híradástechnika

VOLUME LVI.

2001/12

Selected papers

Mobile future

IP network development

Theoretical results

Scientific Association for Infocommunications

Contents

Preface to the English language issue

Young researchers are happy to go abroad for exchange of experiences, learning or even working. This absolutely bring advantages since already in the Middle Ages past-masters sent journeymen for six months or even one year to travel around. In the past ten years we were seeing a revival of this tradition, even its reinforcement. However, the receiving businesses are far from being so open than workshops were 400 – 500 years ago. They are not ready to reveal professional secrets since there is world-wide business competition. This means that exchange of experiences is possible only on a reciprocity basis. That is why when a young professional can present itself with substantial development results behind him doors open up to him and he can participate at professional conferences. However presentation requires not only words but also publications in a language that is understood in the receiving country. Fortunately in the area of telecommunications and IT English is the conventional standard language of international exchange of experiences.

That is why we publish in our English language issue the most valuable professional results of the last half year, i.e. the highest standard articles published in Híradástechnika. In the latest English language issues we could fully implement this principle. However the last period encourages us to publish the works of some foreign authors too. Of course, this will not be to the detriment of our successful authors, but we wish to present in the English language issue the current problems of our days.

More than one year ago the issue of next generation networks (NGN) came to the focus. It is natural that a new generation always arises, but it is not usual that these issues are so much published before the new technology and the new series of equipment are developed. The recession in 2001 made introduction of a new generation even more timely. We are beginning to see the outlines of the changes going to happen in the next few years. All of the different trends emphasize that only such investments can bring profit and payback which can offer broadband both to homes and offices. Both areas are interested in teleworking and shared working, and in homes substantial profit can be generated with entertainment and video. Offices with multiple workstations, primary financial institutions and authorities, are expected to pay more for shared availability of data and information.

In the second half of 2001 various conferences addressed this issue so we decided to publish two determinant forward-looking presentations in the original language. In particular we paid attention to and selected from those which have not been published in articles, but the proceedings included only the projected slides.

The second block addresses the creation of IP networks and their expected development trends. This includes articles from two Hungarian and one foreign authors. Their objective was to create a network that is easy to design with the new methods and satisfies the quality requirements. This is a possible trend with which the upturn of telecommunications and IT may be promoted. Finally we included some theoretical considerations in the third group which may be still not be directly applicable, but their results are expected to have an impact on mobile and broadband transmission.

Of course, we do not want our readers to believe that they are reading the publications of a reliable oracle. The very risky task of making prophecies cannot be committed by a journal. However we can certainly publish opinions that do not contain future dates and quantities. While in a high-standard professional article engineers do not like generalities and are happy to characterize everything with metrics, the views of experienced professionals can be useful for learning trends.


Global Communication Global Communication Global Communication

Mobile Networking in the 21st Century

Josef F. Huber

Siemens AG Vice Chairman

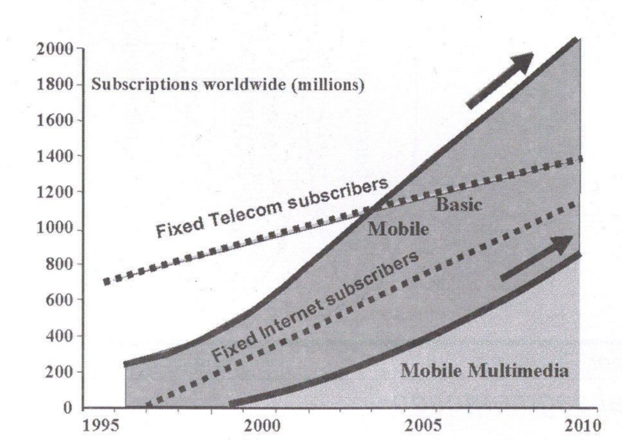
UMTS Forum

Over the last twenty years, the telecommunication's focus changed considerably: from traditional wireline telephony oriented services and infrastructures to data, from homogeneous to heterogeneous networks, from non-intelligent devices to smart handhelds, PDAs and Mobil Computers. Internet with its world-wide web, GSM and all mobile wireless communications in the local areas contributed to the dramatic changes in this field. The forecasts of both sectors' markets were underestimated every year.

The Universal Mobile Telecommunication System (UMTS) has been conceived as as concept to carry the personal communication user into the information society of the new century. Its upcoming information society is characterised by multimedia processing and delivery of information, video and voice, fax, and data. The provision of such broadband multimedia applications in a global mobile radio network is understood as the main objective UMTS a 3rd generation system.

1. Introduction

For the last 15 to 20 years, the mobile market was always underestimated. The forecast figures were moved up every year and ended up every year as being too low. A similar situation appeared with the Internet. It succeeded beyond the expectations of its designers. In very few years from now, the Internet will run out off its present address capacity based upon the protocol version IPv4. Nowadays, the mobile handhelds achieve the Internet capability even under restricted functionality; it should come with no surprise to learn that the number of mobile devices with Internet capability will outnumber the Internet capable Personal Computers. All available predictions signalise such scenarios, see Fig. 1.

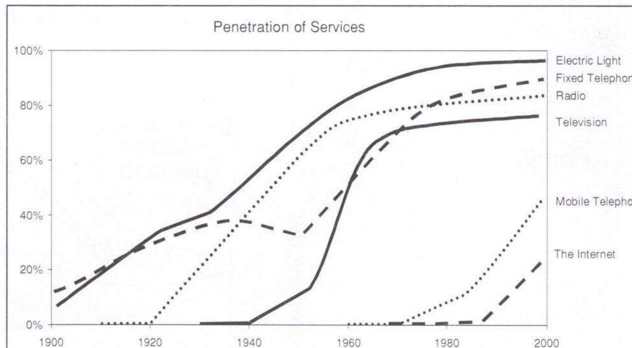


Source: UMTS Forum

Figure 1. Mobile Terminals become Internet enabled

The diagram (Fig. 1) indicates that the Internet will reach a huge mass-market size very soon and this will become the base for the new businesses with new services on the mobile side.

Both mobile and Internet penetrations are now on the exponential portion of the logistic demand curve illustrated in Figure 2. But this does not necessarily mean that the "mobile Internet", a frequently used description of 3G services, will automatically experience the same exponential growth rate. Nevertheless, most forecasts of 3G subscribers and revenues simply assume a dramatic uptake of mobile Internet services. This is more a statement of faith than a reflection of reality. 3G service revenues will only materialise if 3G networks deliver compelling services that satisfy subscriber needs. Adding mobility to the Internet is a necessary but not a sufficient condition for success.



Source: US Bureau of Census, 1996

Figure 2: World-wide Penetration of Services

Not everybody agrees with this proposition. There is a belief within the industry that the introduction of the

Mobile Internet will accelerate the already rapid growth of both Mobile and Internet services.

The Internet was in existence for about three decades before it entered the current phase of mass-market acceptance. The trigger was the introduction of interfaces such as Mosaic and the World Wide Web that transformed the Internet into a user-friendly environment. The PC was merely a cheap alternative to mini-computers until the introduction of mass-market applications such as spreadsheets and word processing. Mobile services were a limited alternative to fixed telecommunication services until national coverage could be guaranteed, and has only become a true mass-market commodity with the introduction of global roaming and prepaid services. In this decade, however, mobile systems will merge with the Internet by taking up IP protocols into the radio and switching part.

The success of the Mobile Internet will not just come from the mere combination of two existing successful phenomena – mobility and the web. The real success of the Mobile Internet will result from the creation of new service capabilities that genuinely fulfil a market need. Creating and meeting market demand requires economies of scale to be present. The ability to benefit from economies of scale is one of the strongest market drivers for mobile Internet services.

2. A "New Focus" in the Mobile Business

The Internet is transitioning from inexpensive medium for advertising, marketing, and customer support to a common platform for transactions and business applications. At the same time, technological and commercial developments are melting together information, communications, commerce and enterta-inment into one large, consolidated industry. Part of the reason for this evolution is because more consumers are accessing the Internet using multiple devices and over multiple communications networks. They are also changing their behaviour and consum-ption patterns.

The wireless access to the Internet is going to bring new customers with new requirements and will therefore change the development of the Internet for several reasons:

- Wireless allows service providers and Internet businesses to increase their mobile culture and total service consumption;
- The mobility and immediacy offered by wireless allows Internet content delivery and commerce to be non-location-specific;
- The person-specific nature of wireless allows companies to develop customer profiles that enable

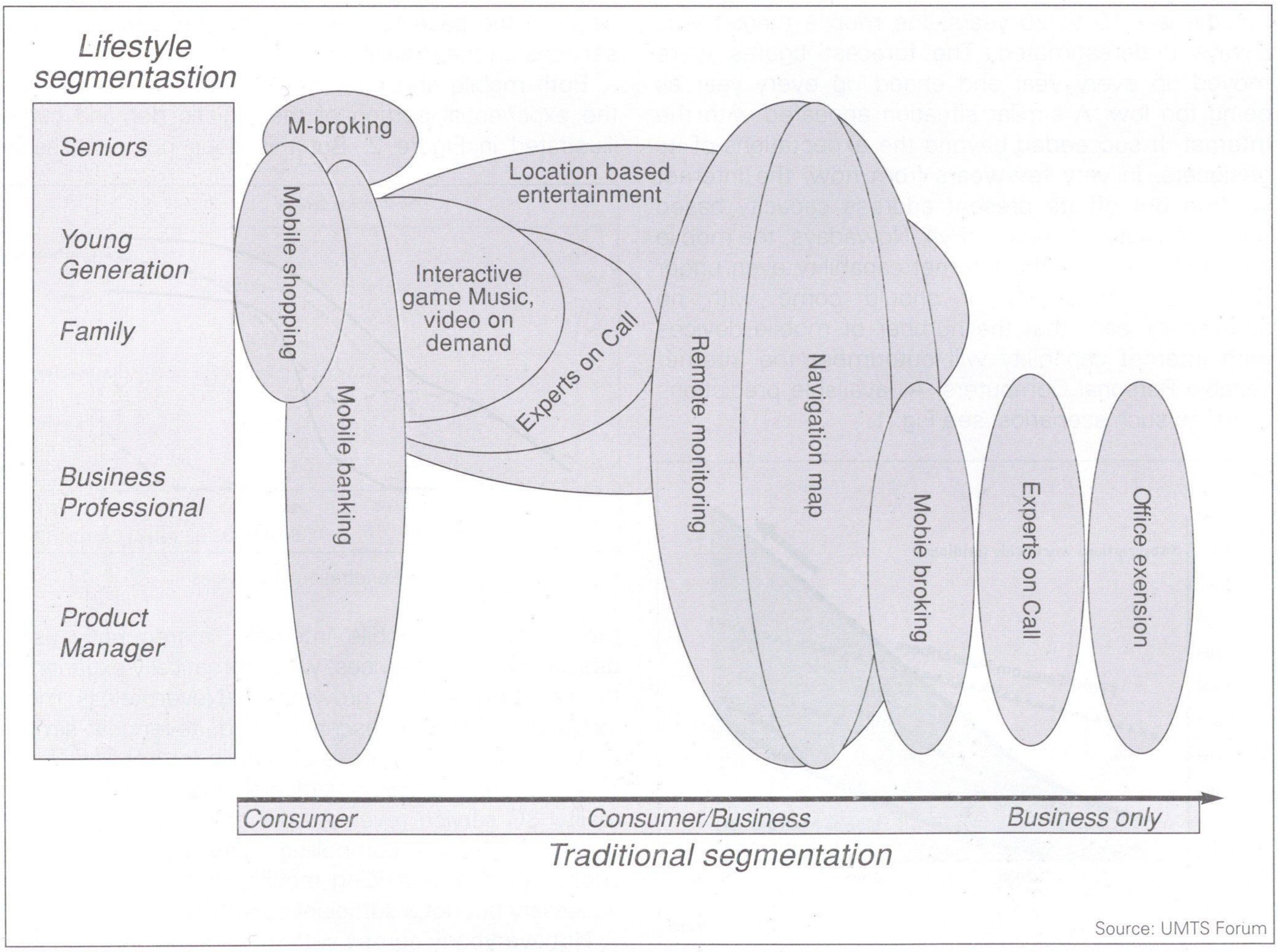


Figure 3: Traditional and lifestyle segmentation

them to narrowcast and distribute better value-added information to customers, and lastly

- Location-based facilities and services provide another tier of customer knowledge that allows Internet businesses to deliver "context" specific services that also improve customer value.

Many services, although they may start with 2G will become more affordable using 3G. Services, which already exist, will be greatly improved with interactivity and mobile multimedia with customer segmentation based on lifestyle management.

While work-oriented applications are believed to drive the 3G market in Europe and in the US, in Japan the high value placed on increasingly more sophisticated consumer-oriented equipment could have a substantial impact. The demand for increased personal productivity is also of importance.

Mobile commerce is not explained as a service in itself. It covers too many services like m-broking, m-shopping, m-auctioning, m-banking, m-cash etc. Typical services are described in the UMTS Forum Market Study [1].

Converging the Telecom and Internet Model

The convergence of mobile networks with the Internet leads to a new model for business considerations, for operators and manufacturers, for standardisation and development. New elements will be involved impacting each other. In addition, the delivery of multimedia services involves a number of new players and activities. For example, the market analysis from Analysys/Intercai [2] highlights the Multimedia Service Provider (MSP) as a new key player in the multimedia value chain. MSPs purchase multimedia information (content) from third party suppliers such as TV programme providers and publishers. Then they format the content to what they believe is appropriate for their customers ("filtering") and store it on a multimedia server or cache server.

Figure 4 indicates the changes in the business values of the single elements as meanwhile seen from a number of studies and market forecasts. It gives a clear signal to the mobile industry of the need to change their positioning with regard to applications, services, and content when planning their future strategic orientation. It includes the possible roles as VAS Providers, Service Providers (ISPs, ASPs and Portal Operators) and Content Providers [3].

In the light of the new business chain, the issue is to consider whether to simply provide a wireless Internet Protocol (IP) pipe to a service offering hosted elsewhere on the Internet, e. g. at a Portal like Yahoo, AOL, T-Online, Excite or Infoseek, or to go for a harmonised end-to-end solution where the UMTS Operator is involved in the applications level. The wireless IP pipe business using tunnelling will tend towards a commodity bit-shifting operation, where cost, coverage and data rate are the only competitive dimensions. Alternatively, by developing and pre-selecting useful Internet-based mobility services with competitive tariffs, the UMTS Operator will become a competent player in this field.

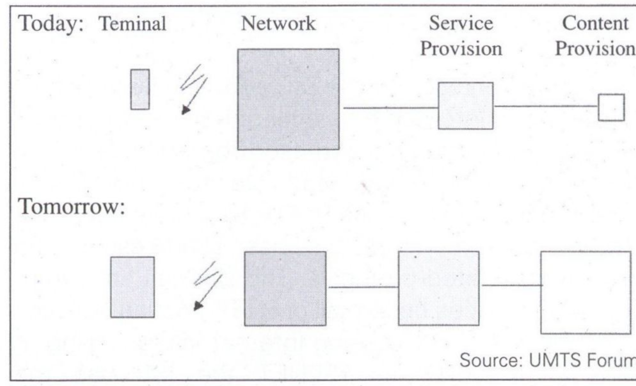


Figure 4: UMTS: Value Chain, Size of Business Changing

At issue is the location of the *subscriber profile records,* which reflect the personalised service choices of the end-user: message filtering options, choice of mobile information and type of mobile device, correlated with name, billing address, mobile phone number and e-mail address. This store of data, three to five years from now, will permit additional returns through selectively targeted mobile e-commerce and advertising. UMTS Operators have three distinct separate or combined possibilities:

Firstly, they charge the subscriber on a *time dependent call basis.* This makes it possible to get a revenue stream via a small additional charge for mobile Internet services and means a business return is possible well before mobile commerce and advertising become feasible. Added to traffic revenue, the Operator is in a commanding position to capture or selectively share this revenue with value chain partners on its own terms.

Secondly, the Operator provides IP-packet transport (e. g. GPRS-based); all necessary to integrate Internet services with IN, voice, data and fax services and allowing *volume based charging* in addition to time.

Thirdly, in future, the Operator will know the subscriber's location using emerging cellular positioning technologies. Positioning adds end-user value through information customisation, e. g. details of the nearest restaurant or automatic conversion of e-mail to speech for a driver of a moving car. In future, location information will enormously increase the revenues combining it with content. This will allow *applications related charging.*

Wireless Internet will become one of the media channels for content providers, and wireless network Operators will join. The WAP-Portal offerings of GSM Operators and i-Mode offerings of NTT DoCoMo are examples of a new strategy. The mobile Portals are unique because they are a solution in which Operators and service providers can manage content and integrate with communication and transactions.

The Internet Service Provider (ISP) Function

The simple definition of the ISP is definitely providing access to the Internet. The level of service, the set of functions supporting the end-user from the ISP varies. The client-server model leads to the *"user@host"* account (Point of Presence PoP) which means that the ISP has to care about the user's addresses and applications-related protocols. The ISP can start from tunnelling services up to mail or HTTP hosting services etc. In all cases of accessing Internet via tunnelling or HTTP, E-mail, FTP or TELNET the ISP has the responsibility for the end-user's account. As a consequence the "All-IP" UMTS network, which will be specified in the worldwide 3rd Generation Partnership Project (3GPP) as the so-called 'Release 5', takes the UMTS Operator into an ISP position. It is combined with extended functionality dealing with the roaming user: "Mobile ISP". The additional complexity comes from inter-Operator roaming.

Portal and Content

Providing the Portal does not simply creating the content. Content feeds are likely to come from existing content providers or other Portals, but with subscriber data being captured by the Operator. It naturally follows that the first applications should focus on building the subscriber base and increasing airtime, namely messaging applications. Next should come applications which drive subscriber profile data capture, such as personalised subscription-based content push and wireless personal information synchronisation. Only after building the profile database can mobile e-commerce and advertising be successful.

For Operators with strong brands and a desire to reduce churn and increase their share of the customer value chain are offering their own Mobile Internet services to users on a mobile Portal platform integrated with billing, customer care and positioning systems. This is known as the 'Mobile Portal' approach, which exists already in a number of GSM networks [1].

3. Convergence of Standards

Extending the Scope of Standardisation

Global standardisation is of fundamental importance for mobile cellular systems and essential for worldwide roaming. This is a key issue for IMT-2000/UMTS. The framework standards for wireline and wireless access, for national and international telecoms infrastructures have been developed and agreed on the ITU level (the Framework Standards) and on the regional/national level (the Specifications).

With second-generation cellular networks, wireless access has already introduced a new set of standards and protocols that add a layer of complexity to application solutions that are not necessarily compatible with the Internet world. This has been seen with the introduction of Voice/Video over IP, WAP, i-mode and similar IP services to mobile networks. In addition, for data processing, data communications, TV and Multimedia ISO, IEC and ITC standards were developed and continue to be valid for all kinds of data and multimedia applications and services.

The following diagrams show the main difference in standardisation between point-to-point traditional services (e.g. voice) and content-based services in the context of the Extended Vision. Fig. 6 shows the standards associated with the UMTS environment linked with the ITU-based fixed network environment.

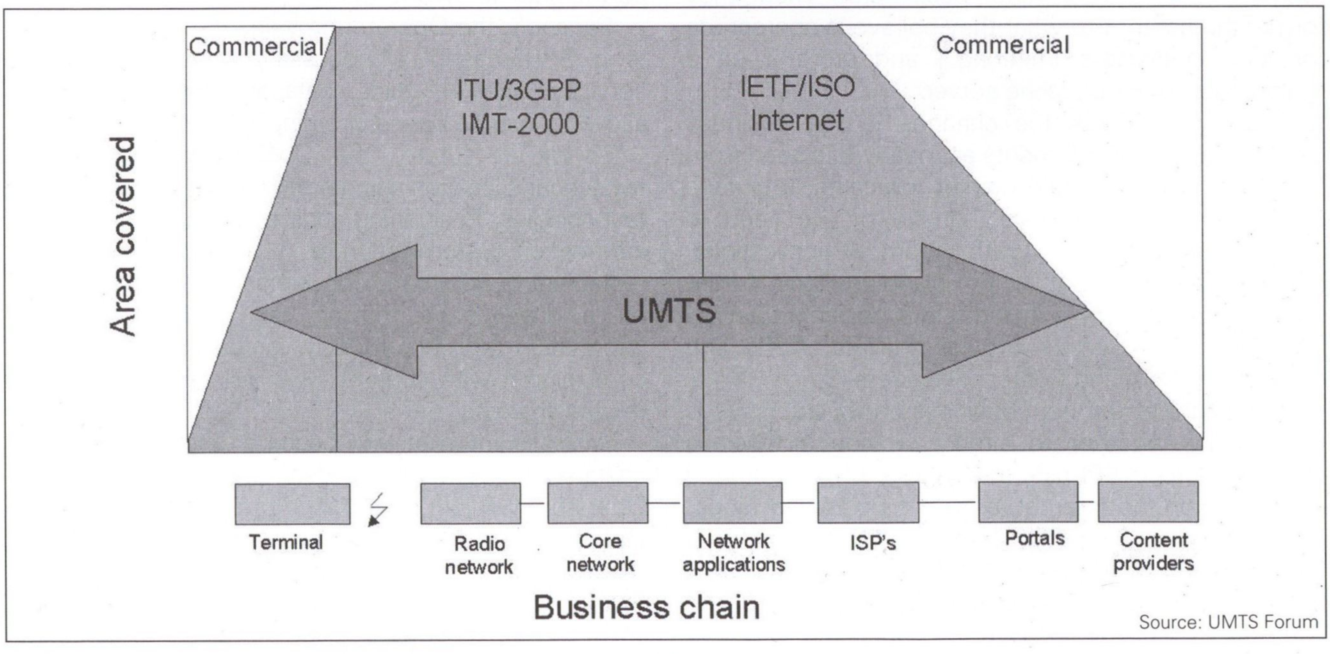


Figure 5: Positioning UMTS – the new Value Chain for Content-based Services

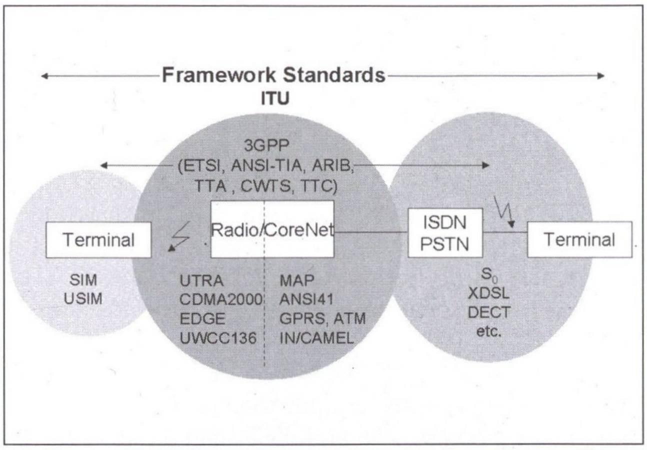


Figure 6: UMTS Standardisation Scenario in the Framework of IMT-2000

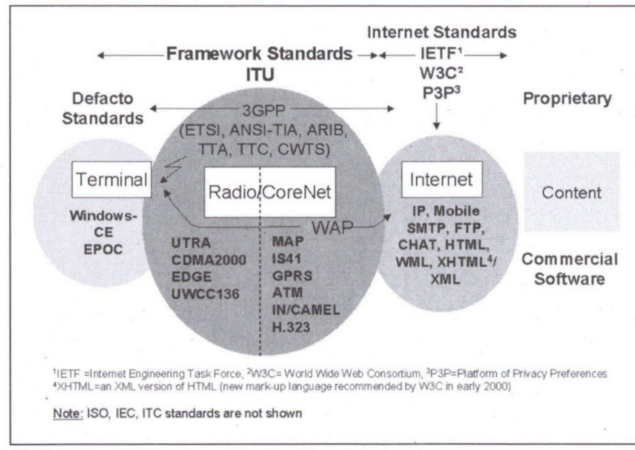


Figure 7. UMTS Standardisation Scenario UMTS/IMT-2000 with Internet

A series of network services are specified entirely in the ITU framework including service interworking, addressing, signalling and QoS standards. The typical configuration is the physical and logical connection from terminal to terminal or person to person. Voice is the dominating service, but data services are set up in the same way in these connection-oriented services. Billing of these services is generally based on the duration of a call.

For UMTS these standards have to be supplemented for those services that will also be offered on the wireline network, such as bitrate harmonisation, circuit switched service interworking, etc. As an example, UMTS provides 384kb/s and 2Mb/s connections, whereas the fixed ISDN network only offers 64kb/s, 128kb/s in a limited way. Also, the harmonisation of terminal interworking characteristics between wireless and wireline terminals may be a standardisation issue for services such as video telephony, VoIP, fax and other telematic services. QoS, security and billing issues also have an impact on standardisation.

Fig. 7 shows the scenario for multimedia services in the scope of the Extended Vision: UMTS is linked with the Internet, the portal and content provisioning. The typical configuration, physically, is from terminal to computer or client to server. This is a "many to one" relationship, because there may be many physical terminals logically connected to one computer. Bitrate conversion is a must. "Connectionless Services" exist on a "user at host" basis. As UMTS will combine the ITU-related standards with the IETF, W3C etc.-related ones, it will be faced with a number of new interworking issues. These relate to quality, security, mobility management, billing, etc. A main issue is the interworking on the protocol layers where "Out-of-band" and "In-band" control functions have to be aligned. The standardisation also has to specify impacts regarding addressing (ITU, IP), which is quite different in IPv4 and IPv6. The Internet transition from IPv4 to IPv6 is not only impacting the Internet itself, but also UMTS. Thus, this topic will play a significant role in the future standardisation within 3GPP. The harmonisation of the UMTS standards between Engineering Task Force (IETF)/3GPP and the Internet impacts the Mobile Multimedia Portal platform as a workable solution in an international mobile environment, especially for the roaming user.

Push E-Mail is another issue. Standards have to be specified based upon the existing IETF SMTP and Instant Message Service (IMS) protocol standard. Standardisation issues need to be discussed in the relevant bodies to guarantee end-user transparency and compatibility.

Mobility

Mobility, in all its many forms, is becoming the watchword of our society. Everything moves faster and faster. IP seems to be the end to end protocol of the future delivery of most services since it will exist in the wireline and wireless world, in office extension environments and in home networks.

The interest in mobile IP as a potential mobility solution for cellular networks leads to a requirement to extend the existing protocol of telecommunication networks. As a network-layer protocol, mobile IP is completely independent of the media over which it runs, i.e. it is independent of technology. This is in keeping with the design philosophy behind the Internet Protocol itself, which was designed to be independent of the underlying characteristics of the links over which it runs.

Mobile IPv6 as standardised by the IETF benefits from the integration of ongoing development of the Internet and cellular standards. It allows the user to keep their home address while roaming because they are always "ON".

IPv6 mobility determined by and optimised for mobile terminals will be one of the major features in cellular networks. It offers "Built-in" IPSec that provides all management applications security, it supports cellular and non-cellular access, and its flexibility allows sharing of resources to support a diversity of both wireline and wireless technologies. It has no impact on location registers since the information required to route packets is managed independently from the information used to locate and authenticate a UMTS user.

Each mobile terminal is always identified by its home address, regardless of its current point of attachment to the Internet. While situated away from its home, a mobile terminal is also associated with a care-of address, which provides information about the mobile node's current location. IPv6 packets addressed to a mobile terminal's home address are transparently routed to its care-of address. The protocol enables IPv6 terminals to cache the binding of a mobile terminal's home address with its care-of address, and to send any packets destined for the mobile terminal directly to it at this care-of address. The first option in providing mobility functionality over heterogeneous networks is an interworking solution. Within the UMTS system, classical MAP-based mobility functions are exploited, while on the IP side, mobile IP can be used. In this situation a complete interworking function has to be developed to allow the user overall mobility across both the mobile and the IP environment with 3GPP Release 5, the mobile IP will be used also in the radio access network.

Figure 8 illustrates a transparent IP mobility solution both within the UMTS system and in the IP environment, and will be implemented after the unique IETF/3GPP mobility support solution becomes available. In this case a transparent end-to-end mobility protocol is provided through the exploitation of integrated IP functions, i.e. mobile IP for the discrete macro-mobility, within and between Core Networks, and cellular IP for the continuous micro-mobility within the Access Network.

Security

Rapid advances in communication technology have accentuated the need for security in the Internet. The IP Security Protocol Working Group (IPSec) has developed mechanisms to protect client protocols of IP. A security protocol in the network layer is developed to provide cryptographic security services that will flexibly support combinations of authentication, integrity, access control, and confidentiality.

The protocol consists of three core components. The IP Authentication Header (AH) verifies the identity of a packet's sender and the authenticity of the packet's contents. The IP Encapsulating Security Payload (ESP) encrypts a packet before transmitting it, and may also encapsulate the original IP packet. It is independent of the cryptographic algorithm. The Internet Key Exchange (IKE) governs the transfer of security keys between senders and receivers. The preliminary goals of IPSec are to pursue host-to-host security, followed by subnet-to-subnet and host-to-subnet topologies. AH and ESP can be used with various authentication and encryption schemes, some of which are mandatory.

With the improved Wireless Access Protocol Release WAP2.0, UMTS services will be using XHTML-based Webaccess. With the improved WAP, the Transport Layer Security (TLS) Protocol will be linked with the Internet Security Layer (SSL), in order to guarantee end-to-end security. The linkage between TLS and SSL takes place at the WAP Gateway (see Fig. 8b) Protocol and cryptographic techniques have also been developed to support the key management requirements of the network layer security. The Internet Key Management Protocol (IKMP) will be specified as an application layer protocol that is independent of the lower layer security protocol.

IPSec gateways work only in tunnel mode, which means no part of the original packet is vulnerable to interception.

Some of the key IPSec features are:

- Signalling: integrity, authentication, anti-replay protection
- User traffic: integrity, authentication, confidentiality
- Visited network resources and traffic: access control, confidentiality
- No Foreign Agents
- IPSec protocols are integrated into IPv6 devices as standard
- Home Address Option eliminates network-ingress filter problems
- Route-optimisation functionality is integral

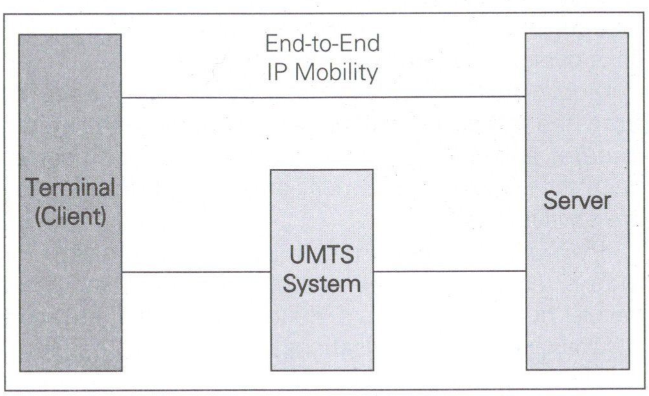


Figure 8a. Transparent IP Mobility

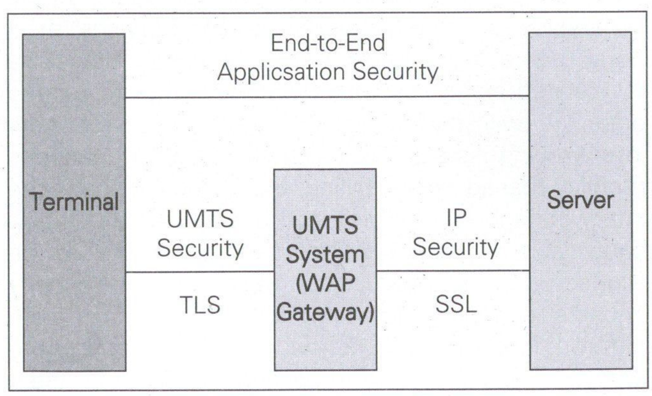


Figure 8b. Security Relationships

Similar to the mobility problem of providing complete and reliable end-to-end functionality across heterogeneous networks, security will follow the same approach.

Security relationships in a heterogeneous network environment can be very complicated, requiring both an interworked and a transparent end-to-end solution. In fact, dedicated mutual authentication and encryption facilities are used to check and secure the transmission segments in each specific environment, i. e. both the UMTS system and the IP segment. End-to-end transparent solutions are required to secure the application layer, via mechanisms such as digital signatures plus non-repudiation techniques. Non-repudiation can be achieved by generating a signature and combining it with some form of user authentication data (such as a PIN). When the user commits to a transaction, the signature will ensure that they cannot deny later that it took place. Of course, this kind of functionality is fundamental in any kind of mobile-commerce transaction.

To guarantee the overall required level of security, global co-ordination is necessary between the solutions adopted at the transport layer and the solutions adopted at other layers such as the applications layer.

The UMTS SIM Card (USIM)

The key element of the subscription to a UMTS service is the USIM. In addition to the well-known functions of the GSM SIM card, the USIM adds UMTS-related extensions and provides compatibility for UMTS-GSM roaming. The issue of electronic money, used for m-commerce, becomes more valuable in contrast to debit cards today. Considering Internet-based services means that the USIM card needs additional functions, which deal with "In-band" control mechanism, identification, security and applications. Other future items will be secure and personal interconnectability with other end-user devices that could be linked with the UMTS terminal, e.g. via Bluetooth.

Quality of Serve (QoS)

Guaranteed QoS is a standard item in today's telecom business, whereas the Internet only offers QoS on a best effort basis to the end-user. This is a concern, as it will be virtually impossible for the user to distinguish any observed delays as being due to the Internet, as opposed to the air interface.

On the other hand, in order for to be able to ensure QoS for the user, it will most likely that voice services will be circuit-switched for a longer period of time then expected. Although it is already decided, to apply the Session Initiation Protocol (SIP) for Voice over IP applications in UMTS Release 5, the standardisation on this topic is not entirely finalised.

The Portal Operator will have to be able to adjust to the varying (on demand) data rates requested by the user, depending on the application. Thus the Portal Operator needs to know the requested QoS, as well as the minimum acceptable QoS, to allocate the necessary resources. Portal Operators and dynamic Content Providers have the ability to deliver, intelligently, relevant information to users based on a series of characteristics and capabilities. Quality of Service (QoS) as requested by the user and as delivered by the Network Operator is one of those parameters on which this intelligence can operate.

