ISSN 1934-7332 (Print)
ISSN 1934-7340 (Online)

Computer Technology and Application

Volume 2, Number 10, October 2011

David Publishing Company
www.davidpublishing.com

From Knowledge to Wisdom

Computer Technology and Application

Volume 2, Number 10, October 2011 (Serial Number 11)

David Publishing Company
www.davidpublishing.com

Publication Information:
Computer Technology and Application is published monthly in hard copy (ISSN1934-7332) and online (ISSN1934-7340) by David Publishing Company located at 1840 Industrial Drive, Suite 160, Libertyville, Illinois 60048, USA.

Aims and Scope:
Computer Technology and Application, a monthly professional academic journal, particularly emphasizes practical application of up-to-date technology in realm of computer and other relevant fields. And articles interpreting successful policies, programs or cases are also welcome.

Manuscripts and correspondence are invited for publication. You can submit your papers via web submission, or E-mail to computer@davidpublishing.com. Submission guidelines and web submission system are available at http://www.davidpublishing.com.

Editorial Office:
1840 Industrial Drive, Suite 160, Libertyville, Illinois 60048
Tel: 1-847-281-9862
Fax: 1-847-281-9855
E-mail: computer@davidpublishing.com

Abstracted / Indexed in:
Database of EBSCO, Massachusetts, USA
Chinese Database of CEPS, Airiti Inc. & OCLC
CSA Technology Research Database
Ulrich's Periodicals Directory
Summon Serials Solutions
Chinese Scientific Journals Database, VIP Corporation, Chongqing, China

Subscription Information:
Price (per year):
Print $450; Online $320; Print and Online $600

David Publishing Company
1840 Industrial Drive, Suite 160, Libertyville, Illinois 60048
Tel: 1-847-281-9862. Fax: 1-847-281-9855
E-mail: order@davidpublishing.com

David Publishing Company
www.davidpublishing.com

Computer Technology and Application

Volume 2, Number 10, October 2011 (Serial Number 11)

Contents

Computer Technology and Application 2 (2011) 757-764

Investor's Portfolio Allocation by Financial Asset Classes Using Fuzzy Logic-Based Approach in Decision Support System

Andrius Jurgutis[1, 2], Rimvydas Simutis[2] and Aušrinė Jurgutienė[3]

1. The Treasury Department, JSC Ūkio Bank, Kaunas LT-44250, Lithuania

2. The Department of Process Control, Faculty of Electrical and Control Engineering, Kaunas University of Technology, Kaunas LT-51367, Lithuania

3. The Department of Economics and International Trade, Faculty of Economics and Management, Kaunas University of Technology, Kaunas LT-51367, Lithuania

Received: July 20, 2011 / Accepted: August 02, 2011 / Published: October 25, 2011.

Abstract: The paper present the fuzzy logic expert system called MADSYS for an investor's portfolio allocation by financial asset classes. MADSYS system will be used in the interface agent (agents) of multi-agent investment management information system. One of the principal tasks of the multi-agent system is to help an investor to make investment decisions and to provide appropriate investment proposals according to the investor's profile. From MADSYS depends a lot of things, namely the multi-agent investment management information system accuracy, proposed investment decisions, the right portfolio allocation of financial assets, reliability and investor satisfaction. The usage of MADSYS system in the multi-agent system makes it more intellectual, i.e. the system will be able to adjust automatically to the changing of investor profile. The MADSYS system may be tried online at the following address: www.sprendimutechnologijos.lt/webapp.

Key words: Fuzzy logic, portfolio allocation, multi-agents system.

1. Introduction

Nowadays more and more intellectual and adaptive management technologies are penetrating into different spheres of human life and activity. Modern mechanisms and their management technologies allow to replace people completing various tasks, automate the management of various processes and so on. The same situation exists in the sphere of the provision of financial services, and especially in the spheres of finance and investments management (consulting) in which intellectual and adaptive management technologies have already found their lodgment and are penetrating deeper, whereas the number of people who trust in these technologies in the financial market is increasing every day [1-3]. In order to get more reliable, accurate and profitable investment management people invoke such elements of artificial intelligence as for example fuzzy logic, genetic programming, neural networks, software agents, etc.

Our performed works and researches are aimed to create an intellectual investment management information system. It is commonly known that investment management is a complicated and complex task coming down to the solution of some parallel and successive tasks (tasks performance), and the control

Rimvydas Simutis, professor, research fields: information technology, artificial intelligence, control systems.

Aušrinė Jurgutienė, Ph.D., MBA, research fields: risk management, portfolio management.

Corresponding author: Andrius Jurgutis, Ph.D., MBA, fix-income dealer, research fields: multi-agents system, intellectual and artificial intelligence systems in finance. E-mail: a.jurgutis@ub.lt.

and management of separate tasks [2-3]. It is our opinion that in order to create a desirable system, software agents are suitable. They compose multi-agent system which is able to perform parallel and successive investment management tasks. These are the samples of such systems: WARREN, MASST [4]. We can visually imagine a multi-agent system as an autonomous virtual investment management and consulting company in which intellectual software agents execute works more economically, quicker than specialists working in classical investment management company, i.e., software agents replace economists, statisticians, finance brokers, analysts etc. Make a long story short a multi-agent securities assessment system taking into consideration the current situation in the financial market and investor's objectives should give advice to an investor when and what security to buy/hold/sell in accordance with the investment's profile it should properly offer to diversify the portfolio, and thus successfully start managing investments.

In their publications concerning multi-agent investment management information systems the authors notice plenty advantages of these systems in different economic, technologic or other aspects [2, 5-7].

This article does not include the information particularly given in our previous publications [2, 5] concerning the advantages, structure, agents' relationship and mechanism of our multi-agent investment management information system. In this article we just briefly remind the essential details. So, during the creation of the multi-agent investment management information system we considered that the system should be universal and suitable for different investors, i.e., it should satisfy and correspond to the behavior of different investors in the market. In order the system to satisfy above mentioned expectations we construct the investment management information system using three types of agents: interface agent, information agents and cooperative agents (task agents) and considering that we can divide the investors into two main groups, i.e., conservative and aggressive investors, we have chosen two possible structures of the system (horizontal and vertical) [2], which thanks to the characteristics of agents can transform one into the other thus adopting to the investor's profile. The main attention in this article is given to fuzzy logic-based system MADSYS used in interface agent. MADSYS solves two tasks: assesses investor's tolerance towards risk (profile) [5, 8-9] and in accordance with the investor's profile offers properly diversify the portfolio into three asset classes with different risk level [10-11]. MADSYS system in interface agent is very important for the whole multi-agent investment management information system, because its appropriate operation influences qualitative work of other agents and the accuracy and reliability of multi-agent system itself, as well as the quality of the satisfaction of investor's needs.

After introducing of the prerequisites for the appearance of the intellectual multi-agent securities assessment information system in the introduction, the article further discusses only how MADSYS diversifies the portfolio. So, the second section of the article explains how the knowledge of portfolio diversification experts is imparted to MADSYS system. The third section of the article represents the structure of the fuzzy logic-based system designed for the portfolio diversification. The fourth section of the article discusses obtained system's characteristics and given systems' results.

2. The Formation of Diversification Rules

This section of the article discusses in more detail how the rules made within fuzzy logic-based system are used by MADSYS and designed for the diversification of the investment portfolio. Describing the operation of fuzzy logic-based system for the diversification of the portfolio we followed the prerequisite that the reader of this article has basic knowledge concerning fuzzy logic-based systems, i.e., knows that the construction of a fuzzy logic-based

expert system consists of four steps: the introduction of fuzziness, the formation of the rules of expert evaluation, aggregation (the summation of the conclusions of evaluation rules), the elimination of fuzziness. So, we are not going to discuss every step in details, i.e., what for and why each step is necessary. Instead we are going to discuss only essential details about each step. Besides, we are going to discuss these steps not sequentially, because we think that in such a manner the article will be more understandable and its structure will be more obvious.

Let's start with the formation of the rules of expert evaluation. Economists and financiers best of all understand portfolio diversification, so it is notable that they are co-authors of this article and the experts who help to establish diversification rules. The formation of diversification rules was fulfilled by five experts working with clients in the sphere of investment management. The model of rules formation was simple, i.e., the experts were asked to diversify the investor's portfolio on the ground of three parameters about the investor (the investor's profile is defined by three variables), i.e., to tell how much asset in percentage terms should be invested into safe investment means, how much asset should be invested into investment means of moderate risk and eventually into risky investment means. The example of safe investment means is deposit, government bonds, investments of moderate risk are corporate bonds, dividend stocks, and risky investments are company stocks, future transactions. Naturally, there can be other partition. The parameters characterizing an investor are the following: income (on the scale 0-100 hecto litas per month (for example, 10 hLt = 1000 Lt)), age (on the scale 0-100 years old) and the evaluation of the tolerance toward risk (expressed on the scale 0-100 or respectively from very low risk tolerance to very high risk tolerance). It is commonly known that in fuzzy logic numeric values of parameters should be related to linguistic variables and in such a manner the introduction of fuzziness is accomplished. We will discuss it in more details in the third section of this article, here we are going to say what fuzzy linguistic values our chosen parameters will have, because they are necessary in order the experts could express diversification rules linguistically. So, the parameter *income* can have three linguistic values: low ("low" is also the mark of the value of this parameter in fuzzy logic-based system), medium ("medium"), high ("high"). The parameter *age* can have three linguistic values: young ("young"), middle-aged ("middle-aged") and elderly ("elderly"). The parameter *risk tolerance* can have one of five values: very low risk ("R0"), low risk ("R1"), medium risk ("R2"), high risk ("R3") and very high risk ("R4"). Naturally, there can be more parameters, however it is necessary to appreciate that in this case the number of expert rules of fuzzy logic-based system will also increase, it will be more difficult to calibrate the model, calculation time will increase, economy will suffer, etc.. In the fourth section of the article we are going to discuss what one or another number of parameters or their values would determine for MADSYS system.

So, the situation and tasks for the experts have been formulated, i.e., to write down linguistically portfolio allocation decisions on the ground of three parameters and their linguistic values. It is practical to write down these decisions in the form of Table 1. In order the tables not to be massive, but easily understandable and easy to fill in we have divided the tables into three groups considering that we will diversify investments into three asset groups of different risk. So, every group of tables expresses diversification decisions according to different risk asset groups. Diversification decision itself is expressed linguistically, i.e., its value specifies how strongly the expert recommends investing into the asset of an appropriate risk group with regard to investor's parameters. The variable of diversification decision can obtain these five linguistic values: low ("Low"), less than medium ("MediumN"), medium ("Medium"), more than medium ("MediumP"), high ("High").

Table 1 The example of the table of expert's diversification decision.

Age \ Risk tolerance	R0	R1	R2	R3	R4
Young	High	High	MediumP	Medium	Medium
Middle-aged	High	High	MediumP	MediumP	Medium
Elderly	High	High	MediumP	MediumP	MediumP

I recommend investing into safe investment means when income is low and...

With a help of the table the expert fully specifies his diversification decision under respectively fixed investor's parameters and with regard to the riskiness of asset group recommends investing. Using the Table 1 and assuming that an investor is young and does not tolerate risk, diversification rule linguistically is the following: "I recommend to invest into safe investment means when the incomes are low and the age is young, and tolerance towards risk is very low High". As is known from fuzzy logic theory if we know the number of parameters and their possible values we can easily calculate the total number of tables and the number of expert rules. In our case the total number of tables is nine, and the total number of expert rules is 135.

Generally, we can write the titles (headings) of all tables as follows: "I recommend to invest into {safe, medium risk, risky} investment means, when the incomes are {low, medium, high} and ... (further tables)". When all experts gave their filled in tables we combined them all and derived nine final diversification decisions tables which included most frequent experts' decisions, thus the tables better represented the experience of all experts.

Finally, when we have the tables of diversification decisions we can easily copy out all 135 expert rules. We implemented fuzzy logic-based system designed for the diversification of portfolio with a help of "Matlab Fuzzy Toolbox", and the fragment of expert rules made with a help of this tool is shown in Table 2.

There the role of portfolio diversification experts is over. Their knowledge in the form of linguistic expert rules was transferred into fuzzy logic-based system.

Table 2 The fragment of diversification rules

The fragment of diversification rules
…
If (pajamos is zemos) and (amzius is jaunas) and (rizikos_lygis is R0) then (saugios_investicijos is High) (1)
If (pajamos is zemos) and (amzius is jaunas) and (rizikos_lygis is R1) then (saugios_investicijos is High) (1)
If (pajamos is zemos) and (amzius is jaunas) and (rizikos_lygis is R2) then (saugios_investicijos is MediumP) (1)
If (pajamos is zemos) and (amzius is jaunas) and (rizikos_lygis is R3) then (saugios_investicijos is Medium) (1)
If (pajamos is zemos) and (amzius is jaunas) and (rizikos_lygis is R4) then (saugios_investicijos is Medium) (1)
If (pajamos is zemos) and (amzius is vidutinio) and (rizikos_lygis is R0) then (saugios_investicijos is High) (1)
…

Further we are going to discuss the structure of fuzzy logic-based system designed for the diversification and member dependence functions which relate (connect) linguistic values of investor's parameters to corresponding numeric values.

3. The System Structure

The structure of fuzzy logic-based system designed for portfolio diversification is shown in Fig. 1.

The structure of fuzzy logic-based system designed for diversification consists of three smaller Mamdani type fuzzy logic-based systems as is shown in Fig. 1. Every of these systems respectively evaluates (assesses) how much to invest into asset class of appropriate risk. For the aggregation we use "centroid" function. In the last block in Fig. 1 we recalculate how to allocate in percentage terms investments through all three risk asset classes, and thus we implement the last step of the construction of fuzzy logic-based system, i.e. the elimination of fuzziness.

Fig. 2 shows the membership functions which are used in three above mentioned fuzzy logic-based systems. Membership functions connect the linguistic values of investor's parameters with an appropriate interval of numeric values and thus fuzziness is introduced.

All three fuzzy logic-based systems use identical membership functions. There were chosen trapezoidal and triangular functions, because they according in our

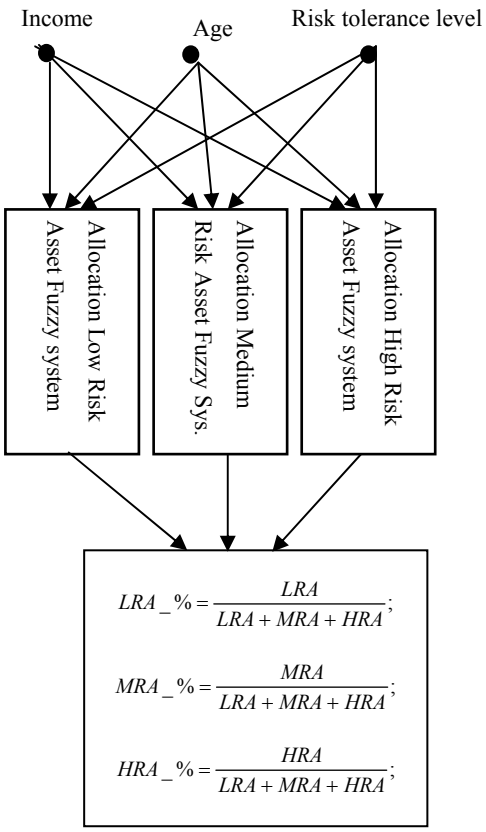


Fig. 1 The structure of fuzzy logic-based system.

opinion best of all allow relating the linguistic values of chosen investor's parameters to appropriate numeric intervals.

Here we stop discussing the steps of the construction of fuzzy logic-based system and are going to move to the discussion of results given by the system and systems' characteristics.

4. The Results and Characteristics of the System

MADSYS system's result is invested assets allocation table offered to an investor by the system. In this table an investor can see how much in percentage he should invest in each of three asset classes of different risk. The results are presented in Table 3.

Table 3 shows the results obtained under the following fixed investor's parameters: incomes are 2500 Lt, age is 31 years old and tolerance towards risk is medium. For such an investor the system offers to invest up to 33 % into safe financial means, up to 50 % into medium risk financial means and up to 17 % into risky financial means. The allocation table is very useful for investors who do not have much experience, who hesitate to make an independent decision where and how to invest. The allocation table is important for all multi-agent securities assessment system. It will let to allocate work among task and information agents

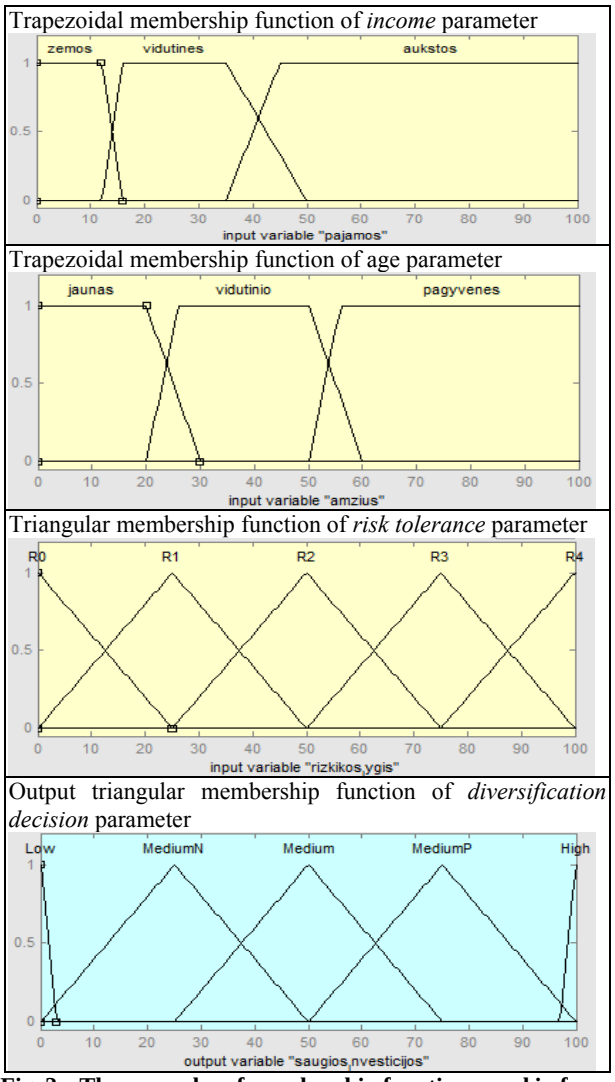


Fig. 2 The examples of membership functions used in fuzzy logic-based systems.

Table 3 The example of the allocation offered by MADSYS system.

Recommended asset allocation	
Safe and liquid investments (e.g., Deposit, government bond, bond fund)	33.33%
Medium-risk investment (e.g., Corporate bonds, SASO, dividend share, funds)	50.00%
Risky and speculative investment (e.g., Stocks, option, futures, ETF, stocks fund)	16.67%

properly. However, these questions are not the theme of this article we will analyze them in other articles.

In speaking of the characteristics of diversification system first of all we are going to analyze diversification decisions provided by the system concerning the various spectrums of fixed investor's parameters. For this purpose we need drawn surfaces which connect two investor's parameters (the third is considered to be fixed) to diversification decisions given by one of fuzzy logic-based systems (refer to Fig. 1). The examples of above mentioned surfaces are given in Figs. 3-5. In these examples investor's *income* parameter is fixed and it always corresponds to medium (average) income, to be more precise 2500 lt. Fig. 3 illustrates the surfaces of diversification decisions using "Allocation Low Risk Asset Fuzzy system" (refer to Fig. 1), Fig. 4 illustrates the surfaces of diversification decisions using "Allocation Medium Risk Asset Fuzzy system", and Fig. 5 illustrates the surfaces of diversification decisions using "Allocation High Risk Asset Fuzzy system".

Obtained surfaces are informative and it is not complex to interpret them. The surfaces are important because of two things. Firstly, they allow assessing the consistency and logic nature of the whole of expert rules and to notice possible mistakes. Secondly, decisions surfaces show that portfolio diversification implemented by fuzzy logic-based systems has many versions of decisions what allows regarding investor's characteristics (profile) individually.

Let's come back to the questions raised in the third section of the article, i.e., what does possible bigger or smaller number of linguistic values of investor's parameters determine? In our case, for example the parameter *age* has three linguistic values (young, middle aged, elderly), system's output parameter *diversification decision* can obtain one of five linguistic values (high, medium high, medium, medium low, low). After conducting of experiments with the system we noticed that bigger choice of linguistic values of investor's parameters determines

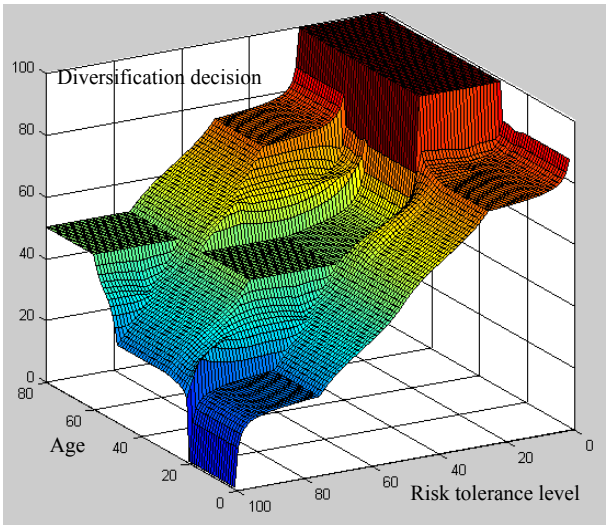


Fig. 3 The surface of diversification decisions using "Allocation Low Risk Asset Fuzzy system".

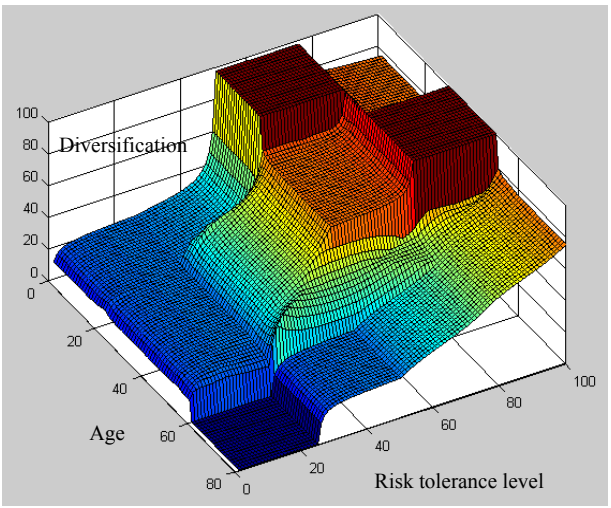


Fig. 4 The surface of diversification decisions using "Allocation Medium Risk Asset Fuzzy system".

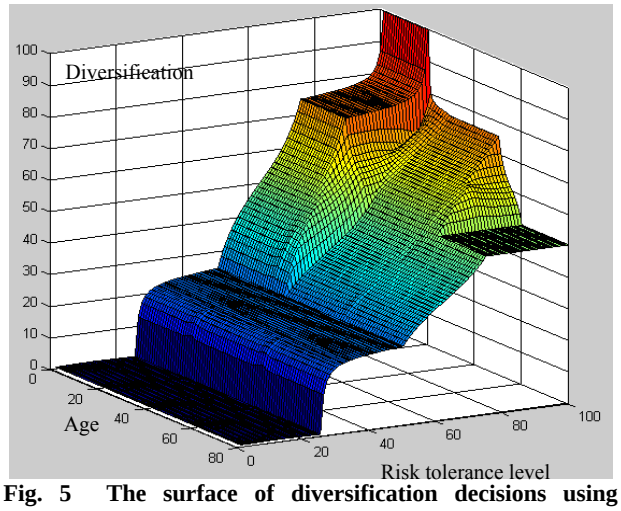


Fig. 5 The surface of diversification decisions using "Allocation High Risk Asset Fuzzy system".

smoother surface of diversification decisions. In other words, if the choice of the linguistic values of the parameters is small, diversification surface will be very "angular" and determine the fact that a small change of parameter's value will significantly change diversification decision what is not good and correct. The example of the surface of diversification decision (when using "Allocation High Risk Asset Fuzzy system" (refer to Fig. 1)), when the parameter diversification decision can obtain only three linguistic values (low, medium, high) is shown in Fig. 6. We can see that comparing Figs. 5-6 the latter is more "angular".

It is also useful to scan diagrams which show diversification decisions for the whole spectrum of tolerance towards risk if two other investor's parameters (*age* and *income*) are fixed. The example of this diagram, when fixed investor's age is 23 years old and incomes are 1300 Lt, is given in Fig. 7.

Fig. 7 shows that if tolerance towards risk amounts 30 % it is offered to invest most part about 80 % of investments into safe investment means, and the rest part into the means of medium risk. Further when tolerance towards risk increases the part of safe investment decreases and proportionally increases the part of medium and risky investments. Such diagrams are important for the calibration of the system, or rather for the connection of linguistic values with numeric values or for the selection of membership functions which was discussed in the third part of the article (refer to Fig. 2). Experiments with the system showed that if this place is not "discerned" the results of the system can be illogical and wrong.

5. Conclusions

The experiments with MADSYS system showed that we can successfully apply fuzzy logic for portfolio diversification task. One of the obtained advantages is that with a help of fuzzy logic we can expand the number of the variants of portfolio diversification and thus satisfy the needs of various investors to the maximum.

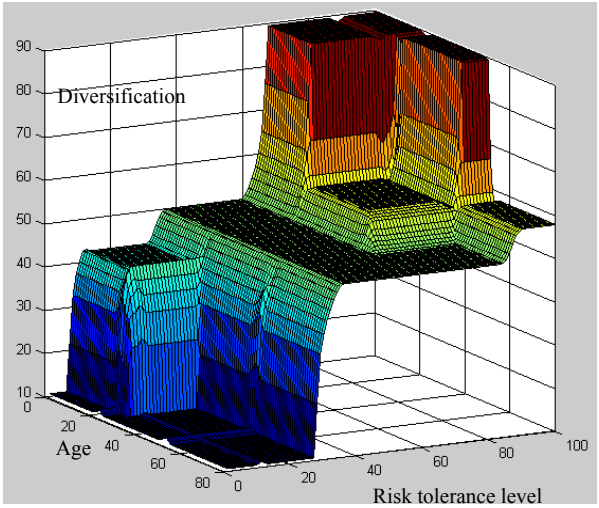


Fig. 6 The surface of diversification decisions using "Allocation High Risk Asset Fuzzy system" when the parameter diversification decision can obtain three linguistic values.

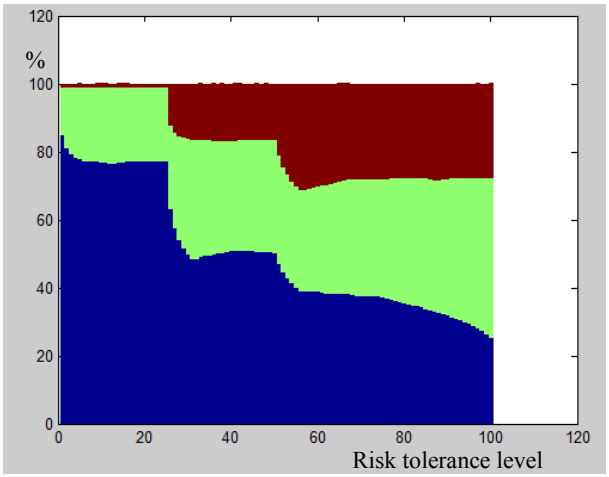


Fig. 7 Diversification decision with two fixed parameters.

For the created multi-agent securities assessment information system portfolio diversification system will allow planning tasks and the work of information agents more rationally, constrict the field of information search and calculation extent.

References

[1] G. Bojadziev, M. Bojadziev, Fuzzy Logic for Business, Finance, and Management, 2nd ed., World Scientific Publishing Co. Pte. Ltd., 2007, p. 232.

[2] A. Jurgutis, R. Simutis, Building of securities valuation IT System using multi-agents approach, in: Proceedings of the 4th International Conference on Electrical and Control Technologies, Kaunas, 2009, pp. 156-161.

[3] M. Wooldridge, An Introduction to Multi-Agent Systems,

John Wiley and Sons, UK, 2006, p. 348.

[4] M. Wooldridge, K. Sycara, N.R. Jennings, A roadmap of agent research and development, Journal of Autonomous Agents ant Multi-Agent Systems 1 (1998) 7-38.

[5] A. Jurgutis, R. Simutis, An investor risk tolerance assessment using interface agent in multi-agents decision support system, in: Proceedings of the 16th International Conference on Information and Software Technologies, Kaunas, 2010, pp. 1-8.

[6] M. Negnevitsky, Artificial Intelligence: a Guide to Intelligent Systems, 2nd ed., Addison-Wesley, 2005, p. 415.

[7] H.S. Nwana, Software Agents: an Overview, Knowledge Engineering Review, Cambridge University Press, 1996, pp. 1-40.

[8] V.J. Callan, M. Johnson, Some guidelines for financial planners in measuring and advising clients about their levels of risk tolerance, Journal of Personal Finance 1 (2002) 31-44.

[9] W.G. Droms, S.N. Strauss, Assessing risk tolerance for asset allocation, Journal of Financial Planning 16 (2003) 72-77.

[10] J. Shanteau, Expert Judgment and Financial Decision Making, in: Proceeding of the First International Stockholm Seminar on Risk Behaviour and Risk Management, 1995, pp. 16-32.

[11] S.N. Sivanandam, S. Sumathi, S.N. Deepa, Introduction to Fuzzy Logic Using MATLAB, Springer, 2007, p. 430.

Computer Technology and Application 2 (2011) 765-773

How the Information and Communications Technology Threaten Our Privacy: The Case of Email and P2P Propagation Models

Han-Wei Hsiao, Cathy S. Lin and Kun-Yu Chen

Department of Information Management, National University of Kaohsiung, Kaohsiung 811, Taiwan, China

Received: June 03, 2011 / Accepted: June 21, 2011 / Published: October 25, 2011.

Abstract: With the rapid growth of information and communication technology (ICT), the violation of information privacy has increased in recent years. The privacy concerns now re-emerge right because people perceives a threat from new ICT that are equipped with enhanced capabilities for surveillance, storage, retrieval, and diffusion of personal information. With the trend in the prevalence and the easy use of ICT, it is of necessary to pay much attention to the issue how the ICT can threaten the privacy of individuals on the Internet. While the Email and P2P (Peer-to-Peer) tools are the most popular ICT, this paper aims at understanding their respectively dissemination patterns in spreading of personal private information. To this purpose, this paper using dynamic model technique to simulate the pattern of sensitive or personal private information propagating situation. In this study, an Email propagation model and a Susceptible-Infected-Removed (SIR) model are proposed to simulate the propagation patterns of Email and P2P network respectively. Knowing their dissemination patterns would be helpful for system designers, ICT manager, corporate IT personnel, educators, policy makers, and legislators to incorporate consciousness of social and ethical information issues into the protection of information privacy.

Key words: Information privacy, propagation model, email, peer-to-peer (P2P).

1. Introduction

With the rapid growth of information and communication technology (ICT), the violation of information privacy has increased in recent years. While there is no affirmative fact to predicate if the ICT has the more mischief or beneficial to our human life, it is for sure that ICT can both benefit humans and come up the threat to the privacy of individuals.

Just as the saying "curiosity killed the cat", the concept of privacy is a fragile perception, especially "other's privacy". For example, recent years in Chinese society there are several recognized cases concerning the spreading of private pictures on the Internet, such as "Chu Mei-feng" sex scandal event in 2001, the sex scandal photos of "Edison Chan" in 2008, and "Zhang Ziyi" sexy beach photo scandal in 2009. Actually these events are quite personal private affairs, but through the ICT, these private pictures/films all have spread like "wildfire" on the Internet. Indeed, the ICT bringing much convenience in data processing, query efficiency, yet the privacy concerns now re-emerge right because people perceives a threat from new ICT that are equipped with enhanced capabilities for surveillance, storage, retrieval, and diffusion of personal information [1-5].

With the trend in the prevalence and the easy use of ICT, it is of necessary to pay much attention to the issue how the ICT can threaten the privacy of individuals on the Internet. To this purpose, this paper aims at investigating two kinds of ICT, the email and P2P tool, to see their respectively dissemination patterns in spreading of personal private information.

Corresponding author: Cathy S. Lin, assistant professor, Ph.D., research fields: information privacy, information ethics, electronic commerce, management information systems. E-mail: cathy@nuk.edu.tw.

The email has the personalized characteristic and the P2P is quite an influential and powerful ICT in sharing information, both of them are extremely pervasiveness Internet applications and almost everyone uses them. Therefore, this study proposes an email propagation model and a Susceptible-Infected-Removed (SIR) model to simulate the diffusion patterns of email and P2P network respectively. Knowing their dissemination patterns would be helpful for system designers, ICT manager, corporate IT personnel, educators, policy makers, and legislators to incorporate consciousness of social and ethical information issues into the protection of information privacy.

The paper is organized as follows: Section 2 discusses the literature reviews concerning email, P2P network, features, and classification, and Epidemic models; section 3 introduces the propagation model; section 4 is simulation results; section 5 presents results and conclusions.

2. Literature Review

2.1 E-mail

According to the "Pew Internet Project Data Memo" on January 28th, 2009, email remains the most popular online activity for Internet users. Compared to the physical post mail service, Email has the advantages of no time limit, free space constraints and zero costs sending and receiving messages among people. The content of email now covers a variety forms can be delivered, including texts, all kinds of files such as music, video, audio, and even computer viruses. Specific to the viruses, many previous studies pointed that email has become a most serious dissemination channel and causes lots of torment to Internet users [6-7]. For the reasons mentioned above, since email is right the common ICT channel for most Internet users delivering their private message and files, there is of a need exploring the diffusion pattern of computer virus via email. Thus, a new email propagation model is proposed in this study to examine this point.

2.2 P2P Network

As the Network technique development, Peer-to-peer (P2P) network has now replaced traditional file sharing tools and become a major platform spreading all kinds of files, including the unauthorized and private files sharing. The P2P software, such as Kazaa, E-Donkey, E-Mule, Morpheus, and Napster, are the hotbed of private files spreading. The privacy risks on P2P since which is quite an easy way transmitting larger media files, which might include private videos and images being molested. The private and even extremely sensitive files are shared unintentionally by Internet users using P2P networks. A HP and the University of Minnesota survey has confirmed this point that only few P2P users aware the files sharing, that means there are more files transmission without users discovered. Obviously, many privacy invasions occur without P2P users' knowledge. Previous study exhibited a mathematical model to simulate the diffusion of private files in P2P network [8]. Following the privacy concerns mentioned above, this study would propose a Susceptible-Infected-Removed (SIR) model to simulate the propagation patterns of P2P network.

2.3 P2P Network Classification

Traditionally, the structure of client-server is one of the main ideas of P2P network. This innovative idea makes every peer both as a client and a server all at once. This has the advantages of making not only a lower cost of maintaining websites but also very efficient in spreading files by separating the computing to each unit in the network. Thus, the "file index" and "file distribution" are considered as two dimensions in the P2P network. Overall, three distinct types of network can be implemented to achieve the network file-sharing: (1) Centralized index & Distributed data; (2) Centralized index & Centralized data; and (3) Distributed index & Distributed data. The first and the third types have been applied to the P2P network. Recent years, a mixed type of those two

types has proposed, for example, the E-Mule.

2.4 P2P Network Features

According to Thommes and Coates [9], P2P network has some unique features. That is, every peer in the P2P network has one file-sharing folder containing files which are publicly available for download by others on the network. When users try to download files, it began with sending out a searching query. Eventually, users will receive a list which contains the searching result that matches the criteria. The list was generated varies among the various P2P network. Then, when a file is decided to download, users would click the file on the list. The connections will be set up among peers. After that, users start downloading the file from others. The features of P2P network makes it available to download different parts of files from different peers at the same time. Finally, when parts of the file were downloaded, they would immediately be placed at the share folder in order to share with others immediately.

