híradástechnika

VOLUME LVIII.

2003/6

Selected Papers

Network Development Concepts

The Quality of Transmission and Switching

Noise Problems

Scientific Association for Infocommunications

Contents

Cover: Sending our messages to north, east, west and south

Foreword

In our English Summer issue, we present a selection of papers published in the first four numbers of this Journal in 2003. In the course of this selection the consideration was to present interesting novelties for the foreign experts, working in Hungary and also for our friends abroad. Our principle governing our semi-annually published English editions was previously the same but now within this general concept we were looking for new possibilities supporting the business success too. The demand for fix terminals decreased, the demand for broadband access networks did not meet the expectations, and the mobile systems with 3G code division multiplexing wideband have not achieved breakthrough either. Due to these experiences, manufacturers and service providers are now both looking for prosperity and for improving business results.

Let us start our survey in the field of network architectures. We are looking for solutions attracting new customers groups and providing access to the network with arbitrary bandwidth anywhere, without any advance booking. Pervasive computing is the first part of a long run development work. At present, we survey and evaluate the possibilities, and our research fellows started a research in order to find solutions combining the features of the existing concepts fulfilling all the expectations. To become economically successful, the concepts of the competitors have to be assessed. The reaction of the users to the tariff and quality proceedings have to be forecasted. To this task we used the Game Theoretical Methods invented by János Neumann.

The growth of wireless network solutions is limited by the number of available frequencies and by the requirement to stay within the scope of international regulations. Two methods will be presented. The first is called the Link Adaptation in MIMO Systems, while the other is strongly connected to this field, but places particular emphasis on secure operation. The preliminary results of the described Space-Time coding procedure are promising.

Another current issue is in close correlation with the improvement of routing and transmission quality. The extremely rapid growth of the available transmission capacity, taking in account George Gilder's statement about the transmission possibilities that are limited neither by distance nor by cost, even so, for the time being, the methods to be applied for accessing these networks are emphasising the switching and the routing technologies. Short-term objectives call also for assessing the quality of IP networks. As seen in the papers on QoS and MPLS, the quality concept applying for these networks cannot be defined as exactly as for line switched systems for which the intelligibility, loss and waiting time can specified be exactly.

In three papers, noise is the critical issue. First, research results applying to component qualification will be presented, and subsequently, the effects of noises will be discussed. It is interesting to note that street noises and their subjective effects are not related linearly. Finally, from the paper entitled "Watermarked Image Demonising", noise reduction methods in picture transmission can be learned, and these methods can be utilised subsequently.

A recession motivated paper in our journal is concerned with the prosperity waves in economics. A memorial paper is dedicated to the centenary of professor László Kozma. His development results are known world-wide, in spite of the fact that he delivered his lectures primarily in Hungary. His everlasting merits in the development of the first register controlled rotary exchanges using revertive impulses are widely acknowledged.

Surveying these samples, arbitrarily selected from Hungarian development results, we can justly expect that in the future, these will affect the market too. However, a prerequisite for this is the establishment of a viable bridge, to be realised between theory and practice, allowing knowledge to be implemented also in materials and services.

Dr. György Lajtha

Pervasive Computing – An Overview

GERGELY BICZÓK, KRISTÓF FODOR, BALÁZS KOVÁCS, ÁGOSTON SZABÓ

students at the Budapest University of Technology and Economics
{bg132, fk206, kb138, sa235}@hszk.bme.hu

Reviewed

At the beginning of the 90's a new field of computer science, the so called ubiquitous computing started to arise based on the ideas of Mark Weiser. The scope of this new field is to invisibly integrate intelligent communicating devices into our everyday lives in a way that the usage of them comes instinctly, without paying real attention on the operation. This article presents the history and the features of this field, including the requirements and the current researches based on the subject.

Introduction

At the early ages of computing many users used to work on the same computer according to some kind of time-sharing model. The costs of the operation were extremely high because of the expenses of frequent hardware failures and the low amount of mainframe computers caused by the unique production. Later came the new era of personal computers, when everyone started to use a computer by his or her own. Development of production technologies, lower prices of hardware elements and the rising demand for comprehensive computer usage could make this change possible.

At this point the usage of modern technologies became no longer the privilege of the military and big research institutes, however, they still were in the first place in developing and introducing new technologies. The appearance of computer networks was a great milestone in the history of computer science because they made the communication simple, data sharing and transmission easy and last but not least, computer usage popular. There is no longer doubt about the importance of the Internet since being a great source of information, an effective medium for delivering data and a source of entertainment as well.

Nowadays the research of portable computer devices is also of great importance. Many companies pay great attention on developing laptops, notebooks and other portable devices because the demand for these equipments has been strongly rising. The reason of this popularity is that the people have got a natural need for the experience of freedom these devices can offer. Because of the special requirements there are problems to be solved such like lowering energy consumption, minimizing size and weight with maximizing speed and capacity. There is a whole industrial branch working on the development and production of portable devices. The new technologies for mobile communication including networks and equipments are still keep spreading around the world.

The appearance of wireless networks together with mobile computing created a whole new aspect in computer science. These two technologies together with integrated services lead to the arising of a new technological paradigm, *pervasive computing*.

This field of computer science is brand new but is obviously of great importance, since the IEEE (Institute of Electrical and Electronics Engineers) started to publish a journal in 2002 titled *Pervasive Computing – Mobile and Ubiquitous Systems*.

The purpose of this article is to introduce this new field of computer science, to describe the circumstances of its evolution and to summarize the achievements so far. It is also of great importance to present the currently running research projects of science institutes and universities and to analyse the expectations and future challenges on pervasive computing applications.

Mark Weiser's vision

The principles of *ubiquitous computing* were first drafted by Mark Weiser. As the leading developer of Xerox he created various prototypes of products based on his unique ideas. He described his vision about the computer of the 21st century in an article [1] published in 1991. He states that information technology should become a natural part of people's everyday lives, usage of devices should be just as evident as, for example, reading. When you read a sign on the street you absorb its information without consciously performing the act of reading. In his opinion the most efficient technologies are the ones that the people, whilst actually using it, are not aware of the usage.

His conception is the opposite of the paradigm of virtual reality, since the latter focuses an enormous apparatus on simulating the world rather than on invisibly enhancing the world that already exists.

In his opinion ubiquitous computing should explore quite different ground from the idea that computers should be autonomous agents that take on our goals [2]. To characterize the difference he describes an example. Suppose you want to lift a heavy object. You can call in your strong assistant to lift it for you, or you can be yourself made effortlessly, unconsciously, stronger and just lift it.

There are times when both are good. Much of the past and current effort for better computers has been aimed at the former; ubiquitous computing aims at the latter.

There has been a big need for developments in the field of mobile computing to actually realize Mark Weiser's conception. Ubiquitous computing needs:

- cheap and low consumption hardware
- some kind of network to connect the devices
- software elements that support distributed operation

At the time he wrote his article the technology needed to implement a system based on his ideas did not yet exist. In the recent decade, a number of scientific achievements were made which are of great importance in connection with materializing Weiser's idea. These include the use of fibre optics in data transmission [21], that provides almost limitless bandwidth, the evolution of human voice controlled systems [22, 23] (which is needed for new generation user interfaces) and the breakthrough in image processing [24].

Nowadays researches based on wireless ad hoc networks are of great interest. Among others, during these researches new technologies are discovered that can be used in the realisation of ubiquitous computing. Various radio interface technologies were developed for supporting communication: IEEE 802.11 [3], HiperLAN and HiperLAN2 [4] and Bluetooth [5, 6]. 802.11 uses MACA (Medium Access Control with Avoidance) [7] in ad hoc mode. There are also various routing algorithms like AODV (Ad hoc On-demand Distance Vector), DSR (Dynamic Source Routing) [8], DSDV (Destination Sequenced Distance Vector) [9] and TORA (Temporarily Ordered Routing Algorithm) [10, 11].

Scenarios on ubiquitous computing

To understand the point of ubiquitous computing Mark Weiser described an example with a girl called Sal, some devices in the background and the intelligent softwares running on these devices.

Sal awakens: she smells coffee. A few minutes ago her alarm clock, alerted by her restless rolling before waking, had quietly asked "coffee?", and she had mumbled "yes." "Yes" and "no" are the only words it knows. The alarm clock sends her request to the coffee machine.

At breakfast Sal reads the news. She still prefers the paper form, as do most people. She spots an interesting quote from a columnist in the business section. She wipes her pen over the newspaper's name, date, section, and page number and then circles the quote. The pen sends the quote to her office.

Electronic mail arrives from the company that made her garage door opener. She lost the instruction manual, and asked them for help. They have sent her a new manual, and also something unexpected – a way to find the old one. According to the note, she can press a code into the opener and the missing manual will find itself. In the garage, she tracks a beeping noise to where the oil-stained manual had fallen behind some boxes. Sure enough, there is the tiny device the manufacturer had affixed in the cover to try to avoid e-mail requests like her own.

Once Sal arrives at work, the foreview helps her to quickly find a parking spot. As she walks into the building the machines in her office prepare to log her in, but don't complete the sequence until she actually enters her office.

The telltale by the door that Sal programmed her first day on the job is blinking: fresh coffee. She heads for the coffee machine.

Coming back to her office, Sal picks up a device and "waves" it to her friend Joe in the design group, with whom she is sharing a virtual office for a few weeks. They have a joint assignment on her latest project. Virtual office sharing can take many forms – in this case the two have given each other access to their location detectors and to each other's screen contents and location. Sal chooses to keep miniature versions of all Joe's shared devices in view and 3-dimensionally correct in a little suite of tabs in the back corner of her desk. She can't see what anything says, but she feels more in touch with his work when noticing the displays change out of the corner of her eye, and she can easily enlarge anything if necessary.

A device on Sal's desk beeps, and displays the word "Joe" on it. She picks it up and gestures with it towards her liveboard. Joe's face appears on the liveboard and they start talking. Joe mentions a colleague called Mary but Sal cannot recall her face. She only remembers that they met on a meeting about a week ago. Sal starts a quick search for the photo and biography of Mary's among the people who participated the same meeting last week as Sal.

This example above shows a world visualized more than ten years ago. In view of the realisations today the visions are nowadays of course different. However, the example above can be a good starting point to understand the concept for the ones who hear about ubiquitous computing for the first time.

In addition to Weiser`s scenario, a lot of other examples can be given. Imagine a person, waiting at the airport who wants to send his previously finished documents as e-mails before his plane takes off. The problem is that the network access point of the gate he is at is highly utilized, a lot of people are browsing the Web. The system detects that there is not enough available bandwith to upload the e-mails before the take-off. It informs the user – through a pop-up window on his PDA's display – about the fact that the access network at the next gate is almost unused, so he can walk there, send the messages and come back comfortably in time.

The system asks the user to set the e-mails in order of importance, to be able to send the most significant mail first. While the user is waiting for the uploading, the system projects his favourite TV comedy to the nearest big screen. After a while it tells the man to head back to his departure gate. The last e-mail is being uploaded during his way back.

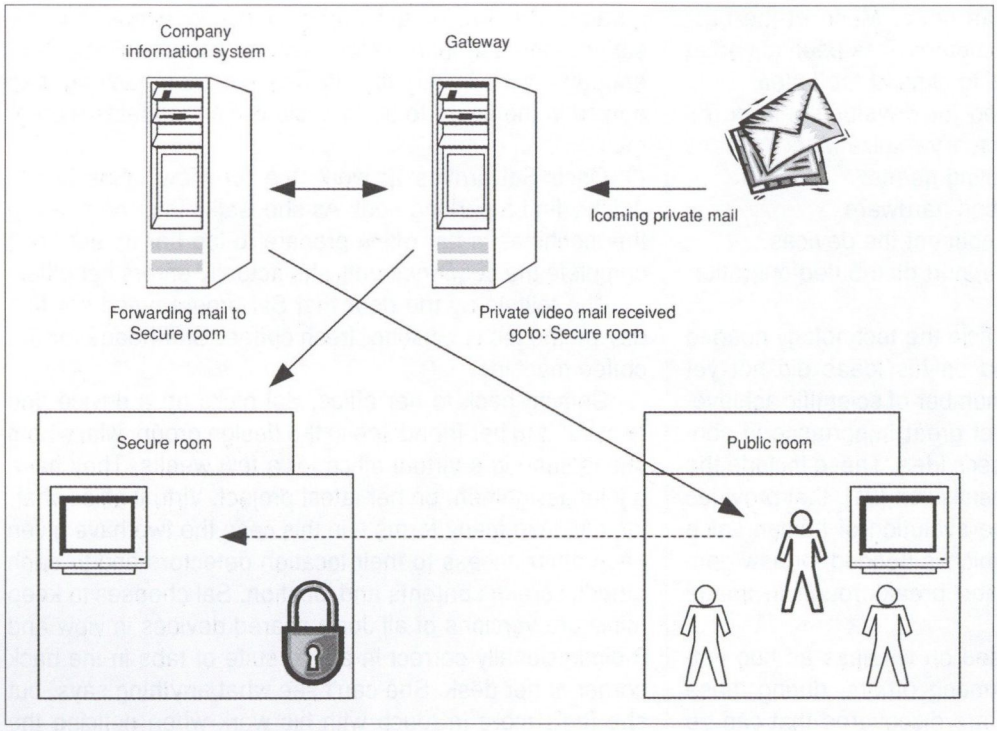


Figure 1 Location- and privacy-aware message redirecting

An other interesting situation is the one when an intelligent system (located inside an office building) redirects the messages (e.g. e-mails), addressed to the employees, by being aware of their actual physical location. It is not a big deal concerning simple text messages, because you can redirect them to the worker's mobile or palmtop, but if it comes to nice images or even videos it is better to watch them on a big screen. In that case, the system alerts the user and guides them to the nearest adequate display. The situation is even more complex if we also consider private or classified messages, you need a screen in a room – which can be locked *(Fig. 1)*.

Ubiquitous or pervasive?

In his articles Mark Weiser named this new field of computer science ubiquitous computing. However, recent documents refer to the same subject as pervasive computing, which naming stems from researchers of IBM. The connection between these two expressions is often mentioned and there is also a whole article dealing with the matter [12]. Some consider the two expressions as synonyms, pervasive computing is simply a new name of pervasive computing. Others think that the two expressions mean quite the same with some differences: pervasive computing is based on a system of small and mobile devices that is used for retrieving information anytime, anywhere (e.g. surfing on the Internet using a cellular phone), while the goal of ubiquitous computing is to hide computer architecture. In this article we consider these expressions as synonyms of each other. In recent documents there is also an another name for this field of computer science – *invisible computing*.

Expectations and challenges

The main goal of pervasive computing is to create a technology that can invisibly assimilate into our everyday lives. There are four main aspects [19] to consider to reach this goal.

The first aspect is the usage and integration of smart spaces. Smart spaces are intelligent computer systems installed in common buildings, rooms, etc. When used efficiently, smart spaces are e.g. able to control the buildings features like heating and lighting of rooms according to the people's position and actions. In another point of view in smart spaces a software used by a person can change behaviour according to the user's position.

The second aspect is *invisibility* – according to the vision of Mark Weiser pervasive computing has got to exclude consciousness from the operation. In practice, a reasonable approximation to this ideal is minimized user distraction. If a pervasive computing environment continuously meets user expectations and rarely presents him or her with surprises it allows interaction nearly on subconscious level.

The third is *local scalability* – as the size of a smart space grows the number of participating devices and hence the number of interactions between the user and the surrounding entities increase. This can lead to lack of bandwidth, more power consumption and hence inconvenience for the users. The presence of multiple users will further complicate the problem. Previous works on scalability ignored physical distance – a web server should handle as many clients as possible regardless of whether they are located next door or across the country. In pervasive computing the number of interactions should decrease if the distance between the user and the smart space increases otherwise the system will be overwhelmed with interactions of little relevance. It is also important to allow users to send requests to a smart space from thousands of kilometers away.

The last one is the *ability of masking areas with uneven conditions*. The penetration of ubiquitous computing is dependent of many non-technical factors like organizational structure, economics and business models. Uniform penetration, if ever achieved, is many years or decades away. Hence the difference between the „smartness" of different areas will be huge. There surely will be offices and buildings with more modern equipment than others. These differences can be jarring to a user, which contradicts the goal of creating an invisible computing infrastruc-

ture. One way to reduce the amount of variation seen by a user is to have his or her personal computing space for „dumb" environments. As a trivial example, a system that is capable of disconnected operation is able to mask the absence of wireless coverage in its environment.

The main arising problems of development and realisation of a pervasive system are:

- tracking user intentions
- exploiting wired infrastructure to relieve mobile devices
- adaptation strategies: applications must adapt to the needs of the system and the system must be able to adapt to the needs of the applications as well (QoS)
- high level energy management, physical and performance planning
- context awareness
- equilibrium between proactivity and invisibility
- security and authentication
- merging of pieces of information from different levels (it might be useful to extend a low-level resource information with a higher level context information)

Ongoing research projects

The goal of this section is to chart the recent efforts in the field of pervasive computing, and to describe the ongoing projects carried out at various universities and IT companies.

Aura

The project Aura [13] is hosted at Carnegie Mellon University (CMU). This university is known for its broad IT research work, they have a number of advanced technologies right at hand which the researchers could integrate into the complex system Aura.

This can effectively speed up implementation so the concept can be materialized faster.

The aim of the researchers is to provide a personalized, invisible information aura for every user involved, which is at service regardless of time and place – that is it moves together with the user.

To accomplish its ambitious goals, research in Aura spans every system level: from the hardware, through the operating system, to applications and end users. Underlying this diversity of concerns, Aura applies two broad concepts. First, it uses *proactivity*, which is a system layer's ability to anticipate requests from a higher layer. Second, Aura is *self-tuning*: layers adapt by observing the demands made on them and adjusting their performance and resource usage characteristics accordingly. Both of these techniques will help lower demand for human attention.

For the actual implementation, some existing technologies *(Fig. 2)* have also been used:

- *Coda* file system provides support for nomadic, disconnectable, and bandwidth-adaptive file access,
- *Odyssey* supports resource registration and monitoring and application-aware adaptation of system elements,
- *Spectra* is an adaptive remote execution mechanism that optimises the execution of remote calls.

Prism is a new system layer that is responsible for capturing and managing user intent. It is layered above applications and provides high-level support for proactivity and self-tuning.

A very interesting video clip can be found on the project's website, it is dealing with the operation of Aura. It can be seen – and heard – that they use human voice for communicating with the system. The test version of the system exists inside the CMU campus.

Oxygen

Oxygen is a project on human-centered computing started at MIT supported by DARPA and Oxygen Alliance [14]. With the help of speech and visual technologies the user can communicate with Oxygen like communicating with a real person and this way they can save a lot of energy. System technologies provides location-independent applicability of user technologies, they can be used at home, in the office or on the move.

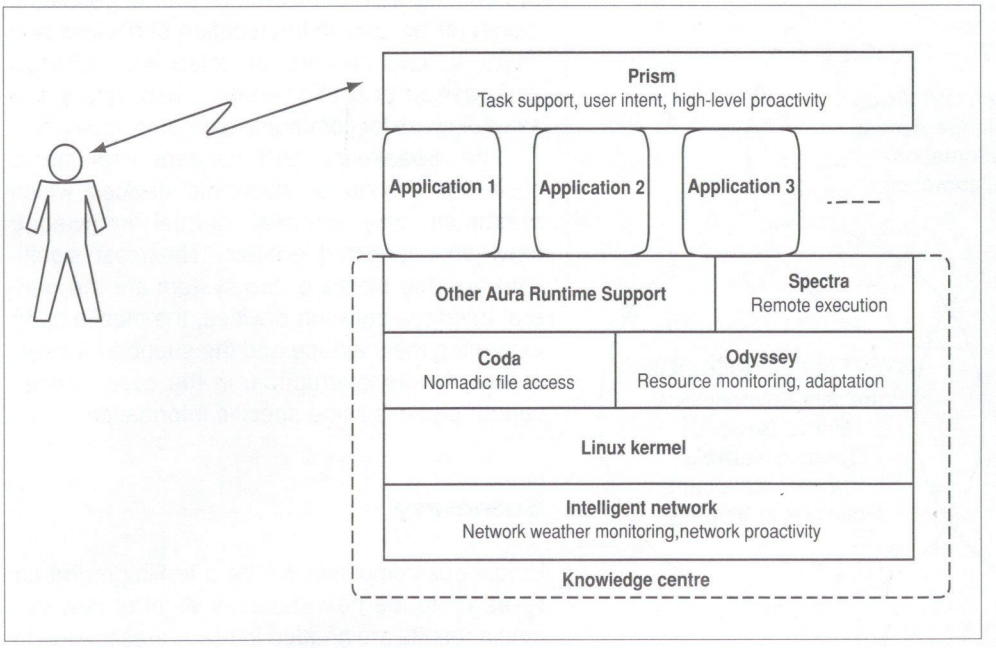


Figure 2
Architecture of the system Aura

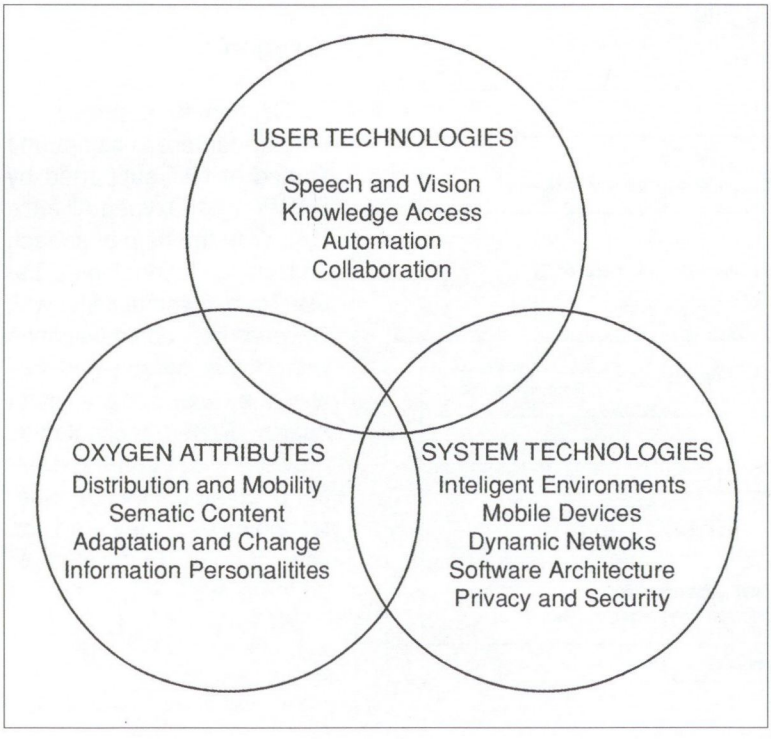
The Oxygen technologies work together and pay attention to several important themes *(Fig. 3)*:

- Distribution and mobility
  – for people, resources, and services.
- Semantic content
  – what we mean, not just what we say.
- Adaptation and change
  – essential features of an increasingly dynamic world.
- Information personalities
  – the privacy, security, and form of our individual interactions with Oxygen.

People access Oxygen through stationary devices (E21s) embedded in the environment or via portable hand-held devices (H21s). These universally accessible devices supply power for computation, communication, and perception in much the same way that wall outlets and batteries deliver power to electrical appliances. Embedded in offices, buildings, homes, and vehicles, E21s enable us to create situated entities, often linked to local sensors and actuators, that perform various functions on our behalf, even in our absence. For example, we can create entities and situate them to monitor and change the temperature of a room. Among other things, H21s can serve as cellular phones, beepers, radios, televisions, geographical positioning systems, cameras, or personal digital assistants, thereby reducing the number of special-purpose gadgets we must carry. To conserve power, they may offload communication and computation onto nearby E21s.

Universally available network connectivity and computational power enable decentralized Oxygen components to perform these tasks by communicating and cooperating much as humans do in organizations.

Figure 3 Oxygen technologies



Components can be delegated to find resources, to link them together in useful ways, to monitor their progress, and to respond to change.

N21 networks support dynamically changing configurations of self-identifying mobile and stationary devices. They allow us to identify devices and services by how we intend to use them, not just by where they are located. They enable us to access the information and services we need, securely and privately, so that we are comfortable integrating Oxygen into our personal lives. N21s support multiple protocols for low-power communication.

The software architecture matches current user goals with currently available software services, configuring those services to achieve the desired goals. When necessary, it adapts the resulting configurations to changes in goals, available services, or operating conditions. Thereby, it relieves users of the burden of directing and monitoring the operation of the system as it accomplishes its goals.

Other projects

Besides the two projects mentioned above, there are more, significant researches going on. The system called *Portolano* [15, 16] is being developed at the University of Washington. It mostly focuses on mapping out a next generation user interface, resource discovery processes and data-driven networks.

The most important feature of the project *Endeavour* [17] (University of Berkeley) is the support of fluid software. This revolutionary concept means that computing and other software functions existing in the network distribute themselves automatically among the devices. Another feature is the exceptional system availability.

Of course, Microsoft cannot be out of exploring this field of computing getting popular. In 1998 they started to work on a study called *EasyLiving* [18] which is now an operating test system. Its core components are: a geometric model for describing the location of devices and users, various sensors for determining location and position of entities, service descriptors and a middleware for communication among devices.

The *Speakeasy* [20] concept implements such cooperation of electronic devices which presumes only minimal mutual knowledge between connected entities. The most significant building blocks of the system are the general interface between devices, the mobile code extending the interface and the support for user-in-the-process interruption, in the case devices cannot process some specific information.

Summary

Ubiquitous computing will be a fertile ground for research in the next decades. A lot of new scientific results are needed in many areas, even in

those that are not closely related to computer systems, if they are intended to transform a dream to reality. These areas include human-computer interactions (specifically focusing on the variety of interfaces and human-centered hardware design), software agents (with paying attention to high-level proactive behaviour) and artificial intelligence (concentrating on decision-making and planning).

The capabilities originating from these areas should be integrated into the future systems that are able to fulfil the four major requirements mentioned earlier in this article. So, pervasive computing comes to existence as an integration of results achieved in a number of separate fields of science.

Acknowledgement

The authors would like to thank the precious help of Miklós Aurél Rónai, Zoltán Turányi and András Valkó.

References

[1] Mark Weiser – The Computer for the 21st Century, Scientific American, September 1991.
[2] Mark Weiser – Some Computer Science Issues in Ubiquitous Computing, Communications of the ACM, July 1993.
[3] IEEE Std 802.11, 1999 Edition, http://standards.ieee.org/catalog/olis/lanman.html
[4] HiperLAN2 overview, http://www.hiperlan2.com/WhyHiperlan2.asp
[5] Jaap Hartsen – BLUETOOTH – The universal radio interface for ad hoc, wireless connectivity, Ericsson Review No. 3, 1998
[6] Bluetooth Baseband Specification, http://www.bluetooth.com
[7] Phil Karn – MACA - A New Channel Access Method for Packet Radio, appeared in the proceedings of the 9th ARRL Computer Networking Conference, London, Ontario, Canada, 1990
[8] S. Das, C. Perkins, E. Royer – Performance Comparison of Two On-demand Ad hoc Routing Algorithms, Proceedings of the IEEE Conference on Computer Communication, March 2000.
[9] C. Perkins, P. Bhagwat – Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers, SIGCOMM'94
[10] Vincent D. Park and M. Scott Corson – A Highly Adaptive Distributed Routing Algorithm for Mobile Wireless Networks, Proceedings of IEEE INFOCOM '97, Kobe, Japan, April 1997.
[11] J. Broch, D. A. Maltz, D. B. Johnson, Y. Hu, J. Jetcheva – A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols, MobiCom '98.
[12] Anne McCrory – Ubiquitous? Pervasive? Sorry, they don't compute, Computer World, March 2000.
[13] D. Garlan, D. P. Siewiorek, A. Smailagic, P. Steenkiste – Aura: Toward Distraction-Free Pervasive Computing, IEEE Pervasive Computing, 2002.
[14] MIT Project Oxygen, Online Documentation, http://oxygen.lcs.mit.edu/publications/Oxygen.pdf
[15] M. Esler, J. Hightower, T. Anderson, G. Borriello – Next Century Challenges: Data-Centric Networking for Invisible Computing. The Portolano Project at the University of Washington, Mobicom '99.
[16] Portolano/Workscape: Charting the new territory of invisible computing for knowledge work, On-line Documentation, http://portolano.cs.washington.edu/proposal/
[17] The Endeavour Expedition: Charting the Fluid Information Utility, Online Documentation, http://endeavour.cs.berkeley.edu/proposal/
[18] B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer – EasyLiving: Technologies for Intelligent Environments. Handheld and Ubiquitous Computing, September 2000.
[19] M. Satyanarayanan – Pervasive Computing: Vision and Challenges, IEEE Personal Communications, August 2001.
[20] W. K. Edwards, M. W. Newman, J. Sedivy, T. Smith Challenge – Recombinant Computing and the Speakeasy Approach, MobiCom'02, September 23-28, 2002, Atlanta, Georgia, USA.
[21] George Gilder – Dark Fibre, Dumb Network, Forbes ASAP, December 1993.
[22] P. Tatai – Open Vocabulary Speech Recognition - Brief State Report on a Research Project, Proceedings of the Polish-Czech-Hungarian Workshop on Circuits Theory, Signal Processing and Applications, September 3-7, 1997, Budapest, pp. 52-57.
[23] Olaszy, G., G. Gordos and G. Németh, "The MULTIVOX multilingual text-to-speech converter", in: G. Bailly, C. Benoit and T. Sawallis (eds.): Talking machines: Theories, Models and Applications, Elsevier, 1992, pp. 385-411.
[24] T. Roska and L. O. Chua, "The CNN Universal Machine: An analogic array computer", IEEE Transactions on Circuits and Systems-II, Vol. 40, pp. 163-173, March 1993.

Game Theoretical Methods

KRISZTINA LÓJA

PhD student
Budapest University of Technology and Economics, Department of Telecommunications and Telematics
loja@math.bme.hu

Reviewed

Game theory is a possible way to model the evolving competition in telecommunications. The aim of this article is to offer a survey of the basic game theoretical methods, illustrating them with a few examples from the area of telecommunications.

With the liberalization of telecommunications an oligopolistic market evolves in both the wired and the mobile telecommunications. While in the case of monopoly or free competition the investor has to maximize a definite function, in an oligopolistic situation he has to take the decision of the other participants into account. To model this problem, non-cooperative game theory provides a repertoire of techniques. We consider a game non-cooperative, if the participants cannot do binding agreements.

First, the Nash-equilibrium, the central equilibrium concept of game theory is introduced. This part illuminates the problems about the pure strategy Nash-equilibrium.

Next we define the mixed strategy Nash-equilibrium with the use of mixed strategies.

In the third section we describe the iterated elimination of dominated strategies which is applicable to static games and gives a Nash-equilibrium, and in the fourth we show a method for dynamic games which leads to a refinement of the Nash-equilibrium.

After that we mention an example of the appearance of a chaotic phenomenon, finally we introduce the Harsányi-transformation, which is applicable to games with incomplete information.

I. Nash-equilibrium

Let the integer $n \geq 2$ be the number of players and let A_k be the strategy set of the k-th player. The strategy is a decision with one or more steps that determine the behavior of the players in every possible situation.

The players decide independently, the k-th player chooses the strategy $a_k \in A_k$ and so his utility is the real number $\pi_k(a_1, \ldots, a_k, \ldots, a_n)$. The strategy profile of the players forms a Nash-equilibrium if none of them has the incentive to deviate unilaterally from his own strategy of the equilibrium profile. Let $a_{-k}=(a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_n)$ the vector originating from the vector $a=(a_1, \ldots, a_k, \ldots, a_n)$ by dropping its k-th element.

Then the strategy profile $(a_1^*, \ldots, a_n^*)$ forms a (weak) Nash-equilibrium if $\pi_k(a_k^*, a_{-k}^*) \geq \pi_k(a_k, a_{-k})$ for all strategies $a_k \in A_k$ $(k=1, 2, \ldots, n)$.

In the case of strict inequalities the strategy profile forms a strong Nash-equilibrium.

The strategy $a_k^* \in A_k$ is the best reply of the k-th player to the $a_{-k} \in A_{-k}$ strategy profile of the other players, if a_k^* maximizes the utility of the k-th player in the case of $a_{-k} \in A_{-k}$. So Nash-equilibrium means that every player gives a (not necessarily unique) best reply to the others' equilibrium strategies.

Let us consider a duopoly for example, in which the two firms make a decision about the magnitude of the network [1]. Let us assume, that the first company has three and the second has four alternatives and to the possible strategies with higher serial number belong a greater magnitude network. The payoffs can be seen in the payoff matrix in *Table 1* (in the cells the first number is the utility of the first company and the second is the utility of the second company).

		Company 2			
		1.	2.	3.	4.
Company 1	1.	2, 6	1, 7	1, 6	0, 5
	2.	3, 5	**2, 6**	1, 5	0, 4
	3.	3, 4	1, 5	1, 4	-1, 2

Table 1 An example of the Nash-equilibrium

If both firms choose the second alternative then a Nash-equilibrium evolves, the payoff of the first firm is 2 and that of the other is 6 and it cannot be improved by neither of them with unilateral changes. (The utilities belonging to the Nash-equilibrium are written here and later on with bold numbers.)

The Cournot-equilibrium and the Bertrand-equilibrium of the oligopol markets are also Nash-equilibria. They determine the price and the product quantity respectively in such a way that no firm has the incentive to deviate from them.

There are several problems with the Nash-equilibrium. There are games without any pure strategy Nash-equilibrium (pure strategy is defined in section II.). An example of it is the rock-paper-scissors game *(Table 2)*. The two players choose simultaneously rock, paper or scissors and if their choices are the same, there is a draw with 0 payoff to both players, in the case of different choices scissors beats paper, rock beats scissors and paper beats rock. The utility of the winner is 1 and that of the loser is –1.

Table 2		Player 2		
		Rock	Paper	Scissors
Player 1	Rock	0,0	-1,1	1,-1
	Paper	1,-1	0,0	-1,1
	Scissors	-1,1	1,-1	0,0

Table 3		Wife	
		Prize fight	Ballet
Husband	Prize fight	**4, 3**	2, 2
	Ballet	1, 1	**3, 4**

Table 4		Prisoner 2	
		Deny	Confess
Prisoner 1	Deny	-1,-1	-10,0
	Confess	0,-10	**-8,-8**

Table 5		Player 2	
		Altruist	Egoist
Player 1	Altruist	3,3	0,4
	Egoist	4,0	**1,1**

Table 2 Rock-paper-scissors
Table 3 Battle of sexes
Table 4 Prisoners' dilemma
Table 5 A version of prisoners' dilemma

It is easy to see that the two players cannot make a decision with no incentive to deviate from it for any of them, so this game has no (pure strategy) Nash-equilibrium.

We can give examples of games in which there exists a Nash-equilibrium, but it is not unique. The game called battle of sexes is a game of this kind. There are two players: the husband wants to go to the prize fight and the wife wants to go to the ballet. It is more important for each of them to be together than their own preferences, but they do not have the opportunity to discuss. The payoff matrix can be seen in *Table 3*.

It is a Nash-equilibrium if both players go to the prize fight and it is also one if they go to the ballet.

There are more realistic examples with the same payoff matrix. Standardization is a problem of this kind. Compatibility of their products is more important for the firms then their own interest. In this case, it is a Nash-equilibrium if all firms adapt themselves to one of them, so they all make products compatible with the products of the chosen firm.

The real problem is not the multiplicity of Nash-equilibria, but our unability to choose from them. There are games with many Nash-equilibria in which the players have some kind of principle to help them to choose a strategy profile.

The game described by Kreps [2] is one of this kind. The two participants get a list of cities from which both of them have to choose a subset with a given element. If the two lists form a correct partition of the original list (so every city is on the lists but not in both), then both participants get a certain amount of money, otherwise they get nothing. In this game every partition is a Nash-equilibrium, they are equivalent in game theoretical point of view, but the two participants can choose only one of them without any previous negotiation, for example with political or geographical considerations. The equilibrium, which can be chosen by the players without any previous discussion from the many Nash-equilibria with guidelines outside of game theory, is called the focal point.

If there exists a Nash-equilibrium and it is unique, it is not certain that it is optimal. The prisoners' dilemma *(Table 4)* is the best known example of this. The story is the following: two suspects are arrested by the police and they are charged of commiting a crime together, but there is not enough evidence against them. They are locked in separate cells (so they cannot communicate) and they have the following opportunities: if neither of them confesses, both are sentenced to one year each, if both confess they are sentenced to 8 years each. If one of them confesses and the other does not, then the former is free and the latter gets 10 years of prison.

This game has a unique Nash-equilibrium, but it is not optimal in Pareto-sense, so there is a pair of strategies which gives more payoff to both players. The equilibrium is (confess, confess), so that each gets a sentence of 8 years although they could get off with one year each.

An alternative way to illustrate this situation is the following *(Table 5)*. Both players have to decide between getting 100 dollars or let the other get 300 dollars. If both players are altruists, both will have 300 dollars, but in the Nash-equilibrium both get only 100 dollars. This game is absolutely equivalent with the game described in table 4.

It is a game with similar construction when the two firms of a duopoly are deciding about the amount of money spending for advertisement. Let us suppose, that the two firms have two alternatives, they spend either a great or a small amount of money for advertisements. When one firm advertises a lot and the other does not, the better known company wins the greater part of the market. Although they are best off when neither of them spends a lot for advertisements, the Nash-equilibrium is spending a lot for both even though it will not be refunded, and this happens in practice, too.

In the prisoners' dilemma, there is a Nash-equilibrium that is not optimal but it is worth to choose, because it is a best reply to all possible decisions of the other players. We can construct a game, where the only Nash-equilibrium will not be choosen by any of the players.

A game of this kind, suggested by Kreps [2], is the following: there are two players, they have to choose independently between X and Y. If both choose Y, they both get 1 dollar, and if one of them chooses X and the other Y then they get nothing, if both of them chooses Y, then in a second level of the game they both have to choose a positive integer. Whoever chooses the greater one gets 250 dollars, the other gets 100 dollars.

The only Nash-equilibrium here, (Y, Y), will probably not be chosen by the players, since (X, X) gives at least 100 dollars for both of them. (Y, Y) with the low utility can be a unique Nash-equilibrium here, because the game following the choice (X, X) does not have any Nash-equilibrium.

If (100, 100) payoffs belonged to the decision pair (X, X) without any further decision, then (X, X) would be the other (the Pareto-optimal) Nash-equilibrium of the game.

II. Mixed strategy Nash-equilibrium

In the examples until now, the players chose a strategy which they followed with probability 1. This is what we call pure strategy. John von Neumann introduced mixed strategies, where the choice among the possibilities is based on chance.

Let P_i be a probability distribution over the strategies of the i-th player. Let the real number $p_i(a_j) \geq 0$ represent the probability of choosing the strategy a_j. For the sake of simplicity, let there be only two players. Then the expected value of the utility of the i-th player is depending on the probability distributions P_1 and P_2.

The Neumann-Morgenstern utility function of the i-th player (i=1, 2) is the following:

$$\pi_i(P_1, P_2) = \sum_{a_1 \in A_1} \sum_{a_2 \in A_2} p_1(a_1) p_2(a_2) \pi_i(a_1, a_2)$$

The Nash-equilibrium can be expanded to mixed strategies in a natural way. The game rock-paper-scissors introduced earlier (with no pure strategy equilibrium) has a unique Nash-equilibrium, where p_i(rock) = p_i(paper) = p_i(scissors) = 1/3, (i =1, 2), so both players choose all the possibilities with 1/3 probability.

