact on the efficiency of the opto-electronic core transport.

We used a flow level simulator, already used for the research of opto-electronic networks [24]. We generated individual flows, and the sum of these demands resulted in the overall traffic demand. Each link was divided into lightpaths of 10 Mbps capacity. This results in 20 lightpaths within each link that offers enough flexibility for multiplexing the flows within the core. Based on earlier work with the simulator we opted for 12 individual flows per lightpath, resulting in a flow capacity of 0.83 Mbps.

Within each scenario – that is for different overall traffic demands – we have simulated several sub-cases, where the load of the main branch was gradually re-distributed among the secondary branches. At first we started with the situation where 30% of the traffic was entering at edges *e1_edge* and *e3_edge* (15% on each of them). From there on, we stepwise directed more and more traffic towards to the secondary branches while the network was able to carry the traffic without loss. In order to be sure on that, we also simulated the next step following this point. The individual flow demands were scheduled randomly. For each situation we run ten simulations and averaged the results.

B. Spanning Trees in the aggregation domain

In order to get the input traffic at the ingresses of the core domain, we had to build the trees that bring the traffic to the edges of the core. For this, we had to build the set of trees that form the basis of the MSTP operation. We used a combination of the TOTEM [25] and BridgeSim [26] tools to simulate these trees.

With the combination of these two tools we could use the topologies created in TOTEM and apply the STP formation protocol implemented in BridgeSim. We generated all the spanning trees that can potentially be used in our scenarios, that is, all the trees that are rooted in one of the three edges and the traffic sources are their leaves. *Fig. 3* presents the tree generated for the northern edge, the other two trees have a similar shape. The actual traffic distribution among the three trees is decided according to our solution and should be enforced by the Knowledge Plane, as mentioned earlier.

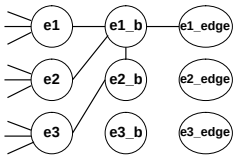


Figure 3: The STP rooted in node e1_edge.

We did not investigate the behavior (delay, blocking, packet loss, etc.) of the aggregation domain, only used it to generate the MSTPs and determine the input traffic for the core domain. In what follows we will be interested only about the traffic distribution among the three ingress nodes.

C. Balancing the load in the core

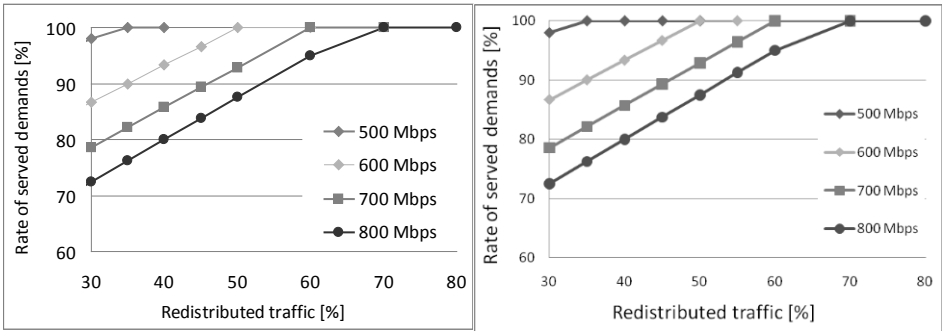


Figure 4: The successfully served flow demands as the function of traffic re-distribution for the meshed core (left) and the dual ring core (right).

The left-hand side of *Fig. 4* presents the ratio of successfully served traffic demands in the meshed core. It can be seen that 500 Mbps total traffic will be served if we redirect 35% of the traffic on the secondary branches. The traffic volumes that must be redirected to achieve a loss-free ratio for the 600Mbps, 700Mbps and 800Mbps traffic scenarios are 50%, 60% and 70%, respectively. These results confirm that if we redistribute the traffic before it hits the core edges, we can balance the core load, thus it is a viable mechanism to actively increase the efficiency of the core traffic engineering process.

We achieved similar results for the dual ring topology, as well (right-hand side of *Fig. 4*). The congestion-free core is achieved for the redistribution of the 35% of traffic for the 500Mbps case and 70% for the 800Mbps (worst) case.

D. Dual opto-electronic model

In the following, we present our simulation results on the effect of our proposal on the efficiency of the opto-electrical dual transport core network.

First we explain the results using the meshed core. The first parameter is the number of lightpaths. We can see in Fig. 5 (on the left) that as the rate of successfully served traffic demands is rising, but is still below 100%, the number of lightpaths is increasing. This is due to the fact that more and more individual flows are in the network and these are following new (alternative) routes. Thus, the increase of this parameter is not a consequence of the decreasing efficiency but of the growth of the core utilization.

This trend is reversing if we keep redistributing the traffic even after all the traffic reaches its destination. This corresponds to the situation depicted in Fig. 4 by the dots on the 100% line. As we already mentioned, for each overall traffic load scenario we simulated two cases when all the demands were served by the core: the "break-even" point and one following step where we increased the traffic redistribution by 10%. The results obtained for these cases are encircled in Fig. 5 and as we can see the number of paths is decreasing. The more the core is loaded, the more these trends are accentuated, therefore it can be seen the best result on the curve corresponding to the highest loads.

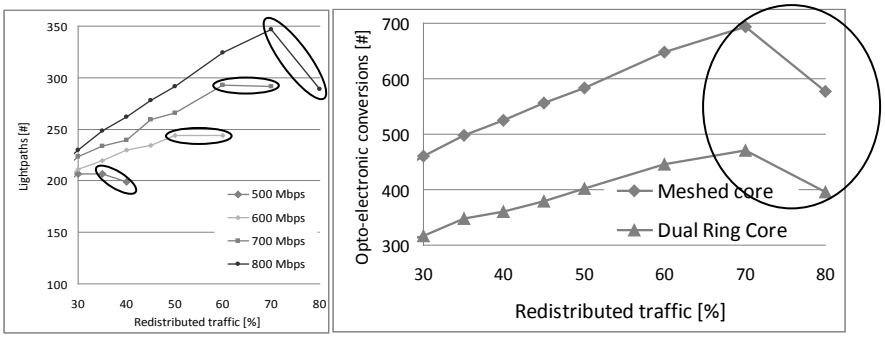


Figure 5: The efficiency of the lightpath management. Number of lightpaths for meshed core (left) and number of opto-electronic conversions (right).

If the primary goal is the minimization of the operations in the electrical layer, the best option is to distribute the traffic. The drawback of this solution is that we have to use the alternative branches. In regular operation the alternative paths are associated with higher costs. Therefore it is the decision of the operator depending on its business model to find the balance between the path costs (that usually are translated into financial costs) and the efficiency of the opto-electronic layer (higher-layer delays are translated into worse QoS).

The right-hand side of *Fig. 5* presents the number of opto-electronic conversions done in the core (values that yielded 100% success rate are highlighted with a circle). We have plotted in the same graph the results for both the meshed core and dual ring core. These conversions can happen only in

the nodes that make a grooming operation and multiple conversions may happen in such a node. Based on our simulations there are several hundreds of such conversions per node. The trend observed for the number of the lightpaths is valid also here, for both core topologies.

6. Conclusion

This paper proposed a traffic management solution that improves the performance of the core network. The aggregation network is supposed to deploy L2 switched technologies, while the core network will use WDM in combination with GMPLS. We have prepared a meshed and a dual ring topology following the principles of the TIGER2 project's reference network and investigated our proposal by means of simulations.

We have shown that if congestion occurs in the core, we can eliminate the congestion just with a proper coordination between the control planes of the aggregation and core domains, redistributing the traffic prior entering the core. We deployed MSTP protocol in the aggregation domain and the traffic redistribution was done using these spanning trees. This solution increases the ratio of successfully served traffic demands, increasing the utilization of the core. The traffic redistribution at the aggregation has positive effects even if there is no congestion in the network, because in such cases it increases the efficiency of the opto-electronic transport layer.

As a conclusion we can say that the cooperation of the control layer of the aggregation and core domains has multiple advantages. In the future we plan to investigate the trade-off between the cost of load balancing and opto-electronic efficiency.

Acknowledgements

This work has been partially funded in the framework of the CELTIC TIGER2 project (CP5-024) as part of the EUREKA cluster program.

References

[1] Osbourne, E., Simha, A., "Traffic Engineering with MPLS", Cisco Press, Indianapolis, ISBN 978-1-58705-031-2, 2003.

[2] Fortz, B., Rexford, J., Thorup, M., "Traffic engineering with traditional IP routing protocols", *IEEE Comm. Magazine*, vol. 40, no. 10, pp. 118-124, 2002.

[3] Dasgupta, S., de Oliveira, J. C., Vasseur, J.-P., "Dynamic traffic engineering for mixed traffic on international networks: Simulation and analysis on real network and traffic scenarios", *Computer Networks*, vol. 52, no. 11, pp. 2237-2258, 2008.

[4] Casellas, R., Martinez, R., Munoz, R., Gunreben, S.,"Enhanced Backwards Recursive Path Computation for Multi-area Wavelength Switched Optical Networks under Wavelength Continuity Constraint", *Journal of Optical Communications and Networking* (*JOCN*), vol. 1, no. 2, pp. A180-A193, 2009.
[5] Ho, K-H. et al, "Inter-autonomous system provisioning for end-to-end bandwidth guarantees", *Comp. Commun.*, vol.30, no. 18, pp. 3757-3777, Dec. 2007.
[6] Sabella, R., Zhang, H., eds.: "Traffic Engineering in Optical Networks", *IEEE Network*, vol.17, no. 2, pp. 6-7, 2003.
[7] Cinkler, T., "Traffic- and λ-Grooming", *IEEE Network*, vol. 17, no. 2., pp. 16-21, 2003.
[8] Liu, K. H., "IP Over WDM", John Wiley & Sons Inc., ISBN: 978-0-470-84417-5, 2002.
[9] Mukherjee, B., "Optical WDM Networks", *Optical Networks Series*, Springer, ISBN: 978-0-387-29055-3, 2006.
[10] Ziegelmann, M., "Constrained Shortest Path and Related Problems: Constrained Network Optimization", VDM Verlag Dr. Müller, ISBN 978-3-8364-4633-4, 2007.
[11] Lee, Y., Mukherjee, B., "Traffic engineering in next-generation optical networks", *IEEE Comm. Surveys and Tutorials*, vol. 6, no. 1-4, pp. 16-33, 2004.
[12] Fang., L., Bitar, N., Zhang, R., Taylor, M., "The Evolution of Carrier Ethernet Services: Requirements and Deployment Case Studies", *IEEE Comm. Mag.*, vol. 46, no. 3, pp. 69–76, 2008.
[13] Occam Networks whitepaper, "Switching Versus Routing in Access Networks", http://www.occamnetworks.com/pdf/SWITCH_VS_ROUT_WP_FINAL.pdf, May, 2007.
[14] Caro, L. F., Papadimitriou, D., Marzo, J. L., "A performance analysis of carrier Ethernet schemes based on Multiple Spanning Trees", *VIII Workshop in G/MPLS networks*, Girona, Spain, Jun. 2009.
[15] CELTIC TIGER2 project homepage, http://projects.celtic-initiative.org/tiger2/
[16] Dorgeulle, F., "Rationales and scenarios for investigations on next generation of access, backhauling and aggregation networks", CELTIC TIGER2 public report D20, Nov. 2009.
[17] Awduche, D. et al., "Overview and Principles of Internet Traffic Engineering", RFC3272, May 2002.
[18] Feamster, N., Borkenhagen, J., Rexford, J., "Guidelines for interdomain traffic engineering", *SIGCOMM Comput. Comm.*, Rev. 33, pp. 19-30, 2003.
[19] Vigoureux, M., et al., "Multi-layer traffic engineering for GMPLS-enabled networks", *IEEE Comm. Mag.*, 0163-6804 vol. 43 (7), pp. 44–50, 2005.
[20] Mukherjee, B., "Optical Communication Networks", McGraw-Hill, ISBN 978-0-070-44435-5, 1997.
[21] Chiu, A., et al., "Unique Features and Requirements for The Optical Layer Control Plane", IETF internet draft, work in progress.
[22] International Telecommunication Union, "OTN – ITU-T Recomm. on ASTN/ASON Control Plane", http://www.itu.int/ITU-T/2001-2004/com15/otn/astn-control.html.
[23] Clark, D., Partridge, C., Ramming, J. Ch., Wroclawski, J. T., "A knowledge plane for the internet", *ACM SIGCOMM 2003*, Karlsruhe, Germany, pp. 3-10, Aug. 2003.
[24] Hegyi, P., Cinkler, T., Sengezer, N., Karasan, E., "Traffic Engineering in Case of Interconnected and Integrated Layers", *IEEE Networks*, Budapest, Hungary, pp. 1-8, Sept. 2008.
[25] Homepage of TOTEM simulator, http://totem.run.montefiore.ulg.ac.be/features.html.
[26] Homepage of BridgeSim simulator, http://www.cs.cmu.edu/~acm/bridgesim/index.html.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 99-113

Energy-Efficient Networking: An Overview

László SZILÁGYI, Tibor CINKLER, Zoltán CSERNÁTONY

Budapest University of Technology and Economics,
Dept. of Telecommunications and Media Informatics,
e-mail: szilagyi@tmit.bme.hu; cinkler@tmit.bme.hu; csernatony@tmit.bme.hu

Manuscript received June 30, 2010; revised October 30, 2010.

Abstract: Recently – for economical and ecological reasons –, energy-efficiency has been receiving an emerging attention from both industrial and fundamental researchers. Given the large-scale growth and resource overprovision of communication networks, the issues related to energy consumption get more significant than ever. In this paper, we provide an overview of some of the latest contributions and most important trends on energy-efficient networking. Since today's networks are heterogeneous to a large extent, different concepts and aspects of networking are discussed in separate sections.

Following the introduction of the fundamental discipline of energy-proportional computing (and also its relationship with resource over-provisioning), we give an overview on the power-saving opportunities of core networks with focusing on two of the most-widely applied techniques for energy management. A comparison of circuit and packet switched networking approaches is also made, along with some considerations taken for the cases in which electrical and optical switching are employed. In case of access networks, in addition to considering the commonly used landline connections, we give a brief description about the popular handoff mechanisms for wireless networks. Finally, we consider energy-saving opportunities on data centers and take a short overlook on cloud computing. Despite all the evident differences, being between distinct segments of networking, some of the methods applied at distinct networking technologies show considerable similarities with each other due to the fact that some of the occurring problems exhibit similar patterns in terms of modeling and problem formulation.

Keywords: energy-efficiency, energy consumption, green networking.

1. Introduction

With the explosive growth of communication networks, energy consumption has risen to a major economical (operational expenditures) and ecological (CO_2 emission) concern in the past few years. About 2 percent of the total CO_2 emission is produced by the Information Technology and Communication sector (ICT) which is more than the contribution of the whole aviation industry [1]. A recent study puts more emphasis on this issue by showing that the rise of energy consumption of large communication systems corresponds to Moore's law [2]. Therefore, power consumption has become a critical factor of communication networks, IT facilities, data centers and high performance network elements. Energy-efficient designing helps cutting the Operating Expenses (OPEX) as well [3], [4], and in addition to that, it also might result in more reliable network elements (due to the decrease of heat dissipation).

In order to save energy in communication networks, first, we have to reveal the reasons of energy wastage in existing systems. Energy inefficiency might come from architectural (SW related) and physical design (HW related). From the energy-efficiency point of view, the most important feature of networking is underutilization. While networks are generally designed to handle peak-time traffic, most of the time, their capacity remains (heavily) unexploited. This is called *over-provisioning*. According to [1], the magnitude of network utilization is 33% for switched voice, 15% for internet backbones, 3~5% for private line networks and 1% for LANs, while the energy consumption of network equipments remains substantial even when the network is idle. A rather physical related issue is that the energy consumption of network elements is not proportional to their utilization, i.e. energy cost is a function of capacity, not throughput. These facts result in high energy wastage.

Today's networks are mostly *configured statically*, running at full performance all the time which is not necessary. In order to achieve higher energy-efficiency, network management methods need to be able to dynamically adopt network characteristics to the actual traffic demands during operation. Switching off underutilized (or idle) parts of the network and dynamically adapting transmission rates (with satisfying certain QoS constraints) are ultimately important approaches of designing a greener network. In order to make energy-aware management possible, network elements also should support these features. On-demand frequency-scaling of CPUs and data storage modules and network interfaces with adjustable transmission rates (rate-adaptation support) are all mandatory for attaining greener network elements.

Finding efficient ways for *cooling* network equipments is also a big challenge. In case of data centers, roughly 50% of electricity is consumed by the

cooling infrastructure, the other 50% used for computing [5], [6]. Increasing cooling efficiency and using alternative cooling methods have a huge contribution to the electric bill of equipment rooms. Employing alternative energy sources (e.g. solar, wind) for supplying network nodes (base stations) is also a matter of interest nowadays.

In this paper, we provide an overview of the latest results concerning energy-efficient networking, discussing the different functional parts of the ICT infrastructure separately. In *Fig. 1*, the estimated share in energy consumption by different areas of ITC can be seen [7]. Energy-efficiency is examined from an operator's point of view with focusing on networking infrastructure and data centers, but leaving PCs, monitors, printers, and other user equipments out of consideration.

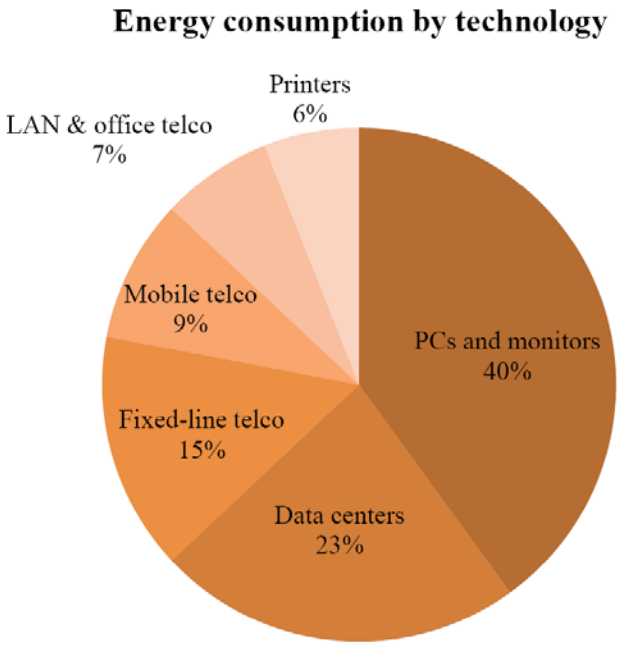


Figure 1: The estimated energy consumption shares of different functional parts of ITC [7].

The paper is organized as follows. Section 2 explains the importance of power consumption dynamic range. In the remaining sections, we provide a closer look at the issues with different areas of communication networks: Section 3 discusses some of the most important features of core networks, while Section 4 and 5 do the same for access networks and data centers, respectively. Finally, in Section 6, we conclude our experiences and reveal a few important directions for future solutions.

2. The dynamic range of power consumption: Rate adaptation versus switching off components

Rate-adaptation is highly important in energy-aware data transmission and data processing for the following reason. For transmission power-rate function is a relationship that gives the required amount of transmission power for a certain rate. Keeping the bit-error probability fixed, the required power is a convex function of the rate for most encoding schemes. This results from Jensen's inequality which states that transmitting data at lower rates and over longer time intervals has less energy cost compared to faster rate transmissions [8]. For data processing variable adaptive rate control means adjusting CPU frequency for the required performance in order to save power.

The *dynamic range* of power consumption is one of the most important attributes of network elements: it is the relative difference of energy consumption of an element between 0% and 100% of relative utilization. For example, study [6] shows that the utilization–power usage characteristics of a system (consisting of servers) shown in *Fig. 2* has a smaller (50%) dynamic range compared to the one on *Fig. 3* (90%). As suggested by the results, given the typical utilization region (10-50%), a 50% save can be achieved in terms of energy consumption when having a dynamic range of 90%. Consequently, network component vendors should increase the dynamic range of their equipments as much as possible. The dynamic range of 100% addresses the case of *energy-proportional computing*.

However – even when applying rate-adaptation –, network components with low dynamic range still tend to consume a considerable amount of energy when being not utilized at all (if the offset of power usage is too high). Even so, in many cases (apart from rate-adaptation), power consumption can be further reduced by *switching off* certain underutilized (or idle) components. The price to pay here is the potential occurrence of additional delay, however – due to the typically high power on consumption – transition between active and passive operation mode cannot be done arbitrarily often.

A real challenge for these solutions is finding a method to decide whether a network element should operate or not, and if so, then determining the operational rate in specific moments.

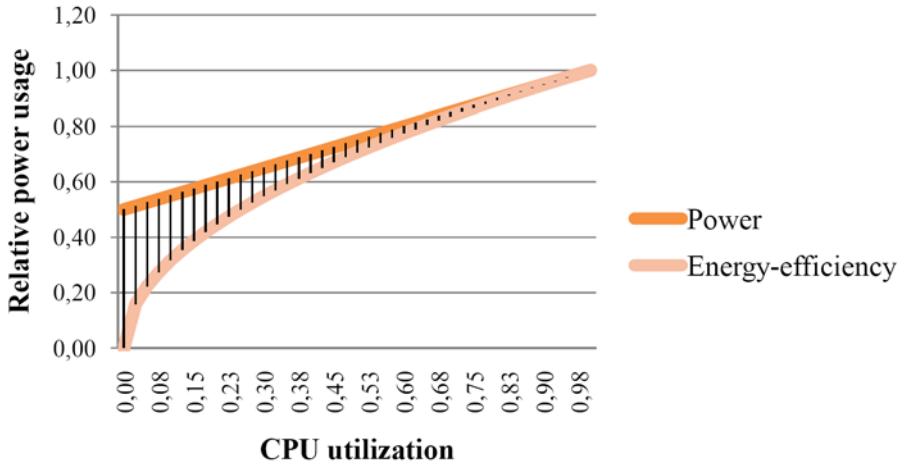


Figure 2: Relative power usage and energy-efficiency for a typical data center scenario.

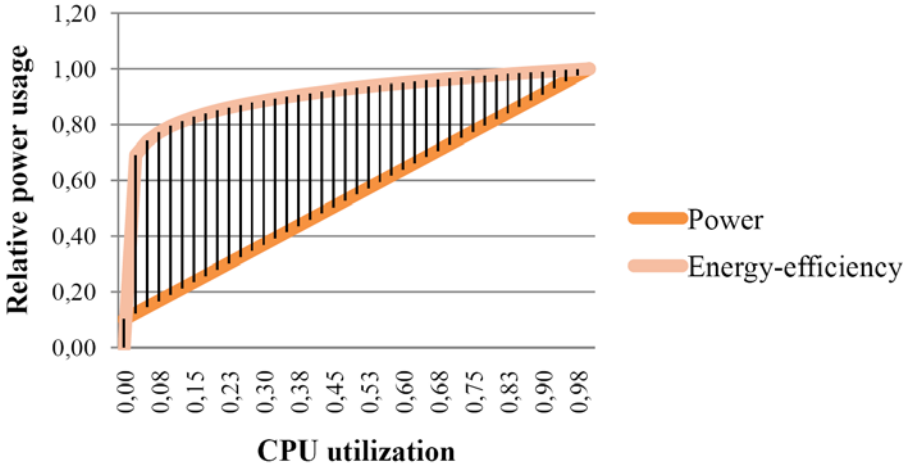


Figure 3: Relative power usage and energy-efficiency for a data center scenario with larger dynamic range [6].

3. Core networks, packet versus circuit-switching

The backbone infrastructure of a communication network transferring aggregated data flows is called a *core network*. Given that the typical rate of utilization shows a considerable variation within a single day, and that (as stated previously) networks are provisioned to handle peak traffic, networks tend to be *underutilized* for most of the time.

An opportunity for energy conservation method is to shut down idle network elements. Since the transition between active and sleeping states requires additional energy and time, inactive periods along with the routing has to be scheduled properly in order to preserve power without exceeding a given value of delay. In [9], a frame-based periodic scheduling is introduced. The network elements work either at a full or at a zero rate. The lengths of active periods are proportional to the traffic loads of network elements while periods cannot be arbitrarily small since transitions consume extra power, however too long periods result in increased delays. In [2], the network is controlled on the basis of traffic statistics: the day is sliced into one-hour long “network configuration periods”, and the state of the system is formalized as a “linear programming problem” with the daily operation scheme being determined to satisfy the specific QoS constraints.

An important feature of core networks is that multiple paths might exist between distinct nodes of them. This fact allows us to distinguish between nodes to determine the ones that can be switched off during low backbone utilization periods. Many high capacity links within a core network actually are compositions of smaller capacity ones. In [10], a method is demonstrated for shutting off cables in bundled links for lower traffic periods. The problem of determining the sublinks to shut down is formulated as an Integer Linear Program (ILP) that is proven to be NP-complete. Given this fact, for tackling this intractability, efficient heuristics have to be proposed. In [10], a fast greedy heuristic solution suggested for a dedicated multicore server is proven to be feasible. Optimization cycles becoming 8 seconds to 50 minutes long (depending on the actual topology) mean a dynamic power-saving solution.

