Acta
Technica
Jaurinensis

Vol. 7., No. 4., pp. 328-343, 2014
DOI: 10.14513/actatechjaur.v7.n4.313
Available online at acta.sze.hu

# Design Luenberger Observer for an Electromechanical Actuator

**Zs. Horváth, Gy. Molnárka**

**Department of Mechatronics and Machine Design**
**Széchenyi István University, Egyetem tér 1, H-9026, Győr, Hungary**
**zsolt2.horvath@audi.hu**

Abstract: In this work we have designed a Luenberger Observer for an Electromechanical Actuator (EMA). We have considered a simple linear model of the Actuator plant with neglecting nonlinearities. The model of the EMA, the Observer and the simulations have been computed in MATLAB®. In our simulation we have examined the dynamics of the planned Observer. With different pole placement proportions and state estimation error conditions, three states have been estimated: Motor current ($X_1$), angular velocity of the throttle plate ($X_2$), and angular position of the throttle plate ($X_3$). Outline of our examination should be to take the conclusions from the influence of the pole placement and error condition on the dynamics of the Observer. Thus determining of an optimal pole values for the proposed Luenberger Observer. Outline of this paper could be used for further research topics (fault detection in Electromechanical Actuator).

*Keywords: observer, observability, EMA plant-model, state space equation, pole-placement*

## 1. Introduction

With increasing number of processing integrated electromechanical systems on current automotive vehicles, such as actuators, sensors and microcomputers, the field of supervision (monitoring), diagnosis and control plays an inportant role. Electromechanical Actuators often have to perform in extreme plant conditions, measurement of needed signals is often not possible or it is too expensive, and that is why their state control and diagnostic is important.

In applications feedback control or diagnosis systems the entire state vector of the system must be controlled. If the state vector can not be measured, what is a tipical case in most complex systems, one suitable approximation to the state vector is needed. That can be substituted into a control law. A system which produces such an approximation, is called an Observer or Luenberger Observer. It was Luenberger [1] who has first

developed an Observer for linear control systems, and after that it was introduced for nonlinear systems too, by Thau [2].

This paper proposes a design for a Luenberger Observer for an electromechanical throttle valve. Many different observer concepts are nowadays succesfully used for modelling and control of Electromechanical Actuators. Nonlinear variable structure system (VSS) observer for throttle systems are presented in e.g. [8]. Paper [3] uses a sliding-mode observer for a robust position control of these Actuators. The most common approaches are using Extended Kalman Filter (EKF) or Unscented Kalman Filter (UKF) [7], which are preferable solutions when the process nonlinearities are strong and noise is associated with the real system. Another study [5] proposes using of the Kalman Filter for the observation of the DC-motor. A further study is applying Luenberger Observer for Sensor Monitoring in Active Front Steering Systems can be found in [11].

The throttle valve is a type of an Electromechanical Actuator, which is advanced in several applications of combustion engine control (intake manifold, exthaust gas recirculation, variable turbine geometry, ect.). The simple EMA plant model is a system of 3rd order.

The presented paper consists of : in Section I the EMA plant model is introduced, then in Section II the fundamentals of Luenberger Observer are revised, in Section III the design of Luenberger Observer is described. Finally in Section IV the main obtained results are presented in this paper.

## 2. Model of an Electromechanical Actuator

The simple model of the EMA *Figure 1.* consists of following parts: DC-motor, gearbox, return spring, throttle plate and sensor for valve position. The DC-motor is supplied with a bipoloar chopper, motor shaft rotation is transmitted through a gearbox to the throttle plate. The position of the throttle plate is given to the ECU (Engine Control Unit). A reset spring places the de-energized throttle plate in its full open position. The position control of the throttle valve and the diagnosis of the actuator is done by the engine ECU (ECU will be not examined in this work).

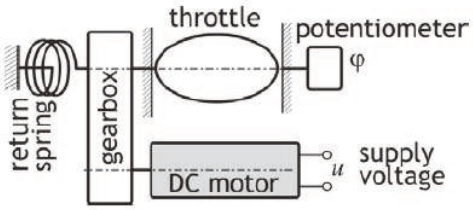



*Figure 1. The scheme of the Electromechanical Actuator (EMA) [6]*

### 2.1 Model Equations of the EMA

The relation between input voltage $u_a$ and current $i$ in the armature circuit can be described as

$$u_a = Ri + L\frac{di}{dt} + K_e n\omega \tag{1}$$

where $L$ is the inductance, $R$ is the resistance in the armature circuit, $K_e$ is the inductive voltage constant, $n$ is the gear ratio and $\omega$ *is* the angular velocity of the throttle plate .

The motion Equations of the throttle plate (related to the plate axel) can be desribed as

$$J\frac{d\omega}{dt} = K_t i - M_o - c\varphi - M_c sign(\omega) - K_V \omega \tag{2}$$

where φ is the angular position of the throttle plate, J is the rotary inertia at the throttle shaft, which is composed of inertia of the DC-motor. Mo denotes a torque due to the prestressed springs, c is a linear spring constant. Mc is a torque of the coulomb friction of the system. Kv is a positive constant for viscousus friction. Finally Kt denotes the motor torque constant.

*Table 1. EMA parameters*

| ***Parameter*** | ***Values*** | ***Units*** |
|---|---|---|
| R | 4.3 | Ohm |
| L | 0.002 | H |
| $K_e$ | 0.017 | Vs/rad |
| $K_t$ | 0.017 | Nm/A |
| J | 0.0016 | $kgm^2$ |
| n | 40 | - |
| $M_c$ | 0.075 | Nm |
| $M_0$ | 0.25 | Nm |
| c | 0.059 | Nm/rad |
| $K_v$ | 0.02 | Nms/rad |
| $K_s$ | 2.73 | V/rad |

Considering the Equations (1) and (2), written down in the standard state-space description and neglecting the nonlinearities of the Coulomb friction (we are controlling the valves only in one direction now), the input voltage (ua) is 12V, the measured variable is the angular position of the plate φ ϵ (0°…90°).

For a linear time-invariant process the state space Equations can be given as

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= C^T x + D^T u \end{aligned} \tag{3}$$

with $u \in \mathbf{R}^r$ , $x \in \mathbf{R}^n$ , $y \in \mathbf{R}^m$

The elements of the Equations are the state vector x(t), input u(t), output y(t), state matrix A, input vector B, output vector CT and the direct feedthrough D.

The first Equation here is the state Equation and the second is the output.
With defining the state vector and input vector, it can be written as

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} i \\ \omega \\ \varphi \end{bmatrix} \quad ; \quad u = \begin{bmatrix} u_a \\ m_t \\ 0 \end{bmatrix} \tag{4}$$

Where $x_1$ is the first component of the state vector $x$, and it represents the motor current, $x_2$ is the angular velocity, and $x_3$ is the angular position of the throttle plate.

In the Equation of the input vector are the input voltage and the sum torque of the Coulomb friction and spring pretension (they are constant).

For the state space description of our model, Equations (1) and (2) are brought into the form

$$\begin{aligned} \frac{di}{dt} &= -\frac{R}{L} i - n\frac{Ke}{L}\omega + \frac{1}{L} u_a \\ \frac{d\omega}{dt} &= n\frac{K_t}{J} i - \frac{K_v}{J}\omega - \frac{M_c + M_0}{J} - \frac{c}{J}\varphi \end{aligned} \tag{5}$$

Eq. (5) written into a form vector-matrix

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -\frac{R}{L} & -n\frac{K_e}{L} & 0 \\ n\frac{K_t}{J} & -\frac{K_v}{J} & -\frac{c}{J} \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} \frac{1}{L} & 0 & 0 \\ 0 & -\frac{1}{J} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_a \\ m_t \\ 0 \end{bmatrix} \tag{6}$$

$$y = \begin{bmatrix} 0 & 0 & K_s \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_a \\ m_t \\ 0 \end{bmatrix}$$

Through substituting the parameter values of the EMA into the Eq. (6) we receive the following state space model of the plant

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}-2150 & -340 & 0\\ 420 & -12.5 & -36.7\\ 0 & 1 & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}+\begin{bmatrix}500 & 0 & 0\\ 0 & -625 & 0\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}u_a\\ 0.32\\ 0\end{bmatrix} \tag{7}$$

$$y=\begin{bmatrix}0 & 0 & 2.73\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}+\begin{bmatrix}0 & 0 & 0\end{bmatrix}\begin{bmatrix}u_a\\ 0.32\\ 0\end{bmatrix}$$

as well as the transfer functions for angular position

$$W_1(s)=\frac{0.68}{3.2\cdot10^6 s^3+6.92\cdot10^{-3}s^2+0.543s+0.252}$$

$$W_2(s)=-\frac{2\cdot10^{-3}s+4.3}{3.2\cdot10^6 s^3+6.92\cdot10^{-3}s^2+0.543s+0.252} \tag{8}$$

$$\mathrm{W(s)}=W_1(\mathrm{s})+W_2(s)$$

The angular position of the plate can be determined as

$$\varphi(\mathrm{s})=W_1(\mathrm{s})\,\mathrm{u}_a(\mathrm{s})+W_2(s)\,\mathrm{m}_t$$
where
$$m_t=M_0+M_c \tag{9}$$

$m_t$ is a load torque during the plate opening.
The parameters of the plant are shown in the Table 1.

Fig. 2 shows the step response of second and third state, $\omega$, $\varphi$ respectively by input voltage of 12V.

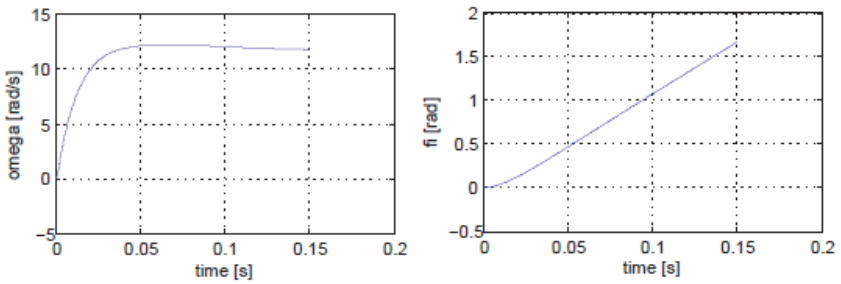



*Figure 2. Step responses of the EMA*
*left: Angular velocity of throttle plate , right: Angular posion of throttle plate*

With getting the step responses, we are able to see the linear process dynamics of the EMA. The left side of *Figure 2* shows the high dynamics of the EMA with a settling

time of ~ 0.05sec. We must take this in consideration by designing the Observer when determining the poles.

## 3. Luenberger Observer

Luenberger was the first who suggested the state Observer for estimating the unknown states of a deterministic linear system in 1971 [1]. The Equation for this contains a correction term for current state estimates, which corrects with an amount that is proportional to the error that we predicted: reducing the current output sum with the actually measured sum. With this output sum reduction we can ensure the stability and convergence of the Observer even in case of unstable state of the observed system.

### 3.1 Observability of Time-Invariant Systems

Observability is a necessary condition for investigating an Observer.

Observability [2] is the issue of whether the state of dynamic system with a known model is uniquely determinable from its inputs and outputs. It is essentially a property of the given system model. A given linear dynamic system model with a given linear input/output model is considered observable if and only if its state is uniquely determinable from the models definition, its inputs and outputs.

The observability of time invariant systems can be characterised with the rank of the observability matrices.

As already mentioned in section 2.1, the elements of the Equation (10) are:

- ***A :*** state matrix
- $C^T$ : output vector

$$M_0 = \begin{bmatrix} C^T \\ C^T A \\ \vdots \\ C^T A^{n-1} \end{bmatrix} \tag{10}$$

for continuous-time systems.

The systems are observable if the dimension of the system state vector is rank n.
The observability matrix for the EMA, using Equation (10) is

$$M_0 = \begin{bmatrix} 0 & 0 & 2.7 \\ 0 & 2.7 & 0 \\ 1146 & -34.1 & -100.3 \end{bmatrix} \tag{11}$$

Here, $M$ has a rank (rank=3) equal to the dimension of $x(t)$ , therefore our system is observable.

### 3.2 Basic Theory of Luenberger Observer

Considering the state space system again

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \quad , \quad x(0) = 0 \\ y(t) &= Cx(t) \end{aligned} \tag{12}$$

with enlarging Eq. (12) through additional input sequence $u_B$

$$\begin{aligned} \frac{d\hat{x}(t)}{dt} &= A\hat{x}(t) + Bu(t) + u_B(\mathrm{t}) \, , \quad \hat{x}(0) = \hat{x}_0 \\ \hat{y}(t) &= C\hat{x}(t) \end{aligned} \tag{13}$$

where

$$u_B(t) = L\left(y(t) - \hat{y}(t)\right) \tag{14}$$

In designing the Observer, the matrices *A,B,C* are fixed, but the n x m matrix *L* is arbitrary. *L* is called as Observer gain. An identity Observer is determined by selection of *L*.
Through substituting Eq. (14) into Eq. (13) and taking in consideration that

$$\begin{aligned} &\hat{y}(t) = C\hat{x}(t) \\ &\text{and} \\ &y(t) = Cx(t) \end{aligned} \tag{15}$$

the Equation (13) is brought into the form

$$\frac{d\hat{x}(t)}{dt} = A\hat{x}(t) + Bu(t) + LC\left(x(t) - \hat{x}(t)\right) \tag{16}$$

With the obtained state matrix of the Observer *[A−LC],* the full Equation of the Luenberger Observer is

$$\frac{d\hat{x}(t)}{dt} = \left(A - LC\right)\hat{x}(t) + Bu(t) + Ly(t) \tag{17}$$

*u* and *y* are the inputs of the Observer, $\hat{x}(t)$ is an estimate of the plant state *x(t).*

The estimation error is defined as

$$e(t) = x(t) - \hat{x}(t) \tag{18}$$

which is not directly measurable in most cases, because *x(t)* is not available as an input to the obeserver.

With taking the first derivate of *e(t)* we can get

$$\begin{aligned}\dot{e}(t) &= \frac{d}{dt}\left(x(t)-\hat{x}(t)\right) = Ax(t)+Bu(t)-A\hat{x}(t)- \\ -Bu(t)&-LC\left(x(t)-\hat{x}(t)\right) = \left(A-LC\right)\left(x(t)-\hat{x}(t)\right)\end{aligned} \tag{19}$$

Equation (19) has proved that the dinamics of the observing process is determined by matrix *[A−LC]*. The eigenvalues of *[A−LC]* are satisfactory, the reason for this is that the Observer has to be able to imitate the plant state.

The dinamic of the error is given by

$$\begin{aligned}&\dot{e}(t) = \left(A-LC\right)e(t)\ , \quad e(0) = x_0 - \hat{x}_0 \\ &\lim_{t\to\infty}\left\|e(t)\right\| = 0\end{aligned} \tag{20}$$

To obtain the accurate state estimates with no regard of $e(0), e(t)$ as $t \to \infty$ exponentially, we must choose all the eigenvalues of the $[A - LC]$ matrix to lie in the left half s - plane. To achieve that, the state of the Observer converges to the state of the Observer system, the Observers eigenvalues must be negative, moreover, they should be a little more negative than the eigenvalues of the observed system. In that case the convergence is faster than other system effects. In theory, these eigenvalues can be moved to minus infinity arbitrarily, to achieve extreme fast convergence. But this leads the Observer to act like a differentiator, become very sensitive to noise and arise other difficulties. The problem of selecting good eigenvalues is still not totally resolved, but a functional practice is that we should place them so, that the Observer can be a little faster than the rest of the closed-loop system [4].

*Figure 3* helps in the interpretation of the Luenberger Observer.
The elements of this Figure are:

- *x :* state vector
- *u :* input
- *y :* output
- ***A*** *:* state matrix
- ***B*** *:* input vector
- C *:* output vector
- *L*: Observer gain
- ***e*** *:* estimation error

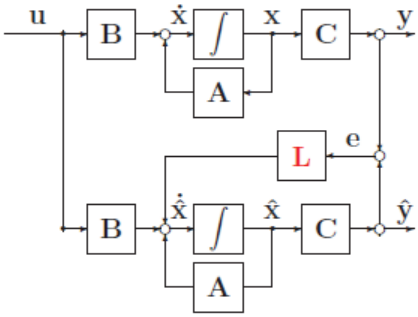



*Figure 3. Scheme of the Luenberger Observer*

The residual is the difference between the measured output and the actual observation. It can be written as

$$r(t) = y(t) - \hat{y}(t) \tag{21}$$

In ideal case is, when *r(t) = 0* during the entire operational interval if the Observer is a perfect model of the plant. However in reality there are uncertainities in the model of the plant represented by *ΔA, ΔB* and *ΔC*. These are neglected in this work. In case when an instrument fault occurs the estimated vector will be $\hat{x}(t)$ upset. That makes the value of the residual nonzero. As a result, the residual can be used also for fault detection of instrument faults.

### 3.3 Design of the Luenberger Observer

For designing an Observer we must know the dynamics of the observed system, defined by the eigenvalues of the system.

As known, the eigenvalues of the plant can be calculated with a characteristical polynom as follows

$$\left|A - \lambda \cdot E\right| = -\lambda^3 - \lambda^2 \cdot 2162.5 - \lambda \cdot 169711.25 - 77937.5 \tag{22}$$

from this, the poles of the plant are

$$\lambda_1 = -2081\ ;\ \lambda_2 = -81.1\ ;\ \lambda_3 = -50 \tag{23}$$

Knowing the poles of the observed system, the next step is defining the Observers poles [A−LC]. Generally the formula is that we define 2-6 times bigger poles for the Observer than the dominant systems poles [4]. The barrier for the bigger poleshifting is the increasing noise in the Observer. My goal in this work is setting up 5 times bigger poleshifting, since I calculate with small noise rate.

The reason for choosing factor 5 is that the eigenvalues of the [A−LC] matrix needs to be prompt compared at beginning of the operating interval too.

As we can see, the EMA plant has a small dominant time constant and therefore small settling–time.

From this the proposed eigenvalues of the Observer

$$\mathrm{P}_0 = \det\left(sI - A + Lc^T\right) = F\left(s\right) = s^n + f_1 s^{n-1} + \ldots + f_{n-1}s + f_n \tag{24}$$

$$P_0 = \left[-2331; -331.1; -250.5\right]^T$$

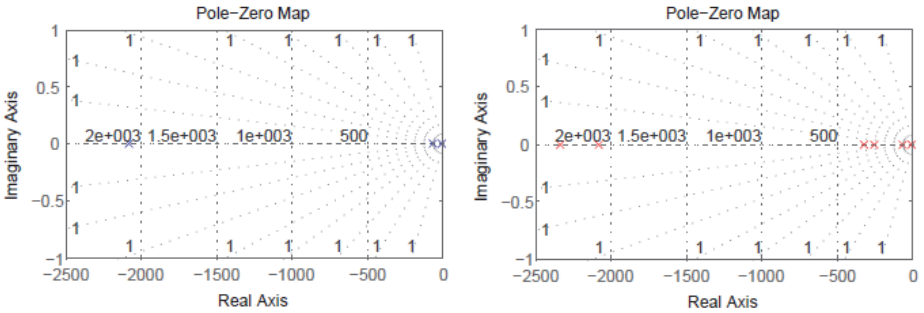



*Figure 4. The pole placement for the Observer showed on the Pole-Zero Map left: poles of the EMA plant , right: desired poles of the Observer*

The next step of the design is determining the feedback gain matrix *L*.

A number of algorithms can be found in the Literature [4], [10] to do that, some of which are built in a popular control system design software. In MATLAB® commands acker or place can be used as feedback to calculate the matrix L.

If the plant model is given in transfer function form, the poles can be easily taken with using the state space canonical form. In the matrix of the canonical form the poles of the plant are placed in the first column ( Eq. (25) ).

$$A_0 = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 \\ -a_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_{n-1} & 0 & 0 & \cdots & 1 \\ -a_n & 0 & 0 & \cdots & 0 \end{bmatrix};$$

$$C_0^T = \left[1, \quad 0, \quad \ldots, \quad 0\right]; b_0 = \left[b_1, \quad b_2, \quad \ldots, \quad b_n\right]^T \tag{25}$$

To determine the Observer feedback matrix in canonical form, the matrix Equation can be described as

$$A_0 - L_0 C^T = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 \\ -a_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_{n-1} & 0 & 0 & \cdots & 1 \\ -a_n & 0 & 0 & \cdots & 0 \end{bmatrix} -$$

$$-L_0 \begin{bmatrix} 1, & 0, & \ldots, & 0 \end{bmatrix} = \begin{bmatrix} -f_1 & 1 & 0 & \cdots & 0 \\ -f_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -f_{n-1} & 0 & 0 & \cdots & 1 \\ -f_n & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{26}$$

$$L = L_0 = \begin{bmatrix} f_1 - a_1, f_2 - a_2, \ldots, f_n - a_n \end{bmatrix}^T$$

From this the Observer feedback matrix is

$$L = \begin{bmatrix} 719670; -129350; 270 \end{bmatrix}^T \tag{27}$$

The Observer gain matrix can be written as follows:

$$\begin{aligned} & A - LC = \\ & \begin{bmatrix} -2150 & -340 & 0 \\ 420 & -12.5 & -36.7 \\ 0 & 1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 1964700 \\ 0 & 0 & -353100 \\ 0 & 0 & 800 \end{bmatrix} \\ & = \begin{bmatrix} -2200 & -300 & -1964700 \\ 400 & 0 & 353100 \\ 0 & 0 & -800 \end{bmatrix} \end{aligned} \tag{28}$$

Finally the state space Equation of the Luenberger Observer resulted from Eq. (28) is determined to be:

$$\begin{aligned} & \frac{d\hat{x}}{dt} = \begin{bmatrix} -2200 & -300 & -1964700 \\ 400 & 0 & 353100 \\ 0 & 0 & -800 \end{bmatrix} x + \\ & + \begin{bmatrix} 500 & 0 & 0 \\ 0 & -625 & 0 \\ 0 & 0 & 0 \end{bmatrix} u + \begin{bmatrix} 719670 \\ -129350 \\ 270 \end{bmatrix} y \end{aligned} \tag{29}$$

## 4. Experimental Results

In the following I will examine the state variables of the EMA model and the function of the Observer group planned for it, with different pole placement proportions and state estimation error conditions. Outline of our examination should be to take the conclusion to the influence of the pole placement and error condition on the dynamics of the observer. The algorithm for the Observer has been computed in MATLAB®. The throttle valve is excited with a step input signal, the input voltage is 12V. Three states have been estimated in the simulation: Motor current (X1), angular velocity of the throttle plate (X2), and angular position of the throttle plate (X3). The pole placement factor is calculated as follows: $k = P_{dom} + P_0$ .

In the first case the pole placement rate (k= −50) and value of the estimation error condition were small ($e_0 = [0.1,0.1,0.4]$ ), showed on the Figure 5.

The Observer follows the state variables of the EMA plant with slow dynamics. This doesn't fulfill the dynamic requirements.

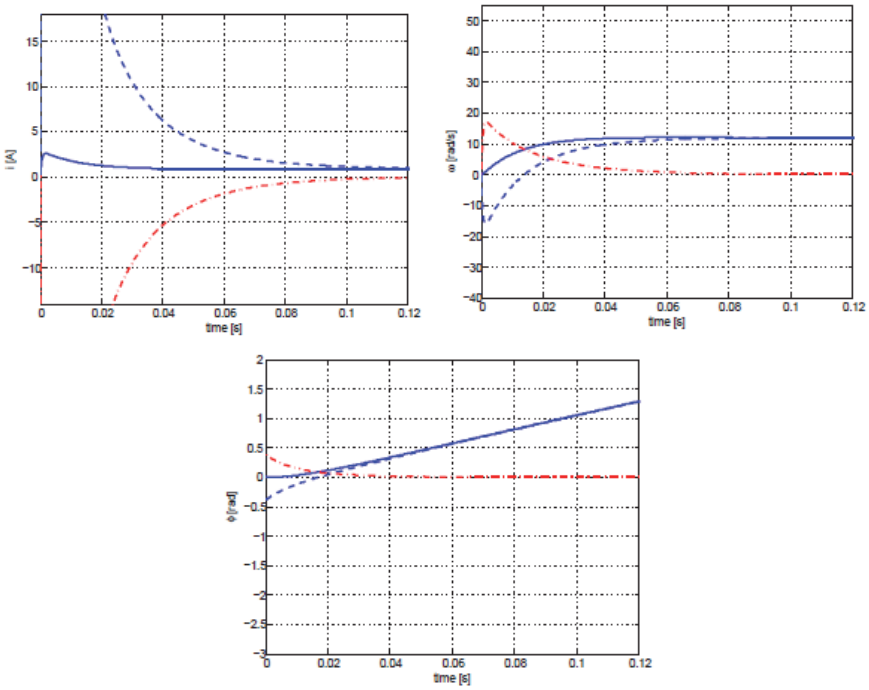



*Figure 5. States of the EMA model with estimated pole placement rate of k=−50 and setting small estimation error condition ( $e_0 = [0.1,0.1,0.4]$ ) blue : state of the plant, dashed blue : estimated state, d.-dashed red : estimation error top left: motor current, top right: angular velocity, bottom : angular position of the plate*

In the second case the pole placement rate *(k= −50)* and the value of the estimation error condition were big *($e_0 = [0.1,0.1,2.5]$ )*, as showed on the *Figure 6.*
The Observer follows the state variables of the EMA plant with slower dynamics than in the first case. That is because of the big estimation error condition. These pole settings doesn't fulfill the dynamic requirements.

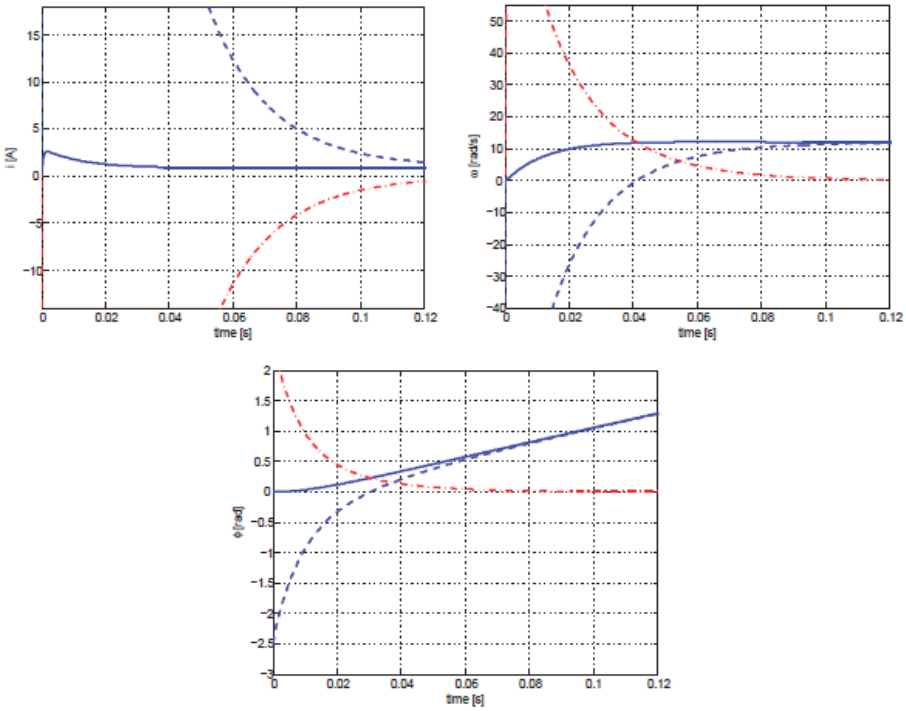



*Figure 6. States of the EMA model with estimated pole placement rate of k= −50 and setting big estimation error condition ( $e_0 = [0.1,0.1,2.5]$)*
*blue : state of the plant, dashed blue : estimated state, d.-dashed red : estimation error*
*top left: motor current, top right : angular velocity,*
*bottom : angular position of the plate*

After this, we have increased the pole placement rate (k= −100). The value of the estimation error condition was small ( $e_0 = [0.1,0.1,0.4]$ ), as showed on the Figure 6. We have experienced, that the Observer follows the state variables of the EMA plant with satisfactory dynamics. This would fulfill the dynamic requirements.

But when the pole placement rate (k= −100) stayed the same, and the value of the estimation error condition has been increased to big ( $e_0 = [0.1,0.1,2.5]$, the Observer follows the state variables of the EMA plant with slower dynamics than in the first case. The estimated state variables of motor current and angular velocity of the throttle plate needed more time to settle than in the case with small estimation error condition. These pole settings doesn't fulfill the dynamic requirements either.

As a result from these, it can be stated, that as far as the estimation error condition is increasing, the Observers settling time will be slower and slower.

In the case when the pole placement rate (k= −200) and the value of the estimation error condition were small ($e_0 = [0.1,0.1,0.4]$ ), we have experienced that the Observer follows the state variables of the EMA plant with adequate dynamics. The settling time was short for every estimated state. This fulfills the dynamic requirements.

In the last case the pole placement rate was the same (k= −200) , but the value of the estimation error condition was big ($e_0 = [0.1,0.1,2.5]$ ), as shown on the Figure 7. The Observer follows the state variables of the EMA plant with a little slower dynamics than in the first case. The estimated state variables of motor current and angular velocity of the throttle plate needed more time for settling on than in the case with small estimation error condition. In spite of this, this pole settling fulfills the dynamic requirements.