4. UMTS – the 3G Multimedia Concept

UMTS – the Universal Mobile Telecommunication System is not a technology. It is a concept encompassing existing upgraded technologies combined with new technologies. This relates to the addition of wideband CDMA and TD-CDMA. On the core network side, it relates to CPRS, to the ISP and Portal function, where the additional features for mobility and personalisation will be added, but also other components will be affected, e.g. for providing security, address translation between the ITU E.164 scheme and the Internet addressing scheme.

The network roadmap of UMTS is shown in Fig. 9: starting with the first Release, called R99, UMTS will re-use core network elements from GSM, especially for the Mobility Management (Home Location Register, Visitor Location Register), for the Circuit switched transport (Mobile Service Switching Centres 3G-MSC), for the packet switched transport (GPRS Service Switching Node SGSN), for the Authentication (AC), Gateway elements like the G-MSC or the G GSN take over the interoperation tasks on the border to other network infrastructures. The Signalling Control Point (SCP) takes care of real-time billing, number translation and value-added services via Intelligent Network Support. For packet switched transport, the GPRS tunnelling protocol (GTP) is used to transport IP packets between terminal and ISP.

Subsequent UMTS Standard Releases will integrate stepwise functions related to IP protocols, Release 5 will introduce the IP Multimedia Services (IMS) based on IPv6 and the IP based UTRA and Core network. Location-based services will already be possible based on Release 99 as the standardisation provides position information of the user's location based on Cell Identification. The accuracy of the position information will improve with the ongoing standards developments.

On the Portal side, existing standards from IETF (Internet Engineering Task Force) and ISO will be used for the implementation of personalised Portals comprising also position information in order to offer location-based services like restaurant finder etc.

In Fig. 10 it is shown, that two radio parts are connected to the core network: the GSM part with the Base Transceiver Station (BTS) and Base Station Controller (BSC) and the UMTS part, called Universal

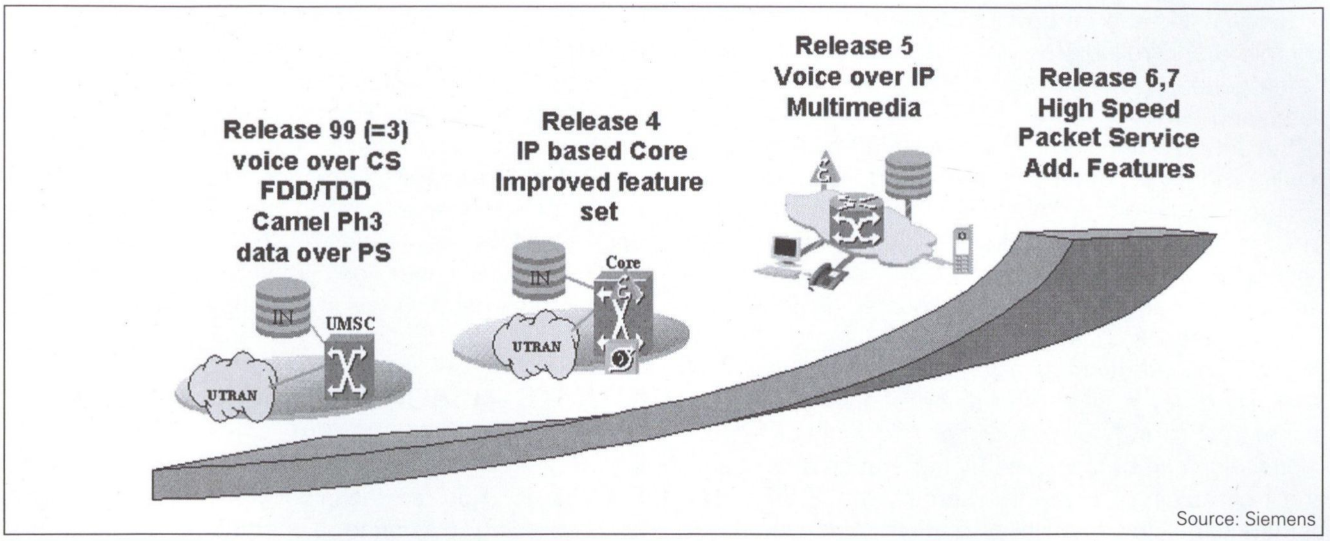


Figure 9. Network Roadmap: UMTS goes IP

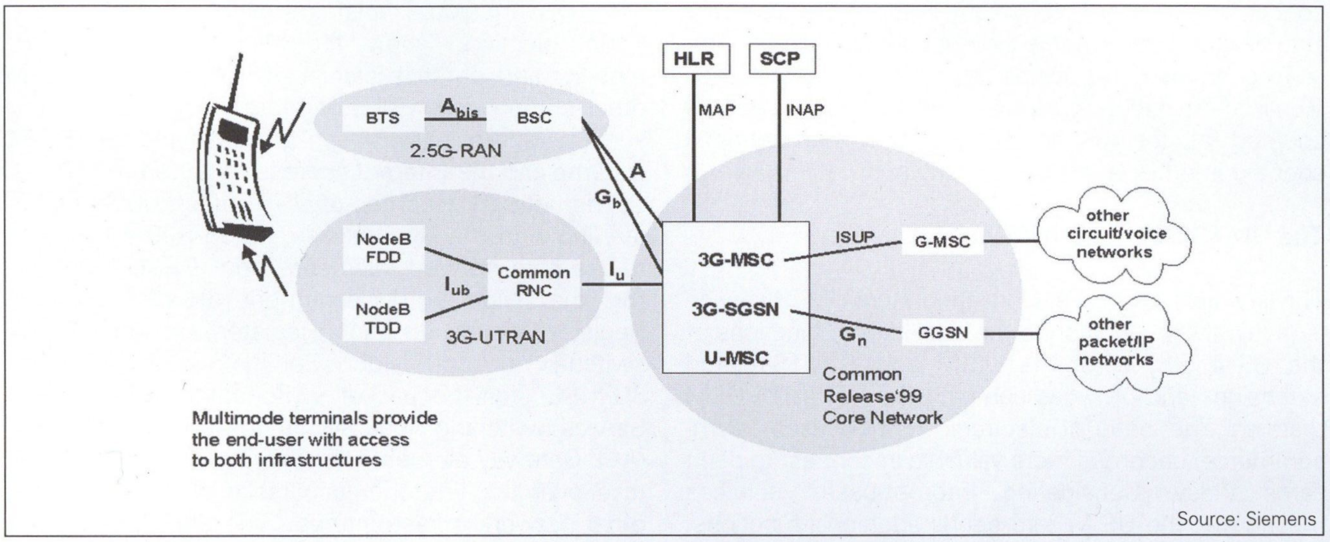


Figure 10. Evolving Proven 2G Network Infrastructure

Terrestrial Radio Access Network (UTRAN). This radio part is dealing with two operational modes, FDD (Frequency Division Duplex) and TDD (Time Division Duplex) in order to enable the network operator for the use of paired and unpaired frequency bands. The respective radio transceiver equipments are located in the Node B, the Radio Network Controller (RNC) has the task to control several Node Bs and to concentrate/distribute the information flows.

The evolution from 2G/GSM to UMTS meanwhile also includes the TDMA standard from the Universal Wireless Cellular Consortium (UWCC), an independent organisation of IS136/IS41 industries (manufacturers and operators). This consortium participates as Market Representation Partner in the 3GPP global standardisation of UMTS. Other Market Representation Partners are: the UMTS Forum, the IPv6 Forum, the Global Mobile Supplier Association GSA and the GSM Association GSMA etc.

As more markets introduce 3G networks, led by Europe, we will begin to see more differentiation in style and design. Even in Europe, broad implementation of mobile Internet browsing is not expected before 2002 with the launching of 3G networks combined with the availability of browser enhanced 3G terminals.

But the move into 3G will happen quickly. The current average global replacement rate is close to 50%; meaning phones are replaced once every two years. We don't see the replacement average increasing much, except perhaps on a regional basis. But it is possible that the majority of users worldwide will be using 3G terminals by 2006 or 2007, making way for the fast adoption of 3G applications [1].

The idea of a mobile phone or PDA today may radically change in just a few years. With numerous committees and standards boards focusing on the mobile and wireless space, and with the conjunction of computing and communications, we are seeing a rapid change in the way terminals are designed and used.

5. Spectrum Situation for Mobile Services in the 21st Century

The discussion on spectrum for mobile communication systems goes back for a number of years. The World Administrative Radio Conference WARC-92 already identified a total of 230 MHz of spectrum in the 2 GHz bands for 3G services on a global basis. These 230 MHz are split into 170 MHz for global terrestrial use and 60 MHz shared with satellite services.

As shown in Fig. 11 the main regions in the world didn't converge fully in the direction of the WARC-92 agreement.

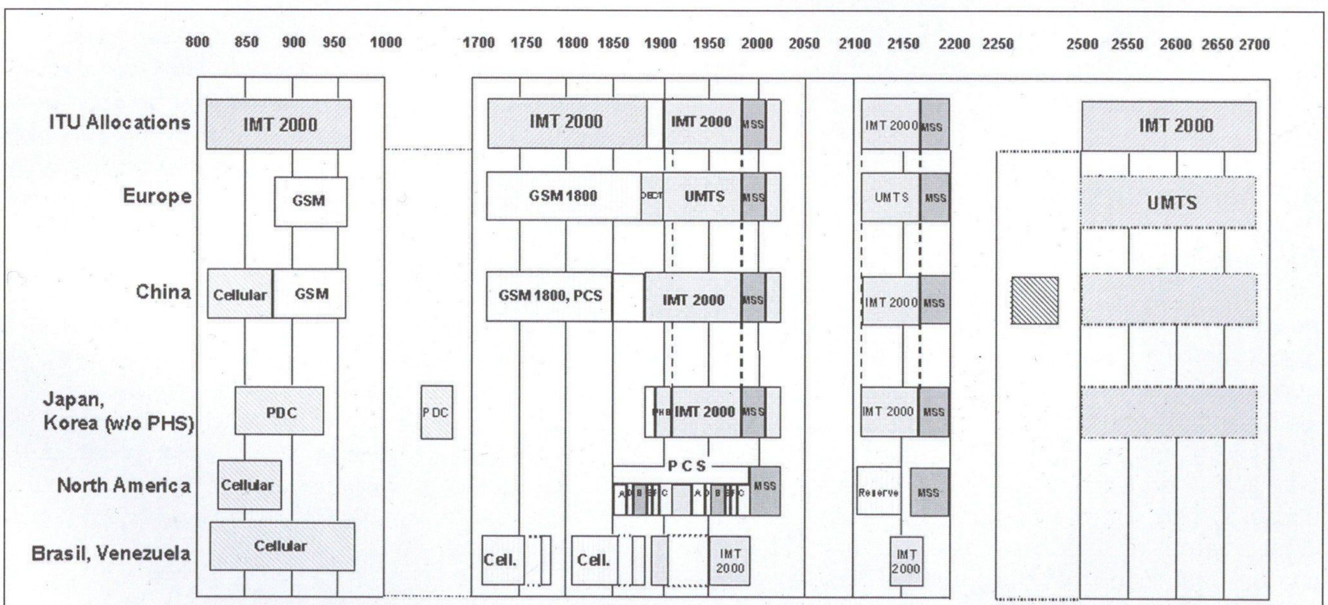
- CEPT in Europe has designated most of the spectrum for UMTS, the European solution for IMT 2000; 155 MHz of spectrum will be designated to terrestrial services with an additional 60 MHz set aside for UMTS satellite services. The remaining 15 MHz of the total 230 MHz identified by WARC-92 is made available to DECT within Europe and is therefore not available for UMTS applications at the present time.
- The Japanese Frequency Administration is going to designate the WARC-92 spectrum for third generation systems in the same way as the Europeans with the difference, that the frequency band 1895 MHz to 1915 MHz is made available to PHS services.
- In North America a different scenario developed: the introduction of PCS services and the auctioning led to a split into licenses of 2 x 15 MHz and 2 x 5 MHz up to 1980 MHz. Questions appear, how such services can be implemented and how they could be harmonised with IMT 2000 services in Europe and in Japan and in the rest of the world.
- The remaining regions in the world continue with the WARC-92 decision. There are no indications in a different direction, which means that the ITU objective, to come to a worldwide harmonised IMT 2000 spectrum allocation, is still valid for the most parts in the world.

Estimations on Total Spectrum Demand

The investigations made up to now on mobile spectrum demand led to an estimate for the total bandwidth needed in the years of 2005 and 2010. The investigations are based on market forecasts, traffic assumptions and population density in various scenarios [5].

The total spectrum demand figures include all services as presently covered by second generation systems (mainly voice) and future multimedia services. The total spectrum demand for terrestrial and satellite systems in the three regions of the worked is shown in Fig. 14. The spectrum demand per operator varies dependent on its market share. In order to give some guidelines to regulators on how much spectrum will be needed per UMTS Operator in the initial phase, the UMTS Forum released its Report No. 5 [8].

Already in the early discussions about 3rd generation mobile services, it was recognised that the considerations for UMTS spectrum in the regions may differ considerably. To enable the administrations a stepwise and more flexible spectrum designation for UMTS, the following concept was developed: Every country needs start-up spectrum for 3G, the minimum bandwidth per operator is 2 x 15 MHz paired and 5 MHz impaired. For full exploitation of the market later on additional spectrum will be needed. The start-up spectrum is called Core Band; the additional spectrum is called Extension Band. This concept gives guidance to the industry for harmonised development of products and services.



Source: Siemens

Figure 11. IMT-2000/UMTS Frequency Spectrum After WRC-2000

The Core Band is derived from the IMT 2000 Band. As the UMTS is understood as an IMT 2000 service, the UMTS Forum recommended, to start defining a Core Band within the ITU, which will be available in parts or in total in most parts of the world. The size of the Core Band is for terrestrial use differing from country to country for satellite use it remains more or less the same.

The Core Band should be designated as a first priority for full mobility applications only.

The Extension Band should be a harmonised band throughout a region and - if possible - beyond. The size in the 2500 MHz range is up to 190 MHz.

For the efficient use of spectrum, different operation modes were standardised for UMTS based on wideband CDMA (WCDMA): Time Division Duplex (TDD) technologies and Frequency Division Duplex (FDD) systems:

WCDMA is a direct sequence cdma scheme using a single carrier with a bandwidth of 5 MHz. With this radio scheme, users' information bits are spread over an artificially broadened bandwidth. This is done by multiplying the information bits with a pseudorandom bit stream, which is several times as fast. The bits in the pseudorandom bit stream are referred to as chips, the stream is known as spreading factor (typical values are 4 – 16). On the receiving side are the correlation receivers. They store exact copies of all the spreading codes. The received data stream is multiplied with these codes, and the information is selected with the same codes used for transmission. The decoded user signal gets increased by the spreading factor; it is called processing gain. Low user bit rates get a lot of processing gain with high spreading ratios, while high user bit rates get less processing gain because of their lower spreading ratios. The spreading and de-spreading allows all the radio base stations in a network to use the same frequency carrier (re-use 1). Strict power control is required for the terminals, so that signals from all terminals arrive at the base station with the same strength.

6. Final Remarks

Although 3G progressed strongly in development and standardisation, there are still many activities necessary to get to the mobile Internet market. The global scenario looks promising: in Japan, In Europe a number of Trial Networks started already in the year 2001 and continued and to public service. The technology and market learning curves are on their way and improving every day. The pressure is high, because many operators paid high license fees and want a return of their capital expenditures. Thus, we can be optimistic, that the future of Mobile Internet will come enabling us to communicate ‚Anywhere, Anytime and instantly!'

References

1. UMTS Forum Report No. 9, 13
2. UMTS Forum Report No. 8, „The Future Mobile Market", March 1999
3. UMTS Forum Report No. 10, „Shaping the Mobile Multimedia Future", October 2000
4. IEEE Communications Magazine, September 2000, Vol. 38, No. 9, UMTS, the Mobile Multimedia Vision for IMT-2000: A focus on Standardisation, Josef F. Huber, Dirk Weiler, Herrmann Brand, Siemens AG
5. UMTS Forum Report No. 6, „UMTS/IMT-2000 Spectrum", December 1998.
6. Third Generation Mobile Communication Systems, Ranjee Prasad, Werner Mohr, Walter Konhäuser. ArtechHouse Publishers, Boston, London.
7. ERC Report ‚Global Circulation of Terminals'
8. UMTS Forum Report No. 5, „Minimum Spectrum per public terrestrial UMTS Operator in the initial phase", UMTS Forum 1998.

Mr. Josef Franz Huber, *born in 939 in Austria, studied Electrical Engineering in Innsbruck and Vienna and finalised his studies in 1965 with the Master of Science degree at the Technical University in Vienna/Austria.*

In 1965, he joined the Siemens Central Laboratories in Munich/Germany, for the development of the first Siemens stored programme controlled text and data switch EDS. In 1980, Josef Huber became head of the Text and Data Development Department for the circuit and packet switching technology, value-added services and satellite data communications "Kopernikus". From 1987 to 1990, his work additionally focused on broadband Metropolitan Area Networks. In 1990, where he built up a mobile development group in Munich and Berlin/Germany, and where he also assumed responsibility on GSM data service developments and radio-related issues.

Presently, M J. F. Huber is active as Senior Vice President in the Mobile Network Business Unit. He was Chairman of the Mobile Group in the German Electrotechnical Industry Associaton ZVEI. In the years 1996 – 1998, he chaired the experts group in the UMTS Forum dealing with global spectrum calculations, methodologies and traffice models. This team was strongly involved in the preparations for the ITU World's Conference 2000. He was deeply involved in the conceptual work for future on IMT-2000 resulting in a number of agreements for IMT-2000 spectrum world-wide and on the extended UMTS Vision, the migration of mobile technologies with the Internet.

He is Vice Chairman of the UMTS Forum dealing with 3rd Generation Mobile Multimedia issues for global applications. His further involvement lies in the German industry association BITKOM as the acting Chairman for the mobile industry sector.

A Vision for an IP Future

ALISTAIR MCBAIN*

Marconi plc, UK

The paper outlines a vision of the New Public Network that is based on a QoS enabled IP infrastructure. A pilot network provides the proof that this vision is moving towards reality. Such a network will deliver tangible benefits to both end users and network operators.

1. The New Public Network

Today we enjoy a great variety of telecommunications services, not just traditional voice calls, but also television and other video applications as well as access to the ubiquitous Internet. As each of these services has been introduced it has often been supported by a new dedicated network infrastructure, leading to a set of networks being built with a number of 'layers', one for each of the services that are offered. Each service layer then contains a dedicated user data plane, most services also require their own control and management planes.

The benefit of this approach is that each of the service dedicated networks can be optimised to deliver the performance that each type of service requires, but this comes at the cost of inefficiencies in total bandwidth utilisation, maintenance, and overall network operations. In particular, there are significant costs involved in managing a proliferation of such multiple networks.

This model is also constrained by the emergence of ever more new services and applications, many of which are impossible to predict. Ten years ago, for example, few people would have anticipated the phenomenal impact that the Internet has had on telecommunications. As an alternative to adding ever more network layers, there is clearly a desire to base services on a common platform that is as future proof as possible.

Technology advances are now making such convergence possible, providing the capability to consolidate these multiple overlay networks into a single converged network that delivers multiple services, including legacy services such as voice in addition to new multimedia based services. Gateways will then deployed to map the legacy services onto the converged network. Another driver towards this network consolidation is that the boundaries that separate service specific networks are blurring. Examples include telephony service which can now be delivered over ATM or IP and video content that is provided as an IP stream (Figure 1).

Many large, incumbent Network Operators are already starting to migrate towards this type of converged network, for example by combining Frame Relay and ATM services over a common core. New Network Operators are often now building a converged network from the outset, thus providing the flexibility and cost advantages that are key to their competitive position against the incumbent operators.

These considerations lead to a network approach that is becoming referred to as the 'New Public Network' (NPN). The New Public Network can be defined as a single converged network that will support current and future services with 'Service Provider' quality. The economies of scale and the elimination of duplicate resources within the NPN will drive down bandwidth costs, while the simplicity of managing a converged network compared to multiple discrete networks will dramatically reduce operations costs.

Which user data plane protocol should be used for the NPN? With the veritable explosion of the Internet, there can now be no doubt that the Internet Protocol (IP) has become well established within the data communications world. In particular, IP delivered across Ethernet now provides a ubiquitous user

* The paper has been presented the ITU Telecom Africa Forum 2001

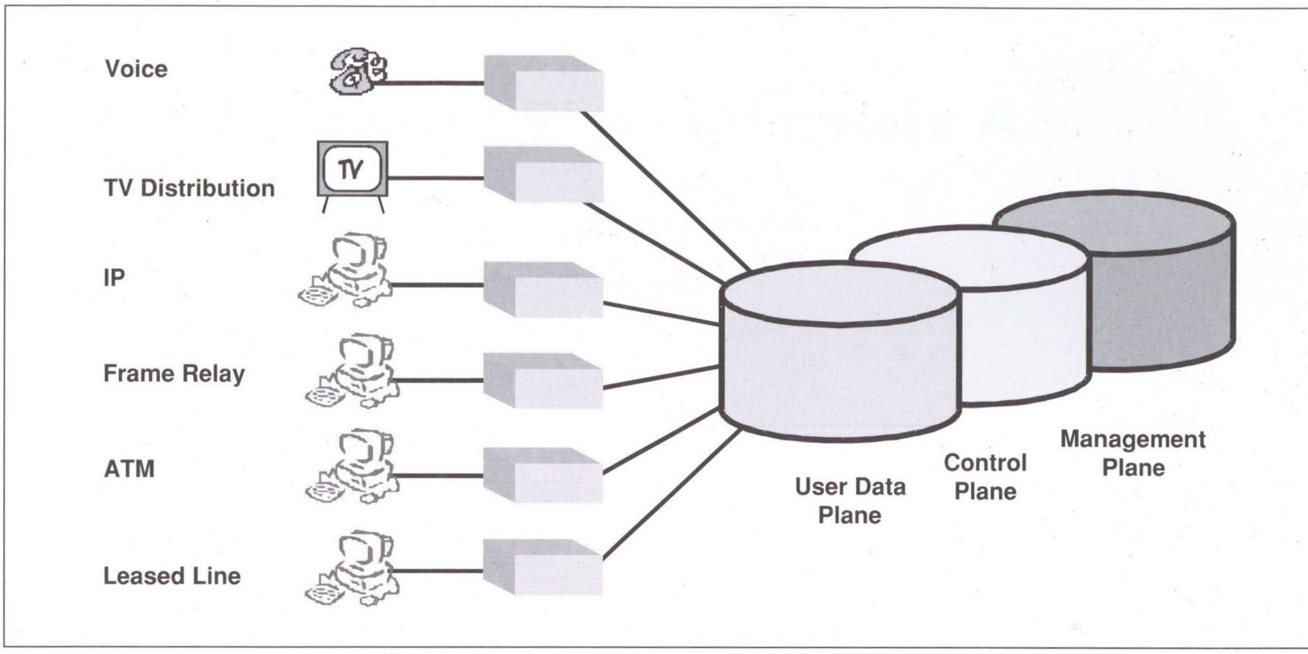


Figure 1.

interface. The influence of IP continues to spread, for example, we are now starting to see the emergence of voice services being carried over IP networks as Voice over IP (VoIP). IP has therefore emerged as the unifying protocol that will support the new generation of multimedia services. Furthermore, the PCs that we already use to access the Internet are fast evolving to become a de facto multimedia terminal device. As a result its therefore now generally accepted that the data fabric of the NGN should be IP based. However, there is a caveat to this conclusion, as will be discussed in the following sections it will be necessary to underpin the IP protocol with a connection orientated layer, such as MPLS, in order to provide the Quality of Service levels that the NGN services demand.

Current networks also contain multiple control planes, for example narrowband call control coupled with the ITU Signalling System No. 7 is dedicated to controlling voice and other 64kbit/s circuit switched calls. The converged control plane of the New Public Network, on the other hand, will feature full support for multimedia services and will require the use of IP based signalling protocols such as H.323 and the Session Initiation Protocol (SIP).

If we are to have an IP based network core for the NPN, then there is merit in using the well established Ethernet interface to provide user access. Some extensions will be required to the basic Ethernet standard, for example Quality of Service (QoS) parameters will need to be supported. Another addition that is required will be the inclusion of a power feed across the interface, to allow the network to line power IP telephones that provide lifeline telephony.

This leads to a vision for the NPN where all services, including telephony, are hosted on an IP based converged network which is accessed via a standard Ethernet interface. In many ways, this will be the realisation of the 'ISDN dream' where networks offer a single user interface which can be used to access all services types which are carried by a common network.

There are, however, a number of technical issues that will need to be addressed in order to turn this vision into reality. For example, it will be important to ensure that the necessary Quality of Service requirements are met within an IP based multiservice environment. The need to add QoS to IP has lead to the development of Multi Protocol Label Switching (MPLS) which is discussed later. MPLS provides one way to successfully address these concerns by adding traffic engineering to the IP routing protocols.

This remainder of this paper outlines some of these technical challenges that are to be faced in realising this vision of an IP based NPN. An early implementation of such a network (in Dubai) is described along with an overview of the resulting benefits that can be gained by both end users and network providers.

2. Public Network Performance

It is very important that the approach that is adopted for an IP based 'network of the future' is of 'carrier class' and is able to support public network services. While there are many 'enterprise class' multimedia based solutions available today they generally do not scale to provide the throughput or provide the availability and other features that are required for a revenue earning public network.

The support of public network performance places a number of requirements on a network, the main ones being highlighted below.

- *IP based with QoS.* As stated above, while the new network will be IP based, this must be IP with QoS guarantees. This important requirement is considered further below.
- *Support for a new range of applications.* There are already many different services which are offered by today's networks. This number will continue to grow as new services are introduced. Many of these new services are likely to be multimedia in nature. While it is obviously impossible to predict the characteristics of new services, which are sometimes referred to as 'As Yet Unknown Services' (AYUS), their support is nonetheless a requirement for the new network!
- *High availability.* In order to carry 'lifeline telephony' the network will need to achieve at least 99.999% (five nines) availability. This requires that there are no single points of failure within the network. In order to achieve this the equipment deployed should support features such as hot swappable hardware and on-line software upgrades.
- *Scaleable.* The design of the network must be economic at both small and large configurations as well as supporting in-service growth. High traffic levels should be handled without degrading service
- *Based on global standards.* It is obviously desirable to have a standards based approach to the NPN. While recognising that the ITU, through its series of Recommendations, defines the current telecommunications infrastructure it is important to also acknowledge the role of other bodies such as the Internet Engineering Task Force (IETF) in shaping the Internet together with the applications which run over IP based networks.

3. Ensuring Quality of Service

One key requirement for the New Public Network is the ability to support voice and other real time services with an acceptable quality. The exact definition of what is 'acceptable' may be an issue for debate, but it is generally agreed that performance levels should remain at least similar to those of traditional voice networks. It should be possible to define 'Service Level Agreements' (SLAs), which will form a contract between a Network Operator and the end user, guaranteeing a certain QoS level.

This leads to the concept of Quality of Service (QoS). QoS can be considered to be a collection of parameters that characterise a connection as being fit for use. From an end user point of view the principal factors which affect the perceived QoS include:

- *Delay*
  The delay across a network is obviously a critical issue for real time services such as voice. It is true that transporting voice over any packet network will introduce a packetisation delay into the connection. However, there is a degree of phobia here which dates back to the times when echo cancellation was expensive to implement. When it is considered that the delays within GSM networks require the use of echo control on every connection it can be appreciated that this is no longer either a significant cost or a source of user concern. In reality, studies show that, with echo cancellation, a round trip delay in the order of 300 ms can be tolerated.
- *Delay variation or jitter*
  Delay variation is a concern both for voice and other real time services as well as for 'near real time' services such as video on demand, where the absolute delay is not important, but play back buffers could empty if the delay variation exceeds certain limits.
- *Error rate (packet loss)*
  Packet loss affects many services, and it is of particular concern for video services. The coding techniques deployed mean that individual packet loss effects may last for many video frames over several seconds. For example, a packet loss rate as low as 3% may translate into a video frame error rate of 30%.

Unfortunately, the standard IP protocol as deployed in the Internet has a very poor reputation with regard to the provision of QoS. It was originally designed to support military and research applications, but is now being deployed on a scale that was never envisaged. It was also only intended to support 'best efforts' traffic and provides no QoS capability. The performance of the public Internet only serves to reinforce this view. As a result it is not possible to guarantee the delivery of real time services such as voice over a conventional IP network in the presence of other data traffic. 'Add-ons' to the basic IP protocol such as IntServ (based on the Resource Reservation Protocol, RSVP) and DiffServ (which makes use of the Type Of Service (TOS) field within the IP header) have only gone some way to addressing this problem. However, the 'Big Internet' still operates, in the main, with no QoS and so cannot be used to carry NGN services.

The Asynchronous Transfer Mode (ATM), on the other hand, does provide the hard bandwidth guarantees that many services require, as well as other traffic management and traffic engineering capabilities that help connectionless IP networks scale. As a result, many parts of the 'Big Internet' now run over ATM backbones which allows Network Operators to provide control over how the bandwidth is used and provide mixed services over the links that it uses.

In recent years there have been a number of initiatives which seek to more closely couple the benefits of IP with the connection orientated capabilities of ATM in order to provide a QoS enabled IP backbones with improved packet forwarding performance. The proprietary nature of most of these initiatives have not led to their wide adoption, but this situation changed in the spring of 1997 when the IETF started work on Multi Protocol Label Switching (MPLS), developing an open set of IP based protocols.

The essence of MPLS is the generation of a short fixed-length 'label' that acts as a shorthand representation of an IP packet's header. MPLS enabled switch/routers within an MPLS network then perform table lookups based on these simple labels to determine where data should be forwarded.

The label summarises essential information about routing an IP packet, for example:

- Next destination
- Precedence
- Virtual Private Network membership
- Quality of Service (QoS) information, i.e. from RSVP
- The route for the packet, as chosen by the Traffic Engineering (TE) algorithms that run within the network.

The emergence of MPLS has given rise to the concept of a 'switch router', a term which acknowledges that such devices provide packet routing based on a connection orientated switch fabric. There are two broad categories of MPLS switch router. At the edge of the network, we require high performance packet classifiers that can apply (and remove) the MPLS labels, these are referred to as MPLS edge label switch routers, or more simply as Label Edge Routers (LERs). Complete analysis of the Layer 3 header is performed only once at the LER. The LER then applies a fixed length MPLS label to the IP packet that is determined by the Layer 3 headers.

Inside a MPLS network core high bandwidth Label Switch Routers (LSRs) are deployed to process the labelled packets At each LSR within the network, only the label from the incoming packet needs to be examined in order to send the packet on its way across the network. Forwarding decisions are made simply by a single table lookup from the fixed-length label. The relative simple functionality of an LSR allows it to process packets quickly, ensuring a high throughput within the network core. At the other end of the network, a LER swaps the label out for the appropriate Layer 3 header data linked to that label.

Label switching thus integrates switching and routing functions, combining the reachability information provided by the router function, plus the traffic engineering benefits achieved by the optimising capabilities of switches. This illustrates another of the key concepts in MPLS. Conventional IP routers are based on the exchange of network reachability information. They contain routing tables which are looked up using the IP header from a packet to decide how to forward that packet. These tables are built by IP routing protocols (e.g. OSPF or BGP-4) which carry IP reachability information in the form of IP addresses. In practice, forwarding (IP header lookup) and control planes (generation of the routing tables) are tightly coupled. The route to the final destination is chosen at each router and may change dynamically.

Since MPLS forwarding is based on labels it is possible to cleanly separate the (label based) forwarding plane from the routing protocol control plane. By separating the two, each can be modified independently. With such a separation, there is no longer a need to change the forwarding mechanisms, for example, to introduce a new routing strategy into the network.

The MPLS protocol is applicable to networks using any Layer 2 switching protocol, but it has particular advantages when operated over ATM networks. MPLS can integrate IP routing with ATM switching in order to offer scalable IP-over-ATM networks.

MPLS certainly has a lot of promise, and will be one of the key enabling technologies for the New Public Network, but ATM will still be used for the next several years to provide hard bandwidth guarantees. However, it is important to remember that IP is not MPLS, rather MPLS allows connection oriented capabilities to be added to IP.

While MPLS will provide a solution to the provision of QoS within the network core, there remains the problem of transporting QoS enabled traffic within the access component of the network. As discussed elsewhere, the ubiquitous Ethernet interface may be regarded as the user interface for the NPN. Here too, recent developments have seen the addition of QoS support with the extension of Ethernet to include the traffic priority mechanism that is embodied within the IEEE 802.1D (formally the 802.1p) standard

4. The IP Vision

Based on the above discussion it is possible to define a vision of the future IP based New Public Network. Such a network will exhibit a number of attributes, which are summarised below.

Single multi service network. The New Public Network will be based on a single infrastructure that will support all services, including voice, video and data. The network should also support the next generation of multimedia services that will become increasingly important in the future.

QoS enabled. As stated above QoS is essential in order to provide support for real time services such as voice over an IP network. MPLS has emerged as the best contender to provide QoS within the network core.

Common user interface. All services must be accessible via a standard user-network interface. This will require that telephones, Set Top Boxes and PCs all support a standard IP based interface. The obvious candidate is 100BASE-T 100 Mbit/s Ethernet accessed via a RJ45 connector.

Cost effective. Lastly, the solution will need to be prove to be economic compared to the alternative of a number of service dedicated networks. Besides the initial costs it is necessary to consider the total cost of ownership of the solution. The potential for a NPN to support new services as they are developed will provide further cost savings.

It could be argued that equipment costs, in particular for the Customer Premise Equipment (CPE), will be significant for an all IP solution. For example, an Ethernet based IP telephone is likely to cost more than a simple analogue instrument. While this may be true in the short term, it is expected that the economics of scale will largely remove this issue since the additional complexity lies largely in silicon components. It is anticipated that any residual cost penalties will be more than balanced by the savings to be achieved from an integrated network. Furthermore, the use of IP telephones opens up possibilities of delivering different services such as videophone, e-mail pick up, web browsing etc from a phone connection and also in delivering phone service in turn to other devices such as PCs and PDAs.

