híradástechnika

VOLUME LVIII.

2003/12

Selected Papers

Network Design Guidelines

Photonic Devices

Evaluation of Quality

Scientific Association for Infocommunications

Contents

Cover: Different articles arranged in a suitable manner shape a harmonic view

Foreword

Every four years Geneva is a meeting point for players of the world of telecommunications and information technology: manufacturers, service providers and often even users. Participants get insight into some important events of the past four years and have impressions of what kind of new devices or services interested players will introduce to the market. Last October ITU organized World Telecom where visitors could not get surprisingly new information neither of the past nor of the future. This may be attributed to the recession but there was a noticeable rearrangement. Important companies of past events were not present and did not present any novelty. At the same time the role of manufacturers and service providers of the Far-East has increased. The presence of these new market players proved that they learned all they could in professional literature and which are determinant elements of today's telecommunications. In this way our special edition in English cannot publish articles which are fundamentally new and lead readers to new fields.

In part recession, in other part the 100th anniversary of the birth of John von Neumann have highlighted risk analysis and game theory. In order to guarantee a successful future service providers review development options and trends. If they want to take risk minimization into account and to survive in competition, they have to rely on relevant mathematical methods. Our first two articles are translations of studies written for the Neumann Centenary. Evolution and practical application of game theory can result in economical success.

The second theme is application of optical fibers. Photonics is not a new technology. George Gilder writes, however, that the huge transmission capacity of optical fibers changes economic relations of telecommunications networks and therefore new devices should be applied. The transmission cost of bits has dramatically decreased so there is no point in saving bandwidth, instead, our primary task is to sell this capacity. This is the central idea of the fundamental paper by György Takács. The success of the application of optical fibers depends greatly on the availability of the necessary fittings. We review the solutions which are either ready or under intensive development. Our authors write about optical switching, broadband amplifiers and about amplifiers based on the Raman principle. This photonics block reports on current abilities but with a view of capabilities of future service development.

The next block introduces three theoretical curiosities. It is surprising and straightforward that noise perception is not always the same as reading on measuring instruments. In his article professor Illényi characterizes this subjective perception by objective indicators. Video compression will play an important role in broadband contents, particularly in the distribution of entertainment services. Contour recognition is a promising technology in this field. This group includes also the mathematical management of textual functions which facilitates the machine based exploration of inherent contradictions or disturbing deficiencies in laws and patents. Finally our periodical offers an article on a typical feature of IP systems.

Summing up one could say that with a wide-spread utilization and economical application of existing devices the industry tries to promote the proliferation of telecommunications products and to enhance services. This is in accordance with the experiments of World Telecom in that we are looking for ensuring the way of continuous development in an era when telecommunications needs no more are several orders higher than the growth of GDP and there are no astonishing innovations every week. At present this sector merges into the general trend of growth and development. With this reality in mind we do our best to ensure the success of the program for the next year.

Dr. György Lajtha

A playful strategy

GYÖRGY BŐGEL

Strategic consultant of KFKI Computer Systems Group,
teacher of the Business School of the Central European University (Budapest),
associate professor of the University of Debrecen
gybogel@kfki.com

Keywords: Strategy, Telecom economy

John von Neumann was dealing with several aspects of applied mathematics and economics. Models of game theory to be found along with border line of economics, mathematics and management can well be used with strategic decisions. In the life of economics, particularly in the field of telecommunications and information technology, a lot of cases can be described and analyzed with the use of game theory. However, this theory has its limitations as well.

John von Neumann is described as a typically cheerful and active society man who organized certain kind of party every week with Klara Dan, his second wife. He stored a lot of jokes in his fantastic memory (according to an anecdote he was able to learn even the telephone directory) and he could any time recall one to vitalize the party. He also liked to play poker though he was not really very good at it. Winner or loser, it does not matter for us: the main point is that he was thinking in course of the game and came to the conclusion that the outcome of the game depends not only on probabilities, i.e. poker is not (only) a *game of chance* but can also be a *game of strategy*. As a mathematician, he started to formalize "bluffing" i.e. the strategy aiming at cheating the other players or at hiding relevant information.

War games

Before von Neumann poker has already inspired another scientist. In 1921 Emile Borel French mathematician published studies on the *theory of games* with the forecast that game theory would play important role in economics and war strategy. He was most interested in the fact whether there exists a *best strategy* for a particular game and if so, how it can be found. This approach was quite a new one but he did not get very far in game theory. That is why most science historians consider von Neumann as the first developer and popularizer of game theory.

Von Neumann published his first article on game theory in 1928 [Neumann 1928]. He also shared the view that this new theory will have a bright future in economics. In 1929 he got a job in Princeton where – together with the Austrian Oskar Morgenstern – he wrote the book [Morgenstern–Neumann 1990] which is taken for as the basic work of game theory. Though this book was written for economists, it became clear soon that it can be quite well used in psychology, sociology, war strategy, sports and many other fields as well.

The relation of John von Neumann with *game theory* thereafter turned out to be very interesting and controversial. Though he was aware that this theory would revolutionize economics, he personally was more interested in *political and military applications*. This could be attributed to "Kriegspiel", a military simulation game very similar to chess which was one of his favorite games in his childhood. At the outbreak of World War II he described the conflict by game theory methods and forecasted the victory of the allied powers. In 1943 he was invited to join *Manhattan Project* dealing with the development of nuclear bomb. His calculations played important role in the project but his models were used also for other purposes such as finding the most secure path for bombers or identifying opposing cities as targets.

In year 1948 the scientist became consultant of RAND Corporation. This organization was founded by military industry companies together with the air force in order to "imagine the unimaginable", e.g. to calculate the possibilities of a *nuclear war* and to develop strategies for this event. That time Neumann was a committed follower of the "pre-emptive strike". He was sure that Soviet spies have already acquired the information necessary for the production of a nuclear bomb and it's a question time when the Soviet Union would become a nuclear power. He thought if Russians are able to build their nuclear arsenal, their war against America cannot be avoided. Based on this theory he proposed for the USA to make a pre-emptive strike on Moscow thereby disrupting the enemy, avoiding a more tragic war and becoming a dominant world power. As one of his often cited saying in the Life Magazine goes: "If you ask whether we are going to bomb them tomorrow, I ask you: why not today?"

Soon after its birth the idea of „pre-emptive war" became impossible. In 1953 the Soviets had much more than three hundred torpedo heads so they could give an efficient riposte to any nuclear strike.

In 1954 John von Neumann became member of the Atomic Energy Commission of the USA. One year later bone cancer was diagnosed with him which – according to the book of William Poundstone [Poundstone 1993] was the consequence of the radioactive sand inbreathed during the nuclear experiments on the Bikini-islands. The ill-

ness did not reduce his activity, not even later when he was able to move in wheel chair only and received the representatives of the air force and the Department of Defense in a sick-ward. This might be the direct reason why many think he was the model (one of the models?) of the odd German wheel-chaired scientist in the movie "Dr. Strangelove", the brilliant, satiric and provocative work of Stanley Kubrick which came to light in 1964, just after the assassination of president Kennedy, the time of the "coldest" war.

It is not by chance that we put this story at the beginning of an article written on the subject of *corporate management*. Today the idea of a "pre-emptive nuclear war" against Moscow sounds insanity. For John von Neumann and some other scientists, however, it was a rational answer to a fatal dilemma in a poker game where cards are replaced by existing and non existing rockets, neither party has trust in the other and never knows whether the other says the truths or is just bluffing. A lesson of the story is that mathematical rationality, rationally built models and decision-making schemas have a great importance but they have also their limitations.

Games in the economy

Putting aside rockets now, we can observe the same interesting relations in the field of economy: from time to time the management profession takes to rational, mathematically based *quantitative methods,* this love, however, begins to decrease in intensity and then gets momentum in another field.

The 1994 Nobel-prize in economics of John Nash, John Harsanyi and Reinhard Selten is a clear indication of the fact that today game theory is an organic part of *economics.* It is also easy to recognize that the competition between companies is in several aspects similar to chess or card games: there are situations in which players make decisions, other players react and so on. Company leaders would like to know the same as Emile Borel or John von Neumann: what is the game we are playing, is there a winner strategy in it, and if so, how it can be identified. They know that the outcomes of their decisions depend also on decisions made by others since any action is followed by a counter-action. The relevant book by Ghemawat Pankaj [Pankaj 1997] is a kind of repository of examples. Strategic decision-making problems are formed in several industries which then are analyzed by the methods of game theory. He concludes that in the 70s and 80s the interest in the theory had highly increased, more and more models were built and tested for particular cases. Pankaj thinks that game theory is having considerable influence on *corporate strategies* and/or on the process of *strategy forming*. He points out also that certain trends within strategy forming are close relatives to game theory, their essence can be described with the tools of the latter.

However, there is an interesting contradiction here: books of management literature written on strategies generally do not even mention the expression "game theory". If we open the "encyclopaedia of strategy" by Henry Mintzberg and his co-authors [Mintzberg et al. 2003], we cannot find game theory in the index. This seems to justify the saying of the Nobel-prize winner Herbert Simon: the abstract and mathematics-based "management science" and the poor "management profession" relate to each other as oil to water. They can be mixed with bold movements but when left alone, they get separated again [Simon 1991, p.146].

Nevertheless, Pankaj is right: game theory is really a very useful tool for the understanding and handling of a lot of management situations and strategic problems: it can be a bridge connecting *economics, mathematics and management*. Fortunately there are some researchers who work on building this bridge. Proceeding from game theory and economics they come to conclusions which may attract company managers and can be formulated with terms of practical management.

According to Adam Brandenburger and Barry Nalebuff [Brandenburger-Nalebuff 1997] game theory gives answer to the most acute management problem: how can we find out the suitable strategy and how can we make a *good decision.* This is especially useful when many mutually related factors have to be taken into account, i.e. when no separate decisions can be made. Today's business is exactly such an environment: its *complexity* is more and more embarrassing. Success or failure can be influenced by factors which decision makers do not even think of. The effects of a change traverse the complex network of people and organizations in a very short period of time. When game theory decomposes games into their components, it can offer a guide in this jungle, it can help in understanding and explaining decision options, identifying possible and chosen strategies, exploring and comparing decision alternatives. As Brandenburger and Nalebuff say: *game theory is a way of thinking.*

Competition and co-operation

Games are characterized by *conflicts* – say Jenő Szép and Ferenc Forgó in the preface of their work on game theory [Szép–Forgó 1974, p.11.]. Generally players with contrasting interests get into conflict in games. It is the interest of each player to guarantee an outcome which is favorable for him with given rules. In course of the game the behavior of a player in a particular situation is determined by its *strategy.* This strategy can be seen as a plan or coherence of behavior which serves as guidance in emerging situations. During the game players can study the behavior of each other, can learn from it or can even *form coalitions* – a thing which is most actual and interesting in today's economic applications.

There are namely co-operative games, too. Some players may give up the independence so that the cumulative benefit of the group be greater than the sum of benefits achieved separately. Co-operating players coordinate their strategy, i.e. their decisions are made not independently.

This means that in games there is *competition and co-operation* – and now we arrived at one of the most interesting problem of corporate management. A well functioning company produces value but it generally does it not alone but as a member of certain labor division system, in cooperation with others. Its own profit depends on two things: on one hand on the profit of the whole system, on the other hand on the proportion it gets from this profit. When speaking about the aggregate profit, other members of the system are considered as *allied,* when speaking about its own share, the same members are *rivals.*

The decision-making problems occurring in such situations are perfectly represented in the classical game called "prisoner's dilemma". In the original story two doers of a robbery are captured by the police but there is no evidence against them. They are kept in two isolated cells. Detectives offer a bargain: if one of them testifies against the other, can freely go away and the other will be seriously punished. If neither of them testifies against the other (i.e. they are cooperating), they get a smaller punishment since there is no real evidence against them. If both of them testify against each other, they will be punished, but less seriously then if only one would testify. The dilemma lies in the fact that each of the prisoners can choose only from two decision alternatives (to testify or not to testify against the other) but they cannot make a good decision unless they know the intention of the other.

This type of sharing of benefits and losses can be found in many *economical situations.* The party ready to cooperate loses if its gesture remains unanswered. The game is created so that cooperation (or synergy) involves less personal benefit than one-sided delation, this latter solution being thus more attractive. The assumption that the benefit of cooperation is less, is not always true in the economy but considering that its benefits present generally on a longer run, the situation becomes realistic. The prisoner game can be a good model for short-term decision-making situations where players have no plans and ideas for a future cooperation.

Let us notice one more thing. If in the given situation both prisoners behave *rationally,* there is chance for the cooperation. Rationality means to chose to best alternative, independently of the decision of others. If the other delates, it's better for you to delate as well. If the other does not speak, the delation pays better. Consequently, if both players are purely rational, both testify and either of them wins. If, however, they decide "irrationally" to cooperate, the situation becomes easier for both of them and their punishment will be minimal. In other words: independent optimization of *parts* brings less result than the optimization of the *whole* system.

As already mentioned above, in the economy there are a lot of situations where one has to decide between "rivalry" and "cooperation", to think over common and individual benefits at the same time, i.e. where game theory models similar to the prisoner's dilemma can be especially useful.

No doubt, members of *value chains* connecting suppliers, manufacturers and vendors are rivals to each other since they have to share the profit of the whole chain. If one of them is able to reach a better position (e.g. can purchase at a lower price from his supplier), he will have a greater slice of the cake which is limited ultimately by the purchasing inclination of the customers of the chain. The individual rationality suggests the following pattern: create competition among your suppliers and monopoly against your customers. In this competition a successful individual action can result in substantial profit, a greater share in the cake of the chain. If members of the chain cooperate, they limit their independence, on the other hand, owing to cooperation, the cake can become bigger, since the efficiency of the chain increases, plans become more accurate, costs and assets decrease and servicing cycle times get shorter. How much rivalry and how much cooperation is required to achieve the optimum? What is the relation between individual and group benefits? This is the dilemma of the supply chains management which can be modeled as a game. It should be noted that it is the advanced information technology and telecommunications that make *supply chains management* viable even at global level.

Let's take another example: the outsourcing of corporate activities. In this game outsourcer and service provider confront. On short-run and seen from a narrow perspective they play a zero-amount game: a profit on one side is the same loss at the other; if the outsourcer reduces service fees, the difference remains with him as profit and presents loss at the service provider side and vice versa. On a longer run, however, with suitable cooperative behavior they are able to make the common cake bigger, to produce more value together, than as rivaling companies. Similarly, with rivalism they can even bring into danger each other.

In a given situation it is difficult to decide who is rival and who is allied, sometimes the same player can be rival and allied as well. This is suggested by the strange title of the book by Brandenburger and Nalebuff: „Co-opetition" [Brandenburger–Nalebuff 1997], which is obviously created from words *co-operation and competition.* In their model economical value is produced in a multi-player network. Companies of this network are sometimes rivals of each other, in other situations they are allied or complementing entities. On the one hand they fight for a bigger individual share, on the other hand – when cooperating – they work together for a bigger cake.

This becomes most clear when the development of a company is linked to the development of firms which are considered as rivals. A young and developing market, such as today's telecommunications sector, shows a lot of examples for this phenomenon. Remember the so-called network effect: the more customers connect to a network, the more it is worth. This means that a telecommunications company may benefit from the development of another since this involves more communications opportunities for his own subscribers as well. The sector is regularly facing the problem that *additional (complementing)* businesses do not develop in the necessary pace: e.g. there is a network but there is not enough content to distribute on it. Content industry, information technology, telecommunications: this three-player game is a classical example of "co-

opetition", left alone other players like people, state, companies, suppliers, etc. Development of the market as a whole (i.e. the cake to be shared) is at least as important as the individual benefit from the collective profit of the sector. Here the development of a strategy needs thinking in a complex eco-system which can be modeled with the tools of game theory. Obviously the prisoner's dilemma may occur as well: is it worth to cooperate if I do not know how the others behave? The example of telecommunications is a good indication of the fact that competing behavior may lead to a result which none of the players requires [Rosenbush et al. 2002].

Limitations and risks

As already pointed out, game theory can be a very useful tool for decision makers in the economy. However, looking back to the start of this article, we should be cautious. We know well that tools like operations research, mathematical statistics or game theory have produced noteworthy results in the field of military strategy. Andrea Gabor presents a detailed description of how the American army was put straight during and after World War II by persons with outstanding mathematical capabilities like Robert McNamara and his team. We know also that the Vietnam War was supported by highly developed modeling techniques, statistical apparatus and methodically controlled logistic systems.

Despite all these factors America was unable to win this war. Regarding the reasons opinions greatly differ but it became clear that advanced optimization models were often and deliberately fed with distorted information and war field reports, on the other hand, mathematical tools were unable to manage important factors such as dauntlessness of the enemy. McNamara who reached as far as the post of president of Ford in his industrial career could even observe how his controlling system based on quantitative indicators was bested by workshop managers, how measurable qualitative indicators divert attention from non-measurable indicators or how the overstraining of cost-benefit analysis can lead to a series of human tragedies like inflaming Ford Pintos in the early 70s.

From observations of Herbert Simon based on experiments – and from our own experiments – we know that *human rationality has limitations* [Simon 1997]: people are not as perfectly rational beings as they are described by some theoretical economists. The number of options, decision alternatives is practically infinite – instead of optimization people often are satisfied with an acceptable solution or in case of decision-making they simply rely on their habits and follow already proven patterns.

Corporate decision-making is a multi-step and multi-player process with different interests, formal and informal agreements, professional discussions and fights for power. Probably just this complexity and diversity excited John von Neumann who was interested in nearly all aspects of life. This is how in a taxi in London he explained the essence of game theory to Jacob Bronowski who worked with him during World War II (cited in [Poundstone 1993, p.6]):

Since I was an enthusiastic fan of chess I asked him the following question by habit: "Game theory is something like chess, isn't it?" "Oh no", he answered, "chess is not a game. Chess is a well defined form of calculations. Maybe you cannot give answers in practice but in theory there must be a solution, a right technique in every situation. Real games are not like this. Real life is different. In real life there is bluffing, doodling, and considering what the other think I am going to do. Just these are the players in games of my theory."

References

[1] Bőgel György – Forgács András (2003): Üzlet, vezetés, informatika. Műszaki Kiadó

[2] Brandenburger, A. – Nalebuff, B. (1997): Co-opetition. Doubleday

[3] Dixit, A. et al. (1999): Games of Strategy. W.W. Norton and Company

[4] Dixit, A. – Nalebuff, B. (1993): Thinking Strategically. W.W. Norton and Company

[5] Forgó – Szép – Szidarovszky (1999): Introduction to the Theory of Games. Kluwer Academic Publishers

[6] Gabor, A. (2000): The Capitalist Philosophers. Three Rivers Press

[7] Mintzberg et al. (2003): The Strategy Process. Prentice Hall

[8] Morgenstern, O. – J. von Neumann (1980): Theory of Games and Economic Behavior. Princeton University Press

[9] Nasar, S. (2001): Egy csodálatos elme. GABO Könyvkiadó

[10] Neumann, J. von (1928): Zur Theorie der Gesellschaftsspiele. Mathematische Annalen

[11] Pankaj, G. (1997): Games Businesses Play. MIT Press

[12] Poundstone, W. (1993): Prisoner's Dilemma. Anchor

[13] Rosenbush, S. et al. (2002): The Telecom Depression: When will it End? Business Week, október 7.

[14] Salamonné Huszty Anna: Jövőkép- és stratégiaalkotás. Kossuth Könyvkiadó

[15] Simon, H. (1991): Models of My Life. Basic Books

[16] Simon, H. (1997): Administrative Behavior. Free Press
Bevezetés a játékelméletbe. Közgazdasági és Jogi Könyvkiadó

[17] Szép Jenő – Forgó Ferenc (1974): Bevezetés a játékelméletbe. Közgazdasági és Jogi Könyvkiadó

[18] Zalai Ernő (2000): Matematikai közgazdaságtan. KJK-Kerszöv

Profit optimisation using business risk analysis and game theory

ROZÁLIA KONKOLY*, DR. ISTVÁN FEKETE**

*PKI Telecommunications Development Institute
konkoly.laszlone@ln.matav.hu
**Matáv Business Solutions LOB
fekete.istvan@ln.matav.hu

Reviewed

Keywords: Risk analysis, Game theory, Telecom economy, Present value calculations

There is an increasing need on the side of the companies to use also mathematical tools for modelling risks, supporting hereby their strategic and business decisions, creating connection between decisions and their expected results.

Introduction

On those areas, where there is competition, and so it is in the telecommunication sector as well, it is extremely important to satisfy customer demands on a high level. Besides, the main aim of all service providers is to maximise the available profit. All these are intended to reach in such a market environment, where competitors are also present, and their aim is also to increase their own profit. To reach the above goals it is necessary that in their business planning process companies should be able to explore and evaluate the risks residing in the competitive environment, namely they should take into consideration the expected strategies of the competitors as well. For this purpose game theory is an appropriate tool. Game theoretical analysis can give an answer for decision-maker questions like "what would happen if...", and helps to decide on "what would be the best step if...". The proper use of game theory occasionally can fling just as much on the prosperity of a company that its profit in a special case will be 10 percent profit and not the same quantity of loss.

In the paper business risk analysis, which is suitable for the identification and evaluation of the risks involved in the competitive environment and to which game theory modelling can be joined, will be shortly presented. The theoretical background and the most important features of game analysis will be shown in separated chapters. Then a case study elaborated for the telecommunication sector will be presented as an illustration how the result of risk analysis regarding the net cost of a given product, can be built into the game model taking into consideration the expected behaviour of the competitors. The presented model was elaborated in order to be able to determine the optimal price strategy.

1. Business risk analysis

The objective of business risk analysis is to explore the external and internal risk factors, which can have either positive or negative effect arising as a consequence of the strategic and business decisions of a company, and to determine their effect numerically by using mathematical methods. On the basis of this companies can prepare their risk management plan. In the competitive market, taking into account the appeasing of the customers' high level needs as well, the procedure substantiates the business planning process by exploring and managing the risks again and again as frequently as necessary, resulting a more and more realistic business plan for the company.

Realising the need for such a methodology, business risk analysis procedure has been developed at the Hungarian Telecommunications Company. The set-up for the risk management plan composed by the results of risk analysis, is part of the process. The feedback to the measurement of process efficiency is also part of the procedure. By applying the developed procedure, risks in relationship with business activity have become manageable, and in case of their occurrence the validated numerical characterisation has resulted in an effective risk management. Also the dynamic follow up of the changes has been made possible.

Figure 1 shows on what levels of planning risk analysis methods has to be applied when preparing business decisions at a company.

The first (upper) level is the level of creating corporate strategy as a result of processing various different pieces of information. The most important ones are the expected market and technological trends, the strategy of competitors, regulations and other externalities referring the planning phase. On this level business risk analysis gives assistance to identify strategic goals:

– The development of pricing strategy should be especially mentioned. In a competitive environment using game theory through the business risk analysis process makes possible to model e.g. the price war of service providers by taking into account the expected price strategy of the competitors as well. This application possibility will be presented in the subsequent chapters.

– Another important area is to support the decision-making related to various strategic capital investments. The question can be the time and schedule of realising the investment. The joint application of real options and game theory being used through business risk analysis can help to decide questions of this type. This methodology has been developed in the Eurescom P901 project (see hereinafter and [4]).

The second level is the level of business planning. In this phase among others corporate level revenues, operating (OPEX) and investment costs (CAPEX) are planned. In the course of marketing planning complex and all-around interrelations involved in the environment of the company can be modelled. Exploring and evaluating the positive and negative risk factors impacting revenues and costs regarding sale activity can do this. This activity improves the information base used for the preparation of the business plan. The model created in this manner is suitable for example to establish the risk level at which top management's profit requirements can be fulfilled, and to determine the tasks to be done for ensuring the realisation of the expected profit.

During the analysis of an investment program business risk analysis gives assistance for the optimisation of the investment portfolio when allocating investment resources. Additionally, it is suitable for determining the priority order of the investments to be planned.

The third level is the project planning level. In general all projects require performing risk analysis. On this level however, especially in case of long (several years) lead-time projects, it is advisable to perform risk analysis in the phase of implementation as well. This investigation can basically cover project lead time and implementation costs.

Of course on this level the components of the proposed methodology must be applied with different content comparing to the case of strategic and business planning level.

On each level textual analysis, sensitivity tests and the identification of the risk benefit level has already been applied previously. To complement these, the application of brainstorming workshops, functional consulting, Monte-Carlo simulation, real option and game theory have been developed. According to the application having been put in year 2002, this methodology development was admitted by the Hungarian Innovation Association with an award declared to be a successful innovation [5].

The methodology of business risk analysis consists of several modules. Among these we intend to shortly present the procedure developed for the practical application of real options, because, according to our information, at present this methodology is still quite new in Hungary [15].

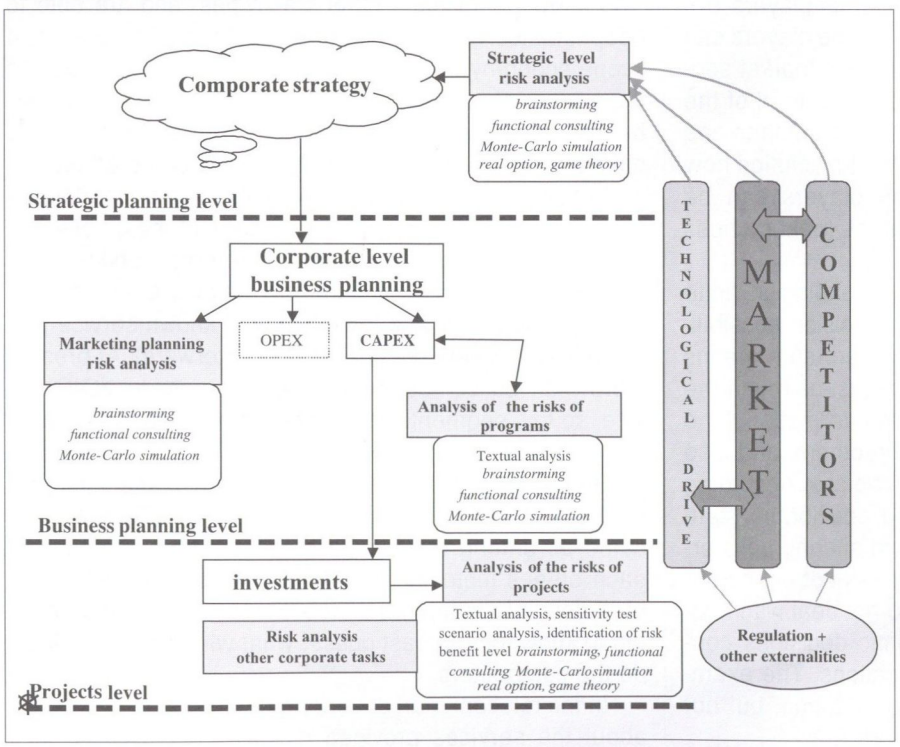
Real option

Most of the evaluation models are static, namely it optimises the decisions made in the time of analysis. There exist models that are able to catch the dynamic nature of the decisions as well, that is the future decision possibilities of the management. Real option is one of these models [18]. The real option model can be applied both on the strategy level and the project planning level for evaluating the strategic capital investment.

The detailed introduction of real option is skipped in this paper, we only present the principles elaborated for practice. It is worth taking into consideration that we should reuse the results gained in earlier phases of the applied risk analysis wherever it is possible.

In case of applying real option, first we have to choose the factor whose stochastic changes basically determine the management decisions. Such factor can be the sale price, technology changes or the stochastic change of the demand regarding to a given product. If we can not find a factor, for the stochastic changes of which there are no long-term public or easily accessible data at our disposal, we can use the net present value (NPV) of the future cash-flow regarding an investment or a strategic capital investment. The values of this factor can be produced by Monte-Carlo simulation. Of course this factor can be used for an optional evaluation only in cases when we can reassuringly prove, that the stochastic variations of the cash-flow generated on this manner follows a special kind of stochastic process.

Figure 1. Application of business risk analysis during the business activity of a company



The possible processes can be e.g. geometric Brownmotion process, Ornstein-Uhlenberg process or a process generating stochastic impulses. These all come under the huge group of the Ito processes. The stochastic changes of an Ito process can be described by the generic equation below:

$$dS = a(S,t)dt + b(S,t)dz,$$

where the dS/dt *change of the process during a small* dt *time follows a normal distribution. The trend of the change is shown by the* a(S,t) *factor, the* b(S,t) *factor shows the stochastic variance, so called volatility.*

The verification regarding to the normal distribution of the stochastic changes of the cash-flow must be performed before every application cases. In many cases however, conspicuous manipulation and not probable effects direct processes. If the process is not manipulated, the next task is to determine the volatility parameter expressing the measure of the stochastic changes.

There are books and papers referring different methods for this. In many cases we do not have enough data from the past, volatility can be determined by using Monte-Carlo simulations again. For performing this, we have to substitute the data, gained by applying Monte-Carlo simulations for the cash-flow, into the new simulation model, followed some algorithms found in the technical literature [15].

Of course the object of the optional investigation is also the effective risk management. The possible risk management actions can be as follows: to postpone the realisation of the investment or strategic capital investment, to cut the investment for more sections, to suspend an ongoing project or to finish it permanently.

Game theoretic models can be applied by using the pay-off function gained while using real option. In most cases we experience some movements in the equilibrium strategies compared to strategies resulted by the static models. So in case of long-run projects it is suggested to use this dynamic model, but we have to carefully verify all the necessary conditions [4, 8].

2. Oligopoly game theory for modelling competition in telecommunication

Game theory is a discipline between mathematics and economics, suitable for analysing the different players' behaviour and the interactions among them. The players can be e.g. companies selling products in a given market segment. Game theory was developed in the first third of the 20th century, in which the work of John von Neumann had a central place. Neumann was looking for the solution how decisions made by rationally thinking players can be brought in the two-person gambling games (poker, chess etc).

Neumann theory concerns to games ending in finite steps and containing a finite number of decision possibilities, zero sum and perfect information. Zero-sum means that as much as one of the players wins just as much the other will loose. The game is of complete information if the players know the possible choices of the others, and also the results got in case of the different choices. According to Neumann theory in these games an equilibrium state can be reached, from which it is no worth altering unilaterally for the players, because their profit can not be increased on this manner. In this paper we do not deal with hazard games, but we investigate economic decisions concentrated on the area of telecommunications. The example below is a finite, complete information game, but not duopoly and not zero-sum.

If companies can apply game theory eligibly in determining their corporate strategies, they can conclude their most reasonable decisions from the expected behaviour of the others. Game theory is able to give answers for questions such as, what will probably happen, and what will probably never happen in a given market segment.

Calculations can be applied in a competitive environment, where market players are constrained by appropriate laws and which they are complying with.

The starting point for game theory applied in the telecommunication area is that by opening the market companies are forced to increase efficiency, and they have to invest more money for research and development than before. This determines the structure of the market, influence pricing and forms of co-operation among the market players. Competition is intensified by the fact that because of the development of technologies various companies turn up on a given market, among which previously would not have happened to be a competition. For example nowadays cable television companies can also provide telecommunication services, and followed the wire-line access, mobile and microwave technology has made the possibility for the fast multimedia and Internet applications. To make easier the market struggle, and to alleviate the burden, larger companies buy and assimilate smaller ones. West-European experiences show that only larger companies can compete efficiently against incumbents having lost their monopoly status.

According to specialists' forecasts in the Hungarian telecommunication market, just like in other large industrial areas, liberalisation will not result a many-players "wild race", but rather more an oligopoly market will appear. In such a market there are only a few companies present, who know about each other's activities, and are able to adapt themselves to the other competitors. The scale and scope economy put the lid on entering of newcomers. As the number of market players will be relatively few, they will be able to influence e.g. the prices, but meanwhile they have to observe the prices and services of the others.

In the present domestic telecommunication market the oligopoly structure prevails e.g. in case of the mobile and leased line services. The number of internet service providers is more than that, though the majority of the internet traffic still appears at a few (4-5) well known service providers (e.g. Axelero, GTS, Enternet, Interware). At present more than 50 companies are able to provide VOIP service, so this segment really does not belongs to the oligopoly structure category.

By using game models elaborated to the oligopoly market it is possible to determine how equilibrium could come off among the market players if they comply with each other's mutually. To set-up a practical model (game of perfect information) it is necessary for the market players to know, but at least guess, what will and what will not do the competitors.

In the telecom market, among others, we have to know about the services provided by the other players, their

quality of service, tariffs, network platform, costs and the number of subscribers. Actually, companies do not agree their activity, because in general they are not allowed to do so, but somehow they must be able to comply with each other. Game theory results give information about how companies can conclude their most reasonable steps from the expected behaviour of the competitors when planning strategies. In reality companies do not harmonise their activities, but in the interest of long-term efficiency, they must be able to comply with the others. Game theory is a suitable tool to determine such kind of equilibrium states, and uses the mathematical optimisation to determine the best joint strategy of the competitors. In case of oligopoly this task means strategy optimisation based on a number of equations representing the pay-off functions, and the number of equations is equal to the number of competitors.

Similarly to a many-player competition players of an oligopoly market make an effort to maximise their own profit, but they are always influencing each others' activities, so it will also be a natural effort to give up autonomy for the sake of increasing the profit. (This situation can be modelled by using co-operative game theory.)

Because of the great deal of information and calculation needs of oligopoly game models, only a few application examples can be found in the literature, but it's usage can be enforced by the liberalised telecommunication environment. By setting up models very carefully companies can get a lot of useful information. If we can build a model based on correct data, we can get useful quantitative results. If data we collected are less accurate we can expect only results showing useable qualitative issues.

Game theory can be used not only for market players. In the control of the world-wide broadband data networks it can be necessary to relate the seizure and dimensioning of network resources with the expected „behaviour" of the data stream [19]. In general, algorithms have to find the optimal solution in the environment of caotically changing traffic environment when network conditions and routing rules change also very often. In the model players can be softwares, users or both. The applied methodology compounds the genious invention of John von Neumann, the computer and also the game theoretic items, in the elaboration of this his work also played a very significant role.

Another important application area is the frequency auction modeling [20]. Description models are very complex in these cases as well, we have multi-stage games with several players, so models are not oligopoly in general.

Research is going on in both topics for developing applicable procedures, but in this paper we do not go into details.

3. Relation between business risk analysis and game theory

Market competition can appear regarding price, investments for developing the infrastructure (e.g. a network), strategic capital investments, assortment or quality of services provided, or other factors (e.g. starting an advertisement campaign, providing different preferential bundles of services). In case of a competition regarding network capacities, companies decide on how big network they should build in order to provide a given service economically. Network capacity determines e.g. the maximal number of subscribers to be connected (or served). The profit is a function of the number of users, the price and the cost elements. In these models price is not the subject of the optimization in general, but price depends on the offer, namely in case of building big networks price must be decreased in order to capacity could be sold, and so the income should be more and more bigger. This demand-price elasticity is expressed in the demand curve well known in economics.

To set up an application model and to perform calculations it is necessary that market players know, but at least guess, what will (or what will not) the competitors do. Therefore it is a must that they should have information about the service, infrastructure and investment data of the others.

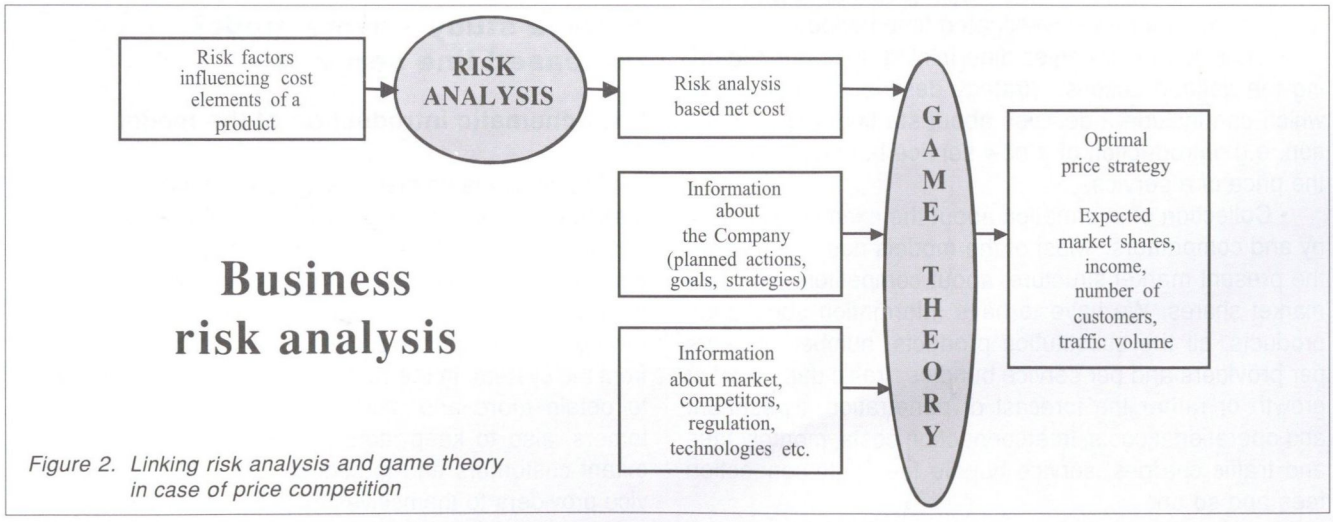


Figure 2. Linking risk analysis and game theory in case of price competition

In case of price competition companies would like to determine the price so that they have as much profit as possible. In this case care must be taken on having the appropriate network measure, because shortage of network resources can not assure service provisioning on the demanded level, and so it can degrade competition. Sometimes it happens that companies would like to optimise both the sizes of the network and the price together. In this case however modelling is more difficult than it is in the other two cases.

From the models mentioned above the schema regarding price competition using risk analysis and game theory together can be seen on *Figure 2.*

The link set up between business risk analysis and game theory consists in the solution that one of the most important data among those necessary for the model, namely the net costs regarding the investigated product are calculated by using risk analysis.

