ITHEA

International Journal

INFORMATION TECHNOLOGIES & KNOWLEDGE

2008 Volume 2 Number 2

International Journal
INFORMATION TECHNOLOGIES & KNOWLEDGE
Volume 2 / 2008, Number 2

IJ ITK is official publisher of the scientific papers of the members of the ITHEA International Scientific Society, the Association of Developers and Users of Intellectualized Systems (ADUIS) and the Association for Development of the Information Society (ADIS)

IJ ITK rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITK as well as the **subscription fees** are given on *www.foibg.com*.
The camera-ready copy of the paper should be received by e-mail: info@foibg.com.
Responsibility for papers published in IJ ITK belongs to authors.

General Sponsor of IJ ITK is the **Consortium FOI Bulgaria** (www.foibg.com).

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Vol.2, Number 2, 2008

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with the **V.M.Glushkov Institute of Cybernetics of NAS, Ukraine**, and the **Institute of Mathematics and Informatics** and the **Institute of Information Technologies, BAS, Bulgaria**.

Publisher: **Institute of Information Theories and Applications FOI ITHEA**
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, www.foibg.com, e-mail: info@foibg.com
Printed in Bulgaria

ISSN 1313-0455 (printed) **ISSN 1313-048X (online)** **ISSN 1313-0501 (CD/DVD)**

ANALYSIS OF INFORMATION SECURITY OF OBJECTS UNDER ATTACKS AND PROCESSED BY METHODS OF COMPRESSION

Dimitrina Polimirova-Nickolova, Eugene Nickolov

Abstract: In this paper a methodology for evaluation of information security of objects under attacks, processed by methods of compression, is represented. Two basic parameters for evaluation of information security of objects – TIME and SIZE – are chosen and the characteristics, which reflect on their evaluation, are analyzed and estimated. A co-efficient of information security of object is proposed as a mean of the coefficients of the parameter TIME and SIZE. From the simulation experiments which were carried out methods with the highest co-efficient of information security had been determined. Assessments and conclusions for future investigations are proposed.

Keywords: Information Security, File Objects, Information Attacks, Methods of Compression, Information Flows, Coefficient of Information Security

ACM Classification Keywords: D.4.6 Security and Protection: information flow controls

Introduction

The development of information systems and technologies extends the necessity of processing, transferring and saving of volume sizable information flows, which are in network TCP/IP environment. These information flows, in the form of file objects, are an object of non-stop attacks according to their information security, which determines the significant necessity for investigation of methods and means for their protection.

A general strategy for protecting file objects could include applying compression methods to objects to achieve decrease in volume size of information flow.

For the purposes of this paper the following reservation can be made: it is enough to investigate only the influence of compression methods on objects exposed to one or more attacks, as the difference in their behavior before and after the attacks when standard and not corporate (government) requirements are used is taken into consideration.

The Problem

The main aim of this paper is to make analysis of the information security of the file objects, found in TCP/IP environment, under information attacks, noting the influence of the compression methods.

The following tasks are set in reaching the aim:

1) to offer a methodology for evaluation of the information security of objects under attack and processed with a method of compression;

2) to set a co-efficient of information security of an object;

3) to find the methods of compression those reach the highest values of the co-efficient of information security.

For the aim of this paper the following work definitions are proposed [1], [2], [3]: 1) as information security we will note the protection of the information in an object from a random or purposeful access aimed at reading, transferring (coping), modifying or destroying the information in it; 2) as file object we will note the whole interconnected data or program records, saved under one name; 3) as information attack we will note an attack in connection with the content of the current information stream; 4) as method of compression we will note the procedure for data encoding aimed at shrinking their volume during the processes of transfer and storage.

1. METHODOLOGY OF EVALUATION OF THE INFORMATION SECURITY.

The methodology for evaluation of the information security of an object supposed to attack and processed with a method of compression will meet the following limitations:

- only the potential sets of attacks, methods and objects will be analyzed. These sets are made by stagely reduction of the known at the moment of study information attacks, methods of compression and file objects by using of matrix transformations. The stages of reduction of the multitudes are described in [4];
- the experiments are conducted at standard users', non-corporations' (governments') requirements;
- in order to simplify the computations the lossy methods of compression are except;
- in conducting the experiments for determining the co-efficient of information security, the objects used have equal or similar starting size.

Upon determining [4] the real relationships between attacks, methods and objects, studies and analysis can be made in the following three directions:

- ✓ evaluation of the *success of the attack*, made on an object processed with a method of compression;
- ✓ evaluation of the *protection by method of compression*, applied on an object, exposed to an attack;
- ✓ evaluation of the *security of an object* exposed to an attack and processed by a method of compression.

This paper is aimed at the possibility to evaluate the security (information security) of objects supposed to information attacks noting the influence of the methods of compression.

1.1 Setting the basic parameters for evaluating the information security.

The information security of an object can be determined as a quantitative value, which depends on several fundamental parameters, which can be represented as ratios of separate values before and after certain impact.

For the purposes of this paper considering the usage of standard users' requirements, not corporations' (governments') requirements it is enough to study and evaluate only the parameters *TIME* and *SIZE*, by marking the difference in the objects behavior before and after applying the method of compression.

The parameter *TIME* (*T*) reflects the evaluation of time for attack at an object before and after the influence of the method of compression. The parameter *SIZE* (*S*) reflects the evaluation of the size of an object before and after its processing with a method of compression.

1.2. Determining the characteristics which influence over chosen parameters.

After determining the main parameters, which will be analyzed and evaluated with regard to the information security of an object, is necessary to determine the basic characteristics, which have influence on the evaluation of the main parameters.

The basic characteristics, which have influence on the evaluation of the parameters BEFORE applying a method of compression to the object, are:

- for the evaluation of the parameter *TIME* the following characteristics can be taken into consideration: *time for examination* and *time for processing*;
- for the evaluation of the parameter *SIZE* will pointed characteristics depending of the category to which file objects belong to. Two basic categories are: DIRECTLY USED (these are objects, which have to be used directly) and NON-DIRECTLY USED (these are objects, requiring secondary processing to become directly used):
  - o the characteristics, which have influence on the evaluation of the parameter *SIZE* for objects belonging to DIRECTLY USED category, are: *characters' size, image's size, video's and audio's size* and *official information's size*;
  - o the characteristics, which have influence on the evaluation of the parameter *SIZE* for objects belonging to NON-DIRECTLY USED category, are: *resolution of the image*, *bit depth, official information's size* (for representatives of the group "graphical objects"); *sample size*, *sample rate*, *official information's size* (for representatives of the group "music and sound").

The basic characteristics, which have influence on the evaluation of the parameters AFTER applying a method of compression to the object, are:

- for the evaluation of the parameter *TIME* the characteristic *time for restoration* is added to these, mentioned above before applying a method of compression to an object;
- for the evaluation of the parameter *SIZE* are specified characteristics, depending of the method of compression applied over the object:

o when statistical methods of compression are applied, the characteristics (in addition to these mentioned above for DIRECTLY USED objects), which have influence on the evaluation, are: *entropy of the message*, *information redundancy*, *level of compression*, *bits of information after compression*, *size of the model for decompression*;

o when dictionary methods of compression are applied, the characteristics (in addition to these mentioned above for DIRECTLY USED objects), which have influence on the evaluation, are: *size of the dictionary*, *entropy of the message*, *information redundancy, level of compression*;

o when image methods of compression are applied the characteristics (in addition to these mentioned above for NON-DIRECTLY USED graphical objects), which have influence on the evaluation, are: *average number of pixel repetitions, average number of sequenced pixels*, *level of compression*;

o when audio methods of compression are applied the characteristics (in addition to these mentioned above for NON-DIRECTLY USED objects from the group "sound and music"), which have influence on the evaluation, are: *level of sample size*, *level of sample rate*, *average number of sequenced zero samples, level of compression*.

1.3. Determining the evaluations of the characteristics, which have influence on the general valuation of the respective parameter.

Each characteristic is necessary to be evaluated with respect to the information security of an object under attack before and after applying a method of compression. To determine these evaluations is taken into consideration additional factors, which have influence on the evaluation of the respective characteristic. After that is necessary to examine each characteristic by providing simulation experiments, which will determine the relationship between the obtained after the examination result and the evaluation of the characteristic with respect to the information security of an object (for example: the increasing of the time of an attack to process an object, increases object's information security, which leads to higher valuation of this characteristic; the increasing of the size of the model for decompression decreases the possibility for better compression of the object, which leads to faster restoration in its original state, respectively to faster braking the protection mechanism of the object as a mean of method of compression, that means lower valuation of this characteristic with respect to the information security of this object). At the end the valuation (*V*) of the respective characteristic is determined.

1.4. Setting a weighted co-efficient for each characteristic.

The weighted co-efficient (*W*) determine the level of influence which each valuation of the respective characteristic have influence on the general evaluation of the parameter to which it belongs to. For determining the weighted co-efficient of the characteristic is used the AHP (Analytic Hierarchy Process) method [5], which consists of four basic stages: 1) determining the characteristics which have to be evaluated; 2) arranging the chosen characteristics in a matrix; 3) comparing each couple of characteristics by preliminarily selected measurement scales for evaluation; 4) determining the respective weights of the characteristics by consecution of mathematical operations.

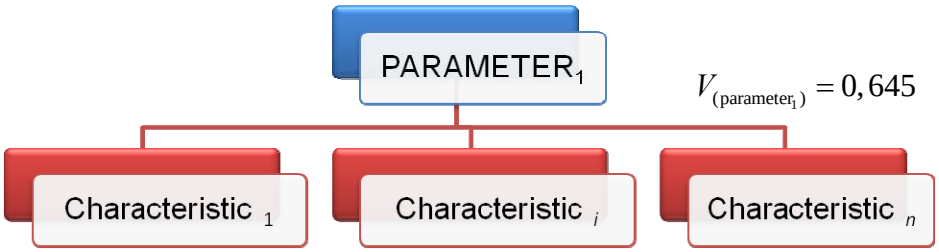
1.5. Estimating the general evaluation of the respective parameter.

The estimating of the general evaluation of the parameter consists of the following stages:

1) determining the evaluation of the characteristics, which have influence on the basic evaluation of the selected parameter $V_{(\text{charact.}_n)} = [0 \div 1]$, where *n* is the number of the characteristics;

2) setting the weighted co-efficient of each characteristic $W_{(\text{charact.}_n)}$, like $\sum_{i=1}^{n} W_i = 1$;

3) determining the evaluation of the parameter as $V_{(\text{parameter}_1)} = \sum_{i=1}^{n} \left(V_{(\text{charact.}_i)} . W_i \right)$ (Figure 1).



$V_{(\text{parameter}_1)} = 0,645$ From 0 to 1

$V_{(\text{charact.}_1)} = 0,4$ $V_{(\text{charact.}_i)} = 0,8$ $V_{(\text{charact.}_n)} = 0,5$ From 0 to 1

$W_{(\text{charact.}_1)} = 0,05$ $W_{(\text{charact.}_i)} = 0,5$ $W_{(\text{charact.}_n)} = 0,45$ $\sum_{i=1}^{n} W_i = 1$

Figure 1 Determining the evaluation of the parameter

2. DETERMINING THE CO-EFFICIENT OF INFORMATION SECURITY.

A co-efficient of information security is compounded to analyze the information security of the objects. It is presented as a variable, formed from the examined above parameters *TIME* and *SIZE*, reflecting the condition of the object before and after applying methods of compression.

2.1. Determining the co-efficient of information security of an object in regard to the evaluation of the parameter *TIME* ($K^{IS(T)}$).

The determination of $K^{IS(T)}$ proceeds over the following stages:

(1) for each relation attack—method—object is determined relatively valuation of the time ($RV_{(T)}$). It presents the number of increases of the value $V_{(T)}$ of an object after processing it with method of compression. The relatively valuation of the time can be represented as a ration of the *valuation-delta* ($\Delta V_{(T)}$) and *valuation-prim* ($V'_{(T)}$) for the security of the object with respect to the time (Formula 1):

$$RV_{(T)} = \frac{\Delta V_{(T)}}{V'_{(T)}} \qquad \text{Formula 1}$$

where $\Delta V_{(T)} = V''_{(T)} - V'_{(T)}$ like $V'_{(T)}$ is the determined valuation of information security of an object in regard to the time before applying the method of compression and $V''_{(T)}$ is the determined valuation of information security of an object in regard to the time after applying the method of compression;

(2) for each object o_f is determined the highest value of relatively valuation of the time ($\max RV_{(T)}$), which presents the highest increase of $V_{(T)}$, achieved by this object in all triple relations;

(3) for each relation attack—method—object is determined the co-efficient of information security with respect to the parameter *TIME* ($K^{IS(T)}$). For each triple relation this co-efficient presents the part of maximum possible value of relatively valuation of the time, which the object is achieved (Formula 2):

$$K^{IS(T)} = \frac{RV_{(T)}}{\max RV_{(T)}} \qquad \text{Formula 2}$$

Graphically $K^{IS(T)}$ can be presented as (Expression 1):

$$K_z^{IS(T)} = f\left(a_i, m_j\right) \quad \text{for each } o_f \qquad \text{Expression 1}$$

where $a_i \in A_{pot}\{a_1, a_2, ..., a_i, ..., a_p\}$, $m_j \in M_{pot}\{m_1, m_2, ..., m_j, ..., m_q\}$, $o_f \in O_{pot}\{o_1, o_2, ..., o_f, ..., o_r\}$, and the index *z* is changing within the bounds of the formula a_p, m_q and o_r.

2.2. Determining the co-efficient of information security of an object in regard to the evaluation of the parameter *SIZE* ($K^{IS(S)}$).

The determination of $K^{IS(S)}$ proceeds over the following stages:

(1) for each relation attack—method—object is determined relatively valuation of the size ($RV_{(S)}$). It presents the number of increases of the value $V_{(S)}$ of an object after processing it with method of compression. The relatively valuation of the size can be represented as a ration of the *valuation-delta* ($\Delta V_{(S)}$) and *valuation-prim* ($V'_{(S)}$) for the security of the object with respect to the size (Formula 3):

$$RV_{(S)} = \frac{\Delta V_{(S)}}{V'_{(S)}} \qquad \text{Formula 3}$$

where $\Delta V_{(s)} = V''_{(s)} - V'_{(s)}$ like $V'_{(s)}$ is the determined valuation of information security of an object in regard to the size before applying the method of compression and $V''_{(S)}$ is the determined valuation of information security of an object in regard to the size after applying the method of compression;

(2) for each object o_f is determined the highest value of relatively valuation of the size ($\max RV_{(S)}$), which presents the highest increase of $V_{(S)}$, achieved by this object in all triple relations;

(3) for each relation attack—method—object is determined the co-efficient of information security with respect to the parameter *SIZE* ($K^{IS(S)}$). For each triple relation this co-efficient presents the part of maximum possible value of relatively valuation of the size, which the object is achieved (Formula 4):

$$K^{IS(S)} = \frac{RV_{(S)}}{\max RV_{(S)}} \qquad \text{Formula 4}$$

Graphically $K^{IS(S)}$ can be presented as (Expression 2):

$$K_z^{IS(S)} = f\left(a_i, m_j\right) \quad \text{for each } o_f \qquad \text{Expression 2}$$

where $a_i \in A_{pot}\{a_1, a_2, ..., a_i, ..., a_p\}$, $m_j \in M_{pot}\{m_1, m_2, ..., m_j, ..., m_q\}$, $o_f \in O_{pot}\{o_1, o_2, ..., o_f, ..., o_r\}$, and the index *z* is changing within the bounds of the formula a_p, m_q and o_r.

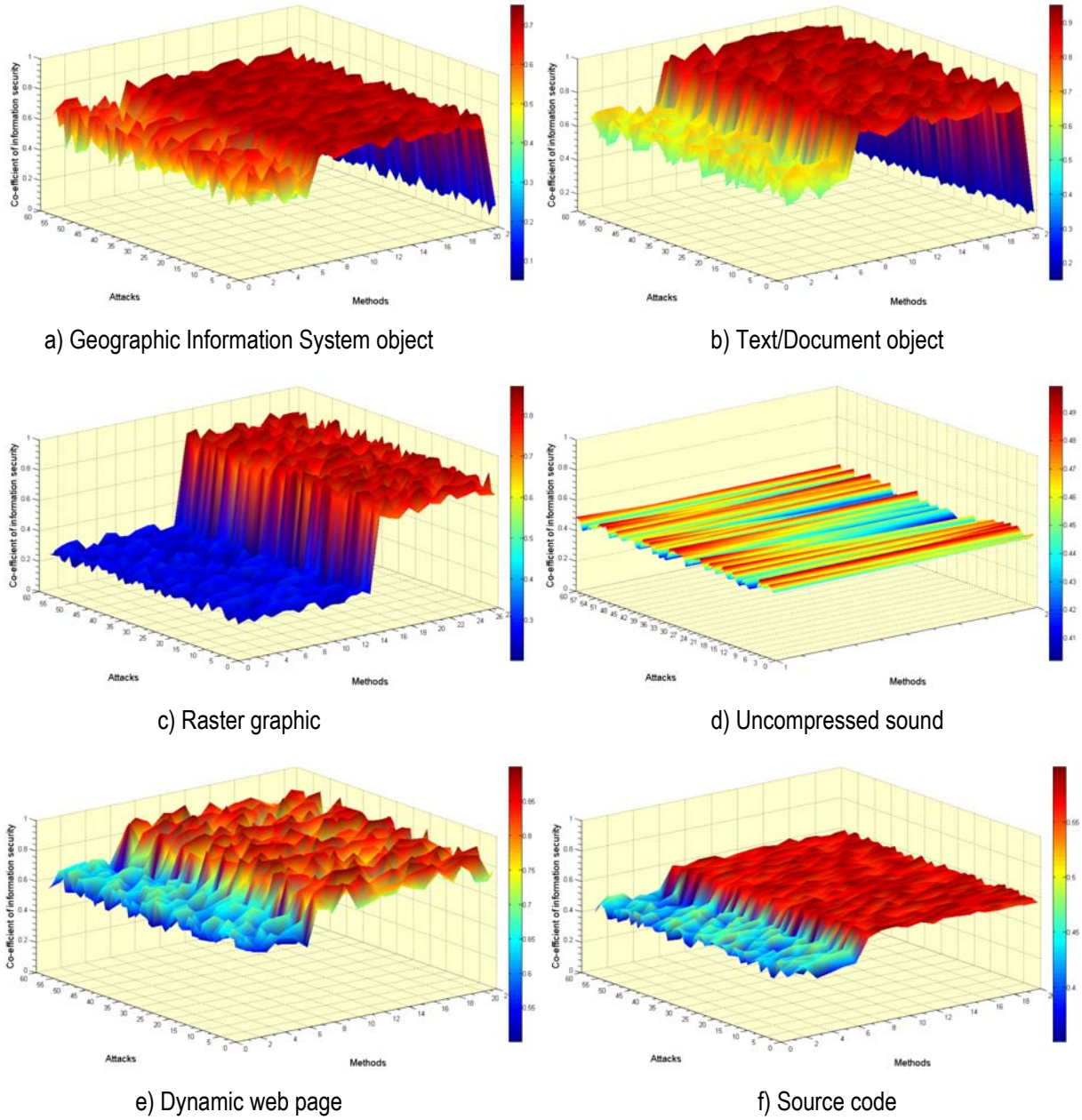
2.3 Determining the co-efficient of information security of an object (K^{IS}) as a mean of the co-efficients for evaluating the two parameters (*TIME* and *SIZE*).

After determining of the co-efficients $K^{IS(T)}$ and $K^{IS(S)}$, for each object can be composed co-efficient of information security. The co-efficient of information security of an object (K^{IS}) can be determined as a mean of co-efficients for valuation of parameters *TIME* and *SIZE* (Formula 5):

$$K_z^{IS} = \frac{1}{n}\sum_{p=1}^{n} K^{IS(p)} \qquad \text{Formula 5}$$

where $K^{IS(p)}$ is the co-efficient of information security of an object in regard to a given parameter *p*, *n* is the number of investigated parameters in regard to information security of an object and *z* is changing within the bounds of the formula a_p, m_q and o_r.

Graphic interpretation for determined values of the K^{IS} for most frequently used file objects is shown on Figure 2a), b), c), d), e), f).



a) Geographic Information System object

b) Text/Document object

c) Raster graphic

d) Uncompressed sound

e) Dynamic web page

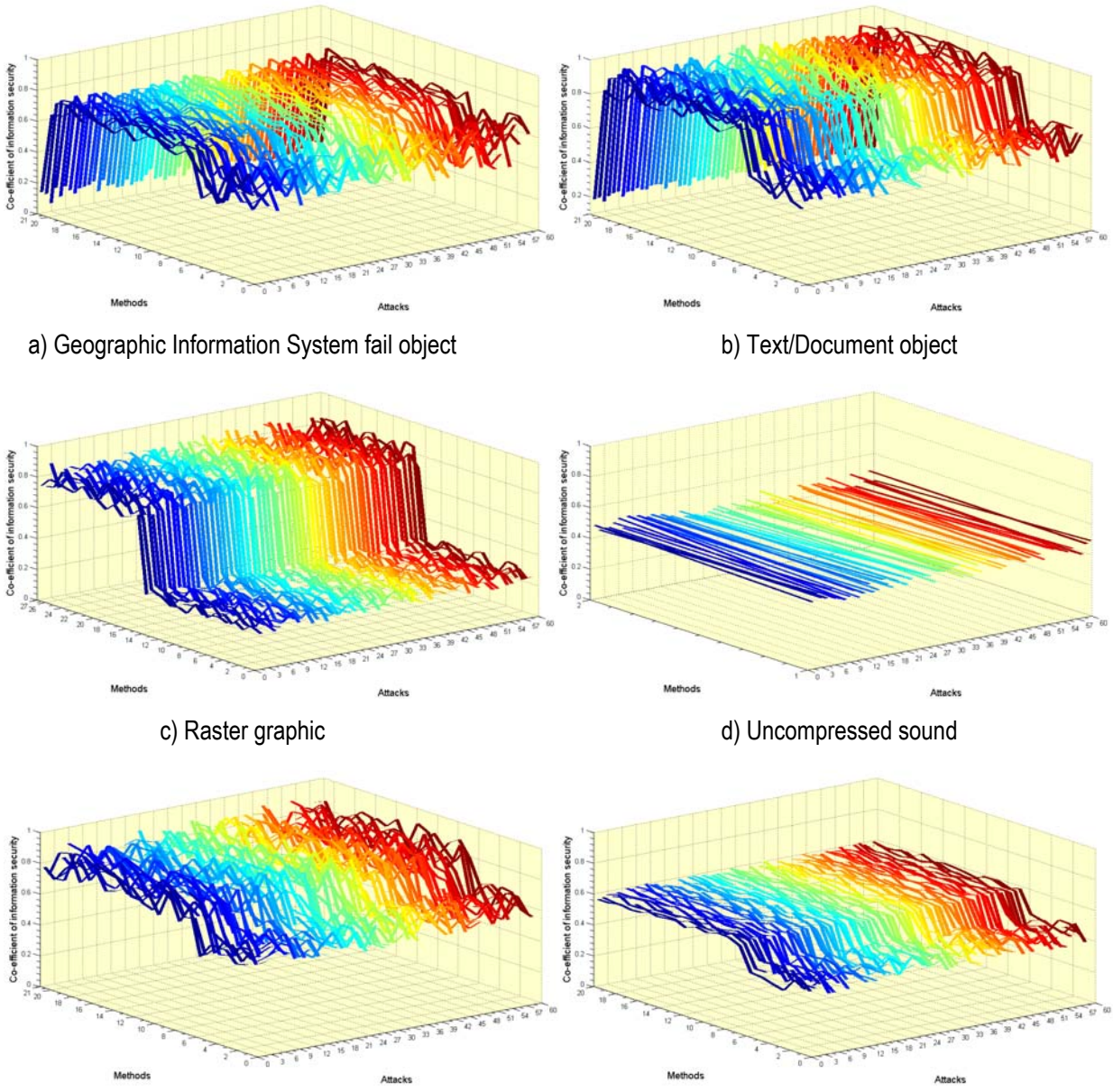
f) Source code

Figure 2 Graphic interpretation for determined values of the co-efficient of information security for different file objects

3. METHODS WITH THE HIGHEST VALUES OF THE CO-EFFICIENT OF INFORMATION SECURITY.

3.1. Determining the methods with the highest values of the co-efficient of information security for each object for the given attack.

After determining K^{IS} for each object we can determine which is the method with the highest value of K^{IS} for the given object and attack. On fig 3a), b), c), d), e), f) we can see a graphical presentation of the change in the co-efficient of information security for given objects in regard to given attacks, determined after applying the given methods for compression.



a) Geographic Information System fail object

b) Text/Document object

c) Raster graphic

d) Uncompressed sound

e) Dynamic web page

f) Source code

Figure 3 Distribution of the co-efficient for informational security for given object and attack, when a given method of compression is applied

3.2. Determining the methods with the highest values of the co-efficient of information security for all objects in regard to given attacks.

Thus for each object can be set up a group of methods of compression, reaching the highest values of K^{IS} with respect to all attacks on which the object can be exposed.

Derived from this particular scientific work the results are the basis for further research in connection with the opportunity to determine the method of compression, which will have the lowest risk in regard to the information security for the given object and attacks, for which it can be applied.

Assessments

1) Parameters used for determining *TIME* and *SIZE* are sufficient for researching information security of objects and computer systems and networks for consumer, not governmental (corporate) needs.

2) Evaluation in regard to the selected objects, which were processed with methods of compression, is positive and the allowances do not affect the derived result.

3) In regard to the methods of compression we used the assessment is positive and the above mentioned experiments can be used and tailored to other methods of compression.

4) We can conclude, looking at the experiments, that with the decreasing size of an object after compression, time needed for an attack to complete its work over the object will increase.

5) As with the co-efficient of information security the best results were obtained from data objects, processed with dictionary methods of compression, and the worst results were obtained with the graphics objects processed with statistical methods of compression.

6) From all 59 methods of compression, 13 of them gave us the highest value of the co-efficient of information security of the object. They are from the group of dictionary methods and image methods of compression.

Bibliography

[1] Elena Ferrari, Bhavani M. Thuraisingham, *Web and Information Security*, IRM Press, 2006, ISBN: 1-59140-589-0, p. 215

[2] http://www.answers.com/file

[3] David Salomon, *Data Compression: The Complete Reference*, Springer, 2006, ISBN: 1846286026, p.1-9

[4] Polimirova, D., Nickolov, E., Nikolov, C., *Investigating The Relations Of Attacks, Methods And Objects In Regard To Information Security In Network TCP/IP Environment*, International Journal "Information Theories & Applications", vol. 1 / 2007, Number 1, ISSN 1313-0455, p. 85-92

[5] Hubert Hasenauer, *Sustainable Forest Management: Growth Models for Europe*, Springer 2006, ISBN: 9783540260981 p.267-269

Authors' Information

PhD Student, Dimitrina Polimirova, Research Associate, National Laboratory of Computer Virology, Bulgarian Academy of Sciences, Phone: +359-2-9733398, E-mail: polimira@nlcv.bas.bg.

Prof. Eugene Nickolov, DSc, PhD, Eng, National Laboratory of Computer Virology, Bulgarian Academy of Sciences, Phone: +359-2-9733398, E-mail: eugene@nlcv.bas.bg.

ICT SECURITY MANAGEMENT

Jeanne Schreurs, Rachel Moreau

Abstract: Security becomes more and more important and companies are aware that it has become a management problem. It's critical to know what are the critical resources and processes of the company and their weaknesses. A security audit can be a handy solution. We have developed BEVA, a method to critically analyse the company and to uncover the weak spots in the security system. BEVA results in security scores for each security factor and also in a general security score. The goal is to increase the security score Ss to a postulated level by focusing on the critical security factors, those with a low security score.

Keywords: Security, Scan, Audit

Introduction

As a consequence of the fast integration of technologies as Internet, Intranet, Extranet, Voice over IP and e-commerce, companies ICT-infrastructure will move to more openness to the outside world and as a consequence

will become more vulnerable for security threats. This offers lots of new opportunities but also creates new threats. That's why focus and responsibility concerning security become even more and more important. The Computer Crime and Security Survey 2005 shows that these are the 10 most frequent attacks or misuses: Virus, insider abuse of net access, laptop/mobile theft, unauthorized access to information, denial of service, abuse of wireless network, system penetration, theft of proprietary info, telecom fraud and financial fraud. Figures show that attacks come from inside as well as from outside the organisation and bring along large costs. Especially unauthorized access and laptop and mobile theft becomes a enormous expense for the companies during the last years. Because of these large costs, companies became more and more aware that they not only deal with a technical problem but also with a management problem. To tackle this management problem, it is quite important to know the ICT-security state your company is in.

ICT security management

Spending each year a certain amount on security measures is not enough. A company needs a total security approach. It is a must to know what are the critical resources and processes of the company and their weaknesses so the can be protected in the right way.

A solution to this is a security audit. A security audit is ideal to detect the weak spots in the ICT security state of the company. Based on the results of the audit, a security policy can be developed, adjusted to the company situation. A security audit can be used to analyse and describe the security level.

1. Security audit checklist

We have developed a security audit, called BEVA. BEVA is a method to analyse critically the company and to uncover the weak spots of the security system. It positions the company on point of the security aspects in the different areas of business functions. We have developed a standard list that covers all aspects of security, structured in 10 domains being:

- Security policy
- Organization of information security
- Asset management
- Human resources security
- Physical and environmental security
- Communications and operations management
- Access control
- Information systems acquisition, development and maintenance
- Information secuirity incident management
- Business continuity management

Each of these areas consists of different security factors. The factors are in their turn tested on the basis of several subcriteia. Our list for the security factors is based on the standard ISO 17799. The 38 security factors are spread over the 10 domains, as set forward in the standard ISO17799 model.

For example you have the domain "access control" and in this domain you have the factors: requirements for access, management of user access, user responsibility, control of network access, control access to OS, control of access to applications and information and use of mobile infrastructure.

For each of the 38 factors, a number of subcriteria are formulated. We developed a list of questions, covering the subcriteria we created. The questions are partly based on the "checklists in information management" SDU publishers. (www.riskworld.net/7799-2.htm).

2. The audit process and the calculation of security factor scores Sfi's and the security score Ss

To collect the information about the current security situation of the company, we start with the questioning of the key persons in the company using the audit checklist questionnaire.

The company determines which systems or processes are critical for them and connected with it, which security factors are important or relevant. An importance rate is given to the security factors from A (low importance) to E (high importance) (see figure 1).

Security Factor Sfi	Importance	Sub Factor	Relevance/weight 1 to 4	Code question	Question	evaluation 1 to 4
Domain: Access control						
Sf20. Business requirements for access controlPremise	B	access control policymanagement	3	20.1	Is the access control policymanagement based on the business security requirements?	3
				20.2	Are aspects of logical and physical access control included?	3
				20.3	Is it clear for users and service providers which rules are applicable?	2
Sf21. User access management	C	registration of users	2	21.1	Is there any formal user registration and de-registration procedure for granting access to multi-user IS and services?	1
		privilege management	1	21.2	are privileges and allocated on need-to-use basis?	3
				21.3	are privileges only allocated after formal authorisation process?	1
		user password management	4	21.4	should the allocation and the reallocation of passwords be controlled through a formal management process?	3
				21.5	are the users asked to sign a statement to keep the password confidential?	1
		review of user access rights	3	21.6	does there exist a process to review user access rights at regular intervals?	4

Figure 1: Questions audit checklist

In BEVA, we express the state of security into scores of the security factor (Sfi's). We do this for all the factors and in the end we give a general security score (Ss) over all security factors. We based our security analysis partly on the Marion-AP method.

Security factor Sfi	Security Subfactor Ssfij	Relevance /weight 1 to 4 w(i,j)	Code question	evaluation 1 to 4	mean evaluation 1 to 4 eval(i,j)	Security factor score Sfis
Domain: Access control						
Sf20. Business requirements for access controlPremise	access control policymanagement	3	20.1	3	2,67	2,67
			20.2	3		
			20.3	2		
		3				
Sf21. User access management	registration of users	2	21.1	1	1	2,25
	privilege management	1	21.2	3	2	
			21.3	1		
	user password management	4	21.4	3	2	
			21.5	1		
	review of user access rights	3	21.6	4	3,5	
			21.7	3		
		10				

Sfis = sum [(w(i,j) * eval(i,j)] / sum w(i,j)

Figure 2: Calculation of the Sf i's

To evolve to a security factor score, the key persons is asked to allocate a weight from 0 to 4 to the subcriteria of the security factors to indicate the relevance. Subsequently the evaluation starts and the list of questions is asked. Each question is given a score between 1 and 4. (see figure 2). The management team evaluates the

company for all aspects on a one to four scale and at the same time measures the importance or relevance of all subfactors.

When the questionnaire is completed, BEVA now calculates the security factor scores (Sf) being:

Sfi s = sum [eval (i,j) * w(i,j)] / sum w(i,k)

If all the factor scores are calculated also a general security score Ss is given:

Ss= sum [eval (1,38) * w(1,38)] / sum w(1, 38)

For example see factor 21 in the example: Sf21:[2*1 + 1*2 + 4*2 + 3*3,5]/10 = 2.25

Ss= in this example 2.66

Based on the evaluated questionnaire and the allocated weights, a realistic picture of the security situation of the company can be created as well general as by factor. The system BEVA creates a graphical output of the correlation diagram between these two variables measured for all aspects. Figure 3 shows the scores of all the security factors.

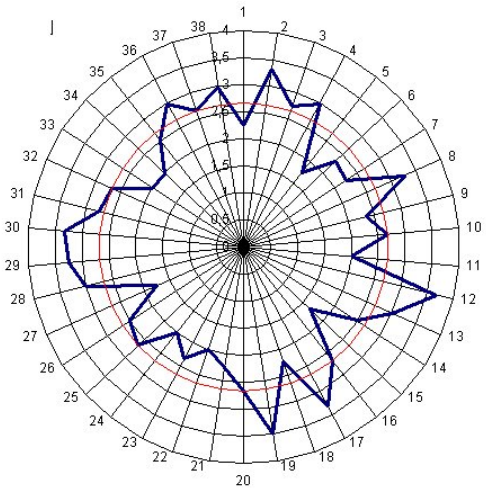


Figure 3: Graph of the security scores

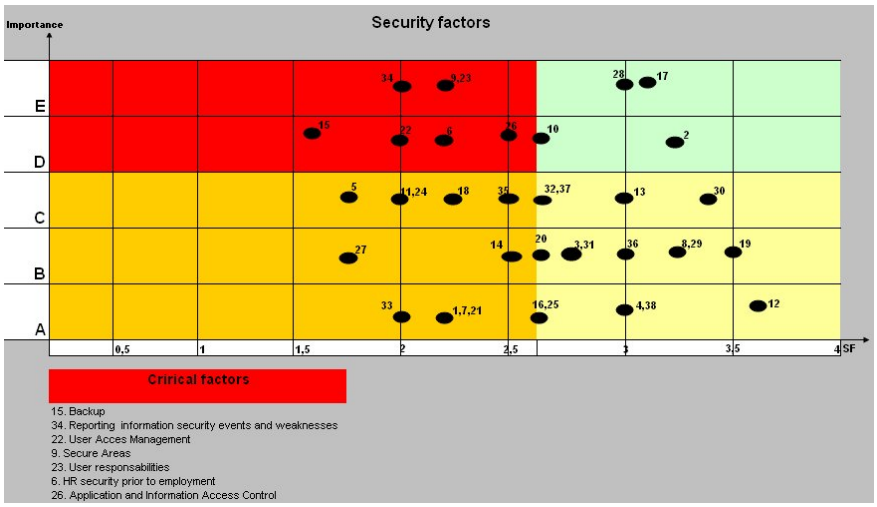
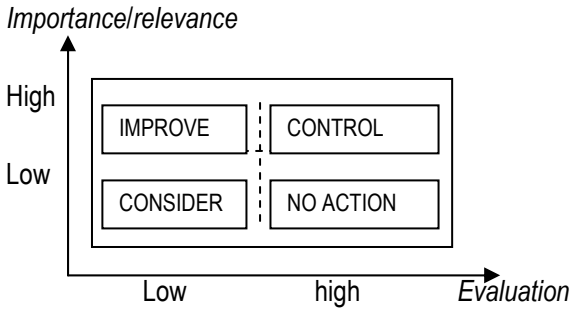


Figure 4: Graph of security factors and their importance

The red line states Ss the general security score. The blue line connects the individual scores of the security factors. Security factors 1, 5, 6, 7, 9, 11, 14, 15, 18, 21, 22, 24, 26, 27, 33 and 34 score beneath the general security score.

Figure 4 combines the scores of the security factor with its importance. For example factor 33 scores low namely 2 but has importance A, low importance. Factor 34 scores also 2 but had importance E, high importance. These differences are well stressed in this graphic. As you can see the *red* area highlights the security factors that score low and have a high importance. The factors lying in this area are critical and need immediate attention.

The *green* area is important and good secured. It is important to continue these actions and follow up these factors well. The *yellow* zone scores good but isn't that important, no action needs to be taken here. The less important factors that don't score well are situated in the *orange* zone. These factors need to be considered but probably with a small piece of the budget.



Now a clear view of the security situation is obtained. Feedback is given to the company and the evaluation states immediate points of action.

3. The occurrence of threats

The yearly organised CSI/FBI-study delivers the following probabilities for the threats (see fig. 5).

Our final goal is to influence the occurrence of the threats, or the probability of the occurrence of them, by implementing selective security measures in the company. This will impact in the long run the security situation.

We must concentrate on the critical security factors, following the results of the audit. If the security factor is critical, than the threats linked with it have a critical risk too.

In figure 6 we figured out the relations between the threats and the security factors

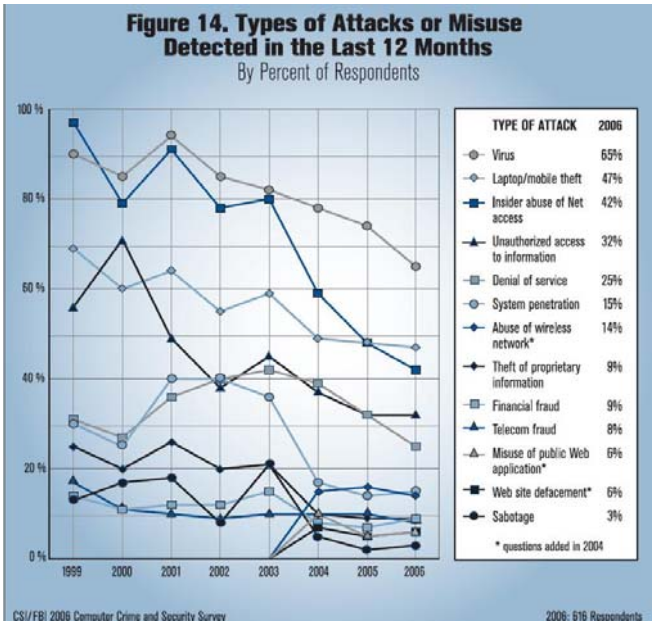


Figure 5: Threats and their occurance

Threat	Virus	Laptop/Mobile theft	Insider abuse of net access	Unauthorized access to information	Denial of Service - aanval	System penetration	Abuse of wireless network	Theft of proprietary information	Financial fraud	Telecom fraud	Misuse of public web application	Website defacement	Sabotage	Illegal software applications on the system (bots, Trojan horses,...)	Phishing	Misuse of the chat	Password sniffing	Exploiting the DNS server of the organization
1.1 Information security aspects of continuity mgt					x								x					
2.1 Business requirements for access control				x		x	x	x				x					x	
2.2 User access management				x		x	x	x	x			x					x	
2.3 User responsibilities				x		x											x	
2.4 Network access control			x				x		x							x		
2.5 OS access control													x					
2.6. Application en information access control											x					x		
2.7 Mobile computing and telenetworking		x					x			x								
3.1 Security requirements of IS				x		x												
3.2 Correct processing in applications				x		x												
3.3 Crypto-graphic controls			x	x		x	x	x										
3.4 Security of system files				x		x												
3.5 Security in development and support processes				x									x	x				
3.6 Technical vulnerability management				x		x							x					
4.1 Secure areas		x						x					x					
4.2 Equipment security		x						x					x					
5.1 Compliance with legal requirements									x									
5.2 Compliance with security policies and standards				x		x												
5.3 Information Systems audit considerations			x															
6.1 Prior to employment									x	x						x		
6.2 During employment			x				x			x	x					x		
6.3 Termination of change of employment		x	x	x		x	x	x	x	x		x						

7.1 Information security policy		x	x	x							x		x			x		
7.2 Internal organization			x															
7.3 External parties				x		x		x	x	x		x	x	x			x	
8.1 Operational Procedures and Operations mgt				x		x		x	x	x					x		x	
8.2 Third party service delivery management				x				x	x	x								
8.3 System planning and acceptance																		
8.4 Protection against malicious and mobile code	x											x		x	x		x	x
8.5 Back-Up management								x										
8.6 Network security management			x	x	x	x	x					x					x	x
8.7 Media Handling								x										
8.8 Exchange of information		x		x				x					x					
8.9 E-commerce					x	x		x	x	x	x				x		x	x
8.10 Monitoring			x	x		x	x		x	x						x		
9.1 Responsibility for assets		x											x					
9.2 Information classification								x					x					
10.1 Information incident management									x	x								

Figure 6: Relation between threats and security factors

4. Security measures and follow up

A next step is to create a list of action points. Taking into account the stated security budget and the factors and their importance, an action plan is suggested. In the CSI study we can find the most used measures. A table is created were the most used measures are related with the threats they prevent.

Measures	Virus	Laptop/Mobile theft	Insider abuse of net access	Unauthorised access to information	Denial of Service - aanval	System penetration	Abuse of wireless network	Theft of proprietary information	Financial fraude	Telecom fraude	Misuse of public web application	Website defacement	Sabotage	Illigal software applications on the system (bots, trojan horses,...)	Phishing	Misuse of the chat	Pasword sniffing	Exploiting the DNS server of the organisation
Firewall				x	x	x					x	x		x				
AntiVirus Software	x													x			x	
AtiSpyware Software														x			x	
Server Based Acces control list				x		x		x										
Intrusion detection system				x		x		x			x	x						
Ecryption for data				x				x									x	
Reusable account system																	x	
Intrusion prevention system				x		x		x			x	x						
Log management software			x						x	x						x		
Application level firewall	x				x									x				
Smart card/ one time password token				x		x		x										
Specialized wireless security							x											
Training personeel		x					x						x					
Endpoint security client software	x													x				
Update server	x			x		x		x				x		x				

Figure 7: Relation between measures and threats

The action plan concerning security will be implemented, taking into account the weakest security factors and of course considering the budget.

After a period of approximately 3 months after implementing the security measures, a new security audit should be taken. The new security score Ss is calculated and compared to the stated aimed Security score using the security measures. If there are security factors that score too low, these should be investigated and adjusted.

Conclusion

The awareness that security is a management problem is everywhere present. It's critical to know what are the critical resources and processes of the company and their weaknesses. Our security audit is a handy solution. We have developed BEVA, a method to critically analyse the company and to uncover the weak spots in the security system. BEVA results in security scores for each security factor and also in a general security score. The goal is to increase the security score Ss to a postulated level by focusing on the critical security factors, those with a low security score. The results of the audit are an ideal start to do risk analysis.

Bibliography

[Shannon, 1949] C.E.Shannon. The Mathematical Theory of Communication. In: The Mathematical Theory of Communication. Ed. C.E.Shannon and W.Weaver. University of Illinois Press, Urbana, 1949.

[Jean-Marc Lamère] la sécurité informatique; Dunod : La méthode MARION (Méthodologie d'Analyse de Risques Informatiques Orientée par Niveaux) www.eisti.fr/~bg/COURSITACT/TXT/m_marion.txt

[Val Thiagarajan B.E, 2005] Information Security Management; BS ISO/ IEC 17799:2005; SANS Audit Check List: author:., M.Comp, CCSE, MCSE, SFS, ITS 2319, IT Security Specialist.

Security Management: A New Model to Align Security with Business Needs; Sumner Blount, CA Security Solutions; August 2006

[Schreurs J, Moreau R.] ICT security management- ECEC 2007 (www.riskworld.net/7799-2.htm)

[Lawrence A. Gordon, Martin P. Loeb, William Lucyshyn and Robert Richardson] 2006 CSI/FBI-study about cybercrime: COMPUTER CRIME AND SECURITY SURVEY

https://event.on24.com/eventRegistration/EventLobbyServlet?target=registration.jsp&eventid=27372&sessionid=1&key=42F39B89EE0B30BA951711A5E7A98EDD&sourcepage=register

http://mediaproducts.gartner.com/gc/webletter/computerassociates/vol3issue3_risk/index.html

Authors' Information

Jeanne Schreurs – prof. Business informatics, Universiteit Hasselt; gebouw D, Agoralaan, 3590 Diepenbeek, Belgium; e-mail: jeanne.schreurs@uhasselt.be

Rachel Moreau - Universiteit Hasselt; gebouw D, Agoralaan, 3590 Diepenbeek, Belgium; e-mail: Rachel.moreau@uhasselt.be

COMPLEX PROTECTION SYSTEM OF METADATA-BASED DISTRIBUTED INFORMATION SYSTEMS

Denis Kourilov, Lyudmila Lyadova

Abstract: A description of architecture and approaches to the implementation of a protection system of metadata-based adaptable information systems is suggested. Various protection means are examined. The system described is a multilevel complex based on a multiagent system combining IDS functional abilities with structure and logics protection means.

Keywords: adaptable information systems, protection mechanisms, metadata, multiagent systems.

ACM Classification Keywords: D.2 Software Engineering: D.2.0 General – Protection mechanisms; K.6 Management of Computing and Information Systems: K.6.5 Security and Protection – Authentication, Insurance, Invasive software (e.g., viruses, worms, Trojan horses), Unauthorized access (e.g., hacking, phreaking); I.2 Artificial Intelligence: I.2.11 Distributed Artificial Intelligence – Multiagent systems.

Introduction

Modern information systems (IS) are developed for various application domains and environments that influence their protection and reliability level. The typical peculiarities of modern IS are:

- *Complexity*. As the complexity of information systems grows, they reveal more and more vulnerabilities that are difficult to disclose and repair.
- *Openness and integrability*. The openness of information systems and their integrability, and their being interconnected with internal IS are potentially responsible for IS intrusion vulnerability.
- *Adaptability and expandability*. IS are flexible enough to be configured for certain working conditions and users' needs, internal developers can also expand the systems' functions. This also creates the risk of malware intrusion.
- *IS are distributed*. IS' subsystems can interconnect via network, posing extra security threats, such as IS and client back-ends attack.