5. The Vision in Reality - Dubai Marina

Given the investments that exist in the current generation of networks, it will obviously be some time before this goal of a ubiquitous IP based network that is described above becomes commonplace. However, we are already starting to witness some steps in this direction with the establishment of some new 'green field' networks that are being built around the principles that are outlined within this paper.

One example of such a network is now under construction at Dubai Marina. Dubai Marina is currently being built on the shores of the Arabian Gulf as a 'city of the future'. It will include homes for some 160,000 people in a range of luxury villas and apartment buildings along with commercial, educational and healthcare facilities. Marconi has been tasked with ensuring that a state of the art communications infrastructure is incorporated into the new city from the outset.

Dubai Marina will be a city that employs sophisticated technology to manage and control the environment and the services that its people will enjoy. Dubai Marina will be a complete city, which will offer virtually all services and activities electronically, built around the principles of e-commerce, e-leisure, e-healthcare, e-business, e-government and e-education. It will be an intelligent city.

The plans for this new development envisage the creation of a powerful economy that is fuelled by technology. Technology will be used in order to reinforce and nurture a sense of community. The lifestyle that this enables will attract entrepreneurs and innovators from the world of business, commerce, finance, technology, science, education, healthcare, tourism and the arts. They will be the people who know how to use technology and information to build a vibrant and progressive society.

Security will be a key benefit of the communications network: technology will evolve to include advanced personal identification, voice recognition and surveillance technologies, people will be secure in the knowledge that their home, business, individual and family security needs are looked after. A fully integrated network also means the city will have information at its fingertips to respond quickly to events, co-ordinating emergency services, such as fire, ambulance and police services.

In Dubai Marina, business will be able to take advantage of the electronic marketplace, bringing customers and suppliers together in smart databases and virtual environments, with ID verification, encryption and online translation. Requirements such as billing, taxation and electronic funds transfer will be made easier, as will be automatic production of accounts and auditing. Electronic currency will be used for purchases, including those made at both 'real' and 'virtual' shops as well as airline reservations and sports facilities.

The services that will be supported include telephony, high-speed Internet access, Video-on-Demand, real-time television from around the world, and video-telephony. Applications will include an extensive library of VoD programmes, video-telephony between properties, portals to home shopping, concierge, information and security systems.

In keeping with the vision of a city of the future, the entire telecommunications infrastructure for this new development will be IP based, delivering voice, television and Internet over a common 100 Mbit/s Ethernet interface fed directly into each home or business location. IP telephones, Set Top Boxes and PCs will directly connect into a switched Ethernet network that is provided within each house and office. QoS enabled switch routers will be deployed within the network to form a highly resilient core which connect in turn to the different content servers and to gateways which link into legacy networks.

Dubai network architecture

At a high level the Dubai Marina architecture consists of three components, CPE, core network and content servers and gateways (Figure 2).

Customer Premise Equipment (CPE)

In line with the NPN vision, the access network within customer premises (i.e. residential homes and businesses) is based on Ethernet. A fibre based 100 Mbit/s Ethernet link is used to connect each customer premise to the network. This access fibre is terminated at the core of the CPE network on a small Ethernet switch that is located within each individual home or business office. 100 Mbit/s Ethernet connections then lead from the switch to RJ45 sockets that are located throughout the house or office which provide the end user interface. QoS is maintained within the CPE network by the support of

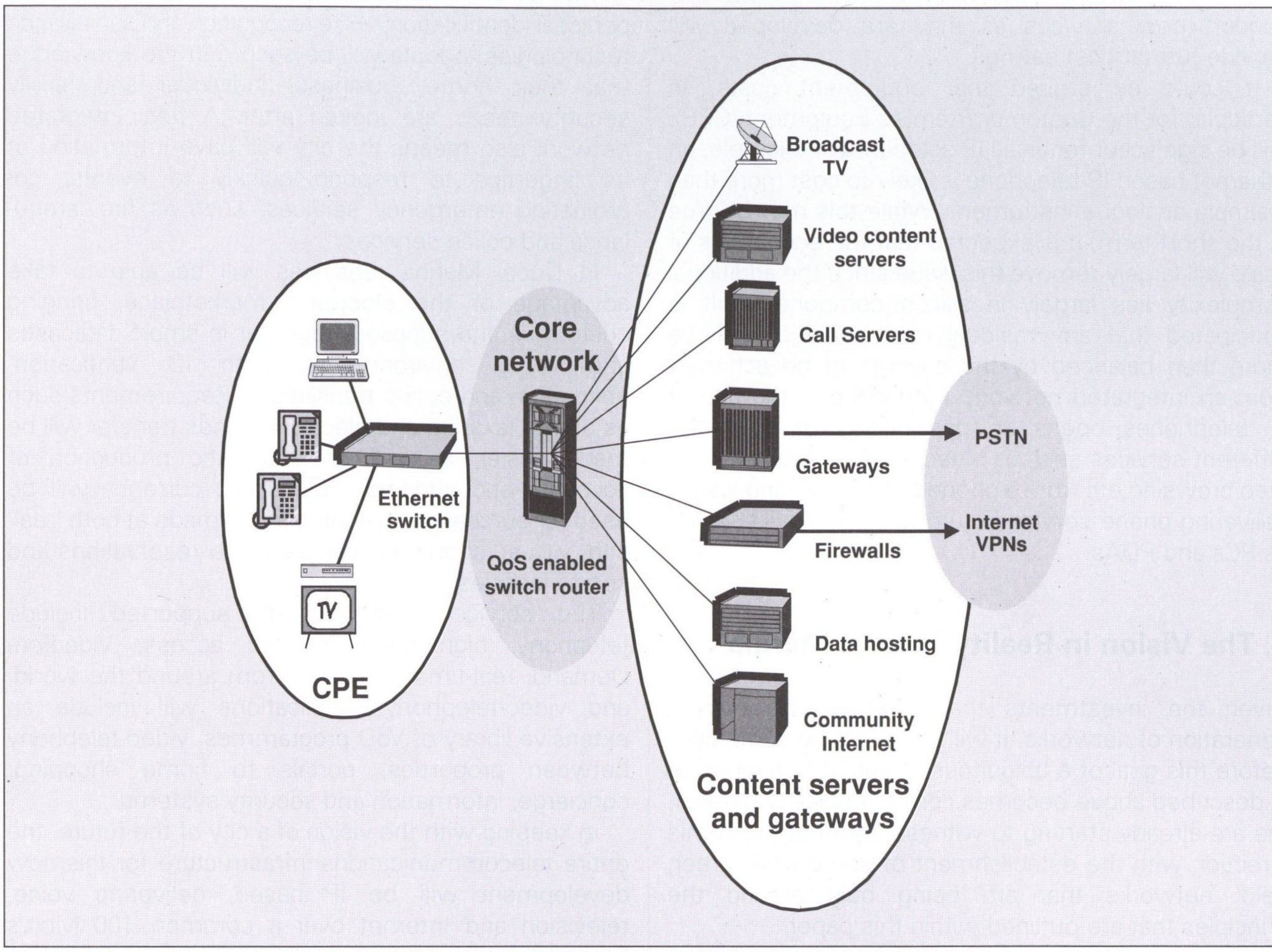


Figure 2.

four service levels within the switch. The Ethernet switch has some additional features to those that are found on traditional enterprise devices. It incorporates a standby battery and the user ports provide line powering to allow lifeline telephony functions to be provided using an IP telephone independently from any local power. Environmental issues are also addressed, for example the Ethernet switch must be able to operate within the temperature range that is experienced in Dubai.

A range of Ethernet based terminal equipment can be used. Most PCs can be directly connected using internal Ethernet interfaces, while several vendors now supply Ethernet based IP telephones. Ethernet equipped Set Top Boxes are also available to provide broadcast TV and Video on Demand services.

Core Network

The network core is initially being built around a QoS enabled IP network which routes traffic to and from servers and gateways. Four QoS types are used to support different services. As the core network grows, it can evolve to become MPLS based. Load balancing is used to share traffic between all the available servers.

Security is a major concern with an IP based network. In order to protect the data and the servers from unauthorised access several layers of firewalls are deployed. These include stateful filtering and intrusion detection firewalls along with packet filters in the load balancers. Denial of Service attacks are intercepted by the load balancing switches, thus adding additional protection to the network. A further firewall located within the Management Network secures the access to internal and future management systems (e.g. billing, etc).

VoIP telephony is supported by a series of call servers within the network core. These servers support H.323 and Media Gateway Control Protocol (MGCP) signalling, they provide standard call control facilities together with a range of advanced features such as call forwarding. Integrated PC support enables additional features including unified messaging. Gateways from the core network are provided to the PSTN to support voice interworking via standard TDM interfaces.

Broadcast television channels are routed through an IP gateway which packetises the content. IP multicasting is deployed to allow multiple customers to receive broadcast channels as required. Arrays of dedicated servers provide Video on Demand content.

Standard Internet based services are supported by the infrastructure including:

- E-mail – SMTP, IMAP, POP and Webmail with virus filtering of incoming e-mails.
- Web browsing – with web caching.
- Web serving.
- Virtual Private Networks (VPNs).

6. Conclusion

This paper has outlined a vision of the New Public Network that is based on a QoS enabled IP infrastructure. The Dubai Marina network provides the proof that this vision is moving towards reality. Such a network will deliver tangible benefits to both End Users and Network Operators. The key advantages of this approach can be summarised as follows.

End User benefits

Multimedia service support. True multimedia applications such as video conferencing are easily supported alongside legacy services, together with the ability to have mixed media calls. For example, a call may start as a conventional voice connection and then convert 'mid call' into a voice/video call.

Common user interface. The provision of all services via a single standard interface type provides great flexibility in terminal deployment. Telephones, set top boxes or PCs can be connected to any available Ethernet socket located within the home or office.

High speed access. The provision of 100 Mbit/s into every home and office provides support for true broadband services. If there are no priority services (e.g. telephony) being used at a given time then all of the bandwidth is potentially available for data applications such as file down loads. Home based workers will experience the same data performance as enjoyed by their office based colleagues.

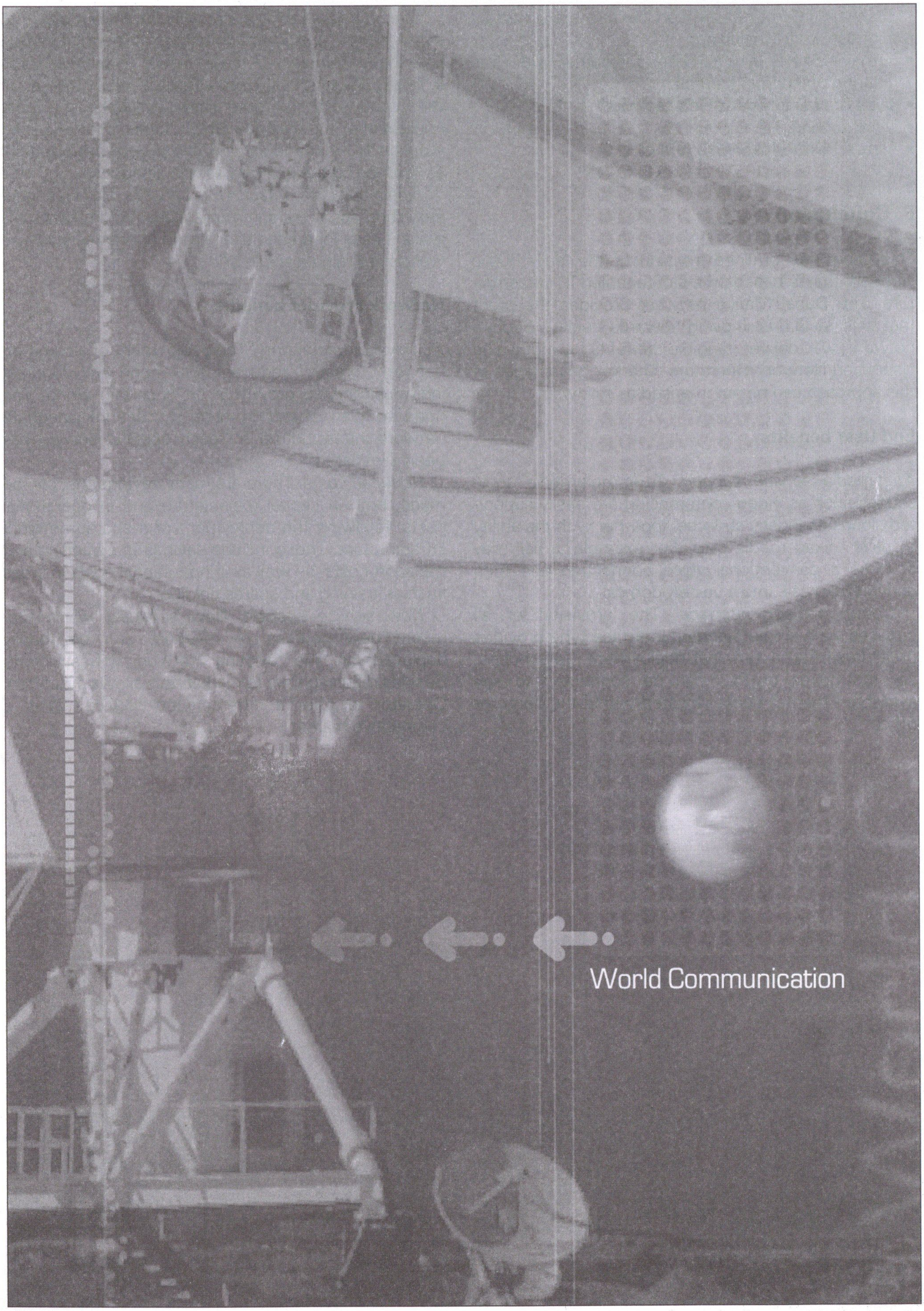
Single point of contact for all services. Having a 'one-stop-shop' for all his communications requirements will make it easy for an end user to add or drop services, as well as providing a single bill.

Network Provider benefits

Removal of uncertainty. The deployment of a common network greatly reduces the need to forecast individual service growth in an uncertain world. The need to plan ahead in order to ensure adequate network infrastructure for each specific service has been eliminated.

Reduced costs. There are many savings that can be made, both in capital expenditure and in operational costs, following the deployment of a shared network infrastructure. These include savings in switching and transport costs, as well as in operational costs in areas such as training and spares holding.

New service support. New services, which represent new sources of revenue, can be readily deployed on the existing network infrastructure. Besides reducing time to market, this allows a Network Operator to offer a differentiated service within a competitive market.


World Communication

Design and Implementation of Traffic Management Functionality for RSVP*

TAMÁS SZÉNÁSI AND GÁBOR RÉTVÁRI[1]

QoS IT Labs of the High Speed Networks Laboratory, Department of Telecommunications and Telematics
Budapest Univercity of Techology and Economics

Given today's need for transmitting multimedia data over the Internet, large efforts have been made to define and deploy various prioritized service extensions (QoS: Quality of Service) for the well-known Internet Protocol suite, retaining the existing functionality. One way to accomplish this daunting task is the so-called Integrated Services (IntServ) architecture.
IntServ implies, that some signaling protocols are invoked to establish, manage and tear down resource allocations along the data forwarding path according to the session's QoS requirements. Such a signaling protocol needs uniform access to the platform dependent lower level QoS forwarding architecture. This calls for a generic Traffic Management (TM) functionality to support QoS over heterogeneous media.
This paper describes the design and implementation of a Linux based TM module and presents a series of measurements to demonstrate the QoS that can be delivered to privileged sessions even in a congested network scenario.

1. Introduction

The Internet Community realized that the traditional, well-known Internet Protocol (IP) infrastructure is no longer capable of providing sufficient services considering the emerging need for multimedia communication. IP has proven to be very effective as long as no privileged service classes are assumed, but it is completely inadequate for the introduction of QoS. Since IP is, by nature, connectionless, it is really difficult to bind network resources permanently to a particular QoS session. One way to overcome this problem is to establish a connection-oriented infrastructure over the connectionless, best-effort IP architecture. Currently, huge research efforts are taken to fulfill this requirement.

The term "QoS IP network" refers herein to a network, that supports both best effort packets and packets with QoS guarantees. The way in which network resources are split between the two classes is irrelevant, except of the assumption that each QoS capable router in the network is able to (1) dedicate some of its resources to satisfy the requirements of QoS packets, and (2) may identify resources that remain available to new QoS flows (e.g.: for admission control purposes).

One vision of the QoS Internet is the "Integrated Services Internet" (IntServ) model: unique QoS flows are handled separately. This implies gaining fine control over QoS guarantees to be delivered for individual IP flows [1]. Nevertheless, it is clear, that the cost of sophisticated resource allocations is the increased amount of state information at routers, which is unfortunately against scalability.

In order to establish signaling connection before data communication takes place, a higher level signaling layer is introduced. The basic component of the signaling architecture is a resource reservation mechanism that creates the data path for a flow and allocates the required resources at routers along the path. Resource reservation involves the definition of the so-called session and facilitates the offering of network resources to packets associated with particular sessions. During communication setup applications that are running at Internet hosts notify the network on the characteristics of their traffic instances, and the network attempts to bind the required amount of resources to the sessions.

Nowadays prevailing reservation setup protocol is RSVP (Resource reSerVation Protocol, [2]). RSVP establishes simplex communication channels in two basic steps. First, the sender emits a reservation setup *(PATH)* message specifying its traffic characteristics and the receiver's address. Selecting an optimal path for the *PATH* message is of crucial importance as this is the way a data forwarding path is assigned for the session's traffic. If a receiver decides to accept the connection, resources are allocated backwardly along the forward path.

* This work is partly sponsored by ETIK – the Inter University Centre for Telecommunications and Informatics

1 e-mails: szenasi@qosip.ttt.bme.hu; retvari@qosip.ttt.bme.hu

Obviously, RSVP needs to communicate with the low level QoS forwarding architecture (e.g.: buffer and packet filter management, packet schedulers, etc.) in order to manipulate resources of the underlying network device.

The IntServ model implies that data forwarding paths for individual IP flows can be selected one at a time. This facilitates both the unique treatment of QoS flows from the routing perspective, and - from the traffic engineering standpoint - the introduction of some sophisticated policies to control network load with rather fine granularity [3]. Therefore, it seems plausible to invoke a special IntServ routing protocol, that can compute routes for IP flows one at a time combining the flow's individual resource requirements and the actual network load state.

A good example of how to extend a conventional intra-domain, link-state, best effort routing protocol with QoS capabilities is the QoSPF routing protocol (QoS extensions to Open Shortest Path First, [4]). Based on the link state information and the actual resource availability in the network, QoSPF computes widest-shortest candidate paths and maintains a QoS routing table (separated from the best effort one). By receiving a new RSVP PATH message routers query the QoS routing module by passing the destination address and the desired QoS of the flow in question. QoS routing picks a route from a pre-computed or on-demand calculated routing table that it deems to be an optimal one according to the session's traffic characteristics and the actual network load.

Several research results [5] have pointed out the potential of QoS routing for improving network utilization and the service level provided to requests with QoS guarantees. The improvement in the service received by users is the increased likelihood of finding a path that meets the requested QoS parameters.

Conversely, the improvement to the network operator is the increased resource efficiency in terms of the number of accommodated QoS flows or simply the higher throughput/goodput. Although the extra-complexity of QoS sensitive routing must not be neglected, a number of previous works presented evidence that the increased protocol and computational overhead of QoS routing is rather tolerable even in large networks.

RSVP and QoSPF often need to consult the low level forwarding architecture. The low level forwarding architecture is in charge of maintaining the traffic buffers, packet filters and forwarding tables. Henceforward we shall refer to this functionality as the QoS forwarding engine and we shall assume, that it is implemented at the operating system's kernel level, therefore it is supposed to be platform dependent. We think, that a separate Traffic Management (TM) functionality is desirable that interfaces between the QoS forwarding engine and the signaling layer and facilitates for RSVP and QoS routing to access the platform dependent QoS forwarding engine through a well defined uniform interface (API: Application Programming Interface). Nevertheless, currently there is no appropriate standard to define this API.

The rest of this paper is structured as follows. In Section 2 we identify the main purposes and attributes of the TM functionality and locate it in the IntServ signaling architecture. In Section 3 we describe a specific design and implementation of the TM module for the Linux operating system according to Section 2. Finally we present some measurement results in Section 4 and conclude our work in Section 5.

2. Traffic Management in the IntServ Signaling Architecture

The introduction and deployment of QoS over the Internet architecture is a somewhat complicated challenge. QoS support in IP implies overall QoS support in each network layer in the protocol stack. Sophisticated scheduling algorithms should be implemented at the data link level to deliver effective QoS for prioritized flows. At the IP layer, appropriate distinction has to be made between QoS and non-QoS packets. At higher levels, signaling protocols are needed to signal QoS among applications. Furthermore, we need real-time transport protocols to reliably carry privileged traffic.

Originally, IP was invented to co-operate with any optional lower layer mechanism, in case the link layer provides a well-defined set of functions. This means that IP can operate over various types of link layer infrastructure, which may (or may not) implement QoS support in different ways. Nevertheless, resource reservation depends on a common access interface to the QoS forwarding engine, therefore RSVP defines a set of routines (an API) for interfacing to a "kernel" traffic control mechanism. This is called LLDAL (link-layer-dependent adaptation layer). It supports QoS over passive media such as leased lines or nowadays shared LAN media. When RSVP establishes resource reservation it has to instruct the QoS forwarding engine to (1) allocate the necessary data buffers, (2) add appropriate packet filters and (3) install the QoS route in the forwarding database in order to assure route pinning. Upon local resource availability changes QoS routing needs to be informed to facilitate the synchronization of routing knowledge among routers. This calls for the maintenance of per interface resource availability information.

The module implementing LLDAL functionality is herein referred to as the so-called Traffic Management (TM) module. Obviously, the actual implementation of the QoS forwarding engine may differ depending on the underlying device, which may be a commercial router architecture, or an operating system in our case. The TM resides between the signaling layer and the kernel and can be thought of as being a uniform interface to the low level QoS forwarding mechanisms (see Figure 1).

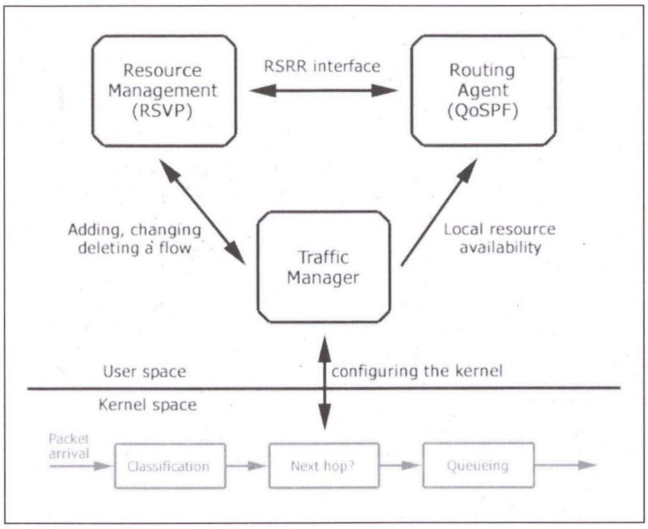


Figure 1. The IntServ signaling architecture

3. Linux Traffic Management

Linux is an open source operating system that comes equipped with full-scale and flexible kernel level traffic control functionality. The Linux traffic control infrastructure consists of a kernel level forwarding mechanism, which takes place in the IP protocol stack, and user level agents that communicate with the kernel via dedicated *NETLINK* sockets [7]. Also, there are sophisticated open source applications to manage the QoS forwarding engine [9].

Kernel Space TM modul

The kernel level QoS forwarding engine consists of a *Classifier* and a *Scheduler* module. The Classifier filters an incoming packet and directs it to the appropriate Scheduler module, which actually inserts it into the outgoing packet stream according to its privileges.

The Classifier examines packet attributes, such as source/destination IP address and port numbers in order to associate a data packet with a QoS flow. If a packet does not belong to any local reservation, it receives the convenient best effort service level. In contrary, if the packet matches any QoS filters, it enters the QoS forwarding engine. The next hop address and the outgoing interface is set to perform route pinning, i.e., to force the packet to the path determined by QoS routing for the corresponding reservation (note, that this QoS route may differ from the best effort paths stored in the kernel routing table). Next hop assignment is performed by using separate Linux routing tables and routing rules.

Packet schedulers maintain the packet queues and schedule packets according to the flow's QoS parameters and the link sharing principles.

Note, that the official Linux kernel release does not implement all the above functionality to perform QoS packet forwarding, and we needed to add some extra features to perform efficient route pinning.

User Space Traffic Management Module

The user level tool that handles the kernel level forwarding engine is the so-called Traffic Management entity.

TM serves various purposes. The main goal of TM is interpreting requests generated at the signaling level and transforming them into a format that is specific to the QoS forwarding engine. The TM's other responsibility is to provide local resource availability information for the QoS routing module. TM also maintains a repository of existing local reservations.

The TM module has three interfaces, one to RSVP, one to the QoS Routing module and a third one to the kernel. All interfaces are implemented using local datagram sockets that utilize the same support layer. This layer performs a simple encapsulation on the data that go through the interface.

The TM – RSVP interface facilitates bi-directional synchronous communication. We used the RSVP code from ISI [8] that already includes interfaces to a hypothetical kernel resident module that performs traffic management. This module implements the functions for adding and deleting reservations and managing filters. This implementation is replaced with calls to the interface to the TM process thus placement of the traffic management functionality is transparent to the rest of the RSVP code. After a request is sent to TM encapsulated in uniform data objects, the interface waits for a response for a reconfigured amount of time and return with an appropriate error if no response is received within the timeout period.

This interface contains the following entry points that through RSVP calls access the local traffic control:

- *TC_Advertise()*: to update the RSVP ADSPEC object carried by PATH messages to reflect local resource availability,
- *TC_AddFlowspec(), TC_DelFlowspec(), TC_ModFlowspec()*: to add, delete and modify reservations, respectively,
- *TC_AddFilter(), TC_DelFilter()*: to manipulate packet filters.

The last two API calls are also responsible for setting the appropriate next hop IP address for the given flow.

The TM side of the interface carries RSVP's requests into effect and, in addition, updates the per-interface information it maintains with the current levels of reservation.

The interface between TM and QoS routing is asynchronous and uni-directional since TM only needs to notify QoS routing of changes in availabile resource at local interfaces. Upon the asynchronous notification is caught by the QoS routing module it makes a decision according to a local update triggering policy whether to send an update to other QoS capable routers.

The third interface is a platform dependent access layer to the QoS forwarding engine implemented by a datagram based *NETLINK* socket. TM interprets the messages received from RSVP, performs parameter mapping and sends the appropriate messages down to the kernel. The kernel reports the results of the operation and TM forwards this information to RSVP.

A good example of the TM in action is resource deallocation. Upon RSVP receives a *RESV-TEAR* that refers to a previously established reservation it calls the *TC_DelFlowspec()* function to free resources. The arguments of the function (interface index, reservation handler) are encapsulated and sent to TM. TM looks up the local reservation registry and commands the kernel to free resources associated with the reservation identifier and informs RSVP on the results. As the amount of available bandwidth has just changed at a local interface, the TM sends a notification to the QoS routing module. QoS routing then makes a decision, whether to generate a new Link State Advertisement to inform adjacent routers on the change of resource availability. This call also involves the *TC_DelFilter()* function call to delete the session's filters and next hop settings.

4. Measurements

Our measurements aim to facilitate the comparison of the services a QoS flow can obtain in a small network topology.

In our first measurement scenario we investigated the behavior of flows requiring QoS in a congested best-effort network. This scenario is believed to represent the "ancient" Internet, where no prioritized treatment could be offered for real time traffic. In the second scenario signaling functionality is present (RSVP). In this case, RSVP invokes a Shortest Path Routing protocol (actually OSPF), thus every flow use the shortest path. Here, only the first reservation can be successfully accommodated in the system by utilizing the majority of the available capacity, and the remaining two flows used the remaining resources. In the third scenario we used RSVP and QoSPF to demonstrate the efficiency of a special QoS Routing protocol that can select an alternative path in the network (even a longer one), if it has enough free capacity to admit a flow.

We chose a network topology depicted in Figure 2. We injected three QoS traffic instances:

- 6Mbit/s CBR traffic in 0-40s time interval
- 7Mbit/s CBR traffic in 5-45s time interval
- 8Mbit/s CBR traffic, 10-50s

Bandwidth at network links is set in a way assuring the admission of one, and only one traffic instance to the three feasible paths. Traffic travels from the Terminal1 to Terminal2.

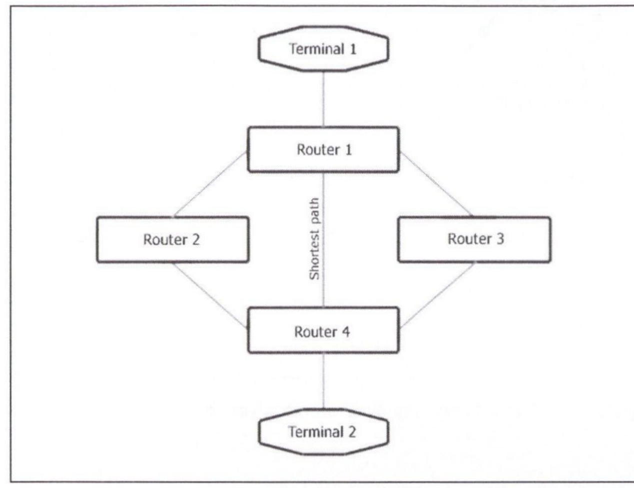


Figure 2. Network topology

Every machine in this scenario is fast enough to handle the 21Mbit/s overall load. They are PC-s with Intel 400MHz Celeron processors, network links are 10Mbit/s Ethernet LANs.

Scenario 1: Best-effort Service

In the first scenario we tried to forward each traffic instance over the best-effort infrastructure without differentiating among sessions. The best-effort routing protocol picked the same shortest path for all the traffic regardless of their nature. Figure 3, and 4 depict the throughput and the delay for every flow experienced at *Terminal2*.

Please note, that delay diagrams are only informational, and only their relative position should be considered, due to the limited precision of the Network Time Protocol (NTP) that was used to synchronize clocks at traffic endpoints.

One may realize that neither flow received the desired QoS service permanently, because of the heavy congestion on the link between *Router1* and *Router2*. Each traffic instance experienced about 0.07 sec mean delay, (0.068s, 0.077s, 0.071s, respectively). The first flow suffered about 37% USP packet loss while the second about 54%, and third about 51%. These measurements underline the experience that the service level provided by best effort service class may degrade seriously in case of congestion. This service level degradation is not tolerable when considering real time multimedia communication.

Scenario 2: RSVP invoking SPF Routing

In this scenario only one flow succeeds to obtain QoS reservation, the one that is started in second. RSVP's admission control to avoid congestion on the critical

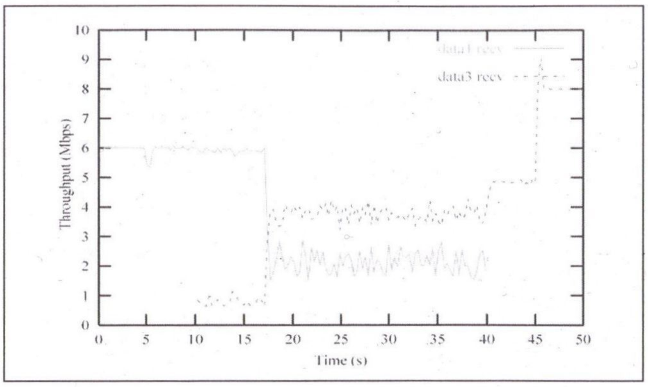


Figure 3. Throughput [Mb/s] diagram in best effort configuration

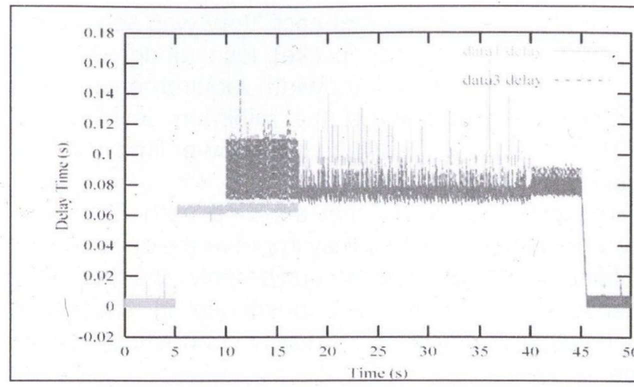


Figure 4. Delay [s] diagram in the best effort configuration

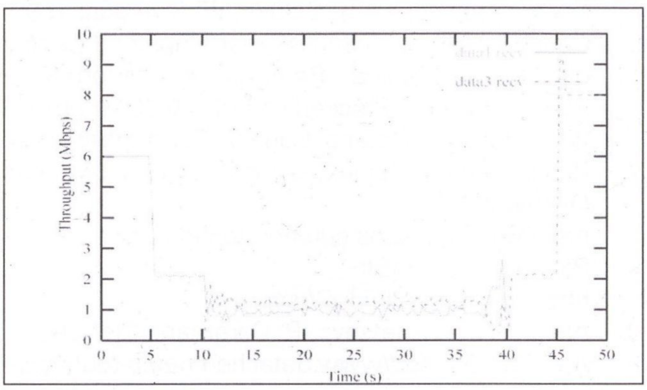


Figure 5. Throughput [Mb/s] diagram in the RSVP – SPF routing

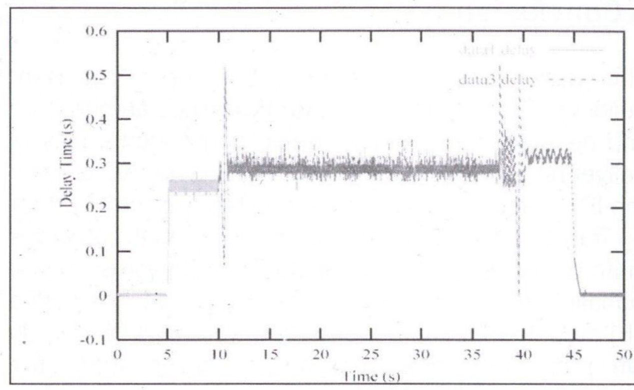


Figure 6. Delay [s] diagram in the RSVP – SPF routin configuration

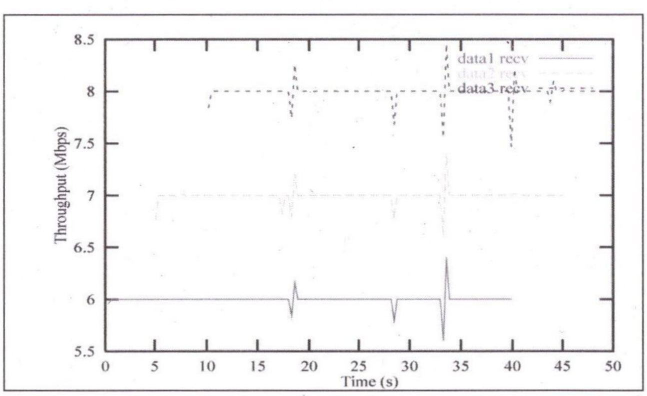


Figure 7. Throughput [Mb/s] diagram in the RSVP – QoS routing configuration

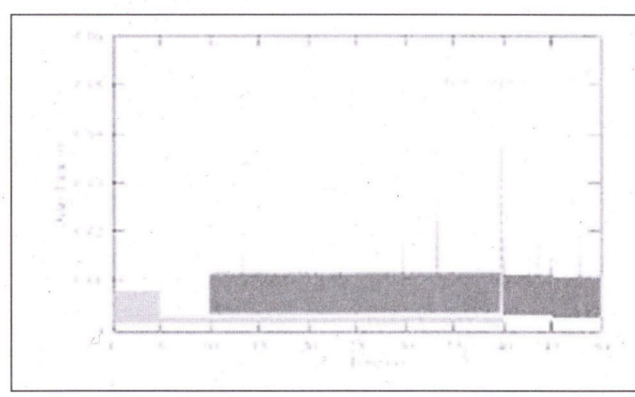


Figure 8. Delay [s] diagram in the RSVP – QoS routing configuration

shortest path from Router1 to Router4 rejects the third flow. Figure 5 and 6 show the measurement results.