Knowing the planned actions, strategies and aims of the investigated company, using market information, data about competitors, regulation and technological information, game theory results complete business risk analysis and the traditional strategic planning tools with the quantitative analysis of the threatens residing in a competitive market. Results can be the expected market share, income, shape of the number of users, shares regarding traffic volume etc.

4. General characterisation of games and expected results

4.1. General characterisation of games [10,11,17]

• Choice of players involved in the model (in order to the model should be treatable, telecommunication providers negligible from the point of the investigated market segment can be eliminated)

• Specification of the time period for the investigation (1 year, 2 years, 3 years etc.) during which the investigated company would like to optimise its profit.

• Definition of admissible steps (actions) followed by the players during the investigated time period

• Definition of strategies determining decisions regarding the defined actions. Strategy describes a behaviour, which can include a decision about starting a special action, e.g. introduction of a new service bundle, change in the price of a service.

• Collection of information about the examined company and competitors. Most of the models need data about the present market structure, about competitors and their market shares. We have to have information about their products, all the substitution products, number of users per providers and per service bundles, traffic data, market growth or rather the forecast of penetration, investment and operational costs, interconnection costs, monthly fees and traffic charges, service bundle fees, interconnection fees and so on.

• Further data are needed for modelling the migration of subscribers to another service and to characterise the users' preferences regarding the choice of another service provider. Marketing research is needed for modelling the cross migrations caused by price differences. Information referring the connection between price and demand can also be used up, if those can be produced from market research data available from the past.

• Elaboration of the strategic interconnection model (in Chapter 6 a detailed introduction of some modelling possibilities developed and applied by us can be seen).

• Composition of the pay-off function, that shows the outcome of the game as a function of the joint strategy-sets of the players. In the models we used the operational cash-flow concerning the investigated time period for presenting the game results according to the following formula: operational cash-flow = operational income – operational cost. In case of models covering several years we calculated the net present value.

• Evaluation of the model, determining equilibrium strategies (dominant strategies, Nash equilibrium, recognising analogies with the most familiar game models, such as prisoners' dilemma)

4.2. Results of a price competition model

• In case of our price model we get the best price strategies referring the information of traffic charges (or price reductions), which are to be followed by the companies included in the model.

• As a result we get the profit for all price-strategy combinations including the optimal price strategies having got as a result of the game, which ensures the possible biggest profit for all the players.

• Network traffic values (decrease or increase in traffic, changes in load distribution etc.) can be followed up in a competitive environment, and these data can be used in the course of network dimensioning.

• From the optimal price strategy price margin can be calculated for the given product. Also the sum of traffic and income to be lost or gained can be calculated.

5. Case study – game model for leased line services

5.1. Schematic introduction of the model

The investigated market segment in this model was the managed leased line service (most preferred bandwidth). The investigation covered those 3 companies having the biggest market share on the market for the given service. In order to the model should be simpler, the 4 companies having not considerable market shares were eliminated from the system. In the competition service providers strive to obtain more and more percentage of the new customers, also to keep more and more percentage of the extant customers and to seduce customers of other service providers to themselves by giving price reductions.

Prior to the game modelling risk analysis has been performed. The main task of business risk analysis was to give a numerical measure for the uncertainties involved in the cost calculation for leased lines. This task was performed in two phases. In the first phase experts of the given speciality has explored the risk factors influencing the value of the elements involved in the cost-based calculation, or rather risk factors having been a danger to the realisation of the actual objectives. At the end the evaluation of the risk factors has been done [12, 13].

In the second part of the analysis the critical factors were construed very thoroughly and then calculated. On the basis of this information Monte-Carlo simulation model [16] has been built, and as a result we have got the mathematical distribution function for the net cost of the given product. From this the mean value for the cost can be calculated. Also the variance or range showing the risk measure can be derived. Finally the results of the analysis can be used for compiling the risk cutback plan.

After simulation we have had the distribution of the net-cost determining the unit cost for the leased line services, and the gained information has been used in the game theoretic optimisation of the monthly fees. Our main aim was to determine the value of the optimal price margin ensuring the maximal profit from salesmanship in a competitive environment.

Namely in the case of products the formation of traffic charges is very important. This is supported by the risk analysis results examining the revenues of the products. During risk analysis among the factors always appears the formation of the tariffs of the competitors as well. Because of this by all means it is reasonable to model the traffic competition among providers, for which a suitable method can be a game theoretic model.

Combining the two methods it will be possible to determine a price structure, which on one hand is based on the numerical effects of the uncertainties arising during the price calculation. On the other hand, by using game theoretical results the combined method is able to give a proposal for an optimal price structure that takes into account how market competition takes an effect on the price policy. This information can be used to determine the optimal price, and rather a price range, in which price reductions resulting gain can be given.

Solving our task the following assumptions has been made to the calculation of the optimal values for the monthly fees:

At the beginning of the investigation period (it was one year) the market leader determines its price. On the course of this some percent decrease compared to the valid list-price regarding the monthly fee can be given. Some times later (in the example presented it was half-year) competitors follow the market leader, namely they also give some price reduction, but it can happen that they do nothing.

Supposing that, if the other two competitors follow the market leader, then they do it at the same time, but prices are determined independently. The reduction of the monthly fee can decrease the income, but the increase in the number of customers as a consequence of the price reduction can fairly compensate this effect. We supposed that leased line subscribers that are business customers on the examined leased line service market consider not only the price. They take into consideration other parameters concerning the quality of service (e.g. yearly service outage time, service availability, time of installation, goodness of customer relations and treatment of customers, technical experiences) as well.

To determine the customers' preferences referring to different providers we used a quantity calculated from the qualifications regarding to the above parameters. In the model presented we used the weighted sum of the qualifications to characterise the preferences the customers use when making a choice among providers.

Figure 3 shows the game in a simplified form. Simplification means that among the five possibilities (0%, 5%, 10%, 15%, 20%) for price reduction we only involved two (0% and 10% reduction) in the figure.

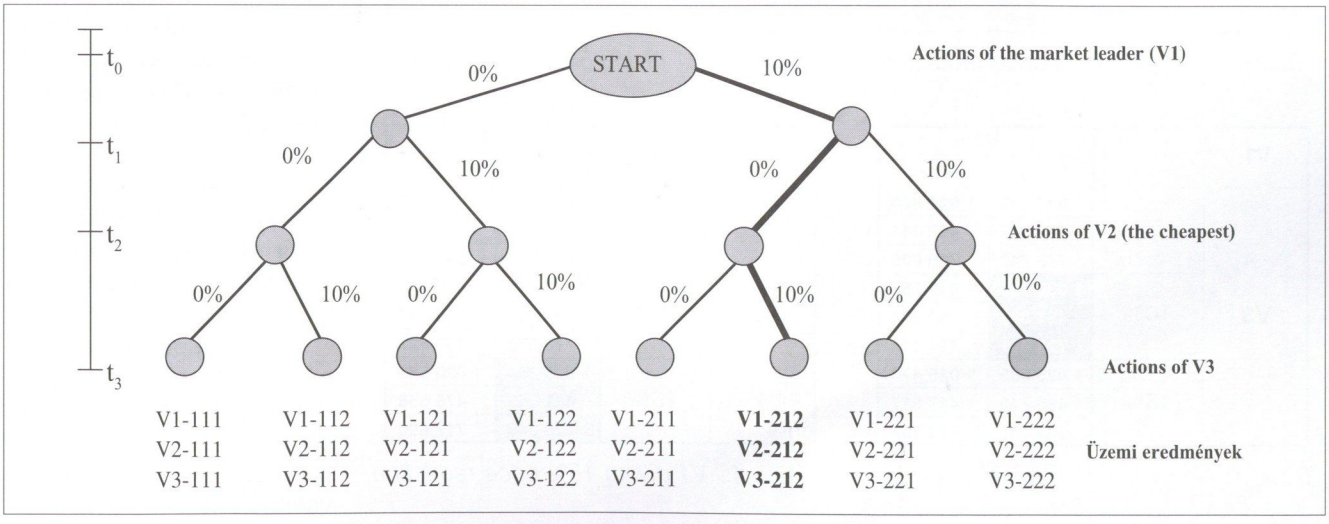
Notations:

V1: market leader company

V2: company providing at the cheapest price (about 20% cheapest than V1 and V3)

V3: company whose price is about 2% higher than that of the market leader's

Figure 3. Schematic game model



The preferences referring to the different providers are the functions of the actual prices. In case of equal prices the qualifications for V1 and V2 were much the same, while the qualification of V2 was lower than that of V1 and V2.

V1-212 denotes the operating cash-flow (revenue minus the operating and investment costs referring the investigated time period) for V1 in case when V1 makes 10%, V2 makes 0%, V3 makes also a 10% price reduction applied as a strategy.

The strategies realising on the other paths can be interpreted similarly. The time axis on the left side of the figure shows the time points of the individual steps. This time axis shows the starting point for the market leader. This point is signed as t_0,in our case it was the beginning of the year. Time points t_1 and t_2 show the time of the steps of V2 and V3. In our example $t_1 = t_2$ and these steps happen at the end of the half-year, t_3 shows the end of the investigated time period (in our example it was the end of the year).

5.2. Results gained from the model

According to the results of the performed calculations the strategy determined by the path drawn by thick lines on Figure 3 proved to be the best joint strategy for the 3 companies. *Tables 1-4* show the values for the pay-off functions in the surroundings of the equilibrium point. (For the sake of perspicuity we present only the significant strategies here. From the strategies of V2 we have not represented the results regarding to the bigger (10-20%) price reductions in the tables, as we will see, these strategies do not mean profitable ones for the company.)

The number triples belonging to the different joint strategies show the pay-off function values for V1, V2 and V3 respectively for the given choice of strategies.

In tables 1-4. we designated the cells representing the Nash equilibrium for V2 and V3 in case of a given strategy of V1 by changing the background of the cells. Nash equilibrium means a set of strategies, from which it is no worthwhile to alter, because in case only V2 or V3 changes unilaterally, it surely gets a worse result.

The 0% price reduction means also dominant strategy for V2. For V2 this is the best strategy in all the cases, whatever is chosen by the V3 company, and whatever was chosen by player V1. So V2 will not alter from this. It is the company who sells on the lowest price, for it further price reduction would not result extra profit.

For V3 there is no strategy, that is the best whatever strategies the other two players follow. For V3 both the 10% and the 15% price reduction will be the best, depending on the strategy choice of the V1.

In case when V1, the market leader made no price reduction (0%) or made a 15% price reduction, for V3 it is the best choice to give a 15% price reduction. When the market leader gives 5% or 10% reduction, then for V3 it is the best to give a 10% reduction. It is a question, what will the market leader choose, if it presumes, that later on the others will optimise their joint strategies. For the market leader the biggest profit would be resulted if it would give a 15% reduction in order to seduce the customers to itself. In this case however the 15% reduction would be the best strategy for V3 as well. At the same time this would deteriorate the profit of V1 compared to the profit it gained when giving only 10% reduction. In this case the 10% reduction will be the best step for V3 as well. That is, the winner strategy set for the V1, V2 and V3 companies will be the 10%, 0% and 10% price reduction respectively.

If companies will make their decisions considering the above thinking, namely that they take into consideration how the others will decide on their best strategies, and on this way optimise their profit, then they tacitly complied with each other.

V1	0%	V2	
		0%	5%
V3	5%	951 254	947 982
		532 064	529 233
		771 084	756 452
	10%	949 001	945 764
		529 004	524 808
		775 479	767 873
	15%	936 773	933 589
		527 364	522 263
		778 258	775 699

Table 1. V1 giving 0% price reduction

V1	5%	V2	
		0%	5%
V3	5%	1 001 171	984 599
		509 027	505 332
		746 211	732 060
	10%	1 000 259	983 717
		506 073	503 678
		750 120	742 509
	15%	998 034	981 521
		504 681	502 136
		748 667	746 194

Table 2. V1 giving 5% price reduction

V1	10%	V2	
		0%	5%
V3	5%	1 032 253	1 021 990
		493 942	492 444
		713 742	700 005
	10%	1 030 490	1 016 253
		491 083	490 298
		728 491	721 355
	15%	1 025 685	1 015 474
		489 740	487 917
		726 787	724 391

Table 3. V1 giving 10% price reduction

V1	15%	V2	
		0%	5%
V3	5%	1 034 293	1 029 547
		485 676	484 987
		697 762	684 173
	10%	1 036 543	1 025 394
		482 866	480 898
		710 579	703 527
	15%	1 028 254	1 020 553
		481 559	478 558
		720 049	717 690

Table 4. V1 giving 15% price reduction

5.3 Results gained by using business risk analysis and game theory together

By using game theory we can determine the optimal price strategy using the main characteristics (minimum and maximum values, expected values) of the net cost distri-

bution gained in the risk analysis process. From the optimal strategies the optimal price margin can be derived. This information can be used both in the determination of the list price, and it is also useful for the salesmanship to decide on the question, what is the range for price reduction that can be given in case of a tender. For company V1 the optimal 10% price reduction is equivalent to an about 6% price margin. It is the task for the experts to elaborate how price reduction is expedient to be realised. There are two possibilities, one of them reduces the list price the other gives preferences. In most cases performing the risk management action plan, which was elaborated during the business risk analysis process, part of the price reduction can be compensated.

6. Game theoretical modelling of strategic interactions

Different methods have been developed how to take into account the effects the competitors' strategies have on each other's strategies in the developed models. In the literature we have found models using mainly the demand curve (see 6.1) [3]. To solve the example shown in Chapter 5 methods presented in chapters 6.2 and 6.3 have been developed and used in our models. Their usage proved to be usable in the course of the game theoretical modelling.

6.1 Demand curve

In the well-known models of economy (e.g. Cournot and Bertrand) demand curve is involved expressing price-demand elasticity. This curve is approximated with a linear function in general. However, this is not only for the sake of simplicity but it can be proved that if the utility curve of the telephone users is a quadratic function then the demand curve will be a linear function. The proof can be found in [3].

The linear demand curve in duopoly case can be written in the following form: *p=a+b(Q1+Q2).*

Here p denotes the price (here it is supposed to be equal for both companies), Q1 and Q2 denotes the quantities produced by the two companies (here it can be the size or measure of the network), "a" and "b" is the two parameters of the linear curve.

Figure 4 shows a demand curve, where the optimal quantity belonging to the optimal price has been represented.

6.2 Migration function

While the demand curve presented in chapter 6.1 shows how companies effect each other's activity through the network capacities, the migration curve has been used to model customer migrations caused by the different prices. If in the market regarding the investigated service there is a possibility to select from more than one providers (see Chapter 5), then first we have to explore the factors, whose

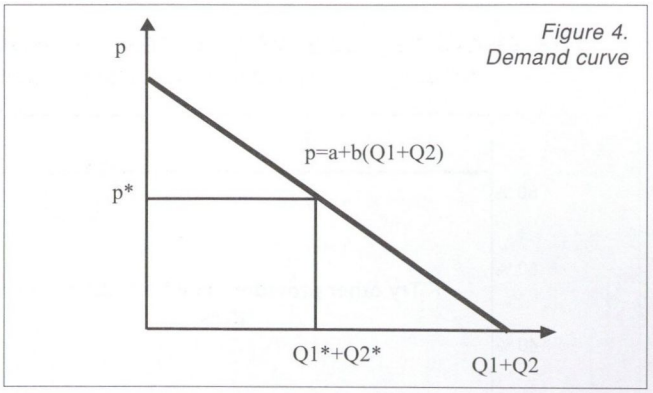


Figure 4.
Demand curve

values and quality are considered by the customers to decide on which service provider to choose. (It can happen that customers can make their choice without a long-term commitment, as in the case of the call-by-call service selection.)

The Bell Research marketing research and consulting company has been carrying-on market research activities in Hungary since 2000. In the press release [1,2] the company reported about a research study that was performed within the scope of the medium (headcount is 50 to 299) and big (headcount is more than 300) companies in Budapest. The study was related to the question whether the above companies would try also other providers or not after the liberalisation. These investigations were applied for the fix telephone service.

Extending the answers given by the 173 companies that were surveyed to the total set of all companies of the above kind, the following results turned up:

- In case of the tariffs are 5-9% less then 17% (268 companies) would change provider
- In case of the tariffs are 10-19% less then 30% (719 companies) would change provider.
- In case of the tariffs are 20-39% less then 44% (1371 companies) would change provider.

In the investigated segment there were no one who would have answered that in case of a 5% tariff difference that the company would come over to another provider. BellResearch calculated from the above data that there might be a 20% difference that medium and big sized companies in Budapest should change their present provider.

The cumulated effect is presented on *Figure 5.*

According to the survey made by BellResearch Figure 5 shows that how big migration (given in percentages on the vertical axis) can be expected in cases of given price differences (given in percentages on the horizontal axis). Under the percentage values, just above the horizontal axis values expressed in the number of companies appear as well. The received data show that above 40% price-difference, migration intensity decreases speedily. This big price difference probably adumbrates in the customer, that he surely will not get the same quality. On the bases of the figure such a service can reflect the fact, that questioning about a more than 60% price difference, which otherwise seems to be a very seductive offer, everybody answers without thinking that yes, of course he would change.

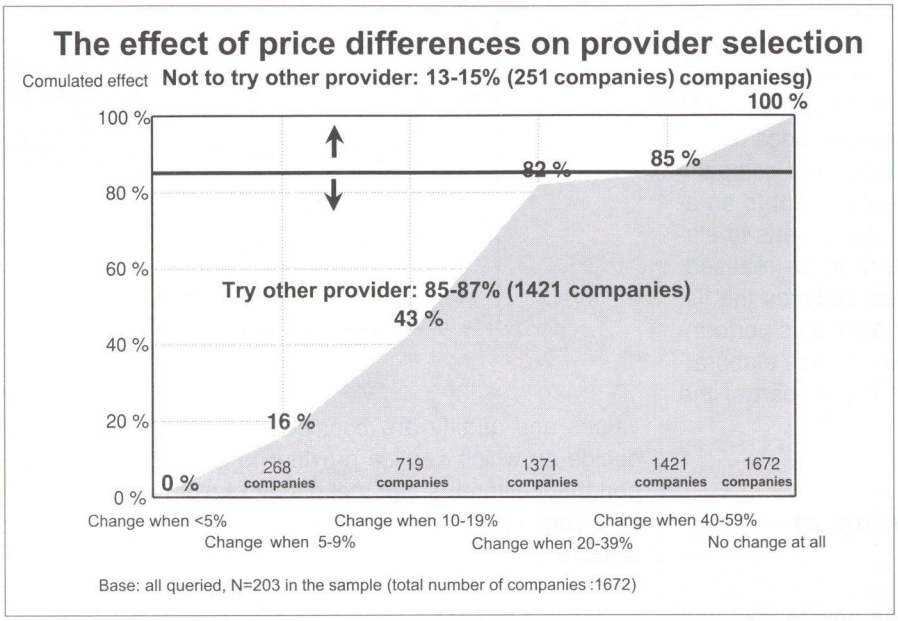


Figure 5. Effects of tariff differences on provider selection [1]

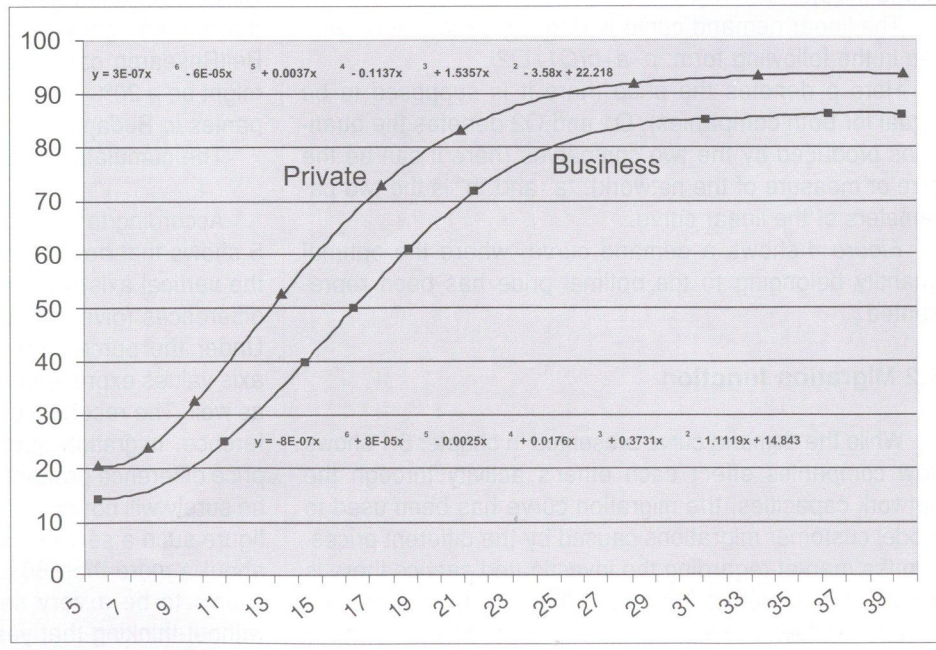
Such a big price difference can not be imagined in reality, neither is it allowed by regulation, nor it could be economical for the providers. The results of surveys based on queries or filling questionnaires can be more and more correct the bigger is the number of enquired entities. In the above survey there were 173 companies asked, and the results were expanded on the bases of their answers for a set of companies which was about 10 times bigger than that involved in the survey.

Probing other providers, so that cutting themselves adrift from the present provider, customers were motivated by several factors. Among these the following reasons can be found. For example there are high tariffs, customer services are not eligible, there are long establishment times, there is no flexibility, service quality is not convenient all the time, sometimes there are accounting problems, slow failure averting and so on. Sometimes companies would try another provider simply from curiosity. Among the above reasons tariffs were mentioned most frequently (76%). Having knowledge about the actual tariff, customers can calculate the difference between the price of his actual provider and the price of another one. The evaluation of the other factors is not so unequivocal for a customer e.g. it is more difficult to compare the time of failure averting or the quality of the services given for the customers.

Using the above data we have prepared an analytical curve (based on a polynomial approximation) which is suitable for our modelling purposes. The points of the curve has been calibrated so that the average values in the ranges shown in the BellResearch study should remain the same, and for the bigger tariff differences a bigger migration percentage should belong. Based on these principles we prepared a migration polynomial valid for the business customers, which is drawn in *Figure 6.* The curve shows the probabilities when a customer is thinking of migration in case of given price differences. For the private customers there were no marketing research data at our disposal. So we constructed a hypothetical curve which shows the price-sensitivity of the private customers, which is stronger than that of the business customers. During the construction of the curve, similarly to the marketing research results, we supposed that there is no migration when the price difference is below 5%. But the actual choice, the customers take into consideration, depends on several other factors as well. It can also happen that after weighing the other factors as well, the customer decide on still remaining at his actual provider in spite of a bigger price difference.

In the example presented in Chapter 5 we prepared the migration curve by using past data available at Matav. This information covered data about giving up the service, including the reasons of why not to continue (e.g. migration to another service but remaining at the present provider, migration to a competitor to the same service, migration to a competitor to another service, terminating the service for good and all).

Figure 6. Migration curves for private and business customers (in analytical form as well)



6.3. Multi-criterion method for evaluating complex systems (KIPA evaluation)

The KIPA method (named after the Kindler-Papp) has been developed in the nineties in the Technical University of Budapest. First it was applied in the operation research for choosing among complex systems. This is one of the most prevalent and reliable methods that can be used to make comparison among "complex systems" (such as companies, company management systems, information systems etc). The differentiation among the systems is carried on according to a complex system containing different aspects.

In our work regarding the leased line services market competition we used the data of a migration curve together with the multi-criterion method for evaluating complex systems [9] for calculating the customer preferences formed about different providers. KIPA evaluation is used generally for qualifying the different companies, but we have not met an application regarding game theoretic models in the literature.

If considering the percentage values of the migration curve as to customers meditate about the change at the given price difference, and take into account other factors as well, then we have to investigate what are the factors considered to decide about the change. In Chapter 6.2, by using the results gained by the marketing research [2], we have shown what were the factors that were considered to leave or not the present provider. In case of the call-by-call provider selection [7] probably there is no need for such deliberation, because customer is not forced to commit himself. In this case we can suppose, that a customer tries to choose another provider. In case of pre-selection [7] a customer will probably consider the other factors qualifying the provider better. But these factors will not be considered uniformly. To this type of modelling KIPA method can be a useful tool, which has been checked in the elaborated game theoretic models [6].

The preparation of the evaluation table is as follows:

Companies receive marks for the factors involved in the evaluation. This can be done by organising a brainstorming workshop. In the same way we add weights to the different factors.

Table 5 shows an „example" evaluation table for two companies, named as COMPANY1 and COMPANY2. Of course, in general, according to the actual application, the set of considered evaluation factors, the weights and also the marks change.

In the lowest row of the table the weighted sum of the values can be found. These values have been used in the calculation of the pay-off functions. This is the function of the factor designated for optimisation (e.g. in our case it is a function of the monthly fee, but any other evaluation factor that is involved in our table can take over this role).

EXAMINED PROVIDERS		COMPANY1	COMPANY2
Evaluation factor	Weight	Received mark	Received mark
Yearly availability	1.4	5.6	4.2
Establishment fee, one-time fee	1.8	9	6.8
Maintenance fee, monthly fee	2.0	**10**	**8**
Establishment time	1.2	3.6	3.3
Consultative Salesmanship, treatment of customers	1.0	3	4
Personal Customer relations	1.0	2	3
Technical Expertise	0.4	4	1.6
Mean failure repairing time	1.2	4.8	3.6
Weighted sum		60.7	50

Table 5. KIPA table for the qualification of COMPANY1 and COMPANY2

Summary

The presented procedure, namely business risk analysis combined with game theory has the following main features and advantages as follows in this short summary:

– By using risk analysis combined with game theory we reached that we are able to take into account the inter-relations among players having different interests. This predominates in a competition between the provider and the customer, or among the providers competing with each other. Though the procedure has been developed for the telecommunications area, with little changes it will be applicable in any other industrial area (transport, trade) in case of market competition.

– We have developed a new practical procedure, which is suitable for modelling the inter-relations among the competitors as well. The developed two methods, for which market research data and past data concerning the behaviour of the customers serve as input data, are especially new approaches:

– The numerical representation of the customer's preferences concerning the provider, which has been used to model how a customer decides about which service provider to choose,

– Generation of the migration function, which models how a customer of a given provider decides on to go over another provider.

The possible directions of developments offer a very wide scope for further research. Both the development for further static and dynamic models and the increase in the accuracy regarding market information, price and cost issues, and also the development for further methods to be able to model inter-relations, and also the joint investigation of many products, hold in reserve interesting challenges for researchers dealing with the topic.

References

[1] Hungarian telecommunication report, BellResearch, 2000.

[2] What can telecommunication liberalisation bring in the capital? BellResearch press release, 2001.

[3] Szidarovszky F., Okuguchi, K.: The theory of oligopoly with multiproduct firms, Springer, Berlin, 1990.

[4] "Extended investment analysis of the telecom operator strategies", Eurescom P901 strategic study. 1999-2000, Confidential. (Participants from Hungary: A. Gyürke, I. Fekete, R. Konkoly)

[5] Development and application of the business risk analysis process, Innovation Blue-ribbon Application 2002. (Innovation acknowledged by a certificate) (Hungarian)

[6] R. Konkoly, I. Fekete: Modeling market competition in case of managed leased line services by using game theoretical methods, Internal study at the Hungarian Telecommunications Company, 2002. (Hungarian)

[7] XL.Act about Telecommunication, year 2001, Hungary

[8] 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, 14-16 December 1999.

[9] Z. Juhász: Review of the Hungarian leased line market, Széchenyi István College, Diploma work, January 2001. (Hungarian)

[10] R. Konkoly, A. Gyürke: Application of game theory in the telecommunication investment analysis, PKI Scientific Days, 23-24 November 1999. (Hungarian)

[11] R. Konkoly, A. Gyürke: Game theoretical application possibilities in telecommunication, PKI Publications, Vol. 44. (Hungarian)

[12] I. Fekete: Analysis & Management of Investment Risks, QSDG Magazine June/July 2000 Vol.3 No.2 pp.43. London, United Kingdom

[13] I. Fekete: Collection and evaluation of risk factors, Hungarian Telecommunication, Vol. 2000/1, pp.43-46. (Hungarian)

[14] I. Fekete: Risk of human resources – fluctuation and its handling using the tools of risk management, Human Innovation 2001. Conference, Organised by Oktáv Rt., Budapest, 4 Oct. 2001.

[15] I. Fekete: Role of risk analysis in determining the cash-flow of investments, Thesis for the doctor's degree, Budapest Technical and Economical University, Budapest 2000. (Hungarian)

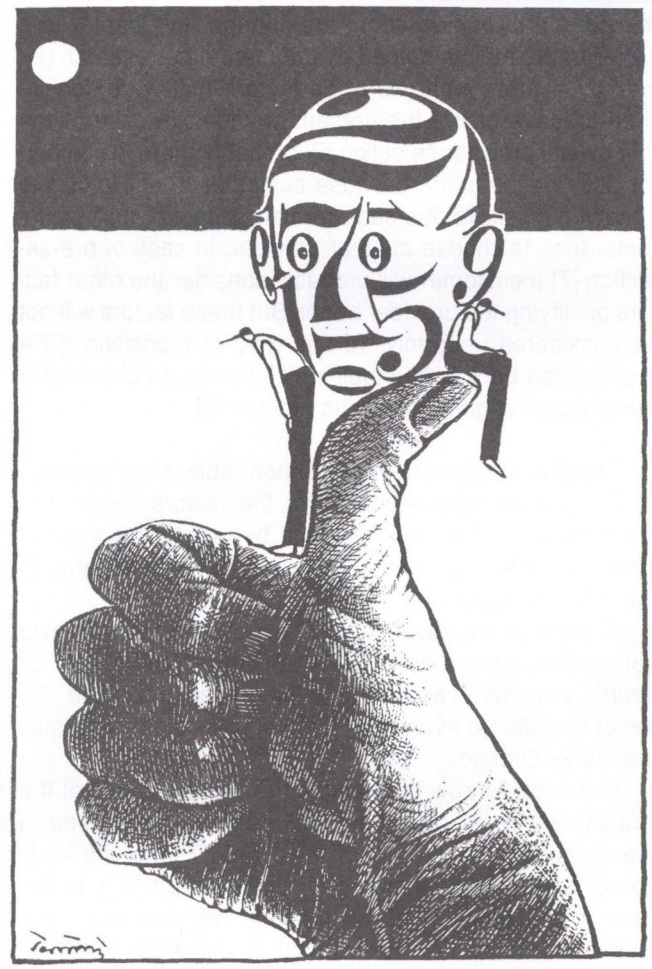
[16] I. Fekete, T. Katona: Modellierung der Erfüllung des Plans vom Jahr 2001 zur Verwertung von Immobilien mit Monte Carlo Simulation, Erfahrungaustausch, Bonn, 05.04.2001. Deutschland

[17] Game theory is not for fun, Daily Economy, Informatics enclosure, Vol. 2000/ 8 (Hungarian)

[18] Timothy A. Luehrman: "Investment Opportunities as Real Options: Getting Started on the Numbers", Harvard Business review, July-August 1998.

[19] A. Vidács, P. Füzesi: Game Theoretic Analysis of Network Dimensioning Strategies in Differentiated Services Networks, Proceeding of IEEE International Conference on Communications (ICC2002), New York, USA

[20] Andrew Byde: Applying Evolutionary Game Theory to Auction Mechanism Design, Intelligent Enterprise Technology Laboratory, HP Laboratories Bristol, www.hpl.hp.com/techreports/2002/HPL-2002-321.pdf



Optical networks and network strategies

GYÖRGY TAKÁCS

Péter Pázmány Catholic University, Faculty of Information Technology

takacs.gyorgy@itk.ppke.hu

Keywords: Gilder theory, Development of transmission bandwidth, Telecom policie

The revolutionary development in telecommunication and information technology is based on several roots. The critical mass of users (sometimes above one billion people) is often mentioned concerning mobile phones and computers, services, applications. Other important issue might be the fierce competition on the global ICT market. A very important factor in this competition to come first to the ICT marketplace with relevant new products and services. There is no ultimate winner in this competition for long run. Day by day the competition starts again and technical novelties push the market into new direction.

1. Introduction

Which directions of technology development promise the best future? How to prepare ourselves to survive the changes or even be leaders? What are the tasks of system suppliers, service providers, users, regulatory bodies, governments to provide better position to the human society by this game? Which fields promise good business, source of tax money, and benefit for end users?

These are really difficult questions. It is absolute impossible to be good fortune teller in this field by eternal statements. The aim of this paper is to stimulate thinking of experts in this field. Might be some elements of this paper are quite provocative. We have to recognise, that bandwidth grows fastest and transmission capacity has less direct limit in growth. Bandwidth is mentioned as a third category, simple technical issue. It has chance, that from enabler category the bandwidth issues will lift to the level of killer category. Tracking back of the stock exchange figures we can detect an opposite trend! Papers of high speed network companies are low now. However there are several optimistic analysers concerning bandwidth.

Facts are reviewed here in a strange sequence. Detailed analysis is done but the customised conclusions remain for readers.

2. Development of technology concerning bandwidth and related fields

The future of networks depends on the development speed and costs of the following items:

- processing one bit,
- transmit one bit,
- store one bit.

The future network philosophy is determined by the winner of the items listed above.

Well known laws summarise the development trends of items above. They all show exponential growth in time. The exponential curve is very steep at all but the value of exponent is highest at bandwidth. Consequently the fastest growth is at bandwidth. The processing power is doubled within 1.5 years, the bandwidth is tripled within one year, and storage capacity is doubled within one year.

The winning position of the bandwidth can be derivated from the value of exponent. An other frequently mentioned fact is that nanotechnology or novel processing ideas are necessary because the physical limits seems to be near in the development possibilities of traditional chip technologies. Hopefully the chip technology will find also new lines of further development to override the limits in

Moore law on processing power	The processing power is doubled in 18 months
Gilder law on bandwidth	**The total bandwidth of global communication system is tripled in 12 months**
Metcalf law on the value of networks	The potential value of networks is proportional with the sqare of the number of users
Shugart law on the price of storage capacity	The price of magnetic storage related to one bit is halved in 18 months
Ruettgers law on storage capacity	The storage capacity is doubled in 12 month
Wacker law on metadata	Any information related to the transaction has higher value than the transaction itself

technology e.g. to solve the thermal problems of big and high speed chips. However such problems do not exist in development of transmission bandwidth.

One carrier frequency in one optical fibre can transmit 10 Gbit/s. In one fibre 1000 carrier can be implemented and in one cable 1000 fibre is feasible. Multiply these figures the result shows that the theoretical transmission capacity of such a cable is about 10^{16} bit/s/. This cable has about one inch diameter, consist of 1000 fibre and in each fibre we can use 1000 carrier. Can we say that the transmission capacity of such a cable is practically infinite? Yes and a simple example may prove it. The normal life time of a human person is about 80 years that means $2,5 \times 10^9$ seconds. The most advanced coding procedures needs about 1Mb/s, to store or transmit good quality video signals. So our example cable can transmit in 1 second so many video signals, which can be watch by 4 persons during 80 years (24 ours per day). The transmission capacity of such cable can be characterised by this 4 life-long video signal transmission in one second.

The same time we can state that we are in the age of practically free bandwidth also. Let us bring an example again to prove it. Hungarian Universities and academic institutions are connected to the HUNGARNET network typically by 1,5 Gigabit/s fibre links. A normal monthly fee paid by institutions for a dark fibre leased line service to transport bits takes 200.000 HUF. The costs of equipment like media converters or routers are not included.

This leased line cost can be considered practically free. A tram ticket takes 125 HUF in Hungary. Using the full download speed 125 HUF takes 27 minutes downloading or downloading of 2430 Gigabits or downloading 6,75 hours good quality video or full content of hard discs of 30 personal computers. So practically the transmission costs can be considered free! That is a quite different question that the theoretical bandwidth is really utilised very low level. The transmission capacity is utilised by few percent during nights or weekends.

As other result of the fantastic development in optical transmission technology we can state, that the geographic distances in the globe are eliminated. The world record in fibre cable attenuation does not exceed 0.001 dB/km. This means that by one hop Hungary and any town of the USA can be directly connected with optical cables! It does not matter whether transatlantic or transpacific routes are used but amplification is not needed.

The NASA homepage has detailed *description of the system producing such optical cable.* On the board of Columbia space shuttle were tested the system several times to produce extreme clean glass fibre in microgravity environment.