2.5 Epidemic Models

Ages ago, diffusion of epidemic disease has been an important issue since 1927, the first propagation model, SIR (Susceptible, Infected, Removed) was proposed by Kermack & McKendrick [10-11]. This model allowed the simulation and prediction of the diffusion of epidemic disease. With the rapid growth of information and communication networks, computers become prevalent in our daily life. In the P2P network, files are transmitted, large amount of computer viruses are also brought huge damages to Internet users. And the dissemination actions of computer viruses on the Internet is quite similar to the epidemic diseases, therefore, this study adopts epidemiological models for disease propagation to simulate P2P network. While the P2P network has the feature of time-dynamic [12], this study adopts the SIR dynamic model to simulate the propagation patterns of P2P network.

3. Propagation Models

3.1 Email Model

3.1.1 Model Description

To examine the diffusion pattern of private files through Email, a new email propagation model is proposed in this study. In the model, the variable θ represents the number of people, not to forward the message that contains privacy information. Since one person might receive the same files from more than one sender, this study considers the volume that recipient sends each time would decrease as the time increase. Hence, in this model, $D(i)$ is the degree formula so as to match the natural phenomenon of human behavior. And a probability mechanism is made to deduct the new recipient which can be received the privacy target more than twice. The above manipulation would make this model more realistic to the real world.

3.1.2 Model Equation

This model aims at investigating the volume changed of new email recipient in a given period. A general form of our model equation is as follows:

$$f(n) = [\sum_{i=1}^{n-1} D(i) f(n-i)] * [\frac{N - F(n-1)}{N}] \quad (1)$$

To make a better understating of the model, the equation would separate into two distinct parts.

$$[\sum_{i=1}^{n-1} D(i) f(n-i)] \quad (2)$$

$D(i)$ represents the volume which its decrease follows a linear recession. It can be any formula that has this identity. Here we simply just make it as θ divided by the times (i) the recipient sends; $f(n-i)$ represents new recipients increased during the period from $(n-i-1)$ to $(n-i)$. $[\sum_{i=1}^{n-1} D(i) f(n-i)]$ which means the summation of new recipients created. In other word, it's the total volume of new recipients created during the period form $(n-1)$ to n.

$$D(i) = \frac{\theta}{i} \quad (3)$$

To simplify the model, we assume that the average

number of sending out a popular private file is two, despite of the fact that it would probably be sent in many times.

$$[\sum_{i=1}^{n-1}\frac{\theta}{i}f(n-i)]=[\frac{\theta}{1}f(n-1)+\frac{\theta}{2}f(n-2)+\frac{\theta}{3}f(n-3)...+\frac{\theta}{n-1}f(n-1)] \quad (4)$$

$$[\frac{\theta}{1}f(n-1)+\frac{\theta}{2}f(n-2)] \quad (5)$$

$f(n-1)$ represents new recipients increased during the period from $(n-2)$ to $(n-1)$, while $f(n-2)$ is from $(n-3)$ to $(n-2)$.

After simplifying, our model would be like Eq. (6):

$$f(n)=[\frac{\theta}{1}f(n-1)+\frac{\theta}{2}f(n-2)]*[\frac{N-F(n-1)}{N}] \quad (6)$$

- $f(n)$: New recipients increased during the period from $n-1$ to n;
- $D(i)$: A degree formula which has the identity of linear recession;
- θ: The average amount of emails recipient sent per time;
- N: Total number of users in the network;
- $f(n-i)$: Represents new recipients increased during the period from $(n-i-1)$ to $(n-i)$;
- $F(n-1)$: Cumulative quantity of email recipients accumulate to time $n-1$.

The second part:

$$[\frac{N-F(n-1)}{N}] \quad (7)$$

One may notice that the new recipients would probably be counted repeatedly, so we set this part to deduct the number which was counted more than twice.

3.2 Peer-to-Peer Model

3.2.1 Epidemic Model of File Diffusion

Among epidemic models, SIR (Susceptible, Infected, Removed) model is the commonly used. This model categorized the population into groups depending on their states, which are susceptible, infected, and removed. The transform rate and were used to transform one state to another. The model initially intended to discuss the propagation of epidemic diseases. Yet in this paper, SIR model were used to represent the network of P2P. It's obvious that the simplicity of the model can't take fully control of the complexity of P2P network. Still, due to the techniques of building dynamic model in SIR would demonstrate the time-dynamic features in P2P network [12].

3.2.2 Model Description

The purpose of this model is to simulate the diffusion of private files in the network of P2P. Three distinctive states were recognized in the model. Each state will be corresponded respectively to the ones in the P2P network. We must draw attention to the definition of users and peers. Users were defined as the people who use P2P software to download files, and the peers were defined as the client software which was used by the users. The followings will explain a little further to the states corresponding to the three in P2P network (see Fig. 1).

- Susceptible

The category represents the one who hasn't been infected. Corresponding to P2P network, it will be regarded as the peer that hasn't the private files downloaded. Volume of the category would change by the increase of time. The number of peers in this category at time *t* is denoted by $S(t)$.

- Infected

The category represents the one who has already been infected. Corresponding to P2P network, it will be regarded as the peer that already has the file downloaded. Volume of the category would change by the increase of time. The number of peers in this category at time *t* is denoted by $I(t)$.

- Removed

The category represents the one who has already died due to the illness. Corresponding to P2P network, the it will be regarded as the peer that already has the file deleted. Volume Of the category would change by the increase of time. The number of peers in this category at time *t* is denoted by $R(t)$.

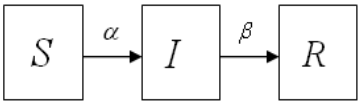


Fig. 1 SIR model.

Assuming that the total number of the peers is N in any given time, and no other states except those three above. So the equation below can be inferred:

$$N = S(t) + I(t) + R(t)$$

From the equation above, we can know that the total volume of the states is zero at any given time.

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$$

In the SIR model, states transformed from one to another by the transform rate α and β, which explains a little further in the following paragraphs:

- α: Transform rate that transforms the susceptible to infected. Corresponding to P2P network, it represents the rate that transforms the state from not having the file to file downloaded;
- β: Transform rate that transforms the infected to removed. Corresponding to P2P network, it represents the rate that transforms the state from having the file downloaded to file deleted.

3.2.3 Model Equation

- Differential equation of infected peers

When transformed from "infected" into "removed", the number of the transformed (βI) will be deducted from the volume of infected peers. Moreover, when transformed from "susceptible" into "infected", the volume of the infected peers (αSI) will increase, and it will be deducted from the susceptible peers. The equation is shown below:

$$\frac{dI}{dt} = \alpha SI - \beta I \quad (8)$$

Corresponding to P2P network, this equation represents the volume of the peers that have the private file downloaded.

- Differential equation of removed peers

When transform from "infected" into "removed", the volume of removed peers (βI) will increase, and relatively be deducted from the infected peers. The equation is shown below:

$$\frac{dR}{dt} = \beta I \quad (9)$$

Corresponding to P2P network, this equation represents the volume of the peers that have the private file deleted.

- Differential equation of susceptible peers

We can recognize from the equations above, once the susceptible peers were infected, the volume (αSI) will be deducted from the susceptible. The equation is shown below:

$$\frac{dS}{dt} = -\alpha SI \quad (10)$$

Corresponding to P2P network, this equation represents the volume of the peers that do not have the private file downloaded.

4. Simulation

4.1 Email

Notion θ:

In the email model, the notion θ was given two different values which are respectively 2 and 4 to observe the diffusion around the network. Assume that only one person has the private files at the very beginning, and the total number of recipient in the network is N = 300,000. Considering one day as a unit 30 days in total. The results of the simulation are shown below.

Result and Discussion:

Notice that once the $f(n)$ reached their maximum; it'll start to go down. That's because the latter the diffusion, the popularity contains more recipients, so it must be deducted from $f(n)$.

As seen in Fig. 2, when the value of θ gets higher (which represents the average amount of email recipient sent per time is larger). The number of new recipient $f(n)$ per day is significant different. When θ=4, the maximum of new recipient is 250,000 and it takes 10 days to complete the files diffusion throughout the network. Yet when θ = 2, the maximum of $f(n)$ is 90,000, which is smaller than the previous one and it takes 15 days to complete the file dissemination.

4.2 Peer-to-Peer Network

Notion α and β:

In this part, we'll discuss two notions respectively

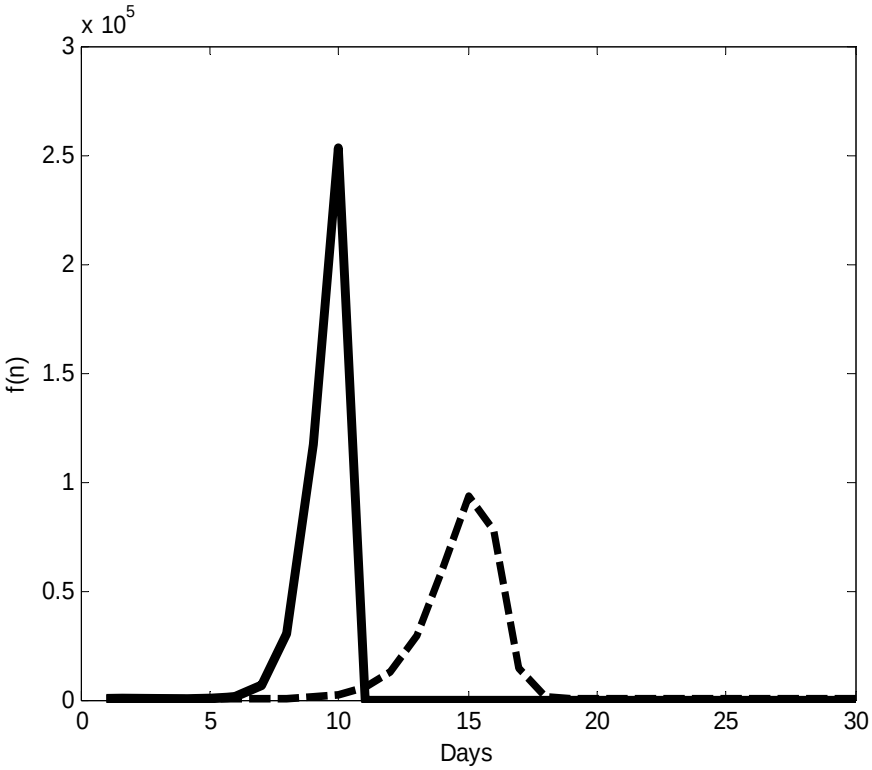


Fig. 2 Dotted line θ = 2; solid line θ = 4.

while the other remains unchanged. Our concern in this paper is the diffusion of private files, so the volume of the state "Infected" will be focused. As mentioned in the previous parts of this paper, the state "Infected" represents the peers already have the private files downloaded. Here we set the total number of peers in the P2P network is N = 300,000. We consider one day as a unit, and there are 30 days in total (as shown in Fig. 3). And only one peer has the private files at the very beginning.

- Exploration in α while β remains unchanged

Under such scenario, we find out that when α gets larger, the volume of the peers will grow faster and the differences between downward slopes are small. The speed of growth was affected by α, but the influence is relatively small to the speed of decline.

- Exploration in β while α remains unchanged

In Fig. 4, the results show that when β gets larger, the volume of the peers will decline faster and the differences between upward slopes are small. The speed of decline was affected by β, yet the influence is relatively small to the speed of growth.

4.3 Email & P2P

Notions:

Due to the convenience of email and the innovative techniques on file-sharing of P2P network, this study considers two kinds of ICTs to observe their respectively dissemination patterns. Mathematical models were developed to simulate the diffusion of private files. The basic assumptions for these two models are as follows: (1) Units (people or peers) in each network is N= 15,000 in total; (2) Only one peer or person has the private files at the very beginning; (3) We consider one day as a unit, and there are 30 days in total; in the email model, θ the average amount of email recipient sent per time is five. And the notions α and β in the P2P model were based on, which α = 5.7870e-005 (corresponds to five download per day) and β =1.1574e-005 (the average time for a peer to delete a private file is 24 hours). Fig. 5 shows the result of the simulation.

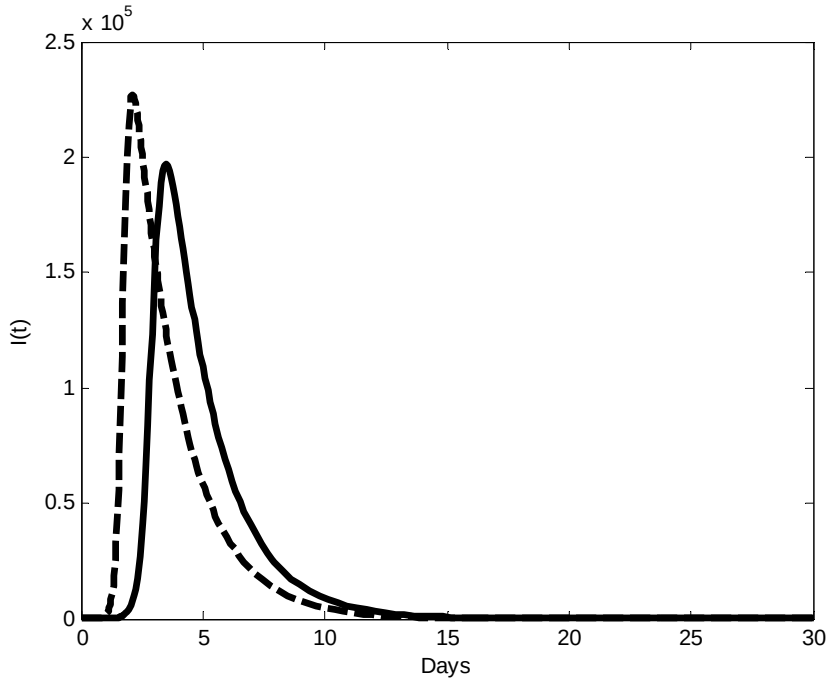


Fig. 3 Solid line: α = 0.000016; dotted line: α = 0.000026.

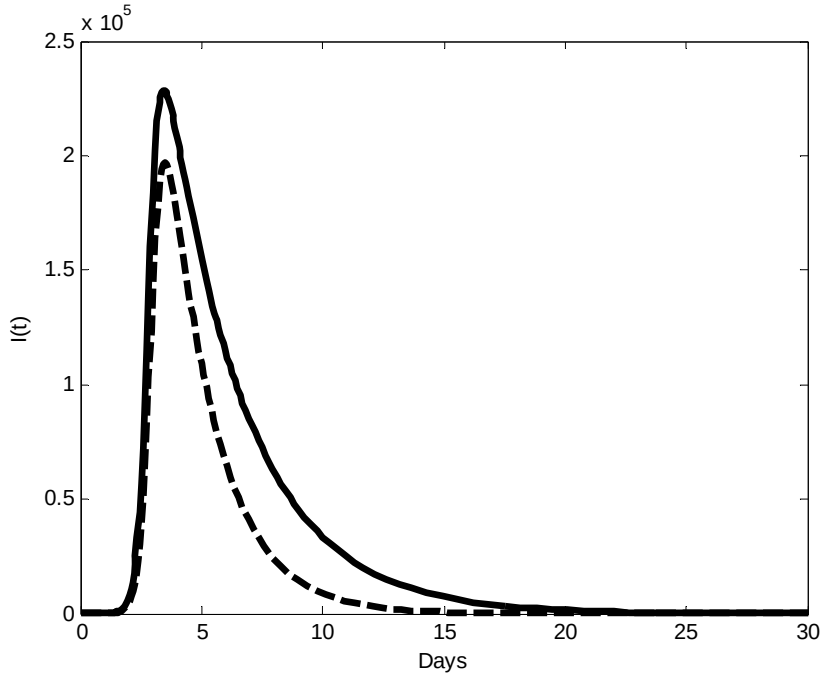


Fig. 4 Solid line: β = 0.3058; dotted line: β = 0.5058.

Result and Discussion:

It is worthy notice that the two lines have slightly different (see Fig. 5). The solid line represents the volume of the peers that own the file while the dotted line represents the cumulative quantity of Email recipients. The simulation results obviously exhibit that the P2P network has a dominant propagation power (in two-day) in dissemination files on the Internet than Email (in nine-day). After that, all Email users and P2P users in their respectively network have already received and/or downloaded the files.

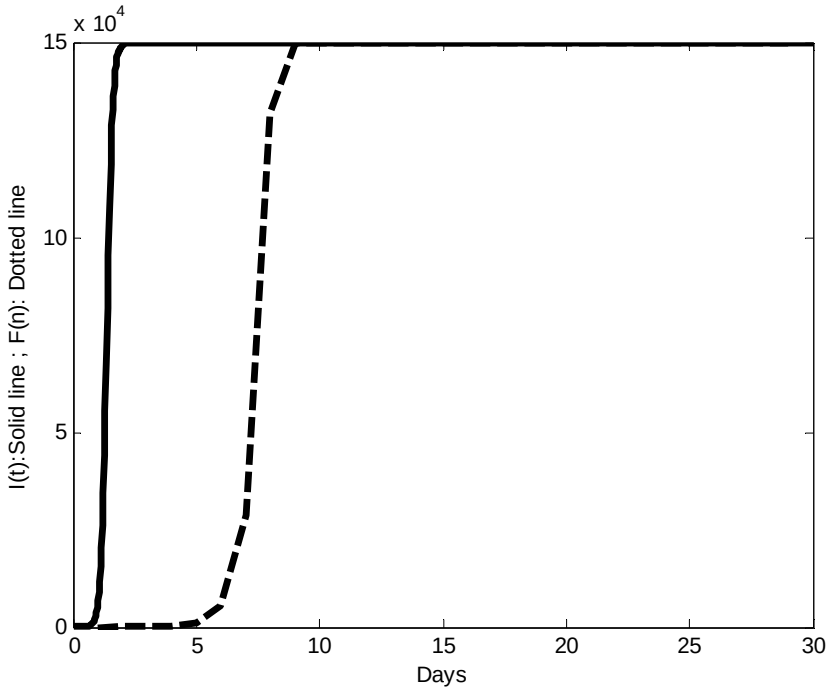


Fig. 5 Solid line: P2P model; dotted line: E-mail model.

5. Results and Conclusions

While the ICT bringing much convenience in the information age, yet the privacy concerns now re-emerge right because people perceives a threat from new ICT that are equipped with enhanced capabilities for diffusion of personal information. But, how soon and how different the ICTs' dissemination patterns? This remains a doubt. The findings of this study demonstrate that the propagation speeds by using Email or P2P are rapidly. Specifically, the speed of spreading privacy information is far faster in P2P network than in Email. And most important, as the increasing of θ or α, the diffusion spreading speed will be enhanced, this finding illustrates that the spread speeds is dependent on the interesting of privacy targets. That is, the dissemination of privacy targets via ICTs, the different interesting level will result in dissimilar spread speeds.

Further, to the different kinds of ICTs, Email and P2P network, just as the simulation results in this study, P2P network has a dominant propagation power (in two-day) in dissemination files on the Internet than Email (in nine-day). This somehow demonstrates that P2P software has a much more enormous impact when dispersed privacy information.

The current study has responded the doubt concerning the ICTs actual impact to information privacy. In the future, a collection of real dataset is needed to empirical validate the two distinct model of Email and P2P network, to apply the models and see their respectively prediction power. Overall, the proposed models in this study have their academic meanings to make a further understanding in these two model and parameters considered in these two models. Also, the practical implications would be that by knowing the propagation patterns of email and P2P network would provide the practicians such as policy makers, legislator, system designers, corporate IT personnel and educators to incorporate consciousness of social and ethical information issues into the protection of information privacy.

Acknowledgment

This study was financially supported by the Research Grant NSC 97-2410-H-390-017 from

Taiwan's National Science Council.

References

[1] R.A. Clarke, Information technology and dataveillance, Communications of the ACM 31 (5) (1988) 498-512.

[2] F.Y. Kuo, C.S. Lin, M.H. Hsu, Assessing gender differences in computer professionals' self-regulatory efficacy concerning information privacy practices, Journal of Business Ethics 73 (2) (2007) 145-160.

[3] R.O. Mason, Four ethical issues of the information age, MIS Quarterly 10 (1) (1986) 5-12.

[4] A. Miller, The Assault on Privacy: Computers, Data Banks and Dossiers, University of Michigan Press, Ann Arbor, MI, 1975.

[5] A.F. Westin, Privacy and Freedom, Atheneum Publishers, New York, 1967.

[6] C. Jin, J. Liu, et al., A novel email virus propagation model, in: Proceedings of the 2008 Workshop on Power Electronics and Intelligent Transportation System, IEEE Computer Society, Washington, 2008, pp. 56-60.

[7] C.C. Zou, D. Towsley, et al., Email virus propagation modeling and analysis, Technical Report: TR-CSE-03-04, University of Massachusetts, Amherst, 2003.

[8] M. Parameswaran, A. Susarla, et al., P2P networking: an information-sharing alternative, Computer 34 (7) (2001) 31-38.

[9] R.W. Thommes, M.J. Coates, Modeling virus propagation in peer-to-peer networks, in: 2005 Fifth International Conference on Information, Communications and Signal Processing, Bangkok, 2005.

[10] W.O. Kermack, A.G. McKendrick, Contributions to the mathematical theory of epidemics, II. The problem of endemicity, in: Proceedings of the Royal Society of London, Series A 138, 1932, pp. 55-83.

[11] W.O. Kermack, A. G. McKendrick, Contributions to the mathematical theory of epidemics, III. Further studies of the problem of endemicity, in: Proceedings of the Royal Society of London, Series A 141, 1933, pp. 94-122.

[12] K. Leibnitz, T. Hoßfeld, et al., Modeling of epidemic diffusion in peer-to-peer file-sharing networks, in: The Second International Workshop on Biologically Inspired Approaches to Advanced Information Technology, 2006, pp. 322-329.

Computer Technology and Application 2 (2011) 774-781

A Multi-Agent Expert System for Prevention of Child Abuse and Neglect

Pedro Sanz-Angulo and Juan José de-Benito-Martín

Department of Business Organization, Marketing and Market Research, Industrial Engineering School, University of Valladolid, Valladolid 47011, Spain

Received: August 31, 2011 / Accepted: September 09, 2011 / Published: October 25, 2011.

Abstract: When a child abuse situation arises, the responsible agencies and entities in charge of response should be capable of providing a fast and personalized solution for the good of the child. This need leads the authors to consider the formation of dynamic virtual organizations tailored to each particular abuse case. In the authors' approach the partner selection of these collaborative networks is done through a software tool that combines two technologies from the field of artificial intelligence, specifically multi-agent systems and expert systems. In addition, these partners come from the breeding environment constituted by all the agencies or individuals, either in a region or locality, which have the potential of response.

Key words: Multi-agent systems, expert systems, prevention of child abuse and neglect, dynamic virtual organizations, VO (virtual organization) breeding environments.

1. Introduction

Child abuse is a universal phenomenon that we must strive to eradicate from our society. This goal, undoubtedly ambitious, requires an intensive work in the education of every individual and, of course, in the recognition of the children's rights.

Although in recent years significant progress has been made in this sense, our efforts must continue as long as abused children exist. We have the responsibility to protect them, using the different resources available. In this sense, we have to search for new solutions and tools that allow us responding promptly and in a personalized way.

To provide this response is necessary to promote a networking model that involves all stakeholders (both entities and people) in the prevention and the solution of child abuse. Every stakeholder must contribute to the collaboration by providing its core business, i.e., what it does best.

However, this network collaboration must face several challenges to become a truly effective instrument to combat the problem of child abuse. Pablo Herrero [1] identifies some of these problems: excessive specialization of the social services, urgency in the response, the pressure, lack of diagnosis, pseudo-agreements, the triangulations and the competition between services, etc.

On the other hand, every case is unique (it is not the same to deal with a sexual abuse case than a peer victimization intervention), so that networks must involve different components in each new situation, and even these components can vary during the performance. It is necessary, therefore, a dynamic and flexible network that allows the continuing evolution of resources and services, as well as the incorporation of new ones.

Juan José de-Benito-Martín, professor, Ph.D., research fields: multi-agent technology, simulation, production, virtual organizations, logistic, etc.

Corresponding author: Pedro Sanz-Angulo, assistant professor, Ph.D., research fields: multi-agent technology and expert systems, production, virtual organizations, logistic and innovation. E-mail: psangulo@eis.uva.es.

These inter-collaborative problems in child abuse domain are difficult to solve without considering the new organizational models and the Information and Communication Technologies (ICT). The ICTs are a key element in order to promote efficient and innovative networking solutions designed to prevent child abuse in all its variants.

In this sense, our work is focused on the creation of Dynamic Virtual Organizations (DVO) tailored to the demands of each abuse situation. The selection of the members is done through a multi-agent expert software tool. It is an approach that solves, or at least reduces, most of these difficulties in a simple and intuitive way.

We also consider that any initiative in this area has to bear in mind the Virtual Organization Breeding Environments (VBE) concept. In child abuse domain these long-term networks are going to be constituted by entities from the different fields (both educational, social, police, judicial and health) involved in Child Abuse Prevention (CAP) in their secondary and tertiary levels.

The use of software agents whose behavior is guided by an expert system is a novelty and an innovation in the field of networking methodology aimed at preventing child abuse. This combination allows optimizing the use of resources and simplifying communication between institutions and professionals, which makes possible a more agile response of the DVO created ad hoc abuse problem.

With our work we try to ensure that networking is not going to be reduced to a simple coordination of institutions as if they were watertight compartment. We, ultimately, are trying to get the best answers to those questions always present in a case of child abuse. It might be mentioned that the networking concept becomes tangible and acquire a dynamic nature according to the needs and requirements of each individual case of abuse.

The paper is organized as follows: Section 2 discusses the child abuse problem and the importance of networking in the response to this problem; section 3 introduces the DVO and VBE concepts and justifies their use for the child abuse problem; section 4 presents the VCAP platform and the IA technologies that we are using in its development; finally, section 5 gives some conclusions and presents future work.

2. Child Abuse and Networking

The United Nations Convention on the Rights of the Child 1989 defines child abuse as [2] “all forms of physical or mental violence, injury or abuse, neglect or negligent treatment, maltreatment or exploitation, including sexual abuse, while in the care of parent(s), legal guardian(s) or any other person who has the care of the child”. In short, we can say that there is child abuse if children’s rights are not respected and there is not a response to their needs [3]; we cannot forget that abuse is, primarily, the lack of good treatment.

But the lack of good treatment can take several forms (physical, sexual, psychological, neglect, etc.), which come in a great variety of types depending on the situation, the involved actors, the degree of intensity, and so on. Each abuse situation is unique and requires, of course, a different approach; it is necessary to define and design an appropriate response to each particular case. In this sense, prevention is the best possible approximation to the problem of child abuse.

Prevention may take place at three different levels[1]: Primary, secondary and tertiary.

- Primary prevention aims to decrease the incidence or onset of abuse cases, through child advocacy and special attention to their needs. It applies to the general population, in order to the community becomes aware about the problem and acquires positive habits and behaviors that prevent the appearance of the child abuse.
- Secondary prevention focuses, for its part, on the concept of risk. This prevention is addressed to social groups, families or individuals classified as “high risk” to avoid certain situations that end up in abuse.

[1] As we shall see below, our work can be perfectly included in the latter two levels of prevention, and especially in the tertiary one.

• Finally, tertiary prevention tries to reduce the duration and severity of the consequences of the problem (stage of rehabilitation or cure). It also requires the intervention in the context, family, etc., to prevent its recurrence.

But to achieve an effective prevention, it is necessary first to promote a networking model that includes all the stakeholders required to take part in the response and solution of child abuse problems. Networking is the way of achieving a collaborative workspace where to reach the pursued goals. It is based on the communication of the agents and institutions, among which there must be an optimum exchange of information, both qualitative and quantitative.

The originators of the sociotherapeutic networking were Speck and Attneave [4]. They created therapeutic teams for intervening in families in crisis, in the United States, in order to break destructive patterns of family relationships and provide support for alternative options. Several years later, in the eighties, Elkaim [5] carried out the first practice at European level in deprived areas of Belgium. Since then, the networking experiences have evolved and multiplied.

In our country, Spain, we find the networking pilot program of Burlada (Pamplona). Its promoters recognize that one of the challenges of any network program is to provide a process of organizing the different institutional levels and professional resources, in order to ensure the creativity and competence of each one of these instances [6].

These and other works, experiences and projects emphasize the need of networking as a mean to achieve the ultimate goal of prevention and resolution of child abuse, establishing guidelines and procedures that normally are constructed ad hoc to the context in which they are developing. They do not possess, in addition, software tools capable of solving the different problems of such kind of relationships, among which the need for a rapid intervention highlights.

Our job serves, precisely, to overcome this deficiency. We pursue the interconnection, in a quick and efficient way, of those professionals and institutions that must be part of the answer. Starting from a simple notification in the system, a dynamic virtual organization is configured in order to respond to the specific abuse case in a personalized way and with little human intervention.

3. The Basis of Our Work

There are several similarities between networking that a child abuse situation requires and the networking of the dynamic virtual organization. Below, we will summarize the more interesting ideas to understand our work, although previously it is necessary to describe the Dynamic Virtual Organization and VO Breeding Environments concepts.

3.1 The DVO in the VBE Context

The Dynamic Virtual Organizations are rapid creation and fast dissolution organizations [7], constructed ad hoc to the opportunity of collaboration. In a DVO a set of business partners come together dynamically, on demand, and in accordance with the requirements and needs [8] of the problem, disappearing when those needs have been met.

However, the creation of a DVO whenever a new collaboration opportunity emerges requires large amounts of time and resources, reducing the agility associated with this business model. The effectiveness of the process depends largely on the availability of adequate information about potential partners, their level of preparedness to engage in a DVO, the existence of trust, etc.

To solve some of these problems, different authors have begun to consider that the formation of such organizations has to take place in the context of what Afsarmanesh and Camarinha-Matos [9] have called Virtual Organization Breeding Environments. The VBEs are clubs of organizations prepared to work in long term relations and from which temporary coalitions emerge able to respond dynamically to the different business opportunities [10].

The VBE simplifies the configuration and establishment of DVOs since it solves, or reduces, many of the obstacles associated with the temporality of such organizations, contributing, thus, to their creation in a more efficient way. Partners of the DVO can, this way, benefit fully from unexpected changes in their context, providing an agile response to the problems or the opportunities.

3.2 Equivalence between DVO and Networking in Child Abuse Domain

We can start, for example, talking about the VO breeding environments. In the case of child abuse domain it is obvious that the VBE would be constituted by any entity or person involved in the solution: schools, social services, health centers, youth services, law enforcement agencies, state security bodies, and so on.

Every one of them brings to the partnership its core business, what they do best, so the VBE contains all necessary processes that the abuse situation requires. In addition, these entities are required to share certain culture of work (action protocols…), ICT infrastructure, etc., to achieve the highest possible performance.

After identifying a new case of abuse (either in the school, or in a health center, or through a complaint at the police station or in the child line, etc.) it is time to use the breeding environment and select the set of institutions and/or persons that are the best prepared to respond to this particular problem.

The union of the different entities that have been selected configures, ultimately, the dynamic virtual organization responsible for providing the personalized answer to the initial child abuse situation. Naturally, different virtual organizations can arise from the breeding environment to provide this response, although the selection of the best one depends on the requirements and needs of the particular situation of abuse and on the features and capacities of the partners.

After the selection of the components of the DVO, partners need to coordinate their actions, they have to work together in order to find an efficient and agile solution to the problem: In other words, operation phase starts. This phase lasts until an adequate response to initial problem has been achieved. This moment coincides, precisely, with the dissolution phase of the dynamic VO.

On the other hand, we must also notice that the entities of a particular DVO can be partners of another consortium which tries to response to a different abuse case. In addition, whenever it is necessary in the DVO performance, more partners can be searched, or even their number can be reduced, and so on, according to the changes of each abuse case. In short, during the operational of the DVO a reconfiguration phase can take place.

4. The VCAP Multi-Agent Expert Platform

Our fundamental objective is to enhance and develop a networking model through the creation of DVOs that are capable of facing the child abuse problems in an agile and efficient way. It is therefore necessary to align the concepts of VBE and DVO to the child abuse domain and provide efficient tools that facilitate both its formation as it's operational.

In order to provide an efficient solution to the formation problem, this paper proposes an approach based on the Multi-Agent System (MAS) paradigm. This technology provides several advantages over other methods previously proposed, especially for its ability to manage complex and distributed problems. But in addition, agents in our MAS have an intelligent behavior through a rule-based Expert System (ES) implemented in their decision module. Finally, the identification of intelligent agents is made through an innovative model based on the VBE concept.

With this new approach the best combination of partners is defined dynamically in real time based on system state, the features of the problem and the predefined objectives of the stakeholders. The identification and assignment of these stakeholders to the child abuse problems emerges from the interaction between the agents, through different negotiation

mechanisms. In this sense, we can introduce the Virtual Child Abuse Prevention (VCAP) concept.

4.1 MAS and ES

In the multi-agent systems the core element is the software agent, a computer entity with some defined targets and able to act flexibly in its environment. These agents are typically autonomous, reactive, pro-active and have social skills, that allow them to interact with each other and their environment by exchanging messages [11].

The characteristic of sociability makes agents a very interesting concept to develop distributed systems where several agents interact each other and with the environment to create multi-agent systems. Each of these agents has the ability to solve a specific part of the problem, so that together they can achieve the desired functionality and operate asynchronously without a central control system.

Moreover, to give an effective response, software agents should behave, think and act like a human expert in the domain would do, but how can we do this? For this purpose we have considered the expert systems. By definition, an expert system is a computer program that simulates the thought process (learning, memorization, reasoning, communication, etc.) of a human expert to solve complex decision problems in a specific domain. By this way, ES can [12] store data and knowledge, draw logical conclusions, make decisions, communicate with human experts, explain their decisions, etc., and, as a consequence of all above, take actions.

Thanks to these systems, and in particular to a rule-based ES, we obtain agents with a rational behavior based on the knowledge and expertise of human experts in the child abuse area. Thus, without resorting to complex mathematical procedures, the agents are able to make the best decisions based on the needs of the problem and on the structured knowledge that has been supplied to them.

The introduction of the expert system is a clear innovation, from the technological point of view, over other existing multi-agent solutions.

4.2 DVEBreeder Platform

In the University of Valladolid, we have developed the DVEBreeder multi-agent expert platform, which uses the VBE concept to build DVEs [13-14] effectively. In this platform the identification of intelligent agents is made through an innovative model [15] and within the VBE context. After identifying the agents capable of performing the required tasks, these negotiate with one another based on the knowledge embodied in its decision modules, resulting in the formation of a DVE.

This platform has been built with various interesting features, such as modularity and flexibility. Thanks to this, we can easily adapt the software tool to the Child Abuse Prevention problem, given the parallels between this problem and the formation of Dynamic VE. Ultimately, the goal pursued in the CAP system (DVE) is none other than finding the best combination of entities and professionals (enterprises) able to respond to the requests of a child abuse case (business opportunity) in an agile and efficient way.

In this system, the breeding environment is composed of all of entities and professionals from a region o locality. Obviously, given the set of features and capabilities of the VBE members, different combinations could arise from this environment in response to a specific case. However, the choice of the most appropriate configuration will depend on many factors among which are: the own case, the capabilities and properties of the breeding environments members, the past experiences, the priority criteria, etc.

Taking into account all of these factors and through the interaction of the different agents the more adequate decision sequence emerges, leading in turn to a DVO. Once the child abuse problem is solved, this configuration will no longer make sense and will disappear, releasing entities and professionals to set up new DVOs. This new approach allows us face the CAP problem from a virtual perspective, giving us the

opportunity to introduce the concept of Virtual CAP system.