Data2, the traffic instance that invokes RSVP to obtain QoS enters in the 5th second of the measurement. It immediately receives its allocated bandwidth share, and makes best effort *Data1* to fall back to about 2 Mbps. Even this service level is further lowered by the third traffic instance entering at the 10th second.

The flow, which received QoS *(Data2)* suffers 0.37% USP packet loss, while the best effort flows see 70% and 73% losses respectively. The mean delay of *Data2* is 0.02 sec, best effort packets are delayed by approx. 0.25 sec, because these packets can only be served after QoS packets were inserted on the wire.

Conclusions are that RSVP can guarantee QoS service for flows that require it and get admitted, however, without appropriate routing support, it may be rather ineffective.

Scenario 3: RSVP co-operating with QoSPF

In this scenario all reservation requests obtain the desired QoS. QoSPF routes each flow along a separate path owing to the sophisticated routing decisions made. The second and third flows utilize the alternative longer paths (via *Router2* and *Router3*) that have enough free capacity to accommodate them.

The diagrams show that each flow was served in a manner as to minimize packet loss (under 0.1% in each case) and assure bandwidth requirements.

Delay is maintained at the minimum level (under 0.01 sec each), which is about the lower limit of NTP's precision in our network.

These results show that applying both RSVP and QoSPF in the same time may improve the service level obtained by QoS sessions considerably, and the Traffic Management functionality, described in this paper works seamlessly even in heavily congested network scenarios.

5 Conclusions

This paper aimed to identify the most important building blocks of the Integrated Services architecture and describe the role that these components play in delivering prioritized services for QoS sessions over the IP infrastructure. Also we pointed out the need for a Traffic Management component that provides uniform access to low level QoS mechanisms for the signaling layer protocols. We described a Linux based TM module that we implemented from scratch. Finally we presented measurements to demonstrate the benefits of the IntServ model that are relevant for future users and network operators.

References

1. Braden , Clark, Shenker: "Integrated Services in the Internet Architecture: an Overview", RFC 1633, IETF, June 1994.
2. Braden, Zhang, Berson, Herzog, Jamin, "Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification", RFC 2205, IETF, March 1997.
3. Crawley, Nair, Rajagopalan, Sandick: "A Framework for QoS-based Routing in the Internet", RFC 2386, IETF, August 1998.
4. Apostolopoulos, Guerin, Kamat, Orda, Przygienda, and Williams: "QoS Routing Mechanisms and OSPF Extensions", RFC 2676, IETF, August 1999.
5. Apostolopoulos, Guerin, Kamat, Tripathi: "Quality of Service Based Routing: A Performance Perspective", In Proceedings of SIGCOM, pp. 17-28, Vancouver, Ontario, Canada, September 1998.
7. Radhakrishan: "Linux - Advanced Networking Overview 1." http://qos.ittc.ukans.edu/howto/index.html
8. RSVP project at ISI: http://www.isi.edu/div7/rsvp/
9. Alexey N. Kuznetsov: IP Command Reference, April 1999 http://www.detached.net/ip-routing/ip-cref/

Reusable Building Blocks for Telecommunications Communication Integration

DR WILLIE DONNELLY

Telecommunications Software Systems Group Waterford Institute of Technology

MR. VINCENT WADE

Dept of Computer Science Trinity College, Dublin

vThis paper proposes a management component framework to support the rapid and flexible construction of an e-Commerce management infrastructure. This management solution is based on a holistic management approach supporting seamless integration of network and application management services (i.e. vertical), as well as integrating management across distinct functional areas (i.e. horizontal). e.g. IP virtual private network provisioning, QoS and accounting/billing functional areas. The target audience of this Framework is broad, covering service providers, management system integrators, and independent software vendors and standardising bodies. This Framework is being evaluated through a focus on the management of business-to-business services over QoS enabled IP networks.

Introduction

The management of telecommunications services and networks has undergone a major paradigm shift over the last number of years. The liberalisation of the telecommunications marketplace has increased the level of competition at all levers of the service value chain; from network connectivity services through to the delivery of more advanced multimedia services.

The second major challenge has been the growth in importance of IP based services and associated emergence of e-commerce/e-business services. The collective global e-Commerce activity is estimated to exceed $6 trillion dollars in 2004. With the increased fragmentation of the telecommunications marketplace resulting in multiple service providers and connectivity providers the emphasis has shifted to an emphasis of managing network and services QoS in line with customer service expectations. However, a key element in successful delivery of customer services is the improved integration and management of the service value chains (i.e. management of business-2-customer (B2C) and business-2-business (B2B) chains). Current e-Business managed solutions, where available, tend to concentrate on only single aspects of the e-Business integration e.g. outsourced accounting management or virtual private network (VPN) services. Such e-Business services would support integrated management solutions (e.g. QoS, accounting, service level agreement, negotiation and management, virtual private network mgt etc), across the B2C and B2B value chain.

This paper proposes a management component framework to support the rapid and flexible construction of an e-Commerce management infrastructure. This management solution is based on a holistic management approach supporting seamless integration of network and application management services (i.e. vertical), as well as integrating management across distinct functional areas (i.e. horizontal). e.g. IP virtual private network provisioning, QoS and accounting/billing functional areas.

The target audience of this Framework is broad, covering service providers, management system integrators, and independent software vendors and standardising bodies. This Framework is being evaluated through a focus on the management of business-to-business services over QoS enabled IP networks.

Open Development framework

Overall, the FORM project is focussed on addressing the interdependencies between the System Development value chain and the Service Provision value chain. A key goal is to develop an architectural framework which can enable each player to develop management systems according to their own service management needs but in a way that is of common benefit.

The System Development chain must address the challenges of integrating separately sourced software to satisfy rapidly changing management system requirements. The software industry is moving towards the (re)use of component-oriented off-the-shelf software and model-driven approaches the software lifecycle. Applying this to the market for communication management software requires new architectural and modelling principles to be shared between Standards Bodies, Independent Software Vendors, System Integrators and System Customers (i.e. the Service Providers). In an open service market the Service Provision chain must support the need for different services provider to collaborate through seamless service management in a rapidly changing

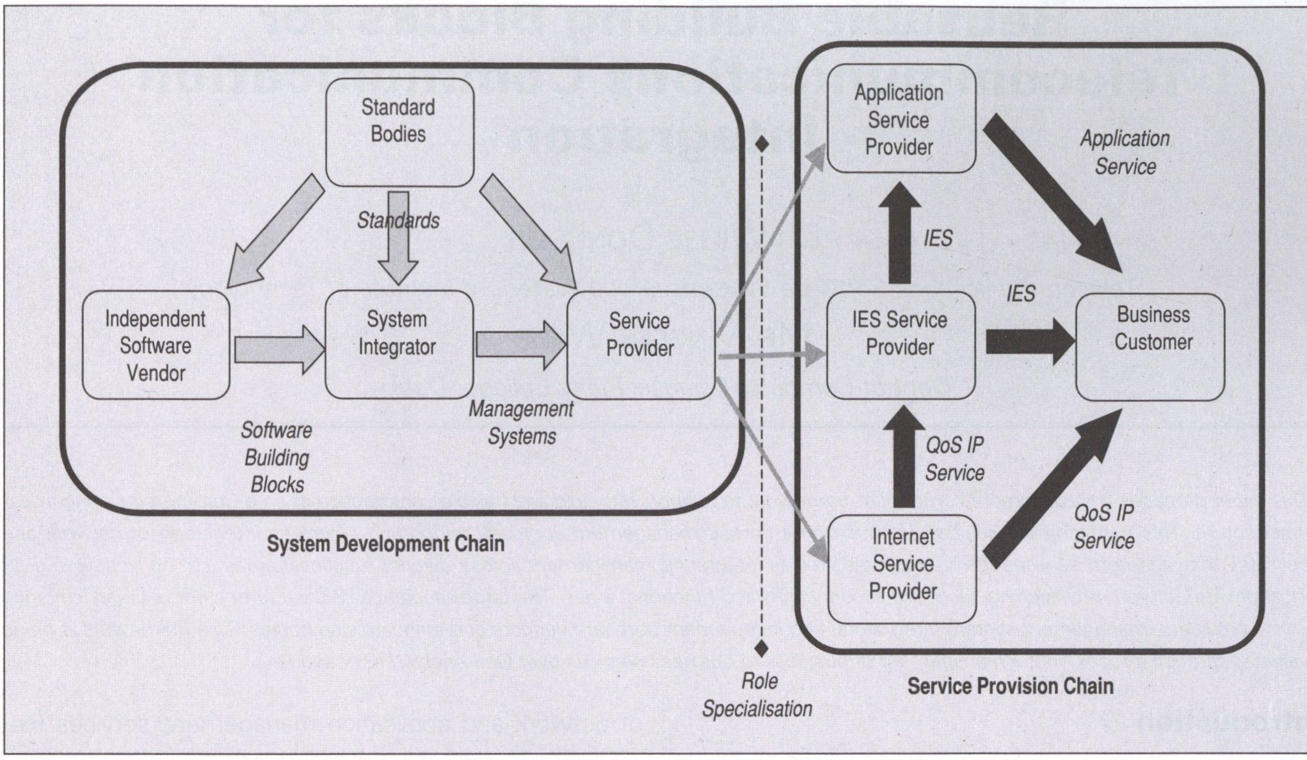


Figure 1. Stakeholder Model for FORM Open Development Framework

technological and service market environment. Enterprises require the ability to dynamically create, reconfigure and dissolve business collaborations. The FORM project explores this requirement through developing and implementing scenarios around the concept of Inter-Enterprise Service (IES) Provider environments. The IES scenarios examined investigate how an IES Provider may manage the dynamic relationships and the quality of service of electronic business interactions between groups of enterprises, the Application Service Providers (ASPs) they use and the Internet Service Providers (ISPs) providing communications infrastructure

Figure 1 depicts the overall stakeholder model for the Open Development Framework. The System Development chain represents the more generic set of stakeholders that drive the core set of generic architectural concepts in the Framework. It is intended that the architectural aspects of the Framework could be applied to any Service Provision chain where component-oriented management systems were required. The Service Provision chain represents the more specific IES related stakeholders used to exercise the Framework in FORM.

The Open Development Framework[D5] is therefore intended to provide guidance to these industry stakeholders in developing and applying software components and systems for the communications management sector and addressing these challenges. As depicted in Figure 2 the Framework is structured into four portions: Logical Architecture, Technology Architecture, a Development Methodology and a Set of Reusable Elements. The Framework is intended to be generic and extensible. In FORM, the generic part of the Framework is extended with others the concerns of the IES Management domain. However it is expected that other users of the Framework will extend the generic part to other domains, e.g. optical network management or mobile service management, and possibly reuse some of the IES Management specific part also.

The concerns addressed by the four portions of the Framework are:

- *Logical Architecture:* The Logical Architecture described the structural concepts of the Framework in a manner independent of any implementation technology. The core structural concept is the software Building Blocks (BB), which implements a number of Contracts via which inter-BBs interactions occur. Systems are described in terms of Business Processes and Business Roles. Reference Points (RP) exist between Business Roles and are decomposed in RP Segments. RP Segments are realised through Contract implementations. To promote reuse of Contract Specifications they are described in a technology neutral format. This involves specifying the information passed via a Contract through reference to an External Information Model.
- *Development Methodology:* The Development Methodology provides the processes and notations needed to develop Building Blocks and assemble systems that conform to the Framework. The primary modelling notation used is UML, though the potential of XML for Contract specification is also being examined. The methodology integrated a number of existing modelling techniques such as

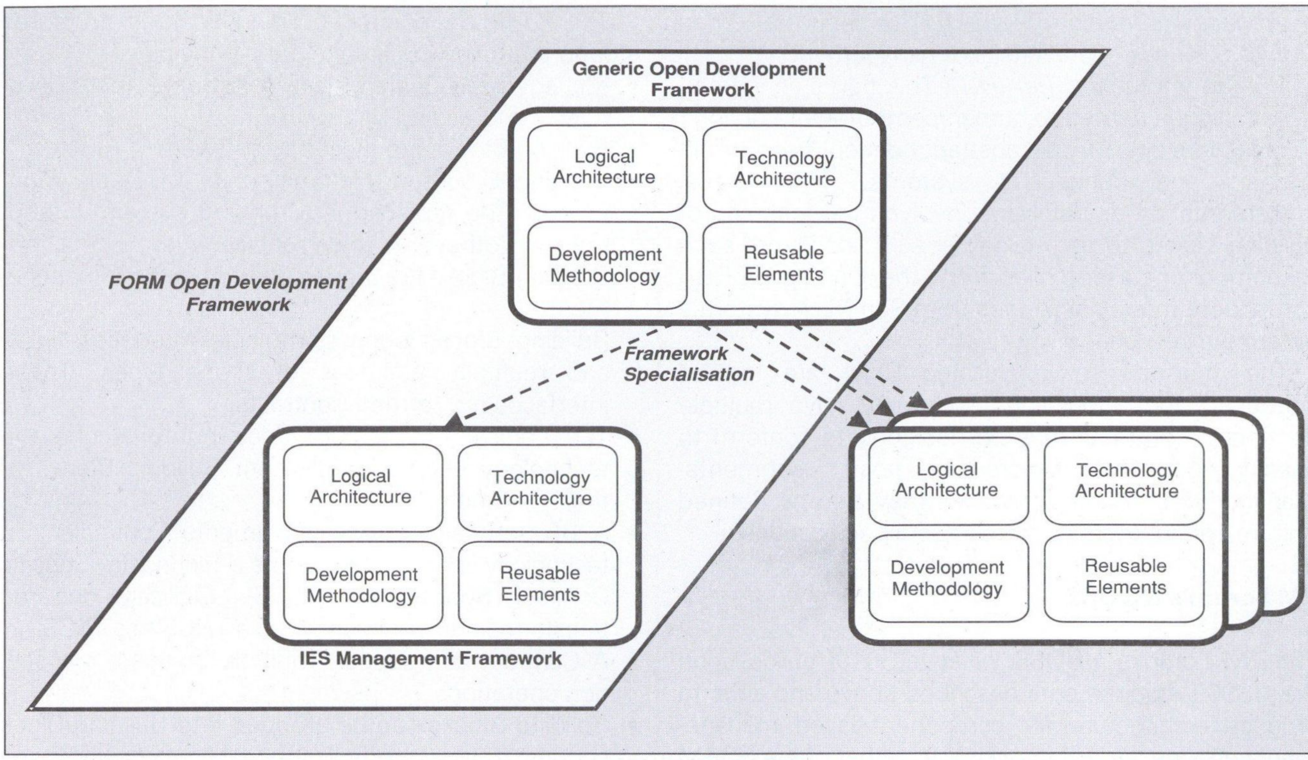


Figure 2. Structure of the Open Development Framework

use case modelling, business process modelling and analysis modelling plus the variety of other modelling approaches supported by UML. The Rational Unified Process is used as a partial template to integrate these techniques.

- *Technology Architecture:* The Technology Architecture addresses how the concepts expressed in the Logical Architecture can be implemented using a range of technologies. For each technology a single mapping between the meta-model for technology neutral Contract specification to the native meta-model of the technology is sought. Adoption to allow interoperability between Contracts implemented in different technologies is also addressed.
- *Reusable Elements:* This portion of the Framework is the repository for reusable products that result when the Framework is applied to a particular application domain, e.g. the IES Management domain addressed in FORM. The principle types of reusable entities are: Business Role definitions; Reference Point Segment specifications and their grouping into Reference Points; Contract specifications and their grouping into BB Specifications and BB implementations. Other types of reusable elements, such as policy and business process definitions are also being investigated.

Specification of Reusable components

The concept of software reusability is not new within the software industry. Centre to the emergence of object orientation software has been the potential for economies through object reuse and inheritance. The availability of "off-the-self" software solutions is rapidly becoming a reality in today's IT industry. However, in today's telecommunications industry the availability reusable software a component for application development is still an aspiration rather than a reality. In addressing the issue of reusable elements the FORM project has looked to a number of initiative in the TMForum workplan for guidance.

TM Forum Generic Requirements for Telecommunications Management Building Blocks

The most comprehensive definition of management-related building blocks currently available is being worked on in the TM Forum's Application Component Team (ACT)[G909]. This is based on the requirements established by Telcordia (formerly Bellcore) in their extensive OSCA/INA analysis. This has already seen some application in the use of INA principles in the work of TINA-C [Mulder]. The ACT however, has, further refined and condensed the OSCA/INA work in [gb909]. The concept of building blocks used here has been adopted for the framework. The ACT work forms the basis of the next generation OSS initiative currently being conducted in the TMF.

The ACT work identifies a "building block" as a deployable unit of interoperating software. In line with contemporary software architectures, building blocks are described as being in one of three computing tiers:

- *Enterprise Information Tier (EIT):* This tier is concerned with the storage and maintenance of enterprise data, i.e., data used by multiple business processes.

- *Process Automation Tier (PAT):* This tier is concerned with business operations and management.
- *Human Interaction Tier (HIT):* This tier is concerned with issues related to human/computer interaction

Building blocks are an abstract concept used in the analysis and design of systems. Their actual implementation as software, involves the mapping of building blocks to technology specific concepts such as component assemblies from the OMG's CORBA component model, or jar files from Sun Microsystem's enterprise java beans.

The interfaces of a Building Block are termed contracts, and a building block may have multiple contracts. Contracts may be designed to conform to specifications of reference point segments. Management system interfaces may also be defined in terms of aggregation of contract specifications

TM Forum's NGOSS

The TM Forum's NGOSS initiative[G910] builds upon the GB909 requirements described above and aims to produce architecture for component-based management systems, an aim very similar to that of the FORM Framework. As a result FORM has been closely tracking and contributing to this initiative. NGOSS is work in progress and at the time of writing there is no architecture document in the public domain. However, FORM has included several concepts from NGOSS into its Framework, namely:

- The explicit specification of information passed via Contracts in an external information model, termed a Shared Information Model in NGOSS.
- The specification of Contracts (and External Information Models) in a technology neutral manner, with mappings being developed to specific interface technologies.

FORM differs from the NGOSS approach as follows:

- The Contract in FORM is not specified as unit of business process as in NGOSS, and may contain several separate operations that may be invoked from process enactment engines, other Building Block or legacy system.
- The FORM Framework does not address the specification of abstract Framework Services of which Building Block implementation make use, as addressed in NGOSS.

Building Blocks General Principles

The FORM project has chose the concept of Building Blocks as defined by TMForum as a starting point for the definition of reusable elements. It has chosen to develop three trial systems addressing Fulfilment, Assurance and Accounting in line with the TMForum Business Process Model. The objective is to apply the principles for Building Block and Contracts definitions to try to identify components of reuse.

The FORM definition of building blocks is structured around the following architectural principles:

- Management Systems are constructed partially or fully from Building Blocks (BB).
- Building Blocks are pieces of software.
- Building Blocks are atomic units of Deployment (one can be replaced in a running system without requiring other BBs to be replaced).
- Building Blocks are atomic units of system management.
- Building Blocks support multiple interfaces types and multiple instances of those types. These interfaces are termed Contracts.
- The Logical Architecture does not prescribe the technology used to implement Building Blocks or their Contracts.
- A BB software release implements Contracts in a technology specific form, terms a Technology Specific Contract Specification (TSCS). Explicitly declared transforms are required to map a TNCS to a TSCS.
- A Contract can support multiple business transaction operations.
- Building Blocks can be grouped into Building Block Groups. This is the typical unit of purchase from a Software Vendor (though sale of individual BBs is possible as a group of size one).
- Building Block Groups must be released with documentation giving the accompanying business context, use cases, analysis model, TNCS and an External Information Model related to the BB and TSCS designs in that Group.
- Building Blocks possess behaviour linking interactions over different contracts. BB design may allow its behaviour to be modified at deployment or runtime. Where this feature is offered it should use a common, policy-based mechanism.
- Reference Points identify boundaries over which interactions occur between two Business Roles identified for a Business Domain (e.g. IES). Reference Points can be mapped to one or more Contract Specifications

FORM Guideline for Developing FORM Building Blocks

The design methodology has been identified as one of the 4 subcomponents of the FORM Framework. This methodology tackles the challenge of providing a component-oriented approach to the design of business process driven management systems for B2B and B2C. The FORM methodology proposes a business process driven approach to the construction of management systems solutions based on re-usable Building Blocks. FORM is developing a methodology consisting of two integrated development guidelines:

- To provide methodological guidelines for design of re-usable management components (termed BUILDING BLOCKS)

- To provide methodological guidelines for construction of B2B and B2C management solutions that implement management business processes using FORM Building Blocks.

The Guideline not only provides advice as to how to model Building Blocks, but also prescribes how such Building Blocks should be represented so as to ensure that the Building Blocks would be easily reusable by other actors (i.e. actors not involved in the development of the Building Block).

More specifically, the objectives of the guideline are to:

- Provide guidance as to Development activities in developing Building Blocks
- Identify artifacts[1] as to what notations to be used
- Identify artifacts to be developed
- Direct tasks/workflows of Building Block Development effort
- Prescribe combination of artifacts to characterise/ communicate usage of Building Block (i.e. re-use)

The approach taken in developing the guideline was to re-use current best practice in software Development and to customise/add features/artifacts where required. Therefore FORM did not develop its own development methodology from scratch, but rather to chose from the most widely accepted methodologies and then augment one as to suit its needs. Several candidate software development processes were reviewed and Rational's Unified Process was chosen. Several reasons underpinned this choice:

1. RUP is widely adopted in the OO software development community
2. RUP employs UML which is the modelling notation adopted widely in industry and standardisation fora
3. RUP is a 'development process framework' and thus is intended to be customised for the development of different specialised artifact
4. RUP claims to support component oriented as well as OO software development.
5. RUP Tool support was available to FORM Partners

Thus the FORM Building Block Guideline is a customisation of known industrial software methodologies and best practices based on experiences in previous research projects and academic work. The FORM Building Block Development Guidelines focuses on the Development of Building Block rather than the mapping of such building blocks into specific technologies or computing platforms.

Rational Unified (Development) Process (RUP)

The Rational Unified Process (RUP) is a software engineering process developed and marketed as a product by Rational Software. RUP is a specialisation of the Unified Software Development Process (USDP). It is delivered online using Web based technology and it consists of more than 1000 hyperlinked pages of text and graphics. It provides a proven disciplined (industrial) process to assigning tasks and responsibilities within a development organisation. RUP aims to capture many of the best practices in software development and then attempts to present them in a tailorable form that is suitable for a wide range of projects.

This section briefly identifies the important aspects of RUP which are utilized to describe the FORM Building Block Guidelines. RUP depicts software development in two dimensions, Phases (Inception, Elaboration, Construction Transition) and Workflows/Development activities (Business Modelling, Requirement Management, Analysis & Development, Test, Deployment[2]). Each Phase can have several iterations and involve some or all development activities. The RUP Phases and activities are depicted in Figure 3.

A development process usually describes who is doing what, how and when. The RUP uses four primary modelling representation elements:

- Workers: who
- Activities: how
- Artifacts: what
- Workflows: when

Building Block Design Guideline

The Building Block Development Guideline customised and extended several aspects of RUP. These extensions were necessary because of the unique domain of FORM, the research project nature of the development work, as well as it guidelines need to explicitly support the Development of management Building Blocks.

The customisations of RUP included:

- Development Phases & Iterations,
- Models to be developed,
- Artifacts to be developed,
- RUP documentation Templates.

Also several aspects were added to RUP, namely

I. An explicit architectural (structural) diagram, called the FORM reference architecture (termed the Business Role Architecture), which indicated reference points (boundaries) between management/administrative roles. This architectural diagram attempts to reconcile the FORM Building Blocks with a logical architecture of the FORM framework.

II. The explicit modelling of Building Blocks. This was considered fundamental to the FORM Framework and did not have a direct equivalent in RUP.

---

1 An artifact is a piece of information that is created, changed and used by actors when performing development activities. An artifact can be a model, a model element or a document.

2 RUP also defines activities for change/configuration management, Project Management and Environment.

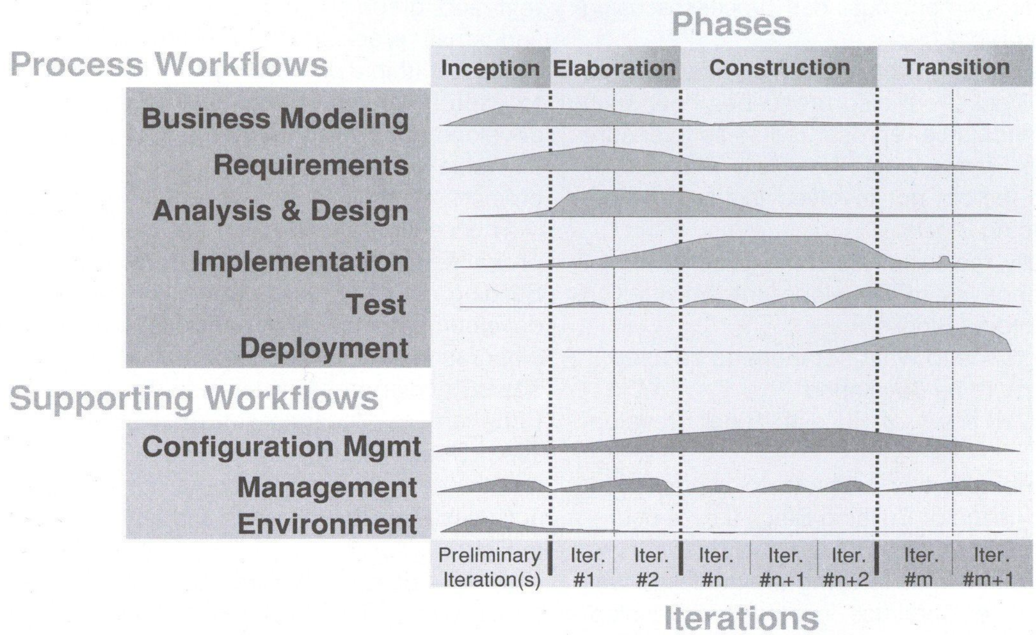


Figure 3. Rational Unified Process

The actors in the development of the FORM Building Blocks are

I. Domain Analyst (principally involved in Business Modelling and Requirements Management)
II. System Architect (principally involved in Reference Architecture development)
III. Building Block Developer (principally involved in analysis & Development activities and Building Block development).

FORM Development Phases & Iterations

FORM customised the RUP phases (Inception, Elaboration, Construction, Transition) to reflect both the goals of the (FORM) research project and the timescale of development available to the project. Only three of the phases of RUP were considered appropriate for the FORM project, namely Inception, Elaboration, and Construction. The following Figure 4 illustrates the FORM project lifecycle model.

It is important to notice that the effort distribution is rather different when compared with the unmodified RUP lifecycle model. This is natural when considering the activities of a research project like FORM which are centred on investigative and analytic activities.

FORM Inception Phase

The Inception Phase of the FORM development effort focused on establishing the boundaries of the Framework and creating a 'vision' document, which outlined the scope of the project. An initial reference architecture was devised for the Framework. Also developed were the business model(s) for the Framework and the initial Use Cases for the functional areas being investigated by the project (namely Fulfilment, Accounting, Billing). During this phase, terminology and the selection of relevant standards/models were identified e.g. IPDR for Accounting, DMTF CIM and IEFT QoS information models for Assurance.

An important aspect of this phase was the initial identification of potential Building Blocks. The method by which the Building Blocks were identified and refined is presented in the next section (concerning Development activities).

The key outcomes for this phase were:

- Vision Documents indicating the scope, context and objectives FORM Development Framework
- Business models, actors and identification of functional domains (areas) of the Framework
- Initial (logical) Reference Architecture with reference points and domain boundaries
- Requirements capture & management
- Use Cases for three functional domains (F-A-B)
- Plan of how development work will proceed

FORM Elaboration Phase

The Elaboration Phase focused on the revision and refinement of the FORM Vision document(s) In this phase the FORM reference architecture was solidified, with testing of proposed computing and testbed

A 2001-ben megjelent számaink magyar tartalomjegyzéke

Content of the year 2001.

2001/7

2001/8

2001/9

Szerzők szerinti cikklista

M

N

O

P

R

S

Sz

Angolnyelvű számok

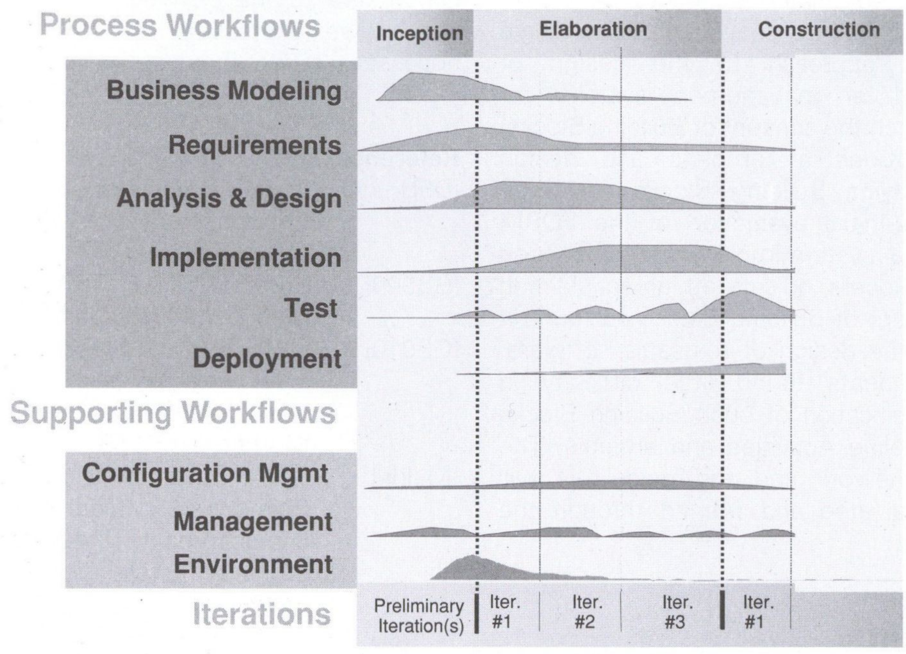


Figure 4. Customised RUP Phases

infrastructures. Also during this phase, the use cases to be supported by the Building Blocks were identified and refined. The concept of Building Block was further refined and a Building Block template specified. This template defined the essential UML models used to characterise/describe the building blocks.

The main results of the elaboration phase are:

- To provide stable system architecture model
- To provide initial Development models for Building Blocks in the FORM Framework

FORM Construction Phase

The construction phase is focused on prototyping the essential aspects of FORM Building Blocks and infrastructure. This culminated in the FORM 1st Trial. Only a limited set of Building Blocks were developed in this phase. In a typical product development project this phase would have many iterations. However, because FORM is a research project and the construction of the Building Blocks at this stage in the project were intended to investigate new technologies and approaches to development, this phase as only iterated once. The milestone at the end of this phase in FORM was the 1st Trial and its evaluation.

The main results of the Construction Phase (1st iteration) are:

- Development of test prototypes to evaluate and validate aspects of the FORM architecture
- Prototyping (partial functionality) of several Building Blocks within functional domains of FORM

Transition Phase

The Transition phase was not considered by the project as this phase is focused on the transition of robust, almost complete prototypes into fully deployed products. Such a phase is outside the scope of EU research project. Also, FORM was deliberately investigating and researching the most risk centred aspects of the Framework, rather than trying to implement a fully functional or complete Framework.

Iterating the FORM Phases

FORM intends to iterate the three phases during the development effort for its next trial phase which has now commenced. The importance of the second iteration of the phases is that it allows a deeper investigation of the application areas within which the Building Blocks are being developed. Also, the Building Block concepts and architecture will be extended and refined.

Conclusion

FORM is developing an Open Development Framework to guide the developers of component-based management systems. This Framework is based on experienced garnered from previous EU projects and standardisation work on component-oriented or management-related architectures.