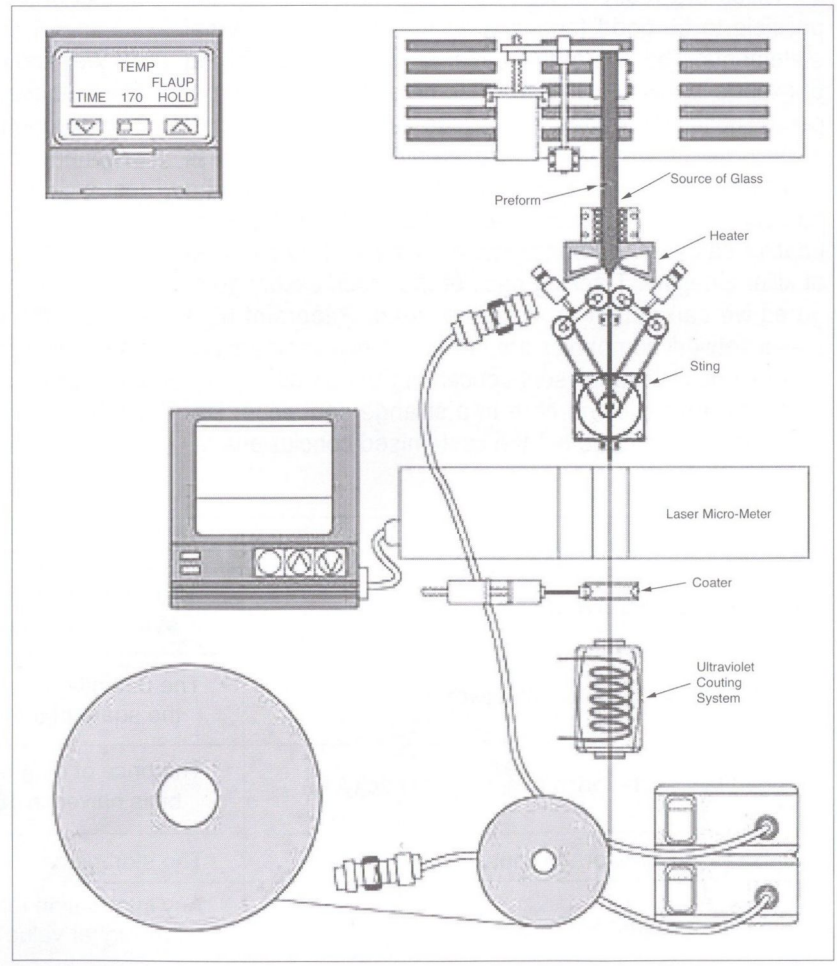
The fibre diameter is precisely controlled by the pulling speed and the fibre is immediately coated. We have to mention that the raw materials are available all around the globe with practically unlimited amount (not the same concerning copper because the copper resources of the globe are more and more limited).

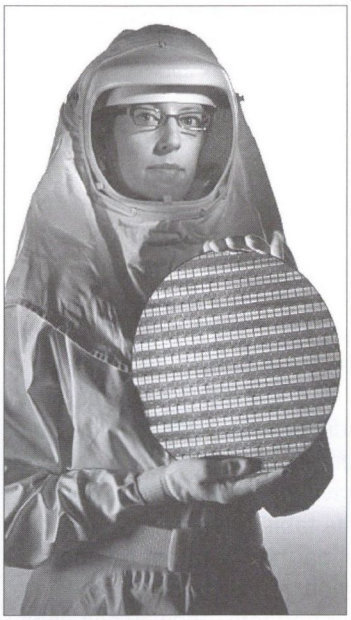
Why can we calculate with real explosion in the field of information technologies as a result of fast development in transmission capacities? Because the real basic practice will changed radically soon. As the development speed of processing power and the memory capacities is high enough this two factors together speed up the demand for transmission capacity too. The new solutions will spread mostly in the best developed areas of the world. So the winners and losers of new results will be even more divided.

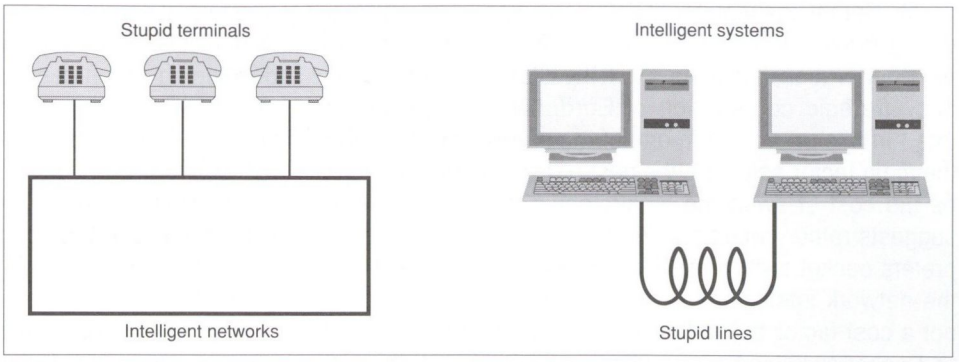
3. Impacts on network and service structure

The unlimited and free bandwidth has serious consequences on the development of network structure, on the network operation and on the services.

The unlimited and almost free goods have no high respects and not so easy to make profitable business on them – except stimulating wasteful applications. Even in the field of electronic engineering can be mentioned other wasteful applications: in 1970 the price of one transistor







was about 1 USD but today in a high scaled CHIP 0,000 000 025 USD costs one transistor. In one chip billions of transistors can be produced and an example *foto* is shown below.

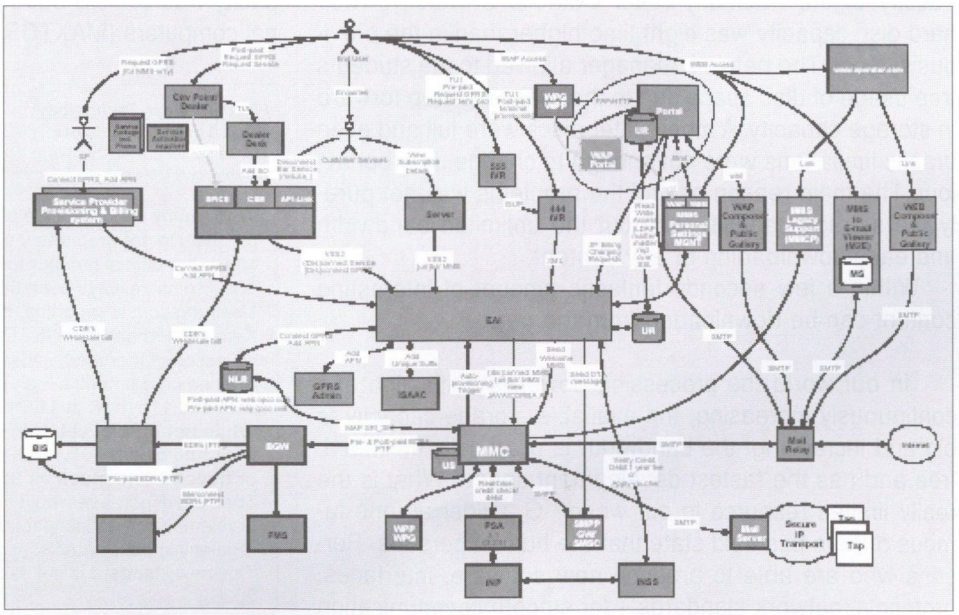
Companies make good business on practically free transistors. Only trick is to stimulate wasteful using of them. Consider please the up to date personal computers! They consist of billions of transistors. In spite of almost free transistors the complete computer has profitable prise. Hysteric campaigns are initiated to increase quickly the number of computers in Hungary! We have to follow the official statistic figures of the global leaders! Remember please the typical application of a personal computer (the application of billion of transistors in each)! Billions of bits are used for games or for screen savers or animated symbols in the help menu. Our fathers used a pack of cards for comparable games instead of billions of transistors with a sophisticated software package on the top of it. The real problem is not only the wasteful application but pulling the development towards the even more wasteful solutions without the critical thinking of other promising principles. It is time rethinking the principles of information technologies! It is time to analyse the consequences of wasteful usage of bandwidth.

4. Conclusion

New network structures and new roles of network nodes come with the wasteful bandwidth applications. In the traditional telecom networks had high level intelligence in the networks and stupid terminals were at the ends. In the computer networks the terminals were very intelligent and the network functions were considered primitive level, mostly as simple transparent bit-pipes (one bit in-one bit out) and similar simple network functions in nodes (controlled by terminals). The main difference is flow control. The essential features are represented in the *figure (above)* below.

One of the global leaders in systems demonstrated a new system concept. This system supports mobile users, standardised network components, different service providers, different application providers, shared network functions, authentication of users, billing systems... The service creation and provision functions, the access and transport functions, the application creation and provision functions are implemented in harmony of competition oriented multi provider environment. Elementary functions to find the mobile users, making the identification, authentication and authorisation process, providing correct data record for fair billing, meantime protect the rights of personal data...

The *figure (down)* below shows an example of implementation of such network functions. The names of individual boxes have not importance but the complexity of network structure is really interesting. The present telecom operators calculate fees mostly based on call minutes, so somehow they charge the used bandwidth. The network intelligence and mostly the content fees are mostly transformed into the charges of bandwidth. This network structure don't suggest development towards the free and unlimited bandwidth and extreme bandwidth stupid networks.



Mr. Carelli, the director of EURESCOM had a very interesting presentation in Budapest in 2001 concerning the next generation networks. (EURESCOM is the research and strategic cooperation of European network operators.) In this presentation one of main messages was that the main factor selecting the best network mode seems to be the cost of switching technologies. Cheap bandwidth suggests rather circuit switching, cheap processing power prefers packet switching. The concentration or distribution the network intelligence is a quite different issue. This is not a cost factor but rather question of technical feasibility. Packet switching of the above mentioned 10^{16} bit/s bit stream is not expensive but not realistic. Even one per cent of this bit flow is more above the processing capabilities of the fastest processing chip.

Relevant new (or re-invented) solutions in the network world are quite common. Remember the suggested network topology by A.G. Bell: this was a fully meshed structure. A very similar structure is suggested by G. Gilder for unlimited bandwidth optical networks. The invention of telephone switchboard by T. Puskás some years after the invention of telephone introduced the rational management with lines or with other words the rationalisation of bandwidth.

Using other words the unlimited bandwidth connections and meshed or ring network topology can be utilised as the traditional radio broadcasting: send the information flow everywhere and the selection of the relevant information is the task of the user at the receiving terminal. It is a logical question where can we put so many fibre cables? The existing duct system has space enough removing the copper cables – there is no risk to destroy pedestrian ways in towns.

Relevant new user behaviour can emerge by the free and unlimited bandwidth. Even the author of this paper has recognised that it is faster downloading a file from the net than find it in an unknown map with unknown name in the hard disc. In spring time of 2003 a new computer series was introduced at the Information Technology Faculty of the Pázmány Péter Catholic University. Their hard disc capacity was eight time higher than in the previous system. The network manager allowed to the students free usage of disc space due to the really big step forward in storage capacity. A month later discs were full and even drastic limitations were not enough to change user behaviour. The main reason behind the problems was not purely the storage capacity limit but the unlimited bandwidth and easy downloading of any content.

Within a few seconds fantastic amount of interesting content can be downloaded from the net.

In our world the processing power is quite high and continuously increasing, the available storage capacity is big and increasing; the bandwidth is practically unlimited, free and has the fastest development speed. What is the really limited resource in our world? G. Gilder as one famous guru in this field state that the human persons. Persons who are able to produce new software, interfaces, protocols, network standards – for smooth communication of information systems, and interoperability with previous systems. New solutions are necessary to utilise of quickly developing resources. Gilder forecasted the quick death of networks having stupid terminals like telephone or television. Gilder predicted high intelligence in the networks but concentrated at the edges. For example in mobile phones the processor, memory, display, camera .... considerable exceed the capabilities a normal PC in the early nineties. These terminal functions are organic part of network intelligence.

The new network philosophy suggested by G. Gilder promise a new age in content provision as well. Any content will be available in any time and any place. The main difference between the traditional TV broadcasting and the new download based content provision can illustrated by the following example. Travel by train is comfortable but you have to adapt to the time table have to go to the railway station, and your travel partners can not be chosen. By car you can departure when you wish, can stop where you want, you can select your partners, and you can change routes even meantime. If any content accessible for you, then the absolute freedom might be available for you to compose the actual and individual information, learning and amusement program without standing from your armchair. For such networks and services new companies are required. The existing firms hope better profit by traditional services. Probably pioneer companies initiate revolution in network and service structures.

What is the connection of bandwidth development to relate other factors? What is the perspective in processor power? A serious limit is the heating. The forced air heating is problematic to keep normal temperature in the processor chip. Water heated PC processors means one possible heating solutions. Several companies demonstrated water-heated processors in CeBIT 2003.

Developments in storage capacities have nice figures also. IBM provided in 2003 personal computers with 400 GB hard discs as well. In 1998 the average hard disc capacity was 5,1 GB but in 2002 36,1 GB in shipped personal computers (MAXTOR).

Fixed Wireless Technology

802.11 To Get Speed Boost?

The IEEE has a study group pushing for new standards to officially bring throughput levels for wireless LANs up to 108Mbps—or faster.

While many vendors have played with the technology of 802.11 products to get proprietary speed boosts of 72 Mbps and even 108 Mbps, the official number for the maximum speed with 802.11a and 802.11g has always been 54 Mbps. But that might change. Unstrung.com is reporting, based on a conversation with Stuart Kerry, chairman of the IEEE's 802.11 Working Group, that a collection of members called the High Throughput Study Group is working on a potential high-performance standard that would boost both 802.11b (now at 11 Mbps) and 802.11a standards.

While not official yet, this standard for increased throughput might be called 802.11n. Proposals say it could go to 108 Mbps or beyond—as much as 320 Mbps.

The speed increase would take place due to the handling of problems such as lost packets, interference, and other issues that regularly impact WLANs.

This new standard is not expected to be complete until 2005 or 2006.

The optical cables today connect network nodes but usually not the user terminals. The last mile links need higher and higher bandwidth. The high speed utilisation of twisted copper pairs means one promising mode of high bandwidth connection. The wireless connections have also revolutionary development. As a new example the start of 802.11n WLAN standard by 300 bps links can bring new waves in wireless connections replacing the 802. 11b version of 11 Mbps links. The fibre connection can be terminated at small WLAN base stations. The end users can connect by short distance high speed radio link providing real full conform.

In 2003 six chip suppliers have full chip set for the 802.11.b standard and a WLAN card for notebook take a couple of ten USD only.

These rapid developments in technology promise big changes in the field of services as well. Analysts predict competition of 3rd generation mobile systems and WI-FI in US. An example copy of headline news tries to indicate below the stile of starting competition:

Issue 11.05 UNWIRED –
A Wired Special Report - May 2003
Good-Bye 3G - Hello Wi-Fi Frappuccino
Cellcos bet big on third-generation wireless - and took a big hit. Now T-Mobile's John Stanton has a grand convergence plan. Starbucks is just the beginning.
By Dan Briody

According to the opinion of some journalists the 3rd generation mobile systems are failed even before the real start. The WLAN based services using their popular name WI-FI have aggressive rollout plans starting from public places. The T-Mobile in US cooperating with STARBUCKS started an attractive combination of services. Sit down; use the internet with your notebook and the free coffee is only an addition to other services. Phone calls naturally are included in the service set. Such service combination might be really promising in a big country where the mobile penetration and the radio coverage are far behind European figures. Most of the services and applications planned in the 3rd generation mobile will be implemented in WI-FI as well. It is a future question whether free coffee comes to the charged internet or free internet usage to the charged coffee.....

The mobile operators (using the new fashion word "cellco-s") have paid in Europe very big sum of money for the 3rd generation service licences or for frequencies. The new competitors use free frequencies for WI-FI! Might be the high frequency fees spoil cellcos or funny small things like coffee can bring the good balance.

One of the interesting consequences of cheap access to big information might be the end of intellectual proprietary rights. The big market turbulence in downloading or changing music records by internet seems to be only beginning of the whole process. Due to the free bandwidth there is no reason to store at home books, films or CD-s. Downloading for momentary use is simpler. Authors frequently support real global free access to their products.

The free bandwidth can make the globe even more comfortable or sustainable. Instead of sitting in the car and driving to work, learning or enjoy the life the teleworking, telelearning and other applications can spare a lot of fuel or can help to reduce the terrible road traffic.

The downloaded content can be clean from dirty advertisements. The media industry in the downloading age can separate real programs and advertisements.

One of the potential utilisation of infinite bandwidth might be to apply many cameras and propagate their signal. A lot of news can be found on the net concerning cameras.

Not Terrorist but voyeur was on the board of luxury ship

An American luxury ship was evacuated on Thursday evening because one of the passengers informed the staff on a suspicious object with wires in the lavatory.
After carefully search it was recognised that no explosive but small camera was in the lavatory for ladies.

HP is working on wearable digital cameras

HP develops digital cameras which can be wear as a part of the clothing and can continuously record important event of the user's life.

Cameras as part of clothing can be sort according to the cloth stile. So records of events can be sorted as events with jacket, events with jeans, and events with swimming dress....

Telelearning is predicted as a real competitor of traditional contact based education. Complete courses are available free on the net (e.g. MIT OpenCourseWare 2002). So called "professor-less" courses and telelearning communities have been started (peer-to-peer communities of students). Experts and professors will be used mostly in charged teleconsulting. One of possible consequence of the wide spreading telelearning, that the certificatory institutions of knowledge will not be the traditional universities. Teachers mention that students can learn a lot from each other. Especially information technology is a field where fathers can learn from sons.

If the real future is the downloading age how could we find the desired content in the endless WEB? We hope that the development of search engines can follow the

content production. In text based search the application of higher level language technology is inevitable. The association based procedures do not fit to the traditional serial processing methods. Content based search of videos and music records are at the very beginning.

5. Summary

In the age of unlimited bandwidth new perspective are for the development of telecommunication and information technology. Good highways stimulate road traffic so high bandwidth networks will transport more and more information serving even wasteful but useful applications. The listed examples and applications demonstrate the actual situation and help to understand the real direction of development. Conclusions sometimes are extravagant but better to bring bigger umbrella and close it in sunshine than only a very small one and become wet even in small rain.

References

[1] Report of the Transatlantic Networking Committee
http://www.uscms.org/s&c/reviews/doe-nsf/2001-11/docs/TAN-report-final.doc.pdf

[2] dr. Takács György:
Next generation networks – George Gilder gondolatai
PKI Tudományos Napok, Bp., 2002. nov. 19-20.

[3] Wireless Networking
Wireless networking news, publications and reviews
http://bengross.com/wireless.html

[4] Mobility Networks Overview
Bringing WLAN to Macro-networks
http://www.mobilitynetworks.com

[5] Good-Bye 3G – Hello Wi-Fi Frappuccino
http://www.wired.com/wired/archive/11.05/unwired/convergence.html

[6] Georg Gilder:
TELECOSM How the Infinite Bandwidth Will Revolutionize Our World
The Free Press, New York 2000.

[7] David Futrelle:
Was Georg Gilder Blinded by the Light?
www.business2.com/articles/web

[8] Charle Sanders:
The Late, Great Telecosm?
www.spectator.org/AmericanSpectatorArticles

[9] Gary Rivlin:
The Madness of King George
www.j-bradford-delong.net/Stray_Notes/gilder_wired_2002-07.html

[10] David S. Isenberg:
The Dawn of the Stupid Network
www.isen.com/papers/Dawnstupid.html

[11] David Isenberg:
Rise of the Stupid Network
www.hyperorg.com/misc/stupidnet.html

[12] Microgravity Fibre-Pulling Apparatus
www.nasatech.com/Briefs/Dec98/MFS26503.html

ITU-News

The International Telecommunication Union published a **new standard that allows content providers to roll out value-added interactive TV (iTV) services** to any network without modification. ITV allows viewers of a football match for example, to display data on a player while a match is in progress. The standard means that content providers can develop interactive material for programmes that can then be distributed worldwide without extra labour or cost. It means the content will stay true to the author's desingn in all markets – a key concern for advertisers. A key feature for industry is the flexibility of the standard that allows operators to design individual content and easily tag-on interactive content to their programmes.

ITU-T J.202 consolidates the work of other standards makers illustrating ITU-T's position for coordination of Information Communication Technology standards.

The **first global index** to rank Information and Communication Technology (ICT) access has turned up some surprises. Slovenia ties France and the Republic of Korea, usually not among the top ten in international ICT rankings, – comes in in fourth. Apart from Canada, ranked 10th, the top ten economies are exclusively Asian and European. The Digital Access Index (DAI) distinguishes itself from other indices by including a number of new variables, such as education and affordability. It also covers a total of 178 economies, which makes it the first truly global ICT ranking.

Countries are classified into one of four digital assecc categories: high, upper, medium and low. Those in the upper category include mainly nations from Central and Eastern Europe, the Caribbean, Gulf States and emerging Latin American nations. Many have used ICTs as a development enabler and government policies have helped them reach an impressive level of ICT access. This includes major ICT projects such as the Dubai Internet City in the United Arab Emirates, the Multimedia Super Corridor in Malaysia and the Cyber City in Mauritius. The DAI will be a useful tool for tracking the future advancement of these ambitious emerging economies.

Optical burst and packet switching

ZSOLT PÁNDI, PHD. STUDENT

Budapest University of Technology and Economics, Department of Telecommunications
pandi@hit.bme.hu

Reviewed

Keywords: Burst switching, Packet switching, Optical systems, Photonics

The increasing volume and ratio of packet switched traffic poses new challenges on optical transmission technology applied in telecommunications networks. Retaining efficient operation of networks under the changing traffic load calls for development of new solutions. The paper focuses on two research directions of increasing importance: Optical Burst Switching (OBS) and Optical Packet Switching (OPS). The basic principles behind the two emerging technologies are introduced and an attempt is made to determine the possible role of these technologies in networking scenarios. Both theoretical and practical issues concerning the implementation are discussed and some of the proposed solutions to these issues are also introduced.

Introduction

Originated in Public Switched Telephone Networks (PSTNs), traditional optical transmission technology applied in telecommunications networks was exclusively of curcuit switched nature. However, the characteristics of the traffic to be transmitted has been significantly changed in the meantime.

In reaction to the substantial increase of the volume of packet switched traffic and of its ratio to all traffic to be transmitted optical technology was forced to approach solutions formed in computer networks. However, several new issues arose related to this. Some of these issues derived from questions raised by packet switched traffic (such as IP traffic) like transparency, scalability and fine granularity. Moreover, as packet switched traffic in general is hard to characterize, bursty and composed mostly of ephemeral connections transmission technology candidates need to be flexible and they need to facilitate easy reconfiguration. Another set of the newly arised issues was related to the fact that optical transmission technology had not been sufficiently prepared for operation in packet switched mode [1].

Another motivation towards a change of discipline was the phenomenon that utilization of transmission capacities provided by fibers – in part due to the changed characteristics of the transmitted traffic – was comparatively low.

During the research for appropriate answers the Automatically Switched Optical Network (ASON) was born, which is capable of establishing large capacity semipermanent channels in a flexible and easily configurable manner, but still does not offer a satisfying solution for exploiting further optical capacity reserves [2].

That is how photonics reached a point where implementing the principle of packet switching or at least elaborating a solution that handles packet switched traffic in a more efficient way was seriously considered.

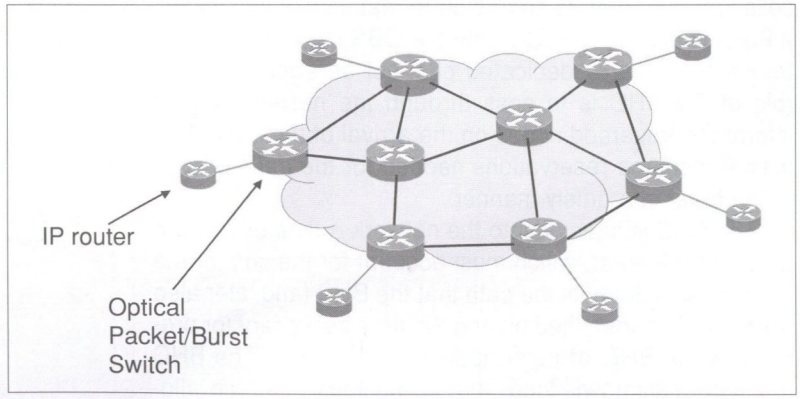
The paper focuses on two emerging solution candidates: Optical Burst Switching (OBS) and Optical Packet Switching (OPS).

Networking scenarios

An optical transmission technology that is capable of handling packet switched traffic in an efficient way may seamlessly cooperate with current network architectures. *Figure 1* depicts an architecture in which an IP-over-OBS ar IP-over-OPS solution is applied. The advantage of such an architecture over the already existing IP-over-SDH or IP-over-WDM solutions is that it not only provides channels for forwarding IP traffic, but it is also capable of adapting to the characteristics of the traffic, thus achieving a more efficient resource utilization [3].

Another proposed architecture is shown in *Figure 2* for MAN scenarios [3]. Optical buffering is not an issue in the network shown, as each node needs to have az E/O converter, which could also be utilized to transfer the burden of buffering from optics to electronics. Another advantage of this architecture is that it recycles the already deployed fibers of metropolitan networks, which often have a ring structure.

Figure 1. IP over OPS/OBS network



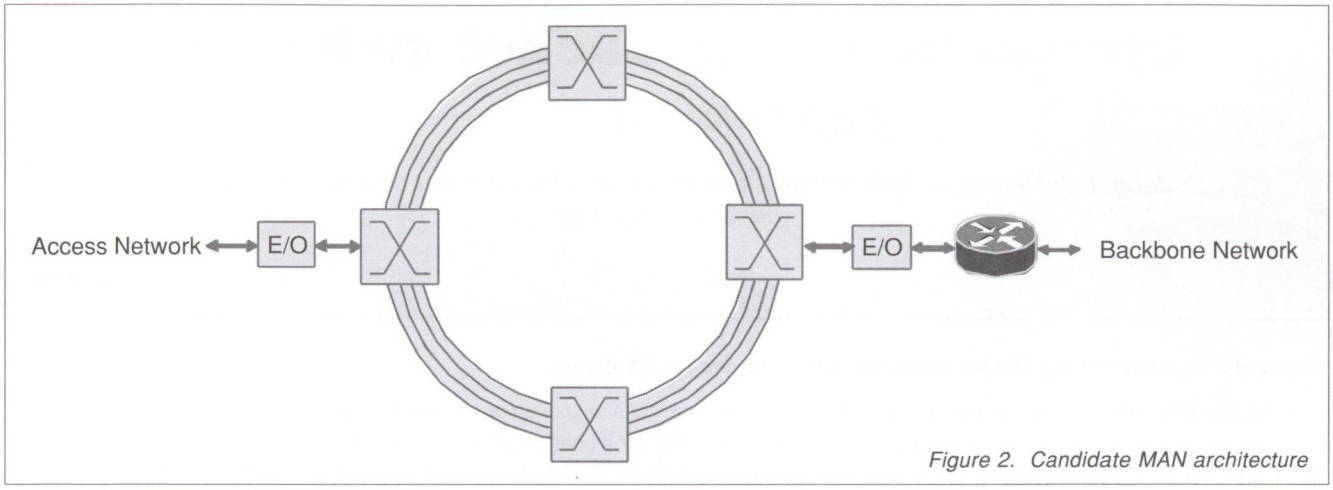


Figure 2. Candidate MAN architecture

As in *Figure 2* the OPS or OBS network is placed in the role of the feeder network in a metropolitan scenario, different LANs, such as university campuses or enterprise networks may be attached to the access nodes, while the rest of the nodes may serve as connection points to the backbone network.

Optical Burst Switching

Optical Burst Switching is an attempt to blend the favorable characteristics of circuit and packet switching in a technology that is based primarily on more or less currently available results of optical transmission related research [4].

OBS networks comprise of two types of nodes: *edge nodes* and *core nodes* and the interconnecting fibers. Edge nodes are the interface of the network towards another network that may use any packet switched technology, while core nodes may only be connected to either core or edge nodes in the OBS network. Between the two types of nodes there are major functional differences.

Edge nodes contain buffers, in which packets are collected that should be forwarded to the same edge nodes of the OBS network. Optical buffering might as well be avoided if buffering capabilities of the technology of the adjacent network are exploited, however, E/O conversion would still be necessary. No matter where buffering takes place, the contents of a buffer is transmitted together as a burst in the OBS network.

When the edge node decides to forward the packets collected in one of its buffers up to that time, it first injects a Burst Header Cell (BHC) into the OBS network, which is transmitted over a dedicated channel on each link. The role of the BHC is to pass through the network and to inform the traversed nodes on the arrival of the burst thus initiate resource reservations needed for the transmission of the burst in a timely manner.

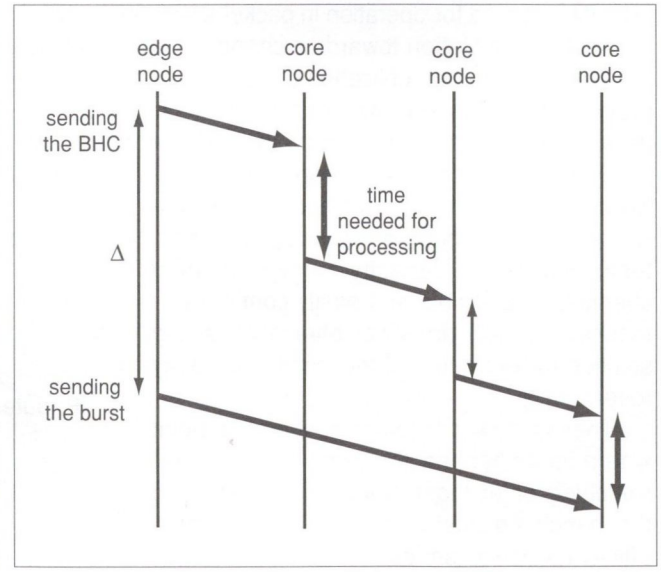
The BHC is injected into the network Δ time before the arrival of the burst, which must account for the sum of the transmission delay of the path that the BHC (and later also the burst) is transmitted on and the time necessary for processing the BHC at each node *(see Figure 3)*. The BHC remains unacknowledged, that is, besides resource allocation no additional measures are taken. The edge node unconditionally injects the burst into the network once the Δ time is over.

The BHC contains the destination (edge) node of the burst, the burst size, the identification of the channel that the burst will arrive on to the node receiving the BHC, and the Δ time.

Having received the BHC, core nodes try to allocate a channel on which the burst can be forwarded to the next node on the path based on the information contained within the BHC and based on the set of channels available at the time of the arrival of the burst. Then they modify the contents of the BHC and their own channel allocation registry, and forward the BHC over a dedicated channel to the next node on the path to be followed by the burst. If no channel can be allocated for the transmission of the burst, the BHC is dropped, consequently, the burst will be lost.

Core nodes, therefore, will already know where to forward the burst by the time it will arrive, and so they have enough time to prepare for the transmission by connecting the appropriate input and output channels, thus avoiding the need for optical buffering. They will release resour-

Figure 3. Sending the BHC and the burst



ces once the amount of data recorded in the BHC passed. This mechanism ensures that the burst itself remains in the optical domain in OBS networks.

To give an estimation for the Δ time let us follow the following train of thoughts. The processing time at each node will be approximately the sum of the times necessary for synchronization to the incoming signal [5], O/E conversion, optical switch setup [6] and generation of the new BHC. If we estimate this sum to be in the range 10-100 μs, and we take into account that light travels 2-20 km over a fiber during this amount of time, we can conclude that transmission delay will be the dominant component in Δ. The foreseen and expected development and speedup of optical devices also supports this argument.

The mechanism described above only illustrates the basic principle of OBS. Several variations of the applied signaling have been elaborated, of which an overview is given in [7] for the interested reader.

Optical Packet Switching

Optical Packet Switching works similarly to packet switching already implemented in electronics [8]. Control information is transmitted together with the packet in its header. As a consequence, packets need to be temporarily stored at each traversed node, at least for the time of processing and regeneration of the header. To facilitate effective exploitation of bandwidths offered by optical transmission solutions involving O/E/O conversions should be avoided, and at least the payload of the packet should remain in the optical domain for the whole time, including storage.

Nodes in the OPS network determine the next node of the path when the header of the packet is being processed, that is, any kind of path selection strategy could be used.

OPS networks fall into two main categories: *slotted* and *unslotted* networks. Slotted networks may only forward packets of the same size, and they also need global clock synchronization, an issue already encountered at the SDH-SONET technology. Unslotted networks, on the other hand, are capable of forwarding packets of arbitrary size.

Figure 4 depicts the functional components of a node in an OPS network. It is important to remark that the figure does not contain the functions necessary for contention resolution, which will be dealt with later on.

Theoretical and practical issues

In what follows the issues will be discussed that need to be resolved in order to be able to implement the introduced technologies. A common characteristic of the preferred solutions is that they omit O/E/O conversion, as this would violate the requirement of transparency, and limit the processing capacity due to the limited operating speed of the electronic devices.

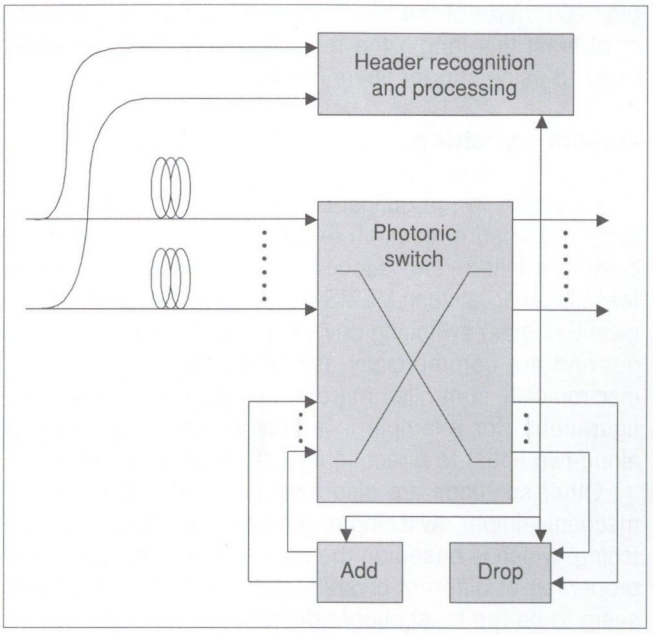


Figure 4. Functional block diagram of an OPS node

Optical buffering

One of the most profound issues is data storage in the optical domain. This capability is necessary at each node in OPS networks, while in OBS networks it might even be spared, as it would only be needed at edge nodes, where there is always another technology present, possibly with buffering capabilities that could be utilized.

Most solutions for optical data storage comprise of Fiber Delay Loops (FDL), which are fibers of different length coiled up, in which the light representing the data to be stored is traveling for the time of the "storage". This time can be controlled by selecting the loop of appropriate length.

The implementations of optical storage based on FDLs may be grouped in different ways. They may be either single-stage or multi-stage, depending on whether the data to be stored is transmitted over a single piece of fiber or more of them during the time of storage. Another type of classification is possible based on the distinction between feed-forward and feedback operation. In the first case data to be stored may only pass each point in the storage device at most once, while in the latter case data may do several loops inside the device.

Figure 5 demonstrates an example for the implementation of an optical storage device [9]. The device shown works as an *n*-port switch with *m* slots available for storage. Data directed to different slots will reappear at one of the switch ports after different time intervals, moreover, delays $d_1..d_m$ may be combined due to the feedback structure.

Applying the estimation mentioned earlier we can suppose that in case of OPS the light representing a packet travels 2-20 km along the fiber while the node is getting prepared for its forwarding. Nevertheless, applying the demonstrated optical storage a single piece of fiber of this length is not necessary, for a packet of 10 kB only occu-

pies 400 meter of the fiber. With the combination of loops of at least this length the required delay might be easily realized using shorter fibers in total.

Optical switching

Switching in the optical domain is probably the single area where more research results of different nature have been published, and successfully implemented offering feasible alternatives. MEMS-based (Micro-ElectroMechanical Systems) switching devices have already been rolled out and are commercially available. These apply small, electronically controlled mirrors arranged in different configurations (for example in a two dimensional matrix or along two lines) to direct light to the appropriate place.

Other solutions are also available, such as traditional mechanical fiber switching or guided-wave solid-state switching, which is based on the controllable light conduction properties of different crystals, but MEMS-based devices seem to be the most successful [8].

Synchronization

Synchronization is a problem of vital importance in slotted OPS networks [8]. On the one hand the receiver needs to synchronize to the clock signal of the incoming packet, and, on the other hand, the output needs to be aligned with the slot timing. The first is facilitated by introducing the *guard time*. The header and the payload of the packet does not fill completely the available time slot, but in between the end points of the time slot, and the header and the payload of the packet there are unused time ranges, called guard time. These are present to help the recevier to synchronize to the signal.

Output synchronization may be achieved in multiple ways. As an example, N 2x2 optical switches in a serial configuration may be used for producing delay with a granularity of $1/2^n$ packet length, if the i^{th} switch either switches the signal directly to its neighbor or diverts it to a fiber loop of $1/2^i$ packet length before forwarding it. Another feasible solution may consist of a strongly dispersive fiber loop and a wavelength converter. Making use of the different propagation speed of light at different wavelengths to produce the required delay it just converts the incoming signal to the appropriate wavelength.

Unfortunately, both solutions have drawbacks: the series of switches may cause severe degradation of signal-to-noise ratio, while the wavelength converter-based solution has a limited granularity.

Optical wavelength conversion

Optical wavelength conversion is necessary both in OBS and OPS to ensure full exploitation of optical resources. Devices providing optical wavelength conversion are still at the stage of laboratory experiments. Most of the researchers investigate the applicability of non-linear effects, which may be grouped as Kerr-effects and scattering effects [10].

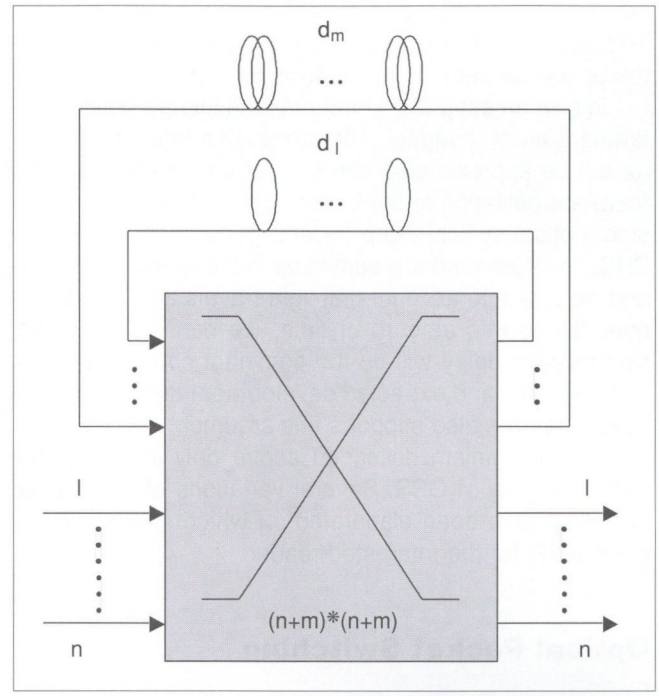


Figure 5. Single-stage, feedback optical storage

There are three different phenomena termed as Kerr effect, each of which is based on the fact that the refraction index of the fiber core changes depending on the intensity of the light to be transmitted. In fact, this is the Kerr-effect itself, which may cause self phase modulation (spreading of a wavelength to adjacent wavelengths), cross-phase modulation (spreading of multiple wavelegths due to their influence on each other) and four-wave mixing (when two or more wavelengths together combine a new wavelegth).