Underutilization means that a link or a network element does not need to operate at its full capacity to satisfy the corresponding QoS constraints. Upon this fact, rate adaptive routers, switches and links can be employed in order to save energy [2], [11], [12]. A suitable hybrid application of element switching-off and rate-adaptation is a promising solution [11].When multiple transmission lines form a single large-capacity link, *rate-adaptation* can be the combination of operations: switching off transmission lines, and tuning the transmission rate of individual lines (width control) [13].

In [14], core networks are examined for their power-efficiency from a different perspective. Today's backbone network infrastructures are composed

by optical links between large distance nodes with the processing and switching of traffic being performed mostly in the electrical domain. While the energy consumption of electrical components is getting lower by every year, large reduction could be achieved in terms of power consumption if the whole processing and switching is executed in the optical domain. Although, optical switching has become a reality (MEMS, CMOS), IP, nowadays' dominant packet switched transmission technology, requires random access memory for buffering which is yet to be implemented purely optically. Large fiber delay lines are not practical as they require power-consuming signal regenerators, not to mention the impractical size of them.

From an energy-efficiency point of view, it is important to note that – despite the connectionless nature of IP – above 90% of the traffic within backbone networks is transported via the connection oriented Transmission Control Protocol (TCP) [14] (consider applications such as IP television, voice over IP, video conferencing or online gaming services, all those by which a very high quality of service is required).

As being shown in *Fig. 4*, within a node of an IP network, more than half of the energy is consumed by the traffic processing and forwarding engine (TP/FE).

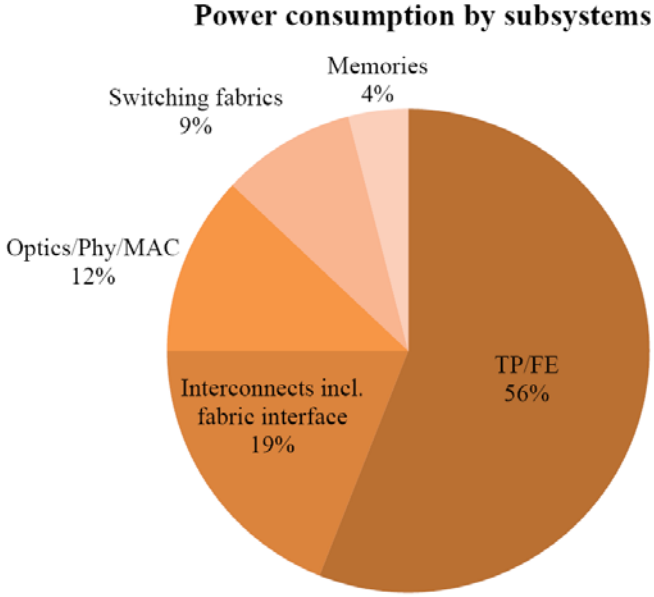


Figure 4: An estimation of power consumption by subsystems (the efficiency of power conversion is not considered) [14].

On the contrary, the implementation of all-optical, circuit-switched nodes does not require complex traffic processing and large memory units (within the electrical domain).

In contrast, all-optical circuit-switched nodes whose implementation doesn't require complex traffic processing, large memories in the electrical domain and optical to electrical to optical conversions consume less than 10% of the power consumed by ordinary optical packet-switched architectures using electronic traffic processing. Similarly, electronic circuit-switched designs consume 43% less power than packet-switched one. Taking all these into account, future energy-efficient core networks nodes might use circuit switching in the core, along with packet switching edge nodes in order to reduce complexity, thus power consumption.

4. Access networks

Since most of the physical elements are located in the segment of access networks, the energy saving by each type of elements is multiplied by a large factor. This makes an important contribution to the reduction of total consumption [15].

Nowadays, the most widely deployed technology for broadband *landline connections* still employs copper lines for bearers. With the continuous increase of bandwidth requirements, broadband penetration and bandwidth demands, more energy is required than ever. Although new transmission technologies, such as VDSL2, allow higher speeds, they induce increased complexity and power consumption [16]. By today's networking technologies, serving high-bandwidth demands together with sustainable energy consumption can only be achieved through progressive optical fiber deployment in FTTCab, FTTB (and also in the longer-term FTTH) architectures which are expected to shorten the copper access network and to boost the overall performance of xDSL systems. The deployment of such systems however requires certain capital expenses to be involved which makes this technological shift a gradual one. Dynamic Spectrum Management [17] and energy-aware solutions of such kind can make copper technologies more sustainable, but they only yield the industry little additional time for making the required technological shift towards optical networks.

Mobile operators with radio access networks (2G, 3G etc.) have to provide services for very large physical areas and for several subscribers as well [2]. Given the necessity for several base stations, operating them requires a large amount of energy. As mobile broadband is expanding rapidly, the energy-efficiency of radio access networks is expected to receive even more significant attendance in the future.

As presented in *Fig. 5*, base stations take most of the energy requirements of cellular networks [18]. For that reason, in the remaining parts of this section, we focus on the issues of radio base stations.

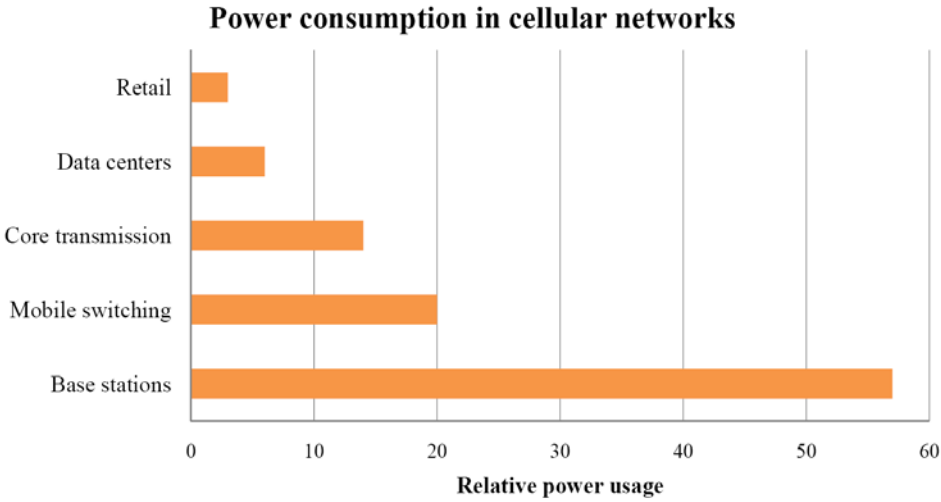


Figure 5: Power consumption in cellular networks [18].

Since radio access networks are also dimensioned for peak traffic loads, a large amount of energy is being wasted due to underutilization. The reduction of traffic level in some portions of a cellular network comes from the combination of two effects: first, the characteristics of the typical day-night behavior of users, and second, the daily swarming of users carrying their mobile terminals around different kinds of areas (residential, office districts, etc.). On the one hand, this induces the need for a large capacity in all areas at peak usage times, on the other hand, it reduces resource requirements during lightly occupied time intervals (day for residential and night for office districts) [15].

Energy can be saved by switching off certain underutilized cells. When some cells are switched off, we assume that radio coverage and service provisioning can be taken care of by the cells which remain active, possibly with a smaller increase in the emitted power, in order to guarantee the availability of service over the whole area. This energy-saving approach assumes that the original network dimensioning is essentially driven by traffic demands, as it normally happens in metropolitan areas, comprising a large number of small cells [15]. In regular cellular topologies, around 25-30% of the energy is possible to save by dynamically powering off active cells during under-utilized (low-traffic) periods [15], [19]. *Fig. 6* illustrates the method itself.

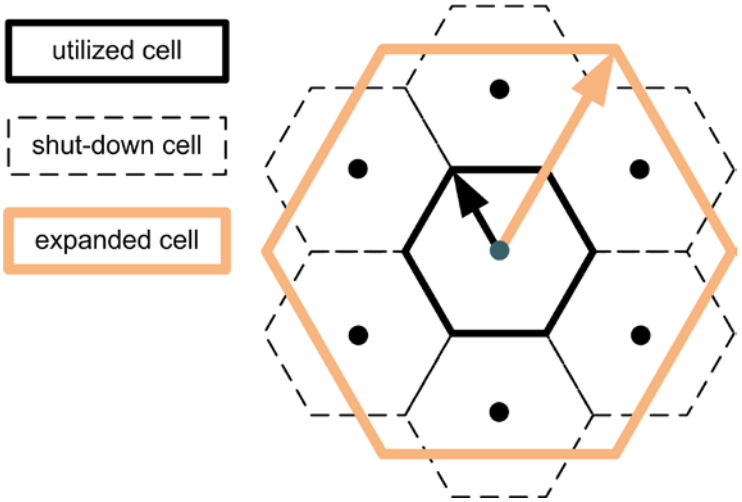


Figure 6: Dynamically powering off cells in cellular networks.

Rate-adaptation is an applicable procedure for wireless links too. Modern wireless devices are equipped with rate-adaptive capabilities which allow the transmitter to adjust its transmission rate over time. This can be achieved in various ways that include adjusting the power level, symbol rate, coding scheme, constellation size and various combinations of these approaches. In order to satisfy QoS requirements, existing systems can change transmission power/coding scheme dynamically depending on the noise floor. If large bandwidth is not required, “low power” transmission over longer time periods might also fit for the actual application [8].

In many areas, multiple access technologies are available. Matching bandwidth requirements with available access technologies, while taking energy-efficiency into consideration can save energy. There are two types of handoffs considered: vertical and horizontal handoffs. *Vertical handoff* is an interworking technique between different networking technologies whose aim originally is to assure the best connectivity to applications for mobile terminals by providing a transparent and seamless roaming between systems of different technologies. *Horizontal handoff* means the switching within the same technology (or layer). Initially, handoff decisions were aimed to be made on the basis of a lot of different factors such as QoS, cost-of-service, etc. [20]. Lately, there has been revealed that by suitably involving energy consumption into the objective (function) of the handoff, energy-efficiency aspects can also be taken into consideration [21], [22], [23], [24]. In *Fig. 7*, an illustration of horizontal and vertical handoff of two base station cells and a WLAN cell can be seen.

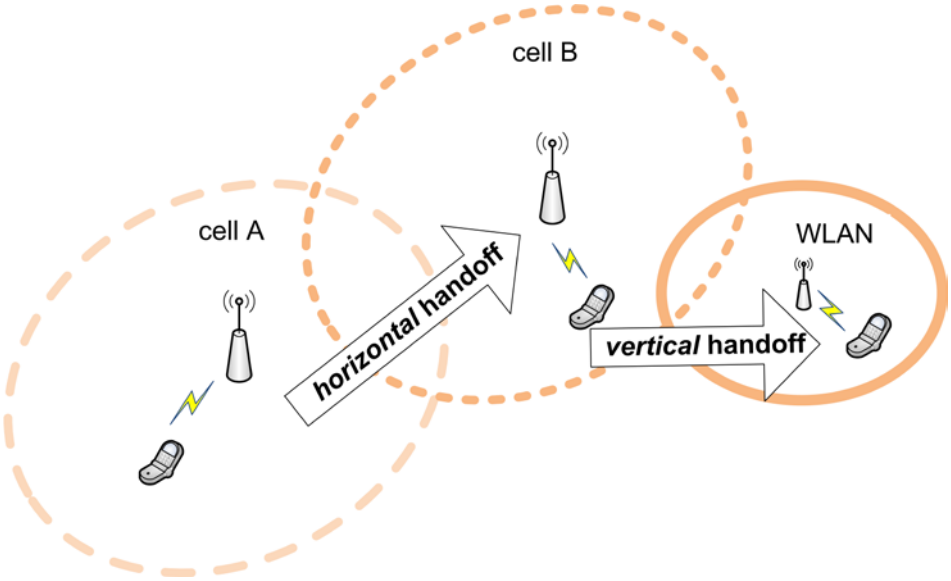


Figure 7: An illustration of vertical and "horizontal" handoff mechanisms between cellular networks and WLAN.

Base stations, exploiting alternative energy resources, already exist [25]. Employing alternative energy resources such as sun and wind for supplying base stations make mobile communication "greener", and in the same time these more sustainable sources ease the deployment of BSs far from the electric grid.

5. Data centers and cloud computing

The majority of networked services are being hosted in data centers. With the rapid growing of the demand for such data-intensive services (e.g. cloud computing, video broadcasting and online social networking), the employment of large *data centers* is of emerging significance. By now, they have evolved to networks consisting of tens of thousands of high data-rate servers. Such systems require special design and management approaches to provide the demanded robustness, performance and reliability.

As indicated in *Fig. 8*, according to [6], servers in a typical large system are being under-utilized in most of the time. As presented in *Fig. 5*, the power consumption is still significant in case of data centers. The problem is with the power consumption dynamic range, as stated above. CPUs are easily tunable, but they are not the major consumers. HDDs and even memory modules are much harder to manage from the aspect of power consumption. Switching off

underutilized servers is not feasible in a large scale for servers, as applications and data are usually distributed over numerous machines. Switching off components has too large a penalty, for HDD-s latency becomes in order of 1000 times bigger, and the spin up consumption consumes energy which can be saved in minutes of being off. In order to increase the efficiency of servers, their utilization should be maximized, while machines should be created with as wide dynamic power range as possible (energy proportional computing). Provided the dynamic range indicated in *Fig. 2* could be achieved, 50% of the energy could be saved compared to servers with characteristics shown in *Fig. 1*.

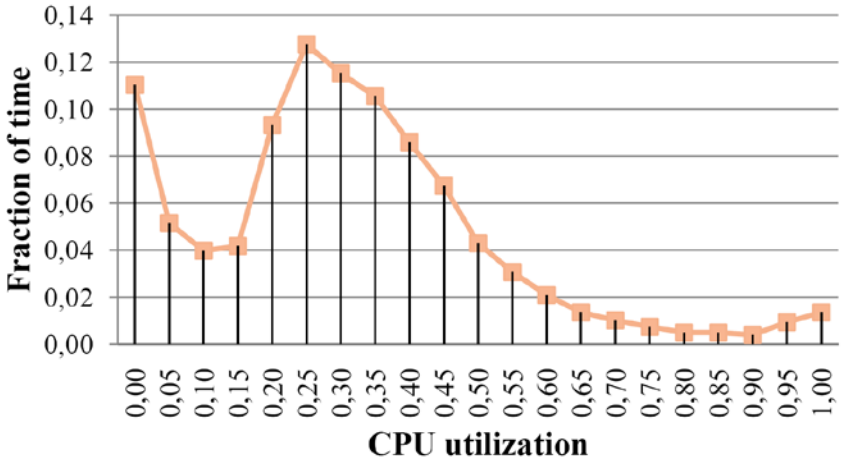


Figure 8: CPU utilization in a typical large-scale system [20].

The architecture of a data center influences the power consumption. Different architectures for a given number of servers mean different links and switching fabrics, different ways of connecting them to each other. The applied links and switching fabrics determine both the throughput and energy consumption. However, more sophisticated structures might provide bigger throughput, smaller delays, higher access speeds, but also require more power. Architects have to make a compromise between throughput and power consumption [26]. Today's popular symmetric architectures (BCube, DCell, Fat-tree, Balanced tree) are quite rigid and make it hard to apply them for the given constraints (throughput, power). Many systems are unnecessarily over-provisioned due to this factor wasting much power. Future architectures should employ more asymmetricity to provide more fine-grained datacenter designs.

It is important to note that the concept of cloud computing itself has got a large potential to make the whole IT industry greener [27]. Concepts like Software as a Service (SaaS) or Infrastructure as a Service (IaaS) helps

companies and governments not only to reduce cost with optimized resource usage, but to keep their business greener. Cloud computing service providers have large scale data centers and server farms and provide computing services over the Internet making companies having neither to invest in their own server parks nor to worry about over-provisioning their systems. The capital expenditure of a startup company can be restricted to investing in thin clients to access the demanded services. Although such infrastructures might be more energy-efficient compared to the legacy approach (i.e. ordinary companies providing a single PC for every employer while operating a more or less over-provisioned server park), large scale data centers still tend to be more energy-efficient.

6. Summary

In this paper, we have revealed a number of possibilities for saving power in different parts of the network. We have started the discussion with presenting the general concept that the over-provisioning of the resources results in suboptimal energy-efficiency by making the energy consumption of the network disproportional to its utilization. In the following, we have given a review on the various power-saving opportunities of underutilized network elements for core networks via sleeping and rate adaptation. Moreover, it has been shown that by deploying circuit-switched all-optical networks, it can be capitalized on that the vast majority of Internet traffic is transported via the connection-oriented TCP. Afterwards, access networking has been discussed: among landline connection technologies, optical-based FTTx solutions seem to be promising, while for wireless networks, horizontal and vertical handoff mechanisms have been presented for adjusting range of radio communication. Finally, energy-efficient ways for operating data centers and the concept of cloud computing have been briefly overviewed.

Based on the state-of-the-art, we suspect that the biggest challenge will be the effective control and cooperation of network components of varying energy-awareness. Today's large-scale communication networks are of great heterogeneity in terms of the employed networking technologies. Consequently, network management software has to handle different networking equipment types and generations at the same time. For both complexity and heterogeneity reasons, there is a fundamental need for a shift to be made from centralized solutions towards a distributed approach. Designing future's energy-efficient network components and architecture needs strong inter-technological cooperation to match HW capabilities with management techniques.

References

[1] Nordman, B., "Energy Use and Savings in Communications", in *Proc. of IEEE ICC*, 2009, Keynote.

[2] Bolla, R. et al., "Energy-Aware Performance Optimization for Next-Generation Green Network Equipment", in *Proc. of ACM SIGCOMM, Workshop on Programmable routers for extensible services of tomorrow*, 2009, pp. 49-54.

[3] Qureshi, A., Weber, R., Balakrishnan, H., Guttag, J. and Maggs, B., "Cutting the Electric Bill for Internet-Scale Systems", in *Proc. of ACM SIGCOMM, Conference on Data communication*, 2009, pp. 123-134.

[4] Odlyzko, A. M., "Data networks are lightly utilized, and will stay that way", *Review of Network Economics*, vol. 2, no. 3, pp. 210-237, 2003.

[5] Fan, X., Weber, W. D., Barroso, L.A., "Power Provisioning for a Warehouse-Sized Computer", in *Proc. of ACM International Symposium on Computer Architecture*, 2007, pp. 13-23.

[6] Barroso, L. A. and Hölzle, U., "The Case for Energy-Proportional Computing", in *Proc. of IEEE Computer*, 2007, pp. 33-37.

[7] Kumar, R. and Mieritz, L., "Conceptualizing 'Green IT' and data centre power and cooling issues", *Gartner Research Paper*, No. G00150322, 2007.

[8] Zafer, M. A., Modiano, E., "A calculus approach to energy-efficient data transmission with quality-of-service constraints", in *Proc. of IEEE/ACM, Transactions on Networking*, vol. 17, issue 3, pp. 898-911, 2009.

[9] Andrews, M., Anta, A. F., Zhang, L., and Zhao, W., "Routing and scheduling for energy and delay minimization in the powerdown model", in *Proc. of IEEE INFOCOM*, 2010, pp. 21-25.

[10] Fisher, W., Suchara, M., and Rexford, J., "Greening backbone networks: reducing energy consumption by shutting off cables in bundled links", in *Proc. of ACM SIGCOMM, Green Networking*, 2010, pp. 29-34.

[11] Nedevschi, S., Popa, L., Iannaccone, G., Ratnasamy, S., and Wetherall, D., "Reducing Network Energy Consumption via Sleeping and Rate-Adaptation", in *Proc. of 5th USENIX Symposium on Networked Systems Design and Implementation*, 2008, pp. 323-336.

[12] Bolla, R., Bruschi, R., Davoli, F., Ranieri, A., "Performance Constrained Power Consumption Optimization in Distributed Network Equipment", in *Proc. of IEEE ICC, Workshop on Green Communications*, 2009, pp. 1-6.

[13] Kant, K., "Power Control of High Speed Network Interconnects in Data Centers", in *Proc. of IEEE ICC*, 2009, pp. 145-150.

[14] Aleksic, S., "Analysis of Power Consumption in Future High-Capacity Network Nodes", in *Proc. of IEEE/OSA JOCN*, vol. 1, issue 3, 2009, pp. 245-258.

[15] Marsan, M. A., Chiaraviglio, L., Ciullo, D. and Meo, M., "Optimal Energy Savings in Cellular Access Networks", in *Proc. of IEEE ICC, Workshop on Green Communications*, 2009, pp. 1-5.

[16] Bianco, C., Cucchietti, F., Griffa, G., "Energy consumption trends in the Next Generation Access Network - a Telco perspective", in *Proc. of INTELEC*, 2007, pp. 737-742.

[17] Cioffi, J. M., Zou, H., Chowdhery, A., Lee, W., and Jagannathan, S., "Greener Copper with Dynamic Spectrum Management", in *Proc. of IEEE GLOBECOMM*, 2008, pp. 1-5.

[18] Zuckerman, D., "Green Communications – Management Included", in *Proc. of IEEE ICC*, 2009, Keynote.

[19] Marsan, M. A., Meo, M., "Energy Efficient Management of two Cellular Access Networks", in *Proc. of ACM SIGMETRICS, Performance Evaluation Review archive*, vol. 37, issue 4, 2010, pp. 69-73.

[20] Hasswa, A., Nasser, N., Hassanein, H., "Generic Vertical Handoff Decision Function for Heterogeneous Wireless Networks", in *Proc. of IFIP Conference on Wireless and Optical Communications*, 2005, pp. 239–243.
[21] Choi, Y. and Choi, S., "Service Charge and Energy-Aware Vertical Handoff in Integrated IEEE 802.16e/802.11 Networks", in *Proc. of IEEE INFOCOM*, 2007, pp. 589-597.
[22] Yang, W.-H., Wang, Y.-C., Tseng, Y.-C., and Lin, B.-S. P., "An Energy-Efficient Handover Scheme with Geographic Mobility Awareness in WiMAX-WiFi Integrated Networks", in *Proc. of IEEE WCNC*, 2009, pp. 2720--2725.
[23] Seo, S. and Song, J., "Energy-Efficient Vertical Handover Mechanism", in *Proc. of IEICE Transactions on Communications*, vol. E92-B, no. 9, 2009, pp. 2964-2966.
[24] Petander, H., "Energy-aware network selection using traffic estimation", in *Proc. of ACM MICNET*, 2009, pp. 55-60.
[25] Barth, U., Wong, P., Bourse, D., "Key Challenges for Green Networking", in *Proc. of Ercim News* 79, 2009, pp. 13.
[26] Gyarmati, L., Trinh, T. A., "How Can Architecture Help to Reduce Energy Consumption in Data Center Networking?", in *Proc. of ACM SIGCOMM*, 2010, pp. 183-186.
[27] Chang, V., Bacigalupo, D., Wills, G., and De Roure, D., "A Categorization of Cloud Computing Business Models" in *Proc. of IEEE/ACM CCGRID*, 2010, pp. 509-512.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 114-122

Localization of the Mobile Calls Based on SS7 Information and Using Web Mapping Service

Virgil CAZACU [1], Laura COBÂRZAN [2], Dan ROBU [3], Florin SANDU [4]

[1] BitDefender, Bucharest, Romania, e-mail: virgil.cazacu@gmail.com
[2] Softvision, Cluj-Napoca, Romania, e-mail: laura.cobarzan@gmail.com
[3] Siemens Program and System Engineering, Brasov, Romania, e-mail: dan.robu@siemens.com
[4] Faculty of Electrical Engineering & Computer Science, "Transilvania" University, Brasov, Romania, e-mail: sandu@unitbv.ro

Manuscript received October 01, 2010; revised October 20, 2010.

Abstract: Localization of the calls is a topic that has been coming up even from the early time of the telephony. Calls made from mobile phones were even more interested to be localized due to their mobility. This paper presents a localization solution that uses information from the mobile network, being a technical solution that ensures the acquisition of the localization information of the calls from the terminals in the mobile network and which is delivering this data to a localization server. The localization solution that is presented has three major features: receiving calls' information from mobile networks and obtaining the localization information from the Signaling System #7(SS7); data processing from signaling frame and IP-transmitting of this information to a localization server; visualization of the call location on the map. Due to client-server architecture, users of the system can access calls locations using digital maps.

Keywords: Localization, mobile networks, service, client-server, integration.

1. Introduction

Localization of the calls is useful not only from the legal point of view but also in case of emergencies as is for example the usage of the short number 112 or 911. In this case, the localization of the person who is in possible danger is vital.

Using SS7 localization approach has drawbacks which are treated in the presented solution: each mobile phone service provider supplies the localization information within the Initial Address Message (IAM) field of the ISDN User Part (ISUP) protocol from SS7 frame in its' own specific format [1]. Thus, the

solution offers the possibility to configure the necessary parameters, depending on the place of deployment.

From the design point of view, the solution ensures the service of acquirement of the localization information for the terminals in the mobile networks and it is delivering this information to a localization server for calls that are selected to be localized. The call processing, localization information extraction and delivering these on the interface to the localization server is performed in almost-real time, delays appearing if the load of the system is high. This solution is adaptable with minimal costs for future changes of the architecture. These changes might include resizing the necessary input/output traffic and the modification of the field from the SS7 signaling frame in which the localization information is transmitted by using parameterized components.