As the following theorem shows, a Nash-equilibrium exists in general. According to the theorem of Nash, every n-person game has at least one pure strategy Nash-equilibrium, if the A_k strategy set is a non-empty convex and compact set in a finite dimensional Euclidean space for all k, and the utility $\pi_k(a_1, \ldots, a_k, \ldots, a_n)$ of the k-th player is continuous in every variable and quasiconcave in a_k (k=1, 2, ..., n). We call a function $f: R^n \to R$ quasiconcave, if its domain is an $x \subseteq R^n$ convex set and all of its upper level-sets $\{ x \in X: f(x) > t \}$ are convex.

According to the theorem of Nash regarding the mixed strategies, if the original (pure) strategy sets are finite, then the game defined on the expanded strategy sets evolving by applying mixed strategies, has at least one mixed strategy Nash-equilibrium.

III. Elimination of dominated strategies

Iterated elimination of dominated strategies

Let $a_1^* \in A_1$ and $a_1 \in A_1$. We say, that a_1^* weakly dominates a_1, if for all $a_2 \in A_2$ strategy the utility of player 1 applying a_1^* is greater or equal with the payoff of choosing a_1, that is $\pi_1(a_1^*, a_2) \geq \pi_1(a_1, a_2)$ for all $a_2 \in A_2$, and there exists an $a_2' \in A_2$, so that $\pi_1(a_1^*, a_2') > \pi_1(a_1, a_2')$. The a_1^* strategy strictly dominates a_1, if $\pi_1(a_1^*, a_2) > \pi_1(a_1, a_2)$ for all $a_2 \in A_2$. In the case of the prisoners' dilemma, for example, confessing strictly dominates denying, because if the other confesses, the first prisoner gets 10 years by denying, but only 8 when he confesses. If the other denies, the first receives one year in prison, when he denies, but he is set free when he confesses.

It is a general assumption in game theory, that the participants are rational, so we assume, that they do not choose a strategy that gives less payoff in every case, over another one, so we can omit the dominated strategies from the examined game. From the game we get in this way, we can also omit the dominated strategies. If we continue this process, it can happen that only one strategy profile remains. Then it is surely a Nash-eqilibrium.

		Player 2		
		x	*y*	*z*
Player 1	*u*	6,3	1,5	0,6
	v	1,7	**2,8**	2,6

Table 6 Iterated elimination

Consider for example the game illustrated in *Table 6*. Since *y* dominates *x*, we can assume, that player 2 will not choose *x*. If we leave out strategy *x*, considering the remaining possibilities, *v* dominates *u*. If we omit *u*, then *y* dominates *z*. So *(v, y)* is the only remaining strategy profile and it is the unique Nash-equlibrium of the game.

This method cannot always be applied. There are games, in which this procedure narrows down the game of examination, but does not lead to a single strategy profile and we can give examples to games, in which no strategy dominates any other. The game called battle of sexes is one of this kind. In that game, there are two pure (and one mixed) strategy equilibria, and this method does not make it easier to choose between them.

		Player 2	
		u	*v*
Player 1	*x*	**1,1**	0,0
	y	0,0	**10,10**

Table 7

In the case of the game depicted in *Table 7*, although there are two Nash-equilibria, the significant difference between the utilities belonging to them helps us to choose between them. On this basis, players will choose the strategy profile *(y, v)*, but the iterated elimination of dominated strategies does not lead to this result, since this game does not have any dominated strategy, *y* does not dominate *x*, and *v* does not dominate *u*.

If we omit only the strictly dominated strategies, then if an equilibrium exists, it is unique. It is not so in the case of iterated elimination of weakly dominated strategies. Rasmusen [3] gives an example for this with the following game *(Table 8)*.

In this example, eliminating the weakly dominated strategies in different orders gives different results. If the order of elimination is s_3, o_3, o_2, s_2, then we get the (s_1, o_1) strategy pair, but the order o_2, s_2, o_1, s_3, gives (s_1, o_3).

		Player 2		
		o_1	o_2	o_3
Player 1	s_1	2,12	1,10	1,12
	s_2	0,12	0,10	0,11
	s_3	0,12	0,10	0,13

Table 8

One-step elimination of dominated strategies

The strategy profile that we get by deleting all the weakly dominated strategies of every player from the game is called a weakly dominated equilibrium. Since, in this case, elimination occurs in one step (so there is no strategy we can eliminate because we already deleted another strategy) with this procedure, at least as many strategies remain as in the iterated case, for example, in the previous game the strategies belonging to the two Nash-equilibria.

IV. Dynamic games

In the examples until now we used the normal form, so we described the games with their payoff matrix. This is a widely used depiction of games, in which the players decide simultaneously. The other description is the tree structure, the extensive form of which is mainly used for sequential (dynamic) games, where the players make decisions in a certain order, knowing the moves the other players made so far. The static games we described so far can also be depicted in extensive form, but the players do not always know in which decision point of the tree they are, so the introduction of information set is needed. The information set of one of the players is the set of decision points, among which the player cannot determine, which the actual decision point is.

In the next example there are two players, both can choose between two possibilities, they can move either left or right. First the first player moves and then, knowing this decision, moves the other. In the figures, the empty circles show the decision points, the payoffs can be seen under the terminal nodes indicated with solid circles.

Applying the backwards induction (in some literature this is called Kuhn-algorithm), we first consider the player, who makes a choice last, and assume that he chooses the outcome of the game in every decision point, that gives him the higher payoff. The player who makes the decision before the last takes only this outcome into account and decides accordingly. Getting along backwards from the end of the game, the course of the game is outlined, and that results in a Nash-equilibrium.

In the example of *Figure 1*, if the first player moves right, then the payoff of the second is 6 if he moves right and 4 moving left, so we assume that he moves right in this decision point. If the first player moves left, then the payoff of the second is 2 moving left and 1 moving right, so we assume that he moves left. Considering all these, the first player moves left and then the second moves left as well, so their payoffs are (5, 2). This way we have reached the unique Nash-equilibrium of the game, though it is not Pareto-optimal, because the decision pair (right, left) would result in a payoff of (10, 4). The utility of both players would be doubled, but that outcome is not stable, because it is worth for the second player to deviate improving his utility to 6.

In general, this in an efficient way to solve a sequential game, but this method also has its limits, of course. For example, the tree of the game can be infinite. In that case, we cannot even begin this procedure. There are such games, to which backwards induction is applicable in principle, but in practice it is unrealisticly complicated. Chess, for example, is a game like this, its tree is finite but we still cannot determine the optimal strategy with this procedure.

This method has the advantage of eliminating the credible threats. Consider the following example *(Figure 2)*: there are two players in a monopol market, the incumbent and a potential new entrant. Let the latter be Player 1. He moves first, he has two possibilities: he either enters the market or not. If he does not enter, the game is over, his payoff is 0 and the payoff of the monopolist is 20. If he enters, then, knowing this decision, Player 2, the monopolist, has two choices. If he decreases the monopolistic-price according to the duopoly, both firms will have a utility of 4. The other possibility of the monopolist, which pros-

Figure 1

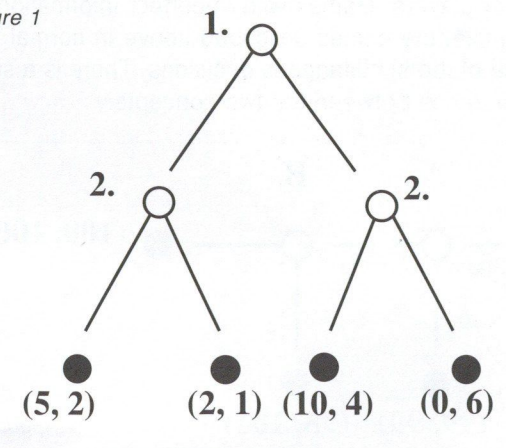
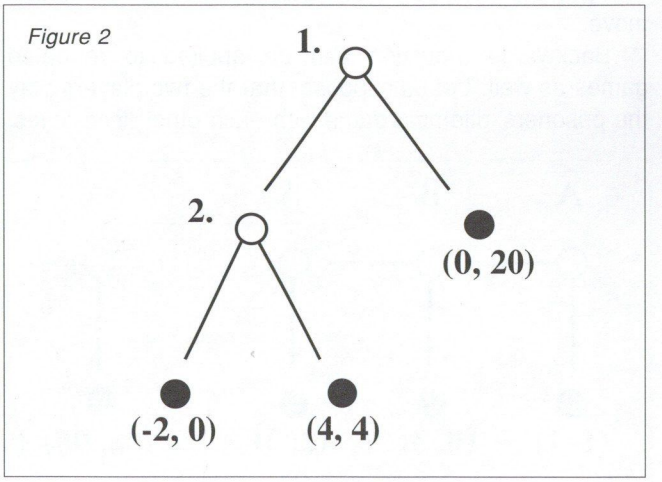


Figure 2



pect he holds out to deter the entrant, that he chooses an irrationally low price, so his payoff will be 0 and the other firm goes bankrupt, his payoff will be –2. Player 1 has to reckon the risk of bankruptcy. But the threat of the incumbent is uncredible, because if the entry has already happened, he will maximize his own profit and will not choose a price less than the duopolistic-price. Bearing this in mind, the new firm, applying backwards induction, will enter the market.

A refinement of Nash-equilibrium for sequential games is the equilibrium conception of subgame perfect Nash-equilibrium (linking with the name of Selten), which eliminates uncredible threats. The subgame of a game is the remaining part of the game, that has already begun, which begins in a decision point of a place known by all the players and includes, in addition, all the decision points following this one, and the corresponding payoffs. So a subgame is a subtree, extending to the terminal nodes and known by all the players. A strategy profile is a subgame perfect Nash-equilibrium, if it is a Nash-equilibrium of the whole game and its corresponding decisions give a Nash-equilibrium for every subgame. Backwards induction gives a constructive proof for the existence of the subgame perfect equilibrium.

The Stackelberg-equilibrium of duopolistic markets is also a subgame perfect Nash-equilibrium. In this case, the two firms determine the product quantity sequentially, so that their decisions are best replies to the other.

We can construct games, so that the solution forecasted to them by backwards induction is right well dubious. The following game depicted in *Figure 3*, the centipede game is linked with the name of Rosenthal.

The tree of the game is finite and known by both players (A and B), so backwards induction can be applied. At the last point (if the game reaches it), Player B will move downwards, because this way his payoff is 101, but it would be 100 if he moved right. For this reason, Player A will move down in the last but one point, because so his payoff is 99, and if he moved right (since B moved down), then his payoff would be 98. Continuing the backwards induction, we find that Player A moves down in the first decision point of the game, the game is over, the utilities of both A and B is 1. Kreps [2] proved with experiments that players who begin this game, hardly ever do this move.

Backwards induction can be applied to repeated games as well. Let us suppose, that the two players play the prisoners' dilemma game with each other finite times, (one hundred times, for example). In the last round, both players will confess, maximizing their own utility. For this reason, neither of them will deny (cooperate) in the round before the last. According to backwards induction, both players will confess in the whole game from the first round, although the experiments show that in this situation cooperation evolves in general.

V. Chaotic phenomena

We can meet chaotic phenomena during the examination of quite simple games. Sato, Akiyama and Farmer [4] show an example for this. Let us consider a generalized version of the already described game rock-paper-scissors *(Table 9)*.

It differs from the original only in the case of draw, where the utility of the first player is ε_x, that of the second is ε_y, where $-1<\varepsilon_x<1$ and $-1<\varepsilon_y<1$. Let us assume, that the two players play this repeatedly throughout several rounds. Let us suppose, that they do not play the optimal strategy maximizing their utility, so they are not rational, but boundedly rational, they develop their strategies over the course of the game. They modify the probabilities in their mixed strategies according to whether the decisions resulted in a win. In this case the sequence of the decisions of the players follows chaotic dynamics.

		Player 2		
		Rock	Paper	Scissors
Player 1	Rock	ε_x, ε_y	-1,1	1,-1
	Paper	1,-1	ε_x, ε_y	-1,1
	Scissors	-1,1	1,-1	ε_x, ε_y

Table 9 Generalized rock-paper-scissors game

VI. Harsányi-transformation for games with incomplete information

In the games so far, the players knew exactly the whole structure of the game, the opportunities of the others and the payoffs. We call games like this games with complete information. We call a game a game with perfect information, if every player knows the decisions made so far by the other players. Games with imperfect information are, for example, the games described above in normal form, because of the simultaneous decisions. There is a significant difference between the two concepts.

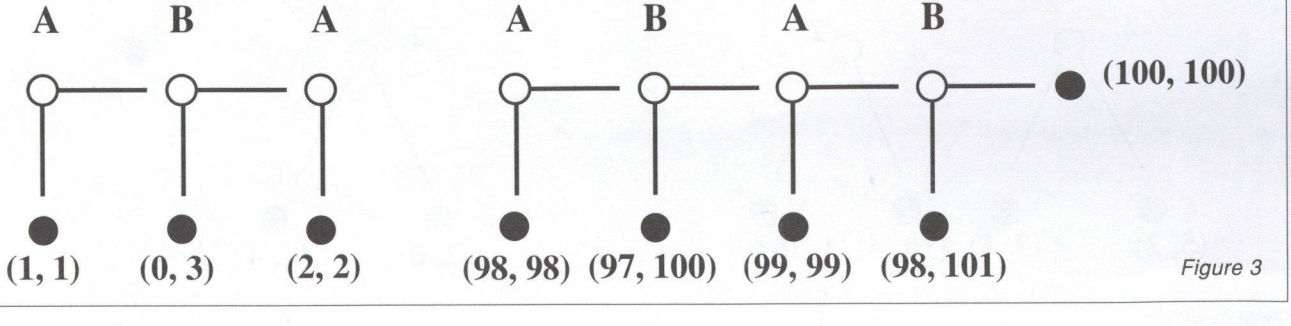


Figure 3

Game theory, as Neumann and Morgenstern created it in 1944 and every game theoretical work until the end of the 1960s, restricted themselves to the examination of games of complete information. Between 1964 and 1970 the U. S. Arms Control and Disarmement Agency employed game theory experts, János Harsányi among others. He was the first who could handle the games with incomplete information.

In this situation, the two players are the American and the Soviet side. Both of them knows only his own position and possibilities, they are not fully aware of the political aim, the military force, peaceful or aggressive intention and other parameters of the other party. So they do not know the payoffs, nor the strategy set of each other. The idea of Harsányi was to suppose that there can be many types of each players according to the above mentioned parameters, and both players know only their own concrete type. A probability distribution can be assigned to the possible types of both players. According to this, chance (as a third player) chooses among the possible types at the beginning of the game, but this move cannot be observed by the players. In this way, the game becomes manageable, the game can be played with the different types of the other party, and the payoffs can be summarized according to the corresponding probabilities. The incomplete information game will only be an imperfect information game, this conversion is called Harsányi-transformation.

This method can be applied naturally not only for military and political problems, but also in economic competitions. The firms know, in general, only their own position and possibilities, but not the others; but they can assign probabilities to the possible types of the other firms. Thus, the method of Harsányi can be applied for the developing telecommunication competition.

Summary

The several described game theoretical methods -despite of their limits- illustrate that the problems of the oligopolistic competition in telecommunications, after a simplification, can be modelled, examined and solved by the different tools of non-cooperative game theory.

References

[1] R. Konkoly, I. Fekete, A. Gyürke: Evaluation of uncertainties in investment projects, Third European Workshop on Techno-economics for Multimedia Networks and Services, Aveiro, Portugal, 1999.

[2] D. Kreps: Game theory and economic modelling Oxford University Press, 1990.

[3] E. Rasmusen: Games and information, Blackwell Publishers, 1989.

[4] Y. Sato, E. Akiyama, J. D. Farmer: Chaos in learning a simple two person game, Proc. Natl. Acad. Sci. USA, Vol. 99, Issue 7, 4748-4751, April, 2002.

News

iPASS – a company offering software for wireless local area network (W-LAN) services – has announced that it now provides software for 1,000 active W-LAN hotspots for 11 different service providers in ten countries. Concentrating on business traveilers, the company has enabled hotspots at 16 major airports in places including Copenhagen, Madrid, Tokyo, Singapore and New York. Ipass has alos set up hotspots in more than 500 hotels and hundreds of coffee shops, resaurants and Internet cafes.

•

Nortel in order to improve handling of traffic on its network. Contactel has installed Nortel s Succession Communication Server 2000 softswitch and its Passport Packet Voice Gateways to deliver voice and broadband services over an IP network infrastructure. After successfully concluding interconnection tests it was decided to run a softswitch in the PSTN.

•

The **SALT** Forum – a group of companies with the shared goal of accelerating the use of speech technologies in telephony systems – has revised its membership structure and organising priciples in response to interest from existing and potential members. Best known as the originator of Speech Application Language Tags (SALT), the Forum s activities were guided by a six-member board of directors. The new organisational structure widens industry representation by introducing a „sponsor" membership class with full voting rights. „Contributor" members will also be granted voting rights within working groupsadding formality to the established practices of the Forum. The SALT Forum s activities will be funded by nominal membership fees paid by sponsors and contributors, which the Forum intends to waive for small companies, individuals and academic institutions.

Link Adaptation in MIMO Systems

ZOLTÁN NÉMETH, SÁNDOR IMRE, FERENC BALÁZS

BME Híradástechnikai Tanszék
imre@hit.bme.hu

Level of requirements for transmission rate in fixed and mobile communication systems is heavily increasing due to the various types of wireless services (data, voice, multimedia, etc.) However, in case of voice and multimedia services efficient – and lossy – compressing techniques are used for reducing the bandwidth, in case of data services these are limited or not available.

1. Introduction

In wireless broadband communication systems widening of the bandwidth is not possible, thus the main purpose is the enhancement of the spectral efficiency, in other words the bitrate per unit bandwidth (b/s/Hz). Among the transmission techniques based on this principle we can mention the various adaptive antenna arrays, the multiple input multiple output (MIMO) systems, the adaptive modulation and coding techniques, level dependent techniques, respectively. The adaptive modulation and coding techniques, also included in the newest standards like the Enhanced Data GSM Evolution (EDGE), are able to adapt continuously to the time-variant radio channel, thus application of them is very promising in the data-centric wireless networks of the future. Among the spectral efficiency heightening algorithms, the adaptive modulation and coding procedures are also called link adaptation (LA) methods.

In the field of link adaptation significant development has achieved in latest time, and according to it the application of these results is widespread in practice. The new solutions seem to be efficient in the spatial, time, frequency, etc. division based multiple transmission antenna systems (e.g. MIMO), and in various multi-carrier techniques – among which the most important one is the orthogonal frequency division multiplexing (OFDM). The main purpose, with application of mentioned principles, is creation of such robust, cost-efficient and as far as possible small complexity systems that meet the requirements of the wireless communication networks of the future.

Firstly in Section 2 of our work we will describe the radio channel and the parameters affecting the quality of the transmission, respectively. The continuous measurement and monitoring of channel parameters is required for link adaptation and reaching the highest efficiency. Problems according to this work are examined in detail in Section 3. The aim of adaptation is heightening the transmission rate, which is attainable in several ways. Some of the possibilities are introduced in Section 4. After recognizing the conditions and having adequate tools (transmission rate improving methods), optimization of the system can be performed by use of adaptation algorithms, as it is shown in Section 5. Finally this article is summarized in Section 6.

2. Channel Features

2.1. Channel Parameters

An ideal link adaptation algorithm, while setting the transmission parameters, makes allowance for every channel parameters being significant in respect of transmission. Contrary to wired line networks in wireless ones the channel features vary randomly and the statistical models describing these features are strongly depend on the given environments. In case of MIMO systems it is practical to apply such channel model in that there are no direct line connection between the transmitter and the receiver (non Line of Sight – nLOS). Among factors affecting the transmission we should mention the channel dispersion, Ricean K-factor, Doppler-effect, cross-polarization attenuation, antenna correlation and the so called condition number [2].

A, Dispersion

The dispersion is an important channel feature, which arises from reflection due to near and far objects, and significantly affects the transmission. The dispersion is usually defined as the mean square of delays of propagation paths. This value is directly proportional to distance and depends on beamwidth and altitude of the antenna, and environments, respectively. Typical value of dispersion is in the interval of 0.1-5 μs as a rule.

B, K-factor

The amplitude of a fading-burdened signal has Ricean distribution, which can be characterized with two parameters: with the constant part of the signal power, P_C, and scattered power, P_S. Quotient of these two values (P_C/P_S) is called Ricean K-factor. The worst case, due to the fading, occurs when there are no constant component (P_C=0).

In this case K equals zero, and the signal becomes Rayleigh distributed. The K-factor is very important design parameter, because it is related to the fading occurrence. For the sake of reliable operation the most critical conditions should be considered in design of both the mobile and fixed communication systems.

C, Doppler Effect

The Doppler spectrums of fixed and mobile channels are different. In case of fixed networks the Doppler shift takes values from the interval of 0.1-2 Hz, and its spectrum is exponential. In mobile applications the value of Doppler shift is in order of 100 Hz having Jake spectrum, which is shown in *Figure 1*. The *S(f)* stands for power density, f_c is the carrier frequency and f_m means the Doppler frequency [5].

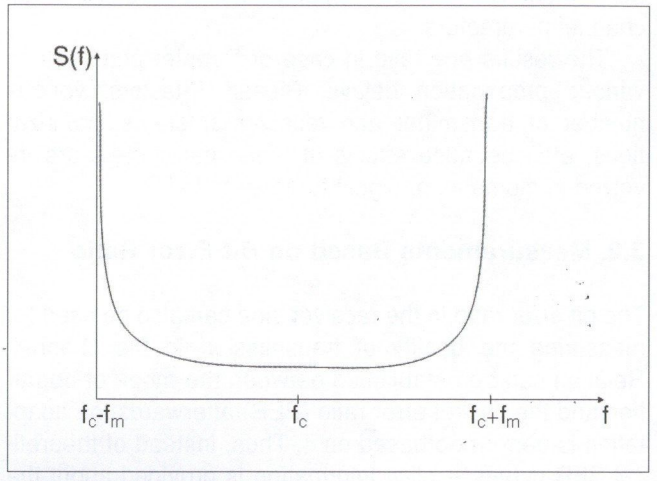


Figure 1 Jake spectrum

D, Cross polarization

Also an important parameter is the cross-polarization discrimination (XPD), which is defined as the quotient of the correctly and incorrectly polarized power. The XPD also means the separation between such two transmission channels that have different orientation of polarization. The larger value has XPD, the smaller part of radiated energy is coupled to other channels having different polarization. Value of XPD is lower in case of higher distance.

E, Antenna Correlation

The correlation between the antennas has significant affection to SIMO, MISO and also MIMO systems. In case the value of the complex correlation coefficient is high (exceeds 0.7), the diversity features become inadequate, furthermore if correlation equals 1, the advances of diversity totally disappears. In practice the correlation coefficient is usually small, takes values from the interval of 0.1-0.5, assumed that the configuration of transmitter and receiver devices are adequately chosen.

F, Condition Number

The condition number of a multiple input multiple output system is defined as the quotient of the largest and the smallest singular values of the channel matrix. In case of spatial multiplexing MIMO system the high capacity can be performed, if the ordinarily continuously varying condition number is low. The value of the condition number is often high in a radio connection having also a direct propagation path (LoS). This problem can be eliminated by use of antenna with dual polarization.

All of the described features affect the state of the channel, and therefore the quality of the transmission. The propagation models depending on the channel parameters are usually classified as what alterations they describe. According to this property two groups of models are distinguished: models of small-scale and large-scale variations. Fluctuation of fieldstrength within small distance or during small time interval due to the multipath propagation is an appropriate example for small-scale variations. In case of broadband signals these rapid variations cause selective fading. The large-scale variations, including path loss and fluctuation of it around its mean value, are modeled as lognormal distributed random variables. These types of variations cause shadow fading.

2.2. Fading Effects

Due to the multipath propagation the radio waves interfere, superpose thus amplitude peaks and attenuation maximums (amplitude minimums, fading) evolve. The appearing time-selecting fading can be characterized by coherence time, which means a time interval within the channel impulse responses are strongly correlated [1, 2]. This feature is inversely related to Doppler propagation, because it measures how slowly varies the channel. The slower the channel change, the larger the coherence time. In case one would like to create a link adaptation method also adjusting to small-scale variations, the channel should be examined with a frequency according to coherence time at least.

Appearance of frequency-selective fading is also an outgrowth of the multipath propagation. The transmitted signals arrive at the receiver via different long paths, thus time difference evolves. Fading appears if this time difference is commensurable with the symbol time interval. This condition is usually fulfilled in case of broadband transmissions.

In case of multiple antenna systems spatial fading is also observable. In such architectures the amplitude of received signal depends on the spatial arrangement of the antennas. From this point of view both the direction of departure (DoD) and direction of arrival (DoA) are important.

3. Measurements of Channel State

3.1. Measurements Based of Signal-to-Noise Ratio

For appropriate link adaptation two subtasks should be solved. In the first step definition and measurement of an index according to the quality of the channel, the so called channel state information (CSI), is required [1]. Afterwards, transmission parameters should be tuned as a function of the state of the channel. There are several measures of channel quality, among which the most important ones are the signal-to-noise ratio and signal-to-interference ratio. Value of these parameters can be measured in the phy-

sical layer (e.g. measurement power level during signal transmission and no communication time interval). Further important measures are the bit error ratio (BER) and packet error ratio (PER), which are provided by the data link layer. Certainly, in case of given conditions the different channel state information also have advances and drawbacks, which should be considered before choosing the most appropriate one. Thereinafter a few possible solutions will be shown.

Setting of the transmitting parameters can be fitted to the average value of the signal-to-noise ratio. This information should be available either the receiver or the transmitter side, however, it is usually measured at the receiver device. Having the average signal-to-noise ratio the bit error ratio should be assigned in the next step. Afterwards according to the signal-to-noise ratio one can choose the optimal mode of operation (e.g. the modulation type) so that, the transmission rate would be the possible maximal, while exceeding the given bit error ratio is not allowed. Finally the transmitter device should be directed changing the mode of operation. Suppose that only time-selective fading is present in the system. The SNR-BER conversion is possible if the averaging of the signal-to-noise ratio is performed within a very small time interval, thus any of time-windows can be treated as a constant fading-free channel.

As long as the instantaneous value of the signal-to-noise ratio is available, according to this we can select the adequate mode of operation. Assuming additive white Gaussian noise and coherent detection the bit error ratio can be expressed. The ideal operation is only theoretic, of course, because the effective refresh time may be longer than the coherence time due to the feedback delay and other derogatory circumstances in practice. In this case calculation of bit error ratio is not possible using the simple AWGN channel model. For the solution of the problem second and higher order statistics of the signal-to-noise ratio is also required [1].

Suppose that the channel state information is measure in a time window, size of which is determined by the link adaptation protocol. In case of multi-carrier modulations, two dimensional time-frequency window is needed. The SNR-BER assigning can be performed by use of probability density function, which is valid in the given time interval. In practice this function cannot be determined with simple analysis, because it is a function of many parameters. Among these parameters one can mention the statistics of the fading regarding to time and frequency, relation between length of the time window and time interval, correlation between size of frequency window and coherence bandwidth, and in case of multi-antenna system the number and polarization of transmitter and receiver antennas [2].

The problem is simplified if k-th order moments of the signal-to-noise ratio are examined in spite time function of it. However, the moments are only approximation of the SNR, they may give enough information for determining the adequate relation between SNR and BER still in case of low value of k. The first order moment (the mean) of the signal-to noise ratio is related to the average power on the receiver side. The second order moment yields information about the time and frequency selectivity within the adaptation window. The higher-order statistics provide further knowledge about the probability density function, however in this case the computational complexity is proportional to the increasing value of k. Determination of adaptation thresholds based on the statistics of signal-to-noise ratio provides a simple and flexible method, because thus the thresholds will be independent from miscellaneous channel parameters.

The results are valid in case of Doppler propagation, various propagation delays, Ricean K-factors, various number of transmitter and receiver antennas, polarizations, etc., because effects of these parameters are involved in moments of signal-to-noise ratio [1].

3.2. Measurements Based on Bit Error Ratio

The bit error ratio in the receiver side can also be used for measuring the quality of transmission in the channel. Relation can be established between the mode of operation and the packet error ratio (PER) afterwards the adaptation is performed based on it. Thus, instead of theoretical BER curves explicit information is provided about the quality of the radio link [1]. Unfortunately our possibilities are limited, because number of samples is also limited within any time window.

The method is based on estimation of the PER, for which some thousands of samples are required to reach adequate reliability. Therefore the rate of adaptation is degraded. If training sequences are not transmitted periodically, then only the large-scale variations can be considered during the adaptation. Further drawback of this method is the strong traffic dependence, which has two consequences. At first controlling the reaction time of the algorithm becomes difficult, and on the other hand possibility of monitoring is lost when no traffic is generated in the channel. In this latter case reinitialization of the adaptation is required.

3.3 Measurements Based on SNR and BER

Parameters describing the channel properties (SNR, BER), as shown, have many advances and drawbacks. Methods based on measuring the signal-to-noise ratio and its statistics provide flexible adaptation. However, these methods depend on the determination of adaptation threshold, consequently may be imprecise. Accuracy of determining thresholds may be improved using higher-order statistics of signal-to-noise ratio. Precision of BER measuring methods is usually satisfactory, but certain amount of traffic should be observed to achieve it. This is a problem especially in low error rate ranges, where accordingly significant additional transmission arises and the adaptation becomes slow. One of the efforts of present research is achieving robustness and accuracy together in case of various channels, adaptation rates and transmission conditions.

4. Improvement of Transmission Capacity

4.1. Diversity Techniques

In this section such a wireless mobile system will be examined in which the transmitter equipment (in this case the base station) contains two antennas. Eventually, by this choice we determine the measure of capacity improvement: our aim is to double the transmission rate. In mobile equipments it is practical to restrict the number of antennas to $N=2$, because higher value of N is at the expense of mobility due to the enlargement in device size.

As mentioned in the introduction, the spectral efficiency can be significantly improved using multi-level modulations. Unfortunately, this solution is very sensitive to interference, therefore it is worth associating with other procedures. In case having multiple antenna devices in both the transmitter and receiver devices also, the spectral efficiency can be enhanced by other methods. In a system including transmitter and receiver devices with N antennas, the improvement of spectral efficiency is in direct proportion to N [3]. This correspondence is valid theoretically, in practice this improvement is usually lower due to the occurrent low signal-to-noise ratio and the limited complexity of the receiver device. While creating multiple input multiple output system our purpose is to approach the theoretical value, which can be performed using various methods.

Among transmission rate improving methods of MIMO systems one can mention the layer-organized multi antenna transmissions and diversity techniques [1, 3]. The first procedure transmits independent data streams which are assigned to each antenna elements. In case of this solution the data rate increases, however, detection of the signals interfering in the radio channel (with the whole complexity of the problem) remains to solve in the receiver device. In diversity solutions space-time block-coding is applied for achieving full diversity. The data rate can be heightened by puncturing the channel code. This method reduces the amount of additional information included in the error correction code and increase the effective bitrate. The difference between the two procedures that while the first method directly creates parallel independent transmissions between antenna pairs, in spite of the second one applies diversity techniques and is able to recover information losing in the channel due to the fading, inadequate channel code or modulation.

Suppose that our purpose is to increase the bitrate of a single input single output (SISO) system to its twice using a multiple input multiple output system. Since layered MIMO systems transmit independent data streams from each antenna, therefore two transmitter antennas are required at least for doubling the bitrate. Certainly, in this case in the receiver device two antennas are also required at least, otherwise the radiated source signal will not be fully recovered. Thus a MIMO architecture is evolved, which is characterized by a 2x2 size channel matrix and includes two transmitter and receiver antennas that assumed to be independent.

A, Layered Scheme

Firstly let us consider the construction of the layered scheme in the transmitter side. Using a serial-to-parallel converter the incident data is divided into two streams, which further on feed the antenna elements. Afterwards, suppose that the bit streams are coded in a convolutional encoder with $R=1/3$ coding rate. Thus encoding the streams separately better performance can be achieved, however in this case more complex receiver device is required. To avoid burst errors it is worth interleaving the bit streams before package formatting and modulation. After the interleaver a so called antenna switching block is inserted, which has to forward the consecutive symbols alternately towards one and another antenna elements. The arrangement is shown in *Figure 2*, where the gray-colored block, the antenna switching block is not indispensable, however, in absence of it the receiver device becomes less efficient [3]. In both cases doubled transmission rate can be achieved compared to the SISO system.

B, Diversity solution

According to the other approach the maximization of diversity order is required, therefore the structure is significantly differs from the former one as one can see on *Figure 3*. At first, the serial-to-parallel converter is unnecessary, and on the other hand to double the bitrate after the convolutional encoding puncturing with $R=2/3$ rate is inserted. Naturally, one can choose other solution and use $R=2/3$ rate convolutional encoder directly, however, the puncturing technique is more practical, because the receiver device becomes simpler to implement due to the fewer code types [3]. One of the most important parts of the transmitter is the 2x2 size space-time encoder that provides the diversity (space-time transmit diversity – STTD).

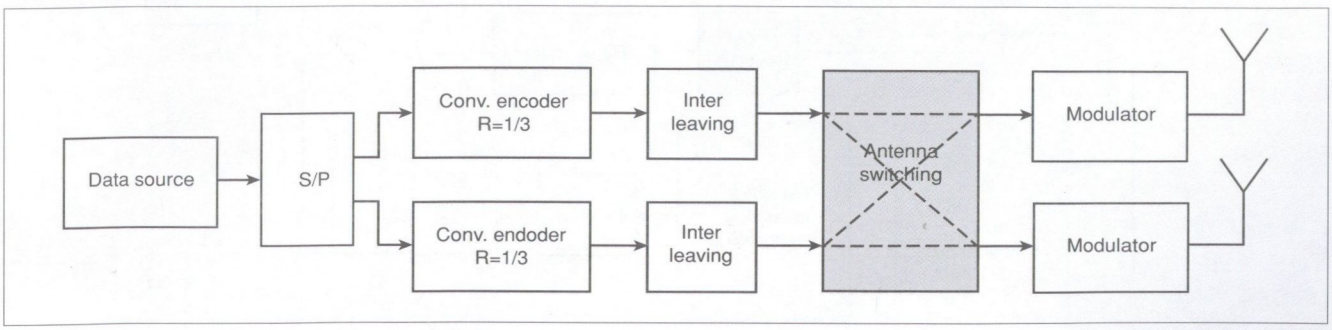


Figure 2 Transmitter device, no diversity

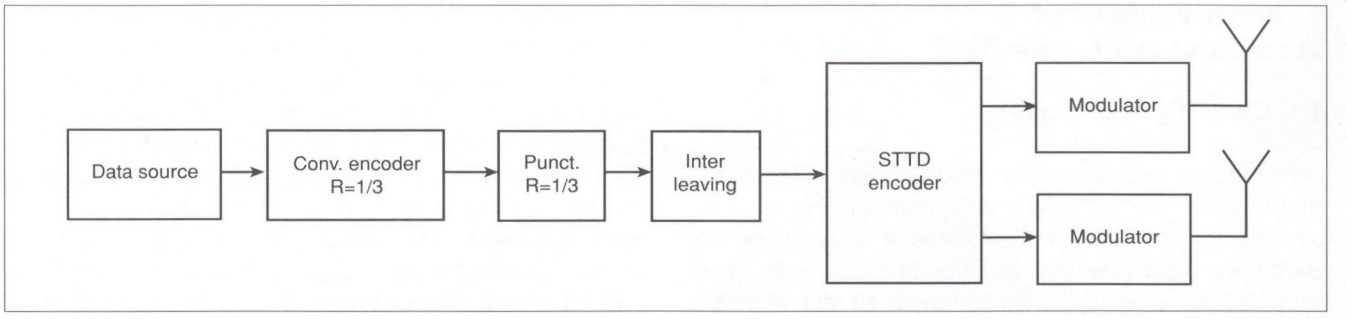


Figure 3 Transmitter device, diversity solution

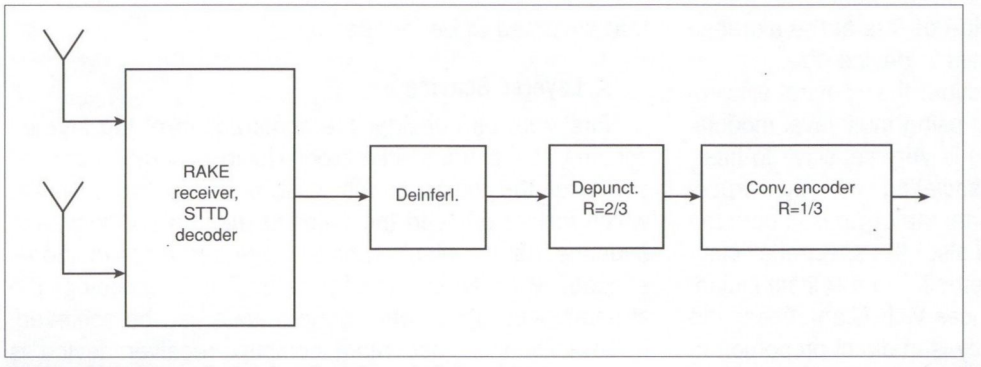


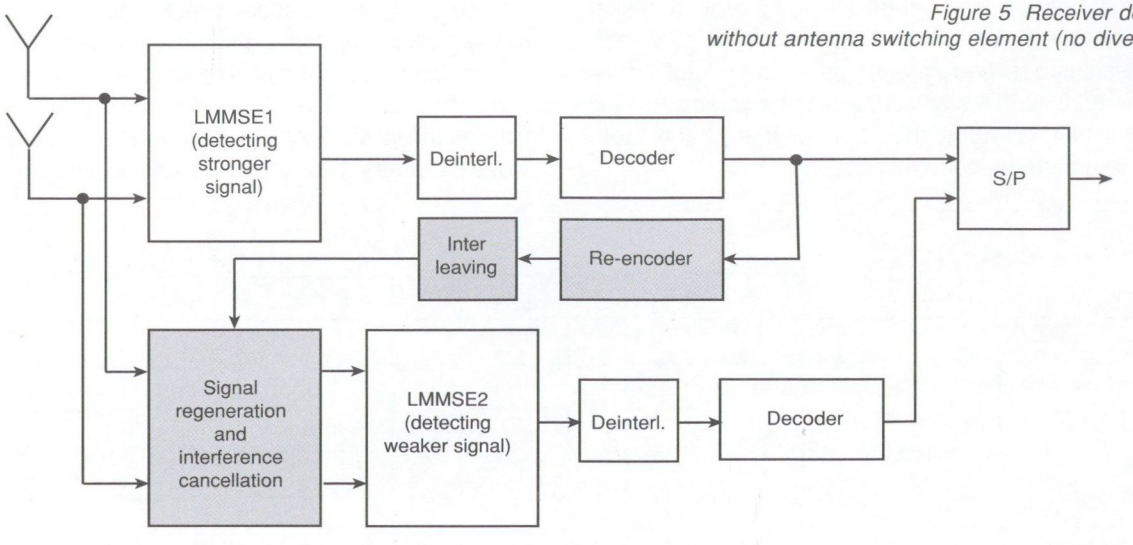
Figure 4 Receiver device for diversity solution

To different transmitter structures belong various receiver architectures. The transmitter operating with puncturing has a receiver pair, which is shown in *Figure 4*. The RAKE receiver also performs the STTD decoding. This device is followed by the deinterleaver and depuncturing units. During depuncturing the bits that have been not transmitted are substituted with zeros, therefore the zero-valued inputs represent maximal uncertainty for the decisions of channel decoder.