Security methods existing today do not allow us to fully protect dynamically adaptable IS that function in distributed environment. They can protect either the program code, or IS data. IS can be adapted with: database dynamic restructuring tools; automatic generation and tuning user interface; query and reporting facilities; business process management tools; connection oriented services for software components created by outside developers. In view of this, there is a growing importance of such problems as IS security, protecting IS resources and software from unauthorized access and distribution. Dishonest users being qualified and having the provided tools at their disposal, can abuse software technologies for adaptable systems.

According to the approach presented in the article, IS software functioning in distributed environment is considered as an integrated software product. Complex protection of IS software and IS defining data and metadata is necessary. Complex protection implies IS data and program code protection and choosing the best licensing scheme.

IS are traditionally considered as composite software complexes, the components of which are set up on network nodes and interconnect via data transfer through communications links. This view of IS generated a congruent approach to IS security management. The approach implies exploiting various security mechanisms to protect network nodes from unauthorized access and resource usage (particularly, protection from malware including different viruses and Trojan software). Besides, this approach includes network interaction channels protection with a range of hardware and software tools (for example, shielding, network traffic analysis, etc). Such an approach to security organization is quite applicable and reasonable for protection of software systems found on separate workstations or within a small network. However, in case of distributed IS that go beyond the bounds of a separate PC or a small local area network, a number of significant disadvantages of the traditional approach can be pointed out:

- *It is difficult to maintain IS security at a proper level*. Since most services for information security support (such as Symantec Intruder Alert, family of ISS RealSecure systems and others) are focused on signature methods of intrusion detection, they require regular updating at all IS nods. This investigation is devoted to the approach to security system management which does not deny services of this kind, but allows dependence on them to be reduced.
- *Information systems are considerably vulnerable to new types of intrusions*, for instance to those based on detection and usage of the IS vulnerabilities that have not been abused yet, including operation system and network software vulnerabilities. The reason for this is signature methods of analysis predominating in today's market of information security systems.
- *There is practically no protection from intrusions that were specifically worked out for hacking a certain IS*. These intrusions are based, particularly, on vulnerabilities and errors in program realization of IS modules. Consequently, there is a need in additional protection agaist attacks performed "within" the IS, with the use of previously hacked modules.
- *There is a need in additional tools to control input and output dataflow*. Input dataflow control implies protection against spam, phishing, malicious ad-ware and other similar external threats. Output dataflow control involves scanning all outcoming information transferred to external systems, thus detecting the protected corporative information.

- *It is difficult to provide a sufficient user authentication level*. In large IS containing protected data and services, it is inefficient to use a standard user identity check method which is based on checking the knowledge of a secret password. In this case there is a risk of information leakage.

The traditional approach to distributed IS security has a lot of drawbacks, because most security tools *do not utilize information about the structure and semantics of the IS under protection*.

Security System Architecture

In this investigation a conceptually new attitude to security system design is suggested. Information system is not considered as a complex of computation nodes that interact while functioning. It is rather regarded as a complex of services provided by the IS components which are implimented in a number of interrelated network nodes. It should be pointed out here that IS as a *complex of services* requires total security instead of protecting separate structure units and data channels.

The approach under consideration is justified due to large IS' enhancement tendency which aims to provide users with daily necessary functions within a single integrated system (for example, such common operations as entering and editing data within business processes automated by IS components, exporting and importing e-mail, data processing and report generating, etc).

The authors suggest an approach to designing an integrated security system for protecting dynamically adapted distributed information systems, based on using metadata defining all aspects of IS functioning. The architecture of an integrated security system includes several levels. It combines various security tools integrated with the IS being protected, using this IS defining metadata.

The suggested security system is a multi-level complex, designed on a *multi-agent system* (MAS) basis. The complex combines the functionality of modern *intrusion detection systems* (IDS) and the tools of *IS structure and program logic security*.

The security system is based on *distributed MAS*. System agents' community is *closed* and *protected* against malicious influence from the outside with the help of its own security mechanisms as well as the way the agents' work is organized. *Each agent is an independent entity that covertly functions within the system under protection*. The information about the agents can be found nowhere in the system beyond the agents community which operate in the system and in the threads where they are run. Hiding is implemented by means of two main methods: agent execution in the threads hidden in the OS core with the help of a security driver, and agent execution in the thread of the protected system with the help of the thread context switching mechanism, actively used by the operating system [2].

All the MAS agents fall into two classes: *analyzer agents* and *sensor agents*. Analyzer agents are intellectual agents built on the basis of the InteRRaP architecture that belongs to the class of multi-layer architectures with vertical layer division [4]. Each layer implements a certain type of an agent's interaction with the environment (system area where this agent operates). The current system-status information is transferred from the lower layers to the higher layers, control is transferred from the higher to the lower layers.

The agent structure is represented by the 3 layers:

- *The layer of behavior* is responsible for reactivity, real-time behavior. The responsibility area of this layer is decision-making under typical circumstances, the examples of which can be user registration, remote network node connection or an intrusion attempt of a known type, the signature (i.e. script) of which is already in the system.
- *The layer of planning (local planning)* – is realization of the *cognitive paradigm* of MAS building. As the information is transferred from the layer of behavior, there is the agent's knowledge base inference on the layer of planning. The aim of this process is to evaluate the class of the current situation and to choose an adequate behavior template (set of responses to the changes in the environment state) in order to further apply it on the level of behavior.
- *The layer of communication (collective planning)* is responsible for realizing the *mechanisms of agents' communication*. This layer represents the possibility of decision-making on the basis of the data arranged by the other agents of the system. It also in charge of controlling team work.

The knowledge used by the agents to evaluate situations is introduced in the *frame paradigm* [3], the rules for decision-making are presented on the condition-action basis.

Sensor agents are used to collect the data about the current state of the security system, of the information system and its modules, as well as the network in which the IS is functioning. These agents are implemented on the basis of Reactive Architecture [4], they serve to collect statistics, post events and detect anomalies on the basic level.

Agents interact on the Contract Net model that implies solving different tasks directing them to the most suitable agents. This model was chosen since it has a number of advantages that make it comply with the requirements as fully as possible:

- Each agent has a functionality system that allows performing some tasks involving no other agents of the system (high level self-efficiency of agents).
- There is a small time interval from when a problem appears to when the process of solving starts.
- There is little possibility of incorrect problem-solving since problems are directed to more competent agents that contain all the necessary functionality.
- There is less overhead expences, as there is no need in every agent regularly analyzing the current system status.
- The high efficiency of system control results from the agents being subject to arrangement into hierarchic structures.

Levels And Mechanisms of Security

Security of Structure and IS program logics is an indispensable part of protection, they are unreasonably ignored by modern information security systems due to the fact mentioned above – they contain no information about semantics of the IS they protect. *IS structure security* serves to prevent hacking program modules from replacing server-side and client-side components of IS, as well as to raise the efficiency of protection measures against unauthorized connections to the IS services. *Program logics protection* is aimed at restraining unauthorized attempts of IS program code modification, which can be intended to error injection for building back doors [5] to arrange subsequent intrusions.

Information about the protected IS semantics is introduced by means of a *hierarchic 3-layer model* that fully describes all the security-critical aspects of the IS.

All the information on IS functioning, and its application environment is distributed among the three layers of the system security model $S = (Str, Ev, Msg)$ where

– *The Layer of structures Str* contains description of the distributed IS structure, including information about network nodes and application domains (IS subsystems), communication channels through which subsystems interconnect. In the model, the level of structures is presented by *P-graph (graph with poles [7]) Str = (N, A)*, where N is a set of vertices with poles representing application domains, network nodes; A is a set of arcs connecting them and representing communication channels.

– *The Layer of events Ev* = $\{T, E, Q, Init(Q), Init(E), Ch, Sch\}$, where T is a set of time moments, E is a finite set of events; $Init(Q): T \to Q$ is a mapping of the initial state; $Init(E): T \to E \times T$ is a mapping of the initial event planning; $Ch: E \times Q \times T \to Q$ is a mapping determining the new state to which the system changes as a result of an event; $Sch: E \times Q \times T \to E \times T$ is planning ratio that represents cause-and-effect relations of the events. The layer of events displays IS operation description in time. This layer comprises data on different states the system may be in, and events causing change of states. Representation of this layer in IS model is a *directed graph*, the vertices of which correspond to IS states at different instants of time. The arcs show events (including those related to receiving messages), causing change of states. This set can be tied with each vertex of structure Str.

– *The layer of messages* Msg comprises description of data which can be shared by the subsystems of IS, and rules for this data conversion. The layer is specified as determination of the layer of events.

The security system is also a multilevel one, it includes the following levels: basic security logics level, privilege control level, inherent security level, system security level.

The multilevel approach to security engineering, above all, makes it possible to *independently design various protection mechanisms*. Particularly, it has become possible to put the “high-level” security logics into practice (for example, activation entry checking), on the ground that it is supposedly impossible for an intruder to modify the program code performing these functions, because the code is protected at another level.

At the level of *basic security logic* essential functionality is realized, irequired from intrusion detection systems in accordance with ISO 15408 Standard: network traffic control, information system services control, anomaly detection.

The main mechanism of this level is an *active audit* subsystem which realizes the statistical and signature approach to activity detection and analysis. These approaches are described in FAU_SAA Security Audit Analysis requirements. The leading function of the the active audit subsystem is detecting anomalies in IS operation. Generally, any intrusion attempt is an anomaly pinpointed by means of a statistical analysis of the IS operation for a long time period. Top priority is given to analyzing how different services operate at the IS server modules.

There is a need in compensating the disadvantages of the statistical approach to analyzing such activities as complex decision-making in the context of the lack of an established empirical facts basis, and difficulty of attack detection in case activity parameters are gradually modified towards those typical of an attack. Aiming at the compensation, within an active audit subsystem, one applies the signature method of malicious activity detection that agrees with FAU_SAA Complex Attack Heuristics requirements.

In this context, signature is understood as a certain sequence of events, characteristic of a system cracking attempt. Efficiency of active audit mechanism is achieved by using the possibilities of a distributed multi-agent system being the basis of the security system. Information sent by sensor agents from various nodes of the network, is received by analyzer agents which are responsible for this data processing and forming the summary of the current system status and its potential security threats. The analysis is produced on the basis of the hierarchic three-layer IS model which also contains information about the security system itself.

The level of privilege control has a function that supports control of the system users' rights based on the stored profiles of activity, according to FAU_SAA.2 Profile Based Anomaly Detection requirements. As a rule, the aim of attacking large corporative IS is obtaining access to confidential data or protected services. It eventually means obtaining high level of privileges in the attacked system [5]. The main mechanism of this level is users' activity analysis subsystem.

In IS, some *user groups* are distinguished, each of them possessing a definite *privilege set*. During the process of the IS configuring and testing, statistical data about activity types of the use groups is collected, and group models represented by activity graphs, are formed. A group model includes the information characterizing the behavior of the user group members when logging into the system, working in the system and logging out of the system. After the group models are built, an individual model is constructed for each user.

The behavior model is a directed graph $G = \{V, A\}$ where $V = \{v_i\}$ is a set of vertices in which the order relation is defined according to the following rule: the element included in set V last, has a higher number in it; $A = \{a_{ij}\}$ is a set of arcs of graph G. Each element $a_{ij} \in A$ is put in correspondence with some weight $w_{ij} \in W$, where W is a set of admissible weights of arcs. Vertices $v_i \in V$ represent values of the controlled parameters. Arcs $a_{ij} \in A$ represent semantic relations among the controlled parameters' values, characterizing the order of adding vertices matching parameter values, i.e. the elements $v_i \in V$, to graph G. Weights $w_{ij} \in W$, appointed to arcs $a_{ij} \in A$, fix semantic distances among the values of the controlled parameters by means of the corresponding vertices incident to these arcs $v_i, v_J \in V$. Semantic distance characterizes the difference among the values of the controlled activity parameter. The model allows controlling the correspondence of parameters to some reference values, "accumulating" the changes for the subsequent analysis.

The models are based on the analysis of different types of *users' activity parameters*:

- *Categorial parameters*. The examples of categorial parameters can be changed files, records in the database, IS services in usage, initiated commands, types of errors, etc. Categorial parameters analysis has an *event-oriented* character.
- *Numeric parameters*. This type comprises any activity parameters, which can be valued numerically - for instance, the quantity of transmitted and requested information, the number of services being in use simultaneously, as well as the number of vertices and arcs of the model.
- *Intensity parameters*. For example, the number of the user's entries into the system during a certain period of time, intensity of the database queries, and the like.
- *Event distribution* parameters. This type may include the frequency ratio of such events as view query and change query, references to certain IS services.

The models are mainly applied via realizing authentication mechanism based on correlating the current user's behavior with statistics on his usual activity parameters. This mechanism is an *addition to the standard authentication mechanisms*; it is aimed at protection from unauthorized access to privileges via the legal users' identity theft.

Individual users' models and group models in dynamically configurable IS can be also utilized for the purposes that are not related to security; for example, user interface automatic generation and configuration, based on statistics of an IS functionality being applied by the user or user group.

The security levels described above are projected, according to the statement that it is impossible for an intruder to modify the program code. On the *inherent security* level some mechanisms are applied to protect the IS program code from analysis and modification.

The key mechanisms of this level are:

- The mechanism of *explicit program code entity control*: it initiates an instant reaction of the security system. Within this mechanism, the application program code is checked for unauthorized changes, and cryptographic security of program modules is realized.
- The mechanism of *implicit control* is used to arrange the deferred system reaction to intrusion, with the purpose of preventing the cracked application from being used. As it becomes evident that an intruder modified the program code or deactivated either security mechanisms of the first levels or the explicit control mechanism, the security system is switched to imitation mode. However, there are no signs of attack detection, but the IS modules that had been abused, get actually isolated, i.e. it is impossible to use them for accessing the key data and IS services.
- The mechanism of *concealing the location of the security system functions*. This mechanism is chiefly aimed at the functions that are responsible for the user feedback. For instance, protected service lockout displays messages on access restriction in case attacks are detected (the so called nag screens). Feedback functions generating this kind of messages are in most cases a convenient starting point for the system hack [8]. Concealment is performed by exporting all the vulnerable functions to the dynamically generated program modules. Besides, the functions generating "dangerous" messages are not saved in the application files, it is difficult enough to detect and modify them.

System security level. The majority of malicious programs can not function without obtaining certain privileges which give access to protected system functions. System functions access is necessary for such tasks as opening network ports (e.g. for interaction with a trojan module installed in the attacked system), executing programs in debug mode (in order to find security breaches), access to protected external memory partitions or address spaces of the executed programs as well as to input/output controllers. Ideally, the code becomes available for execution at ring 0 privilege level. It allows a direct access to any resources of the attacked system, including functions of the operating system kernel and physical units. Kernel level security serves to prevent intruders from access to protected OS functions and OS kernel in particular.

The *main mechanisms* of this level are: the mechanism of processes detection, the mechanism of network interaction control, the mechanism of ring 0 security.

The mechanism of *network interaction control* analyzes network ports status in order to disclose unauthorized attempts to open new ports and change running modes of the active ports.

This mechanism is applied by tracking calls of the corresponding OS kernel functions (it is Native API [8] for Windows operating system), it is performed by means of installing shells, realizing callback interfaces, on these functions. Kernel calls tracking is a sufficient condition for detecting unauthorized access to the OS functions which are potentially dangerous in the context of secure operation management. The reason is that calling any function of application interfaces leads to calling one of the OS kernel functions. Besides, in most cases one OS kernel function comes with several different application interface functions, which are in fact its shells making kernel function calls with a certain set of parameters [8]. The kernel level control can only guarantee security that doesn't depend on possible appearance of new program interfaces and new ways of access to potentially dangerous OS functions.

Security mechanism of the ring 0 privilege level protects the functions performed on the ring 0 privilege level of the system. The greatest security threat is presented by the so-called hacking tool kits [5] – rootkits – that work on the ring 0 privilege level. It is a sort of malware that enables the intruder to obtain almost full control over the infected system and isn't practically subject to detection and liquidation.

A rootkit can be realized in the form of a separate driver or a shell of some OS kernel function. Rootkit intrusion is prevented by *controlling OS kernel function calls* that are responsible for drivers and images uploading to the system. Detecting rootkits installing shells in the OS kernel functions is not technically difficult, as possessing information about the initial structure of the kernel functions is enough to detect unauthorized modification. Within this mechanism, there is *regular verification of the hash functions values* (*that is computed from the program code of the OS kernel functions*) to correspond to the reference values. These values are derived after security system setting-up and authorized changes in the functions. An additional sanction can be tracking attempts to memory access based on addresses matching the OS kernel functions. However, it will inevitably lead to a noticeable decrease in the secured system's productivity, that's why this sanction can be applied only in cases when the maximum level of security is required.

Hidden processes detection mechanism (of the processes invisible on the application level) is aimed at detecting malware that is able to operate on the application level. Hidden processes are disclosed with the help of a *security driver* operating on the system level.

Conclusion

The main efforts of the suggested security system are:

- *Adaptability*. The suggested security system can be adapted to new threats via modification of its knowledge base.
- *Universality*. The suggested security system is based on a multilayered model of the protected IS and therefore can be integrated to nearly any information system.
- *Extensibility*. The suggested security system is knowledge-based therefore its functionality can be extended even without providing changes in its structure or source code.
- *High performance*. Metaknowledge and knowledge on protected IS are used to maximize security system performance.

References

[1] Лядова Л.Н. Архитектура информационной системы «Образование Пермской области» // Математика программных систем: Межвузовский сборник научных трудов / Перм. ун-т. Пермь, 2002. С. 25-35.

[2] Кастер Х. Основы Windows NT и NTFS / Пер. с англ.— М.: Издательский отдел «Русская редакция» ТОО «Channel Trading Ltd.», 1996.

[3] Минский М. Фреймы для представления знаний. М.: Энергия, 1979.

[4] Huhns M., Stephens L. Multiagent Systems and Societies of Agents // Weiss G. Multiagent systems: a modern approach to a distributed artificial intelligence / Massachusetts Institute of Technology.

[5] Хогланд Г., Мак-Гроу Г. Взлом программного обеспечения: анализ и использование кода. М.: Вильямс, 2005.

[6] Лядова Л.Н., Мороз А.А. Модель защиты программного обеспечения от несанкционированного распространения // В кн.: Сборник трудов Второй международной научно-технической конференции «Инфокоммуникационные технологии в науке, производстве и образовании» (Инфоком 2) / Кисловодск, 2006. С. 120-124.

[7] Миков А.И. Автоматизация синтеза микропроцессорных управляющих систем. Иркутск: Изд-во Иркут. ун-та, 1987.

[8] Касперски К. Техника и философия хакерских атак. М.: СОЛОН-Пресс, 2004.

Authors' Information

Denis Kurilov – Perm State University, Graduate student of the Computer Science Department; Bukirev St., 15, Perm-614990, Russia; e-mail: Denis.Kurilov@gmail.com.

Lyudmila Lyadova– Institute of Computing, Deputy Director; Podlesnaya St., 19/2-38, Perm-614097, Russia; e-mail: LNLyadova@mail.ru.

ADVANCE OF THE ACCESS METHODS

Krassimir Markov, Krassimira Ivanova, Ilia Mitov, Stefan Karastanev

Abstract: The goal of this paper is to outline the advance of the access methods in the last ten years as well as to make review of all available in the accessible bibliography methods.

Keywords: Access Methods, Overview of the Access Methods

ACM Classification Keywords: D.4.3 File Systems Management, Access methods

Introduction

The Access Methods (AM) had been available from the beginning of the developing the computer peripheral devices. As many devices there exists so many possibilities for developing different AM we have. Our attention is focused only to the access methods for devices for permanently storing the information with direct access such as magnetic discs, flash memories, etc.

In the beginning, the AM ware functions of the Operational Systems Core or so called Supervisor, and ware executed via corresponded macro-commands in the assembler languages [Stably, 1970] or via corresponding input/output operators in the high level programming languages like FORTRAN, COBOL, PL/I, etc.

Establishing of the first data bases in the 60-ties years of the last century caused gradually accepting the concepts "physical" as well as "logical" organization of the data [CODASYL, 1971], [Martin, 1975]. In 1975 the concepts "access method", "physical" and "logical" are clearly separated. In the same time Christopher Date [Date, 1977] specially remarked:

"The Data Base Management System (DBMS) does not know anything about:

a) physical records (blocks);
b) how the stored fields are integrated in the records (nevertheless that in many cases it is obviously because of theirs physical disposition);
c) how the sorting is realized (for instance it may be realized on the base of physical sequence, using an index or by a chain of pointers);
d) how is realized the direct access (i.e. by index, sequential scanning or hash addressing).

This information is a part of the structures for data storing but it is used by the access method but not by the DBMS. "

Every access method presumes an exact organization of the file which it is operating with and has no relation to the interconnections between the files, respectively – between the records of one file and that in the others files. These interconnections are controlled by the physical organization of the DBMS.

So, in the DBMS we may distinguish four levels:

- access methods of the core (supervisor) of the operation system;
- specialized access methods which upgrade these of the core of the operating system;
- physical organization of the DBMS;
- logical organization of the DBMS.

During the 80-ies years the "Multi-Dimensional Access Methods" had raised. In accordance with them the corresponded "spatial information structures" and the "spatio-temporal information structures" had risen, too. These AM developed the methods of the operating systems via specializing them to the give data models. From different point of view this period had been presented in [Ooi et al, 1993], [Gaede, Günther, 1998], [Arge, 2002], [Mokbel et al, 2003], [Moënne-Loccoz, 2005].

Usually the "one-dimensional" (linear) AM are used in the classical applications, based on the alpha-numerical information, whereas the "multi-dimensional" (spatial) methods are aimed to serve the work with graphical, visual, multimedia information. Now a special attention is given to the multi-dimensional AM.

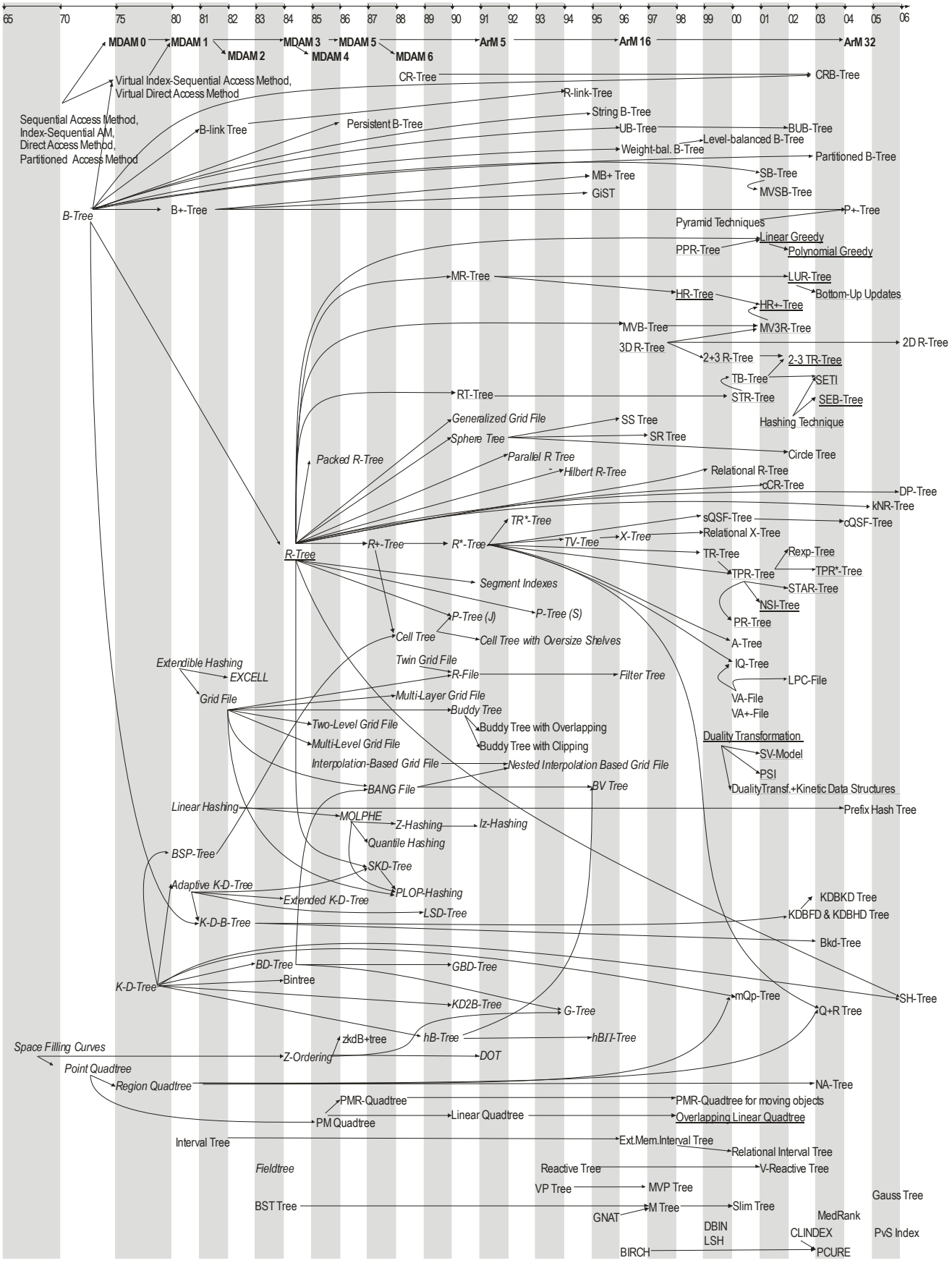


Fig. 1. Genesis of the Access Methods and their modifications
extended variant of [Gaede, Günther, 1998] and [Mokbel et al, 2003]

Maybe one of the most popular analyses is given in [Gaede, Günther, 1998]. The authors presented a scheme of the genesis of the basic multi-dimensional AM and theirs modifications. This scheme firstly was proposed in [Ooi et al, 1993] and it was expanded in [Gaede, Günther, 1998]. An extension in direction to the multi-dimensional spatio-temporal access methods was given in [Mokbel et al, 2003].

This work continues the investigation provided in [Markov, 2006]. The main goal of this paper is to present a new variant of this scheme. It is presented on Fig.1. In it the new access methods created after 1998 are added. The methods presented in [Gaede, Günther, 1998] are marked in italics and methods presented in [Mokbel et al, 2003] are underlined. Access methods, which are given in the two surveys simultaneously, are marked in underlined italics. In the appendix of this paper the corresponded bibliography is given.

The access methods presented on Fig.1 we may classify as follow:

- One-dimensional AM;
- Multidimensional Spatial AM;
- Metric Access Methods;
- High Dimensional Access Methods;
- Spatio-Temporal Access Methods.

One-dimensional Access Methods

One-dimensional AM are based on the concept "record". Let remember that the "record" is a logical sequence of fields which contain data eventually connected to unique identifier (a "key"). The identifier (key) is aimed to distinguish one sequence from another [Stably, 1970]. The records are united in the sets, called "files". There exist three basic formats of the records – with fixed, variable and undefined length.

In the *context-free methods* the storing of the records is not connected to theirs content and depends only on external factors – the sequence, disk address or position in the file. The necessity of stable file systems in the operating systems does not allow a great variety of the context-free AM. There are three main types well known from 60-ies and 70-ies years: *Sequential Access Method (SAM); Direct Access Method (DAM) and Partitioned Access Method (PAM)* [IBM, 1965-68].

The main idea of the *context-depended AM* is that the part of the record is selected as a key which is used for making decision where to store the record and how to search it. This way the content of the record influences on the access to the record.

Historically, from the 60-ies years of the last century the attention is directed mainly to this type of AM. Modern DBMS are built using context-depended AM such as: unsorted sequential files with records with keys; sorted files with fixed record length; static or dynamic hash files; index file and files with data; clustered indexed tables [Connolly, Begg, 2002].

Multidimentional Spatial Access Methods

Multidimensional Spatial Access Methods are developed to serve information about spatial objects, approximated with points, segments, polygons, polyhedrons, etc. The implementations are numerous and include traditional multi-attributive indexing, geographical information systems and spatial databases, content indexing in multimedia databases, etc.

From the point of view of the spatial databases can be split in two main classes of access methods – Point Access Methods and Spatial Access Methods [Gaede, Günther, 1998].

Point Access Methods are used for organizing multidimensional point objects. Typical instance are traditional records, where on every attribute of the relation corresponds one dimension. These methods can be separated in three basic groups:

- Multidimensional Hashing (for instance Grid File and its varieties, EXCELL, Twin Grid File, MOLPHE, Quantile Hashing, PLOP-Hashing, Z-Hashing, etc);
- Hierarchical Access Methods (includes such methods as KDB-Tree, LSD-Tree, Buddy Tree, BANG File, G-Tree, hB-Tree, BV-Tree, etc.);
- Space Filling Curves for Point Data (like Peano curve, N-trees, Z-Ordering, etc).

Spatial Access Methods are used for working with objects which have arbitrary form. The main idea of the spatial indexing of non-point objects is using of the approximation of the geometry of the examined objects to more

simple forms. The most used approximation is Minimum Bounding Rectangle (MBR), i.e. minimal rectangle, which sides are parallel of the coordinate axes and completely include the object. There exist approaches for approximation with Minimum Bounding Spheres (SS Tree) or other polytopes (Cell Tree), as well as their combinations (SR-Tree).

The usual problem when one operates with spatial objects is their overlapping. There are different techniques to avoid this problem. From the point of view of the techniques for organization of the spatial objects Spatial Access Methods can be split in four main groups:

- Transformation – this technique uses transformation of spatial objects to points in the space with more or less dimensions. Most of them spread out the space using space filling curves (Peano Curves, z-ordering, Hibert curves, Gray ordering, etc.) and then use some of point access method upon the transformed data set. For instance UB-Tree [Bayer, 1996], is variant of B-Tree, where keys are region addresses, sorted via "≤" and z-ordering;
- Overlapping Regions – here the data set are separated in groups; different groups can occupy the same part of the space, but every space object associates with only one of the groups. The access methods of this category operate with data in their primary space (without any transformations) eventually in overlapping segments. Methods, which use this technique includes R-Tree, R-link-Tree, Hilbert R-Tree, R*-Tree, Sphere Tree, SS-Tree, SR-Tree, TV-Tree, X-Tree, P-Tree of Schiwietz, SKD-Tree, GBD-Tree, Buddy Tree with overlapping, PLOP-Hashing, etc.;
- Clipping – this technique use eventually clipping of one object to several sub-objects, which will be stored. The main goal is to escape overlapping regions. But this advantage can lead tearing of the objects, extending of the resource expenses and decreasing of the productivity of the method. Representatives of this technique are R+-Tree, Cell-Tree, Extended KD-Tree, Quad-Tree, etc.;
- Multiple Layers – this technique can be examining as variant of the techniques of Overlapping Regions, because the regions from different layers can overlap. But there exist some important differences: first – the layers are organizing hierarchically; second – every layer split primary space in different way; third – the regions of one layer never overlaps; fourth – the data regions are separated from space extensions of the objects. Instances for these methods are Multi-Layer Grid File, R-File, etc.

Metric Access Methods

Metric Access Methods deal with relative distances of data points to chosen points, named anchor points, vantage points or pivots [Moënne-Loccoz, 2005]. These methods are designed to limit the number of distance computation, calculating first distances to anchors, and then finding searched point in narrowed region. These methods are preferred when the distance is highly computational, as e.g. for the dynamic time warping distance between time series. Presentatives of these methods are: Vantage Point Tree (VP Tree), Bisector Tree (BST-Tree), Geometric Near-Neighbour Access Tree (GNNAT), as well as the most effective from this group – Metric Tree (M-Tree) [Chavez et al, 2001].

High Dimensional Access Methods

Increasing of the dimensionality strongly aggravates the qualities of the multidimensional access methods. Usually these methods exhaust their possibilities till dimensions around 15. Only X-Tree reaches the boundary of 25 dimensions, after then this method gives worse results then sequential scanning [Chakrabarti, 2001].

The exit of this situation is based on the data approximation and query approximation in sequential scan. These methods form a new group of access methods – High Dimensional Access Methods.

Data approximation is used in VA-File, VA+-File, LPC-File, IQ-Tree, A-Tree, P+-Tree, etc.

Because in high dimensional access methods the selectivity of the methods makes worse, it is allowed some answers inaccuracy. For query approximation two strategies can be used:

- examine only a part of the database, which is more probably to contain resulting set – as a rule these methods are based on the clustering of the database. Some of these methods are: DBIN, CLINDEX, PCURE;
- splitting the database to several spaces with fewer dimensions and searching in each of them. Here two main methods are used:

1) Random Lines Projection (presentatives of this approach are MedRank, which uses B+-Tree for indexing every arbitrary projection of the database, and PvS Index, which consist of combination of iterative projections and clustering);

2) Locality Sensitive Hashing, which is based on the set of local-sensitive hashing functions [Moënne-Loccoz, 2005].

Spatio-Temporal Access Methods

The Spatio-Temporal Access Methods have additional defined time dimensioning. [Mokbel et al, 2003]. They operate with objects, which change their form and/or position during the time. According to position of time interval in relation to present moment the Spatio-Temporal Access Methods are divided to:

- indexing the past, i.e. methods for operating with historical spatio-temporal data. The problem here is continuously increasing of the information over time. To overcome the overflow of the data space two approaches are used – sampling the stream data at certain time position or update the information only when data is changed. Spatio-temporal indexing schemes for historical data can be split in three categories: first category includes methods that manages spatial and temporal aspects into already existing spatial methods; second can be explained as snapshots of the spatial information in each time instance; the third category focus on trajectory-oriented queries, while spatial dimension lag on second priority. Presentatives of this group are: RT-Tree, 3DR-Tree, STR-Tree, MR-Tree, HR-Tree, HR+-Tree, MV3R-Tree, PPR-Tree, TB-Tree, SETI, SEB-Tree;
- indexing the present. In contrast to previous methods, where all movements are known, here current positions are neither stored nor queried. Some of the methods, which answer of the questions of the current position of the objects are 2+3R-Tree, 2-3TR-Tree, LUR-Tree, Bottom-Up Updates, etc.;
- indexing the future. These methods have to answer on the questions about current and future position of moving object – here are embraced the methods like PMR-Quadtree for moving objects, Duality Transformation, SV-Model, PSI, PR-Tree, TPR-Tree, TPR*-tree, NSI, VCIR-Tree, STAR-Tree, R^{EXP}-Tree.

Conclusion

In this paper we presented a short overview of the current state in the field of development of the access methods. During the last four decades the access methods have been developed toward plenty of modifications of small number basic ideas. It is important to remark that the research has been provided on software as well as on hardware levels. For instance, in [Schlosser et al, 2005] a technology for storing of multi-dimensional data with physically preserving the multi-dimensionality of the data is presented

The developed multi-dimensional index structures are effective for the small number of dimensions (from 2 – 5 up to 10 -15) and are uncomfortable for multi-dimensional spaces which are typical for the contemporary practical problems and the linear scanning may be preferable in many cases [Chakrabarti, 2001]. This is known as "Curse of dimensionality".

The concept of ”curse of dimensionality” was first coined by Richard Bellman [Bellman 1961]. He employed it to describe the problem caused by the exponential increase of the volume with the augmentation of the space dimension when addressing the problem of optimizing functions with several variables. Later, the term was used to indicate, more generally, non-intuitive phenomena observed when the dimension of data increases [Bouteldja et al, 2006].

The survey of the access methods suggests that the context-free multi-dimensional access methods practically are not available. One step in developing such methods is the Multi-domain Access Method introduced in [Markov, 2004].

We have no place to present all access methods in details. The main goal was to collect the basic publications of the most popular access methods. The further survey needs to be provided to present current state of the art in this area.

Appendix 1. Access Methods and Corresponded Publications

Access Method	Publicated in
2+3 R-Tree	**[Nascimento, 1999] M. A. Nascimento, J.R.O. Silva, Y. Theodoridis.** *Evaluation of Access Structures for Discretely Moving Points*. **In Proc. of the Intl. Workshop on Spatio-Temporal Database Management, STDBM, pages 171–188, Sept. 1999.**
2-3 TR-Tree	**[Abdelguerfi et al, 2002] M. Abdelguerfi, J. Givaudan, K. Shaw, R. Ladner.** *The 2-3 TR-tree, A Trajectory-Oriented Index Structure for Fully Evolving Valid-time Spatio-temporal Datasets*. **In Proc. of the ACM workshop on Adv. in Geographic Info. Sys., ACM GIS, pages 29–34, Nov. 2002.**
2D R-Tree	**[Osborn, Barker, 2006] W. Osborn, K. Barker.** *Searching through Spatial Relationships using the 2DR-tree*. **The IASTED Conference on Internet and Multimedia Systems and Applications Honolulu, Hawaii, USA August 14-16, 2006**
3D R-tree	**[Theodoridis et al, 1996] Y. Theodoridis, M. Vazirgiannis, T. Sellis.** *Spatio-Temporal Indexing for Large Multimedia Applications*. **In Proc. of the IEEE Conference on Multimedia Computing and Systems, ICMCS, June 1996.**
Adaptive K-D-Tree	**[Bentley, Friedman, 1979] J. L. Bentley, J. H. Friedman.** *Data structures for range searching*. **ACM Comput. Surv. 11, 1979, 4, 397–409.**
A-Tree (Approximation Tree)	**[Sakurai et al, 2000] Y. Sakurai, M. Yoshikawa, S. Uemura, H. Kojima.** *The a-tree: An index structure for high-dimensional spaces using relative approximation*. **In VLDB, pages 516–526, 2000.**
B+-tree	**[Comer, 1979] D. Comer.** *The ubiquitous B-tree*. **ACM Comput. Surv. 11, 2, 1979, 121–138.**
Balanced Multidimensional Extendible Hash Tree	**[Otoo, 1985] E.J. Otoo.** *Balanced multidimensional extendible hash tree*. **In Proceedings of the fifth ACM SIGACT-SIGMOD symposium on Principles of database systems, Cambridge, Massachusetts, United States, 1985, Pages: 100 – 113**
BANG File	**[Freeston, 1987] M. Freeston.** *The BANG file: A new kind of grid file*. **In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1987, pp. 260–269.**
BD-Tree	**[Ohsawa, Sakauchi, 1983] Y. Ohsawa, M. Sakauchi.** *BD-tree: A new n-dimensional data structure with efficient dynamic characteristics*. **In Proceedings of the Ninth World Computer Congress, IFIP 1983, 1983, pp. 539–544.**
Bintree	**[Tamminen, 1984] M. Tamminen.** *Comment on quad- and octrees*. **Commun. ACM 30, 3, 204–212. 1984**
BIRCH	**[Zhang et al, 1996] T. Zhang, R. Ramakrishnan, M. Livny.** *BIRCH: an efficient data clustering method for very large databases*. **pages 103–114, 1996.**
Bkd-Tree	**[Procopiuc et al, 2003] O. Procopiuc, P. K. Agarwal, L. Arge, J.-S. Vitter.** *Bkd-tree: A Dynamic Scalable kd-tree*. **In Proceedings of International Symposium on Spatial and Temporal Databases, 2003**
B-link Tree	**[Lehman, Yao, 1981] P. Lehman, S. Yao.** *Efficient locking for concurrent operations on B-trees*. **ACM Trans. Database Syst. 6, 4, 1981, 650–670.**
Bottom-up Updates	**[Lee et al, 2003] M. Lee, W. Hsu, C. Jensen, B. Cui, K. Teo.** *Supporting Frequent Updates in R-Trees: A Bottom-Up Approach*. **In Proc. of the Intl. Conf. on Very Large Data Bases, VLDB, Sept. 2003.**
BSP-Tree	**[Fuchs et al, 1980] H. Fuchs, Z. Kedem, B. Naylor.** *On visible surface generation by a priori tree structures*. **Computer Graph. 14, 3, 1980.**
BST-Tree (Bisector Tree)	**[Kalantari, McDonald, 1983] I. Kalantari, G. McDonald.** *A data structure and an algorithm for the nearest point problem*. **IEEE Trans. Software Eng., 9(5):631–634, 1983.**
B-Tree	**[Bayer, McCreicht, 1972] R. Bayer, E. M. McCreicht.** *Organization and maintenance of large ordered indices*. **Acta Inf. 1, 3, 1972, pp. 173–189.**
BUB-Tree (Bounding UB Tree)	**[Fenk, 2002] R. Fenk.** *The BUB-Tree*. **In Proceedings of VLDB Conf. Hongkong, 2002**
Buddy Tree	**[Seeger, Kriegel, 1990] B. Seeger, H.-P. Kriegel.** *The buddy-tree: An efficient and robust access method for spatial data base systems*. **In Proceedings of the Sixteenth International Conference on Very Large Data Bases, 1990, pp. 590–601.**
Buddy Tree with Clipping	**[Seeger, 1991] B. Seeger.** *Performance comparison of segment access methods implemented on top of the buddy-tree*. **In Advances in Spatial Databases, O. Gu¨ nther and H. Schek, Eds., LNCS 525, Springer-Verlag, Berlin/Heidelberg/New York, 1991, 277–296.**
Buddy Tree with Overlapping	**[Seeger, 1991] B. Seeger.** *Performance comparison of segment access methods implemented on top of the buddy-tree*. **In Advances in Spatial Databases, O. Gu¨ nther and H. Schek, Eds., LNCS 525, Springer-Verlag, Berlin/Heidelberg/New York, 1991, 277–296.**
Buffer Tree	**[Arge, 1995] L. Arge.** *The buffer tree: a new technique for optimal I/O-algorithms*. **In Proc.Workshop on Algorithms and Data Structures, pages 334–345. LNCS 955. Springer-Verlag, Berlin, 1995.** **[Arge, 2003] Lars Arge.** *The Buffer Tree: A Technique for Designing Batched External Data Structures.* **Algorithmica, Springer-Verlag New York Inc. 2003**

BV Tree	**[Freeston, 1995] M. Freeston. A** *general solution of the n-dimensional B-tree problem*. **In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1995, pp. 80–91.**
cCR-tree (Cache-Conscious R-Tree)	**[Kim et al, 2001] K. Kim, S.K. Cha, K. Kwon.** *Optimizing multidimensional index trees for main memory access*. **International Conference on Management of Data Proceedings of the 2001 ACM SIGMOD international conference on Management of data, Santa Barbara, California, United States, 2001, Pp: 139 – 150**
Cell Tree	**[Günther, 1988] O. Günther.** *Efficient Structures for Geometric Data Management*. **LNCS 337, Springer-Verlag, Berlin/Heidelberg/New York. 1988.**
Cell Tree with Oversize Shelves	**[Günther, Noltemeier, 1991] O. Günther, H. Noltemeier.** *Spatial database indices for large extended objects*. **In Proceedings of the Seventh IEEE International Conference on Data Engineering, 1991, 520–526.**
Circle Tree	**[Moore, 2002] A.Moore.** *The circle tree – a hierarchical structure for efficient storage, access and multi-scale representation of spatial dat*a. **Presented at SIRC 2002 – The 14th Annual Colloquium of the Spatial Information Research Centre University of Otago, Dunedin, New Zealand, December 3-5th 2002**
CLINDEX	**[Li et al, 2002] C. Li, E. Chang, H. Garcia-Molina, G. Wiederhold.** *Clustering for approximate similarity search in high-dimensional spaces*. **IEEE Transactions on Knowledge and Data Engineering, 14(4):792–808, 2002.**
cQSF Tree (Scalable QSF Tree)	**[Orlandic, Yu, 2004] R. Orlandic, B. Yu.** *Scalable QSF-Trees: Retrieving Regional Objects in High-Dimensional Spaces*. **Journal of Database Management (JDM, IDEAS Group Publishing) Vol. 15⊗in press), 15-page, 2004**
CRB-Tree (Compressed Range B-Tree)	**[Govindarajan et al, 2003] S. Govindarajan, P. K. Agarwal, L. Arge.** *CRB-Tree: An Efficient Indexing Scheme for Range-Aggregate Queries*. **Proceedings of the 9th International Conference on Database Theory, 2003, Pp:143-157**
CR-Tree (Compressed Range Tree)	**[Chazelle, 1988] B. Chazelle.** *A functional approach to data structures and its use in multidimensional searching*. **SIAM J. Comput., 17(3):427–462, June 1988**
DBIN (Density Based Indexing)	**[Bennett et al, 1999] K. P. Bennett, U. Fayyad, D. Geiger.** *Density-based indexing for approximate nearestneighbor queries*. **In KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 233–243, New York, NY, USA, 1999. ACM Press.**
DOT	**[Faloutsos, Rong, 1991] C. Faloutsos, Y. Rong.** *DOT: A spatial access method using fractals*. **In Proceedings of the Seventh IEEE International Conference on Data Engineering, 1991, pp. 152–159.**
DP-Tree	**[Li et al, 2006] M. Li, W.-C. Lee, A. Sivasubramaniam.** *DPTree: A balanced tree based indexing framework for peer-to-peer systems*. **in Proceedings of the 14th International Conference on Network Protocols (ICNP 2006), pages 12-21, November, 2006**
Duality Transformation	**[Kollios et al, 1999] G. Kollios, D. Gunopulos, V. J. Tsotras. On Indexing Mobile Objects. In Proc. of the ACM Symp. on Principles of Database Systems, PODS, pages 261–272, June 1999.**
Duality Transformation with Kinetic Data Structure	**[Agarwal et al, 2000] P. K. Agarwal, L. Arge, and J. Erickson.** *Indexing Moving Points*. **In Proc. of the ACM Symp. on Principles of Database Systems, PODS, pages 175–186, May 2000.**
EXCELL (Extendible Cell)	**[Tamminen, 1982] M. Tamminen.** *The extendible cell method for closest point problems*. **BIT 22, 1982, pp. 27–41.**
Extended K-D-Tree	**[Matsuyama et al, 1984] T. Matsuyama, L.V. Hao, M. Nagao.** *A file organization for geographic information systems based on spatial proximity*. **Int. J. Comput. Vis. Graph. Image Process. 26, 3, 1984, pp. 303–318.**
Extendible Hashing	**[Fagin et al, 1979] R. Fagin, J. Nievergelt, N. Pippenger, R. Strong.** *Extendible hashing: A fast access method for dynamic files*. **ACM Trans. Database Syst. 4, 3, 1979, pp. 315–344.**
Fieldtree	**[Frank, 1983] A. Frank.** *Problems of Realizing LIS: Storage Methods for Space Related Data: The Field Tree*. **Technical Report 71, Institut for Geodesy and Photogrammetry, Swiss Federal Institut of Technology, Zurich, Switzerland, 1983.**
Filter Tree	**[Sevcik, Koudas, 1996] K. Sevcik, N. Koudas.** *Filter trees for managing spatial data over a range of size granularities*. **In Proceedings of the 22th International Conference on Very Large Data Bases (Bombay), 1996, pp. 16–27.**
Gauss-Tree	**[Bohm et al, 2006] C. Bohm, A. Pryakhin, M. Schubert.** *The Gauss-Tree: Efficient Object Identification in Databases of Probabilistic Feature Vectors*. **22nd Int. Conf. on Data Engineering (ICDE'06), Atlanta, GA, 2006**
GBD-Tree	**[Ohsawa, Sakauchi, 1990] Y. Ohsawa, M. Sakauchi.** *A new tree type data structure with homogeneous node suitable for a very large spatial database*. **In Proceedings of the Sixth IEEE International Conference on Data Engineering, 1990, pp. 296–303.**
Generalized Grid File	**[Blanken et al, 1990] H. Blanken, A. Ijbema, P. Meek, B. Van den Akker.** *The generalized grid file: Description and performance aspects*. **In Proceedings of the Sixth IEEE International Conference on Data Engineering, 1990, pp. 380–388.**
GiST (Generalizad Search Tree)	**[Hellerstein et al, 1995] J. M. Hellerstein, J. F. Naughton, A. Pfeffer.** *Generalized Search Trees for Database Systems*. **Proc. 21st Int. Conf. on Very Large Databases, September 1995, pp. 562-573.**