SS7 localization data is sent using "Cell ID" type from ISUP protocol, in the IAM message, "Location number" field and/or "Called number address" field [1].

The next table presents an example of localization data format that is specific for each mobile service provider.

Table 1: Localization data format.

Network code (e.g.72,74 or 4072, 4074)	Services bit (reserved)	Location area code	Cell ID
2-4 digits	1 digit	5 digits	5 digits

Based on these frames, each mobile operator maintains a database with geographic information that can offer information about the caller position based on the positioning string. The database structure is different, according to the telephony provider and contains the equivalent geographical coordinates for the above data from SS7 ISUP frame as it is shown as an example in *Table 2*.

Table 2: Geographical coordinates database structure.

Cell code	District	City	Street	Lat (Grade, Minutes, Seconds)	Long (Grade, Minutes, Seconds)	Azimuth	BSC	LAC
Specific Code	String	String	String	Int (6 digits)	Int (6 digits)	Int (3 digits)	Specific Code	Int (4 digits)

The location of these databases is defined by each mobile operator, which also manages and maintains it. These databases should be interfaced with a solution like the one presented in this paper.

The general solution of the architecture is presented in *Fig.1*.

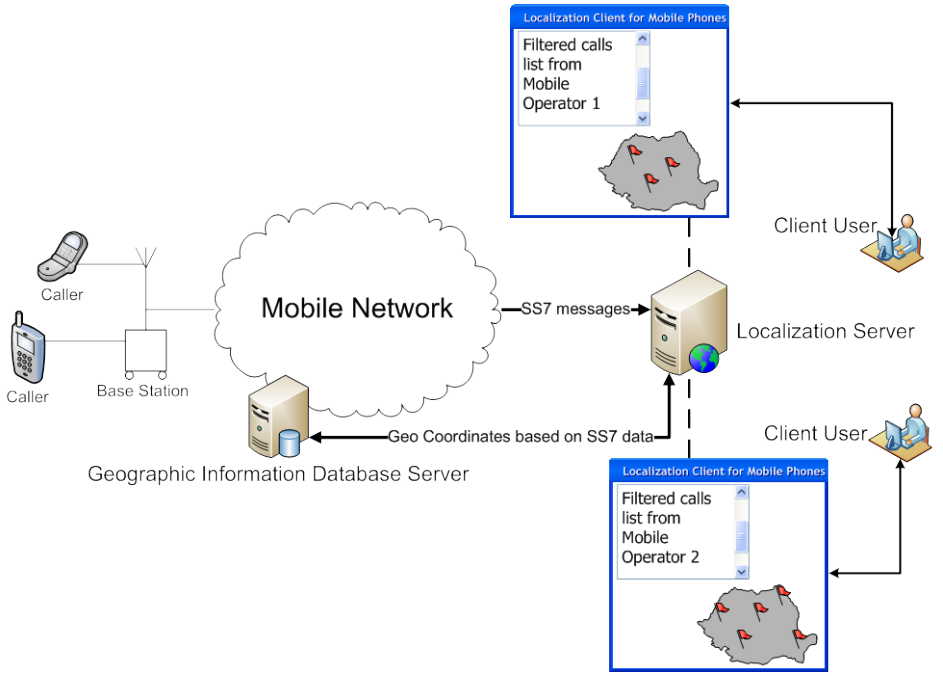


Figure 1: Overview of the solution architecture.

The solution, as the picture above shows, is divided into four modules and it is targeted to be used in operational centers that can coordinate emergency activities:

- Extracting Localization String modules from the Localization Server presumes the definition of rules for parsing SS7 ISUP information. The main service of the server is to define interaction protocols as well as Geo Information Database Server interaction.
- Geographic Information Database Server is maintaining the geo-coordination of the radio cells.
- Client side module is handling communication between Graphical User Interface (GUI) module and Localization Server.
- GUI module is handling specific interface functions and the digital maps using web mapping service available on the market at the solution's implementation time.

2. Description of the main components

This section presents the technical implementation of the solution modules using as examples the case when mobile users are using the Emergency Service 112 and are customers of one of the Romanian mobile operators.

A. Localization Server

In this module, there are two major functionalities: SS7 frame parsing and communication protocols between the client and the Geo Database Server.

The parsing module consists of two parts, one dealing with the SS7 Integrated Services Digital Network User Part (ISUP) communication, while the other being responsible for the protocol message parsing. It is out of the paper's scope to detail the SS7 communication between our solution and the mobile network. The technical approach taken is to use the JAIN ISUP API that gives the possibility to exchange ISUP messages in the form of Java Event Objects [2], [3].

One rule for parsing SS7 information is the fact that independently of the mobile operator, SS7 frames are in standard format and the relevant parameters for our solution can be found under the Initial Address heading. The relevant parameters are presented in the *Table 3*.

Table 3: IAM parameters used for localization.

Parameter Name	Explanation
Calling party number	Nature of address: either National or International. Calling address signal: the telephone number of the caller party (with a 2 digit prefix for international calls).
Called party number	Called address signals. For Emergency cases, the called telephone number is 112.
Location number	The localization string, that contains all the localization information for the given provider.
Cell ID	The mobile operator internal ID for the radio cell where the call is made from.

The phone numbers are received without the prefix digits, so in this implementation the "Calling party number" parameter is taken into account. For national calls a 0 digit and for international calls two 0 digits are inserted at the beginning of the caller number. Also, to determine from which mobile operator the call is performed and knowing that every telephone number begins for example with 07XY, where depending on the XY digits, the solution can extract

the provider of the call based on a table correspondence and on interrogating the portability server, if available.

In the implementation of the module, the extracted information is stored in an object called SS7Object with fields like String callingNumber, String calledNumber, String Nature, String LocalizationString, Date dateCreated, String Provider. SS7Objects are sent to the Localization Server module for further processing.

The communication protocol with the client uses sockets and when new localization objects are received from the SS7 parsing module, the server will send the object to the client in order to use it on the GUI. After sending the localization object, the server is waiting for an answer from the client. If the client does not confirm the reception of the object in the previously defined time frame, the localization server will resend this information. The Localization objects that are not confirmed are maintained in a waiting list. When the localization server receives a message from GUI/Google Maps, it will delete the corresponding object from the "waiting list", meaning that it will not wait for the confirmation for that object.

The communication protocol between Localization Server and the Geo Information Database Server is done by calling the getCoordinatesBy LocalizationString (localization_string) method which takes a string parameter, representing the localization string and as returned value, an object which contains the coordinates of the area from which the call was made. The coordinates will be the latitude and the longitude, each one containing 3 fields: degrees, minutes and seconds. The calling of the class is done using Remote Method Invocation (RMI).

B. Geographical Information Database Server

As it was mentioned earlier in the paper, this server has to be located at each mobile operator since it contains internal information about the place where radio cells are deployed from geographical point of view. For completeness of the solution description, the RMI Database Server will be presented, that has several classes in order to extract the coordinates from the local database.

The Coordinate class is common with the RMI client, the *CoordinateInterface* class which contains the remote method and the *CoordinateImplementation* class, which implements the remote method as shown below:

public Coordinate getCoordinatesByLocalizationString(String localization_string)

The Provider class has a static method String getProvider(String localization_str), which returns the provider, based on the localization string and

the ExtractCoordinate class has a static method that returns the coordinates, based on the provider and the localization string, as it can be seen next capture.

public static Coordinate getCoordinatesByProvider(String provider, String localization_str)

The ExtractFromDatabase class contains one static method for each provider, to extract the coordinates from the local database, based on the localization string, as shown below:

public static Coordinate ExtractVodafone(String localization_string)

A database structure example for this server is presented in *Fig. 2*.

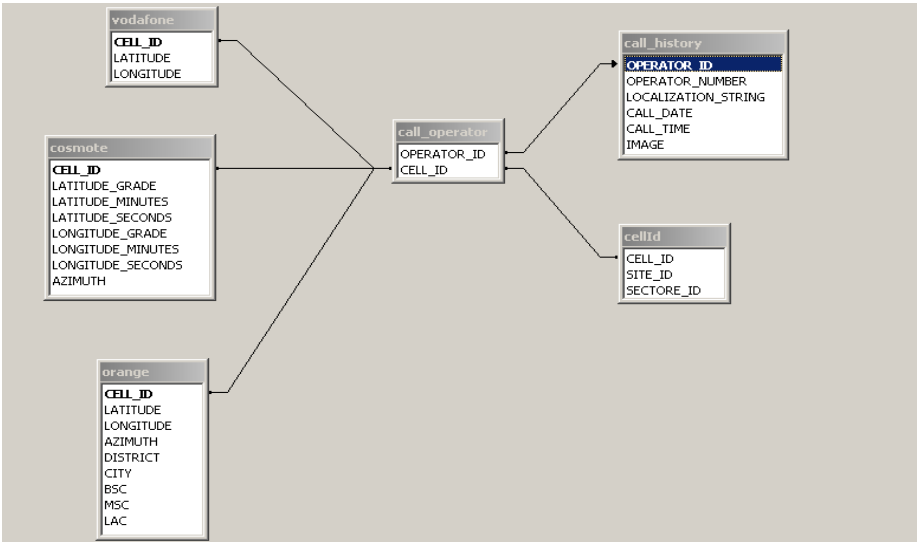


Figure 2: Structure example of geo coordinates database.

C. Client side including GUI

As the communication between the client and the Localization Server is detailed in section B, this part is focused on the usage of web mapping service and user interface.

The graphical user interface presented in this paper has a proof of concept oriented design. The GUI of the solution is composed of two frames: one frame with 4 tabs: View Calls, View Archive, Options and Help. The other frame is displaying the digital map with all its options.

Only one tab is presented in this paper, the View Calls tab which contains the recent calls information in a list. The call information contains the exact time of the call, the caller number, the provider and the location of the call as it is received from the Location Server. Each call has its own checkbox, which will specify if the call was processed or not. When a call is selected in the list, the application is marking automatically the location of the call in the map frame. More calls can be selected simultaneously, so the map can be marked in more locations. Calls that are checked (processed) are deleted from the list and the marks from the map disappear.

In order to integrate a digital map into the solution, Google Maps API was chosen due to several considerations [5].

Google Static Maps API embeds a Google Maps image without requiring JavaScript but the problem is that it returns the map as an image (GIF, PNG or JPEG) in response to a HTTP request via a URL. This way, the benefits of the zoom and navigation facilities disappear.

JXMapViewer embeds mapping abilities into Java application, but at the solution's development time it was not possible to use it with Google Maps or Yahoo since there were legal restrictions.

One other strong reason why the Google Maps API was chosen for integrating the web mapping service was the possibility to control the zoom and navigation features from the application's back-end.

Since Google Maps API uses JavaScript, the *JWebBrowser* class from the chrriis.dj.nativeswing.components package has been used in the development of the Java client application; this offers the possibility to have a native web browser component in the application [4]. Because the client application has to be operating system independent, the web browser component was developed to use the Mozilla engine.

```
NSOption opt = new NSOption(JWebBrowser.useXULRunnerRuntime());
web_browser = new JWebBrowser(opt);
JWebBrowser.useXULRunnerRuntime();
```

The digital map from Google is loaded using the following code line [5].

```
web_browser.navigate(gmapfilelocation.getAbsolutePath());
```

The parameter *gmapfilelocation* points to the file containing the script which loads the map.

In *Fig. 3*, the client GUI is shown together with the calls markers, each marker descriptor containing a string defining the location to place the marker and the visual attributes to use when displaying the mark.

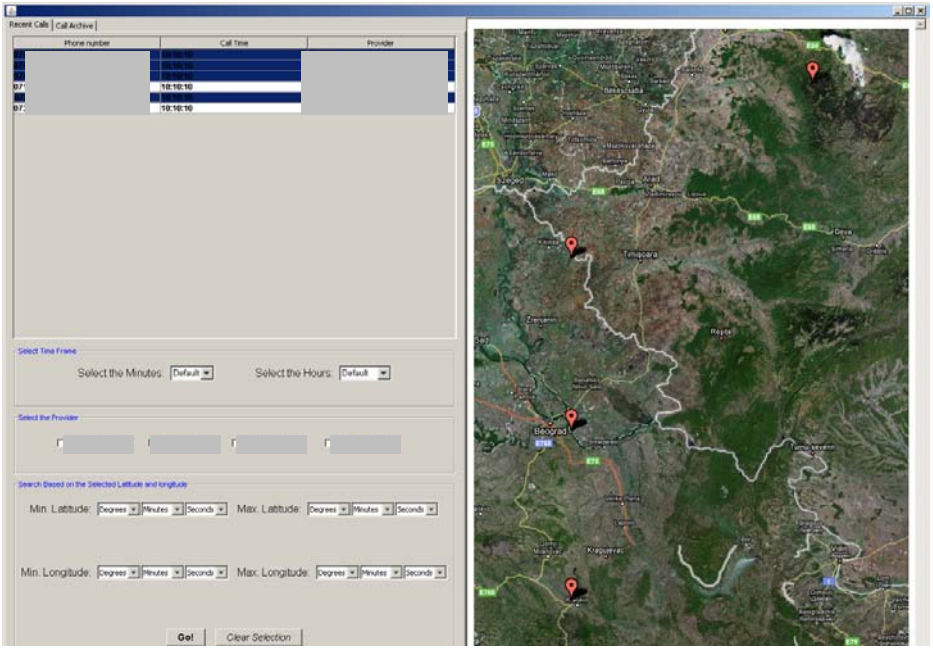


Figure 3: Calls markers on the map.

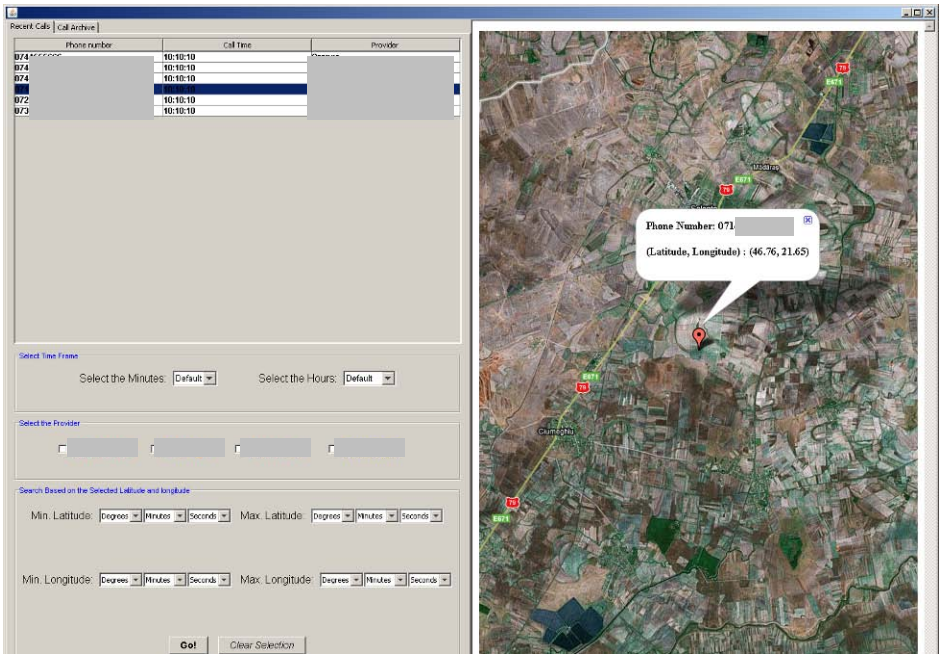


Figure 4: Zoom on caller location.

Figure 4 shows a case when one call is selected from the list and automatically the application zooms in on the location where the caller is positioned. If another call is selected, so that two calls are on the map, the application automatically zooms out exactly as it is necessary to display both callers on the map.

3. Conclusion

The solution of call localization presented in this paper is still under development since topics like high degree of availability or the ability to work in load-balanced and failed-over conditions between locations are not implemented. The application's architecture has been implemented by the authors of the paper and solution allows further improvements, in order to enable features like accepted input traffic of a high number of simultaneous voice calls to be implemented as easy as possible.

But the goal of the research, at least in this phase, was achieved, since the usage of a web mapping service for calls location has been demonstrated by the solution presented in this paper.

Acknowledgements

The authors wish to thank their colleagues who contributed with their effort to achieve the results presented in this paper, especially the colleagues located at the Cluj-Napoca Siemens PSE site.

References

[1] Dryburgh, L., Hewett, J., "Signaling System No. 7 (SS7/C7): protocol, architecture, and services", Cisco Press, 2005.
[2] Jepsen, T. C., Anjum, F., "Java in telecommunications: solutions for next generation networks", John Wiley & Sons, 2001.
[3] *** http://jcp.org/en/jsr/summary?id=ISUP.
[4] *** http://djproject.sourceforge.net/ns/documentation/javadoc/index.html.
[5] *** http://code.google.com/apis/maps/documentation/reference.html.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 123-135

Analysis of Neuroelectric Oscillations of the Scalp EEG Signals

László F. MÁRTON, László SZABÓ, Margit ANTAL, Katalin GYÖRGY

Department of Electrical Engineering, Faculty of Technical and Human Sciences,
Sapientia University, Tg. Mureş, e-mail: martonlf@ms.sapientia.ro

Manuscript received November 15, 2010; revised November 30, 2010.

Abstract: Electroencephalography (EEG) or magneto-encephalography (MEG) are usual methods adopted in clinics and physiology to extract information relative to cortical brain activity. EEG/MEG signals are a measure of the collective neural cell activity on restricted regions of the cortex. The brain activities ensue from the interaction of excitatory and inhibitory populations of neurons. Their kinetics vary depending on a particular task to be fulfilled, on the particular region involved in that task, and in any instant during the task. In order to improve our understanding of EEG/MEG signals, and to gain a deeper comprehension of the neuro-physiological information contained, various mathematical models and methods have been proposed in previous years. Based on these concepts we develop procedures for an optimal, online and hardware based EEG signal processing. EEG recordings are analyzed in an event-related fashion when we want to gain insights into the relation of the EEG and experimental events. An approach is to concentrate on event-related oscillations (EROs). The method suited to analyze the temporal and spatial characteristics of EROs, is the time-frequency analysis namely wavelet transforms. Recently introduced wavelet-based methods for studying dynamical interrelations between brain signals will be discussed.

Keywords: brain-computer-interface (BCI), time frequency analysis, filters, EEG signals, EMG signals, stationarity of signals.

1. Introduction

We propose a novel framework to analyse electroencephalogram (EEG) biosignals from multi-trial visually-evoked potential (VEPs) signals recorded

with a brain-computer interface (BCI). Electromagnetic activities of the neuromuscular system, including electroencephalography (EEG), electro-myography (EMG), and magneto-encephalography (MEG) signals, have been widely used in the study of motor control in humans. VEPs signals contain components of EMG, MEG and EEG [1], [2], [3], [4].

Oscillations are characterized by their amplitude, frequency and phase. The amplitude of a recorded oscillation is typically between 0 and 10 μV. The (cyclic) phase ranges are between 0 and 2π. At every scalp point at every time moment the amplitude and phase of an oscillation can be recorded. The frequency band of the typical recordable cortical oscillations range is usually from 0.1 Hz to 80 Hz (highly correlated to a 256 Hz sampling rate).

Several methods exist to extract oscillations of biologically specific frequency bands from ERO data. Among the most used are band filtering, Fourier analysis, and wavelet analysis. Oscillating potentials derived from a specific scalp surface, originating from the outer layer of cortex (grouped neuron structures in the layer I cortical areas) are called visual-evoked potential (VEP) signals. These signals are related to the brain's response to visual stimulation and have applications in numerous neuropsychological studies. EROs comprise exogenous and endogenous components. Exogenous components are obligatory responses which result on the presentation of physical stimuli. The endogenous components (say P300 component of an ERO signal) manifest the processing activities which depend on the stimuli's role within the task being performed by the subject. Usually, EEG-signals are based on various phenomena like, for example, visual and P300 evoked potentials, slow cortical potentials, or sensory-motor cortical rhythms. The P300 shape is an event related potential, elicited by a generally stochastic, task related stimuli.

2. Materials and methods

We have used a BrainMaster recording system (BrainMaster AT-1 W2.5 Clinical Pro Wideband System, see *Fig. 1.*), battery powered, portable, two channel 2E neurofeedback module hardware, with added cleaning gel and conductive paste. Recordings have been made with 5 gold plated electrodes (2 recording electrodes, 2 reference electrodes and a ground electrode). The used sampling rate was 256 Hz, sufficient for the studied frequency bands.

Recording areas were the left and right hand side of cortical surface at Primary Motor cortex and sometime on Pre-Motor and Supplementary Motor Cortex (Secondary Motor Cortex) on the scalp of young persons. The recorded and sampled signals from two channels (Ch1, Ch2) are stored in text type files and processed using MATLAB (The Math-Works Inc., Natick, MA.) platform. The software package is elaborated by the authors based on concepts presented

in [5]. The test recordings are usually 1 to 1.5 minutes long. The ERO intervals alternate with relax state with a period of usually 10 seconds. The time-frequency method used in this application is the continuous wavelet transform based on Morlet, Paul and DOG ($m = 2$ and $m = 6$) wavelet base functions. The Morlet and Paul bases are providing a complex continuous transform proper for time-frequency component analysis of the recordings. For the subjects of the experiments, during the recordings a deckchair was used to avoid extra EMG noise created by the body stability problems.

Figure 1: The BrainMaster AT-1 System (Brain Master LTD company product image).

In the EEG–EMG experiments, subjects are trained for two different motor tasks: a left-right or up-down movement of the closed or opened eye balls and right or left hand movement. The scenario of the performed task is recorded in the header of the generated file. The recording technique is a not invasive recording method.

3. Data analysis tools

A period of baseline EEG+MEG was recorded at the beginning of each experiment when the subjects rested. During the offline analysis, signals within an approximately 10 s epoch were selected from the rest period and averaged to obtain baseline EEG. Subsequently, the value of the baseline EEG+MEG was subtracted from the entire EEG+MEG data set to acquire baseline corrected signals. The corrected values were saved into files for further processing.

An important fact is that the magnitudes of cyclic visual-evoked potential components are much more detectable in the 0–10 Hz frequency range. This is important for the further analysis.

From the experimental studies of VEPs, relative to the recorded oscillations, the literature clearly depicts the delta (1–4 Hz) and theta (4–10 Hz) ranges as containing main components of power in frequency domain for the waves. We will consider these bands for further identification of the activity patterns [1]. Now, we are considering the important details of wavelet transform used in these processings.

By decomposing a time series into time–frequency space, one is able to determine both the dominant modes of variability and how those modes vary in time. The first tested method was the Windowed Fourier Transform (WFT). The WFT represents one of analysis tool for extracting local-frequency information from a signal. The WFT represents a method of time–frequency localization, as it imposes a scale or 'response interval' T into the analysis. An inaccuracy arises from the aliasing of high- and low-frequency components that do not fall within the frequency range of T window. Several window lengths must be usually analyzed to determine the most appropriate choice of window size to be sure to contain within the window the main, but unknown basic oscillatory components. To avoid this difficult task, in our analysis finally we have used wavelet transform (WT) methods.

The WT can be used to analyze time series that contain nonstationary power at many different frequencies. The term 'wavelet function' is generically used to refer to either orthogonal or nonorthogonal wavelets. The term "wavelet basis" refers only to an orthogonal set of functions. The use of an orthogonal basis implies the use of the discrete wavelet transform (DWT), while a nonorthogonal wavelet function can be used with either the discrete or the continuous wavelet transform (CWT).

A brief description of CWT is following. Assume that the recorded time series, x_n, is with equal time spacing δt (sampling period) and $n = 0…N\text{-}1$. Also assume that one has a wavelet function, $\Psi_0(\eta)$, which depends on a non-dimensional 'time' parameter η.

To be 'admissible' as a wavelet, this function must have zero mean and must be localized in both time and frequency space. An example is the Morlet wavelet, consisting of a plane wave modulated by a Gaussian function:

$$\Psi_0(\eta) = \pi^{-1/4} \cdot e^{i \cdot \omega_0 \cdot \eta} \cdot e^{-\eta^2/2} \tag{1}$$

This is a wavelet basis function, where w_0 is the non-dimensional frequency, here taken to be 6 to satisfy admissibility condition.

The continuous wavelet transform of a discrete sequence x_n is defined as the convolution of x_n with a scaled and translated version of $\Psi_0(\eta)$:

$$W_n(s) = \sum_{i=0}^{N-1} x_i \cdot \Psi^* \left[\frac{(i-n) \cdot \delta t}{s} \right], \tag{2}$$

where the (*) indicates the complex conjugate. By varying the *wavelet scale* ***s*** and translating along the *localized time index* ***n***, one can construct a picture showing both the amplitude of any features versus the scale and how this amplitude varies with time. The subscript 0 on Ψ has been dropped to indicate that this Ψ has also been normalized. It is possible to calculate the wavelet transform using (2), and it is considerably faster to do the calculations in Fourier space. By choosing *N* points, the convolution theorem allows us to do all *N* convolutions simultaneously in Fourier space using discrete Fourier transform (DFT). To ensure that the wavelet transforms at each scale *s* are directly comparable to each other and to the transforms of other time series, the wavelet function at each scale *s* was normalized to have unit energy. Normalization is an important step in time-series analysis and is used at each scale *s*.

