SCIENCE & MILITARY

V E D A A V O J E N S T V O

No 2 | Volume 2 | 2007

Vážení čitatelia,

dostáva sa Vám do rúk prvé periodikum SCIENCE&MILITARY monotematicky zamerané na problematiku komunikačných a informačných technológií. Vzniklo na základe výberu najvýznamnejších vedeckých a odborných príspevkov, ktoré odzneli na 4. medzinárodnej vedeckej konferencii "Komunikačné a informačné technológie - KIT 2007" v dňoch 3. - 5. októbra 2007 v Tatranských Zruboch.

Početné zastúpenie kompetentných odborníkov z Ministerstiev obrany Slovenskej a Českej republiky na konferencii potvrdilo, akú dôležitosť prikladajú implementácii informačných a komunikačných technológií v ozbrojených silách obidvoch krajín. Konferenciu pod záštitou ministra obrany Slovenskej republiky Mgr. Františka Kašického pripravila a zorganizovala Katedra informatiky Akadémie ozbrojených síl generála M. R. Štefánika v spolupráci s asociáciou dodávateľov a používateľov vojenských komunikačných a informačných systémov AFCEA Slovak Chapter.

Počas troch dní, v dvoch paralelne prebiehajúcich sekciách, v 64 vedeckých a odborných príspevkoch, ktoré odzneli, rezonovali otázky základného a aplikovaného výskumu, ale predovšetkým otázky implementácie komunikačných a informačných systémov v Ozbrojených silách Slovenskej republiky.

Pri výbere príspevkov do tohto čísla časopisu kolektív recenzentov pod vedením Dr.h.c prof. Ing. Jána Chmúrneho, DrSc. sa pokúsil vybrať a ponúknuť Vám výber 18 najzaujímavejších tém, ktoré odzneli v 7. odborných blokoch:

- *Digitálne spracovanie signálov*
- *E-learningové technológie a aplikácie*
- *Informačné technológie - technológie, technické prostriedky*
- *Informačné technológie - programovanie, programové prostriedky*
- *Komunikačné technológie - telekomunikačné siete a služby*
- *Komunikačné technológie - systémy a technológie*
- *Rozvoj a implementácia informačných technológií v ozbrojených silách*

Vážení čitatelia, prajem Vám príjemné chvíle pri čítaní tohto špeciálneho vydania časopisu s možnosťou rozšírenia si vedeckého a odborného pohľadu do oblasti vývoja najmodernejších komunikačných a informačných technológií.

Dear Readers,

You are receiving the first issue of SCIENCE&MILITARY focused monothematically on communication and information technologies. The issue contains a selection of most important scientific and specialist articles and contributions presented at the 4th international scientific conference "Communication and Information Technologies - KIT 2007" which was held October 3 – 5, 2007 in Tatranské Zruby, Slovakia.

Numerous representations of experts from the ministries of defense of the Slovak and Czech Republic confirmed the high degree of importance these ministries attach to the implementation of information and communication technologies in the Armed Forces of both countries. The Conference was organized under the auspices of the Minister of Defense of the Slovak Republic Mgr. František Kašický by the Department of Informatics of the Academy of the Armed Forces of general M. R. Štefánik in cooperation with AFCEA – Slovak Chapter.

During three days, in two parallel sessions, in 64 scientific and specialist articles presented, the problems and questions of primary and applied research as well as implementation of communication and information technologies in the Armed Forces of the Slovak Republic were dealt with.

During the process of selection of contributions for this issue of the journal the team of reviewers under the leadership of Dr.h.c. prof. Ing. Ján Chmúrny, DrSc. tried to choose 18 most interesting topics which were presented in 7 special sessions:

- *Digital signal processing*
- *E-learning technologies and applications*
- *Information technologies – technologies, hardware*
- *Information technologies – programming, software*
- *Communication technologies – telecommunication networks and services*
- *Communication technologies – systems and technologies*
- *Development and implementation of information technologies in the Armed Forces*

Dear Readers, enjoy this special issue of the journal, which can help you to enlarge you view of the development of up-to-date communication and information technologies.

Assoc. Prof. Ing. Marcel Harakaľ, PhD.
Head of the Department of Informatics of the AAF

Recenzenti / Reviewers:

*doc. Ing. Ľubomír **ANDRÁŠ**, PhD.*	*AOS Liptovský Mikuláš*
*Ing. Marián **BABJAK, PhD.***	*AOS Liptovský Mikuláš*
*Ing. Julius **BARÁTH**, PhD.*	*AOS Liptovský Mikuláš*
*doc. Ing. Pavel **ČIČÁK,** Phd.*	*STU Bratislava*
*prof. Ing. Ladislav **BUŘITA**, CSc.*	*UO Brno*
*doc. RNDr. Ľubomír **DEDERA**, PhD.*	*AOS Liptovský Mikuláš*
*Ing. Miroslav **ĎULÍK**, PhD.*	*AOS Liptovský Mikuláš*
*Ing. Peter **FUCHS**, PhD.*	*STU Bratislava*
*Ing. Milan **GOTTSTEIN**, PhD.*	*Veliteľstvo vzdušných síl OS SR Zvolen*
*doc. Ing. Marcel **HARAKAĽ**, PhD.*	*AOS Liptovský Mikuláš*
*Dr. h. c. prof. Ing. Ján **CHMÚRNY,** DrSc.*	*AOS Liptovský Mikuláš*
*doc. Ing. Ján **JAKUBEK**, CSc.*	*AOS Liptovský Mikuláš*
*prof. Ing. Jiří **JINDRA**, CSc.*	*VŠE Praha*
*doc. RNDr. Milan **LEHOTSKÝ**, CSc.*	*KU Ružomberok*
*prof. Ing. Dušan **LEVICKÝ**, CSc.*	*TU Košice*
*prof. Ing. Miroslav **LÍŠKA**, CSc.*	*AOS Liptovský Mikuláš*
*doc. Ing. Ján **KOLLÁR**, CSc.*	*TU Košice*
*prof. Ing. Jozef **ŠTULRAJTER**, CSc.*	*AOS Liptovský Mikuláš*
*prof. Ing. Stanislav **MARCHEVSKÝ**, CSc.*	*TU Košice*
*doc. Ing. Martin **MARKO**, CSc.*	*AOS Liptovský Mikuláš*
*prof. Ing. Igor **MOKRIŠ**, CSc.*	*SAV Bratislava*
*doc. Ing. František **NEBUS**, PhD.*	*AOS Liptovský Mikuláš*
*prof. Ing. Dušan **REPČÍK**, CSc.*	*ŽU Žilina*
*doc. Ing. Václav **ŽALUD**, CSc.*	*ČVUT Praha*

MODELING THE QUANTIZATION EFFECTS IN DIGITAL FILTERS VIA SPICE-FAMILY PROGRAMS

Dalibor BIOLEK, Viera BIOLKOVÁ, Zdeněk KOLKA

Abstract: The paper describes an unusual approach of employing the SPICE-family circuit simulation programs in a simple analysis of digital filters, including the quantization effects. For explanation and demonstration, the evaluation version of OrCad PSpice v.15.7 is used which is available free on the Internet.

Keywords: Digital filter, quantization, OrCad, PSpice, modeling, analysis.

1. INTRODUCTION

The design and subsequent analysis of digital filters is well supported by MATLAB combined with the Signal Processing Toolbox. In addition, owners of the Filter Design Toolbox can easily include quantization effects in the design. Additional features are enabled by exporting the designed realization structures into Simulink.

However, not everyone can avail themselves of such expensive software tools. That is why the paper is focused on the analysis of digital filters, including quantization effects, via the SPICE-family programs, namely via the evaluation version of OrCad PSpice from v. 15.7, which is freely available. Compared with MATLAB, another advantage is the graphical tool in the form of schematic editor, enabling easy drawing of filter realization structures with subsequent analysis of frequency responses from the input into an arbitrary node of filter structure. This feature is especially desirable for dynamic range optimization and scaling. The filter structure is created by connecting the blocks of delay, coefficient multiplication, and signal summation.

The techniques of digital filter analysis via Micro-Cap program are described in [1]. These procedures were transferable to the OrCad PSpice program only for idealized analysis, i.e. without considering the quantization effects. The reason was the absence of special mathematical functions for quantization modeling in the OrCad version 10.5 and olders. However, the set-up has changed beginning with version 15.7. In this paper, methods of modeling the quantization effects in OrCad PSpice v. 15.7 are described both for the AC and the transient analyses of digital filters.

2. MODELING THE QUANTIZATION EFFECTS IN ORCAD PSPICE V. 15.7

Starting with version 15.7, OrCad PSpice includes the CEIL, FLOOR, and INTQ functions for behavioral modeling. The FLOOR function returns an integer value of the argument [2]. It can be useful for modeling various methods of data quantization in fixed-point representation. Three of them are shown in Fig. 1 [3], [4].

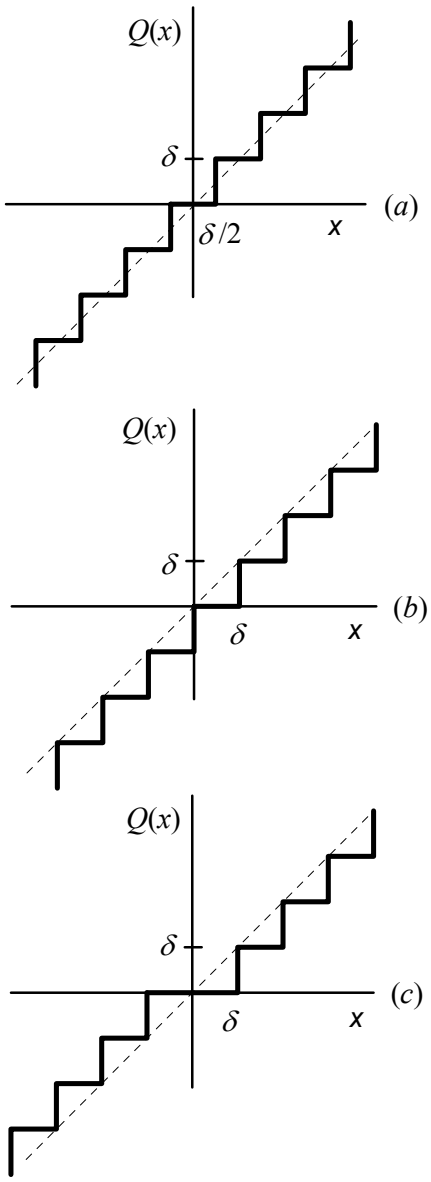


Fig. 1 Quantization methods [3]: (*a*) rounding, (*b*) two's complement truncation, (*c*) one's- complement and sign-magnitude truncation. Here $\delta = 2^{-N}$, N is the number of bits without the sign bit

The (*a*), (*b*), and (*c*) – type quantizations can be modeled in PSpice via FLOOR and other PSpice functions as follows:

$$Q(x) = \delta\ FLOOR(x/\delta + 0.5) \quad (1a)$$
$$Q(x) = \delta\ FLOOR(x/\delta) \quad (1b)$$
$$Q(x) = SGN(x)\delta\ FLOOR(ABS(x)/\delta) \quad (1c)$$

The above methods can be used for modeling the quantization of the filter coefficients, the input data, and the arithmetic rounding errors.

A concrete PSpice implementation by the user function ROUND can be as follows. In this demonstration, rounding method (*a*) is used:

```
.func round(x,N)
+{if(N<=0,x,2^(-N)*FLOOR(x*2^N+0.5))}      (2)
```

The number of bits N is defined by a global parameter. In Eq. (2), the user can discard the quantization after setting $N \leq 0$.

3. IMPLEMENTATION OF BASIC BLOCKS IN PSPICE

Since the basic circuit variables in the SPICE-family programs are voltages and currents, the summing block, the block of coefficient multiplication, and the block of delay by sampling period $T_S = 1/f_s$ can be implemented by controlled sources. For the popular PSpice program, the E-type controlled source can be applied:

Summing block:

```
Esum out 0 value={V(in1)+V(in2)}
```

Block of multiplying by constant A:

```
Emul out 0 in 0 {A}
```

Delay block:

```
Ez out 0 LAPLACE {V(in)} {exp(-s/fs)}
```

Both constant A, representing a concrete filter coefficient, and sampling frequency f_s must be defined as global variables by .param command.

Before the AC analysis of a digital filter compiled from the above blocks its input must be excited by an independent voltage source with the attribute AC = 1. After that, one will observe the frequency dependence of voltages at the outputs of respective blocks. With regard to the principal necessity of solving the linear model, it is necessary to disable the quantization of filter variables, except for the coefficients quantization. Full quantization can be used for the Transient analysis.

It is possible to model the above blocks as SPICE subcircuits and to assign them schematic symbols In CAPTURE. We get a powerful tool for digital filter analysis.

One possible form of simple PSpice library is given below:

```
*Digfil.lib - PSPICE library for digital filters
*analysis including the quantization effects

*quantization function, rounding
.func quant(x,N)
+{if(N<=0,x,2^(-N)*FLOOR(x*2^N+0.5))}
*quantization function, two's complement
*truncation
;.func quant(x,N)
+{if(N<=0,x,2^(-N)*FLOOR(x*2^N))}
*quantization function, ones' -complement and
*sign-magnitude truncation
;.func quant(x,N) {if(N<=0,x,
+SGN(x)*2^(-N)*FLOOR(ABS(x)*2^N))}

*summing block
*this subcircuit uses global parameter Nsum
*which MUST be defined in circuit file
.subckt SUM in1 in2 out params:Nsum={Nsum}
Esum out 0 value={quant(V(in1)+V(in2),Nsum)}
.ends

* multiplier
*this subcircuit uses global parameters Ncoef and
*Nmul which MUST be defined in circuit file
.subckt MUL in out
+params:coef=1 Ncoef={Ncoef} Nmul={Nmul}
Emul out 0
+value={quant(V(in)*quant(coef,Ncoef),Nmul)}
.ends

*delay block
*this subcircuit uses global parameter fs
*which MUST be defined in circuit file
.subckt DELAY in out
Ez out 0 LAPLACE {V(in)} {exp(-s/fs)}
Raux out 0 1T
.ends
```

4. DEMONSTRATION OF PSPICE SIMULATIONS

The following demonstration will be performed on the digital filter in Fig. 2 [1]. It is a cascade IIR 8th - order low-pass filter with a cut-off frequency of 3.4kHz and with a sampling frequency of 22.05kHz. The filter coefficients are given in the PSpice input file in **Appendix**. The results of AC analysis are in Fig. 3. Note that it is necessary to define filter coefficients by at least 9 bits in order to achieve therequired characteristic within the entire frequency range.

The filter impulse response as a result of Transient analysis is in Fig. 4. The filter coefficients are quantized to 9 bits. A comparison of two cases is given, for rounding free arithmetic operation (case (*a*)), and for 7-bit quantization of the data (case (*b*)). For the case (*b*), one can observe the parasitic limit

cycle operation. Note that the waveforms represent the continuous-time representation of discrete-time signal. That is why one must allow for the steady-state values within a concrete sampling action.

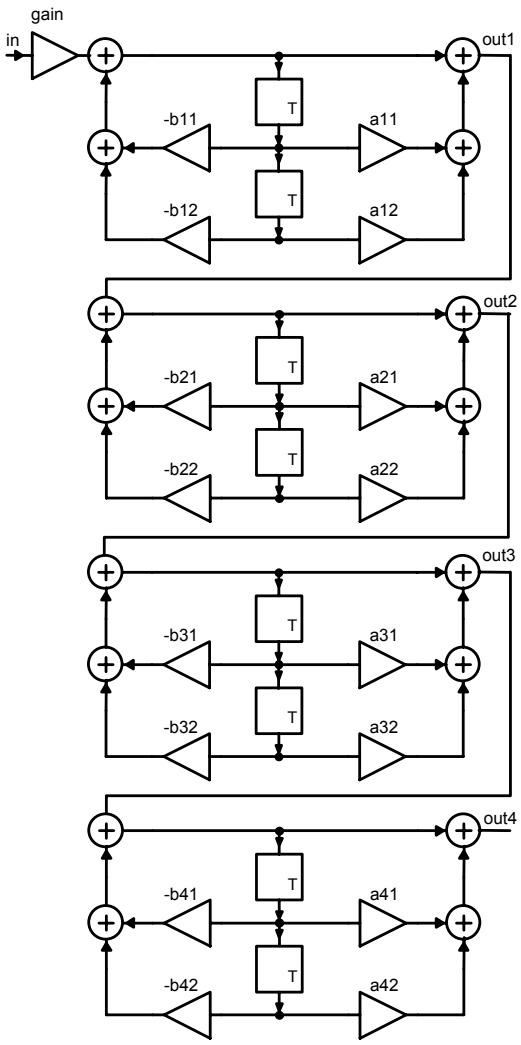


Fig. 2 IIR filter of LP type [1]

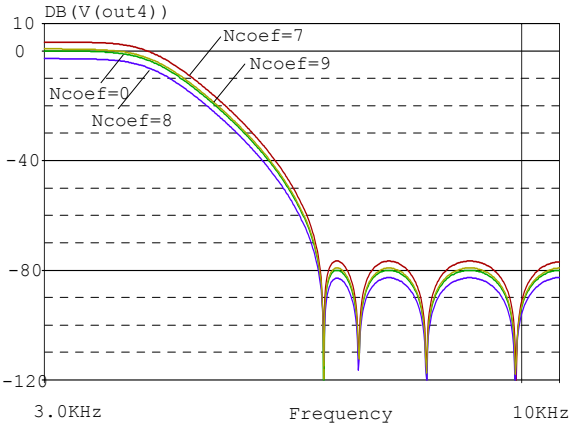


Fig. 3 Frequency responses of filter in Fig. 2 for several numbers of bits for coefficient quantization

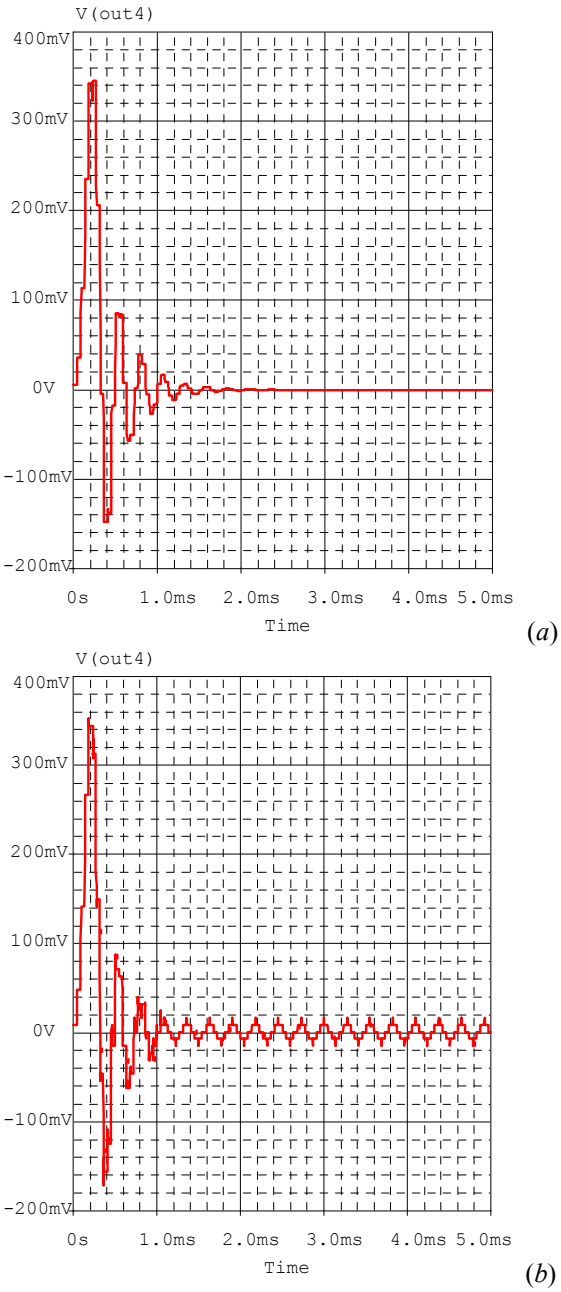


Fig. 4 Impulse responses of filter in Fig. 2 for N_{coef}=9 and (*a*) rounding free arithmetic operation, (*b*) for 7 bits quantization

5. ANALYSIS OF QUANTIZATION NOISE PROPAGATING TO THE FILTER OUTPUT

Regarding the quantization effects, in addition to coefficient quantization, the ratio of powers of output and input noises is frequently analyzed. Denote this ratio as *NNR* (Noise to Noise Ratio). It is useful to differentiate between two kinds of sources of quantization noise: noise at the filter input, generated by analog-to-digital converter, and quantization noise due to multiplying the digital

filter signals and filter coefficients. In the first case, the complete filter is placed between the noise source and the output. In the latter, the transfer path of noise starts at the output of the appropriate multiplier and leads to the filter output.

However, the methodology of *NNR* computation is identical in both cases. In the first step, we find the transfer function $K(z)$ between the source of the noise and the filter output. Then the *NNR* is evaluated by means of the formula [5]

$$NNR = \frac{1}{2\pi j}\oint_L z^{-1}K(z)K(z^{-1})dz = \sum_{z_p} res\{z^{-1}K(z)K(z^{-1})\} \tag{3}$$

The symbol L represents a curve of integration, specified as a circle of unity radius, centered in the origin of the z-domain complex plane. According to the residue theorem [6], the right-side sum represents the sum of residua of complex function within the braces {} in the poles of this function, which are located inside the curve of integration.

As proven in [7], for Mth-order digital FIR filter, the *NNR* is equal to the sum of squares of its coefficients h_i:

$$NNR = \sum_{i=0}^{M} h_i^2 \tag{4}$$

In the case of IIR filters or high-order FIR filters, the computation of *NNR* according to (4) can be complicated. In such cases, we can use another method, which starts from the frequency response of the transfer path between the source of noise and the filter output:

$$NNR = \frac{1}{f_s}\int_0^{f_s} |K(f)|^2 df = \frac{2}{f_s}\int_0^{f_s/2} |K(f)|^2 df \tag{5}$$

The frequency response can be easily obtained by AC analysis. After that, the simulation program must compute the mean square of the absolute value of frequency response within the frequency region from 0 to $f_s/2$ or from 0 to f_s, if appropriate.

An economical realization structure of a 15th-order linear-phase FIR filter is in Fig. 5 [1]. The sampling frequency is 11025Hz. The filter has 16 "symmetrical" coefficients. In Fig. 5, half of them are implemented:

a0= 0.011932832718619923,
a1= 0.0043636441772756003,
a2= -0.008189250603971928 4,
a3= 0.073487283106366652,
a4= 0.027003232751724886,
a5= -0.21586100424115826,
a6=-0.14953110748229956,
a7= 0.26565882334944552.

Fig. 6 summarizes the results of AC analysis in the frequency range up to one half of the sampling frequency. The upper curve illustrates the running average (represented by the „*avg*“ function) of the absolute value of the square of frequency response. According to (5), its value for the frequency fs/2 should be equal to the *NNR*. The value scanned from PROBE is 0.291776, which is – within the frame of all digits above – exactly the same result as those generated from the formula (4).

The method described is general, enabling a fast determination of *NNR* from an arbitrary node of the realization structure to the filter output.

6. CONCLUSION

The above facilities of digital filter analyses via the SPICE-family programs have been verified for many simulation problems. Additional details, e.g. how to simulate large digital structures via

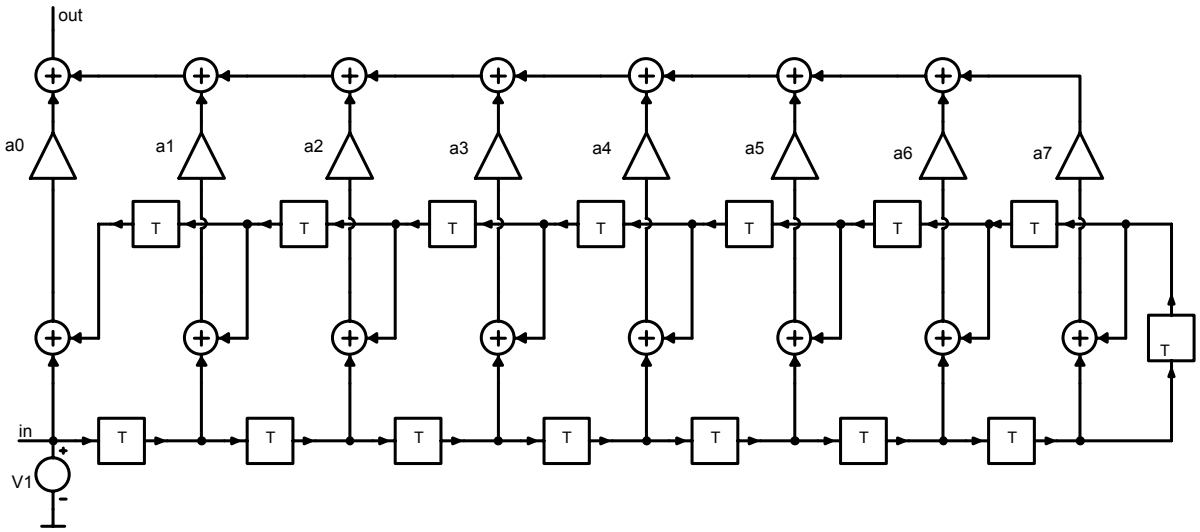


Fig. 5 15th-order bandpass FIR filter

evaluation version of Micro-Cap program, etc., are described in [1].

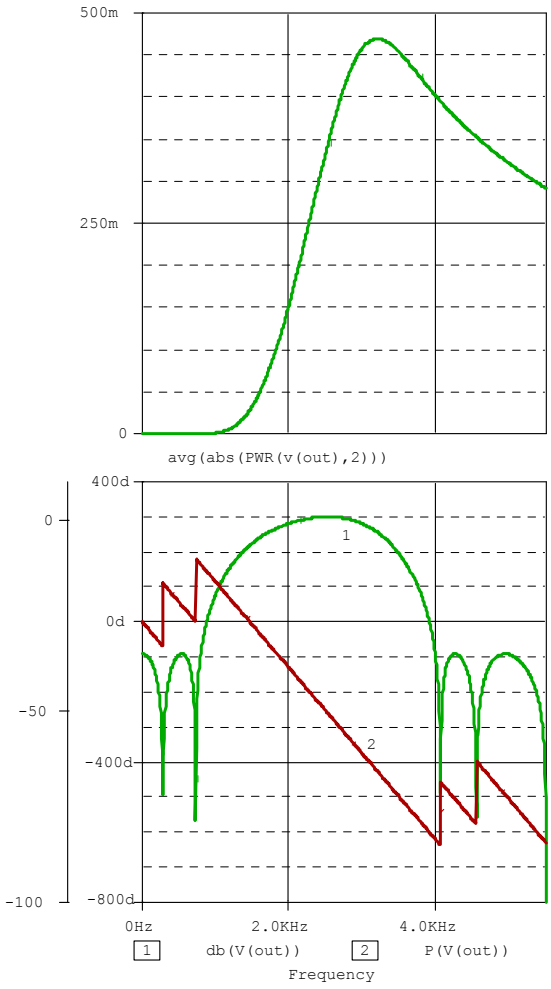


Fig. 6 Frequency responses of the filter from Fig. 5

ACKNOWLEDGMENT

This work is supported by the Grant Agency of the Czech Republic under grants No. 102/05/0771 and No. 102/05/0277, and by the research programmes of BUT MSM0021630503, MSM0021630513, and UD Brno MO FVT0000403.

References

[1] BIOLEK, D., ABUETWIRAT, I. F.: Analysis of digital filters via Spice-Compatible programs, 2006. Elektrorevue, 2006/26, ISSN 1213-1539, http://www.elektrorevue.cz/clanky/06026/english.htm.

[2] PSpice A/D Reference Guide. Product Version 15.7, July 2006.

[3] MITRA, S. K.: Digital Signal Processing. A Computer-Based Approach. Third Edition. McGraw Hill, 2006. ISBN 007-124467-0.

[4] DAVÍDEK, V., PŠENIČKA, B.: Tools for discrete-time signal processing. Exercises. ČVUT Praha, 1994. ISBN 80-01-01139-9 (In Czech).

[5] AHMED, N., NATARAJAN, T.: Discrete time Signals and Systems. Reston, Prentice Hall, 1983.

[6] Cauchy integral theorem and residue theorem: http://en.wikipedia.org/wiki/Residue_theorem

[7] ABUETWIRAT, I.F.: Finite Impulse Response Filter. Diploma work, Inst. of Microelectronics, BUT, 2006.

Appendix - PSpice circuit file: IIR 8-th order low-pass digital filter, full quantization

```
*
.options ITL4 100
.param
+fs 22.05kHz N 7 Ncoef 0 Nsum {N} Nmul {N}
.param a11 1.8880173933198123
+a12 0.99999999999998335
+b11 -0.32655932340433669
+b12 0.039446152567426181
+a21 1.2426043302792451
+a22 1.0000000000000131
+b21 -0.4147385903741958
+b22 0.15472564825866436
+a31 0.62610036993806006
+a32 0.99999999999999201
+b31 -0.57140840958280426
+b32 0.38161259426751648
+a41 0.3148879700393597
+a42 0.99999999999999423
+b41 -0.77820865449734911
+b42 0.74175331275429501
.param gain  0.0053733804257323015

;Vin in 0 DC 1 AC 1 ;input signal for step
                        * response and AC analysis
Ein in 0 value={stp(time)-stp(time-1/fs)}  ;input
                      *signal for impulse response
Xgain in in1 MUL params:coef={gain}
*stage No. 1
Xz11 11 12 DELAY
Xz12 12 13 DELAY
Xb11 12 m11 MUL params:coef={-b11} ;the
                                  *feedback path
Xb12 13 m12 MUL params:coef={-b12}
Xsum11 in1 s11 11 SUM
Xsum13 m11 m12 s11 SUM
Xa11 12 m13 MUL params:coef={a11}  ;the
                                   *forward path
Xa12 13 m14 MUL params:coef={a12}
Xsum12 11 s12 out1 SUM
```

```
Xsum14 m13 m14 s12 SUM
*end of stage No. 1
*stage No. 2
Xz21 21 22 DELAY
Xz22 22 23 DELAY
Xb21 22 m21 MUL params:coef={-b21} ;the
                                    *feedback path
Xb22 23 m22 MUL params:coef={-b22}
Xsum21 out1 s21 21 SUM
Xsum23 m21 m22 s21 SUM
Xa21 22 m23 MUL params:coef={a21}  ;the
                                     *forward path
Xa22 23 m24 MUL params:coef={a22}
Xsum22 21 s22 out2 SUM
Xsum24 m23 m24 s22 SUM
*end of stage No. 2
*stage No. 3
Xz31 31 32 DELAY
Xz32 32 33 DELAY
Xb31 32 m31 MUL params:coef={-b31} ;the
                                    *feedback path
Xb32 33 m32 MUL params:coef={-b32}
Xsum31 out2 s31 31 SUM
Xsum33 m31 m32 s31 SUM
Xa31 32 m33 mul params:coef={a31}  ;the forward
                                              *path
Xa32 33 m34 mul params:coef={a32}
Xsum32 31 s32 out3 SUM
Xsum34 m33 m34 s32 SUM
*end of stage No. 3
*stage No. 4
Xz41 41 42 DELAY
Xz42 42 43 DELAY
Xb41 42 m41 MUL params:coef={-b41} ;the
                                    *feedback path
Xb42 43 m42 MUL params:coef={-b42}
Xsum41 out3 s41 41 SUM
Xsum43 m41 m42 s41 SUM
Xa41 42 m43 MUL params:coef={a41}  ;the
                                     *forward path
Xa42 43 m44 MUL params:coef={a42}
Xsum42 41 s42 out4 SUM
Xsum44 m43 m44 s42 SUM
*end of stage No. 4

;.step param Ncoef list 0 7 8 9
;.AC dec 1000 10 11.025k
.tran 0 5m skipbp
.probe
.inc digfil.lib
.end
```

Summary: The paper shows that digital filters can be successfully analyzed via programs, which are primarily designated for simulating the analog circuits, namely the SPICE-family circuit simulation programs. By means of techniques of behavioral modeling, we can include also the quantization effects into the models of digital filters.

The free evaluation version of OrCad PSpice v.15.7 is used for the implementation. However, one can utilize other similar SPICE-like programs with proper internal mathematical tools for modelling the quantization effects.

prof. Ing. Dalibor BIOLEK, CSc.[1)]
Ing. Viera BIOLKOVÁ[2)]
doc. Dr. Ing. Zdeněk KOLKA[2)]
[1)] UO Brno, Katedra elektrotechniky, Kounicova 65
612 00 Brno, Česká republika
E-mail: dalibor.biolek@unob.cz
[2)] FEKT VUT Brno, Ústav radioelektroniky, Purkyňova
118 612 00 Brno, Česká republika
E-mail: biolkova@feec.vutbr.cz
kolka@feec.vutbr.cz

INTEROPERABILITA V INFORMAČNÍCH SYSTÉMECH

INTEROPERABILITY IN INFORMATION SYSTEMS

Ladislav BUŘITA

Abstract: In the introduction, the meaning of the term "interoperability" is analyzed. The requirements and recommendations for achieving the information systems (IS) interoperability are quoted from the European Interoperability Framework and the recommendation of the US Office of Electronic Government. Predominantly, the aspects of accessibility, multilingualism, security, privacy, subsidiarity, the use of open standards, the assessment of the benefits of open source software, and the use of multilateral solutions are highlighted. It is recommended to use the business-centric methodology, and move to standard mechanisms. To achieve the information systems interoperability, it is necessary to apply the architectural approaches, particularly the implementation of NATO Architecture Framework in the military environment. When constructing an interoperable IS, it is suitable to use metadata, and the model-driven and service-oriented architecture. The current research in IS interoperability is oriented to the semantic WEB, ontologies, information modeling, and knowledge bases.

Keywords: interoperability, information system, European interoperability framework, architecture, metadata, ontology.

1. ÚVOD

Pojem interoperability je velmi široký, jen v pomůckách NATO je zaznamenáno několik definic. V AAP-06 Terminologický slovník pojmů a definic NATO (3 definice):

(1) Schopnost působit ve vzájemné podpoře při plnění stanovených úkolů.

(2) Schopnost ozbrojených sil společně efektivně cvičit a působit při plnění stanovených úkolů.

(3) Schopnost sil dvou nebo více zemí vést výcvik, provádět společná cvičení a společně působit při plnění stanovených úkolů.

V ADatP-02 Slovník NATO - Informační technologie (2 definice):

(1) Schopnost systému, jednotek nebo sil poskytnout služby a přijmout služby od jiných systémů, jednotek nebo sil a použít takto získané služby ke zvýšení efektivity společné činnosti.

(2) Schopnost komunikovat, zpracovávat programy nebo přenášet data mezi různými funkčními jednotkami způsobem, který od uživatele vyžaduje pouze malé nebo žádné znalosti o specifických vlastnostech těchto jednotek.

„Interoperabilita je, v kontextu rozsáhlé aplikace, schopnost systému/produktu spolupracovat s jinými systémy/produkty bez zvláštního úsilí zákazníka/uživatele. Je to schopnost interakce a výměny informací jak uvnitř, tak vně organizace. Její dosažení je jedním ze základních požadavků na systémy podnikové a státní sféry", viz preambule konference IESA-2007 [1].

Interoperabilita, tj. schopnost interakce mezi organizacemi a jejich informačními systémy (IS), musí být řešena minimálně ve třech oblastech/úrovních: DATA, APLIKACE, ORGANIZACE. Nejde tedy pouze o problém informačních technologií (IT), ale jde rovněž o komunikaci a organizační záležitosti. Naopak v otázkách interoperability hrají právě otázky organizační a zejména politické vůle rozhodující roli.

V článku jsou zohledněny i poznatky, které autor získal na mezinárodních konferencích IESA-2007 [1] a ICCRTS-2007 [2], jichž se autor účastnil.

2. POŽADAVKY A DOPORUČENÍ K DOSAŽENÍ INTEROPERABILITY

Jednou z cest, kterou lze dosahování interoperability ovlivnit, je usměrňování dodavatelů, řešitelů, odběratelů, uživatelů a ostatních skupin vydáváním doporučení. Zveřejní se opatření a zásady, které bychom měli respektovat, abychom se snáze k interoperabilitě dobrali. V článku jsou citovaná doporučení pro European Interoperability Framework a US Office of Electronic Government.

2.1 European Interoperability Framework

Dokument Evropský rámec interoperability (European Interoperability Framework – EIF) [3] specifikuje doporučení k dosažení organizační, sémantické a technické interoperability. Navrhuje principy evropské spolupráce ve službách, které by měly zajistit kvalitní fungování eGovernmentu.

Na základě rozboru politických a organizačních předpokladů, na základě analýzy technických a technologických možností informačních a komunikačních technologií (ICT), vydává EIF 17 doporučení:

(1) Administrativa členských států, instituce a agentury Evropské unie (EU) by měly vycházet při tvorbě vlastního interoperabilního rámce evropských služeb eGovernmentu z EIF.

(2) Evropské služby eGovernmentu by měly respektovat principy dostupnosti (Accessibility), vícejazyčnosti (Multilingual), bezpečnosti (Security), ochrany soukromí (Privacy), decentralizace (Subsidiarity), využití otevřených standardů (Use of Open Standards), posouzení přínosu volně dostupného SW (Assess the benefits of Open Source Software) a aplikaci multilaterálních řešení (Use of Multilateral Solutions).

(3) Nasazení služeb eGovernmentu je třeba uvažovat v kontextu organizační, sémantické a technické interoperability.
(4) Požadavky na evropské služby eGovernmentu je třeba posuzovat komplexně, což by mělo vést k jejich správné identifikaci a prioritizaci.
(5) Veřejná správa, která předpokládá implementaci služeb eGovernmentu, by měla analyzovat relevantní procesy. Měla by schválit potřebná interoperabilní rozhraní (Business Interoperability Interfaces-BII).
(6) Pokud má být evropská služba eGovernmentu řešena více státy, je třeba formalizovat výsledná očekávání v podobě smlouvy, například jako Service Level Agreement (SLA). SLA by měla zahrnout dotčená BII a navíc by se měla odsouhlasit bezpečnostní politika služby.
(7) Pro interoperabilní výměnu datových prvků v evropských službách eGovernmentu je třeba odpovědnými správci na národní úrovni:
- *Zveřejnit informaci o odpovídajících datových prvcích, jež výměnu zahrnují.*
- *Zaslat na evropskou úroveň k odsouhlasení data a příslušné datové slovníky. Přitom je třeba vycházet ze společných datových prvků všech služeb.*
- *Předat na evropskou úroveň k odsouhlasení multilaterální mapování tabulek na schválené datové prvky.*

(8) Z hlediska sémantické interoperability je třeba dbát na lingvistické trasování speciálních právních slovníků, jež jsou ve službách využity. Právní a sociální rámec EU předpokládá ekvivalenci směrnic.
(9) Evropské iniciativy směřující k vyvinutí společné sémantiky na základě XML by měly být realizované koordinovaně a v kooperaci se standardizačními organizacemi. Přitom je třeba respektovat schválené základní datové prvky (Core Data Elements).
(10) Do uživatelského přístupu ke službám (Front-office Level) se předpokládá z pohledu technické interoperability zahrnout:
- *Prezentaci a výměnu dat (Data Presentation and Exchange).*
- *Přístupové rozhraní (Accessibility – Interface Design Principles).*
- *Multikanálový přístup (Multi-channel Access).*
- *Znakové sady (Character Sets).*
- *Typ souboru a formáty dokumentů (File Type and Document Formats).*
- *Kompresi souboru (File Compression).*

(11) Do pozadí fungování služeb (Back-office Level) se předpokládá z pohledu technické interoperability zahrnout:
- *Integraci dat a rozhraní (Data Integration and Middleware).*
- *Standardy na bázi XML (XML-based Standards).*
- *Standardy na bázi EDI (EDI-based Standards).*
- *Webové služby (Web Services).*
- *Distribuovanou aplikační architekturu (Distributed Application Architecture).*
- *Komunikační služby (Interconnection Services).*
- *Protokoly přenosu souborů a zpráv (File and Message Transfer Protocols).*
- *Přenos zpráv a bezpečnost (Message Transport and Security).*
- *Služby ukládání zpráv (Message Store Services).*
- *Elektronickou poštu (Mailbox Access).*
- *Adresářové služby (Directory and Domain Name Services).*
- *Síťové služby (Network).*

(12) Aspekty bezpečnosti zajistit na úrovni:
- *Služby bezpečnosti (Security Services).*
- *Obecné služby bezpečnosti – PKI (General Security Services).*
- *Webové bezpečností služby (Web Service Security).*
- *Firewalů (Firewalls).*
- *Ochrany proti virům, červům, trojským koňům a e-mailovým bombám (Protection Against Viruses, Worms, Trojan Horses and E-mail bombs).*

(13) Administrativa členských států, instituce a agentury EU by měly vyvinout a používat společná pravidla technické interoperability evropských sítí, aplikací a služeb eGovernmentu. Výchozí pravidla byla vyvinuta v organizaci Interoperable Delivery of pan-European Services to Public Administrations, Businesses and Citizens (IDABC).
(14) Společná pravidla technické interoperability by měla být vytvořena s respektem na otevřené standardy.
(15) Organizacím a občanům umožnit vkládat zprávy do evropských služeb ve svém mateřském jazyce. Další alternativou je psaní v omezeném rozsahu jazyků (tím je míněna angličtina, francouzština a němčina).
(16) Ty evropské služby, které jsou poskytované přes portál, musí být na nejvyšší úrovni portálu plně mnohojazyčné, stránky druhé úrovně mohou být přístupny v oficiálních jazycích a vnější odkazy s příslušnými národními webovými stránkami mohou být již kromě národního jazyka ještě alespoň v jednom dalším jazyce (například angličtině).
(17) V ostatních případech (než doporučení 16) by měl být k dispozici překladatelský SW, který zajistí hrubý překlad do požadovaného jazyka.

2.2 Doporučení pro interoperabilitu US Office of Electronic Government

Následující doporučení US Office of Electronic Government [4] jsou navržena s cílem podporovat interoperabilitu:

- Business-Centric Methodology (Metodika orientovaná na business)

Tato metodika podporuje sémantickou a pragmatickou interoperabilitu, zaměřuje se na hlavní problémy, na symptomy integrace. Důraz je na business experty s významným zapojením zájmových osob (Communities of Interests - COI). Používá otevřené deklarační postupy, tím se zvyšuje srozumitelnost, přístupnost a umožňuje uživatelské přizpůsobení slovníků a modelů v heterogenním prostředí. Izoluje business přístupy od technologií, snižuje komplexnost problematiky členěním do úrovní.

- Move to Standard Mechanisms (Přejděte na standardní mechanismy)

Vyvarujte se nestandardní syntaxi dat. Existuje nepřeberné množství možností popisu dat, každá z nich má svoje silné a slabé stránky. Z hlediska interoperability může působit těžkosti zejména konverze dat. V aktuálních síťových aplikacích lze doporučit dvě řešení syntaxe dat:

(1) Mezinárodní standard Abstract Syntax Notation (ASN.1).
(2) Průmyslový standard Extensible Markup Language (XML).

Zaznamenejte sémantiku sdílených datových prvků. Rozhraní mezi participujícími systémy eGovernmentu zpravidla obsahují množství datových prvků. Většinou není jednoduché jejich význam pochopit, zejména pokud je syntaxe uvedena ve schématu XML či definici ASN.1. Vhodnější je forma datového slovníku podle mezinárodního standardu ISO/IEC 11179, Information Technology - Metadata Registries (Informační technologie – Slovníky metadat).

Dokumentujte služby rozhraní standardním způsobem. Kromě syntaxe a sémantiky datových prvků v rozhraní je třeba

popsat funkce rozhraní, tzn. jak systémy spolu komunikují. Jedním ze způsobů je popis jazykem definování rozhraní (Interface Definition Language - IDL); například IDL pro architekturu CORBA (Common Object Request Broker Architecture). Podniková sféra se spíše orientuje na WSDL (Web Services Definition Language) nebo ebXML (Electronic Business XML). Dalším všeobecně přijímaným nástrojem je UML.

Podporujte služby rozhraní pro přístup k informacím (Information Discovery), což je proces vyhledání relevantních dat a informačních zdrojů bez prvotní znalosti o těchto zdrojích. Takové služby rozhraní jsou implementovány dle mezinárodního standardu ISO 23950 Protocol for Information Search and Retrieval (Protokol pro vyhledávání informací a přístup k nim). Protokol je běžně používaný v tradičních knihovních systémech. Služba podporuje syntaxi dat dle XML a ASN.1; sémantika dat je zaznamenaná v ISO sémantickém registru (ISO Basic Semantics Register); služba rozhraní je dočasně v CORBA IDL a WSDL a je publikovaná v UDDI (Universal Description, Discovery, and Integration).

- Provide Infrastructure for Visibility (Poskytněte přehled o infrastruktuře)

Služby registru obsahují informace o struktuře, formátu a definici dat. Zpravidla se jedná o databázovou aplikaci, která umožňuje v datech vyhledávat a zjišťovat vztahy mezi nimi. Takto je k dispozici přehled o datech a má podobu metadat. Kritickým aspektem sady spolupracujících služeb je přiřazení slovníkových významů pojmům metadat (Aligning Vocabularies Around Concepts) a umožnění uživatelům v nich navigovat.

- COI – Communities of Interest (Zájmové skupiny)

Zájmové skupiny (Communities of Interest - COI) jsou spolupracující skupiny uživatelů, které sdílejí společné cíle, zájmy, procesy a pracují odsouhlaseným způsobem. Tvoří se institucionálně či účelově. Institucionální COI spolupracují dlouhodobě a mají za realizované operace odpovědnost, poskytují podporu náhodným situacím a krizovým událostem. Účelové COI jsou přechodné, sestavované ad-hoc se zaměřením na náhodné situace a krizové události. Infrastruktura pro spolupráci musí vyhovovat oběma skupinám COI.

3. ASPEKTY DOSAŽENÍ INTEROPERABILITY

Na vytvářené IS jsou kladeny stále vyšší nároky na stabilitu, bezpečnost, použitelnost či otevřenost. K uspokojení požadavků na interoperabilitu je zapotřebí navrhovat systémy s využitím moderních přístupů a ověřených poznatků. Jednou z klíčových dovedností při tvorbě IS je správné rozhodnutí o architektuře takového systému, volba vhodné metodiky a technologie. Rozvoj IS je stále více ovlivňován znalostními přístupy.

3.1 Architektury

S architekturami se lze setkat na různé úrovni abstrakce. Nejvyšší abstrakci v oblasti IS je Enterprise Architecture. Přičemž Enterprise v tomto sousloví znamená souhrn organizací, které mají společné cíle a jsou postaveny na stejném základě. Pod pojmem enterprise si lze představit organizaci, podnik nebo jeho součást (divizi), popř. oddělení nebo také skupinu geograficky oddělených organizací mající společného vlastníka. Termín Architektura je definován různě, pro naše potřeby vyjdeme ze standardu ANSI/IEEE 1471-2000, kde je architektura definovaná jako „Uspořádání systému vyjádřené v jeho komponentách, vzájemných vztazích těchto komponent a vztazích k okolí a popsané principy určujícími jeho návrh a další rozvoj“.

Enterprise architecture je rovněž komplexním rámcem – Comprehensive Framework, který umožní zvládnutí a pochopení struktury a chodu organizace, jejich procesů a informací při respektování celkové strategie. Architekturní rámec je nástrojem pro tvorbu široké škály různých architektur, měl by popsat metody návrhu IS jako soustavu bloků, k tomu by měl poskytnout sadu technik a nástrojů a společný slovník. Dále by měl obsahovat seznam doporučených standardů a sadu povolených produktů, které mohou být použity pro výstavbu těchto bloků.

Enterprise architektura je výchozí pro podřízené architektury, zejména:

- Organizační/business architektura - definující strategii, řízení, organizaci a klíčové procesy.
- Architektura aplikační/softwarovou - poskytující detailní pohled na aplikace, které budou nasazeny, jejich vzájemnou komunikaci a vztah ke klíčovým procesům.
- Technologická architektura - specifikující softwarovou a hardwarovou infrastrukturu, která bude sloužit k nasazení klíčových aplikací organizace.
- Informační/datová architektura - popisující strukturu a organizaci fyzických a logických datových prvků v organizaci a řízení datových zdrojů.