Scattering effects could belong to one of the following two types. In case of Raman stimulated scattering light loses energy when photons collide with molecules of the fiber and this energy is emitted in the form of light at a longer wavelength. In case of Brillouin stimulated scattering light may cause the emission of acoustic waves in the fiber, which may cause light to scatter to different wavelengths.

There are also alternate solutions, such as those applying Semiconductor Optical Amplifiers (SOA), but they are omitted from the present discussion. The interested reader may find additional information and a comprehensive coverage of possible solutions in [11].

Contention resolution

Contention occurs if two different packets or bursts need to be transmitted on the same output channel [8]. This concurrence may be resolved in three dimensions: time (using an optical storage), wavelengths (using an optical wavelength converter) and space (using deflection routing).

The latter may help avoid using optical storage, as well, as in case of hot-potato routing, for example. However, it is important then to control the lifetime of packets or bursts

using a time stamp and not the TTL (Time To Live) header field, which is already successfully applied in IP networks, as the latter would require the modification of the header at each node that the packet or burst passes.

Conclusion

The paper attempted to give an overview of the present status of two current development directions of optical fiber transmission technology: Optical Burst and Packet Switching. It included theoretical and practical issues concerning the implementation of these technologies and also some proposed solutions for these issues.

Discussing every research direction would have been impossible, still it is worth mentioning some of the most important ones of those omitted, that still represent a challenge for researchers: selecting the appropriate size of packets or bursts, handling priorities, transmission of multicast traffic, fault-tolerant route selection, preliminary mathematical performance analysis, developing of even faster switch fabrics and implementation of an optical RAM.

In general, the developments are aiming at solutions without O/E/O conversion in the signal path thus arriving at an architecture that will consist of a smaller numer of layers, in which the optical layer will have an increasing importance and functionality.

Optical Burst Switching unifies the advantages of packet switching and circuit switching and being largely based on already existing technology it provides an answer for the challenge posed by the transmission of packet switched traffic. However, for the implementation of Optical Packet Switching several open issues still need to be resolved.

Acknowledgements

The author wishes to thank Tivadar Jakab, Zsolt Lakatos and Gábor Horváth for their help in studying the literature, and Dr. Tien Van Do for his ideas helping the work. The author also expresses his gratitude to Dr. György Lajtha for his suggestions that considerably improved the overall quality of the paper.

References

[1] Zs. Pándi,
Introduction to Optical Burst and Packet Switching,
Third Hungarian WDM Workshop,
Budapest, Hungary, April 2003

[2] B. F. Craignou et al.,
Network Operator Perspectives on Optical Networks – Evolution towards ASON,
10th International Telecommunication Network Strategy and Planning Symposium (Networks 2002),
Munich, Germany, June 2002

[3] A. Jourdan et al.,
The Perspective of Optical Packet Switching in IP-Dominant Backbone and Metropolitan Networks,
IEEE Communications Magazine, March 2001

[4] J. S. Turner,
Terabit Burst Switching,
Journal of High Speed Networks, Volume 8 (1999)

[5] V. W. S. Chan et. al.,
Architectures and Technologies for High-Speed Optical Data Networks,
IEEE Journal of LightwaveTechnology,
Vol.16., No. 12., pp.2146-2168, December 1998

[6] Q. Yang et. al.,
WDM Packet Routing for High-Capacity Data Networks,
IEEE Journal of LightwaveTechnology,
Vol.19., No. 10., pp.1420-1426, October 2001

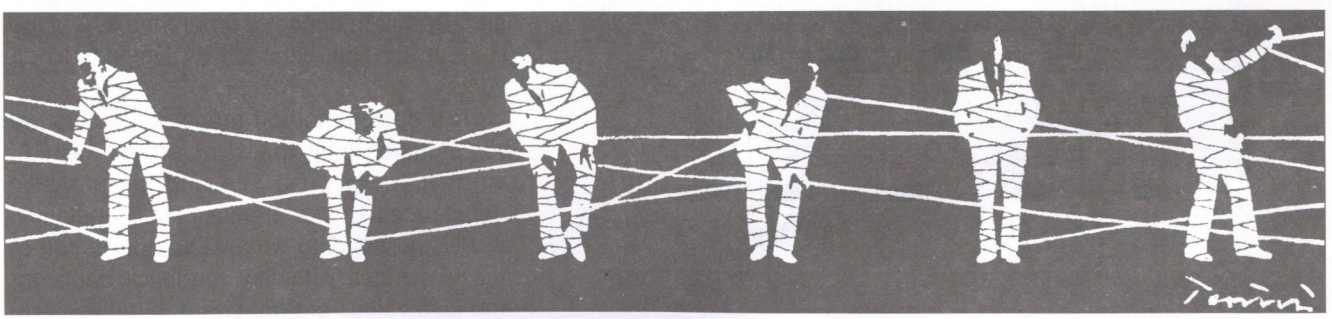
[7] M. Nord et. al.,
OPS or OBS in the Core Network? (COST 266),
Proceedings of the 7th IFIP Working Conference on Opical Network Design & Modelling,
Budapest, Hungary, February 2003

[8] S. Yao et al.,
Advances in Photonic Packet Switching: An Overview,
IEEE Communications Magazine, February 2000

[9] M. J. Karol,
A Shared-Memory Optical Packet (ATM) Switch,
6th IEEE Workshop on Local and Metropolitan Area Networks, San Diego, CA, USA, October 1993

[10] D. Penninckx et al.,
New Physical Analysis of 10 Gb/s Transparent Optical Networks, IEEE Photonics Technology Letters, Volume 15, Issue 5, May 2003

[11] J. M. H. Elmirghani et al.,
All-Optical Wavelength Conversion:
Technologies and Applications in DWDM Networks,
IEEE Communications Magazine, March 2000



Design of wide band distributed amplifiers

ATTILA ZÓLOMY

Budapest University of Technology and Economics
Department of Broadband Infocommunication Systems

Reviewed

Keywords: Broadband amplifiers, Engineering methods, Design of extremely high frequency components

Due to the continuos demand for increased transmission speed, parallel connection of several WDM channels would be necessary to build up a high-speed lightpath between two end nodes of an all-optical network. In this case the electrical circuits of the applied optical transmitters and receivers must have extraordinary bandwidth, and even a so simple function like amplification can be critical. As the distributed amplifier (DA) has the highest bandwidth among the known amplifier structures, it can be the ideal choice. The paper presents a new design method for DAs comprises interstage transmission lines (TLs) between the active devices. It is shown that by proper design of the applied interstage TLs the image impedance and the cut-off frequency (COF) of the input or output line of the DA can be independently tuned. By this way, the phase mismatch between the lines of the DA and thus the gain characteristic can be varied independently from the termination conditions and from the active device's capacitances. In addition to this, the phase mismatch can be influenced by the proper choice of the connection structure of the active device and by the lead inductance value as well.

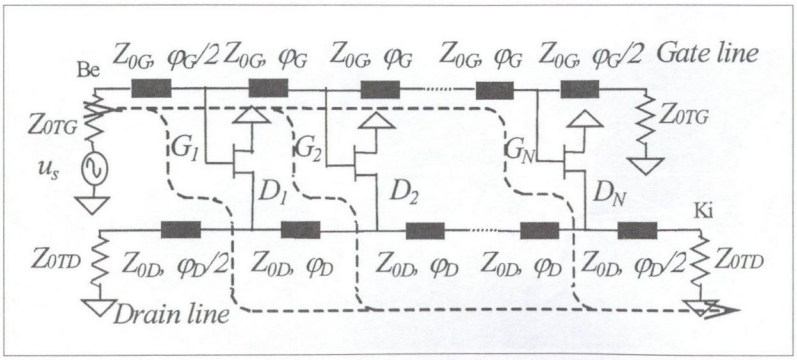
Introduction

The schematic of an N stage DA is shown in *Figure 1.* The K ladder structure of the input line is built up from the input admittance of the active devices and from the interstage inductances or TLs. The output signals of the active devices propagate to the matched right hand side output termination, are collected in phase by the output line.

The extraordinary bandwidth of this amplifier type was demonstrated in several papers. The highest achieved upper bandwidth is 110 GHz [1] using GaAs monolithic technology. Most of the results were yielded mainly by the development of the technology. Huge amount of publications in the field [2] describes only analyses methods and there is a lack of effective synthesis methods. In practice usually computer optimization is applied during the design. However, in presence of the transistor parasitics and especially at high stage numbers it is difficult to find the global minimum of the optimization's error function due to the existence of several local minimums.

More efficient if the effect of the most relevant transistor parasitics is analyzed and than compensated during a systematic design process, which results in a simplified, but more or less well operating structure. This structure can be used as a starting point for the sophisticated computer optimization includes the effect of all parasitics. This approach is used in [3], where the unique systematic design method published in [4] was used as a starting point for the annealing type computer optimization. The mentioned design method is manipulating the active device's losses to eliminate the pole of the power gain at the upper end of the passband in case of phase synchronization between the input and output line. For this, mainly the increase of the input loss (input loss/capacitance ratio) is required, which may increase the noise and restricts the freedom during the active device design. The method is based on analytical calculations were performed on a simplified, lumped element DA structure having an unilateral transistor model, which comprises the input and output capacitances and losses. Phase synchronization is assumed between the lines, which -due to the usually very different input and output capacitance values of the active devices- can be difficult to achieve without introducing further parasitics, in a lumped element DA, if it works between the same (50 Ω) termination impedances. For the design graphical curves derived from analytical expressions, are used. They are accurate at high stage numbers (N≥4). According to the above mentioned reasons the method is well applicable for monolithic DAs at frequencies where the interstage TLs can be approximated by lumped Π networks.

Figure 1. Schematic of DA



The design method presented in this paper modifies only the structure of the DA, which may comprise TLs and the parasitic inductances of the transistor's connections and/or the bonding wires. The application of active devices with low feedback capacitor and loss values are important, but these pro-

perties are also essential in high frequency low noise transistors, anyway. As the active devices are not manipulated, the possibilities to achieve that are wider. The flat gain characteristic is yielded by the proper adjustment of the phase mismatch between the input and output lines of the DA. Good results can be achieved even without the usage of computer optimization methods.

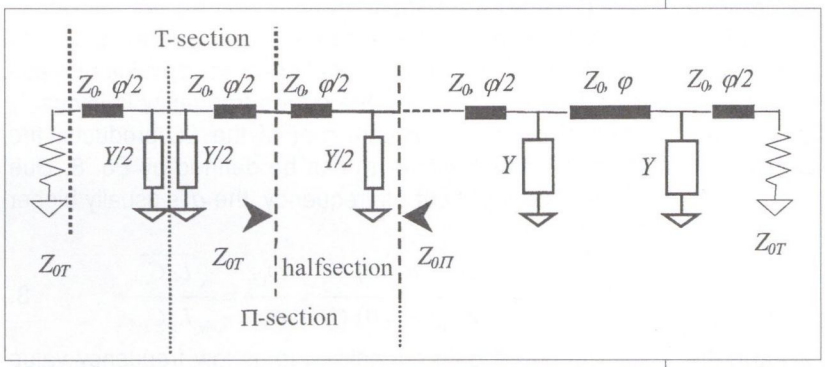


Figure 2. Structure of the lines of the DA

Theoretical background

The general ladder structure of the DA is shown in *Figure 2.* In lumped element case the interstage TLs are replaced by lumped inductances. The ladder structure can be divided into equal T- or Π-sections or further to half sections and can be described easier by the chain (ABCD) matrix.

The chain matrix of any section can be derived from the chain matrix of the building blocks in it, e.g. in case of half section:

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} \cos\left(\frac{\phi}{2}\right) & jZ_0 \sin\left(\frac{\phi}{2}\right) \\ jY_0 \sin\left(\frac{\phi}{2}\right) & \cos\left(\frac{\phi}{2}\right) \end{bmatrix} \begin{bmatrix} \frac{Y}{2} & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix} \qquad 1.$$

In case of lumped element DA the chain matrix of the TL is replaced by that of the inductance in Eq. 1. Knowing the chain matrix any transfer function (voltage-, current-, power gain, phase shift etc.) or the Z_{0T}, $Z_{0\Pi}$ image impedances can be derived very easily:

$$Z_{0T} = \sqrt{\frac{B}{D}\frac{A}{C}}; \quad Z_{0\Pi} = \sqrt{\frac{D}{C}\frac{B}{A}}; \quad e^{-\Theta} = \sqrt{\frac{V_2}{V_1}\left(\frac{-I_2}{I_1}\right)}, \qquad 2.$$

$$\Theta = ar\cosh\sqrt{AD}, \quad \frac{V_2}{V} = \sqrt{\frac{D}{A}}\frac{1}{\sqrt{AD}+\sqrt{BC}}, \quad \frac{-I_2}{I_1} = \frac{A}{D}\frac{V_2}{V_1}$$

where ABCD is the elements of the chain (ABCD) matrix.

Knowing the transfer functions of the input and output T-sections the transfer functions of the k[th] signal path of the DA to the right hand side termination, which goes through the k[th] active device (Figure 1.) can be derived. E.g. the voltage gain is given by Eq. 3.

$$V_K = V_{TG}^{(K-1)} V_{TD}^{(N-K)} \qquad 3.$$

where V_{tr} is the transfer function of a so-called transfer T-section, which has the input at the input (denoted by G) line and has an output at the output (denoted by D) line and involves the whole k[th] transistor. The voltage gain of the whole DA is the sum of all signal paths.

$$V = V_{Tr} \sum_{K=1}^{N} V_{TG}^{(K-1)} V_{TD}^{(N-K)} \qquad 4.$$

In the most simple case, called ideal case, the active device comprises only the input and output capacitances and the transconductance (*gm*) and in the DA lumped element interstage inductances are applied.

The lines of the DA are simple LC ladder networks. The general form of the image impedances, propagation factor and the power gain of an ideal DA is given by Equations 5.

$$P = \frac{gm^2}{4} Z_{0\Pi G} Z_{0\Pi D} e^{re(\theta_G - \theta_D)} e^{-2Nre(\theta_D)} \sum_{K=1}^{N} e^{K(\theta_D - \theta_G)} \sum_{K=1}^{N} e^{K(\theta_D - \theta_G)^*}$$

$$Z_{0r} = \sqrt{\frac{L}{C}\left(1 - \frac{\omega^2}{\omega_c^2}\right)}, \quad Z_{0\Pi} = \sqrt{\frac{L}{C}\left(1 - \frac{\omega^2}{\omega^2}\right)^{-1}}, \qquad 5.b$$

$$\Theta = ar\cosh\left(1 - \frac{2\omega^2}{\omega_c^2}\right), \quad \omega_c = \frac{2}{\sqrt{LC}} \qquad 5.c$$

One can observe that the T and Π image impedance is zero and infinity at the cut-off frequency, respectively. The propagation factor (denoted by θ) is pure imaginary below and pure real above the cut-off frequency. The power gain is proportional to the Π image impedances of the input and output line, to the *gm* of the active devices, and to a sum part, whose frequency response is only depends on the difference between the input and output propagation factors (phase mismatch). As the θ is pure imaginary below cut-off the value of the middle parts are 1. In case of phase synchronisation the gain-bandwidth product (GBP) is proportional to the image impedance-bandwidth product if the frequency dependence of the *gm* is neglected. In this case the power gain has a pole at cut-off. The effect of this pole can be compensated by the proper adjustment of the phase mismatch. However, as the image impedances are proportional to the *L/C* ratio, in case of fixed transistor capacitances and termination impedances, the *L* values and thus the cut-off frequencies i.e. the θ of the lines are determined. Hence, the phase mismatch is also determined and can not be varied. But, if one allow a slight input and output mismatch and change the image impedance values, a limited variation of the phase mismatch can be yielded.

Impedance mismatch problems of DAs

Due to realization problems the lines of the practical amplifiers usually ends in T-section. Hence, the amplifier shows a monotonically decreasing impedance at its input and output. The magnitude of the input (or output) reflec-

tion in case of a Z_L (constant, frequency independent) termination impedance is given by Eq. 6.

$$|\Gamma(\omega)| = \left|\frac{Z_{be}(\omega) - Z_L}{Z_{be}(\omega) + Z_L}\right|, \quad Z_{be}(\omega) = Z_{0T}(\omega) \qquad 6.$$

Hereinafter, the so-called operation bandwidth (OB) of the amplifier, in which the reflection is less than a specified value will be determined. Assuming that $Z_{in}(\omega) \leq Z_L$ around the critical frequency, by substituting the expression of Z_{0T} (eq. 5.b) an operation frequency can be determined up to which the reflection is better than the specified limit in case of a Z_L termination impedance. Expression of the operation frequency relative to the cut-off is given as follows (Eq. 7.)

$$\omega_{\Gamma rel} = \frac{\omega_\Gamma}{\omega_c} = \frac{\sqrt{(1+|\Gamma|)^2 - (Z'_L)^2(|\Gamma|-1)^2}}{(1+|\Gamma|)}, Z'_L = \frac{Z_L}{Z_{AF}} = \frac{Z_L}{\sqrt{\frac{L}{C}}} \qquad 7.$$

E.g. the termination impedance is equal to Z_{LF} and the reflection must be better than –20 dB, the amplifier can be used up to the 57.5% of the lower cut-off of it's lines. This value can be increased by the decrease of the Z_L, but it degrades the originally perfect matching at the low frequency linear section of the image impedances. Thus, the decrease of Z_L can be continued until the reflection remain below the specified limit.

The relative operation frequencies at different reflection values are given by *Table*. In case of first row Z_L=Z_{LF}. In case of second row the Z_L is reduced to the limit. It can be observed that the increase of the relative operation frequency is significant especially at the more restricted reflection limit. As the cut-off frequency is unchanged the increase of the operation frequency is proportional to the increase of the relative operation frequency.

	$\omega_{\Gamma rel}$	
	$\Gamma = -10\,dB$	$\Gamma = -20\,dB$
If Z_L=Z_{AF}	0.854	0.575
if $Z_L = Z_{AF}\frac{1-\Gamma}{1+\Gamma}$	0.963	0.741
$\sum\alpha$ transmission loss	0.9 dB	0.086 dB

However, the introduction of the mismatch reduces the gain due to the effect of the transmission loss. Above a reflection value (approx. –12dB) the decrease of the gain is higher than the increase of the operation frequency, and thus the resulted gain-operation bandwidth product (GOBP) is decreasing.

Similarly to the previous case the relative operation frequency can also be increased by the increase of the image impedance (*L/C* ratio), which usually achieved by the increase of the *L*. But, due to the increase of the *L/C* ratio the gain become higher and parallel with it due to the increase of the *LC* product the cut-off frequency is decreasing. Finally, it can happen that despite of the higher relative operation frequency the real operation frequency become lower.

The mentioned methods are practical to apply at the input line due to its lower bandwidth resulted by the usually higher input capacitances of the active devices.

Effect of phase mismatch

Due to the usually much higher input capacitance of the applied transistors, there is a significant phase mismatch between the input and output lines. As the phase mismatch can be eliminated i.e. the *LC* products can be equalized by the proper variation of the *L* values (image impedance values) and the cut off frequency is inversely proportional to the square root of the *LC* products, the phase mismatch factor (*q*) can be defined by Eq. 8. Due to the lower input cut-off frequency, the *q* is usually bigger than 1.

$$q = \frac{Z_{0TG}(\omega=0)}{Z_{0TD}(\omega=0)}\frac{C_G}{C_D} = \frac{\omega_{CD}}{\omega_{CG}} = \frac{\sqrt{L_G C_G}}{\sqrt{L_D C_D}} \qquad 8.$$

The power gain normalized to its low frequency value is given by Eq. 9. In case of perfect input and output impedance matching.

$$\Delta P(\omega) = \frac{P(\omega)}{P(0)} = \frac{1}{N^2}\frac{e^{re(\theta_G-\theta_D)}e^{-2Nre(\theta_D)}}{\left(\sqrt{1-\omega^2\frac{L_D C_D}{4}}\right)^*\sqrt{1-\omega^2\frac{L_G C_G}{4}}}\left|\sum_{K=1}^{N} e^{K(\theta_G-\theta_D)}\right|^2$$

The last term describing the effect of the phase mismatch is the sum of a geometric progression. Taking into account this and after rearranging, Eq. 9 can be expressed as a function of q and the relative frequency (wrel), which is the frequency normalized to the cut-off of the input DA line as usually it is the lower.

$$\Delta P(\omega_{rel}) = \frac{P(\omega_{rel})}{P(0)} = \frac{1}{N^2}\frac{1}{\left(\sqrt{1-\omega_{rel}^2\frac{1}{q^2}}\right)^*\sqrt{1-\omega_{rel}^2}}\left|\frac{\sin\left(\frac{N}{2}\Delta\theta\right)}{\sin\left(\frac{\Delta\theta}{2}\right)}\right|^2 \qquad 10.$$

where:

$$\Delta\theta = \theta_D - \theta_G = a\cos h\left(1-\frac{2\omega_{rel}^2}{q^2}\right) - a\cos h\left(1-2\omega_{rel}^2\right)$$

It can be observed that the frequency response is only depends on *q* and the stage number (*N*). *Figure 3.* shows Eq. 10 in log scale at several stage numbers in case of *q=2 (C_G=0.4 pF, Z_0=50 Ohm).*

On the X-axis the normal frequency is given. The frequency response is flat only at a given stage number as the phase mismatch compensates the peak around cut-off only at that case.

By the decrease and increase of *N* ascending or descending of the gain can be observed at the upper section of the passband, respectively. By expanding the sin functions in Eq. 10 into taylor series and neglecting the higher order terms an approximate analytical expression can be derived for either *N* or *q*. Eq. 11 is the equation for the *N* to obtain a specified ΔP gain deviation at a given ω_{rel} relative frequency. It is shown in *Figure 4.* vs. the relative frequency at several ΔP gain deviations.

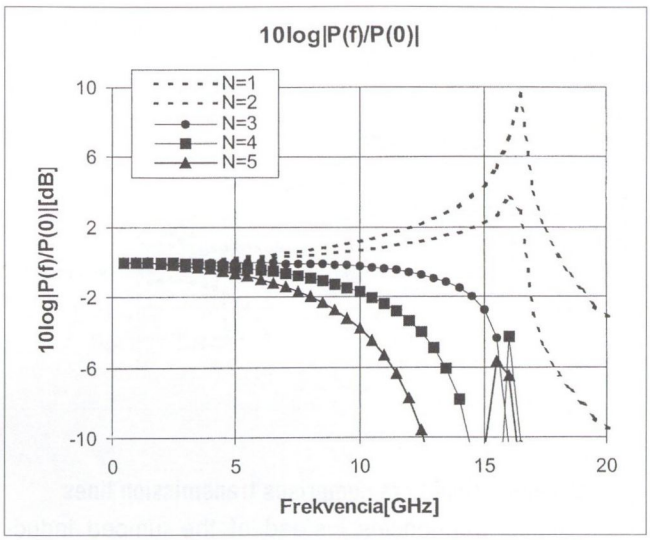


Figure 3. Normalised power gain

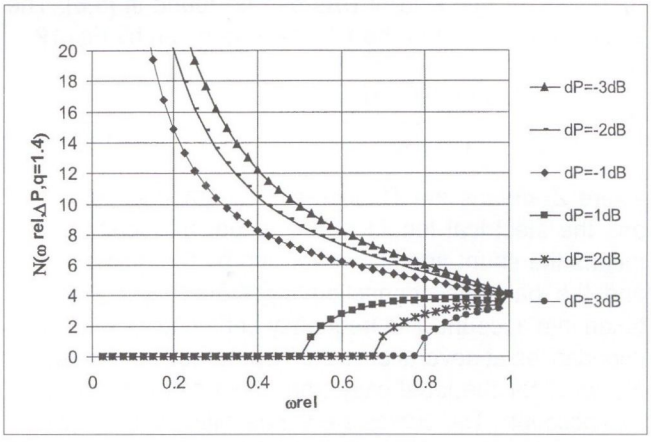


Figure 4. N vs. ω_{rel} to achieve a specified ΔP gain deviation

$$N(\omega_{rel}) = \sqrt{\frac{40 - \sqrt{-320 + (1920 - 80\Delta\theta^2 + \Delta\theta^4)(1-\omega_{rel}^2)^{\frac{1}{4}}\left(1 - \frac{\omega_{rel}^2}{q^2}\right)^{\frac{1}{4}}\sqrt{\Delta P}}}{\Delta\theta^2}}$$

As it can be observed at higher relative frequency values the *ΔN* range, in which the gain deviation is less than the actual *ΔP* is become narrower. Flat gain at the whole passband (up to $\omega_{rel}=1$) can only be achieved at an ideal *N* value (*N=4* in Figure 4), which is depends on the *q* value. Higher the *q* value, lower the ideal *N* and vica-versa.

Effect of parasitic inductance of transistor connections

The connections of the active devices to the lines of the amplifier can be modelled well by series parasitic inductances. The effect of them is significant especially at the input line on the one hand due to the higher input capacitance and on the other hand due to the usually long narrow metal lead necessary in FET devices to connect the gate electrode.

The structure of a line of a lumped element DA comprises the parasitic inductances is given in *Figure 5.*

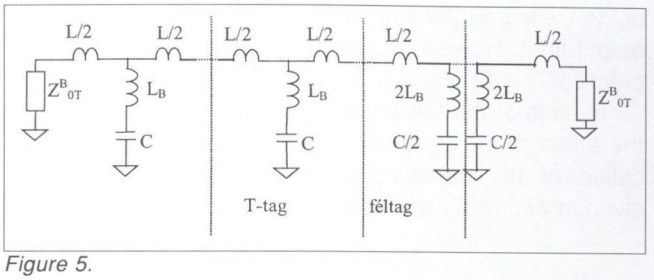


Figure 5.

The image impedances, propagation factor and cut-off frequency is given by Eq. 12.

$$Z^B_{0T} = \sqrt{\frac{L}{C}\left(1 - \omega^2\left(L_B C + \frac{LC}{4}\right)\right)}, Z^B_{0\Pi} = \sqrt{\frac{L}{C}}\frac{1-\omega^2 L_B C}{\sqrt{1-\omega^2\left(L_B C + \frac{LC}{4}\right)}}$$

$$\theta^B = ar\cosh\left(1 - 2\left(\frac{\omega^2 LC}{4\left(1-\omega^2 C L_B\right)}\right)\right), \omega_c^B = \frac{1}{\sqrt{C\left(L_B + L/4\right)}}$$

As it can be concluded the parasitic inductance reduces the cut-off frequency radically and increases the propagation factor (the phase shift per section). According to this, the phenomena is especially harmful at the input line where it further decreases the critical cut-off frequency, and thus the GOBP. Similarly to the ideal DAs the normalised power gain and the expression for the allowed stage number (similar to Eq. 11) can be derived. They are given by Eq. 13 and Eq. 14.

$$\Delta P(\omega_{rel}) \approx \frac{1}{\left(1-\omega_{rel}^2\right)^{3/2}}\frac{q^3}{\left(q^2-\omega_{rel}^2\right)^{3/2}}\left|\frac{\sin\left(\frac{N}{2}\Delta\theta\right)}{\sin\left(\frac{\Delta\theta}{2}\right)}\right|^2$$

$$N(\omega_{rel}) = \sqrt{\frac{40 - \sqrt{-320 + (1920 - 80\Delta\theta^2 + \Delta\theta^4)(1-\omega_{rel}^2)^{\frac{3}{4}}\left(1 - \frac{\omega_{rel}^2}{q^2}\right)^{\frac{3}{4}}\frac{\sqrt{\Delta P}}{q}}}{\Delta\theta^2}}$$

where the *q* is also the ratio of the line cut-off frequencies:

$$q = \frac{\omega^B_{cD}}{\omega^B_{cG}} = \frac{\sqrt{C_G\left(L_G + 4L_{BG}\right)}}{\sqrt{C_D\left(L_D + 4L_{BD}\right)}} \qquad 15.$$

Similarly to Figure 4 the *N* (Eq. 14) vs. relative frequency could also be plotted at several *ΔP* values. If it is done for the same *q* value (1.41) and for $L_B=0.4nH$, the flat gain characteristic can be obtained at *N=7* instead of *N=4* of the ideal case. It means that in the presence of parasitic inductances stronger compensation of the gain peak at cut off is necessary. Beside that, due to the descent of the cut off frequency, the same relative frequency value means lower physical frequency value.

Fortunately, the phenomenon is useful if the primary aim is to increase the gain due to the ascent of the optimum stage number. Beside that, the optimum *N* can be further increased by the proper reduction of *q*, which can be obtained by increasing the parasitic inductance value

of the DA line with the higher cut-off frequency (usually the output line). However, accurate tuning of L_B is difficult in practice.

In case of the DA line with the lower cut-off (input line) the effect of the L_B must be minimised. For this the application of the so-called V-shape connection structure, given in *Figure 6.*, is practical.

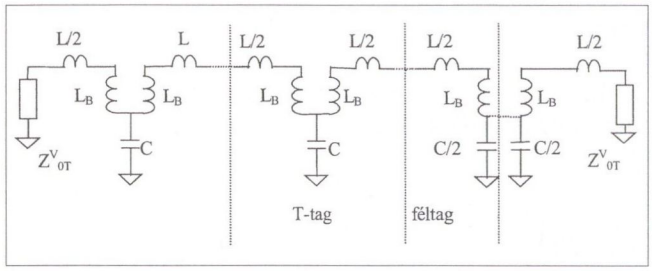


Figure 6. Structure of the V-shape connection

The V-shape connection structure connects the incoming and outgoing interstage inductances (or TLs) to the pin of the active device separately. By this way the parasitic inductances are transformed to the series arm from the shunt arm of the DA line ladder network. The image impedances and the propagation factor can be derived by the previously mentioned method. E.g. the T image impedance and the cut-off frequency is given by E.q. 16.

$$Z_{0T}^{V} = \sqrt{\frac{2L_B + L}{C}\left(1 - \frac{\omega^2(2L_B + L)C}{4}\right)}, \omega_c^{V} = \frac{2}{\sqrt{(L + 2L_B)C}} \quad 16.$$

One can observe, the L_B increases the value of the image impedance and parallel with this the descent of the cut off frequency is not so radical. Beside that, the effect of the L_B can be reduced or may be eliminated by subtracting it from the interstage inductance.

The investigations of this chapter suggests that a so-called hybrid structure, in which a V-shape connected input line and a normal connected output line structure is used, is practical for high frequency active devices applied nowadays. If the interstage inductance of the normal connected output line is neglected beside the effect of the L_{BD} the normalized gain and the expression for the necessary N to achieve ΔP gain deviation at a given ω_{rel} relative frequency is given by Eq. 17 and Eq. 18, respectively.

$$\Delta P(\omega_{rel}) \approx \frac{1}{\left(1-\omega_{rel}^{\ 2}\right)^{1/2}} \frac{q^3}{\left(q^2-\omega_{rel}^{\ 2}\right)^{1/2}} \left| \frac{\sin\left(\frac{N}{2}\Delta\theta\right)}{\sin\left(\frac{\Delta\theta}{2}\right)} \right|^2$$

$$N(\omega_{rel}) = \sqrt{\frac{40 - \sqrt{-320 + (1920 - 80\Delta\theta^2 + \Delta\theta^4)\left(1-\omega_{rel}^2\right)^{\frac{1}{4}}\left(1-\frac{\omega_{rel}^2}{q^2}\right)^{\frac{3}{4}}\frac{\sqrt{\Delta P}}{q}}}{\Delta\theta^2}}$$

The N (Eq. 18) vs. ω_{rel} is shown in *Figure 7.* for the previously used q *(1.41)* and L_B *(0.4nH)* values. As it can be observed flat gain is obtained at *N=5*, which is between the optimum N value of the ideal *(4)* and normal connected case (7).

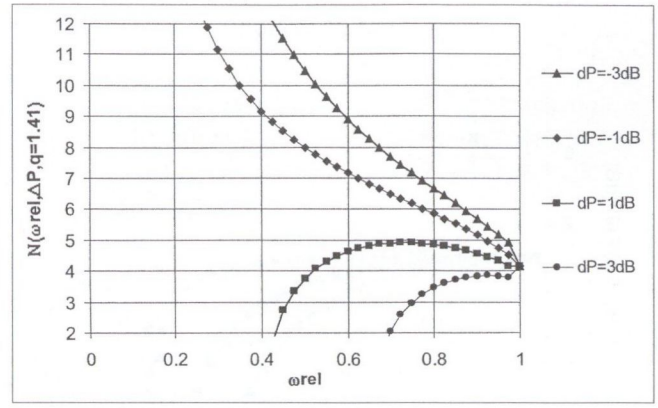


Figure 7.

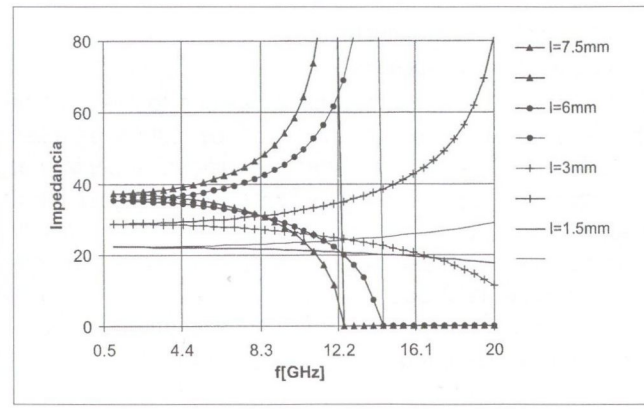
Distributed amplifiers comprises transmission lines

At higher frequencies instead of the lumped inductances application of interstage TLs are necessary. Several results of this kind of DAs can be found in [5,6]. The image impedances of the DA lines are given by Eq. 19.

$$Z_{0T} = Z_0\sqrt{\frac{2_j\sin(\varphi) + Z_0 Y(\cos(\varphi) - 1)}{2_j\sin(\varphi) + Z_0 Y(\cos(\varphi) + 1)}}, Z_{0\Pi} = \frac{2Z_{0T}}{2 + jYZ_0 tan(\varphi)}$$

where Z_0 means the TL characteristic impedance, φ means the electrical length at the actual frequency and Y means the shunt admittance formed by the active device and the parasitic connection inductance (L_B) if that is taken into account. The frequency behaviour of the image impedances at several physical TL lengths (l) are shown in *Figure 8.* for the ideal case (the effect of L_B is not taken into account). The curves are very similar to the curves of the lumped element case, i.e. the Π-impedance has a pole the T-impedance has a zero at the cut-off frequency. Beside that, the lower the physical length the higher the cut-off frequency and lower the image impedances. This is due to the fact that the distributed inductances of the TLs decrease and thus the effect of the fixed transistor capacitances increases. The image impedances can be maintained on a fixed value independently of the l by the proper tune of the TL characteristic impedance (Z_0) i.e. in case of shortening the Z_0 must be higher to make the TL more inductive (to decrease it's distributed capacitance and increase it's distributed inductance).

Figure 8.
Image impedances of a DA line at different TL physical length



The necessary Z_0 value can be computed from the condition, that the DC limit value of the image impedances must be equal to the desired impedance rate (k). Eq. 20 is the expression derived by this way for the ideal and normal connected DA line and Eq. 21 is the equation for the V-shape connected case.

$$Z_0(k,l,C)=\frac{1}{2l}\left(k^2Cv+\sqrt{k^4C^2v^2+4l^2k^2}\right) \qquad 20.$$

$$Z_0(k,l,C,L_B)=\frac{1}{2l}\left(k^2Cv-2L_Bv+\right. \qquad 21.$$
$$\left.+\sqrt{k^2C^2v^2+4l^2k^2+4L_B^2v^2-4L_Bv^2k^2C}\right)$$

It can be shown that the above expressions are monotonically increasing during the shortening of l and tend to infinity (i.e. the TLs become lumped inductances) if l approaches 0. The propagation factor can be calculated by Eq. 22 and for the V-shape connected case by Eq. 23 after the substitution of Eq. 20 and Eq. 21, respectively.

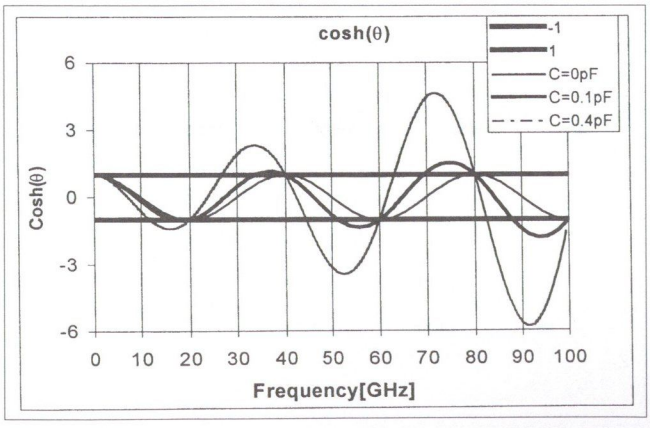
$$\theta=ar\cosh\left(\cos(\varphi)+j\frac{YZ_0}{2}\sin(\varphi)\right) \qquad 22.$$

$$\theta=ar\cos h\left(\left(\cos(\theta)(1+YZ_B)+\right.\right.$$
$$\left.\left.+j\sin(\theta)\left(\left(2Z_B+Y\left(Z_0^2+Z_B^2\right)\right)/2Z_0\right)\right)\right. \qquad 23.$$

The cut-off frequency can be derived from the condition that the propagation factor must be pure imaginary i.e. the magnitude of the argument of the arcosh function must be equal or less than 1. However, these condition leads to transcendental equations for the cut-off frequency, which can not be solved in closed form. An approximate solution can be yielded e.g. by graphical way, which is shown in *Figure 9.* at several capacitance values for the ideal case ($L_B=0$). As it can be seen at nonzero capacitance values several passbands exists. At the beginning of the n^{th} passband the electrical length of the TLs must be $n\pi$ (see Eq. 22), thus the starting frequencies of the passbands are inversely proportional to the l.