As a result from these, it can be stated, that the placement rate k= −200 had to fulfill the dynamic requirements observing the throttle valve in every condition.

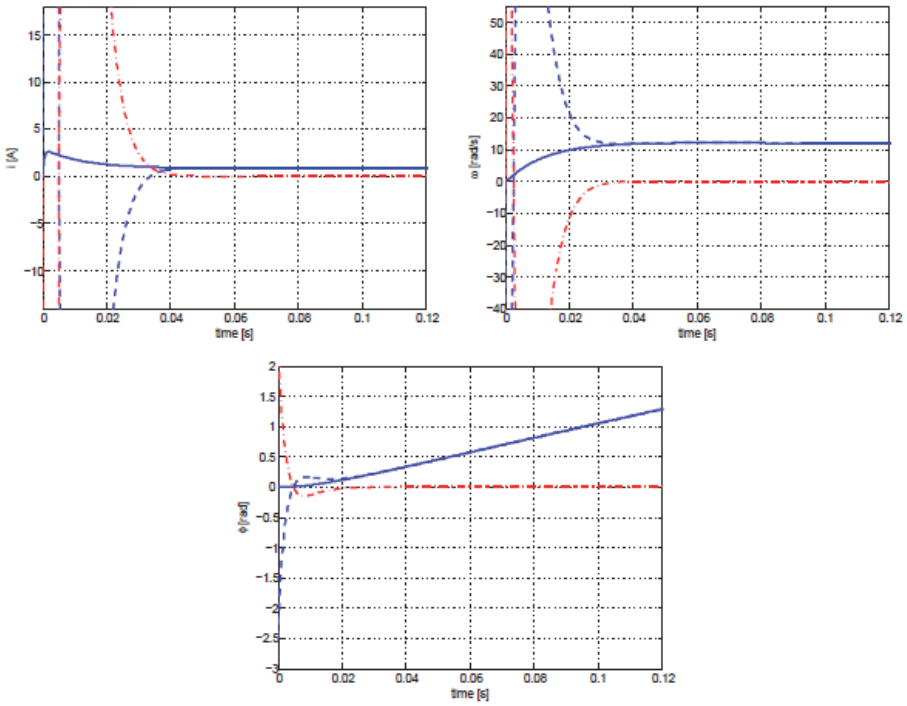



*Figure 7. States of the EMA model with estimated pole placement rate of k=−200 and setting big estimation error condition ( $e_0 = [0.1,0.1,2.5]$)*
*blue : state of the plant, dashed blue : estimated state, d.-dashed red : estimation error*
*top left: motor current, top right : angular velocity,*
*bottom : angular position of the plate*

## 5. Conclusion

This paper demonstrates a design of the Luenberger Observer for an Electromechanical Actuator. For this design, we have considered a simple linear model of the EMA plant with neglecting nonlinearities. The models of the EMA and the Observer were computed in MATLAB®.

In the simulation we have examined the dynamics of the Observer to be planned for it.

With different pole placement proportions and state estimation error conditions three states have been estimated: Motor current (X1), angular velocity of the throttle plate (X2), and angular position of the throttle plate (X3).

In the last case of the simulation with pole placement rate of (k= −200), despite the value of the estimation error condition was big ($e_0 = [0.1,0.1,2.5]$), the Observer has fulfilled the dynamic requirements. It can be stated, that the placement rate k= −200 has to fulfill the dynamic requirements observing the throttle valve in every condition.

As a result of our examination, we can state two things:

Firstly, the more the pole of the Observer is negative placed in the left direction, the Observers settling time will be shorter and shorter. Secondly, as far as the estimation error condition is increasing, the Observers settling time will be slower and slower. With setting the eigenvalues 5 times bigger for the Observer than the eigenvalues of the dominant system, it could be achieved that the convergence of the estimation is faster than others system effects.

Outline of this work is, despite that Luenbergers Observer is not a new method, it gives a simple algorithm to state space observing of Electromechanical Actuators. With further development it could be used for condition monitoring and fault detection. That will be the goal of our investigation.

## References

[1] Luenberger, D.G.: *An Introduction to Observers*, IEEE Transaction on Automatic Control, vol. AC-16, no. 6, 1971

[2] Kailath, T.: *Linear Systems*, Prentice Hall, pp. 79-152, 1980

[3] Reichhartinger, M., Horn, M.: *Robust Position Control of an Electromechanical Actuator for Automotive Applications*, World Academy of Science, Engineering and Technology, vol. 5, no. 11, pp. 1184-1188, 2011

[4] Lunze, J.: *Regelungstechnik 2*, Springer, pp. 337-366, 2010

[5] Padmakumar, S., Vivek A.: *A Comparative Study into Observer based Fault Detection and Diagnosis in DC Motors: Part-I*, World Academy of Science, Engineering and Technology, vol. 3, no. 3, pp. 294-299, 2009

[6] Grepl, R., Lee, B.: *Modelling, Identification and Control of Electronics Throttle Body Using Dspace Tools*, International Conference Technical Computing Prague, CD proceeding, pp. 1-10, 2008

[7] Vašak, M., Petrović, I., Perić, N.: *State Estimation of an Electronic Throttle Body* , IEEE International Conference, vol. 1, pp. 472-477, 2003 DOI: 10.1109/ICIT.2003.1290368

[8] Nakano, K., Sawut, U., Higuci, K., Okajima, Y.: *Modelling an Observer-based Sliding-Mode Control of Electronic Throttle Systems*, ECTI Transactions on Electrical Eng, Electronics, and Communications, vol. 4, no. 1, pp. 22-27, 2006

[9] Kwadzogah, R., Misra, D.: *Simultaneous Identification of Friction and Transfer Function of a DC Servo Positioning System via Simulation*, International Journal of Intelligent Control and Systems, vol. 18, no. 1, pp. 10-16, 2013

[10] Angermann, A., Beuschel, M., Rau, M., Wolfahrt, U.: *Matlab®-Simulink®-Stateflow®*, Oldenburg Verlag, pp. 173-177, 2007

[11] Reinelt, W., Lundquest C.: *Observer Based Sensor Monitoring in Active Front Steering System using Explicit Sensor Faliure Modelling*, 16th IFAC World Congress, vol. 16, no. 1, pp. 1-6, 2005

Acta
Technica
Jaurinensis

Vol. 7., No. 4., pp. 344-355, 2014
DOI: 10.14513/actatechjaur.v7.n4.334
Available online at acta.sze.hu

# Mixing Efficiency Study of Nano and Micro Filled PP Systems

**H. Hargitai[1], D. Török[2]**

**[1]Széchenyi István University, Department of Materials Science and Technology
Egyetem square 1, H-9026 Győr, Hungary
E-mail: hargitai@sze.hu**

**[2]Budapest University of Technology and Economics, Faculty of Mechanical Engineering, Department of Polymer Engineering, Műegyetem rkp. 3, H-1111 Budapest, Hungary
E-mail: torok@pt.bme.hu**

Abstract: In this paper micron and nano sized fillers were compounded with polypropylene homopolymer. First masterbaches were produced with high filler content than samples were injection molded by different filler loading. The effect of using static mixers with different element number was analyzed on the mechanical properties and filler distribution. Tensile properties of PP–talc, PP–boron nitride and PP-graphene systems were investigated and optical microscopic observations were carried out for mapping the distribution of the fillers in the polymer matrix.

*Keywords: polypropylene, boron nitride, talc, graphene, filler distribution, tensile properties, static mixer*

## 1. Introduction

Since the middle of the last decade, polymeric composites have turned out to be commonly used engineering materials and manufactured in large quantities in order to be used in countless applications in sectors such as automotive, electronics, construction, household goods and etc. [1]. Main advantages of these kinds of materials are their low density and the wide range of mechanical properties, in case of some engineering type the high relative strength and stiffness. The main disadvantages for some application are their limited mechanical properties, low thermal and electrical conductivity. To improve their mechanical, thermal or electrical properties, polymers are often reinforced by incorporating fillers into the matrix.

Electronic industrial applications have special requirements because of the heat generated by the products, which need to be dissipating. Because of the remarkable development in electronics polymer composites having good thermal conduction are in the focus of several researches [2][7].

There are several factors which have a great influence on the composite's properties. It is well known that the geometry of the reinforcing particle is an important factor in achieving the appropriate (or optimal) properties of composites. In general, there is an inverse relation between the effectiveness of the reinforcement and the size of the filler (or reinforcing fiber). The greater surface-to-volume ratio of the filler the greater effectiveness can be achieved [4]. With boron nitride (BN) nanopowder, having particle size about 130 nm polypropylene (PP)/BN composite had two times higher thermal conductivity than that of composite's filled with micron-sized powder [8]. An opposite relation was found by Cheewawuttipong et al. They achieved larger thermal conductivity of PP/BN composite using BN with large (7-10μm) size than that using BN with small size (1-2μm). They concluded that the network structure of BN would be typically easy to be created by using BN with large size [3]. Moreover the volume fraction has also strong effect on the properties. Especially in case of conductivity there is a lower limit of applying fillers to form a conductive path on the composite system and by application of smaller particles, especially nano-size fillers better interaction can be achieved and thus higher conductivity [4] [9].

For improving the thermal conductivity of a polymeric material, polymer matrices are usually filled with the thermally conductive fillers like metal powders (aluminum, copper), ceramic powders (alumina diamond, silicon carbide, boron nitride) and carbon-based materials (graphite, carbon fiber) [10]. Zhou et al. found that the combined use of BN particles and alumina short fiber has synergetic effect and resulted in higher thermal conductivity of composites compared to the BN used alone [5].

To enhance both the conductivity and mechanical properties nanofillers are seems to be very effective. Nanometer sized boron nitride, copper; carbon nanotube and synthetic diamond were compared as conductive fillers by Nurul et al. The greater thermal conductivity was achieved by using carbon nanotubes however the entanglements of the filler resulted in reduced tensile properties [4].

Another attractive nanomaterial is graphene, which is a 2D structured material, planar monolayer of carbon atoms owning exceptional charge transport, thermal, optical, and mechanical properties and thus being studied in nearly every field of science and engineering [11]. There are several paper dealing with the graphene and its derivates, such as graphene oxide (GO) or reduced graphene oxide (RGO) application in polymer composites [12][18]. Song et al. found that 2 vol% graphene can be is significantly improve the thermal oxidative stability of PP by the due to the barrier effect of its lamellar structure although considerable enhancement of the mechanical properties of PP was achieved by using very low (0.5 vol%) amount of graphene [18].

One of the most critical aspects of reinforcing is the adhesion on the fiber-matrix interface. Many papers are dealing with the chemical treatment of the reinforcing fiber or chemical compatibilization during the processing to enhance the interaction between the components [13][15]. By the treatment of graphite using strong mineral acids and oxidizing agents graphite oxide can be produced which is an effective reinforcement. Chemically reduced graphene oxide has a pronounced reinforcing effect in poly(vinyl alcohol) , the tensile strength increased three times by adding 3 vol% graphene while elongation at break showed opposing trends with increasing volume fraction [13].

Talc is attractive as a filler for polypropylene (PP) for several reasons, cost effective, acts as nucleating agent in the crystallization process and offers relatively high strength and stiffness [19][21]. However, dispersion and distribution of talc within the polymer have a significant effect on the performance of the composites [19]. There are several way (mechanical or chemical) to reduce agglomeration and improve particle dispersion and distribution [19]. Castillo et al. improved the mechanical properties such as yield strength, elongation at break and toughness of PP/talc composites by grafting acetoxy groups onto the talc surface, although the effect on the modulus is less pronounced [19]. Tang et al investigated graphene oxide-epoxy composites and they found that the highly dispersed RGO resulted in higher strength and fracture toughness of epoxy resin than the poorly dispersed RGO, although no significant differences in both the tensile and flexural moduli was observed regardless of the dispersion level [16].

As it was mentioned above one of the influencing factors to make polymer composites with appropriate mechanical and other physical properties is the distribution of the fillers in the polymer matrix. There are not only chemical methods to achieve good distribution. Using injection molding static mixing nozzles a homogeneous mixing of polymer melt can be created during injection. The resulting high viscous plastics melt flow is homogeneous with regard to colorant, additives and temperature (Fig. 1.). This kind of melt flow mixing of molten polymer prior to injection results in several benefits mainly used in coloring, such as narrower part tolerance, less part distortion, shorter cycle time or improved melt flow [22].

*Figure 1. Stamixco injection molding static mixing nozzle with the elements (left);Empty nozzle provides no mixing (right top) and eight SMN mixing elements create homogeneous mixing in a short length (right bottom) [22].*

In this study micron and nano sized fillers which are used for producing thermally conductive polymer composites were compounded with polypropylene homopolymer. First masterbaches were made with high filler content than samples were injection molded by different filler loading. The effect of using static mixers was analyzed on the mechanical properties and filler distribution.

## 2. Experimental

### 2.1. Materials

All the materials used in this study are commercially available. Homopolymer polypropylene (PP, TIPPLEN H145 F (TVK, Hungary) was used as matrix material. This

grade of PP has a specific gravity of 0.9 g/cm$^3$, a melt flow index of 25 g/10 min (at 230°C and 2.16 kg of load).

Three kind of fillers were used; micron sized talc (QualChem; 35 μm, specific gravity of 2.7 g/cm$^3$) and boron nitride (BN) (Henze Hebofil® 482; specific gravity of 2.1 g/cm$^3$) and nano sized graphite powder Timrex C-Therm 011 (Timcal Ltd., Switzerland).

### 2.2. Sample preparation

Sample preparation was a two-step process, first melt mixing to make masterbaches than injection molding of test parts.

Polypropylene and the fillers were compounded in a LabTech Scientific-type laboratory twin-screw extruder (screw diameter of 20 mm and L/D ratio of 44). The zone temperatures of the extrusion were 190 to 240°C After extrusion, the composite was solidified and cooled in a water bath then pelletized. In the melt mixing step talc-PP, BN-PP and graphene-PP masterbaches were produced; 30 vol% of talc, same amount of boron nitride or 2 vol% of graphene were incorporated separately into the PP melt, producing tree different masterbaches.

Arburg 370S 700-290 Advance injection molding machine with screw diameter of 30 mm has been used for sample preparation by varying the amount of the fillers and mixing conditions. The filler content of the composite samples can be seen in Table 1.

*Table 1. Composition of the filled PP materials*

| ***Filler content [volume%]*** | ***Masterbach compounds*** | ***Injection molded specimen*** | | | |
|---|---|---|---|---|---|
| | | ***No.1*** | ***No.2*** | ***No.3*** | ***No.4*** |
| Talc | 30 | - | 5 | 10 | 20 |
| Boron nitride | 30 | - | 5 | 10 | 20 |
| Graphene | 2 | 0.2 | 0.5 | 1 | 2 |

From the masterbaches plaque specimens having 80x80x2 mm in dimension were injection molded. In the mold a fan gate measuring 1 mm in thickness were used. The injection molding parameters are summarized in Table 2.

*Table 2. Injection molding set-up parameters*

| ***Injection molding parameter*** | ***Value*** |
|---|---|
| Injection volume [cm$^3$] | 49 |
| Injection rate [cm$^3$/s] | 50 |
| Holding pressure [bar] | 80% of the injection pressure |
| Holding time [s] | 10 |
| Residual Cooling time [s] | 10 |
| Clamping force [kN] | 650 |
| Screw rotational speed [m/min] | 15 |
| Melt temperature [°C] | 200 |
| Temperature of the mold [°C] | 50 |

To study the effect of static mixers on the filler distribution Stamixco SMN type static mixers with 5 and 8 elements has been used in the injection molding process. Reference samples were also made without using static mixer.

### 2.3. Test methods

## Tensile test

Three pieces of small size dump-bell shaped 5A type specimens according to the MSZ EN ISO 527-1:1999 standard were cut from the injection molded plates by waterjet equipment parallel with the injection molding direction (Fig. 2). Tensile tests were carried out at room temperature on a standard tensile Zwick machine; model Z020 by using cross-head speed of 2 mm/min and clamping length of 50 mm.

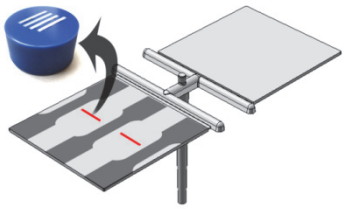


*Figure 2. Preparation of dumbbell specimens from the injection molded plaque plates and the positions (red lines) of the cross-sections for optical microscopy*

Stress-strain curves were obtained for each specimen, where Young's modulus and tensile strength were determined. Tests were carried out on at least five samples for each composite in order to minimize the error.

## Optical microscopic investigations

To study the distribution of the fillers in the PP matrix the cross section in the same position of each sample were analyzed. The polished samples were observed by Zeiss Axio Imager M1 optical microscope with different magnifications (50X, 200X, 500X and 1000X) and imaging modes such as Polarized, Bright Field, Dark Field. Differential Interference Contrast (DIC) technique relies on Polarized illumination was also used to create a 3D effect of the specimen's surface. The prisms placed in the condenser and in the back focal plane of the objective modify the normal extinction resulting from the crossed polarizers.

## 3. Results and discussion

### 3.1. Tensile properties

The tensile strength and modulus of talc and boron nitride filled composites are shown in Figure 3. It was observed that the tensile strengths decreased, while the modulus increased with the filler content for both talc and boron nitride. Decrease of the strength could be due to the week interfacial bond between the components.

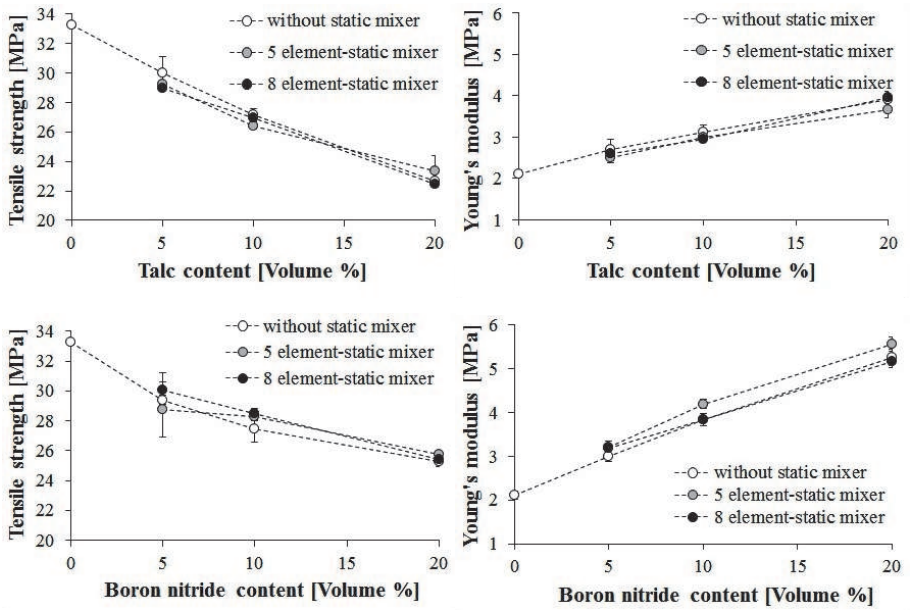



*Figure 3. Tensile strength and Young's modulus of talc (top) and boron nitride (bottom) filled polypropylene composites as a function of filler content made without, with 5 and 8 element static mixers*

Lower decrease of the strength and higher increase of the modulus of boron nitride –PP composites compared to the talc composites was found. By analyzing the effect of the static mixer no significant differences can be observed between the mechanical parameters by using the 5 or 8 element systems. In case of higher filler content a slight increase of properties can be seen in comparison with the composites made without static mixer.

Figure 4. shows the results of the nano-filled graphene-PP composites. The same tendencies with slighter effect are occurred as in case of the talc and BN composites.

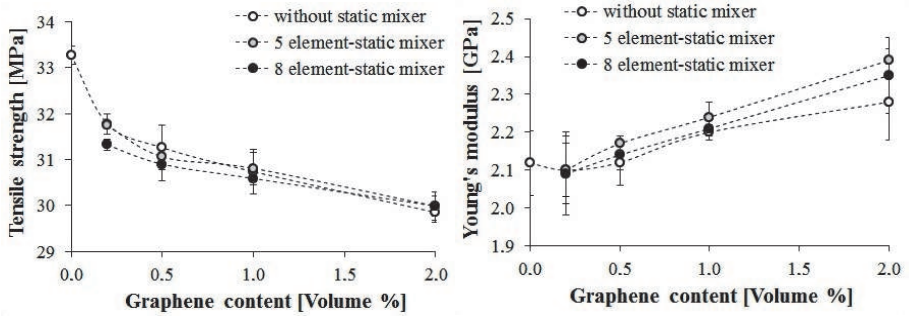



*Figure 4. Tensile strength and Young's modulus of graphene-polypropylene composites as a function of filler content made without, with 5 and 8 element static mixers*

For comparison the effect of the three types of fillers each result are summarized in the diagrams of Figure 5. From the point of view the tensile properties the best results were obtained by applying boron nitride fillers. The same tendencies are clearly seen as the slight effect of the static mixing on the tensile properties of the composites.

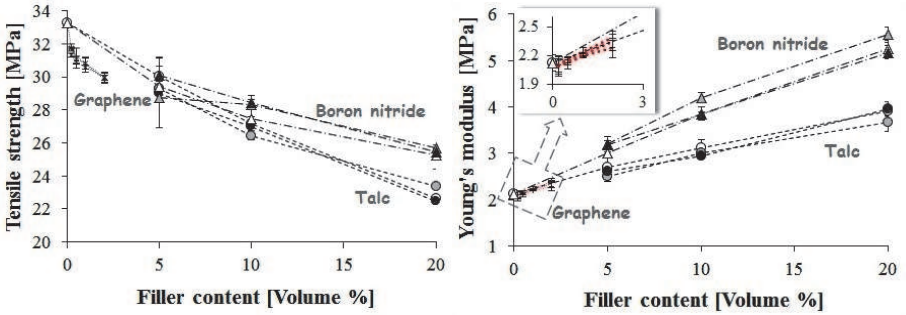



*Figure 5. Tensile strength and Young's modulus of the composites for each filler type as a function of filler content made without, with 5 and 8 element static mixers*

### 3.2. Microscopic analysis

For analysis the distribution of the filler particles in the polypropylene optical microscopic pictures were taken from the polished cross-sectional surfaces of the composites. Bright Field imaging mode or DIC prism in polarized mode was applied to get pictures which can be analyzed.

By evaluating the microscopic images no significant differences were found between the composites made by different mixing modes. No evaluable pictures were taken from the graphene-PP composites. On Figure 6. the slight effect of using 8 element static mixer can be seen in case of boron nitride filled composites for each filler content.

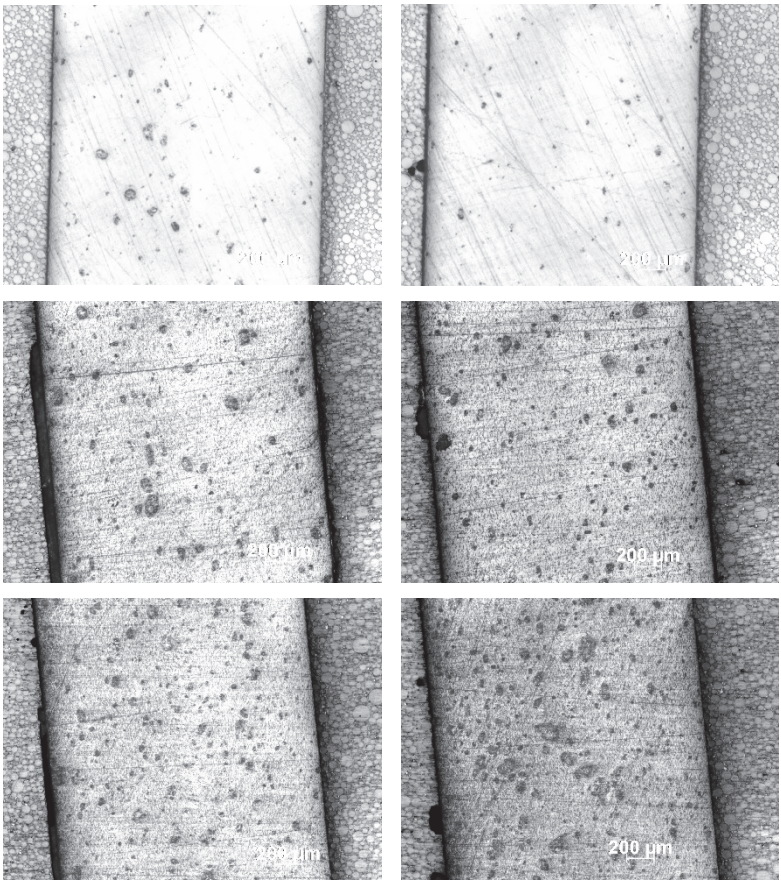



*Figure 6. Optical microscopic images (Bright Field, 50X) of boron nitride-PP composites produced without (left) and with 8 elements static mixer (right) with 5 vol% (top). 10 vol% (middle) and 20 vol% (bottom) filler content*

Differential Interference Contrast (DIC) prism was used on polarized illumination to create 3D-like pictures. Smaller agglomeration sizes can be found by using 8 element static mixer in case of both 5 vol% boron nitride and 5 vol% talc filled composites (Figure 7.).

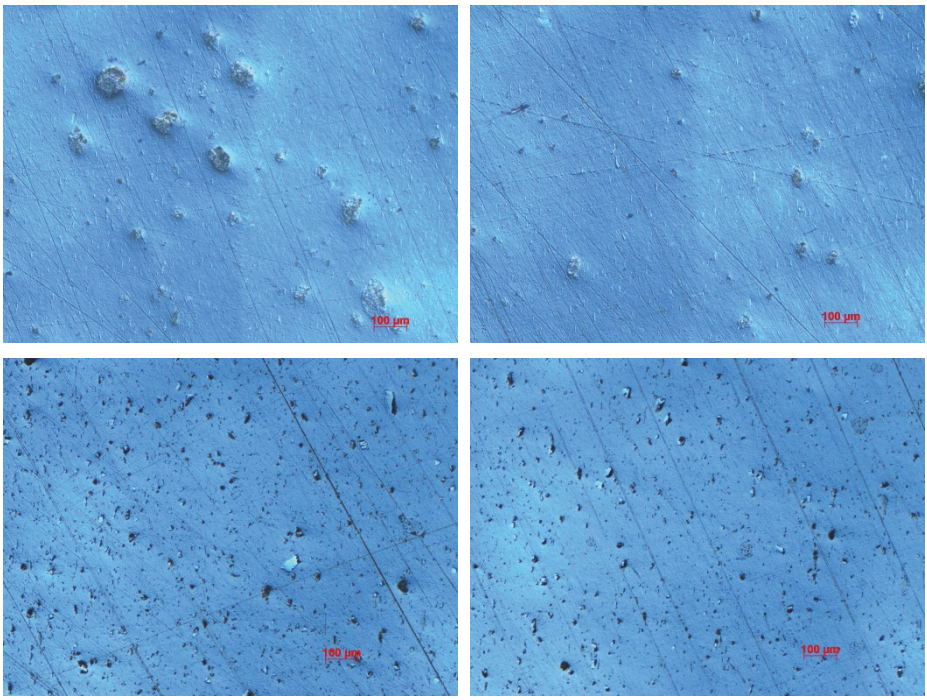



*Figure 6. Optical microscopic images (polarised light and DIC prism, 100X) of 5 vol% boron nitride-PP (top) and 5 vol% talc-PP (bottom) composites produced without (left) and with 8 elements static mixer (right)*

## 4. Conclusions

In our experiments we focused on thermal conductive polymeric composite materials, especially to affect the distribution of the fillers in the polymeric matrix material. Micron sized boron nitride and talc fillers and nanosized graphene were compounded with polypropylene homopolymer. Masterbaches were made with high filler content in laboratory twin screw extruder than samples were injection molded by different filler loading. Static mixers with 5 and 8 elements were used and their effect on the tensile properties and the filler distribution were analyzed.

Based on the tensile test results of PP–talc, PP-boron nitride and PP-graphene systems and the optical microscopic observations of the polished cross-sectional surfaces of the composites the following can be concluded:

- Tensile strength of each composite decreased with the filler content, while an opposite tendency was observed in case of Young's modulus.
- No significant effect of the static mixers was found, but increasing the filler content the use of static mixers showed a slight increase in the properties compared to those, which have made without static mixer.

- In comparison the three different filler composites, the best properties were achieved by incorporating boron nitride into the matrix PP.
- The optical microscopic images show no significant effect of the static mixing.

Based on our experimental results and the review of the scientific literature the positive effect of the fillers is depending on several factors, such as filler dispersion and distribution, filler content and filler size. Beside these factors the strength of the interfacial adhesion may has the greatest impact on the mechanical properties. The week interaction between the components resulted in the decreasing strength of the composites compared to the neat polymer. In further researches the interfacial adhesion should be investigated as well.

## Acknowledgement

This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

The research work presented in this paper was carried out as part of the TÁMOP-4.2.2.A-11/1/KONV-2012-0029 project in the framework of the New Széchenyi Plan. The realization of this project is supported by the European Union, and co-financed by the European Social Fund.