4.3 Analysis of the VCAP Multi-Agent Expert System

The analysis of the platform has been made following the steps defined by Nikraz, Caire and Bahri [16]. Thus, we have identified the different types of agents and users in the system, along with their responsibilities and the acquaintance relationships that exist between them. The result is the agent diagram shown in Fig. 1.

Taking into account the agent container structure of JADE (Java Agent Development Environment), there is a main container on the manager server (see Fig. 2) in which the VCAP Management Agent resides, in addition to the Information Provider Agent, the Yellow Pages Agent (where the Resource Agents register their services) and the rest of agents for the management of the Jade platform.

Also there are multiple remote containers: specifically, so many remote containers as institutions and professionals the breeding environment has. In these containers we can identify the VBE Member Agent and an Access Agent for security issues.

We have considered that all of these entities can provide their core business and, at the same time, be the initiators of the DVO formation process. It is also possible that a third entity or person starts the process, but for now we just consider that the system is open only to the VBE members. Perhaps in the future the platform can be open to everyone, but previously it is necessary to solve different security problems.

The union of these remote containers with the main container comes into a distributed agent platform, as can be seen in Fig. 2.

4.4 VCAP Platform

As we have mentioned, the current development of the VCAP platform is possible thanks to our experience in the creation of the DVEBreeder application. The VCAP platform, like its predecessor, is a software tool

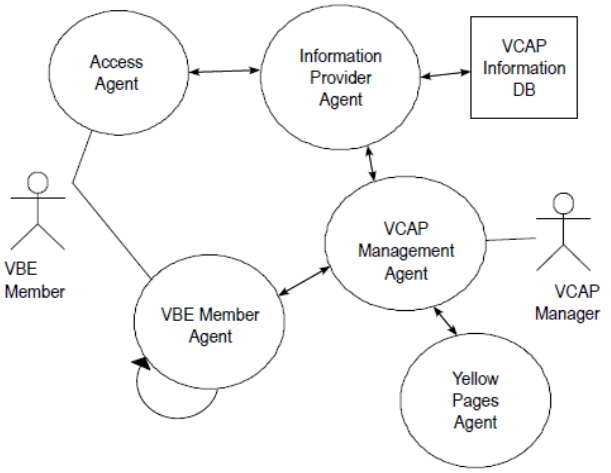


Fig. 1 Agent diagram of the VCAP platform.

that combines multi-agent technology with expert systems: this combination provides us with multiple interesting features.

On one hand, communication between agents is trivial thanks to the platform used in the MAS development, which is none other than Jade. On the other hand, agents' behavior is quite similar to the behaviour human experts can have thanks to the JESS (Java Expert System Shell) expert system, which allows to express the behavior of the agents as a set of facts and rules.

Furthermore, it is flexible, dynamic, scalable and, as we have seen, innovative both from technological, organizational and conceptual standpoints (for its full compliance with the initial model). Agents are autonomous, behaving intelligently and communicating asynchronously. FIPA standards are met, the interface is intuitive, the code reuse is simple thanks to its open structure, etc.

The result is an innovative tool able to solve the CAP problem in an agile and efficient way. Fig. 3 shows a simplified scheme of the platform VCAP in the considered application domain.

5. Conclusions

The rapid development of the Information and Communication Technologies has radically changed the way the problems are being addressed in many areas of research and application. In turn, it has allowed

Fig. 2 Distributed structure of the VCAP platform.

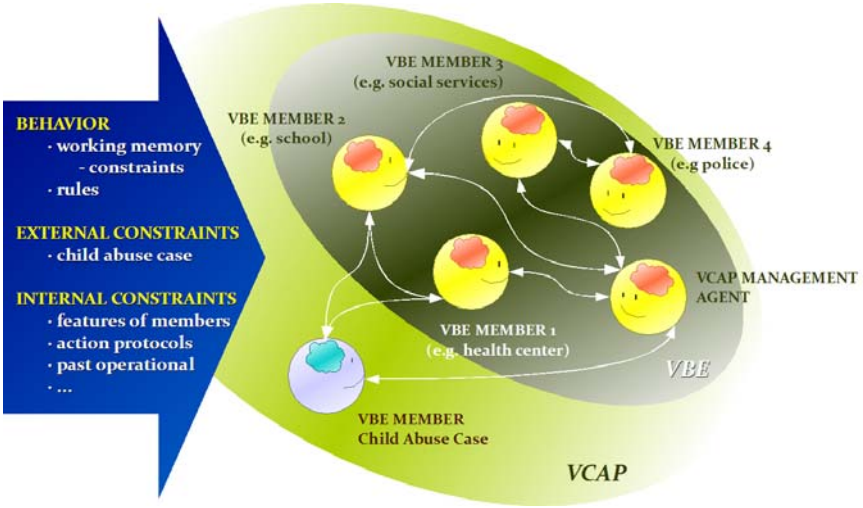


Fig. 3 Simplified scheme of the VCAP platform.

developing an incomparable framework for the creation of new organizational forms with innovative and interesting possibilities, like the DVO paradigm.

This new architecture is characterized by an intensive use of technology and by agility in responding to any new need or change in the environment. This agility is a key element to resolve a problem as complex in contemporary society as the child abuse.

Based on this, the article has presented a multi-agent expert solution in which DVOs emerge in the context of VBEs. For its development, we count with the experience in the implementation of the DVEBreeder for the DVE creation.

Modularity and flexibility of this platform allow us an easy adaptation to any application domain (for example, the installation of solar photovoltaics, the shop floor control problem, the prevention of child abuse cases, etc.). These qualities are certainly a competitive advantage over other similar tools that

exist in the literature, e.g., Refs. [8, 17-18], which are usually built ad hoc the problem they try to solve.

After doing the planning and analysis, we are currently working in design and implementation phases. To do this, we need to define some elements of the platform as e.g., the application domain ontology, service descriptions, agent decision modules, etc. Others, such as the interactions, message templates, interaction protocols or the ontologies for the management of the platform, are very similar to those established in the DVEBreeder tool, which will greatly facilitate the final implementation of the platform.

Acknowledgments

The authors would like to express our gratitude to the managers and members of REA, the Association for the Children and Youth Advocacy of Castile and León, and in particular to Ms Mª Elena Villa Ceinos.

References

[1] P. Herrero, La red inteligente, Zerbitzuan: Gizarte zerbitzuetarako aldizkaria, Revista de Servicios Sociales 43 (2008) 65-72.

[2] United Nations—Office of the High Commissioner of Human Rights, Convention on the Rights of the Child, resolution 44/25 of 20 November 1989, available online at: http://www2.ohchr.org/english/law/crc.htm.

[3] P. Sanz, M.E. Villa, M.A. Manso, J.J.D. Benito, Aplicación de la Tecnología Multiagente y los Sistemas Expertos al trabajo en Red para la Prevención del Maltrato Infantil, IX Congreso Estatal de Infancia Maltratada, Valladolid, España, Noviembre de 2008.

[4] R. Speck, C. Attneave, Redes Familiares, Amorrortu Editores, Buenos Aires, 1974.

[5] M. Elkaïm, Las prácticas de la terapia de red, Barcelona, Gedisa, 1989.

[6] M.D. Miguel, M. Fernández, Detección precoz del maltrato infantil, Programa piloto de trabajo en red, Anales del Sistema Sanitario de Navarra 25 (2) (2002) 25-34.

[7] J. Browne, J. Zhang, Extended and virtual enterprises—similarities and differences, International Journal of Agile Management Systems 1 (1) (1999) 30-36.

[8] E.K. Ouzounis, An agent-based platform for the management of dynamic virtual enterprises, Ph.D. Thesis, University of Berlin, 2001.

[9] H. Afsarmanesh, L.M. C.-Matos, A framework for management of virtual organization breeding environments, in: Collaborative Networks and Their Breeding Environments, Springer, Boston, 2005, pp. 35-48.

[10] L.M.C.-Matos, H. Afsarmanesh, M. Ollus, ECOLEAD: a holistic approach to creation and management of dynamic virtual organizations, in: Collaborative Networks and Their Breeding Environments, Springer, Boston, 2005, pp. 3-16.

[11] M. Wooldridge, Intelligent agents, in: G. Weiss (Ed.), Multiagent Systems, The MIT Press, April 1999.

[12] E. Castillo, E. Alvarez, Expert Systems: Uncertainty and Learning, Elsevier Applied Science and Computational Mechanics Publications, London and New York, 1991.

[13] P. Sanz, Partner selection in DVEs, Ph.D. Thesis, University of Valladolid, 2008.

[14] P. Sanz, J.J.D. Benito, Design and implementation of a multi-agent framework for the selection of partners in dynamic VEs, in: Leveraging Knowledge for Innovation in Collaborative Networks, Springer, 2009, pp. 341-348.

[15] P. Sanz, J.J.D. Benito, Diseño de un modelo global y operativo del proceso de selección de socios en las empresas virtuales dinámicas, Sistemas e Tecnologías de infomação, 2009, pp. 79-84.

[16] M. Nikraz, G. Caire, P.A. Bahri, A methodology for the analysis and design of multi-agent systems using JADE, International Journal of Computer Systems Science and Engineering 21 (2) (2006).

[17] A. Rocha, E. Oliveira, An electronic market architecture for the formation of virtual enterprises, in: Proceedings of the IFIP TC5 WG5.3 / PRODNET Working Conference on Infrastructures for Virtual Enterprises: Networking Industrial Enterprises, Kluwer, 1999, pp. 421-432.

[18] J. Rao, S.A. Petersen, Implementing virtual enterprises using AGORA multi-agent system, Information Systems for a Connected Society, CAiSE'03 Forum, Austria, 2003, pp. 193-196.

Computer Technology and Application 2 (2011) 782-787

Intelligent Transportation System and Night Delivery Schemes for City Logistics

Ivana Cavar, Zvonko Kavran and Natalija Jolic

Faculty of Transport and Traffic Sciences, University of Zagreb, Vukeliceva 4, Zagreb HR-10000, Croatia

Received: August 14, 2011 / Accepted: August 29, 2011 / Published: October 25, 2011.

Abstract: Intelligent Transportation System (ITS) integrates information and communication technologies with location based technologies into roads, vehicles, traffic and transport management systems. Application of ITS can improve situation in major cities where due to the increasing number of residents and level of motorization traffic congestion represents important issue. Another problem arises from this facts and that is that cities have contradictory needs as grooving need for goods within urban areas as well as need for less vehicles in the same area. The urban logistics activities by private companies within urban areas represent an integral part of city logistics that aims to improve the efficiency of urban freight transportation, reduce traffic congestion, mitigate environmental impacts, reduce costs and fuel consumption. This research presents application of ITS technologies and vehicle routing problem for night delivery scheme planning as one of the core techniques for modelling city logistics.

Key words: ITS (intelligent transportation system), city logistics, night delivery scheme, traffic system development.

1. Introduction

City Logistics is a relatively new field of investigation developed as result of challenge brought by increase of quantities of freight that are being delivered within urban areas on daily level. Well developed city logistics are of great interest to all stakeholders involved in these activities. Lately many regulators are developing or implementing different city logistic schemes to provide good quality, efficient, environmental friendly, productive and safe delivery of fright into urban areas.

City logistics include activities of delivering and collecting goods in/from town and city centres (transportation, handling, storage, management of inventory, waste and returns as well as home delivery services) while considering the traffic congestion, safety, environmental impacts and energy consumption. All of this should be done within the framework of a market economy and with respect to the interests of residents, business subjects (shippers and freight carriers) and policy makers. Simplistically, it concerns the means to achieve freight distribution in urban areas by

- Improving the efficiency of urban freight transportation and reducing traffic congestion (by a more efficient pick-up/delivery system that reduces the size of the fleets and maximized the load factor);
- Mitigating environmental impacts (reduction of the emissions and noise, rational land use planning);
- Reductions in delivery fixed and operational costs as well as
- Reductions in fuel consumption or the resort to alternative energies.

Urban driving style decreases impact of freight vehicles on surrounding environment. This is evident due to the facts that urban operating speeds are slower (engine of delivery vehicles is running at consistently lower than optimal speed) and urban fuel consumption and emission levels are higher (constant acceleration/deceleration). All of this is caused by

Corresponding author: Ivana Cavar, Ph.D., research fields: telecommunications, traffic simulations and planning, geographic information systems, ITS. E-mail: ivana.cavar@fpz.hr.

traffic lights, traffic restrictions and congestion as well as the wear of vehicles. Despite this the benefit for the whole society of restriction for freight transport in urban areas is conflicting. For the economic competitiveness of an urban area it is very important that exchange of goods is not too expensive or that regulations are too restrictive. This can lead to the decline in commercial activity in city centre areas as people move to the suburbs in search of better social and commercial activity.

The paper is organized as follows: Section 2 discusses the numerical argumentation for city logistics; section 3 gives description of different delivery schemes studied or implemented worldwide; section 4 introduces night delivery scheme trough the case study and gives comparison of night and day delivery schemes based on number of vehicles involved in delivery and overall distance travelled; section 5 gives conclusion remarks.

2. Importance of City Logistics in Numbers

Worldwide urbanization trend is emptying the countryside and making large cities even larger. This is evident from data showing that in Europe over 75% of the population is living in urban areas [1]. Within the countries members of OECD, the urban population was 50% of the total population in 1950, was 77% in 2000, and should reach the 85% mark by 2020 [2]. Data also show that 80% of GDB is produced in urban areas and 50% of accidents take place there [3]. All of this leads to a high potential for freight transport within an urban context.

When in urban areas freight transportation is compared to non freight transportation it yields results shown in Fig. 1 [3-6].

It is estimated that by 2030 45% of the transport related energy consumption will be due to freight transport and freight transport tonne kilometres are expected to grow by 63% [6].

3. Transportation, ITS and Logistics

Transport system is the most important economic component among the elements of business logistics systems. Well developed transport system in city logistics context can provide improved efficiency, service quality, safety as well as reduction of operation costs. Fig. 2 shows the components of logistics costs [7]. This analysis shows that transportation is the highest cost, which encloses 29.4% of logistics costs, followed by inventory, warehousing cost, packing cost, management cost, movement cost and ordering cost. This figure signifies cost structure of logistics systems and the importance order in improvement processing. The improvement of the item of higher operation costs can get better effects. Hence, logistics managers must comprehend transport system operation thoroughly. In

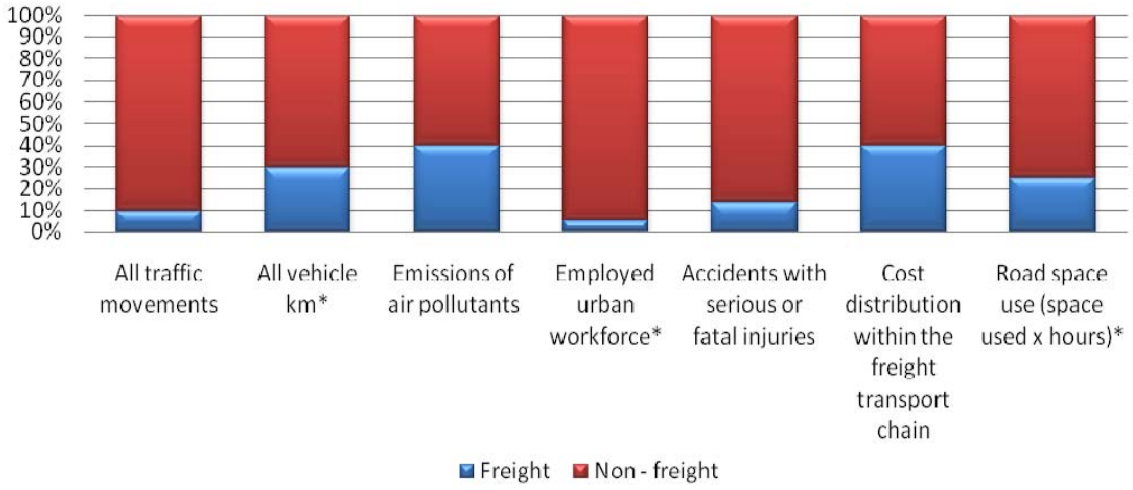


Fig. 1 Share of freight and non freight transportation characteristics within urban areas.

* means that several values can be found in literature and the highest one for freight transportation is considered.

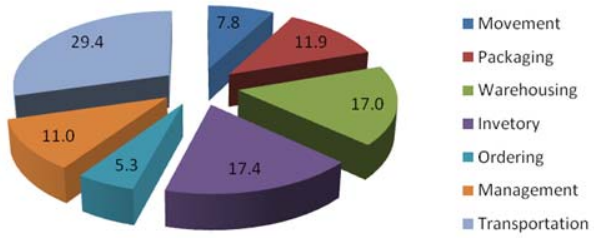


Fig. 2 Cost ratio of logistics items.

literature [8] can be found influence of costs in relation to different stakeholders objectives within city logistics in more details.

Applications of ITS in transport systems are widespread. The most common techniques for city logistics include Global Positioning System (GPS), Geographic Information Systems (GIS), advanced information systems and route guidance systems. GPS provides the service of vehicles positioning and together with two way communication, GIS and route guidance systems it is used to help the control centres to monitor and dispatch vehicles as well as to enable to organize their routes better. Advanced traveller information systems (ATIS) provide the real-time information for both managers and deliverymen and among this information special attention should be given to estimated travel time information within urban areas. The integration of GPS, GIS and advanced information systems provides a high manoeuvrability of transport systems providing a better service quality, reduced unnecessary trips, and increased loading rate.

When considered worldwide there are different approaches for city logistics schemes [6, 9-16]:

• Environmental zones are mainly implemented in European cities. There are different variations on the subject mainly focused on provision of certificates for vehicles that meet defined criteria and therefore have right to use special loading zones (Copenhagen) or on definition of environmental zones that can be entered by vehicles younger then defined age (mainly 8 years) and/or have emission control device installed (Stockholm, Gothenburg, Malmoe, Lund, London, Bristol, Nottingham, Edinburgh, Brussels, etc.).

• "Green" vehicles application for city logistics has been implemented and/or tested in cities around the world. Rotterdam gives an example of usage of electric vehicle city distribution system (ELCIDIS project), while in Osaka there is example of electronic vans with implementation of advanced information system. In Zurich authorities have adopted innovative cargo tram solution for waste collection. They turned out to be cheaper, faster and producing fewer pollutants then trucks. In Graz a project introduced the deployment of electrical vehicles for distribution purposes in the narrow streets of the city centre, especially during large construction works at the city tramline. In Norwich and Stockholm logistic centres has been set up outside of the cities to consolidate goods in clean vehicles to deliver to the city centre.

• Coordinated transport: Cities of Stockholm and Borlänge established logistic centres with aim to reduce energy usage and emissions through coordinated transports to the residents, municipal institutions, private companies, etc.. They succeeded to reduce number of stops for unloading in the area 45 percent. Another example is Public Private Partnership (PPP) for special goods traffic platform implemented in Berlin. It consists of areas along main shopping roads defined as loading zones with "no stopping" sign for private cars contributing to the reduction of congestion and steadier vehicle flow. PPP was also applied in La Rochelle for this scheme.

• Congestion mitigation: In Barcelona multiple-use lanes are introduced with different usage purpose during the different time of the day (regular traffic, loading/unloading services, resident parking, etc.). Paris, Barcelona, Rome Dublin, Turin, London and many other cities have applied night delivery schemes by increasing the proportion of deliveries at night. This provided faster delivery service and minimized the impact of freight on congestion. These schemes also promote the usage of cleaner and quieter vehicles for deliveries. Results from Barcelona shows that this operators are only partially successful (in 45% of cases)

in unloading within the ambient noise conditions.

• Congestion charging: The main aims of congestion charging are to reduce traffic congestion in and around the charging zone. This scheme is implemented in London and tested within some CIVITAS Elan towns (Zagreb, etc.). Also, distance-based truck toll for all heavy commercial vehicles and vehicle combinations is introduced in Germany.

• Information and communication technologies: Internet port information services have been implemented (for longer or shorter time period) in ports like New York and Vancouver. Application of vessel information and other services has helped reduce congestion and wait times at terminal gates at the port. Examples of Advanced Information System application for city logistics can also be found in Tokyo and Osaka. In Malmo and La Rochelle a system for delivering cars and vans that allows an efficient two way communication between the traffic-dispatch and vehicles has been installed. Project MOSCA aims to improve the efficiency of door-to-door transport of goods in urban areas by providing both demand and supply side information in a system. This resulted in a set of tools implements services for shortest path finding, vehicle route planning, on-line vehicle routing planning, urban shop delivery planning, etc.

• Water use: The problem of narrow and frequently congested roads that caused delivery vans often to be blocked or delayed is for the first time solved by floating distribution centre in Amsterdam.

4. Night Delivery Scheme

Night delivery scheme has been successfully implemented in many cities around the world and is highly recommended for medium size and big cities (more than 500000 inhabitants [17]. It ensures faster and more efficient delivery of freight but in the same time minimizes the impact of this activity on city congestion. The problems concerning night deliveries are mainly associated with noise for the city residents, as well as theft and security for both the drivers and the goods. This is mainly solved by the application of "silent" electric vehicles and different surveillance systems. The most important noise sources are shown in Fig. 3 [12].

Also, an important part of delivery cost concerning night delivery schemes is the increase of labour cost (night working bonus). Major advantages of night delivery schemes are evident thought less traffic on roads and higher road speeds during the day, reduced emissions, availability of parking spaces and better organization of the receiving of the goods and less disturbance for the customers from shop owners point of view as well as enhanced road safety.

This scheme incorporates information and communication technologies as well as specialized tools. Implementation of different VRP (Vehicle Routing Problem) algorithms for vehicle routing with travel time estimation procedures decreases efficiency of night scheme delivery planning and is the core technique for modelling city logistics.

4.1 Vehicle Routing Problem

The VRP is a generic name given to a whole class of problems. These problems concern a number of geographically dispersed cities or customers that need to be served by one or more vehicles. Vehicles routes starts/ends at one or more depots. The objective of the VRP is to deliver a set of customers with known demands on minimum-cost vehicle routes originating and terminating at a depot.

The optimized routes have a significant impact on the transport organization by reducing the overall costs and travel time duration. Practical experiences from

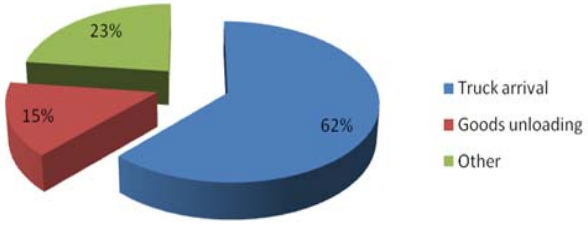


Fig. 3 Night delivery scheme noise sources.

literature show that by optimizing the transport organization by application of VRP solutions a reduction of transportation costs of even up to 20% may be achieved [18].

To solve the VRP problem the exact, heuristic and meta-heuristic algorithms are in use. Application of each of them has its own good and bad sides. The exact algorithms, based on linear programming techniques are limited to solving simple practical problems. With heuristic algorithms the acceptable (not the best) solutions can be achieved relatively fast. The algorithms of improving the heuristic solutions, e.g., 2-opt; try to further optimize an already existing solution by route modifications. The meta-heuristic methods, e.g., simulated annealing, genetic algorithms, tabu search, ant colonies, simulate natural processes and they have proven as very successful optimization methods.

4.2 Case Study

Problem of delivering goods to 45 customers with capacity and time window demand is solved with Clark and Wright algorithm. The distributional centre is situated in industrial area outside of city centre. Costumers are served during day (Fig. 4) and with night delivery scheme (Fig. 5).

Substantial savings (25% vehicles less needed and 17.8% distance less travelled) are made by application of night delivery scheme (Table 1).

This indicates that additional cost produced by night-working bonus for personnel (30% more) can be overcome by number of vehicles less needed for night delivery scheme (9 drivers * hours * 1.3 < 12 drivers * hours). It should be also noted that due to less congestion and higher travel speeds working time will be even less than during the day and therefore savings on application of night delivery scheme vs. day delivery even more significant. Less fuel consumption in this case is evidently pure saving for company that deriver goods. Less congestion and emission, as well as availability of parking spaces during the day and

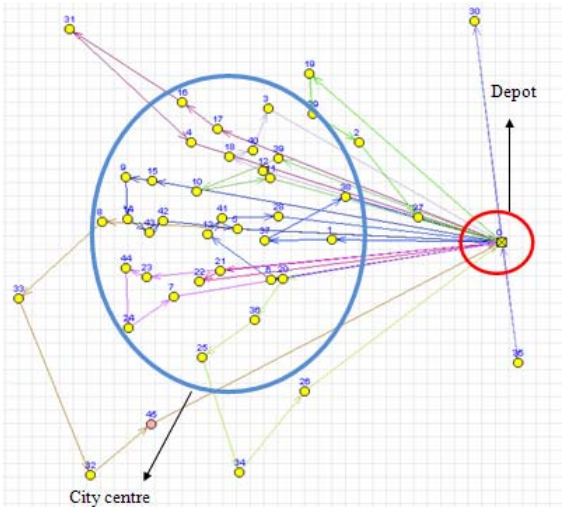


Fig. 4 Day delivery.

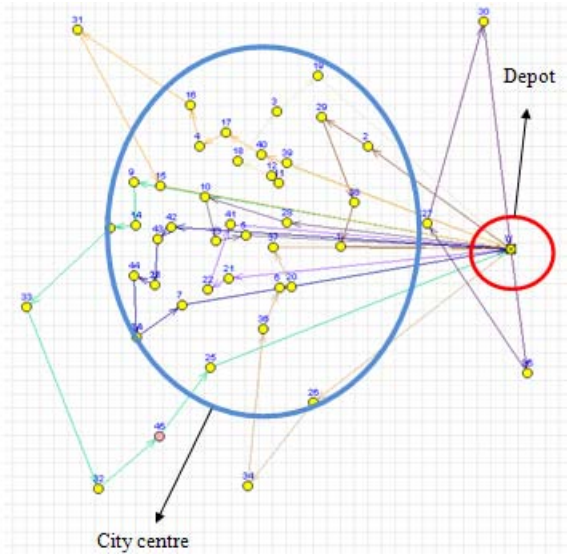


Fig. 5 Night delivery scheme.

Table 1 Day and night delivery results.

Scheme	No of vehicles	Distance travelled
Day delivery	12	9739.75
Night delivery	9	8006.17

enhanced road safety are pure winnings for the residents and city visitors.

5. Conclusions

During the last decades freight transport movements have increased enormously. The distribution of goods based on road services in urban areas contribute to traffic congestion, generates environmental impacts and incurs in high logistics costs. Significant gains can

be achieved through a streamlining of distribution activities resulting in less freight vehicles travelling within the city. Since transportation participates highly in cost ratio for logistic it is essential to make an effort to decrees these expenses. This should be made with high awareness on transportation influence on the city, city logistic and quality of life in it. Taking into account various types of solutions, optimization methods (as VRP) are necessary to satisfy the convergent needs for each stakeholder. Night delivery scheme is cost-effective both from the local authority and the operators' points of view. The difficulties of achieving the necessary noise reductions need to be assessed and the focus now aims to involve vehicle manufacturers, since the noise levels of general vehicle movements are of as much concern as the correct organisation and execution of unloading activities.

References

[1] M. Baratta, Der Fischer Weltalmanach 2001: Zahlen, Daten, Fakten, Fischer Taschenbuch Verlag: Frankfurt am Main, 2000.

[2] Organisation for Economic Co-Operation and Development, Delivering the Goods: 21st Century Challenges to Urban Goods Transport, Technical Report, OECD Publishing, available online at: http://www.oecdbookshop.org.

[3] M. Campanai, Best Practices in City Logistics Solutions in Italy: ICT & Organizational Models, SUGAR: Round Table, available online at: www.sugarlogistics.eu%2Findex.php%3Foption%3Dcom_docman%26task%3Ddoc_download%26gid%3D3%26Itemid%3D55&ei=MyRATtGWHYaBOsC8zfkO&usg=AFQjCNFmUwTsafnaDAMOECGiwzzmJklrBA.

[4] H. Koriath, W. Thetrich, COST 321, Urban Goods Transport: Final Report of the Action, Office for Official Publications of the EC, Bruxelles, Luxembourg, 1998.

[5] ETH Zürich IMPULSE, Operation Forms for the Network Modes, IN-95-SC.001 European Commission, Bruxelles 1996.

[6] Group of authors: Goods Distribution and City Logistics, CIVITAS Thematic Leadership Programme.

[7] Y.H. Chang, Logistical Management, Hwa-Tai Bookstore Ltd., Taiwan, 1998.

[8] E. Taniguchi, D. Tamagawa, Evaluating city logistics measures considering the behavior of several stakeholders, Journal of the Eastern Asia Society for Transportation Studies 6 (2005) 3062-3076.

[9] N. Geroliminis, C.F. Daganzo, A Review of Green Logistics Schemes Used in Cities around the World, available online at: www.metrans.org%2Fnuf%2Fdocuments%2Fgeroliminis.pdf&ei=aCRATtnPPMfqOYe77P4M&usg=AFQjCNFitEbVkxQiCBl56h1JngUi9-6lbA.

[10] European Directorate-General for Energy and Transport available online at: www.smile-europe.org.

[11] Available online at: www.eltis.org.

[12] Silence E-Learning for Decision Makers-Measures to Tackle Noise: the Night Time Delivery Scheme Works in Barcelona, available online at: http://www.silence-ip.org/site/index.php?id=177.

[13] W. Debauche, A Study on the setting-up of lorry dedicated routes in the Brussels capital region, in: Proceedings of the 3rd International Conference on City Logistics, Madeira, Portugal, June 2003, pp. 333-348.

[14] T. Nemoto, An experimental cooperative parcel pick-up system using the internet in the central business district in Tokyo, in: Proceedings of the 3rd International Conference on City Logistics, Madeira, Portugal, June 2003, pp. 309-320.

[15] M. Ruesch, J. Eugster, Strategic development of freight transport in the canton of Zurich: study on behalf of the Canton of Zurich, Final Report, Zurich, 8.10.2001.

[16] E. Taniguchi, S. Kawakatsu, H. Tsuji, New Co-Operative System Using Electric Vans for Urban Freight Transport, Urban Transport and the Environment for the 21st Century VI, WIT Press, 2000.

[17] Group of authors: Innovative Urban Transport Concepts, NICHES, available online at: www.niches-transport.org%2Ffileadmin%2Farchive%2FDeliverables%2F14277_transportconcept_1_BAT_low.pdf&ei=LiVATqWNJ4GbOri41IgH&usg=AFQjCNG-wn804r-xVD85ksGrEWw7kn32Vg.

[18] H. Gold, T. Carić, A. Galić, I. Ćavar, Vehicle routing problem solvingsystem as a CRO-GRID application, in: Proceedings of the 28th International Convention MIPRO 2005—Hypermedia and Grid Systems, Opatija, 2005, pp. 228-233.

Computer Technology and Application 2 (2011) 788-800

Some Tools to Model Ground or Supply Bounces Induced in and out of Heterogeneous Integrated Circuits

Christian Gontrand[1], Olivier Valorge[2], Rabah Dahmani[1], Fengyuan Sun[1], Francis Calmon[1], Jacques Verdier[1] and Paul Dautriche[3]

1. Université de Lyon, INSA- Lyon, INL, CNRS-UMR5270, Villeurbanne F-69621, France

2. R³ Logic, 50 chemin de l'Etoile, Monbonnot-Saint-Martin 38330, France

3. STMicroelectronics, 12 Rue Jules Horowitz, 38019 Grenoble Cedex 9, France

Received: July 19, 2011 / Accepted: August 09, 2011 / Published: October 25, 2011.

Abstract: Electrical ground looks simple on a schematic; unfortunately, the actual performance of a circuit is dictated by its layout (and by its printed-circuit-board). When the ground node moves, system performance suffers and the system radiates electromagnetic interferences. But the understanding of the physics of ground noise can provide an intuitive sense for reducing the problem. Ground bounce can produce transients with amplitudes of volts; most often changing magnetic flux is the cause; in this work, the authors use a Finite-Difference Time-Domain to begin to understand such phenomena. Additionally, predicting substrate cross-talks in mixed-signal circuits has become a critical issue to preserve signal integrity in future integrated systems. Phenomena that involve parasitic signal propagation into the substrate are discussed. A simple methodology to predict the substrate cross-talk and some associated tools are presented. Finally, the authors indicate a stochastic method which could grasp both outer or inner RF (Radio-Frequency) radiations and substrate parasites.

Key words: Electromagnetism, 3D (three-dimensional), integration noise, TSV (through silicon vias), ground or supply bounce, mixed analog-digital integrated circuits, substrate noise, stochastic methodology.

1. Introduction

The continued miniaturization of microelectronics is becoming a challenge from both the technological and financial perspectives. Three-dimensional (3D) integration technology ("more than Moore") offers significant promise of improved performance and form factor without having to necessarily introduce finer devices ("more Moore"). 3D technology relies on electrical interconnection directly between multiple strata that are stacked on top of one another, thereby offering significantly reduced interconnection length as compared with equivalent two-dimensional (2D) systems like Systems-on-Chip (SoC). This path length reduction translates into lower interconnection delay and power. This, in addition to improved form factor and the integration of heterogeneous technologies, is among the most promising benefits of 3-D integration technology [1-2].

Nowadays, mixed-signal integrated circuits (ICs) merge many blocks of different nature (digital, analogue, radiofrequency, AD-DA (analogue/digital-digital/analogue) converters, memory, power management, MEMS (Micro-Electro-Mechanical Systems).

This heterogeneity (towards SoC) means that managing signal integrity is a key issue. Advanced Complementary Metal Oxide Semiconductor (CMOS) processes (65 nm node or less) associated with power voltage decrease (below 1 V) impact the noise margin and the signal to noise ratio. As a consequence, for bulk silicon technologies, the designers have to face this

Corresponding author: Christian Gontrand, professor, research fields: telecommunications, noise, 3D integration. E-mail: Christian.Gontrand@insa-lyon.fr.

problem: "how to integrate different functions on or into the same silicon die without any perturbations between them". As already mentioned in the literature [3-5], the substrate coupling or the substrate noise propagation is basically the main problem to take into account. Many papers and research programs have been published in recent years, but this topic has still to be investigated, as it is a complex problem to quantify with a clear method and easy-to-use tools. As an example, the accurate substrate modelling by means of advanced finite element methods with powerful simulators is time wasting without taking into account the chip in its application—package, printed circuit board (PCB), etc. The context: Silicon technology based upon low or high resistivity substrate, the bulk noise generation and propagation. Subsequently, noise generation will be simplified by digital power supply ringing transmitted to P-substrate or N-wells.

This paper is organized as follows: First of all, we point out the necessity to model accurately through silicon vias (TSVs) that plays a key role for the 3D integration. In the following section, we develop simple analytical models to describe the substrate noise. In the last section, we try to adapt some markovian algorithm, often used in communication theory, to model the parasite switching digital flow statistics of complex circuits.

2. Through Silicon Via

Grounding of connector requires special consideration. The most important issue is to carefully consider how the currents will flow. The inductance and resistance of tracks and wires will cause voltage to be induced in the ground returns, and it is important to consider where these voltages will appear to avoid other circuitry malfunctioning or even being destroyed.

A good advice is: Do not accept that any point is necessarily "ground potential".

For high frequency circuits, if there is an opportunity to earth the metal case of devices, it is usually worthwhile. With PCB and careful supply routing and bypassing, it is possible to pass Electromagnetic Interferences (EMI) emission tests on just a double sided PCB with no case shielding. The overall shield is also desirable on low level analogue circuitry; in that situation, it reduces interference reaching the low level circuitry.

The studied structure, a TSV (see Fig. 1), is relatively simple. The simulator Sentaurus [6] used hereafter is based on the Finite-Difference Time-Domain method (FDTD) [7].