GNAT (Geometric Near-Neighbor Access Tree)	**[Brin, 1995] S. Brin.** *Near neighbor search in large metric spaces*. **In VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases, pages 574–584, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.**
Grid File	**[Nievergelt et al, 1981] J. Nievergelt, H. Hinterberger, K. Sevcik.** *The grid file: An adaptable, symmetric multikey file structure*. **In Proceedings of the Third ECI Conference, A. Duijvestijn and P. Lockemann, Eds., LNCS 123, Springer-Verlag, Berlin/Heidelberg/New York, 1981, pp. 236–251.**
G-Tree	**[Kumar, 1994] A. Kumar.** *G-tree: A new data structure for organizing multidimensional data*. **IEEE Trans. Knowl. Data Eng. 6, 2, 1994, pp. 341-347.**
Hana Tree	**[Kwon, Jeong, 2000] Y. Kwon, C. Jeong.** *Hana Tree: A Dynamic and Robust Access Method for Spatial Data Handling*. **Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 1846/2000. Proceedings of Web-Age Information Management: First International Conference, WAIM 2000, Shanghai, China, June 21-23, 2000.**
Hashing Technique	**[Song, Roussopoulos, 2001] Z. Song, N. Roussopoulos.** *Hashing Moving Objects*. **In Mobile Data Management, pages 161–172, Jan. 2001.**
hBP-Tree	**[Evangelidis et al, 1995] G. Evangelidis, D. Lomet, B. Salzberg.** *The hBP-tree: A modified hB-tree supporting concurrency, recovery and node consolidation*. **In Proceedings of the 21st International Conference on Very Large Data Bases, 1995, pp. 551–561.**
hB-Tree	**[Lomet, Salzberg, 1989] D.B. Lomet, B. Salzberg.** *The hBtree: A robust multiattribute search structure*. **In Proceedings of the Fifth IEEE International Conference on Data Engineering, 1989, pp. 296–304.**
Hilbert R-Tree	**[Kamel, Faloutsos, 1994] I. Kamel, C. Faloutsos.** *Hilbert R-tree: An improved R-tree using fractals*. **In Proceedings of the Twentieth International Conference on Very Large Data Bases, 1994, pp. 500–509.**
HR+-Tree	**[Tao, Papadias, 2001a] Y. Tao, D. Papadias.** *Efficient Historical R-trees*. **In Proc. of the Intl. Conf. on Scientific and Statistical Database Management, SSDBM, pages 223–232, July 2001.**
HR-Tree (Historical R-Tree)	**[Nascimento, Silva, 1998] M. A. Nascimento, J.R.O. Silva.** *Towards historical R-trees*. **In Proc. of the ACM Symp. on Applied Computing, SAC, pages 235–240, Feb. 1998.**
Interpolation-Based Grid File	**[Ouksel, 1985] M. Ouksel.** *The interpolation based grid file*. **In Proceedings of the Fourth ACM SIGACT –SIGMOD Symposium on Principles of Database Systems, 1985, pp. 20–27.**
Interval Tree	**[Edelsbrunner 1980] H. Edelsbrunner.** *Dynamic Rectangle Intersection Searching*. **Institute for Information. Processing Report 47, Technical University of Graz, Austria, 1980.**
IQ-Tree (Independent Quantization Tree)	**[Berchtold et al, 2000] S. Berchtold, C. Bohm, H. V. Jagadish, H.-P. Kriegel, J. Sander.** *Independent quantization: An index compression technique for high-dimensional data spaces*. **In ICDE '00: Proceedings of the 16th International Conference on Data Engineering, page 577, Washington, DC, USA, 2000. IEEE Computer Society.**
KD2B-Tree	**[Oosterom, 1990] P. Oosterom.** *Reactive data structures for geographic information systems*. **Ph.D. Thesis, University of Leiden, The Netherlands. 1990.**
KDB_{FD}-Tree **KDB_{HD}-Tree**	**[Orlandic, Yu, 2002] R. Orlandic, B. Yu.** *A retrieval technique for high-dimensional data and partially specified queries*. **Data & Knowledge Engineering 2002; 42(2):1-21.**
KDB_{KD}-Tree	**[Yu et al, 2003] B. Yu, R. Orlandic, T. Bailey, J. Somavaram. KDBKD-Tree: A Compact KDB-Tree Structure for Indexing Multidimensional Data. International Conference on Information Technology: Computers and Communications, 2003.**
K-D-B-Tree	**[Robinson, 1981] J.T. Robinson.** *The K-D-B-tree: A search structure for large multidimensional dynamic indexes*. **In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1981, pp. 10-18.**
K-D-Tree	**[Bentley, 1975] J. L. Bentley.** *Multidimensional binary search trees used for associative searching*. **Commun. ACM 18, 9, 1975, pp. 509–517.**
kNR-Tree	**[Mondal et al, 2005] A. Mondal, A. K. H. Tung, M. Kitsuregawa.** *kNR-tree: A novel R-tree-based index for facilitating Spatial Window Queries on any k relations among N spatial relations in Mobile environments.* **MDM 2005 05 Ayia Napa Cyprus, 2005**
Level Balaced B-Tree	**[Agarwal et al, 1999] P.K. Agarwal, L. Arge, G.S. Brodal, J.S. Vitter.** *I/O-efficient.dynamic point location in monotone planar subdivisions*. **In Proc. ACM-SIAM Symp. on Discrete Algorithms, 1999, pp.1116-1127**
Linear Hashing	**[Litwin, 1980] W. Litwin.** *Linear hashing: A new tool for file and table addressing*. **In Proceedings of the Sixth International Conference on Very Large Data Bases, 1980, pp. 212–223.** **[Larson, 1980] P. A. Larson.** *Linear hashing with partial expansions*. **In Proceedings of the Sixth International Conference on Very Large Data Bases, 1980, pp. 224–232.**
Linear Quadtree	**[Samet, 1990] H. Samet.** *Applications of Spatial Data Structures*. **Addison-Wesley, Reading, MA. 1990.**
LPC-File (Local Polar Coordinate File)	**[Cha et al, 2002] G.-H. Cha, X. Zhu, D. Petkovic, C.-W. Chung.** *An efficient indexing method for nearest neighbor searches inhigh-dirnensional image databases*. **IEEE Transactions on Multimedia, 4(1):76–87, 2002.**
LSD-Tree	**[Henrich et al, 1989] A. Henrich, H.-W. Six, P. Widmayer.** *The LSD tree: Spatial access to multidimensional point and non-point objects*. **In Proceedings of the Fifteenth International Conference on Very Large Data Bases, 1989, pp. 45–53.**

LSH (Locality Sensitive Hashing)	**[Gionis et al, 1999] A. Gionis, P. Indyk, R. Motwani.** *Similarity search in high dimensions via hashing*. **In VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases, pages 518–529, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.**
	[Indyk, Motwani, 1998] P. Indyk, R. Motwani. *Approximate nearest neighbors: towards removing the curse of dimensionality*. **In STOC '98: Proceedings of the thirtieth annual ACM symposium on Theory of computing, pages 604–613, New York, NY, USA, 1998. ACM Press.**
LUR-Tree (Lazy Update R-Tree)	**[Kwon et al, 2002] D. Kwon, S. Lee, S. Lee.** *Indexing the Current Positions of Moving Objects Using the Lazy Update R-tree*. **In Mobile Data Management, MDM, pages 113–120, Jan. 2002.**
lz-Hashing	**[Hutflesz et al, 1991] A. Hutflesz, P. Widmayer, C. Zimmermann.** *Global order makes spatial access faster*. **In Geographic Database Management Systems, G. Gambosi, M. Scholl, and H.-W. Six, Eds., Springer-Verlag, Berlin/Heidelberg/ New York, 1991, pp. 161–176.**
MB+ Tree	**[Yang et al, 1995] Q. Yang, A. Vellaikal, S. Dao.** *MB+-Tree: A New Index Structure for Multimedia Databases*. **Proceedings of the International Workshop on Multimedia Database Management Systems, August, 1995, pp. 151-158.**
MDAM, ArM	**[Markov, 1984] Kr. Markov.** *A Multi-domain Access Method.* **Proceedings of the International Conference on Computer Based Scientific Research. Plovdiv, 1984, pp.558-563.**
	[Markov, 2004] Kr. Markov. *Multi-Domain Information Model*. **International Journal "Information Theories and Applications", ISSN 1310-0513. Vol. 11, No: 4, 2004, pp. 303-308**
MedRank (Median Rank)	**[Fagin et al, 2003] R. Fagin, R. Kumar, D. Sivakumar.** *Efficient similarity search and classification via rank aggregation*. **In SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pages 301–312, New York, NY, USA, 2003. ACM Press.**
MOLPHE	**[Kriegel, Seeger, 1986] H.-P. Kriegel, B. Seeger.** *Multidimensional order preserving linear hashing with partial expansions*. **In Proceedings of the International Conference on Database Theory, LNCS 243, Springer-Verlag, Berlin/Heidelberg/New York. 1986.**
mQp-Tree	**[Salas, Polo, 2000] M. Salas, A. Polo.** *The mQp-tree: a multidimensional access method based on a non-binary tree*. **Proc of the VIIth Conference on Extending Database Technology, March 2000, Konstanz - Germany**
MR-tree	**[Xu et al, 1990] X. Xu, J. Han, W. Lu.** *RT-Tree: An Improved R-Tree Indexing Structure for Temporal Spatial Databases*. **In Proc. of the Intl. Symp. on Spatial Data Handling, SDH, pages 1040–1049, July 1990.**
M-Tree (Metric Tree)	**[Ciaccia et al, 1997] P. Ciaccia, M. Patella, P. Zezula.** *M-tree: An efficient access method for similarity search in metric spaces*. **In VLDB '97: Proceedings of the 23rd International Conference on Very Large Data Bases, pages 426–435, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.**
Multi-Layer Grid File	**[Six, Widmayer, 1988] H. Six, P. Widmayer.** *Spatial searching in geometric databases*. **In Proceedings of the Fourth IEEE International Conference on Data Engineering, 1988, pp. 496–503.**
Multi-Level Grid File	**[Whang, Krishnamurthy, 1985] K.-Y. Whang, R. Krishnamurthy.** *Multilevel grid files*. **IBM Research Laboratory, Yorktown Heights, NY. 1985.**
MV3R-Tree	**[Tao, Papadias, 2001b] Y. Tao, D. Papadias.** *MV3R-Tree: A Spatio-Temporal Access Method for Timestamp and Interval Queries*. **In Proc. of the Intl. Conf. on Very Large Data Bases, VLDB, pages 431–440, Sept. 2001.**
MVB-Tree (Multi Version B-Tree)	**[Becker et al, 1996] B. Becker, S. Gschwind, T. Ohler, B. Seeger, P. Widmayer. An** *Asymptotically Optimal Multiversion B-Tree*. **VLDB Journal, 5(4):264–275, 1996.**
MVP Tree	**[Bozkaya, Ozsoyoglu, 1997] T. Bozkaya, M. Ozsoyoglu.** *Distance-Based Indexing for High-Dimensional Metric Spaces*. **Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, May 1997, pp. 357-368.**
MVSB-Tree (Multiversion SB-tree)	**[Markowetz et al, 2001] D. Zhang, A. Markowetz, V. Tsotras, D. Gunopulos, B. Seeger.** *Efficient computation of temporal aggregates with range predicates*. **In Proc. Principles Of Database Systems, pages 237–245, 2001.**
NA-Tree (Nine Areas Tree)	**[Chang et al, 2003] Y.-I. Chang, C.–H. Liao, H.-L. Chen.** *NA-Trees: A Dynamic Index for Spatial Data*. **Journal of Information Science and Engineering, 19, 103-139, 2003.**
Nested Interpolation Based Grid File	**[Ouksel, Mayer, 1992] M.A. Ouksel, O. Mayer.** *A robust and efficient spatial data structure*. **Acta Inf. 29, 1992, pp. 335–373.**
NSI-Tree	**[Porkaew et al, 2001] K. Porkaew, I. Lazaridis, S. Mehrotra.** *Querying Mobile Objects in Spatio-Temporal Databases*. **In Proc. of the Intl. Symp. on Advances in Spatial and Temporal Databases, SSTD, pages 59–78, Redondo Beach, CA, July 2001.**
Overlapping Linear Quadtree	**[Tzouramanis et al, 1998] T. Tzouramanis, M. Vassilakopoulos, Y. Manolopoulos.** *Overlapping Linear Quadtrees: A Spatio-Temporal Access Method*. **In Proc. of the ACM workshop on Adv. in Geographic Info. Sys., ACM GIS, pages 1–7, Nov. 1998.**
P+-Tree	**[Zhang et al, 2004] R. Zhang, B. C. Ooi, K.-L. Tan.** *Making the pyramid technique robust to query types and workloads*. **In ICDE'04 : Proceedings of the Twenth International Conference on Data Engineering, pages 313–324, Washington, DC, USA, 2004. IEEE Computer Society.**
Packed R-Tree	**[Roussopoulos, Leifker, 1985] N. Roussopoulos, D. Leifker.** *Direct spatial search on pictorial databases using packed R-trees*. **In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1985, pp. 17–31.**

Parallel R-Tree	**[Kamel, Faloutsos, 1992] I. Kamel, C. Faloutsos.** *Parallel R-trees*. **In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1992, pp. 195–204.**
Partitioned B-Tree	**[Graefe, 2003] G. Graefe.** *Sorting and Indexing with Partitioned B-Trees*. **Conference on Innovative Data Systems Research, 2003.**
PCURE (Paralel Implementation of Clustering Using Representatives)	**[Berrani et al, 2003] S.-A. Berrani, L. Amsaleg, P. Gros.** *Approximate searches: k-neighbors + precision*. **In CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management, pages 24–31, New York, NY, USA, 2003. ACM Press.**
Persistent B-tree	**[Sarnak, Tarjan, 1986] N. Sarnak, R.E. Tarjan.** *Planar point location using persistent search trees*. **Communication of the ACM, 1986, 29:669-679.**
PLOP-Hashing	**[Kriegel, Seeger, 1988] H.-P. Kriegel, B. Seeger.** *PLOP-hashing: A grid file without directory*. **In Proceedings of the Fourth IEEE International Conference on Data Engineering, 1988, pp. 369–376.**
PM Quadtree	**[Samet, Webber, 1985] H. Samet, R. E. Webber.** *Storing a collection of polygons using quadtrees*. **ACM Trans. Graph. 4, 3, 1985, 182–222.**
PMR-Quadtree	**[Nelson, Samet, 1986] R.C. Nelson, H. Samet.** *A Consistent Hierarchical Representation for Vector Data*. **In Proc. of the ACM SIGGRAPH, pages 197–206, Aug. 1986.**
PMR-Quadtree for moving objects	**[Tayeb et al, 1998] J. Tayeb, O . Ulusoy, O.Wolfson.** *A Quadtree-Based Dynamic Attribute Indexing Method*. **The Computer Journal, 41(3):185–200, 1998.**
Point Quadtree	**[Klinger, 1971] A. Klinger.** *Pattern and search statistics*. **In Optimizing Methods in Statistics, S.Rustagi, Ed., 1971, pp. 303–337.**
PPR-Tree	**[Kumar et al, 1998] A. Kumar, V. J. Tsotras, and C. Faloutsos.** *Designing Access Methods for Bitemporal Databases*. **IEEE Trans. on Knowledge and Data Engineering, TKDE, 10(1):1–20, 1998.**
PPR-Tree with Linear Greedy	**[Kollios et al, 2001] G. Kollios, V. J. Tsotras, D. Gunopulos, A. Delis, M. Hadjieleftheriou.** *Indexing Animated Objects Using Spatiotemporal Access Methods*. **IEEE Trans. on Knowledge and Data Engineering, TKDE, 13(5):758–777, 2001.**
PPR-Tree with Polynomial Greedy	**[Hadjieleftheriou et al, 2002] M. Hadjieleftheriou, G. Kollios, V. J. Tsotras, D. Gunopulos.** *Efficient Indexing of Spatiotemporal Objects*. **In Proc. of the Intl. Conf. on Extending Database Technology, EDBT, pages 251–268, Czech Republic, Mar. 2002.**
Prefix Hash Tree	**[Ramabhadran et al, 2004] S. Ramabhadran, S. Ratnasamy, J.M. Hellerstein, S. Shenker.** *Prefix Hash Tree: An Indexing Data Structure over Distributed Hash Tables*. **Submitted to PODC 2004**
PR-Tree	**[Cai, Revesz, 2000] M. Cai, P. Revesz.** *Parametric R-Tree: An Index Structure for Moving Objects*. **In Proc. of the Intl. Conf. on Management of Data, COMAD, Dec. 2000.**
PSI (Parametric Space Indexing)	**[Porkaew et al, 2001] K. Porkaew, I. Lazaridis, S. Mehrotra.** *Querying Mobile Objects in Spatio-Temporal Databases*. **In Proc. of the Intl. Symp. on Advances in Spatial and Temporal Databases, SSTD, pages 59–78, Redondo Beach, CA, July 2001.**
P-Tree (J)	**[Jagadish, 1990] H.V. Jagadish.** *Spatial search with polyhedra*. **In Proceedings of the Sixth IEEE International Conference on Data Engineering, 1990, pp. 311–319.**
P-Tree (S)	**[Schiwietz, 1993] M. Schiwietz.** *Speicherung und anfragebearbeitung komplexer geo-objekte*. **Ph.D. Thesis, Ludwig-Maximilians-Universitat Munchen, Germany (in German). 1993.**
PvS Index	**[Lejsek et al, 2005] H. Lejsek, F. H. Asmundsson, B. P. Jonsson, L. Amsaleg.** *Efficient and effective image copyright enforcement*, **Technical Report. Reykjavík University, 2005.**
Pyramid Technique	**[Berchtold et al, 1998] S. Berchtold, C. Bohm, H.-P. Kriegel.** *The pyramid-technique: Towards breaking the curse of dimensionality*. **In L. M. Haas and A. Tiwary, editors, SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA, pages 142–153. ACM Press, 1998.**
Q+R-Tree	**[Yuni, Prabhakar, 2003] X. Yuni, S. Prabhakar.** *Q+Rtree: efficient indexing for moving object databases*. **Database Systems for Advanced Applications, 2003. (DASFAA 2003). Proceedings. Eighth International Conference on Publication Date: 26-28 March 2003. Pp: 175- 182**
Quantile Hashing	**[Kriegel, Seeger, 1987] H.-P. Kriegel, B. Seeger.** *Multidimensional quantile hashing is very efficient for non-uniform record distributions*. **In Proceedings of the Third IEEE International Conference on Data Engineering, 1987, pp. 10–17.**
R*-Tree	**[Beckmann et al, 1990] N. Beckmann, H.-P. Kriegel, R. Schneider, B. Seeger.** *The R*-tree: An efficient and robust access method for points and rectangles*. **In Proceedings of ACM SIGMOD International Conference on Management of Data, 1990, pp. 322–331.**
R+-Tree	**[Sellis et al, 1987] T. Sellis, N. Roussopoulos, C. Faloutsos.** *The R+-tree: A dynamic index for multi-dimensional objects*. **In Proceedings of the Thirteenth International Conference on Very Large Data Bases, 1987, pp. 507–518.**
Reactive Tree	**[Oosterom, 1993] P. van Oosterom.** *Reactive Data Structures for Geographic Information Systems*. **Oxford University Press, Oxford. 1993.**
Region Quadtree	**[Finkel, Bentley, 1974] R. Finkel, J. L. Bentley.** *Quadtrees: A data structure for retrieval of composite keys*. **Acta Inf. 4, 1, 1974, pp. 1–9.**
Relational Interval Tree	**[Kriegel et al, 2000] H.-P. Kriegel, M. Pötke, T. Seidl.** *Managing Intervals Efficiently in Object-Relational Databases*. **Proc. 26th Int. Conf. on Very Large Databases (VLDB): 407-418, 2000.**
Relational R-Tree	**[Ravi et al, 1999] K.V. Ravi Kanth, S. Ravada, J. Sharma, J. Banerjee.** *Indexing Medium-dimensionality Data in Oracle*. **Proc. ACM SIGMOD Int. Conf. on Management of Data: 521-522, 1999.**

Relational X-Tree	**[Berchtold et al, 1999] S. Berchtold, C. Böhm, H.-P. Kriegel, U. Michel.** *Implementation of Multidimensional Index Structures for Knowledge Discovery in Relational Databases*. **Proc. 1st Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK), LNCS 1676: 261-270, 1999.**
R^{EXP}-Tree	**[Saltenis, Jensen, 2002] S. Saltenis, C. S. Jensen.** *Indexing of Moving Objects for Location-Based Services*. **In Proc. of the Intl. Conf. on Data Engineering, ICDE, Feb. 2002.**
R-File	**[Hutflesz et al, 1990] A. Hutflesz, H.-W. Six, P. Widmayer.** *The R-file: An efficient access structure for proximity queries*. **In Proceedings of the Sixth IEEE International Conference on Data Engineering, 1990, pp. 372–379.**
R-link Tree	**[Ng, Kameda, 1994] V. Ng, T. Kameda.** *The R-link tree: A recoverable index structure for spatial data*. **In Proceedings of the Fifth Conference on Database and Expert Systems Applications (DEXA'94), D. Karagiannis, Ed., LNCS 856, Springer-Verlag, Berlin/Heidelberg/New York, 1994, 163–172.**
R-Tree	**[Guttman, 1984] A. Guttman.** *R-trees: A dynamic index structure for spatial searching*. **In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1984, 47–54.**
RT-Tree	**[Xu et al, 1990] X. Xu, J. Han, W. Lu.** *RT-Tree: An Improved R-Tree Indexing Structure for Temporal Spatial Databases*. **In Proc. of the Intl. Symp. on Spatial Data Handling, SDH, pages 1040–1049, July 1990.**
SB-Tree (Segment B-Tree)	**[Yang, Widom, 2001] J. Yang, J. Widom.** *Incremental Computation and Maintenance of Temporal Aggregates*. **Proceedings of the 17th International Conference on Data Engineering, 2001, Pages: 51 - 60**
SEB-Tree (Start/End Timestamp B-Tree)	**[Song, Roussopoulos, 2003] Z. Song, N. Roussopoulos.** *SEB-tree: An Approach to Index Continuously Moving Objects*. **In Mobile Data Management, MDM, pages 340–344, Jan. 2003.**
Segment Indexes	**[Kolovson, Stonebraker, 1991] C. Kolovson, M. Stonebraker.** *Segment indexes: Dynamic indexing techniques for multi-dimensional interval data*. **In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1991, pp. 138–147.**
SETI (Scalable and Efficient Trajectory Index)	**[Chakka et al, 2003] V. P. Chakka, A. Everspaugh, J. M. Patel.** *Indexing Large Trajectory Data Sets with SETI*. **In Proc. of the Conf. on Innovative Data Systems Research, CIDR, Asilomar, CA, Jan. 2003.**
SH-Tree (Super Hybrid Tree)	**[Dang, 2006] T.K. Dang.** *The SH-Tree: A Novel and Flexible Super Hybrid Index Structure for Similarity Search on Multidimensional Data*. **International Journal of Computer Science & Applications, 2006 Technomathematics Research Foundation Vol. III, No. I, pp. 1 - 25**
SKD-Tree	**[Ooi et al, 1987] B.C. Ooi, K.J. McDonell, R. Sacks-Davis.** *Spatial kd-tree: An indexing mechanism for spatial databases*. **In Proceedings of the IEEE Computer Software and Applications Conference, 1987, pp. 433–438.**
Slim Tree	**[Traina Jr. et al, 2000] C. Traina Jr., A. Traina, B. Seeger, C. Faloutsos.** *Slim-trees: High Performance Metric Trees Minimizing Overlap Between Nodes*. **International Conference on Extending Database Technology (EDBT) 2000, Konstanz, Germany, March 27-31, 2000.**
Space Filling Curves	**[Morton, 1966] G. Morton.** *A computer oriented geodetic data base and a new technique in file sequencing*. **IBM Ltd. 1966.**
Sphere Tree	**[Oosterom, 1990] P. Oosterom.** *Reactive data structures for geographic information systems*. **Ph.D. Thesis, University of Leiden, The Netherlands. 1990.**
sQSF-Tree (Simple QSF-Tree)	**[Yu et al, 1999] B. Yu, R. Orlandic, M. Evens.** *Simple QSF-trees: an efficient and scalable spatial access method*. **Proceedings of the Eighth International Conference on Information and Knowledge Management, p.5-14, November 02-06, 1999, Kansas City, Missouri, United States**
SR Tree (Sphere-Rectangle Tree)	**[Katayama, S. Satoh, 1997] N. Katayama, S. Satoh.** *The sr-tree: an index structure for high-dimensional nearest neighbor queries*. **In SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data, pages 369–380, New York, NY, USA, 1997. ACM Press.**
SS Tree (Similarity Search Tree)	**[White, Jain, 1996] D. A. White, R. Jain.** *Similarity indexing with the ss-tree*. **In ICDE '96: Proceedings of the Twelfth International Conference on Data Engineering, pages 516–523, Washington, DC, USA, 1996. IEEE Computer Society.**
STAR-Tree (Spatio-temporal Self Adjusting R-Tree)	**[Procopiuc et al, 2002] C. M. Procopiuc, P. K. Agarwal, S. Har-Peled.** *STAR-Tree: An Efficient Self-Adjusting Index for Moving Objects*. **In Proc. of the Workshop on Alg. Eng. and Experimentation, ALENEX, pages 178–193, Jan. 2002.**
String B-tree	**[Ferragina, Grossi, 1995] P. Ferragina, R. Grossi.** *A fully-dynamic data structure for external substring search*. **In Proc. ACM Symp. on Theory of Computation, 1995, pp.693-702.**
STR-tree	**[Pfoser et al, 2000] D. Pfoser, C. S. Jensen, and Y. Theodoridis.** *Novel Approaches in Query Processing for Moving Object Trajectories*. **In Proc. of the Intl. Conf. on Very Large Data Bases, VLDB, pages 395–406, Sept. 2000.**
SV-Model	**[Chon et al, 2001] H. D. Chon, D. Agrawal, A. E. Abbadi.** *Storage and Retrieval of Moving Objects*. **In Mobile Data Management, pages 173–184, Jan. 2001.**
TB-Tree (Trajectory-Bundle Tree)	**[Pfoser et al, 2000] D. Pfoser, C. S. Jensen, and Y. Theodoridis.** *Novel Approaches in Query Processing for Moving Object Trajectories*. **In Proc. of the Intl. Conf. on Very Large Data Bases, VLDB, pages 395–406, Sept. 2000.**

TPR*-Tree	**[Tao et al, 2003] Y. Tao, D. Papadias, J. Sun.** *The TPR*-Tree: An Optimized Spatio-Temporal Access Method for Predictive Queries*. **Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003**
TPR-tree (Time Parameterized R-Tree)	**[Saltenis et al, 2000] S. Saltenis, C. S. Jensen, S. T. Leutenegger, M. A. Lopez.** *Indexing the Positions of Continuously Moving Objects*. **In Proc. of the ACM Intl. Conf. on Management of Data, SIGMOD, pages 331–342, May 2000.**
TR*-Tree	**[Schneider, Kriegel, 1992] R. Schneider, H.-P. Kriegel.** *The TR*-tree: A new representation of polygonal objects supporting spatial queries and operations*. **In Proceedings of the Seventh Workshop on Computational Geometry, LNCS 553, Springer-Verlag, Berlin/Heidelberg/New York, 1992, pp. 249–264.**
TR-Tree (Temporal R-Tree)	**[Almeida et al, 1999] V. T. Almeida, J. M. Souza, G. Zimbrao.** *The Temporal R-Tree*. **Technical Report No. ES-492/99, COPPE Sistemas/UFRJ, 1999.**
TSB-Tree	**[Lomet, Salzberg, 1989] D.B. Lomet, B. Salzberg.** *Access Methods for Multiversion Data*. **In Proc. of the ACM Intl. Conf. on Management of Data, SIGMOD, pages 315–324, May 1989.**
TV-Tree (Telescoping Vector Tree)	**[Lin et al, 1994] K. I. Lin, H. V. Jagadish, C. Faloutsos.** *The tv-tree: an index structure for high-dimensional data*. **The VLDB Journal, 3(4):517–542, 1994.**
Twin Grid File	**[Hutflesz et al, 1988b] A. Hutflesz, H.-W. Six, P. Widmayer.** *Twin grid files: Space optimizing access schemes*. **In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1988, pp. 183–190.**
Two-Level Grid File	**[Hinrichs, 1985] K. Hinrichs.** *Implementation of the grid file: Design concepts and experience*. **BIT 25, 1985, pp. 569–592.**
UB-Tree	**[Bayer, 1996] R. Bayer.** *The universal B-tree for multidimensional indexing*. **Tech. Rep. I9639, Technische Universitat Munchen, Munich, Germany. 1996.**
VA+-File	**[Ferhatosmanoglu et al, 2000] H. Ferhatosmanoglu, E. Tuncel, D. Agrawal, A. E. Abbadi.** *Vector approximation based indexing for non-uniform high dimensional data sets*. **In CIKM '00: Proceedings of the ninth international conference on Information and knowledge management, pages 202–209, New York, NY, USA, 2000. ACM Press.**
VA-File (Vector Approximation File)	**[Weber et al, 2000] R. Weber, K. Bohm, H.-J. Schek. Interactive-time similarity search for large image collections using parallel va-files. In ICDE, page 197, 2000.**
VCI R-Tree (Velocity Constrained Indexing R-Tree)	**[Prabhakar et al, 2002] S. Prabhakar, Y. Xia, D. V. Kalashnikov, W. G. Aref, S. E. Hambrusch.** *Query Indexing and Velocity Constrained Indexing: Scalable Techniques for Continuous Queries on Moving Objects*. **IEEE Transactions on Computers, 51(10):1124–1140, 2002.**
VP Tree (Vantage Point Tree)	**[Yianilos, 1993] P. N. Yianilos.** *Data structures and algorithms for nearest neighbor search in general metric spaces*. **In SODA '93: Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms, pages 311–321, Philadelphia, PA, USA, 1993. Society for Industrial and Applied Mathematics.**
V-Reactive Tree	**[Li et al, 2001] J. Li, N. Jing, M. Sun.** *Spatial Database Techniques Oriented to Visualization in 3D GIS*. **In Proceedings of the 2nd International Symposium on Digital Earth. 2001.**
Weight-balanced B-tree	**[Arge, Vitter, 1996] L. Arge, J. S.Vitter.** *Optimal Dynamic Interval Management in External Memory*. **Proc. 37th Annual Symp. on Foundations of Computer Science: 560-569, 1996.**
X-Tree	**[Berchtold et al, 1996] S. Berchtold, D. Keim, H.-P. Kriegel.** *The X-tree: An index structure for high-dimensional data*. **In Proceedings of the 22nd International Conference on Very Large Data Bases, (Bombay) 1996, pp. 28–39.**
Z-Hashing	**[Hutflesz et al, 1988a] A. Hutflesz, H.-W. Six, P. Widmayer.** *Globally order preserving multidimensional linear hashing*. **In Proceedings of the Fourth IEEE International Conference on Data Engineering, 1988, pp. 572–579.**
zkdB+tree	**[Orenstein, 1986] J. A. Orenstein.** *Spatial query processing in an object-oriented database system*. **In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1986, 326–333.**
Z-Ordering	**[Orenstein, Merrett, 1984] J. Orenstein, T.H. Merrett.** *A class of data structures for associative searching*. **In Proceedings of the Third ACM SIGACT–SIGMOD Symposium on Principles of Database Systems, 1984, pp. 181–190.**

Additional bibliography

[Arge, 2002] L. Arge. *External Memory Data Structures*. Part 4, chapter 9 in Handbook of Massive Datasets. J. Abello, P.M. Pardalos, M.G.C. Resende (eds), Kluwer Academic Publishers, 2002, pp. 313-357.

[Bayer, 1996] R. Bayer. *The Universal B-tree for Multidimensional Indexing*. Techical Report I9639, Technische Universitat Munchen, Munich, Germany. 1996.

[Bellman 1961] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press., 1961

[Bouteldja et al, 2006] N. Bouteldja, V. Gouet-Brunet, M. Scholl. *Back to the Curse of Dimensionality with Local Image Descriptors.* CEDRIC Research Report no 1049. July 20, 2006.

[Chakrabarti, 2001] K. Chakrabarti. *Managing Large Multidimensional Datasets Inside a Database System*. Phd Thesis, University of Illinois at Urbana-Champaign. Urbana, Illinois, 2001

[Chavez et al, 2001] E. Chavez, G. Navarro, R. Baeza-Yates, J. Marroquin. *Searching in Metric Spaces*. ACM Computing Surveys, 33(3):273–321, Sept. 2001.

[CODASYL, 1971] Codasyl Systems Committee. *Feature Analysis of Generalized Data Base Management Systems*. Technical Report, May, 1971 / Информационные системы общего предназначения (Аналитический обзор систем управления базами данных). Москва, Статистика, 1975.

[Connolly, Begg, 2002] T.M. Connolly, C.E.Begg. *Database Systems. A Practical Approach to Design, Implementation, and Managemen*t. Third Edition. Addison-Wesley Longman, Inc. – Pearson Education Ltd., 1995, 2002 / Т.Коннолли, К.Бегг. *Базы данных. Проектирование, реализация и сопровождение*. Теория и практика. Москва-Санкт Петербург-Киев, Издателский дом "Вильямс", 2003. 1440 с.

[Date, 1977] C.J. Date. *An Introduction to Database Systems*. Addison-Wesley Inc. 1975. / К.Дейт. *Введение в системы баз данных*. Москва, Наука, 1980.

[Gaede, Günther, 1998] V. Gaede, O. Günther. *Multidimensional Access Methods*. ACM Computing Surveys, Vol. 30, No. 2, June 1998.

[IBM, 1965-68] *IBM System/360, Disk Operating System, Data Management Concepts*. IBM System Reference Library, IBM Corp. 1965, Major Revision, February 1968.

[Markov, 2004] Kr. Markov. *Multi-Domain Information Model*. International Journal "Information Theories and Applications", ISSN 1310-0513. Vol. 11, No: 4, 2004, pp. 303-308

[Markov, 2006] Kr. Markov. Multi-dimensional Context-free Access Method. PhD Thesis. Institute of Mathematics and Informatics, Bulgarian Academy of Sciences. Sofia, 2006.

[Martin, 1975] J.Martin. *Computer Data-Base Organization*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey / Дж. Мартин. *Организация баз данных в вычислительных системах*. Москва, Мир, 1978.

[Moënne-Loccoz, 2005] N. Moënne-Loccoz. *High-Dimensional Access Methods for Efficient Similarity Queries*. Technical Report N:0505, University of Geneva, Computer Vision and Multimedia Laboratory, May 2005.

[Mokbel et al, 2003] M. F. Mokbel, T. M. Ghanem, W. G. Aref. *Spatio-temporal Access Methods*. IEEE Data Engineering Bulletin, 26(2), 40-49, June, 2003.

[Ooi et al, 1993] B.C. Ooi, R. Sacks-Davis, J. Han. *Indexing in Spatial Databases*. Technical Report. 1993.

[Schlosser et al, 2005] S.W. Schlosser, J. Schindler, S. Papadomanolakis, M. Shao, A. Ailamaki, C. Faloutsos, G.R. Ganger. *On Multidimensional Data and Modern Disks*. Proceedings of the 4th USENIX Conference on File and Storage Technology (FAST '05). San Francisco, CA. December 13-16, 2005.

[Stably, 1970] D. Stably. *Logical Programming with System/360*. New York, 1970 / Д.Стэбли. *Логическое программирование в системе/360*. Москва, Мир, 1974.

Authors' Information

Krassimir Markov - Institute of Mathematics and Informatics, BAS, Acad.G.Bonthev St., bl.8, Sofia-1113, Bulgaria; Institute of Information Theories and Applications FOI ITHEA, P.O.Box: 775, Sofia-1090, Bulgaria; e-mail: markov@foibg.com

Krassimira Ivanova - Institute of Mathematics and Informatics, BAS, Acad.G.Bonthev St., bl.8, Sofia-1113, Bulgaria; e-mail: kivanova@math.bas.bg

Ilia Mitov - Institute of Information Theories and Applications FOI ITHEA, P.O.Box: 775, Sofia-1090, Bulgaria; e-mail: mitov@foibg.com

Stefan Karastanev - Institute of Mechanics and Biomechanics, BAS, Acad.G.Bonthev St., bl.4, Sofia-1113, Bulgaria; e-mail: stefan@info.imbm.bas.bg

GENERAL REGRESSION NEURO–FUZZY NETWORK FOR IDENTIFICATION OF NONSTATIONARY PLANTS

Yevgeniy Bodyanskiy, Nataliya Teslenko

Abstract: General Regression Neuro-Fuzzy Network, which combines the properties of conventional General Regression Neural Network and Adaptive Network-based Fuzzy Inference System is proposed in this work. This network relates to so-called "memory-based networks", which is adjusted by one-pass learning algorithm.

Keywords: memory-based networks, one-pass learning, Fuzzy Inference Systems, fuzzy-basis membership functions, neurons at data points, nonlinear identification.

ACM Classification Keywords: F.1 Computation by abstract devices - Self-modifying machines (e.g., neural networks), I.2.6 Learning - Connectionism and neural nets, G.1.2. Approximation – Nonlinear approximation.

Introduction

Nowadays neural networks have wide spreading for identification, prediction and nonlinear objects control problems solving. Neural networks possess universal approximating abilities and capabilities for learning by the data that characterize the functioning of investigating systems. The situation becomes sharply complicated in the case, when the data are fed in real time, their processing must be simultaneous with functioning of the plant, and the plant is nonstationary. It's clear, that conventional multilayer perceptron, that is universal approximator, isn't effective in this case, so Radial Basis Functions Networks (RBFN) can be used as its alternative [1-3]. These networks are also universal approximators, and their output is linearly dependent on tuned synaptic weights. In this case, recurrent least squares method or its modifications can be used for their real time learning. These procedures are second-order optimization algorithms, which provide quadratic convergence to the optimal solution. At the same time, practical application of RBFN is bounded by so-called curse of dimensionality as well as appearance of "gaps" in the space of radial-basis functions (RBF) that lead to appearance of regions where all neurons of the network are inactive.

So-called, space partition of unity, implemented by Normalized Radial Basis Functions Networks (NRBFN), in which output signal is normalized by the sum of outputs of all neurons, is used to avoid such a "gaps" [4]. Given networks are learned using recurrent gradient algorithms that have slow rate of convergence and possibility of getting to the local minima as their common drawback.

Thus, these neural networks and many others that use recurrent learning procedures and united by general name "optimization-based networks" may be inefficient in problems of adaptive identification, prediction and real-time control, when the information is fed for processing with sufficiently high frequency. In this case, these networks have not time to learn and are unable to follow changing parameters of a plant.

The so-called "memory-based networks" are the effective alternative to "optimization-based networks" and General Regression Neural Network (GRNN), proposed by D. F Specht [5], is the brightest representative of these networks.

At the basis of this network lies the idea of Parzen windows [6], kernel estimates of Nadaraya-Watson [7-9] and nonparametric models [10]. Its learning consists of one-time adjustment of multidimensional radial-basis functions (RBF) at points of unit centered hypercube, which are specified unambiguously by the learning set. Therefore, these networks can be referred to, so-called, just-in-time models [4], which are adjusted by one-pass learning algorithm. Being similar to NRBFN by the architecture, GRNN learns much faster, placing the centers of RBF at the points with coordinates that are determined by input signals of a plant using principle "neurons at the data points" [11] and with RBF heights, which coincide with corresponding values of plant output signal. High learning rate of GRNN provides their effective using in the real-time problems solving [12,13].

For the solving of nonlinear plant identification problem

$$y(k) = F(x(k))$$

where *y(k), x(k)*– scalar and *(nx1)*-vector of output and input signals correspondingly in the instant time *k=1,2,…*, *F(•)* – unknown nonlinear operator of the plant, it is necessary to form learning sample $\{x^*(k), y^*(k)\}$, *k=1,2,…,l,* whereupon it is possible to get the estimate $\hat{y}(k)$ of the plant response *y(k)* to arbitrary input signal *x* in the form

$$\hat{y}(x) = \frac{\sum_{k=1}^{l} y^*(k)\varphi(D(k))}{\sum_{k=1}^{l} \varphi(D(k))} \qquad (1)$$

where *D(k)* – distance measure in accepted metrics between *x* and *x*(k)*, $\varphi(\bullet)$ – some kernel function, usually, Gaussian. Conventionally the Euclidean metrics is used as a distance

$$D^2(k) = \sum_{i=1}^{n} \left(\frac{x_i(k) - x_i^*(k)}{\sigma(k)} \right)^2$$

(here $\sigma(k)$ – scalar parameter, which determines the receptive field radius of kernel function $\varphi(\bullet)$), although in more common case it is possible to use Minkowski metrics

$$D^p(k) = \sum_{i=1}^{n} \left| \frac{x_i - x_i^*(k)}{\sigma(k)} \right|^p , \; p \geq 1 .$$

Thus, GRNN converges asymptotically to optimal nonlinear regression surface with the growing of learning sample size [9].

GRNN learning process can be organized easily in real time. In this case the learning pairs *x*(k), y*(k)* are fed to the network sequentially, forming new radial-basis function-neurons. At the same time, the distance between newly formed and already existing functions is estimated gradually. If this distance is smaller than threshold value *r*, that is defined in advance, new neuron isn't included in the network. The main problems concerned with GRNN using are defined by possible curse of dimensionality. Growing of the learning sample size *l* and the difficulties with correct definition of parameter *r*, which is sufficiently difficult to choose and interpret in multidimensional space, are the causes of it.

Neuro-Fuzzy Systems (NFS) are the natural expansion of artificial neural networks [14-15]. They combine the neural networks learning abilities with transparence and interpretability of the Fuzzy Inference Systems (FIS). Generally, FIS represents fuzzy models, which are learned by observations data of plant inputs and outputs, using univariate Fuzzy Basis Functions (FBF) instead of multidimensional RBF. In common case FBF are bell-shaped (usually Gaussian) membership functions, which are used in Fuzzy Logic. Using of bell-shaped FBF allows us to combine local features of the kernel functions with the properties of sigmoidal activation functions that provide global approximation properties [16]. Having the approximating abilities of RBFN [15], NFS subject to curse of dimensionality with less degree, that provides them advantage in comparison with neural networks.

Among Neuro-Fuzzy Systems (NFS) Adaptive Network-based Fuzzy Inference System (ANFIS) have got wide spread [17]. ANFIS has five-layer architecture, whose synaptic weights are tuned similarly to RBFN. The adjusting possibility of FBFs using error back-propagation algorithm is provided in this system too. ANFIS and many other similar neuro-fuzzy systems [4,15,16] are typical representatives of the optimization-based networks family, which are characterized by insufficient learning rate.

Lattice-based Associative Memory Networks (LAMN) [18, 19] are the representatives of memory-based networks, whose output signal is formed on basis of univariate bell-shaped functions uniformly distributed on axes of n-dimensional input space. As a result of aggregation operation multidimensional FBFs are formed, whose centers are also uniformly distributed in multidimensional space, and their layout doesn't depend on characteristics of learning sample.

The goal of this work is the development of General Regression Neuro-Fuzzy Network (GRNFN), which represents by itself NFS and learns as GRNN that provides it approximating properties of ANFIS with learning rate of memory-based networks.

The General Regression Neuro-Fuzzy Network architecture

The architecture of General Regression Neuro-Fuzzy Network is illustrated on Fig. 1 and consists of five sequentially connected layers. First hidden layer is composed of *l* blocks with *n* FBF in each and realizes fuzzification of the input variables vector. Second hidden layer implements aggregation of membership levels that are computed in first layer, and consists of *l* multiplication blocks. Third hidden layer – the layer of synaptic weights that are defined in special way. Fourth layer is formed by two summation units and computes the sums of output signals from the second and third layers. Finally, normalization takes place in fifth (output) layer, as a result of which, the output network signal is computed.

One can see, that the architecture of GRNFN coincides with the architecture of L.-X. Wang–-J.M. Mendel neuro-fuzzy system [20], which, in turn, is the modification of zero-order T. Takagi–M. Sugeno fuzzy inference system [21]. However, if NFS is learned using one or another optimization procedures, GRNFN is adjusted using one-pass learning algorithm.