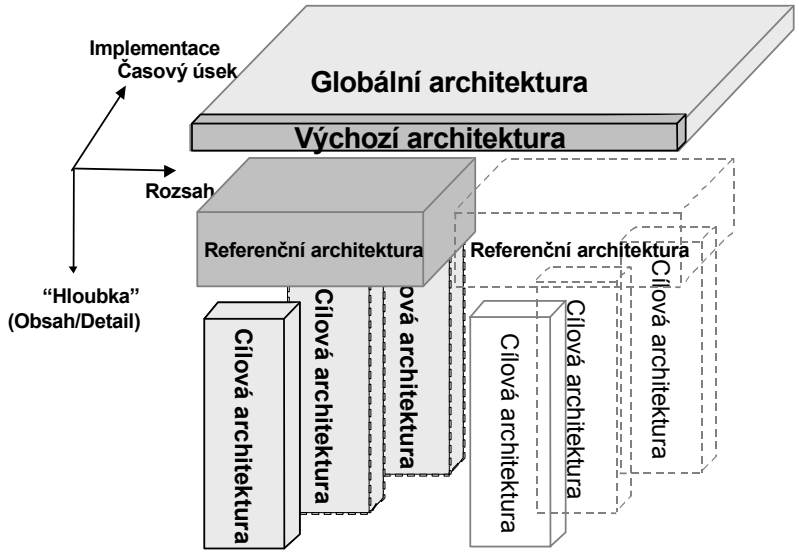
Cílem takto široce chápaného pojmu architektury je udržení konzistence mezi jednotlivými architekturami. Na druhou stranu rozčlenění architektur je nezbytné z důvodu značné složitosti problematiky, zde je nutná specializace a zaměření na dílčí oblasti.

3.2 NATO Architecture Framework

Pro tvorbu C3 (Consultation, Command and Control) systémů a pro NEC (Network Enabled Capability) vydalo NATO Architecture Framework (NAF). NAF specifikuje požadavky na popis C3 systémů s cílem dosažení jejich interoperability. Podle ní je v resortu MO ČR vydána Metodika řízení tvorby architektur C3 systému [5].

NAF zahrnuje čtyři základní architektury, které jsou uspořádané dle logického postupu jejich tvorby - Výchozí architektura (VA), Globální architektura (GA), Referenční architektura (RA) a Cílová architektura (CA), viz Obr. 1. Každá z architektur je definovaná ze tří pohledů - operačního, systémového a technického. Očekává se rozšíření NAF k zahrnutí

servisního přístupu (Service-oriented). Operační pohled je uživatelským pohledem a lze jím získat přehled o prvcích systému, jejich funkcích a chování a je nezávislý na technologii. Systémový pohled vymezuje hranice systému a popisuje jeho strukturu a chování. Technický pohled vychází z přijatých standardů a vymezuje uspořádání technologických komponent systému.



Obr. 1 Základní typy architektur v NAF a vztahy mezi nimi

Architektury se tvoří pomocí formalizovaných šablon (modelů, vzorů, schémat). Z hlediska interoperability je třeba, aby se šablony maximálně shodovaly s běžně aplikovanými vzory. Vhodná podoba šablon vychází především ze standardu UML (Unified Modelling Language).

3.3 Metadata, metainformační systém

Metadata, neboli data o datech, popisují, objasňují, lokalizují či jinak usnadňují přístup k datům, jejich pochopení, spravování a využití. Pojem metadata je chápán v různých komunitách rozdílně. Popisují informační zdroje a jsou uplatněna pro počítačové porozumění informacím. Metadata jsou členěna na:

(1) Popisná (Descriptive) – popisují informační zdroje pro identifikaci a usnadnění vyhledávání. Mohou obsahovat prvky jako název, abstrakt, autoři, klíčová slova.
(2) Strukturní (Structural) – specifikují, jak jsou objekty složeny.
(3) Administrativní (Administrative) – týkají se řízení práce s informačními zdroji (přístupová práva, ochrana dat a metadat, technické detaily o datech).

Metadata popisují relační databáze (Relational Database), datové sklady (Data Warehouse), souborové systémy (File System), obrazová data (Image) a počítačové programy (Program). Metadata významně pomáhají v uspořádání elektronických zdrojů, usnadňují jejich využití a integraci, poskytují digitální identifikaci a podporují ochranu a archivaci. Popis datových zdrojů pomocí metadat vede k jejich porozumění jak lidmi tak počítači. Metadata jsou zpravidla ukládána do centrální databáze, nazývané Datový slovník (Data Dictionary) či Metadatový registr (Metadata Registry), což v organizaci vede ke standardizaci dat. V resortu obrany ČR byl takový registr vytvořen ve spolupráci katedry KIS FVT/UO Brno a Agentury rozvoje informatiky v rámci výzkumného záměru [6].

Registr je informační podporou systému distribuce a správy datových prvků (jednotka dat, která je v daném kontextu považovaná za nedělitelnou) a číselníků (seznam přípustných hodnot datového prvku, obvykle ve formě dvojice kódovaného údaje a hodnoty jeho kódu) mezi jednotlivými IS resortu MO. Je katalyzátorem integrace a prostředkem dosažení interoperability IS. V předpokládaném Průřezovém IS (PRIS) MO zabezpečí centrální správu a jednotné úložiště metadat. Je provozován v celoarmádní datové síti za využití některých služeb serverů ŠIS (Štábní informační systém).

Při analýze a návrhu se uplatnil procesní a architekturní přístup (nástroje Enterprise Architect a UML). Byly popsány procesy, specifikovány role v systému (gestor, správce, garant, uživatel), popsána datová struktura a funkce systému. Registr předpokládá propojení na IS veřejné správy ČR a na NATO (NDAG – NATO Data Administration Group), kde byl i presentován a získal uznání.

3.4 Modelem řízená a servisně orientovaná architektura

Modelem řízená architektura (Model-Driven Architecture – MDA) specifikuje 4 úrovně návrhu a realizace SW:

(1) Model nezávislý na počítači (Computational-Independent Model - CIM).
(2) Model nezávislý na počítačové platformě (Platform-Independent Model - PIM).
(3) Model pro vybranou počítačovou platformu (Platform-Specific Model - PSM).
(4) Zdrojový kód SW aplikace.

Modely se transformují pokud možno automaticky podle standardu QVT (Queries/ Views/ Transformations). MDA je podpořena dalšími standardy, jako jsou UML, Meta-Object Facility, XML Metadata Interchange, Enterprise Distributed Object Computing, Software Process Engineering Metamodel a Common Warehouse Metamodel.

Jedním z hlavních cílů MDA je oddělit návrh od architektury a dosáhnout tak lepší přenositelnosti SW aplikací a dobrého zvládnutí komplexnosti řešeného projektu. MDA je přístup systémového inženýrství, ve kterém jsou modely užity k pochopení SW projektu; jeho návrhu, tvorbě a implementaci, provozování, údržbě a modifikaci.

Servisně orientovaná architektura (Service-Oriented Architecture – SOA) je tvořena sadou služeb, které spolu komunikují. Komunikace může vyvolat buď předání dat nebo může zahrnout více služeb, které spolu koordinují nějakou aktivitu. Služba je dobře definovanou samostatnou funkci, která je opakovatelně použitelná, modulární a dobře ohraničená, odpovídá standardům a tím je interoperabilní.

Stavebními prvky SOA jsou webové služby (WS), tedy bloky SW, jsou dosažitelné v Internetu (intranetu) a jsou popsané standardizovaným způsobem. Automatické skládáním služeb, kdy se kombinuje funkcionalita opakovaně využitelných WS, je významnou výzvou pro praktické využití SOA. Orchestrace WS vyjadřuje kontinuální řízený tok volání WS, které lze popsat např. jazykem WS-BPEL (Business Process Execution Language). Služby operují podle definice WDSL – Web Services Definition Language, která je nezávislá na platformě a programovacím jazyku. Za interoperabilním rozhraní služby je skryta implementace služby v prostředí J2EE nebo .NET. Služby napsané v C++ jsou provozovány v platformě .NET a napsané v jazyce Java na platformě J2EE. Pro MDA byl definován v projektech ATHENA a INTEROP metamodel PIM4SOA, což je platformě nezávislý modelovací jazyk k podpoře SOA.

Úvahy nad praktickým uplatněním SOA nelze vést naivním způsobem. Nejedná se pouze o role tvůrce, poskytovatele a příjemce služeb s registrem UDDI (Universal Description, Discovery, and Integration). Je nutno vyřešit politiku znovupoužití a provozování služeb, jejich bezpečnost, pečlivý popis služeb v datovém slovníku, jejich testování, stanovení garanta za službu apod.

Většinou se hovoří pouze o přínosech SOA k zajištění flexibility, transparentnosti, interoperability a rychlé tvorby SW aplikací. Ale již se méně zdůrazňuje, že komplexnost takto vytvořeného SW se nesníží, naopak dojde k značnému zvýšení. Dále je nezbytné, aby byla garantovaná kvalita služeb, jejich řízení a správa, což nelze bez silného centrálního přístupu zvládnout. Služby je třeba monitorovat, rozvíjet a přijímat potřebná opatření.

Možné řešení nabízí například HP SOA Centres [7], sada SW a rámce pro zvládnutí technologie (Governance Interoperability Framework). Přináší řešení pro tvorbu politik SOA a jejich praktické vynucování, poskytuje datový slovník (System of Records), nástroje životního cyklu služeb v SOA. Umožňuje kvalitní testování a monitorování služeb, zajištění jejich dostupnosti. Uvádí použitelnost SOA z roviny teorie do praktického uplatnění.

3.5 Sémantický WEB a ontologie, informační modelování a znalostní báze

Sémantický web představuje publikovaná data na webu, s příslušným kontextem, se strojově čitelnou informací o jejich významu. Idea sémantického webu vychází z potřeby dát obsahu webu jasný smysl a učinit zde dostupná data srozumitelná strojům. Stroje budou k tomuto účelu využívat popisy informací uvedené na webu, definované slovníky a ontologie. Ontologie je myšlena jako vymezení konceptu, tedy zařazení slova do širších souvislostí - příkladem takového vymezení může být popis významu slov ve výkladovém slovníku. K dokumentům na webu a k dalším internetovým zdrojům by měly být připojeny informace, které budou pro počítače představovat podklad pro vyvozování vztahů mezi těmito informačními zdroji.

Předpovídat, jaké služby může sémantický web poskytovat je stejně obtížné, jak bylo předpovídání v době zrodu webu, kam se vyvine. Celá oblast je objektem rozsáhlého výzkumu, sponzorovaného například z Evropské unie. Velmi obtížným

aspektem budování sémantického webu je vytvoření vhodných ontologií. Tento proces vyžaduje úsilí mnoha rozdílných komunit. Jen tak bude možno vytvořit obecné slovníky, které využijí systémy k rozpoznání obsahu webového dokumentu. Naštěstí vytváření ontologií nevyžaduje globální koordinaci, lze je sestavovat dle přijatých standardů a schémat s využitím jednotných jazyků nezávisle.

Ontologie popisuje všeobecně využitelným způsobem poznanou znalost tak, aby mohla být využita a sdílena v mnoha oblastem různými skupinami uživatelů. Tvoří rámec pro unifikaci různých hledisek pohledu na informace, obsahuje pojmy, vztahy a pravidla pro kombinaci pojmů. Smyslem ontologie je zdokonalení komunikace mezi počítači, mezi lidmi a počítači a mezi lidmi navzájem. Ontologie se z historického hlediska člení na:

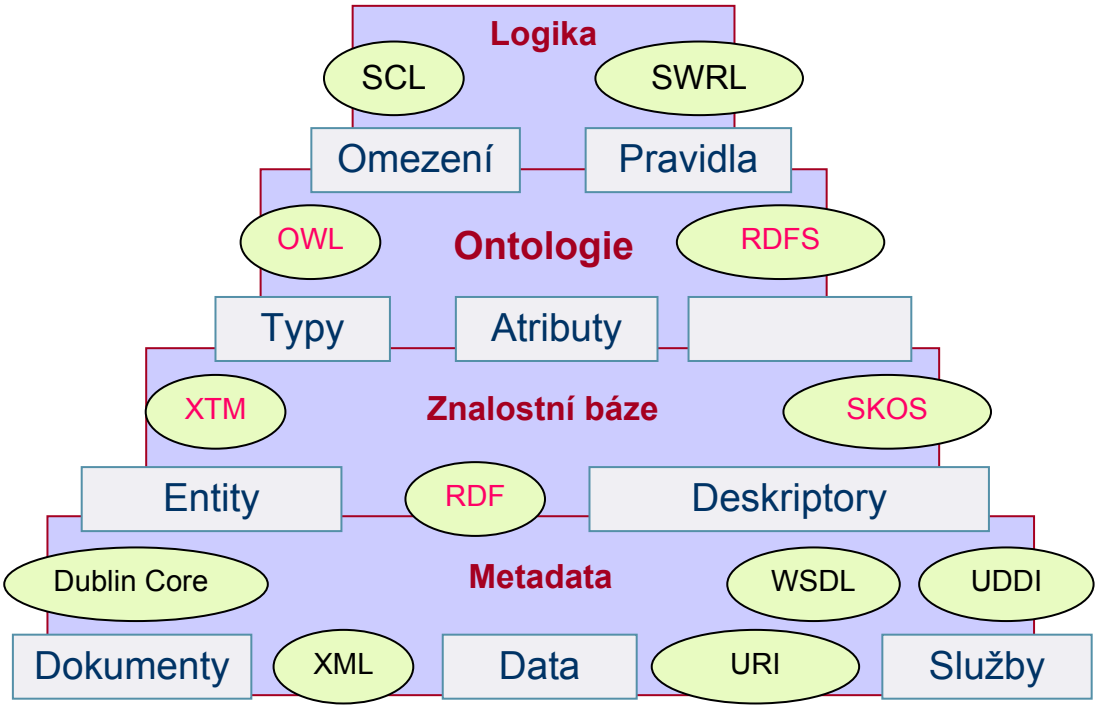
(1) Terminologické, lexikální (obdoba pokročilých tezaurů; knihovnictví).
(2) Informační (rozvinutí koncepčních DB-schémat - abstraktní nadstavba pro pojmové vyhledání; zajištění vyšší úrovně kontroly integrity).
(3) Znalostní (reprezentace znalostí pro umělou inteligenci – logické teorie).

Ontologie a tedy vyjádření významu a obsahu (sémantiky) předmětné oblasti (domény) vyžaduje určitou formální definici:

- Tříd (Classes): Typy objektů v dané oblasti zájmu.
- Vlastností (Properties): Atributy objektů a vztahy mezi nimi.
- Pravidel (Axioms): Vyjádření pravidel a omezení tříd a vlastností.
- Výskytů (Instances) všech výše uvedených konstruktů (volitelně).

Obr. 2 znázorňuje v logické pyramidě obsah, prvky a nástroje informačního modelování a znalostních bází. V článku není prostor na jejich podrobné vysvětlení, proto budou uvedeny pouze jejich názvy s oblastí použití:

- *XML (Extended Markup Language): značkovací jazyk, univerzální syntaxe dokumentů.*
- *URI (Uniform Resource Identifier): jedinečná identifikace informačního zdroje.*
- *Dublin-Core: Metadatová iniciativa, obsahuje soubor metadatových prvků.*
- *WSDL (Web Service Definition Language): Jazyk definování webových služeb.*
- *UDDI (Universal Description, Discovery and Integration): Registr webových služeb.*
- *RDF (Resource Description Framework): Rámec pro popis informačních zdrojů.*
- *XTM (XML Topic Maps): Reprezentace struktury informačních zdrojů v podobě Topic Maps.*
- *SKOS (Simple Knowledge Organisation Systems): Organizace znalost. systémů.*
- *OWL (Web Ontology Language): Jazyk pro zápis ontologie.*
- *RDFS (Resource Description Framework Schema): Popis použití RDF pro tvorbu slovníků.*
- *SCL (Simple Common Logic): Jazyk logiky 1. řádu pro výměnu informací.*
- *SWRL (Semantic Web Rule Language): Jazyk pro zápis pravidel sémantického webu.*



Obr. 2 Pyramida prvků a nástrojů informačního modelování a znalostních bází

4. PROBLÉMY ENTERPRISE INTEROPERABILITY A VÝCHODISKA

Podle [8] nemohou stávající technologie dosáhnout enterprise datové interoperability, neboť veškeré nezávisle vytvořené datové modely, včetně databázových schémat, datových slovníků, metadat, taxonomií a ontologií nejsou interoperabilní. Každý má vlastní pohled, cíle a omezovací pravidla, které vedou k divergenci. Obecná praxe ukazuje, že informatici tráví zbytečně mnoho času interpretací dat a jejich vkládáním do jiných systémů a přípravou rozhraní k propojení systémů.

Tvorba rozsáhlejších domén není cestou k cíli, protože nikdy nelze zahrnout vše. Se zvětšujícím se rozsahem datového modelu, je stále obtížnější, aby se vývojáři modelu sjednotili na společných datových prvcích. To není jen důsledkem počtu tvůrců modelu, ale růstem počtu skupin, které v rámci rozsáhlé domény mají jiná hlediska a používají jinou terminologii.

Násobné mapování rovněž není řešením. To, co lze uskutečnit mapováním datových modelů mezi dvěma nebo několika málo systémy, to je neuskutečnitelné v celém rozsahu zájmové oblasti. Například v pozemních silách armády USA je v architektuře zahrnuto 18 domén, což by znamenalo 18 x 17 = 306 mapování. A to se jedná pouze o pozemní síly, co ještě letectvo, námořnictvo, vnitro, federální agentury, koaliční partneři a komerční sféra.

Standardními jazyky, jako jsou RDF či OWL, se neprodukují interoperabilní data. Jazyky mají mnoho variant významových možností. Na počátku vyjadřovacích možností je taxonomie, kterou tvoří hierarchie pojmů, z nichž každý je podtřídou obecnějšího pojmu. Lidé mohou pojem pochopit pouze v kontextu, avšak jednoduchá hierarchie je k tomu limitovaná. Na konci vyjadřovacích možností lze vyjádřit více významů, dané jazykem jako je OWL, ale ani tyto jazyky neprodukují data a ontologie shodného významu. Konsorcium World Wide Web (W3C) se sice snaží, ale interoperabilitu neřeší. Primárně se zaměřuje na standardy organizace WEBu. Jejich iniciativa v Sémantickém WEBu je směřována na sémantickou interoperabilitu dat v rámci domén a COI.

Interoperabilita je mnohoúrovňová, zejména je kritická sémantická úroveň interoperability dat. Na nejnižší úrovni musí počítač rozpoznat znaky, dobrým standardem interoperability zde může být znaková sada Unicode. Na další úrovni je XML standardním formátovacím jazykem. Vrstva ontologií definuje sémantiku dat, ale interoperabilitu nezajišťuje. Dále nad ní jsou vrstvy důvěry a spolehlivosti (Proof and Trust). Pokud počítač najde a porozumí v sémantickém webu zdroji dat, ještě bude třeba pro zamýšlený účel ověřit, zda lze nalezeným datům důvěřovat.

Současné technologie jsou měřeny dle schopnosti podporovat funkce organizací. Technické komunity používají měřítko Technology Readiness Level (TRL), což je aplikováno na Software Readiness Scale (měřítko připravenosti SW k nasazení) s úrovněmi:

(1) Basic principles observed and reported (Dosaženy základní principy).
(2) Technology concept and/or application formulated (Formulován koncept).
(3) Analytical and experimental critical functions and/or characteristic proof of concept (Ověřen koncept analytických funkcí).
(4) Component and/or breadboard validation in laboratory environment (Experimentováno a ověřeno v laboratorních podmínkách).
(5) Component and/or breadboard validation in relevant environment (Experimentováno a ověřeno v příslušných podmínkách).
(6) System/subsystem model or prototype demonstration in a relevant environment (Demonstrace prototypu v příslušných podmínkách).
(7) System prototype demonstration in an operational environment (Demonstrace prototypu v operačním prostředí).
(8) Actual system completed and 'flight qualified' through test and demonstration (Systém připraven po testech a ověřeních).
(9) Actual system 'flight proven' though successful mission operations (Systém připraven k nasazení).

Podle výše uvedeného měřítka jsou současnými technologiemi ty, které splňují úroveň 8-9, což jsou XML, Metadata, RDF a OWL Language. Přesto neposkytují schůdné řešení pro sdílení dat více doménami velkých organizací. Nezávisle vyvinuté modely nejsou interoperabilní. Velké domény jsou uvnitř interoperabilní, nikoliv vně a znamenají nezvládnutelný problém při dalším rozšiřování.

4.1 Jak dál v dosažení enterprise datové interoperability?

Kandidáti technického řešení enterprise interoperability byly prověřeni pracovní skupinou Cross Domain Semantic Interoperability. Jedná se o jednoduché upper ontologie, jako jsou SUMO, DOLCE, OpenCyc, BFO. Upper ontologie (Top-level či Foundation) jsou pokusem o vytvoření ontologie, která by popsala obecné pojmy, jež jsou napříč doménami shodné. Množina doménových ontologií bude přístupna prostřednictvím upper ontologie, čím může být vytvořena interoperabilní infrastruktura.

V SUMO (Suggested Upper Merged Ontology, například na http://ontology.teknowledge.com/) je mapováno na všech 100 000 pojmů slovníku Wordnet. Velké organizace, jako je ministerstvo obrany či federální vláda by měly standardizovat potřebné upper ontologie a pak vyvinout či mapovat

standardní doménové ontologie v logistice, personalistce, akvizicích, lékařství, …

Dalšími kandidáty byla sada mapovaných upper ontologií. Teoretici namítají, že nelze vytvořit jedinou upper ontologii, aby vyhovovala všem systémům. Hovoří o sadě ontologií se silným mapováním (asi 5).

5. ZÁVĚR

Cílem článku bylo nastínit problémy interoperability IS a názory na jejich řešení. Z obecných hledisek lze odvodit požadavky na interoperabilitu v rámci NEC. Jedná se komplexní a komplikovaný problém, jehož zvládnutí je velkou výzvou pro výzkum, technologie, ale i pro politiky.

Seznam bibliografických odkazů

[1] International Conference Interoperability for Enterprise Software and Applications 2007. Portugal, Madeira, March 2007, www.i-esa.org/i-esa2007.
[2] International Command and Control Research and Technology Symposium. USA, New Port, June 2007, http://www.dodccrp.org/html3 /events_main.html
[3] European Interoperability Framework for pan-European eGovernment Services. Version 1.0. Belgium, Luxembourg: Office for Official Publications of the European Communities, 2004, 26 pp. ISBN 92-894-8389-X.
[4] http://www.usa.gov /Federal_Employees /Electronic_Government.shtml
[5] Metodika řízení tvorby architektur C3 systému resortu Ministerstva obrany. Praha: SKIS MO, 2005.
[6] Rozvoj, integrace, správa a bezpečnost KIS v prostředí NATO. Výzkumný záměr FVT403. Brno: UO/FVT, 2004-2008.
[7] Seminář HP IT Obzory „Nástroje pro strategické řízení IT“. Praha: Village Cinemas Anděl, 12.4.2007.
[8] Data Interoperability across the Enterprise - Why Current Technology Can’t Achieve it. US ARMY: CDSI, 2007, 12 s.

Summary: The article is aimed at outlining the IS interoperability problems and suggesting their solutions. From general aspects, the demands for interoperability can be derived within NEC. It is a complex and complicated problem, the managing of which poses a great challenge for research and technologies, as well as for politicians.

prof. Ing. Ladislav BUŘITA, CSc.
Univerzita obrany, Fakulta vojenských technologií
Katedra komunikačních a informačních systémů
Kounicova 65, 612 00 Brno, Česká republika
E-mail: Ladislav.Burita@unob.cz

NEW FILTERS FOR DISCRETE WAVELET TRANSFORM IN THE JPEG2000 STANDARD

Marcel HARAKAĽ, Ľubomír DEDERA, Július BARÁTH

Abstract: This article presents an application of the discrete wavelet transform in image compression with the JPEG2000 standard. The article also presents properties and some aspects of realization of the discrete wavelet transform as the base transform method in the JPEG2000 standard. Properties of the discrete wavelet transform are derived from results of experimental verification of filters used by the discrete wavelet transform.

Keywords: discrete wavelet transform, image compression, JPEG2000, FIR filter.

1. INTRODUCTION

In the last decade the wavelet transform (WT) [1] belongs to important theoretic and practical tools for developers and research workers. Its application areas include data processing and analysis, but first of all data coding, data compression and image processing. The roots of the WT can be found in the first half of the 20th century, but its practical utilization has begun after derivation and realization of the fast calculation algorithm [4] which uses quadrature-mirror filter (QMF) banks with finite impulse response (FIR). The main problem in selected application areas which utilize the WT (e.g. image compression) is the right choice of FIR filters [3].

Important properties of the WT in coding and compression areas are also utilized in compression standards like JPEG2000 and MPEG4. These standards are using the WT as the basic transformation method in the data preprocessing stage. Despite of the fact that the present standards recommend for WT realization certain types of FIR filters, it is possible to use any other filter types, which allow to reach marvelous results in image compression – higher quality of the reconstructed image.

2. STANDARD FOR STATIC IMAGE COMPRESSION JPEG200

Nowadays JPEG2000 [2] is the most frequently used standard for static image compression. This standard is not optimized only for effective image compression but also for its scalability and interoperability in telecommunication networks including cellular phone networks. The most significant features of JPEG2000 (in comparison with the previous JPEG standard) are:

- higher effectivity of lossy and lossless (200:1) image (black and white, gray, color) compression (with lossless decompression available in all types of computation),
- progressive transmission with the possibility to manage a resolution,
- possibility to choose a region of interest, for coding,
- continuous and by-level compressions,
- state-of-the-art low bit-rate compression performance,
- ownership protection by using watermarks and cryptography,
- option to add additional data (e.g. for a display).

In the terms of realization JPEG2000 is fundamentally more complicated than the JPEG standard. The coding process, using JPEG2000 method, can be divided into three steps: preprocessing, compression, and forming the output data stream [3, 5]. The JPEG2000 coder uses a similar block structure as the JPEG coder, but the discrete cosine transform (DCT) block is replaced by the discrete wavelet transform (DWT) block. The image decomposition using the DWT is realized into space oriented channels and it is represented by coefficients of image details and approximation with various directions. Quantization and coding is realized in bit-layers (code blocks). Entropy coding is based on arithmetic coding instead of Huffman coding. The output data stream created by the coder in the compression process is processed and saved as a file with new structure. The realization progress of color image compression with the JPEG2000 method can be seen in Fig. 1.

3. DISCRETE WAVELET TRANSFORM

The wavelet transform is a relatively new method of applied mathematics with the first experiments dating back to the the years 1982 and 1984. In spite of the fact that the first works connected with the definition of the WT can be found in the beginning of the 20th century (1909, the definition of the Haar basis), its practical usage

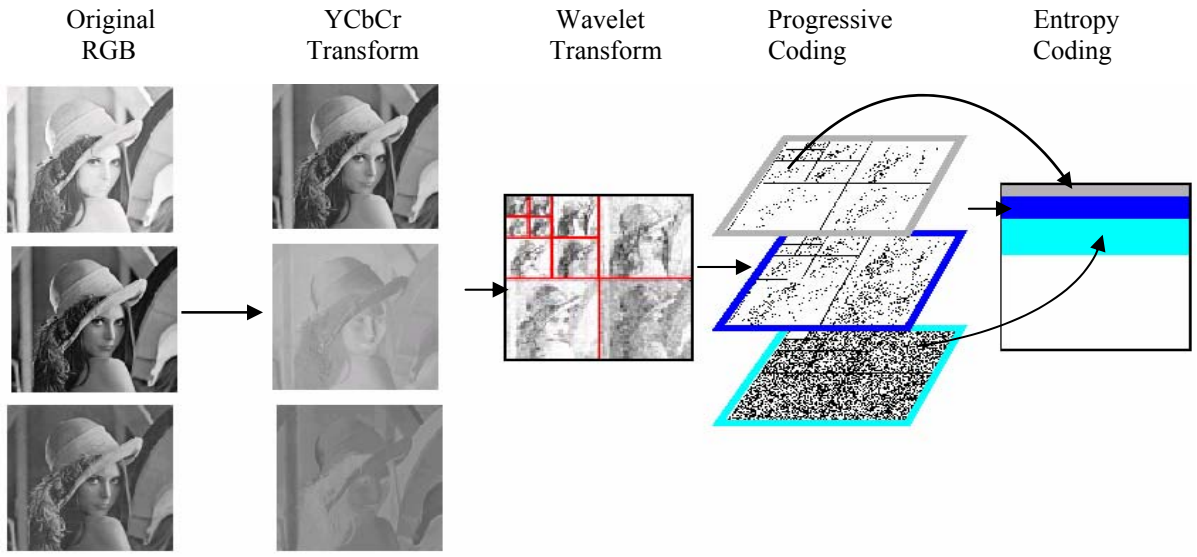


Fig. 1 Realization progress of image compression using JPEG2000 method

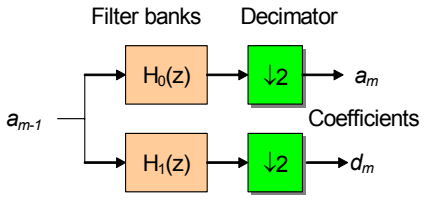
have been established by A. Grosmann and J. Morlet, who studied the continuous WT, with the definition of the notion "wavelet" as the base structural unit in the function decomposition (1). This transform decomposes the signal into function sets, which basis is the parent wavelet function $\psi(t) \in L^2(R)$. Let the function $\psi(t)$ be a wavelet function, along with the function $f(t) \in L^2(R)$. Then the continuous wavelet transform (CWT) of the function $f(t)$ is denoted by $WT_f(a,b)$: $(0,\infty)\times R \rightarrow R$ and defined by the formula

$$WT_f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t)\overline{\psi\left(\frac{t-b}{a}\right)}dt, \qquad (1)$$

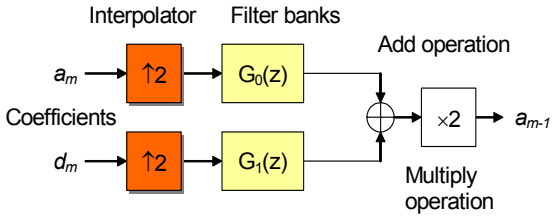
where $a \in R^+$, $b \in R$. Because the coefficient a represents scale and b time-shift, the representation presented here is also called the time-scale space. The continuous wavelet transform is highly redundant representation, uses scale and time-shift factors a, $b \in R$, $a>0$ and the realization of effective algorithms must implement their discretization. The formula (1) is not applicable for numerical computing, because it always contains the continuous variable t. One alternative to eliminate this problem is the usage of discrete functions and the substitution of the integral with the summation. A simple solution is to substitute $a=2$, $b=1$ with the unit sampling frequency. The formula (1) can be then rewritten as

$$w_f(m,n) = 2^{-m/2} \sum_k f(k)\psi(2^{-m}k - n), \qquad (2)$$

where $k, m, n \in Z$.



a) Decomposition of discrete signal



b) Composition of discrete signal

Fig. 2 Technical realization of discrete signal decomposition and composition

The practical computation of the discrete wavelet transform (DWT) by the formula (2) is possible, but it is not effective, because with acending m the count of the samples of the discrete function ψ excessively increases. Therefore, the fast Mallat algorithm [4] is used for numerical computations, which is based on the application of FIR filter banks, which were derived for subband coding. The use of proper combinations of filter banks (Fig. 2) allows executing the signal

decomposition and composition with the arbitrary deepness of decomposition. Then it is possible to use the obtained image detail d_m and the approximation a_m coefficients for realization of the designed compression method. An image in the transformed space, generated by the coefficient sets, is of the same size as the original image. The transform coefficients obtained in the form of a hierarchical tree and pyramidal structures represent gross image approximation and detailed images with different resolution and different spatial operation.

Tab. 1 Parameters of filters bank for realization DWT

Filter name	Type	Coefficients (in order $0, \pm 1, \pm 2, \pm 3, ...$)
Integer (1/3)	H_0	1
	H_1	1, -1/2
Integer (5/3)	H_0	6/8, 2/8, -1/8
	H_1	1, -1/2
Daubechies (9/7)	H_0	0.602949, 0.266864, -0.078223, -0.016864, 0.026748
	H_1	1.115087, -0.591272, -0.057544, 0.091272
CRF (13/7)	H_0	162/256, 80/256, -31/256, -16/256, 14/256, 0, -1/256
	H_1	1, -9/16, 0, 1/16
Swelden (13/7)	H_0	348/512, 144/512, -63/512, -6/512, 18/512, 0, -1/12
	H_1	1, -9/16, 0, 1/16
Float (9/7)	H_0	0.852699, 0.377402, -0.110624, -0.023849, 0.037828
	H_1	0.788486, -0.418092, -0.040689, 0.064539
Float (13/11)	H_0	0.76725, 0.38327, -0.06888, -0.03348, 0.04728, 0.00376, -0.00847
	H_1	0.83285, -0.44811, -0.06916, 0.10874, 0.00629, -0.01418
Float (5/3)	H_0	1.06066, 0.353553, -0.176777
	H_1	0.707107, -0.353553
Float (9/3)	H_0	0.994369, 0.419845, -0.176777, -0.066291, 0.033146
	H_1	0.707107, -0.353553

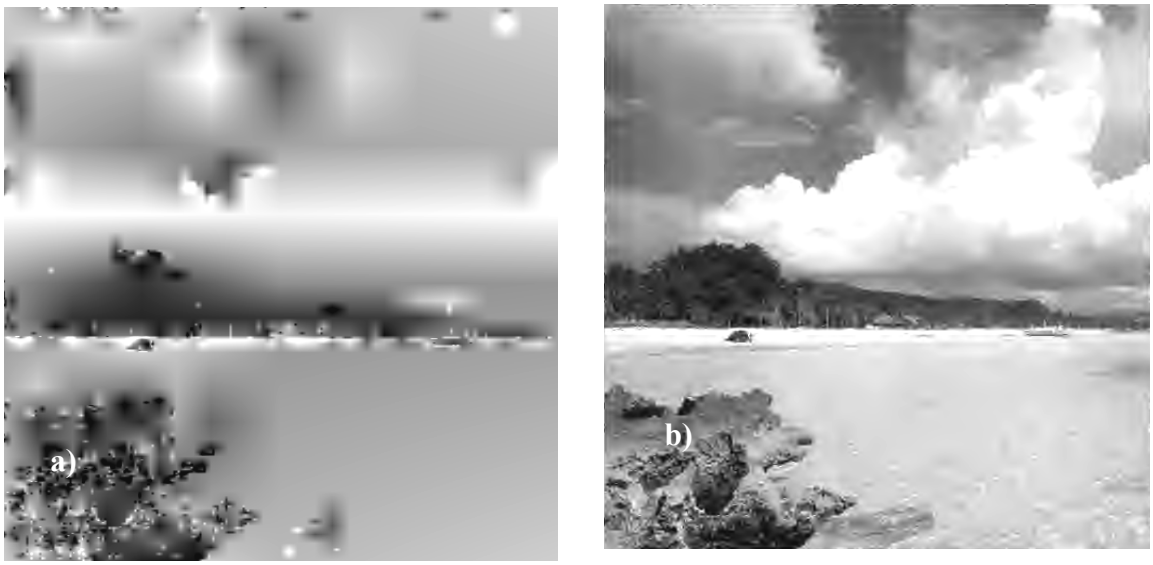


Fig. 3 60% image compression „Landscape“ with filters: a) Integer 1/3, b) Float 9/7

4. EXPERIMENTAL VERIFICATION OF FILTERS FOR REALIZATION OF DWT

In the design of a system realizing the DWT the selection of filter banks (Tab. 1) is the determining factor, which influences the quality of image processing [3]. In the JPEG2000 method the filter type selection can be realized by the mode chosen: lossy and lossless. In the case of the lossy mode, convolution filters are used.

In the case of the lossless mode, DWT is realized on the base of a „lifting" algorithm, which decomposes the filter function on linear partition. Usually, filter banks which are based on wavelet function properties (orthogonality, smoothness, zero moments count, translation invariance, definitiveness and regularity) are used.

In the JPEG2000 method it is recommended by the norm to use the integer 5/3 filter for lossless compression and the Daubechies 9/7 filter for lossy compression. For DWT realization, the Daubechies 10/18 filter, 6/10 filter, Swelden 13/7 filter, integer 2/6 filter, 2/10 filter, float 5/3 filter, 9/7 filter, 9/3 filter, 13/11 filter can be also applied. In the way the filter places the value and parity the realization of the methods are fundamentally different. The preferred methods are those with impair filters, because of their less complicated realization.

Tab. 2 Values PSNR in dB for image „Landscape" with various compression value

Filter type	Compression [%]				
	10	20	30	40	50
Integer (1/3)	29,01	24,89	21,62	19,70	17,47
Integer (5/3)	31,89	26,52	22,92	20,64	18,94
Daubechies (9/7)	29,77	23,57	19,82	18,04	15,60
CRF (13/7)	30,54	24,60	21,75	19,62	18,29
Swelden (13/7)	20,35	19,99	19,45	18,46	17,29
Float (9/7)	38,15	33,66	31,39	29,99	28,98
Float (13/11)	35,27	32,23	30,31	29,11	28,21
Float (5/3)	38,16	34,01	31,92	30,60	29,59
Float (9/3)	**38,37**	**34,23**	**32,14**	**30,84**	**29,84**

Filter type	Compression [%]				
	60	70	80	90	100
Integer (1/3)	15,63	14,51	13,80	12,65	12,98
Integer (5/3)	17,63	15,88	15,19	15,13	14,88
Daubechies (9/7)	15,28	14,90	14,38	14,21	13,80
CRF (13/7)	16,20	15,01	15,03	14,89	14,65
Swelden (13/7)	17,12	16,55	15,16	13,85	13,37
Float (9/7)	28,25	27,65	27,14	26,75	26,37
Float (13/11)	27,47	26,92	26,47	26,07	25,69
Float (5/3)	28,89	28,36	27,91	27,54	27,16
Float (9/3)	**29,10**	**28,57**	**28,14**	**27,75**	**27,38**

We have experimentally verified impair filter properties on a set of multilayer gray images. Depending on the compression level settings and filter types we measured the quality of the reconstructed images by the Peak Signal to Noise Ratio (PSNR) in decibels (dB). The results of both experiments (e.g. table 2 and 3) have shown the best properties of the impair float 9/3 filter.

5. CONCLUSION

In the article properties and some aspects of the realization of the DWT as the base transformation method in the JPEG2000 standard are presented. It also includes an experimental verification of the properties of impair FIR filters designed for the realization of the DWT. The experimental results

on a set of multilayer gray images show that the float 9/3 filter has the best properties within the examined set of impair FIR filters. This important result means higher quality of reconstructed image in comparison with the recommended Daubechies 9/7 filter. Therefore it is possible to take this fact into consideration in the realization of the DWT within the JPEG2000method in the lossy mode.

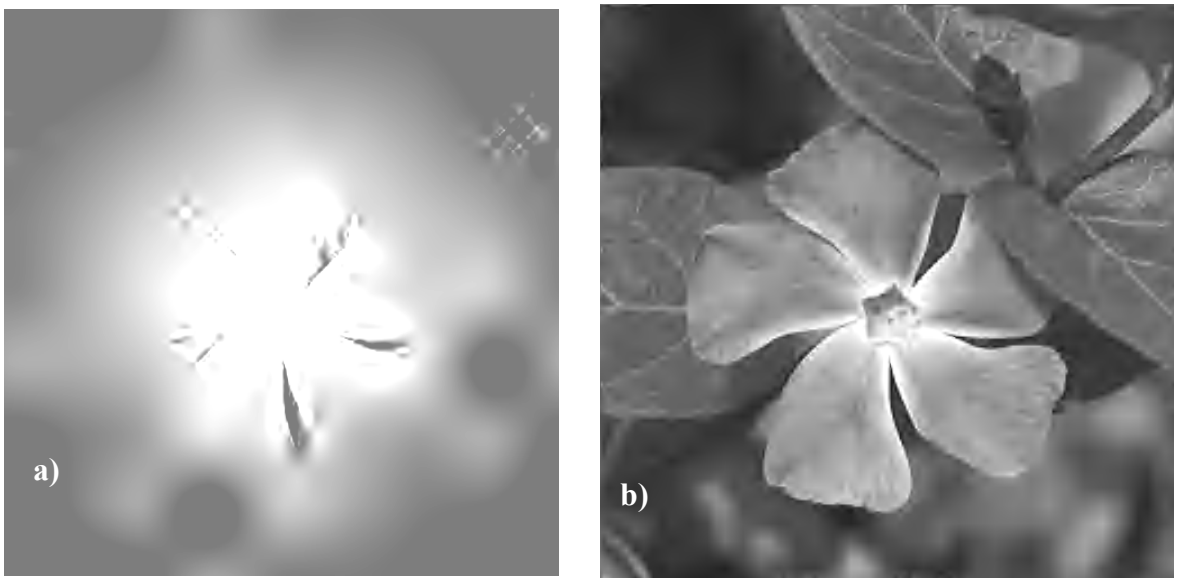


Fig. 4 90% image compression „Flower“ with filters: a) *Swelden 13/7*, b) *Float 9/3*

Tab. 3 Values PSNR in dB for image „Flower“ with various compression value

Filter type	Compression [%]				
	10	20	30	40	50
Integer (1/3)	29,89	24,18	20,93	19,81	19,18
Integer (5/3)	32,14	26,51	22,82	21,32	19,27
Daubechies (9/7)	29,74	22,96	20,59	18,31	17,82
CRF (13/7)	28,48	24,82	21,64	19,14	18,06
Swelden (13/7)	17,12	16,66	16,06	15,25	14,90
Float (9/7)	39,15	35,48	33,60	32,30	31,34
Float (13/11)	33,06	31,70	30,67	29,87	29,18
Float (5/3)	39,38	36,09	34,33	33,18	32,25
Float (9/3)	**39,64**	**36,33**	**34,61**	**33,47**	**32,56**

Filter type	Compression [%]				
	60	70	80	90	100
Integer (1/3)	18,41	17,04	16,22	16,35	16,36
Integer (5/3)	18,10	17,96	17,75	17,63	17,39
Daubechies (9/7)	17,49	17,37	17,31	17,24	17,22
CRF (13/7)	17,86	17,43	17,25	17,16	17,17
Swelden (13/7)	14,33	13,99	13,83	13,73	13,65
Float (9/7)	30,58	29,90	29,33	28,88	28,42
Float (13/11)	28,65	28,18	27,74	27,35	27,03
Float (5/3)	31,50	30,89	30,38	29,99	29,55
Float (9/3)	31,78	31,17	30,67	30,26	29,86

ACKNOWLEDGEMENT

This work was supported by Academic Grant Agency the Academy of the Armed Forces project No. AGA-01-2007 "Acceleration Technologies for High-Performance Computing".

References

[1] CHUI, C. K.: An Introduction to Wavelets. Academic Press. San Diego, CA, 1992.
[2] CHRISTOPOULOS, CH., SKODRAS, A., EBRAHIMI, T.: The JPEG2000 Still Image Coding System: An Overview. IEEE Transactions on Consumer Electronics, Vol. 46, No.4, 2000, pp. 1103-1127.
[3] HARAKAĽ, M.: Hardvérová implementácia waveletovej transformácie pre kompresiu obrazov. Habilitačná práca, Liptovský Mikuláš, 2001.
[4] MALLAT, S. G.: A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Transactions on PAMI, 11, 1989, pp. 674-693.
[5] LEVICKÝ, D.: Multimediálne telekomunikácie – multimédia, technológie a vodoznaky. ELFA s.r.o., Košice, 2002.
[6] HARAKAĽ, M.: Realizácia waveletovej transformácie v štandarde JPEG 2000. Zborník Vojenskej akadémie, roč. 11, č. 2, s. 5-12 Liptovský Mikuláš, 2004.
[7] HARAKAĽ, M., JUREČKA, M., TURČANÍK, M.: A new experimental verification of the wavelet transform in the JPEG2000 standard. In: Proceedings of the DSP – MCOM 2005. The 6th International Conference on Digital Signal Processing and Multimedia Communications: September 13 - 14, 2005, Košice, Slovak Republic. – Košice : Department of Electronics and Multimedia Communications, 2005. – ISBN 80-8073-313-9. – S. 31-34

Summary: The article presents an application of the discrete wavelet transform in image compression with the JPEG2000 standard. The article also presents properties and some aspects of realization of the discrete wavelet transform as the base transform method in the JPEG2000 standard. Properties of the discrete wavelet transform are derived from results of experimental verification of filters used by the discrete wavelet transform. The experimental results on a set of multilayer gray images show that the float 9/3 filter has the best properties within the examined set of impair FIR filters. This important result means higher quality of reconstructed image in comparison with the recommended Daubechies 9/7 filter. Therefore it is possible to take this fact into consideration in the realization of the DWT within the JPEG2000 method in the lossy mode.

doc. Ing. Marcel HARAKAĽ, PhD.
doc. RNDr. Ľubomír DEDERA, PhD.
Ing. Július BARÁTH, PhD.
Akadémia ozbrojených síl generála M. R. Štefánika
Katedra informatiky
Demänová 393
031 01 Liptovský Mikuláš
Slovenská republika
E-mail: harakal@aoslm.sk
dedera@aoslm.sk
julius.barath@aoslm.sk

ACCURATE SIMULATION OF SWITCHED SYSTEMS USING VHDL-AMS

Zdeněk KOLKA, Dalibor BIOLEK, Viera BIOLKOVÁ

Abstract: The paper deals with behavioral modeling of switched systems with discontinuities. The purpose of such models is to obtain the first-order effects to verify analytic calculations or to increase simulation speed. Traditional algorithms for the time-domain analysis implemented in Spice-class simulators are based on the assumption of smoothness and continuity. Abrupt changes of system parameters or even discontinuity during switching cause numerical errors. The VHDL-AMS language brings radically different approach in comparison with Spice [1]. The system of differential algebraic equations can be formulated explicitly and can be structurally modified during simulation. The basic principles will be demonstrated on the model of boost converter with accelerated finding of steady-state solution.

Keywords: Computer simulation, switched systems, behavioral modeling, VHDL-AMS.

1. INTRODUCTION

Numerical algorithms for the time-domain analysis of continuous-time circuits are based on the formulation of circuit equations into the system of differential-algebraic equations (DAEs)

$$F(\dot{\mathbf{x}}, \mathbf{x}, t) = \mathbf{0} \qquad (1)$$

where $\mathbf{x}$ is the vector of circuit variables. System (1) cannot be solved directly, but its solution $\mathbf{x}(t)$ is approximated by a discrete sequence $\mathbf{x}(t_n) \approx \mathbf{x}_n$. To find the $(n+1)th$ step the adjoint algebraic system

$$G(\mathbf{x}_{n+1}, t_{n+1}, \mathbf{X}_n) = \mathbf{0} \qquad (2)$$

should be solved using the Newton-Raphson iteration method. $\mathbf{X}_n$ is the matrix of several past solutions up to $\mathbf{x}_n$, $\mathbf{X}_n = \left\lfloor \mathbf{x}_{n-p}, ..., \mathbf{x}_{n-1}, \mathbf{x}_n \right\rfloor$. Various discretization methods transforming (1) into (2) are discussed in [2].

The time step $h_n = t_n - t_{n-1}$ is automatically adjusted according to the estimation of Local Truncation Error (LTE), which is inversely proportional to h [1]. The minimum timestep is usually limited to

$$h \geq T_{final} \frac{10^{-m}}{RELTOL} \qquad (3)$$

where m is the precision of time representation in simulator, T_{final} is the length of simulation interval and *RELTOL* is the simulator parameter representing required relative accuracy. For usual values we obtain $h \geq T_{final} 10^{-12}$. The lower limit is necessary to maintain the numerical precision of $t_{n+1} = t_n + h_{n+1}$ assignment.

Algorithms of traditional Spice simulators are based on the assumption of continuity of (1) resulting in a smooth solution. VHDL expects piecewise smooth solution with a finite number of discontinuities where

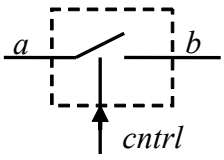
$$\dot{\mathbf{x}}(t_n^-) \neq \dot{\mathbf{x}}(t_n^+) \text{ or even } \mathbf{x}(t_n^-) \neq \mathbf{x}(t_n^+) \; . \qquad (4)$$

The discontinuity (4) can be handled easily by computing the solution at t_n^-, changing the system (1) and its state variables, and restarting the solution from t_n^+ using $\mathbf{x}(t_n^+)$ as the initial condition if the exact time t_n is signaled to the solver explicitly.

Without the explicit signaling the simulator jumps over the exact time instant t_n in case when the LTE is acceptable or ends up with a "Timestep too small" message if limit (3) is reached. It should be noted that system (2) is still expected to be smooth in the neighborhood of $\mathbf{x}_{n+1}$.