The starting frequencies of the passbands can only be approximated. In theory, the higher passbands can also be used for amplification, but in practice several difficulties must be overcome, hence the first passband is usually used and thus will be investigated further.

Figure 9.



It can be shown, that the cut–off frequency is monotonically increases and tends to the cut-off of the lumped element line with the same shunt capacitance and image impedance value as the l tends to zero. This is demonstrated by *Figure 10.* and *Figure 11.*, where the cut-off frequency vs. l is shown for the ideal case and for the V-shape connected line ($L_B=0.3$ *nH*) at several capacitance values ($k=50\ \Omega$, $v=c$). Very similar curves can be drawn for the normal connected case as well.

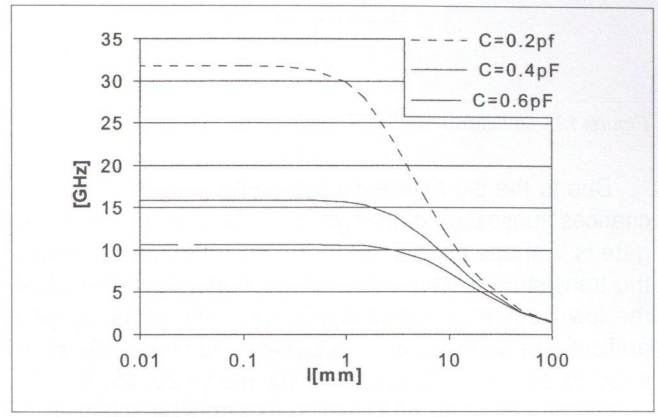


Figure 10. Cut-off frequency vs. l

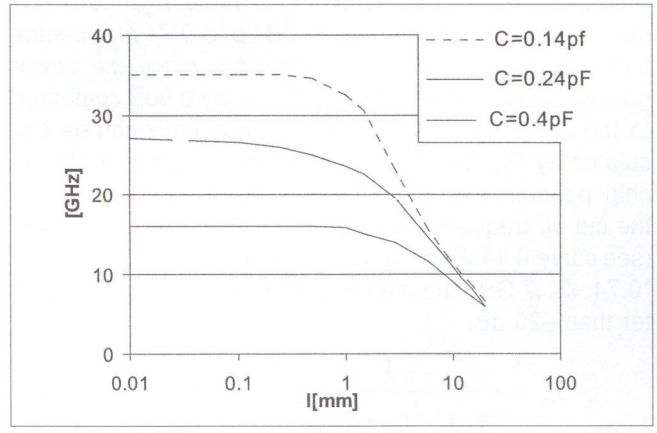


Figure 11.

Hence, with a given active device capacitance and termination impedance (k) value the lumped element DA line has the highest impedance-bandwidth product. As the gain is proportional to the Π-image impedances (see Eq. 5.a, which also valid for distributed element DAs), the lumped element amplifier has the highest GBP or GOBP value as well. Thus, during the design TLs with the minimum required length must be used.

Design Example

The method is demonstrated on a design example, comprises chip MESFETs. The applied unilateral transistor model with the on-chip parasitic inductances is shown in *Figure 12.* It is assumed, that due to technological reasons the lowest value of the connection inductances (f.e. bonding inductance) is $L_B=0.3$ *nH* and the minimum length of the interstage TLs is 0.5mm.

If a minimum gain of 14 dB is required, at least four stage is necessary according to Eq. 5.a (at low freq. the sum terms are equal to N^2), between 50 Ohm impedances.

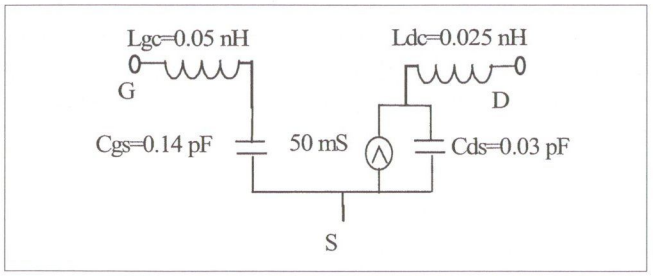


Figure 12. Unilateral MESFET modell

Due to the big difference between the transistor capacitances application of the hybrid structure is practical. The gate is V-shape connected. If the only elements between the transistors were the two connection inductance *($2L_B$)*, the low frequency impedance *(Z_{LF})* of the resulting gate artificial line would be $Z_{LFG}=(2L_B/C)^{0.5}=65\ \Omega$ (Eq. 16). However, to obtain a reflection better than –20 dB, the Z_{LF} should be less than 60 Ω, which can only be achieved by the insertion of TLs with the minimum 0.5mm length and with $Z_0=38\ \Omega$ (Eq. 22). According to Table 1. at 60 Ω *(Z_{LF}* the reflection is better than –20 dB up to *0.74* ω_{rel}. Beside that, due to the higher gate line impedance the power gain will increase by approx. 20% i.e. by 0.8dB compared to the $Z_{LF}=50\ \Omega$ case. The cut-off frequency can be calculated by Eq. 24, due to the significant effect of the on chip inductance on the bandwidth [5]. The degradation of the cut-off frequency due to the TL is approx. 0.708 GHz (see curve 0.14pF in Figure 11). Thus, up to (30.08-0.708) *0.74 =21.7 GHz the matching of the gate line will be better than –20 dB.

$$f_{cG}=\frac{1}{2\pi\sqrt{C_{gs}\left(L_{gc}+2L_B/4\right)}}=30.08GHz \qquad 24.$$

In the drain line the transistors are connected by normal way i.e the connection inductances are in series. If the line were designed for *50 Ω* Z_{LF} the resulting interstage inductance value would be $L_D=0.075\ nH$ and the lumped element cut-off frequency would be (Eq. 25):

$$f_{cD}=\frac{1}{\pi\sqrt{C_{ds}\left(4\left(L_B+L_{dc}\right)+L_D\right)}}=49.56GHz \qquad 25.$$

However, in this case the *q* is 1.69, which causes gain fall at high frequencies according to Eq. 18 if *N=4*. Flat gain can be obtained if *q=1.55*, thus the drain cut-off frequency must be reduced to *45.4 GHz*. To achieve this, a proper TL must replaces the L_D inductance. The necessary length can be read out from the cut-off frequency vs. physical length figure of the normal connected line (*Figure 13.*, curve belongs to *C=0.03 pF and* $L_B=0.3\ nH$). The resulting *l* is *1.45 mm* (in case of $\varepsilon_r=1$, *i.e. v=c*). The required drain TL Z_0 value to maintain a drain line Z_{LF} of *50 Ω* is $Z_{0D}=58$ *Ohm*, according to Eq. 21.

The resulted four stage amplifier structure is shown in *Figure 14.* The simulated results without optimization is shown in *Figure 15.* (Aplac 7.5). They exactly follow the predicted properties (operation frequency, reflection, gain etc.). To a certain extent the effect of the losses and the source inductance can be also compensated by the decrease of *q* (by gain enhancement at high frequencies). The required *q* decrease is fairly low: in case of *Rg=5 Ω* series gate loss and $R_{ds}=250\ \Omega$ parallel drain loss the necessary new value of the *q* is *1.44*. It can be easely achieved by increasing the length of the drain TLs by *0.35mm* and by decreasing the Z_{0D} to *56 Ω*.

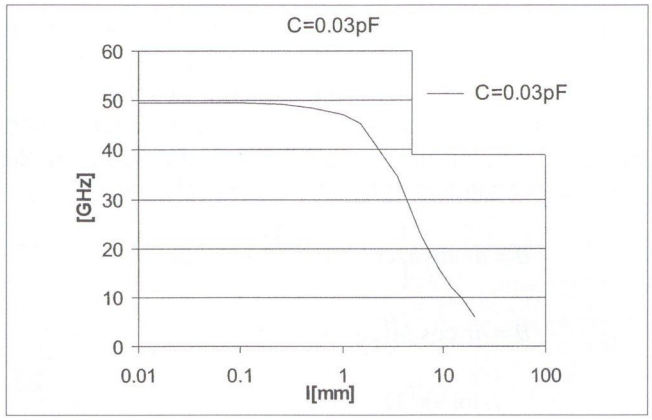


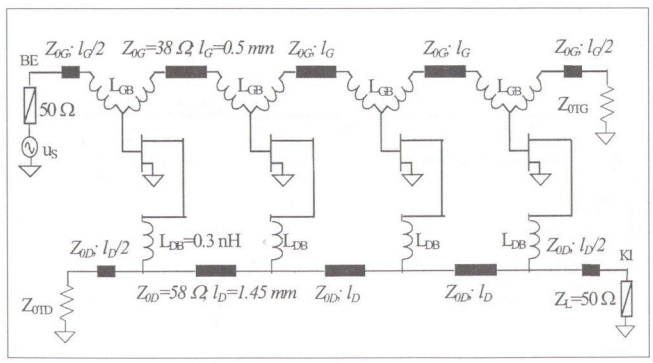
Figure 13.

Only the value of the feedback *(C_{GD})* capacitor is critical. Fortunately, it was observed, that if its value is less than the 40% of the drain capacitance its effect can easily be compensated by computer optimization methods utilizing the resulted structure of the presented method as a starting point.

Summary

Based on theoretical investigation a new systematic design method for distributed amplifiers comprise transmission lines is presented. The resulted amplifier structure has the highest gain-bandwidth product as it minimizes the physical lengths of the applied transmission lines and thus maximizes the cut-off frequency. The impedances of the amplifier lines are adjusted to have low input and output reflection. The so-called hybrid structure is proposed, which utilizes the effect of the parasitic connection induc-

Figure 14. Resulted DA structure



tances in a right manner to adjust the phase mismatch between the input and output lines to a value proper for flat gain in case of the targeted stage number required for the gain specification. The only requirements against the active device are the low value of loss and feedback. The method was demonstrated by a design example and good results were achieved without computer optimization.

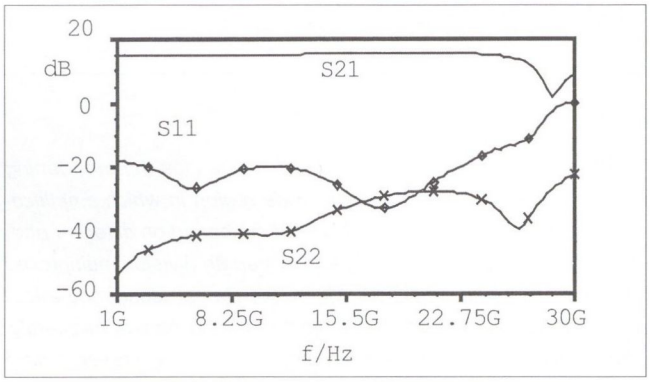


Figure 15. Simulated (not optimized) S-parameters

Acknowledgement

The author wishes to thank Prof. Tibor Berceli and Attila Hilt to the continuos encouragement and help, and Gábor Járó to the useful discussions. He also thanks the financial support of OTKA (No. T-17295, F-024113, T030148, T026557) and E.U. COPERNICUS research projects.

References

[1] Yuhki Imai at al:
"New Distributed Amplifier Design Using Transmission-Gate :FET's",
IEEE Microwave and Guided Wave Letters,
Vol. 6, No. 10, pp.357-359 October 1996.

[2] T. T.Y.Wong:
"Fundamentals of Distributed Amplification",
Artech House, Boston, London, 1993.
ISBN 0-89006-615-9

[3] M.K. Vai:
"Computer-aided design of monolithic MESFET distributed amplifiers",
IEEE Trans. Microwave Theoiy and Tech.,
Vol. 38, pp.345-349,1990.

[4] J.B. Beyer at al:
"MESFET Distributed Amplifier Design Guidelines"
IEEE Trans. Microwave Theory and Tech.,
Vol. 32, pp.268-275, 1984.

[5] A.Zólomy:
"Gain-Bandwidth Performance Comparison of Lumped and Distributed Element Distributed Amplifiers",
MIKON'2000 Conference proceedings,
Vol. 1, pp.101-104, Wroclaw, Poland, May 2000.

[6] A. Zólomy:
"Synthesis Method for Distributed Amplifiers",
MIKON'2002 Conference proceedings CD,
Gdansk, Poland, May 20-22, 2002.

ITU-News

The International Telecommunication Union has launched **Regional Working Parties** on private sector issues to enable ITU Member States, Sector Members, and also potential members to work on regional issues to better cooperate, design, select and implement projects for Information and Communications Technology (ICT) development. To this end, the Telecommunication Development Bureau of the International Telecommunication Union is organizing, together with Senegal's main operator SONATEL and the Ministry of Telecommunication of Senegal, a forum on partnerships with public and private sectors for ICT Development.

The **Radiotelecommunication Assembly** (RA-2003) has spelt out the future direction of the ITU Radiocommunication Sector (ITU-R). The ITU-R plays a vital role in the management of the radio-frequency spectrum and satellite orbits, finite natural resources which are increasingly in demand. The Chairman of the Assembly said „the environment in which ITU operates has been changed drastically in the past 10 years due to the expansion of Internet and wireless communication." He believes this was reflected in the unprecedented number of items that the Assembly was to deal with.

The work of the Study Groups involves developing technical, operational, and procedural bases for efficient use of the radio spectrum and the geostationary-satellite orbit.
The Study Groups are as follows:

- Study Group 1 – Spectrum Management
- Study Group 3 – Radiowave propagation
- Study Group 4 – Fixed-Satellite service
- Study Group 6 – Broadcasting services
- Study Group 7 – Science services
- Study Group 8 – Mobile, radiodetermination, amateur and related satellite services
- Study Group 9 – Fixed service

The work programme approved for the next study period contains some 361 questions with their priority and urgency for completion of studies. It includes studies on matters related to agenda items of World Radio Conferences or requested by WRC resolutions.

Broadband Raman amplifiers in modern telecommunication systems

ZOLTÁN VÁRALLYAY,
Budapest University of Technology and Economics, Department of Atomic Physics, vz423@hszk.bme.hu
GÁBOR VARGA, LÁSZLÓ JAKAB, PÉTER RICHTER
Budapest University of Technology and Economics, Department of Physics

Keywords: Raman effect, Photonic amplifiers, Full optical connections

In order to increase the transmitted amount of data in optical fiber drastically more channels are needed. This causes the broadening of the bandwidth used. Therefore one should apply optical amplifiers which can amplify in a relatively wide region in which amplification is approximately at. Beside the well-known Erbium Doped Fiber Amplifier (EDFA), there exist such amplifiers based on different phenomena and providing better properties in certain areas. They are able to fulfill the requirements of the wavelength division multiplexed (WDM) systems regarding wide amplification region and atness. Among others, Raman amplifiers which use Raman scattering effect are noteworthy since their technological background is solid and their fields of application are the subject of recent papers frequently. We give here an overview about the operation of Raman amplifiers, about their advan-tages and disadvantages and we present here a typical engineering example which occurs often during the design of optical transmission systems and which is analyzed by the model developed by us.

1. Introduction

The phenomena in which an electromagnetic field transfers a given portion of its energy to a lower frequency electromagnetic field was discovered by hindoo physicist Venkata Raman in 1928 [1]. He received the Nobel price for his discovery.

The induced Raman scattering can be used to amplify a signal propagating in a dielectric medium using a higher intensity and frequency electromagnetic field.

The investigations of the possibility of using this principle for amplification in optical fibers begun in the 70s [2, 3, 4]. After the appearance of Erbium Doped Fibers (EDF) which had very good amplification properties and which had the technological background to use it (e.g existence of 980 nm pumping lasers), the investigation of the Raman amplifiers have fallen back. This was also due to the missing of suitable pumping laser with the appropriate intensity and emission frequency.

When the high intensity pumping lasers appeared in the middle of 90s, mostly for EDFA pumping, and they were suitable for cascade Raman amplification as well, the research of Raman amplifiers were reborn again. The possibility of using Raman amplifiers also arose due to the 1480 nm, high intensity (more than 1 W) laser diodes whose maximal amplification region is around 1550 nm (which region is often used in optical systems because of the lack of OH^- absorption). For larger traffic, either the bit rate must be larger in a given channel or the number of channels must be increased. Both case lead to the broadening of the frequency range used.

Therefore, channels will occupy from the recently used conventional band (C-band) to the short-wavelength (S-band) and to the long-wavelength band (L-band) *(see Figure 1)*. EDFAs, are practically not able to amplify under the 1525 nm wavelength considering that the laser transitions of Er^{3+} ions can be found at longer wavelengths. The amplification region of Raman amplifiers depends only on the wavelength of the pump laser. This makes applying more pump lasers with different frequencies (WDM pumping) to broadband amplifier possible even below the amplification range of EDFA.

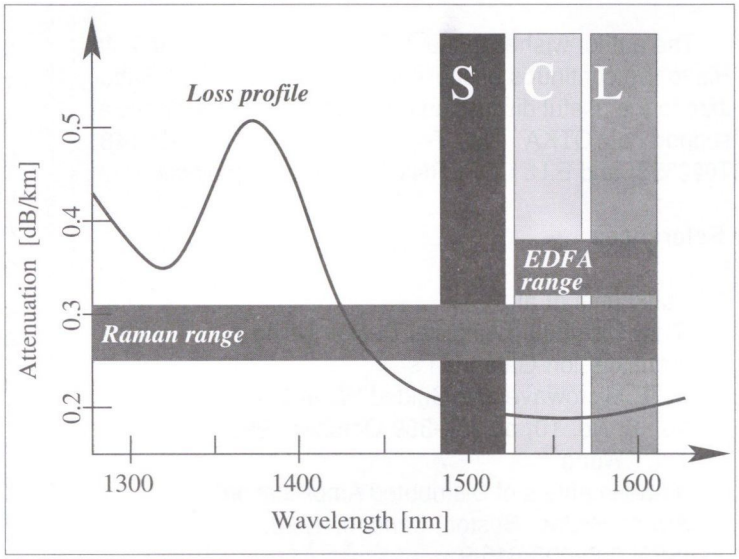


Figure 1. Schematic illustration of loss profile, standardized frequency bands and the amplification ranges of amplifiers. (Considering the Raman amplification range see, for example, [11])

It is important to design and model Raman amplifiers in order to be able to predict the magnitude of the amplification of distinct channels. We give a short overview about the optical amplifiers, the physical background of Raman scattering, the properties of Raman amplifiers and their theoretical and numerical investigations. Finally, we present application of these models in the design of WDM systems.

2. Optical amplifiers

The best known and most frequently used optical amplifiers are EDFAs, Semiconductor Laser Amplifiers (SLA) and Fiber Raman Amplifiers (FRA). Some other rare earth doped fibers also exist.

In the following we overview the requirements related to the using optical amplifiers in WDM systems. EDFA and FRA will be explained separately and finally they will be compared.

2.1. Amplification of WDM systems

The following parameters are considered during the design of WDM systems related to optical amplifiers:

1. noise of amplifiers
2. atness of the amplification
3. dispersion
4. nonlinearity

Every amplifier adds noise to the given system, namely it lowers the Signal-to-Noise Ratio (SNR). In order to characterize this we introduce an amount which compares the ratio of SNR at the input and output of the amplifier:

$$F_n = \frac{(\mathrm{SNR})_{\mathrm{in}}}{(\mathrm{SNR})_{\mathrm{out}}} \qquad (1)$$

This amount is called Noise Figure (NF). An amplifier can be considered better if its values of NF are smaller than other amplifiers' regarding all frequency components.

The second property shows the uniformity of amplification for the different channels belonging to different frequencies. If certain channels are amplified with high gain while the amplification of others are weak this may cause data loss easily. When the intensity of channel decreases under the detectable limit. This has high possibility even at uniform channel spacing because of the presence of crass-talk effect.

An amplifier can be considered good if its amplification profile is relatively at.

The third and fourth parameters are about the dispersion and nonlinearity of the amplifier. We must consider here that a good amplifier should not have significant contribution to the dispersion related and nonlinear effects of the system.

2.2. EDFA

Erbium doped fibers as well as semiconductor lasers work on the basis of population inversion [5, 6]. It is important in this functionality that the Er^{3+} ions can be excited on certain, discrete wavelengths. 980 nm and 1480 nm semiconductor lasers can pump them most effectively. Here, a few times ten mW is enough to get very good amplification (20-30 dB) during a few times ten meters of fiber. Of course, there exist pump laser with different wavelengths in different regions but the efficiency is better around the above mentioned ranges.

The pump light can be sent in forward and backward directions compared to the direction of the signal to be amplified. The rate of the amplification is almost same in both cases but considering the Amplified Spontaneous Emission (ASE) which contributes to the noise, the backward pumping scheme is better.

A very popular arrangement is also when pumps are applied in both directions because the magnitude of inversion level and also the amplification of the signal are mostly homogeneous along the amplifier.

Table 1. Comparisons of EDFA and FRA

	EDFA	FRA
Fiber type	Erbium doped fiber is necessary (mature technology	Same fiber used for telecommunication and amplification
Cost	Expensive	Cheaper, only pump lasers are needed theoretically
Noise	SNR is worst	SRN is better
Pump wawe lengths	980 nm or 1480 nm (for erbium)	Wide range can be used
Amplification range	It can not be used under 1525 nm	Wide range can be used depends on the pump sources only
Width of the amplification range	20 nm (in case of erbium doping)	48 nm and this can be increased using more pumps (see Figure 1 and[11])
Magnitude of amplification	10 dB - 30dB	0-12 dB
Fiber length	Several ten meters	Several kilometers to 100 km

2.3. Raman scattering

Scatterings are divided to two large groups in physics: elastic and non-elastic scattering. In the case of elastic scattering, the frequency of interacting photon with molecules remain unchanged. Typical example is the Rayleigh scattering (see for example [7]).

The Raman scattering is a non-elastic scattering. The incident electromagnetic field loses a part of its energy during the scattering on molecules (Stokes process). This energy difference appears in phonon vib-

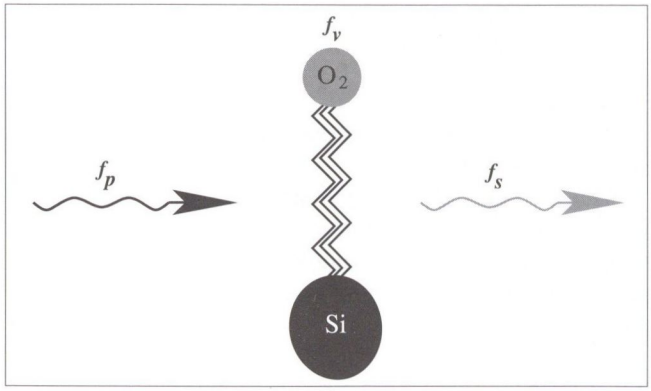


Figure 2. Spontaneous Raman scattering. Frequency down-shifted photon is emitted ($f_p > f_s$) and molecule vibrational states are changed.

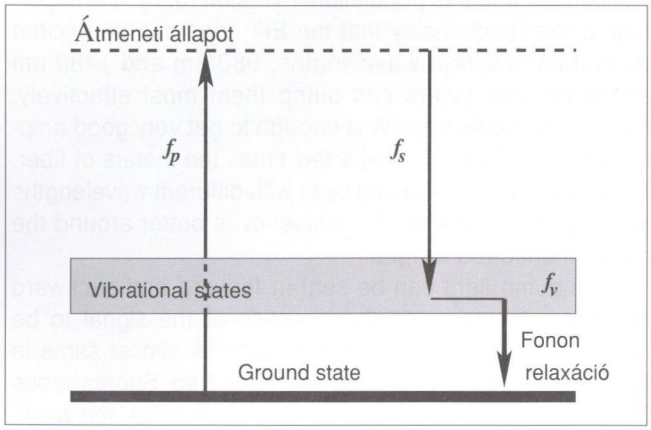


Figure 3. Energy-level diagram of Raman scattering

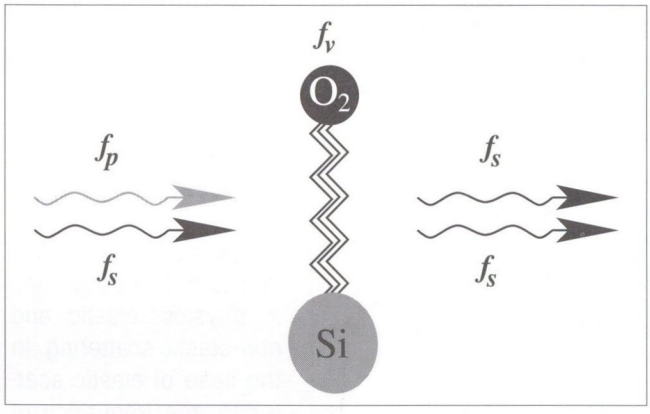


Figure 4. Induced Raman scattering

rations (vibration state of molecules are changed) and a lower frequency photon keeps on *(see Figure 2)*. This phenomena, with the usually used notation in physics, can be visualized by an energy-level diagram in *Figure 3*. The zone of vibrational states are relatively wide here considering that in the case of amorphous materials (such as fused-silica fibers) the vibrational states cover a considerably wide region. **This phenomena makes it possible that a simple silica fiber can be used as a Raman amplifier.**

To do this only a pump source is needed with frequency larger than that of the amplified signal as much as the energy (divided by *h*) taken away by the phonons (molecular vibrations) *(Figure 4)*.

Note that the photon with frequency f_s might also suffer Raman scattering and a frequency shifted photon would be scattered again. However, the occurrence of Raman scattering has a threshold, namely it becomes significant only above a certain intensity. We must note that however, in Dense WDM (DWDM) systems, in case of lots of channels the density of energy could be so high that this effect is not negligible furthermore and can cause more problems.

2.4. Raman amplifiers

The Raman amplifiers as well as EDFAs can be pumped in three different ways forward, backward and bi-directional. However the wavelength of pump laser can vary in a large range and it does not have to have a discrete value in frequency or wavelength.

The basis of Raman amplification is to find one or more pump sources which are so far from the signal which must be amplified that they may get the possible largest amplification during the process. To do this we should know the Raman gain profile as a function of frequency shift for the material used for the amplification. We can see this type of Raman profile in *Figure 5* which is a measurement result of a single mode optical fiber with a core produced from SiO_2 and GeO_2.

We can observe in Figure 5 that the larger the frequency shift is (up to 13.5 THz) the larger gain can be achieved.

For example we consider the case of a signal with 193.4 THz which is about 1550 nm in wavelength, this is the center of the C-band. Then we should choose a pump source which has frequency with about 13.5 THz higher. This is the 206.9 THz frequency light (1490 nm in wavelength).

In case of broadband, multi channel systems, this task is not so easy because different channels are amplified and attenuated with different magnitudes if we apply one pump. Therefore, high intensity differences may appear between different channels after some kilometers of propagation. To avoid this problem more than one pump is necessary. The intensity and frequency distribution of them is a subject of optimization of Raman amplifier design which is discussed bellow.

2.5. Advantages, disadvantages

Table 1 shows some comparisons of EDFA and FRA.

Note that EDFA can not amplify for wavelengths shorter than 1525 nm but there exist some rare-earth doped fiber amplifiers which can be applied in the S-band. This fact also shows that the features of rare-earth doped fibers are dependent on the material used but, in case of Raman amplifiers, the amplified regions depends on the properties of the pump light not on the waveguide medium which means a relatively large freedom in designing amplifiers.

Of course, Raman amplifiers have some disadvantages as well:

- Many pump lasers are necessary (in some cases these may be 8-12) in order to get relatively good atness over a 100 nm wide range [8, 9]
- Compared to EDFA, relatively high pump intensity should be used in order to get considerable amplification (more hundreds mW to 1-2 W)
- Crass-talk can appear among signal and pump channels under certain circumstances.

Note that, recently, Raman amplifiers are used for compensating the losses in telecom-munication systems and to do this some few times ten milliwatt intensity pumps are enough in a few times ten kilometer fibers. Thus, Raman amplifiers are usually combined with other type of amplifiers (mostly EDFA) or they are used as pre-amplifiers.

However, nowadays, more than 10 dB amplification could be reached with Raman am-plifiers in a few times ten kilometer fibers which result shows that using high intensity laser diodes make it possible to use FRAs separately as complete amplifiers in optical systems.

3. Numerical simulation

We solve the equation connected to the light propagation during these simulations which is usually called Nonlinear Schrödinger Equation (NLSE), because of the form of it, or nonlinear Schrödinger type equations. These equations, with some approximations which work well for single-mode optical fibers, can be derived from the Maxwell equations [10]. One of the most simple form of these propagation equations have a form of:

$$\frac{\partial A}{\partial z}+\beta_1\frac{\partial A}{\partial t}+\frac{i\beta_2}{2}\frac{\partial^2 A}{\partial t^2}+\frac{\alpha}{2}A=i\gamma|A|^2A, \qquad (2)$$

where z is the coordinate of the distance which is same as the optical axis, t is the time coordinate, $A = A(z,t)$ is the envelope of the electromagnetic field which is also called, after the applied approximation in the derivations, slowly varying envelope function. β_1 is the inverse of the group-velocity, β_2 is the so-called group velocity dispersion, α is the loss which has a minimum around 1550 nm wavelength, as it can be seen in Figure 1, for fused silica fibers (about 0.2 dBm/km). γ is the so-called nonlinear coefficient. Its value depends on the nonlinear refractive index, on the frequency of the propagating light and on the effective core area at the given frequency (more details can be found in [10]). In this case, the term at the right-hand side of Eq. (2) is called self-phase modulation.

Raman scattering is a nonlinear effect and it can be derived using some considerations connected to the third order optical susceptibility tensor. If we do this some new nonlinear terms will appear in Eq. (2). And this become more complicated if we have more propagating channels and more propagating pumps.

The equations which have similar forms to Eq. (2) can be treated numerically with the so-called Split-Step Fourier (SSF) method which is sometimes also called beam propagation method. The essence of this method is that the above equation can be formally written in the next form

$$\frac{\partial A}{\partial z}=(\hat{L}+\hat{N})\mathrm{A} \qquad (3)$$

where $\hat{L}$ contains the linear coefficients of A and $\hat{N}$ the nonlinear ones. The solution of this is given as

$$A(z+\Delta z,t)=\exp\left[(\hat{L}+\hat{N})\Delta z\right]A(z,t) \qquad (4)$$

where Δz is the step-size. Namely, the solution of the problem can be traced back to a multiplication with an exponential operator. The linear part of it is calculated in the Fourier domain and the nonlinear one in the temporal space and the contact is provided by the Fourier-transform between them.

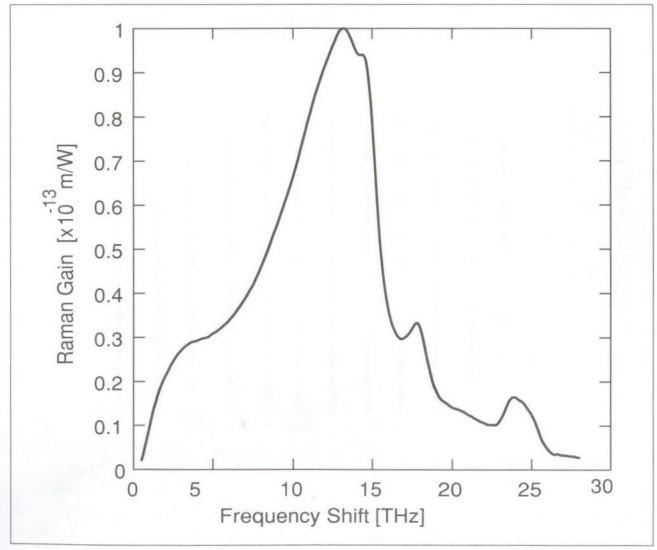
The coupled system of equations describing Raman amplification (see the details in [10] Chapter 8) can not be solved so easily because they can not be rewritten in a form of expo-nential operators. We use here a mixed method in that the linear contribution is calculated with the split-step method but the nonlinear one with a fourth order Runge-Kutta (RK4) method.

Using the measured Raman amplification profile and a loss profile for a given fiber, we performed such calculations using the above explained procedure which made to simulate a WDM system possible with more pumps, arranging them in a way to obtain an optimal behavior considering the amplification of the different channels. Thus, numerical calcula-tions can reach their aim to design Raman amplifiers in a WDM system without further measurements.

4. Computational design

Let there be a twelve channel system in that the channel spacings are large (1 THz) and the center frequency of this propagating WDM signal let be 195 THz. This means we have frequencies 189.5, 190.5, ..., 199.5, 200.5 THz. This large channel spacing was chosen to demonstrate Raman amplification related effects more spectacularly.

Figure 5. Measured Raman amplification as a function of frequency shift



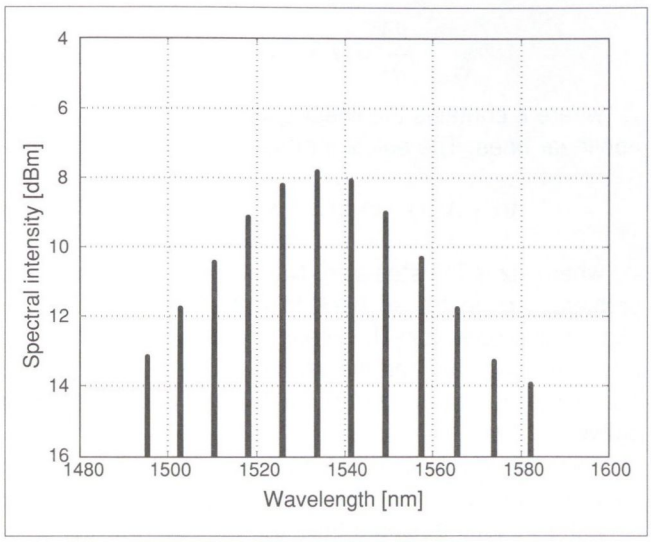


Figure 6. Spectrum of twelve channels with large channel spacings (1 THz) after 50 km propa-gation using 400 mW backward pump

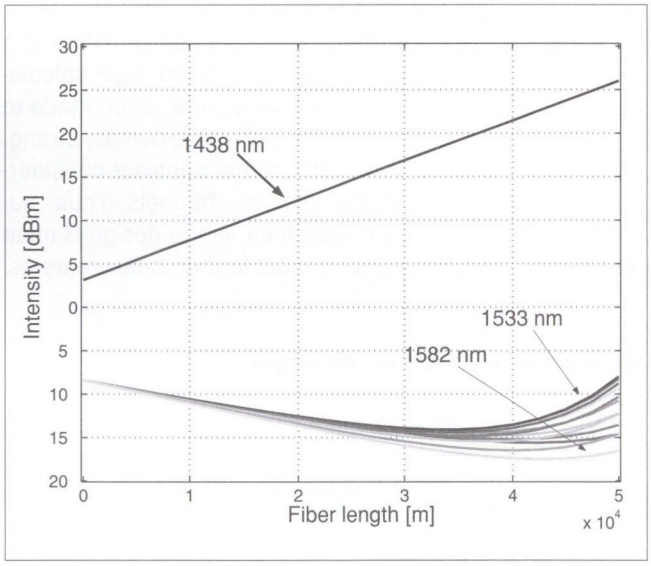


Figure 7. Intensity changes of different frequency components as a function of length

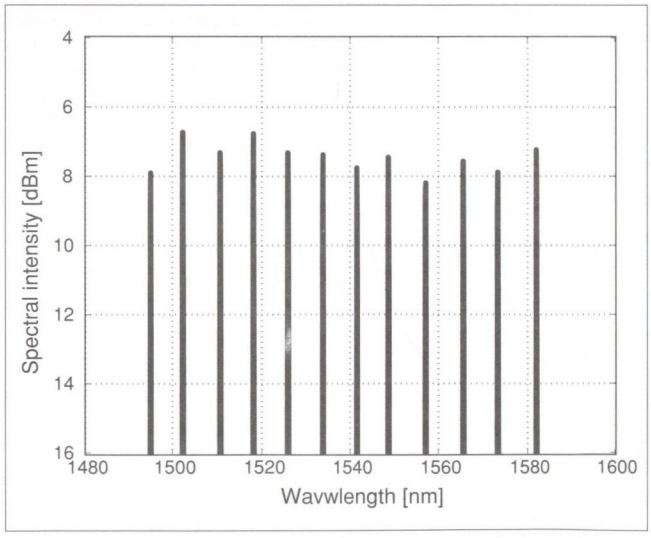


Figure 8. Flattening the spectrum using seven pumps with different intensities and frequencies

If we use only one pump, for instance, at 208.5 THz, the result can be seen in *Figure 6.* The original signal was multiplexed from Gaussian pulse series with the above mentioned frequencies. Each channel had originally a -8.4 dBm intensity (this is the spectral intensity in dBm of a 1 mW Gaussian beam) and the pump was 400 mW (26 dBm).

We can see in this figure that the amplifications of different channels are far from each other and the profile of the resulted spectrum intensities are more or less similar to the Raman amplification profile. (In these calculation, we used an approximated Raman profile with a Lorentzian function instead of the measured one.) Note that the attenuation of different channels were also different. We used a loss profile measured for single-mode fused silica fibers.

We can see also this phenomena in *Figure 7.* where the intensity changes were plotted as a function of length. The intensity differences at the end of the fiber are relatively high similar to the spectrum. The strait line that starts from the end of the fiber (from 26 dBm), is the backward pump.