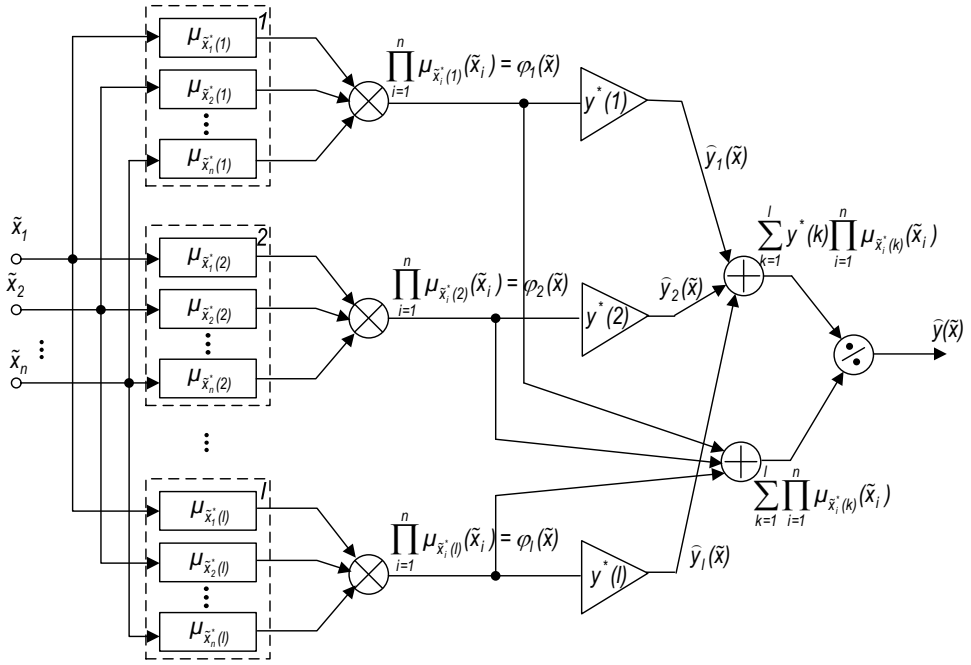


Fig.1 – General Regression Neuro-Fuzzy Network.

The General Regression Neuro-Fuzzy Network learning

Since GRNFN belongs to memory-based networks, its learning is based on principle "neurons at data points" that makes it extremely easy and fast.

Learning sample vectors $x^*(1),...,x^*(k),...,x^*(l)$ are normalized in advance on unit centered hypercube so, that

$$x_i^{*min} \le x_i^*(k) \le x_i^{*max}, \; i = 1,2,...,n,$$

$$-0,5 \le \tilde{x}_i^*(k) \le 0,5 .$$

Mutual recalculation is made according to the next expressions

$$\tilde{x}_i^*(k) = \frac{x_i^*(k) - x_i^{*min}}{x_i^{*max} - x_i^{*min}} - 0,5 ,$$

$$x_i^*(k) = (\tilde{x}_i^*(k) + 0,5)(x_i^{*max} - x_i^{*min}) + x_i^{*min} .$$

For each vector from the learning sample $\tilde{x}^*(k) = (\tilde{x}_1^*(k), \tilde{x}_2^*(k),..., \tilde{x}_n^*(k))^T$ in the first hidden layer own set of fuzzy-basis membership functions $\mu_{\tilde{x}_1^*(k)}, \mu_{\tilde{x}_2^*(k)},..., \mu_{\tilde{x}_n^*(k)}$ is formed, so that centers of $\mu_{\tilde{x}_i^*(k)}$ coincide with $\tilde{x}_i^*(k)$,

k=1,2,…,l. The process of FBF formation is illustrated on Fig. 2. Note that GRNFN contains *nl* fuzzy-basis functions, that can't lead to the curse of dimensionality.

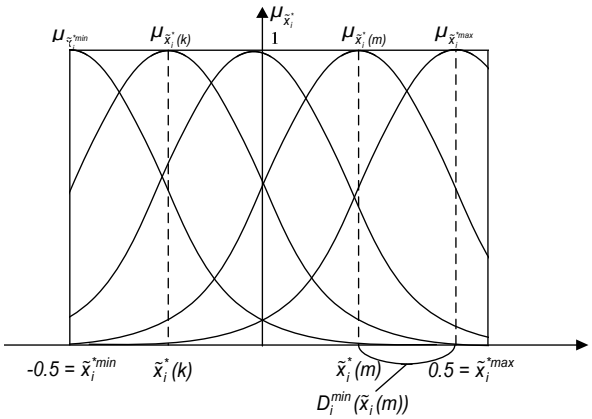


Fig.2 – Fuzzy-basis membership functions.

Theoretically, any kernel function with non-strictly local support can be used as FBF. It allows avoiding of appearance of "gaps" [9]. As such a function one can recommend generalized Gaussian

$$\mu_{\tilde{x}_i^*(k)}(\tilde{x}_i) = \left(1 + \left| \frac{\tilde{x}_i^*(k) - \tilde{x}_i}{\sigma_i(k)} \right|^{2b} \right)^{-1}, \; b \geq 0{,}5 , \tag{2}$$

that is the bell-shaped function, whose shape is defined by the scalar parameter *b* [15]. Let's also note, that *b* defines the metrics $D^{2b}(k)$ too. As for choosing of the width parameter $\sigma_i(k)$, standard recommendation leads to the idea [8], that it must ensures small overlapping of neighboring FBFs. Easy to see, that for Gaussian this recommendation leads to estimate

$$\sigma_i(k) < \frac{l-1}{2 \div 3} .$$

At the same time with FBFs forming in first hidden layer, the synaptic weights are being tuned in the third hidden layer and they are supposed to be equal to the signals of learning sample $y^*(k)$.

Thus, when arbitrary signal $\tilde{x}$ is fed to the input of GRNFN in the first hidden layer membership levels $\mu_{\tilde{x}_i^*(k)}(\tilde{x}_i)$, *i=1,2,…,n*, *k=1,2,…,l* are computed, in the second layer their aggregation is made by forming multidimensional FBFs

$$\varphi_k(\tilde{x}) = \prod_{i=1}^{n} \left(1 + \left| \frac{\tilde{x}_i^*(k) - \tilde{x}_i}{\sigma_i(k)} \right|^{2b} \right)^{-1}, \; k = 1,2,...,l ,$$

in the third layer products $\hat{y}(\tilde{x}) = y^*(k)\varphi_k(\tilde{x})$ are determined, fourth layer computes the values of signals $\sum_{k=1}^{l} y^*(k)\varphi_k(\tilde{x})$ and $\sum_{k=1}^{l} \varphi_k(\tilde{x})$, and, finally, in the output layer the estimate

$$\hat{y}(\tilde{x}) = \frac{\sum_{k=1}^{l} y^*(k)\varphi_k(\tilde{x})}{\sum_{k=1}^{l} \varphi_k(\tilde{x})} = \frac{\sum_{k=1}^{l} y^*(k) \prod_{i=1}^{n} \mu_{\tilde{x}_i^*(k)}(\tilde{x}_i)}{\sum_{k=1}^{l} \prod_{i=1}^{n} \mu_{\tilde{x}_i^*(k)}(\tilde{x}_i)} ,$$

is forming, which coincides with (1) with the only difference, that instead of radial-basis functions multidimensional fuzzy-basis functions are used, that were formed of univariate FBF.

The scheme of fuzzy inference, which is realized by GRNFN can be presented as a logic equations system

$$IF(\tilde{x}_1.IS.A_1(1)).AND.(\tilde{x}_2.IS.A_2(1)).AND.....AND.(\tilde{x}_n.IS.A_n(1)), \quad THEN \quad \hat{y}_1(\tilde{x}) = y^*(1)$$
$$\vdots$$
$$IF(\tilde{x}_1.IS.A_1(k)).AND.(\tilde{x}_2.IS.A_2(k)).AND.....AND.(\tilde{x}_n.IS.A_n(k)), \quad THEN \quad \hat{y}_k(\tilde{x}) = y^*(k)$$
$$\vdots$$
$$IF(\tilde{x}_1.IS.A_1(l)).AND.(\tilde{x}_2.IS.A_2(l)).AND.....AND.(\tilde{x}_n.IS.A_n(l)), \quad THEN \quad \hat{y}_l(\tilde{x}) = y^*(l)$$

where the operator $A_i(k)$ is represented by the membership function (2). Hence, using of neuro-fuzzy approach allows ensuring of obtained results interpretation.

The GRNFN learning process can proceed both in batch mode, when learning sample $\left\{x^*(k), y^*(k)\right\}$ is specified apriori and in real time, when pairs $x^*(k)$, $y^*(k)$ are given sequentially, forming multidimensional FBFs φ_k. It is sufficiently easy to organize the exclusion process of slight information pairs. If for some observation $\tilde{x}^*(m)$ next condition is held

$$\max_i D_i^{min}(\tilde{x}_i(m)) < r < (l-1)^{-1} \tag{3}$$

(here $D_i^{min}(\tilde{x}_i(m))$ – the least distance between $\tilde{x}_i(m)$ and earlier formed neighboring centers of FBFs), then $\tilde{x}^*(m)$ doesn't form function φ_m and is removed from the consideration. Note, that for univariate situation the threshold parameter r and the distance D_i^{max} are significantly easier to define, then in multidimensional case of GRNN.

Operation of GRNFN can be organized simply in the continuous adaptation mode that is essentially important for nonstationary objects identification and control. Here it is possible to use two approaches. The first is – on the sliding window of l observations, when while learning pairs $x^*(l+1)$, $y^*(l+1)$ are being fed to the input of the network, in the first and third layers the pair of $\mu_{\tilde{x}_i^*(1)}$ and $y^*(1)$ is removed, and instead of it the membership function $\mu_{\tilde{x}_i^*(l+1)}$ and weight $y^*(l+1)$ are formed. The second approach is based on inequality (3). In this case newly received pair $x^*(m)$, $y^*(m)$ isn't removed, but replaces the nearest to it in the "old" data.

As far as the learning process operates almost immediately, there is no problem with following properties of tuning algorithm at all.

Numerical experiment

In this experiment, the plant is assumed to be of the form [22]:

$$y(k+1) = f(y(k), y(k-1), y(k-2), u(k), u(k-1)),$$

where the unknown function f has the form

$$f(x_1, x_2, x_3, x_4, x_5) = \frac{x_1 x_2 x_3 x_5 (x_3 - 1) + x_4}{1 + x_2^2 + x_3^2}.$$

The input to the plant is given by $u(k) = sin(2\pi k/250)$ for $k \le 500$ and $u(k) = 0.8sin(2\pi k/250) + 0.2sin(2\pi k/25)$ for $k > 500$, in all 1000 signals. Fig.3(a) shows the output of the plant.

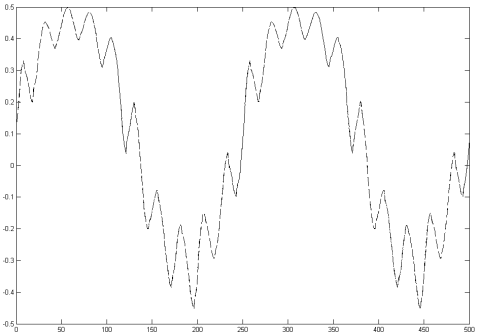


Fig. 3.a) Outputs of the plant.

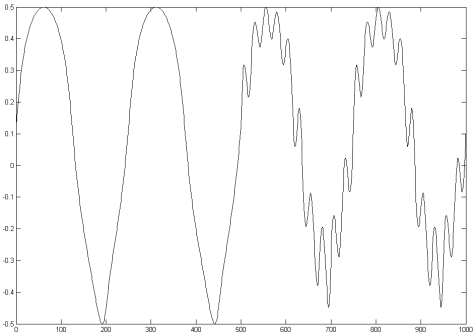


Fig. 3. b) Outputs of the GRNFN (dash-dot line) and GRNN (dashed line) practically coincide.

Two experiments were made. In the first experiment, GRNFN was constructed and learned by first 500 signals, which organized learning sample. After that, the next 500 signals were fed to the network for testing its performance. In addition, this problem was solved using conventional GRNN. The results are shown in Fig.3(b) for last 500 instants. One can see that output signals of GRNFN and GRNN practically agree with test signals and with each other, but numerical analysis shows that GRNFN has accuracy higher by 2%. In the second experiment the distances between all learning signals were computed and compared with threshold value. Only 378 of 500 signals exceeded preassigned threshold value, and they organized learning sample. In this case, GRNFN has the same accuracy. Hence, it is logically to conclude that GRNFN needs less number of signals to be learned in comparison with GRNN.

Conclusions

General Regression Neuro-Fuzzy Network, that is generalization of conventional GRNN and adaptive fuzzy inference systems, is proposed in this work. Network is characterized by computational simplicity, interpretability of the results and ensures high accuracy in the nonlinear nonstationary systems prediction and identification problems.

Bibliography

[1] Moody J., Darken C.J. Fast learning in networks of locally-tuned processing units// Neural Computation.- 1989.-1.-P.281-294.

[2] Park J., Sandberg I.W. Universal approximation using radial-basis-function networks// Neural Computation.-1991.-3.-P.246-257.

[3] Schilling R.J., Carrol J.J., Al-Ajlouni A.F. Approximation of nonlinear systems with radial basis function neural networks// IEEE Trans. on Neural Networks.-2001.-12.-P.1-15.

[4] Nelles O. Nonlinear System Identification.-Berlin: Springer, 2001.-785p.

[5] Specht D.E. A general regression neural network// IEEE Trans. on Neural Networks.-1991.-2.-P.568-576.

[6] Parzen E. On the estimation of a probability density function and the mode//Ann. Math. Stat.-1962.-38.-P.1065-1076.

[7] Nadaraya E.A. About nonparametric probability density and regression estimates// Probability theory and its Application.-1965.-10.-№1.-P199-203.

[8] Bishop C.M. Neural Networks for Pattern Recognition.- Oxford: Clarendon Press, 1995.-482p.

[9] Friedman J., Hastie T., Tibshirani R. The Elements of Statistical Learning. Data Mining, Inference, and Prediction.- Berlin: Springer, 2003.-552p.

[10] Zhivoglyadov V.G., Medvedev A.V. Nonparametric algorithms of adaptation.-Frunze: Ilim, 1974.-135p. (in Russian).

[11] Zahirniak D.R., Capman R., Rogers S.K., Suter B.W., Kabrisky M., Pyati V. Pattern recognition using radial basis function network// Proc. 6-th Ann. Aerospace Application of AI Conf.- Dayton, OH, 1990.-P.249-260.

[12] Seng T.L., Khalid M., Yusof R., Omatu S. Adaptive neuro-fuzzy control system by RBF and GRNN neural networks// J. of Intelligent and Robotic Systems.- 1998.-23.-P.267-289.

[13] Guo X.-P., Wang F.-L., Jia M.-X. A sub-stage moving window GRNN quality prediction method for injection molding process// "Lecture Notes in Computer Science"- V3973.- Berlin-Heidelberg: Springer-Verlag, 2006.-P.1138-1143.

[14] Jang J.-S. R., Sun G.-T. Neuro-fuzzy modeling and control// Proc. IEEE.-1995.-83.-P.378-406.

[15] Jang J.-S. R., Sun G.-T., Mizutani E. Neuro-Fuzzy and Soft Computing.- Upper Saddle River, NJ: Prentice Hall, 1997.-614p.

[16] Cios K.J., Pedrycz W. Neuro-Fuzzy algorithms// In: "Handbook on Neural Computation" – Oxford: University Press, 1997.-D1.3:1-D1.3:7.

[17] Jang J.-S. R. ANFIS: Adaptive-Network-based Fuzzy Inference Systems// IEEE Trans. on Systems, Man, and Cybernetics.-1993.-23.-P.665-685.

[18] Brown M., Harris C.J. Neural networks for modeling and control/ In: Eds. by C.J. Harris "Advances in Intellectual Control".- London: Taylor and Francis, 1994.-P.17-55.

[19] Wang H., Liu G.P., Harris C.J., Brown M. Advanced Adaptive Control.- Oxford: Pergamon, 1995.- 262p.

[20] Wang L.-X., Mendel J.M. Fuzzy basis functions, universal approximation, and orthogonal least squares learning//IEEE Trans. on Neural Networks.-1992.-3.-P.807-814.

[21] Takagi T., Sugeno M. Fuzzy identification of systems and its applications to modeling and control// IEEE Trans. on Systems, Man, and Cybernetics.-1985.-15.-P.116-132.

[22] Narendra K.S., Parthasarathy K. Identification and control of dynamical systems using neural networks// IEEE Trans. on Neural Networks.-1990.-1.-P.4-26.

Authors' Information

Yevgeniy Bodyanskiy – Kharkiv National University of Radio Electronics, Doctor of Technical Sciences, Professor of Artificial Intelligence Department, Head of Control Systems Research Laboratory, IEEE Senior Member; Postal address: CSRL, Ofiice 511, Lenin Av., 14, Kharkiv, 61166, Ukraine; e-mail: bodya@kture.kharkov.ua.

Nataliya Teslenko - Kharkiv National University of Radio Electronics, post-graduate student, research scientist of Control Systems Research Laboratory; Postal address: CSRL, Ofiice 511, Lenin Av., 14, Kharkiv, 61166, Ukraine; e-mail: ntntp@ukr.net.

SMART PORTABLE FLUOROMETER FOR EXPRESS-DIAGNOSTICS OF PHOTOSYNTHESIS: PRINCIPLES OF OPERATION AND RESULTS OF EXPERIMENTAL RESEARCHES

Volodymyr Romanov, Volodimyr Sherer, Igor Galelyuka, Yevgeniya Sarakhan, Oleksandra Skrypnyk

Abstract: In the Institute of Cybernetics of National Academy of Sciences of Ukraine the smart portable fluorometer for express-diagnostics of photosynthesis was designed. The device allows easy to estimate the level of influence of natural environment and pollutions to alive plants. The device is based on real time processing of the curve of chlorophyll fluorescent induction. The principles of operation and results of experimental researches of device are described in the article.

Keywords: Kautsky effect, chlorophyll, chlorophyll fluorescence induction, fluorescence, fluorometer, portable device, vine plant.

ACM Classification Keywords: J.3 Life and Medical Sciences - Biology and Genetics

Introduction

Development of information technologies and microelectronic circuits allows filling of the world market with portable computer devices such as handheld PCs, laptops, media players, medical devices (tonometers, glucometers, cardiographs), navigation devices and so on. The achievements of Ukrainian scientists who work in the field of biosensors combined with modern capabilities of information technologies provided development of devices for express-diagnostics of plant state, evaluation of environmental parameters, exposure of infective diseases etc.

In the context of the program of Presidium of NAS of Ukraine "Development in the field of sensor systems and technology" in Glushkov's Institute of cybernetics of NAS of Ukraine the portable computer device was developed for express-diagnostics of stress factors which influence on the plant's state. The portable device measures chlorophyll fluorescence induction (CFI) without plant destruction. Using the curve of CFI (alike the cardiogram) allows diagnosing influence of one or other influential factor on the plant's state.

Features of biological objects' luminescence

As a result of external influence, different objects, including biological ones, can generate plenty of radiation that is independent of these objects temperature.

All the types of radiation that were caused by some external sources of energy are called luminescence. Duration of luminescence after external influence stopping exceeds period of light fluctuations. Luminescence is conditioned by fluctuations of relatively small number of atoms or molecules of substance that become excited under energy source activity. Radiation is a result of transformation of atoms' or molecules' states into fundamental (unexcited) or less excited (they have less energy) states.

This is well adjusted with quantum theory, according to what every stationary orbit conforms to definite value of atom's energy (Bore's postulate). Being placed on stationary orbits an electron doesn't radiate and doesn't absorb electromagnetic waves. According to the second Bore's postulate radiation and absorption can happen only when atom changes its state from one stationary state to another:

$$h\varpi_{mn} = h\nu_{mn} = E_n - E_m, \quad (1)$$

where ϖ_{mn} or ν_{mn} – photon's frequency, E_m, E_n – energy values of the states *m* and *n*, h – Planck's constant, *m* and *n* – the numbers of energy states. At the same time electron switches from one stationary orbit to another.

Luminescence is defined by the structure of substance energy spectrum, the average time of staying in excited states and rules of selection, which allow absorption or radiation of light of defined frequency. Short-timed luminescence is also called fluorescence. Luminescence which appears during lighting of substance (phosphor) with visible or ultraviolet light is called photoluminescence. Usually process of luminescence satisfies Stocks' rule that claims that wave length λ' of radiated light is greater than wave λ of excited light. According to the quantum theory this means that photon's energy $h\varpi(h\nu)$ is used partially for non-optical processes:

$$h\varpi = h\varpi' + E, \varpi > \varpi', \quad (2)$$

where ϖ' – luminescence's frequency, E – energy waste on another process.

Luminescence is characterized by energy output which equals to ratio of luminescence energy to energy that was absorbed by substance under stationary conditions.

Energy efficiency of photoluminescence increases proportionally to wave length λ of absorbed light up to the definite maximum value at $\lambda = \lambda_{\max}$ and then rapidly decreases to zero at $\lambda > \lambda_{\max}$ (Vavilov's rule). A sharp decrease of energy at $\lambda > \lambda_{\max}$ is explained by the fact that at these wave lengths λ the energy of absorbed photons is not enough for the process of phosphor atoms and molecules transfer to the excited states.

Ratio of luminescence photons number to absorbed photons with fixed energy is called quantum yield of photoluminescence. According to Vavilov's rule, which is under Stocks' rule, quantum yield of photoluminescence doesn't depend on wave length of excited light and rapidly decreases for anti-Stocks radiation.

Intensity of luminescence *I* depends on behavior of elementary processes that causes this radiation. In case of spontaneous luminescence, when radiation starts after light absorption during which atoms or molecules are transmitted to the excited level that is placed higher than the level at which radiation takes place and then these atoms (molecules) are transmitted to the luminescence level, intensity is subordinate to exponential rule

$$I = I_0 \exp(-t/\tau), \quad (3)$$

where I – lighting intensity at the moment *t*, I_0 – lighting intensity in a moment of excited radiation stopping, $\tau \approx 10^{-9} - 10^{-8}\tilde{n}$ – an average duration of excited state of phosphor atoms or molecules. Luminescence of compound molecules and phosphorescence (after lighting) of organic substance are subordinate to the low (3).

Under influence of light there can be happened photochemical transformation of substance (including photosynthesis), which is called photochemical reactions. In a process of such reactions light absorption takes place. Energy is spent on compound molecules and polyatomic ions decomposition to component parts and creation of compound molecules of primary ones. An example of photochemical reactions is decomposition carbon dioxide under influence of light

$$2CO_2 + 2h\varpi \rightarrow 2CO + O_2 \quad (4)$$

Carbon dioxide decomposition takes place in green parts of plants under sun light influence, as photochemical process, which is a part of photosynthesis.

Principles of Operation

One of the most important properties of the molecule of chlorophyll which is the basic pigment of plant cell is ability to fluoresce. For the first time this phenomenon was researched by Kautsky [Kautsky, 1931], [Kautsky, 1937]. Dependence of chlorophyll fluorescence induction on time passed after start of lightning of plant's leaves is known as an induction curve or a chlorophyll fluorescence induction curve (Fig. 1). The form of this curve is rather

sensible to changes in the photosynthetic apparatus of plants during adaptation to different environmental conditions. This fact is a basic for extensive usage of Kautsky effect in photosynthesis research. The advantages of the method of CFI are the following: high self-descriptiveness, expressiveness, noninvasiveness and high sensibility.

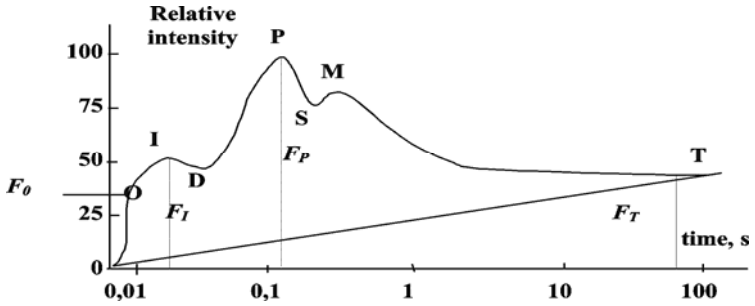


Figure 1. Chlorophyll fluorescence induction curve

The examples of changing form of this curve under influence external factors are listed in [Romanov, 2007]. The organization, scheme, basic components and advantages of the portable computer fluorometer "Floratset" are discussed in [Fedack, 2005] in detail.

Results of Experimental Research

The experimental researches of the "Floratest" were conducted in National Scientific Center "V.E. Tairov's Institute of viticulture and winemaking" of Academy of Agrarian Sciences of Ukraine. The conditions and results of the experimental researches are listed below.

Mature leaves of vine were used in the researches. Under changes of soil watering conditions there were observed sharp changes in behavior of induction transitions of chlorophyll fluorescence which were accompanied by quite essential changes of leaf tissue spectral characteristics.

Determination of fluorescence spectral characteristics was done by placing the device's sensor on the leaf's surface without integrity disturbance directly in a pot or in a field. It allowed to research on plastid and vacuolar pigments in their natural state and in that way approaching to understanding of the biophysical and physiology-biochemical processes which take place in the live leaf, and determination of important sides of photosynthetic activity.

Fluorescence intensity of the sample was determined in relative units.

It is significant that under natural conditions in the middle latitudes the drought is accompanied simultaneously by high temperatures of air, and that intensifies bad influence of ground water lack on agricultural plants.

Even in the first variant of experiment (drought) there appeared considerable changes of the behavior of fluorescence induction comparing to the control samples. Changes show in weakening of penetrability of the chloroplasts' membrane structures. That results in substantial increase of time characteristics of fluorescence induction slow decrease. At the same time noticeable variety differences become apparent. Sharp decrease of its value is typical for profound functional injuries of photosynthetic structures and cells of particular variety entirely.

Accordingly in this stage of drought influence significant variety differences in exsiccate factor resistance of both photosynthetic structures and lamina's parenchymal cells entirely became apparent.

More deep changes of destructive nature may be observed in case of high temperatures (+40 °C), which influence on leaves complementary to drought. In this case for all the varieties being studied significant and almost irreversible functional changes of plastid structures are noted. These functional changes show in sharp decrease of CFI intensity.

Disastrous changes of life activity of vine leaf cells which take place during these processes show in oppression of biosynthetic processes, intensive decomposition of cytoplasmic structures and intensification of oxide catabolism of plant cell's content. The consequence of these processes is decrease of CFI intensity as a result of its oxidizing transformation.

Diagrams of measuring of chlorophyll fluorescence intensity for vine plants under drought conditions and normal conditions accordingly are displayed on figures 2, 3.

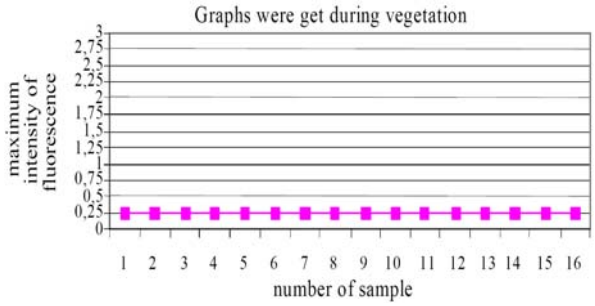


Figure 2. Maximum of CFI intensity of vine plant under drought influence (28-30% insufficient water capacity)

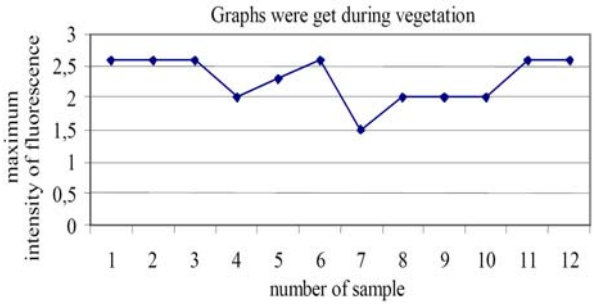


Figure 3. Maximum of CFI intensity of vine plant, control samples (68-70% insufficient water capacity)

Thus, a water deficit (WD) shows up on the Kautsky curve (figure 1) as difference of fluorescence ($F_p - F_0$) decrease. The most credible reason of this is oppression of oxygen emission which is related with slowing down of electrons transfer. Assuming that F_0 almost does not change for the test and control plants, in a maximum point the chlorophyll fluorescence intensity value can define the level of water deficit.

On the figures 4, 5 there are shown the diagrams of measuring of chlorophyll fluorescence intensity for two sort of vine plants (PxP 101-14 and Kober 5BB) during 5 months. The vine plants were under drought influence and normal conditions.

Examples of the practical usage of fluorometer "Floratest" in the National Scientific Center "V.E. Tairov's Institute of viticulture and winemaking" and the graph of CFI on the device's display are shown on figure 6.

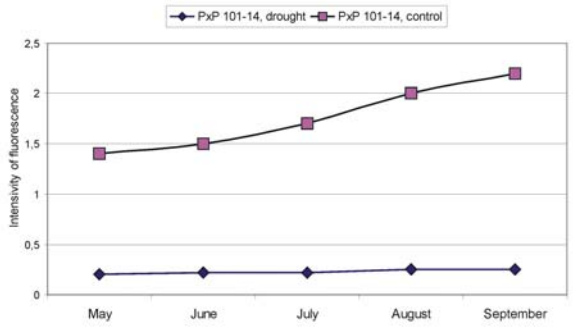


Figure 4. CFI intensity of vine plant (sort PxP 101-14) under drought and normal conditions

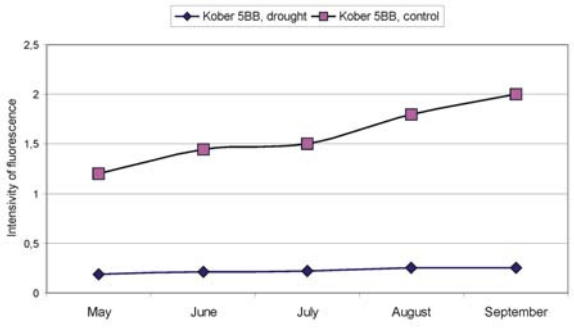


Figure 5. CFI intensity of vine plant (Kober 5BB) under drought and normal conditions

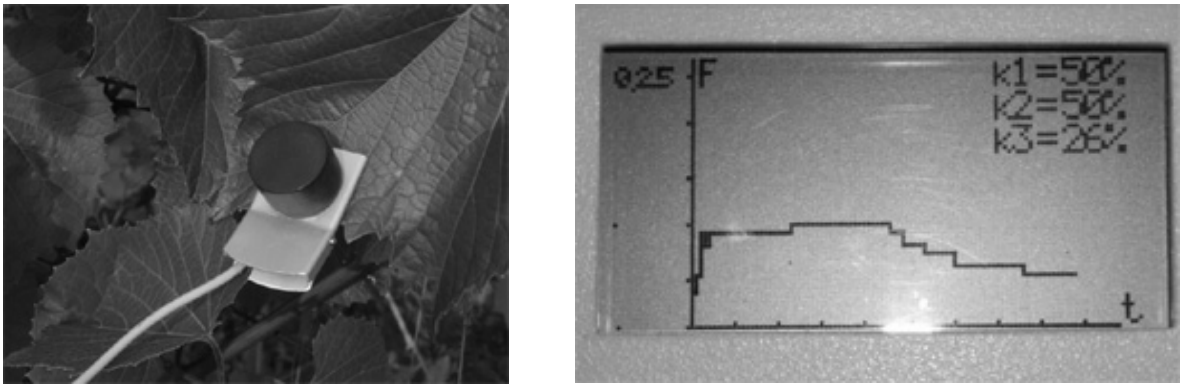


Figure 6. The sensor of the "Floratest" on the vine leaf and the image of CFI on the device's display

Experimental researches of fluorometer "Floratest" in National Scientific Center "V.E. Tairov's Institute of viticulture and winemaking" allow:

- determination of vine plants' state under the stress factor influence accordingly to the parameters of Kautsky curve;
- development of the recommendations on the fluorometer's software update and bringing output information on the device display to the recommendations, which accompany the Kautsky curve;
- development of recommendations on creation of the set of removable sensors, using which both detection of stress factors and express-diagnostics of plant disease can be performed.

Conclusions

- on the basis of modern information technologies and achievements in field of biosensorics original noninvasive portable fluorometer for express-diagnostics of plant state under stress conditions was developed;
- during the fluorometer designing and fast software and hardware tools adaptation to the conditions of exploitation the methods of virtual design created in the Institute of Cybernetics of NAS of Ukraine as a part of virtual laboratory for computer-aided design were used extensively;
- during experimental researches in National Scientific Center "V.E.Tairov's Institute of viticulture and winemaking" of Academy of Agrarian Sciences of Ukraine there were developed methodical tools which allow evaluating the state of vine plants under drought conditions and conditions of insufficient water capacity in express-mode.

Bibliography

[Kautsky, 1931] Kautsky H., Hirsch A. Neue Versuche zur Kohlenstoffassimilation // Natur wissenenschaften. – 1931. – 19. – S. 964

[Kautsky, 1934] Kautsky H., Hirsch A. Das Fluoreszenzverhalten gruner Pflanzen // Biochem Z. – 1934. – 274. – S. 422–434.

[Romanov, 2007] Romanov V., Fedak V., Galelyuka I., Sarakhan Ye., Skrypnyk O. Portable Fluorometer for Express-Diagnostics of Photosynthesis: Principles of Operation and Results of Experimental Researches // Proceeding of the 4th IEEE Workshop on "Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications", IDAACS'2007. – Dortmund, Germany. – 2007, September 6–8. – P. 570–573.

[Fedack, 2005] Fedack V., Kytaev O., Klochan P., Romanov V., Voytovych I. Portable Chronofluorometer for Express-Diagnostics of Photosynthesis // Proceeding of the Third IEEE Workshop on "Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications", IDAACS'2005. – Sofia, Bulgaria. – 2005, September 5–7. – P. 287–288.

Authors' Information

Volodymyr Romanov - Head of department of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine, Doctor of technical sciences, professor; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: dept230@insyg.kiev.ua

Volodimyr Sherer – deputy director of National Scientific Center "V.E. Tairov's Institute of viticulture and wine-making" of Academy of Agrarian Sciences of Ukraine; Doctor of agricultural sciences; 40 let Pobeda Str., 27, Tairovo, Odessa, 65496, Ukraine; e-mail: iviv@te.net.ua

Igor Galelyuka – research fellow of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine

Yevgeniya Sarakhan – research fellow of National Scientific Center "V.E. Tairov's Institute of viticulture and wine-making" of Academy of Agrarian Sciences of Ukraine; 40 let Pobeda Str., 27, Tairovo, Odessa, 65496, Ukraine; e-mail: sarakhan2006@ukr.net

Oleksandra Skrypnyk – software engineer of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine

MATHEMATICAL MODEL AND SIMULATION OF A PNEUMATIC APPARATUS FOR IN-DRILLING ALIGNMENT OF AN INERTIAL NAVIGATION UNIT DURING HORIZONTAL WELL DRILLING

Alexander Djurkov, Justin Cloutier, Martin P. Mintchev

Abstract: Conventional methods in horizontal drilling processes incorporate magnetic surveying techniques for determining the position and orientation of the bottom-hole assembly (BHA). Such means result in an increased weight of the drilling assembly, higher cost due to the use of non-magnetic collars necessary for the shielding of the magnetometers, and significant errors in the position of the drilling bit. A fiber-optic gyroscope (FOG) based inertial navigation system (INS) has been proposed as an alternative to magnetometer -based downhole surveying. The utilizing of a tactical-grade FOG based surveying system in the harsh downhole environment has been shown to be theoretically feasible, yielding a significant BHA position error reduction (less than 100m over a 2-h experiment). To limit the growing errors of the INS, an in-drilling alignment (IDA) method for the INS has been proposed. This article aims at describing a simple, pneumatics-based design of the IDA apparatus and its implementation downhole. A mathematical model of the setup is developed and tested with Bloodshed Dev-C++. The simulations demonstrate a simple, low cost and feasible IDA apparatus.

Keywords: Mathematical Modeling, Measurement-While-Drilling, In-Drilling Alignment

ACM Keywords: Mathematical Modeling

List of Abbreviations

BHA	Bottom-hole assembly	INS	Inertial Navigation System
FOG	Fiber-optic gyroscope	MWD	Measuring-while-drilling
IDA	In-drilling alignment	ZUPT	Zero velocity update
IMU	Inertial Measurement Unit		

Nomenclature:

a	Orifice area (m^2)	Pa	Air Pressure in Chamber A
Aa	Piston area enclosing Chamber A (m^2)	P_b	Air Pressure in Chamber B
Ab	Piston area enclosing Chamber B (m^2)	R	Gas constant for air (287 J/kg/K)
c_p	Constant air pressure specific heat (1003.5 $Jkg^{-1}K^{-1}$)	$T_{a,\,b}$	Cylinder's chamber temperatures (K)
c_q	Orifice Discharge Coefficient	$T_{s,\,ex}$	Air tank temperatures (K)
c_v	Constant air volume specific heat (718.6$Jkg^{-1}K^{-1}$)	V_{da}	Chamber A dead volume (m^3)
m_a	Mass of air in Chamber A (kg)	V_{db}	Chamber B dead volume (m^3)
m_b	Mass of air in Chamber B (kg)	x	Displacement of piston
M	Combined mass of piston, piston rod and IMU (kg)	x_1	Cylinder's stroke

1. Introduction

1.1 Conventional Horizontal Drilling Techniques

Horizontal drilling features several advantages when it comes to oil exploration and production. First, it facilitates the accessibility of reservoirs in complex locations: under riverbeds, mountains and even cities [1]. Secondly, if a particular reservoir is characterized by a large surface area, but is distributed over a thin horizontal layer, a horizontal well will yield a larger contact area with the reservoir and thus lead to a higher productivity and longevity when compared to vertical ones [2]. Present applications of horizontal wells include intersecting of fractures; eliminating of coning problems in wells with gas and water coning problems; the improving of draining area per well in gas production, resulting in a reduction of the number of wells required to drain the reservoir; and providing larger reservoir contact area and enhancing injectivity of an injection well [3].

The drilling of a directional (horizontal) well begins by drilling vertically from the surface to a kick-off point at a predetermined depth. Then, the well bore is deviated intentionally from the vertical at a controlled rate. To

implement this complex drilling trajectory, measurement-while-drilling (MWD) equipment, steerable setup and surveying sensors must be incorporated within the drilling assembly [4]. The drilling assembly utilizes a diamond bit and a mud turbo-drill motor installed in front of a trajectory control sub, nonmagnetic drill collars which include the magnetic surveying sensors, and a drill pipe [5], (Fig.1).

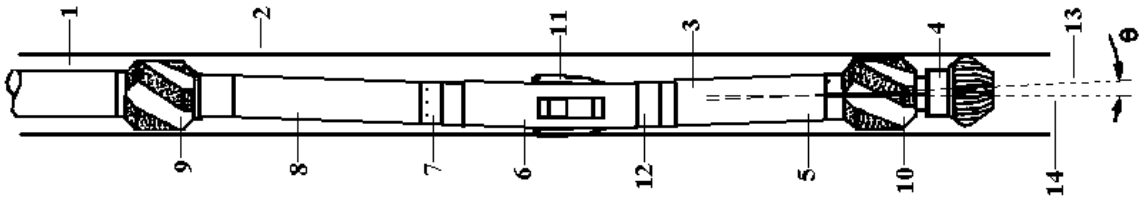


Fig.1: Drilling Assembly: 1 – drill string, 2 – borehole, 3 – bottom hole assembly (BHA), 4 – drill bit, 5 – drilling motor, 6 – trajectory control sub, 7 – bypass sub, 8 – MWD tool included in nonmagnetic collars, 9, 10 – upper and lower stabilizers for centering the drilling assembly in the borehole, 11 – stabilizer blades, 12 – induced bend to provide angular offset (θ) between the axis of the drill bit (13) and the center line (14).

1.2 Principles of Magnetic Surveying

The conventional measurement-while-drilling (MWD) surveying system presently utilizes three-axis accelerometers and three-axis magnetometers fixed in three mutually orthogonal directions [13]. At a certain predetermined surveying stations, the drilling assembly is brought to rest. At that point, the body frame of the MWD surveying system, formed by the axes of the accelerometers and magnetometers, is an angular transformation of the reference (North-East-Vertical) frame. Since the position of the bottom-hole assembly (BHA) is known, the direction and magnitude of Earth's acceleration are known as well. By comparing the acceleration vector formed from the measurements of the three accelerometers with the known vector of Earth's gravitational acceleration in the reference frame, the pitch (θ) and row (Φ) can be calculated (Fig.2) [7].

Then, the measurements from the magnetometers are combined with the calculated pitch and row to determine the azimuth angle (Ψ). The BHA trajectory is then computed by assuming a certain trajectory between the two successive stations.

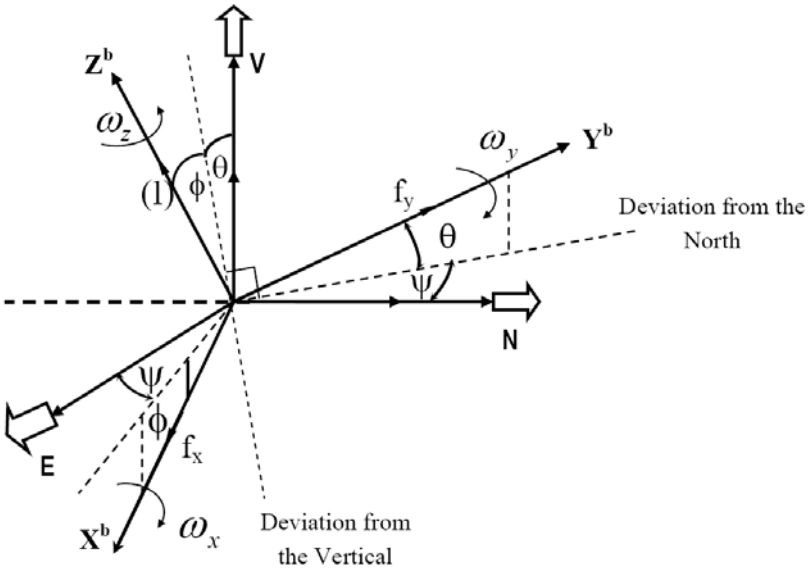


Fig.2: Orientation of the MWD magneto-surveying system with respect to North, East, and Vertical directions: the pitch (θ), the roll (Φ), and the azimuth (Ψ). In the drawing, X^b, Y^b and Z^b form the body frame, with its axes coinciding with the axes of the accelerometers and magnetometers. E, N, and V denote East, North, and Vertical and form the reference frame. The measured accelerations along the axes x, y and z of the body frame are respectively f_x, f_y, and f_z. The measured angular rates in the body frame about the x, y and z axes are respectively ω_x, ω_y, and ω_z.

1.3 Problems with MWD Magneto-Surveying System

Several external factors affect the performance of the magnetic surveying sensors. Such factors encompass the presence of randomly located ore deposits and geomagnetic influences. Moreover, the dynamic behavior of the magnetometers is negatively affected by magnetic interferences from the drill string. This requires the utilization of nonmagnetic collars for protecting the magnetic sensors. Although the accuracy of the magneto-surveying system increases with length of the nonmagnetic collars, this results in heavier and more costly MWD apparatus. Additionally, another source of error is introduced. Since the surveying sensors are located approximately 20 meters away from the drill bit, some rotations of the near-drill bit assembly may not be recognized [6].

1.4 Review of Current Inertial Navigation System (INS) -Based Navigation

In order to avoid the problems associated with magnetometers and non-magnetic collars, an INS based inertial measurement unit (IMU) incorporating a single fiber-optic gyroscope (FOG) and three-axis accelerometer has been proposed [7]. The INS determines the position, velocity and orientation of the drilling assembly in three-dimensional space by integrating the measured components of the acceleration (provided by the accelerometers) and the angular velocity (provided by the gyroscope). However, due to the small errors in the measurements of the accelerometers and the fiber-optic gyroscope, a continuous error growth in the position and the velocity of the BHA is observed [8]. Several approaches to limit this error growth have been proposed.

The first approach is based on continuous surveying with the aid of velocity and altitude updates through a Kalman filter. It has been reported that this method yields an inclination and azimuth angle errors of less than 0.4º and 1º, respectively, over a two-hour experiment. Moreover, the altitude errors have not exceeded ±0.5m over the entire experiment, while the errors along the East and North directions, dependant on the accelerometer bias, have been kept less than 50m and 20m respectively, over a two hour experiment [8].

The second approach was applied when velocity updates were not available. The approach involved the interrupting of the BHA motion at some predetermined station to apply the velocity zero update (ZUPT) for resetting the velocity errors and stopping the growth of position errors. The ZUPT approach was associated with position errors of less than 25m and 100m along the East and North directions respectively [8]. However, these results did not show substantial advantage over standard magnetic surveying.

A third method, called the In-Drilling Alignment Method (IDA), involves the induction of motion on the IMU in the horizontal North-East plane, while the entire bottom-hole assembly (BHA) is at rest. If the acceleration of the IMU at any time during the induced motion is known more precisely than the accuracy of the accelerometers on the IMU, the observations may be used as acceleration updates to align the accelerometers. Separately, an angular motion of the IMU about the axis of its gyroscope may be induced with accurately known angular rate and be used as an update for the gyroscope [9]. Such an IDA apparatus that will perform effectively in bore-hole drilling conditions has not been designed.

The aims of this paper are: (1) to design an In-Drilling Alignment apparatus for testing this newly-proposed concept; and (2) to mathematically model the expected results provided by such an apparatus.

2. Methods and Materials

2.1 Inducing Motion on the IMU in the North-East Horizontal Plane

A pneumatically-based solution is proposed for inducing a motion on the IMU in the North-East horizontal plane while the BHA is at rest. A compact, cylindrical capsule containing an IMU, RF transmitter and a small battery to power the IMU and the transmitter is attached to the end of a piston rod of a pneumatic cylinder via a bearing. The bearing allows the capsule to rotate freely around the cylinder's rod. By correctly regulating the pressure on each side of the piston, desired linear accelerations of the piston rod-IMU assembly can be obtained (Fig.3).

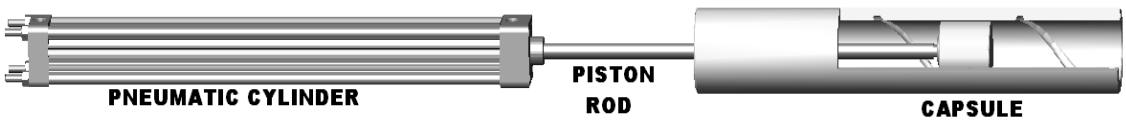


Fig.3: In-Drilling Alignment Apparatus

This linear motion can further be employed for inducing an angular motion on the IMU about the axis of one its gyroscopes. On the exterior surface of the cylindrical capsule, around its axis, ball bearings are positioned in a helical pattern. Similar helical thread is machined on the inner side of a pipe, to allow the bearings on the capsule to smoothly traverse along it. Thus, any linear motion induced on the capsule by the pneumatic cylinder will simultaneously cause an angular motion. If the linear acceleration of the IMU-containing capsule and the angular step of the helical thread are accurately known, then the angular acceleration of the capsule can be calculated easily. This in turn can be integrated to yield the angular rotation rate of the capsule (Fig.4).