## References

[1] Hargitai, H., Rácz, I.: *Applications of Macro- and Microfiller-Reinforced Polymer Composites* in Polymer Composites, vol. 1, Macro- and Microcomposites, Editors: Sabu, T., Kuruvilla, J., Malhotra, S. K., Goda, K., Sreekala, M. S., Wiley-VCH Verlag, Weinheim, pp. 749-791, 2012
DOI: 10.1002/9783527645213.ch23

[2] Suplicz, A., Szabo, F., Kovacs, J. G.: *Injection molding of ceramic filled polypropylene: The effect of thermal conductivity and cooling rate on crystallinity*, Thermochimica Acta, vol. 574, pp. 145-150, 2013
DOI: 10.1016/j.tca.2013.10.005

[3] Cheewawuttipong, W., Fuoka, D., Tanoue, S., Uematsu, H., Iemoto, Y.: *Thermal and Mechanical Properties of Polypropylene/Boron Nitride Composites*, Energy Procedia, vol. 34, pp. 808-817, 2013
DOI:10.1016/j.egypro.2013.06.817

[4] Nurul, M. S., Mariatti, M.: *Effect of thermal conductive fillers on the properties of polypropylene composites*, Journal of Thermoplastic Composite Materials, vol. 26, no. 5, pp. 627-639, 2013
DOI: 10.1177/0892705711427345

[5] Zhou, W., Qi, S., An, Q., Zhao, H., Liu, N.: *Thermal conductivity of boron nitride reinforced polyethylene composites*, Materials Research Bulletin, vol. 42, no. 10, pp. 1863-1873, 2007
DOI: 10.1016/j.materresbull.2006.11.047

[6] Suplicz, A., Kovács, J. G.: *Development of Thermally Conductive Polymer Materials and their Investigation*, Materials Science Forum, vol. 729, pp. 80-84, 2013

DOI: 10.4028/www.scientific.net/MSF.729.80

[7] Suplicz, A., Kovács, J. G.: *Thermally Conductive Polymer Compounds for Injection Moulding: Synergetic Effect of Hexagonal Boron Nitride and Talc*, Journal of Reinforced Plastics And Composites, vol. 32, no. 16, pp. 1234-1240, 2013
DOI: 10.1177/0731684413489755

[8] Muratov, D.S., Kuznetsov, D.V., Il'inykh, I.A., Mazov, I.N., Stepashkin, A.A., Tcherdyntsev, V.V.: *Thermal conductivity of polypropylene filled with inorganic particles*, Journal of Alloys and Compounds, vol. 586, suppl. 1, pp. S451-S454, 2014
DOI: 10.1016/j.jallcom.2012.11.142

[9] Schöne, J., Kotter, I., Grellmann, W.: *Properties of Polypropylen Talc Compounds with Different Talc Particle Size*, Journal of Plastics Technology, vol. 8, no. 2, pp. 231-251, 2012

[10] Hong, J., Park, D. W., Shim, S. E.: *A review on thermal conductivity of polymer composites using carbon-based fillers : carbon nanotubes and carbon fibers*, Carbon Letters, vol. 11., no. 4., pp. 347-356, 2010
DOI : 10.5714/CL.2010.11.4.347

[11] Singh, V., Joung, D., Zhai, L., Das, S., Khondaker, S. I., Seal, S.: *Graphene based materials: Past, present and future*, Progress in Materials Science, vol. 56, no. 8, pp. 1178-1271, 2011
DOI: 10.1016/j.pmatsci.2011.03.003

[12] Kuilla, T., Bhadra, S., Yao, D., Kim, N. H., Bose, S., Lee, J. H.: *Recent advances in graphene based polymer composites*, Progress in Polymer Science, vol. 35, no. 11, pp. 1350-1375, 2010
DOI: 10.1016/j.progpolymsci.2010.07.005

[13] Potts, J. R., Dreyer, D. R., Bielawski, C. W., Ruoff, R. S.: *Graphene-based polymer nanocomposites*, Polymer, vol. 52, no. 1, pp. 5-25, 2011
DOI: 10.1016/j.polymer.2010.11.042

[14] Hu, K., Kulkarni, D. D., Choi, I., Tsukruk, V. V.: *Graphene-polymer nanocomposites for structural and functional applications*, Progress in Polymer Science, 2014
DOI: 10.1016/j.progpolymsci.2014.03.001

[15] Yang, X., Wang, X., Yang, J., Li, J., Wan, L.: *Functionalization of graphene using trimethoxysilanes and its reinforcement on polypropylene nanocomposites*, Chemical Physics Letters, vol. 570, pp. 125-131, 2013
DOI: 10.1016/j.cplett.2013.03.069

[16] Tang, L-C., Wan, Y.-J., Yan, D., Pei, Y-B., Zhao, L., Li, Y-B., Wu, L-B., Jiang, J-X., Lai, G-Q.: *The effect of graphene dispersion on the mechanical properties of graphene/epoxy composites*, Carbon, vol. 60, pp. 16-27, 2013
DOI: 10.1016/j.carbon.2013.03.050

[17] Milani, M. A., González, D., Quijada, R., Basso, N. R.S., Cerrada, M. L., Azambuja, D. S., Galland, G. B.: *Polypropylene/graphene nanosheet nanocomposites by in situ polymerization: Synthesis, characterization and fundamental properties*, Composites Science and Technology, vol. 84, pp. 1-7, 2013
DOI: 10.1016/j.compscitech.2013.05.001

[18] Song, P., Cao, Z., Cai, Y., Zhao, L., Fang, Z., Fu, S.: *Fabrication of exfoliated graphene-based polypropylene nanocomposites with enhanced mechanical and thermal properties*, Polymer, vol. 52, no. 18, pp. 4001-4010, 2011
DOI: 10.1016/j.polymer.2011.06.045

[19] Castillo, L., Barbosa, S., Capiati, N.: *Improved performance of polypropylene/talc composites*, Society of Plastics Engineers Plastics Research Online, 2012
DOI: 10.2417/spepro.004405

[20] Rao Patnaik, K.S.K., Sirisha Devi, K., Kiran Kumar, V.: *Non-isothermal Crystallization Kinetics of Polypropylene (PP) and Polypropylene (PP)/Talc Nanocomposite*, International Journal of Chemical Engineering and Applications vol. 1, no. 4, pp. 346-353, 2010
DOI: 10.7763/IJCEA.2010.V1.60

[21] Castillo, L. A., Barbosa, S. E., Capiati, N. J.: *Influence of talc morphology on the mechanical properties of talc filled polypropylene*, Journal of Polymer Research, vol. 20, pp. 152-160, 2013
DOI 10.1007/s10965-013-0152-2

[22] Schneider, G., Ungvári, Gy.: *Stamixco static mixer for producing homogeious melt during injection molding (in Hungarian)*, Műanyag és Gumi, vol. 44. no. 3, pp. 111-115, 2007

[23] Zsíros, L., Kovács, J. G.: *Optimization of homogenizing capacity of injection moulding machines (in Hungarian)*, Műanyag és Gumi, vol. 50. no. 9, pp. 347-350, 2013

Acta
Technica
Jaurinensis

Vol. 7., No. 4., pp. 356-367, 2014
DOI: 10.14513/actatechjaur.v7.n4.338
Available online at acta.sze.hu

# Design of Small Dual Band Microstrip Antenna for Broadband Applications

**R.S. Parbhane**

**SSGBCOET, Department of Electronics and Communication**
**Bhuswal, 425201, India**
**Email:parbhanerahul67@gmail.com**

Abstract: In this article, a small dual band rectangular slot antenna for broadband application is presented. The antenna is being excited by capacitively coupled probe feed. The antenna consists of rectangular slot from the center of the patch. The rectangular slot is being used to obtain dual frequencies of 2GHz and 3.34GHz.Instead of a rectangular feed strip, antenna is excited by a triangular feed strip with same dimension. Triangular feed strip is used to obtain a broad band frequency range. The size of the antenna is 100×100×1.6mm3with the ground of same dimension. The proposed antenna is simulated and optimized using IE3D simulation software.

*Keywords: return loss, VSWR, FR4, Air Gap, SMA connector, MOM simulation, slotted microstrip antenna*

## 1. Introduction

Today as the research is increasing day by day, so in high performance application such as aircraft, spacecraft, satellite and missile applications, where size, weight, cost, performance, ease of installation and aerodynamic profile are constraints low profile antenna is used. Microstrip antenna is being designed to work in broadband as well as narrow and wideband frequency range [1, 2]. A slotted circular monopole antenna is presented [3] has spike shaped slots with square patch rotated around 45 degree to achieve ultra-wideband (UWB) applications.

Coplanar capacitively coupled microstrip antenna [4] is used for wideband applications with impendence bandwidth of 50%.A printed Egg curved slot antenna[5]for wideband frequency range with gain of 4.1-5.1dBi. Many microstrip antennas which operate for dual and tri-band operation are being reported [6-11]. For example bandwidth enhancement of printed slot antenna [6] having wideband of 2.80 to 11.81 GHz is reported. Antennas reported in [7] have a circular polarization with frequency ratio of 1.11. On the other hand, the antenna in [8] offers for ultra wideband

frequency band from 840MHz to 960MHz and [9] offers for maximum antenna bandwidth.

Antenna reported in [11] offers for a dual band frequency range for a dual square ring slot. Polarization is also main characteristics in design of antenna. Antenna in [12] offers for circular polarization with axial ratio (AR) below 2dB. Antenna proposed in [13, 14] have a slot along the patch to enhance for dual band operation. In the proposed antenna an air gap with dual band frequency with good bandwidth is being proposed

The basic geometry is shown in section 2. The design starts with the selection of center frequency and it may be scaled to any frequency of interest. Design and optimization procedure are shown in section 3. Sections 4 show the Experimental validation along with discussion. Conclusion of this study is shown in section 5.

## 2. Antenna Geometry

The proposed antenna has a rectangular slot in vertical direction from center of the patch. Substrate used for design is FR4 with dielectric constant of 4.4 and thickness of 1.6mm with air gap of 7.5mm. A long pin SMA connector is used to connect the feed strip which couples the energy to the patch by capacitive means. The detailed optimization procedure of the antenna and their optimum dimension characteristics are presented in section 3.

The antenna was designed to operate with center frequency of 2.4 GHz. Rectangular antenna with rectangular slot with all physical parameter is shown in the Fig. 1. which is optimized with the IE3D [15] which is a method of moment (MOM) simulation software. The details dimensions of the optimized antenna are listed in Table 1.

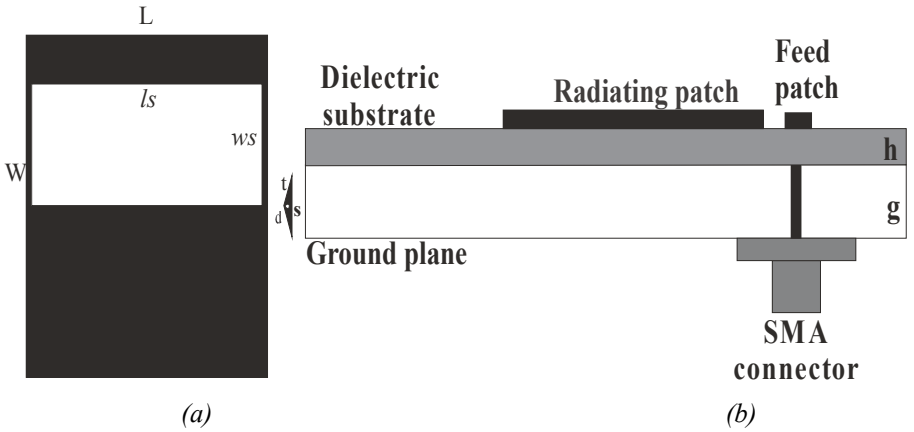



*Figure 1. Geometry of patch antenna. (a) Top view (b) Cross sectional view*

*Table1. Optimized dimensions of the proposed antenna*

| Antenna Parameters | Values with Air Gap(mm) |
|---|---|
| Length of radiator patch(L) | 42.0 |
| Width of radiator patch(W) | 57.0 |
| Length of feed strip(s) | 11.1 |
| Width of feed strip(t) | 1.5 |
| Separation between feed strip from the patch(d) | 0.5 |
| Air gap(g) | 7.5 |
| Slot length(*ls*) | 40.0 |
| Slot width(*ws*) | 20.0 |
| Slot position(p) (from center of patch) | 10.0 |

## 3. Geometry Optimization and Discussions

The optimization process is shown in this section. The key design parameters used for the design are air gap, distance between radiative patch and feed, slot position, sloth width, and slot length. The details are given in the following subsections.

### 3.1. Effect of Air Gap (g)

As shown in Fig. 2., air gap of antenna is being varied from 6.5 mm to 11.5 mm in steps of 1mm. air gap (g) is used to maximize the antenna's bandwidth. Air gap of 7.5 mm is used to maximum bandwidth and maximum gain. As air gap changes, there is shift in the resonance frequency of the antenna. The upper frequency goes on increasing as we increase the air gap (g). From Fig. 2., it may be noted that the upper and lower cutoff frequency varies accordingly as the air gap change. From Table 2. it may be noted that maximum bandwidth is obtained at 1.92-3.52 GHz frequency i.e 65% for the proposed geometry. As the air gap goes on increasing the bandwidth of proposed geometry goes on reducing.

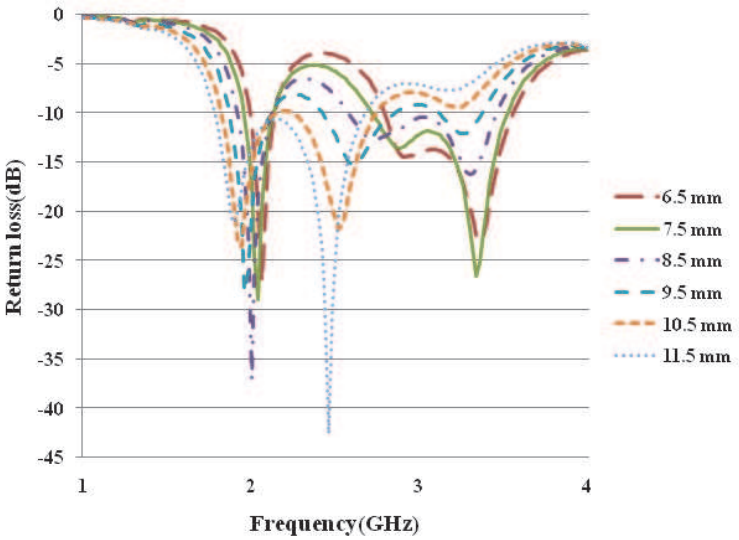



*Figure 2. Return loss characteristics for different air gap.*

*Table 2. Effect of Variation of Air Gap on bandwidth of proposed antenna*

| **Air Gap(g) (mm)** | 6.5 | 7.5 | 8.5 | 9.5 | 10.5 | 11.5 |
|---|---|---|---|---|---|---|
| **Frequency range (GHz)** | 2-3.5 | 1.96-3.52 | 1.92-3.46 | 1.88-3.38 | 1.84-2.74 | 1.8-2.66 |
| **Bandwidth (%)** | 62.5 | 65.0 | 64.2 | 62.5 | 37.5 | 35.8 |

### 3.2. Effect of distance between radiator patch and feed strip (d)

The distance between the patch and feed strip plays an important role in design of the antenna. It does not change the bandwidth of the antenna but shows a change in the depth of $S_{11}$ parameter of the antenna. The distance between radiator and strip is change from 0.4 mm to 0.9 mm in steps of 0.1 mm each. The variation in $S_{11}$ parameter of radiator patch and feed strip is shown in Fig. 3. From Fig. 3., it is clear that the lower cut-off frequency remains constant, only the upper cut-off frequency changes. From Table 3 it is clear that the optimum result for the distance with maximum bandwidth of 65% is obtained at d=0.5mm.

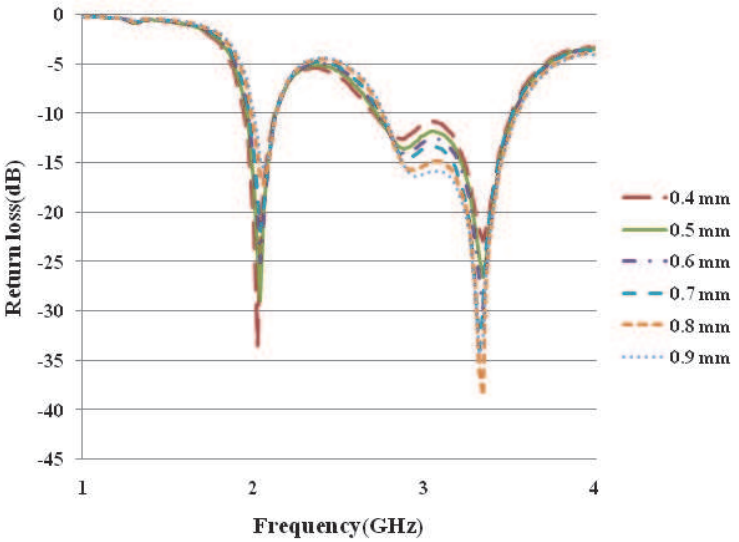



*Figure 3. Return loss characteristics for different feed strip distance*

*Table 3. Effect of variation of distance between patch and strip on bandwidth of proposed antenna*

| **Patch and feed distance (d) (mm)** | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|
| **Frequency range (GHz)** | 1.96-3.52 | 1.98-3.54 | 1.98-3.5 | 2-3.52 | 2-3.52 | 2-3.54 |
| **Bandwidth (%)** | 64.1 | 65.0 | 63.3 | 63.3 | 63.3 | 64.1 |

### 3.3. Effect of slot position (p)

Rectangular slot is being introduced from center of the patch at a position of 10mm to obtain maximum bandwidth. The slot position is varied from 8mm to 13mm in steps of 1mm. The optimized $S_{11}$ parameter is being shown in the Fig. 4. The slot is being introduced in vertical direction to obtain a dual band frequency range. From Fig. 4, it is clear that the lower cut-off frequency of the geometry does not change but there is change in the upper cut-off frequency of the geometry. From Table 4, the optimum slot position having maximum bandwidth is obtained at p=10mm.

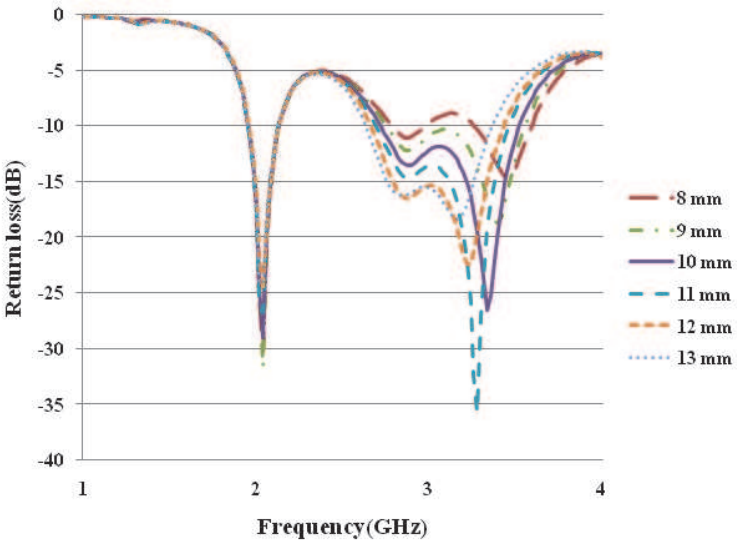



*Figure 4. Return loss characteristics for different slot position*

*Table 4. Effect of variation of slot position on bandwidth of proposed antenna.*

| **Slot position(p) (mm)** | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|
| **Frequency range(GHz)** | 1.96-3.58 | 1.96-3.54 | 1.92-3.54 | 1.96-3.46 | 1.96-3.42 | 1.96-3.36 |
| **Bandwidth (%)** | 67.5 | 65.8 | 67.5 | 62.5 | 60.8 | 58.3 |

### 3.4. Effect of Slot Length (*ls*)

Keeping all the above parameter constant, slot length is varied from 37mm to 42mm in steps of 1 mm each. The $S_{11}$ parameter of slot length variation is shown in Fig. 5. From Fig. 5, as we increase the length of the slot the upper cut-off frequency changes accordingly. As we increase the slot length above 41mm the bandwidth of the antenna reduces as shown in the Table 5. The maximum bandwidth is obtained at the frequency range of 1.96-3.52 GHz which is 65%. The optimum slot length is being obtained at *ls*=40mm without the shift in the resonance frequency.

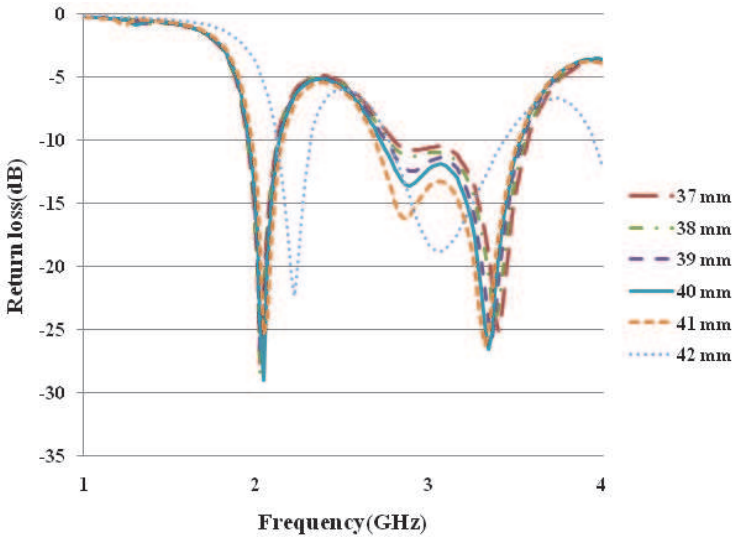



*Figure 5. Return loss characteristics for different length of slot*

*Table 5. Effect of variation of slot length on bandwidth of proposed antenna.*

| **Slot length(*ls*) (mm)** | 37 | 38 | 39 | 40 | 41 | 42 |
|---|---|---|---|---|---|---|
| **Frequency range (GHz)** | 1.96-3.51 | 1.96-3.52 | 1.96-3.5 | 1.96-3.52 | 1.98-3.5 | 2.12-3.42 |
| **Bandwidth (%)** | 64.6 | 64.6 | 64.2 | 65.0 | 63.3 | 54.2 |

### 3.5. Effect of slot width (*ws*)

The slot width is being varied from 18mm to 21mm in steps of 1mm keeping above parameter constant. The $S_{11}$parameter change is shown in Fig. 6. From Fig. 6., it is clear that there is slight change in the upper and lower cut-off frequency of the proposed geometry. As *ws*=22mm shows for maximum depth in the return loss but there is slight shift in the resonance frequency of the antenna. From Table 6, it is being noted that the optimum slot width is obtained at *ws*=20mm without change in the resonance frequency having maximum bandwidth of 65%.

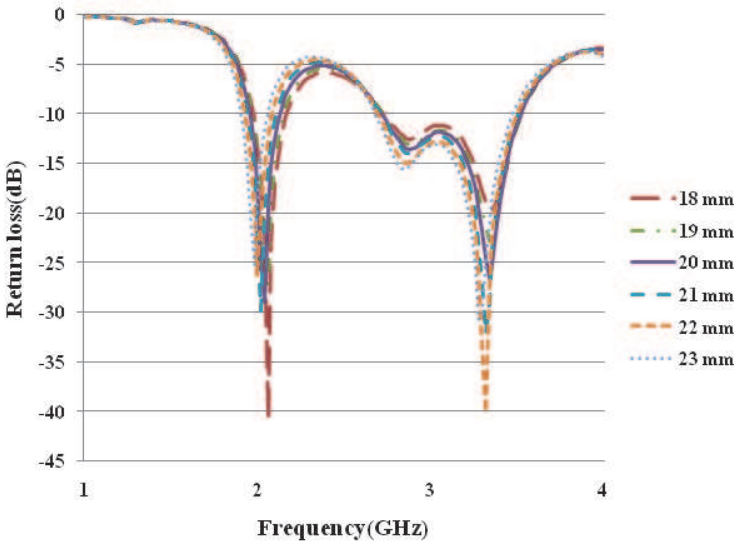



*Figure 6. Return loss characteristics for different width of slot*

*Table 6. Effect of variation of slot width on bandwidth of proposed antenna*

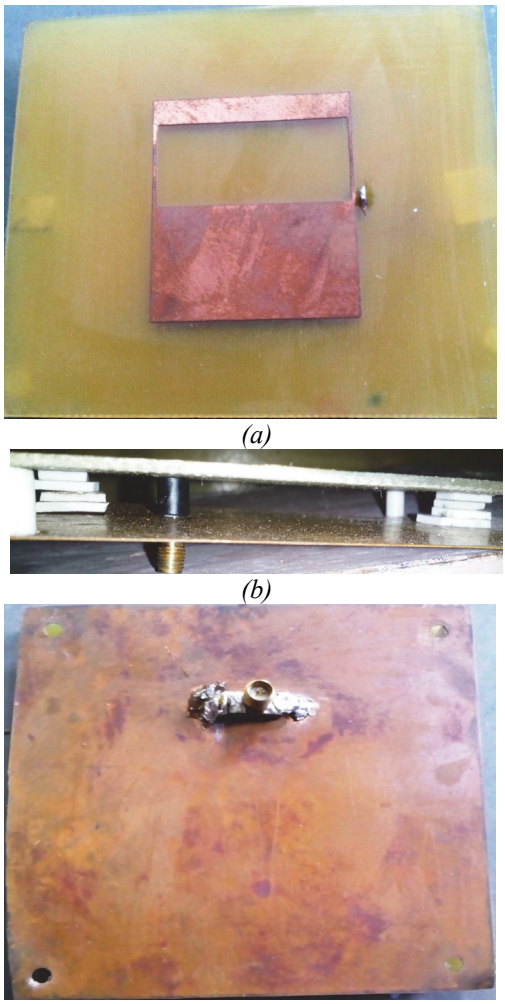
| **Slot width(*ws*) (mm)** | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|
| **Frequency range (GHz)** | 1.98-3.52 | 1.98-3.52 | 1.96-3.52 | 1.96-3.5 | 1.94-3.5 | 1.92-3.46 |
| **Bandwidth (%)** | 64.2 | 64.2 | 65.0 | 64.2 | 65.0 | 64.2 |

From all the cases studied, the optimum set of parameter are d=0.5mm, t=1.5mm, s=10mm, along with slot parameter p=10mm, *ls*=40mm and *ws*=20mm. As IE3D assumes infinite ground and substrate dimension the optimized geometry was resimulated using Ansoft HFSS v.13 [16]. Also, at both the resonant frequency, more than 5dB gain was observed. Detailed studies of parameter have been conducted for various designs.

## 4. Experimental Results and Discussions

The prototype antenna with dimension listed in Fig. 1. with optimum parameter using IE3D presented in Table 1. was fabricated and tested. Return loss comparison is shown in Fig. 8.The substrate used for manufacturing is FR4 glass epoxy with dielectric constant of 4.4 and thickness 1.6 mm along with air gap of 7.5 mm. The substrate is assembled above copper ground plane of dimension 100×100×1.6 $mm^3$.A photography of the antenna is shown in the Fig. 7. Comparison of $S_{11}$ parameter can be seen in Fig. 8. The pin of SMA connector is extended to reach the feed strip and is soldered there.

From Fig. 8. it is clear that the proposed antenna has an operating frequency range at 2.4 GHz. The prototype antenna was tested for $S_{11}$ using Agilent Technologies N9925A. From Fig. 9. it is clear that voltage standing wave ratio (VSWR) of the proposed antenna is below 2 dB at the operating frequency 2.4 GHz. The gain of antenna is above 5dB at the operating frequency shown in Fig. 10. The measured gain fairly agrees with the stimulated gain of the proposed antenna.