2. SWITCHED CIRCUIT MODELING WITH VHDL-AMS

A switched system contains active and passive switches. Fig. 1 shows a VHDL-AMS model of an active (controlled) switch. The switch-on resistance R_A is allowed to be zero.




```
entity sw is
   generic (
    Ra : resistance := 0.0);
  port (
    terminal a,b: electrical;
    signal cntrl: in std_logic);
end entity sw;

architecture ideal of sw is
  quantity v across i through a
to b;
begin
  if cntrl='1' use
    v == i * Ra;
  else
    i == 0.0;
  end use;
  break on cntrl;
end architecture ideal;
```

Fig. 1 Model of ideal active switch

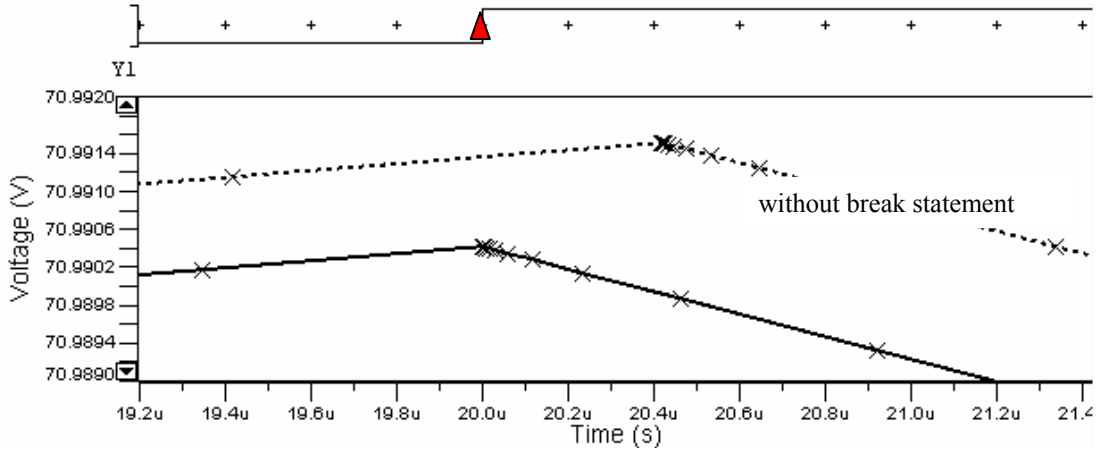


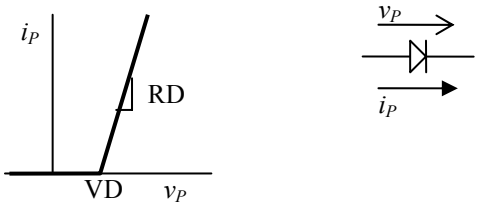
Fig. 2 Effect of break statement (capacitor voltage from Fig. 4a)

Discrete-time signal *cntrl* controls the behavior of the switch through the *if-else* statement in a piecewise manner. During iterations to obtain the solution of (2) for a particular time instant t_{n+1} the value of *cntrl* does not change, i.e. switching does not occur. Thus (2) is continuous. The *if* statement selects which equation will be used for the DAE system.

The exact time instant of changing the switch state should be signaled to the analog solver or else it "jumps over" the event. This is done by the *break* statement, Fig. 2.

The extended syntax of the break statement allows changing the values of state variables of (1) at given time instant. The state-variable transformation $\mathbf{x}(t_n^-) \to \mathbf{x}(t_n^+)$ should be defined explicitly using VHDL-AMS constructs. The simulator cannot derive it from the circuit topology.

The passive switch (diode) can be approximated by a piecewise linear function. Fig. 3 shows the approximation and a fragment of the VHDL-AMS code. This function introduces an *implicit* discontinuity that occurs during iterations of the analog solver. The effect is a numerical error similar to that in Spice simulators.




```
iP == realmax(0.0,(vP-VD)/RD);
```

Fig. 3 Piecewise-linear passive switch

3. MODEL OF BOOST CONVERTER

A VHDL-AMS model can detect the type of analysis being performed. A discrete signal *domain* whose value depends on the type of analysis is predefined in VHDL-AMS. With the *if* statement it is possible to select which equation to use for a particular analysis. This technique will be demonstrated on the time domain model of boost converter with accelerated finding of the steady-state solution. The direct time-domain analysis of switched power supplies in the Spice and the VHDL-AMS simulators results in very long simulation times as the integration step depends on switching transients and the time interval of interest is usually several orders of magnitude longer. Since the steady-state analysis is not available in majority of simulators, the only possibility is to run a sufficiently long transient simulation.

The utilization of the averaged modeling technique results in an incomparably faster analysis [3], [4]. On the other hand, by smoothing the fast switching process we lose information about the output voltage ripple and other characteristics.

The steady-state solution of (1) is characterized by condition

$$\mathbf{x}(T) = \mathbf{x}(0) \qquad (5)$$

where T is the switching period. Finding the steady-state solution is equivalent to finding the appropriate initial conditions $\mathbf{x}(0)$. The utilization of the DC operating point as the initial condition, normally used for the transient analysis, is useless here.

The averaged model represents relations between short-time average values of all quantities. Finding a steady state-solution in the original model is equivalent to finding a DC operating point in the averaged model.

Fig. 4a shows a time-domain model of boost converter from [3]. The following numerical values have been used:

V = 60V Ro = 60Ω C = 1000μF
L = 6mH RL = 3Ω
T = 10μs D = 0.25 (duty ratio)
active switch: RA = 1Ω (Fig. 2)
passive switch: VD = 0.6V, RD = 1Ω (Fig. 3)

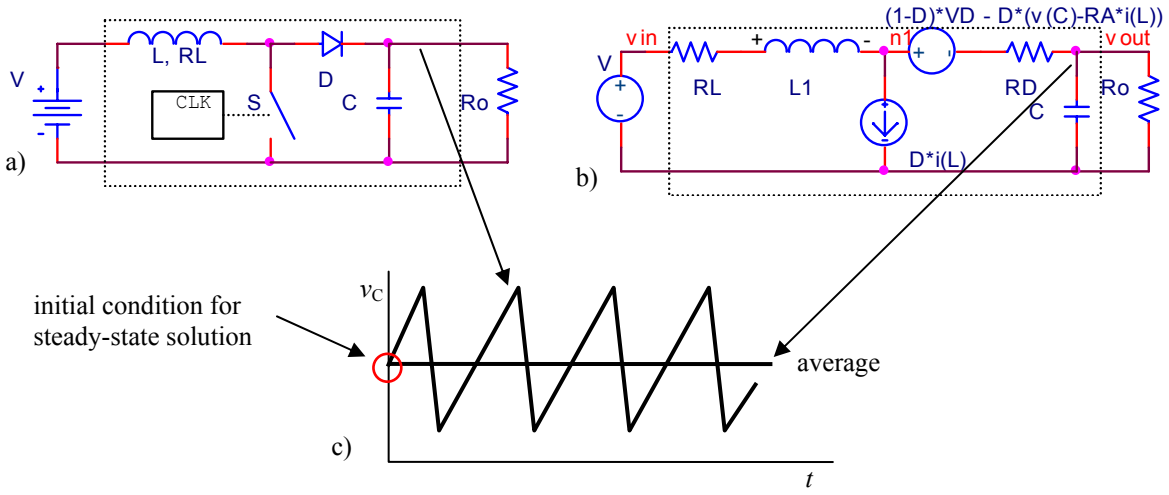


Fig. 4 a) Boost converter; b) its averaged model; c) diagram of capacitor voltage

The technique of PWM switch has been used to obtain the averaged model [3], [4], Fig. 4b. It consists in the replacing both switches by two controlled sources. The model obtained is a linear circuit, which can be used for all basic analyses. It is valid from DC up to the Nyquist frequency ($f_{SW}/2$).

Fig. 5 shows the VHDL-AMS code of the behavioral model. For the DC and the AC domains the averaged model is used. The DC analysis corresponds to the steady-state analysis in the original circuit. For the time domain the original model is used. If the DC (averaged) solution is used as the initial condition we obtain immediately the steady-state in the time domain. The first cycle of switching clock generator is shortened to correctly use the initial condition, Fig. 4c.

```
entity boost is
port (terminal vin, vout, vref:
      electrical;
  quantity D : in real);
end entity boost;

architecture md of boost is
  terminal n1  : electrical;
  signal   clk : bit := '0';
  quantity vL across iL through vin to
n1;
  quantity vA across iA through n1 to
vref;
  quantity vP across iP through n1 to
vout;
  quantity vC across iC through vout to
vref;
begin
--clock generator
Clock: process
begin
  clk <= '0';
  wait for Ts*(1.0-D)/2.0;
  loop
    clk <= '1';
    wait for Ts*D;
    clk <= '0';
    wait for Ts*(1.0-D);
  end loop;
end process Clock;
break on clk;

if domain = time_domain use
  --L, RL
  vL == RL*iL + L*iL'dot;
  --C
  iC == C * vC'dot;
  --active switch
  if clk='1' use
    vA  == iA * RA;
  else
    iA  == 0.0;
  end use;
  --passive switch
  iP == realmax(0.0,(vP-VD)/RD);
else
  -- other domains
  --L, RL
  vL == RL*iL + L*iL'dot;
  --C
  iC == C * vC'dot;
  --active switch
  iA  == D * iL;
  --passive switch
  vP == (1.0-D)*VD
        -D*(vC-RA*iL)+iP*RD;
end use;
end architecture md;
```

Fig. 5 VHDL-AMS code of boost converter (definition of generic constants omitted)

Fig. 6 shows a comparison of the proposed steady- state model with a long transient simulation from zero initial condition. Even after 2000 switching periods the system still has not reached the steady-state.

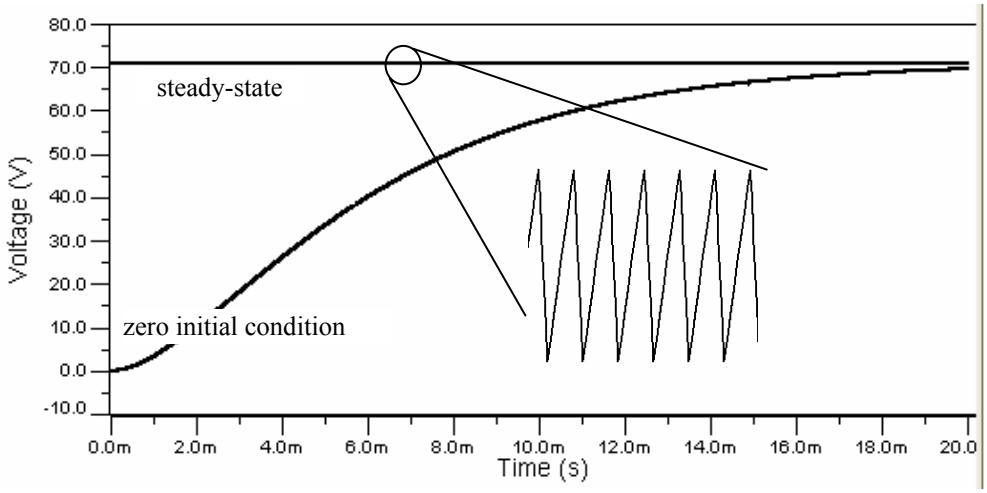


Fig. 6 Comparison the zero initial condition with the AVERAGED DC operating point

5. CONCLUSIONS

A method for rapid finding of the steady-state solution of switched mode power supplies based on averaged modeling has been shown. The method is based on the new possibilities introduced in VHDL-AMS for **behavioral modeling**. For a detailed analysis of circuits on the transistor level the best choice is still the Spice simulator.

ACKNOWLEDGEMENT

This research has been financially supported by the Czech Science Foundation under projects No. 102/05/0771 and No. 102/05/0277, and by the Czech Ministry of Education under research program No. MSM0021630513. The simulation software SystemVision has been provided by Mentor Graphics under the Higher Education Programme.

References

[1] CRISTEN, E., BAKALAR, K. VHDL-AMS – A Hardware Description Language for Analog and Mixed-Signal Applications, IEEE Trans. on Circ. and Syst. II, vol.46, no.10, Oct. 1999, p. 1263-1272.

[2] KUNDERT, K., The Designer's Guide to SPICE and SPECTRE, Kluwer Academic Publishers, 1995.

[3] DIJK, E., SPRUIJT, J. N., SULLIVAN, M. O., KLAASSENS, J. B. PWM-switch Modeling of DC-DC Converters, IEEE Trans on Power Electronics, 1995, vol. 10, no. 6, p. 659-664.

[4] BIOLEK, D., BIOLKOVÁ, V., KOLKA, Z. SPICE Modeling of Switched DC-DC Converters via Generalized Model of PWM Switch, Proc of Radioelektronika 2006, 2006, p. 180-183.

[6] VHDL-AMS Language Reference Manual, IEEE Standard 1076.1 - 1999.

Summary: The paper deals with behavioral modeling of switched systems with discontinuities. The purpose of such models is to obtain the first-order effects to verify analytic calculations or to increase simulation speed. Traditional algorithms for the time-domain analysis implemented in Spice-class simulators are based on the assumption of smoothness and continuity. Abrupt changes of system parameters or even discontinuity during switching cause numerical errors. The VHDL-AMS language brings radically different approach in comparison with Spice. The system of differential algebraic equations can be formulated explicitly and can be structurally modified during simulation. The basic principles will be demonstrated on the model of boost converter with accelerated finding of steady-state solution.

prof. Ing. Dalibor BIOLEK, CSc.[1)]
Ing. Viera BIOLKOVÁ[2)]
doc. Dr. Ing. Zdeněk KOLKA[2)]
[1)] UO Brno, Katedra elektrotechniky, Kounicova 65
612 00 Brno, Česká republika
E-mail: dalibor.biolek@unob.cz
[2)] FEKT VUT Brno, Ústav radioelektroniky, Purkyňova
118 612 00 Brno, Česká republika
E-mail: biolkova@feec.vutbr.cz
kolka@feec.vutbr.cz

SOFTWARE EVOLUTION FROM A META-LEVEL COMPILER PERSPECTIVE

Ján KOLLÁR, Jaroslav PORUBÄN, Peter VÁCLAVÍK,
Jana BANDÁKOVÁ, Michal FORGÁČ

Abstract: From the viewpoint of adaptability, we classify software systems as being nonreflexive, introspective and adaptive. Multiple metalevel concepts are essential demand for a systematic language approach, to build up adaptable software systems dynamically, i.e. to evolve them. Paper presents the software evolution from a computer language perspective. Using this approach the system can be evolved not just through source code changes but even the language itself is evolving through the compiler adaptation defined on meta-levels.

Keywords: Adaptive compiler, program transformation, software evolution, system reflection, metaprogramming.

1. INTRODUCTION

There is an increasing demand for software systems that can be easily configured for a specific deployment environment or they even adjust themselves dynamically to a changed environment at runtime. Today, software needs to be changed on an ongoing basis with major enhancements required on a short timescale (days or weeks) in order to meet new business opportunities and reduce the time to market for new products and services [15]. Software changes now comprise a major portion of software life-cycle costs.

Software maintenance is the modification of a software product after delivery to correct faults, to adapt to a changed external runtime environment, or to adapt to a changed user requirements. According to various studies costs for software maintenance ranges between 50 and 90 percentages [3, 4, 14, 15]. Nowadays companies spend more resources on maintenance of existing software than on development of new software. Maintenance becomes the most expensive software activity.

Software evolution covers programming activity that is intended to generate a new software version from an earlier operational version. Software evolution is the process of conducting continuous software reengineering. Reengineering implies a single change cycle, but evolution can go on forever. In other words, to a large extent, software evolution is repeated software reengineering [15]. System evolution is so common that a development from scratch is the exception.

The main steps for reengineering are to determine what the existing software does, to decide what to modify in the software and how to actually carry out the modifications. Understanding software means to identify and extract the actual, current design of the software.

Functional enhancements are inevitable in the evolution of any successful software. As the business environment changes, users come up with new requirements [6]. In common specifications are created incrementally as functional and non-functional system's requirements evolve. Customers are rarely able to provide a complete specification at any stage of the project [12].

The more general focus of software evolution studies is on the how of evolution. The concern has been, and still is, to find effective abstractions, formalisms, procedures, methods and tools, for example, for performing and improving the evolution process so as to increase productivity, reliability, dependability, adaptability and predictability, to improve quality, to decrease development time and so on [10].

In our research we are concentrated on system evolution involved by a system external runtime change and especially on software evolution based on user requirements change, not on fault fixing software evolution.

In the paper we present the software evolution from a computer language perspective. Using this approach the system is evolved not just through source code changes but even the language itself is evolving through the compiler adaptation defined on meta-levels. This is a step on a way to self-adaptive software – a software that incorporates monitoring and evaluation functions, and can rapidly (at runtime) respond to some sorts of need for change [9]. Only by understanding the relationships and dependencies between entities in the software process (such as specification, design and implementation) can we begin to objectively categorize and potentially automate aspects of software evolution [5]. The meta-level compiler can be useful in many systems like described in [1, 11].

The section 2 presents the principles of metaprogramming. In the section 3 we classify systems from the viewpoint of their degree of self-adaptability. The section 4 introduces the meta-level compiler and its relation to a system evolution. The section 5 concludes the paper.

2. METAPROGRAMMING

Metaprogramming is about writing programs that represent and manipulate other programs or themselves. The prefix "meta" denotes the property of "being about", that is, metaprograms are programs

about programs [1]. The most common metaprogramming tool is a compiler.

Reflection is an entity's integral ability to represent, operate on, and otherwise deal with itself in the same way that it represents, operates on, and deal with its primary subject matter. Reflection is a fundamental concept of self-adaptive systems.

The main idea of applying reflection as a general principle for flexible systems in software engineering is to split a system into two parts: metalevel and a base level. A metalevel provides information about selected system and makes the software self-aware. A base level includes the application logic.

There are two aspects of reflection: introspection and intercession. Introspection is the ability of a program to observe and therefore to reason about its own state. Intercession is a higher degree of reflection, since it is the ability of a program to modify its own execution state or alter its own interpretation or meaning. Both aspects require a mechanism for encoding execution state as data, providing such an encoding is called reification.

Different languages provide different levels of reflection. For example, the Java reflection API, allows a programmer to discover methods and attributes in classes at runtime, and to create objects of classes, whose names are not known until runtime. Similarly, it is possible to call methods and access attributes whose names are not known until runtime, because they may be discovered with the help of reflexive facilities, or they may be computed at runtime. Thus, Java's reflexive facilities primarily support introspection. In contrast to Smalltalk, Java does not allow to directly modifying classes or methods by modifying their metaobjects at runtime, that is, it does not support intercession [1].

Metaobjects are objects that represent methods, execution stacks, the processor, and nearly all elements of the language and its execution environment. Most importantly, regular language code can access and modify these metaobjects.

There is a principal difference between metaclass and metaobject. Java metaclass is static data while metaobject in Smaltalk is dynamic data. As we will see, dynamic metadata is the proposition for building adaptive systems.

Adaptiveness is mostly related to the area of software engineering - object oriented programming, aspect-oriented programming, intentional programming, template programming, etc. On the other hand, the properties of systems are expressed using programming language, specification language, or modeling language, in that solutions of problems are constructively formulated, no matter of language abstraction level. At bottom level of a system, machine code is executed, based on a machine language. At top level, human thoughts arise and they are formulated using a natural language.

3. SYSTEMS BEHAVIOR CLASSIFICATION FROM A META-LEVEL PERSPECTIVE

In this section we classify software systems from the viewpoint of their degree of adaptive behavior to three categories: nonreflexive execution, introspective execution and adaptive (intercessional) execution (Fig. 1).

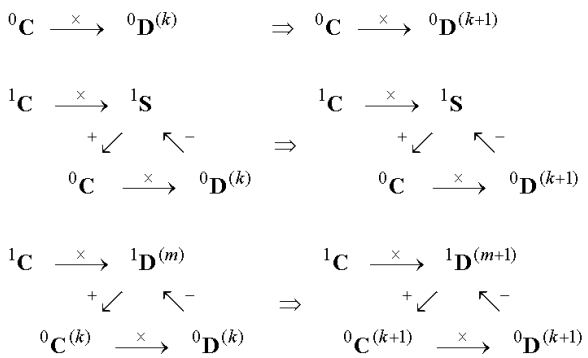


Fig. 1 Classification of software system execution: (a) Nonreflexive execution (b) Introspective execution (c) Adaptive execution

3.1 Nonreflexive Execution

Machine code – the constant set $^{0}\mathbf{C}$ of instructions at base level 0 does not vary during execution, and then the execution changes data $^{0}\mathbf{D}^{(k)}$ (the set of data records on the stack or in the heap) to a new data set $^{0}\mathbf{D}^{(k+1)}$. An execution step is the transformation of configuration shown on the Fig 1 (a). The relation ($\overset{\times}{\rightarrow}$) denotes that many instructions from $^{0}\mathbf{C}$, can access many data records $^{0}\mathbf{D}$. This execution is nonreflexive, because there is no feedback loop from data to code, and no possibility given to code to observe or even to change itself.

3.2 Introspective Execution

In an introspective execution, code ^{0}C constructs and changes data $^{0}\mathrm{D}^{(k)}$, as in nonreflexive execution. In addition to this, each subset of the set $^{0}\mathrm{D}^{(k)}$ refer (which we designate by $\overset{-}{\rightarrow}$) exactly one element of static data ^{1}S, and this data refers ($\overset{+}{\rightarrow}$) a subset of code $^{0}\mathbf{C}$ according Fig. 1 (b). Static data set $^{1}\mathbf{S}$ at level 1 are metalevel static data to the level 0. Since metalevel data (set of records) $^{1}\mathbf{S}$ is static, metacode $^{1}\mathbf{C}$ may produce it once, and then the execution of $^{1}\mathbf{C}$ is finished. Clearly, such metacode cannot be runtime process, and execution of $^{0}\mathbf{C}$ is nonadaptive. However, it is introspective, because of existence of feedback loop from code $^{0}\mathbf{C}$ to code $^{0}\mathbf{C}$ via data set $^{0}\mathbf{D}$ and some metadata element from $^{1}\mathbf{S}$.

3.3 Adaptive Execution

Adaptive execution is defined on the Fig. 1 (c). In this case, metalevel data ${}^{1}\mathbf{D}^{(m)}$ can change in runtime to data ${}^{1}\mathbf{D}^{(m+1)}$, by execution of metalevel code ${}^{1}\mathbf{C}$. This code itself is nonreflexive, since there is no feedback loop via metadata at level 2. On the other hand, new ${}^{1}\mathbf{D}^{(m+1)}$ may result to new ${}^{0}\mathbf{C}^{(k)}$, continuing its execution at level 0.

It means that code at level 0 may be not just introspective, but also adaptive, and this fact is essential for an adaptive execution. It is easy to see, that level 2 may be built upto level 1 similarly as level 1 upto level 0, etc. Such chain of metalevels represent abstractions of previous level, and this abstraction we recognized optimal, provided that

- Cardinality relation $|{}^{(l+1)}\mathbf{D}| << |{}^{l}\mathbf{D}|$ holds, since each metalevel should express its base level concisely, and
- Computational time relation ${}^{(l+1)}\tau << {}^{l}\tau$ hold, which says that computational time at a metalevel should be significantly shorter than that at a base level.

4. META-LEVEL COMPILER

Compiler translates source code into another form suitable for interpretation or execution. Using an adaptive compiler code can be recompiled accordingly to the results of program during its evaluation. Metadata about language in which a program is written are used to adapt to a new system's conditions. Using the meta-level compiler system can be change during the runtime according to current state of the system. Adaptive compiler infrastructure is shown on the Fig. 2.

During the evolution of a software system the source code (program) of a system is updated to reflect the changes of its external environment and user requirements. Program is recompiled and executed in iterative manner during the evolution.

$$(execute \circ translate)\ program_1 \rightarrow (execute \circ translate)\ program_2 \rightarrow (execute \circ translate)\ program_3 \rightarrow \ldots$$

In our approach instead of evolving a program (source code), the compiler and also the computer language are adopted.

$$(execute \circ translate_1)\ program \rightarrow (execute \circ translate_2)\ program \rightarrow (execute \circ translate_3)\ program \rightarrow \ldots$$

where

$$translate_i = metatranslate\ \ metaprogram_i$$

Compiling process usually consists of many phases such as lexical analysis phase, syntactic analysis phase, translation phase and code generation phase. Using the meta-level compiler all phases can be parameterized by meta-levels and configured by metaprograms leading to the modification of the whole system during runtime. In the Fig. 2 the meta-level is represented by Meta-level Lexical Analysis, Meta-level Syntactic Analysis, Meta-level Translator and Meta-level Code Generation which corresponds to mentioned phases of compiling. Compiler itself is the part of the runtime therefore it is possible to change and recompile program in the runtime.

4.1 Experiments

According to the specified meta-compiler architecture some experiments were done with LL(1) expression language proving the feasibility of suggested meta-compiler architecture. As a platform for meta-compiler implementation functional programming language Haskell [13] and Java programming language with JavaCC tool [7] have been chosen [8]. At the runtime the evaluation was reflected and adopted by specified rules.

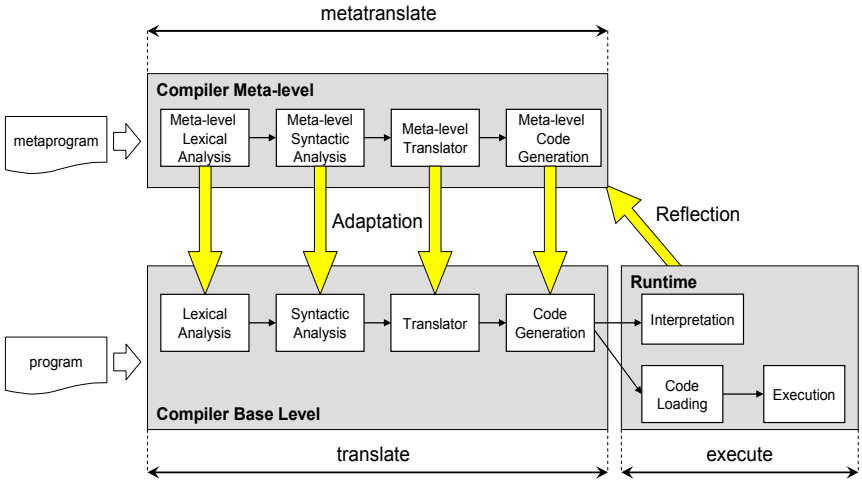


Fig. 2 Adaptive compiler infrastructure

Adoption was based on changing compiler grammar and the language translation rules. It is possible to provide feedback loops from any subsequent phase of language implementation to any preceding phase, via metadata. For example, it is possible to replace names by feedback loop from translator, interpreter, machine code generator, loader, or even from target machine execution, since it is interpreted. Metalevel was formed by abstraction of various grammars (EBNFs) and translation schemes.

5. CONCLUSION

Presented adaptive translator illustrates the ability for further extensions. First, it is possible to provide feedback loops from any subsequent phase of language implementation to any preceding phase, via metadata.

The main contribution of this work, from the viewpoint of our future research, is as follows:

- Domain specific languages can be developed as adaptive language systems for rather metalevel domains than application domains.
- Provided that some level or metalevel is adaptive, it contains feedback loops from data to code via metalevel or metametalevel.
- Even if any level or metalevel is adaptive, it must be still manually initiated. By the way, it is a base principle of control systems. The task of adaptive systems is to reduce this manual work, or shift it to the metalevel or metametalevel.

References

[1] BARÁTH, J., HARAKAĽ, M., DEDERA, Ľ.: Moodle – Experience from Exploitation in the Academy of the Armed Forces. In: CATE 2007, Simulation and Distance Learning 2007: International scientific conference, Brno, Czech Republic, 2007.

[2] CZARNECKI, K., EISENECKER. U.: Generative Programming: Methods, Tools, and Applications. Addison-Wesley Professional, 2000, 864 pp. ISBN 0201309777.

[3] EASTWOOD, A.: Firm Fires Shots at Legacy Systems, Computing Canada, 19(2), 1993, p. 17.

[4] ERLIKH, L.: Leveraging Legacy System Dollars for E-Business, (IEEE) IT Pro, May–June 2000, pp. 17–23.

[5] HEARNDEN, D., BAILES, P., LAWLEY, M., RAYMOND, K.: Automating Software Evolution. In Proceedings of the Principles of Software Evolution, 7th International Workshop on (IWPSE'04), IEEE Computer Society, pp. 95–100, 2004.

[6] JARZABEK, S.: Effective Software Maintenance and Evolution: A Reuse-Based Approach. Auerbach Publisher, 2007, 424 pp. ISBN 0849335922.

[7] Java Compiler Compiler (JavaCC) - The Java Parser Generator, https://javacc.dev.java.net, 24.10.2007.

[8] KOLLÁR, J., PORUBÄN, J., VÁCLAVÍK, P., BANDÁKOVÁ, J., FORGÁČ, M.: Adaptive Language Approach to Software Systems Evolution. In Proceedings of 1st Workshop on Advances in Programming Languages (WAPL'07), Wisla, Poland, October 15-17 2007, pp. 1081-1091, ISSN 1896-7094.

[9] LADDAGA, R., ROBERTSON, P., SHROBE, H.: Self-Adaptive Software: Internalized Feedback. Chapter 26 in Software Evolution and Feedback: Theory and Practice, Wiley, 2006, 612 pp. ISBN 0470871806.

[10] LEHMAN, M. M., RAMIL, J.F.: An Approach to a Theory of Software Evolution. In Proceedings of International Workshop on Principles of Software Evolution (IWPSE), Vienna, 2001, pp. 70-74.

[11] LÍŠKA, M., OČKAY, M.: General-Purpose Computing on Graphics Processing Units: New trends for computational acceleration In: KIT 2007, Tatranské Zruby, Liptovský Mikuláš, 2007.

[12] LUBERS, M., POTTS, C., RICHTER, C.: A Review of the State of the Practice. In Requirements Modeling, Proc. International Requirements Engineering Symposium, Los Alamitos, California, 1993.

[13] THOMPSON, S.: Haskell: The Craft of Functional Programming, 2nd Edition, Addison Wesley, 1999, 512 pp. ISBN 0201342758.

[14] PFLEEGER, S. L.: Software Engineering: Theory and Practice. 2nd edition, Prentice Hall, 2001, 659 pp. ISBN 0130290491.

[15] YANG, H., WARD, M.: Successful Evolution of Software Systems. Artech House Publishers, 2003, 300 pp. ISBN 1580533493.

doc. Ing. Ján KOLLÁR, CSc.
Ing. Jaroslav PORUBÄN, PhD.
Ing. Peter VÁCLAVÍK, PhD.
Ing. Jana BANDÁKOVÁ
Ing. Michal FORGÁČ
Fakulta elektrotechniky a informatiky
Technická univerzita
Letná 9 040 22 Košice
Slovenská republika
E-mail: Jan.Kollar@tuke.sk, Jaroslav.Poruban@tuke.sk, Peter.Vaclavik@tuke.sk, Jana.Bandakova@tuke.sk, Michal.Forgac@tuke.sk

This work was supported by VEGA project No. 1/4073/07 "Aspect-oriented Evolution of Complex SoftwareSystems"

GENERAL-PURPOSE COMPUTING ON GRAPHICS PROCESSING UNITS: NEW TRENDS FOR COMPUTATIONAL ACCELERATION

Miroslav LÍŠKA, Miloš OČKAY

Abstract: GPGPU is a promising trend of using parallel, computational power of GPU for general-purpose computing. We are presenting a simple comparison of CPU and GPU methods and rules, which are computing results. We are showing possibilities how to use GPU for general-purpose computing. Short list of applications is also included.

Keywords: GPU, GPGPU, General-purpose computing, Parallel computing, Compute unified device architecture, Close to metal.

1. INTRODUCTION

Computational power is required in many scientific and commercial applications. Complex simulations are performed on large data sets and results are required in short time. There are varieties of ways how to build a powerful high-performance system. Each uses different hardware architecture and software application interface (API) to achieve fast and accurate computing.

2. CPU

One of possibilities refers to use of parallel supercomputers. This system is comprised of multiple central processing units (CPU) and they are working together. A tightly coupled group of computers or supercomputers also makes a powerful system called computer cluster. Cluster is scalable and can produce a cost-effective high performance or an availability.

These systems are based on CPU (Figure 1). CPU is the component that interprets program instructions and processes data. A CPU, which is manufactured as a single integrated circuit, is known as a microprocessor. The form, design and implementation have changed since the earliest CPUs, but their fundamental operations have remained much the same. Early CPUs were built for specific purpose only. But the standardization trend set the way to mass-production of general purpose processors suitable for many applications. Modern microprocessors can be found in everything from cell phones to children's toys.

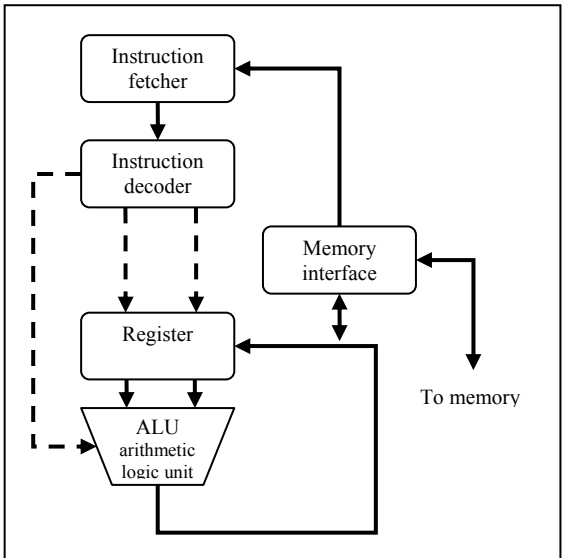


Fig. 1 Block diagram of simple CPU

There are four steps that almost all von Neumann's CPUs use in their operations: fetch, decode, execute and writeback. Fetch retrieves an instruction from memory. Instructions to be fetched must be retrieved from a slow memory, causing CPU to stall while waiting. This issue is solved by caches and pipeline architecture in modern processors. In decode step, instruction is broken up into several parts which are translated into various signals for CPU. After first two steps, the execution step is performed. Cooperation of various portions of CPU performs the requested operation and output will contain the final result. The last and final step, writeback, performs writing of the result to the memory. After the writeback, the entire process repeats again [8].

In more complex CPUs multiple instructions can be processed simultaneously by these four steps. Most CPUs are synchronous sequential devices. By calculating the maximum time that electrical signals can move in various branches of a CPU's circuits, we can estimate the appropriate period of synchronization clock signal. This period must be long enough to handle worst-case scenario of signal movement time. The previous mentioned have the advantage to simplify CPU, but CPU must wait for its slowest elements. This limitation has been compensated by using methods of increasing CPU parallelism. Variety of methodologies make CPU behave less linearly and more in parallel. Main two terms are: instruction level parallelism (ILP) and thread level parallelism (TLP). ILP seeks to increase the rate at which instructions are executed. TLP increases the number of threads as individual simultaneous programs. Parallelism is more natural way how to increase speed of CPU or whole computational system. It is easier to increase speed

by adding a new computational element, instead of increasing the frequency of clock signal.

3. GPU

Graphics processing unit (GPU) is the graphics rendering device for personal computers or other graphics systems. Modern GPUs are very efficient at manipulating and displaying computer graphics, and their highly paralel structure makes them more effective than CPUs. A GPU implements a lot of graphics operations, and running them is much faster than drawing them with common CPU. Rendering effect is performed by software instructions called shaders. Traditional GPU architecture uses three types of shaders: vertex, geometry and pixel. Vertex shader affects only the shape of an object. Geometry shader is used to combine a series of vertices into an object that can be affected by pixel shaders. Pixel shaders affect individual pixels to apply textures, bump maps and effects. These types of shaders are processed within GPU pipeline (Figure 2). Whole process looks like this:

- CPU sends geometry data to GPU
- vertex shader transforms geometry and some lighting calculations are performed
- if geometry shader is included in GPU, some changes of geometry are done
- triangles are transformed to quads
- pixel shader is applied
- visibility test and memory write is performed.

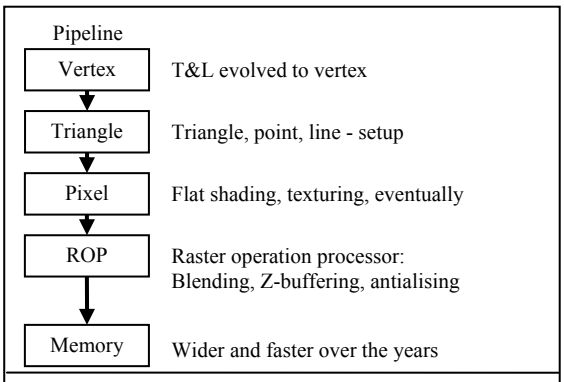


Fig. 2 Traditional GPU Architecture

Unified shader model uses the same instruction set across all shader types, instead of pixel and vertex shaders which each uses different set of instructions. Single floating point processor can work on both pixel and vertex data, as well as new types of data such as geometry, physics and more (Figure 3). The DirectX10 specification unifies the programming specified for vertex, geometry and pixel processing for a better fit for unified shader hardware, providing a unified pool of programmable resources [7].

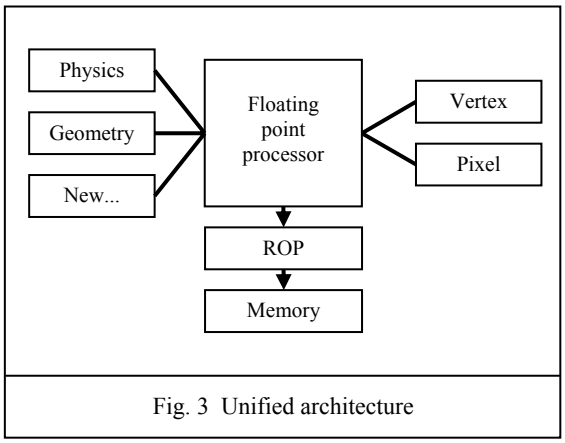


Fig. 3 Unified architecture

4. GPGPU

General-Purpose Computing on Graphics Processing Units (GPGPU) is a trend in computer science that uses the GPU to perform the computations rather than CPU (Figure 4 and 5).

GPU has a massive, floating-point parallel computational power which can be turned into a general-purpose computing power. Main problems using GPUs as general-purpose computing devices were as follows: chips were designed for video game development, the programming model was unusual, the programming environment was tightly constrained and underlying architectures were usually secret. APIs were optimized for graphics operations. Graphics driver was designed to hide hardware. The main purpose of using GPU was to free CPU from handling graphic processing.

Main programming APIs that GPGPU can use are OpenGL and Direct3D. OpenGL is often used in the academic community due to the platform portability and also extension mechanism, which allow vendors to add new features to API as soon as the hardware supports them [9]. DirectX/Direct3D is used in the computer game industry and it is Windows dependant [10]. Either API works perfectly well for GPGPU. GPGPU is only effective at tracking problems that can be solved by using stream processing. GPUs are stream processors that can operate in parallel by running a single kernel on many records in a stream at once. A stream is simply a set of records that requires similar computation. Kernels are the functions that are applied to each element in stream. In GPU, vertices and pixels are the elements in the streams, while vertex and pixel shaders are the kernels to be run on them. For each element we can only read form input, perform operations on it, and write to the output. An important parameter is memory access latency because it will limit computational speed.

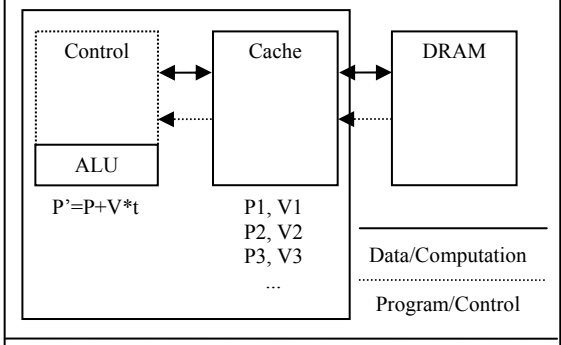


Fig. 4 Problem solved with CPU

Ideal GPGPU applications have large data sets, high parallelism and minimal dependency between data elements [1].

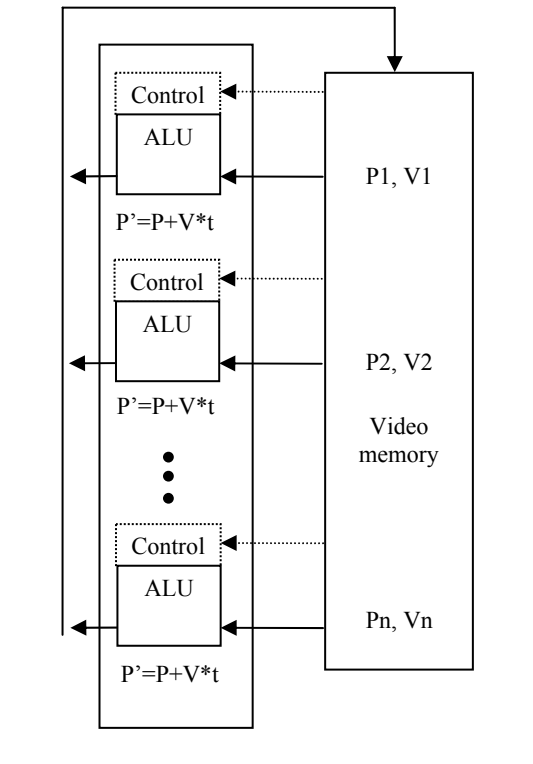


Fig. 5 Problem solved with GPGPU

Recently, several streaming languages have been developed to be used for GPGPU. Brook is an extension of the C language efficiently used for high intensity arithmetic GPGPU operations. Sh is library C++ extension for graphical and general-purpose computations. It is available for downloading under an open-source license [2].

5. APPLICATIONS OF GPGPU

The following examples are showing possibilities how to use GPU for general-purpose processing:

5.1 Physical based simulations

Physical based simulations are usually based on Newtonian physics models and they use GPU as computational tool. At the beginning GPU was used to simulate dynamic phenomena that can be described by partial differential equations, such as boiling or chemical reactions. Kim and Lin used GPUs to simulate ice crystal growth [12]. Lattice-Boltzmann Methods were used for fluid flow and gas simulations. Next step implements full floating point support in GPUs thus enabling possibility to use finite difference and finite element techniques for the solution of systems of partial differential equations [11]. These were used to implement basic cloth and particle systems simulations.

5.2 Signal and image processing

GPU computational power has made graphical hardware an attractive target for audio, video and other signals processing. The most notable application are those related to 2D and 3D segmentation. The segmentation method seeks features embedded in 2D and 3D images [11]. This is commonly used in medical applications such as Computed Tomography where digital geometry processing is used to generate 3D image of an object from large series of 2D X-ray images taken around a single axis of rotation. Signal processing is also used in ray tracing, global illumination and other methods in computer graphics. Ray tracing is a general technique from geometrical optics modeling the path taken by the light by following rays of light as they interact with an optical surface. It is used in 3D rendering systems to produce realistic, high quality output images. Global illumination improves the visual quality of GPU generated images.

5.3 Computer cluster and grid computing

Computer cluster is a parallel system which allows the possibility to use more than one PC to compute the same task. Computers in cluster are computational nods which communicate by passing messages within a cluster. Effectiveness of nods can be increased by adding GPUs as general-purpose computational elements. Grid architecture is also a massive parallel system and GPUs is considered as a source for this virtual computer architecture.

6. SOME OTHER POSSIBLE SOLUTIONS

Major computer graphic cards vendors create programming environments for their GPUs. nVidia came with Compute Unified Device Architecture (CUDA). CUDA is closely bonded with nVidia's

G80 unified shader architecture and brings a complete development solution to stream computing. nVidia makes C compiler available for the parallel GPU applications programming [5]. Applications functionality will scale with new GPUs. The goal of development environment is that developers do not need to learn new language to be able to develop GPU applications. CUDA would be able to share data faster, using shared data cache instead of video memory. GeForce 8800 compared to dual core Conroe running at 2.67 GHZ shows a significant GPU advantage. In some applications GPU is 10 times or much faster if compared with CPU [6].

AMD/ATI also come with their own GPU for general purpose solution called Close to metal (CTM) [3]. CTM is an interface for stream processor products [4]. CTM gives developers an access to the native instruction set and memory of the parallel computational elements in AMD stream processor (general-purpose product) and Radeon series of GPUs. CTM is much more low-leveled than CUDA, but gives a close access to the hardware which is necessary to develop high-level programming tools such as compilers, debuggers, math libraries and many others.

7. CONCLUSION

GPU has become more popular as general-purpose computational system since it was moved to a unified architecture and increased programmability. New solutions are available. There is much more compatibility in graphics gaming world (DirectX 10) and GPGPU than in specific GPU solutions (CUDA, CTM). There is a possibility how to use high-level CUDA for CTM hardware with some third-party middleware in the future or integrate both of them in multi-paradigm systems [13] thus increasing the interoperability of applications.

ACKNOWLEDGEMENT

This work was supported by Academic Grant Agency the Academy of the Armed Forces project No. AGA-01-2007 "Acceleration Technologies for High-Performance Computing".

References

[1] GPGPU, 2007. General-purpose computation using graphics hardware homepage. http://www.gpgpu.org
[2] GPGPU, 2007. The Official GPGPU FAQ http://www.gpgpu.org/wiki/FAQ
[3] AMD/ATI, 2007. ATI CTM Guide. http://ati.amd.com/companyinfo/researcher/documents.html
[4] AMD/ATI, 2007. AMD Stream processor product site. http://ati.amd.com/products/streamprocessor/index.html
[5] NVIDIA, 2007. nVidia CUDA homepage. http://developer.nvidia.com/object/cuda.html
[6] NVIDIA, 2007. nVidia G80 architecture reviews and specification. http://www.nvidia.com/page/8800_reviews.html http://www.nvidia.com/page/8800_tech_specs.html
[7] Beyond3D, 2006. NVIDIA G80: Architecture and GPU Analysis http://www.beyond3d.com/content/reviews/1
[8] Wikipedia, 2007. Wikipedia – The free encyclopedia http://en.wikipedia.org
[9] OpenGL, 2007. OpenGL homepage. http://www.opengl.org/
[10] Microsoft, 2007. DirectX resource center. http://msdn2.microsoft.com/en-us/xna/aa937781.aspx
[11] OWENS, D. J., LUEBKE, D., GOVINDARAJU, N., HARRIS, M., KRÜGER, J., LEFOHN, E. A., PURCELL, T.: A Survey of General-Purpose Computation on Graphics Hardware. In The Eurographics, 2005.
[12] KIM, T., LIN, M.C.: Visual simulation of ice crystal growth. In ACM Siggraph, 2003.
[13] KOLLÁR, J., PORUBÄN, J., VÁCLAVÍK, P., TÓTH, M., BANDÁKOVÁ, J., FORGÁČ, M.: Multi-paradigm Approaches to Systems Evolution, Computer Science and Technology Research Survey, Košice, Elfa s.r.o., 2007, 1, pp. 6-10, 978-80-8086-046-2

Summary: The goal of our contribution was to show new possibilities of accelerating non-graphical, general-purpose computations by using GPUs. GPUs are source of massive parallel computational power that can be used to solve large set of problems.

prof. Ing. Miroslav LÍŠKA, CSc.
Ing. Miloš OČKAY
Department of Informatics the Armed Forces of general M. R. Štefánik
Demänová 393
031 01 Liptovský Mikuláš
Slovak Republic
E-mail: liska@aoslm.sk,
ockaym@aoslm.sk

INFORMATION RETRIEVAL BY ART NEURAL NETWORKS

Igor MOKRIŠ, Roman KRAKOVSKÝ

Abstract: The paper deals with the ART neural networks with unsupervised learning based on adaptive resonance theory ART for processing of text documents in nature language. The paper is focused on the ART neural network description, principle of adaptive resonance and its separate phases in t he learning process. Next, the paper continues with utilization of ART neural networks in the text documents processing with respect to the clustering of document by sequence of keywords. On the end the paper introduces description of an algorithm for automatic generation of ontological construction by means of the projective ART neural network PART based on the association of keywords which is performed by Bayesian network.

Keywords: ART neural networks, clustering of text documents, association of keywords and documents, Bayesian network, ontology construction.

1. INTRODUCTION

In the relation with increase of information expansion on internet the new worldwide phenomenon – abundance of information was constituted. Majority of information is published in natural language. Most often of information is presented in the text documents. Demand for more powerful tools for administration, storage, searching and retrieving information from text documents according to the user requirements is growing. Keywords are usual means to formalization of documents for information retrieval on internet. Nowadays the research in this field focuses on information retrieval from text documents in natural languages based on keywords and its context.

A lot of approaches are proposed for information retrieval systems for text documents. In practice, mainly Boolean, vector space, probability and linguistic models for representation of information retrieval systems were applied. Except them, lots of models based on neural networks exist there. The main advantage of neural networks is the fact that very often it is difficult to find out the relationship between variables, describing processes of information retrieval from text documents. It is possible to avoid these methods by neural networks (NN), the structure of which represents the model or system structure for information retrieval.