To compensate these large differences, visualized in Figure 6 and 7, we must use more pumps with different frequencies. One difficulty to do this that some of them will be placed in frequency where the loss is relatively high (about 1350-1450 nm, see for loss profile Figure 1). Therefore we use bidirectional pumping scheme, with that we reduce the possibility of the decreasing of some channels still at the beginning of the fiber.

We found the following arrangement the most suitable after some optimization: 3 forward direction pump applied with 100, 120 and 140 mW intensities, on 1425, 1415 and 1405 nm wavelengths, and 4 backward pumps used with 60, 65, 80 and 535 mW intensities, on 1490,1460, 1420 and 1400 nm wavelengths. Results can be seen in *Figure 8.* where the observed high intensity variations disappeared. Largest difference is about 1 dBm which shows that the amplification profile is relatively at for the modelled 100 nm band.

The obtained intensities have almost the same magnitude as at the beginning of the fiber, therefore this arrangement is suitable for compensating the loss of the fiber without remarkable changes in intensity differences between channels.

About 120-240 different channels can be modelled within the wavelength range with increasing density of channels (0.05-0.1 THz channel spacing). This simulation can be performed according to the above described method.

5. Conclusion

This article was intended to present the theoretical and practical background of Raman amplifiers including their physics and designing methods. Some practically important examples were presented using numerical simulations.

These simulations are about to optimize the attening of Raman amplification using different intensity and frequency pump sources (WDM pump). We also compared the properties of Raman amplification to other type of optical amplifiers, we discussed the advantages and disadvantages of them. We described and presented here how we can use an optical amplifier based on the Raman scattering in telecommunication systems and how numerical simulation can help to do this.

References

[1] C. V. Raman,
A new radiation. Indian J. Phys., 2:387, 1928.

[2] R. H. Stolen, E. P. Ippen, and A. R. Tynes,
Raman oscillation in glass optical waveguide.
Appl. Phys. Lett., 20(2):62–64, January 1972.

[3] J. Stone, Cw raman fiber amplifier.
Appl. Phys. Lett., 26(4):163–165, February 1975.

[4] C. Lin, L. G. Cohen, R. H. Stolen, G. W. Tasker, and W. G. French,
Near-infrared sources in the 1-1.3 μm region by efficient stimulated raman emission in glass fibers.
Opt. Commun., 20(3):426–428, March 1977.

[5] C. R. Giles and E. Desurvire,
Modeling erbium-doped fiber amplifiers.
J. Lightwave Technol., 9(2):271–283, Feb 1991.

[6] A. Bononi and L. A. Rusch,
Doped-fiber amplifier dynamics: A system perspective.
J. Lightwave Technol., 16(5):945–956, May 1998.

[7] B. E. A. Saleh and M. C. Teich,
Fundamentals of Photonics.
John Wiley and Sons, Inc., 1991.

[8] Y. Emori and S. Namiki,
Broadband raman amplifier for wdm.
IEICE Trans. Electron, E84-C(5):593–597, May 2001.

[9] Y. Emori and S. Namiki,
Broadband raman amplifier for wdm.
IEICE Trans. Comm., E84-B(5):1219–1223, May 2001.

[10] G. P. Agrawal,
Nonlinear Fiber Optics. Academic Press, Inc.,
San Diego, London, 3rd edition, 2001.

[11] IPF Technology. Tunable raman lasers.
http://www.ipfibre.co.uk/Products/Raman Laser/Tunable/tunable frl.htm

ITU-News

New European standards Guidelines Agreed

European standardization policy has been brought up to date as the result of the adoption of new guidelines between the European Comission (EC), the European Free Trade Association (EFTA), and the three official European Standards Organizations (ESOs),CEN, CENELEC and ETSI. These new guidelines were signed in Nice, France, March 2003 by Erkki Liikanen, the EC Commissioner responsible for Enterprise and Information Society, EFTA Secretary-General William Rossier, and senior representatives of the three standards organizations, during the conference on „Accessibility for all", organized by the ESOs.

IEEE Virtual Museum Celebrates one-year anniversary

The IEEE Virtual Museum was conceived of as a site that would enhance the public's understanding of the technologies that underpin modern society and which would place those technologies into social and humanistic contexts. In the past year we have made great inroads into achieving this goal. Currently the IEEE VM contains about 300 unique pages, which are organized into four exhibits. Upon the first week of launch in February 2002 we had several hundred Web „sessions". In the last months of 2002 we've been averaging about 15,000 sessions per week. As we move into our second year of activity we hope to continue growing the IEEE VM despite a hostile economic climate. This expansion includes adding more features and more materials for use by educators.

We wish a merry Christmas
and a happy New Year for every reader

Editorial Board

Auralisation as a technical tool

Relation between measured noise values and acoustic sensation

ANDRÁS ILLÉNYI

Budapest University of Technology and Economics, György Békésy Acoustic Research Laboratory
illenyi@alpha.ttt.bme.hu

Keywords: Noise effects, Subjectiv evaluation, Correlation methods

Noise of electrical signals and other unwanted, annoying or unpleasant sound signals are widely known types of disturbing noise. From point of view of physical definition and measuring technology sound and noise are identical quantities. Its the momentary human judgement that makes difference between them which depends on the actual expectations of the audience. This article presents a review of measuring techniques of acoustical noises.

We will point out that the dBA measure used in measuring practice generally does not follow the results of sensory acoustical judgement. The advanced approach of quality has brought about an interesting *turn-round* in the decade long hang of *acoustical measuring technology*. This states that the operation of a product shall be accompanied by the expected noise, or more precisely, the expected noise and/or vibration spectrum. This new approach allocates different noise and vibration parameters to each product. All this can be done with using the objective methods which take into account features of binaural hearing (auralisation).

Sensory acoustics allocates realistic indexes to acoustical stimuli that can be assessed objectively. We will introduce some important psycho-acoustical sensory measures and will refer also to the way of working with sensory measures and yet objective methods.

Introduction

Under noise is meant the unwanted signal disturbing the study or perception of the useful signal or disturbing the percipient. Disturbing noises exist with all sorts of signal. The most important occurrence of them is with electrical signals and sound signals. In this latter case – according to definition – any unwanted, disturbing or annoying sound is considered as noise (noise pollution). It is well known that for physical definition and measuring technique sound and noise are equal concepts. It is the instantaneous human judgement that makes difference between them. The same sound (e.g. a famous music or crying of a baby) can be agreeable or carry information for one person while the source of an unpleasant noise for another.

This means that in the acoustics – in given circumstances – difference between sound and noise can only be made on the basis of human expectation. Measuring techniques often make this judgement based on statistical average values since some people are more sensitive, others are less sensitive for the general judgement of the noise. The acceptability or the permitted level of noise is specified in international standards and measurement guidelines.

This article is dealing primarily with the measurement of acoustic noise but ordinary measuring techniques are also reviewed. It will be pointed out that the measuring practice which is still using some measures (e.g. dBA) "inherited" from the measuring technique of the 20th century generally does not follow results of sensory acoustics. This would lead to contradictions which cannot be justified by keeping previous measuring techniques.

Change in approach due to sound quality

The demand for and approach to quality typical for several aspects of life has brought about an interesting turnround in the acoustic measuring techniques as well. At the same time there are more and more signs that measures obtained by traditional measuring techniques but using advanced digital measuring technology and data processing are in contradiction with the quality sensed by the ear of user (recipient). Nevertheless it is just this sound quality that is users expect! Now, if sensation as fact contradicts measurement results then we can be sure that the problem is not with facts! More particularly: our measured values may not indicate what we expect from them.

The so-called sensory diagnostics i.e. measurement and interpretation of quantities proportional to subjective sensing has been playing more and more important role since the middle of the 20th century. In the last decade of the past century the concept of "acoustical quality" became official which raised the demand for a brand new measuring technique. Today this is an independent profession allowing for the specification, control and planned technological application of human expectation for quality, in connection with quality management systems and based on standard series ISO 45000 and ISO 17025. As a consequence, the task of an acoustic professional is broader than ever. Formerly an acoustic professional "just" measured and assessed noise and vibration signals and then presented proposals or plans regarding their attenuation. In the product design phase he or she was responsible for keeping noise and vibration levels below limits.

The new approach is quite different.

In order to ensure quality, acoustic developing engineers should realize specified noise and vibration spectrum within given limits for the product being designed. This is a complex task since the objective is not just to remain below a certain limit value. The product shall operate with the specified noise level as well as with the expected noise and vibration spectrum! This means that the new and advanced task allocates different noise and vibration characteristics to each product. This may sound pointless but we should bear in mind that a customer today expects given noise and vibration features from a given product. A saloon-car is expected to be silent but have a "sound of car" to prevent by-passers from stepping in front of it. On the other hand, a sports car should be "sport-like". A vacuum cleaner must function silently but its sound shall persuade the buyer of a suitable suck!

The new task is thus to design and manufacture for a sensory-expected noise and vibration. This is more and more frequently requested and realized. Let us remember just two examples: the designed and realized sound of a closing car door and the pleasant sound of electrical switches. The sound of a product shall be specified after hearing and the approved construction of the expected sound. This is performed with objective methods based on binaural hearing, using the so-called auralisation. The concept of "auralisation" is more and more broadly used in acoustical technology. It covers all processes and techniques which use objective simulation and computer technology to create models offering the expected sound experience of the system as it is modeled in virtual reality [10]. Auralisation means also the objective data processing based on binaural techniques and/or sensory acoustics [11].

In this article we use the widely accepted scientific determination stating that sensory acoustics answers with realistic and objective measures to acoustic stimuli. Hereinafter we review the most important psycho-acoustic sensory measures and show how they can be used for objective methods and how products can be constructed for the expected noise experience. The basis for these processes is the so-called binaural measuring technique.

Measures of sound sensation

In course usual acoustic problems a measuring microphone is placed in the sound space then its signal is averaged in time and weighed in frequency and finally examined in level values such as LA. This measure is in relation with the sensing of our ear, with the possible sensing impairment of our ear and leads to a simple, univalent index. In an indirect way it refers also to the loudness of sound and disturbing noise and to its noisiness as determined through subjective sensing. Despite all these features there is a growing demand for having also objective measures for the analysis of sound events [1, 2].

Studying human hearing as complex physical and sensing process we can state the analysis of sound waves requires the measurement of more than one physical and sensory parameter. It is worth taking account of the level, duration, spectrum, construction in time and space, information content, sensory effects of sound signals and their quantitative indexes. This suggests that one parameter cannot be enough for the successful description of a sound. The human ear is an organ with broad dynamic range, high sensitivity, long and short term memory able to compare sound events. It is able not only to compare sound levels but also to sensitive determination of changes as well as to describe quality in terms of so-called sensory judgement-pairs.

Some typical sensory judgement-pairs:

annoying ⇔ enjoyable	flat ⇔ rough
loud ⇔ quiet	sharp ⇔ soft
suitable ⇔ unsuitable	dull ⇔ clear
exciting ⇔ calm	noisy ⇔ not noisy

These parameters are not interchangeable and cannot substitute each other. On the other hand, they are important parameters and as such have to be determined with psycho-acoustic measurement. This measurement task usually involves the following considerations:

- Usually applied simple physical sound signal parameters, e.g. A-sound pressure level or analysis carried out in 1/1 and 1/3 octave bands do not provide comprehensive information on sound events.
- An all-round analysis requires sensory parameters that can be determined by subjective observation and also further parameters that can be processed with objective methods, such as loudness, sharpness, roughness, pre- and post-masking and simultaneous masking.
- Under normal circumstances human hearing organ is a binaural device as opposed to the monaural system of usual measuring techniques. We have to add: the human ear has two measuring channels which can be used for selective and interactive, comparative analysis at the same time. Furthermore, our hearing organ takes sound signal samples from a space (so-called perturbed sound field) which is rather complex owing to the disturbing effect of the head and the body thereunder. This is rather strange as opposed to usual measuring techniques yet it offers very good noise suppression, signal processing and high spatial selectivity!

During the past decade many important publications [3] appeared in the field of acoustics suggesting that the new, quality based approach to sound analysis cannot disregard sound signal diagnosis in connection with sound signal parameters.

Limitations of traditional acoustic measurements

It is well known that with most acoustic measurements A-sound pressure levels do not provide enough information on whether these levels will have acceptable, unpleasant or even pleasant effect on recipients. For instance, the analysis of diesel engine noise, pink noise and square pulses of the same LA level and similar 1/3 octave spec-

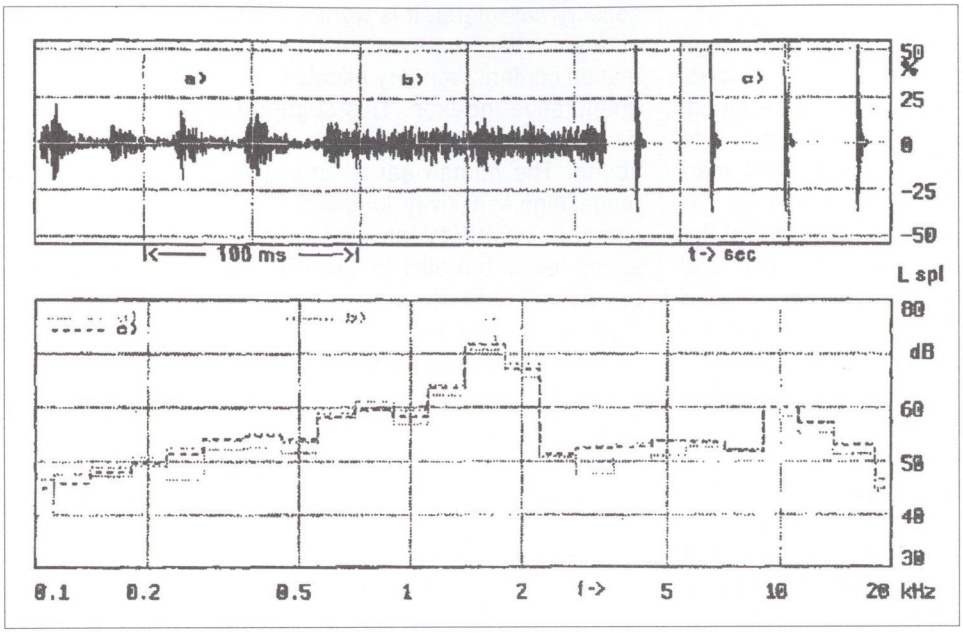


Figure 1. Three different sound signal: dies engine noise (a), pink noise (b) and series of square pulse (c) and their 1/3 octave spectrum producing nearly identical LA levels [7]

trum shows that despite identical physical measurement results the sensory opinion is different in terms of loudness and sound quality of the three noises *(Figure 1)*.

This makes it clear that determination alone of LA level is not enough, no matter if it is averaged or the result of analysis of narrow-band samples. In order to obtain the correlation between measurements and sensory effects, the changes of signal in time and its spectral transients shall also be taken into account. With the same measured LA levels the values of loudness can differ even ten times [3]. Other studies showed that with the same LA levels the loudness value can be even 150% lower if sound signal is 1/3 octave band wide and does not cover the whole sound frequency band. In case of identical sound intensity the sensed loudness depends on the spectrum distribution of the sounds under study.

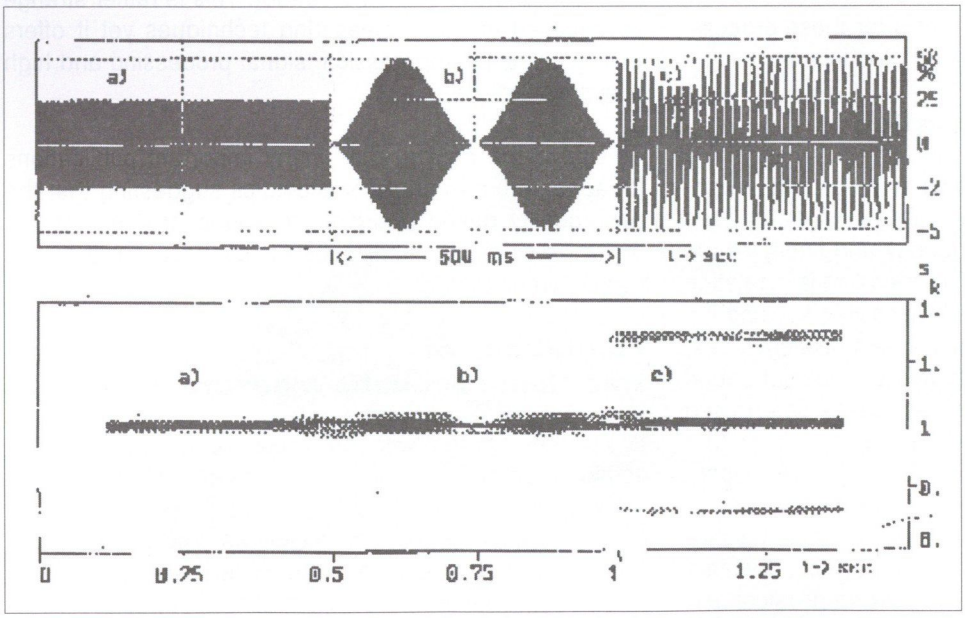
It is also well known that modulation has influence on temporal and spatial changes of the sound signal while average A-sound pressure levels (AL) are the same for the aforementioned sounds. The physically modulated sound is sensed as inequality and this phenomenon is called roughness of the sound. However, when a 4 kHz sinusoidal signal is amplitude modulated with a 4 Hz or 70 Hz signal we find that 4 Hz modulation gives slightly unpleasant feeling whereas the 70 Hz modulation produces a hard, rough sense. Lower part of *Figure 2* also side bands of modulation can be observed.

It is also possible to describe the sensory measures mentioned above in closed mathematical forms. *Loudness,* the psycho-acoustical equivalent of intensity is calculated and measured according to the Zwicker method [1,2,6]. In calculations the loudness of G critical band is expressed in sones as follows:

NG = 20.1(LG – 40) if **LN > 40 phon,** or
LN = 40 + 33.22lgN if **N > 1 sone.**

Here LG is the level value belonging to the bark bandwidth of G frequency group, e.g. critical bands and LN is the phon value belonging to N loudness expressed in sone.

Figure 2. Representations of a 4 kHz sinusoidal signal (a) and its AM versions modulated with 4 Hz (b) and 70 Hz signal in so-called roughness units [7]



ISO standard 532B specifies methods for the calculation of sone loudness of broadband noises in case of frontal free field sone incidence (F) and equal loudness in diffuse field (D) [5, 6]. Several examples are known where measured noises were identical in dBA level values but there were considerable differences between their loudness *(Figure 3)*.

The objective of psycho-acoustical engineering is the quantitative description of hearing events occurring as a result of physically well defined stimuli and their application in the engineering activity [11]. The sound corresponding to the expected sound quality is

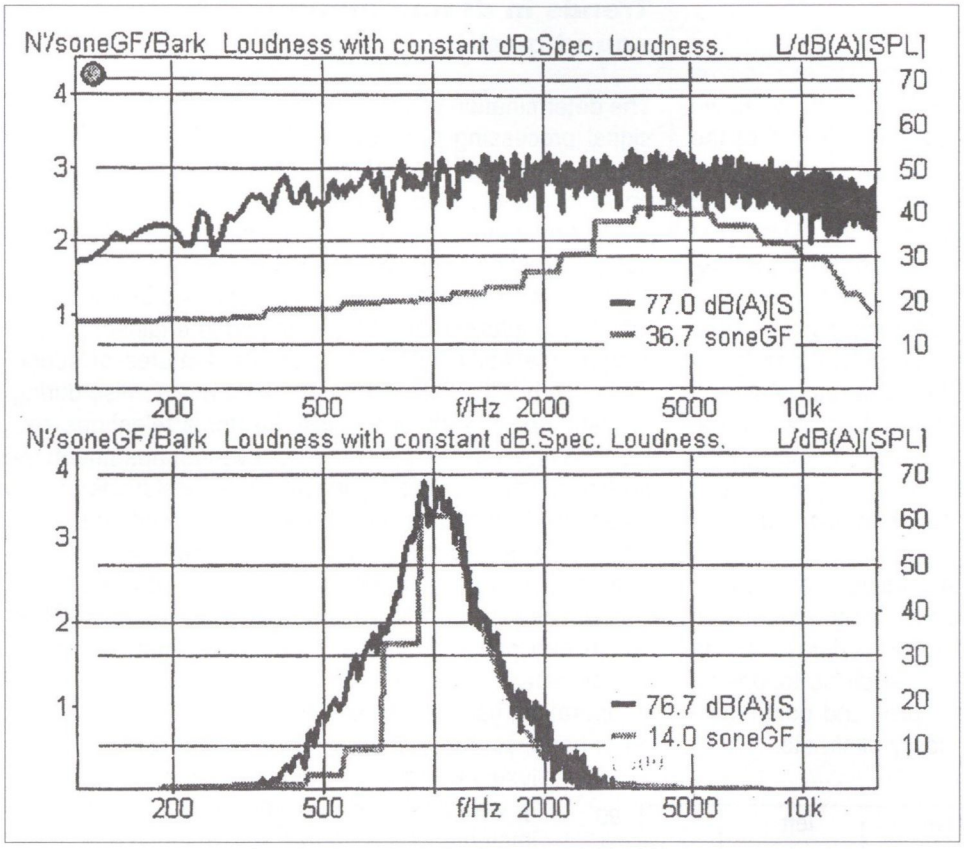


Figure 3. Two nearly identical LA ~70dBA noise levels and their loudness values expressed in sones [7]

synthesized and then compared with the recorded sound or the sound of the product under test. Physically this "sound quality answer" depends typically on the success of the so-called free field equalization. The realization, however, is not easy since equivalents of sound field quantities in the hearing range are frequency and intensity dependent. The relation is of nonlinear character on both scales. In case of white noise masking the masking is dependent on sound intensity and frequency; if f > 500 Hz the slope of masking threshold is -10 dB/octave.

Masking curves 160 Hz narrow band and 1 kHz center band frequency noise are symmetrical to the intermediate frequency with low intensities. Increasing intensity leads to a growing asymmetry if masking noise level > 40 dB (nonlinear extension of masking).

At low frequencies the slope of the rising section reaches 100 dB/octave, this means that realization of sensory effect requires extreme filters! Critical band is a fundamental feature of psycho-acoustics. Its technical equivalent is a series of filters with bandwidth of 100 Hz if f < 500 Hz, for higher frequencies ≅ 0,2f. It can be mentioned for comparison that the constant bandwidth of 1/3 octave filters is 23%, i.e. $0{,}23f_0$ and generally $f = f_0$ the center band frequency.

Measured values of loudness have important role in the determination of further psycho-acoustic parameters. Some of the will be summarized below.

Unbiased annoyance [4], expressed in **au** units:

$$UBA = d(N_{10})^e\,(1+S+f) \qquad [au] \qquad (1)$$

where **d** is a factor depending on part of the day, **N10** is the loudness value in 10% of exceeded measuring time in sone, **s** is a sharpness-specific factor and **0** is a tonal component specific factor which can be taken from tables.

$$d = 1+\left(\frac{N_{10}}{5}\right)^{0.5}, \qquad (2,3)$$

$$S = 1+0.25(N_{10}-1)\cdot\lg(N_{10}+10)$$

Value of **S** *(sharpness)* can be determined by the rate between loudness and the loudness of frequencies affecting loudness. According to other views it depends on rising time of impulsive sounds and the duration of pulses which is therefore characterized by the fluctuation of duration rates and expressed in **vacil** (from word *"vacillation"*).

F is the quantity of sound fluctuation, a sensory quantity which is specific to noise fluctuation. Thus the above simplified expression of UBA can be extended as follows (4):

$$UBA = d(N_{10})^{1.3}\cdot\left\{1+0.25(N_{10}-1)\cdot\lg(N_{10}+10)+0.3F\cdot\frac{1+N_{10}}{0.3+N_{10}}\right\}[au]$$

Here **S** can be expressed with the relation of the sum of partial loudness of critical bands to the complete loudness, where g(z) is the weight function interpreted on the z frequency axis. This can be derived from the roughness of the narrow band signal with constant loudness but the method of calculation goes beyond the scope of this article.

$$S = 0.11\cdot\frac{\int_0^{24\,Bark} N'\cdot z\cdot g(z)\,dz}{N}\,[acum] \qquad (5)$$

$$F = \frac{0.36\int_0^{24\,Bark}\lg\left(\frac{N'_{max}}{N'_{min}}\right)dz}{\frac{T}{0.25[s]}+\frac{0.25[s]}{T}}\,[vacil] \qquad (6)$$

Finally let us remember that N'max and N'min are the specific loudness calculated from the rate of maximum and minimum loudness values (with condition of N'max/N'min <5) and T is the time duration between the two mentioned loudness values measured in seconds.

In practice long term average loudness of fluctuating noise is strongly affected by smaller noise events occurring between intermediate parts. In this case the sensed loudness therefore is higher than averaged loudness. Especially so-called tonal components can affect sound sensation.

Role of the auricle

In addition to knowledge gained in the field of complex and sensory sound analysis, during the past decades we discovered many new aspects regarding the role of the auricles as well. Parts of the head, the body and the auricle being in the way of sound propagation affect sound signal depending on frequency and direction. This takes place in the +20 dB...–30 dB range when the signal obtained in this way is compared to the signal of a measuring microphone placed in the hypothetical center of the head. In case of frontal sound incidence the change in white noise level is ±8...10 dB in the same ear when the head is turned in the horizontal plane. This means that binaural sound analyzing technique is more complex than the usual acoustic measurements, on the other hand it allows for more opportunity for discovering information on the sound signal under test.

In practice this technique is successfully applied for source identification, spatial localization, and with the use of suitable digital signal processing procedures, for the objective analysis of sensory effects such as loudness, annoyance, roughness, sharpness, pre- and post-masking and simultaneous masking, sensory pitch, etc.

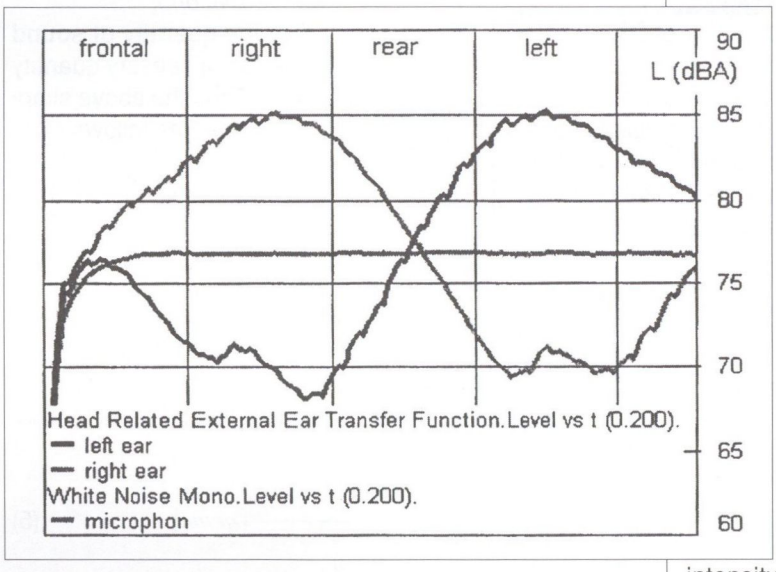


Figure 4. Directional transfer function of white noise in one point of the sound field with single channel microphone (central line). Directional signal specific transfer functions as measured by left hand (strong line) and right hand (light line) microphone of an artificial ear with frontal, rear, right/left incidences [7].

Combining advanced techniques of digital signal processing with similarly advanced procedures of subjective hearing we have already suitable means for the realization of complex tasks of sound quality. Even the examination of sound signals segmented to 2 ms can provide analysis about the relation of complete loudness with complete sound pressure level or with a part thereof and the resulting analysis correlates well with corresponding sensory parameters. These techniques make good use of the physical and signal processing opportunities coming from the perturbation of aural analysis techniques.

Trends in determination of sound quality

The determination of sound quality is based on the known signal processing procedures of human hearing. In this model signal analysis, stored sound information and psycho-acoustic knowledge have important role.

Just as formerly in sound recording then in media technology practice, the sound engineering responsible for quality has to realize acoustic expectations of the customer. In the information society the sound engineer is thus responsible not only for the acoustic features of sound recording but also for any product that emits noise during operation or has affect thereon, i.e. for a telephone set, labor-saving machines, vehicles or even an apartment! Interestingly, the realization of acoustic expectations contributes also to the protection of the environment. At the same time new acoustic analysis and assessment techniques offer more opportunity for the sound engineer.

This new approach reveals quite different aspects of sound and noise. Sound and noise, its inevitable companion means not only sound parameters but also quality, operation, risk and the pollution of environment. On the other hand, sound and noise have become the symbol of luxury design, sport-like product, the acoustically well designed and still environment-friendly implementation. In this sense sound and noise ranks the product from several points of view (acoustics, psycho-acoustics, construction and mental).

Technological, constructional efforts being made for the sake of customers, environment and quality, together with technical development are much more than abstract philosophical considerations. An advanced acoustic approach was born called psycho-acoustics engineering which defines selection, creation and implementation of the expected sound spectrum as construction objectives. Scope of psycho-acoustics engineering is not limited to the study of absolute physical or psycho-acoustical parameters (loudness, sharpness, roughness, intensity fluctuation, etc.). Analysis of spectrum construction in terms of quality is also getting afore which can be realized by the examination of short (a few milliseconds) spectrum [12].

According to the experiences of HEAD Acoustics GmbH (Germany) quality (Q) can be specified by taking account of loudness (N), sharpness (S) and long term (2-4 s) critical band level values FG(i) as well as short term (2 ms) level values FG(i, n) as follows [7]:

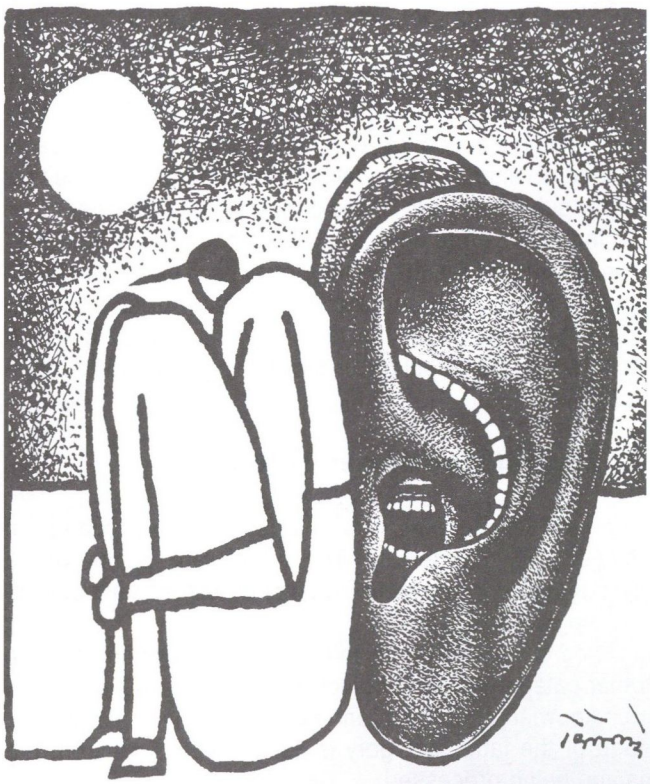
$$Q = \Phi(N,S) + \Phi(\Sigma i = 1K\ 24\{|FG(i-1) - FG(i)| \cdot w1(i,FG(i))\} + \\ + \Sigma n = 1K\ T\ \{|FG(i,n) - FG(i,n+1)| \cdot w2(i,FG(i))\} \qquad (7)$$

where Φ symbolizes a functional relation while $0<n<T$ for all n = 2 ms partial times, and i is one of i = 1...24 critical bands, and w1(i, FG) and w2(i, FG(i)) are weighting factors.

In technical implementation the most frequently used method is artificial head based measurement with sound recording. The digital sound recording made in this way is then modified, synthesized and compared with the original. This work is done using headphones in so-called hearing groups. Sound events are analyzed in details then variants are optimized based on comparison of the two results.

The product documentation contains the expected sound and the way of its technological implementation. Before finalization of the product the technological and acoustical viability of the subjective "target sound" is checked and controlled. During the past decade this methodology has been widely used in car industry. One of the latest lines of technological research is mathematical modeling and simulation of constructional body sound conduction. In this process the analysis of the expected result of constructional vibrations demonstrated on artificial head plays a very important role [10].

In the development and technological process new and powerful devices are available for engineers such as artificial head technology, recording technology analogue to hearing, multichannel signal analysis and source analysis based on transfer routes, psycho-acoustical analysis and sound analysis analogue to human hearing as well as sound quality prediction based on mathematical and digital simulation and synthesis with the use of auralisation [9]. In general we can state that all efforts made for the design and implementation of a machine industrial product serves also for the development of sound recording technology and products. It is becoming clear that professional achievements mentioned above are worth using in multimedia applications.



References

[1] E. Zwicker: Procedure for calculation loudness of temporary variable sound. J. Acoust. Soc. Am. 62, 675-682 pp. (1977)

[2] ISO 532B Procedure for calculating loudness level

[3] K. Genuit: The use of psychoacoustic parameters combined with A-weighted SPL in noise description. Proc INTER_NOISE '99, Fort Lauderdale Florida USA, on CD published by Inst. of Noise Controll Eng. of the USA Inc. paper 252 pdf 1-4 pp.

[4] E. Zwicker: A proposal for defining and calculating the unbiased annoyance. – Contributions to Physiological Acoustics; Results of the Fith Oldenburg Symposium on Psychological Acoustics (Ed. By A. Schick, J. Hellbrück, R. Weber) BIS 1991, Oldenburg, 1087-202 pp.

[5] A. Illényi, P. Korpássy: Correlation between loudness and quality of stereophonic loudspeakers – Acoustica, Vol. 49, (1981) pp. 334-336.

[6] A. Illényi, K. Vicsi, A. Vig: Two channel digital set-up to measure loudness. – Noise & Man; Noise as a Public Health Problem Proc. 6th Int. Congress Nice, 5-9 Julliet 1993, Actes INRETS No 34 ter. Vol. 3., p. 251

[7] K. Genuit: How to influence enviromental Noise based on Pscychoacoustics parameters – Proc. INTER NOISE 2000 (Nice August 27 – 30) Vol. 4. pp. 2273-2278.

[8] H. Fastl: Sound Quality of Electric Razors – Effects of Loudness – Proc. INTER NOISE 2000 (Nice August 27 – 30, 2000) Vol. IV. pp. 2173 – 2177.

[9] K. Genuit: The future of Sound Quality of the interior noise of vehicles – Proc. INTER NOISE 2000 (Nice August 27 – 30, 2000) Vol. 1. 427-432.

[10] K. Genuit: Prediction of sound and vibration based on a virtual vechicle. Proc. INTER NOISE 2002 August 19-21., N 139; invited paper. Ed. By Ahmet Selamet, Rajendra Singh, George C. Mahling, Ohio State Univ. Center for Automotive Reserarch, (2002).

[11] H. Fastl: The Psychoacoustics of Sound-Quality Evaluation. In EAA Tutorium Aurally Adequate Sound-Quality Evaluation; Antwerp, Marc 31 1996)

[12] Illényi A., Csányi K. és mtsaik: Mérnöki pszichoakusztika. Jegyzet. 2001 Bp. BME TTT (ftp://domino.ttt.bme.hu/pub/mpa)

Analysis of methods for edge detection

JIŘÍ ŠŤASTNÝ*, VLADISLAV ŠKORPIL**

* Brno University of Technology, Faculty of Mechanical Engineering, Institute of Automation and Computer Science, stastny@uai.fme.vutbr.cz

** Brno University of Technology, Faculty of Electrical Engineering and Communications, Department of Telecoms, skorpil@feec.vutbr.cz

Keywords: Picture transmission, Compression methods, Edge detection, Image recognition

The paper describes the application of algorithms for edge detection in images. The principles and algorithms given below have been used in an application that was developed at Brno University of Technology and has been programmed in the Microsoft Visual C++ environment. In the development, the Win32 API and MFC libraries were made use of. In the application, any of the eleven implemented algorithms can be used for edge detection.

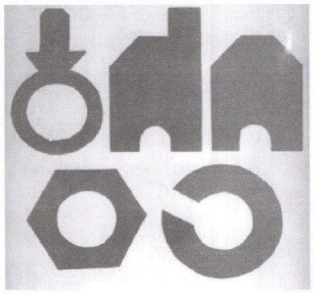
1. Introduction

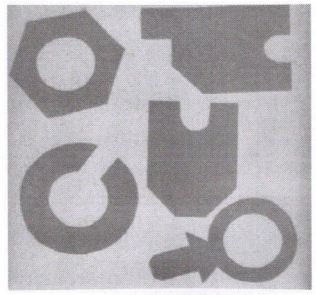
Edge detectors seek to find in an image a sharp change in the intensity of image function. Edges are formed by points where brightness changes abruptly. Edge detection in an image is closely connected with image processing and preprocessing. The application program therefore includes an adjustment of brightness and contrast, gray scale transformation, several filtering methods (median filtering, the Gauss filter, averaging, bar mask)and histogram operations (equalization, automatic search for threshold and maxima). The application contains 11 edge detectors that are used to yield edges (the gradient method [12,15], the Roberts operator [1,15,22], the Laplace operator [1,15], the Prewitt operator [22], the Sobel operator [1,12,15], the isotropic operator [22], the Compass operator [15], statistical methods [4] and [9], the zero passage operator [9,12], the Canny detector [8] and the wavelet transform [5]). These edge detectors were examined from the viewpoint of speed and from the viewpoint of edge detection itself. The evaluation concerned the overall image results, compactness of edges and the amount of residual noise. The Fourier descriptors [14] and object momenta [5] were used for image recognition.