*(a)*

*(b)*

*(c)*

*Figure 7. Fabricated prototype (a) Front side (b) Air Gap (c) Rear side*

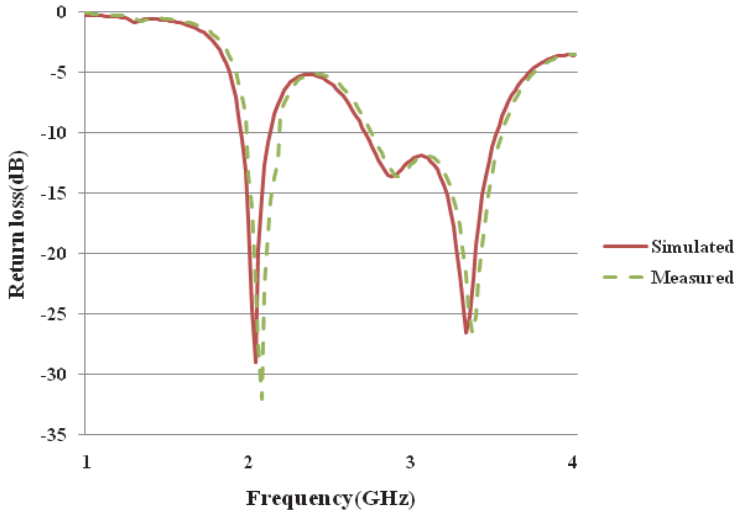



*Figure 8. Return loss characteristics of antenna shown in Figure 1 with stimulated and measured results*

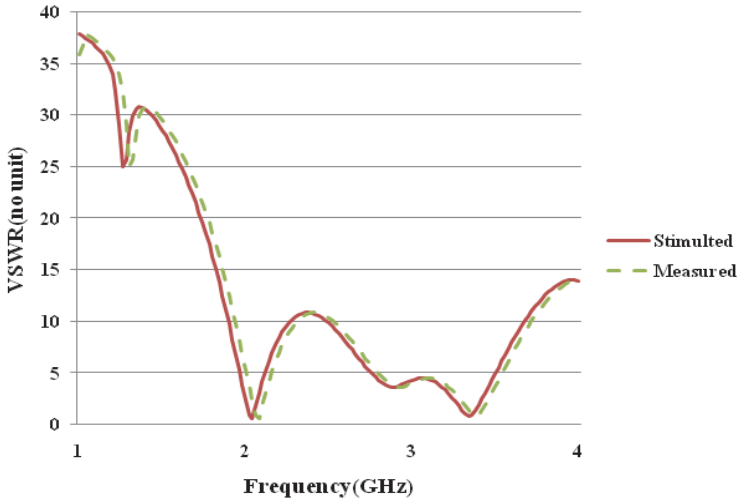



*Figure 9. VSWR versus frequency of proposed antenna shown in Figure 1.*

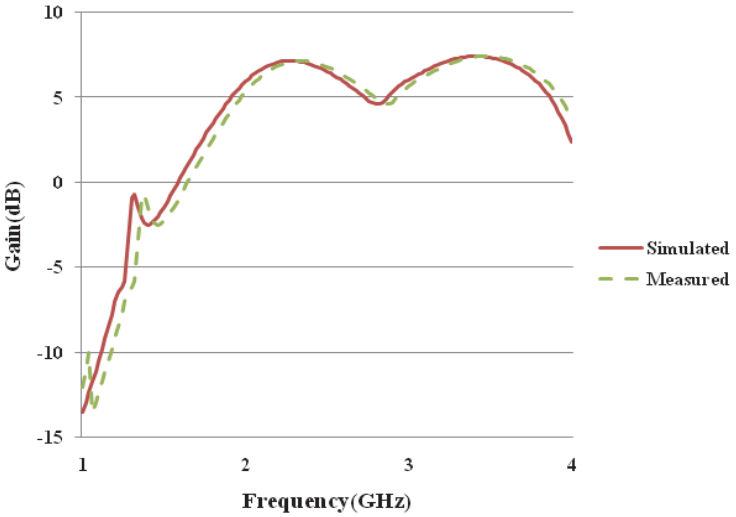



*Figure 10. Gain versus frequency of proposed antenna shown in Figure 1.*

## *5.* Conclusion

Rectangular slot antenna for broadband application with a finite ground is being presented. By varying the slot length and width dual band frequency is being achieved. The antenna presented here offers impedance bandwidth of 5.83 % and 33.3% in the frequency range of 1.96GHz to 2.1GHz and 2.72GHz to 3.52GHz respectively. The gain of the proposed antenna has value greater than 5dB. The proposed antenna is simple, easy to fabricate and need to investigate with less parameter. The antenna presented here is suitable for broadband applications.

## References

[1] Kumar, G., Ray, K.P.: *Broadband Microstrip Antenna*, Artech House, 2003

[2] Balanis, C.A.: *Antenna Theory*, 2nd Edition, John Wiley & Sons, Inc., 2004

[3] Dadime, G.M., Kasabegoudar, V.G.: *A slotted circular monopole antenna for wireless applications*, International Journal of wireless Communications and Mobile Computing, vol. 2, no. 2, pp. 30-34, 2014

[4] Kasabegoudar, V.G., Vinoy, K.J.: *Coplanar Capacitively Coupled Probe Fed Microstrip Antennas For wideband Applications*, IEEE Transactions On antennas and propagation, vol. 58, no. 10, pp. 3131-3138, 2010
DOI: 10.1109/TAP.2010.2055781

[5] Subhanshu Verma, Preetam Kumar: *Printed Egg Curved Slot Antenna for wideband Applications*, Progress in Electromagnetic Research B, vol. 58, pp. 111-121, 2014
DOI: 10.2528/PIERB14010402

[6] Yingying Tan, Liping Yan, Xiang Zhao, Changjun Liu, Kama Huang: *Bandwidth Enhancement of a Printed Slot Antenna with a Diamond Shaped Tuning Stub*, Progress in Electromagnetic Research C, vol. 50, pp. 87-93, 2014
DOI: 10.2528/PIERC14032404

[7] Jianjun Wu, Yingzeng Yin, Zedong Wang, Ruina Lian: *Dual Band Circularly Polarized Antenna with Differential Feeding*, Progress in Electromagnetic Research C, vol. 49, pp. 11-17, 2014
DOI: 10.2528/PIERC14030603

[8] Shaoshuai Zhang, He Huang, Yingzeng Yin: *A Broadband CPW- Fed Circularly Polarized Square Slot Antenna For UHF RFID Application*, Progress in Electromagnetic Research C, vol. 50, pp. 39-46, 2014
DOI: 10.2528/PIERC14021707

[9] Veeresh G. Kasabegoudar, Dibyant S. Upadhyay, K.J Vinoy: *Design Studies of Ultra-Wideband Microstrip Antennas with Small Capacitive Feed*, Hindavi Publishing Corporation International Journal of Antenna and Propagation, Article ID 67503, pp. 1-8, 2007
DOI: 10.1155/2007/67503

[10] Wen-Chung Liu, Chao-Ming Wu, Yang Dai: *Design of Triple Frequency Microstrip Fed Monopole Antenna usion Defected Ground Structure*, IEEE Transactions On antennas and propagation, vol. 59, no. 7, pp. 2457-2463, 2011
DOI: 10.1109/TAP.2011.2152315

[11] G.F.Khodaei, J.Nourinia, C. Ghobadi: *A Practical Miniaturized U-Slot Patch Antenna With Enhanced Bandwidth,* Progress in Electromagnetic Research B, vol. 3, pp. 47-62, 2008
DOI: 10.2528/PIERB07112201

[12] V.G.Kasabegoudar, K. J. Vinoy: *A Broadband Suspended Microstrip Antenna For Circular Polarization*, Progress In Electromagnetic Research, PIER 90, pp. 353-368, 2009

[13] Sathiyamoorthy Murugan, Elamurugan Satish Kumar, Vayanaperumal Rajamani: *Design and Analysis of Double Uslot Loaded Dual Frequency Microstrip Antenna*, Progress In Electromagnetic Research C, vol. 45, 101-112, 2013
DOI: 10.2528/PIERC13090502

[14] V. G. Kasabegoudar*: Dual Frequency Ring Antennas with Coplanar Capacitive Feed*, Progress In Electromagnetic Research C, vol. 23, 27-39, 2011
DOI: 10.2528/PIERC11060104

[15] *Ansoft High Frequency Structure Simulation (HFSS),* Ver. 13, Ansoft Corporation, 2008

[16] *IE3D simulation software*, Zealand Software Inc., Ver. 14.05, CA, 2008

Acta
Technica
Jaurinensis

Vol. 7., No. 4., pp. 368-377, 2014
DOI: 10.14513/actatechjaur.v7.n4.339
Available online at acta.sze.hu

# Finite Element Simulation by Free Software Tools

**M. Kuczmann, T. Budai**

**Széchenyi István University, Department of Automation**
**Egyetem tér 1, H-9026 Győr, Hungary**
**E-mail: kuczmann@sze.hu**

Abstract: This paper presents a finite element based environment for the simulation of two dimensional electromagnetic field problems, especially electric motors. The graphical user interface, the finite element mesh generator and the computational functions are free software tools that have been developed by scientific teams from all over the world. The aim of this work is to realize a new software environment that can be used in the simulation of two dimensional problems as well as in education.

*Keywords: electromagnetic field, finite element method, parallel computation*

## 1. Introduction

A new software tool has been developed for the simulation of two dimensional electromagnetic field problems; the focus of our research is especially on electric motors. The finite element method [1-5] (FEM) has been applied in the numerical field analysis. Simple one dimensional problems can also be solved by the developed tool.

The main aims of the developing team can be summarized as follows. The software tool must be based on free environments and functions realized by other scientific teams from all over the world, i.e. this FEM code is also available to download and to use. The environment of GMSH [6] is a three-dimensional finite element mesh generator with built-in preprocessing and postprocessing facilities, which is a user-friendly interface for FEM applications. GMSH contains a built-in CAD (Computer Aided Design) interface to build up the geometry of the application to be simulated. The postprocessing scheme can also be realized in the frame of GMSH. The geometry of the problem can be imported from many file formats, i.e. from many other CAD applications. The bridge between the geometry with physics and the postprocessing task (i.e. the numerical field analysis) can be realized by the functions of PETSc [7], which is a portable, extensible toolkit for scientific computation. It consists of many functions, e.g. to assemble and to solve large system of equations, and it is written in the C programming language. The FEM simulation is a time consuming and computer consuming task, realization of parallel algorithms is necessary to speed up the analysis. Our goal is to study the possibilities of parallel algorithms [8-10,29] and domain decomposition techniques [11-15]. The different models of the different materials are very important to include in a new environment, here the modelling of permanent magnets [16] and ferromagnetic hysteresis [17] are the most important goals.

## 2. The Finite Element Method, a short introduction

The main steps of FEM simulations are shown in fig. 1 [1]. The first step is the model specification, i.e. to build up the models of the real life problems which simulation require electromagnetic field calculations (the partial differential equations have to be found, which must be solved with prescribed boundary and continuity conditions; it has to be found out, whether it is a linear or a nonlinear problem and how the characteristics look like). After selecting potentials, the weak formulation of these partial differential equations must be worked out. As it is presented in the introduction, the geometry of the problem must be defined by a CAD software tool. The chosen free environment is GMSH [6].

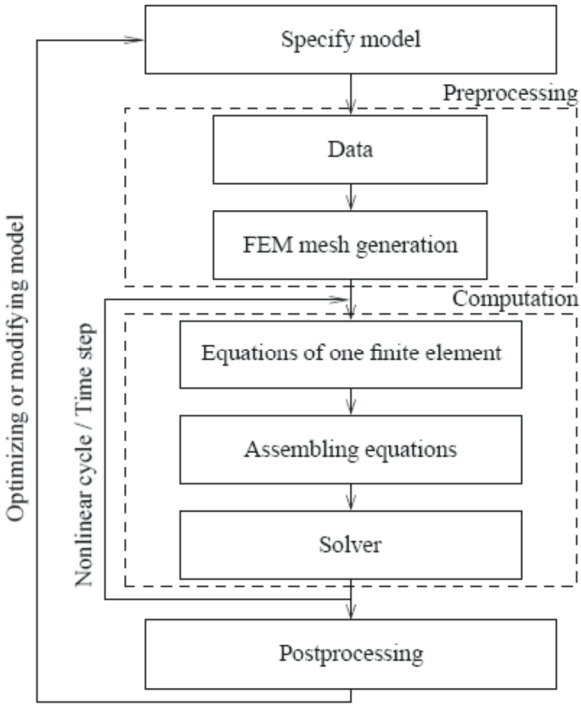



*Figure 1. The main steps of the finite element simulations*

The next step is the preprocessing task. Here the values of different parameters are given, e.g. the material properties, the excitation signal and so on. The geometry can be simplified according to symmetries. The geometry of the problem must be discretized by a FEM mesh [1-5]. The fundamental idea of FEM is to divide the problem region to be analyzed into smaller finite elements with given shape, as triangles or quadrangles in 2D, or tetrahedra, hexahedra or prisms in 3D.

The next step in FEM simulations is solving the problem by the help of PETSc functions [7]. The FEM equations, based on the Galerkin weak formulations, must be set up in the level of one finite element, and then these equations must be assembled

through the FEM mesh [1-5]. Then this global system of equations must be solved [7]. The computation may contain iteration if the constitutive equations are nonlinear. This is the situation when simulating ferromagnetic materials with nonlinear characteristics. Iteration means that the system of equations must be set up and must be solved step by step until convergence is reached. If the problem is time dependent, then the solution must be worked out at every discrete time instant.

The result of computations is the approximated potential value above the mesh. Any electromagnetic field quantity can be calculated by using the potentials at the postprocessing stage. Inductance, energy, force, torque and any other quantities can also be calculated [1-5]. The postprocessing step gives a chance to modify the geometry, the material parameters or the FEM mesh to get more accurate result. The above listed quantities are calculated by functions developed in the frame of PETSc.

## 3. The developed environment

A test environment has been built at the Széchenyi István University to realize the above mentioned technique, it is shown in fig. 2, and it is presented in [18] in more detail. Here a brief presentation is given.

The test-bed system consists of four IBM HS21-8853 blades (nodes) housed in a BladeCenter E Chassis [19]. The free and open-source Linux distribution, Debian Linux 7.4 has been used as operating system on all nodes. To handle the communication between the nodes an open-source implementation of the MPI [20] standard, OpenMPI 1.6.5 [20] has been installed.

The above mentioned PETSc has been used to numerically solve equation-systems according to FEM.

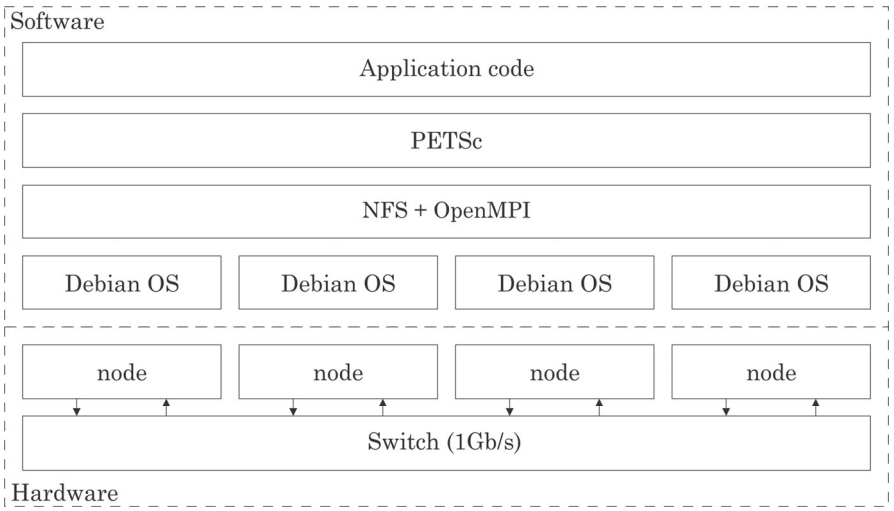



*Figure 2. The completed cluster*

Compilation and execution of parallel C programs using PETSc requires the software packages mentioned above.

In the time of writing the Debian repository [22] contains outdated versions of OpenMPI and PETSc. Another way of installation is to compile these packages from source on the target machine. This consists of installing compiler, setting configuration parameters and starting the compilation process. Since this task is fairly complex and has to be repeated on each node of the cluster and on machines used for development, automation via shell-scripting is strongly advised. To ease the installation on the cluster and to help other members in the research team who wants to develop such software in the future, an automated installer has been developed [18].

The first node has been chosen as master; the other nodes are slaves and have been controlled by the master. All application sources were compiled on and all measurements were controlled by the master machine.

## 4. Illustrative examples

### 4.1. Laminations with ferromagnetic hysteresis

A simple one dimensional problem has been solved by the above mentioned software tool. The schematic view of the problem can be seen in fig. 3, which is a lamination placed into a magnetic field defined by the time varying $\boldsymbol{B}_0$ of one single frequency [23,24]. The thickness $d$ of the lamination is much smaller than the other two dimensions, i.e. a one dimensional model can be set up. The orthogonal components of the electric and the magnetic field are depending only on $x$ as it is shown in the figure.

The following Maxwell's equations must be solved to simulate the lamination core taking eddy currents into account [23,24]: $\nabla \times \boldsymbol{H} = \sigma \boldsymbol{E}$, $\nabla \times \boldsymbol{E} = -\dot{\boldsymbol{B}}$, where $\boldsymbol{H}$, $\boldsymbol{B}$, $\boldsymbol{E}$ and σ are the magnetic field intensity, the magnetic flux density, the electric field intensity, and the conductivity, respectively.

The nonlinear constitutive relationship with static hysteresis is decomposed into a linear term, and a nonlinear residual term, as follows: $\boldsymbol{B} = \mu \boldsymbol{H}_{\mathrm{st}} + \boldsymbol{R}$. Here, $\mu$ is the optimal value of permeability selected as [25,26] $\mu = \frac{2}{\nu_{\max} + \nu_{\min}}$, where $\nu_{\max}$ and $\nu_{\min}$ are the maximum and the minimum slope of the inverse static hysteresis characteristics [27], i.e. the maximum and the minimum reluctivity. The index st in the polarization formulation is for the word static, because the static hysteresis model represents the relationship between $\boldsymbol{B}$ and $\boldsymbol{H}_{\mathrm{st}}$, i.e. $\boldsymbol{H}_{\mathrm{st}} = \mathcal{H}\{\boldsymbol{B}\}$. Excess loss term is ignored by this representation, only the nonlinearity, i.e. hysteresis losses, and eddy current losses are present, $\boldsymbol{H}_{\mathrm{st}} = \boldsymbol{H}_{\mathrm{h}} + \boldsymbol{H}_{\mathrm{eddy}}$. These terms are calculated by the FEM procedure. Decreasing the frequency of excitation, the term $\boldsymbol{H}_{\mathrm{eddy}}$ is decreasing automatically, $\boldsymbol{H}_{\mathrm{h}}$ is the frequency independent term of the magnetic field intensity [24].

The above equations are nonexpansive, and the fixed point iteration scheme through the polarization formulation results in a contraction mapping, meaning convergent iterative process [25,26].

However, excess losses are not present. Excess losses can be represented by the following scheme.

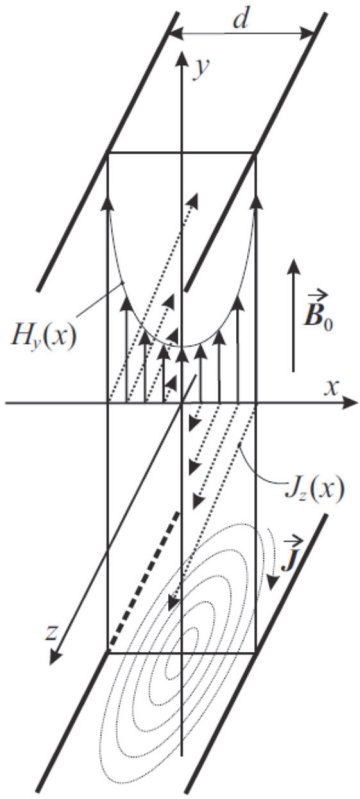



*Figure 3. Electromagnetic field inside laminations excited by external field*

The magnetic field intensity is decomposed into three parts in the more complex model, $\boldsymbol{H} = \boldsymbol{H}_{\mathrm{h}} + \boldsymbol{H}_{\mathrm{eddy}} + \boldsymbol{H}_{\mathrm{exc}}$. The last term is responsible for the anomalous or excess losses [23,24,28]. The advantages and convergence properties of the fixed point technique can be hold by introducing the excess field term $\boldsymbol{H}_{\mathrm{exc}}$ in the Maxwell's equations as follows [24]: $\nabla \times \boldsymbol{H} = \sigma \boldsymbol{E}$, $\nabla \times \boldsymbol{E} = -\dot{\boldsymbol{B}}$, $\boldsymbol{B} = \mu(\boldsymbol{H} - \boldsymbol{H}_{\mathrm{exc}}) + \boldsymbol{R}$, i.e. the decomposition of the nonlinear constitutive relationship must be rewritten. The following nonlinear partial differential equation can be obtained:

$$\nabla \times \nu \nabla \times \boldsymbol{H} + \mu\sigma\dot{\boldsymbol{H}} = \mu\sigma\dot{\boldsymbol{H}}_{\mathrm{exc}} - \sigma\dot{\boldsymbol{R}}. \qquad (1)$$

This problem can be solved numerically by the finite element method. It is noted that the excess field term is present on the right hand side of (1), and the solution of the problem is the total magnetic field intensity, containing the excess field term, too.

The three terms can be seen separately in fig. 4 supplying sinusoidal external magnetic field with three different amplitude of the magnetic flux. A comparison between measured and simulated higher order dynamic minor loops can be seen in fig. 5.

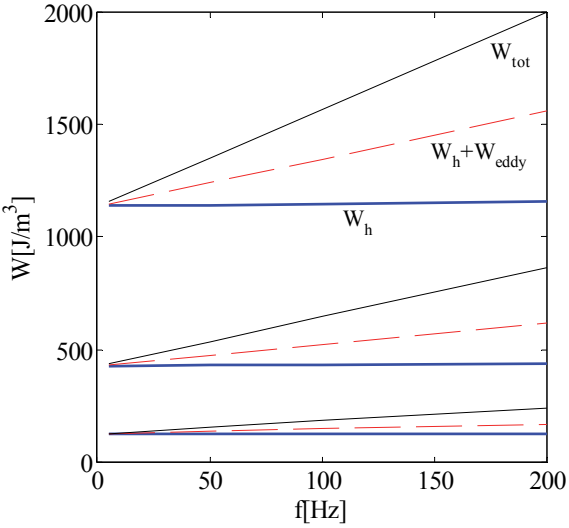



*Figure 4. Losses inside the lamination*

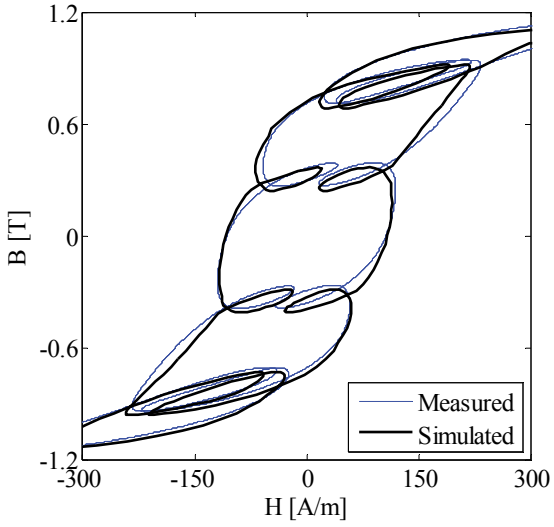



*Figure 5. Comparison between measured and simulated higher order dynamic minor loops*

### 4.2. Three phase transformer core

The model of a three phase test transformer has been built to simulate the complex behaviour of the transformer core. The geometry of the transformer can be seen in fig. 6, and the simulated locus of the magnetic field intensity vector around the two signed areas has been plotted in fig. 7.

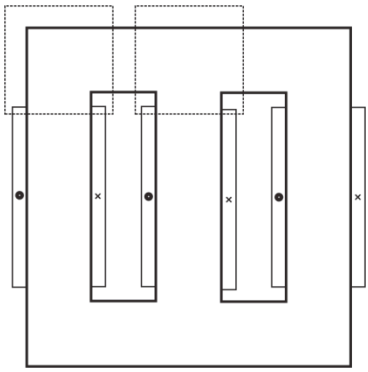


*Figure 6. Three phase transformer core*

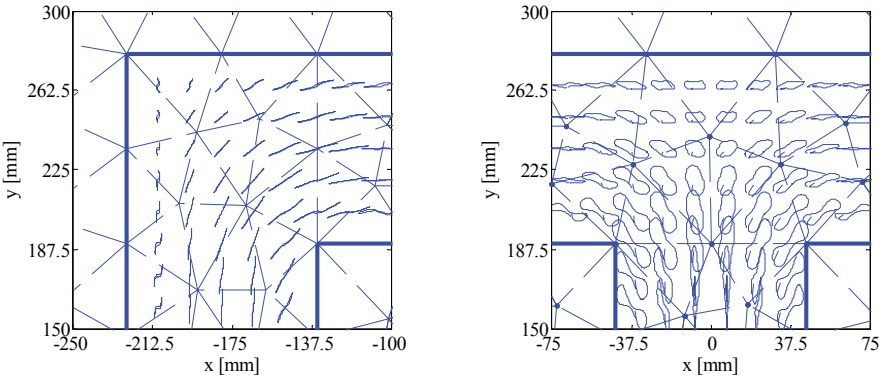



*Figure 7. Locus of magnetic field simulated by the vector Preisach model of hysteresis*

Here, the so-called vector Preisach model of hysteresis [17,24,27] has been implemented in the FEM code combining it with the fixed-point scheme [25,26].

### 4.3. Capacitor

The electric field of a capacitor has been simulated by the above mentioned environment [18]. The linear problem, as a simple and first case study, has been solved as a simple one dimensional, a more difficult two dimensional, and finally, as a three dimensional arrangement to increase the number of unknowns to test the speedup of the tool.

Looking at the achieved speedup for each solver, it is clear that problems which require complex computation steps in the assembly phase can benefit the most from parallel processing.

The deviation of the individual results shows that system-level functions, like caching has no significant impact on the results. Fig. 8. shows the execution times of the sequential and best parallel solvers compared [18].

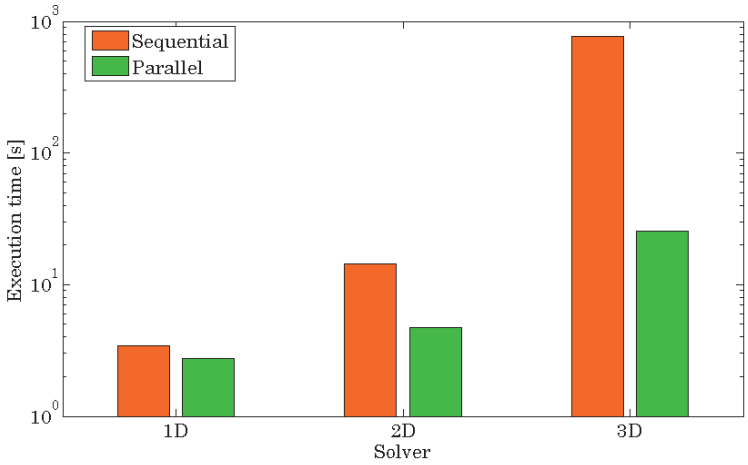



*Figure 8. Execution times of the different solvers*

## 5. Conclusion

The developed code has just started to simulate permanent magnet synchronous machines applied in hybrid vehicles.

The next step of the work is to implement the model of permanent magnets in the code, and to give a graphical user interface to handle hysteresis. Modelling of motor geometry is also under construction, and the insertion of the code into control algorithms and identification tasks is also an open question.

## Acknowledgement

This paper is sponsored by ”TÁMOP-4.2.2.A-11/1/KONV-2012-0012: Basic research for the development of hybrid and electric vehicles -The Project is supported by the Hungarian Government and co-financed by the European Social Fund”.

## References

[1] Kuczmann, M., Iványi, A.: *The Finite Element Method in Magnetics*, Akadémiai Kiadó, Budapest, 2008
DOI: 10.13140/2.1.3104.1927

[2] Bíró, O.: *CAD in Electromagnetism*, Advances in Electronics and Electron Physics, Vol. 82, pp. 1–96, 1991
DOI: 10.1016/S0065-2539(08)60911-7

[3] Bastos, J.P.A., Sadowski, N.: *Electromagnetic Modeling by Finite Element Methods*, Marcel Dekker Inc., New York, 2003

[4] Jin, J.: *The Finite Element Method in Electromagnetics*, John Wiley and Sons, New York, 2002

[5] Meunier, G.: *The Finite Element Method for Electromagnetic Modeling*, John Wiley and Sons, New York, 2008

[6] Geuzaine, C., Remacle, J.F.: *Gmsh: a Three Dimensional Finite Element Mesh Generator with Built-in Pre and Postprocessing Facilities*, International Journal for Numerical Methods in Engineering, Vol. 79, No. 11, pp. 1309–1331, 2009 DOI: 10.1002/nme.2579
[7] PETSc: http://www.mcs.anl.gov/petsc/ (last visited: 2014. 09. 14.)
[8] METIS, http://glaros.dtc.umn.edu/gkhome/views/metis (last visited: 2014. 09. 14)
[9] Molnárka, G., Varjasi, N.: *A Simultaneous Solution for General Linear Equations on a Ring or Hierarchical Cluster*, Acta Technica Jaurinensis, Vol. 3, No. 1, pp. 65-73, 2010
[10] Toselli, A., Vasseur, X.: *Robust and Efficient FETI Domain Decomposition Algorithms for Edge Element Approximations*, COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, Vol. 24, No. 2, pp. 396-407, 2005 DOI: 10.1108/03321640510586033
[11] Kruis, J.: *Domain Decomposition Methods for Distributed Computing*, Saxe-Coburg Publications, Kippen, Stirling, 2006
[12] Nikishkov, G.P.: *Basics of the Domain Decomposition Method for Finite Element Analysis*, in Mesh Partitioning Techniques and Domain Decomposition Methods, Editor: Magoulés, F., Saxe-Coburg Publications, Kippen, Stirling, pp. 119-142, 2007
[13] Farhat, C., Pierson, K., Lesoinne, M.: *The Second Generation FETI Methods and Their Application to the Parallel Solution of Large Scale Linear and Geometrically Nonlinear Structural Analysis Problems*, Computer Methods in Applied Mechanics and Engineering, Vol. 184, No. 2-4, pp. 333–374, 2000 DOI: 10.1016/S0045-7825(99)00234-0
[14] Marcsa, D.: *Domain Decomposition Algorithms for Edge Element Based Parabolic Type Problems*, Acta Technica Jaurinensis, Vol. 7, No. 2, pp. 193-206, 2014 DOI: 10.14513/actatechjaur.v7.n2.278
[15] Farhat, C., Roux, F.X.: *Method of Finite Element Tearing and Interconnecting and its Parallel Solution Algorithm*, International Journal for Numerical Methods in Engineering, Vol. 32, No. 6, pp. 1205-1227, 1991 DOI: 10.1002/nme.1620320604
[16] Kovács, G.: *Up-to-Date Finite Element Based Simulation for Permanent Magnet*, Acta Technica Jaurinensis, Vol. 7, No. 2, pp. 172-182, 2014 DOI: 10.14513/actatechjaur.v7.n2.281
[17] Bertotti, G., Mayergoyz, I.D.: *The Science of Hysteresis*, Academic Press, New York, 2006
[18] Budai, T., Kuczmann, M.: *Parallel Implementation of Finite Element Solvers by the help of MPI*, MSc Conference at the Budapest Univeristy of Technology and Economics, Budapest, pp. 39-42, 2014
[19] IBM BladeCenter E Chassis, www.ibm.com/systems/bladecenter/hardware/chassis/bladee/ (last visited: 2014. 09.14.)
[20] Message Passing Interface Forum, http://www.mpi-forum.org/ (last visited: 2014. 09. 14.)
[21] Open MPI, http://www.open-mpi.org/ (last visited: 2014. 09. 14.)