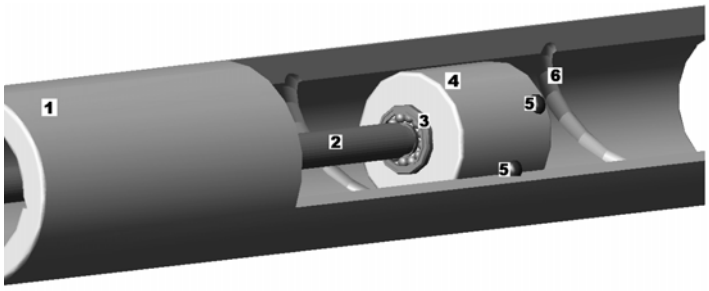


Fig.4: Schematic of the angular motion inducing mechanism: 1-pipe, 2-pneumatic cylinder rod, 3-bearing, 4-capsule enclosing IMU, battery and RF transmitter, 5-ball bearings aligned in a helical pattern over the surface of the capsule, 6-helical thread machined on the interior surface of the pipe.

2.2 Monitoring the Induced Motion of the IMU

The principle of the magnetostrictive effect is employed for monitoring the position of the piston in the pneumatic cylinder. For this purpose, the piston is equipped with tiny magnets, and a special piston position-sensing unit is installed along the cylinder (Fig.5) [11].

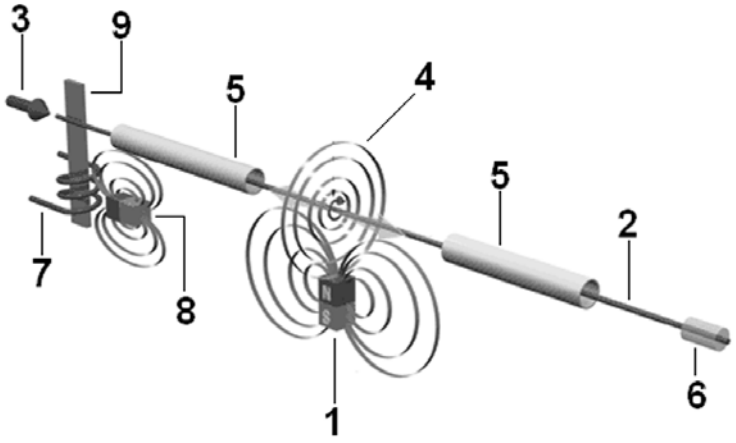


Fig. 5: Schematic of the operation of a magnetostrictive effect-based piston position sensing unit: 1-piston magnet, 2-waveguide, 3-short current pulse, 4-magnetic field around the waveguide due to the current pulse (3), 5-protective casing, 6-dampener, 7-mechanical wave detecting coil, 8-magnet providing a magnetic field in which the detecting coil is located (7), 9-strip along which the deformation wave is transmitted to the coil.

The unit consists of a “waveguide” made of a special nickel-alloy tube through which runs a copper wire. The initiation of a measurement is denoted by a short electric pulse through this wire, which sets up a circular magnetic field around it. At the point along the “waveguide” where the produced field intersects the perpendicular magnetic filed due to the magnets located in the piston of a pneumatic cylinder, an elastic deformation of the nickel-alloy tube is caused according to the magnetostrictive effect. The component of the deformation wave that traverses the “waveguide” toward its back end is dampened, while the component that arrives at the signal converter is transformed into an electric pulse. Since the travel time for the pulse is directly proportional to the

position of the magnetic piston [11], by determining the elapsed time between the initiating pulse and received pulse, the piston's position can be estimated with high accuracy in the order of 5μm [11].

Once the position of the piston is accurately known, a differentiation yields its velocity and acceleration. However, since the IMU capsule is affixed to the piston rod of the pneumatic cylinder, its linear component of motion is completely defined. Moreover, the angular rate of the IMU around the axis of the pneumatic cylinder can be calculated according to:

$$\omega = \boldsymbol{v} \cdot \lambda \tag{1}$$

where (ω) is the angular speed, (v) is the linear speed and (λ) is the angular step of the machined helical thread.

2.3 Pneumatic Setup of the IDA Apparatus

The following simplified pneumatic setup is proposed for inducing and controlling the motion of the IMU (Fig.6).

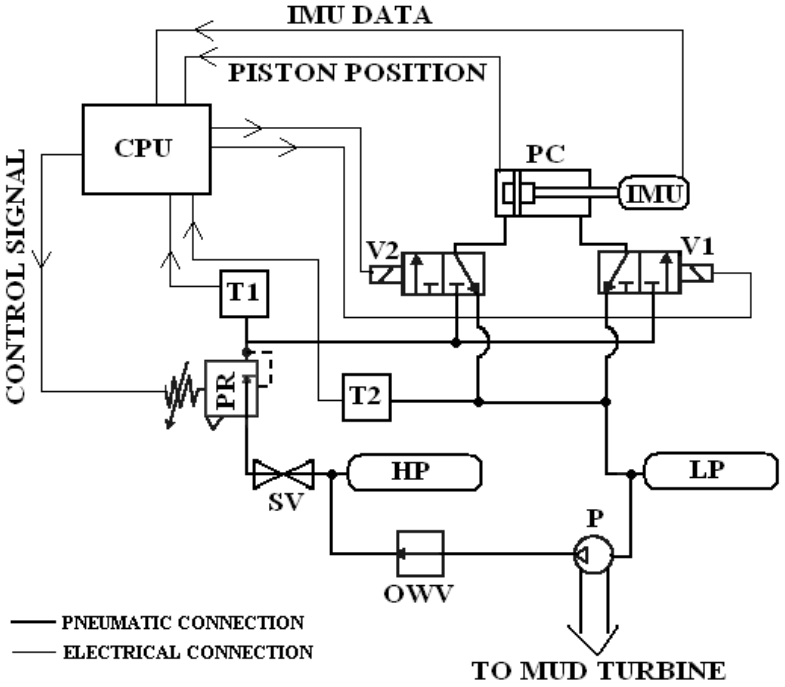


Fig.6: Pneumatic System Setup: HP-high pressure air tank; LP-low pressure air tank; PC-pneumatic cylinder, cushioned at both ends; V1, V2-two way solenoid valves; PR-proportional electric pressure regulator, T1, T2-electric pressure transducers, SV-shutoff valve; OWV-one-way air valve; P-air pump.

Initially, the system comprises a high (HP) and a low (LP) pressurized air tanks. The Central Processing Unit (CPU) can independently control the two solenoid valves (V1) and (V2) through which the pneumatic cylinder is connected to the rest of the pneumatic system. By feeding the appropriate signals to the two valves, the right chamber of the cylinder may be connected to the low-pressurized air tank, and the left to the highly-pressurized (HP) air tank via the electronic pressure regulator (PR). Then the two electric pressure transducers (T1) and (T2) inform the CPU of the air pressure in each chamber of the cylinder. Based on this information, the CPU calculates the necessary regulated pressure and controls the proportional regulator (PR). Once a pressure differential is established across the piston, a linear acceleration on the piston-IMU assembly is induced. A measurement of the piston's position is supplied to the CPU by the magnetostrictive effect-based measuring unit. The three acceleration components and angular rates measured by the IMU are also passed to the CPU where, together with the position of the piston, the data is processed mathematically to align the IMU.

Once the piston of the pneumatic cylinder is near the end of its stroke, the CPU reverses the valves (V1 and V2) and an opposite acceleration is induced. Cushions are provided on both sides of the piston to reduce the severity of the impact with the cylinder's walls.

Eventually, the pressures in the two air tanks will equalize, limiting the number of piston cycles and thus the number of alignment data points. To restart the system, the mud-powered air pump is turned on to pressurize the HP air tank to its initial high pressure. This in turn will bring the LP tank back to its original low pressure. Air is pumped from the LP tank to the HP tank through a special one-way air valve (OWV) that will prevent air from leaking back to the LP tank through the pump P. This resetting procedure is only possible when there is mud flow. Thus, it will be performed during the drilling process. The IDA process takes place when the bottom-hole assembly is at rest.

2.4 Data Manipulation and Transmission

Since the IMU is constantly in motion during the IDA process, wiring the IMU will be impractical and will result in constant stress applied to the wires. To eliminate such problems, RF link is proposed between the IMU and a local receiving module mounted on the exterior surface of the tube through which the IMU is accelerated. Thus, the three components of acceleration and angular rate measured by the IMU are sent to a local RF receiving module and then, together with the cylinder's piston position are wired to the CPU. There, the data is mathematically processed to determine the position of the BHA in the horizontal North-East frame. It is then send to the surface by the conventional method of mud pulse telemetry [3].

2.5 Mathematical Model of the Pneumatic System

To model the pneumatic system extensively, first a model of the pneumatic cylinder for its specific application will be derived. Throughout the entire model, all pneumatic processes are assumed to be adiabatic and the fluid (gas) is treated ideally. It has been shown that such assumptions still provide excellent results for similar applications, while greatly simplifying the model [10].

Let the cylinder be divided into two separate chambers A and B. Also, assume that the piston is moving to the right with speed v, (Fig.7).

The pressure change in chamber A is described by [10]:

$$\dot{P}_a = \left(\dot{m}_a - \frac{P_a A_a}{RT_s} \dot{x} \right) \frac{c_p R T_s}{c_v \left(V_{da} + \left(\frac{x_1}{2} + x \right) A_a \right)} \tag{2}$$

where m_a and P_a are the mass of gas and pressure in chamber A respectively, and A_a is the area of the piston's surface enclosing chamber A.

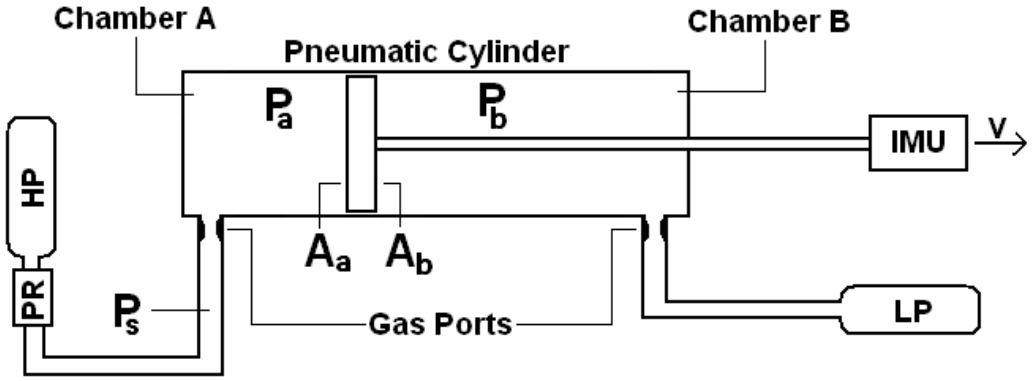


Fig.7: Supplying air to the cylinder: HP-high-pressure tank; LP-low pressure tank, PR-pressure regulator; P_a, P_b – pressure in chamber A and B respectively; P_s-supplied pressure by regulator (PR), A_a, A_b – area of piston common to chamber A and B respectively.

The position of the piston in the cylinder is denoted by x, while x_1 denotes the cylinder's stroke; Vda is the dead volume entitled to chamber A (tubing volume and unused cylinder volume). The temperature of the supplied gas is T_s, and c_p and c_v stand for the constant pressure and volume specific heats of the gas respectively; R is the gas constant. The rate of change of mass of gas in chamber A is given by [10]:

$$\dot{m}_a = \frac{c_q a P_s}{\sqrt{T_s}} \sqrt{\frac{2.8}{R(\gamma - 1)} \left[\left(\frac{P_a}{P_s} \right)^{\frac{2}{\gamma}} - \left(\frac{P_a}{P_s} \right)^{\frac{\gamma+1}{\gamma}} \right]} \quad (3)$$

In (3), c_q is the flow discharge coefficient of the pneumatic cylinder's inlet, a is the area of the inlet; and γ is the specific heat ratio. Similarly, the pressure change model for chamber B is [10]:

$$\dot{P}_b = \left(\dot{m}_b + \frac{P_b A_b}{R T_s} \dot{x} \right) \frac{c_p R T_s}{c_v \left(V_{db} + \left(\frac{x_1}{2} - x \right) A_b \right)} \quad (4)$$

where the variables correspond to the ones defined in Eq.(2), but applicable to chamber B. The rate of change of gas mass in chamber B is quantified similarly [10]:

$$\dot{m}_b = \frac{c_q a P_b}{\sqrt{T_b}} \sqrt{\frac{2.8}{R(\gamma - 1)} \left[\left(\frac{P_{ex}}{P_b} \right)^{\frac{2}{\gamma}} - \left(\frac{P_{ex}}{P_b} \right)^{\frac{\gamma+1}{\gamma}} \right]} \quad (5)$$

where, T_b is the temperature of chamber B, and P_{ex} is the exhaust pressure (pressure of LP tank).

Furthermore, the supplied pressure P_s that appears in Eq. (3) is the regulated pressure that comes from the proportional pressure regulator PR (Fig.6). However, since P_s is estimated by the CPU based only on the readings of the two pressure transducers T1 and T2 (Fig.6), it can be concluded that:

$$P_s = f(T_1, T_2) \quad (6)$$

Additionally, the motion of the IMU-piston assembly can be modeled by [10]:

$$M(\ddot{x} + g') + D\dot{x} = P_a A_a - P_b A_b + \hat{x} k \Delta \quad (7)$$

where M is the total mass of the IMU-containing capsule, piston and rod; x is the position of the piston inside the cylinder; D is some constant dependant on the materials used and the construction of the apparatus; g' is the component of Earth's acceleration parallel to the direction of induced motion on the IMU; k is the elasticity constant for the front and rear bumpers of the piston, and Δ is the change in length of the bumper. Equations 1-7 now completely define the pneumatic system for inducing a linear and angular motion on the IMU.

2.6 Materials

In order to implement the proposed design, the following materials and components were sourced.

- Pneumatic Cylinder (Cat. No. 2.00CJ2MABUS14AC20, Parker Pneumatics, Calgary, Alberta) with magnetostrictive linear position sensor (Cat. No. BTL5M1M0500RSU022KA02, Parker Pneumatics, Calgary, Alberta)
    - Cylinder Bore: 50.8mm
    - Cylinder Stroke: 508mm
    - Both sides cushioned magnetic piston:
        - Simulated Elasticity Constant(k): 20000N/m
        - Simulated Cushion Thickness: 5mm
    - Inlet/Outlet Air Ports
        - Flow Discharge Coefficient: 0.9
        - Port Cross-Section Area: $1.96*10^{-5}$ m^2
    - Dead Volumes
        - Chamber A/B : $1.96*10^{-3}$ m^3
- Electronic Proportional Pressure Regulator (Cat. No. PAR-15 W2154B179B, Parker Pneumatics, Calgary, Alberta)

- Analog Voltage Control (0-10V)
- Simulated Pressure Regulating Function:

```
Arguments (High pressure chamber (HP), Low pressure chamber (LP))
{
if ( HP-LP < 2000Pa AND LP+20kPa < pressure of high-pressure tank)
{
 Regulated Pressure = LP+20kPa
 }
else {Regulated Pressure = HP}
}
```

- Micro-electromechanical (MEM) Inertial Measurement Unit (MEMSense 2693D, Rapid City, SD)
  - Accelerometers (A50)
    - Dynamic Range: ±50g
    - Drift: 0.3g
  - Gyroscopes (-1200C050)
    - Dynamic Range: ±1200º/s
  - Magnetometers (not utilized in the proposed design)
    - Dynamic Rang: ±1.9G
    - Drift: 2700ppm/ ºC
  - Absolute Maximum Ratings:
    - Operation Temperature: -40 ºC to 85 ºC
    - Acceleration (Shock): 2000g for 0.5ms

3. Results

3.1 Motion of the Piston-IMU Assembly

According to the derived model of the pneumatic system and the outlined parameters of each component, a C++ simulation (Bloodshed Dev C++, Bloodshed Software, www.bloodshed.net/devcpp.html) revealed the position of the piston in the pneumatic cylinder as a function of time (Fig.8).

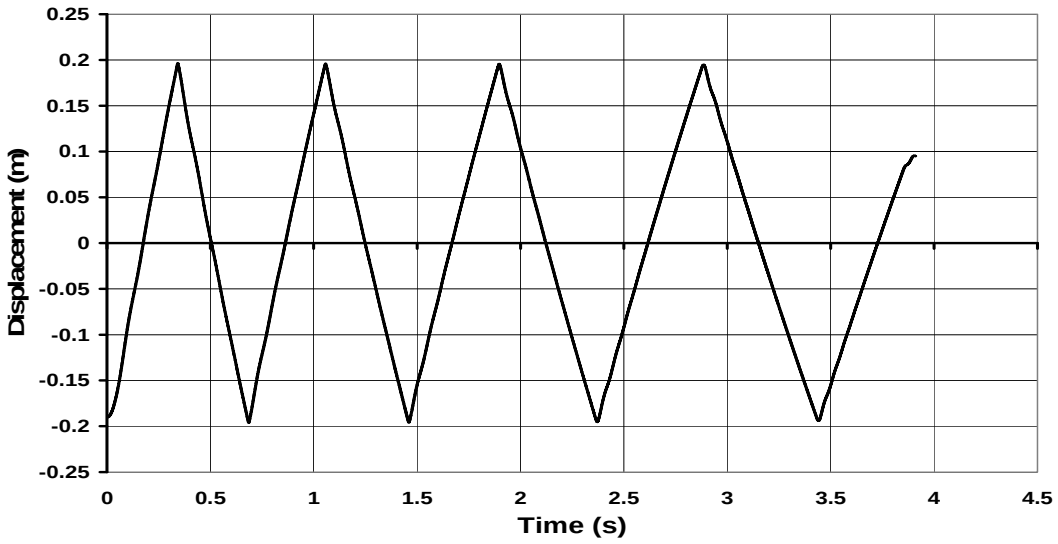


Fig.8: The displacement of the piston inside the pneumatic cylinder as a function of time. The displacement is with respect to the middle of the stroke of the cylinder.

Figure 8 demonstrates that a tank, initially pressurized to ten atmospheres will allow the completion of four full cycles in less than 3.5 seconds. The piston can be then brought to rest during the fifth cycle and locked in place by completely closing the inlet and outlet ports of the cylinder. The acceleration of the piston-IMU assembly was also simulated over the duration of a full cycle (Fig.9).

The constantly changing acceleration of the piston (Fig.9) is due to the specifically implemented function in the simulation, relating the two electronic pressure transducer outputs to the regulated pressure adjusted by the proportional pressure regulator. For a sampling rate of 400Hz, the time intervals of 0 to 0.3 seconds and 0.35 to 0.6 seconds will be proper choices for observations source. The data obtained in these time intervals can then be utilized in aligning the IMU sensors. However, a more gradually changing acceleration of the piston is desired in order to align the IMU more accurately.

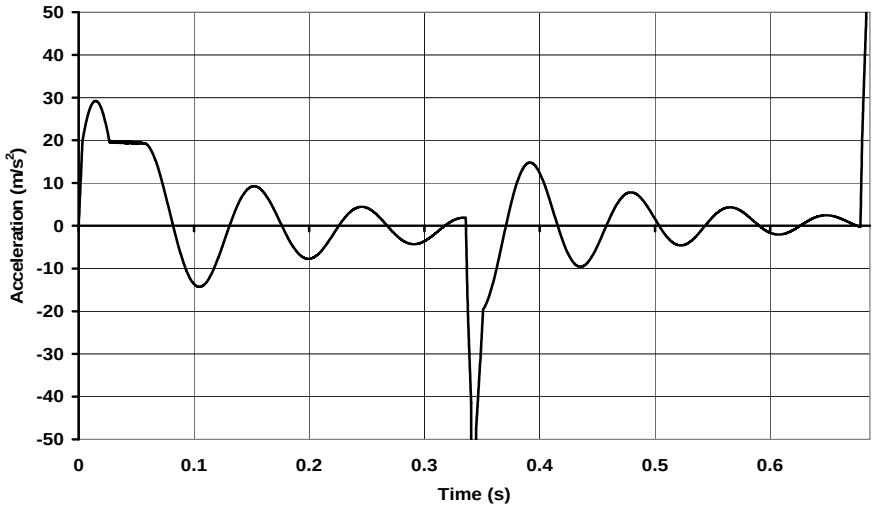


Fig.9: Piston's acceleration as a function of time during one full cycle. The acceleration peaks at 0.34s and 0.68s correspond to the accelerations experienced by the IMU-piston assembly when the piston's bumper collides with the cylinder's wall.

The pressure in each tank as a function of time during the entire induced motion process has also been explored (Fig.10).

It is clearly evident that after 3.8s (for the outlined system parameters), the pressures in the two tanks will equalize, and the induced motion will come to an end. At this point, the mud-powered air pump is turned on to pressurize the high-pressure tank to its initial value. Although the currently implemented pressure regulating function will yield economical use of the fluid (air), a function that will provide more gradual accelerations of the piston is desired.

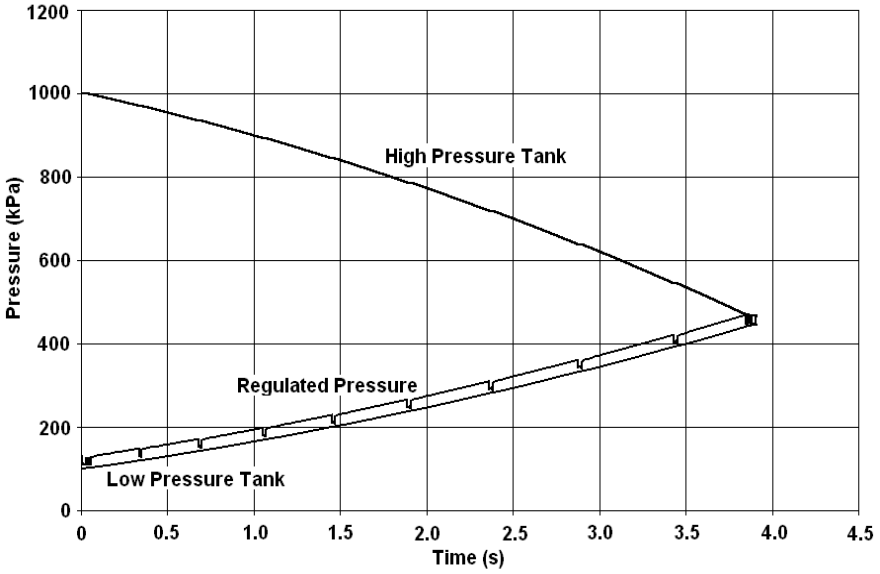


Fig.10: Output of the proportional pressure regulator, and pressures of the high and low-pressure tanks as a function of time over the entire induced motion process.

4. Conclusion

This article focused on designing and quantifying an apparatus that will allow for an effective, simple and low cost aligning of the sensors of an Inertial-Measurement Unit for continuous angle attitude angle information delivery in a downhole drilling environment. A pneumatic solution was proposed, comprising an air-cylinder, two air tanks, air-pump and a proportional pressure regulator. The highly pressurized air-tank is discharged into the low-pressure tank through the air-cylinder. Correct control of the pressure on each side of the piston of the air-cylinder yields the desired accelerations of the IMU-piston assembly. The position of the piston is constantly monitored by a magnetostrictive sensor, which in turn is differentiated to give the acceleration of the IMU-piston assembly. Moreover, by moving the IMU along a helical thread, angular motion is induced on it, whose angular acceleration is a simple function of the linear acceleration. Once the IMU's angular and linear motion components are known, they are utilized in aligning the unit.

A mathematical model of the entire pneumatic system was derived and simulated with C++. It was shown that an air tank with initial pressure of ten atmospheres will yield more than four full alignment cycles of the IMU-piston assembly within a timeframe of four seconds. The induced accelerations on the IMU-piston assembly were in the range of 3g's, except during a collision with the walls of the air-cylinder, where they reach 80g's. Despite the fact that the model showed a feasible design in downhole conditions, a pressure regulating function that will allow more gradual induced accelerations is desired.

Bibliography

[1] J. Burkmann and N. Nickels, "Directional, navigational and horizontal drilling techniques," *Geothermal Resources Council Bull*., vol. V19, no.4, pp. 106-112, 1990.

[2] S.D. Joshi and W. Ding, "The cost benefits of horizontal drilling," In *Proc. American Gas Association*, Arlington, VA, Apr.-May 29-1, 1991, pp.679-684.

[3] S. D. Joshi, "Horizontal Well Technology", *Technology and Industrial Arts*, PennWell Books, 1991

[4] E.K. Fisher and M.R. French, "Drilling the first horizontal well in the Gulf of Mexico: A case history of East Cameron block 278 well B-12," in *Proc. 66th SPE Annu. Technical Conf. Exhibition*, Dallas, TX, Oct.6-9, 1991, pp.111-123

[5] C. Walker, "Drill Bit Steering," US. Patent 5311953, May 17, 1994

[6] B.A. Shelkholeslami, B.W. Schlottman, F.A. Siedel, and D.M. Button, "Drilling and production aspects of horizontal wells in the Austin Chalkl," *J. Petroleum Technol*., pp.773-779, Jul.1991.

[7] A. Noureldin, "New measurement-while drilling surveying technique utilizing a set of fiber-optic rotation sensors," Ph.D. Dissertation, Dept. Elect. Eng., Univ. Calgary, Calgary, AB, Canada, 2002.

[8] A. Noureldin, D. Irvine-Halliday, and M.P. Mintchev, "Accuracy Limitations of FOG-Based Continuous Measurement-While-Drilling Surveying Instruments for Horizontal Wells," *IEEE Trans. On Instr. And Meas.*, vol.51, no.6, Oct. 2002.

[9] E. Pecht, "INS In-Drilling Alignment for improving Observability in Horizontal-Directional Drilling,"Ph.D. Dissertation, Dept. Elect. Eng., Univ. Calgary, Calgary, AB, Canada, 2005.

[10] R. Richardson, A.R. Plummer, M. Brown, "Modeling and simulation of pneumatic cylinders for a physiotherapy robot," School of Mech. Eng., University of Leeds, UK,

[11] O. Sound, "Linear Position Sensor Option for Series 2MA Cylinder," Parker Hannifin Corporation, Des Plaines, IL USA

[12] P. Tubel, C. Bergeron, S. Bell, "Mud pulse telemetry system for downhole measurement-while-drilling," *IEEE Instr. And Meas. Tech. Conf.*, 1992, p 219-23

[13] J.L. Thorogood and D. R. Knott, "Surveying techniques with a solid state magnetic multi-shot device," in *Proc. SPE/IADC Drilling Conf*., New Orleans, LA, Feb. 28-March 3, 1989, pp.841-856.

Authors' Information

Alexander Djurkov – Department of Electrical and Computer Engineering, University of Calgary, Alberta, Canada, T2N 1N4. Phone: (403) 244-2298; e-mail: alexsd_bg@yahoo.co.uk

Justin Cloutier – Imperial Oil Ltd., Calgary, Alberta, Canada; Department of Electrical and Computer Engineering, University of Calgary, Alberta, Canada, T2N 1N4; Phone: (403) 220-2191.

Martin P. Mintchev – Prof., Dr., Department of Electrical and Computer Engineering; University of Calgary; Calgary, Alberta, Canada, T2N 1N4; Department of Surgery, University of Alberta; Edmonton, Alberta T6G 2B7; Phone: (403) 220-5309; Fax (403) 282-6855; e-mail: mintchev@ucalgary.ca

MODELING OPTICAL RESPONSE OF THIN FILMS: CHOICE OF THE REFRACTIVE INDEX DISPERSION LAW

Peter Sharlandjiev, Georgi Stoilov

Abstract: Determination of the so-called optical constants (complex refractive index N, which is usually a function of the wavelength, and physical thickness D) of thin films from experimental data is a typical inverse non-linear problem. It is still a challenge to the scientific community because of the complexity of the problem and its basic and technological significance in optics. Usually, solutions are looked for models with 3-10 parameters. Best estimates of these parameters are obtained by minimization procedures. Herein, we discuss the choice of orthogonal polynomials for the dispersion law of the thin film refractive index. We show the advantage of their use, compared to the Selmeier, Lorentz or Cauchy models.

Keywords: Thin films; Materials and process characterization

ACM Classification Keywords: J.2 Physical Sciences and Engineering (Physics)

Introduction

The problem of estimation of the optical parameters of thin films: physical thickness (D) and complex refractive index N = n – i*k (real refractive index (n) and extinction coefficient (k)) is challenging from mathematical point of view and has technological and scientific importance. Usually, n and k are unknown functions of the wavelength (λ). The task is to evaluate them by the use of measurable quantities, such as film transmittance (T), front side reflectance (R) and/or backside reflectance (R′). Different methods have been proposed but no one has shown yet absolute advantage over the others. We can say that estimation of thin films optical parameters is more of an art, than scientific analysis. There are several steps that have to be fallowed: a) creation of a model, which describes the optical behavior of the film; b) collecting empirical data; c) fitting the postulated model to the data; d) evaluation of the results. The model of the wavelength dependence of the refractive index is of crucial importance: it defines the number of the unknown parameters and their functional relation. Some of the most popular models are named after the scientists that have proposed them: Cauchy, Drude, Selmeier, Lorentz, etc. Cauchy dispersion law is purely empirical:

$$n(\lambda) = A_0 + \frac{A_1}{\lambda^2} + \frac{A_2}{\lambda^4} + ... ,$$

where A_0, A_1, A_2,… are parameters to be determined. The number of terms can reach 10 – 15. Selmeier dispersion is semi-empirical:

$$n(\lambda) = \sqrt{A_0 + \frac{A_1\lambda^2}{\lambda^2 - B_1^2} + ...} ,$$

where A_0, A_1, B_1,… are parameters to be determined. More terms can be added for different oscillator positions.

Once the model is assumed, minimization techniques are applied to estimate the unknown parameters to the optical response of the thin film.

Here we shall consider the use of orthogonal polynomials in the dispersion law representation. We shall simulate a measurable quantity (transmittance) with predefined wavelength dependence of the complex refractive index. Then we shall fit the simulated data to different models of refractive index. Parameters in the dispersion law will be estimated, comparing Cauchy, Selmeier and orthogonal polynomials (OP) approaches.

Models and Computational Procedures

We shall consider a thin homogeneous film with wavelength dependence of the complex refractive index and physical thickness of 350 nm. The spectra of $n(\lambda)$ and $k(\lambda)$ are shown in Figure 1a and 1b, respectively. The choice of $n(\lambda)$ and $k(\lambda)$ is characteristic for many optical materials, such as amorphous semiconductors.

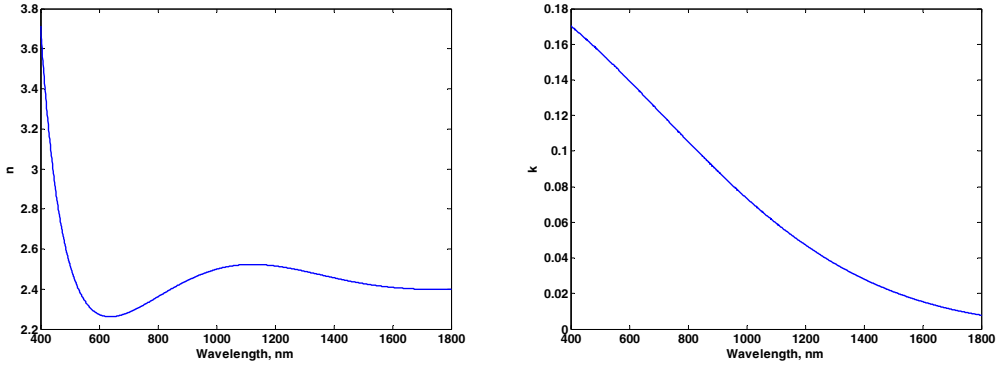


Figure 1. Spectral dependence of the refractive index (a) and extinction coefficient (b)

The simulated measurement is the spectrum of transmission in VIS at normal incidence of light calculated by the help of Abelès characteristic matrix [1], Figure 2. The measurable quantity is T ~ tt* (* stands for complex conjugate). The amplitude transmittance t is a complex quantity, related to N and D by transcendental equations [1]. During the fitting calculations, we have assumed that the physical thickness D and the extinction coefficient k(λ) are known, so that there is no fitting on these quantities. In this way we have strongly reduced parameter interactions. We have used associated Legendre functions *P(m,p;λ)* of degree *p* and order *m = 0,1,...,p*, as orthogonal polynomials. Thus, the dispersion law for n(λ) stands as:

$$n(\lambda) = A_0 P(0, p; \lambda) + A_1 P(1, p; \lambda) + A_2 P(2, p; \lambda) + ...$$

For Cauchy, Selmeier or OP dispersion laws, 8-9 coefficients are need for relevant representation of the refractive index n(λ), shown in Figure1a. The nonlinear data-fitting problem is solved by the Levenberg-Marquardt method (unconstrained minimization) [2].

Results and Discussion

Calculations of the film optical response and preliminary fits showed that Selmeier model of the refractive index has to be disregarded: it cannot describe properly the 'experimental' data, shown in Figure 2. In order to compare the Cauchy and OP models, we have used 8 coefficients in their corresponding presentations, so that the degrees of freedom in the two cases are the same. Levenberg-Marquardt procedures demand initial guess of the unknown parameters. For each fit, we have put 7 of the coefficients equal to zero, while the first one is 20% up of its initial 'true' value. After the termination of the minimization, one step refinement of the estimations is undertaken as well. The results obtained with the two models differ significantly. The residual error of the fit with the Cauchy model is an order of magnitude greater. The residual error of the fit with OP reaches 0.1%, which is equal to the experimental uncertainty of high precision spectral instruments. This means that further improvement of the fit is meaningless. The minimization procedure is much faster in latter case – the consumed CPU time is ~20 times greater for the Cauchy case. In Figure 3 the relative errors of the estimated wavelength dependence of the refractive index are shown. This is an illustration of the performance of the Cauchy and OP models.

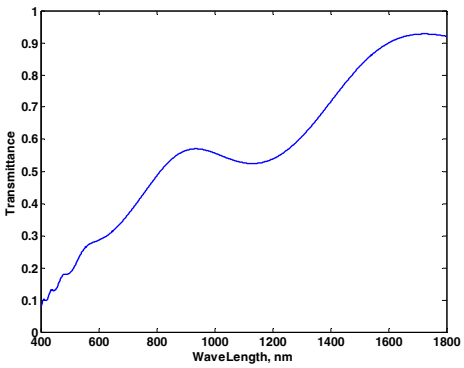


Figure 2. Calculated transmittance of the thin film as 'experimental' data for the fit

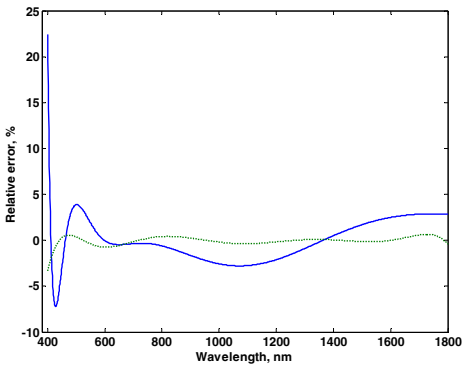


Figure 3. Relative error of refractive index fit: (dots) orthogonal polynomials; (line) Cauchy law

The advantages of fitting orthogonal polynomials to experimental data are well-known [2]. In our case, the situation is more complicated because of the nonlinear functional dependence of the target on the fitting parameters. However, the main advantage of this approach is sustained: due to the orthogonal property, each coefficient in the dispersion law representation can be determined independently from the others. If one has already obtained an evaluation of m^{-th} degree polynomial, an additional term in the dispersion law ((m+1) degree polynomial) requires only one new coefficient to be determined. The other coefficients remain the same, unlike in the Cauchy or Selmeier case. In the Cauchy case, high order polynomials may result in ill conditioned matrices. Besides, the joint confidence region in the parameter space, estimated by the covariance matrix, has minimum volume.

Conclusion

We have shown that the use of orthogonal polynomials in refractive index modeling is effective and highly productive. Although the involved coefficients have no physical meaning, this is also true for the Cauchy and Selmeier models. The OP principal feature is that the number of parameters to be fitted can be kept low at initial steps and then it can be increased, retaining the intermediate results. The orthogonal polynomials approach can be of use in many branches of material science, including photonic crystal design, optimization of elements for effective conversion of solar radiation, etc.

Acknowledgments

This work was partially supported by the National Science Foundation at the Ministry of Education of Bulgaria by grant D01-377/2006.

Bibliography

1. [Born, 1957] E. Born and P. Wolf, Principles of Optics, Ed. Princeton, New York, 1957.
2. [Himmelblau, 1970] D. Himmelblau, Process analysis by statistical methods, Ed. John Wiley & Sons, New York, 1970.

Authors' Information

Peter Sharlandjiev – Central Laboratory of Optical Storage and Processing of Information, BAS, Acad.G.Bontchev St. bl. 101, Sofia-1113, Bulgaria; e-mail: pete@optics.bas.bg .

Georgi Stoilov – Institute of Mechanics, BAS, Acad.G.Bontchev St. bl. 4, Sofia-1113, Bulgaria; e-mail: gstoilov@imbm.bas.bg.

FORMALIZATION OF INTERACTION EVENTS IN MULTI-AGENT SYSTEMS

Dmitry Cheremisinov, Liudmila Cheremisinova

Abstract: The problem of the description of interaction between spatially divided agents in the form of dialogues is explored. The concept of processes synchronization is analyzed to formalize the specification of interaction at the level of events constituting the processes. The approach to formalization of the description of conditions of synchronization when both the independent behavior and the communications of agents can be presented at a logic level is offered. It is shown, that the collective behavior of agents can be specified by the synthetic temporal logic that unites linear and branching time temporal logics.

Keywords: multi-agent system, interaction protocol, time.

ACM Classification Keywords: I.2.11 [Computer Applications]; Distributed Artificial Intelligence, Multiagent systems; D.3.3 [Programming Languages]: Language Constructs and Features – Control structures, Concurrent programming structures

Introduction

A multi-agent system can be considered as the organization of agents (by analogy to the human organization) or, in other words, as some artificial society. It is a computational system in which two or more agents interact or work together to perform a set of tasks or to achieve a set of goals [1]. One of the core concept of multi-agent systems is *interaction*, that is the foundation for cooperative behavior among several autonomous agents. Agent interactions are established through exchanging information in the form of messages that specify the desired performatives of interacting agents. Agent system can operate if the agents have a common understanding of the possible types of messages, then they must know which messages they can expect in a particular situation and what they may do when they got some message. So messages exchanged between agents in some multi-agent system need to follow some standard patterns which are described in agent interaction *protocol*.

Protocols play the central role in agent communication. An interaction protocol defines the rules the dialog among agents conforms to. It constrains the possible sequences of messages that may occur in agent interaction. Interacting agents should comply with an interaction protocol in order to engage permissible sequences of message exchange. When agent sends a message it can expect a response to be among a set of messages indicated by the accepted protocol. The interaction protocol can be assigned by the designer of the multi-agent system otherwise an agent needs to indicate the protocol that it wants to follow before it starts to interact with other members of the system.

It is necessary for any protocol itself to be correct and verifiable. If it is not correct then the agents that follow it may perform contradictory and unexpected actions leading to possible breakdown of the interaction. The central problem of the verification of interactions (dialogues of negotiations) that take place in open (not being cooperative) systems is the problem of conformance inspection between behavior of agents and interaction protocol. That is the protocol must be understandable by all agents of the system and the they behave according to this protocol. The implementation of conformance inspection confront with a problem of identification of dialogue steps between agents. Recognition of the dialogue step which is carried out by two spatially divided agents requires analyzing the concept of interaction of processes.

At the heart of the formal models of a protocol are cooperating sequential processes. Fundamental feature, the proposed protocol models differ, is the degree of synchronization of behaviors of participants of interaction. There is still a need for a proper formalism for the process of synchronization that is suitable for human understanding and automated implementation. In this paper we focus on logical analysis of synchronization of behaviors of interacting participants. The simple yet expressive class of interactions is considered, namely dialogues consisting of separate steps. The considered dialogues involve only two agents. This restriction allows concentrating on the kernel of the problem of synchronization in different formal models of interaction protocols. The agent interaction is considered as interaction between two (or more) processes. And a special case of such interaction is considered, when one of processes outputs at the same time as the other one inputs it. The actions of message exchanging have duration. The concept of a process and an event are analyzed to formalize the specification of interaction at the level of events.

Formalization of the concept of interaction event

Usually collective behaviour of multi-agent system is described as a dialogue of agents which communicate by means of sending and receiving messages. On each step of activity an agent carries out some action depending on its internal state and the received message. As a result of the action the agent changes its internal state and sends some messages to other agents. Speaking informally, the architecture of an agent includes 1) the internal structures of data defining internal states of the agent, 2) mail box containing messages from other agents, 3) integrity restrictions on the agent internal states, 4) actions which the agent can execute, and 5) the program that specifies the control of action execution. Execution of an action consists of 1) changing a current internal state of the agent and 2) sending a message to other agents. The current contents of a mail box consist of the messages received by the agent from other agents on the previous step. The global state of a multi-agent system consists of internal states and contents of mail boxes of all agents of the system [1].

To specify independent behaviors of agents, formalisms of high level abstractness are widely used, for example, such as temporal logic. At the same time the communications between agents are specified by means of the concepts of realization level, such as mail boxes and messages. One of the problems of such segregated approach to interaction lies in that it is extremely difficult to simulate interactions between agents though at the

same time the independent behaviour of separate agents is described completely. This problem arises due to the absence of agent model unifying all aspects of both independent behaviours and the communication. The main reason of the absence of such a general model is that there exists no general conceptual basis unifying all abstractions, connected with collective behaviour of agents.

When analyzing the behaviour of multi-agent system agents are characterized by processes. The process is specified by exhaustive description of potential behaviour of the agent. The process consists of events. Thus, to be in position to analyze the concept of interaction of processes, a suitable axiomatization of the concept of an event is required.

The concept of an event allows abstracting from physical time when describing behaviour of a system. The widespread axiomatization of an event is connected with the assumption, that events have no duration [2]. The behaviour of a multi-agent system consists of some events – steps of dialogue between agents – and is sequential in this sense. For recognition of the step which is carried out jointly by two spatially divided agents, it is impossible to bypass the concept of parallelism.

The models of parallelism known in the literature could be roughly divided into two classes: 1) the models, in which concurrent execution of two processes is described by interleaving of (atomic) events of those processes; 2) models in which causal dependencies between events are set explicitly. Interleaving models are focused on systems with events considered as instantaneous and indivisible. In this case the act of interaction is a complete event which describes participation of all processes cooperating in this act [2]. This act as the step of a dialogue is carried out by two spatially divided agents and represents the event which should have duration and structure.

There is popular opinion the concept of an event having duration is reduced to the concept of an instantaneous event. The following formulation of this assumption is taken from Hoare [3, p. 24]; “The actual occurrence of each event in the life of an object should be regarded as an instantaneous or an atomic action without duration. Extended or time-consuming actions should be represented by a pair of events the first denoting its start and the second denoting its finish.”

Now it is known that this opinion is erroneous, and often splitting, i. e. the use of pairs of instantaneous events to model events having duration, is unnatural. Mutual irreducibility of the concept of an event having duration and the concept of instantaneous event is proved formally and constructively [4]. The formal proof is based on the incomparability of these formalisms describing event systems [5]. Systems of the durational events are described by the causal relation (branching-time temporal logic), systems of instantaneous events – by relation of consequence and parallelism (linear-time temporal logic).

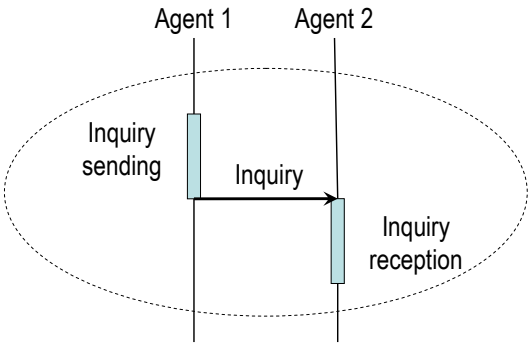


Fig. 1. Structure of the first model of a step of dialogue

The problem under discussion is how processes and events can be assembled together into a system in which the components interact with each other and with their external environment. The elementary structure of the dialogue step is a pair of durational events which constitute the step densely without a time interval between. First event of the pair can be interpreted as “pronouncing” of the message by one of the agents; the second event can be interpreted as “perception” of this message by the other participant of the dialogue. The basic feature of this structure is the assumption of density of the event composition and that members constituting the event belong to behaviours of different agents (fig. 1). Absence of a time interval between of pair of durational events designates that the event of synchronization of corresponding processes is instantaneous.

The first model of an interaction event

Ignoring the functionality of agents, it is possible to consider synchronization of their behaviour as the only goal of interaction. Thus, the dialogue step is the composition of three events, two events are durational ones, and the other is instantaneous event. However the events constituting interaction still remain occurring simultaneously in different processes.

On the one hand, synchronization of agent behaviours occur during the rare moments, in the rest of the time communicating agents behave independently from each other. On the other hand, processes should interchange information about current states to ensure synchronization. Formally it can be reached by splitting of all events constituting agent behaviour on internal and external ones. Only external events of the agent behaviour can be "visible" to the other agents. In this case the specification of the agent behaviour is the cause-effect relation on a set of possible events. In particular, this relation describes the reasons of occurrence of external and internal events.

Let a composition of a durational internal event E and instantaneous external event y is an operation. A composition $y \to E$ of durational and instantaneous events is called as a waiting operation that waits the external event y, and a composition $E \to y$ is called as an acting operation which effect is the realization of external event y. It is necessary to note, that the event sequence in both operations is the same: the first one is durational event, the second is instantaneous event. The mentioned compositions allow considering dependences between events in a composition as cause and effect because from physical reasons, event-consequence occurs behind event-reason without overlapping on time. Waiting operation is a durational event and the reason of its termination is the occurrence of y. Acting operation is a durational event too and it is the reason of occurrence of y. The symbol "$\to$" can be interpreted as cause and effect dependence between events.

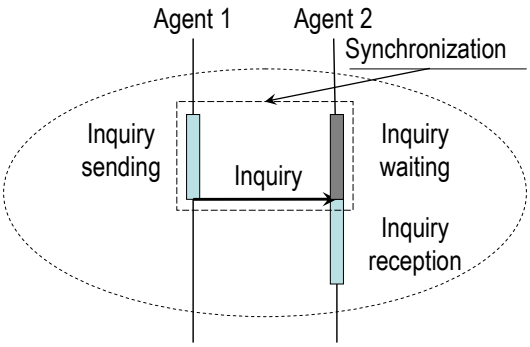


Fig. 2. Synchronization of behaviour of agents

Such treatment of waiting and acting operations is a basis of the formal semantics [6, 7] of PRALU language [8] in which conjunctions of Boolean variables describe external events of operations. PRALU language in this interpretation represents the synthetic temporal logic uniting linear and branching time temporal logics supplied by the assumption of density of time [9]. Temporal formulas of this logic are interpreted as the statements concerning event sequences of two sorts: instantaneous and durational.