The FDTD algorithm: Brief recall

Mawxell's equations relative to the vortical fields, considering media as perfect are

$$\nabla \times \vec{E} = -\mu \cdot \frac{\partial \vec{H}}{\partial t} \quad (1)$$

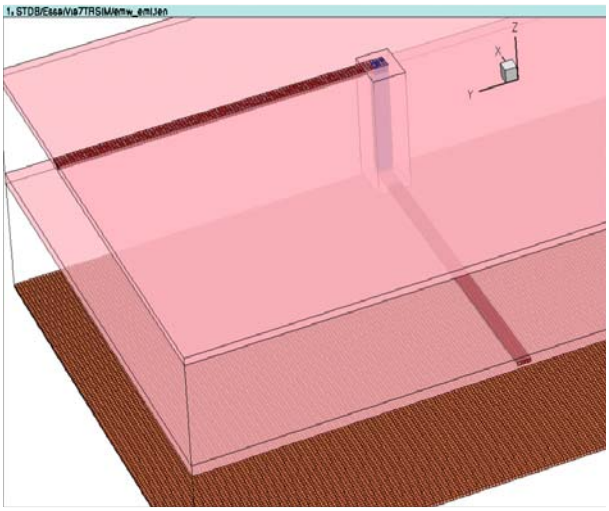
$$\nabla \times \vec{H} = \varepsilon \cdot \frac{\partial \vec{E}}{\partial t} + \sigma \vec{E} \quad (2)$$

where μ denotes the permeability, ε the permittivity, and σ the electric conductivity. Although there are different ways to resolve Maxwell' equations numerically, the FDTD is one of the most suitable schemes for the purpose of analyzing 3-D structures. Yees's pioneering approach [7-8] uses centred differences to approximate space and time derivatives Time derivatives are not computed at the same time-frame, but at different, interlaced intervals. Then, each unknown field component depends only on previously-calculated components of the dual kind: Two sets of equations are thus recovered, each set being composed of independent equations. This so-called "leapfrog" algorithm makes the set of linear equations easy to handle and solve sequentially. For instance the curl relative to the H field can be approximate as

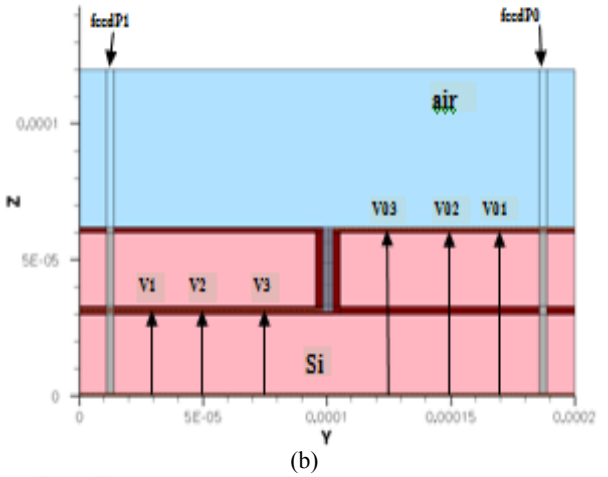
$$\frac{H_{y:i+1/2} - H_{y:i-1/2}}{\Delta x} - \frac{H_{x:i+1/2} - H_{x:i-1/2}}{\Delta y} = \varepsilon \cdot \frac{E_z^{n+1} - E_z^n}{\Delta t} + \vec{J}_{leqz}\left(\vec{E}_z\right) \quad (3)$$

The space grid size must only a fraction of the wave length.

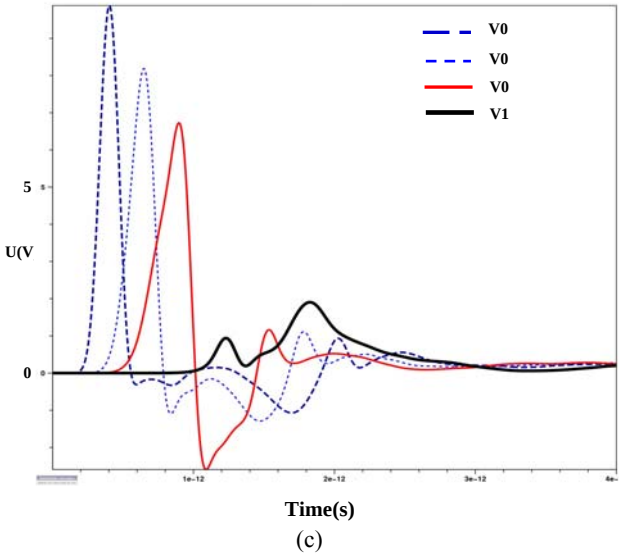
Figs. 1a-1b show a 3D view of a circular hollow via; the characteristic dimensions are



(a)



(b)



(c)

Fig. 1 Two perpendicular striplines apart of a circular via: (a) 3D via and bent striplines (z: vertical direction); (b) Flattened <x, y> plane z-transversal cut; the via is at the center (in grey), between the V3 and V03 probes; (c) Voltages observed at the two ports, for the transient regime.

- Striplines widths: w = 0.2 mm;
- Striplines thickness: t = 50 µm;
- Striplines length: l = 3.6 mm;
- Via height: 0.6 mm;
- Via diameter: 0.6 mm;
- Via thickness: 0.2 mm.

The via goes to a ground plane; additional shielding planes are added on the top and the bottom. Transient simulations are performed for each stripline without being connected to the via, for extraction of incident pulses. Simulations for the striplines connected to the via, with each port excited, are also needed for extraction of scattered pulses (e.g., Fig. 1: two perpendicular strpilines (in <x, z> and <y, z> planes, apart of a via).

We observe some complex oscillations (Fig. 1c); an important conclusion is that will seem necessary to analyse the TSV/device coupling also in an electro-magnetic point of view, because of the high frequency parasitic phenomena involved in complex circuits.

3. Substrate Noise in Mixed-Signal ICs in a Bulk Silicon Process

3.1 Ground and Substrate Noise Mechanisms

One of the main phenomena that induce substrate noise in a mixed-signal circuit is the power and ground supply voltage fluctuations that are transmitted into the substrate through all substrate biasing contacts [8]. The power and ground supply lines are not perfect and introduce several parasitic elements: Resistances, inductances and capacitances. In a large digital circuit, high peaks and fast slew rate on supply current create power-supply noise in the supply network due to *RLC* (Resistance, Inductance, Capacitance) network formed by all these parasitic elements at different levels of a SoC: PCB, package and the circuit itself. Other mechanisms create substrate disturbances in mixed-signal devices [9].

These mechanisms are just referred here and shown in Fig. 2, which represents a single digital inverter.

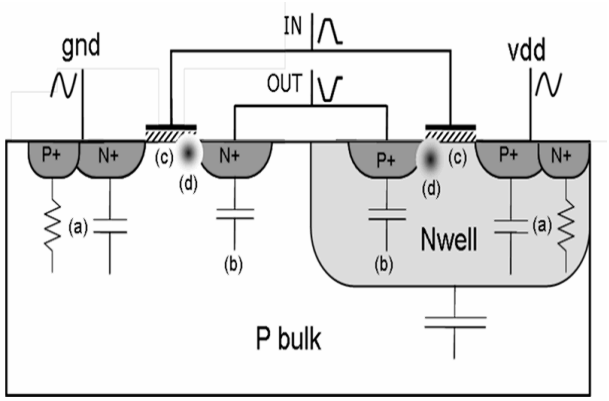


Fig. 2 Substrate noise injection mechanisms in a digital inverter.

The main phenomenon of substrate coupling is briefly described: It is the power supply noise - (a: see figure) -. The signal transitions can also be coupled to the substrate through different physical structures: the input MOS capacitance - (c) - of all transistors, the output drain - (b) - to substrate capacitances and the metal interconnection to substrate capacitances. Other phenomena can induce parasitic substrate currents like impact ionization - (d) -, photon induced current and diode leakage current.

After the substrate noise generation briefly described in the previous section, the propagation has to be quantified through the substrate. Silicon material is characterized by a cutting frequency $f_c = \sigma / 2\pi\varepsilon$, where σ is the conductivity and ε is the permittivity. With high resistivity substrate (> 10 Ω·cm), silicon can be considered as purely ohmic for signal frequency below 10 GHz [10]. Then, the silicon equivalent transfer function can be defined as an attenuation factor that can be calculated with point-to-point equivalent resistance. At this step, one has to consider the nature of the substrate: low or high resistivity silicon substrate.

Considering low resistivity substrate, the substrate model depends on the comparison between the epitaxial layer thickness (W_{epi}) and the distance D between substrate taps. As shown in Figs. 3-4 (case b), when the distance is up to $4 \times W_{epi}$, all the current flowlines pass through the low resistivity substrate that can be considered as a unique point. The resistance variation function of the distance D is represented in

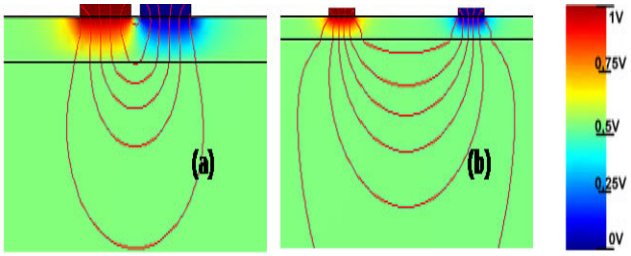


Fig. 3 Current flowlines for low resistivity substrate with epitaxial layer (W_{epi} thickness): (a) $D < 4 \times W_{epi}$; (b) $D > 4 \times W_{epi}$.

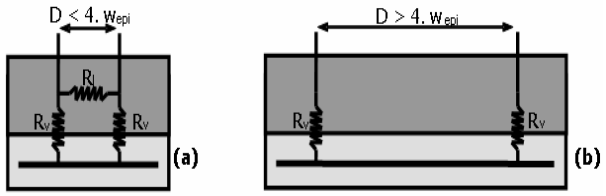


Fig. 4 Equivalent model for low resistivity substrate with epitaxial layer (W_{epi} thickness): (a) $D < 4 \times W_{epi}$; (b) $D > 4 \times W_{epi}$.

Fig. 5 (substrate tap area is 20 × 20 µm², W_{epi} = 5 µm @ 10 Ω·cm, substrate thickness is 300 µm @ 0.05 Ω·cm).

For high resistivity substrate, the die backside connection is first of interest. If the die backside is floating, the resistance between two taps increases with the distance (Figs. 6-8). If the die backside is correctly connected to the ground, the resistance becomes constant when the distance is larger than the substrate thickness $W_{substrate}$ as illustrated in Figs. 6b and 8 (substrate tap area is 20 × 20 µm², $W_{substrate}$ = 50 µm @ 6.7 Ω·cm).

The resistance characteristics function of the distance between the substrate taps presented in this section are calculated with a commercial 3-dimensional finite element method (FEM) simulator in static condition. Two taps (20 × 20 µm²) are placed on the silicon (first tap @ 1 V and second tap @ 0 V). The current between these two taps allows calculating the substrate equivalent resistance.

3.2 Ground and Substrate Noise Modeling

In this approach, we consider the power-supply noise as the unique substrate perturbation source.

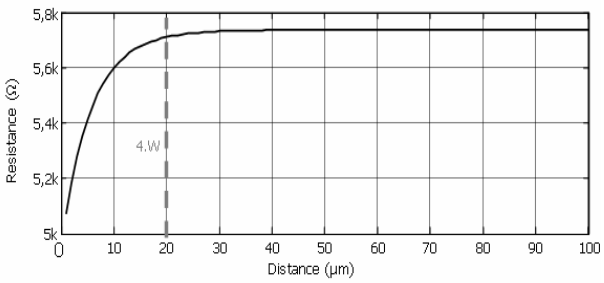


Fig. 5 Equivalent resistance function of the distance for low resistivity substrate with epitaxial layer (W_{epi} thickness).

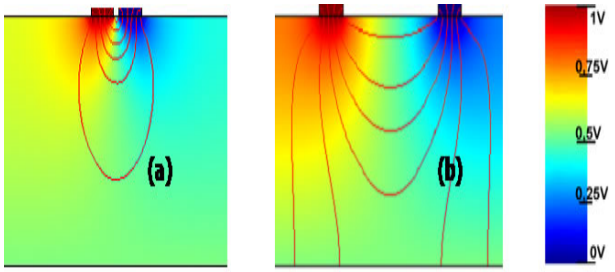


Fig. 6 Current flowlines for high resistivity substrate ($W_{substrate}$ thickness) with grounded backside: (a) $D < W_{substrate}/2$; (b) $D > W_{substrate}/2$.

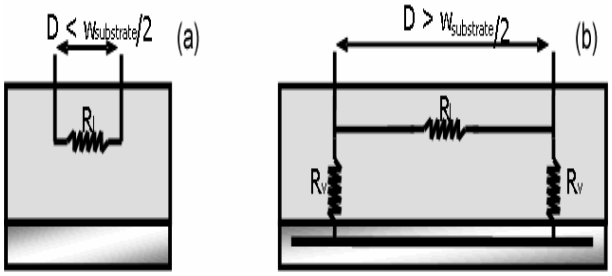


Fig. 7 Equivalent model for high resistivity substrate ($W_{substrate}$ thickness) with grounded backside, (a) $D < W_{substrate}/2$, (b) $D > W_{substrate}/2$.

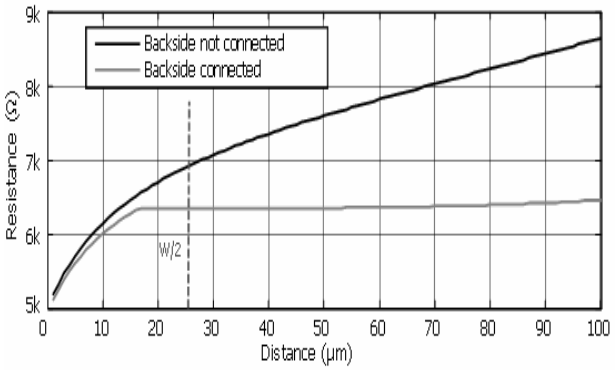


Fig. 8 Equivalent resistance functions of the distance for high resistivity substrate ($W_{substrate}$ thickness) with floating or grounded backside.

In order to model power and ground voltage oscillations, we have chosen to use the Integrated Circuit Emission Model approach (ICEM) [11]. The classical ICEM has been extended with a Substrate Network Sub-Model. Fig. 9 gives an overview of the extended-ICEM model.

The Passive Distribution Network (PDN) Sub-Model is supposed to model the parasitic elements of power supply or signal lines. The Internal Activity (IA) Sub-Model characterizes the dynamic current consumption of the device. The Substrate-Network (SN) Sub-Model is expected to model the substrate propagation of parasitic signals from aggressor parts to victim parts.

3.3 Developed Tools

Basically, the conceptual extended-ICEM methodology can be described in its simplest form by the electrical diagram presented in Fig. 10 (one digital block with one *Vdd-gnd* power supply pair). Nevertheless, some parameters have to be determined for each IC project. So, the authors propose to use software applications developed in JAVA language. The first application deals with quantifying the ground and power supply fluctuations (bounces). This can be done in three steps. Firstly, the user describes the passive elements: R, L, C_{core}, and K.

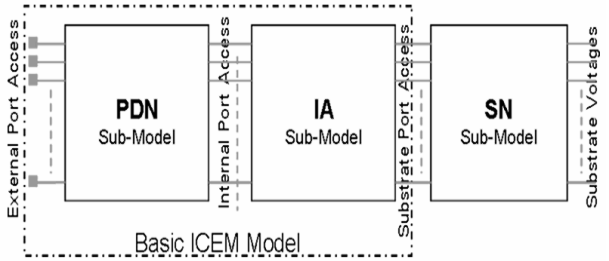


Fig. 9 Conceptual architecture of extended ICEM.

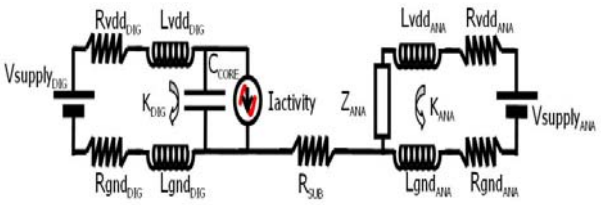
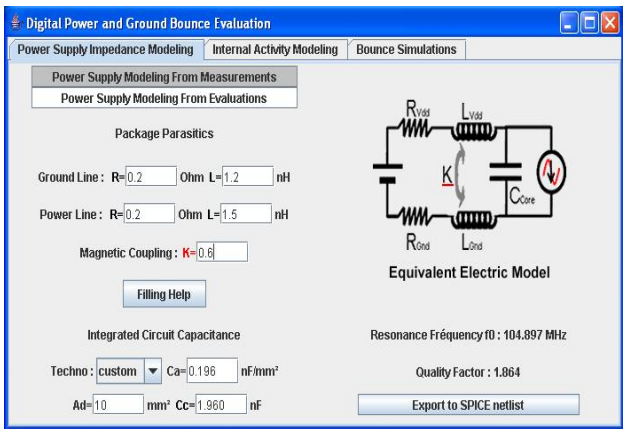


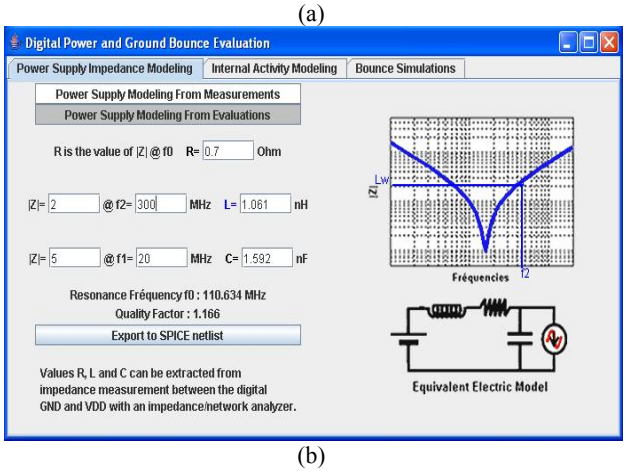
Fig. 10 Basic architecture of extended-ICEM model.

3.4 Application Developed for Ground and Power Supply Bounce Effects

The parameters R and L are the parasitic resistance and inductance of the lines (PCB, socket, package, on chip metal rails). K is the coupling factor (mutual inductance) between the lines (mutual between bonding wires, typical value is around 0.5). In digital IC's, the power and ground parasitic elements are mainly due to the package; power-grid effects can be neglected in a first order approach. Package lead and bonding wire lengths are often longer than on-silicon metal lines and induce mainly parasitic inductances. C_{core} is the equivalent digital capacitance. These passive parasitic elements can be estimated or determined from the S-parameter measurement (Figs. 11a-11b).



(a)



(b)

Fig. 11 Dedicated application to quantify digital ground and power supply bounces (passive distribution network): (a) Supply impedance passive parasitic elements model; (b) Passive parasitic elements can be determined from the S-parameter measurements.

In a second step, the internal activity of the digital block is described in a simple way. This current source is the superposition of periodical current pulses. The characteristics of each pulse (height, width, frequency, rising and falling time) can be easily determined from general digital circuit parameters: its average consumption, its clock frequencies, its rising and falling transition times, its skews and each clock latency (Fig. 12).

The last step concerns the ground and power supply bounce calculations. This is performed in the time domain (analytical resolution of *piece-wise* waveforms) or in the frequency domain (cf. Fourier analysis). Fig. 13 shows an example of simulated *Vdd* power supply ringing for different line resistances (from 0.05 Ω to 0.5 Ω).

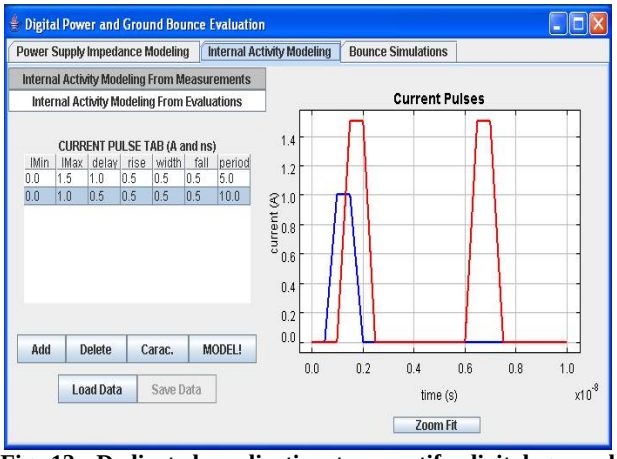


Fig. 12 Dedicated application to quantify digital ground and power supply bounces (internal activity).

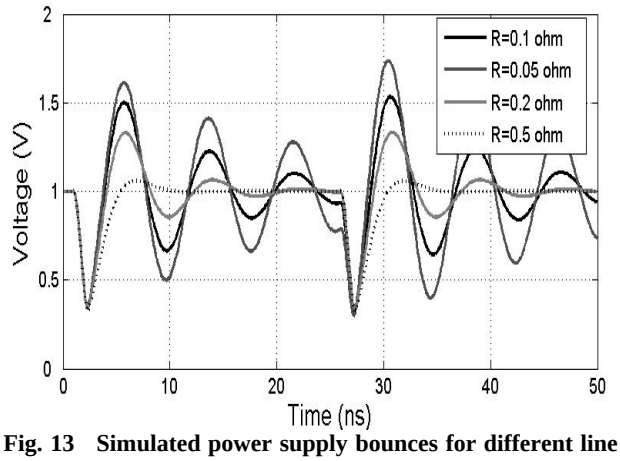


Fig. 13 Simulated power supply bounces for different line resistances.

3.5 Basic Rules to Reduce Digital Power Supply Network Ringing

With the presented tools, the user can estimate how the analog ground will fluctuate due to the activity of the digital(s) block(s) even if the supplies are completely separated (Fig. 14). Moreover, these tools can help to understand malfunctions of digital blocks with an unstable digital power supply.

The passive distribution network presents a resonant frequency $f_{res} = 1/2\pi\sqrt{LC_{core}}$ and a quality factor $Q = L\omega_{res}/R$. The passive distribution network will oscillate for $Q > 0.5$. To suppress or reduce power supply ringing, the designer can control the following parameters: R (line resistance from board or inside chip), L (line inductance), and C_{core} (core capacitance and additionally on-chip decoupling capacitance). Hereafter, we summarize the influence of these parameters on the digital power supply ringing:

• Increasing R leads to a decrease in oscillation amplitude without any influence on the oscillation frequency (Fig. 13);

• Increasing L leads to a decrease in the oscillation frequency without any influence on the oscillation amplitude;

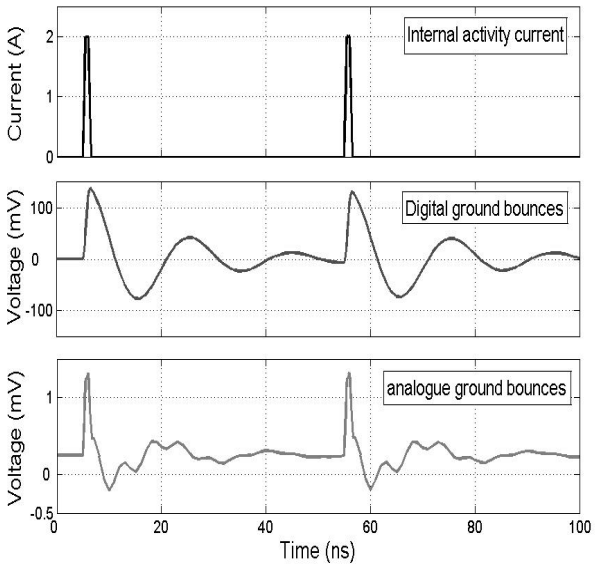


Fig. 14 Simulated analog ground fluctuations.

• Increasing C_{core} leads to a simultaneous decrease of amplitude and frequency of the oscillations.

An intuitive way to decrease power and ground bounces is to add a huge decoupling capacitance on the power network of the PCB. In fact, this external decoupling capacitor just filters the power network parasitic voltage, and is not enough to decrease significantly the interferences. The main disturbing frequency is due to the ringing of the *RLC* path composed of the bonding wire inductance and resistance and by the core capacitance of the digital part of the circuit. Increasing as much as possible the off-chip capacitance will not avoid the on-chip natural ringing. Another simple solution is to add a serial resistor to the power network (on the PCB or inside the chip). The added resistor decreases the quality factor of the *RLC* network and, in the same time, dampers the induced oscillations. But, too great a resistance of the power network induces a high power consumption of the chip and also could cause some functional mistakes. These are the reasons why this kind of noise reducing technique can be applied to a small digital application such as a frequency synthesizer.

The internal activity (digital equivalent current source) has also a direct link with the power and ground oscillation amplitude. For clock frequency below resonant frequency, the power supply ringing will be enhanced when the resonant frequency is a harmonic of the clock frequency (the worst case is for $f_{res} = 2f_{clock}$). For resonant frequency below clock frequency, the passive distribution network acts as a low-pass filter. An increase in the average consumed digital current leads to an increase in the oscillation amplitude. The decrease in the transition times leads to a decrease in the oscillation amplitude.

Parasitic waveform of voltage (or current) flow can be modelled by some stochastic process, instead of efficient, but very time consuming statistical methods (cf. Monte-Carlo); another way is to introduce the Markov chain theory utilization [12].

4. Stochastic Behaviour of the Perturbations

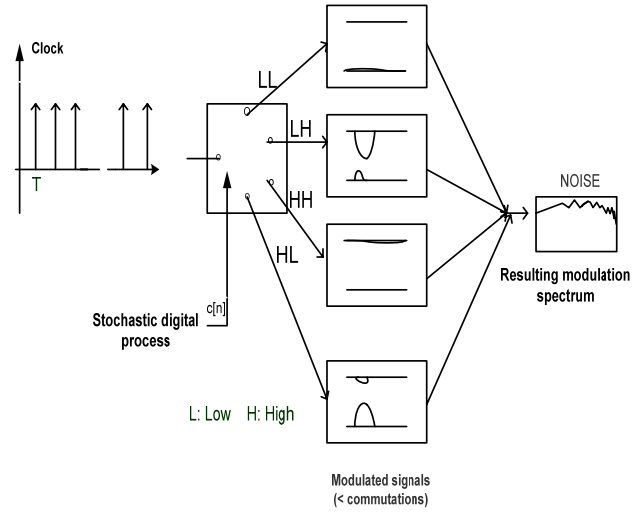
Up to now, the perturbations introduced by the digital part against the analogue one or parasitic radiation waves are key problems, which is hard to handle by the designers, especially in the radio-frequency (RF) range. Substrate noise involved by the digital part, or by any parasitic wave, will degrade the performance of analogue circuits. The in-depth problem is to construct efficient models for the digital part activity, the substrate, and their coupling, in a package that could be embedded in some unique computer aided design (CAD) tool. The well-known tool SPICE [13] is able to model such phenomena: substrate or digital signals; but for a complex technology, with very large digital blocs, it seems too optimistic. Some significant attempts on modelling noise generation injected in the bulk, on a "SPICE" and a process/device point of view have been proposed; but it seems accurate for relatively small circuits and simple parasitic waveforms, travelling through the substrate.

Today CAD tools cannot be used, ab initio, in order to slow down paths that are not timing critical, and so reducing parasitic coupling mechanisms. However, for this goal, a placement option, which can be used during the place/routing stage, exists in most of CAD tools; this option is called "recovery" and is used to downsize the logical gates which are oversized. Its first objective is to reduce the digital gates area, but it also decreases the global noise of a digital block by reducing all transitions that are unnecessarily fast. Though this step could be CPU time consuming, it should be interesting to include it into the mixed signals IC's back-end digital flow.

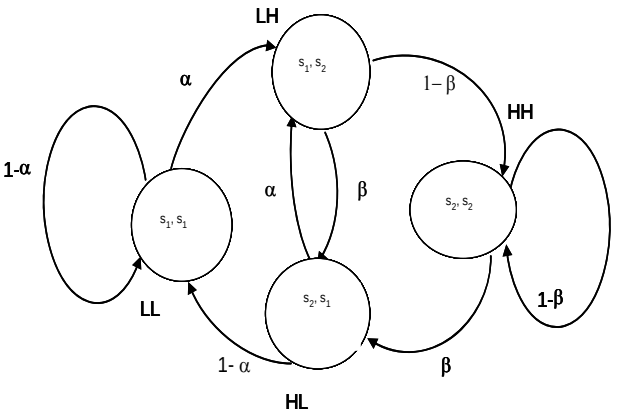
Signal parasitic injections, and their possible interferences, can also originate from physical phenomena, like impact ionization (from CMOS cells or at the P/N junction level), from exterior or interior (RF induction) radiative electromagnetic waves. The interferences can propagate through various ways: power supply grid, signal wires, package pins, bondings, TSVs and obviously the silicon substrate, in-depth or at its surface.

Briefly said, in a Markov process, "the past has no influence on the future if the present is specified."

The stochastic methodology relies essentially on probabilistic modelling of digital switching activity. The ultimate aim is to compute the interference noise power spectrum (see the synoptic block diagram in Fig. 15a). This spectrum should have both discrete and continuous component; the first one is involved at a periodic rate, triggered by the digital circuit clock(s), the continuous component resulting from the stochastic nature of the events. In fine, the resulting spectrum will depend on the shape of the injected interference signals.



(a)



(b)

Fig. 15 (a) Block diagram of the propagation chain of the parasitic signal; (b) Four states sequential machine.

First of all, we will depict the methodology, as generally as possible. The idea described in this paper has been proposed in Ref. [14]; in this work we present its development, work out details and finally present, for the moment, simple examples for the practical application of the methodology we did implemented.

4.1 Methodology

Our subjects of interest are the digital transmission/propagation systems [15-17].

For a general approach, consider M specific noise (parasitic) sources, $x_1(t)$, $x_2(t)$, ..., $x_M(t)$, time-continuous signals, modulated by an encoded message c[n] . It can be written in the general form**:**

$$x(t) = \sum_{n=-\infty}^{\infty} g^{c[n]}(t - nT \pm \phi) \qquad (4)$$

This equation means that the "symbol" c[n], the switching event, emitted at the instant *nT*, assigns the waveform $g^{c[n]}(t - nT)$ (the phase Φ will be equal to zero, in this paper), where T is the period of the digital clock.

The origin of this parasitic signal could be twofold:

(i) From the digital block which can send parasitic signal synchronized with the clock (of T' period);

(ii) From parasitic electromagnetic waves, in the radiofrequency range, coming from the exterior or the deep substrate (of T'' "pseudo"-period), and propagated via some TSV, for instance.

We could state Φ as a random phase next work, independent of the encoded message c[n], either to model some jitter (i) or some delay for parasitic wave injections (ii).

In these two cases, the parasitic signals—currents/voltages and electromagnetic waves can be approximated, by now, as a Gaussian pulse.

Let s_n be a stationary Markov chain, with state set $S = \{\sigma_1, \sigma_2, ..., \sigma_I\}$ and c[n] = $f(s_n)$ be a vector-valued stochastic process obtained as a memoryless function of s_n.

We assume that the Markov chain s_n is ergodic (s_n has only one class of communicating states. No degeneracy is involved in this process; a temporal average is equal to an ensemble one).

The digital signal *s* state can be defined as following:

L → H (or: 0 → 1) : with a probability = α;

H → L (or: 1 → 0): with a probability = β, (L and H mean respectively high and low).

Fig. 15b represents the state diagram of such a "machine".

Let us suppose the "input" *i* is itself a Markov Chain with a transfer probability matrix:

$$\pi = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} = \left(\begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix} \right) \qquad (5)$$

with $\sum_j \pi_{ij} = 1$, *i* = 0 or 1: row number; *j* = 0 or 1: column number.

π_{01} = Pr {j =1, if i =0} =α and π_{10} = Pr {j = 0, if i =1} = β.

Then: π_{00} = Pr {j = 0, if i = 0} =1 – α, and π_{11} = Pr {j =1, if i =1} = 1-β.

A simple digital signal that may switch from low to high with probability α, and low to high with probability β at every clock tick, is modelled by four states (see Fig. 15b); it has a state Transition Probability Matrix (TPM) representing the switching activity, expressed as an $I \times I$ Matrix *P*:

$$P = \begin{bmatrix} 1-\alpha & \alpha & 0 & 0 \\ 0 & 0 & \beta & 1-\beta \\ 1-\alpha & \alpha & 0 & 0 \\ 0 & 0 & \beta & 1-\beta \end{bmatrix} \qquad (6)$$

Our application can be straightly investigated as spectral properties of encoded messages. It is a direct application of the communication theory field, although it concerns, in this work parasitic signal propagation and interference.

The general term of *P* can be understood as:

$P_{ij/kl}$ = Pr {state transition, from *i* to *j*, at *n +1* time/ if transition *k* to l at *n*}.

Our objective is to find the FFT of the corresponding autocorrelation function (cf. second order moment).

This spectral density *w(f)* can be splited in two terms: $w(f) = w_d(f) + w_{cont}(f)$, where *f* is the frequency.

The discrete (W_d) component exhibits spectral lines or jumps at the frequency of the clock and its multiples ($f = n/T$), while the continuous part (w_{cont}) can be straightly derived by setting $z = exp(j*2\pi f)$ into the z-transform of its correlation function:

$$w(z) = \sum_{k=-\infty}^{\infty} R\,cont\,[k]\; z^{-k} \qquad (7)$$

(R_{cont} is explicated hereafter).

Moreover, the discrete random process *S* can be characterized by an autocorrelation matrix:

$R(k) \equiv E\{S_n^T(j)S_{n+k}(i)\} = Pr\{s_{n+k} = \sigma_j,\ s_i = \sigma_n\}$ (T: transpose).

By using the matrix algebra tool, the whole correlation matrix (of the source codes c[n]) can be derived as

$R_c(k) = C^T R(k) C$, where *R(k)* is the joint probability matrix between the random vectors S_n and S_{n+k}, with the "code words "c[n] = $S_n C$, where *C* is a matrix whom each column assigns a specific parasitic waveform.

R[k] can be as well decomposed in two "orthogonal" parts (cf. Vold's theorem [12]), a continuous and a discrete one:

$R[k] = R_{cont}[k] + R_d[k]$, where

$$R_{\text{cont}}[k] = \begin{cases} \Delta H^k & k > 0 \\ \left(H^T\right)^{-k}\Delta & k < 0 \\ \Delta - \sum_{r=0}^{h-1} K_r & k = 0 \end{cases} \qquad (8)$$

and a periodic component (cf. pure tones) with period *h*, given by

$$R_d[k] = \sum_{r=0}^{h-1} K_r . \exp(\ j2\pi rk\ /\ h) \qquad (9),$$

which is cyclic when $h > 1$.

We substitute (8) into (7) and taking into account that the limit of the matrix sum:

$M_i = \sum_{k=1}^{i} H^k z^{-k}$ converges to the matrix $M_\infty = zI - H$ (this result is easy if we use the so-called matrix formula:

$\sum_{k=0}^{\infty} A^k = (I - A)^{-1}$, for any matrix *A*, provided all its eigen values have a modulus less than the unity).

The eigen values λ_h of *H* are the value null, or non-null with a modulus lesser than 1 (λ_h = 0, 0, 2-α-β).

After some cumbersome, bur relatively easy calculations, the continuous part of *w(z)* (cf. Eq. (8)) can be written:

$$w(z) = (z^{-1}I - P^T)^{-1} C\ (zI - P)^{-1} \qquad (10)$$

with $C = \Delta - P^T \Delta P$ (10 bis)

$z = exp(j*2\pi f)$ and *I* is the identity matrix.