[22] Debian Packages Repository, https://packages.debian.org/stable/ (last visited: 2014. 09. 14.)
[23] Dlala, E., Belahcen, A., Fonteyn, K.A., Belkasim, M.: *Improving Loss Properties of the Mayergoyz Vector Hysteresis Model*, IEEE Transactions on Magnetism, Vol. 46, No. 3, pp. 918-924, 2010
DOI: 10.1109/TMAG.2009.2034846
[24] Kuczmann, M., Kovács, G.: *Improvement and Application of the Viscous-Type Frequency-Dependent Preisach Model*, IEEE Transactions on Magnetism, Vol. 50, No. 2, Paper No. 7009404, 2014
DOI: 10.1109/TMAG.2013.2283398
[25] Bottauscio, O., Canova, A., Chiampi, M., Repetto, M.: *Iron Losses in Electrical Machines: Influence of Different Material Models*, IEEE Transactions on Magnetism, Vol. 38, No. 2, pp. 805-808, 2002
DOI: 10.1109/20.996208
[26] Hantila, F.I., Preda, G., Vasiliu, M.: *Polarization Method for Static Fields*, IEEE Transactions on Magnetism, Vol. 36, No. 4, pp. 672-675, 2000
DOI: 10.1109/20.877538
[27] Iványi, A.: *Hysteresis Models in Electromagnetic Computation*, Akadémiai Kiadó, Budapest, 1997
[28] Bertotti, G.: *General Properties of Power Losses in Soft Ferromagnetic Materials*, IEEE Transactions on Magnetism, Vol. 24, No. 1, pp. 621-630, 1988
DOI: 10.1109/20.43994
[29] Iványi, P., Radó J.: *Preprocessing to Parallel Computation*, Budapest: Typotex Kiadó, Budapest, 2014
ISBN: 978-963-279-381-8

Acta
Technica
Jaurinensis

Vol. 7., No. 4., pp. 378–387, 2014
DOI: 10.14513/actatechjaur.v7.n4.342
Available online at acta.sze.hu

# Population and Gradient Based Optimization Techniques, a Theoretical Overview

**G. Friedl, M. Kuczmann**

**Széchenyi István University, Department of Automation**
**9026, Győr, Hungary**
**E-mail: friedl@maxwell.sze.hu**

Abstract: Many practical problems can be modeled only as a nonlinear continuous global optimization problem. It is usually impossible and impractical to solve them exactly. Evolutionary and hybrid algorithms are modern techniques to find optima for complex search spaces. This paper is a short summary about the optimization techniques, especially about the evolutionary based global searching algorithms, and the gradient based local searching algorithms. After the introduction, the second chapter gives a brief summary about population based techniques, especially about genetic and bacterial evolutionary algorithm, and particle swarm optimization. The next chapter discusses two gradient based searching technique, the *steepest descent* and the *Levenberg-Marquardt* method. This paper is a theoretical overview of the basics of my future Ph.D. dissertation.

*Keywords:* *optimization, soft computing, evolutionary algorithms, memetic algorithms*

## 1. Introduction

Soft computing techniques are characterised to provide an inexact solution to computationally hard problems, for which there is no known algorithm, what could calculate the exact solution, and the computation time is non-deterministic. Evolutionary algorithm is a branch of them, and represents many type of global searching techniques. This optimization technique family is named *evolutionary*, because the ideas behind them are drawn from the biological evolution. It is a population (or multi-population) based metaheuristic optimization algorithm. For a problem in analytically indescribable search field with limited computation time, metaheuristics provide a suitable result, a so-called quasi-optimum.

These methods are useful to find the area of global maximum (or minimum) of the search field, but the convergence becomes slower, when the task is to find the local maximum (or minimum) of this area. There are many searching algorithms, which are developed to find

these points on an unknown function as fast, as it is possible. These methods are called gradient based methods, e.g. the steepest descent method, or the Levenberg-Marquardt method. These methods are very useful to find local extreme values within a reasonable time, but if they founded this point, they can not find an another, maybe better point in the search field, these algorithms get stuck at the local extreme values.

To take the advantage of these two types of searching algorithms, hybrid, so-called memetic algorithms were developed. These algorithms use population based algorithms to find a point next to the global maximum or minimum, then apply the gradient based algorithms to achieve fast convergence.

## 2. Population based searching algorithms

Population based algorithms, like genetic algorithm [1], bacterial evolutionary algorithm [2, 3] and particle swarm optimization [4] are developed to find global quasi-optimum of any complex search field. They are all inspired by a natural phenomena. These algorithms operate with a population of possible solutions of the search field. These solutions are called *individuals*. The population is then arranged in pairs as parents, the children or the mutation of them may be a better solution than the parents. The goodness of these individuals are described by their fitness function. The fitness function is an objective function, that summarises the properties of an individual in a single number. There is no unified description of the fitness function, because its definition is hardly depends on the type of the task.

### 2.1. Genetic algorithm

The first evolutionary type, and one of the most applied method is the genetic algorithm [1]. This technique imitates the process of natural selection by modeling the crossover of individuals and randomly occurring mutations. Every individual has a chromosome, that represents a set of properties. A property can be a bit, a numeric value, a color, etc. During the evolution, these chromosomes are continuously changing to create better and better individuals. The genetic algorithm has five main steps [5, 6]. These are the following:

**Initialization:** Creating an initial population of randomly chosen individuals in the search field. This step has to be done only once at the beginning of the algorithm.

**Selection:** Sorting the individuals according to their fitness value, then selecting some of them. The chance of the selection is proportional to the fitness value. The selected ones are called as *parents*.

**Crossover:** Arranging the selected part of the population in pairs. For each pair a random point of their chromosome is selected, and the parents change every properties between each other after this point (see in Fig. 1). The newly created individuals

are called *offspring*. There are modified algorithms, that work with two or more point crossover, e.g. in the case of two point crossover the changed part of the chromosomes between these two point.

**Mutation:** Changing a random property of the chromosomes of each offspring with small probability (see in Fig. 2). The new, mutated value of this point is chosen randomly.

**Substitution:** Substituting some selected members of the population by the offspring. The chance of the selection is inversely proportional to their fitness value. The *strong* individuals survive with a bigger chance. If there is no acceptably good individual, or the iteration/time limit is not reached, the algorithm steps back to the *selection* step.

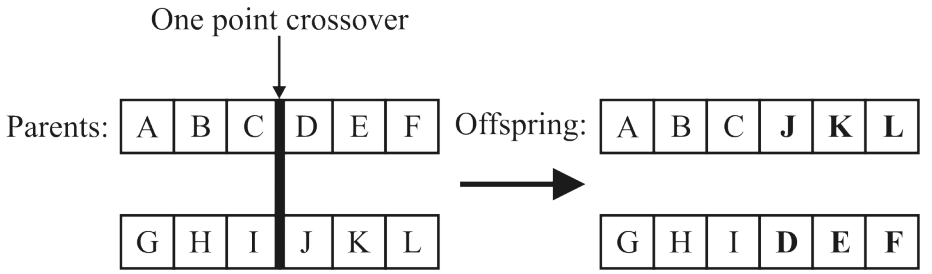



*Figure 1. One point crossover.*

There are different methods to execute the selection step. The two most widely used methods are the *roulette wheel technique* and the *stochastic universal sampling* [6]. If the so-called *elitist strategy* is implemented, the best individual always survives.

The genetic algorithm finds the area of global optimum very effectively. On the other side, the algorithm is very time consuming, the calculation of the fitness function can be very difficult and the convergence of the algorithm gets slower.

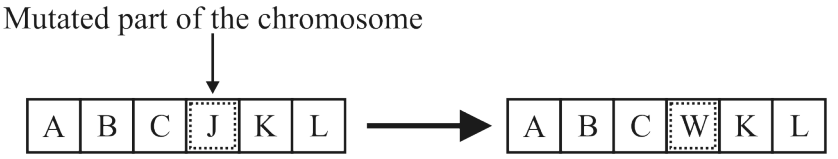



*Figure 2. Mutation of a single property of the chromosome of an individual.*

### 2.2. Bacterial evolutionary algorithm

Another widely used evolutionary type searching method is the bacterial evolutionary algorithm[2, 3], which is based on the microbial evolution. There are two main difference between the bacterial and genetic algorithm:

1. The *crossover* and *selection* step is substituted by a new operator called gene transfer.
2. The *mutation* step is modified as bacterial mutation.

The implementation of this method is easier than the genetic algorithm, but usually it is also more efficient. The bacterial evolution algorithm has three main steps [5, 7]. These are the following:

**Initialization:** Creating an initial population of randomly chosen individuals in the search field. This step has to be done only once at the beginning of the algorithm.

**Bacterial mutation:** Mutating every property of every bacteria in a random order even multiple times. If the original bacterium is a better solution to the problem, then it is restored, otherwise the new individual substitutes the original (see in Fig. 3).

**Gene transfer:** Separating the population into two groups according to the fitness values (superior and inferior group), arranging the members in pairs (one from the superior and one from the inferior group), then transferring genes from the superior individual to the inferior individual.

The usage bacterial evolutionary algorithm is also an effective way to find quasi-optimum. This method usually faster, than the genetic algorithm, but the convergence also gets slower, when the actual best result is close to the global optimum. This slowing is caused by the randomness of the evolution.

### 2.3. Particle swarm optimization

The idea behind the particle swarm optimization method is the social behavior of bird flocking or fish schooling [4]. The individuals are called as *particles*, and the population is called as a *swarm*. The particles placed randomly on the search field are continuously moving. The direction of moving of each particle depends on their actual position, the location of their *local best value*, and the location of the *global best value*. Every particle has its own local best place, which is the place where its fitness value was the best during the moving. The global best place is the place of the best local best place. These particles can be ordered in different topologies. These topologies describe, which other particles will affect the moving direction.

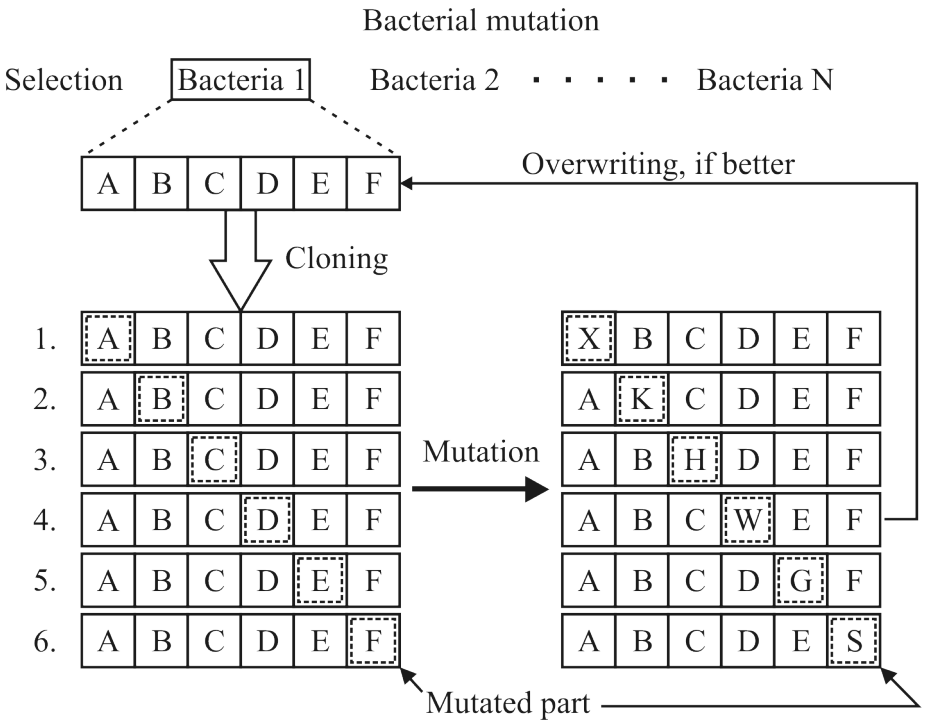



*Figure 3. A step of the bacterial mutation*

In the simplest case, the particle swarm optimization method has three main step. These are the following[5, 8]:

**Initialization:** Creating an initial swarm (or multi-swarm) placing particles in random positions.

**Moving:** The particles are moving in directions affected by their local best values $\boldsymbol{l}$ and the global best value $\boldsymbol{g}$. If $i$ is the iteration number, and the values of the set of parameters at actual location of a particle is denoted as $\boldsymbol{x}_i$, the direction of the moving, and the next place of the particle can be determined by the following equation:

$$\begin{aligned} \boldsymbol{d}_0 &= 0, \\ \boldsymbol{d}_{i+1} &= \varphi_d \boldsymbol{d}_i + \varphi_l r_l (\boldsymbol{l}_i - \boldsymbol{x}_i) + \varphi_g r_g (\boldsymbol{g}_i - \boldsymbol{x}_i), \\ \boldsymbol{x}_{i+1} &= \boldsymbol{x}_i + \boldsymbol{d}_{i+1}, \end{aligned} \tag{1}$$

where $\varphi_d$, $\varphi_l$ and $\varphi_g$ are parameters, $r_l$ and $r_g$ are random values. An explanation for this step can be seen in Fig. 4.

**Update:** Calculating the fitness value of each particle, and if the actual value is better, than at its local best place, the actual place is set as the new local best place. If there is any particle, that is in a better position, than the actual global best place, the global best place is set as the place of the actual best particle.

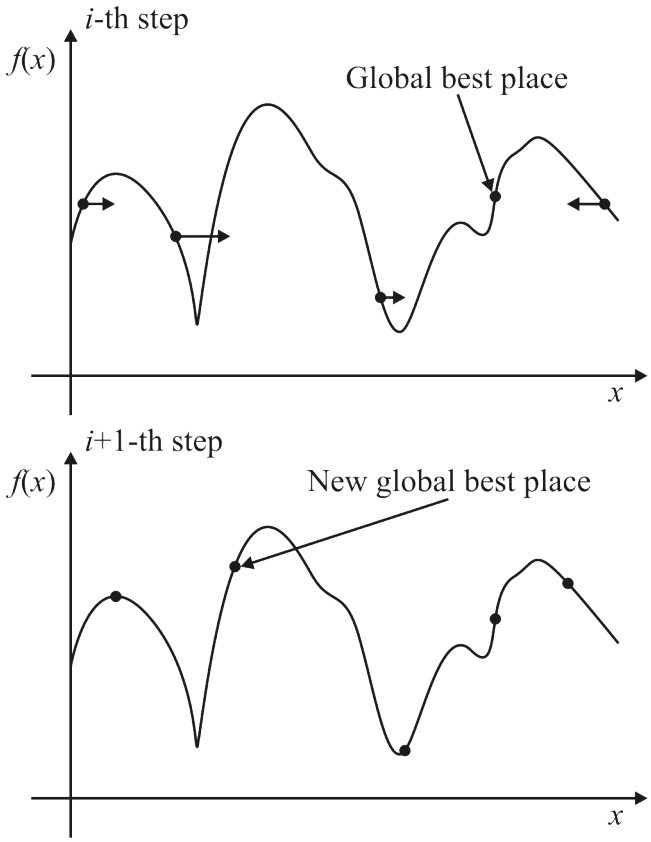



*Figure 4. Moving of the particles to the direction of the global best place.*

The iteration above runs, until a particle is in a better position, than the required criteria, or until the time limit or generation limit exceeded.

### 2.4. Advantages and disadvantages

The population based algorithms can be widely applied, when the complexity of the search field prevents of the usage of other analytical or numerical techniques. They can provide for the most task a quasi-optimal solution. A moderate optimal solution can be find relatively

fast. The algorithms can be easily implemented at low cost. The algorithms are iterative, and the model parameters can be changed adaptively between two iteration.

These algorithms have also some disadvantages as well. The global best solution is not guaranteed in a finite time, and the convergence speed is continuously slowing, because of the randomness of the evolution. In some cases, the model parameters can be only modified by trying.

## 3. Gradient based searching algorithms

Gradient based algorithms, like steepest descent/ascent method [7] and Levenberg-Marquardt method [9, 10] are developed to find the nearest local optimum. These algorithms work with calculating the gradient of the search field at a point in question. The gradient can be calculated analytically from the gradient vector function, otherwise a pseudo-gradient is needed. The pseudo-gradient can be calculated experimentally by producing the derivative of the search field with very small steps. The searching is moving towards based on the gradient, where the optimization needs (to the maximum or minimum).

### 3.1. Steepest descent/ascent method

The steepest descend (or also mentioned as gradient descend) [7, 11, 12] method is a first order local minimum searching technique. The direction of the search is described by the negative gradient of the search field at the current point. If the interest is in finding the local maximum, the step is proportional to the positive of the gradient. In this case, this method is called as steepest ascent method. The size and the direction of the step is the gradient vector multiplied by a so-called bravery factor. The bigger the $\gamma$ bravery factor is, the bigger the chance is to missing/jumping over a local minimum. If the search field is denoted as $F$, and the values of the set of variables in the actual place is denoted as $\boldsymbol{x}_i$, the next place of the searching $\boldsymbol{x}_{i+1}$ from the actual place $x_i$ can be calculated as:

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \gamma \nabla F(\boldsymbol{x}_i). \tag{2}$$

This searching procedure on a single random two dimensional $F$ function can be seen on Figure 5.

This optimization technique is very sensitive to the starting position, because it can stuck at every local minimum, instead of finding the real minimum position of the search field.

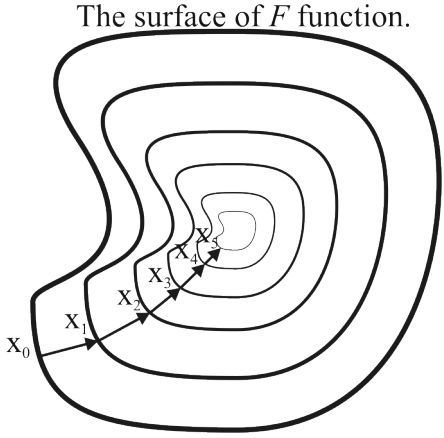



*Figure 5. Minimum searching with steepest descent method.*

### 3.2. Levenberg-Marquardt method

The Levenberg-Marquardt method is developed to find the minimum of a multivariate function $f(\boldsymbol{x})$ that is expressed as the sum of squared errors

$$f(\boldsymbol{x}) = \frac{1}{2}\sum_{j=1}^{m} \boldsymbol{r}_j^2(\boldsymbol{x}), \tag{3}$$

where $\boldsymbol{r}_j(\boldsymbol{x})$ is a residual function for $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ parameter set. The method uses the *Jacobian-matrix* to minimize the error function. Each row of the Jacobian is the gradient vector of the residual functions corresponding to $\boldsymbol{x}$, and can be written as:

$$\mathbf{J} = \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial r_1}{\partial x_1} & \cdots & \frac{\partial r_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_1}{\partial x_n} & \cdots & \frac{\partial r_m}{\partial x_n} \end{pmatrix}. \tag{4}$$

If the Jacobian can not be calculated calculated analytically, because the formula of the gradient vector function is unknown, a pseudo-Jacobian can be calculated experimentally by determining each pseudo-gradient vector function. The basic of the Levenberg-Marquardt method is the linear approximation of the error function $f$ in the neighborhood of $\boldsymbol{x}$. The core equation of parameter changing between two iteration steps of the method is the following:

$$\Delta(\boldsymbol{x}) = -\left(\mathbf{J}\mathbf{J}^T + \lambda \mathbf{I}\right)^{-1} \mathbf{J}^T \boldsymbol{r}. \tag{5}$$

In (5) $\lambda$ is a damping parameter, which can be adaptively changed. Using large values for $\lambda$ parameter results in the gradient descent method, and small values for $\lambda$ parameter results in the *Gauss-Newton* method.

The Levenberg-Marquardt method is the most widely used optimization technique. In the most case, this method has the best convergence properties, but if the analytical formula of the gradient vector functions are unknown, the calculation of the Jacobian can be very time consuming.

### 3.3. Advantages and disadvantages

The rapid convergence is the main advantage of the gradient based methods. This methods find the nearest local minimum/maximum effectively with adaptively variable model parameters.

These methods are not sufficient finding global optima, the effectiveness of them are strongly depends on where the searching algorithm starts. Because of the small steps, they can easily stuck at a single local maximum/minimum value, and can only explore the neighborhood of the starting position.

## 4. Memetic algorithms

The population based algorithms are very effective finding the area of global optima. They provide quasi-optimal solutions within acceptable time, but the speed of the convergence becomes very slow because of the randomness of the evolution, finding the exact global optima is hopeless within finite time. The gradient based methods causing minor modifications of the input parameters, so they can provide fast and accurate result of the place of local optimum, but as it was mentioned before, these algorithms are very sensitive of the starting point, and can easily stuck at a local optima.

Memetic algorithms are hybrid algorithms developed to avoid the disadvantages above by applying them combined. This can be done easily, e.g. applying local searching techniques when the fitness value of the individuals reached a critical value, or using local searches between each generation of the population based algorithms. There are many ways to combine these algorithms, e.g. the combination of genetic algorithm and the steepest descent method is referred as *genetic steepest descent*, the combination of the bacterial evolutionary algorithm and the Levenberg-Marquardt method results in *bacterial memetic algorithm*, etc. The family of memetic algorithms are often referred as *Baldwinian* or *Lamarckian* evolutionary algorithms.

## 5. Conclusion

In this review paper, I summarized the most widely used population and gradient based searching algorithms. These methods will form the theoretical basics of my future Ph.D. dissertation. The main goal of my work is developing evolutionary and hybrid algorithms to design antennas for strictly specified radiation pattern.

## References

[1] Holland, J. H.: *Adaption in Natural and Artificial Systems*, The MIT Press, Massachusetts, 1992

[2] Nawa, N. E., Hashiyama, T., Furuhashi, T., Uchikawa, Y.: *Fuzzy Logic Controllers Generated by Pseudo-Bacterial Genetic Algorithm*, In Proceedings of the IEEE International Conference on Neural Networks, vol. 4, pp. 2408–2413, Houston, 1997
DOI: 10.1109/ICNN.1997.614446

[3] Nawa, N. E., Furuhashi T.: *Fuzzy system parameters discovery by bacterial evolutionary algorithm*, IEEE Transactions on Fuzzy Systems, vol. 7, no. 5, pp. 608–616, 1999
DOI: 10.1109/91.797983

[4] Kennedy, J., Eberhart R.: *Particle swarm optimization* In Proceedings of the IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948, Perth, 1995
DOI: 10.1109/ICNN.1995.488968

[5] Balázs, K.: *Advanced Approaches in the Application Methodologies of Evolutionary Algorithms*. Ph.D. dissertation, Budapest, Hungary, 2013

[6] Sumathi, S., Hamsapriya, T., Surekha P.: *Evolutionary Intelligence - An Introduction to Theory and Applications with Matlab*, Springer, India, 2008

[7] Gál, L.: *Optimalization of Fuzzy Models by Bacterial Type Algorithms (in Hungarian)*, Ph.D. dissertation, Győr, Hungary, 2012

[8] del Valle, Y., Venayagamoorthy, G. K., Mohagheghi, S., Hernandez, J.-C., Harley, R. G.: *Particle Swarm Optimization: Basic Concepts, Variants and Applications in Power Systems*, IEEE Transactions on Evolutionary Computation, vol. 12, no. 2, pp. 171-195, 2008
DOI: 10.1109/TEVC.2007.896686

[9] Levenberg, K.: *A method for the solution of certain non-linear problems in least squares*, Quarterly Journal of Applied Mathematics, vol. 2, no. 2, pp. 164–168, 1944

[10] Marquardt, D.: *An Algorithm for Least-Squares Estimation of Nonlinear Parameters*, SIAM J. Appl. Math., vol. 11, no. 2, pp. 431-441, 1963

[11] Yamada, I.: *The Hybrid Steepest Descent Method for the Variational Inequality Problem over the Intersection of Fixed Points Sets of Nonexpansive Mapping*, Studies in Computational Mathemathics, vol. 8, pp. 473–504, 2001
DOI: 10.1016/S1570-579X(01)80028-8

[12] Snyman, J.: *Practical Mathematical Optimization*, Springer, New York, 2005

Acta
Technica
Jaurinensis

Vol. 7., No. 4., pp. 388-394, 2014
DOI: 10.14513/actatechjaur.v7.n4.344
Available online at acta.sze.hu

# Roller Bench Design for Measurement of Effective Engine Power

**I. Lakatos, P. Dely**

**Széchenyi István University, Department of Road and Rail Vehicle**
**Egyetem tér 1, H-9026, Győr, Hungary**
**e-mail: lakatos@sze.hu**

Abstract: Effective performance of road vehicle engines represents an important segment of vehicle specification. A specific measurement method is required for acquisition and assessment of data coming from diagnostic tools attached to the rolling road. A new method has been elaborated by the first author. This method can be used as a measurement system to measure effective engine power for operational diagnostic purposes. Free rollers are needed instead of loading machine, according to the new approach. This article demonstrates the design principles for the roller bench needed.

*Keywords: drive train, wheel performance, effective performance, free acceleration*

## 1. Introduction

The principle of the measurement method is to accelerate and decelerate the unloaded drive train of the studied vehicle on free rollers (there is no need for a rolling road). Since the decision to measure external characteristics, the measurement must be performed under total load conditions [1, 2, 3, 4, 5, 10].

The kinetic energy change of the system is displayed in the acceleration of the wheel and the rollers, so this element equals to the wheel performance ($P_k$). Though the diverted heat equals to the running loss performance ($P_v$):

$$P_e = P_k + P_v \tag{1}$$

The basic dynamic equation of rotation can be described both for acceleration and deceleration phases:

$$P = \mathrm{M}\omega = (\theta_{red}\varepsilon)\omega = \theta_{red}\frac{d^2\varphi}{dt^2}\frac{d\varphi}{dt} \tag{2}$$

where:

- $\omega$ is the angular velocity of the roller of the rolling road;
- $\Theta_{red}$ is moment of inertia of the drivetrain of the vehicle reduced to the axis of the roller of the rolling road;
- $\varphi$ is angle of rotation of the rollers;
- $\varepsilon$ is angular acceleration of the rollers;
- $t$ is time.

## 2. Theoretical background of the new measurement method

The phases of the measurement are followings:

- Acceleration phase: the drive train of the vehicle on the bench and the rollers of the rolling road are accelerated in the studied gear, up to the rated engine speed with full load (on full blast).
- Deceleration phase: by releasing the clutch, leaving the gear at the given position, we let the car decelerate until it stops.

During the measurement, as there is no external load, the engine has to accelerate the moments of inertia indicated. During deceleration phase the moment of inertia of the engine is separated, so with this exemption the rest of the moments of inertia decelerate the system.

During the test, the angular velocity and the angular acceleration values of the roller are recorded exclusively, see equation (2). In order to calculate power, the moment of inertia of the drivetrain reduced to the roller axis is needed.

According to the aforementioned measurements the diagram displayed in Fig. 1. can be recorded.