Morlet wavelet function $\Psi(\eta)$ is a complex function, the wavelet transform $W_n(s)$ is also complex. The transform can then be divided into the real and imaginary part, or amplitude and phase. Finally, one can define the *wavelet power spectrum* as $|W_n(s)|^2$. The *expectation value* for $|W_n(s)|^2$ is equal to *N* times the expectation value for the discrete Fourier transform of the time series. For a white-noise time series, this expectation value is σ^2/N, where σ^2 is the variance of the noise. Thus, for a white-noise process, the expectation value for the wavelet transform is $|W_n(s)|^2 = \sigma^2$ at all *n* and *s*. Based on this knowledge, the same logic is used to calculate the expected value of red noise. The $|W_n(s)|^2 / \sigma^2$ is the measure of the normalized signal value relative to white noise. As the biological background noise is a red noise type, the normalization method is relative to red noise as it is described in [5] (in a way as it was used in the results of this paper).

An important concept of this study is the so called *Cone of influence* (COI). The cone of influence is the region of the wavelet spectrum in which edge effects become important because of the finite length of signal and used window. The significance of the edge effect is defined as the *e*-factor (power spectrum edge drops by a factor e^{-2}) time of wavelet power at each scale. The edge effects are negligible beyond the COI region. This must be considered for an accurate analysis. In each figure, COI is represented at the edge of the wavelet transforms (lighter area in figures).

Another important factor we have added to this analysis is the *significance level* of the correlation studies. The theoretical white/red noise wavelet power spectra are derived and compared to Monte Carlo simulation results. These spectra are used to establish a null hypothesis for the significance of a peak in the wavelet power spectrum (the question to be answered: is a power peak from a wavelet figure the result of biological events or it is a result of stochastic red/white noise effect?).

The null hypothesis is defined for the wavelet power spectrum as follows. It is assumed that the time series has a mean power spectrum, if a peak in the wavelet power spectrum is significantly above this background spectrum, then it can be assumed to be a true feature with a certain percent of confidence. ('significant at the 5% level' is equivalent to 'the 95% confidence interval').

Our application is highlighting the biological events, with surrounding the significant peaks of correlations at the confidence interval of 95%. It is important that in biological studies the background noise can be modeled by a red (or pink) noise. A simplest model for a red noise is the lag-1 autoregressive [AR (1), or Markov] process. In the following figures, all significant localization of a biological event (VEPs) in time-frequency domain is also statistically significant. This is an important result. What is not within a significant area, is not considered in the results. Another important result in our analysis is the representation of the phase relationship between two recordings. In each cross-wavelet spectrum and cross-coherence spectrum the phase relationship is represented with arrows. A horizontal arrow to the right means that in that time frequency domain the two biological signals are in phase (if that domain is significant at 5% level and is not within COI domain). The opposite arrow orientation has the meaning of opposite phase correlation. The angle of arrows relative to horizontal line is showing the phase angle in that time frequency domain. Our application is calculating and is representing all these phase values. The definition of wavelet-cross-correlation and wavelet-cross-coherence are defined in [5] and [6].

4. Results

The following figures are slim examples from our recordings and their analysis. *Fig. 2* is the amplitude/time representation of a channel signal as a recording on the right hand side of the Motor Cortex area when the left hand has been lifted two times during a 50 s recording session. This figure is showing also the application menu created for this WFT type of analysis. *Fig. 3* is the WFT representation of the *Fig. 2* recording in different frequency bands. The vertical axis is for the frequency, and the horizontal one is the time axis. The frequency bands for the different windows are 0.1 – 50 Hz for the top left, 0.1 – 8 Hz for bottom left, 24 – 30 Hz top right window and finally 30 – 36 Hz for bottom right window. The frequency bands are representative for biological events. A color-code represents the intensity of a time-frequency domain for that WFT decomposition (the corresponding color-code is represented at the right hand side of each window). The shape within each window is characteristic for a real time motion recorded as EMG+EEG. *Fig. 4* is similar with *Fig. 3* but it is represented in 3D for a better visual understanding of the

significant domains contained in this time-series decomposition. Based on these figures, it can be concluded that a 'shape' in time-frequency domain is related to an arm motion in time domain. The automatic identification of these domains (peaks) in time-frequency decomposition it must be related to the corresponding arm motion what has happened in time domain. For this peak identification one can use a pattern recognition procedure.

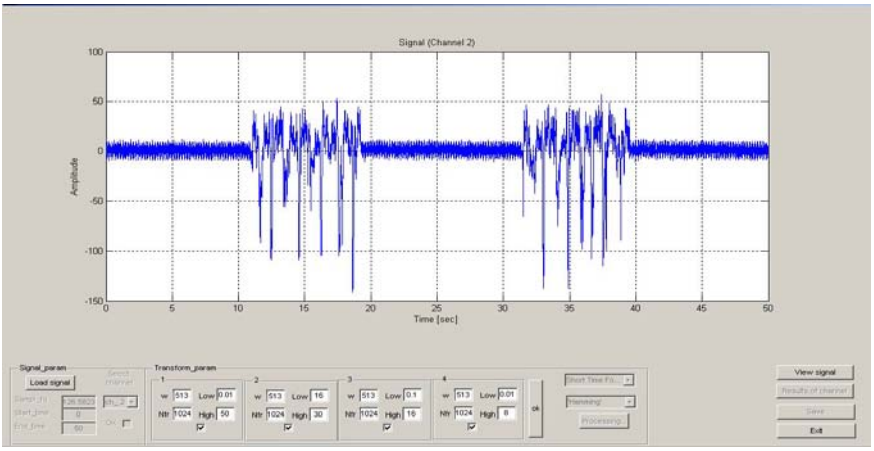


Figure 2: A recorded (right hand side Motor Cortex) amplitude/time representation.

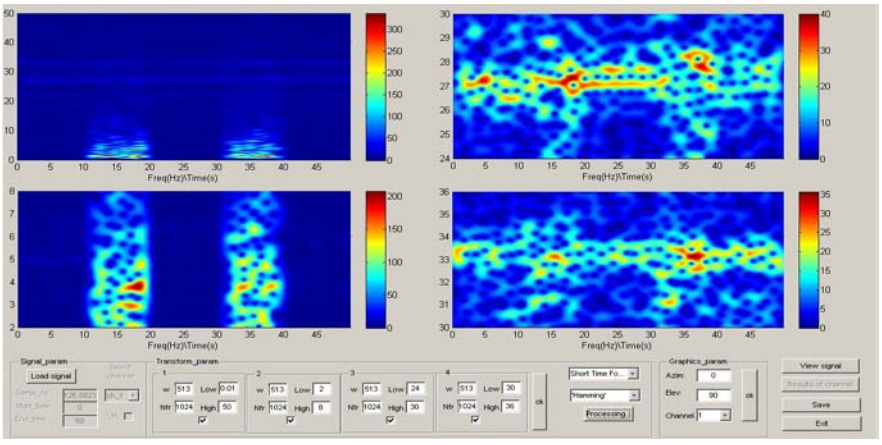


Figure 3: The wavelet Fourier transform (WFT) in four different frequency bands.

As it was mentioned, the same hand lifting time event recording has been done on both cortical sides. The left cortical side recording and decomposition with WFT is not represented here, but it has a similar configuration. *Fig. 5* is the difference in time-frequency domain of the two side signals. It is visible that the two side recordings are not the same as it is known from theory. This

analysis is WFT, and here the COI and significance test was not used. As it was mentioned, the WFT is very sensitive to the window length (T) used in decomposition of the time signal [7], [8]. In the spectrum, not controllable frequency interference is present, but the method is powerful enough to be usable in detection and classification of not very sensitive types of motor actions.

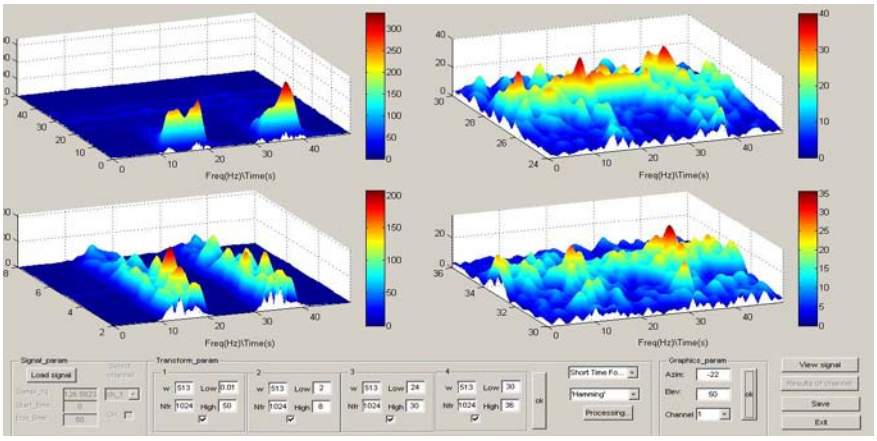


Figure 4: The 3D representation of figure 3 WFT decomposition.

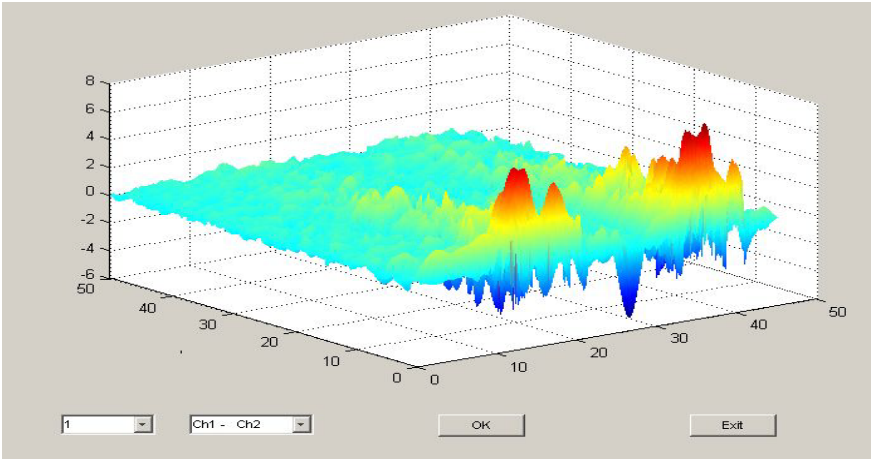


Figure 5: The difference of the two channels recording decomposition by WFT.

A most sensitive procedure is the use of Morlet type of Wavelet transform (WT) with the representation of COI and with significance test of biological events based on wavelet Cross-correlation and wavelet Cross-Coherence [5], [9]. These results are more accurate and with higher resolution in time

frequency in comparison with the previous WFT analysis. The amount of calculation is higher than in case of WFT but optimizing the procedures on hardware based processor units, this method should be very powerful.

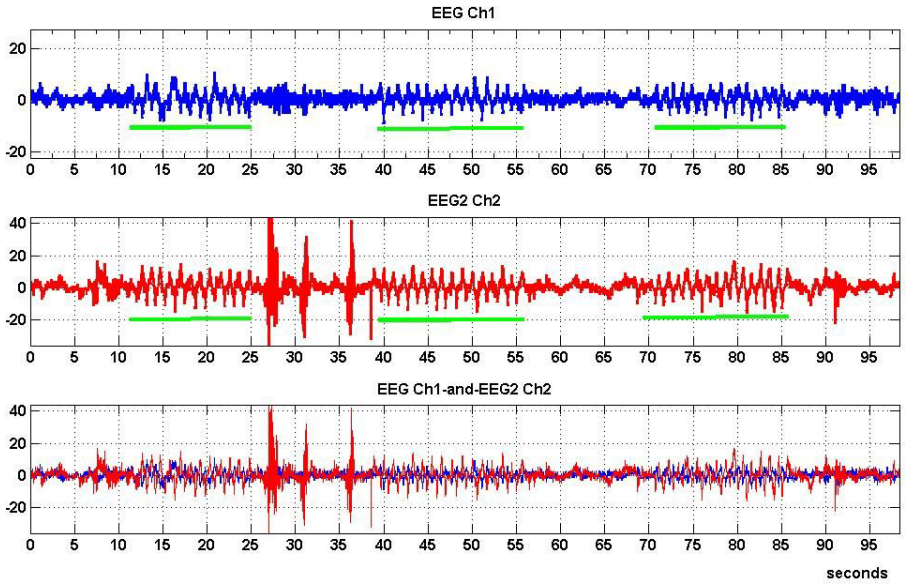


Figure 6: The two channels (Ch1, Ch2) amplitude/time representation of the recordings. The bottom figure is showing the two superimposed signals. The green highlights are the eye movement time sequences (sample size on vertical).

Fig. 6 is representing left-right movement of eye balls with a relax time between them. It is visible the time sequence of left and right hand cortical side EMG+EEG. The whole recording length is about 100 seconds. The high amplitude signals in 25-37 seconds interval in Ch2 recording is an extra EMG, a noise from the experiment point of view. The time sequence is containing three eye movement events. These are between (12, 25) sec, (39, 55) sec and finally (71, 85) seconds. These are highlighted by green line segments. In the third (bottom) window it is visible that the Ch1 and Ch2 recordings are in opposite phase. But this will be obvious from cross correlations calculated and represented in *Fig. 9*.

The next two figures (*Fig. 7* and *Fig. 8*) are the representation of Morlet WT of these channel recordings. The COI and the significant areas are represented. Domains of the signal within these closed contours are significant, outside are not significant (should not be considered biological events). We are considering the two parallel lines delimiting roughly the (0.75 – 1.5) Hz frequency interval. Within these limits we can consider the events of eye movement left-right-left in the detected time intervals. It is very obvious the presence of significant

domains, and they can be easily identified. In *Fig. 8*, in 25 sec to 37 sec interval, the EMG 'noise' is there, but the basic components are present also at much higher frequency domains.

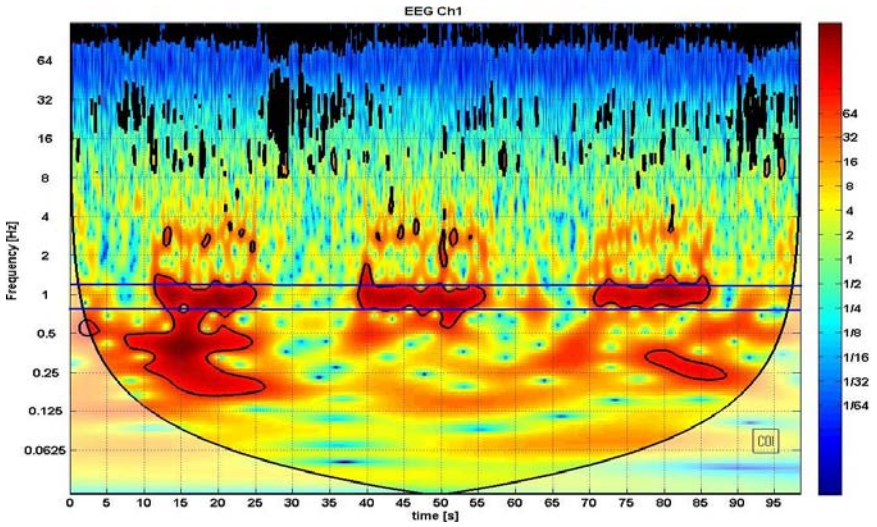


Figure 7: Morlet WT of Ch1 recording with localization of eye left-right movement.

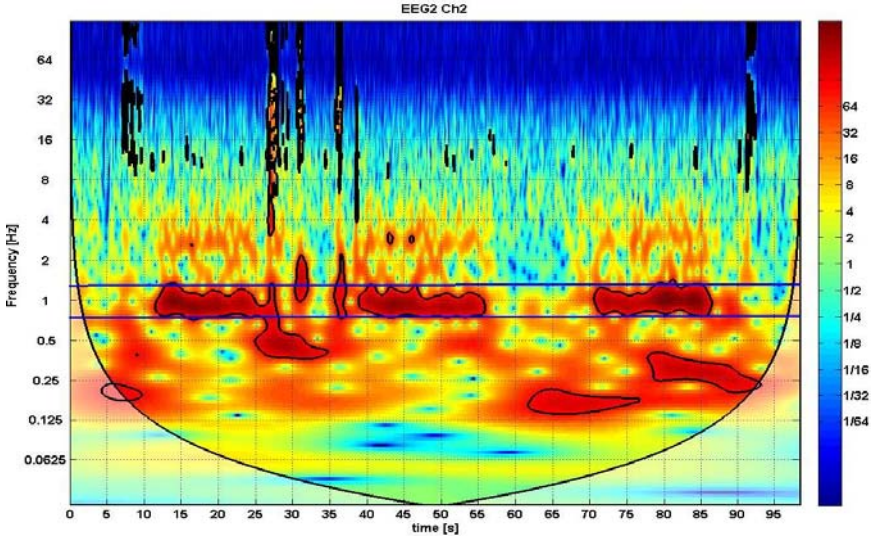


Figure 8: Morlet WT of Ch2 recording with localization of eye left-right movement.

It is very important to discuss, what *Fig. 9* represents. The COI is present as the not significant domain, but also a phase relationship between the two signals is calculated. The arrows' orientation within the significant area is to the left

(0.75 – 1.5Hz). This means that the correlation between the two channels, in this frequency band is in opposition. In the recordings with half way eye movement (left to middle, or right to middle) the phase shift is not opposite but is around of 90/270 degrees. These phase events permit the detection of the direction of eye movements. *Fig. 9* bottom image is the normalized version of the same cross-correlation, the so-called cross-coherence between the two channels. This information about the interrelation of the two recordings is more relevant to characterize the ERO contained visual evoked potentials.

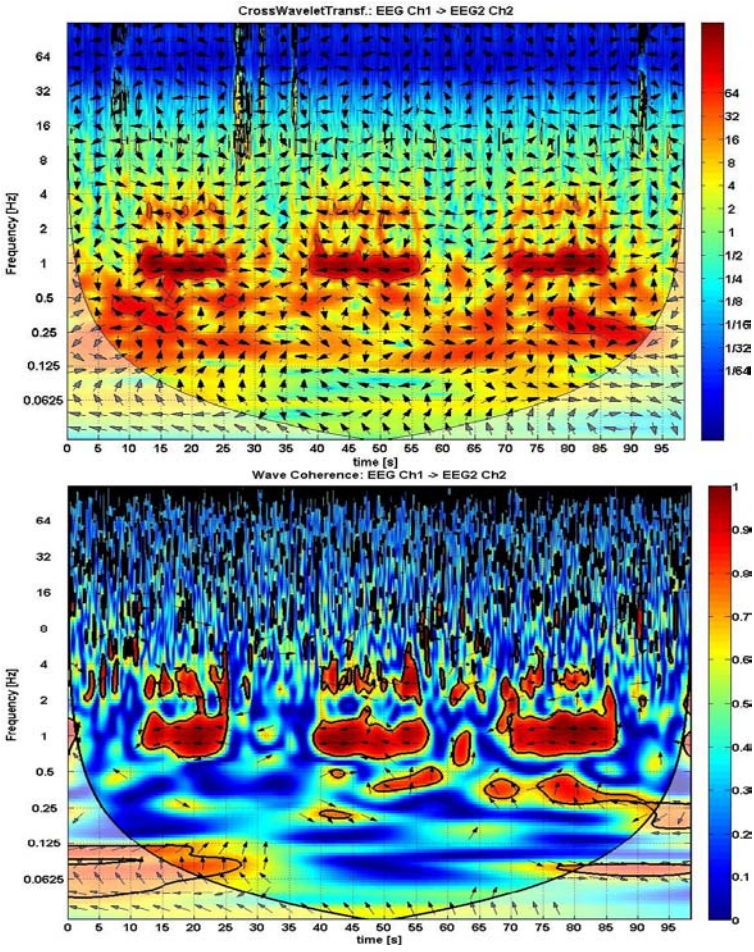


Figure 9: The cross-correlation (up) and the cross-coherence of the two channels signals.

The same analysis has been made also for the up-down eye movement direction. The results are very similar with the left-right movement conclusions,

and are not presented in this paper, but can be considered for technical applications based on EEG+MEG recordings. The cross-coherence matrix is processed to extract the information (signals) for further control tasks.

5. Conclusions

Every characteristic information of EROs (extracted numeric values of the processed VEPs) is usable for further signal processing tasks. Detecting motion related, EEG+MEG signal configuration, the possibly extracted information is usable in external system control tasks. The sensitivity of the electrodes is not the subject of this paper. It is obvious that with a much higher sensitivity of electrodes, it should be possible to record much more detailed signals with deeper correlations. The ideas used in this study are at the beginning of more event sensitive identification and complicated command possibilities [8], [10], [11]. We must also consider the effect of so called grid cells in the cortical area which display regular responses to the position in a virtual, internal 2-D space. This study should be possible using multichannel (>2) recordings, a next step in our research.

References

[1] Nicolelis, M. A. L. and Lebedev, M. A., "Principles of neural ensemble physiology underlying the operation of brain–machine interfaces", *Nature Reviews Neuroscience*, vol. 10, pp 530-540, July 2009.

[2] Hockensmith, G. B, Lowell, S. Y, and Fuglevand, A. J. "Common input across motor nuclei mediation precision grip in humans". *J. Neuroscience*, vol. 25, pp. 4560–4564, 2005.

[3] Caviness, J. N., Adler, C. H., Sabbagh, M. N., Connor, D. J., Hernandez, J. L., and Lagerlund T. D., "Abnormal corticomuscular coherence is associated with the small amplitude cortical myoclonus in Parkinson's disease", *Mov Disorder*, 2003, no.18, pp. 1157–1162, 2003.

[4] Grosse, P., Cassidy, M. J., Brown, P. "EEG–EMG, MEG–EMG and EMG–EMG frequency analysis: physiological principles and clinical applications", *Clin Neurophysiol*, no. 113, pp. 1523–1531, 2002.

[5] Torrence, C., and Compo, G. P., "A Practical Guide to Wavelet Analysis", *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61-78, January 1998.

[6] Yao, B., Salenius, S., Yue, G. H., Brownc, R. W., and Liu, Z. L., "Effects of surface EMG rectification on power and coherence analyses: An EEG and MEG study", *Journal of Neuroscience Methods*, no.159, pp. 215–223, 2007.

[7] Ermentrout, B. G, Gala´n, R. F., and Urban N. N. "Reliability, synchrony and noise" *Review: Trends in Neurosciences* Cell Press, vol. 31, no. 8, pp. 428-434, 2008.

[8] Faes, L., Chon, Ki. H., and Giandomenico, N., "A Method for the Time-Varying Nonlinear Prediction of Complex Nonstationary Biomedical Signals", *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 2, pp. 205-209, February 2009.

[9] Chua, K. C., Chandran, V., Rajendra Acharya, U., and Lim C.M. "Analysis of epileptic EEG signals using higher order spectra" *Journal of Medical Engineering & Technology*, vol. 33, no. 1, 42–50, January 2009.

[10] Guo, X., Yan, G., and He, W. “A novel method of three-dimensional localization based on a neural network algorithm”, *Journal of Medical Engineering & Technology*, vol. 33, no. 3, pp. 192–198, April 2009.

[11] Ajoudani, A., and Erfanian, A., “A Neuro-Sliding-Mode Control With Adaptive Modeling of Uncertainty for Control of Movement in Paralyzed Limbs Using Functional Electrical Stimulation”, *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 7, pp. 1771-1780 July 2009.

[12] Harrison, T. C., Sigler, A., and Murphy, T. H., “Simple and cost-effective hardware and software for functional brain mapping using intrinsic optical signal imaging”, *Journal of Neuroscience Methods*, no. 182, pp. 211–218, 2009.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 136-145

Nonlinear Filtering in ECG Signal Denoising

Zoltán GERMÁN-SALLÓ

Department of Electrical Engineering, Faculty of Engineering, "Petru Maior"
University of Tg. Mureş, e-mail: zgerman@engineering.upm.ro

Manuscript received October 15, 2010; revised November 08, 2010.

Abstract: This paper presents a non-linear filtering method based on the multiresolution analysis of the Discrete Wavelet Transform (DWT). The main idea is to use the time-frequency localization properties of the wavelet decomposition. The proposed algorithm is using an extra decomposition of the identified noise in order to reduce the correlation between the electrocardiogram (ECG) signal and noise. The linear denoising approach assumes that the noise can be found within certain scales, for example, at the finest scales when the coarsest scales are assumed to be noise-free. The non-linear thresholding approach involves discarding the details exceeding a certain limit. This approach assumes that every wavelet coefficient contains noise which is distributed over all scales. The non-linear filter thresholds the wavelet coefficients and subtracts the correlated noise. The used threshold depends on the noise level in each of the frequency bands associated to the wavelet decomposition. The proposed non-linear filter acts by thresholding the detail coefficients in a particular way, in order to eliminate the correlation between the noise and the signal. In this paper, in order to evaluate the proposed filtering method, signals from the MIT-BIH database have been used, and the filtering procedure was performed with added Gaussian noise. The proposed procedure was compared with ordinary wavelet transform and wavelet packet transform based denoising procedures, the followed parameters are the signal to noise ratio and the denoising error.