This Framework tackles the challenge of providing a component-oriented approach to the design of business process driven management systems for B2B and B2C based on the concept of Building Blocks. The Guideline provides a template and design activities for specifying Building Blocks and their associated models. Initial evaluation of the FORM guidelines by the trial system implementers has been positive. Most developers indicate a high satisfaction level with the concept of Building Block and found it useful in assisting the design of re-usable software. Also, most respondents found little difficulty in generating the specification of such Building Blocks based on the Guideline Activities and artifacts. The overall FORM framework and the guidelines will continue to be evaluated and refined through the FORM trail systems.

Acknowledgment

This work was conducted under the partial funding of the EU through the IST project FORM (contract IST-1999-103571). I would like to thank my colleagues in Form and especially Dr. Dave Lewis and who contributed to the development of this framework. The views expressed in this document do not necessarily reflect those of the consortia.

References

[D5] Initial specification of FORM Inter-Enterprise Management Framework http://www.uhc.dk/form/index.htm

[GB909] SMART TMN Technology Integration Map, GB 909, Issues 1.1, TMF, Oct 1998

[GB910] NMF Telecoms Operation Map: A high-level view of end-to-end service fulfilment, service assurance and billing, NMF, Morristown, 1998

[Mulder] TINA Business Model and Reference Points, v4.0, Mulder, H. (ed), TINA baseline document, TINA-C, 1997

Neural Network Based Scalar and Isotropic Vector Hysteresis Operator

MIKLÓS KUCZMANN, AMÁLIA IVÁNYI

Department of Electromagnetic Theory, Budapest University of Technology and Economics

L

The classical Preisach model is one of the most generally applied simulation techniques to simulate the behaviour of magnetic materials, to describe hysteresis phenomena. According to the theory of Weiss, the classical Preisach model assumes that the ferromagnetic materials consist of many elementary interacting domains, and each of them can be represented by a rectangular elementary hysteresis loop. The fundamental concepts of the Preisach model is that different domains have some distribution, which can be described by a distribution function, also called Preisach kernel. Feedforward type artificial neural networks are able to approximate any kind of nonlinear, continuous functions represented by its discrete set of measurements. It is based on the so-called Kolmogorov-Arnold theorem. A neural network (NN) based scalar hysteresis model has been constructed on the function approximation ability of NNs and if-then type rules about hysteresis phenomena. Vectorial generalization to describe isotropic magnetic materials in two and three dimensions with an original identification method has been introduced in this paper. Results are illustrated in figures.

Keywords: hysteresis characteristics, Everett surface, vector hysteresis, feedforward type neural networks, backpropagation training method.

Introduction

Simulation of hysteresis characteristics of magnetic materials has to be included in electromagnetic field simulation software tools to predict the behavior of different type of magnetic equipments. Hysteresis characteristics of magnetic materials can be described by a nonlinear, multivalued relation between the magnetic field intensity $H(t)$ and the magnetization $M(t)$. That is called the hysteresis operator. Many hypotheses have been developed since the last period of magnetic research, as the classical Preisach model [1, 2, 3, 11], the Jiles-Atherton model [1], the Stoner-Wohlfarth model [5] and some approach based on NNs [7, 8, 9].

A hysteresis operator with continuous output built on the function approximation ability of feedforward type NNs and its vectorial generalization for isotropic magnetic materials in two and three dimensions have been introduced in this paper [12-20]. The applicability of the method is illustrated in figures.

Construction of the NN model of hysteresis is based on the first order reversal curves. A method for the formulation of the Everett surface from symmetric minor loops also has been introduced.

The scalar hysteresis operator

The developed NN model of scalar hysteresis characteristics consists of a system of two feedforward-type NNs with bipolar sigmoid transfer functions and a heuristic if-then type knowledge-base about the hysteresis phenomena.

The classical scalar Preisach model has been used to generate different types of training data sets. The virgin curve and a set of the first order reversal branches are available. It has been found that, it is enough to use only the descending (or ascending) branches (Fig.1.), because of the symmetry of hysteresis characteristics. In practice, these data sets can be replaced by measurements [12-20].

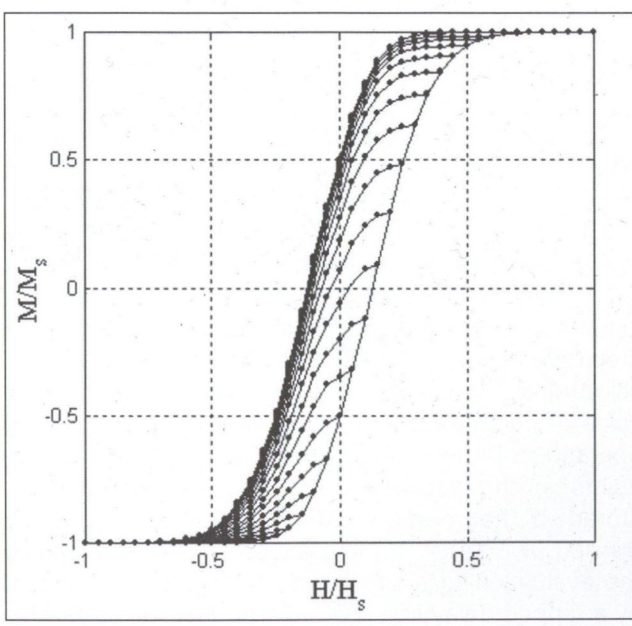


Figure 1. A set of the descending branches

Hysteresis characteristics is a multivalued function, it means difficulties when using feedforward-type NNs. If a new dimension is introduced to the measured and normalized transisition curves, multivalued function results a single-valued surface. The downgrade part of hysteresis characteristics can be described at a turning point $H_{tp}^{(desc)}$ with a negative real parameter $\xi^{(desc)}$, determined as

$$\xi^{(desc)} = -(1 + H_{tp}^{(desc)})/2 \qquad (1)$$

The effect of this preprocessing technique can be seen in Fig.2 for descending curves. After preprocessing, function approximation is worked out by feedforward NNs trained by the Levenberg-Marquardt backpropagation method [10]. The anhysteretic curve with 41 training points can be approximated by a NN with 8 neurons in one hidden layer, and the preprocessed first order reversal branches (about 500-600 data pairs) are approached by a NN with 7, 11 and 6 processing elements in 3 hidden layers. Training of NNs takes about twenty minutes ($MSE = 5.10^{-6}$) on a Celeron 566MHz computer (192Mbyte RAM), using the Neural Network Toolbox of MATLAB.

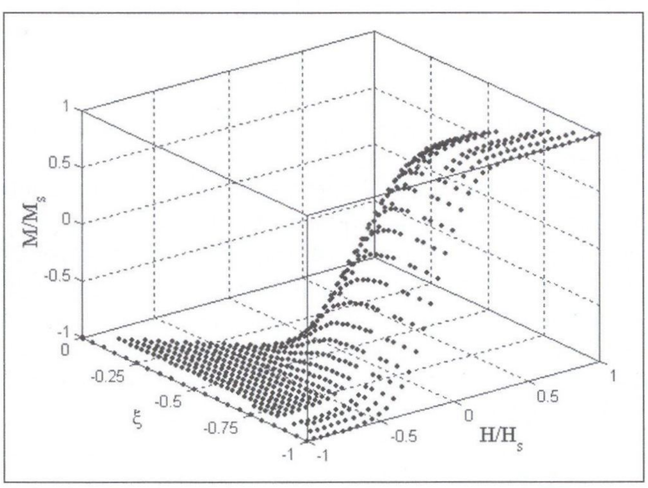


Figure 2. Descending revarsal curves after preprocessing

Applying NNs, relationship between magnetization M and magnetic field intensity H can be performed in analytical formula, $M = \phi\{H\}$.

Memory mechanism of magnetic materials is realized by an additional algorithm based on heuristics. It is the knowledge-base for the properties of hysteresis phenomena. Magnetization at a simulation step responsed by the NN is constructed on the actual value of the magnetic field intensity, the appropriate value of parameter ξ and a set of turning points. The operation of the model is built on a set of turning points in the ascending and descending branches, stored in the memory, which is a matrix with the division $[H_{tp}, M_{tp}, \xi]^T$. Turning points can be detected by the evaluation of a sequence of $\{H_{k-1}, H_k\}$, generated by a tapped delay line (TDL). After detecting a turning point $H_{tp} = H_{k-1}$ and storing it in the memory, the aim is to select an appropriate transition curve for the detected turning point.

Conditions are collected in the selection rules, to choose the suitable NN. After detecting a turning point, generally denoted by $H_{START} = H_{tp}$, the algorithm for first order minor loops can be summarized follows. If $\mathrm{MATRIX}^{(desc)}$ ($\mathrm{MATRIX}^{(asc)}$) has more columns and magnetic field intensity is increasing (decreasing) at the k^{th} simulation step, the actual minor loop must be closed at the last stored value of $H_{GOAL} = H_{tp}^{(desc)}$ ($H_{GOAL} = H_{tp}^{(asc)}$) which can be found in the last column of $\mathrm{MATRIX}^{(desc)}$ ($\mathrm{MATRIX}^{(asc)}$). Denote this column of the appropriate MATRIX with C. The value of normalized magnetization M_k responsed by the neural model at H_{GOAL} must be equal to $M_{tp}^{(desc)}$ ($H_{tp}^{(asc)}$)in the C^{th} column. It is the condition for closing the minor loops, can be satisfied by the correction $\eta = \eta(H)$. After closing a minor loop, the appropriate columns must be erased.

The block representation of the model can be seen in Fig.3.

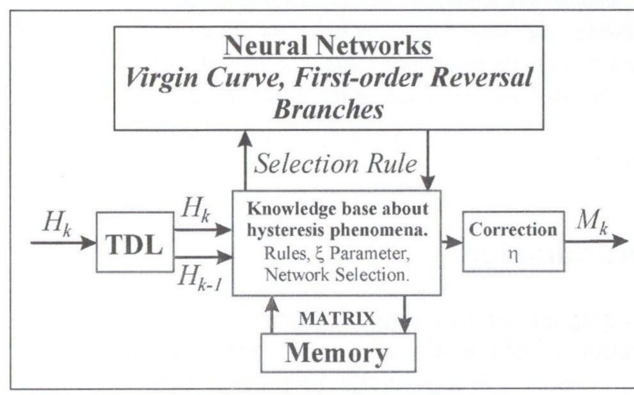


Figure 3. Block representation o the scalar neural network model of hysteresis phenomena

The experimented NN model of hysteresis can be used as a continuous scalar model to simulate the behavior of magnetic materials. Two kind of hysteresis characteristics predicted by the developed model have been compared with the results of the Preisach model as plotted in Fig.4.

Accommodation property also can be simulated, when $H_k = H_k + \alpha M_{k-1}$ is applied as input of the model, where α is the moving parameter [1,11]. An illustration, generated by the NN model can be seen in Fig.5, where $\alpha = 0.001$.

In electromagnetic field calculation software it is favorable to use the Newton-Raphson iteration technique. It requires the value of differential susceptibility, $\chi_{diff} = dM/dH$, which can be performed in analytical form by the chain rule when applying NNs, $dM/dH = d\phi\{H\}dH$

Everett surface from minor loops

The Everett surface can be calculated from measured symmetric minor loops as well. Knowing the Everett surface, magnetization can be formulated as

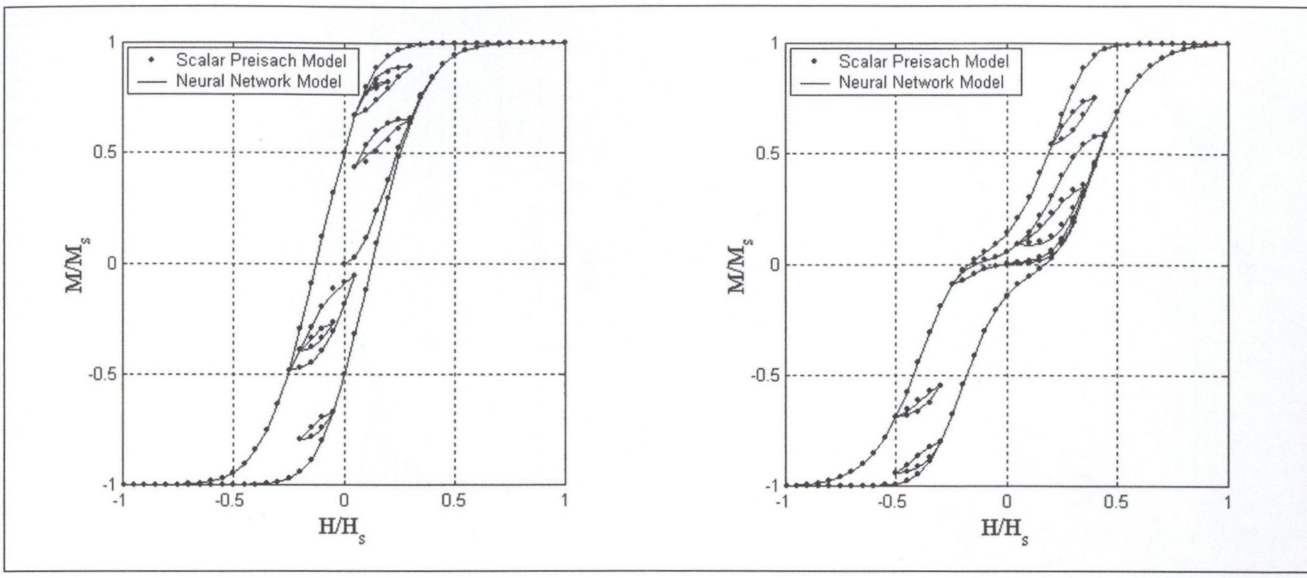


Figure 4. Comparison between neural scolar model and the classical Preisach model

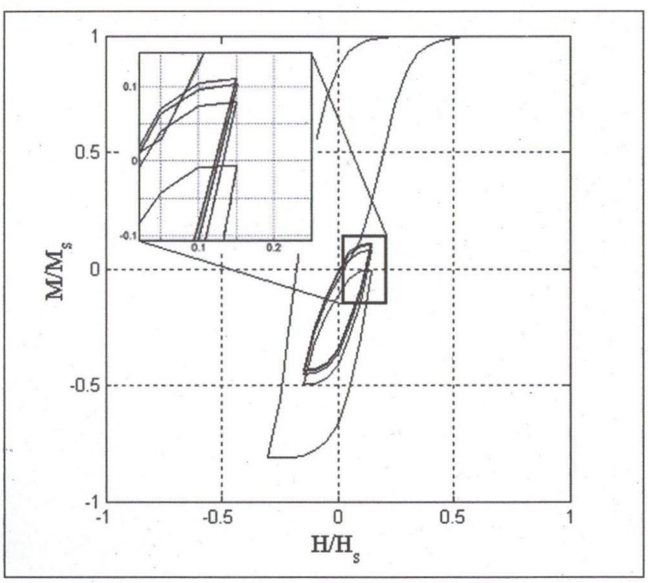


Figure 5. Illustration for accommodation

$$M(t) = -F(\alpha_0, \beta_0) + 2\sum_{k=1}^{n}\left(F(M_k, m_{k-1}) - \right.$$
$$\left. - F(M_k, m_k)\right), \quad (2)$$

where $F(\alpha,\beta)$ is the Everett surface, $\alpha_0 = 1$, $\beta_0 = -1$, and $\{M_k\}_{k=1}^{n}$, $\{m_k\}_{k=1}^{n}$, are the increasing and decreasing sequences of the applied magnetic field intensity [11].

Values in the Everett table, according to $\alpha = 1$ can be calculated as the well known relation

$$F(\alpha, \beta) = \frac{1}{2}\left(M_\alpha - M_{\alpha,\beta}\right), \quad (3)$$

where M_α is a reversal magnetization point in the major hysteresis loop according to the magnetic field intensity α, and $M_{\alpha,\beta}$ is a magnetization value in a reversal curve starting from the reversal point (α, M_α). Illustration for the calculation of Everett surface can be seen in Fig.6.

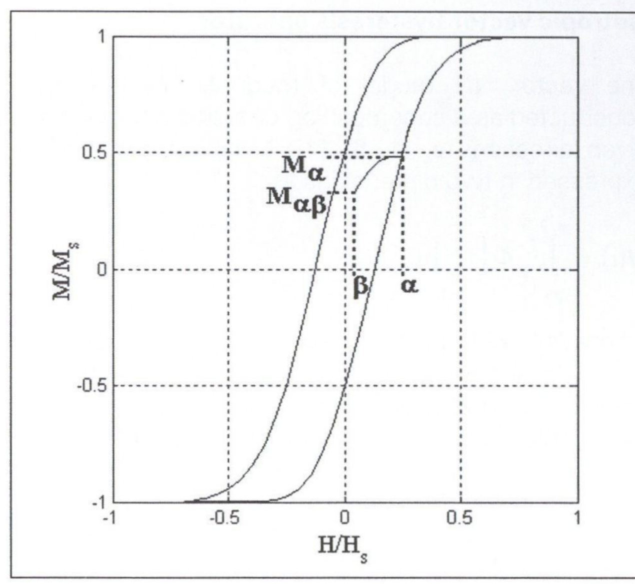


Figure 6. Calculation of the Everett surface

Starting from demagnetized state, increasing the magnetic field intensity while $\alpha = M_K$ ($K = N-1,...,N/2$, where N is the size of the Everett table), (2) can be reformulated as

$$M(t) = -F(\alpha_0, \beta_0) + 2\sum_{k=1}^{n_1}\left(F(M_k, m_{k-1}) - \right.$$
$$\left. - F(M_k, m_k)\right) + 2\left(F(M_K, m_{K-1}) - \right.$$
$$\left. - F(M_K, m_K)\right) \quad (4)$$

According to the definition of the Everett surface, $F(M_K, m_K) = 0$, if, $\alpha = M_K$, so $F(M_K, m_{K-1})$ can be calculated. Otherwise $F(M_K, m_K)$ is the only unknown in (4).

In Fig.7 an example for symmetric minor loops and the corresponding descending branches (points) calculated from the identified Everett surface with comparison with known reversal curves (dashed lines).

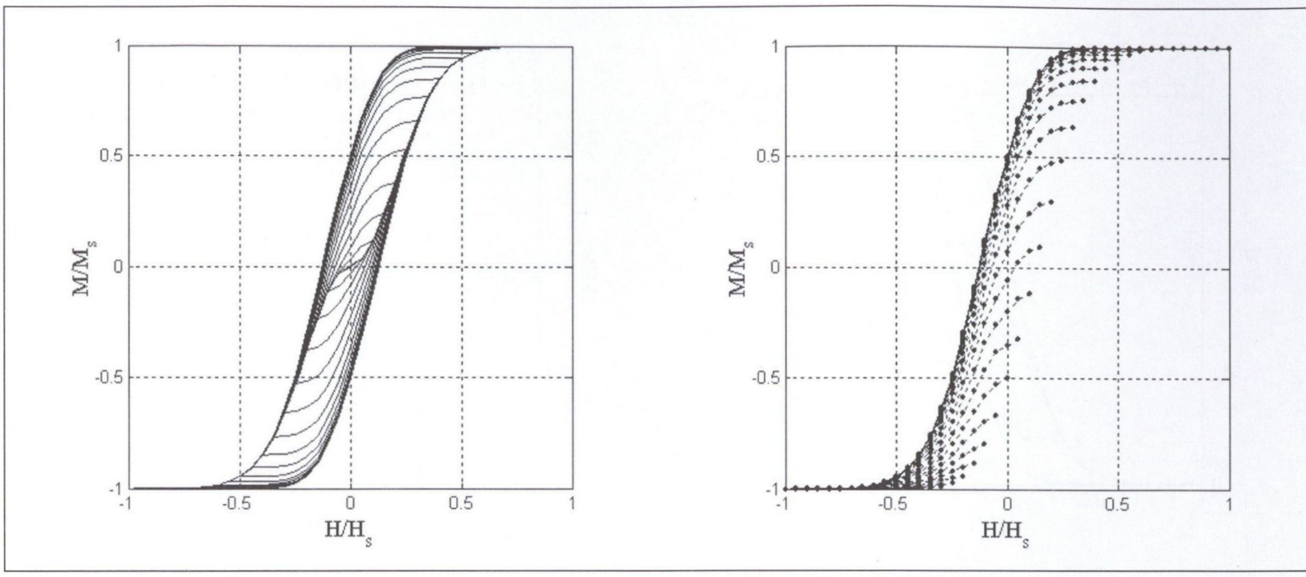


Figure 7. Symmetric minor loops and the corresponding reversal curves

Isotropic vector hysteresis operator

The vector NN model of magnetic hysteresis is constructed as a superposition of scalar NN models in given directions e_ϕ [1, 6, 11]. Magnetization can be expressed in two dimensions as

$$M(t)=\int_{-\pi/2}^{\pi/2} e_\varphi\,\Phi\{H_\varphi\}d\varphi\,, \tag{5}$$

where $M_\phi = \phi\{H_\phi\}$ is the magnetization in the direction e_ϕ, $H_\phi = |H|\cos(\vartheta_H-\phi)$ and ϑ_H is the direction of magnetic field intensity H. NNs $\phi\{H_\phi\}$ depend on the polar angle ϕ if the magnetic material presents anisotropy, otherwise it is independent on it. Firstly, isotropic case has been analyzed. In computer realization it is useful to discretize the interval $[\pi/2,\pi/2]$ as $\phi_i = -\pi/2,+(i-1)\pi/n$, $i = 1,\dots,n$, (Fig.8) and n is the number of directions.

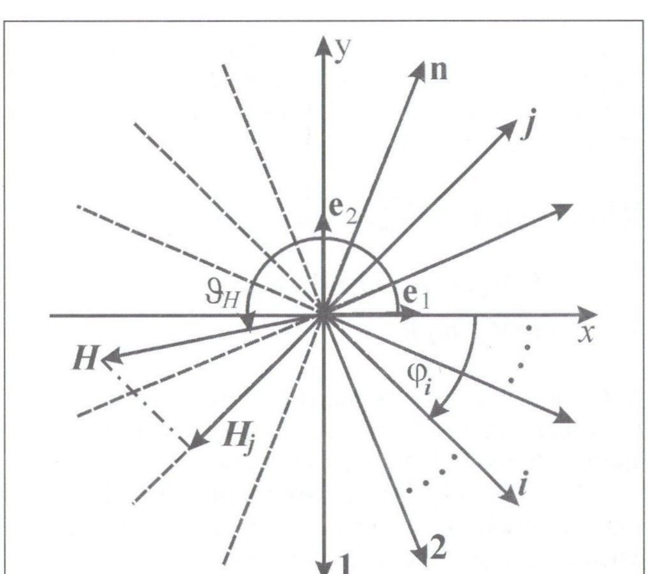


Figure 8. Definition of directions in two dimensions

In three dimensions a similar expression can be obtained,

$$M(t)=\int_{-\pi/2}^{\pi/2}\int_{0}^{\pi} e_{\vartheta,\varphi}\,\Phi\{H_{\vartheta,\varphi}\}d\vartheta\, d\varphi\,, \tag{6}$$

where

$$H_{\vartheta,\varphi}=\begin{bmatrix}a_1 & a_2 & a_3\end{bmatrix}_{\vartheta,\varphi}\begin{bmatrix}H_x & H_y & H_z\end{bmatrix}^{\mathrm{T}}, \tag{7}$$

and the directions are given as $a = a_1e_1 + a_2e_2 + a_3e_3$, $|a| = 1$.

After measuring the Everett surface in the x direction, the following expression can be obtained between the scalar Everett surface $F(\alpha,\beta)$ and the vector Everett function $E(\alpha,\beta)$:

$$F(\alpha,\beta)\cong\sum_{i=1}^{n}\cos\varphi_i\,E(\alpha\cos\varphi_i,\beta\cos\varphi_i)\,. \tag{8}$$

In isotropic case the vector Everett surface is unique for all directions. Expression (8) can be solved numerically, the algorithms is given as follows. Expression (8) can be rewritten in the form,

$$F(\alpha_k,\beta_l)\cong\sum_{i_1=1}^{n_1}\cos\varphi_{i_1}\,E\left(\alpha_k\cos\varphi_{i_1},\,\beta_l\cos\varphi_{i_1}\right)+ \\ +\sum_{i_2=1}^{n_2}\cos\varphi_{i_2}\,E\left(\alpha_k\cos\varphi_{i_2},\,\beta_l\cos\varphi_{i_2}\right), \tag{9}$$

where $\alpha_k = 2(_k-(N+1)/2)/N$, $\beta_l = 2((N+1)/2-l)/N$, $k,l = 1,\dots,N+1$ and the size of Everett table is $(N+1)\times(N+1)$. Expression (9) contains n_1 known and n_2 unknown points in the Everett surface. If $\beta_l = 0$, $l = 1,\dots,N+1$ and assuming linear interpolation in the surface $E(\alpha,\beta)$,

$$F(\alpha_k,0)=\sum_{i_1=1}^{n_1}\cos\varphi_{i_1}\left(E(\alpha_{j-1},0)+\left(E(\alpha_j,0)-\right.\right.$$

$$-E(\alpha_{j-1},0))(\alpha_k\cos\varphi_{i_1}-\alpha_{j-1})/(\alpha_j-\alpha_{j-1}))$$
$$+E(\alpha_{k-1},0)((1+b_k)c_1-a_kc_2)+E(\alpha_k,0)(a_kc_2-b_kc_1), \tag{10}$$

where $c_1=\sum_{i_2=1}^{n_2}\cos\varphi_{i_2}$, $c_2=\sum_{i_2=1}^{n_2}\cos^2\varphi_{i_2}$,

and $a_k = \alpha_k/(\alpha_k-\alpha_{k-1})$, $b_k = \alpha_{k-1}/(\alpha_k-\alpha_{k-1})$. From (10), value of $E(\alpha_k,0)$ can be expressed. A similar relation can be obtained, when $\alpha_k = \beta_k$.

If $\beta\neq 0$, a similar mathematical formulation can be obtained. Firstly, let us assume that, $(\alpha_k\cos\phi_{i_1},\beta_l\cos\phi_{i_1})$ blcos is bounded by known points $A(x_1,y_1,z_1)$, $B(x_2,y_2,z_2)$ and $C(x_3,y_3,z_3)$ in the vector Everett surface. The value of Everett surface in this coordinate can be expressed assuming linear interpolation in the given triangle. Unknown values can be expressed after some poor mathematical formulations.

It is enough to calculate the half of the Everett surface because of symmetry.

First order reversal curves of vector NN model can be calculated from the identified vector Everett surface.

Let us assume that, measured hysteresis curve and the corresponding Everett surface is given in the x direction as plotted in Fig.9.

Simulation results for reversal curves obtained from the identified Everett surface in two dimensions (20 directions) and three dimensions (18 directions) can be seen in Fig.10 and Fig.11. Directions of the three dimensional model are generated by an icosahedron (Fig.12).

In electromagnetic field analysis software it is useful to apply the Newton method to solve the nonlinear partial differential equations. It requires the Jacobian matrix of the disretized form of (5) in two dimensions,

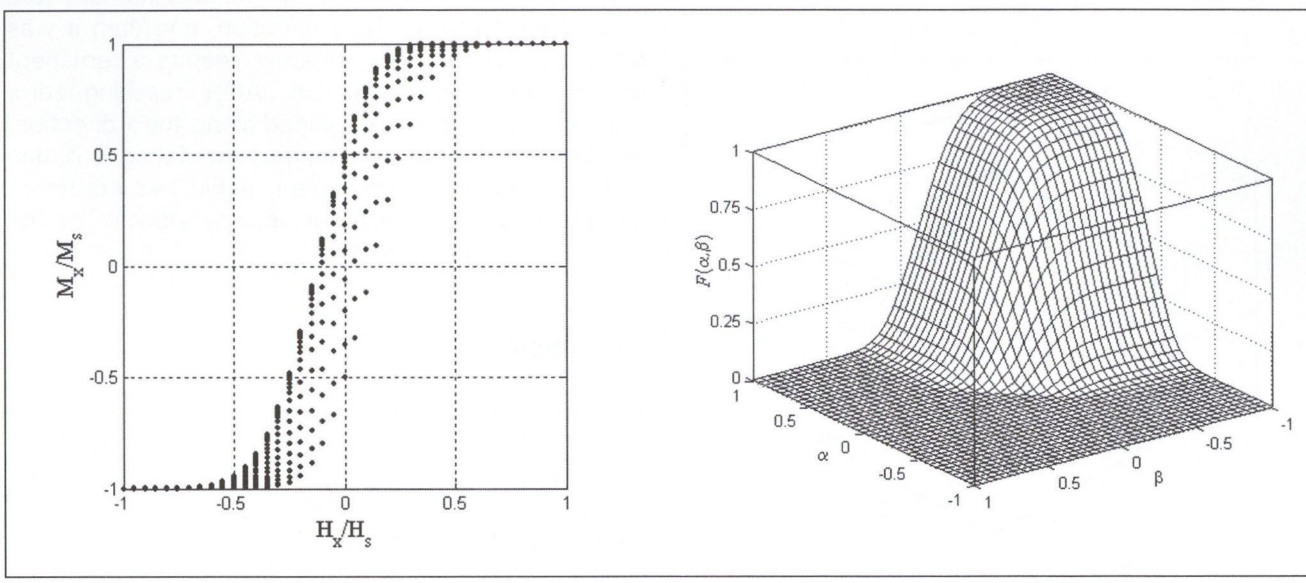


Figure 9. Hysteresis curve and the Everett surface mearured in the x direction

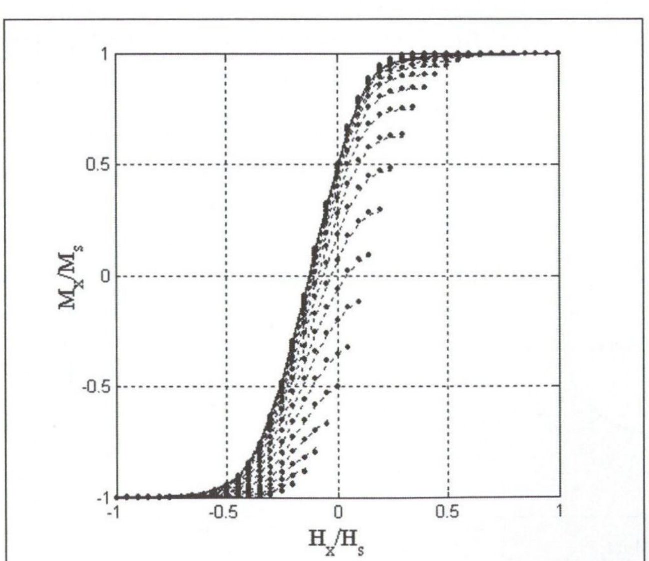


Figure 10. Indentification result in two dimensions

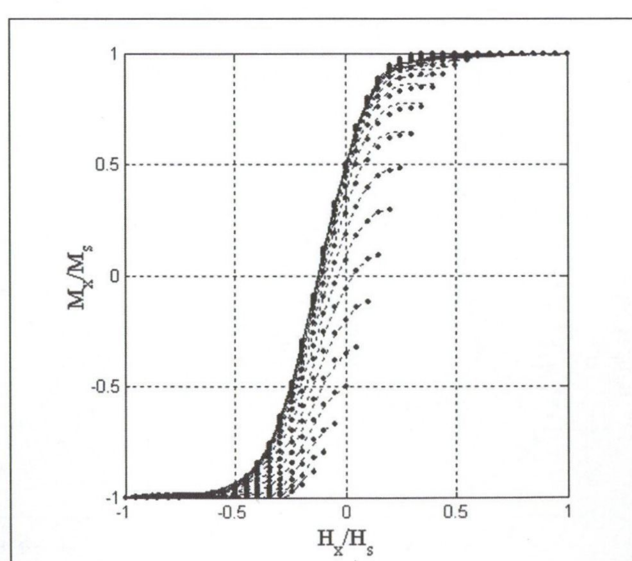


Figure 11. Indentification result in three dimensions

$$\frac{\partial M}{\partial H}=\frac{\pi}{2n}\sum_{k=1}^{n}\frac{d\Phi\{H_{\varphi_k}\}}{dH_{\varphi_k}}\begin{bmatrix}\cos^2\varphi_k & \sin\varphi_k\cos\varphi_k\\ \sin\varphi_k\cos\varphi_k & \sin^2\varphi_k\end{bmatrix}, \quad (11)$$

where $d\phi\{H_{\phi_k}\}/H_{\phi_k}$ can be calculated analytically, when applying NNs. A similar expression can be formulated in three dimensions,

$$\frac{\partial M}{\partial H}=\lambda\sum_{k=1}^{n}\frac{d\Phi\{H_{\vartheta_k,\varphi_k}\}}{dH_{\vartheta_k,\varphi_k}}$$

$$\begin{bmatrix}\cos^2\xi_k & \cos\xi_k\cos\zeta_k & \cos\xi_k\cos\psi_k\\ \cos\xi_k\cos\zeta_k & \cos^2\zeta_k & \cos\psi_k\cos\zeta_k\\ \cos\xi_k\cos\psi_k & \cos\psi_k\cos\zeta_k & \cos^2\psi_k\end{bmatrix} \quad (12)$$

where ξ_k, ζ_k and ψ_k are the angles from the x, y and z directions. Parameter λ is depending on the number of directions, if using icosahedron, because directions are not uniform.

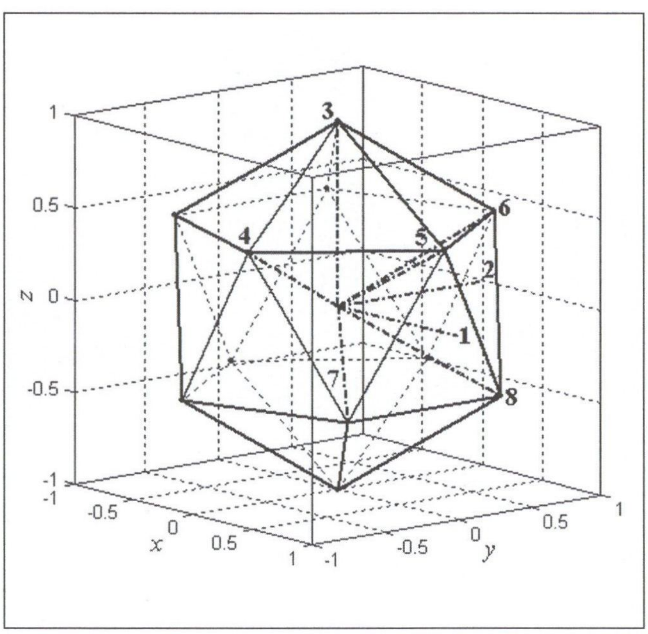


Figure 12. Definition of directions in two dimensions

Some properties of the NN vector isotropic model

Applying rotational magnetic field intensity with different amplitude and with linearly increasing amplitude, output of the 2D vector NN model has been plotted in Fig.13. a, b and c. In Fig.13.c, the vector of magnetization gradually approaches the regime of

$$H(t)=\{H_x(t)=H_m\cos(\omega t),\ H_y(t)=H_m\sin(\omega t)\},$$

uniform rotation. The variation of magnetic field intensity is as follows,

(13)

where H_m is the amplitude of magnetic field strength and ω is the angular velocity [11].