When talking about the existing information retrieval systems mainly the feed-forward neural networks were applied using vector space model and its reduction by means of latent semantic model for representation of documents and keywords. The disadvantage of this approach is huge dimension of matrix keywords appearance for a considerable large amount of documents and inconvenience to express the keywords inside the structure of feed-forward neural network. This kind of networks realizes decision rules for classifying process based on supervised learning. Second principle of classification process is based on the unsupervised learning. It seems that recurrent neural networks are in frame to solve this problem. Inside its structure they realize a predict principle or association principle of clustering, based on unsupervised learning, allowing making the neural network structure smaller and decreasing exact enumeration for information retrieval models. Last years the research of recurrent NN in this field is oriented into ART neural networks.

2. ART NEURAL NETWORKS

A considerable feature of humane memory is an ability to learn new pieces of knowledge without forgetting the information already received. Unsupervised neural network without forgetting „older knowledge“ before invented in 1976 mathematician and neurobiologist S. Grossberg and together with G. Carpenter developed in 1986 to ART1, in 1987 to ATR2 and in 1990 to ART3 neural network [1].

Adaptive resonance theory (ART) was evolved for pattern recognition. It applies the principle of associative memory with unsupervised learning. Models of ART NN could be compared with an adaptive version of k – nearest neighbor method. The main advantage of ART NN is an ability to change stabile and plastic mode without damaging the learned information. Neural network works in a plastic mode, when weight can be modified. A stable mode is state of network when the net has fixed preferences and works as a classifier. Another plus of ART NN is a context sensitivity and ability to eliminate erroneous data [7].

Architecture of ART NN consists of two layers: (fig.1) F_1-compare layer and F_2 -competitive layer [3]. F_1 layer can be divided into two parts: input sub-layer F_{1a} and interface sub-layer F_{1b}. The terms compare and competitive layer are used because in processing of neural network the function of input layer changes to output layer and conversely. Both layers are connected by bottom-up weights w_{ij} and top-down weights t_{ij}, for i = 1,2,3..N to be formed

input neurons and j = 1,2,3..M is amount of output neurons (number of clusters).

ART NN utilizes the principle of back propagation of signal from output layer to the input layer and so propagation signals between nodes inside comparative sub-layer. Beside these sub-layers there exists a supplemental control unit (fig. 2), checking a stream of data in network and performing vigilance test. Between both layers the reset unit R is present, which controls input pattern with vigilance matching. The success in vigilance test determines if input pattern should be clustered in an existing class or a new one should be created.

There are several phases in ART NN during the process of learning and clustering of patterns. It differs from feed - forward neural network in input pattern passing repeatedly between input and output sub-layer and cluster weights are adjusted to let the cluster unit learn pattern. It is described as adaptive resonance, where the weights are adapted in every circulation of patterns between both layers until the condition of the net is stable.

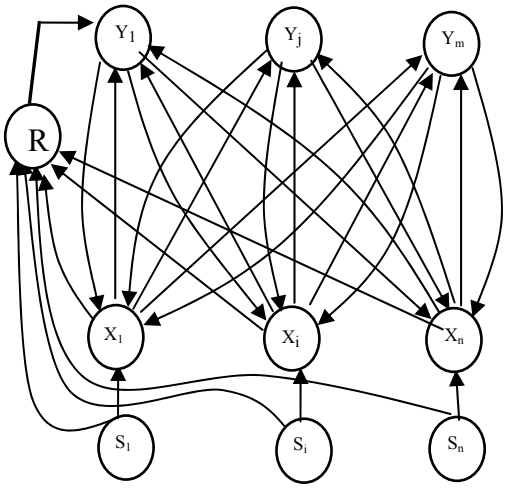


Fig. 1 Structure of ART neural network

Pattern processing in ART NN include this stages [9]:

1. Initialization stage – initialize parameters, input and output vectors. Gain control signal G_1 is adjusted to value 1, that means it is activated. This signal has to control the direction of stream pattern in the neural network. If it is set to G_1=1, the input sub-layer is ready to receive input pattern from environment. Failing that, the network is set to recognition phase and some neuron from recognition layer is activated. Gain control signal G_2 can acquire two stages too. If the signal is set to G_2=1, input pattern is considered to be recognizable. Otherwise pattern didn't pass through vigilance test and each node of F_2 layer is inactivated. Bottom-up weights are set to value $w_{ij} = 1/(1+n)$, where N means number of input nodes. The vigilance parameter ρ has rate between 0 and 1.

2. Recognition stage – signals from input sub-layer are transformed through bottom-up weights towards to recognition layer applying formula $y_j = \Sigma w_{ij} x_i$. Every unit in F_2 layer of ART NN has three sources: signal from F_1 unit (input signal), from F_2 node (a top-down signal) and from G_1 unit. Since there are three possible sources of signal, this requirement is called the two/three rule, where unit must receive two excitatory signals in order to be ‚on'. Input pattern is oscillating among different cluster unites on the output layer and then the neuron nearest to the input vector will be chosen. Among output nodes the side inhibitory signals impact. It means that the winner neuron is supported and the rest of nodes are deactivated. Each neuron possesses the feedback which increases its own output signal. The result of combination of the side inhibition and the feedback is that only one neuron will stay active in a portion. At a moment of the winner neuron recognition, this is sent to compare sub-layer.

3. Compare stage – at a gate of the compare layer two vectors are presented. Reset unit R is responsible to compare of input pattern and compare vector similarity and the achieved result S is subsequently compared with vigilance parameter ρ. When the ratio is S > ρ, responsible cluster for input pattern was found and the control signal G_2 is set to value 1. Failing that, any similar cluster vector was found and the network is trying to find another cluster vector.

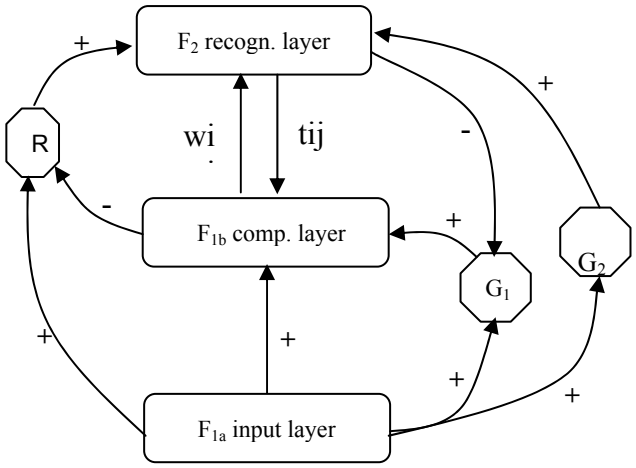


Fig. 2 Principle of adaptive resonance in ART NN (+,- indicate excitatory/inhibitory signals)

4. Trace stage – process continues until all units are inhibited or a satisfactory match is found (a candidate is accepted) and its ratio is higher than or

equal to the vigilance parameter. If none of them passes the vigilance test, a new prototype is created for the current input pattern. This is possible only when it did not run out of total number of clusters that can be formed.

5. Adapt stage – if the winning pattern passing the vigilance test has been found, then the bottom-up and top-down weights for winner node will be updated.

2.1 Division of ART neural networks

Basic division of ART NN is as follows [10]:

- ART1 – unsupervised learning, clustering of binary input pattern,
- ART2 – clustering of binary and analog input pattern,
- ART3 – into ART theory is included model of chemical synapses,
- Fuzzy ART – applies fuzzy sets to ART NN,
- DART–distributed ART NN for unsupervised clustering of analog pattern ,
- Gauss ART – applying Gauss classifier in ART NN,
- ARTMAP – modification of ART NN for supervised learning.

3. UTILIZATION OF ART NEURAL NETWORKS IN TEXT DOCUMENT PROCESSING

ART neural networks can be utilized for documents clustering. A disadvantage of classic ART networks in clustering is that one document is not reliable to be classified into more than one cluster. Modified version of the Fuzzy ART NN, permitting one document to belong to multiple clusters is called KMART system [6]. Document processing by means of this system consists of next stages: pre-processing, where all stop-words and redundant words are removed from all documents. Second stage is used for cluster building. In final step representative keywords for each cluster formed in the previous stage are determined and displayed. In practice it is approximately first 7-10 words as keywords from each cluster are chosen. Inner representation of documents uses the TF – idf weighting scheme.

Testing of KMART system was parallel conducted with processing the same documents with clustering algorithms like Fuzzy ART, SISC, K-Means and Fractionation [8]. Using experiment with 2000 documents from World Wide Web belong to different categories, reached KMART with Fuzzy ART better quality formed cluster. It was compared the cluster formed by the documents against the documents in the original categories and matched the clusters with categories one -to-one. It was also compared the execution times of approach, whereupon of KMART system was linear with the number of documents.

4. TEXT DOCUMENTS PROCESSING BY LINGUISTIC APPROACH AND ART NEURAL NETWORKS

Nowadays for processing of documents knowledge approach based on ontology is used. The purpose of ontology is to describe text documents and their structure by means of domain representation from collection of documents [4]. Reuse domain representation from documents can be achieved by linguistic approach due to ontology languages including XOL, OWL, SHOE, RDF, DAM+OIL [12].

Documents retrieval through context of keywords including ontology and neural network was published in [5]. The study presented novel, automatically generated ontology construction, based on documents processing through artificial neural network and association keywords with documents due to Bayesian network [13, 14, 15].

Methodology of used approach can be divided into next steps:

1. First of all web pages relative to the problem domain were chosen.
2. Next the labels from HTML tags to select keywords were utilized and used WorldNet [11] to determine meaningful keywords called terms.
3. Process continued by calculating of weight entropy of terms and deleted keywords, whose entropy value was equal 0.
4. After above steps a projective adaptive resonance theory (PART) neural network for generating clusters of keywords from documents was used.
5. Finally, the Bayesian network to express the hierarchical nearness between keywords in the documents was used and relation to the ontology construction of document was found.

To analyze pages context blocks and discover information context the entropy can by applied. An advantage of Shannon's information entropy was taken to calculate the keywords entropy, based on keyword web pages matrix. Matrix stored the frequency of selected keywords appearing in documents, applying entropy formula

$$E(T_i) = -\Sigma P_{ij} \log P_{ij} \quad (1)$$

where P_{ij} means the probability, that i-keyword appears in the j web page.

Kernel technology on whole system is Projective Adaptive Resonance Theory (PART) neural network. In order to deal with the feasibility – reliability dilemma in clustering data sets, Yongqiang Cao a Jianhong Wu proposed a new neural network architecture – Projective Adaptive Resonance Theory in 2002 [2]. The main difference between PART and ART neural network is in the input layer. Besides the vigilance test, the PART NN adds the distance test to increase the accuracy of clustering. Matrix, inputted to PART is TF-matrix, considered from terms and documents. Output from PART is tree architecture, where the recursion was added, based on the threshold value. After the PART tree process, a basic tree structure was gotten, that can be used to represent all web pages and Bayesian network was applied to construct complete domain representation. The system finally output an ontology using RFD format through Jean package. Architecture of system for automatically generated ontology is shown in fig. 3.

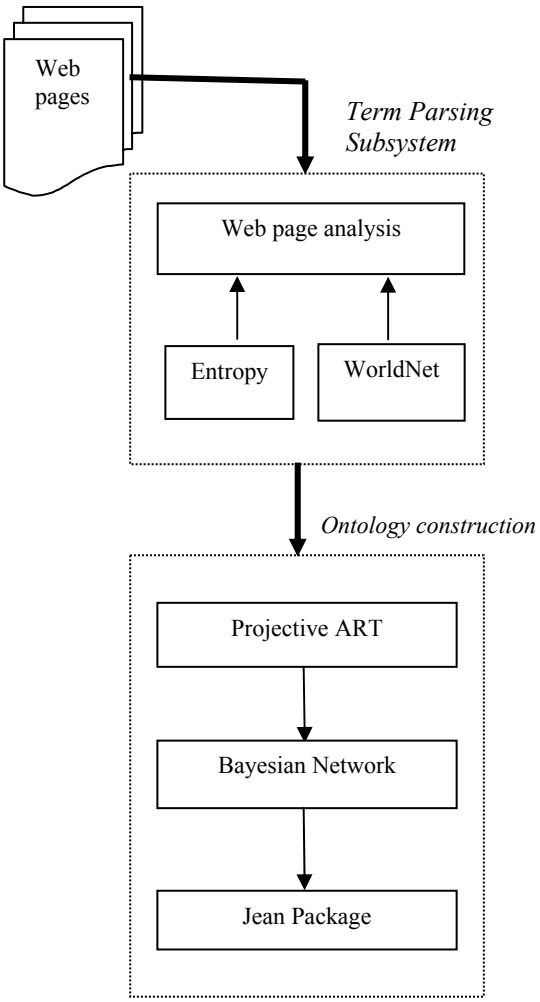


Fig. 3 The ontology generation system architecture

Automatic ontology construction based on Projective PART NN and Bayesian network overcomes lacks of flexibility in clustering, WorldNet and entropy deal, with the lack of knowledge acquisition for future processing.

5. CONCLUSION

Nowadays, assertion of ART neural networks is topic of the research in the field of text document processing. Main advantage of ART NN is possession botch features of stability and plasticity with ability to learn new pieces of knowledge without forgetting older knowledge. Incorporation of this neural network with some known text processing's system could bring better results. All published algorithms and text processing systems on web were used for English language based documents. Next research of information retrieval from text documents is focused to national languages – in our case to Slovak.

References

[1] CARPENTER, G. A., GROSSBERG, S.: Adaptive Resonance Theory. The Handbook of Brain Theory and Neural Networks. MIT Press, 1998.
[2] CAO, Y., WU, J.: Projective ART for Clustering Data Sets in High Dimensional Spaces. Neural Network, No. 15, 2002, pp. 105-120.
[3] FAUSETT, L: Fundamentals of Neural Networks. Prentice Hall, 1994.
[4] HREŇO, J: Generation of Document Abstracts by Domain Approach. [PhD Thesis]. FEI TU Košice, 2006, (in Slovak).
[5] CHEN, R. C., CHUANG, C. H., TSENG, C. C.: Constructing an Ontology Automatically by Projective ART Neural Network. Proc. of Int. Computer Symp. ICS 2006, Taipei, Taiwan, Dec. 4 -6, 2006.
[6] KONDADADI, R., KOZMA, R.: A Modified Fuzzy ART for Soft Document Clustering. Int. Joint Conf. on Neural Networks, Honolulu, Hawaii, 2002, pp. 2545-2549.
[7] KVASNIČKA, V. et al..: Introduction to Neural Network Theory. IRIS 1997, (in Slovak).
[8] LIN, K., KONDADADI, R.: A Similarity - Based Soft Clustering Algorithm for Documents. 7-th Int. Conf. on Database Systems for Advanced Applications (DASFAA), 2001, pp. 40-48.
[9] MARIK, V., et al.: Artificial Intelligence. Academia Praha 2003, (in Czech).
[10] http://web.umr.edu/~tauritzd/art/overview.html, (Boston university, Department of Cognitive and Neural systems).

[11] WorldNet, http://worldnet.princeton.edu.
[12] W3C, http://www.w3c.org.
[13] DENOYER, L., GALLINARI, P.: Bayesian Network Model for Semi-Structured Document Classification. Information Processing and Management, Vol. 40, 2004, pp. 807–827.
[14] MURPHY, K.: A Brief Introduction to Graphical Models and Bayesian Networks, 1998, www.cs.ubc.ca/~murphyk/Bayes/bayes.html
[15] FERRIS, B., FRIEDMAN, S.: Identifier Labeling Using Graphical Models, 2005, www.cs.washington.edu/homes/bdferris/papers/IdentifierLabeling.pdf

ACKNOWLEDGEMENT

This work was supported by Slovak Science and Techn. Assist. Agency under the contract No. APVT-51-024604 and Slovak Science Agency VEGA No. 2/7098/27.

prof. Ing. Igor MOKRIŠ, PhD. [1)]
Ing. Roman KRAKOVSKÝ [2)]
[1)] Institute of Informatics, Slovak Academy of Sciences
Dúbravská cesta 9
845 07 Bratislava
Slovakia
E-mail: mokris@aoslm.sk
[2)] Dept. of Informatics,
Pedagogical faculty of the Catholic University
Námestie A.Hlinku 56/1
034 01 Ružomberok
Slovakia
E-mail: krakovsky@fedu.ku.sk

ANALÝZA PREDIKČNÝCH METÓD STAVU MOBILNÉHO RÁDIOVÉHO KANÁLA

ANALYSIS OF MOBILE RADIO CHANNEL PREDICTION METHODS

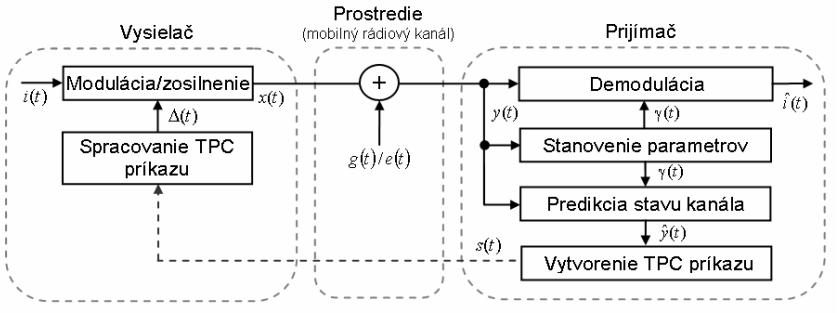
Vladimír PŠENÁK, Vladimir WIESER

Abstract: The effective handling of the radio resources is important to keep declared quality of service (QoS). A basic hybrid link adaptation algorithm updates technical link parameters according to the delayed feedback information from receiver. The fast power adaptation if essential for the third generation of mobile radio network UMTS, because active users are sharing the same carrier frequency (WCDMA). A prediction of future channel state can be used to eliminate feedback delay and power control command transport delay; therefore the link adaptation becomes more efficient. We have analyzed and simulated prediction methods used for prediction of the mobile radio channel state in this article. Implementation of proposed methods into the hybrid link adaptation algorithm should increase the efficiency of data transmission among user equipment and base stations (uplink).

Keywords: hybrid link adaptation, mobile radio channel, prediction methods.

1. ÚVOD

V mobilnej rádiovej komunikácií je efektívne využívanie prideleného frekvenčného pásma jedným z najdôležitejších cieľov. Manažment rádiových zdrojov má k dispozícií obmedzené množstvo rádiových prostriedkov, ktoré je potrebné rozdeľovať efektívne, jednak aby bola udržaná požadovaná kvalita služby (QoS) a zároveň bol obslúžený maximálny počet účastníkov. V mobilných rádiových sieťach založených na viacnásobnom prístupe s kódovým delením CDMA (3GPP UMTS [1]) je eliminácia dlhodobého a krátkodobého úniku signálu adaptáciou technických parametrov (výkonu stanice, modulácie a kódovania) nevyhnutnou podmienkou pre správnu a efektívnu činnosť systému. Kritériom hybridného algoritmu adaptácie spoja musí byť merateľná veličina: bitová chybovosť, pomer signál-interferencia, avšak takýto algoritmus pracuje s dopravným oneskorením, keďže rozhodnutie o zmene technických parametrov vysielača je závislé od spätnej väzby (v podobe príkazu riadenia výkonu) od prijímača. Doplnením rozhodovacieho kritéria o predikovanú informáciu o stave rádiového kanála je možné nepresnosť adaptácie spoja oneskorením obmedziť. Zjednodušený model vzostupnej adaptácie spoja s prediktorom, ktorý vychádza zo systému 3G UMTS je znázornený na obrázku 1 [1], [6]. Zmena výstupného výkonu vysielača o $\Delta(t)$ je riadená TPC príkazom $s(t)$, ktorý je v prijímači generovaný na základe informácie o aktuálnom stave kanála $\gamma(t)$ a predikovanej hodnoty $\hat{y}(t)$. Výstupný signál $x(t)$ je ovplyvňovaný rádiovým kanálom $g(t)$ a šumom $e(t)$ a takto ovplyvnený je tento signál prijatý prijímačom ($y(t)$).



Obr. 1 Zjednodušený model vzostupnej adaptácie spoja s prediktorom

2. VLASTNOSTI MOBILNÉHO RÁDIOVÉHO KANÁLA

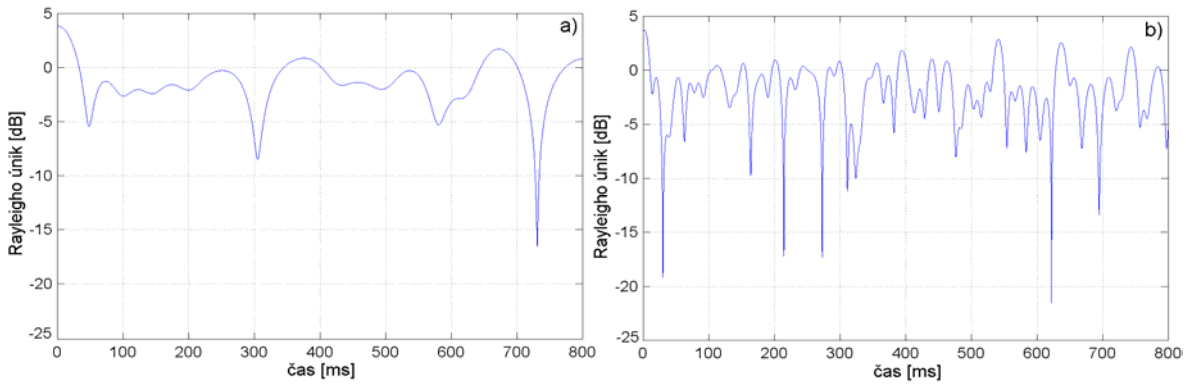
Z hľadiska aplikácie predikčných metód môžeme mobilný rádiový kanál modelovať a popásat jako veľké množstvo horizontálnych rádiových vĺn, ktoré sú prijímané z rôznych smerov. Ak vezmeme do úvahy aj zmenu polarizácie, rôzne vertikálne uhly príjmu a tiež predpoklad, že prijaté vlny nie sú rovinné, dostávame zložitý fyzikálny popis. Prístupnejší spôsob popisu je nahradenie mobilného rádiového

kanála dynamickým systémom, ktorý je popísaný postupnosťou časových vzoriek merania prijatého výkonu alebo postupnosťou komplexných vzoriek časovej impulznej odozvy, pričom táto postupnosť časových vzoriek je hlavnou vstupnou premennou každého prediktora. Podľa odporúčaní [2] delíme mobilný rádiový kanál na základe vlastností prostredia na 3 rôzne typy (tabuľka 1):

Tab. 1 Typy modelov mobilného rádiového kanála

Typ	Max. rýchlosť pohybu stanice	Počet ciest šírenia	Tlmenie ciest šírenia	Oneskorenie ciest
"Vozidlo"	500 km.h^{-1}	6	0 až -20.0 dB	0 až 20 000 ns
"Chodec"	120 km.h^{-1}	5 6	0 až -22.8 dB 0 až -23.9 dB	0 až 410 ns 0 až 3 700 ns
"V budove"	10 km.h^{-1}	6	0 až -32.0 dB	0 až 700 ns

Charakter mobilného rádiového kanála závisí od typu prostredia (počtu prekážok spôsobujúcich rozptyl alebo odraz, t.j. viaccestného šírenia) a taktiež od rýchlosti pohybu mobilnej stanice MS v [m.s^{-1}] (Dopplerov efekt). Aplikované vzťahy popisujúce daný model a výsledky modelovania mobilného rádiového kanála dokazujú, že s narastajúcou rýchlosťou pohybu narastá intenzita krátkodobých únikov signálu (obrázok 2), ktoré majú na adaptáciu spoja najnepriaznivejší vplyv.



Obr. 2 Príklad krátkodobého úniku v mobilnom rádiovom kanále:
a) rýchlosť MS $v = 2{,}8$ m.s^{-1}, b) rýchlosť MS $v = 11{,}1$ m.s^{-1}

3. ROZDELENIE PREDIKČNÝCH METÓD

Predikčné metódy (podľa predikovanej veličiny) delíme na predikciu stavu (dominantných) komplexných ciest šírenia signálu (v závislosti od počtu prstov RAKE prijímača) na základe komplexnej impulznej odozvy kanála a na predikciu celkového výkonu prijímaného signálu. Podľa časovej stálosti predikčných koeficientov delíme predikčné metódy na adaptívne a neadaptívne [3].

Základná neadaptívna metóda je lineárna predikcia komplexnej premennej (FIR prediktor), ktorej výhodou sú dobré zovšeobecnovacie vlastnosti. Tento prediktor je možné použiť pri predikcii jednotlivých komplexných ciest šírenia, pričom výsledná odhadnutá hodnota je dosiahnutá kombináciou a váhovaním predikovaných hodnôt jednotlivých ciest. Preto je do každého prstu RAKE prijímača potrebné implementovať samostatný prediktor, čím sa zvyšujú nároky na pamäť a výpočtové prostriedky.

Zo vzťahu pre jednorozmerný lineárny FIR prediktor zavedením komplexných predikčných koeficientov dostávame vzťah popisujúci FIR prediktor komplexnej premennej (dvojrozmerný) [3]:

$$\dot{\hat{h}}(n+L) = \sum_{i=1}^{M} \dot{\alpha}_i \dot{h}(n-i) \qquad (1)$$

Kde L je predikčný interval, $\dot{\hat{h}}(n+L)$ je predikovaná vzorka, M je počet predikčných koeficientov $\dot{\alpha}$ a $\dot{h}(n)$ je vstupná vzorka. Pre dosiahnutie relevantných výsledkov musí byť počet vstupných vzoriek (veľkosť pamäte prediktora) niekoľkonásobne väčší jako počet predikčních

koeficientov. Hodnoty koeficientov $\dot{\alpha}$ je možné stanoviť viacerými metódami, pričom pri simuláciách bola použitá metóda najmenších štvorcov.

Metóda priamej predikcie stavu kanála (výkonu) založená na jednorozmernom lineárnom FIR prediktore vychádza z informácie o celkovej intenzite prijímaného užitočného signálu, resp. kontrolných dát, ktorých vysielací výkon je prijímacej stanici (ZS) známy. Kvadratický prediktor pracuje s informáciou o celkovom prijatom výkone (suma všetkých čiastkových výkonov sledovaných ciest šírenia Q) užitočného signálu $p(n)$ [4]:

$$p(n) = \sum_{q=1}^{Q} p_q(n) = \sum_{q=1}^{Q} \left| h_q(n) \right|^2 \qquad (2)$$

Potom predikciu výkonu jednej cesty môžeme definovať nasledovne:

$$\hat{p}_q(n+L) = \left| \dot{\hat{h}}_q(n+L) \right|^2 + E\left[\varepsilon_{p,q}(n)\right] \qquad (3)$$

Kde $E\left[\varepsilon_{p,q}(n)\right]$ je stredná hodnota chyby predikcie ε, ktorá vyjadruje rozdiel medzi skutočnou a predikovanou hodnotou.

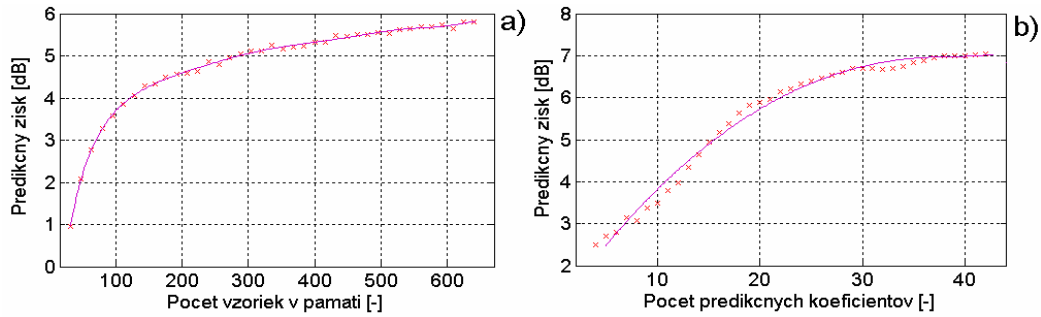
U vyššie uvedených predikčných metód je predikčný interval v nepriamej úmere so vzorkovaciou frekvenciou vstupných hodnôt, teda zvyšovanie predikčného intervalu znamená znižovanie vzorkovaciej frekvencie, čím zároveň dochádza ku strate dôležitej informácie o charaktere predikovanej veličiny. Túto nevýhodu čiastočne kompenzuje priamy iteratívny prediktor, ktorý patrí medzi adaptívne predikčné metódy. Princíp iteratívneho prediktora spočíva v použití predikovanej hodnoty získanej v predchádzajúcej iterácii pre predikciu ďalšej hodnoty. Medzi jeho hlavné výhody patrí možnosť odhadu hodnôt vo viacerých predikčných intervaloch v jednom cykle, bez potreby opätovného návrhu predikčných koeficientov a prevzorkovania vstupných hodnôt. Vzájomné porovnanie predikčných metód je vykonané na základe dosiahnutého predikčného zisku $G(L)$ [4]:

$$G(L) = 10\log_{10} \frac{E\left[\left(h(n) - E\left[h(n)\right]\right)^2\right]}{E\left[\varepsilon^2\right]} \qquad (4)$$

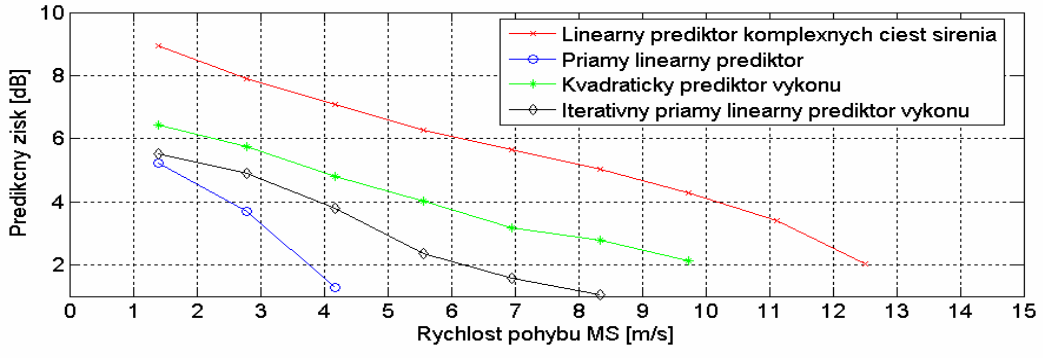
4. POROVNANIE VLASTNOSTÍ PREDIKČNÝCH METÓD

Frekvencia vzorkovania vstupných vzoriek bola stanovená na 10 kHz, čo predstavuje 100 vzoriek na jeden rádiový rámec s dĺžkou 10 ms. Uvažovaný simulačný WCDMA model môže zmeniť vysielací výkon stanice príkazom riadenia výkonu TPC raz počas trvania rádiového rámca. Vplyvom dopravného oneskorenia je TPC príkaz doručený s oneskorením 1 rádiového rámca vzhľadom na aktuálny stav vzostupného rádiového kanála, preto minimálny predikčný interval je $L = 10$ ms. V prípade jednoduchého prediktora je tento interval dosiahnutý podvzorkovaním vstupných vzoriek, v prípade iteračného prediktora je predikčný interval priamoúmerný počtu iterácií.

Na obrázku 3a) je vidieť, že v prípade aplikácie lineárneho prediktora komplexnej premennej pre odhad stavu mobilného rádiového kanála (chodec, 8 m.s-1) je nárast predikčného zisku, ak je veľkosť pamäte prediktora väčšia ako 200 vzoriek, minimálny (pričom ak $L = 10$ ms, potom získanie potrebných vzoriek trvá viac než 2 sekundy). Rovnako s rastúcim počtom predikčných koeficientov rastú nároky na výpočet koeficientov, ale od určitého počtu (približne 30) je nárast predikčného zisku minimálny, obrázok 3b).



Obr. 3 Závislosť predikčného zisku lineárneho prediktora komplexnej premennej: a) od veľkosti pamäte prediktora, b) od počtu predikčných koeficientov



Obr. 4 Závislosť predikčného zisku od rýchlosti pohybu MS

5. ZÁVER

Prezentované simulácie ukazujú (obrázok 4), že použitie jednoduchých predikčných metód pri adaptácií parametrov rádiového spoja má svoj význam v prípade, ak rýchlosť MS nie je príliš veľká s ohľadom na charakter prostredia rádiového kanála (typ prostredia: "v budove" alebo "chodec"). Aplikácia neadaptívnych a adaptívnych prediktorov na navrhnuté simulačné modely kanála vykazovala priaznivé výsledky (predikčný zisk) pri rýchlostiach MS do 50 km.h^{-1}. Pre vyššie rýchlosti je potrebné uvažovať o predikčnej metóde, ktorá nielen aproximuje charakter zmien kanála (krivku krátkodobých únikov), ale aj modeluje jeho vlastnosti v rámci definovaných podmienok. Pri výbere vhodnej predikčnej metódy je potrebné brať ohľad aj na implementovateľnosť do algortimu riadnia spoja, aby dosiahnutý predikčný zisk bol efektívne využitý [5].

Zoznam bibliografických odkazov

[1] CASTRO, Jonathan. The UMTS Network and Radio Access Technology. Anglicko: Willey, 2004 ISBN 0 471 813753

[2] European Telecommunications Standard Institute ETSI TR 101 112 V3.2.0 (1998-04) Universal Mobile Telecommunications System (UMTS): Selection procedures for the choice of radio transmission Technologies of the UMTS

[3] EKMAN, T. Prediction of mobile radio channels. PhD thesis, Uppsala University, Sweden, 2001.

[4] EKMAN, T. Prediction of mobile radio channels – modeling and design. Dissertation for the degree of Doctor of Philosophy in Signal Processing at Uppsala University, Sweden, 2002, ISBN 91-506-1625-0.

[5] WIESER, V., PŠENÁK, V. BER and SIR based hybrid link algorithms performance in mobile radio channel. In. Radioenginnering No.4, Vol. 14, December 2005, p. 81-86, ISSN 1210-2512.

[6] HOLMA, H., TOSKALA, A. WCDMA for UMTS. Radio access for third generation mobile communications. Willey, England, 2000, ISBN 0-471-72051-8.

Summary: The main aim of our contribution was to present and compare applicable prediction methods to estimate future mobile radio channel state according to its previouse and current state. Simulations show that proposed prediction methods are usable while mobile station velocity is bellow 40km.h^{-1}.

Ing. Vladimír PŠENÁK[1)]
Assoc. prof. Vladimír WIESER, PhD.[2)]
[1)] SIEMENS Program and System Engineering s.r.o.
Hurbanova 21, 010 10 Žilina
E-mail: vladimir.psenak@siemens.com
[2)] Žilinská Univerzita v Žiline, Katedra telekomunikácií,
Veľký diel, 010 26 Žilina
E-mail: vladimir.wieser@fel.uniza.sk

Tento príspevok bol podporený vedeckou grantovou agentúrou VEGA v projekte č. 1/4067/07.

DRM BASED ON THE ROBUST DIGITAL WATERMARKING

Radovan RIDZOŇ, Dušan LEVICKÝ

Abstract: The geometrical attacks are still open problem for many digital watermarking algorithms used in present time. Most of geometrical attacks can be described by using affine transforms. This article deals with digital watermarking in images robust against the affine transformations. The new approach to improve robustness against geometrical attacks is presented. The discrete fourier transform and log-polar mapping is used for watermark embedding and for watermark detection. Some attacks against the embedded watermarks are performed and the results are given.

Keywords: digital watermarking, geometrical attacks, discrete fourier transform, log-polar mapping.

INTRODUCTION

Digital multimedia and digital multimedia processing have brought many advantages compared with the analog form of multimedia, for example easy processing and storage, compression and better noise resistance. However, the digital multimedia form established the problems with the ownership rights and making the illegal copies of the multimedia that are identical with the original and may be produced and transmitted easy and with low cost. Another problem is how to protect digital multimedia against unauthorized access during the transmission over communication networks, during processing and storage. There are two core technologies which can be used to protect multimedia content. Content protection of multimedia during the transmission can be solved by using cryptographic methods, which realized encryption of the content of the multimedia. Received multimedia after decryption in the receiver are not protected any more. Multimedia content protection and ownership rights after the transmission and decryption can be realized by digital watermarking. The idea of digital watermarking is embedding imperceptible information into the multimedia.

1. SECURITY OF THE MULTIMEDIA

Dynamic expansion of multimedia processing and transmission in digital form has brought the requirement of multimedia security and protection of ownership rights. These requirements can be divided into three basic groups: ownership rights protection, distribution of illegal copies and unauthorized access to multimedia. From this point of view the multimedia security can be divided into multimedia content protection during the transmission and multimedia content protection after the transmission.

Multimedia content protection during transmission can be realized by cryptographic methods, which secure the information content of multimedia by encryption. But cryptographic methods don't protect information content of multimedia after the decryption. After the decryption in the receiver the information content is no more protected and data may be copying easy and without the quality degradation.

Multimedia content protection after the transmission can be realized by digital watermarking, which performed embedding of the imperceptible information into the multimedia content and this information should not be easy removable by using the basic multimedia processing techniques.

Cryptographic methods and digital watermarking are basic techniques in the field that is called Digital Right Management (DRM). Digital right management is a collection of techniques and technologies that enable technically enforced licensing of digital information, secure transmission, authors and ownership rights for all types of multimedia.

2. IMPLEMENTATION OF THE MULTIMEDIA CONTENT PROTECTION AND OWNERSHIP RIGHTS IN THE DRM SYSTEMS

The core technologies used for DRM systems implementation into the multimedia are encryption techniques and digital watermarking techniques.

The multimedia encryption techniques encrypt the multimedia content with the goal to prevent the access to the multimedia content for unauthorized users.

The digital watermarking techniques in multimedia create metadata that contain information about the protected multimedia content, and then hide this metadata within the content.

Digital watermarking is process of embedding additional information directly into the digital multimedia, also called original data, by making small modifications to them.

Digital watermarking technologies allow users to embed digital code into audio, images, video and printed documents which are imperceptible during normal use but readable by computers and software. The additional information is called watermark. Watermark embedding as a form of the metadata is realized in DRM implementation block.

The process of the digital watermarking should match these four requirements: robustness of the embedded watermark, perceptive transparency of the watermarked data, watermark undetectability and security.

Robustness of the embedded watermark is the resilience of the watermark against the attacks performed by the unauthorized person. The goal of the attacks is to remove embedded watermark and obtain unwatermarked data without content protection.

Perceptive transparency of the watermarked data is the request for the embedded watermark imperceptibility. Watermark embedding should not causes multimedia quality degradation or visible or audible artifacts in the multimedia.

Watermark undetectability is the request for the statistical undetectability of the changes caused by the watermark embedding.

Security is the request for the impossible watermark extraction or removing from watermarked data without the knowledge of the embedded metadata or the key.

Secret or public key could be used during the watermark embedding. Usage of the key improves the security against the manipulation with watermarked data.

On the receiver side two processes can be realized: watermark extraction and watermark detection.

In the case of ***watermark extraction***, the embedded watermark is extracted from tested data and is compared with original watermark.

Watermark detection is the binary decision process. In this case only the presence or absence of the watermark in the tested image is confirmed.

Digital watermarking makes it possible in DRM systems to cover basic functions: author's rights protection in multimedia, authentization or data integrity check, copyright protection, transfer of the controlling and additional information.

In digital watermarking as tools for the protection ownership rights and copy prohibition, there are a lot of processes performed by unauthorized persons which aim to corrupt the embedded information. These processes are called attacks. There are various categorizations of attacks on watermarks. One from the categorization is categorization into four main groups:

- removal attacks,
- geometrical attacks,
- cryptographic attacks,
- protocol attacks.

Removal attacks achieve complete removal of the watermark information from the watermarked data without cracking the security of the watermarking algorithm. This category includes denoising, lossy compression, quantization, remodulation, collusion and averaging attacks.

Geometrical attacks do not remove the embedded watermark itself, but intend to distort the watermark detector synchronization with the embedded information. To the category of the geometrical attacks belong the cropping, flip, rotation, shift, scaling, and translation and so on.

Cryptographic attacks aim at cracking the security methods in watermarking schemes and thus finding a way to remove the embedded watermark information or to embed misleading watermarks. These attacks are very similar to the attacks used in cryptography. There are the brute force attacks which aim at finding secret information through an exhaustive search.

Protocol attacks aim at attacking the entire concept of the watermarking application. This category includes the copy attack and the attacks made by invertible watermarks.

The geometrical attacks on the digital watermarks are still open problem for a many watermark algorithms used in present time. A few approaches to improve the robustness against geometrical attacks have been presented in many papers.

The methods capable to estimate and recover the undergone global affine transformations can be divided into 3 main groups:

Invariant watermarks. The transform invariant domain approach mostly consists in the application of the Fourier-Mellin transform to the magnitude of the original (or cover) image spectrum, associated with a log-polar or a log-log coordinate mapping. The watermarks, which used the Fourier-Mellin transform, are designed to robustness mainly against the rotation, scales and translations.

Template based schemes. In this case, the watermark consists of two parts: template and the self watermark. The template contains no information but is merely a tool used to recover possible transformations in the image. The recovery of the watermark is a two stage process. First, the transformation undergone by the image is determined, and then inversion or compensation for the transformation when decoding the watermark is done. The points of the template may be distributed for example in the DFT domain.

Autocorrelation techniques. The third method for the recovery of geometrical transformations is the use of the auto-correlation function. These methods are based on the adding of the repeated watermarks in the overlapping fashion. At the detection, the estimation of the watermark is performed and the autocorrelation function is calculated. The peaks in the auto-correlation function are obtained due to the repetitive insertion of the watermark. Since the auto-correlation

of the inserted watermark is known, this is compared with the auto-correlation function of the recovered watermark. A transformation matrix is calculated based on the two sets of peaks. This transformation is then inverted and the watermark is decoded.

3. LOG-POLAR MAPPING

As was mentioned before, the exploitation of the features of some transformations which are invariant against the affine transformations can be used to improve robustness of watermarks against the geometrical attacks. Discrete Fourier transform (DFT) fulfill these requests and is often being used in the digital watermarking algorithms.

If the picture is defined as two dimensional function $x(i,j)$ in the Cartesian coordinate system with limitations $0 \le i < N_1$ and $0 \le j < N_2$, the DFT and inverse DFT is defined as follows

$$F(u,v) = \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} x(i,j) \cdot e^{-j(2\pi/N_1)ui} \cdot e^{-j(2\pi/N_2)vj}, \quad (1)$$

$$x(i,j) = \frac{1}{N_1 N_2} \sum_{p=0}^{N_1-1} \sum_{q=0}^{N_2-1} F(u,v) \cdot e^{j(2\pi/N_1)ui} \cdot e^{j(2\pi/N_2)vj}. \quad (2)$$

Affine transforms performed with images in the spatial domain caused the specific changes in the DFT domain.

The picture shift in the spatial domain cause a linear shift in the phase component of the DFT

$$F(k_1,k_2) e^{-j(\gamma_x k_1 + \gamma_y k_2)} \leftrightarrow f(x+\gamma_x, y+\gamma_y). \quad (3)$$

The symbol $\leftrightarrow$ represents the transform relationship between DFT domain and the image spatial domain.

Scaling the axes in the spatial domain causes an inverse scaling in the frequency domain

$$\frac{1}{\rho} F\left(\frac{k_1}{\rho}, \frac{k_2}{\rho}\right) \leftrightarrow f(\rho x, \rho y). \quad (4)$$

The image rotation through an angle θ in the spatial domain causes the DFT representation to be rotated through the same angle

$$F(k_1 \cos\theta - k_2 \sin\theta, k_1 \sin\theta + k_2 \cos\theta) \leftrightarrow \\ \leftrightarrow f(x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta) \quad (5)$$

From equation (3) is clear that spatial shifts affect only the phase representation of the image. Also the equations (4) and (5) can be rewritten by using the specific substitution. The changes in the image caused by scaling and image rotation in the spatial domain can be described by invariant shift after this substitution. This can be performed by the substitution which is called ***log-polar mapping (LPM)***.

Consider a point $(x,y) \in R^2$ and defines $x = e^{\mu}\cos\theta$, $y = e^{\mu}\sin\theta$, where $\mu \in R$ and $0 \le \theta < 2\pi$. The result of this substitution is that for every point (x,y) there is a point (μ,θ) that uniquely corresponds to it.

The new coordinate system (μ,θ) converts the scaling and rotation into the simple translation in the direction of the axis.

Scaling is converted to a translation

$$(\rho x, \rho y) \leftrightarrow (\mu + \log\rho, \theta). \quad (6)$$

Rotation is converted also to a translation

$$(x\cos(\theta+\delta) - y\sin(\theta+\delta), x\sin(\theta+\delta) + y\cos(\theta+\delta)) \leftrightarrow \\ \leftrightarrow (\mu, \theta+\delta) \quad (7)$$

4. PROPOSED WATERMARKING METHOD

Proposed algorithm is based on the combination of the DFT and LPM features. Inserted watermark is in the form of spare matrix which is created depending on the secret key K. On the receiver side, the watermark detection process is performed.

4.1 Watermark embedding

The entries of the watermark embedding process are the original gray scale image *I* and the key *K*, which is in the form of number and is used as the initialization vector for the pseudorandom generator.

The process of the watermark embedding is shown in the *Fig. 1* and can be described in five steps:

- DFT of the original image *I*,
- watermark generation based on the secret key *K*,
- transformation of the key by using inverse LPM.
- watermark embedding into the chosen coefficients of the magnitude spectrum of the DFT.
- inverse DFT.

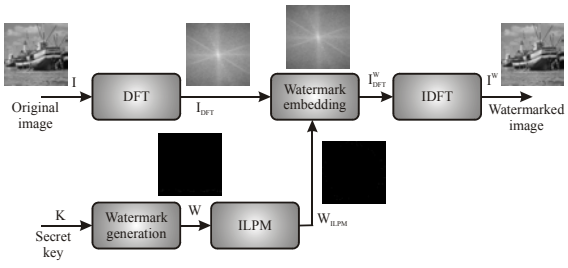


Fig. 1 Watermark embedding process

The calculation of the DFT of the original image is the first step in the watermark embedding process. Pseudorandom sequence is generated based on the secret key K. This secret key is used as the initialization vector for the pseudorandom generator. Generated pseudorandom sequence has normal dispersion and zero mean value and based on the chosen decision level is mapped on the two values (0,1). The size of the sequence is selected based on the desired quality of the watermarked image and on the desired robustness of the embedded watermark.

The size of the watermark has to be the same as the size of the original image. Generated and mapped sequence is situated in the lower part of the watermark. The quality of the watermarked image is also influenced by the location of the sequence in the watermark.

The next step in the embedding process is the transformation of the watermark by using inverse log-polar mapping (ILPM). The ILPM transforms the sequence in the watermark into the concentric ring. In the DFT spectrum the medium frequencies are situated in this region. These frequencies are chosen for the watermark embedding for two reasons: modification of medium frequencies causes less degradation of the watermarked image as the modification of the lower frequencies would caused and if the higher frequencies are modified during the watermark embedding process, the watermark robustness would be low against the attack, mainly against lossy compressions.

The watermark is embedded into the magnitude coefficients of the DFT in the form of local peaks. The process of the embedding is adaptive, that means the watermark is not embedded into the whole picture with the same power. The process of the watermark embedding can be described as

$$I_{DFT}^{W}(i,j)=\begin{cases}\dfrac{\alpha}{9}\displaystyle\sum_{i-1}^{i+1}\sum_{i-1}^{i+1}I_{DFT}(i,j) & if \quad W_{ILPM}(i,j)=1\\ I_{DFT}(i,j) & if \quad W_{ILPM}(i,j)=0\end{cases} \tag{8}$$

where α is the power of the embedded watermark.

4.2 Watermark detection

The detection algorithm does not require the original image and is based on the correlation test between the original and extracted watermark. The algorithm entries are tested image and secret key for the watermark generation. The process of the watermark detection is shown in the *Fig. 2* and can be described in five steps:

- DFT transformation of the tested image,
- position finding of the local maxims,
- LPM transformation of the local maxims,
- watermark generation by using secret key *K*,
- correlation test,
- decision about the presence or absence of the watermark in the tested image.

The DFT is the firs step in the watermark extraction process. The local maxims are searching in the small windows. In the experiments the window with 10x14 pixels was used. The positions of all founded local maxims are saved into the empty matrix. This matrix represents the extracted watermark.