Keywords: Wavelet decomposition, wavelet shrinkage, non-linear filtering.

1. Introduction

The interest in using the wavelet transform to denoise the electrocardiogram (ECG) signals is increasing. The wavelet transform is a useful tool from time–frequency domain, preferred for the analysis of complex signals. The application of this transform to ECG signal processing has been found particularly useful due to its localization in time and frequency domains. The discrete wavelet transform- based approach produces a dyadic decomposition structure of the signals. In this way, the wavelet packet approach is an adaptive method using an optimization of the best tree decomposition structure independently for each signal.

2. Methods

The continuous wavelet transform (CWT) of the signal $x(t)$ is defined as a convolution of a the signal with a scaled and translated version of a base wavelet function, [1]:

$$W_a x(b) = \int_{-\infty}^{+\infty} x(t) \cdot \psi_{a,b}(t)\, dt = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} x(t) \cdot \psi\left(\frac{t-b}{a}\right) dt \tag{1}$$

where the scale '*a*' and translation '*b*' parameters are nonzero real values and the wavelet function is also real. A small value of '*a*' gives a contracted version of the mother wavelet function and then allows the analysis of high frequency components. A large value of the scaling factor stretches the basic function and provides the analysis of low-frequency components of the signal.

The discrete wavelet transform (DWT) is defined as a convolution between the analyzed signal and discrete dilation and translation of a discrete wavelet function. In its most common form, the DWT applies a dyadic grid (integer power of 2 scaling with '*s*' and '*l*') and orthonormal wavelet basis function: .

$$\psi_{(s,l)}(x) = 2^{-\frac{s}{2}} \psi\left(2^{-s} x - l\right) \tag{2}$$

The variables *s* and *l* are integers that scale and translate the mother function ψ to generate wavelets (analyzing functions). The scale index *s* indicates the wavelet's width, and the location index *l* gives its position. The mother wavelets are rescaled, or "dilated" by powers of two, and translated by integer '*l*' values. In this case we have a dyadic decomposition structure. These functions define an orthogonal basis, the so-called wavelet basis [3], [5]. The Discrete Wavelet Transform (DWT) decomposition of the signal into different frequency bands

(according to Mallat's algorithm [2]) can be performed by successive high-pass and low-pass filtering (digital FIR structures) in the time domain followed by downsampling to eliminate the redundancy, as shown in *Fig.1*.

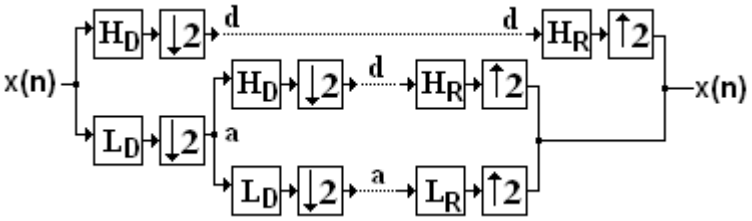


Figure: Second order dyadic scale decomposition and reconstruction.

In discrete wavelet analysis of $x(t)$ is decomposed on different scales, as follows:

$$x(t) = \sum_{j=1}^{K} \sum_{k=-\infty}^{\infty} d_j(k)\psi_{j,k}(t) + \sum_{k=-\infty}^{\infty} a_K(k)\varphi_{K,k}(t), \tag{3}$$

where $\psi_{j,k}(t)$ are discrete analysis wavelets and $\varphi_{K,k}(t)$ are discrete scaling functions, $d_j(k)$ are the detailed wavelet coefficients at scale 2^j and $a_K(k)$ are the approximated scaling coefficients at scale 2^K. The discrete wavelet transform can be implemented by the wavelet and scaling filters:

$$L(n) = \frac{1}{\sqrt{2}}\langle \varphi(t), \varphi(2t-n) \rangle, \tag{4}$$

$$H(n) = \frac{1}{\sqrt{2}}\langle \psi(t), \varphi(2t-n) \rangle = (-1)^n L(n), \tag{5}$$

which are quadratic mirror filters [3]. The estimation of the detail signal at level j will be obtained by convolving the approximation signal at level j-1 with the coefficients $L(n)$. Convolving the approximation coefficients at level j-1 with the coefficients $H(n)$ gives an estimate for the approximation signal at level j. This results in a logarithmic set of bandwidths. *Fig.2* shows the wavelet decomposition tree, the time-frequency blocks and the resulted bandwidths for a third order DWT decomposition. The first stage divides the spectrum into two equal parts. The second stage divides the lower part in quarters and so on. This results in a logarithmic set of bandwidths.

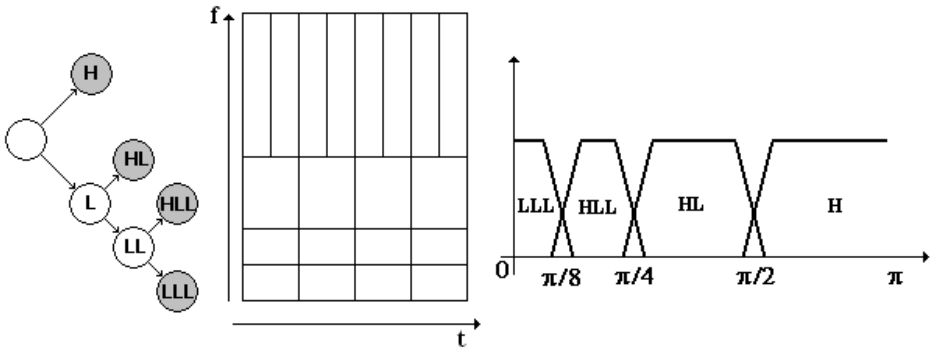


Figure 2: Time-frequency blocks and resulted relative bandwidths for third order dyadic scale decomposition.

The wavelet packet analysis is a generalization of discrete wavelet analysis providing a redundant decomposition procedure, where both detail and approximation signals are split at each level into finer components. This produces a decomposition tree as shown in *Fig. 3*.

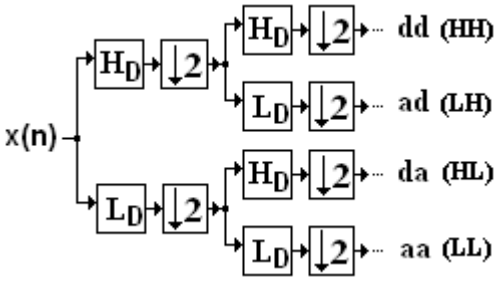


Figure 3. Wavelet packet decomposition tree.

The tree contains several admissible bases one of which is the wavelet basis itself. Having a large but finite library of bases it is possible to extract the best basis relatively to some criterion. The best basis algorithm finds a set of wavelet bases that provide the most desirable representation of the data relative to a particular cost function which may be chosen to fit a particular application [7]. This basis can be any subtree of the initial entire tree. The reconstruction procedure is similar to the inverse wavelet transform.

Figure 4 shows the time-frequency blocks for a second order wavelet packet decomposition, L and H are the low- and highpass filters with downsamplers.

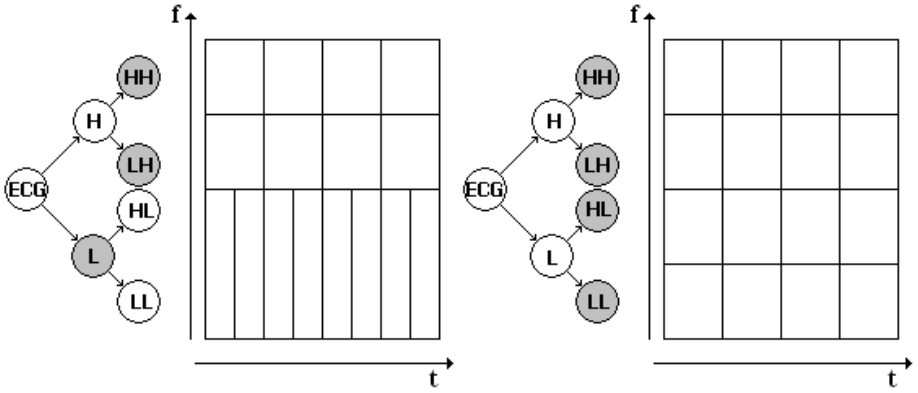


Figure 4: Time-frequency blocks for second order dyadic wavelet packet decomposition.

3. The proposed non-linear filtering method

The goal of any denoising procedure is to extract the useful signal from the noisy one, by eliminating the identified noise. The most simple model of the noisy signal is the superposition of the signal and a Gaussian type of noise. The main idea of non-linear filtering is to use the time-frequency localization properties of the discrete wavelet decomposition. The non-linear denoising approach assumes that every wavelet coefficient contains noise and it is distributed over all scales. The non-linear thresholding means discarding the detail coefficients exceeding a certain limit. There are two types of thresholding, the soft and the hard methods. With hard thresholding the coefficients which are lower than the threshold are set to zero. In soft thresholding, the remaining non-zero coefficients are shrunk toward zero. In this paper we assume that the identified noise is contained by the first order detail coefficients. It can be observed on the first order decomposition of an electrocardiogram signal that the first order detail coefficients shows some correlation with some characteristic points of the signal, like local maxima as shown in *Fig. 5*. The main idea in this work is to reduce this correlation by an extra decomposition followed by a nonlinear thresholding.

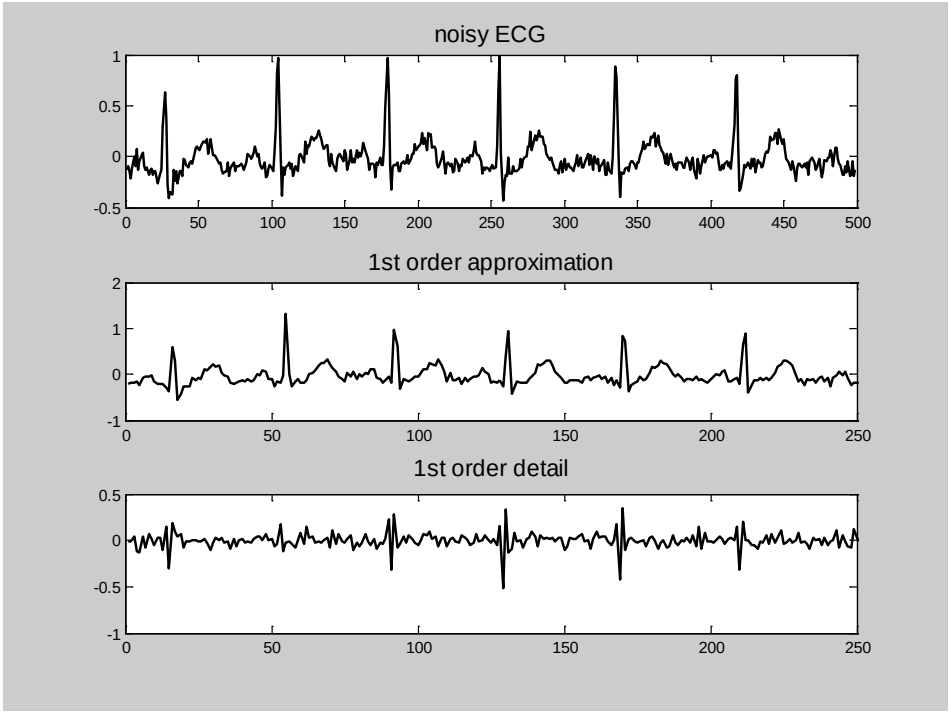


Figure 5: Correlation between first order detail and approximation coefficients.

The proposed non-linear filter acts by thresholding the detail coefficients in a particular way. The estimated noise, n_e, assumed to be the first order detail part is decomposed in a discrete wavelet structure, after that it is reconstructed only from the second order detail coefficients and is subtracted from the initial noise. The proposed smoothing procedure consists of:

1. third level DWT decomposition of the noisy ECG, resulting 3 detail coefficient sets (H, HL, HLL,) and an approximation coefficient set LLL;
2. two step second level decomposition of the estimated noise (H), resulting HHH, LHH, LH;
3. reconstruction of H' only from HHH, LHH;
4. the detail coefficients HL, HLL are thresholded;
5. reconstruction of the filtered signal through a third order Inverse Wavelet Transformation.

The procedure can be seen in *Fig. 6*. Only the detail coefficients were thresholded, the estimated correlation between noise and signal (LH) was removed. Thresholding the average coefficients (LLL) can lead to another type of filtering and can be the subject of another research paper. Both of hard and soft thresholds have been applied [4], [6].

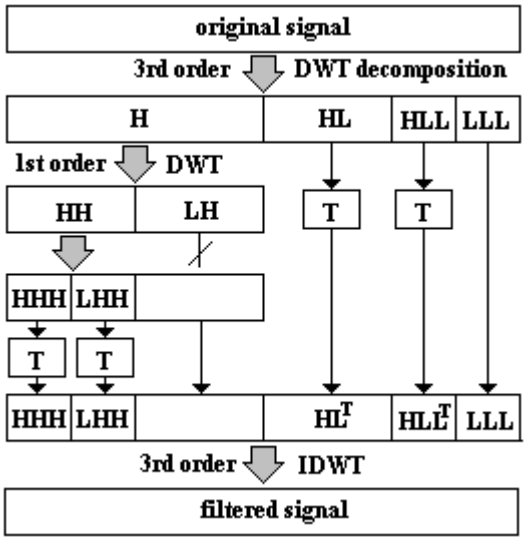


Figure 6: The proposed non-linear filtering procedure.

4. Results

To estimate the ability of this denoising procedure the followed parameters were the obtained signal to noise ratio and the absolute value of the error defined as:

$$SNR_1[dB] = 10\lg\left(\frac{PoriginalECG}{PoriginalECG - PdenoisedECG}\right) \tag{6}$$

$$Error = abs(originalECG - denoisedECG) \tag{7}$$

Figure 7 presents the original signal, the wavelet transform based denoised (soft thresholding) signal and the result of proposed nonlinear filtering. One can see (visual analysis) that the new method preserves more accurate information about signals characteristic points than the DWT based procedure. *Figure 8* shows the obtained signal-to-noise ratios by different denoising methods. The results show that the proposed method performs better denoising if the signal has lower signal-to-noise ratio.

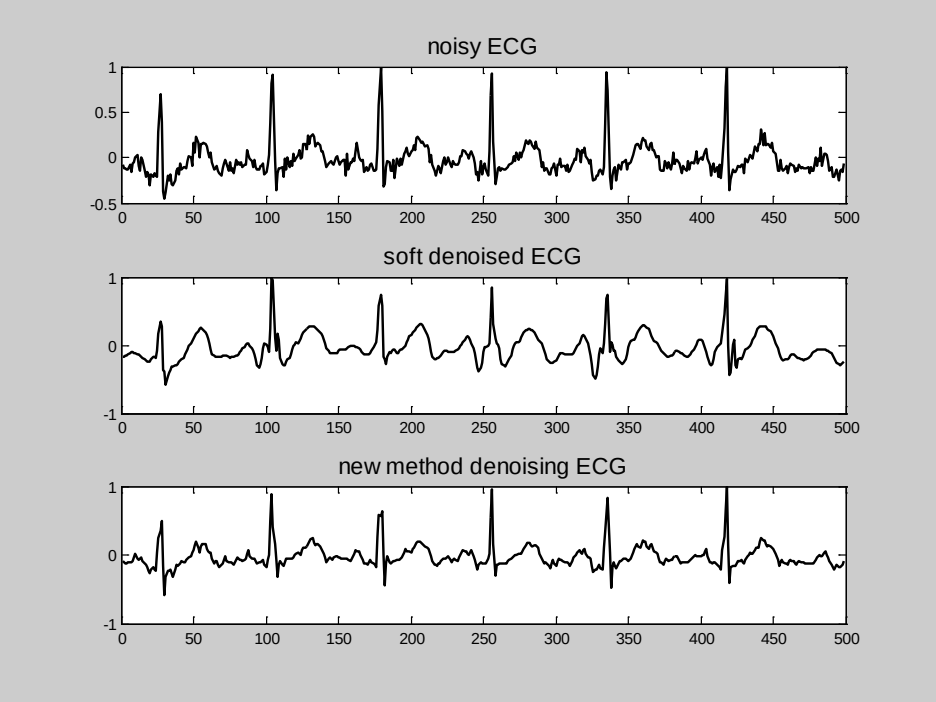


Figure 7: The results obtained using different denoising procedures.

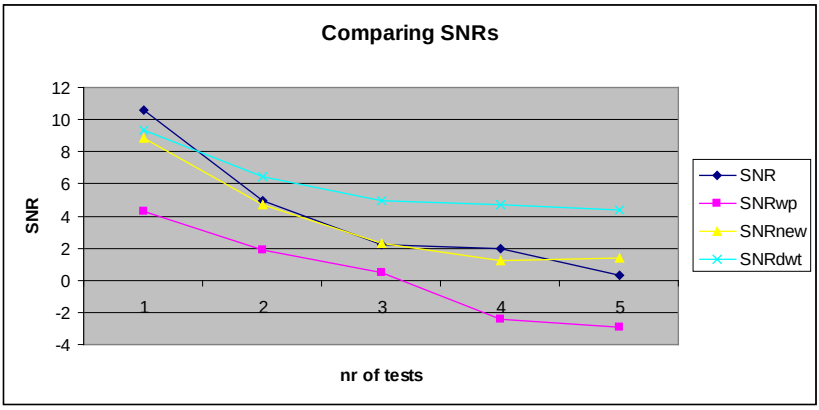


Figure 8: Comparison between signal-to-noise ratios obtained by different denoising methods.

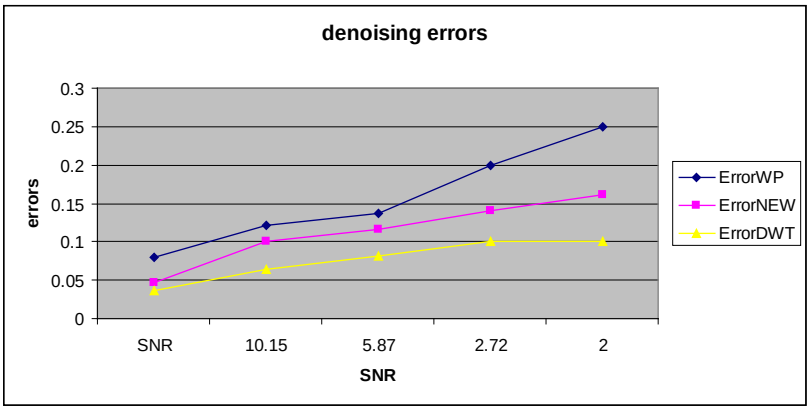


Figure 9: Comparison between different denoising errors.

Figure 7 presents the filtering errors for different methods, the proposed procedure seems to be slightly better than the wavelet packed based denoising method.

5. Conclusions

The main idea was to estimate the correlation between the noise and the signal. The discrete wavelet decomposition algorithm offers a good opportunity to have access to different time-frequency domains in order to perform non-linear filtering. An extra decomposition of the noise was used to reduce this correlation. This method was compared with ordinary wavelet decomposition and wavelet packet decomposition based filtering techniques. Wavelet and wavelet packets based denoising methods gave different performances, due to the different division strategies of the signal decomposition structures.

References

[1] Donoho, D. L., "De-noising by soft-thresholding", *IEEE, Transaction on Information Theory*, vol 41, no 3, pp. 613-627, 1995.

[2] Aldroubi, A., and Unser, M.: "Wavelets in Medicine and Biology", CRC Press New York 1996.

[3] Misiti, M., Misiti, Y., Oppenheim, and G., Poggi, J-M.: "WaveletToolbox. For Use with Matlab. User's Guide", Version 2, The MathWorks Inc 2000.

[4] Coifman, R. R., and Wickerhauser, M. V., "Entropy-based algorithms for best basis selection", *IEEE Transaction on Information Theory*, vol. 38, no 2, pp. 713–718, 1992.

[5] Mallat, S. A., “A theory for multi-resolution signal decompositions: The wavelet representtation”, *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674-693, 1989.

[6] Donoho, D. L., and Johnstone, I. M. “Ideal spatial adaptation by wavelet shrinkage”, *Biometrika, Engineering in Medicine and Biology 27th Annual Conference*, Shanghai, China, September 1-4, vol. 81, 2005, pp. 425-455.

[7] Chang, C. S., Jin, J., Kumar, S., Su, Q., Hoshino, T., Hanai, M., and Kobayashi, N., “Denoising of partial discharge signals in wavelet packets domain”, *IEEE Proceedings of Science, Measurement and Technology*, vol. 152, no. 3, pp. 129-140, 2005.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 146-158

Microstructural Modification of $(Ti_{1-x}Al_xSi_y)N$ Thin Film Coatings as a Function of Nitrogen Concentration

Domokos BIRÓ [1], Sándor PAPP [2], László JAKAB-FARKAS [2]

[1] Department of Mechanical Engineering, Faculty of Technical and Human Sciences, Sapientia University, Tg. Mureş, e-mail: dbiro@ms.sapientia.ro
[2] Department of Electrical Engineering, Faculty of Technical and Human Sciences, Sapientia University, Tg. Mureş, e-mail: spapp@ms.sapientia.ro, jflaci@ms.sapientia.ro

Manuscript received November 02, 2010; revised November 24, 2010.

Abstract: In the past few years a considerable research activity has been directed towards understanding the structure forming phenomena of nanostructured materials. A wide range of composition and structure of multiphase material systems have been investigated in order to allow a fine tuning of their functional properties. We have studied one of the most promising material systems composed of Al, Ti, Si, N, where a significant reduction in grain growth was achieved through control of phase separation process. Nanostructured (Al, Ti, Si)N thin film coatings were synthesized on Si(100) and high speed steel substrates by DC reactive magnetron sputtering of a planar rectangular Al:Ti:Si=50:25:25 alloyed target, performed in Ar/N_2 gas mixture. For all the samples we have started with deposition of a nitrogen-free TiAlSi seed layer. Cross-sectional transmission electron microscopy investigation (XTEM) of as-deposited films revealed distinct microstructure evolution for different samples. The metallic AlTiSi film exhibited strong columnar growth with a textured crystalline structure. Addition of a small amount of nitrogen to the Ar process gas caused grain refinement. Further increase of nitrogen concentration resulted in fine lamellar growth morphology consisting of very fine grains in close crystallographic orientation showing up clusters of the chain-like pearls in a dendrite form evolution. Even higher N concentration produced homogeneous compact coating, with an isotropic structure in which we can observe nanocrystals with average size of ~3nm. The kinetics of structural transformations is explained in the paper by considering the basic mechanism of spinodal decomposition process.

Keywords: Nanostructured (Ti,Al,Si)N thin films, cross-sectional transmission electron microscopy (XTEM) investigation, grain refinement, lamellar growth morphology, spinodal decomposition.

1. Introduction

In the last decade intense research activity was devoted to investigate nanocomposite coating materials, consisting of a nanocrystalline transition metal nitride and an amorphous tissue phase. These coating materials are characterized by high hardness [1], enhanced elasticity [2] and high thermal stability [3], which define their unusual mechanical and tribological properties. Various studies revealed that in multiphase nanocomposite materials the microstructure and the ratio between hardness and elastic modulus H/E are important in the coating performance [4]. Recently the most studied material is the quaternary (Ti, Al, Si)N nitride system revealing the most promising results.

As it has been suggested by *Veprek* [5], in nanocomposite materials the structure and size of the nanocrystalline grains embedded in the amorphous tissue phase together with the high cohesive strength of their interface, are the main parameters which control the mechanical behavior of the coatings. The reported results revealed that adatom mobility may control the microstructure evolution in multi-elemental coating systems, where the substrate temperature and the low energy ion/atom arrival ratio have significant effect on the growth of nanocrystalline grains.

The microstructure and growth mechanism of arc plasma deposited TiAlSiN (35 at.% Ti, 42 at.% Al, 6.5at.% Si) thin films were investigated by *Parlinska et al.* [6, 7]. It was shown that compositionally graded TiAlSiN thin films with Ti-rich zone close to the substrate exhibited crystalline structure with pronounced columnar growth. Addition of Al+Si leads to a grain refinement of the coatings, and a further increase of the Al+Si concentration results in the formation of nanocomposites, consisting of equiaxial, crystalline nanograins surrounded by a disordered, amorphous SiN_x phase.

In our study (Al, Ti, Si)N single layer thin film coatings were deposited on Si(100) and high-speed steel substrates by DC reactive magnetron sputtering. We investigated the micro structural modification of $(Ti_{1-x}Al_xSi_y)N$ thin film coatings as a function of nitrogen concentration by conventional transmission electron microscopy.