The composition of events considered above allows describing independent behaviours of agents. Parallel execution of waiting operation $y \to E$ by one agent and acting operation $E \to y$ by the other one results in synchronization of behaviour of agents during the moment of occurrence of instantaneous event y (fig. 2). By the definition the effect of a waiting operation is its termination at a moment of occurrence of instantaneous external event y.

The line of "life" of the agent consists of pairs waiting and acting operations. The boundary between waiting and acting operations serves as the synchronization event. Here the action consists from "perception" of the accepted message and "pronouncing" new one. Obviously, the occurrence of synchronization depends on duration of acting operations.

The second model of an interaction event

A dialogue step can be considered also as the other composition of some three events. One of them is durational, and the others are instantaneous (fig. 3). In this model of a dialogue step the interaction itself is a durational event. This event, having duration, should have a physical basis. Without loss of a generality it is possible to consider that event of interaction occurs in an environment of agents. In this model interaction event becomes not distributed, but a local one in the external environment.

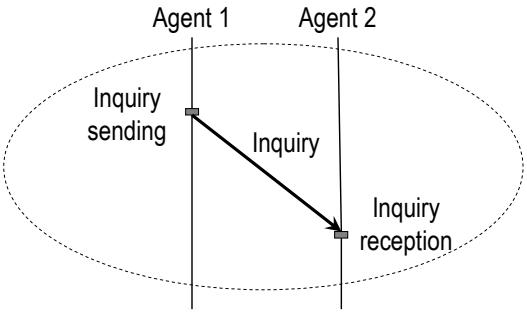


Fig. 3. Structure of the second model of an interaction event

Concept of an environment

From analysis of physical realizations of distributed systems it follows that the synchronization requires a special organization of a system of cooperating agents. Two basic types of the system organization aimed achieving synchronization are known: synchronous and asynchronous systems. The standard definition of the distinction of these types of systems declares that synchronous systems have the same shared “clock”, and in asynchronous systems each agent has its own independent clock. It is obvious, that the shared “clock” belong to an external environment of all agents of the synchronous system.

In traditional interaction theories CCS [3] or CSP [10] the concept of an environment is used implicitly, hence it is not formalized. CCS and CSP rely on the concept of an environment having the following distinguishing features. The environment is considered simply as the other agent. In other words, the environment for the given agent includes all other agents of the system that operate in parallel with this agent. In this case an agent and an environment are objects of the same nature. It is obvious, that this assumption of properties of an environment is not good enough from the point of view of specifying an agent interaction directed to achievement of synchronization. Such an approach is justified only by the following reasoning. It is considered that the concept of an environment concerns with the system realization and it is not represented at the level of agent behaviour.

Our purpose is to offer a formal model of interaction which is not concerned with a system realization. We consider an environment is essentially distinct from agents. The basis of this approach is that the interaction is considered as the communication act consisting of sending and receiving of messages. This formalization of interaction originates from Shannon’s paper about the theory of communication [11] in which interaction is considered as a way to transfer the message from a sender to a receiver through a medium, also called as transfer environment. Physical realization of an environment can be a computer program, a device or a physical environment.

Obviously, synchronization of agent behaviours is impossible without fixing data which are transferred by an environment during agent interaction [12]. Thus, environment serves as a model of transport system to deliver messages. In other words the environment can be considered as the memory that is shared with all agents. This memory is known as a global state of multi-agent system. In its most simple form, the communication can be based on the fixed set of differing signals. In the case of binary signals the representation of a global state is a set of the Boolean variables which values are possible signals. In the case of structural signals an agent environment usually refers as message passing system. The concept of an environment is closely concerned with a notion of autonomy of agents. Autonomy of agents has its focus on freely choosing between actions and on acting independently. Autonomy means also that the agents receive all information only through an environment.

System, in which the behaviour of an environment is deterministic, refers to closed system. In the case of a closed system it is supposed, that the reasons of all events are inside of the system and its behaviour is completely self controlled. If the behaviour of an environment is nondeterministic, the system refers to an opened system. Unlike the memory considered in the theory of finite state machines, the behaviour of the memory of an environment of the open multi-agent system can depend on uncontrollable conditions.

The specification of interaction in the form of description of a message passing system does the description of autonomous behaviour of the separate agent not closed because this description is not enough for understanding of the complete behaviour of the agent. Obviously, most important property of the message passing system is restriction on length of durational events, imposed by this system.

Time as a logic concept

One of the possible semantics of ts and their interaction is that of differential equations. The agents represent relations between continuous-time functions, and the interactions are the continuous-time functions by their nature. In the systems of differential equations a continuous time variable t is used. The task of an execution environment is to find a solution of the system of differential equations, i.e., a set of continuous-time functions that satisfy all the relations. This solution is seeked by “integrating” or “numerically solving” differential equations. The problem with this approach is that it is in general case a hard mathematical problem requiring much effort in searching a solution.

There are several meanings in which computation of “solving differential equations” may be understood. If by “solving” we mean manipulating the mathematical expressions of the differential equations to get a mathematical formula for the closed-form solution, then this (to the extent it is algorithmic) is computation of the familiar

discrete, symbol-manipulating variety. Alternately, we may take “solving” in the extended sense, in which a multy-agent system is “solving” a system of differential equations when the multy-agent system behavior can be described (perhaps approximately) by those differential equations. Here, the description of the multy-agent system by differential equations has nothing to do with whether it is can be computated or not. Multy-agent systems may be described by differential equations, but it may be not computable.

Differential equations can be discretized to get difference equations. In this case, a global clock defines the discrete points at which signals have values (at the ticks). Difference equations are considerably easier to implement in software than differential equations. Their key weaknesses are the global synchronization implied by the clock, and the awkwardness of specifying irregularly timed events and control logic.

In the previous part of the paper time was considered as the logic concept expressed by relations between events through their sequence and order. Time is discrete, because there is an observable time quantization by the events that is fixed in behaviour of an environment.

Time cannot be measured, if we do not impose some restrictions on the duration of events in all components of a multi-agent system. Time is measured if each event in a history of the system behaviour is accompanied with a number that expresses either duration of the event or specifying the moment of time when it occurs. Synchronization of behaviour of agents means that time is measured. But measured time model is not difference equations model.

Measured time can be realized, if we assume, that the duration of all simultaneously executed acting operations in a multi-agent system is identical. It is natural to accept this duration as the unit of time. In this case in the closed systems the duration of waiting operations is expressed by an integer $i \geq 1$. The assumption that duration of all simultaneously executed acting operations is identical holds in synchronous systems (global synchronization implied by the clock). Obviously, this assumption specifies a pairs of interacting waiting and acting operations by the counter number of the appropriate step of time. Synchronous system keeps the assumption that the time is discrete and measured.

Other assumption that allows realizing measured time is that any operation, carried out in parallel to itself, is illegal. In this case realization of any operation in a history of the agent functioning can be accompanied with some counter number of this realization. The formal proof of this statement is in [7]. The function which calculates a counter number of the operation realization (from the start of the system) when this operation starts can be used for measurement of time. Interaction occurs only in pairs of waiting and acting operations which have the same counter number. This is known as a rendezvous condition.

Rendezvous models is a part of Hoare's communicating sequential processes (CSP) [3] and Milner's calculus of communicating systems (CCS) [10]. In these frameworks, rendezvous is atomic, instantaneous action of communication. If two processes are to communicate, and one of them reaches first the point at which it is ready to communicate, then it stalls until the other process will be ready to communicate. Here “atomic” (rendezvous) means that the two processes are simultaneously involved in the exchange, and that the exchange is initiated and completed in a single uninterruptible step. Rendezvous in our framework is an event with duration and structure. A key weakness of rendezvous of CSP and CCS is that maintaining determinacy can be difficult. The proposed framework could break down the problem of specification determinacy.

An asynchronous system keeps the assumption that time is discrete and measured, but rejects the assumption that duration of all simultaneously executed acting operations is identical. The last principle of measuring time differs from that for synchronous system.

Conclusion

The independent behaviour of agents in the majority of models of multi-agent systems is described by means of formalisms of high level abstractness, but the communication is specified by the concepts close to realization. The difference of levels of the description does not allow simulating communications between agents at the level in which their independent autonomous behaviour is described. This problem arises because of absence of agent models that unify all aspects of local behaviour and the communications.

In the paper we suggest to describe the synchronization conditions by specification of event properties which have been not concerned with the realization of these events. Our approach allows specifying both the independent behaviour and the communication at a level of logic. It is shown, that the collective behaviour of

agents can be described by the synthetic temporal logic that unites the linear and branching time temporal logics. Such synthetic logic is one of interpretations of the existing PRALU language.

The transition from the analysis to design in development of the software is always based on mapping or transformation of conceptual models. The use of models during the analysis is inevitable. Analysis models target to describe system of the real world as mapping into some problem area. The concepts used in analysis model concern directly with concepts of system of the real world. On the other hand, the models used for design use an additional level of abstraction and pursue other purpose. Models for designing describe such concepts of the software, as objects, structures and processes which only are indirectly connected with concepts of problem area. The purpose of model for designing consists in masking details of realization and to create a formal basis for its subsequent transformation in the program.

Our contribution in this paper is both a new behavior model of agent interaction in multiaget systems, and the interactinon event abstraction which seems to be a good abstraction for distributed and parallel programming. Distributed systems design is unnecessarily complex because our current conceptual models do not provide the right kinds of abstractions. By adding appropriate abstraction to our models, we can also reduce the conceptual distance between analysis and design.

Acknowledgements

The research was partially supported by the Fond of Fundamental Researches of Belarus (Project **F07-125**).

Bibliography

1. Subrahmanian V.S., Bonatti P., Dix J. et al. “Heterogeneous Agent Systems”, MIT Press, 2000.
2. Brookes S.D., Hoare C. A.R., and Roscoe A.D. “A Theory of Communicating Sequential Processes”, Journal of the ACM, no 31(3), pp. 560–599, 1984.
3. C.A.R. Hoare “Communicating Sequential Processes”, Prentice Hall International Series in Computer Science, 1985.
4. Van Glabbeek R., Vaandrager F. “The Difference between Splitting in n and n+1”, Report CS-R9553, Centre for Mathematics and Computer Science, Amsterdam 1995; Abstract in: Proceedings 3rd Workshop on Concurrency and Compositionality, Goslar, March 5-8, 1991 (E. Best & G. Rozenberg, eds.), GMD-Studien Nr. 191, Sankt Augustin, Germany 1991.
5. Cheremisinov D.I. “The Real Difference between Linear and Branching Temporal Logics”, Workshop on Discrete-Event System Design DESDes ’ 04, University of Zielona Gora Press, Poland, 2004, pp. 103–108, 2004.
6. Cheremisinov D.I. “Formal description of behaviour of the distributed systems”, Minsk: Belarus, 1991, preprints no 38. – 44 p. (in Russian).
7. Cheremisinov D.I. “The morphisms of reactive system formal languages”, Informatics, no 1 (5), pp. 76–88, 2005 (in Russian).
8. A.D. Zakrevskij, “Parallel algorithms for logical control”, Minsk, Institute of Engineering Cybernetics of NAS of Belarus, 202 p., 1999 (in Russian).
9. Cheremisinov D.I. “About interpretation of temporal logic at symbolical verification”, Informatics, no 1, pp. 131–138, 2004 (in Russian).
10. Milner R. “A calculus of communication systems”, LNCS 92, Springer Verlag, 1980.
11. Shannon C.E. “A mathematical theory of communication”, Bell System Technical Journal, vol. 27, pp. 379–423 and pp. 623–656, 1948.
12. Odell J., Parunak H. V. D., Fleischer M., Brueckner S. “Modeling Agents and their Environment”, Proceedings of the Third International Workshop on AgentOriented Software Engineering, Lecture Notes in Computer Science, Springer Verlag (Berlin, D), vol. 2585, pp. 16–31, 2003.

Authors' Information

Dmitry Cheremisinov, Liudmila Cheremisinova – The United Institute of Informatics Problems of National Academy of Sciences of Belarus, Surganov str., 6, Minsk, 220012, Belarus, Tel.: (10-375-17) 284-20-76, e-mail: cher@newman.bas-net.by, cld@newman.bas-net.by

INTELLIGENT CAR PARKING LOCATOR SERVICE

Ivan Ganchev, Máirtín O'Droma, Damien Meere

Abstract: This paper presents an InfoStation-based multi-agent system facilitating a Car Parking Locator service provision within a University Campus. The system network architecture is outlined, illustrating its functioning during the service provision. A detailed description of the Car Parking Locator service is given and the system entities' interaction is described. System implementation approaches are also considered.

Keywords: InfoStations, intelligent agents, multi-agent system, JADE, LEAP.

ACM Classification Keywords: H.3.4 Systems and Software, C.2.1 Network Architecture and Design.

I. Introduction

The InfoStations paradigm is an infrastructural system concept supporting "many-time, many-where" (Frenkiel and Imielinski 1996) wireless communications services. The InfoStation-based system outlined in this paper is established and operates across a University Campus area for the purpose of enhancing the mobile services experience. It allows mobile devices (mobile phones, laptops, personal digital assistants–PDAs) to communicate to each other and to a number of servers through geographically intermittent high-speed connections. In this paper, we detail the underlying network architecture and show how the different components within the architecture collaborate to facilitate one particular service, namely the Parking Locator service. This service allows registered users to locate available parking spaces throughout the campus.

The rest of the paper is organized as follows. Section II presents the InfoStation-based network architecture, illustrating how the architecture functions during service provision. Section III illustrates the Parking Locator service provision outlining sample interactions between system entities. Section IV outlines some implementation issues, and finally Section V concludes the paper.

II. InfoStation-based Network Architecture

The following InfoStation-based network architecture (Ganchev, O'Droma et al. 2003; Ganchev, Stojanov et al. 2006; Ganchev, Stojanov et al. 2006) provides access to a number of very useful services, for users equipped with mobile wireless devices, via a set of InfoStations deployed in key points around a University Campus. The 3-tier network architecture consists of the following basic building entities as depicted in Figures 1 and 2: user mobile devices, InfoStations and an InfoStation Center.

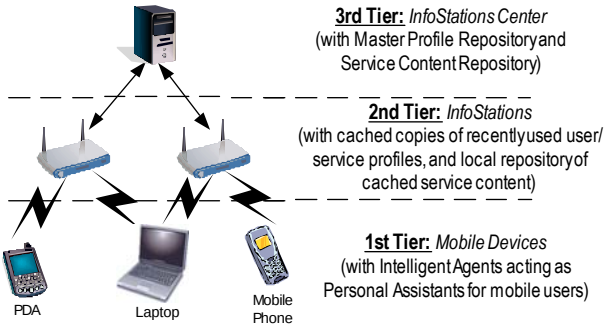


Figure 1. The 3-tier InfoStation-based network architecture

The users request services (through their mobile devices) from the nearest InfoStation via available Bluetooth (IEEE 802.15 WPAN), WiFi (IEEE 802.11 WLAN), or WiMAX (IEEE 802.16) connections. The InfoStation-based system is organized in such a way that if the InfoStation cannot fully satisfy the user request, the request is forwarded to the InfoStation Center, which decides on the most appropriate, quickest and cheapest way of delivering the service to each user according to his/her current individual location and mobile device's capabilities

(specified in the user profile). Figure 2 illustrates some of the main components within each entity of the architecture.

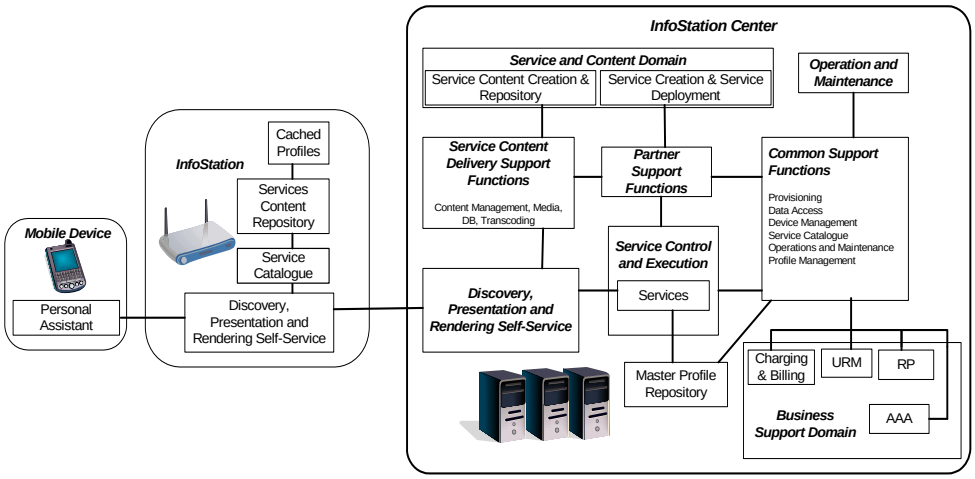


Figure 2: The InfoStation System Architecture

The *InfoStation Center* is concerned with the creation of service content and service creation, deployment, operation, maintenance, control and execution. In addition there are some common support functions that each service requires when initially created, for example device management, profile management, service catalogue etc. The InfoStation Center also houses a repository of all (up-to-date) master profiles relating to both users and services alike. Any changes made by the individual user to his/her own user profile and/or user service profile are forwarded on from the user mobile device, through an InfoStation to the InfoStation Center, where the repository is updated. (Each InfoStation keeps cached copies of all recently used, or updated by users, profiles.) The InfoStation Center also houses the Business Support Domain with a number of components relating to the charging and billing of users, User Relationship Management (URM), Resource Planning (RP) and indeed user Authentication, Authorization and Accounting (AAA).

When a mobile user enters within the range of an *InfoStation*, the Personal Assistant, installed in the user mobile device, and the InfoStation mutually discover each other. This process is facilitated through the Discovery, Presentation and Rendering Self-Service module within the InfoStation. The Personal Assistant sends a request to the InfoStation for user's Authorization, Authentication and Accounting (AAA). This request also includes a description of the mobile device currently being used by the user (or just the device's make and model) as well as any updates of user profile and user service profile (Figure 3). In particular with the Intelligent Parking Locator service, this process may occur a number of times as the user will, more often then not, pass through a number of InfoStation coverage areas with his/her vehicle.

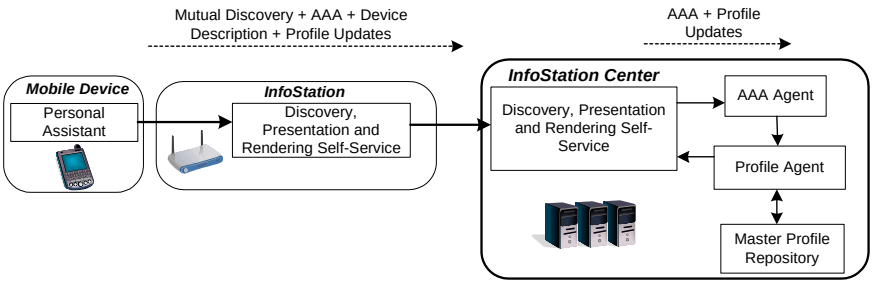


Figure 3: Step 1- Initial AAA and profile updates.

The InfoStation forwards this AAA request to the InfoStation Center along with the profile updates (Figure 3). If the user is successfully authenticated and authorized to utilize the services by the AAA module within the InfoStation Center, a new account record is created for the user. The user profile is analyzed by the InfoStation Center for current user preferences (e.g. applicable services) and device capabilities (utilizing the Composite

Capabilities/Preference Profile – User Agent Profile, UAProf). Then the InfoStation Center makes a service offer to the user in a form of a compiled list of applicable services from the Service Catalogue (Figure 4).

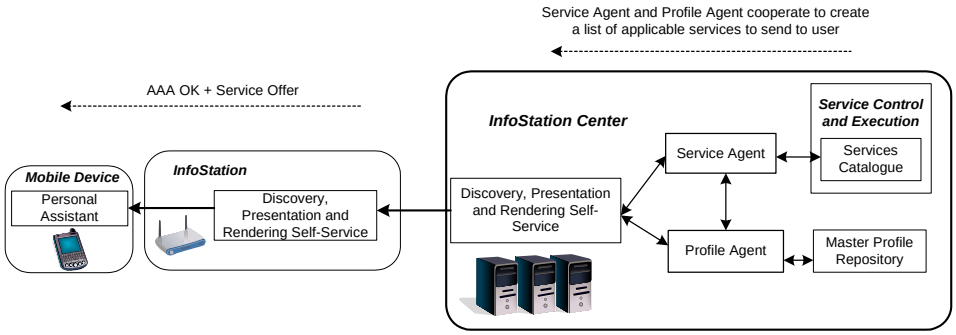


Figure 4: Step 2- Service Offer

This service offer is sent towards the Personal Assistant along with the AAA acknowledgment. The Personal Assistant displays the offer to the user who makes a choice and selects (makes a request for) the service s/he wishes to use. When the Personal Assistant forwards the user service request to the InfoStation, the latter checks its cache for the most up-to-date version of the requested service content (e.g. campus news bulletin). If the InfoStation is able to satisfy fully the user service request, it does so (Figure 5). Otherwise the InfoStation forwards this request to the InfoStation Center, which is better equipped to deal with it.

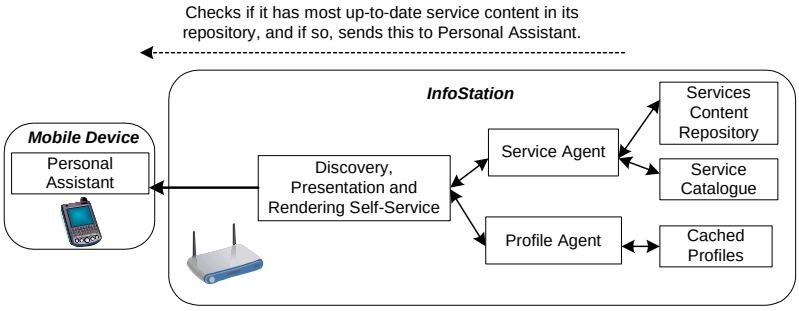


Figure 5: Step 3- Service request satisfied by InfoStation

On the user *mobile device*, the Personal Assistant (*agent*) facilitates the service utilization by the user. This is down to an agent-oriented approach to the implementation of the system. The service migrates onto the users mobile device, allowing the user unhindered access to the service even when out of range of the InfoStation. The Personal Assistant may make a service request while within the range of an InfoStation, then may pass out of the coverage area but will continue to work autonomously, adopting the functionality of the service until the user has completed his/her task. Once the mobile device comes within range of another InfoStation, the Personal Assistant updates and synchronizes the user service profile to reflect any work completed, or any new service requests made by the user while out of range.

In the following section we describe the provision of the Intelligent Parking Locator service in more detail.

III. Intelligent Parking Locator Service

A multi-agent approach (Carabelea and Boissier 2003; Ganchev, Stojanov et al. 2004; Stojanov, Ganchev et al. 2005; Adaçal and Bener 2006; Ganchev, Stojanov et al. 2006) is adopted as most suitable approach to structuring our system. In order to facilitate flexible and adaptable service provision, intelligent agents, residing within each of the three tiers of the system architecture must interact so as to satisfy, in the 'best' possible way, any user requests they might encounter. The following description outlines the entity interactions that take place during the Intelligent Parking Locator service provision. This service allows registered mobile users to gain access to information regarding available parking spaces on the University Campus and reserve a space that best suits them when approaching/entering the campus. However, visitors may also gain access to this service

through prior temporary registration in the system for the duration of their stay. On accessing the service, these visitors would be directed to a visitor's car park.

In the delivery of this service, the content must be adapted and customized according to the capabilities of the user device and the user preferences. For example if the user has access to a resource-rich mobile device (e.g. a laptop or indeed a PDA), s/he may gain access to a graphical representation of the campus, which would greatly assist the user in finding the required parking space. If however the user only has access to a device with limited capabilities (e.g. a mobile phone), then the details of the available parking spaces would be specified in a simple format which 'best' suits the device (e.g. SMS/MMS). This trimming (adaptation) of the services is one way to address the shortcomings of some mobile devices, while still delivering the service.

We use the "Composite Capabilities / Preference Profile" (CC/PP) as the uniform format for the implementation of the user profiles. The master profile repository in the InfoStation Center contains descriptions of all registered user devices, i.e. their capabilities and technical characteristics. During the initial AAA request, the user's Personal Assistant sends as parameters the make and the model of the user device. An agent working on the InfoStation (or the InfoStation Center) reads the corresponding device's description from the repository and according to this, selects and forwards the best format of the service content. However a problem arises when a user uses a non-registered device as s/he might receive the service content in unsuitable format. Thus the user needs first to register any new mobile device s/he wants to use within the system. In this case, during the initial AAA request the Personal Assistant sends a full description of the user device's capabilities towards the InfoStation Center.

Figure 6, depicts a sample interaction between entities involved in the Intelligent Parking Locator service provision. As the user enters the campus area in a vehicle, s/he enters the coverage area of an InfoStation, positioned at the entrance to the campus. The Personal Assistant, installed in the user mobile device, and the InfoStation mutually discover each other. The Personal Assistant sends a request to the InfoStation for user's Authorization, Authentication and Accounting (AAA). During this initial AAA request, the user's Personal Assistant sends also the make and the model parameters of the user device, and any updates of user profile and user service profile. The InfoStation registers the user in its local Virtual Address Book and updates the profile, before forwarding the user request onto the InfoStation Center along with profile updates. In the case of successful AAA, the Profile Agent within the InfoStation Center (updates and) analyses the user profile stored in its Master Profile Repository. The Service Agent, in collaboration with the Profile Agent, creates a list of services applicable to the user and makes a service offer to the user.

However the user may specify in his/her profile that a request for the Parking Locator service be sent automatically after the successful AAA (and profile update) procedure. Or alternatively, if the user makes regular use of the service, the Personal Assistant could proactively anticipate the users request, i.e. once this service becomes available, the Personal Assistant automatically requests the location of parking for the user's vehicle. The InfoStation forwards on the user request to the InfoStation Center. Sensor networks within the car parks constantly update the InfoStation Center as to the availability of spaces. Different time periods of the day require more regular updates, especially from morning to mid-afternoon, as the user would require the information be as up-to-date as possible. However the updates can occur at much larger intervals during the evening and weekends when many more spaces would be available. In the case of Staff user's, the InfoStation Center discerns the location of the user's office from the user profile, and as such compiles a sorted list of available parking spaces according to their proximity to the user's office (desired destination). For Students and Visitors, the InfoStation Center locates parking spaces within the visitor and student car parks. In these cases, the InfoStation Center will also consult the user profiles to order the parking spaces according to criterion such as convenience to final destination and in particular for students, the cost associated with each parking space. The InfoStation Center then determines the approximate position of the user based on the location of the InfoStation from which the request was received. The InfoStation Center then discerns the best directions from user's current location to each of the available spaces.The Service Content agent and Profile Agent cooperate to adapt the content to the format that best suits the current user device capabilities and user preferences (i.e. graphical representation, audio description, text). Once the content is prepared for transfer, the InfoStation Center discerns the most suitable InfoStations to forward the data on to. As the user is most probably accessing the service whilst in transit, there is a good chance the user will pass through a number of InfoStation coverage areas. The InfoStation Center makes allowances for this and forwards the information on to a number of InfoStations in the

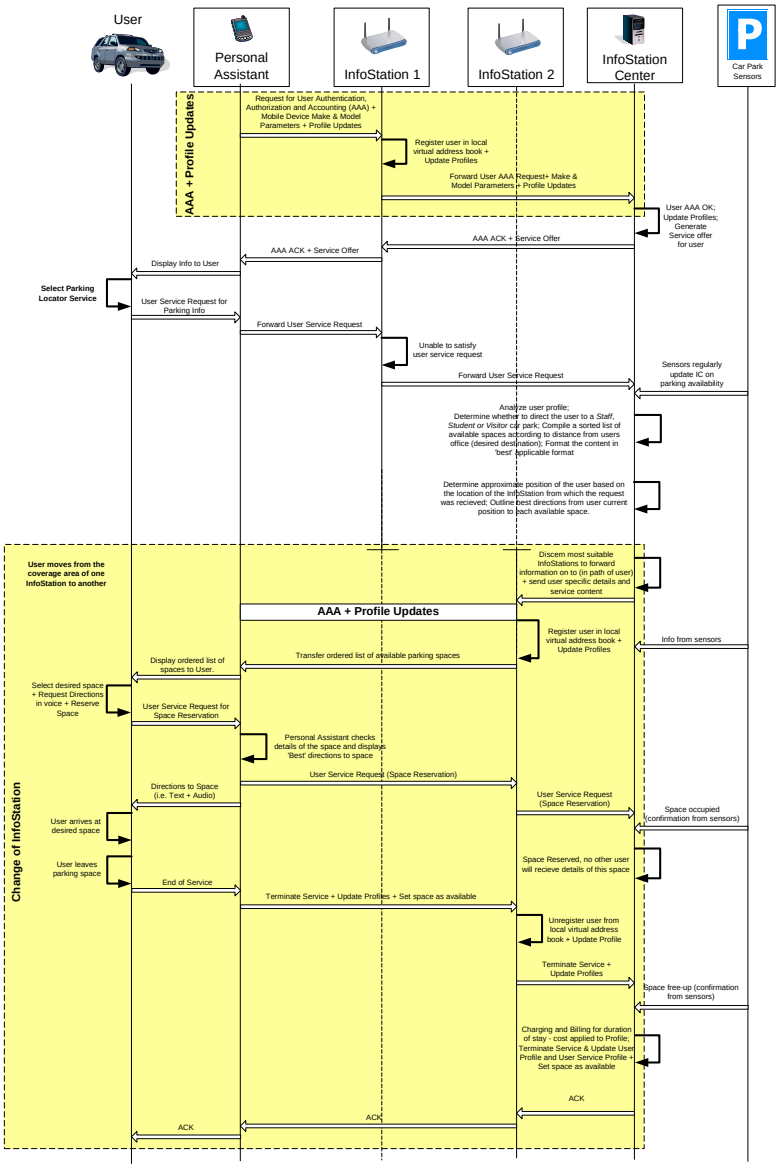


Figure 6: Intelligent Parking Locator Service: Entity Interactions

path of the user (Borràs and Yates 1999; Yuen, Yates et al. 2003), along with specific user details.

As the user moves from the coverage area of one InfoStation to another, AAA and profile update procedures are executed first. The approached InfoStation will have already received information about the user from the InfoStation Center, along with requested service content. As such the InfoStation can account for the user and immediately forward on the requisite content. This reduces the time taken for the InfoStation to provide the service content. This process may happen with a number of InfoStations as the user drives through the campus. As the user leaves the coverage area of an InfoStation, the user service profile is updated, specifying how much of the service content was transferred (if the transaction was not completed). This information is circulated around the InfoStation network, so as to ensure the user's Personal Assistant does not receive the same information a number of times.

Once the user receives the ordered list of parking spaces, s/he chooses a particular parking space, reserves it and request directions to that space. Once a space is chosen, the Personal Assistant examines the details of the space and displays precise directions. An audio explanation accompanying the text description would be best suited to this service, as it would provide the least distraction, allowing the user to concentrate on driving.

The Personal Assistant also forwards on a parking space reservation request to the InfoStation Center. Once the space has been reserved (and it's occupancy confirmed by the sensor network), no other users will be supplied

with details of that space. The InfoStation Center monitors the duration the user occupies the parking space for charging and billing purposes.

When the user leaves the parking space, the sensor network confirms this to the InfoStation Center. The Charging and Billing Module within the InfoStation Center accounts for the duration of the user's stay. A corresponding charge related to parking in that car park, is charged to the user account. Once the user/ service profiles have been updated, the service is terminated.

Another issue to be taken into account is that of Staff specific parking spaces. Certain Staff members (e.g. University President, Vice-Presidents, etc) will be allocated their own specific private parking space. If perhaps an unauthorized user enters the space, the sensors in the car park will alert the InfoStation Center. If this unauthorized user is registered, the InfoStation Center will, if possible, forward on a notification to the user to vacate the space and perhaps provide the location of and available space nearby. If the unauthorized user happens to be unregistered and un-contactable, campus security may be notified.

IV. Implementation

The system is implemented in an agent-oriented manner utilizing the Java Agent DEvelopment (JADE) (JADE; Bellifemine, Poggi et al. 2001; Anghel and Salomie 2003; Bellifemine, Caire et al. 2003; Bellifemine, Caire et al. 2005; Bellifemine, Caire et al. 2006) framework. This allows for the flexible development of multi-agent systems and applications for management of network resources in compliance with the FIPA specifications. The JADE architecture is completely modular and as such, by utilizing specific modules, can be configured to adapt to the requirements of a number of different deployment environments. Within our JADE implementation, one of the most useful modules is the Lightweight Extensible Agent Platform (or LEAP) (Moreno, Valls et al. 2003; Caire and Pieri 2006) Module. This module, or add-on, replaces some parts of the JADE kernel, providing a modified run-time environment, which facilitates the implementation of agents on mobile devices with limited resources. Another very useful aspect of JADE-LEAP is its ability to support split-containers (split run-time environments) on resource-thin devices. The container can be split into two separate sections, a FrontEnd (running on the mobile device itself), and the BackEnd (running from a fixed network entity - a mediator) as illustrated in Figure 7.

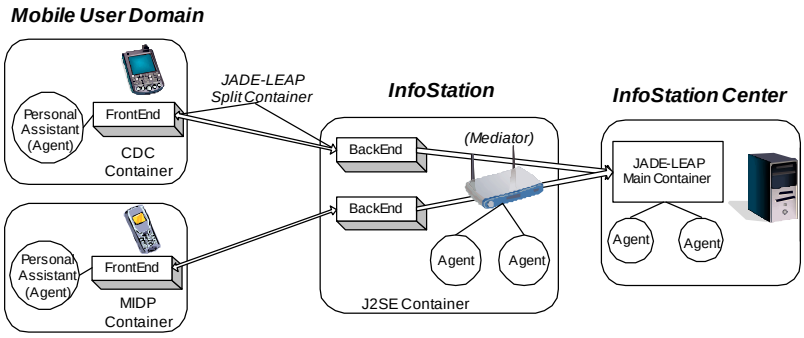


Figure 7. JADE-LEAP split-container execution

This mediator is charged with instantiating and maintaining the BackEnds. In our system, the InfoStations deployed throughout the campus take on these mediator roles. Each FrontEnd is connected to each BackEnd through a bi-directional connection. The splitting of the container into two separate, yet connected, entities is particularly useful in the realm of resource-constrained devices, as the FrontEnd of the container is far more lightweight in terms of the required memory and processing power then the entire container. Due to the geographically intermittent nature of the InfoStation connection, the FrontEnd and the BackEnd may undergo a loss of connection, however the Front-End can detect this and re-establish the connection as soon as possible. Any messages not transmitted due to this temporary disconnection can be buffered and delivered when the connection is re-established. This store-and-forward mechanism (implemented in both the FrontEnd and the BackEnd) is especially important to the efficient facilitation of the Parking Locator Service, where the user will pass in and out of coverage range of a number of different InfoStations, and as such data will have to be buffered and transmitted after a period of time by another InfoStation.

Figure 8. Screenshots of service execution on devices with varying capabilities.

The splitting of the container has no bearing on us, as the same functionality and set of APIs are available to an agent, whether it is contained within a full container or the FrontEnd of a split container. The JADE framework also serves to shield us from the complexity of the distributed environment, allowing the concentration of our efforts on developing the application logic, rather then worrying about middleware issues such as discovery and communication of entities within the system.

The following are two sample screen shots of how this service will appear on two different mobile devices with different capabilities. The screen shot on the left represents a device with the capabilities to show complex graphical information. In this case the device shows a map of the campus and graphically illustrates the path for the user's vehicle to follow, in order to reach the reserved destination parking space.

The device on the right illustrates how a device with more limited capabilities may convey the same information to the user. That particular device's profile will specify its capability to only handle text information during its communications with the InfoStations. As such the InfoStation will provide only the requisite text information.

V. Conclusion

The effectuation of the InfoStation-based Parking Locator service in a University Campus area has been outlined in this paper. The underlying network architecture has been described along with an illustration of how each of the different components within the architecture collaborates to facilitate mobile services. The Parking Locator service has been considered as an example. This service allows registered users to locate available parking spaces throughout the campus and reserve a space that best suits them when approaching/entering the campus area. Details of how service content is tailored to specific devices and how the duration of the user's stay affects the charging and billing for utilization of the service have been outlined.

The multi-agent structure, implemented by means of the Java Agent DEvelopment (JADE) software framework utilizing its Lightweight Extensible Agent Platform (LEAP) module in particular, has been discussed in detail due to its suitability to the proposed system. The benefits of this implementation have been also outlined in detail.

Acknowledgments

Dr. Ivan Ganchev, Dr. Máirtín O'Droma, and Damien Meere wish to acknowledge the financial support of the Ireland's HEA Strategic Initiatives Funding Program 'Technology in Education' for the development of the system.

Bibliography

Adaçal, M. and A. Bener (2006). "Mobile Web Services: A New Agent-Based Framework." IEEE Internet Computing Vol. 10(no. 3): pp. 58-65.

Anghel, C. and I. Salomie (2003). JADE Based solutions for knowledge assessment in eLearning Environments, TILAB & University of Limerick.

Bellifemine, F., G. Caire, et al. (2003). "JADE: A White Paper." exp, Telecom Italia Lab Volume 3(No. 3,).

Bellifemine, F., G. Caire, et al. (2005). JADE Programmers Guide, TILab.

Bellifemine, F., G. Caire, et al. (2006). Jade Administrator's Guide, TILab.

Bellifemine, F., A. Poggi, et al. (2001). JADE: A FIPA2000 Compliant Agent Development Environment. AGENTS '01, Montreal, Quebec, Canada.

Borràs, J. and R. D. Yates (1999). Highway InfoStations. WPMC'99, Amsterdam.

Caire, G. and F. Pieri (2006). LEAP User Guide, TILab.

Carabelea, C. and O. Boissier (2003). Multi-agent platforms on smart devices: Dream or reality? Smart Objects Conference (SOC03), Grenoble, France.

Frenkiel, R. H. and T. Imielinski (1996). "Infostations: The joy of `many-time, many-where' communications." WINLAB Technical Report(WINLAB-TR-119).

Ganchev, I., M. O'Droma, et al. (2003). A model for integration of electronic services into a distributed eLearning center. 14th EAEEIE International Conference, Gdansk, Poland.

Ganchev, I., S. Stojanov, et al. (2006). An InfoStation-Based Multi-Agent System for the Provision of Intelligent Mobile Services in a University Campus Area. IEEE-IS'06, London.

Ganchev, I., S. Stojanov, et al. (2006). An InfoStation-Based University Campus System for the Provision of mLearning Services. IEEE-ICALT '06, Kerkrade, The Netherlands.

Ganchev, I., S. Stojanov, et al. (2004). Enhancement of DeLC for the Provision of Intelligent Mobile Services. 2nd International IEEE Conference on Intelligent Systems (IS'2004), Varna, Bulgaria.

JADE Java Agent Development Framework Project - http://jade.cselt.it.

Moreno, A., A. Valls, et al. (2003). "Using JADE-LEAP to implement agents in mobile devices." TILAB "EXP in search of innovation", Italy.

Stojanov, S., I. Ganchev, et al. (2005). An Approach for the Development of Agent-Oriented Distributed eLearning Center. International Conference on Computer Systems and Technologies - CompSysTech, Varna, Bulgaria.

Yuen, W. H., R. D. Yates, et al. (2003). Effect of Node Mobility on Highway Mobile Infostation Networks. ACM MSWiM 2003, San Diego.

Authors' Information

Dr. Ivan Ganchev – Dip. Eng. (honours), PhD, IEEE (M.), IEEE ComSoc (M.), Lecturer and Deputy Director of the Telecommunications Research Centre, University of Limerick, Ireland. He has served on the TPC of many international conferences including IEEE VTC2007Spring, IEEE Globecom2006, IEEE ISWCS 2006 & 2007. Ivan.Ganchev@ul.ie

Dr. Máirtín S. O'Droma – B.E., PhD, C.Eng., FIEE, IEEE (SM), Senior Lecturer and Director of the Telecommunications Research Centre, University of Limerick, Ireland. He has served on the TPC of many international conferences including IEEE VTC2007Spring, IEEE ISWCS 2006 & 2007. Mairtin.ODroma@ul.ie

Damien Meere – Researcher in the Telecommunications Research Centre in the University of Limerick, Ireland. He is currently pursuing his MEng degree leading to transfer to PhD. Damien.Meere@ul.ie

TRAFFIC OFFERED BEHAVIOUR REGARDING TARGET QOS PARAMETERS IN NETWORK DIMENSIONING

Emiliya Saranova

Abstract: We consider a model of overall telecommunication network with virtual circuits switching, in stationary state, with Poisson input flow, repeated calls, limited number of homogeneous terminals and 8 types of losses. One of the main problems of network dimensioning/redimensioning is estimation of traffic offered in network because it reflects on finding of necessary number of circuit switching lines on the basis of the consideration of detailed users manners and target Quality of Service (QoS). In this paper we investigate the behaviour of the traffic offered in a network regarding QoS variables: "probability of blocked switching" and "probability of finding B-terminals busy". Numerical dependencies are shown graphically. A network dimensioning task (NDT) is formulated, solvability of the NDT and the necessary conditions for analytical solution are researched as well.

The received results make the network dimensioning/redimensioning, based on QoS requirements easily, due to clearer understanding of important variables behaviour.

The described approach is applicable directly for every (virtual) circuit switching telecommunication system e.g. GSM, PSTN, ISDN and BISDN. For packet - switching networks, at various layers, proposed approach may be used as a comparison basis and when they work in circuit switching mode (e.g. VoIP).

Keywords: Overall Network Traffic, Offered Traffic, Virtual Circuits Switching.

ACM Classification Keywords: C.2.1 Network Architecture and Design; C.2.3 Network Operations; C.4 Performance of Systems.

1. Introduction

We consider a model of telecommunication system with virtual circuits switching, in stationary state, with Poisson input flow, repeated calls, limited number of homogeneous terminals and losses due to abandoned and interrupted dialing, blocked and interrupted switching, not available intent terminal, blocked and abandoned ringing and abandoned conversation.

One of the main problems of network dimensioning/redimensioning is estimation of traffic offered in network because it reflects on finding of necessary number of circuit switching lines. There are many different factors that we need to take into account when analyzing traffic. QoS parameters are administratively specified in Service Level Agreement (SLA) between users and operators [1].

Based on the ITU definitions [2] (E.600- 4.1, 4.2, 2.8 and 2.11), as QoS parameters, we use the follow two parameters, dependable from the network macro-state (*Yab*): probability *Pbs* (blocked switching) due to lack of resources and probability *Pbr* of finding B-terminals busy. We denote the target value of blocked switching by *trg.Pbs*.

A network dimensioning task (NDT) is formulated, solvability of the NDT and the necessary conditions for analytical solution are researched as well [3].

In this paper we investigate the behaviour of the traffic offered (*ofr.Ys*) in a network regarding QoS variables *Pbs* and *Pbr*. Numerical dependencies are shown graphically.

The results are useful for finding the range of *ofr.Ys* variability in every concrete case and developing of the best suitable numerical method for finding of the necessary number of equivalent internal switching lines (*Ns*) in dimensioning and redimensioning tasks.

2. Conceptual model and analytical models

The conceptual model [4] of the telecommunication system includes the paths of the call attempts, generated from (and occupying) the A-terminals in the considered network. All assumptions made are described and the base general equations are explained in [4]. A system of equations based on the conceptual model and some dependencies [5] between parameters of the researched telecommunication system, is derived.

3. Network Dimensioning Task

3.1 Formulation of a network dimensioning task (NDT):

To determine the volume of telecommunication resources, based on previous experience in other telecommunication networks, that is enough for serving expected input flow of demands, with prescribed characteristics of QoS, is one of the main problems that often have to be solved by operators. It includes the following two tasks:

1. Finding the values of the designed parameters, describing the designed system state. For example, a system parameter, describing offered traffic intensity of the switching system (*dsn.ofr.Ys*), designed probability to find B terminal “busy” (*dsn.Pbr*), etc…

2. Dimensioned a network. It means to be found the number of internal switching lines, necessary to satisfy a level of QoS that has been administratively pre-determined, and for which the values of known parameters are measured and/or calculated (in the case of any operational network).

For solving of these tasks we have to determine the dependencies of designed offered traffic intensity with respect to QoS – parameters and users' behaviour characteristics in telecommunication system.

3.2 Parameters and aims in the Network Dimensioning Task (NDT):

Given parameters:

Administrative determined parameters: *trg.Pbs, Nab* = *adm.Nab*

Parameters with empirical values: $Fo,\ S_1,\ S_2,\ S_3,\ R_1,\ R_2,\ R_3,\ S_{1z},\ S_{2z},\ Tb$

Aim: To determine the number of equivalent internal switching lines *Ns*; and the values of following

Designed unknown parameters: *dsn.Pbr, dsn.ofr.Ys*

3.3 Analytical solution of the NDT:

In the denoted papers [3], [4] and [5] the follow facts, in different theorems, are proved and the conditions are researched:

1. *Yab* is the intensity of the terminal traffic [4] in the system and in NDT is derived as the function of dynamic parameters with empirical values $Fo,\ S_1,\ S_2,\ S_3,\ R_1,\ R_2,\ R_3,\ S_{1z},\ S_{2z}$

$$Yab = \frac{Fo(1-Pbr)\{S_1 - S_2 Pbr - (S_3 - S_2 Pbr)Pbs\}}{Fo(1+MPbr)\{S_1 - S_2 Pbr - (S_3 - S_2 Pbr)Pbs\} - Pbr\{1 - R_1 - R_2(1-Pbs)Pbr - R_3 Pbs\}} \quad (3.1)$$

,

when $Fo \neq 0$ and $Pbr \neq 0$.[5]

For calling rate of call demand exist in NDT an expression [3]

$$dem.Fa = Fo\{(Nab-1)(1+MPbr)+1+M\}. \quad (3.2)$$

2. If $Pbr \neq 0 \ and \ \ Pbs \neq \frac{S_1 - S_2 Pbr}{S_3 - S_2 Pbr}$ in the NDT, then for *rep.Fa* [3] and *ofr.Ys* [3], an analytical expression exist for its evaluation

$$rep.Fa = \frac{Yab\{R_1 + R_2(1-Pbs)\,Pbr + R_3\,Pbs\}}{S_1 - S_2(1-Pbs)Pbr - S_3\,Pbs} \quad (3.3)$$

$$ofr.Ys = \frac{Yab(1-Pad)(1-Pid)(S_1 z - S_2 z\,Pbr)}{S_1 - S_2(1-Pbs)Pbr - S_3\,Pbs}\ . \quad (3.4)$$

3. In NDT equation [3]

$$A\,Pbr^2 + B\,Pbr + C = 0, \text{ where} \quad (3.5)$$

$$A = R_2(1-trg.Pbs)(Nab-1)$$

$$B = (1-trg.Pbs)(R_2 + dem.Fa\,S_2) + (1 - R_1 + R_3 trg.Pbs)(Nab-1)$$

$$C = 1 - R_1 + R_3\,trg.Pbs - dem.Fa\,(S_1 - S_3\,trg.Pbs),$$

when $Fo \neq 0$ and $trg.Pbs \neq \frac{S_1 - S_2}{S_3 - S_2}$, has at least one solution $Pbr^* \in (0;1)$.