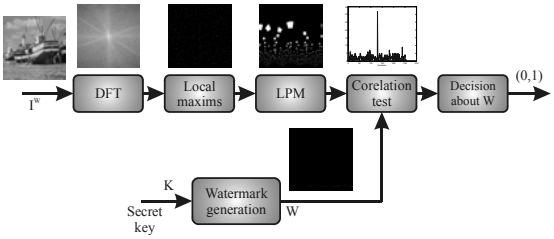
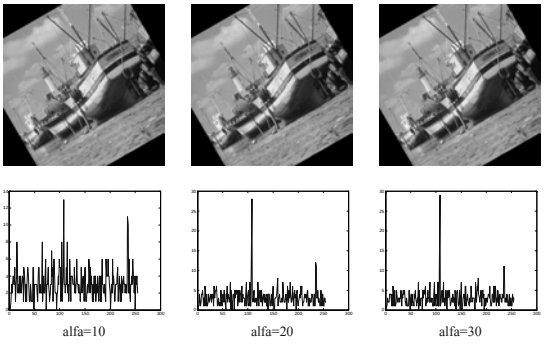


Fig. 2 Watermark detection

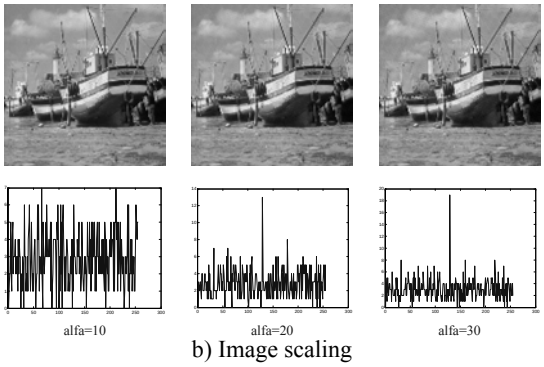
In the next step, the local maxims are transformed by using LPM and the correlation test between the original and tested watermark is calculated. The decision about the watermark presence or absence is based on the chosen level of the correlation function.

5. EXPERIMENTAL RESULTS

Three different powers of the α (alfa) parameter during the watermark embedding process were used. The robustness of the embedded watermarks was tested against some attacks.



a) Rotated images (30 degrees)



b) Image scaling

Fig. 3 Attacked images

Rotated images are shown in the *Fig. 3a*. As can be seen, in the second and third image there is one local peak in the correlation function and the watermark was detected successfully.

The attack by image scaling is shown in the *Fig. 3b*. The image was reduced in dimensions to the half and thereafter enlarged to the original dimensions. As can be seen, the watermark was detected in two cases. The watermark was destroyed in the first case and the watermark detection was unsuccessful.

CONCLUSION

In this paper was shown one approach how to improve robustness of the digital watermarks in gray scale images based on the DFT and LPM. The further work will be oriented on the improving of the robustness against the removal attacks, mainly the lossy compressions, and on the embedding watermarks into the color images based on presented approach.

ACKNOWLEDGEMENTS

The work presented in this paper was supported by Grant of Ministry of Education and Academy of Science of Slovak republic VEGA under Grant No. 1/4054/07.

References

[1] DEGUILLAUME, F.; VOLOSHYNOVSKIY, S.; PUN, T.: A method for the estimation and recovering from general affine transforms in digital watermarking applications, In *Proc. SPIE Vol. 4675, Security and Watermarking of Multimedia Contents IV*, pp. 313–322, 04/2002.

[2] Digital Rights: Management, protection, standardization. IEEE Signal Processing Magazine, Vol. 21, No. 2, March 2004.

[3] RIDZOŇ, R.; LEVICKÝ, D.: Log-polar Mapping in Robust Digital Image Watermarking. In: Radioelektronika 2007: Proceedings of 17th international conference, April 24-25, 2007, Department of Radio Electronics, Brno University of Technology, 2007. p. 525-528. ISBN 978-80-214-3390-8.

[4] RUANAIDH, J.J.K., PUN, T.: Rotation, scale and translation invariant digital image watermarking, in *Proc. IEEE Int. Conf. Image Processing 1997 (ICIP 97), Santa Barbara, CA*, vol. 1, pp. 536–539, Oct. 1997.

[5] VOLOSHYNOVSKIY, S. et al.: Attack Modeling: Towards a Second Generation Watermarking Benchmark, *Sig. Processing, Special Issue on Information Theoretic Issues in Digital Watermarking*, 2001, vol. 81. no. 6, pp. 1177-1214.

[6] ZHENG, D.; ZHAO, J.; El SADDIK, A.: RST Invariant Digital Image Watermarking Based on Log-Polar Mapping and Phase Correlation in *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Authentication, Copyright Protection and Information Hiding*, Vol. 13, Issue 8, pp. 753-765, August 2003.

Ing. Radovan RIDZOŇ
prof. Ing. Dušan LEVICKÝ, CSc.
Department of Electronics and Multimedia Communications
Faculty of Electronics and Informatics
Technical University of Košice
Park Komenského 13, 041 20 Košice
Slovak republic
E-mail: radovan.ridzon@tuke.sk
dusan.levicky@tuke.sk

APLIKÁCIA MORFOLOGICKÝCH FILTROV V SUBPÁSMOVOM KÓDOVANÍ

THE APPLICATION OF MORPHOLOGICAL FILTERS IN SUBBAND CODING

Jozef ŠTULRAJTER, Milan LEHOTSKÝ, Marcel HARAKAĽ

Abstract: Subband coding system using morphological filters in analyzing and synthesizing filter banks is described in this paper a standard method of morphological decomposition is realized by decomposition of the input image to the subimages created by objects of certain size. Perfect reconstruction is reached by adding all subimages. The purpose is to design system BAF/BSF using the bank of morphological systems with perfect reconstruction of the image.

Keywords: morphological filters, morphological transformation, subband coding, filter banks, morphological operations.

1. ÚVOD - ZÁKLADNÉ MORFOLOGICKÉ TRANSFORMÁCIE

Morfologické filtre sú nelineárne filtre, ktoré modifikujú geometrický tvar objektov s využitím matematickej morfológie. Využívajú sa pri zvýraznení objektov, ktoré sú ukryté v pozadí, pri detekcii hrán, rozpoznávaní objektov, pri odstraňovaní impulzového šumu z obrazu, pri analýze textúr a podobne. Vstupný obraz je filtrovaný pomocou morfologického elementu a tento proces je podobný konvolučnému súčinu obrazu a morfologického elementu. Morfologické filtre sú tvorené kombináciou štyroch základných operácii matematickej morfológie, ktoré sú ***dilatácia, erózia, otvorenie a uzavretie***. Morfologická filtrácia je proces, pri ktorom sa na obraz aplikujú základné morfologické operácie v ľubovoľnom poradí.

Nech je funkcia x(n) binárneho obrazu, B je základná morfologická štruktúra a nech x(n) a B sú podmnožinami 2-R Euklidovského priestoru. Potom x(n-**b**) definuje priestorové posunutie vstupného obrazu podľa vektora **b**. Prvá základná morfologická operácia ***dilatácia*** je definovaná vzťahom

$$x(n) \oplus B = \bigcup_{\mathbf{b} \in B} x(n - \mathbf{b}). \qquad (1)$$

To znamená, že dilatácia x(n) podľa B je uskutočnená zjednotením posunutí obrazu x(n) podľa všetkých vektorov $\mathbf{b} \in B$. Doplnkovou operáciou k dilatácii je ***erózia*** a tú možno vyjadriť vzťahom

$$x(n) \ominus B = \bigcap_{\mathbf{b} \in B} x(n + \mathbf{b}). \qquad (2)$$

Výsledkom erózie je prienik všetkých posunutí vstupného obrazu x(n) podľa vektora –**b**, kde $\mathbf{b} \in B$. Vplyv dilatácie a erózie na tvar objektu zo vstupného obrazu je graficky znázornený na obr. 1. Z obrázku je zrejmé, že dilatácia spôsobuje zväčšovanie objektov a naopak erózia spôsobuje zmenšovanie objektov. Napriek tomu, že dilatácia a erózia sú doplnkové operácie, tieto operácie sú nevratné. To znamená, že ak vstupný obraz je filtrovaný pomocou erózie, tak tento obraz nemusí byť obnovený dilatáciou a naopak.

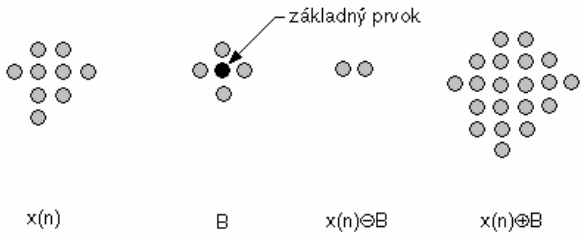
Ďalšou morfologickou operáciou je ***otvorenie. Otvorenie*** je definované ako aplikácia erózie a následne dilatácie na obraze x(n) vyjadrená vzťahom

$$x(n) \circ B = [x(n) \ominus B] \oplus B. \qquad (3)$$

Uzavretie je definované ako proces aplikácie ***dilatácie*** a následnej ***erózie*** na obraz x(n), teda platí

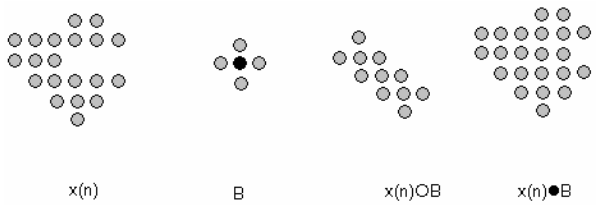
$$x(n) \bullet B = [x(n) \oplus B] \ominus B. \qquad (4)$$

Vplyv ***otvorenia*** a ***uzavretia*** na tvar objektu zo vstupného obrazu je graficky znázornený na obrázku 2. Z obrázku je zrejmé, že ***otvorenie*** môže byť realizované ako posuv oblasti B po obraze tak, aby každý prvok B ležal vo vnútri objektu. Výsledným objektom je zjednotenie prienikov všetkých pozícií oblasti B a objektu zo vstupného obrazu x(n). ***Otvorenie*** odstraňuje detaily z obrazu reprezentovaného vo forme lalokov vystupujúcich z veľkých objektov. Podobne proces **uzavretia** môže byť realizovaný tak, že centrálny (základný) prvok oblasti B sa pohybuje po obrysoch objektu. Výsledkom je zjednotenie objektu z obrazu x(n) a všetkých pozícií oblasti B, čím sa odstránia z obrazu detaily reprezentované lalokmi vstupujúcimi do veľkých objektov. Teda realizáciou ***otvorenia*** sa veľkosť objektov zmenšuje



Obr. 1 Vplyv erózie a dilatácie na tvar objektu z obrazu x(n)

a realizáciou ***uzavretia*** sa veľkosť objektov zväčšuje oproti pôvodnej veľkosti objektov.



Obr. 2 Vplyv otvorenia a uzavretia na tvar objektu zo vstupného obrazu x(n)

Proces morfologickej filtrácie môže byť zovšeobecnený pre obrazy s viac úrovňovými hodnotami obrazového prvku (*op)*. Základné elementy týchto morfologických operácií sú 3R útvary ako napr. gule a valce. Obrazy môžu byť reprezentované taktiež v 3R priestore, kde výška (tretia súradnica) zodpovedá hodnote *op*. Morfologická filtrácia je vlastne posúvanie základného elementu B po povrchu 3R zobrazenia obrazu x(n). Potom dilatácia je určená vzťahom

$$x(n) \oplus B = \max_{\mathbf{b} \in B}[x(n-\mathbf{b}) + B(n)] \cdot \qquad (5)$$

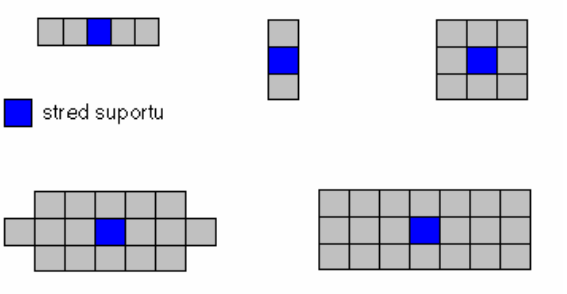
Pri binárnych obrazoch bolo použité zjednotenie a pri viac úrovňových obrazoch sa určuje maximum posunutí obrazu x(n), kde B(**b**) je váhovacia funkcia. Podobne pre ***eróziu*** platí

$$x(n) \Theta B \min_{\mathbf{b} \in B}[x(n+\mathbf{b}) + B(\mathbf{b})], \qquad (6)$$

teda proces ***erózie*** je realizovaný nájdením minimálnej hodnoty posunutí obrazu x(n). ***Otvorenie a uzavretie*** je definované zhodne ako u binárnych obrazov vzťahmi (3) a (4).

2. TYPY MORFOLOGICKÝCH FILTROV

Morfologické filtre sú nelineárne filtre [5, 7]. K nelineárnym filtrom patrí aj mediánový filter. Mediánový filter je realizovaný tak, že z blízkeho okolia vyšetrovaného obrazového prvku (*op)* sa vyberú *op* a zoradia do postupnosti podľa veľkosti hodnôt *op*. Hodnota *op* zo stredu postupnosti je výsledná hodnota mediánovej filtrácie a túto hodnotu nadobúda *op* vo výstupnom obraze, ktorý má tie isté priestorové súradnice ako vyšetrovaný *op* vo vstupnom obraze. Okolie vyšetrovaného bodu nadobúda rôzne tvary a veľkosti (pozri obr. 3). Z obrázku je zrejmé, že je možné realizovať jednorozmernú (1R), dvojrozmernú (2R) a vo všeobecnosti v-rozmernú mediánovú filtráciu. Suport mediánovej filtrácie je oblasť, ktorá definuje tvar a veľkosť okolia vyšetrovaného *op*. Snahou je, aby suport prekrýval nepárny počet *op*, lebo tak je možné jednoznačne určiť stred postupnosti hodnôt *op*.



Obr. 3 Príklady suportov pre mediánovú filtráciu

Mediánova filtrácia je proces, pri ktorom stred suportu postupne prechádza po všetkých *op* vstupného obrazu. Postupne na tie isté priestorové súradnice, ako má stred suportu na vstupnom obraze, sa uloží výsledok mediánovej filtrácie do výstupného obrazu. Mediánové filtre sú vysoko účinné pri odstraňovaní impulzového šumu z obrazu a využívajú sa aj ako dolnopriepustné filtre (DP) s možnosťou použitia aj pri realizácii systému analýzy a syntézy obrazu.

Morfologický filter realizovaný dvojicou operácií *otvorenie-uzavretia* (Open–Closing OC) alebo *uzavretie-otvorenia* (Clos-Opening CO) pri odstraňovaní impulzového šumu z obrazu sa správa podobne ako mediánový filter. Morfologický filter realizovaný dvojicou operácií má výhodu v tom, že môže rozlišovať a teda oddelene odfiltrovať kladné alebo záporné zložky impulzového šumu, čo mediánový filter nedokáže [5].

Algoritmy pre detekciu textúr

Pri subpásmovom kódovaní obrazu použitím banky morfologických filtrov (BMF) je výhodné detekovať a odlišným spôsobom kódovať textúry, ako ostatné časti obrazu. Správna detekcia textúr umožňuje efektívny rozklad vstupného obrazu bez prídavnej informácie. Nesprávna detekcia textúr môže spôsobiť zákmity, čo znehodnotí navrhovaný systém analýzy a syntézy obrazu s použitím BMF.

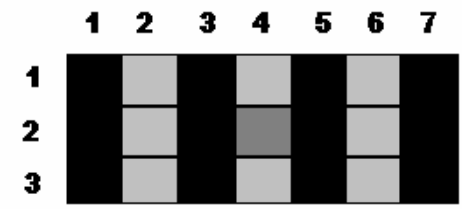
Obrazy sa skladajú z homogénnych oblastí, hrán a textúr. Analýzou vlastností textúr je možné dospieť k algoritmu detekcie textúr. Homogénne oblasti sú zvyčajne veľké priestorové oblasti, ktorých hodnoty *op* nadobúdajú približne rovnaké hodnoty. Naopak, hrany sú malé priestorové oblasti, ktorých hodnoty *op* nadobúdajú veľmi rozdielne hodnoty *op* v smere kolmom na hranu a malé rozdiely v smere pozdĺž hrany. Textúry sú oblasti Oi, ktoré majú rozdielne hodnoty *op* vo všetkých smeroch. Teda pri detekcii textúr budú sledované dva parametre. Prvým je disperzia hodnôt σ_i^2 v oblasti Oi. Disperzia σ_i^2 charakterizuje mieru rozdielnosti hodnôt op v sledovanej oblasti Oi. Pre disperziu platí

$$\sigma_i^2 = \frac{1}{k}\sum_{k=0}^{k-1}[x(k)-\bar{x}]^2 , \qquad (7)$$

kde k je počet *op* nachádzajúcich sa v sledovanej oblasti Oi a $\bar{x}$ je stredná hodnota, pre ktorú platí

$$\bar{x} = \frac{1}{k}\sum_{k=0}^{k-1}x(k). \qquad (8)$$

Simuláciou sa dospelo k záveru, že najvhodnejšia veľkosť sledovanej oblasti Oi je pre k=12. Tvar tejto oblasti je na obr. 4.



Obr. 4 Odporúčaný tvar sledovanej oblasti Oi
Čierne *op* zodpovedajú aktívnym prvkom oblasti.

Ak Oi je homogénna, tak disperzia σ_i^2 tejto oblasti má byť menšia ako hraničná disperzia σ_h^2. Ak $\sigma_i^2 > \sigma_h^2$ tak centrálny bod z Qi patrí hrane alebo textúre.

Druhým parametrom p sa určí, či stredový *op* Oi patrí hrane alebo textúre. K tomu účelu sa počítajú disperzie v štyroch smeroch. Pre oblasť Oi na obrázku 4 môžu byť takéto smery

J=1: (1,1) (2,3) (2,5) (3,7),
J=2: (3,1) (2,3) (2,5) (1,7),
J=3: (2,1) (2,3) (2,5) (2,7),
J=4: (1,5) (2,5) (3,5).

Tieto štyri smery sú vybrané tak, aby reprezentovali všetky možné smery 0, π/4, π/2, 3π/2. Všetky disperzie $\sigma_{i,j}^2$ sú vypočítané podľa (7) . Parameter p je definovaný vzťahom

$$p = \max_i(\sigma_{ij}^2)/\min_i(\sigma_{ij}^2), \qquad (9)$$

kde j definuje smer podľa pre oblasť Oi.

Ako už bolo naznačené, textúry majú mať približne rovnaké disperzie vo všetkých smeroch. Teda ak parameter *p* má vysokú hodnotu, tak stredový *op* z Oi je *op* na hrane. Preto oblasť Oi je považovaná za hranu, ak $p>p_h$, kde p_h je hraničná hodnota parametra p. Z predchádzajúceho je zrejmé, že detekcia textúr môže byť realizovaná nasledujúcim algoritmom:

Algoritmus 1

1. Výpočet σ_i^2 podľa (7),
2. Výpočet p podľa (9),
3. Ak ($\sigma_i^2 > \sigma_h^2$) a súčasne ($p<p_h$), tak stredový *op* z Oi patrí do textúry, *koniec*.
4. Alebo ak ($\sigma_i^2 > \sigma_h^2$) a súčasne ($p>p_h$), tak stredový *op* z Oi patrí hrane, *koniec*.
5. Alebo stredový *op* z Oi patrí homogénnej oblasti.

Pomocou tohto algoritmu je možné detekovať väčšinu textúr. V textúrach sa však objavili izolované *op* a v blízkosti hrán tenké čiary, ktoré spôsobujú zákmity. Preto je potrebné vykonať na obrazoch proces, ktorý by odstránil rušivé *op* a čiary, ktoré vznikli detekciou textúr.

Na odstránenie týchto rušivých *op* a čiar spôsobených impulzovým šumom je najvýhodnejšie použiť mediánový filter s veľkosťou napr. 3x3. Filtrácia prebieha tak, že stred mediánového filtra sa posúva po obraze, pričom stredový prvok filtra prechádza po prvkoch textúry a ani jeden prvok filtra nepatrí do homogénnej oblasti. Tento proces odstráni osamelé *op* a tenké čiary v textúrach. Po detekcii textúr algoritmom 1 sa uskutoční dodatočná úprava obrazu realizáciou algoritmu 2 pre všetky *op* obrazu.

Algoritmus 2

Ak *op* patrí hrane a osem susedných *op* patrí textúre alebo hrane, tak *op* patrí textúre, alebo na *op* sa aplikuje mediánová filtrácia o veľkosti filtra 3x3, pričom stred mediánového filtra leží na klasifikovanom *op*.

Realizáciou algoritmu 2 sa odstráni väčšina rušivých *op* a čiar, čo zabezpečí efektívnu reprezentáciu bez zákmitov obrazu.

3. APLIKÁCIA MORFOLOGICKÝCH FILTROV V BANKÁCH FILTROV PRE SBC

V tejto časti je naznačený návrh maximálne decimovanej banky morfologických filtrov (BMF) s vlastnosťou dokonalej rekonštrukcie obrazu. Lineárne banky filtrov (BF) môžu vytvárať maximálne decimované banky filtrov. Je problematické popísať morfologické filtre takým spôsobom, aby sa získali nutné a postačujúce podmienky dokonalej rekonštrukcie obrazu pri použití BMF. Morfologické dolnopriepustné filtre realizujúce operáciu ***otvorenia a uzavretia*** neumožňujú realizovať dokonalú rekonštrukciu obrazu. Tieto filtre realizujú filtráciu takým spôsobom, že obraz síce nebude obsahovať malé objekty, ale obrysy veľkých objektov sú ostré, čo svedčí o prítomnosti vysokých frekvencií. To

znamená, že vysoké frekvencie sa vždy nachádzajú v subobraze získanom takouto filtráciou. Preto na strane syntézy morfologické filtre nie sú schopné potláčať prekrývanie spektier subobrazov, čo spôsobí nedokonalú rekonštrukciu obrazu.

Preto je potrebné získať nástroj, pomocou ktorého by boli popísané nelineárne filtre. Ďalej je potrebné navrhnúť dvojkanálovú BMF, ktorej filtre prepúšťajú polovicu frekvenčného pásma. Nech 1R lineárny filter s prenosovou funkciou G(z) je dolnopriepustný filter (DP) prepúšťajúci polovicu frekvenčného pásma, ak každý nepárny prvok jeho impulzovej odozvy nadobúda nulovú hodnotu, preto platí

$$g(m)=\begin{cases} 1/2, & pre\ \ m=2i \\ 0, & pre\ m=2i+1 \end{cases} \quad i=0,1,2,\ldots\ldots M/2-1, \tag{10}$$

kde M je dĺžka impulzovej odozvy filtra.

Nech 1R signál x(n) je decimovaný a následne interpolovaný s rádom N=2. Potom nech y(n) je filtrovaný použitím G(z). Nech z(n) je výsledok konvolučného súčinu y(n) a g(n), teda platí

$$z(n) = y(n) * g(n). \tag{11}$$

Potom pre každú párnu vzorku z(n) platí

$$z(2n) = 1/2x(2n). \tag{12}$$

Tento vzťah umožňuje definovať nelineárny filter prepúšťajúci polovicu frekvenčného pásma. Nech vstupný signál x(n), ktorý je decimovaný a interpolovaný s rádom N=2 a výsledkom je signál y(n). Nelineárny filter prepúšťa polovicu frekvenčného pásma, ak každá párna vzorka signálu y(n) sa rovná vstupnému signálu x(n). Nech z(n) je výstup z nelineárneho filtra, potom

$$z(2n) = cx(2n), \tag{13}$$

kde $c \in R$ je ľubovoľná konštanta.

Takéto nelineárne filtre existujú a sú realizované morfologickými procesmi ***erózie, dilatácie*** alebo **mediánovej** filtrácie, ak je správne vybraný suport filtrácie.

Dokonalú rekonštrukciu pre dvojkanálový systém analýzy (BAF) a syntézy (BSF) obrazu s lineárnou bankou filtrov (obrázok 5a) možno dosiahnuť, ak $G_0(z)$ a $H_0(z)$ sú prenosové funkcie DP filtrov prepúšťajúce polovicu frekvenčného pásma a $G_1(z)$ a $H_1(z)$ sú prenosové funkcie hornopriepustných (HP) filtrov. Nech medzi filtrami platí vzťah ako pri dokonale rekonštrukčných filtrov, teda platí

$$H_0(z) = G_1(-z),$$
$$H_1(z) = -G_0(-z). \tag{14}$$

Nech $G_0(z)=1$, potom $H_1(z)= -1$. Takáto banka filtrov je na obrázku 5b. Nech $H_0(z)$ je nahradený nelineárnym filtrom s prenosovou funkciou $M_0(n)$. Keďže sa jedná o nelineárny filter, nedá sa popísať v z - oblasti. Pri lineárnych filtroch prepočet filtra z DP na HP a naopak je realizovaný negáciou premennej, čo možno vyjadriť vzťahom

$$Z^{-1}\{G(-z)\} = (-1)^m g(m). \tag{15}$$

Pre lineárny filter s minimálnou fázou, ktorý prepúšťa polovicu frekvenčného pásma platí

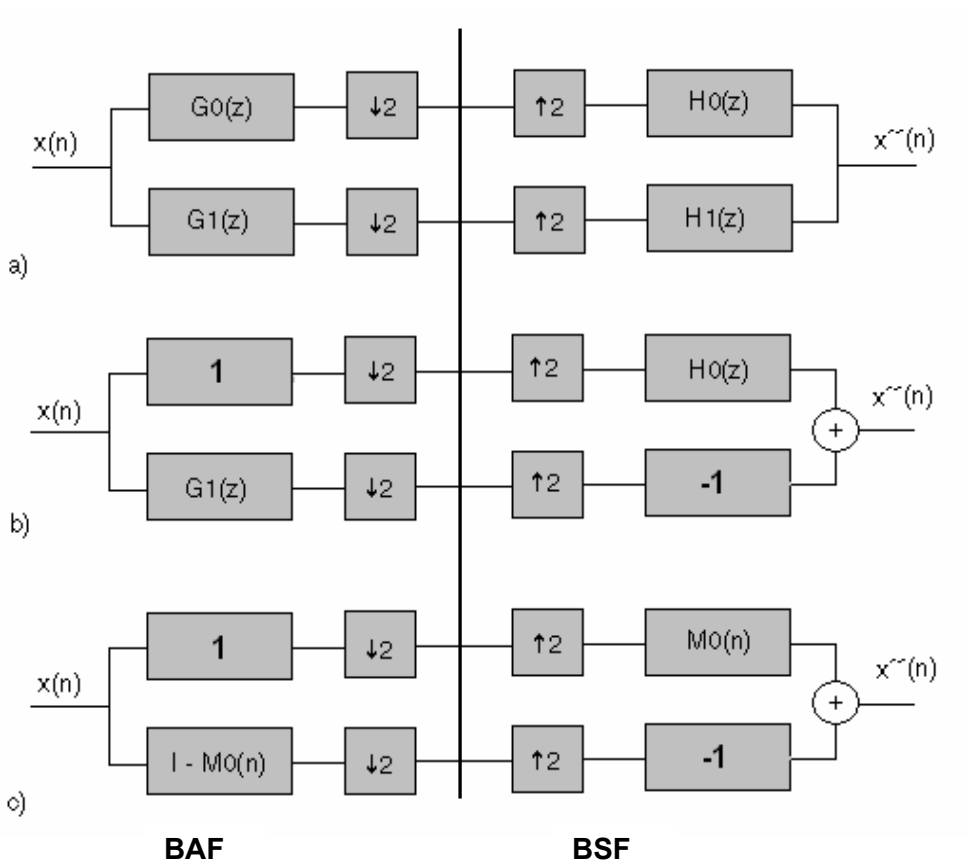
$$G(-z) = 1 - G(z). \tag{16}$$

Táto rovnica môže byť zovšeobecnená pre nelineárny filter. Nech $M_1(n)$ je prenosová funkcia nelineárneho HP filtra, potom platí [3]

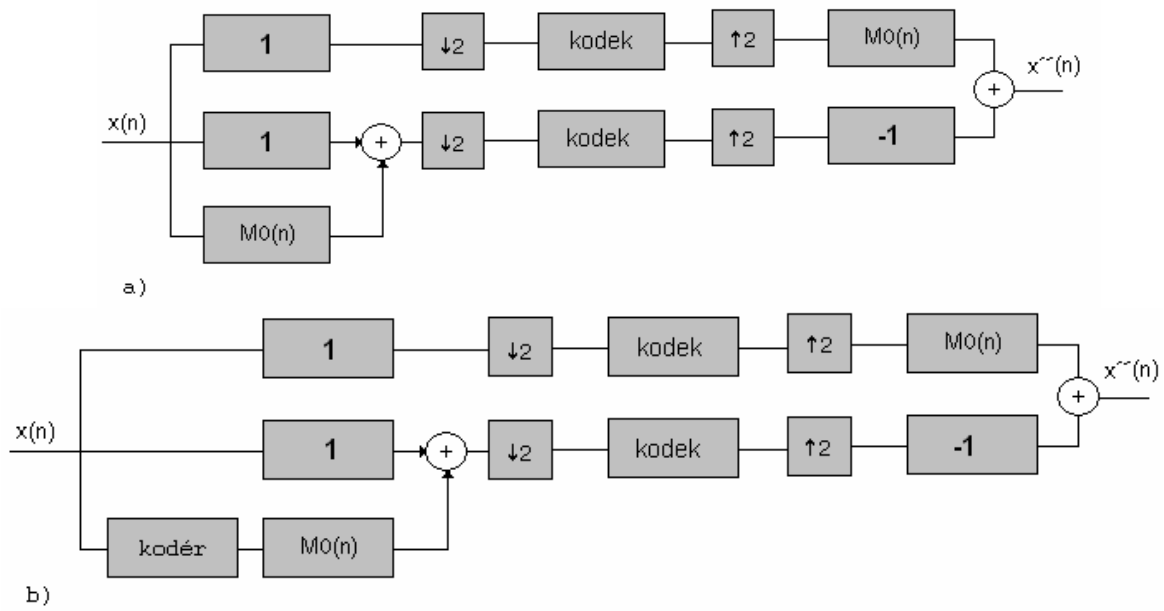
$$M_1(n) = I - M_0(n), \tag{17}$$

kde I je operátor identity. Z toho odvodená banka filtrov na obrázku 5c umožňuje dokonalú rekonštrukciu.

Pri návrhu systému analýzy a syntézy je dôležitou úlohou správny návrh filtrov analýzy a syntézy. Návrh morfologických filtrov je odlišný od návrhu lineárnych filtrov. Pri návrhu morfologických filtrov je vhodné mať vedomosti z oblasti matematickej morfológie. Návrh filtrov môže byť realizovaný aj experimentálne, lebo existuje len obmedzený počet riešení. Väčšina morfologických filtrov je určená suportom a spôsobom spracovania *op* prekrytých suportom. Často je potrebné len vybrať najlepší možný filter porovnávaním ich vplyvu na obrazy. Z výsledkov simulácie s použitím reálnych obrazov je zrejmé, že najvhodnejšie morfologické filtre pre subpásmové kódovanie použitím BMF sú mediánové filtre, lebo umožňujú najúčinnejšie kódovanie pri reprezentácii subobrazov. Teda ostáva určiť veľkosť a tvar suportu mediánového filtra. Nech kódovanie je realizované až po filtrácii na strane analýzy, ako je to naznačené na obrázku 6a. Kódovanie môže zmeniť charakter obrazu, čo na strane syntézy by sa mohlo prejaviť výberom nesprávneho typu filtra. Na obrázku 6b je naznačený systém subpásmového kódovania, kde kódovanie je realizované na strane analýzy ešte pred filtráciou. Dvojnásobné kódovanie HP vetvy nemá vplyv na kvalitu výstupného obrazu. V HP vetve analýzy sa počíta interpolačná chyba medzi vstupným a výstupným obrazom a kvantovaným a filtrovaným obrazom. Táto interpolačná chyba je menšia ako polovica kvantizačného kroku použitého kodéra. Systém na obrázku 6b vykazuje vyššiu účinnosť kódovania ako na obrázku 6a.



Obr. 5 Systémy analýzy a syntézy obrazu znázorňujúce nahradenie maximálne decimovanej banky lineárnych filtrov bankou morfologických filtrov



Obr. 6 Systém subpásmového kódovania obrazu s použitím banky morfologických filtrov, a) kódovanie po filtrácii, b) kódovanie pred filtráciou na strane analýzy.

ZÁVER

Pri simulácii subpásmového kódovania obrazu boli u lineárnej BF použité filtre s veľkou šírku prechodového pásma, čo mierne znižuje zákmity vo výstupnom obraze. Pri porovnaní oboch bánk filtrov je možné použiť objektívne kritéria kvality, ako pomer signál/šum a pod., ale tieto kritéria nie sú vhodné pre porovnanie dvoch úplne odlišných techník systému analýzy a syntézy, pretože skreslenie má iný charakter. Preto vizuálny vnem sa stal subjektívnym kritériom kvality výstupného obrazu.

Pri použití lineárnej BF neboli zákmity s vysokou energiou, ale boli viditeľné a pôsobili rušivo. Nižšia energia zákmitov súvisí s výberom lineárnych filtrov s veľkou šírkou prechodového pásma. Pri použití BMF sa neobjavili zákmity. Textúry boli lepšie spracované lineárnou BF ako BMF. Z hľadiska vizuálneho vnemu BMF sú výhodnejšie pri obrazoch bez textúr a lineárne BF sú výhodnejšie pri spracovaní obrazov s textúrami. Z predchádzajúceho je zrejmé, že najvýhodnejší by bol taký systém, ktorý by prebral všetky výhodné vlastnosti lineárnych a morfologických bánk filtrov. Mohlo by sa prepínať medzi dvoma bankami filtrov, čo si vyžaduje prítomnosť oboch bánk filtrov. Zároveň by bolo potrebné prenášať do systému syntézy prídavnú informáciu o tom, aký typ banky bol použitý pre konkrétny *op* na strane analýzy. Preto je potrebné navrhnúť kritérium, podľa ktorého by sa vybral a použil totožný typ BF pre každý *op* v systéme analýzy a syntézy.

ACKNOWLEDGEMENT

This work was supported by DoD project ŠPP 118_06-RO02_RU21-240 "NATO Network Enabled Capability and implementation in the Armed Forces of the Slovak Republic".

Zoznam bibliografických odkazov

[1] http://www.otolith.com/otolith/olt/sbc.html, 2007.
[2] ŽÁRA, J., BENEŠ, B., FELKEL, P.: Reprezentace obrazu, Spracování obrazu, Filtrace obrazu.: In: Moderní počítačová grafika. Computer Press 1998, 136s.
[3] DZIVÝ, J.: Metódy analýzy a syntézy pre subpásmové kódovanie obrazu. Rigorózna práca – Technická Univerzita, Fakulta elektrotechniky a informatiky, Košice 1996, 63s.
[4] LEON, G. A., WIDJAJA ,I.: Communication Networks. ISBN: 007246352X, cGraw-Hill Companies, The July 2003, 928 pp.
[5] LEHOTSKÝ, M., ŠTULRAJTER, J., REPČÍK, D.: Použitie FIR filtrov a morfologických transformácií v subpás-movom kódovaní obrazov. In.: Zborník z medzinárodnej vedeckej konferencie "Komunikačné a informačné technologie. 2001", 26. - 27. 9. 2001 Tatranské Zruby. Liptovský Mikuláš: Vojenská akadémia, 2001. – s. 276 – 281. ISBN 80 - 8040 - 159 - 4.
[5] LEVICKÝ, D.: Multimediálne telekomunikácie – multimédia, technológie a vodoznaky. ELFA s.r.o., Košice, 2002.
[6] HARAKAĽ, M., LEHOTSKÝ, M., CHMÚRNY, J.: Optimalizácia morfologických transformácií na báze kvadrantového stromu. In: Zborník vedeckej konferencie Nové smery v spracovaní signálov IV, 27.-29.5. 1998, Liptovský Mikuláš.- Liptovský Mikuláš: Vojenská akadémia, 1998, - S. 129-133. ISBN 80-8040-071-7

Summary: Obviously either BAF/BSF linear FIR filters or non-linear IIR filters are used in SBC. The properties of the former ones in image processing are described in the conclusion of the paper. The computational complexity of the latter ones is substantially lower, however, it is necessary to take into account their stability. An interesting view of the problem is to use BMF in SBC (with its properties) and also a method of substituting perfect reconstructing bank of FIR filters using BMF.

prof. Ing. Jozef ŠTULRAJTER, CSc.[1)]
doc. RNDr. Milan LEHOTSKÝ, CSc.[2)]
doc. Ing. Marcel HARAKAĽ, PhD.[1)]
[1)] Katedra informatiky
Akadémia ozbrojených síl generála M. R. Štefánka
Demänová 393, 031 01 Liptovský Mikuláš
Slovenská republika
E-mail: stulrajter@aoslm.sk
harakal@aoslm.sk
[2)] Katedra informatiky
Pedagogická fakulta
Katolícka univerzita
nám. A. Hlinku 56/1, 034 01 Ružomberok
Slovenská republika
E-mail: Milan.lehotsky@fedu.ku.sk

NETWORK ACCESS CONTROL TECHNOLOGIES FOR SECURING INTERNAL NETWORKS

Július BARÁTH, Ľubomír DEDERA, Marcel HARAKAĽ

Abstract: Today, networks must face the threat of their systems being compromised by misuse or malicious access. The paper presented examines the role of Network Access Control (NAC) and compares approaches that can help to:

- reduce the risk of security incidents and increase compliance with security policies by enforcing IT security policies as a prerequisite for network access,
- dramatically reduce the number and severity of security events and aid in regulatory compliance.

Adding Network Access Control (NAC) to an existing network is a dramatic and significant change to the physical network. When NAC is in place, the network is no longer a neutral substrate for moving packets around as quickly as possible. Instead, it becomes a security barrier which can authenticate users, evaluate the security of end-point systems, and apply access control focused on the user and his/her security status. A NAC-enabled network is no longer a utility, like power and water, but must be tailored to fit organizationally into networking, security, and desktop management teams to be effective [1].

Keywords: Network Access Control, network security, Network Admission Control.

1. NAC DESCRIPTION

NAC also known as Network Admission Control defined by Cisco Systems is the approach to use network infrastructure to enforce security policy on the devices (PCs, PDAs) using the network. NAC can find noncompliant endpoint devices, put them to the quarantine and allow them only restricted access to the local network where all necessary patches and updates are located. The main result of the enforced policy is a decrease of damages caused by malicious code, viruses, worms, etc. NAC is part of the Cisco Self-Defending Network. Its goal is to create additional intelligence in the network to automatically identify, prevent, and adapt to security threats.

NAC function can be compared to the antivirus checking solution, where the device wishing to access a network is verified to comply with operating system version and service packs installed, presence of an antivirus software and age of the signature file and other important security related aspects.

A possible NAC solution is depicted in Fig. 1 and consists of four main parts:

- hosts attempting network access,
- network access devices (NADs),
- policy server decision points,
- management system.

A host is attempting to access for the first time a corporate network, which usually has some improvements of local security on place. To ensure that the security policy is on the appropriate level, Cisco implementation of NAC requires the Cisco Trust Agent to be installed on the accessing host. The agent is a software module communicating with operating system and third party security applications (e.g. AV software) and its role is to verify versions, release numbers, and dates of signature files. This information is provided during a connection attempt to the system and used by decision points to grant or deny access to the device.

Network access devices are the active devices of the first contact that enforce admission control policy and they are usually switches, wireless access points, security appliances or routers. The devices get host credentials and pass them to the servers, where network admission decisions are made. Based on the network security policy a host gets permit, deny, quarantine or restrict authorization.

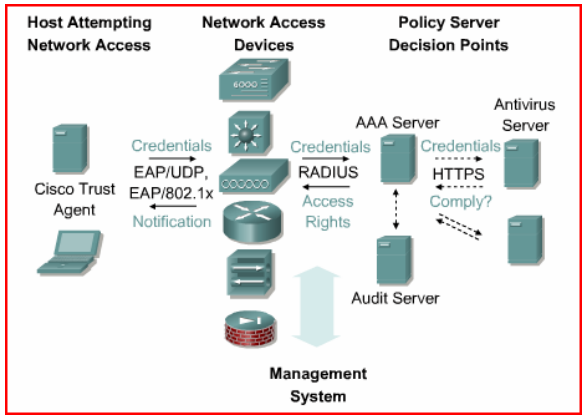


Fig. 1 NAC realization

Policy servers are the devices with knowledge about current versions of operating systems, antivirus programs and other security related applications used, together with the most current versions of updates and signature files required to fulfill the security policy. The policy servers work together with cosponsor application servers and audit servers which aid in assessing systems.

Management system provides monitoring and reporting operational tools.

2. NAC OPERATION

To better understand what happens each time a client connects to the network infrastructure, see the diagram in Fig. 2.

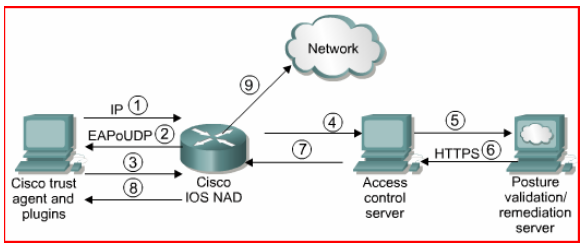


Fig. 2 NAC operation diagram

NAC component interaction occurs as follows:

1. Client sends a packet through a NAC-enabled router.
2. NAD begins posture validation using Extensible Authentication Protocol (EAP) over UDP (EOU).
3. Client sends posture credentials using EOU to the NAD.
4. NAD sends posture to Cisco ACS using RADIUS.
5. Cisco Secure ACS requests posture validation using the Host Credential Authorization Protocol (HCAP) inside an HTTPS tunnel.
6. Posture validation/remediation server sends validation response of pass, fail, quarantine, etc.
7. To permit or deny network access, Cisco Secure ACS sends an accept with ACLs/URL redirect.
8. NAD forwards posture response to client.
9. Client is granted or denied access, redirected, or contained.

In case that a host attempting to gain access to the network has no client software supporting NAC installed, the system prevents this access via NAD. Restriction applied on NAD prevents the host to go anywhere else except the defined network segment where the necessary installation packages reside [2].

3. COMPARISON WITH OTHER SYSTEMS

Microsoft Network Access Protection (NAP) includes client and server components. Administrators can configure IPSec enforcement, 802.1X enforcement, VPN enforcement, DHCP enforcement, or all of them, depending on their needs. NAP provides an infrastructure and an API, which vendors and software developers can use to build their own health validation and limited network access or communication components.

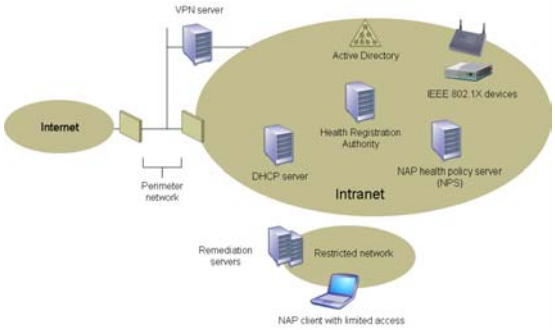


Fig. 3 Example of an intranet that has NAP deployed

- The example of intranet (Fig. 3) is configured for the following:
- Health state validation, health policy compliance, and limited network access for noncompliant NAP clients;

IPSec enforcement, 802.1X enforcement, VPN enforcement, and DHCP enforcement.

Recently Microsoft (23.5.07) also revealed plans to support interoperability between Microsoft's Network Access Protection NAP and the Trusted Computing Group's Trusted Network Connect TNC [3] NAC standard.

Exactly how will TNC/NAP interoperability work? Microsoft has contributed its SOH protocol to TCG, which has published it as a new TNC specification, IF-TNCCS-SOH, and it is now an open standard available for anyone to freely download or implement. Briefly, IF-TNCCS-SOH is a client/server protocol for reporting on the health — that is, security state — of a client before providing a network connection. In particular, the IF-TNCCS-SOH protocol complements IF-TNCCS, which is the existing TNC protocol for such checks. Some examples of health checks might include: presence of the security agent, firewall running, BIOS intact, latest OS patches, up-to-date antivirus software, no malware detected, only approved applications installed.

Microsoft NAP and the TCG TNC architecture interoperate at three points in either a TNC compliant or NAP-protected network:

- Endpoint,
- Policy Decision Point (PDP),
- Policy Enforcement Point (PEP) [4].

4. CONCLUSION

One of the most time-consuming challenges that network administrators face is ensuring that computers that connect to private networks are up to date and meet security policy requirements.This

complex task is commonly referred to as maintaining of computer health. Enforcing requirements is even more difficult when the computers, such as home computers or traveling laptops, are not under the administrator's control. NAC is the approach how to face such challenges and is supported by leading companies in networking and operating systems area [5].

Prior to NAC introduction, administrators had to manage security in different ways. They used domain policies, strong and direct management of user's workstations which was very time consuming and required a lot of human actions. Although there were automatic tools to support administrator's activities, computers and devices which were not included in the domain were not covered and were left out of control. NAC approach covers all devices attempting to connect to the network and provides a higher level of security both for local users and the rest of the network and we expect that this solution will be more and more widely used to secure corporate networks in the near future.

It important to note, that security enhancements and new technologies are developed to provide more reliable, stable and secure computing environments. From the modern army's perspective we see extreme demand on security and continuous movement toward new version of IP protocol – IPv6. More details about some aspects of transition to IPv6 in military environment can be found in [6].

ACKNOWLEDGEMENT

This work was supported by DoD project ŠPP 118_06-RO02_RU21-240 "NATO Network Enabled Capability and implementation in the Armed Forces of the Slovak Republic".

References

[1] SNYDER, J.: NAC Deployment a Five Step Methodology. [Online] Opus One, February 2007. [Cited: July 3, 2007.] www.juniper.net/solutions/literature/white_papers/nac_deployment_opus_one.pdf.

[2] Cisco Systems. Network Security 2.0. Chapter 4. [Online] Cisco Systems, 2004. http://cisco.netacad.net.

[3] TNC Workgroup. Trusted Network Connect Work Group. [Online] 2007. https://www.trustedcomputinggroup.org/groups/network/.

[4] TNC, Microsoft. Standardizing Network Access Control: TNC and Microsoft NAP. Trusted Computing Group: Trusted Network Connect. [Online] May 2007. [Cited: July 3, 2007.] https://www.trustedcomputinggroup.org/groups/network/.

[5] Microsoft Corporation. Introduction to Network Access Protection. [Online] 2007. http://www.microsoft.com/technet/network/nap/napoverview.mspx.

[6] KADERKA, J.: Přechod z IPv4 na IPv6 v AČR. Brno : AFCEA-DELINFO-Univerzita obrany, 2007. 11. mezinárodní konference spojovacího vojska AČR a 9. mezinárodní konference ITTE Komunikace v prostředí NEC. Vols. CD-ROM, p. 10. ISBN 978-80-239-9156-7.

Ing. Július BARÁTH, PhD.
doc. RNDr. Ľubomír DEDERA, PhD.
doc. Ing. Marcel HARAKAĽ, PhD.
Department of Informatics
The Academy of the Armed Forces
of General Milan Rastislav Štefánik
Demänová 393
031 01 Liptovský Mikuláš
Slovak republic
E-mail: julius.barath@aoslm.sk
dedera@aoslm.sk
harakal@aoslm.sk

KVALITA SLUŽBY V IP SIEŤACH PRE MULTIMEDIÁLNE PRENOSY

QUALITY OF SERVICES IN THE IP NETWORKS FOR MULTIMEDIA COMMUNICATIONS

Milan GOTTSTEIN

Abstract: Nowadays, a question about convergence voice and data networks is still more relevant. The best solution for convergent networks is platform based on TCP/IP protocol. These types of nets weren't projected for multimedia transfers in the real time; therefore we need a consistent solution of quality of services. This contribution indicates a possibility, how to solve QoS in IP networks, moreover, shows results and influence of some precautions towards QoS, according an original simulation model.

Keywords: Quality of service – QoS, Voice over IP – VoIP, IP networks, Simulation of QoS in IP networks.

1. KVALITA SLUŽBY PRE KONVERGOVANÉ SIETE

Otázky požiadaviek na niektoré parametre kvality služby sú relatívne nové a súčasné dátové siete ich nemuseli nevyhnutne dodržiavať. Jedná sa o prenos hlasu a ostatných služieb v reálnom čase, ktoré vyžadujú nízke oneskorenie a nepretržitú dopravu relatívne veľkého počtu malých datagramov. Ak sa napríklad pri prenose súborov v sieti niekoľko datagramov oneskorí, dátový tok sa na jednotky sekúnd preruší a pod., užívateľ sťahujúci súbor si toto väčšinou ani nevšimne. Ak však takéto javy nastanú pri telefonovaní cez dátovú sieť, kvalita hovoru sa zníži, alebo nastane nepríjemný výpadok. Podobná nepríjemná situácia nastane pri dočasnom, alebo trvalom preťažení siete. Ak sa sieť preťaží pri dátových prenosoch (sťahovaní súborov), užívateľom sa prenos spomalí a na svoj súbor budú čakať o niečo dlhšie. Ak sa však sieť preťaží pri telefonovaní, účastníkov budú obťažovať nepríjemné výpadky a to nie len pre hovory ktoré sú „navyše", ale pre všetky hovory cez preťažený úsek siete. Z uvedeného vyplýva, že riešenie kvality služby je nevyhnutné pre implementáciu IP telefónie do dátových sietí.