2. Experimental details and characterization technique

Deposition experiments of (Al, Ti, Si)N quaternary nitride coatings were carried out in a laboratory scale equipment by DC driven magnetron sputtering, whose details are reported elsewhere [8]. The three independently operated sputter sources were closely arranged side by side on the neighboring vertical walls of a 75 *l* octagonal all-metal high vacuum chamber. The closely disposed UM magnetrons arranged on an arc segment were highly interacting by their

magnetic fields, leading to a far extended active plasma volume. In the presented deposition experiments only the central magnetron source was active, while the adjacent two magnetron sources contributed only in the closed magnetic field. A high purity planar rectangular target material of alloyed AlTiSi was used. Elemental composition of the PLANSEE GmbH. alloyed target was 50 at.% Al, 25 at.% Ti, and 25 at.% Si, with 165×85×12 mm^3 in size, which was partially covered on the erosion zone with a high purity 99.98% Ti sheet. Prior to deposition in the vacuum chamber a base pressure of $2 \cdot 10^{-4}$ Pa was established by operating a 540 l/s turbo molecular pump.

Polycrystalline high-speed steel (HSS) substrates were used for tribological measurements, and native SiO_2 covered mono-crystalline <100> Si wafers were also used as substrates for XTEM microstructure investigation of the as-deposited (Al, Ti, Si)N single layer thin film coatings. The target-to-substrate distance was kept constant at 110 mm in all runs. The substrates were positioned in static mode on a molybdenum sheet substrate holder, which allowed application of $U_s = -75$ V bias voltage. The Mo sheet was externally heated to a controllable substrate temperature of $T_s = 400$ ºC.

Prior to the starting of the deposition process, the surface of the substrates was plasma-etched by a DC glow discharge in argon for 10 min at 0.8 Pa, while the bias voltage was limited up to 350 V. During the ion etching of substrates, the target surfaces were also sputter cleaned by operating the magnetron unit at limited discharge power (pre-sputtering power of 150 W). The substrate surfaces were shielded during the pre-sputtering. The reactive sputtering process was performed in a mixture of Ar and N_2 atmosphere at 0.28 Pa pressure. During the reactive sputtering process the nitrogen mass flow rate was controlled with an Aalborg DFC 26 flow controller, which contains a solenoid valve. The argon gas throughput ($q_{Ar} = 6.0$ sccm, measured by GFM 17 Aalborg mass flow meter) was adjusted by a servo motor driven mass flow rate controller (MFC-Granville Phillips S 216).

A constant sputtering power with a current density of 10 $mA \cdot cm^{-2}$ was selected for about 10 min sputter cleaning of the targets. During deposition the discharge power at the target surface was raised to 500 W, and the development of coating started with the deposition of a 50 nm thick AlTiSi metallic seed-layer performed in pure Ar atmosphere. In the next step of deposition an (Al, Ti, Si)(N) interlayer with gradient composition was reactively grown, while PC controlled N2 flow rate was increased slowly up to the pre-selected value. The argon gas flow was kept constant at 6.0 sccm. The typical thickness of the coatings was approximately 2 μm.

The experimental conditions for preparation of (Al, Ti, Si)N coatings are listed in *Table 1*. The microstructure and growth morphology of the as-deposited coatings was examined by use of a 100 kV operated JEOL 100U

transmission electron microscope. In order to prepare cross-sectional XTEM samples for transverse observations, the samples were subjected to ion-milling in view of thinning up to electron beam transparency. Thin specimens for XTEM investigations were prepared in a Technoorg-Linda Ltd. model 4IV/H/L ion beam thinning unit. High energy ion beam thinning was completed with a low angle and low energy (200 eV) ion beam process in order to eliminate the amorphous by-products and etching defects induced by the high energy ions.

Table 1: Summary of deposition parameters used for preparation of (Al, Ti, Si)N coatings: P_d- DC magnetron discharge power, q_{N2}- nitrogen mass flow rate, T_s- substrate temperature, U_s- substrate bias voltage.

Samples	P_d [W]	q_{N2} [sccm]	T_s [°C]	U_s [V]
TiS_01	500	-	400	-75
TiS_07	500	1.0	400	-75
TiS_08	500	1.0	400	-75
TiS_04	500	2.0	400	-75
TiS_09	500	2.0	400	-75
TiS_10	500	2.0	100	-75
TiS-06	500	3.0	400	-75
TiS-03	500	4.0	400	-75
TiS-05	500	5.0	400	-75
TiS-02	500	6.0	400	-75

Bright-field (BF) and dark-field (DF) transmission imaging techniques were used for microstructure investigation of the as-prepared samples. The identification of the crystallographic phases and the crystal orientation were also performed by evaluation of selected area electron diffraction (SAED) patterns. The SAED patterns were processed with the ´Process-Diffraction´ software tool developed by *Labar* [9].

3. Results and discussion

In this XTEM study of our prepared (Al, Ti, Si)N coatings, we combined direct imaging and selected area electron diffraction modes (SAED), which

facilitate obtaining information on the microstructure morphology, grain size and crystallographic preferred orientation.

Cross-sectional image of the metallic polycrystalline AlTiSi coating's columnar structured morphology is given in *Fig. 1*. The micrograph shows that the crystallite in a conical shape evolution starts close to the substrate and grows in a competitive mode. The columns with crystalline grains grow through the entire film thickness up to the top surface of the coating. The columnar grains are normally oriented to the substrate's surface. The large AlTi(Si) crystallites of approximately 80 nm in width are separated by the more electron-transparent $TiSi_2$ phase segregated to the grain boundaries. The SAED patterns taken from the bulk region of the film exhibit well-defined spotted diffraction rings (not to be seen here). The SAED pattern taken from the near substrate region of the coating proved crystalline character of the Si doped TiAl film (inset of *Fig. 1.*). The phases that can be derived from the diffraction pattern are mainly fcc-B1 NaCl-type of TiAlSi solid solution crystallites. The simulation of the diffraction rings was performed taking into account an fcc-type structure. It can be clearly seen the <200> preferential growth direction, indicated by significant brightness increase due to reflections from (200) crystallite planes that are oriented in parallel to the growing surface.

The chemical composition of the as deposited TiAlSi thin film was evaluated from EDS spectra analysis, and found to be of 34 at.% Ti, 46 at.% Al and 20 at.% Si.

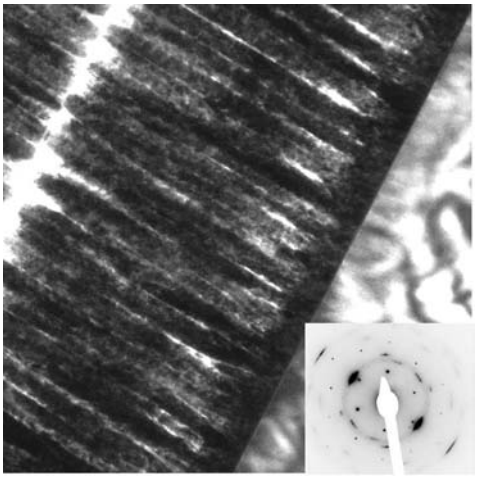


Figure 1: XTEM micrograph showing a cross-sectional view of the columnar structured polycrystalline TiAlSi thin film coating (TiS_01 sample). The inset of SAED electron diffraction pattern indicates an fcc-structured TiAlSi solid solution phase, showing (200) texture evolution in the growing direction.

By adding a small amount of nitrogen as reactive gas in the argon process gas, the growth morphology of the film dramatically changed (*Fig. 2*).

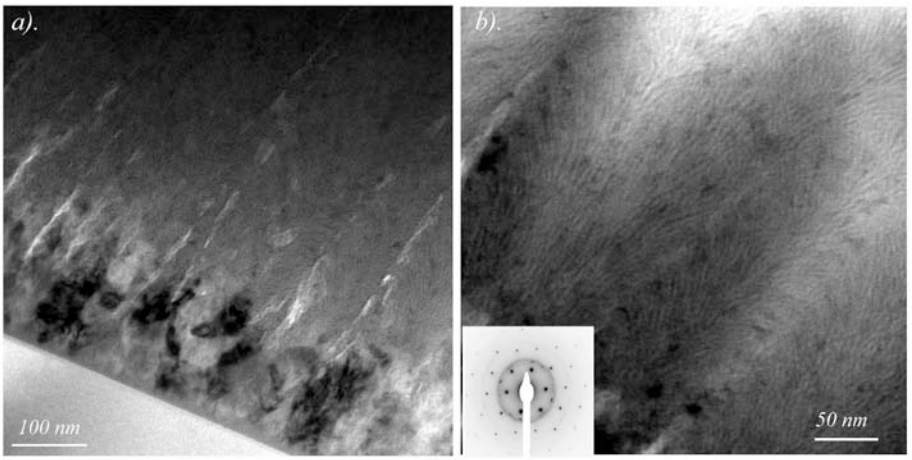


Figure 2: XTEM micrograph and SAED electron diffraction pattern of the (AlTiSi)N coating grown by nitrogen flow rate of q_{N2}=2 sccm (sample TiS_09): a). Bright field (BF) image indicates a weakly columnar structure evolution in close vicinity of transition zone from the ternary TiAlSi sub-layer to the quaternary (AlTiSi)N overgrown layer, b). On the enlarged micrograph slightly curved fine lamellar growth morphology could be identified inside the individual columns.

For a nitrogen flow rate of q_{N2} = 2 sccm the microstructure indicates a weak columnar evolution (*Fig. 2a*). Slightly curved fine lamellar growth morphology could be identified inside of the individual columns (see on the enlarged micrograph, *Fig. 2b*).

Selected area electron diffraction pattern (SAED) performed in close vicinity of transition zone –including also the Si(100) bulk–, claims for a two-phase mixture of fcc-TiAlN nanocrystals embedded in an amorphous tissue phase (inset of *Fig. 2b*). Furthermore, (200) preferential growth in close vicinity of transition zone from the ternary TiAlSi sub-layer to the quaternary (AlTiSi)N overgrown layer was slightly maintained. The presence of continuous reflection rings suggests a grain refinement of the coating with a strong tendency for evolution from the textured polycrystalline phase to a mixture of nanocrystalline AlTi(Si)N phase and possible formation of silicon nitride amorphous tissue phase.

Chemical composition of the as deposited (Ti, Al, Si)N thin films was evaluated from EDS spectra, and found to be 23 at.% Ti, 46 at.% Al, 26 at.%. Si, and about 5 at.% N.

With further increase of nitrogen amount in the coating deposition process (TiS_05 sample performed by q_{N2} =5 sccm nitrogen flow rate, which determined in our reactive magnetron sputtering process a p_N = 0.0016 mbar nitrogen partial pressure) the crystalline character of the coating disappeared and formed an isotropic nanocomposite structure, possibly consisting of nanocrystalline $Ti_3AlN/TiSi_2$ grains of 2…3 nm in size surrounded by Si_xN_y and/or AlN amorphous matrix phase (*Fig. 3a*). The SAED diffraction pattern revealed that an increased nitrogen amount leads to a nanocomposite structure, consisting of equiaxially distributed $Ti_3AlN/TiSi_2$ nanocrystalline grains surrounded by a very thin amorphous Si_3N_4 phase.

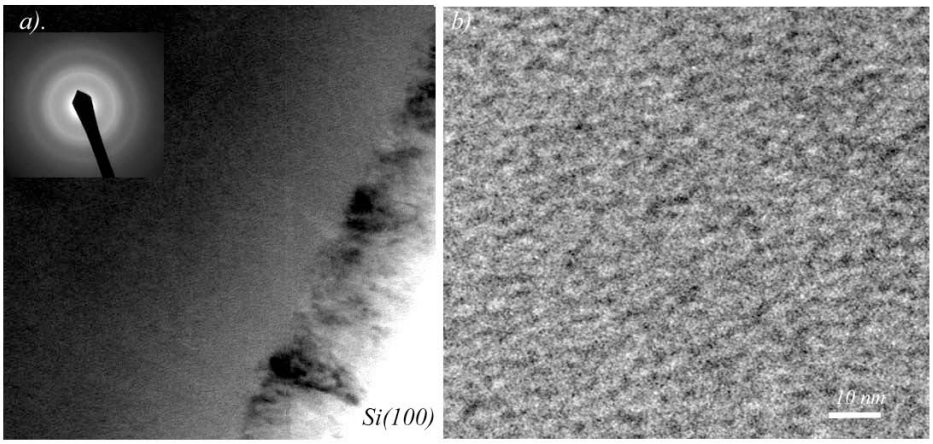


Figure 3: Bright field XTEM micrograph of (AlTiSi)N thin film deposited with an increased nitrogen flow rate (TiS_05 sample, q_{N2}=5 sccm): a). The coating's microstructure indicates the development of a competitive columnar evolution of the ternary TiAlSi sub-layer followed by the growth of the quaternary (AlTiSi)N overgrown layer developed in an isotropic morphology. b). The enlarged micrograph clearly shows a random distribution of very fine *nc*-($Al_{1-x}Ti_x$)N grains, having an average size of ~ 3 nm, with disordered grain limiting boundaries.

The chemical composition of the as deposited thin film was evaluated from EDS spectra analysis, and found to be 12 at.% Ti, 19 at.% Al, 23 at.% Si and 46 at.% N. The oxygen impurity content decreased to about 0.2 % which was related to a prolonged outgassing process of the vacuum chamber and thermal degassing of the substrate by heating to 600 °C prior to the deposition process.

Veprek et. al [10, 11] in their recently published review paper emphasized that ultra-hard nanocomposite nitride phase coatings based on (Ti, Al, Si)N elemental composition can be managed by well-controlled plasma and deposition conditions. The development in a periodic structure of the

nc-$(Al_{1-x}Ti_x)N$/*a*-Si_3N_4 nanocomposite coatings, composed from the uniformly distributed $(Al_{1-x}Ti_x)N$ nanocrystals and amorphous tissue phase of Si_3N_4 with about one monolayer (ML) thickness, was explained by the spontaneous separation in spinodal decomposition and self-organization upon phase segregation process. The strong immiscibility of Si_3N_4 phase in crystalline TiAlN phase, as well the absence of the Ostwald ripening, are appropriate conditions for the development of nanocomposite coatings.

Favvas and *Mitropoulos* [12] in their paper (see also the cited paper therein) compiled the IUPAC definition of spinodal decomposition: "A clustering reaction in a homogeneous, supersaturated solution (solid or liquid) which is unstable against infinitesimal fluctuations in density or composition. Therefore, homogeneous solution separates spontaneously into two phases, starting with small fluctuation and proceeding with decrease in the Gibbs free energy without a nucleation barrier. "

It is known from thermodynamic theory that by fast cooling of a homogeneous solution the diffusion process occurs with net reduction in Gibbs free energy of the system. The free energy of mixing, ΔG^{mix}, defined by the thermodynamic equation of Gibbs has a general form:

$$\Delta G^{mix} = \Delta H^{mix} - T \cdot \Delta S^{mix}, \tag{1}$$

where the enthalpy change ΔH^{mix} is associated with the interactions between the components of the mixture, and the entropy change ΔS^{mix} is associated with the random mixing of components.

At high temperature of the system, with partially miscible components *A* and *B*, the components give rise to a continuous solution (in a liquid or solid phase) due to a complete solubility. At lower temperature the solution becomes unstable and may exists compositional ranges where the co-existing solid phases are more stable.

Therefore, during the cooling of liquid, phase separation proceeds in order to minimize the free energy. If the super-cooling of the homogeneous solution takes place into the coherent spinodal region, defined by the compositional interval between points of inflexion on the free energy diagram as a function of molar composition $G = f(X_A)$, the phase separation process occurs by the decomposition of the homogeneous solution. Decomposition is favored by small fluctuations produced in the chemical composition or infinitesimal perturbations in chemical potential. In accordance with Fick's first law, the diffusion flux of a component, e.g. j_A is proportional to the concentration gradient of the respective component *A*, $\frac{\partial C_A}{\partial x}$:

$$j_A = -D_A \cdot \frac{\partial C_A}{\partial x}, \tag{2}$$

where D_A stands for diffusion coefficient in the first empirical law of Fick.

The diffusion flux j_A can be driven also by the free energy gradient of component *A*:

$$j_A = -C_A \cdot \mu_A \cdot \frac{\partial G_A}{\partial x}, \tag{3}$$

where μ_A stand for the mobility constant of component *A*.

From the above equations the diffusion coefficient D_A can be written as a derivative function of the Gibbs free energy in respect to the concentration:

$$D_A = C_A \cdot \mu_A \cdot \frac{\partial G_A}{\partial C_A} \tag{4}$$

If the diffusion coefficient is positive, $D_A > 0$, i.e. $\frac{\partial G_A}{\partial C_A} > 0$, the chemical potential gradient has the same direction as the concentration gradient, therefore the diffusion flux occurs along the concentration gradient. For diffusion coefficient $D_A < 0$, i.e. $\frac{\partial G_A}{\partial C_A} < 0$, the diffusion flux occurs against to the concentration gradient.

By correlating the phase composition diagram (i.e. a diagram of phases, where the dependence for temperature *T* versus the molar fractions X_A and X_B of components indicates the composition ranges for equilibrium phases) and the diagram of the free energy change versus the molar fraction of the mixture, it can be seen that below the spinodal (where the second derivative of the free energy of mixing versus the molar fraction X_A is zero, $\frac{\partial^2 \Delta G}{\partial X_A^{\,2}} = 0$) the system is unstable (*Fig. 4*).

For negative values of the second derivative of the free energy of mixing versus the molar fraction X_A, e.g. of component *A*, i.e. $\frac{\partial^2 \Delta G}{\partial X_A^{\,2}} < 0$, the homogeneous supersaturated solution is unstable and spinodal decomposition may proceed.

The spinodal decomposition process happens in an unstable region, where further instability is caused by the small fluctuation occurred in the local concentration, while the diffusion process takes place from the lower to higher concentration (i.e. “up-hill” diffusion).

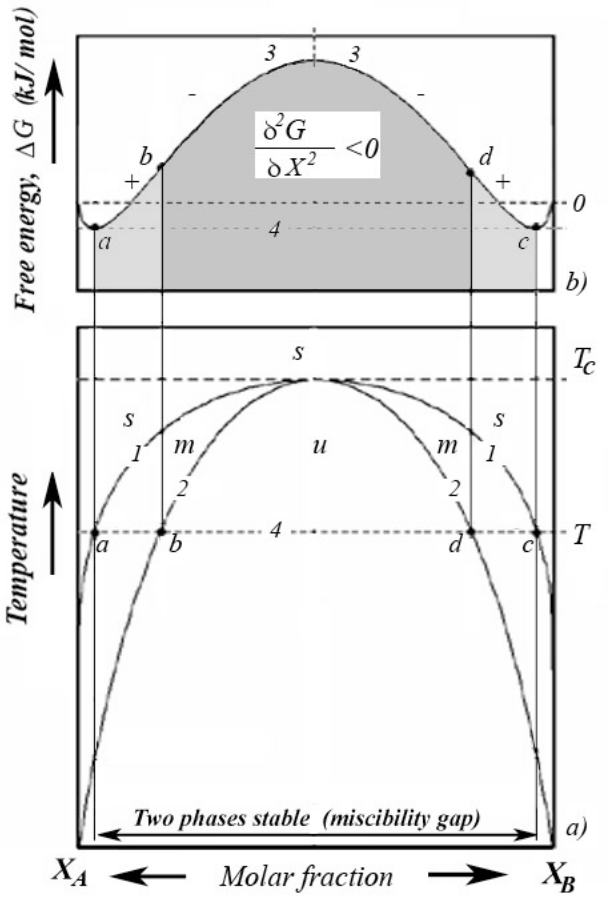


Figure 4: Illustration of the spinodal decomposition on the phase separation mechanism: *a*). Equilibrium phase diagram with the phase boundary limits: for a temperature *T* of the system points *a* and *c* stand for compositional interval of miscibility gap (with two stable solid phases), the interval between points *b* and *d* stand for the spinodal limits; *b*). Diagram of the Gibbs free energy changes in the mixing process of two partially miscible components *A* and *B*, boundary line (1) separates a stable region (*s*) of the homogeneous liquid phase and the metastable region (*m*) of the precipitated solid phases. Line (2) is the limit of spinodal, below of which the thermodynamically unstable system (*u*) turns from one phase (liquid) to a mixture of two phases (solid) system with different elemental composition

Therefore, spinodal decomposition process is a spontaneous reaction, which is the main route to develop periodic structures with uniform size.

Our experimental results on fine lamellar growth morphology of (Ti, Al, Si)N nitride coatings, consisting of chain-like pearls in a dendrite evolution with very fine grains in close crystallographic orientation, may be explained in accordance with Veprek's theory by partial spinodal decomposition and phase segregation during the film's growth while percolation threshold composition is attained by an increased nitrogen activity.

On the other hand, the increase of deposition rate induces a decrease of the surface mobility related to the decrease of the ion-to-atom arrival rate ratio. These particular deposition conditions explain the columnar structure of TiAlSi solid solution crystallites, which can be clearly observed in the XTEM image of TiS_01 sample. Addition of minor amounts of nitrogen leads to an encapsulation of the growing TiAl(Si)N crystallites by process segregated amorphous phase.

From the detailed observation of the SAED diffuse diffraction pattern of sample TiS_05 obtained with an increased nitrogen flow rate, the presence of an amorphous phase surrounding the Ti_3AlN nanocrystallites can be attributed to Si_3N_4 matrix phase (inset of *Fig. 3a*). The formation of amorphous $TiSi_2$ and AlN phase due to the partial segregation of Al and Si atoms should be also considered due to the effect of enhanced ion bombardment provided by the focused plasma beam that is characteristic to the present experimental conditions [8].

When the atomic surface mobility in the growing film is adequate, the segregated atoms can nucleate and develop the new phases controlled by deposition temperature and by the energy transfer from an increased incident ion-to-atom arrival rate ratio [13-15].

Further experiments are in progress to investigate the influence of the deposition temperature on structure evolution of (TiAlSi)N coatings.

4. Conclusions

In the present work it was shown that:

a) Columnar structure of polycrystalline AlTiSi thin film coating evolved by non-reactive DC magnetron sputtering applied to Al:Ti:Si = 50:25:25 alloyed target (performed in pure Ar atmosphere, where the 500 W discharge power, T_s = 400 ºC substrate temperature and U_s = –75 V bias voltage were held constants).

b) Addition of a small amount of nitrogen to the process gas leads to a grain refinement of polycrystalline (Ti, Al, Si)N thin films. Increase of N concentration (q_{N2} = 2 sccm flow rate) resulted in fine lamellar growth *m*orphology of coatings, showing chain-like pearls in a dendrite evolution, consisting of clusters of very fine grains in close crystallographic orientation.

c) Further increase in the nitrogen amount (q_{N2} = 5 sccm) leads to evolution of a nanocomposite coating consisting of crystalline Ti_3AlN nanograins in 2…3 nm size surrounded by an amorphous Si_xN_y covalent nitride and/or AlN matrix phase.

d) Kinetics of the structural transformations were explained by considering the basic mechanism of spinodal decomposition process.

Acknowledgements

The authors are thankful for the financial support of this project granted by Sapientia Foundation – Institute for Scientific Research, Sapientia University. The EDS analyses of the investigated coatings were performed in a CM 20 Philips 200kV TEM electron microscope by Professor P. B. Barna from RITPMS, Budapest. Professor P. B. Barna's contribution to investigating elemental composition and the valuable discussions are highly appreciated.

References

[1] Yoon, J. S., Lee, H. Y., Han, J., Yang, S. H., Musil, J., "The effect of Al composition on the microstructure and mechanical properties of WC–TiAlN superhard composite coating", *Surface and Coatings Technology*, vol. 142-144, pp. 596-602, 2001.

[2] Duran-Drouhin, O., Santana, A. E., Karimi, A., "Mechanical properties and failure modes of TiAl(Si)N single and multilayer thin films", *Surface and Coatings Technology*, vol. 163-164, pp. 260-266, 2000.

[3] Musil, J. and Hruby, H., "Superhard Nanocomposite $Ti_{1-x}Al_xN$ Films Prepared by Magnetron Sputtering" , *Thin Solid Films*, vol. 365, pp. 104-109, 2000.

[4] Ribeiro, E., Malczyk, A., Carvalho, S., Rebouta, L., Fernandes, J. V., Alves, E., Miranda, A. S., "Effect of ion bombardment on properties of d.c. sputtered superhard (Ti,Si,Al)N nanocomposite coatings", *Surface and Coatings Technology*, vol. 151-152, pp. 515-520, 2002.

[5] Veprek, S., "New development in superhard coatings: the superhard nanocrystalline-amorphous compositics", *Thin Solid Films.* vol. 317, pp. 449-454, 1998.

[6] Parlinska-Wojtan, M., Karimi, A., Cselle, T., Morstein, M., "Conventional and high resolution TEM investigation of the microstructure of compositionally graded TiAlSiN thin films", *Surface and Coatings Technology*, vol. 177-178, pp. 376-381, 2004.

[7] Parlinska-Wojtan, M., Karimi, A., Coddet, O., Cselle, T., Morstein, M., "Characterization of thermally treated TiAlSiN coatings by TEM and nanoindentation", *Surface and Coatings Technology*, vol. 188-189, pp. 344-350, 2004.

[8] Biro, D., Barna, P. B., Szekely, L., Geszti, O., Hattori, T., Devenyi, A., "Preparation of multilayered nanocrystalline thin films with composition-modulated interfaces", *Nuclear Instruments and Methods in Physics Research* vol. 590, pp. 99-106, 2008.