In case of layered transmitters device more complex structured receiver is required. Since two interfering signals have to be detected, therefore it is worth detecting first the signal stream that arrives at the receiver in more favorable conditions. This approach corresponds to principle of successive interference canceling methods. In case the antenna switching block is not used in the transmitter device, the receiver algorithm is the following: the incident stronger signal is detected with LMMSE (Linear Minimum Mean Square Error) detector then this one is subtracted from the incident signal. In case error has not occurred the second weaker signal can also be detected error-free. Since the original data streams have independently encoded, hence before second detection re-encoding and re-interlacing of the signal (which has already decoded) is required for adequate interference cancellation. Thus a very reliable interference canceling method is achieved. In *Figure 5* the gray-colored blocks are the additional units required for appropriate interference cancellation.

In case we use an antenna switching unit, two equivalent layers are formed, therefore there is no point in detection of any of the two signals before the other one. As one can see on *Figure 6,* two different LMMSE detectors are used parallely. After decoding, re-encoding and interference cancellation the produced signal is led again into the LMMSE detector of which outputs provides the finally decodable decisions. For this purpose the complexity of the receiver device should be enlarged, however the reliability also increases.

Figure 5 Receiver device without antenna switching element (no diversity)



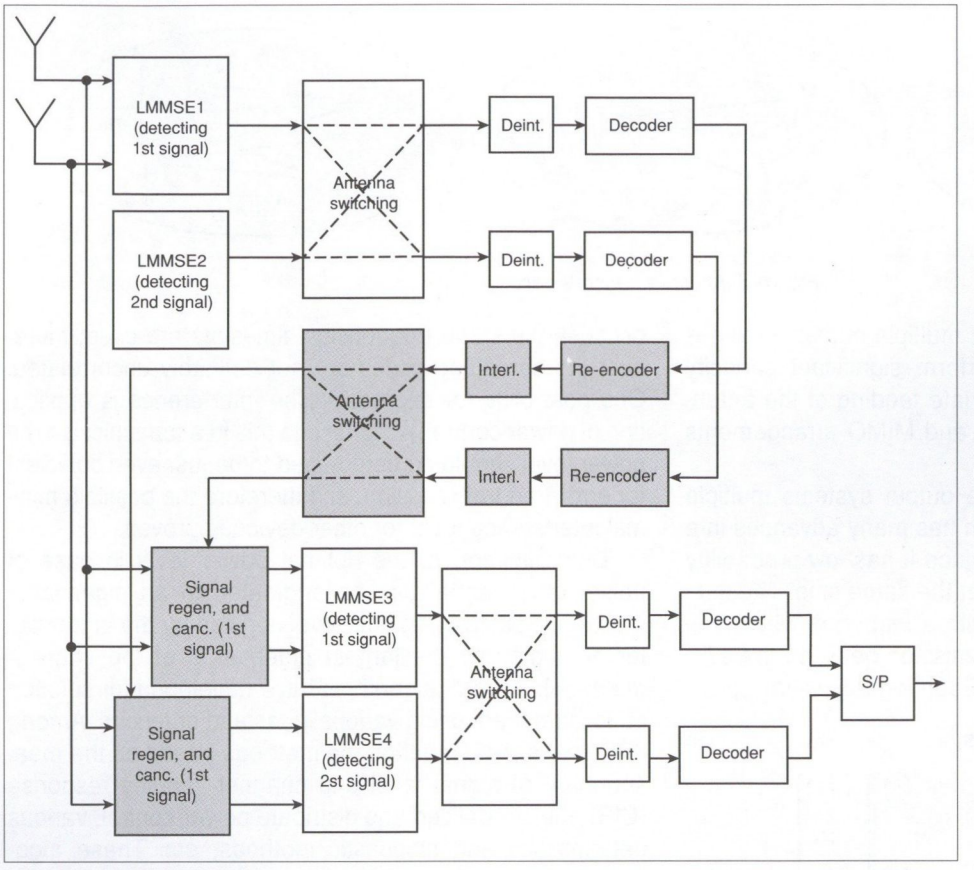


Figure 6 Receiver device with antenna switching element (no diversity)

achieving the higher data rate one has to decide between application of the two methods.

The orthogonal frequency division multiplexing modulation based on the mentioned principles can be implemented efficiently with FFT (Fast Fourier Transformation) algorithms both in the transmitter and receiver side. Each emergent frequency components, the tones can be corresponded to sub-channel in a MIMO system and the state of the channel can be characterized by a matrix. In this case Equation 1 is valid:

$$\mathbf{x} = \mathbf{A}_{OFDM} \cdot \mathbf{s}\,, \qquad (1)$$

where s is the vector of transmitted signals, x means the vector of incident signals in the receiver device and A_{OFDM} stands for the matrix of cross-overs between the sub-channels.

4.2. Multi-carrier systems

Similarly to the intelligent antennas the multi-carrier system can also be applied for improvement of transmission capacity. The basic assumptions and fundamental principles are the followings. In broadband systems frequency selective fading arise due to the multipath propagation. To eliminate this effect it is practical to apply any kind of spread spectrum techniques. Beside frequency hopping and direct sequential spectrum spreading one of the most popular solutions is the orthogonal frequency multiplexing (OFDM).

The OFDM methods divide the original broadband signal into narrow bands and assign them to sub-carriers (tones). Thus the robustness of the system against multipath propagation is improved. In ideal case signal streams of each tone are mutually independent to achieve maximal data rate. However, any of the carriers become unusable due to the fadings, thus information carried in this tone will lost. To avoid this effect dynamic sub-channel allocation should be applied, and the carriers ruined by the fading should not be modulated and data stream should not be divide for them. In this case the instantaneous state of the channel should be known in the transmitter device, for which additional information and computational capacity is required. In spite of this it is more practical to work with adequate redundant coding and interleaving between the sub-channels, thus frequency diversity is provided [1, 2]. However, in this case due to the additional data, the spectral efficiency is reduced, thus for

4.3. Function of MAC sub-layer

In MIMO-OFDM systems, also in spite of all methods straining after robustness, transmission errors occur. The arising errors should be treated with an appropriately efficient medium access control (MAC) layer in a reliable wireless network. For the proper operation an automatic repeat request (ARQ) method is required, in which the transmitter divide the bit stream into adequate sized packages. In case erroneous package appears in the receive device, the transmission should be repeated. The ARQ mechanism can be regarded as such a method that provides time diversity eliminating the effect of noise, interference and fading, respectively [2].

5. Adaptive methods

5.1. Intelligent antenna arrays

Application of adaptive antenna arrays, the so call intelligent antennas, is very promising for enhancement of spectral efficiency in wireless systems, thus these antennas provide good solutions for link adaptation. Intelligent antennas usually mean applying antenna array in one side of the communication. The multiple antennas can be installed into the transmitter device, these are the multiple input single output (MISO) systems, or into the receiver side, these are the single input multiple output (SIMO) solutions. In case we use multiple antennas at both com-

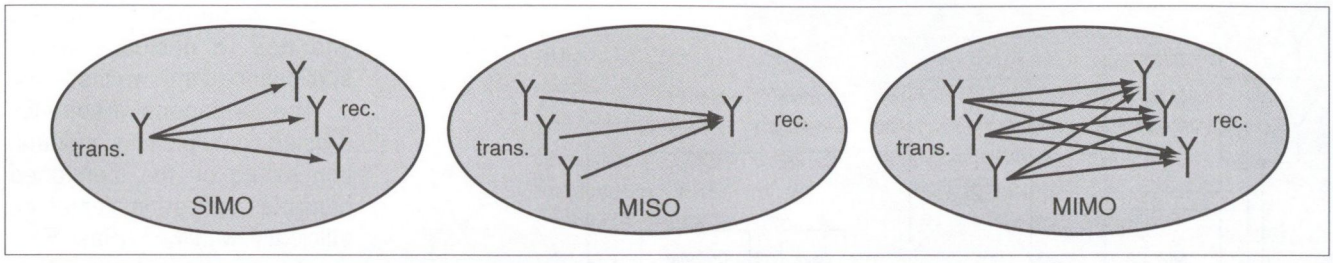


Figure 7 Multiple transmissions

munications ends, multiple input multiple output system is formed, which are able to perform significant capacity improvement assuming appropriate feeding of the antenna elements. The MISO, SIMO and MIMO arrangements are shown in *Figure 7.*

Using multiple input multiple output systems multiple spatial channel is created, which has many advances in a fading burdened environment, since it has low probability that every channel is unusable at the same time. The multiple channel is characterized with a matrix, of which elements correspond to each transmission between the sub-channels [3, 4]. For this model Equation 2 is valid:

$$\mathbf{x} = \mathbf{Hs}, \tag{2}$$

$$\begin{bmatrix} x_1(t) \\ \vdots \\ x_m(t) \end{bmatrix} = \begin{bmatrix} h_{11}(t) & \cdots & h_{1n}(t) \\ \vdots & \ddots & \vdots \\ h_{m1}(t) & \cdots & h_{mn}(t) \end{bmatrix} \begin{bmatrix} s_1(t) \\ \vdots \\ s_n(t) \end{bmatrix}$$

where **H** means the channel matrix, **s** and **x** are the vectors of source and received signals. The best selection for feedings and the signals of transmitter antennas can be determined, considering to elements of the matrix. With any different choices of input bit streams and transmitted signals depending on them, we have different purposes. Firstly, we are intent achieving maximal transmission rate for which independent feeding of each antenna elements is required. This technique is the simple spatial multiplexing, which is efficient if the sub-channels are mutually independent or, in case of weaker restriction, have minimal correlation. Nevertheless if the requirement is not fulfilled, the transmission rate is heavily degrades due to the frequent errors.

Elimination of this problem, that already means the realization of the other main purpose, can be solved with redundancy (spatial, time or frequency overlap in the transmission, redundant coding) installed into the transmission. Thus we attain to so called space-time coding procedures of which main aim is improving spatial diversity and therefore achieving minimal bit error ratio. The created system becomes robust, however, the transmission rate degrades. Due to this method the fading margin can be kept down with 10-20 dB also in case of low number of elements, therefore fading reserve is ensured [1, 3].

5.2. Power Control

Interference burdening of MIMO systems is considerably heavy due to the multiple antennas. This assertion is especially true in case we apply code division multiplexing (CDM), because the same frequencies, timeslots are used, moreover the spreading codes are not definitely uncorrelated. One possibility for decreasing the interference is application of power control, According to this in a transmission the power level should be determined to be just even sufficient for error-free transmission, and therefore the possible minimal interference level for other device is proven.

Determination of the optimal power level in case of noise free channel can be originated in an eigenvalue problem. Setting of optimal level is given by the eigenvector according to the largest eigenvalue of the channel matrix [4]. Several algorithms have developed for solution of power control using various ideas and criterions. Among these we should mention the methods based on the measurement of signal level and channel impulse response (CIR), the centralized and distribute power control, various deterministic and stochastic methods, etc. These algorithms have different advances, for example fast convergence, minimal information demand, but the solution of the mentioned maximization problem, depending only on the number of users and channel characteristic, gives the upper bound for all of them.

By the ensemble application of the power control and the adaptive antenna systems (e.g. MIMO) the capacity of the wireless networks can be enhanced significantly, moreover the convergence of the controlling algorithm becomes faster [4]. For achieving optimal solution the power control should be performed in both communication ends. The procedure is especially useful in cellular mobile systems, where the power level of the base stations should be kept low to decrease interference level for other communicating devices, meanwhile in mobile equipments the limited energy resource is the reason for demand of low power consumption.

5.3. The Link Adaptation

By fundamental assumption of link adaptation the system monitors the state of the radio channel, and according to this sets the transmission parameters, like type of modulation, coding procedure, signaling bandwidth, power level, etc. The purpose of parameter adaptation is achieving the highest spectral efficiency. For this purpose the transmission features should be dynamically adapted to the interference arising between communication devices and other disturbing effects.

The variable describing the state of the channel, for example the signal-to-noise ratio, is considered as a quantified measure; accordingly the channel state is regarded

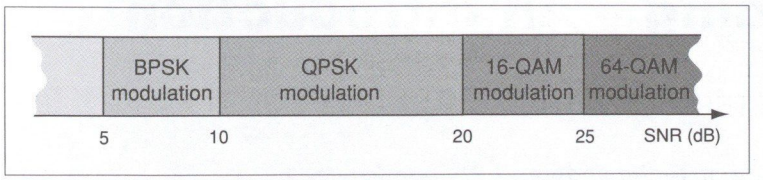


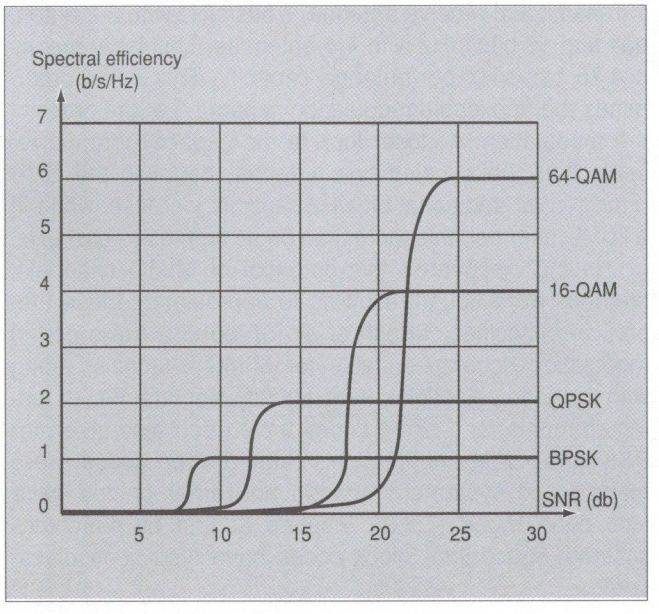
Figure 8 A possible modulation assignment

constant between two assigned values. Each parameter set or mode of operation is assigned to discrete channel states [8]. A possible solution is shown in *Figure 8*, where the different modulation schemes are assigned to different signal-to-noise intervals.

The reason for this assignment is the different noise sensitivity and spectral efficiency of each modulation. The BPSK (Binary Phase Shift Keying) modulation can be used in case of low signal-to-noise ratio, however, its spectral efficiency is small (1 b/s/Hz), meanwhile the 64-QAM spectral efficiency is 6 b/s/Hz, but it can be only applied in case of high signal-to-noise ratio regarding appropriate constraint for bit error ratio. The algorithm of the link adaptation should provide the most efficient transmission. This involves achieving the highest spectral efficiency besides robust operation in case of unfavorable conditions. A system that can select BPSK, QPSK, 16-QAM and 64-QAM modulations as a function of signal-to-noise ratio for performing the transmission, should decide as it is shown in *Figure 9*. The accentuated curve shows the correct decision and the spectral efficiency for various signal-to-noise values.

In absence of adaptation algorithm the design parameters should be chosen to provide the adequate operation in case of unfavorable conditions, like low value of signal-to-noise ratio (worst-case design). If fixed BPSK modulation applied then the robust operation is provide, however, in case of better condition the channel capacity remains unexploited.

Figure 9 Spectral efficiency as a function of modulation and SNR



6. Summary

Due to the wireless services that have increasing demand for transmission rate, the spread of multiple input multiple output systems is expectable in the near future. The main aim of MIMO networks is enhancement of spectral efficiency and transmission rate, respectively. For efficient operation appropriate link adaptation is required, which has several possibilities to be realized. Before performing various solutions, at first, several restrictive features should be examined that arise in the radio channel. Examination all of these parameters would require enormous computational complexity, therefore the so called channel state information should be considered that are functions of the mentioned parameters, but processing them is not so problematic. This requirement is fulfilled in case of signal-to-noise ratio or bit error ratio, for instance. Due to the finite adaptation rate only statistics can be managed instead of instantaneous values. Measurement of these statistics claims for appropriate foresight, and that should be considered that accuracy and celerity of the measurement and flexibility can be improved at expense of one another. Having the information about the state of the channel, using the adequate link adaptation algorithm the spectral efficiency and transmission rate of the MIMO system can be optimized. During adaptation various spectral efficiency enhancing equipments are available. Selection and ensemble use of these possibilities is decided by the adaptation algorithm. Among the equipments one should mention the multi-level modulations, the space-time coding and the multi-carrier techniques (e.g. OFDM). The assignment of channel state and the transmission techniques together with the measurement of the channel state is considerably difficult task, and examination of this problem is an active field of research.

References

[1] S. Catreaux, V. Erceg, D. Gesbert, R. W. Heart Jr.: „Adaptive Modulation and MIMO coding for Broadband Wireless Data Networks", IEEE Communications Magazine, June 2002, Vol. 40, No. 6, pp.108-115

[2] H. Sampath, S. Talwar, J. Tellado, V. Erceg, A. Paulraj: „A Fourth-Generation MIMO-OFDM Broadband Wireless System: Design, Performance, and Field Trial Results", IEEE Communications Magazine, September 2002, Vol. 42, No. 9, pp.143-149

[3] M. J. Heikkilä, K. Majonen: „Comparison of Layered and Diversity Approaches for Increasing WCDMA Data Rates in Frequency-Selective MIMO Channels", IEEE 7th Int. Symp. on Spread-Spectrum Tech&Appl., Prague, Czech Republic, Sept. 2-5, 2002, Vol. 2, pp.333-337

[4] M. Elmusrati, H. Koivo: „Performance Analysis of DS-CDMA Mobile Communication Systems with MIMO Antenna System and Power Control", IEEE 7th Int. Symp. on Spread-Spectrum Tech. & Appl., Prague, Czech Republic, Sept. 2-5, 2002, Vol. 2, pp.541-544

[5] T.S. Rappaport: „Wireless Communications", 1996, New Jersey, Prentice Hall, pp.180-185

Space-Time Coding – An Introduction

PÉTER HORVÁTH

Budapest University of Technology and Economics
Department of Broadband Infocommunication Systems
hp@mht.bme.hu

Reviewed

In addition to the well-known receive diversity techniques, diversity schemes applying several transmit and receive antennas have recently been introduced. These schemes handle the radio channel as a Multiple-Input, Multiple-Output (MIMO) system. Space diversity results from applying several antennas while time "diversity" is also introduced due to the coding used. Hence the designation space-time coding is generally used for such methods that can result in a considerable overall gain. Accordingly, several recent wireless standards include support for space-time coding. This paper is an introduction to space-time coding. In addition, some applications are reviewed.

Introduction

Fading effects limit the ultimate data rate in wireless communication systems. It is known that in a Rayleigh fading channel, the bit error ratio is inversely proportional to the signal-to-noise ratio while in nonfading channels, this decrease is exponential. Conventional countermeasures against fading effects include error-correction coding, interleaving and receive diversity. These methods utilize the temporal and spatial decorrelation of the received signal. Interleaving is effective only if the interleaving affects a larger span than the duration of a fading minimum. Therefore, in a very slowly faded environment, large interleaving depth is needed to be effective, which also increases the incurred delay. Receive diversity requires several receive antennas providing sufficiently uncorrelated signals. When a mobile terminal is surrounded by a relatively large number of nearby scatterers, a half-wavelength separation is often sufficient. At an elevated base station, however, a separation of many wavelengths is needed because the received signals are more correlated [3]. Due to limitations in the mobile terminal size, it is impossible to build in more than one antenna into the equipment.

This problem can be solved by using transmit diversity involving a base station transmitting over several antennas at the same frequency. A simple, uncoded transmit diversity scheme was invented by Alamouti [4], which can be regarded as the dual of the maximum ratio receive combining system. It is intended to be used with linear modulation schemes. This scheme, like space-time schemes, utilizes channel state estimation only at the receiver; and at the transmitter, no channel state information is required. Later, a suitable differential detection scheme was also published (differential means that no channel estimation is required at the receiver). With one receive and two transmit antennas, it results in a two times diversity gain, without having coding gain.

Recently, multi-transmit-multi-receive antenna diversity schemes have been published. Telatar [1], and independently Foschini and Gans [2] derived the obtainable information-theoretical capacity of Multiple-Input, Multiple Output (MIMO) channels. The capacity in an additive white Gaussian noise (AWGN) channel is given by

$$C = \log_2 \det\left(\mathbf{I} + \frac{\rho}{n_T}\mathbf{H}\cdot\mathbf{H}^{*T}\right)$$

where **H** is the channel transfer matrix describing channel gains between transmit and receive antennas, n_T and n_R are the numbers of transmit and receive antennas, respectively, **I** is the identity matrix, ρ is the average signal-to-noise ratio (SNR) at each receiver. The normalizing factor ρ / n_T is required to constrain the overall transmit power.

In a fading channel, capacity is a random variable, therefore capacity with a given probability of outage is specified;. this capacity is larger than the specified value in 99.99% of time. The following example can be found in [2]:

In a slow Rayleigh fading channel with an average SNR of 21 dB, capacity with 1% outage is 1 bit/s/Hz with 1 transmit and receive antenna, 7 bits/s/Hz with 2-2 antennas and 19 bits/s/Hz with 4-4 antennas. This latter means a 4.75 bits/s/Hz per-antenna capacity. This value determines the applicable modulation scheme, i.e. the order of the modulation. A closer look at the capacity formula reveals that, while using one antenna, doubling the SNR results in an additional bit/s/Hz capacity increase, while in a MIMO channel, this factor can be as high as n_T bits/s/Hz.

Having discovered this potential of MIMO channels, methods were sought allowing to approximate these theoretical capacities. Tarokh et al. [5] found a trellis coded modulation scheme suitable for MIMO channels, along with the code selection criteria for Rayleigh and Rician fading channels, and some codes have been also given for PSK and QAM modulation. Later, better codes were found using systematic search, and other criteria have also been found [7,8,9]. Inspired at first by Alamouti's scheme, space-time block codes have also been discovered.

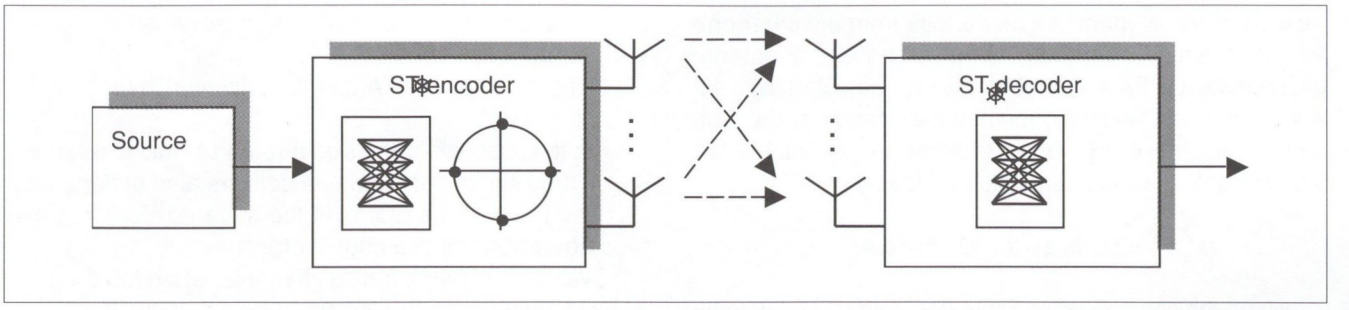


Figure 1 A space-time coded communication system

Most detection algorithms assume good channel estimates in the receiver. Differential detection schemes were invented to avoid the need for the channel state information [10]. These schemes assume slow variation of the channel, and encode the information into the difference between successive blocks, similarly to differential modulation schemes.

What follows is a short overview on space-time trellis codes with some applications.

Space-time trellis coding

A block diagram of a space-time coded communication system is shown in *Figure 1*. Its baseband model comprises n_T transmit and n_R receive antennas.

The information stream is encoded by a space-time encoder, which produces a stream for each transmit antenna, i.e. it produces n_T parallel streams, each of the same data rate as the original stream: $c_t^1, c_t^2, \ldots, c_t^{n_T}$ where t is the time parameter, $t = 1,\ldots,l$ where l is the frame length. They are subsequently mapped on a given constellation. The n_T transmit antennas transmit their own signals simultaneously. The receive antennas receive a noisy, fading-corrupted superposition of the transmit signals. The signal at receive antenna j is given by [5]

$$r_t^j = \sum_{i=1}^{n_T} h_{i,j}^t c_t^i + h_t^j \qquad (1)$$

where $h_{i,j}^t$ is the channel gain between transmit antenna i and receive antenna j at time t, and η_t^j is the additive white Gaussian noise. When the fading is slow (quasistatic), the channel gains can be assumed to be constant during one frame period, so the time dependence of the factors $h_{i,j}$ can be neglected. Then the complex matrix describing the channel is as follows:

$$\mathbf{H} = \begin{pmatrix} h_{1,1} & \cdots & \cdots & h_{n_T,1} \\ h_{2,1} & \cdots & \cdots & \cdots \\ \cdots & h_{i,j} & \cdots & \cdots \\ h_{1,n_R} & \cdots & \cdots & h_{n_T,n_R} \end{pmatrix}$$

At the receiver, the decoding can be realized by choosing the most probable transmitted sequence when $\boldsymbol{H}$ is known (channel state information is available).

We can assign the following expression to a hypothetical transmitted sequence q when sequence r was received:

$$\sum_{t=1}^{l} \sum_{j=1}^{n_R} \left| r_t^j - \sum_{i=1}^{n_T} h_{i,j}^t q_t^i \right|^2 \qquad (2)$$

The most probable sequence is the one that minimizes (2). In order to avoid the need for an exhaustive search, this sequence can be found by using the well-known Viterbi algorithm. If the trellis of the code employed is known, the branch metrics are given by the inner terms of (2). Channel estimates can be gained by inserting known pilot symbols into the data stream, and interpolating between these channel estimates.

It is interesting to derive an upper bound on the error probability. This was done in [5]. In slow Rayleigh fading channels, error occurs if a codeword $c_1^l, c_2^l, \ldots, c_{nT}^l, \ldots, c_1^l, \ldots, c_{nT}^l$ was transmitted but the receiver decides in favor of another codeword $e_1^l, e_2^l, \ldots, e_{nT}^l, \ldots, e_1^l, \ldots, e_{nT}^l$. In [5], a matrix $\mathbf{B}(\boldsymbol{c},\boldsymbol{e})$ is introduced as follows: its element ij equal the difference $e_j^i - c_j^i$ (difference of the two code words in position i, transmit antenna j), and let $\mathbf{A(c,e)} = \mathbf{B(c,e)B(c,e)}^H$. Thus, this matrix depends on the code properties. If rank of this matrix is r and its nonzero eigenvalues are λ_i, an upper bound can be given for the error probability:

$$\Pr(\mathrm{c} \to \mathrm{e}) \le \left(\prod_{i=1}^{r} \lambda_i \right)^{-n_R} \left(\frac{E_s}{4N_0} \right)^{-rxn_R} \qquad (3)$$

Analyzing this formula shows that a diversity gain of r x n_R is achieved (this determines the exponent of the decrease in the error probability when the SNR is increased, i.e. on a conventional logarithmic plot the slope of the error curve. Additionally, a coding gain of

$$G_C = \left(\prod_{i=1}^{r} \lambda_i \right)^{-1/r}$$

can be achieved because equation (3) can be written in the following form:

$$\Pr(\mathrm{c} \to \mathrm{e}) = \left(\frac{E_s / G_C}{4N_0} \right)^{-rxn_R}$$

This is an offset in the logarithmic plot.

Code construction

Proper choice of the codes is essential for an efficient operation of space-time codes. At first, the full trellis of the codes was given to describe a code, as in [5]. In [7], the generator matrix for QPSK codes was introduced, which provided a more compact and analytical description. These forms are equivalent, from the practical point of

view, as either of them can be derived from the other one. As an example, the outputs of a 4-state, 2-antenna encoder for QPSK modulation can be calculated as follows: if in time t the binary input to the encoder is the tuple (a_t,b_t), then the 4-ary output symbols (x^1_t,x^2_t) for the two transmit antennas are calculated as follows:

$$(x^1_t,x^2_t) = (a_t\ b_t\ a_{t-1}\ b_{t-1})\mathbf{G} \pmod 4$$

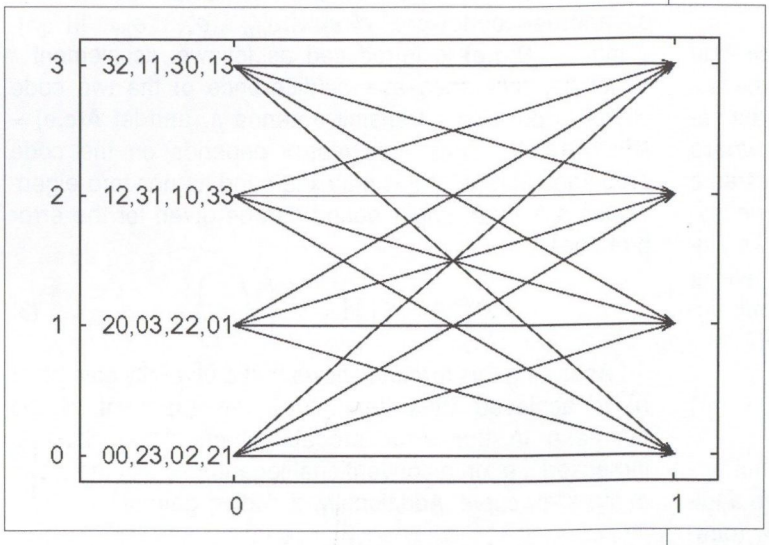
where the corresponding generator matrix [8] is given by:

$$G^T = \begin{pmatrix} 2 & 3 & 2 & 0 \\ 0 & 2 & 1 & 2 \end{pmatrix},$$

and the trellis diagram is depicted in *Figure 2*. Four branches are leaving each node, corresponding to the four possible values of the two input bits in order (11, 10, 01, 00) from the top to the bottom.

The number pairs at the left give the encoder output for antenna 1 and 2, respectively, in the former order of input combinations. In state 3 the value 32 means that for input combination (11) the outputs are 3 and 2 for antenna 1 and 2, respectively.

Figure 2 Trellis of a four-state, two-transmitter QPSK space-time code



As mentioned above, the first code selection criteria were published in [5]. Using the upper bound (3) and assuming a quasistatic Rayleigh fading channel, the error probability is minimum if the code words are selected according to the following conditions:

1. maximizing over all code words the rank r of the matrix **A(c,e)**, and
2. maximizing over all **A(c,e)** the minimum determinant $\left(\prod_{i=1}^{r}\lambda_i\right)$.

Later, it was shown that for a larger number of antennas [8], or over a so-called diagonal channel [11], performance is dominated by the main diagonal of **A(c,e)** and maximum rank codes can perform worse than such codes. For this case, the code construction should follow the following guidelines:

1. the rank of the matrix **A(c,e)** must be large enough ($r \times n_R \geq 4$)
2. maximize over all **A(c,e)** the minimum spur $\left(\sum_{i=1}^{r}\lambda_i\right)$.

For the case when the signal-to-noise ratio is relatively small, the second set of criteria delivers also better codes [9]. The trace of the matrix is the squared Euclidian distance between the two code words.

Over fast Rayleigh fading channels, when the diversity order is small, the following set of criteria is the best [8]:

1. maximize the minimum Hamming distance over all code words and
2. along the path(s) associated with the minimum Hamming distance, the distance $d_p^2 = \prod_{t\in\rho(\mathbf{c},\mathbf{t})} \|\mathbf{c}_t - \mathbf{e}_t\|^2$ (product distance) should be maximized. (ρ**(c,t)** is the set of those time instants t when $||\mathbf{c}_t - \mathbf{e}_t||^2 \neq 0$.)

When manifold diversity is employed, the Euclidian distance dominates the performance as in the slow fading case. Exact performance evaluation can be done by computer simulations. With two transmit and two receive antennas over Rayleigh slow fading channels, a 32-state code can achieve a frame error ratio (FER) of 10^{-2} if the average SNR is 10 dB.

Although space-time codes are thought for mobile use, their application on other areas can be also considered. As an example we mention the millimeter-wave fixed wireless access networks (LMDS or BFWA systems). Here, the biggest problem is the attenuation caused by the precipitation, mostly rainfall. Use of route diversity is the only effective countermeasure. Due to the small wavelength, high-gain, high-directivity antennas must be used at the terminal side. In this case, the channel matrix is dominated by the diagonal entries. Employing the space-time trellis coding in this environment (termed as route-time coding), FER plots for 2-2 antennas can be seen in *Figure 3* [11].

The uppermost curve represents the uncoded case, the other curves represent FER for different numbers of encoder states. Due to the properties of the channel, only twofold diversity can be achieved, FER decreases as the square of the signal-to-noise ratio.

Although this paper considered space-time trellis codes in detail, space-time block codes are also of great importance. Among others, differential schemes are based largely on block codes. In the first publications authors focused on schemes suitable for narrowband, linear modulation systems, prevailing PSK.

Schemes suitable for use with OFDM were also published, and frequency-selective fading channels have also been considered. Moreover, investigations on using space-time turbo coding and performance of concatenated coding schemes have also been discussed in recent papers.

There are many theoretical results on this topic, though these schemes didn't appear in the everyday practice. Nevertheless, many wireless standards include a possibility of some form of space-time coding.

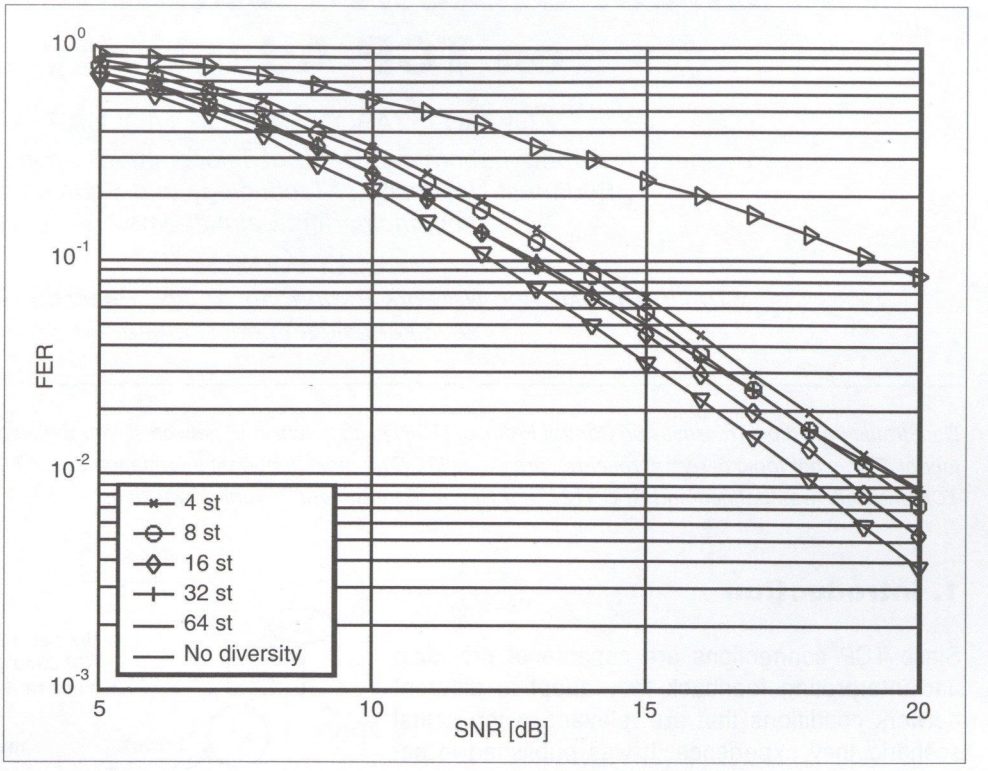


Figure 3 Frame error ratio versus SNR

Summary

Finite bandwidth resources and physical properties of the wireless medium ultimately limit the data rates achievable by conventional means. Hence it is important to look for methods that are capable of increasing data rates significantly without increasing the allocated frequency band and power requirements. Latest research efforts recognized the potential of better exploiting the spatial dimension of the wireless transmission. Combined spatio-temporal processing and coding is an important step into the direction of increased spectral efficiency and decreased outage. Their widespread use will enable notable quality improvement in mobile communications (wireless LANs, mobile telephony) and in broadband wireless access systems.

References

[1] I. E. Telatar, "Capacity of multiple input-multiple output Gaussian channels" Eur. Trans. Telecom., Vol. 10, pp. 585-595, Nov. 1999.

[2] G. J. Foschini, M.J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas", Wireless Personal Comm., Vol. 6, No. 3, pp.311-335, Mar. 1998

[3] J.D. Parsons, The mobile radio propagation channel, Wiley, 2000

[4] S. M. Alamouti, "A simple transmit diversity technique for wireless communications", IEEE Journal on Selected Areas in Communications, Vol. 16, No. 8., October 1998

[5] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: performance criterion and code construction", IEEE Trans. Information Theory, Vol. 44, No. 2, pp.744-765, March 1998

[6] V. Tarokh, H. Jafarkhani, A. R. Calderbank, "Space-time block codes from orthogonal designs", IEEE Trans. Inform. Theory, Vol. 45, No. 5, pp.1456-1467, July 1999

[7] S. Baro, G. Bauch, A. Hansmann, "Improved codes for space-time trellis coded modulation", IEEE Comm. Letters, Vol. 4, pp. 20-22, Jan. 2000

[8] Z. Chen, J. Yuan, B. Vucetic, "Improved space-time trellis coded modulation scheme on slow Rayleigh fading channels", Electron. Lett., Vol. 37, pp. 440-441, Mar. 2001

[9] M. Tao, R. S. Cheng, "Improved design criteria and new trellis codes...", IEEE Comm. Letters., Vol. 5, No. 7, pp.313-315, July 2001

[10] B. M. Hochwald, W. Sweldens, "Differential unitary space-time modulation", IEEE Trans. Comm., Vol. 48, No. 12, pp.2041-2052, Dec. 2000

[11] I. Frigyes, P. Horváth, "Mitigation of rain-induced fading: route diversity vs. Route-Time Coding", PM 4005, COST 280, 4th MC meeting, Prague, 2002

The Effect of Active Buffer Management on TCP Adaptivity

ZOLTÁN SZABÓ, SÁNDOR MOLNÁR
High Speed Networks Laboratory, Dept. of Telecommunications and Telematics
Budapest University of Technology and Economics
{szabo, molnar}@ttt-atm.ttt.bme.hu
ZSOLT KENESI
Traffic Analysis and Network Performance Lab., Ericsson Hungary Ltd
zsolt.kenesi@eth.ericsson.se

Reviewed

Data traffic based on Transmission Control Protocol (TCP) is dominant in IP networks. The detailed analysis of the features and behaviour of TCP is hot topic of recent research programs [3]. (The most important information about TCP can be found in RFC 793 in which TCP originally was defined and RFC 1122 and RFC 2001 that contain further additions.

1. Introduction

Since TCP connections are capable of providing and interpreting feedback they adapt to different network conditions that are relevant to the actual scenario they experience. It was published in papers [1] and [2] that if a TCP connection shares a bottleneck link with non-adaptive background traffic flow, TCP adapts to it inheriting and propagating the correlation structure and statistical properties of the background traffic flow above a characteristic time scale. This time scale of adaptation depends on the end-to-end path properties.

Service providers' IP networks support traditionally public Internet service, so a single best-effort serviceclass was adequate to support all Internet application. With different applications requiring different QoS it is important to introduce QoS support into the network. The QoS architecture of IP networks is the Differentiated Service solution [4].

The DiffServ architecture provides QoS by dividing traffic into different categories and ensures differential treatment of traffic within each class. Active queue management techniques and queue scheduling algorithms are currently implemented in most of the routers (Cisco, Juniper) to support different QoS classes in IP networks.