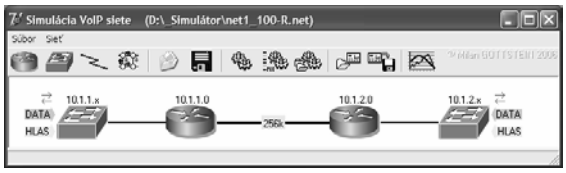
Mnohé tieto vlastnosti sú dané parametrami daného výrobku nasadeného do infraštruktúry siete. Aj protokoly sú štandardné a nie je možné pre danú aplikáciu vymýšľať zvláštne protokoly. Existuje však niekoľko systémových parametrov, ktorých voľba môže ovplyvniť kvalitu služby siete. Jedným z týchto parametrov je fragmentácia dlhých datagramov. Ak sa totiž popri krátkych hlasových datagramoch sieťou prenášajú príliš dlhé dátové datagramy, môže to spôsobiť nežiaduce kolísanie oneskorenia hlasových datagramov, pretože aj pri zavedení priorít hlasu sa neprerušuje vysielanie (dlhého) datagramu. IP protokol má podporu fragmentovania dlhých datagramov, ale otázkou je, akú zvoliť hranicu veľkosti datagramu, od ktorej sa budú datagramy fragmentovať. Tento parameter sa nazýva MTU (Maximum Transmission Unit). Jeho voľba je kompromisom medzi kolísaním oneskorenia hlasových datagramov a zaťažením siete fragmentmi dátových datagramov (zvyšovaním pomeru veľkostí hlavičiek k užitočnej informácii datagramov) a toto optimum môže byť pre každú konkrétnu sieť iné. Ďalšie možnosti zlepšenia kvality služby je využitie špeciálnych metód činnosti a riadenia prvkov siete (na základe štandardných protokolov).

2. SIMULÁCIA KONVERGOVANEJ SIETE

Pre simuláciu niektorých vlastností siete s konvergovanými hlasovými a dátovými službami, som vytvoril simulačný model vo vývojovom prostredí Borland Delphi 7. Nie je to konkurent profesionálnych programov (ako napr. Opnet a pod.), ale bol vytvorený preto, aby bolo možné vykonávať simulácie rôznych experimentálnych metód, ktoré štandardné nástroje pochopiteľne nepodporujú. Aplikácia je vytvorená v modernom vizuálnom vývojovom prostredí s využitím objektovo orientovaného programovania – jednotlivé prvky simulačného modelu sú vytvorené ako triedy, čo umožňuje programátorovi jednoduché a prehľadné doplňovanie ďalších vlastností simulačného modelu.

Pre overenie možností riešenia kvality služby v konvergovaných sieťach boli pomocou uvedeného simulačného programu vykonané simulácie jedného úseku siete a hypotetickej štruktúry siete pre rôzne charakteristiky záťaže.

Najskôr som vykonal simulácie pre veľmi jednoduchú štruktúru siete (resp. jeden spoj), ktorá je na obr. 1 (pohľad na hlavné okno programu).



Obr. 1 Štruktúra siete – spoja pre simuláciu

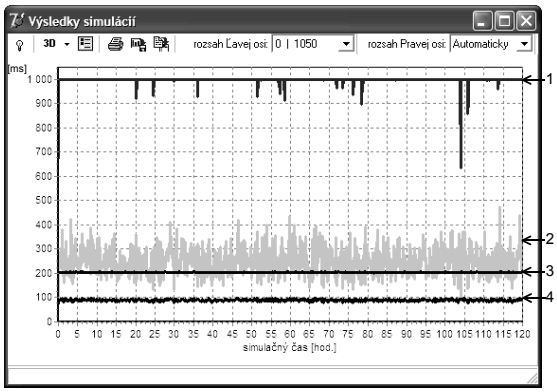
Pre jednotlivé prvky siete je možné v simulačnom programe nastavovať rôzne parametre, ktorých popis a ukážku pre obmedzený rozsah tohto príspevku neuvádzam. Práca v programe je intuitívna, formuláre pre nastavenie parametrov sú doplnené popismi a nápovedou, čo robí program použiteľným aj bez podrobnejšieho návodu na obsluhu (aj keď nebol určený pre širšie využitie).

2.1 Uprednostňovanie hlasových datagramov

Základným princípom zabezpečenia kvality služby v koncepcii diferencovaných služieb [1] je uprednostňovanie hlasových datagramov (datagramov, prenášajúcich informácie v reálnom čase) pred dátovými (ostatnými). Nasledujú výsledky simulácie pre sieť (úsek siete – obr. 1), ktorý je približne v rovnakom pomere prenosových rýchlostí zaťažený hlasovými a dátovými datagramy a stredná intenzita toku datagramov je približne rovnaká, ako kapacita spoja, čím nutne musia vznikať straty (zahodenie datagramov).

Obr. 2 vyjadruje priebehy oneskorenia v [ms] pre prípad, kedy sa nepoužilo uprednostňovanie hlasových datagramov a pre prípad, kedy hlasové datagramy mali prioritu:

1. maximálne oneskorenie datagramov v sieti bez QoS (bez priorít hlasu);
2. stredné oneskorenie datagramov v sieti bez QoS (bez priorít hlasu);
3. maximálne oneskorenie datagramov v sieti s QoS (priorita hlasu);
4. stredné oneskorenie datagramov v sieti s QoS (priorita hlasu).

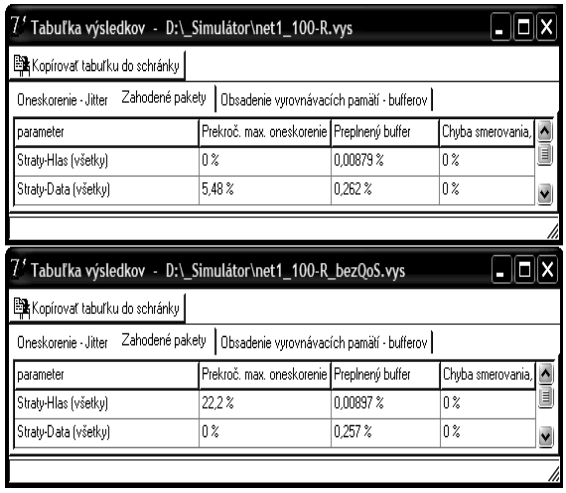


Obr. 2 Priebehy oneskorenia hlasových datagramov pre sieť s QoS a bez QoS

Z výsledkov uvedenej simulácie je vidno výrazné zlepšenie ukazovateľov QoS pri zavedení priority hlasových datagramov. To, že v uvedenom grafe oneskorenie nepresiahne 1 sekundu je dané nastavením, ktoré pri väčšom oneskorení považuje datagram za stratený.

Simulačný program udáva celkové oneskorenie hlasových datagramov – nie len oneskorenie čakaním vo vyrovnávacích pamätiach smerovačov a oneskorenie prenosom, ale aj oneskorenie spôsobené kódovaním hlasu pre vybraný kodek [1].

Stratovosť datagramov je pre lepšiu prehľadnosť zobrazená vo forme tabuľky (obr. 3). Prvá tabuľka vyjadruje stratovosť pre sieť s uprednostňovaním hlasových datagramov a spodná tabuľka v uvedenom obrázku stratovosť pre sieť bez využitia metód QoS. Výsledky ukazujú, že zhoršenie stratovosti dátových datagramov pri zavedení priority hlasových datagramov nie je výrazné.

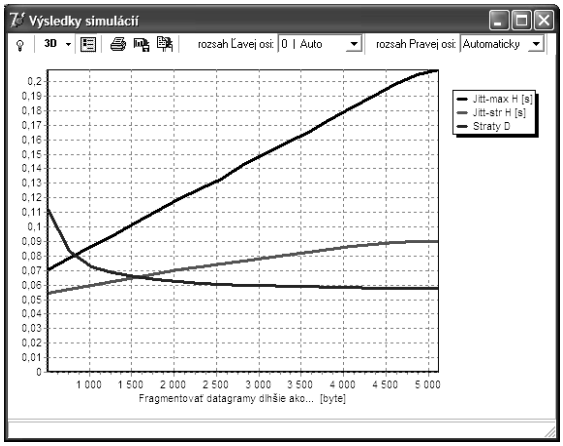


Obr. 3 Stratovosť datagramov pre sieť s QoS a bez QoS

2.2 Vplyv fragmentácie dlhých datagramov

Dlhé dátové datagramy môžu negatívne ovplyvniť oneskorenie hlasových datagramov resp. kolísanie ich oneskorenia (jitter). Riešením tohto problému je fragmentácia dlhých datagramov [1]. Naopak fragmentácia na veľa malých fragmentov navyše zaťažuje sieť (hlavičkami, ktoré musí mať každý fragment) a zvyšuje sa pravdepodobnosť straty datagramu. Nasledujú výsledky simulácie pre sieť (úsek siete), ktorý je približne v rovnakom pomere prenosových rýchlostí zaťažený hlasovými a dátovými datagramy a stredná intenzita toku datagramov je približne rovnaká, ako kapacita spoja – nutne musia vznikať straty. Obr. 4 popisuje priebehy maximálneho a stredného oneskorenia hlasových datagramov (jitter) v [s] a stratovosti dátových datagramov (pomer stratených datagramov ku všetkým) v závislosti na zvolenom parametri fragmentovania. Generované datagramy sú do

veľkosti 5 000 bajtov, tak že hodnota horizontálnej osi nad 5 000 je zároveň hodnotou bez fragmentácie.



Obr. 4 Závislosť oneskorenia hlasu a stratovosti dát na parametri fragmentovania

Z uvedeného grafu je možné určiť, že pre tento prípad je vhodné fragmentovať datagramy väčšie ako cca 1 000 až 1 500 bajtov.

Výsledky simulácie vplyvu fragmentácie na kvalitu služby popisujú praktický vplyv fragmentácie a naznačujú potreby zavedenia fragmentácie dlhých datagramov. Často kladenou otázkou môže byť, či je potrebné pre danú sieť zaviesť fragmentáciu datagramov a akú medzu veľkosti datagramu (MTU) použiť. Pomocou tohto simulačného modelu je možné určiť vplyv fragmentácie pre rôzne prenosové rýchlosti spojov medzi uzlami

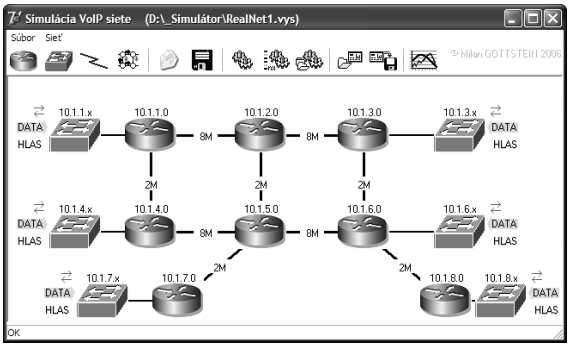
2.3 Niektoré špeciálne metódy QoS

Princíp fragmentovania dlhých datagramov je účinný pre zamedzenie veľkého oneskorenia hlasových datagramov, ktoré prídu práve v čase vysielania dlhého datagramu. Nevýhodou fragmentácie je väčšie zaťaženie siete tzv. „neplatenou záťažou“ (hlavičky) a niekedy nie je vhodné povoľovať fragmentáciu z dôvodu možných útokov na sieť. Pre vysokorýchlostné spoje síce nie je vplyv dlhých datagramov signifikantný, ale pre nižšie rýchlosti je potrebné tento problém riešiť. Niektoré štúdie (uvedené napr. v [2]) sa snažia štatistickými metódami predpovedať príchod hlasového datagramu a podľa toho vo vhodný okamih zahájiť vysielanie dlhého dátového datagramu. Pri stochastickom toku hlasových datagramov z viacerých zdrojov však tieto metódy prestávajú byť účinné. Iná možnosť je vyslať najdlhší datagram z fronty okamžite za hlasovým (kedy je najmenšia pravdepodobnosť príchodu ďalšieho hlasového datagramu). Túto možnosť som implementoval do simulačného modelu ako „špeciálne metódy QoS“, ktoré sa líšia počtom datagramov vo fronte, pre ktoré sa mení poradie. Zo simulácií, ktoré som vykonal na niekoľkých sieťach však vyplýva, že výsledný efekt týchto metód nie je významný a zrejme väčšina takýchto metód zostane iba v abstraktnej rovine ako predmet teoretického bádania.

2.4 Simulácia reálnej siete

Simulácia vyššie uvedenej jednoduchej siete je vhodná pre overenie rôznych metód a prístupov, ale simulačný program umožňuje simulovať reálnu záťaž v rozsiahlej sieti (až 254 uzlov a po drobnej úprave programu aj viac).

Na obr. 5 je príklad reálnej siete, pre ktorú bola vykonaná simulácia.



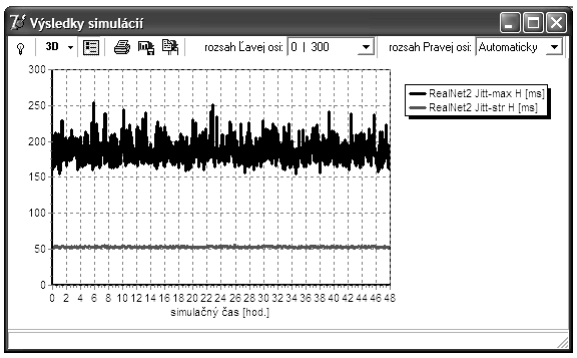
Obr. 5 Štruktúra reálnej siete

Tab. 1 Výpis tabuliek výsledkov simulácie pre jednotlivé smery a úseky siete

Parameter pre spoj medzi uzlami	Max. oneskorenie hlasových datagramov
Jitter-Hlas (1, 3)	95 ms
Jitter-Hlas (4, 6)	100 ms
Jitter-Hlas (3, 4)	159 ms
Jitter-Hlas (1, 4)	160 ms
Jitter-Hlas (1, 6)	228 ms
Jitter-Hlas (3, 6)	228 ms
Jitter-Hlas (1, 7)	*1 000 ms*
Jitter-Hlas (1, 8)	*1 000 ms*
Jitter-Hlas (3, 7)	*1 000 ms*
Jitter-Hlas (3, 8)	*1 000 ms*
Jitter-Hlas (4, 7)	*1 000 ms*
Jitter-Hlas (4, 8)	*1 000 ms*
Jitter-Hlas (6, 7)	*1 000 ms*
Jitter-Hlas (6, 8)	*1 000 ms*
Parameter pre spoj	*Maximálne obsadenie vyrovnávacej pamäte*
Obsad. Buff. (1, 2)	*7,09%*
Obsad. Buff. (2, 5)	*8,24%*
Obsad. Buff. (2, 3)	*9,74%*
Obsad. Buff. (1, 4)	*10,20%*
Obsad. Buff. (4, 5)	*10,30%*
Obsad. Buff. (5, 6)	*11,60%*
Obsad. Buff. (3, 6)	*16,90%*
Obsad. Buff. (5, 7)	*100%*
Obsad. Buff. (6, 8)	*100%*

Z výsledkov simulácie pre uvedenú sieť a zadané požiadavky je bolo zistené, že sieť nevyhovuje požiadavkám na kvalitu služby (pre požadovaný rozsah článku nie sú tieto výsledky uvedené). Z tab. 1, ktorá je výpisom z tabuliek výsledkov simulácie, je možné odhaliť „slabé miesta“ siete. Je to preťaženie spoja medzi uzly 5–7 a 6–8. Ďalej by mohlo priniesť zlepšenie QoS posilnenie spoja 3–6 a ak sa zvýši priepustnosť z uzlov 7 a 8, bude asi kritický aj spoj 5–6.

Na základe týchto poznatkov bola sieť doplnená o ďalšie spoje (4–7 a 5–8) a na obr. 6 je priebeh stredného a maximálneho oneskorenia hlasových datagramov v celej sieti. Tieto výsledky sú už uspokojivé a podľa detailnejších výsledkov by bolo možné hľadať a posilňovať ďalšie kritické miesta siete.



Obr. 6 Výsledky simulácie oneskorenia hlasových datagramov v doplnenej sieti

3. ZÁVER

Pri projektovaní siete je potrebné klásť dôraz na vhodný výber zariadení a návrh optimálnej štruktúry siete, prenosových kapacít a použitých technológií. Uvedené výsledky simulácie slúžia pre získanie prehľadu o vplyve jednotlivých parametrov na kvalitu služby siete a simulačný model by sa prípadne mohol využiť aj pre overenie jednotlivých prístupov v riešení projektu a na overenie parametrov celej siete. Prezentované výsledky dávajú dostatočný dôkaz pre nutnosť riešenia QoS pre hlasové služby a ukazujú smery riešenia. Jednoznačne je tu dokázané, že bez uplatnenia priorít hlasových datagramov nie je väčšina sietí schopná v požadovanej kvalite prenášať hlas. Predimenzovanie kapacít prenosových ciest neprináša efektívne výsledky. Ďalej sa zabúda na možnosť fragmentovania dlhých datagramov. Z vyššie uvedených výsledkov je vidno, že aplikácia fragmentovania dlhých datagramov v smerovačoch siete môže priniesť podstatné zlepšenie kvality služby a to už pri veľkostiach MTU (maximálna veľkosť prenášaného datagramu), ktoré podstatne nezvyšujú zaťaženosť siete. Ostatné metódy správy fronty datagramov, ktoré by mali zlepšiť kvalitu hlasovej služby v danej sieti, zrejme nebudú mať takú efektívnosť, ako priority a fragmentácia.

Zoznam bibliografických odkazov

[1] GOTTSTEIN, M.: Prenos hlasu v IP sieťach. AOS Liptovský Mikuláš 2005.
[2] Zborník konferencie: 7th NATO regional conference on military communications and information systems 2005. 4. - 5. 10. 2005, Military Communication Institute in Zegrze, Poland.
[4] ŠMRHA, P., RUDOLF, V.: Internetworking pomocí TCP/IP. KOPP České Budějovice 1994.
[5] Svět sítí [online]. Dostupné na Internete: http://www.svetsiti.cz/
[6] H.323 Forum [online]. Dostupné na Internete: http://www.h323forum.org/
[7] Open H323 [online]. Dostupné na Internete: http://www.openh323.org/
[8] ITU-T Recommendation H.323 „Packet Based Multimedia Communications Systems“. Telecommunication Standardization Sector of ITU, Geneva, Switzerland, 2002.
[9] ITU-T Recommendation H.245 „Control Protocol for Multimedia Communication“. Telecommunication Standardization Sector of ITU, Geneva, Switzerland, 2002.
[10] ITU-T Recommendation H.225.0 „Call signalling protocols and media stream packetization for packet-based multimedia communication systems“. Telecommunication Standardization Sector of ITU, Geneva, Switzerland, 2002.
[11] Cisco Systems [online]. Oficiálna Web stránka spoločnosti Cisco Systems, Inc. Dostupné na Internete: <http://www.cisco.com/>

Summary: Some of requests towards parameters QoS are relatively new and current data networks hadn’t to follow them perforce, for example: voice transmission and other real time services, which need a low time delay and a continuous transfer relatively a big number of small datagrams. I created a simulation model in Borland Delphi7, because of simulation some features of the net with convergent voice and data services. Said results of simulation are appointed to obtain a view about an influence of individual parameters to QoS. The simulation model should be also used for checking next parameters of the net and single approaches towards project solution. Said results give sufficient evidence about necessity to solve QoS for voice services and show convenient.

Ing. Milan GOTTSTEIN, PhD.
Veliteľstvo vzdušných síl OS SR, Zvolen
E-mail: milan.gottstein@mil.sk

AKTUÁLNÍ PROBLÉMY BEZPEČNOSTI IP TELEFONIE

CURRENT IP TELEPHONY SECURITY PROBLEMS

Jaroslav DOČKAL

Abstract: We often ask this question: Why invest to expensive Cisco etc. devices and not to use free software. This article is looking for an answer for this question by this way: firstly article shows results of laboratory experiments that illustrate how IP telephony is not resistant to network attacks. Secondly it describes possibilities that dispose IP telephony involved into Cisco network infrastructure.

Keywords: Security, IP telephony, attack, Snort, SIP, Call Manager.

1. ÚVOD

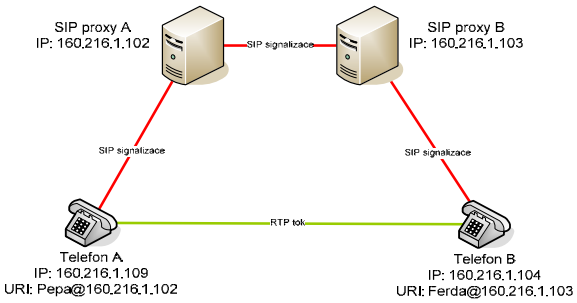
Počítačové systémy se mohou snadno stát objektem negativního působení dobře připravených útočníků. Proto musí být přípravě odborníků správy sítě v oblasti její ochrany věnováno potřebné úsilí. Problémem je, že zatímco u provozu datových sítí lze vycházet z mnohaletých zkušeností, v oblasti IP telefonie jsou tyto poznatky teprve postupně získávány. Je zde tudíž i velký prostor pro zajímavý výzkum.

Cílem příspěvku bylo předat poznatky a zkušenosti, které byly získány na Univerzitě obrany v Brně. Armáda České republiky (AČR) je v oblasti sítí rozsáhle vybavována produkty společnosti Cisco a v souladu s touto koncepcí i Laboratoř sítě Univerzity obrany. S pomocí těchto zařízení jsme se snažili ocenit bezpečnostní vlastnosti Cisco produktů v oblasti IP telefonie a jejich přednosti před volně dostupným softwarem. Bezpečnostní možnosti těchto produktů vyžadují jejich dobrou znalost a kvalifikovanou přípravu uživatelů. Přitom ani samy protokoly spjaté s IP telefonií nejsou triviální, o čemž svědčí například délka RFC popisujícího protokol SIP (nejdelší RFC ze všech) anebo abstrakce zápisu protokolu H.323 (byla použita notace ASN.1).

Při zkoumání bezpečnostních slabin IP telefonie jsme vyšli z taxonomie VOIPSA [9]. Vzhledem k významnosti rozptylu zpoždění vidíme jako zvlášť nebezpečné útoky typu DoS. Na webu VOIPSA (http://voipsa.org/Resources/tools.php) lze nalézt i několik desítek nástrojů pro provádění útoků na hlasové služby. V [3] naleznete popis základních způsobů, jak IP telefonii zabezpečit. My jsme se rozhodli provést základní experimenty a pro naše partnery v AČR zpracovat doporučení, jak postupovat při zabezpečení systémů na bázi Cisco produktů. Omezujeme se zatím na prostředí TCP/IP s protokolem IPv4. Podmínky práce v prostředí IPv6 (speciálně pro potřeby NNEC, viz například [1]) budou předmětem budoucího testování.

2. LABORATORNÍ EXPERIMENTY

Při přípravě laboratorních testů jsme vycházeli z poznatků projektu [8]. Pro hledání slabin jsme použili zprávu [7]. Cílem bylo připravit sadu útočných skriptů a sadu pravidel pro detekci těchto útoků pomocí Snortu. Laboratorní uspořádání ukazuje obr. 1. Byly použity dva softphony X-ten Lite (X_lite-Xten-Win32-1103m-14262.exe), a dva SIP proxy (proxy zajišťují směrování žádostí dle aktuálního umístění adresáta, autentizaci, účtování) používající software SIP Express Router.



Obr. 1 Uspořádání laboratorní sestavy pro potřebu testování

Na počítač s IP addresou 160.216.1.102 byl umístěn generátor sady útoků (/root/iptel/snort-rules-test/snort-rules-test script) a tyto útoky byly cíleny na SIP proxy o stejné adrese. Na počítač o IP adrese 160.216.1.103 byl nainstalován Snort IDS (nakompilovaný z iptel.org) a konfigurován k použití speciální sady pravidel pro hlášení detekovaných útoků (soubor /etc/snort/rules/sip.rules). Pro instalaci Snortu jsme použili systém apt, jen v konfiguračním souboru /etc/snort/snort.conf jsme standardní direktivy nahradili pomocí include /etc/snort/rules/sip.rules). Nejprve jsme otestovali obecné TCP/IP útoky Teardrop (útok, který zneužívá útok nespojitelnými fragmenty paketu), Ping of Death (ping paketem větším než 65 536 byte) a SYN Flood (útok záplavou SYN paketů při navazování TCP spojení), poslední viz:

```
alert tcp any any -> $SIP_PROXY_IP any \
(msg: "TCP SYN packet flooding from single
source"; \
threshold: type both, track by_src, count
200, seconds 20; \
flow:stateless; flags:S,12; sid:5000100; rev:1;)
```

Pak jsme se zaměřili na slabiny protokolu SIP a jeho realizací. Například analogickým útokem k SYN flood je INVITE flood. Pokud neexistuje volaný (404 Not Found message), odpověď je rychlá. Pokud však existuje, výzva je ukládána v paměti nejméně tři minuty. Zde je příklad pravidla Snortu ohlašujícího útok INVITE flood:

```
alert ip any any -> $SIP_PROXY_IP
$SIP_PROXY_PORTS \
(msg:"INVITE message flooding";
content:"INVITE"; depth:6; \
threshold: type both, track by_src, count
200, seconds 60; \
sid:1000100; rev:1;)
```

Obdobně lze vyčerpat zdroje SIP proxy pomocí záplavy zpráv REGISTER. Pokud je na proxy slabá autentizace, může se útočník zaregistrovat místo své oběti. Jiným útokem je přerušení toku zprávou BYE. Útok typu DoS lze realizovat rovněž řadou systémových zpráv, například 504 Server Time-out.

Složitějším DoS útokem je vytvoření smyčky mezi dvěma proxy. Například útočník zaregistruje uživatele userA v doméně A jako userA@domainA.com s kontaktní adresou userB@domainB.com a na serveru v doméně B registruje uživatele userB jako userB@domainB.com s kontaktní adresou userA@domainA.com. Ekvivalentem pole TTL v záhlaví IP paketu je pole Max-Forwards. Kontrolu nekonečných smyček jsme prováděli takto:

```
if (search("^(Contact|m):
.*@(10\.1\.2\.3\|sip-
proxy\.mydomain\.cz)")){
log("LOG: alert: someone is trying to set
aor==contact\n");
sl_send_reply("476", "No Server Address in
Contacts Allowed" );
break;
};
```

Záplavu zpráv lze vytvořit i pomocí jedné z funkcí proxy serveru zvané forking, která je náročná na paměť i zátěž procesoru.

Útok typu SQL injection je ekvivalentní vkládání kódu do HTTP zpráv, falešný kód lze vložit například do polí Username či Realm zpráv VoIP. SQL kód pak může vypadat například takto:

Select password from subscriber where username='myname'; DROP table Subscriber -- ' and realm='147.32.121.11'

Záhlaví je proto třeba kontrolovat, nejlépe je před SIP proxy umístit IDS.

Útočník také může pro DoS útok využít nerozložitelná DNS jména v některém z polí záhlaví, např. v poli Request-URI, To atd. Tento útok se používá pro posílení útoku na zprávu INVITE, zdroje jsou pak vyčerpávány mnohem rychleji. Detekovat ho lze například testováním abnormálního počtu výskytu odpovědí „No such name":

```
alert udp $DNS_SERVERS 53 -> $SIP_PROXY_IP any \
msg:"DNS No such name treshold"; \
content:"|83|"; offset:3; depth:1; \
threshold: type both , track by_src, count 2000, seconds 60; \
sid:1000400; rev:1;)
```

Snort jsme při svých testech spouštěli příkazem snort –i eth0 –A console –c /etc/snort/snort.conf a sbírali odpovědi. Pro debugging jsme používali příkazy typu ser –l 160.216.1.102 –D –E –ddd.

3. OBRANA PROTI ÚTOKŮM V PROSTŘEDÍ CISCO PRODUKTŮ

Útoky proto IP telefonii lze vést třemi směry:

3.1 útoky proti koncovým bodům;
3.2 útoky proti serverům IP telefonie.

3.1 Ochrana koncových bodů

V obecné rovině je třeba využívat možnosti, které poskytují síťové prvky: používat přístupová pravidla (Access Control List – ACL), autentizovat směrovací informaci, využívat detektory průniku a bezpečnostní management sítě. Na Catalystu (např. 3550), k němuž je připojen IP telefon, je třeba pro dané rozhraní nastavit MicroFlow (nejlépe 6kb/s) příkazem mls qos flow-policing.

Dále je třeba využívat možností, které poskytují bezpečnostní možnosti vyšších řad Catalystů, viz [2]:

- Oddělení hlasových a datových VLAN – zabrání se tím útoku např. pomocí nástroje VOMIT. Telefony spolu komunikují přes RTP resp. SRTP, neboli není třeba, aby používaly TCP nebo ICMP.
- Nastavení ACL pro každou VLAN.
- Ochrana jednotlivých portů, tj. povolené MAC adresy resp. jejich počet (proti útoku typu DoS) – IP Source Guard (IPSG). Toto opatření chrání před tzv. IP a MAC spoofingem.
- Ochrana před P2P provozem (KaZaa, Morpheus, Groekster, Napster, iMesh atd.) a hrami (Doom, Quake, Unreal Tournament atd.) pomocí tzv. scavenger-class provozu (mapuje se do DSCP CS1), který má nižší prioritu, než tzv. Best effort.

- Ochrana před neautorizovanými DHCP odpověďmi (DHCP Snooping), například při útoku man-in-the-middle:

```
!Zapnutí sledování na přepínači:
Switch(config)# ip dhcp snooping
Switch(config)# ip dhcp snooping vlan 1
Switch(config)# interface
GigabitEthernet 5/1
Switch(config-if)# ip dhcp snooping
trust
```

- Ochrana před útokem MITM (Man in the Middle) falešnými ARP odpověďmi (generuje např. ettercap či dsniff) – Dynamic ARP inspection:

```
Switch(config)#ip arp inspection vlan 1
Switch(config)#int range f1/1-4, f2/24
Switch(config-if)#ip arp inspection
trust
```

Další skupinou opatření je zodolnění telefonu: podepsání obrazu systému ve firmware, podepsání konfiguračních souborů, vypnutí PC portu, nastavení tlačítek, hlasitosti a webového přístupu. Vypnutí webového přístupu ale vyřadí XML aplikace, lépe je mezi telefonem a serverem pomocí ACL blokovat jiné porty než 80.

Na IP telefonu je uložena celá řada cenných informací: IP adresa a maska, adresa defaultní brány, DHCP serveru, DNS serveru, TFTP serveru, CallManageru, adresáře, logon serveru a XML serveru. Konfiguraci lze chránit pomocí CAPF utility, umožňující generovat dvojici klíčů, šifrovat, dešifrovat podepisovat, kontrolovat podpisy, ukládat, číst a rušit certifikáty i dvojice klíčů, instalovat lze i certifikáty s lokální platností.

SIP si generuje na rozdíl od SCCP vlastní klíč pro každou relaci zajištěnou protokolem SRTP (podpora šifrování se samozřejmě týká jen telefonů Cisco). Protokol SRTP je použit s parametry pro HMAC-SHA-1 autentizaci a AEC-128-CM pro šifrování. Přidává čtyři byte na paket, což přenos hlasu zpožďuje. Pro zabezpečení registračních požadavky se používá protokol TLS (SRTP zatím ne) s parametry: algoritmus podpisu RSA, autentizace ve variantě HMAC-SHA-1 a šifrování ve variantě AES-128-CBC. Mezi dvěma telefony může být i smíšené přenosové prostředí (jeden telefon komunikuje s bezpečnostní branou pomocí SRTP a druhý pomocí RTP).

Cisco telefony podporují Proxy EAPOL-Logoff, který probíhá takto: PC pošle zprávu EAPOL-Logon, authentifikátor se dotáže AAA serveru, ten pošle nový identifikátor VLAN ID. Když pak PC pošle dotaz na DHCP, je vytvořena tabulka „DHCP Snooping Binding“ a monitorována shoda s IP-SG a DAI. V případě odpojení PC telefon vyšle zprávu EAPOL-Logoff. Žádné cizí PC nemůže tuto zprávu poslat a tím je zajištěna ochrana pře útokem zvaným „EAPOL-LogOff DoS“.

Cisco vloni v listopadu provedlo akvizici společnosti Metreos. Získala tím VoIP firewall běžící nad operačním systémem Fedora, který po autentizaci ze specifických IP adres telefonů otevírá příslušné porty výhradně v době volání. Je to ideální softwarový firewall pro volání z domu, pobočky či ve vojenských podmínkách z odloučeného útvaru.

3.2 Ochrana serveru IP telefonie

Cisco používá jako server IP telefonie produkt CallManager, což je silný, ale zároveň komplikovaný nástroj, jehož možností není snadné využívat. Cisco je výrobce, který u svého produktu konstrukčně vychází z koncepce co nejmenší komunikace s hostujícím prostředím. Útoků typu DoS to však nezabrání, například u Windows je přes 80 % útoků cíleno na IIS. Proto Cisco vyžaduje vypnutí těchto služeb ve svých manuálech pro CallManager 4.x.

V roce 2004 vyzvala redakce Network World [4] pět nejznámějších výrobců IP telefonie, aby ji poskytli své produkty k testování. Obdrželi je pouze od dvou – jeden z nich bylo právě Cisco. Nejlepší hodnocení obdrželo Cisco s CallManagerem 4.0. Do něj byly doplněny digitální certifikáty potvrzující identitu síťových zařízení, a šifrování pro komunikaci mezi CallManagerem a IP telefonem. Odolnost systému byla posílena pomocí řešení Cisco Security Agent (CSA). V jednom balíčku je zde spojeno více úrovní zabezpečení, a to prevence před průniky, autentizace IP telefonů, distribuovaný firewall a ochrana před internetovými viry. Cisco je jediný výrobce, jehož IP PBX byla po testování v laboratoři Miercom označena jako SECURE.

CallManagerem 4.1 pak přinesl komunikaci s CallManagerem pomocí SHTTP, použití SSL pro LDAP (SLDAP), plnou TLS a SRTP podporu, SRTP pracující v SRST módu (Survivable Remote Site Telephony), neboli směrovač může v případě ztráty WAN spojení zaskočit za CallManager.

V roce 2006 přišlo Cisco s Unified CallManagerem 5.0, u kterého je hostitelským prostředím Linux. CallManager je zde vybaven celou řadou nových bezpečnostních vlastností spojených převážně s přechodem na protokol SIP. Jsou to především:

- automaticky instalovaný Cisco Security Agent;
- rychlý reset hesla (v případě jeho zapomenutí není třeba obtěžovat správce systému);
- bezpečný reset hesla administrátora bez přerušení poskytování služeb;
- bezpečnostní profily telefonu s protokolem SIP, SCCP či trunkem;
- příjem datumu a času od NTP serveru a jejich další výdej;
- protokol TLS pro autentizaci a šifrování přenosu s SIP telefony;

- protokol SRTP (Secure Real-Time Transport Protocol) pro komunikaci mezi SIP telefony a s bránou MGCP;
- protokol IPSec pro tunelování k branám s IOSem;
- použití SSL pro dotazy na adresář LDAP;
- odchyt a zaslání zašifrované signalizace k Cisco Technical Assistance Center (TAC);
- hostující firewall;
- možnost generovat a rušit certifikáty;
- HTTPS rozhraní místo HTTP;
- rozdělení uživatelů do základní a pokročilé skupiny;
- různé administrátorské účty zaručující pružnost řízení přístupu;
- vypnutí po třiceti minutách pasivity.

V průběhu roku 2007 přišel Cisco unified Communications Manager 6.0, který poskytuje nové informace pro konfiguraci firewallů, přístupových seznamů ACL (Access Control List), a řítení kvality služby. Na počátek roku 2009 byl ohlášen Cisco Unified Communications Manager 7.0 s databází na bázi IBM Informixu s výbornými bezpečnostními vlastnostmi.

Před jakýkoliv server CallManager je potřebné umístit firewall s TLS proxy funkcí pro zašifrovanou signalizaci i data. Firewally PIX a zařízeních ASA navíc umožňují omezit rychlost a prioritní režim LLQ (Low Latency Queuing).

ZÁVĚR

Často padne otázka: Proč investovat do nákladných zařízení Cisca pro IP telefonii, když je na Internetu volně dostupný použitelný software. Pokud opomeneme výkon a možnosti poskytování kvality služby, je zde navíc podstatný rozdíl v úrovni bezpečnostních vlastností. Za to se ovšem platí i rozsahem a hloubkou požadovaných znalostí správy sítě. Ukázat tento rozdíl bylo snahou autora tohoto článku.

Seznam bibliografických odkazů

[1] BARÁTH, J. DEDERA, Ľ. HARAKAĽ, M. LÍŠKA, M. Protokol IPv6 a jeho potreba pre dosiahnutie NNEC, ITTE 2007, Brno, 2.5.2007, ISBN 978-80-239-9156-7.
[2] Using the Catalyst Integrated Security Features (CISF) to Mitigate Network Attacks. Cisco 2004. http://ciscoexpo.eurorscg.bg/images/presentations/cisf3.pdf.
[3] DOČKAL, J. MALINA, R. MARKL, J. VANĚK, T. Bezpečnost internetové telefonie. DSM 6/06. s. 36-42.
[4] Independent Lab Test Report: Security of Cisco CallManagerbased IP Telephony against malicious hacker attacks. Mier Communications, Inc. Report Issued: 24 May 2004.
[5] MIER, E. BIRDSALL, R. THAYER R. Breaking through IP telephony. Network World Lab Alliance, Network World, 05/24/04.
[6] Independent Lab Test Report: Lab test summary – Report 060302. Mier Communications, Inc. March 2006.
[7] Oulu University Secure Programming Group, "PROTOS Test-Suite: c07-sip", University of Oulu, 2005, http://www.ee.oulu.fi/research/ouspg/protos/testing/c07/sip/.
[8] SNOCER project team: General Reliability and Security Framework for VoIP Infrastructures. 2005. http://www.snocer.org/.
[9] Security and Privacy VoIP Taxonomy. VOIPSA 2005. http://voipsa.org/Activities/VOIPSA_Threat_Taxonomy_0.1.pdf.

Summary: This paper provides pieces of knowledge and experiences acquired at the University in Defence, Brno, Czech Republic. The paper describes denial of service attacks (DoS) targeted against VoIP infrastructure based on the SIP protocol and gives examples of "field tested" rules for detecting these attacks by Snort IDS, and available countermeasures. Second part of the paper describes defence against attacks in environment of the Cisco IT telephony products, separately is described defence of endpoints of networks and IP telephony servers.

doc. Ing. Jaroslav DOČKAL, CSc.
Univerzita obrany, Kounicova 65
612 00 Brno
Česká republika
E-mail: Jaroslav.dockal@unob.cz

VYUŽITIE MIKROPÁSIKOVÉHO MENIČA FÁZY AKO NAPÁJAČA DYNAMICKY FÁZOVANEJ ANTÉNOVEJ SÚSTAVY

A MICROSTRIP PHASE SHIFTER AS AN DRIVEN ELEMENT OF DYNAMIC PHASED ARRAY

Ján HARING, Norbert MAJER, Peter POLOHA, Rudolf HRONEC

Abstract: An antenna system is a system compound from simply radiators (dipoles, microstrip antennas), which together form desired radiation pattern. During form a radiation pattern of antenna system the main emphasis is on the width, contour and direction of orientation. By the assistance this way of created patterns it is possible a radiation energy of antenna system to rout into desiderative direction, the reduce interference and increase efficiency entire transmission. In practice are known several techniques which can form radiation pattern of antenna system. The article deals with by description of technique based on feeding particular elements antenna system by different phase signals.

Keywords: Interference, Smart Antenna, Beamsteering, Phase Shifter, Microstrip Line.

1. INTELIGENTNÁ ANTÉNOVÁ SÚSTAVA

Základňové stanice mobilných rádiokomunikačných systémov v súčasnosti využívajú hlavne všesmerové, alebo smerové (sektorové) anténové sústavy. Použitie takýchto systémov je z hľadiska vysielacieho výkonu neefektívne, pretože maximum výkonu je vyžarované aj v iných smeroch, ako sa nachádza účastník. Naviac, výkon vyžiarený v iných smeroch predstavuje zdroj interferenčného signálu pre iných účastníkov. Z týchto dôvodov sa objavila koncepcia inteligentných antén. Na obr. 1 je zobrazené rozdelenie inteligentných anténových sústav podľa princípu ich činnosti. [1, 2, 3]

Obr. 1 Rozdelenie inteligentných anténových sústav podľa princípu ich činnosti

Anténová sústava s prepínanými lalokmi nazývaná aj mnohozväzková anténová sústava obsahujúca len základnú prepínaciu funkciu medzi samostatnými smerovými anténami, alebo vopred definovanými vyžarovacími lalokmi sústavy.

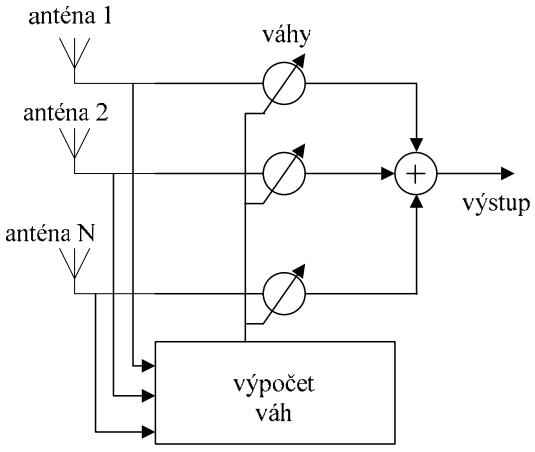
Dynamicky fázovaná anténová sústava využíva sústavu jednoduchých antén a kombinuje signály týchto antén na vytvorenie výstupu. Smer, v ktorom je dosiahnutý maximálny zisk, je riadený nastavením fázy medzi anténami.

V adaptačných anténových sústavách sú zisky a fázy prvkov nastavené pred kombinovaním signálov tak, aby sa zisk sústavy menil dynamicky.

Anténová sústava s optimálnym kombinovaním je sústava, v ktorej zisk a fáza každého prvku sú nastavené tak, aby sa dosiahla optimálna činnosť napr. maximálna hodnota pomeru SNR.

1.1 Princíp inteligentnej anténovej sústavy

Termínom inteligentná anténa sa označujú všetky systémy v ktorých sa používa anténová sústava a charakteristika antény je dynamicky nastavovaná podľa požiadaviek systému.



Obr. 2 Princíp inteligentnej anténovej sústavy

V inteligentných anténach sa u komunikačných systémov používajú lineárne rady anténových prvkov s malým ziskom, ktoré sú prepojené pomocou kombinačnej siete. Bloková schéma takéhoto systému je na obr. 2.

Signály indukované na každej vetve anténovej sústavy sú násobené komplexnou váhou. Jednotlivé váhy sú charakterizované amplitúdou a fázou. Následne sú skombinované a výsledný signál je privedený na výstup systému, kde je možné jeho ďalšie spracovanie.[1, 4]

2. TVAROVANIE VYŽAROVACEJ CHARAKTERISTIKY INTELIGENTNEJ ANTÉNOVEJ SÚSTAVY

Inteligentné anténové sústavy riadia smer vyžarovania, alebo príjmu automaticky tak, aby sa prispôsobili definovaným podmienkam v kanáli. To znamená, že musia zabezpečiť maximálny výkon signálu v žiaducom smere a minimálny výkon signálov z nežiaducich smerov.

Tieto sústavy dokážu nielen sledovať užitočný signál (resp. vysielať žiaducim smerom) a zabezpečiť maximum výstupného pomeru S/I, ale dokážu tiež dosiahnuť minimálnu hodnotu interferencie. Minimálna hodnota interferencie sa dosiahne tvorbou núl v smeroch interferenčných zdrojov.

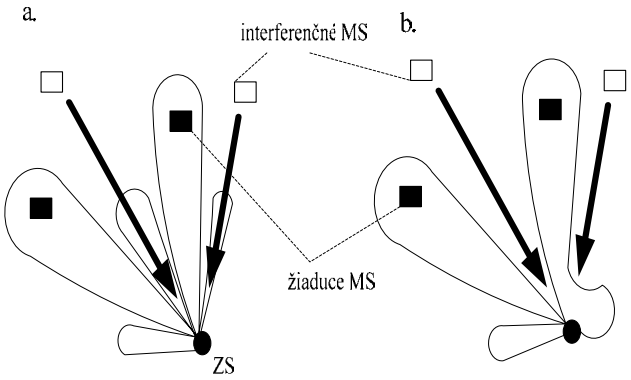
Tvarovanie vyžarovacieho diagramu slúži na zabezpečenie eliminácie interferencie, a teda je možné z tohto pohľadu metódy tvarovania rozdeliť do nasledovných skupín:

1. Tvarovanie lalokov (Beamsteering) – riadiace algoritmy snažiace sa nasmerovať lalok vyžarovacieho diagramu do žiaduceho smeru (obr. 3 a). Tento postup nie je schopný eliminovať interferenčné signály.

2. Tvarovanie núl – na rozdiel od predchádzajúcej metódy je možné okrem nastavenia vyžarovacích lalokov do požadovaného smeru nastaviť aj nuly do smeru interferenčných zdrojov (obr. 3 b). Výsledkom je značné zmenšenie interferencie a následné zväčšenie kapacity systému. [1]

2.1 Tvarovanie vyžarovacej charakteristiky dynamicky fázovanej anténovej sústavy

Celý proces tvarovania vyžarovacej charakteristiky anténovej sústavy spočíva v napájaní

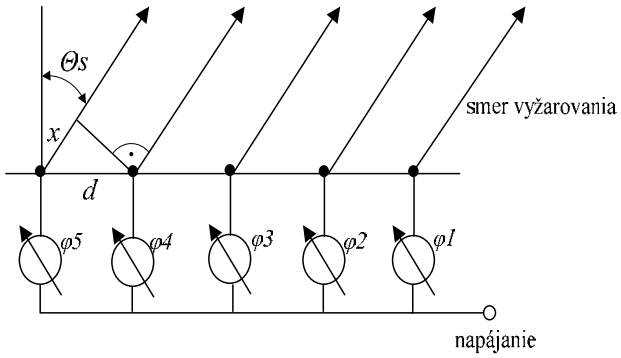


Obr. 3 Tvarovane lalokov (Beamsteering) (a), Tvarovanie núl (b)

jednotlivých elementov signálmi, ktoré sú vzájomne fázovo posunuté. Pri takomto napájaní dochádza ku konštruktívnemu a deštruktívnemu sčítaniu signálov. To v jednoduchosti znamená, že signály ktoré sú vo fáze, budú sčítané konštruktívne a signály sčítané v protifáze, sa vzájomne zrušia. Z toho vyplýva, že pri napájaní jednotlivých žiaričov anténovej sústavy s rozdielnou fázou je možné tvarovať výsledný vyžarovací lalok celej sústavy.

Fázový posuv vstupného signálu sa realizuje v napájači každého elementu anténovej sústavy samostatne. Prvok na túto funkciu určený, sa nazýva menič alebo posúvač fázy (Phase Shifter).

Na obr. 4 je principiálne zobrazená anténová sústava skladajúca sa z piatich elementárnych žiaričov.



Obr. 4 Principiálna schéma inteligentnej anténovej sústavy

Napájanie elementov je sériové, to znamená, že signál z vysielača prenášaný hlavným vedením je pomocou odbočiek distribuovaný do jednotlivých

antén. Fázový posuv tohto signálu sa realizuje samostatne na každej anténe.

Uhol Θs predstavuje odklonenie vyžarovacieho laloku od priameho smeru. Vzdialenosť d predstavuje vzájomný odstup jednotlivých elementárnych antén tvoriacich sústavu. Táto vzdialenosť závisí od vlnovej dĺžky spracovávaného signálu.

Ak platí, že:

$$x = d.\sin\Theta s \tag{1}$$

$$\frac{360°}{\varphi} = \frac{\lambda}{x} \tag{2}$$

Pomocou vzťahov (1), (2) vyplývajúcich z obrázka 4 je možné vytvoriť rovnicu (3) vyjadrujúcu vzťah medzi fázovým posuvom (φ) medzi jednotlivými anténami a uhlom natočenia vyžarovacieho laloka (Θs). [5]

$$\varphi = \frac{360°}{\lambda} d.\sin\Theta s \tag{3}$$

3. MIKROPÁSIKOVÝ MENIČ FÁZY

Menič fázy je prvok nachádzajúci sa v napájači každého elementu tvoriaceho anténovú sústavu. Jeho úlohou je meniť fázu vstupného signálu takým spôsobom, aby bolo možné nasmerovať vyžarovací lalok do potrebného smeru.