[9] Lábár, J. L., "ProcessDiffraction: A computer program to process electron diffraction patterns from polycrystalline or amorphous samples", *Proceedings of the XII EUREM*, Brno (L. Frank and F. Ciampor, Eds.), vol. III., pp. I 379-380, 2000.

[10] Veprek, S., Veprek-Heijman, M. G. J., Karvankova, P., Prochazka, J., "Different approaches to superhard coatings and nanocomposites", *Thin Solid Films,* vol. 476, pp. 1-29, 2005.

[11] Veprek, S., Zhang, R. F., Veprek-Heijman, M. G. J., Sheng, S. H., Argon, A. S., "Superhard nanocomposites: Origin of hardness enhancement, properties and applications", in *Surf. And Coat. Technol.,* vol. 204, pp. 1898-1096, 2009.

[12] E. P. Favvas, A. Ch. Mitropoulos: "What is spinodal decomposition?", *Journal of Engineering Science and Technology Review*, vol. 1, pp. 15-27, 2008.

[13] Carvalho, S., Rebouta, L., Ribeiro, E., Vaz, F., Dennnot, M. F., Pacaud, J., Riviere, J. P., Paumier, F., Gaboriaud, R. J., Alves, E., "Microstructure of (Ti,Si,Al)N nanocomposite coatings", *Surface and Coatings Technology,* vol. 177-178, pp. 369-375, 2004.

[14] Carvalho, S., Rebouta, L., Cavaleiro, A., Rocha, L. A., Gomes, J., Alves, E., "Microstructure and mechanical properties of nanocomposite (Ti,Si,Al)N coatings", *Thin Solid Films,* vol. 398-399 , pp. 391-396, 2001.

[15] Vaz, F., Rebouta, L., Goudeau, P., Pacaud, J., Garem, H., Riviere, J. P., Cavaleiro, A., Alves, E., "Characterization of TiSiN nanocomposite films", *Surface and Coatings Technology,* vol. 133-134 , pp. 307-313, 2000.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 159-165

On Some Pecularities of Paloid Bevel Gear Worm-Hobs

Dénes HOLLANDA, Márton MÁTÉ

Department of Mechanical Engineering, Faculty of Technical and Human Sciences,
Sapientia University, Tg. Mureş,
e-mail: hollanda@ms.sapientia.ro, mmate@ms.sapientia.ro

Manuscript received November 11, 2010; revised November 24, 2010.

Abstract: The paper discusses the generating surfaces of the paloid bevel worm hob used for paloid gear cutting. A pitch modification is required by this type of tool in order to ensure the optimal contact pattern by gearing. As a consequence of this modification the flank line of the plain gear tooth results as a paloid- a more general shaped curve than the theoretical involute of the basic circle. Equations of the generating surfaces result as equations of a generalized arhimedic bevel helical surface presenting those modifications that arise from the variation of the tooth thickness.

The first subsection discusses the essential geometrical peculiarities of the paloid worm hob. Here it is to remark that the most important characteristic of the tool is the variation of the tooth thickness on the rolling tape generator. The tooth thickness has its minimum value at the middle of the generator, and is maximum on the extremities. As a consequence, the generated gear tooth presents an opposite variation of thickness. The thickness variation is realized by moving the relieving tool on an ellipse, but this is not the only possible trajectory to be used.

The second subsection presents the generalized mathematical model of the tooth thickness variation. Starting from the ellipse used by classical relieving technologies, and writing the equation of the ellipse reported to the coordinate system of the paloid hob it results the radius function of the revolved surface of the reference helix. This function is used in its condensed form. With this, the developed mathematical model can be used for other forms of the relieving tool trajectory.

The next paragraphs present the matrix transformations between the coordinate systems of the hob and the relieving tool. Finally the parametric equations of the hob tooth flank are obtained. These equations depend on the radius modification function. Using other relieving trajectories that differ from an ellipse, other tooth flank forms will be obtained. Using this model, the flanks of the cut gear tooth can be easily written for all types of trajectories.

Keywords: paloid bevel worm hob, tooth thickness modification, generating surface.

1. General description

Paloid bevel gears are realized on Klingelnberg type teething machine-tools, using paloid bevel gear worm-hobs [1], [6]. These tools present a straight-shaped edge in their axial section. As a conclusion, the origin surface of the paloid worm hob is an Arhimedic bevel worm having the half taper angle of 30° as shown in *Fig. 1*. [4], [5]. The chip-collecting slots are axially driven. In order to realize the clearance angle on all edges, a helical relieving, perpendicularly oriented to the bevel generator is allowed.

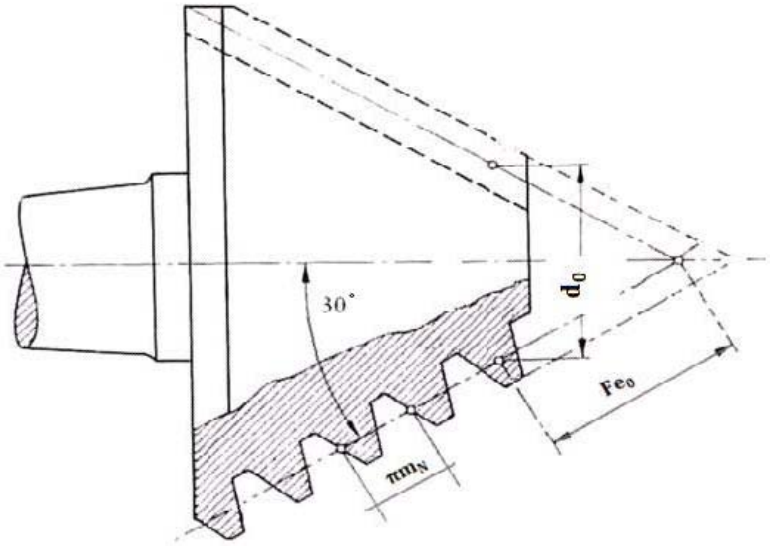


Figure 1: The paloid worm hob characteristic dimensions.

In order to ensure the good positioning of the contact patch by paloid gears, the pitch of the tool is variable but the tooth thickness (measured on the rolling taper generator) must increase at the extremities of the tool. Using this principle, the tooth thickness on the rolling tape generator increases in both directions starting from the point *N* (*Fig. 2*).

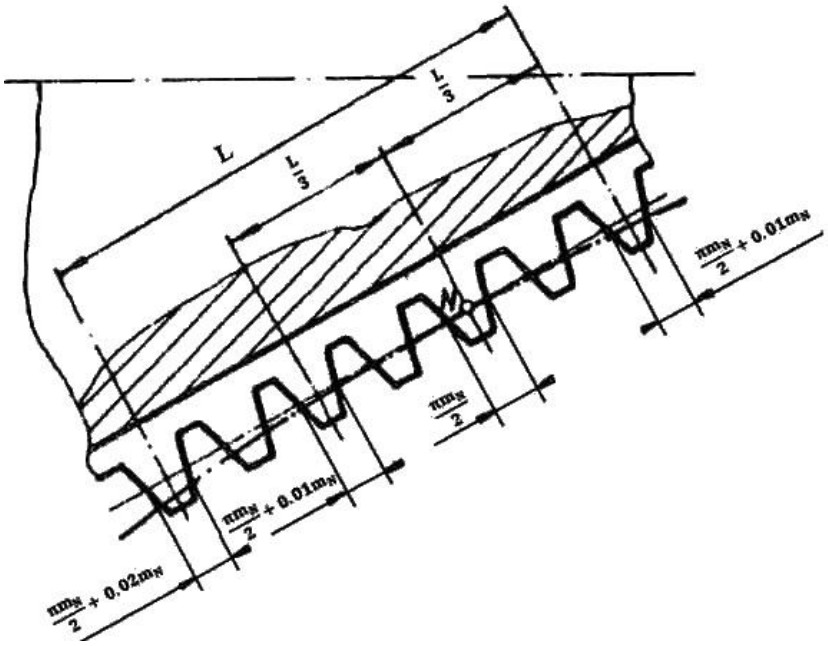


Figure 2: The tooth thickness distribution along the cone generator.

The thickness of teeth in case of gears manufactured using the worm-hob described above results smaller at the tooth extremities and larger in the middle of it. This localizes the contact pattern in the middle of the tooth-flank when gearing. This modification causes the deformation of the tooth line on the plain gear too, transforming the theoretical involute into a more general shaped curve named paloid.

Both the threading and the relieving operation are realized on a relieving machine, where the axis of the worm-hob is declined with a 30° angle reported to the axis of the chuck, in a horizontal plain containing the axis. This way the taper generator becomes parallel with the direction of the longitudinal slider.

The curved generator of the worm-hob is realized through leading the cutter slider by a profiled turning-template. The same template is used by the threading and relieving operations.

Bevel worm hobs present one thread, and allow the cutting of any teeth number by a fixed module and pressure angle. The pitch is normalized along the rolling cone generator and can be calculated using the formula: $p_N = \pi \cdot m_N$. The axial pitch (defined on the worm hob's axis) is determined by $p_A = \pi \cdot m_N \cdot \cos 30^\circ$.

2. The geometry of the tooth thickness modification

Considering the bevel worm-hob as a helical bevel surface, like any bevel thread, it is to mention that next to the axial pitch, a radial pitch can be defined, depending on the axial pitch and the half taper angle of the rolling cone (30^o). If the generator of the worm hob were a straight line, the radial pitch value would satisfy the expression $h = p_A \cdot \mathrm{tg}30^o$. However, the generator is curved and the radial pitch is calculated taking into account the fact that the endpoint of the characteristic radius must be on that generator. Considering the tool's pressure angle α, the distance between the curved and the straight generators (based on the dimensions of the tool profile indicated in [1], [2], [6]) can be calculated using the formula

$$\Delta_1 = (p_N + 0{,}01 \cdot m_N - p_N)/2 \cdot \mathrm{tg}\alpha = 0{,}01 \cdot m_N / 2 \cdot \mathrm{tg}\alpha$$

applicable for point *A* (*Fig 3*). The same distance is indicated related to point *C* (*Fig 3*). In point *D*, the distance discussed above will increase to double as follows from the formula $\Delta_2 = (p_N + 0{,}02 \cdot m_N - p_N)/2 \cdot \mathrm{tg}\alpha = 2\Delta_1$.

The curved generator is usually an arc of ellipse of which major axis (superposed to the *Ox*) is parallel to the rolling cone generator, and it's minor axis (superposed to *Oy*) passes through point *B* (*Fig 3*), situated at a distance of $F_{eo} + 2p_N$ from the top of the rolling cone.

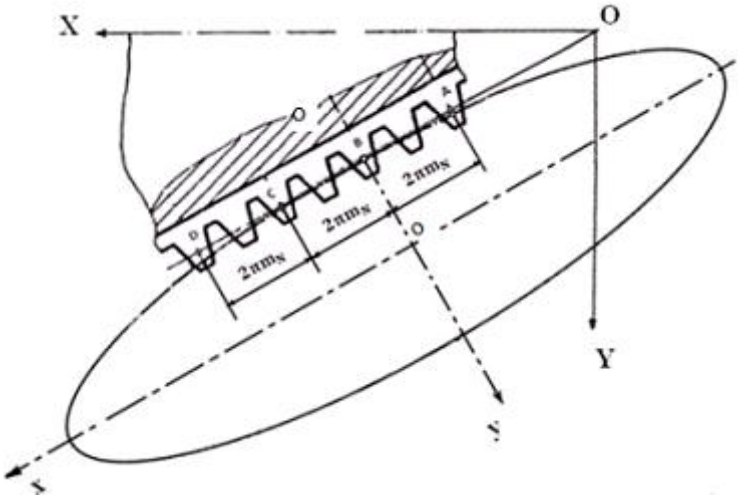


Figure 3: The elliptically curved generator.

It is to remark that the ellipse passes through the following points: $A(-p_N, a-\Delta_1)$; $B(0,a)$; $C(-2.p_N, \Delta_1)$ and $D(-4.p_N, a-\Delta_2)$.

The canonic equation of an arbitrarily shaped ellipse, reported to its axes, is:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-1=0 \tag{1}$$

Writing the equation of the ellipses passing through the points *A*, *B*, *C* and *D* a system of equation will result, of which solution leads to the reference radius endpoint function e.g. $h=f(p_N/\cos 30^o)$. In this paper the radial pitch is accepted having a general form as expressed by the above formula. As a consequence, the relations deduced in the following can be used for other forms of the curved generator too. Using this form, the essential aspects of the dependences will not be influenced by the very sophisticated expressions of the radial pitch [3].

The matrix equation of the bevel worm hob's edge, reported to the self coordinate system, is:

$$r_M=\left\|\begin{array}{c}\left(F_{eo}-\dfrac{\pi\cdot m_N}{4}\right)[\cos 30^\circ+\mathrm{tg}30^\circ\sin(30^\circ-\alpha)]+\lambda\sin(30^\circ-\alpha)\\ \lambda\cos\alpha\\ 0\\ 1\end{array}\right\| \tag{2}$$

This edge is fixed to the mobile coordinate system $O_M X_M Y_M Z_M$ superposed at the initial moment with the fixed coordinate system $OXYZ$. The parametric form of the edge expressed in the mobile system will is denoted by r_M but it keeps the form of r_t given by (2).

The surface of the generator flank of the bevel worm hob can be obtained considering the followings: the edge is fixed to the mobile coordinate system (*Fig 4*) that executes a helical motion reported to the stationary system. The origin moves in the direction of axis with an amount of $p_A\cdot\varphi$, simultaneously with a rotation by angle φ of the worm-hob (representing the workpiece in the above case).

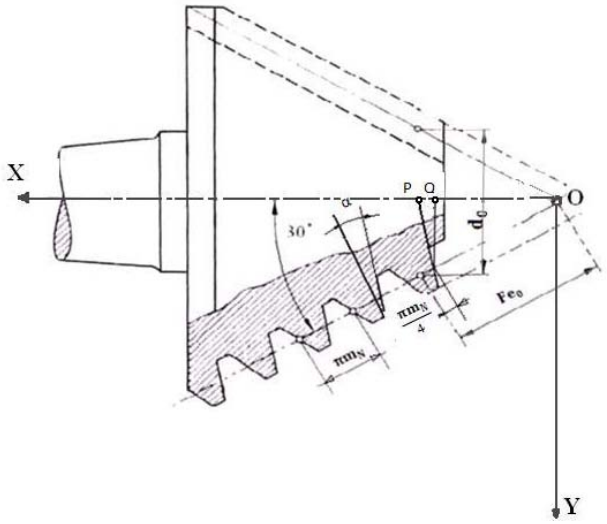


Figure 4: The position of the right edge related to the used reference system.

First the edge must be reported to an auxiliary coordinate system. Between this and the auxiliary system $O_M X_M Y_M Z_M$ there exist only translations by amounts of $p_A \cdot \varphi / 2\pi$ and $h \cdot \varphi / 2\pi$ respectively. A transfer matrix describing the above translations is:

$$M_{aM} = \begin{Vmatrix} 1 & 0 & 0 & \dfrac{p_A}{2\pi}\varphi \\ 0 & 1 & 0 & \dfrac{h}{2\pi}\varphi \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{Vmatrix} \tag{3}$$

After that the edge's equation is transformed to the stationary system, mentioning that auxiliary system executes only a rotation around its X axis, characterized by the matrix

$$M_{Oa} = \begin{Vmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi & 0 \\ 0 & \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 0 & 1 \end{Vmatrix} \tag{4}$$

The pointing vector of the bevel helical surface described by the edge, reported to the stationary reference system is obtained from the following matrix equation

$$r = M_{Oa} \cdot M_{aM.} \cdot r_M = M_{OM} \cdot r_M \tag{5}$$

Multiplying M_{Oa} by M_{aM} it results

$$M_{OM} = \begin{Vmatrix} 1 & 0 & 0 & \frac{p_A}{2\pi}\varphi \\ 0 & \cos\varphi & -\sin\varphi & \frac{h}{2\pi}\varphi.\cos\varphi \\ 0 & \sin\varphi & \cos\varphi & \frac{h}{2\pi}\varphi.\sin\varphi \\ 0 & 0 & 0 & 1 \end{Vmatrix} \tag{6}$$

Finally, the matrix expression of the pointing vector is:

$$\begin{Vmatrix} X \\ Y \\ Z \\ 1 \end{Vmatrix} = \begin{Vmatrix} 1 & 0 & 0 & \frac{p_A}{2\pi}\varphi \\ 0 & \cos\varphi & -\sin\varphi & \frac{h}{2\pi}\varphi.\cos\varphi \\ 0 & \sin\varphi & \cos\varphi & \frac{h}{2\pi}\varphi.\sin\varphi \\ 0 & 0 & 0 & 1 \end{Vmatrix} \cdot \begin{Vmatrix} X_M \\ Y_M \\ Z_M \\ 1 \end{Vmatrix} \tag{7}$$

Expression (7) realizes the matrix form of the right flank of the bevel worm-hob tooth, reported to the stationary system *OXYZ*, attached to the hob.

Analytical expression is obtained after multiplying the matrices in the equation before. Similarly results the equations of the opposite flank, if starting the calculus with the equations of the opposite edge.

References

[1] Krumme, W., "Klingelnberg-Spiralkegelräder Dritte neubearbeitete Auflage", Springer Verlag, Berlin, 1967.

[2] *** "MASINOSTROIENIE"- ENCYCLOPEDY, Vol.VII., Masghiz Moskou.

[3] Máté, M., Hollanda, D. "The Enveloping Surfaces of the Paloid Mill Cutter", in *Proc. 18th International Conference on Mechanical Engineering*, Baia Mare, 23-25 April 2010, pp. 291-294.

[4] Michalski, J., Skoczylas, L. "Modeling the tooth flanks of hobbed gears in the CAD environment", *The International Journal of Advanced Manufacturing Technology*, vol. 36, no. 7-8, 2008.

[5] Shu-han Chen, Hong –zhi Jan, Xing-zu Ming, "Analysis and modeling of error of spiral bevel gear grinder based on multi-body system theory", *Journal of Central South University of Technology*, vol. 15, no. 5, 2008.

[6] Klingelnberg, J. "Kegelräder- Grundlagen, Anwendungen", *Springer Verlag*, 2008.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 2 (2010) 166-176

Kinematic Analysis of a 6 DOF 3-PRRS Parallel Manipulator

Zoltán FORGÓ

Department of Mechanical Engineering, Faculty of Technical and Human Sciences,
Sapientia University, Tg. Mureş, e-mail: zforgo@ms.sapientia.ro

Manuscript received October 14, 2010; revised November 08, 2010.

Abstract: The number of parallel mechanism applications in the industry is growing and the interest of the academia to find new solutions and applications to implement such mechanisms is present all over the world. In this paper, after a summarised group theory presentation, a symmetrical six degrees of freedom mechanism (3-PRRS) will be defined using this theory. Enumerating some possible kinematic chains for Schoenflies-motion, one solution is kept in order to build up the proposed mechanism. The easy way of mathematical modelling is given by the fact that the mechanism can be considered as an extended well known planar Delta manipulator. The double driven joints in each limb ensure the third translation and other two rotations of the moving platform complementing the planar motion of the Delta manipulator. After the kinematical modelling of the presented mechanism, the actuation of the links is considered. A new parallel driven actuation system is presented in order to fulfill the rotation and translation movements required for the PRRS limb actuation. The aspects of singular configurations, which are similar to the planar Delta mechanism singular configurations with some extensions, are considered also in the presented paper. The paper closes by enumerating some major advantages of the proposed 6 degrees of freedom manipulator.

Keywords: parallel mechanism, kinematics, group theory, Lie algebra.

1. Introduction

The number of applications in the industry which use parallel mechanisms are growing and the interest of the academia to find new solutions and applications to implement such mechanisms is present all over the world. The lower degree of freedom mechanisms which are suited for some specific tasks

are preferred because of the architecture simplicity and therefore the easy mathematical modeling and finally, but not at least for economical reasons.

The 6 degrees of freedom (DOF) parallel mechanism is introduced by Steward and Gough [1] and since then many aspects of the mechanism and its application are revealed. During the last decades more attention has been paid to the study of 6 DOF parallel mechanisms, including synthesis and analysis on kinematics, dynamics, singularities, error and workspace. Some milestones in the analysis of those mechanisms are set by Earl and Rooney using a method for synthesis of new kinematic structures [2], Hunt studied the manipulators on the basis of screw theory [3], Tsai is using systematic methodology in [4] and Hervé discussed the structural synthesis of parallel robots using the mathematical group theory [5]. More recently Shen proposed a systematic type synthesis methodology for 6 DOF kinematic structures enumerating 29 parallel structures [6]. Hereby Shen defines the hybrid single-open chains (HSOC) which are able to generate three translations and three rotation angles. Using those HSOCs four 6 DOF manipulators are presented with symmetrical arrangement of the limbs (see No.3-No.6 architectures, *Table 2.* from [6]). According to Tsai [7], the symmetry implies the use of the same number of actuators on the same positions in each limb. Moreover he says that a parallel manipulator is symmetrical if it satisfies the condition that the number of limbs is equal to the number of degrees of freedom of the moving platform. In the case of double actuated limbs (with two actuated joints) the last presented condition can be omitted. So the HSOCs defined by Shen can be replaced by serial chains which enable three translations and three rotations also.

This paper presents some kinematic structures according to the above mentioned criteria without the aim of full discussion about all the possible structures. The geometrical model of one architecture is presented as well.

2. General motion generators

The enumeration of serial topology limbs which enable the spatial motion (three translations and three rotations) is greatly simplified by using the Lie group of rigid body displacement as introduced by Hervé [8]. If each limb of a parallel manipulator generates a subset of possible displacements, which is a Lie subgroup, the intersection set is also a Lie subgroup of the mobile platform. According to this statement if the platform undergoes the spatial, general motion, each limb must ensure the three translations and three rotations. According to *Table 1* $\{D\}$ denotes the general rigid body motions for the 6 DOF mobile platform and $\{L_i\}$ denotes the displacement Lie subgroup of the i^{th} limb. The relation between them is given by:

$$\{L_1\} \cap \{L_2\} \cap \{L_3\} = \{D\}. \tag{1}$$

It is obvious that the only possibility for a true equation (1) is:

$$\{L_1\} = \{L_2\} = \{L_3\} = \{D\}. \tag{2}$$

To obtain simple mechanical structures, better symmetry and good manufacturing for the three limbs the same architecture is considered. For this reason, the analysis of the $\{L_i\}$ displacement Lie subgroup is carried out. The notations for displacement Lie subgroups are recalled in *Table 1* [9].

According to the group theory it can be stated:

$$\begin{aligned} \{L_i\} &= \{D\} = \{T\}\{S(N)\} = \\ &= \{T(\boldsymbol{u})\}\{T(\boldsymbol{v})\}\{T(\boldsymbol{w})\}\{R(N,\boldsymbol{u})\}\{R(N,\boldsymbol{v})\}\{R(N,\boldsymbol{w})\} \quad \forall N. \end{aligned} \tag{3}$$

Table 1: List of displacements Lie subgroups [4].

Lie subgroup	*Description of the subgroup*
$\{E\}$	identity
$\{T(\boldsymbol{u})\}$	translations parallel to the $\boldsymbol{u}$ vector
$\{R(N,\boldsymbol{u})\}$	rotations around the axis determined by N and $\boldsymbol{u}$
$\{H(N,\boldsymbol{u},p)\}$	helical motions with axis (N,$\boldsymbol{u}$) and the pitch p
$\{T(Pl)\}$	translations parallel to the *Pl* plane
$\{C(N,\boldsymbol{u})\}$	cylindrical motions along an axis (N,$\boldsymbol{u}$)
$\{T\}$	spatial translations
$\{G(\boldsymbol{u})\}$	planar gliding motions perpendicular to $\boldsymbol{u}$
$\{S(N)\}$	spherical motions about point *S*
$\{X(\boldsymbol{u})\}$	Schoenflies motions
$\boldsymbol{\{D\}}$	general rigid body motions or displacements

Based on [9] a planar joint has 5 equivalencies as presented below:

$$\{G(\boldsymbol{u})\} = \{R(A,\boldsymbol{u})\}\{R(B,\boldsymbol{u})\}\{R(C,\boldsymbol{u})\}; \tag{4}$$

$$\{G(\boldsymbol{u})\} = \{R(A,\boldsymbol{u})\}\{T(\boldsymbol{v})\}\{R(C,\boldsymbol{u})\} \quad \boldsymbol{v} \perp \boldsymbol{u}; \tag{5}$$

$$\{G(\boldsymbol{u})\} = \{T(\boldsymbol{v})\}\{R(B,\boldsymbol{u})\}\{R(C,\boldsymbol{u})\} \quad \boldsymbol{v} \perp \boldsymbol{u}; \tag{6}$$

$$\{G(\boldsymbol{u})\} = \{T(\boldsymbol{v})\}\{R(B,\boldsymbol{u})\}\{T(\boldsymbol{w})\} \quad \boldsymbol{v},\boldsymbol{w} \perp \boldsymbol{u}; \tag{7}$$

$$\{G(\boldsymbol{u})\} = \{T(\boldsymbol{v})\}\{T(\boldsymbol{w})\}\{R(C,\boldsymbol{u})\} \quad \boldsymbol{v},\boldsymbol{w} \perp \boldsymbol{u}. \tag{8}$$

Considering equations (4) and (8) respectively equality $N \equiv C$ the equation (3) becomes:

$$\{L_i\} = \{T(\boldsymbol{u})\}\{R(A,\boldsymbol{u})\}\{R(C,\boldsymbol{u})\}\{R(B,\boldsymbol{u})\}\{R(B,\boldsymbol{v})\}\{R(B,\boldsymbol{w})\}\ \forall A,B,C; \quad (9)$$

$$\{L_i\} = \{T(\boldsymbol{u})\}\{R(A,\boldsymbol{u})\}\{R(C,\boldsymbol{u})\}\{S(B)\} \quad \forall A,B,C. \quad (10)$$

The above defined $\{L_i\}$ displacement Lie group variants are presented in *Fig. 1*.