The spectral density of the modulated signal *x(t)* can be formally written as

$$S_x(f) = S_x^{\text{cont}}(f) + \sum_k \delta\left(f - f_k^d\right) \cdot S_{x,k}^d \qquad (11)$$

The continuous spectral density of the modulated signal *x(t)* is given by

$$S_x^{\text{cont}}(f) = \frac{1}{T} \cdot G^*(f) \cdot W_x(e^{j2\pi fT}) \cdot G^T(f) \qquad (12)$$

where *G(f)* is the Fourier transform of $g(t)$ and W_x (cf. Eq. (6)) is the Fourier transform de $R_x(\tau)$, with

$$R_x(\tau) = \frac{1}{T} \cdot \sum_{k=-\infty}^{+\infty} \int_{-\infty}^{+\infty} g^*(t) R(k) g(t + \tau - kT) dt \qquad (13)$$

Indeed, note that $R_x(\tau)$ depends on the statistics of the Markov chain s_n, depending on *R(k)*.

On another hand, the noise signal *x(t)* exhibits spectral lines (jumps) at frequency multiples of $f_0 = 1/T$, of amplitude

$$w_x^d(r) = C^* K_r C \quad , \ r = 0,1, \ldots, h\text{-}1 \qquad (14)$$

Then $W_x^d(r)$, with K_r defined by (13) becomes

$$W_x^d(r) = \left|\frac{1}{T} p J_r G_r(f_0)\right|^2 \quad (15)$$

(J_r is the diagonal matrix of general term λ_r. I).

The TPM P has only one eigen value, λ_0 equal to 1; all eigen values of modulus 1 result in Dirac distributions. In our case, only λ_0 results in a train of Dirac distributions at the frequency $1/T$ and its harmonics.

Our problem is analogue to some modulation in digital communication. Then g(t : c[n]) can be straightforwardly written c[n] .f(t), f(t) being the parasitic basic waveform vector. Finally, g(t) ≡ f(t).C. Therefore the continuous part of the spectrum becomes

$$W_x^{cont}(f) = \frac{1}{T}\cdot|F(f)|^2 \cdot Wc(e^{j2\pi fT}), \quad (16)$$

with $w_C(e^{j2\pi fT}) = C^* w(e^{j2\pi fT})C$ (16) bis

And the discrete one:

$$S_x(f) = \frac{1}{T^2}\cdot|F(f)|^2 W_c^d \quad (17)$$

If now, we actually study the propagation of the noise, it is necessary to take into account the transfer function of the "propagation channel", said $T(f)$.

We get the spectrum noise density, taking into account the network parasitic function transfer, T. The final power spectrum will be

$$S_y(f) = T(f)\cdot S_x(f)\cdot T^*(f) \quad (18)$$

with: $S_x(f)$: Power spectral density of the noise sources obtained using modulation theory;

T(2j π f) (generally) multi input (noise sources), multi output linear (number of sensitive points) system.

4.2 Simulation Example

As a check to illustrate the algorithm explicated above, we study the case of signals injected into the analog power grid by one switching digital cell.

We will consider, for the moment, the power (or ground) bounces as the unique source of substrate noise (knowing that this matricial method can be easily generalized). For instance, we present some noise measurements, concerning a 90 nm CMOS technology (see Fig. 16). In a CMOS inverter, the NMOS source transmits the ground bounces and the PMOS source transmits the power fluctuations. The power and ground fluctuations are generated by the large transient-current spikes pulled by the simultaneous switching of multiple digital gates. These spikes, along with the parasitic elements of the package, PCB, pad ring and other parts of the power and ground networks, can introduce resonant pathways and so power and ground ringing. Supply oscillations are directly coupled to the substrate through some P+ biasing. The coupling with the substrate can be: resistive through substrate biasing contacts, capacitive through interconnects and depletion capacitances, non-linear through the MOS body-effect mechanism for instance or other effects related to each basic analogue component.

In fact, the main perturbing frequency seems to be the ringing of the RLC path composed of the bonding wire inductance and resistance and by the core capacitance of the digital part of the circuit; in others words, power (and ground) lines have to be efficiently modelled to predict voltage drops.

The ICEM (Integrated Circuit Emission Model) [11] predicts conducted-mode emission levels on a chip; we can describe every part of the power lines with this model. Each part is modelled by an RLC network (see the schematic in Fig. 11. In our case, and for the future developments, for a digital block, the switching activity current will be applied to one point: the middle point of this block.

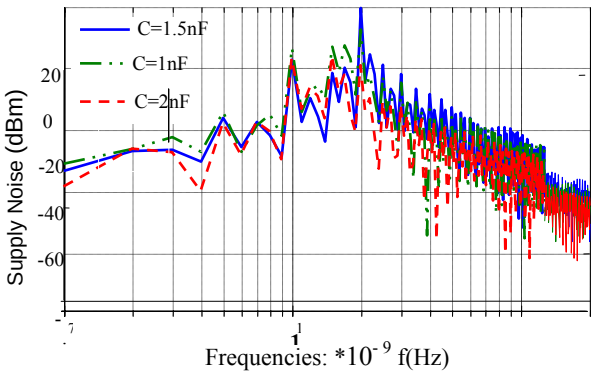


Fig. 16 Typical supply noise measurements, with different core capacities.

As a first example to validate the methodology and provide insight into some aspect of the problem, we consider a gate with only the state transition LH and HL (since the LL and HH state transitions involve a very low parasitic signal); the LH and HL parasitic signal shape will be assumed to be Gaussian versus the time. We extracted the parameter mean values (from the 90 nm CMOS process (see the beginning of this section):

LH: delay: 20 ns; standard deviation: 0.001 ns. V_{peak} = + 0.5V.

HL: delay: 60 ns; standard deviation: 0.01 ns. V_{peak} = - 0.15V.

Fig. 17 shows the first simulations coming from the stochastic algorithm; it depicts the sum of the continuous and discrete parts of the power spectrum of the parasitic waveforms injected into the Vdd line. A periodic behaviour in this spectrum can be seen corresponding to the clock frequency. This figure shows a global result on what can be produced by a commutating gate through a RLC-like network.

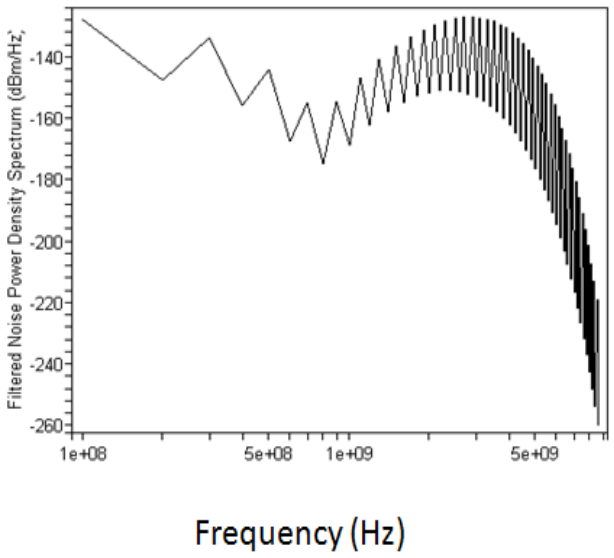


Fig. 17 Noise spectral density: simulation.

5. Conclusions

In this article, after a first approach of bounce creation on PCB or/and in circuit, we have resumed the basics of the substrate noise generation and propagation in mixed-signal IC's. We present the standard IC Emission Model and its extension allowing substrate noise simulations. This model considers only the noise generated in the substrate by the power supply and ground fluctuations of the digital part. Although this approach seems simple, it gives good results regarding test-chip measurements. To reduce substrate noise, the designers can control some parameters concerning, at first, the noise generators. The skew and the transition time can be optimized in order to reduce the average current and then power supply ringing. The passive power supply distribution network, through its parasitic elements RLC, has also a key role. Adding decoupling capacitance (PCB and/or on-chip), resistance (if possible) may reduce oscillation frequency and magnitude. Finally, substrate isolation techniques can be implemented (beginning with the substrate resistivity choice and the die back-side connectivity). Classical guard-rings and triple-wells are often solutions but their efficiency depends on the biasing quality. Furthermore, triple-well will lose its isolation property when its area increases. On another hand, the designer cannot act on exterior parasitic EM waves; a practical solution could be to build some surface (or interior) waveguide, to divert aggressor waves from the sensible circuit parts. General rules cannot be claimed as each design is a specific case and the solutions have to be adapted. Our contribution is pragmatic as it proposes a simple investigation methodology to a complex problem. Nevertheless, some work remains to be continued, for example concerning the way to obtain the transient consumption current of an entire large digital part or an efficient method to simplify a layout and to extract an accurate substrate *RC* network. Moreover, the signal integrity issue in Systems-on-Chip–SoC-or Systems-in-Package–SiP-does not only concern conducted noise but also radiated propagation with electromagnetic effects: A modified ICEM standard model, with a general frequency-dependent RCL network could also answer to this need.

References

[1] V.F. Pavlidis, E.G. Friedman, Three-Dimensional Integrated Circuit Design, Elsevier Inc., 2009.

[2] M. Rousseau, M. Jaud, P. Leduc, A. Farcy, A. Marty Impact of substrate coupling induced by 3D-IC architecture on advanced CMOS technology, in: Proc. of the European Microelectronics and Packaging Conf., Italy, 2009.

[3] B.R. Stanisic, N.K. Verghese, R.A. Rutenbar, L.R. Carley, D.J. Allstot, Addressing substrate coupling in mixed-mode IC's: simulation and power distribution synthesis, IEEE Journal of Solid-State Circuits 29 (3) (1994) 226-238.

[4] M. Nagata, J. Nigai, K. Hijikata, T. Morie, A. Iwata, Physical design guides for substrate noise reduction in cmos digital circuits, IEEE Journal of Solid-State Circuits 36 (3) (2001) 539-549.

[5] J. Briaire, K.S. Krisch, Principles of substrate crosstalk generation in CMOS circuits, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 19 (6) (2000) 645-653.

[6] Sentaurus, Device Electromagnetic Wave Solver, User Guide, Version A-2008.09, SYNOPSYS.

[7] K.S. Yee, Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media, IEEE Transactions on Antennas and Propagation 14 (1966) 302-307.

[8] M. Felder, J. Ganger, Analysis of ground-bounce induced substrate noise coupling in a low resistive bulk epitaxial process: design strategies to minimize noise effects on a mixed-signal chip, IEEE Transactions on Circuits and Systems 46 (11) (1999) 1427-1436.

[9] M. Badaroglu, G.V. Plas, P. Wambacq, L. Balasubramanian, K. Tiri, I. Verbauwhede, S. Donnay, G.G.E. Gielen, H.J.D. Man, Digital circuit capacitance and switching analysis for ground bounce in ICs with a high-ohmic substrate, IEEE Journal of Solid-State Circuits 39 (2004) 1119-1130.

[10] R. Singh, A review of substrate coupling issues and modeling strategies, in: Proceedings of the IEEE 1999 Custom Integrated Circuits Conference, 1999, pp. 491-499.

[11] International Electro-Technical Commission, IEC 62014-3: Integrated Circuits Emission Model (ICEM), Draft Technical Report, IEC, Nov. 2004.

[12] A. Papoulis, S.U. Pillai, Probability, Random Variables, and Stochastic Process, 3rd ed., New-York, McGraw-Hill, 1991.

[13] SPICE: Simulation Program with Integrated Circuit Emphasis, University of California, Berkeley.

[14] A. Demir, P. Feldmann, Modeling and simulation of the interference due to digital switching in mixed-signal ICs, in: Proc. IEEE/ACM International Conference on Computer-Aided Design, November, 1999.

[15] A. Demir, J. Roychowdhury, P. Feldmann, Modelling and simulation of noise in analog/mixed-signals communications systems, in: Proceedings of the IEEE 1999 Custom Integrated Circuits Conference, 1999, pp. 385-393.

[16] G. Bilardi, R. Padovani, G.L. Pierobon, Spectral analysis of functions of markov chains with applications, IEEE Transactions on Communications COM-31 (1983) 853-861.

[17] H. Yasuda, Calculation of power spectra of pulses sequences by means of transition matrices, Electronics and Communications in Japan 54-A (1971).

Computer Technology and Application 2 (2011) 801-812

Investigation of Differently Modulated Optical Signals Transmission in HDWDM Systems

Aleksejs Udalcovs, Vjaceslavs Bobrovs and Girts Ivanovs

Institute of Telecommunications, Faculty of Electronics and Telecommunications, Riga Technical University, Riga LV-1048, Latvia

Received: August 14, 2011 / Accepted: August 29, 2011 / Published: October 25, 2011.

Abstract: The authors' developed combined system model can be considered under the concept of next generation optical network (NGON) as a model for the future design of backbone networks. Such solution can be topical in the result of different telecom operators' optical networks convergence. In this case a necessity to transmit differently modulated signals over a single optical fiber even with different bit rates may occur. This research is performed with OptSim 5.2 simulation software that numerically solves nonlinear Schrödinger equation. The authors have revealed the optimal parameter configuration for developed combined transmission systems and obtained in system's channels detected signals bit-error-rate (BER) correlation diagrams. They represent BER as a function from different system's parameters such as channel output power level, optical amplifier fixed output power level and system's channels allotment in C-band of ITU-T (Telecommunication Standardization Sector of the International Telecommunications Union) recommended spectral grid. As well as these obtained BER values were compared with the results for similar system, where instead of standard single mode fiber (according ITU-T Rec. G.652 D) optical signals are transmitted over non-zero dispersion shifted fiber (ITU-T Rec. G. 655).

Key words: Wavelength division multiplexing (WDM), modulation formats, bit-error-rate.

1. Introduction

Within the last few years strongly arises demand of transmission systems' channels information throughput. This trend is observed mainly due to rising number of worldwide internet user and data volume itself that is requested per user [1-3]. New information services including data, online and broadband services, such as online video conferences and video on demand, and their rapid advance in modern information age only contributes to this trend of increase of demand for information capacity [4]. In order to satisfy and secure appropriate quality of service (QoS) and service level agreement (SLA) required bandwidth for one channel currently is being doubled within two year period [5].

Wavelength division multiplexing (WDM) technology for fiber optic transmission systems has been developed and introduced in order to make use of approximately 60 THz bandwidth that is offered by silica optical fibers [1, 3, 6]. In recent years total information carrying capacity of transmission systems was increased for the account of channels number, channel spacing and per channel bit rates. However in this case must take into the account total amount of optical power coupled into a fiber. Coupled power increase resulting in additional transmission impairments caused by nonlinear optical effects (NOE) and its combination with linear distortion mechanisms [2]. Generally it leads to distorted transmission in some channels of fiber optic transmission system (FOTS) or even to complete failure of system's channels. It means that informative signals cannot be detected on the other end of the fiber with a required

Vjaceslavs Bobrovs, assistant professor, senior researcher, Dr.Sc.Ing., research field: telecommunications.

Girts Ivanovs, professor, Dr.Sc.Ing., research field: telecommunications.

Corresponding author: Aleksejs Udalcovs, researcher, Ph.D. student, research field: telecommunications. E-mail: aleksejs.udalcovs@rtu.lv.

error probability. This reduces system's total carrying capacity and channel's data throughput. Consequently, another system's total transmission capacity increment solution must be found out.

Currently one of the most intensively studied system's total transmission capacity increment solutions is the increasing of system's channel spectral efficiency. Actually it is more efficient utilization of available bandwidth. It means that more informative bits are transmitted using one hertz from available frequency band. It ensures that a smaller number of channels must be used to transmit the same amount of informative bits. Channel's spectral efficiency can be increased in three different ways. The first one, the reduction of used system's channel spacing. This means that a larger number of transmission channels can be allocated in available frequency band. The second one, the increase of per channel bit rate maintaining previously used channel spacing values for separation of transmission channels. And finally the third one is the combination of pervious two ways.

Obviously that it is easier to achieve a larger channel's spectral efficiency if for optical signal modulation and coding some of novel modulation formats are used. This novel (or advanced) modulation formats provide narrower optical signals spectrum or multilevel encoding schemes that ensure more bits per one symbol than it is in traditional modulation formats, for example, on-off keying (OOK) with non-return to zero (NRZ) encoding format (NRZ-OOK) [7-8]. Maximal spectral efficiency, which can be obtained with traditional OOK modulation formats, is about 0.4 bit/s/Hz [7]. It has been reported in Ref. [9-11] that using such novel modulation formats as quadrature amplitude modulation (16-QAM particularly) and orthogonal frequency-division multiplexing(OFDM) together with polarization division multiplexing (PDM) technique it can be achieved SE larger than 6 bit/s/Hz and even reaches 7 bit/s/Hz.

Our study object of this paper is optimal combined WDM system configuration that provides lowest in system's channels detected signals BER values. This developed combined WDM system's model is offered for the future design of backbone optical networks and can be considered under the concept of NGON. Chosen optical signal modulations formats and per channel bit rates, according to authors' thoughts, are the most appropriate and probable at this moment. It was concluded after careful evaluation of current state of optical telecommunication networks, their possible and the most likely development strategy and trends in the future [2, 12-13]. The paper is organized as follows: Section 2 developed combined high density WDM (HDWDM) system model is described as well as simulation scheme and strategy is revealed; section 3 describes the method of measurements and here authors evaluate accuracy of the results obtained by simulations; in section 4, results of this research are discussed and section 5 contains main conclusions.

2. Simulation Model and Schemes

Previously in Ref. [2] as combined FOTS model have been offered three-channel WDM system (see Fig. 1), where three different modulation formats are used for optical carrier signal modulation: differential phase shift keying (DPSK) with non-return-to-zero encoding (NRZ-DPSK), orthogonal binary polarization shift keying (2-POLSK) and on-off keying with NRZ encoding (NRZ-OOK).

In addition in Ref. [2] have been detected that if in such combined system 50, 75 or 100 GHz channel spacing are used for channel separation, then the obtained channels' BER values corresponding to BER values, which are obtained for systems, where only one

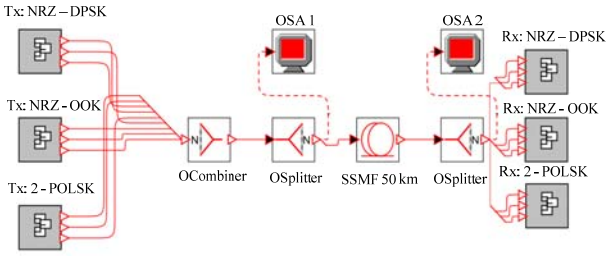
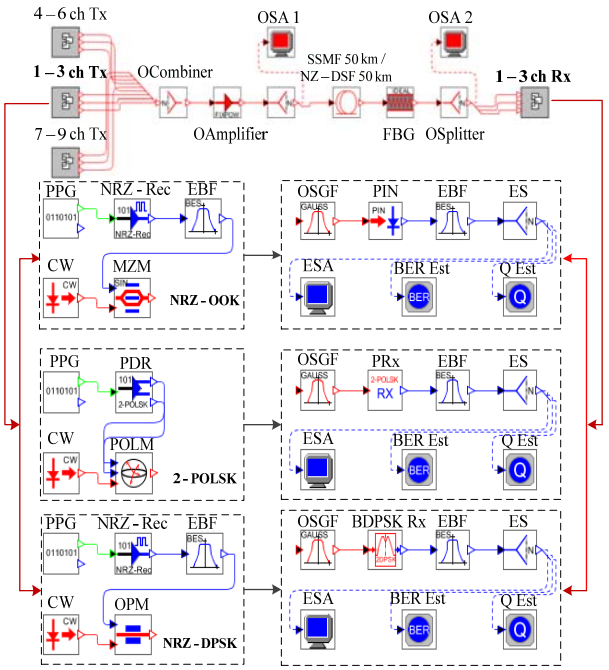


Fig. 1 Purposed model of three-channel combined WDM system.

these formats is used for the optical signal modulation. It's approximately equal to: 10^{-17} for NRZ-DPSK channel, 10^{-25} for NRZ-OOK channel and 10^{-40} for 2-POLSK. While if in three-channel WDM system with 10 Gbit/s per channel bit rate less than 50 GHz channel spacing is used for channel separation then become evident combined transmission features. And in this case obtained BER value for detected signals in combined WDM system depends not only from modulation format, which is used of optical signal modulation in given channel, but also form modulation formats in adjacent channels.

To identify channel that is a source of larger amount of interchannel crosstalk noise than the rest of system's channels six combined systems were investigated. These systems differ from each other only with modulation formats distribution among channels. This distribution scheme is as follows: [NRZ-DPSK (1, 2, 3, 3, 2, 1)]-[NRZ-OOK (2, 1, 1, 2, 3, 3)]-[2-POLSK (3, 3, 2, 1, 2)]. This configuration represents modulation format and channel's number where one on these formats is used. The system's channels central frequencies are anchored to 193.1 THz according to ITU-T Recommendation G.694.1 and the first channel's central frequency is equal to 193.075 THz, the second—193.100 THz and the third—193.125 THz. After this crosstalk source have been detected simulation model were updated in order to find out the optimal modulation format distribution, which provides the lowest in system's channels detected signals' BER values. For this purpose existing transmission systems model were updated to nine-channel WDM system. These channels are grouped by three and these groups have identical transmitter and receiver blocks configuration but with different channels' central wavelengths. It was specially done to take into account linear and nonlinear crosstalk influences to optical signal transmission which are experience central's group channel (1^{st}-3^{rd}) from adjacent groups (4^{th}-6^{th} and 7^{th}-9^{th}). For system's further analysis we will use only channels number 1-3, but 4-6 and 7-9 are used as sources of interchannel crosstalk (see Fig. 2).

Then NRZ-DPSK, 2-POLSK and NRZ-OOK modulated optical signals are combined, optically preamplified with fixed output power erbium-doped fiber amplifier (EDFA) and send over 50 km of single mode optical fiber. There are two different types of single mode fiber used in this research: standard single mode fiber or SSMF (according to ITU-T Recommendation G.652 D) and non-zero dispersion shifted fiber or NZ-DSF (according to ITU-T Recommendation G.655). Then optical signals are filtered with Super Gaussian optical filters, converted to electrical signals and then electrically filtered using Bessel electrical filters. Fiber span length was chosen equal to 50 km in order to avoid increase of amplified spontaneous emission (ASE) noise. Larger amplifier spacing would require gain greater than 10 dB, but this in a prohibitive leads to growth of ASE noise [14].



PPG – Pulse Pattern Generator
CW – Continuous Wavelength
MZM – Mach-Zehnder Modulator
POLM – Polarization Modulator
OPM – Optical Phase Modulator
EBF – Electrical Bessel Filter
OSGF – Optical Super Gaussian Filter
ES - Electrical Splitter
BDPSK Rx – Balanced DPSK Receiver
PRx – POLSK Receiver
PIN – PIN Photodiode
Q Est – Q factor Estimator
BER Est – Bit-Error-Rate Estimator
ESA – Electrical Signal Analyser

Fig. 2 Developed combined WDM system model and its block scheme in OptSim software.

In addition for evaluation of system performance and its further analyze such measurements as detected signals eye diagrams and system's output optical spectrum in the beginning and in the end of optical link were obtained.

3. Measurement Technique and Accuracy

This research is based on powerful and accepted mathematical simulation software OptSim 5.2. It solves complex differential nonlinear Schrödinger equation (NLSE) using split-step Fourier method (SSFM). This equation describes optical signal propagation over the fiber and can be written as Eq. (1) [3]:

$$\frac{\partial}{\partial z}\cdot A+\frac{\alpha^l}{2}\cdot A+j\cdot\frac{\beta_2}{2}\cdot\frac{\partial^2}{\partial t^2}\cdot A-\frac{\beta_3}{6}\frac{\partial^3}{\partial t^3}\cdot A= \\ =j\cdot\gamma\cdot|A|^2\cdot A \quad (1)$$

where *A(t, z)* is complex optical field; *z*—fiber length, [km]; α^l is linear attenuation coefficient of an optical fiber, [km^{-1}]; β_2 is the second order parameter of chromatic dispersion, [ps^2/nm]; β_3 is the third order parameter of chromatic dispersion, [ps^3/nm]; γ is nonlinear coefficient, [W^{-1}·km^{-1}]; *t*—time, [s]. NLSE takes into the account linear and nonlinear affects and they influence to optical signal distortions. The principle of split-step method is better illustrated by (1), which can be written as follows [3, 15]:

$$\frac{\partial}{\partial z}\cdot A(t,z)=\left(\widehat{D}+\widehat{N}\right)\cdot A(t,z) \quad (2)$$

$\widehat{D}$ is linear operator responsible for linear effects such as dispersion and attenuation [3]:

$$\widehat{D}=-\frac{\alpha^l}{2}-j\cdot\frac{\beta_2}{2}\cdot\frac{\partial^2}{\partial t^2}+\frac{\beta_3}{6}\cdot\frac{\partial^3}{\partial t^3} \quad (3)$$

$\widehat{N}$ is nonlinear operator, which takes into account Kerr and other nonlinear effects (NOEs) [3]:

$$\widehat{N}=j\cdot\gamma\cdot|A|^2\cdot A \quad (4)$$

In general split-step method is based on assumption that linear and nonlinear effects affect optical signals independently. This statement can be considered as true if we assume that all fiber length *z* is being divided into sufficiently small spans *Δz*, and only then these linear and nonlinear effects by turns are taken into account for each $\frac{\Delta z}{2}$ segment.

There are two basic algorithms for realization of SSFM: Time domain split step (TDSS) and frequency domain split step (FDSS). These two algorithms differ only with an approach that is being used for calculation of linear operator $\widehat{D}$. While in both cases nonlinear operator $\widehat{N}$ is being calculated in time domain [15].

Operator $\widehat{D}$ is being fully characterized by its impulse response *h(t)* and it is mathematically correct to calculate its influence to *A(t, z)* optical field using products of mathematical convolution. In TDSS case it can be written as follows [15]:

$$A_L[n]=A[n]*h[n]=\sum_{k=-\infty}^{\infty}A[k]\cdot h[n-k] \quad (5)$$

This algorithm calculates this convolution in time domain and precisely obtains time delay values between signals with different wavelength. In OptSim software this TDSS algorithm is realized using finite impulse reaction (FIR) filters. This sophisticated technique provides complete control of an overall mistake that may occur during all process of calculating. By contrast FDSS calculates $\widehat{D}$ in frequency domain but firstly for this algorithm is necessary to calculate fast Fourier transformation (FFT) from *A[n]* signal samples and from *h(t)* impulse reaction. Then it is necessary to use invers FFT (FFT^{-1}) to convert obtained data array to time sample domain. FDSS algorithm can be mathematically described using following equation [15]:

$$A_L'[n]=A[n]\otimes h[n]= \\ =FFT^{-1}\{FFT(A[n])\times FFT(h[n])\} \quad (6)$$

As one can see, then in this case circular convolution is used for obtaining signal sample array $A_L'[n]$. This array may contain fewer samples than it is necessary to obtain actual convolution products—$A_L[n]$ sample array. Hence this algorithm is easier to implement than TDSS and it requires less computation time and resources but serious errors may occur during calculation [3].

Eq. (2) can be solved assuming that $\widehat{D}$ and $\widehat{N}$ operators are independent and fiber span Δz length is small enough (5-100 m depending on the simulation accuracy requirements). Then optical signal after propagation over Δz span can be described in the following manner [15]:

$$A(t, z + \Delta z) \cong \exp\left[\frac{\Delta z}{2} \cdot \widehat{D}\right] \cdot \exp\left\{\Delta z \cdot \widehat{N}\left[A\left(t, z + \frac{\Delta z}{2}\right)\right]\right\} \cdot \exp\left(\frac{\Delta z}{2} \cdot \widehat{D}\right) \cdot A(t, z) \quad (7)$$

For the evaluation of system performance will be used such parameter as Q-factor and BER value. Q = 7.03 (16.94 dB) corresponds to the commonly used reference for BER of 10^{-12}.

Q-factor uncertainty range (see Fig. 3) and BER confidence interval magnitude depends on the total number of simulated bits N_{total} [15]:

$$\mathrm{dev}[Q^*] \equiv \sigma_Q \cong \frac{Q}{\sqrt{2 \cdot N_{total}}} \quad (8)$$

where Q is Q-factor value that can be calculated using following Eq. (9) [3, 15]:

$$\mathrm{Q} = \frac{|\mu_1 - \mu_0|}{\sigma_1 + \sigma_0} \quad (9)$$

where $\mu_{1,0}$ and $\sigma_{1,0}$ are the mean and the standard deviation of the received signal, when a logical "1" and "0" is transmitted, and $\pi \approx 3.14$ [3, 15].

$$\mathrm{BER} = \frac{1}{2} \cdot \mathrm{erfc}\left(\frac{Q}{\sqrt{2}}\right) \quad (10)$$

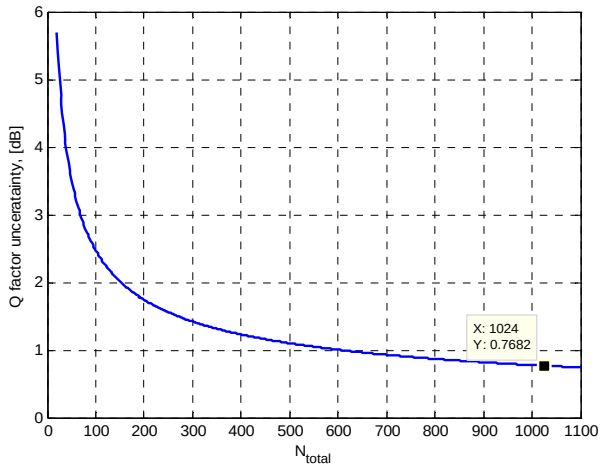


Fig. 3 Q-factor uncertainty as a function form total number of simulated bits.

Using Eq. (8) Q-factor uncertainty range can be expressed as Ref. [15]:

$$\mathrm{range} = 20 \cdot \log_{10}\left(\frac{1 + \sqrt{\frac{2}{N_{total}}}}{1 - \sqrt{\frac{2}{N_{total}}}}\right) \quad (11)$$

As one can see from Fig. 3 Q-factor uncertainty range for 1,024 simulated bits that is used in our schemes is equal to 0.77 dB.

Q-factor and BER value 95% ($\pm 2\sigma_Q$) confidence intervals for 16.94 dB nominal value can be obtained using Eq. (8) and Eq. (10) assuming that we are dealing with Gaussian distribution. For 1,024 of simulated bits these intervals are

$$\mathrm{Q}_{\text{for 16.94 dB}} \in [16.55; 17.31], \text{[dB]} \quad (12)$$

$$\log 10\{\mathrm{BER}_{\text{for } 10^{-12}}\} \in [-12.97; -11.04] \quad (13)$$

As an example we will give BER 95 % confidence interval as a function from the number of simulated bits for 10^{-12} nominal. This value will be used as reference for transmission channel with 10 Gbit/s per channel bitrate (see Fig. 4).

As one can see, when simulating 1,024 bits at BER = 10^{-12}, the confidence interval magnitude is less than ±1 order. It points to the conclusion that OptSim simulation software allows obtaining sufficiently accurate preliminary results and there is no point to increase the total number of simulated bits, because obtained results accuracy does not improve sufficiently.

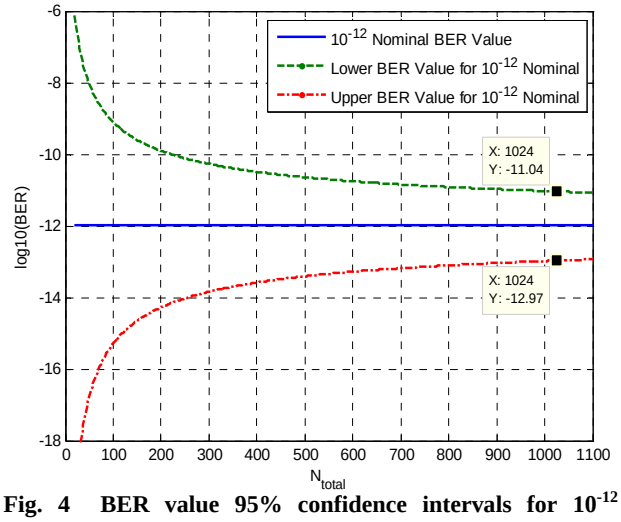


Fig. 4 BER value 95% confidence intervals for 10^{-12} nominal.

4. Results and Discussion

The aim of this paper was to investigate optimal configuration for combined WDM systems where differently modulated optical signals are transmitted. To achieve this goal several objectives must be solved.

Firstly it is necessary to identify channel in [1st: NRZ-DPSK, 10 Gbit/s, 193.075 THz]-[2nd: 2-POLSK, 10 Gbit/s, 193.100 THz]-[3rd: NRZ-OOK, 10 Gbit/s, 193.125 THz] combined WDM FOTS that is a source of larger amount of interchannel crosstalk noise than the rest of channels.

For this purpose six different systems were studied. These systems differ from each other only with modulation formats that are used in each particular system's channels. These systems have following configurations:

[NRZ-DPSK]-[NRZ-OOK]-[2-POLSK];
[NRZ-OOK]-[NRZ-DPSK]-[2-POLSK];
[NRZ-OOK]-[2-POLSK]-[NRZ-DPSK];
[2-POLSK]-[NRZ-OOK]-[NRZ-DPSK];
[2-POLSK]-[NRZ-DPSK]-[NRZ-OOK];
[NRZ-DPSK]-[2-POLSK]-[NRZ-OOK].

In each systems channel were determined signals BER values that further were used for system's performance analyze. The obtained results are summarized below (see Table 1). Using these results for each configuration system's average detected signals BER values were calculated. As one can see, sufficiently smaller BER value is for the third combined system configuration then it is for the rest of possible configuration. The third configuration is as follows: [1st: NRZ-OOK\193.075 THz]-[2nd: 2-POLSK\193.100 THz]-[3rd: NRZ-DPSK\193.125 THz]. After careful analysis of these obtained results it was found that investigated combined system channel, where for optical signals modulation NRZ-DPSK modulation format is used, is a source of larger amount of interchannel crosstalk than channels, where NRZ-OOK or 2-POLSK format is used. This was concluded based on obtained NRZ-OOK and 2-POLSK channels BER results for different system's configurations. This become evident if we analyze obtained BER values for the fourth, fifth and sixth system.

Table 1 BER values for different combined systems channels.

f (THz)	1st system	2nd system	3rd system
193.075	3×10^{-24}	2×10^{-8}	1×10^{-40}
193.100	9×10^{-12}	9×10^{-25}	1×10^{-18}
193.125	1×10^{-40}	1×10^{-13}	3×10^{-27}
Average	3×10^{-12}	7×10^{-9}	5×10^{-19}
f (THz)	4th systems	5th system	6th system
193.075	1×10^{-40}	1×10^{-21}	4×10^{-27}
193.100	2×10^{-8}	6×10^{-25}	8×10^{-14}
193.125	3×10^{-25}	6×10^{-12}	1×10^{-40}
Average	5×10^{-9}	2×10^{-12}	3×10^{-14}

Firstly let's focus to the fifth system's BER values. As one can see from this configuration scheme then in this case NRZ-DPSK modulated optical signals are transmitted in central system's channel. As a result detected signals BER values in adjacent channels are sufficiently higher than they are in cases, when NRZ-OOK or 2-POLSK modulated optical signals are located further from NRZ-DPSK channel as it is in the sixth system. Comparing BER results obtained for 2-POLSK modulated signals in the fourth and sixth system (1×10^{-40} and 9×10^{-14} respectively), we can conclude that in combined system detected signals BER value decreases if channel, where these signals are transmitted, is located further from NRZ-DPSK channel.

To assess NRZ-DPSK channel created crosstalk impact to optical signals transmission in all others combined system's channels previously studied three-channel combined systems model (see Fig. 1) was modified and supplemented with 2 × 3 channels that have appropriate system's configuration (see Fig. 2). As before, in system channels detected signals BER values were obtained for six different combined system configurations (see Table 2).