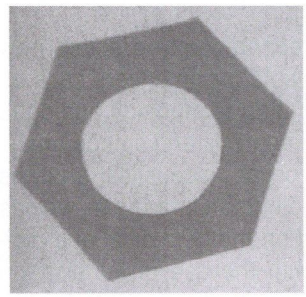
2. Test scenes and environment used

Technological scenes *(Figure 1.)* were used for the comparison of individual algorithms, segments from technological scenes *(Figure 2.)* were used to evaluate details and the picture of Lena (Figure 2.) was used for the comparison with other papers on the Internet (the most frequently used picture when presenting edge detectors).

Figure 1. Technological scenes used in the comparison of results obtained for Etalon and Scene1







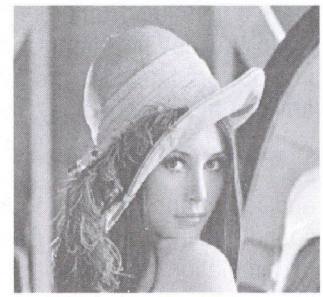


Figure 2. Nut and Lena

Measuring the time of individual functions has been built into the program. The time is displayed in the image frame. The values given in the tables were obtained by threefold execution of the function tested and by calculating the arithmetic mean. This should provide for possible fluctuations caused by programs running on the background. Evaluating the quality of edge detection results is highly subjective, it is based on individual assessment.

Individual edge detectors were compared as to detection speed and results on their outputs. The result should lead to a selection and recommendation of an edge detector best suited for the technological scenes given. A comparison of detectors of real images can be found in [21].

3. Edge detection methods

Edge detectors seek in an image any sharp change in image function intensity. The edges are points where brightness changes abruptly. Edge detectors can be divided into two categories.

One category detects in the image known objects, the other category detects objects in the image without previous knowledge of the scene being processed. An advantage of the latter category is its independence from the image being processed.

3.1. Gradient method

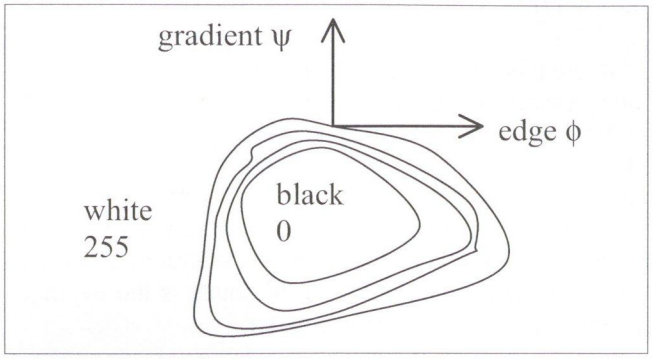


Figure 3. Gradient direction and edge direction

A change in the image function can be described by a gradient [12,15]. The gradient is the direction of maximum growth of image function. Gradient direction *(Figure 3.)* gives the direction of maximum growth of function (from black to white).

The edges are normals to the gradient direction.

The mathematical tool for studying changes in the function of two variable are the partial derivatives:

$$|grad\ \ g(x,y)| = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2} \qquad (1)$$

$$\psi = \arg\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) \qquad (2)$$

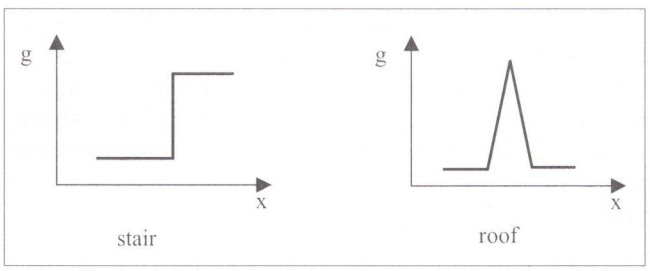


Figure 4. The examples of edge profiles

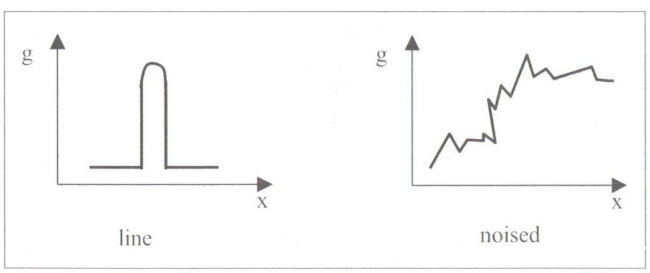


Figure 5. Further examples of edge profiles

Edges can be classified by means of a one-dimensional brightness profile in the gradient direction in a given pixel. Typical examples are shown in *Figures 4 and 5.* The first three profiles, i.e. jump edge, roof edge and narrow line give the theoretical waveform of the brightness profile. In an actual image there are noise-affected edges, as can be seen in Figure 5.

Gradient operators can be divided into three categories:

- Operators that approximate derivatives via differences.
- Operators based on seeking out edges at places where the second derivative of image function goes through zero.
- Operators trying to approximate the image function via a simple parametric model such as a polynomial of two variables.

3.2. The Roberts operator

This is the oldest and very simple operator [1,15,22], which only uses the 2x2 neighbourhood.

Its convolution masks are given:

$$h_1 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad h_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \qquad (3)$$

The magnitude of gradient is calculated according to the relation (4):

$$|\Delta g(i,j)| = |g(i,j) - g(i+1,j+1)| + |g(i,j+1) - g(i,j+1)|$$

The main disadvantage of the Roberts operator is its high sensitivity to noise.

The Laplace operator [11,15] approximates the second derivative of image function. It is invariant with respect to rotation and gives the edge magnitude and not its direction. The convolution cores used for the 3x3 neighbourhood are:

$$h_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad h_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix} \qquad (5)$$

The resultant gradient is calculated by the formula:

$$G = \sqrt{\left(h_1^2 + h_2^2\right)} \qquad (6)$$

3.3. The Prewitt operator

This operator [22] approximates the first derivative. The gradient is estimated in the 3x3 neighbourhood for eight directions. All the eight convolution masks are calculated. One convolution mask is then selected, namely that with the largest gradient module. (7)

$$h_1 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}, \quad h_2 = \begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix},$$

$$h_3 = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}, \quad h_4 = \begin{bmatrix} -1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad h_5 = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

$$h_6 = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}, \quad h_7 = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & -1 \end{bmatrix}, \quad h_8 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -1 \end{bmatrix}$$

If we want to detect edges in only one direction, we a convolution mask that corresponds to the given direction.

3.4. The Sobel operator

Edge detection after Sobel [1,12,15] highlights all the edges in the image irrespective of the direction. The algorithm is applied as a vector sum of two directional edge operators. The resultant image is transformed from the initial image such that spots with constant brightness values are transformed into black patches while spots with changing brightness are transformed into white patches.

The convolution masks of the Sobel operator are given:

$$h_1 = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}, \quad h_2 = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{8}$$

h_1 is used to detect vertical edges while core h_2 yields horizontal edges in an image.

The formula for the calculation of amplitude:

$$M = \sqrt{\left(h_1^2 + h_2^2\right)} \tag{9}$$

The formula for an approximate calculation of amplitude:

$$M = |h_1| + |h_2| \tag{10}$$

The isotropic operator [22]

Gradient convolution cores of the isotropic operator (11):

$$h_1 = \begin{bmatrix} -1 & 0 & 1 \\ -sqrt(2) & 0 & sqrt(2) \\ -1 & 0 & 1 \end{bmatrix}, \quad h_2 = \begin{bmatrix} -1 & -sqrt(2) & -1 \\ 0 & 0 & 0 \\ 1 & sqrt(2) & 1 \end{bmatrix}$$

h_1 is used to detect vertical edges while core h_2 yields horizontal edges in an image.

The formulae for calculating the amplitude are the same as for the Sobel operator.

The Compass operator

In contrast to all the preceding operators, the Compass operator realizes the inverse output image. Its convolution masks are:

$$h_1 = \begin{bmatrix} -1 & 1 & 1 \\ -1 & -2 & 1 \\ -1 & 1 & 1 \end{bmatrix}, \quad h_2 = \begin{bmatrix} -1 & -1 & 1 \\ -1 & -2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{12}$$

The remaining 6 convolution masks are formed by clockwise rotation.

3.5. Statistical methods

The method published in [4] and [9] is again based on the mask passing through input image (thresholded in advance) g(i, j) while evaluating for each point (i, j) the edge value on the basis of the content of submatrix b of matrix g. Checking the existence of an edge in submatrix b is done by calculating the sum of differences or the standard deviation or by following the submatrix form.

Calculation and significance of sum of differences DS. The sum of differences in submatrix b_{mem} of image matrix g is defined as (13):

$$DS = HDS + VDS = \sum_{i=1}^{m}\sum_{j=2}^{n}|b(i,j) - b(i,j-1)| + \sum_{i=2}^{m}$$

where b is the submatrix of matrix g, HDS is the sum of horizontal differences, and VDS is the sum of vertical differences. DS corresponds approximately to the number of edges in the submatrix, 1/DS thus corresponds to the measure of image compactness in the submatrix and it will be maximum if the values of intensity in the submatrix are the same. Calculation and significance of standard deviation. The standard deviation of a set of points is the average difference from their arithmetic mean.

$$MD = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(b(i,j) - \phi\right) \tag{14}$$

$$\phi = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}b(i,j) \tag{15}$$

where ϕ is the arithmetic mean of intennsity values in submatrix b. The standard deviation reaches its maximum if submatrix b contains $m^2/2$ white values and $m^2/2$ black values. It thus reflects the measure of bimodal distribution in the submatrix. Calculation and significance of shape in the submatrix. Since the calculation of 1/DS indicates the diagonal edges in the submatrix more markedly than the vertical (horizontal)edges, we define another measurement of bimodal distribution by claculating the non-linear function

$$B(f,m) = f\left(2m^2 - f\right) \tag{16}$$

where f is the occurrence frequency of points whose shade lies above a pre-determined threshold T, and m is the size of the submatrix. The function B(f,m) will acquire its maximum if there are $m^2/2$ α-values and $m^2/2$ β-values in submatrix b (α-value is a point with the shade above threshold T, β-value is a point with the shade below threshold T). It has been found experimentally that to emphasize the diagonal edges of submatrix b_{mem} it is of advantage to express B as:

$$B(f,m) = f\left(2m(m-1) - f\right) \tag{17}$$

Threshold T can be determined as the arithmetic mean or the median of the brightness of matrix b points.

Methods of deviations. This method evaluates the value of edge EM (edge morit) at point (i,j) in submatrix b_{mem} of input image matrix g on the basis of calculating the quantities MD and DS using the relation:

$$EM(i,j,m) = \frac{MD(i,j,m)}{DS(i,j,m)} \tag{18}$$

which is proportional to the size of edge in submatrix b_{mem}. The value EM is segmented by thresholding in order to obtain binary image.

Method of shapes. This method evaluates the value of edge EM (edge morit) at point (i,j) in submatix b_{mem} of input image matrix g on the basis of calculating the quantities B and DS using the relation:

$$EM(i,j,m) = \frac{B(f,i,j,m)}{DS(i,j,m)} \tag{19}$$

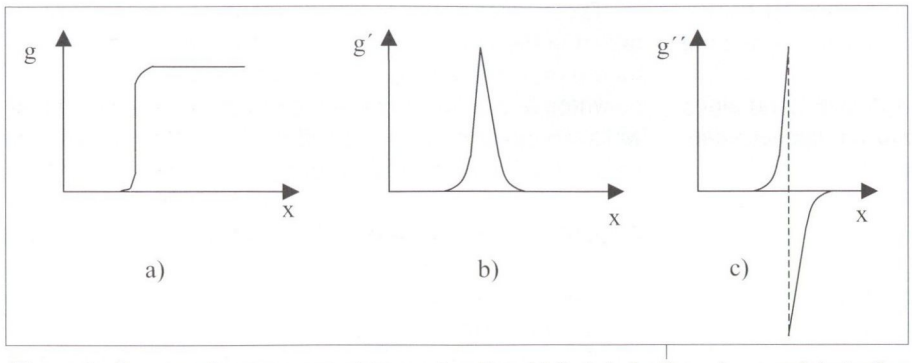


Figure 6. Passage through zero: a) image function, b) first derivative, c) second derivation

3.6. The zero passage operator

The main disadvantage of the the Sobel, Prewitt and other operators is their dependence on a concrete scene, the size of object in the image (the size of convolution mask must correspond to the size of significant details in the image)and, above all, on the sensitivity to image noise. The edge detection technique is based on passage through zero [9,12] of the second derivative of image function *(Figure 6c)*, when the edge corresponds to the steep change in image function. The first derivative of image function *(Figure 6b)* has its maximum where the edge is and therefore the second derivative can be zero in the same place.

In practice it is much simpler to find the passage through zero than to seek the function maximum. First, the image must be smoothed out and noise removed. The key problem in this edge detector is how to calculate the second derivative of image function. The estimate of second derivative (for example using the Laplace operator) exhibits much sensitivity to noise.

This problem is solved by a convolution filter of the following properties:

- The filter should be smooth and, in the frequency spectrum, corresponding roughly to bandpass filter in order to limit the possible number of frequencies at which the passage through zero can occur.
- The requirement for precision of localizing the edge in a plane assumes that the filter will only react to points in immediate vicintiy if the edge.

By the Marr theory [17] the solution consists in such a linear filter whose coefficients in the convolution mask correspond to the 2D Gaussian distribution:

$$G(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}} \qquad (20)$$

where x, y are the coordinates in the image, and σ is the mean quadratic deviation, which is the only parameter and says in how large a neighbourhood the filter operates. Pixels closer to the centre are of greater weight in filtering while the effect of pixels more than 3σ distant is negligible.

For edge detection it is necessary to determine the second derivative G" and the convolution mask coefficients can then be calculated as (21)

$$h(x,y) = c\left(\frac{x^2+y^2-\sigma^2}{\sigma^4}\right)e^{-\frac{x^2+y^2}{2\sigma^2}}$$

where c is the constant ensuring that the sum of all the coeefficient in the mask is zero. The convolution mask can be calculated in advance for a chosen size of neighbourhood σ . For example, a 5x5 mask calculated by means of formula (19) will have the following form (22):

$$h = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & -2 & -1 & 0 \\ -1 & -2 & 16 & -2 & -1 \\ 0 & -1 & -2 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{bmatrix}$$

3.7. The Canny detector

Canny proposed a new approach to edge detection, which is optimum for the detection of edges in an image with white noise. It is based on the calculation of the first and the second derivative (see the preceding zero passage detector).

It is the only detector that tries to meet simultaneously the following three requirements, which are used as standard quality criteria also for the other boundary detectors [8]:

- Good detection of object boundaries in the image – the edge detector must create new boundaries and leave out points with only minimum probability.
- Good localization of edges in the image – the boundaries found must lie as close to the original boundaries as possible.
- Only a single response to the passage through edge – in order not to create double edges.

The Canny operator [15] is actually the result of mathematical optimization of a suitable detector aimed at meeting the above requirements. For edge detection it makes use of the zero passage, which is applied in the direction of local maximum gradient.

The derivation of the Canny detector is built up on several ideas [7]:

1) The edge detector is specialized for 1D signal and for the first two optimum criteria.

2) If a third criterion is added (multiple response), the best solution is found via numerical optimization. The resultant filter can be effectively approximated (with an error of less than 20%)by the first derivative of the Gaussian smoothing filter with standard deviation • [12, 15].

3)The detector is the same in up to two dimensions. The edge is given by its position, orientation and, possibly, its thickness. Since convolution by large Gaussian co-

res is very time-consuming it is replace by two 1D convolutions. One dimension is for the direction of axis x and the other for the direction of axis y.

G is the 2D Gaussian function (see [12, 15])and let us assume that we can use the convolution of operator Gn, which is the first derivative of G in direction n.

$$G_n = \frac{\partial G}{\partial n} = n * \nabla G \tag{23}$$

Direction n should be perpendicular to the edge. If g is the image, the normal to edge n is calculated as

$$n = \frac{\nabla(G * g)}{|\nabla(G * g)|} \tag{24}$$

The position of the edge is then on the local maximum in direction n of operator G_n applied to image g.

$$\frac{\partial}{\partial n} G * g = 0 \tag{25}$$

Substituting in equation (25) for G_n from equation (23) we obtain:

$$\frac{\partial^2}{\partial n^2} G * g = 0 \tag{26}$$

Equation (26) shows how to find local maximum in a direction perpendicular to the edge.

This operator is often referred to as non-maximal suppression. Since in equation (26) convolution and differentiation are associative operations, we can first calculate the convolution of image g with Gaussian function G and then calculate the second directional derivative using the estimate of direction n calculated from equation (24). The edge thickness (the magnitude of the gradient of brightness function of image g)is given by:

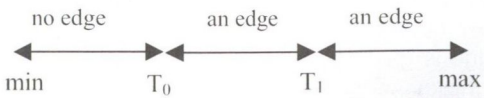
$$|G_n * g| = |\nabla(G * g)| \tag{27}$$

4) A false response to a single edge (multiple response to a single edge) is caused by „streaking" noise. This noise is a problem in most edge detectors. Streaking can be removed by thresholding with hysteresis [9]:

Algorithm – thresholding with hysteresis

1. Determine the hysteresis for thresholding the image of edges in the interval [T_0 T_1].
2. Mark all pixels whose edge values exceed T_1 as real pixels of the future ede.
3. Go through all pixels whose values are in the interval [T_0 T_1].
4. If in the neighbourhood of the examined pixel there is a pixel marked as edge pixel, mark the pixel being examined also as an edge.
5. Repeat going through the image until you obtain a stable output image.

Figure 7. Filtering an edge detector with hysteresis



The correct operator size depends on the objects contained in the image. Different scales of the Canny detector are represented by different quadratic deviations σ. An operator with a large scale is good at suppressing noise while an operator with a small scale is very sensitive to details (that means also noise)in the image.

Algorithm – the Canny edge detector

1. Repeat steps 2) to 6) for increasing values of mean quadratic deviation σ.
2. Calculate convolution g with Gaussian function of magnitude σ .
3. For each pixel, remove local direction normal n to the edge according to equation (24).
4. Find the edge position using equation (26) (non-maximum value).
5. Calculate the edge importance from equation (27).
6. Threshold the image edges via hysteresis in order to remove false any response to the edge

The Canny edge represents a complicated but significant edge detector, which finds application in practice. We can mention its application to the processing of image information from computer tomography [8].

3.8. Wavelet transform

The wavelet transform [5] can be considered an extension of the Fourier analysis of periodic oscillations to a general non-periodic case. While the Fourier series expresses data as a mixture of harmonic sinusoidal oscillations, the wavelet transform expresses data as a mixture of damped oscillations (wavelets). The wavelet transform is based on a suitable change of the width of „window" in time and using its shape to obtain their optimum ratio for object recognition. For low frequencies the „window" is wider, for higher frequencies it is narrower.

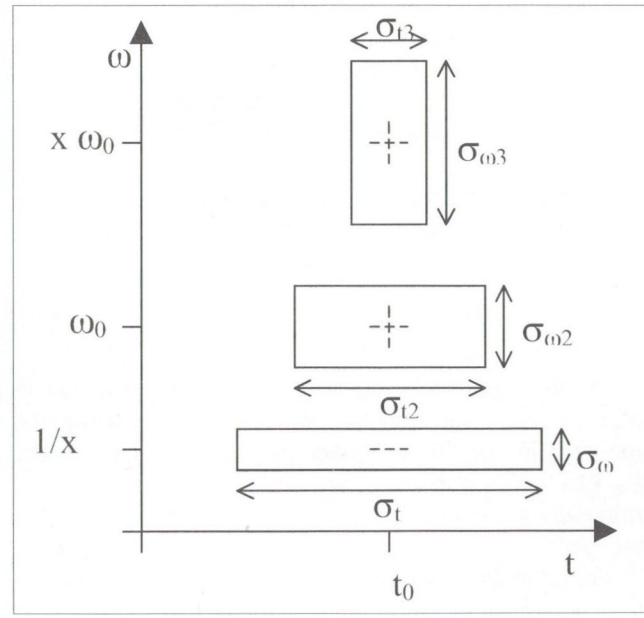


Figure 8. Time-frequency resolution of wavelet transform

Wavelet transform and edge detection

The main advantage of the wavelet transform in edge detection is the choice of the size of details to be detected. The size of detected details is set by the wavelet scale. In the case of discrete wavelet transform it is done by multiple passage of the signal through the wavelet filter.

When processing a 2D image the wavelet transform is performed separately for the horizontal and the vertical function. In this way, horizontal and vertical edges are detected. Edge detection can be conducted in two ways. We can recognize edges via zero passage analysis (LP), in which case it corresponds to the second derivative of input function or via the detection of maxima (HP). The detection is thus based on recognizing individual function maxima. In the latter case a threshold must be chosen according to which thresholding will be performed.

The wavelets used in the program have been taken over from [5]and compared with the other edge detectors. For interest 's sake, only the CRF (13, 7)and SWE (13, 7)wavelets were added, namely for the purpose of using them in the JPEG2000 compression.

3.9. Edge Detection Speed

Table 1. gives the times of all implemented edge detectors that were run on the test computer.

The tests were conducted on a scene (Figure 1). The results show the speed to be dependent on the hardware used. It is obvious that the processor and the speed of graphics card memory will have the main effect on the speed of computation.

	Parameters	Time [s]
Canny	sigma 0,5, tlow 0,25, thigh 0,7	0.91
Thresholding	thres. 135	0.036
Method of deviations	thres. 135, mask 3	1.34
Sobel operator	thres. 199	0.16
Isotropic operator	thres.199	0.284
Laplace operator	thres. 120	0.18
Prewitt operator	thres. 199	0.13
Compass operator	thres. 255	0.21
Gradient method	thres. 35	0.047
Zero passage operator	thres. 230	0.07
WT CDF (1,1)	thres. 26	0.51

Table 1. Speed of individual edge detectors

It follows from the above values that edge detection speed is directly proportional to the complexity of the methods applied. From the viewpoint of detection quality, simple methods, characterized by a high speed, do not give the best results. A logical step is then to choose detectors that give good results in the available time.

4. Edge detection precision

Unless stated otherwise, edge detection took place without prior image modification or preprocessing. Examples of the edge detectors used:

4.1. Thresholding

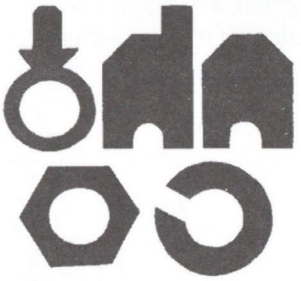


Figure 9. Image thresholding

The effect of illumination uniformity on thresholding results is evident from the figure. *Figure 9* – the scene was illuminated uniformly, , thresholding was performed without any problems.

The result can be used for further processing – the scene periphery is not illuminated to the same extent as the centre. Fringes appear on the periphery, which make further application impossible. The problem can be resolved by better illumination. The properties of an image cannot be improved significantly by a general program modification.

Figure 10. Thresholding – detail a) nut T =120, b) Lena T =130

Figure 10a) – noise residue due to imperfect illumination can be seen in the left--hand part.

Figure 10b) is properly segmented, it preserves all the important image contours. Thresholding is the simplest and also the fastest method of image segmentation. If the thresholding result is similar to that in Figure 10a), it can be used as the input image of edge detector. The results are then much better than if the edge detector is used by itself.

4.2. Gradient method

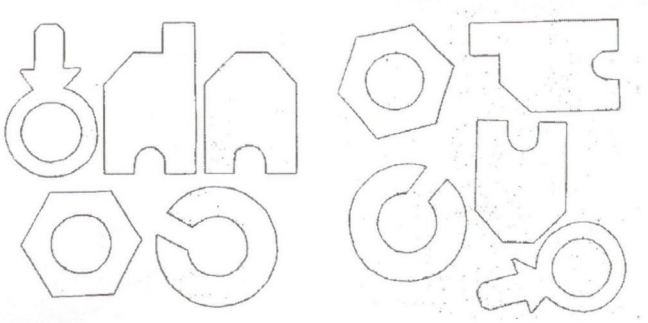


Figure 11.
Gradient method: a) T =35, b) T =25

The operator does not produce compact edges (they are often interrupted). The edges consist of arrays of single points. A great amount of noise remains in the image area.

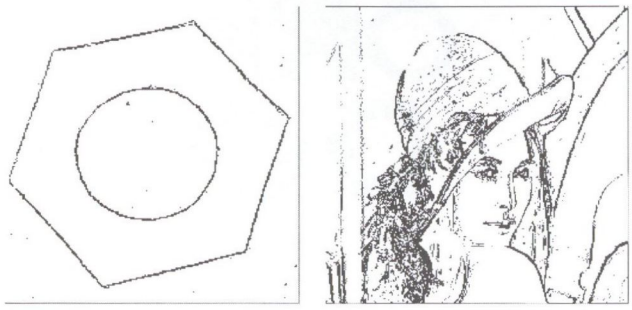


Figure 12.
Gradient method – detail a) T =25, b) WT CDF (1,1) T=20

Figure 12a) – doubling of edges appears in object corners. *Figure 12b)* shows clearly the arrays of points mentioned above. The results are not very good. The results can be improved by combination with the thresholding method.

Combining thresholding with gradient detector yields a compact edge without residual noise in the image. Provided that appropriate thresholding conditions are fulfilled, a very fast edge detector with good results is obtained.

4.3. Zero passage operator

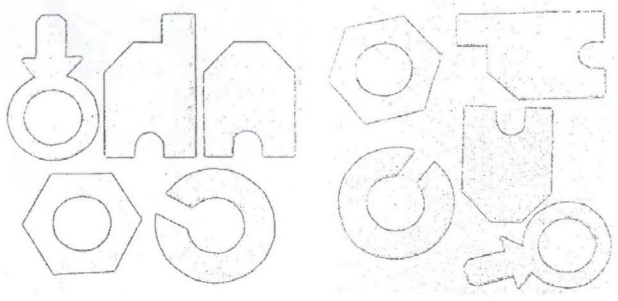


Figure 13.
Zero passage operator: a) T=230, b) T =210

In the case of zero passage operator the points are much smaller than for the gradient method.

Also, the amount of noise in the image has increased.

Using the thresholding method results in compact edges without noise in the image. The edges are, however, thicker. Also, the detection speed gets reduced due to greater time demands of this method.

4.4. Prewitt operator

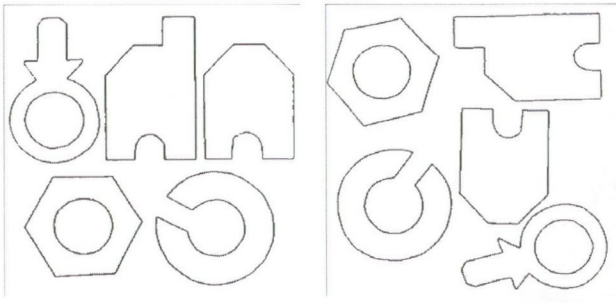


Figure 14.
Prewitt operator: a) T =199, b) T =144

The operator yields comparatively compact edges, which are interrupted at several points. The edges are frayed in some places. The edge thickness is several pixels. The image area is clear, without noise. This is a mediocre edge detector.

4.5. Laplace operator

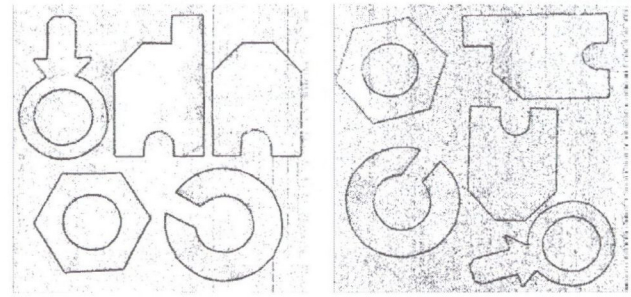


Figure 15.
Laplace operator: a) T =120, b) T =100

The operator doubles the edges. The edges are not compact. There is much noise in the image and it overcasts the edges proper. Preprocessing via filtering will partially suppress the noise.

4.6. Sobel and isotropic operators

Despite different convolution masks, the two operators yield the same results. The Sobel operator is faster than the isotropic operator. Edges detected by these operators are compact, with occasional fraying. The image is clear, without noise. The Sobel operator can be used with advantage for edge detection in a technological scene.

4.7. Compass operator

This is not an edge detector but a surface detector and is thus an analogy to the method of thresholding. It is, however, four times slower and gives much worse results than thresholding does. Its application is not suitable for the scenes tested.

4.8. Wavelet transform CDF (1, 1)

Edges detected by Wavelet transform CDF (1,1), are incompact, with noise in the image. For the scenes tested it would be necessary to resort to image preprocessing.

4.9. Method of deviations

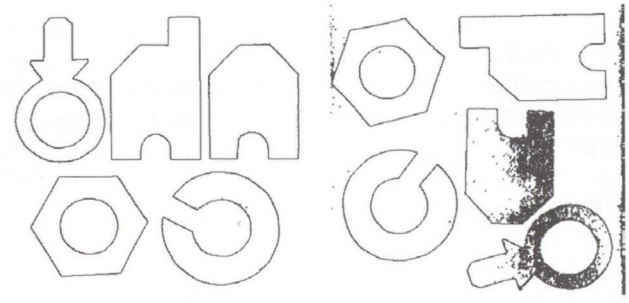


Figure 16.
Method of deviations: a) T =135, b) T =120

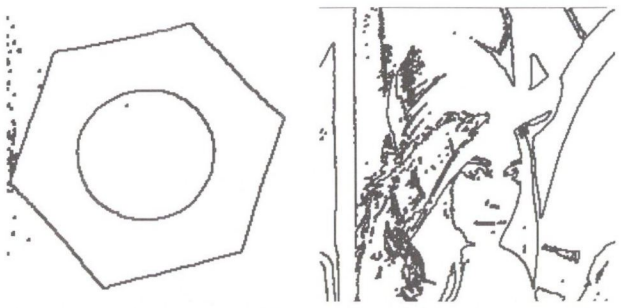


Figure 17.
Method of deviations – detail a) T =120, b) T =120

The method of deviations depends on thresholding. If the conditions (given for the thresholding test)are fulfilled, the method yields compact smooth edges without interruption, without noise in the image. Any noise that might appear during thresholding will be multiplied in the method of deviations.

4.10. Canny detector

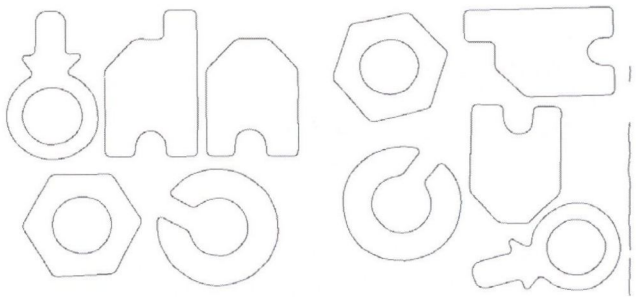


Figure 18.
Canny detector a) σ = 5; tlow =0.250; thigh = 0.7;
b) σ = 5; tlow = 0.250; thigh = 0.7

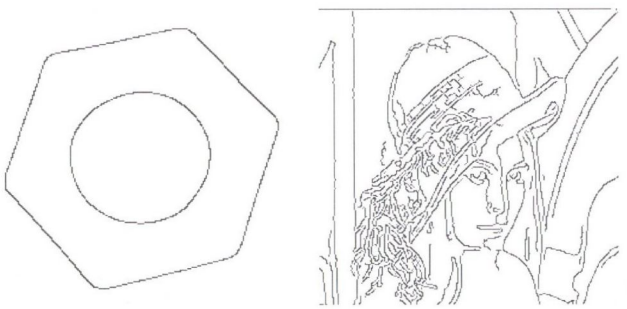


Figure 19.
Canny detector – detail a) σ =5; tlow = 0.250; thigh = 0.7;
b) σ = 1; tlow = 0.250; thigh = 0.7

This is perhaps the presently most elaborate edge detector, which gives very good results. The high quality of edges is paid for by the longest processing time of all the edge detectors tested. The resultant edges are compact, continuous and smooth and their thickness is 1 pixel. The image is without noise. This detector has everything (excepting speed) that can be required of an edge detector. For a simple scene (the technological scene set)it is as satisfactory as the Sobel edge detector or the Prewitt operator. If successful thresholding can be assured, it would be better to combine thresholding with the gradient method. For images of greater complexity (such as Lena), the Canny detector is the most suitable.

5. Object recognition

Two implemented methods for recognizing objects in an image are compared in the application. Reference standards taken from the image Etalon are prepared on the hard disk.

They are text files, where the first item is the name of the object, followed by descriptors. In the case of FD the value is 40, for object momenta it is 7. *Table 2* gives the times of object recognition (inclusive of following the boundary and joining the edges). The long times of object momenta are due to filling the areas where the simplest algorithm is used. To reduce these times, the following variants can be suggested:

To use another algorithm to fill the area (line filling).

To use thresholding or object colouring. Then it would no longer be necessary to fill the area.

The speed of recognition via FD is very good but the results are not as good as with object momenta.

	Type of recognition	Computer [s]
Canny	FD	0.09
	Momenta	1.07
Isotropic operator	FD	0.11
	Momenta	1.12
Prewitt operator	FD	0.107
	Momenta	1.087
Sobel operator	FD	0.084
	Momenta	0.81

Table 2. Speed of object recognition

The great effect of the speed of FD and object momenta is apparent from the above values.

6. Conclusion

The rate of success was examined by way of performing recognition on all technological scenes with edge detectors given in the comparison of edge detector speeds. For each edge detector, new reference standards were prepared from the edges detected in the initial scene (Etalon).

With the FD method, the rate of success in finding objects in the image was only 34%. Its rate of incorrect classification of the object, namely 33%, is also too high. In [14] the author gives 95%rate of successful classification but for a case when a maximum raster of 64x64 pixels is used, which is not comparable with present-day resolution values. It is with finer rasters that much more changes in the line slope of non-orthogonal edges take place and thus other FDs are calculated. This problem arises exactly in non-orthogonal objects such as nut or washer. In view of the above facts, the classification using FD can be recommended for objects of the type of nut or washer. If the problem of smoothing non-orthogonal edges could be solved, FD would be applicable to general objects too.

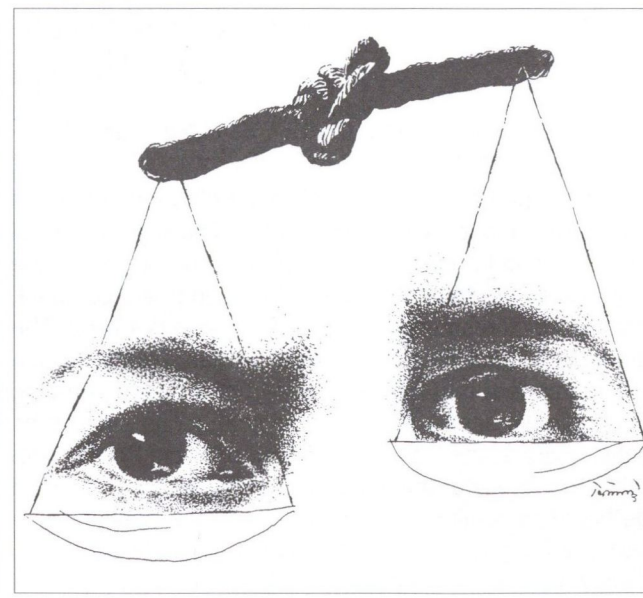
The method of object momenta gives very good results, 74% of successfully recognized objects. It is very interesting that this method does not exhibit any incorrect classification. The greatest problem experienced with the method was the recognition of the object of a key in the image. If the problem of low speed could be resolved, this method would be very good to apply.

Acknowledgments

This paper has been supported by the research project CEZ:J22/98:26100009 Non-Traditional Methods for Investigating Complex and Vague Systems of Brno University of Technology.

References

[1] HLAVÁC, V. and SONKA, M.
Computer Vision. GRADA,
Prague 1993
[2] GONZALES, R., C. and WOODS, R., E.
Digital Image Processing Addison-Wesley Publ. Co.,
New York 1993
[3] ZÁRA, J. and BENES, B.
Model Computer Graphic. Computer Press,
Prague 1996
[4] SMITH, M., W. and Davis, W., A.
A new Algorithm for Edge Detection. John Wiley,
New York 1974
[5] STASTNY, J.
Application of the Wavelet Transform to Edge Detection. In Proceedings of the 2nd International Conference APLIMAT 03.
Bratislava 2003
[6] SVITÁK, R.
Edge Detection on Images. [Project-online]. ZU FAV.
http://herakles.zcu.cz/~rsvitak/school/zvi_rsvitak.pdf
Plzen 2001
[7] SONKA, M. and HLAVÁC, V. and BOYLE, R.
Image Processing, Analysis and Machine Vision. PWS,
Boston 1998
[8] NAGY, I. and HUSTÁK, J.
Processing of Image Information from CT [Research Report]. ÚT AV CR a ÚMT FSI VUT,
http://www.cbmi.cvut.cz/konference/skelet/skelet2000/Nagy_Hustak.doc
Brno 2001
[9] JEZEK, B.
Computer Graphics II. [Lectures-online]. UHK FIM
http://lide.uhk.cz/home/fim/ucitel/fujezeb1/www/
Hradec Králové 2002
[10] SOCHOR, J.
Computer Graphics [Lectures-online]. MUNI FI
http://www.fi.muni.cz/usr/sochor/,
Brno 2002
[11] BULB, M.
Programming [Online].
www.freesoft.cz/projekty/vyhen/clanky/prog/bres.html
Prague 2000
[12] SONKA, M.
Course Digital Image Processing. [Online].
css.engineering.uiowa.edu/~DIP
Prague 2002
[13] PELIKÁN, J.
Completing of Indiscrete Area [Online]. KSVI MFF UK
http//sun3.ms.mff.cuni.cz/~pepca/,
Prague 2000
[14] BOGR, J. (1984)
Image Recognizing by the Fourier Descriptor Method [PhD Thesis]. VUT,
Brno 1984
[15] HIPR, J.
Image processing learning resources. [Online].
www.dai.ed.ac.uk/HIPR2,
Prague 2000
[16] SNOREK, M. and JIRINA, M.
Neural Networks and Computers. CVUT,
Prague 1996
[17] ZIOU, D. and TABBONE, S.
Edge Detection Techniques
[18] Internet Locator [Online]
<www. google. com>
[19] FOJTÍK, J.
Algorithms of aero snaps vectorisation. [Online].
cmp.felk.cvut.cz/~fojtik/vectoris/vektoris.htm,
Prague 2001
[20] FISHER, R.
World and Scene Representations. [Online].
www.dai.ed.ac.uk/CVonline/repres.htm,
London 2002
[21] HEALTH, M. and SARKAR, S.
Edge detection comparison. [Online].
marathon.csee.usf.edu/edge/edge_detection.html,
New York 1996
[22] KHOROSWARE, J.
Edge detection I. [Online].
www.ee.bgu.ac.il/~greg/graphics/special.html,
London 2000



Textual functions
New prospects in e-administration

JÓZSEF VÁRKONYI

t_ford10@elender.hu

Keywords: Evolution of texts, Contradiction findings, Control of legal texts

It was 12 years ago that I discovered that any textual knowledge such as laws can be expressed in a full, exact and solid manner by an "IF – THEN" based formalism. Some obvious advantages of this realization became clear soon. When, for instance, this formalization is applied on legal texts, a defect or gap in the law, discrepancies and inaccuracies can be discovered clearly. There is an even more exciting field of application: using the formalized version of a law, and a suitable system framework, the behavior and mechanism of the law can be examined by computer simulations.