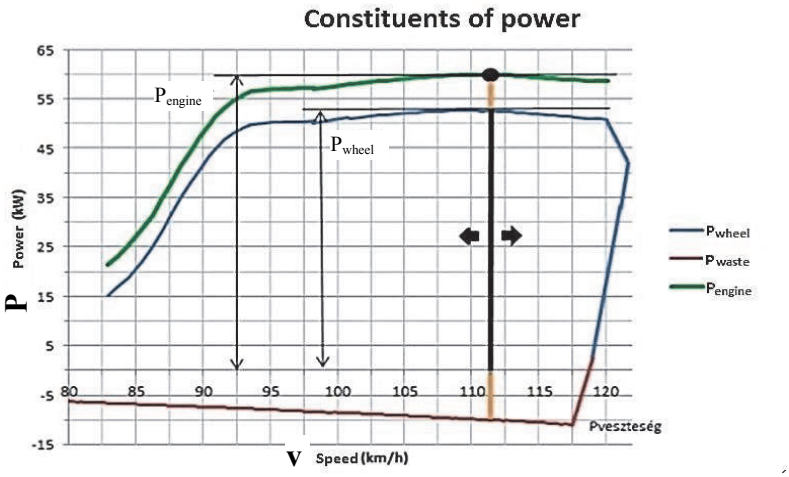



*Figure 1. Characteristic curve recorded during the measurement*

As can be seen in Fig. 1., the effective power of the engine is calculated as the sum of the wheel power and the waste power. Because the moment of inertia of the drivetrain is unknown, the power constituent values have to be determined by measurement.

The new measurement technology solves the above problem by means of an additive mass. Thus, the necessity of a loading machine in the roller bed can be eliminated and the price of the roller bench can be reduced significantly.

## 3. Validating the external characteristic curve of the engine without rolling road test (new measurement method)

The main points of the measurement method which has been elaborated are as follows:

In case of not possessing a rolling road just a roller bed two free acceleration measurements have to be conducted.

Therefore, an additive mass is needed to perform the measurement. The layout of the Schenck W280 roller bench with rotating mass in the department laboratory is shown in Fig. 2.

In this case, the additive mass is not detachable, consequently, it needs to be modified in order to adapt the machine to the new measurement. After the modification, diagrams in Fig 3. can be produced.

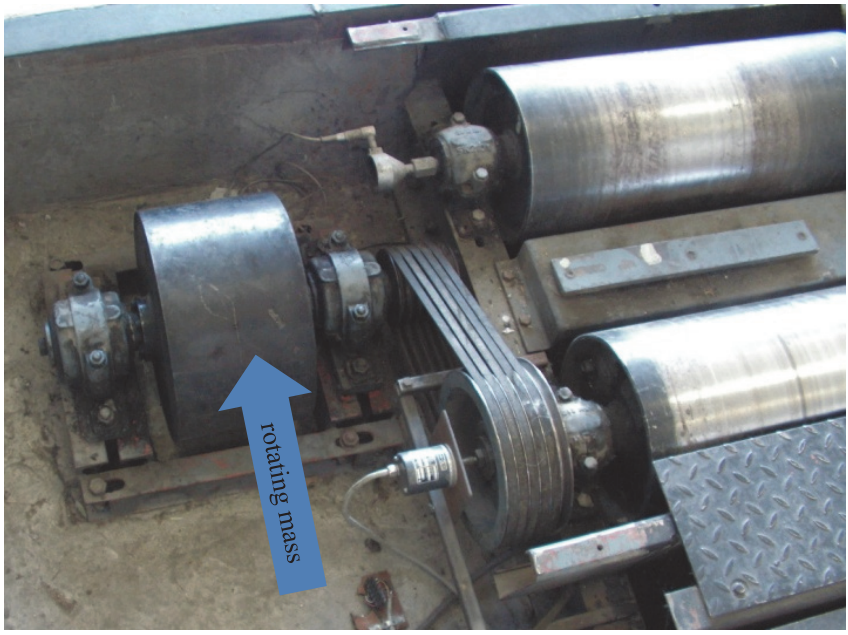



*Figure 2. Roller bench with the rotating mass*

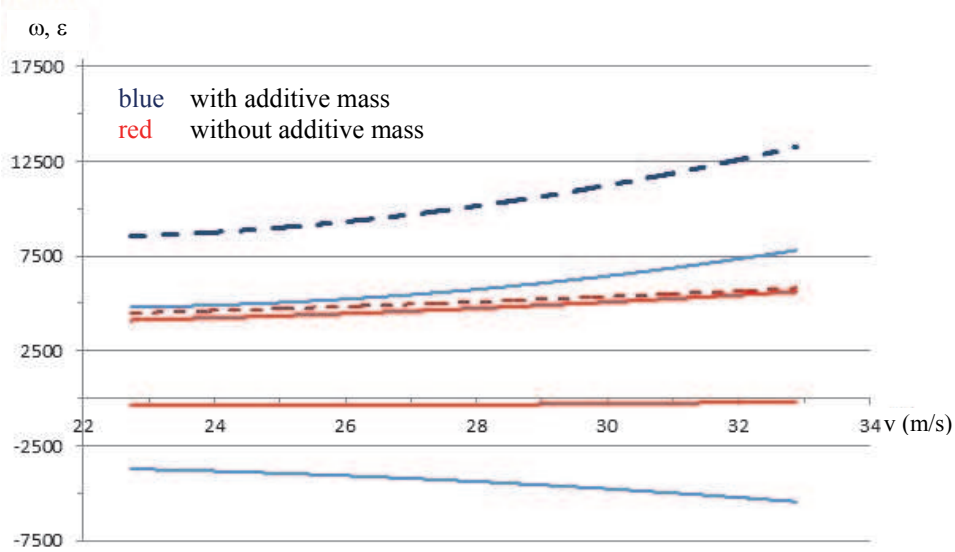



*Figure 3. Diagrams of two subsequent measurements*

## 4. Rundown experiments with additive mass

The engine is running at a given speed (i.e. in the given case is the 2/3 of the nominal engine speed value) and then the fuel supply is stopped. The speed of the engine gradually decreases and it comes to a halt.

The measurement was performed in one given gearshift lever position, with additive mass in the first run and without mass in the second. Both measurements produced a graph of free rundown. Considering the Fig. 4. two graphs, it is now possible to calculate the moment of inertia of the drivetrain.

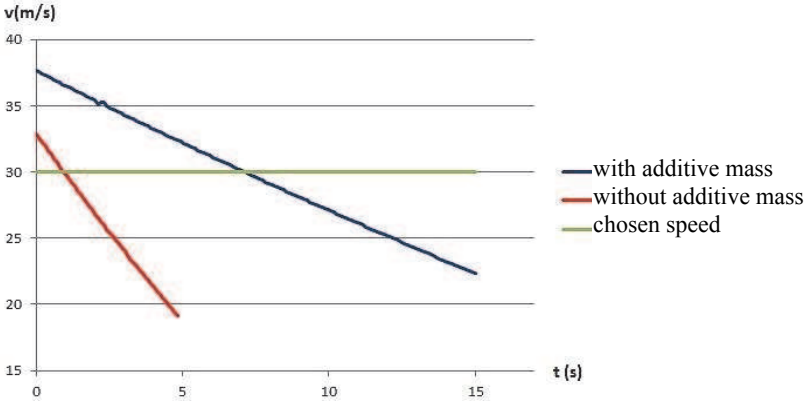



*Figure 4. Rundown graphs*

It can easily be observed that the rundown time of the system is significantly longer when the additive mass is coupled to the rollers (Fig. 5.).

## 5. Modification of the SCHENCK W 280 type roller bench into detachable-additive-mass type

### 5.1. Selection of electromagnetic clutch

When using the aforementioned new measurement method it is imperative that one should be able to couple or decouple the additive mass by means of a clutch easily. The optimum solution for the above purpose is offered by an electromagnetically controlled clutch. The transfer of the required torque depends on the moment of inertia of the flywheel and on the maximum acceleration value.

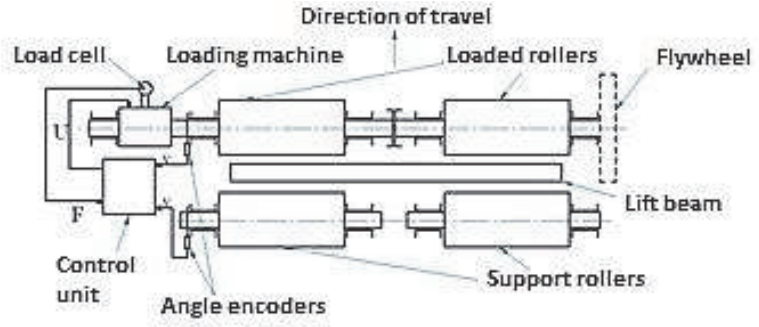



*Figure 5. Layout of the roller bench*

The appropriate clutch size can be calculated based on the maximum acceleration value of the vehicle.

The clutch must be built into the drive pulley of the flywheel of the present roller bed.

After the modification, there is no need any serious interference to couple or decouple the additive mass to or from the rollers, just a suitable electronic control unit with a software is needed.

### 5.2. Modification of the roller bench

During the modification of the roller bench, the present pulley must be modified in order to enable the mounting of the electromagnetic clutch inside (Fig 6.).

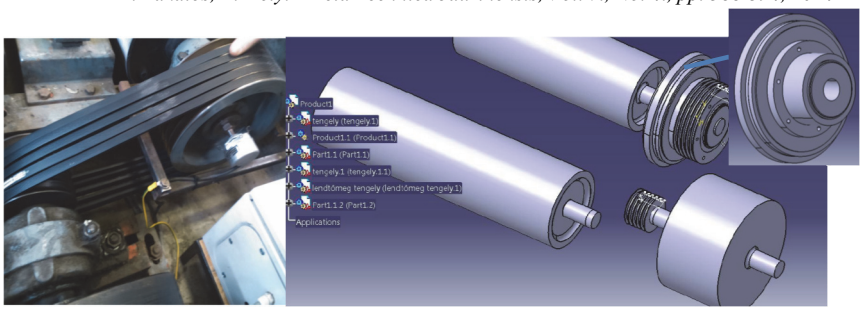



*Figure 6. Design of possible layout of electromagnetic clutch*

The roller bench modified or built according to the design (shown in fig. 6.) is capable of measuring the effective power of a vehicle engine. Provided, the additive mass and appropriate measurement software is used. The main advantage of the method is that high measurement accuracy can be achieved with the help of diagnostic tools and a budget-saving machinery.

## 6. Summary

Several advantages emerge with elaborating a new measurement system [6, 7, 8, 9, 11, 12, 13]:

- There is no need for rolling road, therefore simpler and cheaper measurement devices can be developed.
- The measurement is more accurate since empirical correction factors are excluded from the calculations.

The new method described in the present article provides an available opportunity for professional garages, as it is capable of defining the diagnostic performance of the engine with the required accuracy.

It is a very important achievement that with the help of diagnostic tools (without removal of the engine) an exact result of the effective performance of the engine can be achieved, as it significantly differs from the performance result of the wheels that can be measured on the rolling road (Fig. 5.)

## 7. Acknowledgement

TÁMOP-4.2.4.A/2-11/1-2012-0001 identification number of the National Excellence Program - was the establishment and operation system for domestic students, researchers and staff support as a priority project by the convergence program. The project is funded by the EU, the European Social Fund co-funded.

## References


[1] Lakatos, I.: *Analyse der Zusammenhängen zwischen indizierten Motorkennwertwen und Rollprüfstanduntersuchungsergebnissen*, Járművek, vol. 49, no. 6, pp. 31-34, 2002

[2] Lakatos, I.: *Comparative measures on rolling road dynamometers*, XXV. microCAD International Scientific Conference, Miskolc, Hungary, pp. 57-64, 2011

[3] Lakatos, I.: *Gasoline engine diagnostic on chassis dynamometers*, XIX. microCAD International Scientific Conference, Miskolc, Hungary, pp. 27-32, 2010

[4] Lakatos, I.: *Instacioner Engine Performance Measure on Rolling Road Dynamometers*, XIX. microCAD International Scientific Conference, Miskolc, Hungary, pp. 33-38, 2010

[5] Lakatos, I.: *Measuring Engine Performance with Diagnostic Tools*, Innovation and Sustainable Surface Transportation Conference, Budapest, Hungary, CD proceeding, 2010

[6] Péter, T.: *Reduction of parameters of spatial non-linear vehicle swinging systems, for identification and optimisation purposes*, Perodica Polytechnica: Transportation Engineering, vol. 16, no. 1-2, pp. 131-141, 1988

[7] Péter, T., Bokor, J.: *Modeling of vehicle traffic systems and research of their control (in Hungarian),* A Jövő Járműve: Járműipari Innováció, vol. 1, no.1-2, pp. 19-23, 2006

[8] Peter, T, Bellay, A.: *Integral Transformations of Road Profile Excitation for Variable Vehicle Speeds*, Vehicle Systems Dynamics, vol. 15, no. 1, pp. 19-40, 1986

[9] Péter, T.: *Fuzzy and Anytime Signal Processing Approaches for Supporting Modeling and Control*, In: Rudas I.J. (ed.), ICCC 2005: IEEE 3rd International Conference on Computational Cybernetics, pp. 339-344, 2005
DOI: 10.1109/ICCCYB.2005.1511598

[10] Lakatos, I.: *Diagnostic measurement for the effective performance of motor vehicles*, Acta Polytechnica Hungarica, vol. 10, no. 3, pp. 239-249, 2013
DOI: 10.12700/APH.10.03.2013.3.16

[11] Gáspár, P., Bokor, J., Szabó, Z., Fülep, T., Szauter, F., Mihály, A.: *Integrated robust control design for in-wheel-motor vehicles*, FISITA 2014 World Automotive Congress, 2-6 June 2014, Maastricht, 2014

[12] Zezyulin, D.V., Makarov, V., Ogorodnov, S., Belyakov, V.: *Methodology of Roadway Impacts Modelling to Predict the Fatigue Life of Vehicles Parts*, Acta Technica Jauriensis, vol. 7, no. 3, pp 267-279, 2014
DOI: 10.14513/actatechjaur.v7.n3.277

[13] Szauter, F., Péter, T., Bokor, J.: *Examination of Complex Traffic Dynamic Systems*, Acta Technica Jaurinensis Series Transitus, vol. 6, no. 3, pp 136-142, 2013

Acta
Technica
Jaurinensis

Vol. 7., No. 4., pp. 395-403, 2014
DOI: 10.14513/actatechjaur.v7.n4.329
Available online at acta.sze.hu

# Some Considerations on Serious Road Traffic Injuries

**P. Holló, D. Kiss**

**Széchenyi István University, H-9026 Győr, Egyetem tér 1.**
**Phone: +36 96 503 400, fax: +36 96 503 400**
**e-mail: hollo@sze.hu; kdiana@sze.hu**

Abstract: In the EU in 2011 the number of serious road injuries was more than 250,000 and the death tolls were 28,000. In the last period (between 2001 and 2011) the number of those who lost their lives as a result of road accident decreased ordinarily by 43 percent in the EU countries on average, whereas that of the seriously injured by 36 percent – in the light of these countries' own definitions differing from one another. The MAIS3+ is the adopted common EU definition, that is all 3-grade or above values according to the Maximum Abbreviated Injury Scale (MAIS). Although the definition seems professionally justified, in our view further clarification is necessary. For the years 2014 and 2015 the EU has already drawn up specific tasks for the member states. Since the Baltic States were particularly successful in the field of road safety in the last 10 years, it is certain that in the future they can do a lot in order to have the number of serious road accident victims significantly reduced and also internationally compared and assessed.

*Keywords: road safety, serious injuries, MAIS3+*

## 1. Introduction

As a consequence of road accidents the number of people killed was 28.126 and that of people injured was 1.432.235 in the 28 member states of the EU in 2012. For every death on Europe's roads there are an estimated 4 permanently disabling injuries such as damage to the brain or spinal cord, 8 serious and 50 minor injuries. [5].

In the EU the road accident is considered as number one mortality cause in the age group of 45 years and younger ones. Road accident is similarly the cause of most hospitalizations.

Beyond human sufferings injuries cause tremendous loss for the national economy, too. In the EU this is estimated to 2% of the GDP. In 2012 this amount was 250 billion Euro. In worldwide dimension, according to WHO data, this is approximately equal to 580 million dollars/year. [9].

On the priority list the most frequent serious injuries are the head and brain impairments then follow the traumas of the lower limbs and the vertebral column. Mainly vulnerable road users (pedestrians, cyclists, motorcyclists) or the most vulnerable age groups (elderly people, children) are the victims of such injuries.

Such kind of injuries can be experienced on every road types, however most of them occur in built-up areas and their victims are the vulnerable road users. Mainly because of higher speed, injuries are even more serious outside built-up areas.

In the last period (between 2001 and 2011) the number of those who lost their lives as a consequence of road accident decreased by 43 percent in the EU countries on average, whereas that of the seriously injured by 36 percent [4].

The Figure below illustrates the change in EU fatalities between 2001 and 2013 [5].


*Figure 1: The number of road accident fatalities in the EU between 2001 and 2013[5]*

Comparing the two data there are many who note that while in the period in question in the EU on average the number of fatalities decreased by 43%, that of the seriously injured by 36% "only".

"Only", in our opinion is unjustifiable because one must not forget that several passive safety devices (e.g. the safety belt, the airbag, etc) are the cause different kinds of injuries while saving the life of those involved in accidents.

To put it in another way: the "price" of survival mostly involves the endurance of the consequences of some injury.

To set a more moderate, numerical target in order to change the number of seriously injured seems to be more realistic.

In most highly motorized countries a dramatic decrease in the number of accident fatalities can be observed, i.e. primarily it was not the probability of the occurrence of road accidents with personal injury that decreased but the probability of survival increased. In other words: it seems that the development of passive safety is more successful than that of the active safety. (So-called "risk compensation" is likely to have

a role in this which means that the devices meant to enhance active safety – while generating a false sense of safety in the driver – unfavourably affect the driver's behaviour and lead to higher levels of risk-taking.)

Such trends can be observed in Hungary, too. (See Fig. 2) One can recognize the dramatic decrease in the number of road fatalities (passive safety) but only a small decrease in the number of personal injury accidents (active safety).

No doubt those active safety devices are very important, too, which – even without the driver's knowledge – support accident prevention and, as a consequence, the avoiding of serious injuries as well. Such an active safety device is for example the electronic stability programme (ESP) which by separate braking of different wheels is responsible for correcting the stability loss of the vehicle (under- or over-steering) in critical situations.


*Figure 2: Number of road motor vehicles, of personal injury accidents and the victims killed as a consequence between 1976 and 2012 (The main road safety phases)*

## 2. Definitions in Hungary

The EU has recognized the importance of serious injuries however, the absence of a uniform EU definition made impossible the setting of a common numerical target.

Before describing the development accomplished in this area a brief overview of the domestic situation is given below.

Until 2011 according to national accident statistics those injured were considered as serious cases whose recovery was beyond 8 days.

The experts of accident analysis have already previously found that this 3-degree scale (fatal, serious and slight injuries) is completely improper for the appropriate classification of traumatic injuries, because, for example a person already entirely healthy on the ninth day was considered as seriously injured as the one who was forced to end his life in a wheelchair.

The AIS scale (Abbreviated Injury Scale) [1] which makes a more suitable and a more relevant comparison possible has been used long ago in the domain of public health nonetheless that its application requires high level medical knowledge.

Before dealing with this, one has to see how the definitions of the accident statistics changed.

As of 2011 the Hungarian Central Statistical Office (KSH) adopted the following definitions [7]:

Serious injury: means an injury suffered in the course of an accident, and which

- requires hospitalization for more than 48 hours within seven days as of the date of the injury, or
- causes some fracture, with the exception of the fractures of fingers, toes and nose, or which
- involves lacerations causing severe haemorrhage or nerve, muscle or tendon damages, or
- causes damage to the internal organs, or
- involves second or third degree burns or harms as a result of which more than 5% of the body surface area burns.

This definition – which has been introduced in the spirit of the harmonization of the different transport modes – raises doubts on the one hand, in connection with the homogeneity of time series, and on the other hand, with respect to data's verification.

Not to mention the fact that one cannot expect the police officer visiting the scene of the accident to judge the outcome in a professional way, since no such training has been obtained.

Despite the change in definition no significant change appears in the decreasing tendency, so in addition to various definitions of terms the number of injuries seems to be comparable without correction factors. (Figure 4)

Causing a road traffic accident with serious injury is considered as a criminal offence, consequently it is the subject of a more severe judgement and the administrative proceedings related to the case differ also from the milder cases.

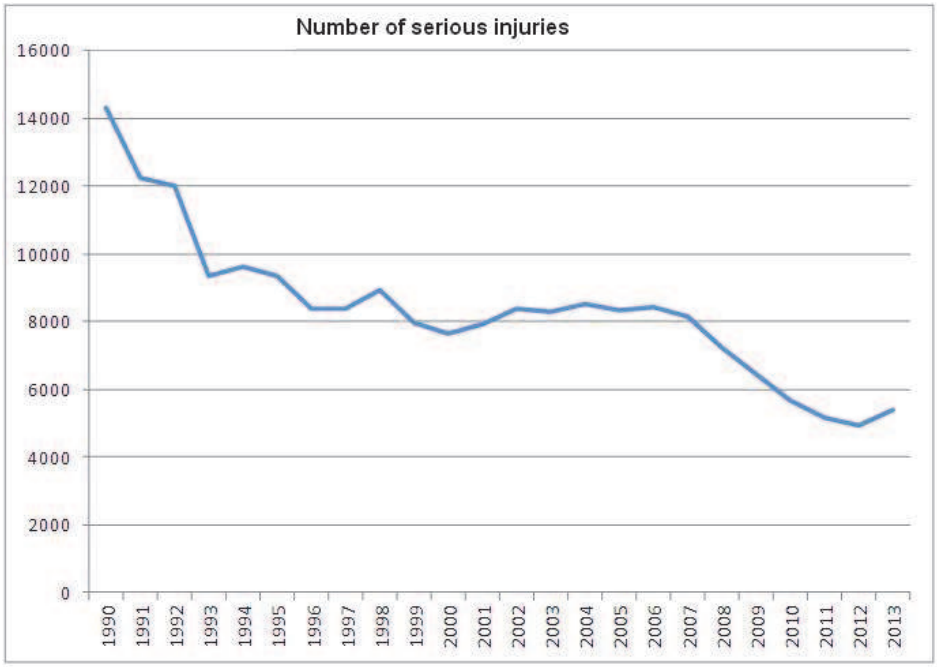



*Figure 4: Number of serious injuries resulting from road accidents between 1990 and 2013 [5]*

## 3. Definitions used in other countries

It is just the absence of a uniform definition that makes difficult the international comparison of the data related to serious injuries and the definition of the numerical EU target.

This is one reason why the international comparison of road safety is still limited to comparing the data of road accident fatalities, because the definition of the fatally injured (the so-called 30-day definition) is widely uniform. In case of some countries where this is not used, its lack can be solved by the application of the so-called correction factor.

Some examples to illustrate different definitions of the seriously injured:

**The period of hospitalization:**

In most countries 24 hours

In Poland: 7 days

**Type of injury:**

Sweden: a person who has suffered some fracture, contusion, rupture, severe cut, shock, or internal injury.

**Incapacity:**

Austria, Switzerland.

**Time of recovery:**

Japan: more than 30 days.

To compare the data of different countries is further complicated by the circumstance that there is a significant discrepancy between the statistical data which are based on police investigation on the spot and the data recorded in the medical databases. This is the so-called underreporting, which is due to the fact that in the event of serious or light injuries in some cases the participants do not call police. According to some studies only about 70% of the data relating to serious injuries are recorded in the databases of the police. [6]. From the point of view of data deficiency, too, the figures relating to fatal injuries can be considered as the most complete and reliable ones. Not to mention the fact that these are the most tragic consequences of road accidents.

The uniform definition could eliminate the differences manifested in data deficiency [2].

The first step taken in this direction has been made by IRTAD (International Road Traffic and Accident Database, the accident and road traffic database of the OECD countries) with the introduction of the definition of the "hospitalized person". This referred to injured who spent minimum 24 hours in hospital. [3]. (Today you may already know that besides its benefits it wasn't precise enough and did not really spread).

It seems increasingly that only a reliable, professional national public health database can provide a complete picture on the data relating to the injured of the road accidents.

## 4. Brief information about the AIS scale

To encode the severity of injuries the following codes are used in the AIS 2005 [1] updated in 2008 (Table 1.):

*Table 1: AIS code and description*

| AIS code | description |
|---|---|
| 1 | Minor (slight, insignificant) |
| 2 | Moderate (moderate) |
| 3 | Serious (serious) |
| 4 | Severe (very serious) |
| 5 | Critical (dangerous, life-threatening) |
| 6 | Maximal (fatal, life-incompatible) |

The common EU-wide definition that until now has been adopted by all organizations (the EU High Level Group, IRTAD, ETSC, etc.) is the

MAIS3+,

i.e. all number 3 values or those beyond this according to the Maximum Abbreviated Injury Scale (MAIS).

In our view this definition is not precise enough

On the one hand, it includes the AIS6 value, too, which practically means those who died on the spot. Thus, there is an inconsistency (overlapping) in the definition and at least a theoretical risk that the fatal victims are taken into account twice.

On the other hand, the interval is open from one side and closed from the other side which is inconsistent in our opinion.

We consider that the precise definition of serious injuries can range from MAIS3 to MAIS5.

Based on the above the EU has defined the following tasks:

- In 2014 the member states have to make arrangements for being prepared for the use of the new definition
- In 2015 the member states have to provide information concerning the first, serious injury data

Subsequently the EU sets a numerical target and determines a strategy for reducing the number of serious injuries between the years 2015 and 2020.
The Forum of European Road Safety Research Institutes (FERSI) established a working group called the "Severely injured road users in crash statistics" when recognized the challenges of the tasks and the existing gaps of the research. Dr. Péter HOLLÓ is member of this group.

According to our information the EU received the so-called "position paper" well and is ready to cooperate with FERSI in solving the existing problems in the field.

It's definitely worth mentioning that an ongoing EU project is just aimed at creating a uniform European system in order to record the injured persons' data in a professional and reliable way [10].

This is the JAMIE (Joint action on monitoring injuries in Europe) project, which will be completed by mid 2014. As we are informed not even the suggested whole data content (FDS) will include the data describing the injury's severity. The National Health Development Institute is representing Hungary in the consortium.

The EU High Level Group deems that the following solutions are feasible to resolve the outlined tasks:

- Further collection of police data, application of correction factors to estimate the real number of the injured,
- Collection of the data at hospitals using the MAIS codes.
- Linking the two data sources (police and hospitals).

In our opinion the first solution may only be a temporary one and determining the correction factors must be based on a representative sample. It goes without saying that data recorded by the police forming the bases of numerous activities are still very much needed. (Limitations: underreporting, not always precisely defined accident causes, etc.) This may be the reality in the near future.

The second solution requires the establishment of the collection system of the national health data and professional application of the AIS codes. To our knowledge this cannot

be expected in the near future in Hungary. Without the collection of former police data this is not sufficient either.

The third option gives the optimal long-term solution providing the most complete picture about the seriously injured. In most countries linking of the two datasets cannot be done by using the name of the injured person (protection of personality rights) which complicates this process.

Close co-operation and common work of the police and hospitals (moreover, of the polyclinics and family physicians) are indispensable for preparing precise statistics.

No "medical" accomplishment can be expected from the police officer arriving at the scene of an accident to determine on the basis of what has been witnessed the severity of an injury. While having an almost permanent contact with the police, neither a doctor's working time nor the intensive stress of work allow to harmonize the number and severity of the injuries. Co-operation between these two work-fields needs necessarily the development of such an information background that would allow the simple, fast but the more accurate recording.

Currently there are only a few EU member states that have the data meeting all the requirements (Sweden, the Netherlands, Austria, etc.)

In some countries helped by the appropriate algorithms the ICD codes are transformed into AIS, or MAIS codes, which is also a possible solution [8].

## 5. Challenges

The EU expectations for 2014, such as the new uniform definition, too, seem somewhat premature. See some challenges which can be outlined already now:

- The definition should be clarified.
- Development of a national public health database containing also the severity data needs much time and expenditure. It requires the increased co-operation of the ministry of health and home affairs.
- It is not decided yet whether the overburdened health workers should be involved in the encoding process, or is there any other idea to solve the problem. (The application requires a high level expertise.) Universities, research institutes could come into question.
- The quantified target for reducing the number of seriously injured can be developed only if it is known already the real number of the occurrence of serious injuries.

Overall it may be concluded that co-ordinated measures are needed which have to cover the division of responsibilities, collaboration, legislation, enforcement and even many other areas.