As well as using these data two different channels average BER values were calculated: system's average BER that takes into account all system channels (1st-9th); central group channels' average BER that takes into account only channels number one to three.

Table 2 BER values for different 9-channel combined systems' channels.

No.	f (THz)	1st system	2nd system	3rd system
4	193.000	5×10^{-23}	2×10^{-11}	1×10^{-40}
5	193.025	5×10^{-11}	4×10^{-24}	1×10^{-11}
6	193.050	1×10^{-13}	3×10^{-16}	9×10^{-24}
1	193.075	4×10^{-22}	2×10^{-8}	1×10^{-10}
2	193.100	5×10^{-10}	2×10^{-23}	4×10^{-15}
3	193.125	3×10^{-16}	3×10^{-13}	4×10^{-23}
7	193.150	4×10^{-24}	8×10^{-8}	1×10^{-11}
8	193.175	7×10^{-9}	2×10^{-23}	4×10^{-17}
9	193.200	1×10^{-40}	1×10^{-14}	4×10^{-29}
Average (1st-3rd)		2×10^{-10}	7×10^{-9}	4×10^{-11}
Average (1st-9th)		9×10^{-10}	1×10^{-8}	2×10^{-11}
No.	f (THz)	4th systems	5th system	6th system
4	193.000	1×10^{-40}	3×10^{-31}	9×10^{-27}
5	193.025	3×10^{-11}	2×10^{-23}	6×10^{-15}
6	193.050	3×10^{-24}	2×10^{-11}	3×10^{-11}
1	193.075	1×10^{-11}	2×10^{-15}	4×10^{-23}
2	193.100	1×10^{-8}	3×10^{-26}	1×10^{-14}
3	193.125	4×10^{-22}	3×10^{-10}	5×10^{-10}
7	193.150	1×10^{-12}	2×10^{-15}	3×10^{-22}
8	193.175	3×10^{-8}	5×10^{-24}	2×10^{-13}
9	193.200	1×10^{-23}	3×10^{-12}	1×10^{-40}
Average (1st-3rd)		3×10^{-9}	9×10^{-11}	2×10^{-10}
Average (1st-9th)		4×10^{-9}	3×10^{-11}	5×10^{-11}

As one can see from obtained data (see Table 2), then the lowest average BER values for 1st till 3rd and 1st till 9th channel are for the third combined system configuration and they are equal to $BER_{1st\text{-}3rd} = 4 \times 10^{-11}$ and $BER_{1st\text{-}9th} = 2 \times 10^{-11}$ respectively. But the highest BER values are for the second configuration and they are equal to $BER_{1st\text{-}3rd} = 7 \times 10^{-9}$ and $BER_{1st\text{-}9th} = 1 \times 10^{-8}$. So, BER difference between the best and worst case scenario, corresponding to [(NRZ-OOK)\193.075 THz]-[(2-POLSK)\193.100 THz]-[(NRZ-DPSK)\193.125 THz] and [(NRZ-OOK)\193.075 THz]-[(NRZ-DPSK)\193.100 THz]-[(2-POLSK)\193.125 THz] configuration respectively, is approximately three orders (see Figs. 5-6).

In these both cases channel with highest detected signal error probability is the first one, where by the way NRZ-OOK modulated optical signals are transmitted. Comparing BER values obtained for NRZ-OOK and 2-POLSK modulated optical signals

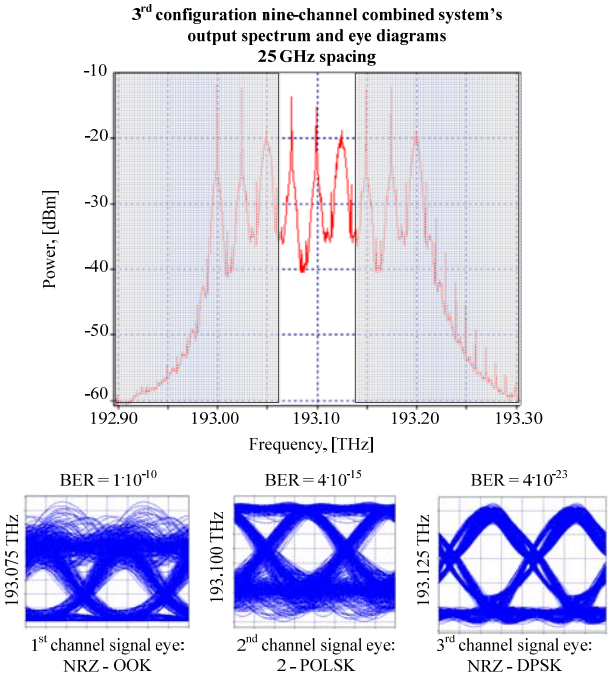


Fig. 5 Nine-channel combined system's with the third configuration output spectrum and eye diagrams in case of 10 Gbit/s per channel bitrates and 25 GHz channel spacing.

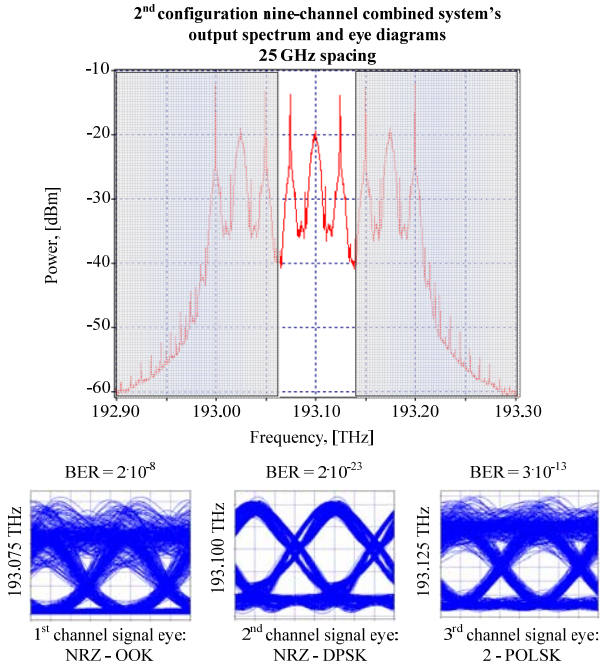


Fig. 6 Nine-channel combined system's with the second configuration output spectrum and eye diagrams in case of 10 Gbit/s per channel bitrates and 25 GHz channel spacing.

for these two systems configuration, we have to conclude that these values differ by no more than two orders (1×10^{-10} and 2×10^{-8} in NRZ-OOK case and 4

$\times 10^{-15}$ and 3×10^{-13} for 2-POLSK channels). As for NRZ-DPSK channel then the resulting BER values differences in both cases are not significant: 4×10^{-23} and 2×10^{-23} (see the 1st channel in Fig. 5 and the 2nd channel in Fig. 6).

As a result, for further research of optimal combined system configuration will be used as a starting point nine-channel combined WDM system with the third configuration.

Previously it has been detected that channel, where NRZ-DPSK modulated optical signals are transmitted, is larger amount of interchannel crosstalk source than NRZ-OOK or 2-POLSK channels. So, to reduce that type of noise it has been decided to decrease optical power level radiated by distributed feedback lasers (DFB) in continuous wavelength (CW) regime that are used in these channels. Using OptSim Scan Parameter simulation regime in system's channel detected signals BER value as function from NRZ-DPSK channel laser output power level was obtained (see Fig. 7).

As previously, using these BER results for each system channel average BER value for central channels were calculated. It revealed that in system channels detected signals average BER values are below 10^{-12} if NRZ-DPSK channels' lasers output power level is in the range from 3.5 to 4.5 dBm. The lowest average channels' BER value is reached if these lasers output power is equal to 3.5 dBm. In this case $BER_{1st\text{-}3rd} = 3 \times 10^{-18}$ and the worst channel is the second one (2-POLSK) and its $BER_{2nd} = 1 \times 10^{-17}$ (see Fig. 7).

Assuming that we are dealing with one sector of ultra-long haul backbone optical network, it was decided to supplement this model of combined WDM system with additional optical element—fixed output power optical amplifier. It allowed take into an account ASE noise arising from EDFA which is the most widely used optical amplifier. To find out optimal amplifier output power level, that provides minimal channels' BER values, BER correlation diagram for each were obtained. It represents in systems channels detected signals BER values as a function from amplifier fixed output power level (see Fig. 8). Let us note that, in this case NRZ-DPSK channel laser output power level remains unchanged as it was in initial combined WDM system model in Ref. [2].

As one can see from Fig. 8, then BER value for the system's first channel varies around 10^{-11} value, for the second channel around 10^{-16} and for the third—10^{-24}. Knowing that the worst combined system's channel is the second one, where 2-POLSK modulated optical signals are transmitted, then was decided to choose amplifier output power level that provides minimal BER value exactly in this channel.

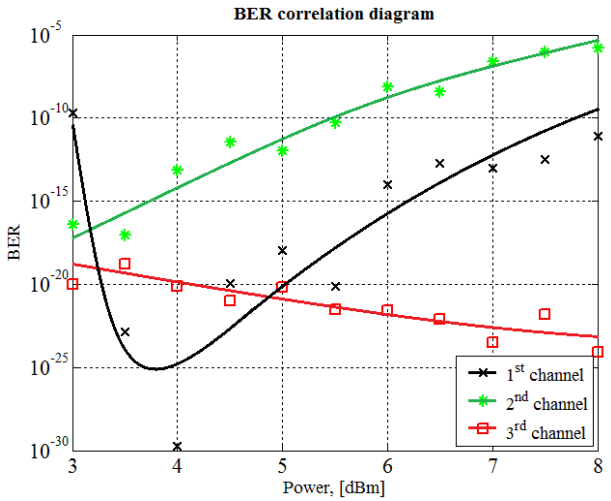


Fig. 7 BER as a function from NRZ-DPSK channel laser radiated output power level.

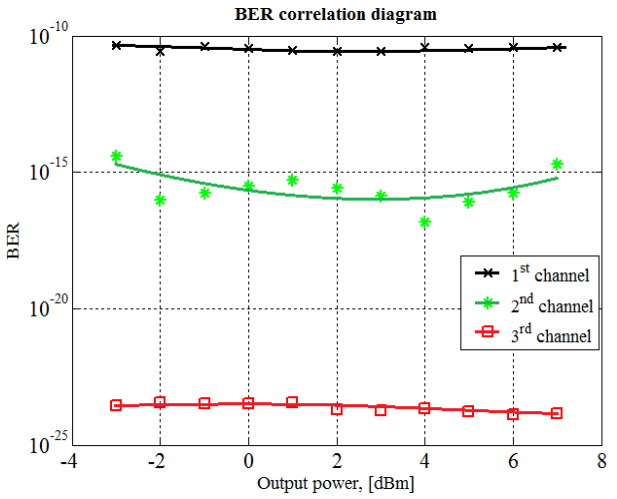


Fig. 8 BER as a function form optical amplifier fixed output power level.

Consequently, optical amplifier fixed output power level equal to 4 dBm was chosen. This level provides in the second system's channel detected signals $BER_{2nd} = 1 \times 10^{-17}$.

If both these optimal parameters are used in combined system model configuration, then in system channels detected signals BER values are well below the maximal acceptable BER threshold 10^{-12}, that is defined for 10 Gbit/s per channel bitrate (see Table 3). Channels BER values for revealed optimal system configuration were obtained for two types of single mode fiber. The first fiber was standard single mode fiber (SSMF) according to ITU-T Recommendation G.652 D and the second was non-zero dispersion shifted fiber (NZ-DSF) according to ITU-T Recommendation G.655.

For these two cases BER results as well as detected signals eye diagrams were compared one to another (see Fig. 9).

To summarize results that was reported in Table 3, as well as in Fig. 9, must be stated the fact that in system channels detected signals BER values decreased significantly comparing to initial unoptimized system.

BER value for the first system channel sufficiently dropped from 1×10^{-10} to 5×10^{-26} if SSMF is used and to 9×10^{-20} if NZ-DSF is used. Exactly for this first channel, where NRZ-OOK modulated optical signals are transmitted, experiencing the most radical BER value improvement comparing to 2-POLSK and NRZ-DPSK channels. In these channels detected signal BER values do not improve so noticeably (see eye diagrams in Figs. 5 and 9). The second channel's BER value decreases from 4×10^{-15} to 7×10^{-19} for SSMF and to 6×10^{-21} for NZ-DSF, but the third channel's BER value variation is not essential from 4×10^{-23} to 6×10^{-22} if SSMF is used. But if in this system instead of SSMF NZ-DSF is used then it is possible to obtain lower BER values for NRZ-DPSK channels. In this channel detected signals BER value decreases to 1×10^{-40}.

Table 3 BER value for different system configurations.

System description	1st	2nd	3rd
Initial	1×10^{-10}	4×10^{-15}	4×10^{-23}
$P_{NRZ\text{-}DPSK} = 3.5$ dBm	1×10^{-23}	1×10^{-17}	2×10^{-19}
FIXPOWER = 4 dBm	4×10^{-11}	1×10^{-17}	2×10^{-24}
Optimal and SSMF	5×10^{-26}	7×10^{-19}	6×10^{-22}
Optimal and NZ-DSF	9×10^{-20}	6×10^{-21}	1×10^{-40}

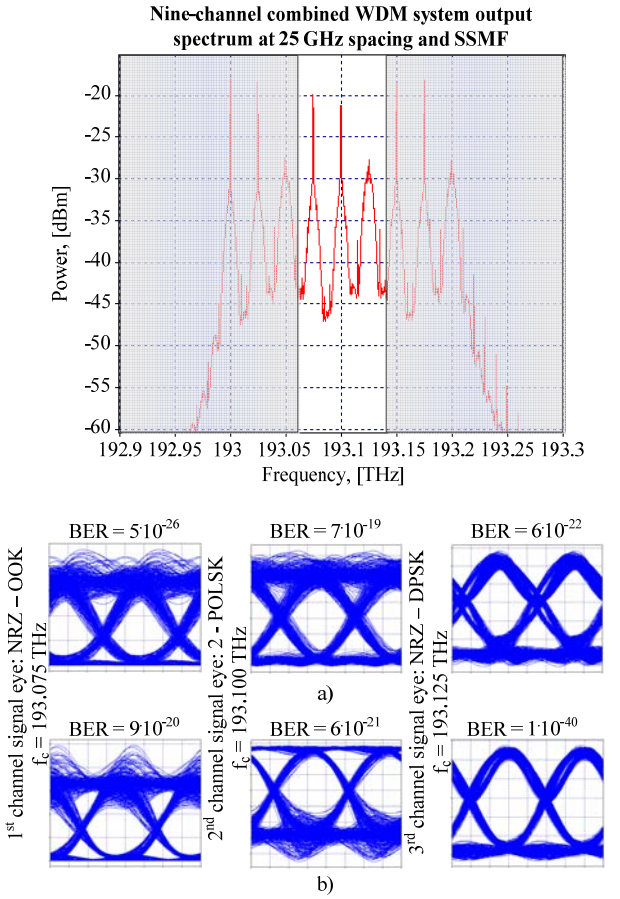


Fig. 9 Optimal configuration nine-channel combined WDM system output optical spectrum and eye diagrams: (a) SSMF; (b) NZ-DSF.

In addition, coherence between detected signals BER values and channels' central frequency position in C-band (191.6-195.9 THz) was investigated. As well as, for each system channel the worst and the best position in C-band, that provides the highest and the lowest possible signals BER values, respectively, for previously found optimal combined system's configuration, was revealed (see Table 4). As previously, this research was held for two types of optical fiber: SSMF (G.652 D) and NZ-DSF (G.655).

Table 4 Best and worst channels positions in C-band and their BER values.

No.	Fiber	B/W	f_c (THz)	BER
1st	SSMF (G.652 D)	Best	195.725	8×10^{-26}
		Worst	192.350	2×10^{-16}
2nd		Best	193.950	3×10^{-31}
		Worst	192.150	2×10^{-20}
3rd		Best	195.100	8×10^{-25}
		Worst	192.400	7×10^{-19}
1st	NZ-DSF (G.655)	Best	193.475	1×10^{-39}
		Worst	193.925	5×10^{-26}
2nd		Best	191.925	2×10^{-23}
		Worst	195.750	3×10^{-14}
3rd		Best	C-band	1×10^{-40}
		Worst		

It showed that depending on channel central frequency the first channel's BER value varies around nominal value of 10^{-20} if SSMF is used and around 10^{-30} if NZ-DSF. But BER values obtained for the second channel and NZ-DSF are for several orders worse comparing to transmission over SSMF. They vary around 10^{-20} and 10^{-25}, respectively.

In addition, as one can see from Table 4, then BER value obtained for NZ-DSF and the worst case of 2nd channel central frequency is approximately for six orders larger comparing to the worst case of SSMF. These let us conclude that 2-POLSK modulated signals are not suitable for transmission over NZ-DSF fiber in [NRZ-OOK]-[2-POLSK]-[NRZ-DPSK] combined WDM systems and they are sufficiently distorted at appropriate channel central frequency. As result system's average BER value is significantly higher than it might be in case of SSMF (see Table 5). In same time the lowest 1st-3rd channels average BER value is gained with NZ-DSF and it is about two orders lower than in case of SSMF. As for the third channel, then in case of SSMF and different channel central frequencies obtained BER values vary somewhere around 10^{-23}. Whereas in case of NZ-DSF these values remain constant and approximately equal to 10^{-40} in all C-band. It allows to judge about NRZ-DPSK modulated optical signals transmission suitability over NZ-DSF single mode optical fiber.

Table 5 Minimal and maximal in system's 1st-3rd channel detected signals BER values.

Characteristics	BER	2nd channel f_c (THz)	1st-3rd channel
MIN average BER/ SSMF	3×10^{-23}	195.750	8×10^{-26}
			2×10^{-24}
MAx average BER/ SSMF	6×10^{-17}	192.375	8×10^{-23}
			2×10^{-16}
			7×10^{-28}
MIN average BER/ NZ-DSF	5×10^{-25}	192.825	7×10^{-19}
			1×10^{-31}
MAx average BER/ NZ-DSF	1×10^{-14}	195.750	1×10^{-24}
			1×10^{-40}
			1×10^{-27}
			3×10^{-14}
			1×10^{-40}

Let is note, that Corning LEAF non-zero dispersion shifted fiber characteristics and parameters were used in OptSim in order to obtain mathematical model of a NZ-DSF. This fiber is the world's most widely deployed NZ-DSF and is specially optimized for high-speed and high capacity long-haul and metro networks [16].

5. Conclusions

The most appropriate realization of high-density combined WDM fiber optic transmission system configuration for differently modulated optical signals has been investigated. As a model of combined WDM systems have been offered WDM system, where three different optical signal modulation formats are used. This model has following initial configuration: [1st: NRZ-DPSK, 10 Gbit/s, 193.075 THz]-[2nd: NRZ-OOK, 10 Gbit/s, 193.100 THz]-[3rd: 2-POLSK, 10 Gbit/s, 193.125 THz]. In this paper obtained results are summarized below as recommendations and conclusions for the future design of backbone optical networks.

Channels, where NRZ-DPSK modulated optical signals are transmitted, are sources of larger interchannel crosstalk noise comparing to NRZ-OOK or 2-POLSK signals. Consequently, in these channels detected signals BER value decreases if channel, where these signals are transmitted, is located further from

NRZ-DPSK channels. It is possible to minimize NRZ-DPSK created signals crosstalk by reduction the output power of lasers that are used in these channels or by the use of optimal modulation format allocation for particular transmission channel.

Optimal modulation format allocation for particular channel is [(NRZ-OOK)\193.075 THz]-[(2-POLSK)\ 193.100 THz]-[(NRZ-DPSK)\193.125 THz]. It provides in system channels detected signals average BER values not higher than 2×10^{-11}. Whereas [(NRZ-OOK)\193.075 THz]-[(NRZ-DPSK)\193.100 THz]-[(2-POLSK)\193.125 THz] that is the worst possible modulation formats distribution provides BER values not higher than 1×10^{-8}. In these both cases NRZ-OOK modulated signals are the most distorted and as result channels, where in this way modulated optical signals are transmitted, are the worsts combined systems' channels with highest BER values. As well as, detected signals BER values in case of these two combined system's configurations differ for NRZ-OOK and 2-POLSK signals no more than by two orders (1×10^{-10} and 2×10^{-8} ; 4×10^{-15} and 3×10^{-13} respectively) and for NRZ – DPSK this difference is not significant (4×10^{-23} and 2×10^{-23}).

Lasers output power level in the range from 3.5 to 4.5 dBm, that are used in channel, where NRZ-DPSK modulated signals are transmitted, provides average BER value well below necessary threshold of 10^{-12}. If these lasers output power is equal to 3.5 dBm, then the lowest average BER value is obtained and it's approximately equal to 10^{-18}. In this case the worst system's channel is the second (2-POLSK) and it's BER = 10^{-17}. To simulate ASE noise arising from EDFA combined system model was supplemented with fixed output power optical amplifier. Based on obtained channel BER correlation diagrams, this power level was chosen equal to 4 dBm. It provides the lowest possible BER values for 2-POLSK signals detected in the second system's channel (around 10^{-18}) that is the worst channel in developed combined system if SSMF is used. But if optical signals are transmitted over NZ-DSF, then the worst system channel is the first one where NRZ-OOK modulation format is used and its BER varies around 10^{-19}.

Investigating coherence between BER values and channels' central frequencies position in C-band, it have been stated several facts. Firstly, obtained channels central frequency values for each system channel the worst and the best position in C-band, that provide the highest and the lowest possible detected signals BER values, respectively (see Table 4). Secondly, 2-POLSK modulated signals are not suitable for transmission over NZ-DSF fiber in investigated combined WDM system, because these signals are sufficiently distorted at appropriate channel central frequency. If combined system's channels are allocated around nominal of 195.750 THz, then obtained average detected signals BER value is about 10^{-14} and it is the worst possible case. Whereas the lowest possible average BER value (5×10^{-25}) can be reached anchored combined systems channels frequency grid to 192.825 THz. Thirdly, NRZ-DPSK modulated optical signals are well suitable for transmission over NZ-DSF. As result, in these channels detected signals BER values are not higher than 10^{-40} regardless to channel central frequency positioning in C-band.

Acknowledgment

This work has been supported by the European Social Fund within the project "Support for the implementation of doctoral studies at the Riga Technical University" and by the European Regional Development Fund in Latvia within the project Nr. 2010/0270/2DP/2.1.1.1.0/10/APIA/VIAA/002.

References

[1] C. Peucheret, Fibre and component induced limitations in high capacity optical networks, Doctoral Thesis, 2004, pp. 1-8.

[2] A. Udalcovs, V. Bobrovs, G. Ivanovs, Investigation of allowed channel spacing for differently modulated optical signals in combined HDWDM systems, Lithuanian

Journal of Electronics and Electrical Engineering 6 (112) (2011) 19-24.

[3] A.C. Wietfeld, Modeling, simulation and analysis of optical time division multiplexing transmission systems, Doctoral Thesis, 2004, pp. 9-18.

[4] V. Bobrovs, S. Spolitis, A. Udalcovs, G. Ivanovs, Investigation of chromatic dispersion compensation methods for combined HDWDM systems, Latvian Journal of Physics and Technical Sciences 5 (4) (2011) 12-28.

[5] A.K. Dutta, N.K. Dutta, M. Fujiwara, WDM Technologies: Optical Networks, USA, Elsevier Inc., 2004, 336 p.

[6] ITU-T Recommendation G.694.1, Spectral grids for WDM applications: DWDM frequency grid, 06/2002.

[7] C. Xu, X. Liu, X. Wei, Differential phase-shift keying for high spectral efficiency optical transmission, IEEE Journal of Selected Topics in Quantum Electronics 10 (2) (2004) 281-293.

[8] C. Zhang, P. Zhang, L. Geng, A survey of advanced modulation formats for high-speed transmission, in: Proceedings of 2010 International Conference on Computer Application and System Modeling, 2010, Vol. 8, pp. 248-252.

[9] A. Sano, H. Masuda, T. Kobayashi, M. Fujiwara, K. Horikoshi, E. Yoshida, et al., Ultra-high capacity WDM transmission using spectrally-efficient PDM 16-QAM modulation and C- and extended L-band wideband optical amplification, Journal of Lightwave Technology 29 (4) (2011) 578-586.

[10] Y. Miyamoto, Ultra High Capacity Transmission for Optical Transport Network, OFC/NFOEC 2011.

[11] H. Takahashi, A.A. Amin, S.L. Jansen, I. Morita, H. Tanaka, Highly spectrally efficient DWDM transmission at 7.0 b/s/Hz using 8 × 65.1-Gb/s coherent PDM-OFDM, Journal of Lightwave Technology 28 (2010) 406-414.

[12] K. Kikuchi, Coherent transmission systems, in: 34th European Conference on Optical Communication, 2008, pp. 1-39.

[13] S. Bottacchi, A. Beling, A. Matiss, M.L. Nielsen, A.G. Steffan, G. Unterborsch, et al., Advanced photo receivers for high-speed optical fiber transmission systems, IEEE Journal of Selected Topics in Quantum Electronics 16 (5) (2010) 1099-1112.

[14] J.P. Gordon, L.F. Mollenauer, Effects of fiber nonlinearities and amplifier spacing on ultra-long distance transmission, Journal of Lightwave Technology 9 (2) (1991) 171-173.

[15] RSoft Design Group, Inc., OptSim User Guide, USA, 2008, p. 404.

[16] Corning LEAF Fiber, available online at: http://www.corning.com/opticalfiber/products/LEAF_fiber.aspx.

Computer Technology and Application 2 (2011) 813-817

Information Interaction Is the Fifth Type of Fundamental Interactions

Igor Gurevich

The Institute of Informatics Problems of the Russian Academy of Sciences, Hetnet Consulting Corp., Moscow 119333, Russia

Received: August 19, 2011 / Accepted: September 02, 2011 / Published: October 25, 2011.

Abstract: Information is an integral part of the Universe. By its physical essence information is heterogeneity of matter and energy. Therefore information is inseparably connected with matter and energy. The universal measure of information in physical heterogeneity is the Shannon information entropy. An information approach along with a physical one allows to obtain new, sometimes more general data in relation to data obtained on the ground of physical rules only. The results presented in this paper show the effectiveness of informational approach for studying the interactions in the Universe. The paper shows that, along with the physical interactions the gravitational, electromagnetic, strong, weak interactions exists fifth type of fundamental interactions—information interaction, whose magnitude is not dependent on distance. The existence of information interaction is determined by the entanglement of quantum states, of quantum subsystems. The magnitude of information interaction is measured in bits.

Key words: Physical interactions, quantum system, entangled states, information interaction.

1. Introduction

There are four types of physical interactions: gravitational, electromagnetic, strong and weak. The force of each type of interaction depends on the distance between interacting objects. For example this force for gravitational interaction is as the inverse square of the distance between the objects, for the strong interaction it decreases exponentially with the distance, interaction between quarks is proportional to the distance. The type of dependence of interaction force on distance is defined by relevant physical law [1].

Entangled states, subsystems of quantum system give birth to the fifth type of interaction—information interaction. "Entanglement is a unique quant mechanical resource which plays key part in a number of most interesting applications of quantum calculations and quantum information; this is a kind of iron in the bronze age of classical world. Entanglement is considered a fundamental resource of nature compaired in importance to energy, information, entropy or any other fundamental resource" [2]. The magnitude of information interaction is not dependent on distance [3-4].

Corresponding author: Igor Gurevich, candidate of technical sciences, senior researcher of the institute of informatics problems RUS, main designer hetnet consulting corp., teacher in the Moscow physical-technical institute, Russia, research fields: theoretical informatics, physical informatics, network of classical and quantum computers. E-mail: iggurevich@gmail.com.

2. Estimate of Information Interaction of Entangled States, Subsystems

To make an estimate of connection between two q-bits it is suggested in the work [5] to use mutual information ("информацию связи" in Russian [6-7]). It is shown that information interaction of two q-bits is in the range of (0, 1), and the maximal interaction between two entangled q-bits is equal to one bit.

As interaction of entangled states is measured in information units it is natural to consider this interaction to be information.

In general connection (more precisely as it will be clear later—the magnitude of interaction) between entangled subsystems A and B of quantum system $A +$

B, as well as the connection between two q-bits can be characterized by mutual information.

Mutual information of two quantities A и B communication is defined as follows [6-7]: $I_{AB} = N_A + N_B - N_{AB}$, where I_{AB} is mutual information of subsystem A and B is $I_{AB} = N_A + N_B - N_{AB}$, N_A , N_B —uncertainty (information entropy) of subsystems A, B, N_{AB} is joint uncertainty (joint information entropy) of system $A + B$.

Joint uncertainty (joint information entropy) of the joint allocation of events x and y is equal to $N_{xy} = -\sum_{i,j} p_{ij} \log_2 p_{ij}$, where p_{ij} is probability of joint execution of the event i for x and j for y, $\sum_{i,j} p_{ij} = 1$.

For estimating mutual information of two subsystems of arbitrary system we use representation of the system in the form of Schmidt decomposition [8-9].

We present the wave function (probability amplitude) ψ_{AB} of the system $A + B$ consisting of two entangled subsystems A and B in the form of $\psi_{AB} = \sum_{i=1}^{d} c_i \left|\alpha_i\right\rangle \left|\beta_i\right\rangle$, where *d* is dimensionality of subsystem *A* and *B* (dimensionality of relevant Hilbert spaces); $\left|\alpha_i\right\rangle, \left|\beta_i\right\rangle$ – are orthogonal basis vectors of subsystems A and B; while c_i – are amplitudes of vectors $\left|\alpha_i\right\rangle \left|\beta_i\right\rangle$.

The matrix of joint allocation of basis states of probabilities for subsystems of the system $A + B$ while using Schmidt decomposition is equal to

$$P_{\text{совм. }AB} = \begin{pmatrix} |c_1|^2 & 0 & 0 & 0 \\ 0 & |c_2|^2 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & |c_d|^2 \end{pmatrix}$$

and vectors of P_A^T allocation probabilities of P_B^T realization for basis vectors $\left|\alpha_i\right\rangle$, $\left|\beta_i\right\rangle$ of subsystems A, B are equal, that is

$$P_A^T = \left(|c_1|^2, \ |c_2|^2, \ \ldots, \ |c_d|^2\right)$$

$$P_B^T = \left(|c_1|^2, \ |c_2|^2, \ \ldots, \ |c_d|^2\right)$$

That said the joint uncertainty N_{AB} of subsystems A, B for the system $A + B$, as well as uncertainties N_A, N_B of subsystems A, B discretely are equal, that is

$$N_{AB} = -\sum_{i=1}^{d} |c_i|^2 \log_2 |c_i|^2$$

$$N_A = N_B = -\sum_{i=1}^{d} |c_i|^2 \log_2 |c_i|^2$$

and the mutual information I_{AB} (magnitude of information interaction) of subsystems A, B for the system $A + B$ is equal to

$$I_{AB} = N_A + N_B - N_{AB} = -\sum_{i=1}^{d} |c_i|^2 \log_2 |c_i|^2 \text{ bits.}$$

From the law of uncertainty conservation [3-4] it follows that if a system is at the state ψ, then at changes of coordinates and orientation of q-bits, subsets of q-bits, subsystems, entangled states in general, the uncertainties of entangled states remain.

Q-bits contained in the interlinked position can be also moved relative to each other with any speed without changing the uncertainty which explains conservation of magnitude of information interaction.

By using the connection between information and entropy $S = k \cdot I$ ([6-7] as well as between entropy and energy (potential) $dE = T \cdot dS - P \cdot dV$ [10-11] information interaction of subsystems A and B can be described in accepted units of energy (ergs, joules). The energy (potential) of interaction of subsystems A and B for the system $A + B$ (energy which is necessary to break information interaction between A and B in the accepted units of energy is equal to

$$E_{AB} = S \cdot T = k \cdot T \cdot I_{AB}$$

It is evident that maximum information interaction I_{AB} of subsystems A, B for the system системы $A + B$ is equal to $I_{AB\max} = \log_2 d$ bits.

In as much as for unentangled subsystems A, B $I_{AB\min} = 0$, то then information interaction $F_{I\,AB}$ of subsystems A, B for the system системы $A + B$ lies in the range of $I_{AB\min} \le F_{I\,AB} \le I_{AB\max}$ or $0 \le F_{I\,AB} \le \log_2 d$.

3. Examples of Estimation of Information Interaction of Entangled States

Example 1:

Let joint allocation of probabilities of subsystems A, B states for A, B the system системы $A + B$ is equal to

$$p_{ij} = \frac{1}{n^2},\ i, j = 1,..., n,$$

and

$$P = \begin{Vmatrix} \frac{1}{n^2} & \dots & \frac{1}{n^2} \\ \dots & \dots & \\ \frac{1}{n^2} & \dots & \frac{1}{n^2} \end{Vmatrix}.$$

Joint uncertainty of subsystems A, B is

$$N_{AB} = -\sum_{i,j} p_{ij} \log_2 p_{ij} = -n^2 \frac{1}{n^2} \log_2 \frac{1}{n^2} = 2 \log_2 n$$

As this takes place allocations of probabilities of subsystems A, B (p_i, q_j) states and uncertainties of subsystems A, B $(N_A,\ N_B)$ are equal to

$$p_i = \frac{1}{n},\ i = 1,..., n,\ q_j = \frac{1}{n},\ j = 1,..., n$$

$$N_A = -\sum_i p_i \log_2 p_i = -n \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n$$

$$N_B = -\sum_j p_j \log_2 p_j = -n \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n$$

Information interaction of subsystems A, B for the system $A + B$ is

$$I_{AB} = N_A + N_B - N_{AB} = \log_2 n + \log_2 n - 2 \log_2 n = 0$$

In this case subsystems A, B of the system $A + B$ informationly do not interact.

Example 2:

Let the joint allocation of probabilities of subsystems A, B states for the system $A + B$ is equal to $p_{ij} = \frac{1}{n}\delta_{ij}$, $i, j = 1,..., n$, and

$$P = \begin{Vmatrix} \frac{1}{n} & 0 & 0 & 0 \\ 0 & \frac{1}{n} & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \frac{1}{n} \end{Vmatrix}.$$

Joint uncertainty of subsystems A, B is $N_{AB} = -\sum_{i,j} p_{ij} \log_2 p_{ij} = -n \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n$.

Probabilities of subsystems A, B (p_i, q_j) states and uncertainties of subsystems A, B $(N_A,\ N_B)$ is equal to

$$p_i = \frac{1}{n},\ i = 1,..., n$$

$$q_j = \frac{1}{n},\ j = 1,..., n$$

$$N_A = -\sum_i p_i \log_2 p_i = -n \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n$$

$$N_B = -\sum_j p_j \log_2 p_j = -n \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n$$

Information interaction of subsystems A, B for the system $A + B$ is equal to

$$I_{AB} = N_A + N_B - N_{AB} = \log_2 n + \log_2 n - \log_2 n = \log_2 n$$

In the given example information interaction of subsystems A, B for the system $A + B$ equals to the maximal possible information interaction.

Example 3:

Let the system $A+B$, of the subsystem A, B contains $\approx N$ bits of information. There at the number of q-bits in each of the interacting subsystems is approximately equals to $\approx N$ and the dimensionality of the relevant Hilbert space equals to and the information interaction of subsystems does not exceed the value of $F_{I\,AB}=\log_2 d=N$ bits. Information interaction of subsystems is determined by the volume of information in the subsystems [3-4, 12-13].

Inasmuch as there are 10^{90} bits of information in the Universe [3-4, 12-13] то one can estimate the maximal information interaction of its two subsystems.

Let the maximal possible volume of information (q-bits) in each of the interacting subsystems approximately equals to 10^{90}, and since the dimensionality of the relevant Hilbert space equals to $d=2^{10^{90}}$, then maximal possible information interaction of subsystems in the Universe does not exceed 10^{90} bits.