Let us suppose that the magnetic field intensity was first increased along the y direction, and then it was decreased to zero. This process results a remanent magnetization in y direction. After reaching zero, magnetic field intensity is varied along the x direction. The orthogonal remanent component of magnetization can be reduced as it can be seen in Fig.14 for different remanent values. It is an anisotropy induced by the magnetic prehistory [4,11].

Conclusions

A NN model for magnetic hysteresis based on the function approximation ability of NNs has been experimented. The anhysteretic magnetization curve and a set of the first order reversal branches must be measured on a real magnetic material. Introducing an additional parameter ξ solves a fundamental problem of simulating hysteresis characteristics, that is the multivalued property. The magnetization becomes a single valued function of two variables and an if-then

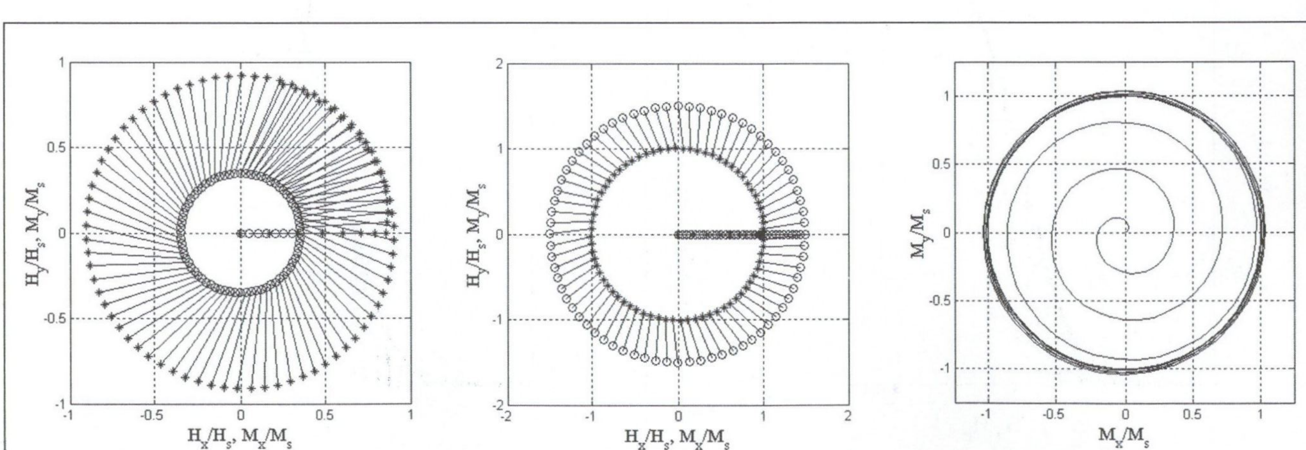


Figure 13. Simulated H and M loci for different conditions

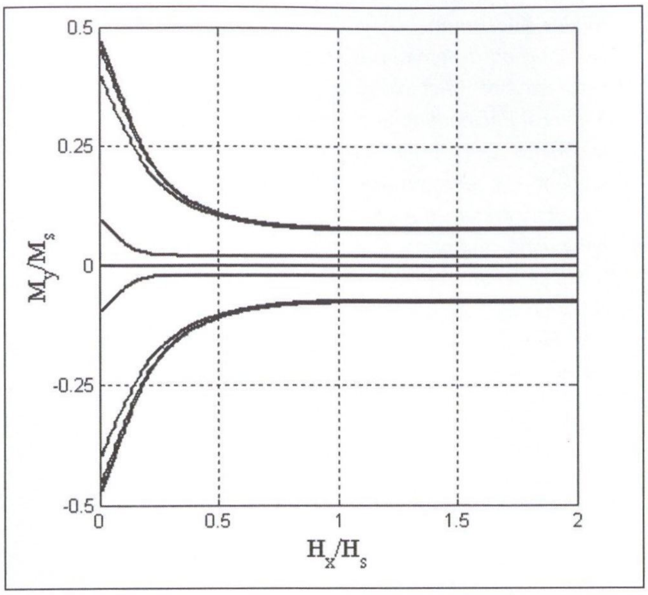


Figure 14. Induced anisotropy

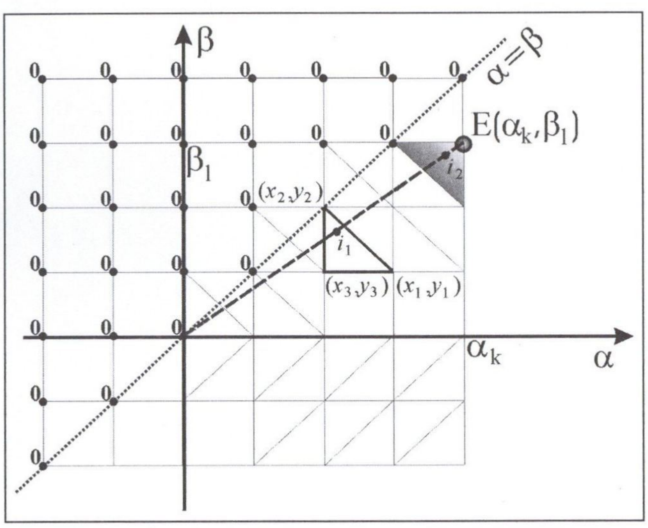


Figure 15. Illustration for the identification process

type knowledge-base can be used for simulating different phenomena of magnetic materials.

This method has been generalized in two and three dimensions with an original identification process. Vector model is based on the Mayergoyz type technique, but the identical scalar models are constructed on the identified scalar NN model of hysteresis.

Acknowledgement. The research work is supported by the Hungarian Scientific Research Fund, OTKA 2002, Pr. No. T 034 164/2002.

References

1. A. Iványi: Hysteresis Models in Electromagnetic Computation, Akadémia Kiadó, Budapest, 1997.
2. J. Füzi: Parameter Identification in Preisach Model to Fit Major Loop Data, in Applied Electromagnetics and Computational Technology, in Series of Studies in Applied Electromagnetics and Mechanism, vol.11, IOS Press, 1997, pp.77-82.
3. J. Füzi–E. Helerea–D. Oltenau–A. Iványi–H. Pfützner: Experimental Verification of a Preisach-Type Model of Magnetic Hysteresis, Studies in Applied Electromagnetics and Mechanics, Vol. 13, Non-linear Electromagnetic Systems, 8th ISEM Conference Braunschweig, IOS Press, 1997, pp.479-482.
4. J. Füzi–A. Iványi: Isotropic vector Preisach particle, Physica B, vol. 275, 2000, pp.179-182.
5. Zs. Szabó–A. Iványi: Computer-Aided Simulation of Stoner-Wohlfarth Model, Journal of Magnetism and Magnetic Materials, vol. 215-216, 2000, pp.33-36.
6. C. Ragusa–M. Repetto: Accurate Analysis of Magnetic Devices with Anisotropic Vector Hysteresis, Physica B 275, 2000, pp.92-98.
7. C. Serpico–C. Visone: Magnetic Hysteresis Modeling via Feed-Forward Neural Networks, IEEE Trans. on Magn., vol.34, 1998, pp.623-628.
8. A. A. Adly–S. K. Abd-El-Hazif: Using Neural Networks in the Identification of Preisach-Type Hysteresis Models, IEEE Trans. on Magn., vol.34, 1998, pp.629-635.
9. A. A. Adly–S. K. Abd-El-Hafiz–I. D. Mayergoyz: Identification of Vector Preisach Models from Arbitrary Measured Data Using Neural Networks, Journal of Applied Physics, Vol.87, No.9, 2000, pp.6821-6823.
10. C. Christodoulou–M. Georgiopoulos: Applications of Neural Networks in Electromagnetics, Artech House, Norwood, 2001.
11. I. D. Mayergoyz: Mathematical Models of Hysteresis, Springer, 1991.
12. M. Kuczmann–A. Iványi: Scalar Hysteresis Model Based on Neural Network, IEEE International Workshop on Intelligent Signal Processing, Budapest, Hungary, May 24-25, 2001, pp.143-148.
13. M.Kuczmann–A. Iványi: Vector Neural Network Hysteresis Model, Physica B, vol.306, 2001, pp.143-148.
14. M.Kuczmann–A. Iványi: A New Neural Network Based Scalar Hysteresis Model, Record of the 13th Compumag Conference on the Computation of Electromagnetic Fields, Lyon-Evian, France, July 2-5, 2001, Vol.2/4, PC1-9, pp.30-31.
15. M.Kuczmann–A. Iványi: Neural Network Based Scalar Hysteresis Model, Proceedings of the 10th International Symposium on Applied Electromagnetics and Mechanics, Tokyo, Japan, May 13-16, 2001, pp. 493-494.

16. M.Kuczmann–A. Iványi: Scalar Neural Network Hysteresis Model, Proceedings of the XI. Internatonal Symposium on Theoretical Electrical Engineering, Linz Austria, August 19-22, 2001, ISTET131.
17. M.Kuczmann–A. Iványi–T. Barbarics: Neural Network Based Simulation of Scalar Hysteresis, Proceedings of the X. International Symposium on Electromagnetic Fields in Electrical Engineering, Cracow, Poland, September 20-22, 2001, pp. 413-416.
18. M. Kuczmann, A. Iványi, Neural Network Model of Magnetic Hysteresis, submitted to Compel.
19. M. Kuczmann–A. Iványi: Neural Network Simulation of Scalar Magnetic Hysteresis, Proceedings of the 4th Japan-Central Europe Joint Workshop on Energy and Information in Non-Linear Systems, Brno, Czech Republic, November 10-12, 2000, pp.102-105.
20. M. Kuczmann–A. Iványi: Neural Network Based Scalar and Isotropic Vector Hysteresis Operator (in Hungarian), submitted to Híradástechnika.

Miklós Kuczmann

He obtained diploma in electrical engineering at the Budapest University of Technology and Economics at the Faculty of Electrical Engineering and Information Technology in 2000. He joined the Department of Electromagnetic Theory as a PhD student. His research field is the simulation of the behavior of magnetic materials applying artificial neural networks and soft computing methods, electromagnetic field calculation and nondestructive testing. He got the price of SIEMENS for diploma thesis. Member of the Scientific Association for Infocommunications and the Hungarian Electrotechnical Association.

Motion estimation in environments corrupted by impulse noise

PETER RADOCZI AND RASTISLAV LUKÁČ

Department of Electronics and Multimedia Communications, Technical University of Košice

In this paper, the motion estimation algorithms are studied in term of their performance in the environments corrupted by the impulse noise. In order to achieve an unbiased view about the motion estimation performance in noisy environments, a wide range of the impulse noise corruption from no corruption to 20% of corrupted image points was considered. Besides three well-known motion estimation algorithms with the fast search such as three step algorithm, conjugate direction algorithm and two-dimensional logarithmic algorithm, a new method minimising a number of searching points is provided.

1. Introduction

A boisterous development of image processing and especially multimedia applications [1],[2],[3],[4],[8] such as Internet, CD-ROM, digital television, videoconferencing, etc. brings a requirement for an effective storage, processing and transmission of digital data. In the case of image component of multimedia data, the requirement is related to a compression of visual information. In order to achieve high compression rates, there is necessary to eliminate redundant image data expressed in the form of similar spatial and temporal frequencies.

Since a motion video or in other words an image sequence represents a three-dimensional (3-D) image signal [6],[10],[11] or a time sequence of two-dimensional images, besides the spatial frequencies in the frames, the redundant image data are related to temporal frequencies of an image sequence, too. For that reason in the image sequence coding, the statistical redundancies can be categorised into the spatial and the temporal domains.

A transform coding is one of the most general ways to reduce the spatial redundancy, which is called an intraframe coding, whereas the reduction of the temporal redundancy is referred to as interframe techniques. For the purpose of reducing temporal redundancies, motion estimation techniques, which estimate the displacement of objects between successive frames, have been successfully applied. In the image sequence coding [2],[5],[7],[13],[14] the interframe coding methods with the prediction of the object motion are used. For simplification of implementation, a motion compensation with a motion estimation is generally performed by stepwise translation of objects in the image. In the first stage of coding, the displacement of object is estimated by using motion estimation methods. The main methods of motion estimation are pel recursive techniques and block matching techniques. The first one estimate the motion vector on pel-by-pel basis, whereas the second one estimate the motion vector on block-by-block basis. The result of this step is an identification of block in the current frame with the most similarity in the previous frame. The offset between both blocks is the motion vector for motion compensated prediction. Consequently, the prediction error and motion vectors for each block has to be transmitted, where the prediction error is coded the DCT transform (in MPEG or H.26X standards) and transformation coefficient of DCT are quantized and coded by VLC method.

Usually, the performance of motion estimation algorithms is relatively comparable, mainly in the case of the noise-free image signals. However, an original signal is often corrupted by a noise [6],[9],[10],[11] that can represent a danger for an accurate motion compensation and thus, the decline of compression rates, too. If the present noise has an impulse character such as the impulse noise or bit errors, the performance of motion estimation decreases rapidly. For that reason, this paper focuses on the comparison between full search [15] and three well-known motion estimation algorithms based on fast search, namely a three step algorithm [7], a conjugate direction algorithm [13], a two-dimensional logarithmic algorithm [5] and a new motion estimation algorithm called R algorithm in the dependence on a wide range of the impulse noise corruption ranged from 0% (no corruption) to 20% corruption.

The rest of this paper is organised as follows. In the next section, a principle of motion compensation with motion estimation is described. Selected motion estimation algorithms including a new method are described in Section 3. Test sequences, mathematical models of the used noise corruption and the experimental results including a number of figures and tables are presented in Section 4. Finally, the achieved results and main ideas are concluded in Section 5.

2. Block matching motion estimation

The block matching algorithm (BMA) estimates the motion vector in a block-by-block basis. In the BMA, a current frame is divided into blocks of size $(N \times N)$ pixels. The block of pixels in the current frame is compared with the corresponding blocks within a search area of size $(N+2p) \times (N+2p)$ pixels in the previous frame, where p is the maximum displacement allowed. The motion vector of the current block is found. We briefly describe the operation of the BMA between two consecutive frames in Figure 1.

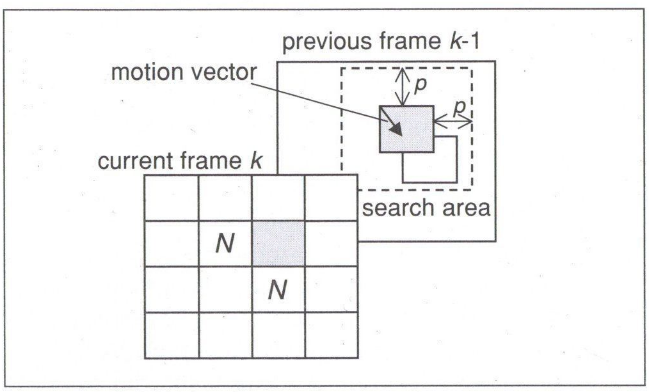


Figure 1. Principle of block matching

Let us assume $x_k(m,n)$ to be location of the pixel of the current block in the current frame, and $x_{k-1}(m+i,n+j)$ to be the location of the pixel in the candidate block in the previous frame, shifted by the i pixels and j lines within the search area. For the best match, the motion vector (i,j) represents the estimate of displacement in horizontal and vertical direction, respectively. The accuracy of motion estimation depends on the matching criteria (cost function) applied in the search area. The most popular ones are briefly described as follows [15]:

- *Cross-Correlation Function*

  In this measure, the highest $NCCF(i,j)$ within the search area, represents the best match.

$$NCCF(i,j) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} x_k(m,n)\, x_{k-1}(m+i,n+j)}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N} x_k^2(m,n)}\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N} x_{k-1}^2(m+i,n+j)}} \qquad (1)$$

- *Mean square error (MSE)*

  Reversal, for the smallest $MSE(i,j)$ within the search area, (i,j) represents the motion vector of the block. MSE is simpler than NCCF in computational complexity.

$$MSE(i,j) = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}[x_k(m,n) - x_{k-1}(m+i,n+j)]^2 \qquad (2)$$

- Mean absolute difference (MAD)

  In this criterion, the motion vector is determined by the smallest MAD(i,j) for all possible displacement (i,j) within the search area. The MAD is often applied due to its lower computational complexity.

$$MAD(i,j) = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left|x_k(m,n) - x_{k-1}(m+i,n+j)\right| \qquad (3)$$

3. Block matching algorithms

In this Section, there will be used three well known fast search-based motion estimation algorithms such as the three step algorithm [7], the conjugate direction algorithm [13], the two-dimensional logarithmic algorithm [5] and a new algorithm called R algorithm as described below. All presented fast search algorithms eliminated the positions in SA by principles of the quadrant monotonic function.

- *Three step algorithm*

  In the Three step search algorithm (TSA) all eight positions surrounding the co-ordinate with a step size of $p/2$ are searched first. At each minimum position the search step size is halved and the next eight new positions are searched. This method searches 25 positions to locate the best match.

- *Conjugate direction algorithm*

  In Srinivasan and Rao's conjugate direction search (CDS) method, at every iteration of the direction search, two conjugate directions with a step size of one pel, centered at the minimum position are searched.

- *Two-dimensional logarithmic algorithm*

  To reduce the computational complexity, Jain and Jain developed a two-dimensional logarithmic search algorithm (TDLOG) to track the direction of minimum mean square error distortion measure. In their method, the distortion for five initial positions, one at the center of the co-ordinate and four at co-ordinates $(\pm p/2, \pm p/2)$ of the search area, are computed first. In the next step, three more positions with the same step size in the direction of the previous minimum are searched. The step size is halved and the above procedure continues until the step size becomes 1. Finally all nine positions are searched.

- *R algorithm*

  In order to minimise a number of searching points on the assumption that the displacement is minimal, a new algorithm called R algorithm is proposed. In introduction of the searching procedure are search points situated into the centre of SA (Figure 2). The distance between two pixels is 1 pixel. After the first step there can occur following cases:

1. The point having a minimum distortion is the central point. Then the motion vector is zero and algorithm stops.

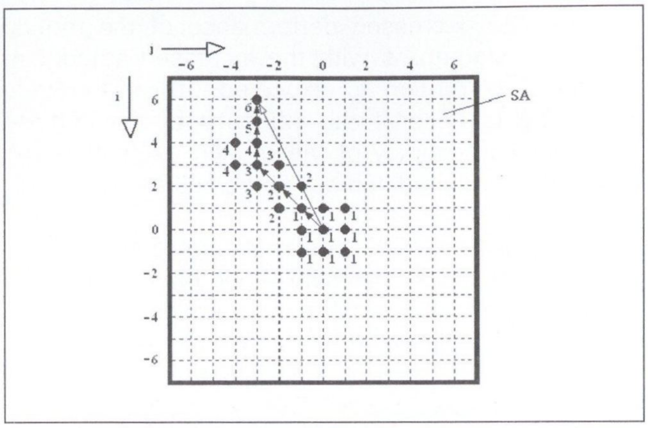


Figure 2. R Algorithm

2. The point with a minimum distortion is located into some of orthogonal directions. Then, a part of the well-know orthogonal search algorithm is applied into the particular orthogonal direction.
3. The point with a minimum distortion is located into some of corners of reduced SA *(-1,1)*. In the next step, there are 3 search points located in the directions to the boundary of SA, namely two points in orthogonal and one in diagonal direction *((1,-2);(2,2);(2,-1))*. After the second step, there can occur 3 cases repeatedly:
   a) If values of the distortion of 3 new search points in the second step, they are higher as corner's point in the first step, then the location of the corner's point defines the motion vector.
   b) If the point with minimum distortion is located into the some of orthogonal directions, the search algorithm returns to the second step.
   c) If the point with minimum distortion is located into the diagonal direction, then the search algorithm returns to the third step.

4. Experimental results

- Test image sequences

In this paper, we use three image sequences (Figure 3). Each sequence consists of 30 frames with a resolution of 256x256 pixels with 8-bits/pixel gray-scale quantization. The complexity of an image sequence is mirrored by the amount of motion and by a number of image details and edges. The first sequence called Salesman (Figure 3a,b) includes a number of details and edges. The considerable motion of the man in a foreground increases complexity of this sequence. The sequence Susie (Figure 3c,d) is the most elementary, there are many monotonous fields, the problem areas can be hair. Accordingly, the minimal motion is observed. Small objects such as vegetation create background of the sequence People (Figure 3e,f). In addition, there is a large complex

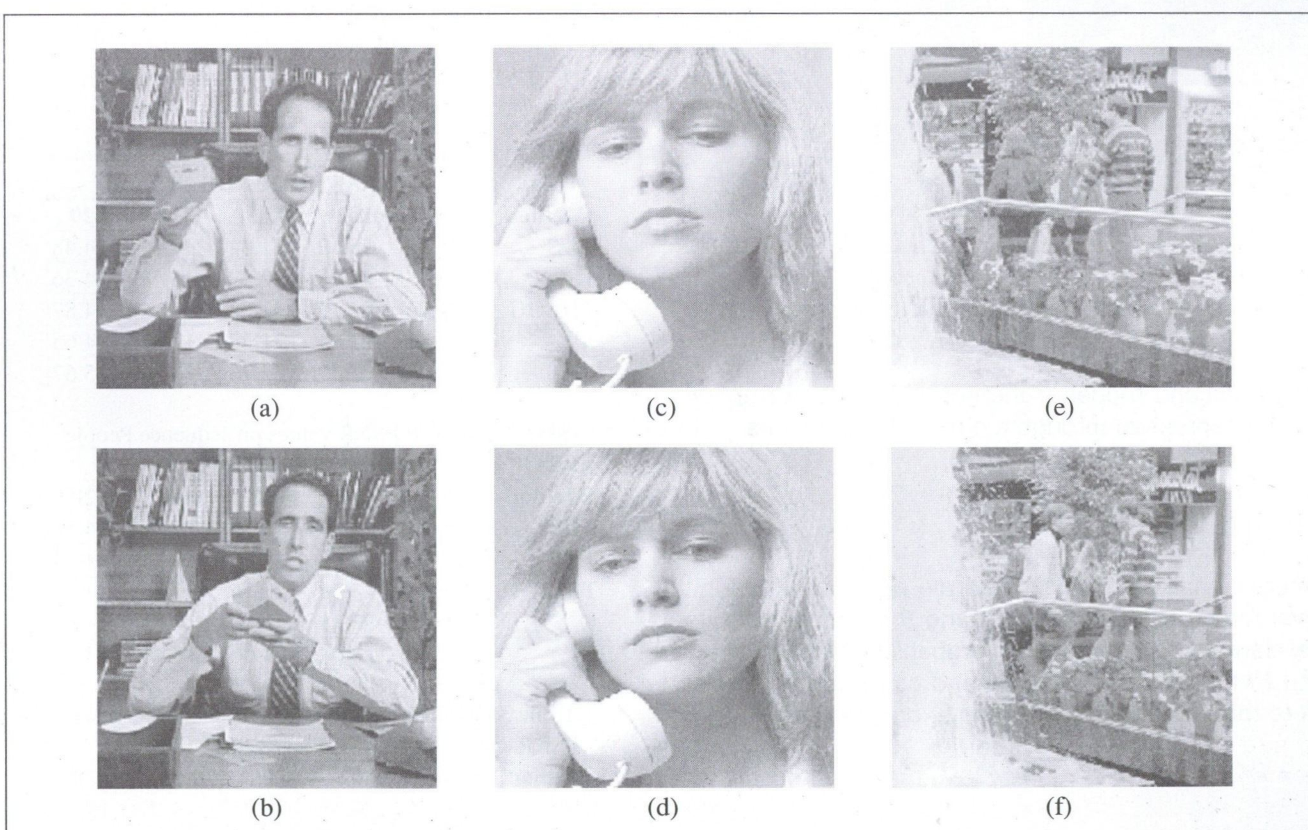


Figure 3. Original image sequences with various amount of the motion and signal-details (edge density) complexity
(a) Salesman – 5th frame (b) Salesman – 25th frame (c) Susie – 5th frame
(d) Susie – 25th frame (e) People – 5th frame (f) People – 25th frame

motion caused by the motion of people and camera, simultaneously.

- Noise corruption

To achieve a noise corruption, we use the impulse noise. In general, mathematical model of the impulse noise can be expressed as [10,11]

$$x_{i,j} = \begin{cases} \upsilon & \text{with probability } p \\ o_{i,j} & \text{with probability } 1-p \end{cases} \qquad (4)$$

where i,j characterise the sample position, $o_{i,j}$ is the sample from the original image, $x_{i,j}$ represents the sample from the noisy image, p is a corruption probability and υ is a noise vector of intensity random values.

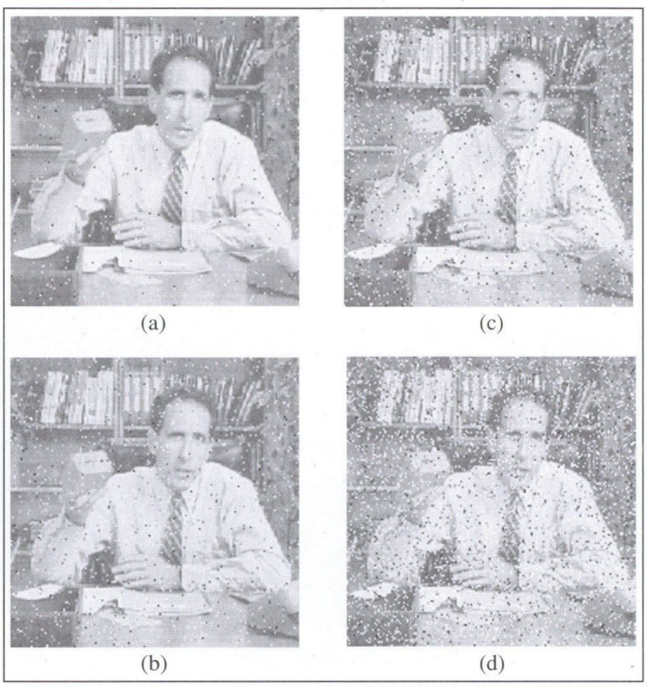


Figure 4. 5th frame of noisy image sequence Salesman (a corruption is modeled by the impulse noise) (a) p=0.02 (b) p=0.05 (c) p=0.10 (c) p=0.15

The second model of the impulse noise corruption can be expressed through the model of bit errors

$$^{*}k_{i,j}^{m} = \begin{cases} k_{i,j}^{m} & 1-p_{\upsilon} \\ 1-k_{i,j}^{m} & p_{\upsilon} \end{cases} \qquad (5)$$

where i,j characterise the sample position, m is a bit level forced to be between 1 and B (a number of bits per sample), p is a bit error probability and finally $\{k\}$ and $\{^{*}k\}$ characterise original and corrupted bit levels. Note that the original sample is expressed as

$$o_{i,j} = k_{i,j}^{1}2^{B-1} + k_{i,j}^{2}2^{B-2} + \ldots + k_{i,j}^{B-1}2 + k_{i,j}^{B} \qquad (6)$$

whereas a sample from noisy image is defined by

$$x_{i,j} = {}^{*}k_{i,j}^{1}2^{B-1} + {}^{*}k_{i,j}^{2}2^{B-2} + \ldots + {}^{*}k_{i,j}^{B-1}2 + {}^{*}k_{i,j}^{B} \qquad (7)$$

Since, the decreased performance of the motion estimation algorithms with the increased amount of the noise corruption is expected, it can be very interesting to observe the influence of the impulse noise on the accuracy of the motion estimation and the average required number of the search points.

- Performance of motion estimation algorithm

Presented block matching algorithms have been tested on the above-mentioned image sequences (Figure 3 and Figure 4). As the cost function was chosen the MSE error criteria and the size of block was equal to 8x8 pixels. The degree of reduction interframe redundancy was evaluated by Peak-signal-to-noise-ratio (PSNR).

$$PSNR = 10\log_{10}\left\{\frac{(255)^2}{\frac{1}{256\times256}\sum_{i=1}^{256}\sum_{j=1}^{256}(x_{i,j}-x_{i,j}^{C})^2}\right\} [dB] \qquad (8)$$

where $x_{i,j}$ are values of pixels into the current frame and $x_{i,j}^{C}$ are values of pixels into the previous frame after motion compensation.

The achieved results related to PSNR are summarised in Table 1, Table 2 and Table 3. It can be seen that all fast motion estimation algorithms achieved PSNR value nearly to PSNR values by the Full Search (FS) algorithm. Mainly, by the third sequence Susie, the difference is really near to zero. On the other side, the higher the noise value the increased the difference. According to PSNR parameter the best performances achieved the TDLOG and R algorithms, respectively.

Table 1 Obtained PSNR values on sequence Salesman

PSNR	NOISE PROBABILITY *p*				
Method	**0**	**0.02**	**0.05**	**0.10**	**0.20**
CDS	34.63	23.69	21.02	17.15	14.48
TSA	33.57	22.78	20.38	16.78	14.36
TDLOG	34.76	24.15	21.58	17.60	14.87
R	34.81	23.84	21.23	17.37	14.65
FS	35.07	25.30	22.76	18.56	15.67

Table 2 Obtained PSNR values on sequence People

PSNR	NOISE PROBABILITY *p*				
Method	**0**	**0.02**	**0.05**	**0.10**	**0.20**
CDS	29.43	22.87	19.75	17.15	14.53
TSA	28.06	22.01	19.10	16.74	14.34
TDLOG	29.53	23.26	20.16	17.58	14.90
R	29.23	23.38	20.18	17.63	14.99
FS	29.99	24.07	21.06	18.46	15.70

Table 3 Obtained PSNR values on sequence Susie

PSNR	NOISE PROBABILITY *p*				
Method	**0**	**0.02**	**0.05**	**0.10**	**0.20**
CDS	36.54	24.68	20.96	18.10	15.26
TSA	36.17	24.03	20.36	17.79	15.05
TDLOG	36.49	25.16	21.33	18.46	15.44
R	36.59	24.84	21.21	18.33	15.46
FS	36.63	26.51	22.69	19.70	16.51

Now, the performance of the motion estimation algorithms is evaluated (Table 4, Table 5 and Table 6) through the average required number of search points (ANSP). There are visually greater differences among the motion estimation algorithms than it is in the case of PSNR evaluating. The algorithms are referred to the number of search points of FS algorithm. It can be seen that all fast algorithms achieve the considerable reduction of the search points in comparison with FS algorithm. According to ANSP, the low amount of points was achieved by CDS and R algorithms, respectively.

Table 4 Obtained ANSP values on sequence Salesman

ANSP	NOISE PROBABILITY *p*				
Method	**0**	**0.02**	**0.05**	**0.10**	**0.20**
CDS	8.86	9.41	10.02	10.27	10.35
TSA	25.00	25.00	25.00	25.00	25.00
TDLOG	15.58	17.68	19.64	19.65	19.71
R	8.74	9.51	10.44	10.31	10.29
FS	169.00	169.00	169.00	169.00	169.00

Table 5 Obtained ANSP values on sequence People

ANSP	NOISE PROBABILITY *p*				
Method	**0**	**0.02**	**0.05**	**0.10**	**0.20**
CDS	11.98	11.63	11.60	11.35	11.14
TSA	25.00	25.00	25.00	25.00	25.00
TDLOG	19.36	19.90	20.22	20.37	20.25
R	11.29	11.15	11.25	11.12	10.98
FS	169.00	169.00	169.00	169.00	169.00

Table 6 Obtained ANSP values on sequence Susie

ANSP	NOISE PROBABILITY *p*				
Method	**0**	**0.02**	**0.05**	**0.10**	**0.20**
CDS	7.68	9.24	10.07	10.50	10.62
TSA	25.00	25.00	25.00	25.00	25.00
TDLOG	14.53	17.85	19.55	20.16	20.29
R	8.36	9.74	10.33	10.69	10.50
FS	169.00	169.00	169.00	169.00	169.00

Generally, the motion estimation algorithm can be evaluated by both parameters, the PSNR value and the average required number of the search points. According to the first parameter, the best results were achieved by the TDLOG algorithm and the R algorithm. If the motion estimation was evaluated through the second parameter, the best performance was provided by the CDS algorithm and the R algorithm.

5. Conclusion

In order to find the block with the minimum distortion, the full search algorithm exploited the process based on the matches of all possible displaced candidate blocks within the search area in the previous frame. The main disadvantage of this procedure is a very high computational complexity.

The reason, why the fast block matching algorithms provided lower performances than the FS procedure did, is that in the search area does not exist only one global minimum, but in the search area often exist several local minimum values of the cost function. Currently, the location of these local minimum values is ultimate for finding an optimal point. The correct estimation of the motion vector can be executed by the assumption that all pixels in block do the same move. Reliability of the estimation displacement of the block depends on block size and largeness of the movement. Using less block size increases the probability that in the SA will exist more identical blocks.

Since the useful information is often degraded by a noise, this paper has focused on the performance of the fast motion estimation algorithms in environments corrupted by the impulse noise. In general, the performance of all motion estimation algorithms decreases with the increased impulse noise probability. The achieved results have shown that the performance of a new R algorithm is comparable with the performance of the FS algorithm and three well-known fast motion estimation algorithms. In addition, the proposed method has provided the smaller average required number of the search points (ANSP) for a wide range of the impulse noise corruption. For that reason, the proposed method has provided the best compromise between the PSNR and ANSP not only for the motion estimation in the ideal noise-free situation, however, it can be used successfully in the noisy environments, too.