## References

[1] AIS 2005: *Abbreviated Injury Scale 2005*, Update 2008, Association for the Advancement of Automotive Medicine, 2008

[2] Broughton, J., Keigan, M., Yannis, G., Evgenikos, P., Chaziris, A., Papadimitriou, E., Bos, N.M., Hoeglinger, S., Pérez, K., Amoros, E., Holló, P., Tecl, J.: *Estimation of the real number of road casualties in Europe*, Safety Science, vol. 48, pp. 365-371, 2010

[3] Derriks, H.M., Mak, P.M.: IRTAD special report: *underreporting of road traffic casualties*, 2007

[4] ETSC Response to the European Commission Staff Working Document *"First Milestone towards a Serious Injury Strategy"*, Brussels, 2013

[5] European Comission: *Statistics – accidents data: Trends*, http://ec.europa.eu/transport/road_safety/specialist/statistics/index_en.htm (last visited at 20. June 2014)

[6] European Commission (2013): *Towards a European road safety area: policy orientations on road safety* http://europa.eu/rapid/press-release_MEMO-13-232_hu.htm (last visited at 5. June 2014)

[7] Hungarian Central Statistical Office*: Number of personal injury accident* http://www.ksh.hu/docs/hun/xstadat/xstadat_evkozi/e_ods001.html (last visited at 20. June 2014)

[8] Niels Bos*: Injury coding systems and conversations between them: AIS, ICD*. Better safety data for better road safety outcomes, Buenos Aires, Argentina, 13-14 November 2013

[9] World Health Day: *Road safety is no accident!* http://www.who.int/mediacentre/news/releases/2004/pr24/en/ (last visited at 5. June 2014)

[10] W.H.J.Rogmans: *Joint action on monitoring injuries in Europe (JAMIE)*, Rogmans Archives of Public Health, vol. 70, no. 19, 2012

Acta
Technica
Jaurinensis

Vol. 7., No. 4., pp. 404-427, 2014
DOI: 10.14513/actatechjaur.v7.n4.340
Available online at acta.sze.hu

# Performance and Stability Analysis of Free NAT64 Implementations with Different Protocols

**S. Répás, P. Farnadi, G. Lencse**

**Széchenyi István University, Department of Telecommunications**
**Egyetem tér 1, H-9026 Győr, Hungary**
**E-mail: repas.sandor@sze.hu, farnadi@tilb.sze.hu, lencse@sze.hu**

Abstract: Due to its rapid spread, Internet has now outgrown the address space provided by IPv4. The transition to the new IPv6 protocol is extremely slow. The different transition techniques are intended to facilitate the transition to the new version of Internet Protocol. The NAT64 transition technique is one of the most suitable solutions. In this paper, both the performance and the stability of two NAT64 gateway implementations are examined by ICMP, TCP and UDP protocols. The tested two free implementations, the TAYGA on the Linux system and the PF of the OpenBSD system can be effectively used in a production environment, and can facilitate the deployment of the IPv6 protocol.

*Keywords: NAT64, TAYGA, PF, performance, stability*

## 1. Introduction

Due to the exhaustion of the IPv4 address pool [1], the internet service providers will not be able to provide IPv4 addresses to their customers in the near future. The application of the new version of the Internet Protocol, the IPv6 can provide practically infinite number of addresses for the huge number of Internet-enabled devices. The specification of the IPv6 protocol exists since 1998 [2], but the widespread application of it is still pending. While the IPv6 protocol can solve the address exhaustion problem, the application of the new protocol raises another problem. Because of the different IP header structure and addressing scheme, the direct communication between an IPv4 and IPv6 hosts is impossible. The most important hindering factor of the rapid deployment of the IPv6 protocol is the combination of this incompatibility and the huge number of the installed IPv4 only devices. To solve this problem and speed up the IPv6 widespread implementation one can use the so called IPv6 transition techniques. These techniques enable the communication between hosts in a mixed IPv4 and IPv6 network. Over the years several transition techniques have been developed. According to the authors' opinion, the combination of a DNS64 [3] server and a NAT64 [4] gateway is currently one of the most useful techniques to speed up the IPv6 deployment. The appropriate choice of high performance and very stable DNS64 and NAT64 implementations can be crucial for the service providers. This paper deals with the stability and performance of

two different open source NAT64 implementations with the three most important protocols over: ICMP, UDP and TCP.

The remainder of this paper is organized as follows: first, the operation of the DNS64+NAT64 solution is described, second, TAYGA under Linux and Packet Filter of OpenBSD are introduced, third, the NAT64 performance and stability research results are surveyed, fourth, the description of the test network and the testing method of each protocols are given, fifth, the test results are introduced, sixth, our results are summarized and discussed, and finally, our conclusions are given.

## 2. The DNS64 and NAT64 transition techniques

In the current phase of the IPv6 deployment it is a very important task to provide connectivity between an IPv6 only client and an IPv4 only server. To solve this problem one can use the DNS64+NAT64 combination. DNS64 is an extension of the DNS server, and NAT64 is similar to the "normal" Network Address Translation [5] process, but with address family translation. The operation of DNS64+NAT64 is shown in Fig. 1.

The explanation of the communication process is the following:

- Step 1: The IPv6 only client sends a query to its name server about the IPv6 address of the destination server with the DNS name of the server (query the AAAA record [6] of the destined server).
- Step 2: The DNS64 server tries to resolve the DNS name.
- Step 3:
  - If the DNS server resolves the given name to an IPv4 address, but not IPv6: The DNS64 server generates an answer with an AAAA record with a synthesized IPv6 address. This IPv6 address contains the given IPv4 address of the server at the last 32 bits, while the first 96 bits can be a network specific prefix or the NAT64 well-known prefix. This special IPv6 address is called IPv4-Embedded IPv6 Address.
  - If the DNS server could resolve the given name to an IPv6 address (the given name has an AAAA resource record) then the DNS64 server acts as normal. (This case is not shown in the figure.)
- Step 4: The DNS64 server sends back as an answer for the query of its client.
- Step 5: The client sends the IPv6 packet through its (NAT64) gateway.
- Step 6:
  - If the destination address of the packet contains the special prefix, the gateway knows that the IPv6 packet is destined to an IPv4 server. The NAT64 gateway makes a stateful network address and address family translation between IPv6 and IPv4 and sends the resulted IPv4 packet to the IPv4 only server with its own public IPv4 address as source address. The NAT64 gateway needs to have both IPv4 and IPv6 addresses, to make this process happen.

– If the destination address of the packet does not contain the prefix, the gateway forward the packet normally to its next hop. (This case is not shown in the figure.)

- Step 7: The IPv4 only server sends its answer to the IPv4 address of the NAT64 gateway.
- Step 8: The NAT64 gateway assembles the IPv6 version of the received packet using the payload of the IPv4 packet and its own state table and then sends the IPv6 packet to the originating client.

The whole process is intended to be transparent for the client computer. This IPv6 transition solution is compatible with the majority of the wide spread application layer protocols that work in client-server model (e.g. HTTP, SMTP, POP3, IMAP4, SSH, etc.) but it has issues with those protocols that transfer IP addresses (e.g. FTP, SIP) and usually does not work with peer-to-peer applications (e.g. BitTorrent), see: [7], [8] and [9].

For a more detailed but still easy to follow introduction to DNS64+NAT64, see [10] and for the most accurate and detailed information, see the relating RFCs [3] and [4].

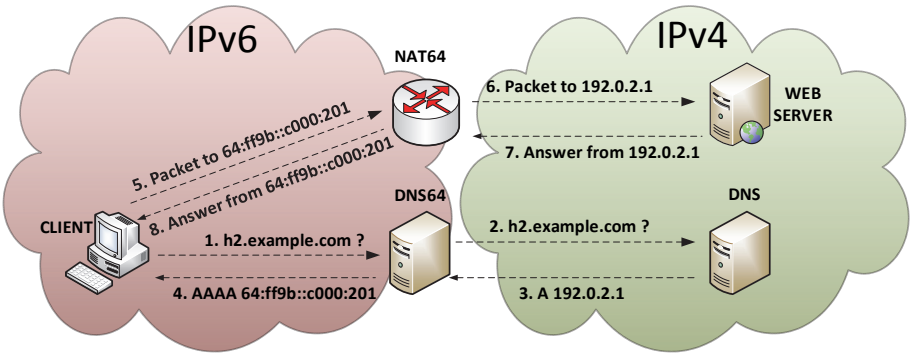



*Figure 1. Operation of DNS64+NAT64 (Based on: [11])*

## 3. The examined NAT64 implementations

### 3.1. TAYGA

TAYGA [12] is a free stateless NAT64 implementation for Linux under GPLv2 license. The main goal of its developers was to provide a production quality NAT64 service where a dedicated NAT64 device would be overkill. The latest release of TAYGA is 0.9.2. According to its authors, TAYGA could never come close to offering the power and flexibility available in the packet filter of Linux (iptables), so instead TAYGA turns IPv6 into IPv4 in the most transparent manner possible, allowing existing IPv4-only tools to be used to further manipulate sessions flowing through it. It means that by itself it can create only a one-to-one mapping between IPv6 and IPv4 addresses. For this reason TAYGA is used together with a stateful NAT44 packet filter (iptables under Linux): TAYGA maps the source IPv6 addresses to different IPv4 addresses from a suitable size of private IPv4 address range, and the stateful NAT44 packet filter performs an SNAT (Source Network Address Translation) from the private IPv4 addresses to the public IPv4

address of the NAT64 gateway. In the reverse direction, the stateful NAT44 packet filter "knows" which private IPv4 address belongs to the reply packet arriving to the IPv4 interface of the NAT64 gateway. After the NAT44 translation, TAYGA can determine the appropriate IPv6 address using its one-to-one address mapping and then it rewrites the packet to IPv6. To work with TAYGA, a suitably large private IPv4 dynamic address pool should be provided.

### Packet Filter, PF

The Packet Filter was introduced in the OpenBSD system in 2001. PF is a very popular firewall application in the BSD systems. The PF of OpenBSD supports stateful and stateless mode at the same time. It supports various types of packet manipulation, and it supports IPv4 and IPv6 stateful NAT for many years. Since OpenBSD 5.1 it supports stateful NAT64, too [13]. Stateful behaviour means that it does not need the help of a stateful NAT44 packet filter to work as a complete NAT64 gateway. This nature of PF can speed up the NAT64 translation.

## 4. A Short Survey of the Current Research Results

Though NAT64 is addressed in high number of current research papers, only a relatively few of them deals with the performance analysis of its different implementations. For example, [14] gives a good summary of the NAT64 papers until 2012 and it states that "two papers were found that deal with NAT64 performance issues". And also many of those that deal with the performance of NAT64 implementations do it in the way that they examine the performance of a given NAT64 implementation together with a given DNS64 implementation. For instance, they measure the performance of the TAYGA NAT64 implementation with the TOTD DNS64 implementation in [15]. The authors of two other papers [16] and [17] measure the performance of the Ecdysis NAT64 implementation, which uses its own DNS64 implementation. However, we have already shown that as DNS64 and NAT64 are two distinct services, they may be and thus should be analyzed separately to be able to choose the best suiting implementations for someone's purposes [18]. We have published our first results on the separate performance analysis of the TAYGA NAT64 implementation and of the BIND DNS64 implementation in [19]. Later on, we compared the performance of the BIND and TOTD DNS64 implementations under Linux, OpenBSD and FreeBSD in [20]. We tested the stability and the performance of the TAYGA and of the PF NAT64 implementations using ICMP in [18]. The aim of our current research it to test their stability and compare their performance using also TCP and UDP protocols, as they are used over IP in real life applications.

## 5. The test environment for the NAT64 performance measurements

### 5.1. The topology of the test network

The test network was built up in the Infocommunications Laboratory of the Department of Telecommunications, Széchenyi István University. The topology of the network is shown in Fig. 2. All of the network elements were connected with 1000Base-TX network connections. For this purpose, a 3Com Baseline 2948-SFP switch was used. The

responder computer can be seen on the top of the figure. The role of the computer was to answer all of the measurement traffic destined to it. For this reason, a high performance workstation computer was used for this role. The central element of the network is the NAT64 server. This gateway was built from the lowest performance computer in the laboratory, because serious overload situation was necessary during the measurement process. In the bottom of the picture eight pieces of high performance test clients are shown. These workstation computers played the role of the high number of clients during the different test scenarios.

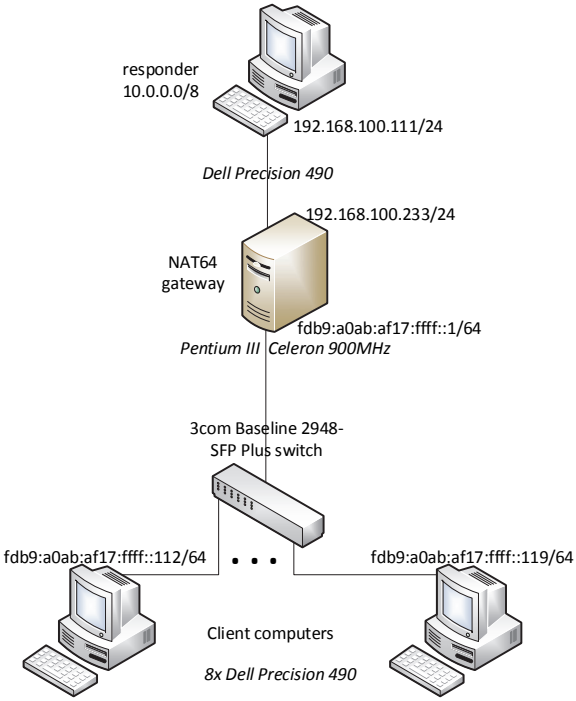



*Figure 2. Topology of the test network*

### 5.2. The hardware configuration of the computers

The configuration of the responder computer was the following:

- DELL 0GU083, Intel 5000X chipset mainboard
- Two Dual Core Intel(R) Xeon(R) CPU 5160 3.00GHz dual core microprocessors
- 4x1 GB 533 MHz DDR2 SDRAM (quad channel)
- Broadcom NetXtreme BCM5752 Gigabit Ethernet PCI Express network card
- Debian 6.0.7

- KDE 4.4.5 graphic environment

The configuration of the NAT64 gateway computer was the following:

- Intel D815EEA2 mainboard
- 900 Intel Pentium III (Coppermine) microprocessor
- 256 MB, 133 MHz SDRAM
- Two 3Com 3c940 Gigabit Ethernet PCI network cards

The configuration of all of the client computers was the following:

- DELL 0GU083, Intel 5000X chipset mainboard
- Two Dual Core Intel(R) Xeon(R) CPU 5140 2.33GHz dual core microprocessors
- 4x1 GB 533 MHz DDR2 SDRAM (quad channel)
- Broadcom NetXtreme BCM5752 Gigabit Ethernet PCI Express network card
- Debian 6.0.7

### 5.3. The software configuration of the computers

The NAT64 gateway performance was monitored with the commands as seen in Figures 3 and 4. The settings of the network interfaces are shown in Figures 5, 6 and 7.

```
dstat -t -c -m -l -p --unix --output load.txt
```

*Figure 3. NAT64 gateway computer performance monitoring on Linux system*

```
vmstat –w 1 > load.txt
```

*Figure 4. NAT64 gateway computer performance monitoring on OpenBSD system*

The settings of the TAYGA on Linux are shown in Fig. 8. The starting of TAYGA was automatized with a shell script. The script is shown in Fig. 9. The NAT64 function was enabled on the OpenBSD system in a modification in the `/etc/pf.conf` file as shown in Fig. 10.

```
#Internal
auto eth0
iface eth0 inet6 static
address fdb9:a0ab:af17:ffff::1
netmask 64
up sleep 1
up echo 0 > /proc/sys/net/ipv6/conf/eth1/autoconf
up echo 0 > /proc/sys/net/ipv6/conf/eth1/accept_ra
post-up sleep 1
post-up /root/nat64-config.sh

#External
auto eth1
iface eth1 inet static
address 192.16.100.233
netmask 255.255.255.240
gateway 192.168.100.1

pre-up echo 1 > /proc/sys/net/ipv6/conf/all/forwarding
```

*Figure 5. Network interface settings in the `/etc/network/interfaces` file on the Linux system*

```
inet 192.168.100.233 255.255.255.0
!route add –inet 10.0.0.0/8 192.168.100.111
```

*Figure 6. External network interface settings in the `/etc/hostname.sk0` file on the OpenBSD system*

```
inet6 fdb9:a0ab:af17:ffff::1 64
```

*Figure 7. Internal network interface settings in the `/etc/hostname.sk1` file on the OpenBSD system*

```
tun-device nat64
ipv4-addr 172.16.0.1
prefix fdb9:a0ab:af17:ffff:ffff:ffff::/96
dynamic-pool 172.16.0.0/12
data-dir /var/db/tayga
```

*Figure 8. TAYGA settings in the `/usr/local/etc/tayga.conf` file*

On the responder computer, all of the IP traffic destined to the 10.0.0.0/8 network were redirected to the local address of the Ethernet interface of the computer with iptables command as seen on Fig. 11.

```
#!/bin/bash
tayga --mktun
ip link set nat64 up

#Private IPv4 address space for stateless NAT
ip addr add 172.16.0.1 dev nat64
ip route add 172.16.0.0/12 dev nat64

#For NAT64 unique local IPv6 address space
ip addr add fdb9:a0ab:af17:ffff::2 dev nat64
ip route add fdb9:a0ab:af17:ffff:ffff:ffff::/96 dev nat64
tayga

#Enable packet forwarding
echo 1 > /proc/sys/net/ipv4/ip_forward
echo 1 > /proc/sys/net/ipv6/conf/all/forwarding

#Enable stateful NAT44
iptables -t nat -A POSTROUTING -o eth0 -j MASQUERADE

#The used addresses are routed to the responder
ip route add 10.0.0.0/8 via 192.168.100.111
```

*Figure 9. The `nat64-config.sh` shell script on Linux system*

```
set limit states 40000
pass in on sk1 inet6 from any to fdb9:a0ab:af17:ffff:ffff:ffff::/96 \
  af-to inet from 192.168.10.233
```

*Figure 10. The `/etc/pf.conf` settings on OpenBSD system*

```
iptables -t nat -A PREROUTING -d 10.0.0.0/8 -j DNAT \
  --to-destination 192.168.100.111
```

*Figure 11. iptables settings on the responder computer*

## 6. The measurement process and scripts

### 6.1. ICMP

## Responder

During the preliminary measurements, some fluctuations were experienced in the load of the NAT64 gateway. This behavior was caused by the limited size of the connection tracking table on the responder computer. This problem could be solved with two methods. First, the size of the table could be drastically increased. Second, the timeout value of a record in the table could be decrease to a low value. For the measurements, the second solution was chosen. It is shown on Figure 12.

```
echo 1 > /proc/sys/net/netfilter/nf_conntrack_icmp_timeout
```

*Figure 12. Setting the timeout value of ICMP packets to 1s in the conntrack table*

## Client computers

All of the measurements were made with 1, 2, 4 and 8 client computers simultaneously. The measurement of the execution time and the starting of the experiment was initiated with the command line on the Fig. 13. The synchronized start of the scripts on the client computers were done by using the "Send Input to All Sessions" function of the Konsole terminal program of the KDE graphical environment on a separated "controller" computer.

```
~# /usr/bin/time –f "%" –o runtime –a ./icmptest.sh experiment_ID
```

*Figure 13. Starting the measurement*

The test script is shown in figure 14.

```
#!/bin/bash
mkdir $1
i=`cat /etc/hostname | grep -o .$`
for b in {0..255}
do
  mkdir $1/$b
  for c in {0..248..8}
  do
    ping6 -c10 -i0  fdb9:a0ab:af17:ffff:ffff:ffff:10.$i.$b.$c \
      >> $1/$b/nat64p-10-$i-$b-$c &
    ping6 -c10 -i0  fdb9:a0ab:af17:ffff:ffff:ffff:10.$i.$b.$((c+1)) \
      >> $1/$b/nat64p-10-$i-$b-$((c+1)) &
    ping6 -c10 -i0  fdb9:a0ab:af17:ffff:ffff:ffff:10.$i.$b.$((c+2)) \
      >> $1/$b/nat64p-10-$i-$b-$((c+2)) &
    ping6 -c10 -i0  fdb9:a0ab:af17:ffff:ffff:ffff:10.$i.$b.$((c+3)) \
      >> $1/$b/nat64p-10-$i-$b-$((c+3)) &
    ping6 -c10 -i0  fdb9:a0ab:af17:ffff:ffff:ffff:10.$i.$b.$((c+4)) \
      >> $1/$b/nat64p-10-$i-$b-$((c+4)) &
    ping6 -c10 -i0  fdb9:a0ab:af17:ffff:ffff:ffff:10.$i.$b.$((c+5)) \
      >> $1/$b/nat64p-10-$i-$b-$((c+5)) &
    ping6 -c10 -i0  fdb9:a0ab:af17:ffff:ffff:ffff:10.$i.$b.$((c+6)) \
      >> $1/$b/nat64p-10-$i-$b-$((c+6)) &
    ping6 -c10 -i0  fdb9:a0ab:af17:ffff:ffff:ffff:10.$i.$b.$((c+7)) \
      >> $1/$b/nat64p-10-$i-$b-$((c+7))
  done
done
```

*Figure 14. icmp-test.sh shell script*

Each client sent 256*256*10=655360 ICMP Echo request packets to 256*256=65536 different destination IP addresses during one experiment. The body of the cycle contains 8 ping commands which were started concurrent. With the 8 simultaneous commands, we were able to provide sufficiently large load on the NAT64 gateway.

### 6.2. TCP

## Responder

On the responder computer, the Apache 2 web server was used to respond the queries sent over TCP. The Apache 2 web server was installed from the Debian repository by the `apt-get` command. The preparation of the files with different sizes was made by the dd command. For example to generate a 100 byte long file, the following command line was used:

```
dd if=/dev/zero of=/var/www/file bs=100 count=1
```

The timeout value of the TCP records in the connection tracking table was decreased to 1s by issuing the following command:

```
 echo 1 > /proc/sys/net/netfilter/nf_conntrack_tcp_timeout_time_wait
```

## Client computers

In this case the `tcp256.sh` shell script was started synchronously, similarly to the ICMP measurement. The `tcp256.sh` script is shown in Fig. 15.

The `tcp.sh` script was invoked 256 times by the `tcp256.sh` script. The `tcp.sh` script was responsible for downloading the files from the responder computer. The two scripts downloaded altogether 256*256=65536 files from 65536 IP addresses. The `tcp.sh` script is shown in Fig. 16.

```
#!/bin/bash
i=`cat /etc/hostname | grep -o .$`
for b in {0..255}
do
    mkdir $1/$b
    /usr/bin/time -f "%e" -o output.txt -a ./tcp.sh $i $b $1
done
```

*Figure 15. `tcp256.sh` shell script*

```
#!/bin/bash
for c in {0..255}
do
  wget -m http://[fdb9:a0ab:af17:ffff:ffff:ffff:10.$1.$2.$c]/file \
    -P $3/$2
done
```

*Figure 16. `tcp.sh` shell script*

### 6.3. UDP

### 6.4. Network

For the UDP measurements, the OpenVPN software with UDP VPN tunnel was selected. The logical network diagram is shown in Fig. 17. The Responder computer played the role of the OpenVPN server, while all of the Client computers were the clients of the OpenVPN server.

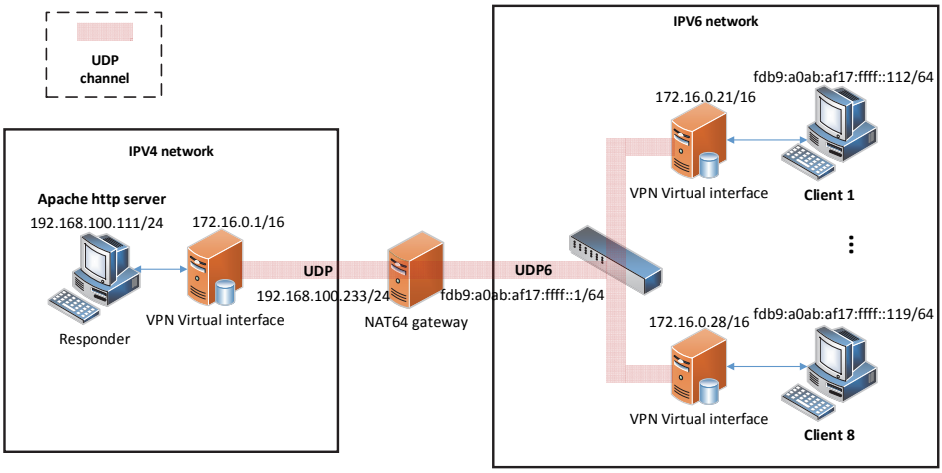



*Figure 17. Network diagram of the UDP measurement*

### 6.5. Responder

The OpenVPN server was installed from the Debian repository by the `apt-get` command. For the working OpenVPN environment, the following tasks were performed:

- Step 1: Editing the `/usr/share/doc/openvpn/examples/easy-rsa/2.0/vars` file, according to Fig. 18.
- Step 2: Creation of the `/etc/openvpn/keys` directory with the `mkdir` command.
- Step 3: Copying the `openssl.cnf`, `whochopensslcnf`, and `pkitool` files from the `/usr/share/doc/openvpn/examples/easy-rsa/2.0/` to the newly created `/etc/openvpn/keys` directory by `cp` command.
- Step 4: Building the certificate authority by issuing the commands on Fig. 19.
- Step 5: Creating the certificates by issuing the commands on Fig. 20.
- Step 6: Copying the `ca.crt`, `dh1024.pem`, `server.crt` and `server.key` files to the `/etc/openvpn/` directory.
- Step 7: Extraction of the `/usr/share/doc/openvpn/examples/sample-configfiles/server.conf.gz` file into the `/etc/openvpn` directory and modifying the `server.conf`, according to Fig. 21.

```
export KEY_COUNTRY="HU"
export KEY_PPROVINCE="GYMS"
export KEY_CITY="GYOR"
export KEY_ORG="TILB"
export KEY_EMAIL=openvpnadmin@tilb.sze.hu
```

*Figure 18. Modification of the /usr/share/doc/openvpn/examples/easy-rsa/2.0/vars file*

```
. /usr/share/doc/openvpn/examples/easy-rsa/2.0/vars
. /usr/share/doc/openvpn/examples/easy-rsa/2.0/clean-all
. /usr/share/doc/openvpn/examples/easy-rsa/2.0/build-ca
. /usr/share/doc/openvpn/examples/easy-rsa/2.0/build-dh
```

*Figure XVIX. Creating the Certificate Authority*

```
cd /usr/share/doc/openvpn/examples/easy-rsa/2.0/
./build-key-server server
./build-key client
```

*Figure 20. Creating the certificates*

```
proto udp
dev tun
server 172.16.0.0 255.255.0.0
duplicate cn
group
nogroup
```

*Figure 21. The modifications of the /etc/openvpn/server.conf file*

### 6.6. Clients

The OpenVPN clients were installed from the Debian repository by the `apt-get` command. For the working OpenVPN environment, the following tasks were performed:

- Step 1: Copying the `client.crt`, `ca.crt`, `client.key` files from the `/etc/openvpn/keys/` directory of the responder computer into the `/etc/openvpn/` directory of the client computers.
- Step 2: Copying the `/usr/share/doc/openvpn/examples/sample-configfiles/client.conf` file into the `/etc/openvpn/` directory and modifying it, according to Fig. 22.

```
dev tun
proto udp6
remote fdb9:a0ab:af17:ffff:ffff:ffff:192.168.100.111
```

*Figure 22. The modifications of the /etc/openvpn/client.conf file*

In this case the `udp256.sh` shell script was started synchronously on the client computers, similarly to the ICMP and TCP measurements. The `udp256.sh` script is shown in Fig. 23.

The `udp.sh` script was invoked 64 times by the `udp256.sh` script. The `udp.sh` script was responsible the downloading of the files from the responder computer, see Fig. 24. The core of the "`for`" cycle downloaded four files parallel. The two scripts downloaded altogether 64*256=16384 files from the same IP address.

```
#!/bin/bash
mkdir $1
for b in {0..63}
do
  /usr/bin/time -f "%e" -o output.txt -a ./udp.sh $b $1
done
```

*Figure 23. udp256.sh shell script*

```
#!/bin/bash
for c in {0..252..4}
do
  wget http://172.16.0.1/file -P $2/$1/$c &
  wget http://172.16.0.1/file -P $2/$1/$((c+1)) &
  wget http://172.16.0.1/file -P $2/$1/$((c+2)) &
  wget http://172.16.0.1/file -P $2/$1/$((c+3))
done
```

*Figure 24. udp.sh shell subscript*

## 7. NAT64 performance results

### 7.1. ICMP

## TAYGA

The results can be found in Table 1. Row 1 shows the number of clients that executed the test script. (The load of the NAT64 gateway was proportional to the number of the clients.) The packet loss ratio is displayed in the second row. Rows 3, 4 and 5 show the average, the standard deviation and the maximum values of the response time (expressed in milliseconds), respectively. The following two rows show the average and the standard deviation of the CPU utilization of the test computer. The last row shows the number of forwarded packets per seconds.

*Table 1. TAYGA, ICMP NAT64 performance*

| ***1*** | ***Number of clients*** | | ***1*** | ***2*** | ***4*** | ***8*** |
|---|---|---|---|---|---|---|
| 2 | ***Packet loss (%)*** | | 0.002 | 0.003 | 0.005 | 0.007 |
| 3 | ***Response time (ms)*** | ***average*** | 1.67 | 3.65 | 7.42 | 28.47 |
| 4 | | ***std. deviation*** | 0.56 | 0.94 | 1.70 | 9.21 |
| 5 | | ***maximum*** | 30.00 | 34.70 | 52.40 | 87.24 |
| 6 | ***CPU utilization (%)*** | ***average*** | 88.89 | 95.64 | 99.55 | 100.00 |
| 7 | | ***std. deviation*** | 2.16 | 1.77 | 0.89 | 0.00 |
| 8 | ***Traffic volume (packets/s)*** | | 3825 | 3928 | 4097 | 4128 |

Evaluation of the results:

- The packet loss ratio was slightly increased with higher traffic. With one client the NAT64 gateway produced 0.002% packet loss, whereas with 8 clients it was 0.007% packet loss.
- The CPU utilization was very high on the gateway even with one client. With 2 and 4 clients the behavior of the system remained predictable and the averages of the response times were only slightly more than doubled (3.65/1.67=2.19; 7.42/3.65=2.03). With 8 clients, the system remained stable, but the average CPU usage reached the 100%, while the average of the response times was almost four fold (28.47/7.42=3.84).
- The number of served queries were increased while the gateway computer had free CPU capacity. When the CPU usage reached the 100%, the response time increased unexpectedly.