Jeho konštrukcia môže byť riešená mnohými spôsobmi. Ako najvhodnejšie riešenie bola zvolená konštrukcia meniča fázy s použitím odbočiek.

Menič fázy je realizovaný ako mikropásikové vedenie s odbočkami. Sú použité štyri odbočky, ktoré vytvárajú fázový posuv 22.5°, 45°, 90° a 180°.

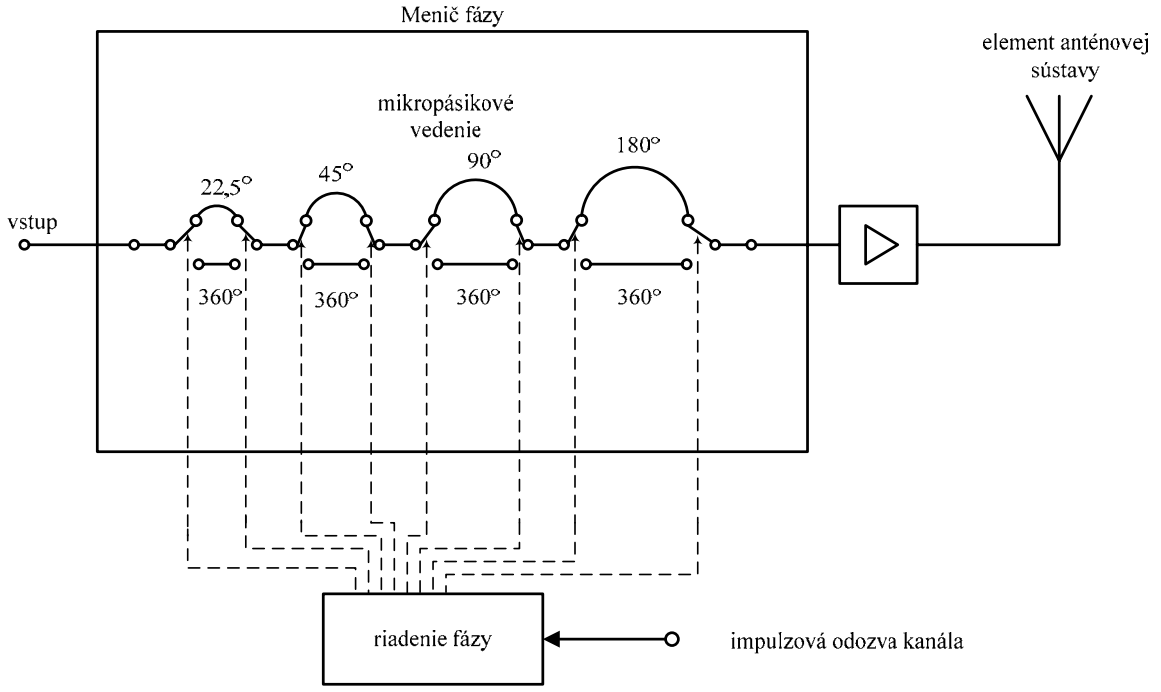
Zapojenie jednotlivých odbočiek je možné kombinovať. To znamená, že na každom elementárnom žiariči je možné vytvoriť fázový posuv od 22,5° do 337,5°.

Pri praktickej realizácii sa predpokladá, že zariadenie bude pracovať na frekvencii 1,95 GHz. Ako substrát je použitý materiál FR4 (skloepoxidový laminát s 35μm vrstvou medi a ε_r = 4,34).

Jednotlivé kombinácie odbočiek je možné voliť pomocou PIN diód, ktoré plnia funkciu prepínačov. Tieto diódy sú ovládané na základe požiadaviek na šírku, tvar a smer výsledného vyžarovacieho laloka sústavy.

Riadiaci algoritmus realizuje zmenu fázového posuvu signálu na jednotlivých elementárnych žiaričoch tvoriacich anténovú sústavu na základe informácií získaných z impulzovej odozvy kanála a z algoritmu zisťovania smeru príchodu signálu (Direction of Arrival – DOA). Spracovanie programu na riadenie fázového posuvu sa realizuje v programovateľnom mikroprocesore (napr. DSP).

Takto je možné hlavný vyžarovací lalok nasmerovať do potrebného smeru a vytvoriť nuly do smerov, v ktorých môže vzniknúť interferencia. Návrh meniča fázy pre jeden element anténovej sústavy je na obr. 5.



Obr. 5 Bloková schéma navrhovaného meniča fázy

Zoznam bibliografických odkazov

[1] WIESER, V.: Mobilné rádiové siete 2. Žilina: Edis, 2004, 297 s, ISBN 80-8070-345-0.
[2] ČEPEL, P., WIESER, V.: Trendy systémov inteligentných antén. Bratislava: Zborník 10. medzinárodnej konferencie COFAX Telekomunikácie 2004, 2004, ISBN 80-967019-6-7.
[3] COOPER, M., GOLDBURG, M.: Intelligent Antennas: Spatial Division Multiple Access. Annual Review of Communications, http://www.arraycomm.com.
[4] GODARA, L .:Smart Antennas. Londýn: CRC Press, 2004, 458 s, ISBN 8493-1206-X.
[5] WOLFF, Ch.: Radargrundlagen. Berlín: 2007, http://www.radartutorial.eu
[6] MAILLOUX, R .: Phased Array Antenna Handbook. Norwood: Artech House, 2005, 496 s, ISBN 1-58053-689-1.

Summary: A task of the article to outline a reader problems antenna systems and their control. The most important phase of control of antenna systems is to create and to form radiation pattern. There exist several elements and procedures, which are used at the control. One of possibilities is use the phase shifter. The designed phase shifter is appropriated (by its simplicity and price) for wide possibilities of use. By assistance large numbers of branches it is possible to make very accurate phase shift at the individual elements of antenna system. The automatic control of branches of converter of phase based on impulsion response of channel is meanwhile in the state of preparations and experiments.

Ing. Ján HARING
Ing. Norbert MAJER
Ing. Peter POLOHA,
doc. Ing. Rudolf HRONEC, PhD.
Katedra telekomunikácií, Elektrotechnická fakulta
Žilinská univerzita
Univerzitná 8215/1
010 26 Žilina
Slovenská republika
E-mail: haring@fel.utc.sk
majer@fel.utc.sk
poloha@fel.uniza.sk
hronec@fel.utc.sk

PREDIKČNÝ ALGORITMUS ŠÍRENIA ELEKTROMAGNETICKÝCH VĺN PODĽA ODPORÚČANÍ ITU-R VHODNÝ NA IMPLEMENTÁCIU V INFORMAČNOM SYSTÉME

RADIO WAVE PROPAGATION PREDICTION ALGORITHM BASED ON ITU-R RECOMMENDATIONS SUITABLE FOR IMPLEMENTATION IN INFORMATION SYSTEM

Marek HOVANEC, Martin MARKO

Abstract: An accurate prediction of the field strength and propagation loses is necessary for proper base stations deployment and a proper frequency planning along with meeting criteria of electromagnetic compatibility. The communication system establishment time is crucial especially in military area. Therefore, it is suitable to use computer applications for radio system projection to decrease the necessary deployment time. In the article the outcomes of dissertation thesis are briefly discussed. The primary aim of dissertation is to work out a propagation prediction algorithm based on ITU recommendations suitable for implementation in C2 information system. Some results of measurements carried out in Slovak republic are discussed. Accuracy, advantages, disadvantages and usability of existing models for a practical application in military information systems are evaluated.

Keywords: radio wave propagation, prediction methods.

ÚVOD

Vo vojenskej oblasti je významným a kritickým hľadiskom čas zriadenia komunikačného systému, kde je vo veľkej miere vhodné použiť podporné výpočtové aplikácie zjednodušujúce proces projektovania spojenia.

Článok stručne zoznamuje s výsledkami doktorandskej práce orientovanej na metódy zvýšenia efektívnosti rádiového spojenia s dôrazom na jeho projektovanie s využitím VISÚ a vybranými otázkami praktického využitia výsledkov práce. Ťažiskom práce je návrh predikčného algoritmu podľa odporúčaní ITU-R vhodného na implementáciu v informačnom systéme podpory velenia a riadenia (C2).

Algoritmus pre serverovú aplikáciu predikčného modelu je založený na procedúre odporúčaní ITU a je upravený pre podmienky SR resp. Európy.

1. PRISPÔSOBENIE PARAMETROV RÁDIOVÉHO SPOJA PREVÁDZKOVÝM PODMIENKAM

Snaha o čo najväčšiu efektívnosť rádiovej prevádzky - maximalizáciu prenosových rýchlostí, minimalizáciu energetických nárokov, zvýšenie odolnosti proti úmyselnému a neúmyselnému rušeniu, bezpečnosť, spoľahlivosť rádiového spoja, minimalizácia šírky frekvenčného pásma potrebného na prenos modulovaného signálu, elektromagnetická kompatibilita rádiových systémov a vplyv týchto ukazovateľov na cenu výstavby a prevádzku systému je nemysliteľný bez aplikácie jednej zo základných metód zvyšovania efektívnosti – metódy prispôsobenia rádiového spoja prevádzkovým podmienkam.

Ide o projektovanie rádiového spojenia, činnosť zameranú na optimalizáciu rozmiestnenia rádiových prostriedkov na dosiahnutie požadovaných parametrov spoja vo zvolenom frekvenčnom pásme, čase a priestore. Projektovanie rádiových spojov je založené na čo najpresnejšej predikcii šírenia rádiových vĺn a výpočte (odhade) úrovne signálu v bode príjmu. Hlavným cieľom je správnym výberom vysielacích a prijímacích stanovíšť, vybavených vhodným typom antén, dosiahnuť efektívne rádiové spojenie so zvoleným typom modulácie pri racionálne energeticky využitom spoji s ohľadom na vyžarovaný výkon vysielača a dodržaní zásad EMC.

Otázka predikcie zahŕňa širokú problematiku a potrebný aparát pre jej riešenie. Ide predovšetkým o vyšetrenie cesty signálu od vysielača k prijímaču, vyjadrenie útlmu signálu na trase šírenia, predikciu mnohocestného šírenia a v určitých prípadoch aj oneskorenia signálu vo vzťahu k chybovosti v digitálnych mobilných rádiových komunikáciách.

Správnosť výstupného riešenia algoritmu na projektovanie rádiového spoja je v najväčšej miere podmienené presnosťou popisu fyzikálnych javov, ku ktorým dochádza pri dopade rádiovej vlny na rozhranie dvoch prostredí, alebo na výrazné terénne nerovnosti. Ide predovšetkým o popis difrakcie, refrakcie, rozptylu, odrazu a lomu rádiových vĺn.

2. MODELY NA PREDIKCIU ŠÍRENIA RÁDIOVÝCH VĺN

Existujúce modely je možné charakterizovať ako množinu matematických výrazov, diagramov a algoritmov využívaných na reprezentáciu rádiových charakteristík v konkrétnom prostredí.

Všetky doteraz známe metódy sú založené na geografických bázach dát s rozličným rozlíšením a klasifikáciou. Je možné rozlíšiť tri základné prístupy k predikcii intenzity elektrického poľa:

- empirické modely založené na štatistických výsledkoch sérií meraní popisujúcich šírenie rádiových vĺn a s tým súvisiace straty pri konkrétnych typoch terénu,
- deterministické modely využívajúce matematické a geometrické metódy na určenie intenzity elektrického poľa,
- zmiešané modely.

Hlavnou výhodou empirických modelov je, že sú implicitne vzaté do úvahy všetky existujúce vplyvy prostredia bez nutnosti ich samostatného skúmania. Na druhej strane, presnosť takýchto modelov nezávisí iba od presnosti vykonaných meraní, ale taktiež od podobnosti analyzovaného prostredia a prostredia, v ktorom boli štatistické merania realizované.

Deterministické modely sú založené na aplikácii známych fyzikálnych princípov šírenia elektromagnetických vĺn a z tohto dôvodu môžu byť aplikované na rôzne prostredia bez zníženia presnosti. Na rozdiel od štatistických modelov nie sú založené na výsledkoch sérii meraní, ale skôr na detailných znalostiach skúmaného prostredia. V praxi je ich realizácia obmedzená obrovskými nárokmi na bázy dát geografických informačných systémov spojenými v mnohých prípadoch s náročným až nemožným zberom údajov pre dostatočne presné aplikácie. Algoritmy používané deterministickými modelmi sú obyčajne komplikované a dosahujú nízku výpočtovú efektívnosť. Vytvorené predikčné modely je možné rozdeliť na [1]:

- makrobunkové,
- mikrobunkové,
- pikobunkové.

Na riešenie problematiky predikcie intenzity elektrického poľa bolo vyvinuté mnoho metód a modelov, z ktorých najznámejšie sú:

- Okumurova Hatov model,
- COST 231 -Walfisch-Ikegamiho model,
- ITU (CCIR) metódy,
- Longley-Riceho model,
- Leeho model,
- empirické mikrobunkové modely,
- HCM model (Harmonised Calculation Method),
- Leeho mikrobunkový model,
- metódy „Ray Tracing“ a „Ray Launching“,
- modely založené na aplikácii neurónových sietí.

Ich detailný popis je nad rámec tohto materiálu a čitateľ ich nájde napr. v publikáciách [2], [3] a odporúčaniach ITU-R.

Správnosť výberu algoritmu na projektovanie rádiového spoja a výstupného riešenia je v najväčšej miere determinované presnosťou popisu fyzikálnych javov, ku ktorým dochádza pri dopade rádiovej vlny na rozhranie dvoch prostredí, alebo na výrazné terénne nerovnosti. Ide predovšetkým o popis difrakcie, refrakcie, rozptylu, odrazu a lomu rádiových vĺn. Doposiaľ bolo navrhnuté množstvo predikčných modelov a algoritmov. Pričom však nie je možné vybrať univerzálny model, resp. model s najvyššou presnosťou. Modely sa navzájom líšia prístupom k riešeniu, zložitosťou a presnosťou. Medzi modelmi existuje aj rozdiel v nárokoch na požadované množstvo a presnosť informácií o teréne, keďže na riešenie intenzity elektrickej zložky elektromagnetického poľa resp. strát je v niektorých prípadoch postačujúci parameter zvlnenia terénu a teda nie je potrebná explicitná informácia o podrobnom tvare a vlastnostiach terénu.

Modely využívajúce možnosti výpočtovej techniky sú v súčasnej dobe najčastejšie využívané na projektovanie rádiových spojov, pričom vo väčšine prípadov je predikčný algoritmus založený na kombinácii teoretických a empirických modelov. V týchto prípadoch je dôležitá podrobnosť bázových dát o teréne a jeho nadstavbe. Potom všeobecným kritériom presnosti je miera prispôsobenia predikčného algoritmu vlastnostiam terénu v oblasti záujmu, čo je špecifické aj pre ozbrojené sily.

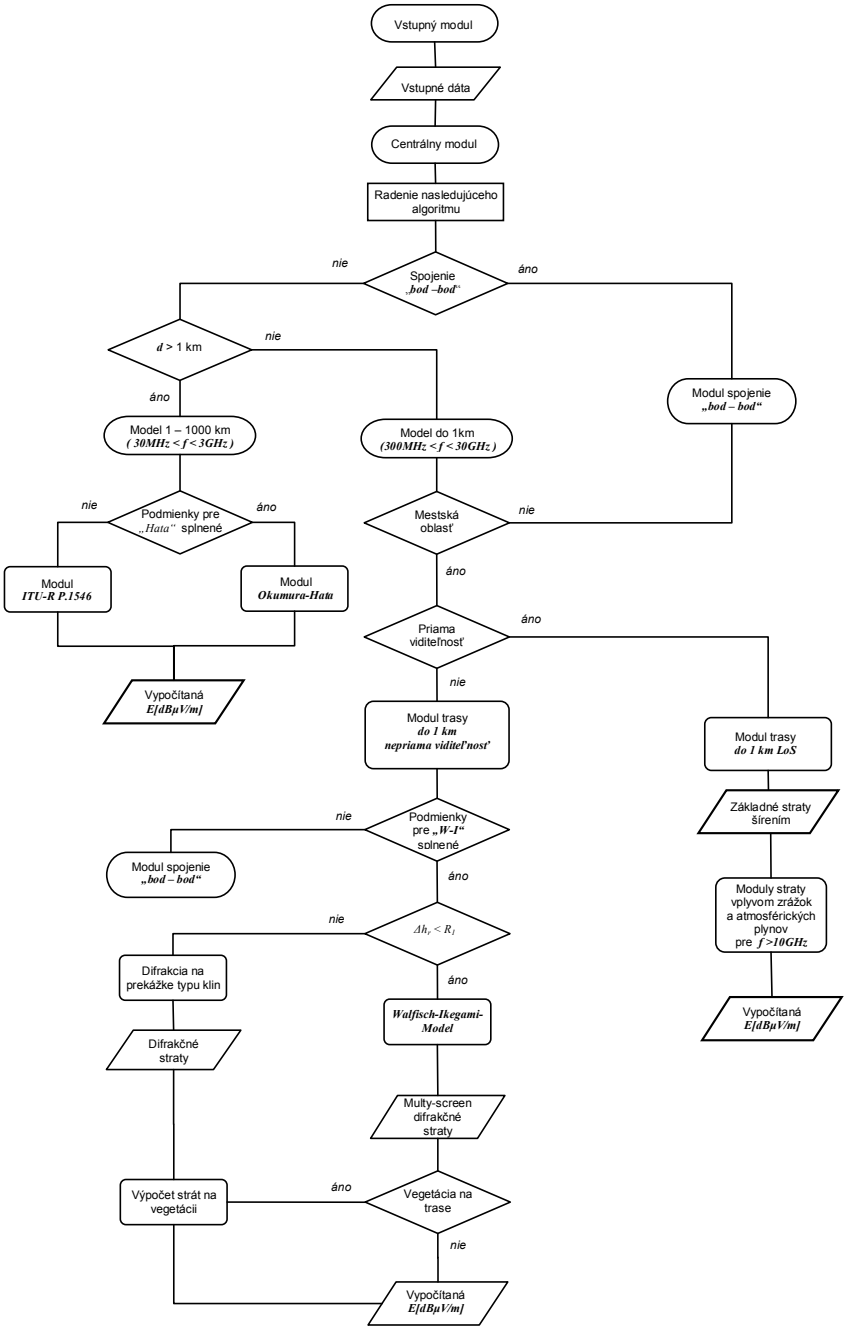
3. CIELE PRÁCE

Na základe podrobných analýz predikčných modelov s prihliadnutím na špecifiku a vývojové tendencie vo využívaní frekvenčných pásiem pridelených ozbrojeným silám i konkrétnu v implementáciu nových rádiokomunikačných systémov boli v práci vytýčené ciele:

1. Algoritmizácia serverovej úlohy na projektovanie rádiokomunikačných systémov s cieľom zvýšenia ich efektívnosti využitím modelov na predikciu elektrickej zložky elektromagnetického poľa.
2. Navrhnutý model a algoritmy modifikovať a optimalizovať pre serverovú úlohu s využitím VISÚ.
3. Spresniť požiadavky na Vojenský informačný systém o území Slovenskej republiky pre navrhnuté serverové aplikácie na projektovanie spojenia.
4. Verifikovať presnosť predikčných metód praktickými meraniami na území Slovenskej republiky.

4. ALGORITMIZÁCIA SERVEROVEJ ÚLOHY

Variabilita prostredia, v ktorom Ozbrojené sily SR pôsobia, siaha od husto osídlených oblastí – v prípade projektovania rezortných rádiokomunikačných sietí v miestach posádok, cez plánovanie spojenia pre poľné podmienky od nížinných rovinatých oblastí až po plánovanie spojenia v horských oblastiach. Dostupnosť projektových služieb širokej skupine užívateľov (veliteľov, náčelníkov zložiek), pri využití maximálního množstv a informácií o teréne a vysokej výpočtovej výkonnosti je možné zvýšiť realizáciou projektového algoritmu ako serverovej úlohy. s využitím vojenského informačného systému o území (VISÚ). Je potrebné si uvedomiť, že z dôvodu konečnej presnosti digitalizovaného modelu terénu (VISÚ) dochádza k nepresnostiam v stanovení intenzity elektrickej zložky elektromagnetického poľa. Je možné hovoriť o odchýlkach vplyvom horizontálnej alebo vertikálnej chyby terénnych tvarov a terénnych predmetov. K tomu je nutné ešte pripočítať samotnú nepresnosť v určení okamžitej polohy pri automatizovanom projektovaní spojenia. Uvedená analýza je súčasťou doktorandskej práce autora.



Obr. 1 Algoritmus pre serverovú aplikáciu predikčného modelu – 1. časť

Vstupom a výstupom počítačovej realizácie predikčného algoritmu musí byť čo najmenšia skupina údajov tak, aby ťažisko výpočtovej výkonnosti bolo položené na možnosti servera. Podľa ITU-R P.1144 sú vstupnými údajmi na výpočet intenzity elektrického poľa pomocou odporúčania ITU-R P.1546 geografické dáta o terénnych tvaroch a terénnych predmetoch, klasifikácia trasy, vzdialenosť, výška vysielacej antény, frekvencia, pomerná percentuálna miera času, výška prijímacej antény, TCA, percento oblasti.

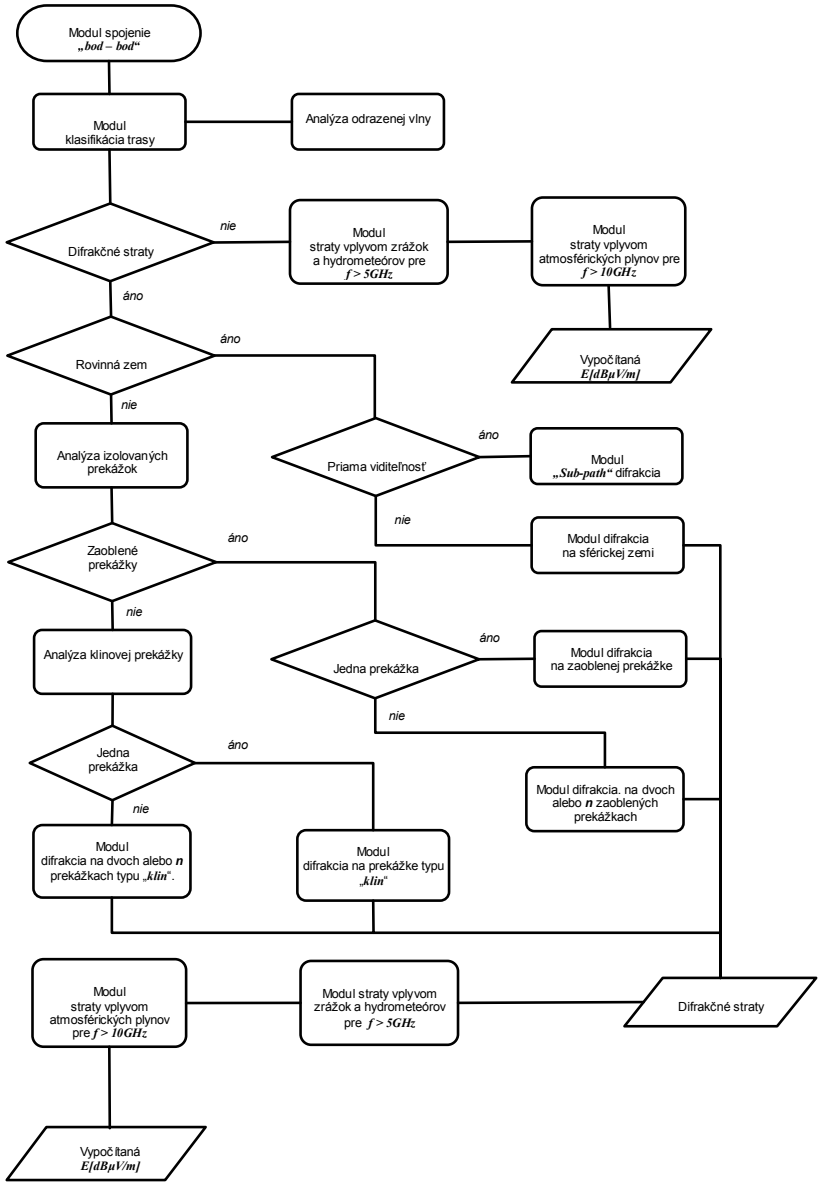
Hrubý algoritmus pre serverovú aplikáciu predikčného modelu je na obr. 1. a 2, kde sú uvedené väzby medzi jednotlivými modulmi.

Rešpektuje odporúčania ITU-R platné pre územie Slovenskej republiky a Európy.

5. VERIFIKÁCIA VÝSLEDKOV A MERANIA

Doktorandská práca v prílohách obsahuje množstvo meraní a výpočtov, ktoré boli uskutočnené v okolí Liptovského Mikuláša pri realizácii spojenia na mieste (bod - bod) i za pohybu pri definovaných rýchlostiach meracieho vozidla (MMP). Ako príklad sú tabuľke č. 1 zobrazené priemerné odchýlky od nameranej hodnoty pre každú z jednotlivých metód.

Je zrejmé, že predikované hodnoty podľa jednotlivých metód majú veľké rozdiely. Najnižšiu priemernú odchýlku dosiahla česká metóda RDK-2.



Obr. 2 Algoritmus pre serverovú aplikáciu predikčného modelu – 2. časť (modul spojenie „bod-bod“)

Približne o 3 dB vyššiu hodnotu má metóda ITU-R P.1546 pri zohľadnení korekčného faktora „Clearance angle„. Podľa uvedeného hodnotenia treťou v poradí je „klasická“ metóda ITU-R P.1546.

Hodnoty strednej odchýlky ostatných metód prevyšujú 20 $dB\mu Vm^{-1}$. Podľa tohto merania má Okumurova – Hatova metóda strednú chybu 24,04 $dB\mu Vm^{-1}$, avšak ňou predikované hodnoty v mestskom prostredí pre vzdialenosť približne 6 – 10 km dosiahli hodnoty veľmi blízke nameraným hodnotám. Výsledky dosiahnuté pomocou metódy HCM sú veľmi podobné resp. zhodné s výsledkami metódy ITU-R P.370.

Tab. 1 Namerané hodnoty

Metóda	Stredná chyba [dBµV/m]	Stredná kvadratická odchýlka [dBµV/m]
ITU-R P.1546	17,39	21,85
ITU-R P.1546 CA	13,13	18,33
Okumura-Hata	24,04	35,50
RDK - 2	10,58	13,52
ITU-370	26,60	30,51
ITU-370 CA	27,09	36,32
ITU-370 DH	26,63	30,52
ITU-370 CA DH	27,11	36,33
HCM	27,11	36,33

ZÁVER

Práca vytvára predpoklady pre efektívnejšie projektovanie rádiového spojenia v rámci systémov velenia a riadenia v ozbrojených silách s využitím vojenského informačného systému o území. Boli verifikované predikčné algoritmy, ktoré boli v prevažnej miere vytvorené na komerčné účely. Napriek svojmu širokému rozsahu však nemohla vyriešiť všetky otázky, predovšetkým v oblasti použitia predikčných algoritmov vo vojenských aplikáciách a ich verifikácii.

Predikčný algoritmus je navrhnutý modulárne, tak aby v prípade nutnosti editácie algoritmu ho bolo možné pozmeniť napr. doplniť alebo vykonať úpravu niektorej funkcionality, úpravu niektorého z predikčných resp. difrakčných modelov prípadne aj doplniť úplne nový predikčný model vo forme samostatného modulu.

Zoznam bibliografických odkazov

[1] NESKOVIC, A. ET AL.: Modern approaches in modeling of mobile radio systems propagation environment. Ieee communications surveys 2000, www.comsoc.org/pubs/surveys

[2] LEE, W.C.Y.: Mobile communications engeneering. 2nd ed. New York: McGraw Hill, 1998.

[3] LEBHERZ, M., WIESBECK, W., KRAN K,W.: A versatile wave propagation model for the VHF/UHF range considering three-dimensional terrain," Ieee trans. on antennas and propagation, vol. 40, no. 10, 1992

[4] OKUMURA, Y.: Field strength and its variability in VHF and UHF land -mobile services. Land-mobile communications engineering. New York: Ieee press, 1983

[5] HAR D., WATSON A. M., CHADNEY. G.: Comment on diffraction loss of rooftop-to-street in COST 231- Walfisch-Ikegami model. Ieee trans. vehic. tech., vol . 48, no. 5, sep. 1999.

Summary: The main aim of dissertation is to improve efficiency of radio systems projection in the area of C2 information systems. The fundamental presumption for projecting tasks is utilization of military geographic information systems.

The prediction algorithm is modular so that it could be edited and adjusted to any actual circumstances. Any prediction model can be easily integrated into prediction algorithm as an individual module and whichever module can be corrected separately without any undesirable impact on the other modules or prediction algorithms.

kpt. Ing. Marek HOVANEC[1)]
doc. Ing. Martin MARKO, CSc.[2)]
[1)] Stredisko riadenia a prevádzky komunikačných a informačných systémov
Trenčín
Slovenská republika
E-mail: marek.hovanec@mil.sk,
m.hovanec@centrum.sk
[2)] Katedra elektroniky
Akadémia ozbrojených síl generála M. R. Štefánika
Demänová 393
031 01 Liptovský Mikuláš
Slovenská republika
E-mail: marco@aoslm.sk

NETWORK ENTRY PROCEDURE IN WIMAX

Pavel MACH, Robert BESTAK

Abstract: The article analyses the newest trends in wireless networks with focus on a standard IEEE 802.16-2004 that is also known as WiMAX (fixed WiMAX). Both, point to multipoint and Mesh network topologies are considered and the main differences between them are described. Besides, a new network topology can be distinguished when introducing relay stations to the network. Thus, several new schemes of node association procedure may be specified. The article describes the standard association procedures together with the new schemes when considering relay stations.

Keywords: WiMAX, Network entry, PMP, Mesh, Relay.

1. INTRODUCTION

The WiMAX (Worldwide Interoperability for Microwave Access) is a broadband wireless technology based on IEEE 802.16 standard, namely on IEEE 802.16-2004 [1] that is intended for fixed scenarios. Beside the version IEEE 802.16-2004, a version IEEE 802.16e [2] also exists which enriches the former standard about mobility features. The WiMAX technology is designed to provide wireless last mile broadband access in the Metropolitan Area Network (MAN) delivering performance comparable to a cable system such as DSL (Digital Subscriber Line).

Two connections modes are supported by WiMAX (see Fig. 1): i) Point to Multipoint (PMP), which is mandatory, and Mesh, which is optional. The PMP mode represents a classical cellular network structure where Subscriber Stations (SSs) are directly connected to a Base Station (BS). In comparison with the PMP mode, the Mesh mode makes possible direct communication between individual SS and each node has a capability to serve as a relay station.

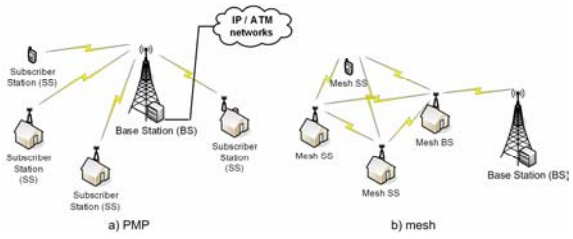


Fig. 1 Comparison of PMP and Mesh modes

Nowadays, the IEEE group is working on a new WiMAX standard labeled as IEEE 802.16j [3], [4] which introduced a new network node know as relay station (RS). The basic idea is to provide the date transmission via intermediate relay stations and thus extent range (a relay is placed at the cell edge) or increase throughput (a relay is placed within the cell radius). Integration of RSs into IEEE 802.16 is described in [5].

If a new SS wants to enter into the WiMAX network, an association procedure has to be proceeded. The existing WiMAX standard specifies individual network entry steps for PMP and Mesh network structure but not for the relay based structure. Due to the RSs, several new possibilities of network enter procedure can be introduced.

The rest of the paper is organized as follows. Section 2 describes the network entry procedure of SS in PMP mode and briefly characterize each phase. The next section focuses on the SS association in Mesh mode and basic differences compared to PMP mode are discussed. Section 4 describes the possible network entry scenarios for SS in case of relay based network architecture. Finally, the last section gives our conclusions.

2. SS NETWORK ENTRY IN PMP MODE

Fig. 2 depicts individual steps that are carried out during the network entry procedure. In the first phase, a SS scans for a downlink (DL) channel. Once the SS receives at the least one of the Downlink map message (DL-MAP), the synchronization process is accomplished. The SS remains synchronize with the DL channel as long as it receives the DL-MAP and Downlink Channel Descriptor (DCD) MAC messages that are periodically broadcasted by the BS. From the DCD message that characterize the DL channel and its burst profiles are obtained downlink channel parameters.

After successful synchronization with the DL channel, the SS waits for the Uplink Channel Descriptor (UCD) and the Uplink Map (UL-MAP) that are broadcast messages reciprocal to the DCD and the DL-MAP messages for the DL channel. The SS determines from the UCD whether it may use the uplink channel or not. If the uplink channel is suitable, the SS acquires additional parameters from the UCD and then waits for the UL-MAP. The UL-MAP message informs how the next uplink frame will be allocated. The UCD and UL-MAP messages are periodically sent by the BS. As long as the SS

receives the messages, it is assumed that the SS has valid uplink parameters.

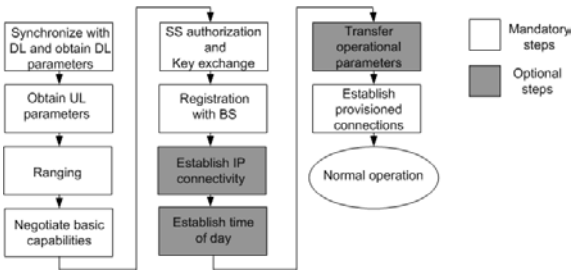


Fig. 2 SS initialization procedure in PMP mode

The next phase in the initialization procedure is an initial ranging. The Initial ranging procedure is the process during which the SS acquires correct transmission parameters such as time offset and transmitted power level together with Basic and Primary Management connection identifiers (CIDs). The SS scans the UL-MAP messages in order to find out an initial ranging interval which is allocated by the BS and which contains one or more transmission opportunities for SSs. When the Initial ranging transmission opportunity occurs, the SS sends the Ranging Request (RNG-REQ) message using the initial ranging CID.

After the BS successfully receives the RNG-REQ message, the SS responses with the RNG-RSP. This message assigns to the SS Basic and Primary management CIDs and further the message contains information about the power level adjustment, offset frequency adjustment and eventually any timing offset corrections. The RNG-RSP also includes information about ranging status that continues as long as the ranging status is not successful proceed or aborted by the BS.

Once successful ranged, the SS should inform the BS of its basic capabilities by transmitting the SS Basic Capability Request (SBC-REQ) message with its capabilities flag set to "on". The SBC-REQ message is sent via basic CID that is assigned to this SS in the RNG-RSP message. The SBC-REQ contains only those capability parameters that are necessary for effective communication between the SS and BS during the remain part of the initialization procedure. In response to the SBC-REQ message, the BS generates the SS Basic Capability Response (SBC-RSP) message through which informs the SS whether those capability parameters can be supported by the BS. The capability parameters that are not supported by the BS are set to "off" in the SBC-RSP message.

After the basic capabilities negotiation phase, a SS authorization and key exchange phase follow. This phase of initialization can be divided into two parts, i) SS authorization and Authorization Key (AK) exchange, and ii) Traffic Encryption Key (TEK) exchange (more detail about the authorization can be found in [1] or [2]).

If the SS is successfully authorized by the BS, it must be registered. The registration is a process through which the SS is allowed to enter into the network. The BS sends to the controlled SSs their Secondary Management CID. Once the BS receives the confirmation that a SS is controlled by the BS, the following steps are fulfilled: i) establishing of IP connectivity, ii) establishing of time and iii) transferring of operation parameters.

In order to obtain the IP address, the SS initiates DHCP procedure by sending a broadcast DHCP discover packet. After the SS MAC Address is checked, DHCP servers answer with DHCP offer packets. In the next step, the SS sends to the selected DHCP server the DHCP request packet. The IP connectivity is successful established upon receiving of DHCP response packet. If the SS has a configuration file containing further configuration parameters, the DHCP response contains its name. The establishment of IP connectivity is performed on the SS's Secondary Management Connection.

The time request and response are transferred via UDP datagrams. The current local time is created by combination of received time from the time server and the time offset received in DHCP response. The setting of time is performed on the SS's Secondary Management Connection.

Operation parameters are obtained from a configuration file that is downloaded via TFTP. The SS informs the BS about the successful download of the file via the TFTP-CPLT message that is sent on the SS's Primary Management Connection. The TFTP-CPLT is periodically transmitted until the TFTP-RSP message is received.

The final phase of network entry procedure is establishing of provisioned connections where service flows with specific QoS parameters (such as latency, jitter, throughput guarantee, etc.) are created. A service flow is a MAC transport service that provides unidirectional transport of packets either on the uplink or downlink direction. The service flow creation can be initiated by BS (mandatory capability) or by SS (optional capability).

3. SS NETWORK ENTRY IN MESH MODE

The network entry procedure in case of Mesh mode is partly different in comparison with the PMP mode (see Fig. 3). At the beginning of initialization, the mesh node listens to the network configuration messages called MSH-NCFG. These messages are periodically transmitted by all stations that are present in the network. The messages contain information about i) coarse synchronization to the network, ii) basic network parameters and iii) list of neighbors.

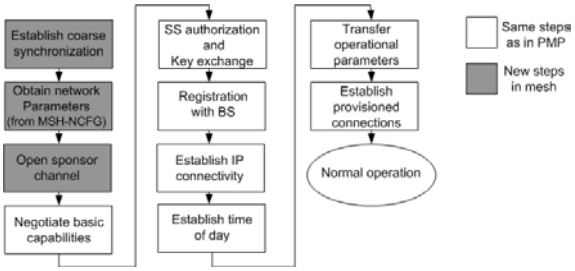


Fig. 3 SS initialization procedure in Mesh Mode

Subsequently, the SS selects a potential sponsoring node (from list of neighbors) and starts to the negotiation procedure by transmitting the MSH-NENT message of NetEntryRequest type [1]. The sponsoring node can either accept or refuse the request for sponsoring channel opening. The acceptance is signalized via MAC address advertisement in the MSH-NCFG message which include a field Net EntryOpen IE (Information Element). A sponsoring channel is immediately opened by the sponsoring node upon reception of the MSH-NENT message of NetEntryAck type. If the sponsoring node denies the access, the SS has to choose another adjacent node and the whole procedure has to be repeated.

Once the sponsoring channel is opened, the new SS performs similar operations as is described in the previous section (negotiation of basic capabilities, authorization, registration, etc.). The only difference is that the exchange of messages occurs between the new SS and sponsoring node (instead of between SS and BS as is in the PMP mode) and if necessary between the sponsoring node and BS. The network entry procedure is terminated by sending the MSH-NENT message of NetEntryClose type which is confirmed by the MSH-NENT message of NetEntryAck type.

4 SS NETWORK ENTRY IN RELAY BASED SCENARIO

The association process of SS can be divided into several scenarios, as shown in Fig. 4.

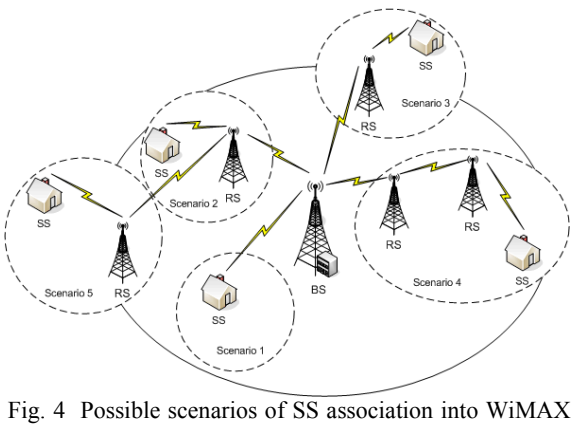


Fig. 4 Possible scenarios of SS association into WiMAX multihop relay network

The classification of these scenarios depends on several parameters such as the number of hops between the SS and BS or whether a SS is in the range of BS or not. To be standard compliant, none or only minimal modifications are required.

4.1 One hop scenario

In one hop scenario (scenario 1 in Fig. 4), a SS connects directly to the BS. The SS only receives messages broadcasted by the BS. Thus, the association procedure is the same as is discussed in section "Network entry in PMP mode", i.e. no special modifications have to be done.

4.2 Single RS scenarios

A SS connects to the BS via one intermediate RS. In this case, only one RS is allowed to be in the forwarding path. There are two possible association procedures according to scenario:

a) SS is in the range of BS – When the SS is in the range of BS (scenario 2 in Fig. 4), the SS receives MAC frame not only by BS but also by RS. This case corresponds to the situation when RSs are deployed in order to increase system capacity.

To satisfy the backward compatibility with legacy standards like IEEE 802.16-2004 or IEEE 802.16e, a RS has to behave towards the SS as a regular BS. This implies that decentrally controlled RS should broadcast its own control information (UL/DL MAP, DCD/UCD). In case of centrally controlled RS, a RS has to retransmit these messages from BS. Since the SS does not distinguish between BS and RS, the SS tries to associate to stations (i.e. RS) with better signal quality.

However, there are two examples when this decision doesn't have to be the most efficient way in terms of the end to end throughput:

- SS detects a MAC frame with a stronger signal from a RS than from the BS and therefore the SS associate to the RS. Nevertheless, the path to the BS through the RS involves two hops and the overall capacity may be lower one than in case of the direct link to the BS.
- SS detects a MAC frame with a stronger signal from the BS than from a RS and therefore, the SS establishes the link with the BS. Nevertheless, this network attachment decision is optimum for the DL data transmission and not necessarily for the UL direction.

The efficient way how to prevent these drawbacks is to implement a signaling and routing mechanism that determines to which stations (BS, RS) the SS should associate. Therefore, the overall network entry procedure is to be extended by another step (labeled as routing in Fig. 5) that takes place either before or after the establishment of

provisioned connections. The former possibility may prolong the whole entry procedure. On the other hand, the attachment to the most profitable node is reached on the first attempt and thus there is no need to cancel current connections and establish new ones.

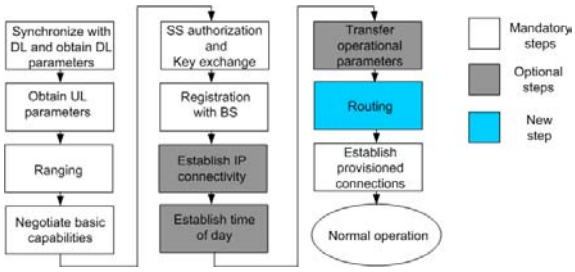


Fig. 5 Modified SS network procedure in relay based network architecture

b) SS is not in the range of BS – This scenario (scenario 3 in Fig. 4) corresponds to a situation where a RS is intended to extend the coverage area of the BS. Since the SS is out of BS range, no direct MAC frames can reach the SS and only messages re-broadcasted by the centrally controlled RS or sent by the decentrally controlled RS reaches the SS. Thus, there is no need to decide to which stations SS should associate like in the previous scenario. The following network entry procedure is the same as in the first scenario, but all messages have to be sent via the intermediate RS.

4.3 Multiple RSs scenarios

A SS is connected to the BS via more than one RS. As in the single RS scenario, the multiple RSs scenario may be divided into two cases:

a) SS is in the range of BS – Similar to scenario 2, a SS receives broadcasted messages from the BS and RS (scenario 4 in Fig. 4). The only difference is that the RS is not connected to BS directly but via one or more RSs. However, this scenario is not optimal in most cases since the number of hops between the BS and SS may be of a great value (more than two hops).

b) SS is not in the range of BS – The same scenario as the previous one but the SS doesn't receive MAC frames from the BS but from RSs (scenario 5 in Fig. 4).

ACKNOWLEDGMENT

This research work is supported by grant of Czech Ministry of Education, Youth and Sports No. MSM6840770014.

References

[1] IEEE: IEEE Std 802.16-2004, IEEE Standard for Local and metropolitan area networks, Part 16: Air Interface for Fixed BWA Systems, October 2004

[2] IEEE: IEEE Std 802.16e-2005, IEEE Standard for Local and metropolitan area networks, Part 16: Air Interface for Fixed and Mobile BWA Systems, Amendment for Physical and Medium Access Layers for Combined Fixed and Mobile Operation in Licensed Bands, February 2006.

[3] IEEE: IEEE Baseline Document 802.16j-2006, Air Interface for Fixed and Mobile Broadband Wireless Access Systems: Multihop Relay Specification, 2007.

[4] MACH, P., BESTAK, R.: Wireless Mesh and Relay Networks. In Research in Telecommunication Technology 2006 - Proceedings [CD-ROM]. Brno: Vysoké učení technické v Brně, 2006, vol. I, s. 450-454. ISBN 80-214-3243-8.

[5] HOYMANN, CH., KLAGGES, K.: MAC frame concepts to support multihop communication in IEEE 802.16 Networks.

Summary: The paper focuses on association procedures in WiMAX system based on IEEE 802.16 standards. Description of connection setup in PMP and Mesh modes are described. Network entry procedure in case of relay based network architecture is investigated and new possible scenarios for entering SSs into WiMAX network are determined together with the proposed modification to the existing SS association procedure.

Ing. Pavel MACH
Ing. Robert BESTAK, PhD.
Czech Technical University
Technicka 2
166 27 Prague 6
Czech Republic
E-mail: machp2@fel.cvut.cz
bestar1@fel.cvut.cz

VLASTNOSTI OPTICKÝCH PÁSEM 850 nm A 1550 nm Z POHLEDU JEJICH VYUŽITÍ PRO BEZKABELOVÉ OPTICKÉ SPOJE

PROPERTIES OF OPTICAL WAVELENGTHS 850 nm AND 1550 nm FROM VIEW OF THEIR USE FOR WIRELESS OPTICAL LINKS

Aleš PROKEŠ

Abstract: In the paper, the parameters of an optical wireless communication link which are most dependent on the optical wavelength are discussed. The paper deals with the comparison of the atmospheric attenuation at 850 nm and 1550 nm caused by scattering by the particles present in the atmosphere and with comparison of the optical receiver sensitivity in dependence on a bit rate for both wavelength. Presented calculations are demonstrated on the connection of avalanche and PIN photodiodes with a high-impedance amplifier using a MOSFET and with a transimpedance amplifier using a bipolar junction transistor. It is assumed that the silicon photodiodes work at a wavelength of 850 nm and the InGaAs photodiodes at 1550 nm.

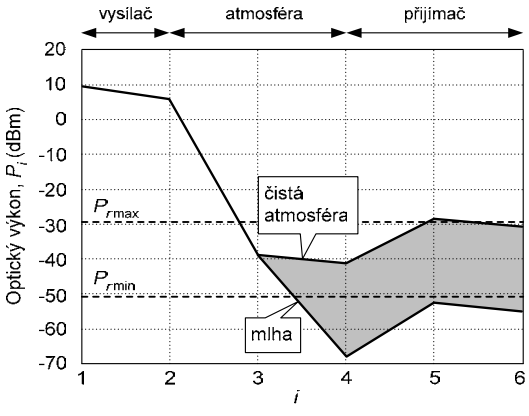
Keywords: Atmospheric attenuation, avalanche photodiode, PIN photodiode, meteorological visibility, optical receiver sensitivity.

1. ÚVOD

Bezkabelové optické komunikační systémy, nejčastěji označované zkratkou FSO (Free Space Optics), jsou vhodné pro přenos dat typu bod-bod na vzdálenost od stovek metrů do jednotek kilometrů při přenosových rychlostech do jednotek gigabitů. FSO systémy mají několik výhod vůči optovláknovým spojům: rychlá a snadná instalace, nízká cena (není třeba provádět výkopy přes komunikace), a také vůči radiovým komunikačním systémům: velká šířka pásma umožňující vysoké přenosové rychlosti, bezlicenční provoz (není třeba povolení pro využívání kmitočtového spektra), neexistující vzájemné interference několika spojů a obtížný odposlech. Nevýhodou FSO je omezená dostupnost spoje způsobená fluktuacemi útlumu atmosféry v důsledku sněžení, deště nebo mlhy. V pozemských aplikacích se bezkabelové optické spoje používají jako telekomunikační prostředky pro tzv. „poslední míli“ nebo jako spoje v síti LAN např. mezi dvěma budovami.

Ačkoli existuje několik vhodných optických vlnových délek pro přenosy v atmosféře [1], návrháři FSO obvykle používají optická spektrální okna v oblasti 850 nm nebo 1550 nm. Důvodem této volby je dobrá dostupnost laserových diod a fotodiod pro obě vlnové délky. Rozhodnutí které okno je výhodnější pro FSO je velmi obtížné, neboť existuje mnoho faktorů závislých na vlnové délce, které ovlivňují systémové parametry spoje jako například vysílaný optický výkon, útlum atmosféry, citlivost přijímače a pro většinu aplikací také cena použitých komponentů. Příklad výkonového diagramu optického spoje instalovaného na vzdálenost přibližně 1 km je na obr. 1.