In our research we focus on the analysis of the effects of widely spread quality of service mechanisms on TCP adaptivity.

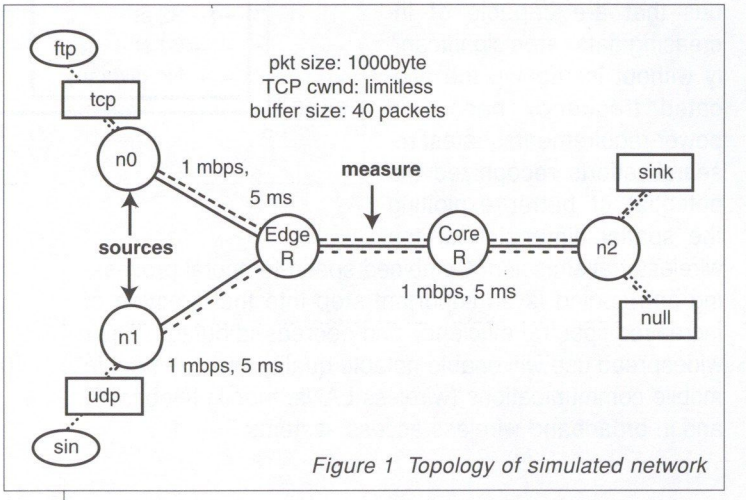


Figure 1 Topology of simulated network

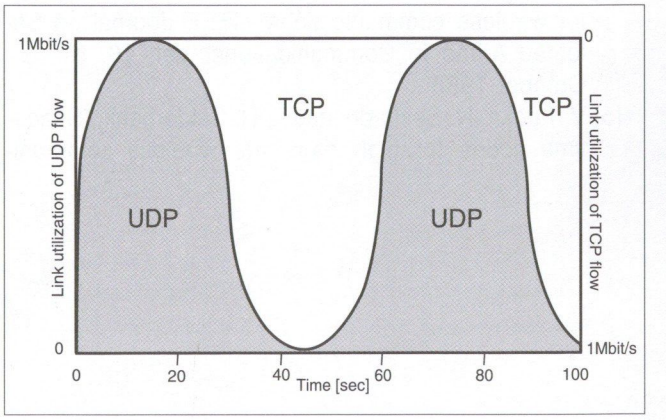
2. Characteristics of TCP adaptivity

In our work we analysed TCP adaptivity that is described by characteristics of adaptivity firstly published in [2]. The applied topology is depicted in *Figure 1.*

Since TCP connections are capable of providing and interpreting feedback they adapt to different network conditions that are relevant to the actual scenario they experience. It was published in [2] that if a TCP connection shares a bottleneck link with a non-adaptive background traffic flow, TCP adapts to it inheriting and propagating the correlation structure and statistical properties of the background traffic flow above a characteristic time scale. This time scale of adaptation depends on the end-to-end path properties, i.e. round-trip time, widow size, etc.

In the cited paper the measure of adaptivity, i.e., the ratio of power spectra of TCP and background traffic at a given frequency, was investigated.

Figure 2 Fluctuated UDP and totally adapted TCP traffic on bottleneck link



During the analysis the rate of non-adaptive UDP background traffic changed as a function of sine wave with constant amplitude, as we discuss below.

The power spectrum of sine wave of a given frequency f consists of a single frequency component at f. If TCP is able to adapt to the fluctuations of the background traffic flow, it changes as a function of sine wave of a given frequency f, see *Figure 2*. In this case, the same frequency component f should appear as a significant spike in the power spectrum of the TCP traffic, as well. Note that the background traffic utilised half of the bandwidth in average at each frequency setting.

The measure of adaptivity is defined as spectral density of the adapting TCP rate process divided by the spectral density of the background traffic at a given frequency.

An important question is how this measure changes with the frequency of background traffic, i.e., with time-scale of fluctuations. It is possible to plot the adaptivity characteristics of TCP by performing the above mentioned experiment for a wide range of time-scales.

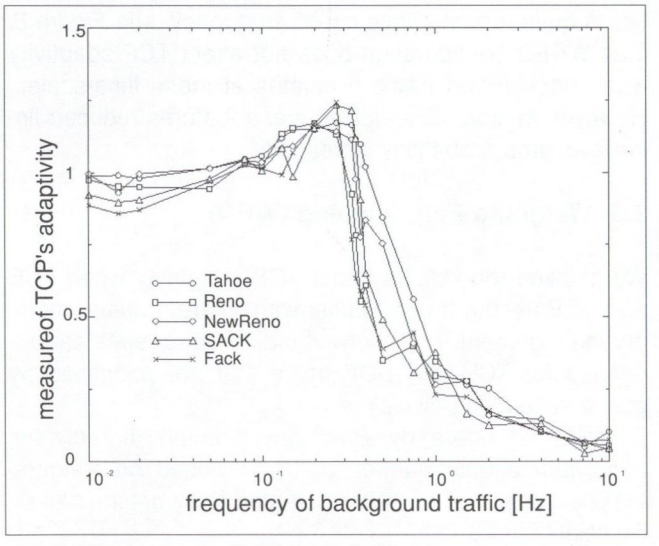
Figure 3 shows the characteristics of adaptivity. It can be observed that TCP adaptivity is almost invariant to TCP versions. If the time-scale the background traffic fluctuates on is large ($f_{background} < 0.1$Hz), the measure of adaptivity is close to 1. It means that TCP is able to adapt almost perfectly to the background traffic. However, on smaller time-scales ($f_{background} > 0.3$Hz) TCP cannot adapt to the changes of the background traffic, therefore the ratio of the TCP and UDP frequency component is very small.

In the range of 0.1-0.3 1/sec a resonance effect can be observed. At these time-scales TCP is more aggressive, and gains even higher throughput than what is left unused by the non-adaptive background flow.

3. QoS Mechanisms

The Differentiated Service architecture [4] provides QoS over IP networks by classifying traffic and applying different QoS classes.

Figure 3 Characteristic of TCP adaptivity



In our measurement we differentiated the TCP and non-adaptive background UDP traffic so that the flows were classified by their transmission protocol. We defined different QoS classes for non-adaptive UDP flow and for TCP-based data traffic.

Previously reported [1] characteristics of TCP adaptivity describes how TCP was able to adapt to periodically changing background traffic while TCP and UDP met in a shared buffer. The tail-drop buffer size is 40 packets in the configuration. There is no queue management algorithm implemented in tail-drop queues, they drop packets when the queue is full. Tail-drop queues treat all traffic in the same way.

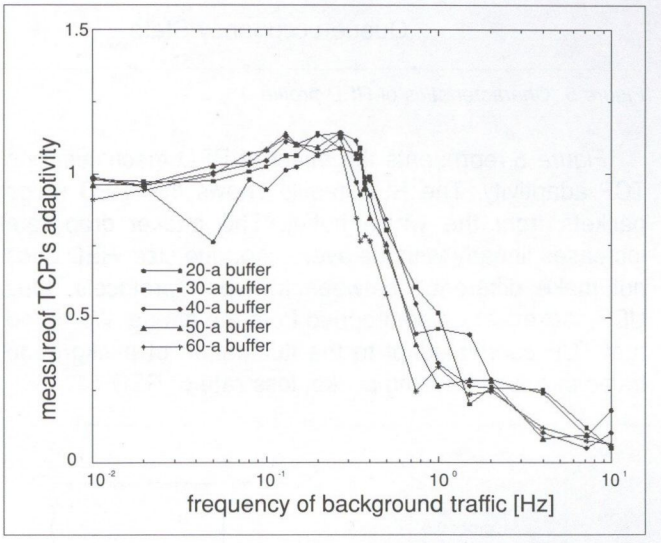


Figure 4 Effect of changed buffer size on TCP adaptivity

In practice, buffer size is an important setting in the network configuration. *Figure 4* shows the effect of buffer size on the curve of TCP adaptivity.

It can be observed that TCP adaptivity almost independent of the buffer size therefore we kept 40 packets for the initial buffer size in the following scenarios.

3.1 Random Early Detection (RED)

RED mechanism was proposed by Sally Floyd and Van Jacobson in the early 1990s to address network congestion in a responsive rather than reactive manner. RED offers a widespread, effective congestion-avoidance mechanism. Like Tail Drop, RED treats all traffic equally in a given buffer. But RED takes advantage of the congestion control mechanism of TCP. RED indicates the reduction of the source transmission rate by randomly dropping packets prior to periods of high congestion. RED uses a packet drop profile to control the aggressiveness of its packet discard process. The drop profile defines a range of drop probabilities across a range of queue occupancy states, see *Figure 5*.

In light congestion phase the RED starts to drop packets and it will decrease its transmission rate until all the packets reach their destination which indicates that the congestion is cleared.

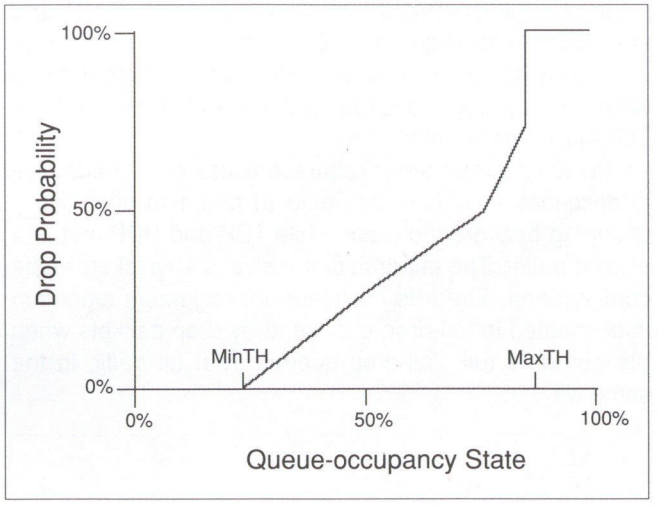


Figure 5 Characteristics of RED profile

Figure 6 represents the effect of RED mechanism on TCP adaptivity. The RED profile shows that RED drops packets from the whole buffer. The packet drop rate increases linearly with the average queue size. RED does not make difference between transport protocols, thus UDP packets are also dropped from the queue. We found that TCP cannot adapt to the fluctuation of background traffic due to increasing packet loss rate of RED.

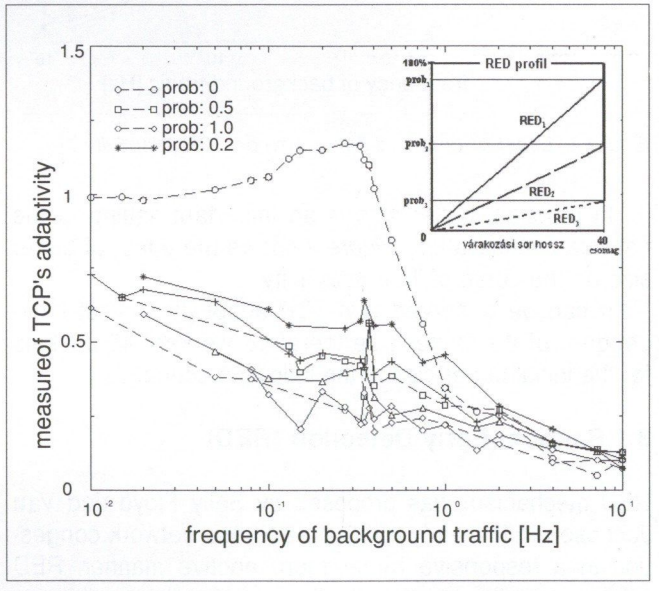


Figure 6 Effect of RED mechanism on TCP adaptivity

The result clearly shows that using RED for UDP traffic is pointless. In the following we analyse the effect of another RED mechanism which differentiate between TCP and UDP protocol.

3.2. Weighted Random Early Detection (WRED)

To support QoS in IP network it is essential to differentiate the TCP-based, loss-sensitive data traffic from real-time, delay-sensitive UDP stream even if these flows meet in a shared buffer. WRED is an extension to RED that allows you to assign different RED drop profiles to different types

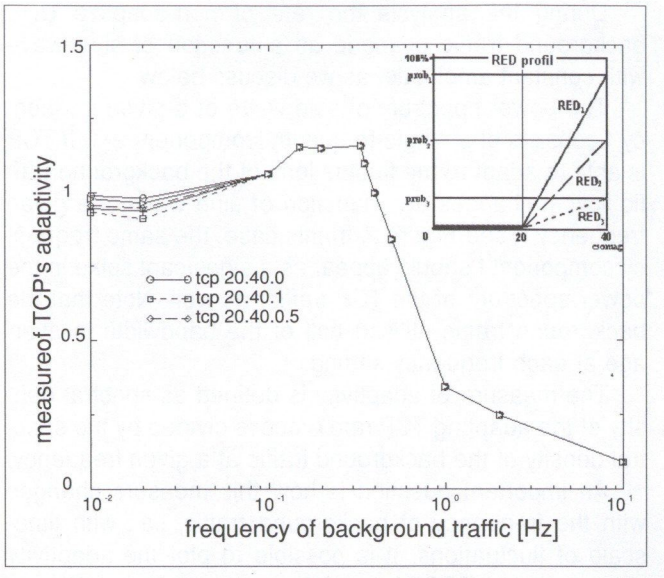


Figure 7 Effect of RED mechanism on TCP adaptivity (less aggressive drop profile for TCP)

of traffic in the same queue. In our work we defined different RED profile to TCP and UDP flow.

Applying WRED active queue memory management in the shared buffer queuing we was able to define different QoS classes with different RED parameters for TCP and UDP in the shared buffer.

In this case we analysed the effects of WRED on TCP features. To push an advantage of WRED we defined different RED profiles for TCP and no RED profile for UDP, which means that, UDP flow was able to utilise the whole shared buffer without packet loss.

In the following we present the effects of different WRED configurations on TCP adaptivity. We defined a less aggressive RED profile for TCP and no RED profile for UDP. This setting results in that UDP queuing works as Tail Drop mechanism and TCP flow does not loss packet below the queue occupancy level of 50 percent. *Figure 7* illustrates that this WRED configuration has no effect on TCP adaptivity.

In the second scenario more aggressive RED profile is applied. It can be noticed that the adaptivity characteristics is quite robust above cut-off frequency, see *Figure 8*. This WRED configuration does not affect TCP adaptivity when background traffic fluctuates at these time-scales. However, in small time-scales these features reduced linearly as drop probability increased.

3.3. Weighted Fair Queuing (WFQ)

We showed the results about TCP adaptivity when TCP and UDP used a shared buffer with different queue memory management. In IP networks different queues can be defined for TCP and UDP traffic that are prioritised by queue scheduling disciplines.

WFQ [7] offers dynamic, fair queuing that divides bandwidth across queues of traffic based on weights. WFQ ensures that all traffic is treated fairly determined by its weight.

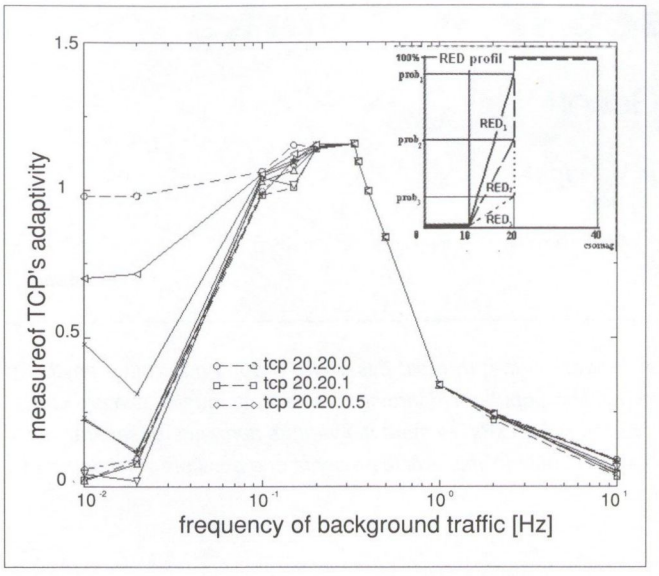


Figure 8 Effect of RED mechanism on TCP adaptivity (more aggressive drop profile for TCP)

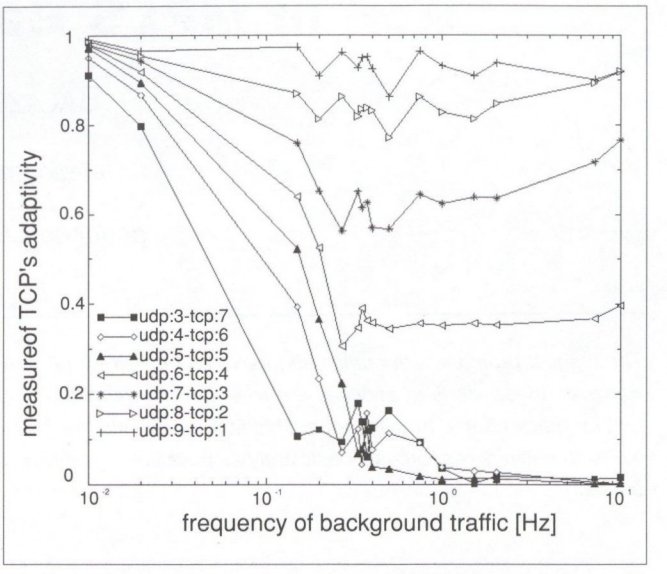


Figure 9 Effect of WFQ mechanism on TCP adaptivity

WFQ is an automated scheduling method that provides fair bandwidth allocation to all network traffic based on the applied priority. WFQ is a flow-based algorithm that simultaneously schedules interactive traffic to the front of a queue to reduce response time and fairly shares the remaining bandwidth among high-bandwidth flows.

We also analysed the effect of WFQ with different priorities for UDP flow on TCP adaptivity.

Figure 9 shows that TCP adaptivity is dependent on priority levels allocated to UDP and TCP applying WFQ. In case of higher priority for UDP serving TCP flow is able to adapt better to background traffic on all time-scales.

Setting WFQ queue scheduling discipline in our configuration we defined different buffer for TCP and UDP flow instead of one shared buffer, which caused variation on TCP adaptivity as Figure 8 shows.

We observed that in case of TCP and UDP flows are separated into different buffers TCP does not suffer heavy packet losses and able to send packets when UDP packets are not transmitted.

Summary

In our research we focused on how TCP adaptation is affected by active buffer management techniques. Our results show that in case the TCP shares bottleneck with non-adaptive background traffic flow, i.e., UDP traffic, the characteristics of adaptivity can be modified by applying RED or WRED queuing mechanisms and the measure of variation depends on different RED and WRED settings.

If the traffic is separated into different buffers and served by WFQ, TCP adaptivity is almost perfect at each time-scale.

All results can also contribute to the better understanding of the fractal properties of TCP traffic.

References

[1] A. Veres, Zs. Kenesi, S. Molnár, G. Vattay, TCP's Role in the Propagation of Self-Similarity in the Internet, appear in Computer Communications, Special issue on Performance Evaluation of IP Networks and Services.

[2] A. Veres, Zs. Kenesi, S. Molnár, G. Vattay, On the Propagation of Long-Range Dependence in the Internet, ACM Computer Communication Review, Vol. 30, No. 4, pp. 243-254, October, 2000.

[3] Gary R. Wright, W. Richard Stevens, TCP/IP Illustrated, Volume 2 The Implementation, Addison-Wesley Publishing Company 1995

[4] S. Blake, D. Black, M. Carlson, An Architecture for Differentiated Services, Request for Comments 2475, December, 1998

[5] Sally Floyd, Van Jacobson, Random Early Detection Gateways for Congestion Avoidance, IEEE/ACM Transactions on Networking, 1993

[6] Chuck Semeria, Supporting Differentiated Service Classes: Active Queue Memory Management, Juniper Networks, White Paper, 2002, http://www.juniper.net

[7] H. Zhang, Service Disciplines For Guaranteed Performance Service in Packet Switching Networks, Proceedings of the IEEE, 83(10), Oct 1995

QoS in MPLS Based IP Networks

BALÁZS GÓDOR

Development Manager
MATÁV-PKI
godor.balazs@ln.matav.hu

Reviewed

The biggest problem with multimedia services offered by the Internet network is that in most cases, ISPs don't guarantee anything. However, these kinds of services are free, but not reliable enough to become popular. Regarding a service, its quality is more important for many people than its price. This fact represents the greatest challange for ISPs: to meet the various demands of several users by using network resources as efficiently as possible. By utilizing Multi-VC technology, this article presents one possible solution for this problem.

Most user applications present different network requirements. Consider the difference between voice transmission and FTP. Voice traffic is quite sensitive to delay and jitter while FTP is not at all influenced by these parameters. The network should meet strict requirements in order to make it suitable for the transmission of both delay sensitive and insensitive traffic. This may lead to wasteful solutions since in several cases, users are apt to provide excessive resources.

1. QoS in IP networks

QoS refers to the ability of a network to provide improved service to specific network traffic by means of various essential technologies. In particular, QoS provides improved and more predictable network service by offering the following features:

- Supporting dedicated bandwidth
- Improving loss characteristics
- Avoiding and managing network congestion
- Shaping network traffic
- Setting traffic priorities across the network

There are three important IP QoS service models:

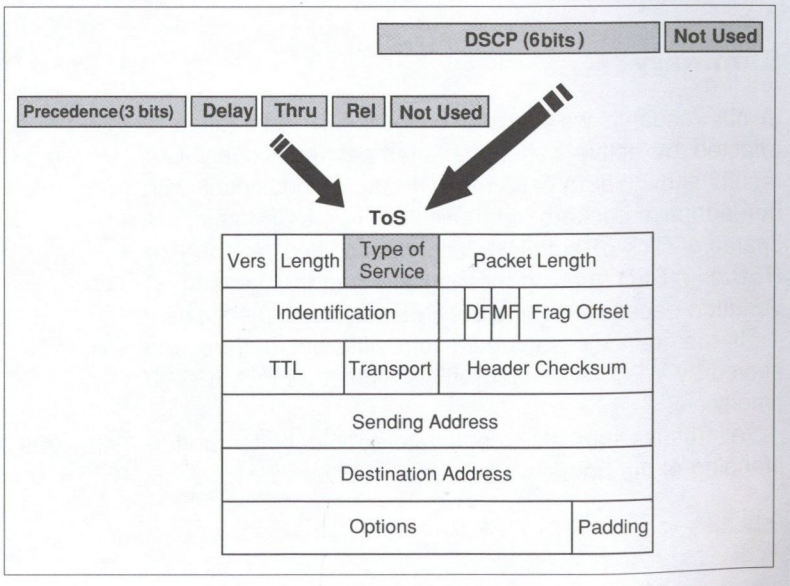
- IntServ
  - RSVP
- DiffServ
- ECN

Let us take a closer look at the know-how of DiffServ. In contrast with IntServ (reserving network resources for flows generated by applications), DiffServ reserves network resources for traffic classes. Think of two classes. Traffic belonging to the first one does not get any special treatment. It gets the best-effort service that has been the only network service since the early days of IP. Traffic belonging to the other class do get some guarantee for delay, jitter or packet loss originating from the network. In the case of DiffServ, some bits in the header of the IP packets show the class they belong to. DiffServ is much simpler than IntServ because an extra protocol is present to show the class of a flow. These bits, called DSCP,.are the first six bits of the ToS byte in the IP header.

In theory, DSCP can be used to differentiate between 2^6 (=64) classes but in practice, much less is used. The DSCP tells the router something about how the packet should be treated from a QoS aspect. Formally, the DSCP identifies a "per-hop behaviour" (PHB). There are some standard PHBs, and it is also possible to define local PHBs within a network [1]. Some of the standard PHBs are the following:

- Default: No special treatment, equivalent to best-effort.
- Expedited Forwarding (EF): Packets marked EF should be forwarded with minimal delay and experience low loss. This could be realized by putting all such packets in a dedicated EF queue, and ensuring that the arrival rate of packets to the queue is less than the service rate (RFC2598).

Figure 1 IP header



- Assured Forwarding (AF): A set of AF PHBs is defined in the following way: Each PHB in the set is AFxy for a range of values of x and y. The value of x is referred to as the AF class, and typically selects a queue for the packet, while the value of y determines the drop preference for the packet. The greater the value of y the bigger the possibility that the packet in question will be dropped. The recommended number of AF PHBs is 12, representing four AF classes with three drop preference levels in each. [1]. Packets belonging to the same class (Afx1, Afx2, Afx3) will be placed into the same queue. (RFC2597)

The DiffServ architecture is an elegant way to provide service discrimination within a commercial network. Customers willing to pay more will see their applications receive better service than those paying less. A traffic class is a predefined aggregate of traffic. Compared with IntServ, traffic classes in DiffServ are accessible without signalling which means they are readily available to applications without any setup delay. Consequently, traffic classes can provide qualitative or relative services to applications that cannot state their requirements quantitatively. This conforms to the original design philosophy of the Internet. An example of qualitative service is "traffic offered at service level A will be delivered with low latency," while a relative service could be "traffic offered at service level A will be delivered with higher probability than traffic offered at service level B." Quantitative services can also be provided by DiffServ. A quantitative service might be "90 percent of in-profile traffic offered at service level C will be delivered." [6]. The lack of the signalling protocol on one hand is an advantage, but on the other hand, it can be a disadvantage as well.

The entering packets can be bound to any destination in the Internet, and may thus be routed towards any border router of the domain (except the one where it entered). Because of the lack of a signalling protocol, a substantial proportion of the entering packets might, in the worst case, all exit the domain through the same border router. We can see that unless resources are massively overprovisioned in both interior and border routers, traffic and network dynamics can cause momentary violation of service agreements, especially those related to quantitative services. On the other hand, massive overprovisioning results in a very poor statistical multiplexing gain, and is therefore inefficient and expensive [6].

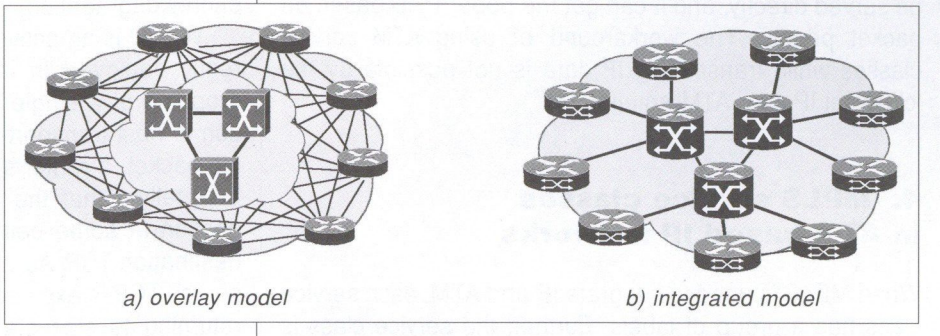
2. MPLS based IP networks

MPLS can be applied between the second and third OSI network layers. In practice, IP is dominant in the third layer, and the second layer is either packet switched (PoS, Ethernet, etc.) or ATM. Now we are going to examine networks where MPLS is responsible for the co-operation between IP (layer 3) and ATM (layer 2). MPLS is a combination of label switched packet forwarding and network layer routing. The application of MPLS gives the network more commercial efficiency and better scalability [2].

Integrating the fast but less intelligent switching and the intelligent routing, less third layer packet processing is needed in a given network. If only second layer processing takes place at each hop in the core network, the delay will be decreased, and also less processor-power will be consumed by the network elements [3].

With the help of MPLS, IP and ATM can be integrated more efficiently. This is important because IP networks are normally connected through ATM PVC-s but this solution is neither scalable nor managable. Think of the routing protocols (especially OSPF) where big amount of routing updates must be processed.

Figure 2 IP over ATM



a) overlay model *b) integrated model*

In case of the overlay model (*Figure 2/a* – MPLS not used), the ATM network is transparent to the IP network. From the point of view of IP routing, a PVC, passing through several hops, is equivalent to another PVC passing through one hop. The breakdown of an ATM node in this architecture can lead to IP node failures too, further it will be quite hard to locate the erroneous ATM node because of the IP transparency.

With the application of MPLS in the integrated model, ATM switches can also be treated as IP nodes *(Figure 2/b)*. They can be identified at IP level so the network becomes not only more scalable but more reliable too [4].

3. ATM based MPLS networks

MPLS nodes (LSRs) can be examined in a layered architecture comprising two layers. The upper one is the control plane while the lower is called data plane. The task of the data plane is to forward packets according to forwarding tables while the control plane is responsible for the validity of the information at the forwarding tables. An ATM MPLS node comprises also two physical parts: a router (LSC) for the control functions and an ATM switch for the packet forwarding. In some equipment, both functions are realized within one hardware. These equipment provide different services within the same network at the same time. This way, an MPLS-ATM switch can act either like a traditional ATM switch (responding to ATM signalling protocols) or like an MPLS node. It is suitable to provide high bandwidth, QoS and also real time services simultaneously.

The interface (VSI) between the router (LSC) and ATM switch has been standardized by the MSI (Multiservice Switching Forum). With the help of the VSI, the LSC is able to modify the ATM VPI/VCI values according to the labels, i.e. to control the switching field. From the LSC aspect, the interfaces of the ATM switch (through the VSI) look like its own interfaces [5].

MPLS uses labelled connections, a new category in addition to the swiched (SVC) and permanent (PVC) connection types. The bandwidth of trunks between ATM switches can be shared arbitrarily by the SVCs, PVCs and the LSC [3].

By the integration of IP and ATM by MPLS, packets belonging to MPLS connections can be placed into queues differing from those comprising SVC or PVC packets. For this reason, it is not necessary to use 'ATM Forum service classes' as translational points. MPLS traffic can be served directly, and it can get the poper PHB based on packet priority. The workaround of using ATM service classes while transmitting IP data is not possible by the 'classical IP over ATM' solution [4].

4. MPLS service classes in ATM based IP networks

When MPLS is used to integrate IP and ATM, each service class has a group of labels. Further, the service class is coded in each label. At most four labels can be assigned to each source-destination pair according to service classes that are provided on this path. The four labels mean four individual service classes and four parallel LVCs. Considering the labels, backbone LSRs use some kind of queuing in order to provide the proper bandwidth and containers for the service classes. Cells belonging to different service classes will be placed into differing queues. This way, we have more control over the delay. Weights assigned to service classes are relative and not absolute.

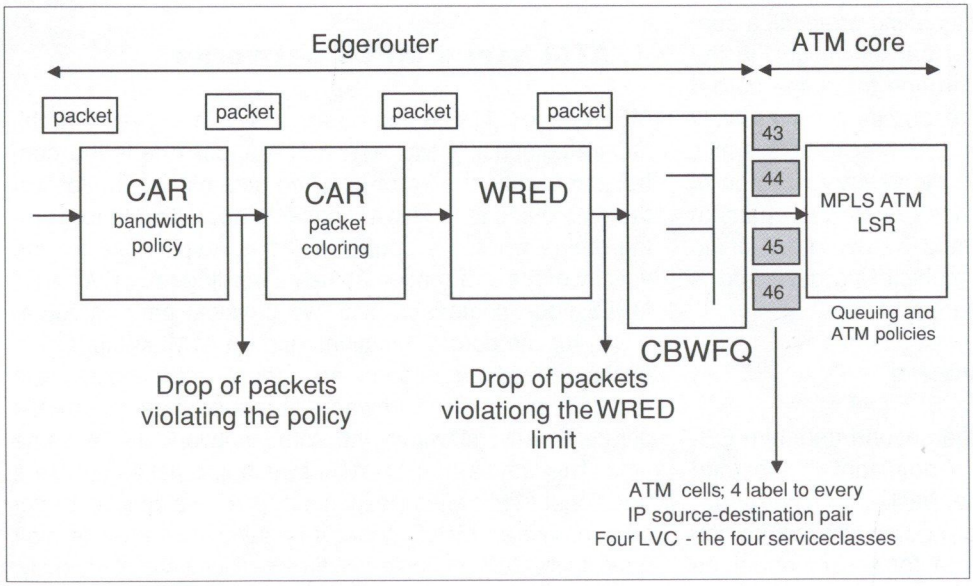
When a high priority traffic is not using the bandwidth allocated for it, and a low priority traffic is being congested, the low priority traffic may utilize the bandwidth left free by the high priority traffic. *Figure 3* shows IP QoS multiple LVCs.

The first 'CAR' box in Figure 3 is responsible for the bandwidth allocation to the service classes. The second 'CAR' box is resonsible for the colouring of packets. This means that packets not violating the bandwidth policy are going to be coloured, i.e. the priority bits will be set somehow in the ToS byte of the IP packet header. Due to this setting, packets (of a given bit pattern) belonging to individual service classes will be recognizable anywhere in the network. Packets with different colours will be treated in a different way (different PHBs – cf. DiffServ). With CAR you can define policies, and define procedures applied when packets are violating the given policy (for example: drop, priority degradation).

WRED is an active queue management technique currently deployed in large IP networks. With WRED, the dropping of a single packet is enough to signal congestion to host transport-layer protocols. By discarding a single packet, a router sends an implicit warning to a source TCP telling that the discarded packet experienced congestion at some point along the end-to-end path to the destination TCP. As a response to this implicit warning, the source TCP is expected to reduce its transmission rate (by returning to slow-start or fast recovery with congestion avoidance) so that the router's queue buffer does not overflow [7]. The service class is defined by the precedence bits; in the case of ATM based MPLS, only 4 classes can be applied. Therefore, four parallel LVCs can be used between a given source and destination. Within each LVC you can create two further classes using the CLP bit at the header of the ATM packet. To sum up, 8 classes can be realized.

5. Case study – QoS in ATM based MPLS backbone

Figure 3 Multiple LVCs



The primary goal with the demonstration network was to check the proper functioning of Mutli-VC technology, i.e. to investigate whether QoS policies work in a real ATM based MPLS network. Although the size of an ATM PVC can be reduced arbitrarily, the size of an LVC can not be reduced.

This means that you cannot allocate a defined very small bandwith for an LVC. It has been a serious drawback during the examinations, because on leased lines from CE routers (Exeter, Sevilla, Venice), 8Mb/s was the maxi-

mum configurable bandwidth, much less than the 155 Mb/s backbone bandwidth. This is not an optimal situation because traffic arrives from CE sections to the backbone, and their behaviour has to be examined there, compared to each other in case of congestion.

The expression 'compared to each other' must be emphasised here, because high bandwidth (155 Mb/s) traffic generated by the traffic generator (TG on *Figure 4.*) will be mixed in the backbone with the low bandwidth (8 Mb/s) CE traffic, and their respective behaviour there should be investigated.. These traffic flows differ by orders of magnitude so we really have to be on guard to be able to analyse the results properely.

Configuring Multi-VC

You need the following configuration on the ATM 1/0 interface of Bremen and Dortmund PE routers:

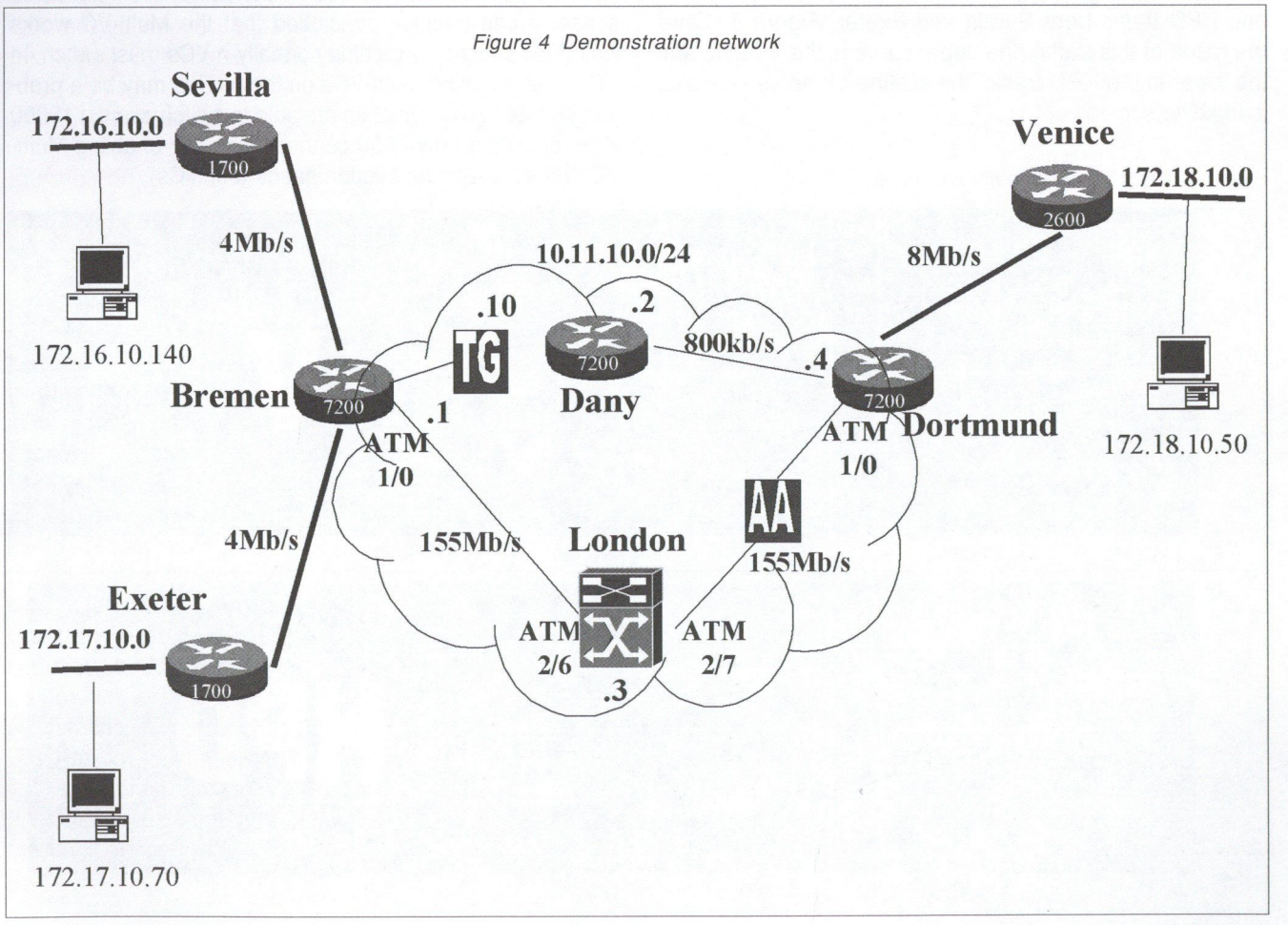
```
interface ATM1/0.1 tag-switching
  ip unnumbered Loopback0
  service-policy output LLQ3      ! congestion policy
  mpls label protocol ldp         ! LDP is the Label
                                    Distribution Proto.
  tag-switching atm multi-vc
  tag-switching atm vpi 10-15 vci-range 33-65535
  tag-switching ip
```

Regarding packets arriving from CE routers to the core, the ATM 1/0 interface of Bremen is an outgoing interface. So it is reasonable to define the bandwidth policy here. This policy can be defined at London LSR as well:

```
tag-switching atm cos available 91
tag-switching atm cos standard 1
tag-switching atm cos premium 3
tag-switching atm cos control 5
```

With the configuration lines above (on interfaces ATM 2/6 and 2/7 in London LSR), the bandwidth distribution between service classes can be defined. Keywords 'available', 'standard' and 'premium' represent service classes. In the demonstration network, premium class was used for voice (V) traffic, standard class for high priority data (HPD), and available class for best-effort (BE) traffic. The reason why exactly three classes are implemented is the following. With these three classes, the network traffic of a typical company can easily be modelled. When communicating, emloyees usually make phone calls (V = voice traffic), send e-mails or faxes, or browse over the Internet (BE = best-effort traffic). Furthermore, also high priority data (HPD) forwarding can be provided. This can be useful for the management of a company, enabling it to handle time sensitive transactions too. Packets belonging to these three classes have different bit patterns in their precedence field of the ToS byte.