If the value of *Pbr* is determined thereby, we say that *Pbr* is determined on the base of the NDT and we denote it *dsn.Pbr*. Based on it (when we know *trg.Pbs* and *dsn.Pbr*), in NDT, we may evaluate design values of unknown parameters *dsn.Yab*, *dsn.ofr.Ys* e.t.c. For example, the expression for *ofr.Ys* is

$$dsn.ofr.Ys = \frac{(1+dsn.Pbr\,(Nab-1))(1-Pad)(1-Pid)\ (S_{1Z}-S_{2Z}\ dsn.Pbr)}{S_1+S_2(1-trg.Pbs)\,dsn.Pbr+S_3\,trg.Pbs}. \qquad (3.6)$$

4. In the NDT, only one solution of the equation

   Erl_b (Ns,dsn.ofr.Ys) =trg.Pbs regarding the number of switching lines *Ns* exists.

The expression *Erl_b (Ns,dsn.ofr.Ys)* is the famous Erlang B-formulae.

Trg.Pbs $\in$(0; 1] is in advance administratively determined target value of blocking probability, providing of GoS [3].

It is proved [3] that only one solution of *Ns* exists, fulfilling the equation (4.3.1) and corresponding to the determined administratively in advance value of the blocking probability *trg.Pbs* $\in$(0; 1].

The number of internal switching lines *Ns* and the values of *dsn.ofr.Ys* are calculated on the conditions of the theorems in [3], [4] and [5]. Algorithm and computer program for calculating the values of the NDT parameters are worked out.

4 Parameters dependency in NDT

4.1. Dependency – used definitions.

As mathematical approach are used partial derivatives of researched parameters [3] and following definitions:

Let $P(x_1, x_2, ..., x_n)$ is tuple (ordered set) consists of variables $x_1, x_2, ..., x_n$.

Tuple P_0 of empirical or evaluated parameters' values is $P_0(x_1^0, x_2^0, ..., x_n^0)$, where the parameters' values are $x_1^0, x_2^0, ..., x_n^0$.

Let parameter *A* depends on tuple *P* then $A(P) = f(x_1, x_2, ..., x_n)$ where $x_1 \in D_1$, $x_2 \in D_2 ..... x_n \in D_n$. The value of A in P_0 is $A(P_0) = f(x_1^0, x_2^0, ..., x_n^0)$.

Range of parameters' values of *A* according parameter x is amplitude of it $Max\,A(x_k^s) - Min\,A(x_k^p)$, where $Max\,A(x_k^s)$ and $Min\,A(x_k^p)$ are absolute maximum resp. minimum received in points $x_k^s \in D_k$ and $x_k^p \in D_k$.

$$Range(A \mid x_k) = Max\,A(x_k^s) - Min\,A(x_k^p)$$

4.2. Research of parameters dependency in NDT regarding GoS parameters probability of blocking due to *Pbs* and *Pbr*. Knowing functional dependencies from a parameter, we may estimate the parameters' importance and necessary accuracy of its measurement.

We consider *ofr.Ys* - dependency regarding *Pbs* and *Pbr* because *Ns* is direct dependent on it.

We denote the tuple of lost probabilities *L*= (*Pad, Pid, Pis, Pns, Par, Pac*).

Then *ofr.Ys*= *ofr.Ys* (*Fo, L, Pbs, Pbr*). We assume that in NDT the users behaviour is ordinary and then *Fo* and *L*= (*Pad, Pid, Pis, Pns, Par, Pac*) have fixed mean values, empirical received from measurements in operational system.

Theorem 1: Function *ofr.Ys* is increasing regarding *Pbs* and has not extremum regarding *Pbs* in NDT.

Proof: From equations (3.4) and $Yab = 1 + Pbr(Nab - 1)$

$$ofr.Ys(Fo,L,dsn.Pbr,Pbs)=\frac{(1+Pbr\,(Nab-1))(1-Pad)(1-Pid)\;(S_{1Z}-S_{2Z}\;Pbr)}{S_1+S_2(1-Pbs)\,Pbr+S_3Pbs} \qquad (4.1)$$

follows

$$\left.\frac{\partial ofr.Ys}{\partial Pbs}\right|_{dsn.Pbr}=(1-Pad)(1-Pid)(S_{1z}-S_{2z}Pbr)\left.\frac{(S_3-S_2Pbr)(1+Pbr(Nab-1))}{(S_1-S_2(1-Pbs)Pbr-S_3Pbs)^2}\right|_{dsn.Pbr}$$

From $S_3=(1-Pad)(1-Pid)[PisTis-Tbs+(1-Pis)[Pns\,Tns+(1-Pns)[Tcs+2\,Tb]]]=$
$=(1-Pad)(1-Pid)[PisTis-Tbs+(1-Pis)[Pns\,Tns+(1-Pns)[Tcs+Tbr]]]+S_2$

follows $S_3>S_2$ therefore $S_3-S_2Pbr>0$.

Therefore *ofr.Ys* increases concerning *Pbs* for each $dsn.Pbr$ and has not extremum in NDT.

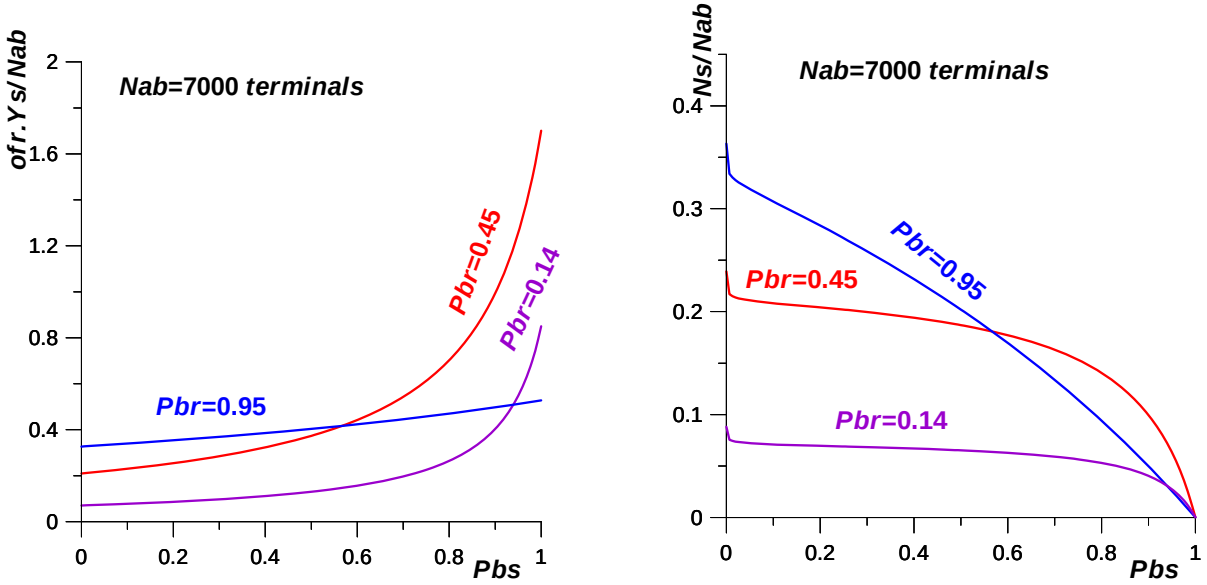


Fig. 1. and *Fig. 2. Dependencies of offered traffic ofr.Ys/Nab and Ns/Nab on target blocking probability due to insufficient lines value (trg. Pbs).*

Let *trg.Pbs* is determined in advance and fixed. Then *ofr.Ys*= *ofr.Ys* (*Fo, L, Pbs, Pbr*) is function regarding *Pbr* only.

Theorem 2: If *ofr.Ys* has extremum regarding *Pbr* in NDT, then the equation (4.2) is in force:

$$L\,Pbr^2+M\,Pbr+N=0,\ \text{where} \qquad (4.2)$$

$$L=S_{2Z}S_2(1-trg.Pbs)(Nab-1)$$

$$M=2S_{2Z}(S_3\,trg.Pbs-S_1)(Nab-1)$$

$$N=S_{1Z}[(Nab-1)(S_1-S_3\,trg.Pbs)+S_2(1-trg.Pbs)]-S_{2Z}(S_1-S_3\,trg.Pbs)$$

Proof: From equations (3.1.7) and (3.1.8) in analytical model [3] follows

$ofr.Ys=(1-Pad)(1-Pid)(S_{1}z-S_{2}z\,Pbr)Fa$, respectively

$$\left.\frac{\partial ofr.Ys}{\partial Pbr}\right|_{trg.Pbs}=(1-Pad)(1-Pid)\left(\frac{\partial Fa}{\partial Pbr}(S_{1Z}-S_{2Z}Pbr)-S_{2Z}Fa\right)\Bigg|_{trg.Pbs} \qquad (4.3)$$

where
$$Fa = \frac{1 + Pbr(Nab-1)}{S_1 - S_2(1-Pbs)Pbr - S_3 Pbs} \tag{4.4}$$

$$\frac{\partial Fa}{\partial Pbr} = \frac{(Nab-1)(S_1 - S_3 Pbs) + S_2(1-Pbs)}{(S_1 - S_2(1-Pbs)Pbr - S_3 Pbs)^2}. \tag{4.5}$$

Based on the analytical condition for existing of local extremum of *ofr.Ys* regarding *Pbr* follow the equation

$$S_{2Z} Fa = \frac{\partial Fa}{\partial Pbr}(S_{1Z} - S_{2Z} Pbr). \tag{4.6}$$

We may substitute (4.5) to (4.6) and after algebraic transform we receive regarding *Pbr* equation (4.2).

With this we proved that if *ofr.Ys* has local extremum regarding *Pbr* in NDT then *Pbr* is a solution of eq. (4.2). We will prove that exist at least one solution of eq. (4.2) in conditions of NDT.

Theorem 3: For equation (4.2) exist a solution $Pbr^* \in (0;1)$ in NDT with analytical conditions (4.7)-(4.8).

Proof: 1) If in eq. (4.2) coefficient *L* = 0 and *M≠0* (i.e. *Tb* ≠ *Tbr*), then exist only one solution $Pbr^* \in (0;1)$ on the conditions, following from the system:

$$\begin{cases} 0 < \dfrac{S_{1Z}[(Nab-1)(S_1 - S_3 trg.Pbs) + S_2(1-trg.Pbs)] - S_{2Z}(S_1 - S_3 trg.Pbs)}{2S_{2Z}(Nab-1)(S_1 - S_3 trg.Pbs} < 1 \\ S_{2Z} S_2 (1-trg.Pbs)(Nab-1) = 0 \\ \\ Tb \neq Tbr \end{cases} \tag{4.7}$$

2) When coefficient *L* ≠ 0 in eq. (4.2) and discriminant *D* ≥ 0

$$\begin{aligned} D = [&-S_{2Z}(S_1 - S_3 trg.Pbs)(Nab-1)]^2 - S_{2Z} S_2 (1-trg.Pbs)(Nab-1) \\ &[S_{1Z}[(Nab-1)(S_1 - S_3 trg.Pbs) + S_2(1-trg.Pbs)] - S_{2Z}(S_1 - S_3 trg.Pbs)] \end{aligned} \tag{4.8}$$

then exist at least one solution of eq. (4.2).

The condition $Pbr^* \in (0;1)$ is equivalent to *N* (*L*+*M*+*N*) ≠ 0 or to the follow system:

$$\begin{cases} [-2S_{2Z}(S_1 - S_3 trg.Pbs)(Nab-1)]^2 - S_{2Z} S_2 (1-trg.Pbs)(Nab-1) \\ [S_{1Z}[(Nab-1)(S_1 - S_3 trg.Pbs) + S_2(1-trg.Pbs)] - S_{2Z}(S_1 - S_3 trg.Pbs)] \geq 0 \\ \{S_2(1-trg.Pbs)[S_{2Z}(Nab-1) + S_{1Z}] + (S_{1Z} - 2S_{2Z})(Nab-1)(S_1 - S_3 trg.Pbs)\} \\ S_{1Z}[(Nab-1)(S_1 - S_3 trg.Pbs) + S_2(1-trg.Pbs)] - S_{2Z}(S_1 - S_3 trg.Pbs) \neq 0 \\ S_{2Z} S_2 (1-trg.Pbs)(Nab-1) \neq 0 \end{cases} \tag{4.9}$$

Numerical analysis shows that the bigger root in this case, fulfills the conditions above.

Theorem 4: *ofr.Ys* has a local maximum regarding *Pbr* if $Pbr^* = \dfrac{S_1 - S_3 trg.Pbs}{S_2(1-trg.Pbs)}$ when *Tb*>*Tbr* in NDT.

Proof: Investigation of equation (4.2) shows that when $S_{2Z} < 0$ (if coefficient *L* < 0), respective *Tb*>*Tbr,* then *ofr.Ys* has local maximum when $Pbr^* = \dfrac{S_1 - S_3 trg.Pbs}{S_2(1-trg.Pbs)}$ in NDT. Note that *Tb*>*Tbr* fulfils in all real situations.

Consequence 1: The value of local maximum of *ofr.Ys* regarding *Pbr* is

$$max\, ofr.Ys(Pbr)\big|_{trg.Pbs} = ofr.Ys(Pbr^*)\big|_{trg.Pbs} = \frac{[1+Pbr^*(Nab-1)](1-Pad)(1-Pid)\,Ts}{S_1 - S_3}$$

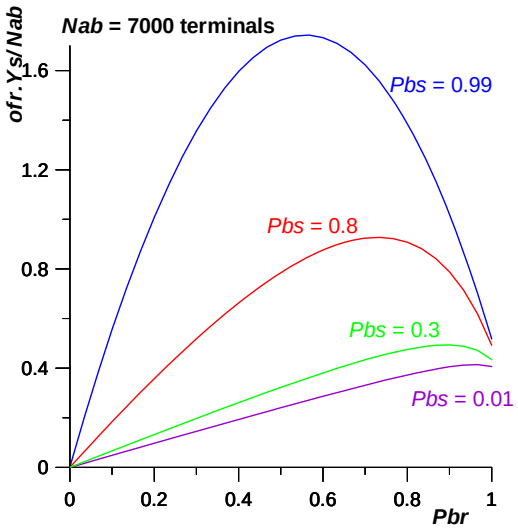
when target *trg.Pbs* is determined administratively in advance.

Consequence 2: In NDT the absolute maximum of *ofr.Ys* regarding *Pbr,* coincides with relative maximum in Pbr^*.

The absolute minimum of *ofr.Ys* is

$$min\, ofr.Ys(Pbr)\big|_{trg.Pbs} = \lim_{Pbr\to 0} ofr.Ys(Pbr)\Big|_{trg.Pbs} = \frac{(1-Pad)(1-Pid)\,S_{1Z}}{S_1}.$$

It is researched and numerical results are shown graphically on *Fig.3* and *Fig.4*.



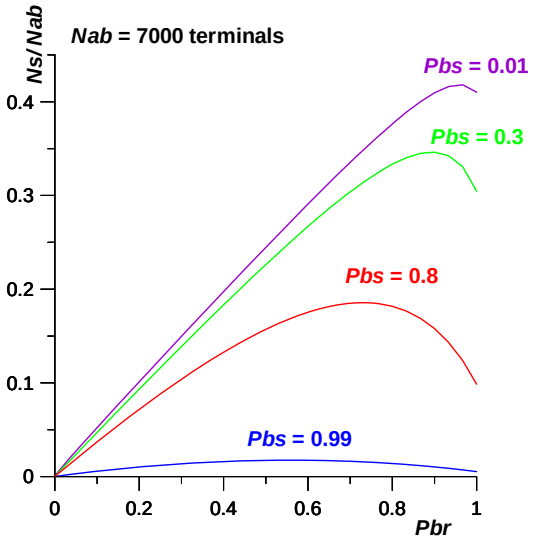


Fig. 3. and *Fig. 4. Dependency of offered traffic ofr.Ys/Nab and Ns/Nab on blocking probability of finding B – terminal busy Pbr.*

The range of *ofr.Ys* is

$$Range(ofr.Ys)\big|_{trg.Pbs} = [ofr.Ys(Pbr^*) - ofr.Ys(0)]\big|_{trg.Pbs} =$$

$$(1-Pad)(1-Pid)\left(\frac{(Nab-1)(1+Pbr^*(Nab-1))}{S_1 - S_3}(S_{1Z} - S_{2Z}Pbr^*) - \frac{S_{1Z}}{S_1}\right)\Bigg|_{trg.Pbs}$$

5. Conclusions

1. Detailed normalized conceptual model, of an overall (virtual) circuit switching telecommunication system is considered.
2. The target blocking probability *trg.Pbs* and probability of finding B – terminal busy (*Pbr*) as GoS – parameters in network dimensioning task, are used.
3. The behaviour of the traffic offered regarding QoS variables *Pbs* and *Pbr* is investigated. Function *ofr.Ys* is increasing regarding *Pbs* and has not extremum regarding *Pbs* in NDT. *Ofr.Ys* has a local maximum regarding *Pbr* which coincides with its absolute maximum. The conditions of these facts are investigated.
4. The results are useful for finding the range of *ofr.Ys* in every concrete case and developing of the best suitable numerical method for finding of the necessary number of equivalent internal switching lines (Ns) in dimensioning and redimensioning tasks.

5. The received results make the network dimensioning/redimensioning, based on QoS requirements easily, due to clearer behaviour of the important variables.
6. Numerical experiments are made and the results are graphically shown.
7. The described approach is applicable directly for every (virtual) circuit switching telecommunication system (like GSM and PSTN) and may help considerably for ISDN, BISDN and most of core and access networks dimensioning. For packet switching systems, like Internet, proposed approach may be used as a comparison basis and when they work in circuit switching mode.

Bibliography

1. Iversen V. B., 2004. Teletraffic Engineering and Network Planning, Technical University of Denmark, pp.125, 127;
2. ITU-T Recommendation E.600: Terms and Definitions of Traffic Engineering (Melbourne, 1988; revised at Helsinki, 1993);
3. Saranova E. T. ,2006. Dimensioning of telecommunication network based on quality of services demand and detailed behaviour of users- доклад на IV международна конференция " Information research, applications, and education i.tech", Варна, България, 20-25 юни 2006, изд. FOI- COMMERCE- Publisher 2006, ISBN-13: 978-954-16-0036-8, pp. 245- 256.
4. S. A. Poryazov, E. T. Saranova. ,2006. Some General Terminal and Network Teletraffic Equations in Virtual Circuit Switching Systems. Chapter in: A. Nejat Ince, Ercan Topuz (Editors). "Modeling and Simulation Tools for Emerging Telecommunications Networks: Needs, Trends, Challenges, Solutions", Springer Sciences+Business Media, LLC 2006, pp. 471-505. Printed in USA, Library of Congress Control Number: 2006924687. ISBN-13: 978-0387-32921-5 (HB) pp. 471-505;
5. Saranova E. T., 2006. Redimensioning of Telecommunication Network based on ITU definition of Quality of Services Concept, In: Proceedings of the International Workshop "Distributed Computer and Communication Networks", Sofia, Bulgaria, 2006, Editors: V. Vishnevski and Hr. Daskalova, Technosphera publisher, Moscow, Russia, 2006, p. 12;
6. ITU-T Recommendation E.501: Estimation of traffic offered in the network. (Previously CCITT - Recommendation, revised 26. May 1997);
7. Saranova E. T., 2007. Influence of some Users' Behaviour Parameters over Network Redimensioning – Proceedings of V International Conference " Information research, applications, and education i.tech", Varna, Bulgaria, june 2007, FOI- COMMERCE- Publisher 2007,

   ISSN 1313-1109, FSBN 978-954-16-2004-5
8. Poryazov S. A. 2005. What is Offered Traffic in a Real Telecommunication Network? COST 285 TD/285/05/05; 19th International Teletraffic Congress, Beijing, China, August 29- September 2, 2005, accepted paper No 32-104A.;
9. ITU-T Recommendation E.734 (10/96**),** *Methods for allocating and dimensioning Inteligent Network (IN) resources;*

Author's Information

Emiliya Saranova – e-mail: saranova@hctp.acad.bg, Emiliya@cc.bas.bg
Institute of Mathematics and Informatics - Bulgarian Academy of Science, Sofia, Bulgaria
College of Telecommunication and Posts, Sofia, Bulgaria

VOIP TRAFFIC SHAPING ANALYSES IN METROPOLITAN AREA NETWORKS

Rossitza Goleva, Mariya Goleva, Dimitar Atamian, Tashko Nikolov, Kostadin Golev

Abstract: This paper represents VoIP shaping analyses in devices that apply the three Quality of Service techniques – IntServ, DiffServ and RSVP. The results show queue management and packet stream shaping based on simulation of the three mostly demanded services – VoIP, LAN emulation and transaction exchange. Special attention is paid to the VoIP as the most demanding service for real time communication.

Keywords: Packet network, IP, Quality of Service, VoIP, shaping.

ACM Classification Keywords: C.4 Performance of Systems, C. Computer Systems Organization, C.2 Computer-Communication Networks

Introduction

IP networks and their Quality of the Service are challenging area for investigation. In spite of the fact that they are easy for use, enough cheap and quite useful in human life there is recently high demand for IP network use instead of all other kind of communication. Real time and non real time services and applications interwork on the same infrastructure. Different services have different quality requirements. The quality offered by the network depends on the traffic. In this paper we analyze the traffic shaping effect of the three mostly used techniques – IntServ, DiffServ, and RSVP. The analyses are made on the basis of the three popular services – VoIP, LAN emulation, transaction exchange [Jha], [Janevski], [Pitts], [Ralsanen]. The shaping effect is estimated under typical queueing circumstances. The model uses queues and priorities specific for the IntServ, DiffSerf, and RSVP. The reason is to investigate the effect that can be reached without implementation of the expensive shaping devices. This fractional shaping phenomenon is important in small to medium wire and wireless Metropolitan Area Networks (MAN) that grow rapidly. Changing circumstances in ad hoc networks also can apply the results presented.

Traffic sources

The traffic sources generate combination of three types of services in the network – Voice over IP, LAN emulation and transaction exchange. The size of the example network is typical for the business area. Some assumptions are made for every traffic source. In Voice over IP (VoIP) service silence and talk intervals are exponentially distributed with equal mean values [Jha], [Pitts]. There are authors who use talk to silence ratio of ½. Others do prefer to use on-off model for voice service. The behavior of the VoIP end-user is supposed to be similar to the phone user. The limits for waiting times are calculated under consideration of end-to-end delay bounds for every service [Lavenberg], [Iversen]. The same is valid for queue length. Servicing times per packets are fixed for LAN connection of 100 Mbps. Table 1 represents traffic sources parameters in the model.

LAN emulation is modeled with sessions. Sessions are established for any Internet connections. Packet rate is higher in comparison to the VoIP. Session duration is low. The traffic source is behaving as on-off model with exponential duration of the silence and transmission intervals [Lavenberg]. Transaction exchange is specific with few packets exchange. The service is not time demanding. Sessions are short and similar to the datagram exchange.

Table 1. Traffic Sources Parameters

No	Parameter	VoIP	LAN emulation	Transactions
1.	Pear rate, packets per second	10	164	0
2.	Mean call/ session duration, sec	180	20	10
3.	Mean duration between calls/sessions, sec	360	10	15
4.	Mean talk/ silence duration, sec	20	5	2
5.	Distribution of call/series duration	Exponential	Exponential	Exponential
6.	Maximal waiting time, sec	0.00072	0.6	1

7.	Maximal number of waiting packets	210	1804	2
8.	Traffic sources	5000	500	1500
9.	Priorities	High	Medium	Low
10.	Packet length, bytes	800	800	800

Number of traffic sources is taken from the typical business area. Packets are taken to be long. IP packets of 800 bytes carry up to 80 milliseconds voice. This means that quality voice can be transmitted only in the area with up to 2-3 hops. Therefore, we design VoIP service for regional connectivity. More precision investigation can be done with up to 200 bytes voice packets.

Integrated Services

Integrated Services (IntServ) is a complex technique that ensures Quality of Service in IP networks. It is applied usually in access routers or gateways and tried to serve packets from different services in a different ways depending on the quality requirements. IntServ classifies services into three main classes depending on the traffic requirements [Janevski]:

- Elastic application;
- Tolerant real-time applications;
- Intolerant real-time applications.

Elastic applications are served with “best effort” discipline [Tanenbaum]. They are served without any guarantee of quality level like transaction exchange. Tolerant real-time applications are delay sensitive and usually require high bandwidth. Token bucket model with peak rate control is a proper model for such traffic. LAN emulation is usually modeled this way. Some authors propose token bucket that controls series length and mean rate for more accuracy. Many authors propose the two token buckets to be connected in a cascade as it is shown on Figure 1 [Ralsanen]. Intolerant real-time applications require low delays and almost guaranteed bandwidth. The model with two cascaded token buckets is compulsory for such traffic [Jha]. VoIP service is intolerant to the quality degradation service. IntServ simulation model is based on two cascaded token buckets that bound peak rate, series length and mean rate of the traffic (Figure 1). The model is approximated as a black box that changes the characteristics of the data at output in specific for IntServ way. As a result after approximation and few calculations it is easy to derive simpler model with one FIFO queue, priorities, fixed rate at the output and different limits for waiting times in the queue. The resulting model is represented on Figure 2. This is the model that has been simulated further. Table 2 represents main data for model behavior.

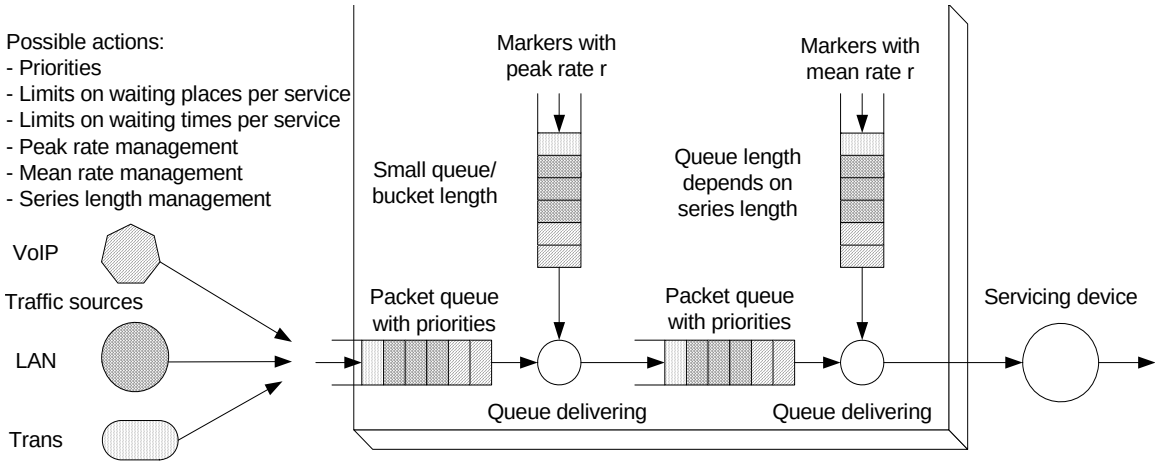


Figure 1. Black box IntServ model approximation

Differentiated Services

Differentiated Services (DiffServ) is another quality management technique that is more applicable for core networks. Due to its nature DiffServ applies its rules on aggregated traffic. After appropriate marking of the aggregated packets they are gathered in the way that is defined for their class. There are three main types of services we highlighted in this paper [Pitts]:

- Premium service with low delay, low loss, guaranteed bandwidth like VoIP;
- Assured service with less requirements to the delay and loss in comparison to the premium service like LAN emulation;
- Olympic service with no time requirements at all like transaction exchange.

The model from Figure 2 with different parameters is used to represents DiffServ application. The parameters are shown on Table 2.

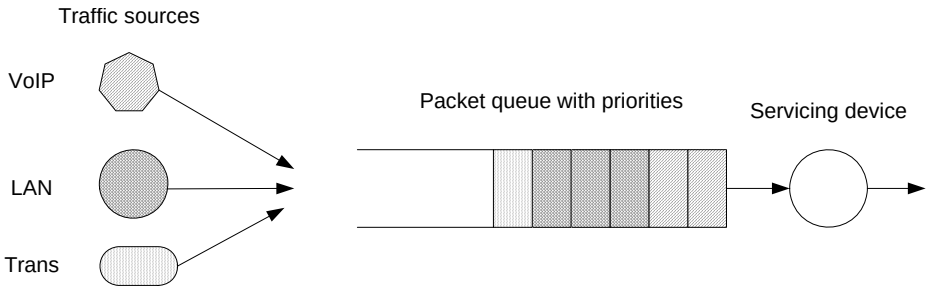


Figure 2. Final IntServ model with input data, bounds for waiting times and queue length specific for service type

RSVP

Resource Reservation Protocol (RSVP) is a technique useful for delay sensitive traffic like VoIP. Three types of services are identified for RSVP like:

- Wildcard filter that is applied to gather maximal requirements for given interface like LAN emulation;
- Shared explicit that is applied to gather maximal requirements for the interface taking into account called address. Transaction exchange is modeled as shared explicit service;
- Fixed filter that requires full reservation for quality sensitive services like VoIP.

The model simplified for IntServ and DiffServ procedures is applied with specific parameters for RSVP. Characteristics of the derived model are shown on Table 2.

Table 2. Model characteristics

No	Parameter	IntServ	DiffServ	RSVP
1.	Queue length, packets	2016	1840	1840
2.	VoIP queue length fraction, packets	210	200	200
3.	LAN queue length fraction, packets	1804	1640	1640
4.	Transaction queue length fraction, packets	2	2	2
5.	Maximal waiting time for VoIP, sec	0,000716	0,0303	0,07508
6.	Maximal waiting time for LAN, sec	0,6	0,27876	0,69
7.	Maximal waiting time for transactions, sec	1	1	1
8.	Priority for VoIP	Highest	Highest	Highest
9.	Priority for LAN	Medium	Medium	Medium
10.	Priority for transactions	Low	Low	Low

Results

Simulation is performed on C++ language. The pseudo exponential pseudo deterministic characteristics of the traffic sources are reached after usage of combination between many random generators [Kleinrock], [Iversen], [Lavenberg]. The queue behavior is complex due to the priorities and limits for waiting times. Many parameters have been derived from the model like time and space loss probabilities, probabilities to wait for different types of traffic, statistical data for probability distribution functions and probability density functions of the packets intervals, queue lengths, waiting times at many interface points of the model like output of the traffic sources, input and output of the queue. Statistical accuracy of the derived results is proven by Student criterion. IntServ, DiffServ and RSVP have different way to gather with packets and this influences the way they drop packets and shape them.

On Figure 3, 4 and 5 observations of packet intervals at the input and output of the queue are shown. It is interesting for shaping estimation. The effect of fast servicing in RSVP can be seen from Figure 3. The delay variation of the packet intervals is becoming smoother and tends to constant value. Similar result is visible for IntServ on Figure 4. On Figure 5 shaping of the IntServ and DiffServ is seen.

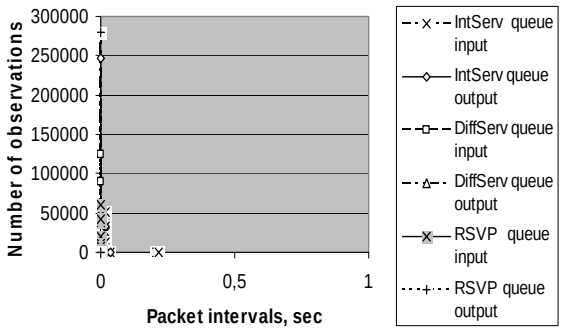


Figure 3. Delay variation reduction in RSVP

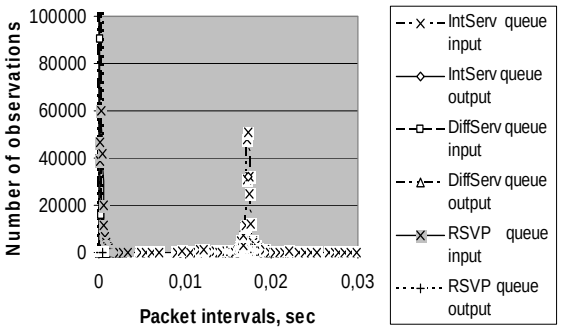


Figure 4. Delay variation reduction in IntServ

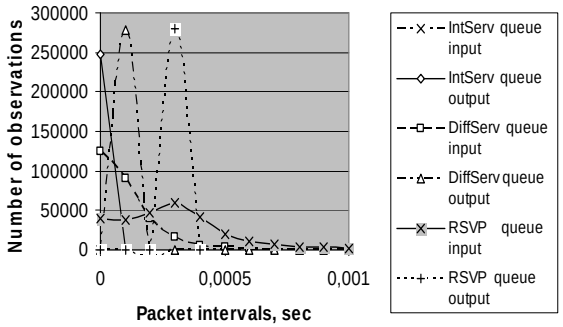


Figure 5. Delay variation reduction in IntServ and DiffServ

Interesting results that influence directly interfaces and queue management are derived on the basis of queue length per service type. The queue fraction of the three services is observed. It is visible from Figure 6 that for services with highest priority like VoIP IntServ is the most proper shaping mechanism. With some not quite accurate approximation the distribution of the queue length can be considered exponential. Figure 7 represents the observations for LAN service. Because of the less critical waiting times and low priority the distribution tends to be deterministic.

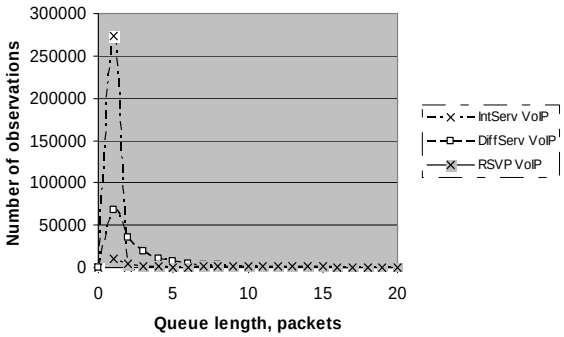


Figure 6. Observations of queue length in VoIP

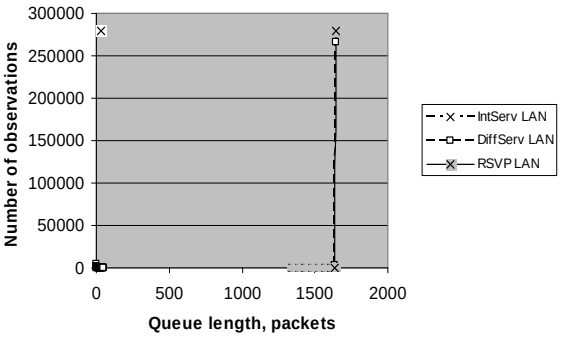


Figure 7. Observations of queue length in LAN

On Fugure 8 and 9 the observations of packet intervals only between voice packets are shown. The statistical multiplexing effect and shaping phenomenon are due to the high priority of the voice traffic in comparison to the priority of the data traffic. On Figure 10 the shaping effect of the three techniques is visible. On Figure 11 and 12 only effect of DiffServ is obvious in different observation scales.

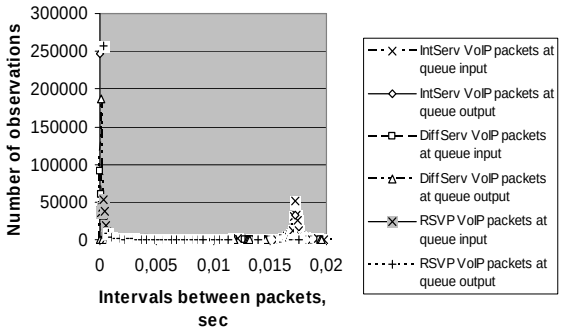


Figure 8. Observations of intervals between VoIP packets

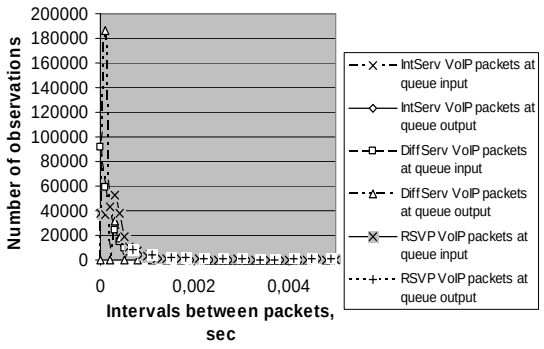


Figure 9. Observations of intervals between VoIP packets for DiffServ

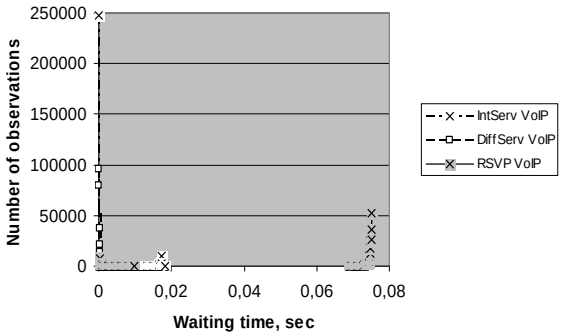


Figure 10. Observations of waiting times for VoIP packets

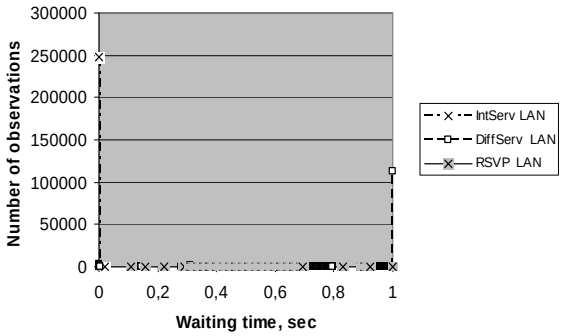


Figure 11. Observations of waiting times for LAN packets

Conclusion

In this paper we show observations of the packet intervals at the queue input and queue output as well as statistical data of queue length, waiting times and loss per service type (Table 3). These results demonstrate the specific characteristics of the queue as a packet shaper in three QoS management algorithms IntServ, DiffServ, RSVP. The shaping effect is possible for priority service types.

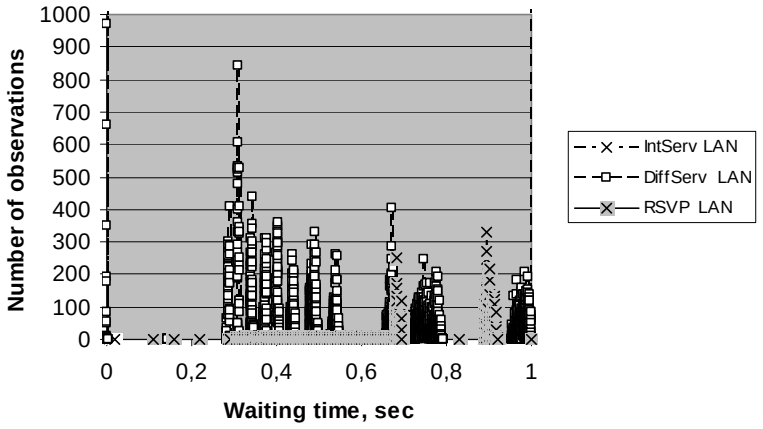


Figure 12. Observations of waiting times for LAN packets

The low delays for priority services types are due to the bigger delays for non priority service types. The waiting times are redistributed due to the QoS algorithm and priority. The deterministic nature of the packet streams suppress shaping and increase losses. The statistical multiplexing effect is very limited due to the deterministic streams. Mean values of the queue lengths, probability to wait, loss probability due to the lack of space and waiting time bounds per discipline

and per service redistribution are visible from Table 3. They can be used for configuration planning of the time and space limits in the router interfaces.

The results demonstrate the capability of IntServ to define excellent service for its higher priority applications. It is promising in access networks. DiffServ shows excellent resource management and utilization and therefore is better for core services. RSVP is a good counterpart of IntServ in access networks.

The authors refine the simulation model with more traffic sources and more precise generation of the packets from these sources based on the observation of the real traffic. MMPP and geometric/ Weibull distributions are also considered. Limits criteria for queue management are under investigation.

Table 3. Queue length, waiting times, loss probability

Mechanism	IntServ	DiffServ	RSVP
Overall mean queue length, packets	37.97523	1586.961	1798.409
Mean queue length of VoIP fraction, packets	1	2.24207	156.9017
Mean queue length of LAN fraction, packets	35	1583.561	1639.675
Mean queue length of Trans fraction, packets	2	1.98331	2
Overall loss probability due to the lack of space	0.0094	0.77815	0.91222
VoIP packets loss probability due to the lack of space	0	0	0.33540
LAN packets loss probability due to the lack of space	0	0.89373	0.98747
Transaction packets loss probability due to the lack of space	1	0.99574	1
Overall loss probability due to waiting time bound	0.98865	0	0
VoIP packets loss probability due to waiting time bound	0.98109	0	0
LAN packets loss probability due to waiting time bound	1	0	0
Transaction packets loss probability due to waiting time bound	0	0	0
Overall probability to wait	0.00191	0.22074	0.08767
VoIP packet probability to wait	0.0189	0.9996	0.66433
LAN packet probability to wait	0	0.105130	0.01244
Transaction packet probability to wait	0	0.00339	0
Overall interface occupancy, fraction	0.13668	0.13644	0.13704
Interface occupancy due to the VoIP traffic, fraction	0.13621	0.05046	0.08955
Interface occupancy due to the LAN traffic, fraction	0.00048	0.08588	0.04749
Interface occupancy due to the transaction traffic, fraction	0	0.00009	0

Acknowledgements

This paper is sponsored by the Ministry of Education and Research of the Republic of Bulgaria in the framework of project No 105 “Multimedia Packet Switching Networks Planning with Quality of Service and Traffic Management”.

Bibliography

[Jha, 2002] Jha, S., M. Hassan, “Engineering Internet QoS”, Artech Hose, 2002.

[Janevski, 2003] Janevski, T., “Traffic Analysis and Design of Wireless IP Networks”, Artech House, 2003.

[Kleinrock, 1976] Kleinrock, Leonard, “Queueing Systems”, Volumes I and II, John Wiley and Sons, 1976.

[Iversen, 2005] Iversen, V., “Teletraffic Engineering Handbook”, ITU-D, 2005.

[Lavenberg, 1983] Lavenberg, Stephen S., Editor, “Computer Performance Modeling Handbook”, Academic Press, 1983, ISBN 0-12-438720-9.

[Pitts, 2000] Pitts, J., J. Schormans, “Introduction to IP and ATM Design and Performnce”, John Wiley&Sons, Ltd., 2000.

[Ralsanen, 2003] Ralsanen, V., “Implementing Service Quality in IP Networks”, John Wiley & Sons, Ltd., 2003.

[Tanenbaum. 2003] Tanenbaum, Andrew S., “Computer Networks, Second Edition”, Prentice-Hall International, Inc., 2003, ISBN 0-13-166836-6.

Authors' Information

Rossitza Iv. Goleva – Assistant-Professor; Department of Telecommunications, Technical University of Sofia, Bulgaria, Kl. Ohridski blvd. 8, Sofia. 1756, Bulgaria; e-mail: rig@tu-sofia.bg

Mariya At. Goleva – Student; Department of Communication Networks, University of Bremen, Germany; e-mail: mgoleva@gmail.com

Dimitar K. Atamian - Assistant-Professor; Department of Telecommunications, Technical University of Sofia, Bulgaria, Kl. Ohridski blvd. 8, Sofia. 1756, Bulgaria; e-mail: dka@tu-sofia.bg

Tashko Nikolov – Assistant-Professor, Ph.D. Department of Telecommunications, Technical University of Sofia, Bulgaria, Kl. Ohridski blvd. 8, Sofia. 1756, Bulgaria; e-mail: tan@tu-sofia.bg

Kostadin At. Golev – Developer; Bianor Ltd., Bulgaria; e-mail: kotseto@gmail.com

STUDY OF QUEUEING BEHAVIOUR IN IP BUFFERS

Seferin Mirtchev

Abstract: It is unquestioned that the importance of IP network will further increase and that it will serve as a platform for more and more services, requiring different types and degrees of service quality. Modern architectures and protocols are being standardized, which aims at guaranteeing the quality of service delivered to users. In this paper, we investigate the queueing behaviour found in IP output buffers. This queueing increases because multiple streams of packets with different length are being multiplexed together. We develop balance equations for the state of the system, from which we derive packet loss and delay results. To analyze these types of behaviour, we study the discrete-time version of the "classical" queue model M/M/1/k called Geo/Gx/1/k, where Gx denotes a different packet length distribution defined on a range between a minimum and maximum value.

Keywords: delay system, queueing analyses, discrete time queue, IP traffic modelling; packet size distribution.

ACM Classification Keywords: G.3 Probability and statistics: queueing theory, I.6.5 Model development

Introduction

The initial motivation for this paper is the necessity of traffic engineering in IP networks. Many analyses of Internet traffic behaviour require accurate knowledge of the traffic characteristics for purposes ranging from a management of the network quality of service to modelling the effect of new protocols on the existing traffic mix.

Modern architectures and protocols are being standardized, which aims at guaranteeing the quality of service delivered to users. The proper functioning of these protocols requires an increasingly detailed knowledge for statistical characteristics of IP packets. The amount of information flowing through the network also increases, and the challenge is to obtain the accurate information from a huge set of data packets.