Historic overview

Any case falling into the "scope of interpretation" of the law could be simulated by computer with a full and adequate conformance to the contents of the law under test. The process also demonstrated that the computer manages the formalized version of the legal text as if it were the mathematical formulation of a technical task. The specific formalization offered practical advantages which were then justified by several experimental applications. However, the "phenomenon" itself seemed to be an ad hoc matter both for the legal and the IT community. Legal and information technology professionals generally did not believe that there can be a methodically correct way for textual knowledge and laws to be formalized and modeled as described above. Initially this skepticism was justified in that formalization had been carried out by instinct for a long time.

The breakthrough came in 1997 when the research work resulted in the recognition that the formalization of texts can be determined exactly within the framework of function and set theory. Based on this recognition one could already determine notions of *textual function variable* and *textual function shape*.

Following this recognition all phenomena in this field became interpretable and also the mapping procedure became more conscious and "stable". Finally, we could identify potential scopes and perspectives of practical utilization. Thanks to a fortunate coincidence, it was in that time that problems like legal harmonization, modernization of public administration, then information society and especially e-administration became very popular, all of which can be efficiently supported by the use of textual functions.

Year 1999 marked a milestone in the process. In this year a close co-operation was formed with Institute of Control and Information Technology at the Budapest Technical University led by prof. dr. Péter Arató, then in year 2001 also with Institute of Law and State Affairs at Pázmány Péter University. With this latter university I held lectures for PhD students in law (they were enthusiastic about the methodology and the subject as well).

The textual function shape

The easiest way to present a textual function shape is to show an application, e.g. mapping a law.

1. Text of the law is decomposed into elementary and/or partial interrelations, i.e. into basic levels of interrelations;

2. At the basic levels of interrelations factors (notions) are determined and methodized depending on whether they are of conditional nature i.e. are *input type notions* (factors) or represent the *output* of a basic interrelation under test. In other words, we determine that at the level of notions "what are dependent of what". (This is a logical procedure in a text destined for practical use.)

3. Factors determined in this way are categorized according to their content and the resulting sets of notion are given a name typical of their content. In this way, elements of individual sets of notion are formed by the original notions as found in the text. E.g. in case of the law on income tax, all itemized income type such as those coming from employment, intellectual activity etc. are counted to the set "type of income". This means that the set of "type of income" contains only elements which are mentioned as type of income in the text of the law.

4. In the next phase sets of notion are handled as *function variables* whereas original notions are interpreted and handled as possible *values* of function variables. Function variables can be *independent variables* (if they are input variables) or *dependent variables* (if they are output variables). From another point of view: function variable can be of *textual type* if its possible values are of textual type, or *numeric* if its value is numeric. E.g. in the law on income tax, the amount of revenue is a numerical independent variable while taxable value is a numeric dependent variable. On the other hand, "type of income" is a textual function variable (because its value is "from

employment" or "from intellectual activity" or ...), which is an independent variable since different tax base calculation algorithms shall be applied depending on different types of revenue. The *interpretation range* referred to independent variable named "type of revenue" is formed by the aggregate value allowed by the text of law while the *set of value* of tax base is formed by the entirety of potential values.

In the text of the law notions like "revenue from employment" or "revenue from intellectual activity" etc. figure, their function variable representation follows as applicable:

type of revenue = revenue from employment,
type of revenue = revenue from intellectual activity.

5. Basic and partial interrelations of the law are re-constructed from function variable shape of the applied notions i.e. from function variables, but in "IF ... THEN ..." rules i.e. in the form of a textual function. In this way one can obtain a closed, coherent function system organized in a tree structure that represents fully and exactly the original text by clear assignment and mapping of values of sets of notion.

Let's take some rules describing basic interrelations in the field of income tax:

Tax base, employment
IF type of revenue = revenue from employment
THEN amount of tax base = 1 x amount of revenue

Tax base, intellectual 1
IF type of revenue = revenue from intellectual activity
OR revenue from other source
AND way of cost reckoning = non itemized
THEN amount of tax base = 0.9 x amount of revenue

Tax base, intellectual 2
OR revenue from other source
AND way of cost reckoning = itemized
THEN amount of tax base = amount of revenue – amount of costs

Experiences suggest that this very simple and logical formalism is apt for the accurate and closed representation of any law. In this opportunity the operative role is thus fulfilled by the systematic formulation of textual function variables. This manifests itself in the arrangement of notions into closed sets which contain always a finite number of elements and these elements cover exactly and optimally the whole set of notion of the law. It is important to know that:

• These are textual function variables, not logical ones. This means that we are not confined by bivalence. In this way formalism becomes flexible, nonetheless we can create a closed, logical order both in relation of the applied notions and their complex interdependencies.

• The creation of a function format ready to run on a computer does not require computer skills and the resulting function format can be directly read and interpreted by a person who knows the original legal text. This means that the given part of the process does not involve programming in the traditional sense. Furthermore, computer simulation can be carried out by anyone who is generally able to use a computer. This is to suggest that all computer related works within the use of textual functions in legal and administrative fields are limited to the creation of the framework (or competing frameworks on the market) that are suitable for the execution of simulation. Obviously, creation of "contents" ready to run on computer can fully belong to the competency of legal and administrative authorities. This is however, a mathematical (and not information technology) task. On the other hand, the mathematical character of the problem opens up a great new perspective for the information technology.

Textual function shape and e-administration

In early 1998, after several experimental applications the Communications Authority, Hungary (CAH), charged me of the formalization of government orders regarding the licensing of certain telecommunications services and of facilitating the computer simulation of legal texts. This task lasted about two months in course of which we faced several inconsistencies part of which were known by the professional of the Authority. Finally, by playing some specific cases selected by CAH we could demonstrate the functioning of simulations of laws. (This involved to simulate any case which was allowed by the content of the law.)

Simulation is helping us to determine in an objective manner that a certain company with given data and specifications wishing to provide telecommunications services will get the license or the application will be refused. Similarly, the adequacy and the outcomes of a frequency auction can be determined. Using the function-like version, the text of the law was represented for the issue under test and then for a specific subject in a customized way.

In this regard CAH staff realized a potential scope of application: the procedure could and should be used at customer service centers of CAH to inform applicants – in full, clear and customized manner – about legal requirements of licensing. This would result not only in a better and clear customer information service but also in more complete applications. (Owing to changes in the staff this work was cancelled.)

The recognition outlined above led to a novel approach of e-administration: full texts of laws governing licensing and other public administration procedures should be made available over the Internet. (In this case "full texts" refers not only to the law itself but also to closed and coherent function shapes of regulatory statutes at different levels.) In this construction not only customer information but also the draw-up and presentation of the application can be realized in an electronic environment. (Obviously the necessary related documents shall be made available through other means.)

Applications used with CAH and other organizations, however, highlighted an easily detectable and not Hungary-specific problem: owing to consistency problems laws are not suitable for electronic use. Though formalization can reveal gaps, defects and contradictions in the text of

the law, making the function consistent is not enough. The law itself can be modified only by authorized legislators while the function shape cannot be neither "better" nor "worse" than the prevailing law.

The use of the textual function shape allows for a unique solution in course of legislation. The reality of this recognition was justified by a project recently completed with the Ministry of the Interior. (The solution was successfully experimented in the preparation of law on civil firearms, i.e. in a socially delicate field.) First a closed, coherent, ready-to-run model was set up using the function shape and the text was formulated based on this model. In course of the social debate currently in progress we assess incoming opinions against this coherent function system (structure and process description). Opinions do not affect the structure and its bases just certain branches. Following this method experts can easily include acceptable opinions in the right place of the text.

The solution outlined above allows for the creation of laws which can readily be used in electronic, e.g. Internet environment. This feature can significantly enlarge the scope of e-administration, then – following due generalization – that of the information society.

The information technology environment

In course of experiments I have been using up to now a borrowed system framework for simulation of laws. This program system – ALLEX PLUS 2.0 – is a PROLOG based expert system framework developed in the late 80's by a team led by dr. Iván Futó as a joint work of companies SZÁMALK and SZKI. This framework was designed for use under DOS. In the simulation it meets all essential criteria but the compilation of knowledge base is rather uncomfortable and its screen presentation lags far behind today's elegant and smart solutions. The main problem is that it is not suitable for use on the Internet so it cannot be used in e-administration either.

It would be necessary to develop a system framework which would replace ALLEX and could operate over the Internet. Prof. dr. Péter Arató (Institute of Control and Information Technology at the Budapest Technical University) and his colleagues became familiar with the theme and were ready to help me in starting the development work. We agreed that a new program system was needed, nonetheless because on the international market there is no proven solution in the field of law which could compete with us either at product or publication level. (There are, of course, experiments based on artificial intelligence but these did not result in a general conclusion or methodology.) However, applications of the Institute for system development were steadily refused.

It should be noted that – as suggested by explanations of textual functions – there are here two fields which could easily be separated: a formal linguistic and methodology section and an information technology section. (This separation is justified by the fact that the linguistic and methodology section could be used even 200 years ago.) The linguistic-methodological part of the project was completed about four years ago and was "operational" from IT point of view. It was demonstrated in an ALLEX environment and was "ready to sell" to an investors' group. It would represent, however, quite a different scientific, business and prestige value if it offered not only a proven methodology but also a turn-key, business-like solution which could be used "industrially" in a modern information technology environment in the field of law, public administration and e-administration and Hungary were able to introduce it to the civilized world heading toward the information society.

ITU-News

A number of countries in the region expressed interest in hosting the **AFRICA 2004** event, and three detailed offers were received, from Algeria, Egypt and Senegal. A series of consultations and negotiations was undertaken, with particular consideration given to infrastructure, accomodation, transport and conference and exhibition facilities. Given the excellent offer made by Egypt as well as its significant progress in IT and telecom in the past few years, and on advice from the ITU TELECOM Board, Yoshio Utsumi accepted the offer of the Government of the Arab Republic of Egypt to host the event. The ITU TELECOM Board, which represents exhibitors'views, provides strategic advice to the Secretary-General. AFRICA 2004 will be held from 3 to 7 May 2004 at the Cairo International Conference Centre (CICC), with a Press and VIP day on Sunday 2 May. The venue is ideally located and offers advanced forum and exhibition facilities, which will be enhanced by plans to build a new permanent exhibition hall offering a further 8,000m^2 of net exhibition space to be ready by the end of 2003. The CICC was previously used by ITU to stage the AFRICA TELECOM 94 event, which was attended by nearly 12,000 participants and opened by His Excellency Mr. Mohamed Hosni Mubarak, President of the Arab Republic of Egypt, on 25 April 1994. The AFRICA 2004 exhibition will feature a comprehensive range of telecommunications-related products and services, while the associated forum will focus on the latest telecommunication developments and growth and will provide a platform for telecommunication leaders to share their ideas on future trends and discuss appropriate strategies for the developing as well as the industrialized world.

Simple inter-domain propagation algorithm for the ProFIS architecture

ANDRÁS GULYÁS, ISTVÁN PATAKI

High Speed Networks Laboratory, Department of Telecommunications and Telematics
Technical University of Budapest,
pi205@axelero.hu

Keywords: Internet, Packet propagation, Optimal routing

This paper deals with providing Quality of Service (QoS) over IP based networks. We are going to give a brief survey about this topic, and present our work at this area. There are many solutions of the problem, but the standardization of the methods is not finished yet. At the moment there are two kinds of approaches of the reservation problem. The distributed method handles the network nodes independently, and get the nodes making their own admittance decisions along the reservation path (i.e. Border Gateway Reservation Protocol BGRP [1]). The centralized way -we discuss in details-, which collects the network nodes into domains, and handles them using a network manager. Generally there are two significant parts of the network management: intra domain, and inter-domain. This article focuses on making reservations over several domains, which is the part of the inter-domain functions.

1. Contents

First we give a short overview of the QoS providing over IP networks, and it's reason for the existence. In section two we discuss the principles need to be taken to provide QoS over IP networks. In the rest of this section we describe IntServ and DiffServ, and the QoS architecture using Bandwith Broker. At the end of this part we deal little with the ProFIS architecture. In the third chapter we introduce our inter-domain communication protocol for the ProFIS, and at the end of the document we are giving a summary of our work.

2. Motivation and brief survey of providing QoS

Recent times the performance of the personal computers increases likewise the number of real-time Internet and multimedia applications. In case of these applications *best effort* traffic is not enough for satisfying the quality claim of the users. The best effort guarantee is an elementary provision of the Internet. The network elements try their best to deliver the packets to their destination without any bounds on delay, jitter, and latency, but they cannot give any guarantee for the delivery. These "guarantees" are not sufficient for i.e. a videoconference, because delay over a limit, or jitter can cut down or bust the interactivity and usability. The goal of the Internet Service Providers (ISPs) is to satisfy the quality demand of the customers and ensure the same sort of QoS and reliability over IP networks as in the circuit switched networks. By applying packet classification they can deliver different kind of services on the same link without the suffering of the important flows. The IP QoS is one of the most important research areas in our days. The development is driven by the increasing demands of the customers for service quality and reliability. Most of the technological challenges have been solved, now it is a matter of standardizing the technologies, and making the system scaleable. To solve the problem of scalability is one of the most important challenges because of the rapid growth of the modern Internet.

Main negative of the Public Switched Telephone Network (PSTN) that it can only deliver one kind of data. The IP technology is more flexible than any circuits switched provision as it can carry different kind of traffic on the same link. The Internet is a complicated, heterogeneous system, which contains lot of different Autonomous System (AS) with different routing algorithms and different QoS technologies, applications. The number of the QoS architectures is high, but the interoperability and the standardizing are still not solved.

Since the actual Internet architecture does not provide mechanisms for resource management and isolation of the flows, all of the running services suffer in the congestion periods. Hence, in order to provide quality of service, an important step is to implement admission control mechanisms.

Other shortcoming of the current IP networks is that IP does not have the technical support for offering premium services. Each transmitted packet in the network is treated in the same way, the treating functions do not depend on the carried information of the packet, only on the destination address.

2.1. IP QoS principles

To solve the problems described above and provide QoS over IP networks, different principles have been developed:

- Packet classification is necessary to make the network capable to differentiate between the various categories of the traffic. The traffic must be classified based on its requested parameters like delay, jitter, bandwidth, price etc
- Isolation: scheduling and policing. The packets of the different categories must be scheduled different ways, and treated according to the policies. Each packet in the same QoS class is treated using the same method.

• High resource utilization is very important because of the economy view of the service provider.

• Call admission is required for accurate resource management in order to avoid congestion.

Although there are several approaches to the problem, two main QoS models can be considered for deployment: Integrated Services (IntServ) and Differentiated Services (DiffServ). These models are widely researched and accepted. The models or their combination seem to be good solution for providing QoS on the Internet.

3. IntServ

First in 1993 Integrated Services was developed by Internet Engineering Task Force (IETF). In IntServ, a signaled QoS model is defined, where resource requirements are signaled from an endpoint, and the network device honors the signaling and reserves resources for flows. The protocol used for signaling here is Resource Reservation Protocol (RSVP) [5] . In addition to provide QoS per flow, in this model, admission control of traffic flows is available because of the inherent signaling ability.

The RSVP has two main messages. The PATH and the RESV message.

End stations, proxy devices, or voice gateways can initiate signaling by sending a PATH message with filter specification. The filter specification contains source, destination addresses and port numbers along with bandwidth requirement for the "flow", which can be defined as traffic going user to user, session to session, gateway to gateway or proxy device to any other device in the network. When the PATH message traverses the network along each hop the network device performs a check, initializes a state for the flow and forwards the Path message toward the destination. When the PATH message reaches the destination, the destination device responds with a RESV message including the bandwidth requirement for the flow. The RESV message traces its path back to the source. Along each hop the network device performs admission control. If there is enough available bandwidth, it admits the flow by assigning resources to it such as queues, weights, etc. and forwards the RESV message upstream toward the source. When the RESV message reaches the source, the signaling process is complete. When the service starts, network devices classify the traffic, recognize that it belongs to a reserved flow and put the packets into appropriate assigned queues so that the traffic gets the treatment it needs or signaled for.

Using IntServ and RSVP the flows must be administrated in every node along the path, and they are identified by their source and destination address. The number of the Internet hosts increases very quickly and ere now overshot the 120,000,000. Resource reservation schemes must scale well with the growth of the number of the hosts, because a router may be able to handle tens of thousand of reservations at same time. In case of the today's Internet the flow number can be a million in each router of the path, and such a big flow number debase the performance of the routers. Hence this method is not scalable for large networks because of the quick growth of the flows. One possibility to solve the scalability problem is that you do not distinguish between the packets, just the group of the packets. Other possibility is to group the nodes and so put hierarchy in the system. In the following chapter we examine the DiffServ model, which was developed driven by the demand of scalability.

4. DiffServ

DiffServ [1] as defined in the IETF RFC 2475 is a model that allows deployment of QoS in a simple fashion with network devices only handling traffic at an aggregate level rather than per flow.

The six most significant bits in the TOS byte of IP header is defined as DiffServ Code Point (DSCP). Packets are marked with a certain value depending on the type of treatment the packet must receive in the network device. Traffic is aggregated into traffic classes that require the same treatment and marked with a DSCP value in the TOS byte. DiffServ defines the DS domain, which is a continuos set of DiffServ capable nodes. The complexity of the network was moved to the edges of the domains. Filters are configured at the network ingress to identify the traffic and mark the traffic with the appropriate DSCP. The ingress routers are also responsible for policing and shaping. So inside this points a queue must be set up and a drop policy must be defined. Besides this, a policier-shaper must be configured and aggregated bandwidth must be allocated to the queue.

The size of the DSCP is six bit, so it is able to manage up to 64 different behaviors. Hence the traffic of a DiffServ managed network from any source to any destination must fit into one of the 64 behaviors. In the DiffServ architecture each packet, which has the same DSCP, get the same treatment irrespective from the source and destination of the packet. So in a DiffServ node requires less entry inside a node than the flow entries in the same node in case of IntServ. It seems to be a good idea to manage the DS domains, because a domain manager is able to use the resources of the domain more effectively, and makes possible to serve the customer's claims. The domain manager of the DiffServ model is called Bandwidth Broker as introduced in RFC 2638 [6].

4.1. Bandwidth Broker

Each domain has at least one BB, which is the manager of the domain's resources. The BB knows the topology, and has correct information about the currently reserved and free link capacities. The main functions of the BB are the following. Managing the resources of the own domain. These functions are the intra-domain functions. Another group of the BB functions is the inter-domain communication, it includes the communication of the availability information with the adjacent domain's Bandwidth Brokers, and the negotiation based on the received availability information.

Within the range of the intra-domain functions the BB manages the resources of the domain. If the BB receives a reservation from an end user, or an adjacent BB, checks the topology file. Inspects every link along the path, if the needed bandwidth fits in the unreserved capacities. If the claim can be satisfied the reservation can be admitted into the domain. After the decision was done the BB reserves the resources for the admitted reservation. Sets up the border routers. The main advantage of the DiffServ, and it makes it to a scalable architecture, that only the ingress routers need to be configured, because the routers inside the domain only forward the packets along the path towards the egress router according to the predefined per hop behavior. Because of the scalability of th earcitecture a good way to build a scalable network is to apply Bandwidth Broker managed DiffServ domains.

We consider two functions as as the inter-domain functions of the BB. Diffusing of the availability information towards the BBs of the adjacent DiffServ domain. The diffusion is important because the BB, and the customers of the domain have only information regarding to the available resources of the own domain. The diffusion is the way to get information about the available resources of the other domains. After receiving the availability information from the adjacent BB, the customer, or the BB can send a request to the BB about the resources he wants to reserve. If the request can be satisfied, the BB will get the resources from the adjacent domain. Consequently the negotiation is duable using this two method.

In the following section we describe Telia's DiffServ based BB managed architecture in few words, and after that present our work, which issued in a realized ProFis architecture with inter-domain communication functions.

4.2. The ProFIS architecture

The ProFIS architecture is a system specification for providing end-to-end QoS over IP based networks. This concept uses the idea of the Bandwidth Broker, which means that the network domains are treated in a centralised way. The job of the BB is to handle the resources of a specified domain. The BB must determine whether the received bandwidth demands can be admitted to the network or not, and using this information it have to configure the border routers as well (intra-domain management). This approach can be prosperous, because equipment, which has knowledge about the whole domain, may handle the resources effectively. In the next part of this paper, we are going to describe an algorithm, which realises an inter-domain communication for the ProFIS architecture. The method is designed for especially this architecture, but we suppose that it contains useful parts for all the Bandwidth Broker architectures.

5. Inter-domain communication

First of all we have to mention, that this communication method is designed for handling aggregated demands. We presume, that the reservations are not made by edge users, but network providers, so they arrive periodically and not at a random time. This is a common case of reservations in backbone networks. These reservations have high bandwidth demands, and low bandwidth fluctuation. Considering this we can say that the state of the whole system scarcely differs from the previous state. We call this state of the system as constant state hereafter. The condition of the subsistence of the constant state is that the demands of the users are considerably constant, they send demand specifications periodically and there is no configuration change in the network.

At first, we will define a propagation algorithm. The goal of this is to provide the edge domains of the network, with proper information about the other available edge domains. We call this information as availability information (AI). One AI refers to one edge domain. This contents the name of the edge domain, the amount of the bandwidth, the average delay, the maximal delay, the delay, and the loss ratio. The ProFIS concept defines propagation steps for the event, when the constant state is set and some availability parameter, i.e. loss, delay, changes in the network. We made this theory complete by defining a method for the case when the configuration changes in the network i.e. we connect another transit domain to the network with another BB. The special of case of this, when the whole system stands up. We implemented a method to make the setting up of the system fast.

After this we are going to define a method, to reach and hold up the constant state of the system. The main idea is that the users send out demand specifications periodically at the edges of the network, and we generate automatic demand specifications in the internal network by aggregating the demands.

5.1. Availability Information (AI) propagation

5.1.1. AI propagation in a constant state

The goal of this process is to provide each BB with the correct AIs. We expect that by the end of this propagation all the BBs will possess exactly one AI for each available edge domain in each service class and it will be the best. At this point we chose from the AIs in the basis of the bandwidth and delay parameter.

Considering that the system is in a constant state, we presume that each BB has one AI for each edge domain and all the BBs are stood up, so we only have to deal with the case when the some of the network parameters change. (This change is reported by a measurement system.) In this case the BB calculates the AIs again and spreads them to the other BBs. Let's see what happens when an AI arrives to a BB:

We have to decide whether to store the received AI or not. We compare this AI with the stored AI in the same service class.

1. If the AI comes from the same BB as the stored we have to store the new in every case even if it worse. This can happen if somewhere the delay grows in the network.

2. If the AI comes from different BB we store the AI, which has greater bandwidth. If the bandwidths are equal we chose the one, which is favorable in terms of delay. If the delays are equal we chose the stored.

According to the two points above, the BB is able to determine which AI is the better, so at the end of the full process there will be only one AI for each edge domain in each service class. (We can store the AIs in a database from which the users can download the list of the available domains)

If we stored the received AI we must propagate it to all BBs because we made an AI change. The propagation happens in the following steps:

1. Calculating delay parameter for the given route. We gain delay data from the database of a measurement system connected to the BB. The system is able to measure delay between two interfaces in the domain. We use the link, which leads to the BB, which sent the AI as the ingress link, and the link, which goes to the next BB in the given route as egress link. We add the delay for this route to the delay parameter we received in the AI.
2. We send this AI to the specified BB.

If these steps are done for each AI, all the BBs will possess exactly one AI for each edge domain in each service class and will be informed about the network parameter changes. It is important to mention that there can be loops in the network, and with a spreading method like this there, can be infinite loops in the propagation. Storing only one AI for each edge domain, which is the best, solves this problem here. If we received the best there will be no more AI change so the propagation stops.

5.1.2. AI propagation for the standing up

Now we are going to complete the method above with different propagation steps to make the system to handle network configuration change. We expect from this process all the BBs will have the proper AI even there is a configuration change in the network.

For doing this we define a special AI called NewAI and we make difference between the two AIs in the header field. The BB sends out NewAIs when it is standing up. The BB stands up in the following way:

At first it sends out the local AIs to the other BBs. We call an AI as local AI if the edge domain is directly connected to the domain, which is managed by the BB. The BB sends out the first AI as a NewAI, and sends out the other AIs as simple AI. By this step we reached that all the BBs will have information about the newly connected edge domains.

Finally the BB must gain the information about all the available edge domains, which are connected to the network. For solving this problem we use the fact that the other BBs are in the constant state so each BB possess exactly one AI for each edge domain. Now the BB is going to query the whole AI database from the neighboring BB. For this aim we use the special treatment of the NewSD. This method hardly differs from handling an AI. If the BB propagates the NewAI to the BB from which the NewAI is received than it will send the whole AI database to it. If the BB propagates to another direction, than makes a simple AI from the NewAI. Treating a NewAI is the following:

1. The BB determines whether the received AI is better then the stored.
2. The BB calculates the correct delay parameters and sends out a simple AI to all the BBs instead of the sender of the NewAI. If the propagation goes to the sender than the BB sends out the whole AI database to that BB.

Now the newly switched on BB will possess all the information about the available domains.

Using the idea of the NewAIs we can able to handle the problem of network errors, and network configuration changes. We can do it in using the "soft-state" principle. We provide the AIs with a validity time. The AIs have to be refreshed periodically else the referring edge domains are not available in the network.

Since the BBs store only one AI for each edge domain this communication method cannot realizes load balancing in the network, because the users are able to make reservations only for the route, which is evolved during the propagation. If a link is full on a route between two edge domains, the system will not be able to satisfy the demands even if there are free resources along other routes.

5.2. DS processing

This process realizes the bandwidth reservation mechanism. Let's see in a few steps how it works. We consider one cycle for one term to send the demand specifications.

1. The users send out the demand specifications from the edge domains. (At the beginning of the cycle)
2. Aggregating the DSs according to the destination information. Archiving the demands and sending out aggregated DSs automatically to the appropriate domains. (In the middle of the cycle)
3. On the basis of all the received demand specifications, sending the demands to the BB's intra-domain part, which will decide whether the demand can be satisfied or not, and configures the border routers. (At the end of the cycle)

This method handles the BBs independently, so there is no time synchronization between them. This means that different BBs achieve these points independently at a random time. This way can happen that a BB configures its domain just before it receives a demand specification. This demand can be satisfied only in the next term. There can be an unfortunate case, when an event like this happens along a whole route. Satisfying this kind of demand can suffer a huge delay. The principle of the ProFIS architecture, that the state of the system hardly differs from the previous state, covers this problem. We can say that in a constant state, it does not matter when the demands are satisfied, because there is a demand specification like that from previous states of the system. This way can happen that the user feels that his demand is served but it can occur that this is the affect of a DS, which has been sent few terms ago.

We made the step No. 3 in a simple way, that we give the reservations to the inter-domain communication in the sequence the arrived to the inter-domain module. This can cause that some demands can be fully achieved, but some demands cannot be satisfied because there are no free resources. The system informs the user about the unsuccessful reservation. There can be other solutions to make the sequence just by using game theoretical considerations.

6. Conclusion

As a result of this work we have given a suggestion how to realize End-to-End QoS over DiffServ capable network using Bandwidth Broker. It is important to mention that the introduced solution is only one of the several possibilities. The main advantages are that the number of the messages is low (each edge domain sends only one DS in any term in a constant state) and the system converges to a constant state rapidly. The system is scalable it consumes the network resources more efficient. Disadvantages are the sensitivity to the forecasts, which are not correct, other problem is that the DS message propagation time can be too long if the processing time points are quite different in each BB.

Finally we notice that the inter domain protocol is recommended to be a standardized communication protocol. It is important because there can be several kind of implementations of these methods. We suggest using Extended Markup Language[7] (XML) [7] for this purpose. XML is a standardized language, it is easy to understand by human reading, and it is compatible with the World Wide Web. The legibility for humans is important because some parts of the DS processing cannot be automatized. For example some decisions that are made by the machine are not overlaps with the financial considerations. For this reason, an interface has to be implemented for human interaction.

References

[1] Ping P. Pan, Ellen L. Hahne, Henning G. Schulzerinne "BGRP: Sink-Tree-Based Aggregation for Inter-Domain Reservations"
[2] S. Blake et al., „An Architecture for Differentiated Service", IETF RFC, rfc2475.txt
[3] Zsuzsanna Ladányi, András Gulyás, István Pataki, Gábor Balogh, László Nagy, Péter Füzesi, „Implementing End-to-End QoS over DiffServ Networks", HSNLAB Workshop 2001 spring, poster
[4] Integrated Services in the Internet Architecture: an Overview. R. Braden, D. Clark, S. Shenker. June 1994, IETF RFC, rfc1633.txt
[5] R. Braden, Ed., L. Zhang, S. Berson, S. Herzog, S. Jamin." Resource ReSerVation Protocol (RSVP) Version 1 Functional Specification." RFC 2205, IETF 1997, szeptember
[6] A Two-bit Differentiated Services Architecture for the Internet. K. Nichols, V. Jacobson, L. Zhang. July 1999.
[7] Extensible Markup Language (XML) 1.0 (2nd Edition) W3C Recommendation 6 October 2000, http://www.w3.org/TR/2000/REC-xml-20001006

ITU-News

The United Nations will convene the first-ever **World Summit on the Information Society (WSIS) in Geneva** from 10 to 12 December of this year. Like ITU TELECOM World 2003, the Summit is being organized by the International Telecommunication Union (ITU), the UN specialized agency for telecommunications. The Summit is seeking ways to extend the benefits of the information and communication technology revolution to all of humanity, by:

– connecting developing countries with technology;
– boosting Internet security, fighting cybercrime and cyberterrorism;
– energizing the ICT sector with millions of new users / buyers;
– upgrading health / social services through cyber solutions;
– guaranteeing the free exchange of information;
– examining the need for global policies for cyberspace.

ITU in collaboration with ISO and ETSI is organized a workshop offering the automotive and telecommunication industries an opportunity to exchange ideas on the future of communication technologies in motor vehicles. This workshop will be a first attempt to bring all interested parties together from around the world to forge standards that will expanded markets, promoted-innovation and ensured that in-car communication technology moves forward at a rapid pace. Examples of the types of technology to be examined include, speech recognition, e-calls – that generate automatic calls to the emergency services in the case of an accident, predictive technology to prevent accidents and location finding technology. The goal was to create a better understanding between the two industry sectors, and to combine efforts and skills to create standards for mutual benefit.

Summaries • of the papers published in this issue

A PLAYFUL STRATEGY

John von Neumann was dealing with several aspects of applied mathematics and economics. Models of game theory to be found along with border line of economics, mathematics and management can well be used with strategic decisions. In the life of economics, particularly in the field of telecommunications and information technology, a lot of cases can be described and analyzed with the use of game theory. However, this theory has its limitations as well.

PROFIT OPTIMISATION USING BUSINESS RISK ANALYSIS AND GAME THEORY

There is an increasing need on the side of the companies to use also mathematical tools for modelling risks, supporting hereby their strategic and business decisions, creating connection between decisions and their expected results. On those areas, where there is competition, and so it is in the telecommunication sector as well, it is extremely important to satisfy customer demands on a high level. Besides, the main aim of all service providers is to maximise the available profit. All these are intended to reach in such a market environment, where competitors are also present, and their aim is also to increase their own profit.

OPTICAL NETWORKS AND NETWORK STRATEGIES

The revolutionary development in telecommunication and information technology is based on several roots. The critical mass of users (sometimes above one billion people) is often mentioned concerning mobile phones and computers, services, applications. Other important issue might be the fierce competition on the global ICT market. A very important factor in this competition to come first to the ICT marketplace with relevant new products and services. There is no ultimate winner in this competition for long run. Day by day the competition starts again and technical novelties push the market into new direction.

OPTICAL BURST AND PACKET SWITCHING

The increasing volume and ratio of packet switched traffic poses new challenges on optical transmission technology applied in telecommunications networks. Retaining efficient operation of networks under the changing traffic load calls for development of new solutions. The paper focuses on two research directions of increasing importance: Optical Burst Switching (OBS) and Optical Packet Switching (OPS). The basic principles behind the two emerging technologies are introduced and an attempt is made to determine the possible role of these technologies in networking scenarios. Both theoretical and practical issues concerning the implementation are discussed and some of the proposed solutions to these issues are also introduced.

DESIGN OF WIDE BAND DISTRIBUTED AMPLIFIERS

Due to the continuos demand for increased transmission speed, parallel connection of several WDM channels would be necessary to build up a high-speed lightpath between two end nodes of an all-optical network. In this case the electrical circuits of the applied optical transmitters and receivers must have extraordinary bandwidth, and even a so simple function like amplification can be critical. As the distributed amplifier (DA) has the highest bandwidth among the known amplifier structures, it can be the ideal choice. The paper presents a new design method for DAs comprises interstage transmission lines (TLs) between the active devices.

BROADBAND RAMAN AMPLIFIERS...

In order to increase the transmitted amount of data in optical fiber drastically more channels are needed. This causes the broadening of the bandwidth used. Therefore one should apply optical amplifiers which can amplify in a relatively wide region in which amplification is approximately at. Beside the well-known Erbium Doped Fiber Amplifier (EDFA), there exist such amplifiers based on different phenomena and providing better properties in certain areas. They are able to fulfill the requirements of the wavelength division multiplexed (WDM) systems regarding wide amplification region and atness. Among others, Raman amplifiers which use Raman scattering effect are noteworthy since their technological background is solid and their fields of application are the subject of recent papers frequently.

AURALISATION AS A TECHNICAL TOOL

Noise of electrical signals and other unwanted, annoying or unpleasant sound signals are widely known types of disturbing noise. From point of view of physical definition and measuring technology sound and noise are identical quantities. Its the momentary human judgement that makes difference between them which depends on the actual expectations of the audience. This article presents a review of measuring techniques of acoustical noises. We will point out that the dBA measure used in measuring practice generally does not follow the results of sensory acoustical judgement. The advanced approach of quality has brought about an interesting turn-round in the decade long hang of acoustical measuring technology.

ANALYSIS OF METHODS FOR EDGE DETECTION

The paper describes the application of algorithms for edge detection in images.The principles and algorithms given below have been used in an application that was developed at Brno University of Technology and has been programmed in the Microsoft Visual C++ environment.In the development,the Win32 API and MFC libraries were made use of.In the application,any of the eleven implemented algorithms can be used for edge detection.

TEXTUAL FUNCTIONS – NEW PROSPECTS IN E-ADMINISTRATION

It was 12 years ago that I discovered that any textual knowledge such as laws can be expressed in a full, exact and solid manner by an "IF – THEN" based formalism. Some obvious advantages of this realization became clear soon. When, for instance, this formalization is applied on legal texts, a defect or gap in the law, discrepancies and inaccuracies can be discovered clearly. There is an even more exciting field of application: using the formalized version of a law, and a suitable system framework, the behavior and mechanism of the law can be examined by computer simulations.

SIMPLE INTER-DOMAIN PROPAGATION ALGORITHM FOR THE PROFIS ARCHITECTURE

This paper deals with providing Quality of Service (QoS) over IP based networks. We are going to give a brief survey about this topic, and present our work at this area. There are many solutions of the problem, but the standardization of the methods is not finished yet. At the moment there are two kinds of approaches of the reservation problem. The distributed method handles the network nodes independently, and get the nodes making their own admittance decisions along the reservation path. The centralized way – we discuss in details –, which collects the network nodes into domains, and handles them using a network manager. Generally there are two significant parts of the network management: intra domain, and inter-domain. This article focuses on making reservations over several domains, which is the part of the inter-domain functions.

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.

Egy fő-rendszerintegrátorral minden egyszerűbb

A Matáv egyedüli infokommunikációs fővállalkozóként fogja össze a legjobb informatikai partnereket személyre szabott üzleti megoldásaihoz.

Bízza ránk a csapatépítést!
A Matáv tagvállalataival – Axelero, BCN, Cardnet, CompArgo, MatávCom, SafeCom, X-Byte – és partnereivel kidolgozott komplex üzleti megoldásai választ adnak Önnek mind informatikai, mind távközlési igényeire, az összehangolt elemek pedig hatékonyan támogatják cége üzletmenetét.

ALBACOMP

hp invent

IBM

Microsoft

www.matav.hu

matáv

a szavakon túl