All quantum objects, quantum systems and subsystems—bozons and fermions are subject to information interaction. Vacuum due to its universal aspect seems to be the way of transmitter (carrier) of information interaction (defining the carrier of information interaction is the subject of future research).

Note: Information interaction can not be treated as a consequence and/or characteristics of the known fundamental physical interactions: gravitational, electromagnetic, strong, weak, though the entangled states are made with the use of these interactions, first of all with the use of electromagnetic interaction. Such treatment is not possible to use due to the fact that information interaction does not depend on distance while all known types of interaction do.

4. Properties of Information Interaction

Information interaction (mutual information) of subsystems A, B of arbitrary system $A+B$, being in the state ψ, possesses the following main properties.

The information interaction of subsystems A, B is scalar.

The magnitude of information interaction of subsystems A, B is symmetrical: $I_{AB}=I_{BA}$.

The magnitude of information interaction of subsystems A, B is not negative.

The magnitude of information interaction of subsystems A, B does not exceed the value of $I_{AB\max}=N_A=N_B$.

At changes of coordinates and orientation also speed of q-bits, subsets of q-bits, subsystems, entangled states in general, the uncertainties of entangled states remain.

5. Decoherence of the Entangled States, Subsystems Are Reducing the Magnitude of the Information Interaction

It should be noted that in general intensity of information interaction decreases with time. The course of it is explained by decoherence of entangled states determined by interaction with external environment. In this section the ideas of article [9] are used. “Decoherence consists in attenuation (disappearance) of off-diagonal elements of systems density matrix as a result of orthogonalization of environment states corresponding to various basis states of computer. Thus for quantum computers interaction of register with uncontrolled environment, ambiguities in parameters values of control pulses, as well as uncontrolled interaction of q-bits is the origin of decoherence state of quantum computer as a whole”.

In general, external environment is the source of decoherence of any entangled states, subsystems. In article [9] the estimate of decoherence speed is given which was received for the model of q-bits system at the environment of interacting oscillators.

Systems with coherent environment of the type of state like “Schrodinger cat” $\left|0_1 0_2 ... 0_n\right\rangle \pm \left|1_1 1_2 ... 1_n\right\rangle$

are subject to maximal quick decoherence. At the same time there are states that are free from decoherence. Such are the states where for one half of q-bits the state is equal to $|0\rangle$, and for another half it is equal to $|1\rangle$.

As an example of such systems for even-numbered n, there is a system of the type $|0_1 0_2 ... 0_{n/2} 1_{n/2+1} 1_{n/2+2} ... 1_n\rangle \pm |1_1 1_2 ... 1_{n/2} 0_{n/2+1} 0_{n/2+2} ... 0_n\rangle$.

In particular the states of Bell $|0_1 1_2\rangle \pm |1_1 0_2\rangle$ are free from decoherence.

Note that the vectors $|0_1 0_2 ... 0_{n/2} 1_{n/2+1} 1_{n/2+2} ... 1_n\rangle$ and $|1_1 1_2 ... 1_{n/2} 0_{n/2+1} 0_{n/2+2} ... 0_n\rangle$, which present the states free from decoherence possess equal energy. In the same way the vectors $|0_1 1_2\rangle$ и $|1_1 0_2\rangle$, containing the states of Bell free from decoherence possess equal energy.

Thus in a general case the decoherence leads to decreasing and after the expiry of some time to disappearance of information interaction.

6. Conclusions

The paper shows that, along with the gravitational, electromagnetic, strong, weak interactions exist the fifth type of fundamental interactions—information interaction, whose magnitude is not dependent on distance.

The existence of information interaction is determined by the entanglement of quantum states, of subsystems.

In general connection (the magnitude of interaction) between entangled subsystems A and B of quantum system *A + B,* as well as the connection between two q-bits can be characterized by mutual information.

The magnitude of information interaction is measured in bits.

In a general case the decoherence leads to decreasing and after the expiry of some time to disappearance of information interaction.

References

[1] I.L. Buchbinder, Fundamental interactions (in Russian), Soros Educational Journal (in Russian) 3 (5) (1997) 66-73.

[2] M. Nielsen, I. Chuang, Quantum computation and quantum information (in Russian), Mir, Moscow, 2006.

[3] I.M. Gurevich, Law of Informatics—A Basis of Researches and Designing of Complex Communication and Management Systems (in Russian), Ecos, Moscow, 1989.

[4] I.M. Gurevich, Informatics Laws—A Basis of a Structure and Cognitive of Complex Systems (in Russian), 2nd ed., Torus Press Moscow, 2007.

[5] I.M. Gurevich, Information characteristics of the entangled states (in Russian), Information Technology, № 5, Moscow, 2006.

[6] C.E. Shannon, A mathematical theory of communication, Bell System Technical Journal T. 27 (1948) 379-423 & 623-656.

[7] R.L. Stratonovich, Information Theory (in Russian), Soviet Radio, Moscow, 1975.

[8] K.A. Valiev, A.A. Kokin, Quantum Computers: Hope and Reality (in Russian), Scientific and Publishing Center "Regular and Chaotic Dynamics", Moscow-Izhevsk, 2004, p. 320.

[9] K.A. Valiev, Quantum computers and quantum computing (in Russian), Successes of Physical Sciences 175 (1) (2005).

[10] L. Brillouin, Science and Information Theory (in Russian), Fizmatgiz, Moscow, 1960.

[11] L.D. Landau, E.M. Lifshitz, Statistical Physics (in Russian), Science, Moscow, 1964.

[12] Penrose R. The Emperor's New Mind, Oxford University Press, 1989.

[13] S. Lloyd, Computational capacity of the universe, Phys. Rev. Lett. 88 (2002).

Computer Technology and Application 2 (2011) 818-827

Interactive Protein Data Clustering

Terje Kristensen[1, 2] and Vemund Jakobsen[2]

1. The Department of Computer Engineering, Faculty of Engineering, Bergen University College, Nygårdsgaten 112, Bergen 5020, Norway

2. The Company Pattern Solutions AS, Nygårdsgaten 112, Bergen 5020, Norway

Received: August 25, 2011 / Accepted: September 09, 2011 / Published: October 25, 2011.

Abstract: In this paper, the authors present three different algorithms for data clustering. These are Self-Organizing Map (SOM), Neural Gas (NG) and Fuzzy C-Means (FCM) algorithms. SOM and NG algorithms are based on competitive learning. An important property of these algorithms is that they preserve the topological structure of data. This means that data that is close in input distribution is mapped to nearby locations in the network. The FCM algorithm is an algorithm based on soft clustering which means that the different clusters are not necessarily distinct, but may overlap. This clustering method may be very useful in many biological problems, for instance in genetics, where a gene may belong to different clusters. The different algorithms are compared in terms of their visualization of the clustering of proteomic data.

Key words: Datamining, self-organizing map, neural gas, fuzzy c-means algorithm and protein clustering.

1. Introduction

An important step in the data analysis process is to organize the data into meaningful structures to uncover their natural grouping(s). Once the data has been organized, it can be used in subsequent steps of analysis such as hypothesis forming or decision making. This method of exploring and organizing data can be done automatically by using an algorithmic approach known as data clustering (or cluster analysis). Data clustering has been studied extensively over the years and has several areas of application, including market research, pattern recognition, image analysis and machine learning [1]. In these areas, one often has to do analysis without a priori information about the data. Data clustering can in such cases be useful for organizing and uncovering relationships hidden in the data before doing further analysis.

Analyzing large amounts of data can be difficult and is a time-consuming task if done manually. It is therefore necessary to develop tools to support analysis and visualization of large multi-dimensional data sets. These tools may provide structured views of the data, and potentially reveal previously unknown information. Although there are several algorithms that can be used for this purpose, they are not always equally suited for the problem at hand. Using different algorithms on the same problem will often produces different results, and the best way to analyze data is therefore to compare the different algorithmic approaches and their solutions.

The three algorithms and their visualization are

- Self-Organizing Map (SOM), which is proven to be a powerful tool for visualizing high-dimensional data;
- Neural Gas (NG) [2-3], that can be a less constrained version of SOM;
- Fuzzy C-means (FCM), which opposed to the other algorithms, allow objects to belong to several clusters.

There are many systems that support analysis and visualization of clusters based on SOM, NG and FCM, but not in one combined system. In the paper we describe a system that explores data by using these clustering methods and their visualizations. The data to

Vemund Jakobsen, master, research fields: datamining, adaptive systems, machine learning.

Corresponding author: Terje Kristensen, professor, CEO, cand.real, research fields: machine learning, pattern recognition, datamining, multi-agent systems. E-mail: tkr@hib.no.

be clustered is high-dimensional with more than three dimensions. To be able to interpret the data, the dimension has to be reduced to either two or three dimensions. Finally, the algorithms have been applied to interpret the clustering of protein data found in the blood serum.

The organization of the paper is as follows: Section 2 defines hard and soft clustering, in section 3 and 4 self-organizing map and neural gas algorithms are described. In section 5 the fuzzy center means algorithm (FCM) is presented. Section 6 presents the complexity of the different algorithms. The system that is developed to carry out the on-line clustering is presented in section in section 7. In section 8 the colour model is presented for mapping similar weights to similar colours. Finally, section 9 presents a case of clustering of protein mass spectrometry data (MS) data. Section 10 gives the conclusions.

2. Main Types of Clustering

2.1 Hard Clustering

In hard clustering (or crisp clustering) the objects belong to one and only one cluster, and each cluster contains at least one object. In mathematical terms it can be defined as follows:

Let *D* be the set of data consisting of *n* vectors, that is

$$D = \{x_1, x_2, \ldots, x_n\} \quad (1)$$

We now define the k-clustering of *D* as the partition of *D* into *k* sets (clusters) *C1*, ..., *Ck* such that the following conditions are satisfied

$$Ci \neq \emptyset, \quad i = 1, \ldots, k \quad (2)$$

$$\cup Ci = D \quad (3)$$

$$Ci \cap Cj = \emptyset, \quad i \neq j, \quad i, j = 1, \ldots, k \quad (4)$$

2.2 Soft Clustering

In soft clustering or Fuzzy clustering an approach based on fuzzy logic is used [4]. The objects may then belong to several objects with a certain degree of membership. This means for instance that an object in the center of a cluster may have a higher probability of belonging to that cluster than an object on the edge of the cluster. It can be defined as follows: let *D* be the set of data consisting of *n* vectors. A fuzzy clustering of *D* into *k* clusters is characterized by *k* functions u_j where

$$u_j: D \rightarrow [0, 1], \quad j = 1, 2, \ldots, k, \quad \text{and} \quad (5)$$

$$\sum_{j=1}^{k} u_j(x_i) = 1, \; i = 1, 2, .., n, \quad 0 < \sum_{j=1}^{k} u_j(x_i) < n \quad (6)$$

The functions u_j are the membership functions. Vectors with membership values close to unity have a high degree of membership in the corresponding cluster, and vectors with values close to zero have a low degree of membership in the corresponding cluster.

3. Self-Organization Map Algorithm

The Self-Organizing Map (SOM) network consists of two fully connected layers: the input layer (representing input vectors) and the output layer. The output layer consists of nodes that are arranged in a map. Moreover, each node has a specific position in the map and also an associated weight vector of the same dimension as the input vector.

An important feature of the SOM algorithm is that it preserves the topological structure of the input space. This means that data items that are close in the input space are mapped to nodes that are close in the map [5]. The SOM algorithm accomplishes this by calculating a neighbourhood function within the map. All nodes within this neighbourhood are adjusted towards the input vector.

The idea of SOM is to adjust the nodes until they represent the input distribution. The nodes represent clusters that reflect how the data is distributed as given in Fig. 1. The SOM algorithm begins by setting the weights to random values. It then proceeds to the three processes given below.

3.1 Competitive Process

Let $x = (x_1, x_2, \ldots, x_d)^T$ be an object selected at random from the data set, where *d* is the dimension of the data set. Let $w_i = (w_{i1}, w_{i2}, \ldots, w_{id})^T$ be the weight vector of node *i*. The node that is most similar, or closest,

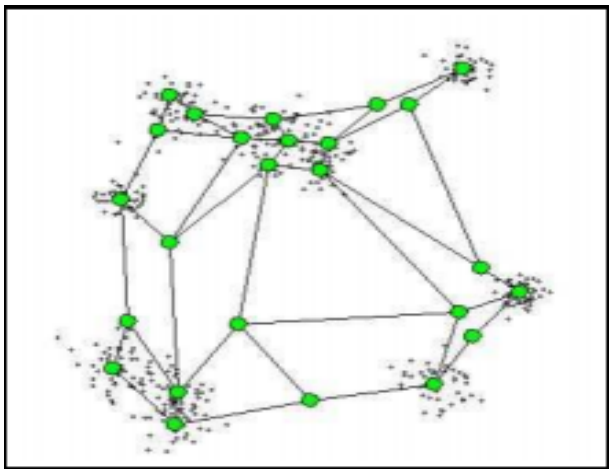


Fig. 1 SOM network approximating the input distribution.

according to some distance measure to the input vector, is then determined. This node is referred to as the Best Matching Unit (BMU) [6-7].

3.2 Cooperative Process

In this process a topological neighbourhood is defined where the BMU locates the center of the neighbourhood. The neighbourhood is usually an exponential function, typically a Gaussian function defined as

$$h(t) = \exp(-d^2_{c,j}/2\sigma^2(t)) \quad (7)$$

where t is the current iteration, $d_{c,j}$ is the lateral distance at node j and σ is the radius or with of the neigbourhood function specified by

$$\sigma(t) = \sigma_0 \exp(-t/\lambda_1) \quad (8)$$

where λ_1 is a time constant that allows the exponential function to decay with increasing time, and σ_0 is the radius at time t_0. The lateral distance d_c in the two-dimensional case is defined as

$$d_c = ||x - w_c(t)|| = \min\{||x - w_i(t)||\}, 1 \le i \le k \quad (9)$$

3.3 Adaptive Process

In the adaptive process the weights of all nodes within the neigbourhood, as well as BMU, are adjusted according to the rule:

$$w_i(t+1) = w_i(t) + \alpha(t)h(t)[\mathbf{x}(t) - \mathbf{w_i}(t)] \quad (10)$$

where $w_i(t)$ and $x(t)$ are the weight and input vector at time t. The learning rate $\alpha(t)$ is defined as

$$\alpha(t) = \alpha_0 \exp(-t/\lambda_2) \quad (11)$$

where λ_2 is a time constant that brings the learning rate close to zero with increasing time t, and α_0 is the learning rate at time t_0.

3.4 Choice of Parameters

SOM leads to an organized representation of input objects provided that the parameters of the algorithm are selected properly. The adaption process may be decomposed into two phases; the ordering phase and the convergence phase. In the ordering phase the topological ordering of the weight vectors takes place. The topological ordering produces a rough ordering of weight vectors when the neighbourhood is big and the learning rate high. In the convergence phase the map is fined-tuned and provides an accurate statistical quantification of the input space. During this phase the neighbourhood is small and the learning rate low. Some guidelines for selecting the parameters are given below:

- The learning rate is often set initially close to 0.1 and decreases gradually to above 0.01;
- The neigbourhood $h(t)$ starts with a large radius covering most of the nodes in the grid. The initial radius σ_0 can be set equal to the "radius" of the lattice;
- The number of iterations are difficult to set and depends on both the size and dimensionality of the data set.

In general the topological ordering of the weights may require as many as 10000 iterations or more [8]. After the weights have been topologically ordered, the map needs to be fine-tuned. The number of iterations constituting the convergence phase should be at least 500 times the number of nodes in the lattice. The algorithm in pseudo code is given in Fig. 2.

4. The Neural Gas Algorithm

The NG algorithm is, similar to SOM, an algorithm that uses unsupervised competitive learning. It starts by initializing a set W containing N nodes, each with weights set to random values. It then proceeds, in a similar way as SOM, to the three processes given below.

Algorithm 1: Pseudocode of the SOM algorithm
Input: A set of input vectors D = {$\mathbf{x_1}$, $\mathbf{x_2}$, . . . , $\mathbf{x_n}$}
Output: A set of weight vectors W = {$\mathbf{w_1}$, $\mathbf{w_2}$, .. . ,$\mathbf{w_n}$}
1: Set the parameters α_0, σ_0, τ_1, τ_2, k and t_{max}
2: Initialize all w ∈ W to random values
3: **for** t = 1 to t_{max} **do**
4: Select random **x** ∈ D
5: find **w** such that d(**x, w**) = min{ d(**x, w**) | **x** ∈ D}
6: **for all w** in the neigbourhood h **do**
7: Update the weights **w** = **w** + αh(**x**-**w**)
8: Reduce learning rate α
9: **end for**
10: **end for**

Fig. 2 The SOM algorithm in pseudo code.

Like SOM, NG aims to preserve the topological structure of the input space as shown in Fig. 3. It also uses a neighbourhood function. However, instead of computing the neigbourhood within a lattice, it computes the neighborhood within the input space. The nodes are therefore ranked by their distance from an input object. After ranking, the node weights are adjusted according to an adapted rule, with the closest node being adjusted most. As a consequence, similar input objects are more likely to be mapped to nodes that are close (i.e., similar) in the NG network (output space). In addition, the NG adaption rule minimizes a global cost function. Such a cost function does not exist for the SOM adaption rule [9].

Since NG lacks a fixed output space, it can achieve better results that SOM. However, this also limits the applications of NG to data projection and visualization, and therefore only a few visualization schemes have been developed.

4.1 Competitive Process

Let $x = (x_1, x_2, \ldots, x_d)^T$ be an object selected at random from the data set, where *d* is the dimension of the object. Sort all nodes of the weights space *W* according to their distance to *x*, with the nearest node coming first and the farthest as the last. That is, for sorted W = {Cm, Co, Ck, . . .} with corresponding weights w_m, w_o, w_k, …, the following relation is given

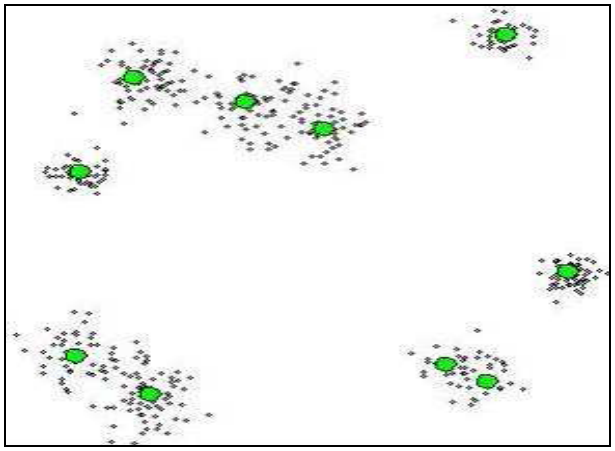


Fig. 3 NG network approximating the input distribution. The network consists of green points and the input data of black points.

$$\|x - w_m\| \le \|x - w_0\| \le \|x - w_k\| \ldots.$$

where the norm is usually the Euclidean norm.

4.2 Cooperative Process

In this process the neigbourhood function could be an exponential function given by

$$h(k_i) = \exp(-k_i / \lambda(t)),\ t = 1, 2, 3 \ldots, \quad (12)$$

where k_i is the rank index in *W* and *t* is the current iteration. $k_i = 0$ means the closest node w_i to the input vector *x* and $k_i = N\text{-}1$ the least closed node to the input vector *x*. In the equation above $\lambda(t)$ is defined by

$$\lambda(t) = \lambda_0(\lambda_f / \lambda_0)^{(t / t_{max})} \quad (13)$$

where λ_0 and λ_f are the initial and final with of the neigbourhood, respectively, and t_{max} the maximum number of iterations.

4.3 Adaptive Process

In the adaptive process the weights of all the nodes in *W* are adjusted according to the rule:

$$w_i(t + 1) = w_i(t) + \varepsilon(t)h(t)[x(t) - w_i(t)] \quad (14)$$

where $\varepsilon(t)$ is given by

$$\varepsilon(t) = \varepsilon_0(\varepsilon_f / \varepsilon_0)^{(t / t_{max})} \quad (15)$$

The algorithm in pseudo code is given in Fig. 4.

5. The Fuzzy C-Means Algorithm

In traditional clustering the clusters are disjoint. Fuzzy C-means clustering is a fuzzy-based clustering algorithm that allows objects to belong to several

Algorithm 2: Pseudocode of the NG algorithm
Input: A set of input vectors D = {$\mathbf{x_1}$, $\mathbf{x_2}$, ... , $\mathbf{x_n}$}
Output: A set of weight vectors W = {$\mathbf{w_1}$, $\mathbf{w_2}$, ... ,$\mathbf{w_n}$}
1: Set the parameters λ_0, λ_f, , ε_f , h_λ, N and t_{max}
2: Initialize all w ∈ W to random values
3: **for i** = 1 to t_{max} **do**
4: Select random **x** ∈ D
5: **for** j = 1 to t_{max} **do**
6: Sort the weight vectors in W according to their distance
7: Update the weights $\mathbf{w} = \mathbf{w} + \varepsilon\, h_\lambda(k(\mathbf{x}, W))(\mathbf{x}-\mathbf{w})$
8: Reduce learning rate ε
9: **end for**
10: **end for**

Fig. 4 The NG algorithm in pseudo code.

clusters with different degrees of membership. In many practical situations fuzzy clustering is more natural to use than hard clustering.

In FCM the clusters are represented by the cluster centers. Each of the data has a degree of membership in each cluster center. In a plotting, the vertical axis represents the membership function value $u_j(x)$ corresponding to the cluster j and the horizontal axis represents the data items x to be clustered. In Fig. 5 there are three membership functions corresponding to the three clusters, denoted as low, medium and high.

FCM is based on fuzzy-partition and can be described as follows:

Let $C = \{c_1,c_2,...,c_k\}$, be a set of cluster centers and $D = \{x_1,x_2,...,x_n\}$ the set of given data where each object $x_i = \{x_{i1},x_{i2},...,x_{id}\}$ is a d-dimensional vector and U a $k \times k$ matrix. The matrix U is called a fuzzy k-partition and has the following constraints:

$$u_{ji} \in [0.1] , 1 \le j \le k, \quad 1 \le i \le n \tag{16}$$

$$\sum_{j=1}^{k} u_{ji} = 1, \; 1 \le i \le n \tag{17}$$

$$\sum_{i=1}^{n} u_{ji} > 0, \quad 1 \le j \le k \tag{18}$$

where u_{ji} is the membership value of object x_i in cluster j. The constraint (17) implies that every object has some degree of membership in at least one cluster and

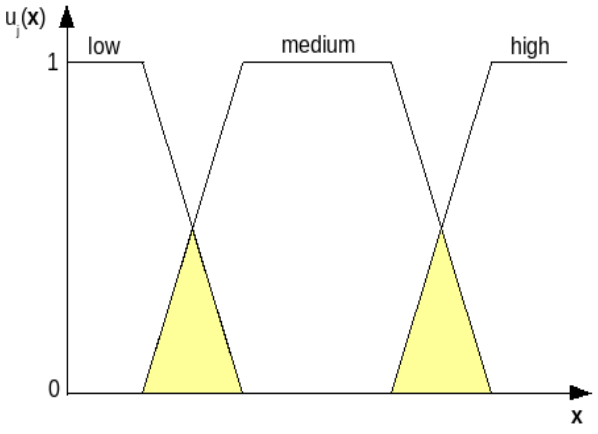


Fig. 5 The membership function of three clusters.

that the sum of the objects' membership values is 1. The last constraint implies that each cluster contains at least one object to some degree, i.e., there are no empty clusters.

The aim of the FCM algorithm is to find an optimal fuzzy k-partition and corresponding cluster centers by minimizing the objective function [10].

$$J(U,C) = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ji}^{\,m} \|x_i - c_j\|^2 , \; 1< m < \infty \tag{19}$$

In (19) m is a fuzzification constant that influences the membership values, c_j is a d-dimensional cluster center and ||.|| is the norm expressing the similarity between objects and cluster centers. The value of m determines the amount of fuzzification. The larger m is, the fuzzier the cluster membership is. An m value close to 1 expresses hard clustering.

The algorithm starts by generating randomly a fuzzy k-partition U^0. It then proceeds to update the cluster centers c_j by calculating

$$c_j(t) = \frac{\sum_{i=1}^{n} (u_{ji}(t))^m xi}{\sum_{i=1}^{n} (u_{ji}(t))^m}, \quad t=1, 2 ,3,.... \tag{20}$$

where t is the current iteration step. Given the new cluster centers, the membership values are updated according to:

$$u_{ji}(t+1) = \frac{1}{\sum_{i=1}^{n} \left(\frac{\| x_i - c_j(t) \|}{\| x_i - c_h(t) \|}\right)^{\frac{2}{m-1}}} \tag{21}$$

The process stops when $||U(t+1) - U(t)|| \leq \varepsilon$ where $||.||$ is the matrix norm and ε is a termination criterion between 0 and 1, or the maximum number of iterations t_{max} is reached. The algorithm in pseudo code is given in Fig. 6.

6. Complexity

6.1 Complexity of SOM

Within the neigbourhood the weights of the SOM algorithm are adjusted by the updating procedure. This requires updating of every component of the weight vector and has *O(d)* complexity where *d* is the dimension of the weight vector. Adjusting the SOM map requires looping through the whole map where *k* is the number of nodes. Finding the BMU and adjusting the map are operations performed in every iteration. For a total of *n* iterations the SOM complexity is *O(nkd)*.

6.2 Complexity of NG

The NG training process is very similar to the training process of SOM. To implement NG we have to compute the Euclidean distance and current neighbourhood, that means

- Compute the Euclidean distance between the input vector and every node;
- Sort the nodes in ascending order according to the computed distance;
- Adjust the weights. The closer a node is to an input vector the more the weight is adjusted.

With respect to the points above the complexity of computing the distances d is *O(kd)*, where *k* is the number of nodes and is the dimension of the input vector. The nodes are then sorted according their distances. In an actual implementation of NG we may use "merge sort" with a complexity of *O(klogk)*, where *k* is the number of elements. Adjusting the node weights requires looping through the list of nodes adjusting the weights. The gives a complexity of *O(kd)* as in SOM. For a total number of n iterations the total complexity becomes *O(nkd + nklogk)*.

Algorithm 3: Pseudocode of the FCM algorithm
Input: A set of input vectors D = {$\mathbf{x_1}$, $\mathbf{x_2}$, . . . , $\mathbf{x_n}$}
Output: The fuzzy partition of U and corresponding cluster centers
1: Set the parameters ε, m, k, t_{max}
2: Initialize U to random values such that
$\sum_{j=1}^{k} u_{ji} = 1$, i =1,2,3,….
3: **repeat**
4: Calculate the cluster centers c_j = 1,2,3, . ..k

$$c_j(t) = \frac{\sum_{i=1}^{n} (u_{ji}(t))^m xi}{\sum_{i=1}^{n} (u_{ji}(t))^m}, \quad t=1, 2, 3, \ldots.$$

5: **for** i = 1 to n **do**
6: **for** j = 1 to k **do**
7: Calculate $d(\mathbf{x}_i, \mathbf{c}_j) = ||\mathbf{x}_i - \mathbf{c}_j||$
8: **if** $d(\mathbf{x}_i, \mathbf{c}_j) > 0$ **then**
9: $u_{ji} = \frac{1}{\sum_{i=1}^{n} \left(\frac{||x_i - c_j(t)||}{||x_i - c_h(t)||}\right)^{\frac{2}{m-1}}}$
10: **else**
11: $u_{ji} = 1$
12: **end if**
13: **end for**
14: **end for**
15: **until** $||U(t+1) - U(t)|| \leq \varepsilon$ OR the maximum number of iterations t_{max} is achieved

Fig. 6 The FCM algorithm in pseudo code.

6.3 Complexity of FCM

The FCM algorithm is very different from SOM and NG algorithm. To implement FCM algorithm one has to

- Select an input sample from the data set;
- Update the cluster centers;
- Update the membership values;
- Compute the distance between the current membership matrix and the previous and check for termination;
- Copy the membership matrix for use in the next iteration.

The updating of cluster centers and membership values are the most computationally expensive operations. Each cluster center is the mean of all input vectors weighted by their degree of belonging to the cluster. Updating one cluster center takes *O(nd)*, where *n* is the number of input vectors and *d* is the

dimensionality of input vector (or cluster center). Since this is done for every cluster center, the total complexity becomes *O(knd)*, where *k* is the number of cluster centers.

The complexity of finding the Euclidean distance between cluster centers and input vector is *O(d)*, where *d* is the dimensionality of the input vector. We need to compute the distance between the input value and every cluster centers which takes *O(kd)* time, where k is the member of cluster centers. For n input vectors and every cluster center the total complexity is $O(k^2nd)$ in the worst case.

The complexity of computing the matrix distance is done by minimizing their distance for every time the cluster centers and membership values have been updated. When this time is lower than a threshold value, the algorithm stops. The distance between the current membership matrix and the one in the previous iteration is computed in the same way as the Euclidean distance. This matrix norm is known as Frobenius norm [11] and has a complexity of *O(kn)*, where *k* is the number of cluster centers and n is the number of input vectors. The total complexity of FCM for a total of *t* iterations is therefore $O(tk^2nd)$ in the worst case, and thus more computationally heavy than both SOM and NG.

7. The System

In our system the SOM, NG and FCM algorithms can be run simultaneously in background while exploring and revealing the cluster relationships hidden in the data. By using threads in java one is able to run the different algorithms simultaneously.

By comparing the different clustering methods the user can get new insights into analysis of the data at hand. The system mainly performs two tasks:

(1) Running the SOM, NG and FCM algorithm on a data set;

(2) Visualizing the SOM, NG and FCM clustering of data.

The first task involves the following steps:

- Importing a data set into a data structure;
- Choosing an algorithm and setting the parameters;
- Running the algorithm on the data set.

The second task involves:

- Translating the trained network to a graphical display such that the cluster relationships are revealed;
- Displaying additional tools on request.

7.1 Functional requirements

The input to the system is supposed to be a text file containing input objects. The column represents the attributes of the input objects. We assume that the attributes currently have numerical values. The attributes may be separated by a delimiter which can be any character or just a blank.

In the visualization step we have to translate the internal data structure onto a graphical display by mapping each weight vector to a colour such that similar vectors are reflected in similar colours. Each algorithm is visualized in a separate window to make it easier to compare the different types of clustering. The system is scaled well and is able to handle data sets of different size (number of items) and dimensionality. The parameter values (iterations, nodes, etc.) are not fixed, but depend on the dimensionality of the data set.

8. The Colour Model

The mapping of similar weights to similar colours is rather difficult since the weights are of higher dimension than the colour space. To be able to map objects of an arbitrary dimension onto colours we need to use an approach proposed by Schatzman [12]. In this approach an HSB model was developed where H = hue, S = saturation and B = brightness. This model describes the points inside an inverted cone (see Fig. 7).

In a representation the hue, saturation and brightness are floating point values between 0 and 1. In the mapping of node weights to colours, the attribute values must be in the range from 0 to 1. To map a set of d-dimensional weight vectors to colours, one first has to divide the HSB colour circle at *d* equally spaced angles, given *d* colours:

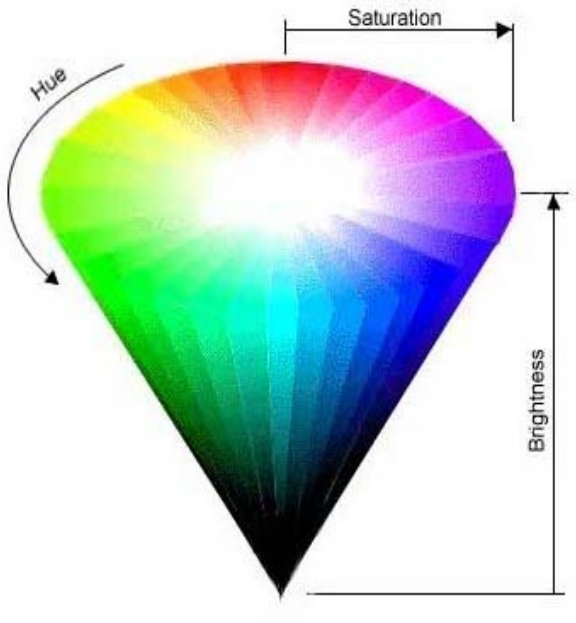


Fig. 7 Colour mapping model.

$$c_1, c_2,..., c_d \quad (22)$$

Each of these colours is a three-dimensional vector with red, green and blue component.

$$c_i = (c_{ir}, c_{ig,}, c_{ib})^T \quad (23)$$

Any d dimensional vector $w_i = (w_1,w_2,...,w_d)^T$ can now be mapped to a colour c_w by computing:

$$cw_r = \sum c_{ir} w_i / \sum c_{ir} \quad (24)$$

$$cw_g = \sum c_{ig} w_i / \sum c_{ig} \quad (25)$$

$$cw_b = \sum c_{ib} w_i / \sum c_{ib} \quad (26)$$

The mapping of every node in the map to a colour does not increase the overall complexity of the algorithm. In SOM, for instance, to make d colours give a complexity of $O(d)$. The mapping of k nodes to colours requires $O(kd)$ which is less than the overall complexity of SOM.

The SOM, NG and FCM algorithms require different visualization regimes that reflect their behavior. Furthermore, implementation of NG and FCM, as opposed to SOM require dimension reduction to be mapped to the nodes (or cluster centers) onto a two-dimensional surface and therefore requires more computation.

9. Protein Clustering

The three algorithms have been applied on clustering of protein mass spectrometry (MS) data [13-15] where the aim is to describe large-scale sets of proteins, their structure and function. MS has now become a platform for both identification and quantification of proteins. The basic principle of MS is to generate ions separated by their mass-to-charge ratio (m/z) which can be detected qualitatively and quantitatively by their m/z value and abundance.

MS produces large amounts of high-dimensional data that is difficult to analyze [16]. In this paper we use the three clustering algorithms to group the MS protein data in order to detect if a change in the environment causes a change in the clustering of the proteins of the human blood serum.

9.1 The Data Set

The entire data set is made of ten smaller data sets containing MS data of proteins (or protein fragments) in the human serum, where each measurement is described by many attributes. These 10 data sets are further divided into two groups where five groups contain MS data representing proteins in the blood serum when the individual is sitting (S) and the other five while the individual is lying down (L). Our goal is to detect differences of these samples of data by comparing the clustering of S and L.

9.2 The Clustering of the Data Set

The difference in magnitude between the entries in the data sets were large, so they were preprocessed using a procedure known as z-score transformation [17].

The SOM clustering also contains the density of the input distribution. This is based on an idea developed by Simula [18] and aims to show the density of the input distribution in terms of map nodes. This is done by first locating the input vector's BMU, after training. Then, the BMU's "hit" counter is incremented. Inside each square, the hit counter of the corresponding node is printed. If a node has zero hits the square remains entirely blank. This makes it easier to detect borders.

The following parameter values were used in the

different clustering algorithms:

SOM	NG	FCM
#iter:50 000	#iter:50 000	#iter:100
map size:10 x 12	# nodes:120	#cluster centers:3
Init radius: 6	init radius 60	fuzzification:2
Init learn_ rate:0.1	init learn_ rate:0.5	stop crit:0.001
	final radius:0.01	
	final learn_rate:0.005	

9.3 Visualization of the Sets L and S

Figs. 8-12 show the SOM, NG and FCM clustering of the *L* and *S* data set.

In the SOM clustering there is a red cluster in *L* that is larger than the corresponding one in *S*, but the red cluster in *S* seems to be more compact as we can see in Figs. 8-9.