References

1. Z. Bojkovic and D. Milovanovic: "Challenges of Information Processing in Multimedia Communications," Proc. ECMCS'2001 EUURASIP Conference in Budapest, Hungary, 2001, pp.133-140
2. B. Furht, J. Greenberg and R. Westwater: "Motion estimation algorithms for video compression," Kluwer Academic Publishers, Massachusetts 1997.
3. M. Grgic, M. Ghanbari and S. Grgic, "Texture-based Image Retrieval in MPEG-7 Multimedia System," Proc. of the IEEE Region 8 EUROCON'2001 Conference - International Conference on Trends in Communications (Volume 2/2), Bratislava, 2001, Slovakia, pp. 365-368
4. M. Grgic, S. Grgic and B. Zovko-Cihlar: "Image Retrieval Based on Texture Features Extraction," Proc. ECMCS'2001 EURASIP Conference in Budapest, Hungary, 2001, pp. 152-155.
5. J. R. Jain and A. K. Jain: "Displacement Measurement and Its Application in Interframe Coding," IEEE Transaction on Communications, Vol. COM-29, No. 12, December 1981, pp. 1799-1808.
6. R. P. Kleihorst, R. L. Lagendijk and J. Biemond: "Noise Reduction of Image Sequences Using Motion Compensation and Signal Decomposition," IEEE Trans. on Image Processing, Vol.4, 1995, pp.274-284.

7. T. Koga–K. Iinuma–A. Hirano–Y. Iijima–T. Ishiguro: "Motion-Compensated Interframe Coding for Video Conferencing," Proc. Nat. Telecommun. Conf., 1981 New Orleans, LA, pp. G5.3.1-G5.3.5.
8. C. Kotropoulos, and I. Pitas: "Intelligence in Modern Communication Systems," Proc. of COST 254, Neuchatel, 1999, pp. 1-12.
9. R. Lukáč: "Impulse Detection by Entropy Detector (H - Detector)," Journal of Electrical Engineering, Vol.50, No.9-10, 1999, pp.310-312.
10. R. Lukáč: "Impulse Detectors for Noised Sequences," Radioengineering, Vol.10, No.2, July 2001, pp.24-31.
11. R. Lukáč–S. Marchevský, "LUM Smoother with Smooth Control for Noisy Image Sequences," EURASIP Journal on Applied Signal Processing, Vol.2001, No.2, 2001, pp.110-120.
12. R. Lukáč–S. Marchevský: "Boolean Expression of LUM Smoothers," IEEE Signal Processing Letters, Vol.8, No.11, November 2001, pp. 292-294.
13. R. Srinivasan–K. R. Rao: "Predictive Coding Based on Efficient Motion Estimation," IEEE Transaction on Communications, Vol. COM-33, No. 8, August 1985, pp.888-896.
14. P. Radoczi: "Algebraic and Trigonometric Approximation in Block Matching," Proc. DSP-MCOM 2001, the 5th International Scientific Conference, Digital Signal Processing and Multimedia Communications, Košice, November 2001, pp. 60-62.
15. P. Radoczi: "Influence of Boundary Properties to Search Area in Block Matching Motion Estimation," Proc. Radioelektronika 2001, Brno - Czech Republik, 2001, pp. 378-381.

Peter Radoczi (Ing.) *received the M.Sc. (Ing.) degree with honours at the Technical University of Košice, Slovak Republic, at the Department of Electronics and Multimedia Communications in 1999. Currently, he is a PhD. student at the same department. His research interest includes motion estimation and multimedia processing.*

Rastislav Lukáč (Ing, Ph.D.) *received the M.Sc. (Ing.) degree with honours at the Technical University of Košice, Slovak Republic, at the Department of Electronics and Multimedia Communications in 1998. In 2001 he finished PhD. study. Currently, he is an assistant professor at the Department of Electronics and Multimedia Communications at the Technical University of Košice. His research interest includes image filtering, impulse detection, neural networks and permutations.*

Congestion Measure: Generalization, Analysis and Application

D. D. LUONG, J. BÍRÓ

High Speed Networks Laboratory
Department of Telecommunication and Telematics
Budapest University of Technology and Economics

In end-to-end congestion control protocols, sources continuously obtain feedback from the network, detect the level of congestion along their network path and adjust their sending rate accordingly. The level of congestion can be represented by the loss rate in TCP Tahoe/ Reno, queuing delay in TCP Vegas or marking probability in ECN-capable protocols. We call them the congestion measures. Our aim in this paper is to generalize the con-gestion measures and investigate their common properties, especially the so-called „population effect", where the total congestion level tends to increase with the number of active connections. This better understanding of the congestion measures promotes the use of ECN-like router mechanisms to avoid the negative impact of popula-tion effect. We also propose the use of congestion measure in available bandwidth estimation and dynamic path selection algorithms.

Keywords: congestion measure, end-to-end congestion control protocols, available bandwidth, path selection.

I. Introduction

TCP Tahoe or TCP Reno always probes the unused bandwidth of the network by increasing their sending rate. When it notices packet losses via loss signal (duplicate ACK or time-out), it decreases the sending rate and retransmits the lost packets. The well-known principle of TCP is the additive-increase, multiplicative-decrease behavior. Assuming that losses are random, by different approaches [1], [2], [3] researchers have obtained the throughput as a function of loss rate. The simplest form of throughput formula is the following

$$Throughput = \frac{cMSS}{RTT\sqrt{p}}$$

where MSS and RTT are the Maximum Segment Size and the Round Trip Time, p is the loss rate and c is a constant which depends on the current TCP implementation, the ACK strategy (delayed or not) [2].

TCP Vegas adopts another mechanism to detect congestion [6]. It estimates the amount of extra data and controls its congestion window (cwnd) accordingly. Extra data is the fraction of data in transit that stays in the buffers of routers. The basic idea is that the amount of extra data should not be too large or too small. When the amount of extra data becomes larger than a threshold, β meaning that the level of buffer occupancy gets large and overflow may occur, TCP Vegas decreases its $cwnd$ linearly. When the amount of extra data is smaller than another threshold α ($\alpha < \beta$), the $cwnd$ is increased by one per round-trip time. Otherwise, the $cwnd$ remains unchanged. The decision of the window size is done once each RTT. In [5], D. D. Luong et al. have shown that by setting the two thresholds α and β close to each other, in the stable state the throughput of TCP Vegas is inversely proportional to the total upward queuing delay (queuing delay experienced by the acknowledgements is excluded).

Recently, several end-to-end congestion control mechanisms are proposed to achieve global optimization. Each end user has its own utility function. Resources (bandwidth in this case) need to be distributed among users in such a way to maximize the overall objective function, which is the sum of utilities of all active flows [9].

A window-based congestion control protocol for global optimization is described in [14], which estimates the end-to-end propagation delay and queuing delay, then adjusts the window size accordingly. The authors prove that the applied mechanism guarantees the convergence and stability. In the stable state, the window size and therefore, the sending rate is a function of queuing delay. Another congestion control protocol proposed in [15] adjusts the transmission rate in accordance with the loss rate or ECN marking probability. These two protocols do not require any modifications in routers. On the contrary, S. Low et al. found a congestion control protocol [12], [13], which requires the co-operation of routers. A so-called "link price" is defined at each link to indicate the level of congestion. The sender obtains the total price of the network path via Random Exponential Marking mechanism (REM). Then the transfer rate is the function of the total price.

All protocols mentioned above have an essential common point: their sending rate is altered according to the level of congestion estimated via measures generated by the routers eg. the loss rate, the ECN marking probability, the queuing delay. We call them the congestion measures. Our aim in this paper is to generalize the congestion measures and investigate their common properties, especially the so-called "population effect", where the total congestion level tends to increase with the number of active connections. This better understanding of the congestion measures promotes the use of ECN-like router mechanisms to avoid the negative impact of population effect.

The congestion measures are to indicate the level of congestion within the network and have been only used in congestion control protocols. Our opinion here is that the congestion measures can be applied in other areas of networking. In this paper, we show how to use congestion measures to estimate roughly the available bandwidth for elastic applications. Moreover, the congestion measures can be used as a link state information in dynamic path selection algorithms.

Throughout this paper, we have to use some new terms to describe things which proper denomination we could not find in the literature. All new terms and their definition are listed in the Appendix. We apologize in advance for any confusion resulting from our terminologies.

The rest of this paper is organized as follows. Section II gives the general network model and the definition of the additive congestion measure. We analyze some properties of the additive congestion measures in Section III. We show the application of congestion measures to the available bandwidth estimation and the path selection in Section IV. Section V demonstrates some simulation results. This paper is concluded in Section VI.

II. The model

Let's consider a network with a set of links L. Each link $l \in L$ has a finite capacity $\mu_l > 0$. Let F denote the set of flows which have data to transfer over the network, each flow $f \in F$ goes through a non-empty subset of L, called a route (or a path). We define an indicator function

$$I(f,l)=\begin{cases}1, & l \in \text{the route of flow} f\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

Let λ_f be the bandwidth allocation of the flow f. Any feasible bandwidth allocations must sat-isfy the capacity constraint

$$\sum_{f\in F} I(f,l)\lambda_f \le \mu_l, \quad \text{for all } l \in L \tag{2}$$

Link l is the congested link if the aggregate source rate is exactly equal to its capacity. That is the inequality symbol is replaced by the equality symbol in (2). The congestion control mechanisms always try to use all of the available bandwidth. Therefore, any flow must have at least one congested link on its network path. Note that the congested link should not be confused with the bottleneck link of a network path, which has smallest capacity on that path.

Denote s_l be the level of congestion measure at link l. We assume that s_l is not negative. Furthermore, s_l is equal to 0 if l is not a congested link. These assumptions are consistent with all types of congestion measure mentioned in Section I. Let s_f be the level of end-to-end congestion measure experienced by flow f. In the rest of this paper we call the level of congestion measure simply the congestion level.

We assume that the end-to-end congestion control protocol applied by the flows converges and stabilizes. In the stable state, each flow f have the rate of λ_f

$$\lambda_f = R_f(s_f) \tag{3}$$

where R_f is strictly decreasing function and have positive values if s_f is positive. Different types of R_f are shown in Table I.

Definition 1: A congestion measure is additive if the end-to-end congestion level experienced by a flow is equal to the sum of conges-tion levels at all links along the path of that flow.

$$s_f = \sum_{l\in L} I(f,l)s_l \quad \text{for all } f \in F \tag{4}$$

Obviously, the queuing delay, used in TCP Vegas [5], [6] or in J. Mo's protocol [14], is an additive congestion measure since end-to-end queuing delay is the sum of queuing times at relevant routers. Similarly, link's price, used in S. H. Low's protocol, is also additive and clearly proved in Page 2 of [13].

The loss rate, used in TCP Tahoe/Reno or in protocol proposed by S. Kunniyur [15], is not additive. Let's consider a flow f passing through m_f links with loss probabilities $p_1, p_2, \ldots, p_{mf}$ on these links, then the end-to-end loss probability pf can be computed by

$$p_f = 1-\prod_{i=1}^{m_f}(1-p_i) = \sum_{i=1}^{m_f} p_i - \sum_{i=1}^{m_f-1}\sum_{j=i+1}^{m_f} p_i p_j + \sum_{i=1}^{m_f-2}\sum_{j=i+1}^{m_f-1}\sum_{k=j+1}^{m_f} p_i p_j p_k - \ldots \tag{5}$$

With small loss probabilities $p_1, p_2, \ldots, p_{mf}$

$$p_f \approx \sum_{i=1}^{m_f} p_i \tag{6}$$

Remark 1: In case of small link loss probabilities, the loss rate can be considered as an additive conges-tion measure

Protocol	Rf(s)	The meaning of s	Remarks	Ref.
TCP Reno	$\frac{cMSS}{RTT\sqrt{s}}$	loss rate	c: constant, MSS: Maximum Segment Size	[1]
TCP Vegas	$\frac{\gamma}{s}$	total upward queuing delay	γ: constant	[5]
Low's, Mo's and Kun-niyur's mechanisms	$U_f^{\prime -1}(s_f)$	link price, total upward queuing delay and loss rate, re-spectively.	U_f is the utility function of the flow f	[12], [14], [16]

Table I. Transmission rate as a decreasing function of congestion level.

III. Some properties of congestion control protocols using additive congestion measure

A. The population effect

To characterize congestion measures, we consider a simple network with a single bottleneck link and n different competing flows. When the system stabilizes, the total transmission rate of all flows is exactly equal to the capacity of the bottleneck link

$$\mu_{bott} = \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} R_i(s) \tag{7}$$

where R_i is the rate function of flow i and s is the congestion level at the bottleneck link in the stable state. For simplicity we assume that all flows have the same rate function R.

$$\mu_{bott} = nR(s) \tag{8}$$

Then

$$s = R^{-1}\left(\frac{\mu_{bott}}{n}\right) \tag{9}$$

R is a strictly decreasing function (Section II) and then so is R^{-1}. Equation (9) implies that the congestion level at the bottleneck link increases if the number of active flows increases. We call this the population effect. Table II shows the connection between the number of flows and the congestion level in some particular cases.

The population effect shown above could cause performance degradation if the congestion measure used in a congestion control protocol is loss rate or queuing delay. The high loss rate caused by a large number of connections results in a low bandwidth utilization. Similarly, the increase of queuing delay yields a high end-to-end delay, which is a critical performance factor in multimedia applications. ECN is a realizable solution to avoid these impacts of the population effect, since the increase of marking rate only makes the sending rates throttle down as one expects from a congestion control protocol, it does not have any negative impacts on performance factors eg. loss rate, delay.

Rate function R	Congestion level s
$\frac{cMSS}{RTT\sqrt{s}}$	$\frac{cMSS}{RTT\sqrt{\mu_{bott}}}\sqrt{n}$
$\frac{\gamma}{s}$	$\frac{\gamma}{\mu_{bott}} n$

In networks containing more than 1 link, the change of congestion levels is more complicated. When a new connection is established (or an old one is terminated), the link's congestion level does not always increase (decrease). We illustrate this via the network configuration in Fig. 1. Connection 0 goes through 2 bottleneck links with the same capacity, while connections 1 and 2 go through link 1 and link 2, respectively. The link's congestion levels in the stable state are s_1 at link 1 and s_2 at link 2. The congestion level s2 must increase when the new connection (con. 3) is established. We assume that the congestion measure in this example is additive. If s_1 is unchanged or increases, the transmission rate of connection 1 does not increase while the rate of connection 0 decreases. In this case, there are unused capacity of link 1, meaning that connection 1 has no congested link. Consequently, s_1 must decrease.

Individual link's congestion level can increase or decrease after the establishment of new connections. But how about the total congestion level which is equal to the sum of congestion levels of all links? In the following, we derive an upper bound and a lower bound of the total congestion level and show that these bounds increase strictly with the establishment of new connections.

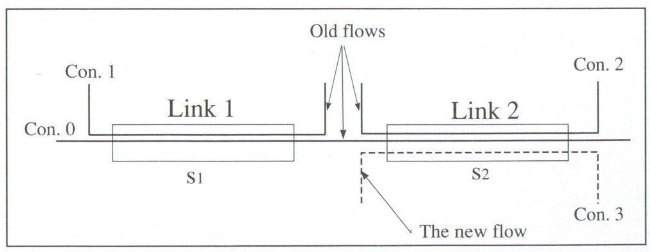


Figure 1. A network model with 2 bottleneck links.

Proposition 1: If all flows use the same rate function R and the congestion measure is addi-tive, we have

$$s_l \le R^{-1}\left(\frac{\mu_l}{n_l}\right) \quad (10)$$

where n_l denotes the number of the flows passing through link l

Proof: If link l is not a congested link, the congestion level s_l is equal to 0 and the inequality described above is obvious. Otherwise, in the case of congested link, we have

$$\sum_{f\in F} I(f,l)\lambda_f = \sum_{f:I(f,l)=1} R(s_f) = \mu_l \quad (11)$$

Since the congestion measure is additive and R is a decreasing function

$$R(s_f) = R\left(\sum_{l'\in L} I(f,l')s_{l'}\right) \le R(s_l), \forall f: I(f,l)=1 \quad (12)$$

From (11) and (12)

$$\mu_l = \sum_{f:I(f,l)=1} R(s_f) \le n_f R(s_l) \quad (13)$$

Then

$$R(s_l) \ge \frac{\mu_l}{n_l} \quad (14)$$

Or

$$s_l \le R_{-1}\left(\frac{\mu_l}{n_l}\right) \quad (15)$$

Proposition 2: If all flows use the same rate function R and the congestion measure is addi-tive, we have

$$\sum_{l'\in L} s_{l'} \ge R^{-1}\left(\min_{l\in L}\frac{\mu_l}{n_l}\right) \quad (16)$$

Proof: From the capacity constraint (2), for any link l we have

$$\sum_{f\in F} I(f,l)\lambda_f = \sum_{f:I(f,l)=1} R(s_f) \le \mu_l \quad (17)$$

Since the congestion measure is additive

$$R(s_f) = R\left(\sum_{l'\in L} I(f,l')s_{l'}\right) \ge R\left(\sum_{l'\in L} s_{l'}\right) \quad (18)$$

From (17) and (18)

$$\mu_l \ge \sum_{f:I(f,l)=1} R(s_f) \ge \sum_{f:I(f,l)=1} R\left(\sum_{l'\in L} s_{l'}\right) =$$

$$= n_l R\left(\sum_{l'\in L} s_{l'}\right) \quad (19)$$

Which means

$$R\left(\sum_{l'\in L} s_{l'}\right) \le \frac{\mu_l}{\mu_l} \quad (20)$$

This is true for all link l, so

$$R\left(\sum_{l'\in L} s_{l'}\right) \le \min_{l\in L}\frac{\mu_l}{n_l} \quad (21)$$

Or

$$\sum_{l'\in L} s_{l'} \ge R^{-1}\left(\min_{l\in L}\frac{\mu_l}{n_l}\right) \quad (22)$$

Remark 2: An upper bound and a lower bound of the total congestion level can be given by

$$R^{-1}\left(\min_{l\in L}\frac{\mu_l}{n_l}\right) \le \sum_{l'\in L} s_{l'} \le \sum_{l\in L} R^{-1}\left(\frac{\mu_l}{n_l}\right) \quad (23)$$

Conclusively, in a general network, the lower bound and the upper bound of total congestion level increases with the set-up of new connections. We have the similar results when the rate functions are different (see the Appendix). We regret that we could not claim any more precise statement about the evolution of the total congestion level itself. But by simulation in Section V, we provide evidence that suggest the strict increase of the total congestion level when new connection are established.

B. The relation between congestion measure and global fairness optimization

The most common rate sharing scheme for competing flows is that the bandwidth should be shared as equally as possible and this results in the max-min fairness. In spite of the fact that the notion of fairness in resource sharing usually indicates the max-min fairness, there is no economic motivation for this criterion.

F. Kelly [9] argues that bandwidth should rather be shared in such a way to maximize an objective function representing the overall utility of the flows in the network. Each flow f with rate allocation λ_f has a utility $U_f(\lambda_f)$, which is an increasing, strictly concave and continuously differentiable function of λ_f [11]. The overall utility is assumed to be additive, meaning that it is $\Sigma_{f\in F} U_f(\lambda_f)$. The rate sharing scheme under this model is the solution of the following optimization problem

P:

maximize $\Sigma_{f\in F} U_f(\lambda_f)$
subject to $\Sigma_{f\in F} I(f,l)\lambda_f \le \mu_l$ for all $l \in L$
over $\lambda_f \ge 0$ for all $f \in F$

Proposition 3: If a congestion control protocol stabilizes and the relevant congestion signal is additive, the rate sharing of that protocol is the

In other words, the estimation becomes more accurate when the networks are more heavily loaded. As shown in Fig. 5, the relative errors get closer to the line $y = 0$ when the number of active flow gets larger. The straight lines in this figure are also the linear fits and we can see that they come closer and closer to the the line $y = 0$.

C. 3 Path selection

As mentioned in Subsection III-B, if each flow f has a utility function U_f, then the control protocol with rate functions $U_f'^{-1}$ maximizes the overall utility $\Sigma_{f\in F} U_f(\lambda_f)$ of all flows. In that case, the path of each flow is already selected. If we consider the optimization on both routing and rate distribution level, the overall utility is a good measure to evaluate the path selection algorithms for elastic application.

We use the utility function $U(\lambda) = 5log(\lambda)$ for all flows. Then the corresponding rate func-tion is $R(s) = \frac{5}{s}$. Three path selection algorithms are evaluated: minimum hop count, Shortest-Widest path and Widest-Shortest path. As we can see in Fig. 6, the Shortest-Widest path and Widest-Shortest path have nearly equal overall utilities and both of them outperform the min-imum hop count algorithm. In a heavily loaded network, the Widest-Shortest path is the best choice to achieve efficient resource utilization (Fig. 7).

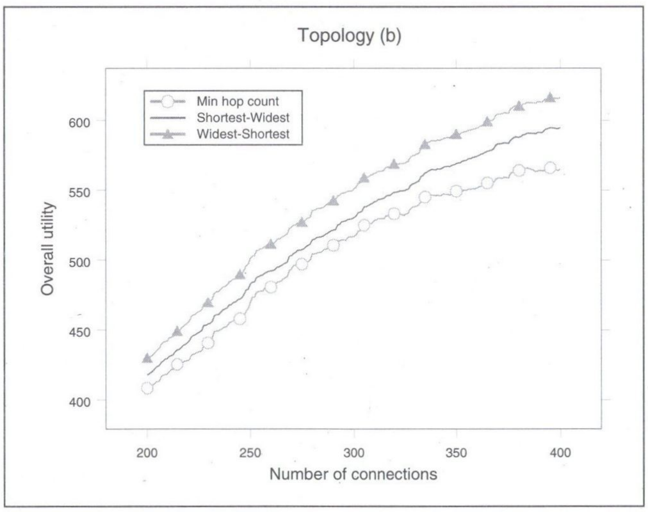


Fig. 7. Path selection algorithms in heavily loaded network-Topology (b)

VI. Conclusion

This paper generalizes and investigates a common characteristic of end-to-end congestion control protocols: the congestion measure. We have shown an important property of the con-gestion measure: the population effect, where the total congestion level tends to increase when new connections are established. If the congestion measure is the loss rate or queuing delay, the population effect could lead to the network performance degradation. This observa-tion promotes the use of ECN-like marking mecha-nisms.

Up to the present, the congestion measures have been only used in congestion control proto-cols. Our opinion is that the congestion measures, presenting loading state of the network, can be applied in other areas of networking. In this paper, we have proposed two applications:

1. The available unused bandwidth is not a sufficient information to predict the performance of elastic applications. We have proposed the use of end-to-end congestion level to estimate the future transfer rate for best effort applications.
2. We have also shown that congestion measure can be used as a cost of links in dynamic path selection.

Simulation results have confirmed that the elastic bandwidth estimation using end-to-end con-gestion level is fairly accurate when the network path contains several competing flows. It has been also shown by simulation that path selection using congestion measure as a link's cost outperforms the static shortest path algorithm.

Appendices

I. Some Terminoligies

Congestion measure: a measure to represent the level of congestion. Some examples of conges-tion measure are the loss rate, ECN marking probability and the queuing delay.
Total congestion level: is the sum of congestion levels of all network links.
Rate function R_f: is a mapping from the end-to-end congestion level s_f to the transmission rate λ_f of a flow f

$\lambda_f = R_f(s_f)$

Additive congestion measure: A congestion measure is additive if the end-to-end congestion level experienced by a flow is equal to the sum of congestion level at all links along the path of that flow.
Population effect: a phenomenon where the congestion level of links tends to increase with the number of active connections.
Elastic available bandwidth: the expected bandwidth available for a certain elastic application.

II. A general upper bound and lower bound for total congestion level

When the rate functions are different, we introduce the following functions:

$$R_{l,F}(s) = \sum_{f\in F} I(f,l) R_f(s)$$

Note that $R_{l,F}$ is a strictly decreasing function because of the same property of F_f.

Proposition 4: If the congestion measure is additive, we have

$$s_l \leq R_{l,F}^{-1}(\mu_l) \tag{27}$$

Proposition 5: If the congestion measure is additive, we have

$$\sum_{l' \in L} s_{l'} \geq \max_{l \in L} R_{l,F}^{-1}(\mu_l) \tag{28}$$

The proof of Proposition 4 and Proposition 5 is similar to the proof of Proposition 1 and Proposition 2.

Remark 3: An upper bound and a lower bound of the total congestion level can be given by

$$\max_{l \in L} R_{l,F}^{-1}(\mu_l) \leq \sum_{l' \in L} s_{l'} \leq \sum_{l \in L} R_{l,F}^{-1}(\mu_l) \tag{29}$$

Consider a new connection f^* to be established. Denote F^* be the new set of connections. Then

$$F^* = F \cup \{f^*\}$$

From the definition of $R_{l,F}$ above:

$$R_{l,F^*}(s) = R_{l,F}(s) + I(f^*,l)\, R_{f^*}(s)\, R_{l,F}(s) \tag{30}$$

Since R_{l,F^*} and $R_{l,F}$ are strictly decreasing functions, then

$$\max_{l \in L} R_{l,F}^{-1}(\mu_l) \leq \sum_{l' \in L} s_{l'} \leq \sum_{l \in L} R_{l,F}^{-1}(\mu_l) \tag{31}$$

Inequality (31) and Remark 3 imply that in case of different rate functions the upper bound and lower bound of the total congestion level also increase when a new connection is set up.

References

1. T. J. Ott, Kemperman and M. Mathis, The stationary behavior of ideal TCP congestion avoidance, ftp:// ftp. bellcore. com/ pub/ tjo/ TCPwindow. ps, August, 1996.
2. AM. Mathis, J. Semke, J. Mahdavi, and T. Ott, The macroscopic behavior of the TCP congestion avoidance algorithm, Communication Review, vol. 27, July 1997.
3. J. Padhye, V. Firoiu, D. Towsley and J. Kurose, Modeling TCP Throughput: A Simple Model and its Empirical Validation, Proceedings of SIGCOMM'98, Vancouver, August 1998.
4. G. Hasegawa, M. Murata, and H. Miyahara, Fairness and Stability of Congestion Control Mechansisms of TCP, In Proc. of INFOCOM '99, pages 1329 1336, March 1999.
5. D. D. Luong, J. Biro , On the Proportional Fairness of TCP Vegas, accepted to Globecom 2001.
6. D. S. Brakmo, S. W. O'Malley and L. L. Peterson, TCP Vegas: New techniques for congestion detection and avoidance, Proc. of SIGCOMM '94, 1994.
7. S. Floyd, M. Handley, J. Padhye, and J. Widmer, Equation-Based Congestion Control for Unicast Applications, Proc. of SIGCOMM '2000, August 2000
8. W Richard Stevens, TCP/ IP Illustrated, Volume 1, The Protocols, Addison-Wesley, July 1997.
9. F. P. Kelly, A. K. Maulloo and D. K. H. Tan, Rate control in communication networks: shadow prices, proportional fairness and stability, Journal of the Operational Research Society 49 (1998), 237-252.
10. L. Massouli, J. Roberts, Bandwidth sharing: objectives and algorithms, In Proc. of INFOCOM 1999.
11. S. Shenker, Fundemental design issues for the future Internet, IEEE Journal Selected Areas Communication, 13,1176-1188.
12. S. H. Low and D. E. Lapsley, Optimization Flow Control, I: Basic Algorithm and Convergence, IEEE/ ACM Transactions on Networking, 7(6): 861-75, Dec. 1999.
13. S. Athuraliya and S. H. Low, Optimization Flow Control, II: Implementation, http:// netlab. caltech. edu/ pub. html,
14. J. Mo and J. Walrand, Fair end-to-end window-based congestion control, IEEE/ ACM Transactions on Networking, vol. 8, no. 5, Oct. 2000
15. S. Kunniyur and R. Srikant, End-to-End Congestion Control Schemes: Utility Functions, Random Losses and ECN Marks, In Proc. of INFOCOM 2000.
16. K. Kar, S. Sarkar and L. Tassiulas, A Simple Rate Control Algorithm for Maximizing Total User Utility, In Proc. of INFOCOM 2001.
17. R. W. Pike, Optimization for engineering systems, Van Nostrand Reinhold Company Inc., New York 1986.
18. Q. Ma, P. Steenkiste and H. Zhang, High-bandwidth Traffic in Max-min Fair Share Networks, In Conference proceedings of ACM SIGCOMM'96, pp 206-217, Stanford, California, August, 1996
19. Q. Ma and P. Steenkiste, On Path Selection for Traffic with Bandwidth Guarantees, In the Proceedings of IEEE Interna-tional Conference on Network Protocols, to appear, Atlanta, Georgia, October 1997
20. R. L. Carter, M. Crovella, On the network impact of dynamic server selection, Computer Networks 31(23-24): 2529-2558 (1999)

Outsourcing in Poland – a Case Report

Marec Béla Steffens[1]

Paper presented at the European Regional Meeting of the International Telecommunications Society, Dublin, 1-3 Sept. 2001 and at the Economic Meeting of the Deutsche Gesellschaft für Osteuropakunde (German Society for Eastern European Studies), Berlin, 25-27 Nov. 2001; in English forthcoming in Osteuropa-Wirtschaft

1. How it Started: Siemens and ZWUT

Siemens AG acquired the traditional Polish telecommunications manufacturer ZWUT S.A. in 1993 [7., 8]. The relationship had started with a joint venture called CEWIS, founded in 1991 with 49 % Siemens shares. Soon thereafter the Polish government decided that the aim should no longer be to create joint ventures as islands within the economy. As role models, the joint ventures looked attractive, but the large state-owned companies did not have the resources to follow the example. So the government went for a comprehensive solution and decreed that only those foreign companies could qualify as suppliers for the national telecommuni-cations network who would take an active part in the privatisation of the industry. That is, the foreign company had to buy one of the local manufacturers, to commit large investments and to start production.

Through this process, after due diligence and public tendering, Siemens AG bought ZWUT in Warsaw together with Elwro, a Wroclaw-based computer manufacturer, since these companies were offered in double-pack. Meanwhile Siemens has invested some 120 million USD in ZWUT who became the market leader among telecommunications networks manufacturers in Poland. Elwro was sold in 2000.

Already before ZWUT was acquired by Siemens, the company had reduced its workforce from 4,500 to 2,350. It was clear from the beginning that for manufacturing, selling and servicing the Siemens product, the state-of-the-art public switching system EWSD, even fewer people would be needed. It was a privatisation condition that no employees were to be dismissed during the first two years after acquisition, though rather generous premiums were offered to those who resigned voluntarily.

Another means to reduce the headcount of the acquired company was outsourcing. This term had just then become fashionable in the West, starting with Eastman Kodak in 1989 (Wißkirchen 1999a: V), but in Eastern and Central Europe there were even more opportunities for it. Due to the chronic shortages in a socialist economy, companies had always striven for autarky: every task that was performed within the own organisation reduced the dependency on the insecure outside world. Thus a company like ZWUT harboured several activities that could rather easily be brought into the form of an independent company.

In a couple of cases, ZWUT employees and junior managers appreciated the opportunity to venture into capitalism. ZWUT, then called ZWUT S.A. a Siemens Company, supported them with loans and with the sale of used equipment at favourable prices. Some of these companies do the service for the old ZWUT-installed base, others produce accessories like Main Distribution Frames, racks or tools, still another one is a master partner for Siemens business and enterprise telecommunications solutions, etc. Even trade union representatives turned into entrepreneurs when they, through public offering, purchased ZWUT's recreation facility.

In other cases, buildings and machinery were sold and employees transferred to external companies

[1] Dr Marec Béla Steffens, .arec-bela.steffens@siemens.pl. Thanks to Andreas Hoffmann, CFO then of ZWUT and now of Siemens Sp. z o.o., for helpful comments.

from other branches of the economy. The number of ZWUT employees fell to below 500 in 1995. Due to the success in business, it rose again to about 825, with a better age and education structure than ever before. All this went without strikes because the trade unions realised there was no alternative.

2. Background Reasons for Outsourcing

After the achievements of a heroic start phase, ZWUT in 2000 had, in Bertolt Brecht's words, to face the toils of the plain before in 2001 the sectoral crisis reached Poland. In order to bundle central functions, ZWUT was per 1 Oct. 01 integrated with Siemens Sp. z o.o. (= Ltd.), the Siemens regional company in Poland. Now the telecommunication networks business is under one organisational roof with the other in-formation and communication activities and in fact most of Siemens business in Poland. At his first visit to Siemens Telefongyár, the Hungarian pendant to ZWUT, the author was astonished to see the cleaners wearing tags with the name of a different company and asked why Siemens Telefongyár with its then huge staff numbers had to hire outsiders for cleaning. However, the cleaners were ex-Telefongyár workers involved in an outsourcing project supported by the management.

Outsourcing is a hot topic still, backed now by the general fashion which made it a buzzword almost as mighty as e-commerce. Headquarters, not content with profit figures reported to them quarter by quarter, also bother a lot with how the profit is generated. One of the nicest figures, much easier to grasp than, say, EBITDA or EVA,[2] is headcount. It can easily be compared among countries (if one neglects the little fact that one head in Bangladesh does not cause the same cost as in Switzerland).

But, polemics aside, outsourcing is more than just a possibility to impress headquarters with a reduced headcount figure. Managers (and employees) cannot be experts on all things, and it is worthwhile to consider which are a company's core functions and which are supporting only. Not an easy task to define this. Schneider (2000: 167) recalls the case when IBM outsourced the development of an operating system for personal computers to a then insignificant company called Microsoft. Plus, the term core business is too often used as an excuse for poor business.

Nevertheless, there are functions which are necessary but not part of the competence that gives a company a competitive advantage on the market, and these tasks can be transferred to an outsourcing partner whose core competencies lie in the respective areas. This helps the company to focus management attention on the crucial areas and to benefit from the partner's experience.

3. Outsource What and to Whom?

In the years 2000-01, ZWUT S.A. a Siemens Company, with the personal involvement of this author in all but one, completed 10 outsourcing projects affecting 51 employees. Almost all divisions of the company were affected: sales and service, general administration and information technology management. Rather simple tasks were outsourced as well as quite complex ones.