Figure 4 Demonstration network



By entering the two configurations described above and by providing suitable traffic routing, parallel LVCs are established between source and target addresses (Multi-VC mode). The network is thus ready to provide quality services.

Packets belonging to different service classes are simultaneously present in the core of the network, but receive 'differentiated service'. In the following example Multi-VC was used to realize a DiffServ model in an ATM based IP backbone. Let us see now the practical significance of this model on hand of the following example.

High bandwidth (155Mb/s–20MB/s), low priority traffic is generated by a TG (traffic generator) to the target station Dany. On their way, these packets had to pass through the backbone. The backbone could thus be loaded so as to provide ideal conditions for congestion. Simultaneously, the Dortmund-Venice CE section did not get overloaded.

This is advantageous because the 4Mb/s HPD and the 4Mb/s V traffic from Sevilla and Exeter to Venice will not be congested at the Dortmund-Venice section (since the capacity of Dortmund-Venice section is just 8Mb/s; equal to the sum of all traffic coming from CE routers Sevilla and Exeter). This is why the analyzer software running on the Venice PC will show the actual traffic relations in the backbone.

The *Table (above)* shows the allocated bandwidth for each type of traffic in the backbone. Next, we started generating the high bandwidth BE traffic by the TG (in order to establish congestion) and the low bandwidth (4Mb/s) V and HPD traffic from Sevilla and Exeter. *Figure 5* shows the result of this traffic (the upper curve is the V traffic and the lower is the HPD traffic; the scaling of the vertical axis is in kbyte/sec).

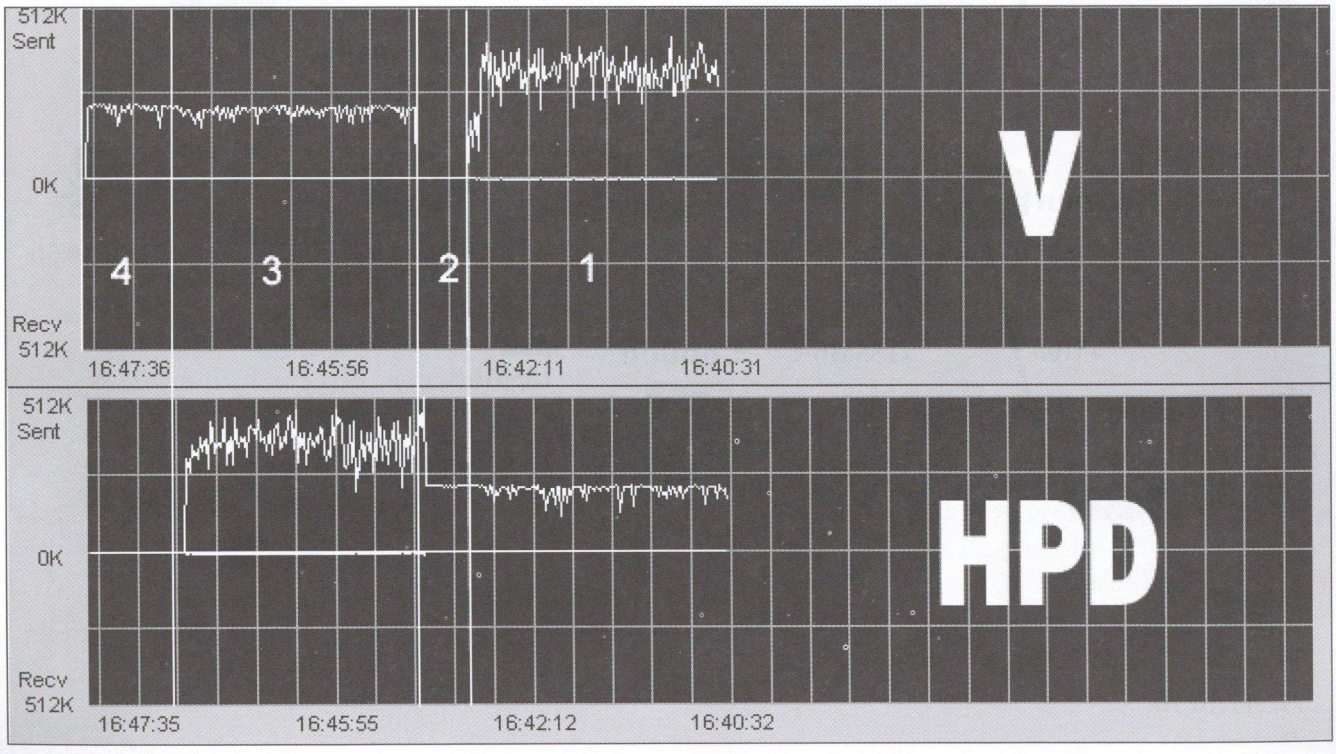
	%	[kB/s]	[kb/s]
Best-Effort	71	14129	112677
High Priority Data	1	199	1587
Voice	3	595	4761

In the first section of Figure 5, HPD received 1% and V received 3% of the total core bandwidth (155Mb/s) in case of congestion, according to the service policy. HPD, arriving at a rate of 500kB/s, could fully occupy the allocated bandwidth (~200kB/s). Because of congestion, the policy limits the band of HPD at that rate (200 kB/s) so a straight border line can be seen in Figure 5. Simultaneously, voice is arriving at a rate of 500kB/s, and 3% (~600kB/s) is allocated, and this is not limited (no congestion). This rate corrsponds to the maximum possible speed in these circumstances. These results confirm the proper functioning of the QoS policies.

In the second section of Fig. 5, V traffic is not generated, only HPD taffic. At the background, of course, as in every case during the investigations, there is a 20MB/s BE traffic loading the backbone. So the second section contains only the BE and HPD traffic (HPD is still congested), and the curve of HPD is much straighter than before at 200kB/s. This confirms again the proper functioning of the QoS policy.

In Sections 3 and 4 of the Figure (to be sure that the results are correct), the HPD and V traffic have been inverted in the policy map. The result is the same, but in a reverse sense. It can thus be concluded that the Multi-VC works well. ATM switches switching normally n VCs must switch 4n VCs after enabling Multi-VCs on them. This may be a problem because ATM switches are able to switch at most 16000 VCs, providing only 4000 connections after enabling Multi-VCs (since every connection needs four VCs).

Figure 5 Traffic relations in the backbone



Summary

The warranties of service providers apply for traffic classes, and the resources are also allocated for classes. Traffic flows belonging to classes must share the resources and the bandwidth allocated for the entire class. For this reason, DiffServ is not appropriate for providing real-time data services. It's not a good idea to control e.g. a power plant over a DiffServ network.

If someone needs real-time data service in an IP network, in advance resource allocation for flows is indispensable. Nevertheless, DiffServ is a quite popular QoS model (DiffServ has been realized in the above case study too). The reason for its popularity is its simplicity (no need for an additional signalling protocol) and scalability. However, DiffServ is not a universal solution among the QoS techniques.

REFERENCES

[1] MPLS Technology and Applications
Bruce Davie, Yakov Rekhter
Academic Press, 2000, USA, ISBN 1 55860 656 4

[2] MPLS and VPN achitectures
Ivan Pepelnjak, Jim Guichard
Cisco Press, ciscopress.com

[3] IP platform értéknövelt szolgáltatások kialakítására
(IP platform for establishing value added services)
Dr. Varga Balázs, Géczi Csaba, Onder Zoltán
2000.06.27. Budapest, Matáv PKI

[4] Híradástechnika
Címkék az Interneten (Labels on the Internet)
Gáspár Csaba, Láposi Levente, Tapolcai János, Laborci Péter
LVII. évf. 2002/2. szám, HU-ISSN 0018-2028

[5] MPLS study
Géczi Csaba, Dr. Varga Balázs
1999.10.29. Budapest, Matáv PKI

[6] The Internet: A Global Telecommunications Solution?
Laurent Mathy, Christopher Edwards, David Hutchinson
IEEE Network, July/August 2000, 0890-8044/00
http://www.comp.lancs.ac.uk/computing/users/laurent/papers/full/ieeenetwork00.pdf

[7] Supporting Differentiated Service Classes:
Active Queue Memory Management
White Paper, February 2002
Chuck Semeria, Juniper Networks
http://www.juniper.net/solutions/literature/white_papers/200021.pdf

[8] RFCs : (http://www.ietf.org)
2597 Assured Forwarding PHB Group. J. Heinanen, F. Baker, W. Weiss, J. Wroclawski. June 1999.
(Format: TXT=24068 bytes) (Updated by RFC3260)
(Status: PROPOSED STANDARD)
2598 An Expedited Forwarding PHB. V. Jacobson, K. Nichols, K. Poduri. June 1999.
(Format: TXT=23656 bytes) (Obsoleted by RFC3246)
(Status: PROPOSED STANDARD)
3270 Multi-Protocol Label Switching (MPLS)
Support of Differentiated Services. F. Le Faucheur, L. Wu, B. Davie, S.Davari, P. Vaananen, R. Krishnan, P. Cheval, J. Heinanen. May 2002.
(Format: TXT=137960 bytes)
(Status: PROPOSED STANDARD)

ABBREVIATIONS

ACL	Access Control List
AF	Assured Forwarding
ATM	Asynchronous Transfer Mode
b/s	bit per second
B/s	byte per second
CAR	Committed Access Rate
CBWFQ	Class-Based Weighted Fair Queuing
CE	Customer Edge
CLP	Cell Loss Priority
CoS	Class of Service
DiffServ	Differentiated Services
DSCP	DiffServ Code Point
ECN	Expressed Congestion Notification
EF	Expedited Forwarding
FTP	File Transport Protocol
IntServ	Integrated Services
IP	Internet Protocol
LDP	Label Distribution Protocol
LSC	Label Switch Controller
LSP	Label Switched Path
LSR	Label Switch Router
LVC	Label Virtual Connection
MPLS	Multiprotocol Label Switching
OSI	Open System Interconnection
OSPF	Open Shortest Path First
PE	Provider Edge
PHB	Per Hop Behaviour
PoS	Packet over Sonet
PVC	Permanent Virtual Connection
PVP	Permanent Virtual Path
QoS	Quality of Service
RED	Random Early Detection
RSVP	Resource Reservation Protocol
SVC	Switched Virtual Connection
ToS	Type of Service
VCI	Virtual Channel Identifier
VPI	Virtual Path Identifier
VSI	Virtual Switch Interface
WRED	Weighted RED

Stability and Noise Characteristics Improvements of Mode-Locked Laser Sources

TAMÁS BÁNKY

Department of Broadband Infocommunication Systems, Budapest University of Technology and Economics,
banky@mht.bme.hu

Reviewed

The study, hereby, introduces a method, which aims to improve the efficiency of the millimeterwave signal generation technique based upon the principles of active optical mode-locking. A complete experimental setup is to be shown, used for both the investigation and the dramatic suppression of the effects of phase noise content in the mode-locker radio frequency signal.

I. Introduction

The structure of the paper is as follows. The center topic of the paper (Section III. and IV.) is concerned with the effects of the phase noise on the performance parameters of the mode-locked laser sources, and also with a new way of denfense against them. The detailed buildup of the examined system appears in Section II. Here the basic principles of our work are to be mentioned, too. Section III. will point on the need for the high spectral purity mode-locker RF signal, with the help of the quality analysis of the output signal (as a function of the RF source's noise level). Chapter IV. will give a solution which makes significant (~10dB) improvements possible in the phase-purity of the RF oscillator's output signal. Section V. concludes the major results of the whole work that has been done and points on the opportunities hiding behind this given application (see Chapter II.).

Actively Mode-Locked Lasers (AMLL) are more and more frequently used in areas where generation of short, controlled optical pulses or remote production of microwave signals (on an optical way) are of interest. (These two cases are principally identical.) The AMLL structure basically takes an optical cavity, which has a resonance frequency in the microwave region. The optical cavity has two functionally distinguished sections inside: one behaves as an amplification medium, while the other as a modulator. The amplification media is responsible for the lasing effect, while the modulator performs amplitude or phase modulation over the optical signal inside (the modulation frequency should be equal to the optical cavity's resonance-frequency). This modulation procedure "locks together" the several longitudinal modes (rising with arbitrary phases). (In Section III. a short mathematical description can be found, which proves the need for this procedure.)

A strict condition for reaching the mode-locking is the tuning of the RF frequency to the cavity resonance. When these are mistuned, both *amplitude-* and *time-jitter* can be observed in the optical pulses (and also *widening* of the pulses in time domain representation). These facts give importance to the investigations on the effects of the noisy mode-locker source, (where the existing phase noise at a given moment can be handled as an instantaneous mistuning of the mode-locker source's frequency).

As the subject of our study we have chosen the active mode-lockink technique, since this is the basic construction, which can be found in applications used for generation of synchronized optical pulses*.

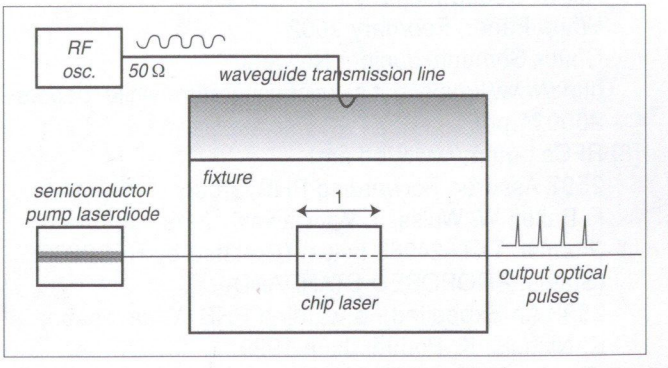
II. System Build-up

The system [1] that we have made our investigations on, was consisted of a usual active mode-locking setup: a lithium-niobate based solidstate laser was designed to have its longitudinal modes spaced by 5GHz from each other. These modes were locked together with the help of a dielectric resonator microwave oscillator which should have an oscillation frequency at: $f_{RF} = \nu_{q+1} - \nu_q = c/2nl$.

As it is shown in *Figure 1*, this RF signal penetrates inside the 13.5 mm long Nd:$LiNbO_3$ laser crystal through a waveguide transmission line. This transmission line also provide proper imedance matching between the laser crystal and the 50Ω output impedance of the RF oscillator.

From the optical side of the system we have to pump the chip laser by another laser source. This semiconductor pump laser (Nd:YVO_4) works at 814 nm with an output power level of 280 mW.

Figure 1 Experimental setup for the investigations on AMLLs



* passive *mode-locking: uses saturable absorber material; its structure is even more simple; the optical pulses of this case are not synchronized;* hybrid *mode-locking and* soliton *mode-locking: pulses are synchronized; both are using active mode-locking as a basic building block.*

By designing such a system the question can pop up whether the output signal quality is sensitive to the RF source's noise level or not (if yes, how much), so whether we do need to employ high spectral purity (and so more complex and more expensive) RF source or not. By introducing some computer calculation results, the next section will reveal that the answer is yes to both of the questions above.

III. Effects of Noisy RF Source

By locking the modes, we are able to harmonize the phase of the living longitudinal modes. When we reach to a state like that, we can observe periodic train of optical pulses at the output of our crystal laser, (else we would get random intensity fluctuation in the output optical signal. The output signal [2] in the time domain can be described as:

$$e(t) = \sum_n E_n e^{j[(\omega_0 + n\omega)t + \phi_n]} \tag{1}$$

where ω_n and Φ_n are the frequency and phase of the *n*th longitudinal mode, respectively, and

$$\omega = 2\Pi\left(v_q - v_{q-1}\right) = \frac{2\Pi}{T} = \frac{\Pi c}{l}$$

This signal (if we assume its periodicity in the time domain) should be identical to itself shifted by $T=1/\nu$:

$$e(t) = \sum_n E_n e^{j\left[(\omega_0 + n\omega)\left(t + \frac{2\Pi}{\omega}\right) + \phi_n\right]} = \tag{2}$$

$$= \sum_n E_n e^{j[(\omega_0 - n\omega)t + \phi_n]} e^{j\left[2\Pi\left(\frac{\omega_0}{\omega} + n\right)\right]} = e(t)$$

Eq.2. expressively represents that the condition for the output signal's periodicity is, the locked relation of the Φ_n phases. This is the point we should handle with special care if we want to get even pulses.

The realtive position of these phases (and so the locking range, the output optical power, the waveform, etc.) [3][4] will also be reactive to the quality of the millimeter-wave mode-locking source.

The noise dependence of these parameters was analysed through computer aided numerical calculations in *VPIsystems™* 's optical simulation environment. The simulation results are depicted in *Figure 2.*

We took two cases (one will appear as red, the other as yellow colored graphs in Figure 2). Basically, the only difference between them is the noise level of the mode locking RF source (in the „yellow" case it is 10dB lower). We used this noise level step between the two cases to test the overall effects of our new approach (introduced in Section IV) with which we could reach about 10dB phase-noise reduction.

Fig. 2/a shows the output optical signal in the time domain for the first 200nsec of the system operation. We should take note on the difference between the range of the amplitude fluctuations in the two different cases. The time and amplitude jitter improvement of the pulses, due to the phase-noise suppression, is remarkable (can be easily seen by comparing the eye diagrams *Fig.2/b1* and *Fig.2/b2).*

The output optical spectrum of the mode-locked laser source is also visualized (in *Fig.2/c1* and *Fig.2/c2*).

The difference between the optical spectrums will naturally appear in the microwave signal when detected *(Fig.2/d).*

And this is what we were curious about: now we can state that the noise power of the microwave oscillator that is responsible for the mode-locking effect, does strongly affect the quality of the output optical signal. This signal quality can have high importance when the mode-locked laser source is used for either narrow optical pulse train generation or optical generation of microwave frequencies.

These results made us to take special care of the phase purity of the RF oscillator's signal. In the next section we will introduce a new technique, which we used to suppress the phase noise level of the mode-locking signal.

IV. A New Method for Phase-Noise Suppression

According to the well-known concept [9] the phase noise is originating from the active elements' flicker *(1/f)* noise. The signal of this noise source modulates the oscillator, so perturbing the frequency and amplitude of its output signal. (With other words: through the active element's nonlinearities the baseband signal of the noise source will be upconverted to the neighborhood of the oscillation frequency.)

That's why we can avoid (or at least reduce the effects of) this mixing process when we reduce the level of the mixing product (or reduce the level of the modulating noise signal).

Since the signal of the noise source has a power spectral density function degrading by $1/f^{\alpha}$ $(\alpha \sim 1)$ [8], our problem of noise compensation should be a *low frequency* (LF) *problem* [5].

So, we can utilize LF signal controlling loop to reduce the instant level of the input noise. We just need good RF/LF isolation from the main oscillator circuit; and phase-sensitive LF feedback loop with signal conditioning. (Naturally, for the design and insertion of this, we have to exactly know the phase-shift, amplification and the location of the internal noise sources, etc. of the applied active element.)

Appropriately inserted lowfrequency feedback with -180° overall phase shift and 1 amplification, can so effectively compensate the actual signal of the flicker noise source.

We have tested this concept also in Agilent's ADS2001 RF simulation environment. The simulated system is shown in *Figure 3 (on the next page).*

As active element, we used Agilent's ATF-38143 low-noise pHEMT. We provided the simulator the noisy nonli-

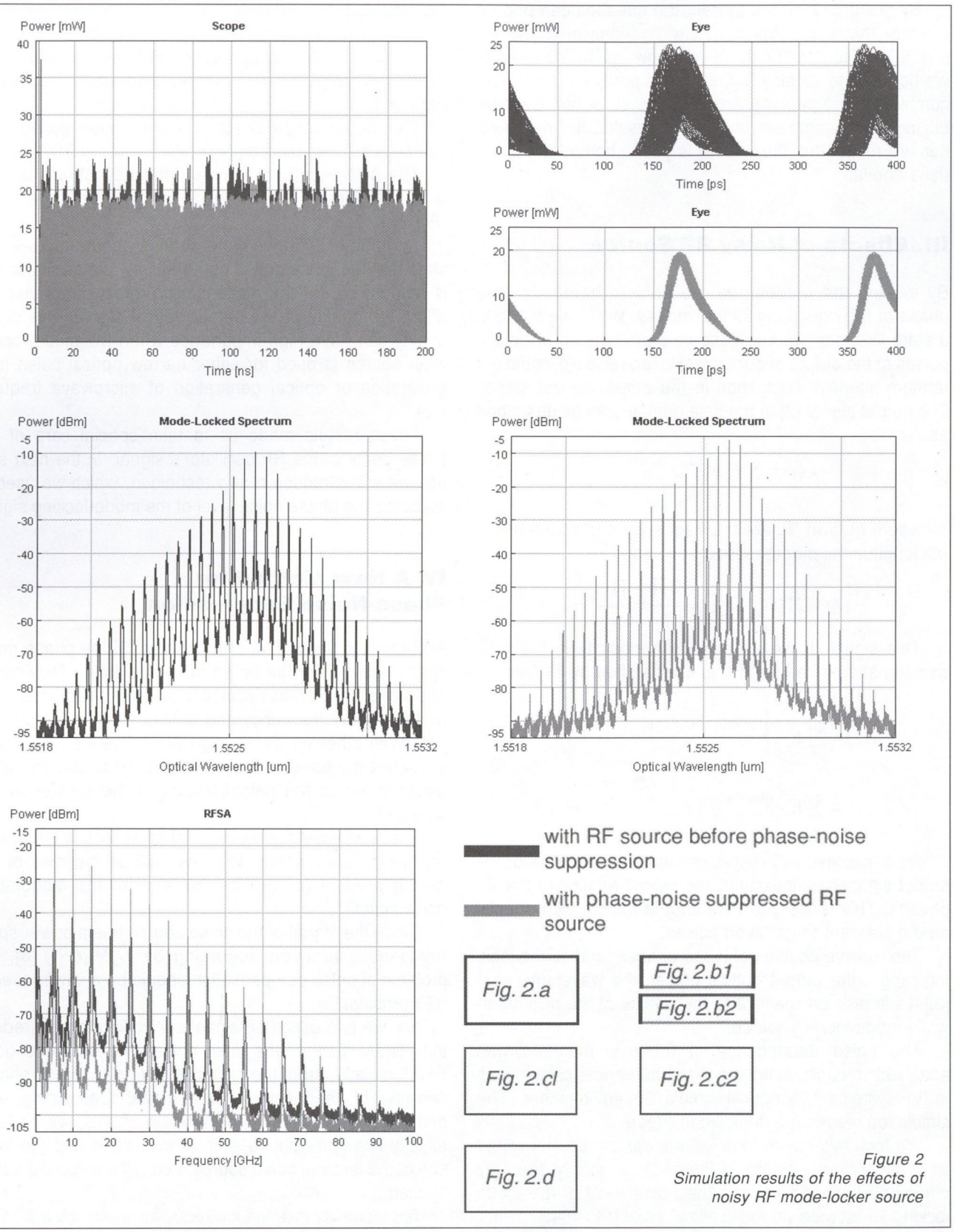


Figure 2
Simulation results of the effects of noisy RF mode-locker source

near Statz-model for that transistor. The upper loop is responsible for the satisfaction of the oscillation conditions and for the output power splitting. The dielectric resonator (since due to its good phase characteristics we have utilized Dielectric Resonator Oscillator) was modeled by a series attenuator bandpass-filter (BPF) pair.

For the oscillation, the phase condition is set by the length of the two 50Ω transmission lines. The low frequency noise components were driven to the lower loop by a circulator – low-pass filter (LPF) combination. This way the RF part (upper loop) wasn't disturbed by the LF part (lower loop).

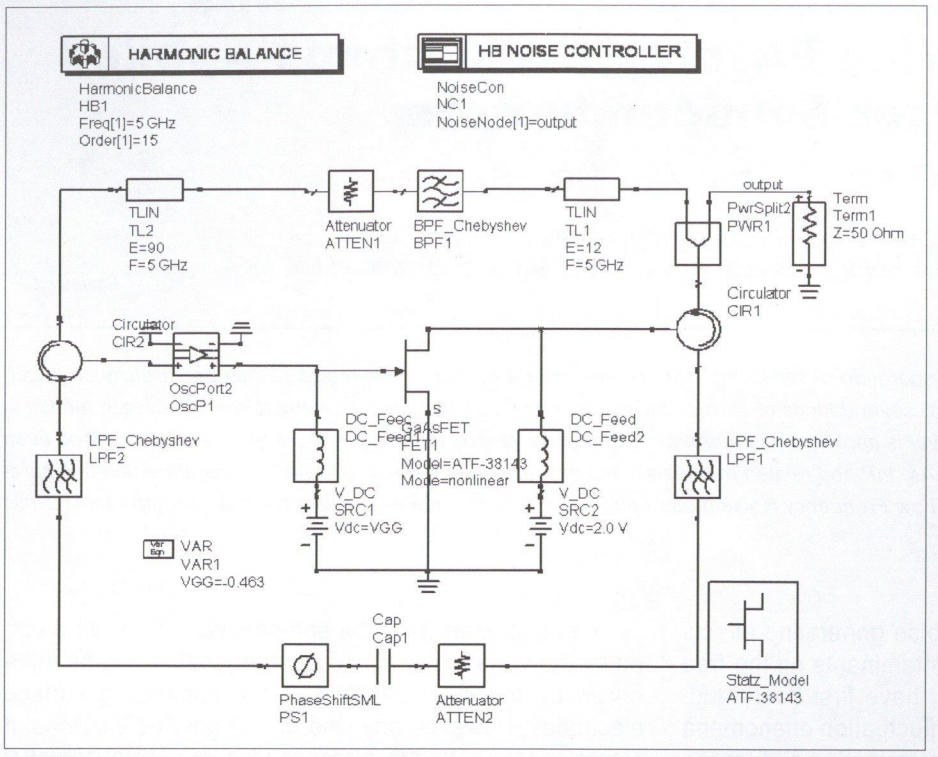


Figure 3 5GHz oscillator structure completed by the low-frequency noise suppression feedback

On *Figure 4.* Harmonic Balance based noise simulation results are shown:

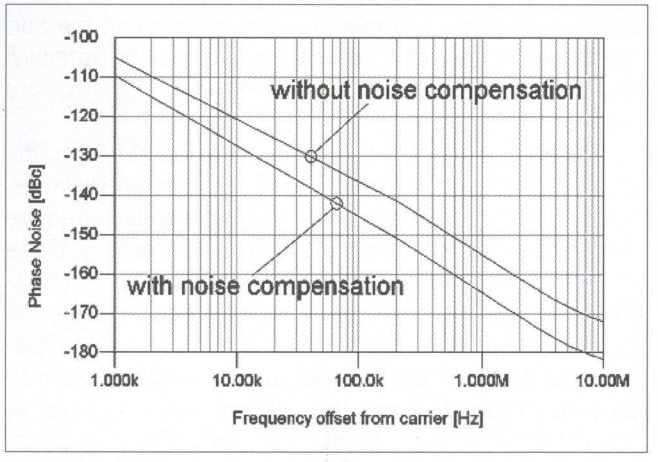


Figure 4 ADS SSB phase noise simulation results

As it can be read from Figure 4. the low-frequency compensation method gave significant phase noise reduction results (6-10dB reduction depending on the offset frequency from f_0).

V. Conclusion

The study above investigates a recently more and more popular, effective millimeter-wave signal generation technique, using the active mode-locking of solidstate laser sources. We have tested the effects of the mode-locking RF signal's phase noise on the overall performance of a given system which utilizes a Nd:$LiNbO_3$ crystal laser and a dielectric resonator microwave oscillator for mode-locking purpose. For overcoming the above mentioned negative effects we developed a new, easy-to-apply approach for the reduction of the phase-noise level at our millimeter-wave signal source.

With this approach we reached significant phase-noise suppression (about 10dB) and so we could enhance the quality (timing and shape of the pulses) of the signal at the optical system's output. This quality enhancement are shown by the graphs in Section III.

References

[1] W. D. Jemison, P. R.Herczfeld, W. Rosen, A. Vieira, A. Rosen, A. Paolella, A. Joshi: „Hybrid Fiberoptic-Millimeter Wave Links", IEEE Microwave Magazine Vol.1, Number 2, June 2000

[2] Amnon Yariv: „Optical Electronics in Modern Communications" Oxford University Press, Inc., 1997.

[3] P. Laporta, S. Longhi, M. Marchesi, S. Taccheo, O. Svelto: „2.5GHz and 5GHz Harmonic Mode-Locking of a Diode-Pumped Bulk Erbium-Ytterbium Glass Laser at 1.5 Microns" IEEE Photonics Tech. Lett., Vol.7., No.2. Feb. 1995.

[4] D. Novak, D. Y. Kim, Hai-Feng Liu, Z. Ahmed, Y. Ogawa: „Locking Range of a Hybrid Mode-Locked Monolithic DBR Semiconductor Laser at Millimeter-Wave Frequencies" IEEE Microw. And Guided Wave Lett., Vol.6., No.9., Sept 1996

[5] M. Prigent, J. Obregon, "Phase Noise Reduction in FET Oscillators by Low-Frequency Loading and Feedback Ciruitry Optimization" IEEE Trans. Microwave Theory and Tech., Vol. MTT-35, No. 3, pp. 349-352, March 1987.

[6] T. Berceli, "Nonlinear Active Microwave Circuits" Akadémiai Kiadó, Budapest, pp. 31-68, 1987.

[7] M. Regis, O.Llopis, J. Graffeuil, "Nonlinear Model-ing and Design of Bipolar Transistors Ultra-Low Phase-Noise Dielectric-Resonator Oscillators" IEEE Trans. Vol. MTT-46, no.10, pp.1589-1593, Oct. 1998.

[8] D. B. Leeson, "A simple Model of Feedback Oscillator Noise Spectrum" Proc. IEEE, Vol. 54, pp. 329-330, 1966.

[9] A. Demir, A. Mehrotra, J. Roychowdhury, "Phase Noise in Oscillators: A Unifying Theory and Numerical Methods for Characterisation" Proceedings of the 35th Annual Conference on Design Automation Conference, pp. 26-31, 1998.

Low-frequency Noise Measurements for Investigating Passivation Methods Applied for Semiconductors

PÉTER GOTTWALD, *PhD* AND BÉLA SZENTPÁLI, *PhD*

gottwald@goliat.eik.bme.hu, szentpa@mfa.kfki.hu
Research Institute for Technical Physics and Material Science MTA MFA

Reviewed

To achieve high stability and low noise operation of semiconductor devices, their electrical active region should be carefully protected from any external influence. To this end, several kinds of surface passivation methods are used whereby a special surface modification or coating by a suitable dielectric layer is applied. The parameters of the passivation technology should be carefully optimized even for compound semiconductors (e.g. GaAs, InP and related materials), that are rather sensitive to elevated temperature, ion bombardment and surface contaminations. The Low Frequency Noise measurement (LFN) technique has proved to be an effective tool for this optimization.

In using the LFN technique, the noise generating effects of several materials, applied as contaminants on the free surface of GaAs, InP and InGaAs, have first been studied. It was observed that electrical fluctuation phenomena were characteristic both for the contaminating and for the semiconductor materials.

In the course of further investigations, the LFN technique was applied to characterize and optimize the commonly used passivation technologies applied for the above semiconductors. We now present new results in addition to those included in our previous review [1]. These apply to the passivation of InP by a Photo Enhanced Vapour Deposited (PVD) SiO_2 or Si_3N_4 layer on one side. On the other side, an anomalous additional low frequency noise was observed for GaAs as a particular effect of unknown origin. This spectacular effect was observed during the passivation by PVD SiO_2 and during the passivation by an HF sputtered Si_3N_4 layer.

1. Introduction

The low-cost, high reliability solutions in modern electronics are based on Integrated Solid-State Devices. Apart from the different technologies used in manufacturing, these devices are all realized by a relatively thin active layer grown on a thicker semiconductor substrate. The active layer thickness falls into the micrometer range, while the average area of a single circuit element is some 10 μm^2 or less.

The number of the integrated circuit elements varies in a very large range, depending on the type of the circuit. However, it is characteristic that the electrical signal levels (voltage and also power) connected by a single circuit element become lower as the geometrical dimensions are decreased. However, to meet a general requirement, the operation speed can be increased also by decreasing the geometrical dimensions. As a consequence, in such circuits low potential variations take place closely under the surface of the semiconductor material.

In normal operation, the potential variations are created by the variation of electrical charges that can be influenced by the electrical field of the conducting surface electrodes. However, any kind of charges, as e.g. ions, if closely enough to the surface, also generate potential changes in the active semiconductor region. Randomly moving, thermally vibrating surface charges induce undesirable charge carrier modulations into the active layer. Thus, slower drift of the electrical parameters or electrical fluctuation phenomena (noise) may result. Therefore, it is extremely important to minimize the influence of the surroundings on the semiconductor active region. Normally, suitable separating dielectric layers (e.g. SiO_2, Si_3N_4) are used for this purpose.

Because of the high quality native oxide of silicon, the passivation of silicon based devices is simpler. However, the microwave and optoelectronic applications require other ("direct band") semiconductor materials, e.g. GaAs and InP or any of the related compounds. Apart from silicon, they have no high quality native oxide, and the different materials have rather different surface state behaviour. Additionally, a lot of special surface modifications may easily be caused when any coating dielectric layer is grown on the surface. Electrically, the surface modifications appear as additional allowed electron states not far from the middle of the forbidden gap, able to capture and emit electrons causing thus charge carrier fluctuation. This is called subsequently as generation-recombination (G-R) process, and the corresponding energy states will be designated as G-R levels or deep levels.

Due to the above facts, the passivation technology should always be carefully optimized for each particular case.

Both for the previous [1] and also for the current investigations and technology optimizations, the same test structure of planar resistors has been used, designed to have an enhanced surface sensitivity.

In our planar test resistors, a remarkable part of the fluctuation is due to the resistor surface [2], [3], [4], beside the resistance fluctuation of bulk effect origin.

The surface connected fluctuations can be separated as they do not follow the classical $1/f$ like spectral intensity that is characteristic for the bulk effect part [5]. This allows us to detect and analyze the surface effects caused by the passivation technology.

Our previous review [1] will now be completed by the following new results:

A) Passivation of InP by SiO_2 and Si_3N_4 layers grown by Photo-Enhanced Chemical Vapour Deposition (PVD).

B) Resonant type additional fluctuation, as a particular effect caused by an unknown flaw in the passivation technology by GaAs, observed very rarely for GaAs by PVD SiO_2 and high-frequency sputtered Si_3N_4 passivating layers.

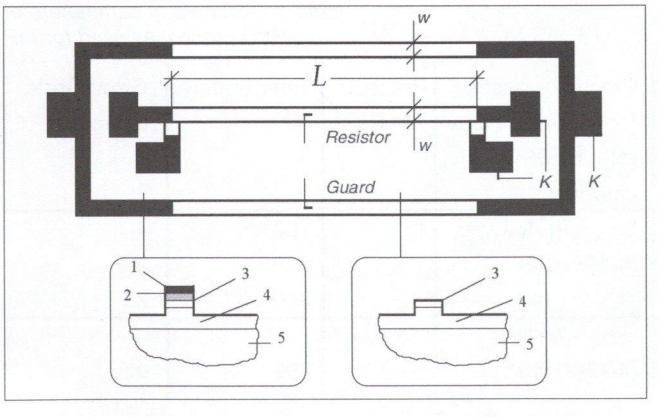
2. The Test Chip

The layers of the test chip mentioned earlier are shown in *Figure 1*, as applied for the investigations treated in part **A**. For the investigations described in part B, the test chip is different: the highly doped layer marked by **2** is omitted, while the other layers are GaAs (and not InP).

The main part of the test chip is the strip-shaped resistor body which is a thin n-type layer held by a thicker semi-insulating substrate. Both ends of the resistor have two contacts realizing thus the well known four-point (force/sense) measurement technique.

Around the resistor strip, there is a closed pattern of similar structure that will be called subsequently guard ring. This has special importance in the undesirable case in which the uncovered free surface of the substrate around the resistor strip starts to conduct weakly. It will be shown later that some passivation technologies are able to produce such an effect. In this case, the guard ring helps to determine the surface conductance value.

Figure 1 Structure of the test chip
The strip width w is 40 μm, the length to width ratio is 20. (Length L means the distance between the resistive contacts.) The resistive metallizations are marked by black areas denoted by "K". The layers of the structure are as follows:
***1** - resistive metallization, **2** - n-type, highly doped, 330 nm thick InGaAs layer which improves the contact resistance,*
***3** - n-type, 280 nm thick active InP layer with a doping level of 10^{17} cm^{-3} , **4** - undoped, 340 nm thick InP buffer layer,*
***5** – semi-insulating InP substrate.*



To realize the test chips, standard MeSFET technology has been applied using photolithography, selective chemical etching and the lift-off technology for the resistive contacts. The technology of these contacts was optimized for minimum specific contact resistance in each case. The whole surface of the chip was thereafter covered by the passivation layer in which windows were etched over the contacts.

3. The Low Frequency Noise Measurement Technique

According to theory, electrical resistors inherently exhibit low frequency fluctuations. The mean square value of the resistance fluctuation in a bandwidth of 1 Hz, around a given frequency f, is called subsequently as spectral density, and will be denoted by S_R. For ideal cases, following relation was derived by Hooge [5]:

$$S_R / R^2 = \alpha / fN,$$

where R is the resistance value, N means the total number of charge carriers in the resistor, and α is the so-called Hooge parameter that was thought earlier to be a universal constant. However, it has subsequently been proved that its value depends on the material and also on the geometrical dimensions of the sample, and varies in the range from 10^{-6} to 10^{-2}. It should be noted that the resistance fluctuation due to the above equation concerns the number and/or mobility fluctuation of the charge carriers that is due to a bulk effect, and does not cover other fluctuations of surface origin [5]. Therefore, beside the $1/f$ -like fluctuation of bulk origin, the measurable spectral intensity has practically always an additional component that is generated, in most cases, by several surface processes.

If the surface related component is caused by some generation-recombination (G-R) process, due to different G-R levels, then the additional spectral intensity component will be the sum of the corresponding so-called Lorenzian components. Such a component has following frequency dependence:

$$\Delta S(f) = S_0 / [1 + (2\pi\tau f)].$$

The two new parameters S_0 and τ are the low-frequency "plateau" value of the additional spectral intensity and the time constant of the corresponding G-R process, respectively. Both new parameters are typically temperature dependent. By analyzing these temperature-dependencies, the concentration and the energy of the corresponding deep level can be determined [6].

It is important to note that other kinds of additional fluctuations may also exist. They are generated e.g. by particular free surface processes, or even by the passivation process, if not optimized or processed correctly. The latter is illustrated in part B.

In the practical measurement set-up, a moderate DC inspection current I_0 converts the resistance fluctuation into a noise voltage. The noise voltage was measured in the frequency range of 1.6 Hz to 20 kHz by a Brüel &

Kjaer digital frequency analyzer. The temperature was increased in the range from 0°C to 80°C in 10°C steps.

Assuming that the square values of both the noise and the DC voltages on the sample are proportional to the square value of the DC inspection current, all results have been transformed thereafter to correspond to a DC voltage of 1 V on the sample. This ensures that the different results are directly comparable.

The essence of the investigation is the separation and analysis of the temperature dependent components of the measured noise spectrum whereby a rigorous curve-fitting computer aided method was used. It was seen that a very high accuracy has already been reached by assuming one $1/f$ and two dominant G-R noise components having a Lorenzian spectrum. Thus, the additional noise can be separated as the sum of the Lorenzian components.