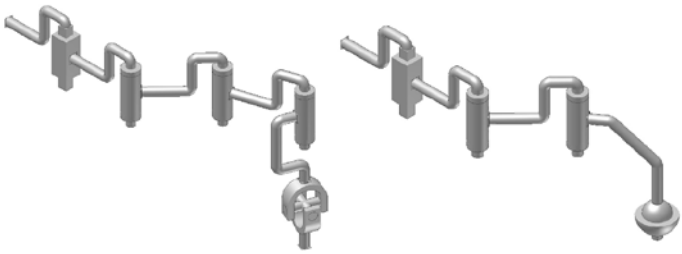


Figure 1: The $\{L_i\}$ displacement Lie group variants incorporating the X-motion generator.

The X-motion (or Shoenflies motion) generator can be easily observed, due to equation (9) and *Fig. 1a*. Considering primitive Schoenflies-motion generators [10] equivalences can be applied. Extending those generator family members with the universal joint as seen in *Fig. 1*, new generators for {D} displacement Lie group can be introduced. However, this enumeration is out of the topic of this paper. Because of the reduced link number and simplicity, in further investigation, the *Fig. 1b* variant is preferred. Using other geometrical constraints the architecture is presented in [11] also. The schematic design of such a limb for a 6 DOF manipulator is presented in *Fig. 2b*. The index *i* is introduced because the same type of limbs are used for moving the manipulator platform.

3. Kinematics of 3-PRRS mechanism

The general setup for the parallel mechanism having three translations and three rotations for the end effector (denoted by P) is presented in *Fig. 2c*. For simplicity the mechanism is presented from top view. The geometrical parameters used in the mathematical modelling are enumerated in sketches b) and c) from *Fig. 2*. Further the real number values x_N, y_N and z_N are introduced as the coordinates of a point N in the Cartesian space $0x_0y_0z_0$.
Using the equivalency between sketches a) and b) from *Fig. 2* it can be stated:

$$\boldsymbol{C_iB_i} = \boldsymbol{C_iD_i} + \boldsymbol{D_iB_i} = C_iD_{ix} \cdot \boldsymbol{i} + C_iD_{iy} \cdot \boldsymbol{j} + D_iB_i \cdot \boldsymbol{k}\,, \tag{11}$$

$$\boldsymbol{C_iB_i} = C_iB_{ix} \cdot \boldsymbol{i} + C_iB_{iy} \cdot \boldsymbol{j} + D_iB_i \cdot \boldsymbol{k}\,, \tag{12}$$

where $\boldsymbol{i}$, $\boldsymbol{j}$ and $\boldsymbol{k}$ are the unit vectors of the x_0, y_0 and z_0 Cartesian axes. The setup of the mechanism (based on the projection of the manipulator on the $0x_0y_0$ plan – top view from *Fig. 2*) suggests a planar Delta manipulator.

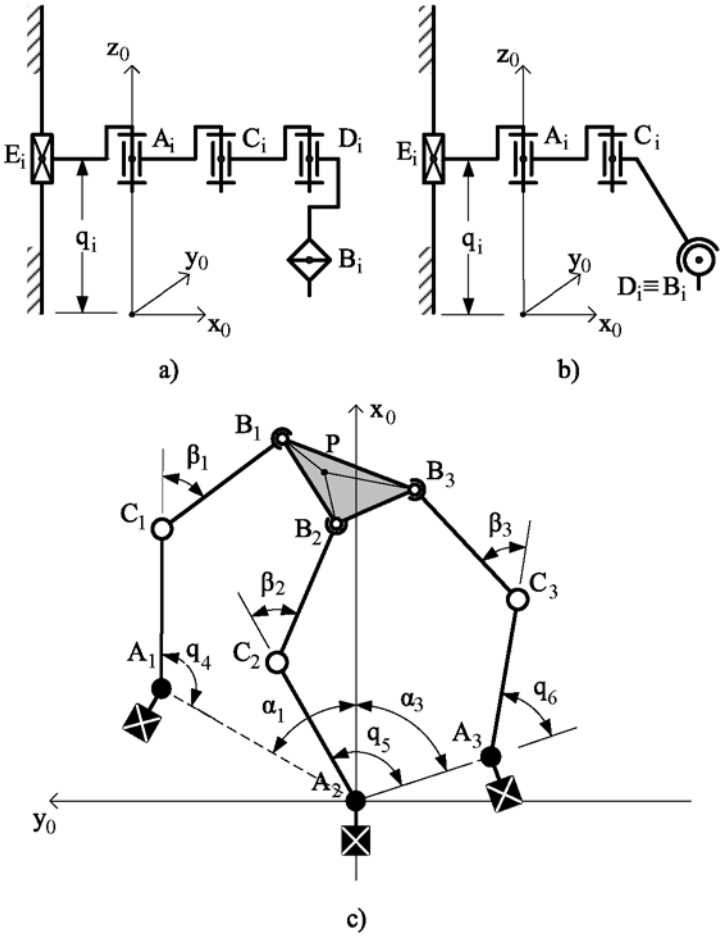


Figure 2: Schematic design of i[th] limb of the 6 DOF manipulator (a, b), and the top view of the proposed mechanism (c). The shaded couplings are the active joints (one prismatic and one rotation for each limb), and the white ones are passive bonds.

For this reason the mathematical modelling of the proposed mechanism is made easily and it is like a well known planar Delta manipulator modelling [7] with some completions. These completions must be made due to the fact, that it

is possible to rotate the platform around the x_0 and y_0 axes too, and so the projections of the B_iB_{i+1} platform length are variable. Through these paragraphs the inverse and direct kinematics of the proposed mechanism is defined, and issues about singular configurations are presented as well.

At the beginning the closure equation is considered for the three limbs:

$$\boldsymbol{OA_i} + \boldsymbol{A_iC_i} + \boldsymbol{C_iB_i} = \boldsymbol{OP} + \boldsymbol{PB_i} \text{ where } i = 1,2,3\,. \tag{13}$$

In case of inverse kinematic modelling the right side of equation (13) is given through the coordinates of the characteristic point (denoted by *P*) and through the three rotation angles around the axes of the fixed $0x_0y_0z_0$ system: $\boldsymbol{X} = [x_P \;\; y_P \;\; z_P \;\; \alpha \;\; \beta \;\; \gamma]^{\mathrm{T}}$. The task is to determine the robot parameters $\boldsymbol{q} = [q_1 \;\; q_2 \;\; q_3 \;\; q_4 \;\; q_5 \;\; q_6]^{\mathrm{T}}$ from the left side of the equation. Assuming that vector $\boldsymbol{a}$ has the components $\boldsymbol{a_{xy}}$ parallel to the $0x_0y_0$ plan and $\boldsymbol{a_z}$ parallel to the z_0 axis, equation (13) becomes:

$$\begin{cases} \boldsymbol{OA_{ixy}} + \boldsymbol{A_iC_{ixy}} + \boldsymbol{C_iB_{ixy}} = \boldsymbol{OP_{xy}} + \boldsymbol{PB_{ixy}} \\ \boldsymbol{OA_{iz}} + \boldsymbol{A_iC_{iz}} + \boldsymbol{C_iB_{iz}} = \boldsymbol{OP_z} + \boldsymbol{PB_{iz}} \end{cases} \text{ where } i = 1,2,3\,. \tag{14}$$

In order to determine the q_i translational parameters (i=1,2,3), introduced in *Fig. 2b*, the second equation from (14) is considered:

$$q_i \cdot \boldsymbol{k} + \boldsymbol{C_iB_{iz}} = z_P \cdot \boldsymbol{k} + \boldsymbol{PB_{iz}}\text{, respectively in scalar form} \tag{15}$$

$$q_i = z_P + PB_{iz} + C_iB_{iz} \tag{16}$$

Hence C_iB_{iz} is a constant geometrical parameter of the manipulator, the first two terms from the right side of equation (16) contain the general parameters because $PB_{iz} = PB_{iz}(\alpha,\beta,\gamma)$.

To obtain the q_{i+3} rotation joints parameters (i=1,2,3) the first equation from (14) is recalled and presented in scalar form:

$$\begin{cases} OA_i\cos\alpha_i + A_iC_i\cos(q_{i+3}+\alpha_i-\pi) + C_iB_i\cos(q_{i+3}+\alpha_i-\pi-\beta_i) = x_P + PB_{ix} \\ OA_i\sin\alpha_i + A_iC_i\sin(q_{i+3}+\alpha_i-\pi) + C_iB_i\sin(q_{i+3}+\alpha_i-\pi-\beta_i) = y_P + PB_{iy} \end{cases}, \tag{17}$$

where $PB_{ix} = PB_{ix}(\alpha,\beta,\gamma)$ and $PB_{iy} = PB_{iy}(\alpha,\beta,\gamma)$ respectively i=1,2,3. To eliminate the β_i parameter belonging to a passive joint, the equations are rearranged, and summing the square of the two equations in (17) yields:

$$e_{1i}\cdot\sin(q_{i+3}+\alpha_i-\pi) + e_{2i}\cdot\cos(q_{i+3}+\alpha_i-\pi) + e_{3i} = 0\,, \tag{18}$$

where

$$\begin{cases} e_{1i} = -2 \cdot A_i C_i \cdot \left(y_P + PB_{iy} - OA_i \sin\alpha_i \right); \\ e_{2i} = -2 \cdot A_i C_i \cdot \left(x_P + PB_{ix} - OA_i \cos\alpha_i \right); \\ e_{3i} = \left(x_P + PB_{ix} - OA_i \cos\alpha_i \right)^2 + \left(y_P + PB_{iy} - OA_i \sin\alpha_i \right)^2 + {A_i C_i}^2 - {C_i B_i}^2. \end{cases} \quad (19)$$

Solving equation (18) by using the substitutions:

$$\begin{cases} \sin\left(q_{i+3} + \alpha_i - \pi\right) = \dfrac{2t_i}{1+t_i^2} \\ \cos\left(q_{i+3} + \alpha_i - \pi\right) = \dfrac{1-t_i^2}{1+t_i^2} \end{cases} \text{ where } t_i = \tan\frac{q_{i+3} + \alpha_i - \pi}{2}, \quad (20)$$

the q_{i+3} parameters (i=1,2,3) are given by:

$$q_{i+3} = \pi - \alpha_i + 2\tan^{-1}\frac{-e_{1i} \pm \sqrt{e_{1i}^2 + e_{2i}^2 - e_{3i}^2}}{e_{1i} - e_{2i}}. \quad (21)$$

Equations (16) and (21) define the robot parameters in case of inverse kinematics. To obtain the general coordinates of the mechanism it is necessary to calculate the position of the $B_i(x_{Bi},y_{Bi},z_{Bi})$ joints (i=1,2,3) in Cartesian space and knowing the geometrical dimensions of the mobile platform the $\boldsymbol{X} = [x_P \;\; y_P \;\; z_P \;\; \alpha \;\; \beta \;\; \gamma]^{\mathrm{T}}$ vector is obvious. The x_{Bi}, y_{Bi}, and z_{Bi} values are defined through the following nine equations:

$$\begin{cases} \left(x_{Ci} - x_{Bi}\right)^2 + \left(y_{Ci} - y_{Bi}\right)^2 = C_i B_i^2 \\ \left(x_{Bi} - x_{Bj}\right)^2 + \left(y_{Bi} - y_{Bj}\right)^2 = d^2 - \left(z_{Bi} - z_{Bj}\right)^2 \\ z_{Bi} = z_{Ci} - D_i B_i \end{cases} \text{ for } \begin{matrix} i = 1,2,3 \\ j = \begin{cases} i+1, if\ i = 1,2 \\ 1, if\ i = 3 \end{cases} \end{matrix} \quad (22)$$

where $x_{Ci} = x_{Ci}(q_{i+3})$, $y_{Ci} = y_{Ci}(q_{i+3})$, $z_{Ci} = z_{Ci}(q_i)$ respectively $C_i B_i$ and $D_i B_i$ are constant geometrical dimensions using i = 1,2,3. It is important to mention, that in present case the forward kinematics deals with only 8 solutions (as by the planar Delta robot).

To complete the kinematic calculations the relation between the actuated joints and the platform velocities is needed and obtained through:

$$J_X \cdot \dot{\boldsymbol{X}} = J_q \cdot \dot{\boldsymbol{q}}, \quad (23)$$

where the matrices:

$$J_X = \begin{bmatrix} b_{1x} & b_{1y} & b_{1z} & e_{1y}b_{1z} - e_{1z}b_{1y} & e_{1z}b_{1x} - e_{1x}b_{1z} & e_{1x}b_{1y} - e_{1y}b_{1x} \\ b_{2x} & b_{2y} & b_{2z} & e_{2y}b_{2z} - e_{2z}b_{2y} & e_{2z}b_{2x} - e_{2x}b_{2z} & e_{2x}b_{2y} - e_{2y}b_{2x} \\ b_{3x} & b_{3y} & b_{3z} & e_{3y}b_{3z} - e_{3z}b_{3y} & e_{3z}b_{3x} - e_{3x}b_{3z} & e_{3x}b_{3y} - e_{3y}b_{3x} \end{bmatrix}, \quad (24)$$

$$J_q = \begin{bmatrix} b_{1z} & 0 & 0 & a_{1x}b_{1y} - a_{1y}b_{1x} & 0 & 0 \\ 0 & b_{2z} & 0 & 0 & a_{2x}b_{2y} - a_{2y}b_{2x} & 0 \\ 0 & 0 & b_{3z} & 0 & 0 & a_{3x}b_{3y} - a_{3y}b_{3x} \end{bmatrix}, \quad (25)$$

can be written using the notations $\boldsymbol{a} = \boldsymbol{A_i C_i}$, $\boldsymbol{b} = \boldsymbol{C_i B_i}$ and $\boldsymbol{e} = \boldsymbol{P B_i}$. Equation (23) can be considered for calculation of direct and inverse kinematics.

4. Parallel drive actuation of a manipulator limb

To assure the parallel mechanism concept for the 3-PRRS manipulator the parallel drive of the three limbs must be realized. Therefore a toothed belt drive H-shaped system can be applied as it can be seen in *Fig. 3*. At the bottom of the mechanism the actuated pulleys (gray color fill) induce the motion in the mechanism by the q_i^M and q_{i+3}^M driving parameters. The values q_i and q_{i+3} are set as the output parameters.

Using the parallel drive system two rotation inputs are transformed into translation and rotation output. The relation between them is given by the following equation:

$$\begin{bmatrix} q_i \\ q_{i+3} \end{bmatrix} = \begin{bmatrix} \frac{r}{2} & -\frac{r}{2} \\ -\frac{r}{2R} & -\frac{r}{2R} \end{bmatrix} \cdot \begin{bmatrix} q_i^M \\ q_{i+3}^M \end{bmatrix}. \quad (26)$$

The inverse geometry calculus can be performed using the following equation:

$$\begin{bmatrix} q_i^M \\ q_{i+3}^M \end{bmatrix} = \begin{bmatrix} \frac{1}{r} & -\frac{R}{r} \\ -\frac{1}{r} & -\frac{R}{r} \end{bmatrix} \cdot \begin{bmatrix} q_i \\ q_{i+3} \end{bmatrix}. \quad (27)$$

Using the above formulation and considering equations (16) and (21) the inverse geometry is obtained in the following form:

$$\overline{q}^{M} = \begin{bmatrix} q_1^M \\ q_2^M \\ q_3^M \\ q_4^M \\ q_5^M \\ q_6^M \end{bmatrix} = \begin{bmatrix} \frac{1}{r} & 0 & 0 & -\frac{R}{r} & 0 & 0 \\ 0 & \frac{1}{r} & 0 & 0 & -\frac{R}{r} & 0 \\ 0 & 0 & \frac{1}{r} & 0 & 0 & -\frac{R}{r} \\ -\frac{1}{r} & 0 & 0 & -\frac{R}{r} & 0 & 0 \\ 0 & -\frac{1}{r} & 0 & 0 & -\frac{R}{r} & 0 \\ 0 & 0 & -\frac{1}{r} & 0 & 0 & -\frac{R}{r} \end{bmatrix} \cdot \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \\ q_6 \end{bmatrix} = A \cdot \overline{q} \,. \quad (28)$$

Due to the characteristic setup of the driving mechanism the equations for the kinematics are obtained in similar way:

$$\dot{\overline{q}}^{M} = A \cdot \dot{\overline{q}} \text{ and } \quad \dot{\overline{q}} = A^{-1} \cdot \dot{\overline{q}}^{M} \,. \quad (29)$$

In accordance with the formulated equations the dynamics of the manipulator can be calculated easily, and will be presented in a further paper.

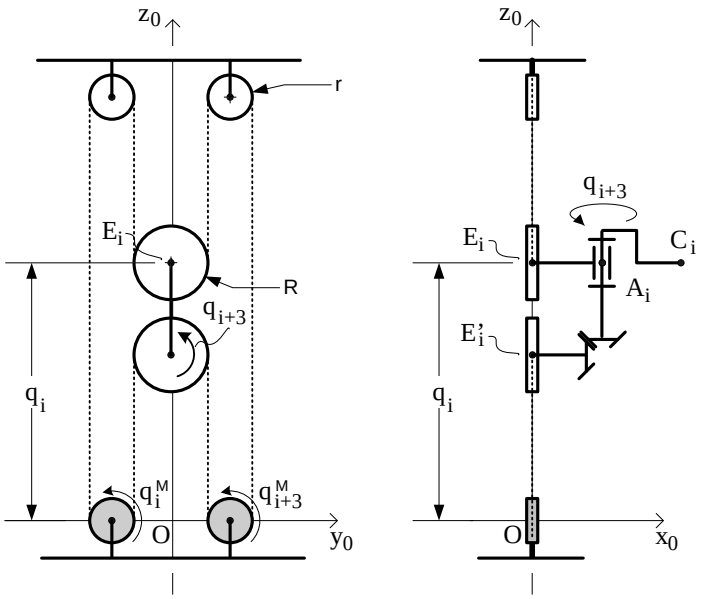


Figure 3: Schematic design of the belt mechanism for one, double drive link.

5. Singular configurations

The singularity analysis of this mechanism can be done based on the matrices from (24) and (25). Inverse kinematic singularities occure in case of $a_{ix}b_{iy} - a_{iy}b_{ix} = 0$ (i=1,2,3) which defines the workspace boundaries. An other possibility is $b_{iz} = 0$ (i=1,2,3) but it can be avoided through geometrical design, because it is a constant value. Direct kinematic singularities occure when at the same time it can be stated that $b_{ix} = 0$ or $b_{iy} = 0$ (i=1,2,3), which means that the C_iB_i links are parallel. The same type of singularities can be found for coexistence of $e_{ix}b_{iy} - e_{iy}b_{ix} = 0$ (i=1,2,3) in case of coliniar $\boldsymbol{C_iB_i}$ and $\boldsymbol{PB_i}$ vectors. Both direct kinematic singularity cases can be avoided by careful geometrical design. The implemented parallel drive mechanisms have no singular configurations, and this kind of calculations can be omitted.

6. Conclusions

This paper deals with a 6 degrees of freedom manipulator architecture using the group theory. The mobile platform is connected to the base through three PRRS limbs, each being double actuated on the first and second joint levels. The inverse geometrical calculations are performed through equations (16) and (21), hence the direct modelling is presented through the equation system (22). The relation between the robot and general velocities is stated by the equation (23). Some aspects about the singular configurations are introduced in the paper based on the equation mentioned before. As it can be seen in the figures presented in this paper the architecture is the extension of the well known planar Delta robot to a 6 DOF mechanism. The mathematical model of the spatial manipulator reflects this fact very well. The simple setup of the presented mechanism assures a good manufacturability and needs a relatively easy control algorithm considering some other 6 DOF manipulators.

References

[1] Stewart, D. A., “Platform with six degrees of freedom”, in *Proceedings on Institution of Mechanical Engineering*, 1965, vol. 180, pp. 371-386.

[2] Dasguta, B. and Mruthyunjaya, T. S., “The Stewart platform manipulator: a review” *Mechanism and Machine Theory*, vol. 35, pp. 15-40, 2000.

[3] Hunt, K. H., “Structural kinematics of in-parallel-actuated robot arms”, *ASME Journal of Mechanical Design*, vol. 105, pp. 705-712, 1983.

[4] Tsai, L. W., and Joshi, S., “Kinematics and optimization of a Spatial 3-UPU Parallel manipulator”, *ASME Journal of Mechanical Design*, vol. 122, pp. 439-446, 2000.

[5] Hervé, J. M., "Design of parallel manipulators via the displacement group", in *Proceedings of the 9th World Congress on Theory of Machine and Mechanisms*, Milano, 1985, pp. 2079-2082.

[6] Shen, H., Yang., T., and Ma, L., "Synthesis and structure analysis of kinematic structures of 6-DOF parallel robotic mechanisms", *Mechanism and Machine Theory*, vol. 40, pp. 1164-1180, 2005.

[7] Tsai, L. W., "Robot Analysis – The Mechanics of Serial and Parallel Manipulators", John Wiley & Sons, 1999.

[8] Hervé, J. M., "The Lie group of rigid body displacements, a fundamental tool for mechanism design", *Mechanism and Machine Theory*, vol. 34, pp. 719-730, 1999.

[9] Hervé, J. M., "The planar-spherical kinematic bond: Implementation in parallel mechanisms", http://www.parallemic.org/Reviews/Review013p.html.

[10] Lee, C. C. and Hervé, J. M., "Type synthesis of primitive Schoenflies-motion generators", *Mechanism and Machine Theory*, vol. 44, pp. 1980-1997, 2009.

[11] Olea, G., Plitea, N., and Takamusa, K., "Kinematical analysis and simulation of a new parallel mechanism for robotics application", in *ARK Piran, Piran*, 25-29 June 2000, pp. 403-410.

ACKNOWLEDGEMENT

The Editors would like to acknowledge the contributions of all who coordinated and performed the double-blind peer review of the manuscripts submitted to the journal. Besides the members of the Editorial Board, the following researchers made significant efforts to complete this process:

László BAKÓ
Sándor Tihamér BRASSAI
József DOMOKOS
Katalin GYÖRGY
Piroska HALLER
Tünde JÁNOSI-RANCZ
Lajos KENÉZ
Nimród KUTASI
László Ferenc MÁRTON
Márton MÁTÉ
István PAPP
Sándor PAPP
László SZABÓ
László SZILÁGYI
Tamás VAJDA

Acta Universitatis Sapientiae

The scientific journal of Sapientia University publishes original papers and surveys
in several areas of sciences written in English.
Information about each series can be found at
http://www.acta.sapientia.ro.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering

Sapientia University

Scientia Publishing House

ISSN 2065-5916
http://www.acta.sapientia.ro

Information for authors

Acta Universitatis Sapientiae, Electrical and Mechanical Engineering publishes only original papers and surveys in various fields of Electrical and Mechanical Engineering. All papers are peer-reviewed.

Papers published in current and previous volumes can be found in Portable Document Format (PDF) form at the address: `http://www.acta.sapientia.ro`.

The submitted papers must not be considered to be published by other journals. The corresponding author is responsible to obtain the permission for publication of co-authors and of the authorities of institutes, if needed. The Editorial Board is disclaiming any responsibility.

The paper must be submitted both in MSWord document and PDF format. The submitted PDF document is used as reference. The camera-ready journal is prepared in PDF format by the editors. In order to reduce subsequent changes of aspect to minimum, an accurate formatting is required. The paper should be prepared on A4 paper (210 $\times$ 297 mm) and it must contain an abstract of 200-250 words.

The language of the journal is English. The paper must be prepared in single-column format, not exceeding 12 pages including figures, tables and references.

The template file from http://www.acta.sapientia.ro/acta-emeng/emeng-main.htm may be used for details.

Submission must be made only by e-mail (`acta-emeng@acta.sapientia.ro`).

One issue is offered to each author free of charge. No reprints are available.

Contact address and subscription:
Acta Universitatis Sapientiae, Electrical and Mechanical Engineering
RO 400112 Cluj-Napoca
Str. Matei Corvin nr. 4.
E-mail: acta-emeng@acta.sapientia.ro

Publication supported by

Printed by Gloria Printing House
Director: Péter Nagy