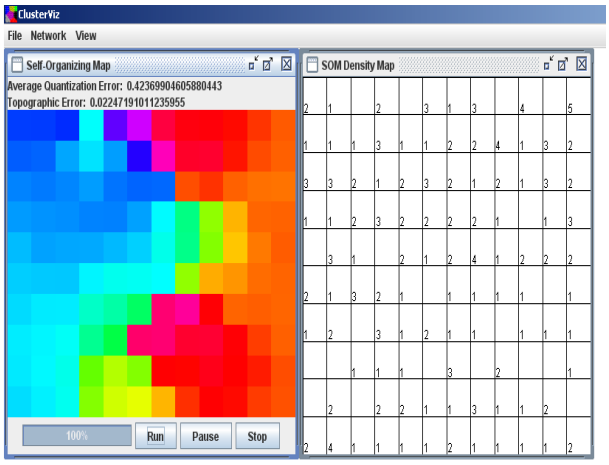


Fig. 8 SOM clustering of *L* data set.

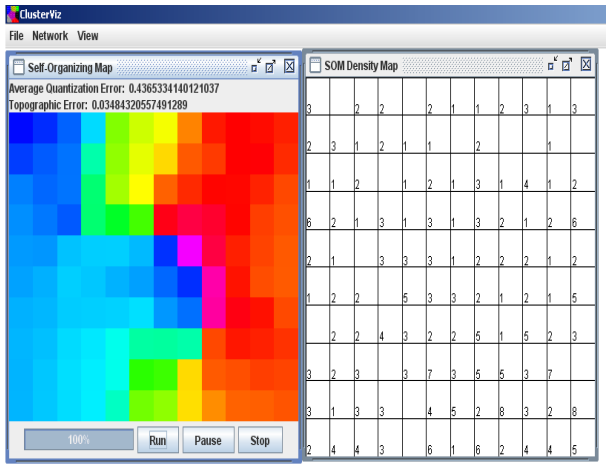
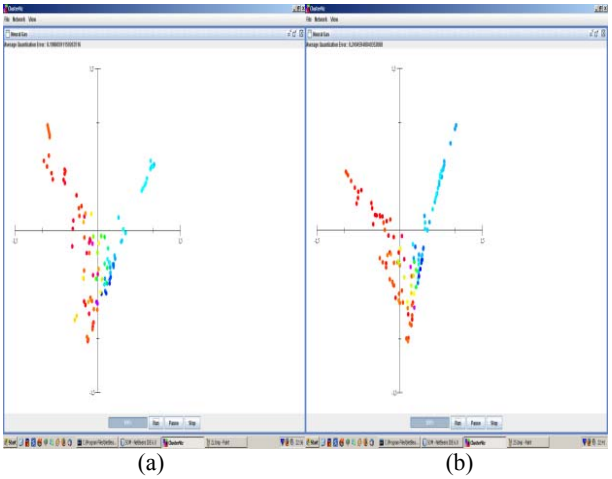


Fig. 9 SOM clustering of *S* data set.

For NG we notice that the blue cluster of *S* is a little more compact the blue cluster of *L*, but the blue cluster of *S* is larger. In addition, NG also contains a red cluster which is similar to the SOM cluster found in both *L* and *S* as seen in Fig. 10.



(a) (b)

Fig. 10 (a) NG clustering of *L* data; (b) NG clustering of *S* data.

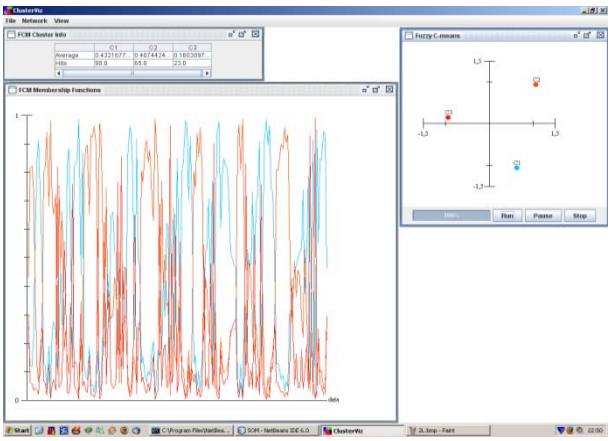


Fig. 11 The FCM clustering of *L* data.

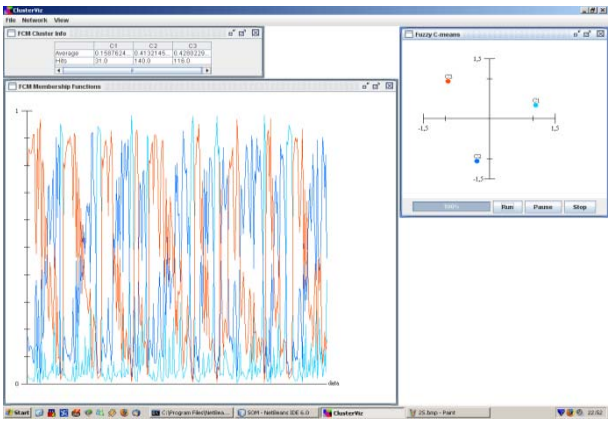


Fig. 12 The FCM clustering of *S* data.

For the FCM clustering we observe in Fig. 11 that L contains two red clusters represented by the cluster centers denoted by c_2 and c_3. On the contrary, in Fig. 12, the FCM clustering in S contains only one red cluster represented by c_3. The corresponding c_3 red cluster of L has a low hit number (not seen in the Fig. 11), so most of the data is contained in the c_2 and c_3 clusters of L. In S, in Fig. 12, the situation is opposite, where the blue clusters c_2 and c_3 contain most of the data.

10. Conclusions

A clustering system has been developed for multi-dimensional data. The system provides functionality for visualizing SOM, NG and FCM algorithms and for viewing them side by side to identify similarities in their structures. The system has been demonstrated on proteomic data of the blood serum. Use of all the different methods indicates that there exist structural differences in the protein structure when blood tests are taken in lying or sitting positions.

This may be an important aspect to remember in connection with determination of Alzheimer or Parkinson disease based, for instance, on genetic data. The micro-array gene expression data may be different when blood samples are taken from patients in lying and sitting positions.

References

[1] P.A. Estevez, W. Lucky, Automatic equalization for digital communication, Bell Syst. Tech. J. 44 (4) (1965) 547-588.

[2] P.A. Estevez, Data projection and visualization using self-organizing neural networks, EVIC 2005.

[3] P.A. Estevez, C.J Fiuer, K. Saito, Cross-entropy embedding of high-dimensional data using neural gas model, Neural Networks 18 (2005) 727-737.

[4] T.J. Ross, Fuzzy Logic with Engineering Applications, McGraw-Hill, USA, 1995.

[5] S. Haykin, Neural Networks and Learning Machines, 3rd ed., Prentice Hall, 2009.

[6] T. Kohonen, O. Simula, Engineering applications of the self-organizing map, in: Proceedings of the IEEE, 1996 Vol. 84, pp. 1368-1372.

[7] S. Obayashi, D. Saisaki, Visualization and data mining of pareto solutions using self-organizing map, in: Second International Conference on Evolutionary Multi-Criterion Optimization (EMO 2003), Faro, Portugal, LNCS 2632, Springer-Verlag Berlin, Heidelberg, 2003, pp. 796-809.

[8] G. Gan, C. Ma, J. Wu, Data Clustering: Theory, Algorithms and Applications, Society for Industrial and Applied Mathematics, 2007.

[9] T.M. Martinetz, S.G. Berlkovich, K.J. Schulten, Neural gas network for vector quantization and its application to time-series, IEEE Transaction on Neural Networks 4 (559) (1993).

[10] S. Nasciento, B. Mirkin, F. Moura-Pires, A fuzzy clustering model of data and fuzzy-means, in: Proceedings of IEEE Conference on Fuzzy Systems 1 (2000) 302-307.

[11] C.D. Meyer, Matrix Analysis and Applied Linear Algebra, Society for Industrial and Applied Mathematics, 2001.

[12] J. Schatzmann, Using self-organizing maps to visualize clusters and trends in multidimensional datasets, Master Thesis, Imperial College, London, U.K., 2003, pp. 17-24.

[13] V. Jakobsen, Data clustering with visualization, Master Thesis in Software Engineering, Department of Informatics, University of Bergen, 2010, Bergen, Norway.

[14] V. Jakobsen, T. Kristensen, Interactive clustering of proteomic data—a comparison between self-organizing map and neural gas, in: Proceedings of 6th Conference on Data Mining, Las Vegas, USA, July 12-15, 2010, pp. 350-356.

[15] T. Kristensen, V. Jakobsen, A.A. Bjørkum, Different visualization schemes of protein data clustering, in: Proceedings of Fourteenth International Conference on Cognitive and Neural Systems, Boston, USA, May 19-22, 2010.

[16] J.H. Gross, Mass Spectrometry, Springer, 2004.

[17] C. Cheadle, M.P. Vawter, W.J. Freed, K.G. Becker, Analysis of micro array data using Z score transformation, J. Mol. Diagn. 5 (2) (2003) 73-81.

[18] O. Simula, P. Vasara, J. Vesanto, R.-R. Helminen, Self-organizing map in industry analysis, in: L.C. Jain, V.R. Vemuri (Eds.), Intelligent Techniques in Industry, 1999, pp. 87-112.

Computer Technology and Application 2 (2011) 828-834

Numerical Analysis of Slag Splashing in a Steelmaking Converter

Miguel Barron and Isaias Hilerio

Departamento de Materiales, Universidad Autonoma Metropolitana Azcapotzalco, Av. San Pablo 180, Col. Reynosa-Tamaulipas Mexico, D.F. 02200, Mexico

Received: August 05, 2011 / Accepted: August 17, 2011 / Published: October 25, 2011.

Abstract: Some variables that influence the slag splashing phenomenon in an oxygen steelmaking converter are numerically analyzed in this work. The effect of lance height, jet velocity, jet exit angle and slag viscosity on the washing and ejection mechanisms of slag splashing is studied employing transient two-dimensional computational fluid dynamics simulations. A parameter here called average slag volume fraction is proposed for the quantitative evaluation of the slag splashing efficiency. Besides, a qualitative comparison is made between the computational fluid dynamics results and physical model results from literature.

Keywords: Computational fluid dynamics, basic oxygen furnace, oxygen steelmaking, refractory lining, slag splashing.

1. Introduction

Wear of refractory lining in oxygen converters for raw steel manufacturing is a factor which greatly influences the production costs in current industrial plants. In recent years slag splashing has emerged as a new technology to extend the lifetime of the refractory lining of converters by reducing the wear associated to thermal and chemical attack by slag and mechanical impact by scrap [1]. In the slag splashing process, molten slag remaining in the bottom of the vessel after the draining of steel is splashed to the converter sidewalls using a supersonic jet of nitrogen. Molten slag freezes at the converter walls forming a protective coating that decreases the wear of the refractory lining. Nitrogen is injected into the converter through a water-cooled vertical lance which has several inclined convergent-divergent nozzles.

Three main stages have been identified in the formation of the slag protective coating: transport of molten slag to the converter sidewalls, adherence of the molten slag to the sidewalls, and freezing and hardening of the slag layer [2]. Related to the transport of the molten slag to the converter sidewalls, two mechanisms are present: wash coating, and ejection coating [3]. The first one occurs due to the bulk movement of the molten slag to rise above the initial level, and the second one due to the ejection of slag droplets which adhere to the vessel sidewalls [4].

Several experimental studies on slag splashing are reported in the literature. These studies employ physical scale models of the converter, and cold liquids and air replacing molten slag and nitrogen, respectively [3, 5-6]. In Ref. [3] it is reported that large nozzle inclination and lance heights increase the splashing, and the main mechanism of splashing changes from ejection to washing as the viscosity of the liquid is increased. In Ref. [5] it is reported that the amount of slag splashed is increased as the lance is raised but beyond at a critical value of the lance height splashing decreases. In Ref. [6] it is shown that when the air flow rate is increased or the liquid viscosity is decreased, the lower regions of the walls are splashed at a much greater rate. It is also shown that by

Corresponding author: Miguel Barron, Ph.D., research fields: process modeling and simulation, synchronization. E-mail: bmma@correo.azc.uam.mx.

changing the lance height, more liquid is deposited in the upper regions of the vessel at the expense of lower regions.

Numerical studies of gaseous jets impinging on a liquid surface are carried out using Computational Fluid Dynamics (CFD) simulations [7-9]. Agreement is reported between numerical and water model results, unfortunately these studies are mainly focused on the jet-surface interaction and the prediction of the surface topography, rather than on liquid splashing. Few CFD studies are reported in the literature on the direct simulation of slag splashing using dimensions of actual converters and properties of molten slags.

In this work it is analyzed the influence of some variables on the slag splashing process using CFD simulations. The variables considered here are the lance height, jet velocity, jet exit angle and slag viscosity. The dimensions of the converter correspond to an actual industrial converter of 150 metric tons. A dimensionless parameter here called *average slag volume fraction* is proposed for the quantitative evaluation of the slag splashing efficiency. A qualitative comparison is made between the CFD results and experimental results from literature.

The rest of the paper is organized as follows: Section 2 describes the slag splashing process, the coating mechanisms and the mathematical model; section 3 provides the physical dimensions of the converter, the values of the considered variables, and gives details about the conditions of the computer runs; section 4 presents and discusses the computer results comparing them with results from literature; finally, the concluding remarks are in section 5.

2. Mathematical Model

In the slag splashing process, molten slag resting in the converter bottom is splashed to the converter sidewalls using a supersonic jet of gaseous nitrogen. The momentum of the nitrogen jet is transferred to the slag, which causes the slag to be stirred and ejected by the action of a standing wave and high shear forces, respectively [3]. Predominance of one of the coating mechanisms, i.e., washing or ejection, depends on factors such as jet characteristics (velocity, exit angle), operating conditions (lance height, molten slag depth) and slag properties (viscosity, density, temperature). Inertial, gravitational, viscous and interfacial forces are acting on gaseous nitrogen and molten slag causing a nonisothermal multiphase flow. To model this complex system, equations which govern fluid flow, heat and mass transfer, turbulence and multiphase flow are required. Fortunately, nowadays such equations are well known, however, numerical solutions are mandatory.

The flow of an incompressible newtonian fluid is governed by the Navier-Stokes equations, which in vector form are expressed as follows [10]:

$$\frac{\partial(\rho u_i)}{\partial t}+\frac{\partial(\rho u_i u_j)}{\partial x_j}=-\frac{\partial p}{\partial x_i}+\frac{\partial}{\partial x_j}\left\{\mu_{eff}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)\right\} \quad (1)$$

where ρ is the fluid density, u_i is the i^{th} component of the fluid velocity u, t is time, x_j is j^{th} spatial coordinate, p is pressure, and μ_{eff} is the effective fluid viscosity. To maintain the mass balance in the system, the continuity equation [10] $\partial u_j / \partial x_j = 0$ must be solved. Turbulence is simulated by means of the classical two equations K-ε model [11]:

$$\rho v_j \frac{\partial K}{\partial x_j}=\frac{\partial}{\partial x_j}\left(\frac{\mu_t}{\sigma_K}\frac{\partial K}{\partial x_j}\right)+\mu_t\frac{\partial v_j}{\partial x_i}\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}\right)-\rho\varepsilon \quad (2)$$

$$\rho v_j \frac{\partial \varepsilon}{\partial x_j}=\frac{\partial}{\partial x_j}\left(\frac{\mu_t}{\sigma_\varepsilon}\frac{\partial \varepsilon}{\partial x_j}\right)+C_1\mu_t\frac{\varepsilon}{K}\frac{\partial v_j}{\partial x_i}\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}\right)-C_2\frac{\varepsilon}{K}\rho\varepsilon \quad (3)$$

In the Navier-Stokes equations the effective viscosity is determined from $\mu_{eff}=\mu_0+\mu_t$,where μ_0 is the laminar viscosity and μ_t is the turbulent viscosity. μ_t is obtained from $\mu_t=\rho C_\mu K^2/\varepsilon$, where K and ε are calculated by solving Eqs. (2)-(3), and C_μ = 0.09. Boundary conditions for K and ε at the inlet nozzle are calculated as follows [12]: $K_{in}=0.01U_{in}^2$, $\varepsilon_{in}=2K_{in}^{3/2}/D_n$

where U_{in} and D_n are the inlet nominal velocity and the nozzle diameter, respectively.

The Volume of Fluid (VOF) model to issue the multiphase flow is based on the assumption that two or more phases are not interpenetrating. For each additional phase q its volume fraction α_q is introduced as a variable. In each control volume the volume fractions of all phases sum to unity. The tracking of the interface between the phases is accomplished by solving the continuity equation for each phase [13]:

$$\frac{\partial \alpha_q}{\partial t} + \vec{v} \cdot \nabla \alpha_q = 0 \quad (4)$$

3. Computer Simulations

The physical dimensions of the considered converter, shown in Table 1, correspond to an actual industrial converter of 150 metric tons of capacity. The coordinate system and meshing of the converter are shown in Fig. 1. Just one nozzle, locate in the left side of the converter, is considered.

Some simplifications are needed in order to tackle the complex numerical task without excessive computer costs. An isothermal two-dimensional system is assumed. The coupled Navier-Stokes equations, the continuity equation, the turbulence model and the VOF model are numerically solved using the CFD technique. Transient isothermal two-dimensional computer simulations are carried out using a time step of 0.0001 s and a mesh consisting of 14075 trilateral cells. Total integration time is 5 s. The variables considered relevant to the slag splashing phenomenon which are studied in this work are the lance height, jet velocity, jet exit angle and slag viscosity. Table 2 shows the values of the above variables employed in the computer simulations, which correspond to actual values in industrial converters.

4. Analysis of Results

Numerical results are shown in this Section for an integration time of 5 s. In Fig. 2 one can observe the distribution of phases in the converter for the following values of variables: lance height, 2 m; jet velocity, 1.5 Mach; jet exit angle, 10°; slag viscosity, 0.1 kg/(m s). The blue and the red phases represent nitrogen and molten slag, respectively. Besides, Fig. 2 is depicted the reference velocity vector. Two recirculatory zones are present. The red vector corresponds to the incoming supersonic nitrogen jet. The two mechanisms of slag coating, i.e., washing and ejection, are clearly seen and indicated in this figure. The slag splashing pattern

Table 1 Dimensions of the converter.

Parameter	Value
Height	7.2 m
Diameter	4.8 m
Lance diameter	0.2 m
Nozzle diameter	0.043 m
Slag depth, h_s	0.5 m
Number of nozzles	1

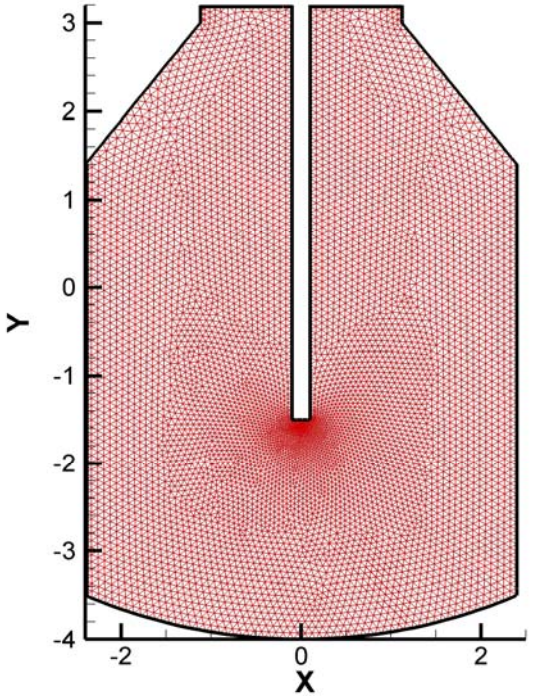


Fig. 1 Coordinate system (m) and meshing of the converter.

Table 2 Values of the considered variables.

Variable	Values
Lance height, h	1,2,3 m
Jet velocity, v	1.5,2 Mach
Jet exit angle, α	5,10,15 degrees
Slag viscosity, μ	0.01,0.1,0.5 kg/(m s)

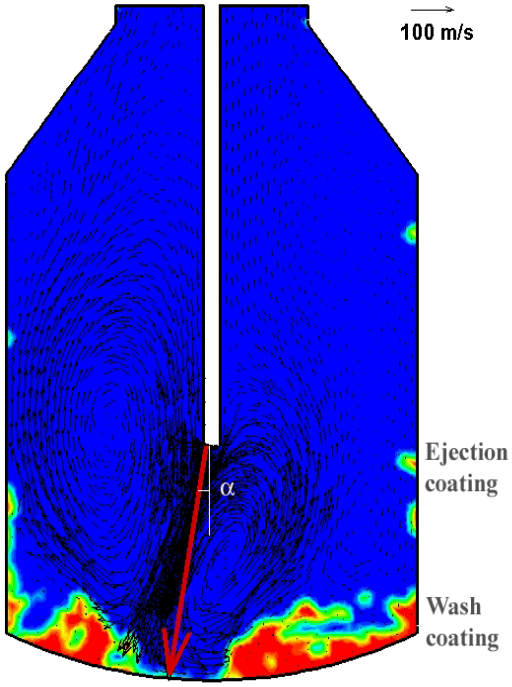


Fig. 2 Distribution of phases and velocity vectors in the converter. Mechanisms of slag coating are marked. Blue phase is nitrogen, red phase is molten slag.

shown in Fig. 2 is similar to those reported in the literature [3-4, 14].

A dimensionless parameter $\bar{v}_s$ here called *average slag volume fraction* is proposed for the quantitative evaluation of the slag splashing efficiency. This parameter is calculated for the left vertical sidewall, and is defined as follows:

$$\bar{v}_s = \frac{\int_{H\,min}^{H\,max} v_s(h)dh}{H_{max} - H_{min}} \quad (5)$$

where v_s is the local value of the slag volume fraction for a given time, and H_{min}= -3.5 m and H_{max}= 1.4 m (see Fig. 1) are the vertical coordinates at the start and end of the left vertical sidewall, respectively. The physical meaning of $\bar{v}_s$ corresponds to an average volume fraction of molten slag in the left sidewall due to the slag splashing by the combined action of washing and ejection mechanisms. This parameter attempts to measure the efficiency of slag splashing in this way: $\bar{v}_s = 0$ means absolutely no coating of the left sidewall, whereas $\bar{v}_s = 1$ implies full coating of the aforementioned sidewall.

The variables studied are the lance height, jet velocity, jet exit angle and slag viscosity with the values considered in Table 2.

4.1 Influence of Lance Height

Figs. 3-4 show the distribution of phases and the average slag volume fraction as a function of the lance height, respectively. Splashing is increased as the lance height is increased, however, beyond a certain value of the lance height, splashing decreases. This behavior is consistent with that reported in physical model experiments [3, 5, 15]. Small lance height causes a deep cavity in the molten slag and a large recirculation zone [4] which favours the prevalence of the wash coating mechanism. Large lance height promotes drop generation [16-17] and the ejection coating mechanism is favoured.

4.2 Influence of Jet Velocity

The depth of the cavity formed in the molten slag depends on the momentum of the nitrogen jet. In its turn, the jet momentum depends on the nitrogen mass flow rate and the velocity. Figs. 5-6 show the effect of the jet velocity on the distribution of phases and slag splashing. As the jet velocity is increased the slag splashing is increased. Low jet velocity promotes the stirring of the slag, and the wash coating mechanism dominates. High jet velocity generates high shearing forces and high drop generation [16-17], therefore in this case the ejection mechanism becomes dominant. Anyway, as the jet velocity is increased the slag coating efficiency is raised, as is seen in Fig. 6. Results of physical models corroborate these findings [4-5].

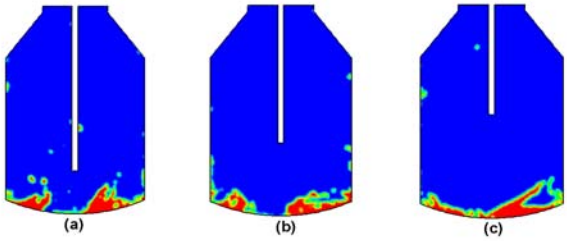


Fig. 3 Effect of lance height on slag splashing. (a), (b) and (c) correspond to lance heights of 1, 2 and 3 m, respectively, for v = Mach 1.5, α = 10°, μ = 0.1 kg/(m s), h_s = 0.5 m.

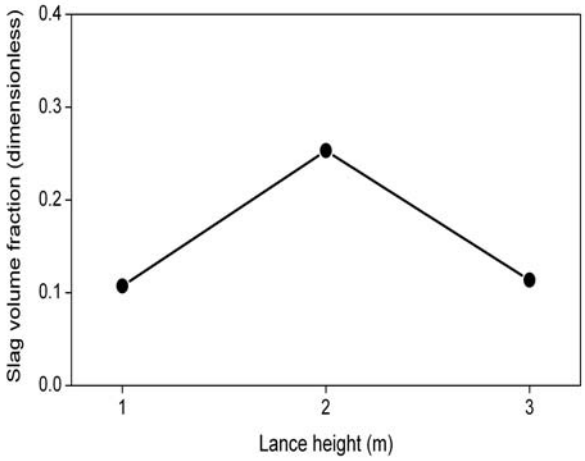


Fig. 4 Average slag volume fraction as a function of the lance height.

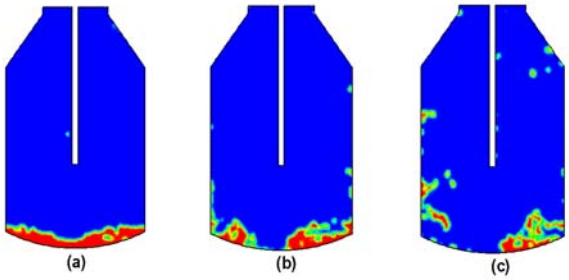


Fig. 5 Effect of the jet velocity on the slag splashing: (a) Mach 1, (b) Mach 1.5, (c) Mach 2, for h = 2 m, α = 10°, μ = 0.1 kg/(m s), h_s = 0.5 m.

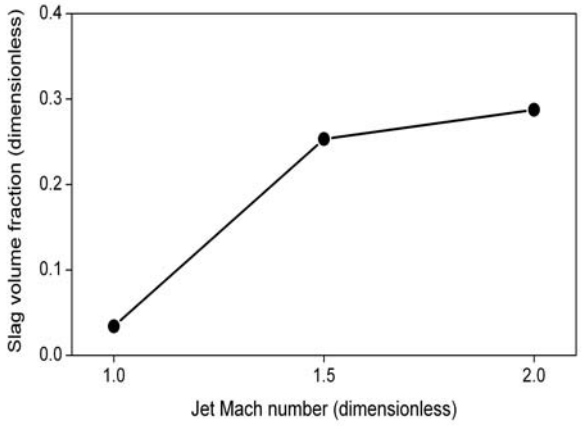


Fig. 6 Average slag volume fraction as a function of the jet velocity.

4.3 Influence of Jet Exit Angle

When the lance nozzle is inclined the nitrogen jet goes out from the nozzle and enters into the converter with an angle α (see Fig. 2). As the exit angle is increased, the vertical component of the jet velocity is decreased and the horizontal component is increased, as is observed in Figs. 7-8. In this case the shearing forces are increased and an overall increase in the amount of slag splashed is observed [3-4]. Of course, when the angle becomes too large and goes beyond a critical value, the jet does not impact the slag anymore. The value of the critical angle depends on the lance height and the molten slag depth. In Ref. [4] it is reported that the critical value of the jet exit angle is 45°. Beyond 45° the slag splashing would be decreased.

4.4 Influence of Slag Viscosity

The influence of the slag viscosity is observed in Figs. 9-10. In accordance to Fig. 10, as viscosity is increased the efficiency of slag coating is decreased. This is explained by the fact that for higher viscosities higher shearing forces are needed for drop generation and the formation of a standing wave. From this point of view, a low viscosity of the molten slag would be desirable, unfortunately slags with low viscosity have little adherence to the converter sidewalls and tend to flow down [4]. On the other hand, the viscosity of a molten slag mainly depends on temperature, composition and the presence of solid phases. Then, in some plants the slag is subjected to a conditioning process for optimizing its viscosity through the addition of materials such as magnesia and ferrous oxide [18]. A comparison among Figs. 4, 6, 8 and 10 show that the slag viscosity is, among the variables studied, the variable which gives the highest slag coating efficiency.

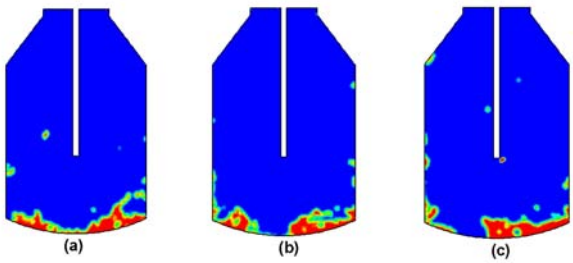


Fig. 7 Effect of the jet exit angle on the slag splashing: (a) 5°, (b) 10°, (c) 15°, for v = Mach 1.5, h = 2 m, μ = 0.1 kg/(m s), h_s = 0.5 m.

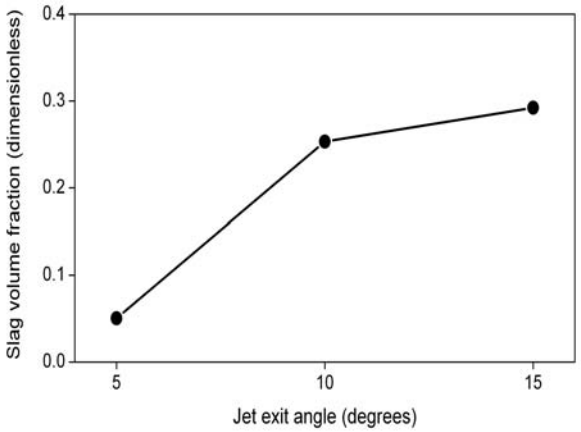


Fig. 8 Average slag volume fraction as a function of the jet exit angle.

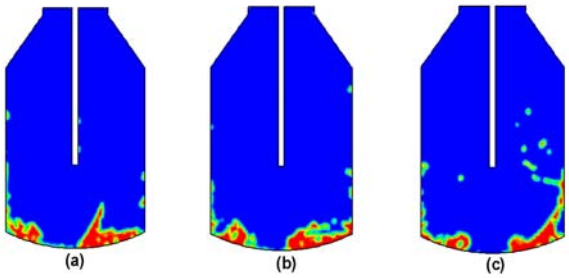


Fig. 9 Effect of the slag viscosity on the slag splashing: (a) 0.01, (b) 0.1, (c) 0.5 kg/(m s), for v = Mach 1.5, h = 2 m, α = 10°, h_s = 0.5 m.

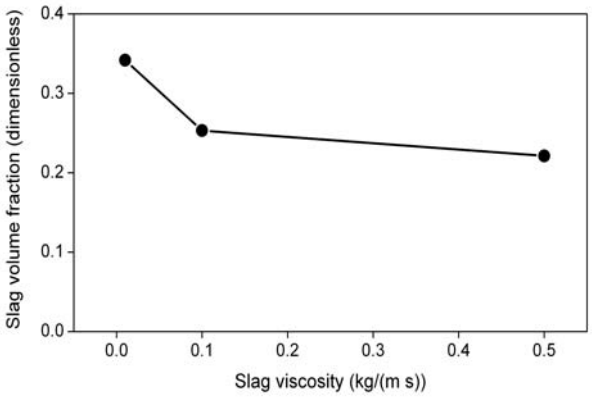


Fig. 10 Average slag volume fraction as a function of the slag viscosity.

5. Conclusions

The slag splashing phenomenon has been studied by means of Computational Fluid Dynamics simulations. A parameter to carry out a quantitative evaluation of the coating efficiency of the slag splashing process is proposed. The effect of variables such as the lance height, jet velocity, jet exit angle and slag viscosity on the coating efficiency is elucidated. Numerical results of this work are in qualitative agreement with those reported in the literature for physical models. Among the studied variables, the most influential one is, apparently, the slag viscosity.

References

[1] C.J. Messina, J.R. Paules, The worldwide status of BOF slag splashing practices and performance, in: Proceedings of ISS Steelmaking Conference, Pittsburgh, PA, 1996, 153-155.

[2] T.R. Galiullin, E.V. Protopopov, V.V. Sokolov, A.G. Chernyatevich, Gas-jet conditions in the slag coating of oxygen-converter linings, Steel in Translation 38 (2008) 97-100.

[3] K.D. Peaslee, Physical modelling of slag splashing in the BOF, Iron and Steel Engineer 73 (1996) 33-37.

[4] K.C. Mills, Y. Su, A.B. Fox, Z. Li, R.P. Thackray, H.T. Tsai, A review of slag splashing, ISIJ International 45 (2005) 619-633.

[5] M.J. Luomala, T.M.J. Fabritius, E.O. Virtanen, T.P. Siivola, T.L.J. Fabritius, H. Tenkku, J.J. Härkki, Physical model study of selective slag splashing in the BOF, ISIJ International 42 (2002) 1219-1224.

[6] K.D. Peaslee, W. Chen, Important factors for effective slag splashing, in: Proceedings of CIM Conference, Edmonton, Canada, 2004.

[7] L. Salinas, R. Fuentes, Interaction between a gaseous vertical descending jet and a liquid surface—a theoretical and experimental study, in: Proceedings of the fourth International Conference COPPER99-COBRE 99 Smelting, Thechnology Development, Process Modeling and Fundamentals, Phoenix, AZ, 1999.

[8] A. Nguyen, G. Evans, Computational fluid dynamics modelling of gas jets impinging onto liquid pools, Applied Mathematical Modelling 30 (2003) 1472-1484.

[9] M. Ersson, A. Tilliander, L. Jonsson, P. Jonsson, A mathematical model of an impinging air jet on a water surface, ISIJ International 48 (2008) 377-384.

[10] R.B. Bird, W.E. Stewart, E.N. Lightfoot, Transport Phenomena, 2nd ed., Wiley, New York, 2002.

[11] B.G. Thomas, Q. Yuan, S. Sivaramakrishnan, T. Shi, S.P. Vanka, M.B. Assar, Comparison of four methods to evaluate fluid velocities in a continuous slab casting mold, ISIJ International 41 (2001) 1262-1271.

[12] G. Solorio-Díaz, R.D. Morales, J. Palafox-Ramos, L. García-Demedices, A. Ramos-Banderas, Analysis of fluid flow turbulence in tundishes fed by a swirling ladle shroud,

ISIJ International 44 (2004) 1024-1032.

[13] C.W. Hirt, B.D. Nichols, Volume of fluid (VOF) method for the dynamics of free boundaries, Journal of Computational Physics 39 (1981) 201-225.

[14] K.M. Goodson, N. Donaghy, R.O. Russell, Furnace refractory maintenance and slag splashing, Iron and Steelmaker 22 (1995) 31-34.

[15] O. Olivares, A. Elias, R. Sanchez, M. Diaz-Cruz, R.D. Morales, Physical and mathematical models of gas-liquid fluid dynamics in LD converters, Steel Research 73 (2002) 44-51.

[16] N. Standish, Q.L. He, Drop generation due to an impinging jet and the effect of bottom blowing in the steelmaking vessel, ISIJ International 29 (1989) 455-461.

[17] Q.L. He, N. Standish, A model study of droplet generation in the BOF steelmaking, ISIJ International 30 (1990) 305-309.

[18] V.A. Sheremet, A.V. Kekukh, S.V. Troshii, A.P. Stovpchenko, A.S. Brodskii, O.I. Pavlyuchenkov, Experience in the use and comprehensive maintenance of converter refractory linings, Refractories and Industrial Ceramics 47 (2006) 75-77.

Computer Technology and Application
Volume 2, Number 10, October 2011
David Publishing Company
1840 Industrial Drive, Suite 160, Libertyville, IL 60048
Tel: 1-847-281-9862; Fax: 1-847-281-9855
http://www.davidpublishing.com
computer@davidpublishing.com