4. Results

4.1 Passivation of InP by PVD grown SiO_2 and Si_3N_4 layers

To avoid thermal decomposition of the InP through elevated surface temperature during layer growth, only low temperature techniques may be used. Such a technology is the *Chemical Vapour Deposition* (CVD) whereby the layer is formed on the surface from its ionic components. To produce the necessary ionic mixture (plasma) from the precursor gases, thermal energy is not allowed in this case. Therefore, to supply the necessary decomposition energy, instead of heat, either a high frequency electric field or an effective interaction with high-energy photons may be applied. The first technology is the *Plasma Enhanced* CVD (briefly PECVD), while the second one is the *Photo Enhanced* CVD (briefly PVD).

Two main PVD versions exit. In the first one (called as excimer-PVD), a high-energy excimer UV light-source is applied, and the photons interact directly with the precursor gases causing thus photolysis. Otherwise, in the second version, an indirect photolysis is generated. For this reason, the vapour of a suitable material (mostly Hg) is mixed to the precursor gases, and the atoms of the additional material will then be excited by an UV light of suitable wavelength (for Hg: λ = 253.7 nm). Then the excited atoms interact with the precursor gases. This classical method is called Hg-sensitized PVD.

Although all of the above methods allow a relatively low layer growth temperature (regularly under 300°C), they have disadvantages too. Thus, the PECVD has proved to be not the best choice for InP because of the damaging ion bombardment, and partially because of the higher deposition temperature usually required (commonly higher than 200°C). Otherwise, the PVD technique may result in a better passivation quality, especially for InP. However, using Hg for sensitizing, Hg may become incorporated into the layer, thus deteriorating its quality. Since the incorporation shows a remarkable temperature-dependence, a strict optimization of the technology is always necessary. Lastly, if excimer PVD is used, the generated ozone and the direct bombardment by high-energy photons act disadvantageously.

Our investigations concern the passivation layers deposited by the Hg-sensitized PVD.

Before the layer growth procedure, rigorous chemical surface cleaning has been carried out. *Table 1.* summarizes the most important layer-growth parameters. (Instead ofr Si_3N_4, SiN_x will subsequently be used, since in practice, there is always a small deviation from the ideal stochiometry.)

As can be seen from the table, SiN_x layers were grown only at a medium temperature (200°C). Namely, at lower temperature (150°C), the adherence becomes poorer, and the microscopic layer-structures show considerable non-uniformity. Otherwise, if the growth temperature is higher (e.g. 300°C) then the growth rate is extremely low (~0.5 nm/min) [9].

Similarly, the quality of the SiO_2 layers grown at lower temperature (150°C) has not been satisfactory. Moreover, the surface of the semi-insulating InP substrate around the resistor strip seems to conduct weakly. The surface conduction scatters from sample to sample in a range from 20 to 200 kOhm/❑. (This dimension means the resistance of a square-like part of a thin conducting layer between the opposite edges, and is called sheet or surface resistance.)

Nevertheless, when surface conductance develops, the layer is not satisfactory for a high quality passivation purpose. Therefore, only the SiO_2 layers grown at higher temperatures (200 and 300°C), and by SiN_x the layers deposited at 200°C has been treated in detail.

4.1.1 Comparison of the LFN spectra

Passivation by SiO_2 – At 200°C deposition temperature, the layer was more uniform. Still, for two of the five investigated samples, some surface conductance was indicated due to a sheet resistance well above 1 MOhm/❑. However, when the deposition temperature was raised to 300°C, the surface conduction completely disappeared.

Table 1 Technology parameters of PVD processes used for InP

Layer	Thickness nm	Precursor gases Pressure and flow-rate at 10^5 Pa, cm^3/min	Pressure 10^2 Pa	Temperature °C	Growth-rate nm/min
SiO_2	150	SiH_4 +Ar+N_2O (4/196/120)	1	150 200 300	4 2.5 2
SiN_x	100	SiH_4 +Ar+NH_4 (4/196/100)	1	200	0.5

Figure 2 shows the characteristic temperature dependent LFN spectra for deposition temperatures 200 and 300°C, respectively.

As an important phenomenon, for the deposition temperature of 200°C, a tendentious noise enhancement was observed in the whole frequency range, as by the measurement the sample temperature was raised to 70°C or above. However, re-measuring this enhancement was regularly not possible since in the course of the very first measurement, the moderate heat treatment cancels this effect. Otherwise, the 300°C deposition temperature results in a smooth and moderate noise in the whole frequency and temperature range.

These results may be explained by the fact that the PVD did not practically cause any surface damage. However, Hg can react with the N_2O that is used by the deposition. As a consequence, HgO is generated, infiltrates into the deposited layer, and may cause problems. This occurs especially at lower deposition temperatures. Namely, when Hg diffuses into the InP, it may act as a shallow acceptor [8], and can cause surface conduction. Otherwise, when it remains in the interface region close to the surface, additional noise may easily be generated.

Other investigations – using RBS analysis – have shown [9] that at higher deposition temperatures, the HgO incorporation rapidly diminishes. Accordingly, at the 300°C deposition temperature, the surface conduction disappears, and the LFN results improve. Nevertheless, from a further increase of the deposition temperature no further improvement may be expected, since the phosphorous loss may really start to increase, and also the adhesion of the deposited layer degrades. Thus, investigating the homogeneity, reproducibility, surface conduction and noise, the optimum result was found to be around 300°C.

Figure 2 Temperature dependence of the LFN of PVD SiO_2 passivated InP samples. The temperature step is 10°C. The deposition temperatures are 200 and 300°C, respectively. For the 200°C deposition temperature, the measurements have been carried out in 10° intervals in the 0 to 80°C temperature range. At 70 and 80°C, an additional noise enhancement is started. For the 300°C deposition temperature, no such enhancement has been observed at 70 and 80°C.

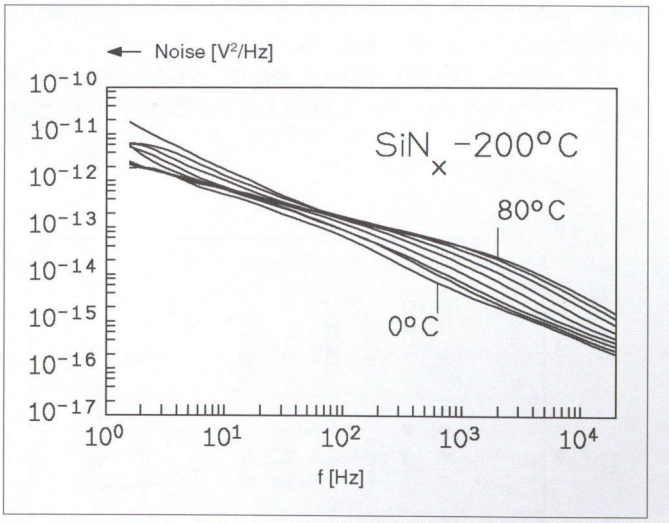


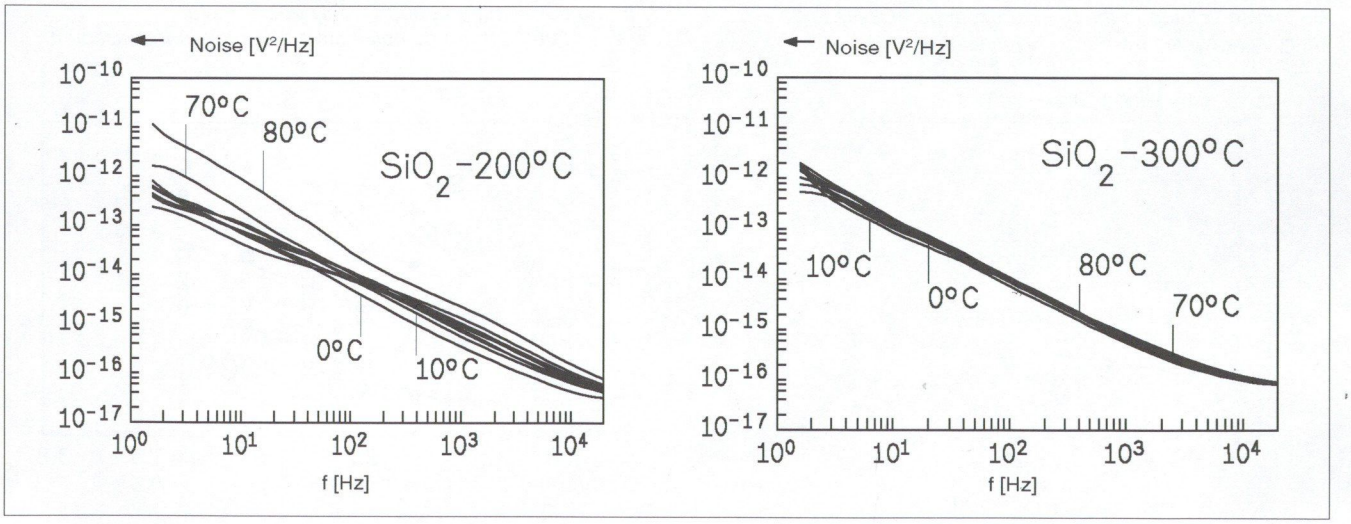
Figure 3 Temperature dependence of the LFN for PVD SiN_x passivated InP samples. The deposition temperature is 200°C. The temperature step is 10°C. The typical dependence of the curves on. temperature is due to one significant G-R noise component

Passivation by SiN_x – The alternative solution of the passivation is the PVD SiN_x deposition. In this case, SiH_4 és NH_4 are used as precursor gases. *Figure 3* shows the temperature dependent noise spectra for a sample passivation by PVD SiN_x, deposited at 200°C substrate temperature. In comparison with the results obtained for SiO_2, a similar but remarkably stronger temperature dependence could be observed.

Comparing this result with the LFN obtained for SiO_2, higher noise and stronger temperature dependence are obviously seen. As a further complication, in this case a week surface conductance is also present, which corresponds to a surface resistance value of about 0,5 to 1 MOhm/❑.

4.1.2 The LFN components

The above direct comparison of the LFN spectra provided useful information. However, better characterization can be achieved by the comparison and analysis of the individual LFN components. For this reason, the spectra



were split into a $1/f$ noise component and two G-R noise components, as explained in Sec. 3.

$1/f$ noise – The $1/f$ noise is represented in the form of K_0/f. *Figure 4* shows the temperature dependence of the coefficient K_0.

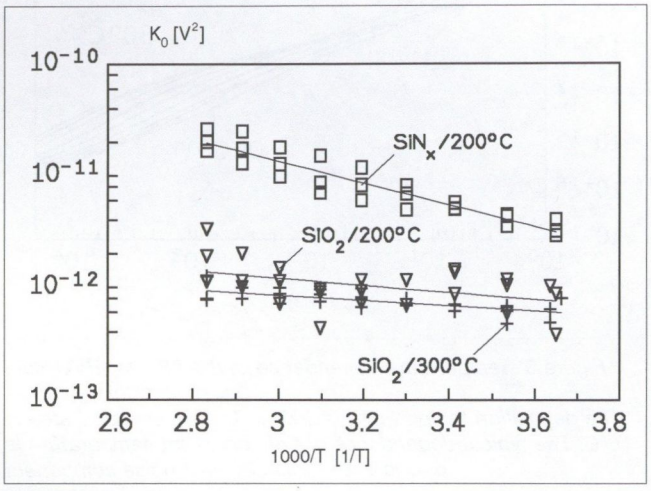
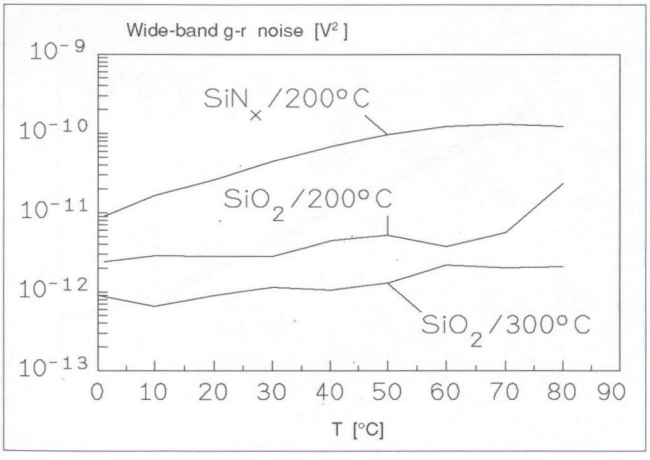


Figure 4 Temperature dependence of the coefficient K_0 of the $1/f$ component.

As shown by Figure 4, the $1/f$ noise is higher and the temperature dependence is stronger for SiN_x passivation than for SiO_2 passivation The much higher value of the $1/f$ noise may be considered as a result of the surface conductance. We should note that the physical origin of the $1/f$ shape of this additional noise is still not clarified. Anyway, in other cases, we have already measured $1/f$ additional noise for InP [10]. Consequently, the superiority of SiO_2 passivation over SiN_x is clear, but the better result requires a 300°C deposition temperature.

G-R noise components – From the practical aspect, the G-R noise intensity is a very sensitive indication of any structural imperfection of a semiconductor material. Therefore, the sum of the separated G-R noise spectra has been integrated in a frequency range of 2 Hz to 20 kHz yielding the variance (mean-square value) of the wideband G-R noise voltage.

Figure 5 Temperature dependence of the wideband G-R noise voltage in the frequency range of 2 Hz to 20 kHz. Calculated curves from the LFN measurement results.



In *Figure 5,* the temperature dependence of the total G-R noise voltage variance, as an average of three characteristic samples, is plotted for each of the investigated deposition technologies. Figure 5 shows that SiO_2 passivation results in a much lower wideband G-R noise. Therefore, it is a more convenient solution for passivation than the use of SiN_x. Additionally, for SiO_2, the superiority of the higher deposition temperature is also clear.

4.1.3 - Deep levels

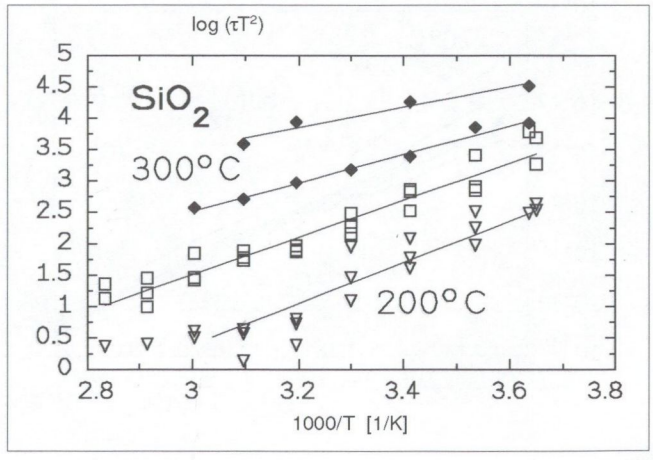
As already defined, the deep levels are *allowed* energy states nearly in the middle of the *forbidden* gap of the semiconductor material. These are mostly caused by specific crystal defects.

Technology-processes acting on the semiconductor surface (as e.g. the passivation too) may likely generate deep levels, closely under the surface. If just the passivation-caused deep levels have to be investigated, and hence the structure must not be further modified or patterned according to a measurement method, only the LFN technique can be applied. (It should be noted that the well-known DLTS method does not meet this requirement.) Thus, it is of great importance that the deep level parameters can be determined from the LFN measurement results.

The basis for this is the behaviour of the deep levels that randomly – but statistically in a well-defined manner – capture and emit charge carriers, generating thus the Lorenzian LFN components. From the temperature dependence of the parameter τ of a Lorenzian component, both the energy level and the capture cross-section can be determined. For this reason, the temperature dependence should be plotted (using the absolute temperature T) as $\log(\tau T^2)$ vs. $1000/T$. Theoretically, this is a straight line called Arrhenius plot. As known [6], from the slope of the plot the deep level energy, and from the value at $1000/T \rightarrow 0$, the capture cross-section can be determined.

Figure 6 shows the Arrhenius plots for passivation with SiO_2. Since the G-R noise for the 300°C deposition temperature is rather low, the uncertainty of the determined results is higher.

Figure 6 Arrhenius plots for the PVD - SiO_2 passivation, when two G-R noise components are split. For the passivation at 300°C, only one sample, and for 200°C, three samples have been taken into account.



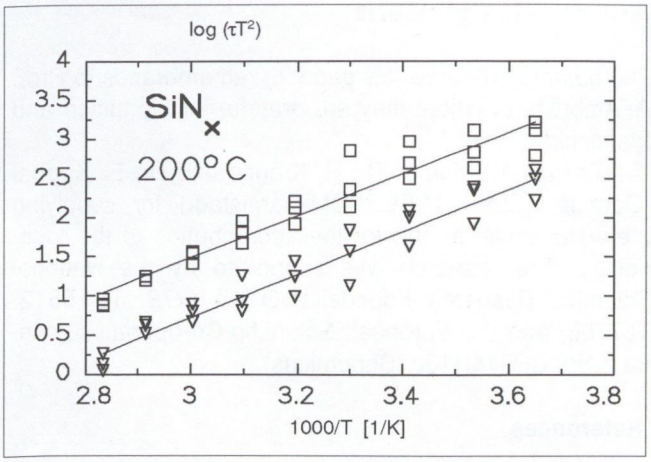


Figure 7 Arrhenius plots for the PVD- SiN_x passivation at 200°C, when two G-R noise components are split. Three samples have been taken into account.

Regarding this fact, the importance of the detailed analysis is less in this case. However, the obtained (somewhat uncertain) results are: two deep level energies at about 320 and 450 meV with the corresponding capture cross-sections of about 10^{-19} and $10^{-16}cm^2$. Comparing these with the results published elsewhere, there are two known levels in InP, namely at 350 meV [10] and at 450 meV [11]. In spite of the uncertainty, the level at 320 meV may physically correspond to the known level at 350 meV, due to the P vacancies in the lattice. This is a reminder that at 300°C, the P out-diffusion starts, and the higher deposition temperature is inconvenient.

For the 200°C deposition temperature, three samples are involved in the investigation, yielding thus more reliable results. The lower plot in the Figure gives a level at 650 meV with a capture cross-section of $10^{-11}cm^2$, while for the upper plot, the corresponding values are 590 meV and $10^{-13}cm^2$. These are practically identical with the levels at 580 and 680 meV, published in [10].

As the SiN_x passivation gives a considerably higher G-R noise than the SiO_2 passivation, the results are more accurate.

Figure 7 shows the Arrhenius plots, calculated by averaging the results of three characteristic samples. The corresponding two levels are nearly at the same energy (550 meV), but the capture cross-sections are different (10^{-13} and $10^{-14}cm^2$). We note that there is a known level in InP at 560 meV [11].

4.2. Resonant-like additional noise measured on passivated GaAs samples

In the previous investigations, the LFN methods were used to characterize and optimize the passivation technology. There are also cases in which an unknown flaw in the passivation technology can be detected from an unusual alteration of the LFN results.

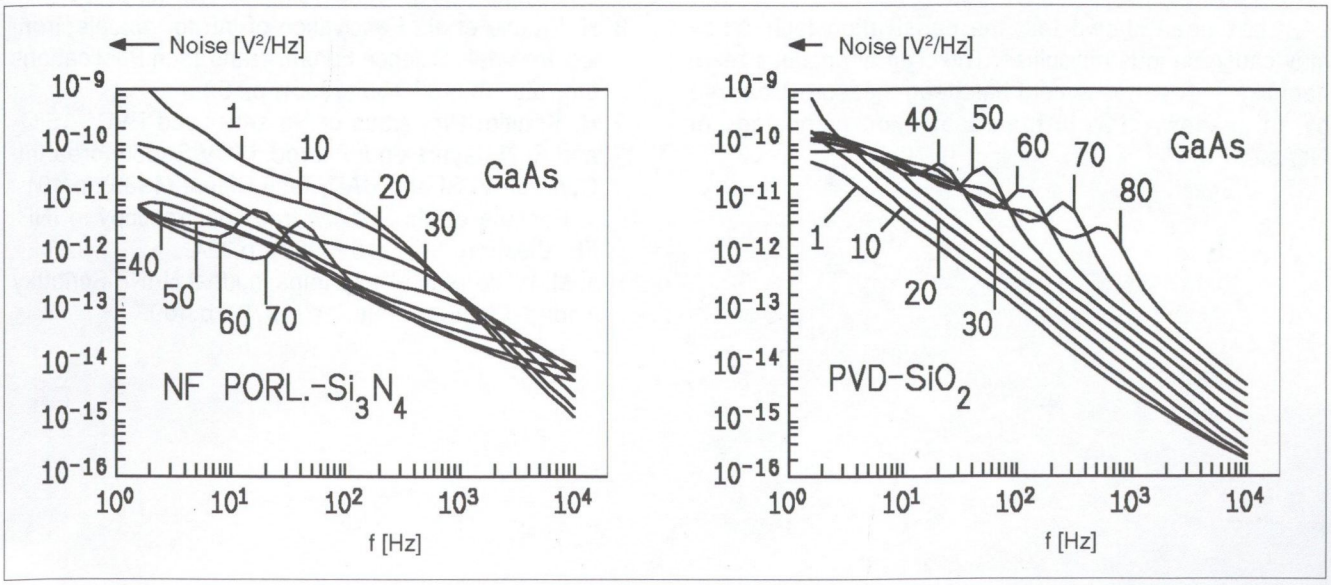
Two very surprising results have been measured in our practice, showing a resonant-like anomalous additional noise on the LFN spectra [7]. Both were measured for GaAs in the following cases:

- Passivation by RF-sputtered Si_3N_4 (target-temperature: 30°C, N_2 pressure: $5.2x10^{-3}$ mbar, acceleration voltage: 1.4 kV, process-time 10 min.)
- Hg sensitised photo-CVD of SiO_2 layer [3]. (target-temperature: 150°C, thickness: 100 nm)

The measurement results are plotted in *Figure 8*. The characteristic features of the anomalous additional noise have been found as follows:

- If the samples were illuminated by visible light, the resonant enhancement has been hidden in both cases.
- The frequency f_r of the peak enhancement is decreased if the electric field strength in the sample is increased. This decrease is only slight for the samples

Figure 8 Anomalous peak enhancement in the LFN spectra of surface passivated GaAs samples in the temperature range of 10 to 80°C, measured just after processing. Results for passivation by sputtered Si_3N_4 and for passivation by PVD of SiO_2 layer are given. It is seen that the f_r resonant frequency is temperature dependent.



passivated by Si_3N_4, but considerably greater for the passivation by photo-CVD of SiO_2. Below a DC field strength of about 60 V/cm in the sample, the anomalous enhancement vanished in both cases.

- A long-term drift of the shape of the spectra has been observed in a 3-years interval, while the frequency of the peak enhancement shifted toward higher frequencies.

Although the physical explanation of the anomalous effects is still not clear due to the temperature dependence of f_r, an Arrhenius plot of (τT^2) has been prepared, using the relation $\tau = (2\pi f_r)^{-1}$. Calculation of the activation energy resulted in a well-defined level at about 600 meV for the passivation by PVD of SiO_2, just after processing. On the other hand, for the samples passivated by Si_3N_4, a level at about 880 meV has been found after processing, but 3 years later a lower value of 600 meV has been determined.

The above anomalous effects show that even for very different passivation technologies, serious problems can occur leading to difficulties and degradation of the electrical behaviour. It was also noted that the LFN measurement technique is able to detect such problems.

Summary

By using suitably designed test chips, it has been shown that the LFN measurement technique is an effective tool for investigating and optimizing the passivation technologies.

It was shown that for InP, the PVD SiO_2 passivation gives better results than the passivation by Si_2N_4. For SiO_2, the best result has been achieved with the 300°C deposition temperature. In this case, no surface conduction is developed, the $1/f$ noise is low and approaches the bulk value known for InP. Additionally, the wideband G-R noise is also the lowest. Analyzing the temperature dependence of the G-R noise, deep level analysis has been carried out too.

It has been shown that the passivation technology may cause serious difficulties. Thus, an anomalous resonant-like LFN enhancement has been measured, caused by an unknown flaw in the passivation technology of GaAs.

Acknowledgement

The authors prepared this paper in remembrance to Prof. A. Ambrózy to whom they are grateful for his tuition and friendship.

Thanks are due to Dr. H. Kräutle and Dr. E. Kuphal (German TELEKOM/BERKOM-Darmstadt) for supplying the wafer material and for their contribution to the technology. The research was supported by the National Scientific Research Foundation/OTKA (773, TO15612, T37706) and the European Scientific Co-operation (contract No. CP94/01180 /Copernicus).

References

1. P. Gottwald: Low-frequency noise measurements as a diagnostic tool in the semiconductor technology. Journal on Communications, 46 (1995), No. 2, p.3.
2. A. Ambrózy et al.: Surface effects on the low frequency noise of thin GaAs layers. Proc. Noise in Physical Systems and $1/f$ Fluctuation, ICNF'91, Kyoto (1991), p.23.
3. P. Gottwald et al.: Comparison of Photo- and Plasma-Assisted Passivating Process Effects on GaAs Devices by Means of Low-Frequency Noise Measurements. Solid-St. Electronics, 38 (1995), p.413.
4. P. Gottwald et al.: Damage characterisation of InP after Reactive Ion Etching using the low-frequency noise measurement technique. Solid-St. Electronics, 41 (1997) p.539.
5. F. N. Hooge: $1/f$ is no surface effect, Phys. Lett. A, A-29 (1969), p.139.
6. L. Loreck et al.: Deep level analysis in (AlGa) As-GaAs 2-D electron gas devices by means of low-frequency noise measurements. IEEE Electron Dev. Lett., EDL-5 (1984), p.9.
7. P. Gottwald et al.: Anomalous additional low-frequency noise of surface origin generated in thin GaAs and InP layers. Proc. of the 1st Int. Conf. on Unsolved Problems of Noise, (Ed.: Doering Ch. R., Kiss L. B., Shlesinger M. F.), World Scientific (1997), p.122.
8. H. Kräutle et al.: Passivation of InP for optoelectronics. Materials Science Forum, Trans Tech Publications Ltd.: Zürich, 185-188 (1995), p.199.
9. H. Kräutle: Properties of Hg sensitized PVD – SiO_2 and Si_3N_4 layers on InP. Proc. IVCV Seoul Korea Int. Conf. on VLSI and CAD, Tech Digest (1989), p.401.
10. J. Bonnafe et al.: Shallow trap spectroscopy in INP-FE. Electron. Lett., 16 (1980), p.324.
11. A. M. White et al.: Deep traps in ideal N-INP Schottky diodes. Electron. Lett., 14 (1978), p.409.

Watermarked Image Denoising by Impulse Detection Based Approaches

RASTISLAV LUKAC

Slovak Image Processing Center, Dobsina, Slovak Republic
e-mail: lukacr@ieee.org

The noise introduced to image information during the image transmission over noisy channel affects not only visual data, however, it also degrades hidden data such as the embedded digital watermark. Thus, the digital watermark extracted from noisy watermarked images should provide the inefficient quality and prohibit worse authentication, labelling, monitoring or protection often necessary in multimedia and Internet applications. Since, the image filtering is stated often as the balance between the noise suppression and signal-detail preservation and the used filter introduces an estimation error, the watermarked image denoising often results in the significant watermark corruption. This paper focuses on the impulse detection strategy for watermarked images corrupted by impulse noise so that the well-known median filter affects the noisy information only, whereas the desired image features are invariant to a filtering operation. The proposed strategy allows to preserve the embedded watermark, significantly.

1. Introduction

The development of multimedia and Internet technologies is accompanied with the request for copyright protection and authentication of original digital contents [2],[6],[18], [19].

The reason is that the digital copies are identical to the original. Thus, the digital watermarking based on the embedding information data directly into digital contents with an imperceptible form for human audio/visual system finds a great application.

In general, the embedded watermark does not spoil the quality of the original information and should not be perceptible. It must be difficult for an attacker to remove it and should be robust to signal processing including various filtering algorithms, image enhancement methods, etc [19].

In many situations, the noise affects the transmitted data especially through bit errors [1] defined by

$$ {}^{*}k_{i,j}^{m} = \begin{cases} k_{i,j}^{m} & \text{with probability } 1-p_v \\ 1-k_{i,j}^{m} & \text{with probability } p_v \end{cases} \tag{1} $$

where p_v is the bit change probability, $k_{i,j}^{m}$ and ${}^{*}k_{i,j}^{m}$ are binary values {0,1} of B-bit watermarked sample $o_{i,j}''$ and noisy watermarked sample $x_{i,j}$ given by

$$ o_{i,j} = k_{i,j}^{1} 2^{B-1} + k_{i,j}^{2} 2^{B-2} + \ldots + k_{i,j}^{B-1} 2 + k_{i,j}^{B-1} \tag{2} $$

$$ x_{i,j} = {}^{*}k_{i,j}^{1} 2^{B-1} + {}^{*}k_{i,j}^{2} 2^{B-2} + \ldots + {}^{*}k_{i,j}^{B-1} 2 + {}^{*}k_{i,j}^{B} \tag{3} $$

The simplified definition of the bit errors results in frequently used mathematical formula for the random valued impulse noise [1],[3],[10] expressed as

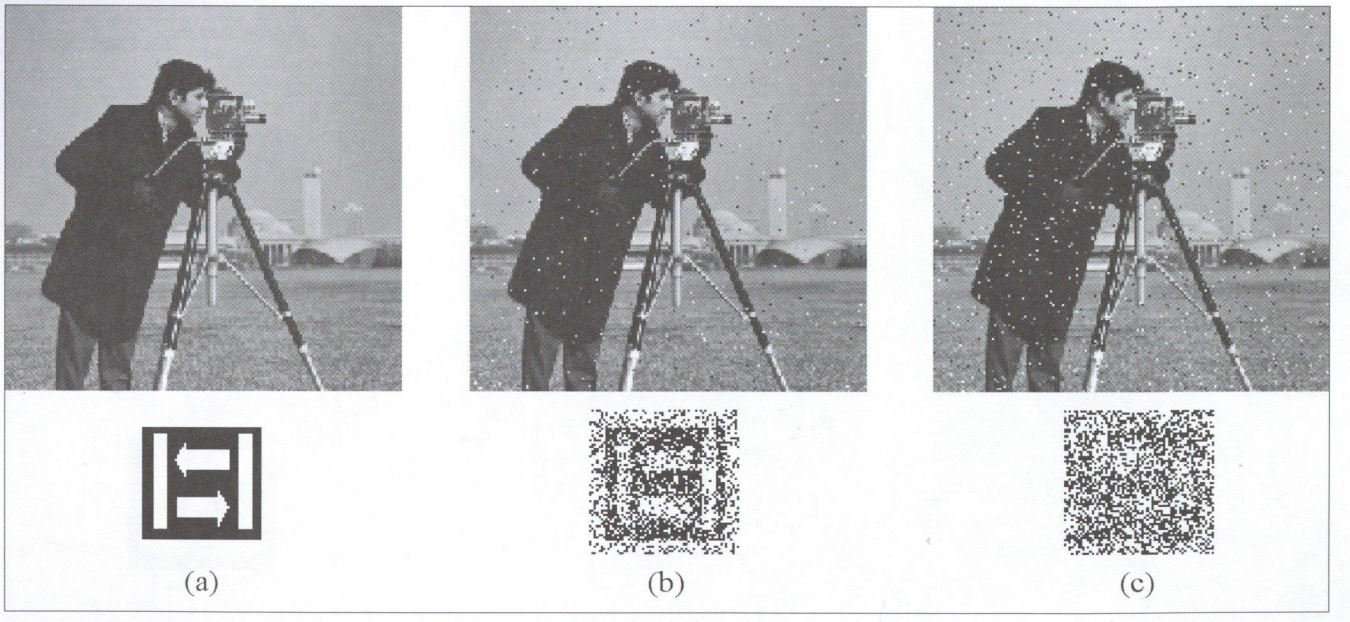


Figure 1 Watermarked images and their extracted watermarks
(a) noise-free ($p_v = 0$) (b) impulse noise $p_v = 0.02$ (c) impulse noise $p_v = 0.05$

$$^{*}x_{i,j} = \begin{cases} o''_{i,j} & \text{with probability } 1-p_v \\ v & \text{with probability } p_v \end{cases} \quad (4)$$

where $x_{i,j}$ is the noisy image sample, $o''_{i,j}$ describes the sample from the noise-free watermarked image, i, j are indices of the sample location, v is the random value from <0,255> and p_v is the impulse probability.

Fig.1/b and *Fig.1/c* show some examples how the impulse noise or bit errors affect the visual quality of extracted watermarks. For that reason, very important task is related to searching for the smoothing function [7],[8]. Note that the smoothing (filtering) function should respect [1],[10] the desired (noise-free) samples and provide the robust estimate for noisy samples, simultaneously. In addition, the hidden information should be preserved with a maximum possible measure [11].

2. Digital Watermarking in DCT Domain

In this Section, the discrete cosine transform (DCT) watermarking algorithm is described. Although there were developed some approaches for the watermark embedding on the image spatial domain [14], the watermark embedding on the image transform domain such as DCT [12],[16], discrete Fourier transformation [19], discrete wavelets transformation [17], etc. is more popular. The reason is that the spread spectrum watermark embedded in a suitably chosen low-medium frequency range results in increased security, invisibility, robustness to lossy compression.

2.1 Watermark embedding

Let $\mathbf{o}_{MxN} = \{o(m,n)\}$ characterize an original image with a size $M \times N$ and m,n denote a sample position bounded by $0 \le m \le M-1, 0 \le n \le N-1$. Let $o(m,n)$ be an image sample represented by 8 bits (256 quantization levels) and $\mathbf{c}_{MxN} = \{c(m,n)\}$ characterize the spectral transformation (DCT, DWT), for two-dimensional (2-D) case described as

$$c(m,n) = \sum_{k=0}^{M-1}\sum_{l=0}^{M-1} \Phi(m,n;k,l)o(k,l) \quad (5)$$

where $c(m,n)$ is a spectral coefficient and $\Phi(m,n;k,l)$ is a transformation kernel. In order to simplify the following expressions, consider a number of samples equal to $S = MN$ and the mapping from $\mathbf{o}_{MxN}$ to $\mathbf{o} = \{o(n)\}$ and from $\mathbf{c}_{MxN}$ to $\mathbf{c} = \{c(n)\}$, where $\mathbf{o}$ and $\mathbf{c}$ are one-dimensional (1-D) vectors of S original samples and spectral coefficients, respectively, achieved by the simple reordering of 2-D arrays to 1-D vectors. Then, the transformation process can be expressed as

$$\mathbf{c} = \mathbf{T}_{SxS}\mathbf{o} \quad (6)$$

where $\mathbf{T}_{SxS}$ characterizes a transformation matrix created from the 2-D transformation kernel.

Let $\mathbf{w}_{KxL} = \{w(m,n)\}$ be a 2-D digital image watermark, i.e. binary data represented by $w(m,n) \in \{-1,1\}$, and $\mathbf{w} = \{w(n)\}$ be a vector of $P=KL$ binary samples achieved by the simple reordering of the 2-D watermark array to the 1-D vector.

After the selection of P spectral coefficients $\{c(i_1), c(i_2), \ldots, c(i_P)\} \in \mathbf{c}$, it is necessary to quantize the selected spectral coefficients to the set $\mathbf{c}_Q = \{c_Q(i_1), c_Q(i_2), \ldots, c_Q(i_P)\}$. Then, the embedding process is characterized by [12]

$$c'(i_k) = \begin{cases} c_Q(i_k) + \Delta & \text{for } w(k) = 1 \\ c_Q(i_k) - \Delta & \text{for } w(k) = -1 \end{cases} \quad \text{for } k = 1,2,\ldots,P \quad (7)$$

where Δ is embedded intensity. The watermarked spectral coefficients are given as follows

$$d(n) = \begin{cases} c'(i_k) & \text{for } n \in \{i_1, i_2, \ldots, i\} \\ c(n) & \text{otherwise} \end{cases} \quad \text{for } n = 1,2,\ldots,S \quad (8)$$

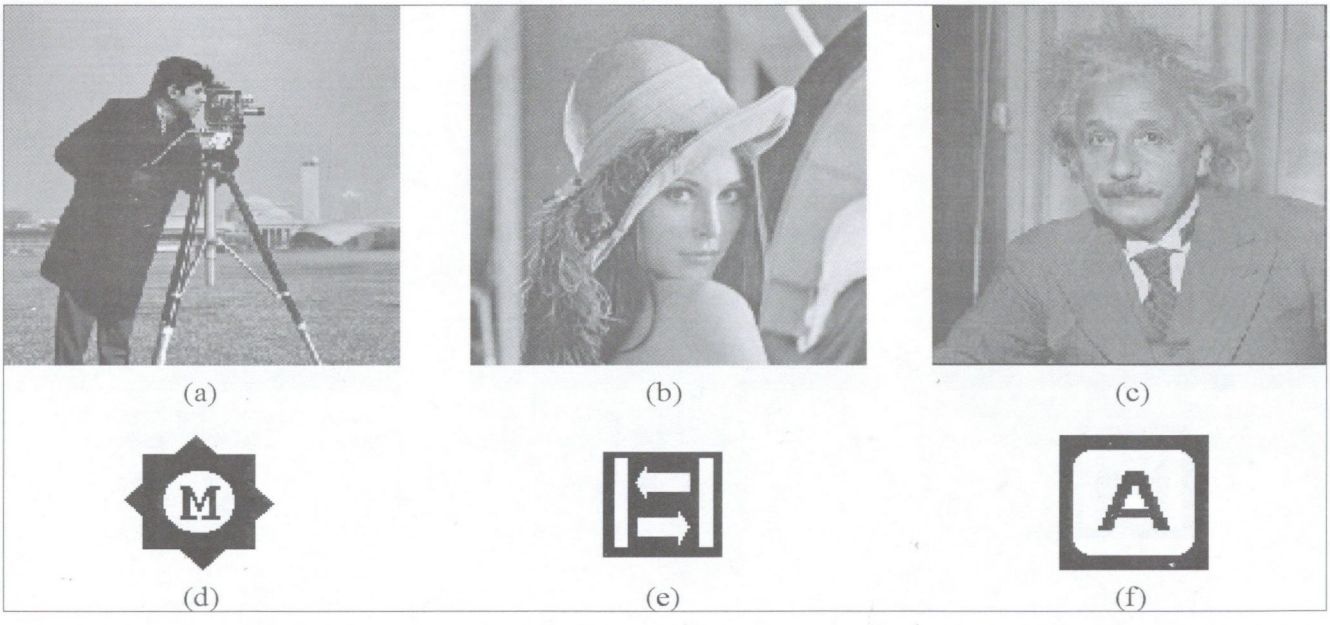


Figure 2 Test images and watermarks
(a) Cameraman (b) Lena (c) Einstein (d) w1 (e) w2 (f) w3

Applying the inverse spectral transformation, the 1-D vector

$$\mathbf{o}' = \mathbf{T}_{S\times S}^{-1}\mathbf{d} \tag{9}$$

corresponding with a watermarked image $\mathbf{o}'_{MxN} = \{o'(m,n)\}$ is achieved. Now, it is necessary to quantize the samples $\{o'(m,n)\}$ to 256 levels, i.e. to use the 8 bit representation written as

$$\mathbf{o}''_{M\times N} = f_Q\left(\mathbf{o}'_{M\times N}\right) + \delta_{M\times N} \tag{10}$$

where $\mathbf{o}''_{MxN} = \{o''(m,n)\}$ is the resulting quantized watermarked image, $f_Q(.)$ is the quantization and $\delta_{MxN} = (m,n)\}$ represents the quantization error whose elements are independent.