Vysílaný optický výkon P_2 závisí na výkonu laseru P_1, na vazebních ztrátách mezi optickou soustavou vysílače a laserovou diodou a na útlumu v čočkách (ztráty ve skle). Ztráty způsobené šířením ve volném prostoru (P_2 - P_3) jsou funkcí vzdálenosti hlavic FSO a divergence svazku [2]. V atmosférickém kanálu je světlo tlumeno vlivem absorpce, rozptylu na částicích a vlivem turbulence ($P_3 – P_4$). Tyto jevy se mění v závislosti na povětrnostních podmínkách a je velmi těžké je předpovědět. Optický výkonový zisk ($P_4 – P_5$) na vstupu přijímače je dán poměrem efektivních ploch optických soustav vysílače a přijímače. Vazba mezi čočkou a fotodiodou, útlum čočky a odraz na čočce způsobují další přídavné ztráty v přijímači ($P_5 – P_6$). Pro správnou detekci signálu musí být zajištěno, aby výkon na aktivní ploše fotodiody byl uvnitř intervalu omezeného citlivostí přijímače $P_{r\,\min}$ a saturací přijímače $P_{r\,\max}$. Je zřejmé, že na obr. 1 není tato podmínka splněna pro mlhu.



Obr. 1 Příklad výkonového diagramu FSO

Určitá závislost útlumu nebo zisku na vlnové délce se dá najít ve všech zobrazených intervalech ale v mnoha případech lze dosáhnout stejných výsledků v obou optických pásmech volbou vhodných optických komponentů. Proto je dále věnována pozornost pouze útlumu atmosféry a citlivosti přijímače.

2. ÚTLUM ATMOSFÉRY

Útlum laserového svazku v atmosféře je popsán Beers-Lambertovým zákonem [3]

$$\tau(\lambda, L) = \frac{P(\lambda, L)}{P(0)} = \exp\left[-\gamma(\lambda)\alpha L\right], \quad (1)$$

kde $\tau(\lambda, L)$ je propustnost atmosféry ve vzdálenosti L od vysílače pracujícího na vlnové délce λ, $P(\lambda, L)$ je optický výkon ve vzdálenosti L, $P(0)$ je optický výkon zdroje, $\gamma(\lambda)$ je celkový koeficient útlumu atmosféry a pro konstantu α platí $\alpha = 1/10\log_{10}(e)$, jestliže $\gamma(\lambda)$ je vyjádřeno v decibelech na jednotku délky, nebo $\alpha = 1$, jestliže $\gamma(\lambda)$ je vyjádřeno v neperech na jednotku délky. Koeficient útlumu atmosféry je obecně tvořen absorpčním a rozptylovým členem, avšak v případě vlnových délek používaných spoji FSO je obvykle uvažován pouze rozptyl na částicích atmosféry [3]. Pak koeficient útlumu atmosféry (v dB/km) může být vyjádřen ve tvaru

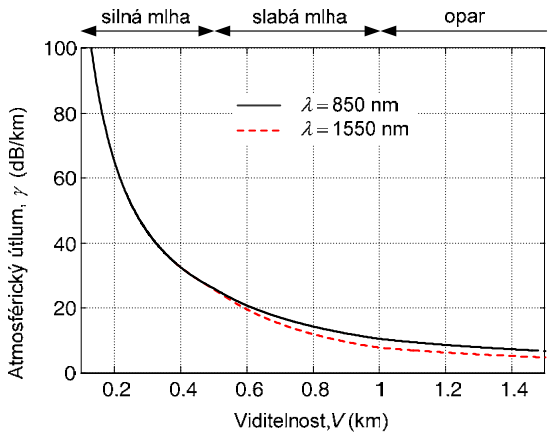
$$\gamma(\lambda) = \frac{13}{V}\left(\frac{\lambda}{550}\right)^{-q}, \quad (2)$$

kde V je meteorologická viditelnost (v km), λ je vlnová délka (v nm) a koeficient

$$q = \begin{cases} 0.16V + 0.34 & \text{pro} \quad 1km < V < 6km \\ V - 0.5 & \text{pro} \quad 0.5km < V < 1km \\ 0 & \text{pro} \quad V < 0.5km \end{cases} \quad (3)$$

je dán rozložením velikostí částic.

Ačkoli jsou známy i další vztahy popisující útlum atmosféry v mlze, dešti nebo při sněžení [4], rovnice (2) a (3) jsou používány poměrně často, protože odpovídají velmi dobře realitě, zvláště v případě mlhy, která je kritická pro činnost FSO. Grafické vyjádření (2) je ukázáno na obr. 2.

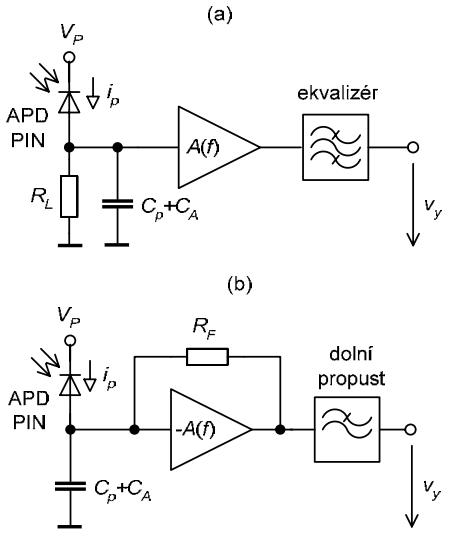


Obr. 2 Závislost útlumu atmosféry na meteorologické viditelnosti

Je vidět, že útlum atmosféry nezávisí na vlnové délce v silné mlze, která redukuje viditelnost pod 500 m. Ve slabé mlze a oparu je útlum na vlnové délce 850 nm nepatrně vyšší (asi 3 dB/km) než útlum na 1550 nm. Tento rozdíl však nemá podstatný význam, neboť optické spoje mají obvykle rezervu na vlivy počasí dovolující eliminovat i silnou mlhu a proto slabá mlha nemá na jejich činnost vliv.

3. CITLIVOST PŘIJÍMAČE

Primární faktor určující citlivost přijímače je úroveň šumu generovaného fotodiodou a následným zesilovačem. Nejčastěji používané zesilovače lze rozdělit do dvou skupin: vysokoimpedanční (VZ) a transimpedanční (TZ). VZ zesilovač zobrazený na obrázku 3(a) umožňuje dosáhnout vysokou citlivost, protože velký vstupní odpor R_L vykazuje nízký termální šumový proud. Avšak šířka pásma je omezena velkou časovou konstantou $\tau = (C_P + C_A)R_L$, kde C_P je kapacita fotodiody a C_A je vstupní kapacita zesilovače. Šířka pásma zesilovače je pak zvýšena na požadovanou hodnotu ekvalizérem. Zesilovač v TZ zapojení používá zpětnovazební odpor podél invertujícího zesilovače, jak je ukázáno na obrázku 3(b). Relativně nízká hodnota zpětnovazebního odporu R_F zvýší šířku pásma zesilovače, avšak zvýší i úroveň termálního šumu. Šumová šířka pásma pak může být redukována na minimální přípustnou hodnotu dolní propustí na výstupu zesilovače.



Obr. 3 Typická zapojení vstupních obvodů přijímače. Vysokoimpedanční zesilovač (a) a transimpedanční zesilovač (b)

Typický modulační formát pro nekoherentní detekci je klíčování optické nosné OOK (On-Off Keying). OOK detektor obvykle porovnává výstupní napětí zesilovače s určitým prahovým napětím. Jestliže nastavení optimálního prahového napětí je založeno na předpokladu že výstupní šum zesilovače má Gaussovské rozložení hustoty pravděpodobnosti a je použito symetrické zapojení

komparátoru [5], lze citlivost přijímače (minimální optický výkon přijímaného signálu pro danou pravděpodobnost vzniku chyby P_b) vyjádřit ve tvaru

$$P_{r1} = \frac{2Q}{R}\left[qF(M)QR_bI_2 + \left[I_{ni}^2/M^2 + 2qF(M)I_2R_b\left[2qF(M)Q^2R_bI_2 + I_{DM}\right]\right]^{1/2}\right], \quad (4)$$

kde Q je faktor daný inverzní distribuční funkcí pravděpodobnosti P_b, R je responsivita fotodiody, M je multiplikativní koeficient (zisk) lavinové fotodiody, dále označované zkratkou APD (Avalanche Photodiode), $q = 1.6\cdot10^{-19}$ C je náboj elektronu, I_{DM} je proud za tmy podléhající multiplikativnímu procesu, $F(M)$ je činitel přídavného šumu APD závisející na multiplikativním koeficientu a ionizačním poměru k_i tak, že platí

$$F(M) = k_iM + (1-k_i)(2-1/M), \quad (5)$$

I_2 je váhovací funkce, která závisí na tvaru vstupního pulsu a pulsu na výstupu fotodiody resp. zesilovače a R_b je rychlost přenosu (v bit/s).

Výpočet efektivní hodnoty šumového proudu vztaženého ke vstupu zesilovače I_{ni} je uveden v mnoha publikacích jako např. v [6], [7] nebo [8]. Pro zesilovač s bipolárním tranzistorem na vstupu platí

$$I_{ni} = \left[\frac{4kT}{R_1}I_2R_b + 2q\frac{I_C}{\beta}I_2R_b + 4kTr_b(2\pi C_P)^2I_3R_b^3 + \frac{2qV_T^2}{I_C}[2\pi(C_A+C_P)]^2I_3R_b^3\right]^{1/2}, \quad (6)$$

kde R_1 je zpětnovazební odpor v TZ zapojení nebo zatěžovací odpor fotodiody ve VZ zapojení, β je proudový zisk, I_3 je váhovací funkce (podobná I_2), r_b je odpor báze tranzistoru, $V_T = kT/q$ je teplotní napětí, C_A je složena z kapacit malosignálového π-modelu transistoru [6] C_π a C_μ a C_P zahrnuje kapacitu fotodiody a parazitní kapacity spojů na vstupu zesilovače.

Podobně efektivní hodnota šumového proudu vztaženého ke vstupu zesilovače s tranzistorem FET je dána vztahem

$$I_{ni} = \left[\frac{4kT}{R_1}I_2R_b + 2qI_LI_2R_b + 4kT\frac{\Gamma}{g_m}[2\pi(C_A+C_P)]^2\left(1+\frac{f_c}{R_b}\frac{I_f}{I_3}\right)I_3R_b^3\right]^{1/2}, \quad (7)$$

kde I_L je proud sestávající z proudu hradla tranzistoru FET I_g a nenásobeného proudu za tmy I_{DN} fotodiody, g_m je transkonduktance FET, f_c je lomová frekvence l/f šumu, Γ je numerická konstanta daná technologií výroby tranzistoru FET a I_f je váhovací funkce. Kapacita C_A je složena z kapacit hradlo-emitor C_{gs} a hradlo-kolektor C_{gd} tranzistoru. Citlivosti vypočítané podle (4) pro transimpedanční zapojení s bipolárním tranzistorem (TZB) a pro vysokoimpedanční zapojení s tranzistorem MOSFET (VZF) jsou porovnány na obrázcích 4(a) a 4(b). Váhovací funkce $I_2 = 1.05$, $I_3 = 0.52$ a $I_f = 0.61$ byly zvoleny pro NRZ kód filtrovaný Butterworthovým filtrem třetího řádu [8]. Předpokládá se že zpětnovazební odpor v TZB zapojení se mění nepřímo úměrně s přenosovou rychlostí. Součin R_1R_b je tedy konstantní a jeho hodnota byla zvolena 75 kΩ·Mbits/s. Ostatní hodnoty použité pro výpočet jsou: $Q = 6$ ($P_b = 10^{-9}$), $\beta = 100$, $r_b = 20\ \Omega$, $C_\pi = 1.0$ pF, $C_\mu = 0.05$ pF a $I_C = 0.4$ mA. Zatěžovací odpor fotodiody u VZF zapojení je 500 kΩ. Pro MOSFET byly zvoleny následující hodnoty: $I_g = 0.01$ nA, $C_{gs} = 0.4$ pF, $C_{gd} = 0.05$ pF, $\Gamma = 1.1$, $f_c = 10$ MHz, a $g_m = 30$ mS. Parametry fotodiod použité pro výpočet jsou shrnuty tab. 1. Výpočet vychází z předpokladu, že Si fotodiody pracují na vlnové délce 850 nm, zatímco InGaAs fotodiody jsou použity pro 1550 nm.

Tab. 1 Parametry fotodiod použité pro výpočet citlivosti přijímače

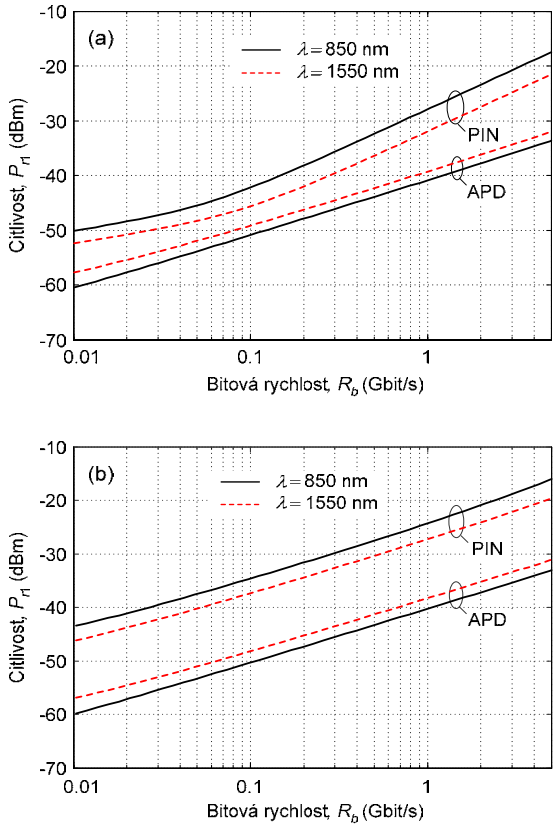
	Si PIN	Si APD	InGaAs PIN	InGaAs APD
C_P [pF]	1.0	1.0	0.5	0.5
I_{DM} [nA]	0	0.01	0	1.5
I_{DN} [nA]	1	0.5	2.5	2
M [-]	1	150	1	30
k_i [-]	0	0.02	0	0.4
R [A/W]	0.55	0.55	0.9	0.9

Z obrázků 4(a) a 4(b) je zřejmé, že APD umožňují zvýšení citlivosti přibližně o 10-17 dB vůči PIN fotodiodám. Přijímač s PIN fotodiodou vykazuje vyšší citlivost na 1550 nm díky větší responsivitě InGaA s fotodiody vzhledem k Si fotodiodě. Přijímače s APD jsou citlivější na 850 nm, protože Si APD ve srovnání s InGaAs APD mají nižší činitel přídavného šumu, vyšší zisk a generují nižší proud za tmy, který je zdrojem výstřelového šumu. VZF zapojení je vhodnější, jestliže je použita PIN fotodioda, protože dosahuje vyšší citlivosti ve srovnání s TZB zapojením. Rozdíl je znatelný obzvláště na nižších přenosových rychlostech.

4. ZÁVĚR

Pro většinu aplikací je požadována komunikační dostupnost spoje větší než 99 %. Díky povětrnostním podmínkám v mnohých lokalitách po celém světě tento požadavek způsobuje, že FSO musí být navrženy pro spolehlivou činnost v mlze, kde je útlum svazku nezávislý na vlnové délce použitého světla. Proto je nižší útlum atmosféry ve slabé mlze a oparu na 1550 nm (okolo 3 dB/km) ve srovnání s 850 nm nepodstatný. Dostupnost spoje může být částečně zlepšena volbou vhodné fotodiody. Přijímač s PIN fotodiodou je vhodnější pro 1550 nm, zatímco přijímač s APD by měl být preferován pro pásmo v okolí 850 nm. Rozdíl v citlivostech pro obě vlnové délky je 2-5 dB v závislosti na přenosové rychlosti.

Je zřejmé, že obě vlnové délky nabízejí srovnatelné vlastnosti spoje a volba jedné z nich není jednoznačná. Rozhodující úlohu proto mohou hrát ostatní kriteria jako například mezní limit výkonu laseru zajišťující bezpečnost zraku, který je asi padesátkrát vyšší pro 1550 nm vůči 850 nm, dále cena optických a elektrooptických komponentů nebo dostupnost optických měřících přístrojů.



Obr. 4 Citlivost optického přijímače v zapojení VZF (a) a v zapojení TZB (b)

PODĚKOVÁNÍ

Práce byla podporována výzkumným programem MSM0021630513 *Elektronické komunikační systémy a technologie nových generací* a grantem GAČR 102/06/1358 *Metodika návrhu optických bezkabelových spojů s vysokou spolehlivostí*.

Seznam bibliografických odkazů

[1] MANOR, H., ARNON, S.: Performance of an optical wireless communication system as a function of wavelength. *Applied Optics*, Jul. 2003, Vol. 42, No. 21, p. 4285-4294.

[2] KOLKA, Z., WILFERT, O.: Statistical model of free-space optical data link. In: *Proc. of The International Symposium on Optical Science and Technology*, Denver, 2004, p. 203 - 213.

[3] KIM, I. I., MCARTHUR, B., KOREVAAR, E.: Comparison of laser beam propagation at 785 nm and 1550 nm in fog and haze for optical wireless communications. In: *Proc. of SPIE - Vol. 4214 Optical Wireless Communications III*, Feb. 2001, p. 26-37.

[4] Al NABOULSI, M. SIZUN, H., de FORNEL, F.: Propagation of optical and infrared waves in the atmosphere. In: *Proc. of the XXVIIIth URSI General Assembly*, New Delhi, Oct. 2005, [CD-ROM].

[5] PROKES, A., WILFERT, O.: Analysis and Comparison of Various Free-Space Optical Receiver Configurations. In: *Proc. of the Conference Security and Defence, Advanced Free-Space Optical Communication Techniques and Applications III*. Stockholm: SPIE, 2006, p. 6399-D1 – D10.

[6] MUOI, T. V: Receiver design for optical-fiber systéme. *Journal of Lightwave Technology*, Vol. 2, Apr. 1984, p. 243-265

[7] PERSONICK, S. D.: Receiver Design for Digital Fiber Optic Communication Systems, Part I and II. *Bell System Technical Journal*. July – August 1973, vol. 52, no. 6, p. 843-886.

[8] ALEXANDER, B. S.: *Optical Communication Receiver Design*. Washington: SPIE Optical Engineering Press, 1997.

Summary: For most applications, the requirement of the communication link availability is greater than 99 %. Due to weather conditions in many localities in the world this demand causes that the FSO has to be designed for reliable operation in fog, where the attenuation of the optical beam is independent of the wavelength. Hence the lower attenuation in haze (about 3 dB/km) at 1550 nm in comparison with 850 nm is unimportant. Link availability can be slightly improved by the choice of proper photodiode. The PIN photodiode based receiver is more

appropriate for the 1550 nm wavelength, while APD based receiver is preferable in the 850 nm wavelength. The difference in sensitivity for both optical windows is 3-5 dB in dependence on the data rate.

doc. Ing. Aleš PROKEŠ, Ph.D.
Vysoké učení technické v Brně, FEKT UREL
Purkyňova 118
612 00 Brno
E-mail: prokes@feec.vutbr.cz

ROZVOJ MODERNÍCH RÁDIOVÝCH SYSTÉMŮ S KMITOČTOVÝM SKÁKÁNÍM; INTEGRACE RÁDIOVÝCH SÍTÍ A ELEKTRONICKÉHO BOJE V RÁMCI FH SYSTÉMŮ

DEVELOPMENT OF MODERN RADIO FREQUENCY HOPPING SYSTEMS; INTEGRATION OF RADIO NETWORKS AND ELECTRONIC WARFARE OF FH SYSTEMS

Andrej LÚČ, Juraj HRABOVSKÝ, Michal HALUZA

Abstract: The contribution is oriented to a development of FH telecommunication technology in the tactical radio systems. The content of contribution is based on Electronic Warfare (EW) requirements with a focus on modern military radio stations. Weapon systems call for reconstruction of communication systems. The content of contribution treats of the creation of radio networks. In the second part the contribution describes the jamming of FH radio systems. The analysis of FH modern communication and jamming systems are in the frequency and the time domain.

Keywords: Electronic Warfare, Recognition, Jamming, Frequency Hopping System, Pseudo-Random Sequence, Correlation, System Gain.

1. POŽADAVKY NA ROZVOJ RÁDIOVÝCH SÍTÍ A JEJICH INTEGRACE

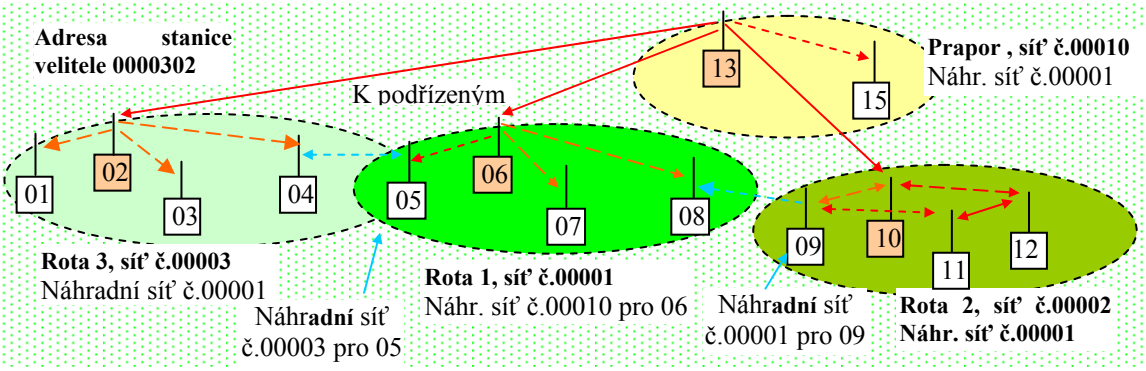
1.1 Základní požadavky na rozvoj rádiových systémů

Vojenský taktický rádiový systém musí zabezpečit **adresný, integrovaný a utajený** rádiový přenos pro mobilní účastníky. Současně by měl být odolný **proti rádiovému průzkumu** (nevyzrazovat vlastní činnost rádiem) a hlavně by měl být odolný proti **úmyslnému rušení.** Rádiový systém by neměl vyzrazovat vlastní činnost rádiovým signálem. Z uvedeného vyplývají **základní požadavky** pro **rozvoj** vojenských rádiových systémů.

1.2 Rozvoj taktických rádiových sítí

Vysoká mobilita a značný rozvoj vysoce účinných zbraňových systémů si vyžádal rychlý rozvoj rádiových sítí. Vývoj rádiových sítí vychází z uvedených požadavků a byl proveden ve třech hlavních směrech:

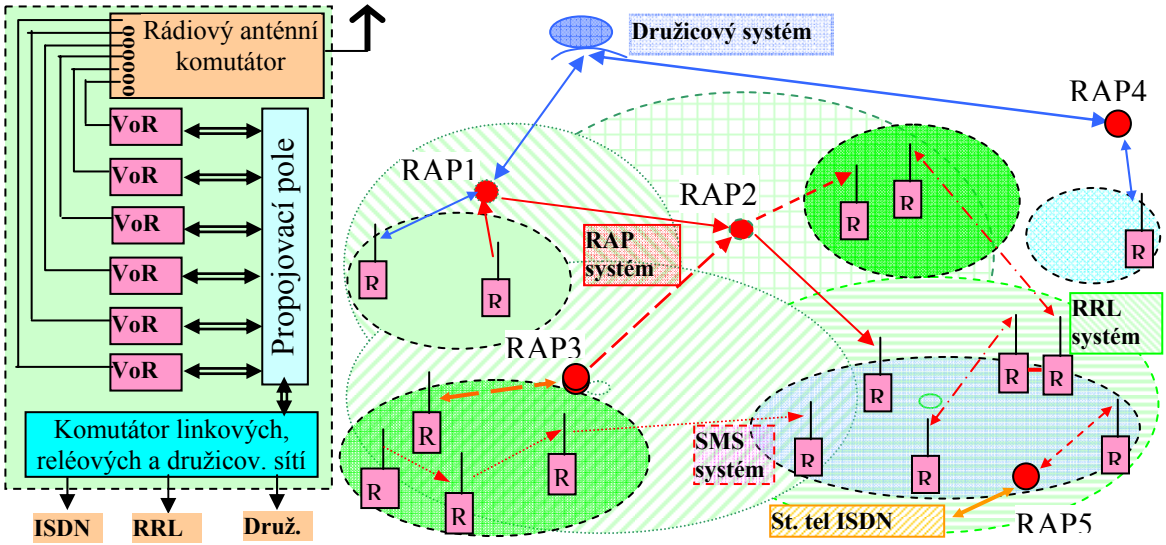
a. Rádiová síť musí být **adresná**, aby bylo možno se dovolat komukoliv ve vlastních a sousedních rádiem dosažitelných sítích, viz obr. 1. Současně aby rádiový systém umožnil integraci v bojových podmínkách spojit několik sítí do jedné se společnou, např. nadřízenou síti, viz obr. 1. Je zřejmé, že i opačný postup je možný.
b. **Integrovaný a mobilní** rádiový systém musí zabezpečit přenos informace mezi mobilními účastníky i mimo dosah rádiových stanic. Toto je dosaženo zavedením rádiových přístupových bodů (RAP – Radio Access Point) do rádiového systému, viz obr. 2. Dostáváme tak **integrovaný** systém, schopný vyhledat a přenést informace přes RAP do vzdálených rádiových sítí, dále do radioreléového systému, standardního telefonního systému a také do družicového systému. Přenos zpráv musí být zabezpečen i za přesunu. Je tak vytvořen **integrovaný** a **mobilní** rádiový systém.



Obr. 1 Využití rádiových sítí k částečné integraci spojovacího systému (v dosahu rádiových stanic)

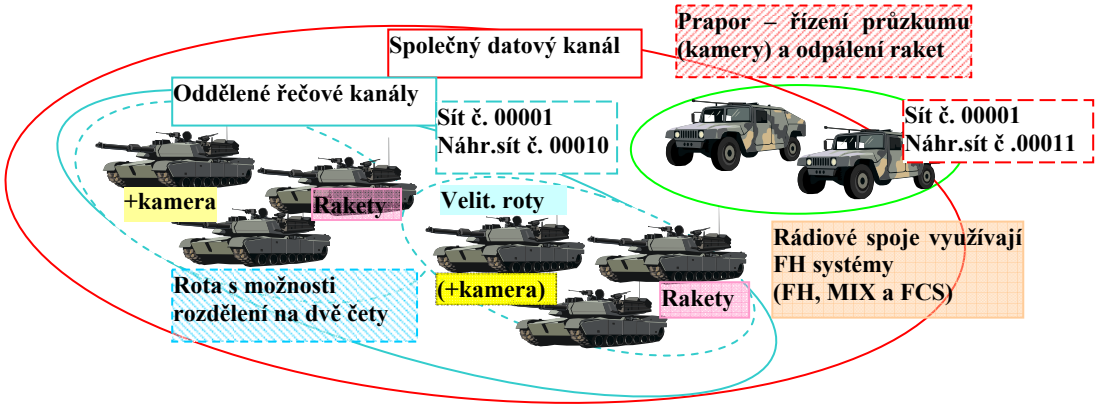
c. Rozvoj digitální a programovací techniky umožnil provést **digitalizaci** celého **bojiště**, umožnil přenášet velký počet informací v reálném čase. **Programovací technika** umožnila vysokou automatizaci řízení, zavést vysokou schopnost při řešení všech problémů.

Z těchto důvodů se značně zvýšil požadavek na přenos dat a to na všech úrovních. Např. průzkumník s kamerou potřebuje přenášet obrazový signál a současně stručný komentář. Jeho rádiová stanice potřebuje přenášet řeč i datový signál současně, viz obr. 3.



Obr. 2 Konfigurace 5-ti cestného RAPu. Vytváření spojů a jejich integrace: přes systém RAP; družicový systém; standardní systém; ISDN; radioreléový systém.

Plně nezávislé oddělení datového signálu od hlasového přinese lepší využití společného FH kanálu. V rotě se tak ušetří polovina rádiových prostředků. Data z kamery se přenášejí na velení praporu. Velitel roty může také sledovat průzkumný kanál. Při větším počtu kamer je možno postupně přepínat jednotlivé kamery tak, aby bylo sledováno více důležitých míst. Hlasová část kanálu je využívaná pro velení roty. Oba signály, hlasový a datový, jsou přenášeny nezávisle na sobě. Synchronizace kanálu musí být zachována.



Obr. 3 Velitel roty řídí tankový útok proti nepříteli hlasem a velitel praporu řídí průzkum a odpaluje rakety na klíčové body protivníka. Obě uvedené operace jsou prováděny na společném kanálu. Velení tankové roty je provedeno hlasem a průzkum a odpalování raket je provedeno na stejném kanálu z pozice praporu.

1.3 Rozvoj dalších možností taktických rádiových FH systémů

Digitalizace rádiových FH systémů umožnila doplnit rádiové stanice o další funkce, které umožňují zkvalitnit a rozšířit možnosti rádiových taktických stanic, jako jsou:

- Automatická změna módu stanice – přechod z FH na FCS (volba volného kmitočtu) podle typu rušení.
- Priorita velitele ve vlastní rádiové síti zabezpečí veliteli kdykoliv velet (vypne vysílání podřízených)
- Vytvoření možnosti spojit standardní jednokanálovou stanici FM se stanicí v režimu FH – kanálová výzva.
- Přenos důležité zprávy, např.chemický poplach a pod., při silném rušení. Je to provedeno s vysokým zabezpečením přenosu.
- Kontrola protistrany; jestli je správná osoba na druhé straně spoje.
- Spojení s jednotlivcem, nebo skupinou v síti (spojení s několika adresami společně.
- Vytvoření radioreléové stanice VKV-VKV, nebo VKV-KV.

K lepšímu pochopení diskutovaných FH systémů je na obr. 4a nakreslené bolkové schéma FH systému tak, aby bylo ukázáno schéma možného zapojení a činnosti stanice podle výše uvedených požadavků. Jsou zde stručně naznačené cesty zpracování řečového i datového signálu. Zabezpečený přenos datového signálu je proveden řetězením Read-Salamonovým a konvolučním kódem s prokládáním. Uvedené prokládání je ukázané na obr. 4b. Toto prokládání na obr. 4 je provedeno tak, aby hop, který skočí např. na plně zarušený kanál (např. na vlastní blízkou stanici, i přesto musí být opraven.

1.4 Stručné srovnání nabízených rádiových stanic FH

Realizace a kvalita integrovaných sítí závisí hlavně na možnostech použitých rádiových prvků (rádiových stanic). V současné době nejpoužívanější (nejprodávanější) jsou rádiové systémy od firmy THALES (PR4G F@stnet, HF 3000 SKYF@ST) a HARRIS (FALCON 2 (3)). O stanici od firmy Rohde&Schwarz M3T jsem nezískal žádné novinky. Pro vytvoření částečně ukrytého spoje, i přes jeho malé kmitočtové rozprostření , je možno použít systém KONSBERG (MR200).

Základní nevýhodou systému FALKON 2 pro jeho realizaci v integrovaném systému je, že jeho přenos v základním režimu není adresný. Stanice nemá svou adresu. Druhá základní nevýhoda systému FALCON 2 pro realizaci integrovaného systému je skutečnost, že není realizována priorita velitele.

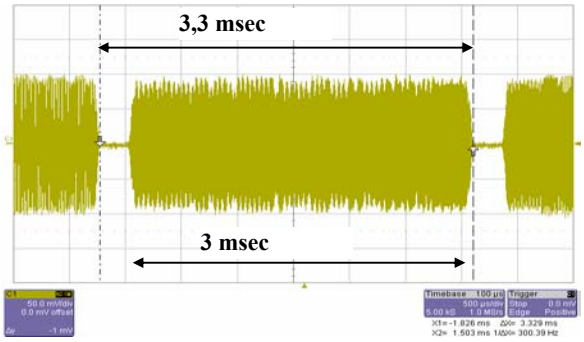
Systémy PR4G, nebo HF 3000 SKYF@ST obsahují a respektují svoji adresu. Systémy od firmy THALES využívají adresy sítě a stanice. Jsou schopny realizovat výše uvedené požadavky. V případě, že je volána jen jedna konkrétní stanice, signál je přenesen jen na danou adresu stanice, která je dána adresou sítě a adresou stanice v síti. Z výše uvedených skutečností můžeme rádiové systémy od fy THALES pokládat za systémy schopné vytvářet integrované digitální systémy. Také jejich odolnost vůči rušení je na velmi vysoké úrovni. Uvedené systémy od fy THALES mají zabezpečený přenos přes stanici RAP a dokáží přenášet také krátké zprávy pomocí stanic PR4G.

2.RUŠENÍ FH SYSTÉMU

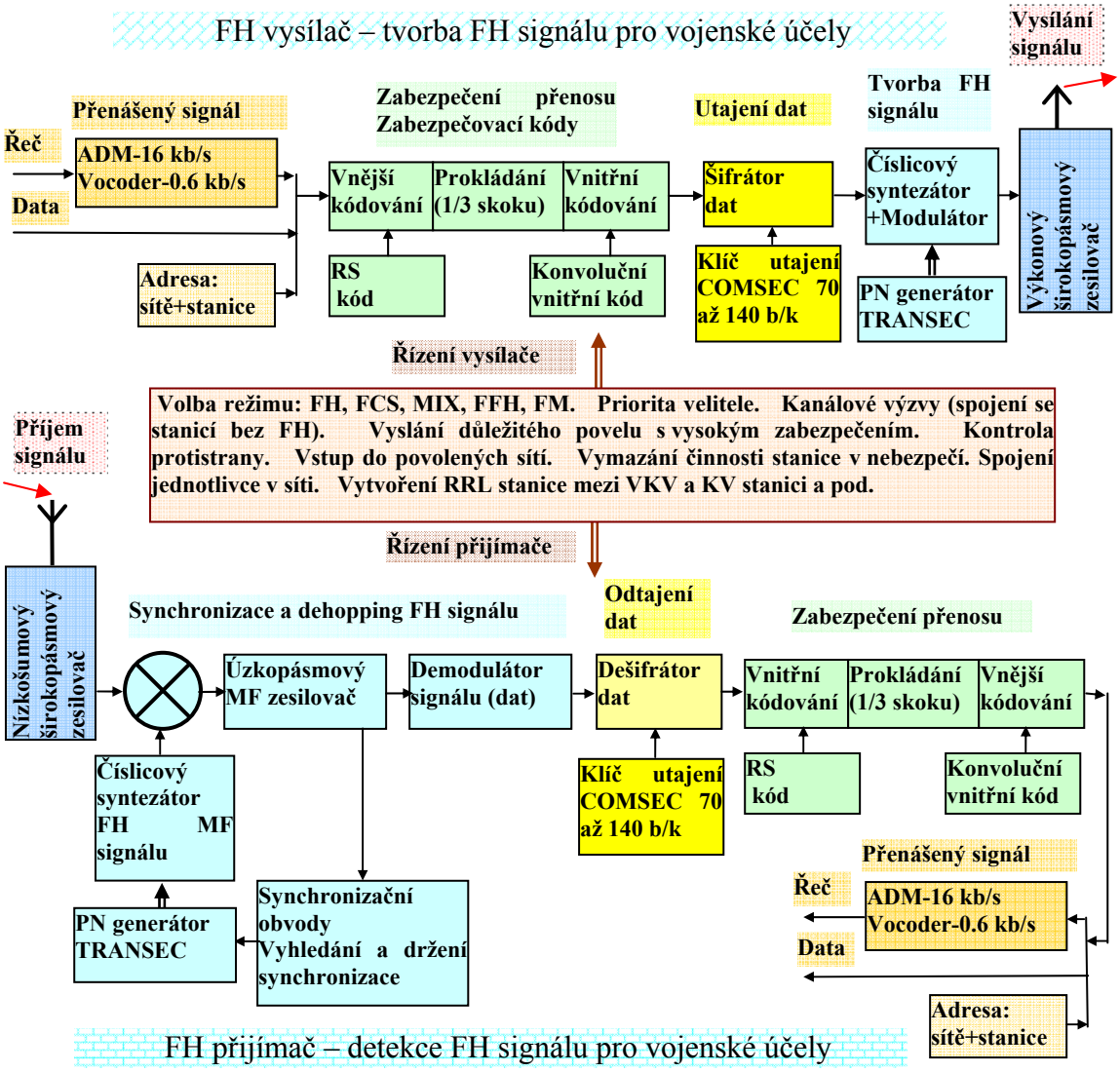
2.1 Průzkum a úvod k rušení FH systému

Rychlost kmitočtového skákání v moderních FH systémech je řádově několik stovek skoků za sec. Pro stručnost budeme v uvedených rozborech uvažovat nejrozšířenější FH systémy: THALES – PR4G F@STNET, 300 skoků/s; HARRIS – FALCON 2 (RF5800), 300skoků/s; ROHDE&SCHWARZ – M3TR, 500 skoků/s, KONGSBERG – MRR, ~200 skoků/s. Parametry pro rozbor a měření se budou provádět se stanicí F@STNET

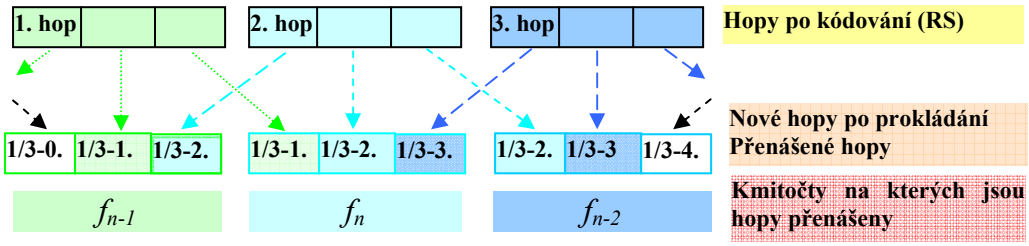
Z obr. 5 vidíme, že rychlost skákání je ~300 skoků/s, délka aktivního skoku je ~3 ms a doba potřebná k přepnutí na jiný kmitočet je ~0.33 ms.



Obr. 5. Rychlost skákání, aktivní délka skoku a doba potřebná na přeladění na další kmitočet



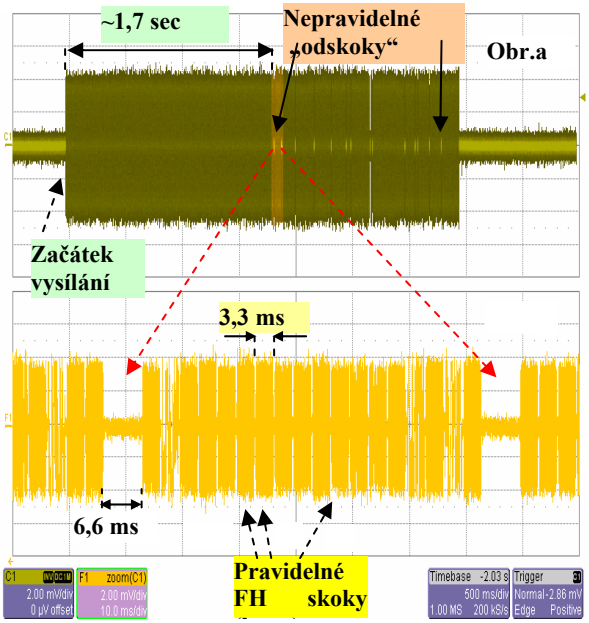
Obr. 4a Tvorba a zpracování FH signálu TRANSEC). Spreading a despreading FH signálu. Utajení a zabezpečení přenosu (COMSEC); vnější a vnitřní kódování (Read-Salamonuv a konvoluční kód) + překládání dat (1/3+1/3+1/3 z prvního, druhého a třetího hopu). Digitalizace řeči:adaptivní delta modulace ADM (SVDM), vokodér (0.6,.2, 2.4 a 4,8 kb/s).



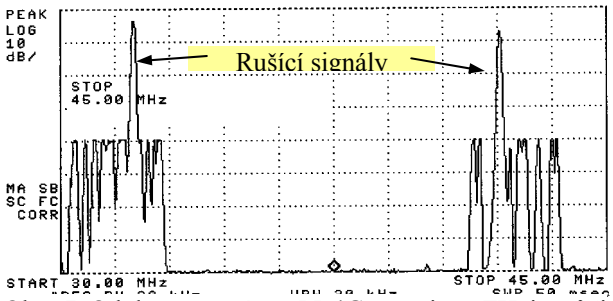
Obr. 4b Překládání 1/3 hopu do předchozího a 1/3 do následujícího hopu. K opravení plně zarušeného hopu postačí znát předchozí a následující hop.

Z obr. 6 vidíme, že FH skoky se pravidelně opakují a po zahájení vysílání ~1,7 sec systém si nepravidelně „odskočí“ na sledované kmitočty (výzvy a alarmu). Odskok trvá po dobu dvou skoků , tj ~6.6 msec.

Vidíme, že FH systém v našem případě skáče ve dvou kmitočtových podrozsazích a to od 30,5 do 32,5 MHz a 41 až 43 MHz. Vzdálenost sousedních skoků v druhém pásmu byla nastavena na hodnotu 100 kHz. Hustota skákání ve druhém pásmu je menší, což ukazuje i naměřený obr. 7



Obr. 6 Obrázek potvrzuje periodičnost skoků a po určité době od zahájení vysílání systém odskakuje na důležité sledované kmitočty

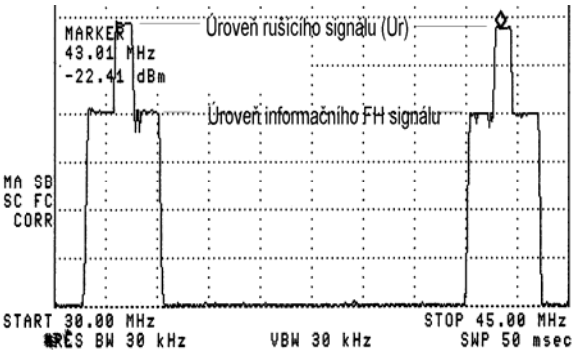


Obr. 7 Odolnost systému PR4G v režimu FH je vůči silnému úzkopásmovému rušení značně odolný. Ani dva úzkopásmové signály o 30 dB větší než FH signál nezarušil rádiový přenos

2.2 Rušení FH systému se spojitým širokopásmovým signálem

Systém PR4G je značně odolný vůči rušení úzkopásmovým signálem v porovnání s jinými FH stanicemi. Rušení na jednom kanálu (kmitočtu) i velmi silným signálem systém PR4G praktický nelze zarušit. I když rušící signál je větší o více než 1000 krát, jak je ukázáno na obr. 7, systém PR4G bude nezarušený.

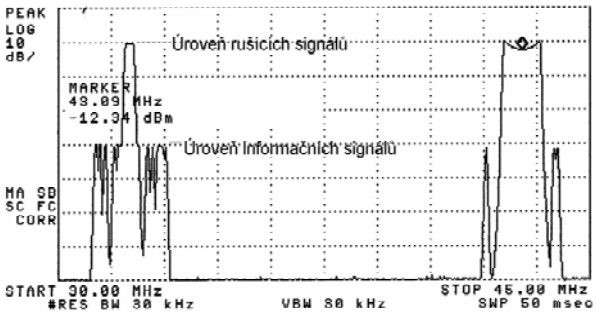
Jestliže rušící signál je širokopásmovější, ne však větší než 1/3 FH pásma, avšak rušící výkon může být větší o hodnotu až 20 dB, ani v takovém případě FH signál nebude zarušen, viz obr. 8.



Obr. 8 Rušení FH signálu s rušícím signálem o šířce pásma menším než 30% a větší ampliudu rušícího signálu o ~20 dB než je hodnota přenášeného signálu. I takk zarušení PR4G systému nedošlo

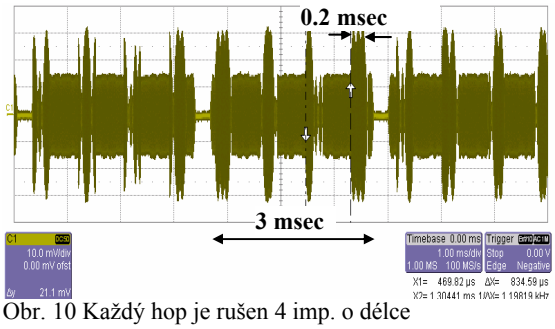
Při použití režimu MIX = FH+FCS nedošlo k zarušení v uvedených rušících systémů. Při použití režimu MIX nedošlo k zarušení ani v případě rušení uvedeného na obr. 9. Při zarušení více než 20 % vysílaného signálu došlo k automatickému přepnutí do režimu. V režimu FCS (Free Channel Selection – vyhledej volný kanál) je při každém zahájení vysílání vyhledán nový nerušený kanál. Dobře naprogramovaný režim MIX je téměř nezarušitelný.

Úspěšné zarušení FH systému nastalo až za podmínek širokopásmového signálu, kdě bylo zarušeno více než 1/3 pracovního kmitočtového pásma, jak ukazuje obr. 9. Amplitudová velikost rušícího signálu je ~1000 krát (30 dB). Při menší amplitudě rušícího signálu byl signál částečně srozumitelný z důvodu, že druhé podpásmo odsahuje jen 4x menší počet nosných kmitočtů, rozteč kmitočtů je 100 kHz.

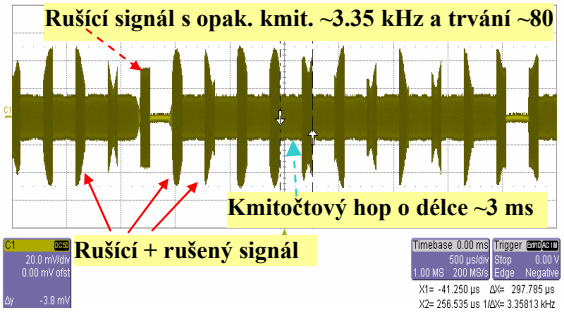


Obr. 9 K zarušení datového přenosu došlo až při rušení větším než FH signál až o 30dB a šířka je větší něž 1/3 pracovního pásma FH systémem.

Prověření zarušitelnosti bylo provedeno i v časové rovině, obr. 10 a obr. 11.



Obr. 10 Každý hop je rušen 4 imp. o délce 0.2 msec, tj. 0.8 msec; systém FH nebyl zarušen



Obr. 11 Rušení FH syst. s impuls. menším než 0.03 msec opakovacím kmitočtem 10 krát větším než je kmitočet skákání. Systém FH nebyl zarušen.

K zarušení systému FH potřebujeme rušit alespoň 1/3 vysílaného FH signálu.. Potvrdila se skutečnost, že signál FH musí být rušen po dobu více než 1/3 délky hopu, nebo zarušit plně každý třetí hop.

Seznam bibliografických odkazů

[1] LÚČ, A.:. Vývoj taktických FH systémů. Konference: Rozvoj komunikačních FH systémů. VA Brno 2003.
[2] LÚČ, A.: Moderní radiové stanice pro automatizované systémy velení a řízení. Konference: Moderní rádiové stanice pro velení a řízení. UO Brno 2005.
[3] LÚČ, A. a kol.: Směry rozvoje moderních rádiových stanic a vytváření rádiových sítí NRC. Mezinárodní konference: Komunikace v prostředí Network Enabled Capability. UO Brno 2007.
[4] LÚČ, A.: Rádiové ochranné rušiče proti teroristickým útokům. Mez. kongres informačních technologií v krizovém řízení „Interop-soft“ Brno 2007.

Summary: Presentation of this article discusses common the problems of FH (Frequency Hopping) systems. The authors try to give the newest information of modern FH systems. Submitted contribution deals with the following problems:

- Direct access to the radio networks.
- Integration and mobile of digital radio systems.
- The automatic FH mode changes into the FCS (Free Channel Selection) mode.
- Priority of commander in the home network.
- A possibility to communicate by a FH station with a standard radio station (out of FH).

The article discusses also the jamming of FH systems and other problemsIng.

prof. Ing. Andrej LÚČ, CSc.
Ing. Juraj HRABOVSKÝ
Ing. Miachal HALUZA
URC Systems s.r.o
Pražákova 49
619 00 Brno
Česká republika
E-mail: andrej.luc@urc-systems.cz

THE ACADEMY OF THE ARMED FORCES
of General Milan Rastislav Stefanik

in Liptovsky Mikulas

Department of Electronics

and
collaborating body

SLOVAK SOCIETY OF ELECTRICAL ENGINEERS
chapter Liptovsky Mikulas

organize

international scientific conference

NEW TRENDS IN SIGNAL PROCESSING IX

28 - 30 May 2008
Hotel Tatranske Zruby

Information on conference can be found at: http://www.aoslm.sk/nsss2008/