The packet queueing in an IP router arises because multiple streams of packets from different input ports are being multiplexed together over the same output port. A key characteristic is that the packets have different length. The minimum header size in IPv4 is 20 octets, and in IPv6, it is 40 octets. The maximum packet size depends on the specific sub-networks technology: 1500 octets in Ethernet and 1000 octets are common in X.25 networks. The packet length distribution measured from the real traces exhibits the well-known multi-mode behaviour, with peaks for very short packets and for the different maximum transfer units in the network, with a dominating peak at 1500 bytes, due to the size of Ethernet frame. This specific packet length distribution has a direct impact on the service time and we need a different approach to the queueing analysis.

Discrete-time queueing systems have been a research topic for several decades now and there are many reference works on discrete-time queueing theory. Over the years, different methodologies have been developed to assess the performance of queueing systems. The two main analytical approaches are the matrix analytic method and the transform method for discrete and for continuous-time analyses. Many authors have considered the Geo/G/1 queueing system [Pitts, 2000], [Mirtchev, 2006], [Vicari, 1996], [Zang, 2001].

In [Atencia, 2005] is carried out a complete study of a discrete-time single-server queue with geometrical arrivals of both positive and negative customers. Negative arrivals are used as a control mechanism in many telecommunication and computer networks. [Atencia, 2006] is concerned with the study of a discrete-time single-server retrial queue with geometrical inter-arrival times and a phase-type service process. An iterative algorithm to calculate the stationary distribution of Markov chain is given.

[Salvador, 2004] is proposed a traffic model and a parameter fitting procedure that are capable of achieving accurate prediction of the queuing behaviour for IP traffic exhibiting long-range dependence. The modelling process is a discrete-time batch Markovian arrival process (dBMAP) that jointly characterizes the packet arrival process and the packet size distribution. In the proposed dBMAP, packet arrivals occur according to a discrete-time Markov modulated Poisson process (dMMPP) and each arrival is characterized by a packet size with a general distribution that may depend on the phase of the dMMPP.

[Cao, 2004] is presented an introduction to bandwidth estimation and a solution to the problem of the best-effort traffic for the case where the quality criteria specify negligible packet loss. The solution is a simple statistical model, which is built and validated using queueing theory and extensive empirical study.

It has been shown [Dan, 2005] that in the case of real-time communications, for which small buffers are used for delay reasons, short range dependence dominates the loss process and so the Markov-modulated Poisson process (MMPP) might be a reasonable source model. They have presented an exact mathematical model for the loss process of a MMPP+M/Ek/1/K queue and have concluded that the packet size distribution affects the packet loss process and thus the efficiency of forward error correction.

In this paper, we investigate the basic queueing behaviour of packets found in IP output buffers. This queueing is complicated because multiple streams of packets are being multiplexed together. The traffic is being generated from the packets of varying sizes that arrive for transmission on the link. The packets can queue up and loss if their size is bigger than the free positions of the buffer. The quality metrics for the best-effort traffic on the Internet are the packets loss and delay. To analyze these types of behaviour, we study the discrete-time version of the "classical" queue model M/M/1/k called Geo/Gx/1/k, where Gx denotes a different packet length distribution. We developed balance equations for the state of the system, from which we derived packets loss and delay.

Balance equations for the queue model Geo/Gx/1/k

Let us consider a single server finite queue delay system *Geo/Gx/1/k* with a geometric distributed inter-arrival time and different distributions of the packet length: truncated geometric, binomial, discrete uniform and discrete triangular. These packet length distributions are defined on a range between a minimum and maximum value.

We consider queueing phenomena in discrete-time queueing systems. That is, we assume a fundamental time unit (time slot), the time to transmit an octet (byte), T_b. Customers arrive in the queueing system under consideration during the consecutive slots, but they can only start service at the beginning of slots. That is, service of customers is synchronized with respect to slot boundaries. Further, customer service times are integer multiples of the slot length, which implies that customers leave the system at slot boundaries. During the consecutive slots, packets arrive in the system, are stored in a finite capacity queue and are served by a single server on a first in first out (FIFO) basis (fig.1).

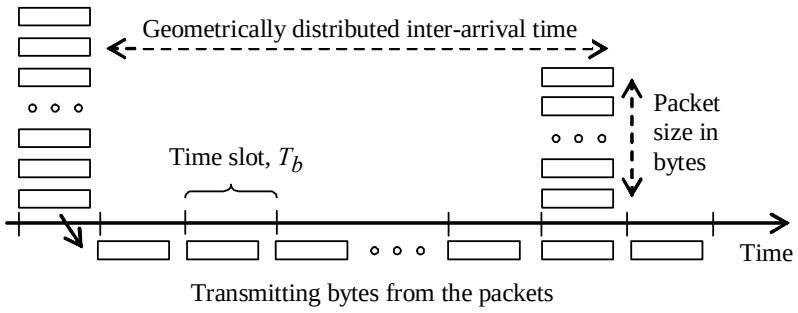


Fig.1. Timing of events in the Geo/Geo/1/k queueing system

We use a Bernoulli process for the packet arrivals, i.e. a geometrically distributed number of slots between arrivals. Let the probability that a packet arrives in an octet slot is p.

In this model, we assume a truncated geometric distribution at variable packet sizes with a minimum value m_1 and a maximum value m_2, as the first kind of distribution.

Let the probability that a packet completes service at the end of an octet slot is q. We define the probability that the packet size is n octets:

$$b_n = \frac{q(1-q)^{n-m_1}}{q\sum_{r=0}^{m_2-m_1}(1-q)^r}, \qquad m_1 \le n \le m_2 \ . \tag{1}$$

The mean number of bytes in the packet by definition is

$$b = \sum_{i=m_1}^{m_2} i\, b_i \approx 1/q + m_1 \ . \tag{2}$$

The second kind of a packet size distribution is binomial

$$\begin{aligned} b_n &= \binom{m_2 - m_1}{n - m_1} q^{n-m_1}(1-q)^{m_2-n}, \qquad m_1 \le n \le m_2 \\ b &= m_1 + (m_2 - m_1)q \end{aligned}, \tag{3}$$

The third kind of a packet size distribution is discrete uniform

$$\begin{aligned} b_n &= \frac{1}{m_2 - m_1 + 1} \quad for\ all\ values\ of\ n, \qquad m_1 \le n \le m_2 \\ b &= (m_2 + m_1)/2 \end{aligned}, \tag{4}$$

The next kind of a packet size distribution is discrete triangular. When the mode is equal to the minimum value, we have linear decreasing distribution with the following probabilities that the packet size is n octets and the mean number of the bytes in the packet

$$\begin{aligned} b_n &= \frac{m_2 - n + 1}{\sum_{r=m_1}^{m_2} m_2 - r + 1} \ , \qquad m_1 \le n \le m_2 \\ b &= m_1 + (m_2 - m_1)/3 \end{aligned}, \tag{5}$$

When the mode is equal to the maximum value, we have linear increasing discrete triangular distribution

$$\begin{aligned} b_n &= \frac{n - m_1 + 1}{\sum_{r=m_1}^{m_2} r - m_2 + 1} \ , \qquad m_1 \le n \le m_2 \\ b &= m_1 + 2(m_2 - m_1)/3 \end{aligned}, \tag{6}$$

Thus we have a batch arrival process with geometrically distributed inter-arrival times. That is, the number of slots that separate consecutive slots where there are customer arrivals, constitute a series of independent and identically geometric distributed random variables. The probability no octets arriving in a time slot is

$$a_0 = 1 - p \ . \tag{7}$$

The probability that n octets arriving in a time slot is

$$a_n = p\, b_n, \qquad m_1 \le n \le m_2 \ . \tag{8}$$

The mean packet service time is the octet transmission time multiplied by the mean number of octets

$$\tau = T_b \sum_{i=m_1}^{m_2} i\, b_i = T_b\, b, \ s \ . \tag{9}$$

The mean arrival rate is

$$\lambda = p/T_b, \quad packets/s \ . \tag{10}$$

Therefore, the offered traffic is given by

$$A = \lambda\tau = p \sum_{i=m_1}^{m_2} i\, b_i, \ erl \ . \tag{11}$$

We define the state probability P_i of being of state *i*, as the probability that there are *i* octets in the system at the end of any time slot. For the system to contain *i* bytes at the end of any time slots it could have contained any of *0,1,2,…..,i+1* at the end of the previous slot. State *i* can be reached from any of the states *0* up to *i* by a precise number of arrivals. To move from *I +1* to *i* requires that there are no arrivals.

We can write the first equation by considering all the ways in which it is possible to reach the empty state

$$P_0 = P_0 a_0 + P_1 a_0 \quad . \tag{12}$$

Similarly, we find a formula for the next state probabilities by writing the balance equations

$$P_i = P_{i+1} a_0 , \qquad 1 \le i \le m_1 - 1 \quad . \tag{13}$$

We continue with this process when the packet arrives in a time slot with length between m_1 and m_2.bytes

$$\begin{aligned}
P_{m_1} &= (P_0 + P_1) a_{m_1} + P_{m_1+1} a_0 \\
P_{m_1+1} &= (P_0 + P_1) a_{m_1+1} + P_2 a_{m_1} + P_{m_1+2} a_0 \\
&\quad o \;\; o \;\; o \\
P_{m_2} &= (P_0 + P_1) a_{m_2} + P_2 a_{m_2-1} + \ldots + P_{m_2-m_1+1} a_{m_1} + P_{m_2+1} a_0 \\
P_{m_2+1} &= P_2 a_{m_2} + P_3 a_{m_2-1} + \ldots + P_{m_2-m_1+2} a_{m_1} + P_{m_1+2} a_0 \\
&\quad o \;\; o \;\; o \\
P_{k-1} &= P_{k-m_2} a_{m_2} + P_{k-m_2+1} a_{m_2-1} + \ldots + P_{k-m_1} a_{m_1} + P_k a_0 \\
P_k &= P_{k-m_2+1} a_{m_2} + P_{k-m_2+2} a_{m_2-1} + \ldots + P_{k-m_1+1} a_{m_1} + P_{k+1} a_0
\end{aligned} \tag{14}$$

Then using the fact that all the state probabilities must sum to 1

$$\sum_{i=0}^{k+1} P_i = 1 \quad , \tag{15}$$

We can solve the system equations (12), (13), (14) and 15 and calculate the state probabilities.

Performance Measures

The carried traffic is equivalent to the probability that the system is busy

$$A_o = 1 - P_0 , \; erl \quad . \tag{16}$$

The packet congestion probability is the ratio of lost traffic (offered minus carried traffic) to offered traffic

$$B = \left(A - A_o\right) / A \quad . \tag{17}$$

The mean number of bytes and packets present in the system in steady state by definition is

$$L_b = \sum_{j=1}^{k+1} j P_j , \; bytes; \qquad L_p = L_b / b , \; packets \; . \tag{18}$$

From the Little formula, we have the normalized mean system time of the bytes (time is measured in time slots)

$$\frac{W_b}{T_b} = \frac{L_b}{T_b \lambda b} = \frac{L_b}{A} \quad . \tag{19}$$

Numerical Results

In this section, we give numerical results obtained by a Pascal program on a personal computer. The described methods were tested on a computer over a wide range of arguments.

Figures 2 and 3 show the stationary probability distribution in a single server queue *Geo/Gx/1/k* with 0.8 and 0.7 erl offered traffic respectively, 1000 waiting positions, 30 bytes minimum packet length, 80 bytes maximum packet length and different packet length distributions: discrete uniform, truncated geometric, binomial, discrete triangular decreasing and discrete triangular increasing. We can see that the probability distributions are almost linear decreasing in logarithmic scale and the influence of the packet length distribution kind on the stationary probability is negligible even though in case of discrete triangular increasing packet length distribution.

Figures 4 and 5 illustrate the dependence on the packet congestion probability from the queue length when the offered traffic is 0.7 erl, the range of packet length is from 30 to 80 bytes and different packet length distributions. When the queue length is big the packet congestion probability is almost linear decreasing in logarithmic scale.

The packet length distribution in defined rage is not so essential. The main reason for this behaviour is the fact that the packet length is limited.

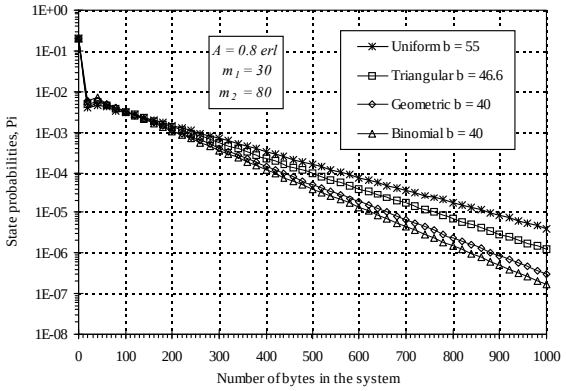


Fig.2. Graph of the state probability distributions for a finite queue with Geometric, Binomial, Uniformly and Triangular decreasing packet length distribution

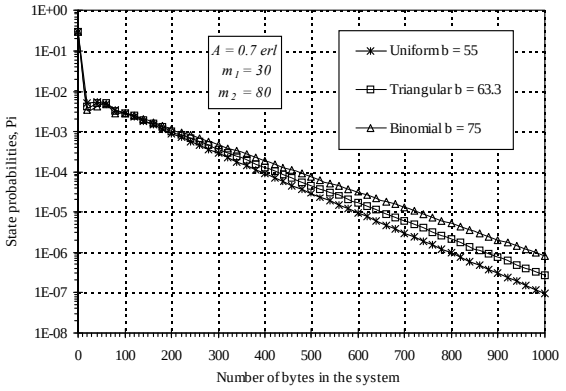


Fig.3. Graph of the state probability distributions for a finite queue with Binomial, Uniformly and Triangular increasing packet length distribution

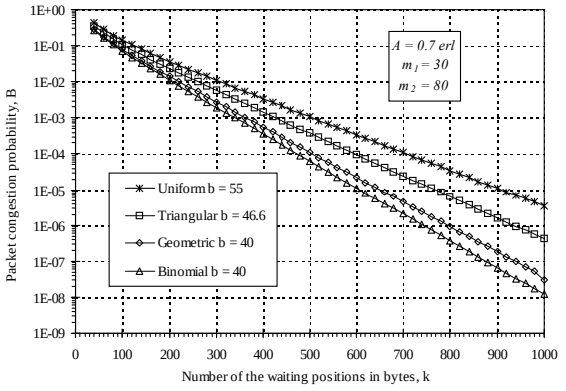


Fig.4. Packet congestion probability in the Geo/Gx/1/k with different packet length distributions and mean packet lengths between the minimum and the average value

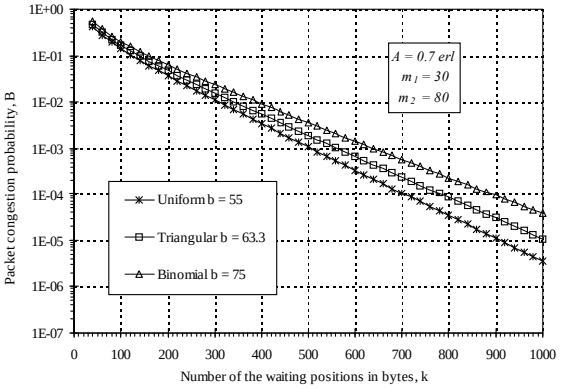


Fig.5. Packet congestion probability in the Geo/Gx/1/k with different packet length distributions and mean packet lengths between the average and the maximum value

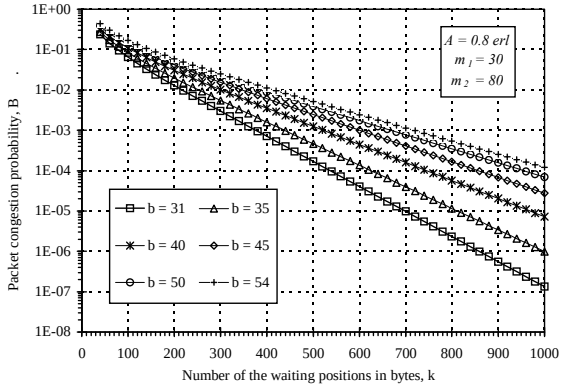


Fig.6. Packet congestion probability in discrete time single server queue with a truncated geometric packet length distribution and different mean packet lengths between the minimum and the average value

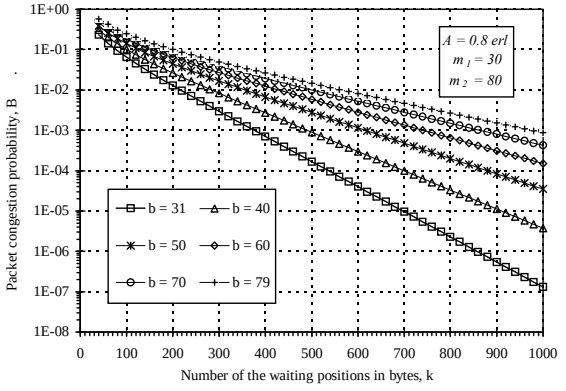


Fig.7. Packet congestion probability in discrete time single server queue with a binomial packet length distribution and different mean packet lengths between the minimum and the maximum value

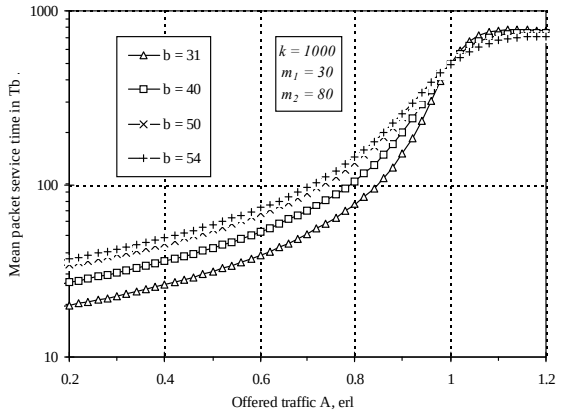


Fig.8. Normalized mean system time of the bytes in discrete time single server queue with a truncated geometric packet length distribution and different mean packet lengths

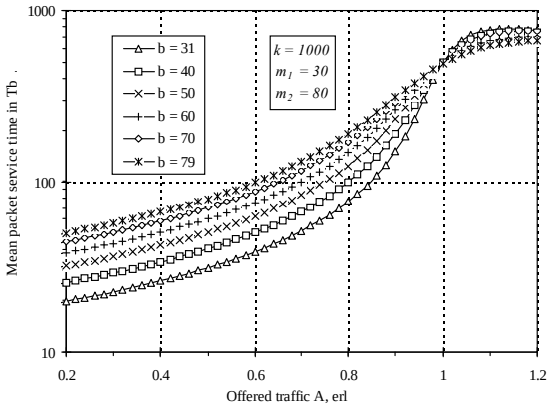


Fig.9. Normalized mean system time of the bytes in discrete time single server queue with a binomial packet length distribution and different mean packet lengths

Figures 6 and 7 compare the packet congestion probability when the offered traffic is 0.8 erl, the range of packet length is from 30 to 80 bytes, truncated geometric and binomial distribution accordingly and different mean packet size. We can see that the influence of the mean packet length on the packet congestion probability is big.

Figures 8 and 9 present the normalized mean system time of the bytes (W/T_b) as function of the traffic intensity when the queue length is 1000 bytes, the range of packet length is from 30 to 80 bytes, truncated geometric and binomial distribution accordingly and different mean packet size. The influence of the mean packet size on the mean system time is significant when the offered traffic is smaller then 1 erl.

Conclusion

In this paper, different distributions of the packet length: truncated geometric, binomial, discrete uniform and discrete triangular are used and explained. A basic discrete-time single server teletraffic system Geo/Gx/1/k is examined in detail.

The proposed approach provides a unified framework to model discrete-time single server queue. Numerical results and subsequent experience have shown that this approach is accurate and useful in both analyses and simulations of traffic systems.

The importance of a single server queue in a case of a geometric input stream and different distributions of the packet length comes from its ability to describe behaviour that is to be found in more complex real queueing systems. It is the case in a general traffic system, which is an important feature in designing telecommunication networks and systems.

The results presented here add a new aspect to the evaluation of the discrete-time queueing system, and serve as a basis for future research on guaranteeing the quality of service

In conclusion, we believe that the presented formulas will be useful in practice.

Acknowledgements

This paper is sponsored by the National Science Funds of MES - Bulgaria in the framework of project **BY-TH-105/2005** “Multimedia Telecommunications Networks Planning with Quality of Service and Traffic Management”.

Bibliography

[Atencia, 2005] Atencia I. and P. Moreno. A single-server G-queue in discrete-time with geometrical arrival and service process. Perform. Eval. 59: pp. 85-97 (2005)

[Atencia, 2006] Atencia I, P. Bocharov and P. Moreno. A discrete-time Geo/PH/1 queueing system with repeated attempts. Информационные процессы, Том 6, N: 3, стр. 272-280 (2006).

[Cao, 2004] Cao J., W. Cleveland and D. Sun. Bandwidth estimation for best-effort Internet traffic. Statist. Sci., Volume 19, Number 3 (2004), pp. 518-543.

[Dan, 2005] Dan G., V. Fodor, and G. Karlsson, "Packet size distribution: an aside?" in Proc. of QoS-IP'05, pp. 75–87, February 2005.

[Farber, 2002] Farber J., S. Bodamer and J. Charzinski. Measurement and Modelling of Internet Traffic at Access Networks, Proceedings of the EUNICE'98,1998,196-203.

[Janevski, 2003] Janevski T., D. Temkov, A. Tudjarov: Statistical Analysis and Modelling of the Internet Traffic. ICEST Sofia, 2003, pp. 170-173.

[Mirtchev, 2006] Mirtchev S., G. Balabanov and S. Statev, New Teletraffic Models in the IP Networks. National Conference with Foreign Participation, Telecom'2006, Varna, Bulgaria, 2006 (in Bulgarian).

[Pitts, 2000] Pitts J.and J. Schormans. Introduction to IP and ATM Design and Performance - 2nd Ed., John Wiley & Sons, 2000.

[Salvador, 2004] Salvador P., A. Pacheco and R. Valadas. Modelling IP traffic: joint characterization of packet arrivals and packet sizes using BMAPs. Computer Networks, Volume 44, Issue 3 , 2004, pp. 335-352.

[Vicari, 1996] Vicari N. and P. Tran-Gia. A numerical analysis of the Geo/D/N queueing system. Technical Report 04, COST-257, 1996.

[Zang, 2001] Zhang Z. and N.Tian. Discrete Time Geo/G/1 Queue with Multiple Adaptive Vacations, Queueing Systems, Volume 38, Number 4, August 2001, pp. 419-429.

Authors' Information

Seferin Mirtchev – Technical University of Sofia, Kliment Ohridski St., N:8, Bl.1, Sofia-1000, Bulgaria; e-mail: stm@tu-sofia.bg

TOWARDS USEFUL OVERALL NETWORK TELETRAFFIC DEFINITIONS

Stoyan Poryazov

Abstract. A detailed conceptual and a corresponding analytical traffic models of an overall (virtual) circuit switching telecommunication system are used. The models are relatively close to real-life communication systems with homogeneous terminals. In addition to Normalized and Pie-Models Ensue Model and Denial Traffic concept are proposed, as a parts of a technique for presentation and analysis of overall network traffic models functional structure; The ITU-T definitions for: fully routed, successful and effective attempts, and effective traffic are re-formulated. Definitions for fully routed traffic and successful traffic are proposed, because they are absent in the ITU-T recommendations; A definition of demand traffic (absent in ITU-T Recommendations) is proposed. For each definition are appointed: 1) the correspondent part of the conceptual model graphical presentation; 2) analytical equations, valid for mean values, in a stationary state. This allows real network traffic considered to be classified more precisely and shortly. The proposed definitions are applicable for every telecommunication system.

Keywords: Overall Network TrafficTheory, ITU-T Definitions, Virtual Circuits Switching.

ACM Classification Keywords:

1. Introduction

The first what we need for usable Overall Network and Terminal Traffic Theory, is a complete set of clear, precise and useful definitions, particularly for overall network characteristics.

State of the art: Expressions "offered traffic" and "demand traffic" are not found in "ETSI Publications Download Area" [http://pda.etsi.org/pda/queryform.asp] and in [ANSI 2001]. The ITU-T definition of offered traffic is not valid for real telecommunication systems [Poryazov 2005] and that one for demand traffic is simply absent, despite usage in ITU-T Recommendations of expression "demand traffic" three times, and of "traffic demand" – 50 times.

Objective of the research: To trigger discussions towards establishing stable fundaments of Overall Network Traffic Theory.

Methods used: Conceptual telecommunication network modeling, influenced by Structural Programming approach. The reasoning is illustrated with circuit switching network models, because:

1) They are relative simple;
2) We have an existing Overall Network Teletraffic Model, consisting conceptual and correspondent analytical models;
3) "…the teletraffic theory of the Internet with dimensioning methods is mainly the topic of the future." [Molnar 2006];
4) We are discussing base traffic concepts and definitions, which have to be valid in any telecommunication system.

All assumptions, notations and equations, not mentioned here, are explaneted in [Poryazov 2005] and, in more details, in [Poryazov, Saranova 2006].

2. Conceptual and reference models

2.1. Normalized structure of traffic models

In this paper three types of virtual devices are used: base, comprising base devices (enforsing group limitations on the comprised base devices, e.g. maximal sum of capacities) and aggregating base devices (used in the reasoning only).

2.1.1. Base Virtual Devices and Their Parameters

In the normalized models, used in this paper, every base virtual device, except the switches, has no more than one entrance and/or one exit. Switches, as a rule, have one entrance and two exits, but, as exception, may have more. The structural normalization is possible for every computer program [Bohm&Jacopini 1966] and therefore for every model presentable as a computer program (e.g. computer simulation model). We will use base virtual device types with names and graphic notation shown on Fig.1. For every device we propose the following notation for its parameters: Letter *F* stands for calling rate (frequency) of the flow [calls/sec.], *P* = probability for directing the calls of the external flow to the device considered, *T* = mean service time, in the device, of a served call attempt [sec.], *Y* = intensity of the device traffic [Erl].

2.1.2. The Virtual Base Device Names

In the conceptual model each virtual device has a unique name. The names of the devices are constructed according to their position in the model.

The model is partitioned into service stages (**d**ialing, **s**witching, **r**inging and **c**ommunication).

Every service stage has branches (**e**nter, **a**bandoned, **b**locked, **i**nterrupted, **n**ot available, **c**arried), correspondingly to the modeled possible cases of ends of the calls' service in the branch considered.

Every branch has two exits (**r**epeated, **t**erminated) which show what happens with the calls after they leave the telecommunication system. Users may make a new bid (repeated call attempt), or to stop attempts (terminated call attempt).

In virtual device name construction, the corresponding bold first letters of the names of stages, branches end exits above are used in the order shown below:

Virtual Device Name = <BRANCH EXIT><BRANCH><STAGE>

A parameter's name of one virtual device is a concatenation of parameters name letter and virtual device name. For example, "*Yid*" means "traffic intensity in interrupted dialing case"; "*Fid*" means "flow (call attempts') rate in interrupted dialing case"; "*Pid*" means "probability for interrupted dialing"; *Tid* = "mean duration of the interrupted dialing"; "*Frid*" = "repeated flow call attempts' rate, caused by (after) interrupted dialing". All expression "device modeling the service of **r**epeated attempts after **i**nterrupted **d**ialing" is sometimes notated with {rid}.

2.1.3. The Paths of the Call attempts

We consider call attempts generated from terminals and correspondent to content and signaling terminal traffics. In this paper, we ignore the internal network signalization.

Figure 1 shows the paths of the call attempts, generated from (and occupying) the A-terminals in the proposed network traffic model and its environment. *Fo* is the intent rate of call attempts of one idle terminal; *M* is a constant, characterizing the BPP flow of demand attempts (*dem.Fa*). In this paper we assume *M = 0.*

In the model in Fig. 1, some of the blocks are numbered. These are Reference Points, e.g. the input point of call attempts into the model is virtual switch with Reference Point 2 (RP2). Comprising devices ("**a**", "**s**" and "**b**") are notated with graphical blocks.

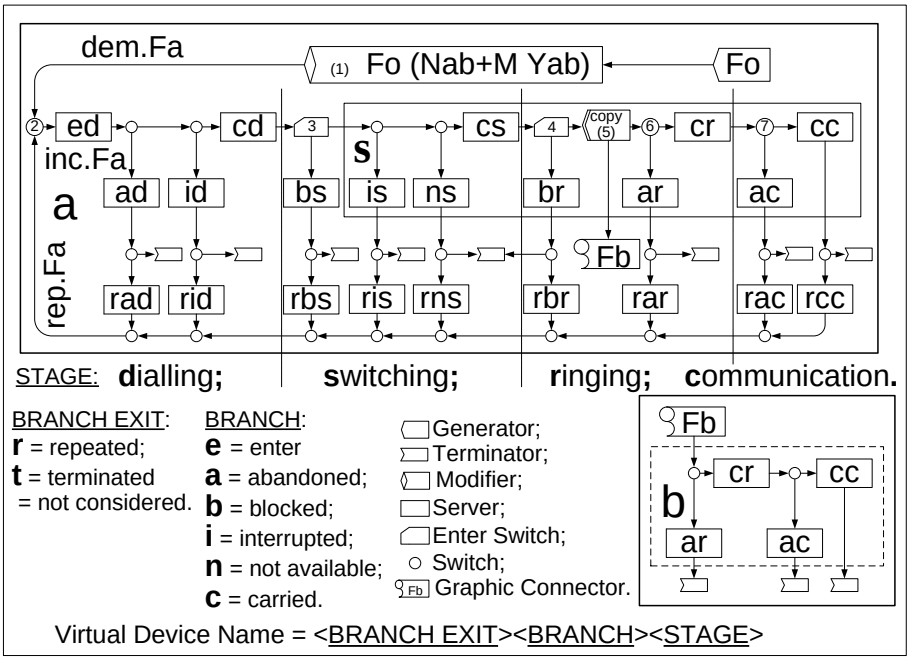


Figure 1. Conceptual model of the telecommunication system and its environment, including: the paths of the call attempts, occupying A-terminals (**a**-device), switching system (**s**-device) and B-terminals (**b**-device); base virtual device types, with their names and graphic notation. Some of switches, on the Carried Communication Branch are numbered as Reference Points.

2.2. Demand and repeated call attempts

2.2.1. The next definitions in [ITU E.600] are connected with demand traffic definition:

E.600, Definition 2.2: **call intent**: The desire to establish a connection to a user;

E.600, Definition 2.3: **call demand**: A call intent that results in a first call attempt;

E.600, Definition 2.4: **call attempt:** An attempt to achieve a connection to one or more devices attached to a telecommunications network;

E.600, Definition 2.5: **first call attempt**: The first attempt of a call demand that reaches a given point of the network;

E.600, Definition 2.6: **repeated call attempt; reattempt**: Any of the call attempts subsequent to a first call attempt related to a given call demand.

Following definitions above, we'll use the following shortenings, as definitions:

2.2.2. Definition **D2.2.1**: "**demand attempts**" for the first call attempts, from all considered call demands, that reach a given point of the network. (The calling rate of demand attempts, generating from calling (A) terminals and incoming in the Reference Point 2 into the network model presented in Fig. 1, we note with $dem.Fa$);

2.2.3. Definition **D2.2.2**: "**repeated attempts**" are call attempts subsequent to first call attempts related to all considered call demands, that reaches a given point of the network.

2.2.4. The calling rate of these repeated attempts, generating from, and occupying calling (A) terminals and incoming in the Reference Point 2 into the network model presented in Figure 1, we note with $rep.Fa$. In this notation, the rate of all, incoming in the network, attempts (Fa) is:

$$Fa = dem.Fa + rep.Fa \,. \tag{1}$$

The traffic of A-terminals, correspondent to Fa is notated with Ya .

2.3. Pie-Model concept

The Pie-Model concept is known through so called "pie-charts". In pie-models all call attempts incoming to the network, e.g. in (RP2) in Figure 1, are distributed into branches with beginning RP2 and with end – the base virtual device considered, inclusively. The all virtual devices in the branch are considered as one device, aggregating them. The name of this aggregative device is the name of the last device in the branch, following from suffix ".p", standing for "pie". For example, the branch "carried communication" has last virtual device "cc" in the normalized model. The corresponding aggregation "pie-device" is named "cc.p" and has main parameters *Pcc.p*, *Tcc.p* and *Ycc.p*. These parameters may be expressed easy by means of normalized devices, e.g. for holding time (Tcc.p) we have (normalized base devices have suffix ".n", standing for "normalized"):

$$\begin{aligned} &Pcc.p = (1-Pad.n)(1-Pid.n)(1-Pbs.n)(1-Pis.n) \\ &\qquad (1-Pns.n)(1-Pbr.n)(1-Par.n)(1-Pac.n)\,. \\ &Tcc.p = Ted.n + Tcd.n + Tcs.n + Tcr.n + Tcc.n\,. \\ &Ycc.p = Fcc.p\ Tcc.p = Fa\ Pcc.p\ Tcc.p = (dem.Fa + rep.Fa)Pcc.p\ Tcc.p. \end{aligned} \tag{2}$$

2.4. Ensue-Model concept

Let us consider the call attempts outgoing the device "carried switching" {cs.n} and incoming in RP4 in Figure 1. They have ensured continuation of their way in four possible branches. The set consisting of all base virtual devices, that may be occupied from the call attempts, after their leaving an appointed virtual device, we consider as an aggregative virtual device. The name of this aggregative device is the name of the appointed base device, following from suffix ".e", standing for "ensue"[1]). The parameters of this "ensue device" may be expressed by means of normalized devices, e.g. for ensued traffic intensity ($Ycs.e$) and ensued holding time ($Tcs.e$) we have:

$$\begin{aligned} &Ycs.e = Ybr.n + Yar.n + Ycr.n + Yac.n + Ycc.n\,. \\ &Tcs.e = Pbr.n\ Tbr.n + (1-Pbr.n)\ Tb\,, \end{aligned} \tag{3}$$

where Tb is the occupation time of the B-terminals (see Figure 1).

3. Effective traffic related definitions

Let us consider the following ITU-T definitions, copied from [ITU E.600] and commented by the author:

3.1. E.600, Definition 2.10: **fully routed call attempt; successful call attempt:** A call attempt that receives intelligible information about the state of the called user.

Comment: User's states are, for example, "present" or "absent" and they are different from the terminal's states, e.g. "not available"/"available" (in mobile networks) and "busy"/"free".

3.2. E.600, Definition 2.11: **completed call attempt; effective call attempt:** A successful call attempt that receives an answer signal.

3.3. This definition refers rather the wanted terminal (responding equipment) state, because there are "calls on which an answer signal was received, although the called subscriber did not answer" [ITU E.422].

3.4. E.600, Definition 2.12: **successful call:** A call that has reached the wanted number and allows the conversation to proceed.

3.5. E.600, Definition 5.7: **effective traffic:** The traffic corresponding only to the conversational portion of effective call attempts.

3.6. Answer signal in effective call attempts (see E.600, 2.11 Definition) don't means effective conversation, because: the user may absent (not answering condition); another user may repots that the wanted user is not near by; the quality of connection may be unacceptable, etc., so, an "abandoned conversation" case may occur. The referring to the "successful call" (see 3.4. E.600, 2.12 Definition) is more appropriate.

In view of mentioned and other discrepancies in the ITU-T definitions above, it is necessary to give new editions of the most of them:

3.7. Definition **D3.1**: **fully routed call attempt:** A call attempt that receives intelligible information about the state of the called terminal.

3.8. This means that fully routed attempts are reached RP4 in Fig. 1. In other words, they have created traffic *Ycs.p* and ensure ensued traffic *Ycs.e* with mean holding time Tcs.e (see equations (3)).

[1] ensue = happen afterwards; occur as a result [COD 99].

3.9. Definition **D3.2**: **fully routed traffic:** The portion of the traffic corresponding to the fully routed attempts, from the moment they reach the called terminal.

3.10. Definition D3.2 is in conformity with definition D3.1 and the value of the fully routed traffic is *Ycs.e*.

3.11. **D3.3**: **successful call attempt:** A fully routed call attempt that receives intelligible information about the state of the called user.

3.12. The successful attempt is reached RP6 in Fig. 1.

3.13. **D3.4**: **successful traffic:** The portion of the traffic corresponding to the successful attempts, from the moment they occupy the called terminal.

3.14. The B-terminal occupation happens in RP6. Following the reasoning in definitions D3.1 and D3.2, the value of successful traffic is $Yar.n + Ycr.n + Yac.n + Ycc.n$.

3.14. Definition **D3.5**: **effective call attempt:** A call attempt that has reached the called terminal and allows the communication with a user to proceed.

3.15. The effective attempt is reached the RP7 on Fig. 1 and a conversation (abandoned or carried) is occurring.

3.16. Definition **D3.6**: **effective traffic:** The portion of the traffic corresponding to the effective call attempts, from the moment of beginning the communication with a user.

3.17. Following the reasoning in definitions D3.1 and D3.2, the value of effective traffic is $Ycc.n$, because the abandoned communication is difficult to be accepted as "effective". This is a strict definition. Some administrations might prefer a broad definition - to include the abandoned communications in definition also, because effective traffic is known as "cost effective traffic". In this case, the effective traffic is $Ycr.e = Yac.n + Ycc.n$.

3.18. Definition D3.6 doesn't contradict to E.600, definition 5.7. It's only reformulated in order to reflect packet switching reality.

3.19. The (strict) effective attempts are moving only along the branch, corresponds to the {cc.p} (from RP2 to {cc.n}, see Figure 1) we call it "the *Carried Communication Branch*". It's parameters are: $Pcc.p, Tcc.p, Ycc.p$ (see equations (2)).

3.20. The "*ineffective attempts*" are all call attempts which are not effective.

4. Denial traffic concept

4.1. Every call attempt is generated with a will for success and it is moving in the Carried Communication Branch. In the normalized model every virtual switch, in the Carried Communication Branch, has two exits, so there are only two possibilities for a call attempt: 1) to continue its way towards {cc}; 2) to become ineffective and deflects of the effective way (to be failed in that switch point).

4.2. Definition **D4.1**: **denial traffic**: The portion of the traffic corresponding to the ineffective call attempts, created after call attempt deflection from the Carried Communication Branch.

4.3. In other words, denial traffic is served after the call attempt's failure in a point of Carried Communication Branch.

4.4. The denial traffic is real traffic, a part of ineffective traffic, corresponding to the ineffective call attempts. In the model in Figure 1, denial traffic is served in 8 devices: {ad.n}, {id.n}, {bs.n}, {is.n}, {ns.n}, {br.n}, {ar.n}, and {ac.n} (the including of {ac.n} in the list depends on the accepted effective traffic definition, see 3.17 above). The blocking is only a cause for denial traffic appearance.

4.5. The denial traffic concept is a next generalization step, following a generalization tendency in ITU-T: "End-to-end connection, party, and multi-party set-up failure, …, can occur from a lack of resources due to insufficient dimensioning or failure from other errors. End-to-end failure from a lack of resources due to insufficient dimensioning can be considered as a special case of the set-up failure probability." [ITU I.358].

5. Carried traffic concept

Let us consider a portion of the network on Figure 1, named "switching stage" (between the two vertical dotted lines). That portion of the network consists of four virtual devices: blocked switching {bs.n}, with traffic intensity $Ybs.n$; interrupted switching {is.n}, with traffic $Yis.n$; not switching (incorrect number, etc.), {ns.n} with traffic $Yns.n$; carried switching {cs.n} with traffic $Ycs.n$.

5.1. ITU concept for equivalent offered traffic [ITU E.501] is based on the carried traffic, defined in [ITU E.600]: E.600, Definition **5.1**: **traffic carried:** The traffic served by a pool of resources.

5.2. E.600, Definition 5.5 doesn't reflect the difference between the parts of the served traffic: carried and denial traffics. The distinguishing is necessary for service assessment and optimization. The ratio carried/served traffic is a good efficiency indicator.

5.3. Obviously $Yis.n$ and $Yns.n$ are denial traffics, following of attempt's termination and possible repeated attempts. These traffics are real, they load switching system and must be taken into considerations in dimensioning, but it is a little forcedly to name them "carried"[2], better leave for them the name "denial".

5.4. There is a big distinction between carried traffic ($crr.Ys$) in the cases of circuit switching and packet switching networks. In the circuit switching, the carried traffic coincides with the traffic corresponded to the carried in the switching system attempts and the denial traffic ($Ycs.n + Ycs.e$, see equations (3)). In the packet switching networks, carried packets occupy switching system for a relative short time ($Tcs.n$ and correspondent $Ycs.n$).

The presented above gives grounds for a common, for circuit and packet switching networks, definitions, with illustrations based on the system presented on Fiure 1. These definitions are in force not only for switches.

6. Target traffic related definitions

In traffic engineering, many parameters have target values, which are interpreted as design objectives, see [ITU E.726].

6.1. The usual target value of blocking probability is zero (in our example, $trg.Pbs.n = 0$).

The traffic corresponding to this target value, in ITU-T recommendations is named "offered traffic": ITU E.600

6.2. The natural generalization of the target traffics is demand traffic concept. Since nobody demands unproductive attempt's occupation (and correspondent repeated attempts), the next definition is proposing:

Definition **D6.2.1: demand traffic**: The traffic that would be carried, in the overall network, from the demand attempts, if they all are served as current effective attempts. Consequently: $Pcc.p = 1$ and $rep.Fa = 0$.

6.2.1. Following definitions D2.2.1, D2.2.2, D6.2.1 and equations (2), putting $Pcc.p = 1$ and $rep.Fa = 0$ for the demand traffic ($dem.Ya$) of A-terminals, we receive:

$$dem.Ya = dem.Fa \;\; Tcc.p \,. \tag{4}$$

6.2.2. The only difficulty, in evaluation of the demand traffic, through measurements in the real systems, is the estimation of $dem.Fa$, because it is connected with determination of repeated attempts flow.

6.2.3. Demand traffic is a dream target value for users and in the traffic management. Together with effective, carried and served traffics, it is useful for overall network performance evaluation.

6.2.4. The phrase "demand traffic" is used three times, without any definition, in the ITU-T Recommendations; "traffic demand" is used 50 times.

7. Conclusions

7.1. In addition to normalize and pie-models [Poryazov 2001], ensue model and denial traffic concept are proposed, as a parts of a technique for presentation and analysis of overall network traffic models functional structure. This allows real network traffic considered to be classified more precisely and shortly.

7.2. The ITU-T definitions for: fully routed, successful and effective attempts and effective traffic are re-formulated in order to avoid some discrepancies and to reflect packet switching reality. Definitions for fully routed traffic and successful traffic are proposed, because they are absent in the ITU-T recommendations.

7.3. A definition of demand traffic (absent in ITU-T Recommendations) is proposed. Together with effective, carried and served traffics, it is useful for overall network performance estimation.

7.4. For each definition are appointed:

1) the correspondent part of the conceptual model graphical presentation;

2) analytical equations, valid for mean values, in a stationary state.

[2] carry: 1. support or hold up, esp. while moving; 2. convey with one from one place to another; 3. have on one's person (carry a watch); 4. conduct or transmit (pipe carries water; wire carries electric current)…[COD 99]

7.5. The ITU-T definitions are needed a careful over-thinking for accuracy and completeness, because ITU is the base body for common fundamental concepts acceptation. Most of discussed in this paper terms are absent in [ANSI 2001] and ETSI definitions.

Bibliography

[ANSI 2001] ATIS Committee T1A1 Performance and Signal Processing. ATIS Telecom Glossary 2000. T1.523-2001. Approved February 28, 2001, American National Standards Institute, Inc. (http://www.its.bldrdoc.gov/projects/devglossary/)

[Bohm&Jacopini, 1966] Bohm, C., G. Jacopini. Flow diagrams, Turing machines and languages with only two formation rules. Comm. ACM, 9 (1966), pp. 366-371.

[COD 99] Concise Oxford Dictionary 9th Edition, Oxford, 1999.

[Engset 1918] Engset, T., 1918. The Probability Calculation to Determine the Number of Switches in Automatic Telephone Exchanges. English translation by Mr. Eliot Jensen, Telektronikk, juni 1991, pp 1-5, ISSN 0085-7130. (Thore Olaus Engset (1865-1943). "Die Wahrscheinlichkeitsrechnung zur Bestimmung der Wählerzahl in automatischen Fernsprechämtern", Elektrotechnische zeitschrift, 1918, Heft 31.)

[ITU E.501] ITU-T Recommendation E.501: Estimation of Traffic Offered in The Network. (26th of May 1997).

[ITU E.526] ITU-T Recommendation E.526. Dimensioning a circuit group with multi-slot bearer services and no overflow inputs. (approved: 1993).

[ITU E.600] ITU-T Recommendation E.600: Terms and Definitions of Traffic Engineering. (Melbourne, 1988; revised at Helsinki, 1993).

[ITU E.726] ITU-T Recommendation E.726. Network grade of service parameters and target values for B-ISDN (13 March 2000).

[Iversen 2006] Iversen Villy B. Teletraffic Engineering and Network Planning. Lyngby, Denmark, June 20, 2006, pp. 354. http://oldwww.com.dtu.dk/education/34340/material/telenookpdf.pdf (Access 26.11.2006).

[Molnar 2006] Molnár, Sándor. Traffic models and teletraffic dimensioning. Chapter in: "Scientific Association for Infocommunications. Telecommunication Networks and Informatics Services". On-line book, Budapest, Hungary, 03.01.2006. (http://www.hte.hu/index.php?option=com_content&task=view&id=69&Itemid=102&lang=en)

[Poryazov 2001] Poryazov, S. A., 2001. On the Two Basic Structures of the Teletraffic Models. Conference "Telecom'2001" - Varna, Bulgaria, 10-12 October 2001 – pp. 435-450).

[Poryazov 2005] Poryazov, S. A. What is Offered Traffic in a Real Telecommunication Network? 19th International Teletraffic Congress, Beijing, China, August 29- September 2, 2005, Volume 6a, Liang X.J., Xin Z. H., V.B. Iversen and Kuo G. S.(Editors), Beijing University of Posts and Telecommunications Press, pp. 707-718.

[Poryazov, Saranova 2006] S. A. Poryazov, E. T. Saranova. Some General Terminal and Network Teletraffic Equations in Virtual Circuit Switching Systems. Chapter in: A. Nejat Ince, Ercan Topuz (Editors). "Modeling and Simulation Tools for Emerging Telecommunications Networks: Needs, Trends, Challenges, Solutions", Springer Sciences+Business Media, LLC 2006, pp. 471-505. Printed in USA, Library of Congress Control Number: 2006924687.

Author's Information

Stoyan Poryazov – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Block 8, 1113 Sofia, Bulgaria, phone: (+359 2) 979 28 46; fax: (+359 2) 971 36 49, e-mail: stoyan@cc.bas.bg

TABLE OF CONTENTS OF VOLUME 2, NUMBER 2