Besides the set of quantized spectral coefficients $\mathbf{c}_Q = \{c_Q(i_k)\}$, where $k = 1,2,...,P$, the watermark detection key contains the set $\mathbf{E}_{SxP} = \{e(m,n)\}$ that determines the watermark embedding position, i.e. [12]

$$e(m,n) = \begin{cases} 1 & \text{if } (m,n) = (i_k,k) \\ 0 & \text{otherwise} \end{cases} \quad \text{for } m = 1,2,...,S \text{ and } k = 1,2,...,P \tag{11}$$

Note that the set of watermarked spectral coefficients $\mathbf{d} = \{d(n)\}$ can be described as

$$\mathbf{d} = \mathbf{c}'_Q + \Delta\mathbf{E}_{S\times P}\mathbf{w} \tag{12}$$

where $\mathbf{c}'_Q = \{c'_Q(n)\}$ is given by

$$c'_Q(n) = \begin{cases} c_Q(i_k) & \text{if } n = i_k \\ c(n) & \text{otherwise} \end{cases} \quad \text{for } k = 1,2,...,P \tag{13}$$

2.2 Watermark extraction

In the watermark extracting process, the set of watermarked image intensities $\mathbf{o}'' = \{o''(n)\}$, $1 \le n \le S$, corresponding with the watermarked image $\mathbf{o}''_{MxN}$, is transformed into spectral coefficients, $\mathbf{d}' = \{d(n)\}$ i.e.

$$\mathbf{d}' = \mathbf{T}_{S\times S}\mathbf{o}'' \tag{14}$$

The watermarked coefficients $\mathbf{u} = \{u(n)\}$ can be picked up by the following expression

$$\mathbf{u} = \mathbf{E}'_{P\times S}\mathbf{d}' \tag{15}$$

where $\mathbf{E}'_{SxP}$ characterizes the transpose of the matrix $\mathbf{E}_{SxP}$ (11), however the condition

$$\mathbf{E}'_{P\times S}\mathbf{E}_{S\times P} = \mathbf{I}_{P\times P} \tag{16}$$

where $\mathbf{I}_{PxP}$ is the unit matrix, must be satisfied.

Considering the expressions (12) and (13), the set of watermarked spectral coefficients is given by

$$\mathbf{d}' = \mathbf{c}'_Q + \Delta\mathbf{E}_{S\times P}\mathbf{w} + \mathbf{T}\mathbf{d} \tag{17}$$

where $\delta = \{\delta(n)\}$ is the 1-D vector of quantization errors achieved by the simple reordering of the 2-D matrix δ_{MxN}.

In the sense of the definitions (15),(17) and

$$\mathbf{c}_Q = \mathbf{E}'_{P\times S}\,\mathbf{c}'_Q \tag{18}$$

the extracting process with the key information $\mathbf{c}_Q$ and $\mathbf{E}_{SxP}$ contains steps such as (15) and

$$\mathbf{w}' = \begin{cases} 1 & \text{if } g(\mathbf{u} - \mathbf{c}_Q) \ge 0 \\ -1 & \text{if } g(\mathbf{u} - \mathbf{c}_Q) < 0 \end{cases} \tag{19}$$

Using (15) and (17), the correct watermark extraction means that

$$\mathbf{w}' = \mathbf{w} \tag{20}$$

3. Impulse Detection Strategy

If the noise corruption is modeled as non-Gaussian or impulsive [1],[3], the popular class of linear filters exhibits worse noise attenuation characteristics. A sufficient tool for the suppression of all impulse noise types was proved by a large family of nonlinear order-statistic filters. A nonlinearity of order-statistic filters is given by the ordering [1] of the input samples spanned by a filter window. By this operation, the atypical image samples are removed to the borders of the ordered set. Probably the most popular is the well-known median filter characterised by the excellent robustness against the impulse noise. However it often introduces to an image too much smoothing resulting in a blurring that can be more objectionable than that of the original noise.

That was the reason why many interested order-statistics based filter classes [4],[10],[13], e.g. weighted median filters, LUM smoothers, stack filters, etc. that improve the signal-detail capability of the median filter [4] were developed. Since, the nonlinear filters do not satisfy the superposition property [1], the optimal filtering situation, where only noisy samples are filtered whereas the desired image features will be invariant to a filtering operation can never be fully obtained. However, there exists a way, how to design the method taking the advantage from the optimal filtering situation. This way leads to a connection of the robust median and the impulse detector. If the impulse detector marks the central sample from the filter window as the noise (impulse), the median filter will provide the estimate. Otherwise, the system performs no smoothing, i.e. an identity operation.

In general, each impulse detection [13] is based on the following inequality

$$Val \ge Tol \tag{21}$$

where *Val* corresponds with the detection value that is compared with the threshold value *Tol*. According to the character of the threshold value *Tol*, the impulse detectors can be divided into two classes, i.e. a class with the fixed threshold and a class with the adaptive threshold.

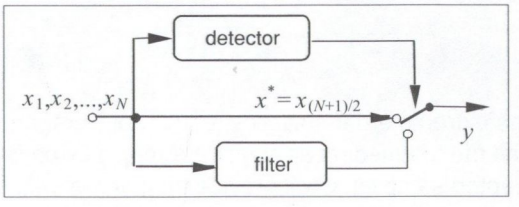


Figure 3 Impulse detector based filtering.

In order to use the detection condition for filtering purposes, it is necessary to specify the following rule:

$$\text{IF} \quad Vol \geq Tol \quad \begin{matrix} \text{THEN} & y = y_{MF} \\ \text{ELSE} & y = x_{(M+1)/2} \end{matrix} \tag{22}$$

If the inequality (21) is valid, the central sample $x_{(N+1)/2}$ is probably corrupted and it will be replaced with the median value y_{MF} from the input set $x_1, x_2, \ldots, x_N$ spawned by the filter window of a window size N. If the condition (21) is not satisfied, then the central sample $x_{(N+1)/2}$ is probably noise-free and no changes will be performed.

3.1 Median detector (MD)

Let $x_1, x_2, \ldots, x_N$ be a discrete-time continuos-valued input set determined by a filter window and $x_{(1)}, x_{(2)}, \ldots, x_{(N)}$ an ordered set so that

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)} \tag{23}$$

$$x_{(i)} \in \{x_1, x_2, \ldots, x_N\} \quad \text{for} \quad i = 1, 2, \ldots, N \tag{24}$$

The median filter output [1],[4],[13] is given by

$$y_{MF} = med\{x_1, x_2, \ldots, x_N\} \tag{25}$$

$$= x_{((N+1)/2)}(n) \tag{26}$$

where *med* is a median operator, N is a window size and $x_{((N+1)/2)}$ is the central sample from the ordered sequence. Note that the description $x_{(N+1)/2}$ characterizes the input central sample.

The advantage of the median operator lies in the robust estimate in the environments corrupted by impulse noise. It is provided by an ordering operation that removes atypical image samples frequently introduced as impulses or outliers to the borders of the ordered set and thus, the median value is probably noise-free sample.

For that reason, the median value is useful not only for the noise filtering, however, it can be used successfully in the detection inequality (21), where the detection value is given by the absolute difference between the median output y_{MF} and the central sample $x_{(N+1)/2}$, i.e.

$$Val = \left| y_{MF} - x_{(N+1)/2} \right| \tag{27}$$

and *Tol* is the fixed threshold value optimally set as *Tol* = 60

3.2 Order-statistic detector (OD)

Now, consider the set of n mid-positioned ordered samples $x_{((N+1)/2)}$, $x_{(1+(n+1)/2)}, \ldots, x_{(N-(n+1)/2)}$, and the simple trimmed mean defined as

$$\mu_n = \frac{1}{n} \sum_{i=(n+1)/2}^{N-(n+1)/2} x_{(i)} \tag{28}$$

Since, the extreme order-statistics (usually outliers) are excluded and the trimmed mean μ_n is determined by probably no corrupted samples, it will provide the precise value necessary to be compare with the central sample $x_{(N+1)/2}$. Thus, the detector value can be stated as follows

$$Val = \left| \mu_n - x_{(N+1)/2} \right| \tag{29}$$

In the OD detection scheme [15] of the window size N = 9, the threshold value necessary in the detector-filter structure (22) is optimally chosen as *Tol* = 45 for the number of considered mid-positioned samples n = 5.

3.3 Sigma concept based detector (SD)

The last presented detection-filter structure is based on the sigma filtering concept [9], i.e. the simple statistical measures such as the mean value and standard deviation are utilized.

The detection operator is defined as

$$Val = \left| \mu - x_{(N+1)/2} \right| \tag{30}$$

where μ is a sample mean of observed data $x_1, x_2, \ldots, x_N$ and the adaptive threshold value *Tol* is given by

$$Tol = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2} \tag{31}$$

If the detection value is greater than or equal to the standard deviation or the adaptive threshold *Tol* the central sample is probably distorted because it is more different from other input samples.

4. Experimental Results

Figure 2 shows the original test images Cameraman, Lena, Einstein and three used watermarks. The test images have a resolution 256x256 image samples with an 8 bit per sample representation and the used watermarks have a size of 64x64 samples.

The used objective criterion is given by the *Peak-Signal to Noise Ratio* (PSNR) defined as follows

$$PSNR = 10 \log_{10} \frac{255^2}{\frac{1}{KL} \sum_{i=1}^{K} \sum_{j=1}^{L} \left(w_{i,j} - w'_{i,j} \right)^2} \tag{32}$$

where K, L represent watermark dimensions, i, j determine the time position, $w_{i,j}$ and $w'_{i,j}$ are samples from the original watermark and the extracted watermark, respectively.

Thus, the difference between these watermarks will be computed for the watermark extracted from noisy and filtered images, too. It will be interesting to observe that the undesired filter behavior can degrade the quality of extracted watermarks to the measure corresponding with the original noise. In addition, the quality of extracted watermarks will decrease with the increased impulse noise corruption P_v.

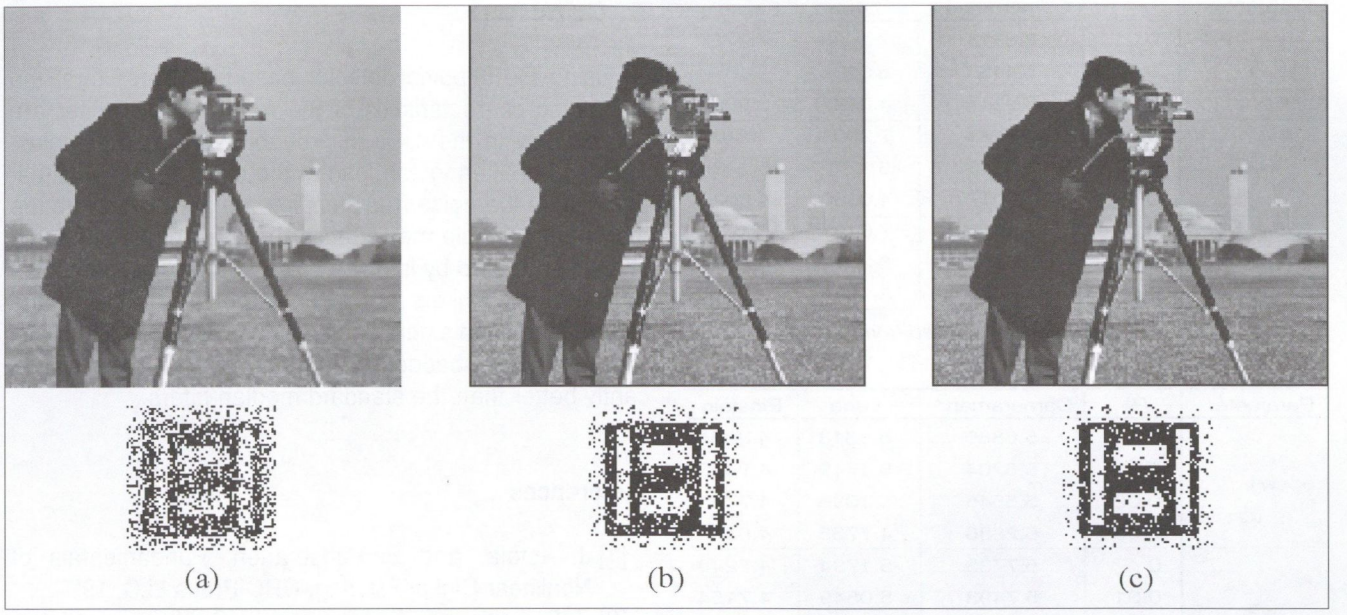


Figure 4 Filtered watermarked images and extracted watermarks related to no corruption.
(a) median filter (b) OD approach (c) SD approach

The noisy watermarked images and the corresponding extracted watermark are shown in Figure 1. It can be easily seen that the increased noise corruption P_v increases the watermark degradation. Since, there exist situations, where the watermarked images are transmitted with no corruption, *Figure 4* shows the filter influence on watermarked data, only. The median filter affects the whole image without the additional information about the sample corruption, it introduces to an image too much smoothing resulting in a blurring. For that reason, the extracted watermark *(Fig.4/a)* is characterized by significant corruption. The proposed impulse detector based median filters affect only the atypical samples especially at the image edges and thus, they preserve the hidden watermark *(Fig.4/b,c)*. In the case of noisy watermarked image filtering *(Fig.5 and Fig.6)*, the median filter suppresses the noise excellently, however its estimation error introduced to the output images prohibits practically recognizing the useful information of extracted watermarks. If the performance of the impulse detector-based median filters is observed, the best results are provided by OD approach. It preserves the signal-details excellently, however similarly to the case of noise-free image filtering *(Fig.4c)*, a small estimation error especially at the image edges slightly influences the hidden data. Numerically, the achieved results are provided in *Tables 1-5*. It can be seen that the median filter provides two times worse PSNR than that of the OD approach. This difference decreases with the increased noise corruption P_v, where the impulse detection characteristics of proposed methods get worsen.

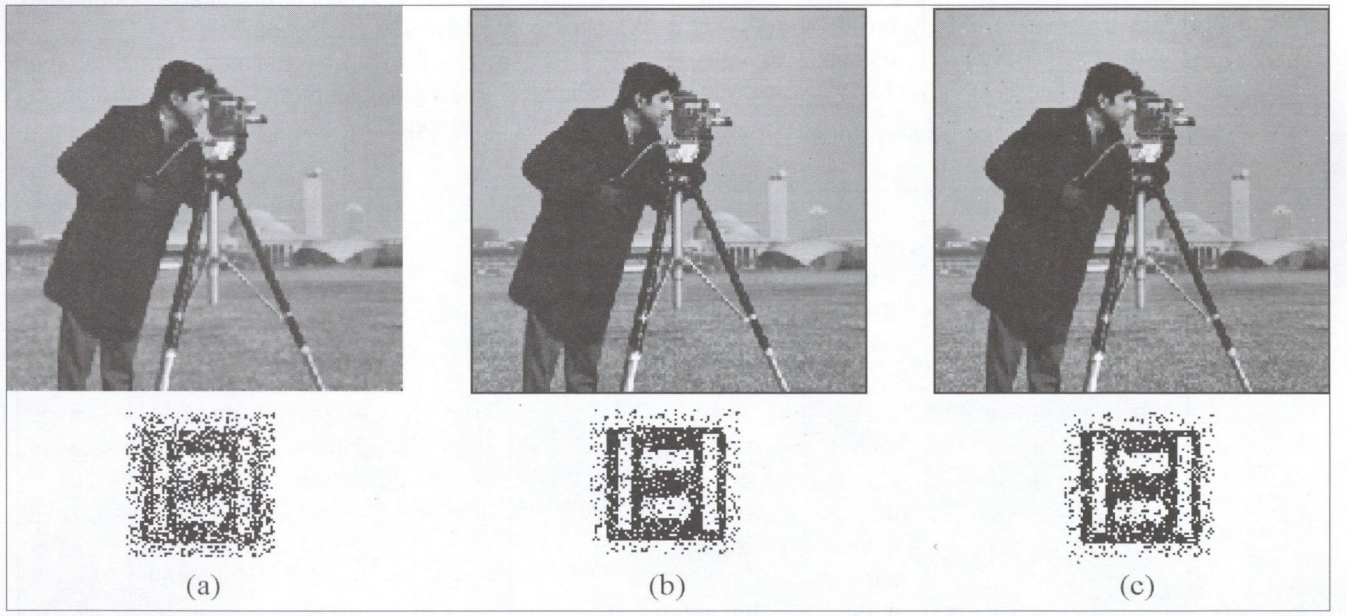


Figure 5 Filtered watermarked images and extracted watermarks related to impulse noise p_i = 0,02.
(a) median filter (b) MD approach (c) OD approach

Parameter	P_v	Cameraman	Lena	Einstein
w1	0.01	7.6913	7.1023	7.2111
	0.02	5.4455	5.0638	5.1870
	0.05	3.9147	4.0121	3.9862
w2	0.01	7.7163	7.2004	7.2480
	0.02	5.4940	5.1272	5.1936
	0.05	4.0017	4.0380	4.0027
w3	0.01	7.4913	7.0499	7.1924
	0.02	5.6392	5.3010	5.2087
	0.05	4.0044	3.9899	3.9965

Table 1 PSNR of watermarks extracted from noisy images.

Parameter	P_v	Cameraman	Lena	Einstein
w1	0	5.6860	5.3318	4.8851
	0.01	5.5204	5.1719	4.8493
	0.02	5.5546	5.1096	4.7786
	0.05	5.2886	4.7755	4.6561
w2	0	5.7733	5.1794	4.7248
	0.01	5.7493	5.0549	4.7154
	0.02	5.5546	5.0753	4.7122
	0.05	5.4753	4.7028	4.4830
w3	0	5.6509	5.2352	4.9443
	0.01	5.5393	5.0958	4.8558
	0.02	5.5128	5.0076	4.8461
	0.05	5.4269	4.6966	4.7217

Table 2 PSNR achieved by median filter.

Parameter	P_v	Cameraman	Lena	Einstein
w1	0	8.5345	9.0140	10.9914
	0.01	8.2771	8.8209	10.3704
	0.02	8.0959	8.4004	9.9536
	0.05	7.9085	7.5805	9.0993
w2	0	8.5421	8.9803	11.2100
	0.01	8.3185	8.6004	10.7065
	0.02	8.1648	8.4446	10.3372
	0.05	7.9479	8.0078	9.2039
w3	0	8.3784	8.9887	11.2664
	0.01	8.1971	8.6920	10.8761
	0.02	8.0415	8.2989	10.4885
	0.05	7.6171	7.8113	9.0309

Table 3 PSNR achieved by MD approach.

Parameter	P_v	Cameraman	Lena	Einstein
w1	0	9.8602	11.0048	14.4801
	0.01	8.3408	9.9730	12.4883
	0.02	7.6356	8.4818	9.1688
	0.05	6.4575	6.4952	7.3499
w2	0	10.1466	11.0450	14.6623
	0.01	8.7085	10.0121	12.2922
	0.02	8.0213	8.4595	9.7587
	0.05	6.6837	6.7035	7.1583
w3	0	9.8808	10.8344	14.4210
	0.01	8.9912	9.7804	12.4036
	0.02	7.7351	8.0078	9.8397
	0.05	6.3464	6.4294	7.4372

Table 4 PSNR achieved by OD approach.

5. Conclusion

By using the impulse detector based median filters, the smoothing characteristics of the robust median filter are utilized only in the case of probably corrupted samples. This kind of filtering can lead to the optimal filtering situation, where the noise-free samples are preserved with the maximum possible measure. In the case of watermarked images corrupted by impulse noise, the proposed impulse detector-based filters are able to preserve not only the image edges and signal details, however, these filters can preserve the embedded watermark (hidden data) significantly better than the standard median filters.

References

[1] J. Astola, and P. Kuosmanen, Fundamentals of Nonlinear Digital Filtering, CRC Press LLC, 1997.

[2] Z. Bojkovic, and D. Milovanovic, Challenges of Information Processing in Multimedia Communications, Proceedings of the 3rd EURASIP Conference on Digital Signal Processing and Multimedia Communications and Services ECMCS-2001 in Budapest, Hungary, September 11-13, 2001, pp.133-140.

[3] C. Boncelet, Image Noise Models, in Handbook of Image & Video Processing (ed. A. Bovik), Academic Press, 2000, pp.325-336.

[4] T. Chen, K.K. Ma, and L.H. Chen, Tri-State Median Filter for Image Denoising, IEEE Transactions on Image Processing, Vol.8, No.12, December 1999, pp. 1834-1838.

[5] T. Chen, H.R. Wu, Adaptive Impulse Detection Using Center-Weighted Median Filters, IEEE Signal Processing Letters, Vol.8, No.1, January 2001, pp.1-3.

[6] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, Secure Spread Spectrum Watermarking for Multimedia, IEEE Transactions on Image Processing, Vol.6, No. 12, 1997, 1673-1687.

[7] J. Glasa, On Derivatives Estimation of Smoothed Digital Curves, Computers and Artificial Intelligence, Vol.19, 2000, pp.335-349.

Parameter	P_v	Cameraman	Lena	Einstein
w1	0	6.0804	6.0206	5.5928
	0.01	6.1138	5.9920	5.6407
	0.02	6.1063	5.9117	5.6821
	0.05	6.1716	5.7176	5.8787
w2	0	6.3327	5.9533	5.8338
	0.01	6.3882	5.9899	5.9125
	0.02	6.4669	6.0206	5.9449
	0.05	6.3190	5.9283	5.9117
w3	0	6.3281	5.9700	5.7613
	0.01	6.2184	6.0136	5.8173
	0.02	6.3647	6.0164	5.7853
	0.05	6.1323	5.9700	5.8176

Table 5 PSNR achieved by SD approach

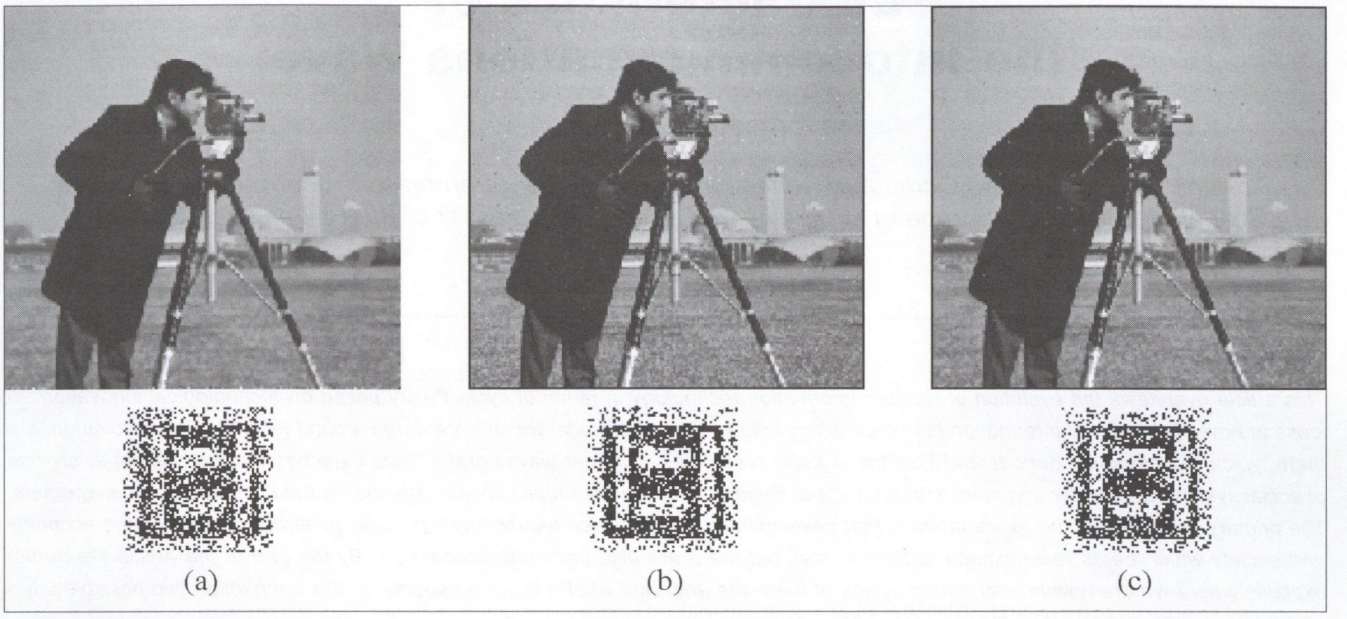


Figure 6 Filtered watermarked images and extracted watermarks related to impulse noise
(a) median filter (b) OD approach (c) SD approach

[8] J. Glasa, Bit-Level Systolic Arrays for Digital Contour Smoothing by Abel-Poisson Kernel, Parallel Processing Letters, Vol.3, 1993, pp.43-51.

[9] J.S. Lee, Digital Image Smoothing and the Sigma Filter, Computer Vision, Graphics, and Image Processing, Vol.24, No.2, November 1983, pp.255-269.

[10] R. Lukac, and S. Marchevsky, LUM Smoother with Smooth Control for Noisy Image Sequences, EURASIP Journal on Applied Signal Processing, Vol.2001, No.2, 2001, pp.110-120.

[11] R. Lukac, Details Preserving LUM FTC Filter for Noisy Images with Hidden Information, Proceedings of the Conference, Training and Workshops on Electronic Imaging & Visual Arts EVA 2002 in Florence, Italy, March 18-22, 2002, pp.104-109.

[12] A. Miyazaki, and A. Okamoto, Analysis of Watermarking Systems in the Frequency Domain and its Application to Design of Robust Watermarking Systems, Proceedings of the 2001 IEEE International Conference on Image Processing ICIP 2001 in Thessaloniki, Greece, October 7-10, 2001, Vol.2, pp.506-509.

[13] V. Moucha, S. Marchevsky, R. Lukac, and C. Stupak, Digital Image Signal Filtering. Editorial Center of VLA gen. M. R. Stefánika in Kosice, Kosice 2000.

[14] N. Nikolaidis, and I. Pitas, Robust Image Watermarking in the Spatial Domain, Signal Processing, Vol.66, No.3, 1998, pp.385-403.

[15] J. Park, and L. Kurz, Image Enhancement Using the Modified ICM Method, IEEE Transactions on Image Processing, Vol.5, No.5, May 1996, pp.765-771.

[16] M Swanson, M. Kobayashi and H. Tewfik, Multimedia Data-Embedding and Watermarking Technologies, Proceedings of the IEEE, Vol.86, No.6, June 1998, pp.1064-1087.

[17] D. Taskovski, S. Bogdanova, and M. Bogdanov: A Low Resolution Content Based on Watermarking Images in Wavelet Domain, Proceedings of 2nd IEEE Region 8-EURASIP Symposium on Image and Signal Processing and Analysis ISPA'01 in Pula, Croatia, June 19-21, 2001, pp.604-608.

[18] A. Z. Tirkel, C. F. Osborne, and T. E. Hall, Image and Watermark Registration, Signal Processing, Vol.66, No.3, 1998, pp.373-383.

[19] G. Voyatzis, and I. Pitas, Image Watermarking for Copyright Protection and Authentication, in Handbook of Image & Video Processing (ed. A. Bovik), Academic Press, 2000, pp.733-745.

* ***Rastislav Lukac*** *received the Diploma in Telecommunications degree with honors in 1998 and the Ph.D. degree in 2001, both at the Technical University of Kosice, Slovak Republic. From February 2001 till August 2002 he was an assistant professor at the Department of Electronics and Multimedia Communications at the Technical University of Kosice. Since August 2002 he is a researcher in Slovak Image Processing Center in Dobsina, Slovak Republic.*

Recently, his research interests include nonlinear digital filters, impulse detection, color image processing, image sequence processing and the use of Boolean functions and permutation theory in filter design.

Dr. Lukac is a member of the IEEE Signal Processing Society. He is an active member of Review and Program Committees at some European conferences and a reviewer for some scientific journals.

Characteristics of the Infocommunications Wave

GYÖRGY BŐGEL

Strategic advisor at KFKI Computer Technology Ltd., member of professors' board at the Business School of the Central European University and associate professor of the University of Debrecen

gbogel@kfki.hu

"...life itself is cyclic..." (András Bródy)

This article overviews the evolution of modern information technology in terms of cycle theory based on technological innovation . It looks at how the history of information high technology follows the cycle model which is centered around technological innovation, are there typical development periods and how the IT cycle resembles to former waves and if there have been brought about a "change of social system" on a wider and more profound scale. Regarding this latter subject special attention is paid to corporate management. The primary statement of the examination is that development of information technology has really generated a wave in the economy and society which bears resemblance to former ones but there are also certain discrepancies. By the end of the article the author explains where we are now in information cycles (if there are any) and what are our prospects. In this connection two possible interpretations and scenarios are outlined.

1. Innovation waves in the economy and society

General features of technological innovation waves

The scope of important technological innovations (such as steam engine, railway or telephone) goes far beyond limits of technology or economy. Each wave has a *bearing industry* to which other leading industries join: e.g. steam engine forms an "innovation bundle" with railway, railway related equipment and steam-driven manufacturing tools. The development of these industries is based on certain *inputs* (on iron and coal in our example). The given period has its basic transporting and communication means (railway, telegraph, steamboat) which themselves have influence on the rate of spreading of innovations. Technological innovations are accompanied by organizational and managerial innovations: railway companies were founded mainly in the form of corporation.

It is easy to recognize that the propagation of new procedures and products are affected by several factors, such as maturity of former technologies (car industry could not have boosted without electricity and railway), the abundance and availability of raw materials, availability of energy and infrastructure, quantity and qualification of manpower, consumer habits and inclination for purchasing, legal system and public regulation. Without a certain harmony among these factors the wave cannot start.

Based on a work by Carlota Perez [2000] let's have a look at how the "change of system" driven by technological innovations have taken place. According to the Venezuelan researcher events follow a specific pattern. The "implosion" of the new technology requires dramatic price reduction of certain *basic inputs* (e.g. coal, iron, steal, electricity, oil, microchip). This allows for the take-up of certain industries using low cost sources in high quantities. This attracts more businesses to the production of the input and owing to bulkiness will further improve economical indicators. A new "planning space" opens for engineers and entrepreneurs with low-cost and readily available inputs, new procedures and models. The spreading of input-based products (railway, electrical appliances, car, etc.) starts off which leads to an "implosion" of additional industries dealing with their sale, service and support, and also new infrastructure will be needed. The self-exciting interplay of industries producing inputs, "bearing" new products, providing support and infrastructure, the progress on the learning curve, the efficiency of mass production, the extra profit attracting innovators gives an impetus to economy. This phenomenon is called *positive growth spiral* by Bill Gates, who uses this expression to characterize the young information technology industry in his book published in 1995. George Gilder observed the same process regarding optical fibers and forecasted possible future deployments.

For the design, production, distribution and use of new products and technologies new managing, learning and other systems, ideologies, rules, lifestyles, culture, governance and political order are needed. The massively growing and highly profitable new industries do their best to comply with these requirements: they are distributing, promoting, lobbying, explaining, showing example, etc. Since the old order inevitably *resists*, the new harmony is born in heavy debates, sometimes in blood and tears. Workers and employees of falling, declining industries and related systems are laid off, skills, competences and expertises lose their value, social and political tensions occur, heavy debates begin in the street, cafes and the parliament on all sorts of regulatory, economical, political, duties and

* *The original, version of this article was published in Competitio, the internal periodical of the Department of Economics of the University of Debrecen.*

other public administration issues. The development of the events are affected by specific *local circumstances* (political and geographical situation, historical heritages, language, personality of important persons, etc.) as well. The wave begins to propagate, extends and brings about a "system change".

Schumpeter [1980] calls this process "creative destruction" suggesting thereby that something perishes but at the same time something new is created, the declining phase of one wave is overlapping the ascending phase of another.

Life cycle of technologies

The life cycle of technologies leading to "system change" can be divided into the following typical phases (though, as the above description suggests, they are rather bundles of technology deriving from the same roots and they include related work management methods as well):

a) Latency. The new technology is in the *laboratory phase.* (Depending on the age, under "laboratory" a cloister, the closet of a castle, a barn or even a garaged should be meant.) The first prototypes are born which are rather immature solutions due to in-house implementation. The first product presentations are held and first patents filed. All sorts of application experiments are carried out but works on technical history (e.g. [Greguss, 1985]) make it clear that inventors and scientists often have no idea of what the innovation can be used for, not mentioning the general public. This phase can last quite a long while the scope of invention is small and new products can hardly be seen.

b) Proving. The technical feasibility is proved. New technologies and products seem to be viable not only from technical but also from *business* point of view: applications attracting mass market are found, there is demand for them, the return on investments is guaranteed, profit can be made. New on innovative products and technologies spread over, there is wide interest in them both in consumers and business circles. This is the moment in which the great socio-economical wave begins to propagate.

c) Implosion. The interplay among inputs as well as bearing, carrying and additional industries leading to the positive growth spiral mentioned above has started. There is a general enthusiasm in innovator circles, the hope of high profit is very attractive, the *business foundation fever* gets high. More and more new enterprises are established. As András Bródy notes, the burstiness of innovation – this "burst" happens in the *implosion* phase – is not the *cause* but the *consequence* of the cyclic course of economy [Bródy 1983]. This is to say on one hand that from economic point of view "innovation" means no the invention of something but its wide-spread application, on the other hand it indicates that the "innovation burst" is a function of cyclic movement of other factors of economy (e.g. rate of interest, investments).

d) Growth. New technologies and products become generally known and accepted, part of everyday life and form a *dominant system.* The scope of applications extends dramatically. Periods of style waves and "craziness" add to the generally increasing trend, observers forecast even the coming of general and lasting golden age. (The nature of mania is analyzed in details by Kindleberger [2000].) Profound changes take place in society, politics, corporate management, education, culture, legislation and regulation.

e) Slow-down. Dominant technologies get matured, their development is not of revolutionary but of evolutionary character. The era of fast enrichment closes down, saturating markets and heavy competition bring down profitability. Certain sources run dry and become more expensive. Signs of surplus appear in the market. Even bankrupts are normal. This again is a difficult period, especially for those who are accustomed to the euphoria, high growth rates and revenues of the former period, and the associated social status and public interest. They feel to loose their ground. This reminds us the inflection point of Grove [1997]: this is the bend where dramatic moves, structural adaptation is needed. The slow-down is getting more and more obvious in the whole economy, the society is anxious, pessimistic forecasts tend to prevail. Meanwhile germs of new technologies appear which, after the proving period, can "implode" and form the world to their own shape.

f) Maturity. At the end of a wave the "declining" section should come but we want to avoid this word. There are cases where at the end of a wave the technology and associated products simply *disappear.* Owing to an impulse, however, a renascence can happen as well, because say a new application is found for the product. In this case the life cycle starts from the beginning. There is one more opportunity: though the underlying technology is different but the product stays with us and we can use it further: the age of the great railway constructions is over but we could not live without railway even today.

Wave in the car industry

Having gone through the general description of the phases of technological life cycles let's turn to the study of the latest, more or less complete technological wave. Cutting-edge products of this wave were internal combustion engine and car.

a) Latency. Though internal combustion engine was invented in the mid-1800s, it had no special influence on the industry and economy for a long time. Manufacturing of internal combustion engine-driven cars took place in small car production workshops both in the USA and Europe. Driving cars was the amusement of the rich.

b) Prove. Between 1908 and 1914 Henry Ford introduced several innovations in his car manufactory.

(Remember: he got a lot of help from a Hungarian engineer József Galamb.) The more important of them concerned not the construction of the product but the way of production. Ford replaced traditional, individual elements by standardized, interchangeable parts which were produced using single purpose machines. It was realized that cars can be stabilized as standardized products, the mass manufacturing can be organized rationally and a successful business model can be built on machine-production. Henry Ford himself acted also as socio-philosopher [Ford 1989], many of his ideas were realized but he had also some ambiguous ideas [Baldwin 2001]. Using the Greiner model [1995] which describe typical development phases of businesses we should say that Ford transferred car manufacturing from "creativity" phase to "professional management" phase.

c) Implosion. In the 20s of the past century internal combustion engine reached an *exclusive role*. Hundreds of car manufactories were replaced by car factories working on the basis of Ford's principles. Thanks to standardization, new work organization methods, consistent rationalization the productivity of work had seen a spectacular growth. Petrol as "basic input" became a product of stable quality with low price and general availability. Between 1908 and 1927 the manufactured number of Ford T-model amounted to 15 million. Its price matched more and more to the pocket of the man-of-the-street: in 1908 it cost 850 dollars, in 1913 the price went down to 600 dollars while in 1916 it could be purchased for just 360 dollars.

d) Growth. Between 1938 and 1980 the car production of the world extended considerably. Some small workshops specialized in specific models continued staying in the market but *factory giants* were at the top of the industry. Along with car industry the associated bearing, additional and infrastructural industries were also extended. Innovations in car making followed one after the other. The market segmentation started which led to differentiation of products. General Motors could react faster to these developments than Ford [Chandler 1969].

Schultz, who is studying investments into human capital demonstrated that in the US population the per capita number of constant school years increased by more than six times between 1900 and 1957 [Schultz 1971], the growth sharply intensified after 1929 and got an impetus after the World War II. This means the "knowledge based economy" dates far more back than the end of the 20th century and has nothing to do with information technology. As we will see later, the "new wave" could not have started without new "knowledge bases" and "knowledge organizations". (Probably those are right who say economy has always been based on "knowledge".)

e) Slow-down. During the 70s of the past century the growth slowed down, profit rate in the American car industries began to decrease. Some companies went bankrupt, great acquisitions and mergers started, the market became *oligopolistic*. All over the world 70s and 80s passed in *crisis environment*. In 1973 the OPEC countries brought about an oil crisis which clearly showed how the world became dependent on a limited natural resource. Oil crisis was followed by a general price increase. More and more developing countries run into critical debts. In course of the 60s unemployment rate was low, in the early 80s, however, it varied between 5 to 10% in developed countries. András Bródy, who investigated in details the causes of slow-down and general bad-feeling, gives the following forecast at the early 80s: "If theories and views outlined in my book are correct, then a general stagnation is waiting for us during the next two decades, i.e. until the end of the century." ([Bródy 1983] p.161.) More profound and critical recession is to come – he says. In his book just a few sentences refer to the microelectronic industry where the motion is above the average.

f) Maturity. The car industry did not disappear, of course, after the slow-down phase: there are even scenarios for reform, adaptation and symbiosis. Huge companies dominate the industry, profit rates are low, notable innovations are rare. The importance of the industry remained unchanged but it was not this industry that led the next innovation wave and the "long upswing" of the 90s. Late phases of the car industry wave overlapped the beginning and upswing of the info-communications wave.

2. Features of the info-communications wave

The car industry wave was followed by computer technology, information technology and telecommunications wave. The first phase of its history follows the classical pattern but it cannot be considered as completed: there are several scenarios for the present and the near future.

Latency

The period of latency was rather long in the information technology industry. Let's go back in time. Blaise Pascal constructed a calculating machine in 1642 and Charles Babbage introduced machines carrying out complex tasks between 1820 and 1860.

During World War II the British government charged Alan Turing, mathematician of the Manchester University, of the development of a machine which is able to decipher the Enigma military code of the Germans. We should not forget about the fact that governments of developed countries recognized the military importance of the matter and showed intense interest in the electronics, telecommunications and then computer technology industry which manifested in considerable financial support and great state orders from the very beginning. Despite all sorts of popular stories the root of modern information technology was not in small garages but in university departments and research institutes. In other words: it was born in knowledge centers enjoying public and military support inherited from the previous wave. Machines like ENIAC,

EDVAC and UNIVAC were constructed at the University of Pennsylvania. The development of basic architecture of today's computers (CPU, memory devices, input/output devices) was performed also by a university scientist, John Neumann. The wave of the information technology innovation could not have started without a new and very strong network connecting state (money and order), universities (knowledge) and corporate research & development division. Until as far as the 50s even IBM's market analysts indicated that just a few machines can be sold on the market. The same "latency" happened to Internet as well. It existed already in the 60s under the name of ARPANET and was a university developed and deployed network with the financial support of Pentagon.