## Packet Filter

The results can be found in Table 2. Evaluation of the results:

- The packet loss rate was always very low.
- Due to the doubling of the load on the NAT64 gateway, the response time ratios were as follows: 0.7/0.48=1.45; 1.43/0.7=2.04; 3.19/1.42=2.23. The CPU utilization increased with more clients, but the system remained stable in the serious overload situation with 8 clients.
- The number of served queries increased while the gateway computer had free CPU capacity. From 1 to 2 clients, there was 64.13% increase in the number of answered queries, whereas from 2 to 4 clients there was 22.73% increase. With 8 clients, the NAT64 gateway reached its maximum capacity, the CPU usage was close to 100%, and the number of served queries showed only a slight increase (4.35%).

*Table 2. PF, ICMP NAT64 performance*

| ***1*** | ***Number of clients*** | | ***1*** | ***2*** | ***4*** | ***8*** |
|---|---|---|---|---|---|---|
| 2 | ***Packet loss (%)*** | | 0.001 | 0.002 | 0.007 | 0.008 |
| 3 | ***Response time (ms)*** | ***average*** | 0.48 | 0.70 | 1.43 | 3.19 |
| 4 | | ***std. deviation*** | 0.22 | 0.46 | 0.77 | 1.15 |
| 5 | | ***maximum*** | 21.30 | 30.50 | 44.00 | 43.40 |
| 6 | ***CPU utilization (%)*** | ***average*** | 49.86 | 77.23 | 91.17 | 96.47 |
| 7 | | ***std. deviation*** | 6.64 | 3.91 | 1.82 | 1.51 |
| 8 | ***Traffic volume (packets/s)*** | | 8536 | 14010 | 17195 | 17944 |

## Comparison of the ICMP results

Comparing the performance results with ICMP packets of TAYGA and PF, we can state the followings:

- With one client, the throughput of PF is significantly higher than that of TAYGA: 8536/3825=223.16%, in addition to that, PF reaches this performance with significantly lower usage of the CPU. With 2, 4 and 8 clients the ratios of the throughputs of the two implementations are: 14010/3928=356.67%; 17195/4097=419.69%; 17944/4128=436.69%. (Fig. 25.)
- A similar phenomenon could be seen with the average response times: 1.67/0.48=3.48; 3.65/0.7=5.21; 7.42/1.42=5.23; 28.47/3.19=8.92 (Fig. 26.).
- In addition, the maximum values of the response times of PF are better, too.

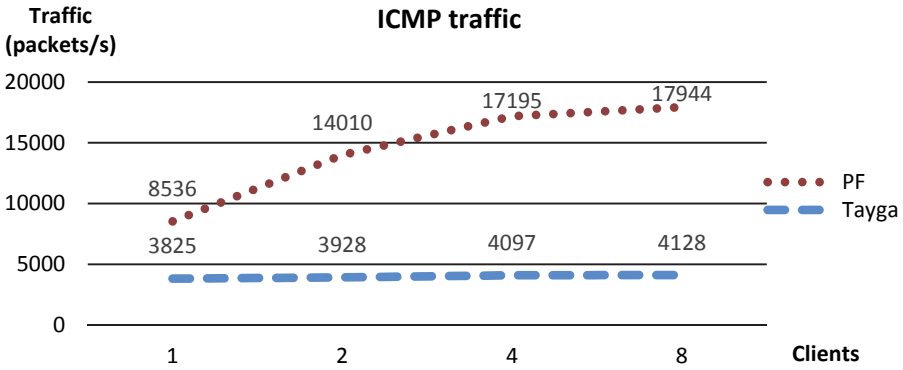



*Figure 25. ICMP throughput comparison (higher is better)*

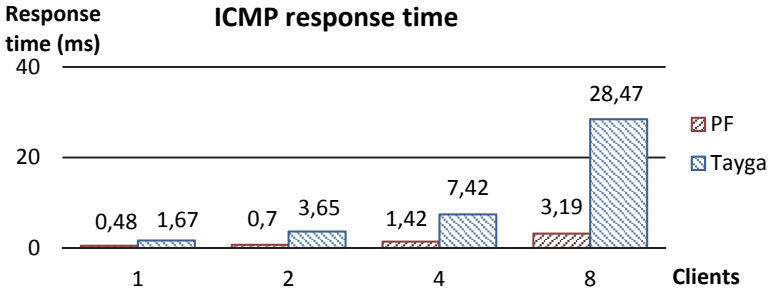



*Figure 26. ICMP response time comparison (lower is better)*

### 7.2. TCP

The performance measurements of the two NAT64 gateway implementations with TCP protocol were done by using 8 different file sizes. This method generated 8 times more data. It is more practical to interpret these results with graphs.

## TAYGA

The average number of served queries can be seen on Fig. 27. The horizontal axis (x) shows the size of the files. The sizes were doubled during the measurements from 100 to 12800 bytes. The vertical axis (y) shows the throughput of the NAT64 gateway (files/sec),

whereas the graph depth (z) corresponds to the number of clients. In the intersection points of different values of the x and z axes, the y values show that how many files with a given size and at a given number of clients the system was able to serve. The average of the response times and the average values of the CPU utilization are shown on Figures 28 and 29, respectively. Evaluation of the results:

- With one client, the CPU utilization is strictly increasing from 55% to 75% in the 100 – 12800 bytes file size range. The system was able to serve nearly the same number of files of size from 100 to 3200 bytes. Then the curve started to break down slowly, with 6400 bytes it served by 21.51% less and with 12800 bytes it served by 25.8% less than with the previous size (see Fig. 27).
- With two clients, the CPU still had free capacity, thus the response time just slightly increased, the system remained stable. The throughput of the system with 100 bytes files increased by 683/405=68.64% compared to one client. But with 12800 byte files, this difference is only 292/233=25.32%.
- With 4 and 8 clients, the system does not have free capacity, thus the response times are increasing continuously.

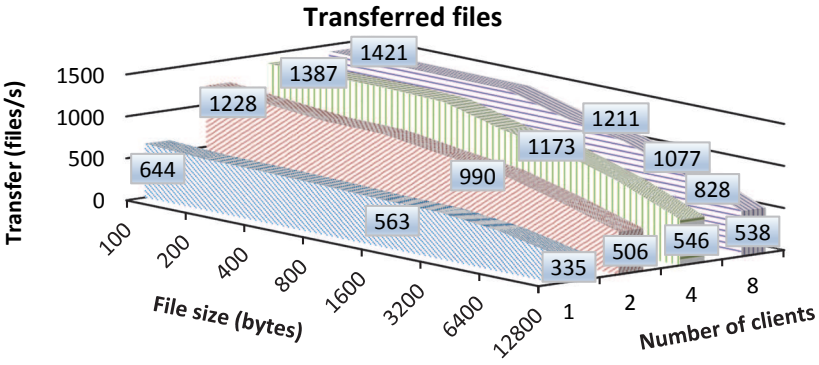



*Figure 27. Number of transferred files per second by TAYGA NAT64 gateway with TCP protocol*

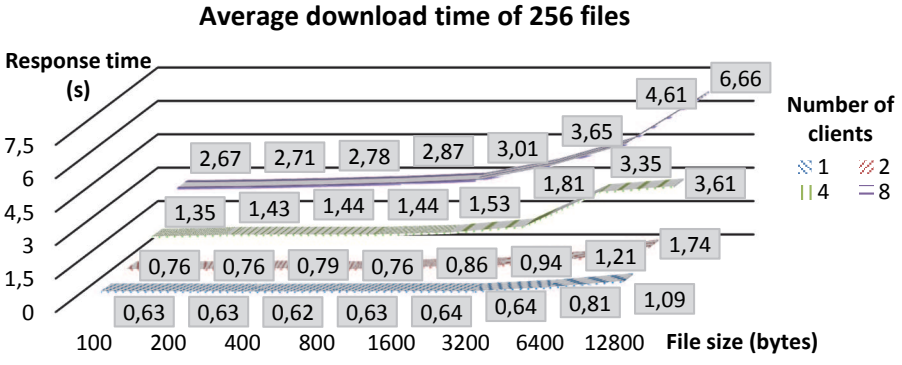



*Figure 28. Average download time of 256 files with TAYGA and TCP protocol*

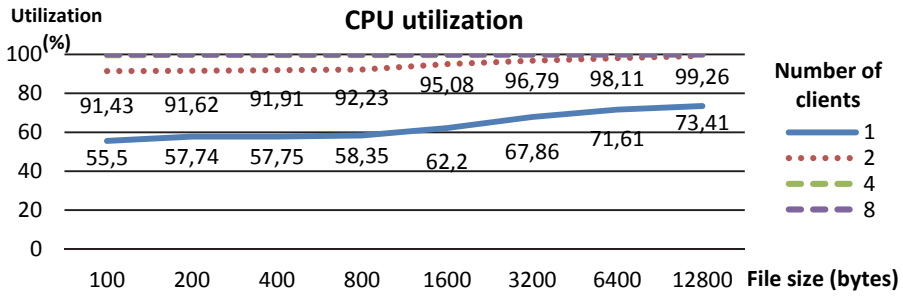



*Figure 29. Average CPU utilization with TAYGA*

## PF

The number of served queries, the average of the download times and of the CPU utilization are shown on Figures 30, 31 and 32, respectively. Evaluation of the results:

- With one client, the utilization of the CPU is 18.82%, and with the maximum file size it is 32.02%. This is the reason of the low value of the average and maximum of the transfer time. With larger size of the transferred files the number of them just slightly decreases.
- With two clients, the CPU utilization is 33.88% in the beginning of the range, whereas it is 57.81% at the end of the range. Thus the average and maximum values of the response times only slightly increase. The throughput of the system is almost doubled compared to one client (~190%).
- With four clients, the CPU utilization is 59.79% at the beginning of the range, whereas it reaches 88.91% at the end of the range. The number of the transferred files strictly decreases from the value of 192% at the beginning of the range until 162% at the end of it – compared to the values of the measurement with two clients.
- With eight clients, the system reaches the end of its capacity. At the beginning of the range it uses 93.04% of its CPU, while transfers about 70% more files compared to the case with 4 clients. With the increasing of the size of the transferred files to 1600 bytes, the CPU utilization increases, too. From this point the system starts to behave unpredictable, the response time sharply increases, while the CPU utilization starts to fluctuate. Thus, the number of the transferred files of the system decreases to 1055 files/seconds at 12800 bytes from 2875 files at 100 bytes.

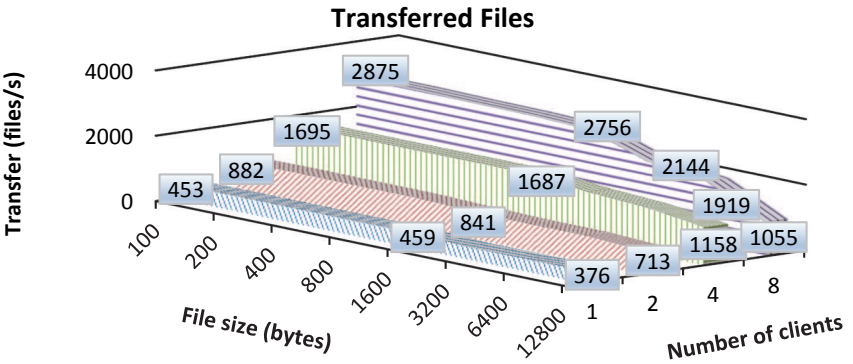



*Figure 30. Number of transferred files per second by PF NAT64 gateway with TCP protocol*

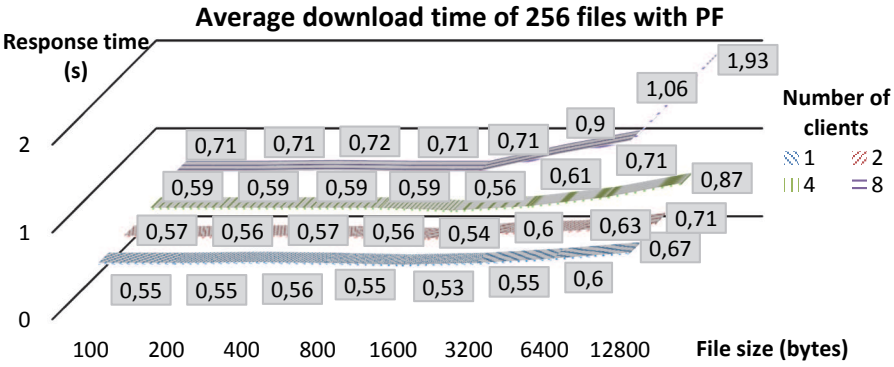



*Figure 31. Average download time of 256 files with PF*

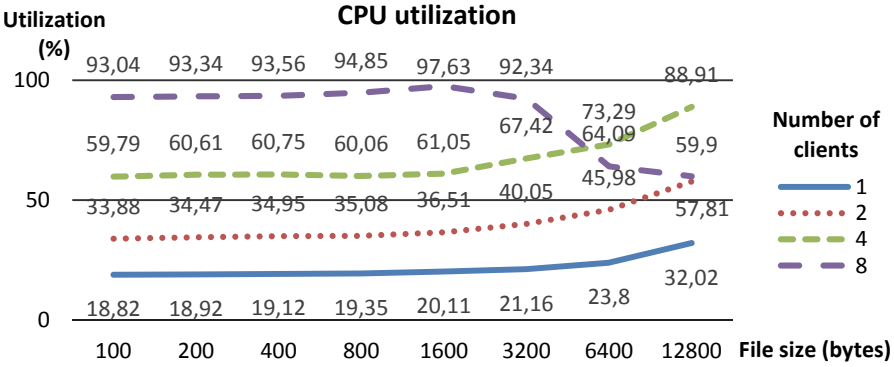



*Figure 32. Average CPU utilization with PF*

## Comparison of the TCP results

Comparing the performance results of TAYGA and PF with transfer via TCP protocol, we can state the followings:

- With one, two and four clients the CPU utilization of PF of OpenBSD was much lower than that of TAYGA on Linux system. To serve one client with 100 bytes files, PF needs 18.82% CPU time, whereas the TAYGA needs 55.5%. Thus the average and maximum values of the transfer times are also better with PF.
- In download time with 12800 bytes files the advantage of the PF is about 3.5 times.
- In serious overload situation, the TAYGA remains stable, where the behavior of the PF is unpredictable but its performance is still higher than that of TAYGA.

### 7.3. UDP

## TAYGA

The number of the transferred files, the average of the download times and of the CPU utilization are shown on Figure 33, 34, 35, respectively. Evaluation of the results:

- With one client, the utilization of the CPU is 46.04% and it starts to increase at 800 bytes long files, and it grows until 70.41% at the end of the range. The average values of the download times doubled to the end of the range. With larger size of the transferred files the number of them slowly decreases from 644 to 335, which means -48%.
- With two clients, the CPU utilization is between 77.85% and 96.28%. Thus the average and maximum values of the response times increase from 1600 bytes. With larger size of the transferred files the number of them slowly decreases from 1228 to 506, which means -59%.
- With four and eight clients, the response times greatly increase compared to the two measurement with two clients. The main cause of this phenomenon is the high processor utilization, which is about 100% at the whole range. With four and eight clients the number of the transferred files decreases with 61% and 62% to the end of the range.

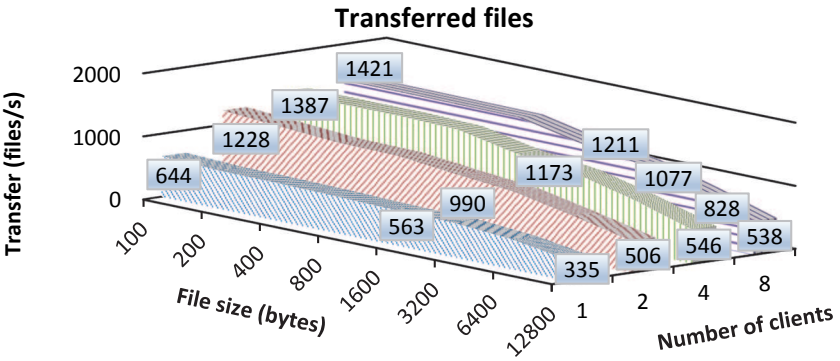



*Figure 33. Number of transferred files per second by TAYGA NAT64 gateway with UDP protocol*

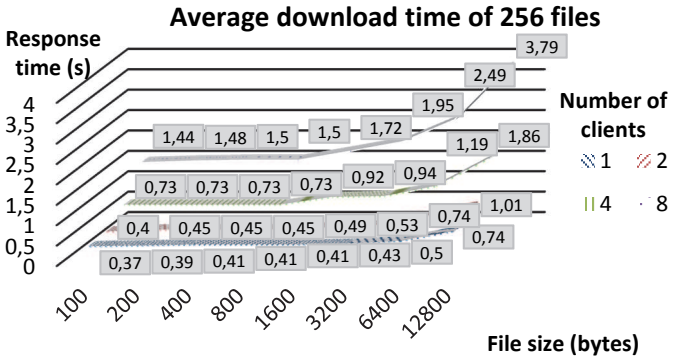



*Figure 34. Average download time of 256 files with TAYGA*

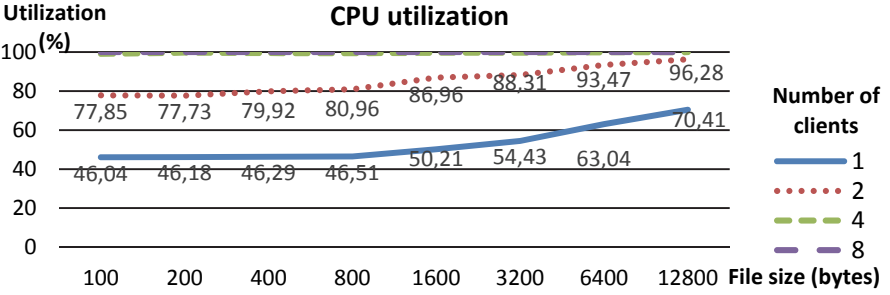



*Figure 35. Average CPU utilization with TAYGA*

## PF

The number of the transferred files, the average of the download times and of the CPU utilization are shown on Figure 36, 37, 38, respectively. Evaluation of the results:

- With one client the utilization of the CPU is 19.47% and it starts to increase at 800 bytes long files, until 32.35% at the end of the range. The average values of the response times increase by 46.87% to the end of the range. With larger size of the transferred files the number of them slowly decreases from 747 to 446, which means -40%.
- With two clients, the CPU utilization is between 32.67% and 48.87%. The average values of the response times increase by 47% to the end of the range. The response times increase by 3.13% through 15% at the whole range compared to the one client case, while the number of the transferred files increases by 92.5% through 95%.
- With four clients, the utilization of the CPU starts at 59.41% and starts to increase between 1600 and 3200 bytes long files, until 75.59% at the end of the range. The response times are almost doubled in the range. The response times increase by 6.1% at 100 bytes, whereas 26% at 12800 bytes, compared to the measurement with two clients, while the number of transferred files increase to 195.91% with 100 bytes files and 167% with 12800 bytes files.
- With eight clients, the average utilization of the processor starts at 76.31%, and finishes at 82.41%. Comparing with the 4 clients case, the average of the response times increases by 37% at 100 bytes, and by 75% at the end of the range, while the throughput of the NAT64 gateway increases by 43% and 18%.

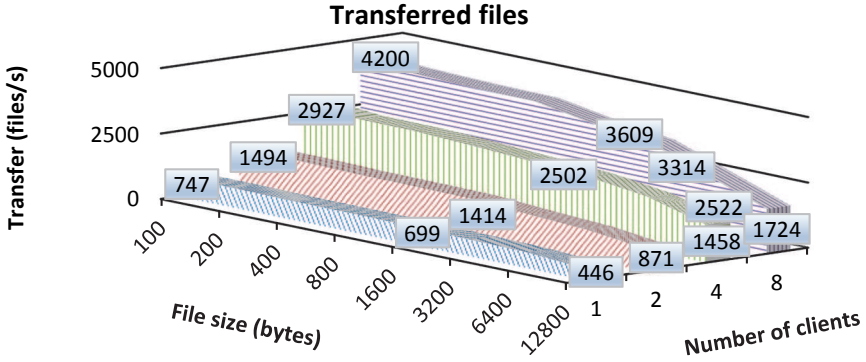



*Figure 36. Number of transferred files per second by PF NAT64 gateway with UDP protocol*

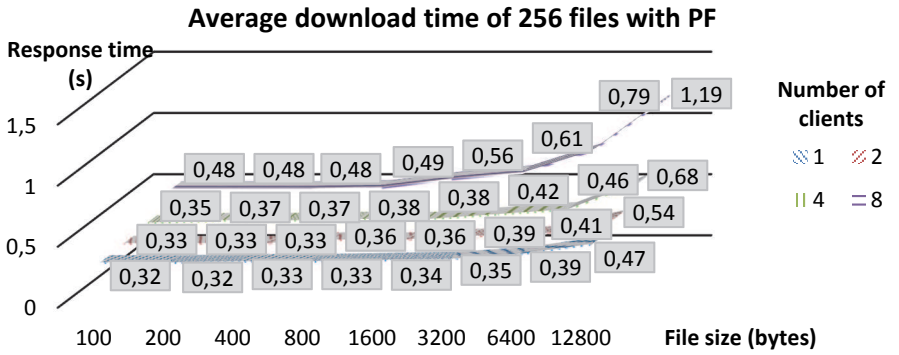



*Figure 37. Average download time of 256 files with PF*

## Comparison of the UDP results

With one client, the PF based NAT64 gateway can transfer by about 20% more files then TAYGA at the whole range, with less than the half of the CPU usage of the TAYGA. With more clients PF gains even greater superiority over TAYGA. The advantage of PF with eight clients in the number of transferred 100 bytes long files is 295% and it is 320% with 12800 bytes files. TAYGA used all of its computing power with 4 clients, whereas PF cannot reach the 83% with eight clients. Both of the implementations proved their stability during the measurements.

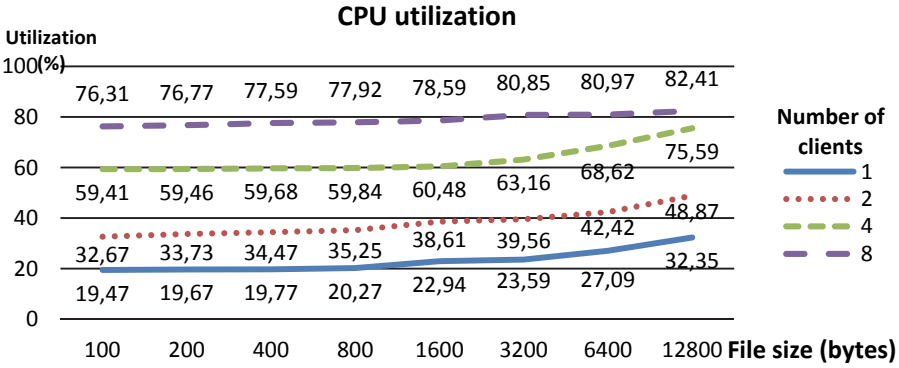



*Figure 38. Average CPU utilization with PF*

## 8. Conclusions

Both of the two NAT64 implementations were found stable enough to be used in a production environment, whereas PF of OpenBSD showed the best overall performance characteristics.

With ICMP traffic, both implementations behaved stable, and the throughput of PF was significantly better than that of TAYGA. In the serious overload situation PF outperformed TAYGA more than 4 times by means of the number of forwarded packets

per second. With the TCP protocol, the dominance of PF is reduced to 3.5 times, which is still a significant value. However, PF became unpredictable under very high volume of TCP traffic. PF produced its relatively lowest values with UDP traffic, "only" 3.2 times outperformed TAYGA on Linux, and remained stable all of the time.

As for their response times, PF was 8.9 times faster than TAYGA with ICMP protocol at the highest load, whereas this proportion was 3.45 with TCP and 3.18 with UDP protocol.

## Acknowledgement

The work of Sándor Répás was supported in the framework of TÁMOP 4.2.4. A/2-11-1-2012-0001 "National Excellence Program – Elaborating and operating an inland student and researcher personal support system convergence program" The project was subsidized by the European Union and co-financed by the European Social Fund.

The work of Gábor Lencse was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0012: "Smarter Transport" – IT for co-operative transport system – The Project is supported by the Hungarian Government and co-financed by the European Social Fund.

## References

[1] Geoff, H.: *IPv4 Address Report*, http://www.potaroo.net/tools/ipv4/

[2] Deering, S., Hinden, R.: *Internet Protocol, Version 6 (IPv6) Specification*, (RFC 2460), 1998

[3] Bagnulo, M., Sullivan, A., Matthews, P., Beijnum, I.: *DNS64: DNS extensions for network address translation from IPv6 clients to IPv4 servers*, IETF, ISSN: 2070-1721 (RFC 6147), 2011

[4] Bagnulo, M., Matthews P., Beijnum, I.: *Stateful NAT64: Network address and protocol translation from IPv6 clients to IPv4 servers*, IETF, ISSN: 2070-1721 (RFC 6146), 2011

[5] Egevang, K., Francis, P.: *The IP Network Address Translator (NAT)*, IETF, RFC 1631, 1994

[6] Thomson, S., Huitema, C., Ksinant, 5., Souissi, M.: *DNS Extensions to Support IP Version 6*, IETF, RFC 3596, 2003

[7] Škoberne, N., Ciglarič, M.: *Practical Evaluation of Stateful NAT64/DNS64 Translation,* Advances in Electrical and Computer Engineering, vol. 11, no. 3, pp. 49-54, 2011
DOI: 10.4316/AECE.2011.03008

[8] Bajpai, 5., Melnikov, N., Sehgal, A., Schönwälder, J.: *Flow-based Identification of Failures Caused by IPv6 Transition Mechanisms,* in Dependable Networks and Services, Springer LNCS, vol. 7279, pp. 139-150, 2012,
DOI: 10.1007/978-3-642-30633-4_19

[9] Répás, S., Hajas, T., Lencse, G.: *Application Compatibility of the NAT64 IPv6 Transition Technology* in 37th International Conference on Telecommunications and Signal Processing (TSP 2014), Berlin, Germany, pp. 49-55, 2014

[10] Bagnulo, M., Garcia-Martinez, A., Beijnum, I.: *The NAT64/DNS64 tool suite for IPv6 transition*, IEEE Communications Magazine, vol. 50, no. 7, pp. 177-183, 2012
DOI: 10.1109/MCOM.2012.6231295

[11] MRO: *IPv6 transition mechanisms: NAT64*, http://en.wikipedia.org/wiki/IPv6_transition_mechanisms#mediaviewer/File:NAT64.svg
[12] *TAYGA: Simple, no-fuss NAT64 for Linux*, http://www.litech.org/tayga/
[13] Theo de Raadt: *OpenBSD 5.1*, 2012, ISBN 978-0-9784475-9-5, http://www.openbsd.org/51.html
[14] Hodzic, E., Mrdovic, S.: *IPv4/IPv6 Transition Using DNS64/NAT64: Deployment Issues*, in IX International Symposium on Telecommunications (BIHTEL), Sarajevo, Bosnia and Herzegovina, 2012
[15] Llanto, K. J. O., Yu, W. E. S.: *Performance of NAT64 versus NAT44 in the Context of IPv6 Migration*, in International MultiConference of Engineers and Compuer Scientists (IMECS 2012), Hong Kong, pp. 638-645, 2012
[16] Monte, C. P. et al: Implementation and evaluation of protocols translating methods for IPv4 to IPv6 transition, Journal of Computer Science & Technology, vol. 12, no. 2, pp. 64-70, 2012
[17] Yu, S., Carpenter, B. E.: *Measuring IPv4 – IPv6 translation techniques*, Technical Report 2012-001, Department of Computer Science, The University of Auckland, January 2012
[18] Lencse, G., Répás, S.: *Performance analysis and comparison of the TAYGA and of the PF NAT64 implementations*, in 36th International Conference on Telecommunications and Signal Processing (TSP 2013), Rome, Italy, pp. 71-76, 2013
DOI: 10.1109/TSP.2013.6613894
[19] Lencse G., Takács, G.: *Performance Analysis of DNS64 and NAT64 Solutions*, Infocommunications Journal, vol. 4, no. 2, pp. 29-36, 2012
[20] Lencse G., Répás, S.: *Performance analysis and comparison of different DNS64 implementations for Linux, OpenBSD and FreeBSD*, in 27th IEEE International Conference on Advanced Information Networking and Applications (AINA-2013) Barcelona, Spain, pp. 877-884, 2013
DOI: 10.1109/AINA.2013.80