The outsourcing partners are all kind of companies: one is another Siemens entity, Siemens Business Services; Sirti Poland is the subsidiary of an Italian company experienced in sub-contracting and project execution. Prime Car Management (PCM, today a subsidiary of General Motors but not at the time when we started outsourcing) and Patron-1 are car rental companies who took over our staff in the framework of a car pool leasing scheme. Solid Security is what the name suggests, so doing reception and internal mail distribution was an extension of their scope close to their main area of services. Simplicom on the other hand is a company founded by two former junior managers of ZWUT: we realised that with the cost structure of ZWUT we could not be competitive in the segment of telecommunications services for small businesses and enterprises. Now there is a vehicle with a better cost structure and flexibility, and by outsourcing to Simplicom a couple of fixed network installers as well we created a stable stream of revenues most welcome in the start-up phase. Last but not least, one outsourced employee founded a one-man company.

In all cases we took care not to diminish our competence in the eye of our customers and not to create units that could compete with us in the market. Where the outsourcing partners were to interact with our customers, which was the case in Simplicom's sales of small telecommunications systems to business and enterprise customers, we would coordinate the activities and accordingly withdraw our own sales staff.

It was not always easy to decide what is a core competence. Managers who are made responsible for their teams' results tend to keep things and people under their direct control. Sometimes they had to be convinced that outsourcing would give them no worse grip of the people doing the work – at times even a better one because an outsourcing partner fearing to lose the business can be a better address for complaints than middle management within the company. In the end we said that wherever services are offered in the market at good quality and lower

2 EBITDA = Earnings Before Interest, Tax, Depreciation and Amortisation; EVA = Economic Value Added, that is, after-tax profit after imputed cost of equity and actual cost of loans.

In other words, the estimation becomes more accurate when the networks are more heavily loaded. As shown in Fig. 5, the relative errors get closer to the line $y = 0$ when the number of active flow gets larger. The straight lines in this figure are also the linear fits and we can see that they come closer and closer to the the line $y = 0$.

C. 3 Path selection

As mentioned in Subsection III-B, if each flow f has a utility function U_f, then the control protocol with rate functions $U_f'^{-1}$ maximizes the overall utility $\Sigma_{f\in F}U_f(\lambda_f)$ of all flows. In that case, the path of each flow is already selected. If we consider the optimization on both routing and rate distribution level, the overall utility is a good measure to evaluate the path selection algorithms for elastic application.

We use the utility function $U(\lambda) = 5log(\lambda)$ for all flows. Then the corresponding rate func-tion is $R(s) = \frac{5}{s}$. Three path selection algorithms are evaluated: minimum hop count, Shortest-Widest path and Widest-Shortest path. As we can see in Fig. 6, the Shortest-Widest path and Widest-Shortest path have nearly equal overall utilities and both of them outperform the min-imum hop count algorithm. In a heavily loaded network, the Widest-Shortest path is the best choice to achieve efficient resource utilization (Fig. 7).

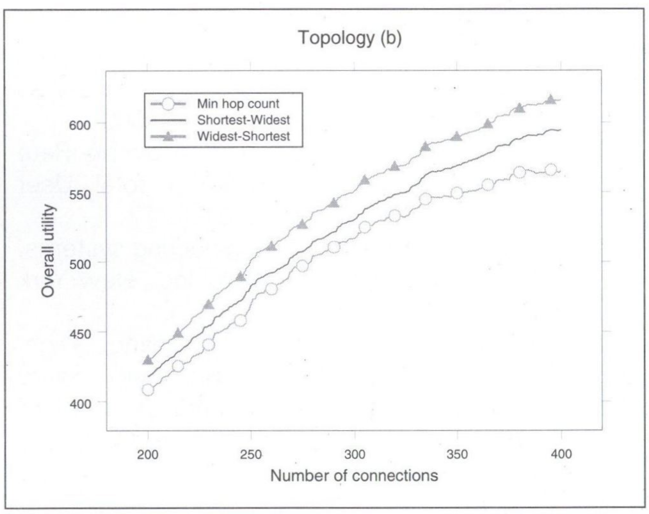


Fig. 7. Path selection algorithms in heavily loaded network-Topology (b)

VI. Conclusion

This paper generalizes and investigates a common characteristic of end-to-end congestion control protocols: the congestion measure. We have shown an important property of the con-gestion measure: the population effect, where the total congestion level tends to increase when new connections are established. If the congestion measure is the loss rate or queuing delay, the population effect could lead to the network performance degradation. This observa-tion promotes the use of ECN-like marking mecha-nisms.

Up to the present, the congestion measures have been only used in congestion control proto-cols. Our opinion is that the congestion measures, presenting loading state of the network, can be applied in other areas of networking. In this paper, we have proposed two applications:

1. The available unused bandwidth is not a sufficient information to predict the performance of elastic applications. We have proposed the use of end-to-end congestion level to estimate the future transfer rate for best effort applications.
2. We have also shown that congestion measure can be used as a cost of links in dynamic path selection.

Simulation results have confirmed that the elastic bandwidth estimation using end-to-end con-gestion level is fairly accurate when the network path contains several competing flows. It has been also shown by simulation that path selection using congestion measure as a link's cost outperforms the static shortest path algorithm.

Appendices

I. Some Terminoligies

Congestion measure: a measure to represent the level of congestion. Some examples of conges-tion measure are the loss rate, ECN marking probability and the queuing delay.

Total congestion level: is the sum of congestion levels of all network links.

Rate function R_f: is a mapping from the end-to-end congestion level s_f to the transmission rate λ_f of a flow f

$\lambda_f = R_f(s_f)$

Additive congestion measure: A congestion measure is additive if the end-to-end congestion level experienced by a flow is equal to the sum of congestion level at all links along the path of that flow.

Population effect: a phenomenon where the congestion level of links tends to increase with the number of active connections.

Elastic available bandwidth: the expected bandwidth available for a certain elastic application.

II. A general upper bound and lower bound for total congestion level

When the rate functions are different, we introduce the following functions:

$$R_{l,F}(s) = \sum_{f\in F} I(f,l)R_f(s)$$

Note that $R_{l,F}$ is a strictly decreasing function because of the same property of F_f.

Proposition 4: If the congestion measure is additive, we have

$$s_l \le R^{-1}_{l,F}(\mu_l) \quad (27)$$

Proposition 5: If the congestion measure is additive, we have

$$\sum_{l' \in L} s_{l'} \ge \max_{l \in L} R^{-1}_{l,F}(\mu_l) \quad (28)$$

The proof of Proposition 4 and Proposition 5 is similar to the proof of Proposition 1 and Proposition 2.

Remark 3: An upper bound and a lower bound of the total congestion level can be given by

$$\max_{l \in L} R^{-1}_{l,F}(\mu_l) \le \sum_{l' \in L} s_{l'} \le \sum_{l \in L} R^{-1}_{l,F}(\mu_l) \quad (29)$$

Consider a new connection f^* to be established. Denote F^* be the new set of connections. Then

$$F^* = F \cup \{f^*\}$$

From the definition of $R_{l,F}$ above:

$$R_{l,F^*}(s) = R_{l,F}(s) + I(f^*,l)\, R_{f^*}(s)\, R_{l,F}(s) \quad (30)$$

Since R_{l,F^*} and $R_{l,F}$ are strictly decreasing functions, then

$$\max_{l \in L} R^{-1}_{l,F}(\mu_l) \le \sum_{l' \in L} s_{l'} \le \sum_{l \in L} R^{-1}_{l,F}(\mu_l) \quad (31)$$

Inequality (31) and Remark 3 imply that in case of different rate functions the upper bound and lower bound of the total congestion level also increase when a new connection is set up.

References

1. T. J. Ott, Kemperman and M. Mathis, The stationary behavior of ideal TCP congestion avoidance, ftp:// ftp. bellcore. com/ pub/ tjo/ TCPwindow. ps, August, 1996.
2. AM. Mathis, J. Semke, J. Mahdavi, and T. Ott, The macroscopic behavior of the TCP congestion avoidance algorithm, Communication Review, vol. 27, July 1997.
3. J. Padhye, V. Firoiu, D. Towsley and J. Kurose, Modeling TCP Throughput: A Simple Model and its Empirical Validation, Proceedings of SIGCOMM'98, Vancouver, August 1998.
4. G. Hasegawa, M. Murata, and H. Miyahara, Fairness and Stability of Congestion Control Mechansisms of TCP, In Proc. of INFOCOM '99, pages 1329 1336, March 1999.
5. D. D. Luong, J. Biro , On the Proportional Fairness of TCP Vegas, accepted to Globecom 2001.
6. D. S. Brakmo, S. W. O'Malley and L. L. Peterson, TCP Vegas: New techniques for congestion detection and avoidance, Proc. of SIGCOMM '94, 1994.
7. S. Floyd, M. Handley, J. Padhye, and J. Widmer, Equation-Based Congestion Control for Unicast Applications, Proc. of SIGCOMM '2000, August 2000
8. W Richard Stevens, TCP/ IP Illustrated, Volume 1, The Protocols, Addison-Wesley, July 1997.
9. F. P. Kelly, A. K. Maulloo and D. K. H. Tan, Rate control in communication networks: shadow prices, proportional fairness and stability, Journal of the Operational Research Society 49 (1998), 237-252.
10. L. Massouli, J. Roberts, Bandwidth sharing: objectives and algorithms, In Proc. of INFOCOM 1999.
11. S. Shenker, Fundemental design issues for the future Internet, IEEE Journal Selected Areas Communication, 13,1176-1188.
12. S. H. Low and D. E. Lapsley, Optimization Flow Control, I: Basic Algorithm and Convergence, IEEE/ ACM Transactions on Networking, 7(6): 861-75, Dec. 1999.
13. S. Athuraliya and S. H. Low, Optimization Flow Control, II: Implementation, http:// netlab. caltech. edu/ pub. html,
14. J. Mo and J. Walrand, Fair end-to-end window-based congestion control, IEEE/ ACM Transactions on Networking, vol. 8, no. 5, Oct. 2000
15. S. Kunniyur and R. Srikant, End-to-End Congestion Control Schemes: Utility Functions, Random Losses and ECN Marks, In Proc. of INFOCOM 2000.
16. K. Kar, S. Sarkar and L. Tassiulas, A Simple Rate Control Algorithm for Maximizing Total User Utility, In Proc. of INFOCOM 2001.
17. R. W. Pike, Optimization for engineering systems, Van Nostrand Reinhold Company Inc., New York 1986.
18. Q. Ma, P. Steenkiste and H. Zhang, High-bandwidth Traffic in Max-min Fair Share Networks, In Conference proceedings of ACM SIGCOMM'96, pp 206-217, Stanford, California, August, 1996
19. Q. Ma and P. Steenkiste, On Path Selection for Traffic with Bandwidth Guarantees, In the Proceedings of IEEE Interna-tional Conference on Network Protocols, to appear, Atlanta, Georgia, October 1997
20. R. L. Carter, M. Crovella, On the network impact of dynamic server selection, Computer Networks 31(23-24): 2529-2558 (1999)

Outsourcing in Poland – a Case Report

Marec Béla Steffens[1]

Paper presented at the European Regional Meeting of the International Telecommunications Society, Dublin, 1-3 Sept. 2001 and at the Economic Meeting of the Deutsche Gesellschaft für Osteuropakunde (German Society for Eastern European Studies), Berlin, 25-27 Nov. 2001; in English forthcoming in Osteuropa-Wirtschaft

1. How it Started: Siemens and ZWUT

Siemens AG acquired the traditional Polish telecommunications manufacturer ZWUT S.A. in 1993 [7., 8]. The relationship had started with a joint venture called CEWIS, founded in 1991 with 49 % Siemens shares. Soon thereafter the Polish government decided that the aim should no longer be to create joint ventures as islands within the economy. As role models, the joint ventures looked attractive, but the large state-owned companies did not have the resources to follow the example. So the government went for a comprehensive solution and decreed that only those foreign companies could qualify as suppliers for the national telecommuni-cations network who would take an active part in the privatisation of the industry. That is, the foreign company had to buy one of the local manufacturers, to commit large investments and to start production.

Through this process, after due diligence and public tendering, Siemens AG bought ZWUT in Warsaw together with Elwro, a Wroclaw-based computer manufacturer, since these companies were offered in double-pack. Meanwhile Siemens has invested some 120 million USD in ZWUT who became the market leader among telecommunications networks manufacturers in Poland. Elwro was sold in 2000.

Already before ZWUT was acquired by Siemens, the company had reduced its workforce from 4,500 to 2,350. It was clear from the beginning that for manufacturing, selling and servicing the Siemens product, the state-of-the-art public switching system EWSD, even fewer people would be needed. It was a privatisation condition that no employees were to be dismissed during the first two years after acquisition, though rather generous premiums were offered to those who resigned voluntarily.

Another means to reduce the headcount of the acquired company was outsourcing. This term had just then become fashionable in the West, starting with Eastman Kodak in 1989 (Wißkirchen 1999a: V), but in Eastern and Central Europe there were even more opportunities for it. Due to the chronic shortages in a socialist economy, companies had always striven for autarky: every task that was performed within the own organisation reduced the dependency on the insecure outside world. Thus a company like ZWUT harboured several activities that could rather easily be brought into the form of an independent company.

In a couple of cases, ZWUT employees and junior managers appreciated the opportunity to venture into capitalism. ZWUT, then called ZWUT S.A. a Siemens Company, supported them with loans and with the sale of used equipment at favourable prices. Some of these companies do the service for the old ZWUT-installed base, others produce accessories like Main Distribution Frames, racks or tools, still another one is a master partner for Siemens business and enterprise telecommunications solutions, etc. Even trade union representatives turned into entrepreneurs when they, through public offering, purchased ZWUT's recreation facility.

In other cases, buildings and machinery were sold and employees transferred to external companies

[1] Dr Marec Béla Steffens, .arec-bela.steffens@siemens.pl. Thanks to Andreas Hoffmann, CFO then of ZWUT and now of Siemens Sp. z o.o., for helpful comments.

from other branches of the economy. The number of ZWUT employees fell to below 500 in 1995. Due to the success in business, it rose again to about 825, with a better age and education structure than ever before. All this went without strikes because the trade unions realised there was no alternative.

2. Background Reasons for Outsourcing

After the achievements of a heroic start phase, ZWUT in 2000 had, in Bertolt Brecht's words, to face the toils of the plain before in 2001 the sectoral crisis reached Poland. In order to bundle central functions, ZWUT was per 1 Oct. 01 integrated with Siemens Sp. z o.o. (= Ltd.), the Siemens regional company in Poland. Now the telecommunication networks business is under one organisational roof with the other in-formation and communication activities and in fact most of Siemens business in Poland. At his first visit to Siemens Telefongyár, the Hungarian pendant to ZWUT, the author was astonished to see the cleaners wearing tags with the name of a different company and asked why Siemens Telefongyár with its then huge staff numbers had to hire outsiders for cleaning. However, the cleaners were ex-Telefongyár workers involved in an outsourcing project supported by the management.

Outsourcing is a hot topic still, backed now by the general fashion which made it a buzzword almost as mighty as e-commerce. Headquarters, not content with profit figures reported to them quarter by quarter, also bother a lot with how the profit is generated. One of the nicest figures, much easier to grasp than, say, EBITDA or EVA,[2] is headcount. It can easily be compared among countries (if one neglects the little fact that one head in Bangladesh does not cause the same cost as in Switzerland).

But, polemics aside, outsourcing is more than just a possibility to impress headquarters with a reduced headcount figure. Managers (and employees) cannot be experts on all things, and it is worthwhile to consider which are a company's core functions and which are supporting only. Not an easy task to define this. Schneider (2000: 167) recalls the case when IBM outsourced the development of an operating system for personal computers to a then insignificant company called Microsoft. Plus, the term core business is too often used as an excuse for poor business.

Nevertheless, there are functions which are necessary but not part of the competence that gives a company a competitive advantage on the market, and these tasks can be transferred to an outsourcing partner whose core competencies lie in the respective areas. This helps the company to focus management attention on the crucial areas and to benefit from the partner's experience.

3. Outsource What and to Whom?

In the years 2000-01, ZWUT S.A. a Siemens Company, with the personal involvement of this author in all but one, completed 10 outsourcing projects affecting 51 employees. Almost all divisions of the company were affected: sales and service, general administration and information technology management. Rather simple tasks were outsourced as well as quite complex ones.

The outsourcing partners are all kind of companies: one is another Siemens entity, Siemens Business Services; Sirti Poland is the subsidiary of an Italian company experienced in sub-contracting and project execution. Prime Car Management (PCM, today a subsidiary of General Motors but not at the time when we started outsourcing) and Patron-1 are car rental companies who took over our staff in the framework of a car pool leasing scheme. Solid Security is what the name suggests, so doing reception and internal mail distribution was an extension of their scope close to their main area of services. Simplicom on the other hand is a company founded by two former junior managers of ZWUT: we realised that with the cost structure of ZWUT we could not be competitive in the segment of telecommunications services for small businesses and enterprises. Now there is a vehicle with a better cost structure and flexibility, and by outsourcing to Simplicom a couple of fixed network installers as well we created a stable stream of revenues most welcome in the start-up phase. Last but not least, one outsourced employee founded a one-man company.

In all cases we took care not to diminish our competence in the eye of our customers and not to create units that could compete with us in the market. Where the outsourcing partners were to interact with our customers, which was the case in Simplicom's sales of small telecommunications systems to business and enterprise customers, we would coordinate the activities and accordingly withdraw our own sales staff.

It was not always easy to decide what is a core competence. Managers who are made responsible for their teams' results tend to keep things and people under their direct control. Sometimes they had to be convinced that outsourcing would give them no worse grip of the people doing the work – at times even a better one because an outsourcing partner fearing to lose the business can be a better address for complaints than middle management within the company. In the end we said that wherever services are offered in the market at good quality and lower

2 EBITDA = Earnings Before Interest, Tax, Depreciation and Amortisation; EVA = Economic Value Added, that is, after-tax profit after imputed cost of equity and actual cost of loans.

Area	Outsourcing to	Transfer date	Out-sourced staff
Reception	Solid Security	1 Aug. 00	1
Internal mail	Solid Security	1 Aug. 00	1
Service: mobile networks installers	Sirti	1 Aug. 00	7
Service: fixed networks installers	Simplicom	15 Sept. 00	13
IT network administration	Siemens Business Services	29 Sept. 00	10
Driver	Self-employer	29 Sept. 00	1
Business and Enterprise sales & service	Simplicom	div. in 00/01	7
Car pool	PCM and Patron 1	1 Dec. 00	2
SAP administration	Siemens Business Services	1 July 01	8
Service: documentation centre	Simplicom	1 Aug. 01	3
Total			**51**

prices than our cost, it shows that this cannot be one of our core competencies. At the same time, we must keep competence enough to control the outsourcing partners and their work results. Otherwise we would be in danger, as the customers rightfully make us responsible for all sub-contractors' and outsourcing partners' performance. Accordingly, to bring an example from service, all installation jobs were outsourced except for the latest technologies, but we kept the supervisors as well as the commissioning and trouble-shooting specialists.

Quite from the beginning the top management made clear that cost savings were one objective of outsourcing, but not the only one and rather in medium-term perspective. Companies who rely on outsourcing as a means of radical cost-cutting often overlook that a significant part of the cost will remain: residual overhead cost, transaction cost and control cost of outsourcing. They should never be underestimated (Müller/Pran-gen-berg 1997: 34 ff.).

4. The Staff Side

Acceptance of outsourcing within the affected staff was remarkably high: they would have preferred to stay at ZWUT as before but realised that their tasks had no future within the company. They acknowledged that we ensured them quite fair outsourcing conditions. We were aware that outsourcing can also be a result of dissatisfaction with the level of in-house services and might therefore be regarded as some kind of sanction by the employees concerned – an approach not likely to result in successful outsourcing (Horchler 1996: 197, Mocsnik 1999: 336). Therefore we stressed at every opportunity that we aimed at maintaining the service and knowledge of the people concerned even if they would no longer be on our payroll. This was under-stood and even praised by the outsourcing partners.

The outsourced employees are now in an environment where their specific task is a core competence. They can exchange experience with others working in the same field, and they are valuable to their new employer because they know best the specific requirements of their previous employer – the new one's valuable customer. And in case the previous employer has a reduced demand for the outsourced services – seasonally or permanently – the outsourcing company will usually have a number of customers and thus be able to shift personnel between them. An additional advantage is that it is the outsourcing partner who has to deal with illness, vacation and even insufficient performance up to potential labour disputes. Of course this will affect the price as well.

As we wanted to have the employees supporting the outsourcing scheme, we con-tractually required the new employer to give a job guarantee for twelve months at the same conditions as were offered by ZWUT. These would include fringe benefits like 13th salary, an additional bonus, a salary increase at least once a year, private leave, training, participation in a medical scheme or a compensation payment for this, etc. We even required the new employers to have the employment contracts checked by us to make sure that our transferred staff would be treated in a fair manner. If the cost for transferred personnel were thus higher than for comparable staff otherwise employed by the outsourcing partner, we accordingly accepted the price for the outsourced services to be higher than on the market – at least for the initial twelve months. At last, by using trained employees accustomed to working for us, the quality should be higher than market average.

Some issues, however, we could not demand to be settled for our staff in the new environment in the same manner as they are in ZWUT: for travel, car and mobile phone regulations, the transferred employees will have to follow the standards of the new company. It would raise bad feelings against them if they for example stayed in a different hotel than their colleagues working on the same installation job. But we checked the applicable regulations of the outsourcing partner and excluded such companies from the scheme whose regulations were far below our level. And in order to compensate the employees for certain losses and to encourage their participation, we paid a special outsourcing bonus of three monthly salaries to every transferred employee (except to those transferred to another Siemens subsidiary) at the time of transfer. All this means we de facto applied Art. 23-1 of the Polish Labour Code in a rather extensive interpretation although legal analysis was equivocal as to whether this article would apply to our outsourcing cases at all.[3]

By limiting the job guarantee to twelve months, we copied the law, and indeed we did not want to limit our flexibility – which in fact is a main advantage of outsourcing – by giving a guarantee for a longer period. And it must be considered that an employee not affected by outsourcing at all does not have such guarantee either, not even for twelve months. According to the Polish Labour Law, he or she can always be dismissed at three months' notice with a compensation depending on the length of the working career, amounting to three monthly salaries maximum.

On the other hand, even if we avoided a binding commitment, we sent a strong signal to the new employers. In the service contracts where we gave the outsourcing partners twelve months' work for the transferred employees we stated that for considering future suppliers we would regard it favourably if they kept our employees for longer. We want the outsourcing partners to regard the transferred employees as a key to future orders from us.

Much has been written about information policy before and during outsourcing. It is certainly bad when rumours are floating around while no official information is available. On the other hand, before official statements are made, two preconditions must be fulfilled. First, it must be made sure that the management speaks with one voice, that outsourcing is not seen as a hobbyhorse of Finance and Business Administration but understood as a measure that will help the whole company. Second, already in the first information round there has to be a clear concept for the affected employees. Telling them "you will be outsourced but I cannot say what this means for you" will not help at all.

5. Defining the Concrete Tasks

Some outsourcing projects covered tasks which are quite easy to define, like reception and internal mail distribution. In other cases, we would simply hire installers to work under the instructions of our supervisors. Then the outsourcing partner would not bear the responsibility for the work results – except, of course, in case the staff did not follow the supervisor's instructions. Then we had a project were a complete task was outsourced, namely the installation of Base Transceiver Stations for a GSM network. In this case the outsourcing partner had performed the same task with its own crews for some time before. So the task did not have to be defined anew and the division of responsibilities was clarified in the contract.

The situation was quite different in IT services where we outsourced complex tasks not described before. Our outsourcing partner here was Siemens Business Services (SBS). We did two separate outsourcing projects with them: one for the operation and management of our IT network, and the other for SAP administration. The structure was in both cases the same. SBS and ZWUT staff did an extensive due diligence and compiled detailed Service Level Agreements. The business units should know what services they are entitled to and what cost are associated with that. Having to pay real money to an external party makes people think twice whether a certain service is actually needed for the business or just 'nice to have.'

In fact, it is a new policy now within ZWUT that all central departments are to conclude Service Level Agreements with the business units as their customers. But with an outsourcing partner it is even easier to define effective sanctions, such as penalties. Such penalties are not an aim in itself: being serviced at the desired level is better for the company's business than cashing in a penalty. But they help making clear that underperformance will not be taken lightly. And, as generally in negotiating contracts, it is easier to define consequences in advance on an abstract level than when the situation already occurred (Steffens 1998).

During the six-month transition period, penalties applied at a reduced rate only. We also used this transition period for necessary adaptations to the Service Level Agreements. It is helpful to leave room for such adaptations, but if the concept is too generally vague, outsourcing will not work at all. In fact it is one of the dangers and of the beauties of outsourcing that it forces us to do our homework. [1., 10.] To neglect this homework because the area will be outsourced anyway is the worst thing to do, for how can the outsourcing partners know if we do not tell them? It is like working on the computer – we get from the computer what we key in, not what we had in mind. Likewise, we get from the outsourcing company what

[3] Art. 23-1 deals with the transfer of "an employment establishment or a part thereof" to another employer, but it is unclear what constitutes "a part thereof" and what does not. – Whereas the legal side does not reveal much of Poland's Solidarnosc tradition, the courts not unfrequently do, and due to the frequent legal changes and the very limited number of reference court cases the results of a legal case in Poland appear to be even less predictable than elsewhere.

we wrote into the contract, not what we diffusely wanted. Or, if we do not get what is written in the contract, it should be easy to sanction the outsourcing partner, but blaming them for what is not written down will not work at all. A lack of clarity here can lead to a complete failure of the project, even to "backsourcing".

For example, for us it was clear that we wanted to outsource complete processes – the employees and the work done by them. In most cases this was well understood, but SBS naturally did not know our processes as well as we did. So, wherever our expectations were not clearly written in the contract, we faced a problem and had to renegotiate to make sure that the outsourced employees would continue to perform the complete process e.g. of ordering personal computers as they had done within ZWUT. Without that, we would have had to build up a double structure which of course would have been counterproductive.

6. Due Diligence and Prices

In all outsourcing projects, pricing was developed in a transparent manner. In the IT outsourcing case with SBS, for example, we agreed on what would be part of a Standard Workplace Office. Whatever was not part of this standard configuration would be charged separately. Cost were laid open: actual personnel cost, deprecia-tion and maintenance fees paid to third parties were documented. Future cost were estimated in mutual agreement, for example replacement for those staff members who were not outsourced but were involved in the relevant tasks for part of their time. We would not outsource all IT staff because we needed some strategic competence to represent our users towards the service provider SBS and to determine the future development path for our IT landscape. We kept such competence in every outsourcing area, because otherwise we would become com-pletely dependent on the outsourcing partner. Not at all can outsourcing make managerial control obsolete. [2.]

Even the margins were openly discussed with all outsourcing partners. This was not difficult because both sides knew the cost and could calculate the revenue from the outsourcing fees. We made it clear that we would accept a reasonable margin, but we would rather stop an outsourcing project than accept too high a margin.

While our other outsourcing contracts have an initial lifetime of one year, in the SBS projects we agreed on three years because of the substantial investment SBS had to make. However, continued unsatisfactory performance after certain warning periods would allow us to terminate prematurely.

For due diligence, we found a fair compromise regarding the liability for possible errors: we as ZWUT took the responsibility for all material we provided and all answers we gave. SBS as outsourcing partner could however not make us liable for wrong processing of our data or for any questions they may have failed to ask.

Of course we could not do without working assumptions which would then be verified within the transition phase. Here again it shows how necessary it is to do one's homework beforehand. To enter into IT outsourcing with contradictory inventory records complicates the task considerably, because the outsourcing partner has to base capacity and revenue planning on doubtful assumptions. What is important here is to state the following in the contract: whatever is found during the transition phase, or whatever changes are required later, this may be a reason for price changes according to the principles of the contract, but does not give the outsourcing partner the right to abandon the contract.

E.g. prices can be based on the assumption that there is a certain number of work stations. But this number can drastically change, for example in case of mergers or carve-outs. Then prices may change, and the outsourcing partner may be granted a certain time before the same service level can be guaranteed for a much higher number of workstations. But the increase or decrease should not be a reason to terminate the contract or to renegotiate it from scratch.

As part of the deal, we sold our IT network infrastructure to SBS. Not personal computers, though: due to technical progress they are frequently exchanged, and if SBS had to take this risk, service prices would go up accordingly. For the network infrastructure, SBS included the depreciation in their cost calculation, and SBS is obliged to make the necessary investment in order to give ZWUT the required service level. We do not any longer deal with what is technically necessary, we pay for getting a specific service level, and SBS has to buy and to do whatever is necessary to achieve this. We assumed that the investment level will tend to be stable: decreasing prices will be matched by increasing functional and capacity demands. Nevertheless we agreed on an annual review: if new investment is higher or lower than depreciation, SBS prices will be adjusted accordingly.

For the sales price, we used independent experts who established a market price. Selling at book value would have brought us in conflict with the arm's length principle, because SBS and ZWUT are affiliated

[4] Wißkirchen (1999b: 288 f.) argues that non-core processes are typically inefficient because of rather low management attention – optimising them can therefore be an alternative to outsourcing. Or, as Fischer (2000: 101) puts it: if you immediately expect from outsourcing a significantly better performance at a considerably reduced price, you have done something wrong in the past.

companies. In sales to external parties, book values may be used. In any case, as desirable as the immediate cash-flow effect may be (Horchler 1996: 65 f.), it does not help to go for a high sales price because it would of course be reflected in the outsourcing fee.

7. The Location Poland

Was our task made more difficult by the fact that the location was Poland? The answer is a clear no. It was no problem for us to find qualified outsourcing partners. The legal side, in spite of rapid changes in Polish legislation and a high degree of uncertainty regarding official interpretation of the law, did not provide particular obstacles. Employees were cooperative. Intercultural issues may arise, but in outsourcing processes not more often than in any other part of the business. Thus, out-sourcing even in a complex areas like telecommunications and IT, can be done in Poland just as well as elsewhere, and no other but the general caveats of outsourcing need to be observed.

References

1. Fischer, Thomas (2000): Mangelhafte Planung lässt Outsourcing-Projekte scheitern (Insufficient Planning Causes Outsourcing Projects to Fail), in: Computerwoche 44 No. 27, p. 101 f.
2. Hodel, Marcus (1999): Outsourcing-Management kompakt und verständlich (O. M. concise and understandable), Braunschweig/Wiesbaden: Vieweg
3. Köhler-Frost, Wilfried (1999): Evaluierung und Auswahl des problemgerechten Outsourcing-Partners – Erfolgsfaktoren, Vorgehensweisen, Markttransparenz, (Evaluation and Selection of the Problem-Adequate Outsourcing Partner – Success Factors, Procedures, Market Transparency), in: Wißkirchen (1999a), S. 187-204
4. Mocsnik, Johann (1999): Outsourcing der Logistik einer Brauerei – eine Reise in die Zukunft unter besonderer Berücksichtigung von Mensch und Strategie (Outsourcing the Logistics of a Brewery – a Journey into Future under Special Consideration of Man and Strategy), in: Wißkirchen (1999a), S. 317-342
5. Schneider, Thomas (2000): Kernkompetenzen und Outsourcing (Core Competencies and Outsourcing), in: Controller-Magazin 25, 166-170
6. Steffens, Marec Béla (1997): Modernizing Telecommunications in Central and Eastern Europe. A Business Perspec-tive, in: Welfens, Paul J J/Yarrow, George (eds.): Telecommunications and Energy in Systemic Transformation, Berlin etc.: Springer, pp. 197-212
7. Steffens, Marec Béla–Sundrum, Lisa (1993a): Joint Ventures in der Telekommunikations-industrie – Erfahrungs-bericht eines Groß-unternehmens in Osteuropa (Joint Ventures in the Telecommunications Industry – A Large Company's Experience in Eastern Europe), in: Osteuropa-Wirtschaft 38, pp. 47-55, in Hungarian in: Magyar Távközlés 4 (1993) No. 11, pp. 9-12
8. Steffens, Marec Béla–Sundrum, Lisa (1993b): "Political Change Opens Telecommuni-ca-tion Markets, in: Telcom Report International 16 No. 4, pp. 19-22
9. Steffens, Marec Béla (1998): Verhandeln in China – Ein Leitfaden in 60 Punkten (Negotiating in China – A Guideline in 60 Points), in: Harvard Business Manager 20 No. 6, pp. 19-24, in English in: China Economic Review 9 (1999) No. 1, pp. 20 f., and No. 2, pp. 18 f., in Hungarian in: Magyar Távközlés 10 (1999) No. 3, pp. 54-58
10. Wißkirchen, Frank (ed., 1999a): Outsourcing-Projekte erfolgreich realisieren (Realising Outsourcing Projects Successfully), Stuttgart: Schäffer-Poeschel
11. Wißkirchen, Frank (1999b): Beurteilung der Vorteilhaftigkeit von Outsourcing unter Berücksichtigung von Prozeßkosten und Transaktionskosten (Assessment of the Advantageousness of Outsourcing in Consideration of Process and Transaction Cost), in: Wißkirchen (1999a), S. 283-313

Tartalom

Editorial Office
Communications and IT Sciences Association
HTE Budapest, Kossuth Lajos tér 6-8.
Phone: 00 36 353 1027 • Fax: 00 36 353 0451
e-mail: hte@mtesz.hu

Articles can be sent also to the following address
BME Microwave Communication Department
Budapest, Goldmann Gy. tér 3.
Phone: 00 36 463 1559 • Fax: 00 36 463 3289
e-mail: zombory@mht.bme.hu

Subscription
Communications and IT Sciences Association
HTE Budapest, Kossuth Lajos tér 6-8.
Phone: 00 36 353 1027 • Fax: 00 36 353 0451
e-mail: hte@mtesz.hu

Subscription rates in 2001
For Hungarian business customers HUF 30,000 per year (including VAT), for Hungarian residential customers
HUF 6000 per year (including VAT). For foreign subscribers 12 issues 150 USD, single copies 15 USD.

www.hte.hu

Publisher: MÁRIA MÁTÉ

Design by: Kocsis és Szabó Kft.

HU ISSN 0018-2028