Prove

The manufacturing technology of computers was more and more stabilized. Capacity of computers increased at a large pace: ENIAC in 1946 could execute 45 arithmetic operations per second, the IBM 360/75 model in 1965 was able to perform nearly 1.5 million operations. In the early 50s IBM was surprised to see the sale of 1,800 of its 650 model while market analysts forecast demand for a few dozen. The company went into motion and made quite a good profit with its mainframes. First in the economy these computers were used for business purposes but later production control and industrial process control solutions became more and more frequent.

IBM focused on mainframes. The first minicomputers were introduced on the market in 1963 by Digital Equipment Corporation, which were then followed by microcomputers. They were suitable for many purposes, e.g. they facilitated automation of industrial processes creating thereby the basis for the vision of a fully automated workshop. Special software and hardware were also developed for these computers.

It was thus proven that the information technology industry just taking its wings is able to make viable products for which there is real and considerable business demand. More and more companies bought computers. Their advent did not question work and corporate organization methods of Ford, moreover, they were justified together with centralized bureaucratic structures. Market was dominated by mainframes which were generally used by a separate functional unit ("electronic data processing department") for standard purposes such as payroll accounting or billing. No "system change" occurred yet, it will start in fact with the appearance of desktop computers using Intel microprocessor.

Implosion

Mainframes fit quite well into Ford's value grid and structures. IBM, their main manufacturer showed little interest in desktop computers. Market of this latter was overrun by new entrepreneurs like Atari, Apple, Commodore and Radio Shack. They offered diverse software and peripherals. Lot of people started to play and work with these low-cost computers. Information technology left corporate data processing centers, universities and military basis and entered the everyday life of people.

Recognizing the danger of delay, in the early 80s IBM developed its own desktop computer using standard, off-the-shelf parts, Microsoft's operating system and other Microsoft software. While most new companies mentioned in the previous paragraph perished, PC quickly and successfully headed for the upper market segments.

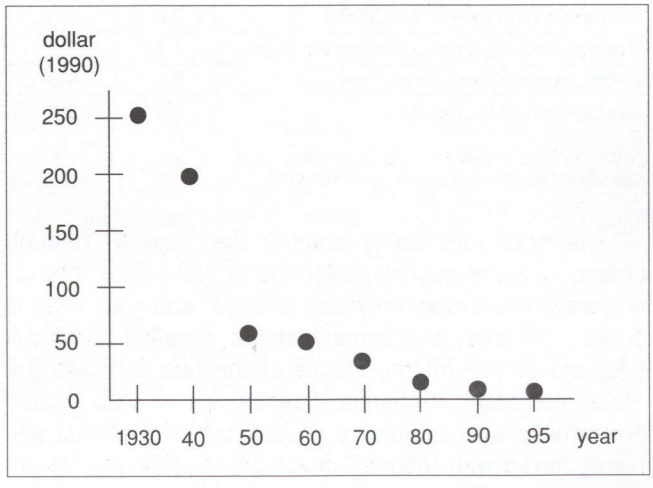
So this was the *cutting-edge product* which was produced using microchips, a *source* with dramatically decreasing price. According to Moore's Law, capacity of chips increased exponentially. Then *bearing, additional* and *infrastructural* industries appeared or closed up, digitization of telecommunications and convergence with computer technology started. In the 70s telecommunications tariffs were much lower than before *(Figure 1)*. Successful innovators realized high profit, it was at this time that many leading persons of the sector established their richness. In the down-hearted atmosphere and slow-down period of the 70s info-communications industry gave positive signals to the world: come here, you find gold here!

Growth

Year 90s were the golden age of information technology industry. In the United States in 1955 there were 1,000, in 1965 there were 30,000, in 1975 there were 100,000 whereas in 1985 there were more than one million computers. By the turn of the century the number of sold desktop computers exceeded one billion, that of Internet users exceeded half billion. These figures suggest that the take-off point where growth curve turns suddenly steep, can be put to the early 80s.

This dynamic growth required one more thing. In the beginning information technology industry was built up in a *vertical* structure: companies like IBM, DEC or Wang produced from chips to peripherals everything themselves while their products were not compatible with each other. This hindered development and growth.

Figure 1
Tariff of three-minutes New York–London telephone call
Source: Global Economic Prospects and the Developing Countries 1995. World Bank



Andrew Grove, the Hungarian-born Intel-chief write in a book [Grove 1997], how, among what tensions and difficulties transformed information technology industry from vertical into *horizontal,* and how the industry got impetus from specialization and the general compatibility.

The last decade of the century passed in the atmosphere of unbroken and rapid growth. Statistical figures show more steep curves than in the same phase of car industry wave. There are several reasons for this and they all contribute to the typical run of the phase. *Networking* plays an important role in the growth, this can also be observed in the info-communications sector. One could say, this is one of the important laws of information economics [Shapiro-Varian 1999]. During the past decade this sector was given new and new technical innovation pulses. In the mid-90s, for example, the easy-to-use browser "imploded" the Internet [Bõgel 1999]. The internationalization and globalization of the economy had started already at time of the previous wave, in this new wave one can count on less and less geographical limits, moreover, most information products can easily be transported through well constructed networks. The typical cost structure of info-communications sector also contributed to the growth: huge amounts invested into research, development and network construction cannot return unless in case of a rapid rise in quantity. We should also remember the media which has a very important role in shaping public opinion [Samuelson 2002].

Because of technical, historical, political and other factors the rate of growth was different in individual regions of the world. No doubt, the United States was the engine of world economy, where in the late 90s clearly the source of spectacular growth was the info-communications sector *(Table 1).* An immense amount of capital went into the information sector, information technology devices and services (including hardware, software and telecommunications) play more and more important role in business investments and expenditures of people. A "large-scale info-communications industry" was formed, big computer factories and software houses, Internet service providers and telecommunications companies gained dominant position on product and capital markets.

	1991-1995	1995-2000
Average annual growth of GDP (%)	3,0	4,3
Average annual growth of productivity (%)	1,7	2,8
Average unemployment rate (%)	6,6	4,8
Average annual inflation (%)	3,3	2,3

Table 1 Figures of the US economy
Source: US Department of Commerce, March 2001

The next question is whether the "regular" (though unique in many aspects) information technology innovation wave has brought "system change" and if so, what is it like. The clear assessment of the situation is difficult because of the *Internet-mania* of the late 90s and the associated stock exchange "balloon" wit all their related phenomena and symptoms. In this period irrational elements mixed with "normal" development, though – as we already saw it – this is not unusual in this phase, some ad hoc irrationality is part of the normal course of things. Market assessment of computer technology, telecommunications and mainly dotcom companies went far from reality. The press was full with articles which forecast seismic changes and spectacular visions suggesting that the reader inevitably falls out of competition unless he or she does not hurry to buy a new device, does not re-organize completely the company, does not enter the e-business or does not buy the shares of a brand new dotcom company. Some companies and persons were celebrated as heroes. Voices of clear-headed observers [e.g. Mandel 2000] were not heard in the all-pervasive enthusiasm. Since then, many have analyzed causes and consequences: events have developed approximately according to similar phenomena in previous times as described by Galbraith [1997] and Kindleberger [2000]. Anyway, the "dotcom mania" placed a "humpback" on statistical curves and removed the sharp boundary between illusion, myth and reality. (Investments associated to Y2K problem emphasized this feature.)

Let's put up the question again: has the information technology wave brought real "system change"? The answer is *yes* but there is a lot of uncertainty regarding content, direction and completeness of changes: no doubt there were and are changes in the system, however, it is difficult to say, which of them will be profound and long-lasting and which of them will prove temporary or even scenical.

Changes in companies

Let's take *companies* first and see if there have been changes in Ford's company model as consequences of technological innovations. It is easy to declare that in info-communications sector there have been created new or novel *business models* such as "portal", "electronic marketplace", "demand aggregator" or "content provider" [Weill-Vitale 2001]. However, in our study changes in division of labor and coordination are much more interesting. Scanning the literature (e.g. [Kocsis-Szabó 2000]; [Drucker 2002]; [Ranadivé 1999]; [The Economist 2002]), roughly the following picture shows up about the "ideal type" of *electronics based* companies (using many types of information technology devices and applications), *integrated* companies (those which integrate their information technology applications along processes and projects) *extended* companies (connected to partners via electronic devices) and *real time* companies:

a) Networked organization system: different working groups communicate and cooperate directly with each other.
b) The company itself works as part of one or more network(s). In this case attention of management is directed to the external world, the establishment of external relations, alliance systems, partnerships becomes one of the most important management task.

c) Electronic sales channels of the company become available from any point of the world and resources are procured at the most favorable place.
d) Self-supplying is not a virtue: only those functions are built up which can be performed more efficiently or cost-effectively than other companies, or which are especially important from point of view of control or security.
e) The function is flexible, roles are often re-arranged, boundary lines between groups, functions and decision-making levels are not always clear. Decision-making system is decentralized which is allowed for and required by many factors: necessary information can be delivered quickly, easily and completely to any place; highly qualified professionals work at the operative level; the control is fast and efficient. The organization's structure is flat with small number of management levels, the traditional forward – sum up – report function of medium level managers is replaced by information technology applications. Within management roles the operative function becomes less important while so called *coach* functions are increasingly important. Planning is essentially not an "upside down" but rather a "down upside" process.
f) Customization is more and more typical of production: flexible systems are able to build different versions using well chosen components, according to customers' need.
g) In addition to "physical" devices (such as building, machine, raw material, etc.) the importance of information and knowledge is growing: the "intellectual capital" consisting of several components is an important part of corporate access.
h) Employees have no permanent, stable place within organization: they are frequently redirected, they wirk in different teams. With their electronic devices they can connect anywhere and any time to corporate network, and get the necessary information for starting their work.
i) In terms of stocks the company thinks not in weeks or months but in hours and minutes, the quantity of stock is minimal, procurement is of just-in-time nature.

It can be easily recognized that this model strongly differs from the company type formed in the 20th century which is ready to centralize and bureaucratize, separating control and execution. (Consulting older literature, this resembles perhaps to "adhocracy" described by Mintzberg [1979], the time of which has come owing to technological development.). Patterns in this regard are Toyota, Cisco and Dell: the first is the pioneer of mass customization and "electronics based lean management", the second gave alone one third of electronic commerce of the United States in 1997 while the third one is the personificator of an electronic business model which is eliminating agents and minimizing losses.

The question is now how real companies approximate the ideal type described above. In this regard the picture is rather mixed. It cannot be excluded that type is a point of attraction to which companies are heading but have difficulties in adapting their inflexible culture. It is also possible that development takes a different route. Information technology involves *full centralization* and *full decentralization* as well.

Social, political and institutional changes

Recurrent and important themes of works [e.g. Dyson 1998] dealing with the social consequences of information technology are as follows:

a) Communities. Modern telecommunications devices can contribute to the flexible organization of new communities with now limits while can facilitate the breakdown of others as well. This latter process is a latent one, its deepness and direction are hard to calculate, and they have serious dangers. From point of view of content the technology is neutral. The "social being" man of workplace communities which are disintegrating owing to modernization, streamlining and continuous "reengineering" is easy to integrate into other communities (hobby circles, sects, political parties and other movements) with the use of modern communications devices. Experiences and reports suggest that several information technology or Internet based communities show specific cultural signs and layered patterns [Lessard-Baldwin 2000] which can spread over to other fields as well through different channels (family, workplace, etc.).

b) Governance and politics, role of the state. There are heavy debates on whether the state may or should interfere in the development of e-economy and the underlying info-communications sector, and if the answer is yes, how. Public expenditures in terms of GDP have increased nearly everywhere. However, in the 70s the era of "information society" comes in with different views. During the past two decades it became a general view that the role of state should be limited and the rate of taxes and public expenditures reduced. Surprisingly, the system of state property and centralized planning was refused not only by conservative and neo-liberal political groups but also by socialist and social-democratic parties and they all preferred free competition.

Deregulation, market liberalization and strengthening of private property have surely played an important role in the proliferation of new technologies (just think of telecommunications sector). The Internet itself became the symbol of development without state intervention. We have just vague ideas on what the social, political and cultural life will be like in the era of "information society". States show serious interest in the electronic commerce and info-communications sector in several ways such as economics policy, investor, tax collector, regulator or user (electronic government, electronic democracy, etc.). In certain countries the associated lobbying activity is quite successful. In the United State, for example, vice president Al Gore heavily supported the program of construction of information

"super-highways" but the European Union and many countries have also "e-strategy". Some less developed countries – such as India [Bõgel 2001] – have great hopes in e-strategy.

c) Education. In course of the past years it became clear that innovation in information technology would not spread quickly unless the most important input – a great number of professionals able to use and develop the technology – is available. This, however, requires considerable changes in contents and methodology of education. It is also clear that the spread of electronic systems and massive use of advanced electronic services require more and more "computer literate" people. Electronization entered already the field of educational methodology: the increasing number of Internet based courses are special but most promising products of e-world, using the Internet for preparing the homework is an everyday action. The general level of education has increased, in the United States, for example, more than half of the adult population has a higher education degree.

d) Poverty and unemployment. In the 90s the United States saw the most rapid development of information technology sector and as a consequence – following the cutback period in the beginning of the decade – unemployment decreased considerably (see Table 1). Just before the stock market balloon went flat, the employment was full and there was even manpower shortage in certain jobs. This means that leading industries of the wave absorbed the available manpower. It is worth to look into the details. We can observe that ad hoc cut-backs jeopardized no so much blue collars as those who's work was easy to algorithmized – i.e. carried by computers – independently of the level of the required knowledge. Revenue statistics show that the income bracket extended in favor of highly qualified information technology and telecommunications professional, though there was a great spreading within this group [Lessard-Baldwin 2000]. In other countries these interrelations were less sharp. It can be stated, however, that in the United States the winners of the increasing efficiency were not investors but employees. This explains why, despite former similar events, economic downturn did not result in a considerable decrease in the consumption of the population.

e) Protection of intellectual property. This one of the most exiting legal issue of the information technology wave. Costs of copying and delivering of intellectual products – manuscript, music, software, etc. – decreased dramatically while delivery and distribution networks work with practically no control. This creates an obvious tension in the music industry where more and more heavy attacks jeopardize traditional products and distribution channels. The problem is a real threat, its legal, cultural and business implications are incalculable in the near future.

f) Security. Though connected directly and indirectly to advanced technology, it is not the terrorist attack in New York that brought out this issue. (Note that means of telecommunications "make sense" of terrorist attacks since they make events visible and "experienceable" for a mass of people guaranteeing thereby the necessary psychological effects. Members of terrorist groups exchange messages by means of mobile phones and the Internet, they place their messages on web sites and move their money through electronic channels.) During the past few years "white collar" computer crime has increased spectacularly for which legislators were not prepared. Some countries depend on their own technological systems so much that they have to take serious measures against "computer attacks".

g) Culture. Implications of information technology (computer, television, Internet, digital games, etc.) is spectacular and clear. In the development there are positive and negative trends as well, sometimes they go to the extremities. New arts are born (e.g. computer graphics), computerization creates a virtual world around us. Publication and promotion opportunities have been extended, "showing up" becomes more and more easy, meanwhile it is more and more difficult to separate value from void. An easily observable phenomenon is the thinning of intellectual content (remember the lot of television channels). Penetration of the English language seems to be stopless, providing a dominant position for the culture of Anglo-Saxon countries.

The above mentioned examples suggest that the information technology wave has had a profound effect on many aspects of economy and society, important and essential processes of change started, however, the "system change" is far from completion and even its directions cannot be calculated.

Slow-down and maturity (?)

The gathering of the last two phases as well as the question mark after them are to suggest that the rest of this article is mostly guessing: the picture is antinomic, not clear, the historical perspective is not enough for making an accurate analysis. No doubt, at the beginning of the new millennium there are signs of slow-down and economic downturn in the developed countries but it cannot be stated clearly that events follow the course of technological waves described above or we are witnessing something else. It is well-known that there were smaller waves in previous large waves, e.g. between 1950 and 1972 there were four "rapid" and five "slow" sections. Again, no doubt that the info-communications sector plays an important role in the current slow-down (it is enough to look at corporate business reports and stock exchange statistics). What we do not know is whether the problems are long-lasting or not.

In the following paragraphs two specific interpretations follow – with no commitment and not excluding other possible explanations and forecasts.

Scenario A: We are in the slow-down and maturity phase of information technology wave as events in the world suggest it. The further development of events will follow the pattern described in section 1. The era of revolutional technological changes is finished, some over-hyped inventions found cool reception. The era of two-digit growth rate is over, a wave of bankruptcy swept over the leading sector, a lot of companies disappeared, acquisitions and mergers are everyday deals. The industry is in the process of consolidation and the time is near when some large companies will dominate the oligopolistic market. Telecommunicaitons and information technology firms laid off a lot of employees, just in some special areas is there demand for manpower.

Profit rate decreased to a fraction of that of the former years, especially in lower segments of the value chain where products are becoming more and more undifferentiated mass articles. On desktop computers – the leading product of the wave – only those companies are able to make some profit which transfer the manufacture to low-cost countries of Asia. In the demand there are signs of set-backs. Market of mobile phones is nearing saturation, there are clear signs of a decreasing turnover. Most companies are over large information technology investments, they will spend much less for computerization in the future. This decrease can be clearly observed in statistics.

This set-back has taken short many companies (in 2000 the growth of sales of Nokia was near 50% while in 2001 it remained below 10%), business decisions were based on the golden age which proved to be a serious mistake. Some companies (like Cisco) were able to survive with temporary accumulation of assets while others ran into enormous debts, like most leading telecommunications companies *(Table 2).*

The "system change" associated with the wave has already started but will not go much further. In the coming period slowly changing, less flexible elements (law, culture) will fall into line with technology but impulses become weaker and no overturn is to come.

On the other hand, information technology sector has no special cause for anxiety: signs suggest that it will successfully co-exist with the next wave of innovation. The next wave may be that of health or biotechnology or perhaps energy, but one thing is sure: any of them will need information technology. The sector feels this adaptation trend so it is frequently emphasized that professionals need "vertical" knowledge related to the particular sector or industry. Superfluous manpower of information technology companies goes over to companies of the "old economy" having a positive effect on the latter. Surviving large information technology companies will be a sort of public service firms which are no longer in the center of the world but without which the world cannot exist.

Scenario B: Development is slowing but the information technology wave is in the rising phase and remains there for a while. Current problems are caused by the "humpback" imposed on the trend, i.e. the Internet fever. This latter has fulfilled its positive role: attracted a lot of capital and human knowledge to the sector, many ideas were tried and implemented, the infrastructure was built out quickly. Now the mania is over, the after-shocks of the Y2K problem are decaying, the downturn is temporary and the leading sector re-gains its dynamism. Growth curves will no more be exponential but continue rising sharply. Though some new innovation fields appeared – such as genetics and nanotechnology – but they do not start a new wave, rather give support to the growth of the existing wave and extend its rising phase.

Reserves of technological innovation are by no means exhausted. Moore's Law on the capacity growth of chips will remain valid for a long time. Number of Internet users is rising continuously, it's enough to have a look at the latest European figures. The penetration of mobile technologies is hindered by the huge debts of telecommunications companies and the lack of content but these are temporary problems and revolutionary changes are coming soon. We are just at the door-step of the "real time" world. Though with some delay bur the general increase of qualification and the change of generation will result in a development and user mass which is able to make use of the potential of new technologies.

Productivity figures provide good basis for optimism. In the United States in course of the 90s the annual average increase of this very important index [Bródy 1983] 0.5% higher than in the previous ten years. The average increase of productivity between 1995 and 2000 was 2.5% while it amounted to 3.3% in 2000 and attained 5.2% in fourth quarter of 2001. The productivity increased even when output decreased – this is unprecedented in the history of economic downturns of the past 50 years. Increasing numbers indicate that information technology started to heavily affect companies, it shows its abilities. If the "Internet humpback" from stock exchange trends is removed, we can see that business results of sectors backed by information technology (e.g. financial services, health) show a heavy increase between 1995 and 2002 [Mandel 2002]. However, information technology companies should learn that secret of their continuous success lies in increasing efficiency with customers.

The increasing productivity bears positive message for the world: purchase infor-

Table 2 Debt rate of telecommunications companies (June 2002)
Source: Newsweek, 10 June 2002, p.52.

Company	Dept/market value (%)
France Telecom	152
Deutsche Telekom	101
Nippon Telephone	68
AT&T	59
Sprint PCS	58
British Telecom	53

mation technology because it pays off! Opportunities are great: in the United States, the homeland of information technology just 60% of companies have implemented or are in the process of implementation of internet-based business solutions. In other countries the perspective is even wider. Currently the electronic commerce accounts only for a fraction of the global turnover. Maybe interested companies are unable to realize the added value coming from the increased efficiency but they can feel that without information technology they could fatally drop behind. The amount of money spend for information technology are really decreasing but curves are rising again *(Figure 2)*. Though slowly, inflexible company structures are changing and later everything gets going.

Since technological innovation goes on, "system change" does not stop either. The increasing rate of manpower related costs in overall corporate expenditures suggests that the most important resource is really human knowledge (in 2002 in the United States this rate attained a record level of 87%). The development of education, culture and social life will follow the development of technology so forecasts regarding information society are not Utopian at all. Underdeveloped countries connecting to the mainstream of global information exchange will be given a new chance for catching up: we can see the role of information technology in the re-accelerated development of "small tigers" in Asia.

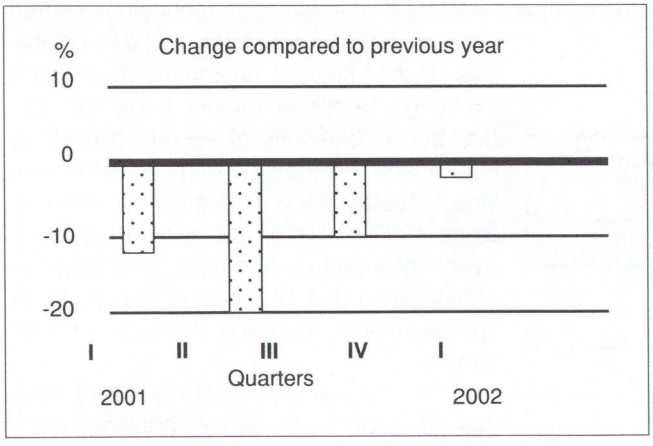
Conclusion

The history of information technology follows quite well the model of innovation cycles described at the beginning of the article. Typical features of former waves can also be observed in the implosion and growth phases, but there are some new peculiarities as well. It can be seen, for example, that:

- owing to network effects, advanced mass communications and other factors growth curves are very steep, the acceptance of technological novelties is faster than in previous waves;
- the wave is global almost from the very beginning;
- limitations of growth are of not financial character: knowledge is the most important resource which is available without limitation in the given period, but otherwise is of different character than financial resources;
- similarly to other waves, in the implosion and growth phase also irrational elements, manias and fashions appear, the "additional wave" caused by these latter factors show very sharp increase and decrease.

We can state that information technology has brought about "system change" but this process is not yet completed and it may have different outcomes. Current slowdown phenomena can be interpreted in several ways. Depending on case analysis and forecasts the expectations regarding the duration of information technology cycle can be different. Anyway, the sector has very good chances to "adapt" to the next cycle and to co-exist and develop with new leading industries.

Figure 2 Development of expenditures on information technology equipment (USA)
Source: DRI-WEFA, Business Week, 13 May 2002, p.15.



References

[1] Az emberiség krónikája (1990) Officina Nova
[2] A technika krónikája (1991) Officina Nova
[3] Baldwin, N. (2001): Henry Ford and the Jews. Public Affairs
[4] Beevor, A. (2002): The Fall of Berlin. Viking Press
[5] Benedek Zoltán (2002): Nyakkendõs bányászok. CEO, április és június
[6] Berend T. Iván (1982): Válságos évtizedek. Gondolat Könyvkiadó
[7] Bõgel György (1999): Miért a Netscape? Vezetéstudomány, szeptember
[8] Bõgel György (2001): Buddha mosolyog. Az indiai szoftveripar sikereirõl. CEO, október
[9] Bródy András (1983): Lassuló idõ. Közgazdasági és Jogi Könyvkiadó
[10] Chandler, A. (1969): Strategy and Strucure. MIT Press
[11] Christensen, C. (1997): The Innovator's Dilemma. Harvard Business School Press
[12] Christensen, C. (1999): Innovation and the General Manager. Irwin/McGraw-Hill
[13] Dertouzos et al. szerk. (1989): Made in America. MIT Press
[14] Drucker, P. (2002): Managing in the Next Society. St. Martin's Press
[15] Dyson, E. (1998): Release 2.1. Broadway Books
[16] Freeman, C. – Louca, F. (2002): As Time Goes by. Oxford University Press
[17] Ford H. (1989): Today and Tomorrow. Productivity Pr
[18] Galbraith, J. (1997): The Great Crash 1929. Mariner Books
[19] Galbraith, J. (1978): The New Industrial State. Houghton Mifflin
[20] Gates, B. (1995): The Road Ahead. Viking

[21] Goldenberg, B. (2002): CRM Automation. Prentice Hall
[22] Greiner, L. (1995): Evolution and Revolution as Organizations Grow. Harvard Business Review, február
[23] Greguss Ferenc (1985): Élhetetlen feltalálók, halhatatlan találmányok. Móra Könyvkiadó
[24] Grove, A. (1997): Only the Paranoid Survive. HarperCollinsBusiness
[25] Heizer, J. – Render, B. (2001): Operations Management. Prentice Hall
[26] Kerekes Tibor (2002): Biztonságos (?) hálózatok. Alma Mater, Budapesti Műszaki és gazdaságtudományi Egyetem, augusztus
[27] Kindleberger, C. (2000): Manias, Panics, and Crashes. John Wiley & Sons
[28] Kocsis Éva – Szabó Katalin (2000): A posztmodern vállalat. Oktatási Minisztérium
[29] Kuhn, T. (1996): The Structure of Scientific Revolutions. University of Chicago Press
[30] Lessard, B. – Baldwin, S. (2000): NetSlaves. McGraw-Hill
[31] Lowry Miller, K. (2002): The Giants Stumble. Newsweek, július 8.
[32] Mandel, M. (2000): The Coming Internet Depression. Basic Books
[33] Mandel, M. (2002): The Boon behind the Bubble. Business Week, 2002. július 15.
[34] Michaels, E. – Handfield-Jones, H. – Axelrod, B. (2001): The War for Talent. Harvard Business School Press
[35] Mintzberg, H. (1979): The Structuring of Organizations. Prentice-Hall
[36] Ono, T. (1988): Toyota Production System. Productivity Pr
[37] Perez, C. (2000): Technological Revolutions, Paradigm Shifts and Socio-Institutional Change. In E. Reinert szerk.: Evolutionary Economics and Income Equality. Aldershot: Edward Elgar
[38] Ranadivé, V. (1999): The Power of Now. Osborne McGraw-Hill
[39] Samuelson, R. (2002): The Media's Heavy Hand. Newsweek, július 1.
[40] Schein, E. (1997): Organizational Culture and Leadership. Jossey-Bass
[41] Schultz, T. (1971): Beruházás az emberi tőkébe. Közgazdasági és Jogi Könyvkiadó
[42] Schumpeter, A. (1980): A gazdasági fejlődés elmélete. Közgazdasági és Jogi Könyvkiadó
[43] Shapiro, C. – Varian, H. (1999): Information Rules. Harvard Business School Press
[44] Polónyi István – Tímár János (2001): Tudásgyár vagy papírgyár? Új Mandátum
[45] Weill, P. – Vitale, M. (2001): Place to Space. Harvard Business School

News

E-Health is the use of telecommunication tools, coupled with medical expertise, to deliver diagnostic, therapeutic and educational services to individuals living some distance from medical facilities. However solutions currently available have often been developed on a proprietary or ad-hoc basis posing implementation challenges to the agencies involved. Standardization in e-health is seen as a way to increase levels of interoperability, but has so far not produced much in the way of implementable procedures. The workshop aims to bring together key players to discuss how to define a framework for standardization in e-health and take that forward. Sessions will provide a combination of case studies highlighting end-user requirements and discussion on the technical issues. *(ITU Headquarters, Geneva)*

•

COLT has launched end-to-end, international Ethernet services over SDH – and claims to be the first company in Europe to offer local area networking services across national borders. The company is offering two Ethernet services: „EuroLANLink" and „CityLANLink", both of which allow businesses to provide point-to-point connectivity between sites in any of the 32 European cities in which COLT is present.

•

According to research consultancy **TeleGeography**, the undersea cablebuilding boom has finally come to a halt. The recent completion of trensoceanic systems by **Tyco Telecom** and **Cable and Wireless** marks the end of a construction craze that led to a 30-fold increase in communications capacity across the Atlantic and Pacific from 1998 onwards. TeleGeography says that USD3bn is earmarked for new submarine cables entering service in 2003-down from USD13bn in 2001. Even though the end of the building boom should help in reducing the current glut of submarine capacity, TeleGeography suggests that continuing pressures on pricing and a slow market for the purchase of subsea capacity means that submarine cable operators still face tough times in the next couple of years. Unless demand grows or sharks bite...

Centenary of dr. Ladislas Kozma (1902-1983)

GYULA HORVÁTH, *Consulting engineer*
e-mail: horgyul@hdsnet.hu

Dr. **Ladislas Kozma** was a great man in the Hungarian telecommunications trade. He started working at the Hungarian United Incandescent Lamps Limited as electrician. He was very intelligent and industrious, in his free-time, he studied intensively the circuits of the (electro-mechanical) telephone exchange, he translated from English its technical documentation. The directors discovered rapidly his talent, his serious personality and his assiduity in work.

Being a son of financially not well established parents, the leading managers of the Company clubbed together to cover his studies on the German Technical University at Brno. After graduation, at 1930, he was invited to Bell Telephone Manufacturing Co. (BTM), Antwerp. This was the European headquarter of the mighty International Standard Electric Co., where the famous Rotary switching system has been developed. There he took part in the further development of this system, in planning a number of telephone exchanges produced there and studied the theory of traffic. His ingenious activity resulted in 27 patents holding his name as author or co-author.

In the year 1938 he was entrusted to form and lead a research group devoted to the elaboration of a computer, composed of standard components (relays, etc.) of the actual automatic telephone exchange. The computer No. 1 was very slow, did not included memory, it was not capaz to perform a division. Computer No. 2 was faster, it produced a quotient/sec performing a division. Another ten patents bear witness to this product. In 1940, this computer was sent to the USA but did not arrived, supposedly due to some war event. In 1942 dr. Kozma returned to Hungary, where the anti-Jewish laws allowed him to work only as a technician. Later he was deported to forced labour. Fortunately, he escaped extermination and returned to home in mid-1945. He joined immediately Standard Electric Company, Budapest, a subsidiary of International Standard Electric Co., New York. As the chief of the Telephone Engineering Department, his goal was to develop the enterprise up to world-class level. He personally teached the young engineers (the author is one of them) and technicians to the theoretical and practical knowledge of the automatic telephone exchange and to the English technical language.

He disputed thoroughly traffic theory and made clear that this is a mathematical model of telephone traffic. Grading was at that time in the focus of investigation. While others tried to elaborate a formula useful when grading is calculated, BTM constructed a simulator for this purpose. The big tables obtained by simulation, was multiplied and distributed by dr. Kozma among the young engineers he instructed.

Similarly he treated mathematical logic (Boolean algebra) as a model of the operation of digital resources. He teached us to apply it, instead of designing relay circuits intuitively, applying only some basic relations as e.g. „two make relay contact in series realize a logical AND relation".

Dr. Kozma practised our English on letters too, from which we obtained an insight into the correspondence between the company and BTM and learned how to arrange some everyday problems. Following the practice of ISEC, he made frequently a tour in the laboratories and at the drawing-boards visiting each employee. One colleague remembers of what impressioning his visit was, how much he profited from it.

As far as the future was concerned, Dr. Kozma introduced us into the crossbar technics, he explained the working of ENIAC and the expected capabilities of eletronic computers. For example, he indicated that the electronic computer may be used to compute the whole payroll of the company. This looked as a dizzy perspective. In addition to the scientific and technical details, he mentionned examples for cultural, social, economical and political issues related to telecommunications.

Simultaneously he took part in the foundation of the post-graduated education of technicians and in the foundation the chair of telecommunication at the Budapest Technical University. He was its first professor and among thr first winners of the Kossuth-prize, founded in 1948. Every year, this prize is awarded to those who rendered the best qualified contribution to the progress of the nation.

He capitalized his ample experience helping the company in meeting the extraordinary requirements the telephone exchanges to be delivered as war damage compensation must fulfill. Farther he elaborated a cordless toll exchange, doing its system engineering work and designed circuits. It was a matured member of the famous Rotary system, developed by BTM. It was manufactured till 1970.

Some months after having been decorated with the Kossuth-prize, he was imprisoned for 5 years as a result of a show trial. Two decades later, he wrote a book on this period of life. It is a shocking story, especially the description of the incertitude in which he was continuously kept. During the last two years of this pitiless period, an office was arranged in the prison where he, together with other condemned engineers worked. He got tasks also from the nationalized Company. One of them was the design of a test-circuit. As an annex, he delivered a list of errors he found in the circuit to be tested. His former apprentices appreciated this gesture well, saying „he is teaching us even from the prison". Indeed, various problems of telephone exchanges quasi engrossed his thoughts during the deportation and later in the prison as well. This way of thinking contributed very much to his survival.

After having been released in 1954, he was rehabilitated and continued his service as university professor. The then existing political and economy system in Hungary as an environment did not supported his endeavour to reach world-class level. At first he concentrated his efforts to teaching. With his teaching staff, he elaborated the mature content of the teaching-line of telephony. As a result, he and the staff wrote a full set of notes for the students.

He included the economic bearings of telephony in his course: he declared emphatically that the user must be regarded as the most important person, in his interest are working the telecommunications equipments and everybody, who participates in its production.

He considered illustration as an important adjunct to the lecture. As an attractive example one cannot neglect that supported by the Hungarian Academy of Science, he constructed and built an electromechanical digital computer, proceeding on his way. This had an important didactic value: due to the relays and other components of the Rotary system, the computer was relatively slow, so the students were able to follow its operation by looking on. There was keys to stop the computer in order to observe its state. Many of his students obtained a basic knowledge on computer technics, some of them choosed computer technics as occupation.

He was very keen in transferring knowledge: anybody from his staff or any of his students reached him for consultation at any times when not occupied. He followed the work of his staff encouraging each member to do his best and helped them in proceeding on their scientific progress. He aimed and attained that every member of his staff teach well, do research succesfully, and be able to apply his skill in the practical life too.

He was keen also in making his students familiar with practical engineer's work. He extended this endeavour to his staff, recommending to find a temporary appropriate part-time job and assisted them in landing the job. He used the exams as an event of crucial importance for education. He tried patiently to manage to prove the maximal level of knowledge of the student examined. He was a demanding, but a just examiner. Once happened that an unprepared student tried to get dr. Kozma to let him pass the examen. He asked:"Why is this so important for you?" The student answered: „My assignement depends on it." „Why is your assignement important?" „For the sake of my salary." „If you wish earn many money, then you would better be a football player." „I am not expert in it." „In this subject neither" was the prompt answer.

He expounded his view everywhere and at all times frankly. Consequently, he did not get any official reward for the realization of his electromechanical computer, nor his sharp critique on the state of the Hungarian telephone network met goodwill. Inversely, he deserved by this work a posthumous (1996) credit issued by the IEEE Computer Society recognizing Ladislas Kozma as a „Computer pioneer for development of the 1930 relay machine, and going on to build early computer in post-war Hungary."

The third area where he was also a pioneer, is speech processing. Working on this field, he construed an automaton produced language statistics capable to perform 80 statistical tasks simultaneously.

He organized visits to important telecommunications plants (factories and operators) for his students, he himself often accompanied them. He emphasized there the significance of the construction and technology. By other cases too, he tried to get into personal acquaintance with his students. When consulting, he explained all details patiently and took care for that no question remains unanswered. His most favoured event was the gatherings of the Scientific Student's Association Conference, where he got round to everybody for conversation. These meetings were very popular among the students.

Initially, when the staffs of the chair were small, family meetings was organized in rotation at the staff members. There revealed itself his tender humour. His jokes never offended anybody. He liked arts, and music. He had a fair collection of disks.

He reiterated his efforts in order to improve the poor telephone service in Hungary. He wrote articles, delivered lectures on his matter. As some acknowledgement of his scientific activity, in 1961 he was elected to be an associate of the Hungarian Academy of Sciences, and in 1976, he was elected to be an ordinary member of the Academy.

The impression of dr. Kozma on his environment, yielded in that his pupils and followers are feeling today that they all are members of a scientific-engineering school, created by his personality. They are distinguished by their open-minded approach to engineering problems, their breadth of view and their method of work. Those, who have been instructed at Standard Electric Co., are now over 80 (if alive), while the younger former pupils are working successfully in today's Hungarian governmental institutions and private enterprises. Dr. Kozma created another scientific-engineering school in the computer-technic. A number of generalists and specialists in the Hungarian computer science, operators and leaders made once acquaintance with computer technology at the electromechanical computer of dr. Kozma. All these colleagues from two generation, are indebted to their unforgettable master, whose centenary was celebrated on 28. November 2002.

The celebration took place at the Budapest University of Technology and Economics, in its newest and biggest lecture room, which was named after him by this opportunity. Near to its entrance, the bust of him was unveiled. Also a commemorative plaque was fixed on another building of this University, where he founded about half a century ago the chair for telecommunication and where he created his electromechanical computer (MESZ1).

The body of the Memorial Day was composed of 3 verbal commemorations, presented by the professor of his chair, who characterized him as the creator of a scientific-engineering school, and by a representative of the Hungarian telecommunications industry, another of the Hungarian computer industry. They recalled a number of personal experiences with him. Thereafter spoke five lecturers, summarizing some aspects of his life-work: infocommunication networks, signalling systems, protocols, speech processing, informatics, progress in telecommunications, giving account of the actual state in Hungary of the disciplines cultivated by him. In the memorial issue (2002/11) of this periodical of the Scientific Association for Infocommunications, the extended verbal commemorations and the lectures has been published, followed by four scientific publications dedicated to the memory of dr. Kozma. These last publications gave also an overview, dealing digitalisation, mobility and up to date traffic theory, and some aspects of the history of telecommunication in Hungary have also been mentioned.

The attendants on the Memorial Day paid due homage to his engineering, scientific and pedagogic life-work, to his persistence and other human virtues.

Editoral Office (Subscription and Advertisements):
Scientific Association for Infocommunications
H-1055 Budapest V., Kossuth Lajos tér 6-8.
Phone: +36 353 1027, Fax: +36 353 0451
e-mail: hte@mtesz.hu • www.hte.hu

Articles can be sent also to the following address:
BME Department of Broadband Infocommunication System
H-1111 Budapest XI., Goldmann György tér 3.
Tel.: +36 463 1559, Fax: +36 463 3289,
e-mail: zombory@mht.bme.hu

Publisher: MÁRIA MÁTÉ • Manager: András Dankó

HU ISSN 0018-2028 • Layout: MATT DTP Bt. • Printed by: Regiszter Kft.